

ESTADÍSTICA USANDO EXCEL



Preencha a **ficha de cadastro** no final deste livro e receba gratuitamente informações sobre os lançamentos e as promoções da Elsevier.

Consulte nosso catálogo completo, últimos lançamentos e serviços no site **www.elsevier.com.br**

ESTADÍSTICA USANDO EXCEL

© 2005, Juan Carlos Lapponi

Todos os direitos reservados e protegidos pela Lei nº 9.610 de 19/12/1998.
Nenhuma parte deste livro, sem autorização prévia por escrito da editora,
poderá ser reproduzida ou transmitida sejam quais forem os meios empregados:
eletrônicos, mecânicos, fotográficos, gravação ou quaisquer outros.

Editoração Eletrônica: Estúdio Castellani

Copidesque: Lígia Paixão

Revisão Gráfica: Roberto Mauro Facce e Carlos Maurício da Silva Neto

Projeto Gráfico

Elsevier Editora Ltda.

Conhecimento sem Fronteiras

Rua Sete de Setembro, 111 – 16º andar

20050-006 – Centro – Rio de Janeiro – RJ – Brasil

Rua Quintana, 753 – 8º andar

04569-011 – Brooklin – São Paulo – SP – Brasil

Serviço de Atendimento ao Cliente

0800-0265340

sac@elsevier.com.br

ISBN 978-85-352-1574-8

Nota: Muito zelo e técnica foram empregados na edição desta obra. No entanto, podem ocorrer erros de digitação, impressão ou dúvida conceitual. Em qualquer das hipóteses, solicitamos a comunicação ao nosso Serviço de Atendimento ao Cliente, para que possamos esclarecer ou encaminhar a questão.

Nem a editora nem o autor assumem qualquer responsabilidade por eventuais danos ou perdas a pessoas ou bens, originados do uso desta publicação.

CIP-Brasil. Catalogação na fonte.
Sindicato Nacional dos Editores de Livros, RJ

L322e

Lapponi, Juan Carlos

Estatística usando Excel / Juan Carlos Lapponi.

– Rio de Janeiro: Elsevier, 2005 – 8ª reimpressão.
il.

Inclui bibliografia

ISBN 978-85-352-1574-8

1. Excel (Programa de computador). 2. Estatística.

I. Título.

04-2744.

CDD — 005.369

CDU — 004.42

O Autor

JUAN CARLOS LAPPONI

Engenheiro pela Faculdade de Engenharia da Universidade de Buenos Aires e Doutor em Engenharia pela Escola Politécnica da Universidade de São Paulo. Professor dos cursos de MBA em Finanças Corporativas e MBA em Gestão Empresarial da FGV Management da Fundação Getulio Vargas e do MBA em Administração para Engenheiros do IMT–Instituto Mauá de Tecnologia.

Obras do Autor

Matemática Financeira com Aplicações em Microcomputadores e Planilha de Cálculo, Ebrás 1987.

As seguintes obras foram publicadas pela Editora Lapponi

Lotus 1-2-3 em Modelos para Avaliação Econômica de Projetos de Investimento, 1989.

Novas Funções Financeiras para Lotus 1-2-3, 1991.

Matemática Financeira Usando Excel, versão 4, 1993.

Matemática Financeira Usando Excel 4 e 5, 1994.

Estatística Usando Excel 4 e 5, 1995.

Matemática Financeira Uma Abordagem Moderna, terceira edição 1995.

Avaliação de Projetos de Investimento – Modelos em Excel, 1996.

Matemática Financeira Usando Excel 5 e 7, 1996.

Estatística Usando Excel 5 e 7, 1997.

Matemática Financeira, 1998.

Excel & Cálculos Financeiros – Introdução à Modelagem Financeira, 1999.

Estatística Usando Excel, 2000.

Todas as obras anteriores estão esgotadas.

A seguir a relação das obras atuais da Editora Lapponi.

Projetos de Investimento – Construção e Avaliação do Fluxo de Caixa, 2000.

Matemática Financeira Usando Excel – Como Medir Criação de Valor, 2002.

Edição da Editora Elsevier – Campus

Modelagem Financeira com Excel, Elsevier - Campus, primeira edição 2004.

Estatística Usando Excel, Elsevier - Campus, quarta edição 2005.

Prefácio

Estatística Usando Excel ensina Estatística e explica como aplicar os conceitos e analisar resultados por meio de exemplos resolvidos com os procedimentos tradicionais de cálculo e o Excel. Nesta quarta edição de *Estatística Usando Excel*, boa parte dos temas da edição anterior foi reescrita e ampliada, melhorando a sequência e a compreensão dos temas. Novos temas foram adicionados, ampliando os conceitos estatísticos como, por exemplo, a tabela de probabilidades conjuntas e totais, os cálculos inversos com a distribuição normal, o poder do teste de hipóteses, a análise de variância com dois fatores, a regressão linear múltipla, a projeção por ajuste de polinômio, taxa média e reta de regressão, a construção de um ou mais *boxplot* com os recursos próprios do Excel, e outros temas mais.

Nesta nova edição, os exemplos em sua maioria foram resolvidos integrando os conceitos, os procedimentos de cálculo e a análise dos resultados. No desenvolvimento dos exemplos, são explicados os conceitos, os procedimentos de cálculo utilizando as fórmulas, as tabelas estatísticas, as funções e as ferramentas de análise estatísticas do Excel e as planilhas, os modelos e os simuladores desenvolvidos pelo autor. Essa integração torna o aprendizado de *Estatística* eficaz e mostra que há diversos caminhos para alcançar o mesmo resultado, incluindo a superposição de recursos do Excel.

A maioria das planilhas, modelos e simuladores da edição anterior ganhou um novo layout a fim de facilitar a compreensão dos conceitos, a realização de cálculos e a análise dos resultados. Foram adicionadas novas planilhas, novos modelos e novos simuladores em Excel, por exemplo, os modelos *Amostragem sem reposição*, construção de *Histogramas* e *Análise estatística numérica* sem limitação de tamanho de amostra. Também, os modelos da *Distribuição amostral*, da visualização das *Propriedades da média*, do *Teste de hipóteses* com novo gráfico descritivo da decisão para cada um dos três procedimentos, do *Ajuste manual* da reta de regressão, do *Gráfico das distribuições* apresentadas com visualização contínua do comportamento da curva em função dos parâmetros, e outros modelos mais. Os simuladores do *Lançamento de uma moeda* com até 10.000 lançamentos, do *Coefficiente de correlação* visualizando sua relação com o gráfico dos pontos das amostras, do *Teorema central do limite* variando o tamanho da amostra e o número de repetições, do *Intervalo de confiança* etc. Ao mesmo tempo, o leitor aprende a utilizar recursos do Excel, por exemplo, a construção de gráficos e histogramas, o registro de uma função e de uma fórmula como matriz, as ferramentas de análise, os comandos *Atingir Meta*, *Linha de tendência*, *Solver* e outras funções e comandos como a *Formatação condicional* etc.

Nesta nova edição foi mantido o objetivo de facilitar o *autodesenvolvimento* do leitor por meio de exemplos resolvidos, integrando procedimentos de cálculo e mais problemas propostos com respostas. Temas complementares de *Estatística* e de Excel foram adicionados em apêndices de capítulos para não interferir no aprendizado. Alguns deles podem ser utilizados como trabalhos extras, por exemplo, as demonstrações de fórmulas, os modelos para combinação linear de variáveis aleatórias com aplicações em finanças, a variável aleatória do VPL de um investimento e a formação de uma carteira de investimento utilizando o *Solver*, a determinação dos coeficientes de regressão utilizando o *Solver* e outros mais.

Todas as planilhas, os modelos estatísticos e os simuladores desenvolvidos em Excel 2002, compatíveis com as versões 2000 e 2003, bem como parte dos exemplos e problemas resolvidos estão incluídos na página do livro, no site da Editora.

O conteúdo deste livro será útil para:

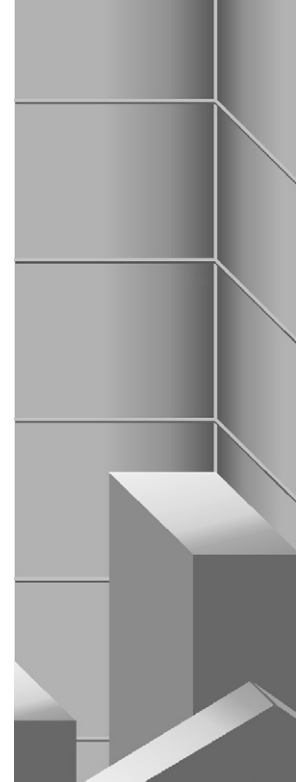
- Estudantes que cursam *Estatística* nas diversas áreas do conhecimento e em diferentes níveis de graduação como, em ordem alfabética, Administração, Biologia, Contabilidade, Economia, Engenharia, Finanças, Marketing, Medicina etc.
- Estudantes que necessitam aprimorar ou complementar seus conhecimentos de *Estatística* utilizando o Excel.
- Profissionais das diversas áreas que utilizam os conceitos de *Estatística* e necessitam, ou gostariam, de utilizar as funções estatísticas, as ferramentas de análise, planilhas, modelos e simuladores de estatística em Excel.
- Todos aqueles que poderão utilizar as planilhas, os modelos e os simuladores de estatística em Excel da forma como estão na página do livro, no site da Editora, ou modificando-os, para atender às suas necessidades.
- Alunos de áreas correlatas que utilizarão estatística e desejam antecipar seu aprendizado e agregar valor ao seu conhecimento visando ao mercado de trabalho.
- Usuários de Excel que desejam conhecer e aprender a utilizar os recursos de *Estatística* disponíveis.

Queremos agradecer a todos os professores e alunos que utilizaram as edições anteriores deste livro e que, com seu apoio, nos estimularam para apresentar esta quarta edição de *Estatística Usando Excel*. Agradecemos também a todos aqueles que participam de nosso constante desenvolvimento.

JUAN CARLOS LAPPONI
Agosto 2004

Capítulo 1

DADOS, VARIÁVEIS E AMOSTRAS



Um exemplo de *Estatística* é o Censo 2000 realizado pelo IBGE cujo primeiro resultado mostra que a população do Brasil no ano 2000 era de 169.799.170 pessoas. Depois, a população nos anos 1980, 1990, 1996 e 2000 classificadas por sexo, por grandes grupos de idade e por situação de domicílio em % está registrada na tabela¹ da Figura 1.1.

POPULAÇÃO TOTAL E PROPORÇÃO DA POPULAÇÃO POR SEXO, GRANDES GRUPOS DE IDADE E SITUAÇÃO DE DOMICÍLIO				
	1980	1990	1996	2000
População total	119.002.706	146.825.475	157.070.163	169.799.170
Por sexo (%)				
Homens	49,68	49,36	49,3	49,22
Mulheres	50,31	50,63	50,69	50,78
Por grandes grupos de idade (%)				
0-14 anos	38,2	34,72	31,54	29,6
15-64 anos	57,68	60,45	62,85	64,55
65 e mais	4,01	4,83	5,35	5,85
Por situação do domicílio (%)				
Urbana	67,59	75,59	78,36	81,25
Rural	32,41	24,41	21,64	18,75

FIGURA 1.1 Resultados do Censo 2000 realizado pelo IBGE.

Dos resultados registrados na tabela da Figura 1.1 pode-se deduzir como essas proporções evoluíram com o passar do tempo, as tendências de crescimento, mas não permitem medir a força dessas tendências. Uma forma de analisar essas tendências é medir a variação desses crescimentos durante os

¹ Informações obtidas em *Brasil em Síntese* no site www.ibge.gov.br do IBGE – Instituto Brasileiro de Geografia e Estatística.

anos definidos nas colunas da tabela. Na planilha *Censo 2000* incluída na pasta **Capítulo 1** foi calculada a taxa de crescimento de cada item utilizando o procedimento de média geométrica como mostra a tabela da Taxa de Crescimento Figura 1.2. Por exemplo, a média geométrica anual da população entre

os anos 1980 e 1990 é 2,12% resultado obtido com a seguinte fórmula $\left(\frac{146.825.475}{119.002.706}\right)^{\frac{1}{10}} - 1 = 0,0212$.

Esse procedimento de cálculo foi utilizado para obter os resultados restantes da tabela da Figura 1.2.²

Taxa de crescimento – Média geométrica anual					
	1990/1980	1996/1990	2000/1996	2000/1980	2000/1990
População total	2,12%	1,13%	1,97%	1,79%	1,46%
Por sexo					
Homens	-0,065%	-0,020%	-0,041%	-0,047%	-0,028%
Mulheres	0,063%	0,020%	0,044%	0,047%	0,030%
Por grandes grupos de idade					
0-14 anos	-0,95%	-1,59%	-1,57%	-1,27%	-1,58%
15-64 anos	0,47%	0,65%	0,67%	0,56%	0,66%
65 e mais	1,88%	1,72%	2,26%	1,91%	1,93%
Por situação do domicílio					
Urbana	1,12%	0,60%	0,91%	0,92%	0,72%
Rural	-2,79%	-1,99%	-3,52%	-2,70%	-2,60%

FIGURA 1.2 Taxa de crescimento utilizando a média geométrica anual.

Análise dos resultados

Os resultados da tabela da Figura 1.2 mostram que:

- A população total continua crescendo, entretanto a média geométrica da taxa de crescimento anual diminui, pois durante os anos 1980 e 1990 a média geométrica foi de 2,12% ao ano e durante os anos 1990 e 2000 foi de 1,5% ao ano.
- Quanto à classificação por sexo, a população de mulheres continua sendo maior que a dos homens com tendência de aumentar essa diferença. De 1980 a 2000 a população de homens tem diminuído com taxa média geométrica de -0,047% ao ano, e a população de mulheres tem aumentado, curiosamente, com taxa média geométrica +0,047% ao ano.
- Quanto à classificação por grandes grupos de idade entre 1980 e 2000, a população entre 0 e 14 anos diminuiu com taxa média geométrica de -1,27% ao ano, a população entre 15 e 64 anos aumentou com taxa média geométrica de 0,56% ao ano, e a população com mais de 65 anos aumentou com taxa média geométrica 1,91% ao ano.
- Quanto à classificação por situação de domicílio 1980 e 2000, a população com domicílio urbano aumentou com taxa média geométrica de crescimento positiva de 0,9% ao ano e a população com domicílios rurais diminuiu com taxa média geométrica de crescimento negativa de -2,7% ao ano.

Você pode conhecer a estimativa da população do Brasil e do Mundo minuto a minuto. Enquanto redigimos esta seção, da página do IBGE na Internet copiamos a informação registrada a seguir:

² O procedimento de projeção pela taxa média geométrica é apresentado no Capítulo 16.

Estimativas da População

no dia 16/7/2004 às 14 horas e 46 minutos

Somos agora no Brasil: 179.203.116 habs.

Somos agora no Mundo: 6.160.714.635 habs.

Projeções

A análise desses resultados não se esgota nas poucas medidas que realizamos na planilha Censo 2000, pois a partir desses resultados surgem perguntas relacionadas, primeiro, com as causas que vêm provocando esses resultados e, depois, com as projeções futuras que se podem extrair desses resultados. Por exemplo, enumerando as causas que vêm provocando a diminuição da população jovem e aumentando a população adulta com destaque às pessoas com mais de 65 anos e, olhando para o futuro, também poderiam ser enumeradas as possíveis consequências dessas tendências. Um resultado rápido das consequências futuras pode-se resumir da seguinte forma: em longo prazo a população será mais velha e crescerá menos como mostra a projeção da Figura 1.3.³



FIGURA 1.3 Projeção de Indicadores Sociais realizadas pelo IBGE.

Decisões

Os resultados estatísticos ajudam a tomar decisões com base em poucos dados.⁴ O processo estatístico de amostragem ou censo gera informações que auxiliam na realização de previsões ou projeções e é, ou deve ser, uma das preocupações das atividades de negócios e governamentais. Nas empresas é necessário prever as vendas, os estoques, os custos, o fluxo de caixa etc. para um determinado período como é o orçamento anual do próximo ano. Na administração pública faz-se necessário prever o número de habitantes, a arrecadação, os custos dos serviços prestados etc. Voltando ao Censo 2000, o seguinte trecho é um exemplo do que dizemos “...O estadista tem o dever de governar com olho no futuro, antecipando-se em dar respostas a problemas que explodirão depois de seu mandato....”⁵

³ Do artigo de Nilson Brandão Jr. e Alexandre Rodrigues: *População: mais velha e crescendo menos*, publicado no jornal O Estado de São Paulo em 14/04/2004.

⁴ O primeiro relato de um esforço ambicioso e influente de utilização do processo estatístico de amostragem foi realizado em 1664 em Londres por John Graunt que passara toda sua vida adulta como mercador de aviamentos. Veja Bernstein P. *Desafio aos Deuses – A Fascinante História do Risco* – Editora Campus, 1997.

⁵ Da coluna de Suely Caldas *A Previdência pede socorro! – Com a população idosa crescendo a galope, multiplica-se o déficit previdenciário*. Publicado no jornal O Estado de São Paulo em 18/04/2004.

Nas empresas que desejarem continuar crescendo no mercado em que atuam os desafios não são muito diferentes. As tendências dos índices mostram riscos, oportunidades e desafios. Enquanto o *cliente* dos serviços da administração pública é formado praticamente por todos os habitantes do país, o *cliente* das empresas privadas é uma parte desses habitantes. Por exemplo, o gerente de marketing necessita determinar o tamanho do mercado de seu novo produto, mas a população desse produto nem sempre coincide com a população do país, como descreve o seguinte trecho de um editorial: “*Que a afirmação, repetida à exaustão, de que o Brasil é um mercado constituído por 170 milhões de consumidores é uma falácia não é novidade. ... 40 milhões de pessoas, ou 23,5% da população do País, com rendas média e alta, que participam plenamente do mercado consumidor. ... Do consumo depende o crescimento sustentado da economia. As pessoas com rendas média e alta, segundo a pesquisa, já atingiram o limite de sua capacidade de consumo. A expansão das atividades dependeria, portanto, dos 130 milhões de pessoas que compõem as faixas mais baixas de rendimento ...*”⁶

A disciplina *Estatística*

O Censo 2000 nos deu a oportunidade de apresentar a utilização da *Estatística* sem entrar nos detalhes dos procedimentos de amostragem, resumo e análise dos dados e inferência, destacando algumas possíveis consequências futuras dessas projeções referentes a uma pequena parte das variáveis pesquisadas. Note que a análise realizada partiu do resumo das informações coletadas em questionários compostos de várias páginas utilizadas no censo.

EXEMPLO 1.1

No seu primeiro dia de trabalho, o novo gerente geral Ricardo pediu ao chefe de vendas Carlos o relatório de vendas do mês anterior. No mesmo dia, Carlos entregou o relatório solicitado contendo 65 páginas com 32 registros de vendas diárias em cada página. Carlos explicou que era um relatório completo onde cada registro de venda diária continha a data, o nome do comprador, o valor bruto, o desconto, o valor líquido, o prazo para pagamento e outras informações relevantes sobre a venda como o nome do vendedor etc.

Da forma como estão apresentadas as informações não será possível obter conclusões sobre as vendas do mês anterior senão for realizada alguma classificação desses dados. Para uma análise inicial, Ricardo definiu o valor das vendas diárias e suas datas como os dados relevantes, as variáveis da análise. Depois de resumir os dados dessas duas variáveis, Ricardo constatou que 38% das vendas diárias representam 70% das vendas do mês, e que 73% das vendas foram realizadas nos primeiros quinze dias do mês.

O Exemplo 1.1 mostra a necessidade de resumir as informações, pois da forma como os dados disponíveis estão apresentados não será possível obter conclusões. Algumas vezes os dados disponíveis são resumidos como os do Censo 2000 apresentado anteriormente, ou as informações disponíveis não são as requeridas, ou a quantidade de informações disponíveis é também um desafio para obter resultados. Resumir as informações do Exemplo 1.1 é necessário devido à variabilidade ou falta de uniformidade dos dados. Se, por exemplo, os valores das vendas da empresa forem constantes as respostas seriam obtidas de forma simples sem necessidade de realizar análises estatísticas e, conseqüentemente, a necessidade de estudar estatística seria bem menor.

Para obter as respostas requeridas foi necessário, primeiro, estabelecer quais indivíduos, pessoas, objetos ou coisas deveriam ser analisados e, depois, definir as características que deveriam ser medidas para obter as respostas procuradas.

⁶ Do Editorial *A falácia do mercado de consumo* publicado no jornal *O Estado de São Paulo* em 8/12/2003.

O objetivo da *Estatística Descritiva* é organizar, resumir, analisar e interpretar observações disponíveis.

Para alguns estudiosos a *estatística* é uma arte; para outros a *estatística* é a simples aplicação do *bom senso*. Em qualquer caso, a estatística ajuda a tomar decisões com informações incompletas, tendo presente que o sucesso da decisão dependerá da habilidade do analista para compreender os resultados das informações contidas nos dados. A primeira parte do processo decisório é a *estatística descritiva* e a outra é a *inferência estatística*.

O objetivo da *Inferência Estatística* é obter respostas corretas de questões específicas, atendendo a um determinado grau de acerto.

Origem dos dados

A Estatística lida com dados, números dentro de um contexto. Entretanto, a utilização de estatística é mais do que trabalhar com números, pois embora a organização dos números e a construção de gráficos possa ser mecanizada com softwares e modelos, as ideias e bons julgamentos, por enquanto, não podem ser automatizados. O analista deve ter o hábito de perguntar, por exemplo, o que mostram os resultados dentro de um determinado contexto? Quais as respostas que os dados podem dar a perguntas específicas?

Tenha em mente que durante a apresentação da disciplina Estatística é realizada uma análise explanatória de dados conhecidos, não havendo, em geral, nenhuma pergunta *in mente*, salvo situações como a do Censo 2000 apresentado na qual você consegue vivenciar os resultados apresentados. Entretanto, na prática diária da estatística são procuradas respostas a perguntas específicas, por exemplo, quais indivíduos (pessoas, animais, taxas de juros e outras coisas) devem ser estudados? Que variáveis devem ser medidas? Nesses casos, em geral, os dados devem ser gerados.

Os dados requeridos pela análise são obtidos pesquisando dados disponíveis, ou gerando novos dados. Em geral, os dados disponíveis são gerados e divulgados por instituições e empresas para muitas finalidades, as pesquisas do IBGE, de anuários, Internet, jornais, revistas etc. A procura dessas informações toma bastante tempo, porém com pouco desembolso de dinheiro. Entretanto, na geração de novos dados as respostas desejadas serão obtidas de amostras cujos indivíduos responderão a perguntas bem elaboradas e registradas num questionário. A procura dessas novas informações toma bastante tempo exigindo maior desembolso de dinheiro.

Depois de coletados, os dados poderão ter a necessidade de serem ajustados, pois nem sempre os dados coletados estarão no formato correto. Por exemplo, as vendas em \$ podem refletir variações combinadas de quantidade e de preço, devendo ser necessário retirar um desses efeitos, ou os dois, ajustando as quantidades considerando o crescimento da população e ajustando os preços para moeda constante deflacionando os dados com um índice adequado. Outro caso são as rentabilidades de investimentos que refletem mudanças econômicas como a inflação e os prazos diferentes, devendo ser necessário retirar esses efeitos.

Os dados ajustados são representativos do processo sob análise; entretanto, as unidades dos dados podem ser difíceis de analisar, por exemplo, o histograma do faturamento em \$ de uma empresa, ou a análise simultânea de várias séries de dados com unidades diferentes. Uma forma de facilitar a análise é transformar as séries de dados nas suas respectivas séries de taxas de crescimento, de forma unitária ou percentual, seja um grupo de séries de dados com unidades diferentes ou uma única série. Nesse procedimento, em geral, perde-se o primeiro dado.

Dados e variáveis

Quanto a sua origem, os dados ou observações podem ser obtidos de:

- **Respostas de Pesquisas.** Quem aplica a pesquisa não tem nenhum controle intencional sobre os fatores que influenciam as respostas: a contagem de habitantes de um país, o cadastro dos clientes de um banco, a aceitação de um produto por um determinado tipo de consumidor etc.
- **Respostas de Experimentos.** Quem aplica o experimento tem controle intencional sobre os fatores que influenciam as respostas: o teste de estabilidade de produtos perecíveis frente a diferentes valores de temperatura e umidade, o desgaste de componentes de equipamentos mecânicos em condições especificadas e fora de elas etc.

Unidade elementar é qualquer pessoa, objeto ou coisa que faça parte de uma população.

Dado é o resultado de investigação, cálculo ou pesquisa, do dicionário Houaiss.

Variável é toda característica que pode assumir diversos valores conforme pessoa, objeto ou coisa.

As respostas de uma pesquisa ou um experimento são a matéria-prima da análise estatística em que os dados ou observações são obtidos medindo as características de uma pessoa, objeto ou coisa. O conjunto dessas respostas ou observações forma uma unidade elementar que, em geral, está composta de uma ou mais características denominadas variáveis. Por exemplo, cada questionário do Censo 2000 é uma unidade e cada resposta dentro desse questionário é uma variável.

EXEMPLO 1.2

A tabela seguinte registra parte do Cadastro de Funcionários de uma empresa.

Nome	Idade	Cargo	Sexo	Peso	Escolaridade
João	27	Supervisor	M	62 kg	2º Grau
Alex	38	Chefe	M	78 kg	1º Grau
Estela	34	Gerente	F	65 kg	3º Grau
Ana	32	Secretária	F	58 kg	3º Grau

Quais são as unidades elementares e as variáveis deste cadastro? Cada uma das seis variáveis de cada funcionário da empresa, Nome, Idade, Cargo, Sexo, Peso e Escolaridade, compõem uma unidade elementar, tendo a tabela quatro unidades elementares.

Número de variáveis

A unidade elementar de informação pode conter qualquer número de variáveis e a análise estatística pode ser classificada de acordo com esse número de variáveis, por exemplo:

- **Uma única variável.** São exemplos deste tipo de informação:
 - As vendas mensais de uma loja.
 - As projeções realizadas por 20 analistas financeiros sobre o valor da taxa de juros nos próximos 12 meses.

- O lucro líquido trimestral de uma empresa.
- O saldo médio dos clientes de um banco comercial etc.

Os métodos estatísticos para resumir cada uma dessas variáveis são: o histograma, a média, o desvio padrão etc.

- **Duas variáveis**⁷. São exemplos deste tipo de informação:
 - Os valores mensais do faturamento e do lucro líquido da empresa.
 - A rentabilidade diária de uma ação e a rentabilidade diária do índice da carteira teórica da Bolsa de Valores.
 - A rentabilidade anual de um investimento e a taxa anual de inflação.
 - O salário e a idade dos entrevistados numa pesquisa de clientes potenciais de um fabricante de refrigerantes etc.

Além dos métodos estatísticos para resumir cada uma dessas variáveis há também o objetivo de utilizar métodos estatísticos para verificar e medir a força da relação entre duas variáveis, a projeção de uma variável em função da outra etc.⁸

- **Três ou mais variáveis**⁹. São exemplos deste tipo de informação:
 - A relação entre o PIB e duas ou mais variáveis econômicas.
 - Cadastro dos clientes de um banco: idade, escolaridade, profissão, número de bancos que opera, residência etc.
 - Cadastro dos funcionários de uma empresa: nome, sexo, escolaridade, tempo de casa, cargo etc.
 - Resultados da colheita de um determinado tipo de cultura: área cultivada, região, umidade e tipo do solo, clima durante o cultivo, quantidade e qualidade do fertilizante usado, tipo de preparação da terra, cuidados e qualidade da mão de obra etc.

Neste caso, além dos métodos estatísticos para resumir cada uma dessas variáveis há também o objetivo de utilizar métodos para verificar a existência de relação entre uma e as restantes variáveis, o grau de relação entre as variáveis, a projeção de uma variável em função das restantes etc.

Classificação dos dados

Como o procedimento estatístico a ser aplicado dependerá da natureza dos dados¹⁰ ou das observações de cada variável, deve-se desenvolver a habilidade de distinguir os tipos de dados possíveis e suas unidades de medida. Quanto a sua natureza, as observações ou dados se classificam em quantitativas discretas e contínuas, qualitativas nominais e ordinais, de corte transversal e séries temporais.

- **Dados quantitativos**. Refere-se a quantidades medidas numa escala numérica, em geral, acompanhadas de alguma unidade de medida e podem ser de dois tipos:
 - **Dados discretos**. Referem-se aos valores numéricos que assumem somente números inteiros positivos 0, 1, 2, 3 Os dados discretos resultam, em geral, de contagens: a quantidade de vendas diárias de uma empresa, o número de filhos das famílias de uma região do país, o número de movimentos da conta corrente dos clientes de um banco comercial, a quantidade de peças defeituosas em um lote de produção, o número de transações financeiras com erro de lançamentos, o número de acidentes nas estradas durante as férias anuais de verão etc.

⁷ Denominado como *análise bidimensional*.

⁸ As variáveis são classificadas em *dependentes* e *independentes* conforme a situação, podendo uma mesma variável, em épocas diferentes, assumir um dos dois tipos.

⁹ Denominado como *análise multidimensional*.

¹⁰ Do dicionário Houaiss, *dado* é o resultado de investigação, cálculo ou pesquisa.

- **Dados contínuos.** Referem-se aos valores numéricos que assumem qualquer valor do conjunto dos números reais. Os dados contínuos resultam, em geral, de medições que podem ter grande precisão: o valor das vendas diárias de uma empresa, a estatura dos alunos da terceira série, o valor dos depósitos e retiradas da conta corrente dos clientes de um banco comercial, o consumo mensal de energia elétrica, o tempo necessário para realizar uma tarefa repetitiva, o tempo de espera para ser atendido em um serviço de saúde pública etc.
- **Dados qualitativos.** Refere-se às observações não numéricas e são classificados em nominais e ordinais:
 - **Dados nominais.** Esses dados não têm ordenamento nem hierarquia. Por exemplo, o sexo dos funcionários registrados no cadastro da empresa, o estado civil, o nome das empresas que têm ações negociadas na Bolsa de Valores, cidade de residência do respondente etc.
 - **Dados ordinais.** Esses dados são equivalentes aos nominais, porém incluindo uma ordem, uma hierarquia. Por exemplo, o cargo dos funcionários registrados no cadastro da empresa: presidente, diretor, gerente etc.; a resposta a um questionário de pesquisa onde há uma escala para escolher: bom, regular e ruim; as posições das cinquenta maiores empresas por vendas durante um ano: primeira, segunda etc.

Escala de medição dos dados

Da forma como foi apresentada a classificação dos dados das variáveis não é suficiente. As seguintes quatro escalas de classificação adicionam novas informações às anteriores.

- **Escala Nominal.** Valores numéricos numa escala nominal apenas dão nome a uma categoria ou classe; os números são utilizados somente para diferenciar os objetos, categorias ou nomes. Por exemplo, numa pesquisa de mercado realizada nas regiões Sul e Sudeste do Brasil, o variável *estado de nascimento* do entrevistado foi codificada da seguinte forma: 1=Rio Grande do Sul, 2=Santa Catarina, 3=Paraná, 4=São Paulo e 5=Rio de Janeiro. Embora o código tenha transformado um nome em um número, este número não mantém todas as propriedades dos números; por exemplo, não se podem estabelecer relações como $3 > 2$ ou $1 + 2 = 3$ ou $3 - 2 = 1$ como o leitor pode confirmar substituindo cada número pelo *estado* correspondente.
- **Escala Ordinal.** Valores numa escala ordinal dão nome e ordem a um objeto, categoria ou classe; os números se utilizam para diferenciar em ordem de superioridade seguindo algum critério de hierarquia. Em uma pesquisa a variável *instrução do entrevistado* foi codificada assim: 1=Sem Instrução, 2=Primeiro Grau, 3=Segundo Grau, 4=Terceiro Grau, 5=Mestre e 6=Doutor. Neste caso, na transformação de um nome em um número, o número mantém algumas propriedades dos números; por exemplo, podem ser estabelecidas relações do tipo $3 > 2$ (o grau de instrução 3 é maior que o grau de instrução 2), porém não se podem estabelecer relações do tipo $2 + 3 = 5$ como o leitor pode confirmar substituindo cada número pelo grau de instrução correspondente. Ao estudar as medidas de ordenamento *percentil* e *quartil* se poderá ver que são medidas na escala ordinal, pois elas mostram o desempenho de cada elemento de uma variável com relação aos outros elementos sem preocupação de determinar quanto melhor ou pior foi o desempenho.
- **Escala de Intervalos.** Valores numa escala de intervalos eliminam a limitação da escala ordinal, estabelecendo intervalos iguais onde é possível ordenar as medições e, ao mesmo tempo, explicar em quanto difere uma observação de outra. Por exemplo, o aumento de temperatura de ontem para hoje é de cinco graus, de 20 para 25 graus centígrados. Podemos dizer que hoje está mais quente do que ontem. Essa escala de medida tem uma unidade de medida, um zero arbitrário¹¹ e a distância entre duas medições nessa escala tem um significado preciso. Outro exemplo de escala de intervalos são os tempos dos calendários gregorianos e outros tipos.

¹¹ O zero da escala de graus centígrados é o ponto de congelamento da água no nível do mar; entretanto, essa temperatura medida na escala de graus Fahrenheit é 32 graus.

- **Escala Proporcional.** Valores numa escala proporcional eliminam a limitação da escala intervalar estabelecendo um zero da própria categoria, denominado como zero absoluto. Por exemplo, peso zero claramente significa falta de peso, o peso de uma caixa de 86 kg é o dobro do de uma caixa de 43 kg, e 33 peças rejeitadas de um lote de produção representam o triplo do lote de produção com onze peças rejeitadas.

Tipos de variáveis

As variáveis podem ser obtidas de duas formas.

- **Séries temporais.** As observações são dados de uma mesma variável em diferentes períodos de tempo: o valor do PIB anual de um país, a taxa mensal de desemprego numa região, as cotações diárias de uma ação, a rentabilidade mensal de uma empresa, a demanda de energia elétrica diária na região Sudeste medida às 18h etc.
- **Corte transversal numa data ou período.** Se na coleta dos dados não for considerada a sequência temporal; por exemplo, amostras da quantidade produzida e do preço médio dos produtos, ou das vendas e do investimento em propaganda, a média de apartamentos vendidos durante o último mês pelas primeiras dez imobiliárias da cidade, o número de operações fechadas por cinco ações numa determinada data etc.

População e amostra

A contagem da população em 2000 realizada pelo IBGE foi de 169.799.170. Em termos estatísticos, a contagem do censo foi realizada consultando a *população* do Brasil. Outro exemplo, a partir de uma amostra das contas de energia elétrica dos consumidores residenciais do Brasil, o consumo de energia elétrica pode ser relacionado com as condições sócioeconômicas dos consumidores.

População é o conjunto total unidades elementares de pessoas, objetos ou coisas sobre as quais se querem obter informações.

Um subconjunto de unidades elementares selecionadas de uma população é denominado *amostra*.

Uma população pode ser formada por todos os habitantes de um país, ou de um estado, ou de um município etc. Um exemplo de pesquisa de uma população completa é o censo demográfico do Brasil realizado pelo IBGE. A análise das vendas de um segmento da economia, por exemplo, o de montadoras de carros, durante o mesmo ano é outro exemplo de população. Entretanto, nem sempre é conveniente obter informações de todas as pessoas, objetos ou coisas de uma população. Os resultados de uma pesquisa de intenção de voto de todos os eleitores do país numa eleição presidencial não conseguiriam captar do que os partidos políticos necessitam, pois o tempo necessário para coletar todas as opiniões comprometeria os resultados, além de ser muito cara para a finalidade que se propõe. Em alguns casos, a restrição de consultar toda a população é econômica, como é o caso da determinação da vida útil das lâmpadas que obrigaria a testar todas as lâmpadas produzidas, não restando nenhuma para venda! Dessa maneira, o procedimento recomendado é escolher uma *amostra representativa* de um lote de lâmpadas produzidas.

Deve-se ter presente que nem sempre um censo oferecerá melhores resultados do que uma amostra. Em muitos casos a obtenção de informações de uma amostra da população é mais adequada, pois ela é

mais rápida de ser aplicada, concluída, de obter e utilizar os resultados e, conseqüentemente, tem custo menor. Os erros possíveis de serem cometidos na realização de uma amostragem podem ser evitados ou corrigidos aplicando técnicas adequadas e estabelecendo resultados com estimativa de erro, por exemplo, um intervalo de confiança.

Uma *amostra representativa* tem as mesmas características da população de onde foi retirada.

Muitas aplicações de estatística utilizam amostras retiradas de uma população da qual se deseja obter respostas, tendo presente que a amostra é um subconjunto representativo da população.

EXEMPLO 1.3

O objetivo é estimar o número de palavras contidas neste livro de estatística, considerando apenas as páginas dos capítulos, sem considerar o Sumário, o Índice etc.

Solução. Há diversas formas de estimar o número de palavras contidas no livro. A seguir apresentamos quatro procedimentos diferentes.

1. O primeiro procedimento começa pela escolha de uma página do livro e a contagem do número de linhas dessa página. Depois, selecionamos três linhas da página escolhida e contamos o número de palavras contidas nas três linhas. Em sequência, calculamos a média de palavras por linha e, com esse valor, calculamos o número de palavras por página, utilizando o número de linhas já definido. Finalmente, multiplicando o número estimado de palavras por página pelo número total de páginas obtemos uma estimativa do número de palavras do livro. Antes de o leitor fazer reparos quanto a este procedimento, sugerimos que continue com os outros dois procedimentos.
2. O segundo procedimento começa pela escolha de uma página, segue com a contagem do número de palavras contidas na página escolhida e termina com a multiplicação deste valor pelo número de páginas do livro.
3. O terceiro procedimento, um pouco melhor que o anterior, começa pela escolha de cinco páginas diferentes do livro. Segue com o cálculo da média de palavras por página. Finalmente, o número de palavras do livro é estimado como o resultado da multiplicação da média de palavras por página pelo número de páginas do livro.
4. O último procedimento é a contagem de todas as palavras do livro de estatística, página por página. É o caso de realizar o censo de palavras do livro.

As estimativas do número de palavras do livro dos três primeiros procedimentos do Exemplo 1.3 deverão ser diferentes, sendo que a estimativa da amostra de maior tamanho será mais próxima do resultado da contagem de todas as palavras no quarto procedimento. Estendendo essas conclusões, pode-se atestar confiança na estimativa de uma amostra se os elementos da amostra forem escolhidos assegurando que todos os participantes que formam a população tenham a mesma oportunidade ou chance de serem escolhidos. A amostra de uma população retirada dessa forma é denominada *amostra aleatória* de tamanho n cujas premissas são:

1. Cada unidade elementar da população tem a mesma probabilidade de ser escolhida numa amostra de tamanho n , sendo que cada unidade elementar será escolhida de forma independente das outras unidades.
2. Todas as amostras extraídas possíveis de tamanho n de uma população têm a mesma probabilidade de serem selecionadas.

Uma *amostra aleatória* de tamanho n retirada de uma população é uma das muitas possíveis e igualmente prováveis combinações de n unidades elementares que podem ser retiradas de uma população. Portanto, qualquer amostra de tamanho n tem a mesma probabilidade de ser selecionada.

Dígitos e números aleatórios

As expressões números aleatórios e dígitos aleatórios são utilizadas como sinônimos, entretanto há uma diferença entre essas duas expressões que é importante conhecer.

- Os números aleatórios são independentes e distribuídos uniformemente no intervalo de números reais entre 0 e 1, ou de forma mais técnica no intervalo (0, 1).
- Os dígitos aleatórios são os números do conjunto {0, 1, 2, 3, ..., 9} distribuídos uniformemente.

O agrupamento de vários dígitos aleatórios forma um número aleatório. Por exemplo, na *Tabela de Números Aleatórios* apresentada no capítulo *Tabelas* no final do livro, os números aleatórios são formados com dígitos aleatórios.

Os números aleatórios são o ingrediente básico e necessário no procedimento de simulação da maioria dos sistemas discretos. Em geral, as linguagens de programação têm uma sub-rotina ou função para gerar um número aleatório. O Excel dispõe de duas funções e uma ferramenta de análise para gerar números aleatórios. Os números aleatórios gerados por esses meios são também conhecidos como pseudonúmeros aleatórios, pois eles são gerados através de um procedimento que pode ser reproduzido o que pode introduzir um desvio da premissa dos números aleatórios serem independentes e uniformemente distribuídos. Há testes que medem os desvios dos números aleatórios gerados por esses procedimentos.

Funções do Excel

O Excel dispõe das funções matemáticas ALEATÓRIO e ALEATÓRIOENTRE para gerar números aleatórios e da ferramenta de análise *Amostragem* para extrair amostras com reposição de uma população e da ferramenta de análise *Geração de Número Aleatório* que será apresentada em um capítulo posterior. Tentando evitar aborrecimentos ao leitor provenientes de uma instalação incompleta do Excel, sugerimos que veja o Apêndice 1 deste capítulo *Preparando o Excel para Começar*. Da mesma maneira, sugerimos que veja o Apêndice 2 *Como Registrar uma Função na Planilha Excel*.

Aleatório()

A função matemática ALEATÓRIO¹² retorna um grupo de números aleatórios entre 0,00...0 e 1,00...0 com a quantidade de casas decimais depois da vírgula definida pelo leitor, por exemplo, 0,236; 0,86945 etc. Se o nome da função for inserido sem o acento ortográfico, o Excel aceita e registrará a função com letras maiúsculas e com o acento ortográfico. Na célula C4 da planilha **Funções** incluída na pasta **Capítulo_1** foi registrada a fórmula =ALEATÓRIO(). É importante ter presente que toda vez que a planilha for recalculada a função ALEATÓRIO gerará um novo grupo de números entre 0,00...0 e 1,00...0.

¹² Em inglês, a função ALEATÓRIO é RAND.

Em alguns casos será necessário gerar números aleatórios inteiros entre dois limites, um inferior e o outro superior, por exemplo, entre 000 e 999. Para esses casos podem ser utilizadas fórmulas matemáticas como mostraremos numa seção posterior, ou utilizar a função matemática ALEATÓRIOENTRE do Excel.

Aleatórioentre(*inferior*; *superior*)

A função estatística ALEATÓRIOENTRE¹³ retorna um número aleatório inteiro entre os valores dos argumentos¹⁴ *inferior* e *superior* definidos na função. O argumento *inferior* e o argumento *superior* são, respectivamente, o menor inteiro e o maior inteiro que a função ALEATÓRIOENTRE retornará. Diferente da função ALEATÓRIO, se o nome dessa função for inserido sem o acento ortográfico o Excel não aceitará a função retornando o valor de erro #NOME? na célula.

Inserindo a fórmula =ALEATÓRIOENTRE(0;599)¹⁵ numa célula vazia de uma planilha, o Excel retornará um número inteiro entre 000 e 599, valores dos argumentos inferior e superior, respectivamente. Toda vez que a planilha for recalculada a função ALEATÓRIOENTRE gerará um novo número aleatório dentro do mesmo intervalo. Na célula C5 da planilha **Funções** incluída na pasta **Capítulo_1** foi registrada a fórmula =ALEATÓRIOENTRE(0;599) como mostra a Figura 1.4. No Apêndice 1 *Como Registrar um Função no Excel* o leitor encontrará os procedimentos de registro de funções numa planilha de Excel.

FIGURA 1.4 Funções ALEATÓRIO e ALEATÓRIOENTRE.

	A	B	C	D	E
1	Funções ALEATÓRIO e ALEATÓRIOENTRE				
2					
3					
4					
5					
6					

Funções Matemáticas		
ALEATÓRIO	0,0864	=ALEATÓRIO()
ALEATÓRIOENTRE	580	=ALEATÓRIOENTRE(0;599)

Antecipando um pouco o conhecimento de distribuições de frequências, os grupos de números gerados pelas duas funções apresentadas têm distribuição uniforme, sendo que com a função ALEATÓRIO será gerada uma distribuição uniforme contínua e com a função ALEATÓRIOENTRE, uma distribuição uniforme discreta.

Fórmulas com a função Aleatório do Excel

Como foi antecipado, é possível gerar números aleatórios entre dois limites utilizando fórmulas. As três fórmulas seguintes geram números aleatórios entre os limites *inferior* e *superior* utilizando a função geradora de números aleatórios ALEATÓRIO. Os exemplos seguintes estão registrados na planilha **NA com fórmulas** incluída na pasta **Capítulo 1**.

- =ALEATÓRIO()*(*superior-inferior*)+*inferior*

Essa fórmula gera números aleatórios com decimais entre o limite *superior* e o limite *inferior* informados. Por exemplo, na célula E5 foi registrada a fórmula =ALEATÓRIO()*(C4-C3)+C3 que gera números aleatórios com decimais entre 0 e 599, valores informados nas células C3 e C4. O resultado da célula E5 foi formatado com duas casas decimais como se pode ver na Figura 1.5.

¹³ Em inglês, a função ALEATÓRIOENTRE é RANDBETWEEN.

¹⁴ Argumentos são os valores que uma função usa para realizar operações e cálculos. Os argumentos desta função são: *mínimo* e *máximo*.

¹⁵ Se esta função não estiver disponível e retornar o erro #NOME?, instale e carregue o suplemento *Ferramentas de análise*. Veja o Apêndice 1 deste capítulo.

- $\text{=INT}(\text{ALEATÓRIO}() * (\text{superior} - \text{inferior}) + \text{inferior})$

A fórmula geradora de números aleatórios é a fórmula anterior. Nesta nova fórmula foi incluída a função matemática INT.

- $\text{INT}(\text{número})$

A função INT retorna o valor registrado no argumento *número* arredondado para baixo até o número inteiro mais próximo.

Por exemplo, a fórmula $\text{=INT}(\text{ALEATÓRIO}() * (\text{C4} - \text{C3}) + \text{C3})$ registrada na célula E6 gera números aleatórios sem decimais entre 0 e 599, valores informados nas respectivas células C3 e C4. O resultado da célula E6 foi formatado sem casas decimais.

	A	B	C	D	E	F	G	H	I
1	Geração de números aleatórios com fórmulas								
2									
3		Inferior	0						
4		Superior	599						
5		Número aleatório entre 0 e 599			68,84	$\text{=ALEATÓRIO}() * (\text{C4} - \text{C3}) + \text{C3}$			
6		Número aleatório entre 0 e 599			386	$\text{=INT}(\text{ALEATÓRIO}() * (\text{C4} - \text{C3}) + \text{C3})$			
7		Núm_dígitos	2						
8		Número aleatório entre 0 e 599			166,6500	$\text{=TRUNCAR}(\text{ALEATÓRIO}() * (\text{C4} - \text{C3}) + \text{C3}; \text{C7})$			
9									

FIGURA 1.5 Geração de números aleatórios com fórmulas e a função ALEATÓRIO().

- $\text{=TRUNCAR}(\text{ALEATÓRIO}() * (\text{superior} - \text{inferior}) + \text{inferior}; \text{núm_dígitos})$

A fórmula geradora de números aleatórios é a fórmula anterior adicionada da função matemática TRUNCAR.

- $\text{TRUNCAR}(\text{núm}; \text{núm_dígitos})$

A função matemática TRUNCAR¹⁶ retorna o valor do argumento *núm* truncado com a quantidade de dígitos especificados no argumento *núm_dígitos*. Se *núm_dígitos* for igual a zero, o resultado da função TRUNCAR é equivalente ao da função INT.

Simulação da retirada de um número de uma urna

Os números gerados pelas duas funções apresentadas têm distribuição uniforme, sendo que a função ALEATÓRIO gerará uma distribuição uniforme contínua e a função ALEATÓRIOENTRE uma distribuição uniforme discreta. O que significa isso? Considere que uma urna tenha dez bolas pequenas numeradas de 0 a 9. A seguir suponha que você retira uma bola, verifica o número que identifica a bola, por exemplo, o número 3, registra esse número numa coluna de uma planilha Excel e por último retorna a bola para a urna. A seguir, mexe as bolas dentro da urna, retira uma nova bola e repete o procedimento anterior. Suponha que continua com esse procedimento até completar um número bastante grande de extrações. Como resultado, em longo prazo os dez números que formam o conjunto {0, 1, 2, 3, ..., 9} terão sido retirados o mesmo número de vezes, ou seja, cada um dos números deverá ter sido retirado 10% do total de retiradas ou amostras. Tecnicamente, todos os dez números terão a mesma frequência e, representando os dez número em um diagrama de barras verticais, observe que todas as barras têm a mesma altura, pois se trata de uma distribuição uniforme discreta.

Esse procedimento poderá ser simulado no Excel gerando números aleatórios do conjunto {0, 1, 2, 3, ..., 9} com a fórmula $\text{=ALEATÓRIOENTRE}(0,9)$, repetindo essa fórmula o número de vezes necessárias.

¹⁶ Em inglês, a função TRUNCAR é TRUNC. Como informação adicional, a função ARRED(*núm*; *núm_dígitos*), em inglês ROUND, dá um resultado equivalente ao da função TRUNCAR, porém, arredondando no lugar de truncar o resultado. O Excel dispõe de outras funções: ARREDONDAR.PARA.BAIXO, ARREDONDAR.PARA.CIMA e ARREDMULTB, em inglês, respectivamente, ROUNDDOWN, ROUNDUP e MROUND.

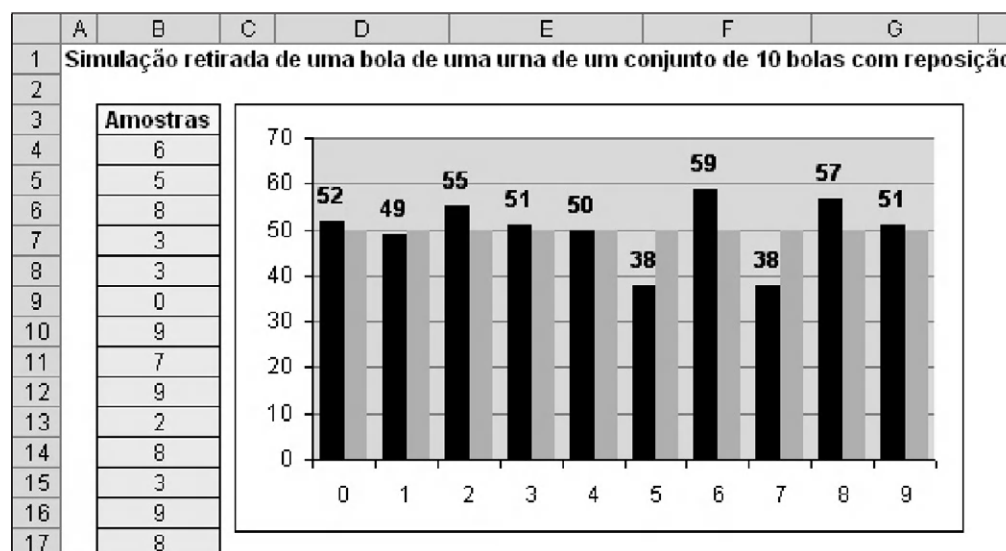
Tecnicamente declaramos que do conjunto de números $\{0, 1, 2, 3, \dots, 9\}$ retiramos um determinado número adequado de amostras aleatórias de tamanho $n=1$ com reposição. O longo prazo não é um valor determinado ou finito e, na prática, esse valor pode ser 500 como utilizamos na simulação seguinte, ou 1.000 ou maior que esse valor. Deve-se entender que quanto menor for o número de amostras da simulação, maior será o desvio dos valores das frequências observadas em comparação com os valores das frequências esperadas. A Figura 1.6 mostra o gráfico de barras verticais do resultado de uma simulação de 500 retiradas com reposição de uma bola de uma urna contendo dez bolas numeradas de zero a nove onde se pode ver que, nesse caso, duas bolas alcançaram o valor 50; 10% do número de retiradas.

Na planilha **Simulação** incluída na pasta **Capítulo 1** foi construído o modelo que gera 500 números aleatórios ou amostras do conjunto $\{0, 1, 2, 3, \dots, 9\}$, conta os resultados e constrói o gráfico de barras verticais denominado histograma. Vejamos o procedimento de construção do modelo:

- Na célula B4 foi registrada a fórmula `=ALEATÓRIOENTRE(0;9)` que gera um número aleatório entre 0 e 9.
- Depois, essa fórmula foi copiada até a célula B503. Os resultados de cada uma das 500 células do intervalo B4:B503 é uma amostra aleatória com reposição de tamanho $n=1$ retirada da população $\{0, 1, 2, 3, \dots, 9\}$.
- No intervalo D4:E14, oculto detrás do gráfico, foi construída a tabela de frequências absolutas, tema que será apresentado no Capítulo 2.

Pressionando a tecla de função F9 a planilha será recalculada, novas amostras serão geradas, uma nova tabela de distribuição de frequências absolutas será registrada e o histograma será atualizado.

FIGURA 1.6
Simulação de 500
retiradas
de uma bola com
reposição.



Analisando as frequências absolutas observadas na Figura 1.6, barras pintadas de cor mais escura, verificamos que seus valores se situam ao redor de 50. Entretanto, a frequência esperada de cada um dos dez números é 50, barras pintadas de cor mais clara no histograma. A diferença entre as frequências observadas e as frequências esperadas correspondentes pode ser atribuída à variabilidade amostral, a falhas do gerador de números aleatórios ou ao reduzido tamanho da amostra.¹⁷

¹⁷ Pela lei dos grandes números, 500 amostras representam um número pequeno, tema tratado no Capítulo 5.

Amostragem

Seguindo alguns critérios de seleção, o subconjunto escolhido de uma população é denominado amostra. Há dois tipos de amostras quanto à forma de serem extraídas da população, a amostra probabilística e as restantes que não são probabilísticas. Na amostra probabilística todos os componentes da população têm alguma chance de serem selecionados, escolhidos. Como nas amostras não probabilísticas alguns componentes da população não têm nenhuma chance de serem selecionados, deverá ser definido algum critério de escolha.

Um exemplo de amostra probabilística, também denominada amostragem aleatória, é a amostragem realizada na seção anterior quando simulamos a retirada de uma bola de uma urna contendo dez bolas, repondo a bola extraída depois de registrar seu resultado. Porém, esse tipo de amostragem tem mais uma característica, pois todos os elementos da população têm a mesma chance de serem selecionados. Esse procedimento de amostragem é denominado amostragem probabilística simples. Entretanto, pode ocorrer que uma amostra desse tipo não seja representativa da população.¹⁸ Por exemplo, em uma população formada por 50% de mulheres e 50% de homens, a amostragem probabilística simples pode resultar numa amostra de 65% de mulheres e 35% de homens. Nesse caso a amostra continua sendo aleatória mas não é representativa.

Na descrição da simulação da retirada de um número de uma urna foi registrado o procedimento que repetimos: retira uma bola da urna, registra o número da bola numa coluna de uma planilha Excel e por último retorna a bola para a urna. Em vez de voltar a bola para a urna, o procedimento poderia ser o de manter as bolas retiradas fora do processo de seleção. A primeira amostra é denominada amostra probabilística simples com reposição, ou simplesmente amostra com reposição, e a outra amostra probabilística simples sem reposição ou amostra sem reposição. Resumindo:

- Na amostragem com reposição, a unidade selecionada retorna para a população. Portanto, em cada nova seleção a população mantém a mesma quantidade de unidades elementares.
- Na amostragem realizada sem reposição, a unidade selecionada não retorna para a população. Portanto, em cada seleção a população é reduzida de uma unidade elementar.

Observe que, em geral, as amostragens são realizadas sem reposição e os cálculos estatísticos nos dois tipos de amostragens são os mesmos. Por exemplo, numa pesquisa de intenção de voto para escolha do governador do estado se espera que cada pessoa seja entrevistada apenas uma vez. Se o tamanho da população for suficientemente maior que o tamanho da amostra, recomendado mais de vinte vezes, os resultados estatísticos das amostras com e sem reposição não serão muito diferentes, pois a chance de escolher o mesmo elemento é muito pequena. Contudo, deve-se tomar cuidado com populações pequenas quando comparadas com o tamanho da amostra a ser extraída.

Geração de amostras probabilísticas simples

Como deve ser gerada uma amostra probabilística simples? Os exemplos a seguir mostram procedimentos e ferramentas.

EXEMPLO 1.4

O objetivo do diretor da escola primária é avaliar o conhecimento de matemática adquirido pelos alunos da sexta série no final do primeiro semestre. No lugar de aplicar um teste aos 35 alunos, ele prefere aplicar o teste numa amostra de seis alunos. Qual o procedimento adequado de amostragem?

Solução. Os 35 alunos da sexta série estão registrados no caderno de presença diária identificados pelo nome em ordem alfabética. O procedimento de amostragem probabilística simples de seis alunos é parecido com o procedimento de um sorteio que a seguir descrevemos:

1. Começamos por preparar 35 pequenos papéis iguais, por exemplo, uma folha de papel quadrada com três a quatro centímetros de lado.
2. Em cada papel registramos o nome completo de cada aluno.
3. Depois, os 35 papéis com os nomes registrados são colocados numa urna adequada, porém sem dobrar os pequenos papéis.
4. Iniciamos o sorteio remexendo os papéis dentro da urna antes de retirar o primeiro papel sorteado.
5. Continuamos com esse procedimento até completar a retirada dos seis papéis planejados.
6. Divulgamos o nome dos seis alunos escolhidos para serem avaliados.

A amostragem do Exemplo 1.4 é sem reposição. Na instrução da amostragem foi estabelecido que os 35 papéis com os nomes dos alunos registrados sejam depositados numa urna adequada sem nenhuma dobra. Deixamos para você analisar as características da extração dos papéis não dobrados comparando com o caso dos mesmos papéis dobrados, por exemplo, em quatro partes. Analise se nos dois procedimentos, papéis com e sem dobra, atende-se à premissa de que cada unidade elementar da população tenha a mesma probabilidade de ser escolhida numa amostra de tamanho 6.

Quando a população for muito grande o procedimento do Exemplo 1.4 não é adequado, pois a seleção dos papéis será mais difícil de realizar e a amostra extraída se afastará das premissas de amostra aleatória. Nesse caso deve-se utilizar a Tabela de Números Aleatórios¹⁹ apresentada em Tabelas no final deste livro. O Exemplo 1.5 mostra a descrição do procedimento de amostragem sem repetição. O Exemplo 1.6 mostra o trabalho direto com a tabela.

EXEMPLO 1.5

O objetivo da auditoria interna da empresa é verificar se o Setor de Contas a Pagar cumpre com as rotinas estabelecidas pela empresa para pagamento de fornecedores. Deve-se estabelecer o procedimento de seleção de quinze processos dos últimos 600 realizados.

Solução. A amostragem que deve ser realizada é do tipo probabilístico simples e sem reposição, pois interessa analisar somente quinze processos diferentes. No processo de seleção dos componentes da amostra é utilizada a Tabela de Números Aleatórios apresentada no capítulo Tabelas no final deste livro. O procedimento de amostragem é:

1. Numerar os últimos 600 processos de pagamento de 000 até 599. Os processos escolhidos serão analisados de trás para a frente das aprovações, começando pela última aprovação que autorizou o pagamento.
2. Escolher um número aleatório qualquer na Tabela de Números Aleatórios e tomar nota dos três últimos algarismos.
3. Deslocar-se na tabela por linha ou por coluna ou pulando entre elas até escolher o próximo número aleatório e tomar nota dos três últimos algarismos.
4. Repetir o procedimento anterior até completar a seleção dos quinze números aleatórios contendo três algarismos diferentes. Os três algarismos que formarem números maiores a 599 ou serem repetidos não serão considerados durante a seleção.
5. Separar os quinze processos de pagamento identificados pelos quinze números aleatórios escolhidos.

Há casos em que é necessário extrair amostras de uma população identificada por dados qualitativos, observações não numéricas classificadas em nominais e ordinais, como mostra o exemplo seguinte.

¹⁹ Apenas como comentário, a lista telefônica de assinantes é uma boa geradora de dígitos aleatórios, considerando apenas os quatro últimos dígitos.

EXEMPLO 1.6

A professora de artes quer analisar o resultado de pintar uma figura geométrica qualquer dividida em oito partes utilizando quatro cores escolhidas aleatoriamente da população formada pelas seguintes quatro cores {*amarelo, vermelho, azul, verde*} e utilizando o Excel.

Solução. A figura seguinte mostra a solução registrada na planilha **Pintura** incluída na pasta **Capítulo 1**.

[illegible]

A amostragem que deve ser realizada é do tipo probabilístico simples com reposição, pois interessa analisar o arranjo de quatro cores em oito partes de uma figura geométrica. O resultado mostrado em cada célula do intervalo E4:E11 da planilha **Pintura** é a cor resultante para cada uma das oito partes de uma figura geométrica qualquer. Os resultados de cada uma das oito células desse intervalo têm duas partes diferentes, a primeira se relaciona com a seleção aleatória do nome da cor e a segunda com a formatação da cor da célula correspondente ao nome da cor.

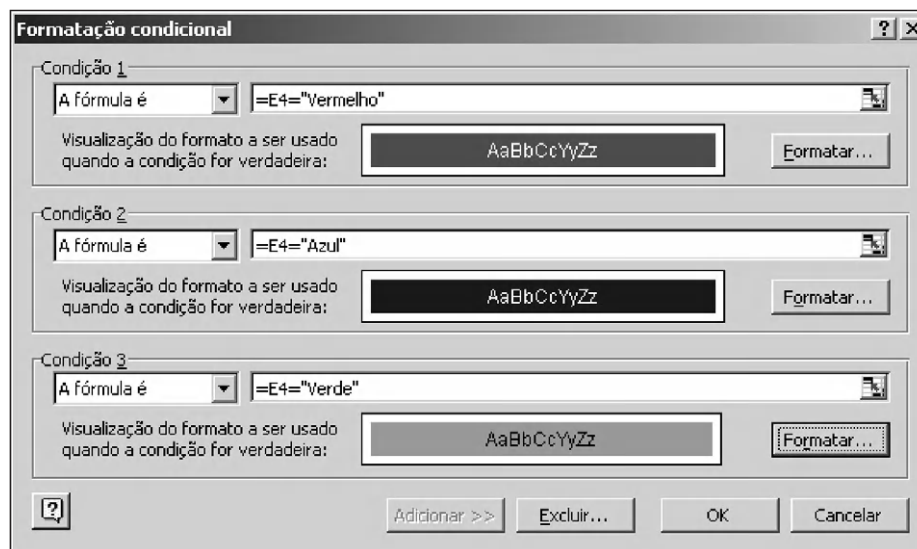
Seleção da cor de cada uma das oito partes da figura geométrica.

- No intervalo B4:B7 foram registrados os nomes das quatro cores pintando cada célula com a cor correspondente ao nome registrado.
- A fórmula =ÍNDICE(\$B\$4:\$B\$7;ALEATÓRIOENTRE(1;4)) foi registrada na célula E4 e depois foi copiada até a célula E11. Essa fórmula seleciona de forma aleatória uma das quatro cores utilizando as funções ÍNDICE e ALEATÓRIOENTRE. Toda vez que a planilha for recalculada a função ALEATÓRIOENTRE seleciona um dos quatro números {1, 2, 3, 4}. Com essa informação a função ÍNDICE seleciona a cor correspondente registrada no intervalo B4:B7 da planilha, sendo que o número 1 corresponde à cor registrada na célula B4 (Amarelo), o número 2 corresponde à cor registrada na célula B5 (Vermelho) e da mesma forma com os números 3 e 4.

- **ÍNDICE**(*matriz; núm_linha; núm_coluna*)

A função ÍNDICE²⁰ retorna um valor ou a referência a um valor do argumento *matriz*, tabela ou intervalo que neste caso é o intervalo \$B\$4:\$B\$7 que recebeu os cifrões para facilitar a cópia da fórmula em todo o intervalo E4:E11.

- O argumento *núm_linha* seleciona a linha na matriz a partir da qual um valor deverá ser retornado, se *núm_linha* for omitido, o argumento *núm_coluna* será obrigatório. Neste caso, a seleção da linha é realizada pela fórmula ALEATÓRIOENTRE(1;4).
- O argumento *núm_coluna* seleciona a coluna na matriz a partir da qual um valor deverá ser retornado; se *núm_coluna* for omitido, *núm_linha* será obrigatório. Neste caso, este argumento foi omitido.




Formatar a cor da célula com o nome da cor registrada na célula.

A formatação da cor da célula correspondente ao nome da cor é realizada com o comando *Formatação condicional* do Excel procedendo como segue:

- Selecione o intervalo E4:E11 e pinte as células de cor amarela forte e a fonte de cor preta com negrito.
- Clique na célula E4 e no menu **Formatar** selecione **Formatação condicional**.
- Na **Condição 1** selecione **A fórmula é** e ao lado registre a fórmula `=E4="Vermelho"` como mostra a figura na página seguinte. Isso indica que sempre que a condição `E4="Vermelho"` for verdadeira o Excel formatará a célula E4 como especificado a seguir; caso contrário, a célula continuará com a cor amarela forte e fonte de cor preta com negrito.
- Clique no botão **Formatar** e selecione as seguintes alternativas.
 - Fonte. Mantendo o corpo, escolher **Negrito** com cor branca.
 - Borda. Não realizar nenhuma seleção.
 - Padrões. Escolher a cor vermelha para a célula.
- Na **Condição 2** selecione **A fórmula é** e ao lado registre a fórmula `=E4="Azul"` como mostra a figura acima. Depois proceda como na Condição 1 mudando apenas a cor da célula para azul e a cor da fonte para branco.
- Na **Condição 3** selecione **A fórmula é** e ao lado registre a fórmula `=E4="Verde"` como mostra a figura anterior. Depois proceda como na Condição 1 mudando apenas a cor da célula para verde sem necessidade de mudar a cor da fonte.
- Por último pressione o botão **OK**. Para conferir o resultado pressione a tecla de função F9 e verifique a seleção do nome da cor e a formatação da cor da célula.

Para copiar a formatação condicional da célula E4 no intervalo E5:E11 proceda como segue:

- Selecione a célula E4.
- No menu Editar selecione **Copiar** ou pressione as teclas **Control+C**, ou pressione o ícone copiar .
- Selecione no intervalo E5:E11.
- No menu **Editar** selecione **Colar especial**. No grupo **Colar** da caixa de diálogo **Copiar especial** selecione **Formatos**.
- Para terminar pressione o botão **OK**.

A figura seguinte mostra outra forma de utilizar a função ÍNDICE, registrada a partir da célula J1 da planilha **Pintura** incluída na pasta **Capítulo 1**. Neste caso não é utilizada a base de dados do intervalo E4:E11, sendo os quatro elementos da população de cores {"Amarelo";"Vermelho";"Azul";"Verde"} registrados como matriz na própria fórmula como a da célula M4:

`=ÍNDICE({"Amarelo";"Vermelho";"Azul";"Verde"}; ALEATÓRIOENTRE(1;4))`

que depois foi copiada até a célula M11. Para terminar, as células do intervalo E4:E11 receberam a formatação condicional copiada da célula E4.

	J	K	L	M	N	O	P	Q	R	S
1	Utilizando somente fórmulas									
2										
3			Parte	Cor						
4			1	Verde	=ÍNDICE({"Amarelo";"Vermelho";"Azul";"Verde";ALEATÓRIOENTRE(1;4))					
5			2	Vermelho						
6			3	Azul						
7			4	Amarelo						
8			5	Verde	=ÍNDICE({"Amarelo";"Vermelho";"Azul";"Verde";ALEATÓRIOENTRE(1;4))					
9			6	Vermelho						
10			7	Azul						
11			8	Verde						
12										

Nos dois casos, pressionando a tecla de função F9 serão obtidas novas combinações de cores.

A fórmula do segundo procedimento do Exemplo 1.6 deve ser utilizada em populações pequenas, pois em populações grandes o registro de todos os nomes pode ser muito trabalhoso, sujeito a erros de registro e até a estourar a capacidade de armazenamento das células do Excel. O exemplo seguinte mostra outra forma de realizar uma amostragem probabilística com reposição.

EXEMPLO 1.7

A tabela seguinte registra a relação das 50 Maiores Empresas Privadas por Vendas do Brasil no ano 2002.²¹ O objetivo é retirar uma amostra aleatória sem reposição de tamanho 10 utilizando a tabela de números aleatórios. A tabela das maiores empresas está registrada na planilha **50 Maiores 2002** incluída na pasta **Capítulo 1** no material disponibilizado no site da editora.

Ordem	Empresa – Ramo	Vendas	Ordem	Empresa – Ramo	Vendas
1	TELEMAR – Telecomunicações	\$ 6.303,7	26	GERDAU – Siderurgia e metalurgia	\$ 2.078,9
2	TELEFÔNICA – Telecomunicações	\$ 5.480,5	27	LIGHT – Serviços públicos	\$ 2.003,6
3	CBB/AMBEV – Alimentos, bebidas e fumo	\$ 5.329,8	28	USIMINAS – Siderurgia e metalurgia	\$ 1.891,8
4	VOLKSWAGEN – Automotivo	\$ 5.295,2	29	REFAP – Química e petroquímica	\$ 1.886,1
5	PETRÓLEO IPIRANGA – Atacado e comércio exterior	\$ 4.214,1	30	VARIG – Serviços de transporte	\$ 1.868,6
6	SHELL – Atacado e comércio exterior	\$ 4.096,8	31	BRASKEM – Química e petroquímica	\$ 1.793,3
7	GENERAL MOTORS – Automotivo	\$ 4.092,7	32	SADIA – Alimentos, bebidas e fumo	\$ 1.760,4
8	CARREFOUR – Comércio varejista	\$ 4.044,9	33	TELESP CELULAR – Telecomunicações	\$ 1.752,1
9	BRASIL TELECOM – Telecomunicações	\$ 3.975,9	34	CASAS BAHIA – Comércio varejista	\$ 1.690,7
10	GRUPO PÃO DE AÇÚCAR – Comércio varejista	\$ 3.837,5	35	IBM – Tecnologia e computação	\$ 1.591,8
11	EMBRATEL – Telecomunicações	\$ 3.668,3	36	DAIMLERCHRYSLER – Automotivo	\$ 1.557,2
12	VALE DO RIO DOCE – Mineração	\$ 3.418,0	37	CPFL – Serviços públicos	\$ 1.551,2
13	BUNGE ALIMENTOS – Alimentos, bebidas e fumo	\$ 3.158,1	38	COPERSUCAR – Atacado e comércio exterior	\$ 1.550,5
14	FIAT – Automotivo	\$ 3.121,4	39	SIEMENS – Eletroeletrônico	\$ 1.528,8
15	ELETROPAULO METROPOLITANA – Serviços públicos	\$ 3.078,0	40	COPESUL – Química e petroquímica	\$ 1.465,8

Ordem	Empresa – Ramo	Vendas	Ordem	Empresa – Ramo	Vendas
16	EMBRAER – Automotivo	\$ 2.945,3	41	TAM – Serviços de transporte	\$ 1.397,0
17	TEXACO – Atacado e comércio exterior	\$ 2.805,2	42	BASF – Química e petroquímica	\$ 1.355,1
18	NESTLÉ – Alimentos, bebidas e fumo	\$ 2.762,7	43	COSIPA – Siderurgia e metalurgia	\$ 1.340,0
19	CARGILL – Alimentos, bebidas e fumo	\$ 2.709,1	44	PERDIGÃO AGROINDUSTRIAL – Alim., beb. e fumo	\$ 1.336,2
20	ESSO – Atacado e comércio exterior	\$ 2.688,5	45	NOKIA – Eletroeletrônico	\$ 1.300,0
21	ITAIPÚ BINACIONAL – Serviços públicos	\$ 2.529,6	46	BUNGE FERTILIZANTES – Química e petroquímica	\$ 1.297,5
22	UNILEVER – Farmacêutico, higiene e cosméticos	\$ 2.456,9	47	SONAE – Comércio varejista	\$ 1.156,5
23	FORD MOTOR – Automotivo	\$ 2.387,6	48	KLABIN PAPEL CELULOSE – Papel e celulose	\$ 1.155,1
24	SOUZA CRUZ – Alimentos, bebidas e fumo	\$ 2.375,9	49	PONTO FRIIO – Comércio varejista	\$ 1.153,3
25	CSN – Siderurgia e metalurgia	\$ 2.160,4	50	MAKRO – Atacado e comércio exterior	\$ 1.127,2

Solução. Começando em qualquer ponto da tabela, a escolha dos números aleatórios pode ser realizada por coluna, por linha ou pulando entre elas. Escolhemos como ponto de partida o número aleatório 0617 da coluna 2 da linha 11, como mostra a seguinte tabela parcial de números aleatórios.

8395	0617	4946	5390	8008	2785	7629	3176	5114	1410
3069	5769	3617	1149	0276	5783	2837	7487	8159	3478
1859	8790	3106	7156	5673	6967	0812	1603	1330	5588
9645	7574	2954	5940	6263	6559	9450	2281	1362	3000
1136	6008	0598	8617	2380	0960	4412	7829	2840	8729

Como a população tem 50 elementos ou empresas para realizar as seleções serão utilizados os dois últimos algarismos de cada número aleatório da tabela acima.

- Do número 0617 são escolhidos 17.
- Do número 4946 os algarismos 46.
- A seguir deveríamos selecionar o número 5390, mas como 90 é maior que 50 continuamos até o número 8008 escolhendo 08.
- Continuamos este procedimento de escolha até completar a amostra de tamanho 10 identificada com a seguinte relação de números de ordem da tabela das 50 empresas: 17, 46, 08, 29, 14, 10, 30, 03, 12 e 50.

Com os números aleatórios selecionados foi construída a tabela seguinte.

Amostra	Números	Empresa – Ramo	Vendas
1	17	TEXACO – Atacado e comércio exterior	\$ 2.805,2
2	46	BUNGE FERTILIZANTES – Química e petroquímica	\$ 1.297,5
3	08	CARREFOUR – Comércio varejista	\$ 4.044,9
4	29	REFAP – Química e petroquímica	\$ 1.886,1
5	14	FIAT – Automotivo	\$ 3.121,4
6	10	GRUPO PÃO DE AÇÚCAR – Comércio varejista	\$ 3.837,5
7	30	VARIG – Serviços de transporte	\$ 1.868,6
8	03	CBB/AMBEV – Alimentos, bebidas e fumo	\$ 5.329,8
9	12	VALE DO RIO DOCE – Mineração	\$ 3.418,0
10	50	MAKRO – Atacado e comércio exterior	\$ 1.127,2

EXEMPLO 1.8

Construa um modelo para extrair uma amostra probabilística simples com reposição de dez empresas da tabela das cinquenta primeiras empresas privadas por vendas no ano 2002.

A	B	C	D	E	F	G	H	I
1	As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame							
2								
3	Ordem	Empresa - Ramo	Vendas					
4	1	TELEMAR - Telecomunicações	\$ 6.303,7					
5	2	TELEFÔNICA - Telecomunicações	\$ 5.480,5					
6	3	CBB/AMBEV - Alimentos, bebidas e fumo	\$ 5.329,6					
7	4	VOLKSWAGEN - Automotivo	\$ 5.295,2					
8	5	PETROLEO IPIRANGA - Atacado e comércio exterior	\$ 4.214,1					
9	6	SHELL - Atacado e comércio exterior	\$ 4.096,8					
10	7	GENERAL MOTORS - Automotivo	\$ 4.092,7					
11	8	CARREFOUR - Comércio varejista	\$ 4.044,9					
12	9	BRASIL TELECOM - Telecomunicações	\$ 3.975,9					
13	10	GRUPO PAO DE AÇÚCAR - Comércio varejista	\$ 3.837,5					
14	11	EMBRATEL - Telecomunicações	\$ 3.688,3					
15	12	VALE DO RIO DOCE - Mineração	\$ 3.418,0					
16	13	BUNCE ALIMENTOS - Alimentos, bebidas e fumo	\$ 3.158,1					

Amostragem com Reposição			
Amostra	Empresa - Ramo	Vendas	
1	43	COSIPA - Siderurgia e metalurgia	\$ 1.340,00
2	45	NOKIA - Eletroeletrônico	\$ 1.300,00
3	22	UNILEVER - Farmacêutico, higiene e cosméticos	\$ 2.456,90
4	33	TELESP CELULAR - Telecomunicações	\$ 1.752,10
5	10	GRUPO PAO DE AÇÚCAR - Comércio varejista	\$ 3.837,50
6	18	NESTLÉ - Alimentos, bebidas e fumo	\$ 2.762,70
7	26	USIMINAS - Siderurgia e metalurgia	\$ 1.091,60
8	2	TELEFÔNICA - Telecomunicações	\$ 5.480,50
9	6	SHELL - Atacado e comércio exterior	\$ 4.096,80
10	46	NINGE FERTILIZANTES - Química e petroquímica	\$ 1.297,50

Pressione a tecla F9 para gerar novos dígitos aleatórios

Solução. Nas colunas B, C e D da planilha **Amostragem com Reposição** incluída no **Capítulo 1** foram registradas a **Ordem**, a **Empresa – Ramo** e as **Vendas** das 50 maiores empresas por vendas no ano 2002, dados copiados da planilha **50 Maiores 2002**. A partir da célula F5 foi construída a tabela que extrairá as amostras aleatórias de tamanho dez utilizando a função ALEATÓRIOENTRE com limite inferior 1 e limite superior 50. Para facilitar o controle, na coluna F foi registrada a ordem da amostragem.

- Na coluna G são gerados os números aleatórios entre os limites 1 e 50. Na célula G5 foi registrada a fórmula =ALEATÓRIOENTRE(1;50) que depois foi copiada até a célula G14.
- A fórmula registrada na célula H5 =PROCV(\$G5;\$B\$4:\$D\$53;2) foi copiada até a célula H14. A partir dos números aleatórios gerados na coluna G, estas fórmulas *procuram* o nome da empresa amostrada na tabela das 50 empresas. No Apêndice 3 deste capítulo está descrita a função PROCV de procura vertical e sua equivalente função PROCH para procura horizontal.
- Finalizando, a fórmula =PROCV(\$G5;\$B\$4:\$D\$53;3) registrada na célula I6 procura o valor das vendas da empresa amostrada; depois essa fórmula foi copiada até a célula I15.
- Cada vez que for pressionada a tecla de função F9 será obtido um novo grupo de 10 amostras que poderá conter mais de uma vez uma mesma empresa. Sugerimos que o leitor se familiarize com este procedimento e com o significado da amostragem probabilística simples com reposição retirando amostras sucessivas com F9.

Como complemento, a partir da linha 18 da planilha **Amostragem com Reposição** foi construída outra tabela de amostragem utilizando a função ÍNDICE apresentada no Exemplo 1.6. A fórmula geradora de números aleatórios ALEATÓRIOENTRE(1;50) não pode ser utilizada dentro da função ÍNDICE, pois com o mesmo número aleatório serão extraídos dois dados da tabela da população, os campos *Empresa – Ramo* e *Vendas*.

Das dez empresas amostradas no Exemplo 1.8 três delas foram selecionadas duas vezes, pois todas as amostras extraídas com esse procedimento são realizadas com reposição. Para tentar selecionar amostras sem reposição com o mesmo modelo anterior e de forma manual, a planilha deverá ser recalculada tantas vezes quanto seja necessário até conseguir uma amostra com dez empresas diferentes.

Ferramentas de análise do Excel

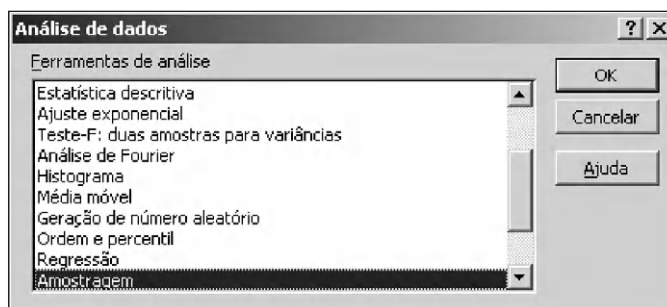
Até esta parte do livro utilizamos algumas das muitas funções estatísticas da planilha Excel²² sendo que algumas delas estão sempre disponíveis quando o aplicativo Excel é carregado, e as outras funções ficam disponíveis depois de instalar o suplemento *Ferramentas de análise* como é mostrado no Apêndice 1 deste capítulo.

²² O Excel também dispõe de funções financeiras, matemáticas, de engenharia etc.

O Excel também dispõe de um conjunto de ferramentas para análise de dados denominadas de forma genérica como *Ferramentas de análise*. Essas ferramentas apresentam soluções integradas de análises estatísticas. Para ver a relação de ferramentas de análise disponíveis dentro da planilha Excel, depois de selecionar **Análise de dados** dentro do menu **Ferramentas** o Excel apresentará a caixa de diálogo da Figura 1.7.

- Pressionando o botão **Ajuda** dessa caixa de diálogo o Excel apresentará a página *Sobre as ferramentas de análise estatística* pertencente à *Ajuda do Excel*.

FIGURA 1.7 Caixa de diálogo das **Ferramentas de análise**.



Na caixa de diálogo **Análise de dados** selecione o nome da ferramenta de análise que deseja utilizar, por exemplo, *Amostragem* e depois pressione o botão **OK**. A seguir o Excel apresentará uma caixa de diálogo com o nome da ferramenta selecionada, **Amostragem**, onde você informará os dados requeridos e definirá, em geral, as opções de análise e de resultados desejados. As caixas de diálogos das ferramentas incluem um botão de **Ajuda** onde poderão ser obtidas algumas informações sobre as opções das análises. Se a opção **Análise de dados** não estiver disponível, você precisará carregar o programa suplementar de *Ferramentas de análise* como é mostrado no Apêndice 1 deste capítulo.

Ferramenta de análise *Amostragem*

O Excel dispõe da ferramenta de análise *Amostragem* para extrair amostras probabilísticas simples com reposição de uma população de valores numéricos com distribuição uniforme e discreta. Também dispõe da ferramenta de análise *Geração de Número Aleatório* para extrair amostras probabilísticas simples com reposição de uma população de valores numéricos com outros tipos de distribuições, incluindo a uniforme, tema apresentado no Apêndice 1 do Capítulo 8. Antes de utilizar a ferramenta *Amostragem*²³ deve-se preparar uma planilha com os dados numéricos da população que será amostrada e registrados numa coluna de onde será retirada a amostra. Para compreender o uso da ferramenta *Amostragem*, o Exemplo 1.8 foi resolvido na planilha **Ferramenta Amostragem** incluída na pasta **Capítulo 1**. Depois de copiar os dados da planilha **50 Maiores 2002** proceda como segue:

- Depois de selecionar **Análise de dados** dentro do menu **Ferramentas** o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, Figura 1.7.
- Escolhendo a ferramenta **Amostragem** e depois pressionando o botão **OK** você receberá a caixa de diálogo **Amostragem** mostrada na Figura 1.8, depois de selecionadas algumas opções.
 - Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Amostragem* pertencente à *Ajuda do Excel*.

²³ Em inglês, a ferramenta de análise AMOSTRAGEM é *SAMPLING*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo da ferramenta *Amostragem*, como mostra a Figura 1.8, são:

- **Intervalo de entrada:** Informar o intervalo de células da planilha onde os dados estão registrados, neste caso o intervalo D3:D53 que inclui a célula onde foi registrado o título *Vendas*, ou rótulo no Excel.
- **Rótulos.** Seleccionamos este item, pois o intervalo informado D3:D53 inclui o título *Vendas*.

No quadro **Método de amostragem** há duas escolhas:

- Escolhendo **Periódico** e informando o **Período**, serão retiradas amostras com período constante. Por exemplo, informando 5 na caixa **Período**, a ferramenta *Amostragem* retirará 10 amostras com periodicidade 5, começando pelo quinto dado da tabela. De outra maneira, retirará o primeiro dado do quinto lugar da tabela, depois o dado na posição 10 e assim sucessivamente até o dado registrado na posição 50.
- Escolhendo **Aleatório** serão retiradas amostras probabilísticas simples até completar o número de amostras registrado na caixa **Número de amostras**, neste caso 10.

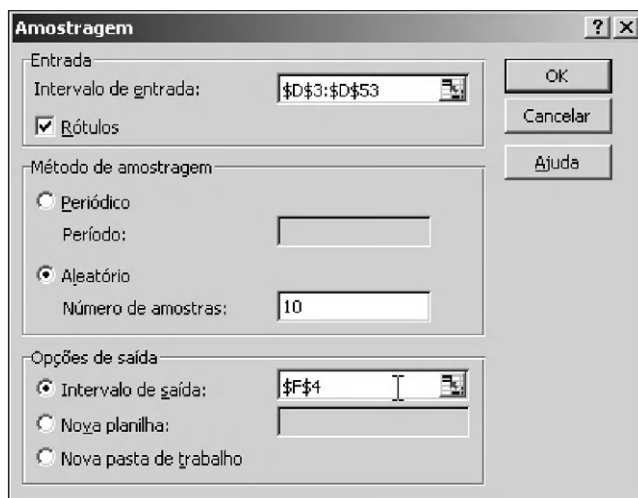


FIGURA 1.8 Caixa de diálogo **Amostragem** probabilística simples.

No quadro **Opções de saída** deve ser obrigatoriamente informado um endereço, a partir do qual a ferramenta *Amostragem* registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três *botões de opção* que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, neste caso F4. Depois de clicar com o botão esquerdo do mouse dentro da caixa correspondente, o endereço pode ser registrado digitando F4, ou *clikando* com o botão esquerdo do *mouse* na célula F4, neste caso será registrado o endereço com os dois cifrões, \$F\$4. Esse endereço é o da célula superior esquerda da tabela que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
- Se não for informado nenhum endereço, a ferramenta inserirá uma nova planilha com o nome **Plan** seguido de um número sequencial. Ao escolher essa alternativa na pasta **Capítulo 1**, a ferramenta inserirá a planilha **Plan1**.

- Há a alternativa de informar o nome da planilha na caixa desta alternativa. Ao registrar o nome *Teste* a ferramenta inserirá na mesma pasta uma nova planilha com o nome *Teste*.
- **Nova pasta de trabalho.** Os resultados serão apresentados numa nova pasta e a partir da célula A1 da planilha *Plan1*.

A Figura 1.9 mostra uma amostragem probabilística simples de tamanho dez extraída com a ferramenta *Amostragem*. Cada vez que for ativada a ferramenta *Amostragem* será extraída, em geral, uma amostra diferente. Essa ferramenta é útil para gerar amostras aleatórias com reposição de tamanho determinado pelo leitor e a partir de uma lista de dados; entretanto, a ferramenta extrai somente valores numéricos.

FIGURA 1.9
Amostragem
probabilística simples
com a ferramenta
Amostragem.

	A	B	C	D	E	F	G	H
1	As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame				Ferramenta de Análise <i>Amostragem</i>			
2								
3		Ordem	Empresa - Ramo	Vendas		Vendas		
4		1	TELEMAR - Telecomunicações	\$ 6.303,7		5480,5		
5		2	TELEFÔNICA - Telecomunicações	\$ 5.480,5		4096,8		
6		3	CBB/AMBEV - Alimentos, bebidas e fumo	\$ 5.329,8		1127,2		
7		4	VOLKSWAGEN - Automotivo	\$ 5.295,2		6303,7		
8		5	PETRÓLEO IPIRANGA - Atacado e comércio exterior	\$ 4.214,1		3837,5		
9		6	SHELL - Atacado e comércio exterior	\$ 4.096,8		1465,8		
10		7	GENERAL MOTORS - Automotivo	\$ 4.092,7		5295,2		
11		8	CARREFOUR - Comércio varejista	\$ 4.044,9		2805,2		
12		9	BRASIL TELECOM - Telecomunicações	\$ 3.975,9		1886,1		
13		10	GRUPO PÃO DE AÇÚCAR - Comércio varejista	\$ 3.837,5		1465,8		
14		11	EMBRATEL - Telecomunicações	\$ 3.668,3				
15		12	VALE DO RIO DOCE - Mineração	\$ 3.418,0				
16		13	BUNGE ALIMENTOS - Alimentos, bebidas e fumo	\$ 3.158,1				
17		14	FIAT - Automotivo	\$ 3.121,4				
18		15	ELETROPAULO METROPOLITANA - Serviços públicos	\$ 3.078,0				
19		16	EMBRAER - Automotivo	\$ 2.945,3				
20		17	TEXACO - Atacado e comércio exterior	\$ 2.805,2				
21		18	NESTLÉ - Alimentos, bebidas e fumo	\$ 2.762,7				
22		19	CARGILL - Alimentos, bebidas e fumo	\$ 2.709,1				
23		20	ESSO - Atacado e comércio exterior	\$ 2.688,5				
24		21	ITAIPU BINACIONAL - Serviços públicos	\$ 2.529,6				
25		22	UNILEVER - Farmacêutico, higiene e cosméticos	\$ 2.456,9				
26		23	FORD MOTOR - Automotivo	\$ 2.387,6				
27		24	SOUZA CRUZ - Alimentos, bebidas e fumo	\$ 2.375,9				

Amostragem periódica

Vendas

4214,1

3837,5

3078

2688,5

2160,4

1868,6

1591,8

1465,8

1300

1127,2

A Figura 1.10 mostra a caixa de diálogo *Amostragem* com os dados para selecionar uma amostra periódica com periodicidade cinco na população das 50 maiores empresas.

FIGURA 1.10 Caixa de
diálogo *Amostragem*
periódica com
periodicidade cinco.

Amostragem

Entrada

Intervalo de entrada:

☒ Rótulos

Método de amostragem

☒ Periódico

Período:

☐ Aleatório

Número de amostras:

Opções de saída

☒ Intervalo de saída:

☐ Nova planilha:

☐ Nova pasta de trabalho

OK

Cancelar

Ajuda

Nas dez amostras registradas a partir da célula F17, Figura 1.9, observe que a primeira amostra retirada corresponde ao valor da quinta posição do intervalo D3:D53. A amostra seguinte ao valor da posição dez e assim sucessivamente até a última amostra que corresponde ao último registro da tabela, a posição dez, como se pode verificar comparando os valores extraídos com os valores extraídos com as vendas da população. Se a amostragem for repetida com os mesmos dados as amostras serão as mesmas. O procedimento de seleção desta ferramenta não acompanha a recomendação técnica de realizar uma amostragem probabilística simples nos cinco primeiros dados da tabela que correspondem à periodicidade cinco. A partir desse primeiro resultado será aplicada a periodicidade desejada. Também se deve tomar cuidado com a periodicidade escolhida, pois esse valor definirá o tamanho da amostra; por exemplo, se for escolhida a periodicidade dez no nosso exemplo será recebida uma amostragem de tamanho cinco.

Amostragens aleatórias sem reposição

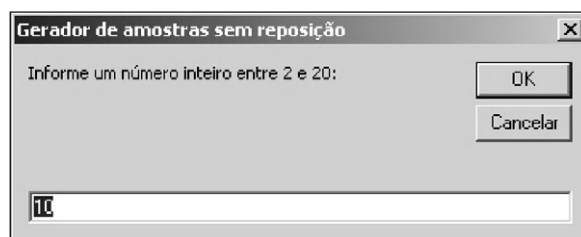
Para selecionar amostras sem reposição com os procedimentos de amostragem com reposição, a planilha deverá ser recalculada tantas vezes quanto seja necessário até conseguir uma amostra com dez empresas diferentes. Para facilitar o procedimento de amostragem sem reposição foi construído o modelo do qual se pode extrair de duas a vinte amostras sem reposição da tabela das 50 primeiras empresas, como mostra o Exemplo 1.9.

EXEMPLO 1.9

Construa um modelo para extrair uma amostra probabilística simples sem reposição de dez empresas da tabela das 50 primeiras empresas privadas por vendas no ano 2002.

Solução. Começamos por preparar a planilha denominada **Amostragem sem Reposição** incluída na pasta **Capítulo 1**, com o mesmo *layout* da planilha utilizada para extrair amostras com reposição. A diferença com aquela planilha está na escolha dos números aleatórios da coluna **Ordem** que não podem ser repetidos. Como a seleção de *números aleatórios não repetidos* não pode ser realizada com os recursos da planilha foi construído um procedimento combinando os recursos da planilha Excel com *macros* em VBA. A *macro* principal é ativada com o botão **Nova Amostragem** e a operação do modelo é a seguinte:

- Depois de pressionar o botão **Nova Amostragem** o modelo apresenta a caixa de entrada de dados **Gera-dor de amostras sem reposição** solicitando a informação do tamanho da amostra, valor entre 2 e 20 com ambos limites incluídos, como mostra a figura seguinte depois de informar o valor 10 que também é o valor *default* dessa caixa.



- Depois de pressionar o botão **OK** é ativada a macro que selecionará as dez amostras desejadas, como mostra a figura seguinte.

A	B	C	D	E	F	G	H	I
As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame								
				Nova Amostragem				
Ordem	Empresa - Ramo		Vendas	Amostragem sem Reposição				
1	Amostra	Empresa - Ramo	Vendas					
2	1	TELEMAR - Telecomunicações	\$ 6.303,7	1	21	ITAIPU BINACIONAL - Serviços públicos	\$ 2.529,6	
3	2	TELEFÔNICA - Telecomunicações	\$ 5.480,5	2	48	KLABIN PAPEL CELULOSE - Papel e celulose	\$ 1.155,1	
4	3	CBBIAMBEV - Alimentos, bebidas e fumo	\$ 5.329,8	3	6	SHELL - Atacado e comércio exterior	\$ 4.096,8	
5	4	VOLKSWAGEN - Automotivo	\$ 5.295,2	4	46	BUNGE FERTILIZANTES - Química e petroquímica	\$ 1.297,5	
6	5	PETRÓLEO IPIRANGA - Atacado e comércio exterior	\$ 4.214,1	5	31	BRASKEM - Química e petroquímica	\$ 1.793,3	
7	6	SHELL - Atacado e comércio exterior	\$ 4.096,8	6	18	NESTLÉ - Alimentos, bebidas e fumo	\$ 2.762,7	
8	7	GENERAL MOTORS - Automotivo	\$ 4.092,7	7	8	CARREFOUR - Comércio varejista	\$ 4.044,9	
9	8	CARREFOUR - Comércio varejista	\$ 4.044,9	8	24	SOUZA CRUZ - Alimentos, bebidas e fumo	\$ 2.375,9	
10	9	BRASIL TELECOM - Telecomunicações	\$ 3.975,9	9	11	EMBRATEL - Telecomunicações	\$ 3.668,3	
11	10	GRUPO PÃO DE AÇÚCAR - Comércio varejista	\$ 3.837,5	10	49	PONTO FRIO - Comércio varejista	\$ 1.153,3	
12	11	EMBRATEL - Telecomunicações	\$ 3.668,3					
13	12	VALE DO RIO DOCE - Mineração	\$ 3.418,0					

As características desse *modelo* de amostragem sem reposição são:

- Toda vez que for pressionado o botão **Nova Amostragem** deverá ser informado o tamanho da amostra desejada, um valor entre 2 e 20 com ambos extremos incluídos. Se for informado um valor fora desse intervalo o modelo apresentará uma caixa de diálogo informando esse dado incorreto. Se o valor informado for correto o modelo selecionará uma nova amostra sem repetição, em geral, diferente da anterior.
- No intervalo de células da planilha K4:K25 é realizado o controle da macro para a seleção das amostras sem reposição.
- O código da macro pode ser visto dentro do Editor de VBA, pressionando simultaneamente as teclas ALT + F11 dentro da planilha Excel.

No Apêndice 4 você encontra um modelo para retirada de amostras sem reposição que tem a vantagem de poder selecionar qualquer informação contida na célula, incluindo valores não numéricos.

Outros tipos de amostragens

Na caixa de diálogo da ferramenta de análise *Amostragem* pode-se escolher um dos dois métodos de amostragem incluídos, *Periódico* e *Aleatório*. Para realizar as amostragens apresentadas na Figura 1.9, escolhemos o procedimento *Aleatório* para a amostra registrada a partir da célula F4 e o procedimento *Periódico* para a amostra registrada a partir da célula F17 com periodicidade cinco, começando pela observação da população na quinta posição, sendo esse tipo de amostragem denominado *amostragem sistemática*. Uma variante recomendada desse tipo de amostragem é escolher a primeira observação de forma aleatória. Por exemplo, antes de iniciar a amostragem com reposição da tabela das *50 Primeiras Empresas por Vendas* o número de empresas cinquenta é dividido pelo tamanho da amostra dez, obtendo cinco grupos contendo dez empresas cada um. Do primeiro grupo de dez empresas uma delas é selecionada de forma aleatória, por exemplo, a amostra da posição seis, e em sequência são extraídas as empresas das posições 16, 26, 36 e 46. Em alguns casos a amostragem sistemática pode ser melhor que a simples amostragem aleatória, pois essa amostragem colhe observações em toda a extensão da população. Em outros casos, este tipo de amostragem pode colher eventos periódicos com o mesmo período da amostragem e comprometer a amostra. Por exemplo, se da máquina de produzir comprimidos com 36 punções retirarmos um comprimido a cada 36 comprimidos produzidos, a variabilidade dos comprimidos da amostra será menor que o da população.

Se algumas características da população forem conhecidas antes da amostragem será possível dividir a população em camadas sem superposição,²⁴ ou estratos, e extrair uma amostra aleatória com melhores resultados com representatividade de cada camada ou estrato. Na amostragem por *conglomerado*, em vez de sortear a população são sorteados territórios desde os estados, municípios, bairros e domicílios até a pessoa que será entrevistada. Outro procedimento é a amostragem por *cotas* em que não há sorteio, apenas se tomam amostras proporcionais ao tamanho de cada grupo previamente definido, homem, mulher etc.

²⁴ Sem superposição significa que a interseção dos conjuntos é vazia.

Como são feitas as pesquisas²⁵

O Datafolha não tem como ouvir todos os eleitores no Brasil. Assim, utiliza métodos estatísticos para aferir a intenção de voto de modo que os entrevistados representem o conjunto do eleitorado.

A Pesquisa

Antes de iniciar uma pesquisa, o Datafolha sabe quantas pessoas vai ouvir e o perfil de cada grupo, para que o conjunto do eleitorado seja representado na mostra de forma proporcional. Antes de sair às ruas, os entrevistadores sabem quantas pessoas em cada lugar têm de ouvir e quais são os lugares. Estando lá, o entrevistador escolhe aleatoriamente os entrevistados, sempre respeitando as faixas de sexo e de idade que compõem o conjunto do eleitorado.

Amostra

O Datafolha faz estudos prévios para saber como é composto o conjunto do eleitorado. O objetivo é que a amostra seja representativa do total de eleitores. Dessa forma, os resultados obtidos na pesquisa podem ser, estatisticamente, ampliados para os milhões de eleitores no Brasil (ou, os eleitores de cada Estado pesquisado).

Como é feito o estudo prévio?

Antes de fazer uma pesquisa, o Datafolha colhe informações nos TREs dos Estados para saber, no conjunto dos eleitores, quantos são homens, quantos são mulheres, quantos estão em cada faixa de idade pesquisada, quantos moram na capital e quantos moram no interior.

Margem de erro

Como não é possível ouvir todos os eleitores, os resultados obtidos na pesquisa são aproximados. Chama-se margem de erro o intervalo – para mais ou para menos – que deve ser considerado para os dados divulgados; por exemplo, a margem de erro é de dois pontos percentuais para São Paulo. Qualquer valor dentro desse intervalo deve ser considerado correto. Na pesquisa para os Estados, a margem de erro da pesquisa também é de dois pontos percentuais para Rio, Minas e Rio Grande do Sul. Para o Distrito Federal, é de três pontos.

Voto espontâneo

É aquele em que não há estímulo. O entrevistador pergunta: “Em quem você gostaria de votar no segundo turno da eleição?”

Voto estimulado

Neste tipo de pergunta, o entrevistado é estimulado. O pesquisador pergunta: “Se a eleição para governador fosse hoje, em quem você votaria: X ou Y.” O entrevistador diz, e a resposta é anotada.

Urna eletrônica

As eleições com urnas eletrônicas podem alterar o resultado final em relação à intenção de voto. Na votação manual, o eleitor recebe a cédula em que constam os nomes e os números dos candidatos ao go-

²⁵ Adaptado do caderno *Eleições* do jornal Folha de São Paulo, 18/10/1998.

verno. Ele marca um “X” no seu candidato. Na votação eletrônica, o eleitor precisa saber o número de seu candidato (e não apenas seu nome). A urna pede que ele digite o número. Se ele não souber, pode errar o voto. Assim mesmo que ele tenha a intenção de votar num candidato (e a pesquisa captou essa intenção) ele pode errar no momento da votação e acabar votando em outro candidato ou anular seu voto.

Votos válidos

São aqueles obtidos sem computar as abstenções (número de eleitores que não votaram), os votos brancos e os nulos. Quando o primeiro colocado numa eleição consegue 50% mais um voto dos votos válidos, não há segundo turno.

As pesquisas e os votos válidos

Os institutos de pesquisa usam o critério “votos válidos” apenas no final do período eleitoral. Isso porque o número de indecisos no início do processo eleitoral é normalmente tão grande que esse grupo certamente terá um peso no resultado final. Ou seja, uma parte dos que dizem não ter candidato vai acabar escolhendo algum. No final, o número de pessoas sem candidatos está mais consolidado (são aqueles que devem anular ou votar em branco).

Esclarecendo os métodos do ibope²⁶

GZM. Quais são os critérios para escolha desses domicílios na coleta de índices (de audiência)?

Dora. Quando se desenha uma amostra de audiência é preciso representar a situação da cidade onde você está pesquisando. Nós nos baseamos em dados do IBGE para sabermos quantos domicílios existem na Grande São Paulo, como eles estão divididos por regiões, qual o percentual de pessoas por sexo, por faixa etária, presença de crianças no domicílio etc. A partir daí, a gente faz a seleção do domicílio. Só que o IBGE não tem um levantamento de classe socioeconômica, e o Ibope passou a fazer o Levantamento Socioeconômico (LSE). Com isso, agregamos mais uma variável. Para fazer parte de uma amostra, o domicílio precisa preencher todos esses requisitos. Quanto mais representativa for a amostra, mais próximo você está de um resultado real.

²⁶ Trecho da entrevista da diretora do Ibope Dora Câmara ao jornalista Gonçalo Junior publicada no jornal *Gazeta Mercantil*, 14/01/2000.

Apêndice 1

Preparando o Excel antes de começar

No livro serão utilizadas *funções e ferramentas de análise* disponíveis no Excel que nem sempre são incorporadas ao iniciar o Excel. Tentando evitar aborrecimentos provenientes de uma instalação incompleta do Excel, sugerimos que o leitor realize a verificação a seguir.

Excel versão 2000

- No menu **Ferramentas** escolha **Suplementos**. O Excel apresentará a caixa de diálogo **Suplementos** com os **Suplementos** disponíveis.
- Os suplementos **Ferramentas de análise** e **Ferramentas de análise-VBA** devem estar selecionados como mostra a Figura 1.11.
- Aproveite e também selecione o suplemento **Solver** que será utilizado neste livro.

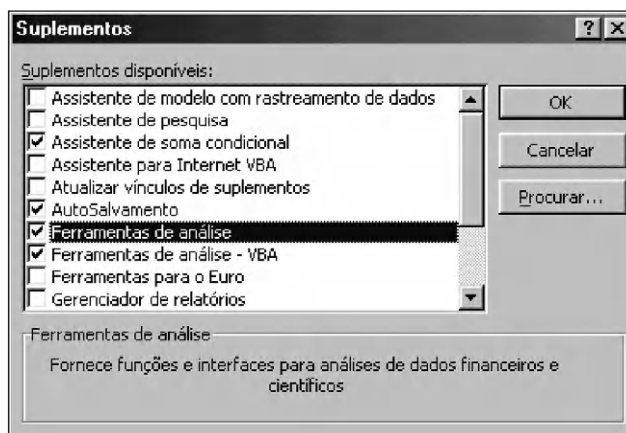


FIGURA 1.11 Caixa de mensagem *Suplementos*.

Excel versões 2002 e 2003

- No menu **Ferramentas** escolha **Suplementos**. O Excel apresentará a caixa de diálogo **Suplementos** com os **Suplementos** disponíveis.
- Os suplementos **Ferramentas de análise** e **Ferramentas de análise-VBA** devem estar selecionados como mostra a Figura 1.12. Depois de pressionar o botão **OK** as ferramentas de análise, bem como as funções especiais, estarão sempre disponíveis quando o aplicativo Excel for carregado.
- Aproveite e também selecione o suplemento **Solver** que será utilizado neste livro.

Para todas as versões do Excel

Se os suplementos **Ferramentas de análise**, **Ferramentas de análise-VBA** e **Solver** não aparecerem na caixa de diálogo **Suplementos**, então os dois suplementos não foram instalados junto com o Excel. Você deverá instalar esses arquivos incluídos no programa de instalação do Excel ou Microsoft Office correspondente.



FIGURA 1.12 Caixa de mensagem *Suplementos*.


Apêndice 2

Como registrar uma função na planilha Excel

Uma função do Excel pode ser registrada numa célula da planilha utilizando um dos três procedimentos seguintes:

- Digitando a fórmula, começando pelo sinal = seguido do nome da função requerida e os argumentos entre parênteses. Este procedimento exige que se lembre o nome da função, os argumentos necessários e sua sequência.
- Copiando a fórmula de outra célula onde a função tenha sido usada anteriormente. Este procedimento facilita a digitação, porém exige que se lembre o significado dos argumentos necessários.
- Utilizando o procedimento *Colar função* do Excel que elimina as desvantagens dos dois procedimentos anteriores.

O procedimento *Colar função* para registrar a função matemática ALEATÓRIOENTRE entre os limites 0 e 599 é o seguinte:

- Posicionar o cursor na célula onde será registrada a função.
- No menu **Inserir** escolher **Função**. O Excel apresentará a caixa de diálogo **Colar função**. No lugar de utilizar o menu **Inserir** se pode ativar a caixa de diálogo **Colar função** diretamente pressionando o botão  que deve ser incorporado na *Barra de Ferramentas Padrão*,²⁷ acompanhando o procedimento de *Adição de botões*.

²⁷ Outra forma de ativar a caixa de diálogo **Colar função** é abrir o menu **Inserir** e depois escolher **Função**.

- Na caixa **ou Selecione uma categoria**: escolha **Matemática e trigonométrica**, Figura 1.13.
- Na caixa **Selecione uma função** escolher **ALEATÓRIOENTRE**.
- Depois de pressionar o botão **OK** aparecerá a caixa de diálogo **Argumentos da função ALEATÓRIOENTRE** onde serão preenchidos os dados, Figura 1.14.

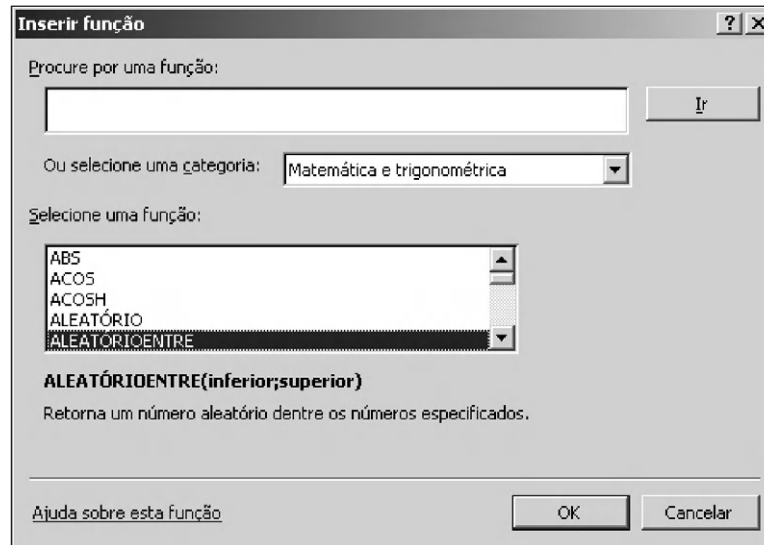


FIGURA 1.13
Selecione a função
ALEATÓRIOENTRE.

Perceba que ao mesmo tempo em que os dados são registrados:

- A caixa de diálogo descreve a função escolhida bem como cada argumento que está sendo registrado e à direita de cada campo é apresentado o valor informado.
- Depois de informar os argumentos da função **ALEATÓRIOENTRE**, na linha seguinte ao último dado é apresentado o resultado do cálculo da função **ALEATÓRIOENTRE**, neste caso 175, valor que *deveria* ser repetido na última linha **Resultado da fórmula** que neste caso é 559. Acreditamos que esta diferença seja provocada pelo resultado de outra rodada de cálculo, pois cada vez que o Excel for recalculado o resultado da função **ALEATÓRIOENTRE**, em geral, será diferente.
- Durante o preenchimento dos dados, na *barra de fórmulas* do Excel é construída a fórmula **=ALEATÓRIOENTRE(0;599)** que será inserida na célula escolhida. Finalmente, pressionando o botão **OK** o resultado da função aparecerá na célula onde foi registrada a fórmula.

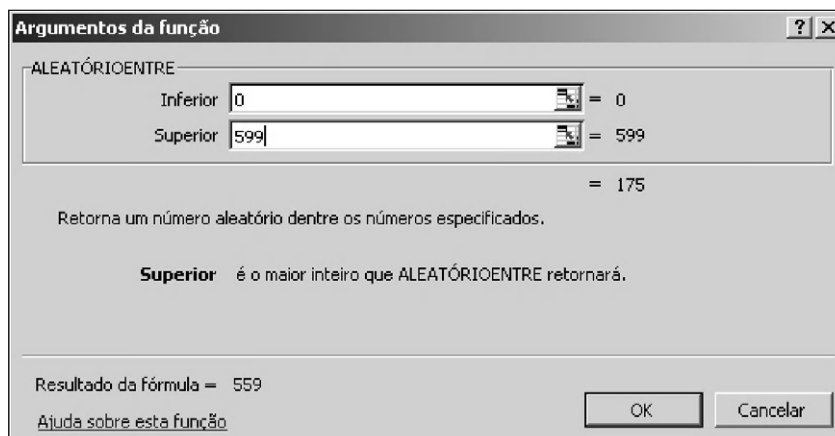


FIGURA 1.14 Caixa
de diálogo da função
ALEATÓRIOENTRE.

Apêndice 3

A função PROCV

Numa tabela com várias colunas, 1, 2, 3, ..., n , a função PROCV primeiro localizará um valor determinado na primeira coluna da esquerda da tabela e, depois, selecionará e retornará um valor registrado na mesma linha de uma coluna especificada à direita da primeira coluna da tabela. A sintaxe dessa função é:

`PROCV(procura;tabela;coluna;tipo_de_procura)`

Analisemos os quatro argumentos da função:

- No argumento *procura* deve ser informado o valor a ser localizado na primeira coluna do argumento *tabela*. Este argumento pode ser um valor numérico, uma referência ou uma sequência de caracteres de texto.
 - Se o valor registrado no argumento *procura* for menor do que o menor valor registrado na primeira coluna da *tabela*, a função PROCV retornará o valor de erro #N/D.
- No argumento *tabela* deve ser informada o intervalo de células da *tabela*, recomendando-se utilizar um nome de intervalo. Os valores na primeira coluna de *tabela* podem ser texto, números ou valores lógicos. Textos em maiúsculas e minúsculas são equivalentes.
 - Se o argumento *tipo_de_procura* for VERDADEIRO, os valores na primeira coluna de *tabela* deverão ser registrados em ordem ascendente, pois do contrário, a função PROCV poderá não retornar o valor correto:
 - Sendo valores numéricos, na ordem: -2, -1, 0, 1, 2, ... ,
 - Sendo caracteres de texto na, ordem de A a Z.
 - Sendo valores lógicos, ordem: FALSO, VERDADEIRO.
 - Se *tipo_de_procura* for FALSO, não será necessário ordenar a *tabela*.
- O argumento *coluna* é o número da coluna da *tabela* onde será selecionado e retornado o valor procurado, sendo a primeira coluna da *tabela* a número um.
 - Se *coluna*=1, a função PROCV retornará o valor na primeira coluna da *tabela*.
 - Se *coluna*=2, a função retornará o valor na segunda coluna da *tabela*.
 - Se *coluna* for menor do que 1, PROCV retornará o valor de erro #VALOR!; e se *coluna* for maior do que o número de colunas da *tabela* a função PROCV retornará o valor de erro #REF!.
- O argumento *tipo_de_procura* é um dos dois valores lógicos, FALSO ou VERDADEIRO, e especifica o tipo de correspondência, exata ou aproximada.
 - Se o argumento *tipo_de_procura* for VERDADEIRO ou omitido, a função PROCV retornará uma correspondência aproximada. De outra maneira, se não for encontrada uma correspondência exata, a função selecionará o menor valor mais próximo do valor informado no argumento *procura*.
 - Se o argumento *tipo_de_procura* for FALSO, a função PROCV procurará uma correspondência exata. Se nenhuma correspondência for encontrada, a função PROCV retornará o valor de erro #N/D. Se a função PROCV não localizar o valor registrado no argumento *procura* e *tipo_de_procura* for FALSO, a função retornará o valor #N/D.

A Função PROCH

O Excel dispõe também da função PROCH equivalente à função apresentada, porém procurando valores localizados em linhas. Numa tabela com várias linhas, 1, 2, 3, ..., n , a função PROCH primeiro localizará um valor determinado na primeira linha superior da tabela e, depois, selecionará e retornará um valor registrado na mesma coluna de uma linha especificada mais abaixo da primeira linha da tabela. A sintaxe desta função é:

PROCH(*procura;tabela;linha;tipo_de_procura*)

O significado dos argumentos é equivalente ao da função PROCV, porém operando com linhas.

Apêndice 4

Outro modelo para amostragem sem reposição

Na planilha **Apêndice 4** incluída na pasta **Capítulo 1** foi construído um procedimento de amostragem sem reposição que se pode aplicar a planilhas que contenham séries de dados de onde se deve extrair uma amostra sem reposição. Proceda como segue, Figura 1.15:

- Nas colunas B, C e D foram repetidos os dados já utilizados e referentes *as 50 primeiras empresas privadas*. Serão extraídas amostras sem reposição das vendas do intervalo D4:D53.
- Na célula F4 foi registrada a fórmula =ALEATÓRIO() que depois foi copiada até a célula F53. Perceba que o número de células com a fórmula =ALEATÓRIO() é o mesmo que o da população D4:D53. Lembre-se também de que cada vez que for recalculada a planilha será gerada uma nova série de números aleatórios.

	A	B	C	D	E	F	G	H	I
1	As 50 primeiras empresas privadas por vendas em 2002 - Revista Exame								
2									
3		Ordem	Empresa	Vendas		N.Aleat.		Amostra selecionada	
4		1	TELEMA	\$ 6.303,7		0,6860785		TEXACO - Atacado e comércio exterior	\$ 2.805,2
5		2	TELEFÔ	\$ 5.480,5		0,1880333		COPERSUCAR - Atacado e comércio exterior	\$ 1.550,5
6		3	CBB/AME	\$ 5.329,8		0,8661437		VOLKSWAGEN - Automotivo	\$ 5.295,2
7		4	VOLKSW	\$ 5.295,2		0,0290058		PONTO FRIO - Comércio varejista	\$ 1.153,3
8		5	PETRÓL	\$ 4.214,1		0,0773903		BONAE - Comércio varejista	\$ 1.156,5
9		6	SHELL -	\$ 4.096,8		0,261369		TELESP CELULAR - Telecomunicações	\$ 1.752,1
10		7	GENERA	\$ 4.092,7		0,1166721		COSIPA - Siderurgia e metalurgia	\$ 1.340,0
11		8	CARREF	\$ 4.044,9		0,0482277		KLABIN PAPEL CELULOSE - Papel e celulose	\$ 1.155,1
12		9	BRASIL	\$ 3.975,9		0,8553086		SHELL - Atacado e comércio exterior	\$ 4.096,8
13		10	GRUPO	\$ 3.837,5		0,19673		CPFL - Serviços públicos	\$ 1.551,2
14		11	EMBRAT	\$ 3.668,3		0,2965299			

FIGURA 1.15

Amostragem sem reposição, utilizando a função ALEATÓRIO.

- A fórmula =ÍNDICE(\$D\$4:\$D\$53;ORDEM(F4:\$F\$4:\$F\$53)) foi registrada na célula I4 e depois copiada até completar o tamanho da amostra, neste caso dez, célula I13. Essa fórmula utiliza a função ÍNDICE, que já foi apresentada neste capítulo, e a função ORDEM.

- **ORDEM(valor; amostra; ordem)**

A função estatística ORDEM²⁸ retorna a posição do argumento *valor* da *amostra* considerando a *ordem* informada:

- Se *ordem* for igual a 0 ou omitida, os valores da amostra serão classificados em ordem decrescente.
- Se *ordem* for diferente de 0, igual a 1, os valores da amostra serão classificados em ordem crescente.
- Se o argumento *amostra* tiver valores repetidos a função ORDEM informará a posição do primeiro valor que encontrar na sua procura, considerando o ordenamento escolhido.
Por exemplo, o objetivo da fórmula ORDEM(F4;\$F\$4:\$F\$53) é definir a posição do número aleatório da célula F4 dentro do intervalo F4:F53, a posição 17, um resultado do grupo de 1 a 50. Depois, a função ÍNDICE selecionará as vendas da empresa localizada na posição 5, neste caso, a empresa TEXACO.
- A fórmula =ÍNDICE(\$C\$4:\$C\$53;ORDEM(I4;\$D\$4:\$D\$53)) foi registrada na célula H4 e depois copiada até completar o tamanho da amostra, neste caso dez, célula H13. Então, deve ser utilizada a função ORDEM porque a função ÍNDICE reconhece somente valores numéricos e não títulos. Ademais, amarrar a fórmula com a resposta da célula F4 garante que se trata da mesma seleção, como foi mostrado no Exemplo 1.8.

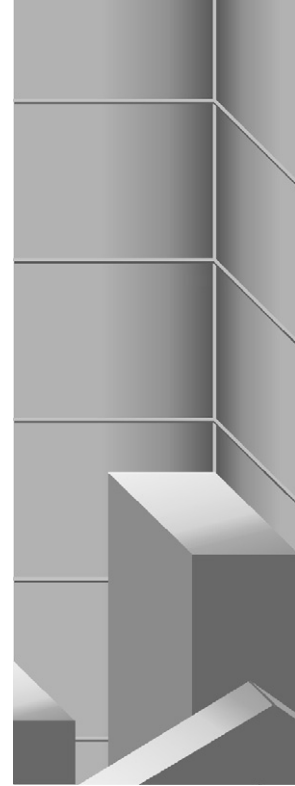
Entende-se que a função ALEATÓRIO gerará a quantidade de números aleatórios diferentes que for necessária,²⁹ que neste exemplo são 50 números aleatórios diferentes, premissa que não foi totalmente confirmada pelo autor.

²⁸ Em inglês, a função ORDEM é RANK.

²⁹ O procedimento apresentado foi baseado na informação registrada no site <http://www.staff.city.ac.uk/r.j.gerrard/excelfaq/faq.html#sample>. Nesse endereço há outras informações úteis navegando-se através de seus links.

Capítulo 2

DESCRIÇÃO DE AMOSTRAS COM TABELAS E GRÁFICOS



A obtenção de informação faz parte da gestão dos negócios. Por exemplo, o gerente de produção está interessado em monitorar continuamente a qualidade do produto produzido, comparando-o com os padrões estabelecidos; o gerente de produtos está interessado em conhecer a aceitação de um novo produto distribuindo amostras grátis e registrando os retornos dos consumidores etc. Para tentar conhecer uma ou mais características dessa população, é extraída uma amostra de uma população, conforme orientado no Capítulo 1. Quando o tamanho da amostra é grande, maior do que 15 a 20 observações, a simples inspeção das observações não será suficiente para obter as conclusões desejadas. Esses dados coletados devem ser organizados ou resumidos com o objetivo de facilitar a análise e a interpretação das observações. Neste capítulo, você aprenderá a agrupar os dados em tabelas de frequências e histogramas, procedimentos que fazem parte da *Estatística Descritiva*.

EXEMPLO 2.1

O gerente do departamento de uma instituição financeira deseja analisar o número diário de operações fechadas nos últimos dois anos por um operador de seu departamento de *opções* de ações negociadas na Bolsa de Valores. Na tabela a seguir foi registrada uma amostra probabilística simples de tamanho 26, extraída das operações diárias fechadas pelo Operador B nos últimos dois anos. O objetivo é obter as possíveis conclusões dos registros dessa tabela.

14	12	13	11	12	13	16	14	14	15	17	14	11
13	14	15	13	12	14	13	14	13	15	16	12	12

Solução. Aplicando inicialmente apenas o bom senso, pode-se constatar que:

- O número de operações fechadas por dia é um número do conjunto {11, 12, 13, 14, 15, 16, 17}.
- O Operador B fechou entre 11 e 17 operações por dia.
- O número diário máximo de operações fechadas pelo Operador B é 17, e o número mínimo é 11.
- O *intervalo* ou *range* das operações fechadas por dia é seis, valor obtido como resultado da subtração $17 - 11 = 6$. Embora o intervalo mostre que o número de negócios fechados por dia é variável, esse mesmo valor não consegue mostrar nada sobre a frequência do número diário de negócios. Se o número diário de operações fechadas fosse constante, não seria necessário aplicar conceitos estatísticos para obter respostas. Entretanto, como os valores da variável não são constantes, o primeiro passo é pesquisar a origem das variações.

Embora tenham sido obtidas algumas conclusões, o simples ordenamento dos dados não permite obter maiores conclusões, pois ainda nos deparamos com a mesma quantidade de dados. Precisamos agrupar os dados de alguma maneira, tendo em mente que esse procedimento não deve interferir na obtenção de conclusões. Uma forma prática e eficiente é agrupar os dados de acordo com suas frequências de repetição, cujo procedimento dá origem às tabelas de frequências ou distribuições de frequências.

Tabelas de frequências de dados quantitativos discretos

Iniciamos este tema com a construção de tabelas de frequências de uma amostra com dados quantitativos discretos que, em geral, medem contagens representadas por números inteiros positivos 0, 1, 2, 3, ..., n , por exemplo, o número de pessoas atendidas em um determinado período, o número de transações financeiras realizadas pela Internet em um determinado banco, a quantidade de peças defeituosas em um lote de produção etc. Depois será tratada a construção de tabelas de frequências de uma amostra com dados quantitativos contínuos que podem assumir qualquer valor do conjunto dos números reais, por exemplo, o peso dos alunos da quarta série dos alunos da rede escolar de uma determinada região, as vendas diárias de uma empresa, o consumo mensal de energia elétrica, a rentabilidade diária das ações mais negociadas na Bolsa de Valores etc. Embora a classificação dos dados quantitativos pareça fácil, a separação entre discretas e contínuas nem sempre é clara.

Tabela de frequências absolutas

Se as observações da amostra do número diário de operações fechadas do Exemplo 2.1 forem agrupadas considerando as repetições de cada observação, poderemos obter mais informações dessa amostra.¹

A *frequência* do valor de uma variável é o número de repetições desse valor.

A *tabela de frequências absolutas* de uma variável é uma função formada pelos valores da variável e suas respectivas frequências; conhecida também como *distribuição de frequências absolutas*.

O par formado por cada valor da variável e sua frequência correspondente determina a *tabela de frequências absolutas* da variável ou *distribuição de frequências absolutas*.

EXEMPLO 2.2

Continuando com o Exemplo 2.1. Construa a tabela de frequências absolutas do número de operações fechadas por dia pelo operador B.

Solução. Para realizar a classificação de forma manual, não é necessário, previamente, ordenar os valores da variável de forma crescente. Na primeira coluna da tabela a seguir, foram registrados os valores do número de operações fechadas por dia e em ordem crescente: 11, 12, 13, 14, 15, 16 e 17. Na segunda coluna, foi realizada a seleção manual da ocorrência de cada um dos valores da primeira coluna da tabela. Por exemplo, o primeiro número 14 da amostra foi registrado com a marca I na linha 14 da segunda coluna da tabela, o segundo número 12 foi registrado com a marca I na linha 12 da segunda coluna da tabela, e assim sucessivamente até o último valor 12 da amostra. Para facilitar a contagem, o quinto valor selecionado de cada valor é representado por uma linha transversal definindo um grupo de cinco seleções do mesmo número. Uma nova seleção do mesmo valor inicia um novo grupo, como se pode ver nas linhas dos valores 13 e 14. Para completar a tabela, na última linha da última coluna, é registrada a soma das frequências absolutas cujo resultado 26 deve ser igual ao número de observações da amostra, também 26.

¹ A variável pode pertencer a uma amostra ou uma população.

Operações fechadas por dia	Seleção	Frequências absolutas
11	II	2
12	IIII	5
13	IIII I	6
14	IIII II	7
15	III	3
16	II	2
17	I	1
	Total	26

Essa tabela de frequências absolutas foi construída na planilha **Tabelas de Frequências**, incluída na pasta **Capítulo 2**.

EXEMPLO 2.3

Analisar os resultados da tabela de frequências absolutas do Exemplo 2.2.

Solução. Da tabela de frequências absolutas do Exemplo 2.2 podemos chegar às seguintes conclusões:

- O número máximo 17 de operações diárias fechadas pelo Operador B aconteceu em apenas um dia da amostragem.
- Entretanto, o valor mínimo 11 repetiu-se em dois dias.
- Em seis dias da amostragem, o Operador B fechou 13 operações por dia, e, em sete dias da amostragem, fechou 14 operações por dia.
- Os valores das frequências de cada observação mostram um contorno crescente da observação 11 até a 14 e decrescente desde esse valor até o 17.

A tabela de frequências absolutas resume uma série de valores numéricos em uma simples classificação de frequências muito útil para descrever características importantes do conjunto de dados da amostra. As duas tabelas de frequências seguintes possibilitarão incluir outras características não mostradas pela primeira tabela.

Tabela de frequências relativas

A tabela de frequências do Exemplo 2.2 agrupa valores absolutos que permitem chegarmos a conclusões como, em cinco dias da amostra, o Operador B fechou 12 operações. Esse tipo de resultado não permite avaliar, por exemplo, se essa frequência doze é alta ou baixa, pois nesse resultado não há nenhuma informação sobre o tamanho da amostra. Conseguiremos extrair mais informação da variável se suas frequências forem expressas como porcentagem do tamanho da amostra.

A *frequência relativa* do valor de uma variável é o resultado de dividir sua frequência absoluta pelo tamanho da amostra.

A *tabela de frequências relativas* de uma variável é uma função formada pelos valores da variável e suas respectivas frequências relativas; conhecida como *distribuição de frequências relativas*.

O par formado por cada valor da variável e sua frequência relativa correspondente determina a *tabela de frequências relativas* da variável ou *distribuição de frequências relativas*, em valores unitários ou percentagem.

EXEMPLO 2.4

Continuando com o Exemplo 2.1. Primeiro construa a tabela de frequências relativas da variável número de operações fechadas por dia pelo operador B e, depois, analise os resultados.

Solução. As duas primeiras colunas da tabela seguinte repetem a tabela das frequências absolutas construída no Exemplo 2.2. Na terceira coluna, foi registrado o resultado da divisão do valor de cada frequência absoluta por 26, o tamanho da amostra. Para completar a tabela, foi adicionada uma linha onde foi registrado o total de cada coluna de frequência. Os resultados dessa última linha devem ser iguais ao número de observações da amostra, 26, na coluna de frequências absolutas, e 100%, na coluna de frequências relativas, pois o resultado 100% indica que todas as observações da amostra estão contidas nessas frequências.

Operações fechadas por dia	Frequências absolutas	Frequências relativas %
11	2	7,69%
12	5	19,23%
13	6	23,08%
14	7	26,92%
15	3	11,54%
16	2	7,69%
17	1	3,85%
Total	26	100,00%

Essa tabela de frequências absolutas foi construída a partir da linha 14 da planilha **Tabelas de Frequências**, incluída na pasta **Capítulo 2**. Da tabela de frequências relativas, chegamos a estas conclusões:

- Em 3,85% dos 26 dias amostrados, o Operador B fechou 17 negócios por dia.
- Em 7,69% dos dias amostrados, o Operador B fechou 11 negócios por dia.
- Durante 26,92% dos dias da amostra, o Operador B fechou 14 negócios.

Um ponto importante que precisa ser ressaltado é que analisando o procedimento do Exemplo 2.4, observamos que a construção da tabela de frequências relativas é realizada com os dados registrados na tabela de frequências absolutas. No sentido inverso, a construção da tabela de frequências absolutas poderá ser realizada com os dados registrados na tabela de frequências relativas se for conhecido o tamanho da amostra.

Tabela de frequências acumuladas

As distribuições de frequências absolutas e relativas apresentadas são muito úteis para organizar e resumir os dados das observações em forma de tabela, permitindo detectar as características relevantes dos valores da variável amostrada. Em alguns casos, o interesse da análise reside em conhecer os valores da variável menores ou maiores a um determinado valor, por exemplo, o número de dias em que o Operador B fechou menos do que 15 operações por dia etc.

A *frequência acumulada* do valor de uma variável é a soma das frequências absolutas ou relativas desde o valor inicial da variável.

A *tabela de frequências acumuladas* ou *distribuição de frequências acumuladas* de uma variável é uma função formada pelos valores da variável e suas respectivas frequências acumuladas.

Por exemplo, se conhecermos a distribuição das peças rejeitadas por lote de produção, poderemos conhecer o número de lotes que tiveram uma rejeição maior ou menor do que um determinado número de peças. Essa informação pode ser obtida da distribuição de frequências acumuladas, ou *ogiva*,² formada pela acumulação dos valores absolutos ou relativos da distribuição inicial.

EXEMPLO 2.5

Continuando com o Exemplo 2.1. Construa a tabela de frequências acumuladas da variável número de operações fechadas por dia pelo operador B.

Solução. Na primeira coluna da tabela seguinte, foram registrados os valores do número de operações fechadas por dia e em ordem crescente: 11, 12, 13, 14, 15, 16 e 17. Para cada valor da variável:

- Na segunda coluna, foram acumuladas as frequências absolutas do Exemplo 2.2 desta forma:
 - A frequência acumulada absoluta até 12 negócios fechados por dia é igual a $7=2+5$.
 - A frequência acumulada absoluta até 13 negócios fechados por dia é igual a $13=2+5+6$. Repetindo esse procedimento até a última linha da tabela, completamos a distribuição de frequências acumuladas absolutas.
 - A frequência acumulada absoluta da última linha deverá sempre ser igual ao tamanho da amostra, nesse caso, 26.
- Na terceira coluna, foram acumuladas as frequências relativas do Exemplo 2.4 desta forma:
 - A frequência acumulada relativa até 12 negócios fechados por dia é igual a $26,92\%=7,69\%+19,23\%$.
 - A frequência acumulada relativa até 13 negócios fechados por dia é igual a $50\%=7,69\%+19,23\%+23,08\%$. Repetindo esse procedimento até a última linha da tabela, completamos a distribuição de frequências acumuladas relativas.
 - A frequência acumulada absoluta da última linha deverá sempre ser igual a 100%, pois o resultado 100% indica que todas as observações da amostra estão contidas nessas frequências

Operações fechadas por dia	Frequências acumuladas	
	Absolutas	Relativas %
11	2	7,69%
12	7	26,92%
13	13	50,00%
14	20	76,92%
15	23	88,46%
16	25	96,15%
17	26	100,00%

Essa tabela de frequências absolutas foi construída a partir da linha 25 da planilha **Tabelas de Frequências** incluída na pasta **Capítulo 2**.

² Como a distribuição de frequências acumuladas sempre é crescente, quando a distribuição é representada com uma poligonal, o desenho se assemelha à ogiva de um foguete.

Das tabelas de frequências acumuladas absolutas e relativas do Exemplo 2.5, temos as seguintes conclusões:

- Ao afirmar que o operador B fechou 14 ou menos operações por dia em 76,92% dos dias da amostra, foi incluído nessa afirmativa o fechamento de 14 operações por dia. Diferente das seguintes declarações:
 - O operador B fechou menos de 14 operações por dia em 50% dos dias da amostra; o fechamento de 14 operações não está incluído.
 - O operador B fechou menos de 15 operações por dia em 76,92% dos dias da amostra; o fechamento de 15 operações por dia não está incluído.
- Ao afirmar que em 23,08% dos dias o operador B fechou 15 ou mais operações por dia, está incluído nesse resultado o fechamento de 15 operações por dia. Verifique que esse último resultado (23,08%) é o complemento do operador ter fechado menos de que 15 operações por dia (76,92%), pois o resultado da soma desses dois valores é 100%.
- Ao afirmar que em 61,54% dos dias o operador B fechou entre 13 e 15 operações, incluindo esses valores, estamos realizando os seguintes cálculos:
 - Em 88,46% dos dias, o operador B fechou 15 ou menos operações.
 - Em 26,92%, fechou 12 ou menos operações, ou fechou menos de 13 operações.
 - Portanto, em $61,54\% = 88,46\% - 26,92\%$ dos dias o operador B fechou entre 13 e 15 operações, incluindo esses valores.

Outro ponto importante a ser destacado é que, analisando o procedimento do Exemplo 2.5, observamos que:

- A construção da tabela de frequências acumuladas absolutas é realizada com os dados registrados na tabela de frequências absolutas. No sentido inverso, a construção da tabela de frequências absolutas poderá ser realizada com os dados registrados na tabela de frequências acumuladas absolutas. E da mesma maneira para as frequências relativas.
- A construção da tabela de frequências acumuladas relativas pode ser realizada com os dados registrados na tabela de frequências acumuladas absolutas se for conhecido o tamanho da amostra. No sentido inverso, a tabela de frequências acumuladas absolutas poderá ser construída com os dados registrados na tabela de frequências acumuladas relativas se for conhecido o tamanho da amostra.

A função Frequência do Excel

O Excel dispõe de muitas funções estatísticas que reduzem o tempo de cálculo e asseguram resultados exatos. O Apêndice 1 registra algumas dessas funções relacionadas com a determinação do valor máximo, do valor mínimo e a contagem de observações de uma amostra. A função estatística FREQUÊNCIA do Excel é de grande ajuda na construção das tabelas de frequências de uma amostra.

FREQUÊNCIA(*matriz_dados*; *matriz_bin*)

A função estatística FREQUÊNCIA³ retorna uma matriz vertical contendo a distribuição de frequências da amostra definida no argumento *matriz_dados* de acordo com a seleção registrada no argumento *matriz_bin*. Portanto:

- No argumento *matriz_dados*, deve ser informado o intervalo da planilha em que foram registradas as observações da amostra.

³ Em inglês, a função FREQUÊNCIA é FREQUENCY.

- No argumento *matriz_bin*, deve ser informado o intervalo da planilha dos valores definidos pelo usuário para selecionar, ou agrupar, as observações da amostra.
- Ao realizar a seleção dos valores da variável, a função FREQUÊNCIA não considera as células vazias ou com texto.

Um detalhe importante: se o nome da função FREQUÊNCIA for inserido com letras minúsculas ou maiúsculas ou sem os acentos ortográficos, felizmente, o Excel aceitará e registrará a função com letras maiúsculas e com os acentos ortográficos.

Com essa função, é possível construir a tabela de frequências absolutas e acumuladas absolutas, de acordo com a forma de registrar essa função:

- Se for registrada como *matriz_coluna*, a função FREQUÊNCIA retornará a tabela da distribuição de frequências absolutas, apresentada como *matriz_coluna*.
- Se for registrada como fórmula, a função FREQUÊNCIA retornará a tabela de frequências acumuladas absolutas.

A descrição da função FREQUÊNCIA mostra que há duas formas de registrá-la na planilha Excel, obtendo, nos dois casos, resultados estatísticos diferentes, ambos importantes e de nosso interesse. Para compreender como deve ser utilizada, será novamente resolvido o Exemplo 2.1 utilizando a função FREQUÊNCIA, repetindo o enunciado.

EXEMPLO 2.6

O gerente do departamento de uma instituição financeira quer analisar o número diário de operações fechadas nos últimos dois anos por um operador de seu departamento de *opções* de ações negociadas na Bolsa de Valores. Na tabela a seguir, foi registrada uma amostra probabilística simples de tamanho 26 e extraída das operações diárias fechadas pelo Operador B nos últimos dois anos. Construa tabela de frequências absolutas do número de operações fechadas por dia pelo operador B utilizando a função FREQUÊNCIA do Excel.

14	12	13	11	12	13	16	14	14	15	17	14	11
13	14	15	13	12	14	13	14	13	15	16	12	12

Solução. A amostra do número de operações fechadas por dia foi registrada no intervalo B4:B29 da planilha **Função Frequência** incluída na pasta **Capítulo 2**. Para a construção da tabela de frequências absolutas, serão utilizados os valores do número de operações fechadas por dia em ordem crescente: 11, 12, 13, 14, 15, 16 e 17; esses valores foram registrados no intervalo D4:D10. Na descrição, foi visto que função FREQUÊNCIA retornará a tabela da distribuição de frequências absolutas apresentada como *matriz_coluna*. Para trabalhar com registros em forma de *matriz*, devemos proceder desta forma:

- Posicionar o mouse na célula E4 e selecionar o intervalo E4:E11. Observe que o intervalo selecionado contém uma linha a mais do que o intervalo em que estão registrados os valores do argumento a *matriz_bin*, intervalo D4:D10.

	A	B	C	D	E	F
1	Função Frequência					
2						
3		Dados		Tabela de Frequências Absolutas		
4		14			=frequência(B4:B29;D4:D10)	
5		12		12		
6		13		13		
7		11		14		
8		12		15		
9		13		16		
10		16		17		
11		14				

- A seguir, digite a fórmula =frequência(B4:B29;D4:D10) sem pressionar a tecla **Enter**, como mostra a figura anterior. Note que o nome da função foi inserido com letras minúsculas e sem os acentos ortográficos, pois felizmente o Excel aceitará e registrará a função com letras maiúsculas e com os acentos ortográficos. Em vez de digitar a fórmula, você pode utilizar o assistente do Excel *Colar função* apresentado no Apêndice 2 do Capítulo 1, que possui mais vantagens em comparação à digitação direta na célula.
- Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**. Mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, obtemos os resultados apresentados na próxima figura, na qual as fórmulas receberam as chaves { }. Você pode usar esse procedimento se utilizar o assistente do Excel *Colar função*.

	A	B	C	D	E	F	G	H
1	Função Frequência							
2								
3		Dados		Tabela de Frequências Absolutas				
4		14		11	2	{=FREQUÊNCIA(B4:B29;D4:D10)}		
5		12		12	5			
6		13		13	6			
7		11		14	7			
8		12		15	3			
9		13		16	2			
10		16		17	1			
11		14			0	{=FREQUÊNCIA(B4:B29;D4:D10)}		
12		14						

Podemos notar que as fórmulas do intervalo E4:E11 são todas iguais a {=FREQUÊNCIA(B4:B29;D4:D10)}, sendo que as chaves { } indicam que as fórmulas fazem parte da mesma *matriz*. Por último, o valor zero na célula E11 informa que nenhum dos valores da variável deixou de ser classificado. De outra maneira, o objetivo da última célula E11 é informar quantos valores da variável não foram classificados.⁴ Como exercício, verifique que a partir das frequências absolutas é possível construir a tabela de frequências acumuladas absolutas da mesma amostra, como foi realizado no intervalo I4:I10 da planilha **Função Frequência** incluída na pasta **Capítulo 2** desta forma:

- Na célula I4 foi registrada a fórmula =E4, pois ambas as frequências têm o mesmo valor.
- Na célula I5 foi registrada a fórmula =I4+E5, que depois foi copiada até a célula I10.

Compare os resultados.

Utilizando a função FREQUÊNCIA como matriz coluna, obtemos a tabela de frequências absolutas da série de dados, adicionando a vantagem de controlar a quantidade de dados que não foram classificados.

EXEMPLO 2.7

Continuando com o Exemplo 2.6. Construa a tabela de frequências acumuladas absolutas do número de operações fechadas por dia pelo operador B utilizando a função FREQUÊNCIA do Excel.

Solução. A partir da linha 13 da planilha **Função Frequência** incluída na pasta **Capítulo 2**, foi construída a tabela de frequências acumuladas absolutas da amostra registrada no intervalo B4:B29. Se a função FREQUÊNCIA for registrada como fórmula única, a função dará como resultado a frequência acumulada dos valores iguais ou menores do que o valor informado no argumento matriz_bin. Como exemplo, se numa célula vazia da planilha referida for registrada a fórmula =FREQUÊNCIA(B4:B29;12), a função retornará o valor 7, a frequência do número de negócios fechados iguais ou menores a doze.

⁴ Sugerimos que você procure se informar sobre o uso das *matrizes* na *ajuda on-line* do Excel, incluindo as rotinas para modificação das fórmulas.

Para obter a tabela de frequências acumuladas absolutas da amostra registrada no intervalo B4:B29, faça o seguinte:

- Registre os valores do número de operações fechadas por dia em ordem crescente: 11, 12, 13, 14, 15, 16 e 17 no intervalo D15:D21.
- Na célula E15, registre a fórmula =FREQUÊNCIA(\$B\$4:\$B\$29;D15) que deverá ser copiada até a célula D21. Sobre os cifrões registrados nos endereços do intervalo B4:B29, veja o Apêndice 2 deste capítulo.
- Depois de pressionar **Enter**, a função retornará o valor 2. A seguir, copie essa fórmula até a célula D21.

	A	B	C	D	E
12		14			
13		15			
14		17			
15		14		11	2
16		11		12	7
17		13		13	13
18		14		14	20
19		15		15	23
20		13		16	25
21		12		17	26
22		14			

A figura mostra a tabela de frequências acumuladas absolutas construída com a função FREQUÊNCIA construída a partir da linha 13 da planilha **Função Frequência** incluída na pasta **Capítulo 2**. Como exercício, verifique que a partir das frequências acumuladas absolutas é possível construir a tabela de frequências absolutas da mesma amostra, como foi realizado no intervalo F15:F21 da planilha **Função Frequência** incluída na pasta **Capítulo 2**, procedendo desta forma:

- Na célula F15, foi registrada a fórmula =E15, pois ambas as frequências têm o mesmo valor.
- Na célula F16, foi registrada a fórmula =E16 – E15 e, depois, foi copiada até a célula F11.

Compare os resultados.

Construção das tabelas de frequências numa planilha Excel

Na planilha **Resultados de Frequências** incluída na pasta **Capítulo 2**, foram construídas as tabelas de frequências. No intervalo B5:B30, foi registrada a amostra do Exemplo 2.1, com os dados ordenados de forma crescente. O ordenamento crescente dos dados foi realizado apenas para visualizar o procedimento de cálculo da função FREQUÊNCIA quando registrada em uma única célula.

A partir das frequências acumuladas absolutas, é possível, também, construir as tabelas de frequências absolutas, relativas e acumuladas relativas da amostra como foi realizado na planilha **Resultados de Frequências**, Figura 2.1:

- No intervalo D5:D11, foram registrados os valores do número de operações fechadas por dia em ordem crescente: 11, 12, 13, 14, 15, 16 e 17.
- Na célula E5, foi registrada =FREQUÊNCIA(\$B\$5:\$B\$30;D5) e, depois, foi copiada até a célula E11. Como resultado, no intervalo E5:E11 estão registradas as frequências acumuladas procuradas.
- As frequências absolutas são registradas no intervalo F5:F11 a partir das frequências acumuladas absolutas registradas no intervalo E5:E11.
 - Na célula F5, foi registrada a fórmula =E5, pois ambas as frequências têm o mesmo valor.
 - Na célula F6, foi registrada a fórmula =E6-E5 e depois foi copiada até a célula F11.
- As frequências relativas são registradas no intervalo G5:G11 a partir das frequências absolutas registradas no intervalo F5:F11 e da contagem de valores do intervalo B5:B30. Na célula G5, foi registrada a fórmula =F5/CONT.NÚM(\$B\$5:\$B\$30) e copiada até a célula G11.

- As frequências acumuladas relativas são registradas no intervalo H5:H11 a partir das frequências relativas registradas no intervalo G5:G11.
- Na célula H5, foi registrada a fórmula =G5, pois as ambas frequências têm o mesmo valor.
- Na célula H6, foi registrada a fórmula =H5+G6, que depois foi copiada até a célula H11.

No intervalo D13:F15 da planilha, foi construído um modelo que, na célula F15, retorna, a partir do valor observado registrado na célula D14, o resultado da frequência selecionada na célula E14. Por exemplo, registrando 15 na célula D14, obteremos o valor 11,54% se na célula E14 for selecionado **Relativa**, uma das quatro frequências possíveis de selecionar, como mostra a Figura 2.1.

FIGURA 2.1

Construção de tabelas de frequências.

	A	B	C	D	E	F	G	H
1	Resultados de Frequências							
2								
3		Dados		Tabelas de Frequências				
4		Ordenados		Seleção	Acumul. Abs.	Absolutas	Relativas	Acumul. Rel
5		11		11	2	2	7,69%	7,69%
6		11		12	7	5	19,23%	26,92%
7		12		13	13	6	23,08%	50,00%
8		12		14	20	7	26,92%	76,92%
9		12		15	23	3	11,54%	88,46%
10		12		16	25	2	7,69%	96,15%
11		12		17	26	1	3,85%	100,00%
12		13						
13		13						
14		13		Dado	Tipo Frequência			
15		13		15	Relativa			
16		13		Freq. Relativa <= 15		11,54%		

As colunas de frequências construídas na planilha Excel estão em uma ordem diferente da utilizada durante sua apresentação. Depois de construídas, as colunas das frequências podem ser permutadas na ordem desejada. Também, a construção das tabelas com Excel poderia começar pela construção inicial da tabela de frequências absolutas, tarefa que deixamos para você, lembrando que a planilha **Resultados de Frequências** foi protegida, sem senha, exceto nas células D14 e E14.

Tabelas de frequências de dados quantitativos contínuos

A construção das tabelas de frequências do Exemplo 2.1 foi relativamente fácil, pois os dados da variável são quantitativos e discretos, que resultam de contagens, com uma quantidade pequena de observações e a maior parte delas repetidas. Entretanto, se os dados da variável forem contínuos, que resultam de medições que podem ter grande precisão, a aplicação do procedimento anterior será trabalhosa e de baixa eficiência, pois poucos ou até nenhum dos dados poderão apresentar frequência. Nesse caso, o procedimento recomendado para variáveis com valores contínuos é trabalhar com *classes* de valores. O método começa pela definição da quantidade, dos limites e da amplitude das classes onde serão selecionados os valores da variável.

Na construção da tabela de frequências, leve em consideração que:

- Não há uma regra exata para determinar o número de classes, apenas orientações práticas para o analista. Por exemplo, para uma amostra de tamanho n , a quantidade de classes k recomendada pode ser obtida de:
 - $k = \sqrt{n}$, arredondando o resultado para o valor inteiro menor ou maior.
 - $k = 1 + 3,322 \times \log(n)$, arredondando o resultado para o valor inteiro menor ou maior.

- O número de classes é o menor valor inteiro k , que satisfaz à condição $2^k \leq n$. Na realidade, essa fórmula é igual à fórmula anterior na condição $2^k = n$.
- A determinação da quantidade de classes tem um pouco do procedimento de tentativa e erro na procura da distribuição que melhor represente os valores da variável. A quantidade de classes para diversos valores do tamanho de amostra utilizando as três fórmulas é apresentada no intervalo B3:E24 da planilha **Quantidade de Classes**, incluída na pasta **Capítulo 2**. Informando o tamanho de amostra na célula B27, a planilha apresenta os resultados pelos três métodos no intervalo C27:E27.
- Ao trabalhar com classes, a tabela de frequências não retém a identidade de cada observação individual, provocando perda de informação. Os valores da variável são transformados em uma nova variável cujos novos valores são os limites dos intervalos das classes.

O exemplo a seguir mostra como proceder para construir tabelas de frequência absolutas utilizando classes.

EXEMPLO 2.8

As vendas diárias em milhares de uma empresa estão registradas na tabela a seguir. O objetivo é construir a tabela de frequências absolutas e relativas e as respectivas frequências acumuladas.

280	305	320	330	310	340	330	341	369	355	370	360	370
365	280	375	380	400	371	390	400	370	401	420	430	

Solução. O procedimento para construir a tabela de frequências absolutas utilizando classes é o seguinte:

Determinação da quantidade de classes

Como premissa inicial, é conveniente que todas as classes tenham a mesma largura, denominado também de intervalo ou amplitude da classe. A quantidade de classes deve ser fixada de forma que as classes representem adequadamente a distribuição de valores da variável sob estudo. Um número pequeno de classes gera amplitudes de classes grandes e vice-versa, podendo gerar distorções indesejáveis. Como vimos, não há uma regra única para escolher a quantidade de classes, apenas regras práticas que orientam o analista. Nesse caso, aplicando qualquer uma das três fórmulas apresentadas, o número de classes sugerido para uma amostra de tamanho 25 é igual a cinco, $k = 5$.

Determinação da amplitude das classes

Os valores máximo e mínimo da amostra são, respectivamente, 430 e 280, e o intervalo de variação é 150, resultado da diferença entre os valores máximo e mínimo da amostra $150 = 430 - 280$. A amplitude das cinco classes é igual a 30, valor obtido como resultado da divisão do intervalo de variação pela quantidade de classes, $\frac{430 - 280}{5} = 30$.

Preparação da tabela de seleção

Com os resultados anteriores, é construída a tabela de seleção com três colunas: a primeira, que identifica a classe, de um a cinco, as duas últimas, que registram o limite inferior e o limite superior das cinco classes, respectivamente.

Classe	Limite inferior	Limite superior
1	280	310
2	310	340
3	340	370
4	370	400
5	400	430

Analisemos os limites das classes dessa tabela:

- Da forma como foram registrados os limites, parece que o limite superior de uma classe é igual ao limite inferior da classe seguinte. O valor 310 da amostra deve ser classificado na primeira classe, com os limites 280-310, ou na segunda classe com os limites 310-340? O valor 310 deve ser classificado na classe 310-340, pois o limite superior de cada classe não inclui o próprio valor; o limite superior é aberto, com exceção da última classe.
- Como prática corrente, o limite inferior da primeira classe deve conter a observação de menor valor da amostra e o limite superior da última classe, o maior. Nada impede utilizar os valores mínimo e máximo da amostra, respectivamente, como limite inferior da primeira classe e o limite superior da última classe.

Seleção dos dados e construção das tabelas de frequências

A seleção dos valores da variável nas classes estabelecidas é executada da forma como foi realizada com os dados discretos, obtendo as seguintes distribuições de frequências absolutas e relativas cujos resultados foram obtidos na planilha **Exemplo 2.8**, incluída na pasta **Capítulo 2**.

Classe	Frequências absolutas	Frequências relativas	Frequências acum. abs.	Frequências acum. rel.
280-310	3	12,00%	3	12,00%
310-340	4	16,00%	7	28,00%
340-370	6	24,00%	13	52,00%
370-400	7	28,00%	20	80,00%
400-430	5	20,00%	25	100,00%
Total	25	100%		

É importante ressaltar que os dados do Exemplo 2.8 facilitaram a obtenção das classes, bem como seus limites, pois, em geral, a determinação da quantidade de classes e amplitudes é um processo de tentativa e erro, procurando o equilíbrio entre a quantidade e a amplitude das classes para conseguir a distribuição de frequências que melhor represente a amostra. Quando cada classe estiver formada por apenas um valor, por exemplo, a quantidade de operações fechadas do Exemplo 2.1, diz-se que não há perda de informação. Entretanto, no caso do Exemplo 2.8, há perda de informação, pois os valores das vendas diárias não são considerados individualmente; eles estão agrupados em classes. O Exemplo 2.9 mostra como utilizar a função FREQUÊNCIA para obter as tabelas de frequências.

EXEMPLO 2.9

Continuando com o Exemplo 2.8. O objetivo é construir a tabela de frequências absolutas e relativas e as respectivas frequências acumuladas utilizando a função FREQUÊNCIA do Excel.

Solução. Antes de utilizar a função FREQUÊNCIA com classes, devemos rever a forma de seleção dessa função. Se a função FREQUÊNCIA for registrada como fórmula única:

- A função retornará a frequência acumulada dos valores iguais ou menores ao valor informado no argumento *matriz_bin* da função, considerando o limite superior da classe como fechado. De outra maneira, o limite superior de cada classe inclui o próprio valor.
- Tecnicamente, o limite superior é aberto, com exceção da última classe. De outra maneira, o limite superior de cada classe não inclui o próprio valor.

	A	B	C	D	E	F	G	H
1	Exemplo 2.9							
2								
3		Amostra	Limites					
4		280	Tec. Inferior	Tec. Superior	Excel			
5		305	280	310	309,9			
6		320	310	340	339,9			
7		330	340	370	369,9			
8		310	370	400	399,9			
9		340	400	430	430			
10		330						
11		341	Limites					
12		369	Tabela de Frequências					
13		365	Excel	Absolutas	Acumul. Abs.	Relativas	Acumul. Rel.	
14		370	309,9	3	3	12,00%	12,00%	
15		360	339,9	4	7	16,00%	28,00%	
16		370	369,9	6	13	24,00%	52,00%	
17		365	399,9	7	20	28,00%	80,00%	
18		280	430	5	25	20,00%	100,00%	
19		375		0				

Para operar com o Excel mantendo o limite superior da classe aberto, o limite superior utilizado na função FREQUÊNCIA deverá ser menor do que o limite teórico. A diminuição do valor do limite superior dependerá dos valores dos dados, por exemplo, se todos os valores da amostra forem números inteiros, a diminuição de 0,1 será suficiente. Contudo, se alguns valores da amostra forem números com uma casa decimal, deverá ser utilizada uma diminuição de 0,01.

Na planilha **Exemplo 2.9**, incluída na pasta **Capítulo 2**, foram construídas as quatro tabelas de frequências do Exemplo 2.9 partindo da tabela de frequências absolutas. Na primeira coluna **Tec. Superior**, foram registrados os limites superiores de cada classe em ordem crescente a partir da primeira classe. Na coluna **Excel**, também foram listados os limites superiores de cada classe, porém ligeiramente menores do que seus equivalentes teóricos, subtraindo 0,10 de cada limite teórico, com exceção da última classe que permanece com o mesmo limite

Histograma

As quatro tabelas de frequências apresentadas resumem os valores de uma amostra, ajudando na sua análise e permitindo inferir sobre a população de onde foi extraída a amostra. O *Histograma* visualiza a tabela de frequências de uma amostra, ou variável, em um gráfico de barras verticais, aumentando a compreensão dos resultados e análises.

Histograma é o gráfico de barras verticais das frequências dos valores de uma amostra ou variável.

Vejamos algumas características gerais da construção dos histogramas. As barras verticais do histograma têm a mesma largura, e o comprimento ou altura das barras é proporcional à frequência de cada valor ou classe representada. Na forma do contorno do histograma, reconheceremos distribuições simétricas e não simétricas, e essa particularidade ajudará no processo de inferência que será realizado.

O histograma é construído a partir da tabela de frequências correspondente, que deverá ser previamente construída. Em vez de mostrar a construção manual do histograma, a seguir mostraremos como construir um histograma com o Excel. Depois será apresentada a ferramenta de análise *Histograma*, que constrói automaticamente o histograma e, ao mesmo tempo, pode apresentar outras respostas conforme a escolha prévia do leitor.

Construção do histograma com Excel


Começamos com a construção do histograma de frequências absolutas de uma amostra com dados quantitativos discretos utilizando a amostra do Exemplo 2.1, deixando para depois a construção do histograma de frequências de uma amostra com dados quantitativos contínuos.

FIGURA 2.2 Assistente de gráfico – etapa 1 de 4 – tipo de gráfico.



O primeiro passo é a preparação da planilha **Construção Histograma**, incluída na pasta **Capítulo 2**, contendo a amostra e a tabela de frequências absolutas calculadas na mesma planilha. Para facilitar a preparação dessa planilha, pode-se economizar tempo copiando a planilha **Função Frequência** com o procedimento apresentado no Apêndice 3 deste capítulo. Na planilha copiada, são removidos os registros desnecessários mantendo apenas a tabela de frequências absolutas.

Depois de preparar a tabela de frequências absolutas, o próximo passo é construir o histograma correspondente. Uma forma rápida de construir o gráfico é a seguinte:

- Selecione as células das frequências absolutas que serão utilizadas no gráfico, intervalo E3:E10, incluindo o título da coluna.
- Clique no ícone assistente de gráfico  e siga as instruções da caixa de diálogo **Assistente de gráfico**. Na **etapa 1 de 4 – tipo de gráfico** do assistente, mantenha-se na página **Tipos padrão** e selecione o tipo de gráfico **Colunas** e o subtipo de gráfico **Colunas agrupadas**, como mostra a Figura 2.3.
- Ao pressionar o botão **Manter pressionado para exibir exemplo**, você verá o gráfico esperado, porém com os valores da amostra apenas a sequência de números 1, 2, ..., 7, que identifica as sete barras verticais.
- Depois de pressionar o botão **Avançar**, será exibida a caixa de diálogo **Assistente de gráfico – etapa 2 de 4 – dados de origem do gráfico**, com duas páginas com os nomes **Intervalo de dados** e **Sequência**.
- Na guia **Intervalo de dados**, deverá estar selecionado **Colunas**, e na caixa **Intervalo de dados** aparecerá o endereço do intervalo previamente selecionado com a referência do nome da planilha da pasta, nesse caso, **Construção Histograma**, Figura 2.3 esquerda.

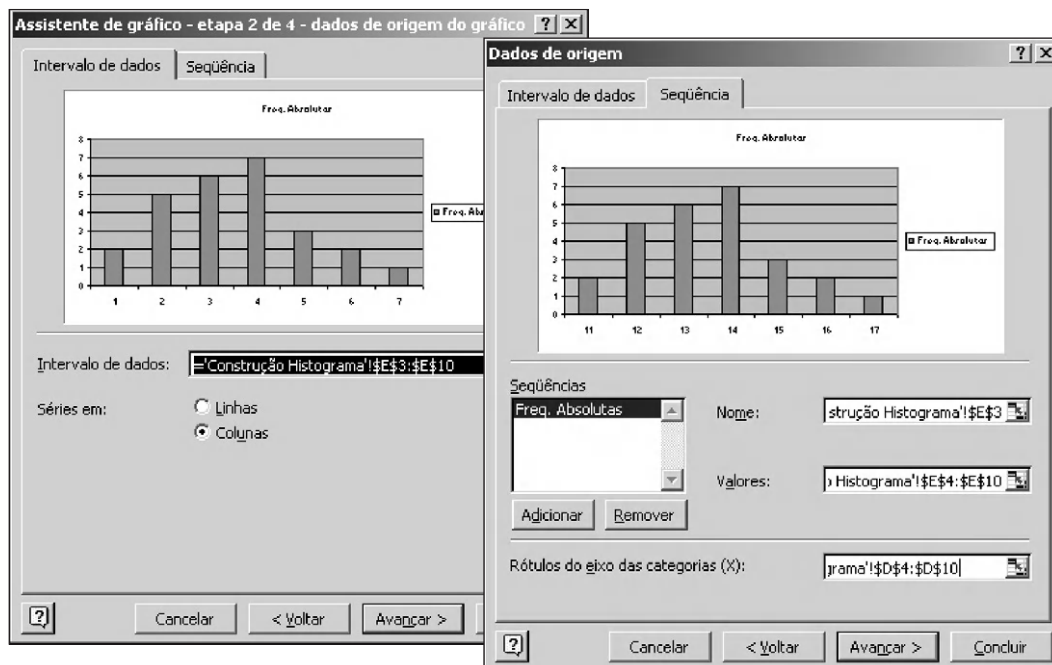


FIGURA 2.3 Assistente de gráfico – etapa 2 de 4 – dados de origem.

- Selecione a guia **Sequência** com a construção do gráfico e:
 - Na caixa de listagem **Sequências**, é exibido o título **Freq. Absolutas** registrado no intervalo da planilha E3:E10, Figura 2.3 à direita.
 - Na caixa **Nome**, está registrada a célula \$E\$3 com a referência do nome da planilha da pasta, neste caso, **Construção Histograma**.
 - Na caixa **Valores**, está registrada a fórmula do intervalo da planilha E4:E10 referente ao eixo de ordenadas ou frequências.
 - A caixa **Rótulos do eixo das categorias (X)** está em branco e deve ser preenchida com os dados do intervalo D4:D10. Para isso proceda desta forma:
 - Posicione o cursor do mouse nessa caixa vazia.
 - Depois, com o mouse, apenas selecione o intervalo D4:D10. A Figura 2.3 à direita mostra o intervalo D4:D10 depois de ser registrado e depois de os valores desse intervalo serem registrados no gráfico. Agora o gráfico mostrado é o histograma que queremos.

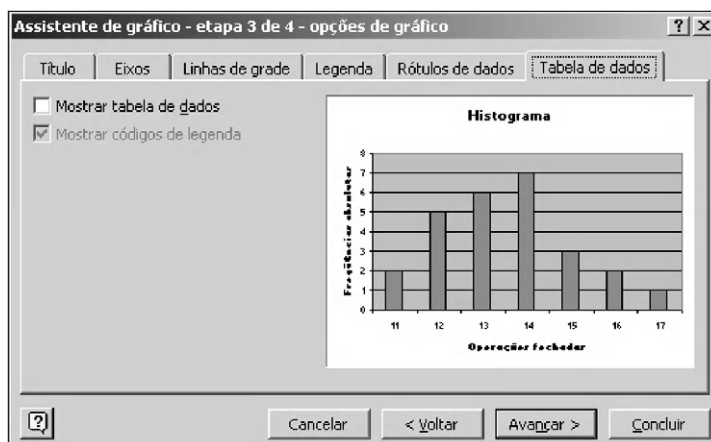


FIGURA 2.4 Assistente de gráfico – etapa 3 de 4 – opções de gráfico.

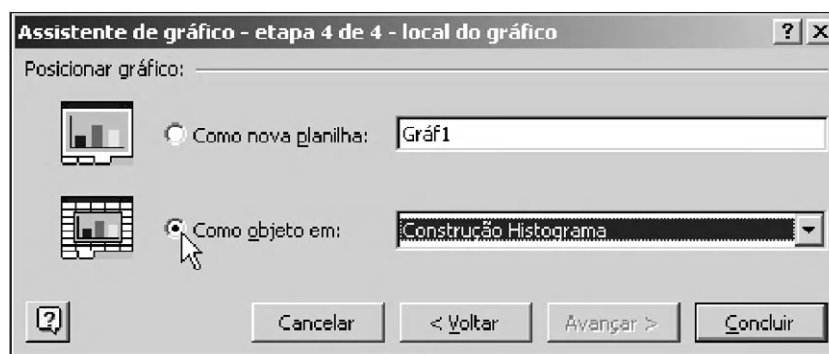
Novamente, depois de pressionar o botão **Avançar**, o Excel exibirá a caixa de diálogo **Assistente de gráfico – etapa 3 de 4 – opções de gráfico** contendo seis páginas e o gráfico desenhado na própria caixa de diálogo. Essa etapa do assistente permitirá realizar mudanças na apresentação do gráfico, Figura 2.4. Uma característica interessante dessa etapa é que, conforme você muda as configurações, elas aparecem no gráfico da própria caixa de diálogo.

- Na primeira página **Título**, procedemos como segue:
 - **Título do gráfico.** Aparece o nome *Freq. Absolutas*, pois é o nome da coluna dos valores informados. Substituímos esse nome pelo nome *Histograma*.
 - **Eixo das categorias (X).** Registramos *Operações fechadas*.
 - **Eixo dos valores (Y).** Registramos *Frequências absolutas*.
Observe que à medida que for registrando as letras dos títulos, o gráfico da caixa de diálogo vai incorporando essas letras. A Figura 2.4 mostra a caixa de diálogo com o gráfico depois de completar os registros. Nas outras cinco páginas, é possível realizar outras mudanças e, ao mesmo tempo, visualizar seus resultados no gráfico.
- Na página **Eixos**, é possível modificar as escalas dos dois eixos, bem como alterar o tipo de informação incluída no eixo X.
- Na página **Linhas de grade**, é possível adicionar ou retirar linhas de grade nos dois eixos. Nesse caso, mantemos somente as linhas de grade principais dos valores Y e desmarcamos todas as demais opções.
- A página **Legenda** refere-se à legenda *Freq. Absolutas*, posicionada à direita do gráfico. Nesse caso, desmarcamos a opção **Mostrar legenda**. A legenda desaparece e o gráfico fica maior.
- Na página **Rótulo de dados**, é possível incluir os valores das ordenadas ou das abscissas.
- Na página **Tabela de dados**, é possível incluir a tabela dos dados combinada com os valores do eixo de abscissas, alternativa que deixamos para o leitor verificar.

Antes de continuar verifique a visualização do gráfico para certificar-se de que está como deseja. Completadas as escolhas anteriores, pressionando o botão **Avançar**, o Excel exibirá a caixa de diálogo **Assistente de gráfico – etapa 4 de 4 – local do gráfico**, Figura 2.5:

- Posicione o gráfico **Como nova planilha**. Escolhendo essa opção, o Excel criará a planilha de gráfico **Gráf1** ou com o nome que você registrar nessa caixa.
- Posicione o gráfico **Como objeto em**. Escolhendo essa opção, o Excel construirá o gráfico na planilha de cálculo registrada, nesse caso **Construção Histograma**, ou em outra planilha que escolher.

FIGURA 2.5 Assistente de gráfico – etapa 4 de 4 – local do gráfico.



Pressionando o botão **Concluir**, o Excel construirá o gráfico na planilha **Construção Histograma** mostrada na Figura 2.6.

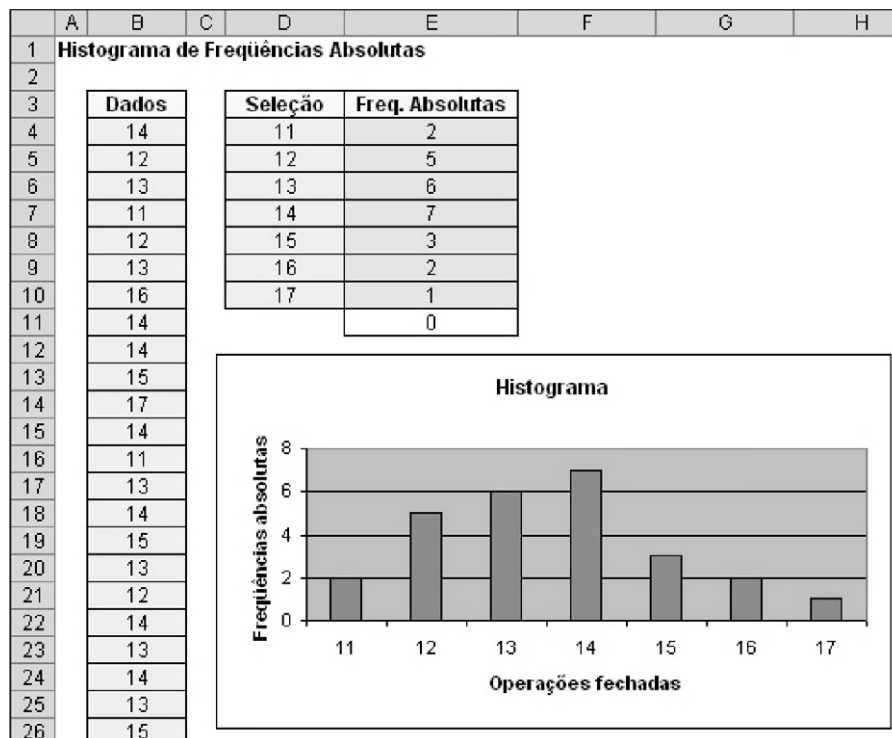




FIGURA 2.6
Histograma de frequências absolutas do Exemplo 2.1.

Todos os histogramas numa única planilha

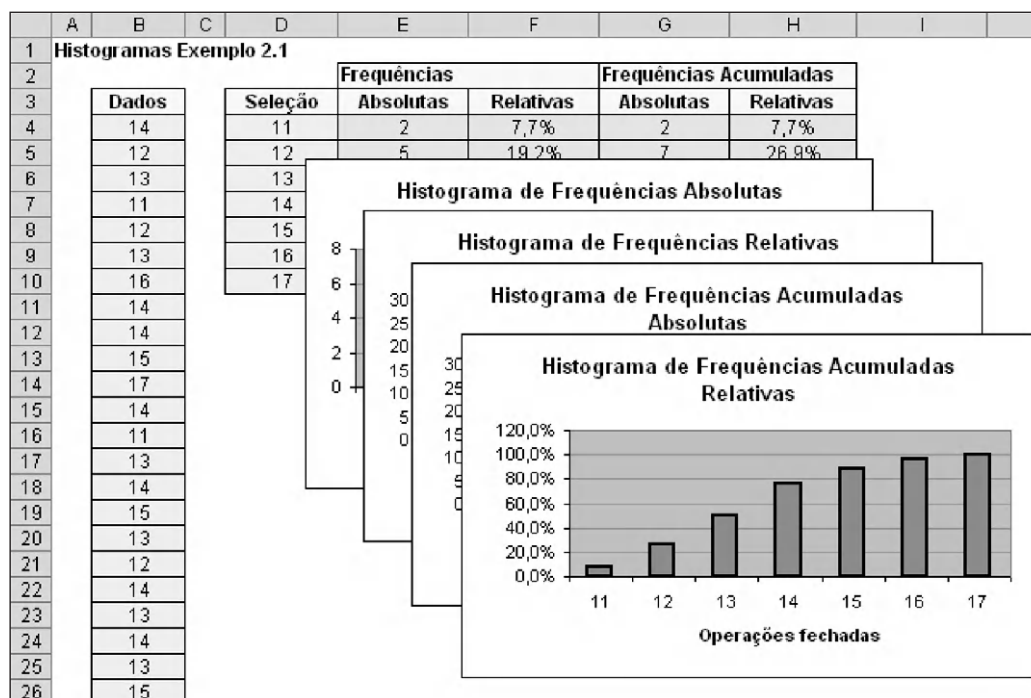
Seguindo o roteiro apresentado anteriormente, é possível construir os outros três histogramas, de frequências relativas, de frequências acumuladas absolutas e de frequências acumuladas relativas. Deve-se cuidar para construir corretamente as tabelas de frequências correspondentes. Também é importante lembrar que as formas dos histogramas de frequências absolutas e frequências relativas são a mesma, mudando apenas a escala de ordenadas dos gráficos, situação que também ocorre com os histogramas de frequências acumuladas absolutas e frequências acumuladas relativas. Essa semelhança ajudará na construção de todos os histogramas em uma única planilha.

O primeiro passo é a preparação da planilha **Histogramas Exemplo 2.1**, incluída na pasta **Capítulo 2**, contendo a amostra e as quatro tabelas de frequências calculadas na mesma planilha. Para facilitar a preparação dessa planilha, pode-se fazer uma cópia da planilha **Construção Histograma** e, depois, construir as três tabelas de frequências restantes, a partir da tabela de frequências absolutas. Por último, os títulos devem ser adequados, mantendo o gráfico de frequências absolutas. Uma alternativa de construção do histograma de frequências relativas é repetir o procedimento apresentado na seção anterior, assunto que deixamos por sua conta. Outra forma é fazer uma cópia do histograma já construído procedendo assim:

- Selecione o histograma de frequências absolutas e no menu **Editar** selecione **Copiar**, ou com as teclas **Control+C**, ou pressionando o ícone copiar .
- Depois, selecione uma célula vazia da mesma planilha e no menu **Editar**, selecione **Colar** ou com as teclas **Control+V**, ou pressionando o ícone colar .
- A seguir, clique com o botão esquerdo do mouse em cima do gráfico copiado para selecioná-lo. Com o cursor em cima do novo gráfico, clique com o botão direito do mouse e, no menu apresentado, selecione **Dados de origem**.

- Na caixa de diálogo **Dados de origem** apresentada pelo Excel:
 - Selecione a página **Intervalo de dados**. Na caixa **Intervalo de dados** estará selecionado o intervalo do gráfico de frequências absolutas. Para substituir esse intervalo com o cursor do mouse, selecione o intervalo F3:F10 correspondente às frequências relativas.
 - Escolha a página **Sequência** e, na caixa **Rótulos do eixo das categorias (X)**, que deverá estar vazia, com o cursor do mouse, selecione o intervalo D4:D10.
 - Pressione o botão **OK** para concluir.

FIGURA 2.7
Histograma de frequências relativas do Exemplo 2.1.



Como o nome do gráfico permaneceu o mesmo do gráfico copiado, será necessário mudar esse nome. Há dois procedimentos.

- Clicar com o botão esquerdo do mouse em cima do título do histograma e selecionar a palavra que deverá ser substituída, *Absolutas* neste caso. A seguir, digite *Relativas* e, para sair, clique com o botão esquerdo em qualquer lugar da planilha.
- Ou clicar com o botão esquerdo do mouse em cima do gráfico copiado para selecioná-lo e, depois, com o cursor em cima do novo gráfico, clicar com o botão direito do mouse e, no menu apresentado, selecionar **Opções de gráfico**. Na página **Título**, escolher a página **Título do gráfico** na qual aparece *Histograma de Frequências Absolutas*. A seguir, substituir *Absolutas* por *Relativas* e pressionar o botão **OK**.

O procedimento anterior é repetido para os dois últimos histogramas de frequências acumuladas, absolutas e relativas. A Figura 2.7 mostra a planilha **Histogramas, Exemplo 2.1**, com os quatro histogramas.

Qualquer um dos histogramas poderá receber modificações; por exemplo, você poderá mudar a cor de fundo das barras, ou a cor das próprias barras escolhendo cores únicas ou mesclas de cores:

- Para mudar a cor da área do histograma, clique com o botão esquerdo do mouse em cima da área do histograma e depois com o botão direito para selecionar **Formatar área de plotagem**. Na caixa de

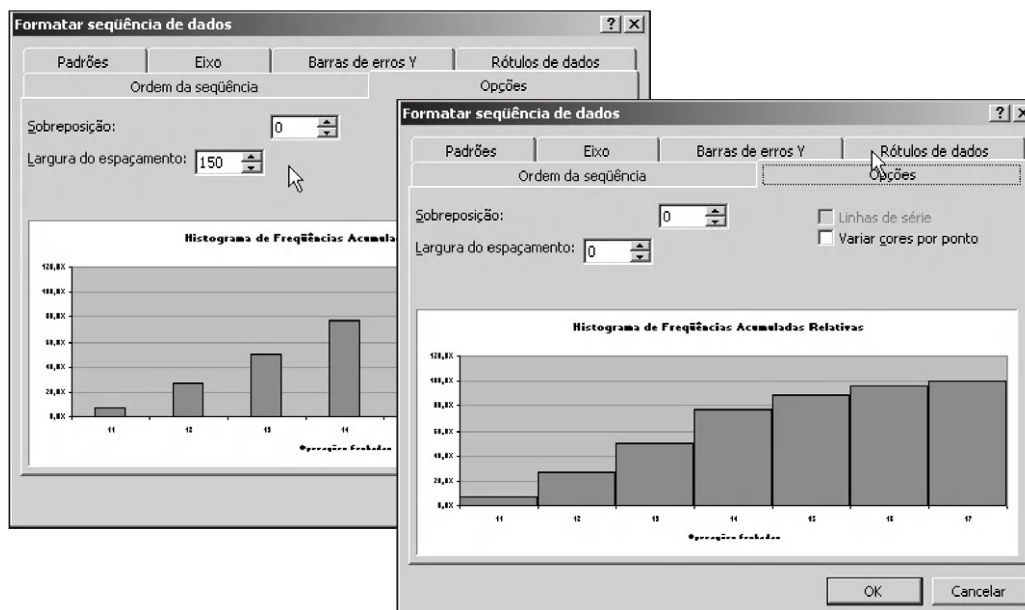


FIGURA 2.8 Mudando a largura das barras verticais.

diálogo apresentada pelo Excel **Formatar área de plotagem** na página **Padrões**, é possível escolher **Borda** ou **Área** e, dentro desta última, incluir efeitos de preenchimento na área do histograma pressionando o botão com o mesmo nome.

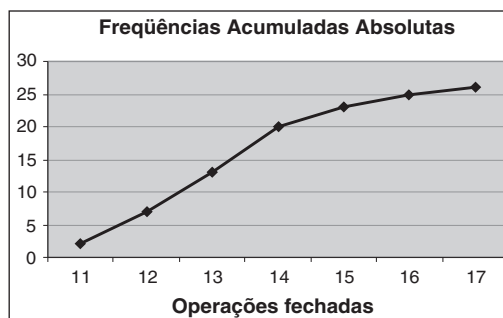
- Para mudar a cor das barras ou colunas do histograma, clique com o botão esquerdo do mouse em cima de uma das colunas do histograma e depois com o botão direito para selecionar **Formatar sequência de dados**. Na caixa de diálogo apresentada pelo Excel **Formatar sequência de dados** na página **Padrões**, é possível escolher **Borda** ou **Área** para mudar as cores procedendo de forma parecida à apresentada anteriormente.
- Na página **Opções**, é possível aumentar a largura das barras ou colunas. Por exemplo, pressionando o botão giratório até registrar o valor zero na caixa **Largura do espaçamento**, consegue-se aumentar as barras do histograma até não ficar nenhum vazio, como mostra a Figura 2.8. Um histograma sem espaços vazios entre as barras ou colunas é bem aceito. Sugerimos que você teste as outras opções desta página, por exemplo, a escolha de **Variar cores por pontos**.

Gráfico poligonal – ogiva

As barras ou colunas verticais dos histogramas construídos podem ser substituídas por uma linha, recebendo o nome de *poligonal*. Esse tipo de representação é interessante no caso do histograma de frequências acumuladas. Por exemplo, a poligonal da distribuição de frequências acumuladas do Exemplo 2.1, que se acostuma denominar *ogiva*, foi construída na planilha **Histogramas, Exemplo 2.1**, incluída na pasta **Capítulo 2**, procedendo como segue, Figura 2.8:

- Faça uma cópia do gráfico *Frequências Acumuladas Absolutas*.
Clique no gráfico e depois de clicar com o botão direito do mouse selecione **Tipo de gráfico**. Na caixa **Tipo de gráfico**, selecione a guia **Tipos padrão** e depois, na caixa **Tipo de gráfico**, primeiro selecione o gráfico **Linha** e depois selecione o gráfico **Linhas com marcadores exibidos a cada valor de dado** e, para terminar, pressione OK.
- O gráfico construído é o apresentado na Figura 2.8 depois de mudar algumas formatações e títulos, como já explicado.

FIGURA 2.9 Poligonal das frequências acumuladas.

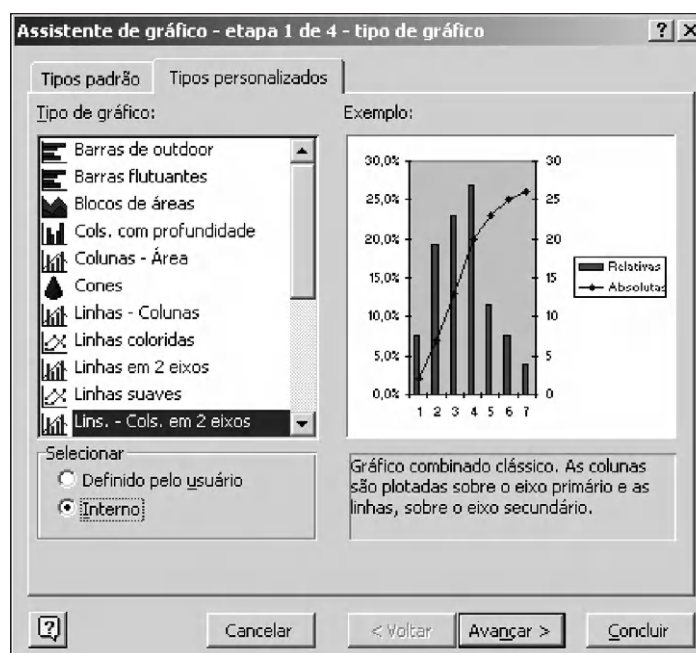


Histograma combinado

Os recursos do Excel permitem também construir o histograma combinado de frequências relativas e frequências acumuladas absolutas, ou outra combinação adequada, como foi realizado na planilha **Histogramas Exemplo 2.1**, incluída na pasta **Capítulo 2**, procedendo como segue.

- Selecione as células das frequências absolutas que serão utilizados no gráfico, intervalo F3:F10 e G3:G10, incluindo o título de cada coluna. Para selecionar dois intervalos ao mesmo tempo, primeiro selecione um dos intervalos e, a seguir, mantendo pressionada a tecla **Ctrl**, selecione o segundo intervalo.
- Clique no ícone assistente de gráfico e siga as instruções da caixa de diálogo **Assistente de gráfico**. Na etapa 1 de 4 – tipo de gráfico do assistente, selecione a página **Tipos personalizados** e o tipo de gráfico **Lins. – Cols. em dois eixos**, Figura 2.10, que mostra o gráfico que será construído pelo Excel.

FIGURA 2.10
Escolha de **Tipo personalizado** de gráfico.



Depois de pressionar o botão **Avançar**, será exibida a caixa de diálogo **Assistente de gráfico – etapa 2 de 4 – dados de origem do gráfico** contendo duas páginas com os nomes **Intervalo de dados** e **Sequência**.

- Na guia **Intervalo de dados**, deverá estar selecionado **Colunas** e, na caixa **Intervalo de dados**, aparecerá o endereço do intervalo previamente selecionado com a referência do nome da planilha da pasta, neste caso, **Histogramas Exemplo 2.1**.
- Selecione a guia **Sequência** com a construção do gráfico e:
 - Na caixa de listagem **Sequências**, aparecerão os títulos registrados no intervalo da planilha F3:F10 e G3:G10, os nomes **Relativas** e **Absolutas**, Figura 2.11.
 - Na caixa **Nome**, está registrada a célula \$F\$3 com a referência do nome da planilha da pasta, neste caso, **Histogramas Exemplo 2.1**.
 - Na caixa **Valores**, está registrada a fórmula do intervalo da planilha F4:F10 referente ao eixo de ordenadas ou frequências.
 - A caixa **Rótulos do eixo das categorias (X)** está em branco e deve ser preenchida com os dados do intervalo D4:D10.
 - A caixa **Rótulos do eixo das segundas categorias (X)** está em branco e deve ser preenchida, também, com os dados do intervalo D4:D10.

Pode-se verificar que o gráfico mostrado no assistente é o histograma que esperamos, Figura 2.11.

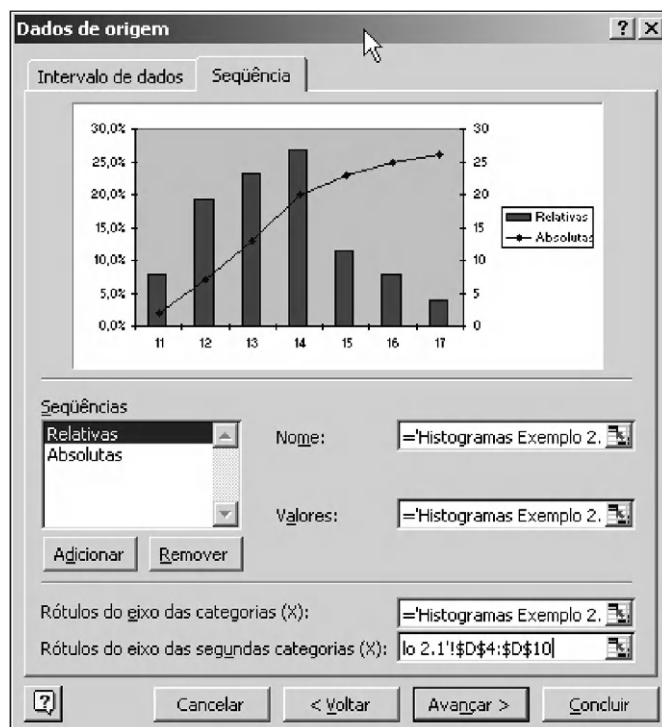
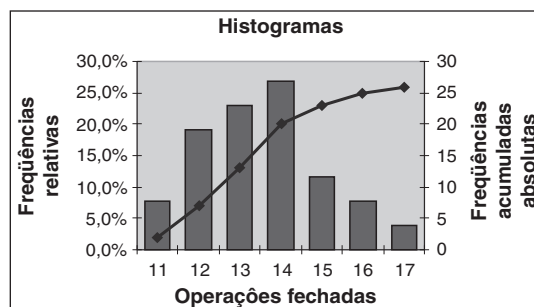


FIGURA 2.11

Assistente de gráfico – etapa 2 de 4 – dados de origem.

Depois de pressionar o botão **Avançar**, o Excel exibirá a caixa de diálogo **Assistente de gráfico – etapa 3 de 4 – opções de gráfico** contendo seis páginas e o gráfico desenhado na própria caixa de diálogo. Essa etapa do assistente permitirá realizar mudanças na apresentação do gráfico da mesma forma como já foi apresentado, porém para dois eixos de ordenadas. Tenha em mente que uma característica interessante dessa etapa é que, conforme você muda essas configurações, elas aparecem no gráfico da própria caixa de diálogo. Portanto, antes de continuar, verifique a visualização do gráfico para certificar-se de que está como deseja. A Figura 2.12 mostra o histograma concluído depois de alguns ajustes, espessura da linha, aumento da largura das colunas ou barras, ajuste dos corpos das fontes dos títulos etc.

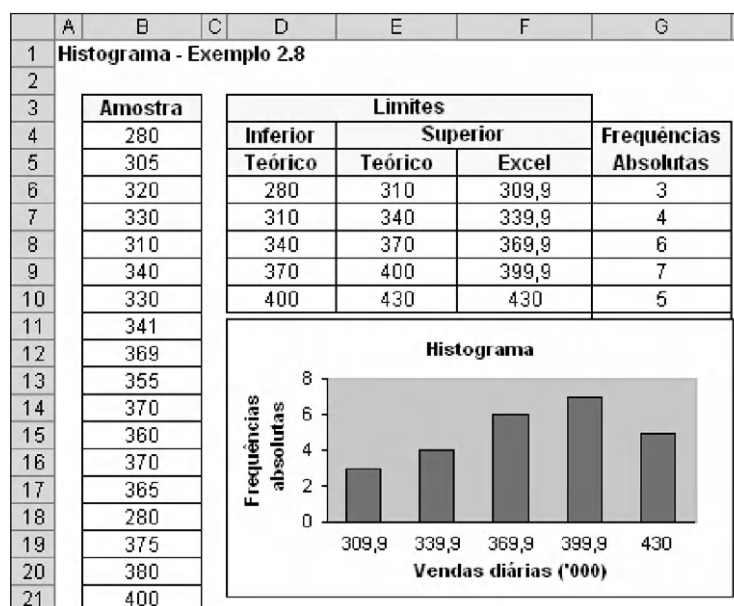
FIGURA 2.12 Histograma combinado.



Histograma com dados quantitativos contínuos

Agora será apresentada a construção do histograma de uma amostra contendo dados quantitativos contínuos. O procedimento de construção do histograma é o mesmo, o que muda é a forma de preparar os limites das classes para trabalhar corretamente com a planilha Excel. O primeiro passo é a preparação da planilha **Histograma Exemplo 2.8**, incluída na pasta **Capítulo 2**, contendo a amostra e a tabela de frequências absolutas e relativas calculadas na mesma planilha. Para facilitar a preparação dessa planilha, pode-se fazer uma cópia da planilha **Exemplo 2.9** com o procedimento apresentado no Apêndice 3 deste capítulo. Depois de copiada, na nova planilha, são apagados os registros desnecessários mantendo apenas as tabelas de frequências absolutas. Vimos que, para utilizar o Excel e manter o limite superior da classe aberto, o limite superior deverá ser inferior ao limite teórico, como apresentado no Exemplo 2.9 e repetido no intervalo D3:F10 da planilha **Histograma Exemplo 2.8**, Figura 2.13.

FIGURA 2.13
Histograma dados quantitativos contínuos.



Nas duas primeiras colunas da tabela apresentada na Figura 2.13, foram registrados o limite inferior e superior de cada classe em ordem crescente a partir da primeira classe. Como a mínima variação dos valores da amostra é uma unidade, na terceira coluna, denominada Excel, foi registrado o limite superior de cada classe igual ao limite superior registrado na segunda coluna subtraído do valor 0,1, com exceção da última classe. Para construir o histograma de frequências absolutas, utilizamos o procedimento já apresentado utilizando as colunas Excel e Frequências Absolutas da tabela. Analisemos o histograma de frequências absolutas da Figura 2.13, na qual os valores do eixo de abscissas se referem ao limite superior de cada classe.

- A barra vertical com o valor 369,9 e frequência absoluta 6 indicam que a frequência dos valores menores ou iguais a 369,9 e maiores do que 339,9 é 6. Da mesma maneira, o número de valores maiores ou iguais a 340 e menores do que 370 é 6.
- Considerando a variação mínima igual a 1 entre os valores da amostra *Vendas diárias* e a redução 0,10 utilizada para definir os limites superiores das classes para construir o histograma com o Excel, poderíamos formatar a escala de abscissas sem a parte decimal e representar o histograma com os limites teóricos como mostra a Figura 2.14. Nesse caso, a barra vertical com o valor 370 e frequência absoluta 6 indicam que a frequência dos valores menores ou iguais a 370 e maiores do que 340 é seis.
- Como todos os valores da amostra estão distribuídos entre o valor mínimo 280 e máximo 430, deve-se entender que não há nenhum vazio entre as barras do histograma; as classes preenchem todo o espaço, como mostra a Figura 2.14.

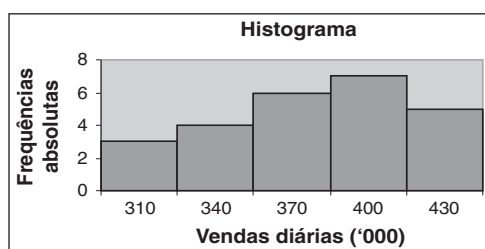


FIGURA 2.14 Histograma com barras mais largas.

O aumento da largura das colunas do histograma da Figura 2.14 foi realizado com o procedimento já apresentado do Excel e a seguir resumido:

- Clicando com o botão esquerdo do mouse em cima de uma barra qualquer do gráfico, todas as barras serão automaticamente selecionadas.
- Com o mouse em uma barra qualquer do gráfico, pressione o botão direito do mouse e, no menu, selecione **Formatar sequência de dados**.
- A caixa de diálogo **Formatar sequência de dados** tem seis páginas. Selecione a página **Opções** e, depois, na caixa **Largura do espaçamento**, registre o valor zero. Para finalizar, pressione o botão OK.
- Nas outras cinco caixas de diálogo, você poderá realizar outras modificações, por exemplo, mudar a cor das barras, do fundo do gráfico etc.

Ferramenta de análise *Histograma*

A partir de uma amostra registrada em uma planilha Excel, uma série de valores de uma amostra registrados em uma ou mais colunas contíguas, a ferramenta de análise *Histograma* retornará soluções integradas das tabelas de frequências e histogramas, registrados a partir do endereço informado pelo usuário. A amostra que será analisada com a ferramenta de análise *Histograma* deve estar registrada em uma planilha, como a de nome **Ferramenta Histograma**, incluída na pasta **Capítulo 2**, onde:

- No intervalo B3:B29, foram registrados os valores numéricos da amostra do Exemplo 2.1, incluindo o nome *Amostra* na célula B3. Os valores da amostra podem ser registrados em uma linha, uma coluna ou combinando linhas e colunas, contanto que sejam contíguos e possíveis de identificá-los com um único intervalo.
- No intervalo D4:D10, foram registrados os valores de seleção utilizados no Exemplo 2.1, incluindo o nome *Seleção* na célula D3.

Para utilizar a ferramenta *Histograma*:⁵

- Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, Figura 1.7 do Capítulo 1.
- Escolhendo a ferramenta **Histograma** e depois pressionando o botão **OK**, será exibida a caixa de diálogo **Histograma** mostrada na Figura 2.15 depois de selecionadas algumas opções.
- Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel exibirá a página *Sobre a caixa de diálogo Histograma* pertencente à *Ajuda do Excel*.

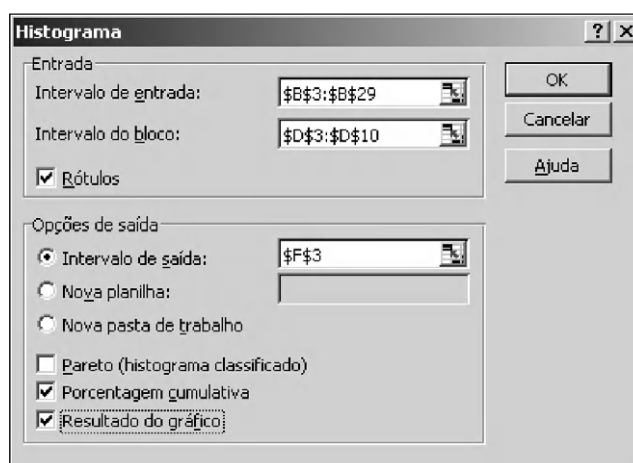


FIGURA 2.15 Caixa de diálogo da ferramenta *Histograma*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo da ferramenta *Histograma* são:

- **Intervalo de entrada.** Informar o intervalo de células da planilha na qual os dados estão registrados, nesse caso, o intervalo B3:B29 que inclui a célula onde foi registrado o título *Amostra*, ou rótulo no Excel.
- **Intervalo do bloco.** A informação deste intervalo é opcional, porém com resultados diferentes, como será mostrado. Nesse momento, foi registrado o intervalo D3:D10, que inclui a célula onde foi registrado o título *Seleção*.
 - Se não for informado nenhum intervalo do bloco, o Excel definirá os limites das classes, tendo presente que, em qualquer caso, a ferramenta *Histograma* considera os limites superiores das classes como fechados, de forma coerente com a função FREQUÊNCIA.
- **Rótulos.** Selecionamos este item, pois os intervalos informados B3:B29 e D3:D10 incluem títulos, respectivamente, *Amostra* e *Seleção*.

Na primeira parte do quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta *Histograma* registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três botões de opção que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso F3. Depois de clicar com o botão esquerdo do mouse dentro da caixa correspondente, o endereço pode ser registrado digitando F3, ou *clcando* com o botão esquerdo do *mouse* na célula F3. Nesse caso, será registrado o endereço com os dois cifrões, \$F\$3. Esse endereço é o da célula superior esquerda da tabela de frequências que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes.

⁵ Em inglês, a ferramenta HISTOGRAMA é HISTOGRAM.

- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
 - Se não for informado nenhum endereço, a ferramenta inserirá uma nova planilha com o nome **Plan**, seguido de um número sequencial. Escolhendo essa alternativa na pasta **Capítulo 2**, a ferramenta inserirá a planilha **Plan1**.
 - Há a alternativa de informar o nome da planilha na caixa desta alternativa. Registrando o nome *Teste*, a ferramenta inserirá na mesma pasta uma nova planilha com o nome **Teste**.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha **Plan1**.

Em continuação, no quadro **Opções de saída**, há três alternativas não excludentes de resultados possíveis, sendo possível selecionar qualquer combinação delas, incluindo a alternativa de não selecionar nenhuma. Essas seleções são realizadas em três caixas de seleção. Se não for selecionada nenhuma das três alternativas, a ferramenta *Histograma* apresentará a tabela de frequências absolutas, em ordem crescente de valores da amostra.

- **Pareto (histograma classificado).** Essa alternativa deve ser selecionada quando se deseja construir o gráfico de frequências absolutas em ordem decrescente de valores de frequências. O diagrama de *Pareto* é tratado com mais detalhe mais adiante neste capítulo.
- **Porcentagem cumulativa.** Selecionando essa alternativa, a ferramenta adicionará à tabela de frequências absolutas, que sempre será construída, a coluna da tabela de frequências acumuladas relativas.
- **Resultado do gráfico.** Selecionando essa alternativa, a ferramenta construirá o gráfico das tabelas de frequências escolhidas. Se desejar incluir no histograma que a ferramenta construirá a poligonal das frequências acumuladas relativas, selecione a alternativa anterior.

Depois de pressionar o botão **OK**, a ferramenta *Histograma* apresentará os resultados solicitados nas seleções realizadas, como mostra a Figura 2.16. É importante destacar que o layout da planilha foi melhorado, ajustando a largura de algumas colunas, mudando as dimensões do gráfico, as cores, o corpo da fonte dos títulos etc.

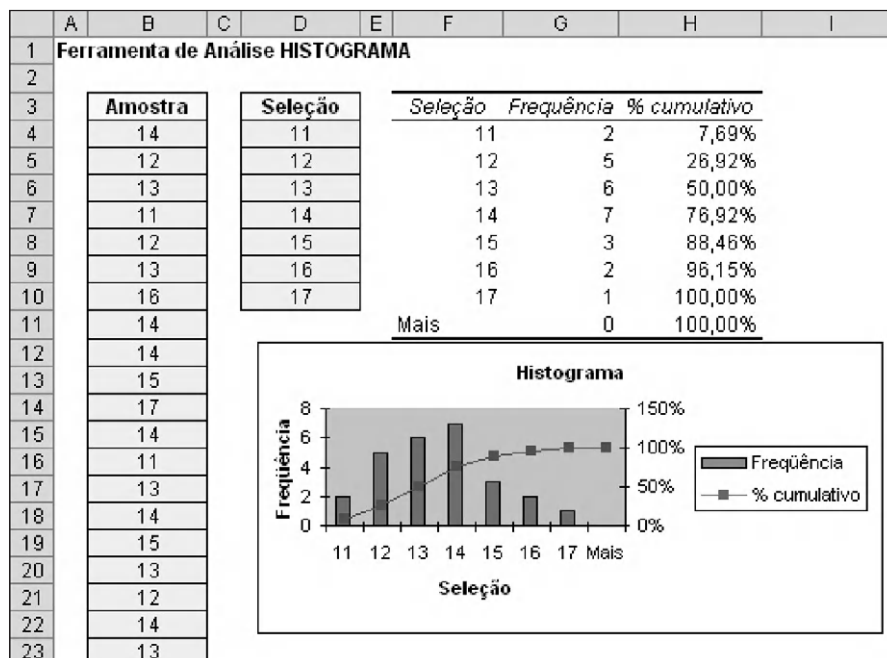


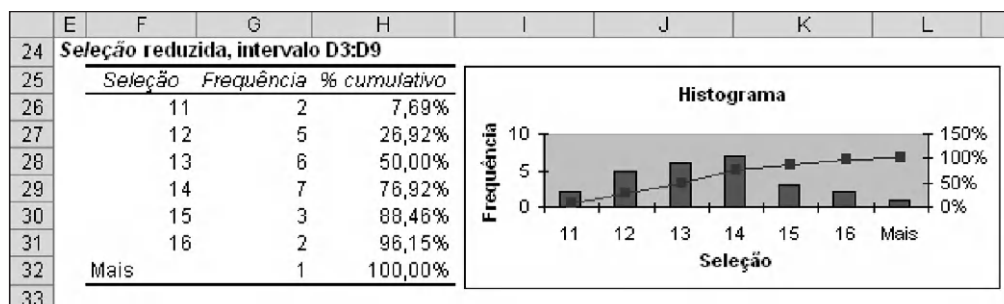
FIGURA 2.16
Histograma do Exemplo 2.1 com a ferramenta *Histograma*.

Como escolher o intervalo de seleção

Talvez você esteja estranhando a última linha *Mais* da coluna *Seleção* com frequência zero, construída pela ferramenta *Histograma*, bem como a última seleção *Mais* do histograma sem coluna. Da forma como foi selecionado o intervalo *Seleção*, a ferramenta adicionou, por sua conta, o valor *Mais* nas tabelas de frequências e no gráfico. É isso aí! Se for informada a tabela completa de seleção de valores, a ferramenta adicionará mais um valor que denomina *Mais*. Para que a ferramenta de análise *Histograma* construa os mesmos gráficos que obtivemos utilizando somente os recursos da função *FREQUÊNCIA* e os gráficos do Excel, não se deve informar o maior valor de seleção, nesse caso registrado na célula D10.

Informando o intervalo D3:D9 no **Intervalo do bloco**, mantendo selecionadas **Porcentagem cumulativa** e **Resultado do gráfico** e escolhendo a célula F25 da planilha **Ferramenta Histograma** para registrar as tabelas de frequências, a ferramenta *Histograma* apresentou os resultados mostrados na Figura 2.17. A ferramenta continua apresentando a última barra com *Mais*, porém agora se refere aos valores superiores a 16, que, nesse exemplo, é o próprio valor 17, pois o limite superior de cada classe é considerado fechado ou o limite superior de cada classe inclui o próprio valor registrado.

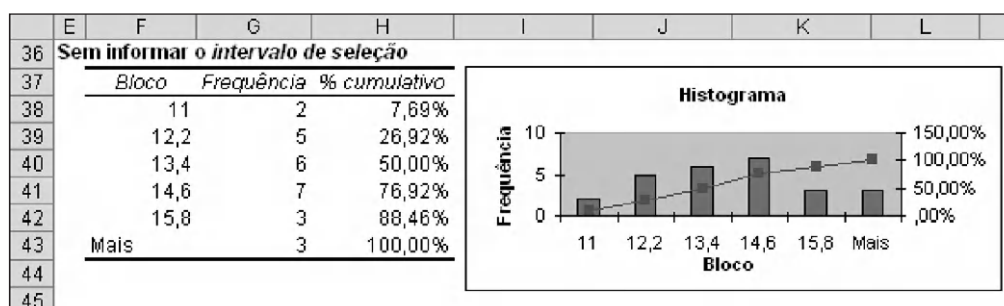
FIGURA 2.17
Gráfico com
intervalo reduzido.



Sem informar o intervalo de seleção

Dissemos que a informação no **Intervalo do bloco** é opcional, porém com resultados diferentes. Omitindo apenas essa informação, a partir da linha 36, a ferramenta *Histograma* construirá um gráfico usando critérios próprios e semelhantes aos apresentados na construção de distribuições de frequências com classes. Portanto, omitindo a informação na caixa **Intervalo do bloco**, mantendo selecionadas **Porcentagem cumulativa** e **Resultado do gráfico** e escolhendo a célula F37 da planilha **Ferramenta Histograma** para registrar as tabelas de frequências, a ferramenta *Histograma* apresentou os resultados a partir dessa célula, como mostrado na Figura 2.18. Como não foi informado o intervalo de seleção, a ferramenta apresentou seus resultados de seleção na coluna de título *Bloco*, célula F37.

FIGURA 2.18
Ferramenta *Histograma*,
omitindo o intervalo
de seleção.



Analisemos os resultados da Figura 2.18. A ferramenta *Histograma* formou seis classes construídas como segue, conclusões baseadas na observação de resultados utilizando a ferramenta *Histograma*.

- Como a amostra tem 26 observações, as três fórmulas apresentadas para determinação do número de classes recomendam escolher cinco classes. O número de classes escolhido pela ferramenta de análise *Histograma* é igual ao resultado de somar um ao valor cinco obtido por uma das três fórmulas. Entretanto, como o limite superior da primeira classe é o valor mínimo da amostra, podemos concluir que a ferramenta *Histograma* cria seis colunas, porém com amplitudes correspondentes a cinco classes.
 - O intervalo de variação é seis, resultado da diferença entre o valor máximo observado e o mínimo, $6=17-11$.
 - As cinco últimas classes têm a mesma amplitude igual a 1,20, valor obtido como resultado de dividir o intervalo de variação seis pelo número cinco.
- O limite superior da segunda classe é $12,2=11+1,2$ e os limites superiores das três classes seguintes são obtidos de forma equivalente.
- O limite superior da sexta e última classe é o maior valor da amostra.

Comparando o histograma da Figura 2.17 com o da Figura 2.18, a forma da distribuição de frequências do primeiro histograma representa a amostra de forma mais adequada, pois tanto o perfil da subida quanto o da descida é mais contínuo, enquanto no segundo histograma aparece um patamar constante nos últimos dois valores. Entendemos que as ferramentas de análise do Excel devem ser utilizadas da forma como foram desenvolvidas, salvo que seus resultados apresentem erros. Sugerimos que a ferramenta de análise *Histograma* seja utilizada sem especificar o **Intervalo do bloco**. Você deverá analisar se os resultados da ferramenta estão dentro de sua expectativa de aceitação. Se não for assim, lembre-se de que o que fornece a ferramenta *Histograma* pode ser conseguido, como foi mostrado, com os recursos das funções estatísticas e dos gráficos do Excel.

Gráfico de Pareto

Terminando o século XVIII, o economista italiano Vilfredo Pareto mostrou que 80% da terra na Itália pertencia a 20% da população, confirmação socioeconômica que posteriormente teve aplicação universal.⁶ Por exemplo, 80% dos estoques de uma empresa são preenchidos por 20% dos produtos; 20% dos clientes são responsáveis por 80% das vendas, 80% das informações de que os usuários precisam estão nos primeiros 20% dos resultados das páginas de busca de Internet etc. Sobre defeitos e perdas, verifica-se que não são muitos os fatores que realmente causam desvios indesejáveis.

O que é o gráfico de Pareto? As barras ou colunas do histograma de frequências absolutas são desenhadas com os valores das observações ordenadas de forma crescente. No gráfico de Pareto, as barras ou colunas do histograma de frequências absolutas estão desenhadas com as frequências ordenadas de forma decrescente, primeiro a coluna de maior frequência e por último a de menor frequência. Na planilha **Gráfico de Pareto**, incluída na pasta **Capítulo 2**, foi construído o gráfico de Pareto utilizando a ferramenta de análise *Histograma*. A Figura 2.19 mostra as tabelas e o gráfico de Pareto tendo informado o intervalo de seleção D3:D9 e selecionado as três alternativas, **Pareto**, **Porcentagem cumulativa** e **Resultado do gráfico**. A ferramenta *Histograma* apresenta a tabela de frequências absolutas, acumuladas absolutas e, em continuação, a tabela da distribuição de Pareto absoluta e acumulada (intervalo I4:K11).

Modelo histogramas

A determinação da quantidade de classes tem um pouco do procedimento de tentativa e erro na procura da distribuição que melhor represente os valores da amostra ou variável. Realizar esse processo de aproximação de forma manual é muito trabalhoso. O uso de planilhas do Excel diminui um pouco esse trabalho, sobretudo com a ferramenta de análise *Histograma*. Todavia, esse procedimento não é práti-

⁶ Hitoshi Kume – *Métodos Estatísticos para a Melhoria da Qualidade* – Editora Gente, 1993.

co, pois devem ser informados outros dados como o intervalo de seleção para construir um histograma com outra quantidade de classes. Esses inconvenientes são eliminados no *Modelo Histograma* construído pelo autor na pasta **Modelo Histogramas**, que pode ser encontrada na página do livro, no site da Editora. A Figura 2.20 mostra esse modelo para uma amostra de tamanho $n=1.000$.

FIGURA 2.19

Gráfico de Pareto com a ferramenta *Histograma*.

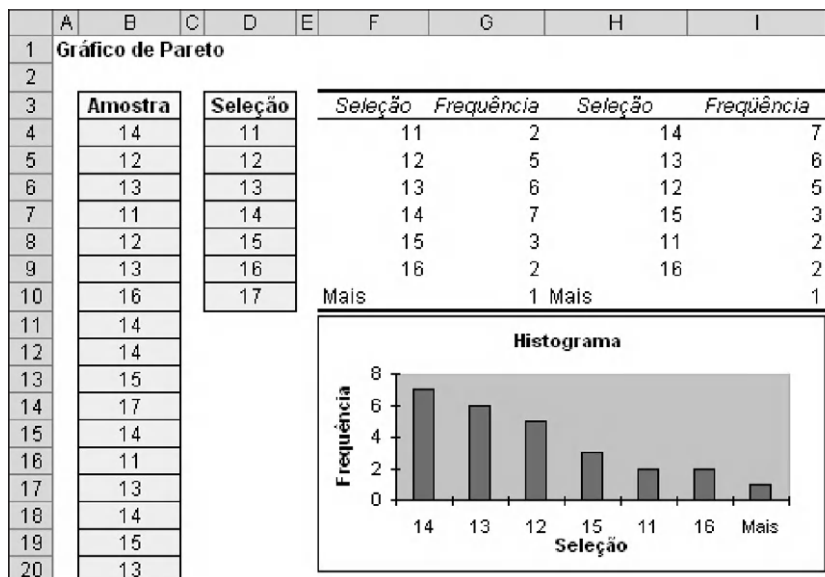
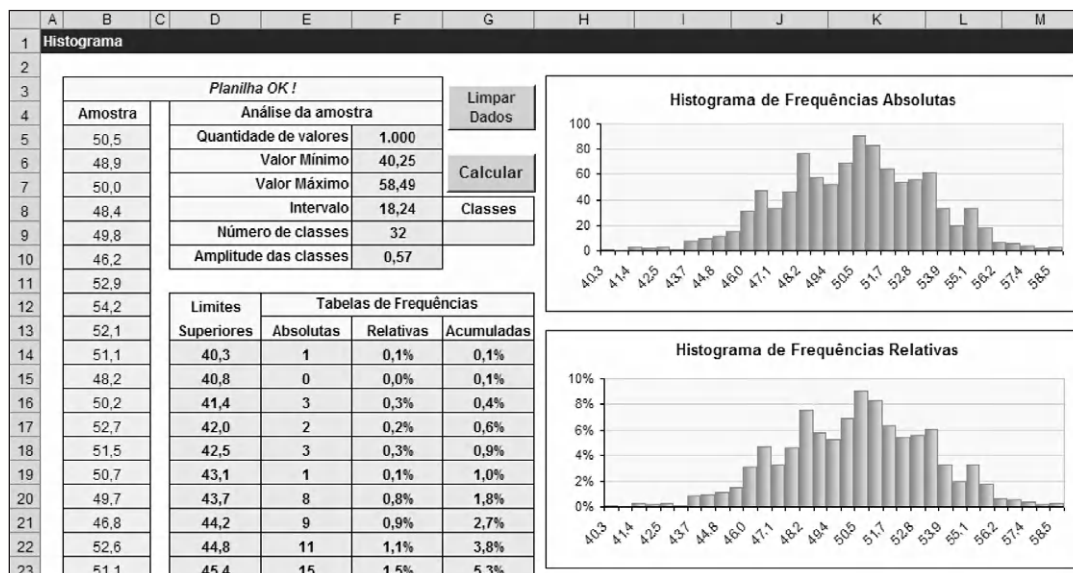


FIGURA 2.20 Modelo Histogramas.



O *Modelo Histogramas* constrói três histogramas, frequências absolutas, frequências relativas e frequências relativas acumuladas, a partir das respectivas tabelas também construídas na planilha. Para operar o *modelo*:

- Recomenda-se zerar os dados e resultados pressionando o botão **Limpar Dados**.
- Informe a série de valores numéricos a partir da célula B5. Não há limite de tamanho da amostra, apenas os limites impostos pela planilha Excel e a memória do microcomputador que está sendo utilizando.
- Depois de informar a amostra, pressione o botão **Calcular**. O *modelo* fornecerá os resultados do intervalo F5:F10 e construirá as tabelas de frequências e os histogramas. A quantidade de classes é determinada com a fórmula $k = \sqrt{n}$, utilizando a parte inteira do resultado.

- Querendo testar um número de classes diferente do sugerido pelo modelo, deve-se, primeiro, informar o novo valor na célula G9 e depois pressionar o botão **Calcular**. O *modelo* realizará todos os cálculos e mostrará os novos resultados do intervalo F5:F10 e construirá as tabelas de frequências e os histogramas. Querendo voltar para o cálculo automático, deve-se limpar a célula G9 e depois pressionar **Calcular**.
- Preste atenção aos avisos que o *modelo* apresenta na célula mesclada E3.

Dados qualitativos

O procedimento com dados qualitativos é mais simples do que com dados quantitativos. Consideremos os resultados populacionais do Censo 2000 apresentados no Capítulo 1. Consideremos a classificação por grandes grupos de idade no ano 2000, cuja planilha **Censo 2000** da pasta **Capítulo 1** foi copiada na pasta **Capítulo 2**, mantendo o mesmo nome da planilha **Censo 2000**. O intervalo I3:J6 registra a população por três grandes grupos de idade no ano 2000, resultados ligados na mesma planilha **Censo 2000**. Analisemos os gráficos da Figura 2.21.

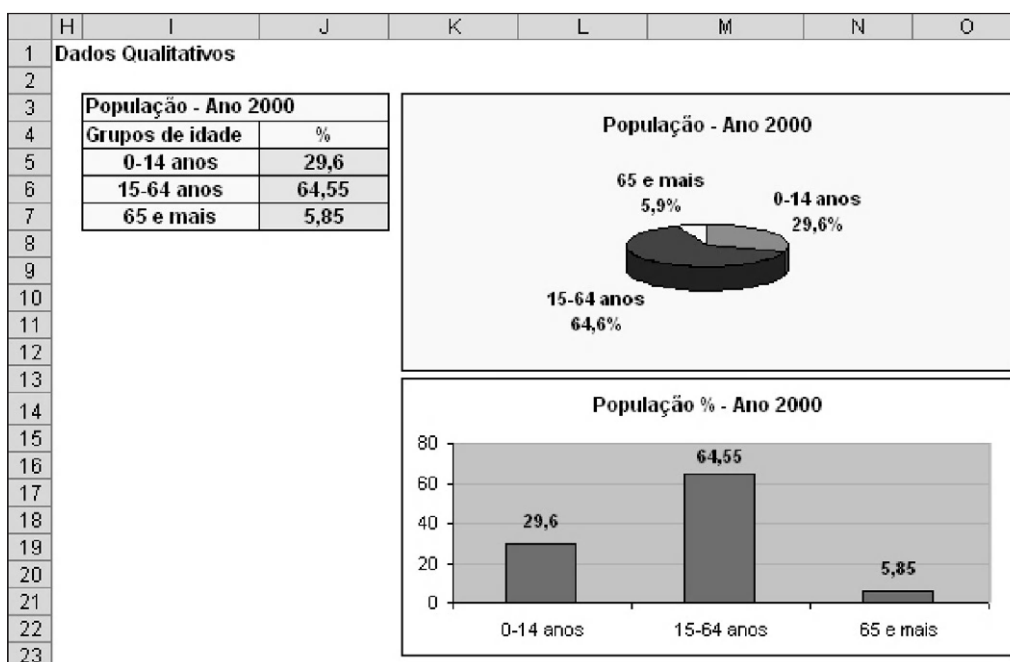


FIGURA 2.21
População por três grandes grupos de idade, ano 2000.

- O gráfico denominado *Pizza* representa a frequência dos grupos de idades como setores de um cilindro da pequena altura. Selecionando **Opções de gráfico**, é possível mudar as formatações do gráfico incluindo títulos e legendas.
- O gráfico de barras verticais representa a frequência dos grupos de idades como altura das barras. Selecionando **Opções de gráfico**, é possível mudar as formatações do gráfico incluindo títulos e legendas.

Considerando que os temas apresentados neste capítulo permitirão navegar pelos tipos de gráficos, bem como alterar as formatações, deixamos por conta do leitor as tentativas de mudar os tipos dos gráficos da Figura 2.21 utilizando dados qualitativos.

Problemas

Na planilha **Problemas**, incluída na pasta **Capítulo 2**, estão registrados problema com seus enunciados e soluções.

Apêndice 1

Funções estatísticas do Excel

Algumas medidas realizadas em uma amostra dão resultados intermediários de um procedimento de cálculo estatístico, por exemplo, a contagem da quantidade de dados de uma amostra, seu valor mínimo etc. Alguns desses resultados podem ser obtidos utilizando funções estatísticas do Excel, como mostrado na planilha **Funções estatísticas**, incluída na pasta **Capítulo 2**, aplicadas na amostra do Exemplo 2.1, Figura 2.22.

Uma característica comum das funções que serão apresentadas, exceto a função CONTAR.VAZIO, são os 30 argumentos (*núm1*; *núm2*; ... ; *núm30*) utilizados para registrar os valores de intervalos. Na apresentação da primeira função MÁXIMO, será mostrado como utilizar esses argumentos, procedimentos que se repetem com as demais funções com o mesmo tipo de argumentos. As sintaxes dessas funções estatísticas são apresentadas a seguir.

MÁXIMO(*núm1*; *núm2*; ... ; *núm30*)

A função estatística MÁXIMO⁷ retorna o valor *máximo* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.⁸ Se o nome da função MÁXIMO for inserido com letras minúsculas ou maiúsculas sem o acento ortográfico, o Excel aceitará e registrará a função com letras maiúsculas e com o acento ortográfico. Por exemplo, a função MÁXIMO aplicada na amostra do Exemplo 2.1 dará como resultado 17. Para obter esse resultado, a função MÁXIMO pode ser utilizada das seguintes maneiras, Figura 2.22:

- Registrando os valores da amostra em um intervalo de células da planilha.
 - Se os valores da variável estiverem registrados em um único intervalo, ou intervalos contíguos, apenas será necessário informar um único intervalo no argumento *num1*. Por exemplo, registrando a fórmula =MÁXIMO(B4:C16)
 - Se os valores da variável estiverem registrados em intervalos não adjacentes, será necessário informar o endereço de cada intervalo em cada argumento *núm1*; *núm2*; ... ; *núm30*. Por exemplo, na célula F6, a fórmula =MÁXIMO(B4:B11;B12:B16;C4:C6;C7:C16) registra três intervalos nos três primeiros argumentos da função MÁXIMO *núm1*; *núm2*; *núm3*
- Registrando os valores da amostra como *matriz* na própria fórmula da função, evitando registrar os valores da amostra em um intervalo de células da planilha.
 - Na célula G6, os valores foram registrados em uma única matriz:
=MÁXIMO({14;12;13;11;12;13;16;14;14;15;17;14;11;
13;14;15;13;12;14;13;14;13;15;16;12;12})
 - Na célula G7, os valores foram registrados em quatro matrizes:
=MÁXIMO({14;12;13;11;12;13;16};{14;14;15;17;14};
{11;13;14;15;13;12;14;13;14;13;15;16};{12;12})

⁷ Em inglês, a função MÁXIMO é MAX.

⁸ Assemelhados são os intervalos definidos por nomes, células vazias, valores lógicos, representações em forma de texto de números, por exemplo, VALOR("10")=10. Os argumentos que são valores de erro ou texto que não podem ser traduzidos em números geram erros.

MÍNIMO(núm1; núm2; ... ; núm30)

A função estatística MÍNIMO⁹ retorna o valor *mínimo* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células da planilha contendo valores numéricos ou semelhantes. Se o nome da função MÍNIMO for inserido com letras minúsculas ou maiúsculas sem o acento ortográfico, o Excel aceitará e registrará a função com letras maiúsculas e com o acento ortográfico. A função MÍNIMO pode ser registrada de diversas formas equivalentes às descritas na função MÁXIMO descrita anteriormente, Figura 2.22.

	A	B	C	D	E	F	G
1	Funções estatísticas						
2							
3		Amostra				Dados informados como	
4		14	13		Funções Estatísticas	Intervalo	Matriz
5		12	14				
6		13	15		MÁXIMO	17,00	17,00
7		11	13			17,00	17,00
8		12	12				
9		13	14		MÍNIMO	11,00	11,00
10		16	13			11,00	11,00
11		14	14				
12		14	13		CONT.NÚM	26,00	26,00
13		15	15			26,00	26,00
14		17	16				
15		14	12		CONT.VALORES	26,00	26,00
16		11	12			26,00	26,00
17							
18					CONTAR.VAZIO	0	
19							

FIGURA 2.22 Como utilizar as funções de estatística no Exemplo 2.1.

MÁXIMO(núm1; núm2; ... ; núm30)

A função estatística MÁXIMO¹⁰ é equivalente à função anterior MÁXIMO. A diferença está relacionada com os valores registrados nos argumentos *núm1*; *núm2*; ... ; *núm30* que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO. Deixamos que você pesquise na *Ajuda* do Excel.

MÍNIMO(núm1; núm2; ... ; núm30)

A função estatística MÍNIMO¹¹ é equivalente à função anterior MÍNIMO. A diferença está relacionada com os valores registrados nos argumentos *núm1*; *núm2*; ... ; *núm30* que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO. Deixamos que você pesquise na *Ajuda* do Excel.

CONT.NÚM(valor1; valor2; ... ; valor30)

A função estatística CONT.NÚM¹² retorna a quantidade de valores numéricos das observações *valor1*; *valor2*; ... ; *valor30*. Cada um dos argumentos *valor* pode conter ou fazer referência a vários tipos de dados; entretanto, apenas os valores numéricos serão considerados na contagem. Se o nome da função CONT.NÚM for inserido com letras minúsculas ou maiúsculas sem o acento ortográfico, o Excel acei-

⁹ Em inglês, a função MÍNIMO é MIN.

¹⁰ Em inglês, a função MÁXIMO é MAXA.

¹¹ Em inglês, a função MÍNIMO é MINA.

¹² Em inglês, a função CONT.NÚM é COUNT.

tará e registrará a função com letras maiúsculas e com o acento ortográfico. A função CONT.NÚM pode ser registrada de diversas formas equivalentes às descritas na função MÁXIMO tratada anteriormente, Figura 2.22.

CONT.VALORES(*valor1; valor2; ... ; valor30*)

A função estatística CONT.VALORES¹³ retorna a quantidade de observações não vazias contidas em *valor1; valor2; ... ; valor30*. Cada um dos argumentos *valor* pode ser qualquer tipo de informação, incluindo texto vazio (""), porém excluindo as células em branco. A função CONT.VALORES pode ser registrada de diversas formas equivalentes às descritas na função MÁXIMO tratada anteriormente, Figura 2.22.

CONTAR.VAZIO(*intervalo*)

A função CONTAR.VAZIO¹⁴ retorna o número de células vazias contidas no *intervalo* informado. Células com fórmulas que forneçam um texto vazio ("") também são contadas; entretanto, células com valores nulos não são contadas. Aplicando a função CONTAR.VAZIO para, obter o número de células vazias da série do Exemplo 2.1 dará como resultado 0, como mostra a Figura 2.22.

Apêndice 2

Fixando o endereço de células

Na célula E15 da planilha **Função Frequência**, foi registrada a fórmula =FREQUÊNCIA(\$B\$4:\$B\$29;D15) que depois foi copiada até a célula D21. Se o intervalo da amostra fosse registrado sem os dois cifrões em cada endereço de célula, não teria sido possível copiar a fórmula de forma correta, pois o intervalo B4:B29 mudaria o endereço em cada célula que fosse copiada. Portanto, os cifrões utilizados no intervalo fixam as duas células do intervalo \$B\$4:\$B\$29, que facilitará a cópia da fórmula até a célula D21. Vejamos as quatro possibilidades de referenciar uma célula em uma fórmula:

- \$B\$4 Tanto a linha quanto a coluna são *absolutas*. Se a fórmula que contém essa referência for copiada em qualquer célula da planilha, o endereço \$B\$4 permanecerá inalterado.
- B\$4 A linha é *absoluta* e a coluna é *relativa*. Se a fórmula que contém essa referência for copiada em qualquer célula da planilha, o endereço será alterado mantendo a linha inalterada e adotando a coluna da nova célula.
- \$B4 A coluna é *absoluta* e a linha é *relativa*. Se a fórmula que contém essa referência for copiada em qualquer célula da planilha, o endereço será alterado mantendo a coluna inalterada e adotando a linha da nova célula.
- B4 A linha é *relativa* e a coluna *relativa*. Se a fórmula que contém essa referência for copiada em qualquer célula da planilha, as duas coordenadas do endereço serão alteradas.

¹³ Em inglês, a função CONT.VALORES é COUNTA.

¹⁴ Em inglês, a função CONTAR.VAZIO é COUNTBLANK. Na Ajuda do Excel, a função CONTAR.VAZIO é classificada como função de informação; entretanto, no menu Inserir é classificada como função Estatística.

A escolha do tipo de referência de uma célula pode ser facilmente incluída no endereço de uma célula utilizando a tecla de função [F4] como segue:

- Em uma célula qualquer da planilha digite, por exemplo, =E13 sem inserir a fórmula na planilha.
- Depois, pressionando a tecla de função [F4], a fórmula passa a ser =\$E\$13; pressionando novamente a tecla, obtemos =E\$13, e pressionando novamente a tecla =\$E13.

Esse procedimento também pode ser utilizado, dentro de uma fórmula já construída durante sua edição.

Apêndice 3

Cópia de uma planilha

Para realizar a cópia de uma planilha na mesma ou em outra pasta proceda desta forma:

- Posicione o cursor em qualquer célula da planilha que será copiada, por exemplo, a célula F1 da planilha **Quantidade de Classes**.
- No menu **Editar**, selecione **Mover ou copiar planilha**. O Excel apresentará a caixa de diálogo da Figura 2.23. Nessa caixa de diálogo foram selecionadas:
 - **Para pasta**. Escolhemos a própria pasta. Você poderá escolher qualquer pasta que estiver aberta ou uma nova pasta.
 - **Antes da planilha**. Escolhemos **Exemplo 2.8**. Essa escolha significa que a cópia da pasta será posicionada depois da pasta **Quantidade de Classes**.
 - **Criar uma cópia**. Deve-se selecionar para copiar a planilha, pois, do contrário, o Excel simplesmente moverá a planilha atual para a nova posição.
- Pressionando o botão **OK**, o Excel criará uma nova planilha idêntica à planilha **Quantidade de Classes**, porém com o nome **Quantidade de Classes (2)**.
- Para concluir, mude o nome da planilha procedendo assim:
 - Selecione a planilha **Quantidade de Classes (2)**.
 - Clique com o botão direito do mouse. No menu, selecione **Renomear** e, a seguir, digite o novo nome da planilha. Como alternativa, consegue-se o mesmo efeito clicando duas vezes seguidas em cima da guia da planilha selecionada.

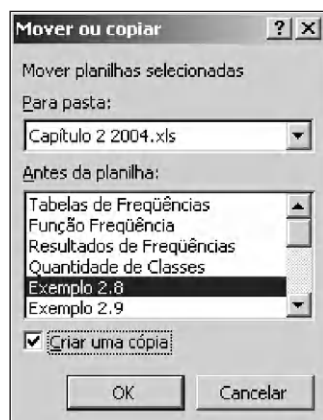
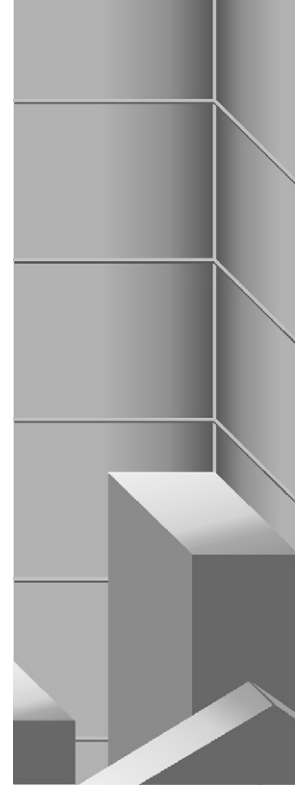


FIGURA 2.23 Caixa de Diálogo de Mover ou copiar.

Capítulo 3

MEDIDAS DE TENDÊNCIA CENTRAL



Para tentar conhecer uma ou mais características de uma população, extraímos uma amostra dessa população, conforme descrito no Capítulo 1. Em geral, quando o tamanho da amostra é grande, maior do que quinze dados, a simples inspeção das observações não será suficiente para obter as características relevantes desses valores. Para facilitar a análise e a interpretação, esses dados devem ser organizados ou resumidos, por exemplo, em tabelas de frequências e histogramas, como foi apresentado no Capítulo 2. As *medidas de ordenamento* e as *medidas de posição* são os métodos numéricos para resumir e analisar os valores de uma série de dados numéricos, seja uma amostra ou a própria população, denominados como *medidas de tendência central*. No Capítulo 4, serão apresentadas as *medidas de dispersão*.

Ordenamento de dados

Em algumas situações, o objetivo é conhecer a *posição* de um determinado valor numérico em relação aos demais valores da amostra; por exemplo, qual a posição de um determinado candidato a *trainee* comparando seu *QI* com os *QIs* dos outros candidatos que concorrem? O *QI* desse candidato é baixo ou alto? Quantos candidatos têm *QI* maior do que o candidato sob análise? Ou, quão maior é o *QI* do candidato? Outro exemplo, o retorno de 15% ao ano é baixo ou alto quando comparado com as rentabilidades das aplicações do mercado financeiro durante o mesmo período? Quantos retornos do mercado financeiro são maiores do que 15%?

Para responder a perguntas desse tipo, primeiro, os valores da série de dados devem estar ordenados em ordem crescente ou decrescente. Depois, deve-se estabelecer um critério que permita definir a posição de um determinado valor da série dentro da própria série de valores numéricos.

EXEMPLO 3.1

Ordene de forma crescente os valores da amostra registrada na tabela a seguir:

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

Solução. Depois de ordenar de forma crescente os onze valores numéricos da amostra, a seguir são associados os números 1, 2, ..., 11 aos valores ordenados como mostra esta tabela:

Amostra	15	18	19	24	27	31	32	38	39	42	43
Ordem	1	2	3	4	5	6	7	8	9	10	11

Agora, o valor 15 tem a posição 1, o 19 a posição 3 e o 43 a posição final 11.

De forma geral, o Exemplo 3.1 mostra que os n valores numéricos de uma amostra ordenada de forma crescente foram associados à série dos números naturais 1, 2, 3, ... até n . Foi estabelecida uma relação de ordem entre os valores numéricos da amostra.

EXEMPLO 3.2

Determine a *ordem* de cada valor da amostra seguinte:

27	32	64	65	58	62	59	54	29	30	26	48	47
46	43	38	29	32	35	37	31	43	45	42	37	36

Solução. Depois de ordenar os valores da amostra de forma crescente, foi associada a série de números 1, 2, ..., 26 aos valores como mostra a tabela seguinte.

Amostra	26	27	29	29	30	31	32	32	35	36	37	37	38
Ordem	1	2	3	4	5	6	7	8	9	10	11	12	13
Amostra	42	43	43	45	46	47	48	54	58	59	62	64	65
Ordem	14	15	16	17	18	19	20	21	22	23	24	25	26

O procedimento de ordenamento em ordem crescente utilizado no Exemplo 3.2 foi o mesmo que o do Exemplo 3.1. No primeiro exemplo, o trabalho manual foi facilitado pelo pequeno tamanho da amostra. No último exemplo, o ordenamento manual é menos eficiente, pois é mais trabalhoso e está sujeito a erro de seleção dos valores da amostra. O comando de classificação do Excel ajudará a ordenar séries de valores em ordem crescente ou decrescente.

EXEMPLO 3.3



Ordene de forma crescente os dados do Exemplo 3.2 utilizando o Excel.

Solução. Primeiro, os dados da amostra do Exemplo 3.2 foram registrados na coluna B da planilha **Exemplo 3.3**, incluída na pasta **Capítulo 3**. A seguir, o intervalo B4:B30 foi copiado no intervalo C4:C30, adicionando o título **Amostra ordenada** como se pode ver na figura a seguir. O ordenamento dos valores da amostra pode ser realizado na própria coluna B; entretanto, a amostra foi copiada na coluna C para manter a amostra inicial e destacar o procedimento de ordenamento do Excel.

	A	B	C	D	E	F	G	H	
1	Exemplo 3.3								
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									

	A	B	C
1	Exemplo 3.3		
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Para ordenar a amostra da coluna C procedemos assim:

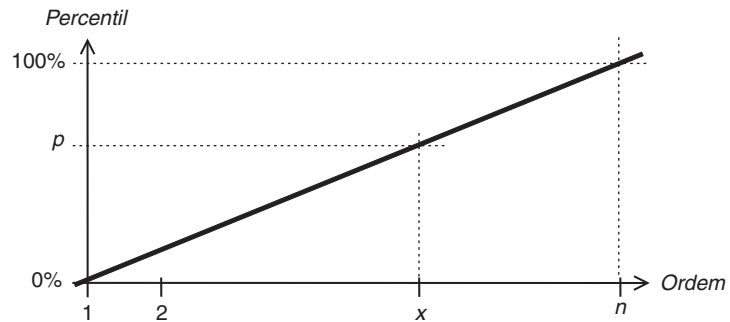
- Selecione o intervalo C4:C30, incluindo o título **Ordenada** da célula C4.
- Depois de escolher **Classificar** no menu **Dados**, o Excel apresentará a caixa de diálogo que detecta outros valores ao redor do intervalo selecionado, mostrando duas opções **Expandir a seleção** e **Continuar com a seleção atual**; selecione esta última opção e depois pressione o botão **Classificar...**
 - Em vez de utilizar o menu **Dados – Classificar**, é possível pressionar o ícone  para classificar em ordem crescente, e o ícone  para classificar em ordem decrescente.
- Em sequência, o Excel exibirá a caixa de diálogo **Classificar** com as seguintes escolhas: no grupo **Minha lista tem** a opção **Linha de cabeçalho**, na caixa **Classificar por** foi selecionado **Ordenada**, a opção **Crescente** e o intervalo C5:C30 estará selecionado, como mostra a figura à esquerda. Verifique que a célula C4 foi retirada da seleção do intervalo, pois informamos que o intervalo C4:C30 contém uma linha de cabeçalho. Essas escolhas estão de acordo com o intervalo da amostra informado.
- Como teste, se no grupo **Minha lista tem** for selecionada a opção **Nenhuma linha de cabeçalho**, mantendo a opção **Crescente**, na caixa **Classificar por** aparecerá **Coluna C** e o intervalo C4:C30 estará selecionado. Nesse caso, a célula C4 foi incluída na seleção do intervalo, pois informamos que o intervalo C4:C30 não contém uma linha de cabeçalho.

Depois de pressionar o botão **OK**, os valores da amostra são ordenados de forma crescente no mesmo intervalo C5:C30 da planilha, como mostra a figura à direita. Para obter mais informações sobre o comando classificar, na ajuda do Excel, procure *Classificar uma lista*, onde encontrará suporte para realizar classificações em mais de uma coluna, classificando valores numéricos ou nomes e semelhantes na ordem crescente (A até Z ou 0 até 9) ou ordem decrescente (Z até A) ou (9 até 0).

Percentil

Os Exemplos 3.1 e 3.2 mostram o mesmo procedimento de ordenamento para duas listas de valores numéricos com quantidade de valores diferentes, sendo que há amostras com quantidades maiores de dados. É conveniente dispor de um procedimento que, mantendo o ordenamento crescente dos dados da amostra e a associação com os números naturais, tenha uma mesma medida e permita realizar comparações. A Figura 3.1 mostra uma relação entre a série de números naturais 0, 1, 2, ... n no eixo de abscissas com uma escala de 0% a 100% no eixo de ordenadas, sendo que 0% corresponde ao primeiro dado da amostra ordenada de forma crescente, e 100% ao último dado da amostra ordenada.

FIGURA 3.1 Ordenamento dos n valores de uma amostra.



Os valores da escala de ordenadas são denominados *percentil*, sendo que o menor valor do percentil é 0% e o maior valor 100%; dessa maneira, qualquer dado da amostra estará sempre entre o percentil 0% e 100%, como se pode ver na Figura 3.1, na qual o valor com *ordem* x corresponde ao percentil p . A relação entre as ordens dos n dados da amostra ou variável e todos os valores de percentil entre 0% a 100% é regida pela seguinte relação geométrica:

$$\frac{n-1}{100\% - 0\%} = \frac{x-1}{p-0\%}$$

Nessa relação, n é a quantidade de dados da amostra, x é a ordem de um determinado dado da amostra ordenada de forma crescente, e p é o percentil correspondente em porcentagem. Dessa relação, obtemos as fórmulas de p e x .

- O percentil p em porcentagem do dado da amostra ou variável com ordem x é obtido com a fórmula: $p = \frac{x-1}{n-1} \times 100\%$. Qual é o significado do resultado p ? O dado de ordem x é maior do que os primeiros p dados da amostra e, ao mesmo tempo, menor do que os restantes $(1-p)$ dados da amostra.
- Da mesma maneira, conhecido o percentil p de um dado da amostra, sua ordem x é calculada com a fórmula: $x = (n-1) \times \frac{p}{100} + 1$.

Resumindo, agora dispomos de uma relação entre uma escala de 0% a 100% (eixo de ordenadas) e a série de números naturais 0, 1, 2, ... n que representam uma série de dados quantitativos ou amostra ordenada de forma crescente (eixo de abscissas), sendo que 0% (percentil 0%) corresponde ao primeiro dado da amostra, e 100% (percentil 100%) corresponde ao último dado da amostra.

EXEMPLO 3.4

Calcule o percentil dos dados da amostra do Exemplo 3.1.

Solução. A partir da ordem de cada dado da amostra do Exemplo 3.1 foi calculado o percentil correspondente. Por exemplo, o dado 18 tem ordem $x=2$ e percentil $p=10\%$, resultado obtido com a fórmula:

$$p = \frac{x-1}{n-1} \times 100\%$$

$$p = \frac{2-1}{11-1} \times 100\% = 10\%$$

Repetindo esse procedimento de cálculo, foi construída a tabela a seguir:

Amostra	15	18	19	24	27	31	32	38	39	42	43
Percentil	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

O percentil do dado 32 do Exemplo 3.4 é 60%. Qual o significado do resultado $p=60\%$? O percentil 60% significa que o dado ordenado 32 é maior do que os primeiros 60% dos dados ordenados de forma crescente da amostra e, ao mesmo tempo, menor do que os demais 40% dos dados da amostra. Sem dúvida que a quantidade exata de dados da amostra do Exemplo 3.1 facilitou o cálculo do percentil de cada dado, pois é um múltiplo de 10%.

EXEMPLO 3.5

Determine a ordem do percentil 10%, 50% e 80% da amostra do Exemplo 3.1.

Solução. Para $p=50\%$, obtemos a ordem $x=6$ como resultado da fórmula:

$$x = (n - 1) \times \frac{p}{100} + 1$$

$$x = (11 - 1) \times \frac{50}{100} + 1 = 6$$

Portanto, consultando a tabela de dados ordenados do Exemplo 3.1, a posição 6 está ocupada pelo valor 31. Continuando com o exemplo:

- Para $p=10\%$, a ordem é $x=2$, que se refere ao valor 18.
- Para $p=80\%$, a ordem é $x=9$, que se refere ao valor 39.

Tenha em mente que há diversas formas de relacionar um conjunto de dados ordenados de forma crescente com o respectivo percentil. A forma apresentada é a utilizada pelas funções estatísticas do Excel.

EXEMPLO 3.6

Determine a ordem dos dados da amostra do Exemplo 3.2, depois, para cada ordem, calcule o percentil correspondente e, por último e a partir desse resultado, obtenha a ordem utilizando o Excel e as fórmulas apresentadas.

Solução. Primeiro foi feita uma cópia da planilha **Exemplo 3.3** que recebeu o nome **Exemplo 3.6**. A seguir:

- Na coluna D, foi registrada a ordem de cada dado ordenado da coluna C, do número um até o 26. Esse preenchimento pode ser realizado de duas formas:
 - Registre os números 1 e 2, respectivamente, nas células D5 e D6. Depois, com o mouse, selecione as duas células e arraste a alça de preenchimento das células selecionadas até a célula D30. Essa alternativa pode provocar mudanças das formatações de células que receberão a cópia dos valores.
 - A alternativa é a seguinte: registrar o número 1 na célula D5, no menu **Editar**, selecionar **Preencher** e, a seguir, **Sequência** que apresentará a caixa de diálogo **Sequência**, cuja figura é mostrada a seguir depois de preencher os dados necessários para registrar os números 1 a 26. Depois de pressionar **OK**, esse comando preenche os valores solicitados. Essa alternativa também pode provocar mudanças das formatações de células que receberão a cópia dos valores.



- Continuando, na célula E5, foi registrada a fórmula $= (D5-1)/(\$D\$30-1)$ que calcula o percentil do dado da amostra com ordem igual a um. Depois, essa fórmula foi copiada até a célula E30, completando o cálculo do percentil da ordem dos dados restantes da amostra. Na coluna F, foi calculada a ordem de cada percentil registrado na coluna E. Na célula F5, foi registrada a fórmula: $= (\$D\$30-1)*E5+1$, que, depois, foi copiada até a célula F30. A próxima figura mostra a **Planilha 3.6** depois de completar o registro das fórmulas.

	A	B	C	D	E	F
1	Exemplo 3.6					
2						
3						
4		Amostra	Amostra Ordenada	Ordem	p	x
5		27	26	1	0,00%	1
6		32	27	2	4,00%	2
7		64	29	3	8,00%	3
8		65	29	4	12,00%	4
9		58	30	5	16,00%	5
10		62	31	6	20,00%	6
11		59	32	7	24,00%	7

EXEMPLO 3.7

Continuando com os dados e resultados do Exemplo 3.2, quais os dados da amostra com percentil 50% e 77%?

Solução. Para o percentil $p=50\%$, obtemos a ordem $x=13,50$, resultado obtido com a fórmula:

$$x = (26 - 1) \times \frac{50}{100} + 1 = 13,50$$

Na tabela do Exemplo 3.2 ou do Exemplo 3.6 ou na planilha Excel correspondente, observa-se que não há ordem 13,50. Entretanto, tendo presente que na definição de percentil foi estabelecida uma relação linear com a ordem, é possível realizar uma interpolação linear entre as ordens definidas. Dessa maneira, se para $x=13$ o dado da amostra é 38 e para $x=14$ é 42, a ordem $x=13,50$ corresponderá ao dado $40=38+(42-38)\times 0,50$, valor que também não pertence à amostra. Com o mesmo procedimento, para o percentil $p=77\%$, obtém-se a ordem $x=20,25$ e o dado correspondente $49,50=48+(54-48)\times 0,25$, que também não pertence à amostra.

EXEMPLO 3.8

Os retornos acumulados nos últimos doze meses dos primeiros vinte fundos de investimento estão registrados em ordem crescente na segunda coluna da tabela da figura a seguir. Nessa tabela, foi adicionada uma coluna com a ordem dos retornos, de um a vinte. Calcule o percentil de cada retorno e, a partir dele, verifique a ordem desse retorno utilizando funções estatísticas do Excel.

Solução. As funções estatísticas ORDEM.PERCENTUAL e PERCENTIL do Excel retornam, respectivamente, o percentil e a ordem. Começamos por conhecer as sintaxes dessas duas funções

- **ORDEM.PORCENTUAL(matriz; valor; núm_ decimais)**

A função estatística ORDEM.PORCENTUAL¹ retorna o percentil do argumento *valor*, considerando a *matriz* ordenada de forma crescente. Se a *matriz* tiver valores repetidos, a função informará o percentil do primeiro valor que encontrar. O argumento *núm_decimais* define o número de casas decimais do resultado; se omitido, o resultado terá três casas decimais. Perceba que não será necessário ordenar previamente os dados da amostra, pois a função ORDEM.PORCENTUAL ordena os dados da amostra de forma crescente antes de calcular.

¹ Em inglês, a função ORDEM.PORCENTUAL é PERCENTRANK.

A fórmula =ORDEM.PORCENTUAL(\$C\$4:\$C\$23;C4;6) foi registrada na célula D4 e depois copiada até a célula D23. Agora, no intervalo D4:D23 está registrado o percentil de cada retorno do intervalo C4:C23. Os cifrões no intervalo da matriz foram adicionados para poder copiar essa fórmula até o último dado da amostra, e o número seis de casas decimais foi para comparar esses resultados. A função ORDEM.PORCENTUAL também pode ser registrada como matriz em uma coluna da planilha:

- Selecione o intervalo G4:G23.
- Digite a fórmula =ORDEM.PORCENTUAL(C4:C23;C4:C23;6) sem pressionar a tecla Enter.
- Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**; mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, obtemos os mesmos resultados do intervalo D4:D23 no qual as fórmulas receberam as chaves { }. As fórmulas matriciais não utilizam cifrões e ocupam menos memória da unidade de processamento comparada a com o registro individual de fórmulas.

	A	B	C	D	E
1	Retorno anual de fundos de investimento				
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					

• PERCENTIL(matriz; k)

A função estatística PERCENTIL² retorna o valor que divide a *matriz* em duas partes, uma menor do que o argumento *k* e a outra maior do que *k*. O argumento *k* é um valor entre 0 e 1, correspondendo respectivamente a 0% e 100% da quantidade de dados da *matriz*. Observe que não será necessário ordenar previamente os dados da amostra, pois a função PERCENTIL ordenará os dados da amostra de forma crescente antes de calcular. Nem sempre o resultado da função percentil é um valor da amostra. Por exemplo, o valor correspondente ao percentil 75% da amostra do Exemplo 3.1 é 38,50, resultado obtido por interpolação linear a partir da relação linear entre a ordem e o percentil de cada valor da amostra, como vimos no Exemplo 3.7.

A fórmula =PERCENTIL(\$C\$4:\$C\$23;D4) foi registrada na célula E4 e depois copiada até a célula E23. Agora, no intervalo E4:E23 está registrado o retorno do percentil registrado no intervalo D4:D23. Os cifrões no intervalo da matriz foram adicionados para poder copiar essa fórmula até o último dado da amostra. A função PERCENTIL pode ser também registrada como matriz em uma coluna da planilha:

- Selecione o intervalo H4:H23.
- Digite a fórmula =PERCENTIL(C4:C23;D4:D23) sem pressionar a tecla Enter.

² Em inglês, a função PERCENTIL é PERCENTILE.

- Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**. Depois de pressionar as três teclas simultaneamente, obtemos os mesmos resultados do intervalo E4:E23, no qual as fórmulas receberam as chaves { }.

Outras funções estatísticas relacionadas com esse tema podem ser encontradas no Apêndice 1 deste capítulo.

Se o administrador de um fundo equivalente não listado na tabela afirma que o retorno acumulado nos últimos doze meses de seu fundo foi 30,2%, então seu percentil é $p=57,9\%$ e, conseqüentemente, o retorno do seu fundo é maior do que 57,9% dos primeiros fundos da tabela e menor do que os 42,1% dos demais fundos. Observe que um fundo com retorno de 32,52% tem percentil 80%; dessa maneira, o retorno desse fundo é maior do que 80% dos fundos da amostra e menor do que os restantes 20% dos fundos com seus retornos ordenados de forma crescente. Note que, enquanto o percentil 80% é uma medida relativa, pois somente avalia o desempenho do fundo em relação aos outros fundos, o retorno do fundo de 32,52% é uma medida absoluta. O ordenamento com percentil não representa uma escala intervalar constante, pois trata apenas com posições de valores ordenados.

Quartil

Na relação entre a escala de 0% a 100% e a série de números naturais 0, 1, 2, ... n que representam uma série de dados de uma amostra ordenada de forma crescente, o primeiro dado da amostra é o percentil 0%, e o último dado da amostra é o percentil 100%. Também há outras formas de definir referências fixas, por exemplo, cada 10% ou *decil*, ou cada 12,5% ou *octil*, ou cada 25% ou *quartil* que será apresentado a seguir. Dividindo os valores ordenados da variável em quatro quartos iguais, obtém-se um quartil para cada quarto definido desta forma:

- O primeiro quartil Q_1 é o percentil 25%. O valor da amostra do primeiro quartil Q_1 é maior do que 25% dos valores menores e menor do que 75% dos demais valores maiores.
- O segundo quartil³ Q_2 é o percentil 50%. O valor da amostra do segundo quartil Q_2 é maior do que 50% dos valores menores e menor do que 50% dos demais valores maiores. O segundo quartil é também a *mediana* que divide a área da distribuição de frequências em duas partes iguais a 50%.
- O terceiro quartil Q_3 é o percentil 75%. O valor da amostra do terceiro quartil Q_3 é maior do que 75% dos valores menores e menor do que 25% dos demais valores maiores.

Da fórmula do percentil, obtêm-se as fórmulas dos três quartis utilizadas pelo Excel, como mostrado a seguir.

- Conhecido o percentil p de um dado da amostra ordenada, sua ordem x é calculada com a fórmula $x = (n - 1) \times \frac{p}{100} + 1$. No primeiro quartil, $p=25\%$ ou $1/4$, a fórmula passa a ser

$$x = (n - 1) \times \frac{1}{4} + 1 = \frac{n + 3}{4}$$

- A fórmula da ordem no segundo quartil $p=50\%$ é $x = \frac{n + 1}{2}$.
- A fórmula da ordem no terceiro quartil $p=75\%$ é $x = \frac{3 \times n + 1}{4}$.

Se o resultado de x não for um número inteiro, o valor do dado da amostra ou variável será obtido com interpolação linear como já apresentado.

³ A mediana divide a área da distribuição de frequências em duas partes iguais a 50%.

EXEMPLO 3.9

Calcule o primeiro, segundo e terceiro quartis dos retornos do Exemplo 3.8.

Solução. A função estatística QUARTIL do Excel retorna o valor do quartil informado. Começemos por conhecer a sintaxe dessa função.

• QUARTIL(*matriz; quarto*)

A função estatística QUARTIL⁴ retorna o dado da *matriz* ordenada correspondente ao argumento *quarto* identificado da seguinte maneira:

- Se *quarto*=0, a função retornará o primeiro ou menor valor da *matriz*.
- Se *quarto*=1, 2 ou 3, a função retornará o valor da matriz correspondente e respectivamente, ao primeiro, segundo ou terceiro quartis.
- Se *quarto*=4, a função retornará o último ou maior valor da *matriz*.

Enquanto a função QUARTIL fornece resultados de posições definidas na amostra ordenada, a função PERCENTIL dá os resultados para qualquer posição de 0 a 1, ou 0% a 100%. No entanto, nem sempre o retorno da função QUARTIL é um dado da amostra.

A próxima figura mostra o cálculo de todos os resultados da função QUARTIL utilizando os retornos dos fundos de investimento da planilha **Exemplo 3.8** a partir da linha 26.

	A	B	C	D	E	F
26	Exemplo 3.9					
27						
28		Percentil	Quartil	Retorno		
29		0%		16,300%	=QUARTIL(C4:C23;0)	
30		25%	Primeiro	24,775%	=QUARTIL(C4:C23;1)	
31		50%	Segundo	28,200%	=QUARTIL(C4:C23;2)	
32		75%	Terceiro	32,125%	=QUARTIL(C4:C23;3)	
33		100%		36,700%	=QUARTIL(C4:C23;4)	
34						

Analisemos os cinco resultados da função estatística QUARTIL, lembrando que nem sempre o retorno é um dado da amostra.

- Os resultados da função QUARTIL para o argumento *quarto* igual a zero ou quatro coincide, respectivamente, com o primeiro (menor) ou último (maior) dado da amostra ordenada.
- O retorno do primeiro quartil é 24,775%, valor que não consta na série de retornos. Nesse caso, o valor do quartil foi obtido com a interpolação linear $0,24775 = 0,2470 + (0,2480 - 0,2470) \times (0,25 - 0,21053) / (0,26316 - 0,21053)$.
- Os retornos do segundo e do terceiro quartil foram obtidos da mesma forma que o do segundo quartil.

Ferramenta de análise *Ordem e Percentil*

A partir de uma amostra quantitativa discreta registrada em uma planilha Excel, uma série de valores registrados em uma ou mais colunas contíguas, a ferramenta de análise *Ordem e percentil* retornará, a partir do endereço informado pelo usuário, uma tabela com a posição ordinal e percentual de cada dado da amostra, permitindo analisar a posição relativa dos valores em um conjunto de dados.

Para utilizar a ferramenta de análise *Histograma Ordem e Percentil*,⁵ a amostra que será analisada deve estar registrada em uma planilha como a **Ferramenta Ordem e Percentil** incluída na pasta **Capítulo 3**, sendo que:

- No intervalo B3:B29 foram registrados os valores numéricos da amostra do Exemplo 3.2, incluindo o nome *Amostra* na célula B3. Os valores da amostra podem ser registrados em uma linha, uma coluna ou combinando linhas e colunas, contanto que sejam contíguos e possíveis de identificá-los com um único intervalo.

⁴ Em inglês, a função QUARTIL é QUARTILE.

⁵ Em inglês, a ferramenta ORDEM E PERCENTIL é RANK AND PERCENTILE.

- Selecione o intervalo B3:B29.
- Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro.
- Escolhendo a ferramenta **Ordem e percentil**, depois de pressionar o botão OK, você receberá a caixa de diálogo **Ordem e percentil** mostrada na Figura 3.2, depois de selecionadas algumas opções.
 - Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Ordem e percentil* pertencente à *Ajuda do Excel*.

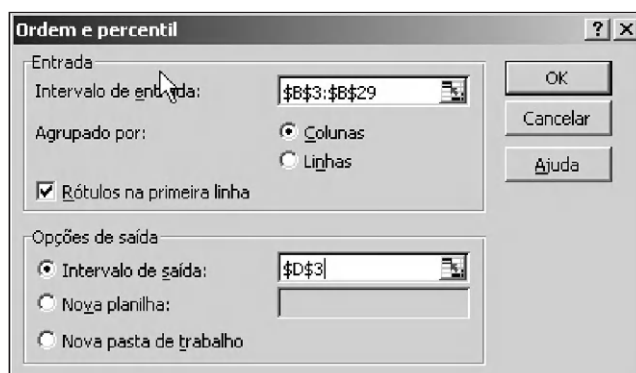


FIGURA 3.2 Caixa de diálogo da ferramenta *Ordem e percentil*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo da ferramenta *Ordem e percentil* são:

- **Intervalo de entrada.** Informe o intervalo de células da planilha no qual os dados estão registrados; nesse caso, o intervalo B3:B29 que inclui a célula onde foi registrado o título *Amostra*, ou rótulo no Excel.
- **Agrupado por.** Selecionamos **Colunas**, pois a amostra foi registrada em uma coluna. Em geral, o Excel selecionará automaticamente depois de ter informado o intervalo da amostra.
- **Rótulos na primeira linha.** Tendo escolhido **Colunas** no item anterior, necessariamente selecionaremos **Rótulos na primeira linha**, pois na primeira célula da série foi incluído o nome *Amostra*.

FIGURA 3.3
Ferramenta Ordem e Percentil resolvendo o Exemplo 3.6.

	A	B	C	D	E	F	G
1	Ferramenta de Análise ORDEM E PERCENTIL						
2							
3		Amostra		Ponto	Amostra	Ordem	Porcentagem
4		27		4	65	1	100,00%
5		32		3	64	2	96,00%
6		64		6	62	3	92,00%
7		65		7	59	4	88,00%
8		58		5	58	5	84,00%
9		62		8	54	6	80,00%
10		59		12	48	7	76,00%
11		54		13	47	8	72,00%

No quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três botões de opção que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso, D3, que é o endereço da célula superior esquerda da tabela de respostas que a ferra-

menta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes. Podem ser encontradas mais informações no Capítulo 1 ou na *Ajuda* do Excel.

- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha **Plan1**.

Depois de pressionar o botão OK, a ferramenta *Ordem e percentil* apresentará os resultados solicitados nas seleções realizadas, como mostra a Figura 3.3. A partir da célula D3 da planilha, a ferramenta registra a tabela de resultados cuja análise é realizada a seguir.

- Na coluna E (*Amostra*) da tabela, a ferramenta registrou os dados da *amostra* ordenados de forma decrescente.
- Na coluna D (*Ponto*), foi registrada a posição de cada dado da coluna E registrado na coluna B. Por exemplo, o valor 62 registrado na célula E6 tem a posição 6 (célula D6) na amostra da coluna B, ou o valor 62 é o sexto dado da amostra da coluna B, célula B9.
- Na coluna F (*Ordem*), foi registrada a ordem de cada dado da amostra registrada na coluna E da tabela. Se na amostra há valores repetidos, a classificação manterá *ordem* do primeiro valor não repetido. A *ordem* é calculada com a função estatística ORDEM, apresentada no Apêndice 1 do Capítulo 3.
- Na coluna G (*Porcentagem*), foi registrado o *percentil* de cada dado da amostra ordenada de forma decrescente. Esses valores foram calculados com a função estatística ORDEM.PORCENTUAL já apresentada.

Medidas de tendência central

No Capítulo 2, mostramos como apresentar dados numéricos de forma agrupada utilizando tabelas de frequências e histogramas. A parte inicial deste Capítulo 3 mostrou como trabalhar com as posições relativas dos dados ordenados de uma amostra utilizando percentil e quartil. Os exemplos desenvolvidos no Capítulo 2 mostram que os dados tendem a se agrupar ao redor de um ponto central, mostrando a oportunidade de definir novas medidas que podem representar toda a amostra ou variável. A *mediana* é uma das medidas de tendência central cuja definição coincide com o percentil 50%, ou o segundo quartil, de uma série de dados ordenados de forma crescente. As outras medidas de tendência central são a *moda* e a *média aritmética* ou simplesmente *média*.

Mediana

A *mediana Md* é uma medida de tendência central cuja definição coincide com o percentil 50%, ou o segundo quartil, de uma série de dados ordenados de forma crescente. A mediana *Md* é um valor localizado na posição central, tal que 50% dos valores são menores do que *Md*, e os demais 50% são maiores.

Depois de ordenar os n valores da variável de forma crescente, a *Md* é determinada de acordo com o tipo do número n :

Se n for um número ímpar, a *Md* será o valor da variável situado na posição $(n+1)/2$.

Se n for um número par, a *Md* será igual ao resultado da divisão por dois da soma dos valores das posições $(n/2)$ e $(n/2)+1$. Nesse caso, a *Md* poderá não ser um valor da variável.

Note que a quantidade de dados da amostra acima de Md é igual à quantidade de dados da amostra abaixo dele, seja n par ou ímpar. De outra maneira, a mediana Md divide a área da distribuição de frequências em duas partes iguais a 50%.

EXEMPLO 3.10

Calcule a *mediana* da amostra do Exemplo 3.1.

Solução. Para facilitar o trabalho, os dados da amostra são repetidos a seguir.

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

A tabela a seguir mostra os 11 valores da amostra ordenados de forma crescente, identificando o valor da mediana dentro de um círculo.

15	18	19	24	27	31	32	38	39	42	43
----	----	----	----	----	----	----	----	----	----	----

Como a quantidade de dados da amostra $n=11$ é um número ímpar, o valor da mediana é $Md=31$, que corresponde ao dado da posição $6=(11+1)/2$. O mesmo resultado foi obtido com a função MED do Excel, como mostra a figura a seguir, referente à planilha **Cálculo da Mediana** da pasta **Capítulo 3**.

	A	B	C	D	E	F
1	Exemplo 3.10					
2						
3		Amostra				
4		31		Mediana	31,00	=MED(B4:B14)
5		38				
6		19				
7		27				

A mediana foi obtida com a fórmula =MED(B4:B14) registrada na célula E4.

• MED(núm1; núm2; ... ; núm30)

A função estatística MED(núm1; núm2; ... ; núm30) retorna a mediana dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou semelhantes. Nesse exemplo, a amostra do intervalo B4:B14 foi registrada no primeiro argumento núm1. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.

EXEMPLO 3.11

Calcule a *mediana* da amostra do Exemplo 3.2.

Solução. Para facilitar o trabalho, os dados da amostra são repetidos a seguir.

27	32	64	65	58	62	59	54	29	30	26	48	47
46	43	38	29	32	35	37	31	43	45	42	37	36

A tabela a seguir mostra os 26 valores da amostra ordenados de forma crescente, identificando os valores que fazem parte do cálculo da mediana dentro de um círculo.

26	27	29	29	30	31	32	32	35	36	37	37	38
42	43	43	45	46	47	48	54	58	59	62	64	65

Como a quantidade de dados $n=26$ é um número par, o valor da mediana será igual ao resultado da divisão por dois da soma dos valores das posições $(n/2)=13$ e $(n/2)+1=14$. O valor da mediana é $Md=40$, resultado obtido de $(38+42)/2$. O mesmo resultado foi obtido com a função MED do Excel, como mostra a figura a seguir referente à planilha **Cálculo da Mediana** da pasta **Capítulo 3**.

	A	B	C	D	E	F	G
17	Exemplo 3.11						
18							
19		Amostra					
20		27		Mediana	40,00	=MED(B20:B45)	
21		32					
22		64					
23		65					

Analisando os resultados dos exemplos anteriores, podemos chegar a algumas conclusões interessantes:

- Na amostra do Exemplo 3.10, acima da $Md=31$, há cinco dados da amostra e, abaixo dela, também há cinco dados, e a mediana é um valor da amostra.
- Da mesma forma, na amostra do Exemplo 3.11, acima da $Md=40$, há 13 dados da amostra e, abaixo dela, também há 13 dados; entretanto, a Md não é um valor da amostra.
- A mediana divide a distribuição de frequências em duas áreas iguais, ou duas áreas com a mesma quantidade de valores ordenados da amostra ou variável ou, de outra maneira, a mediana Md divide a área da distribuição de frequências em duas partes iguais a 50%.
- Se o maior valor da amostra for duplicado, o valor Md não será alterado, pois está relacionado apenas com a ordem da série de valores. A mediana é uma medida, resistente, ela é menos sensível à presença de valores suspeitos, dados bastante diferentes da maioria dos dados coletados na mesma amostra. A eliminação de dados suspeitos não deverá afetar a mediana, o que não ocorrerá com a média que será afetada.

Moda

A tabela de frequências absolutas do Exemplo 2.1 do Capítulo 2 mostra que o número de operações diárias fechadas pelo Operador B com maior frequência da série de dados dessa amostra é 14 operações. Essa é a medida de tendência central denominada *moda* Mo , nesse exemplo $Mo=14$.

Moda é o valor da amostra ou variável que mais se repete; ou valor com mais frequência.

EXEMPLO 3.12

Calcule a moda Mo da amostra do número de operações fechadas diariamente pelo Operador B do Exemplo 2.1, cujos dados repetimos.

14	12	13	11	12	13	16	14	14	15	17	14	11
13	14	15	13	12	14	13	14	13	15	16	12	12

Solução. A tabela de frequências absolutas do Exemplo 2.1 mostra que $Mo=14$, o número de operações diárias fechadas pelo Operador B com maior frequência. O mesmo resultado foi obtido com a função **MODO** do Excel, como mostra a figura a seguir referente à planilha **Cálculo da Moda** da pasta **Capítulo 3**.

	A	B	C	D	E	F	
1	Exemplo 3.12						
2							
3		Amostra					
4		14	14	14	13		
5		12	14	15	15		
6		13	15	13	16		
7		11	17	12	12		
8		12	14	14	12		
9		13	11	13			
10		16	13	14			
11							
12		Moda	14,00	=MODO(B4:B10;C4:C10;D4:D10;E4:E8)			
13							

O valor da moda foi obtido com a fórmula registrada na célula C12=**MODO**(B4:B10;C4:C10;D4:D10;E4:E8).

• **MODO(núm1; núm2; ... ; núm30)**

A função estatística **MODO**(núm1; núm2; ... ; núm30) retorna a moda dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Nesse exemplo, a amostra foi registrada nos quatro primeiros argumentos núm1, núm2, núm3 e núm4. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.

EXEMPLO 3.13

Determine a moda Mo da amostra do Exemplo 3.2.

Solução. Para facilitar a determinação da moda, os dados ordenados de forma crescente da amostra são repetidos e identificados a seguir.

26	27	29	29	30	31	32	32	35	36	37	37	38
42	43	43	45	46	47	48	54	58	59	62	64	65

Na amostra da tabela apresentada detectamos quatro modas, com dois dados cada uma com áreas pintadas, $Mo=29, 32, 37$ e 43 . O resultado obtido com a função **MODO** do Excel na planilha **Cálculo da Moda** da pasta **Capítulo 3** é 32.

As amostras ou variáveis com valores quantitativos contínuos costumam não apresentar moda; por exemplo, a série das 50 maiores empresas privadas por venda mostrada no Capítulo 1 não tem moda. A amostra ou variável com uma única *moda* é denominada *unimodal*, com duas modas é *bimodal* etc. A *moda* também é uma medida resistente, pois está relacionada apenas com a frequência de um ou mais dados da amostra. Por exemplo, a mudança de um dado da amostra poderá não afetar a moda Mo .

Média

A medida de posição mais utilizada é a *média aritmética* ou simplesmente *média* de uma amostra ou variável.

Média \bar{X} é o resultado da divisão da soma dos valores das observações ou dados $X_1, X_2, \dots, X_i, \dots, X_n$ da amostra X pela quantidade de dados n :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

As características importantes da *média* são:

- A unidade de medida da média é a mesma que a dos valores da amostra.
- O resultado da multiplicação da média \bar{X} pela quantidade n de valores da amostra X é igual à soma dos n valores da amostra.

No Apêndice 3 você encontra informações e como utilizar o símbolo somatória Σ .

EXEMPLO 3.14

Calcule a média da amostra do Exemplo 3.1.

Solução. A média da amostra é igual a \bar{X} , resultado obtido com a fórmula e também resolvido na planilha **Cálculo da Média** da pasta **Capítulo 3**.

	A	B	C	D	E	F	G	H
1	Exemplo 3.14							
2								
3		Amostra						
4		31						
5		38						
6		19						
7		27						
8		24						
9		42						
10		32						
11		18						
12		43						
13		15						
14		39						
15								

$$\bar{X} = \frac{\sum_{i=1}^{11} X_i}{11} = \frac{31 + 38 + \dots + 39}{11} = 29,82$$

Média	29,82	=SOMA(B4:B14)/CONT.NÚM(B4:B14)
Média	29,82	=MÉDIA(B4:B14)

O cálculo da média da amostra é realizado de três formas diferentes.

- De forma manual, utilizando a fórmula que define a média da amostra.
- Com funções do Excel equivalentes à fórmula que define a média da amostra utilizando a fórmula =SOMA(B4:B14)/CONT.NÚM(B4:B14) registrada na célula E13.
 - **SOMA(núm1; núm2; ... ; núm30)**

A função matemática SOMA(núm1; núm2; ... ; núm30) retorna a soma dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.
 - Com a função estatística MÉDIA do Excel utilizando a fórmula =MÉDIA(B4:B14) registrada na célula E14.
 - **MÉDIA(núm1; núm2; ... ; núm30)**

A função estatística MÉDIA(núm1; núm2; ... ; núm30) retorna a média aritmética dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Nesse exemplo, a amostra do intervalo B4:B14 foi registrada no primeiro argumento *núm1*. Se o nome da função MÉDIA for inserido com letras minúsculas ou maiúscu-

las ou sem o acento ortográfico, o Excel aceitará e registrará a função com letras maiúsculas e com o acento ortográfico. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.

EXEMPLO 3.15

Calcule a média da amostra de operações diárias fechadas pelo Operador B e explicar seu significado, Exemplo 2.1.

Solução. Aplicando a definição de média da população temos o resultado obtido com a seguinte fórmula.

$$\bar{X} = \frac{1}{26} \sum_{i=1}^{26} X_i = \frac{1}{26} \times (14 + 12 + \dots + 12) = \frac{352}{26}$$

$$\bar{X} = 13,54$$

Qual o significado da média igual a 13,54?

- A média tem a mesma unidade de medida que os valores da amostra.
- A média 13,54 é a quantidade equivalente de operações fechadas diariamente pelo operador B, pois o resultado da multiplicação da média pelo número 26 é igual a 352, a soma dos 26 valores da variável.

Propriedades da média

A média é a medida de posição mais utilizada porque tem propriedades importantes, como as que serão apresentadas. Para mostrar essas propriedades, precisamos utilizar algumas expressões matemáticas. Suponha uma amostra ou variável X com n dados ou observações, não necessariamente ordenados, e identificados pela sequência de valores $X_1, X_2, \dots, X_i, \dots, X_n$, onde X_1 é o primeiro dado, X_2 é o segundo dado, X_i é um dado qualquer da amostra, e assim sucessivamente até o último dado X_n . Denomina-se *desvio de um dado* X_i de uma amostra o resultado da diferença entre X_i e a média \bar{X} da amostra X . Em termos matemáticos $= X_i - \bar{X}$.

Primeira propriedade

A soma dos desvios de uma amostra ou variável é sempre igual a zero.

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

Essa propriedade é útil para verificar ou confirmar o resultado do cálculo da média de uma amostra ou variável, como também no desenvolvimento de provas matemáticas que apresentam a soma de desvios com relação à média. A primeira propriedade da média também pode ser utilizada para determinar a média de uma amostra, como mostra o Exemplo 3.16.

EXEMPLO 3.16

Determine o valor da média da amostra do Exemplo 3.1 aplicando a primeira propriedade da média e utilizando o Excel.

Solução. No intervalo B3:B14 da planilha **Média com Atingir Meta**, incluída na pasta **Capítulo 3**, foi registrada a amostra do Exemplo 3.1. Depois foram adicionados os registros mostrados na próxima figura.

	A	B	C	D	E
1	Cálculo da média utilizando a primeira propriedade				
2					
3		Amostra	Desvios		Determinação da Média
4		31	21,00		10,000
5		38	28,00		
6		19	9,00		
7		27	17,00		
8		24	14,00		
9		42	32,00		
10		32	22,00		
11		18	8,00		
12		43	33,00		
13		15	5,00		
14		39	29,00		
15		Soma	218,0000		
16					

- Na célula E5 será determinado o valor da média da amostra.
- Na célula C4 foi registrada a fórmula =B4-\$E\$5 que calcula o desvio do dado da amostra registrado na célula B4 com relação à média registrada na célula E5. Depois essa fórmula foi copiada até a célula C14.
- Na célula C15 foi registrada a fórmula =SOMA(C4:C14) que retorna a soma de todos os desvios.

Pela primeira propriedade da média, verificamos que o valor 10 registrado na célula E5 não é o valor da média da amostra, pois a soma dos desvios é diferente de zero. Da forma como foi preparada a planilha, poderemos encontrar o valor da média de forma manual, registrando diferentes valores na célula E5 até conseguir zerar o valor da célula E5, procedimento trabalhoso e cansativo. Essa resposta pode ser encontrada rapidamente utilizando o comando *Atingir Meta* da seguinte forma:

- Posicione o cursor do Excel na célula C15.
- No menu **Ferramentas** do Excel, selecione **Atingir meta**. Será exibida a caixa de diálogo **Atingir meta**.
- Nessa caixa de diálogo, informe os dados, como mostra a figura a seguir.
 - **Definir célula.** Nessa caixa é registrado o endereço da célula que contém a fórmula cujo resultado será definido na caixa seguinte. Posicionando o cursor do Excel na célula C15, nessa caixa aparecerá esse endereço. A célula C15 deve obrigatoriamente conter uma fórmula.
 - **Para valor.** Nessa caixa, registramos o resultado desejado na célula C15 endereço definido em *Definir célula*, nesse caso o valor 0. Para acessar a caixa **Para valor**, basta pressionar a tecla **Tab** ou clicar na caixa.
 - **Alternando célula.** Nessa caixa é registrado o endereço da célula que deverá ser alterada para que a célula C15 atinja o valor desejado 0, ou o endereço da célula que contém o valor que se deseja ajustar. Esse dado pode ser registrado, depois de posicionar o cursor nesta caixa, clicando na própria célula E5, ou digitando o endereço da célula E5 na própria caixa.
- Depois de completar as informações, clique em **OK**, e o comando Atingir Meta inicia o processo de busca da

	A	B	C	D	E
1	Cálculo da média utilizando a primeira propriedade				
2					
3		Amostra	Desvios		Determinação da Média
4		31	21,00		10,000
5		38	28,00		
6		19	9,00		
7		27	17,00		
8		24	14,00		
9		42	32,00		
10		32	22,00		
11		18	8,00		
12		43	33,00		
13		15	5,00		
14		39	29,00		
15		Soma	218,0000		
16					

Atingir meta

Definir célula:

Para valor:

Alternando célula:

OK Cancelar

solução desejada. Concluído o processo de busca, o Excel apresentará a caixa de diálogo **Status do comando atingir meta**, informando que foi encontrada uma solução, o *Valor de destino* 0 registrado na caixa **Para valor** e o *Valor atual* encontrado na célula C15.

	A	B	C	D	E
1	Cálculo da média utilizando a primeira propriedade				
2					
3		Amostra	Desvios		Determinação da Média
4		31	1,18		29,818
5		38	8,18		
6		19	-10,82		
7		27	-2,82		
8		24	-5,82		
9		42	12,18		
10		32	2,18		
11		18	-11,82		
12		43	13,18		
13		15	-14,82		
14		39	9,18		
15		Soma	0,0000		
16					

Status do comando atingir meta	
Atingir Meta com a célula C15 encontrou uma solução.	<input type="button" value="OK"/>
Valor de destino: 0	<input type="button" value="Cancelar"/>
Valor atual: 0,0000	<input type="button" value="Etapa"/>
	<input type="button" value="Pausar"/>

Segunda propriedade


A soma dos quadrados dos desvios com relação à própria média de uma variável ou amostra é sempre um valor mínimo.

$$\sum_{i=1}^n (X_i - \bar{X})^2 \Rightarrow \text{mínimo}$$

No Capítulo 4 será mostrado como medir a variabilidade dos dados de uma amostra utilizando os desvios dos dados com relação à média, onde a soma dos quadrados dos desvios é utilizada na definição de variância.

Visualização das propriedades

No caminho ficou a pergunta: qual é o significado de *mínimo*? A resposta está, inicialmente, na própria declaração da propriedade. Que a soma dos quadrados dos desvios com relação à média da própria variável ou amostra seja um valor mínimo significa que se os desvios fossem calculados com relação a qualquer outro valor diferente da média da amostra, a nova soma dos quadrados dos desvios seria maior do que a primeira. Demonstra-se que somente a própria média da amostra ou variável satisfaz à condição de mínimo, como se pode ver no Apêndice 3 deste capítulo. Também há a possibilidade de compreender essa propriedade de forma visual com a planilha **Visualização Propriedades** incluída na pasta **Capítulo 3**, como mostra a Figura 3.4, utilizando a amostra do Exemplo 3.1.

- No intervalo B5:B15 foi registrada a amostra do Exemplo 3.1.
- Na célula D26 foi calculada e registrada a verdadeira média da amostra utilizando a função estatística MÉDIA.
- No intervalo D21:D24, foi incluído o controle giratório , recurso disponível no Excel para aumentar ou diminuir o valor da célula D17, as possíveis médias da amostra. Para aumentar o valor do parâmetro da célula D17, clique na seta para cima do controle e, para diminuir, clique na seta para baixo.
- No intervalo C5:C15, foram calculados os desvios dos valores da amostra com relação ao valor registrado na célula D17. A soma dos desvios foi registrada na célula D18.

- No intervalo D5:D15 foram calculados os quadrados dos desvios cuja soma foi registrada na célula D19.

Na planilha, foram construídos dois gráficos que permitem visualizar o que ocorre quando informamos valores arbitrários da média da amostra. No primeiro gráfico, *Ajuste manual do valor da Média*, clicando na seta para cima ou na seta para baixo do controle giratório, a reta se desloca no sentido vertical do gráfico. Esse ajuste pode ser visualizado de duas formas:

- O primeiro procedimento é comparar os comprimentos das retas tracejadas verticais entre os pontos acima e os pontos abaixo da reta horizontal, que representa a possível média da amostra. Considerando positivos os comprimentos dos valores situados acima da reta horizontal, e negativos os valores abaixo da mesma reta, a soma desses comprimentos tem de ser igual a zero, de acordo com a primeira propriedade.
- O segundo procedimento é acompanhar a variação do valor da soma dos quadrados dos desvios registrada na célula D19.

O segundo gráfico, *Soma quadrado dos desvios Vs. Médias*, mostra a parábola dos valores da soma dos quadrados dos desvios para diversos valores arbitrários da média. O valor de média registrado na célula

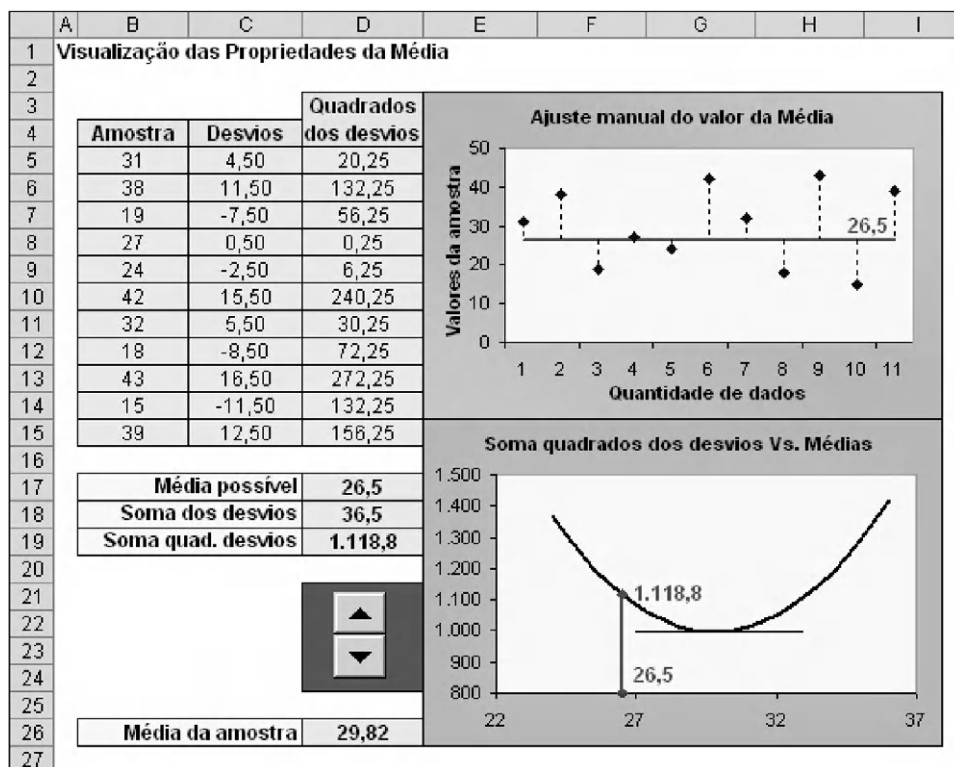


FIGURA 3.4
Visualização das propriedades da média.

D17 é destacado nessa parábola, facilitando a compreensão do procedimento de procura do mínimo. Resumindo, ao clicar na seta para cima ou na seta para baixo do controle giratório, um novo valor arbitrário de média é registrado, a reta do primeiro gráfico se desloca na vertical, o ponto que representa o novo valor arbitrário de média se desloca na parábola e os valores dos desvios mudam, intervalo D18:D19. Uma reta horizontal de espessura fina localizada na parte inferior da parábola é a tangente à curva no ponto de mínimo.

Análise do resultado da média

Analisando o procedimento de cálculo da média, pode-se concluir que:

- Todos os valores da variável são incluídos no cálculo da média.
- A média é um valor único.
- A média está posicionada de forma equilibrada entre os valores ordenados da amostra. De outra maneira, os valores da amostra se distribuem ao redor da média. Os gráficos da planilha **Visualização Propriedades** ajudam a compreender o que descrevemos.
- A média não é uma medida resistente, como a mediana ou a moda, pois ela é sensível à presença de dados suspeitos ou extremos; dados com valores bastante diferentes da maioria dos dados coletados na mesma amostra. Nesse caso, a média será uma medida distorcida da tendência dos valores da amostra, como mostra o Exemplo 3.17. Ao mesmo tempo, a eliminação de dados suspeitos deverá também afetar a média.
- Nas amostras ou variáveis com histograma simétrico, os valores da mediana, a moda e a média, coincidem, seus valores são iguais. Sugerimos que você tenha em mente essa representação ao analisar a formação da média e as variações ou dispersões dos valores da variável ao redor da média, tema que será apresentado no Capítulo 4.

Você deve ter percebido que alguns termos foram utilizados como sinônimos, ou quase. Por exemplo, dados e observações, amostra e variável etc. Poucas vezes nos referimos à amostra e à população como sinônimos, embora o procedimento de cálculo e o resultado da média, e apenas ela, sejam os mesmos. Entretanto, no caso de população e amostra deve-se manter essa separação para identificar a origem das variáveis, pois:

- *Parâmetros* são as medidas numéricas de uma população, identificados com letras gregas, μ para a média e σ para o desvio padrão (tema do próximo capítulo).
- *Estatísticas* são as medidas numéricas de uma amostra, identificadas com letras do nosso alfabeto, \bar{X} para a média e S para o desvio padrão.

Média da população X é o resultado da divisão da soma dos valores $X_1, X_2, \dots, X_i, \dots, X_n$ da variável X pela quantidade de valores N :

$$\mu_X = \frac{\sum_{i=1}^N X_i}{N} = \frac{1}{N} \sum_{i=1}^N X_i$$

EXEMPLO 3.17

A tabela a seguir registra uma amostra ordenada de 28 retornos de diversos investimentos no mesmo período. Analise a média dessa amostra e detecte dados suspeitos.

-2,1%	10,1%	10,6%	16,3%	16,3%	20,4%	21,0%
23,6%	24,7%	24,8%	26,2%	26,6%	27,0%	27,8%
28,6%	30,2%	30,3%	30,7%	32,0%	32,5%	32,6%
34,3%	35,5%	36,7%	52,9%	59,5%	76,2%	114,7%

Solução. Na planilha **Exemplo 3.17** incluída na pasta **Capítulo 3**, foram calculadas a mediana e a média dos retornos de diversos investimentos no mesmo período, respectivamente, $Md = 28,17\%$ e $\bar{X} = 32,31\%$. Analisando a série de retornos desses diversos investimentos ordenados verificamos que:

- A série de retornos tem valores extremos, por exemplo, o primeiro retorno igual a $-2,1\%$ e o último igual a $+114,7\%$. Recalculando a média sem considerar os dois valores extremos, temos $\bar{X}=30,27\%$, retorno mais próximo da mediana.
- Recalculando a média sem considerar o penúltimo valor da série $76,2\%$, temos $\bar{X}=28,44\%$, próximo da mediana.

Análise das medidas de tendência central

Embora média, mediana e moda sejam medidas importantes de tendência central por serem fáceis de serem obtidas e úteis para obter informações sobre uma amostra, elas devem ser utilizadas de acordo com a análise desejada. Analisemos, primeiro, as principais vantagens e desvantagens dessas medidas.

MODA

Vantagens	Desvantagens
Fácil de calcular.	Pode estar afastada do centro dos dados.
Não é afetada pelos dados extremos da amostra.	Difícil de incluir em funções matemáticas.
Pode ser aplicada em qualquer escala: nominal, ordinal, intervalar e proporcional.	Não utiliza todos os dados da amostra.
	A amostra pode ter mais de uma moda.
	Algumas amostras podem não ter moda.

MEDIANA

Vantagens	Desvantagens
Fácil de calcular.	Difícil de incluir em funções matemáticas.
Não é afetada pelos dados extremos da amostra.	Não utiliza todos os dados da amostra.
É um valor único.	
Pode ser aplicada nas escalas: ordinal, intervalar e proporcional.	

MÉDIA

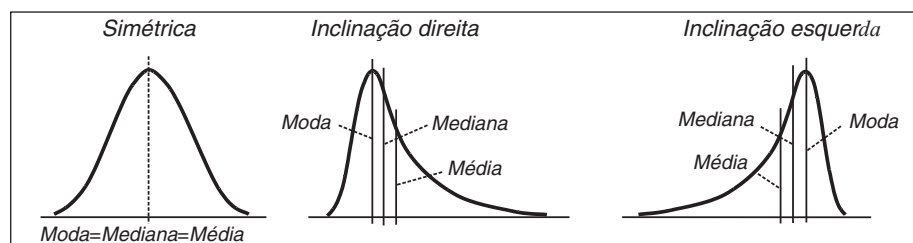
Vantagens	Desvantagens
Fácil de compreender e aplicar.	É afetada pelos dados extremos da amostra.
Utiliza todos os dados da amostra.	É necessário conhecer todos os dados da amostra.
É um valor único.	
Fácil de incluir em funções matemáticas.	
Pode ser aplicada nas escalas: intervalar e proporcional.	

Agora comparemos os valores dessas medidas em três formas diferentes do contorno de uma distribuição de frequências de uma amostra ou variável. A distribuição da esquerda da Figura 3.5 mostra uma distribuição de frequências simétrica ao redor da média. Na distribuição simétrica de frequências, os valores de média, mediana e moda coincidem. As outras duas distribuições da Figura 3.5 não são si-

métricas, e as medidas de tendência central têm posições relativas diferentes entre si, antecipando a forma da distribuição de frequências da amostra ou variável:

- Na figura do meio a distribuição tem inclinação para a direita, simplesmente *inclinação direita* ou *positiva*. A moda está na posição do pico da distribuição, e a mediana, que divide a distribuição em duas áreas iguais, situa-se à direita da moda, pois a distribuição tem inclinação para a cauda direita. Como a média é uma medida afetada pelos dados extremos da amostra, ela estará situada à direita da mediana. Utilizando os valores das medidas, verifica-se a seguinte relação $Média > Mediana > Moda$. Como nem sempre uma amostra ou variável terá moda, a análise da forma de distribuição poderá ser realizada com as outras duas medidas, $Média > Mediana$. Ou seja, se a média é maior do que a mediana, a distribuição deve ter inclinação para a direita.
- De forma equivalente, na distribuição da direita da Figura 3.5, a distribuição tem inclinação para a esquerda, simplesmente *inclinação esquerda* ou *negativa*. A moda está na posição do pico da distribuição, e a mediana, que divide a distribuição em duas áreas iguais, está situada à esquerda da moda, pois a distribuição tem inclinação para a cauda esquerda. Como a média é uma medida afetada pelos dados extremos da amostra, ela estará situada à esquerda da mediana. Utilizando os valores das medidas, verifica-se a seguinte relação $Média < Mediana < Moda$. Como nem sempre uma amostra ou variável terá moda, a análise da forma de distribuição poderá ser realizada com as outras duas medidas, $Média < Mediana$. Ou seja, se a média é menor do que a mediana, a distribuição deve ter inclinação esquerda.

FIGURA 3.5
Distribuições de frequências, simétrica e inclinada.



Qual das três medidas de tendência central utilizar? A escolha da medida depende da aplicação.

- Quando procuramos conhecer valores totais, será utilizada a média. Por exemplo, em controle de qualidade, a média é utilizada para determinar se o processo opera ao redor de um valor esperado ou alvo. Dá-se preferência à média pelas suas propriedades matemáticas.
- Se a amostra apresentar valores extremos, uma distribuição com acentuada inclinação, a mediana será mais adequada, pois não é afetada pelos dados extremos, como a média. Se quisermos conhecer o valor típico dos salários de uma determinada categoria de trabalhadores, será utilizada a mediana. Por exemplo, se os salários pesquisados da categoria são \$500, \$1.800, \$2.000, \$2.200 e \$2.500, a mediana é \$2.000 e a média \$1.800. Portanto, o valor da média tende na direção dos valores extremos e a mediana não é afetada por esses valores extremos.
- A moda é um valor típico de uma amostra ou variável. Por exemplo, na distribuição do consumo de um mesmo produto com diferentes apresentações, a moda mostra a apresentação mais consumida, como é o caso do número de calçados, o tamanho de calças etc.

Média ponderada

O cálculo da média de uma amostra é realizado com todos os dados da amostra. Todos os dados recebem a mesma importância ou o mesmo peso; eles têm uma distribuição uniforme e discreta. Contudo,

os valores repetidos poderiam ser agrupados como mostra o cálculo da média do Exemplo 3.15 que repetimos.

$$\bar{X} = \frac{1}{26} \times (2 \times 11 + 5 \times 12 + 6 \times 13 + 7 \times 14 + 3 \times 15 + 2 \times 16 + 1 \times 17)$$

Realizando a operação indicada nessa expressão:

$$\bar{X} = \frac{2}{26} \times 11 + \frac{5}{26} \times 12 + \frac{6}{26} \times 13 + \frac{7}{26} \times 14 + \frac{3}{26} \times 15 + \frac{2}{26} \times 16 + \frac{1}{26} \times 17$$

$$\bar{X} = 0,0769 \times 11 + 0,1923 \times 12 + 0,2308 \times 13 + 0,2692 \times 14 + \dots + 0,0385 \times 17$$

$$\bar{X} = 13,54$$

O agrupamento dos dados repetidos formam a *média ponderada*, que é a distribuição de frequências relativas de X , veja Exemplo 2.4 do Capítulo 2.

A *média ponderada* \bar{X} da amostra ou variável X é obtida com:

$$\bar{X} = \frac{\sum_{i=1}^n w_i \times X_i}{\sum_{i=1}^n w_i}$$

Nessa expressão, X_i é o dado repetido e w_i seu peso ou frequência.

Algumas conclusões importantes:

- O cálculo da média ponderada é um caso particular do cálculo da média aritmética.
- Os pesos formam a distribuição de frequências relativas da variável.
- No cálculo da média aritmética, a quantidade de dados da variável é conhecida; entretanto, no caso da média ponderada, a quantidade de valores da variável não é explícita.
- Uma vantagem do procedimento da média ponderada é poder definir os pesos de cada dado numa previsão, lembrando que a soma dos pesos deve ser sempre igual a um ou 100%.

EXEMPLO 3.18

O capital da empresa foi captado de três fontes, ações, financiamentos de longo prazo e debêntures, cada um com seu próprio custo definido por uma taxa anual de juros. O objetivo é calcular o custo médio ponderado do capital captado pela empresa, considerando as informações na tabela a seguir:

Capital da empresa	Participação	Taxa de juros
Acionistas	\$1.000.000	12%
Financiamentos	\$600.000	8%
Debêntures	\$400.000	14%

Solução. O capital da empresa é \$2.000.000, obtido como resultado da soma dos três capitais. O custo médio anual CM do capital da empresa é 11,20%.

$$CM = \frac{\sum_{i=1}^3 w_i \times X_i}{\sum_{i=1}^3 w_i}$$

$$CM = \frac{1.000.000 \times 12\% + 600.000 \times 8\% + 400.000 \times 14\%}{1.000.000 + 600.000 + 400.000} = 11,20\%$$

Na planilha **Exemplo 3.18**, incluída na pasta **Capítulo 3**, são apresentadas outras formas de cálculo utilizando diversos recursos do Excel.

	A	B	C	D	E
1	Exemplo 3.18				
2					
3		Pesos	Taxas		
4		50%	12%		
5		30%	8%		
6		20%	14%		
7					
8		CM	11,20%		
9					
10		Utilizando a função SOMARPRODUTO			
11		CM	11,20%		
12					
13		Utilizando a função SOMA registrada como matriz			
14		CM	11,20%		
15					
16		Utilizando o produto de matrizes			
17		Taxas / Pesos	50%	30%	20%
18		12%			
19		8%			
20		14%		CM	11,20%
21					

O resultado do custo médio de capital CM foi obtido da seguinte forma:

- Na célula C8, foi registrada a fórmula =B4*C4+B5*C5+B6*C6
- Na célula C11, foi registrada =SOMARPRODUTO(B4:B6;C4:C6)⁶.
- Na célula C14, foi registrada a fórmula =SOMA(B4:B6*C4:C6), inserida como matriz.
- A fórmula =MATRIZ.MULT(C17:E17;B18:B20)⁷ foi registrada na célula E20.

Problemas

Problema 1

Determine a quantidade de valores e os valores mínimo e máximo da amostra:

5	7	3	4	2	8	9	12
---	---	---	---	---	---	---	----

R: $n=8$, Mínimo=2 e Máximo=12

⁶ Em inglês, a função SOMARPRODUTO é SUMPRODUCT.

⁷ Em inglês, a função MATRIZ.MULT é MMULT.

Problema 2

Continuando com o Problema 1, determine a ordem e o percentil do valor 7.

R: $Ordem=5$ e $Percentil=57,1\%$

Problema 3

Continuando com o Problema 1, qual o valor da amostra com percentil 85,7%?

R: $Valor=9$

Problema 4

Continuando com o Exemplo 3.2, determine o percentil das observações cujas ordens são $x=1, 4, 10$ e 22.

R: $p=0\%, 12\%$ e 84% .

Problema 5

Continuando com o Problema 4, qual o valor da amostra com $p=32\%$?

R: $x=9$.

Problema 6

Repita os Problemas 1, 2 e 3 considerando a amostra a seguir: você escolhe o valor do segundo.

15	16	12	18	22	21	17	16
12	16	18	21	19	18	16	

Problema 7

Continuando com o Problema 6, quais os valores do primeiro quartil, do segundo quartil e do terceiro quartil?

R: $Q_1=16$; $Q_2=17$ e $Q_3=18,50$

Problema 8

Calcule os *quartis* da amostra registrada na próxima tabela.

10	15	14	23	21	18	11	12	14	15	23	12	15
----	----	----	----	----	----	----	----	----	----	----	----	----

R: $Q_1=12$ $Q_2=15$ e $Q_3=18$

Problema 9

Continuando com o Problema 8, qual o percentil do valor 15?

R: $p(15)=50\%$

Problema 10

Continuando com o Problema 8, qual o percentil dos valores 10 e 21?

R: $p(10)=0\%$ e $p(21)=83\%$

Problema 11

Continuando com o Problema 8, qual o valor com percentil 35% e 63%?

R: $X(p=35\%)=14$ e $X(p=63\%)=15$

Problema 12

A tabela a seguir registra uma amostra do número de gerentes operacionais que respondem diretamente a um diretor em empresas do ramo químico. Calcule:

- Os *quartis* da amostra.
- Quais os percentis dos valores 8 e 11?
- Quais os valores com percentis 40% e 75%?

7	7	9	8	7	13	10	14	8	9	8	6
9	9	10	11	7	8	9	6	8	11	12	10

R: a) $Q_1=7,75$ $Q_2=9$ e $Q_3=9,75$ b) $p(8)=26\%$ e $p(11)=82\%$
 c) $(p=40\%)=8$ e $X(p=75\%)=10$

Problema 13

A tabela a seguir registra os retornos das aplicações mais tradicionais do mercado financeiro. Calcule a *ordem* e o *percentil* de cada retorno.

Aplicação	Retorno mensal %
Ouro	-1,74%
Inflação	0,10%
Curto prazo	0,52%
Dólar paralelo	0,87%
CDB para <\$5.000	1,15%
Caderneta de poupança	1,16%
FRF 30 dias	1,30%
FRF 60 dias	1,49%
CDB para >\$100.000	1,58%
Bolsa RJ	2,12%
Bolsa SP	2,99%

Problema 14

Continuando com o Problema 13. No mesmo mês, o retorno do produto financeiro *FourA* foi 1,85% ao mês. Qual o percentil do retorno 1,85%? Explique o significado desse percentil.

R: O produto *FourA* tem percentil $p=83,3\%$. O retorno desse produto é maior do que os 83,3% primeiros retornos da tabela, e menor do que os 16,7% restantes.

Problema 15

Continuando com o Problema 13. Para que o gerente de produtos do *Banco* possa afirmar que o retorno de fundo *TREAL* é maior do que os 75% primeiros produtos da tabela, qual deve ser o retorno desse produto?

R: 1,54% ao mês

Problema 16

A tabela seguinte registra o salário bruto mensal dos operadores de oito empresas do mesmo ramo. Qual o percentil e o significado do salário \$1.050?

\$1.250	\$980	\$1.050	\$1.165	\$1.175	\$1.220	\$1.100	\$1.050
---------	-------	---------	---------	---------	---------	---------	---------

R: $p=14,0\%$

Problema 17

Continuando com o Problema 16. Quando Carlos reivindicou aumento de salário o chefe afirmou que nada podia fazer, pois seu salário está entre o segundo e o terceiro quartis de sua categoria. Qual deve ser o salário de Carlos?

R: O salário de Carlos está no intervalo de \$1.132,50 (Q_2) até \$1.186,30 (Q_3).

Problema 18

Calcular a média da variável do Exemplo 3.2 considerada como população.

R: $\mu=42,11$

Problema 19

Calcule a média, a moda e a mediana da amostra registrada na tabela seguinte.

10	15	14	23	21	18	11	12	14	15	23	12	18	16	15
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

R: $\bar{X}=15,62$ $Mo=15$ e $Md=15$

Problema 20

Calcule a média, a moda e a mediana das notas finais da *Turma C* da disciplina Estatística registradas na tabela a seguir.

89,5	74,7	99,4	84,9	96,5	82,1	77,7	92,7	59,1	74,7	91,0	100,0	77,6	98,5	2,2	60,8
83,1	20,1	84,2	70,1	90,8	97,5	78,2	31,7	98,1	99,0	94,3	73,4	85,7	94,1	61,0	77,8

R: $\bar{X}=78,1$ $Mo=74,7$ e $Md=83,7$

Problema 21

Calcule a média, a moda e a mediana da série de dados do Problema 13.

R: $\bar{X}=1,05\%$ $Mo=\text{Não tem}$ e $Md=1,16\%$

Problema 22

Calcule a média, a moda e a mediana dos dados da relação das 50 maiores empresas listadas na pasta Capítulo 1.

R: $\bar{X}=\$2.550,5$ $Mo=\text{Não tem}$ e $Md=\$2.119,7$

Problema 23

A tabela a seguir registra o lucro bruto em \$milhares no primeiro trimestre do ano dos vinte maiores hotéis. Calcular a média, a moda e a mediana do lucro.

619,7	475,5	356,5	338,5	336	310,5	258	223	209,7	198,4
190,5	189,3	176,9	162,4	155,5	155,5	149	143	141,9	136,6

R: $\bar{X}=\$246,3$ $Mo=\$155,5$ e $Md=\$194,5$

Problema 24

Continuando com o Problema 23, calcule os três *quartis*.

R: $Q1=\$155,5$ $Q2=\$194,5$ $Q3=\$316,9$

Problema 25

Com os resultados do Problema 23, explique a forma da distribuição do lucro bruto dos vinte maiores hotéis.

R: Como os resultados do Problema 23 verificam a condição $\mu > Md$, a distribuição do lucro tem inclinação positiva.

Problema 26

O hotel TRI não participa do grupo de hotéis do Problema 23. Se no mesmo período o lucro bruto foi igual a \$190, determine o percentil do lucro dessa empresa e explique o significado desse valor.

R: O lucro da empresa TRI tem percentil $p=45\%$; portanto, o lucro da empresa é maior do que as 45% primeiras empresas listadas em ordem crescente de lucro, e menor do que as 55% demais empresas listadas.

Problema 27

Continuando com o Problema 23. Para que seja possível afirmar que o lucro bruto de um hotel foi maior do que o lucro das 60% primeiras empresas listadas, qual deverá ser o lucro desse hotel?

R: Lucro=\$215,1 milhares

Problema 28

Mensalmente a empresa fabrica 40 lotes de 100.000 parafusos cada um. Ao escolher uma amostra aleatória de oito lotes, o controle de qualidade verificou o seguinte número de parafusos com defeito em cada lote:

Amostra	1	2	3	4	5	6	7	8
# Defeitos	300	550	480	980	1.050	350	450	870

Estime o número de parafusos com defeito em um mês de trabalho.

R: A média de defeitos é 628,75 parafusos por lote, isto é, 0,62875% de cada lote de 100.000 parafusos. Como durante um mês de produção serão produzidos 4.000.000 de parafusos, a projeção mensal do número de parafusos com defeitos será igual a 25.150 por mês.

Problema 29

A revista de negócios de maior circulação informou que os salários anuais de seus leitores têm média de \$2.200.000 e mediana \$800.000.

- Desenhe a distribuição de frequências dos salários anuais dos leitores.
- Explique a forma dessa curva.

Problema 30

Na empresa de contabilidade trabalham sete funcionários e o dono da empresa. No ano passado, o rendimento anual dos dois contadores *seniores* foi de \$60.000 cada um e dos cinco contadores *juniors* foi de \$25.000 cada um. Se o rendimento anual do dono da empresa de contabilidade foi \$255.000:

- Calcule a média, a moda e a mediana dos rendimentos anuais.
- Desenhe a curva da distribuição das frequências dos rendimentos anuais e explique sua forma.

R: Média=R\$62.500 e Mediana=R\$25.000

Problema 31

As duas tabelas seguintes registram a remuneração total dos executivos das empresas brasileiras incluindo o salário fixo, a remuneração variável e os seguintes benefícios quantificados: assistência médica, assistência odontológica, automóvel, previdência privada e alimentação.⁸

	Empresas com faturamento mensal acima de \$100 milhões					
	Presidente	Dir. financeiro	Dir. comercial	Dir. industrial	Dir. de RH	Dir. marketing
Primeiro quartil	\$30.911	\$18.973	\$14.750	\$15.084	\$13.944	\$12.703
Mediana	\$37.328	\$20.521	\$17.974	\$19.991	\$15.235	\$18.026
Terceiro quartil	\$40.538	\$21.663	\$20.116	\$20.638	\$19.118	\$18.582

	Empresas com faturamento mensal entre \$25 e \$100 milhões					
	Presidente	Dir. financeiro	Dir. comercial	Dir. industrial	Dir. de RH	Dir. marketing
Primeiro quartil	\$25.998	\$13.305	\$12.746	\$13.523	- -	\$11.250
Mediana	\$29.654	\$15.225	\$14.762	\$13.940	- -	\$12.765
Terceiro quartil	\$31.282	\$18.026	\$15.801	\$15.902	- -	\$16.579

Analise os resultados registrados acima e responda às seguintes perguntas:

- Que percentagem dos entrevistados de cada categoria pesquisada se encontram entre o primeiro e o terceiro quartis?
- Por que o intervalo entre a mediana e o primeiro quartil de remuneração da categoria Presidente é diferente do intervalo entre o terceiro quartil e a mediana? Explique essa diferença.
- Repita a comparação anterior com as outras categorias.
- Apresente os resultados das empresas com faturamento mensal acima de \$100 milhões em um gráfico e analise sua forma.

Problema 32

A rede de restaurantes AQUeAGORA, especializada em almoços pelo sistema *refeição por quilo*, tem 30 lojas distribuídas em diversos bairros de São Paulo, todas com o mesmo padrão e capacidade de atendimento. A tabela a seguir apresenta o número de refeições servidas pelas 30 lojas em um dia típico.

290	243	295	275	216	253
266	232	256	224	252	298
316	247	234	278	270	280
226	233	298	278	266	278
252	269	239	325	240	295

Pede-se realizar uma análise dos dados, considerando que a experiência no gerenciamento desse tipo de negócio mostra que o ponto de equilíbrio de uma loja é de 250 refeições por dia.

Apêndice 1

Funções de procura e ordenamento do Excel

O cálculo das medidas de ordenamento utilizando o Excel pode ser realizado utilizando expressões matemáticas e procedimentos combinados com os recursos da planilha, as funções estatísticas e a ferramenta de análise *Ordem e Percentil* do Excel. Na planilha **Funções de Ordenamento**, incluída na pasta **Capítulo 3**, está registrada a utilização de cada função utilizando a amostra do Exemplo 3.1, como se pode ver na Figura 3.6. As sintaxes dessas funções estatísticas são apresentadas a seguir.

CORRESP(*valor; matriz; tipo*)

A função de procura e referência CORRESP⁹ retorna a posição relativa do argumento *valor* especificado no argumento *matriz* de valores em uma ordem específica. A procura é realizada conforme o argumento *tipo*:

- Se *tipo*=1, então a função CORRESP selecionará o menor valor da *matriz* que for maior ou igual ao *valor* em uma *matriz* previamente ordenada de forma decrescente.
- Se *tipo*=0, então a função CORRESP selecionará o primeiro valor da *matriz* que for exatamente igual ao *valor* especificado, sem necessidade de a *matriz* estar ordenada.
- Se *tipo*=-1, então a função CORRESP selecionará o maior valor da *matriz* que for menor ou igual ao *valor* especificado, em uma *matriz* previamente ordenada de forma crescente.

A função CORRESP é parecida com as funções PROCV e PROCH com a diferença de retornar a posição de um valor em um intervalo em vez do valor propriamente dito. O argumento *matriz* pode ser informado como um intervalo de células no qual foi registrada previamente a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado declarando todos os valores da amostra {31;38;19;27;24;42;32;18;43;15;39}.

ORDEM(*valor; amostra; ordem*)

A função estatística ORDEM¹⁰ retorna a posição do argumento *valor* da *amostra*, considerando a *ordem* informada:

- Se *ordem* for igual a 0 ou omitida, os valores da amostra serão classificados em ordem decrescente.
- Se *ordem* for diferente de 0, igual a 1, os valores da amostra serão classificados em ordem crescente.

Se o argumento *amostra* tiver valores repetidos, a função ORDEM informará a posição do primeiro valor que encontrar na sua procura, considerando o ordenamento escolhido.

⁹ Em inglês, a função CORRESP é MATCH.

¹⁰ Em inglês, a função ORDEM é RANK.

ORDEM.PORCENTUAL(*matriz; valor; núm_decimais*)

A função estatística ORDEM.PORCENTUAL¹¹ retorna o *percentil* do argumento *valor*, considerando a *matriz* ordenada de forma crescente. Se a matriz tiver valores repetidos, a função informará o percentil do primeiro valor que encontrar. O argumento *núm_decimais* define o número de casas decimais do resultado; se for omitido, o resultado terá três casas decimais. O argumento *matriz* pode ser informado em qualquer ordem, pois a função ORDEM.PORCENTUAL ordena os valores da amostra de forma crescente antes de calcular. O argumento *matriz* pode ser informado como um intervalo de células onde previamente foi registrada a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado declarando todos os valores da amostra {31;38;19;27;24;42;32;18;43;15;39}.

PERCENTIL(*matriz; k*)

A função estatística PERCENTIL¹² retorna o valor que divide a *matriz* em duas partes, uma menor do que o argumento *k* e a outra maior do que *k*. O argumento *k* é um valor entre 0 e 1,0% e 100%, ou o valor do percentil em que a matriz ordenada será dividida. A função PERCENTIL ordena os valores da *matriz* de forma crescente antes de calcular. Nem sempre o resultado da função percentil é um valor da amostra. O argumento *matriz* pode ser informado como um intervalo de células no qual previamente foi registrada a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado como {31;38;19;27;24;42;32;18;43;15;39}, declarando todos os valores da amostra.

QUARTIL(*matriz; quarto*)

A função estatística QUARTIL¹³ retorna o valor da *matriz* correspondente ao argumento *quarto* identificado da seguinte maneira:

- Se *quarto*=0, a função retornará o primeiro ou menor valor da matriz.
- Se *quarto*=1, 2 ou 3, a função retornará o valor da matriz correspondente e, respectivamente, ao primeiro, segundo ou terceiro *quartil*.
- Se *quarto*=4, a função retornará o último ou maior valor da matriz.

A função QUARTIL ordena os valores da *matriz* de forma crescente antes de calcular. Enquanto a função QUARTIL fornece resultados de posições definidas na amostra ordenada, a função PERCENTIL retorna os resultados para qualquer posição de 0 a 1, ou 0% a 100%. No entanto, nem sempre o retorno da função QUARTIL é um dado da amostra. O argumento *matriz* pode ser informado como um intervalo de células no qual previamente foi registrada a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado declarando todos os valores da amostra {31;38;19;27;24;42;32;18;43;15;39}.

MENOR(*matriz; k-ésimo*)

A função estatística MENOR¹⁴ retorna o *k-ésimo* menor valor da *matriz* ordenada de forma crescente. Para uma mesma *matriz*, o resultado dessa função dependerá do valor do argumento *k-ésimo*:

- Se *k-ésimo*=1, então o menor valor será o primeiro valor da *matriz* ordenada de forma crescente.
- Se *k-ésimo*=2, então o menor valor será o segundo valor da *matriz* ordenada de forma crescente e assim sucessivamente, até o último valor da *matriz*.

¹¹ Em inglês, a função ORDEM.PORCENTUAL é PERCENTRANK.

¹² Em inglês, a função PERCENTIL é PERCENTILE.

¹³ Em inglês, a função QUARTIL é QUARTILE.

¹⁴ Em inglês, a função MENOR é SMALL.

FIGURA 3.6 Como utilizar as funções de ordenamento.

	A	B	C	D	E	F
1	Funções de ordenamento					
2						
3		Amostra		CORRESP(valor; matriz; tipo)		
4		31		valor	18	
5		38		tipo	0	
6		19		CORRESP	8	
7		27		Como matriz		
8		24		CORRESP	8	
9		42				
10		32				
11		18		ORDEM(valor; amostra; ordem)		
12		43		valor	38	
13		15		tipo	1	
14		39		ORDEM	8	
15				Como matriz		
16				ORDEM	Não se aplica	
17						
18						
19				ORDEM.PORCENTUAL(matriz; valor; núm_decimais)		
20				valor	38	
21				núm_decimais	1	
22				ORDEM.PORCENTUAL	70,00%	
23				Como matriz		
24				ORDEM.PORCENTUAL	70,00%	
25						

Na função MENOR, não é necessário informar a série ordenada de forma crescente. O argumento *matriz* pode ser informado como um intervalo de células no qual previamente foi registrada a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado declarando todos os valores da amostra {31;38;19;27;24;42;32;18;43;15;39}.

MAIOR(matriz; k-ésimo)

A função estatística MAIOR¹⁵ dá o *k-ésimo* maior valor da *matriz* ordenada de forma crescente. Para uma mesma *matriz*, o resultado dessa função dependerá do valor do argumento *k-ésimo*:

- Se *k-ésimo*=1, então o maior valor da *matriz* será o último valor da *matriz* ordenada de forma crescente.
- Se *k-ésimo*=2, então o maior valor da *matriz* será o penúltimo valor da *matriz* e assim sucessivamente, até o primeiro valor da *matriz*.

O argumento *matriz* pode ser informado como um intervalo de células no qual previamente foi registrada a amostra, por exemplo, o intervalo B4:B14 da Figura 3.6; ou pode ser informado declarando todos os valores da amostra {31;38;19;27;24;42;32;18;43;15;39}. Na função MAIOR, não é necessário informar a série ordenada de forma crescente.

¹⁵ Em inglês, a função MAIOR é LARGE.

Apêndice 2

O símbolo somatório

Suponha uma amostra ou variável X com n dados ou observações identificados pela sequência de valores $X_1, X_2, \dots, X_i, \dots, X_n$, onde X_1 é o primeiro dado, X_2 é o segundo dado, X_i é um dado qualquer da amostra, e assim sucessivamente, até o último dado X_n .

A soma desses valores representada com $X_1 + X_2 + \dots + X_i + \dots + X_n$ se pode expressar simbolicamente com $\sum_{i=1}^n X_i$, pois $\sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_i + \dots + X_n$. A letra grega Σ , sigma maiúscula, indica que devem ser somadas expressões da forma X_i começando com $i=1$ até $i=n$.

Outro exemplo: a expressão simbólica da soma $R = 2^1 + 2^2 + 2^3 + 2^4$ é $R = \sum_{i=1}^4 X^i = 2^1 + 2^2 + 2^3 + 2^4$.

Vejam algumas propriedades de interesse, tendo presente que as propriedades se aplicam sempre nos dois sentidos da igualdade.

- O resultado de somar n vezes a constante c é o resultado do produto de n vezes a constante c . Com o símbolo somatório $\sum_{i=1}^n c = n \times c$.
- Se cada valor da sequência $X_1, X_2, \dots, X_i, \dots, X_n$ for multiplicado pela constante c , o resultado dessa soma será $\sum_{i=1}^n cX_i = c \times \sum_{i=1}^n X_i$.
- A soma algébrica das sequências $X_1, X_2, \dots, X_i, \dots, X_n$ e $Y_1, Y_2, \dots, Y_i, \dots, Y_n$ é $\sum_{i=1}^n (X_i \pm Y_i) = \sum_{i=1}^n X_i \pm \sum_{i=1}^n Y_i$. Há casos em que as propriedades anteriores do somatório são combinadas $\sum_{i=1}^n (cX_i + Y^{2i}) = \sum_{i=1}^n cX_i + \sum_{i=1}^n Y^{2i} = c \sum_{i=1}^n X_i + \sum_{i=1}^n Y^{2i}$.
- Somatórios múltiplos. A seguinte expressão é formada por três somatórios.

$$\sum_{i=1}^3 \sum_{j=1}^3 X_{i,j} = \sum_{i=1}^3 X_{i,1} + \sum_{i=1}^3 X_{i,2} + \sum_{i=1}^3 X_{i,3}$$

Essa expressão desenvolvida é:

$$\sum_{i=1}^3 \sum_{j=1}^3 X_{i,j} = X_{1,1} + X_{1,2} + X_{1,3} + X_{2,1} + X_{2,2} + X_{2,3} + X_{3,1} + X_{3,2} + X_{3,3}$$

Essas expressões representam a soma dos dados da seguinte tabela, onde i representa a linha e j a coluna.

i/j	Coluna 1	Coluna 2	Coluna 3
Linha 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$
Linha 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$
Linha 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$

Apêndice 3

Prova do mínimo da soma dos quadrados dos desvios

Denominando o desvio como D e z a qualquer número possível de ser a média da amostra X , a soma dos quadrados dos desvios será medida com a expressão $D = \sum_{i=1}^n (X_i - z)^2$. Para calcular o mínimo dessa função, primeiro deve-se calcular a primeira derivada da função D .

$$D = \sum_{i=1}^n (X_i^2 - 2zX_i + z^2)$$

$$\frac{dD}{dz} = \frac{d}{dz} \sum_{i=1}^n X_i^2 - \frac{d}{dz} \sum_{i=1}^n 2zX_i + \frac{d}{dz} \sum_{i=1}^n z^2$$

Depois, a primeira derivada deve ser igualada a zero.

$$\frac{dD}{dz} = 0 - 2 \sum_{i=1}^n X_i + 2 \sum_{i=1}^n z = 0$$

Na última expressão simplificada $-\sum_{i=1}^n X_i + nz = 0$, reconhecemos que a segunda parcela é a soma

dos dados da amostra. O valor de z é o próprio valor da amostra de X já definido como $z = \frac{\sum_{i=1}^n X_i}{n}$. O valor encontrado é realmente um mínimo, pois sua segunda derivada é positiva, como mostrado a seguir:

$$\frac{d^2 D}{dz^2} = \frac{d^2}{dz^2} \left(-2 \sum_{i=1}^n X_i + 2nz \right)$$

$$\frac{d^2 D}{dz^2} = 2n > 0$$

Apêndice 4

Funções de tendência central do Excel

O cálculo das medidas de tendência central utilizando o Excel pode ser realizado utilizando expressões matemáticas e procedimentos combinados com os recursos da planilha e funções estatísticas. Na planilha **Funções de Tendência Central**, incluída na pasta **Capítulo 3**, está registrada a utilização de cada função utilizando a amostra do Exemplo 3.15, como se pode ver na Figura 3.7. Uma característica comum das funções a seguir, exceto a função MÉDIA.INTERNA, são os 30 argumentos (*núm1*; *núm2*; ... ; *núm30*) utilizados para registrar os valores de intervalos. Na apresentação da primeira função SOMA, será mostrado como utilizar esses argumentos, procedimentos que se repetem com as demais funções com o mesmo tipo de argumentos. As sintaxes dessas funções estatísticas são apresentadas a seguir.

SOMA(*núm1*; *núm2*; ... ; *núm30*)

A função matemática SOMA¹⁶ retorna a *soma* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou semelhantes.¹⁷ Por exemplo, a função SOMA aplicada aos valores da amostra do Exemplo 3.15 dá como resultado 352. Para obter esse resultado, a função SOMA pode ser utilizada das seguintes maneiras, Figura 3.7:

- Registrando os valores da amostra em um intervalo de células da planilha.
 - Se os valores da variável estiverem registrados em um único intervalo, ou intervalos contíguos, apenas será necessário informar um único intervalo no argumento *num1*. Por exemplo, na célula F6 foi registrada a fórmula =SOMA(B4:C17), Figura 3.7.
 - Se os valores da variável estiverem registrados em intervalos não adjacentes, será necessário informar o endereço de cada intervalo no lugar de cada *núm* de *núm1*; *núm2*; ... ; *núm30*, até um máximo de 30. Por exemplo, a fórmula =SOMA(B4:C8;B9:B17;C9:C15) registrada na célula F7 tem três intervalos nos três primeiros argumentos da função SOMA *núm1*; *núm2*; *núm3*.
- Registrando os valores da amostra como *matriz* na própria fórmula da função, evitando registrar os valores da amostra em um intervalo de células da planilha.
 - Na célula G6, os valores foram registrados em uma única matriz:

=SOMA({14;12;13;11;12;13;16;14;14;15;17;14;11;13;14;15;

13;12;14;13;14;13;15;16;12;12})
 - Na célula G7, os valores foram registrados em quatro matrizes:

=SOMA({14;12;13;11};{12;13;16;14;14;15;17;14;11;13};

{14;15;13;12;14;13;14;13;15};{16;12;12})

correspondentes aos quatro primeiros argumentos da função SOMA *núm1*; *núm2*; *núm3*; *núm4*.

¹⁶ Em inglês, a função SOMA é SUM.

¹⁷ Assemelhados são os intervalos definidos por nomes, valores lógicos, representações em forma de texto de números; por exemplo, com a função de texto VALOR("10")=10.

MÉDIA(núm1; núm2; ... ; núm30)

A função estatística MÉDIA¹⁸ retorna a *média aritmética* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Um detalhe importante: se o nome da função MÉDIA for inserido com letras minúsculas ou maiúsculas sem o acento ortográfico, o Excel aceitará e registrará a função com letras maiúsculas e com o acento ortográfico. A função MÉDIA pode ser registrada de diversas formas equivalentes às descritas na função SOMA mencionada anteriormente, Figura 3.7.

MÉDIAA(núm1; núm2; ... ; núm30)

A função estatística MÉDIAA¹⁹ é equivalente à função anterior MÉDIA. A diferença está relacionada com os valores registrados nos argumentos *núm1*; *núm2*; ... ; *núm30* que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO. Deixamos para o leitor pesquisar na Ajuda do Excel.

FIGURA 3.7 Como utilizar as funções de tendência central.

	A	B	C	D	E	F	G
1	Funções de tendência central						
2							
3		Amostra				Dados informados como	
4		14	14		Função Matemática	Intervalo	Matriz
5		12	15				
6		13	13		SOMA	352,00	352,00
7		11	12			352,00	352,00
8		12	14				
9		13	13		Funções Estatísticas		
10		16	14				
11		14	13		MÉDIA	13,54	13,54
12		14	15			13,54	13,54
13		15	16				
14		17	12		MED	13,50	13,50
15		14	12			13,50	13,50
16		11					
17		13			MODO	14,00	14,00
18						14,00	14,00
19							

MED(núm1; núm2; ... ; núm30)

A função estatística MED²⁰ retorna a *mediana* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um dos *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. A função MED pode ser registrada de diversas formas equivalentes às descritas na função SOMA anteriormente, Figura 3.7.

MODO(núm1; núm2; ... ; núm30)

A função estatística MODO²¹ retorna o *modo* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha que contém valores numéricos ou asseme-

¹⁸ Em inglês, a função MÉDIA é AVERAGE.

¹⁹ Em inglês, a função MÉDIAA é AVERAGEA.

²⁰ Em inglês, a função MED é MEDIAN.

²¹ Em inglês, a função MODO é MODE.

lhados. Quando a série tem mais de uma moda, a função reconhece apenas uma delas. A função MOD pode ser registrada de diversas formas equivalentes às descritas na função SOMA anteriormente, Figura 3.7.

MÉDIA.GEOMÉTRICA(núm1; núm2; ... ; núm30)

A função estatística MÉDIA.GEOMÉTRICA²² retorna a *média geométrica* dos valores da amostra. Cada um dos núm pode ser um intervalo de células de uma planilha que contém valores numéricos ou *assemelhados*. A média geométrica Mg é definida como $Mg = (X_1 \times X_2 \times \dots \times X_n)^{1/n}$ com os valores X_i maiores do que zero. Comparando com a *média aritmética*:

- A média geométrica é menos afetada por valores extremos.
- A média geométrica é uma medida mais central quando os valores da variável apresentam uma taxa constante de crescimento.
- Para um mesmo grupo de valores, a média geométrica é sempre menor do que a média aritmética.

A função MÉDIA.GEOMÉTRICA pode ser registrada de diversas formas equivalentes às descritas na função SOMA anteriormente, Figura 3.7. Uma aplicação frequente da *média geométrica* é o cálculo da taxa equivalente de juros de uma operação financeira formada por n operações com taxas de juros diferentes, como mostrado no Capítulo 16, utilizando a fórmula:

$$Mg = \left((1 + i_1) \times (1 + i_2) \times \dots \times (1 + i_n) \right)^{1/n}$$

$$i = Mg - 1$$

MÉDIA.HARMÔNICA(núm1; núm2; ... ; núm30)

A função estatística MÉDIA.HARMÔNICA²³ retorna a *média harmônica* dos valores da amostra. Cada um dos núm pode ser um intervalo de células de uma planilha que contém valores numéricos ou *assemelhados*. A *média harmônica* é uma medida útil quando os valores se referem a mudanças de uma magnitude, e seu valor é sempre menor do que o da média geométrica do mesmo conjunto de valores.

- A *média harmônica* é a inversa da média aritmética das inversas dos valores da amostra: $Mh = \frac{1}{\frac{1}{n} \times \sum_{i=1}^n \frac{1}{X_i}}$.
- De outra maneira, a inversa da *média harmônica* Mh é a média da inversa dos valores da amostra: $\frac{1}{Mh} = \frac{1}{n} \times \sum_{i=1}^n \frac{1}{X_i}$.

A função MÉDIA.HARMÔNICA pode ser registrada de diversas formas equivalentes às descritas na função SOMA anteriormente, Figura 3.7.

MÉDIA.INTERNA(matriz; porcentagem)

A função estatística MÉDIA.INTERNA²⁴ retorna a *média aritmética* da *matriz* de valores, tendo previamente excluído, de ambos extremos da *matriz*, uma *porcentagem* de valores informada como valor unitário. É uma média reduzida útil para remover dados extremos, *suspeitos*, de uma amostra.

²² Em inglês, MÉDIA.GEOMÉTRICA é GEOMEAN.

²³ Em inglês, MÉDIA.HARMÔNICA é HARMEAN.

²⁴ Em inglês, MÉDIA.INTERNA é TRIMMEAN.

Capítulo 4

MEDIDAS DE DISPERSÃO

No Capítulo 3, foi mostrado que a média e a mediana determinam um valor central de uma amostra ou variável. Enquanto a mediana localiza a posição do dado ou observação situada no centro da amostra ordenada de forma crescente, e sem considerar os valores da variável, a média determina o valor central considerando todos os valores da variável. Por exemplo, as amostras $X=\{28, 29, 30, 31, 32\}$ e $Y=\{21, 25, 29, 34, 41\}$ têm o mesmo número de dados e, também, a mesma média 30. Entretanto, os desvios são diferentes, pois os desvios da variável X são $-2, -1, 0, 1$ e 2 , e os desvios da variável Y são $-9, -5, -1, 4$ e 11 . A comparação dessas duas amostras aponta a variabilidade ou dispersão de seus dados com relação à média como uma medida importante para descrever uma amostra ou variável. Esse raciocínio poderia ser repetido em variáveis com medianas iguais, porém com menor aplicação do que a média.

Você deve ter em mente que, se não houver variabilidade, a maior parte das medidas estatísticas não teria utilidade. Há várias formas de medir a variabilidade dos dados de uma variável. Uma primeira tentativa é medir o intervalo ou range de variação, definido como o resultado da diferença entre os valores máximo e mínimo da amostra ou variável, como apresentado no Exemplo 2.1 do Capítulo 2.

EXEMPLO 4.1

Determine o intervalo de variação da seguinte amostra:

31	38	19	27	24	42	32	18	43	15	39
----	----	----	----	----	----	----	----	----	----	----

Solução. Os valores mínimo e máximo são, respectivamente, 15 e 43. O intervalo ou range de variação dos dados da amostra é $28=43-15$.

O resultado do Exemplo 4.1 mostra que os dados da amostra se distribuem dentro do intervalo de variação igual a 28. O conhecimento desse intervalo não auxilia muito na tentativa de medir a dispersão dos dados da variável, pois seu cálculo envolve apenas os valores extremos, deixando de considerar os demais valores da variável que também são importantes.

Desvio absoluto médio

No Capítulo 3, vimos que os desvios dos dados de uma amostra ou variável medem sua dispersão ao redor de sua média. Portanto, a tentativa inicial de quantificar a variabilidade seria calcular a soma de todos os desvios, isto é $\sum_{i=1}^n (X_i - \bar{X})$. No entanto, pela primeira propriedade da média, a soma dos desvios é sempre igual a zero. Tentando manter o conceito desvio como medida de variabilidade, pode-se utilizar a média dos valores absolutos¹ dos desvios, procedimento denominado *desvio absoluto médio* ou simplesmente *DAM*.²

O *Desvio absoluto médio-DAM* é obtido da expressão:

$$DAM = \frac{1}{n} \times (|X_1 - \bar{X}| + |X_2 - \bar{X}| + \dots + |X_n - \bar{X}|)$$

$$DAM = \frac{1}{n} \times \sum_{i=1}^n |X_i - \bar{X}|$$

onde X_i é um valor genérico e \bar{X} é a média da variável ou amostra.

EXEMPLO 4.2

Calcule o desvio absoluto médio da amostra do Exemplo 4.1.

Solução. A resposta foi obtida na planilha **Exemplo 4.2**, incluída na pasta **Capítulo 4**, como mostra a figura a seguir.

- No intervalo B4:B14 foi registrada a amostra.
- Na célula G5, foi calculada a média da amostra com =MÉDIA(B4:B14), retornando o valor 29,82.
- Na célula C6 foi calculado o desvio do dado 31 da amostra registrando a fórmula =B4-\$G\$5, retornando o valor 1,18. Depois, essa fórmula foi copiada até a célula C14.
 - O valor de média que mostra a célula G5 é 29,82, valor arredondado com duas casas decimais. Entretanto, o valor exato e registrado na memória do Excel é 29,8181818181818. Ao mesmo tempo, no cálculo dos desvios, o Excel utiliza o valor exato da média. Portanto, você poderá encontrar diferenças entre o resultado final do *DAM* obtido manualmente com a média e os desvios arredondados e o obtido com o Excel sem arredondar nenhum resultado intermediário.
- Na célula D4, foi calculado o valor absoluto do desvio do dado 31, calculado na célula C4, registrando a fórmula =ABS(C4) que retornou o valor 1,18. Depois essa fórmula foi copiada até a célula D14.
 - Em vez de utilizar duas colunas para calcular o desvio absoluto, poderia ter sido utilizada uma única coluna registrando na célula C4; por exemplo, a fórmula combinada =ABS(B4-\$G\$5) que depois seria copiada.
 - A função matemática **ABS(número)**³ retorna o valor absoluto do argumento *número* que pode ser qualquer número do campo real. Pode-se dizer que o valor absoluto de um número é o próprio número sem o respectivo sinal, seja positivo ou negativo.
- Na célula G6 foi registrada a fórmula =SOMA(D4:D14) que retorna o resultado da soma dos desvios absolutos igual a 92,18.

¹ O valor absoluto de um número é o valor desse número considerado positivo.

² Este procedimento é apenas um registro, pois o *DAM* não ajuda na compreensão da dispersão, nem apresenta as vantagens matemáticas da variância e do desvio padrão.

³ Em inglês, a função ABS é ABS.

	A	B	C	D	E	F	G	H
1	Exemplo 4.2							
2								
3		Amostra	Desvio	Desvio Absoluto		Resultados		
4		31	1,18	1,18				
5		38	8,18	8,18				
6		19	-10,82	10,82				
7		27	-2,82	2,82				
8		24	-5,82	5,82				
9		42	12,18	12,18				
10		32	2,18	2,18				
11		18	-11,82	11,82				
12		43	13,18	13,18				
13		15	-14,82	14,82				
14		39	9,18	9,18				
15								

Média	29,82
Soma dos Desvios Absolutos	92,18
DAM	8,38

Função DESV.MÉDIO	8,38
--------------------------	-------------

=DESV.MÉDIO(B4:B14)

Com os resultados parciais obtidos, pode-se calcular o $DAM=8,38$:

- Manualmente a fórmula $DAM = \frac{\sum_{i=1}^{11} |X_i - \bar{X}|}{11} = \frac{92,18}{11} = 8,38$
- Registrando a fórmula =G6/CONT.NÚM(D4:D14) na célula G7 da planilha.

Uma forma direta de obter o resultado desejado é utilizar a função estatística DESV.MÉDIO do Excel que retorna o desvio absoluto médio da amostra informada. Na célula G9 foi registrada a fórmula =DESV.MÉDIO(B4:B14). No Apêndice 1, você encontrará a descrição completa dessa e de outras funções que serão apresentadas neste capítulo.

Comparado com a tentativa de medir a variabilidade com o *intervalo*, o DAM é a média dos desvios absolutos e utiliza todos os valores da variável ou amostra. Entretanto, o valor absoluto dos desvios é um resultado difícil de compreender e não aceita tratamento matemático com as propriedades, por exemplo, do quadrado do desvio que será utilizado a seguir.

Variância

Mantendo os *desvios* para medir a variabilidade de uma variável, o procedimento recomendado é utilizar a soma dos quadrados dos desvios, pois seu resultado é um valor mínimo, como mostrou a segunda propriedade da média apresentada no Capítulo 3.

Seja a variável $X = X_1, X_2, \dots, X_N$ uma população. Define-se variância σ_X^2 da variável X da população contendo N dados:

$$\sigma_X^2 = \frac{1}{N} \times ((X_1 - \mu_X)^2 + (X_2 - \mu_X)^2 + \dots + (X_n - \mu_X)^2)$$

$$\sigma_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X)^2$$

Seja a variável $X = X_1, X_2, \dots, X_n$ uma amostra. Define-se a variância S_X^2 da variável X da amostra contendo n dados:

$$S_X^2 = \frac{1}{n-1} \times ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2)$$

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

EXEMPLO 4.3

Calcule a variância da amostra e da população do Exemplo 4.1 utilizando as fórmulas e as funções estatísticas do Excel.

Solução. A resposta foi obtida na planilha **Exemplo 4.3**, incluída na pasta **Capítulo 4**, como mostra a figura seguinte e tendo presente as características de arredondamento dos resultados intermediários e finais já comentadas.

- No intervalo B4:B14 foi registrada a amostra, na célula G4 foi calculada quantidade de dados da amostra e na célula G5 foi calculada a média da amostra utilizando a fórmula =MÉDIA(B4:B14).
- No intervalo C4:C14 foram calculados os desvios e no intervalo D4:D14, os quadrados dos desvios começando por registrar a fórmula =C4^2 na célula D4. Depois essa fórmula foi copiada até a célula D14.
- Na célula G8 foi calculada e registrada a soma dos quadrados dos desvios igual a 997,64 com a fórmula =SOMA(D4:D14).
- Utilizando a função matemática SOMAQUAD não é necessário construir a coluna dos quadrados dos desvios. A fórmula =SOMAQUAD(C4:C14) registrada na célula G9 retorna a soma dos quadrados dos valores registrados no intervalo C4:C14. No Apêndice 1, você encontrará a descrição completa dessa e de outras funções que serão apresentadas.

	A	B	C	D	E	F	G	H	I
1	Exemplo 4.3								
2									
3		Amostra	Desvio	(Desvio)²		Resultados			
4		31	1,18	1,40		n	11	=CONT.NÚM(B4:B14)	
5		38	8,18	66,94		Média	29,82	=MÉDIA(B4:B14)	
6		19	-10,82	117,03					
7		27	-2,82	7,94		Soma (Desvios)²			
8		24	-5,82	33,85		Fórmula	997,64	=SOMA(D4:D14)	
9		42	12,18	148,40		Função SOMAQUAD	997,64	=SOMAQUAD(C4:C14)	
10		32	2,18	4,76					
11		18	-11,82	139,67		Variância amostra			
12		43	13,18	173,76		Fórmula	99,76	=G6/(G4-1)	
13		15	-14,82	219,58		Função VAR	99,76	=VAR(B4:B14)	
14		39	9,18	84,31					
15						Variância população			
16						Fórmula	90,69	=G6/G4	
17						Função VARP	90,69	=VARP(B4:B14)	
18									

Cálculo da variância da amostra. Com os resultados parciais obtidos, pode-se calcular o valor da variância da amostra $s_X^2 = 99,76$, utilizando:

- $$s_X^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{11-1} = \frac{997,64}{10} = 99,76$$
- Manualmente a fórmula $s_X^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{11-1} = \frac{997,64}{10} = 99,76$
 - Registrando a fórmula =G6/(G4-1) na célula G12 da planilha.
 - Utilizando a função estatística VAR, registrando a fórmula =VAR(B4:B14) na célula G13.

Cálculo da variância da população. Com os resultados parciais obtidos, pode-se calcular o valor da variância da amostra $\sigma_X^2 = 90,69$, utilizando:

- $$\sigma_X^2 = \frac{\sum_{i=1}^{11} (X_i - \mu_X)^2}{11} = \frac{997,64}{11} = 90,69$$
- Manualmente a fórmula $\sigma_X^2 = \frac{\sum_{i=1}^{11} (X_i - \mu_X)^2}{11} = \frac{997,64}{11} = 90,69$
 - Registrando a fórmula =G6/G4 na célula G16 da planilha.
 - Utilizando a função estatística VARP, registrando na célula G17 a fórmula =VARP(B4:B14).

O procedimento de cálculo manual da variância é bastante trabalhoso quando comparado com a utilização das funções estatísticas do Excel; entretanto, essas funções apenas auxiliam o cálculo e podem obscurecer o conceito. O Apêndice 3 deste capítulo mostra como utilizar doze funções para banco de dados ou listas de valores, conhecidas genericamente como BDFunções (*banco_dados*; *campo*; *critérios*). Algumas dessas doze funções são equivalentes às apresentadas. Ademais, esse apêndice apresenta também as funções SUBTOTAL, CONT.SE e SOMASE úteis para realizar operações com bancos de dados ou listas de valores.

Relação entre as variâncias

A partir das definições das variâncias da amostra e da população, o Exemplo 4.3 mostra os procedimentos de cálculo, incluindo as funções estatísticas VAR e VARP. Verifique que uma das variâncias pode ser obtida da outra se o tamanho da amostra também for conhecido. Para facilitar a relação entre as variâncias da população e da amostra repetimos a seguir suas fórmulas.

$$\sum_{i=1}^N (X_i - \mu_X)^2 = N \times \sigma_X^2$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = (n-1) \times S_X^2$$

Como os dois primeiros membros dessas expressões são iguais, é possível igualar os dois segundos membros, o que nos leva à seguinte igualdade:

$$N \times \sigma_X^2 = (n-1) \times S_X^2$$

Portanto, conhecida uma das variâncias, é possível calcular a outra, sendo necessário também conhecer o tamanho da amostra.

$$\sigma_X^2 = \frac{(n-1)}{N} \times S_X^2 \text{ e } S_X^2 = \frac{N}{n-1} \times \sigma_X^2$$

EXEMPLO 4.4

Calcule a variância da população a partir da variância da amostra do Exemplo 4.3, sabendo que o tamanho da amostra é 11.

Solução. A variância da população $\sigma_X^2 = 90,69$ pode ser obtida com a fórmula:

$$\sigma_X^2 = \frac{n-1}{N} \times S_X^2$$

$$\sigma_X^2 = \frac{10}{11} \times 99,76 = 90,69$$

Em vez de tentar memorizar a fórmula de transformação entre as variâncias, recomenda-se ter presente a seguinte orientação:

- A variância da amostra foi obtida como resultado da divisão da soma dos quadrados dos desvios pela quantidade de valores da amostra ($n-1$). Para obter o valor da variância da população, será necessário multiplicar a variância da amostra por $(n-1)$ e, em sequência, dividi-la por n .
- A variância da amostra será o resultado da multiplicação da variância da população por n e, em sequência, divida-a por $(n-1)$.

Características da variância

O procedimento de cálculo utilizando a soma dos quadrados dos desvios é bastante trabalhoso. No Apêndice 2, mostramos um procedimento de cálculo da variância que utiliza somente os dados da amostra e os quadrados desses dados, não sendo necessário utilizar a média e os desvios. Contudo, esse procedimento de cálculo perde força quando comparado com a utilização das funções estatísticas do Excel. A fórmula e o resultado da variância têm características importantes.

- A variância é sempre um número positivo.
- As fórmulas para a amostra e para a população têm o mesmo numerador, a soma dos quadrados dos desvios.
- A variância de uma variável considerada como população é a média aritmética dos quadrados dos desvios.
- A variância de uma variável considerada como amostra é também um tipo de média, pois a soma dos quadrados dos desvios é dividida pela quantidade de dados da variável menos um.⁴
- Para a mesma amostra de tamanho n , a variância da amostra é sempre maior do que a da população. Na medida em que o tamanho n da amostra aumenta, para n maior do que 30, o valor da variância da amostra se aproxima do valor da variância da população.
- Da mesma forma que a média, a variância é afetada pelos valores extremos da variável, ela não é uma medida resistente.
- Uma desvantagem da variância é sua unidade de medida, o quadrado da unidade de medida dos dados da amostra ou variável; outra desvantagem é operar com os valores dos desvios ampliados, pois os desvios são elevados ao quadrado.

Regras operacionais da variância

Há propriedades operacionais muito práticas. Para evitar muitos símbolos nas fórmulas, as variâncias serão representadas como $Var(X)$. Sendo a , b e c constantes, sempre se verifica:

- Se $Y = a$, $Var(Y) = 0$
- Se $Y = aX$, $Var(Y) = a^2 Var(X)$
- Se $Y = X + a$, $Var(Y) = Var(X)$
- Se $Y = X + Z$, $Var(Y) = Var(X) + Var(Z) + 2 Cov(X, Z)$
- Se $Y = aX + bZ$, $Var(Y) = a^2 Var(X) + b^2 Var(Z) + 2 ab Cov(X, Z)$

Desvio padrão

Para definir da variância nos valem da segunda propriedade da média: *a soma dos quadrados dos desvios é sempre um valor mínimo*, como foi apresentado no Capítulo 3. Uma desvantagem da variância é sua unidade de medida, o quadrado da unidade de medida dos dados da amostra ou variável; outra desvantagem é ampliar os desvios, pois são elevados ao quadrado. Por exemplo, se a amostra do Exemplo 4.3 se refere a peças rejeitadas por lote, a unidade de medida da variância da amostra será 99,76 peças rejeitadas ao quadrado, o que não faz muito sentido. Como a unidade de medida da variância não explica nada sobre as características dos valores da amostra, é definido o *desvio padrão* que mantém a unidade de medida dos valores da variável.

⁴ No cálculo da variância da amostra S^2 , deve-se dividir por $(n-1)$ em vez de n para corrigir a tendência de S^2 subestimar σ^2 ; para que S^2 seja um estimador *não viesado*.

O *desvio padrão* da variável X é a raiz quadrada positiva de sua variância. Dessa maneira:

O *desvio padrão* considerado como população é: $\sigma_X = +\sqrt{\sigma_X^2}$.

O *desvio padrão* considerado como amostra é: $S_X = +\sqrt{S_X^2}$.

Essas definições mostram que para determinar o desvio padrão é necessário conhecer o valor da variância correspondente, da amostra ou da população.

EXEMPLO 4.4

Calcular o desvio padrão da amostra e da população do Exemplo 4.1 utilizando as fórmulas e as funções estatísticas do Excel.

Solução. A resposta foi obtida na planilha **Exemplo 4.4**, incluída na pasta **Capítulo 4**, como mostra a figura a seguir e tendo presente as características de arredondamento dos resultados intermediários e finais já comentadas. O registro da amostra, os cálculos dos resultados intermediários e a obtenção dos valores das variâncias da amostra e da população foram realizados da mesma forma como foi apresentado no Exemplo 4.3. Esse procedimento é necessário para mostrar o cálculo do desvio padrão a partir de sua definição, ou a partir do conhecimento da variância correspondente, amostra ou população. No entanto, esse procedimento de cálculo perde força quando comparado com a utilização das funções estatísticas do Excel.

	A	B	C	D	E	F	G	H	I
1	Exemplo 4.4								
2									
3		Amostra	Desvio	(Desvio)²		Resultados			
4		31	1,18	1,40		n	11		
5		38	8,18	66,94		Média	29,82		
6		19	-10,82	117,03		Soma (Desvios)²	997,64		
7		27	-2,82	7,94		Variância amostra	99,76		
8		24	-5,82	33,85		Variância população	90,69		
9		42	12,18	148,40					
10		32	2,18	4,76		Desvio padrão amostra			
11		18	-11,82	139,67		Fórmula	9,99	=RAIZ(G7)	
12		43	13,18	173,76		Função DESVPAD	9,99	=DESVPAD(B4:B14)	
13		15	-14,82	219,68					
14		39	9,18	84,31		Desvio padrão população			
15						Fórmula	9,52	=RAIZ(G8)	
16						Função DESVPADP	9,52	=DESVPADP(B4:B14)	
17									

Cálculo do desvio padrão da amostra. O valor do desvio padrão da amostra $S_X = 9,99$ pode ser obtido:

- Manualmente a fórmula $S_X = +\sqrt{S_X^2} = +\sqrt{99,76} = 9,99$
- Registrando a fórmula =RAIZ(G7) na célula G11 da planilha.
 - A função matemática **RAIZ(número)**⁵ retorna a raiz quadrada positiva do argumento *número* que deve ser qualquer número positivo.
- Utilizando a função estatística DESVPAD ao registrar na célula G12 a fórmula =DESVPAD(B4:B14).

Cálculo do desvio padrão da população. O valor da desvio padrão da população $\sigma_X = 9,52$ pode ser obtido:

- Manualmente pela fórmula $\sigma_X = +\sqrt{\sigma_X^2} = +\sqrt{90,69} = 9,52$
- Registrando a fórmula =RAIZ(G8) na célula G15 da planilha.
- Utilizando a função estatística DESVPADP ao registrar na célula G16 a fórmula =DESVPADP(B4:B14).

5 Em inglês, a função RAIZ é SQRT.

Se a amostra do Exemplo 4.4 se refere à quantidade mensal de peças rejeitadas, o desvio padrão da amostra será 9,99 peças rejeitadas, pois o desvio padrão tem a mesma unidade dos dados da amostra ou variável. Da mesma maneira, o desvio padrão da população é $\sigma_x = +\sqrt{90,69} = 9,52$ peças rejeitadas. O procedimento de cálculo manual do desvio padrão é bastante trabalhoso quando comparado com a utilização das funções estatísticas do Excel; entretanto, essas funções apenas auxiliam o cálculo e podem obscurecer o conceito.

Relação entre os desvios padrão

A partir das definições dos desvios padrão da amostra e da população, o Exemplo 4.4 mostra os procedimentos de cálculo, incluindo as funções estatísticas DESVPAD e DESVPADP. Nesse caso, também, verifica-se que um dos desvios padrão pode ser obtido do outro se o tamanho da amostra também for conhecido. Em alguns casos é necessário operar com os valores do desvio padrão da população e do desvio padrão da amostra de uma variável, tentando sempre usar uma forma prática de obter um valor do outro. Da mesma forma como foi mostrada a relação entre a variância da amostra e a variância da população, as expressões a seguir mostram a relação entre os desvios padrão da população e da amostra.

$$\sigma_x = \sqrt{\frac{n-1}{N}} \times S_x \text{ e } S_x = \sqrt{\frac{N}{n-1}} \times \sigma_x$$

O procedimento recomendado para obter o valor de um desvio padrão em função do outro é, primeiro, realizar essa operação com as variâncias equivalentes, evitando carregar uma fórmula com o símbolo de raiz quadrada. Da mesma forma que a variância, as características do desvio padrão são:

- O desvio padrão é sempre um número positivo.
- Se os dados de uma variável forem iguais, o desvio padrão será zero.
- O desvio padrão não é uma medida resistente, pois é afetada pelos valores extremos da variável.

Significado do desvio padrão

O desvio padrão depende da soma dos quadrados dos desvios dos dados da variável com relação a sua média. Portanto, quanto menor for o desvio padrão, mais os valores da variável se aproximarão de sua média. Analisando a expressão do desvio padrão, podemos chegar a conclusões importantes:

- Qualquer dado da amostra ou variável com desvio menor do que o desvio padrão da variável estará mais próximo da média do que qualquer outro valor com desvio maior.
- Quanto mais os dados se afastarem da média, maior serão os desvios e, conseqüentemente, maior será o desvio padrão da variável.
- Duas variáveis com médias iguais e desvios padrão diferentes têm distribuições de frequências com formas diferentes. A distribuição da variável com maior desvio padrão será mais aberta do que a da variável com menor desvio padrão.

Qual a proporção de dados incluídos em um intervalo de desvios padrão ao redor da média de uma variável ou amostra? O *Teorema de Chebyshev* dá uma resposta para uma variável com qualquer tipo de distribuição de frequências.

Teorema de Chebyshev. Para qualquer conjunto de dados de uma amostra ou população, a proporção mínima de valores que se encontram dentro de k desvios padrão ao redor da média é pelo menos igual a $1 - \frac{1}{k^2}$, sendo k uma constante maior do que 1.

A próxima tabela mostra a proporção mínima de dados dentro de k desvios padrão ao redor da média. Por exemplo, 75% dos dados de uma amostra ou variável estão distribuídos no intervalo de dois desvios padrão ao redor da média; entre menos dois e mais dois desvios padrão ao redor da média.

k	1,5	2	2,5	3	3,5	4
Proporção de dados	0,56	0,75	0,84	0,89	0,92	0,94

Outro exemplo, pelo menos sete dos onze dados da amostra do Exemplo 4.1 estão distribuídos no intervalo de dois desvios padrão ao redor da média 29,8; isto é, entre menos dois desvios padrão ($9,8=29,8-2 \times 9,99$) e mais dois desvios padrão ($49,8=29,8+2 \times 9,99$) ao redor da média. Verifique que no Exemplo 4.1 todos os dados estão distribuídos no intervalo de dois desvios padrão ao redor da média.

Regra prática

Pelo teorema de Chebyshev, é possível determinar a proporção mínima de dados de uma variável dentro de um determinado número de desvios padrão ao redor da média. A partir da média \bar{X} e o desvio padrão S_X de uma amostra ou variável X , a *Regra Prática* permite estabelecer a proporção de valores distribuídos no intervalo $\bar{X} \pm k \times S_X$, considerando a forma da distribuição de frequências da variável X .

Regra Prática

A variável X tem n dados com média \bar{X} e desvio padrão S_X .

$\bar{X} \pm 1 \times S_X$. Em uma distribuição simétrica com forma de sino, a porcentagem de dados contidos no intervalo de um desvio padrão ao redor da média é 68%. Para uma distribuição assimétrica com acentuada inclinação para um lado, essa porcentagem se aproxima de 90%.

$\bar{X} \pm 2 \times S_X$. Em uma distribuição simétrica com forma de sino, a porcentagem de dados contidos no intervalo de dois desvios padrão ao redor da média é 95%. Para uma distribuição assimétrica com acentuada inclinação para um lado, a porcentagem se aproxima de 100%.

$\bar{X} \pm 3 \times S_X$. Para todas as distribuições, a porcentagem de dados contidos no intervalo de três desvios padrão ao redor da média será próxima de 100%.

A *Regra Prática* atende à maioria das distribuições; entretanto, há casos em que será necessário construir o histograma para conhecer a forma da distribuição da amostra. A partir das conclusões obtidas da aplicação da *Regra Prática*, será possível determinar a forma do histograma, da distribuição de frequências dos dados como mostra o Exemplo 4.5.

EXEMPLO 4.5

Determine a porcentagem dos dados da amostra do Exemplo 4.1 incluídos no intervalo de um, dois e três desvios padrão ao redor da média.

Solução. Na planilha **Exemplo 4.5**, incluída na pasta **Capítulo 4**, foram determinadas as quantidades de dados incluídos nos intervalos de um, dois e três desvios padrão ao redor da média, como mostra a figura seguinte.

- Na célula H5, foi calculada a quantidade de dados, na célula H6, a média e, na célula H7, o desvio padrão da amostra utilizando as funções estatísticas correspondentes.
- No intervalo H10:J11, foram calculados os valores dos limites inferiores e superiores dos intervalos de um, dois e três desvios padrão ao redor da média, acompanhando a expressão $\bar{X} \pm k \times S_X$ cujas fórmulas do primeiro intervalo são as seguintes:
 - Com a fórmula =H6-H7 registrada na célula H10, foi calculado o limite inferior do intervalo de um desvio padrão.
 - Com a fórmula =H6+H7 registrada na célula H11, foi calculado o limite superior do intervalo de um desvio padrão. Para os demais limites, procede-se da mesma forma, considerando o número de desvios padrão adequados.

	A	B	C	D	E	F	G	H	I	J
1	Exemplo 4.5									
2										
3		Amostra	1 DP	2DP	3DP		Resultados			
4		14	1	1	1					
5		18	1	1	1					
6		12	0	1	1					
7		17	1	1	1					
8		19	0	1	1					
9		17	1	1	1					
10		14	1	1	1					
11		18	1	1	1					
12		10	0	0	1					
13		17	1	1	1					
14		18	1	1	1					
15										

n	11
Média	15,64
Desvio padrão	2,80

	1 DP	2 DP	3 DP
Limite Inferior	12,83	10,03	7,23
Limite Superior	18,44	21,24	24,04
Quantidade	8	10	11
Porcentagem	73%	91%	100%

Com os limites estabelecidos, nas colunas do intervalo C4:E14, são selecionados os dados contidos em cada intervalo utilizando as seguintes fórmulas:

- **Um desvio padrão ao redor da média, coluna C.** Na célula C4 foi registrada a fórmula =SE(E(B4>=\$H\$10;B4<=\$H\$11);1;0), que depois foi copiada até a célula C14.
- **Dois desvios padrão ao redor da média, coluna D.** Na célula D4 foi registrada a fórmula =SE(E(B4>=\$I\$10;C4<=\$I\$11);1;0), que depois foi copiada até a célula D14.
- **Três desvios padrão ao redor da média, coluna E.** Na célula E4 foi registrada a fórmula =SE(E(B4>=\$J\$10;D4<=\$J\$11);1;0), que depois foi copiada até a célula E14.

Para terminar, no intervalo H12:J12 são contados os dados contidos no intervalo de um, dois e três desvios padrão ao redor da média, e no intervalo H13:J13 são calculadas as respectivas porcentagens, obtendo os seguintes resultados $\bar{X} \pm 1 \times S_X = 73\%$, $\bar{X} \pm 2 \times S_X = 91\%$ e $\bar{X} \pm 3 \times S_X = 100\%$. Portanto, 73%, 91% e 100% dos dados ou observações se distribuem, respectivamente, no intervalo de um, dois e três desvios padrão ao redor da média.

Medida relativa de dispersão

O desvio padrão tem duas características importantes:⁶

⁶ A variância também tem essas duas características.

- Considera que os desvios se distribuem de forma homogênea ao redor da média.
- É uma medida absoluta.

A comparação da dispersão de duas ou mais distribuições pelo simples confronto de seus desvios padrão nem sempre é suficiente, pois as amostras ou populações podem ter unidades diferentes ou, tendo a mesma unidade, seus valores de média podem estar bastante afastados.

O *coeficiente de variação CV* é o resultado de dividir o desvio padrão da variável pela sua média:

$$CV_{pop} = \frac{\sigma_X}{\mu_X} \quad CV_{amo} = \frac{S_X}{\bar{X}}$$

A medida relativa de dispersão *coeficiente de variação CV* permite a comparação de distribuições, pois seu resultado é o desvio padrão por unidade de média. Em alguns casos, o resultado do CV é apresentado multiplicado por 100, em porcentagem. Comparando duas variáveis, a variável que tiver menor CV tem menor dispersão ou variabilidade.

EXEMPLO 4.6

A tabela a seguir registra os retornos mensais dos investimentos A e B durante os últimos seis meses. Interessa conhecer qual dos dois investimentos apresentou maior dispersão.

Solução. Na planilha **Exemplo 4.6**, incluída na pasta **Capítulo 4**, foi resolvido o exemplo, começando pelo cálculo das médias e dos desvios padrão dos retornos dos dois investimentos e terminando pelo cálculo do coeficiente de variação de cada investimento.

	A	B	C	D	E	F	G
1	Exemplo 4.6						
2							
3	Retornos		Resultados				
4	A	B					
5	5%	6%					
6	9%	7%					
7	15%	9%					
8	12%	7%					
9	9%	6%					
10	6%	8%					
11							

Como o CV do investimento A é maior do que o CV do investimento B, a variabilidade⁷ do investimento A foi maior do que a do investimento B.

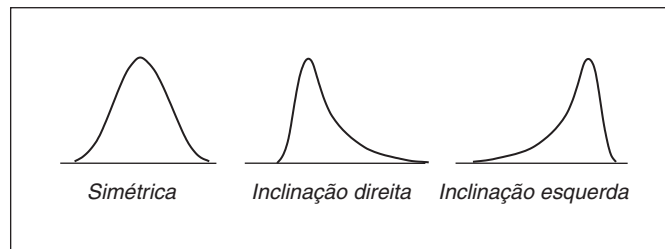
Análise da forma da distribuição de uma amostra

Como nem todas as amostras ou variáveis têm moda ou um único valor de moda, a mediana e a média são as medidas mais usuais de tendência central. Em uma distribuição simétrica de frequências, a média e a mediana têm o mesmo valor. Se os valores da média e da mediana forem diferentes, a distribuição será assimétrica e quanto mais os valores da média e da mediana se afastarem, maior será a inclina-

⁷ Em finanças, a variabilidade é o *risco* do investimento; o investimento A apresentou mais risco do que o investimento B.

ção da distribuição na direção de uma das caudas. Por exemplo, se um ou mais dados da amostra forem valores maiores do que a maioria dos demais dados, então a média será maior do que a mediana e a distribuição de frequências terá inclinação direita ou positiva, conforme mostra a Figura 4.1. Da mesma forma, é possível analisar para o lado esquerdo. A inclinação de uma distribuição é medida pelo *coeficiente de inclinação* da distribuição.

FIGURA 4.1
Distribuições de frequências, simétrica e inclinada.



Apesar de duas amostras ou variáveis apresentarem a mesma dispersão e inclinação, essas características não serão suficientes para supor que as duas distribuições tenham a mesma forma, atributo denominado *curtose*. A curtose é medida pelo *coeficiente de curtose* que compara a distribuição de frequências da amostra com a distribuição normal.

EXEMPLO 4.7

A tabela a seguir registra uma amostra ordenada de 28 retornos de diversos investimentos no mesmo período. Calcule e analise a forma da distribuição dessa amostra.

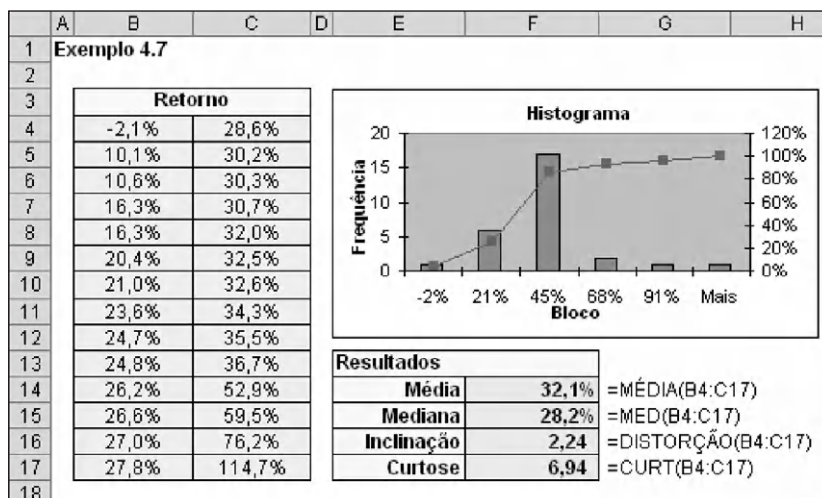
-2,1%	10,1%	10,6%	16,3%	16,3%	20,4%	21,0%
23,6%	24,7%	24,8%	26,2%	26,6%	27,0%	27,8%
28,6%	30,2%	30,3%	30,7%	32,0%	32,5%	32,6%
34,3%	35,5%	36,7%	52,9%	59,5%	76,2%	114,7%

Solução. Na planilha **Exemplo 4.7**, incluída na pasta **Capítulo 4**, foi analisada a forma da distribuição da amostra anterior registrada no intervalo B4:C17, como mostra a figura seguinte. Analisemos os resultados registrados nessa planilha.

- O histograma foi construído utilizando a ferramenta de análise *Histograma*, depois de ajustar a formatação do gráfico, os títulos e as escalas. O histograma mostra que a distribuição apresenta inclinação para a direita.
- No intervalo F14:F15, foram calculadas a média e a mediana, respectivamente iguais a 32,1% e 28,2%. Como a média é maior do que a mediana, a distribuição tem inclinação para a direita.
- O *coeficiente de inclinação* igual a 2,24 foi calculado com a função estatística DISTORÇÃO do Excel, registrando a fórmula =DISTORÇÃO(B4:C17) na célula F16. O resultado positivo mostra que a distribuição tem inclinação para a direita. Se o resultado fosse negativo, a inclinação seria negativa, e se fosse igual a zero, a distribuição seria simétrica.

• DISTORÇÃO (núm1; núm2; ... ; núm30)

A função estatística DISTORÇÃO (núm1; núm2; ... ; núm30) retorna o coeficiente de inclinação dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Nesse exemplo, a amostra do intervalo B4:C17 foi registrado no primeiro argumento núm1. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.



- O coeficiente de curtose igual a 6,94 foi calculado com a função estatística CURT do Excel registrando a fórmula =CURT(B4:C17) na célula F17. O resultado positivo mostra que a distribuição de frequências será concentrada ao redor da média, distribuição com pico. Se o resultado fosse negativo, a distribuição seria achatada, plana, e se fosse igual a zero, a distribuição de frequências seria a própria distribuição normal.

• CURT (núm1; núm2; ... ; núm30)

A função estatística CURT (núm1; núm2; ... ; núm30) retorna o coeficiente de curtose dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. Nesse exemplo, a amostra do intervalo B4:C17 foi registrada no primeiro argumento núm1. Mais informações sobre essa função e outras formas de utilizá-la estão disponíveis no Apêndice 1 deste capítulo.

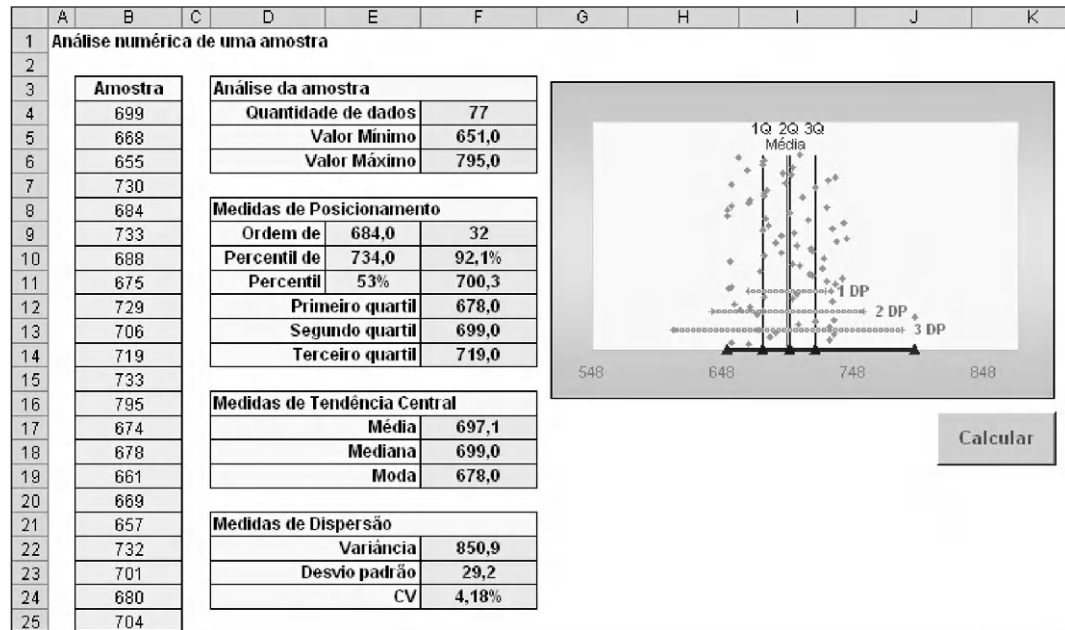
Modelo análise numérica

A determinação isolada de medidas estatísticas numéricas leva à obtenção de respostas parciais. O conjunto dessas medidas melhora a compreensão e a visualização das medidas numéricas em um gráfico complementa a análise da amostra. Realizar esse processo de medição de forma manual é muito trabalhoso; entretanto, utilizando a planilha Excel, consegue-se diminuir um pouco esse trabalho. O conjunto desses resultados é apresentado no *Modelo Análise Numérica* construído pelo autor na pasta **Modelo Análise Numérica** que está disponível na página do livro, no site da Editora. A Figura 4.2 mostra esse modelo para uma amostra de tamanho $n=77$, incluindo o gráfico que destaca as medidas numéricas mais importantes.

O *Modelo Análise Numérica* calcula as medidas mais importantes e constrói um gráfico com os dados da amostra, os intervalos de um, dois e três desvios padrão ao redor da média, a identificação de uma linha (no eixo de abscissas) com as cinco medidas estatísticas que ajudam a descrever a forma da distribuição de frequências, e a identificação de linhas verticais da média e do primeiro, segundo e terceiro quartis. Para operar o *modelo*:

- Recomenda-se zerar os dados da amostra diretamente na planilha.
- Informar a série de valores numéricos a partir da célula B4. Não há limite de tamanho da amostra, apenas os limites impostos pela planilha Excel e a memória do microcomputador utilizada.
- Depois de informar a amostra, pressione o botão **Calcular**. O *modelo* fornecerá os resultados do intervalo F4:F24 e construirá ou atualizará o gráfico.
- No intervalo E9:F11, é possível obter respostas específicas para um dado da amostra quanto à sua posição no intervalo E9:F9, ao seu percentil no intervalo E10:F10 e ao dado referente a um determinado percentil no intervalo E11:F11.
- Preste atenção ao aviso de recálculo que o *modelo* apresenta na célula mesclada H2.

FIGURA 4.2 Modelo
Análise Numérica.



Ainda, no intervalo D26:F29, não mostrado na Figura 4.2, o modelo apresenta a contagem e a proporção de dados dentro de um, dois e três desvios padrão ao redor da média.

Ferramenta de análise *Estatística Descritiva*

A partir de uma amostra quantitativa discreta registrada em uma planilha Excel, a ferramenta de análise *Estatística descritiva* retornará uma tabela com um grupo de resultados estatísticos. Para utilizar a ferramenta de análise *Estatística descritiva*,⁸ a amostra que será analisada deve estar registrada em uma planilha como a *Ferram. Estatística Descritiva*, incluída na pasta *Capítulo 4*, com a amostra do Exemplo 4.1, onde:

- No intervalo B3:B14 foram registrados os valores numéricos da amostra, incluindo o nome Amostra na célula B3. Os valores da amostra devem ser registrados em uma coluna identificados com um único intervalo. Essa ferramenta de análise pode gerar tabelas para mais de uma amostra simultaneamente com a condição de terem o mesmo tamanho e serem registradas em intervalos contíguos.
- Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 deste livro.
- Escolhendo a ferramenta **Estatística descritiva** e depois pressionando o botão **OK**, você receberá a caixa de diálogo **Estatística descritiva** mostrada na Figura 4.3 depois de selecionadas algumas opções.
 - Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Estatística descritiva* pertencente à *Ajuda do Excel*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo da ferramenta *Estatística descritiva* são, conforme apresentado na Figura 4.3:

- **Intervalo de entrada:** Informe o intervalo de células da planilha, no qual os dados estão registrados, nesse caso, o intervalo B3:B14 que inclui a célula onde foi registrado o título *Amostra*, ou rótulo no Excel.

⁸ Em inglês, a ferramenta ESTATÍSTICA DESCRITIVA é DESCRIPTIVE STATISTICS.

- **Agrupado por:** Seleccionamos **Colunas**, pois a amostra foi registrada em uma coluna. Em geral, o Excel selecionará automaticamente depois de ter informado intervalo da amostra.
- **Rótulos na primeira linha.** Tendo escolhido **Colunas** no item anterior, necessariamente selecionaremos **Rótulos na primeira linha**, pois na primeira célula da série foi incluído o nome *Amostra*.

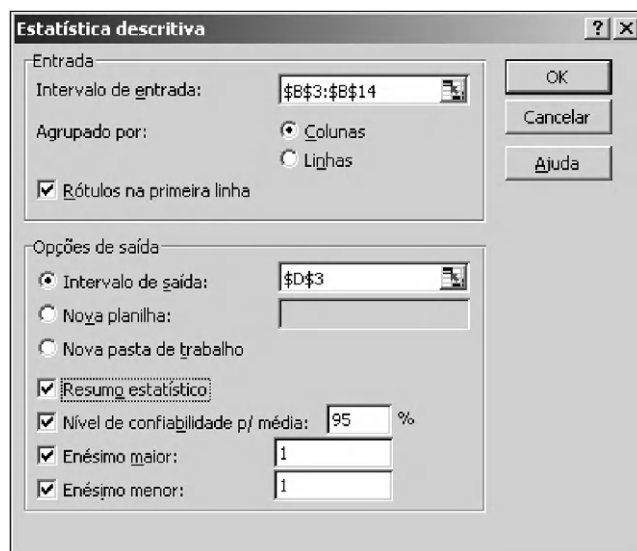


FIGURA 4.3 Caixa de diálogo da ferramenta **Estatística descritiva**.

Na primeira parte do quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta *Estatística descritiva* registrará os resultados. Há três alternativas nas quais não é necessário informar esse endereço, identificadas por três *botões de opção* que aceitam a escolha de uma única alternativa:

- **Intervalo de saída:** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso D3. Depois de clicar com o botão esquerdo do mouse dentro da caixa correspondente, o endereço pode ser registrado digitando D3, ou *clcando* com o botão esquerdo do *mouse* na célula D3; nesse caso, será registrado o endereço com os dois cifrões, \$D\$3. Esse endereço é o da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes.
- **Nova planilha:** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
 - Se não for informado nenhum endereço, a ferramenta inserirá uma nova planilha com o nome **Plan** seguido de um número sequencial; por exemplo, escolhendo essa alternativa na pasta **Capítulo 4**, a ferramenta inserirá a planilha **Plan1**.
 - Há a alternativa de informar o nome da planilha na caixa dessa alternativa; por exemplo, registrando o nome *Teste*, a ferramenta inserirá na mesma pasta uma nova planilha com o nome **Teste**.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha **Plan1**.

Em continuação, no quadro **Opções de saída**, há quatro alternativas não excludentes de resultados possíveis. Nelas é possível selecionar qualquer combinação marcando nas quatro *caixas de seleção*, com a condição de selecionar pelo menos uma delas.

FIGURA 4.4
Resumo estatístico
da ferramenta
Estatística Descritiva.

	A	B	C	D	E	F	G
1	Ferramenta de Análise ESTATÍSTICA DESCRITIVA						
2							
3		Amostra		Amostra			Resultados com fórmulas
4		31					
5		38	Média	29,81818			29,81818
6		19	Erro padrão	3,011548			3,011548
7		27	Mediana	31			31
8		24	Modo	#N/D			#N/D
9		42	Desvio padrão	9,988175			9,988175
10		32	Variância da amostra	99,76364			99,76364
11		18	Curtose	-1,46891			-1,46891
12		43	Assimetria	-0,11748			-0,11748
13		15	Intervalo	28			28
14		39	Mínimo	15			15
15			Máximo	43			43
16			Soma	328			328
17			Contagem	11			11
18			Maior(1)	43			43
19			Menor(1)	15			15
20			Nível de confiança(95,0%)	6,710148			6,710148
21							

- **Resumo estatístico:** Marcando este item, a ferramenta de análise apresentará o resumo estatístico completo, conforme apresentada na Figura 4.4.
- **Nível de confiabilidade p/a média:** A resposta dessa seleção será compreendida ao estudar *Estimação* no Capítulo 11 deste livro. Neste caso, registramos 95, que representa 95% de *intervalo de confiança*.
- **Enésimo maior:** escolhendo este item e informando o valor 1, a ferramenta fornecerá o maior valor da *Amostra* ordenada de forma crescente. Se for informado o valor 2, então a ferramenta apresentará o penúltimo valor da amostra, e assim sucessivamente. **Enésimo maior** retorna o mesmo resultado da função estatística MAIOR, apresentada no Capítulo 3.
- **Enésimo menor:** Escolhendo este item e informando o valor 1, a ferramenta fornecerá o menor valor da *Amostra* ordenada de forma crescente. Se for informado o valor 2, então a ferramenta apresentará o segundo elemento da série, e assim sucessivamente. **Enésimo menor** retorna o resultado da função estatística MENOR, apresentada no Capítulo 3.

Depois de realizar as escolhas e pressionar o botão OK, a ferramenta registra os resultados a partir da célula D3, Figura 4.4.

Análise dos resultados

No intervalo G5:G20, foram registrados os mesmos resultados do intervalo E5:E20 da ferramenta de análise, porém calculados com fórmulas e funções estatísticas, algumas delas já conhecidas. Nem todos os resultados registrados na tabela da Figura 4.4 foram apresentados até o momento no livro; por exemplo, *Erro padrão* e *Nível de confiança (95%)*, mas que a seguir é feita uma introdução.

- **Erro padrão.** O *Erro padrão* é o *erro amostral* S_e estudado na Distribuição Amostral, no Capítulo 10 deste livro. O valor de S_e é calculado com a expressão, também registrado na célula G6 da planilha Ferram. Estatística Descritiva:

$$S_e = \frac{S_x}{\sqrt{n}} = \frac{9,988175}{\sqrt{11}} = 3,0011548$$

- **Nível de confiança(95%).** O *Nível de confiança* é estudado no Capítulo 11, sendo 95% o percentual de acerto da estimativa da média da população. O resultado 6,710148 da célula E20 é o *erro de estimação* com distribuição t , $t_{(1-0,95)/2} \times S_e = 2,228139 \times 3,011548 = 6,710148$. Esse resultado foi calculado na célula G20 com a função estatística INVT e a fórmula =INVT(0,05;10)*G7, que será apresentada no Capítulo 11.

EXEMPLO 4.8

Analise as distribuições de frequências das amostras A e B registradas na tabela seguinte utilizando a ferramenta de análise *Estatística descritiva*.

A	100	120	120	120	120	120	120	140	140	140	140	160	160	180
B	88,6	108,5	108,6	128,5	128,6	128,5	128,6	148,6	148,5	148,6	148,6	148,6	148,6	168,6

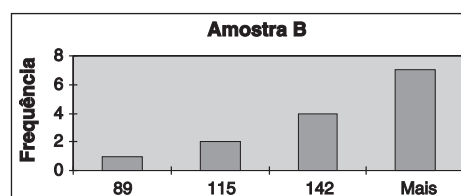
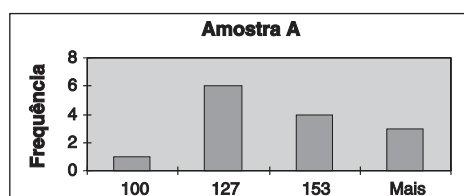
Solução. As amostras A e B e os resultados da ferramenta de análise foram registrados na planilha **Exemplo 4.8**, incluída na pasta **Capítulo 4**, como mostra a figura seguinte depois de ajustar as larguras das colunas e a formatação dos resultados. Analisando os resultados das medidas estatísticas, verificamos que as amostras A e B têm o mesmo valor de média igual a 134,29, medianas diferentes, respectivamente, 130 e 138,55, e desvio padrão praticamente iguais, respectivamente, 21,38 e 21,39. Comparando somente as médias e os desvios padrão, aparentemente, parece que as amostras têm a mesma forma de distribuição. Entretanto, a diferença de medianas mostra que não é assim.

- Como a média da amostra A é maior do que a mediana, pode-se deduzir que a distribuição de frequências da amostra A tem inclinação positiva. Essa inclinação também é confirmada pelo resultado *Assimetria* igual a 0,67 que, por ser positivo, indica a inclinação positiva da distribuição.

No caso da amostra B, ocorre o contrário: ela tem inclinação para a esquerda, como confirmado também pelo resultado *Assimetria* igual a -0,66 que, por ser negativo, indica a inclinação negativa da distribuição.

	A	B	C	D	E	F	G	H
1	Exemplo 4.8							
2								
3	Amostra A		Amostra B		Amostra A		Amostra B	
4	100	88,6			Média	134,29	Média	134,29
5	120	108,5			Erro padrão	5,71	Erro padrão	5,72
6	120	108,6			Mediana	130,00	Mediana	138,55
7	120	128,5			Modo	120,00	Modo	148,60
8	120	128,6			Desvio padrão	21,38	Desvio padrão	21,39
9	120	128,5			Variança da amostra	457,14	Variança da amostra	457,50
10	120	128,6			Curtose	0,19	Curtose	0,18
11	140	148,6			Assimetria	0,67	Assimetria	-0,66
12	140	148,5			Intervalo	80,00	Intervalo	80,00
13	140	148,6			Mínimo	100,00	Mínimo	88,60
14	140	148,6			Máximo	180,00	Máximo	168,60
15	160	148,6			Soma	1880,00	Soma	1880,00
16	160	148,6			Contagem	14,00	Contagem	14,00
17	180	168,6						
18								

Para facilitar a confirmação da análise anterior, com a ferramenta de análise *Histograma*, foram construídos os histogramas a partir da linha 20 da planilha **Exemplo 4.8**. Analisando os histograma, verifica-se que as distribuições são diferentes, pois enquanto a distribuição de frequências da amostra A tem inclinação para a direita, a da amostra B é para a esquerda.



EXEMPLO 4.9

Continuando com o Exemplo 4.8. Analise as distribuições das amostras A e B considerando as seguintes cinco medidas de posição, mínimo, primeiro quartil, mediana, terceiro quartil e máximo.

Solução. As amostras A e B e os resultados da ferramenta de análise foram registrados na planilha **Exemplo 4.9**, incluída na pasta **Capítulo 4**, como mostra a próxima figura. No intervalo F6:G10, estão registrados os resultados: *Mínimo*, Q_1 , *Mediana*, Q_3 e *Máximo* de cada amostra. Note que essas cinco medidas estão registradas em ordem crescente dos resultados. Analisando esses resultados, obtemos:

- As duas amostras têm o mesmo *intervalo* igual a $80 = 180 - 100 = 168,8 - 88,6$.
- A diferença entre o terceiro quartil e o primeiro quartil das duas amostras é o mesmo valor e igual a 20. Esse resultado mostra que 50% dos dados em cada amostra se distribuem entre os dois quartis.
- A *mediana* de cada amostra está situada no centro de Q_1 e Q_3 .
- A diferença entre o Q_1 e o *Mínimo* da amostra A é 20, enquanto a da amostra B é 39,9.
- Da mesma maneira, a diferença entre o *Máximo* e o Q_3 da amostra A é 40, e a da amostra B é 20.

	A	B	C	D	E	F	G
1	Exemplo 4.9						
2							
3	Amostra A		Amostra B		Resultados		
4	100	88,6					
5	120	108,5					
6	120	108,6					
7	120	128,5					
8	120	128,6					
9	120	128,5					
10	120	128,6					
11	140	148,6					
12	140	148,5					
13	140	148,6					
14	140	148,6					
15	160	148,6					
16	160	148,6					
17	180	168,6					
18							

	Amostra A	Amostra B
Mínimo	100	88,6
Q_1	120	128,5
Mediana	130,0	138,6
Q_3	140	148,6
Máximo	180	168,6

Intervalo entre Q_1 e Q_3

Os resultados do Exemplo 4.9 ajudarão a compreender o intervalo entre o primeiro quartil e o terceiro quartil, denominado *IEQ*,⁹ e as vantagens do diagrama *Boxplot* que será apresentado em sequência. O primeiro quartil, a mediana e o terceiro quartil avaliam a forma da parte central e a variabilidade da distribuição de frequências da amostra. O *IEQ* é o resultado da diferença entre o terceiro quartil Q_3 e o primeiro quartil Q_1 :

$$IEQ = Q_3 - Q_1$$

As características importantes do *IEQ* são:

- É uma medida simples, fácil de ser calculada e automatizada.
- Mede a distribuição da metade dos dados da amostra situados ao redor da mediana.
- É uma medida resistente, pois não é afetado pelos dados extremos da amostra ou variável.
- É parecido com o *intervalo*; entretanto, essas três medidas Q_1 , *mediana* e Q_3 dão mais informações.

⁹ Em inglês, IEQ é IQR – *InterQuartile Range*.

- Contudo, essa medida não é suficiente para avaliar a variabilidade de uma amostra ou variável, pois envolve apenas os valores centrais, deixando de considerar os valores extremos que também são importantes, os restantes 50% dos dados.

Boxplot

Embora os três resultados Q_1 , mediana e Q_3 mostrem a forma da distribuição de 50% dos valores ao redor da mediana de uma amostra ou variável, o conjunto formado por esses cinco resultados:¹⁰ *mínimo*, Q_1 , *mediana*, Q_3 e *máximo* permitirão obter muitas informações sobre a forma da distribuição de frequências.

O *boxplot*¹¹ é a forma gráfica para mostrar o conjunto dos cinco resultados estatísticos e obter informações diretas sobre a forma da distribuição de frequências da amostra ou variável. O *boxplot* da Figura 4.5, planilha à esquerda, mostra que a inclinação da amostra A é positiva ou para a direita, confirmando o resultado obtido no Exemplo 4.9. O *boxplot* da Figura 4.5, planilha à direita, mostra que a inclinação da amostra B é negativa ou para a esquerda, confirmando também o resultado obtido no Exemplo 4.9. No gráfico do *boxplot*, foi incluída uma linha (no eixo de abscissas) com as cinco medidas estatísticas que ajudam a descrever a forma da distribuição de frequências, como mostrado no **Modelo Análise Numérica**. Observe que cada amostra tem um *boxplot* diferente que registra:

- Uma medida de tendência central, a *mediana*.
- Duas medidas de variabilidade ou dispersão, o *intervalo* e o *IEQ*.
- O tipo de *inclinação* por comparação da *mediana* com relação aos valores extremos.
- Os possíveis dados suspeitos.

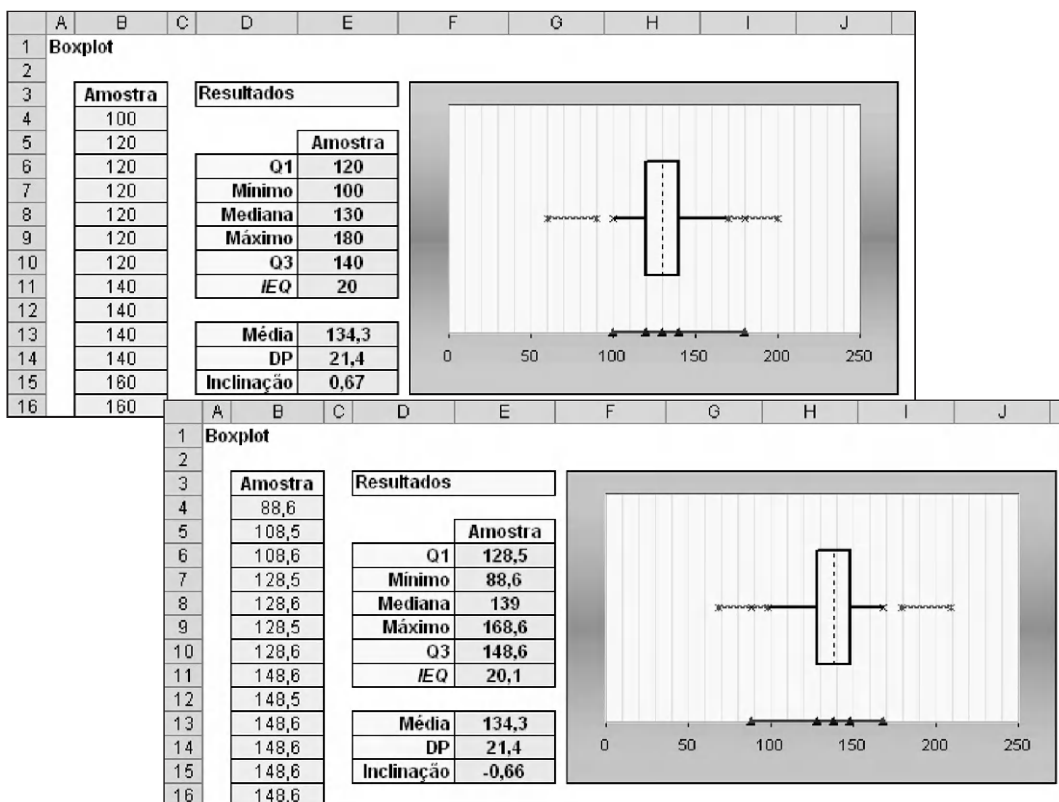


FIGURA 4.5 Boxplot das amostras A e B, Exemplo 4.9.

¹⁰ Em inglês, conhecido como *Five-number Summary*.

¹¹ Preferimos manter o nome *boxplot* em inglês.

Dado suspeito

É denominado *dado suspeito*¹² o dado de uma amostra extremamente diferente da maioria dos dados da amostra. Como qualquer amostra pode conter *dados suspeitos*, é importante estar preparado para detectá-lo e analisar sua causa.

- Se o dado suspeito tiver sua origem em um erro de registro; por exemplo, o valor medido 135 foi registrado como 2.135. Nesse caso, o erro pode ser corrigido e a característica suspeita pode ser eliminada do dado amostrado.
- O que fazer se o dado suspeito foi corretamente amostrado e registrado? Se a população está sendo amostrada através de uma pesquisa de indivíduos de uma determinada população, um dado suspeito poderá ser originado por um indivíduo que não pertence à população definida. O dado suspeito também pode ser evidência de um acontecimento extraordinário ou uma variabilidade não esperada da variável. Em qualquer caso, os dados suspeitos sem causa aparente associada à população devem ser retirados da amostra, registrando esse evento.

O valor X de uma variável é considerado *possível suspeito* se estiver no intervalo $Q_1 - 3 \times IEQ < X < Q_1 - 1,5 \times IEQ$ ou no intervalo $Q_3 + 1,5 \times IEQ < X < Q_3 + 3 \times IEQ$.

O valor X de uma variável é considerado *suspeito* se $X < Q_1 - 3 \times IEQ$ ou $X > Q_3 + 3 \times IEQ$.

Uma estratégia para tratar dados suspeitos e outras irregularidades é utilizar métodos numéricos resistentes que pouco são afetados pelos dados suspeitos. Uma das aplicações do *IEQ* é a detecção de valores suspeitos de uma variável. Embora o *IEQ* ajude a retirar um dado da amostra por considerá-lo suspeito, essa decisão deve ser acompanhada de um criterioso julgamento.

EXEMPLO 4.10

Calcule o *IEQ* das amostras A e B do Exemplo 4.9 e verifique a existência de dados suspeitos.

Solução. A figura a seguir mostra a resolução deste exemplo na planilha **Exemplo 4.10**, incluída na pasta **Capítulo 4**. A primeira parte dos resultados é igual ao Exemplo 4.9, adicionando o intervalo F11:G11 para o cálculo do *IEQ* de cada amostra. Depois, no intervalo E13:I15, foram calculados os limites dos dados suspeitos indicados nesta tabela.

	A	B	C	D	E	F	G	H	I
1	Exemplo 4.10								
2									
3		Amostra A	Amostra B		Resultados				
4		100	88,6						
5		120	108,5						
6		120	108,6						
7		120	128,5						
8		120	128,6						
9		120	128,5						
10		120	128,6						
11		140	148,6						
12		140	148,5						
13		140	148,6						
14		140	148,6						
15		160	148,6						
16		160	148,6						
17		180	168,6						
18									

	Amostra A	Amostra B
Mínimo	100	88,6
Q_1	120	128,5
Mediana	130,0	138,6
Q_3	140	148,6
Máximo	180	168,6
IEQ	20	20,1

	$Q_1 - 3 \times IEQ$	$Q_1 - 1,5 \times IEQ$	$Q_3 + 1,5 \times IEQ$	$Q_3 + 3 \times IEQ$
Amostra A	60	90	170	200
Amostra B	68,2	98,4	178,8	208,9

12 Em inglês, *dados suspeitos* são *outliers*.

- Amostra A.
 - Na cauda inferior da distribuição, são suspeitos os valores menores do que 60, e os valores entre 60 e 90 são possíveis suspeitos. Como o valor mínimo é 100, essa amostra não tem valores suspeitos nessa região.
 - Na cauda superior da distribuição, são suspeitos os valores maiores do que 200, e os valores entre 170 e 200 são possíveis suspeitos. O único valor possível de suspeita é o valor máximo 180.
- Amostra B.
 - Na cauda inferior da distribuição, são suspeitos os valores menores do que 68,2. Os valores entre 68,2 e 98,4 são possíveis suspeitos como o valor mínimo é 88,6.
 - Na cauda superior da distribuição, são suspeitos os valores maiores do que 208,9, e os valores entre 178,8 e 208,9 são possíveis suspeitos. Nenhum valor deve ser considerado suspeito.

Os intervalos de detecção de valores suspeitos foram adicionados ao *boxplot* da planilha **Boxplot**, como mostram as planilhas da Figura 4.5 referentes às amostras A e B. Verifique que:


- Nos extremos da distribuição, são representados os segmentos dos valores potencialmente suspeitos, linhas de cor vermelha.
- A amostra A não tem valores suspeitos na cauda inferior; entretanto, pode ter valores suspeitos na cauda superior da distribuição.
- A amostra B tem valores suspeitos na cauda inferior; entretanto, pode não ter valores suspeitos na cauda superior da distribuição.

Boxplot com Excel

O *boxplot* de uma amostra também pode ser construído utilizando os recursos gráficos do Excel. Na planilha **Boxplot com Excel**, incluída na pasta **Capítulo 4**, foram repetidos os dados e os resultados da planilha **Exemplo 4.9**, fazendo uma cópia dessa planilha. Depois, as posições dos resultados dos cinco números, *mínimo*, Q_1 , *mediana*, Q_3 e *máximo* foram mudadas para a nova sequência dos cinco resultados, Q_1 , *mínimo*, *mediana*, *máximo* e Q_3 .

Construção de um Boxplot

Depois de ter mudado as posições dos cinco resultados na planilha **Boxplot com Excel** proceda assim:¹³

- Selecione o intervalo E5:F10 da planilha **Boxplot com Excel**.
- Clique no ícone assistente de gráfico  e, na página **Tipos padrão** de gráficos, selecione o tipo de gráfico **Linha** e o subtipo de gráfico **Linhas com marcadores exibidos a cada valor de dado**.
- Depois, clique no botão **Avançar**. Na guia **Intervalo de dados** você deverá selecionar **Linhas** apesar de os dados estarem registrados em colunas, como mostra a Figura 4.6, à esquerda. Depois clique no botão **Concluir**.

Agora temos um gráfico como o mostrado na Figura 4.6, à esquerda. Para construir a forma do *boxplot* proceda desta forma:

- Clique duas vezes seguidas com o botão esquerdo do mouse em cima de um dos pontos do gráfico construído. Aparecerá a caixa de diálogo **Formatar sequência de dados**.
- Na caixa de diálogo **Formatar sequência de dados**, selecione a guia **Opções**. Nessa página, marque as caixas **Linhas de máximo/mínimo** e **Barras superiores/inferiores** como mostrado na Figura 4.6, à direita.

¹³ Adaptado de Hunt N. – *Boxplots in Excel* em <http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm>.

- Para terminar, ajuste a formatação do gráfico da forma que achar mais conveniente, mudando a posição da legenda, a cor do fundo do gráfico, a identificação dos cinco pontos etc.

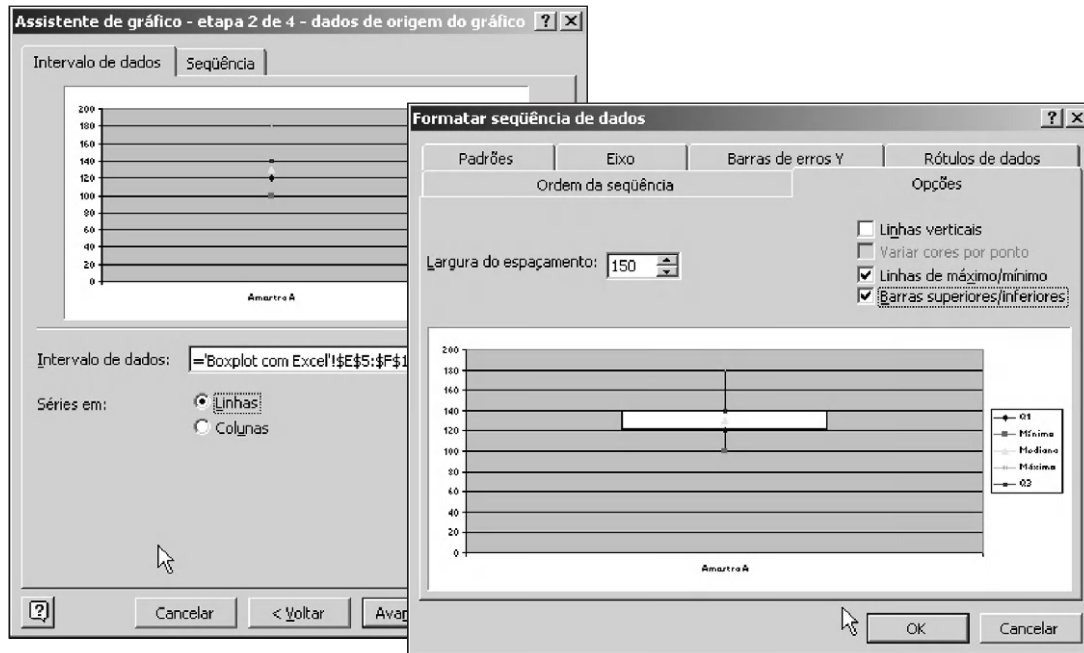



FIGURA 4.6 Construção de um *boxplot* com Excel.

Construção de dois ou mais Boxplot

O procedimento é parecido com o de um *boxplot* anterior e pode ser utilizado para mais de dois *boxplot*:

- Selecione o intervalo E5:G10 da planilha **Boxplot com Excel**.
- Clique no ícone assistente de gráfico  e, na página **Tipos padrão** de gráficos, selecione o tipo de gráfico **Linha** e o subtipo de gráfico **Linhas com marcadores exibidos a cada valor de dado**.
- Depois clique no botão **Avançar**. Na guia **Intervalo de dados**, deverá selecionar **Linhas**, apesar de os dados estarem registrados em colunas, como mostrado na Figura 4.7, à esquerda. Depois clique no botão **Concluir**.

Agora temos um gráfico como o mostrado na Figura 4.7, à esquerda. Para construir a forma de dois *boxplot*, siga este procedimento:

- Clique duas vezes seguidas com o botão esquerdo do mouse na primeira linha do gráfico construído. Será exibida a caixa de diálogo **Formatar sequência de dados**.
- Na caixa de diálogo **Formatar sequência de dados**, selecione a guia **Padrões** e, no quadro **Linha**, marque **Nenhuma** e depois pressione OK. Verifique se, com essa instrução, a linha que ligava os dois pontos foi removida.
- Repita o procedimento anterior com as quatro linhas restantes.
- Na caixa de diálogo **Formatar sequência de dados**, selecione a guia **Opções**. Nessa página, marque as caixas **Linhas de máximo/mínimo** e **Barras superiores/inferiores** como mostra a Figura 4.7, à direita.
- Para terminar, ajuste a formatação do gráfico da forma que achar mais conveniente, mudando a posição da legenda, a cor do fundo do gráfico, a identificação dos cinco pontos etc.

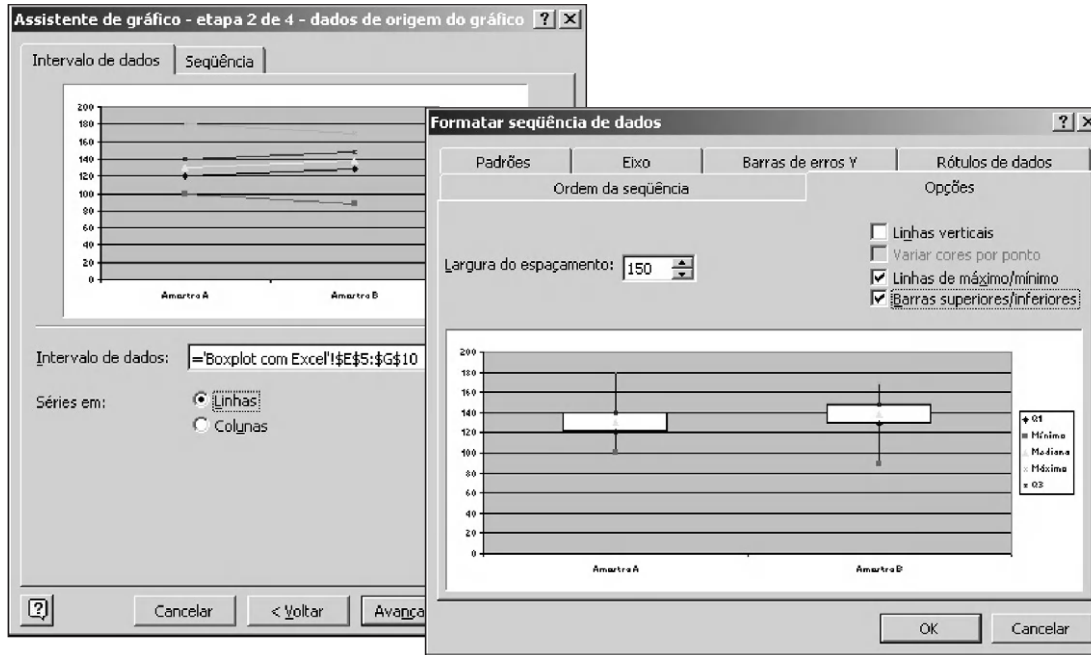


FIGURA 4.7 Construção de dois *boxplot* com Excel.

Problemas

Problema 1

Calcule a variância e o desvio padrão da amostra registrada na tabela seguinte:

10	15	14	23	21	18	11	12	14	15	23	12	15
----	----	----	----	----	----	----	----	----	----	----	----	----

R: $S^2=19,09$ e $S=4,37$

Problema 2

Calcule a variância e o desvio do Problema 1, considerando, como população.

R: $\sigma^2=17,62$ e $\sigma=4,20$

Problema 3

Repita o Problema 2, calculando a variância e o desvio padrão da população a partir da variância e do desvio padrão da amostra e utilizando as fórmulas.

Problema 4

A tabela a seguir registra uma amostra do número de gerentes operacionais que respondem diretamente a um diretor em empresas do ramo químico. Calcule a média e o desvio padrão do número de gerentes por empresa:

7	7	9	8	7	13	10	14	8	9	8	6
9	9	10	11	7	8	9	6	8	11	12	10

R: $\bar{X}=9$ e $S=2,09$

Problema 5

Calcule a variância e o desvio padrão da amostra registrada na tabela:

10	15	14	23	21	18	11	12	14	15	23	12	18	16	15
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

R: $S^2=16,74$ e $S=4,09$

Problema 6

A tabela seguinte registra as notas finais de um grupo de alunos da disciplina Estatística. Calcule a variância e o desvio padrão da amostra.

89,5	74,7	99,4	84,9	96,5	82,1	77,7	92,7	59,1	74,7	91,0	100	77,6	98,5	2,2	60,8
83,1	20,1	84,2	70,1	90,8	97,5	78,2	31,7	98,1	99,0	94,3	73,4	85,7	94,1	61,0	77,8

R: $\bar{X}=78,14$ e $S=23,15$

Problema 7

Continuando com Problema 6. Calcule a mediana da amostra e analise a inclinação da distribuição.

R: $Md=83,65$. A distribuição tem inclinação para a esquerda, pois $\bar{X} < Md$, como mostra o *coeficiente de inclinação* igual a 1,87.

Problema 8

Continuando com o Problema 6, determine a porcentagem das notas finais do grupo de alunos que estão incluídos em um, dois e três desvios padrão.

R: $\bar{X} \pm 1 \times S=91\%$; $\bar{X} \pm 2 \times S=91\%$ e $\bar{X} \pm 3 \times S=97\%$.

Problema 9

Repita o Problema 8, excluindo as observações 2,2; 20,1; e 31,7.

R: $\bar{X} \pm 1 \times S=66\%$; $\bar{X} \pm 2 \times S=97\%$ e $\bar{X} \pm 3 \times S=100\%$.

Problema 10

Calcule a variância e o desvio padrão dos retornos da tabela seguinte.

Aplicação	Retorno mensal %
Ouro	-1,74%
Curto prazo	0,52%
Dólar paralelo	0,87%
CDB para <\$5.000	1,15%
Caderneta de poupança	1,16%
FRF 30 dias	1,30%
FRF 60 dias	1,49%
CDB para >\$100.000	1,58%
Bolsa RJ	2,12%
Bolsa SP	2,99%

R: $S^2=0,00015$ e $S=1,22\%$

Problema 11

Continuando com o Problema 10, determine a porcentagem dos retornos incluídos em um, dois e três desvios padrão.

R: $\bar{X} \pm 1 \times S=80\%$; $\bar{X} \pm 2 \times S=90\%$ e $\bar{X} \pm 3 \times S=100\%$.

Problema 12

Calcule o coeficiente de variação dos retornos do Problema 10.

R: $CV=1,07$

Problema 13

Os retornos anuais das ações X e Y durante os últimos cinco anos estão registrados na tabela seguinte. Qual dos dois retornos tem maior dispersão?

X	Y
12%	12%
15%	16%
12%	15%
11%	9%
14%	13%

R: A dispersão do retorno da ação Y é maior do que a dispersão da ação X.

Problema 14

Continuando com o Problema 13. Calcule os coeficientes de variação de X e Y. Qual é a ação com maior risco?

R: $CV_X=0,13$ e $CV_Y=0,21$

Problema 15

As taxas de juros cobradas nos empréstimos para compra de eletrodomésticos em oito das maiores lojas da cidade estão registradas na tabela seguinte. Calcule a média, a variância e o desvio padrão das taxas de juros.

6,00%	4,80%	5,30%	4,75%	4,10%	5,40%	3,90%	5,20%
-------	-------	-------	-------	-------	-------	-------	-------

R: $\bar{X}=4,93\%$ $S^2=0,0000482$ e $S=0,69\%$

Problema 16

A tabela seguinte registra uma amostra do tempo que os caixas do banco gastam para realizar as transações dos clientes. Calcule a média, a variância e o desvio padrão da amostra.

2,5	8,0	4,5	7,5	2,0	11,0	4,0	5,0	8,0	6,5	3,5
-----	-----	-----	-----	-----	------	-----	-----	-----	-----	-----

R: $\bar{X}=5,68$ minutos, $S^2=7,61$ e $S=2,76$ minutos

Problema 17

Para conhecer o número de horas por semana que os principais executivos das maiores empresas do país trabalham, a empresa de consultoria realizou uma pesquisa com doze executivos escolhidos alea-

toriamente dentre as 500 maiores empresas. Calcule a média, a variância e o desvio padrão da amostra registrada na tabela a seguir.

60	66	64	62	58	62	62	60	62	60	64	66
----	----	----	----	----	----	----	----	----	----	----	----

R: $\bar{X}=62,17$ hs/sem. $S^2=6,15$ e $S=2,48$ hs/sem.

Problema 18

Ao comparar os retornos de duas ações, a ação que apresentar maior *coeficiente de variação* terá maior risco. A tabela seguinte registra os retornos da Ação A e da Ação B durante cinco anos. Determine a ação com maior risco.

Ação A	Ação B
9,00%	12,00%
10,00%	10,50%
12,00%	9,50%
10,50%	11,00%
9,50%	12,50%

R: A Ação A teve maior *coeficiente de variação* e, portanto, maior risco.

Problema 19

Calcule a variância e o desvio padrão da amostra das notas finais da Turma C da disciplina Estatística registradas no Problema 6.

R: $Var=535,90$ e $S=23,15$

Problema 20

Determine os cinco números: *mínimo*, Q_1 , *mediana*, Q_3 e *máximo* da amostra do Problema 19.

R: $Min=2,20$; $Q_1=74,38$; $Med=83,65$; $Q_3=94,15$; $Max=100$

Problema 21

Construa o *boxplot* do Problema 19.

Problema 22

Com os resultados do Problema 21, analise a distribuição de frequências dessa amostra.

Problema 23

Repita o Problema 22 utilizando o **Modelo Análise Numérica**.

Problema 24

Verifique a existência de dados suspeitos na amostra do Problema 19.

Problema 25

Construa o *boxplot* da amostra do Problema 10, analise a distribuição e verifique a existência de dados suspeitos.

Problema 26

A rede de restaurantes AQUIeAGORA, especializada em almoços pelo sistema *refeição por quilo*, tem 30 lojas distribuídas em diversos bairros de São Paulo, todas com o mesmo padrão e capacidade de atendimento. A tabela seguinte apresenta o número de refeições servidas pelas 30 lojas em um dia típico.

290	243	295	275	216	253
266	232	256	224	252	298
316	247	234	278	270	280
226	233	298	278	266	278
252	269	239	325	240	295

Construa o *boxplot*, analise a distribuição e verifique a existência de dados suspeitos.

Problema 27

Repita o Problema 26 utilizando as vendas das 50 primeiras empresas por vendas em 2002, cujos dados estão registrados na planilha **Problemas** deste capítulo.

Apêndice 1

Funções de medida de dispersão do Excel

O cálculo das medidas de dispersão utilizando o Excel pode ser realizado utilizando expressões matemáticas e procedimentos combinados com os recursos da planilha e funções estatísticas. Na planilha **Funções de Dispersão**, incluída na pasta **Capítulo 4**, está registrada a utilização de cada função utilizando a amostra do Exemplo 4.1, como se pode ver na Figura 4.8. Uma característica comum das funções a seguir são os 30 argumentos (*núm1*; *núm2*; ... ; *núm30*) utilizados para registrar os valores de intervalos. Na apresentação da primeira função DESV.MÉDIO, será mostrado como utilizar esses argumentos, procedimentos que se repetem com as demais funções com o mesmo tipo de argumentos. As sintaxes dessas funções estatísticas são apresentadas a seguir.

DESV.MÉDIO(*núm1*; *núm2*; ... ; *núm30*)

A função estatística DESV.MÉDIO¹⁴ retorna o *desvio absoluto médio* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados.¹⁵ Por exemplo, a função DESV.MÉDIO aplicada aos valores do Exemplo 4.1 retorna o resultado 8,38. Para obter esse resultado, a função DESV.MÉDIO pode ser utilizada das seguintes maneiras:

- Registrando os valores da amostra em um intervalo de células da planilha.
 - Se os valores da variável estiverem registrados em um único intervalo, ou intervalos contíguos, apenas será necessário informar um único intervalo no argumento *num1*. Por exemplo, na célula E6 foi registrada a fórmula =DESV.MÉDIO(B4:B14), conforme apresenta a Figura 4.8.
 - Se os valores da variável estiverem registrados em intervalos não adjacentes, será necessário informar o endereço de cada intervalo no lugar de cada *núm* de *núm1*; *núm2*; ... ; *núm30*, até um máximo de 30; por exemplo, na célula E7 foi registrada a seguinte fórmula =DESV.MÉDIO(B4:B7;B8:B12;B13:B14).
- Registrando os valores da amostra como *matriz* na própria fórmula da função, evitando registrar os valores da amostra em um intervalo de células da planilha.
 - Na célula F6, os valores foram registrados em uma única matriz:
=DESV.MÉDIO({31;38;19;27;24;42;32;18;43;15;39})
 - Na célula F7, os valores foram registrados em três matrizes:
=DESV.MÉDIO({31;38;19};{27;24;42;32;18;43};{15;39}) correspondentes aos três primeiros argumentos da função DESV.MÉDIO *núm1*; *núm2*; *núm3*.

DESVQ(*núm1*; *núm2*; ... ; *núm30*)

A função estatística DESVQ¹⁶ retorna a *soma dos quadrados dos desvios* dos valores numéricos *núm1*; *núm2*; ... ; *núm30* com relação à média. Cada um desses *núm* pode ser um intervalo de células de uma

¹⁴ Em inglês, a função DESV.MÉDIO é AVEDEV.

¹⁵ Assemelhados são os intervalos definidos por nomes, valores lógicos, representações em forma de texto de números, por exemplo, com a função de texto VALOR("10")=10.

¹⁶ Em inglês, a função DESVQ é DEVSQ.

planilha contendo valores numéricos ou assemelhados. A função DESVQ pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO, mencionada anteriormente, conforme mostrado na Figura 4.8.

VARP(*núm1*; *núm2*; ... ; *núm30*)

A função estatística VARP¹⁷ retorna a *variância da população* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha contendo valores numéricos ou assemelhados. A função VARP pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO citada anteriormente.

VAR(*núm1*; *núm2*; ... ; *núm30*)

A função estatística VAR¹⁸ retorna a *variância da amostra* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha que contém valores numéricos ou *assemelhados*. A função VAR pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO definida anteriormente.

DESVPADP(*núm1*; *núm2*; ... ; *núm30*)

A função estatística DESVPADP¹⁹ retorna o *desvio padrão da população* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha que contém valores numéricos ou assemelhados. A função DESVPADP pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO mencionada anteriormente.

DESVPAD(*núm1*; *núm2*; ... ; *núm30*)

A função estatística DESVPAD²⁰ retorna o *desvio padrão da amostra* dos valores numéricos *núm1*; *núm2*; ... ; *núm30*. Cada um desses *núm* pode ser um intervalo de células de uma planilha que contém valores numéricos ou assemelhados. A função DESVPADP pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO detalhada anteriormente.

VARPA(*núm1*; *núm2*; ... ; *núm30*)

A função estatística VARPA²¹ é equivalente à função anterior VARP. A diferença está relacionada com os valores registrados nos argumentos *núm1*; *núm2*; ... ; *núm30* que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO.

VARA(*núm1*; *núm2*; ... ; *núm30*)

A função estatística VARPA²² é equivalente à função anterior VAR. A diferença está relacionada com os valores registrados nos argumentos *núm1*; *núm2*; ... ; *núm30* que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO.

¹⁷ Em inglês, a função VARP é VARP.

¹⁸ Em inglês, a função VAR é VAR.

¹⁹ Em inglês, a função DESVPADP é STDEVP.

²⁰ Em inglês, a função DESVPAD é STDEV.

²¹ Em inglês, a função VARPA é VARPA.

²² Em inglês, a função VARA é VARA.

FIGURA 4.8 Aplicando as funções de medidas de dispersão no Exemplo 4.1.

	A	B	C	D	E	F
1	Funções de medição de dispersão					
2						
3		Amostra			Dados informados como	
4		31		Funções Estatísticas	Intervalo	Matriz
5		38				
6		19		DESV.MÉDIO	8,38	8,38
7		27			8,38	8,38
8		24				
9		42		DESV0	997,64	997,64
10		32			997,64	997,64
11		18				
12		43		VARP	90,69	90,69
13		15			90,69	90,69
14		39				
15				VAR	99,76	99,76
16					99,76	99,76
17						
18				DESVADP	9,52	9,52
19					9,52	9,52
20						
21				DESVAD	9,99	9,99
22					9,99	9,99
23						
24				DISTORÇÃO	(0,12)	(0,12)
25					(0,12)	(0,12)
26						
27				CURT	(1,47)	(1,47)
28					(1,47)	(1,47)
29						

DESVADPA(núm1; núm2; ... ; núm30)

A função estatística DESVPADPA²³ é equivalente à função anterior DESVPADP. A diferença está relacionada com os valores registrados nos argumentos núm1; núm2; ... ; núm30 que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO.

DESVADA(núm1; núm2; ... ; núm30)

A função estatística DESVADA²⁴ é equivalente à função anterior DESVPAD. A diferença está relacionada com os valores registrados nos argumentos núm1; núm2; ... ; núm30 que, nesta função, além de números, podem ser valores lógicos e de texto, como VERDADEIRO e FALSO.

DISTORÇÃO(núm1; núm2; ... ; núm30)

A função estatística DISTORÇÃO²⁵ retorna o *coeficiente de inclinação* dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha que contém valores numéricos ou *assemelhados*. A fórmula utilizada pela função DISTORÇÃO para calcular o coeficiente de inclinação é:

$$\text{Coeficiente de Inclinação} = \frac{n}{(n-1) \times (n-2)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right)^3$$

²³ Em inglês, a função DESVPADPA é STDEVPA.

²⁴ Em inglês, a função DESVADA é STDEVA.

²⁵ Em inglês, a função DISTORÇÃO é SKEW.

O coeficiente de inclinação é o resultado da comparação da distribuição de frequências dos valores informados com a distribuição normal, apresentada no Capítulo 8, e seu resultado deve ser interpretado como segue. Se o coeficiente de inclinação for igual a zero, então a distribuição de frequências é simétrica, se for negativo, a distribuição de frequências terá inclinação para a esquerda ou negativa, e se for positivo, a distribuição de frequências terá inclinação para a direita ou positiva. A função DISTORÇÃO pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO, definida anteriormente.

CURT(núm1; núm2; ... ; núm30)

A função estatística CURT²⁶ retorna o *coeficiente de curtose* dos valores numéricos núm1; núm2; ... ; núm30. Cada um desses núm pode ser um intervalo de células de uma planilha que contém valores numéricos ou *assemelhados*. A fórmula utilizada pela função CURT para calcular o coeficiente de curtose é a seguinte:

$$\text{Coeficiente de Curtose} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

O coeficiente de curtose é o resultado da comparação da distribuição de frequências dos valores informados com a distribuição normal apresentada no Capítulo 8, e seu resultado deve ser interpretado como segue. Se o coeficiente de curtose for igual a zero, então a distribuição de frequências será a própria distribuição normal; se for negativo, a distribuição será achatada, plana; e se for positivo, a distribuição de frequências será concentrada ao redor da média, distribuição com pico. A função CURT pode ser registrada de diversas formas equivalentes às descritas na função DESV.MÉDIO mencionada anteriormente.

Apêndice 2

Outra forma de calcular a variância

O cálculo da variância da variável X pode ser realizado utilizando apenas os valores da variável, sem necessidade de calcular a média e os desvios da variável. Se na fórmula da soma dos quadrados dos desvios desenvolvemos o quadrado do binômio indicado, obtemos a seguinte igualdade:

$$\sum_{i=1}^N (X_i - \mu_X)^2 = \sum_{i=1}^N (X_i^2 - 2X_i\mu_X + \mu_X^2)$$

Continuando com o desenvolvimento algébrico, obtemos:

$$\sum_{i=1}^N (X_i - \mu_X)^2 = \sum_{i=1}^N X_i^2 - 2\mu_X \sum_{i=1}^N X_i + \sum_{i=1}^N \mu_X^2$$

²⁶ Em inglês, a função CURT é KURT.

No segundo membro dessa expressão reconhecemos que $\sum_{i=1}^N X_i = N\mu_X$ e $\sum_{i=1}^N \mu_X^2 = N\mu_X^2$. Dessa maneira, o segundo membro pode ser reescrito da seguinte forma: $\sum_{i=1}^N X_i^2 - 2\mu_X N\mu_X + \mu_X^2 = \sum_{i=1}^N X_i^2 - N\mu_X^2$.

Voltando à primeira fórmula, formamos a igualdade que nos interessa:

$$\sum_{i=1}^N (X_i - \mu_X)^2 = \sum_{i=1}^N X_i^2 - N\mu_X^2$$

Ainda, pela definição de média da população $\mu_X = \frac{\left(\sum_{i=1}^N X_i\right)}{N}$. Substituindo essa relação na expressão da soma dos quadrados dos desvios, teremos:

$$\sum_{i=1}^N (X_i - \mu_X)^2 = \sum_{i=1}^N X_i^2 - N = \frac{\left(\sum_{i=1}^N X_i\right)^2}{N^2}$$

Agora, o cálculo da soma dos quadrados dos desvios depende somente dos dados da amostra e dos quadrados desses dados. Dessa maneira, as expressões das variâncias são:

- Da população: $\sigma_X^2 = \frac{1}{N} \left\{ \sum_{i=1}^N X_i^2 - \frac{\left(\sum_{i=1}^N X_i\right)^2}{N} \right\}$
- Da amostra: $S_X^2 = \frac{1}{n-1} \left\{ \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \right\}$

Para calcular a variância, será necessário gerar a série dos quadrados dos valores da variável, não sendo necessário calcular a média nem os desvios. Na realidade, esse procedimento de cálculo perde sua força quando comparado com a utilização das funções estatísticas do Excel, como mostra a planilha **Apêndice 2**, incluída na pasta **Capítulo 4**. Essa expressão da variância será utilizada no Apêndice 1 do Capítulo 9.

Apêndice 3

Funções para banco de dados do Excel

As funções estatísticas apresentadas até este momento foram utilizadas para obter alguma medida estatística de uma amostra ou variável, atendendo a algumas especificações dessas funções:

- Os dados foram registrados em um intervalo de células da planilha e a fórmula com a função em outra célula fora daquele intervalo.
- Os dados da amostra foram registrados como matriz na própria fórmula da função em uma única célula da planilha.

Há situações em que os dados ou variáveis para análise fazem parte de uma tabela contendo outras variáveis. Por exemplo, os resultados mensais significativos de uma empresa durante um ano estão registrados na planilha **Funções Banco de Dados**, incluída na pasta **Capítulo 4**, conforme apresenta a Figura 4.9. Os resultados estão registrados em uma tabela com as colunas identificadas com os nomes *Mês*, *Vendas*, *Custos*, *Lucro Bruto* e *Lucro Líquido*. A tabela com os resultados da empresa é denominada *banco de dados*, e cada uma de suas colunas é denominada *campo*; em termos técnicos, cada linha da tabela é uma unidade elementar de informação que contém quatro variáveis. Para essas situações, o Excel dispõe de funções denominadas genericamente *BDfunções* e equivalentes a algumas das funções apresentadas neste capítulo.

	A	B	C	D	E	F
1	Funções estatísticas para banco de dados					
2						
3						
4		Mês	Vendas	Custos	Lucro Bruto	Lucro Líquido
5		jan/2004	\$ 6.423	\$ 3.270	\$ 3.153	\$ 2.193
6		fev/2004	\$ 5.467	\$ 3.649	\$ 1.818	\$ 1.198
7		mar/2004	\$ 5.191	\$ 3.381	\$ 1.810	\$ 1.156
8		abr/2004	\$ 6.315	\$ 3.513	\$ 2.802	\$ 1.844
9		mai/2004	\$ 6.080	\$ 3.316	\$ 2.764	\$ 1.754
10		jun/2004	\$ 6.195	\$ 3.768	\$ 2.427	\$ 1.664
11		jul/2004	\$ 6.131	\$ 3.564	\$ 2.567	\$ 1.746
12		ago/2004	\$ 6.386	\$ 3.258	\$ 3.128	\$ 2.072
13		set/2004	\$ 6.014	\$ 3.122	\$ 2.892	\$ 1.931
14		out/2004	\$ 5.993	\$ 3.706	\$ 2.287	\$ 1.522
15		nov/2004	\$ 5.237	\$ 3.533	\$ 1.704	\$ 1.129
16		dez/2004	\$ 5.374	\$ 3.115	\$ 2.259	\$ 1.457

FIGURA 4.9 Resultados mensais da empresa.

EXEMPLO 4.11

Calcule a média, o desvio padrão e o valor máximo das *Vendas* da empresa durante o primeiro mês dos quatro trimestres do ano 2004, e cujos resultados estão registrados na tabela da Figura 4.9.

Solução. Os resultados foram obtidos de diversas formas, a partir da célula H1 da planilha **Funções Banco de Dados**, incluída na pasta **Capítulo 4**, como mostra a figura seguinte. A média das vendas da empresa nos primeiros meses dos quatro trimestres do ano 2004 é igual a \$6.215,50, resultado obtido:

- Calculando com a função estatística MÉDIA, registrando na célula K4 a fórmula =MÉDIA(C4;C7;C10;C13).
- Calculando com a função estatística SUBTOTAL, registrando na célula K5 a fórmula =SUBTOTAL(1;C4;C7;C10;C13). Com a função SUBTOTAL, é possível obter 11 resultados diferentes informando um número de 1 a 11 no primeiro argumento da função, como será apresentado mais adiante neste apêndice.
- Calculando com a função estatística para banco de dados BDMÉDIA, registrando na célula K6 a fórmula =BDMÉDIA(B3:F15;C3;I3:I7). A função BDMÉDIA é uma das doze funções para listas ou banco de dados disponíveis no Excel e denominadas genericamente BDFunções, pois todas elas utilizam a mesma sintaxe, BDFunção(*banco_dados; campo; critérios*).
 - No argumento *banco_dados*, deve ser informado o intervalo do banco de dados incluindo a primeira linha com os títulos, neste exemplo B3:F15.
 - No argumento *campo*, deve ser informado o nome da coluna do banco de dados onde será aplicada a função. Neste exemplo, pode ser informado o texto "Vendas", entre aspas duplas, ou o endereço da célula C3.
 - No argumento *critérios*, deve ser registrada a especificação da escolha dos dados. Neste exemplo, no intervalo I3:I7 foi construída a tabela de meses, ou linhas, que identificam os valores correspondentes da coluna *Vendas* o argumento *campus* da função. Como alternativa, pode-se utilizar a fórmula =BDMÉDIA(B3:F15;"Vendas";I3:I7) para obter o mesmo resultado Ou, como um número que represente a posição da coluna dentro da lista, começando com 1 para a primeira coluna, 2 para a segunda coluna e assim sucessivamente, até esgotar as colunas do banco de dados.

	H	I	J	K	L	M
1	Exemplo 4.11					
2						
3		Mês		Cálculo da média		
4		jan/2004		\$ 6.215,50	=MÉDIA(C4;C7;C10;C13)	
5		abr/2004		\$ 6.215,50	=SUBTOTAL(1;C4;C7;C10;C13)	
6		jul/2004		\$ 6.215,50	=BDMÉDIA(B3:F15;C3;I3:I7)	
7		out/2004				
8				Cálculo do desvio padrão, como população		
9				\$ 165,53	=DESVPA(D4;D7;D10;D13)	
10				\$ 165,53	=SUBTOTAL(8;C4;C7;C10;C13)	
11				\$ 165,53	=BDESVPA(B3:F15;C3;I3:I7)	
12						
13				Determinação do valor máximo		
14				\$ 6.423,00	=MÁXIMO(C4;C7;C10;C13)	
15				\$ 6.423,00	=SUBTOTAL(4;C4;C7;C10;C13)	
16				\$ 6.423,00	=BDMÁX(B3:F15;C3;I3:I7)	
17						
18				\$ 6.423,00	=BDMÁX(B3:F15;"Vendas";I3:I7)	
19				\$ 6.423,00	=BDMÁX(B3:F15;2;I3:I7)	
20						

A partir das linhas 8 e 13 da planilha **Funções Banco de Dados**, foram calculados, respectivamente, o desvio padrão e o valor máximo das *Vendas* da empresa durante o primeiro mês dos quatro trimestres do ano 2004, utilizando as três funções apresentadas e adequadas para esses cálculos.

Incluindo outros critérios

Com as funções para banco de dados operamos a distância sem necessidade de definir intervalos dentro do banco de dados. A tabela de *critérios* pode incluir condições lógicas nos *campos* do *banco de dados*. Sem esgotar este assunto, a seguir mostraremos outra forma de incluir critérios.

EXEMPLO 4.12

Calcule a média das vendas da empresa durante o primeiro mês dos quatro trimestres do ano 2004, considerando somente os meses com lucro líquido maior ou igual a \$1.600.

Solução. Para calcular a média das vendas dos primeiros meses dos quatro trimestres do ano 2004, considerando apenas as vendas dos meses com *Lucro Líquido* igual ou maior do que \$1.600, no intervalo O3:P7, foi construída a tabela com os campos *Mês* e *Lucro Líquido*, sendo que, neste último cálculo, foram registradas as restrições de seleção de cada mês, a fórmula ≥ 1600 . A média das vendas da empresa durante o primeiro mês dos quatro trimestres do ano 2004, considerando somente os meses com lucro líquido maior ou igual a \$1.600 é igual a \$6.289,67, resultado obtido com $=BDMÉDIA(B3:F15;C3;O3:P7)$, fórmula registrada na célula R4.

	N	O	P	Q	R	S	T
1	Exemplo 4.12						
2							
3	Mês		Lucro Líquido	Cálculo da média, com restrições			
4	jan/2004		>=1600	\$ 6.289,67	=BDMÉDIA(B3:F15;C3;O3:P7)		
5	abr/2004		>=1600				
6	jul/2004		>=1600				
7	out/2004		>=1600				
8							

Resumo das funções de banco de dados do Excel

O Excel dispõe de doze funções orientadas para banco de dados, denominadas genericamente **BDfunções**, pois cada uma dessas funções tem os mesmos três argumentos: *banco de dados*, *campo* e *critérios*. Sua sintaxe geral é:

BDfunção(banco_dados; campo; critérios)

- O argumento *banco_dados* é o intervalo de células que delimita a tabela com as informações, que pode ser uma lista ou um banco de dados. Um banco de dados é uma lista de dados na qual cada linha é um registro formado por um ou mais campos identificados por um nome na primeira linha de cada coluna. O argumento *banco_dados* pode ser informado como um intervalo de células ou como um nome representando o intervalo.
- O argumento *campo* define o nome da coluna do banco de dados que será utilizada para realizar um cálculo ou uma seleção, podendo ser informado:
 - Como texto, por exemplo, “Vendas” ou “Lucro Líquido”.
 - Como endereço da célula onde está registrado nome do campo.
 - Como um número que represente a posição da coluna dentro da lista, começando com 1 para a primeira coluna, 2 para a segunda coluna e assim sucessivamente, até esgotar as colunas do banco de dados.
- O argumento *critérios* é o intervalo de células que especifica a forma de seleção. Pode ser informado qualquer intervalo, sempre que ele incluir pelo menos um título de coluna e ao menos uma célula abaixo desse título que especifique alguma condição para seleção nessa coluna.

A seguir, são apresentadas as sintaxes das doze funções para bancos de dados disponíveis no Excel. As primeiras onze funções foram registradas com o mesmo argumento (B3:F15;C3;I3:I7) no intervalo K23:L34 da planilha **Funções Banco de Dados**, incluída na pasta **Capítulo 4**, cujos resultados são mostrados na Figura 4.10.

BDMÉDIA(*banco_dados; campo; critérios*)

A função BDMÉDIA²⁷ retorna a média dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDCONTAR(*banco_dados; campo; critérios*)

A função BDCONTAR²⁸ retorna a quantidade de células contendo números da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDCONTARA(*banco_dados; campo; critérios*)

A função BDCONTARA²⁹ retorna a quantidade de células não vazias da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados

BDMÁX(*banco_dados; campo; critérios*)

A função BDMÁX³⁰ retorna o valor máximo da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDMÍN(*banco_dados; campo; critérios*)

A função BDMÍN³¹ retorna o valor mínimo da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDMULTIPL(*banco_dados; campo; critérios*)

A função BDMULTIPL³² retorna o resultado da multiplicação dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDEST(*banco_dados; campo; critérios*)

A função BDEST³³ retorna o desvio padrão da amostra dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDDSVPA(*banco_dados; campo; critérios*)

A função BDDSVPA³⁴ retorna o desvio padrão da população dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDSOMA(*banco_dados; campo; critérios*)

A função BDSOMA³⁵ retorna a soma dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

²⁷ Em inglês, a função BDMÉDIA é DAVERAGE.

²⁸ Em inglês, a função BDCONTAR é DCOUNT.

²⁹ Em inglês, a função BDCONTARA é DCOUNTA.

³⁰ Em inglês, a função BDMÁX é DMAX.

³¹ Em inglês, a função BDMÍN é DMIN.

³² Em inglês, a função BDMULTIPL é DPRODUCT.

³³ Em inglês, a função BDEST é DSTDEV.

³⁴ Em inglês, a função BDDSVPA é DSTDEVP.

³⁵ Em inglês, a função BDSOMA é DSUM.

	H	I	J	K	L	M	N	O
21	Bdfunções							
22								
23		Mês		Função	Resultado			
24		jan/2004		BDMÉDIA	\$ 6.215,50	=BDMÉDIA(B3:F15;C3;I23:I27)		
25		abr/2004		BDCONTAR	4	=BDCONTAR(B3:F15;C3;I23:I27)		
26		jul/2004		BDCONTARA	4	=BDCONTARA(B3:F15;C3;I23:I27)		
27		out/2004		BDMÁX	\$ 6.423	=BDMÁX(B3:F15;C3;I23:I27)		
28				BDMÍN	\$ 5.993	=BDMÍN(B3:F15;C3;I23:I27)		
29				BDMULTIPL	1,49035E+15	=BDMULTIPL(B3:F15;C3;I23:I27)		
30				BDEST	\$ 191,14	=BDEST(B3:F15;C3;I23:I27)		
31				BDDESIPA	\$ 165,53	=BDDESIPA(B3:F15;C3;I23:I27)		
32				BDSOMA	\$ 24.862	=BDSOMA(B3:F15;C3;I23:I27)		
33				BDVAREST	36.534,33	=BDVAREST(B3:F15;C3;I23:I27)		
34				BDVARP	27.400,75	=BDVARP(B3:F15;C3;I23:I27)		
35								
36				BDEXTRAIR	#NÚM!	=BDEXTRAIR(B3:F15;C3;I23:I27)		
37				BDEXTRAIR	\$ 6.423	=BDEXTRAIR(B3:F15;C3;I23:I24)		
38								

FIGURA 4.10

Aplicação das Bdfunções.

BDVAREST(*banco_dados; campo; critérios*)

A função BDVAREST³⁶ retorna a variância da amostra dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDVARP(*banco_dados; campo; critérios*)

A função BDVARP³⁷ retorna a variância da população dos valores da coluna *campo* do *banco_dados* que coincide com os *critérios* especificados.

BDEXTRAIR(*banco_dados; campo; critérios*)

A função BDEXTRAIR³⁸ extrai do *banco_dados* um único registro da coluna *campo* que coincide com os *critérios* especificados. A seguir, apresentamos como se deve utilizar essa função:

- A fórmula =BDEXTRAIR(B3:F15;C3;I23:I27) registrada na célula L36 retorna como resultado o valor de erro #NÚM!, conforme mostrado na Figura 4.10. Isso ocorre porque a função BDEXTRAIR não consegue identificar um valor único no intervalo I23:I27, no qual há quatro datas possíveis.
- A fórmula =BDEXTRAIR(B3:F15;C3;I23:I24) registrada na célula L37 retorna o resultado \$6.423, pois no intervalo I23:I24 há apenas uma única data Jan/2004.

Outras funções do Excel

O Excel dispõe também das funções matemáticas SUBTOTAL, CONT.SE e SOMASE que realizam operações equivalentes às apresentadas para banco de dados.

SUBTOTAL(*número_função; ref1; ref2; ...; ref29*)

A função SUBTOTAL³⁹ retorna o resultado das primeiras onze funções do grupo de Bdfunções. O argumento *número_função* é um número de 1 a 11 que identifica a função que deverá ser utilizada no cálculo de subtotais do banco de dados, de uma lista ou grupo de valores, como mostra a Figura 4.11. Os argumentos *ref1; ref2; ...; ref29* são intervalos de células de uma planilha, ou referências, sobre os quais será calculado o subtotal.

³⁶ Em inglês, a função BDVAREST é DVAR.

³⁷ Em inglês, a função BDVARP é DVARP.

³⁸ Em inglês, BDEXTRAIR é DGET.

³⁹ Em inglês, SUBTOTAL é SUBTOTAL.

<i>número_função</i>	Função equivalente
1	MÉDIA
2	CONT.NÚM
3	CONT.VALORES
4	MÁXIMO
5	MÍNIMO
6	MULT
7	DESVPAD
8	DESVPADP
9	SOMA
10	VAR
11	VARP

FIGURA 4.11 Significado do argumento *número_função*.

A Figura 4.12 mostra os onze resultados possíveis da função SUBTOTAL, registrados a partir da célula J39 da planilha **Funções Banco de Dados**, incluída na pasta **Capítulo 4**. Por exemplo, para calcular a média das vendas da empresa do Exemplo 4.11 referentes aos primeiros meses dos quatro trimestres do ano 2004, na célula L42 foi registrada a fórmula =SUBTOTAL(1;C4;C7;C10;C13), cujo resultado é \$6.215,50.

O leitor atento deve ter percebido que a função SUBTOTAL pode ser utilizada como substituta de algumas das funções básicas apresentadas nos Capítulos 3 e 4 do livro. Como ajuda, a partir da célula H10 da planilha **Funções de Dispersão**, incluída na pasta **Capítulo 4**, foram registradas fórmulas utilizando a função SUBTOTAL ao lado da função equivalente original. Uma vantagem da utilização da função SUBTOTAL é que com um único nome de função poderíamos agrupar onze funções, com a vantagem de ter de lembrar a tabela de equivalência da Figura 4.11, que também não é muito amigável.

	J	K	L	M
39	Função SUBTOTAL			
40				Função
41		Número	Resultado	Equivalente
42		1	\$ 6.215,50	BDMÉDIA
43		2	4	BDCONTAR
44		3	4	BDCONTARA
45		4	\$ 6.423	BDMÁX
46		5	\$ 5.993	BDMÍN
47		6	1,49035E+15	BDMULTIPL
48		7	\$ 191,14	BDEST
49		8	\$ 165,53	BDESVPA
50		9	\$ 24.862,00	BDSOMA
51		10	36.534,33	BDVAREST
52		11	27.400,75	BDVARP
53				

FIGURA 4.12
Resultados com a função
SUBTOTAL.

CONT.SE(*intervalo*; *critérios*)

A função CONT.SE⁴⁰ retorna o número de células não vazias da série de dados definida no argumento *intervalo* e que atendem a *critérios* definidos em forma de texto. Por exemplo, gostaríamos de conhecer, na tabela de resultados da Figura 4.9, em quantos meses do ano 2004 o lucro líquido da empresa foi igual ou maior do que \$1.500. O resultado foi obtido com a função CONT.SE, registrando a fórmula

⁴⁰ Em inglês, CONT.SE é COUNTIF.

=CONT.SE(F4:F15;">=1500") na célula K57 da planilha **Funções Banco de Dados**. Portanto, em oito meses do ano 2004, a empresa registrou lucro líquido igual ou maior do que \$1.500.

SOMASE(*intervalo; critérios; intervalo_soma*)

A função SOMASE⁴¹ retorna a soma de valores das células que atendem a um determinado critério.

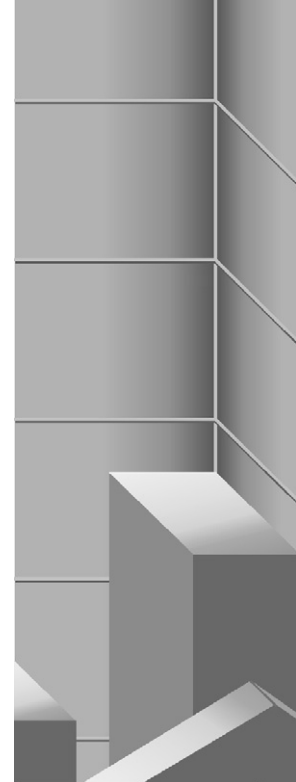
- No argumento *intervalo* é registrado o intervalo de células utilizado para aplicar o critério de seleção.
- No argumento *critérios* é registrado um número, expressão ou texto, que define como as células serão selecionadas.
- No argumento *intervalo_soma* é registrado o intervalo das células que poderão ser somadas, sendo somadas somente as células correspondentes ao argumento *intervalo* que atendam ao argumento *critérios*. Se *intervalo_soma* for omitido, serão somadas as células do argumento *intervalo*.

Por exemplo, gostaríamos de conhecer, da empresa cujos resultados estão registrados na tabela de resultados da Figura 4.9, o total das vendas com lucro líquido igual ou maior do que \$2.000 durante o ano 2004. O resultado foi obtido com a fórmula =SOMASE(F4:F15;">=2000";C4:C15) registrada na célula K62 da planilha **Funções Banco de Dados**. Portanto, o total das vendas com lucro líquido igual ou maior do que \$2.000 durante o ano 2004 foi \$12.809.

⁴¹ Em inglês, SOMASE é SUMIF.

Capítulo 5

PROBABILIDADE



Os quatro primeiros capítulos apresentaram os temas amostragem, descrição gráfica de amostras e medidas numéricas de posição, de tendência central e de dispersão ou variabilidade. Esses conhecimentos permitem analisar séries de dados e obter algumas conclusões sobre como esses dados se distribuem em todo seu intervalo de variação ou ao redor de sua média. O tema deste capítulo ajudará a descrever a informação amostrada, facilitará a apresentação desses resultados e outorgará uma ferramenta útil para realizar inferências sobre a população de onde foi extraída a amostra.

Pela própria experiência de vida, sabemos que o resultado do lançamento de uma moeda pode ser cara ou coroa, descartando a moeda falsa com duas caras, ou duas coroas, ou aquela que possa ficar de pé apoiada na sua borda. Também, periodicamente recebemos informações como a seguinte: na pesquisa de intenção de voto para o segundo turno da eleição para governador, 43% dos eleitores da amostra preferem o candidato A, 37% dos eleitores preferem o candidato B e os demais 20% dos eleitores não sabem. Qual a característica comum do lançamento de uma moeda e da pesquisa de intenção de voto? O resultado não pode ser previsto com antecedência! Por quê? Porque o resultado variará toda vez que lançarmos uma moeda ou extrairmos outra amostra para a pesquisa de intenção de voto.

Entretanto, se o lançamento da moeda for repetido um número muito grande de vezes, perceberemos uma tendência dos resultados. O gráfico da Figura 5.1, um dos muitos gráficos possíveis, representa a proporção de caras em uma simulação de 1.500 lançamentos de uma moeda. O resultado dessa simulação em particular mostra que a proporção de caras tende a 50%, lembrando que esse gráfico foi especialmente escolhido para essa apresentação, pois, tecnicamente, a simulação de 1.500 lançamentos é um número pequeno de tentativas.

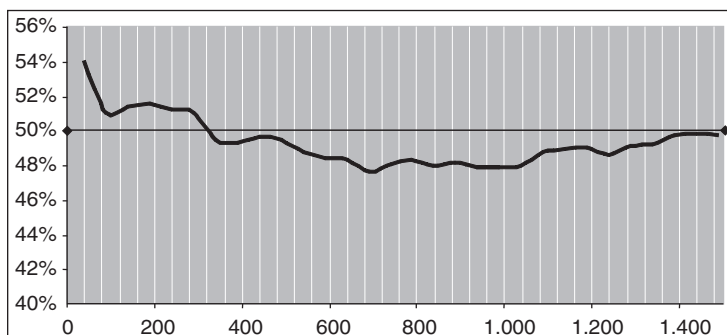


FIGURA 5.1 Proporção de caras no lançamento de uma moeda 1.500 vezes.

Da mesma maneira, se a pesquisa de intenção de voto fosse repetida para um número muito grande de amostras diferentes, também perceberemos uma tendência dos resultados do candidato A e do candidato B. Nos exemplos apresentados, destacam-se dois pontos:

- O lançamento da moeda e a pesquisa de intenção de voto são experimentos aleatórios. Embora os resultados de um experimento aleatório sejam incertos, a longo prazo os resultados têm uma distribuição de frequências definida.
- Depois de repetir um experimento aleatório um número muito grande de vezes, a proporção de ocorrência de um dos resultados é denominada *probabilidade*.¹

A determinação da probabilidade de um dos resultados possíveis de um experimento repetindo-o um número muito grande de vezes não é um procedimento geral, além de ser trabalhoso e dispendioso. O primeiro passo será apresentar o resumo dos conceitos que ajudarão a estabelecer regras gerais.

Experimentos e eventos

Todo processo desenvolvido para realizar observações e obter dados com um determinado objetivo é denominado experimento. O conjunto formado por todos os resultados possíveis de um experimento é denominado espaço amostral do experimento. Um experimento é aleatório quando pode resultar em um dos resultados do espaço amostral sem que se seja possível prever com certeza qual resultado será observado.

Se apesar de conhecer todos os resultados de um experimento não for possível antecipar seu resultado, esse experimento é denominado *experimento aleatório*.

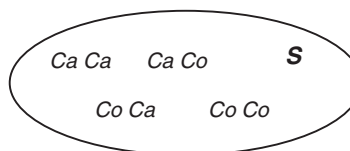
Espaço amostral é o conjunto de todos os possíveis e diferentes resultados de um experimento aleatório.

A análise de um experimento aleatório começa pela identificação de todos os resultados possíveis. Por exemplo, no experimento do lançamento de duas moedas seu espaço amostral é formado pelos quatro resultados possíveis *CaCa*, *CaCo*, *CoCa* e *CoCo*, ou o conjunto *S* dos resultados possíveis $S = \{CaCa, CaCo, CoCa, CoCo\}$. Cada resultado desse espaço amostral *S* é denominado *ponto amostral*.

Eventos

O *diagrama de Venn* é uma forma gráfica de representar o espaço amostral *S*. A Figura 5.2 mostra o diagrama de Venn do espaço amostral *S* do lançamento de duas moedas, o conjunto $S = \{CaCa, CaCo, CoCa, CoCo\}$

FIGURA 5.2 Diagrama de Venn do lançamento de duas moedas.



¹ Do dicionário Houaiss. *Probabilidade*: 1. perspectiva favorável de que algo venha a ocorrer; possibilidade, chance. 2. grau de segurança com que se pode esperar a realização de um evento, determinado pela frequência relativa dos eventos do mesmo tipo numa série de tentativas.

Do lançamento de duas moedas, sempre será obtido um único resultado denominado *evento elementar* do espaço amostral S . Os quatro elementos do espaço amostral S são *eventos elementares*, pois nenhum deles pode ser particionado ou dividido.

Evento elementar é um resultado único do espaço amostral.

Evento é um subconjunto formado por um ou mais resultados do espaço amostral.

Um subconjunto do espaço amostral S é denominado evento. Por exemplo, o *evento dos resultados que têm exatamente apenas uma cara* é descrito pelo subconjunto do espaço amostral $A = \{CaCo, CoCa\}$, como mostra o *diagrama de Venn* da Figura 5.3. Tenha em mente que um evento pode ser particionado, dividido, em seus eventos elementares.

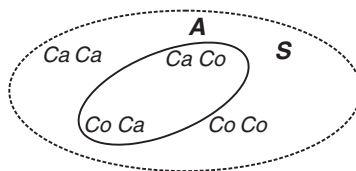


FIGURA 5.3 Diagrama de Venn do evento A .

Operações com eventos

A primeira operação é o *complemento* de um evento. Por exemplo, o complemento do evento A é o subconjunto B formado pelos elementos do espaço amostral não incluídos no evento A . Dessa maneira, o complemento do evento $A = \{CaCo, CoCa\}$ é o evento $B = \sim A = \{CaCa, CoCo\}$, como mostra o diagrama de Venn da Figura 5.4.

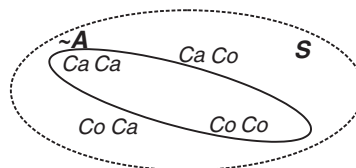


FIGURA 5.4 Diagrama de Venn do complemento de A .

Outras duas operações importantes são a *união* e a *interseção*. Dois ou mais eventos do mesmo espaço amostral podem ser agrupados em operações de união e interseção, como mostra a Figura 5.5. Nos eventos A e B pertencentes ao mesmo espaço amostral S :

- A operação *interseção* dos eventos A e B gera um novo evento formado pelos elementos comuns aos dois conjuntos. Essa operação é representada com $A \cap B$, onde o símbolo \cap representa a operação interseção.
- A operação *união* dos eventos A e B gera um novo evento formado pelos elementos comuns e não comuns dos dois conjuntos. Essa operação é representada com $A \cup B$, onde o símbolo \cup representa a operação união.

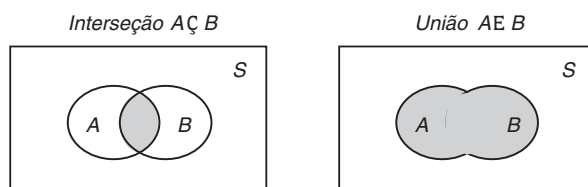


FIGURA 5.5 Operações com eventos.

Algumas conclusões das operações com eventos:

- A união de um evento A e seu complemento $\sim A$ é o próprio espaço amostral S , ou com símbolos $A \cup \sim A = S$.
- A interseção de um evento A e seu complemento $\sim A$ é o conjunto vazio \emptyset , ou com símbolos $A \cap \sim A = \emptyset$.

Eventos mutuamente excludentes e coletivamente exaustivos

Os resultados possíveis do lançamento de uma moeda são apenas dois, os eventos elementares *Cara-Ca* e *Coroa-Co*. Pela própria característica do experimento, se o resultado de um lançamento for cara, esse resultado não poderá ser coroa ao mesmo tempo, pois são *eventos mutuamente excludentes*. A união de eventos elementares forma o espaço amostral, pois são *eventos coletivamente exaustivos*. Portanto, verifica-se que os eventos A e B pertencentes ao mesmo espaço amostral S :

- São mutuamente excludentes se sua interseção for vazia: $A \cap B = \emptyset$, pois os dois eventos não têm nenhum elemento em comum.
- São coletivamente exaustivos se a união dos eventos formarem o espaço amostral: $A \cup B = S$, onde cada evento pode ter elementos repetidos no outro evento.

EXEMPLO 5.1

Análise os resultados do lançamento de uma moeda.

Solução. Como o espaço amostral do lançamento de uma moeda tem apenas dois eventos, os eventos elementares Ca e Co são eventos mutuamente excludentes, eventos complementares e eventos coletivamente exaustivos.

EXEMPLO 5.2

A nota final do curso de *estatística* pode ser: conceito A , conceito B ou conceito C . Analise os resultados dessas notas.

Solução. O espaço amostral da nota final de *estatística* é formado por três eventos elementares: conceito A , conceito B e conceito C . Os três conceitos são eventos mutuamente excludentes e coletivamente exaustivos, pois quando agrupados formam o espaço amostral de todos os conceitos. Não são eventos complementares, pois o complemento do conceito A é a união do conceito B e do conceito C .

Probabilidade

Depois de apresentar os conceitos de experimento e eventos, o objetivo é dirigido para a avaliação do sucesso de ocorrer um determinado evento do espaço amostral de um experimento aleatório. Por exemplo, no lançamento de uma moeda, um número muito grande de vezes, o sucesso de ocorrer o evento *Cara* é medido pela probabilidade $P(\text{Cara})$, um valor dentro do intervalo $(0, 1)$, incluindo ambos os limites.

A probabilidade de sucesso $P(A)$ do evento A é um número entre zero e um. Tendo presente que a probabilidade $P(A)$ está associada à proporção de sucessos do evento A :

Se $P(A)=0$, o evento A nunca ocorrerá, pois é um evento impossível.

Se $P(A)=1$, o evento A sempre ocorrerá, pois é um evento certo.

O valor da probabilidade $P(A)$ de um evento A no intervalo $(0, 1)$ deve ser interpretado como mostra a tabela seguinte, considerando que o experimento aleatório é repetido um número grande de vezes.

$P(A)$	Significado de $P(A)$
1	Sempre ocorre
0,90	Ocorre 90% das vezes e não ocorre em 10% das vezes
0,50	Ocorre 50% das vezes e não ocorre em 50% das vezes
0,15	Ocorre 15% das vezes e não ocorre em 85% das vezes
0	Nunca ocorre

Essa tabela mostra que:

- A soma das probabilidades de todos os possíveis resultados de um experimento aleatório é sempre igual a um.
- A probabilidade de um evento ocorrer é igual ao complemento desse mesmo evento não ocorrer. Se $P(A)$ é a probabilidade de ocorrer o evento A , então a probabilidade desse evento não ocorrer será o complemento $(1-P(A))$.
 - Por exemplo, se depois de repetir um número muito grande de vezes um experimento aleatório com espaço amostral $\{C, D\}$, o evento C ocorreu em 68% das vezes, o que significa que o evento C não ocorreu em 32% das vezes, que é a probabilidade de ocorrer o evento D .
 - Esse resultado mostra que a probabilidade do evento C ocorrer é igual à probabilidade complementar de ocorrer o evento D , isto é, $(1-P(D))=1-0,68=0,32$ ou 32%.

Conhecido o significado de probabilidade de um evento, o próximo passo é mostrar como determinar a probabilidade de um evento. Tradicionalmente há três formas de determinar a probabilidade de um evento, o procedimento teórico (probabilidade clássica ou *a priori*), o procedimento de frequência relativa (probabilidade *a posteriori*) e o procedimento de probabilidade subjetiva.

Probabilidade teórica de eventos

A probabilidade teórica de um evento é obtida utilizando procedimento de contagem. Por exemplo, qual a probabilidade de obter *cara* no lançamento de uma moeda? Nesse caso, o espaço amostral tem apenas dois eventos elementares mutuamente excludentes, *cara* e *coroa*. Considerando que os eventos *cara* e *coroa* são igualmente prováveis e não há nenhuma condição que estabeleça que um dos dois resultados tenha alguma preferência nem que um seja mutuamente dependente do outro, a probabilidade teórica de obter *cara* é obtida como resultado de dividir o número de eventos que atendem à condição *cara* pelo número total de eventos possíveis:

$$P(\text{caras}) = \frac{\text{Número de eventos favoráveis}}{\text{Número de eventos possíveis}}$$

Então, a probabilidade de obter *cara* será 0,5 ou 50%, resultado obtido de:

$$P(\text{caras}) = \frac{1}{2} = 0,50 \text{ ou } 50\%$$

Esse resultado mostra que a probabilidade de obter *coroa* é também 0,5, resultado obtido como o complemento $(1-P(\text{cara}))=1-0,50=0,50$ ou 50%.

Qual a probabilidade teórica de obter o número dois no lançamento de um dado? O espaço amostral do lançamento de um dado tem seis eventos elementares mutuamente excludentes {1, 2, 3, 4, 5, 6}. Como os seis resultados são igualmente prováveis, a probabilidade teórica de obter qualquer um dos eventos elementares é 1/6. Resumindo:

- Quando os eventos de um experimento são igualmente prováveis, a probabilidade de qualquer evento pode ser obtida como um cálculo teórico de contagem. Em geral, se o número de eventos elementares for m , a probabilidade de qualquer evento elementar será $1/m$.
- A probabilidade teórica de um evento é o limite de sua frequência relativa, assunto a ser tratado a seguir. Pela *lei dos grandes números*, ao aumentar o número de experimentos, a frequência relativa de cada evento se aproximará de seu valor teórico.

Frequência relativa

Citando Peter Bernstein, “... *Apenas em raros casos a vida imita os jogos de azar, em que podemos determinar as probabilidades de um resultado antes que um evento chegue a ocorrer – a priori, nas palavras de Jacob Bernoulli. Na maioria dos casos, temos de estimar as probabilidades com base no que aconteceu após o fato – a posteriori. A própria noção de a posteriori implica a experimentação e graus de crenças mutáveis. ...*”²

A probabilidade $P(A)$ de ocorrer o evento A de um experimento aleatório pode ser obtida como a porcentagem de ocorrência do evento A , depois de repetir o experimento um número muito grande de vezes. Por exemplo, repetindo um número muito grande de vezes o lançamento de uma moeda, a frequência relativa do evento *cara* será obtida como resultado de dividir o número de caras observadas pelo número de repetições do experimento. Nesse caso, a frequência relativa do evento *cara* é a própria probabilidade $P(\text{Cara})$.

$$P(\text{caras}) = \frac{\text{Número de caras observadas}}{\text{Número de repetições do experimento}}$$

Experimentalmente, pode-se observar que à medida que o número de repetições do experimento aumenta, a frequência relativa de um evento tende a chegar a um determinado valor que definimos como probabilidade desse evento, como é possível observar experimentalmente utilizando o *modelo* da planilha **Simulação** deste capítulo, mudando o número de lançamentos de uma moeda. A probabilidade teórica de um evento seria o limite de sua frequência relativa e, pela lei dos grandes números, ao aumentar o número de experimentos, a frequência relativa de cada evento se aproximará de seu valor teórico.

Nem sempre os eventos de um experimento são igualmente prováveis; por exemplo, o preço de uma ação daqui a um ano, comparado com o preço de hoje, poderá subir ou baixar, incluindo neste último evento o evento permanecer constante. Na construção do espaço amostral de eventos não igualmente prováveis, devem ser atendidas as seguintes condições.

- Os eventos do espaço amostral devem ser mutuamente excludentes e coletivamente exaustivos. O espaço amostral do preço da ação daqui a um ano contém dois eventos mutuamente excludentes com probabilidades diferentes.
- A soma das probabilidades dos eventos deve ser igual a um; por exemplo, se a probabilidade do evento subir for 70%, a probabilidade do evento baixar deverá ser 30%.

Simulador lançamento de uma moeda

Na planilha **Simulação**, incluída na pasta **Capítulo 5**, foi construído o modelo que simula o lançamento de uma moeda. Para facilitar a compreensão dos resultados, o modelo permite escolher um das três

quantidades de lançamentos programados, 1.500, 3.000 e 10.000 vezes, como mostra a Figura 5.6. Na simulação do lançamento da moeda, foram utilizados a ferramenta de análise *Geração de número aleatório* e o tipo de distribuição discreta que gera os números aleatórios 0 e 1 com probabilidade de 50% para cada um. Esses valores representam, respectivamente, os eventos coroa e cara. Todo o procedimento de simulação, da amostragem à construção do gráfico, passando pelas tabelas de acumulação do número de caras, foi mecanizado utilizando macros do VBA³ Excel.

Os resultados importantes são dois: o número acumulado de lançamentos da moeda e a porcentagem de caras acumuladas ou a frequência relativa do evento cara representado pelo número 1. O modelo deve ser utilizado da seguinte forma:

- Na *caixa de grupo Número de lançamentos da moeda*, selecione o número de lançamentos desejados clicando no botão de opção correspondente.
- Pressione o botão **Nova Simulação** para ativar a macro que realizará a simulação completa do lançamento de uma moeda a quantidade de vezes selecionada.

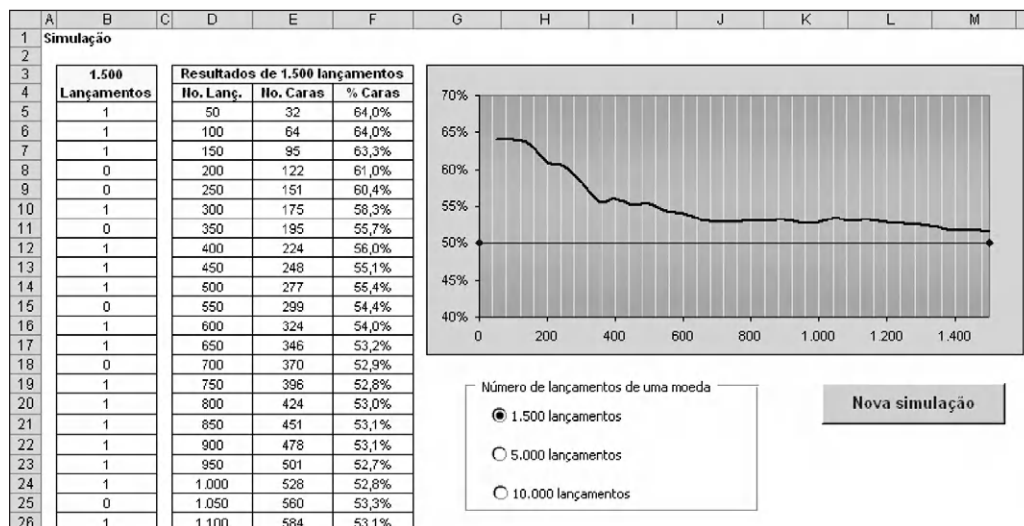


FIGURA 5.6 Simulação de 1.500 lançamentos de uma moeda.

Análise dos resultados da simulação

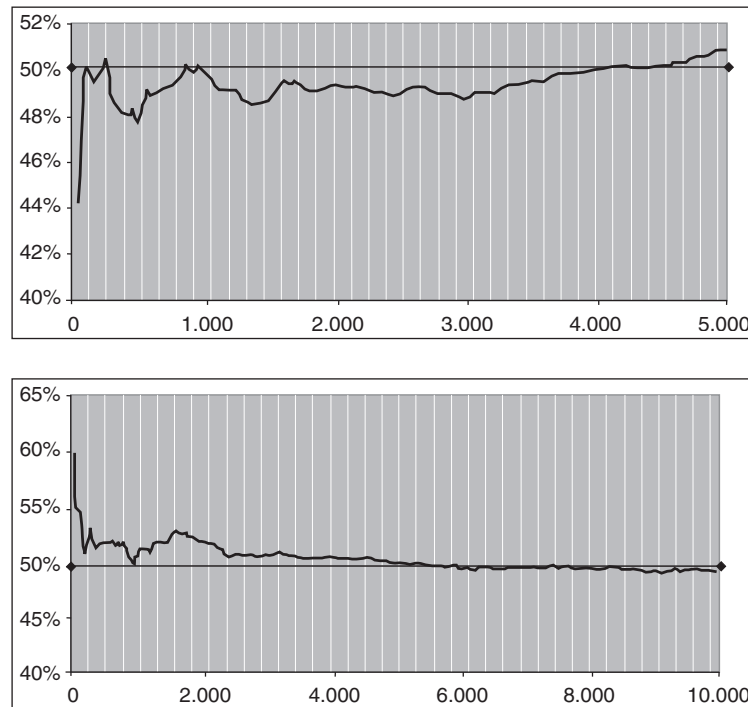
A probabilidade teórica de obter cara no lançamento de uma moeda é 0,50 ou 50%. Entretanto, esse resultado não significa que depois de lançar uma moeda, por exemplo, cem, mil, dez mil ou mais vezes seguidas ocorrerão exatamente 50% de caras e 50% coroas. Os gráficos registrados nas Figuras 5.1, 5.6 e 5.7 mostram a frequência relativa do evento cara para simulações com 1.500, 5.000 e 10.000 lançamentos de uma moeda.

- O gráfico das frequências relativas de caras da Figura 5.1 começa ao redor de 54%, segue com oscilações abaixo de 50% e termina com um valor um pouco abaixo de 50% depois de completar 1.500 lançamentos.
- O da Figura 5.6 começa ao redor de 64% e continua diminuindo com pequenas oscilações até concluir ao redor de 52% depois de completar 1.500 lançamentos.
- O primeiro gráfico da Figura 5.7, para 5.000 lançamentos, começa com 44%, permanece oscilando a maior parte da simulação abaixo de 50% e termina com um valor ao redor de 51%. O segundo

gráfico, de 10.000 lançamentos, começa com 60% e permanece acima de 50% nos primeiros 5.500 lançamentos aproximadamente, depois permanece abaixo de 50% até concluir com um valor ao redor de 49%.

FIGURA 5.7

Frequências relativas de 5.000 e 10.000 lançamentos de uma moeda.



Sugerimos que você realize várias simulações seguidas para cada quantidade de lançamentos programados e tente se sensibilizar com os resultados, primeiro em cada grupo de lançamentos, 1.500, 5.000 e 10.000, e depois tentando comparar os resultados entre esses grupos. Tente perceber que a probabilidade teórica de um evento seria o limite de sua frequência relativa e, pela lei dos grandes números, ao aumentar o número de experimentos a frequência relativa de cada evento se aproximará do seu valor teórico. Contudo, os exemplos obtidos com o modelo de simulação mostram que 1.500 ou 10.000 lançamentos podem apresentar resultados parecidos, o que nos faz pensar que a quantidade de lançamentos não tem tamanho ou há algum conceito que está fugindo ao nosso raciocínio. Voltemos para Peter Bernstein “... Suponha que você atire uma moeda repetidamente. A lei dos grandes números não diz que a média de suas jogadas se aproximará de 50% à medida que você aumentar o número de jogadas; a matemática elementar diz isto, poupando-lhe a tediosa tarefa de atirar a moeda repetidamente. Pelo contrário, a lei enuncia que aumentar o número de jogadas aumentará igualmente a probabilidade de que a razão entre as caras e o total de jogadas se desviará de 50% abaixo de uma quantidade especificada, por menor que seja. ... Não se está em busca da média real de 50%, mas da probabilidade de que o erro entre a média observada e a média real seja inferior a, digamos, 2% – em outras palavras, de que o aumento do número de jogadas aumenta a probabilidade de que a média observada não se desvie em mais de 2% da média real. ... Isso não significa que não haverá erro após um número infinito de jogadas Tudo que a lei nos informa é que a média de um grande número de jogadas diferirá por menos de que certa quantidade especificada da média real mais provavelmente do que a média de um pequeno número de jogadas. Além disso, sempre haverá uma possibilidade de que o resultado observado difira da média real por uma quantidade maior do que o limite especificado. ...” Esse conceito está presente na Estimação da média da população a partir de uma amostra representativa, tema tratado no Capítulo 11 deste livro. Apenas como ideia instigante, às vezes se diz que se o número de experimentos tender a infinito, a frequência relativa tenderá ao valor teórico; no entanto, parece que também pode não ser suficiente.

Lei de Benford

Neste momento, tomamos um desvio do tema que estamos tratando para mostrar uma aplicação interessante da análise de 0s e 1s da série de resultados gerados pelo *modelo* construído na planilha *Simulação*. Em continuação, reproduzimos parte de um artigo de jornal.⁴

“O professor Dr. Theodore P. Hill pede sempre uma lição de casa especial para seus alunos de matemática, no Instituto de Tecnologia da Geórgia. Parte deles deve lançar uma moeda duzentas vezes e registrar fielmente seu resultado, enquanto a outra simplesmente deve fingir que jogou a moeda e inventar um resultado para os duzentos supostos arremessos. No dia seguinte, para espanto dos alunos, Hill consegue, com uma breve olhada nos trabalhos, apontar quase todos os que fraudaram os lançamentos. A verdade, disse ele em uma entrevista, *é que a maioria das pessoas não sabe quais são as reais probabilidades de um exercício como esse e, portanto, não consegue inventar dados convincentes.* ... As previsões de probabilidades são muitas vezes surpreendentes. No caso da experiência com o lançamento da moeda ... em algum ponto de uma série de duzentos arremessos de moeda, ou cara ou coroa aparecerá seis ou mais vezes seguidas. Aqueles que fraudaram um resultado não sabiam disso e evitaram simular longas sequências de caras ou coroas, porque, erroneamente, pensaram ser improvável.”

Primeiro sugerimos que você verifique a afirmação do professor Hill na coluna B do *modelo Simulação*. Depois que continue com o artigo.

... Hill integra o cada vez maior contingente de estatísticos, contadores e matemáticos que estão convencidos do poder assombroso do teorema matemático conhecido como Lei de Benford. O teorema é uma maneira poderosa e relativamente simples de apontar o dedo da suspeita para fraudadores, autores de desfalques, sonegadores de impostos, contadores negligentes e até *bugs* de computador....”

Essa linha de trabalho faz parte da *Lei de Benford* em homenagem ao Dr. Frank Benford que em 1938 divulgou a constatação de que as páginas da tabela de logaritmos dos números começando com o algarismo 1 estavam mais sujas e desgastadas, acreditando que esse resultado não era de nenhuma preferência pessoal por esses números da tabela. Numa análise de 20.229 conjuntos de números de diferentes categorias de informações, todos esses registros seguiam o mesmo padrão de probabilidade do primeiro algarismo. Para explicar essa constatação, considerando que certeza absoluta é definida como 1 e a impossibilidade absoluta como 0, Benford apresentou a seguinte fórmula $\log_{10}(1 + 1/d)$, que retorna a probabilidade do algarismo $d=1, 2, \dots, 9$ ser o primeiro de um grupo de algarismos. Aplicando essa fórmula, a frequência do algarismo 1 é 30,6%, a frequência do algarismo 2 é 17,6% e a frequência dos demais algarismos continua decrescendo até 4,6% para o algarismo 9, resultados constatados em diversas pesquisas.⁵ Observe que as informações pesquisadas não são respostas de eventos aleatórios, nos quais todos os algarismos têm a mesma probabilidade de ocorrência, como mostrado no Capítulo 1 com a geração de números aleatórios e a planilha *Simulação* deste capítulo.

Árvore de possibilidades

A *árvore de possibilidades* é a representação gráfica dos eventos elementares de um espaço amostral. Essa representação é muito útil para organizar os cálculos e os resultados de experimentos com mais de uma etapa, por exemplo, o lançamento de uma moeda três vezes seguidas. Em cada um dos três lança-

⁴ Aplicação do teorema pode indicar fraudes de Malcom, W. Browne artigo publicado no jornal *O Estado de São Paulo* em 9/8/1998.

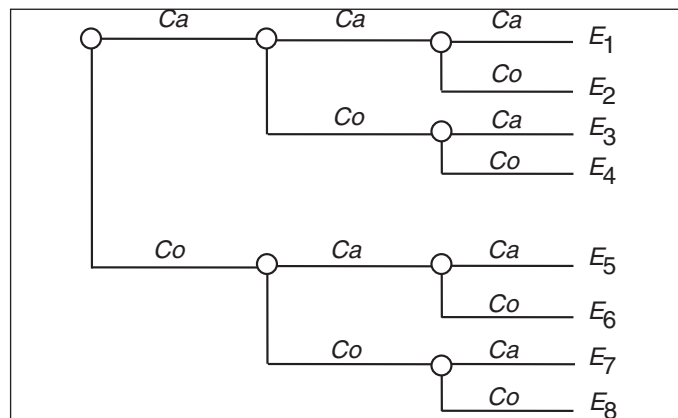
⁵ Mais informação sobre a Lei de Benford podem ser obtidas no site <http://www.rexswain.com/benford.html> com o artigo completo de Malcom W. Browne publicado no jornal *The New York Times* em 4/8/1998. Também em outros endereços conseguidos pelo Google ou outro mecanismo de busca equivalente.

mentos, há duas possibilidades de resultados, cara e coroa. Ao analisar a sequência dos três lançamentos, deve-se raciocinar da seguinte forma:

- Os resultados do segundo lançamento serão combinados com os resultados do primeiro. O resultado do cara do primeiro se combinará com os dois resultados do segundo e, da mesma forma, o resultado coroa do primeiro. Dessas combinações resultam quatro eventos elementares.
- Os resultados do terceiro lançamento serão combinados com os quatro resultados dos dois primeiros lançamentos, resultando oito eventos elementares.

A árvore de possibilidades da Figura 5.8 mostra os oito eventos elementares do espaço amostral S referente ao lançamento de uma moeda três vezes seguidas: $S=\{E_1, E_2, \dots, E_i, \dots, E_8\}$. Repetindo o experimento um número muito grande de vezes, a frequência relativa dos oito eventos será próxima de $1/8$, pois, no limite, quando o número de experimentos tender a infinito, a frequência relativa dos eventos será $1/8$. De outra maneira, os oito eventos têm a mesma probabilidade $1/8$, pois são eventos igualmente prováveis.

FIGURA 5.8 Árvore de possibilidades.



Regra da soma

Os eventos de um mesmo espaço amostral podem ser combinados aplicando as regras das operações união e interseção de conjuntos. Ao mesmo tempo, a probabilidade de uma combinação de eventos pode ser obtida das probabilidades dos eventos, como mostra a regra da soma de eventos mutuamente excludentes.

Sejam dois eventos mutuamente excludentes A e B com probabilidades $P(A)$ e $P(B)$. A probabilidade $P(A \text{ ou } B)$ de ocorrer A ou B é igual à soma das probabilidades dos eventos: $P(A \text{ ou } B) = P(A) + P(B)$.

EXEMPLO 5.3

Calcule a probabilidade de ocorrer apenas uma *cara* no lançamento de uma moeda três vezes seguidas.

Solução. Analisando os resultados da árvore de possibilidades da Figura 5.8, podemos ver que apenas os eventos elementares E_4 , E_6 e E_7 têm uma *cara*. Pela regra da soma de eventos mutuamente excludentes, a probabilidade de acontecer E_4 ou E_6 ou E_7 é igual $37,5\%$, resultado obtido da soma das probabilidades desses eventos:

$$P(E_4 \text{ ou } E_6 \text{ ou } E_7) = P(E_4) + P(E_6) + P(E_7)$$

$$P(E_4 \text{ ou } E_6 \text{ ou } E_7) = 1/8 + 1/8 + 1/8 = 3/8 = 0,375 \text{ ou } 37,50\%$$

EXEMPLO 5.4

Continuando com o lançamento de uma moeda três vezes seguidas. Qual a probabilidade de ocorrerem duas ou mais caras?

Solução. Analisando os resultados da árvore de possibilidades, verificamos que os eventos elementares E_1 , E_2 , E_3 e E_5 têm duas ou mais cara. A probabilidade de ocorrerem duas ou mais caras é 50%, resultado obtido de:

$$P(E_1 \text{ ou } E_2 \text{ ou } E_3 \text{ ou } E_5) = P(E_1) + P(E_2) + P(E_3) + P(E_5) \\ P(E_1 \text{ ou } E_2 \text{ ou } E_3 \text{ ou } E_5) = 1/8 + 1/8 + 1/8 + 1/8 = 4/8 = 0,50 \text{ ou } 50\%$$

Probabilidade condicional

As probabilidades estudadas até este momento são denominadas probabilidades incondicionais porque as únicas condições estabelecidas se referem ao experimento, resultados igualmente prováveis etc. Em alguns casos, interessa rever a probabilidade de um evento, pois há informações adicionais que podem afetar o resultado. Por exemplo, dentro do mesmo espaço amostral, a probabilidade de que aconteça o evento X tendo acontecido, ou sabendo que aconteceu, o evento Y é obtida a partir do espaço amostral reduzido, definido pelo evento Y . A probabilidade $P(X/Y)$ é denominada probabilidade condicional.

EXEMPLO 5.5

Sabendo que no lançamento de três moedas não aconteceram duas nem três coroas, qual a probabilidade que as três moedas sejam cara?

Solução. Começamos por lembrar que a probabilidade de obter três caras no lançamento de uma moeda três vezes seguidas é $1/8$ ou 12,50%. Qual é a vantagem da informação de que não aconteceram nem duas nem três coroas? Ao tomar conhecimento antecipado dessa informação que denominamos evento Y , deduzimos que o evento Y está formado pelos eventos elementares: $\{E_1, E_2, E_3, E_5\}$. O evento que as três sejam cara que denominamos X está formado por um único evento elementar $X=\{E_1\}$. Portanto, a probabilidade de que aconteça X sabendo que aconteceu Y é a probabilidade condicional $P(X/Y)=25\%$, obtida com a fórmula:

$$P(X/Y) = \frac{1}{4} = 0,25$$

Observe que ao tomar conhecimento do evento Y do Exemplo 5.5, o espaço amostral inicial formado por oito eventos elementares foi reduzido para quatro eventos elementares. Portanto, ao tomar conhecimento antecipado do evento Y , o espaço amostral foi reduzido e, conseqüentemente, a probabilidade das três moedas serem cara aumentou de 12,50% para 25%.

A probabilidade condicional $P(X/Y)$ entre os eventos X e Y pode ser obtida como resultado da divisão da probabilidade conjunta $P(X \text{ e } Y)$ pela probabilidade do evento Y : $P(X/Y) = \frac{P(X \text{ e } Y)}{P(Y)}$.

EXEMPLO 5.6

Uma urna contém três bolas, duas verdes V e uma branca B . Duas bolas são retiradas em sequência, uma por vez, e sem reposição. Calcule a probabilidade de que a segunda bola seja verde sabendo que a primeira também foi verde.

Solução. O objetivo é determinar a probabilidade condicional $P(X/Y)$, onde Y é o evento *primeira bola de cor verde* e o evento X *segunda bola de cor verde*. O espaço amostral inicial é $S=\{VV, VB, BV\}$. Ao tomar conhecimento de que a primeira bola foi verde, o espaço amostral do evento Y passa a ser: $Y=\{VV, VB\}$. Desses dois possíveis resultados, apenas nos interessa um, o evento VV . Portanto, a probabilidade condicional $P(Y/X)$ é igual a 50%, obtida com a fórmula: $P(X/Y) = \frac{1}{2} = 0,50$.

Probabilidades conjunta e total

Um mesmo espaço amostral pode ser analisado de diversas formas. Por exemplo, das respostas de 1.000 funcionários a uma pesquisa interna da empresa de serviços, na tabela seguinte foram registrados os resultados do hábito de fumar dos funcionários classificados por sexo, mulher e homem.

	Mulher	Homem
Fuma	68	82
Não fuma	462	388

Para analisar as informações dessa tabela é melhor construir a tabela a seguir com os mesmos resultados, porém considerando a população de 1.000 funcionários, registrando os valores unitários ou como percentagens. À primeira tabela, foram adicionados outros resultados obtidos dos anteriores e registrados nas novas coluna e linha adicionadas.

	Mulher	Homem	Total
Fuma	0,068	0,082	0,15
Não fuma	0,462	0,388	0,85
Total	0,53	0,47	1,00

A tabela construída é denominada *tabela de probabilidades conjuntas e marginais* e é uma forma prática de calcular a probabilidade condicional de dois eventos. Entretanto, analisemos primeiro os resultados:

- O primeiro resultado 0,068 indica que 6,8% das mulheres fumam. Esse resultado representa a probabilidade conjunta: Fuma e é Mulher.
- Da mesma forma, 38,8% dos homens não fumam. Esse resultado representa a probabilidade conjunta: não Fuma e é Homem.
- O total 0,15 da coluna *Total* é o resultado da soma das probabilidades conjuntas 0,068 mais 0,082. O resultado 0,15 ou 15% é a probabilidade total, ou marginal, de Fuma ou, de outra maneira, 15% dos que responderam tem o hábito de fumar.
- Da mesma forma, o total 0,53 da linha *Total* é o resultado da soma das probabilidades conjuntas 0,068 mais 0,462. O resultado 0,53 ou 53% é a probabilidade total de Mulher ou, de outra maneira, 53% dos que responderam são mulheres.
- Como controle, a soma das quatro probabilidades conjuntas deve ser sempre igual a 1 ou 100% e, da mesma maneira, a soma das probabilidades da linha *Total* e da coluna *Total* deve ser sempre igual a 1 ou 100%.

Com os resultados dessa tabela é possível obter probabilidades condicionais, por exemplo, a probabilidade de que o respondente da pesquisa seja mulher sabendo que não fuma. Essa pergunta pode ser representada da seguinte forma $P(\text{Mulher}/\text{Não fuma})$.

	Mulher	Homem	Total
Fuma	0,068	0,082	0,15
Não fuma	0,462	0,388	0,85
Total	0,53	0,47	1,00

Como o evento conhecido é *Não fuma*, primeiro, na tabela selecionamos a linha *Não fuma*, que representa o espaço amostral reduzido, depois de ter tomado conhecimento do evento *Não fuma*. Depois, calculamos a probabilidade $P(\text{Mulher}/\text{Não fuma})=0,5435$ ou 54,35%, dividindo a probabilidade conjunta 0,462 pela probabilidade total 0,85. Esse mesmo resultado pode ser obtido diretamente com a fórmula, utilizando os resultados da última tabela.

$$P(X/Y) = \frac{P(XeY)}{P(Y)}$$

$$P(\text{Mulher}/\text{NãoFuma}) = \frac{0,462}{0,85} = 0,5435$$

Deixamos para o leitor verificar que a probabilidade de o respondente da pesquisa não fumar sabendo que é mulher é $P(\text{Não fuma}/\text{Mulher})=0,8717$ ou 87,17%. Há outras possíveis perguntas, por exemplo, a probabilidade condicional $P(\text{Homem}/\text{Fuma})$ ou a $P(\text{Fuma}/\text{Homem})$ etc.

EXEMPLO 5.7

Dos eventos C e D de um mesmo espaço amostral são conhecidas as probabilidades $P(C \text{ e } D)=0,125$, $P(C)=0,50$ e $P(D)=0,25$. Construa a tabela de probabilidades conjuntas e marginais e depois calcular a probabilidade condicional $P(\text{Não } C/D)$.

Solução. Começamos por registrar os dados dos eventos C e $\text{Não } C$, e D e $\text{Não } D$, na tabela seguinte.

	D	Não D	Total
C	0,125		0,50
Não C			
Total	0,25		1,00

Sabendo que a soma das probabilidades da linha *Total* e da coluna *Total* devem ser sempre igual a 1 ou 100%, completamos os resultados que faltam nessa linha e nessa coluna. Da mesma maneira, as probabilidades conjuntas que faltam podem ser obtidas aplicando as regras das somas, lembrando que a soma das probabilidades conjuntas deve ser sempre igual a 1 ou 100%. Dessa maneira, obtemos a tabela seguinte de probabilidades conjuntas e totais.

	D	Não D	Total
C	0,125	0,375	0,50
Não C	0,125	0,375	0,50
Total	0,25	0,75	1,00

A probabilidade condicional $P(\text{Não } C/D)$ é calculada da seguinte forma. Como o evento conhecido é D , primeiro, na tabela selecionamos a coluna D que representa o espaço amostral reduzido. Depois, calculamos a probabilidade $P(\text{Não } C/D)=0,50$ ou 50% dividindo a probabilidade conjunta 0,125 pela probabilidade total 0,25.

É importante destacar que a tabela de probabilidades conjuntas e marginais pode ter mais de duas linhas ou colunas, dependendo dos valores possíveis de cada evento.

Regra do produto

Da fórmula da probabilidade condicional, obtém-se a importante regra do produto das probabilidades $P(XeY) = P(X/Y) \times P(Y)$.

EXEMPLO 5.8

Repetindo o enunciado do Exemplo 5.6, uma urna contém três bolas, duas verdes V e uma branca B . Duas bolas são retiradas em sequência, uma por vez. Calcule a probabilidade conjunta de que as duas bolas sejam verdes.

Solução. Embora não seja necessário, começamos por construir a tabela das probabilidades conjuntas e totais. Os títulos das duas linhas da tabela se referem à primeira retirada de uma bola, e os títulos das duas colunas se referem à segunda retirada de uma bola. No miolo da tabela, estão registrados os resultados possíveis depois das duas retiradas seguidas sem reposição de uma bola da urna.

	Verde	Branca
Verde	2	2
Branca	2	0

A probabilidade conjunta de que as duas bolas sejam verdes, ou $P(X \text{ e } Y) = 2/6 = 1/3$, está registrada na primeira célula desta tabela.

	Verde	Branca	Total
Verde	$2/6 = 1/3$	$1/6$	$4/6 = 2/3$
Branca	$2/6 = 1/3$	0	$2/6 = 1/3$
Total	$4/6 = 2/3$	$1/6$	6

Repetindo o cálculo da probabilidade de que a segunda bola seja verde sabendo que a primeira também foi verde, Exemplo 5.6, da tabela anterior obtemos o valor 0,50 como resultado da divisão de $2/6$ por $4/6$.

Regra do produto – Eventos Independentes

Se um evento não depender da ocorrência de outro evento anterior, os eventos são definidos como eventos independentes. Nesse caso, se os eventos X e Y são independentes, então a probabilidade condicional de um evento será dada pela expressão $P(X) = P(X/Y)$, e a probabilidade do produto de dois eventos independentes será $P(XeY) = P(X) \times P(Y)$, que é a regra do produto de eventos independentes.

EXEMPLO 5.9

Qual a probabilidade de ocorrerem três caras no lançamento de três moedas?

Solução. A probabilidade de cada lançamento é 0,50. A probabilidade de ocorrerem três caras será 12,50%, obtida da fórmula:

$$P(XeYeZ) = P(X) \times P(Y) \times P(Z)$$

$$P(XeYeZ) = 0,50 \times 0,50 \times 0,50 = 0,125$$

Vejamos algumas conclusões:⁶

- A probabilidade condicional entre dois eventos X e Y é regida pela expressão $P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$, onde $P(Y) \neq 0$.
- Se os eventos X e Y forem mutuamente excludentes, então $P(X \cap Y) = 0$ e, conseqüentemente, $P(X/Y) = 0$. Portanto, $P(X/Y) \neq P(X)$ e os eventos serão necessariamente dependentes.
- Se os eventos X e Y verificarem que $P(X \cap Y) \neq 0$, os dois eventos poderão ser independentes, pois não podemos afirmar que sejam independentes salvo que se verifique a igualdade $P(X \cap Y) = P(X) \times P(Y)$ para cada par de valores.

Técnicas de contagem

Listar e contar os eventos elementares do experimento aleatório lançamento de uma moeda três vezes seguidas é um procedimento simples, pois o número de resultados do espaço amostral desse experimento é pequeno. Entretanto, se o experimento fosse o lançamento de um dado três vezes seguidas ou o lançamento de uma moeda oito vezes seguidas, o procedimento de listar e contar todos os possíveis resultados seria trabalhoso. As técnicas de contagem⁷ ajudam a determinar, sem necessidade de enumeração direta, o número de resultados possíveis de um espaço amostral. Para facilitar o procedimento de cálculo, as técnicas de contagem serão apresentadas combinadas com as funções matemáticas e estatísticas do Excel.

EXEMPLO 5.10

Determine o número de resultados possíveis do lançamento de um dado três vezes seguidas.

Solução. A contagem é realizada em três passos:

- Cada lançamento de um dado tem seis resultados possíveis $\{1, 2, 3, 4, 5, 6\}$.
- Os seis resultados do segundo lançamento serão combinados com cada um dos seis resultados do primeiro lançamento totalizando 36 possíveis resultados.
- Os seis resultados do terceiro lançamento serão combinados com cada um dos 36 resultados acumulados dos dois lançamentos anteriores, totalizando 216 resultados.

O Exemplo 5.10 mostra o procedimento de contagem realizado com a fórmula da multiplicação, se de uma determinada ocorrência há m resultados e, em sequência, de outra ocorrência há n resultados, então há mn resultados associados. Utilizando os dados do Exemplo 5.10, concluímos que o número de resultados do lançamento de um dado três vezes seguido é $6 \times 6 \times 6 = 6^3 = 216$

EXEMPLO 5.11

A placa dos carros que circulam em todo o país é formada por três letras seguidas de quatro algarismos de zero a nove. Determinar o número de placas possíveis considerando que podem ser utilizadas 22 letras em cada posição, e o primeiro algarismo não pode ser zero.

Solução. O número de placas possíveis é 9.583.200 obtido com a fórmula:

$$22 \times 22 \times 22 \times 9 \times 10 \times 10 = 22^3 \times 9 \times 10^2 = 9.583.200$$

⁶ Você pode passar este assunto, sem perda de continuidade com o resto do livro.

⁷ Conhecidas também como *Análise Combinatória*.

Permutações

Os resultados dos Exemplos 5.10 e 5.11 mostram que a fórmula da multiplicação retorna o número de resultados associados de dois ou mais grupos. A fórmula da permutação dá o número de *arranjos* de um mesmo grupo.

EXEMPLO 5.12

Calcule o número de permutações das cinco letras *a, b, c, d e e* tomadas três a três, quatro a quatro e cinco a cinco.

Solução. Para contar o número de permutações procedemos assim:

- A primeira letra pode ser qualquer uma das cinco letras *a, b, c, d e e*.
- A segunda letra pode ser qualquer uma das quatro letras restantes.
- A terceira letra pode ser qualquer uma das três letras restantes.

O número de permutações das cinco letras *a, b, c, d e e* tomadas três a três é 60, há 60 palavras de três letras distintas, resultado obtido com a fórmula da multiplicação $60 = 5 \times 4 \times 3$. Esse resultado pode ser obtido com a função PERMUT do Excel.

- **PERMUT(*n*; *r*)**

A função estatística PERMUT⁸ retorna o número de arranjos de *n* elementos tomados em grupos de *r*. Os valores de *n* e *r* são números inteiros positivos; entretanto, a função PERMUT aceita números fracionários que são truncados para números inteiros antes de calcular os fatoriais.

Neste exemplo, a fórmula =PERMUT(5;3) registrada em uma célula vazia de uma planilha Excel retorna o resultado 60. Na planilha **Funções para Contagem**, incluída na pasta **Capítulo 5**, estão registradas as formas de utilizar a função PERMUT como as que serão apresentadas a seguir.

- A quarta letra pode ser qualquer uma das duas letras restantes.
 - O número de permutações das cinco letras *a, b, c, d e e*, tomadas quatro a quatro, é 120, resultado obtido com a fórmula da multiplicação $120 = 5 \times 4 \times 3 \times 2$. Com a fórmula =PERMUT(5;4), tem-se o resultado 120.
- A quinta letra é a última letra restante.
 - O número de permutações das cinco letras *a, b, c, d e e*, tomadas cinco a cinco, é também 120, resultado obtido com a fórmula da multiplicação $120 = 5 \times 4 \times 3 \times 2 \times 1$. Com a fórmula =PERMUT(5;5), tem-se 120.

Os resultados do Exemplo 5.12 correspondem ao número de permutações de cinco letras tomadas três a três, quatro a quatro e cinco a cinco. De forma geral, o número $P(n,r)$ de permutações de *n* objetos associados em grupos de *r* é calculado com a fórmula:

$$P(n,r) = n \times (n-1) \times \cdots \times (n-r+1)$$

Tendo presente que o fatorial do número natural *n* é o produto de todos os *n* primeiros números inteiros e positivos e é representado pelo símbolo $n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$, definindo que $0! = 1$, a fórmula do número de permutações pode ser escrito com fatoriais:

$$P(n,r) = \frac{n!}{(n-r)!}$$

Aplicando esta última fórmula para calcular o resultado do Exemplo 5.12:

$$P(5,3) = \frac{5!}{(5-3)!} = 60$$

⁸ Em inglês, a função PERMUT é PERMUT.

Esse resultado pode ser obtido com a função FATORIAL do Excel.

• FATORIAL(*n*)

A função matemática FATORIAL⁹ retorna o fatorial do número *n* sendo *n* um número não negativo. Se *n* não for inteiro, será truncado para um número inteiro antes de realizar o cálculo.¹⁰ Por exemplo:

- O fatorial de *n*=5 é 5!=5×4×3×2×1=120, resultado que também pode ser obtido com a fórmula =FATORIAL(5) digitada em qualquer célula vazia da planilha Excel.
- Para resolver a primeira questão do Exemplo 5.12, a fórmula =FATORIAL(5)/FATORIAL(5-3) registrada numa célula do Excel retornará o número de permutações 60.

Na planilha **Funções para Contagem**, incluída na pasta **Capítulo 5**, estão registradas as formas de utilizar a função FATORIAL e as outras funções do Excel utilizadas neste capítulo.

Vejam um caso especial da permutação. Se *x*=*r*, o número de permutações será igual a

$$P(n, n) = \frac{n!}{(n - n)!} = n!, \text{ que é a própria expressão do fatorial de } n, \text{ que representa o número de permutações de } n \text{ objetos tomados todos ao mesmo tempo, como mostrado no Exemplo 5.12. Essa condição mostra que a fórmula } =\text{PERMUT}(5;5) \text{ é equivalente a } =\text{FATORIAL}(5).$$

Combinações

O resultado *b, c, d* como os resultados *c, b, d* e *d, c, b* fazem parte dos 60 resultados da permutação de cinco objetos identificados pelas letras, *a, b, c, d* e *e* tomados três a três do Exemplo 5.12. Como esses três resultados têm as mesmas letras *b, c* e *d*, deduzimos que, na contagem das permutações, a ordem dos objetos é importante. Há casos em que o que interessa é o próprio objeto sem interessar a ordem de como foi obtido; nesse caso, o tipo de contagem é denominada *combinação*.

Por exemplo, vimos que o número de permutações de cinco letras *a, b, c, d* e *e* tomadas três a três sem considerar a ordem das letras é igual a 60. Mas nesse resultado estão incluídas todas as permutações possíveis de três letras que é igual a 6=3×2×1. O número de combinações será igual a 10, resultado obtido da divisão do número de permutações pelo número de permutações de três letras, como mostra a fórmula:

$$\frac{P(5,3)}{3!} = 10$$

De forma geral, o número *C*(*n*,*r*) de combinações de *n* objetos associados em grupos de *r* é calculado com a fórmula:

$$C(n, r) = \frac{n!}{r!(n - r)!}$$

Portanto, o resultado da combinação de cinco letras associadas em grupos de três letras é

$$C(5,3) = \frac{5!}{3!(5 - 3)!} = 10.$$

Esse resultado pode ser obtido com a função COMBIN do Excel.

⁹ Em inglês, a função FATORIAL é FACT.

¹⁰ O Excel dispõe também das funções FACTDOUBLE e MULTINOMIAL, em inglês FACTDOUBLE e MULTINOMIAL.

- **COMBIN($n; x$)**

A função matemática COMBIN¹¹ retorna o número de combinações de x objetos tomados x a x , considerando que a ordem dos objetos não interessa. Os valores de n e x são números inteiros positivos; entretanto, a função COMBIN aceita números fracionários que são truncados para números inteiros antes de calcular os fatoriais. Por exemplo, o número de combinações de cinco objetos tomados três a três é dez, valor obtido registrando a fórmula =COMBIN(5;3) numa célula vazia do Excel. Verifique que:

- A fórmula =FATORIAL(5)/(FATORIAL(5-3)*FATORIAL(3)) registrada numa célula vazia de Excel retorna o resultado 10.

- Das fórmulas $P(n,r)$ e $C(n,r)$, obtém-se a igualdade $C(n,r) = \frac{P(n,r)}{r!}$. Da mesma forma, pode-se ver

$$\text{que } COMBIN(n;x) = \frac{PERMUT(n;x)}{FATORIAL(x)}.$$

Na planilha **Funções para Contagem**, incluída na pasta **Capítulo 5**, estão registradas as formas de utilizar a função COMBIN e as outras funções do Excel utilizadas neste capítulo.

Problemas

Problema 1

No lançamento de uma moeda dez vezes seguidas ocorreram dez coroas. Se a moeda for lançada mais uma vez, qual a probabilidade de que seja cara? Por quê?

R: $P(\text{cara})=50\%$

Problema 2

Suponha que depois de lançar uma moeda dez vezes seguidas, a frequência relativa do evento *cara* seja 70%. É razoável aceitar esse resultado? Por quê?

R: Sim.

Problema 3

Jogue um dado e observe o resultado. Se o experimento for repetido um número muito grande de vezes, que proporção do total de lançamentos terá o resultado observado no primeiro lançamento do dado? Por quê?

R: 1/6

Problema 4

Se depois de lançar um dado doze vezes seguidas, a frequência relativa do resultado cinco for 75% é razoável aceitar esse resultado? Por quê?

R: Sim.

Problema 5

Continuando com o lançamento de uma moeda três vezes seguidas, qual a probabilidade de obter pelo menos duas *coroas*?

R: $P(\text{pelo menos duas coroas})=50\%$

¹¹ Em inglês, a função COMBIN é COMBIN.

Problema 6

Suponha que depois de lançar uma moeda cem mil vezes seguidas a frequência relativa do evento *cara* seja igual a 0,70. É razoável aceitar esse resultado? Por quê?

R: Sim, porém com baixíssima probabilidade de ocorrer.

Problema 7

Qual a probabilidade de ocorrerem três coroas no lançamento de três moedas?

R: $P(\text{as três moedas com coroa}) = 12,50\%$

Problema 8

Continuando com o lançamento de uma moeda três vezes seguidas, qual a probabilidade de obter as três moedas com a mesma face?

R: $P(\text{as três moedas com a mesma face}) = 25\%$

Problema 9

No lançamento de um dado, qual a probabilidade de obter: a) um número menor do que cinco e b) um número par?

R: a) $P(\text{número menor do que cinco}) = 4/6$ b) $P(\text{um número par}) = 3/6 = 1/2$

Problema 10

Uma moeda é lançada duas vezes seguidas. Sabendo que o resultado de uma das moedas foi *cara*, qual a probabilidade que a outra moeda seja também *cara*?

R: $P = 1/3$

Problema 11

Um homem tinha dois gatos, um preto e um branco. O branco era macho. Qual é a probabilidade de que o outro fosse macho?¹²

R: $P = 1/2$

Problema 12

Um homem tinha dois gatos. Um deles, pelo menos, era macho. Qual é a probabilidade de que os dois fossem machos?¹³

R: $P = 1/3$. Analise como o Problema 8.

Problema 13

Semanalmente são sorteados seis números de um grupo de 60 números. Quantos são os resultados possíveis de um sorteio semanal?

R: Resultados possíveis: 50.063.860

Problema 14

Continuando com o Problema 13. Se você concorrer nesse sorteio, qual a probabilidade de acertar o prêmio?

R: $P = 1/50.063.860$, considerando todos os resultados igualmente prováveis.

¹² Exemplo de *O Enigma de Sherazade* de Raymond Smullyan, Jorge Zahar Editor, 1997.

¹³ Veja nota de rodapé 12.

Problema 15

Semanalmente são sorteados cinco números de um grupo de 80 números. Quantos são os resultados possíveis de um sorteio semanal e qual a probabilidade de acertar o prêmio?

R: Resultados possíveis: 24.040.016 $P=1/24.040.016$, considerando todos os resultados igualmente prováveis.

Problema 16

Um fabricante de microcomputadores decidiu vender pela Internet unidades padronizadas definidas pelo comprador. Para começar, estabeleceu as seguintes alternativas: dois tipos de *CPU*, duas *memórias RAM*, três capacidades de *discos rígidos* e quatro tipos de *monitores*. Quantas configurações são possíveis de montar?

R: 48 configurações

Problema 17

A probabilidade de um estudante obter o conceito máximo *A* no primeiro teste de estatística é 25%, e a probabilidade de obter o mesmo conceito *A* no segundo teste é também 25%. Sabendo que a probabilidade de obter *A* nos dois testes é 15%, qual a probabilidade do estudante obter menos do que *A* no segundo teste, sabendo que no primeiro teste obteve o conceito *A*?

R: $P(\text{Não } A/A)=0,10/0,25=40\%$

Problema 18

Continuando com o Problema 17. Qual a probabilidade do estudante obter menos do que *A* nos dois testes?

R: $P(\text{Não } A \text{ e Não } A)=0,10/0,25=40\%$

Problema 19

Uma pesquisa de mercado mostrou que 80% das casas pesquisadas têm um aparelho de TV em cores e que 30% das casas pesquisadas têm um forno de micro-ondas. A pesquisa mostrou também que 20% das casas pesquisadas têm um aparelho de TV em cores e um forno de micro-ondas. Qual a porcentagem das casas pesquisadas que não têm nenhum dos dois?

R: 10%

Problema 20

Qual a porcentagem das casas pesquisadas que não têm um aparelho de TV em cores, porém tem um forno de micro-ondas?

Problema 21

O gerente do departamento de atendimento de uma revendedora de carros agrupou as reclamações dos clientes no último mês em: *Cliente Atendido* e *Não Atendido*, e *Cliente Exigente* e *Normal*, como registrado na tabela seguinte:

Cliente	Exigente	Normal
Atendido	3	56
Não atendido	17	24

Escolhendo aleatoriamente um cliente, calcule a probabilidade de que:

- O cliente tenha sido atendido sabendo que é um cliente *Exigente*.
- O cliente não tenha sido atendido sabendo que é um cliente *Normal*.

R: a) $P(\text{Atendido}/\text{Exigente})=15\%$ b) $P(\text{Não Atendido}/\text{Normal})=30\%$

Problema 22

A gerência de vendas da rede de Magazines classificou as compras de 100 clientes por tipo de produto comprado e por idade do comprador e os resultados estão registrados na tabela seguinte:

	<30	30-40	41-50	>50
Eletrodomésticos	12	10	11	14
Vestário	10	7	8	6
Lazer	1	3	5	13

Determine:

- A probabilidade de que um cliente tenha mais que 40 anos.
- A probabilidade de um cliente ter mais que 50 anos, sabendo que comprou um produto de lazer.
- A probabilidade de um cliente ter mais que 40 anos, sabendo que comprou um produto de vestário.

R: a) 57% b) 59,1% c) 45,2%

Problema 23

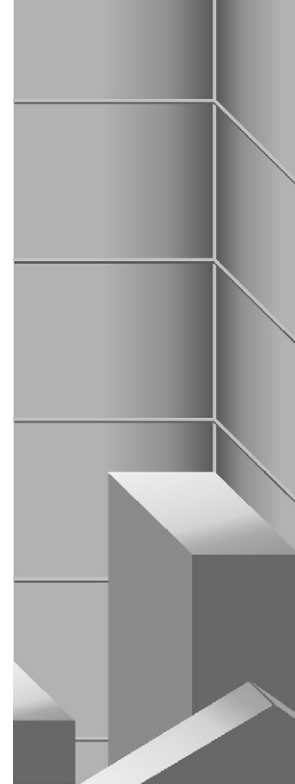
Em uma simulação de 1.000 lançamentos de uma moeda realizada com o *modelo* da planilha *Simulação*, qual das duas sequências de 1s seguintes têm mais chance de ocorrer, A ou B? Por quê?

$A = \{ \dots 001111011111101100\dots \}$

$B = \{ \dots 01110111011111111\dots \}$

Capítulo 6

CORRELAÇÃO



Até este momento, foram analisados os dados de uma amostra ou variável pertencente a uma população. Outra análise importante é determinar como uma variável se relaciona com outras variáveis da mesma população. Neste capítulo, será mostrada uma forma de medir quanto e de que maneira se relacionam duas variáveis. Há muitos exemplos de relações entre amostras, por exemplo, geralmente os meios de comunicação divulgam informações de variáveis relacionadas obtidas de resultados de pesquisas como:

- Nas Instituições de Ensino Superior – IES há uma relação direta entre a qualidade do ensino e a taxa de inadimplência. A taxa de inadimplência das IES que obtiveram conceitos A e B no Provão é 12,1%, nas que obtiveram C é 16% e nas que obtiveram D e E a inadimplência é de 21,9%.¹
- O frio está para o setor farmacêutico como o Dia das Mães está para o comércio. As vendas de medicamentos não controlados, como analgésicos, antigripais e vitaminas, disparam.²
- O faturamento das empresas de energia nos Estados Unidos é diretamente influenciado pela temperatura, especialmente no inverno. Um inverno brando reduz a demanda de energia para calefação e pode diminuir drasticamente o lucro.³

A partir desses exemplos, você poderá encontrar outras relações como, por exemplo, reduzindo o custo, o preço do produto será reduzido e será possível aumentar a quantidade vendida, ou funcionário com maior escolaridade terá mais chance de crescer na empresa etc. Em qualquer caso, é importante lembrar que a informação recebida nem sempre é corretamente exposta, como no caso da relação direta entre qualidade do ensino das IES e a taxa de inadimplência que, na realidade, mantém uma relação inversa, pois a instituição com maior conceito corresponde à menor taxa de inadimplência.

Os gráficos de dispersão da Figura 6.1 mostram dois tipos de relação entre as variáveis X e Y. O gráfico de dispersão da esquerda mostra uma relação direta ou positiva, tendência destacada pela declividade positiva da elipse tracejada, enquanto o gráfico de dispersão da direita mostra uma relação inversa ou negativa, tendência também destacada pela declividade negativa da elipse tracejada.

¹ A Pressão da inadimplência, artigo de P. de Athayde publicado na revista *Carta Capital* de 15/10/2003.

² Frio chega e venda de remédios dispara, artigo de C. Silva publicado no jornal *O Estado de São Paulo* de 25 de maio de 2004.

³ Alugue o sol artigo de Cláudio Gradilone publicado na revista *Exame* em 6/2/2002. Divulga a proteção financeira com derivativos de clima ou *weather derivatives*.

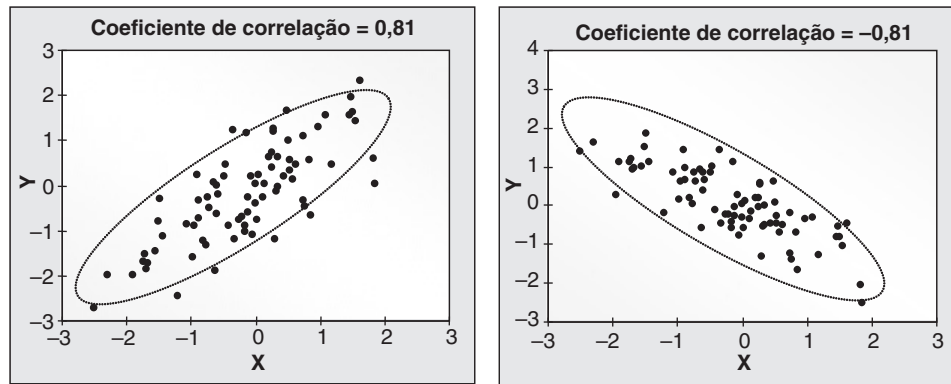
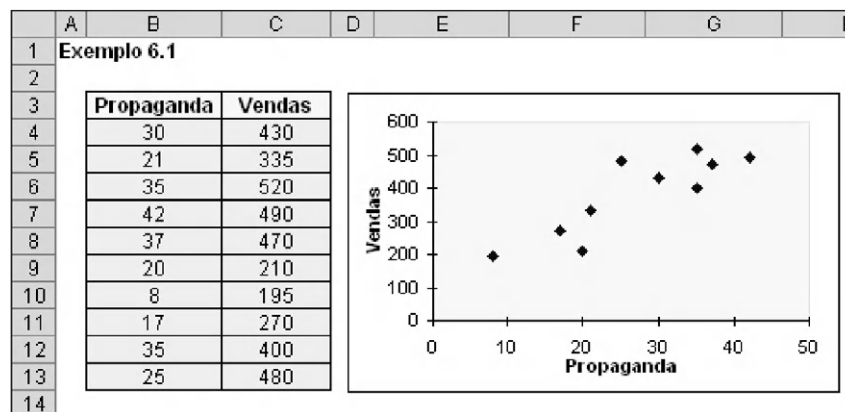


FIGURA 6.1 Dois tipos de relações entre duas variáveis.

EXEMPLO 6.1

O diretor de vendas da rede de varejo nacional está analisando a relação entre o investimento em propaganda e as vendas da empresa utilizando os dados registrados no intervalo B3:C13, incluindo os títulos, da planilha **Exemplo 6.1**, incluída na pasta **Capítulo 6**. Analise a relação entre essas duas amostras.

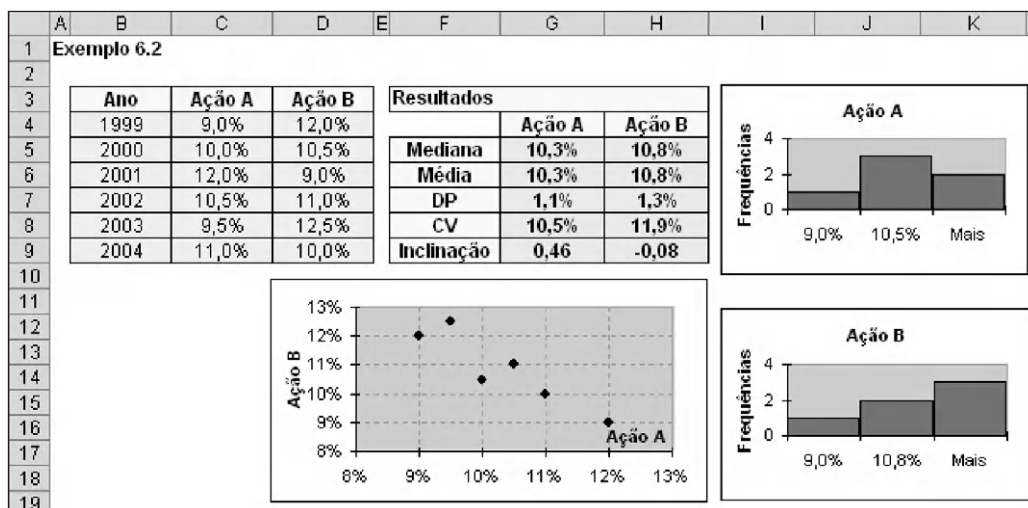
Solução. Com os dados da tabela, foi construído o gráfico de dispersão como, mostra a figura a seguir. Os registros dos dez pares de valores investimento e vendas mostram uma clara relação direta ou positiva, pois à medida que o investimento em propaganda aumenta, as vendas também aumentam, e vice-versa. Esse gráfico mostra que as duas variáveis estão correlacionadas de forma positiva.



EXEMPLO 6.2

Os retornos anuais durante os últimos seis anos da Ação A e da Ação B negociadas na Bolsa de Valores estão registrados na planilha **Exemplo 6.2**, incluída na pasta **Capítulo 6**. Realize uma análise estatística desses retornos e da relação entre eles.

Solução. No intervalo B3:D9 foram registrados os retornos das ações. No intervalo F4:H9 foram calculadas e registradas as medidas estatísticas mediana, média, desvio padrão, coeficiente de variação e coeficiente de inclinação, como mostra a próxima figura.



As medidas estatísticas dos retornos das duas ações são parecidas, exceto o coeficiente de inclinação, que indica formas diferentes das distribuições dos retornos como fica confirmado pelos histogramas construídos com a ferramenta de análise *Histograma*. Os histogramas dos retornos mostram que essas duas amostras têm particularidades que as medidas estatísticas não conseguem capturar. Para ver a diferença entre os retornos das duas ações, na mesma planilha foi construído o gráfico de dispersão que mostra a relação negativa entre os retornos das duas ações. Analisemos o gráfico de dispersão:

- Partindo do ano 1999, retornos 9% e 12%, respectivamente Ação A e Ação B, no ano 2000, o retorno da Ação A aumentou para 10%, enquanto o retorno da Ação B diminuiu para 10,50%. No ano 2001, os retornos mantiveram a mesma tendência do ano 2000.
- Nos anos 2002 e 2003, os retornos inverteram a tendência anterior. Enquanto o retorno da Ação A diminuiu, o retorno da Ação B aumentou.
- No ano 2004, os retornos das ações A e B inverteram novamente a tendência dos dois anos anteriores.

As medidas estatísticas dos retornos anuais das ações A e B do Exemplo 6.2 são parecidas, porém o gráfico de dispersão mostra que os retornos têm tendências opostas. A covariância e o coeficiente de correlação medem a tendência e a força da relação linear entre as duas variáveis ou amostras.

Covariância

O coeficiente de correlação pode ser calculado diretamente com a fórmula do coeficiente de Pearson; entretanto, preferimos iniciar este assunto definindo primeiro a covariância de duas variáveis, apresentação parecida à realizada com o desvio padrão, definindo primeiro a variância.

A covariância σ_{XY} das variáveis $X = X_1, X_2, \dots, X_N$ e $Y = Y_1, Y_2, \dots, Y_N$, consideradas como população é:⁴

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y)$$

⁴ Quando necessário, as variáveis são separadas com vírgula, $\sigma_{X,Y}$ e $S_{X,Y}$.

A covariância S_{XY} das variáveis $X = X_1, X_2, \dots, X_n$ e $Y = Y_1, Y_2, \dots, Y_n$, consideradas como amostra é:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}) \times (Y_i - \bar{Y})$$

EXEMPLO 6.3

Calcule a covariância das variáveis do Exemplo 6.1.

Solução. Os dados desse exemplo foram registrados na planilha **Exemplo 6.3**, incluída na pasta **Capítulo 6**, como mostra a figura seguinte incluindo os resultados. O primeiro resultado necessário e o valor das médias das duas variáveis, calculadas e registradas nas células D16 e D17. Depois:

- Na célula D4, foi registrada a fórmula =B4-\$D\$16 que calcula o desvio do primeiro dado da variável *Propaganda*. Essa fórmula foi copiada até a célula D13. O mesmo procedimento foi utilizado para calcular e registrar os desvios da variável *Vendas*.
- Na célula F4 foi registrada a fórmula =D4*E4 que retorna o produto dos desvios do primeiro dado. Essa fórmula foi copiada até a célula D13.

	A	B	C	D	E	F
1	Exemplo 6.3					
2						
3		Propaganda	Vendas	X-Méd.X	Y-Méd.Y	(X-Méd.X)*(Y-Méd.Y)
4		30	430	3,00	50,00	150
5		21	335	(6,00)	(45,00)	270
6		35	520	8,00	140,00	1120
7		42	490	15,00	110,00	1650
8		37	470	10,00	90,00	900
9		20	210	(7,00)	(170,00)	1190
10		8	195	(19,00)	(185,00)	3515
11		17	270	(10,00)	(110,00)	1100
12		35	400	8,00	20,00	160
13		25	480	(2,00)	100,00	-200
14						
15		Resultados				
16		Média Propaganda	27,00			
17		Média Vendas	380,00			
18		Soma produtos dos desvios	9.855,00			
19		Covariância população	985,50	=D18/CONT.NÚM(B4:B13)		
20		Covariância amostra	1.095,00	=D18/(CONT.NÚM(B4:B13)-1)		
21						
22		Covariância população	985,50	=COVAR(B4:B13;C4:C13)		
23						

No intervalo de resultados:

- Na célula D18 foi registrada a fórmula =SOMA(F4:F13) que calcula a soma dos produtos dos desvios.
- Com a fórmula =D18/CONT.NÚM(B4:B13), registrada na célula D19, é calculada a covariância da população $\sigma_{XY}=985,50$. E na célula D20 foi registrada a fórmula =D18/(CONT.NÚM(B4:B13)-1) que calcula a covariância da amostra $S_{XY}=1.095,00$.

O resultado da covariância da população também pode ser obtido com a função COVAR do Excel, registrando a fórmula =COVAR(B4:B13;C4:C13) na célula D22. A sintaxe da função COVAR é a seguinte:

• COVAR(matriz1; matriz2)

A função estatística COVAR⁵ retorna a covariância da população dos valores registrados nos argumentos *matriz1* e *matriz2*. Esses argumentos podem ser registrados como intervalos de uma planilha, como já mos-

⁵ Em inglês, COVAR é COVAR.

trado, tomando o cuidado de verificar que as duas variáveis tenham a mesma quantidade de dados. Também é possível registrar os argumentos como *matriz* na própria fórmula da função, evitando registrar os valores da amostra num intervalo de células da planilha como foi feito na célula D29, registrando a fórmula =COVAR({30;21;35;42;37;20;8;17;35;25};{430;335;520;490;470;210;195;270;400;480})

Características da covariância

A covariância mede a tendência e a força da relação linear entre duas variáveis. Das expressões da covariância para população e para amostra temos seguintes características:

- As duas amostras ou variáveis devem ter o mesmo número de dados.
- Os pares de dados ocorrem ao mesmo tempo, são pares casados. Embora possa parecer redundante, tenha presente que não se pode mudar a ordem de uma única variável; a mudança de ordem deverá ser realizada nas duas amostras sem descartar os pares de dados.
- A covariância é a média dos produtos dos desvios das duas amostras ou variáveis, obtida como resultado da divisão:
 - No caso de população, da soma dos produtos dos desvios pela quantidade de dados das variáveis.
 - No caso de amostra, da soma dos produtos dos desvios pela quantidade de dados das variáveis menos um.⁶
- Os numeradores das expressões da covariância para população e para amostra são iguais, o resultado da soma dos produtos dos desvios.

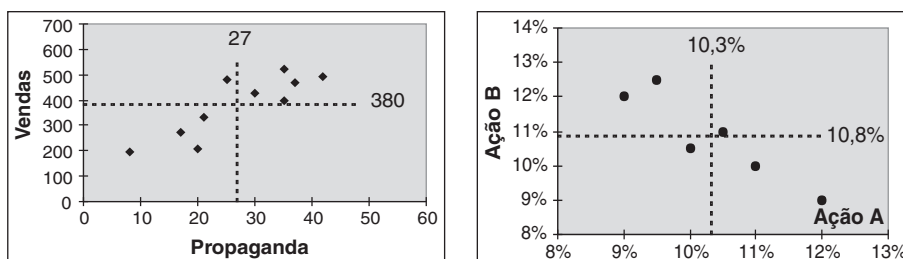


FIGURA 6.2 Análise dos gráficos de dispersão dos Exemplos 6.1 e 6.2.

- A covariância pode assumir qualquer valor do conjunto dos números reais, pois pode ser nula, negativa ou positiva. Baseada na definição dos produtos dos desvios,⁷ uma explicação intuitiva é que a covariância é a medida do afastamento simultâneo das respectivas médias. Se ambas as variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva e, nos outros casos, poderá ser negativa, como mostram os gráficos de dispersão da Figura 6.2.
- O gráfico da esquerda mostra que a maioria dos pares de valores do Exemplo 6.1 tem os dois valores acima de sua média correspondente, provocando a covariância positiva, resultado que pode ser confirmado no Exemplo 6.3.
- O gráfico da direita mostra que a maioria dos pares de valores do Exemplo 6.2 tem um valor acima da média e o outro abaixo da média correspondente, provocando a covariância negativa, como poderá ser confirmado no Exemplo 6.4.
- Da mesma forma que a variância, a covariância é afetada pelos valores extremos da variável, ela não é uma medida resistente.

⁶ Equivalente ao caso da variância da amostra S^2 , Capítulo 4.

⁷ Copeland T. – *Opções Reais*, Editora Campus 2001.

- A unidade de medida é o resultado do produto das unidades dos valores das variáveis; no caso do Exemplo 6.1, a unidade é o binômio *vendas e investimento* e, no Exemplo 6.2, o binômio % e %, ambas sem nenhum significado prático.

Regras operacionais da covariância

As propriedades⁸ mais importantes da covariância são:

- Outra forma de calcular a covariância é com a seguinte fórmula para a população que tem a vantagem de não ter de calcular os desvios. No Apêndice 1 você encontra a demonstração desta fórmula:

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \mu_X \mu_Y$$

- O Exemplo 6.3 mostra como calcular a covariância da população e a covariância da amostra, procedimentos que diferem apenas no valor do divisor da soma dos produtos dos desvios. De forma equivalente ao realizado com as variâncias da população e da amostra, a partir das expressões das covariâncias, pode-se estabelecer a seguinte igualdade:

$$\sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y) = \sigma_{XY} \times N = S_{XY} \times (n - 1)$$

Portanto, a expressão de equivalência entre as duas covariâncias é:

$$S_{XY} = \frac{N}{n - 1} \sigma_{XY}$$

A covariância da amostra do Exemplo 6.3 pode ser obtida a partir do valor da covariância da população com a fórmula:

$$S_{XY} = \frac{10}{10 - 1} (985,5) = 1.095,00$$

- A covariância de uma variável, e ela mesma, é a própria variância da variável, seja no caso de população ou amostra. Como $Y = X$, então:

$$\sigma_{XX} = \frac{\sum_{i=1}^N (X_i - \mu_X) \times (X_i - \mu_X)}{N} = \frac{\sum_{i=1}^N (X_i - \mu_X)^2}{N} = \sigma_X^2$$

- A permutação das variáveis não altera o resultado da covariância, se os pares de valores não forem alterados: $\sigma_{XY} = \sigma_{YX}$.
- Há outras propriedades operacionais muito práticas. Por exemplo, representando a covariância como $Cov(X, Y)$ e sendo a , b e c constantes, sempre se verifica:
 - $Cov(X, a) = 0$
 - $Cov(X, -Y) = -Cov(X, Y)$
 - $Cov(aX, Y) = a Cov(X, Y)$

⁸ Estas propriedades aplicadas com a covariância considerada como população também se aplicam com a covariância considerada como amostra.

- [illegible]

Para mostrar a diferença de procedimento entre resultados de população e amostra dos dados:

- No intervalo G5:H5 foram calculados e registrados os desvios padrão, dos retornos considerando-os como população. Depois, no intervalo G6:H6 os desvios padrão considerando os retornos como amostras.
- No intervalo G7:G8 foi calculada a covariância, primeiro, como população utilizando a função COVAR do Excel, e depois como amostra, a partir do primeiro resultado.
- No intervalo G9:G10 foi calculado o coeficiente de correlação, primeiro, utilizando os resultados parciais de população, e depois como amostra, utilizando também os resultados parciais, porém como amostra. Com esse cálculo, verifica-se o mesmo valor de coeficiente de correlação $r=-0,9203$.
- Utilizando os dados como a população, temos $r_{AB}=-0,9203$ da seguinte forma:

$$r_{AB} = \frac{-0,00010694}{0,009860 \times 0,011785} = -0,9203$$

- Utilizando os dados como a amostra, também se obtém $r_{AB}=-0,9203$ da seguinte forma:

$$r_{AB} = \frac{-0,00012833}{0,010801 \times 0,012910} = -0,9203.$$

O coeficiente de correlação pode ser obtido com a função CORREL do Excel registrando na célula G12 a fórmula =CORREL(C4:C9;D4:D9). A sintaxe da função CORREL é a seguinte:

• **CORREL(matriz1; matriz2)**

A função estatística CORREL⁹ retorna o coeficiente de correlação dos valores registrados nos argumentos *matriz1* e *matriz2*. Esses argumentos podem ser intervalos de uma planilha, como mostrado anteriormente, tomando o cuidado de verificar que as duas variáveis tenham a mesma quantidade de dados. Também se podem registrar os argumentos como *matriz* na própria fórmula da função, evitando registrar os valores da amostra em um intervalo de células da planilha, como feito na célula G14, registrando a fórmula =CORREL({0,09;0,1;0,12;0,105;0,095;0,11}; {0,12;0,105;0,09;0,11;0,125;0,1})

O coeficiente de correlação pode também ser obtido com a função PEARSON do Excel como foi obtido com a fórmula = PEARSON(C4:C9;D4:D9) registrada na célula G16. A sintaxe da função PEARSON é a seguinte:

• **PEARSON(matriz1; matriz2)**

A função estatística PEARSON¹⁰ retorna o coeficiente de correlação dos valores registrados nos argumentos *matriz1* e *matriz2*. Esses argumentos podem ser intervalos de uma planilha, como mostrado anteriormente, tomando o cuidado de verificar que as duas variáveis tenham a mesma quantidade de dados. Também se podem registrar os argumentos como *matriz* na própria fórmula da função, evitando registrar os valores da amostra em um intervalo de células da planilha, como feito na célula G18 registrando a fórmula =CORREL({0,09;0,1;0,12;0,105;0,095;0,11}; {0,12;0,105;0,09;0,11;0,125;0,1})

O resultado da função PEARSON é o mesmo que o da função CORREL, porém utilizando os valores das variáveis como mostra a fórmula seguinte.

$$r = \frac{n \times \sum_{i=1}^n X_i \times Y_i - \sum_{i=1}^n X_i \times \sum_{i=1}^n Y_i}{\sqrt{n \times \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \times \sqrt{n \times \sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i\right)^2}}$$

⁹ Em inglês, CORREL é CORREL.

¹⁰ Em inglês, PEARSON é PEARSON.

Características do coeficiente de correlação

Vejamos algumas características importantes do coeficiente de correlação:

- A fórmula do coeficiente de correlação pode ser apresentada sem incluir a covariância, como mostram as fórmulas a seguir, que dão o mesmo resultado do coeficiente de correlação, embora utilizem medidas estatísticas diferentes:

- População: $r_{XY} = \frac{1}{N} \sum_{i=1}^n \left(\frac{X_i - \mu_X}{\sigma_X} \right) \left(\frac{Y_i - \mu_Y}{\sigma_Y} \right)$

- Amostra: $r_{XY} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right)$

Essas duas fórmulas se assemelham à fórmula de Pearson.

- Se a variável Y é a mesma variável X, então o coeficiente de correlação é igual a um, como mostramos a seguir.

$$r_{XX} = \frac{\sigma_{XX}}{\sigma_X \times \sigma_X} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

- A permutação das variáveis não altera o resultado do coeficiente de correlação, se os mesmos pares de valores forem mantidos $r_{XY} = r_{YX}$.
- Da mesma forma que a covariância, o coeficiente de correlação é afetado pelos valores extremos da variável, ele não é uma medida resistente.
- Se as variáveis X e Y forem estatisticamente independentes, então o coeficiente de correlação dessas variáveis será igual a zero. Entretanto, se o resultado do coeficiente de correlação das variáveis X e Y for igual a zero, não se poderá afirmar que as duas variáveis sejam estatisticamente independentes. Para confirmar essa independência, deve-se verificar se todos os pares de valores das variáveis X e Y cumprem a condição: $P(XeY) = P(X) \times P(Y)$.

Análise dos valores do coeficiente de correlação

Na planilha **Análise**, incluída na pasta **Capítulo 6**, é analisada a tendência e a força da relação linear entre duas variáveis ou amostras X e Y medida pelo coeficiente de correlação.

Variáveis perfeitamente correlacionadas de forma positiva

A Figura 6.3 mostra o comportamento de duas amostras X e Y perfeitamente correlacionadas em sentido positivo. O coeficiente de correlação dessas amostras calculado na célula F3 é igual a $r=+1$.

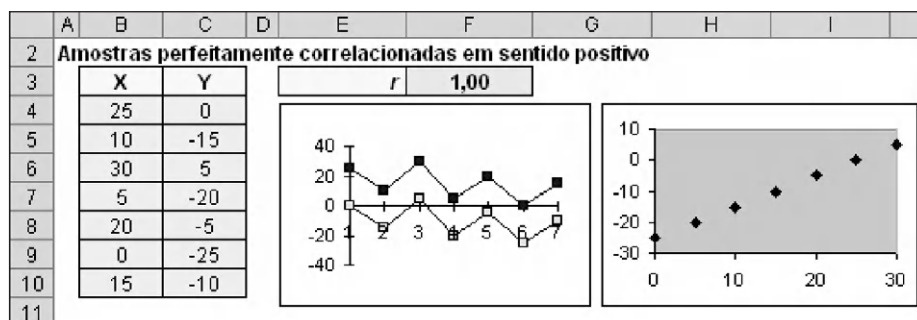


FIGURA 6.3 Amostras perfeitamente correlacionadas em sentido positivo, $r=+1$.

A tabela dos dados das variáveis X e Y mostra que qualquer par de valores dessas variáveis é obtido do par anterior adicionando ou subtraindo o mesmo valor. Por exemplo, o segundo par de valores (10,-15) é o resultado de subtrair 15 de cada valor do primeiro par (25,0). O terceiro par de valores é o resultado de adicionar 20 ao segundo par, e assim sucessivamente, até completar todos os pares de valores. Da análise dos dois gráficos da Figura 6.3:

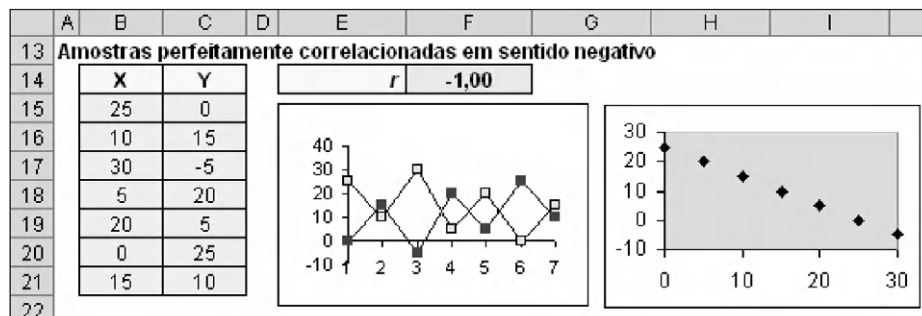
- O gráfico sequencial dos pares de valores, lado esquerdo da Figura 6.3, mostra que os valores das variáveis X e Y formam duas linhas paralelas, com acréscimos e decréscimos iguais e na mesma direção. As duas séries estão perfeitamente correlacionadas de forma positiva.
- O gráfico de dispersão, lado direito da Figura 6.3, mostra que os incrementos dos valores das duas variáveis X e Y são iguais e na mesma direção, sejam acréscimos ou decréscimos. Todos os pares de valores são pontos de uma reta com declividade 45° e, portanto, $r=+1$.

Se os incrementos entre pares são proporcionais e na mesma direção, sejam acréscimos ou decréscimos, os pontos formados pelos pares de valores fazem parte de uma reta com declividade positiva e, portanto, $r=+1$.

Variáveis perfeitamente correlacionadas de forma negativa

A Figura 6.4 mostra o comportamento de duas amostras X e Y perfeitamente correlacionadas em sentido negativo. O coeficiente de correlação dessas amostras calculado na célula F14 é igual a $r=-1$.

FIGURA 6.4 Amostras perfeitamente correlacionadas em sentido negativo, $r=-1$.



Neste caso, também, a tabela dos dados das variáveis X e Y mostra que qualquer par de valores das variáveis X e Y é obtido do par anterior. Por exemplo, o segundo par de valores (10,15) é o resultado de subtrair 15 do valor de X do primeiro par (25,0) e adicionar o valor 15 no primeiro valor de Y . Da mesma maneira, o terceiro par de valores (30,-5) é o resultado de adicionar 20 ao valor de X do segundo par e subtrair 20 do segundo valor de Y , e assim sucessivamente, até completar todos os pares de valores. Da análise dos dois gráficos da Figura 6.4:

- O gráfico sequencial dos pares, lado esquerdo da Figura 6.4, mostra que os valores das variáveis X e Y formam duas linhas opostas, os acréscimos e decréscimos são iguais, porém em direções opostas. As duas séries estão perfeitamente correlacionadas de forma negativa.
- O gráfico de dispersão, lado direito da Figura 6.4, mostra que os incrementos dos valores das duas variáveis X e Y são iguais e em direções opostas. Todos os pares de valores são pontos de uma reta com declividade 135° e, portanto, $r=-1$.

Se os incrementos são proporcionais e em direções opostas, os pontos formados pelos pares de valores fazem parte de uma reta com declividade negativa e, portanto, $r=-1$.

Variáveis não correlacionadas

A Figura 6.5 mostra o comportamento de duas amostras X e Y não correlacionadas. O coeficiente de correlação dessas amostras calculado na célula F25 é igual a $r=0$. Os pares de valores do gráfico de dispersão não apresentam nenhuma tendência.

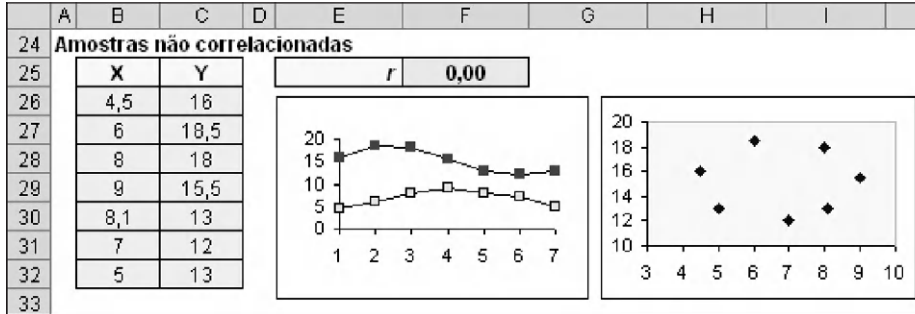


FIGURA 6.5
Amostras não correlacionadas, $r=0$.

Simulador coeficiente de correlação

As três análises apresentadas mostram que o coeficiente de correlação de duas variáveis X e Y com n pares de valores será um valor dentro do intervalo $-1 \leq r_{XY} \leq +1$. Dentro desse intervalo $(-1, +1)$, há um número muito grande de valores do coeficiente de correlação possíveis, que são consequência do afastamento simultâneo dos pares de valores de suas respectivas médias.

A planilha **Simulador de Correlação**, incluída na pasta **Capítulo 6**, ajudará a compreender a formação do coeficiente de correlação, como mostra a Figura 6.6 com o gráfico de dispersão de duas amostras aleatórias X e Y com 100 pares de dados. As características do simulador são:

- A análise pode ser realizada com amostras aleatórias contendo 50, 100 ou 150 pares de dados, ou pontos do gráfico de dispersão, opções que ajudam a compreender a formação do coeficiente de correlação. Cada vez que for escolhida uma *Quantidade de pares de valores*, selecionando o botão de opção correspondente, será ativada a macro que gera as séries aleatórias bivariadas normais.¹¹
- O acionamento da *Barra de rolagem*, localizada na parte inferior do gráfico, gera séries com novos coeficientes de correlação cujo valor é registrado na parte superior do gráfico. Por exemplo, o coeficiente de correlação das séries de dados da Figura 6.6 é $r=0,85$.
- O gráfico inclui a *reta Média X* e a *reta Média Y* que facilitam a visualização da formação do sinal do coeficiente de correlação, como apresentado anteriormente.
- O *modelo* pode gerar séries de valores para o valor de coeficiente de correlação informado na célula I5. Para isso, informe o valor do coeficiente em I5 e depois pressione o botão *r*. Verifique que a barra de rolagem se posicionou no valor registrado em I5.

A Figura 6.7 mostra seis gráficos com coeficientes de correlação diferentes e dentro do intervalo $(-1,1)$.

- No gráfico esquerdo da primeira linha, as amostras têm uma correlação positiva perfeita, $r=1$. Os pontos estão em uma mesma reta com declividade positiva. No gráfico da Figura 6.6, os pontos mostram uma correlação positiva, e as amostras têm correlação positiva, $r=0,85$. A maioria dos pontos está próxima de uma reta com declividade positiva.
- No gráfico direito da primeira linha, as amostras têm uma fraca correlação positiva, $r=0,32$. Os pares de valores formam uma nuvem com ligeira tendência de declividade positiva.

¹¹ *Discrete-Event System Simulation*, Banks J. et al – Prentice Hall, 2ª ed., 1996.

- No gráfico esquerdo da segunda linha, não há relação entre as variáveis, $r=-0,01$. Os pontos formam uma nuvem sem nenhuma tendência.
- No gráfico direito da segunda linha, as amostras têm uma fraca correlação negativa, $r=-0,32$, apresentando uma nuvem de pontos com ligeira tendência de declividade negativa.
- Na terceira e última linha, o gráfico da esquerda mostra que as amostras têm uma boa correlação negativa, $r=-0,85$. Coeficientes de correlação inferiores a esse valor e se aproximando de -1 mostrariam uma forte correlação negativa. A maioria dos pontos está próxima de uma reta com declividade negativa.
- No gráfico direito da terceira linha, as amostras têm uma correlação negativa perfeita, $r=-1$. Os pontos estão em uma mesma reta com declividade negativa.

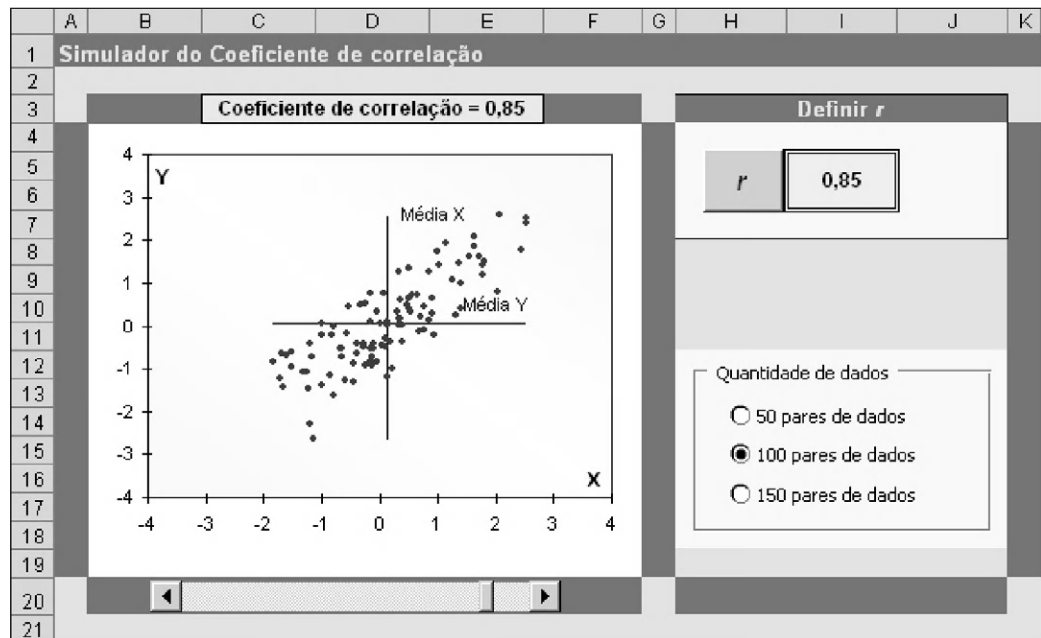


FIGURA 6.6
Simulador de
Correlação.

Alguns cuidados com os resultados

É importante ter em mente duas características do coeficiente de correlação:

- Mantendo os mesmos pares de valores, a permutação das variáveis não altera o resultado do coeficiente de correlação $r_{XY} = r_{YX}$.
- O valor r_{XY} é uma medida da tendência e da força da relação linear entre as variáveis X e Y .

Quando uma cozinheira varia a dosagem de fermento na produção de pão e consegue um melhor sabor, pode-se concluir que a melhoria do sabor foi causado pela nova dosagem de fermento. Nesse caso, há uma relação causa-efeito. Contudo, em geral, uma forte correlação não é sinônimo de uma relação causa-efeito entre as amostras ou variáveis. Há situações em que um coeficiente de correlação próximo de um ou de menos um não significa que a maioria dos pares de valores esteja contida em uma reta. Como será mostrado, o simples conhecimento do coeficiente de correlação não é suficiente devido a anomalias na dispersão dos dados, sendo recomendado construir o gráfico de dispersão das amostras para melhor compreender o resultado.

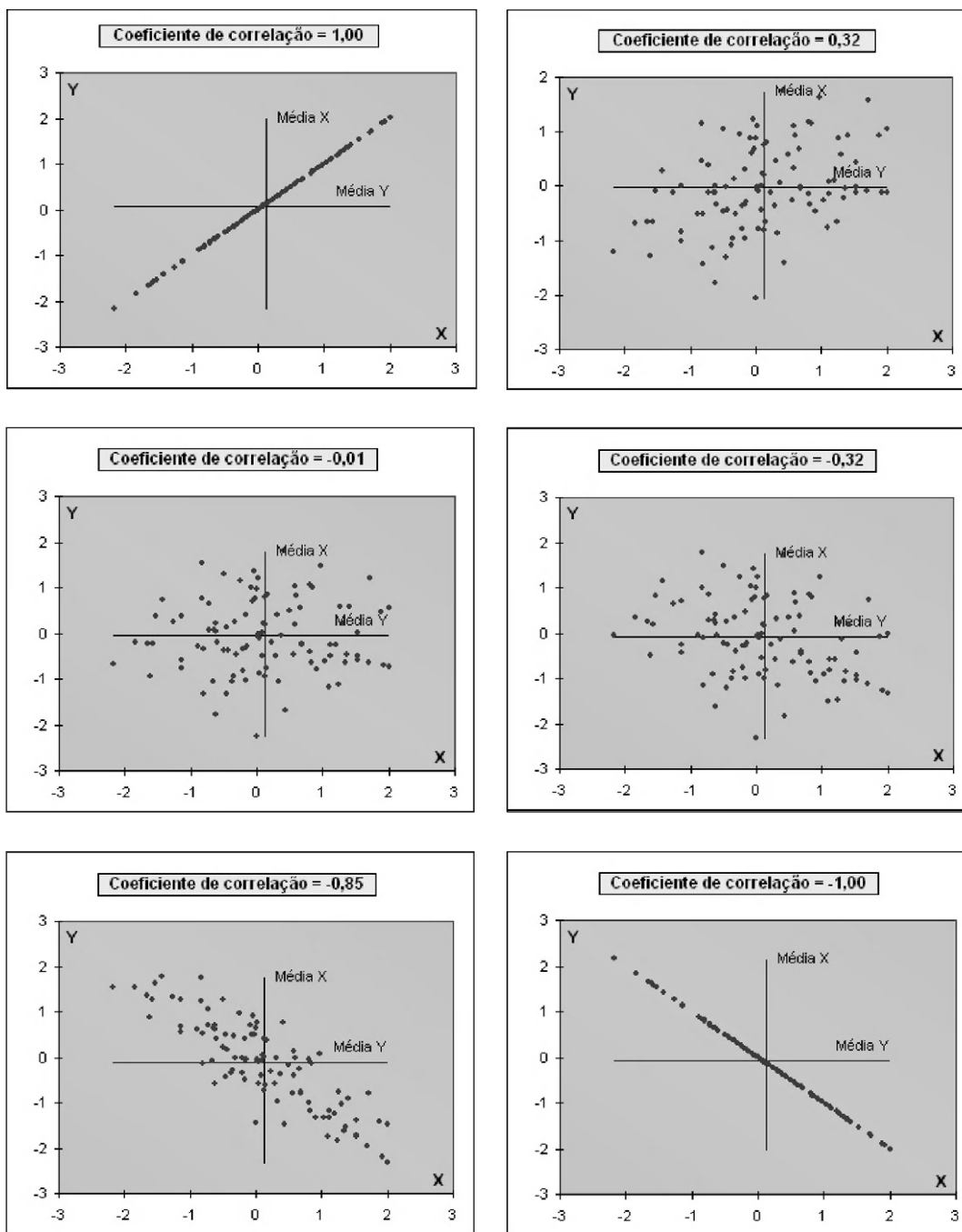


FIGURA 6.7 Simulação de valores do Coeficiente de Correlação de duas amostras.

Correlação e causalidade

As características descritas anteriormente mostram que o coeficiente de correlação não mede a relação causa-efeito entre as variáveis, apesar de essa relação poder estar presente. Por exemplo, uma correlação fortemente positiva entre as variáveis X e Y não autoriza afirmar que variações da variável X provocam variações na variável Y , ou vice-versa. O coeficiente de correlação sozinho não identifica a relação causa-efeito entre as duas variáveis; entretanto, na regressão linear, a relação causa efeito é definida no início da análise. Em alguns casos, a relação causa-efeito pode ser provocada por um ou mais fatores ocultos, uma variável não considerada na análise. Por exemplo, suponha que o número de vendas diá-

rias de um jornal e a produção diária de ovos tenham uma forte correlação positiva. Não se pode afirmar que o aumento da produção de ovos seja a causa do aumento do número de jornais vendidos, nem que o aumento do número de jornais vendidos resulte no aumento da produção de ovos! Para compreender a forte e positiva correlação, devem-se procurar fatores ocultos, por exemplo, o aumento de riqueza da população, que resulta em aumento de demanda dos dois produtos ao mesmo tempo, jornais e ovos.

Anomalias do coeficiente de correlação

Se o coeficiente de correlação for igual a mais um, os pares de valores das variáveis fazem parte de uma reta com declividade positiva. À medida que os pares de valores se afastam dessa reta, o coeficiente de correlação diminuirá de mais um em direção a menos um, passando pelo valor zero, simulação que pode ser facilmente realizada com o *modelo Simulador de Correlação* apresentado na seção anterior. Quanto a essa imagem de linearidade, você deve tomar alguns cuidados, pois há casos em que um coeficiente de correlação próximo de um ou de menos um não significa que a maioria dos pares de valores esteja contida em uma reta. A seguir, comentamos alguns casos registrados com mais detalhes na planilha *Anomalias*, incluída na pasta *Capítulo 6*.

Coeficiente de correlação próximo de +1

A Figura 6.8 mostra alguns casos comentados a seguir.

- Os pares de valores estão contidos numa curva crescente, por exemplo, como a função matemática $Y = 0,10 \times X^2$ mostrada no gráfico esquerdo da primeira linha da Figura 6.8. Nesse caso, o coeficiente de correlação das variáveis X e Y no intervalo $1 \leq X \leq 10$ é igual a 0,97.

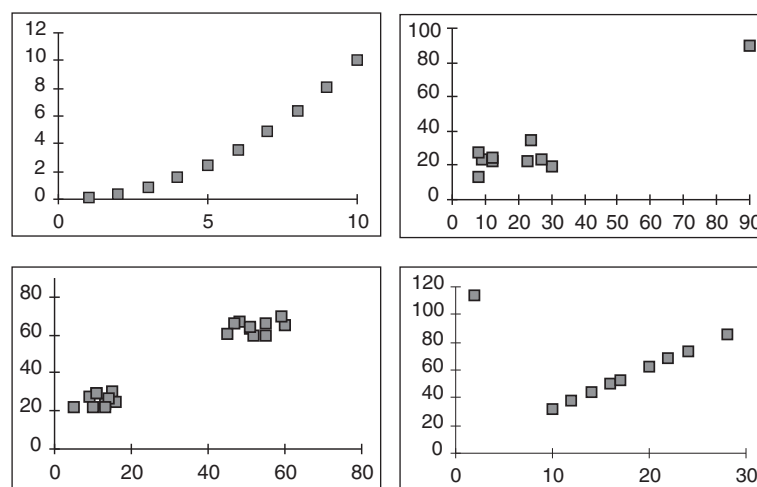


FIGURA 6.8 Anomalias no resultado do coeficiente de correlação.

- O coeficiente de correlação dos pontos do gráfico da direita da primeira linha da Figura 6.8 é 0,92. Um dos pares de valores é um dado suspeito, (90, 90), pois ele está bastante afastado dos demais pares que estão distribuídos, formando uma nuvem de pontos com coeficiente de correlação pequeno, próximo de zero. Essa forte correlação positiva é provocada pelo dado suspeito que gera uma forte tendência positiva.
- Se o primeiro par de valores (90, 90) registrado no intervalo B18:C18 da planilha *Anomalias* for substituído pelo novo par (9, 32) mais próximo da nuvem de pontos, o coeficiente de correlação diminuirá para próximo de zero.

- No gráfico esquerdo da segunda linha da Figura 6.8, os dados estão distribuídos em dois grupos com $r=0,98$. Em geral, amostras de populações diferentes podem provocar essa situação, os dois grupos geram uma tendência de declividade positiva nesse caso.

Coeficiente de correlação próximo de zero

O coeficiente de correlação das variáveis do gráfico direito da segunda linha é igual a zero com um dado suspeito, pois apenas um par está completamente afastado dos demais pares que estão contidos numa reta com declividade positiva. Removendo o dado suspeito, o coeficiente de correlação será igual a +1.

Analise as consequências dessas anomalias:¹²

- Embora seja recomendado excluir os dados suspeitos, esses dados não devem ser simplesmente desprezados. Deve-se dar a devida atenção à causa de tais anomalias, pois esses dados suspeitos podem ser úteis para descobrir a causa dessa ocorrência.
- A escala dos eixos dos gráficos deve ser escolhida adequadamente. Se a escala for mal escolhida, isso poderá prejudicar a interpretação dos resultados. Para evitar que a dispersão apresente tendência diferente, levando a conclusões incorretas, sugere-se que o limite inferior e superior da escala do eixo sejam próximos dos correspondentes valores mínimo e máximo dos dados.

Como conclusão, o simples conhecimento do coeficiente de correlação não é suficiente se não for construído o gráfico de dispersão e, em alguns casos, obtidas outras medidas estatísticas.

Tabelas de covariâncias e de coeficientes de correlação

A covariância e o coeficiente de correlação sempre se referem a duas variáveis ou amostras. Quando há mais de duas variáveis¹³, é possível aplicar os conceitos estatísticos considerando as variáveis duas a duas. Nesse caso, as covariâncias e os coeficientes de correlação são registrados em uma tabela ou matriz de tamanho definido pelo número de variáveis.¹⁴ Para três variáveis A , B e C , as possíveis covariâncias das três variáveis tomadas duas a duas estão registradas na tabela da Figura 6.9.

	A	B	C
A	$\sigma_{A,A}$	$\sigma_{A,B}$	$\sigma_{A,C}$
B	$\sigma_{B,A}$	$\sigma_{B,B}$	$\sigma_{B,C}$
C	$\sigma_{C,A}$	$\sigma_{C,B}$	$\sigma_{C,C}$

FIGURA 6.9 Tabela de covariâncias de três variáveis.

A tabela da Figura 6.9 pode ser simplificada, considerando que:

- A covariância $\sigma_{A,A}$ é a própria variância de A . Portanto, as covariâncias da diagonal principal da tabela são as variâncias das variáveis.
- A permutação das variáveis não altera o resultado da covariância, pois $\sigma_{A,B}=\sigma_{B,A}$. Como as covariâncias da tabela são simétricas com relação à diagonal principal, a tabela de covariâncias pode ser representada com a diagonal principal e apenas uma das duas metades, por exemplo, a parte inferior como mostra a Figura 6.10.

¹² Métodos Estatísticos para Melhora da Qualidade de Kume H. – Editora Gente, 1993.

¹³ Como é o caso da combinação linear de variáveis aleatórias, tema do Capítulo 9.

¹⁴ Para n variáveis, o número de covariâncias ou coeficientes de correlação diferentes é igual a $C(n,2) = \frac{n!}{2!(n-2)!}$

	A	B	C
A	σ_A		
B	$\sigma_{B,A}$	σ_B	
C	$\sigma_{C,A}$	$\sigma_{C,B}$	σ_C

FIGURA 6.10 Tabela de covariâncias, simplificada.

A tabela dos coeficientes de correlação da Figura 6.11 é obtida da tabela de covariâncias da Figura 6.10, substituindo $\sigma_{B,A}$ pelo seu equivalente $r_{B,A}$ e, da mesma forma, as outras duas covariâncias. As variâncias registradas na diagonal da tabela devem ser substituídas pelo valor um, pois para a variável A,

por exemplo, verifica-se que $r_{A,A} = \frac{\sigma_{A,A}}{\sigma_A \times \sigma_A} = \frac{\sigma_A^2}{\sigma_A^2} = 1$.

	A	B	C
A	1		
B	$r_{B,A}$	1	
C	$r_{C,A}$	$r_{C,B}$	1

FIGURA 6.11 Tabela dos coeficientes de correlação.

Ferramenta de análise *Covariância*

O Excel dispõe da ferramenta de análise *Covariância* para construir tabelas de covariâncias, como mostrada na planilha *Ferramenta Covariância*, incluída na pasta **Capítulo 6**, com os dados do Exemplo 6.1 registrados no intervalo B3:C13 incluindo os nomes das variáveis. O procedimento da ferramenta de análise *Covariância* é:

- No menu **Ferramentas**, escolha **Análise de Dados** e na caixa de diálogo **Análise de dados** escolha *Covariância* na lista de **Ferramentas de análise**. Depois pressione o botão **OK**.
- Para calcular a covariância das duas amostras, preencha a caixa de diálogo **Covariância** como mostra a Figura 6.12.
 - Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Covariância* pertencente à *Ajuda do Excel*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo da ferramenta *Covariância* são:

- **Intervalo de entrada.** Informe o intervalo de células da planilha onde os dados estão registrados, nesse caso o intervalo B3:C13, que inclui as células nas quais foram registrados os títulos *Propaganda* e *Vendas*.
- **Agrupado por.** Selecione **Colunas**, pois as amostras foram registradas em coluna. Em geral, o Excel selecionará automaticamente depois de ter informado o intervalo da amostra.
- **Rótulos na primeira linha.** Tendo escolhido **Colunas** no item anterior, necessariamente selecionaremos **Rótulos na primeira linha**, pois nas primeiras células das séries foram registrados os títulos *Propaganda* e *Vendas*.

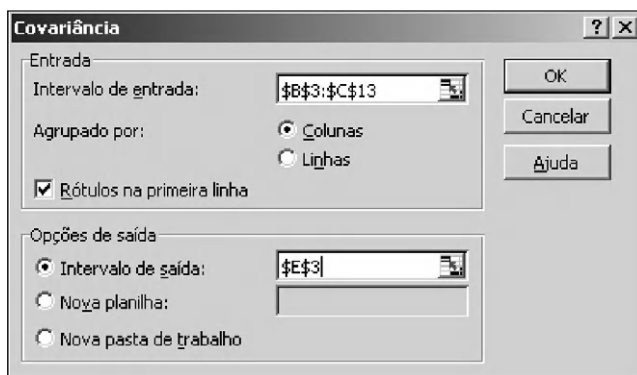


FIGURA 6.12 Caixa de diálogo da ferramenta Covariância.

No quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta *Covariância* registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três *botões de opção* que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso E3. Depois de clicar com o botão esquerdo do mouse dentro da caixa correspondente, o endereço pode ser registrado digitando E3, ou *clcando* com o botão esquerdo do *mouse* na célula E3. Nesse caso, será registrado o endereço com os dois cifrões, \$E\$3. Esse endereço é o da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
 - Se não for informado nenhum endereço, a ferramenta inserirá uma nova planilha com o nome **Plan** seguido de um número sequencial; por exemplo, escolhendo essa alternativa na pasta **Capítulo 6**, a ferramenta inserirá a planilha **Plan1**.
 - Há a alternativa de informar o nome da planilha na caixa desta alternativa; por exemplo, registrando o nome *Teste*, a ferramenta inserirá na mesma pasta uma nova planilha com o nome *Teste*.
- **Nova pasta de trabalho.** Os resultados serão apresentados numa nova pasta e a partir da célula A1 da planilha **Plan1**.

	A	B	C	D	E	F	G
1	Ferramenta de análise Covariância						
2							
3		Propaganda	Vendas				
4		30	430				
5		21	335				
6		35	520				
7		42	490				
8		37	470				
9		20	210				
10		8	195				
11		17	270				
12		35	400				
13		25	480				
14							

		<i>Propaganda</i>	<i>Vendas</i>
Propaganda		101,2	
Vendas		985,5	12995

FIGURA 6.13 Resolução do Exemplo 6.1 com a ferramenta Covariância.

Depois de realizar as escolhas e pressionar o botão **OK**, a ferramenta registra a *tabela de covariâncias* a partir da célula E3, Figura 6.13. Verifique que a covariância como as variâncias obtidas com a ferramenta de análise *Covariância* referem-se à população.

Ferramenta de análise *Correlação*

A ferramenta de análise *Correlação* tem o mesmo formato e procedimento operacional da ferramenta *Covariância*. Dessa maneira, serão apresentadas apenas as diferenças importantes. Para utilizar a ferramenta de análise *Correlação*, foi preparada a planilha **Ferramenta Correlação**, incluída na pasta **Capítulo 6**, com os dados do Exemplo 6.1 registrados no intervalo B3:C13, incluindo os nomes das variáveis. O procedimento da ferramenta de análise *Correlação* é:

- No menu **Ferramentas**, escolha **Análise de Dados** e, na caixa de diálogo **Análise de dados**, escolha *Correlação* na lista de **Ferramentas de análise**. Depois pressione o botão **OK**.
- Para calcular a correlação das duas amostras, preencha a caixa de diálogo *Correlação*, como mostra a Figura 6.14.
- Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Correlação* pertencente à *Ajuda do Excel*.

Depois de realizar as escolhas e pressionar o botão **OK**, a ferramenta registra a *tabela de correlações* a partir da célula E3, Figura 6.15.

FIGURA 6.14 Caixa de diálogo da ferramenta *Correlação*.

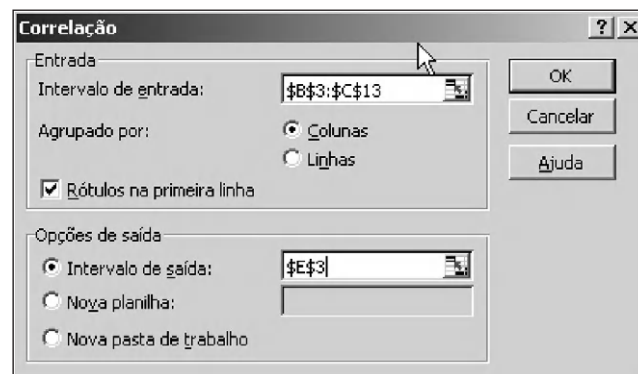


FIGURA 6.15 Resolução do Exemplo 6.1 com a ferramenta *Correlação*.

	A	B	C	D	E	F	G
1	Ferramenta de análise Correlação						
2							
3		Propaganda	Vendas			Propaganda	Vendas
4		30	430			Propaganda	1
5		21	335			Vendas	0,85936613
6		35	520				
7		42	490				
8		37	470				
9		20	210				
10		8	195				
11		17	270				
12		35	400				
13		25	480				
14							

EXEMPLO 6.5

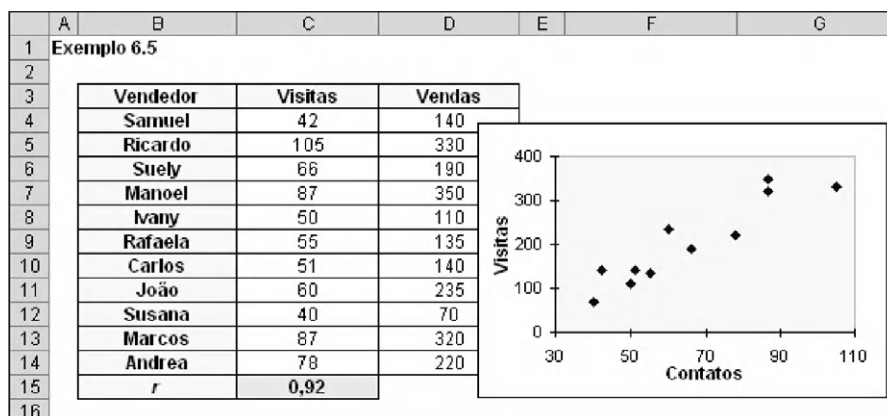
A venda dos produtos tem forte relação com as visitas realizadas pelos vendedores aos seus clientes, é o que afirma o gerente de vendas. A empresa tem onze vendedores e, como regra, eles visitam seus clientes uma vez por mês. Para tentar confirmar sua afirmativa, o gerente preparou a tabela com as visitas realizadas e as vendas de cada vendedor durante o mês passado. O objetivo é verificar se os dados confirmam a crença do gerente.

Solução. Na planilha **Exemplo 6.5**, incluída na pasta **Capítulo 6**, foram registradas as séries de dados, construído o gráfico de dispersão e calculado o coeficiente de correlação das duas amostras, como mostra a figura seguinte.

- O gráfico de dispersão mostra que a maioria dos pares de valores das amostras *Visitas* e *Vendas* se aproxima de uma reta com declividade positiva, confirmando a existência de uma relação forte entre as visitas dos

vendedores e as vendas dos produtos oferecidos. Podemos aceitar que mais visitas geram mais vendas, mas sem poder definir o número de visitas.

- O coeficiente de correlação 0,92 mostra uma forte correlação positiva entre as duas variáveis e parece que confirma a crença do gerente de vendas.



EXEMPLO 6.6

Construa a tabela dos coeficientes de correlação da Tabela de Índices de Preços¹⁵ registrada na planilha **Índices de preço**, incluída na pasta **Capítulo 6**. Embora esses índices tenham metodologias e períodos de coletas de preços diferentes e as séries sejam pequenas, apenas dez dados cada uma, é uma oportunidade interessante para aplicar a ferramenta de análise *Correlação* e analisar as relações entre as séries de índices.

Solução. Com a ferramenta de análise *Correlação*, foi construída a tabela dos coeficientes de correlação a partir da célula B16 da planilha. Analisemos os resultados de maior destaque:

- A menor correlação $r=0,26$ ocorre entre os índices IPCA-E e ICV, que mostra uma fraca correlação entre esses dois índices de preço.
- O índice IPCA-E mantém uma baixa correlação com os outros índices.
- As duas maiores correlações ocorrem com os índices IGPM e IGP-DI e IPA-M e IGP-DI.

	A	B	C	D	E	F	G	H	I	J
1	Índices de preços (%)									
2										
3										
4		Mês	IBGE INPC	Fipe IPC	DIEESE ICV	IBGE IPCA	FGV IGP-DI	FGV IGPM	FGV IPA-M	IBGE IPCA-E
5		jun/03	-0,06	-0,16	-0,26	-0,15	-0,70	-1,00	-1,82	0,22
6		jul/03	0,04	-0,08	0,35	0,20	-0,20	-0,42	-0,75	-0,18
7		ago/03	0,18	0,63	-0,15	0,34	0,62	0,38	0,20	0,27
8		set/03	0,82	0,84	1,26	0,78	1,05	1,18	1,54	0,57
9		out/03	0,39	0,63	0,47	0,29	0,44	0,38	0,36	0,66
10		nov/03	0,37	0,27	0,26	0,34	0,48	0,49	0,57	0,17
11		dez/03	0,54	0,42	0,32	0,52	0,60	0,61	0,64	0,46
12		jan/04	0,83	0,65	1,46	0,76	0,80	0,88	0,98	0,68
13		fev/04	0,39	0,19	-0,18	0,61	1,08	0,69	0,79	0,90
14		mar/04	0,57	0,12	0,47	0,47	0,93	1,13	1,33	0,40
15										
16			INPC	IPC	ICV	IPCA	IGP-DI	IGPM	IPA-M	IPCA-E
17		INPC	1							
18		IPC	0,70	1						
19		ICV	0,82	0,60	1					
20		IPCA	0,91	0,68	0,69	1				
21		IGP-DI	0,80	0,65	0,42	0,90	1			
22		IGPM	0,89	0,67	0,57	0,90	0,97	1		
23		IPA-M	0,90	0,66	0,59	0,91	0,96	1,00	1	
24		IPCA-E	0,62	0,52	0,26	0,62	0,68	0,59	0,59	1
25										

15 Tabela publicada no jornal *O Estado de São Paulo*, 02 de junho de 2004.

Problemas

Problema 1

O dono da oficina mecânica regulou seu carro e, em um dia sem muito movimento na estrada, realizou as medições de consumo de combustível registradas na tabela seguinte para seis velocidades diferentes. Construa o gráfico de dispersão e analise o comportamento das duas variáveis, *Velocidade* e *Consumo*.

Velocidade	Consumo – Km/l
70	10,2
80	9,7
90	9,1
100	8,3
110	7,8
120	7

Problema 2

Calcule a covariância da população e da amostra da relação *Velocidade* e *Consumo* do Problema 1.

R: $Cov(Vel, Con) = -18,75$ e $Cov(Vel, Con) = -22,5$

Problema 3

Calcule o coeficiente de correlação da relação *Velocidade* e *Consumo* do Problema 1 utilizando a fórmula para os dois casos, população e amostra.

R: $r = -0,9975$

Problema 4

O seguinte trecho foi extraído do jornal:¹⁶ “Para cada 1% de aumento no PIB o impacto no nível de emprego é de 0,4%. ... Entretanto, a criação de vagas formais ainda não é suficiente para reduzir significativamente o desemprego no País ... Para isso o PIB precisaria crescer em média 5% ao ano”. Responda às seguintes perguntas:

- Qual é o tipo de relação entre PIB e Emprego?
- Qual deve ser o impacto no emprego de um crescimento de PIB de 5% ao ano?

R: a) Correlação positiva. b) 2%

Problema 5

Supondo que durante cinco anos se mantenha a relação do Problema 4, para cada 1% de aumento no PIB, o impacto no nível de emprego é de 0,4%.

- Construa a tabela com o crescimento do PIB, começando por 100, e o crescimento do Emprego, começando por 60, mantendo ambas as variáveis com crescimento anual constante.
- Construa o gráfico de dispersão correspondente.
- Calcule o valor do coeficiente de correlação.

R: c) $r = 0,999991$ arredondando $r = 1$

Problema 6

Analisando o resultado do Problema 5, você concorda com as seguintes afirmações?

- As variáveis PIB e Emprego têm uma perfeita correlação positiva.
- Os pares de dados são pontos de uma linha reta com declividade positiva.

¹⁶ Criação de emprego acompanha alta do PIB, artigo de Cleide Silva publicado no jornal O Estado de São Paulo de 2 de junho de 2004.

Problema 7

A tabela seguinte registra os retornos das ações tipo ON e PN de um grupo de empresa. Com esses dados:

- Construa o gráfico de dispersão correspondente.
- Calcule a covariância da população e o coeficiente de correlação.
- Analise os resultados e verifique se há algum dado suspeito.

ON%	PN%	ON%	PN%
37,5	20,9	212,5	367,1
-45,0	5,4	46,3	6,9
0,0	49,4	11,1	45,4
31,5	31,1	43,0	27,8
-1,0	30,0	67,0	43,1
20,1	28,0	9,4	13,4

R: $Cov(ON, PN) = 5.083,84$ e $r = 0,8942$

Problema 8

Repita o Problema 7, porém sem considerar os retornos da empresa com ON% de 212,5 e PN% de 367,1.

Problema 9

O professor passou para os alunos uma folha com duas séries de dados para calcular o coeficiente de correlação e explicar o tipo de relação entre as duas séries. Seu colega rapidamente calculou o valor do coeficiente de correlação igual a zero e afirmou que as duas séries não apresentam nenhuma relação. Concorde com a afirmação de seu colega de que as duas séries não apresentam nenhuma relação? Por quê?

Problema 10

Na tabela seguinte, foram registrados sete pares de valores correspondentes aos resultados de um teste de aptidão. Com esses dados e sem construir o gráfico de dispersão nem calcular o coeficiente de correlação:

- Você conseguiria definir a relação e a tendência dessas duas séries?
- Você conseguiria determinar o valor desse coeficiente?
- Se for possível, qual o valor do coeficiente de correlação?

A	0	6	4	12	8	2	10
B	35	26	29	17	23	32	20

Problema 11

Continuando com o Problema 10.

- Construa o gráfico de dispersão correspondente.
- Calcule o coeficiente de correlação.

Problema 12

A diferença de idades dos irmãos Ana e João é de 5 anos. Considerando a série de dados dos anos de Ana, começando por 10 e terminando com 15, e a série de dados de João, começando com 6 e terminando com 11. Sem construir o gráfico de dispersão nem calcular o coeficiente de correlação:

- Você conseguiria definir a relação e a tendência dessas duas séries?
- Você conseguiria determinar o valor desse coeficiente?
- Se for possível, qual o valor do coeficiente de correlação?

Problema 13

Analise a relação entre as amostras X e Y registradas na tabela seguinte, sugerindo começar pela construção do gráfico de dispersão.

X	10	15	18	12	9
Y	21	15	12	18	20

Problema 14

Os *prêmios e preços de exercícios* de cinco séries de opções de compra com mesmo vencimento estão registrados na tabela seguinte. Construa o gráfico de dispersão, calcule o coeficiente de correlação e analise os resultados.

Prêmios	Preços de exercício
\$257,52	\$2.100
\$99,25	\$2.200
\$38,17	\$2.300
\$14,65	\$2.400
\$5,61	\$2.500

R: $r=-0,8933$

Problema 15

Na planilha **Problemas**, incluída na pasta **Capítulo 6**, está registrada a tabela com a relação dos dez maiores e melhores grupos de supermercados no ano 1991, porém sem indicar os nomes das empresas. Construa e analise a tabela dos coeficientes de correlação.

Problema 16

Na planilha **Problemas**, incluída na pasta **Capítulo 6**, está registrada a tabela com a relação dos dez maiores e melhores grupos de supermercados no ano 1998, porém sem indicar os nomes das empresas. Construa e analise a tabela dos coeficientes de correlação.

Problema 17

Tomando como base os resultados e as análises das empresas do Problema 15, analise a evolução dessas empresas entre os anos 1991 e 1998, comparando os resultados de 1998 com os de 1991.

Apêndice 1

Outra forma de calcular a covariância

Partindo da fórmula da covariância que repetimos em seguida:

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y)$$

Desenvolvendo o produto indicado temos:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y) = \frac{1}{N} \sum_{i=1}^N (X_i Y_i - X_i \mu_Y - \mu_X Y_i + \mu_X \mu_Y)$$

$$\frac{1}{N} \sum_{i=1}^N (X_i \mu_X) \times (Y_i - \mu_Y) = \frac{1}{N} \left(\sum_{i=1}^N X_i Y_i - \sum_{i=1}^N X_i \mu_Y - \sum_{i=1}^N \mu_X Y_i + \sum_{i=1}^N \mu_X \mu_Y \right)$$

Simplificando as parcelas do segundo membro temos:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y) = \frac{1}{N} \left(\sum_{i=1}^N X_i Y_i - \mu_Y \sum_{i=1}^N X_i - \mu_X \sum_{i=1}^N Y_i + N \mu_X \mu_Y \right)$$

$$\frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y) = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y$$

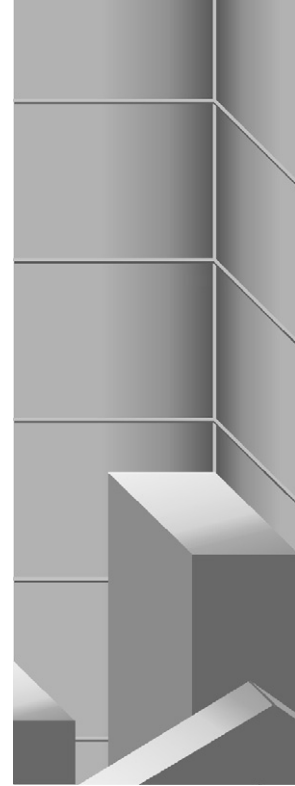
Depois de simplificar as três últimas parcelas do segundo membro temos:

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y) = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \mu_Y \mu_X$$

Essa última expressão será utilizada no Apêndice 1 do Capítulo 9.

Capítulo 7

VARIÁVEIS ALEATÓRIAS E DISTRIBUIÇÕES DISCRETAS



O resultado do lançamento de uma moeda pode ser utilizado para tomar decisões; por exemplo, o árbitro de uma partida de futebol sorteia quem inicia o primeiro tempo do jogo e ainda o ganhador do sorteio escolhe a metade do campo onde sua equipe iniciará o jogo. Outras vezes, o resultado da moeda é para realizar uma tarefa, agradável ou não etc. Embora o resultado do sorteio possa ser utilizado com diferentes finalidades, o experimento lançamento aleatório de uma moeda permanece o mesmo, mantendo os mesmos resultados.

Lembremos que um experimento é aleatório se não for possível antecipar seu resultado, apesar de conhecer todos os resultados possíveis que define o espaço amostral do experimento. Portanto, cada vez que o experimento for repetido, seu resultado pertencerá a esse espaço amostral sendo cada resultado denominando ponto amostral, que não pode ser particionado nem dividido. Ainda no Capítulo 5, foi visto que evento elementar é aquele que contém um único ponto amostral com uma determinada probabilidade de ocorrer. Em vez de operar com o espaço amostral, agora utilizaremos um conceito mais amplo denominado *variável aleatória*, que adota valores de acordo com os resultados de um experimento aleatório.

Variável aleatória VA é uma variável cujo valor é o resultado numérico de um experimento aleatório.

Uma VA é uma função formada por valores numéricos definidos sobre o espaço amostral de um experimento, tendo presente que:

- Para cada resultado do experimento aleatório corresponderá apenas um único valor numérico da VA. Todavia, um valor numérico da VA poderá corresponder a um ou mais resultados de um experimento.
- Dependendo dos valores numéricos, a variável aleatória poderá ser discreta ou contínua.¹

¹ Por exemplo, o número de peças rejeitadas por lote em uma linha de produção é uma VA discreta, e o lucro líquido mensal de uma empresa é uma VA contínua. Entretanto, nem sempre a separação entre variável discreta e variável contínua fica clara.

- Se os valores numéricos da VA se referem a contagens, então a VA será uma variável aleatória discreta.
- Se os valores numéricos da VA pertencem ao conjunto dos números reais, então a VA será uma variável aleatória contínua.

Variáveis aleatórias discretas

Para definir uma VA de forma completa, será necessário especificar as probabilidades e os valores dos eventos elementares do espaço amostral do experimento aleatório. Iniciamos apresentando as funções de probabilidade com uma variável aleatória discreta.

EXEMPLO 7.1

Defina a variável aleatória X que representa o número de caras do experimento aleatório: lançamento de uma moeda três vezes seguidas.

Solução. O número de caras possíveis do experimento lançamento de uma moeda três vezes seguidas são 0, 1, 2 e 3 caras que fazem parte dos oito eventos elementares do espaço amostral.

Determinação dos valores x da variável aleatória X .

O conjunto formado por esses quatro números $\{0, 1, 2, 3\}$ é o conjunto dos valores numéricos x da variável aleatória X . Relacionando os oito eventos elementares do espaço amostral com os valores x da VA do experimento:

- Se $x=0$, nenhuma cara, o único resultado possível é: CoCoCo.
- Se $x=1$, uma cara, os resultados possíveis são: CaCoCo, CoCaCo e CoCoCa.
- Se $x=2$, duas caras, os resultados são: CaCaCo, CaCoCa e CoCaCa.
- Se $x=3$, todas caras, o único resultado possível é: CaCaCa.

Determinação da probabilidade dos valores x da variável aleatória X .

Para determinar as probabilidades dos valores x de X , deve-se considerar que:

- Os oito resultados ou eventos elementares do experimento aleatório são igualmente prováveis com probabilidade $1/8$ ou 12,50%.
- Como os eventos elementares são mutuamente excludentes, pela regra da soma, a probabilidade de ocorrer uma cara será $3/8$ ou 37,50%, obtido como soma das probabilidades dos três eventos elementares com uma cara:

$$P(x = 1) = P(\text{CaCoCo}) + P(\text{CoCaCo}) + P(\text{CoCoCa})$$

$$P(x = 1) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

- Da mesma maneira, a probabilidade de ocorrerem duas caras será $3/8$ ou 37,50%, e a probabilidade de ocorrerem três caras será $1/8$ ou 12,50%.

Definição da variável aleatória X .

A variável aleatória X está definida pelos seus valores numéricos x_i e suas probabilidades associadas $p(x_i)$, como apresentado na tabela seguinte.

x	$p(x)$
0	12,5%
1	37,5%
2	37,5%
3	12,5%

A tabela do Exemplo 7.1 mostra que uma VA é representada pela sua distribuição de probabilidades. Observe que:

- Essa VA foi obtida a partir de uma população conhecida.
- Uma VA representa uma distribuição de frequências relativas, como mostra o histograma da Figura 7.1.

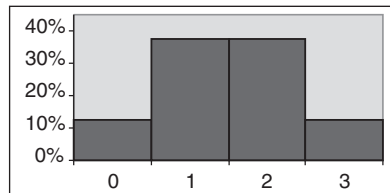


FIGURA 7.1
Histograma
do Exemplo 7.1.

VA de cenários

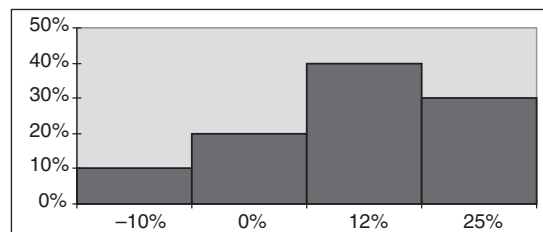
Em alguns casos, a variável aleatória pode ser gerada a partir de cenários definidos pela opinião de um grupo de pessoas acerca de um determinado assunto, por exemplo, os resultados da empresa nos próximos dois anos, o retorno do mercado de ações nos próximos doze meses etc.

EXEMPLO 7.2

Na reunião anual, antes de terminar o ano, o consenso do grupo de analistas definiu os retornos possíveis do mercado de ações nos próximos doze meses e suas probabilidades associadas de acordo com quatro possíveis cenários e os resultados registrados na tabela seguinte.

Cenário	Retorno	Probabilidade
Ruim	-10%	10%
Regular	0%	20%
Bom	+12%	40%
Excelente	+25%	30%

Com a distribuição de frequências relativas apresentada nessa tabela, os analistas definiram a variável aleatória X cujo histograma é mostrado a seguir.



Definição da VA discreta

Os resultados dos Exemplos 7.1 e 7.2 ajudam a estabelecer a definição de variável aleatória discreta:

- A VA discreta X tem o conjunto de valores $X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$.

- Cada valor x_i de X tem associada a probabilidade $p(x_i)$, formando a distribuição de frequências registrada na tabela seguinte:

X	x_1	x_2	\dots	x_i	\dots	x_n
$p(x)$	$p(x_1)$	$p(x_2)$	\dots	$p(x_i)$	\dots	$p(x_n)$

As probabilidades $p(x_i)$ de cada x_i atendem às seguintes premissas:

- Todos os valores x de X têm um valor de probabilidade no intervalo $(0, 1)$ ou de outra maneira $0 \leq p(x_i) \leq 1$.
- A soma das probabilidades de todos os x de X é sempre igual a um; de outra maneira:

$$p(x_1) + p(x_2) + \dots + p(x_n) = \sum_{i=1}^n p(x_i) = 1$$

Valor esperado da VA

Observe que na distribuição de frequências da variável aleatória não é conhecido o número de dados utilizados. Ao mesmo tempo, as frequências relativas da VA do Exemplo 7.2 podem pertencer a variáveis com tamanhos diferentes. Por exemplo, em uma variável com cem dados e as frequências relativas do Exemplo 7.2:

- O valor $-0,10$ teria frequência 10.
- O valor $0,00$ teria frequência 20.
- O valor $+0,12$ teria frequência 40.
- O valor $+0,25$ teria frequência 30.

Como é possível repetir esse exemplo para outra variável com qualquer tamanho de dados, pode-se entender que uma VA é um resumo de uma das muitas séries de dados equivalentes cuja distribuição da população é idêntica à distribuição da amostra. Portanto, todas as variáveis que tiverem a mesma distribuição de frequências² terão as mesmas medidas descritivas, por exemplo, a mesma média. O conceito *valor esperado* aplicado em uma VA é a medida descritiva equivalente à média de uma amostra, ou média ponderada. O valor esperado, 11,30%, do Exemplo 7.3, é a própria média da variável.

Seja a variável aleatória X com valores numéricos x_1, x_2, \dots, x_n e probabilidades associadas $p(x_1), p(x_2), \dots, p(x_n)$. O valor esperado $E[X]$ da variável X é definido por:

$$E[X] = x_1 p(x_1) + x_2 p(x_2) + \dots + x_i p(x_i) + \dots + x_n p(x_n)$$

$$E[X] = \sum_{i=1}^n x_i p(x_i)$$

EXEMPLO 7.3

Calcule o *valor esperado* $E[X]$ da variável aleatória X do Exemplo 7.2.

Solução. O valor esperado de X é $E[X]=11,30\%$, obtido com a fórmula:

² Distribuição de frequências pode ser utilizada como sinônimo de distribuição de probabilidades.

$$E[X] = \sum_{i=1}^4 x_i p(x_i) = x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3) + x_4 p(x_4)$$

$$E[X] = (-0,10) \times 0,10 + 0 \times 0,20 + 0,12 \times 0,40 + 0,25 \times 0,30$$

$$E[X] = 0,1130$$

Observe que, no procedimento manual de cálculo, os valores da VA e suas probabilidades associadas são utilizados de forma unitária em vez de porcentagens. Uma forma prática de calcular $E[X]$ é registrada na planilha **Exemplo 7.3**, incluída na pasta **Capítulo 7**.

	A	B	C	D	E	F
1	Exemplo 7.3					
2						
3		Cenário	X	p(x)	x.p(x)	
4		Ruim	-10%	10%	-0,0100	
5		Regular	0%	20%	0,0000	
6		Bom	12%	40%	0,0480	
7		Excelente	25%	30%	0,0750	
8						
9		Resultados				
10		E[X]	0,1130	=SOMA(E4:E7)		
11		E[X]	0,1130	=C4*D4+C5*D5+C6*D6+C7*D7		
12		E[X]	0,1130	=SOMAPRODUTO(C4:C7;D4:D7)		
13						

Nas colunas C e D da planilha, foram repetidos os dados registrados na planilha Exemplo 7.2 da mesma pasta, mantendo a formatação de porcentagem nas células.

- Na célula E4, foi registrada a fórmula =C4*D4, que depois foi copiada até a célula E7. Essas fórmulas calculam e registram o resultado do produto de cada valor da variável X pela sua probabilidade associada.
- Na célula C10, foi registrada a fórmula =SOMA(E4:E7), que calcula a soma dos produtos e é igual a $E[X]=0,1130$. Nesse caso, mantivemos a formatação unitária do resultado, entretanto, o leitor poderia formatar a célula em %.

Como alternativa de cálculo, que evita o registro dos resultados parciais das parcelas da fórmula do valor esperado no intervalo E4:E7:

- Na célula C11, foi registrada a fórmula =C4*D4+C5*D5+C6*D6+C7*D7, que calcula a soma dos produtos diretamente em uma única célula da planilha.
- O resultado do valor esperado pode ser obtido utilizando a função SOMAPRODUTO do Excel. Na célula C12, foi registrada a fórmula =SOMAPRODUTO(C4:C7;D4:D7).

Simulador média de longo prazo

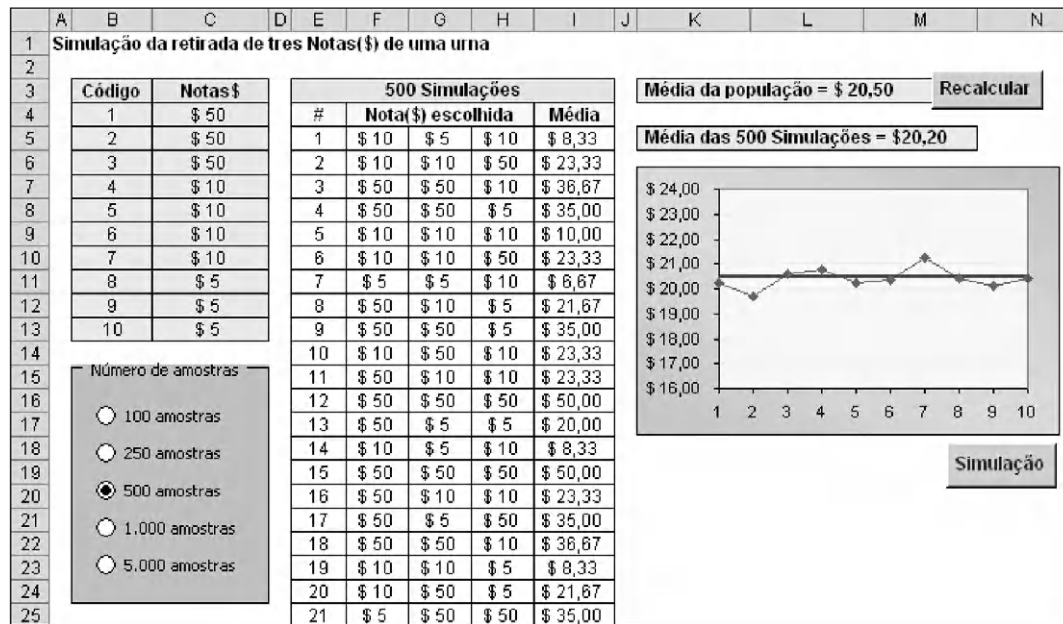
Qual é o significado do valor esperado 11,30% do Exemplo 7.3? Esse valor deve ser percebido da seguinte maneira: se o experimento aleatório for repetido um número muito grande de vezes, então a média de todos os resultados será igual a 11,30%. Por isso, o valor esperado é também denominado *média de longo prazo*. Para compreender esse conceito, foi construída a planilha **Simulação**, incluída na pasta **Capítulo 7**, e baseada na retirada de Notas(\$) de uma urna.

Em uma urna foram depositadas dez bolas iguais representando Notas(\$) de dinheiro:³ três bolas de valor \$50, quatro bolas de valor \$10 e três bolas de valor \$5. Sendo as probabilidades de retirar uma bola de valor \$50, \$10 ou \$5, respectivamente, 3/10, 4/10 e 3/10, o valor esperado da população é \$20,50, obtido com a fórmula: $E[X] = \frac{3}{10} \times \$50 + \frac{4}{10} \times \$10 + \frac{3}{10} \times \$5 = \$20,50$.

³ Usamos bolas no lugar das notas para facilitar a retirada de amostras aleatórias probabilísticas com reposição de uma urna.

O experimento consiste em retirar da urna uma amostra aleatória probabilística de três bolas com reposição e depois calcular e registrar a média dessa amostra. Por exemplo, se na primeira rodada foram retiradas as bolas \$50, \$5 e \$10, a média dessa amostra será \$21,67. Se o experimento for repetido um número muito grande de vezes, pela lei dos grandes números, a média das médias das amostras retiradas tenderá ao valor \$20,50, que corresponde ao valor esperado da população formada pelas três bolas.

FIGURA 7.2 Simulação da retirada de três notas de uma urna.



- As dez Notas(\$) formadas por três notas de \$50, quatro notas de \$10 e três notas de \$5 foram identificadas na planilha com os números um a dez, como mostra o intervalo B4:C13 da Figura 7.2.
- As amostragens aleatórias são realizadas utilizando os números aleatórios gerados pela função =ALEATÓRIOENTRE(1;10), como se pode ver a partir da linha cinco das fórmulas das colunas F, G e H.
- Depois de seleccionar um dígito aleatório entre um e dez, com a função PROCV é seleccionado o valor \$ correspondente na tabela do intervalo B4:C13.
- A partir da célula I5 é calculada e registrada a média de cada amostra.

A planilha está preparada para gerar 100, 250, 500, 1.000 ou 5.000 amostras, conforme seja seleccionado o botão de opção correspondente dentro da caixa de grupo *Número de amostras*. Cada vez que for seleccionado um número de amostras, é ativada uma macro que prepara a planilha para essa nova quantidade e apresenta:

- O resultado dessa simulação na célula K5.
- Em sequência, são realizadas dez simulações completas cujos resultados são apresentados no gráfico, onde aparecem as médias dessas dez simulações seguidas e a linha reta com a média da população \$20,50.

O botão **Recalcular** é utilizado para realizar apenas uma simulação, e o botão **Simulação** realiza dez simulações seguidas e atualiza o gráfico. O gráfico mostra que à medida que o tamanho da amostra aumenta, de 100 para 250, 500, 1.000 e 5.000, a média das médias das amostras diminui sua variabilidade e cada vez mais se aproxima do valor esperado da população. Realizando amostragens extremas, esse

fato fica bem acentuado, o que mostra que o valor esperado é uma média de longo prazo, ressaltando que, embora 5.000 seja um número grande de amostras, não é suficiente para aceitá-lo como longo prazo. No Capítulo 5, vimos que, na realidade, não se está em busca da média real de \$20,50, mas da probabilidade de que o erro entre a média observada e a média da população seja inferior a um certo erro tolerado. Tudo que a lei nos informa é que a média de um grande número de simulações diferirá por menos que certa quantidade especificada da média real e mais provavelmente do que a média de um pequeno número de simulações. Isso não significa que não haverá erro de um número muito grande de simulações.

EXEMPLO 7.4

O seguro de vida da seguradora LIFE para pessoas com menos de 40 anos é \$200.000, devendo-se pagar \$600 por ano. Se a probabilidade de uma pessoa com menos de 40 anos morrer no próximo ano for 0,1%, qual a expectativa do lucro anual da seguradora?

Solução. Os dados mostram que a variável aleatória X deste seguro tem dois eventos elementares:

- A probabilidade de a pessoa não morrer durante o ano é $p(x_1)=99,9\%$. Nesse caso, a seguradora ganha $x_1=\$600$.
- A probabilidade de a pessoa morrer durante o ano é $p(x_2)=0,1\%$. Nesse caso, a seguradora perde $x_2=-\$199.400=\$600-\$200.000$

O valor esperado do lucro anual da seguradora é $E[X]=\$400$ obtido de:

$$E[X] = \sum_{i=1}^2 x_i p(x_i) = x_1 p(x_1) + x_2 p(x_2)$$

$$E[X] = -\$199.400 \times 0,001 + \$600 \times 0,999 = \$400$$

Qual o significado desse resultado? Se a quantidade de seguros desse tipo vendidos anualmente pela seguradora for um número muito grande, então o lucro médio anual da seguradora será igual a \$400 por seguro vendido.

Variância e desvio padrão da variável aleatória discreta

O conceito valor esperado aplicado em uma VA é a medida descritiva equivalente à média de uma amostra, ou média ponderada. Considerando que a variância é a média dos desvios ao quadrado, o conceito de valor esperado pode ser utilizado para calcular a variância e depois o desvio padrão de uma variável aleatória.

Seja a variável aleatória X com valores numéricos x_1, x_2, \dots, x_n e probabilidades associadas $p(x_1), p(x_2), \dots, p(x_n)$. Definimos como:

Variância da variável X :

$$\sigma_X^2 = (x_1 - E[X])^2 p(x_1) + \dots + (x_n - E[X])^2 p(x_n)$$

$$\sigma_X^2 = \sum_{i=1}^n (x_i - E[X])^2 p(x_i)$$

Desvio padrão de X : $\sigma_X = \sqrt{\sigma_X^2}$

Deve-se destacar que:

- A variância pode ser apresentada como $\sigma_X^2 = E[(X - E[X])^2]$. Essa forma de variância de X como valor esperado é obtida da própria definição de valor esperado de X , substituindo a variável X pela variável $(X - E[X])^2$, onde $E[X]$ passa a ser $E[(X - E[X])^2]$.
- A variância da variável aleatória X pode ser obtida, também, com a fórmula $\sigma_X^2 = E[X^2] - E[X]^2$, como mostrado no Apêndice 1.

EXEMPLO 7.5

Continuando o Exemplo 7.3, calcule a variância e o desvio padrão dessa VA aplicando o conceito de valor esperado.

Solução. A variância de X é $\sigma_X^2 = 0,01274$, obtida com a fórmula, depois de conhecido o valor esperado $E[X]$:

$$\sigma_X^2 = (x_1 - E[X])^2 p(x_1) + \dots + (x_n - E[X])^2 p(x_n)$$

$$\sigma_X^2 = (-0,10 - 0,1130)^2 \times 0,10 + \dots + (0,25 - 0,1130)^2 \times 0,30$$

$$\sigma_X^2 = 0,01274$$

Observe que, no procedimento manual de cálculo, os valores da VA e suas probabilidades associadas são utilizados de forma unitária em vez de percentagens. Uma forma prática de calcular a variância de X é registrada na planilha **Exemplo 7.5**, incluída na pasta **Capítulo 7**.

	A	B	C	D	E	F	G
1	Exemplo 7.5						
2							
3							
4		Cenário	X	p(x)	x-E[X]	(x-E[X])².p(x)	
5		Ruim	-0,10	0,10	-0,2130	0,004537	
6		Regular	0,00	0,20	-0,1130	0,002554	
7		Bom	0,12	0,40	0,0070	0,000020	
8		Excelente	0,25	0,30	0,1370	0,005631	
9		Resultados					
10		E[X]	0,1130	=C4*D4+C5*D5+C6*D6+C7*D7			
11		Variância	0,0127	=SOMA(F4:F7)			
12		Desvio Padrão	0,1129	=RAIZ(C11)			
13							
14		Resultados diretos					
15		Variância	0,0127	{=SOMA(((C4:C7-SOMA(C4:C7*D4:D7))^2*D4:D7))}			
16		Desvio Padrão	0,1129	{=RAIZ(SOMA(((C4:C7-SOMA(C4:C7*D4:D7))^2*D4:D7)))}			
17							

Nas colunas C e D da planilha, foram repetidos os dados registrados na planilha Exemplo 7.2 da mesma pasta, mantendo a formatação porcentagem nas células.

- Na célula C10, foi calculado o valor esperado, como já foi mostrado.
- Na célula E4, foi registrada a fórmula =C4-\$C\$10 que depois foi copiada até a célula E7. Essas fórmulas calculam e registram os desvios da VA.
- Na célula F4, foi registrada a fórmula =E4^2*D4 que depois foi copiada até a célula F7. Essas fórmulas calculam e registram os quadrados dos desvios multiplicados pela probabilidade de cada valor da VA.
- Na célula C10, foi registrada a fórmula =SOMA(F4:F7) que retorna o valor da variância igual a 0,0127, valor arredondado na célula, mas não na sua memória
- O desvio padrão igual a 0,1129 foi calculado a partir da variância registrando a fórmula =RAIZ(C11) na célula C12. Nesse caso, mantivemos a formatação unitária do resultado; entretanto, você poderia formatar a célula em %, pois a unidade de medida do desvio padrão é a dos valores da VA.

Como alternativa de cálculo que evita o registro dos resultados parciais das parcelas da fórmula da variância no intervalo E4:F7:

- A fórmula $\{=SOMA(((C4:C7-SOMA(C4:C7*D4:D7))^2*D4:D7))\}$ foi registrada na célula C15. Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**; mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, temos os resultados apresentados na figura seguinte onde as fórmulas receberam as chaves $\{ \}$.
- O desvio padrão pode ser obtido da variância aplicando a raiz quadrada. Também pode ser obtido registrando a seguinte fórmula como matriz $\{=RAIZ(SOMA(((C4:C7-SOMA(C4:C7*D4:D7))^2*D4:D7)))\}$, sem necessidade de registrar a variância em outra célula.

Distribuição binomial

Muitas variáveis aleatórias têm apenas dois possíveis resultados ou eventos elementares:

- O técnico do controle de qualidade sempre retira uma amostra de dez peças de cada lote recebido do fornecedor. O número de peças que não atendem à especificação é uma variável aleatória X .
- O número de respostas *sim* a uma pergunta da pesquisa aplicada em 1.800 pessoas é uma variável aleatória X .
- O número de ações que ontem subiram comparadas com as 50 ações mais negociadas é uma variável aleatória X .

Nos três exemplos, o número de vezes em que um resultado ocorre durante um determinado número de repetições do experimento é a variável aleatória X .

Premissas de um experimento binomial

O experimento é repetido n vezes, e os n resultados do experimento são independentes.

O experimento tem apenas dois possíveis resultados ou eventos mutuamente exclusivos: *sucesso* ou *falha*.

A probabilidade de *sucesso* do experimento é π e se mantém constante durante as n repetições do experimento. A probabilidade de *falha* do experimento é $(1-\pi)$.

EXEMPLO 7.6

O gerente da loja estima que de dez vendas realizadas, três são microcomputadores e sete equipamentos eletrônicos. Qual a probabilidade de que uma das quatro próximas vendas seja um microcomputador?

Solução. Começamos por determinar as quatro próximas vendas e depois suas probabilidades de ocorrência.

- Sendo E a venda de um equipamento eletrônico e M a de um microcomputador, os quatro possíveis resultados (eventos elementares) são: $EEEM$, $EEME$, $EMEE$ e $EEEE$.
- Dos dados do gerente, deduzimos que 70% das vendas realizadas são de equipamentos eletrônicos E e 30% de microcomputadores M . Se a sequência de venda de um M for $EEEM$, sua probabilidade será igual a:

$$P(EEEM) = 0,70 \times 0,70 \times 0,70 \times 0,30$$

$$P(EEEM) = 0,70^3 \times 0,30 = 0,1029$$

O resultado $P(EEEM)=10,29\%$ foi obtido aplicando a regra do produto, pois os eventos são independentes. Repetindo o mesmo procedimento para a sequência de venda $EEME$, sua probabilidade será igual a:

$$P(EEME) = 0,70 \times 0,70 \times 0,30 \times 0,70$$

$$P(EEME) = 0,70^2 \times 0,30 \times 0,70 = 0,1029$$

As probabilidades das duas sequências restantes têm o mesmo valor obtido das seguintes fórmulas:

$$P(EMEE) = 0,70 \times 0,30 \times 0,70 \times 0,70 = 0,70 \times 0,30 \times 0,70^2 = 0,1029$$

$$P(MEEE) = 0,30 \times 0,70 \times 0,70 \times 0,70 = 0,30 \times 0,70^3 = 0,1029$$

Considerando que os quatro eventos são mutuamente excludentes, a probabilidade de que uma das quatro próximas vendas seja um microcomputador é igual a 41,16%, resultado obtido da regra da soma com a seguinte fórmula, onde $x=1$ identifica a venda de um microcomputador:

$$P(x = 1) = P(EEEM) + P(EEME) + P(EMEE) + P(MEEE)$$

$$P(x = 1) = 0,1029 + 0,1029 + 0,1029 + 0,1029$$

$$P(x = 1) = 0,4116$$

Fórmula da distribuição binomial

O Exemplo 7.6 mostra que a probabilidade $P(x=1)$ pode ser obtida contando os possíveis resultados, agrupando-os na seguinte fórmula.

$$P(x = 1) = 4 \times 0,30^1 \times 0,70^3$$

Utilizando o conceito de combinações:

$$P(x = 1) = \frac{4!}{1!(4-1)!} \times 0,30^1 \times 0,70^{4-1} = 0,4116$$

De forma geral, a probabilidade $P(x)$ de conseguir em n experiências x sucessos com probabilidade π é medida pela fórmula:

$$P(x) = \frac{n!}{x!(n-x)!} \times \pi^x \times (1-\pi)^{n-x}$$

Em cada experiência binomial, será possível obter a probabilidade associada aplicando essa fórmula, sempre que o tamanho da amostra for pequeno comparado com o tamanho da população, em geral menor do que 5%.

Probabilidade da distribuição binomial

Se em n experiências ocorrem $x=1, 2, \dots, n$ sucessos com probabilidade π , a variável X terá distribuição binomial⁴ de probabilidades.

$$P(x) = \frac{n!}{x!(n-x)!} \times \pi^x \times (1-\pi)^{n-x}$$

Nessa expressão $x! = x(x-1)(x-2)\dots(2)(1)$ e $0!=1$

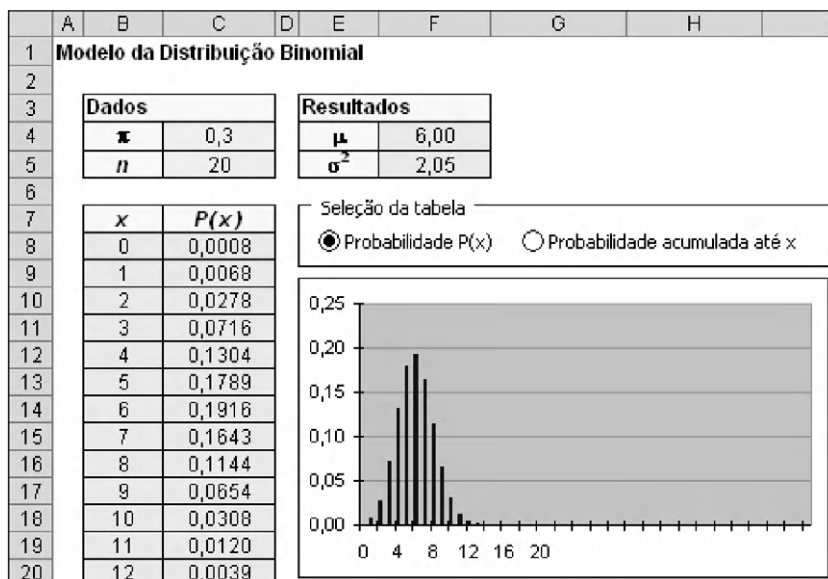


FIGURA 7.3
Distribuição binomial.

A Figura 7.3 mostra a planilha **Modelo Distribuição Binomial**, incluída na pasta **Capítulo 7**. Informando os valores da probabilidade de sucesso π na célula C4 e o número de experimentos ou tentativas na célula C5, limitadas a 50, a planilha calcula a média e a variância nas células F4 e F5, respectivamente, e a probabilidade escolhida na caixa de grupo a partir da célula C8 da tabela:

- **Probabilidade $P(x)$.** Fornecerá a probabilidade de ocorrerem x sucessos, de 0 até n , em n tentativas com a probabilidade de sucesso registrada em C4.
- **Probabilidade Acumulada até x .** Fornecerá a probabilidade acumulada de ocorrerem até x sucessos em n tentativas com a probabilidade de sucesso registrada em C4, como mostra a Figura 7.3.

EXEMPLO 7.7

Uma moeda é lançada dez vezes seguidas. Calcule a probabilidade de conseguir três caras.

Solução. A probabilidade de conseguir três caras é 0,3125 ou 31,25%, resultado obtido com a fórmula:

$$P(x=3) = \frac{10!}{3!(10-3)!} \times 0,50^3 (1-0,50)^{10-3}$$

$$P(x=3) = 10 \times 0,5^5 = 0,3125$$

⁴ A distribuição binomial costuma ser representada com o símbolo $B(n, \pi)$.

Esse resultado pode ser obtido com a função estatística DISTRBINOM do Excel registrando a fórmula =DISTRBINOM(3;5;0,5;FALSO) em uma célula vazia de qualquer planilha.

• **DISTRBINOM(núm_s; tentativas; probabilidade_s; cumulativo)**

A função estatística DISTRBINOM retorna a probabilidade ou a probabilidade acumulada do número de tentativas bem-sucedidas *núm_s*, conforme o valor do argumento cumulativo.

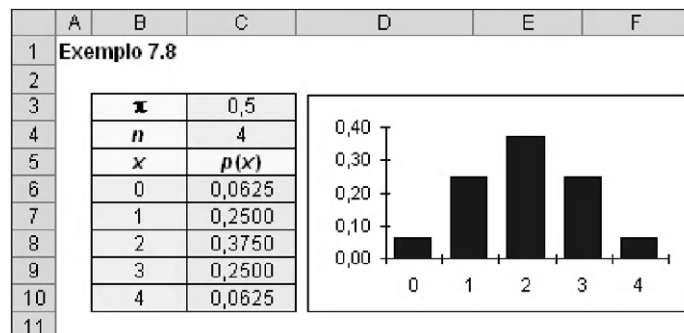
- Se o argumento *cumulativo* for FALSO, a função retornará a probabilidade do número de sucessos *núm_s* com *probabilidade_s* de sucesso para um número de *tentativas* independentes. Neste exemplo, a função retorna a probabilidade 0,3125 de conseguir três sucessos com probabilidade 0,5 em um experimento com cinco tentativas.
- Se o argumento *cumulativo* for VERDADEIRO, a função retornará a probabilidade acumulada do número máximo de sucessos *núm_s* com *probabilidade_s* de sucesso para um número de *tentativas* independentes.

EXEMPLO 7.8

Uma experiência com distribuição binomial foi repetida quatro vezes seguidas. Considerando a probabilidade de sucesso $\pi=0,50$:

- Calcule as probabilidades de todos os possíveis sucessos x .
- Construa o gráfico da distribuição de probabilidades.

Solução. Na planilha **Exemplo 7.8**, incluída na pasta **Capítulo 7**, foi construída a tabela de probabilidades com a fórmula $P(x) = \frac{4!}{x!(4-x)!} 0,50^x (1-0,50)^{4-x}$, e ao lado foi construído o histograma.



O primeiro resultado da tabela, probabilidade de $x=0$, pode ser obtido com a função estatística DISTRBINOM do Excel, registrando em uma célula vazia de qualquer planilha a fórmula =DISTRBINOM(0;4;0,5;FALSO), resultando no valor 0,0625. O segundo resultado, probabilidade de $x=1$, pode ser obtido com a fórmula =DISTRBINOM(1;4;0,5;FALSO), resultando no valor 0,25. Da mesma forma, os demais resultados da tabela. A partir da linha 12 da planilha **Exemplo 7.8**, foram registradas as fórmulas com a função DISTRBINOM utilizando os dados da tabela construída.

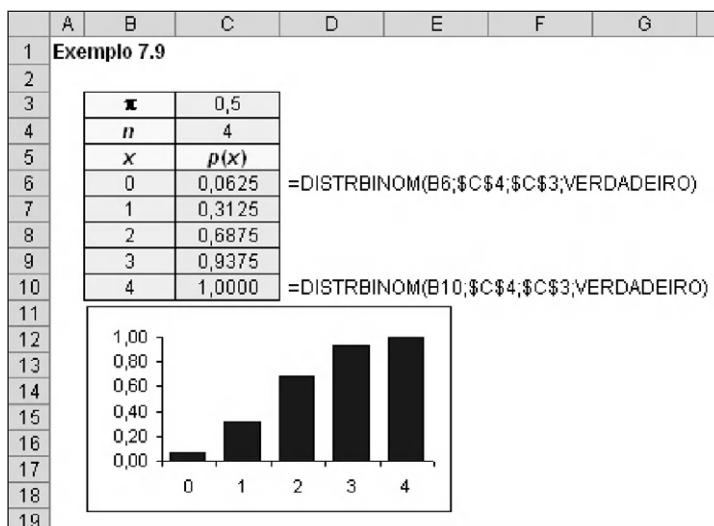
EXEMPLO 7.9

Continuando com o Exemplo 7.8, calcule:

- A probabilidade que x seja menor do que 2.
- A probabilidade que x seja menor ou igual a 2.

Solução. Na planilha **Exemplo 7.9**, incluída na pasta **Capítulo 7**, foi construída a tabela utilizando a função DISTRBINOM. Na célula C6, foi registrada a fórmula =DISTRBINOM(B6;\$C\$4;\$C\$3;VERDADEIRO) com o argumento *cumulativo* igual a VERDADEIRO. Depois essa fórmula foi copiada até a célula C10. A figura seguinte mostra também o gráfico de probabilidades acumuladas. Vejamos as respostas.

- A probabilidade que x seja menor do que 2 é a probabilidade acumulada até $x=1$, pois não deve ser incluída a probabilidade quando $x=2$. Esse resultado $P(x<2)=0,3125$ pode ser obtido registrando em qualquer célula vazia a fórmula =DISTRBINOM(1;4;0,5;VERDADEIRO).
- A probabilidade que x seja menor ou igual a 2 é a probabilidade acumulada até $x=2$. A tabela mostra que $P(x\leq 2)=0,6875$. Também esse resultado pode ser obtido registrando a fórmula =DISTRBINOM(2;4;0,5;VERDADEIRO).



Modelo probabilidade de sucesso

Os exemplos anteriores mostram tabelas de probabilidades e histogramas da distribuição binomial para a ocorrência de x sucessos com probabilidade π durante n experiências. Vejamos algumas conclusões:

- Em um experimento com distribuição binomial, o número de resultados é igual a $n+1$, pois $0 \leq x \leq n$.
- A probabilidade de sucesso π varia entre 0+ e próximo de 1, pois nos casos extremos o experimento não seria aleatório.

Para cada valor de probabilidade de sucesso π , há uma distribuição binomial de probabilidades diferente, mantendo os outros parâmetros inalterados. A planilha **Modelo probabilidade** da Figura 7.4 mostra essa característica. Com a barra de rolagem dessa planilha, pode-se acompanhar a variação da probabilidade de que ocorra um determinado sucesso e a probabilidade acumulada até esse definido sucesso em função da probabilidade de sucesso π variável no intervalo (0, 1) de um experimento com dez tentativas. Quanto à forma da distribuição de probabilidade, você verificará que, para $\pi=0,5$, a distribuição é sempre simétrica, independente do valor do número n de tentativas. Ao mesmo tempo, para valores de $\pi<0,5$, a distribuição de probabilidade apresentará inclinação positiva, para a direita, acentuando-se à medida que se aproxima de zero. Inversamente, para valores de $\pi>0,5$, a distribuição de probabilidade apresentará inclinação negativa, para a esquerda, acentuando-se à medida que se aproxima de um.

Média e variância da distribuição binomial

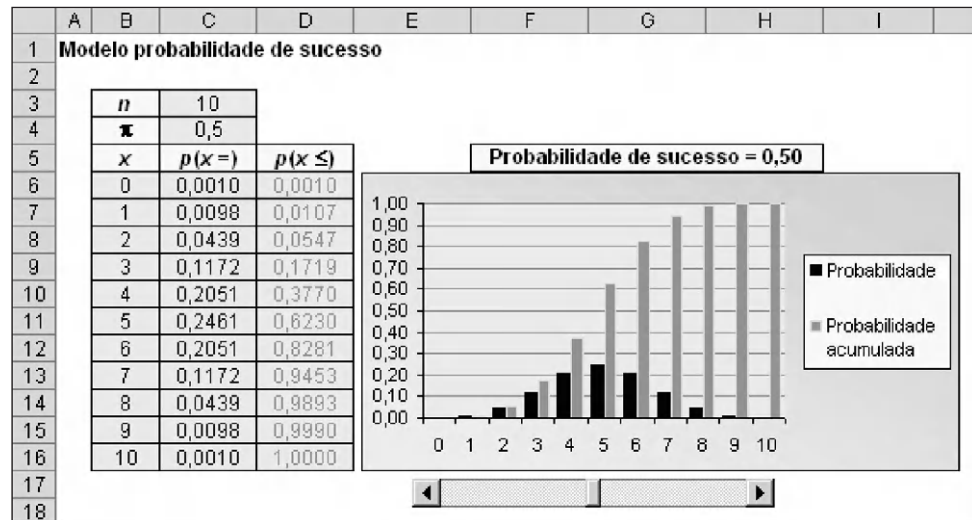
Como já foi visto, as distribuições de probabilidades têm associadas duas medidas estatísticas, a média e a variância, sem considerar o desvio padrão que é obtido da variância. Aplicando os conceitos de valor esperado nas distribuições discretas, substituindo a expressão $P(x)$ da distribuição binomial naquelas expressões, obteremos o valor esperado, a variância e o desvio padrão da distribuição binomial. Observe que esses resultados não dependem do número de sucessos x .

Parâmetros da distribuição binomial

A média, a variância e o desvio padrão são obtidos com:

$$\mu = n \times \pi \quad \sigma^2 = n \times \pi \times (1 - \pi) \quad e \quad \sigma = \sqrt{n \times \pi \times (1 - \pi)}$$

FIGURA 7.4 Modelo probabilidade de sucesso.



EXEMPLO 7.10

São realizadas dez experiências com probabilidade de sucesso $\pi=0,10$. Considerando que o experimento tem distribuição binomial, calcule a média e o desvio padrão.

Solução. Aplicando as fórmulas temos:

$$\mu = n \times \pi = 10 \times 0,1 = 1$$

$$\sigma = \sqrt{n \times \pi \times (1 - \pi)} = \sqrt{10 \times 0,10 \times (1 - 0,10)} = 0,9487$$

EXEMPLO 7.11

Você tem uma carteira com quinze ações. No pregão de ontem, 75% das ações na Bolsa de Valores caíram de preço. Supondo que as ações que perderam valor têm distribuição binomial:

- Quantas ações da sua carteira você espera que tenham caído de preço?
- Qual o desvio padrão das ações da carteira?

Solução. Como 75% das ações caíram de preço, o número de ações da carteira que devem ter sofrido queda de preço será $11,25=0,75 \times 15$. O desvio padrão é 1,67, obtido com a fórmula:

$$\sigma = \sqrt{n \times \pi \times (1 - \pi)} = \sqrt{15 \times 0,75 \times (1 - 0,75)} = 1,67.$$

EXEMPLO 7.12

Continuando com o Exemplo 7.11.

- Qual a probabilidade de que as quinze ações da carteira tenham caído?
- Qual a probabilidade que tenham sofrido redução de preço *exatamente* dez ações?
- Qual a probabilidade que treze ou mais ações tenham sofrido redução de preço?

Solução. Para calcular a probabilidade de que as quinze ações da carteira tenham sofrido redução de preço, aplicamos a expressão geral:

$$P(x = 15) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \binom{15}{15} 0,75^{15} (1 - 0,75)^{15-15}$$

$$P(x = 15) = \frac{15!}{15!(15-15)!} \times 0,75^{15} \times 0,25^0 = 0,0134$$

De forma equivalente, a probabilidade que tenham sofrido redução de preço *exatamente* dez ações é $P(x=10)=0,1651$, e a probabilidade que treze ou mais ações tenham sofrido queda de preço é obtida com $P(x \geq 13) = P(x=13) + P(x=14) + P(x=15) = 0,2361$.

EXEMPLO 7.13

A quantidade de vezes em que o número seis foi obtido no lançamento de um dado vinte vezes seguidas tem distribuição binomial. Calcule:

- A probabilidade de conseguir três seis.
- A probabilidade de conseguir mais de três seis.

Solução. Utilizando a função DISTRBINOM:

- Registrando a fórmula =DISTRBINOM(3;20;1/6;FALSO), temos a probabilidade $P(x=3)=0,2379$ de conseguir três seis em um experimento de 20 lançamentos de um dado.
- Registrando a fórmula =DISTRBINOM(3;20;1/6;VERDADEIRO), temos a probabilidade $P(x \leq 3)=0,5665$ de conseguir até três seis em um experimento de 20 lançamentos de um dado. Para obter $P(x > 3)$, aplicamos a regra do complemento $P(x > 3) = 1 - P(x \leq 3) = 0,4335$.

EXEMPLO 7.14

Uma urna contém 10 bolas, sendo 2 verdes e 8 brancas. Realizando 15 retiradas com reposição, calcule:

- A probabilidade de retirar cinco bolas verdes.
- A probabilidade de conseguir até cinco bolas verdes.
- A média e a variância desse experimento.

Solução. A probabilidade de sucesso de retirar uma bola branca é 0,20. Utilizando a função DISTRBINOM:

- Registrando a fórmula =DISTRBINOM(5;15;0,2;FALSO), temos a probabilidade $P(x=5)=0,1032$ de conseguir cinco bolas verdes em um experimento de 15 retiradas com reposição.
- Registrando a fórmula =DISTRBINOM(5;15;0,2;VERDADEIRO), temos a probabilidade $P(x \leq 5)=0,9389$ de conseguir até cinco bolas verdes em um experimento de 15 retiradas com reposição.
- A média é igual a 3, e a variância, igual a 2,4.

Os resultados anteriores, bem como a maioria dos resultados com a distribuição binomial, podem ser obtidos com a planilha **Cálculo Prob. Binomial**, incluída na pasta **Capítulo 7**, como mostra a figura seguinte.

	A	B	C	D	E	F
1	Cálculo de Probabilidades Binomiais					
2						
3		Dados			Resultados	
4		π	0,2		$P(x = 5)$	0,1032
5		n	15		$P(x \leq 5)$	0,9389
6		x	5		$P(x > 5)$	0,0611
7					μ	3,0
8					σ^2	2,40
9						

Antecipando alguns conceitos, deve-se registrar que se n for adequadamente grande e, para valores de probabilidade de sucesso que não sejam próximos de 0 nem de 1, o teorema central do limite⁵ permitirá aproximar a distribuição binomial utilizando a distribuição normal. A média e o desvio padrão serão obtidos com as fórmulas da distribuição binomial.

Tabela da distribuição binomial

Na planilha **Distribuição Binomial** da pasta **Tabelas** disponível na página do livro no site da Editora, você encontrará a Tabela da Distribuição Binomial. Escolhendo na caixa de grupo:

- *Probabilidade $P(x)$.* A tabela fornecerá a probabilidade de ocorrerem x sucessos em n tentativas com probabilidades de sucesso π definidas no intervalo C6:M6.
- *Probabilidade Acumulada até x .* A tabela fornecerá a probabilidade acumulada de ocorrerem até x sucessos em n tentativas, com as probabilidades de sucesso definidas no intervalo C6:M6, conforme apresentado na Figura 7.5.

A tabela foi limitada até 50 experiências, começando por 1, e, na Figura 7.5, pode-se verificar o resultado da primeira questão do Exemplo 7.14. As colunas do intervalo C:M fornecem as probabilidades desejadas para probabilidades determinadas no intervalo C6:M6. Na coluna O, denominada *Teste*, é possível calcular qualquer probabilidade para uma determinada probabilidade de sucesso, informada na célula O6 e o número de tentativas registradas na célula C4.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Tabela da Distribuição Binomial														
2															
3															
4	<div><div>n</div><div>15</div></div>		<div><div>Seleção da tabela</div><div><div><input checked="" type="radio"/> Probabilidade P(x)</div><div><input type="radio"/> Probabilidade acumulada até x</div></div></div>												
5															
6	<div><div>$x \backslash p$</div></div>	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	Teste		
7	0	1,0000	0,2059	0,0352	0,0047	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002		
8	1	0,0000	0,3432	0,1319	0,0305	0,0047	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000	0,0025		
9	2	0,0000	0,2669	0,2309	0,0916	0,0219	0,0032	0,0003	0,0000	0,0000	0,0000	0,0000	0,0130		
10	3	0,0000	0,1285	0,2501	0,1700	0,0634	0,0139	0,0016	0,0001	0,0000	0,0000	0,0000	0,0426		
11	4	0,0000	0,0428	0,1876	0,2186	0,1268	0,0417	0,0074	0,0006	0,0000	0,0000	0,0000	0,0963		
12	5	0,0000	0,0105	0,1032	0,2061	0,1859	0,0916	0,0245	0,0030	0,0001	0,0000	0,0000	0,1598		
13	6	0,0000	0,0019	0,0430	0,1472	0,2066	0,1527	0,0612	0,0116	0,0007	0,0000	0,0000	0,2010		
14	7	0,0000	0,0003	0,0138	0,0811	0,1771	0,1964	0,1181	0,0348	0,0035	0,0000	0,0000	0,1949		
15	8	0,0000	0,0000	0,0035	0,0348	0,1181	0,1964	0,1771	0,0811	0,0138	0,0003	0,0000	0,1470		

FIGURA 7.5 Amostra parcial da Tabela da Distribuição Binomial.

Outras funções do Excel

Com a função DISTRBINOM, pode ser calculada a probabilidade de um determinado número de sucessos x de um experimento binomial, ou a probabilidade acumulada até esse número de sucessos x . O Excel também dispõe de outras funções estatísticas, como mostrado a seguir utilizando o Exemplo 7.15.

EXEMPLO 7.15

Seja uma experiência com distribuição binomial com $n=4$ e probabilidade de sucesso $\pi=0,60$. Calcule a probabilidade de ter três sucessos e a probabilidade de ter de um até três sucessos, ambos os limites incluídos.

Solução. A probabilidade de ter três sucessos é $P(x=3)=0,3456$, valor obtido com a fórmula =DISTRBINOM(3;4;0,60;FALSO). Da mesma maneira, a probabilidade de ter de um até três sucessos, ambos os limites incluídos, e o resultado da soma $P(x=1)+P(x=2)+P(x=3)$. Dessa maneira, a probabilidade procurada é $P(1 \leq x \leq 3)=0,8448$. Esse valor pode ser obtido, também, com a função DISTRBINOM com a fórmula =DISTRBINOM(3;4;0,60;VERDADEIRO)-

DISTRBINOM(0;4;0,60;FALSO)

PROB(intervalo_x; intervalo_prob; limite_inferior; limite_superior)

A função estatística PROB⁶ retorna a probabilidade acumulada entre os argumentos *limite inferior* e o *limite superior*, ambos incluídos. O argumento *intervalo_x* de valores e o argumento *intervalo_prob* se referem à tabela de probabilidades $P(x)$ construída para esse experimento. Na planilha **Outras funções**, incluída na pasta **Capítulo 7**, é mostrado como utilizar a função PROB, como mostra a Figura 7.6 com os dados do Exemplo 7.15.

- No intervalo B3:C10 foram registrados os valores de π e n e calculadas as probabilidades para os cinco valores de x , de 0 a 4.
- No intervalo F4:F5, foram registrados os valores do limite inferior e o limite superior de x .
- Na célula F8, foi registrada =PROB(B6:B10;C6:C10;F4;F5), fórmula utilizada para calcular a probabilidade acumulada $P(1 \leq x \leq 3) = 0,8448$. Observe que a probabilidade acumulada $P(1 \leq x \leq 3)$ é obtida com a fórmula: $P(1 \leq X \leq 3) = P(X \leq 3) - P(X = 0) = 0,8704 - 0,0256 = 0,8448$.
- O mesmo resultado é obtido informando os dados em forma de matriz, registrando na célula F9 a fórmula:
PROB({0;1;2;3;4};{0,0256;0,1536;0,3456;0,3456;0,1296};E38;E39).
- Com a função DISTRBINOM, registrando na célula F10 a fórmula:
=DISTRBINOM(F5;C4;C3;VERDADEIRO)-
SE(F4=0;0;DISTRBINOM(F4-1;C4;C3;VERDADEIRO))

Observe que foi subtraído um do valor do argumento *núm_s* da segunda parcela da fórmula da função DISTRBINOM. No entanto, quando o limite inferior de x for zero, esse argumento será zero, pela função condicional SE.

	A	B	C	D	E	F	G	H
1	Função PROB							
2								
3		x	0,6		Dados			
4		n	4		Limite inferior	1		
5		x	$P(=x)$		Limite superior	3		
6		0	0,0256					
7		1	0,1536		Resultados			
8		2	0,3456		$P(1 \leq x \leq 3)$	0,8448		=PROB(B6:B10;C6:C10;F4;F5)
9		3	0,3456			0,8448		
10		4	0,1296		Com DISTRBINOM	0,8448		
11								

FIGURA 7.6 Utilizando a função estatística PROB.

CRIT.BINOM(tentativas; probabilidade_s; alfa)

A função estatística CRIT.BINOM⁷ retorna o menor número de sucessos para o qual a distribuição binomial acumulada é maior ou igual ao argumento *alfa*. Para valores exatos de probabilidade acumulada, a função estatística CRIT.BINOM é inversa da função estatística DISTRBINOM com o argumento *cumulativo* VERDADEIRO.

Aplicando a função CRIT.BINOM no Exemplo 7.15, se *alfa*=0,50, o número de sucessos menor ou igual a 0,50 é dois. A partir da coluna J da planilha **Outras funções**, foi construído o modelo para calcular esse resultado. Para verificar o resultado dessa função, ao lado, no intervalo N4:M9, foi construída a tabela de probabilidades acumuladas desse experimento.

⁶ Em inglês, a função PROB é *PROB*.

⁷ Em inglês, a função CRIT.BINOM é *CRITBINOM*.

FIGURA 7.7 Utilizando a função CRIT.BINOM.

	J	K	L	M	N	O
1	Função CRIT.BINOM					
2						
3	Dados					
4	x	0,6		x	P(<=x)	
5	n	4		0	0,0256	
6	alfa	0,50		1	0,1792	
7				2	0,5248	
8	Resultado			3	0,8704	
9	x	2		4	1,0000	
10						

Em outra aplicação, a função CRIT.BINOM determina o número máximo de peças defeituosas de um lote de produção sem rejeitar o lote inteiro.

Distribuição de Poisson

Depois da distribuição binomial, a distribuição de Poisson é a distribuição de probabilidade discreta mais utilizada, pois pode ser aplicada a muitos casos práticos nos quais interessa o número de vezes em que um determinado evento pode ocorrer durante um intervalo de tempo ou em um determinado ambiente físico, denominados sobre o nome de área de oportunidade.⁸ Por exemplo, o número de acidentes de carros por dia em uma grande cidade como São Paulo, o número de chamadas telefônicas por hora recebida na central telefônica durante o período normal de operação de uma empresa, o número de defeitos de soldagem em seis metros de tubo, o número de garrafas mal fechadas por trinta minutos na máquina de enchimento de cerveja, o número de comprimidos rejeitados por hora pela máquina de compressão devido ao peso fora de especificação etc.

Em um *processo de Poisson* podem ser observados eventos discretos em uma área de oportunidade de tal forma que, reduzindo suficientemente essa área de oportunidade:

- A probabilidade de observar apenas um sucesso no intervalo é estável.
- A probabilidade de observar mais de um sucesso no intervalo é zero.
- A ocorrência de um sucesso em qualquer intervalo é estatisticamente independente da ocorrência em qualquer outro intervalo.

A distribuição de Poisson é caracterizada apenas pelo parâmetro λ . Enquanto a variável aleatória do processo de Poisson X se refere ao número de sucessos por área de oportunidade, o parâmetro λ se refere ao valor esperado, ou média, do número de sucessos por área de oportunidade.

Probabilidade da Distribuição de Poisson

A probabilidade $P(x)$ de ocorrência de x conhecido λ é $P(x) = \frac{e^{-\lambda} \times \lambda^x}{x!}$.

λ é o número esperado de sucessos.

$x=0, 1, 2, \dots, \infty$ é o número de sucessos.

e constante aproximadamente igual a 2,7182....

A média e a variância são iguais a $\mu=\lambda$ e $\sigma^2=\lambda$.

⁸ Uma área de oportunidade pode ser um intervalo de tempo, espaço ou área na qual mais de uma ocorrência de um evento pode ocorrer.

EXEMPLO 7.16

As lâmpadas de iluminação da área de manufatura da montadora são substituídas em uma média de oito lâmpadas por dia. Se a distribuição de frequências das lâmpadas substituídas for do tipo *Poisson*:

- Qual a probabilidade de amanhã cinco lâmpadas precisarem ser substituídas?
- Qual a probabilidade de amanhã nenhuma lâmpada precisar ser substituída?
- Qual a probabilidade de amanhã no máximo cinco lâmpadas precisarem ser substituídas?

Solução. Dos dados, deduzimos que o número esperado de trocas diárias de lâmpadas é $\lambda=8$. A probabilidade de amanhã cinco lâmpadas precisarem ser substituídas é $P(x=5)=9,16\%$, resultado obtido com a fórmula:

$$P(x=5) = \frac{e^{-8} \times 8^5}{5!} = 0,091604$$

A probabilidade de amanhã nenhuma lâmpada precisar ser substituída é $P(x=0)=0,033\%$, resultado obtido com a fórmula:

$$P(x=0) = \frac{e^{-8} \times 8^0}{0!} = 0,000335$$

A probabilidade de amanhã no máximo (ou até) cinco lâmpadas precisarem ser substituídas é $P(x \leq 5) = P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4) + P(x=5) = 19,12\%$, resultado obtido com a fórmula:

$$P(x \leq 5) = \sum_{i=0}^5 \frac{e^{-8} \times 8^i}{i!} = 0,1912$$

Vimos que, para probabilidades de sucesso menores do que 0,50, a distribuição binomial tem inclinação para a direita, quanto mais inclinada, maior a chance de a probabilidade se aproximar de zero. Se a probabilidade de sucesso for muito pequena e o número de experiências grande, no limite, será obtida a *distribuição de Poisson*.

A Figura 7.8 mostra a planilha **Modelo Distribuição de Poisson** incluída na pasta **Capítulo 7**. Informando o número esperado de sucessos λ na célula C4, a planilha calcula a média, a variância e as probabilidades escolhidas na caixa de grupo a partir da célula C8:

- **Probabilidade $P(x)$.** Fornecerá a probabilidade de ocorrerem x sucessos, com o número esperado de sucessos registrado em C4.
- **Probabilidade Acumulada até x .** Fornecerá a probabilidade acumulada de ocorrerem até x sucessos com o número esperado de sucessos registrado em C4, como mostra a Figura 7.8.

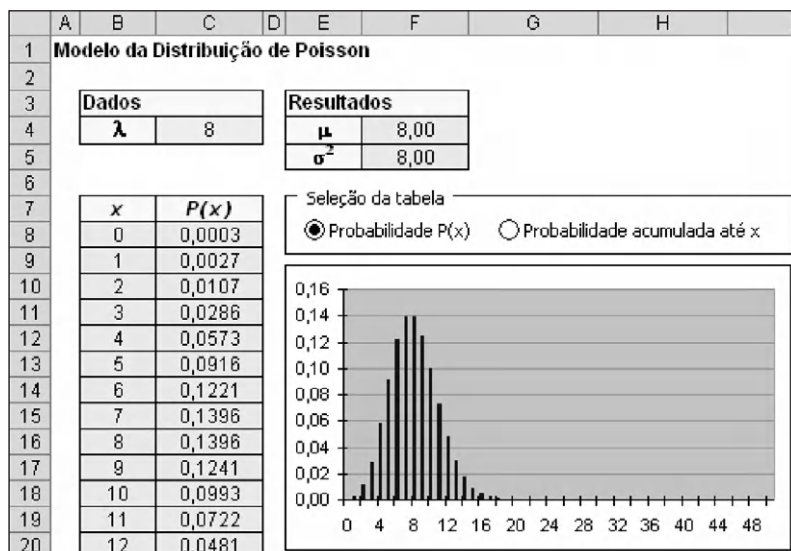


FIGURA 7.8
Distribuição de Poisson.

Tabela da distribuição de Poisson

A planilha **Distribuição de Poisson** da pasta **Tabelas** disponível na página do livro, no site da editora, contém a Tabela da Distribuição de Poisson. Escolhendo na caixa de grupo:

- **Probabilidade $P(x)$.** A tabela fornecerá a probabilidade de ocorrerem x sucessos, com o número esperado de sucessos λ registrado no intervalo C6:L6.
- **Probabilidade Acumulada até x .** A tabela fornecerá a probabilidade acumulada de ocorrerem até x sucessos com o número esperado de sucessos λ , registrado no intervalo C6:M6, como mostra a Figura 7.9.

Como λ pode assumir qualquer valor positivo e a tabela registra somente 10 valores diferentes de λ , você terá de registrar a unidade do valor esperado na célula C4, informação limitada ao intervalo (0, 25). As colunas do intervalo C:L fornecem as probabilidades desejadas para o número esperado de sucessos do intervalo C6:M6. Na coluna N, denominada *Teste*, você poderá calcular qualquer probabilidade para o número esperado de sucessos informado na célula N6.

FIGURA 7.9 Tabela da Distribuição de Poisson.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Tabela da Distribuição de Poisson													
2														
3														
4			Seleção da tabela											
5														
6														
7														
8														
9														
10														
11														
12														
13														
14														
15														

EXEMPLO 7.17

O número de reclamações de malas não recebidas no terminal nacional da maior companhia aérea é de cinco por dia. Se a distribuição de frequências das malas extraviadas é do tipo Poisson:

- Qual a probabilidade de que em qualquer dia sejam extraviadas exatamente duas malas?
- Qual a probabilidade de que em qualquer dia sejam extraviadas três ou menos malas?
- Qual a probabilidade de que em qualquer dia sejam extraviadas três ou mais malas?

Solução. Dos dados, deduzimos o número esperado de malas extraviadas por dia, $\lambda=5$. A probabilidade de que em qualquer dia sejam extraviadas exatamente duas malas é $P(x=2)=8,42\%$, resultado obtido com a fórmula:

$$P(x=2) = \frac{e^{-5} \times 5^2}{2!} = 0,0842$$

Esse resultado pode ser obtido na Tabela da Distribuição de Poisson, procedendo como foi explicado. Também pode ser obtido com a função estatística POISSON do Excel.

• POISSON(x ; *média*; *cumulativo*)

A função estatística POISSON retorna dois tipos de probabilidades, conforme o valor do argumento *cumulativo*. Se o argumento *cumulativo* for FALSO, a função retornará a probabilidade do número de sucessos x , considerando o argumento *média* esperada de sucessos. O resultado dessa questão $P(x=2)=8,42\%$ é obtido registrando a fórmula =POISSON(2;5;FALSO). Se o argumento *cumulativo* for VERDADEIRO, a função retornará a probabilidade acumulada até x , considerando o argumento *média*.

A probabilidade de que em qualquer dia sejam extraviadas três ou menos malas é $P(x \leq 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3) = 26,50\%$. Esse resultado pode ser obtido calculando as parcelas

dessa soma ou utilizando a função POISSON, registrando a fórmula =POISSON(3;5;VERDADEIRO). Esse resultado também pode ser obtido com a planilha **Cálculo Prob. Poisson**, como mostrado a seguir.

	A	B	C	D	E	F
1	Cálculo Probabilidade de Poisson					
2						
3		Dados			Resultados	
4		λ	5		P(x=3)	0,1404
5		x	3		P(x≤3)	0,2650
6					μ	5,00
7					σ²	5,00
8						

A probabilidade de que em qualquer dia sejam extraviadas três ou mais malas é $P(x \geq 3) = P(x=3) + P(x=4) + P(x=5) = 87,53\%$. Esse resultado pode ser obtido calculando as parcelas dessa soma ou utilizando a função POISSON, tendo presente que essa função retorna a probabilidade acumulada desde $x=0$. Para resolver este exemplo, teremos de utilizar o complemento, que algebricamente é dado pela fórmula $P(x \geq 3) = 1 - (P(x=0) + P(x=1) + P(x=2))$. A probabilidade procurada será obtida com a =1-POISSON(2;5;VERDADEIRO).

Outras distribuições discretas

O Excel também dispõe das funções da *distribuição binomial negativa* e a *distribuição hipergeométrica*. Para apresentar a distribuição binomial negativa, faremos uma análise do que foi apresentado na distribuição binomial. O ponto de partida é o processo de Bernoulli, definido como o experimento aleatório cujo espaço amostral tem apenas dois possíveis resultados mutuamente excludentes denominados sucesso e falha, sendo π a probabilidade de sucesso. Se o processo Bernoulli for repetido n vezes, considerando que as experiências são independentes, então a variável aleatória X que define o número de sucessos do experimento terá distribuição binomial. Observe que, na distribuição binomial, o número de experimentos n é definido antecipadamente.

Em vez de repetir o experimento um número determinado de vezes, pode-se estabelecer que o experimento seja repetido até conseguir o primeiro resultado *sucesso*. Nesse caso, a variável aleatória X que define o número de experimentos necessários até conseguir o primeiro resultado sucesso tem uma distribuição geométrica. Ampliando as premissas da distribuição geométrica, em vez de repetir o experimento até conseguir o primeiro resultado *sucesso*, a distribuição binomial negativa⁹ permite determinar a probabilidade de que será necessário realizar exatamente n experimentos para obter x resultados de sucesso com probabilidade π .

$$P(x) = \binom{n-1}{x-1} \pi^x (1-\pi)^{n-x}$$

Com essa expressão, pode-se calcular a probabilidade de que o x -ésimo resultado de sucesso com probabilidade π será obtido até n experimentos, tendo presente que, no conjunto de n experimentos independentes, há somente dois resultados possíveis e a amostragem é realizada sem reposição.¹⁰ A média e a variância de uma distribuição binomial negativa são obtidas com as seguintes expressões:

$$\mu = \frac{x}{\pi} \quad \sigma^2 = r \frac{(1-\pi)}{\pi^2}$$

⁹ Também conhecida como Distribuição de Pascal.

¹⁰ Sheskin D.J. – *Handbook of Parametric and Nonparametric Statistical Procedures* – Editora Chapman & Hall/CRC – 2ª edição, 2000.

EXEMPLO 7.18

A probabilidade de que uma copiadora consiga reproduzir uma cópia aceitável é 25%. Qual a probabilidade de que a quinta cópia aceitável seja reproduzida depois de doze reproduções?

Solução. A probabilidade de que exatamente doze reproduções serão necessárias antes que sejam conseguidas cinco cópias aceitáveis é 4,30%, resultado obtido com a fórmula anterior.

$$P(x=5) = \binom{12-1}{5-1} 0,25^5 (1-0,25)^{12-5} = 0,0430$$

Esse resultado também pode ser obtido registrando numa planilha a fórmula:
`=FATORIAL(11)/(FATORIAL(4)*FATORIAL(11-4))*0,25^5*0,75^7`

A partir da linha 12 da planilha **Outras funções**, incluída na pasta **Capítulo 7**, foram realizados os mesmos cálculos. Na célula O15, com a fórmula anterior, porém utilizando os dados do intervalo L15:L17 e, na célula O16, utilizando a função COMBIN já apresentada. Na célula O17, com a função estatística DIST.BIN.NEG do Excel.

• **DIST.BIN.NEG(núm_f; núm_s; probabilidade_s)**

A função estatística DIST.BIN.NEG¹¹ retorna a probabilidade de que o *num_s* resultado de sucesso com probabilidade *probabilidade_s* será obtido depois de ocorrer o número de falhas *núm_f*. Observe que se o *número de falhas* for igual a zero, a função DIST.BIN.NEG dá o mesmo resultado da função BINOMDIST, considerando que o número de experimentos é igual ao número de sucessos e o argumento *cumulativo* FALSO, por exemplo, DIST.BIN.NEG(0;5;0,25)=DISTRBINOM(5;5;0,60;FALSO).

	J	K	L	M	N	O
12	Função DIST.BIN.NEG					
13						
14	Dados			Resultados		
15	π		0,25	$P(x = 5)$		0,0010
16	n		5	$P(x = 5)$		0,0010
17	x		5	$P(x = 5)$		0,0010
18						

A *distribuição hipergeométrica* não é uma distribuição derivada da distribuição binomial, pois os experimentos são dependentes. Em uma população composta de N objetos que podem ser classificados em duas categorias, C_1 e C_2 , de forma que na população há N_1 em C_1 e N_2 em C_2 . Desejamos retirar uma amostra sem reposição de n objetos dessa população, selecionando x objetos de C_1 e $(n-x)$ objetos de C_2 . A probabilidade de selecionar exatamente x objetos requeridos de C_1 e $(n-x)$ de C_2 é dada pela fórmula:

$$p(x) = \frac{\binom{N_1}{x} \binom{N_2}{n-x}}{\binom{N}{n}}$$

A média e a variância de uma *distribuição hipergeométrica* são obtidas com as seguintes expressões:

$$\mu = n \times \frac{N_1}{N} \quad \sigma^2 = n \times \left(\frac{N_1}{N} \right) \times \left(1 - \frac{N_1}{N} \right) \times \left(\frac{N-n}{N-1} \right)$$

¹¹ Em inglês, a função DIST.BIN.NEG é NEGBINOMDIST.

EXEMPLO 7.19

Qual a probabilidade de selecionar dois meninos e uma menina de uma turma de nove estudantes compostos de cinco meninos e quatro meninas?

Solução. A probabilidade de selecionar dois meninos e uma menina de uma turma de nove estudantes compostos de cinco meninos e quatro meninas é 47,62%, resultado obtido com a fórmula anterior.

$$P(x=2) = \frac{\binom{5}{2} \binom{4}{3-2}}{\binom{9}{3}} = 0,4762$$

A partir da linha 20 da planilha **Outras funções**, incluída na pasta **Capítulo 7**, foram realizados os mesmos cálculos. Na célula O23, utilizando a função COMBIN, já apresentada. Na célula O24, a função estatística DIST.HIPERGEOM do Excel.

• **DIST.HIPERGEOM(exemplo_s; exemplo_núm; população_s; núm_população)**

A função DIST.HIPERGEOM¹² retorna a probabilidade de acontecer o número de sucessos do argumento *exemplo_s*, conhecido o tamanho da amostra do argumento *exemplo_núm*, o número de sucessos do argumento *população_s* e o tamanho da população *núm_população*.

	J	K	L	M	N	O
1	Função CRIT.BINOM					
2						
3	Dados					
4	x		0,6	x		P(<=x)
5	n		4	0		0,0256
6	alfa		0,50	1		0,1792
7				2		0,5248
8	Resultado			3		0,8704
9	x		2	4		1,0000
10						

Problemas

Problema 1

A tabela a seguir registra a variável aleatória discreta Y. Calcule a média, a variância e o desvio padrão dessa variável.

Y	45	56	82	122
p(y)	15%	23%	48%	14%

R: $E[Y] = 76,07$ $\sigma_Y = 23,44$

Problema 2

O retorno monetário para os próximos doze meses de uma ação foi estimado pela variável aleatória X registrada na seguinte tabela. Qual o valor esperado e o desvio padrão?

X	100	125	145	170	200
p(x)	10%	20%	40%	15%	15%

R: $E[X] = 148,50$ $\sigma_X = 28,60$

¹² Em inglês, a função DIST.HIPERGEOM é HYPGEOMDIST.

Problema 3

A estimativa dos preços de uma ação para os próximos doze meses é a variável aleatória *Preço*, registrada na tabela. Calcule o valor esperado e o desvio padrão.

Preço	10	14	19	24	30
p(Preço)	10%	25%	35%	20%	10%

R: $E[\text{Preço}] = \$18,95$; $\sigma_{\text{Preço}} = \$5,61$

Problema 4

O lucro líquido estimado (em milhões) da empresa para o próximo ano e suas respectivas probabilidades considerando quatro cenários estão registrados na tabela seguinte. Calcule o valor esperado e o desvio padrão.

Cenários	Lucro líquido	Probabilidade
Excelente	10	20%
Bom	5	40%
Sofrível	1	25%
Ruim	-4	15%

R: $E[\text{Lucro}] = \$3,65$; $\sigma_{\text{Lucro}} = \$4,40$

Problema 5

O seguro de vida para pessoas com menos de 65 anos é \$100.000, devendo-se pagar \$1.850 por ano. Se a probabilidade de uma pessoa com menos de 65 anos morrer no próximo ano for 1,55%, qual a expectativa do lucro anual da seguradora?

R: $E[\text{Lucro}] = \$300,00$

Problema 6

Os dados históricos das vendas de um televisor importado mostram que, durante o período de garantia de um ano, 80% dos televisores não apresentaram nenhum problema, 15% requereram algum conserto e regulagem, e os demais 5% deveram ser substituídos. O lucro nos três casos foi, respectivamente, \$85, \$20 e -\$35 (considerando o valor da venda como sucata). Calcule o *lucro esperado* na venda de 100 unidades.

R: $\text{Lucro esperado} = 100 \times \$69,25 = \$6.925,00$

Problema 7

No lançamento de uma moeda três vezes seguidas, estamos interessados em conhecer a probabilidade de obter:

- Três caras.
- Menos do que duas caras.

R: a) $P(x=3) = 12,50\%$ b) $P(x < 2) = 50\%$

Problema 8

A companhia de aviação afirma que 95% dos seus voos chegam no horário. Se for extraída uma amostra de dez voos dos registros dos últimos três meses, calcule a probabilidade de:

- Pelo menos oito voos chegam no horário.
- Entre sete e nove voos chegam no horário.

R: a) 98,85% b) 40,02%

Problema 9

No transporte de carros novos entre o pátio da montadora e a concessionária, 3% dos carros transportados sofrem alguma avaria na sua pintura. Se uma concessionária recebe 50 carros, calcule a probabilidade de:

- Nenhum dos carros transportados sofrer avaria na pintura.
- Dois ou mais carros sofrerem avaria na pintura.

R: a) 21,81% b) 44,47%

Problema 10

Em um experimento com distribuição binomial em que são realizadas 120 experiências com probabilidade de sucesso 0,45, qual a média e o desvio padrão?

R: $\mu=54$ e $\sigma=5,45$

Problema 11

Suponha que os registros históricos mostram que 30% de uma população é imune a uma determinada doença de inverno. Se uma amostra probabilística de dez pessoas é retirada dessa população, qual a probabilidade de ter exatamente quatro pessoas imunes?

R: $P(x=4)=20,01\%$

Problema 12

Continuando com o problema 11.

- Qual a probabilidade de ter menos de três pessoas imunes a essa doença?
- Qual a probabilidade de ter cinco ou mais pessoas imunes a essa doença?

R: a) $P(x<3)=38,28\%$ b) $P(x\geq 5)=4,73\%$

Problema 13

A média diária de ajustes dos instrumentos de controle de processo da planta química é de seis instrumentos por dia. Considerando que a distribuição de frequências dos ajustes dos instrumentos é do tipo Poisson, qual a probabilidade de amanhã ter de ajustar:

- Exatamente quatro instrumentos?
- Menos de cinco instrumentos?
- Cinco ou mais instrumentos?
- Nenhum instrumento?

R: $P(x=4)=13,4\%$; $P(x<5)=P(x\leq 4)=28,5\%$; $P(x\geq 5)=71,5\%$; $P(x=0)=0,25$

Problema 14

O gerente do banco afirma que, em média, sua agência tem de administrar a devolução de oito cheques por dia por falta de fundos. Considerando que a distribuição do número de cheques devolvidos é do tipo Poisson, qual a probabilidade de amanhã ter de devolver:

- Menos de oito cheques?
- Exatamente oito cheques?
- Oito ou mais cheques?

R: $P(x<8)=45,30\%$; $P(x=8)=13,96\%$; $P(x\geq 8)=54,70\%$;

Problema 15

O erro de digitação cometido pelos caixas é 0,35 por hora. Qual a probabilidade que um caixa cometa dois erros numa hora?

R: $P(x=2)=4,32\%$

Apêndice 1

Outra fórmula da variância

Para mostrar como utilizar outra fórmula de cálculo da variância com o conceito de valor esperado $\sigma_X^2 = E[X^2] - E[X]^2$, novamente foi resolvido o Exemplo 7.5, utilizando essa fórmula na planilha Apêndice 1 do Capítulo 7, Figura 7.10.

- Na segunda e terceira colunas da tabela foram registrados os dados da variável aleatória X .
- Na quarta coluna, foram calculadas e registradas as quatro parcelas do valor esperado da variável aleatória e, na quinta e última coluna, as parcelas do valor esperado do quadrado da variável aleatória X .
- No intervalo C10:C11 foram calculadas as parcelas da fórmula da variância.

FIGURA 7.10
Resolução do Exemplo 7.5 com outra fórmula da variância.

	A	B	C	D	E	F	G
1	Exemplo 7.5						
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

Cenário	X	$p(x)$	$x.p(x)$	$x^2.p(x)$
Ruim	-10%	10%	-0,0100	0,001000
Regular	0%	20%	0,0000	0,000000
Bom	12%	40%	0,0480	0,005760
Excelente	25%	30%	0,0750	0,018750

Resultados	
$E[X]^2$	0,0128
$E[X^2]$	0,0255
Variância	0,0127
Desvio Padrão	0,1129

Resultados diretos	
$E[X]^2$	0,0128
$E[X^2]$	0,0255
Variância	0,0127
Desvio Padrão	0,1129

$\{=SOMA(C4:C7*D4:D7)^2\}$
$\{=SOMA((C4:C7)^2*D4:D7)\}$
$\{=SOMA((C4:C7)^2*D4:D7)-SOMA(C4:C7*D4:D7)^2\}$
$\{=RAIZ(SOMA((C4:C7)^2*D4:D7)-SOMA(C4:C7*D4:D7)^2)\}$

Na célula C12 foi calculada a variância de X com a fórmula:

$$\sigma_X^2 = E[X^2] - E[X]^2$$

$$\sigma_X^2 = 0,0255 - (0,0128)^2 - 0,0127$$

O desvio padrão de X foi calculado na célula C13, como resultado da raiz quadrada positiva da variância. No intervalo C16:C19, foram obtidos os mesmos resultados do intervalo C10:C13, porém utilizando fórmulas matriciais, que têm a vantagem de retornar o resultado utilizando apenas os dados da variável aleatória registrados no intervalo C3:D7, sem necessidade de construir as últimas duas colunas da tabela.

Ainda é possível obter o resultado da variância, ou do desvio padrão, com uma única fórmula registrada em uma célula vazia da planilha Excel e sem necessidade de construir a tabela com os dados e os

resultados intermediários. Por exemplo, a fórmula seguinte registrada na célula C22 da planilha **Apêndice 1** retorna o resultado da variância sem utilizar nenhum dado nem resultado intermediário registrado nessa planilha.

$$= \text{SOMA}(\{-0,1;0;0,12;0,25\}^2 * \{0,1;0,2;0,4;0,3\}) - \text{SOMA}(\{-0,1;0;0,12;0,25\} * \{0,1;0,2;0,4;0,3\})^2$$

Apêndice 2

Covariância como valor esperado

Na tabela da Figura 7.11, estão registradas as séries de valores da variável X e da variável Y cujas médias são $\mu_Y=3$ e $\mu_X=10,8$.

Y	2	4	3	4	2
X	10	12	10	10	12

FIGURA 7.11 Amostras X e Y .

Para calcular a variância, utilizamos a seguinte fórmula conhecida:

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X) \times (Y_i - \mu_Y)$$

Substituindo os dados das séries X e Y , temos:

$$\sigma_{XY} = \frac{1}{5} \times \left((10 - 10,8)(2 - 3) + (12 - 10,8)(4 - 3) + (10 - 10,8)(3 - 3) + (10 - 10,8)(2 - 3) + (12 - 10,8)(4 - 3) \right)$$

Realizando as operações indicadas, temos o resultado da covariância:

$$\begin{aligned} \sigma_{XY} &= (10 - 10,8)(2 - 3) \times 0,2 + (12 - 10,8)(4 - 3) \times 0,2 + \\ &\quad + (10 - 10,8)(3 - 3) \times 0,2 + (10 - 10,8)(2 - 3) \times 0,2 + (12 - 10,8)(4 - 3) \times 0,2 \\ \sigma_{XY} &= 0 \end{aligned}$$

Embora a covariância seja igual a zero, não podemos afirmar que as variáveis X e Y sejam independentes, como foi apresentado no Capítulo 6. Entretanto, esse não é o objetivo deste tema. As variáveis X e Y têm valores repetidos, e elas podem se apresentar como uma tabela de probabilidades conjuntas, também conhecida como *tabela de contingências*, como mostra a tabela da esquerda da Figura 7.12, com os valores das variáveis aleatórias X e Y classificados. A tabela da direita da Figura 7.12 mostra os mesmos resultados da tabela da esquerda, porém com referência ao número total de pares de valores, nesse caso, cinco. Ou, de outra maneira, os valores do miolo da tabela da direita são as probabilidades conjuntas.

	X	
Y	10	12
2	1	1
3	1	0
4	1	1

	X	
Y	10	12
2	0,20	0,20
3	0,20	0
4	0,20	0,20

FIGURA 7.12 Tabela de contingências das variáveis X e Y.

Observe que as probabilidades conjuntas registradas na tabela da direita são os coeficientes dos produtos dos desvios do cálculo da covariância. Isso mostra que, na fórmula da covariância, está incluída a probabilidade conjunta de cada par de valores, característica representada pela seguinte fórmula:

$$\sigma_{XY} = \sum_{i=1}^r \sum_{j=1}^s (x_i - E[X]) \times (y_j - E[Y]) \times p(x_i, y_j)$$

Nessa expressão, os limites superiores r e s dos somatórios definem a quantidade de valores das variáveis aleatórias X e Y , respectivamente.

- Se a covariância for calculada diretamente com os pares de valores, teremos n pares e, consequentemente, n parcelas a serem somadas.
- Se a covariância for calculada com a tabela de contingências, a quantidade de valores de cada uma das variáveis poderá ter, no máximo, n valores, que corresponde ao caso em que todos os valores de cada variável são diferentes. Nesse caso, o tamanho da tabela será $n \times n$ com n células iguais a um e $(n^2 - n)$ células vazias ou iguais a zero.
- Se as variáveis tiverem valores repetidos, o número de colunas (e linhas) da tabela será menor do que n , daí que, na fórmula da covariância, utilizamos os valores r e s como limites das somas das parcelas.

Resumindo, sejam as variáveis aleatórias $X = x_1, x_2, \dots, x_r$ e $Y = y_1, y_2, \dots, y_s$ com valores esperados $E[X]$ e $E[Y]$ e probabilidades conjuntas $p(x_1, y_1), \dots, p(x_r, y_s)$. A covariância das variáveis X e Y é definida como:

$$\sigma_{XY} = \sum_{i=1}^n (x_i - E[X]) \times (y_i - E[Y]) \times p(x_i, y_i)$$

Outra forma da fórmula da covariância

Uma representação mais formal da covariância expressada como valor esperado é a da expressão:

$$\sigma_{XY} = E[(X - E[X]) \times (Y - E[Y])]$$

Desta última expressão, demonstra-se que:

$$\sigma_{XY} = E[X \times Y] - E[X] \times E[Y]$$

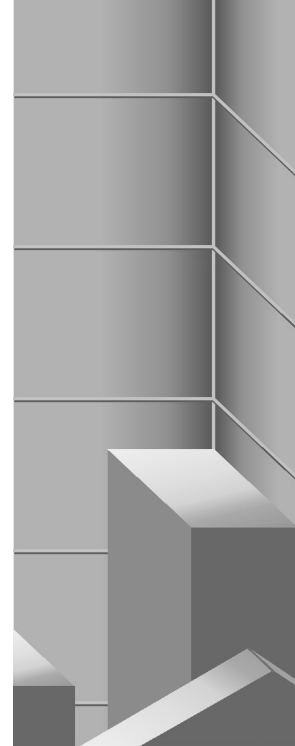
Como verificação, quando as duas variáveis são iguais, por exemplo X , obtém-se a expressão da variância.

$$\sigma_{XY} = \text{Var}(X) = E[X \times X] - E[X] \times E[X]$$

$$\sigma_{XY} = \text{Var}(X) = E[X^2] - E[X]^2$$

Capítulo 8

DISTRIBUIÇÕES CONTÍNUAS



No Capítulo 7, foi mostrado que, para definir uma VA de forma completa, é necessário especificar os valores dos eventos elementares do espaço amostral do experimento aleatório e suas probabilidades associadas. Se os valores da variável aleatória podem assumir qualquer valor do conjunto dos números reais, a variável aleatória é denominada *VA contínua*. Por exemplo, os preços dos carros usados, os salários dos empregados de uma determinada categoria, as rentabilidades mensais das ações etc. Lembre que não é possível registrar todos os valores de uma VA contínua em uma lista, tabela ou histograma, pois o número de valores possíveis é muito grande. Dessa maneira, a distribuição de probabilidades de uma variável aleatória contínua é definida por uma curva contínua e não por pontos discretos de uma tabela.

Seja a variável aleatória contínua X , onde definimos a função $f(x)$ denominada *função densidade de probabilidade*, com as seguintes propriedades:

- A probabilidade da variável aleatória X é sempre definida em um intervalo de valores dessa variável X , por exemplo, (x_1, x_2) .
- A probabilidade da variável aleatória X é medida pela área sob a curva da função densidade $f(x)$ em um determinado intervalo.
- A área total sob a curva $f(x)$ é igual a um ou 100%. Dessa maneira, o valor $f(x)$ da função densidade não mede a probabilidade do valor x da variável aleatória X .

Variável aleatória contínua

Para a variável aleatória contínua X que assume valores do conjunto dos números reais, há uma função matemática $f(x)$ com as seguintes premissas:

- A função densidade de probabilidade $f(x)$ é sempre positiva, $f(x) \geq 0$ para todo x pertencente a X .
- A área sob a função $f(x)$ entre os limites menos infinito e mais infinito da variável aleatória contínua X é igual a um ou 100%, $\int_{-\infty}^{+\infty} f(x) dx = 1$.
- A probabilidade da VA contínua X dentro do intervalo (a, b) com ambos os limites incluídos é medida pela área definida pela função $f(x)$ entre os limites a e b , $P(a \leq x \leq b) = \int_a^b f(x) dx$. Portanto:

- A probabilidade de uma VA contínua é sempre calculada dentro de um intervalo de valores, por exemplo, a e b .
- Um ponto $f(x)$ da função densidade não é a probabilidade do valor x da variável aleatória X , pois, por exemplo, o ponto $f(x=a)$ da função densidade é zero.¹

Como devem ser representados os limites do cálculo da probabilidade de uma variável aleatória contínua dentro do intervalo (a, b) , $P(a \leq X \leq b)$ ou $P(a < X < b)$? As duas representações podem ser utilizadas, incluindo a representação com limites mistos, por exemplo: $P(a \leq X < b)$ ou $P(a < X \leq b)$! Neste livro, utilizaremos a representação $P(a \leq X \leq b)$.

Valor esperado e variância da variável aleatória contínua

Para variáveis discretas, no Capítulo 7 foi definido o valor esperado da variável aleatória X com a fórmula $E[X] = \sum_{i=1}^n x_i p(x_i)$. Essa fórmula mostra que o valor esperado de X é a soma dos produtos dos valores da VA pelas respectivas probabilidades associadas. Como num intervalo da VA contínua X há uma infinidade de valores, no lugar de somar, deve-se integrar o produto $x \cdot f(x)$ entre os limites do intervalo definido.

Neste capítulo, serão apresentadas distribuições de variáveis contínuas, começando pela distribuição uniforme.

Seja a VA contínua X com função densidade de probabilidade $f(x)$:

- Valor esperado de X : $\mu_X = \int_{-\infty}^{+\infty} x f(x) dx$
- Variância de X : $\sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x) dx$ e
- Desvio padrão de X : $\sigma_X = +\sqrt{\sigma_X^2}$

Distribuição uniforme

Uma distribuição de variável aleatória contínua é a *distribuição uniforme* cuja função densidade de probabilidade é constante dentro de um intervalo de valores da variável aleatória X . Dessa maneira, cada um dos possíveis valores que X com distribuição uniforme pode assumir tem a mesma probabilidade de ocorrer.

A variável aleatória X tem *distribuição uniforme* de probabilidades no intervalo (a, b) se a função densidade $f(x)$ for $f(x) = \frac{1}{b-a}$, com as seguintes condições $b \geq a$ e $a \leq x \leq b$.

A representação gráfica da distribuição uniforme é um retângulo com base definida pelos valores a e b , que estabelecem os limites de valores possíveis da variável aleatória X , Figura 8.1.

¹ Pela terceira propriedade: $P(a \leq X \leq a) = \int_a^a f(x) dx = 0$.

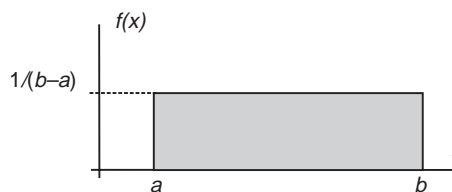


FIGURA 8.1
Distribuição uniforme
de probabilidades.

Da definição da distribuição uniforme, podemos deduzir:

- A área do retângulo é igual a um, pois a base é $(b-a)$ e a altura $1/(b-a)$.
- A probabilidade da variável aleatória X dentro do intervalo (a,b) é igual a 1 ou 100%.
- A probabilidade de $P(x)$ dentro do intervalo (c,d) , ambos os valores dentro do intervalo (a,b) , é

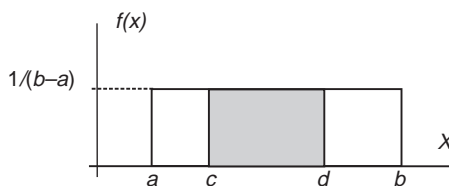


FIGURA 8.2 Relação
de áreas da distribuição
uniforme.

$$P(c \leq X \leq d) = \frac{d-c}{b-a}, \text{ como mostra a Figura 8.2.}$$

A média e a variância da VA com distribuição uniforme contínua dentro do intervalo (a,b) são obtidas utilizando as definições e realizando as integrações necessárias para o intervalo $(a \leq X \leq b)$.

A *média* μ_X e a *variância* σ_X^2 da variável aleatória X com *distribuição uniforme* de probabilidades no intervalo (a,b) são:

- *Média:* $\mu_X = \frac{a+b}{2}$
- *Variância:* $\sigma_X^2 = \frac{(b-a)^2}{12}$

A Figura 8.3 mostra a planilha **Modelo Distribuição Uniforme** incluída na pasta **Capítulo 8**. Informando no intervalo da planilha C4:C5, respectivamente, o valor mínimo e o valor máximo da distribuição, o gráfico da planilha mostra o contorno da distribuição, linha azul clara, e, no intervalo F4:F6, são registrados os valores da média, da variância e do desvio padrão dessa distribuição.

Depois, informando o intervalo (c,d) no intervalo da planilha C8:C9, o gráfico da planilha mostrará o contorno da área de probabilidade calculada e a célula F9 retornará a probabilidade desejada, como mostra a Figura 8.3.

EXEMPLO 8.1

Calcule a média e o desvio padrão da variável aleatória X com distribuição uniforme no intervalo $(100, 200)$.

Solução. A média da variável aleatória contínua X é 150, obtida com a fórmula:

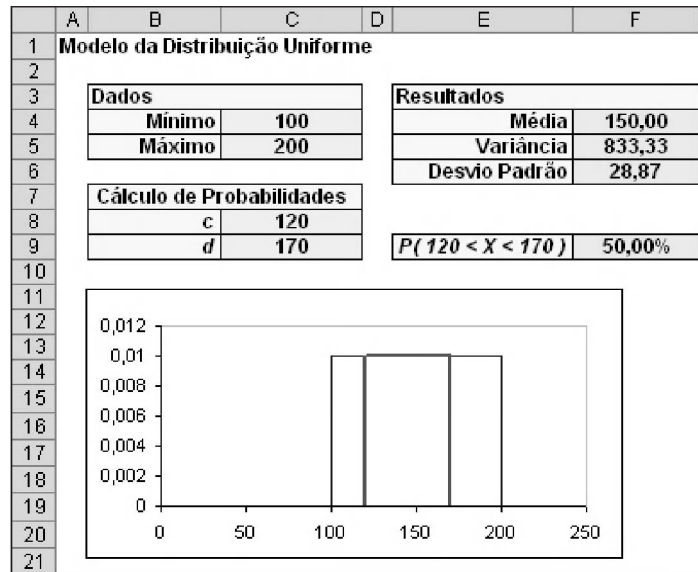
$$\mu_X = \frac{a+b}{2} = \frac{100+200}{2} = 150. \text{ Da mesma maneira, a variância é } 833,33 \text{ foi obtida com a fórmula:}$$

$$\sigma_X^2 = \frac{(b-a)^2}{12} = \frac{(200-100)^2}{12} = 833,33. \text{ O desvio padrão é igual a: } \sigma_X = \sqrt{833,33} = 28,87. \text{ Todos esses resulta-}$$

dos podem ser visualizados na planilha mostrada na Figura 8.3.

FIGURA 8.3

Distribuição uniforme contínua.

**EXEMPLO 8.2**

Qual a probabilidade de um valor de X pertencer ao intervalo $(120, 170)$?

Solução. A probabilidade de um valor da variável X se encontrar entre 120 e 170 é $P(120 \leq X \leq 170) = 0,50$, resultado obtido com:

$$P(120 \leq X \leq 170) = \frac{170 - 120}{200 - 100} = 0,50$$

Distribuição normal

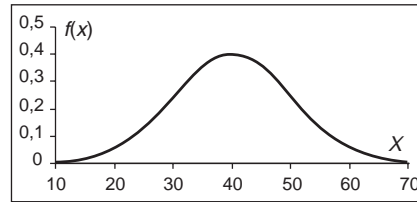
A *distribuição normal* é uma das distribuições fundamentais da moderna teoria estatística. A vantagem da distribuição normal reside na facilidade de defini-la com apenas dois parâmetros, a média μ e o desvio padrão σ da distribuição, por exemplo, a curva da distribuição normal $f(x)$ para $\mu=40$, $\sigma=10$ e valores da variável aleatória no intervalo $(10, 70)$ é mostrada na Figura 8.4. Uma das características importantes é que a partir desses dois parâmetros será possível calcular, por exemplo, a percentagem de valores que deverão estar acima ou abaixo de um determinado valor da variável aleatória, ou entre esses dois valores definidos etc.

A variável aleatória X tem distribuição normal de probabilidades com parâmetros média $-\infty < \mu < +\infty$ e variância $0 < \sigma^2 < +\infty$, se a função densidade $f(x)$ for:²

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Analisando a fórmula da função densidade $f(x)$, observe que, para cada par de parâmetros μ e σ , há uma curva diferente de $f(x)$ ou que, para qualquer outro par de parâmetros μ e σ , a curva $f(x)$ será diferente. Portanto, não há apenas uma única distribuição normal e sim uma família de distribuições normais representadas como $N(\mu, \sigma)$.

² Os valores das constantes matemáticas da fórmula são $e=2,71828...$ e $\pi=3,1416...$


FIGURA 8.4

Distribuição normal, parâmetros $\mu=40$ e $\sigma=10$.

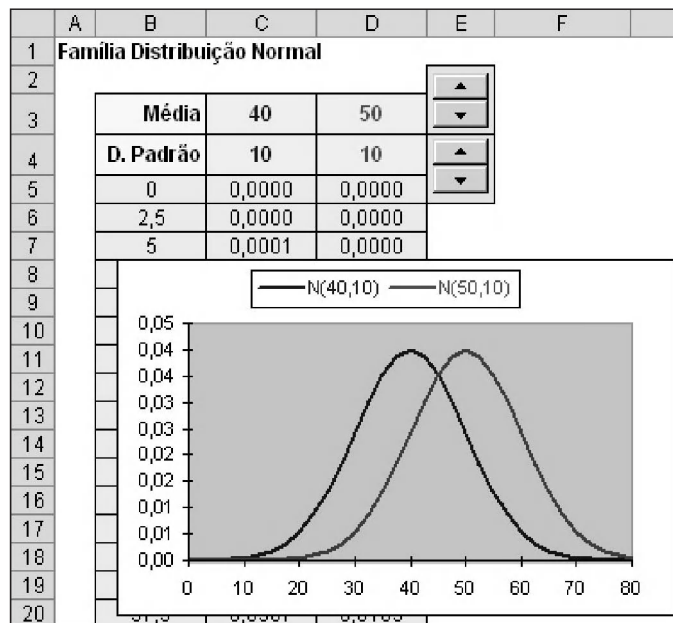
Propriedades da Distribuição Normal

A distribuição normal da variável aleatória X com média μ_X e variância σ_X^2 tem as seguintes propriedades:

- $f(x)$ tende a zero quando x tende a mais ou a menos infinito, ou $f(x) \rightarrow 0$ quando $x \rightarrow \pm\infty$
- A área total sob a curva é 100%, e cada metade da curva tem 50% da área total, pois a curva é simétrica ao redor da média μ_X .
- A probabilidade $P(a \leq X \leq b)$ é a área sob a curva no intervalo (a, b) .

Influência dos parâmetros na forma da distribuição normal

O que ocorreria com a forma da distribuição normal $N(40, 10)$ da Figura 8.4, se o valor da média for mudado de 40 para 50? A forma da distribuição permaneceria a mesma, porém com a média deslocada de 40 para 50 e definindo a nova distribuição normal $N(50, 10)$.


FIGURA 8.5

Distribuições normais para $\sigma=10$ e dois valores de média.

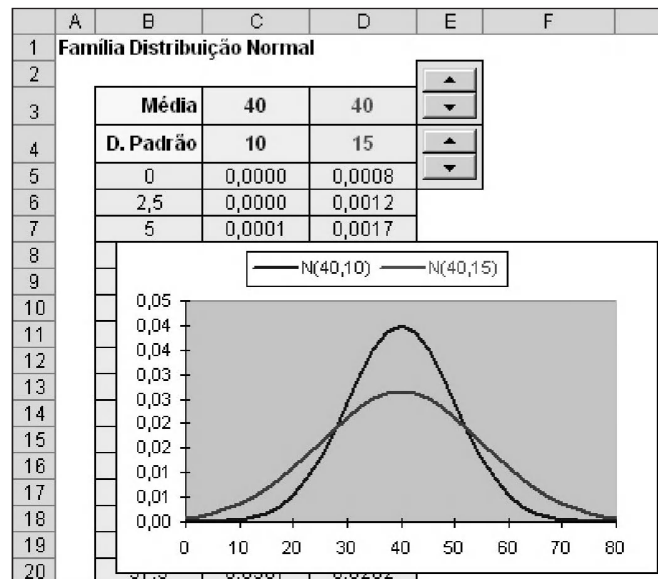
Na planilha **Família DN**, incluída na pasta **Capítulo 8**, é possível visualizar o comportamento da forma da distribuição normal quando muda o parâmetro média e o parâmetro desvio padrão. A tabela de cada gráfico da distribuição normal foi construída utilizando a fórmula da função $f(x)$ dessa distribuição. As duas curvas de distribuição normal da Figura 8.5 aceitam mudanças dos valores de média e desvio padrão, porém de forma diferente. No intervalo C3:C4, é possível mudar os parâmetros da curva de cor azul registrando novos valores; no entanto, os parâmetros da curva vermelha são mudados

acionando um dos dois controles giratórios por vez, célula E3 e E4. Observe que, ao variar o valor da média de 40 para 50, mantendo constante o desvio padrão $\sigma=10$, a distribuição normal mantém sua forma e se desloca para a direita. Da mesma maneira, para valores de média menores a 40, por exemplo $\mu=30$, a distribuição manterá sua forma e se deslocará para a esquerda de $\mu=40$.

Para analisar o comportamento da forma da distribuição normal variando o desvio padrão, lembremos primeiro o significado do desvio padrão apresentado no Capítulo 4. A terceira regra prática estabelece que, em todas as distribuições, a porcentagem de valores contidos dentro de três desvios padrão ao redor da média será próxima de 100%. Como a distribuição normal é simétrica ao redor da média, qualquer que for o valor do desvio padrão, praticamente 100% dos valores da variável aleatória estarão contidos dentro de três desvios padrão ao redor da média. Se na distribuição normal $N(40, 10)$, o desvio padrão for mudado para 15, a forma da curva da $N(40, 15)$ será mais aberta do que a anterior $N(40, 10)$. A Figura 8.6 mostra que a distribuição aumentou sua base e, conseqüentemente, diminuiu sua altura para manter a área de 100%. Da mesma maneira, pode-se verificar que, para desvios padrão menores do que 10, a distribuição diminui sua base e aumenta sua altura.

FIGURA 8.6

Distribuições normais para $\mu=40$ e desvios padrão diferentes.



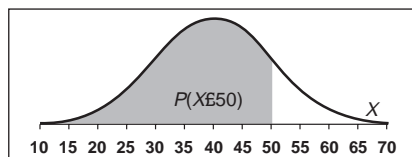
Cálculo de probabilidade

A probabilidade $P(a \leq X \leq b)$ da variável aleatória contínua X ser igual ou maior do que a e, ao mesmo tempo, menor ou igual a b é obtida da área definida pela função $f(x)$ entre os limites a e b , sendo $b > a$. O procedimento de cálculo passa pela integração da função $f(x)$ no intervalo (a,b) , procedimento bastante trabalhoso. Todavia, utilizando a função estatística `DIST.NORM` do Excel, esse cálculo se tornará mais simples e, neste momento, nos permitirá compreender o processo de cálculo da probabilidade $P(X)$ da variável aleatória contínua X com distribuição normal. Os exemplos a seguir mostram os procedimentos básicos de cálculo com a distribuição normal.

EXEMPLO 8.3

Os resultados do experimento formam uma variável aleatória X com distribuição normal $N(40, 10)$. Qual a probabilidade de um resultado do experimento ser menor ou igual a 50?

Solução. Os resultados do experimento têm distribuição normal com média $\mu=40$ e desvio padrão $\sigma=10$. Estamos procurando o valor da probabilidade $P(X \leq 50)$, a área sob a curva da distribuição normal no intervalo $(-\infty, 50)$, como mostra a figura seguinte.



O valor dessa área é o resultado da integração da função $f(x)$ entre menos infinito e 50. Uma forma prática de calcular a probabilidade $P(X \leq 50)$ é utilizando a função estatística DIST.NORM do Excel.

• **DIST.NORM(x ; média; desv_padrão; cumulativo)**

A função DIST.NORM³ retorna a função densidade da distribuição normal ou a probabilidade acumulada de menos infinito até o valor do argumento x , conforme o valor registrado no argumento *cumulativo*.

- Se no argumento *cumulativo* for registrado o valor FALSO, a função estatística DIST.NORM retornará a valor $f(x)$ para o valor x informado no primeiro argumento da função, considerando os parâmetros média e desvio padrão da distribuição, registrados no argumento *média* e no argumento *desv_padrão*.
- Se no argumento *cumulativo* for registrado o valor VERDADEIRO, a função DIST.NORM retornará a probabilidade acumulada de menos infinito até x , considerando os parâmetros média e desvio padrão da distribuição, registrados no argumento *média* e no argumento *desv_padrão*. Nesse caso, a função DIST.NORM calcula a integral da função $f(x)$ no intervalo $(-\infty, x)$.

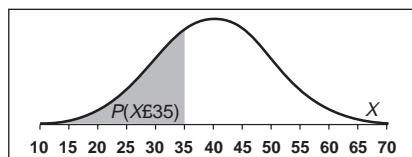
Neste exemplo, a probabilidade $P(X \leq 50)$ é igual a 84,13%, resultado obtido com a fórmula =DIST.NORM(50;40;10;VERDADEIRO), registrada em uma célula vazia de qualquer planilha Excel. Esse resultado tem o seguinte significado:

- A probabilidade de um resultado do experimento (ou valor da variável aleatória X) com distribuição normal $N(40, 10)$ ser menor ou igual a 50 é 84,13%.
- Ou podemos dizer que 84,13% dos resultados do experimento (ou valores da variável aleatória X) com distribuição normal $N(40, 10)$ estão dentro do intervalo $(-\infty, 50)$.

EXEMPLO 8.4

Continuando com a distribuição normal do Exemplo 8.3, qual a probabilidade de um resultado do experimento ser igual ou menor do que 35?

Solução. Este problema é parecido com o anterior, pois devemos calcular a probabilidade $P(X \leq 35)$ obtida da área sob a curva no intervalo $(-\infty, 35)$, como mostra a figura a seguir.



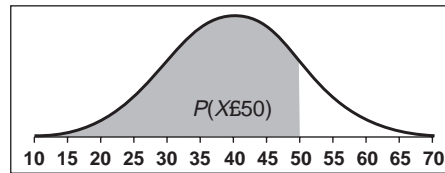
A probabilidade $P(X \leq 35)$ é igual a 30,85%, resultado obtido com a fórmula =DIST.NORM(35;40;10;VERDADEIRO), registrada em uma célula vazia de qualquer planilha Excel. Ou seja, a probabilidade de um resultado do experimento (ou valor da variável aleatória X) com distribuição normal $N(40, 10)$ ser menor ou igual a 35 é 30,85%, ou 30,85% dos resultados do experimento (ou valores da variável aleatória X) com distribuição normal $N(40, 10)$ estão dentro do intervalo $(-\infty, 35)$.

EXEMPLO 8.5

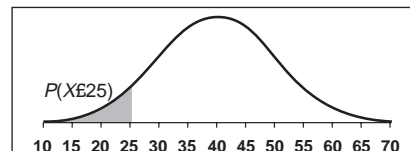
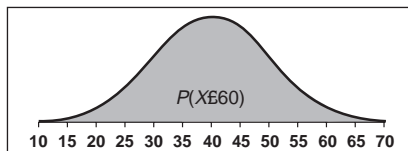
Continuando com a distribuição normal do Exemplo 8.3, qual a probabilidade de um resultado do experimento estar dentro do intervalo (25, 60)?

³ Em inglês, o nome da função DIST.NORM é NORMDIST.

Solução. A probabilidade de um resultado de o experimento pertencer ao intervalo (25, 60) é definida pela área sob a curva da distribuição normal definida entre os limites 25 e 60, como mostra a figura seguinte.



Como a função a função `DIST.NORM` retorna a probabilidade acumulada de menos infinito até x , o cálculo da probabilidade procurada $P(25 \leq X \leq 60)$ deverá ser realizado em duas partes $P(25 \leq X \leq 60) = P(X \leq 60) - P(X \leq 25)$. Primeiro deve ser calculada a probabilidade $P(X \leq 60) = 0,977250$, que corresponde à área da figura à esquerda. Depois deve ser calculada a probabilidade $P(X \leq 25) = 0,066807$, que corresponde à área da figura à direita. Subtraindo a área da figura da direita da área da figura da esquerda será obtida a primeira figura que representa a probabilidade procurada $P(25 \leq X \leq 60)$.



A probabilidade $P(25 \leq X \leq 60)$ é 91,04%, obtida como resultado da diferença das duas probabilidades anteriores $0,977250 - 0,066807 = 0,910443$ ou 91,04%. Esse resultado pode ser obtido com a função `DIST.NORM`, registrando a fórmula seguinte `=DIST.NORM(60;40;10;VERDADEIRO)-DIST.NORM(25;40;10;VERDADEIRO)`

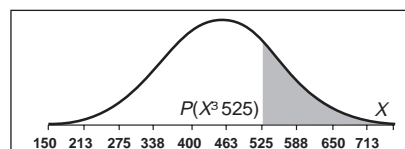
O resultado tem o seguinte significado:

- A probabilidade de um resultado do experimento (ou valor da variável aleatória X) com distribuição normal $N(40, 10)$ pertencer ao intervalo (25, 60) é 91,04%.
- Ou podemos dizer que 91,04% dos resultados do experimento (ou valores da variável aleatória X) com distribuição normal $N(40, 10)$ estão dentro do intervalo (25, 60).

EXEMPLO 8.6

A Gerência de Desenvolvimento da empresa aplicou um teste de conhecimentos gerais. O resultado mostrou que as respostas podem ser representadas com uma distribuição normal $N(450, 100)$. Se João Pedro obteve 525 pontos, que percentagem de funcionários tiraram mais pontos do que ele?

Solução. Agora você deve calcular a probabilidade $P(X \geq 525)$, como mostra a próxima figura. Como a função `DIST.NORM` retorna a probabilidade acumulada de menos infinito até x , este problema deverá ser resolvido pelo complemento $P(X \geq 525) = 1 - P(X \leq 525)$.



Registrando a fórmula `=1-DIST.NORM(525;450;100;VERDADEIRO)`, obtém-se $P(X \geq 525) = 22,66\%$. Ou seja, 22,67% dos funcionários que participaram do teste de conhecimentos gerais tiraram mais pontos do que João Pedro.

Resultados importantes da distribuição normal

A fórmula da função densidade $f(x)$ mostra que, para cada par de parâmetros μ e σ , há uma curva diferente de $f(x)$, ou para qualquer outro par de parâmetros μ e σ , a curva $f(x)$ será diferente. Embora não haja apenas uma única distribuição normal e sim uma família de distribuições normais representadas com $N(\mu, \sigma)$, elas mantêm algumas propriedades em comum, por exemplo, a porcentagem de resultados ao redor da média.

Um desvio padrão ao redor da média.

Qual a porcentagem de valores localizados dentro de um desvio padrão ao redor da média 200 da variável aleatória X com distribuição normal $N(200, 10)$? De outra maneira, um valor de X incluído no intervalo de um desvio padrão ao redor da média deve estar no intervalo (190, 210). A probabilidade de um valor da variável aleatória X estar dentro de um desvio padrão ao redor da média é $P(190 \leq X \leq 210) = 0,6827$ ou 68,27%, resultado obtido com a fórmula
`=DIST.NORM(210;200;10;VERDADEIRO)– DIST.NORM(190;200;10;VERDADEIRO)`

Esse resultado mostra que 68,27% dos valores da variável X com distribuição normal se distribuem no intervalo de um desvio padrão ao redor da média. Ou que a probabilidade de um resultado do experimento com distribuição $N(200, 10)$ pertencer ao intervalo (190, 210) é 68,27%.

Dois desvios padrão ao redor da média.

Qual a porcentagem de valores da variável X com distribuição normal $N(200, 10)$ localizados dentro de dois desvios padrão ao redor da média? Um valor de X incluído no intervalo de dois desvios padrão ao redor da média deve estar no intervalo (180, 220). A probabilidade de um valor da variável aleatória X estar dentro de dois desvios padrão ao redor da média é $P(180 \leq X \leq 220) = 0,9545$ ou 95,45%, resultado obtido com a fórmula:

`=DIST.NORM(220;200;10;VERDADEIRO)– DIST.NORM(180;200;10;VERDADEIRO)`

Esse resultado mostra que 95,45% dos valores da variável X com distribuição normal se distribuem no intervalo de dois desvios padrão ao redor da média. Ou que a probabilidade de um resultado do experimento com distribuição $N(200, 10)$ pertencer ao intervalo (180, 220) é 95,45%.

Três desvios padrão ao redor da média.

Qual a porcentagem de valores da variável X com distribuição normal $N(200, 10)$ localizados dentro de três desvios padrão ao redor da média? Um valor de X incluído no intervalo de três desvios padrão ao redor da média deve estar no intervalo (170, 230). A probabilidade de um valor da variável aleatória X estar dentro de três desvios padrão ao redor da média é $P(170 \leq X \leq 230) = 0,9973$ ou 99,73%, resultado obtido com a fórmula: `=DIST.NORM(230;200;10;VERDADEIRO)– DIST.NORM(170;200;10;VERDADEIRO)`

Esse resultado mostra que 99,73% dos valores da variável X com distribuição normal se distribuem no intervalo de três desvios padrão ao redor da média. Ou que a probabilidade de um resultado do experimento com distribuição $N(200, 10)$ pertencer ao intervalo (170, 230) é 99,73%.

A Figura 8.7 mostra o gráfico da distribuição normal $f(x)$ construído na planilha **Modelo Distribuição Normal** incluído na pasta **Capítulo 8** e incluindo a representação dos intervalos dos três tipos de desvios padrão utilizando os dados da apresentação dos resultados importantes da distribuição normal. Com os controles giratórios, são informados os valores da média na célula C3 e o desvio padrão na célula C4 e, ao mesmo tempo, a planilha recalcula a tabela do intervalo B6:C31, de acordo com o tipo de probabilidade escolhida na caixa de grupo:

FIGURA 8.7 Gráfico $f(x)$ da DN, incluindo os intervalos dos desvios padrão.

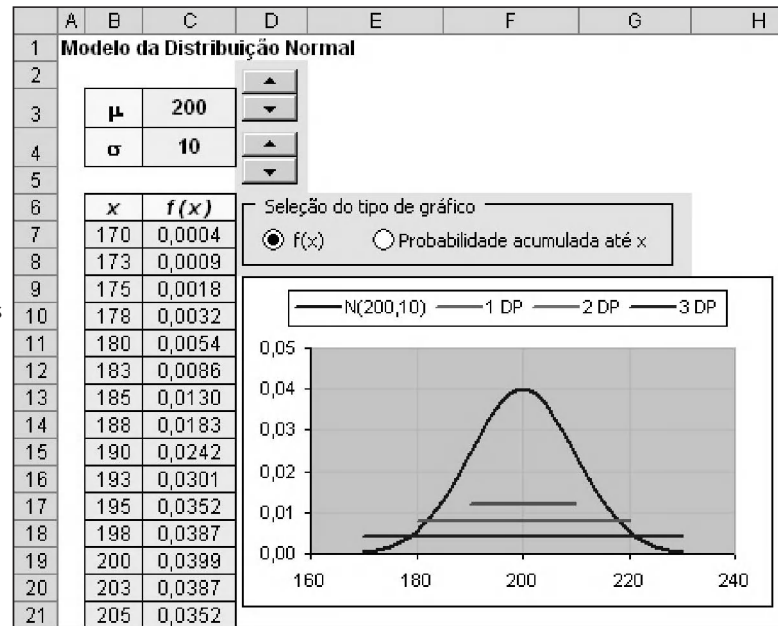
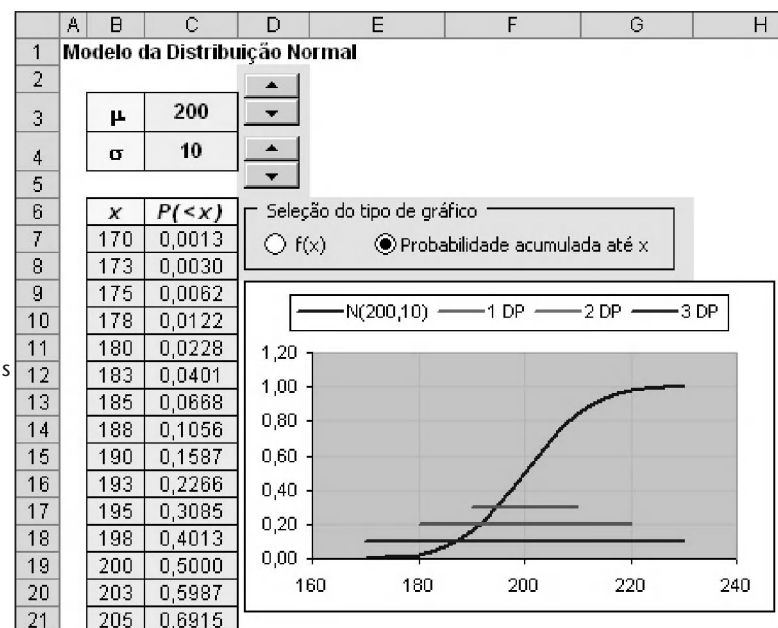


FIGURA 8.8 Probabilidade acumulada da DN, com os intervalos dos desvios padrão.



- $f(x)$. Fornecerá o valor da função $f(x)$ para o valor x .
- *Probabilidade acumulada até x* . Fornecerá a probabilidade acumulada de menos infinito até x .

A Figura 8.8 mostra o gráfico da probabilidade acumulada de menos infinito até x da DN, incluindo a representação dos intervalos dos três desvios padrão.

Modelo DN

O modelo estatístico **Modelo DN**, incluído na pasta **Capítulo 8**, facilita os cálculos de probabilidade com a distribuição normal $N(\mu, \sigma)$, como mostra a Figura 8.9, calculando a probabilidade de um valor da variável aleatória com distribuição $N(200, 10)$ pertencer ao intervalo $(180, 220)$. No *modelo*:

- As células do intervalo C3:C6, pintadas de cor azul, aceitam somente dados, e as células do intervalo C7:D7, pintadas de cor verde, são células de resultados.
- Nas células D5 e D6 são selecionados os tipos de limite do intervalo para cálculo da probabilidade desejada. Assim, no exemplo da Figura 8.9, o limite inferior é ≥ 180 e o limite superior ≤ 220 . Os dados registrados no intervalo C5:C6 são utilizados para definir os limites nas duas caixas de combinações.
 - Na caixa de combinação da célula D5, é possível escolher os limites \leq e \geq . Na caixa de combinação da célula D6, é possível escolher os limites \leq , \geq e *Não*.
- Na célula D7, é registrado o resultado numérico e, no intervalo B7:C7, é informado o texto da probabilidade calculada.

	A	B	C	D
1	Modelo DN			
2				
3		μ	200	
4		σ	10	
5			180	X1 \geq 180 ▼
6			220	X2 \leq 220 ▼
7			$P(180 \leq X \leq 220)$	95,45%
8				

FIGURA 8.9
Modelo DN
calculando um
resultado importante.

Distribuição normal padronizada

Na primeira parte da apresentação da distribuição normal foi mostrado que:

- A probabilidade $P(X \leq a)$ é o resultado de integrar a função $f(x)$ da distribuição normal $N(\mu, \sigma)$ entre os limites $(-\infty, a)$.
- O cálculo de probabilidades é bastante trabalhoso, pois não há apenas uma única distribuição normal e sim uma família de distribuições normais $N(\mu, \sigma)$.

A função estatística DIST.NORM do Excel e o **Modelo DN** reduzem sensivelmente o procedimento de cálculo; entretanto, o procedimento clássico de cálculo de probabilidades utiliza a *distribuição normal padronizada* obtida da distribuição normal $N(\mu, \sigma)$, realizando a mudança da variável X .

A variável aleatória *desvio padrão normalizado* Z de uma *distribuição normal padronizada* é definida pela expressão $Z = \frac{X - \mu}{\sigma}$

A nova variável Z realiza cálculos de probabilidades com uma única curva de distribuição denominada distribuição normal padronizada $N(0, 1)$. Depois da transformação da variável, a função densidade de $f(x)$ passa a ser $f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$.

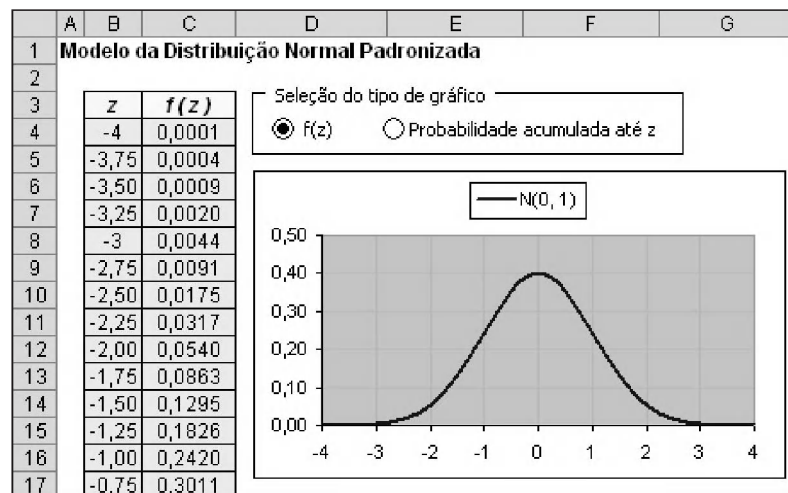
Propriedades da Distribuição Normal Padronizada

A distribuição normal da variável aleatória Z com média μ_Z e variância σ_Z^2 tem as seguintes propriedades:

- $\mu_Z = 0$ e $\sigma_Z^2 = 1$
- $f(Z) \rightarrow 0$ quando $Z \rightarrow \pm \infty$
- A curva é simétrica ao redor da média $\mu_Z = 0$. A área total sob a curva é 100%, e cada metade da curva tem 50% da área total.
- A probabilidade $P(Z_1 \leq Z \leq Z_2)$ é a área sob a curva no intervalo (Z_1, Z_2) .

A planilha **Modelo DN Padronizada**, incluída na pasta **Capítulo 8** mostra a curva única da variável Z com distribuição normal $N(0, 1)$. A tabela e o gráfico foram construídos com os limites de quatro desvios padrão ao redor da média, Figura 8.10. Uma das vantagens da distribuição normal padronizada é a facilidade de visualizar resultados como, por exemplo, a percentagem de valores da variável Z com um, dois e três desvios padrão ao redor da média. Outra vantagem é que, com a passagem para o desvio padrão normalizado Z , a distribuição Z agrupa a família da distribuição normal em uma única distribuição, o desvio padrão normalizado.

FIGURA 8.10
Distribuição normal padronizada.



Depois de realizar a transformação, a curva da distribuição normal padronizada Z tem a mesma forma que a da distribuição normal, porém com média e desvio padrão, respectivamente, zero e um, representada como $N(0, 1)$. A seguir, será mostrado o procedimento de cálculo de probabilidade utilizando a distribuição Z .

EXEMPLO 8.7

Qual a probabilidade de um valor de X ser menor ou igual a 52,4, sabendo que a variável tem distribuição normal (40, 10)?

Solução. O objetivo é determinar a probabilidade $P(X \leq 52,4)$. Para utilizar a distribuição normal padronizada, o primeiro passo é determinar $Z=1,24$, resultado obtido com a fórmula:

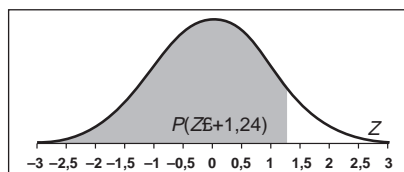
$$Z = \frac{X - \mu}{\sigma} = \frac{52,4 - 40}{10} = +1,24$$

Esse resultado pode ser obtido com a função PADRONIZAR do Excel.

• **PADRONIZAR(*x*; *média*; *desv_padrao*)**

A função estatística PADRONIZAR⁴ retorna o desvio padrão normalizado Z considerando os argumentos x , $média$ e $desv_padrao$ e utilizando a fórmula conhecida $Z = \frac{x - \mu}{\sigma}$. Digitando a fórmula =PADRONIZAR(52,4;40;10) numa célula vazia da planilha Excel, obtém-se o resultado procurado $Z=1,24$.

Neste momento, temos a equivalência $P(X \leq 52,4) = P(Z \leq 1,24)$. A probabilidade que estamos procurando se



encontra na área sombreada da figura a seguir.

O valor da probabilidade $P(Z \leq 1,24)$ pode ser obtido com a função DIST.NORM, registrando a fórmula =DIST.NORM(1,24;0;1;VERDADEIRO) numa célula da planilha Excel para obter o resultado procurado $P(Z \leq 1,24) = 0,8925$. Esse resultado também pode ser obtido com a função DIST.NORMP do Excel para a distribuição normal padronizada.

• **DIST.NORMP(*z*)**

A função estatística DIST.NORMP⁵ retorna a probabilidade acumulada da distribuição normal padronizada Z , de menos infinito até o valor registrado no argumento z .

Registrando a fórmula =DIST.NORMP(1,24) numa célula da planilha Excel, obtém-se a probabilidade $P(Z \leq 1,24) = 0,8925$. Para obter esse resultado, primeiro foi obtido o valor de z ; contudo, poderíamos evitar esse cálculo registrando a fórmula =DIST.NORMP(PADRONIZAR(52,4;40;10)), sendo que, no lugar do argumento z , foi incluída a função PADRONIZAR que calcula o valor z .

Tabela Z

Para mostrar a utilização da Tabela Z, serão utilizados os dados do Exemplo 8.7. A tabela Z_DISTR foi construída na pasta Tabelas disponível na página do livro, no site da Editora, e reproduzida parcialmente na Figura 8.11. Essa tabela começa no valor $z=0$, com o valor 0,5000 correspondente à probabilidade acumulada da metade negativa da curva, região de valores de z negativos. Tanto a primeira coluna quanto os títulos dos cabeçalhos das demais colunas registram valores de z . Na primeira coluna, são registrados os valores de z , variando com intervalo de um décimo, por exemplo, 0; 0,10; 0,20, e assim por diante, até o valor $z=4$. Nos cabeçalhos das colunas, são registrados os valores de z variando com intervalo de um centésimo, por exemplo, 0,00; 0,01; 0,02, e assim por diante, até o valor $z=0,09$.

Vejamos como se procura o valor da probabilidade $P(Z \leq 1,24)$ na tabela Z. A localização do valor z na tabela começa pela seleção da linha 1,20 e depois da coluna 0,04. Na interseção dessa linha e dessa coluna, tem-se $z=1,24$ e, nessa célula, o valor procurado $P(Z \leq 1,24) = 0,8925$, como mostra a Figura 8.11. Qual o significado do resultado $P(Z \leq 1,24)$? O resultado 0,8925 é a probabilidade $P(-\infty \leq Z \leq 1,24)$, correspondente à área da curva de Z no intervalo $(-\infty, +1,24)$. Quando o valor de z não estiver explicitamente registrado na planilha, deverá ser realizada uma interpolação linear entre os dois valores mais próximos, um deles o maior seguinte, e o outro, o menor anterior ao valor de z procurado.

⁴ Em inglês, o nome da função é STANDARDIZE.

⁵ Em inglês, o nome da função é NORMSDIST.

Z	0,00	0,01	0,02	0,03	0,04	0,05
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115

FIGURA 8.11 Tabela da distribuição Z, listagem parcial.

A curva da distribuição Z tem a mesma forma que a curva da distribuição X, com a mudança dos valores do eixo X para o eixo Z. Sugerimos que você tente visualizar a mesma curva de distribuição normal com os dois eixos X e Z. Como há diversos cálculos de probabilidade, a tabela da Figura 8.12 apresenta o resumo dos procedimentos de cálculo de probabilidade utilizando a tabela da distribuição normal padronizada Z.

Tipo de cálculo	Resultado da tabela Z
$P(Z \leq a); a > 0$	Direto da tabela
$P(Z \geq a); a > 0$	$1 - P(Z \leq a)$
$P(Z \leq a); a < 0$	$1 - P(Z \leq a)$
$P(Z \geq a); a < 0$	$P(Z \leq a)$

FIGURA 8.12 Procedimentos de cálculo com a Tabela Z.

A tabela da Figura 8.12 mostra como operar com a tabela quando se procura o valor de probabilidade diferente do registrado na tabela. Essas regras registradas na tabela utilizam a propriedade de simetria da distribuição normal. Por exemplo, a probabilidade $P(Z \geq a)$, sendo a positivo, corresponde à cauda superior da curva a partir do valor a . A área que procuramos é a área complementar de $P(Z \leq a)$, que pode ser obtida da tabela. Portanto, o valor de $P(Z \geq a)$ procurado é o resultado de $(1 - P(Z \leq a))$.

EXEMPLO 8.8

Se a variável X tem distribuição normal com parâmetros $N(40, 10)$, qual a probabilidade $P(X \leq 50)$?

Solução. O valor equivalente de $X=50$ na distribuição normal padronizada é $Z=+1$, resultado obtido com a fórmula $Z = \frac{X - \mu}{\sigma} = \frac{50 - 40}{10} = +1$. Da tabela Z , para $Z=+1$, obtém-se $P(X \leq 50) = P(Z \leq 1) = 0,8413$ ou 84,13%.

O mesmo resultado é obtido com:

- A fórmula =PADRONIZAR(50,4;10) registrada em uma célula de uma planilha Excel.
- Lembrando da propriedade da distribuição normal, é possível realizar um cálculo mental. Por exemplo, o valor $X=50$ se situa a um desvio padrão positivo da média. Sabemos que 68,26% dos valores de X se distribuem no intervalo de um desvio padrão ao redor da média; portanto, a metade dessa percentagem, 34,13%, somada à metade com desvio negativo de 50% resulta no valor de distribuição procurado, 84,13%.

EXEMPLO 8.9

Continuando com a variável X do Exemplo 8.8, qual a probabilidade $P(X \leq 35)$?

Solução. Para $X=35$, obtém-se $Z=-0,50$, porém na tabela não podemos procurar $P(Z \leq -0,50)$. Contudo, a terceira linha da tabela da Figura 8.12 registra que a probabilidade $P(Z \leq a)$, sendo a negativo, deve ser calculada como complemento utilizando a fórmula $1 - P(Z \leq |a|)$, onde $|a|$ significa que deve ser utilizado o valor a como positivo. Nesse caso, $1 - P(Z \leq |a|) = 1 - P(Z \leq +0,50) = 1 - 0,6915 = 0,3085$.

EXEMPLO 8.10

Continuando com a variável X do Exemplo 8.8, qual a probabilidade $P(25 \leq X \leq 60)$?

Solução. Os valores de Z para $X=25$ e $X=60$ são $Z_1=-1,50$ e $Z_2=+2,0$. Como a probabilidade $P(-1,5 \leq Z \leq +2,0)$ não é obtida diretamente da tabela Z , ela será o resultado da diferença $P(Z \leq +2) - P(Z \leq -1,5)$, procedimento que deve ser realizado por partes:

- A probabilidade $P(Z \leq +2,0)$ é igual a 0,9772, obtida diretamente da tabela.
- A segunda parte é $P(Z \leq -1,5) = 1 - P(Z \leq +1,5) = 1 - 0,9332 = 0,0668$.
Finalmente, $P(-1,5 \leq Z \leq +2,0) = 0,9772 - 0,0668 = 0,9104$ ou 91,04%.

Modelo cálculos com DN

O modelo estatístico Cálculos DN, incluído na pasta Capítulo 8, é equivalente ao Modelo DN, porém incluindo os valores de Z e o gráfico da distribuição que mostra as áreas sob a curva que estão sendo utilizadas no cálculo requerido. A Figura 8.13 mostra o modelo calculando a probabilidade de um valor da variável aleatória com distribuição $N(200, 10)$ pertencer ao intervalo (190, 210). No modelo:

- As células do intervalo C3:C6 aceitam somente dados.
- Nas células D4 e D5, são selecionados os tipos de limite do intervalo para cálculo da probabilidade desejada. Assim, no exemplo da Figura 8.13, o limite inferior é ≥ 190 e o limite superior ≤ 210 . Os dados registrados no intervalo C5:C6 são utilizados para definir os limites nas duas caixas de combinações.
 - Na caixa de combinação da célula D5, é possível escolher os limites \leq e \geq . Na caixa de combinação da célula D6, é possível escolher os limites \leq , \geq e Não.
- Nas células do intervalo F5:F6, são registrados os valores de Z correspondentes aos limites registrados em C5:C6.
- No intervalo B7:E7, é registrado o resultado informando o texto da probabilidade calculada.

- O gráfico da distribuição normal mostra a(s) área(s) sob a curva que está(ão) sendo utilizada(s) no cálculo requerido.

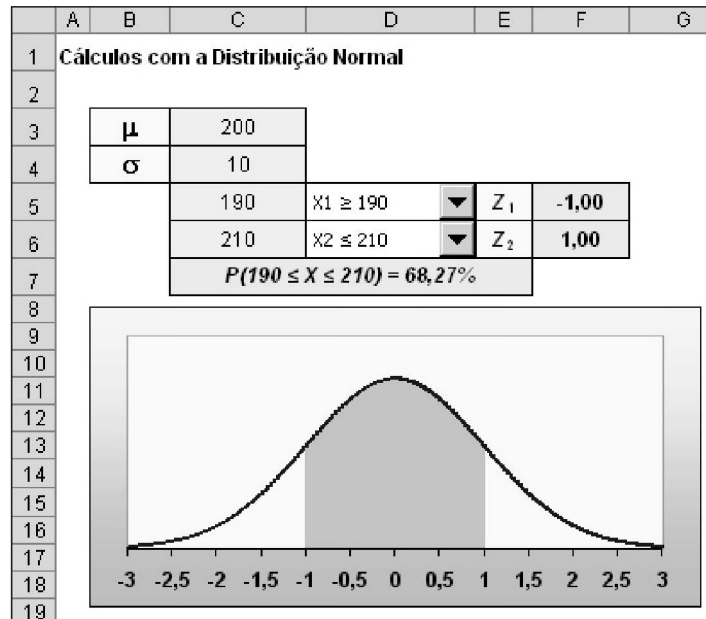


FIGURA 8.13 Modelo Cálculos DN.

Outros cálculos com a distribuição normal

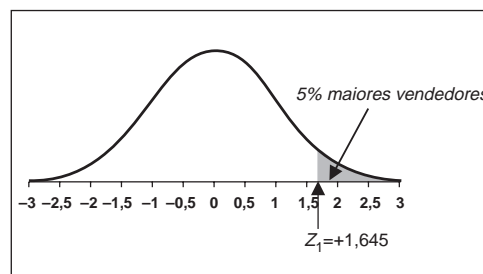
O objetivo dos exemplos de distribuição normal apresentados até este momento foi o cálculo da probabilidade de que ocorra um determinado evento, sendo conhecidos os parâmetros da distribuição normal, a média e o desvio padrão. Há outros problemas com a distribuição normal, por exemplo, o cálculo inverso e a determinação dos parâmetros da distribuição normal a partir de constatações práticas.

Cálculo inverso

O cálculo inverso calcula o valor x da variável aleatória X correspondente a uma determinada probabilidade de ocorrência.

EXEMPLO 8.11

Jota afirma que está entre os 5% maiores vendedores da empresa, pois seu total de vendas no ano passado foi de \$1.350.000. Considerando que as vendas de todos os vendedores têm a distribuição normal $N(\$1.250000, \$100.000)$, verifique se a afirmação do vendedor Jota é correta.



Solução. Os 5% maiores vendedores da empresa estão localizados no final da cauda superior da distribuição normalizada da figura acima. Qual o valor Z_1 que verifica a relação $P(Z \geq Z_1) = 5\%$? Como a tabela não tem registrada essa parte da área da curva, deveremos procurar a probabilidade complementar 0,95 obtida como resul-

tado da diferença (1-0,05). Procurando no miolo da tabela Z, verificamos que o valor 0,95 não coincide com nenhum dos valores registrados na tabela Z, pelo que será necessário realizar uma interpolação. Como o valor 0,95 situa-se entre 0,9495 ($Z=1,64$) e 0,9505 ($Z=1,65$), interpolando entre esses valores obtém-se $Z=1,645$, que corresponde à probabilidade $P(Z \geq 1,645)=0,05$. Com os dados disponíveis, a $N(\$1.250.000, \$100.000)$ e o valor $Z=1,645$, calculamos o valor de venda mínimo x correspondente a 5% dos maiores vendedores, utilizando a fórmula do desvio padrão normalizado Z:

$$Z = \frac{x - \mu}{\sigma} = \frac{x - \$1.250.000}{\$100.000} = 1,645$$

Dessa expressão, temos que o valor de venda mínimo x correspondente a 5% dos maiores vendedores é $x=\$1.414.500$, como mostra a fórmula:

$$x = 1,645 \times \$100.000 + \$1.250.000 = \$1.414.500.$$

Concluindo, para pertencer ao grupo dos 5% maiores vendedores da empresa, seria necessário vender pelo menos \$1.414.500. Como o vendedor Jota vendeu \$1.350.000, ele não pertence ao grupo dos 5% maiores vendedores.

- Verifique que o salário de Jota está no grupo dos 15,87% maiores vendedores da empresa, resultado obtido pelo cálculo direto da probabilidade a partir do valor de vendas anuais, $P(X \geq 1.350.000)=0,1587$. Como ajuda, verifique que o valor de vendas de Jota está a um desvio padrão da média da distribuição.

Outra forma de calcular o valor de venda mínimo X correspondente a 5% dos maiores vendedores é utilizando a função estatística INV.NORM.

• INV.NORM(probabilidade; média; desv_padrão)

A função INV.NORM⁶ retorna o valor de x correspondente aos argumentos *probabilidade*, *média* e *desv_padrão*. A função INV.NORM é a inversa da função DIST.NORM, com argumento *cumulativo* VERDADEIRO. Para calcular o valor de venda mínimo x correspondente a 5% dos maiores vendedores, registramos a fórmula =INV.NORM(0,95;1250000;100000) em uma célula vazia da planilha Excel. Essa fórmula retorna o resultado \$1.414.485, resultado um pouco diferente do obtido da tabela Z com valores arredondados.

- No cálculo de x , o Excel utiliza um procedimento iterativo até alcançar um erro de $\pm 3 \times 10^{-7}$. Entretanto, se até 100 iterações não for possível encontrar o resultado, a função INV.NORM retornará o resultado #N/A.

O Excel também dispõe da função INV.NORMP que pode ser utilizada.

• INV.NORMP(probabilidade)

A função INV.NORMP⁷ retorna o valor Z da *probabilidade* informada, valor entre zero e um. A função DIST.NORMP considera que a *probabilidade* informada se refere à probabilidade acumulada de menos infinito até Z , pois esta função é a inversa da função DIST.NORMP. Para calcular o valor de venda mínimo x correspondente a 5% dos maiores vendedores, registramos a fórmula =INV.NORMP(0,95)*100000+1250000, retornando \$1.414.485, o mesmo resultado obtido com a função INV.NORM.

- Neste caso, também, no cálculo de Z , o Excel utiliza um procedimento iterativo até alcançar um erro de $\pm 3 \times 10^{-7}$. Entretanto, se até 100 iterações não for possível encontrar o resultado, a função DIST.NORMP retornará o resultado #N/A.

Outra forma de resolver o Exemplo 8.11 é utilizando a planilha **Cálculo Inverso**, incluída na pasta **Capítulo 8**, como mostra a Figura 8.14. Este modelo é parecido com os anteriores, porém adaptado para retornar o valor de x nas duas possíveis respostas selecionadas na caixa de combinação da célula D6.

⁶ Em inglês, o nome da função INV.NORM é NORMINV.

⁷ Em inglês, o nome da função INV.NORMP é NORMSINV.

FIGURA 8.14 Planilha **Cálculo Inverso**, resolvendo o Exemplo 8.11.

	A	B	C	D	E	F
1	Cálculo Inverso					
2						
3		μ	1.250.000			
4		σ	100.000			
5		$P(X \geq 1.414.485)$	5,00%			
6			1.414.485	$X \geq 1.414.485$	Z	1,64
7						

EXEMPLO 8.12

Os registros históricos da loja mostram que a demanda mensal do sabonete especial Alfa tem distribuição normal com média 2.400 e desvio padrão 230. Como o valor médio do ticket de compra desses compradores é o mais alto da loja, o gerente quer garantir que 99% dessas vendas sejam atendidas. Calcule o estoque que a loja deve ter no início de cada mês.

Solução. Para que 99% das vendas do sabonete especial Alfa sejam atendidas, a loja deverá ter no início de cada um estoque de 2.935 unidades, resultado obtido com a fórmula =INV.NORM(0,99;2400;230). Esse resultado pode ser obtido com a planilha **Cálculo Inverso**, como mostra a figura seguinte.

	A	B	C	D	E	F
1	Cálculo Inverso					
2						
3		μ	2.400			
4		σ	230			
5		$P(X \leq 2.935)$	99,00%			
6			2.935	$X \leq 2.935$	Z	2,33
7						

Cálculo dos parâmetros da distribuição normal

O Exemplo 8.13 mostra outro tipo de problema com a distribuição normal, no qual a variável aleatória tem distribuição normal, porém seus parâmetros não são conhecidos.

EXEMPLO 8.13

Como costuma ocorrer, o diretor de novos projetos necessita, para ontem, da estimativa preliminar do valor do investimento do lançamento do novo produto. Quando pergunta ao gerente de novos projetos da empresa, que tem muita experiência na avaliação desse tipo de projeto, bastante frequente na empresa, ele responde que a estimativa do investimento se situa entre \$1.500.000 e \$2.000.000, com 50% de probabilidade de acerto. Qual o valor dos parâmetros dessa distribuição, considerando que a variável investimento tem distribuição normal?

Solução.

- **Determinação da média.** A estimativa mínima \$1.500.000 e a estimativa máxima \$2.000.000 do investimento definem a média \$1.250.000 do investimento, pois a variável aleatória investimento tem distribuição normal.
- **Determinação do desvio padrão.** Como a estimativa do investimento se situa entre \$1.500.000 e \$2.000.000, com 50% de probabilidade de acerto, devemos entender que elas se distribuem ao redor da média em duas metades iguais de 25% de cada lado da média. Dessa maneira:

- O desvio padrão normalizado Z do investimento máximo \$2.000.000 é $Z=0,675$, valor obtido da tabela Z , procurando a probabilidade 0,75 (0,50+0,25), ou da fórmula =INV.NORMP(0,75) que retorna o valor 0,6745. Ademais, pela simetria da distribuição normal, o desvio padrão normalizado Z do limite inferior \$1.500.000 será $Z=-0,6745$.

Se na fórmula do desvio padrão normalizado Z , substituirmos os dados:

$$0,6745 = \frac{\$2.000.000 - \$1.750.000}{\sigma}$$

Obteremos o desvio padrão da estimativa do investimento, que é \$370.645:

$$\sigma = \frac{\$250.000}{0,675} = \$370.645$$

Os resultados mostram que a distribuição normal da estimativa do investimento é $N(1.750.000, 370.645)$.

Os parâmetros da distribuição normal do Exemplo 8.13 podem ser obtidos com a planilha **Cálculo Parâmetros DN**, incluída na pasta **Capítulo 8**, como mostra a Figura 8.15. Esse modelo também é parecido com os anteriores, porém adaptado para retornar os valores dos parâmetros da distribuição normal, utilizando um dos dois valores, *Mínimo* e *Máximo*, selecionado na caixa de combinação da célula D6.

	A	B	C	D
1	Cálculo dos Parâmetros da DN			
2				
3		Mínimo	1.500.000,00	
4		Máximo	2.000.000,00	
5		Probabilidade	50,00%	
6		Z	0,6745	X = 2.000.000,00 ▼
7		μ	1.750.000,00	
8		σ	370.650,68	
9				

FIGURA 8.15 Planilha **Cálculo Parâmetros DN**, resolvendo o Exemplo 8.13.

EXEMPLO 8.14

Para definir o preço unitário de um novo produto, o gerente do produto costuma analisar dois cenários, um otimista e o outro pessimista. No caso do novo detergente em cubos para máquina de lavar louças, ele definiu:

- O preço do cenário otimista de \$25 por pacote, considerando que a probabilidade de exceder esse valor seja de 5%.
- O preço do cenário pessimista de \$18 por pacote, considerando que a probabilidade de reduzir esse valor seja de 5%.

Considerando que o preço unitário tenha distribuição normal, qual o valor dos parâmetros dessa distribuição?

Solução. Os dados do exemplo mostram que o preço médio unitário do novo detergente é \$21,50, resultado obtido como média dos preços dos dois cenários. As probabilidades de 5%, em ambos os extremos da distribuição, de exceder o valor otimista e de reduzir o valor pessimista, definem a probabilidade de 90% entre os preços dos cenários pessimista e otimista. A probabilidade 90% se distribui ao redor da média em duas metades iguais, 45% de cada lado. Dessa maneira, o desvio padrão Z do preço do cenário otimista \$25 é $Z=1,645$, resultado obtido com a fórmula =INV.NORMP(0,95). Continuando com o procedimento de cálculo, obtém-se o desvio padrão da estimativa do investimento igual a \$2,13. Esse resultado também pode ser obtido com a planilha **Cálculo Parâmetros DN**, como mostra a figura a seguir, utilizando o limite inferior.

	A	B	C	D
1	Cálculo dos Parâmetros da DN			
2				
3		Mínimo	18,00	
4		Máximo	25,00	
5		Probabilidade	90,00%	
6		Z	-1,6449	X = 18,00
7		μ	21,50	
8		σ	2,13	
9				

O próximo Exemplo 8.15 mostra como calcular a média da distribuição normal conhecendo o desvio padrão da distribuição e atendendo a certas especificações de valor da variável aleatória e sua probabilidade de ocorrência.

EXEMPLO 8.15

O produto farmacêutico colocado é enchido em frascos por uma máquina automática que pode ser ajustada em qualquer volume entre dez e vinte centímetros cúbicos. O volume do produto é uma variável aleatória com distribuição normal com desvio padrão de 0,4 centímetro cúbico. A especificação do controle de qualidade exige que pelo menos 98% dos frascos contenham 16 centímetros cúbicos ou mais. Para qual o volume que a máquina deve ser ajustada?

Solução. A incógnita do problema é a média da distribuição normal, pois a máquina automática pode ser ajustada em qualquer volume entre dez e vinte centímetros cúbicos. Especificar que 98% dos frascos encheidos contenham pelo menos 16 centímetros cúbicos é equivalente a afirmar que 2% das ampolas encheidas terão volume inferior a 16 centímetros cúbicos. Portanto:

- O desvio padrão Z da probabilidade acumulada 2% de menos infinito até 16 é $Z = -2,054$, resultado obtido com a fórmula `=INV.NORMP(0,02)`.
- Substituindo os dados disponíveis na fórmula do desvio padrão normalizado teremos $-2,054 = \frac{16 - \mu}{0,40}$.

A máquina de enchimento deverá ser regulada no valor 16,82 centímetros cúbicos, resultado obtido com a fórmula $\mu = 16 + 2,054 \times 0,40 = 16,82$.

A média da distribuição normal do Exemplo 8.15 pode ser obtida com o modelo construído a partir da linha 11 da planilha **Cálculo Parâmetros DN**, incluída na pasta **Capítulo 8**, como mostra a Figura 8.16. Este modelo também é parecido com os anteriores, porém adaptado para retornar o valor da média da distribuição normal, utilizando um dos dois lados da distribuição normal. Selecionando a relação adequada na caixa de combinação da célula D16, de acordo com a área da probabilidade utilizada, o modelo calcula os valores de Z e da média.

FIGURA 8.16 Planilha **Cálculo Parâmetros DN**, resolvendo o Exemplo 8.15.

	A	B	C	D
11	Cálculo da Média da DN			
12				
13		σ	0,40	
14		X	16	
15		Probabilidade	98,00%	
16		Z	-2,054	X ≥ 16
17		μ	16,82	
18				

Distribuição exponencial

A *distribuição exponencial* é uma distribuição contínua aplicada em muitos problemas de empresas nas áreas de serviços e manufaturas, em geral denominados problemas de *fila de espera*. Quando os serviços prestados por uma empresa para clientes externos ou internos são de duração variável, a distribuição exponencial é indicada para analisar esses experimentos, por exemplo, a duração do atendimento do caixa de um banco ou de postos de saúde, o tempo de operação sem interrupção de um equipamento etc. A distribuição exponencial é definida pelo único parâmetro λ , denominado média, que mede a média de chegadas por hora, por exemplo, ou de serviços por minuto ou alguma outra unidade de tempo.

Probabilidade da Distribuição Exponencial

- A função densidade de probabilidade da distribuição exponencial é $f(x) = \lambda e^{-\lambda x}$ com $\lambda > 0$, $x \geq 0$ e o número $e = 2,71828...$
- A média é igual a $\mu = 1/\lambda$ e o desvio padrão igual a $\sigma = 1/\lambda$.
- A probabilidade acumulada de zero até a é $P(X \leq a) = 1 - e^{-\lambda a}$.
- A probabilidade acumulada complementar é $P(X \geq a) = e^{-\lambda a}$, ou como complemento da anterior $P(X \geq a) = 1 - P(X \leq a)$.

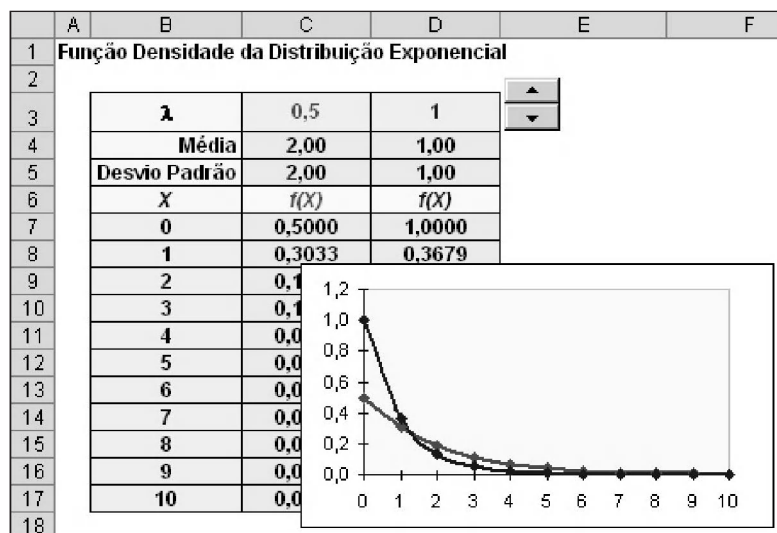


FIGURA 8.17 Função densidade da Distribuição Exponencial.

A distribuição exponencial é definida pela função densidade de probabilidade $f(x) = \lambda e^{-\lambda x}$. As características mais importantes são:

- A distribuição não é simétrica, como mostra a Figura 8.17, para dois valores do parâmetro λ . Essas distribuições foram construídas na planilha **Função Densidade DE**, incluída na pasta **Capítulo 8**, utilizando a fórmula da função densidade.
- A variável aleatória X assume somente valores positivos, ou $x \geq 0$.
- Comparando com a distribuição normal, enquanto esta é completamente definida por dois parâmetros, média e desvio padrão, a distribuição exponencial é definida por apenas um único parâmetro λ .

Acionando o controle giratório, você pode analisar o comportamento da curva da distribuição exponencial em função do parâmetro λ .

EXEMPLO 8.16

O prazo de operação de uma máquina de embalagem de frascos, sem interrupções para manutenção, tem distribuição exponencial com média de duas horas. Qual a probabilidade de essa máquina conseguir operar mais de uma hora sem interrupção?

Solução. A probabilidade da máquina de embalagem de frascos conseguir operar uma hora ou mais sem interrupção é $P(x \geq 1)$. Da distribuição exponencial acumulada complementar com média de duas horas e $\lambda = 0,50$ com a fórmula $P(x \geq 1) = e^{-0,50 \times 1} = 0,6065$ ou 60,65%.

Esse resultado também pode ser obtido com a função estatística DISTEXPON do Excel.

- **DISTEXPON(*x*; *lambda*; *cumulativo*)**

A função estatística DISTEXPON⁸ retorna a função densidade de x ou a probabilidade acumulada de zero até x , conforme o argumento *cumulativo*. Se *cumulativo* for FALSO, a função estatística DISTEXPON retornará a função densidade $f(x) = \lambda e^{-\lambda x}$. Se *cumulativo* for VERDADEIRO, a função DISTEXPON retornará a probabilidade acumulada de zero até x , ou $P(X \leq x)$ valor obtido com a fórmula $P(X \leq x) = 1 - e^{-\lambda x}$.

Registrando a fórmula $=1 - \text{DISTEXPON}(1;0,5;\text{VERDADEIRO})$ em uma célula vazia de uma planilha Excel, temos o resultado $P(X \geq 1) = 0,6065$.

- Essa fórmula calcula o complemento, pois a função DISTEXPON não dá a probabilidade requerida.

O primeiro gráfico da Figura 8.18 mostra o gráfico da distribuição exponencial $f(x)$. Esse gráfico foi construído na planilha **Modelo Distribuição Exponencial**, incluído na pasta **Capítulo 8**. Informando o valor de λ na célula C3, a planilha recalcula a tabela do intervalo B6:C16, de acordo o tipo de probabilidade escolhida na caixa de grupo:

- $f(x)$. Fornecerá o valor da função $f(x)$ para o valor x .
- *Probabilidade acumulada até x* . Fornecerá a probabilidade acumulada de zero até x .

Selecionando o botão de opção *Probabilidade acumulada até x* , a planilha recalcula a tabela do intervalo B6:C16. O segundo gráfico da Figura 8.17 mostra o gráfico da probabilidade acumulada de zero até x .

EXEMPLO 8.17

O atendente de serviços em garantia da distribuidora de carros atende a uma média de quatro clientes por hora. Qual a probabilidade de que um cliente requeira menos de 15 minutos?

Solução. Para iniciar os cálculos, é necessário ter uma única medida de tempo. Como unidade de tempo, pode-se escolher uma hora ou quinze minutos, com os mesmos resultados. Por exemplo:

- Escolhendo como unidade de tempo o intervalo de 15 minutos, a média de atendimentos será $\lambda = 1$ cliente cada 15 minutos. A probabilidade $P(X \leq 1)$ de que um cliente requeira menos de 10 minutos é 0,6321, resultado obtido com a fórmula $=\text{DISTEXPON}(1;1;\text{VERDADEIRO})$ registrada em uma célula da planilha Excel.
- Escolhendo a unidade de tempo uma hora, a média de atendimentos será $\lambda = 4$ clientes por hora, e o intervalo de 15 minutos será equivalente ao intervalo de $x = 15/60$ por hora. A probabilidade $P(X \leq 15/60)$ de que um cliente requeira menos de 15/60 hora é 0,6321, resultado obtido registrada numa célula da planilha Excel a fórmula $=\text{DISTEXPON}(10/60;3;\text{VERDADEIRO})$.

A probabilidade de que um cliente requeira menos de 15 minutos é 63,21%.

⁸ Em inglês, a função DISTEXPON é EXPONDIST.

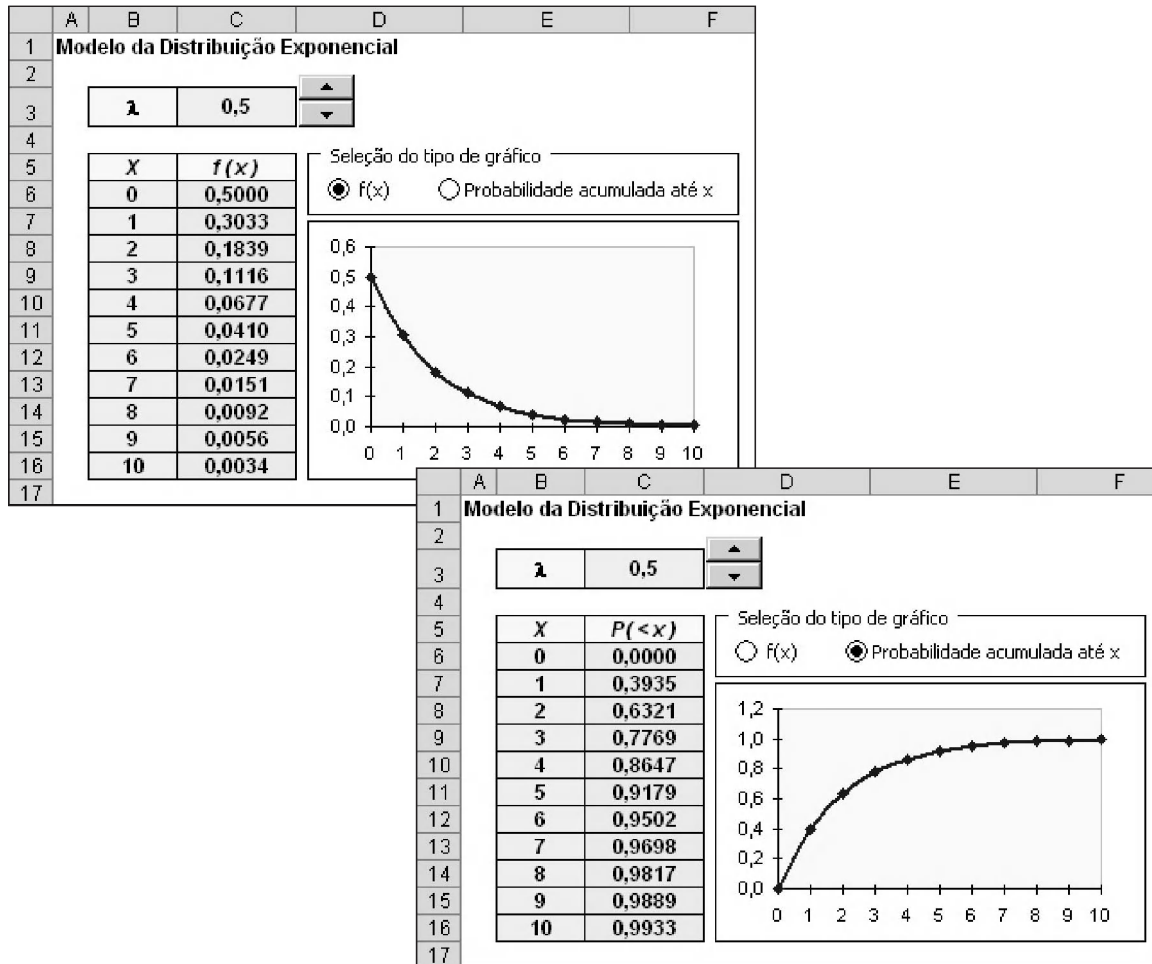


FIGURA 8.18 Gráficos da Distribuição Exponencial.

Distribuição lognormal

A distribuição normal está presente em muitas situações. Uma das aplicações que interessa neste momento é utilizada agora, porém será apresentado com mais detalhe no Capítulo 10. Por exemplo, um cabo de aço trançado utilizado num elevador, ou alguma outra aplicação de tração, é formado por muitos fios de aço que, adequadamente entrelaçados, conferem uma forte resistência ao cabo. A força ou a capacidade do cabo Y é a soma das capacidades individuais dos fios de aço y_i :

$$Y = y_1 + y_2 + y_3 + \cdots + y_i + \cdots + y_n$$

$$Y = \sum_{i=1}^n y_i$$

Se o número de fios n que formam o cabo for adequadamente grande, apesar de a distribuição da capacidade dos fios de aço não ser normal, a distribuição da capacidade do cabo será normal, pelo *Teorema Central do Limite*, que será apresentado no Capítulo 10. Esse exemplo mostra que uma variável aleatória Y , definida como a soma de n variáveis aleatórias y_i pode ser descrita com uma distribuição normal, atendendo a alguns requisitos. Outra situação frequente aparece no caso de uma variável aleatória X definida pelo produto de n variáveis aleatórias x_i , como mostra a fórmula:

$$X = x_1 \times x_2 \times x_3 \times \cdots \times x_i \times \cdots \times x_n$$

$$X = \prod_{i=1}^n x_i$$

Um exemplo prático da multiplicação de variáveis aleatórias é a determinação da taxa de retorno de um ativo durante um mês, obtido como resultado da multiplicação da variação de preços diários desse ativo. Aplicando o logaritmo natural aos dois membros dessa fórmula:

$$\ln X = \ln x_1 + \ln x_2 + \ln x_3 + \cdots + \ln x_i + \cdots + \ln x_n$$

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

Se os termos do segundo membro cumprem com os requisitos necessários, a analogia com a soma anterior é clara, podendo-se afirmar que a variável aleatória $\ln X$ tem distribuição normal. Se $Y = \ln X$, pode-se dizer que a variável aleatória Y tem distribuição normal, e a variável aleatória X tem distribuição lognormal.

A variável aleatória X com valores positivos tem *distribuição lognormal* com função densidade de probabilidade $f(x)$ se a variável aleatória Y definida como $Y = \ln(X)$ tem distribuição normal com média μ_Y e desvio padrão σ_Y .

$$\begin{cases} f(x) = \frac{1}{x\sigma_Y\sqrt{2\pi}} e^{-\frac{1}{2} \times \frac{(\ln x - \mu_Y)^2}{\sigma_Y^2}} & \text{para } x > 0 \\ f(x) = 0 & \text{para } x \leq 0 \end{cases}$$

A média e a variância de X com distribuição lognormal são:

$$\begin{aligned} \mu_X &= e^{\mu_Y + \frac{\sigma_Y^2}{2}} \\ \sigma_X^2 &= e^{2\mu_Y + \sigma_Y^2} \times (e^{\sigma_Y^2} - 1) \end{aligned}$$

Na planilha **Distribuição Lognormal**, incluída na pasta **Capítulo 8**, foram construídas duas funções densidade de probabilidade $f(x)$ da distribuição lognormal. Registrando outros valores de média e desvio padrão da distribuição normal $Y = \ln X$ no intervalo de células C4:D5, será possível analisar o comportamento dessas funções, ajustando o intervalo do eixo x das curvas com valores diferentes registrados na célula D7. No intervalo de células G4:H5, a planilha fornece a média e o desvio padrão de cada distribuição lognormal, como mostra a Figura 8.19.

A distribuição lognormal é muito utilizada em engenharia de confiabilidade⁹ para descrever falhas causadas por fadiga de material, incertezas e taxas de falhas e uma variedade de outros fenômenos. Ainda tem a propriedade de que se duas variáveis aleatórias têm distribuição lognormal, a função gerada pelo produto dessas duas variáveis também terá distribuição lognormal.¹⁰ Também é bastante utilizada

⁹ Do Houaiss, 3. capacidade de uma unidade funcional desempenhar, sem falhas ou avarias, dada tarefa sob certas condições e dentro de um período determinado.

¹⁰ Lewis E. E. – *Introduction to Reliability Engineering* – John Wiley, Second Edition, 1996.

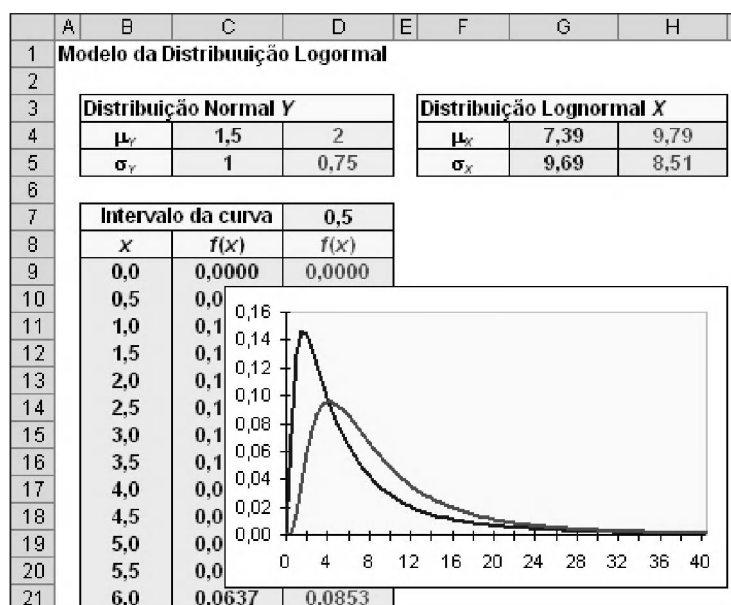


FIGURA 8.19 Função densidade da distribuição lognormal.

em opções de ativos da teoria moderna de finanças. Por exemplo, analisando a variável aleatória retorno de um investimento em ações:

- A relação entre o resgate e a aplicação pode ser maior do que um, sem nenhuma limitação até onde o próprio mercado permitir.
- Entretanto, a relação entre o resgate e a aplicação pode ser menor do que um até o limite de não resgatar nada e perder a aplicação realizada, provocando uma distribuição de retornos assimétrica.

O Excel dispõe das funções estatísticas DIST.LOGNORMAL e INVLOG para cálculos com a distribuição lognormal. Exemplos com funções estão registrados a partir da linha um da coluna J da planilha Distribuição Log-normal incluída, na pasta Capítulo 8.

DIST.LOGNORMAL(x ; *média*; *desv_padrao*)

A função estatística DIST.LOGNORMAL¹¹ retorna a probabilidade acumulada de zero até x , conhecidos os argumentos *média* e *desv_padrao*.

INVLOG(*probabilidade*; *média*; *desv_padrao*)

A função estatística INVLOG¹² retorna o valor de x para do argumento *probabilidade*, conhecidos os argumentos *média* e *desv_padrao*. A função INVLOG é a função inversa da função DIST.LOGNORMAL.

Como a distribuição lognormal é relacionada com a distribuição normal, a probabilidade acumulada de zero até x na distribuição lognormal, com parâmetros μ e σ , é igual à probabilidade acumulada de menos infinito até $\ln x$ da distribuição normal, com média μ e desvio padrão σ :

$$P(X \leq x) = \text{DIST.LOGNORMAL}(x, \mu, \sigma) = \text{DIST.NORM}\left(\frac{\ln(x) - \mu}{\sigma}\right)$$

¹¹ Em inglês, a função DIST.LOGNORMAL é LOGNORMDIST.

¹² Em inglês, a função INVLOG é LOGINV.

Essa igualdade pode ser verificada nas células L6 e L7 da planilha **Distribuição Lognormal**, como mostra a Figura 8.20.

Da mesma maneira, o cálculo de x para uma determinada probabilidade acumulada, considerando os parâmetros da distribuição lognormal, tem a seguinte equivalência com a distribuição normal:

$$x = \text{INVLOG}(p, \mu, \sigma) = e^{[\mu + \sigma \times \text{INV.NORM}(p)]}$$

Essa igualdade pode ser verificada nas células L14 e L15 da planilha **Distribuição Lognormal**, como mostra a Figura 8.20.

FIGURA 8.20 Funções DIST.LOGNORMAL e INVLOG.

	J	K	L	M	N	O
1	Função DIST.LOGNORMAL					
2						
3		μ_y	1,5			
4		σ_y	1			
5		x	4			
6		$P(x \leq 4)$	0,4547	=DIST.LOGNORMAL(L5;L3;L4)		
7		$P(x \leq 4)$	0,4547	=DIST.NORMP((LN(L5)-L3)/L4)		
8						
9	Função INVLOG					
10						
11		μ_y	1,5			
12		σ_y	1			
13		$P(x)$	0,4547			
14		x	4,00	=INVLOG(L13;L11;L12)		
15		x	4,00	=EXP(L11+L12*INV.NORMP(L13))		
16						

Problemas

Problema 1

A amostra X foi retirada de uma população com distribuição uniforme com valores máximo e mínimo iguais a 35 e 125, respectivamente. Calcule a probabilidade de um dado de uma amostra ser maior do que 50.

R: $P(X \geq 50) = P(50 \leq X \leq 125) = 0,83$

Problema 2

Continuando com o Problema 1. Calcule as probabilidades $P(X \leq 65)$; $P(55 \leq X \leq 105)$ e $P(X \leq 35)$.

R: $P(X \leq 65) = P(35 \leq X \leq 65) = 0,33$; $P(55 \leq X \leq 105) = 0,56$ e $P(X \leq 35) = 0$

Problema 3

Continuando com o Problema 1. Calcule a média e o desvio padrão dessa distribuição uniforme.

R: $\mu = 80$ e $\sigma = 25,98$

Problema 4

Repita os Problemas 1, 2 e 3 utilizando a planilha **Distribuição Uniforme**.

Problema 5

Devido à elevada volatilidade do mercado, a estimativa do preço da Ação X para os próximos 12 meses mostra que deverá ser um valor dentro do intervalo \$10 e \$50, com igual probabilidade para todos os valores de preço dentro desse intervalo. Calcule:

- a. A média e o desvio padrão do preço da ação.
- b. A probabilidade de que o valor da ação seja maior do que \$40.

R: a) $\mu=\$30$ e $\sigma=\$11,55$ b) $P(X\geq 40)=25\%$

Problema 6

A população tem distribuição normal com média igual a 20 e desvio padrão 5. Qual a probabilidade de um valor de uma amostra retirada dessa população:

- a. Ser menor ou igual a 22?
- b. Pertencer ao intervalo (19, 22)?

R: a) $P(X\leq 22)=65,54\%$ b) $P(19\leq X\leq 22)=23,47\%$

Problema 7

Seja a variável aleatória X com distribuição normal (100; 25). Qual a probabilidade de que $x>125$, $x<50$ e $75\leq x\leq 150$?

R: $P(X\geq 125)=15,87\%$ $P(X\leq 50)=2,28\%$ $P(75\leq x\leq 150)=81,86\%$

Problema 8

Os resultados da pesquisa de custo de vida de famílias de baixa renda mostram que a despesa mensal com alimentação tem distribuição normal com média \$500 e desvio padrão \$90. Sem utilizar nenhum recurso de cálculo, apenas desenhando a curva da distribuição normal, que porcentagem dessas famílias tem despesas mensais com alimentação:

- a. Até \$500?
- b. Até \$410?
- c. Maiores do que \$590?
- d. Maiores do que \$680?
- e. Menores do que \$680?

R: a) 50,00% b) 15,87% c) 15,87% d) 2,28% e) 97,72%

Problema 9

Continuando com o Problema 8, que porcentagem dessas famílias tem despesas mensais com alimentação:

- a. Entre \$400 e \$500?
- b. Entre \$200 e \$300?
- c. Entre \$490 e \$580?

R: a) 36,67% b) 1,27% c) 35,72%

Problema 10

Continuando com o Problema 8, que porcentagem dessas famílias em despesas mensais com alimentação:

- a. Superiores a \$400?
- b. Inferiores a \$400?
- c. Inferiores a \$585?
- d. Maiores a \$585?

R: a) 86,67% b) 13,33% c) 82,75% d) 17,25%

Problema 11

Se os salários anuais dos auxiliares de escritório de uma grande empresa têm distribuição normal com média \$12.500 e desvio padrão \$2.800, qual a proporção dos auxiliares de escritório que ganham:

- a. Mais que \$14.500?
- b. Menos que \$11.000?
- c. Entre \$10.000 e \$14.000?

R: a) $P(X \geq 14.500) = 23,75\%$ b) $P(X \leq 11.000) = 29,61\%$ c) $P(10.000 \leq X \leq 14.000) = 51,80\%$

Problema 12

O peso das latas de pêssego em calda tem distribuição normal com média 1.000 gramas e desvio padrão 40 gramas. Se for retirada uma lata de um lote grande de latas, calcule a probabilidade de conter:

- a. Menos que 990 gramas.
- b. Mais que 1.060 gramas.
- c. No intervalo (950, 1.050).

R: a) $P(X \leq 990) = 40,13\%$ b) $P(X \geq 1.060) = 6,68\%$ c) $P(950 \leq X \leq 1.050) = 78,87\%$

Problema 13

Os depósitos mensais na caderneta de poupança do banco têm distribuição normal com média \$500 e desvio padrão \$150. Se um depositante realizar um depósito, qual a probabilidade de que esse depósito seja:

- a. Menor ou igual a \$650?
- b. Maior ou igual a \$650?
- c. Entre \$250 e \$650?

R: a) 84,13% b) 15,87% c) 79,36%

Problema 14

A comissão anual das vendedoras domiciliares de uma importante empresa de cosméticos tem distribuição normal com média \$30.000 e desvio padrão de \$7.500. Se aleatoriamente for escolhida uma vendedora domiciliar, calcule a probabilidade de que essa vendedora domiciliar ganhe:

- a. Mais do que \$35.000 por ano.
- b. Menos do que \$12.000 por ano.
- c. Entre \$20.000 e \$30.000 por ano.

R: a) $P(X \geq 35.000) = 25,25\%$ b) $P(X \leq 12.000) = 0,82\%$ c) $P(20.000 \leq X \leq 30.000) = 40,88\%$

Problema 15

O fabricante da máquina de enchimento de refrigerantes afirma que o volume das garrafas tem média de 610 ml com desvio padrão de 8 ml. Qual a probabilidade de uma garrafa de refrigerante conter:

- a. Menos de 600 ml?
- b. Mais de 600 ml?
- c. Entre 600 e 620 ml?

R: a) 10,56% b) 89,44% c) 78,87%

Problema 16

A análise estatística de um investimento mostrou que seu resultado líquido mensal é uma variável aleatória X com média \$10.000 e desvio padrão \$4.000. Sabendo que a variável X tem distribuição normal, qual a probabilidade de o resultado líquido X ser:

- a. Menor do que \$5.000?
- b. Maior do que \$15.000?
- c. Entre \$5.000 e \$15.000?

R: a) $P(X \leq 5.000) = 10,56\%$ b) $P(X \geq 15.000) = 10,56\%$ c) $P(5.000 \leq X \leq 15.000) = 78,88\%$

Problema 17

As vendas diárias da empresa têm distribuição normal com média \$100.000 e desvio padrão \$28.000. Calcule a probabilidade que num dia qualquer as vendas sejam:

- a. Menores do que \$72.000
- b. Maiores do que \$115.000
- c. Entre \$90.000 e 110.000

R: a) 15,87% b) 29,61% c) 27,90%

Problema 18

A pesquisa de salários mensais dos *trainees* de empresas do mesmo ramo mostrou que os salários têm distribuição normal com média \$950 e desvio padrão \$185. Calcule a probabilidade de um *trainee* ganhar:

- a. Menos de \$800 por mês.
- b. Mais de \$1.200 por mês.
- c. Entre \$850 e \$1.150 por mês.

R: a) 20,87% b) 8,83% c) 56,58%

Problema 19

Continuando com o Problema 18. Pedrão, um dos *trainees*, afirma que seu salário está exatamente no limite superior do terceiro quartil de todos salários. Qual o valor desse salário?

R: \$1.074,78

Problema 20

Continuando com o Problema 18. Aproveitando a onda, Raul, outro dos *trainees*, afirma que seu salário se situa exatamente em 2 desvios padrão positivos. Qual o valor desse salário?

R: \$1.320,00

Problema 21

As vendas mensais durante os últimos 50 meses têm distribuição normal, com média igual a \$500 mil e desvio padrão \$80 mil. Se para o próximo mês a empresa estabeleceu uma meta de vendas de \$550 mil, considerando que os dados históricos se repetem no futuro próximo, calcule:

- a. A probabilidade de ficar abaixo da meta.
- b. A probabilidade de superar a meta.
- c. A probabilidade de as vendas se situarem entre 80% e 110% da média.

R: a) $P(X \leq 550) = 73,40\%$ b) $P(X \geq 550) = 26,60\%$ c) $P(400 \leq X \leq 550) = 62,84\%$

Problema 22

O saldo diário de caixa da empresa durante os últimos 400 dias tem distribuição normal, com média \$110.000 e desvio padrão \$40.000. Calcule:

- a. A probabilidade do saldo diário de caixa ser menor do que \$100.000.
- b. A probabilidade do saldo diário de caixa ser negativo.

R: a) $P(X \leq \$100.000) = 40,13\%$ b) $P(X \leq \$0) = 0,30\%$

Problema 23

Continuando com o Problema 22. Analise o resultado da probabilidade do saldo diário de caixa ser negativo.

R: O valor $P(X \leq \$0) = 0,30\%$ indica que entre um e dois dias dos 400 dias o saldo diário de caixa foi menor do que zero.

Problema 24

Continuando com o Problema 22. Aceitando a probabilidade de 10% de ter saldo diário de caixa negativo, calcule a nova média do saldo diário de caixa.

R: $\mu = \$51.200$ (Valor obtido com $Z = -1,28$)

Problema 25

Querendo aplicar os conhecimentos de Estatística, o Engenheiro de Obra manteve o seguinte diálogo com o Mestre de Obra:

Eng – Quantos azulejos, em média, um azulejista consegue assentar por dia?

Mestre – Um oficial azulejista experiente consegue assentar, em média, 8 metros quadrados de azulejos por dia de 8 horas.

Eng – Como todos os dias são diferentes, quais seriam os valores máximo e mínimo?

Mestre – Em geral, esse azulejista assenta entre 6 e 10 metros quadrados de azulejos por dia, devido aos recortes etc.

Eng – Considerando que o número de dias do mês seja 100%, que porcentagem dos dias do mês o azulejista consegue assentar entre 6 e 10 metros quadrados de azulejos por dia?

Mestre – Em geral, esse azulejista assenta entre 6 e 10 metros quadrados de azulejos por dia em 85% dos dias do mês.

Considerando que os metros quadrados de azulejos assentados por dia tenham uma distribuição normal, calcule a média e o desvio padrão dessa distribuição.

R: Média=8 e Desvio Padrão=1,39

Problema 26

O Controle de Qualidade assegura que a duração do tipo de pneu mais vendido tem distribuição normal com média 60.000 km e desvio padrão 1.300 km. Embora não haja reclamações ligadas com a duração desse tipo de pneu, para aumentar as vendas, o gerente de marketing está propondo garantir aos compradores um valor mínimo de quilometragem. Se essa campanha publicitária estima um aumento do lucro da empresa em 5%, considerando a substituição de 3% dos pneus vendidos, qual a quilometragem mínima que deve ser assegurada ao comprador?

R: Quilometragem mínima de 57.554 km

Problema 27

O gerente do controle de qualidade da empresa de refrigerantes afirma que as garrafas têm média 610 ml e desvio padrão 8 ml com distribuição normal. Qual a probabilidade:

a. De uma garrafa conter menos de 600 ml de refrigerante?

b. Mais de 600 ml de refrigerante?

c. Entre 600 e 620 ml de refrigerante?

R: a) 10,56% b) 89,44% c) 78,87%

Problema 28

O gerente do posto bancário dentro de uma empresa verificou que o saldo médio das contas correntes aumentou depois do contato pessoal realizado com grupos de clientes. Se o saldo médio tem a distribuição normal $N(\$650, \$228)$, qual a probabilidade de o saldo médio de uma conta ser:

a. Menor ou igual a \$330?

b. Maior ou igual a \$350?

R: a) $P(X \leq 330) = 66,95\%$ b) $P(X \geq 350) = 1,42\%$

Problema 29

Depois de realizar mais de mil simulações, a distribuição do valor presente líquido do projeto de investimento, simplesmente *VPL*, aproxima-se de uma normal com média \$245.000 e desvio padrão \$83.500. Qual a probabilidade de que o *VPL* seja:

- a. Negativo?
- b. Menor do que \$25.000?

R: a) $P(VPL \leq 0) = 0,17\%$ b) $P(VPL \leq \$25.000) = 0,42\%$

Problema 30

Suponha que a estatura de uma certa população de pessoas seja aproximadamente normal, com média de 177 centímetros e desvio padrão de 7,8 centímetros. Qual a probabilidade de que a estatura de uma pessoa seja:

- a. Maior do que 175 centímetros?
- b. Maior do que 180 e menor do que 190 centímetros?
- c. Menor do que 170 centímetros?

R: a) 60,12% b) 30,25% c) 18,47%

Problema 31

O *QI* (quociente de inteligência) dos estudantes de uma escola tem uma distribuição aproximadamente normal, com média 106 e variância de 260. Qual a probabilidade de que o *QI* de um estudante seja:

- a. Menor que 99?
- b. Maior que 128?
- c. Entre 95 e 117?

R: a) 33,21% b) 66,79% c) 50,49%

Problema 32

Para ser membro da Mensa International (mesa, em latim), por exemplo, o candidato precisa obter uma pontuação que o encaixe entre os 2% mais “inteligentes” do mundo – cerca de 132 pontos no teste de *QI*. A média da população, em geral, varia em torno de 100.¹³ Considerando que o *QI* seja uma variável aleatória com distribuição normal, qual o desvio padrão?

R: $\sigma_{QI} = 15,58$

Problema 33

Continuando com o Problema 32. Qual o *QI* de um participante que pertence ao seletor grupo dos 0,1% mais “inteligentes” do mundo?

R: $QI = 148,15$

Problema 34

Continuando com o Problema 32. Qual o *QI* de um participante que pertence ao grupo dos:

- a. Um em trinta mil mais “inteligentes” do mundo?
- b. Um em um milhão mais “inteligentes” do mundo?

R: a) $QI = 162,13$ b) $QI = 174,06$

Problema 35

Quando uma empresa afirma que seu processo atende à especificação Seis Sigma ela quer informar que a variabilidade de seu processo, produto ou serviço é menor do que 3,4 falhas por milhão de oportuni-

¹³ Do artigo de Bel Moherdau “Gênios de carteirinha” publicado na *Revista da Folha* em 16/01/2000.

dades. Verifique que esse limite de falhas por milhão equivale a 99,99666% de perfeição. *Pista:* tenha em mente que as falhas ocorrem em ambas as caudas da distribuição normal e a percentagem de perfeição se distribui por igual ao redor da média da distribuição normal.

Problema 36

Seja a distribuição exponencial com média $\lambda=12$, qual a probabilidade de que a próxima chegada ocorra em:

- a. Menos do que $X=0,2$?
- b. Mais do que $X=0,2$?
- c. Entre 0,1 e 0,2?

R: a) 90,93% b) 9,07% c) 21,05%

Problema 37

Em dias normais e durante o expediente normal, o caixa automático do banco recebe 18 clientes por hora. A partir do momento da chegada de um cliente, qual a probabilidade de que o próximo cliente chegue dentro dos próximos quatro minutos?

R: $\lambda=18$, $x=4/60$ e $P(x)=69,88\%$

Problema 38

Nos domingos à tarde, os carros chegam ao pedágio da estrada com uma média de 120 carros por hora. A partir do momento da chegada de um carro, qual a probabilidade de que o próximo carro chegue:

- a. Dentro de um minuto?
- b. Dentro de dois minutos?
- c. Dentro de meio minuto?

R: $\lambda=120$, $x=1/60$ a) $P(x)=86,47\%$ b) $P(x)=98,17\%$ c) $P(x)=63,21\%$

Apêndice 1

Geração de números aleatórios

Desde o início do livro, tivemos oportunidade de gerar e utilizar números aleatórios em várias aplicações. No Capítulo 1, foram definidos e gerados dígitos e números aleatórios utilizando a função ALEATÓRIO e a função ALEATÓRIOENTRE do Excel. Ainda nesse capítulo, foi realizada a simulação da retirada de uma bola com reposição de uma urna com dez bolas que apresentam uma distribuição uniforme discreta, utilizando a ferramenta *Amostragem*, e o modelo para amostragem sem reposição. Na simulação do lançamento da moeda do Capítulo 5, foi utilizada a ferramenta de análise *Geração de número aleatório*, com o tipo de distribuição discreta para gerar os números aleatórios 0 e 1 com probabilidade de 50% para cada um. Na simulação realizada para extrair *Notas* de uma urna no Capítulo 7, foi utilizado o procedimento de geração de números aleatórios de uma distribuição de frequências uniforme.

A geração de números aleatórios é utilizada nos processos de simulação com diversas distribuições teóricas disponíveis e algumas definidas para determinada situação. Os números aleatórios gerados são utilizados para selecionar valores de uma variável aleatória a partir de sua distribuição de frequências acumuladas, como mostrado a seguir com duas distribuições conhecidas.

Distribuição Normal. A Figura 8.21 mostra a curva da distribuição de frequências acumuladas da distribuição normal $N(40, 15)$ construída na planilha *Geração Dígitos Aleatórios*, incluída na pasta *Capítulo 8*.

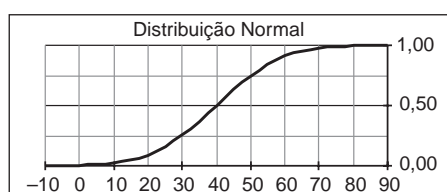


FIGURA 8.21
Distribuição normal acumulada.

Um número aleatório gerado no intervalo $(0, 1)$ define uma reta horizontal que na intersecção com a curva determina o valor da variável correspondente no eixo de abscissas; por exemplo, o número aleatório 0,25 define o valor da variável 29 (valor aproximado pela baixa resolução do gráfico).

Distribuição Uniforme. A Figura 8.22 mostra a curva da distribuição de frequências acumuladas da distribuição uniforme no intervalo $(-10, 90)$ construída na planilha *Geração Dígitos Aleatórios*, incluída na pasta *Capítulo 8*.

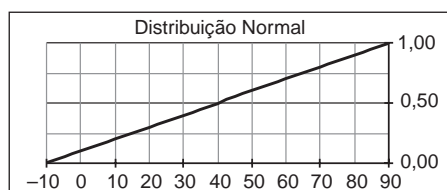


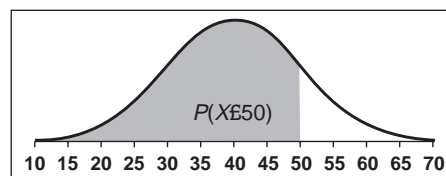
FIGURA 8.22
Distribuição uniforme acumulada.

Um número aleatório gerado no intervalo (0, 1) da distribuição uniforme define uma reta horizontal que, na intersecção com a curva, determina o valor da variável correspondente no eixo de abscissas; por exemplo, o número aleatório 0,25 define o valor da variável 15 (valor aproximado pela baixa resolução do gráfico). Como se pode ver nas figuras 8.21 e 8.22, o eixo de ordenadas que representa o valor da probabilidade acumulada é o mesmo para as duas distribuições, pois os valores da distribuição de frequências acumuladas relativas estarão sempre no intervalo (0, 1); entretanto, os correspondentes valores das variáveis são diferentes.

A Figura 8.23 mostra as duas amostragens no mesmo gráfico depois de gerar o número aleatório 0,75 com a função ALEATÓRIO, por exemplo. Das distribuições de frequências acumuladas, obtém-se o valor 51 na distribuição normal e o valor 66 na distribuição uniforme, ambos valores aproximados pela baixa resolução do gráfico, situação que não ocorre na prática, pois é utilizado o valor gerado sem passar pelo gráfico.

FIGURA 8.23

Distribuições normal e uniforme acumuladas.



Ferramenta de análise *Geração de Número Aleatório*

A ferramenta de análise *Geração de número aleatório*¹⁴ gera dígitos aleatórios extraídos de distribuições de frequências teóricas, por exemplo, a distribuição normal, a distribuição uniforme, a distribuição binomial etc., ou de distribuições de frequências definidas pelo usuário. Essa ferramenta retorna números aleatórios gerados da forma como foi apresentado anteriormente. Para compreender a utilização dessa ferramenta de análise, serão gerados números aleatórios das distribuições normal e uniforme.

Distribuição normal

A geração de números aleatórios com a distribuição normal, utilizando a ferramenta *Geração de número aleatório*, foi realizada na planilha *Geração Números Aleatórios*, incluída na pasta **Capítulo 8**:

- Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis.
- Escolhendo a ferramenta **Geração de número aleatório** e depois pressionando o botão OK, será exibida a caixa de diálogo correspondente, mostrada na Figura 8.24, depois de selecionadas algumas opções.
 - Pressionando o botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Geração de número aleatório* pertencente à *Ajuda do Excel*.

As informações que devem ser registradas na caixa de diálogo da ferramenta *Geração de número aleatório* são:

- **Número de variáveis.** Informar o número de colunas de valores que se deseja na tabela de saída. Neste exemplo informamos 2, pois queremos gerar duas séries de dígitos aleatórios com a distribuição normal.

¹⁴ Em inglês, a ferramenta *Geração de Número Aleatório* é *Random Number Generation*.

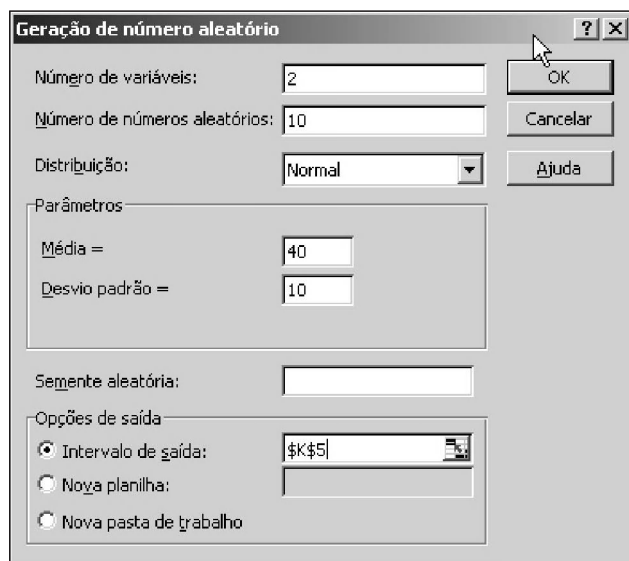


FIGURA 8.24 Geração de número aleatório, distribuição normal.

- **Número de números aleatórios.** Informar o número de valores desejados em cada coluna definida. Neste exemplo informamos 10, pois queremos gerar duas séries com 10 dígitos aleatórios cada uma.
- **Distribuição.** Seleccionamos Normal. Neste caso, o Excel apresenta o quadro **Parâmetros** no qual foi informado o valor da **Média** = 40 e o valor do **Desvio padrão** = 10.
- **Semente aleatória.** Pode-se informar um valor a partir do qual os dígitos aleatórios serão gerados. A vantagem deste procedimento é que, em uma nova geração de dígitos aleatórios, utilizando o mesmo valor de **Semente**, o Excel gerará os mesmos números aleatórios anteriores. Neste exemplo, deixamos em branco.

Na primeira parte do quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta *Geração de número aleatório* registrará os resultados. Há três alternativas para o caso de não informar esse endereço, identificadas por três *botões de opção* que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, neste caso K5. O Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
 - Se não for informado nenhum endereço, a ferramenta inserirá uma nova planilha com o nome **Plan**, seguido de um número sequencial, por exemplo, escolhendo essa alternativa na pasta **Capítulo 2**, a ferramenta inserirá a planilha **Plan1**.
 - Há a alternativa de informar o nome da planilha na caixa desta alternativa; por exemplo, registrando o nome *Teste*, a ferramenta inserirá na mesma pasta uma nova planilha com o nome **Teste**.
- **Nova pasta de trabalho.** Os resultados serão apresentados numa nova pasta e a partir da célula A1 da planilha **Plan1**.

Completadas as informações, depois de pressionar o botão **OK**, o Excel registra os valores solicitados a partir da célula K5, uma tabela com duas séries com dez números aleatórios cada uma, Figura 8.26.

Distribuição uniforme

A partir da célula N5 da planilha **Geração Números Aleatórios**, foram registradas duas amostras geradas com a distribuição uniforme utilizando a ferramenta de análise *Geração de número aleatório*, Figura 8.25. Completadas as informações, depois de pressionar o botão **OK**, o Excel registra os valores solicitados a partir da célula N5, uma tabela com duas séries com dez números aleatórios cada uma, Figura 8.26.

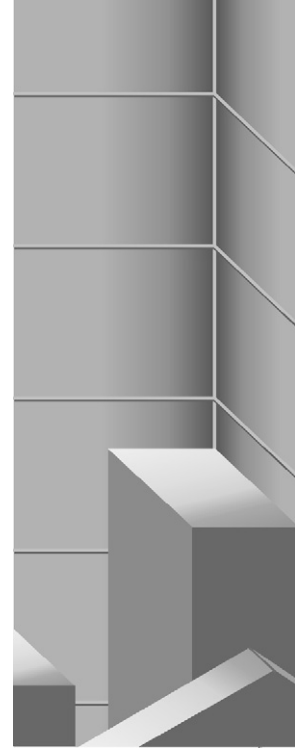
FIGURA 8.25 Geração de número aleatório, distribuição uniforme.

FIGURA 8.26 Números aleatórios gerados com a ferramenta de análise.

	J	K	L	M	N	O
1	Geração de Números Aleatórios					
2						
3		Distribuição Normal			Distribuição Uniforme	
4		X1	X2		Y1	Y2
5		36,9976784	27,2231683		79,8770104	2,34168523
6		42,4425731	52,7647354		62,2800378	59,5791498
7		51,9835022	57,331331		8,96420179	69,1314432
8		18,1641236	37,6581876		17,1370586	77,1974853
9		50,9502253	29,1329935		65,981933	-0,8536027
10		33,0979584	23,0956767		31,892758	-5,5137791
11		21,5308911	30,223705		1,50852992	14,1218299
12		32,2649295	18,8206878		70,9045686	79,7488327
13		34,3207513	35,9595243		76,5596484	83,4415723
14		41,3485305	36,3450705		18,8460952	50,4815821
15						

Capítulo 9

COMBINAÇÃO LINEAR DE VARIÁVEIS ALEATÓRIAS



Até este momento, analisamos variáveis isoladas ou a relação entre duas delas. Há uma aplicação importante quando uma ou mais variáveis aleatórias são combinadas para formar outra variável aleatória. Essa combinação pode ser realizada com as operações de soma, diferença, produtos e divisões. Neste capítulo, apresentaremos a transformação linear de uma variável em outra, e a combinação de duas ou mais variáveis utilizando as operações de soma e diferença.

Transformação linear

Uma variável aleatória pode ser registrada em diferentes unidades de medida, por exemplo, a variável lucro mensal de uma empresa pode ser registrada em real, em dólar americano, em marco alemão etc. Da mesma maneira, as distâncias podem ser medidas em quilômetros ou em milhas, a temperatura em graus Celsius ou em graus Fahrenheit etc. A conversão dos valores de uma variável em uma unidade em outra unidade de medida é denominada *transformação linear de uma variável*.

Transformação Linear de uma Variável

A *transformação linear* modifica a variável $X = (X_1, X_2, \dots, X_i, \dots, X_n)$ na variável $Y = (Y_1, Y_2, \dots, Y_i, \dots, Y_n)$, de forma que cada valor de X tenha seu valor correspondente de Y por meio da função $Y_i = a + bX_i$, sendo a e b números reais.

Analisando a fórmula da transformação linear podemos observar que:

- A adição do coeficiente a desloca todos os valores de X na mesma quantidade a , seja aumentando-os ou diminuindo-os.
- A multiplicação pelo coeficiente b muda o tamanho de todos os valores de X , amplificando-os ou diminuindo-os.

EXEMPLO 9.1

A coluna B da planilha a seguir registra o lucro mensal em USD\$ de uma empresa americana durante o ano 2004. Para comparar o resultado com uma empresa brasileira equivalente, o lucro mensal daquela empresa foi transformado R\$ utilizando a taxa de conversão 2,96 R\$/USD\$, como mostra a coluna C da tabela construída na planilha **Exemplo 9.1**, incluída na pasta **Capítulo 9**. Nessa tabela, foram adicionados os resultados das médias e dos desvios padrão das duas séries de valores monetários.

	A	B	C	D
1	Exemplo 9.1			
2				
3			Taxa conversão	R\$ 2,96
4		Mês	Lucro em USD\$	Lucro em R\$
5		jan/04	\$ 1.200	R\$ 3.552
6		fev/04	\$ 1.300	R\$ 3.848
7		mar/04	\$ 1.500	R\$ 4.440
8		abr/04	\$ 1.450	R\$ 4.292
9		mai/04	\$ 1.280	R\$ 3.789
10		jun/04	\$ 1.420	R\$ 4.203
11		jul/04	\$ 1.560	R\$ 4.618
12		ago/04	\$ 1.320	R\$ 3.907
13		set/04	\$ 1.500	R\$ 4.440
14		out/04	\$ 1.600	R\$ 4.736
15		nov/04	\$ 1.650	R\$ 4.884
16		dez/04	\$ 1.780	R\$ 5.269
17		Média	\$ 1.463	R\$ 4.331
18		DP	\$ 170	R\$ 503
19				

Analisando os resultados das médias das duas séries de valores monetários, verifica-se que a relação entre essas medidas é a própria taxa de conversão \$2,96 R\$/USD\$. O mesmo ocorre com os resultados dos desvios padrão. No cálculo do *Lucro em R\$* a partir do *Lucro em USD\$* do Exemplo 9.1, foi aplicada a transformação linear $\text{Lucro (R\$)} = 2,96 \times \text{Lucro (USD \$)}$. Nesse caso, os coeficientes da fórmula $Y_i = a + bX_i$ são $a=0$ e $b=2,96$.

EXEMPLO 9.2

Na coluna B da tabela a seguir, foram registradas as temperaturas medidas em graus Celsius C na cidade de São Paulo durante um dia de verão.

	A	B	C	D
1	Exemplo 9.2			
2				
3		Hora	C°	F°
4		10:00	28,5	83,3
5		11:00	29,3	84,7
6		12:00	32,5	90,5
7		13:00	33,6	92,5
8		14:00	32,0	89,6
9		15:00	31,0	87,8
10		16:00	29,8	85,6
11		Média	30,96	87,72
12		DP	1,85	3,33
13				

Para incluir esses registros em um relatório em inglês, é necessário transformá-los em graus Fahrenheit F , utilizando a transformação linear $F_i = 32 + 1,80 \times C_i$, como mostra a coluna C da tabela anterior, construída na planilha **Exemplo 9.2**, incluída na pasta **Capítulo 9**.

Analisando a relação entre as duas médias, verificamos que esse resultado não é nenhum dos coeficientes utilizados na transformação. Entretanto, a relação entre os dois desvios padrão é o coeficiente 1,80, utilizado na transformação. No Exemplo 9.2, foi aplicada a transformação linear: $F_i = 32 + 1,80 \times C_i$ com $a=32$ e $b=1,80$.

Consequências da transformação linear

De forma geral, os valores da média μ_Y e do desvio padrão σ_Y da variável Y da transformação linear $Y_i = a + X_i$ são obtidos da média μ_X e do desvio padrão σ_X da variável X com as seguintes expressões

$$\begin{cases} \mu_Y = a + b + \mu_X \\ \sigma_Y = |b| \times \sigma_X \end{cases}$$

- No Exemplo 9.1, foi aplicada a transformação linear $Y_i = bX_i$, com $a=0$. Nesse caso, a média μ_Y e o desvio padrão σ_Y da variável Y são obtidos da média μ_X e do desvio padrão σ_X da variável X com as seguintes relações $\begin{cases} \mu_Y = b + \mu_X \\ \sigma_Y = |b| \times \sigma_X \end{cases}$.
- No Exemplo 9.2, foi aplicada a transformação linear $Y_i = a + bX_i$, com $a=32$ e $b=1,8$. Nesse caso, os valores da média μ_Y e do desvio padrão σ_Y da variável Y são obtidos da média μ_X e do desvio padrão σ_X da variável X com as seguintes relações $\begin{cases} \mu_Y = a + b \times \mu_X \\ \sigma_Y = |b| \times \sigma_X \end{cases}$.

Resumindo:

- O coeficiente a muda a média, porém não muda o desvio padrão.
- O coeficiente b muda a média e o desvio padrão, sendo que no desvio padrão deve ser utilizado o valor absoluto de b , pois o efeito não depende do sinal desse coeficiente.
- A transformação linear não muda a forma da distribuição da variável. Se a distribuição de X for simétrica, a distribuição continuará simétrica e, da mesma maneira, a inclinação positiva da distribuição continuará positiva depois da transformação linear.

Combinação linear

Uma aplicação importante é a combinação de uma ou mais variáveis aleatórias cujo resultado gera outra variável aleatória. Nesta parte, analisaremos a combinação de variáveis aleatórias com as operações de soma e diferença. A combinação linear de variáveis aleatórias gera a função de variáveis aleatórias $H = f(X_1, X_2, \dots, X_n)$, que é uma nova variável aleatória.¹

A função $H = a_1X_1 + a_2X_2 + \dots + a_nX_n$ é uma variável aleatória formada pela combinação linear das variáveis aleatórias X_1, X_2, \dots, X_n e as constantes a_1, a_2, \dots, a_n .

O objetivo é obter a média, a variância e o desvio padrão de H , função de n amostras ou variáveis aleatórias, aplicando um dos dois seguintes procedimentos:²

- Utilizando os próprios valores das variáveis aleatórias.
- Utilizando as medidas estatísticas média e desvio padrão, ou variância, das próprias variáveis ou amostras e as covariâncias, ou coeficientes de correlação, das combinações de variáveis aleatórias ou amostras tomadas duas a duas.

¹ Este assunto pode ser deixado para uma leitura posterior.

² As variáveis aleatórias têm a mesma unidade, por exemplo, cm, % etc.

Utilizando os valores das amostras ou variáveis aleatórias

Os valores da média, da variância e do desvio padrão da variável H podem ser obtidos combinando os valores das variáveis aleatórias como mostram estes exemplos.

EXEMPLO 9.3

Sejam as variáveis X_1 e X_2 cujos valores estão registrados na planilha **Exemplo 9.3**, incluída na pasta **Capítulo 9**. Combine as variáveis X_1 e X_2 para obter a função $H = f(X_1, X_2) = 5X_1 + 2X_2$ e analise os resultados obtidos.

Solução. Analisando os resultados da tabela, pode-se concluir que:

- O desvio padrão da variável X_1 é menor do que o da variável X_2 .
- O coeficiente de correlação $r = -0,06$ mostra que não há correlação entre as variáveis X_1 e X_2 .
- A média da combinação linear ou variável H é a média ponderada das médias das duas variáveis $\mu_H = 5 \times 6 + 2 \times 14,5 = 59$.
- O desvio padrão de H não mantém nenhuma relação linear com os desvios padrão das variáveis X_1 e X_2 .

	A	B	C	D	E
1	Exemplo 9.3				
2					
3			X ₁	X ₂	5×X ₁ +2×X ₂
4			4	12	44
5			6	15	60
6			5	18	61
7			7	17	69
8			6	14	58
9			5	16	57
10			8	12	64
11			6	15	60
12			8	14	68
13			3	13	41
14			6	15	60
15			8	13	66
16		Média	6,0	14,5	59,0
17		Variância	2,33	3,25	68,00
18		DP	1,53	1,80	8,25
19		r	-0,06		
20					

EXEMPLO 9.4

Os retornos anuais durante os últimos seis anos da Ação A e da Ação B negociadas na Bolsa de Valores estão registrados na tabela seguinte. Realize a análise estatística dos retornos das duas ações.

	A	B	C	D
1	Exemplo 9.4			
2				
3		Ano	Ação A	Ação B
4		1999	9,0%	12,0%
5		2000	10,0%	10,5%
6		2001	12,0%	9,0%
7		2002	10,5%	11,0%
8		2003	9,5%	12,5%
9		2004	11,0%	10,0%
10		Média	10,3%	10,8%
11		DP	1,08%	1,29%
12		CV	0,10	0,12
13		r	-0,92	
14				

Solução. As medidas estatísticas calculadas na planilha **Exemplo 9.4**, incluída na pasta **Capítulo 9**, mostram que:

- O desvio padrão da Ação A é menor do que o da Ação B, e o risco da Ação A é menor do que o da Ação B, pois seu coeficiente de variação é menor.
- Os retornos das duas ações mantêm uma correlação fortemente negativa. Os retornos das ações variam em direções opostas, se o retorno da Ação A aumenta, o da Ação B diminui, e vice-versa.

EXEMPLO 9.5

Continuando com o Exemplo 9.4. Suponha que no início de 1999 tenha sido formada uma carteira de investimento dividindo o capital entre as duas ações, investindo 50% na Ação A e os outros 50% na Ação B. Analise o retorno da carteira e compare com os resultados das duas ações.

Solução. O intervalo E5:E10 da planilha **Exemplo 9.5**, incluída na pasta **Capítulo 9**, registra os retornos da carteira formada com 50% do capital investido na Ação A e os outros 50% na Ação B, e o intervalo E11:E13 calcula as medidas estatísticas da carteira.

	A	B	C	D	E
1	Exemplo 9.5				
2					
3			Ação A	Ação B	
4		Ano	50%	50%	Carteira
5		1999	9,0%	12,0%	10,5%
6		2000	10,0%	10,5%	10,3%
7		2001	12,0%	9,0%	10,5%
8		2002	10,5%	11,0%	10,8%
9		2003	9,5%	12,5%	11,0%
10		2004	11,0%	10,0%	10,5%
11		Média	10,3%	10,8%	10,6%
12		DP	1,08%	1,29%	0,26%
13		CV	0,10	0,12	0,02
14		r	-0,92		
15					

Os retornos anuais da carteira de investimento são a média ponderada dos retornos das ações; por exemplo, $R_{1999} = 0,50 \times 9\% + 0,50 \times 12\% = 10,5\%$ é o retorno do ano de 1999. Dos resultados registrados na tabela, pode-se concluir que a média dos retornos da carteira durante os seis anos é igual à média ponderada das médias dos retornos das duas ações, como mostra a fórmula:

$$R = 0,50 \times 10,333\% + 0,50 \times 10,833\% = 10,6\%$$

O desvio padrão da carteira não mantém nenhuma relação linear com os desvios padrão dos retornos das ações. Ademais, o desvio padrão e o coeficiente de variação dos retornos da carteira são significativamente menores do que os das ações, mostrando que a variabilidade dos retornos da carteira é bem menor do que a das duas ações. De outra maneira, o risco de investir na carteira 50% na Ação A e 50% na Ação B, teria sido bem menor do que o de investir 100% em uma das duas ações; a diversificação diminui o risco do investimento.

Utilizando as medidas estatísticas das amostras ou variáveis aleatórias

Na seção anterior, o valor da média e o valor do desvio padrão da função H foram obtidos combinando os pares de valores das duas variáveis aleatórias, procedimento que pode ser estendido para n variáveis. Nesta parte, o valor da média e o valor do desvio padrão da função H serão obtidos utilizando somente a média, o desvio padrão, ou variância, e a covariância, ou o coeficiente de correlação, de cada uma das variáveis aleatórias que participam da combinação linear, procedimento que poderá ser estendido para

n variáveis aplicando as propriedades da combinação linear.³ O primeiro passo é definir a função H em função das variáveis aleatórias do problema, por exemplo:

- Para medir o erro cometido na medição do perímetro de um componente em uma montagem de dispositivos de precisão, será necessário definir a função H do perímetro.
- Para um projeto de investimento, será necessário definir a função H do valor presente líquido do fluxo de caixa do investimento.
- Para avaliar uma carteira de investimento, será necessário estabelecer a função H do retorno da carteira.

Procedimento para Determinar a Variável Aleatória H

- Definir a função H do problema.
- Definir a expressão da média ou valor esperado de H .
- Definir a expressão da variância de H e, depois, do desvio padrão.

Por que é mencionada a variância se o desvio padrão é mais adequado para medir dispersão? Lembre-se de que:

- O desvio padrão é a raiz quadrada positiva da variância.
- Operando com o desvio padrão, será necessário carregar sua expressão com a raiz quadrada proveniente de sua definição, complicando sem necessidade os desenvolvimentos e os cálculos matemáticos.

Combinação linear de duas variáveis aleatórias

A combinação linear das variáveis aleatórias X_1 e X_2 gera a função H , identificada pela sua média ou valor esperado e pela sua variância, valores que deverão ser obtidos das propriedades das variáveis aleatórias.

Combinação Linear de Duas Variáveis Aleatórias

- A função H é a combinação linear $H = a_1X_1 + a_2X_2$.
 - As constantes a_1 e a_2 podem assumir qualquer valor do conjunto dos números reais.
 - As variáveis aleatórias X_1 e X_2 têm médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2 .
- A média ou valor esperado de H é obtida com a expressão:⁴

$$\mu_H = \mu_{a_1X_1 + a_2X_2} = a_1\mu_1 + a_2\mu_2$$

- A variância da variável H é:⁵

$$\sigma_H^2 = \sigma_{a_1X_1 + a_2X_2}^2$$

$$\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 + 2 \cdot a_1 \cdot a_2 \cdot \sigma_{12}$$

³ São apresentadas e demonstradas no Apêndice 1 deste capítulo.

⁴ As demonstrações da fórmula da média e da variância de H foram realizadas no Apêndice 1 deste capítulo.

⁵ O símbolo σ_{12} é a covariância das variáveis X_1 e X_2 .

Sendo σ_{12} a covariância das variáveis X_1 e X_2 . A variância de H também pode ser calculada com o coeficiente de correlação r_{12} :

$$\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 + 2 \cdot a_1 \cdot a_2 \cdot r_{12} \cdot \sigma_1 \cdot \sigma_2$$

- O desvio padrão de H é obtido com: $\sigma_H = +\sqrt{\sigma_H^2}$.

EXEMPLO 9.6

As médias e as variâncias das variáveis X_1 e X_2 estão registradas na tabela seguinte. Calcule a média, a variância e o desvio padrão da variável aleatória $H=0,5 \cdot X_1 + 0,5 \cdot X_2$, sabendo que o coeficiente de correlação das variáveis X_1 e X_2 é $r_{12}=-0,20$.

	X_1	X_2
Média	0,10	0,20
Variância	0,065	0,088
Desvio padrão	0,2550	0,2966

Solução. As medidas estatísticas de H são:

- A média de H é igual a 0,15, resultado obtido com a fórmula:

$$\mu_H = 0,5 \times 0,10 + 0,5 \times 0,20 = 0,15$$

- A variância de H é igual a 0,03075, resultado obtido com a fórmula que utiliza o coeficiente de correlação:

$$\sigma_H^2 = 0,5^2 \times 0,065 + 0,5^2 \times 0,088 + 2 \times 0,5 \times 0,5 \times 0,255 \times 0,2966 \times (-0,20) = 0,03069$$

- O desvio padrão de H é igual a 0,1752, resultado obtido com:

$$\sigma_H = +\sqrt{0,03069} = 0,1752$$

A planilha Combinação duas VA, incluída na pasta Capítulo 9, ajuda no cálculo dos parâmetros da função H com duas variáveis aleatórias, como mostra a Figura 9.1, resolvendo o Exemplo 9.6.

	A	B	C	D	E	F	G
1	Combinação de duas Variáveis Aleatórias						
2							
3		Dados				Resultados	
4			X_1	X_2			H
5		Peso	0,50	0,50		Média	0,1500
6		Média	0,1000	0,2000		Variância	0,030686
7		DP	0,2550	0,2966		DP	0,1752
8		r	-0,2000				
9							

FIGURA 9.1 Modelo para combinação de duas VA.

Análise de resultados importantes

Pela simples inspeção das fórmulas da combinação linear de duas variáveis aleatórias, pode-se verificar que o coeficiente de correlação, ou a covariância, somente impacta a variância e, consequentemente, o

desvio padrão de H . Há importantes valores extremos do coeficiente de correlação que interessam analisar, pois têm aplicações muito úteis. A seguir, mostramos alguns resultados importante do Exemplo 9.6, utilizando o *modelo* da planilha **Combinação duas VA**.

- **Coeficiente de correlação $r=0$.** Neste caso, a variância da função H passa a ser $\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2$. Se X_1 e X_2 forem variáveis aleatórias independentes, a covariância dessas variáveis será igual a zero e, consequentemente, o coeficiente de correlação também será zero. Pode-se ver que as variâncias se somam, com um fator em cada uma delas.
 - No caso do Exemplo 9.6, a variância de H é 0,0385, resultado obtido com a fórmula $\sigma_H^2 = 0,5^2 \times 0,065 + 0,5^2 \times 0,088 = 0,03825$. Neste caso, a variância de H aumentou, pois no cálculo anterior, o coeficiente de correlação fracamente negativo diminuiu a variância de H . O mesmo raciocínio pode ser repetido com o desvio padrão de H .
- **Coeficiente de correlação $r=1$.** As duas variáveis estão perfeitamente correlacionadas em sentido positivo, e a variância da função H passa a ser $\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 + 2 \cdot a_1 \cdot a_2 \cdot \sigma_1 \cdot \sigma_2$. Essa fórmula pode ser transformada em $\sigma_H^2 = (a_1 \cdot \sigma_1 + a_2 \cdot \sigma_2)^2$, mostrando que a variância de H é, salvo um fator, o quadrado da soma dos desvios padrão das variáveis combinadas.
 - No caso do Exemplo 9.6, a variância é igual a 0,0761, resultado maior do que o anterior, pois agora foi adicionado um fator positivo.
- **Coeficiente de correlação $r=-1$.** As duas variáveis estão perfeitamente correlacionadas em sentido negativo, e a variância da função H passa a ser $\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 - 2 \cdot a_1 \cdot a_2 \cdot \sigma_1 \cdot \sigma_2$. Essa fórmula pode ser transformada em $\sigma_H^2 = (a_1 \cdot \sigma_1 - a_2 \cdot \sigma_2)^2$, mostrando que a variância de H é, salvo um fator, o quadrado da diferença dos desvios padrão das variáveis combinadas.
 - No caso do Exemplo 9.6, a variância é igual a 0,00043, resultado bem menor do que os anteriores, pois foi somado um forte fator negativo.
 - Ativos financeiros com coeficiente de correlação fortemente negativo diminuem o risco de uma carteira de investimento, pois reduzem a variabilidade dos retornos.
 - Montagem de duas peças que exijam, por exemplo, uma espessura tão uniforme quanto possível poderia ser produzida realizando uma combinação seletiva das peças com $r=-1$, que compensem a soma de espessuras e reduzam a variação dos lotes produzidos.⁶

Combinação linear de uma variável aleatória

A transformação linear apresentada no início do capítulo é uma combinação linear com uma única variável aleatória X cujos resultados podem ser obtidos da combinação linear com duas variáveis.

Combinação Linear de Uma Variável Aleatória

- A função H é a combinação linear $H = a_1 + a_2 X$.
 - As constantes a_1 e a_2 podem assumir qualquer valor do conjunto dos números reais.
 - A variável aleatória X tem média μ_X e desvio padrão σ_X .
- A média da variável H é $\mu_H = \mu_{a_1+a_2X} = a_1 + a_2\mu_X$
- A variância da variável H é $\sigma_H^2 = \sigma_{a_1+a_2X}^2 = a_2^2 \cdot \sigma_X^2$
- O desvio padrão de H é $\sigma_H = +\sqrt{\sigma_H^2} = |a_2| \sigma_X$

6 Kume H. – *Métodos Estatísticos para Melhoria da Qualidade* – Editora Gente, 1993.

EXEMPLO 9.7

Calcule a média e o desvio padrão da variável aleatória $H=10+5.X$, sabendo que a média e a variância de X são, respectivamente, 0,124 e 0,0454.

Solução. Os resultados solicitados são:

$$\mu_H = 10 + 5 \times 0,124 = 10,62$$

$$\sigma_H^2 = \text{Var}(10 + 5X) = 5^2 \times 0,0454 = 1,135$$

$$\sigma_H = \sqrt{1,135} = 1,065$$

EXEMPLO 9.8

A quantidade de estações gráficas vendidas mensalmente por José é normalmente distribuída com média 0,80 estação por mês e variância 0,25. Mensalmente, José recebe \$5.000 mais \$10.000 por estação gráfica vendida. Calcule a remuneração média por mês e o desvio padrão correspondente.

Solução. A expressão da remuneração mensal do vendedor de estações gráficas é: $H = \$5.000 + \$10.000 \times n$, sendo n a variável aleatória quantidade de estações gráficas vendidas. José recebe \$13.000 por mês com desvio padrão de \$5.000, resultados obtidos das seguintes fórmulas:

$$\mu_H = \$5.000 + \$10.000 \times 0,8 = \$13.000$$

$$\sigma_H^2 = \$10.000^2 \times 0,25 = 25.000.000$$

$$\sigma_H = \sqrt{25.000.000} = \$5.000$$

Como a venda de estações gráficas é normalmente distribuída, observe que, em 68% dos meses, o vendedor receberá entre \$8.000 e \$18.000.

Combinação linear de n variáveis aleatórias

Da mesma forma como foi feito nas combinações com uma e duas variáveis aleatórias, o procedimento pode ser estendido para a combinação linear com n variáveis aleatórias.

EXEMPLO 9.9

Seja a função $H = f(X_1, X_2, X_3) = a_1X_1 + a_2X_2 + a_3X_3$. Calcule a média e o desvio padrão da variável H conhecendo as medidas estatísticas e os coeficientes de correlação das variáveis X_1 , X_2 e X_3 registradas nas tabelas seguintes.

	a	Média	D. padrão
X_1	1	1.000	50
X_2	-2	500	100
X_3	3	800	200

	X_1	X_2	X_3
X_1	1		
X_2	0,8	1	
X_3	-0,9	-0,85	1

Solução.

- A função H é $H = a_1X_1 + a_2X_2 + a_3X_3 = 1 \times X_1 - 2 \times X_2 + 3 \times X_3$
- A média $\mu_H = 2.400$ foi obtida com a fórmula:

$$\mu_H = a_1\mu_1 + a_2\mu_2 + a_3\mu_3$$

$$\mu_H = 1 \times 1.000 - 2 \times 500 + 3 \times 800 = 2.400$$

- A variância de H deve ser calculada com a fórmula:

$$\begin{aligned} \sigma_H^2 &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_3^2\sigma_3^2 + \\ &+ 2a_1a_2\sigma_1\sigma_2r_{12} + 2a_1a_3\sigma_1\sigma_3r_{13} + \\ &+ 2a_2a_3\sigma_2\sigma_3r_{23} \end{aligned}$$

Substituindo os dados nessa fórmula:

$$\begin{aligned} \sigma_H^2 &= 1 \times 50^2 + (-2)^2 \times 100^2 + 3^2 \times 200^2 + \\ &+ 2 \times 1 \times (-2) \times 50 \times 100 \times 0,80 + 2 \times 1 \times 3 \times 50 \times 200 \times (-0,90) + \\ &+ 2 \times (-2) \times 3 \times 100 \times 200 \times (-0,85) \\ \sigma_H^2 &= 402.500 + 134.000 = 536.500 \end{aligned}$$

A variância $H=536.500$ é o resultado da soma de duas parcelas. A primeira parcela (402.500) é a contribuição das variâncias das três variáveis aleatórias, ou variância pura, e a segunda parcela (134.000) é a contribuição das correlações.

- O desvio padrão de H é $\sigma_H = \sqrt{\sigma_H^2} = \sqrt{536.500} = 732,46$

A planilha **Combinação três VA**, inclua na pasta **Capítulo 9**, ajuda no cálculo dos parâmetros da função H com três variáveis aleatórias, como mostra a Figura 9.2, resolvendo o Exemplo 9.9.

FIGURA 9.2 Modelo para combinação de três VA.

	A	B	C	D	E	F	G	H
1	Combinação de três Variáveis Aleatórias							
2								
3		Dados				Resultados		
4			X_1	X_2	X_3		H	
5		Peso	1,00	-2,00	3,00		Média	2.400,00
6		Média	1.000,00	500,00	800,00		Variância	536.500,00
7		DP	50,00	100,00	200,00		DP	732,46
8								
9								
10			X_1	X_2	X_3			
11		X_1	1					
12		X_2	0,80	1				
13		X_3	-0,90	-0,85	1			
14								

Combinação Linear de n Variáveis Aleatórias

- A função H é a combinação linear $H = a_1X_1 + a_2X_2 + \dots + a_nX_n$.
- As constantes a_1, \dots, a_n podem assumir qualquer valor do conjunto dos números reais.
- As variáveis aleatórias X_1, \dots, X_n têm médias μ_1, \dots, μ_n e desvios padrão $\sigma_1, \dots, \sigma_n$.
- A média da variável H é obtida de:

$$\mu_H = \mu_{a_1X_1 + a_2X_2 + \dots + a_nX_n}$$

$$\mu_H = a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n = \sum_{i=1}^n a_i\mu_i$$

- A variância da variável H é obtida de:

$$\sigma_H^2 = \sigma_{a_1X_1 + a_2X_2 + \dots + a_nX_n}^2$$

$$\begin{aligned} \sigma_H^2 = & a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 + \dots + a_n^2 \cdot \sigma_n^2 + \\ & + 2 \cdot a_1 \cdot a_2 \cdot \sigma_{12} + \dots + 2 \cdot a_1 \cdot a_n \cdot \sigma_{1n} + \\ & + 2 \cdot a_2 \cdot a_3 \cdot \sigma_{23} + \dots + 2 \cdot a_2 \cdot a_n \cdot \sigma_{2n} + \\ & + \dots + \\ & + 2 \cdot a_{n-1} \cdot a_n \cdot \sigma_{(n-1)n} \end{aligned}$$

Com os coeficientes de correlação a expressão anterior será:

$$\begin{aligned} \sigma_H^2 = & a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2 + \\ & + 2a_1a_2\sigma_1\sigma_2r_{12} + \dots + 2a_1a_n\sigma_1\sigma_nr_{1n} + \\ & + 2a_2a_3\sigma_2\sigma_3r_{23} + \dots + 2a_2a_n\sigma_2\sigma_nr_{2n} + \\ & + \dots + \\ & + 2a_{n-1}a_n\sigma_{n-1}\sigma_nr_{(n-1)n} \end{aligned}$$

- O desvio padrão de H : $\sigma_H = \sigma_{a_1X_1 + a_2X_2 + \dots + a_nX_n} = \sqrt{\sigma_H^2}$
- Se as n variáveis aleatórias forem independentes, então todas as covariâncias são iguais a zero e, conseqüentemente, os coeficientes de correlação também serão iguais a zero. Dessa maneira, a única medida estatística de H que muda é a da variância que passa a ser $\sigma_H^2 = a_1^2 \cdot \sigma_1^2 + a_2^2 \cdot \sigma_2^2 + \dots + a_n^2 \cdot \sigma_n^2$.

Distribuição de H

A distribuição normal pode ocorrer em diversos contextos, por exemplo, nos dados extraídos de um experimento, na explicação de um experimento etc. Ao mesmo tempo, a soma de um número de efeitos aleatórios, sem dominância de nenhum deles sobre o resultado total, produz uma variável aleatória com distribuição normal. Dessa maneira, deve-se esperar que a função H definida como soma de variáveis aleatórias tenha distribuição normal, considerando que nenhuma das variáveis domina as restan-

tes. Pelo teorema central do limite, se as n variáveis aleatórias X_i que participam da combinação linear tiverem distribuição normal, então a nova variável H também terá distribuição normal, para qualquer valor de n . Entretanto, se as n variáveis aleatórias X_i que participam da combinação linear não tiverem distribuição normal, então a nova variável H terá distribuição normal se o número n de variáveis aleatórias for adequadamente grande. Nos outros casos, a distribuição normal poderá ser adotada como uma aproximação conveniente, pois a grande vantagem é a possibilidade de realizar inferências com a função H . Portanto, conhecendo os parâmetros da função H obtida como combinação linear de variáveis aleatórias será possível realizar cálculos de probabilidade.

EXEMPLO 9.10

Continuando com o Exemplo 9.9. Considerando que a distribuição de H seja normal, qual a probabilidade de que H seja igual ou maior do que 1.800?

Solução. A probabilidade de H ser igual ou maior do que 1.800 é 79,37%, ou $P(H \geq 1.800) = 79,37\%$, resultado obtido registrando em uma célula da planilha Excel a fórmula `=1-DIST.NORM(1800;2400;732,46;VERDADEIRO)`. Registrando a fórmula `=1-DIST.NORMP(PADRONIZAR(1800;2400;732,46))`, também se obtém o mesmo resultado.

EXEMPLO 9.11

Continuando com o Exemplo 9.9. Considerando que as três variáveis são independentes:

- Calcule a variância e o desvio padrão da função H .
- Qual a probabilidade de que H seja igual ou maior do que 1.800?

Solução.

- Na resolução do exemplo 9.9, destacamos que a variância de H é o resultado da soma de duas parcelas. A primeira parcela igual a 402.500 se refere à contribuição das variâncias das três variáveis aleatórias e é o único resultado que interessa para calcular o desvio padrão de H igual a 634,43, com a fórmula $\sigma_H = \sqrt{402.500} = 634,43$.
- A probabilidade de $P(H \geq 1.800)$ é igual a 82,79%, resultado obtido com a fórmula `=1-DIST.NORM(1800;2400;634,43;VERDADEIRO)`. Registrando a fórmula `=1-DIST.NORMP(PADRONIZAR(1800;2400;634,43))` também se obtém o mesmo resultado.

Modelo combinação linear de VA's

A Figura 9.3 mostra o *modelo* para combinação linear de até seis variáveis aleatórias, construído na planilha **Modelo Combinação de VA's** incluído na pasta **Capítulo 9**, resolvendo o Exemplo 9.9 com três VA's.

- Os dados são registrados nas células pintadas de cor azul. As células pintadas de cor verde retornam resultados e as demais células pintadas de cor amarela registram títulos.
- No intervalo de células C5:E10, são informados os parâmetros das variáveis aleatórias que participam da combinação linear: constante a , média e desvio padrão de cada uma das seis VA's.
 - Se a combinação linear for realizada com menos de seis variáveis aleatórias, deve-se verificar que as células dos dados das demais variáveis estejam vazias ou com valores zero.
- Na tabela dos coeficientes de correlação, são informados os valores de r da combinação de variáveis tomadas duas a duas, sendo que a diagonal principal já está preenchida com valores iguais a um, como foi mostrado no Capítulo 6.

No quadro *Função H*, o *modelo* fornece os resultados da nova variável aleatória H . No quadro *Probabilidade de H*, na célula G14 o modelo retorna a probabilidade selecionada na caixa de combinação,

considerando o valor registrado na célula G13. Os cálculos de probabilidade são realizados, considerando que a distribuição da variável H é normal.

Aplicações tradicionais de combinação linear de variáveis em finanças são apresentadas nos Apêndices 2 e 3 deste capítulo.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Modelo para Combinação Linear até Seis Variáveis Aleatórias											
2												
3		Dados das variáveis aleatórias X				Tabela dos coeficientes de correlação						
4		α	μ	σ		X1	X2	X3	X4	X5	X6	
5		X1	1,00	1.000	50	X1	1					
6		X2	-2,0000	500	100	X2	0,80	1				
7		X3	3,0000	800	200	X3	-0,90	-0,85	1			
8		X4				X4				1		
9		X5				X5					1	
10		X6				X6						1
11												
12		Função H				Probabilidade de H						
13		Média de H	2.400,00			0	$H \geq 0$			Z = -3,28		
14		Variância pura	402.500			$P(H \geq 0) = 99,95\%$						
15		Variância das correlações	134.000									
16		Variância de H	536.500									
17		Desvio padrão de H	732,46									
18												

FIGURA 9.3 Modelo para combinação linear até seis variáveis aleatórias.

Problemas

Problema 1

As vendas mensais da empresa dos últimos 24 meses têm média R\$320.000 por mês e desvio padrão R\$65.000 por mês. Calcule os novos valores de média e desvio padrão, considerando como moeda de referência a libra inglesa, utilizando a taxa de conversão 5,55 R\$/£\$ ou 0,1802 £\$/R\$.

R: Corrigir resultados $\mu = \text{£\$ } 57.664$ $\sigma = \text{£\$ } 11.713$

Problema 2

A taxa de juro mensal i tem média 0,80% por mês e desvio padrão 0,1% por mês. Se a relação entre o valor aplicado P e o resgatado F numa operação com prazo de um mês é estabelecida com a fórmula $F = P \times (1 + i)$, qual a média e o desvio padrão do fator $(1 + i)$ da fórmula?

R: $\mu_{(1+i)} = 1,008$ $\sigma_{(1+i)} = 0,001$

Problema 3

Continuando com o Problema 2. Se forem aplicados \$10.000 por um mês, qual a média e o desvio padrão do resgate F calculado com esta fórmula: $F = \$10.000 \times (1 + i)$?

R: $\mu_F = \$100.800$ $\sigma_F = \$10$

Problema 4

Com as variáveis aleatórias X_1 e X_2 cujas medidas estatísticas estão registradas na tabela seguinte foi formada a variável aleatória $H=2.X_1+3.X_2$. Calcule a média e o desvio padrão de H .

	X_1	X_2
Média	10	8
Variância	64	16
r	0,60	

R: $\mu_H=44,00$ $\sigma_H=25,11$

Problema 5

Repita o Problema 4, considerando que as variáveis são independentes.

R: $\mu_H=44,00$ $\sigma_H=20,00$

Problema 6

Repita o Problema 4, considerando o coeficiente de correlação $-0,60$.

R: $\mu_H=44,00$ $\sigma_H=13,02$

Problema 7

Com as variáveis aleatórias X_1 e X_2 cujas medidas estatísticas estão registradas na tabela seguinte, foi formada a variável aleatória $H=-X_1+2.X_2$. Calcule a média e o desvio padrão de H .

	X_1	X_2
Média	1,2	3,5
Variância	9	16
r	-0,85	

R: $\mu_H=5,56$ $\sigma_H=11,22$

Problema 8

Repita o Problema 7, considerando que as variáveis são independentes.

R: $\mu_H=5,56\%$ $\sigma_H=8,773\%$

Problema 9

Repita o Problema 7, considerando o coeficiente de correlação $0,85$.

R: $\mu_H=5,56\%$ $\sigma_H=5,29\%$

Problema 10

Sabendo que a variável aleatória X tem média 2 e variância 1, calcule a média e o desvio padrão da nova variável aleatória $H=-3,6+8.X$.

R: $\mu_H=12,40$ $\sigma_H=8,00$

Problema 11

O preço de venda unitário do produto é $P=\$1.250$, o custo unitário é $C=\$500$ e o custo fixo mensal é $CF=\$120.000$. Se o número de produtos vendidos n por mês tem distribuição normal com média 200 unidades e desvio padrão de 35 unidades, qual a média e o desvio padrão do lucro L medido com fórmula: $L = -CF + (P - C) \times n$?

R: $\mu_L=\$30.000$ $\sigma_L=\$26.250$

Problema 12

Continuando com o Problema 11. Qual a probabilidade de o lucro da empresa ser positivo, maior ou igual a zero?

$$R: P(L \geq 0) = 87,35\%$$

Problema 13

Jean aplicou \$10.000 durante 180 dias com taxa de juro pós-fixada i . A estimativa da taxa de juro i para 180 dias tem distribuição normal, com média igual a 10% aos 180 dias e desvio padrão de 0,3%. Qual a média e o desvio padrão do resgate R calculado com $R = \$10.000 \times (1 + i)$?

$$R: \mu_R = \$11.000 \quad \sigma_R = \$30$$

Problema 14

Continuando com o Problema 13. Qual a probabilidade de o resgate R da aplicação ser maior do que \$11.030?

$$R: P(R \geq 11.030) = 15,87\%$$

Problema 15

Continuando com o Problema 13. Qual a probabilidade de $R \leq \$10.990$?

$$R: P(R \leq 10.990) = 36,94\%$$

Problema 16

Calcule a média e o desvio padrão da variável aleatória H , formada pela combinação linear de três variáveis aleatórias, cujos dados estão registrados nas tabelas seguintes.

	a	Média	D.Padrão
x_1	1	100	30
x_2	1	150	50
x_3	1	130	45

	x_1	x_2	x_3
x_1	1		
x_2	0,65	1	
x_3	0,70	0,90	1

$$R: \mu_H = 380 \quad \sigma_H = 115,39$$

Problema 17

Continuando com o Problema 16. Calcule a probabilidade de H ser maior do que 300, considerando que a distribuição da variável H seja normal.

$$R: P(H \geq 300) = 75,59\%$$

Problema 18

Os dados das quatro variáveis aleatórias que participam da combinação linear estão registrados nas tabelas seguintes. Calcule a média e o desvio padrão da função H .

	a	Média	D.padrão
x_1	-1,00	15.000	200
x_2	0,89	4.500	400
x_3	0,80	8.500	800
x_4	0,71	10.000	1.000

	x_1	x_2	x_3	x_4
x_1	1			
x_2	0,50	1		
x_3	0,80	0,90	1	
x_4	1,00	0,50	0,40	1

$$R: \mu_H = 2.905 \quad \sigma_H = 1.243,72$$

Problema 19

Continuando com o Problema 18. Considerando que a variável H tem distribuição normal, qual o valor da probabilidade $P(H \geq 3.500)$?

R: $P(H \geq 3.500) = 31,62\%$

Problema 20

Os dados da combinação linear de três variáveis aleatórias estão registrados nas tabelas seguintes. Calcule a média e o desvio padrão da função H .

	α	Média	D.padrão
x_1	0,9091	1.400	219,09
x_2	0,8264	1.800	273,86
x_3	0,7513	2.500	438,18

	x_1	x_2	x_3
x_1	1		
x_2	0,80	1	
x_3	0,50	0,65	1

R: $\mu_H = 4.638,51$ $\sigma_H = 658,65$

Problema 21

Continuando com o Problema 20, calcule a probabilidade de H ser menor do que 4.000, considerando que a distribuição de H seja normal.

R: $P(H \leq 4.000) = 16,62\%$

Problema 22

Repita o Problema 20, considerando as variáveis aleatórias independentes.

R: $\mu_H = 4.638,51$ $\sigma_H = 446,39$

Problema 23

Continuando com o Problema 22. Considerando que a distribuição de H seja normal, qual a probabilidade de H ser maior do que 4.000.

R: $P(H \leq 4.000) = 7,63\%$

Problema 24

Para comparar os salários dos gerentes da filial, a matriz americana informou que os salários anuais dos gerentes americanos têm média US\$36.000 e desvio padrão US\$5.800. O gerente de RH da filial converteu os salários em R\$, aplicando a taxa de câmbio 1 USD\$=3,00 R\$, obtendo média R\$108.000 e desvio padrão igual a R\$17.400. Do ponto de vista da transformação, os resultados em reais devem ser aceitos?

R: Resultados corretos.

Problema 25

Analisando os resultados estatísticos do processo produtivo com média de 12,5 e desvio padrão 0,26, o supervisor do controle de qualidade verificou que, no resultado da média, foi esquecido de adicionar o valor de referência 100. Calcule os novos valores de média e de desvio padrão.

R: Média=112,5 e Desvio Padrão=0,26

Apêndice 1

Propriedades para duas variáveis aleatórias

A seguir, são apresentadas algumas propriedades importantes da combinação linear de variáveis aleatórias.

- A média do produto da constante a pela variável aleatória X é igual ao produto da constante pela média da variável aleatória. Pela definição do valor esperado, pode-se escrever que:

$$\mu_{aX} = \frac{\sum_{i=1}^N ax_i}{N} = a \times \frac{\sum_{i=1}^N x_i}{N}$$

$$\mu_{aX} = a \times \mu_X$$

- A média da soma de duas variáveis aleatórias é igual à soma das médias das duas variáveis aleatórias.

$$\mu_{X_1+X_2} = \frac{\sum_{i=1}^N (x_{1_i} + x_{2_i})}{N} = \frac{\sum_{i=1}^N x_{1_i}}{N} + \frac{\sum_{i=1}^N x_{2_i}}{N}$$

$$\mu_{X_1+X_2} = \mu_{X_1} + \mu_{X_2}$$

- A variância do produto da constante a pela variável aleatória X é igual ao produto do quadrado da constante a pela variância da variável X . Partindo da expressão da variância de X :

$$\sigma_X^2 = \frac{\sum_{i=1}^N (x_i - \mu_X)^2}{N} = \frac{\sum_{i=1}^N (x_i^2 - 2x_i\mu_X + \mu_X^2)}{N}$$

Desenvolvendo a soma indicada no *somatório* temos:

$$\sigma_X^2 = \frac{\sum_{i=1}^N x_i^2}{N} - \frac{2\mu_X \times \sum_{i=1}^N x_i}{N} + \frac{\sum_{i=1}^N \mu_X^2}{N}$$

Analisando cada parcela dessa expressão obtemos:

$$\sigma_X^2 = \mu_{X^2} - 2\mu_X^2 + \mu_X^2$$

$$\sigma_X^2 = \mu_{X^2} - \mu_X^2$$

Multiplicando a variável X pela constante a obtemos:

$$\sigma_{aX}^2 = \mu_{(aX)^2} - \mu_{aX}^2 = \mu_{a^2X^2} - \mu_{aX} \times \mu_{aX}$$

$$\sigma_{aX}^2 = a^2 \mu_{X^2} - a^2 \times \mu_X^2 = a^2 \times (\mu_{X^2} - \mu_X^2)$$

$$\sigma_{aX}^2 = a^2 \sigma_X^2$$

- Expressão da variância da combinação linear de duas variáveis aleatórias X_1 e X_2 . Da definição de variância podemos escrever:

$$\sigma_{a_1X_1+a_2X_2}^2 = \mu_{(a_1X_1+a_2X_2)^2} - \mu_{a_1X_1+a_2X_2}^2$$

$$\sigma_{a_1X_1+a_2X_2}^2 = \mu_{(a_1^2X_1^2+a_2^2X_2^2+2a_1a_2X_1X_2)^2} - \mu_{a_1X_1+a_2X_2}^2$$

$$\sigma_{a_1X_1+a_2X_2}^2 = a_1^2 \mu_{X_1^2} + a_2^2 \mu_{X_2^2} + 2a_1a_2 \mu_{X_1X_2} - (a_1^2 \mu_{X_1}^2 + a_2^2 \mu_{X_2}^2 + 2a_1a_2 \mu_{X_1} \mu_{X_2})$$

Aplicando as propriedades anteriores, temos:

$$\sigma_{a_1X_1+a_2X_2}^2 = a_1^2 \mu_{X_1^2} + a_2^2 \mu_{X_2^2} + 2a_1a_2 \mu_{X_1X_2} - (a_1^2 \mu_{X_1}^2 + a_2^2 \mu_{X_2}^2 + 2a_1a_2 \mu_{X_1} \mu_{X_2})$$

Finalmente, temos a expressão da variância procurada:

$$\sigma_{a_1X_1+a_2X_2}^2 = a_1^2 (\mu_{X_1^2} - \mu_{X_1}^2) + a_2^2 (\mu_{X_2^2} - \mu_{X_2}^2) + 2a_1a_2 (\mu_{X_1X_2} - \mu_{X_1} \mu_{X_2})$$

$$\sigma_{a_1X_1+a_2X_2}^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + 2a_1a_2 \sigma_{X_1X_2}$$

Diminuindo os índices das variáveis temos a expressão:

$$\sigma_{a_1X_1+a_2X_2}^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2a_1a_2 \sigma_{12}$$

Apêndice 2

Análise do valor presente líquido de um projeto de investimento

Na avaliação de um projeto de investimento com o *método do valor presente líquido*, ou simplesmente VPL, é utilizada a fórmula:

$$VPL = -I + \frac{FC_1}{(1+k)^1} + \frac{FC_2}{(1+k)^2} + \dots + \frac{FC_n}{(1+k)^n} = -I + \sum_{t=1}^n \frac{FC_t}{(1+k)^t}$$

Na expressão do fluxo de caixa do projeto, participam o desembolso inicial I (investimento) realizado em $t=0$ e com sinal negativo, os n retornos anuais FC identificados pelo índice $t=1, 2, \dots, n$, e a taxa requerida de juro k . Os dados utilizados na fórmula do VPL são estimativas definidas antes de realizar o investimento e estão sujeitas à variação durante a execução se o projeto for realizado. Dessa maneira, pode-se considerar que essas estimativas são variáveis aleatórias com suas respectivas medidas estatísticas (média e variância ou desvio padrão) e que se relacionam entre si de alguma maneira (coeficiente de correlação). Para simplificar essa apresentação, consideraremos como variável aleatória somente os n capitais FC e o investimento I . Pode-se ver que a fórmula do VPL é uma combinação linear de n variáveis aleatórias, como fica destacado reescrevendo a fórmula da seguinte forma:

$$VPL = -I + \frac{1}{(1+k)^1} FC_1 + \frac{1}{(1+k)^2} FC_2 + \dots + \frac{1}{(1+k)^n} FC_n$$

Como as variáveis aleatórias $I, FC_1, FC_2, \dots, FC_n$ são definidas pelas suas respectivas medidas estatísticas e os coeficientes de correlação que estabelecem a relação entre elas, a função VPL é uma variável aleatória H . O Exemplo 9.12 mostra como calcular os parâmetros de VPL

EXEMPLO 9.12

O fluxo de caixa do projeto de investimento está registrado na primeira parte da tabela seguinte, onde anualmente cada capital é uma variável aleatória com valores de média e desvio padrão definido. Os capitais do fluxo de caixa começam na data inicial $t=0$, quando é realizado o desembolso do investimento, e seguem com os retornos anuais de $t=1$ até $t=4$. Na tabela ao lado, estão registrados os coeficientes de correlação entre as variáveis.

t	Capitais	Média	D. padrão		I	CF_1	CF_2	CF_3	CF_4
0	$-I$	-\$12.500	\$1.300	I	1				
1	CF_1	\$3.000	\$330	CF_1	0	1			
2	CF_2	\$3.800	\$450	CF_2	0	0,90	1		
3	CF_3	\$4.500	\$630	CF_3	0	0,82	0,85	1	
4	CF_4	\$5.600	\$750	CF_4	0	0,70	0,78	0,80	1

Determine os parâmetros da variável aleatória VPL e a probabilidade de que VPL seja maior do que zero, considerando a taxa requerida de juro de 8% ao ano.

Solução. A função *VPL* do projeto de investimento está registrada a seguir.

$$VPL = -I + \frac{1}{(1+k)^1} FC_1 + \frac{1}{(1+k)^2} FC_2 + \frac{1}{(1+k)^3} FC_3 + \frac{1}{(1+k)^4} FC_4$$

Cálculo da média do *VPL*.

A média do *VPL* é obtida com a seguinte fórmula:

$$\mu_{VPL} = 1 \cdot \mu_I + \frac{1}{1+i} \mu_{FC_1} + \frac{1}{(1+i)^2} \mu_{FC_2} + \frac{1}{(1+i)^3} \mu_{FC_3} + \frac{1}{(1+i)^4} \mu_{FC_4}$$

Substituindo os dados, obtém-se a média de *VPL* igual a \$3.524,08.

$$\begin{aligned} \mu_{VPL} &= -10.200 + \frac{1}{(1+0,08)^1} 3.000 + \frac{1}{(1+0,08)^2} 3.800 + \\ &\quad + \frac{1}{(1+0,08)^3} 4.500 + \frac{1}{(1+0,08)^4} 5.600 \\ \mu_{VPL} &= 3.524,08 \end{aligned}$$

Cálculo da variância do *VPL*.

A variância do *VPL* deste projeto de investimento é obtida com a fórmula:

$$\begin{aligned} \sigma_{VPL}^2 &= (-1)^2 \cdot \sigma_I^2 + \frac{1}{(1+i)^2} \sigma_{FC_1}^2 + \frac{1}{(1+i)^4} \sigma_{FC_2}^2 + \frac{1}{(1+i)^6} \sigma_{FC_3}^2 + \frac{1}{(1+i)^8} \sigma_{FC_4}^2 + \\ &\quad + 2(-1) \frac{1}{1+i} r_{I,FC_1} \cdot \sigma_I \cdot \sigma_{FC_1} + 2(-1) \frac{1}{(1+i)^2} r_{I,FC_2} \cdot \sigma_I \cdot \sigma_{FC_2} + \\ &\quad + 2(-1) \frac{1}{(1+i)^3} r_{I,FC_3} \cdot \sigma_I \cdot \sigma_{FC_3} + 2(-1) \frac{1}{(1+i)^4} r_{I,FC_4} \cdot \sigma_I \cdot \sigma_{FC_4} + \\ &\quad + 2 \frac{1}{1+i} \frac{1}{(1+i)^2} r_{FC_1,FC_2} \cdot \sigma_{FC_1} \cdot \sigma_{FC_2} + 2 \frac{1}{1+i} \frac{1}{(1+i)^3} r_{FC_1,FC_3} \cdot \sigma_{FC_1} \cdot \sigma_{FC_3} + \\ &\quad + 2 \frac{1}{1+i} \frac{1}{(1+i)^4} r_{FC_1,FC_4} \cdot \sigma_{FC_1} \cdot \sigma_{FC_4} + 2 \frac{1}{(1+i)^2} \frac{1}{(1+i)^3} r_{FC_2,FC_3} \cdot \sigma_{FC_2} \cdot \sigma_{FC_3} + \\ &\quad + 2 \frac{1}{(1+i)^2} \frac{1}{(1+i)^4} r_{FC_2,FC_4} \cdot \sigma_{FC_2} \cdot \sigma_{FC_4} + 2 \frac{1}{(1+i)^3} \frac{1}{(1+i)^4} r_{FC_3,FC_4} \cdot \sigma_{FC_3} \cdot \sigma_{FC_4} \end{aligned}$$

Substituindo os dados, obtém-se a variância de *VPL* igual a 4.285.760.

$$\begin{aligned} \sigma_{VPL}^2 &= 1.300^2 + \frac{1}{1,08^2} 330^2 + \frac{1}{1,08^4} 450^2 + \frac{1}{1,08^6} 630^2 + \frac{1}{1,08^8} 750^2 + \\ &\quad - 2 \frac{1}{1,08} \times 0 \times 1.300 \times 330 - 2 \frac{1}{1,08^2} \times 0 \times 1.300 \times 450 + \\ &\quad - 2 \frac{1}{1,08^3} \times 0 \times 1.300 \times 630 - 2 \frac{1}{1,08^2} \times 0 \times 1.300 \times 750 + \\ &\quad + 2 \frac{1}{1,08} \frac{1}{1,08^2} \times 0,90 \times 330 \times 450 + 2 \frac{1}{1,08} \frac{1}{1,08^3} \times 0,82 \times 330 \times 630 + \\ &\quad + 2 \frac{1}{1,08} \frac{1}{1,08^3} \times 0,82 \times 330 \times 750 + 2 \frac{1}{1,08^2} \frac{1}{1,08^3} \times 0,70 \times 450 \times 630 + \\ &\quad + 2 \frac{1}{1,08^2} \frac{1}{1,08^4} \times 0,70 \times 450 \times 750 + 2 \frac{1}{1,08^3} \frac{1}{1,08^4} \times 0,80 \times 630 \times 750 \\ \sigma_{VPL}^2 &= 4.285.760 \end{aligned}$$

O desvio padrão de *VPL* é igual a \$2.070,21, resultado obtido como a raiz quadrada positiva da variância de *VPL*.

$$\sigma_{VPL} = \sqrt{\sigma_{VPL}^2} = \sqrt{4.285.760} = 2.070,21$$

Com o *modelo* da planilha **Modelo Combinação de VAs**, incluído na pasta **Capítulo 9**, também é possível resolver este exemplo, como mostra a figura seguinte. Para calcular a probabilidade de que o valor do *VPL* seja maior do que zero, na célula G3 do *modelo* foi registrado valor zero retornando na célula G14 o resultado $P(VPL \geq 0) = 95,56\%$, mostrando uma probabilidade bastante alta. Apenas para você conferir esse resultado, se numa célula do Excel for registrada a fórmula =1-DIST.NORM(0;3524,08;2070,21;VERDADEIRO), será obtido o mesmo resultado.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Modelo para Combinação Linear até Seis Variáveis Aleatórias											
2												
3		Dados das variáveis aleatórias X				Tabela dos coeficientes de correlação						
4		α	μ	σ		X1	X2	X3	X4	X5	X6	
5		X1	-1,00	10.200	1.300	X1	1					
6		X2	0,9259	3.000	330	X2	0,00	1				
7		X3	0,8573	3.800	450	X3	0,00	0,90	1			
8		X4	0,7938	4.500	630	X4	0,00	0,82	0,85	1		
9		X5	0,7350	5.600	750	X5	0,00	0,70	0,78	0,80	1	
10		X6				X6						1
11												
12		Função H				Probabilidade de H						
13		Média de H		3.524,08		0	H \geq 0			Z = -1,70		
14		Variância pura		2.486.223		$P(H \geq 0) = 95,56\%$						
15		Variância das correlações		1.799.536								
16		Variância de H		4.285.760								
17		Desvio padrão de H		2.070,21								
18												

Apêndice 3

Formação de uma carteira de investimento

Os exemplos anteriores mostraram que ao investir o capital disponível em dois ou mais ativos ou projetos, o indivíduo ou empresa *diversifica* seu capital com o objetivo de diminuir o risco e manter o mesmo retorno esperado. A teoria que será utilizada para formar carteiras foi iniciada por Harry Markowitz, com a publicação *Portfolio Selection*, no *Journal of Finance*, em 1952 e, posteriormente, Merton Miller e William Sharpe realizaram contribuições importantes. Aos três professores, foi outorgado o *Prêmio Nobel de Economia* de 1990. Iniciamos o tema com a resolução de um exemplo e seguimos com a utilização do comando Solver do Excel. Assim, serão introduzidas novas respostas na formação de uma carteira de investimento.

EXEMPLO 9.13

A tabela seguinte registra as medidas estatísticas de três ativos e seus pesos na formação de uma carteira de investimento.

	Ativo 1	Ativo 2	Ativo 3
Peso	50,00%	15,00%	35,00%
Média	22,00%	10,00%	17,00%
D.padrão	60,00%	26,46%	35,00%

As covariâncias dos três ativos estão registradas na tabela seguinte. Calcule o retorno, a variância e o desvio padrão dessa carteira de investimento.

	Ativo 1	Ativo 2	Ativo 3
Ativo 1	0,36		
Ativo 2	0,02	0,07	
Ativo 3	0	0,04	0,1225

Solução. O valor esperado da carteira é igual a 18,45%, valor obtido depois de substituir os dados do exemplo na fórmula conhecida.

$$E[R_p] = 0,50 \times 22\% + 0,15 \times 10\% + 35\% \times 17\% = 18,45\%$$

A variância da carteira é 0,11378, resultado obtido depois de substituir os dados do exemplo na fórmula conhecida, lembrando que são utilizadas as covariâncias, valores obtidos com a fórmula $\sigma_{xy} = r_{xy} \sigma_x \sigma_y$.

$$\begin{aligned} \sigma_p^2 &= 0,50^2 \times 0,36 + 0,15^2 \times 0,07 + 0,35^2 \times 0,1225 + 2 \times 0,50 \times 0,15 \times 0,02 + 2 \times 0,50 \times 0,35 \times 0 + \\ &\quad + 2 \times 0,15 \times 0,35 \times 0,04 \\ \sigma_p^2 &= 0,11378 \end{aligned}$$

O desvio padrão da carteira é igual a 0,3373 ou 33,73%.

A planilha **Formação de Carteira**, incluída na pasta **Capítulo 9**, é utilizada para resolver o Exemplo 9.13 e apresentar os outros tópicos do tema carteiras. A Figura 9.4 mostra esse *modelo* copiado da planilha **Modelo Combinação de VAs** com novos títulos, novos cálculos e resultados.

Formação de uma carteira utilizando o solver

Com três ativos é possível formar muitas carteiras mudando a proporção dos ativos selecionados. Dessas carteiras, sem dúvida, o investidor escolherá a carteira com menor risco ou variância mínima. De outra maneira, a carteira de *variância mínima* é a carteira com a menor variância de todas as carteiras possíveis, que podem ser formadas com os ativos sem especificar antecipadamente o retorno da carteira. Para formar uma carteira com três ativos, o objetivo do investidor é definido de forma matemática, como segue:

$$\begin{aligned} \text{Minimizar} \Rightarrow \sigma_p^2 &= w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2 + w_3^2 \cdot \sigma_3^2 + \\ &\quad + 2 \cdot w_1 \cdot w_2 \cdot \sigma_{1,2} + 2 \cdot w_1 \cdot w_3 \cdot \sigma_{1,3} + 2 \cdot w_2 \cdot w_3 \cdot \sigma_{2,3} \end{aligned}$$

Sujeita à restrição $w_1 + w_2 + w_3 = 1$, sendo w_i a proporção do investimento no ativo i .


	A	B	C	D	E	F	G	H	I	J	K	L
1	Carteira até Seis Ativos											
2												
3	Dados dos ativos				Tabela dos coeficientes de correlação							
4		w	μ	σ		Ativo 1	Ativo 2	Ativo 3	Ativo 4	Ativo 5	Ativo 6	
5	Ativo 1	50,00%	22,00%	60,00%	Ativo 1	1						
6	Ativo 2	15,00%	10,00%	26,46%	Ativo 2	0,13	1					
7	Ativo 3	35,00%	17,00%	35,00%	Ativo 3	0,00	0,43	1				
8	Ativo 4				Ativo 4				1			
9	Ativo 5				Ativo 5					1		
10	Ativo 6				Ativo 6						1	
11												
12	Resultados da carteira				Probabilidade de H							
13	Média da Carteira		18,45%		0	H \geq 0		Z = -0,55				
14	Variância pura		0,10658		P(H \geq 0) = 70,78%							
15	Variância das correlações		0,00720									
16	Variância da Carteira		0,11378		Soma dos ativos da Carteira				1,00			
17	Desvio padrão da Carteira		33,73%		Retorno esperado da Carteira							
18												

FIGURA 9.4
Modelo para
Formação de
Carteira de
investimento.

Carteira de variância mínima

O comando Solver do Excel é uma ferramenta de otimização muito prática na análise e na formação de carteiras, como será mostrado com o Exemplo 9.13. O ponto de partida é o registro das medidas estatísticas dos ativos, como foi feito na planilha **Formação de Carteira** mostrada na Figura 9.4. Na célula K16 dessa planilha foi registrada a fórmula =SOMA(C5:C10), correspondente à restrição da minimização, que sempre deverá retornar o valor 1, pois a soma das porcentagens do investimento em cada ativo tem de ser igual a 100%. Para obter a proporção do investimento nos três ativos que formam a carteira de variância mínima, procede-se como segue:

- Posicione o cursor da planilha na célula E17 e depois no menu **Ferramentas** escolha **Solver**.⁷
- Na caixa de diálogo **Parâmetros do Solver** da Figura 9.5.
 - Ao ter selecionado a célula E17, esse endereço aparecerá na caixa **Definir célula de destino**.
 - Selecione o botão de opção **Mín**.
 - No quadro **Células variáveis**: registre o intervalo das células dos pesos dos ativos C5:C7. Para inserir esse intervalo, selecione a célula C5 e, mantendo pressionado o botão esquerdo, arraste o mouse até a célula C7, quando soltará o botão.

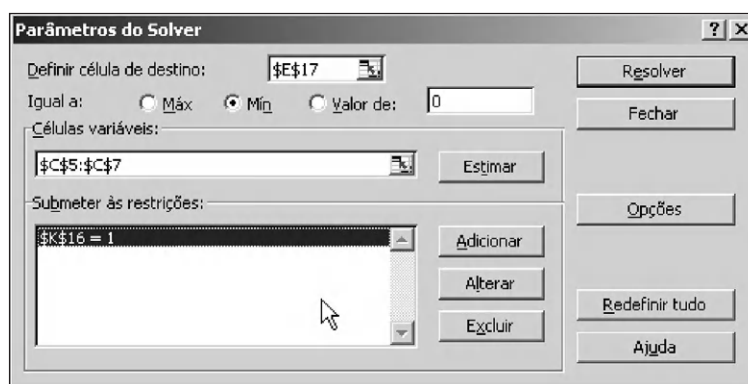


Figura 9.5 Caixa de diálogo dos Parâmetros do Solver.

⁷ O Solver é um *Suplemento* que nem sempre é incorporado ao iniciar o Excel. Para obter mais informações, veja o Apêndice 1 do Capítulo 1, ou a *Ajuda* do Excel.

- No quadro **Submeter as restrições**, pressione o botão **Adicionar**. Na caixa de diálogo **Adicionar restrição**, registre a restrição mostrada na Figura 9.6. Repare que a restrição já aparece na caixa de diálogo da Figura 9.5.

FIGURA 9.6 Caixa de diálogo das restrições do Solver.

Alterar restrição

Referência de célula:

Restrição:

OK Cancelar Adicionar Ajuda

Depois de pressionar o botão **Resolver**, o comando *Solver* registra os resultados no intervalo C5:C6 e apresenta a caixa de diálogo da Figura 9.7, onde está registrado que *O Solver encontrou uma solução. Todas as restrições e condições otimizadas foram atendidas.*

FIGURA 9.7 Caixa de diálogo Resultados do Solver.

Resultados do Solver ? X

O Solver encontrou uma solução. Todas as restrições e condições otimizadas foram atendidas.

☒ Manter solução do Solver
☐ Restaurar valores originais

Relatórios

- Resposta
- Sensibilidade
- Limites

OK Cancelar Salvar cenário... Ajuda

Pressionando o botão OK, os resultados são registrados no intervalo C5:C7, como mostra a Figura 9.8.


	A	B	C	D	E	F	G	H	I	J	K	L
1	Carteira até Seis Ativos											
2												
3	Dados dos ativos				Tabela dos coeficientes de correlação							
4		<i>w</i>	μ	σ		Ativo 1	Ativo 2	Ativo 3	Ativo 4	Ativo 5	Ativo 6	
5	Ativo 1	12,17%	22,00%	60,00%	Ativo 1	1						
6	Ativo 2	62,24%	10,00%	26,46%	Ativo 2	0,13	1					
7	Ativo 3	25,59%	17,00%	35,00%	Ativo 3	0,00	0,43	1				
8	Ativo 4				Ativo 4				1			
9	Ativo 5				Ativo 5					1		
10	Ativo 6				Ativo 6						1	
11												
12	Resultados da carteira				Probabilidade de <i>H</i>							
13	Média da Carteira	13,25%			0	<i>H</i> ≥ 0		<i>Z</i> = -0,56				
14	Variação pura	0,04048			<i>P</i> (<i>H</i> ≥ 0) = 71,18%							
15	Variação das correlações	0,01577										
16	Variação da Carteira	0,05625		Soma dos ativos da Carteira				1,00				
17	Desvio padrão da Carteira	23,72%		Retorno esperado da Carteira								
18												

FIGURA 9.8 Carteira de investimento com variância mínima.

Portanto, para formar uma carteira de investimento com variância mínima sem especificar o retorno da carteira, o investidor deverá comprar 12,17% do capital disponível do Ativo 1, comprar 62,24% do Ativo 2 e comprar 25,59% do Ativo 3, sendo a soma dessas proporções igual a 100%, com o mínimo desvio padrão da carteira de 23,72%. Verifique que qualquer outra combinação de proporção de ativos formará uma carteira com desvio padrão maior do que 23,72%.⁸

Carteira de mínima variância para um definido retorno da carteira

O retorno esperado da carteira da Figura 9.8 é 13,25% com o menor risco possível com esses ativos, desvio padrão 23,72%. Além de ter uma carteira com mínimo risco, o investidor gostaria que essa carteira tivesse um retorno predefinido. Em outros termos, para um determinado retorno da carteira definida pelo investidor μ_p^* , o objetivo é escolher a carteira com mínimo risco, ou mínimo desvio padrão. Nessas condições, para formar uma carteira com três ativos, o objetivo do investidor é definido de forma matemática como segue:

$$\text{Minimizar} \Rightarrow \sigma_p^2 = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2 + w_3^2 \cdot \sigma_3^2 + 2 \cdot w_1 \cdot w_2 \cdot \sigma_{1,2} + 2 \cdot w_1 \cdot w_3 \cdot \sigma_{1,3} + 2 \cdot w_2 \cdot w_3 \cdot \sigma_{2,3}$$

$$\text{Sujeita às restrições} \begin{cases} w_1 \cdot \mu_1 + w_2 \cdot \mu_2 + w_3 \cdot \mu_3 = \mu_p^* \\ w_1 + w_2 + w_3 = 1 \end{cases}$$

Continuando com o procedimento anterior, na célula K17 da planilha **Formação de Carteira** foi registrado o retorno esperado da carteira igual a 16%. A seguir, no menu **Ferramentas** escolha **Solver** e, na caixa de diálogo **Parâmetros do Solver**, adicione a nova restrição que mostra a Figura 9.9.

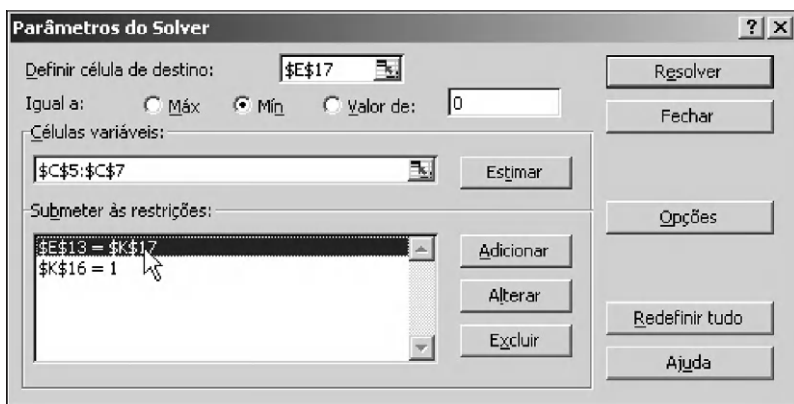


FIGURA 9.9 Caixa de diálogo do Solver com a nova restrição.

Depois, pressione o botão **Resolver** e aceite registrar os novos resultados no intervalo C5:C7, como mostra a Figura 9.10. Verifique que, para esse retorno de 16%, qualquer outra combinação de proporção de ativos formará uma carteira com desvio padrão maior do que 25,71%.

Carteira de mínima variância e sem venda a descoberto

Vejamos outro exemplo. Se o retorno esperado da carteira for 20%, você verificará que as proporções dos ativos da carteira indicarão que se deve comprar 37,85% do capital disponível do Ativo 1, vender


⁸ O comando *Solver* apresenta outras respostas que não são apresentadas neste livro.

–15,82% do Ativo 2 e comprar 77,97% do Ativo 3, sendo a soma dessas proporções igual a 100%, com o mínimo desvio padrão da carteira de 33,99%. A operação de venda é denominada descoberta e inclui um risco maior para o investidor. O próximo passo é construir uma carteira sem venda de ativos cujo objetivo do investidor, para uma carteira com três ativos, é definido matematicamente como segue:

$$\text{Minimizar} \Rightarrow \sigma_p^2 = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2 + w_3^2 \cdot \sigma_3^2 + 2 \cdot w_1 \cdot w_2 \cdot \sigma_{1,2} + 2 \cdot w_1 \cdot w_3 \cdot \sigma_{1,3} + 2 \cdot w_2 \cdot w_3 \cdot \sigma_{2,3}$$

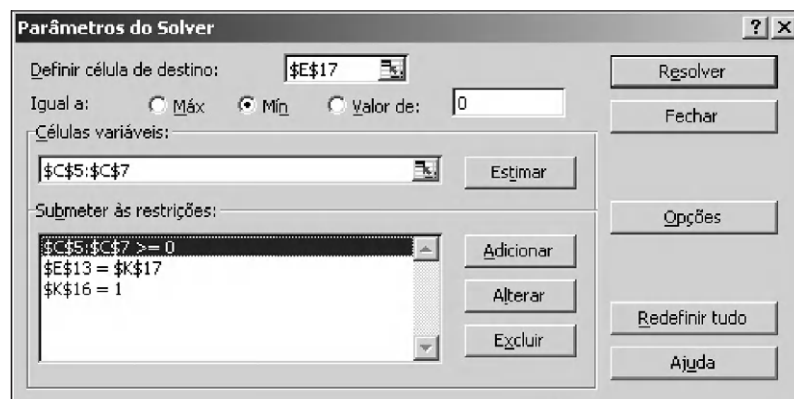
$$\text{Sujeita às restrições} \begin{cases} w_1 \cdot \mu_1 + w_2 \cdot \mu_2 + w_3 \cdot \mu_3 = \mu_p^* \\ w_1 + w_2 + w_3 = 1 \\ w_1 \geq 0, w_2 \geq 0, w_3 \geq 0 \end{cases}$$

FIGURA 9.10 Carteira com retorno definido e mínima variância.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Carteira até Seis Ativos											
2												
3	Dados dos ativos				Tabela dos coeficientes de correlação							
4		w	μ	σ		Ativo 1	Ativo 2	Ativo 3	Ativo 4	Ativo 5	Ativo 6	
5	Ativo 1	22,63%	22,00%	60,00%	Ativo 1	1						
6	Ativo 2	30,45%	10,00%	26,46%	Ativo 2	0,13	1					
7	Ativo 3	46,92%	17,00%	35,00%	Ativo 3	0,00	0,43	1				
8	Ativo 4				Ativo 4				1			
9	Ativo 5				Ativo 5					1		
10	Ativo 6				Ativo 6						1	
11												
12	Resultados da carteira					Probabilidade de H						
13	Média da Carteira		16,00%			0	H ≥ 0		Z = -0,62			
14	Variância pura		0,05190			P(H ≥ 0) = 73,32%						
15	Variância das correlações		0,01419									
16	Variância da Carteira		0,06608			Soma dos ativos da Carteira				1,00		
17	Desvio padrão da Carteira		25,71%			Retorno esperado da Carteira				16,00%		
18												

Continuando com o procedimento anterior, na célula K17 da planilha Formação de Carteira foi registrado o retorno esperado da carteira igual a 20%. A seguir, no menu Ferramentas, escolha Solver e, na caixa de diálogo Parâmetros do Solver, adicione a nova restrição que mostra a Figura 9.11.

FIGURA 9.11 Caixa de diálogo do Solver com restrição de venda de ativos.



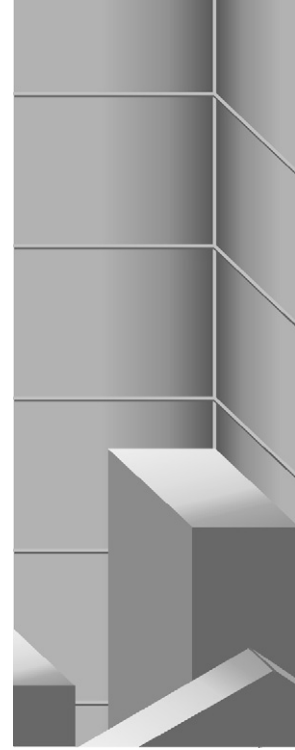
Depois, pressione o botão **Resolver** e aceite registrar os novos resultados no intervalo C5:C7, como mostra a Figura 9.12. Para um retorno esperado de 20% sem venda de ativos, as proporções dos ativos da carteira indicarão que o investidor deverá comprar 60% do capital disponível do Ativo 1, não comprar nada do Ativo 2 e comprar 40% do Ativo 3. Neste caso, também, observe que qualquer outra combinação de proporção de ativos formará uma carteira com desvio padrão maior do que 38,63%

	A	B	C	D	E	F	G	H	I	J	K	L
1	Carteira até Seis Ativos											
2												
3	Dados dos ativos				Tabela dos coeficientes de correlação							
4		w	μ	σ		Ativo 1	Ativo 2	Ativo 3	Ativo 4	Ativo 5	Ativo 6	
5	Ativo 1	60,00%	22,00%	60,00%	Ativo 1	1						
6	Ativo 2	0,00%	10,00%	26,46%	Ativo 2	0,13	1					
7	Ativo 3	40,00%	17,00%	35,00%	Ativo 3	0,00	0,43	1				
8	Ativo 4				Ativo 4				1			
9	Ativo 5				Ativo 5					1		
10	Ativo 6				Ativo 6						1	
11												
12	Resultados da carteira				Probabilidade de H							
13	Média da Carteira		20,00%		0	H ≥ 0		Z = -0,52				
14	Variância pura		0,14920		P(H ≥ 0) = 69,77%							
15	Variância das correlações		0,00000									
16	Variância da Carteira		0,14920		Soma dos ativos da Carteira			1,00				
17	Desvio padrão da Carteira		38,63%		Retorno esperado da Carteira			20,00%				
18												

FIGURA 9.12 Carteira com retorno definido, restrição de venda e mínima variância.

Capítulo 10

DISTRIBUIÇÃO AMOSTRAL



Com a média \bar{X} de uma amostra extraída de uma população será estimada a média μ dessa população. No entanto, como você pode imaginar, de uma mesma população pode-se tomar muitas amostras diferentes do mesmo tamanho. A principal preocupação em uma inferência estatística é obter conclusões sobre a população e não sobre a amostra, como é possível observar nos seguintes exemplos:

- Determinação da vida média de uma lâmpada fluorescente especificada pelo fabricante. Essa determinação pode fazer parte:
 - Do procedimento de controle de qualidade da empresa. Se a vida média das lâmpadas fluorescentes de uma amostra retirada de um lote de produção não atender à especificação estabelecida, então o lote deverá ser rejeitado.
 - Do procedimento de um órgão de defesa do consumidor. Se a vida média das lâmpadas fluorescentes da amostra retirada de diversos pontos de venda atender à especificação do fabricante, então a reclamação dos consumidores não deverá ser aceita.
- Avaliação de um novo produto. Antes de seu lançamento, em muitos casos, o novo produto é distribuído a um grupo de consumidores potenciais que respondem a um questionário. Se os resultados dos questionários mostrarem que o novo produto foi bem aceito, então o grupo de marketing terá suporte para defender o lançamento desse novo produto.
- Previsão do tempo médio de espera dos clientes no caixa de um banco. Se o tempo médio de espera de uma amostra de clientes for maior do que o tempo médio afirmado pelo gerente da agência, então será bastante provável que as reclamações dos clientes tenham fundamento.

Um denominador comum nos três casos apresentados é que as decisões que deverão ser tomadas serão apoiadas em informações incompletas. No dia a dia, estamos acostumados a tomar decisões com informações incompletas suportadas pela própria experiência ou a partir de amostras. Por exemplo, o procedimento de degustar uma porção de fruta ou queijo antes de comprar. Ao aprovar a amostra degustada e comprar uma quantidade da fruta ou do queijo, estamos aceitando que o resto do lote de fruta ou peça de queijo tem a mesma característica que apreciamos na amostra; entretanto, a experiência mostra que é mais fácil acertar no caso do queijo do que no da fruta, salvo que o pedaço de queijo comprado seja de outra peça não amostrada. Qualquer que for a decisão tomada, estará sendo aplicada a *distribuição das médias das amostras*.

Formação da distribuição

A coordenadora do ensino de primeiro grau tem interesse em conhecer a estatura média dos alunos da primeira série da rede escolar. Se a variável estatura estivesse registrada no cadastro dos alunos, seria fácil calcular a média das estaturas dos alunos da primeira série. Contudo, essa informação não está disponível. Uma tentativa de alcançar o objetivo é estimar a média de todos os alunos utilizando a média de uma amostra dos alunos da primeira série, tendo presente que essa amostra será representativa da população; isto é, a amostra possuirá características similares às que seriam observadas na população se estivesse disponível. Para testar a ideia, a coordenadora preparou dez funcionários com a tarefa individual de selecionar aleatoriamente trinta alunos da primeira série da escola designada, medir a estatura dos trinta alunos e finalmente calcular e registrar a média dessa amostra. Terminada a tarefa, a coordenadora receberá as dez médias amostrais $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_{10}$ que, em geral, serão diferentes entre si devido à *variabilidade amostral* e, também, essas dez médias deverão ser diferentes da média da população.

Parâmetro é uma medida numérica que descreve uma população.

Estatística é uma medida numérica que descreve uma amostra.

Analisando o resultado da tarefa dos dez funcionários, pode-se observar:

- O parâmetro *média* da população é um valor único e desconhecido.
- A estatística *média* da amostra é um valor conhecido, porém pode variar de amostra para amostra. Se os dez funcionários realizarem novas amostragens aleatórias do mesmo tamanho, as médias das novas amostras não deverão ser iguais às dez primeiras. Apesar da média da população não ter mudado, a média da amostra dependerá de cada amostra.

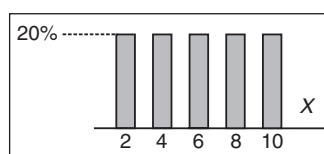
Com as médias das amostras, é possível construir a distribuição de frequências das médias das amostras denominada *distribuição amostral*, cuja média se denomina *média amostral* e seu desvio padrão, *erro padrão*. Embora os parâmetros *média* e *desvio padrão* da população não sejam conhecidos, para ajudar na compreensão da *distribuição amostral*, inicialmente, esses parâmetros serão considerados conhecidos.

EXEMPLO 10.1

Suponha que uma urna contém cinco bolas numeradas com os números 2, 4, 6, 8 e 10. Considerando que da urna serão retiradas amostras com reposição de tamanho $n=2$, o objetivo é determinar o valor esperado da média de todas as combinações possíveis de serem formadas.

Solução. Analisemos a população formada pelas cinco bolas {2, 4, 6, 8, 10} dentro da urna.

- Cada uma das cinco bolas tem a mesma probabilidade de ser escolhida, sendo 20% a probabilidade de uma bola ser escolhida.
- A distribuição de frequências relativas da população {2, 4, 6, 8, 10} é uma distribuição discreta e uniforme com média igual a seis, $\mu=6$, como mostra o histograma da figura seguinte.

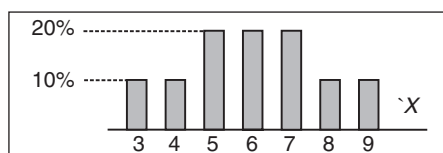


Analisemos o experimento de retirar amostras de tamanho $n=2$ com reposição. O número total de amostras diferentes é dez, resultado obtido da contagem das combinações de cinco bolas tomadas duas a duas. A primeira linha da tabela seguinte registra todas as combinações possíveis, ou amostras de tamanho dois, que podem ser extraídas da população.

Amostras	2, 4	2, 6	2, 8	2, 10	4, 6	4, 8	4, 10	6, 8	6, 10	8, 10
Média \bar{X}	3	4	5	6	5	6	7	7	8	9

A segunda linha da tabela registra as médias \bar{X} das amostras correspondentes à mesma coluna da primeira linha da tabela. A tabela seguinte registra a distribuição de frequências das médias das amostras \bar{X} , tendo presente que todas as amostras são igualmente prováveis.

Média \bar{X}	3	4	5	6	7	8	9
$P(\bar{X})$	0,1	0,1	0,2	0,2	0,2	0,1	0,1



Os resultados da tabela anterior formam o histograma da *distribuição amostral*, registrado na figura seguinte. Essa tabela define a variável aleatória \bar{X} cujo valor esperado $E[\bar{X}] = 6$ é obtido da expressão já conhecida:

$$E[\bar{X}] = 3 \times 0,1 + 4 \times 0,1 + 5 \times 0,2 + 6 \times 0,2 + 7 \times 0,2 + 8 \times 0,1 + 9 \times 0,1$$

$$E[\bar{X}] = 6$$

Os resultados do Exemplo 10.1 permitem obter as seguintes conclusões:

- A média da população é seis, $\mu=6$. A média μ da população é um valor único e constante.
- Cada amostra de tamanho $n=2$ tem sua própria média \bar{X} . A média da amostra depende de cada amostra extraída. Qualquer inferência realizada sobre a média da população utilizando uma única amostra estará sujeita a alguma incerteza, pois a média de cada amostra pode ser diferente.
- A média das dez médias amostras é igual a seis, $\mu_{\bar{X}} = 6$. A média amostral é uma média de longo prazo.
- A média amostral $\mu_{\bar{X}}$ coincide com a média da população μ e, neste exemplo, $\mu_{\bar{X}} = \mu = 6$. Temos a primeira conclusão importante: *a média das médias das amostras é a própria média da população*.¹

Definição da distribuição amostral

Dos histogramas do Exemplo 10.1, pode-se ver que a distribuição de frequências das médias das amostras, denominada *distribuição amostral*, é diferente da distribuição de frequências da população. Enquanto a distribuição da população é uniforme, a distribuição amostral não é uniforme, apresentando forma simétrica com tendência para a distribuição normal. Quanto à variabilidade, a média amostral varia no intervalo (3, 9), enquanto a população varia no intervalo (2, 10). A dispersão da distribuição amostral é menor do que a dispersão da população.

¹ Esta afirmação, baseada nos resultados de apenas um exemplo, é formalmente demonstrada, assunto que não é abordado neste livro.

Média e desvio padrão da distribuição amostral

A distribuição da população X é normal, com média μ e variância σ^2 . Dessa população, é retirada a amostra probabilística $\{X_1, X_2, \dots, X_i, \dots, X_n\}$ de tamanho n . A média \bar{X} dessa amostra é a combinação linear de n variáveis aleatórias independentes com a função $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_i + \dots + X_n)$. No Capítulo 9, vimos que a média de uma combinação linear é a soma das médias das variáveis.

$$\mu_{\bar{X}} = \mu_{\left(\frac{1}{n}(X_1 + X_2 + \dots + X_i + \dots + X_n)\right)} = \frac{1}{n}\mu_{X_1} + \frac{1}{n}\mu_{X_2} + \dots + \frac{1}{n}\mu_{X_n}$$

Como as variáveis aleatórias X_i têm distribuição normal com média μ e desvio padrão σ :

$$\mu_{\bar{X}} = \frac{1}{n}\mu + \frac{1}{n}\mu + \dots + \frac{1}{n}\mu = \frac{1}{n}n \times \mu$$

$$\mu_{\bar{X}} = \mu$$

Da mesma maneira, a variância da combinação linear de n variáveis aleatórias independentes $\sigma_{\bar{X}}^2$ é obtida da expressão:

$$\sigma_{\bar{X}}^2 = \sigma_{\left(\frac{1}{n}(X_1 + X_2 + \dots + X_i + \dots + X_n)\right)} = \frac{1}{n^2}\sigma_{X_1}^2 + \frac{1}{n^2}\sigma_{X_2}^2 + \dots + \frac{1}{n^2}\sigma_{X_n}^2$$

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{1}{n^2}n\sigma^2$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

O desvio padrão ou *erro padrão* é obtido da variância descrita anteriormente:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Média e Desvio Padrão da Distribuição Amostral

As médias \bar{X} das amostras de tamanho n retiradas de uma população, com média μ e desvio padrão σ , formam a distribuição amostral com os seguintes parâmetros:

- O valor esperado $\mu_{\bar{X}}$ é igual à média da população: $\mu_{\bar{X}} = \mu$.
- O desvio padrão das médias amostrais $\sigma_{\bar{X}}$ é igual a: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

EXEMPLO 10.2

A estatura dos alunos da primeira série do primeiro grau tem distribuição normal, com média de 105 centímetros e desvio padrão de 26 centímetros. Qual a média e o erro padrão de uma amostra aleatória de 26 retirada dessa população de alunos?

Solução. Como a média amostral é igual à média da população, deduzimos que $\mu_{\bar{X}} = 105$. Da mesma maneira, o desvio padrão ou *erro padrão* é igual a 5,81 centímetros, resultado obtido com a fórmula:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{X}} = \frac{26}{\sqrt{20}} = 5,81$$

A expressão do $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ mostra que o erro padrão da distribuição de \bar{X} diminui quando aumenta o tamanho da amostra n . Isso significa que à medida que n aumenta e mais informações são utilizadas, a média da amostra \bar{X} se aproxima da média da população μ .

Forma da distribuição de \bar{X}

Ao estudar as medidas estatísticas descritivas, foi observado que a forma da distribuição da variável é importante. Observe que a distribuição amostral do Exemplo 10.1 é simétrica, embora a distribuição da população seja uniforme. De forma geral, a forma da distribuição amostral depende da forma da distribuição da população.

- Se a distribuição da população for normal $N(\mu, \sigma)$, a distribuição amostral \bar{X} também será normal $N(\mu, \sigma/\sqrt{n})$ qualquer que seja o tamanho n da amostra.
- Se a distribuição da população não for normal, à medida que o tamanho da amostra aumentar, a distribuição da média amostral se aproximará da distribuição normal. Pelo teorema central do limite, a distribuição das médias \bar{X} de amostras de tamanho suficientemente grande² poderá ser considerada normal $N(\mu, \sigma/\sqrt{n})$, qualquer que seja a forma da distribuição da população.

Teorema Central do Limite

Se de uma população com parâmetros (μ, σ) for retirada uma amostra de tamanho n suficientemente grande, a distribuição de \bar{X} será aproximadamente normal $N(\mu, \sigma/\sqrt{n})$, qualquer que seja a forma da distribuição da população.

O teorema central do limite é muito importante, pois permite utilizar a distribuição normal para realizar inferências da média amostral \bar{X} , qualquer que seja a forma da distribuição da população. Como aplicação prática, podemos dizer que a soma de um número de efeitos aleatórios, sem dominância de nenhum deles sobre o resultado total, produz uma variável aleatória com distribuição normal. Por exemplo:

- No Capítulo 8, vimos que um cabo de aço trançado utilizado em um elevador é formado por muitos fios de aço que, adequadamente entrelaçados, conferem uma forte resistência ao cabo, cuja capaci-

² Para amostras com $n > 30$. Se a distribuição da população for aproximadamente simétrica, a amostra poderá ter menos de 30 valores, por exemplo, 15 a 25.

dade é igual à soma das capacidades individuais dos fios de aço. Se o número de fios que formam o cabo for adequadamente grande, apesar de a distribuição da capacidade dos fios de aço não ser normal, a distribuição da capacidade do cabo será normal.

- No Capítulo 9, destacamos que você deve esperar que a função H definida como soma de variáveis aleatórias tenha distribuição normal, considerando que nenhuma das variáveis domina as restantes. Pelo teorema central do limite, se as n variáveis aleatórias que participam da combinação linear tiverem distribuição normal, então a nova variável H também terá distribuição normal, para qualquer valor de n . Todavia, se as n variáveis aleatórias não tiverem distribuição normal, então a nova variável H terá distribuição normal se o número de variáveis aleatórias for adequadamente grande.

Simulador teorema central do limite

O teorema central do limite pode ser verificado na prática, realizando simulações repetidas de um experimento, como o lançamento de um dado que tem seis resultados possíveis $\{1, 2, 3, 4, 5, 6\}$, com a mesma probabilidade $1/6$ de ocorrer. O histograma desse experimento mostra que sua distribuição de frequências é discreta e uniforme, com média $\mu = 3,50$ e variância $\sigma^2 = 2,917$ e desvio padrão $\sigma = 1,71$, resultados obtidos no intervalo de células B2:G6 da planilha **Teorema Central do Limite**, incluída na pasta **Capítulo 10**, como mostra a Figura 10.1.

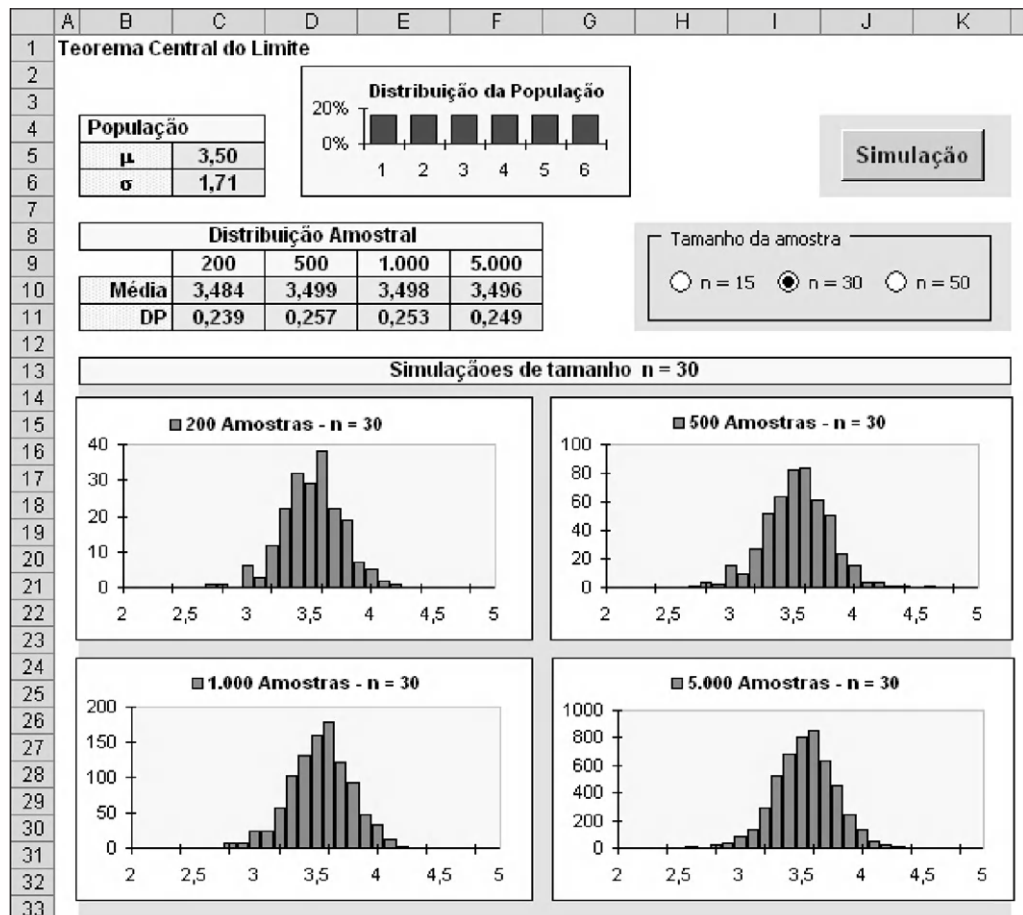


FIGURA 10.1

Simulação Teorema Central do Limite para $n=10$.

O modelo permite realizar simulações para três tamanhos de amostra, $n=15$, 30 e 50 resultados aleatórios de um dado, obtidos com a fórmula $=\text{ARRED}(\text{ALEATÓRIO}() * (6-1) + 1; 0)$, registrada em cada

célula da tabela de trinta lançamentos de um dado repetido cinco mil vezes. A última coluna da tabela registra a média \bar{X} de cada lançamento de $n=15, 30$ e 50 dados. Cada vez que o botão **Simulação** for pressionado, o *modelo* atualiza os seguintes resultados para o tamanho n previamente definido:

- Calcula a média e o desvio padrão das médias amostrais, intervalo de células C10:F11.
- Calcula a tabela de frequências absolutas e apresenta o histograma de 200, 500, 1.000 e 5.000 lançamentos, como mostra a Figura 10.1 no caso de $n=30$.

Os quatro histogramas mostram que, ao aumentar o número de simulações, acentua-se a concentração das médias amostrais ao redor da média da população. Também, se for aumentado o tamanho da amostra para $n=50$, essa concentração ao redor da média começa a apresentar menos simulações, confirmando que à medida que n aumenta, têm-se mais informações, como mostra a Figura 10.2.

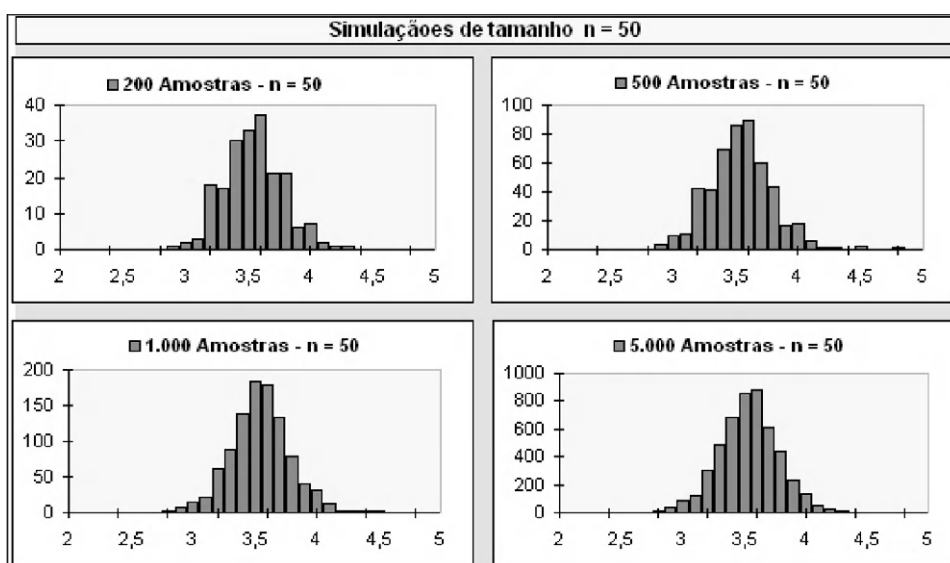


FIGURA 10.2
Simulação Teorema Central do Limite para $n=30$.

Repetindo o que já foi apresentado, a soma de um número de efeitos aleatórios, sem dominância de nenhum deles sobre o resultado total, produz uma variável aleatória com distribuição normal.

Correção pela população ser finita

As premissas incluídas nas expressões apresentadas estabelecem que a população é suficientemente grande, os valores da amostra são independentes e a amostragem é realizada com reposição. Se numa população pequena for realizada uma amostragem sem reposição de tamanho maior do que 5% do tamanho da população, no cálculo do erro padrão deverá ser incluído o fator de correção finita $\sqrt{\frac{N-n}{N-1}}$.

Se a relação entre o tamanho da população N e o tamanho da amostra n for menor do que 20, o erro padrão $\sigma_{\bar{X}}$ deverá ser calculado com a fórmula $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$,

sendo $\sqrt{\frac{N-n}{N-1}}$ denominado *fator de correção finita*.

A tabela seguinte mostra valores do coeficiente de correção finita em função do tamanho n da amostra numa população de tamanho $N=1.000$.

$N=1.000$	
n	Fator
5	0,9980
10	0,9955
30	0,9854
50	0,9752
100	0,9492
250	0,8665
500	0,7075

Análise da média de uma amostra

Quanto maior for o tamanho n da amostra, mais a média amostral será próxima da média da população, pois à medida que n aumenta, o erro padrão diminui e, no limite, quando n tender à própria população, o erro padrão tenderá a zero. As propriedades da distribuição amostral \bar{X} asseguram que a média de uma amostra é uma boa estatística para ser utilizada na inferência da média da população μ da qual foi extraída. Ao mesmo tempo, o teorema central do limite estabelece que se o tamanho da amostra n for suficientemente grande, a distribuição da média amostral será normal, qualquer que seja a forma da distribuição da população. Portanto, o teorema central do limite permite aplicar a distribuição normal para obter respostas da média de uma amostra de tamanho suficientemente grande retirada de uma população qualquer.

Se $\{X_1, X_2, \dots, X_i, \dots, X_n\}$ é uma amostra aleatória extraída de uma população infinita com média μ e variância σ^2 , então, para n suficientemente grande, a distribuição de $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ é normal padronizada.³

EXEMPLO 10.3

Os registros históricos de produção de frascos com detergente mostram que o volume de enchimento realizado pela máquina automática é normalmente distribuído com média igual a 150 e desvio padrão de 0,50 centímetro cúbico. Se for retirada uma amostra de tamanho $n=9$, qual a probabilidade da média dessa amostra ser menor ou igual a 149,75 centímetros cúbicos?

Solução. Deve-se encontrar a probabilidade $P(\bar{X} \leq 149,75)$. Sabemos que a distribuição amostral \bar{X} é normal, com média igual à média da população μ e desvio padrão σ/\sqrt{n} . O resultado $P(\bar{X} \leq 149,75)=0,0668$ foi obtido registrando a fórmula =DIST.NORM(149,75;150;0,5/RAIZ(9);VERDADEIRO) numa célula vazia da planilha Excel. Portanto, a probabilidade da média da amostra de tamanho $n=9$ ser menor ou igual a 149,75 centímetros cúbicos é 6,68%. Esse resultado também pode ser obtido com a distribuição normal padronizada, calculando primeiro o desvio padrão normalizado $Z=1,50$, resultado obtido com a fórmula:

³ Freund J. E. – *Mathematical Statistics* – Prentice Hall, 5ª edição, 1992.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{149,75 - 150}{\frac{0,5}{\sqrt{9}}} = -1,50$$

O resultado $P(\bar{X} \leq 149,75) = P(Z \leq -1,5) = 0,0668$ pode ser obtido com a *tabela Z*, procedimento que deixamos por conta do leitor, ou com a função estatística do Excel =DIST.NORMP(-1,5).

- Pode ser utilizada a fórmula =DIST.NORMP((149,75-150)/(0,5/RAIZ(9))) ou a fórmula =DIST.NORMP(PADRONIZAR(149,75;150;0,5/RAIZ(9))), que é equivalente à anterior. Essas duas fórmulas evitam o cálculo intermediário do desvio normal padronizado Z .

O Exemplo 10.3 mostra o procedimento de cálculo com a média da amostra, sendo conhecidos a média μ e o desvio padrão σ da população.

- Da população é retirada uma amostra aleatória com média \bar{X} e tamanho n suficiente para atender às premissas do teorema do limite central e para a distribuição amostral seja normal.
- A média da distribuição amostral é igual a média da população $\mu_{\bar{X}} = \mu$ e o erro padrão igual a $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- Com a distribuição amostral $N(\mu, \sigma/\sqrt{n})$ são calculadas probabilidades do tipo $P(\bar{X} \leq \mu)$ ou $P(\bar{X} \geq \mu)$.

Quanto ao resultado do Exemplo 10.3, qual o significado da probabilidade $P(\bar{X} \leq 149,75) = 0,0668$? Tendo presente que a distribuição amostral fornece a distribuição teórica de todas as possíveis amostras de tamanho nove, a probabilidade 6,68% é a proporção de amostras de tamanho nove que têm média menor ou igual a 149,75 centímetros cúbicos.

EXEMPLO 10.4

Repita o Exemplo 10.3, considerando que a média $\bar{X}=149,75$ foi obtida de uma amostra com tamanho $n=25$.

Solução. Para $n=25$, obtém-se $P(\bar{X} \leq 149,75) = 0,0062$, resultado obtido com a fórmula =DIST.NORM(149,75;150;0,5/RAIZ(25);VERDADEIRO). Portanto, a probabilidade da média da amostra de tamanho $n=25$ ser menor ou igual a 149,75 centímetros cúbicos é 0,62%. Esse resultado também pode ser obtido com a distribuição normal padronizada, calculando primeiro o desvio padrão normalizado $Z=2,50$, resultado obtido com a fórmula:

$$Z = \frac{149,75 - 150}{\frac{0,5}{\sqrt{25}}} = -2,5$$

O resultado $P(\bar{X} \leq 149,75) = P(Z \leq -2,5) = 0,0062$ pode ser obtido com a *tabela Z*, procedimento que também deixamos por sua conta, ou com a função estatística do Excel =DIST.NORMP(-2,5). O Exemplo 10.3 mostra outras formas de utilizar essa última função estatística.

Qual o significado do valor de probabilidade 0,62% do Exemplo 10.4? Tendo presente que a distribuição amostral fornece a distribuição teórica de todas as possíveis amostras de tamanho 25, a probabilidade 0,62% é a proporção de amostras de tamanho 25, que têm média menor ou igual a 149,75 centímetros cúbicos.

Vale perguntar: a probabilidade 0,62% confirma que o volume de enchimento realizado pela máquina automática seja 150 centímetros cúbicos?

- Se for retirado apenas um único frasco de detergente como amostra, a probabilidade de o volume desse frasco ser menor ou igual a 149,75 centímetros cúbicos é $P(\bar{X} \leq 149,75) = 0,3085$, resultado obtido com a fórmula =DIST.NORM(149,75;150;0,5;VERDADEIRO). A probabilidade 30,85% é muito alta, comparada com as probabilidades com a média de uma amostra dos exemplos anteriores.
- A probabilidade $P(\bar{X} \leq 149,75) = 6,68\%$ da amostra do Exemplo 10.3 não é suficiente para afirmar que a máquina esteja enchendo realmente com média igual a 150, pois a proporção de amostras de tamanho nove com média menor ou igual a 149,75 centímetros cúbicos não é pequena.
- A probabilidade $P(\bar{X} \leq 149,75) = 0,62\%$ da amostra do Exemplo 10.4 é pequena, o que nos leva a aceitar que o volume médio de enchimento da máquina seja 150.
- A aceitação do volume de enchimento da máquina automática depende do tamanho da amostra n , pois quanto maior for n , maior será a chance de aceitação.

O erro padrão é uma medida da distância entre a média da amostra \bar{X} e a média da população μ . A aceitação do lote de produção dependerá do tamanho da amostra n , pois quanto maior for n , maior será a chance de aceitar o lote. Como o tamanho da amostra influencia diretamente o erro padrão da distribuição amostral, quanto maior for o tamanho da amostra n , menor será o erro padrão e, no limite, o erro padrão tenderá a zero, pois n tenderá a ser a própria população.

EXEMPLO 10.5

O peso das latas de pêssego em calda tem distribuição normal com média 1.000 gramas e desvio padrão 40 gramas. Se for retirada uma amostra de 12 latas de um lote grande de latas, qual a probabilidade de haver uma média amostral menor do que 975 gramas?

Solução. Deve-se calcular a probabilidade $P(\bar{X} \leq 975)$. Procedendo como nos exemplos anteriores, a probabilidade de ter uma média amostral menor do que 975 gramas é 1,52% ou $P(\bar{X} \leq 975) = 0,0152$. Esse resultado pode ser obtido com a fórmula =DIST.NORM(975;1000;40/RAIZ(12);VERDADEIRO), registrada numa célula vazia da planilha Excel. Esse resultado também pode ser obtido com a distribuição normal padronizada e a tabela Z e com a função DIST.NORMMP. Outra forma de obter esse resultado é utilizando o modelo **Distribuição Amostral**.

Modelo distribuição amostral

Com o modelo construído na planilha **Distribuição Amostral**, incluído na pasta **Capítulo 10**, é possível resolver o Exemplo 10.5 e os dois anteriores, como mostra a Figura 10.3. Os dados devem ser informados nas células pintadas de cor azul, pois as células de cor verde somente retornam resultados e as células de cor amarela apresentam títulos. No *modelo*:

- Nas células do intervalo C3:C7, são registrados os dados numéricos requeridos para os cálculos.
- A *Média da Amostra* registrada na célula C7 é utilizada para definir os limites na caixa de combinação da célula B7. Nessa caixa de combinação, é possível escolher os limites \leq e \geq , utilizados para o cálculo da probabilidade requerida.
- Nas células do intervalo C9:C11, são apresentados resultados intermediários e o resultado final da probabilidade requerida.

O gráfico do *modelo* apresenta duas distribuições normais, automaticamente identificadas pelos seus respectivos parâmetros. A distribuição com maior dispersão é a distribuição da população, desenhada em cor azul, e a outra é a distribuição das médias amostrais. Conforme os dados do problema que está sendo resolvido, você deverá ajustar a escala do eixo de abscissas pressionando o botão **Ajuste**

escala para um ajuste inicial. O ajuste fino da escala deverá ser feito no próprio eixo, clicando duas vezes em cima do eixo ou nos valores do eixo, depois de desproteger a planilha. As curvas das duas distribuições no mesmo gráfico ajudam a compreender a diferença entre a probabilidade do evento com amostra e a probabilidade do evento com apenas uma unidade, bem como o comportamento das duas curvas para outros valores de desvios padrão. Analisar o comportamento das curvas para diferentes tamanhos de amostra auxiliará a compreender o procedimento de amostragem. Para terminar, informando o tamanho de amostra igual ao valor um o modelo se comporta, parcialmente, como o **MODELO DN** do Capítulo 8.

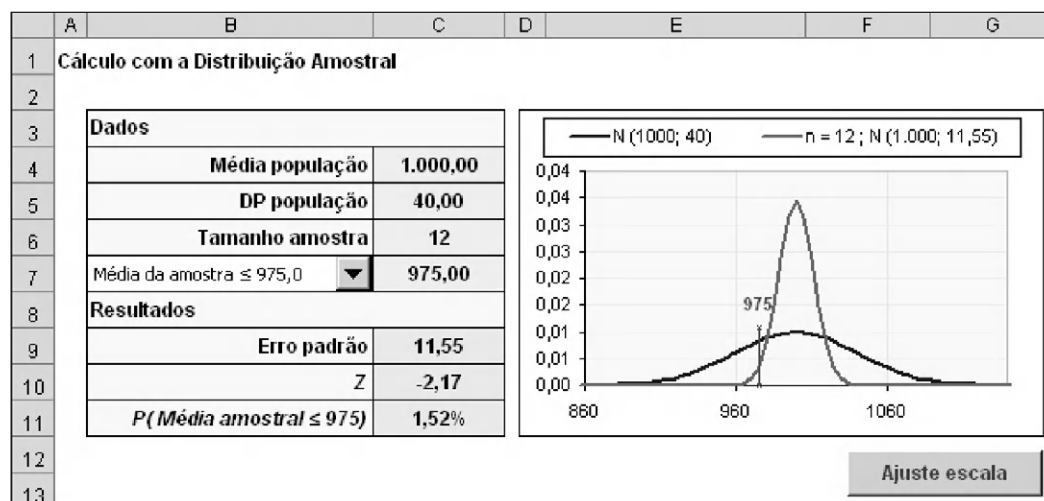


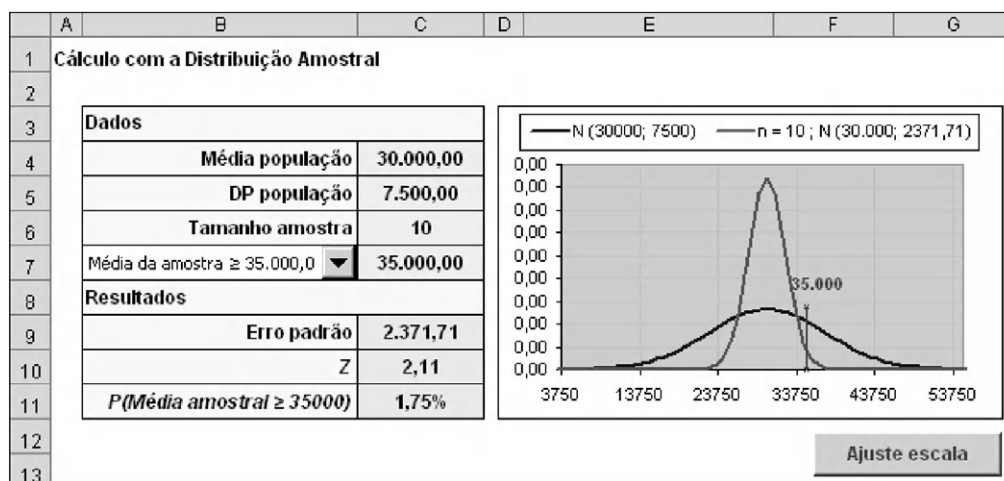
FIGURA 10.3 Modelo Distribuição Normal resolvendo o Exemplo 10.5.

A linha vertical vermelha do gráfico representa a média da amostra. Essa reta permite comparar as áreas das duas distribuições referentes às probabilidades de exceder o valor dessa média da amostra.

EXEMPLO 10.6

A comissão anual das vendedoras domiciliares de uma importante empresa de cosméticos tem distribuição normal com média \$35.000 e desvio padrão de \$7.500. Se aleatoriamente for escolhida uma amostra de dez vendedoras domiciliares, calcule a probabilidade que a média dessa amostra seja:

- Mais do que \$35.000 por ano.
- Menos do que \$29.000 por ano.



Solução. Os resultados $P(\bar{X} \geq 35.000) = 1,75\%$ e $P(\bar{X} \leq 29.000) = 33,66\%$ foram obtidos com o modelo **Distribuição Amostral**, como mostra a figura referente à primeira questão.

EXEMPLO 10.7

Continuando com o Exemplo 10.5. Determine o intervalo ao redor da média da população que inclua 90% das médias das amostras de tamanho $n=16$.

Solução. O peso das latas de pêssego em calda tem distribuição normal, com média 1.000 gramas e desvio padrão 40 gramas. Os 90% das médias das amostras de tamanho $n=16$ distribuídas ao redor da média da população definem as duas caudas da distribuição, com 5% cada uma, considerando o erro padrão igual a 10 gr. O limite inferior que define a cauda inferior é 983,55 gramas, resultado obtido com a fórmula $=\text{INV.NORM}(0,05; 1000; 40/\text{RAIZ}(16))$. Para a cauda superior, registramos a fórmula $=\text{INV.NORM}(0,95; 1000; 40/\text{RAIZ}(16))$, que retorna o limite superior igual a 1.016,45 gramas. Com esses resultados, podemos dizer que 90% das médias dos pesos das amostras de tamanho 16 se encontram entre os valores 983,55 gramas e 1.016,45 gramas.

Esses resultados também podem ser obtidos, com o modelo **Distribuição Amostral**, com ajuda do comando **Atingir Meta** do Excel. Começamos por registrar os dados da população e o tamanho da amostra, registrando também um valor qualquer na célula C7, que é a incógnita de nosso exemplo, como mostra a figura a seguir. Posicione o cursor do Excel na célula C11 do modelo.

- No menu **Ferramentas** do Excel, escolher **Atingir meta**.
- Na caixa de diálogo **Atingir meta**, informar os dados solicitados.

	A	B	C	D	E	F
1	Cálculo com a Distribuição Amostral					
2						
3		Dados				
4		Média população	1.000,00			
5		DP população	40,00			
6		Tamanho amostra	16			
7		Média da amostra $\leq 1.000,0$	1.000,00			
8		Resultados				
9		Erro padrão	10,00			
10		Z	0,00			
11		$P(\text{Média amostral} \leq 1000)$	50,00%			
12						

Atingir meta

Definir célula:

Para valor:

Alternando célula:

OK Cancelar

- Clique em **OK** e o comando **Atingir meta** encontrará a solução. O Excel apresentará a caixa de diálogo **Status do comando attingir meta** com o resultado da célula C7 igual a 983,55 e, ao mesmo tempo, observe que na célula C11 está registrado o valor 5%.

	A	B	C	D	E	F	G
1	Cálculo com a Distribuição Amostral						
2							
3		Dados					
4		Média população	1.000,00				
5		DP população	40,00				
6		Tamanho amostra	16				
7		Média da amostra $\leq 983,6$	983,55				
8		Resultados					
9		Erro padrão	10,00				
10		Z	-1,64				
11		$P(\text{Média amostral} \leq 983,55)$	5,00%				
12							

Status do comando attingir meta

Atingir Meta com a célula C11 encontrou uma solução.

Valor de destino: 0,05

Valor atual: 5,00%

OK Cancelar Etapa Pausar

Para terminar, clique em **OK**. O valor encontrado, 983,55, será registrado na célula C7. Se você clicar em **Cancelar**, o valor original 1.000 permanecerá na célula C7. Na obtenção do resultado desejado, o comando **Atingir Meta** utiliza um processo iterativo conhecido como *método de tentativa e erro*, apoiado por *algoritmos* de convergência. Lembre-se de que o resultado será obtido com o grau de exatidão definido no menu do Excel **Ferramentas – Opções – Cálculo** na caixa **Nº máx. de alterações**. Ainda, embora não se aplique neste exemplo, alguns problemas podem ter mais de uma solução, e o comando **Atingir meta** apresenta apenas uma delas. Para terminar, repita o procedimento para o limite superior procurando a probabilidade de 95%.

Problemas

Problema 1

Na população formada pelos números {1, 2, 3, 4}.

- Determine a quantidade de amostras com reposição de tamanho $n=2$ que podem ser formadas com essa população.
- Identifique todas as amostras de tamanho $n=2$ e calcule suas médias.
- Compare a média da população com a média das médias das amostras.

R: a) A quantidade de amostras de tamanho $n=2$ é igual a 6. b) Por sua conta. c) A média da população e a média amostral são iguais a 2,50.

Problema 2

Na população formada pelos números {20, 22, 24, 26, 28}.

- Construa o histograma da população.
- Determine a quantidade de amostras com reposição de tamanho $n=2$ que podem ser formadas com a população.
- Identifique todas as amostras de tamanho $n=2$ e calcule suas médias.
- Construa e analise o histograma da distribuição amostral.
- Compare a média e o desvio padrão da população com os das amostras.

R: a) Por conta do leitor. b) A quantidade de amostras de tamanho $n=2$ é igual a 10. c) Por sua conta. d) Por conta do leitor. e) A média da população e a média amostral são iguais a 24. A dispersão da população é maior do que a da distribuição amostral.

Problema 3

A tabela a seguir registra a média e o desvio padrão de quatro populações. Calcule a média amostral e o erro padrão de uma amostra de tamanho $n=100$.

População		Distribuição amostral	
Média	Desvio padrão	Média	Erro padrão
10	20	10	2
20	10	20	1
50	300	50	30
100	200	100	20

Problema 4

A média e o desvio padrão da população são, respectivamente, 60 e 10. Calcule a média amostral e o erro padrão de amostras aleatórias de tamanho $n=10, 25, 50, 75, 100, 500, 1.000$.

R: A média para todos os tamanhos é a mesma e igual a 60. Os desvios padrão são (na mesma ordem): 3,16; 2; 1,41; 1,15; 1; 0,45; 0,32.

Problema 5

O gerente da agência bancária verificou que o saldo médio das contas correntes aumentou. Considerando todos os clientes da agência, a média e o desvio padrão do saldo médio das contas correntes são, respectivamente, \$325 e \$114. Se for retirada uma amostra aleatória de 100 contas correntes, qual a probabilidade de a média dos saldos médios ser menor ou igual a \$330?

R: 66,95%

Problema 6

Continuando com o Problema 5, qual a probabilidade de a média dos saldos médios ser maior ou igual a \$350?

R: 1,42%

Problema 7

A garrafa de vinho branco importado é vendida na maior parte dos supermercados do país. Levantamentos realizados pelo distribuidor em todos os pontos de vendas mostraram que o preço médio de venda é USD \$15, com desvio padrão de USD \$2,50. Se for retirada uma amostra aleatória em 45 pontos de venda, qual a probabilidade de a média do preço de venda da garrafa de vinho ser menor ou igual a USD \$14?

R: 0,36%

Problema 8

Continuando com o Problema 7, qual a probabilidade de a média do preço de venda da garrafa de vinho ser maior ou igual a USD \$15,5.

R: 8,99%

Problema 9

O fabricante de pneus assegura que a duração do pneu mais vendido tem média de 60.000 km, com desvio padrão 5.000 km. Como os distribuidores não estão convencidos, o fabricante ofereceu aos revendedores a oportunidade de separar, aleatoriamente, 40 pneus para verificar os resultados afirmados pelo fabricante. Supondo que a afirmação do fabricante seja confirmada, descreva a distribuição da média amostral da duração do pneu.

R: A distribuição amostral é a normal $N(60.000, 790,57)$

Problema 10

Continuando com o Problema 9. Se a afirmação do fabricante for verdadeira, qual a probabilidade de os 40 pneus da amostra terem média de duração menor igual a 57.500 km? Analise o resultado.

R: 0,08%

Problema 11

Uma máquina de enchimento de latas de refrigerantes consegue produzir 12.000 latas por hora, com média de 330 ml e desvio padrão de 3 ml. Responda às seguintes questões, considerando a distribuição normal:

- Qual a probabilidade de uma lata conter menos de 328 ml?
- Qual a probabilidade de uma amostra aleatória de 30 latas conter menos de 328 ml?

R: a) 25,25% b) 0,01%

Problema 12

O retorno anual de todas as ações negociadas na Bolsa de Valores durante o ano passado apresentou média de 5,45% ao ano e desvio padrão 32,3% ao ano. Supondo que a distribuição dos retornos seja bem próxima da normal:

- Que percentagem de ações teve retorno nulo ou negativo?
- Se forem escolhidas cinco ações de forma aleatória do conjunto das ações negociadas na Bolsa de Valores, qual o retorno e o desvio padrão dessa carteira durante o ano passado?
- Qual a probabilidade do retorno dessa carteira ser menor ou igual a zero durante o ano passado?

R: a) 56,70% b) $\mu_{\bar{x}} = -5,45\%$ e $\sigma_{\bar{x}} = 14\%$ c) 64,70%

Problema 13

A montadora de carros afirma que a média de consumo do seu novo modelo tem distribuição normal com média de 15,9 km por litro e desvio padrão de 0,8 km por litro.

- Calcular a probabilidade da média de uma amostra de tamanho $n=25$ ser menor ou igual a 15,5 km/litro.
- Suponha que uma amostra aleatória de 25 carros fabricados na mesma época apresentou média amostral de 15 km/litro. Você acredita que a declaração da montadora deve ser aceita?

R: a) 0,62% b) As possíveis respostas são: 1) A média da amostra é possível, porém pouco provável; 2) Os parâmetros da população mudaram; 3) Há algum erro na afirmação ou nos resultados divulgados.

Problema 14

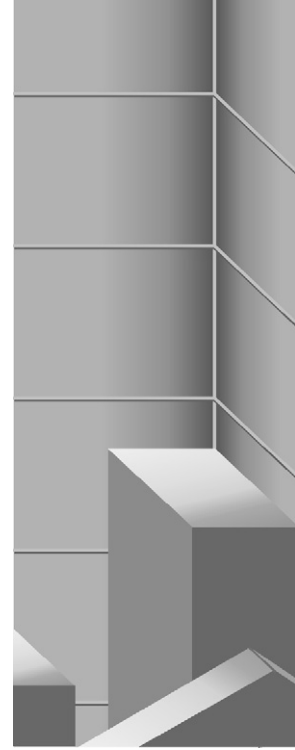
O fabricante da máquina de enchimento de refrigerantes afirma que o volume das garrafas tem média de 750 ml com desvio padrão de 16 ml. Numa amostra de 16 garrafas, qual a probabilidade de a média da amostra ser:

- Menos de 740 ml?
- Mais de 740 ml?
- Entre 742 e 758 ml?

R: a) 0,62% b) 99,38% c) 95,46%

Capítulo 11

ESTIMAÇÃO



O teorema central do limite apresentado no Capítulo 10 é muito importante, pois mostra como utilizar a distribuição normal para realizar inferências da média amostral \bar{X} , qualquer que seja a forma da distribuição da população. Como aplicação prática, dissemos que a soma de um número de efeitos aleatórios, sem dominância de nenhum deles sobre o resultado total, produz uma variável aleatória com distribuição normal. O teorema central do limite mostra que a média \bar{X} de uma amostra de tamanho n suficientemente grande, retirada de uma população com média μ e desvio padrão σ , tem distribuição normal com média igual à média da população $\mu_{\bar{X}} = \mu$ e desvio padrão $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Neste capítulo, mostraremos como *estimar* a média de uma população a partir de uma única amostra¹ aleatória retirada da população. Contudo, se as médias de amostras do mesmo tamanho extraídas de uma população, em geral, não coincidirão entre si nem com a média da população, que precisão devemos esperar de uma única amostra? Deveremos definir o erro máximo da estimativa e sua probabilidade de ocorrência. Há dois tipos de estimativas:

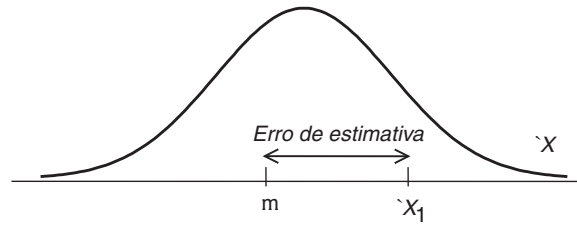
- **Estimativa pontual.** Estimativa pontual de um parâmetro da população é o valor obtido por cálculo de uma amostra retirada da população. Por exemplo, a média \bar{X} de uma amostra aleatória retirada de uma população é uma estimativa pontual da média da população μ .
- **Estimativa intervalar.** A estimativa está incluída num intervalo considerando um grau de acerto denominado *intervalo de confiança* que contém a estimativa pontual. Portanto, a média de uma amostra aleatória retirada de uma população é o valor inicial da média dessa população.

Confiança da estimativa

Para facilitar a compreensão do procedimento de estimativa da média, nesta análise inicial, a média μ e o desvio padrão σ da população serão considerados conhecidos. Se da população for retirada a amostra aleatória X_1 com média \bar{X}_1 , em geral, a média dessa amostra não coincidirá com a média da população μ , como mostra a distribuição normal das médias amostrais da Figura 11.1. A diferença entre a média μ e a média amostral \bar{X}_1 é denominada *erro de estimativa* ou *margem de erro* que pode ser medida a partir de qualquer um dos dois valores.

¹ As estimativas devem ser *não viesadas* (o valor esperado do estimador deve ser igual ao parâmetro da população) e devem ter *variância mínima* (menor valor de variância de todas as possíveis estimativas não viesadas).

FIGURA 11.1 Erro de estimativa.



A média \bar{X}_1 da amostra X_1 é uma boa estimativa da média da população μ , pois é uma amostra aleatória de tamanho n suficientemente grande e, como foi mostrado no Capítulo 10, a média da amostra X_1 tem distribuição normal com parâmetros $N(\mu, \sigma/\sqrt{n})$.

Intervalo de confiança

Devido à variabilidade amostral, as possíveis amostras aleatórias de mesmo tamanho retiradas da mesma população terão médias diferentes. Como estimar a média de uma população com apenas uma amostra? Qual a confiabilidade de uma estimativa pontual? O *intervalo de confiança* definirá de forma objetiva a credibilidade da estimativa.

Intervalo de confiança é o intervalo de valores que contém a média² da população com uma determinada probabilidade de acerto. O intervalo de valores é construído de uma amostra aleatória retirada da população.³

Se o tamanho da amostra n for suficientemente grande, a média \bar{X}_1 da amostra X_1 terá distribuição normal com média igual à da população e, como já foi apresentado no Capítulo 10:

- Da distribuição das médias amostrais, tem-se que $\mu_{\bar{X}} = \mu$ e $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.
- A probabilidade da média \bar{X}_1 estar incluída, por exemplo, no intervalo de dois desvios padrão ao redor da média é 95,44%, representado como $P(\bar{X}_1 \pm 2\sigma_{\bar{X}}) \leq 0,9544$.
- De outra maneira, 95,44% das médias amostrais se situam ao redor de mais ou menos dois desvios padrão.
- Portanto, a média da população μ estará incluída no intervalo de mais ou menos dois desvios padrão em 95,44% das vezes, considerando que o experimento será repetido um número grande de vezes.

Desse modo, o intervalo de confiança define o percentual de todas as amostras possíveis que satisfaz à margem de erro. Continuando com o exemplo anterior, retirando um número muito grande de amostras do mesmo tamanho, poderemos dizer que em 95,44% das vezes a média da população estará incluída no intervalo $\bar{X} \pm 2\sigma_{\bar{X}}$ ou:

$$\bar{X} - 2\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 2\sigma_{\bar{X}}$$

Definindo o *intervalo de confiança* de 95,44%, teremos:

$$P(\bar{X} - 2\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 2\sigma_{\bar{X}}) = 0,9544$$

² Esta definição se aplica a qualquer parâmetro da população.

³ A probabilidade de a média da população estar contida no intervalo de confiança é denominada, também, *nível de confiança* que é medido de forma unitária.

Como consequência direta desse resultado, devemos entender também que, em 4,56% das vezes, a média da população *não* estará incluída no intervalo definido por dois desvios padrão.

EXEMPLO 11.1

De uma população com média desconhecida e desvio padrão 16, foi retirada uma amostra aleatória de tamanho $n=64$ cuja média é 50. Qual a média da população, considerando o intervalo de confiança 95%?

Solução. O tamanho da amostra $n=64$ é suficiente para aceitar que a média da amostra \bar{X} tenha distribuição normal. Numa distribuição normal, a probabilidade 95% ao redor da média permite estabelecer o desvio padrão normalizado⁴ $Z=-1,96$, resultado obtido com a fórmula $=\text{INV.NORMP}(0,025)$ e $Z=+1,96$, resultado obtido com a fórmula $=\text{INV.NORMP}(0,975)$, e ambos os resultados expressos de forma geral $Z=\pm 1,96$. Da expressão de Z :

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A média da população μ é obtida com:

$$\mu = \bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

O símbolo \pm significa que a expressão inclui os dois limites de Z do intervalo. Substituindo os dados do exemplo, teremos:

$$\mu = 50 \pm 1,96 \frac{16}{\sqrt{64}} = 50 \pm 1,96 \times 2$$

$$\mu = 50 \pm 3,92$$

Analisemos o resultado:

- A média da população é $\mu = 50 \pm 3,92$, com intervalo de confiança de 95%.
- O *erro de estimativa* ou *margem de erro* da estimativa é 3,92.
- A média da população é um valor do intervalo de 46,08 até 53,92.
- A probabilidade de a média da população estar contida no intervalo $\pm 1,96$ desvios padrão normalizados ao redor da média amostral é 95%. Esse resultado permite afirmar que em 95% das vezes a média da população se situará entre 46,08 e 53,92

O erro de estimativa (ou margem de erro) e o intervalo de confiança são funções do tamanho da amostra. No Exemplo 11.1, foi definido o intervalo de confiança e o tamanho da amostra, ficando determinado o erro de estimativa. O Exemplo 11.2 mostrará o erro de estimativa, registrando o limite inferior e o limite superior correspondente, para cinco diferentes valores de intervalos de confiança do Exemplo 11.1.

EXEMPLO 11.2

Continuando com o Exemplo 11.1. Determinar os limites do erro de estimativa da média da população, considerando os intervalos de confiança 80%, 85%, 90%, 95% e 99%.

Solução. Os resultados de Z registrados na segunda coluna da tabela seguinte foram obtidos com a função esta-

⁴ Denominados, também, como valores críticos.

tística INV.NORMP. Por exemplo, para o intervalo de confiança 80%, o limite inferior Z é $-1,2816$, resultado obtido com a fórmula =INV.NORMP(0,1), e o limite superior Z é $+1,2816$, resultado obtido com a fórmula =INV.NORMP(0,9). Podemos simplificar o procedimento de cálculo tendo presente que, pela simetria da distribuição normal, poderíamos calcular apenas um dos dois limites e repetir o valor com sinal oposto para o outro limite. Os limites do erro de estimativa da média da população foram obtidos da mesma forma que os do Exemplo 11.1.

IC	$\pm Z$	Limite inferior	Limite superior
80%	$\pm 1,2816$	47,44	52,56
85%	$\pm 1,4395$	47,12	52,88
90%	$\pm 1,6449$	46,71	53,29
95%	$\pm 1,9600$	46,08	53,92
99%	$\pm 2,5758$	44,85	55,15

Analisando os resultados dos Exemplos 11.1 e 11.2, surge uma conclusão importante, afirmar com 90% de probabilidade que a média da população está incluída no intervalo de 46,71 até 53,29 pode ser *verdadeiro* ou *falso*, como mostra o Exemplo 11.3.

EXEMPLO 11.3

Para estimar a média da população, foram retiradas três amostras de tamanho $n=64$ e médias 50, 49 e 53. Analisar as três estimativas da média da população, considerando o intervalo de confiança 90% e sabendo que a média e o desvio padrão são conhecidos e iguais a, respectivamente, 50 e 10.

Solução. Como o intervalo de confiança 90% define os desvios padrão normalizados $Z=\pm 1,645$, a média da população é:

$$\mu = \bar{X} \pm 1,645 \frac{10}{\sqrt{64}} = \bar{X} \pm 1,645 \times 1,25$$

$$\mu = \bar{X} \pm 2,056$$

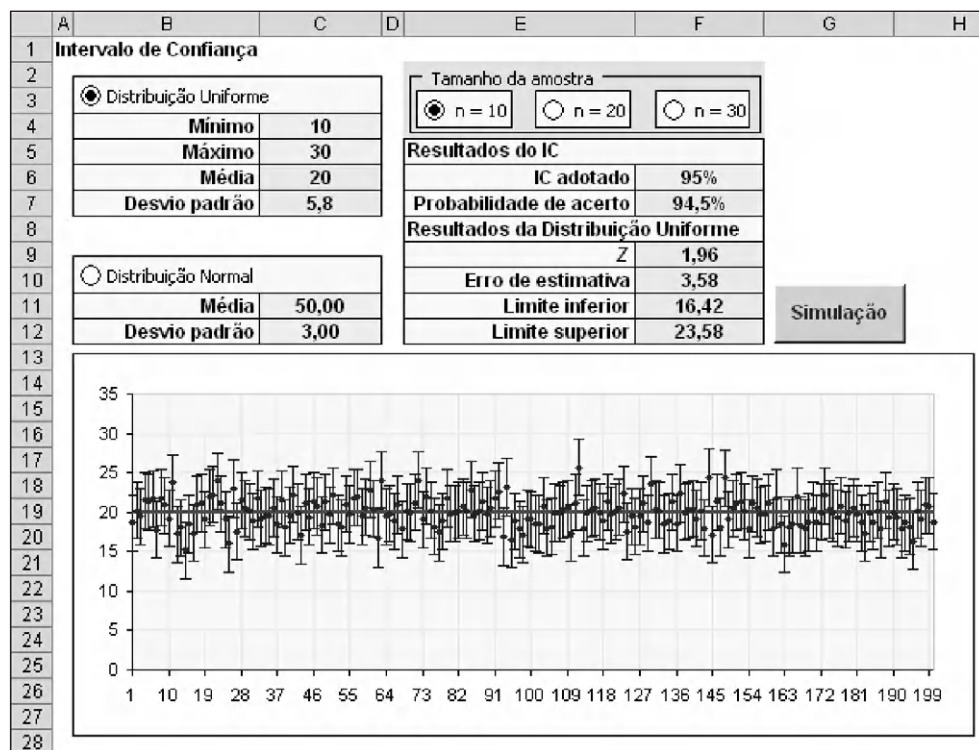
Para qualquer valor de média da amostra, o erro de estimativa é o mesmo e igual a 2,056. Analisemos os intervalos das três médias.

- Se a média da amostra for igual a 50, então a média da população estará no intervalo $\mu = 50 \pm 2,056$, ou entre os limites 47,944 e 52,056. Neste exemplo, afirmar que a média da população está contida no intervalo 47,944 e 52,056 é *verdadeiro*, pois a média da população é 50.
- Se a média da amostra for igual a 49, a média da população estará no intervalo $\mu = 49 \pm 2,056$, ou entre os limites 46,944 e 51,056. Afirar que a média da população está contida no intervalo 46,944 e 51,056 também é *verdadeiro*.
- Se a média da amostra for igual a 53, a média da população estará no intervalo $\mu = 53 \pm 2,056$, ou entre os limites 50,944 e 55,056. Neste exemplo, afirmar que a média da população está contida no intervalo 50,944 e 55,056 é *falso*.

Simulador intervalo de estimação

A estimativa da média da população é um processo aleatório, com os valores *verdadeiro* e *falso* associados a uma distribuição de frequências do verdadeiro valor, incluído o conceito de longo prazo. A planilha **Intervalo de Confiança**, incluída na pasta **Capítulo 11**, ajudará a compreender o que foi exposto. O *modelo* está preparado para extrair da distribuição uniforme, ou da distribuição normal, 200 amostras aleatórias de três tamanhos diferentes $n=10, 20$ e 30 , que podem ser selecionadas na caixa de opções.

- A Figura 11.3 mostra o *modelo* com os registros parciais e o gráfico dos intervalos das estimativas de 200 amostras aleatórias de tamanho dez, extraídas de uma população com distribuição contínua e uniforme entre os limites 10 e 30, considerando o intervalo de confiança de 95%, valor registrado na célula F6. A célula F7 registra a probabilidade de acerto dos 200 intervalos, nesse caso, 94,5%, resultado obtido da contagem dos intervalos que contêm o valor da média da população conhecida e registrada na célula C6.
- A Figura 11.4 mostra o *modelo* com os registros parciais e o gráfico dos intervalos das estimativas de 200 amostras aleatórias de tamanho 30, extraídas de uma população com distribuição normal $N(50, 3)$, considerando o intervalo de confiança de 90%, valor registrado na célula F6. A célula F7 registra a probabilidade de acerto dos 200 intervalos, nesse caso 90,5%.


FIGURA 11.3

Simulação de 200 amostras com a distribuição uniforme.

Toda vez que for selecionado um outro tamanho de amostra, o modelo será recalculado. Clicando no botão **Simulação**, o modelo gerará um novo grupo de 200 amostras de tamanho e da distribuição selecionada. Observe que:

- Todos os intervalos de variação da média são iguais, duas vezes o erro de estimativa, ou margem de erro. No entanto, os limites da estimativa da média da população são aleatórios.
- O aumento do tamanho da amostra diminui a diferença entre o IC estabelecido na célula F6 e a probabilidade de intervalos que contêm a média da população registrada na célula F7.
- Ao mesmo tempo, a simulação de 200 amostragens é uma quantidade pequena para imaginar que, pela lei dos grandes números, ao aumentar o número de experimentos, essa diferença será próxima do seu valor teórico.
- Analisando o gráfico, é possível contar os intervalos que não contêm a média da amostra que, nesse caso, é conhecida. Uma melhor visualização pode ser obtida aumentando o comprimento do gráfico. Para isso, primeiro a planilha deverá ser desprotegida, desta forma: no menu **Ferramentas**, selecione **Proteger** e depois **Desproteger planilha**.

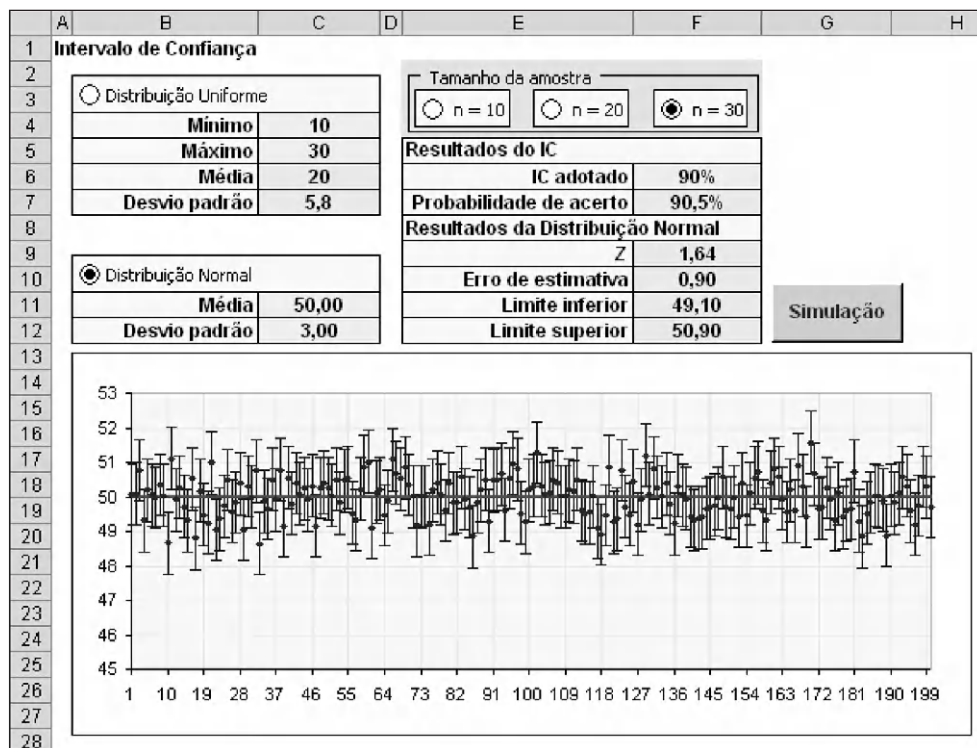


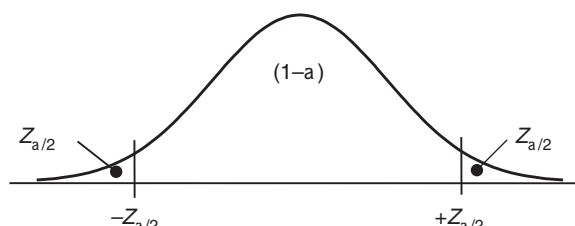
FIGURA 11.4 Simulação de 200 amostras com a distribuição normal.

Incluindo o erro tolerado

O intervalo de confiança determina a probabilidade de acerto da estimativa, por exemplo, se $IC=90\%$, a probabilidade de acerto será 90% e, conseqüentemente, a probabilidade de erro α será 10% . Dessa maneira, o erro α no processo de estimativa define o intervalo de confiança $IC=(1-\alpha)$, medindo ambos com valores unitários.

Distribuindo o erro α nas duas caudas da distribuição normal, o erro em cada cauda será $\alpha/2$. O erro α é denominado, também, *erro tolerado* ou *nível de confiança*, mudando a forma de construir o intervalo de confiança da média. A Figura 11.5 mostra o erro tolerado α nas duas caudas e o desvio padrão normalizado identificado como $Z_{\alpha/2}$.

FIGURA 11.5 Erro tolerado α .



A estimativa da média com intervalo de confiança $(1-\alpha)\times 100$, sendo conhecido o desvio padrão da população, é registrada na fórmula seguinte:

$$\mu = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

A relação entre o erro tolerado α , o intervalo de confiança $(1-\alpha)$ e o desvio padrão normalizado $Z_{\alpha/2}$ é apresentada na tabela da Figura 11.6, cujos resultados mostram que, quanto maior for o erro tolerado α , menor será o desvio padrão normalizado $Z_{\alpha/2}$ e, conseqüentemente, menor será o erro de estimativa.

α	$\alpha/2$	$(1-\alpha) \times 100$	$Z_{\alpha/2}$
0,20	0,100	80%	$\pm 1,282$
0,15	0,075	85%	$\pm 1,444$
0,10	0,050	90%	$\pm 1,645$
0,05	0,025	95%	$\pm 1,960$
0,01	0,005	99%	$\pm 2,576$

FIGURA 11.6 Tabela de comparação.

EXEMPLO 11.4

De uma população com desvio padrão de 2,50, foi retirada uma amostra aleatória de tamanho $n=36$, cuja média é 10,60. Estime a média da população com intervalo de confiança de 95%.

Solução. Como o intervalo de confiança é 95%, o desvio padrão normalizado dos dois limites é $Z=\pm 1,96$. A estimativa da média da população é $10,6 \pm 0,8167$, resultado obtido de:

$$\mu = 10,6 \pm 1,96 \frac{2,5}{\sqrt{36}} = 10,6 \pm 1,96 \times 0,4167$$

$$\mu = 10,6 \pm 0,8167$$

A margem de erro, ou erro de estimativa, também pode ser obtida registrando a fórmula =INT.CONFIANÇA(0,05;2,5;36) em uma célula da planilha Excel, que retorna o resultado 0,8167.

- **INT.CONFIANÇA(alfa; desv_padrão; tamanho)**

A função estatística INT.CONFIANÇA retorna o erro de estimativa (ou margem de erro) utilizando a distribuição Z e os argumentos *alfa* correspondente ao erro tolerado, o argumento *desv_padrão* e o *tamanho* da amostra.⁵

O mesmo resultado pode ser obtido registrando numa célula do Excel a fórmula =INV.NORMP(0,975)*2,50/RAIZ(36). Essa fórmula utiliza o valor de Z do limite superior, e a fórmula =INV.NORMP(0,025)*2,50/RAIZ(36) utiliza o limite inferior, retornando o mesmo resultado anterior, porém com sinal negativo.

A planilha **Modelo Estimativa Média com Z**, incluída na pasta **Capítulo 11**, realiza o cálculo da média da população, como mostra a figura seguinte.

- No quadro com o nome **Dados** são informados todos os dados necessários para realizar a estimativa da média da população, intervalo B3:C7. A célula C7 não aceita valores do tamanho de amostra menores do que 31.
- No quadro com o nome **Resultados**, o modelo registra:
 - O erro de estimativa na célula F4.
 - Os limites do intervalo da média da população nas células F5 e F6.
 - A estimativa da média é apresentada de forma numérica na célula F7.
 - No intervalo E8:F9, a estimativa da média é apresentada por extenso.

⁵ A utilização da distribuição Z é uma aproximação da utilização da distribuição t para amostras de tamanho $n>30$, como será apresentado mais adiante neste capítulo.

	A	B	C	D	E	F
1	Estimativa da Média da população - Distribuição Z					
2						
3		Dados			Resultados	
4		Alpha	0,05		Erro de estimativa	0,82
5		Média da Amostra	10,60		Limite Inferior	9,78
6		Desvio Padrão	2,50		Limite Superior	11,42
7		Tamanho amostra	36		Média da População	10,6 ± 0,82
8					Maior ou igual a 9,78 e menor ou igual a	
9					11,42 com intervalo de confiança de 95%	
10						

Desvio padrão da população desconhecido

Embora nos exemplos anteriores o desvio padrão da população tenha sido considerado conhecido, na maioria dos casos, esse desvio padrão é desconhecido. Como na amostra aleatória extraída da população, pode-se calcular sua média e seu desvio padrão, dessa amostra S_x é razoável adotar o desvio padrão da amostra como a melhor estimativa disponível do desvio padrão da população. Para amostras de tamanho suficientemente grande, em geral $n > 30$, o erro padrão é medido com a expressão:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

Entretanto, como a variabilidade das amostras gera, também, variabilidade no valor do desvio padrão amostral, como garantir que a estimativa do desvio padrão da população atende ao conceito de intervalo de confiança? A teoria e as simulações realizadas confirmam essa estimativa, considerando amostras de tamanho suficientemente grande, em geral, maiores do que 30, independente da forma de distribuição da amostra. Portanto, a estimativa da média da população será obtida com a expressão:

$$\mu = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Tenha em mente que:

- O tamanho da amostra é suficientemente grande, em geral $n > 30$.
- Se o desvio padrão da população for conhecido, o erro de estimativa é constante para qualquer amostra. Contudo, quando o desvio padrão da população não for conhecido, o intervalo da estimativa não será constante, podendo variar de amostra para amostra.

EXEMPLO 11.5

Com a intenção de estimar o número de ações negociadas por dia na bolsa de valores, foi realizada uma amostragem aleatória de 81 dias, cuja média é 12,8 milhões de ações por dia, e seu desvio padrão é 2,7 milhões de ações por dia. Estime a média da população com intervalo de confiança de 95%.

Solução. Aplicando a expressão da média obtemos:

$$\mu = \bar{X} \pm Z_{0,05/2} \times \frac{S_x}{\sqrt{n}} = 12,8 \pm 1,96 \times \frac{2,7}{\sqrt{81}}$$

$$\mu = 12,8 \pm 0,59$$

A média do número de ações negociadas por dia na bolsa de valores é um valor entre 12,21 e 13,39, com intervalo de confiança de 95%.

Escolha do tamanho da amostra

Vimos que o erro de estimativa (ou margem de erro) e o intervalo de confiança são funções do tamanho da amostra. No Exemplo 10.1, foi definido o intervalo de confiança e o tamanho da amostra, ficando determinado o erro de estimativa. Se o intervalo de confiança for definido, quanto maior for o tamanho da amostra, menor será a margem de erro. Essa relação está definida pela segunda parcela da expressão da fórmula da média da população, o erro de estimativa, que a seguir repetimos $e = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$, na qual

foi incluída a letra e para representar o erro de estimativa e completar a fórmula. Em alguns casos, interessa realizar estimativas com um erro aceitável ou erro de estimativa definido.

Definida a precisão da estimativa, as únicas variáveis livres possíveis de escolher são o tamanho da amostra e o intervalo de confiança (ou erro tolerado α), pois o desvio padrão será obtido da própria amostra, ou da população. Conhecidos o desvio padrão da população e o intervalo de confiança, para um erro de estimativa definido e , o tamanho da amostra é determinado com a expressão:

$$n = \left(\frac{Z_{\alpha/2} \times \sigma}{e} \right)^2$$

Se o desvio padrão da população não for conhecido, deverá ser utilizado o desvio padrão da amostra S_x com a mesma expressão. Todavia, como o tamanho da amostra n pode ser determinado, se o desvio padrão da amostra é desconhecido? Um caminho é determinar o desvio padrão de uma amostra piloto, a mais representativa possível.

EXEMPLO 11.6

De uma amostra de tamanho $n=30$, com média 10,50 e desvio padrão 2,45, foi estimada a média da população $\mu=10,50 \pm 0,89$, com intervalo de confiança 95%. Querendo apresentar a estimativa da média da população com um erro de estimativa $e=0,50$, qual deve ser o tamanho da amostra?

Solução. O tamanho da amostra deverá ser $n=93$, resultado obtido de:

$$n = \left(\frac{1,96 \times 2,45}{0,50} \right)^2 = 92,24 \approx 93$$

Esse resultado também pode ser obtido com o modelo da planilha **Modelo Tamanho da Amostra**, incluído na pasta **Capítulo 11**, como mostra a figura a seguir.

	A	B	C	D	E	F
1	Determinação do tamanho da amostra					
2						
3		<input type="radio"/> Alfa	<input checked="" type="radio"/> IC		Resultados com IC = 95%	
4		Intervalo de confiança	95%		Z	1,96
5		Desvio Padrão	2,45		Tamanho da amostra	93
6		Erro de estimativa	0,50			
7						

Nesse modelo, é possível obter resultados utilizando o intervalo de confiança IC ou o nível de confiança $Alfa$, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E3:F3 sempre que o IC for menor do que 80% e $Alfa$, maior do que 20%. O resultado do tamanho da amostra, célula F5, foi arredondado para a unidade seguinte.

O cálculo do tamanho da amostra para populações finitas pode ser encontrado no Apêndice 2 deste capítulo.

Estimativa da média com a distribuição t

Quando o desvio padrão da população não for conhecido, a estimativa da média da população deverá ser realizada com a *distribuição t* , acompanhando o procedimento apresentado com a distribuição Z .⁶ Em alguns casos, não é possível retirar amostras grandes, pois os dados disponíveis são poucos, o custo unitário da amostragem é alto, o tempo disponível não é suficiente etc. Como a forma da distribuição das médias de amostras pequenas dependerá da forma da distribuição da população, o desvio padrão da amostra não será uma boa estimativa do desvio padrão da população. Portanto, para realizar a estimativa da média da população com amostras pequenas, a distribuição da população deverá ser normal.⁷ Nessas condições, a estimativa da média da população será realizada com a distribuição t , conhecida como *distribuição de Student*.

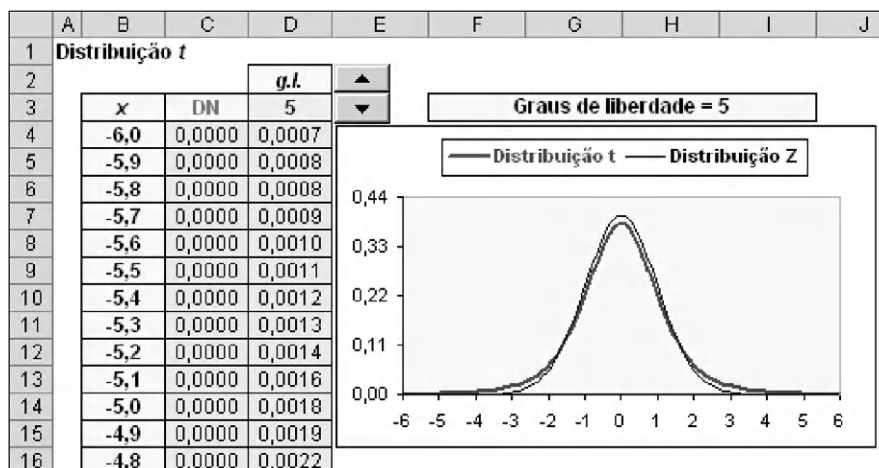
Características da Distribuição t

- Distribuição contínua e simétrica com média igual a zero.
- Há uma família de distribuições t , em função do *grau de liberdade* $gl=(n-1)$.
- É mais aberta e as caudas são um pouco mais altas do que as da distribuição Z . Para amostras com $gl>30$, a distribuição t se aproxima da distribuição Z .

Na planilha **Modelo Distribuição t** , incluída na pasta **Capítulo 11**, foi construída a função densidade da distribuição t que adota forma diferente em função do parâmetro grau de liberdade gl .⁸ Clicando no controle giratório, o número de graus de liberdade varia de um até 60 e, ao mesmo tempo, a forma da distribuição t muda. Para fins comparativos, na mesma planilha foi construída a função densidade da distribuição normal padronizada. Para graus de liberdade maiores do que 30, praticamente, a curva da distribuição t coincide com a curva da distribuição normal $N(0, 1)$. Entretanto, para graus de liberdade menores do que 30, a curva da distribuição t diminui sua altura e aumenta sua largura, mostrando que as caudas da distribuição t são mais altas do que as da distribuição normal padronizada, como mostra a Figura 11.7 para $gl=5$, que corresponde a um tamanho de amostra igual a seis.

FIGURA 11.7

Distribuição t , comparação com a distribuição normal $N(0, 1)$.



⁶ Com a distribuição Z , obtém-se um resultado aproximado para amostra com $n>30$.

⁷ É recomendado verificar a forma da distribuição para confirmar a premissa de normalidade da amostra; por exemplo, construindo seu histograma. Se a inclinação da distribuição da população não for acentuada e o tamanho da amostra não for pequeno, poderá ser utilizada a distribuição t com $(n-1)$ graus de liberdade e desvio da população desconhecido.

⁸ O gl pode ser assim entendido. No cálculo da variância da amostra, é utilizada a média da amostra. Determinada a média da amostra, apenas um dos n valores da amostra ficará determinado pela média da amostra, sendo que os restantes $(n-1)$ valores poderão variar.

Para estimar a média da população considerando as duas caudas da distribuição t e $(n-1)$ graus de liberdade, aplica-se a expressão:⁹

$$\mu = \bar{X} \pm t_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

A tabela da Figura 11.8 compara os valores críticos das distribuições Z e t para $n=31$ e diversas probabilidades de erro α para as duas caudas da distribuição, ou $\alpha/2$, para a cauda superior, e correspondente a $\alpha/2$, para a cauda inferior da distribuição t .

α	$1-\alpha$	$Z_{\alpha/2}$	$t_{\alpha/2}$
0,20	0,80	1,281	1,310
0,15	0,90	1,645	1,697
0,10	0,95	1,960	2,042
0,01	0,99	2,576	2,750

FIGURA 11.8 Valores críticos de Z e t .

No capítulo *Tabelas* no final do livro, você encontrará a tabela dos valores críticos t_c da distribuição t para diferentes intervalos de confiança e graus de liberdade. Analisando os valores críticos da tabela da distribuição t , observa-se que, ao estabelecer amostras de tamanho $n=30$ como limite entre amostras pequenas e grandes, os valores críticos das distribuições t e Z são bem próximos. Para valores maiores do que 30, as diferenças diminuem, porém com menor velocidade, de forma que a escolha do limite $n=30$ é uma escolha de bom senso, não introduzindo erros consideráveis e explicando a simplificação de usar Z para amostras maiores.

EXEMPLO 11.7

A amostra {7, 4, 2, 5, 7} foi retirada de uma população com distribuição normal. Estime a média da população, considerando o intervalo de confiança 90%.

Solução. O primeiro passo é calcular a média e o desvio padrão da amostra, respectivamente, 5 e 2,12. A estimativa da média da população utilizando a distribuição t é realizada com a fórmula:

$$\mu = \bar{X} \pm t_{\alpha/2} \times \frac{S_X}{\sqrt{n}}$$

O t crítico para duas caudas, considerando o intervalo de confiança 90% e $gl=5-1=4$, é igual a $t=2,132$, valor obtido da tabela da distribuição t , como mostra a figura a seguir.

Duas caudas	Intervalo de confiança		
	80%	90%	95%
$g.l. = 1$	3,078	6,314	12,706
2	1,886	2,920	4,303
3	1,638	2,353	3,182
4	1,533	2,132	2,776
5	1,476	2,015	2,571
6	1,440	1,943	2,447

⁹ Se o tamanho n da amostra for maior do que 5% do tamanho N da população, deverá ser aplicado o fator de correção finita.

Considerando o desvio padrão da amostra como o melhor estimador do desvio padrão da população, a estimativa da média da população é $5 \pm 2,02$, com intervalo de confiança 90%, resultado obtido de:

$$\mu = 5 \pm 2,132 \times \frac{2,12}{\sqrt{5}}$$

$$\mu = 5 \pm 2,02$$

O valor do *t* crítico para as duas caudas, considerando o intervalo de confiança 90% e *gl*=4, pode ser obtido utilizando a função INVT do Excel, registrando a fórmula =INVT(0,1;4), que retorna o resultado 2,131846, arredondando para 2,132.

• **INVT(probabilidade; graus_liberdade)**

A função estatística INVT retorna o *t* crítico da distribuição *t* referente aos argumentos *probabilidade* e *graus_liberdade*, considerando que a probabilidade se refere às duas caudas da distribuição. Mais informações sobre esta função e a função DISTT são apresentadas no Apêndice 1 deste capítulo.

EXEMPLO 11.8

Uma amostra das horas trabalhadas semanalmente por doze executivos escolhidos aleatoriamente dentre as 500 maiores empresas está registrada na tabela seguinte.

60	66	64	62	58	62	62	60	62	60	64	66
----	----	----	----	----	----	----	----	----	----	----	----

Estime a média da população, considerando o intervalo de confiança 95% e a distribuição de frequências das horas trabalhadas semanalmente aproximadamente normal.

Solução. A média e o desvio padrão da amostra são, respectivamente, 62,17 horas e 2,48 horas, resultados obtidos com as fórmulas:

$$\begin{aligned} &= \text{MÉDIA}(\{60;66;64;62;58;62;62;60;62;60;64;66\}) \\ &= \text{DESVPAD}(\{60;66;64;62;58;62;62;60;62;60;64;66\}) \end{aligned}$$

Como a amostra é pequena, com distribuição aproximadamente normal, para o intervalo de confiança de 95%, ou nível de confiança de 5%, nas duas caudas e *gl*=11 graus de liberdade, o *t* crítico é igual a 2,201, resultado obtido com a fórmula =INVT(0,05;11). A média da população é $62,17 \pm 1,576$, com intervalo de confiança 95%, resultado obtido de:

$$\mu = 62,17 \pm 2,201 \times \frac{2,48}{\sqrt{12}}$$

$$\mu = 62,17 \pm 1,576$$

A planilha **Modelo Estimativa Média com t**, incluída na pasta **Capítulo 11**, realiza o cálculo da média da população, como mostra a figura:

- No quadro **Dados**, são informados os dados necessários para realizar a estimativa da média da população utilizando a distribuição *t*, intervalo B3:C7.
- No quadro **Resultados**, o modelo registra os resultados:
 - O número de graus de liberdade *g.l.* na célula F4, o valor crítico de *t* na célula F5 e o erro de estimativa na célula F6.
 - Os limites do intervalo da média da população nas células F7 e F8.
 - A estimativa da média é apresentada de forma numérica na célula F9, e no intervalo E10:F11, por extenso.

	A	B	C	D	E	F
1	Estimativa da Média da população - Distribuição t					
2						
3		Dados			Resultados	
4		Alfa	5%		g.l.	11
5		Média da Amostra	62,17		t	2,20
6		Desvio Padrão	2,48		Erro de estimativa	1,58
7		Tamanho amostra	12		Limite Inferior	60,59
8					Limite Superior	63,75
9					Média da População	62,17 ± 1,58
10					Maior ou igual a 60,59 e menor ou igual a	
11					63,75 com intervalo de confiança de 95%	
12						

Modelo geral para estimativa da média

Na planilha **Modelo Geral Estimativa Média**, incluída na pasta **Capítulo 11**, foi construído o *modelo* para estimativa da média de uma população com a distribuição *t* e a distribuição *Z* como mostra a Figura 11.9, resolvendo o Exemplo 11.7. O modelo conta com três caixas de grupo com duas opções cada uma:

- Na caixa de grupo **Distribuição da população**, pode-se escolher **Normal** ou **Não é normal** clicando no botão de opção correspondente.
- Na caixa de grupo **Fator de correção finita**, pode-se escolher **Incluir** ou **Não incluir**, clicando no botão de opção correspondente. Se for selecionado **Incluir**, o modelo solicitará que na célula I4 seja informado o tamanho da população.
- Na caixa de grupo **Estimativa com a**, pode-se escolher **Dist. Normal** ou **Dist. T**, clicando no botão de opção correspondente.

Ainda, com o modelo, é possível obter resultados utilizando o intervalo de confiança *IC* ou o nível de confiança *Alfa*, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E9:F9, sempre que o *IC* for menor do que 80% e *Alfa*, maior do que 20%. O **Modelo Geral** fornece os resultados sempre que a distribuição da população for normal ou, caso não seja normal, para tamanho de amostra maior ou igual a 31. Se os dados não atenderem às premissas do modelo, os resultados relevantes não serão apresentados na planilha.

	A	B	C	D	E	F
1	Modelo Geral para estimativa da Média da população					
2						
3		Distribuição da população			Fator de correção finita	
4		<input checked="" type="radio"/> Normal	<input type="radio"/> Não é normal		<input type="radio"/> Incluir	<input checked="" type="radio"/> Não incluir
5		Estimativa com a				
6		<input type="radio"/> Dist. Normal	<input checked="" type="radio"/> Dist. t			
7						
8						
9		Dados	<input type="radio"/> Alfa	<input checked="" type="radio"/> IC	Resultados com IC = 90%	
10		Intervalo de confiança	90%		t	2,13
11		Média da Amostra	5,00		Erro de Estimativa	2,02
12		Desvio Padrão	2,12		Limite Inferior	2,98
13		Tamanho amostra	5		Limite Superior	7,02
14					Média da População	5 ± 2,02
15					Maior ou igual a 2,98 e menor ou igual a	
16					7,02 com intervalo de confiança de 90%	
17						

FIGURA 11.9 Modelo Geral resolvendo o Exemplo 11.7.

Como regra geral, para estimar a média da população com desvio padrão desconhecido, a distribuição da população deve ser normal, utilizando o *modelo* da Figura 11.9, bem como a tabela *t* ou qualquer outro procedimento. Entretanto, se a inclinação da distribuição da população for pequena, para amostras de tamanho suficientemente grande, $n > 30$, a distribuição *t* poderá ser utilizada para estimar a média da população com desvio padrão desconhecido. Se a distribuição da população não for conhecida e o tamanho da amostra for pequeno, será necessário verificar a premissa de normalidade da população avaliando, por exemplo, a forma da distribuição de frequências ou outro método, assunto não tratado neste livro.

Problemas

Problema 1

Sejam $\bar{X}=10$, $\sigma=4,50$ e $n=16$. Estime a média da população com 95% de intervalo de confiança, considerando que a população tem distribuição normal.

R: $\mu=10 \pm 2,20$

Problema 2

Repita o Problema 1 com intervalo de confiança de:

a. 90%

b. 99%

R: a) $\mu=10 \pm 1,85$ b) $\mu=10 \pm 2,90$

Problema 3

Uma amostra aleatória de tamanho $n=36$ tem média 28,35 e desvio padrão 7,5. Estime a média da população com intervalo de confiança de:

a. 95%

b. 90%

R: a) $\mu=28,35 \pm 2,45$ b) $\mu=28,35 \pm 2,056$

Problema 4

A instituição financeira administra muitas carteiras de investimentos de pessoas físicas consideradas como *investidores médios*. Dessa população, foi retirada uma amostra aleatória de 30 carteiras cujos retornos em porcentagem estão registradas na tabela seguinte.

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

Calcule a estimativa pontual do retorno anual das carteiras, o desvio padrão da amostra e o erro padrão.

R: A média do retorno anual das carteiras no ano passado foi 10,47% e, portanto, a estimativa pontual da média do retorno anual das carteiras é igual 10,47%. O desvio padrão da amostra é 2,45% e o erro padrão 0,45%.

Problema 5

Continuando com o Problema 4. Explique o significado do erro padrão da estimativa da média.

R: O erro padrão $S_{\bar{X}} = S_X / \sqrt{n}$ é uma estimativa de $\sigma_{\bar{X}} = \sigma / \sqrt{n}$. O erro padrão indica que todas as possíveis médias das amostras de igual tamanho \bar{X} que podem ser retiradas da população se distribuem ao redor da média da população, com desvio padrão igual a 0,45%.

Problema 6

Continuando com o Problema 4. Considerando que a distribuição da população é normal, estime a média da população com intervalo de confiança de 95%.

R: Como a distribuição da população é normal e o tamanho da amostra suficientemente grande, podemos aplicar diretamente a distribuição Z. A estimativa da média dos retornos anuais das carteiras no ano passado é $10,47 \pm 0,88$, com valores entre 9,59% e 11,35%.

Problema 7

Continuando com o Problema 4. Calcule a probabilidade de a média dos retornos de uma amostra ser:

- Igual ou maior do que 11%.
- Igual ou maior do que 12%.

R: a) 11,80% b) 2. 0,03%

Problema 8

Continuando com o Problema 4. Calcule a probabilidade de a média dos retornos de uma amostra ser:

- Igual ou menor do que 10%.
- Igual ou menor do que 9%.

R: a) 14,67% b) 0,05%

Problema 9

De uma população com desvio padrão igual a 12, foi retirada uma amostra aleatória de tamanho $n=100$, com média igual a 81. Estime a média da população para os intervalos de confiança: 90%, 95% e 99%.

R: $81 \pm 1,97$ $81 \pm 2,35$ $81 \pm 3,09$

Problema 10

Uma amostra aleatória de 40 contas de pessoas físicas na filial de um banco apresentou saldo médio de \$1.400 com desvio padrão de \$300.

- Estime a média da população com intervalo de confiança de 95%.
- Repita a) com intervalo de confiança 99%.
- Qual a probabilidade de a média da população ser menor do que \$1.300?

R: a) $1.400 \pm 92,97$ b) $1.400 \pm 122,18$ c) 1,75%

Problema 11

Se o desvio padrão de uma amostra piloto de uma população for igual a 18, e a estimativa da média da população deve ser realizada com erro de estimativa igual a seis, qual deverá ser o tamanho da amostra considerando o intervalo de confiança 95%?

R: $n \cong 35$

Problema 12

O desvio padrão de uma população é 12. Qual o tamanho da amostra se o erro de estimativa deve ser igual a dois e o intervalo de confiança 90%?

R: $n \cong 97$

Problema 13

Continuando com o Problema 6. Qual o tamanho da amostra para um erro de estimativa igual a 0,5%?

R: $n \cong 93$

Problema 14

O desvio padrão da duração de uma reunião de um *consultor* é uma hora. Determine o tamanho da amostra para estimar o tempo médio das reuniões do *consultor* com erro de estimativa de 0,30 hora, considerando o intervalo de confiança de 95%.

R: $n \cong 43$. Pelo tamanho da amostra, não é necessário que a população tenha distribuição normal

Problema 15

Repita o Problema 6 aplicando a distribuição t .

R: $10,47 \pm 0,92$, um pouco maior do que a estimativa com a distribuição Z .

Problema 16

A amostra $\{1, 3, 5, 7, 19\}$ foi retirada de uma população com distribuição normal. Estime a média da população considerando o intervalo de confiança 95%.

R: $\mu = 7 \pm 7,22$

Problema 17

Por erro no registro dos dados do Problema 16, o último dado é 9 em vez de 10. Estime a média da população considerando o intervalo de confiança 95% e analise o efeito de um dado suspeito sobre a estimativa da média da população.

R: $\mu = 5 \pm 2,77$

Problema 18

Um mês antes de terminar o ano, foram consultados 25 profissionais de mercado sobre a estimativa da taxa real de juros para o próximo ano. As 25 respostas formaram uma variável aleatória com média 14% e desvio padrão 6,5% ao ano.

a. Calcule os valores máximo e mínimo da taxa real estimada para o próximo ano, considerando o intervalo de confiança 95%.

b. Se a taxa real de juros for 15,5%, que podemos dizer desse resultado comparado com o valor estimado?

c. Se a taxa real de juros for 10%, que podemos dizer desse resultado comparado com o valor estimado?

R: a) Aplicando a distribuição t , a estimativa da média da população com 95% de intervalo de confiança é $\mu = 14 \pm 2,68$. b) A taxa real de juros do ano está incluída no intervalo da estimativa. c) A taxa real de juros do ano não está incluída no intervalo da estimativa. Tenha em mente que o intervalo da estimativa é garantido em 95% das vezes! Outro aspecto importante a ser considerado é que durante o ano devem ter acontecidos eventos que mudaram as características da população sobre a quais os profissionais basearam suas projeções.

Problema 19

De uma população com distribuição normal e desvio padrão igual a 5, foi retirada uma amostra aleatória de tamanho 80 e média 24. Estime a média da população com intervalo de confiança de 90%.

R: $\mu = 24 \pm 0,9195$

Problema 20

Repita o Problema 19, considerando os intervalos de confiança: 95% e 99%.

R: $\mu(95\%) = 24 \pm 1,0957$ $\mu(99\%) = 24 \pm 1,4399$

Problema 21

Refaça os Problemas 19 e 20, considerando o tamanho da amostra $n=36$.

R: $\mu(90\%)=24\pm 1,3707$ $\mu(95\%)=24\pm 1,6333$ $\mu(99\%)=24\pm 2,1465$

Problema 22

Repita o Problema 21, considerando o tamanho da amostra igual a 25.

R: $\mu(90\%)=24\pm 1,7109$ $\mu(95\%)=24\pm 2,0639$ $\mu(99\%)=24\pm 2,7970$

Problema 23

O interesse da revista de carros é estimar a média de consumo em *quilômetros por litro* de um novo modelo de carro da montadora líder do mercado de carros populares. Uma amostra aleatória de 16 carros do novo modelo de carro apresentou média de 14,8 e desvio padrão de 2 quilômetros por litro. Estime a média da população com intervalo de confiança de 95%, considerando que a população tem distribuição normal.

R: $\mu=14,8\pm 1,0657$ km/l.

Problema 24

Repita o Problema 23, considerando que o intervalo de confiança de 90%.

R: $\mu=14,8\pm 0,8765$ km/l.

Problema 25

O gerente do controle aéreo do aeroporto de São Paulo está interessado em conhecer o tempo de aterrissagem dos aviões modelo 737, medindo esse tempo entre o instante em que o piloto inicia a operação de descida e o instante em que o avião abandona a pista principal. Se uma amostra aleatória de 33 aviões tem média de 21 minutos e desvio padrão 4,5 minutos, quais as médias da população, considerando os intervalos de confiança 90% e 95%?

R: $\mu(90\%)=21\pm 1,29$ minutos. $\mu(95\%)=21\pm 1,54$ minutos.

Problema 26

Continuando com o Problema 25. Qual o tamanho da amostra para ter um *erro de estimativa* de 0,80 minuto, com intervalo de confiança de 95%.

R: $n=121,55\approx 122$.

Problema 27

Repita o Problema 26 para o intervalo de confiança de 90%.

R: $n=85,61\approx 86$.

Problema 28

Depois de entrevistar 16 gerentes juniores de uma grande empresa com centenas de profissionais nesse cargo, o analista de salários de uma empresa de recrutamento obteve a média de salários anuais igual a \$33.500, com desvio padrão de \$8.150. Considerando que os salários têm distribuição normal, estime a média dos salários anuais dos gerentes juniores dessa empresa, com intervalo de confiança de 95%.

R: $\mu=\$33.500\pm 4.342,83$

Problema 29

Repita o Problema 28, considerando o intervalo de confiança de 90%.

R: $\mu=\$33.500\pm 3.571,84$

Problema 30

Analise os resultados diferentes das duas pesquisas sobre o projeto de lei para fechamento de bares e restaurantes depois da uma da manhã.

- *Maioria defende fechamento de bares à 1h em SP.*¹⁰ “A maioria absoluta dos paulistanos é favorável ao projeto que proíbe o funcionamento na cidade de bares e restaurantes que não tenham isolamento acústico, seguranças e estacionamento. Foi o que disseram 67% dos 630 entrevistados pelo Datafolha na última quinta feira na cidade de São Paulo. A margem de erro é de quatro pontos percentuais, para mais ou para menos. Uma das principais explicações para um percentual tão grande de apoio à medida é que 80% da população da cidade não costuma frequentar bares e restaurantes após a 1h e, portanto, não se sente prejudicada pela restrição ao funcionamento das casas”.
- *56% reprovam lei que fecha bares à 1 hora.*¹¹ “A maioria dos moradores de São Paulo é contra o fechamento dos bares à 1 hora. Pesquisa InformEstado feita na capital mostra que 56% da população não concorda com o projeto de lei aprovado na semana passada pela Câmara Municipal. A resistência à medida vem, principalmente, dos jovens de 18 a 29 anos (67,5% contra), dos que têm instrução superior (62,8%), dos homens (59,1%) e dos que ganham mais de vinte salários mínimos (63,4%). A lei encontra apoio de 43% dos entrevistados. Ele é maior entre as mulheres (45,4% a favor do fechamento), os que têm mais de 50 anos (55,1%), o primeiro grau incompleto (67,1%) e entre os que recebem até cinco salários mínimos (53,5%) ... O InformEstado entrevistou 622 pessoas. ... A margem de erro é de quatro pontos percentuais”.

¹⁰ Artigo de José Roberto de Toledo publicado no jornal *Folha de São Paulo* em 27/06/99.

¹¹ Artigo de Marcelo Godoy publicado no jornal *O Estado de São Paulo* em 27/06/99.

Apêndice 1

Funções estatísticas do Excel

O Excel dispõe das funções estatísticas DISTT e INVT para a distribuição t cujas sintaxes são:

DISTT(t ; graus_liberdade ; caudas)

A função estatística DISTT¹² retorna a probabilidade de o valor positivo t ser excedido, considerando os argumentos graus_liberdade e o número de caudas da distribuição t . Se o argumento caudas for igual a 1, a função DISTT retornará a probabilidade correspondente a uma cauda da distribuição, e se for igual a 2, retornará a probabilidade correspondente às duas caudas da distribuição. Por exemplo, sejam $t=2,042$ e $gl=30$:

- Para caudas igual a 2, a fórmula =DISTT(2,042;30;2) retornará o valor 0,05, que é a soma de probabilidades das duas caudas $\alpha/2 + \alpha/2 = 0,05$. Essa probabilidade é a área sob a curva entre cada t crítico definido e seu correspondente limite extremo da distribuição, como se pode ver na primeira distribuição da esquerda da Figura 11.10.
- Para caudas igual a 1, a fórmula =DISTT(2,042;30;1) retornará o valor 0,025, que é a probabilidade α de uma das caudas. Essa probabilidade é a área sob a curva entre o t crítico definido e seu correspondente limite extremo da distribuição, como se pode ver na distribuição da direita da Figura 11.10, escolhendo a cauda superior.

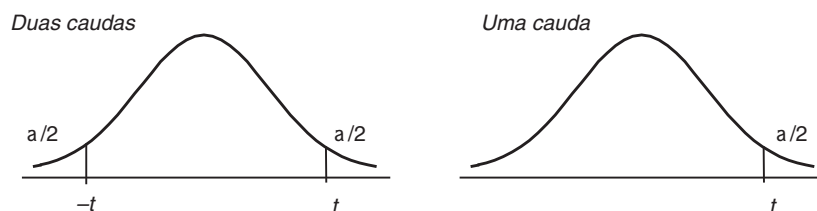


FIGURA 11.10 Valores críticos da distribuição t , conhecido α .

INVT(probabilidade ; graus_liberdade)

A função estatística INVT¹³ retorna o t crítico da distribuição t referente aos argumentos probabilidade e graus_liberdade , considerando que a probabilidade se refere às duas caudas da distribuição. O retorno da função INVT do Excel é o resultado de um procedimento iterativo até alcançar um erro de $\pm 3 \times 10^{-7}$; se em 100 iterações não for possível obter o resultado, a função INVT apresenta #N/A. Por exemplo, no caso de duas caudas, se $\text{probabilidade}=0,05$ e $gl=30$, a função estatística INVT(0,05;30) retornará o valor 2,042. Esse exemplo mostra que a função INVT é a função inversa da DISTT quando o argumento caudas é igual a 2.

No caso de realizar cálculos com a função INVT em uma cauda da distribuição, o valor do argumento probabilidade deverá ser informado como o dobro do valor do problema, pois o procedimento de cálculo da função INVT divide a probabilidade informada por dois.

¹² Em inglês, a função DISTT é TDIST.

¹³ Em inglês, a função INVT é TINV.

Na planilha **Funções DISTT e INVT**, incluída na pasta **Capítulo 11**, foram construídos dois modelos para a utilização e a compreensão das duas funções apresentadas, como mostra a Figura 11.11.

- O primeiro modelo calcula a probabilidade, considerando a escolha realizada na caixa de combinação, *Duas caudas* ou *Uma cauda*, argumentos previstos na própria função DISTT.
- O segundo *modelo* calcula o *t crítico* considerando as duas caudas da distribuição, tendo presente que no caso de realizar cálculos com a função INVT em uma cauda da distribuição, o valor do argumento *probabilidade* deverá ser informado como o dobro do valor do problema, pois o procedimento de cálculo da função INVT divide a probabilidade informada por dois.

FIGURA 11.11 Resultados com as funções DDIST e INVT.

	A	B	C	D	E	F
1		Função DISTT			Função INVT	
2						
3		t	2,042		P(t>)	0,050
4		n	30		n	30
5			Duas caudas ▼		g.l.	29
6		g.l.	29		t	2,042
7		P(t>2,042)	0,050			
8						

Apêndice 2

População finita

No Capítulo 10, foi visto que se em uma população finita for realizada uma amostragem, sem reposição de tamanho, maior do que 5% do tamanho da população, no cálculo do erro padrão deverá ser incluído o fator de correção finita $\sqrt{\frac{N-n}{N-1}}$. Incluindo o fator de correção finita na fórmula do tamanho da amostra, temos a seguinte expressão:

$$e = Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n_f}} \sqrt{\frac{N-n_f}{N-1}}$$

Como essa última expressão se refere ao tamanho da amostra de uma amostragem sem reposição, de tamanho maior do que 5% do tamanho da população incluindo o fator de correção finita, o tamanho da amostra será identificado com n_f . A expressão seguinte mostra uma simplificação de símbolos tendo presente que $n = \left(\frac{Z_{\alpha/2} \times \sigma}{e} \right)^2$:

$$n_f = \left(Z_{\alpha/2} \times \frac{\sigma}{e} \right) \times \frac{N-n_f}{N-1}$$

Realizando as passagens necessárias para obter o tamanho da amostra n_f :

$$n_f = \frac{nN}{N - 1 + n}$$

EXEMPLO 11.9

A média e o desvio padrão de uma amostra de tamanho $n=42$ são, respectivamente, 45 e 12,3. Qual deve ser o tamanho da amostra para ter um erro de estimativa de 2,7 se o tamanho da população é 1.000 e o intervalo de confiança 95%?

Solução. O tamanho da amostra considerando que a população não é finita é $n=79,725$, resultado obtido com a fórmula:

$$n = \left(\frac{1,96 \times 12,3}{2,7} \right)^2 = 79,725$$

O tamanho da amostra considerando a população finita $N=1.000$ é $n_f=74$, resultado obtido com a fórmula:

$$n_f = \frac{79,725 \times 1.000}{1.000 - 1 + 79,725} = 73,91$$

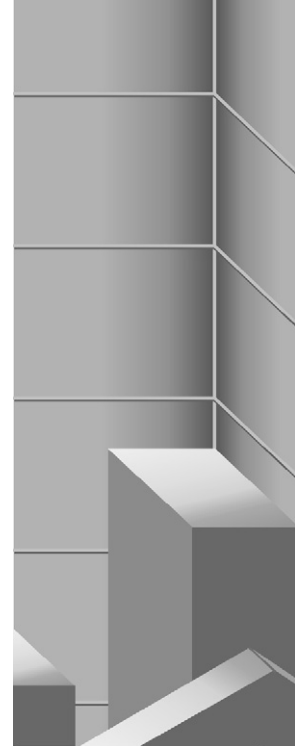
Esse resultado também pode ser obtido com o modelo da planilha **Modelo Tamanho da Amostra 2**, incluído na pasta **Capítulo 11**, como mostra a figura seguinte. Nesse modelo, é possível obter resultados utilizando o intervalo de confiança *IC* ou o nível de confiança *Alfa*, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E3:F3 sempre que o *IC* for menor do que 80% e *Alfa* maior do que 20%. O resultado do tamanho da amostra, célula F5, foi arredondado para a unidade seguinte.

	A	B	C	D	E	F
1	Determinação do tamanho da amostra com população finita					
2						
3		Dados	<input type="radio"/> Alfa	<input checked="" type="radio"/> IC	Resultados com IC = 95%	
4		Intervalo de confiança		95%	Z	1,96
5		Desvio Padrão		12,30	Tamanho da amostra	74
6		Erro de estimativa		2,70		
7		Tamanho população		1.000		
8						

Se o tamanho da população for muito grande comparado com o da amostra, o resultado desse modelo será o mesmo que o do **Modelo Tamanho da Amostra**.

Capítulo 12

TESTE DE HIPÓTESES



No Capítulo 11, foi mostrado como *estimar* a média de uma população a partir de uma única amostra aleatória retirada dessa população. Neste capítulo, será estudado outro tipo de inferência estatística, o *teste de hipóteses*. A estimativa da média de uma população é realizada porque sua média não é conhecida. Entretanto, o teste de hipóteses é realizado para verificar se a média afirmada deve ou não ser aceita, pois a média da população é conhecida. Nesses dois tipos de inferência, são utilizadas amostras para estimar ou confirmar o parâmetro da população.

EXEMPLO 12.1

O gerente financeiro da empresa administradora de cartões de crédito definiu a renda média mensal dos associados de \$2.500 para ser utilizada como premissa durante a preparação do orçamento do próximo ano. Durante a primeira reunião do orçamento anual, o gerente de marketing contestou o valor da renda média mensal adotado, *afirmando* que a atual renda média mensal dos associados é maior do que \$2.500. O gerente-geral solicitou que seja verificado o valor adotado de \$2.500, pois a maior parte do lucro da empresa administradora de cartões de crédito depende da renda dos associados.

A afirmação do gerente de marketing precisa ser provada para tornar-se *significativa*. Uma forma de obter provas ou evidências seria aplicar um questionário de pesquisa em toda a população de associados da empresa, procedimento que implicará um aumento significativo de despesas e demorará mais do que o tempo disponível para completar o orçamento anual. O procedimento escolhido foi aplicar o questionário de pesquisa em uma *amostra* aleatória representativa dessa população de associados da empresa administradora de cartões de crédito.

EXEMPLO 12.2

Na tentativa de verificar a afirmação do gerente de marketing, foi realizada uma pesquisa da renda mensal em uma amostra de 50 associados escolhidos aleatoriamente na população de associados. O resultado da pesquisa mostrou que a variável aleatória *renda mensal* tem média \$2.590 e desvio padrão \$285. Que conclusões podemos tirar desses dados da amostra?

Solução. Com os conhecimentos que temos até este momento, podemos dizer que o aumento da renda mensal de \$2.500 para \$2.590 pode ser proveniente:

- Da própria variabilidade das médias amostrais.
- De um aumento real dos salários dos associados.

Continuamos sem condições de definir se o aumento do valor da renda mensal medido pela pesquisa é realmente um aumento de renda dos associados ou é apenas um das muitas possíveis rendas mensais provenientes da variabilidade amostral.

Os resultados do Exemplo 12.2 mostram que ainda não temos condições de aceitar a afirmativa do gerente de marketing, apesar de a renda mensal \$2.590 da amostra ser maior do que a renda mensal incluída na preparação do orçamento anual. Neste momento, apenas temos evidências para concluir que:

- Se o valor \$2.590 for atribuído à própria variabilidade amostral, a renda mensal adotada \$2.500 deverá ser *aceita*.
- Se o valor \$2.590 for atribuído a um aumento da renda mensal dos associados, a renda mensal adotada \$2.500 deverá ser *rejeitada*.

Hipóteses

Ao adotar \$2.500 como renda média mensal dos associados, o gerente financeiro da administradora de cartões de crédito realizou uma *afirmação* sobre um parâmetro da população formada por todos os associados. Essa afirmação é denominada *hipótese* sobre o valor de um parâmetro de uma população. A hipótese não é necessariamente *verdadeira*, ela pode ser *correta* ou *errada*, sendo necessário obter uma amostra para ajudar a definir sua validade.

No teste de hipóteses são utilizadas duas hipóteses:

- A *hipótese nula* H_0 é a hipótese sobre a qual devem ser obtidas evidências para rejeitá-la.
- A *hipótese alternativa* H_1 é a hipótese sobre a qual devem ser obtidas evidências para aceitá-la.

A *hipótese nula* e a *hipótese alternativa* descrevem dois possíveis estados mutuamente excludentes, pois as duas hipóteses não podem ser aceitas ou rejeitadas ao mesmo tempo.

- A hipótese nula H_0 é o valor correntemente aceito até que se tenham evidências de que esse valor não é mais correto. A hipótese H_0 é uma afirmação ou ponto de partida do teste de hipóteses.
- A hipótese alternativa H_1 será somente aceita se surgirem evidências de que o valor da hipótese nula não é mais correto.

Aceitar a hipótese H_1 é uma posição mais forte do que aceitar a hipótese H_0 , pois é necessário obter evidências.

A convenção utilizada no teste de hipóteses é definir:

- A hipótese nula H_0 : $\mu = \mu_0$, sendo μ_0 o valor afirmado de um parâmetro da população. No Exemplo 12.1: $\mu_0 = \$2.500$.
- A hipótese alternativa H_1 : $\mu \neq \mu_0$ representa a conclusão do teste caso a hipótese nula seja rejeitada. No Exemplo 12.1: $\mu_0 \neq \$2.500$.

EXEMPLO 12.3

Análise os resultados da pesquisa o Exemplo 12.2 incluindo os conceitos da hipótese nula e da hipótese alternativa.

Solução. As hipóteses do Exemplo 12.4 são:

$$H_0: \mu = \$2.500$$

$$H_1: \mu \neq \$2.500$$

Aparentemente, a média amostral \bar{X} igual a \$2.590 não é muito diferente da média da população \$2.500. Como critério inicial, entendemos que se a probabilidade de ocorrência de \$2.590 for pequena, então a hipótese alternativa deverá ser rejeitada, mantendo-se a hipótese nula \$2.500.

A análise dos resultados do teste de hipóteses do Exemplo 12.3 foi realizada de forma intuitiva. Embora pareça correta, o resultado do Exemplo 12.3 é incompleto, pois não define o valor de uma probabilidade pequena. Deve-se dispor de um procedimento que mostre claramente quando aceitar H_0 e rejeitar H_1 , ou vice-versa. Em outras palavras, o objetivo é estabelecer um critério que permita distinguir entre diferenças casuais e diferenças reais; um critério que distinga se a diferença entre a afirmação sobre o valor do parâmetro e o valor medido pela amostra pode ou não ser atribuída à variação amostral.

Testes de hipóteses em uma cauda e nas duas caudas

Os testes de hipóteses podem ser aplicados em uma das duas caudas ou nas duas caudas da distribuição de frequências adotada.

- Um teste de hipótese em uma cauda da distribuição é um teste no qual a hipótese alternativa H_1 define a mudança em alguma direção da hipótese nula H_0 , incluindo na especificação um dos símbolos “ \leq ” ou “ \geq ”.
- Um teste de hipótese em duas caudas da distribuição é um teste no qual a hipótese alternativa H_1 define uma mudança da hipótese nula H_0 sem especificar nenhuma direção, incluindo na especificação o símbolo “ \neq ”.

Os testes de hipóteses deste capítulo serão aplicados nas duas caudas da distribuição, pois não é necessário realizar testes de hipóteses em uma cauda para poder afirmar que a média amostral é significativamente maior ou menor do que o valor utilizado como referência. Essa forma de proceder não limita a aplicação dos testes de hipóteses se as conclusões forem obtidas corretamente. Se o resultado da comparação da média amostral \bar{X} com o valor de referência μ_0 for *significativo*, pois há evidências de que H_0 seja falsa, então o resultado do teste de hipóteses nas duas caudas deverá ser utilizado como segue:

- Se $\bar{X} > \mu_0$, então a média amostral \bar{X} é significativamente maior do que o valor de referência μ_0 .
- Se $\bar{X} < \mu_0$, então a média amostral \bar{X} é significativamente menor do que o valor de referência μ_0 .

O teste de hipóteses sobre o parâmetro¹ média da população pode ser realizado pelos três procedimentos seguintes, que devem dar a mesma decisão:

- Aplicando o intervalo de confiança.
- Aplicando a estatística t ou Z .
- Aplicando o p -value.

¹ Devido à variabilidade dos valores das médias amostrais, as *estatísticas* tendem a se aproximar em vez de se igualarem ao valor do *parâmetro*.

Os dados necessários para a realização de um teste de hipóteses sobre a média de uma população são:

- A média μ_0 da população estabelecida na hipótese nula H_0 .
- O tamanho n e a média \bar{X} da amostra retirada da população.
- O desvio padrão σ da população. Se o desvio padrão σ da população não for conhecido, ele deverá ser estimado com o desvio padrão S da amostra retirada da população.

Teste de hipóteses com o intervalo de confiança

O conceito e o procedimento de estimação da média da população estudado no Capítulo 11 será aplicado no teste de hipóteses. Lembremos que para estimar a média de uma população, começamos por retirar uma amostra de tamanho adequado dessa população. A média e o desvio padrão dessa amostra e o intervalo de confiança adotado determinarão o intervalo de valores que deverá incluir a média da população. Por exemplo, escolhendo o intervalo de confiança 95%, teremos condições de afirmar que a probabilidade da média da população estar incluída no intervalo estimado da média é 95%. Enquanto no procedimento de estimação da média de uma população, o objetivo é estimar o valor da média da população, no teste de hipóteses, a média da população é um valor conhecido.

O procedimento a seguir deve ser utilizado para realizar um teste de hipóteses utilizando o intervalo de confiança:

1. Estabeleça as hipóteses nula e alternativa, H_0 e H_1 .
2. Adote o intervalo de confiança, por exemplo, 95%. Em geral, nos testes de hipóteses se trabalha com o erro tolerado, denominado também de *nível de significância* α , relacionado com o intervalo de confiança pela expressão $(1-\alpha)$, em valores unitários.
3. Estime a média da população. Com a média da amostra \bar{X} , o desvio padrão σ da população ou da amostra S_X , e o tamanho n da amostra retirada da população, construímos o intervalo de valores no qual se espera que a média da população esteja incluída. Para amostras grandes, em geral $n > 30$, pode ser utilizada a distribuição Z e, para amostras pequenas, a distribuição t , lembrando que a distribuição da população é normal, como foi visto no Capítulo 11.
4. Verifique se a média μ_0 , estabelecida na hipótese nula H_0 , está incluída no intervalo da estimativa.
 - Se a média μ_0 estiver incluída, então há evidências de que μ_0 seja a média da população. Logo, deve-se aceitar a hipótese nula H_0 e rejeitar a hipótese alternativa H_1 .
 - Se a média μ_0 não estiver incluída, então há evidências de que μ_0 não seja a média da população. Logo, deve-se rejeitar a hipótese nula H_0 e aceitar a hipótese alternativa H_1 .

EXEMPLO 12.4

Continuando com o Exemplo 12.2. Verifique se a afirmação do pessoal de marketing é significativa, considerando o intervalo de confiança de 95%.

Solução. Aplicando o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu = \$2.500$$

$$H_1: \mu \neq \$2.500$$

2. O intervalo de confiança é 95% e, conseqüentemente, o erro tolerado ou nível de significância α é 0,05 ou 5%.
3. Como o tamanho da amostra n é 50, será utilizada a distribuição Z . Para o intervalo de confiança 95% e conseqüente nível de significância de 5%, os valores críticos de Z são $Z = \pm 1,96$. Lembrando o que foi visto no Capítulo 11, o Z crítico positivo pode ser obtido com a fórmula =INV.NORMP(0,975). O intervalo dos

valores no qual se espera que a média da população esteja incluída é $\$2590 \pm \79 , resultado obtido com a fórmula:

$$\mu = \bar{X} \pm Z_{0,05/2} \times \frac{S_X}{\sqrt{n}}$$

$$\mu = \$2590 \pm 1,96 \times \frac{\$285}{\sqrt{50}} = \$2590 \pm \$79$$

E os limites, inferior e superior, do intervalo da estimativa da média da população são, respectivamente, \$2.511 e \$2.669.

4. Como o valor da média da população \$2.500 não está incluído no intervalo da estimativa da média da população, a média \$2.590 da amostra é significativamente maior do que \$2.500. Concluindo, há evidências que recomendam rejeitar H_0 e aceitar H_1 ou, de outra maneira, há evidências que recomendam aceitar \$2.590 como renda mensal dos associados.

A partir do resultado da estimativa da média da população μ , há formas equivalentes de divulgar a conclusão do teste de hipóteses:²

- Se a média μ_0 definida na hipótese nula H_0 estiver contida no intervalo da estimativa da média da população, então pode-se dizer que:
 - A hipótese nula H_0 deve ser aceita e a hipótese alternativa H_1 deve ser rejeitada. Aceitar a hipótese nula H_0 significa que não há evidências suficientes para rejeitá-la e, portanto, H_0 deve ser verdadeira. Observe que não é afirmado que a hipótese nula H_0 seja verdadeira.
 - A média da amostra \bar{X} não é significativamente diferente de μ_0 .
 - É razoável aceitar que a diferença entre a média da amostra \bar{X} e μ_0 seja somente devida à amostra aleatória escolhida.
 - O resultado não é estatisticamente significativo.
- Se a média μ_0 , definida na hipótese nula H_0 , não estiver contida no intervalo da estimativa da média da população, então pode-se dizer que:
 - A hipótese nula H_0 deve ser rejeitada e a hipótese alternativa H_1 deve ser aceita. Aceitar a hipótese alternativa H_1 significa, apenas, que há evidências de que H_0 seja falsa.
 - A média da amostra \bar{X} é significativamente diferente de μ_0 .
 - Não é razoável aceitar que a diferença entre a média \bar{X} da amostra e μ_0 seja somente devida à amostra aleatória escolhida.
 - O resultado é estatisticamente significativo. Que o resultado seja estatisticamente significativo quer dizer que as evidências contra a hipótese nula alcançaram o erro tolerado ou nível de significância α .

EXEMPLO 12.5

Os consultores de empresas afirmam que os principais executivos das 500 maiores empresas do país trabalham 64 horas por semana. A tabela seguinte registra as horas trabalhadas de uma amostra de doze executivos escolhidos aleatoriamente dentre as 500 maiores empresas.

60	66	64	62	58	62	62	60	62	60	64	66
----	----	----	----	----	----	----	----	----	----	----	----

Verifique se a afirmação dos consultores é significativa, considerando que a distribuição de frequências das horas trabalhadas é aproximadamente normal e o intervalo de confiança 95%.

² Dentro de cada grupo, as afirmações são equivalentes.

Solução. Aplicando o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu=64$$

$$H_1: \mu \neq 64$$

2. O intervalo de confiança é 95% e o nível de significância α é 0,05 ou 5%.
3. A média e o desvio padrão da amostra são, respectivamente, 62,17 horas e 2,48 horas. Como a amostra é pequena com distribuição aproximadamente normal, para o nível de significância 0,05 nas duas caudas e $gl=11$ graus de liberdade, o t crítico é igual a 2,201, resultado obtido com a fórmula =INVT(0,05;11). O intervalo dos valores no qual se espera que a média da população esteja incluída é $62,17 \pm 1,576$, resultado obtido com a fórmula:

$$\mu = \bar{X} \pm t_{0,05/2} \times \frac{S_X}{\sqrt{n}}$$

$$\mu = 62,17 \pm 2,201 \frac{2,48}{\sqrt{12}} = 62,17 \pm 1,576$$

E os limites, inferior e superior, do intervalo da estimativa da média da população são, respectivamente, 60,59 e 63,75 horas.

4. Como a média da população 64 não está incluída no intervalo da estimativa da média da população, a média 62,17 da amostra é significativamente menor do que 64. Concluindo, há evidências que recomendam rejeitar H_0 e aceitar H_1 ou, de outra maneira, há evidências que recomendam rejeitar 64 e aceitar 62,17 horas. As evidências sugerem que, tolerando um erro de 5% na afirmação, os executivos trabalham menos do que 64 horas por semana.

Modelo TH com intervalo de confiança

Na planilha **Modelo TH com Intervalo**, incluída na pasta **Capítulo 11**, foi construído o *modelo* para teste de hipóteses de média aplicando o intervalo de confiança com a distribuição t e a distribuição Z , como mostra a Figura 12.1, resolvendo o Exemplo 12.4. O modelo conta com duas caixas de grupo com duas opções cada uma:

- Na caixa de grupo **Distribuição da população**, pode-se escolher **Normal** ou **Não é normal** clicando no botão de opção correspondente.
- Na caixa de grupo **Estimativa com a**, pode-se escolher **Dist. Normal** ou **Dist. t** clicando no botão de opção correspondente.

Ainda, com o modelo, é possível obter resultados utilizando o intervalo de confiança IC ou o nível de significância $Alfa$, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E7:F7 sempre que o IC for menor do que 80% e $Alfa$ for maior do que 20%.

O **Modelo TH com Intervalo** fornece os resultados sempre que a distribuição da população for normal ou, caso não seja normal, para tamanho de amostra maior ou igual a 31. Se os dados não atenderem às premissas do modelo, os resultados relevantes não serão apresentados na planilha. Para melhorar a compreensão da decisão, no intervalo E10:F10 foi incluído o resultado da decisão por extenso, aceitar ou rejeitar a hipótese nula que no Exemplo 12.4 é rejeitada.

O gráfico incluído no *modelo* mostra o intervalo de confiança em cor verde, que indica o intervalo de aceitação da hipótese nula. Os dois segmentos em cor vermelha mostram os intervalos de rejeição da hipótese nula. O pequeno triângulo em cor escura posiciona o valor da hipótese nula. A Figura 12.2 mostra a resolução do Exemplo 12.5.

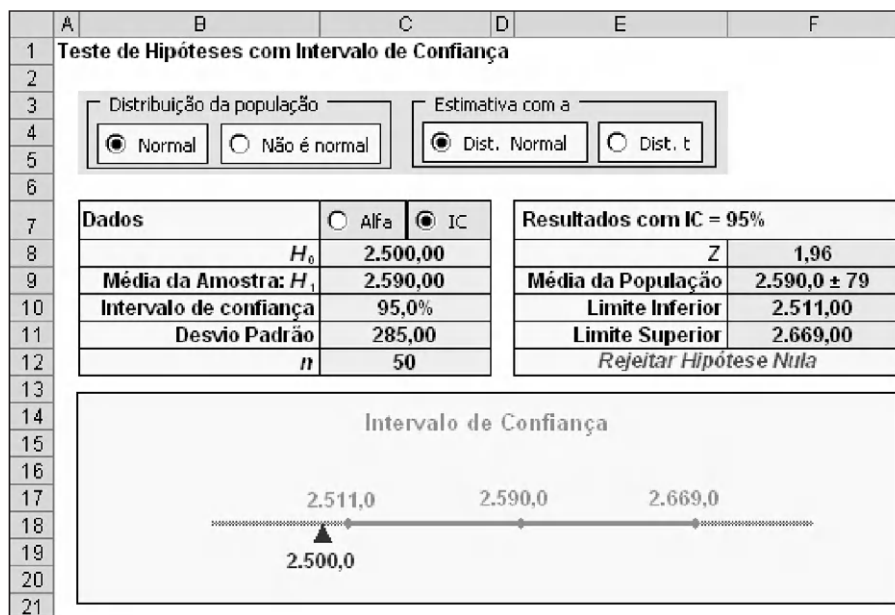


FIGURA 12.1
Modelo TH com
Intervalo resolvendo
o Exemplo 12.4.

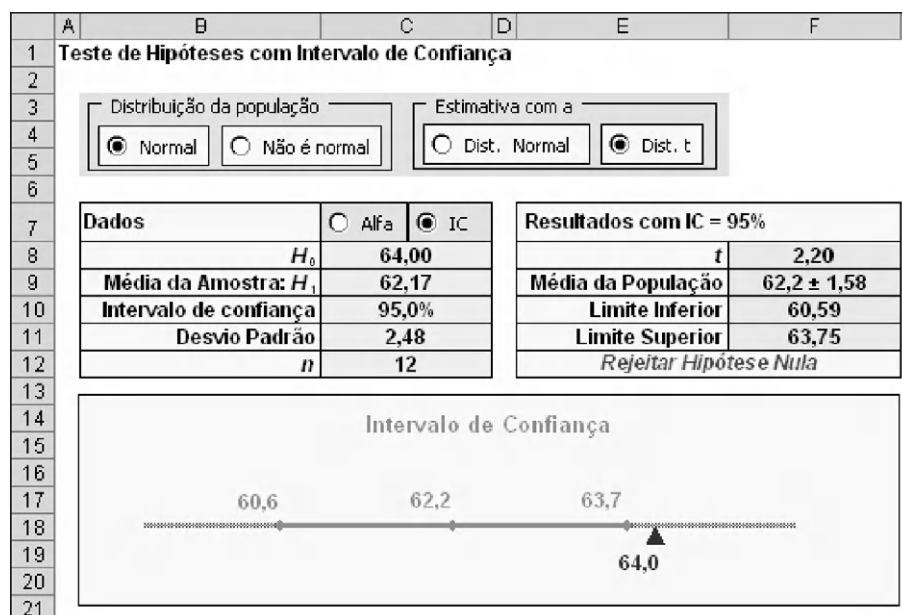


FIGURA 12.2
Modelo TH com
Intervalo resolvendo
o Exemplo 12.5.

Como regra geral, sempre que o desvio padrão da população não for conhecido e a população tiver distribuição normal, o teste de hipóteses deverá ser realizado com o desvio padrão da amostra e a distribuição t com $(n-1)$ graus de liberdade. Entretanto, se a inclinação da distribuição da população for pequena, para amostras de tamanho suficientemente grande, $n > 30$, a distribuição Z poderá ser utilizada para estimar a média da população utilizando o desvio padrão da amostra. Se a distribuição da população não for conhecida e o tamanho da amostra for pequeno, será necessário verificar a premissa de normalidade da população avaliando, por exemplo, a forma da distribuição de frequências ou outro método, assunto não tratado neste livro.

Teste de hipóteses com as distribuições Z e t

Ao estabelecer o intervalo de confiança da estimativa da média da população de 95%, por exemplo, ao mesmo tempo, são automaticamente definidos o nível de significância de 5%, o Z crítico 1,96 ou o t crítico correspondente ao número de graus de liberdade de cada tipo de problema. Nesta parte, será mostrado como utilizar os desvios padrão normalizados das distribuições Z e t para realizar testes de hipóteses.

Estabelecidas as duas premissas do teste de hipóteses, o valor observado Z_o ou t_o será comparado com o valor crítico Z_c ou t_c correspondente ao nível de significância α estabelecido, sendo que a escolha do tipo de distribuição Z ou t dependerá do tamanho da amostra e do julgamento do analista. O procedimento desse segundo método de teste de hipóteses da média de uma população é equivalente ao do intervalo de confiança:

1. Estabeleça as hipóteses nula e alternativa, H_0 e H_1 .
2. Adote o nível de significância α , valor relacionado com o intervalo de confiança $IC=(1-\alpha)$, em valores unitários. Por exemplo, o nível de significância $\alpha=0,05$ ou 5% é equivalente ao intervalo de confiança 95%.
3. Escolha a distribuição a ser utilizada. Dependendo do tamanho da amostra e do julgamento do analista, deverá ser escolhida a distribuição Z ou a distribuição t . Do nível de significância α estabelecido, será obtido o valor crítico Z_c ou t_c correspondente, utilizando a tabela Z ou a tabela t , ou a função estatística INV.NORMP para a distribuição Z e a função estatística INVT para a distribuição t .
 - Os dois valores críticos de Z ou de t definem o intervalo de aceitação da hipótese nula, situação semelhante ao teste de hipóteses utilizando o intervalo de confiança. As duas regiões laterais fora da anterior definem o intervalo de rejeição da hipótese nula.
4. Estime a média da população. Com a média da amostra \bar{X} , o desvio padrão σ da população ou da amostra S_X , o tamanho n da amostra retirada da população e a média da população μ_0 definida na hipótese nula, será calculado o valor observado Z_o ou t_o :

$$Z_o = \frac{\bar{X} - \mu_0}{S_X / \sqrt{n}} \quad \text{ou} \quad t_o = \frac{\bar{X} - \mu_0}{S_X / \sqrt{n}}$$

5. Compare o valor observado Z_o ou t_o com o valor crítico Z_c ou t_c . Como o valor observado pode ser positivo ou negativo, devem-se estabelecer critérios de decisão diferentes para aceitar ou rejeitar a hipótese nula.
 - Os dois possíveis resultados do teste de hipóteses na cauda superior da distribuição são:
 - Se $Z_o < Z_c$ ou $t_o < t_c$, o valor observado está dentro da área de aceitação da hipótese nula. Nesse caso, deve-se aceitar a hipótese nula; há evidências de que μ_0 seja a média da população.
 - Se $Z_o > Z_c$ ou $t_o > t_c$, o valor observado está fora da área de aceitação da hipótese nula. Nesse caso, não se deve aceitar a hipótese nula; há evidências de que μ_0 não seja a média da população.
 - Para a cauda inferior, têm-se dois resultados equivalentes:
 - Se $Z_o < Z_c$ ou $t_o < t_c$, o valor observado está fora da área de aceitação da hipótese nula e, nesse caso, não se deve aceitar a hipótese nula.
 - Se $Z_o > Z_c$ ou $t_o > t_c$, o valor observado está dentro da área de aceitação da hipótese nula e, nesse caso, deve-se aceitar a hipótese nula.

Querendo estabelecer e aplicar um critério único no teste de hipóteses incluindo as duas caudas da distribuição, deve-se comparar os valores absolutos $|Z_o|$ e $|Z_c|$. Se $|Z_o| > |Z_c|$ ou $|t_o| > |t_c|$, o valor observado está fora da área de aceitação da hipótese nula. Nesse caso, não se deve aceitar a hipótese nula; há evidências de que μ não seja a média da população. A Figura 12.3 apresenta as regiões de aceitação e rejeição da hipótese nula H_0 , para o nível de significância α de 0,05, nos três casos possíveis, na cauda superior, na cauda inferior e nas duas caudas da distribuição Z , sendo o raciocínio equivalente para a distribuição t .

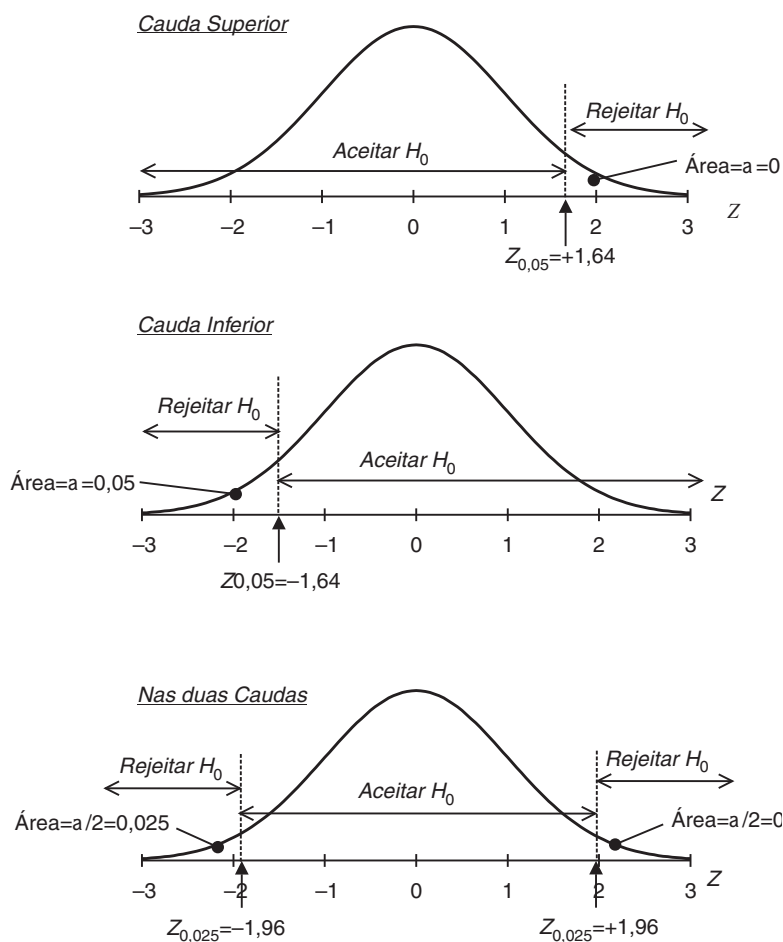


FIGURA 12.3 Regiões de aceitação e rejeição da hipótese nula.

EXEMPLO 12.6

Resolva o Exemplo 12.3 aplicando a distribuição Z .

Solução. Seguindo o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu = 2.500$$

$$H_1: \mu \neq 2.500$$

2. O intervalo de confiança 95% define o nível de significância $\alpha = 0,05$ nas duas caudas, ou 0,0250 em cada cauda.
3. Como o tamanho da amostra n é 50, será utilizada a distribuição Z . Para o intervalo de confiança 95% e consequente nível de significância de 5%, os valores críticos de Z são $Z = \pm 1,96$, resultado obtido com a fórmula =INV.NORMP(0,975).

4. Com a média da população \$2.500 definida na hipótese nula, a média da amostra \$2.590, o desvio padrão \$285 e o tamanho da amostra 50, é calculado o valor observado $Z_o = +2,23$, resultado obtido com a fórmula:

$$Z_o = \frac{2.590 - 2.500}{285/\sqrt{50}} = +2,23$$

Esse resultado pode também ser obtido utilizando a função estatística registrando a fórmula =PADRONIZAR(2590;2500;285/RAIZ(50)).

5. Como $Z_o = +2,23$ é maior do que $Z_c = +1,96$, deve-se rejeitar a hipótese nula e aceitar a hipótese alternativa. Há evidências de que a média da renda mensal da população de associados da empresa seja maior do que \$2.500.

Vimos que o valor de Z que divide as regiões de aceitação e rejeição é denominado valor crítico Z_c . Por exemplo, o valor crítico $Z_c = +1,96$ corresponde ao nível de significância de 0,05 para testes nas duas caudas, distribuindo 0,025 em cada cauda. Para testes de hipóteses numa cauda, o valor crítico $Z_c = +1,645$ corresponde ao nível de significância =0,05 para testes na cauda superior, e o valor crítico $Z_c = -1,645$ para testes na cauda inferior. Um resumo de valores críticos para a distribuição Z está registrado na Figura 12.4.

α	Z		
	Cauda inferior	Cauda superior	Duas caudas
0,10	-1,281	+1,281	$\pm 1,645$
0,05	-1,645	+1,645	$\pm 1,960$
0,02	-1,960	+1,960	$\pm 2,326$
0,01	-2,326	+2,326	$\pm 2,576$

FIGURA 12.4 Relação de α e Z para uma e duas caudas.

Para testes de hipóteses com a distribuição t , procede-se da mesma forma como foi realizado com a distribuição Z . Entretanto, como o valor crítico t_c depende do número de graus de liberdade $gl=n-1$, para cada valor de nível significância α , tem-se mais de um valor de t_c .

EXEMPLO 12.7

Resolva o Exemplo 12.5 aplicando a distribuição t .

Solução. Seguindo o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu=64$$

$$H_1: \mu \neq 64$$

2. O intervalo de confiança 95% define o nível de significância $\alpha=0,05$ nas duas caudas, ou 0,0250 em cada cauda.
3. Como a distribuição da população é aproximadamente normal e o tamanho da amostra n é 12, será utilizada a distribuição t . Para o número de graus de liberdade igual a 11, obtém-se o valor crítico t_c igual a -2,201. Esse resultado pode ser obtido com a fórmula =INVT(0,05;11), porém adicionando o sinal negativo devido à simetria da distribuição.
4. Com a média da população 64 definida na hipótese nula, a média da amostra 62,17, o desvio padrão 2,48 e o tamanho da amostra 12, é calculado o valor observado t_o igual a -2,5562, resultado obtido com a fórmula:

$$t_o = \frac{62,17 - 64}{2,48/\sqrt{12}} = -2,5562$$

Esse resultado pode também ser obtido utilizando a função estatística, registrando a fórmula =PADRONIZAR(64;62,17;2,48/RAIZ(12)).

5. Utilizando as duas formas de decidir, ambas com o mesmo resultado.

- Como $t_o = -2,56$ é menor do que $t_c = -2,20$, deve-se rejeitar a hipótese nula e aceitar a hipótese alternativa.
- Como alternativa, como $|t_o| = |-2,56|$ é maior do que $|t_c| = |-2,20|$, deve-se rejeitar a hipótese nula e aceitar a hipótese alternativa.

Há evidências de que os principais executivos das 500 maiores empresas do país trabalhem menos do que 64 horas por semana.

Modelo TH com valores críticos de Z e t

Na planilha **Modelo TH com Z e t**, incluída na pasta **Capítulo 11**, foi construído o *modelo* para teste de hipóteses de média comparando o valor observado com o valor crítico da distribuição *t* e da distribuição *Z*, como mostra a Figura 12.5 resolvendo o Exemplo 12.6. O modelo conta com duas caixas de grupo com duas opções cada uma:

- Na caixa de grupo **Distribuição da população**, pode-se escolher **Normal** ou **Não é normal**, clicando no botão de opção correspondente.
- Na caixa de grupo **Estimativa com a**, pode-se escolher **Dist. Normal** ou **Dist. T**, clicando no botão de opção correspondente.

Ainda, com o modelo é possível obter resultados utilizando o intervalo de confiança *IC* ou o nível de significância *Alfa*, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E7:F7 sempre que o *IC* for menor do que 80% e *Alfa* for maior do que 20%.

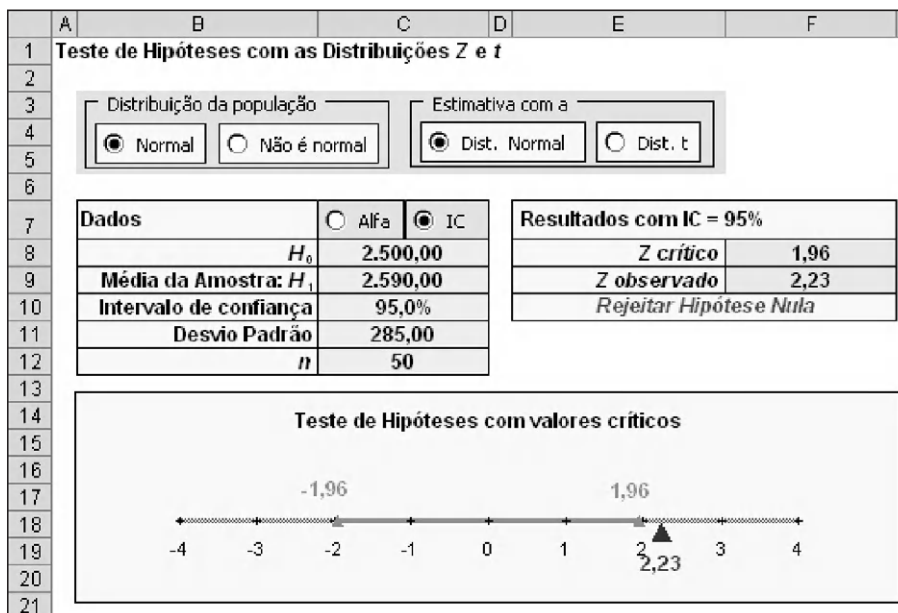
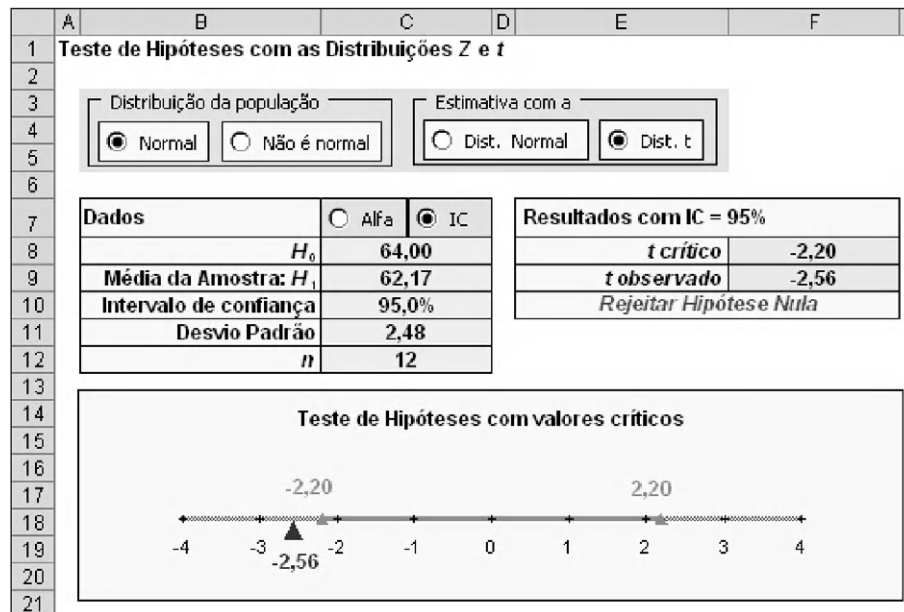


FIGURA 12.5
Modelo TH com
Z e t resolvendo o
Exemplo 12.6.

O Modelo TH com Z e t fornece os resultados sempre que a distribuição da população for normal ou, caso não seja normal, para tamanho de amostra maior ou igual a 31. Se os dados não atenderem às premissas do modelo, os resultados relevantes não serão apresentados na planilha. Para melhorar a compreensão da decisão, no intervalo E10:F10 foi incluído o resultado da decisão por extenso, aceitar ou rejeitar a hipótese nula que no Exemplo 12.6 é rejeitada.

A região verde do gráfico incluído no *modelo* mostra o intervalo de aceitação, e os dois segmentos em cor vermelha mostram os intervalos de rejeição da hipótese nula. O pequeno triângulo em cor escura posiciona o valor crítico observado, nesse caso Z_o . A Figura 12.6 mostra a resolução do Exemplo 12.7.

FIGURA 12.6
Modelo TH com
Z e t resolvendo o
Exemplo 12.7.



Teste de hipóteses com *p-value*

Nos testes de hipóteses com as distribuições Z ou t, o valor observado é comparado com o valor crítico da distribuição escolhida. Essa comparação é realizada depois de o analista ter adotado o nível de significância α , que define a região de rejeição da hipótese nula, independente dos resultados da amostra.

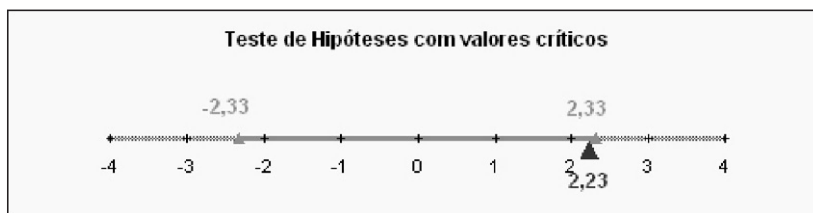
EXEMPLO 12.8

Repita o Exemplo 12.6, considerando $\alpha=0,02$ nas duas caudas e aplicando o teste de hipóteses com o valor crítico da distribuição Z.

Solução. Lembremos as hipóteses: $H_0: \mu=\$2.500$ e $H_1: \mu\neq\$2.500$. Aplicando o procedimento conhecido:

- O $Z_{observado}$ é o mesmo do Exemplo 12.6, $Z_o=+2,233$.
- Para o nível de significância $\alpha=0,02$ nas duas caudas, o Z crítico positivo é $Z_c=+2,326$, resultado obtido com a fórmula =INV.NORMP(0,99).
- Utilizando as duas formas de decidir, ambas com o mesmo resultado.
 - Como $Z_o=2,233$ é menor do que $Z_c=2,326$, deve-se aceitar a hipótese nula e rejeitar a hipótese alternativa.
 - Como $|Z_o|=2,233$ é menor do que $|Z_c|=2,326$, deve-se aceitar a hipótese nula e rejeitar a hipótese alternativa.

A decisão pode ser obtida com o modelo **TH com Z e t**, cujo gráfico é mostrado a seguir.



O Exemplo 12.8 mostra que, ao diminuir o nível de significância de 0,05 para 0,02, a média da amostra deixou de ser significativa, devendo-se aceitar a hipótese nula, como mostra a Figura 12.7.

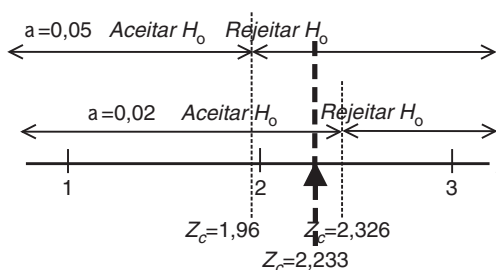


FIGURA 12.7 Teste de hipóteses do Exemplo 12.8 para $\alpha=0,05$ e $\alpha=0,02$.

O nível de significância α estabelece o erro tolerado e, ao mesmo tempo, define a região de rejeição da hipótese nula, do valor crítico até infinito, nas duas caudas da distribuição. Para um determinado nível de significância α , o resultado do teste de hipóteses será estatisticamente significativo se a hipótese nula for rejeitada. Portanto, se um determinado nível de significância α rejeitar a hipótese nula, então qualquer nível de significância maior do que esse α também rejeitará a hipótese nula. No Exemplo 12.6, para $\alpha=0,05$, o resultado do teste de hipóteses é estatisticamente significativo, pois rejeita a hipótese nula. Então, qualquer nível de significância maior do que 5% também rejeitará a hipótese, ou o resultado do teste de hipóteses será significativo, como se pode observar praticamente mudando o valor de alfa na planilha **Modelo TH com Z e t**. Entretanto, para o $\alpha=0,02$ no Exemplo 12.8, o resultado não é estatisticamente significativo, pois aceita a hipótese nula.

Pode-se observar que o nível de significância correspondente ao Z observado é o maior valor de nível de significância que rejeita a hipótese nula, pois, para valores menores, o resultado do teste não é significativo.³ Dessa maneira, apenas rejeitarão a hipótese nula os níveis de significância menores do que o nível de significância correspondente ao Z observado. Todo esse raciocínio foi desenvolvido para mostrar as bases do *p-value*.

Definição do *p-value*

O *p-value*⁴ é definido como a probabilidade de qualquer média da amostra ser mais extrema do que a média da amostra \bar{X} extraída para o teste, sem rejeitar a hipótese nula. Do exposto e da definição de *p-value* temos:

- O *p-value* é o nível de significância observado.
- Se o *p-value* for maior ou igual a α , então a hipótese nula será aceita.
- Se o *p-value* for menor ou igual a α , então a hipótese nula será rejeitada. Quanto menor for o *p-value*, mais forte será a evidência para rejeitar a hipótese nula.

³ Este valor é denominado *α de reversão* da decisão de aceitar a hipótese nula.

⁴ Mantemos o nome *p-value* em inglês por ser normalmente utilizado.

- A decisão do teste de hipóteses será resultado da comparação do p -value com o nível de significância α que o analista julgar mais adequado.

Cálculo do p -value

De uma população com média μ_0 foi retirada uma amostra de tamanho n com média \bar{X} e desvio padrão S_X .

Cálculo do p -value com a distribuição Z.

Dos dados da amostra, é obtido o Z observado Z_0 .

- Teste nas duas caudas da distribuição Z. O p -value é a probabilidade calculada com um dos dois procedimentos seguintes:
 - Se $\bar{X} \geq \mu_0$, então $p\text{-value} = 2 \times P(Z \geq Z_0)$.
 - Se $\bar{X} \leq \mu_0$, então $p\text{-value} = 2 \times P(Z \leq Z_0)$.

Utilizando os valores absolutos de Z e de Z_0 , pode ser utilizada uma única fórmula $p\text{-value} = 2 \times P(|Z| \geq |Z_0|)$.

- Teste em uma cauda na distribuição Z. O p -value é a probabilidade calculada com um dos dois procedimentos seguintes:
 - Se $\bar{X} \geq \mu_0$, então $p\text{-value} = P(Z \geq Z_0)$.
 - Se $\bar{X} \leq \mu_0$, então $p\text{-value} = P(Z \leq Z_0)$.

Utilizando os valores absolutos de Z e de Z_0 , pode ser utilizada uma única fórmula $p\text{-value} = P(|Z| \geq |Z_0|)$.

Cálculo do p -value com a distribuição t.

Dos dados da amostra é obtido o t observado t_0 .

- Teste nas duas caudas da distribuição t. O p -value é a probabilidade calculada com um dos dois procedimentos a seguir:
 - Se $\bar{X} \geq \mu_0$, então $p\text{-value} = 2 \times P(t \geq t_0)$.
 - Se $\bar{X} \leq \mu_0$, então $p\text{-value} = 2 \times P(t \leq t_0)$.

Utilizando os valores absolutos de t e de t_0 , pode ser utilizada uma única fórmula $p\text{-value} = 2 \times P(|t| \geq |t_0|)$.

- Teste numa cauda na distribuição t. O p -value é a probabilidade calculada com um dos dois seguintes procedimentos:
 - Se $\bar{X} \geq \mu_0$, então $p\text{-value} = P(t \geq t_0)$.
 - Se $\bar{X} \leq \mu_0$, então $p\text{-value} = P(t \leq t_0)$.

Utilizando os valores absolutos de t e de t_0 , pode ser utilizada uma única fórmula $p\text{-value} = P(|t| \geq |t_0|)$.

O procedimento de decisão utilizando o p -value é o seguinte:

1. Calcule o Z ou t observado e o p -value.
2. Escolha o nível de significância α .
3. Se o p -value for menor do que α , deve-se rejeitar a hipótese nula.

O procedimento *p-value* ajuda a compreender a força da decisão. O procedimento apresenta como resultado o nível de significância observado, deixando por conta do analista a decisão de escolher o máximo α tolerado.

Também facilita a apresentação de resultados usando *softwares*, como o leitor poderá ver na planilha Modelo TH com *p-value*.

EXEMPLO 12.9

Realize o teste de hipóteses do Exemplo 12.6 com o *p-value*.

Solução. Seguindo o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu = 2.500$$

$$H_1: \mu \neq 2.500$$

2. Com a média da população \$2.500 definida na hipótese nula, a média da amostra \$2.590, o desvio padrão \$285 e o tamanho da amostra 50, é calculado o valor observado $Z_o = +2,23$, resultado obtido com a fórmula:

$$Z_o = \frac{2.590 - 2.500}{285 / \sqrt{50}} = +2,23$$

Esse resultado pode também ser obtido utilizando a função estatística, registrando a fórmula =PADRONIZAR(2590;2500;285/RAIZ(50)).

3. O cálculo do *p-value* deve ser realizado com $p\text{-value} = 2 \times P(Z \geq Z_o)$. Para calcular a probabilidade de Z ser maior do que o Z observado Z_o , foi utilizada a fórmula =(1-DIST.NORMP(2,233)) que retornou $P(Z \geq 2,233) = 0,012774$. O resultado procurado é $p\text{-value} = 2 \times 0,012774 = 0,0255$.
4. O intervalo de confiança 95% define o nível de significância $\alpha = 0,05$ nas duas caudas, ou 0,0250 em cada cauda.
5. Como o *p-value* é menor do que o nível de significância 0,05, a hipótese nula deve ser rejeitada. O $p\text{-value} = 0,0255$ é o maior valor de que rejeita a hipótese nula, como o leitor pode verificar na Figura 12.7.

EXEMPLO 12.10

Resolva o Exemplo 12.5 com o *p-value*.

Solução. Seguindo o procedimento apresentado:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu = 64$$

$$H_1: \mu \neq 64$$

2. Com a média da população 64 definida na hipótese nula, a média da amostra 62,17, o desvio padrão 2,48 e o tamanho da amostra 12, é calculado o valor observado t_o igual a -2,5562, resultado obtido com a fórmula:

$$t_o = \frac{62,17 - 64}{2,48 / \sqrt{12}} = -2,5562$$

Esse resultado pode também ser obtido utilizando a função estatística registrando a fórmula =PADRONIZAR(64;62,17;2,48/RAIZ(12)).

3. O cálculo do *p-value* deve ser realizado com $p\text{-value} = 2 \times P(t \leq t_o)$. Para calcular a probabilidade de t ser menor do que o t observado t_o , foi utilizada a fórmula =DISTT(2,5562;11;1), informando o valor absoluto de t_o . Essa fórmula retornou o resultado $P(t \leq -2,5562) = 0,01335$ e $p\text{-value} = 2 \times 0,01335 = 0,0267$.
4. O intervalo de confiança 95% define o nível de significância $\alpha = 0,05$ nas duas caudas, ou 0,0250 em cada cauda.
5. Como o *p-value* é menor que o nível de significância 0,05, a hipótese nula deve ser rejeitada. O $p\text{-value} = 0,0267$ é o maior valor de que rejeita a hipótese nula, como você pode verificar na Figura 12.9. Há evidências de que os principais executivos das 500 maiores empresas do país trabalhem menos do que 64 horas por semana.

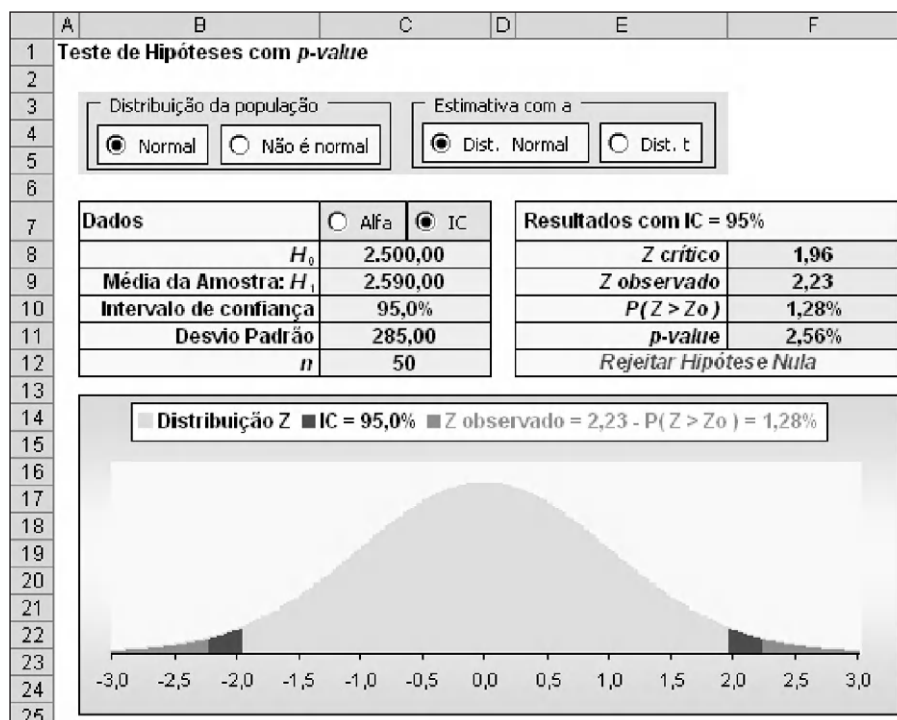
Modelo TH com p -value

Na planilha **Modelo TH com p -value**, incluída na pasta **Capítulo 11**, foi construído o *modelo* para teste de hipóteses de média comparando o p -value com o nível de significância, utilizando a distribuição Z ou t , como mostra a Figura 12.8, resolvendo o Exemplo 12.9 com a distribuição Z . O modelo conta com duas caixas de grupo com duas opções cada uma:

- Na caixa de grupo **Distribuição da população**, escolha **Normal** ou **Não é normal** clicando no botão de opção correspondente.
- Na caixa de grupo **Estimativa com α** , escolha **Dist. Normal** ou **Dist. T**, clicando no botão de opção correspondente. Observe a diferença das formas das duas distribuições para tamanho de amostras menores do que 30.

Ainda, com o modelo, é possível obter resultados utilizando o intervalo de confiança IC ou o nível de significância $Alfa$, selecionando o botão de opção requerido. Para evitar interpretações errôneas, o título dos resultados destaca o tipo de dado selecionado, mudando para a cor vermelha com letras amarelas as células do intervalo E7:F7 sempre que o IC for menor do que 80% e $Alfa$ for maior do que 20%.

FIGURA 12.8
Modelo TH com p -value resolvendo o Exemplo 12.9.



O **Modelo TH com p -value** fornece os resultados sempre que a distribuição da população for normal ou, caso não seja normal, para tamanho de amostra maior ou igual a 31. Se os dados não atenderem às premissas do modelo, os resultados relevantes não serão apresentados na planilha. Na célula F10, o modelo apresenta a probabilidade de exceder o valor observado, seja em sentido positivo ou negativo. Na célula F11, é apresentado o resultado p -value. Para melhorar a compreensão da decisão, no intervalo E12:F12 foi incluído o resultado da decisão por extenso, aceitar ou rejeitar a hipótese nula que no Exemplo 12.9 é rejeitada.

A região central do gráfico (pintada de cor verde claro) incluído no *modelo* não é utilizada no cálculo do p -value. Ao definir o intervalo de confiança ou o nível de significância nas duas caudas da distribui-

ção escolhida ficam definidas duas áreas da distribuição (pintadas de cor vermelha), uma em cada cauda. Com o valor observado, é calculada a probabilidade de exceder esse valor, seja em sentido positivo ou negativo, definindo uma área das duas caudas do lado adequado da distribuição. Entretanto, como o cálculo do *p-value* é o resultado de duplicar aquele resultado de probabilidade, o gráfico do modelo apresenta as duas caudas da distribuição pintadas de cor verde escuro. Para a decisão com o gráfico, sempre que o gráfico mostrar alguma região das caudas pintadas de cor vermelha, a hipótese nula deverá ser rejeitada, decisão que coincidirá com a obtida da comparação de probabilidades das células C10 e F11 e descrita por extenso na célula E12. O leitor atento perceberá que certas decisões serão mais bem definidas no resultado numérico do que no gráfico da distribuição, devido apenas à baixa resolução do gráfico construído com Excel para esse tipo de decisão. A Figura 12.9 mostra a resolução do Exemplo 12.10 com a distribuição *t*.

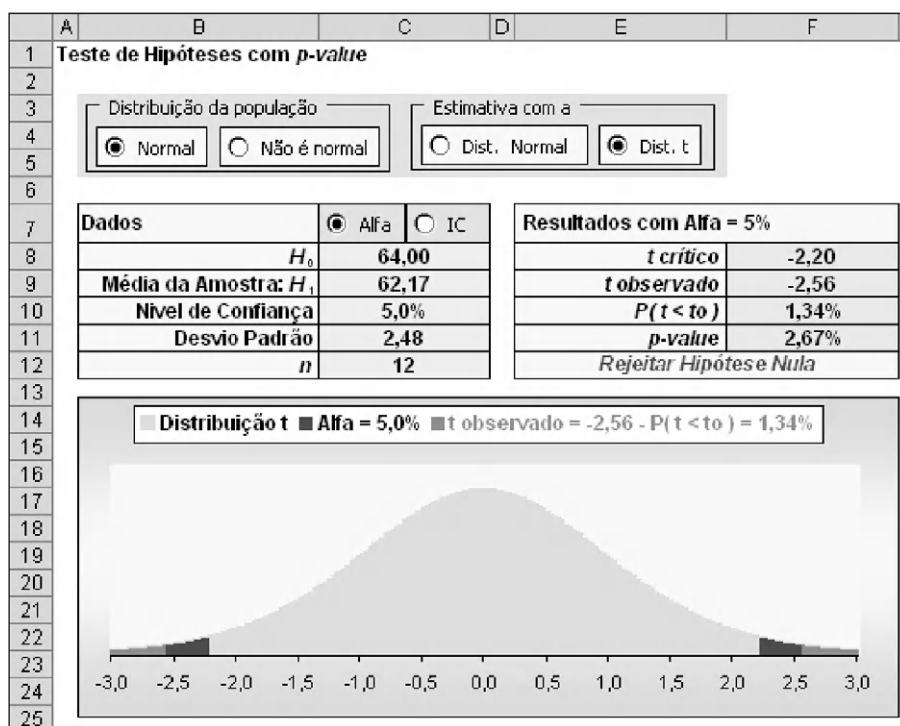


FIGURA 12.9
Modelo TH com
p-value resolvendo
o Exemplo 12.7.

Função teste Z

EXEMPLO 12.11

A amostra aleatória da tabela seguinte foi retirada de uma população com distribuição normal, com média de 31,3 e desvio padrão igual a 5,2.

41	35	25	36	40	36	24	37	28	35	27	33
36	27	33	32	32	42	31	43	30	30	38	38
26	34	42	45	26	23	32	43	22	37	26	32

Verifique se deve aceitar o valor da média da população considerando o nível de significância $\alpha=0,05$ nas duas caudas da distribuição.

Solução. Para realizar o teste de hipóteses aplicando o *p-value* é necessário conhecer o tamanho e a média da amostra. Da amostra registrada no intervalo B4:B39 da planilha **Função TESTZ**, incluída na pasta **Capítulo 12**, foram obtidos os valores do tamanho e da média da amostra, respectivamente 36 e 33,25, resultados obti-

dos nas células E8 e E9 utilizando as funções estatísticas do Excel registradas nessas células. Seguindo o procedimento apresentado para teste de hipóteses com o *p-value*:

1. O teste de hipóteses é estabelecido da seguinte maneira:

$$H_0: \mu=31,30$$

$$H_1: \mu \neq 31,30$$

2. Com a média da população 31,30, a média da amostra 33,25, o desvio padrão da população 5,2 e o tamanho da amostra 36, é calculado o valor observado $Z_o=2,25$ com a fórmula =PADRONIZAR(33,25;31,3;5,2/RAIZ(36)), resultado obtido na célula E11 dessa planilha.
3. O cálculo do *p-value* deve ser realizado com $p\text{-value}=2 \times P(Z \geq Z_o)$. Para calcular a probabilidade de Z ser maior do que o Z observado, foi utilizada a fórmula =1-DIST.NORMP(2,25), que retornou $P(Z \geq 2,25)=0,0122244$ na célula E12. O *p-value* é 0,02445 ou 2,44%, resultado obtido na célula E13.
4. Como o *p-value* é menor do que o nível de significância 0,05, a hipótese nula deve ser rejeitada.

O procedimento de cálculo anterior foi realizado na planilha **TESTEZ**, incluída na pasta **Capítulo 12**, como mostra a figura seguinte.

	A	B	C	D	E
1	Função TESTEZ				
2					
3		Amostra	Dados		
4		41	Média População		31,3
5		35	DP população		5,2
6		25			
7		36	Resultados		
8		40	Tamanho Amostra		36
9		36	Média da Amostra		33,25
10		24	DP amostra		6,26
11		37	Z observado		2,2500
12		28	P (Z >= 2,25)		1,2224%
13		35	p-value		2,44%
14		27	Função TESTEZ		1,2224%
15		33			

Na célula E14, foi obtido o resultado da função estatística TESTEZ incluída no Excel para calcular o *p-value* nas duas caudas, resultado que não coincide com o esperado, como o leitor pode ver nessa planilha.

• **TESTEZ(matriz; x; desv_padrao)**

A função estatística TESTEZ⁵ retorna a probabilidade de que aconteça um valor mais extremo do que o Z observado, calculado com os valores da amostra registrados no argumento *matriz*, a média informada no argumento *x* e o desvio padrão registrado no argumento *desv_padrao*. Tenha-se presente:

- O argumento *x* é a média da população, valor afirmado na hipótese nula do teste de hipóteses.
- Se o argumento *desv_padrao* for omitido, a função TESTEZ utilizará o desvio padrão da amostra.
- O retorno da função TESTEZ é o mesmo retorno da fórmula seguinte:

A função utiliza a seguinte fórmula $=1-DIST.NORMP\left(\frac{\bar{X}-x}{\sigma/\sqrt{n}}\right)$, utilizando o desvio padrão correspon-

dente. Essa fórmula retorna a probabilidade de que aconteça um valor mais extremo do que a média da amostra \bar{X} .

- Portanto, a função TESTEZ retorna somente a probabilidade na cauda direita da distribuição.⁶ Para obter o *p-value* nas duas caudas, deve-se multiplicar por dois o resultado da função.
- Se o Z observado for negativo, a função TESTEZ retornará a probabilidade complementar.

⁵ Em inglês, a função TESTEZ é ZTEST.

⁶ A descrição da função TESTEZ da *ajuda on-line* do Excel não é clara, pois define o resultado dessa função como o *p-value* nas duas caudas da distribuição Z quando, na realidade, o valor retornado pela função se refere a uma cauda.

Erros no teste de hipóteses

A decisão de um teste de hipóteses é aceitar ou rejeitar a hipótese nula H_0 . Qualquer que seja a decisão, procura-se tomar a decisão correta, sabendo que isso não será possível 100% das vezes. Como a decisão tomada é baseada em amostras, nunca teremos certeza de qual das duas hipóteses é a hipótese *realmente* verdadeira, salvo que seja amostrada toda a população. Deve-se lembrar de que:

- Se H_0 for rejeitada, o teste de hipóteses não afirma que H_0 seja falsa, o teste de hipóteses sugere que há evidências de que H_0 seja falsa.
- O que se pode afirmar é que, por exemplo, para o nível de significância 0,10, espera-se que em 90% das vezes a conclusão obtida seja correta. De outra maneira, se o teste fosse repetido um número muito grande de vezes, espera-se que a proporção de acertos seja 90%.

Foi visto que o procedimento *p-value* apresenta como resultado o nível de significância observado, deixando por conta do analista a decisão de escolher o máximo α tolerado, o erro admitido pelo analista. O nível de significância α é uma medida do risco admitido no caso de rejeitar a hipótese nula sendo ela verdadeira. O Exemplo 12.12 ajuda a compreender quando ocorre um erro.

EXEMPLO 12.12

O controle de qualidade da montadora de carros estabeleceu que, de cada lote de oito mil eixos fornecidos pelo fabricante de autopeças, deverá ser retirada uma amostra de 40 eixos. O lote será aprovado somente se a proporção de eixos fora de especificação for menor do que 5%. Analise os possíveis resultados.

Solução. Analisemos duas situações limites.

- *A amostra tem quatro eixos fora de especificação.* Como a porcentagem de eixos fora de especificação é 10%, o lote de oito mil eixos será rejeitado.
 - Admitindo que a população de oito mil eixos fosse verificada, suponha que no lote de oito mil eixos estão fora de especificação somente os quatro eixos da amostra; isto é, apenas 0,50% do lote. Nesse caso, rejeitar o fornecimento foi um *erro*, pois o lote de oito mil eixos deveria ter sido aprovado.
 - Analisando como teste de hipóteses, a amostra forneceu evidências para *rejeitar a hipótese nula quando, na realidade, deveria ter sido aceita*. Em termos técnicos, foi cometido um *erro tipo I*.
- *A amostra tem um eixo fora de especificação.* Como a porcentagem de eixos fora de especificação é 2,5%, o lote de oito mil eixos deverá ser aceito.
 - Admitindo que a população de oito mil eixos fosse verificada, suponha que 10% do lote de oito mil estão fora de especificação. Nesse caso, aceitar o fornecimento foi um *erro*, pois deveria ter sido rejeitado.
 - Analisando como teste de hipóteses, a amostra forneceu evidências para *aceitar a hipótese nula quando, na realidade, deveria ter sido rejeitada*. Em termos técnicos, foi cometido um *erro tipo II*.

A tabela da Figura 12.10 relaciona os resultados do teste de hipóteses obtidos de uma amostra (linhas) com os verdadeiros resultados da população, (colunas).

	H_0 verdadeira	H_0 falsa
Aceita H_0	Decisão correta	<i>erro tipo II</i>
Rejeita H_0	<i>erro tipo I</i>	Decisão correta

FIGURA 12.10 Tipos de erro no teste de hipóteses.

Analisemos as probabilidades desses erros.

- O *erro tipo I* ocorre quando a hipótese nula é rejeitada sendo realmente verdadeira. Como a hipótese nula será rejeitada se o *p-value* for menor do que α , o risco de cometer o *erro tipo I* pode ser reduzido, diminuindo o nível de significância α . Portanto, a probabilidade de cometer o *erro tipo I* é o próprio nível de significância α , valor controlado pelo analista. Para o nível de significância α , a probabilidade de α de cometer um *erro tipo I*:

$$P(\text{Ocorrer erro tipo I}) = P(\text{Rejeitar } H_0 \text{ quando } H_0 \text{ for Verdadeira}) = \alpha$$

- O *erro tipo II* ocorre quando a hipótese nula é aceita sendo realmente falsa. Reduzindo o nível de significância α , poderá ser aceita a hipótese nula quando realmente for falsa e, ao mesmo tempo, aumentarão as chances de cometer o erro tipo II, aceitar H_0 quando realmente ela é falsa. A probabilidade de β de cometer um erro tipo II é:

$$P(\text{Ocorrer erro tipo II}) = P(\text{Aceitar } H_0 \text{ quando } H_0 \text{ for Falsa}) = \beta$$

Entre os dois tipos de erros há relação, quando α aumenta, β diminui e vice-versa, quando α diminui, β aumenta. O caminho para reduzir α e β simultaneamente é aumentar o tamanho da amostra.⁷ A tabela da Figura 12.11 mostra as probabilidades dos dois tipos de erro.

	Quando H_0 for verdadeira	Quando H_0 for falsa
Probabilidade de aceitar H_0	$1-\alpha$	β
Probabilidade de rejeitar H_0	α	$1-\beta$

FIGURA 12.11 Quantificação dos erros no teste de hipóteses.

Poder do teste

Vejamos uma aplicação dos erros no teste de hipóteses, aplicado ao primeiro exemplo do capítulo, destacando os seguintes resultados:

- Para o intervalo de confiança de 95%, o Exemplo 12.4 mostrou que a hipótese nula deve ser rejeitada, pois há evidências de que recomendam a renda mensal de \$2.500 dos associados da empresa administradora de cartões de crédito.
- Diminuindo o nível de significância α de 5% para 2%, o Exemplo 12.8 mostra que a hipótese nula deve ser aceita. Depois, no Exemplo 12.9, foi mostrado que o *p-value* igual a 2,55% é o maior valor que rejeita a hipótese nula do exemplo da renda mensal dos associados.

Em todos os exemplos apresentados, demos mais atenção ao erro tipo I, rejeição da hipótese nula verdadeira. Não demos atenção à probabilidade de cometer um erro tipo II, aceitação da hipótese nula falsa. Enquanto para um teste simples, o erro tolerado α é definido pelo analista, pois é uma medida do risco aceitado por rejeitar a hipótese nula verdadeira, a probabilidade β pode assumir valores diferentes.

⁷ Cometer um *erro tipo I* é mais sério do que cometer um *erro tipo II*.

EXEMPLO 12.13

As hipóteses do Exemplo 12.4 são:

$$H_0: \mu = \$2.500$$

$$H_1: \mu \neq \$2.500$$

Da amostra de tamanho 50 extraída da população, a média mensal da renda dos associados é \$2.590 e seu desvio padrão \$285.

Solução. O primeiro passo para estudar o *poder* do teste é a determinação dos limites do intervalo de aceitação da hipótese nula considerando, nesse caso, a distribuição $N(2.500; 285)$ e o intervalo de confiança de 95%.

- Com a distribuição Z , obtêm-se seus valores críticos iguais a $-1,96$ no limite inferior, resultado obtido com a fórmula $=\text{INV.NORMP}(0,025)$ e $+1,96$ no limite superior, resultado obtido com a fórmula $=\text{INV.NORMP}(0,975)$.
- O erro da estimativa 79,00 foi obtido com a fórmula $1,96 \times \frac{285}{\sqrt{50}} = 79$.
- O limite inferior do intervalo de aceitação da hipótese nula é \$2.421, resultado da diferença \$2.500-\$79, e o limite superior \$2.579=\$2.500+\$79.

Agora suponhamos que a hipótese nula H_0 seja falsa, que a média da renda mensal dos associados não seja igual a \$2.500. Nesse caso, a renda pode ser igual a qualquer valor diferente de \$2.500, valor que desconhecemos. Entretanto, sabemos que se H_0 for falsa, haverá muitos valores maiores ou menores do que \$2.500. Adotando como média verdadeira, por exemplo, o valor \$2.450, a probabilidade β de cometer um erro tipo II – ou a probabilidade de falhar na rejeição da hipótese nula falsa – será definida pela área da distribuição normal $N(2.450; 285)$ entre o limite inferior \$2.421 e o limite superior \$2.579, limites que definem a região de aceitação da hipótese nula da distribuição $N(2.500; 285)$. A probabilidade $P(2.421 \leq \mu \leq 2.579)$, considerando a distribuição normal $N(2.450; 285)$, é igual a 0,7634 ou 76,34%, resultado obtido com a fórmula:

$$=\text{DIST.NORM}(2579;2450;285/\text{RAIZ}(50);\text{VERDADEIRO}) - \text{DIST.NORM}(2421;2450;285/\text{RAIZ}(50);\text{VERDADEIRO})$$

Então, a probabilidade β de falhar na rejeição da H_0 falsa (ou de ocorrer um erro tipo II) é $\beta=76,34\%$, considerando que a média verdadeira seja $\mu=\$2.450$. Consequentemente, a probabilidade de rejeitar a hipótese nula quando ela é realmente falsa, será igual à probabilidade complementar de β medida como $1-\beta$, conhecida como *poder do teste* (veja a tabela da Figura 12.11). Nesse exemplo, a probabilidade de tomar a ação adequada de rejeitar H_0 , quando ela é realmente falsa, é 23,66%, resultado obtido com a fórmula $1-\beta=1-0,7634$.

A probabilidade β de cometer um erro tipo II depende de quatro fatores:

- Do valor do parâmetro definido na hipótese nula do teste.
- Do valor real do parâmetro.
- Do nível de significância α .
- Do tamanho n da amostra.

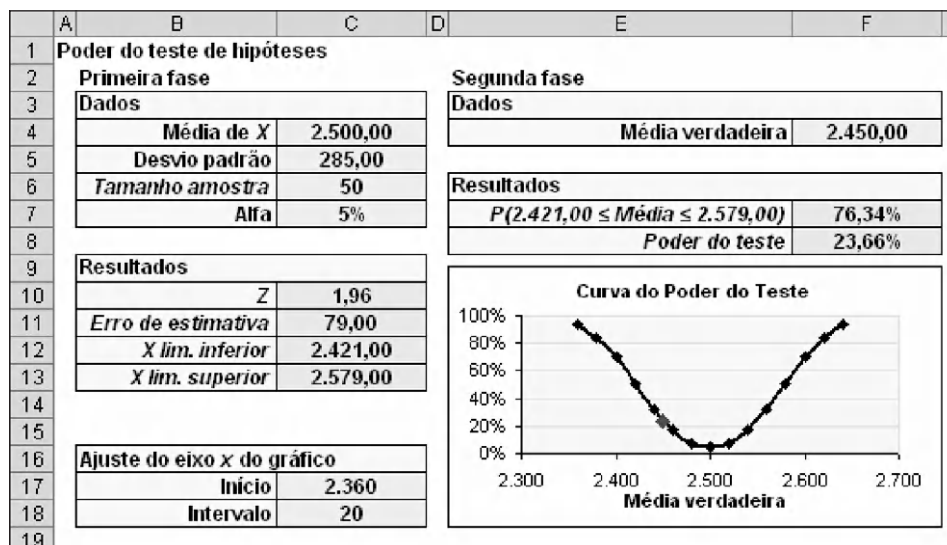
Tendo definido os valores do tamanho n da amostra e do nível de significância α , antes de realizar o teste de hipóteses, é possível obter valores da probabilidade β de cometer um erro tipo II em função de possíveis valores verdadeiros do parâmetro declarado na hipótese nula. O objetivo é conhecer quão bem o teste de hipóteses controla um erro tipo II, ou qual a probabilidade de rejeitar a hipótese nula se realmente for falsa. Essa informação é obtida da probabilidade complementar de β , ou $1-\beta$, denominada *poder*⁸ do teste contra um possível valor verdadeiro do parâmetro declarado na hipótese nula.

Para um determinado teste de hipóteses, é possível definir possíveis valores verdadeiros do parâmetro declarado na hipótese nula e, para cada um deles, calcular a probabilidade $1-\beta$, gerando a *função po-*

⁸ A probabilidade *poder* é correntemente utilizada com seu nome em inglês *power*.

der e seu correspondente gráfico da *curva do poder do teste*. Na planilha **Poder com Z**, incluída na pasta **Capítulo 12**, foi construído o modelo para análise nas duas caudas da distribuição, como mostra a Figura 12.12 com os dados e resultados do Exemplo 12.13.

FIGURA 12.12 Poder do teste de hipóteses.



Nesse modelo, você poderá ver como os fatores mencionados modificam o poder do teste.

Problemas

Os sete problemas seguintes devem ser resolvidos utilizando o procedimento de teste de hipóteses com o intervalo de confiança.

Problema 1

Afirma-se que a média da população é 125 e seu desvio padrão 36. Uma amostra aleatória retirada dessa população de tamanho $n=49$ tem média igual a 114. Considerando o intervalo de confiança de 95%, verifique se deve ser aceita a afirmação de que a média da população seja 125?

R: Rejeitar Hipótese Nula. No intervalo de confiança de 95%, ou nível de significância 0,05, a média 125 não está incluída no intervalo $114 \pm 10,08$.

Problema 2

Repetir o Problema 1, considerando o intervalo de confiança igual a 98%.

R: Aceitar Hipótese Nula. A média 125 está incluída no intervalo $114 \pm 11,96$.

Problema 3

A média de uma amostra aleatória de tamanho $n=38$ é 38,75 e o desvio padrão 3,2. Considerando o intervalo de confiança de 95%, podemos afirmar que a média da população seja igual a 37,5?

R: Rejeitar Hipótese Nula. A média da população 37,5 não está incluída no intervalo $38,75 \pm 1,017$.

Problema 4

Uma amostra aleatória de tamanho $n=18$ tem média 83 e desvio padrão 4,8. Considerando o intervalo de confiança de 95%, pode-se afirmar que a média da população seja igual a 80?

R: Rejeitar Hipótese Nula. A média da população 80 não está incluída no intervalo $83 \pm 2,387$.

Problema 5

Repita o Problema 4 considerando o intervalo de confiança de 99%.

R: *Aceitar Hipótese Nula*. A média da população 80 está incluída no intervalo $83 \pm 3,279$.

Problema 6

Repita o Problema 1, considerando que o tamanho da amostra é $n=22$.

R: *Aceitar Hipótese Nula*. Considerando que a distribuição da população seja normal, com a distribuição t concluímos que a média da população 125 está incluída no intervalo $114 \pm 15,96$.

Problema 7

Refaça o Exemplo 12.11 utilizando o procedimento de teste de hipóteses com o intervalo de confiança.

Os sete problemas seguintes devem ser resolvidos utilizando o procedimento de teste de hipóteses com Z ou t .

Problema 8

Refaça o Problema 1 com a distribuição adequada.

R: *Rejeitar Hipótese Nula*. No nível de significância 0,05, o Z crítico nas duas caudas é $-1,96$ e o Z observado $-2,1389$. Como $|Z_o| > |Z_c|$, deve-se aceitar a hipótese alternativa.

Problema 9

Repetir o Problema 8, considerando o nível de significância de 0,02.

R: *Aceitar Hipótese Nula*.

Problema 10

O vendedor afirma que a média dos negócios fechados diariamente é \$15.000. Para verificar a afirmação do vendedor, o gerente de vendas realizou uma amostragem aleatória das vendas de 45 dias obtendo média \$13.450 e desvio padrão \$3.500. A afirmação do vendedor deve ser aceita considerando o nível de significância 0,05?

R: *Rejeitar Hipótese Nula*. Como o valor absoluto do Z observado é $-2,97$ e o valor do Z crítico $-1,96$, há evidências de que a média dos negócios fechados diariamente pelo vendedor *não* seja igual a \$15.000.

Problema 11

Refaça o Problema 4 com a distribuição adequada.

R: *Rejeitar Hipótese Nula*. O t observado é $+2,6517$ e, para o nível de significância 0,05, o t crítico nas duas caudas é $+2,1098$. Como $|t_o| > |t_c|$, devemos aceitar a hipótese alternativa.

Problema 12

A montadora de carros afirma que a média de consumo de seu novo modelo de carro popular é 15,90 quilômetros por litro. Se uma amostra aleatória de 16 carros do novo modelo tem média de 14,8 com desvio padrão igual a 2, verificar a validade da afirmação realizada pela montadora de carros, considerando que a distribuição do consumo dos carros é normal e o nível de significância 0,05.

R: *Rejeitar Hipótese Nula*. O t observado é $-2,20$ e o t crítico $-2,1315$. Há evidências de que a afirmação da montadora não deve ser aceita.

Problema 13

A empresa especializada em investimentos afirma que no ano passado a média dos retornos reais das ações do segmento de comércio foi 15% ao ano. Uma amostragem aleatória de 31 ações do segmento de comércio apresentou média de 13,6% com desvio padrão 4,35%. A afirmação da empresa deve ser aceita, considerando o nível de significância 0,05?

R: *Aceitar Hipótese Nula*.

Problema 14

Refaça o Exemplo 12.11 utilizando o procedimento de teste de hipóteses com Z ou t .

Os problemas seguintes devem ser resolvidos utilizando o procedimento de teste de hipóteses com p -value.

Problema 15

Afirma-se que a média da população é 200. Uma amostra aleatória retirada dessa população de tamanho $n=36$ tem média de 208 e desvio padrão 35. Verificar se deve ser aceita a afirmação, considerando o nível de significância 0,05.

R: Aceitar Hipótese Nula. O Z observado é 1,3714 e o p -value = $2 \times P(Z \geq 1,37) = 0,1702$. Como 0,1702 é maior do que 0,05, a hipótese nula deve ser aceita: a média da população é 200, com um nível de significância 0,05.

Problema 16

Refaça o Problema 12 utilizando o p -value.

R: Rejeitar Hipótese Nula. O t observado é $-2,20$. O p -value é igual 0,0439, valor obtido de $2 \times P(t \geq 2,20)$. Como 0,0439 é menor do que 0,05, a hipótese nula deve ser rejeitada.

Problema 17

Refaça o Problema 13 utilizando o p -value.

R: Aceitar Hipótese Nula.

A maioria dos problemas seguintes não tem registrados os resultados. O leitor deve utilizar seus conhecimentos para obter as respostas.

Problema 18

O controle de qualidade da empresa de refrigerantes afirma que as novas linhas de enchimento de latas de refrigerantes conseguem produzir 12.000 latas por hora com média de 330 ml e desvio padrão de 5 ml e distribuição normal. Um dos distribuidores regionais informou que numa amostra de 40 latas obteve a média da amostra de 320 ml. Verifique essa reclamação considerando o teste de hipóteses adequado com intervalo de confiança de 95%.

R: O teste mostra que se deve rejeitar a hipótese nula.

Problema 19

Continuando com o Problema 18. Você acredita que a reclamação do distribuidor deve ser aceita? Por quê?

Problema 20

O fabricante de pneus assegura que a duração do pneu mais vendido tem média 60.000 km com desvio padrão 3.500 km. Como os distribuidores não estão convencidos, o fabricante ofereceu aos revendedores a oportunidade de separar, aleatoriamente, 36 pneus para verificar os resultados afirmados pelo fabricante. O teste realizado pelos revendedores apresentou a média de duração dos pneus igual a 59.500 km. Analisar esse resultado considerando o teste de hipóteses adequado com intervalo de confiança de 95%.

R: Aceitar a hipótese nula.

Problema 21

Considerando todos os clientes da agência, a média e o desvio padrão do saldo médio das contas correntes são, respectivamente, \$300 e \$100. O gerente da agência bancária desconfia que o saldo médio das contas correntes de sua agência diminuiu. Uma amostra aleatória de tamanho 60 mostrou saldo médio de \$285. Analisar esse resultado, considerando o teste de hipóteses adequado com intervalo de confiança de 95%.

R: Aceite a hipótese nula. Não há evidência significativa de que o saldo médio tenha diminuído.

Problema 22

Continuando com o Problema 21.

- Qual o saldo médio mínimo que não rejeita a hipótese nula?
- Qual o tamanho da amostra máximo que não rejeita a hipótese nula?
- Qual o nível de significância máximo que não rejeita a hipótese nula?

Capítulo 13

TESTES DE HIPÓTESES COM DUAS AMOSTRAS

O teste de hipóteses da diferença das médias de duas populações é frequentemente utilizado para determinar se é ou não razoável concluir que as médias das duas populações são diferentes. Por exemplo, é de interesse do controle de qualidade determinar se o mesmo produto fornecido por dois fornecedores diferentes apresenta a mesma quantidade de peças com defeitos. Ao médico do laboratório farmacêutico interessa determinar se o novo remédio para controle de diabetes é eficiente acompanhando dois grupos de pacientes, o primeiro grupo que recebeu o remédio e o outro que recebeu apenas *placebo*, produto com a mesma forma, porém sem o elemento ativo. O gerente de compras pode estar interessado em determinar se o mesmo produto fornecido por dois fornecedores diferentes apresenta o mesmo prazo real de entrega. Da mesma forma, o gerente de salários necessita conhecer se os salários da mesma categoria de trabalhadores têm o mesmo valor em duas cidades diferentes. Os exemplos mostram o objetivo do analista em determinar se há diferença entre as médias de duas populações independentes, lembrando que as respostas de um grupo são independentes das respostas do outro grupo.

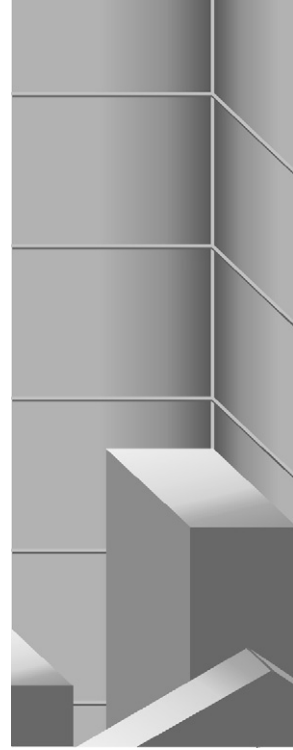
Teste de hipóteses para diferença entre médias

As premissas iniciais do teste de hipóteses para diferenças entre médias podem ser apresentadas da seguinte forma:

- Há duas populações independentes, denominadas X_1 e X_2 , com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2 , sendo que ambas as populações medem a mesma variável.
- Uma amostra aleatória é extraída de cada população. As duas amostras têm tamanhos n_1 e n_2 e médias \bar{X}_1 e \bar{X}_2 .
- A diferença das duas médias $\bar{X}_1 - \bar{X}_2$ é uma nova variável aleatória maior do que zero se $\bar{X}_1 > \bar{X}_2$, e menor do que zero se $\bar{X}_1 < \bar{X}_2$.

Na distribuição de frequências da diferença das médias $\bar{X}_1 - \bar{X}_2$:

- O valor esperado, ou média, de $\bar{X}_1 - \bar{X}_2$ é igual à diferença das médias das populações, $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$.



- A variância de $\bar{X}_1 - \bar{X}_2$ é igual a $\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2$, pois as variáveis são independentes. Utilizando as propriedades da distribuição amostral apresentadas no Capítulo 10, deduzimos:

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

As hipóteses do teste que deve ser aplicado têm a seguinte forma:

$$H_0: \mu_1 - \mu_2.$$

$$H_1: \mu_1 \neq \mu_2.$$

As hipóteses podem ser consideradas, também, como segue:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

À medida que forem apresentados os procedimentos de teste de hipóteses, serão adicionadas novas premissas.

Amostras grandes

Qual é a forma da distribuição da diferença de duas médias? A resposta é dada pelo teorema central do limite apresentado no Capítulo 10. Se for retirado um número grande de amostras das duas populações, a distribuição da diferença das duas médias será aproximadamente normal. Para amostras grandes, $n > 30$, o Z observado Z_o é obtido da normalização da diferença entre as duas médias utilizando a expressão:

$$Z_o = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sendo as variâncias das populações desconhecidas, as variâncias das amostras fornecerão uma boa aproximação, sendo o denominador da fórmula seguinte o erro amostral:

$$Z_o = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

A partir do Z observado e do nível de significância α adotado, pode-se aplicar um dos procedimentos de teste de hipóteses do Capítulo 12, como mostrado no Exemplo 13.1 utilizando o *p-value*.

EXEMPLO 13.1

As variâncias das populações 1 e 2 são, respectivamente, 15 e 30. As amostras independentes 1 e 2 foram retiradas das populações 1 e 2, e seus valores estão registrados nas colunas B e C da planilha **Exemplo 13.1**, incluída na pasta **Capítulo 13**. Realize o teste de hipóteses da diferença das médias considerando o nível de significância $\alpha=0,05$.

Solução. Os dados fornecidos mostram que a hipótese nula deve ser aceita.

	A	B	C	D	E	F	G
1	Comparação de duas médias						
2							
3		Amostra 1	Amostra 2		Dados	Amostra 1	Amostra 2
4		103	117		Variação	15,00	30,00
5		107	115				
6		114	104		Resultados	Amostra 1	Amostra 2
7		109	112		n	40	60
8		114	112		Média	107,05	108,42
9		102	108				
10		101	105		Teste de Hipóteses		
11		108	108		Alfa	5,0%	
12		104	112		Z observado	-1,46	
13		107	100		p-value	14,40%	
14		107	108		Decisão	Aceitar Ho	
15		107	101				

Analise os dados e resultados apresentados na planilha.

- O *modelo* registra as medidas estatísticas tamanho e média de cada amostra, intervalo F7:G8.
- Na célula F11, deve-se registrar o nível de significância *Alfa*.
- Na célula F12, o modelo registra o Z observado, resultado obtido com:

$$Z_o = \frac{107,05 - 108,42 - 0}{\sqrt{\frac{15}{40} + \frac{30}{60}}} = -1,46$$

- Na célula F13, é calculado o *p-value* para duas caudas da distribuição. Como o *p-value* é maior do que o nível de significância $\alpha=0,05$, a hipótese nula deve ser aceita, pois há evidências de que a diferença de médias não seja significativa. Na célula F14, é apresentada a decisão por extenso, *Aceitar Ho* ou *Rejeitar Ho*.

Este procedimento com a distribuição Z deve ser aplicado quando as variâncias das populações são conhecidas, o que, na prática, é difícil de ocorrer. Daí que se o tamanho de uma das amostras for igual ou menor do que 31, o *modelo* não apresentará os títulos e resultados relevantes. Como em geral as variâncias das populações não são conhecidas, é recomendado utilizar o procedimento com a distribuição *t*.

Ferramenta de análise teste Z: duas amostras para médias

A ferramenta de análise *Teste Z: duas amostras para médias*¹ realiza uma análise estatística e o teste de hipóteses da diferença das médias de duas populações independentes. A Figura 13.2 mostra essa ferramenta aplicada no Exemplo 13.1, na planilha **Exemplo 13.1**, a partir da célula I2.

Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro. Escolhendo a ferramenta **Teste Z: duas amostras para médias** e depois de clicar no botão **OK**, você receberá a caixa de diálogo com o mesmo nome, mostrada na Figura 13.1, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Teste Z: duas amostras para médias* pertencente à *Ajuda do Excel*.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo dessa ferramenta são:

- **Intervalo da variável 1:** Informe o intervalo de células da planilha no qual os dados da Amostra 1 estão registrados; nesse caso, o intervalo B3:B43, que inclui a célula onde foi registrado o título *Amostra 1*, ou rótulo no Excel.
- **Intervalo da variável 2:** Informe o intervalo de células da planilha no qual os dados da Amostra 2 estão registrados; nesse caso, o intervalo C3:C63, que inclui a célula onde foi registrado o título *Amostra 2*.

¹ Em inglês, a ferramenta *Teste-Z: duas amostras para média* é *Z-Test: two-sample for means*.

- **Hipótese da diferença da média:** Insira o número que se deseja para a mudança nas médias das amostras. O valor zero indica que as médias das amostras são hipoteticamente iguais. Neste exemplo, foi informado o valor zero, pois a hipótese nula é $\mu_1 - \mu_2 = 0$.
- **Variância da variável 1 (conhecida):** Informe o valor requerido, nesse caso 15.
- **Variância da variável 2 (conhecida):** Informe o valor requerido, nesse caso 30.
- **Rótulos:** Selecione esta caixa, pois na primeira célula de cada intervalo da variável foi incluído o nome da amostra.
- **Alfa:** Informe o nível de significância *alfa* do teste de hipóteses, nesse caso 0,05. A ferramenta de análise não requer que seja estabelecido se o teste deve ser realizado em uma cauda ou nas duas caudas da distribuição, pois a ferramenta de análise apresentará os dois resultados para o mesmo *alfa*.

FIGURA 13.1 Caixa de diálogo *Teste-Z: Duas amostras para média*.

No quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três botões de opção que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso I2, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes. Mais informações podem ser obtidas no Capítulo 4 ou na Ajuda do Excel.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha **Plan1**.

Depois de completar as informações e clicar em OK na caixa de diálogo, o Excel apresentará a tabela com os resultados, Figura 13.2. Comparando os resultados obtidos com a ferramenta e o *modelo* da mesma planilha, temos:

- Na célula J9, é registrado o *Z observado* igual a -1,4610.
- Na célula J12, é registrado o *p-value* igual a 0,1440 para duas caudas da distribuição.
- Como o *p-value* é maior do que o nível de significância 0,05, a hipótese nula deve ser aceita, pois a diferença de médias não é significativa.

	H	I	J	K
1	Ferramenta de análise			
2	Teste-z: duas amostras para médias			
3				
4			<i>Amostra 1</i>	<i>Amostra 2</i>
5	Média		107,05	108,416667
6	Variância conhecida		15	30
7	Observações		40	60
8	Hipótese da diferença de média		0	
9	z		-1,46102812	
10	P(Z≤z) uni-caudal		0,0720039	
11	z crítico uni-caudal		1,64485348	
12	P(Z≤z) bi-caudal		0,1440078	
13	z crítico bi-caudal		1,95996279	
14				

FIGURA 13.2

Ferramenta *Teste-Z: duas amostras para média*, Exemplo 13.1.

Problemas

Problema 1

As amostras 1 e 2 foram retiradas das populações independentes 1 e 2. Os resultados estatísticos das amostras estão registrados na tabela seguinte. Verifique se há evidência de que as médias das populações sejam diferentes, considerando o nível de significância $\alpha=5\%$ e utilizando o *modelo* da planilha *Modelo Z*, incluído na pasta *Capítulo 13*.

	Amostra 1	Amostra 2
<i>n</i>	50	40
Média	1.120	1.075
Variância	6.400	11.025

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(Z \geq Z_0)=0,0251$.

Problema 2

As amostras 3 e 4 foram retiradas das populações independentes 3 e 4. Os resultados estatísticos das amostras estão registrados na tabela seguinte. Verifique se há evidência de que as médias das populações sejam diferentes, considerando o nível de significância $\alpha=5\%$.

	Amostra 3	Amostra 4
<i>n</i>	36	55
Média	325	312
Variância	1.024	1.225

R: Aceitar a hipótese nula, pois $p\text{-value}=2 \times P(Z \geq Z_0)=0,0679$. A planilha *Modelo Z*, incluída na pasta *Capítulo 13*, apresenta os seguintes resultados:

	A	B	C
7			
8	Teste de Hipóteses		
9		<i>Alfa</i>	5,00%
10		<i>Z observado</i>	1,8254
11		<i>p-value</i>	6,79%
12		<i>Decisão</i>	Aceitar H_0
13			

Problema 3

A cooperativa de pequenos supermercados tem dois grandes distribuidores que fornecem a maioria dos produtos, *Fornecedor 1* e *Fornecedor 2*. Na última reunião de associados, foi levantada a possibilidade de mudar um dos fornecedores, pois aparentemente seu prazo de entrega é maior do que o do outro. Os registros mantidos na cooperativa mostram que os desvios padrão da população da variável prazo de entrega em horas do *Fornecedor 1* e *Fornecedor 2* são, respectivamente, 5 e 4 horas. Verifique se há evidência da reclamação dos associados da cooperativa, considerando os dados registrados na planilha **Problema 3**.

R: Rejeitar a hipótese nula, pois $p\text{-value}=0,65\%$. Esse problema está resolvido na planilha **Problema 3**, incluída na pasta **Capítulo 13**, mostrando que o prazo de entrega do *Fornecedor 2* realmente é maior do que o do outro fornecedor.

Amostras pequenas e variâncias das populações iguais

Se o tamanho das amostras for pequeno, às premissas do teste de hipóteses da diferença das médias de duas populações independentes deverão ser adicionadas as seguintes premissas:

- As populações devem ter distribuição normal.
- As variâncias das populações são presumivelmente iguais, $\sigma_1^2 = \sigma_2^2$.
- Deve-se utilizar a distribuição t com $(n_1 + n_2 - 2)$ graus de liberdade.
- Sendo as variâncias das populações desconhecidas, a variância da distribuição da diferença das duas médias S_p^2 , denominada variância agrupada,² será obtida com a fórmula seguinte, onde S_1^2 e S_2^2 são as variâncias das duas amostras:

$$S_p^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}$$

O t observado t_o é obtido com a expressão:

$$t_o = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Se as duas amostras tiverem o mesmo tamanho, $n_1 = n_2$, a variância agrupada S_p^2 será igual

$$S_p^2 = \frac{S_1^2 + S_2^2}{2}.$$

EXEMPLO 13.2

As variâncias das populações 1 e 2 são presumivelmente iguais. As amostras independentes 1 e 2 foram retiradas das populações 1 e 2, e seus valores estão registrados nas colunas B e C da planilha **Exemplo 13.2**, incluída na pasta **Capítulo 13**. Realize o teste de hipóteses da diferença das duas médias, considerando o nível de significância $\alpha=0,05$.

Solução. Os dados fornecidos mostram que a hipótese nula deve ser rejeitada.

² Em inglês, conhecida como *pooled estimate of variance*.

	A	B	C	D	E	F	G
1	Comparação de duas médias - Populações com variâncias iguais						
2							
3		Amostra 1	Amostra 2		Resultados	Amostra 1	Amostra 2
4		20	25		n	5	6
5		23	26		Média	23,00	25,33
6		26	25		Variância	5,00	1,07
7		22	24				
8		24	27				
9			25				
10							
11							
12							
13							
14							
15							

Analise os dados e os resultados apresentados na planilha.

- O *modelo* registra as medidas estatísticas tamanho, média e variância de cada amostra, intervalo F5:G7.
- Na célula F9, deve-se registrar o nível de significância *Alfa*.
- Na célula F10, o modelo calcula os graus de liberdade *gl* aplicando a fórmula $n_1 + n_2 - 2 = 9$. E na célula F11, o modelo calcula a variância agrupada.
- Com os resultados anteriores, na célula F12 é registrado o *t observado*, calculado com a fórmula já apresentada.
- Na célula F13, é calculado o *p-value* para duas caudas da distribuição igual a 4,72%. Como o *p-value* é menor do que o nível de significância $\alpha=0,05$, deve-se rejeitar a hipótese nula, pois a diferença das médias das populações é significativa. Na célula F13, é apresentada a decisão por extenso, *Aceitar Ho* ou *Rejeitar Ho*.

Em vez de utilizar o modelo descrito anteriormente, com os valores das amostras registrados na planilha, pode-se obter o *p-value* utilizando a função estatística TESTET.

• TESTET(matriz1; matriz2; caudas; tipo)

A função estatística TESTET³ retorna o *p-value* do teste de hipóteses da diferença de duas médias a partir dos valores das amostras registrados nos argumentos *matriz1* e *matriz2*. No cálculo do *p-value*, são considerados as caudas da distribuição *t* e o tipo de relacionamento das populações.

- Se o argumento *caudas*=1, a função TESTET retornará o *p-value* numa cauda da distribuição *t*, e se *caudas*=2, retornará o *p-value* nas duas caudas da distribuição *t*.
- Se o argumento *tipo*=1, a função TESTET retornará o *p-value*, considerando que as duas populações são dependentes; se *tipo*=2, retornará o *p-value*, considerando que as duas populações são independentes e têm a mesma variância; por último, se *tipo*=3, retornará o *p-value*, considerando que as duas populações são independentes e têm variâncias diferentes.

Na célula C19 da planilha **Exemplo 13.2**, foi utilizada a função TESTET registrando a fórmula =TESTET(B5:B9;C5:C10;2;2), que retornou o *p-value* igual a 0,0472 ou 4,72%. Verifique que o terceiro argumento da função TESTET, *caudas*=2, requer o teste de hipóteses nas duas caudas, e o argumento *tipo*=2 retornará o *p-value*, considerando que as duas populações são independentes e têm a mesma variância. O resultado da função mostra que a hipótese nula deve ser rejeitada. Ao utilizar a função TESTET, somente será necessário registrar os dados das duas amostras, eliminando a necessidade de realizar os cálculos auxiliares para obter o *p-value*. Para tomar a decisão de aceitar ou rejeitar a hipótese nula, o analista deverá comparar o *p-value* retornado pela função TESTET com o nível de significância adotado. Comparando com o modelo, este adiciona o suporte para tomar a decisão de aceitar ou rejeitar a hipótese nula. Na função TESTET, os valores das duas variáveis podem ser também informados como matrizes, por exemplo, na célula C20 da mesma planilha foi registrada a fórmula =TESTET({20;23;26;22;24};{25;26;25;24;27;25};2;2), retornando o mesmo resultado. Observe que registrando os valores das amostras como matrizes todos os dados e o resultado foram incluídos numa única célula da planilha Excel. Para tomar a decisão de aceitar ou rejeitar a hipótese nula, o analista deverá comparar o *p-value* retornado pela função TESTET com o nível de significância adotado.

3 Em inglês, a função TESTET é TTEST.

Ferramenta de análise teste-T: duas amostras presumindo variâncias equivalentes

A ferramenta de análise *Teste-t: duas amostras presumindo variâncias equivalentes*⁴ realiza análises estatísticas e teste de hipóteses da diferença das médias de duas populações independentes com variâncias iguais. A Figura 13.4 mostra essa ferramenta aplicada no Exemplo 13.2, na planilha **Exemplo 13.2**, a partir da célula I2.

Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como foi mostrado na Figura 1.7 do Capítulo 1 do livro. Escolhendo a ferramenta **Teste-t: duas amostras presumindo variâncias equivalentes** e depois clicando no botão **OK**, você receberá a caixa de diálogo com o mesmo nome mostrado na Figura 13.3, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel exibirá a página *Sobre a caixa de diálogo Teste-t: duas amostras presumindo variâncias equivalentes* pertencente à *Ajuda do Excel*.

FIGURA 13.3

Teste-t: duas amostras presumindo variâncias equivalentes.

A caixa de diálogo "Teste-T: duas amostras presumindo variâncias equivalentes" apresenta os seguintes campos e opções:

- Entrada:**
 - Intervalo da variável 1: \$B\$4:\$B\$9
 - Intervalo da variável 2: \$C\$4:\$C\$10
 - Hipótese da diferença de média: 0
 - ☒ Rótulos
 - Alfa: 0,05
- Opções de saída:**
 - ☒ Intervalo de saída: \$I\$2
 - ☐ Nova planilha:
 - ☐ Nova pasta de trabalho

Botões: OK, Cancelar, Ajuda.

As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo dessa ferramenta são:

- **Intervalo da variável 1:** Informe o intervalo de células da planilha no qual os dados da Amostra 1 estão registrados, nesse caso, o intervalo B4:B9, que inclui a célula onde foi registrado o título *Amostra 1*, ou rótulo no Excel.
- **Intervalo da variável 2:** Informe o intervalo de células da planilha no qual os dados da Amostra 2 estão registrados, nesse caso, o intervalo C4:C10, que inclui a célula onde foi registrado o título *Amostra 2*.
- **Hipótese da diferença da média:** Insira o número que se deseja para a mudança nas médias das amostras. O valor zero indica que as médias das amostras são hipoteticamente iguais. Neste exemplo, informamos o valor zero, pois a hipótese nula é $\mu_1 - \mu_2 = 0$.
- **Rótulos:** deve-se selecionar, pois na primeira célula de cada amostra foi incluído o nome dessa amostra.
- **Alfa:** deve-se informar o nível de significância *alfa* do teste de hipóteses, nesse caso, 0,05. A ferramenta de análise não requer que seja estabelecido se o teste deve ser realizado numa cauda ou nas duas caudas da distribuição, pois a ferramenta de análise apresentará os dois resultados para o mesmo *alfa*.

Na primeira parte do quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados.

⁴ Em inglês, a ferramenta *Teste-t: duas amostras presumindo variâncias equivalentes* é *t-Test: two-sample assuming equal variances*.

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso I2, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta. Mais informações podem ser obtidas no Capítulo 4 ou na Ajuda do Excel.
- **Nova pasta de trabalho.** Os resultados serão apresentados numa nova pasta e a partir da célula A1 da planilha Plan1.

	H	I	J	K
1		Ferramenta de análise		
2		Teste-t: duas amostras presumindo variâncias equivalentes		
3				
4			<i>Amostra 1</i>	<i>Amostra 2</i>
5		Média	23	25,3333333
6		Variância	5	1,06666667
7		Observações	5	6
8		Variância agrupada	2,81481481	
9		Hipótese da diferença de média	0	
10		gl	9	
11		Stat t	-2,29676286	
12		P(T<=t) uni-caudal	0,02362449	
13		t crítico uni-caudal	1,83311386	
14		P(T<=t) bi-caudal	0,04724899	
15		t crítico bi-caudal	2,26215889	
16				

FIGURA 13.4

Teste-t: duas amostras presumindo variâncias equivalentes.

Depois de completar as informações e clicar em OK na caixa de diálogo, o Excel apresenta a tabela com os resultados, Figura 13.4. Comparando os dados obtidos com a ferramenta e o modelo da mesma planilha, temos:

- Na célula J8, é registrado o valor da variância agrupada igual a 2,815.
- Na célula J11, é registrado o *t observado* igual a -2,2968.
- Na célula J14, é registrado o *p-value* igual a 0,0472 para duas caudas da distribuição.
- Como o *p-value* é menor do que o nível de significância 0,05, a hipótese nula não deve ser aceita, pois a diferença de médias é significativa.

Problemas

Problema 4

Os resultados estatísticos da *Amostra 1* e da *Amostra 2* estão registrados na tabela seguinte. As duas amostras foram retiradas de duas populações independentes que têm a mesma variância. Realize o teste de hipóteses da diferença das duas médias considerando o nível de significância $\alpha=5\%$.

	Amostra 1	Amostra 2
<i>n</i>	10	15
Média	125	138
Variância	144	225

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=3,15\%$. A planilha **Modelo t**, incluída na pasta **Capítulo 13**, apresenta os seguintes resultados.

	A	B	C	D
1	Modelo t - Comparação de duas médias			
2				
3	Populações com variâncias			
4	<input checked="" type="radio"/> Iguais <input type="radio"/> Diferentes			
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

Problema 5

Repita o Problema 4 com os resultados estatísticos da *Amostra 3* e da *Amostra 4* registradas na tabela seguinte.

	Amostra 3	Amostra 4
<i>n</i>	19	12
Média	18,5	14,8
Variância	28	35

R: Aceitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=0,080$.

Problema 6

Para verificar os resultados do Problema 3, foram realizadas duas amostras mais recentes cujos resultados estão registrados na tabela seguinte. Realize o teste de hipóteses da diferença das duas médias, considerando que as duas populações têm variâncias iguais e o nível de significância $\alpha=5\%$.

	Fornecedor 1	Fornecedor 2
<i>n</i>	16	27
Média	37,1	40,9
Variância	30,3	35,6

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=4,42\%$.

Amostras pequenas e variâncias das populações diferentes

O procedimento do teste de hipóteses da diferença das médias de duas populações com variâncias desconhecidas, ou presumindo que sejam diferentes, tem as mesmas premissas do procedimento do teste de hipóteses com variâncias iguais, incluindo as seguintes alterações de cálculo:

- Deve ser utilizada a estatística teste t^* definida com a expressão:⁵

⁵ Em inglês, conhecida como *separate-variance t^* test*.

$$t^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- O teste t^* pode ser aproximado ao teste t obtendo o número de graus de liberdade gl com a expressão:

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 - 1}}$$

Como em geral o resultado de gl não é um número inteiro, deve ser adotado o número inteiro mais próximo.

EXEMPLO 13.3

As amostras 1 e 2 foram retiradas das populações independentes 1 e 2 com variâncias diferentes e seus valores estão registrados nas colunas B e C da planilha **Exemplo 13.3**, incluída na pasta **Capítulo 13**. Realize o teste de hipóteses da diferença das duas médias considerando o nível de significância $\alpha=5\%$.

Solução. Os dados e resultados mostram que a hipótese nula deve ser aceita.

	A	B	C	D	E	F	G
1	Comparação de duas médias - Populações com variâncias diferentes						
2							
3		Amostra 1	Amostra 2		Resultados	Amostra 1	Amostra 2
4		10	13		<i>n</i>	6	5
5		12	12		Média	11,83	13,60
6		14	15		Variância	2,57	1,80
7		13	13				
8		12	15				
9		10			Teste de Hipóteses		
10					<i>Alfa</i>	5,00%	
11					<i>gl</i>	9	
12					<i>t* observado</i>	-1,990	
13					<i>p-value</i>	7,77%	
14					<i>Decisão</i>	Aceitar <i>H</i> ₀	

Analisemos os dados e os resultados apresentados na planilha.

- O *modelo* registra as medidas estatísticas, tamanho, média e variância de cada amostra, intervalo F4:G6.
- Na célula F9, deve-se registrar o nível de significância *Alfa*.
- Na célula F10, o modelo calcula os graus de liberdade *gl* aplicando a expressão apresentada. O arredondamento é realizado com a função matemática do Excel ARRED, com o argumento *núm_dígitos* igual a zero.
- Com os resultados anteriores, na célula F11 é calculado o valor do *t* observado* aplicando a expressão apresentada.
- Na célula F12, é calculado o *p-value* 7,77% para duas caudas da distribuição.
- Como o *p-value* é maior do que o nível de significância de 5%, deve-se aceitar a hipótese nula, pois a diferença das médias das populações não é significativa. Na célula F13, é apresentada a decisão por extenso, *Aceitar H*₀ ou *Rejeitar H*₀.

Em vez de utilizar o modelo anterior, a partir dos valores das amostras pode-se obter o *p-value* utilizando a função estatística TESTET. Na célula C19 da planilha **Exemplo 13.3**, foi utilizada a função TESTET registrando

a fórmula $\text{=TESTET}(B4:B9;C4:C8;2;3)$, que retornou o p -value igual a 7,77%. Observe que o terceiro argumento da função TESTET, *caudas*=2, requer o teste de hipóteses nas duas caudas, e o argumento *tipo*=3 retornará o p -value considerando que as duas populações são independentes e têm variâncias diferentes. O resultado da função mostra que a hipótese nula deve ser aceita. Ao utilizar a função TESTET, somente será necessário registrar os dados das duas amostras, eliminando a necessidade de realizar os cálculos auxiliares para obter o p -value. Para tomar a decisão de aceitar ou rejeitar a hipótese nula, o analista deverá comparar o p -value retornado pela função TESTET com o nível de significância adotado. Comparando com o modelo, este adiciona o suporte para tomar a decisão de aceitar ou rejeitar a hipótese nula.

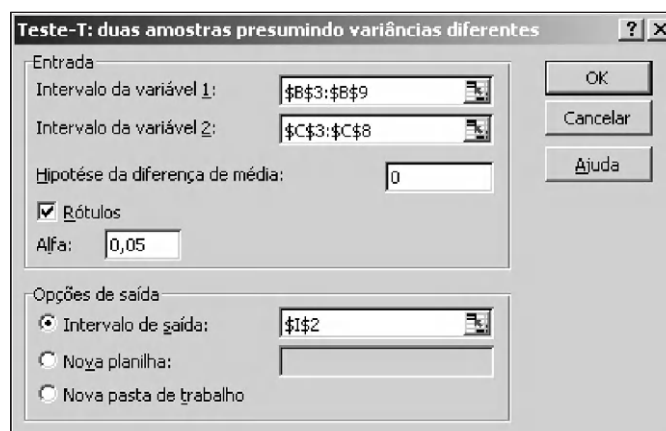
Na função TESTET, os valores das duas variáveis podem ser também informados como matrizes, por exemplo, na célula C20 da mesma planilha, foi registrada a fórmula: $\text{=TESTET}(\{10;12;14;13;12;10\};\{13;12;15;13;15\};2;3)$, retornando o mesmo resultado. Note que registrando os valores das amostras como matrizes, todos os dados e o resultado foram incluídos em uma única célula da planilha Excel. Para tomar a decisão de aceitar ou rejeitar a hipótese nula, o analista deverá comparar o p -value retornado pela função TESTET com o nível de significância adotado.

Ferramenta de análise teste-T: duas amostras presumindo variâncias diferentes

A ferramenta de análise *Teste-t: duas amostras presumindo variâncias diferentes*⁶ realiza análises estatísticas e teste de hipóteses da diferença das médias de duas populações independentes com variâncias diferentes. A Figura 13.6 mostra essa ferramenta aplicada no Exemplo 13.3, na planilha **Exemplo 13.3**, a partir da célula I2.

FIGURA 13.5

Teste-t: duas amostras presumindo variâncias diferentes.



Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro. Escolhendo a ferramenta **Teste-t: duas amostras presumindo variâncias diferentes** e depois clicando no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostrado na Figura 13.5, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Teste-t: duas amostras presumindo variâncias diferentes* pertencente à *Ajuda do Excel*.

Como o procedimento de trabalho dessa ferramenta é o mesmo das ferramentas anteriores, somente serão mostrados alguns detalhes. Depois de completar as informações na caixa de diálogo, clicando no botão **OK**, o Excel exibirá a tabela com os resultados, Figura 13.6.

- Na célula J10, é registrado o t^* observado igual a -1,990.

⁶ Em inglês, a ferramenta *Teste-t: duas amostras presumindo variâncias diferentes* é *t-Test: two-sample assuming unequal variances*.

- Na célula J13, é registrado o *p-value* igual a 0,077739 para duas caudas da distribuição.
- Como o *p-value* é maior que o nível de significância 0,05, a hipótese nula deve ser aceita, pois a diferença de médias não é significativa.

Compare esses resultados com os obtidos no *modelo* construído na mesma planilha Exemplo 13.4.

	H	I	J	K
1		Ferramenta de análise		
2		Teste-t: duas amostras presumindo variâncias diferentes		
3				
4			Amostra 1	Amostra 2
5		Média	11,8333333	13,6
6		Variância	2,56666667	1,8
7		Observações	6	5
8		Hipótese da diferença de média	0	
9		gl	9	
10		Stat t	-1,99045678	
11		P(T<=t) uni-caudal	0,03886968	
12		t crítico uni-caudal	1,83311388	
13		P(T<=t) bi-caudal	0,07773936	
14		t crítico bi-caudal	2,26215889	
15				

FIGURA 13.6

Teste-t: duas amostras presumindo variâncias diferentes.

Problemas

Problema 7

A tabela seguinte registra os resultados estatísticos de duas amostras retiradas de duas populações independentes com variâncias diferentes. Realize o teste de hipóteses da diferença das duas médias considerando $\alpha=5\%$ e aplicando o *modelo* da planilha Modelo t incluído na pasta Capítulo 13.

	Amostra 1	Amostra 2
n	15	24
Média	36,50	38,10
Variância	3,80	5,00

R: Rejeitar a hipótese nula, pois *p-value*=2,46%. A planilha Modelo t, incluída na pasta Capítulo 13, apresenta os seguintes resultados.

	A	B	C	D
1		Modelo t - Comparação de duas médias		
2				
3		Populações com variâncias		
4		<input type="radio"/> Iguais	<input checked="" type="radio"/> Diferentes	
5				
6				
7		Dados	Amostra 1	Amostra 2
8		n	15	24
9		Média	36,50	38,10
10		Variância	3,80	5,00
11				
12		Teste de Hipóteses		
13		Alfa	5,00%	
14		gl	33	
15		t observado	-2,355	
16		p-value	2,46%	
17		Decisão	Rejeitar Ho	
18				

Problema 8

Repita o Problema 7, considerando as amostras da tabela seguinte.

	Amostra A	Amostra B
<i>n</i>	13	10
Média	108,5	101,4
Variância	113,6	125,9

R: Aceitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_o)=0,141$.

Problema 9

Repita o Problema 6, considerando que as duas populações têm variâncias diferentes e o nível de significância $\alpha=5\%$.

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_o)=4,14\%$.

Amostras emparelhadas

Quando for necessário comparar, por exemplo, as vendas diárias de duas filiais que operam com os mesmos produtos, ou os resultados de um treinamento, confrontando o conhecimento antes e depois do treinamento, os procedimentos de teste de hipóteses para diferença das médias utilizados até este momento não podem ser aplicados, pois se referem a duas populações independentes. Agora, necessitamos analisar duas populações relacionadas, isto é, duas populações dependentes. Nesse caso, a variável de interesse será a diferença entre os pares das duas amostras, no lugar das próprias amostras, que devem ter o mesmo tamanho.

Como premissa, a população das diferenças tem distribuição aproximadamente normal, e a amostra das diferenças é extraída aleatoriamente da população das diferenças. O procedimento é o seguinte:

- Das duas variáveis X_1 e X_2 definidas pelos valores $X_{11}, X_{12}, \dots, X_{1n}$ e $X_{21}, X_{22}, \dots, X_{2n}$, é formada a nova variável D das diferenças entre esses valores $D_1 = X_{11} - X_{21}, \dots, D_j = X_{1j} - X_{2j}, \dots, D_n = X_{1n} - X_{2n}$.
- Na variável D , é calculada a média \bar{D} e a variância S_D^2 .
- O t observado é calculado⁷ com a fórmula $t_o = \frac{\bar{D} - 0}{\frac{S_D}{\sqrt{n}}}$.

Definido o nível de significância α , é realizado o teste de hipóteses.

$$H_o: \mu_D = 0$$

$$H_1: \mu_D \neq 0$$

EXEMPLO 13.4

As amostras 1 e 2 foram retiradas das populações relacionadas 1 e 2, e seus valores estão registrados nas colunas B e C da planilha **Exemplo 13.4**, incluída na pasta **Capítulo 13**. Verifique se há diferença entre as médias dessas populações relacionadas, considerando o nível de significância $\alpha=5\%$.

Solução. Os dados e os resultados mostram que a hipótese nula deve ser rejeitada.

⁷ Se $n > 30$, é possível aplicar a distribuição Z. Entretanto, pode-se manter a distribuição t lembrando que, para $n > 30$, a distribuição t se aproxima da distribuição Z.

	A	B	C	D	E	F	G
1	Comparação de duas médias - Amostras Emparelhadas						
2							
3		Amostra 1	Amostra 2	D		Resultados	D
4		12	14	-2		n	9
5		10	10	0		Média	-1,44
6		14	18	-2		Variância	2,28
7		8	10	-2			
8		12	12	0		Teste de Hipóteses	
9		15	18	-3		α	5,00%
10		15	15	0		gl	8
11		12	12	0		t observado	-2,871
12		9	13	-4		p-value	2,08%
13						Decisão	Rejeitar H_0
14							

Analisemos os dados e os resultados apresentados na planilha.

- No intervalo D4:D12, foi construída a amostra *D* cujos valores foram obtidos como diferenças dos valores das amostras.
- O *modelo* registra as medidas estatísticas, tamanho, média e variância da amostra *D*, intervalo G4:G6.
- Na célula G9, deve-se registrar o nível de significância *Alfa*.
- Na célula G10, o *modelo* calcula os graus de liberdade com a expressão $gl=n-1$.
- Na célula G11, é calculado o *t observado*.
- Na célula G12, é calculado o *p-value*=2,08% para duas caudas da distribuição. Como o *p-value* é menor do que o nível de significância $\alpha=5\%$, a hipótese nula deve ser rejeitada, pois a diferença de médias é significativa.

Na célula C19 da planilha Exemplo 13.2, foi utilizada a função TESTET registrando a fórmula =TESTET(B4:B12;C4:C12;2;1), que retornou o *p-value* igual a 0,0208 ou 2,08%. Observe que o terceiro argumento da função TESTET, *caudas*=2, requer o teste de hipóteses nas duas caudas, e o argumento *tipo*=1 retornará o *p-value* considerando que as duas populações são dependentes. Na função TESTET, os valores das duas variáveis podem ser também informados como matrizes, por exemplo, na célula C20 foi registrada a fórmula =TESTET({12;10;14;8;12;15;15;12;9};{14;10;16;10;12;18;15;12;13};2;1) retornando o mesmo resultado. Note que registrando os valores das amostras como matrizes todos os dados e o resultado foram incluídos em uma única célula da planilha Excel. Para tomar a decisão de aceitar ou rejeitar a hipótese nula, o analista deverá comparar o *p-value* retornado pela função TESTET com o nível de significância adotado.

Ferramenta de análise teste-T: duas amostras em par para médias

A ferramenta de análise *Teste-t: duas amostras em par para médias*⁸ realiza análises estatísticas e teste de hipóteses de duas médias informadas como séries de valores. A Figura 13.8 mostra essa ferramenta aplicada no Exemplo 13.4, na planilha Exemplo 13.4, a partir da célula I2.

Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro. Escolhendo a ferramenta **Teste-t: duas amostras em par para médias** e depois clicando no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostrado na Figura 13.7, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Teste-t: duas amostras em par para médias* pertencente à *Ajuda do Excel*.

⁸ Em inglês, a ferramenta *Teste-t: duas amostras em par para médias* é *t-Test: two-sample for means*.

Como o procedimento de trabalho dessa ferramenta é o mesmo das ferramentas anteriores, somente serão mostrados alguns detalhes. Depois de completar as informações na caixa de diálogo, clicando no botão OK, o Excel apresenta a tabela com os resultados, Figura 13.8.

- Na célula J11, é registrado o t observado igual a $-2,871$.
- Na célula J14, é registrado o p -value igual a $0,0207899$ para duas caudas da distribuição. Como o p -value é menor do que o nível de significância $0,05$, a hipótese nula não deverá ser aceita, pois a diferença de médias é significativa.

FIGURA 13.7 Caixa de diálogo *Teste-t: duas amostras em par para médias*.

FIGURA 13.8 *Teste-t: duas amostras em par para médias*.

	H	I	J	K
1		Ferramenta de análise		
2		Teste-t: duas amostras em par para médias		
3				
4			<i>Amostra 1</i>	<i>Amostra 2</i>
5		Média	11,888889	13,333333
6		Variância	6,361111	7,25
7		Observações	9	9
8		Correlação de Pearson	0,8344343	
9		Hipótese da diferença de média	0	
10		gl	8	
11		Stat t	-2,8712197	
12		P(T<=t) uni-caudal	0,010395	
13		t crítico uni-caudal	1,8595483	
14		P(T<=t) bi-caudal	0,0207899	
15		t crítico bi-caudal	2,3060056	
16				

Problemas

Problema 10

O gerente da oficina de carros afirma que seu procedimento de regulagem dos motores consegue reduzir o consumo de combustível sem diminuir a potência do motor. Sua afirmação está baseada no resultado do questionário, um número entre 10 e 15, que os donos dos carros preenchem antes e depois da regulagem do motor, como mostra a tabela seguinte.

Carro	1	2	3	4	5	6	7	8
Antes	10	12	13	11	14	12	10	9
Depois	15	13	14	10	13	13	14	13

Verifique se o procedimento de regulagem foi eficiente, considerando o nível de significância $\alpha=5\%$.

R: Aceitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=8,76$. Esse resultado pode ser obtido com a fórmula a seguir registrada numa célula vazia do Excel =TESTET({10;12;13;11;14;12;10;11};{15;13;14;10;13;13;14;13};2;1). A função TESTET retornou o $p\text{-value}$ igual a 0,087623, mostrando que não se deve rejeitar a hipótese nula, ou que não há evidência significativa para afirmar que o procedimento de regulagem proposto reduza o consumo de combustível. Este problema está resolvido na planilha Problemas 10 a 12, incluída na pasta Capítulo 13.

Problema 11

No início de cada ano a instituição financeira contrata *trainees* que recebem treinamento dentro da própria instituição. Para avaliar os resultados do programa de treinamento, os *trainees* realizam um teste inicial de conhecimento. Depois da conclusão do programa de treinamento, os *trainees* realizam um novo teste equivalente ao teste inicial. Foram escolhidos aleatoriamente 10 *trainees* cujas notas de 0 a 100 dos dois testes estão registradas na tabela seguinte. Verifique se o programa de treinamento foi eficiente, considerando o nível de significância $\alpha=5\%$.

Trainee	1	2	3	4	5	6	7	8	9	10
Teste inicial	65	60	45	50	48	62	66	42	45	56
Teste final	80	65	92	78	74	58	72	83	90	78

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=0,2586\%$. O programa de treinamento foi eficiente. Este problema está resolvido na planilha Problemas 10 a 12, incluída na pasta Capítulo 13.

Problema 12

O gerente do departamento médico da empresa recebeu a tabela seguinte com o peso antes e depois da dieta de dois meses que dez funcionários da empresa aceitaram realizar. Verifique a eficiência do programa de redução de peso considerando o nível de significância $\alpha=5\%$.

Funcionário	1	2	3	4	5	6	7	8	9	10
Antes	120,3	99,4	78	84,2	87	79	121	142,6	83	100
Depois	110,5	95,2	72,2	84	87	74	116,5	131,2	84,5	93,5

R: Rejeitar a hipótese nula, pois $p\text{-value}=2 \times P(t \leq t_0)=0,6917\%$. O programa de redução de peso foi eficiente. Este problema está resolvido na planilha Problemas 10 a 12, incluída na pasta Capítulo 13.

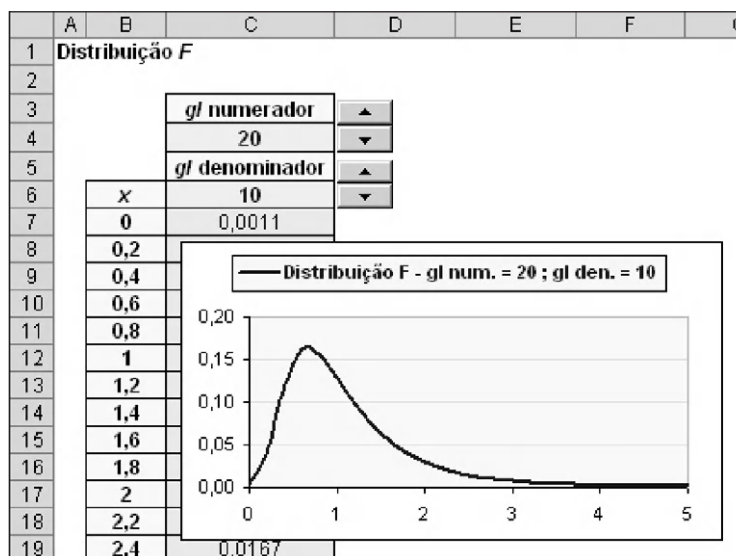
Distribuição F

Na parte inicial deste capítulo, mostramos que o procedimento de teste de hipóteses para a diferença das médias de duas populações é frequentemente utilizado para determinar se é ou não razoável concluir que as médias das duas populações são diferentes. Também é frequente verificar se é ou não razoável concluir que as variâncias das duas populações são diferentes. Para verificar se duas populações independentes têm a mesma variância, é utilizada a estatística da relação das variâncias das amostras S_1^2/S_2^2 retiradas de duas populações. Se as distribuições das duas populações forem normais, então a relação S_1^2/S_2^2 tem distribuição F . Sempre que as distribuições das populações forem normais, a distribuição F será também utilizada para comparar simultaneamente duas ou mais médias, procedimento denominado *análise da variância*, apresentado no Capítulo 14.

A Figura 13.9 mostra uma das possíveis distribuições F dependentes de dois parâmetros, o número de graus gl do numerador e o número de graus gl do denominador. A curva da distribuição F foi construída com a função estatística $DISTF$ na planilha **Distribuição F**, incluída na pasta **Capítulo 13**. Clicando nos dois botões giratórios, um por vez, você poderá ver o comportamento da curva em função dos dois parâmetros. As principais características da distribuição F são:

- A distribuição F é contínua e sempre positiva com valores no intervalo $(0, +\infty)$. A distribuição F tem inclinação positiva.
- Há uma família de distribuições F identificadas por dois parâmetros, graus de liberdade do numerador v_1 e graus de liberdade do denominador v_2 . A forma final da distribuição depende dos graus de liberdade v_1 e v_2 , como mostra a Figura 13.9.

FIGURA 13.9 Família de distribuições F .



EXEMPLO 13.5

Calcule o F crítico F_c da distribuição F com probabilidade de 5% na cauda superior de superar o valor do F crítico, considerando que o número de graus de liberdade do numerador é 6 e o do denominador 10.

Solução. Tradicionalmente, os cálculos com a distribuição F são realizados utilizando tabelas equivalentes às apresentadas para a distribuição Z e para a distribuição t . No capítulo Tabelas, no final do livro, você encontrará duas tabelas da distribuição F , uma para o nível de significância $\alpha=0,01$ e a outra para $\alpha=0,05$. Na pasta **Tabelas** do Excel, incluída na página do livro, no site da Editora, você encontrará a planilha **F_DISTR**, que permite construir a tabela de valores críticos de F para qualquer valor de nível de significância registrado na célula C4 dessa planilha. Continuando com o exemplo, a tabela seguinte apresenta parte da tabela da distribuição F

para o nível de significância $\alpha=0,05$. Nos cabeçalhos das colunas, estão registrados os graus de liberdade do numerador e, nos cabeçalhos das linhas, os graus de liberdade do denominador.

Nível de significância $\alpha=0,05$

	1	2	3	4	5	6	7	8
1	161	199	216	225	230	234	237	239
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85

O F_c da distribuição F com 6 graus de liberdade do numerador e 10 graus de liberdade do denominador, correspondente ao nível de significância 0,05 na cauda superior obtido da *tabela*, F é 3,22. Para informar os valores que participam do F crítico, costuma-se escrever $F_c(\alpha;V_1;V_2)=F_c(0,05;6;10)=3,22$.

O F crítico pode ser obtido com a função estatística INVF do Excel.

• **INV(probabilidade; gl_numerador; gl_denominador)**

A função estatística INVF⁹ retorna o F crítico F_c da distribuição F para uma dada *probabilidade* na cauda superior da distribuição F , e os graus de liberdade do numerador e do denominador, respectivamente, os argumentos *gl_numerador* e *gl_denominador*. A função INVF é a função inversa da DISTF, que será apresentada a seguir. Como o cálculo do F crítico é um procedimento iterativo, se depois de realizar 100 iterações não for alcançado o resultado com um erro de $\pm 3 \times 10^{-7}$, a função INVF retornará o valor de erro #N/A.

Nesse exemplo, o F crítico pode ser obtido registrando numa célula vazia de uma planilha Excel a fórmula =INV(0,05;6;10), que retornará o $F_c=3,2172$. A fórmula =DISTF(3,2173;6;10) registrada em uma célula da planilha retornará o valor 0,0500, que é a probabilidade $P(F \geq 3,2173)$ na cauda superior da distribuição F , de outra maneira, a função DISTF retornou o nível de significância de 5%.

• **DISTF(x; gl_numerador; gl_denominador)**

A função estatística DISTF¹⁰ retorna a probabilidade $P(F \geq x)$ de ser superado o valor do argumento x na cauda superior da distribuição F , para os graus de liberdade do numerador e do denominador, respectivamente, os argumentos *gl_numerador* e *gl_denominador*.

Para facilitar o cálculo com as funções estatísticas DISTF e INVF na planilha **Funções F**, incluída na pasta **Capítulo 13**, você tem dois modelos, como mostra a Figura 13.10, utilizando os dados do Exemplo 13.5.

- **Cálculo da probabilidade na cauda superior da distribuição F .** Esse modelo retorna a probabilidade de um determinado valor x ser superado na cauda superior da distribuição, sendo conhecidos os graus de liberdade do numerador e do denominador e utilizando a função DISTF.

⁹ Em inglês, a função estatística INVF é FINV.

¹⁰ Em inglês, a função estatística DISTF é FDIST.

- **Cálculo do F crítico.** Esse modelo retorna o F crítico para uma determinada probabilidade na cauda superior, sendo conhecidos os graus de liberdade do numerador e do denominador e utilizando a função INVF.

FIGURA 13.10

Utilizando as funções F .

	A	B	C	D	E	F
1	Cálculo da probabilidade na cauda superior da distribuição F					
2						
3		Dados			Resultados	
4		x	3,217		$P(F \geq 3,217)$	5,00%
5		gl numerador	6			
6		gl denominador	10			
7						
8						
9		Cálculo do F crítico				
10						
11		Dados			Resultados	
12		Probabilidade	5,00%		$F_{\text{crítico}}(5,00\%; 6; 10)$	3,217
13		gl numerador	6			
14		gl denominador	10			
15						

Os resultados mostrados na Figura 13.10 se referem à cauda superior da distribuição F , procedimentos de cálculo utilizados para construir a tabela da distribuição F e as funções do Excel. Como se deve proceder se for necessário realizar cálculos na cauda inferior da distribuição F ? Denominando F_s ao F crítico da cauda superior e F_i ao F crítico da cauda inferior, para o nível de significância α , demonstra-se que:

$$F_i(v_1; v_2) = \frac{1}{F_s(v_2; v_1)}$$

Nessa expressão, v_1 é o número de graus de liberdade do numerador, e v_2 é o número de graus de liberdade do denominador. Observe que, para o cálculo do F crítico na cauda inferior, é utilizado o procedimento de cálculo do F crítico da cauda superior, porém permutando os graus de liberdade. A Figura 13.11 mostra o *modelo* que calcula o F crítico nas duas caudas da distribuição F . Esse modelo foi construído a partir da linha 17 da planilha **Funções F**, incluída na pasta **Capítulo 13**. Ao selecionar a cauda clicando no botão de opção correspondente, os nomes das células B21 e B22 permutam o nome de denominador por numerador e vice-versa.

FIGURA 13.11

Cálculo do F crítico nas duas caudas.

	A	B	C	D	E	F
17	F crítico nas duas caudas					
18						
19		Dados			F crítico na cauda	
20		Probabilidade	5,00%		<input type="radio"/> Superior	<input checked="" type="radio"/> Inferior
21		gl denominador	6			
22		gl numerador	10		$F_{\text{crítico}}(5,00\%; 10; 6)$	0,246
23						

Teste F

Como já mencionado, frequentemente precisamos verificar se é ou não razoável concluir que as variâncias das duas populações são diferentes. O teste F é um teste de hipóteses utilizado para verificar se as variâncias de duas populações com distribuição normal são diferentes, ou para verificar qual das duas populações com distribuição normal têm mais variabilidade. De outra maneira, conhecidas duas

amostras com qualquer tamanho, o teste F dá condições para determinar se as duas amostras pertencem à mesma população. O procedimento estatístico é o seguinte:

1. De duas populações com distribuição normal são retiradas duas amostras aleatórias com variâncias σ_1^2 e σ_2^2 .
2. O teste tem as hipóteses:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Essas hipóteses podem ser reescritas como:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

3. Se as variâncias das populações não forem conhecidas, as variâncias das amostras deverão ser utilizadas, pois são as melhores estimativas das respectivas variâncias das populações. Com as variâncias será calculado o F observado $F_o = \frac{S_1^2}{S_2^2}$. Como regra, a variância do numerador de F_o deve ser a da amostra que tiver maior variância, pois, com essa escolha, F_o sempre será maior do que 1 e, consequentemente, somente será utilizada a cauda superior da distribuição F . A partir desse momento, o índice i sempre identificará a amostra com maior variância.
4. Se n_1 e n_2 forem os tamanhos das amostras aleatórias retiradas das populações, a distribuição F terá $v_1 = n_1 - 1$ graus de liberdade do numerador e $v_2 = n_2 - 1$ graus de liberdade do denominador.
5. Com o nível de significância α e os graus de liberdade do numerador e do denominador, será obtido o F crítico F_c da tabela da distribuição F ou com a função INVF.
6. A decisão do teste de hipóteses será realizada, conforme a Figura 13.12:
 - Comparando o F observado F_o e o F crítico F_c . Se $F_o > F_c$, a hipótese nula deverá ser rejeitada; caso contrário, a hipótese nula será aceita.
 - Comparando o $p\text{-value} = P(F \geq F_o)$ e o nível de significância adotado α . Se $p\text{-value} < \alpha$, a hipótese nula deve ser rejeitada; caso contrário, se $p\text{-value} > \alpha$, a hipótese nula deverá ser aceita.

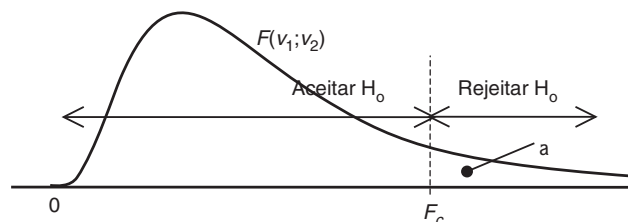


FIGURA 13.12
Decisão com a
Distribuição F .

As formas equivalentes de anunciar a conclusão do teste de hipóteses aplicando a distribuição F :

- Se $F_o < F_c$ ou $p\text{-value} > \alpha$ então:
 - A hipótese nula H_0 deve ser aceita.
 - As variâncias das populações não são significativamente diferentes.

- É razoável aceitar que a diferença entre as variâncias das populações seja devida somente à variabilidade amostral.
- O resultado não é estatisticamente significativo.
- Se $F_o > F_c$ ou $p\text{-value} < \alpha$, então:
 - A hipótese nula H_0 não deve ser aceita.
 - As variâncias das populações são significativamente diferentes.
 - Não é razoável aceitar que a diferença entre as variâncias das populações seja devida apenas à variabilidade amostral.
 - O resultado é estatisticamente significativo.

EXEMPLO 13.6

Verifique se há diferença nas variâncias de duas populações com distribuição normal, conhecendo as medidas estatísticas registradas na tabela seguinte e extraídas dessas populações e considerando o nível de significância $\alpha=5\%$.

	Amostra 1	Amostra 2
n	17	21
Média	2,00	1,14
Variância	1,35	0,61

Solução. Começamos por estabelecer as hipóteses:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

$$\text{Com as variâncias das amostras temos } F_o = \frac{S_1^2}{S_2^2} = \frac{1,35}{0,61} = 2,2131.$$

Como os tamanhos das amostras aleatórias retiradas de populações normais são $n_1=17$ e $n_2=21$, a distribuição F possui $v_1=16$ graus de liberdade do numerador e $v_2=20$ graus de liberdade do denominador.

- O F crítico é igual a $F_c=2,18398$, resultado obtido registrando em uma célula de uma planilha Excel a fórmula $=\text{INV}(0,05;16;20)$, com $\alpha=5\%$.
- Como $F_o > F_c$, a hipótese nula não deve ser aceita, há evidência de que a diferença entre as variâncias é significativa.
- O $p\text{-value}$ é igual a 4,72%, resultado obtido registrando em uma célula de uma planilha Excel a fórmula $=\text{DISTF}(2,2131;16;20)=0,047246$.
- Como ou $p\text{-value} < \alpha$, a hipótese nula não deve ser aceita.

Esse exemplo também pode ser resolvido utilizando o modelo da planilha **Modelo F**, incluída na pasta **Capítulo 13**. As células pintadas de azul são células de dados, as células verdes são resultados e as demais células amarelas são células de títulos.

- Nos intervalos C5:D7, são informadas as medidas estatísticas das amostras. Você não deve ter a preocupação de registrar na primeira coluna a amostra com maior variância, pois o modelo está preparado para reconhecê-la e mudar os títulos das células C4 e C5.
- Na célula C8, deve ser informado o nível de significância *Alfa*.

	A	B	C	D	E	F	G
1	Teste F - Diferença entre duas variâncias						
2							
3	Dados				Resultados		
4		Amostra 1	Amostra 2		Fo: Var1/Var2	2,21	
5	n	17	21		gl-numerador	16	
6	Média	2,00	1,14		gl-denominador	20	
7	Variância	1,35	0,61		F crítico	2,18	
8	Alfa	5,00%			p-value	4,72%	
9							Rejeitar Ho
10							

Os resultados fornecidos pelo *modelo* são:

- Na célula G4, é calculado e registrado o valor do $F_{\text{observado}}$ F_o , calculado com a maior das duas variâncias no seu numerador.
- Nas células G5 e G6, são registrados os graus de liberdade identificando a amostra com maior variância.
- Na célula G7, é calculado o F crítico e na célula G8 o p -value.
- Na célula G9, é registrada por extenso a decisão da análise comparando os valores necessários para isso. Observe que os dois procedimentos sempre dão a mesma decisão.

Os exemplos anteriores partiram das medidas estatísticas das amostras extraídas de duas populações normais. A seguir, o Exemplo 13.7 mostrará como realizar o mesmo teste utilizando a função FTEST e a ferramenta de análise *Teste-F: duas amostras para variâncias*.

EXEMPLO 13.7

As amostras 1 e 2 registradas nas colunas H e I da planilha **Funções F**, incluída na pasta **Capítulo 13**, foram retiradas de duas populações com distribuição normal. Verifique se há diferença nas variâncias, considerando o nível de significância $\alpha=5\%$.

Solução. A figura a seguir mostra como obter resultados operando com a função TESTF na planilha **Funções F**.

	G	H	I	J	K
1	Função TESTEF				
2					
3		Amostra 1	Amostra 2		Função TESTEF
4		2,09	0,58		0,118
5		3,60	0,10		
6		-0,16	1,68		
7		1,07	1,24		
8		2,80	1,64		
9		1,33	0,31		
10		2,88	2,08		
11		1,30	1,20		
12		2,45	1,79		
13			1,20		
14					

• TESTEF(matriz1; matriz2)

A função estatística TESTEF¹¹ retorna a probabilidade na cauda superior da distribuição F da relação das variâncias das amostras registradas nos argumentos *matriz1* e *matriz2*. Em outras palavras, a função TESTEF retorna o p -value do F observado sem registrar esse resultado.

Para calcular o p -value na célula K4 da planilha **Funções F**, foi registrada a fórmula =TESTEF(H4:H12;I4:I13), retornando o resultado 0,118. Tome cuidado com esse resultado, pois, na realidade, o valor 0,118 é o dobro do

¹¹ Em inglês a função estatística TESTEF é FTEST.

resultado correto, 0,05915. De outra maneira, a especificação da função TESTEF estabelece que retorno é o p -value ou a probabilidade $P(F \geq F_0)$ referente à cauda superior da distribuição F ; entretanto, como você pode verificar utilizando o modelo F , o resultado dessa função é igual a $2 \times P(F \geq F_0)$, conforme se pode também confirmar com o resultado da ferramenta de análise *Teste-F: duas amostras para variâncias*.

Ferramenta de análise teste-F: duas amostras para variâncias

A ferramenta de análise *Teste F: duas amostras para variâncias*¹² realiza análises estatísticas e teste de hipóteses de duas variâncias a partir dos valores de duas amostras.

FIGURA 13.13 Caixa de diálogo *Teste-F: duas amostras*.

Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro. Escolhendo a ferramenta **Teste F: duas amostras para variâncias** e depois clicando no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostrado na Figura 13.11, depois de selecionadas as opções do Exemplo 13.7 na planilha **Funções F** a partir da célula M2. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel exibirá a página *Sobre a caixa de diálogo Teste-f: duas amostras para variâncias* pertencente à *Ajuda do Excel*.

Como o procedimento de trabalho desta ferramenta é o mesmo das ferramentas anteriores, somente serão apresentados os resultados. Depois de completar as informações na caixa de diálogo, Figura 13.13, clicando no botão **OK**, o Excel apresenta a tabela com os resultados, conforme a Figura 13.14.

FIGURA 13.14 *Teste-F: duas amostras para variâncias*.

	L	M	N	O
1		Ferramenta de análise		
2		Teste-F: duas amostras para variâncias		
3				
4			<i>Amostra 1</i>	<i>Amostra 2</i>
5		Média	1,92888889	1,182
6		Variância	1,32611111	0,43681778
7		Observações	9	10
8		gl	8	9
9		F	3,0358451	
10		P(F<=f) uni-caudal	0,05914965	
11		F crítico uni-caudal	3,22958726	
12				

¹² Em inglês, a ferramenta *Teste-F: duas amostras para variâncias* é *F-Test: two samples for variances*.

Problemas

Problema 13

Verifique se há diferença significativa entre as variâncias das populações cujas medidas estatísticas das amostras estão registradas na seguinte tabela, considerando $\alpha=5\%$ na cauda superior e utilizando o Modelo F.

	Amostra 1	Amostra 2
<i>n</i>	7	8
<i>Média</i>	56	59
<i>Variância</i>	144	25

R: Rejeitar a hipótese nula, pois $p\text{-value}=1,83$.

Problema 14

O percurso entre os dois aeroportos da cidade é realizado por ônibus seguindo dois trajetos diferentes. Estamos interessados em conhecer se há diferença entre os tempos despendidos nos dois trajetos, considerando o nível de significância de 5% na cauda superior. Para isso, foram realizadas as medições registradas na seguinte tabela.

	Trajeto 1	Trajeto 2
<i>n</i>	16	12
<i>Média</i>	105	116
<i>Variância</i>	10,50	8,3

R: Aceitar a hipótese nula, pois $F_o=1,265 < F_c=2,719$. Não há diferença significativa entre os tempos despendidos pelos ônibus nos dois trajetos.

Problema 15

Há interesse em conhecer se existe diferença entre as variabilidades dos retornos diários de uma ação antes e depois do *Plano Real*, considerando o nível de significância de 5% na cauda superior da distribuição F. Para isso, foram registrados os retornos durante 21 dias antes do início do plano, e os retornos durante 28 dias depois do início do plano.

	Antes	Depois
<i>n</i>	21	28
<i>Média</i>	0,035	0,011
<i>Variância</i>	0,02	0,012

R: Aceitar a hipótese nula, pois $F_o=1,67 < F_c=1,97$. Não há diferença significativa entre as variabilidades dos retornos diários antes e depois do *Plano Real*.

Distribuição χ^2

Na maioria dos testes de hipóteses aplicados até este momento, a distribuição da população era conhecida. Há casos em que a distribuição da população não é conhecida e se deseja verificar se um grupo de va-

lores segue um determinado modelo de probabilidade teórico utilizando a distribuição Qui-Quadrado, ou χ^2 . A Figura 13.15 mostra uma das possíveis distribuições χ^2 dependente do número de graus de liberdade gl . A curva da distribuição χ^2 foi construída com a função estatística DIST.QUI na planilha **Distribuição Qui-Quadrado**, incluída na pasta **Capítulo 13**. Clicando no botão giratório é possível variar o número de graus de liberdade da distribuição. A fórmula `=DIST.QUI($B5;C$4)-DIST.QUI($B6;C$4)` foi registrada na célula C5 e depois copiada até a célula C30, como mostra a Figura 13.15.

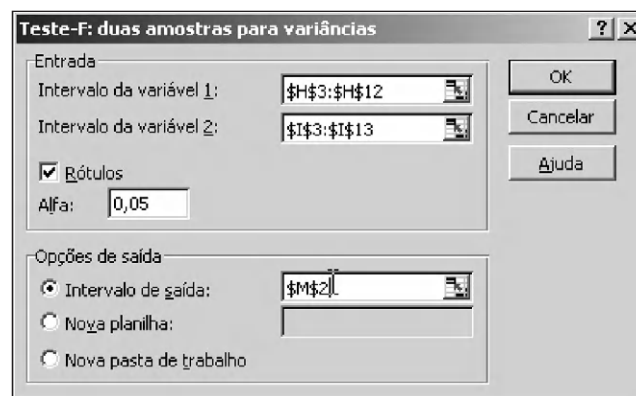
- **DIST.QUI(x ; graus_liberdade)**

A função estatística DIST.QUI¹³ retorna a probabilidade $P(\chi^2 \geq x)$ de superar o valor do argumento x na cauda superior da distribuição qui-quadrado, considerando os *graus_liberdade* especificados nesse argumento; a função retorna o *p-value* na cauda superior da distribuição.

As principais características da distribuição χ^2 são:

- A distribuição χ^2 é contínua e sempre positiva com valores no intervalo $(0, +\infty)$, a distribuição χ^2 tem inclinação positiva.
- Há uma família de distribuições χ^2 identificadas pelo parâmetro graus de liberdade gl .

FIGURA 13.15
Distribuição
qui-quadrado.



Para mostrar uma das aplicações da distribuição χ^2 , voltemos à simulação da retirada de uma bola de uma urna contendo dez bolas numeradas realizada no Capítulo 1, utilizando a função ALEATÓRIOENTRE cujos retornos têm distribuição uniforme discreta. Que significa que grupos de números gerados por essa função tenham distribuição uniforme discreta? Lembrando o processo descrito no Capítulo 1, considere que uma urna tenha dez bolas pequenas numeradas de 0 a 9 que formam o conjunto $\{0, 1, 2, 3, \dots, 9\}$. Suponha que realiza 500 extrações com reposição de uma bola dessa urna e toma nota do número correspondente, por exemplo, numa coluna de uma planilha do Excel. Como foi visto, a longo prazo, a frequência esperada de cada bola será 10% do total de retiradas ou amostras. Tecnicamente, todos os dez números terão a mesma frequência e sua distribuição será uniforme e discreta.

Contudo, as frequências observadas numa simulação de 500 amostras se situarão dentro de um intervalo ao redor de 10%, como mostra o gráfico de barras verticais da Figura 13.16, construído na planilha **Simulação**, incluída na pasta **Capítulo 1**. Na Figura 13.16, as barras verticais com a mesma altura representam as frequências esperadas e, as outras, as frequências observadas correspondentes. A diferença entre as frequências observadas e as correspondentes frequências esperadas pode ser atribuída à variabilidade amostral, a falhas do gerador de números aleatórios ou ao reduzido tamanho da amostra.

13 Em inglês, a função DIST.QUI é CHIDIST.

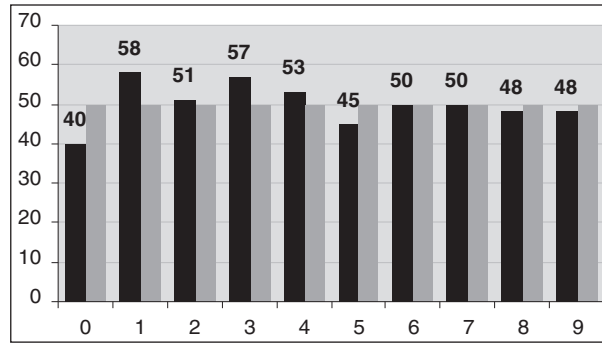


FIGURA 13.16 Simulação de 500 retiradas de uma bola com reposição.

Teste de hipóteses

Para verificar se a diferença entre as frequências observadas e as correspondentes frequências esperadas pode ser atribuída à variabilidade amostral ou à função ALEATÓRIOENTRE, aplica-se o teste de hipóteses a seguir, desconsiderando o tamanho da amostra:

H_0 : A distribuição das frequências observadas é discreta uniforme.

H_1 : A distribuição das frequências observadas *não* é discreta uniforme.

Para analisar e verificar se os dígitos dos números gerados pela função ALEATÓRIOENTRE¹⁴ são realmente aleatórios, foi construído o *modelo* registrado na planilha **Teste Qui-Quadrado**, incluída na pasta **Capítulo 13**, como mostra a Figura 13.17. Clicando no botão **Nova Simulação**, o *modelo* gera uma nova série de 500 amostras. Para realizar o teste de hipóteses, foi construída a seguinte planilha:

- A fórmula =ALEATÓRIOENTRE(0;9) foi registrada na célula B4. Depois, essa fórmula foi copiada até a célula B503.
- No intervalo D5:E15, foi construída a tabela de frequências absolutas, como mostrado no Capítulo 2.
- Nas dez células do intervalo F5:F14, foi registrado o valor 50.
- Na célula G5, foi registrada a fórmula =(E5-F5)^2/F5, que depois foi copiada até a célula G14, formando a coluna *Estatística*. Essa fórmula retorna o resultado de $\frac{(O_i - E_i)^2}{E_i}$, onde:

- O_i é a frequência observada do número i nas 500 amostras, que varia para cada nova simulação.
- E_i é a frequência esperada do número i nas 500 amostras.

- Na célula E19, foi registrada a soma dos resultados do intervalo G5:G14, que varia em cada nova simulação. Esse resultado é o valor observado $\chi_o^2 = \sum_{i=0}^9 \frac{(O_i - E_i)^2}{E_i}$ da estatística χ^2 com $(k-1)$ graus de

liberdade, sendo k a quantidade de categorias ou números aleatórios. Nesse caso, o número de graus de liberdade é $(k-1)=10-1=9$, resultado obtido na célula E19, a partir da quantidade de números aleatórios calculada na célula E18.

- Para o nível de significância $\alpha=5\%$ registrado na célula E17, o valor crítico χ_c^2 da distribuição qui-quadrado para nove graus de liberdade é 16,919. Esse resultado pode ser obtido:
 - Da tabela da distribuição qui-quadrada, registrada no capítulo *Tabelas* no final do livro, para cinco valores do nível de significância α . Na pasta **Tabelas** do Excel, na página do livro, no site da Editora, você encontrará a planilha **Q_DISTR**, que permite construir a tabela de valores críticos de F para qualquer valor de nível de significância registrado em qualquer célula do intervalo

¹⁴ Alguns autores denominam de números *pseudoaleatórios*.

FIGURA 13.17 Teste Qui-Quadrado aplicado à função ALEATÓRIOENTRE.

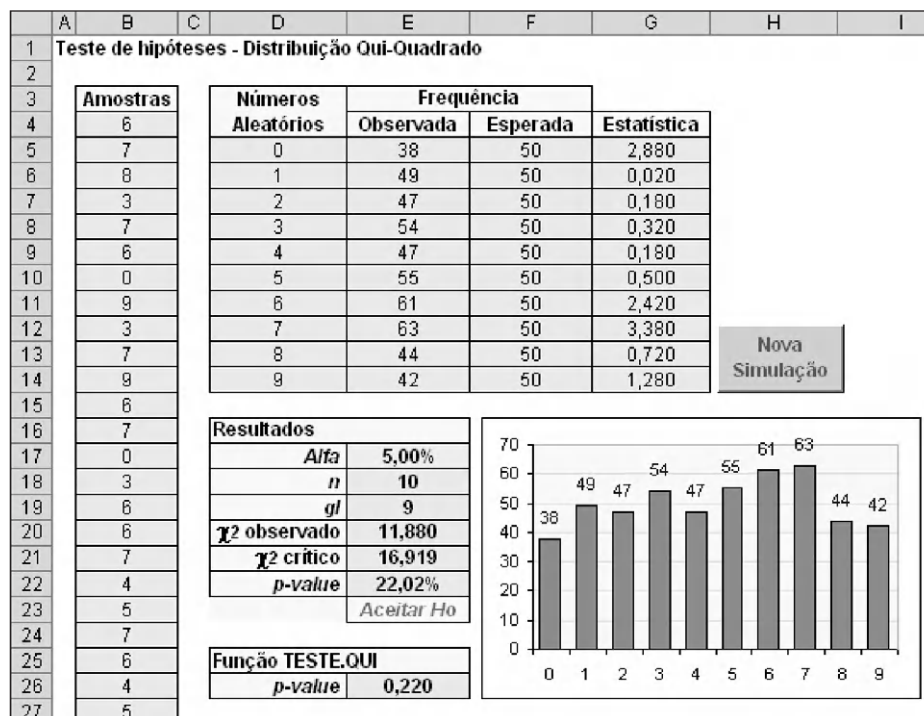


FIGURA 13.18 Tabela da distribuição qui-quadrado, parcial.

q.l.	Nível de significância				
	0,10	0,05	0,03	0,01	0,005
1	2,71	3,84	5,02	6,63	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,81	9,35	11,34	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,64	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,53	20,09	21,95
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19

C4:G4. Continuando com o exemplo, a tabela da Figura 13.18 apresenta parte da tabela da distribuição qui-quadrada.

- Com a função INV.QUI, registrando a fórmula =INV.QUI(0,05;9), que retorna o valor 16,919.
- INV.QUI(*probabilidade; graus_liberdade*)

A função estatística INV.QUI¹⁵ retorna o valor crítico na cauda superior da distribuição qui-quadrada para a *probabilidade* e os *graus_liberdade* especificados. A função INV.QUI é a função inversa da DIST.QUI.

- O procedimento de cálculo pode ser simplificado utilizando a função estatística TESTE.QUI do Excel.
- TESTE.QUI(*intervalo_observado; intervalo_esperado*)

A função estatística TESTE.QUI¹⁶ retorna o *p-value* do valor observado na cauda superior da distribuição qui-quadrada para o *intervalo_observado* e o *intervalo_esperado* especificados. Essa função retorna o mesmo resultado que a função DIST.QUI, porém sem necessidade de realizar os cálculos da coluna G da planilha Teste Qui-Quadrado. Na célula E26 da planilha Teste Qui-Quadrado, foi registrada a fórmula =TESTE.QUI(E5:E14;F5:F14), que retorna o mesmo valor de *p-value* 0,2202 obtido anteriormente, como mostra a Figura 13.17.

¹⁵ Em inglês, a função INV.QUI é CHINV.

¹⁶ Em inglês, a função TESTE.QUI é CHITEST.

- Como na simulação realizada e registrada na figura 13.18, o valor observado $\chi_o^2 = 11,88$ é menor do que o valor crítico $\chi_c^2 = 16,92$, a hipótese nula deve ser aceita, e a distribuição dos valores observados é uniforme. Portanto, considerando essa amostra, a função ALEATÓRIOENTRE do Excel deve ser aceita. Observe que recalculando a planilha várias vezes, poucas amostras rejeitarão a hipótese nula. A longo prazo, no nível de significância = 5%, espera-se que 95% das amostras aceitem a hipótese nula.
- Na célula E23, foi registrada por extenso a decisão da análise comparando os valores necessários para isso. Observe que os dois procedimentos de análise sempre dão a mesma decisão.

Como o tamanho da amostra é importante, sugerimos que você repita esse procedimento construindo uma planilha para 1.000 ou mais amostras, considerando a análise apresentada no Capítulo 5 sobre a lei dos grandes números. O exemplo anterior mostra uma das aplicações da distribuição qui-quadrado denominada teste de aderência ou qualidade de ajustamento. Outras áreas de aplicação da distribuição qui-quadrado estão relacionadas com a tabela de contingências, por exemplo, a verificação entre as proporções de duas populações, a determinação da independência de duas variáveis aleatórias discretas etc.

EXEMPLO 13.8

O diretor de RH da empresa multinacional deseja conhecer se o hábito de fumar observado dos funcionários classificados por sexo na nova planta em outro país é diferente do hábito de fumar dos funcionários da matriz, considerado como hábito esperado. Os resultados da pesquisa estão registrados nos intervalos B3:D5 e B7:D9 da figura seguinte. Considerando o nível de significância de 5%, verifique se há diferença entre as duas filiais.

Solução. Na planilha **Exemplo 13.8**, incluída na pasta **Capítulo 13**, foi resolvido o exemplo como mostra a figura seguinte. Para verificar se a diferença entre as proporções do hábito de fumar das duas plantas, aplicamos o teste de hipóteses seguinte:

H_0 : Não há diferença entre as duas plantas.

H_1 : Há diferença entre as duas plantas.

Para obter o valor observável da estatística qui-quadrado:

- Na célula F8, foi registrada a fórmula $= (C4 - C8)^2 / C8$. Depois, essa fórmula foi copiada no intervalo F8:G9. O valor observado da distribuição qui-quadrado é 1,538, resultado calculado com a fórmula $= \text{SOMA}(F8:G9)$, registrada na célula C14.
- O número de graus de liberdade é o resultado da multiplicação do número de colunas c menos um, vezes o número de linhas l menos um da tabela de probabilidades conjuntas e marginais, ou $(c-1) \times (l-1)$. Nesse caso, como a tabela tem duas linhas e duas colunas, $gl=1$, esse valor foi registrado na célula C13.

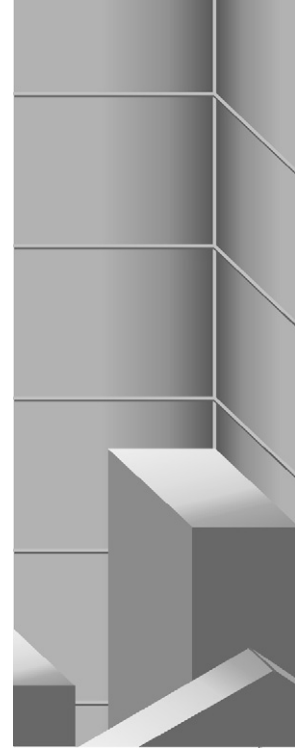
	A	B	C	D	E	F	G
1	Exemplo 13.8						
2							
3		Hábito esperado	Mulher	Homem			
4		Fuma	68	82			
5		Não fuma	462	388			
6							
7		Hábito observado	Mulher	Homem	Estatística		
8		Fuma	75	75	0,6533	0,6533	
9		Não fuma	455	395	0,1077	0,1241	
10							
11		Resultados					
12		Alfa	5,00%				
13		gl	1				
14		χ^2 observado	1,538				
15		χ^2 crítico	3,841				
16		p-value	21,49%				
17			Aceitar H_0			Função TESTE.OUI	
18						p-value	0,2149

- Para o nível de significância de 5% registrado na célula C12 e um grau de liberdade, o valor crítico da distribuição qui-quadrado é 3,841, como se pode obter da tabela da distribuição da Figura 13.18, ou utilizando a função INV.QUI registrando na célula C15 a fórmula =INV.QUI(0,05;1).
- Na célula C16, foi registrada a fórmula =DIST.QUI(C14;C13), que retorna o *p-value* igual a 21,49%. Como o valor observado 1,538 é menor do que o valor crítico 3,841, a hipótese nula deve ser aceita e, portanto, não há evidências significativas de que a proporção dos funcionários que fumam e não fumam classificados por sexo entre os funcionários da matriz e da nova unidade sejam diferentes. Da mesma maneira, sendo o *p-value* maior do que o nível de significância, a hipótese nula deve ser aceita. Na célula C17, foi registrada por extenso a decisão da análise, comparando os valores necessários para isso. Observe que os dois procedimentos sempre dão a mesma decisão.

Nesse caso, também, o procedimento de cálculo pode ser simplificado utilizando a função estatística TESTE.QUI do Excel. Na célula G17 da planilha **Exemplo 13.8**, foi registrada a fórmula =TESTE.QUI(C4:D5;C8:D9) que retorna o mesmo valor de *p-value* 0,2149 obtido anteriormente, como mostra a figura anterior.

Capítulo 14

ANÁLISE DA VARIÂNCIA



Os dois capítulos anteriores introduziram os testes de hipóteses referentes à média de uma população, para a diferença entre médias de duas populações e a comparação de variâncias de duas populações. Neste capítulo, o procedimento de teste de hipóteses será utilizado para comparar as médias de mais de duas populações. Embora o nome não mostre o objetivo real do procedimento, a *análise da variância* ou ANOVA é um teste de hipóteses de médias de duas ou mais populações, procedimento muito útil para comparar, por exemplo:

- A eficiência de diversas marcas de remédios para o tratamento de uma mesma doença, o controle de pressão alta.
- O consumo em km/litro de um modelo de carro abastecido com combustíveis do mesmo tipo, porém de marcas diferentes.
- A eficiência de uma lavoura tratada com diferentes fertilizantes.
- O tempo de reação de uma pessoa em função de estímulo de luz de quatro cores diferentes.

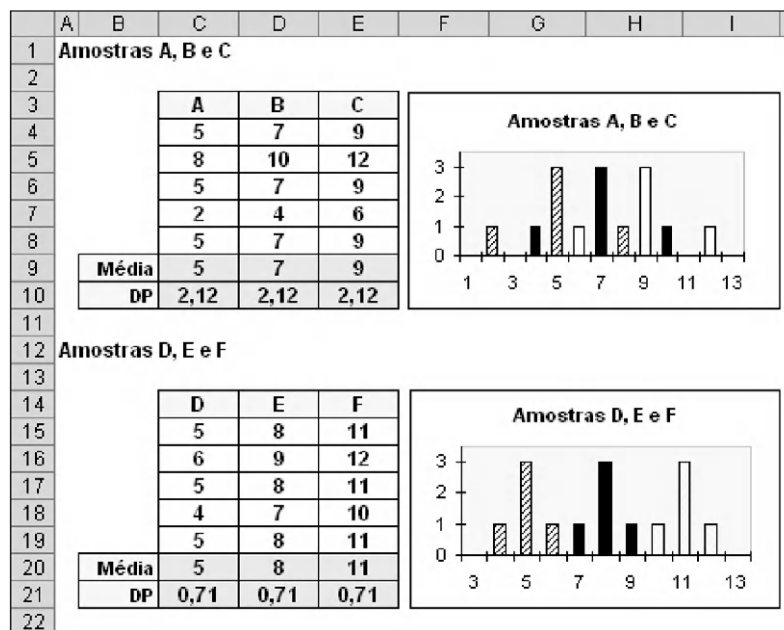
Como introdução, a tabela da Figura 14.1 registra dois grupos com três amostras cada um, A-B-C e D-E-F. Em cada grupo, as três amostras aleatórias independentes de tamanho $n=5$ foram retiradas de três populações com distribuição normal e variâncias iguais. Nas duas últimas linhas da tabela, foram registradas as medidas estatísticas das seis amostras (média e desvio padrão), resultados obtidos na planilha **Início**, incluída na pasta **Capítulo 14**. Com o objetivo de analisar as médias e os desvios padrão dos dois grupos de amostras, também foram construídos os histogramas das amostras de cada grupo.

Primeiro grupo, amostras A, B e C. As três amostras têm o mesmo desvio padrão, a diferença entre suas médias é igual a dois e ainda:

- A diferença entre as médias das três amostras A, B e C corresponde a 0,94 desvio padrão das amostras. Dessa maneira, a média da amostra B está contida no intervalo de 0,94 desvio padrão ao redor da média da amostra A e da amostra C.
- A diferença entre as médias das amostras A e B, e também das amostras B e C, é pequena considerando a dispersão dos valores dessas amostras, como se pode ver no histograma correspondente.

FIGURA 14.1

Medidas estatísticas e histogramas de dois grupos de amostras.



Portanto, apesar de ter retirado as amostras de três populações com a mesma média, não se pode esperar que as médias das três amostras sejam iguais. Ao mesmo tempo, a diferença entre as três médias é apenas consequência da variação amostral? Como há uma grande variabilidade nos valores de cada amostra, os resultados desse grupo mostram evidências de que as médias das três populações *não* são diferentes.

Segundo grupo, amostras D, E e F. As três amostras têm o mesmo desvio padrão, a diferença entre suas médias é igual a três e ainda:

- As médias dessas três amostras do segundo grupo são diferentes das médias das amostras do primeiro grupo, porém a diferença entre elas é constante, e os desvios padrão são menores.
- A diferença entre as médias das três amostras D, E e F corresponde a 4,22 desvios padrão das amostras. Dessa maneira, a média da amostra E não está incluída no intervalo de três desvios padrão ao redor da média das outras duas amostras, como se poder ver no histograma correspondente.
- A diferença entre as médias das amostras D e E, e também das amostras E e F, é grande, considerando a dispersão dos valores dessas amostras, como se pode ver no histograma correspondente.

Da mesma forma que no grupo anterior, não se pode esperar que as médias das três amostras sejam iguais, apesar de ter retirado as amostras de três populações com a mesma média. A diferença entre as três médias é apenas consequência da variação amostral? Como não há grande variabilidade nos valores de cada amostra, os resultados desse grupo mostram evidências de que as médias das três populações sejam diferentes.

Conceituação da análise da variância

O objetivo da análise de variância é avaliar se as diferenças observadas entre as médias das amostras são estatisticamente significantes. Como já foi mostrado em outras ocasiões, esse objetivo pode ser colocado de outra maneira: uma variação de médias das amostras pode ser consequência da variação amostral ou é uma boa evidência da diferença entre as médias das populações? A variabilidade total das amostras pode ser dividida em duas partes ou fontes de variabilidade.

- A primeira parte de variabilidade é proveniente das populações serem diferentes, denominada variabilidade *entre*. Quanto maior for a variabilidade *entre*, mais forte é a evidência de as médias das populações serem diferentes.
- A segunda parte de variabilidade é causada pelas diferenças *dentro* de cada amostra, denominada variabilidade *dentro*. Quanto maior for a variabilidade *dentro*, maior será a dificuldade para concluir se as médias das populações são diferentes.

Premissas da Análise da Variância

As populações têm a mesma variância.

As amostras são retiradas de populações com distribuição normal.

As amostras são aleatórias e independentes.

O teste de hipóteses da análise de variância é estabelecido como:

- A hipótese nula H_0 afirma que as k populações têm a mesma média.
- A hipótese alternativa H_1 afirma que nem todas as médias das k populações são iguais.

A classificação dos testes de análise da variância é feita de acordo com o número de fatores de interesse ou que influem na variável dependente.¹ Por exemplo, na verificação da eficiência do crescimento de uma lavoura tratada com quatro tipos de fertilizantes, cada um dos fertilizantes é um fator. Da mesma maneira, na comparação do consumo de carros abastecidos com o mesmo tipo de combustível, porém de três marcas diferentes, cada marca de combustível é um fator. Por que é denominado análise da variância o procedimento que compara médias de grupos diferentes? Porque na preparação das variabilidades *entre* e *dentro* são utilizados os quadrados dos desvios dos valores das amostras, que fazem parte da definição da variância. De maneira formal, o teste de hipóteses para k níveis de um fator é estabelecido da seguinte forma.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_1 : \text{Nem todas as populações têm a mesma média.}$$

Pelas premissas, as variâncias das populações são iguais, e a hipótese nula afirma que as médias das populações são idênticas, enquanto a hipótese alternativa afirma que há alguma diferença entre as médias, independente do número de populações que difiram. A distribuição F conduzirá a decisão de aceitar ou rejeitar a hipótese nula, comparando o F observado F_o , calculado com a expressão:

$$F_o = \frac{\text{Variância entre}}{\text{Variância dentro}} = \frac{S_b^2}{S_w^2}$$

com o F crítico F_c correspondente ao nível de significância α adotado. Podem ser comparados, também, o p -value de F_o e o nível de significância α .

¹ Como o início do procedimento de análise da variância se deu na agricultura, no teste de hipóteses permanecem algumas definições dessa área, como o termo *tratamento*, que define a causa ou fonte de variação dentro de um conjunto de dados.

EXEMPLO 14.1

De três populações normais com variâncias iguais, foram retiradas três amostras aleatórias independentes, como mostra a planilha seguinte. Calcule o F observado F_o .

Solução. Este exemplo foi resolvido na planilha **Exemplo 14.1**, incluída na pasta **Capítulo 14**. No intervalo C10:E12, foram calculadas e registradas três medidas estatísticas das amostras: tamanho n , média e variância. O procedimento de cálculo do F observado F_o é apresentado a seguir, tendo presente que k é o número de amostras ou tratamentos, e n_k é o tamanho ou número de valores de cada amostra.

	A	B	C	D	E	F	G	H
1	Exemplo 14.1							
2								
3								
4			Amostra 1	Amostra 2	Amostra 3		Resultados	
5			10	6	14		k	3
6			8	9	13		Variância dentro	8,83
7			5	8	10		Grande Média	11,07
8			12	13	17		Variância entre	33,47
9			14		16		F observado	3,79
10			11				Alfa	5,00%
11		n	6	4	5		gl numerador	2
12		Média	10,00	9,00	14,00		gl denominador	12
13		Variância	10,00	8,67	7,50		p-value	5,30%
14							Decisão	Aceitar Ho

Cálculo da variância dentro S_w^2 . O primeiro passo é calcular a média das variâncias das amostras utilizando a expressão:

$$S_w^2 = \frac{\sum_{j=1}^k (n_j - 1) \times S_j^2}{n_T - k}$$

Nessa expressão, n_T é o número total de dados das amostras, e $(n_T - k)$ é o número de graus de liberdade. As variâncias das amostras S_j^2 foram multiplicadas por $(n-1)$, pois inicialmente foram calculadas como amostras. Substituindo os dados do exemplo:

$$S_w^2 = \frac{(6-1) \times 10 + (4-1) \times 8,67 + (5-1) \times 7,5}{15-3}$$

$$S_w^2 = \frac{106}{12} = 8,83$$

Observe que o numerador da expressão S_w^2 é a soma dos quadrados dos desvios de cada amostra com relação à sua própria média.

Cálculo da variância entre S_b^2 . O primeiro passo é calcular a *grande média* $\bar{\bar{X}}$, correspondente à média de todas as observações utilizando a expressão:

$$\bar{\bar{X}} = \frac{X_1 + \dots + X_{n_1+n_2+\dots+n_k}}{n_1 + n_2 + \dots + n_k}$$

Substituindo os dados do exemplo na fórmula $\bar{\bar{X}}=11,07$, resultado obtido com:

$$\bar{\bar{X}} = \frac{10+8+\dots+17+16}{15} = 11,07$$

A grande média $\bar{\bar{X}}$ pode ser calculada, também, como média ponderada das médias das k amostras aplicadas a este exemplo:

$$\bar{\bar{X}} = \sum_{i=1}^3 \bar{X}_i \frac{n_i}{n_T}$$

$$\bar{X} = 10 \frac{6}{15} + 9 \frac{4}{15} + 14 \frac{5}{15} = 11,07$$

Conhecida a grande média, a seguir é calculada a variância das médias das amostras com relação à grande média, utilizando a expressão:

$$S_b^2 = \frac{\sum_{j=1}^k n_j \times (\bar{X}_j - \bar{X})^2}{k-1}$$

Substituindo os dados do exemplo, teremos:

$$S_b^2 = \frac{6 \times (10 - 11,07)^2 + 4 \times (9 - 11,07)^2 + 5 \times (14 - 11,07)^2}{3-1}$$

$$S_b^2 = \frac{66,93}{2} = 33,47$$

Cálculo do F observado F_o . Substituindo os resultados anteriores temos o F observado F_o :

$$F_o = \frac{S_b^2}{S_w^2}$$

$$F_o = \frac{33,47}{8,83} = 3,79$$

O F observado mede a variabilidade *entre* por unidade de variabilidade *dentro*, ou quantas vezes a variabilidade das médias das amostras é maior do que a variabilidade amostral. O resultado do Exemplo 14.1 mostra que a variabilidade *entre* os grupos é 3,79 vezes maior do que a variabilidade das amostras.

O resultado do F observado $F_o=3,79$ permite afirmar que as populações sejam diferentes? Dependerá da comparação do F observado F_o com o F crítico F_c correspondente ao nível de significância α adotado, ou comparando o p -value com o nível de significância.

EXEMPLO 14.2

Continuando com o Exemplo 14.1. Verifique se as médias das três populações são iguais, considerando o nível de significância de 5%.

Solução. No Exemplo 14.1 foi obtido o F observado $F_o=3,79$ com os graus de liberdade:

- Do numerador $v_1 = k - 1 = 3 - 1 = 2$.
- Do denominador $v_2 = n_T - k = n_1 + \dots + n_k - k = 6 + 4 + 5 - 3 = 12$.

O resultado do teste de hipóteses pode ser realizado de duas formas diferentes, obtendo a mesma conclusão de aceitar a hipótese nula:

- Comparando o F observado $F_o=3,79$ já determinado com o F crítico a determinar. Para o nível de significância 5%, o $F_c=3,8853$ foi obtido com a função estatística INVF registrando a fórmula =INV(0,05;2;12) numa célula da planilha Excel. Como o valor observado $F_o=3,79$ é menor do que o valor crítico $F_c=3,8853$, a hipótese nula deve ser aceita, pois as médias das amostras não são significativamente diferentes entre si.
- Comparando o p -value a determinar com o nível de significância de 5%. O p -value igual a 0,053 se refere à probabilidade $P(F \geq 3,79)$, cujo resultado foi obtido com a função estatística DISTF, registrando em uma célula da planilha Excel a fórmula =DISTF(3,79;2;12). Como o p -value 5,3% é maior do que o nível de significância 5%, a hipótese nula deve ser aceita.

Este exemplo está resolvido na planilha **Exemplo 14.1**, incluída na pasta **Capítulo 14**, cuja figura já foi apresentada no Exemplo 14.1.

Tabela ANOVA

Os resultados dos exemplos anteriores podem ser agrupados numa tabela denominada ANOVA² que representa o procedimento natural de cálculo, pois o objetivo da tabela ANOVA é obter o F observado utilizando um procedimento numérico.

Sejam k amostras independentes de tamanhos diferentes³ $X_1 = \{X_{1_1}, \dots, X_{n_1}\}$, $X_2 = \{X_{1_2}, \dots, X_{n_2}\} \dots X_k = \{X_{1_k}, \dots, X_{n_k}\}$, com médias $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$.

Cálculo da variância dentro S_w^2 . O numerador da variância *dentro* é a soma dos quadrados dos desvios dos valores de cada amostra com relação à sua própria média.

$$S_w^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2 + \dots + \sum_{i=1}^{n_k} (X_i - \bar{X}_k)^2}{n_1 + n_2 + \dots + n_k - k}$$

Nessa expressão:

- $n_1 + n_2 + \dots + n_k$ é a soma da quantidade de dados das k amostras.
- $n_1 + n_2 + \dots + n_k - k$ determina os graus de liberdade do denominador.
- O numerador da expressão S_w^2 , soma dos quadrados dos desvios de cada amostra com relação à sua própria média, é denominada SSE .⁴
- O resultado da divisão de SSE pelo número de graus do denominador é denominado MSE .⁵ Com esses novos símbolos, $S_w^2 = \frac{SSE}{n_T - k} = MSE$.

Cálculo da variância entre S_b^2 . A variância *entre* pode ser determinada de duas formas, utilizando a grande média $\bar{\bar{X}}$ e utilizando o resultado SSE obtido anteriormente.

- Denominando grande média $\bar{\bar{X}}$ à média geral de todos os dados das amostras calculada com a seguinte expressão:

$$\bar{\bar{X}} = \frac{X_1 + X_2 + \dots + X_{n_1 + \dots + n_k}}{n_1 + n_2 + \dots + n_k}$$

A variância *entre* S_b^2 é obtida com a expressão:

$$S_b^2 = \frac{n_1(\bar{X}_1 - \bar{\bar{X}})^2 + \dots + n_k(\bar{X}_k - \bar{\bar{X}})^2}{k - 1}$$

O numerador da expressão S_b^2 , soma do produto do tamanho de cada amostra vezes os quadrados dos desvios da média de cada amostra com relação à grande média, é denominado SST .⁶ O resultado da divisão de SST por $(k-1)$ é denominado MST ,⁷ ou $S_b^2 = \frac{SST}{k - 1} = MST$.

² ANOVA significa *Analysis of Variance*.

³ *One-way analysis*, em inglês.

⁴ Em inglês, SSE é *Sum of Square Errors*.

⁵ Em inglês, MSE é *Mean Square of Error*.

⁶ Em inglês, SST é *Sum of Squares for Treatments*.

⁷ Em inglês, MST é *Mean Square for Treatments*.

- Denominamos SS^8 a soma dos quadrados dos desvios dos dados das amostras com relação à grande média, como mostra a expressão:

$$SS = \sum_{i=1}^{n_1 + \dots + n_k} (X_i - \bar{\bar{X}})^2$$

Demonstra-se que $SS=SSE+SST$, de onde se obtém o valor de SST . Como $SST=SS-SSE$, a expressão S_b^2 será igual a:

$$S_b^2 = \frac{SS - SSE}{k - 1}$$

Essa expressão mostra que a soma dos quadrados de todos os desvios com relação à grande média SS pode ser dividido em duas somas de quadrados de desvios, SSE referente à soma dos quadrados dos desvios dos dados de cada amostra com relação às suas próprias médias, e SST referente à soma dos quadrados dos desvios da média de cada amostra com relação à grande média.

Cálculo do F observado.⁹ Substituindo as expressões obtidas na fórmula do F observado $F_o = S_b^2 / S_w^2$, temos a expressão.

$$F_o = \frac{\frac{SST}{k-1}}{\frac{SSE}{n_T - k}} = \frac{MST}{MSE}$$

Em função do exposto, é formada a tabela ANOVA, Figura 14.2.

Fonte	gl	SS	MS	F
Entre	$k-1$	SST	$MST = SST/(k-1)$	$F_o = MST/MSE$
Dentro	$n_T - k$	SSE	$MSE = SSE/(n_T - k)$	
Total	$n_T - 1$	SS		

FIGURA 14.2 Tabela ANOVA.

EXEMPLO 14.3

Resolva o Exemplo 14.1 utilizando a tabela ANOVA.

Solução. Na planilha **Exemplo 14.3**, incluída na pasta **Capítulo 14**, foi resolvido esse exemplo, como mostra a figura seguinte. Para realizar o teste de hipóteses, há dois procedimentos:

- Como o F observado 3,789 é menor do que o F crítico 3,885 correspondente ao nível de significância adotado de 5%, a hipótese nula deve ser aceita.
- Como o p -value 5,3% do F observado 3,79 é maior do que o nível de significância adotado, 5%, a hipótese nula deve ser aceita.

⁸ Em inglês, SS é *Sum of Squares*. Esta igualdade pode ser demonstrada.

⁹ Esta expressão será útil na validação da regressão linear, Capítulo 15.

	A	B	C	D	E	F	G	H	I	J
1	Exemplo 14.3				Tabela ANOVA					
2										
3		Amostra 1	Amostra 2	Amostra 3		Fonte	gl	SS	MS	Fo
4		10	6	14		Entre	2	66,93	33,47	3,7887
5		8	9	13		Dentro	12	106,00	8,83	
6		5	8	10		Total	14	172,93		
7		12	13	17		Teste de hipóteses				
8		14		16		Alfa	5,00%			
9		11				F crítico	3,885			
10						p-value	5,30%			
11						Decisão	Aceitar Ho			
12		Resultados parciais								
13		Grande Média		11,07						
14		SS		172,93						
15		SSE		106,00						
16		SST=SS - SSE		66,93						

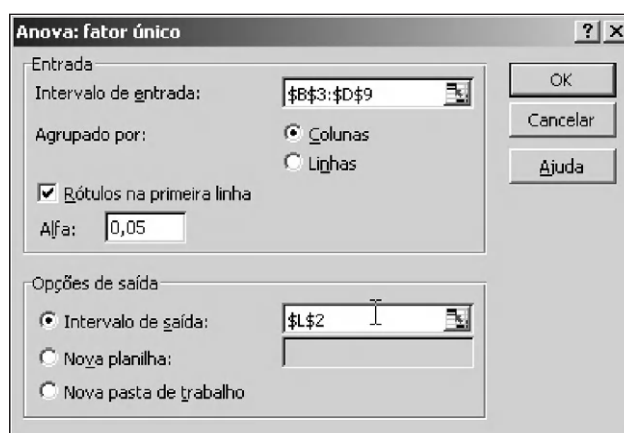
Observe que se o nível de significância do Exemplo 14.3 fosse, por exemplo, $\alpha=1\%$, a decisão continuaria aceitando a hipótese nula. Por quê? Porque o *p-value* é o máximo nível de significância que aceita a hipótese nula.

Ferramenta de análise anova: fator único

O Excel não dispõe de funções estatísticas para análise da variância, porém dispõe de ferramentas de análise. A ferramenta de análise *Anova: fator único* constrói a tabela *Anova* para o teste de hipóteses sobre a igualdade da média de três ou mais populações.

Como o procedimento de trabalho dessa ferramenta é o mesmo que as das ferramentas anteriores, somente serão apresentados alguns detalhes. Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 deste livro. Escolhendo a ferramenta **Anova: fator único** e depois de clicar no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostrado na Figura 14.3, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel exibirá a página *Sobre a caixa de diálogo Anova: fator único* pertencente à *Ajuda do Excel*.

FIGURA 14.3 Caixa de diálogo da ferramenta *Anova: Fator Único*.



No quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três botões de opção que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, nesse caso L2, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá. Mais informações podem ser obtidas no Capítulo 4 ou na Ajuda do Excel.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha Plan1.

K	L	M	N	O	P	Q	R
1	Ferramenta de análise						
2	Anova: fator único						
3							
4	RESUMO						
5	Grupo	Contagem	Soma	Média	Variância		
6	Amostra 1	6	60	10	10		
7	Amostra 2	4	36	9	8,666667		
8	Amostra 3	5	70	14	7,5		
9							
10							
11	ANOVA						
12	Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
13	Entre grupos	66,933333	2	33,466667	3,7886792	0,053035	3,8852903
14	Dentro dos grupos	106	12	8,8333333			
15							
16	Total	172,93333	14				
17							

FIGURA 14.4
Resultados do
Exemplo 14.1
com Anova:
fator Único.

Depois de completar as informações e clicar em OK na caixa de diálogo, o Excel exibirá a tabela com os resultados, conforme a Figura 14.4. No quadro RESUMO, a ferramenta de análise apresenta algumas medidas estatísticas das três amostras. No quadro ANOVA, a ferramenta apresenta os resultados já conhecidos, com pequena diferença de layout e da simbologia empregada no texto.

Para concluir esse tema, se for realizada uma análise de variância nos dois grupos de amostras apresentadas no início deste capítulo, amostras A, B e C e amostras D, E e F, você verificará que, para o nível de significância de 5%, deve-se rejeitar a hipótese nula nos dois grupos, havendo evidências de que as médias das populações de cada grupo seriam diferentes. Reduzindo o nível de significância para 3%, a análise de variância mostraria que haveria evidências de que as populações de onde foram retiradas as amostras A, B e C têm médias iguais, permanecendo inalterada a decisão de rejeitar a hipótese nula do grupo de amostras D, E e F, situação que você poderá comprovar resolvendo os Problemas 3 e 4 deste capítulo.

Problemas

Problema 1

As populações das três amostras registradas na tabela seguinte atendem às premissas da análise da variância. Verifique se as médias das populações são iguais para o nível de significância de 5%.

Amostra 1	Amostra 2	Amostra 3
2,87	3,23	2,25
2,16	3,45	3,13
3,14	2,78	2,44
2,51	3,77	3,27
1,8	2,97	2,81
3,01	3,53	1,36
2,16	3,01	

R: Rejeitar a hipótese nula.

Problema 2

Resolva o Problema 1 com a ferramenta de análise *Anova: Único Fator*.

R: Este problema foi resolvido na planilha **Problemas** da pasta **Capítulo 14**.

Problema 3

Nas três amostras A, B, e C mostradas no início do capítulo, verifique se as médias das populações são iguais, considerando o nível de significância de 3%.

R: *Aceitar a hipótese nula*. Este problema foi resolvido na planilha **Problemas** da pasta **Capítulo 14**.

Problema 4

Nas três amostras D, E, e F mostradas no início do capítulo, verifique se as médias das populações são iguais, considerando o nível de significância de 3%.

R: *Rejeitar a hipótese nula*.

Problema 5

Afirma-se que o número de carros roubados por dia não depende da região da cidade. Para verificar essa afirmação, a cidade foi dividida em quatro zonas e, durante 10 dias, foram registrados os carros roubados nas quatro zonas, conforme registrado na tabela seguinte. Verifique essa afirmação considerando o nível de significância de 5%.

Zona 1	Zona 2	Zona 3	Zona 4
12	12	10	13
15	11	12	15
14	13	14	14
12	18	12	15
15	15	11	17
18	14	13	14
12	13	10	13
14	12	12	14
12	11	13	15
11	10	11	16

R: *Rejeitar a hipótese nula*. Há diferenças entre as quatro zonas.

Problema 6

Para tentar maximizar a quantidade de quilômetros por litro rodados pela frota de veículos da empresa, o gerente de manutenção testou três tipos diferentes de combustíveis em carros da mesma marca. A tabela a seguir registra os quilômetros por litro de dezoito carros com três marcas de combustível diferentes. Verifique se há diferenças entre os combustíveis, considerando o nível de significância 5%.

Combustível 1	Combustível 2	Combustível 3
12,8	12,0	13,1
12,6	12,2	13,3
12,9	12,0	13,0
13,5	11,5	12,8
11,6	11,8	12,6
12,2	12,3	12,9

R: *Rejeitar a hipótese nula*. Há diferenças entre os combustíveis.

Problema 7

A velocidade de tração durante o teste de resistência do cabo de aço foi contestada pelo gerente de produção. Foram realizados dez testes de resistência do cabo para quatro velocidades diferentes, resultados registrados na tabela seguinte. Verifique se a velocidade de tração gera diferenças nos resultados do teste de resistência do cabo de aço, considerando o nível de significância 5%.

Velocidade 1%	Velocidade 10%	Velocidade 20%	Velocidade 30%
363,9	366,0	366,1	363,7
365,8	366,1	364,1	363,9
366,2	365,0	364,1	366,8
365,1	364,8	365,5	365,9
365,0	363,5	364,5	365,3
363,4	365,6	365,2	365,7
363,3	365,3	364,3	365,6
366,3	365,1	364,4	366,2
365,4	366,0	366,5	365,0
366,4	364,0	364,7	364,4

R: Aceitar a hipótese nula. A velocidade de tração não afeta o resultado do teste de resistência do cabo de aço.

Anova com dois fatores

Na primeira parte deste capítulo, foi apresentada a análise da variância com um fator, ou *Anova* com um fator, na qual é avaliado apenas um fator de interesse ou que influi na variável dependente. Nesta parte, serão avaliados dois fatores de interesse que influem em uma variável dependente, seja de forma isolada ou simultaneamente. Na análise da variância com dois fatores, por exemplo, os fatores *A* e *B* podem influir na variável dependente de forma isolada, denominados efeitos principais, e de forma combinada, efeito de uma combinação específica dos fatores *A* e *B*. Cada fator tem um número de níveis; por exemplo, o fator *A* pode ter dois tipos diferentes de processos, e o fator *B* três dosagens diferentes de um determinado aditivo para acelerar a secagem. Não será realizada uma apresentação detalhada como a da primeira parte, mas serão destacadas as premissas e como utilizar e obter conclusões dos resultados da ferramenta *Anova: fator duplo com repetição*. De maneira formal, o teste de hipóteses para dois fatores *A* e *B* tem três hipóteses nulas:

H_0 : Não há efeito principal do fator *A*.

H_0 : Não há efeito principal do fator *B*.

H_0 : Não há combinação de efeitos.

H_1 : Há efeito em cada um dos três casos.

EXEMPLO 14.4

A empresa de porte médio manufatura autopeças para o mercado de reposição está tentando reduzir o tempo de produção de cada peça. O gerente de pesquisas testou dois processos diferentes e três dosagens de um novo aditivo químico para acelerar a secagem. Os tempos obtidos estão apresentados na tabela a seguir e registrados na planilha **Exemplo 14.4**, incluída na pasta **Capítulo 14**. Utilizando a ferramenta de análise do Excel, realize uma análise da variância, considerando o nível de significância de 5%.

	A	B	C	D
1	Anova com dois fatores			
2				
3		Aditivo	Processo 1	Processo 2
4		Dosagem 1	2,5	2,9
5			2,8	2,6
6			2,9	2,8
7			2,7	2,3
8			2,7	2,9
9		Dosagem 2	2,9	2,8
10			2,7	2,9
11			2,8	2,8
12			2,6	2,9
13			3	2,6
14		Dosagem 3	2,6	2,4
15			2,7	2,7
16			2,8	2,7
17			2,5	2,1
18			2,9	2,5
19				

Solução. Na planilha anterior estão definidos dois fatores de análise, o fator *Aditivo* com três níveis de dosagem e o fator *Processo* com dois tipos. Esses dois fatores formam seis grupos de resultados com cinco observações cada um e identificados nas duas colunas, denominadas Processo 1 e Processo 2, e nos três grupos de cinco linhas cada um, denominado Dosagem 1, Dosagem 2 e Dosagem 3. Nesse tipo de análise da variância, os grupos devem ter o mesmo número de observações ou repetições, nesse caso cinco. O teste de hipóteses para o fator *Aditivo* e o fator *Processo* tem três hipóteses nulas:

H_0 : Não há efeito principal do fator *Aditivo*.

H_0 : Não há efeito principal do fator *Processo*.

H_0 : Não há combinação dos efeitos *Aditivo* e *Processo*.

H_1 : Há efeito em cada um dos três casos.

A ferramenta de análise *Anova: fator duplo com repetição* constrói a tabela *Anova* para o teste de hipóteses com dois fatores. Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 deste livro. Escolhendo a ferramenta *Anova: fator duplo com repetição*, e depois de clicar no botão **OK**, será exibida a caixa de diálogo com o mesmo nome mostrado na Figura 14.3, depois de ter selecionado as opções do exemplo. As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo dessa ferramenta são:

- **Intervalo de entrada:** Nesse caso, foi informado o intervalo B3:D18. Nesse intervalo, devem estar incluídos os nomes e níveis dos dois fatores.
- **Linhas por amostra:** Deve-se informar o número de linhas contidas em cada amostra, tendo presente que cada amostra deve conter o mesmo número de linhas, pois cada linha representa uma replicação dos dados. Nesse exemplo, foi informado o valor 5.
- **Alfa:** Deve-se informar o nível de significância *alfa* do teste de hipóteses, nesse caso 0,05.

Anova: fator duplo com repetição

Entrada

Intervalo de entrada:

Linhas por amostra:

Alfa:

Opções de saída

☒ Intervalo de saída:

☐ Nova planilha:

☐ Nova pasta de trabalho

Buttons: OK, Cancelar, Ajuda

No quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados, neste caso F1.

Depois de completar as informações e clicar em **OK**, na caixa de diálogo, a ferramenta de análise *Anova: fator duplo com repetição* apresentará dois grupos de resultados, na primeira parte a tabela *RESUMO* e na segunda parte a tabela *ANOVA*, mostrada a seguir, deixando a primeira para depois.

	E	F	G	H	I	J	K	L
29	ANOVA							
30		Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
31		Amostra	0,222	2	0,111	3,11215	0,062843	3,402832
32		Colunas	0,048	1	0,048	1,345794	0,257425	4,259675
33		Interações	0,074	2	0,037	1,037383	0,369734	3,402832
34		Dentro	0,856	24	0,035667			
35								
36		Total	1,2	29				
37								

A seguir, mostramos como analisar os resultados da tabela ANOVA para realizar o teste de hipóteses:

- **Teste da combinação de fatores.** O ponto de partida é a análise dos resultados da linha *Interações*, que é o resultado da combinação dos dois fatores. Como o *p-value* 0,3697 (ou 36,97%) registrado na célula K33 é maior do que o nível de significância 5%, a hipótese nula deve ser aceita. A aceitação da hipótese nula indica que a combinação dos fatores *Aditivo* e *Processo* não é significativa ou, de outra maneira, não há evidência suficiente de que a combinação de efeitos provocada pelos dois fatores influencie o tempo de produção. Em vez de utilizar o *p-value*, pode-se comparar o *F* observado 1,037, registrado na célula J33, com o *F* crítico 3,403, registrado na célula L33, que também mostra a aceitação da hipótese nula.
 - Se o resultado do teste for significativo, pois a hipótese nula seria rejeitada, então o procedimento de análise deverá continuar se aprofundando com os efeitos das seis possíveis combinações dos dois fatores. Deve-se ter presente que se o efeito da combinação de fatores é significativa e qualquer efeito principal deverá ser tratado com cautela, assunto que não será tratado no livro.

A seguir passamos para a análise dos efeitos dos fatores de forma isolada.

- **Teste do fator *Aditivo*.** O título *Amostra* registrado na tabela ANOVA, linha 31 da planilha, refere-se aos resultados do Fator *Aditivo*. Como o *p-value* 0,0628 (ou 6,28%), registrado na célula K31, é maior do que o nível de significância 5%, a hipótese nula deve ser aceita. A aceitação da hipótese nula indica que o fator *Aditivo* não influencia o tempo de produção das autopeças ou, de outra maneira, não há evidência suficiente de que o fator *Aditivo* influencia o tempo de produção da autopeça. Em vez de utilizar o *p-value*, pode-se comparar o *F* observado 3,11, registrado na célula J31, com o *F* crítico 3,40, registrado na célula L31, que também mostra a aceitação da hipótese nula.
- **Teste do fator *Processo*.** O título *Colunas* registrado na tabela ANOVA, linha 32 da planilha, refere-se aos resultados do Fator *Processo*. Nesse caso, também, a hipótese nula deve ser aceita. Deixamos que você realize as análises comparativas do *p-value* com o nível de significância, e do *F* observado com o *F* crítico.
 - Observe que se o nível de significância for maior do que 6,3%, o fator *Aditivo* passaria a ter influência no tempo de produção das autopeças, enquanto o fator *Processo* continuaria sem ter influência nesse tempo.

Outros resultados registrados na tabela ANOVA são:

- **Dentro.** Na célula G34, é registrado o resultado da soma dos quadrados dos desvios dos dados de cada um dos grupos com relação à sua própria média.
- **Total.** Na célula G35, é registrado o resultado da soma dos quadrados dos desvios de todos os dados com relação à grande média. Também é o resultado da soma do intervalo G31:G34 da planilha.
- **gl.** Essa coluna registra os graus de liberdade de cada grupo de resultados da coluna *SQ* da tabela, para sua linha correspondente.
- **MQ.** Cada linha dessa coluna registra o resultado da divisão da soma dos quadrados dos desvios da coluna *SQ* pelo número de graus de liberdade correspondente da coluna *gl*. Por exemplo, o resultado 0,111 (registrado na célula I31) é o resultado da divisão de 0,222 (célula G31) por 2 (célula H31).
 - Os resultados dessa coluna são utilizados para obter o *F* observado da coluna *F* da tabela ANOVA. Dividindo qualquer um dos três valores do intervalo I31:I33 pelo valor registrado em I34, obtém-se o valor correspondente no intervalo J31:J33, procedimento mostrado no intervalo I38:L41 da planilha **Exemplo 14.4**.

A ferramenta de análise *Anova: fator duplo com repetição* apresenta dois grupos de resultados, na primeira parte a tabela *RESUMO* e na segunda parte a tabela *ANOVA*, que acabou de ser apresentada. Agora serão analisados os resultados registrados na tabela *RESUMO*.

- A tabela tem quatro blocos, três deles identificados pelas dosagens do fator *Aditivo* e o último como *Total*.
- Os três primeiros blocos têm quatro colunas e quatro linhas. Na primeira coluna, estão registrados os títulos *Contagem*, *Soma*, *Média* e *Variância*. Na segunda coluna, estão registradas as medidas estatísticas correspondentes do Processo 1, na terceira coluna do Processo 2 e na última coluna, as correspondentes ao total do bloco. Esses resultados foram reproduzidos no intervalo K5:M8, utilizando as funções do Excel.
- O último bloco tem três colunas e quatro linhas. Na primeira coluna, estão registrados os títulos *Contagem*, *Soma*, *Média* e *Variância*. Na segunda coluna, estão registradas as medidas estatísticas correspondentes do Processo 1 e, na terceira coluna, do Processo 2. Esses resultados foram reproduzidos no intervalo K23:L26, utilizando as funções do Excel.

	E	F	G	H	I
1		Anova: fator duplo com repetição			
2					
3		RESUMO	Processo 1	Processo 2	Total
4		<i>Dosagem 1</i>			
5		Contagem	5	5	10
6		Soma	13,6	13,5	27,1
7		Média	2,72	2,7	2,71
8		Variância	0,022	0,065	0,0387778
9					
10		<i>Dosagem 2</i>			
11		Contagem	5	5	10
12		Soma	14	14	28
13		Média	2,8	2,8	2,8
14		Variância	0,025	0,015	0,0177778
15					
16		<i>Dosagem 3</i>			
17		Contagem	5	5	10
18		Soma	13,5	12,4	25,9
19		Média	2,7	2,48	2,59
20		Variância	0,025	0,062	0,0521111
21					
22		<i>Total</i>			
23		Contagem	15	15	
24		Soma	41,1	39,9	
25		Média	2,74	2,66	
26		Variância	0,0225714	0,0597143	
27					

O Excel também dispõe da ferramenta de análise *Anova: fator duplo sem repetição*, que utiliza somente uma observação por cada bloco de combinação de fatores, sem repetições de observações. Esse teste é utilizado nos casos em que não é possível repetir as experiências ou, sendo possível, se seu custo é elevado comparado com o valor dos resultados obtidos. A utilização da ferramenta de análise *Anova: fator duplo sem repetição* não será apresentada no livro; entretanto, seu procedimento é parecido com o anterior, com a exceção de não incluir o resultado dos efeitos da combinação dos fatores.

Capítulo 15

REGRESSÃO LINEAR

O coeficiente de correlação não mede a relação causa-efeito entre duas variáveis, apesar de essa relação poder estar presente. Por exemplo, uma correlação fortemente positiva entre as variáveis X e Y não autoriza afirmar que variações da variável X provocam variações na variável Y , ou vice-versa. O coeficiente de correlação sozinho não identifica a relação causa-efeito entre as duas variáveis; entretanto, em uma regressão linear, a relação causa-efeito deve ser definida no início da análise. Este capítulo se inicia com a apresentação da relação linear simples entre duas amostras ou variáveis aleatórias e termina com a apresentação da relação de dependência linear múltipla entre três ou mais amostras ou variáveis aleatórias. Na regressão linear simples, será deduzida e analisada a reta que melhor explica essa relação, tendo previamente definido a variável independente e a variável dependente. A regressão linear múltipla será apresentada através de um exemplo resolvido com a ferramenta de análise *Regressão*.

Todos os dias, a *mídia* se encarrega de informar resultados de análises e pesquisas do tipo: o valor da empresa depende do lucro futuro, a taxa de juros depende da inflação, o salário depende da escolaridade do trabalhador etc. O objetivo da análise de regressão é encontrar uma função linear que permita:

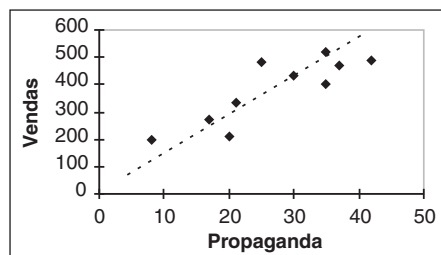
- Descrever e compreender a relação entre uma variável dependente e uma ou mais variáveis independentes.
- Projetar ou estimar uma variável em função de uma ou mais variáveis independentes; por exemplo, as vendas para diferentes valores de investimento em propaganda, a demanda em função do preço unitário e do investimento em propaganda etc.

EXEMPLO 15.1

O objetivo do diretor de vendas de uma rede de varejo é analisar a relação entre o investimento realizado em propaganda e as vendas das lojas da rede, para realizar projeções de vendas de futuros investimentos em propaganda. A tabela a seguir registra uma amostra representativa extraída dos registros históricos das lojas de tamanho equivalente, com os valores de Propaganda e Vendas em milhões. Analise a possibilidade de definir um modelo que represente a relação entre as duas variáveis ou amostras.

Propaganda	30	21	35	42	37	20	8	17	35	25
Vendas	430	335	520	490	470	210	195	270	400	480

Solução. Para analisar a relação entre as duas variáveis na planilha **Exemplo 15.1**, incluída na pasta **Capítulo 15**, foi construído o gráfico de dispersão das vendas anuais em função do investimento anual em propaganda. Nesse gráfico, pode-se ver que, nos últimos dez anos, o aumento de investimento em propaganda gerou aumento das vendas, e vice-versa.



O gráfico de dispersão apresentado mostra que as vendas e o investimento em propaganda estão correlacionados de forma positiva, com um coeficiente de correlação próximo de +1. Uma reta como a linha tracejada nesse gráfico de dispersão poderá ser utilizada para realizar projeções das vendas futuras em função do investimento em propaganda.

A tendência de crescimento positivo dos dados do Exemplo 15.1 sugere um modelo representado por uma linha reta, como a desenhada em linha tracejada no gráfico de dispersão do exemplo, lembrando que:

- A linha tracejada foi ajustada tentando equilibrar os pontos acima da reta com os pontos abaixo dela.
- Essa reta é uma das muitas retas possíveis que poderiam ser ajustadas.

Modelo ajuste de uma reta

O ajuste de uma reta é um modelo linear que relaciona a variável dependente y e a variável independente x por meio da equação de uma reta do tipo $y = a + bx$. É importante lembrar que, da mesma forma como a média resume uma variável aleatória, a reta de regressão resume a relação linear entre duas variáveis aleatórias e, conseqüentemente, da mesma forma como a média varia entre amostras do mesmo tamanho extraídas da mesma população, as retas também variarão entre amostras da mesma população. Continuando com o Exemplo 15.1, o objetivo é ajustar uma reta a partir dos valores das amostras retiradas da população, considerando que o investimento em propaganda é a variável independente x , e as vendas anuais, a variável dependente y . Uma primeira forma de fazer isso é ajustar manualmente essa reta, tentando equilibrar os pontos acima e abaixo dela, como foi feito no gráfico de dispersão do Exemplo 15.1. Como esse procedimento permite o ajuste de diversas retas, é necessário estabelecer um objetivo de eficiência de ajuste possível de medir, como mostrado a seguir.

- Uma primeira forma é ajustar uma reta horizontal de valor igual à média \bar{y} dos valores da variável dependente y , que é uma reta de regressão com $b=0$. Esse critério não necessita de regressão; entretanto, será uma referência útil para medir o grau de explicação¹ da reta de regressão.
- Outra forma é ajustar uma reta que divida os pontos observados de forma que a soma dos desvios seja nula. Contudo, como há muitas retas que cumprem com essa condição, esse critério não poderá ser utilizado.
- Lembrando a definição de variância, outra forma é ajustar uma reta de forma que se minimize a soma dos quadrados dos desvios.

¹ Mais adiante será utilizada para definir o *coeficiente de determinação*.

O modelo construído na planilha **Modelo Ajuste da reta**, incluída na pasta **Capítulo 15**, permite realizar o ajuste da reta de forma manual utilizando os dados do Exemplo 15.1. O gráfico da Figura 15.1 mostra a reta horizontal (declividade zero) com intercepto 380, que é a média da variável dependente y , ou Vendas. Ao mesmo tempo, a célula H16 da planilha registra o valor 129.950, que representa a soma dos quadrados dos desvios dos dez valores da amostra y com relação à reta de regressão. Clicando no botão giratório do grupo **Intercepto**, a reta se deslocará de forma paralela ao eixo de abscissas, subindo ou descendo. Observe que, se aumentar ou diminuir o valor do intercepto ao redor de 380, mantendo a declividade igual a zero, a soma dos quadrados dos desvios sempre aumentará. Por quê? Porque o valor de intercepto é a própria média da amostra y , e esse valor é sempre um mínimo, pela segunda propriedade da média apresentada no Capítulo 3.

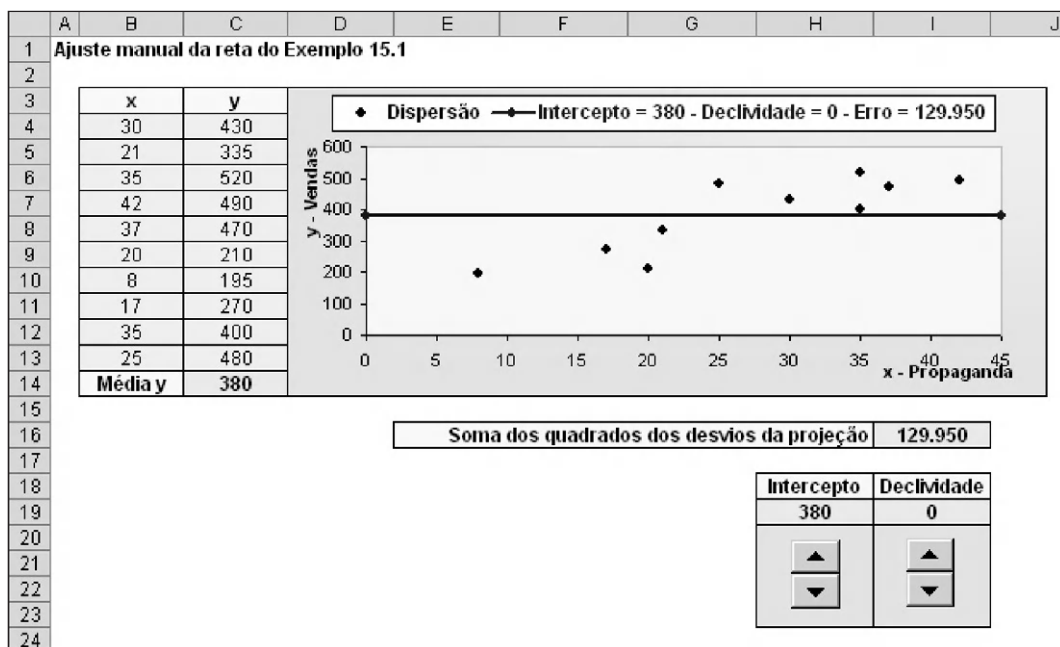


FIGURA 15.1
Modelo Ajuste da
reta, Exemplo 15.1.

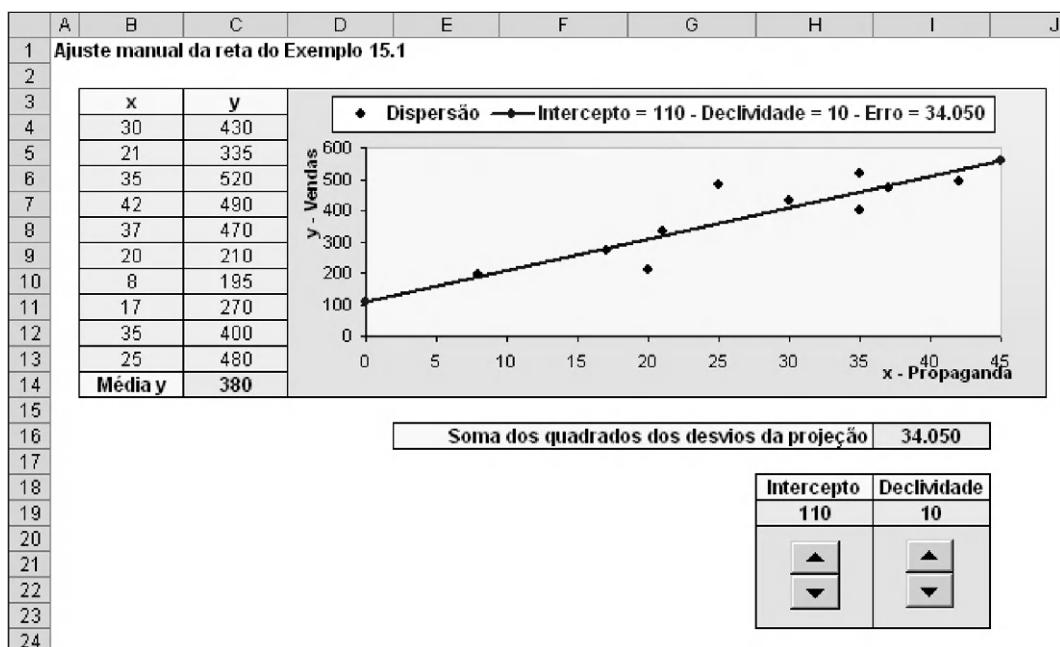


FIGURA 15.2
Melhor ajuste manual
da reta de regressão
do Exemplo 15.1.

O modelo mostrado na Figura 15.1 foi preparado para o Exemplo 15.1, com o objetivo de mostrar o que ocorre com a tentativa de ajuste manual com pouca precisão da reta de regressão. Clicando nos dois controles giratórios, é possível tentar outros valores de intercepto combinado com valores diferentes de declividade. Por exemplo, fixando um valor menor de declividade, será possível ajustar o valor de intercepto que minimiza o resultado da célula H16, ou conseguir a soma mínima dos quadrados dos desvios dos dez valores da amostra y , com relação à reta de regressão. Dentre todos esses possíveis valores mínimos, há de haver um que seja o menor de todos. A Figura 15.2 mostra o mínimo encontrado manualmente, intercepto 110, declividade 10 e erro 34.050.

O procedimento manual para encontrar a soma mínima dos quadrados dos desvios dos dez valores da amostra y com relação à reta de regressão é bastante trabalhoso com resultado aproximado. É claro que, melhorando a escala dos controles giratórios, será possível ajustar essa aproximação. No entanto, sempre haverá um erro, pois nem todos os pares de valores das duas amostras estarão contidos na reta ajustada, eles estarão distribuídos ao redor dessa reta. Somente se os pares de valores formassem uma reta, o erro seria zero. Todavia, o procedimento manual de ajuste da regressão tem o crédito de visualizar o caminho para estabelecer o critério de ajuste da reta de regressão. O objetivo é encontrar os coeficientes a e b da reta de regressão que minimizam a soma dos quadrados dos desvios dos valores da amostra y com relação aos correspondentes valores \hat{y} da reta de regressão.

Linha de tendência do Excel

Uma forma prática de ajustar e obter a equação de uma reta de regressão é usando o comando *Linha de tendência*² do Excel. Neste momento, será apresentada uma parte do comando linha de tendência, pois, no próximo capítulo esse comando será apresentado de forma mais completa. Para construir a linha de tendência, deve-se registrar em uma planilha Excel a tabela com os dados das duas amostras e o gráfico de dispersão construído, como foi feito na planilha **Linha de tendência** da pasta **Capítulo 15**. O procedimento é o seguinte:

- Selecione a trajetória dos pontos do gráfico de dispersão, clicando uma vez em um dos pontos do gráfico. Os pontos do gráfico mudarão de cor.
- Depois de escolher **Adicionar linha de tendência** no menu **Gráfico**, será exibida a caixa de diálogo **Adicionar linha de tendência** com duas guias, **Tipo** e **Opções**.
- No quadro **Tendência/tipo de regressão** de **Tipo**, selecione *Linear*, como mostra a Figura 15.3.
- Na folha **Opções** da caixa de diálogo, selecione **Exibir equação no gráfico**, como mostra a Figura 15.4. Depois de clicar no botão **OK**, o Excel construirá a reta ajustada e registrará no mesmo quadro sua equação. Esses valores estão registrados em um bloco que pode ser mudado de posição, como foi feito na Figura 15.5.

Com a equação $\hat{y} = 9,7381x + 117,07$ obtida com o comando linha de tendência, será possível representar o comportamento das vendas em função do investimento em propaganda com um modelo linear e realizar projeções. Mas qual o critério utilizado pelo comando linha de tendência para obter os coeficientes dessa reta de regressão? Você já deve ter deduzido a resposta a essa questão: os coeficientes a e b da reta de regressão minimizam a soma dos quadrados dos desvios dos valores da amostra y com relação aos valores correspondentes \hat{y} da reta de regressão. Outra questão: quão bem a reta representa o fenômeno amostrado se alguns dos pontos do gráfico de dispersão não estão contidos na reta de regressão? Essa questão será respondida mais adiante neste capítulo.

² Em inglês, o comando Linha de Tendência é *Trendline*.

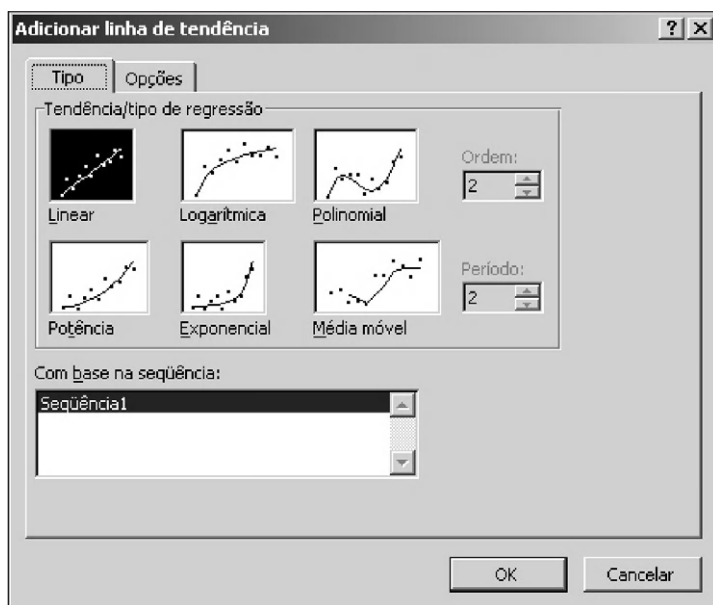


FIGURA 15.3 Caixa de diálogo Adicionar linha de tendência.

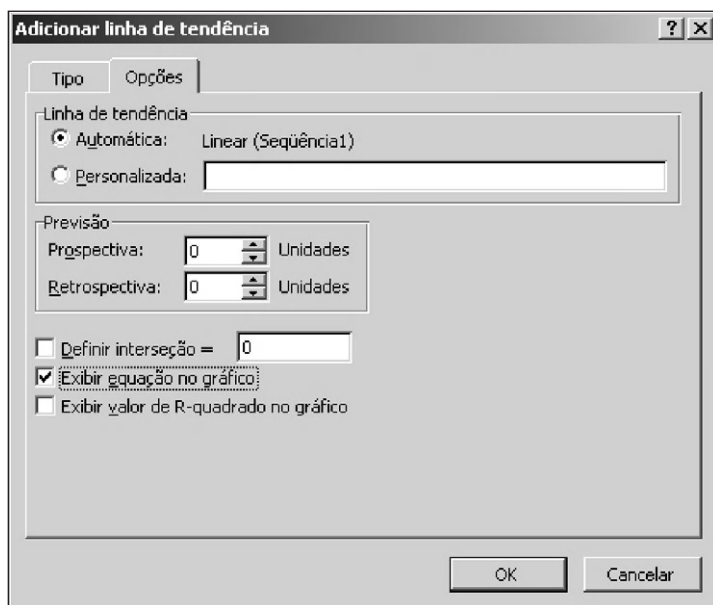


FIGURA 15.4 Folha Opções da linha de tendência.

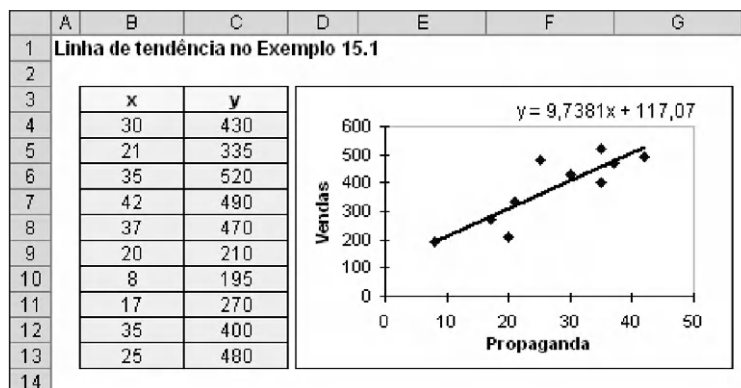


FIGURA 15.5 Reta ajustada e sua equação, Exemplo 15.1.

Coeficientes de regressão

No ajuste da reta de regressão, o procedimento manual utilizado com o *modelo* da Figura 15.1 nos aproximou do procedimento de ajuste e do melhor critério para encontrar os coeficientes de regressão a e b . Depois, o comando linha de tendência forneceu diretamente a equação da reta de regressão. Agora, mostraremos como os valores dos coeficientes de regressão a e b podem ser obtidos utilizando os dados das variáveis.

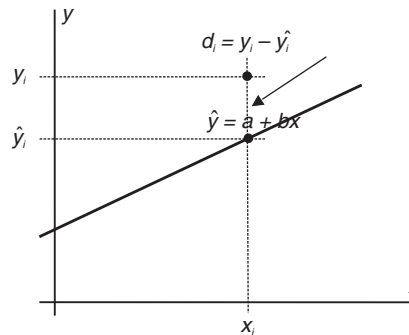
A reta de regressão é representada pela equação $\hat{y} = a + bx$, sendo \hat{y} a variável dependente e x a variável independente. Os coeficientes a e b são os coeficientes de regressão com o seguinte significado:

- O coeficiente b é a declividade da reta e define o aumento ou a diminuição da variável y por unidade de variação da variável x .
- A constante a é o intercepto y , sendo igual³ ao valor de \hat{y} para $x=0$.

No modelo matemático da reta ajustada, verifica-se que:

- Para um único valor x_i poderão ocorrer um ou mais valores de y_i amostrados. Por exemplo, no gráfico de dispersão do Exemplo 15.1, para $x=35$, há dois valores da variável dependente $y=400$ e $y=520$.
- Há apenas um único \hat{y}_i projetado para cada valor de x_i , porém há observações que não são pontos da reta.
- Para cada valor de x_i , há uma diferença entre o valor amostrado y_i e o valor projetado \hat{y}_i . Essa diferença é denominada desvio d_i , como registrado na Figura 15.6.

FIGURA 15.6 Desvio do valor projetado.



O gráfico da Figura 15.6 mostra que, em geral, para cada valor de x_i , o valor observado e o valor projetado serão diferentes. Isto é, ocorre um *desvio* d_i medido pela diferença entre o valor observado e o valor projetado:

$$d_i = y_i - \hat{y}_i$$

O desvio $y_i - \hat{y}_i$ também é denominado *resíduo*. Incluindo a equação da reta de regressão na fórmula do desvio:

$$\begin{aligned} d_i &= y_i - (a + bx_i) \\ d_i &= y_i - a - bx_i \end{aligned}$$

O objetivo é obter os coeficientes a e b da reta $\hat{y} = a + bx$, a partir dos n pares de valores das amostras. Que critério deverá ser aplicado para obter os coeficientes a e b ? Como foi visto durante o ajuste ma-

³ Em alguns casos o valor de $x=0$ não tem significado prático.

nual da reta, quanto menor for a soma de todos os desvios, melhor será o ajuste da reta, ou o *poder de explicação* do modelo. O procedimento utilizado é denominado *método dos quadrados mínimos*, que parte da soma dos quadrados dos desvios:

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

O objetivo é determinar os coeficientes a e b da reta de regressão que minimizam a soma dos quadrados dos desvios. De outra maneira, encontrar a e b de forma que a soma dos quadrados dos desvios seja um *mínimo*. De forma matemática:

$$\text{minimizar} \Rightarrow \sum_{i=1}^n (y_i - a - bx_i)^2$$

Uma forma prática de encontrar esse mínimo é utilizar o comando Solver do Excel, como apresentado no Apêndice 1 deste capítulo, mostrando que será necessário preparar uma planilha para utilizar o Solver. Entretanto, uma maneira elegante é encontrar o mínimo relativo da função aplicando os conceitos de cálculo diferencial para obter as fórmulas dos coeficientes de regressão a e b , como é realizado no Apêndice 2 deste capítulo, e cujas expressões repetimos a seguir.

$$\hat{y} = a + bx \text{ sendo, } \begin{cases} a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n} \\ b = \frac{n \sum_{i=1}^n x_i \times y_i - \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \end{cases}$$

Se os n pares de valores das duas amostras formassem uma reta,⁴ então a equação da reta ajustada representaria esses n pares de valores. Contudo, nem todos os n pares de valores estarão contidos na reta, eles estarão distribuídos ao redor da reta ajustada. A minimização da soma dos quadrados dos desvios é apenas uma propriedade desejada de ajuste da reta e, portanto, não garante que se tenha a *melhor* reta ajustada. O método de ajuste pelo método dos quadrados mínimos é preferível, pois:

- Obtém as melhores estimativas, pois elas serão não viesadas.
- Onera os desvios maiores, fato desejável que evita grandes desvios.
- Permite realizar testes de significância na equação de regressão.
- A reta de regressão passa pelo ponto formado pelos valores das médias das duas amostras.

EXEMPLO 15.2

Com os dados das amostras do Exemplo 15.1, obtenha a reta de regressão linear.

Solução. Na planilha **Exemplo 15.2**, incluída na pasta **Capítulo 15**, foram construídas as colunas necessárias para calcular os coeficientes de regressão, como mostra a figura seguinte.

⁴ As duas amostras têm coeficiente de correlação +1 ou -1.

	A	B	C	D	E	F	G	H	I
1	Exemplo 15.2								
2									
3		x	y	x.y	x^2			Resultados com fórmulas	
4		30	430	12.900	900			b	9,74
5		21	335	7.035	441			a	117,07
6		35	520	18.200	1.225			Equação	
7		42	490	20.580	1.764			y = 117,07 + 9,74 x	
8		37	470	17.390	1.369			Resultados com funções	
9		20	210	4.200	400			a	117,07
10		8	195	1.560	64			=INTERCEPÇÃO(C4:C13;B4:B13)	
11		17	270	4.590	289			b	9,74
12		35	400	14.000	1.225			=INCLINAÇÃO(C4:C13;B4:B13)	
13		25	480	12.000	625				
14		270	3.800	112.455	8.302				
15									

Utilizando as fórmulas. Esses cálculos são mostrados a seguir, lembrando de calcular b antes do coeficiente a .
Coeficiente b . Substituindo os resultados parciais obtidos na planilha da figura apresentada:

$$b = \frac{10 \times 112.455 - 270 \times 3.800}{10 \times 8.302 - 270^2} = 9,7381$$

Coeficiente a . Substituindo os resultados parciais obtidos na planilha da figura apresentada:

$$a = \frac{3.800 - 9,7381 \times 270}{10} = 117,07$$

Portanto, a equação da reta de regressão procurada é $\hat{y} = 117,07 + 9,74x$.

Utilizando as funções estatísticas do Excel. O valor dos coeficientes a e b também pode ser obtido utilizando as funções estatísticas, respectivamente, INTERCEPÇÃO e INCLINAÇÃO, como mostra a figura apresentada anteriormente. Com essas funções, não será necessário calcular b antes de a .

Coeficiente a . Com a fórmula =INTERCEPÇÃO(C4:C13;B4:B13), registrada na célula H10, foi obtido o coeficiente de regressão a .

• **INTERCEPÇÃO(val_conhecidos_y; val_conhecidos_x)**

A função estatística INTERCEPÇÃO⁵ retorna o coeficiente de regressão a da reta de regressão linear $\hat{y} = a + bx$, considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar essa função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente y , e o argumento *val_conhecidos_x*, aos valores da variável independente x . Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

Coeficiente b . Com a fórmula =INCLINAÇÃO(C4:C13;B4:B13), registrada na célula H12, foi obtido o coeficiente de regressão b .

• **INCLINAÇÃO(val_conhecidos_y; val_conhecidos_x)**

A função estatística INCLINAÇÃO⁶ retorna o coeficiente b da reta de regressão linear $\hat{y} = a + bx$, considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar esta função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente y e o argumento *val_conhecidos_x*, aos valores da variável independente x . Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

Para construir a reta de regressão, deve-se projetar pelo menos dois pontos dessa reta utilizando a equação de regressão que repetimos, $\hat{y} = 117,07 + 9,74x$. A figura a seguir mostra o procedimento realizado a partir da célula K1 da planilha **Exemplo 15.2**.

⁵ Em inglês, a função estatística INTERCEPÇÃO é INTERCEPT.

⁶ Em inglês, a função estatística INCLINAÇÃO é SLOPE.

- | | J | K | L | M | N | O |
|----|--|-----------------|---|---|---|---|
| 1 | Construção da reta de regressão | | | | | |
| 2 | | | | | | |
| 3 | | Projeção | | | | |
| 4 | | 409,21 | | | | |
| 5 | | 321,57 | | | | |
| 6 | | 457,91 | | | | |
| 7 | | 526,07 | | | | |
| 8 | | 477,38 | | | | |
| 9 | | 311,83 | | | | |
| 10 | | 194,98 | | | | |
| 11 | | 282,62 | | | | |
| 12 | | 457,91 | | | | |
| 13 | | 360,52 | | | | |
| 14 | | | | | | |

Gráfico de dispersão com uma linha de regressão linear. O eixo X representa o tempo em horas (0 a 50) e o eixo Y representa o consumo de combustível em litros (0 a 600). Os dados são representados por pontos pretos. A equação da reta de regressão é $y = 117,07 + 9,74 x$.

	A	B	C	D	E	F	G	H	I	J	K																													
1	Exemplo 15.3																																							
2																																								
3	<table border="1"> <thead> <tr> <th>x</th> <th>y</th> </tr> </thead> <tbody> <tr><td>30</td><td>430</td></tr> <tr><td>21</td><td>335</td></tr> <tr><td>35</td><td>520</td></tr> <tr><td>42</td><td>490</td></tr> <tr><td>37</td><td>470</td></tr> <tr><td>20</td><td>210</td></tr> <tr><td>8</td><td>195</td></tr> <tr><td>17</td><td>270</td></tr> <tr><td>35</td><td>400</td></tr> <tr><td>25</td><td>480</td></tr> </tbody> </table>		x	y	30	430	21	335	35	520	42	490	37	470	20	210	8	195	17	270	35	400	25	480	<table border="1"> <thead> <tr> <th colspan="2">Resultados</th> </tr> <tr> <th colspan="2">Com a fórmula</th> </tr> <tr> <th>x</th> <th>y</th> </tr> </thead> <tbody> <tr><td>20</td><td>311,83</td></tr> <tr><td>30</td><td>409,21</td></tr> <tr><td>45</td><td>555,29</td></tr> </tbody> </table>		Resultados		Com a fórmula		x	y	20	311,83	30	409,21	45	555,29	= $\$C\$15+\$C\$16*\text{E6}$	
x	y																																							
30	430																																							
21	335																																							
35	520																																							
42	490																																							
37	470																																							
20	210																																							
8	195																																							
17	270																																							
35	400																																							
25	480																																							
Resultados																																								
Com a fórmula																																								
x	y																																							
20	311,83																																							
30	409,21																																							
45	555,29																																							
8			<table border="1"> <thead> <tr> <th colspan="2">Com a função PREVISÃO</th> </tr> <tr> <th>x</th> <th>y</th> </tr> </thead> <tbody> <tr><td>20</td><td>311,83</td></tr> <tr><td>30</td><td>409,21</td></tr> <tr><td>45</td><td>555,29</td></tr> </tbody> </table>		Com a função PREVISÃO		x	y	20	311,83	30	409,21	45	555,29	=PREVISÃO(E12,\$C\$4:\$C\$13,\$B\$4:\$B\$13)																									
Com a função PREVISÃO																																								
x	y																																							
20	311,83																																							
30	409,21																																							
45	555,29																																							
15	<table border="1"> <tr> <td>a</td> <td>117,07</td> </tr> <tr> <td>b</td> <td>9,74</td> </tr> </table>		a	117,07	b	9,74	<table border="1"> <thead> <tr> <th colspan="2">Com a função TENDÊNCIA</th> </tr> <tr> <th>x</th> <th>y</th> </tr> </thead> <tbody> <tr><td>20</td><td>311,83</td></tr> <tr><td>30</td><td>409,21</td></tr> <tr><td>45</td><td>555,29</td></tr> </tbody> </table>		Com a função TENDÊNCIA		x	y	20	311,83	30	409,21	45	555,29	=TENDÊNCIA(\$C\$4:\$C\$13,\$B\$4:\$B\$13;E18;VERDADEIRO)																					
a	117,07																																							
b	9,74																																							
Com a função TENDÊNCIA																																								
x	y																																							
20	311,83																																							
30	409,21																																							
45	555,29																																							
22			<table border="1"> <thead> <tr> <th colspan="2">Função TENDENCIA, matriz</th> </tr> <tr> <th>x</th> <th>y</th> </tr> </thead> <tbody> <tr><td>20</td><td>311,83</td></tr> <tr><td>30</td><td>409,21</td></tr> <tr><td>45</td><td>555,29</td></tr> </tbody> </table>		Função TENDENCIA, matriz		x	y	20	311,83	30	409,21	45	555,29	{=TENDÊNCIA(C4:C13;B4:B13;E24:E26;VERDADEIRO))																									
Função TENDENCIA, matriz																																								
x	y																																							
20	311,83																																							
30	409,21																																							
45	555,29																																							

No intervalo B3:C13, foram registradas as duas amostras, e as projeções para $x=20$, 30 e 45 foram realizadas utilizando três procedimentos.

Projeção utilizando a equação da reta de regressão. Para realizar as projeções de \hat{y} utilizando a equação da reta $\hat{y} = 117,07 + 9,74x$, é necessário calcular os coeficientes de regressão. No intervalo C15:C16, foram calculados os coeficientes utilizando as funções estatísticas correspondentes, apresentadas no Exemplo 15.2. No intervalo E5:F8, foram realizadas as projeções solicitadas, procedendo como segue. Na célula F6, foi registrada a fórmula $=\$C\$15+\$C\$16*E6$, que, depois, foi copiada até a célula F8.

Projeção utilizando a função estatística PREVISÃO. No intervalo E11:F14, foram realizadas as projeções utilizando a função PREVISÃO. Na célula F12, foi registrada a fórmula $=PREVISÃO(E12;\$C\$4:\$C\$13;\$B\$4:\$B\$13)$, que, depois, foi copiada até a célula F14.

• **PREVISÃO(*x*; *val_conhecidos_y*; *val_conhecidos_x*)**

A função estatística PREVISÃO⁷ retorna o valor projetado \hat{y} para o valor registrado no argumento x , considerando a reta de regressão linear simples $\hat{y} = a + bx$ correspondente aos valores das amostras, informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar esta função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o argumento *val_conhecidos_y* se refere aos valores da variável dependente y e o argumento *val_conhecidos_x*, aos valores da variável independente x . Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

Não querendo registrar os valores das amostras, a projeção pode ser realizada em uma única célula da planilha, registrando, por exemplo, a fórmula:

$=PREVISÃO(20;\{430;335;520;490;470;210;195;270;400;480\};$
 $\{30;21;35;42;37;20;8;17;35;25\})$

que retorna o valor 311,83, resultado correspondente à projeção de $x=20$, como se pode ver na célula E29 da planilha.

Projeção utilizando a função estatística TENDÊNCIA. No intervalo E17:F20, foram realizadas as projeções utilizando a função TENDÊNCIA. A fórmula $=TENDÊNCIA(\$C\$4:\$C\$13;\$B\$4:\$B\$13;E18;VERDADEIRO)$ foi registrada na célula F18 e depois copiada até a célula F20.

• **TENDÊNCIA(*val_conhecidos_y*; *val_conhecidos_x*; *x*; *constante*)**

A função estatística TENDÊNCIA⁸ retorna o valor projetado \hat{y} da reta de regressão linear simples para um único ou um grupo de valores de x informados no argumento x , considerando a reta de regressão linear simples $\hat{y} = a + bx$, correspondente aos valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar esta função, deve-se tomar o cuidado de fornecer os valores na ordem correta, o argumento *val_conhecidos_y* se refere aos valores da variável dependente y e o argumento *val_conhecidos_x*, aos valores da variável independente x . Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números. Se o argumento *constante* for VERDADEIRO ou omitido, a função retornará um único ou um grupo de valores da reta de regressão $\hat{y} = a + bx$. Se for FALSO, a função TENDÊNCIA fornecerá os resultados da reta de regressão $\hat{y} = bx$, considerando $a=0$.

Não querendo registrar os valores das amostras, a projeção pode ser realizada numa única célula da planilha, registrando, por exemplo, a fórmula:

$=TENDÊNCIA(\{430;335;520;490;470;210;195;270;400;480\};$
 $\{30;21;35;42;37;20;8;17;35;25\};20;1)$

que retorna o valor 311,83, resultado correspondente à projeção de $x=20$, como se pode ver na célula E30 da planilha.

Uma das vantagens da função TENDÊNCIA é construir *matrizes* de resultados, como é mostrado no intervalo E23:F26 da planilha **Projeção**. Para trabalhar com registros em forma de *matriz*, deve-se proceder como segue:

- Posicionar o mouse na célula F24 e selecionar o intervalo F24:F26. A seguir, registre a fórmula $=TENDÊNCIA(C4:C13;B4:B13;E24:E26;1)$. O valor 1 do último argumento é equivalente a VERDADEIRO, como 0 a FALSO. Em vez de digitar, você poderia utilizar o assistente de função do Excel.

⁷ Em inglês, a função estatística PREVISÃO é FORECAST.

⁸ Em inglês, a função estatística TENDÊNCIA é TREND.

- Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**; mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, as fórmulas receberam as chaves { }.

Para terminar, a função TENDÊNCIA tem mais aplicações do que as apresentadas nesta parte, por exemplo, se *val_conhecidos_x* for omitido, no seu lugar a função considerará a matriz {1.2.3....}, do mesmo tamanho que *val_conhecidos_y*. Sugerimos que você consulte a *Ajuda* do Excel para conhecer todas as aplicações possíveis da função TENDÊNCIA.

As medidas estatísticas e os coeficientes de regressão

Embora as expressões dos coeficientes de regressão a e b não mostrem que estão sendo utilizadas medidas estatísticas das séries de valores de onde foram obtidos, esses conceitos estão presentes nessas expressões. Se nas expressões de a e b forem realizadas transformações algébricas adequadas, obteremos outra forma de calcular a e b , como mostram as expressões seguintes deduzidas no final do Apêndice 2.

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \frac{\sigma_{xy}}{\sigma_x^2} \end{cases}$$

Lembrando que $\sigma_{xy} = r_{xy}\sigma_x\sigma_y$, o coeficiente b poderá ser calculado com a expressão $b = \frac{r_{xy}\sigma_x\sigma_y}{\sigma_x^2}$.

Prescindindo dos índices do coeficiente de correlação, as expressões dos coeficientes de regressão com coeficiente de correlação r serão:

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = r \frac{\sigma_y}{\sigma_x} \end{cases}$$

Como regra geral, recomenda-se ter presente que:

- O coeficiente b é o resultado da divisão da covariância das variáveis pela variância da variável independente. De outra maneira, o coeficiente b é o resultado da multiplicação do coeficiente de correlação das variáveis pelo resultado da divisão do resultado da multiplicação do desvio padrão da variável dependente pelo desvio padrão da variável independente.
- O coeficiente a é o resultado da subtração do produto do coeficiente b pela média da variável independente da média da variável dependente.

EXEMPLO 15.4

Calcule os coeficientes de regressão do Exemplo 15.1, utilizando as medidas estatísticas das amostras.

Solução. Este exemplo está resolvido na planilha **Exemplo 15.4**, incluída na pasta **Capítulo 15**, como mostra a seguinte figura.

	A	B	C	D	E	F	G
1	Exemplo 15.4						
2							
3		x	y		Medidas estatísticas		
4		30	430			x	y
5		21	335		Média	27	380
6		35	520		DP	10,60	120,16
7		42	490		r	0,86	
8		37	470				
9		20	210		Projeção y=f(x)		
10		8	195		b	9,738	=F7*G6/F6
11		17	270		a	117,070	=G5-F10*F5
12		35	400				
13		25	480		Projeção x=f(y)		
14					b	0,076	=F7*F6/G6
15					a	-1,818	=F5-F14*G5
16							

Uma vantagem adicional desse procedimento de cálculo é a possibilidade de calcular as duas possíveis retas de regressão linear com as mesmas medidas estatísticas, permutando apenas as posições das variáveis. Por exemplo, se y for a variável independente e x a variável dependente, os coeficientes da reta de regressão $\hat{x} = f(y)$ serão calculados com as fórmulas:

$$\begin{cases} a = \bar{x} - b\bar{y} \\ b = \frac{\sigma_{xy}}{\sigma_y^2} = r \frac{\sigma_x}{\sigma_y} \end{cases}$$

Observe que os coeficientes da reta de regressão $\hat{x} = f(y)$, com as amostras do Exemplo 15.1, são $a = -1,818$ e $b = 0,076$, resultados obtidos no intervalo F14:F15 da planilha Exemplo 15.4.

Medidas de variação

Nem todos os valores das amostras estão contidos na reta de regressão e quanto mais afastados estiverem, pior a reta representará a relação entre as amostras. A reta obtida pelo método dos quadrados mínimos é um resumo útil da tendência entre as variáveis, pois não explica perfeitamente os dados. Quão útil é a reta de regressão obtida pelo procedimento apresentado? Para responder a essa pergunta, primeiro será analisada a característica dos desvios. Vamos supor que escolhemos como modelo de regressão a reta de regressão horizontal $\hat{y} = \bar{y}$, a equação que representa a média da variável dependente y . Nesse caso, o coeficiente b da reta de regressão é igual a zero e, conseqüentemente, o coeficiente de correlação também é nulo. Embora a reta da média pouco explique, ela é um ponto importante de partida para medir variações. Analisando a reta de regressão com os coeficientes a e b , pode-se ver que a maioria dos valores das amostras está dispersa ao redor da reta, como mostra a Figura 15.7 para um par de valores fora da reta.

Na Figura 15.7 definimos:⁹

- *Variação total* é o resultado da soma dos quadrados dos desvios dos valores y com relação à média

$$\bar{y} : SST = \sum_{i=1}^n (y_i - \bar{y})^2.$$

⁹ Em inglês, SST, SSE e SSR são, respectivamente, *Total Sum of Squares*, *Regression Sum of Squares* e *Error Sum of Squares*.

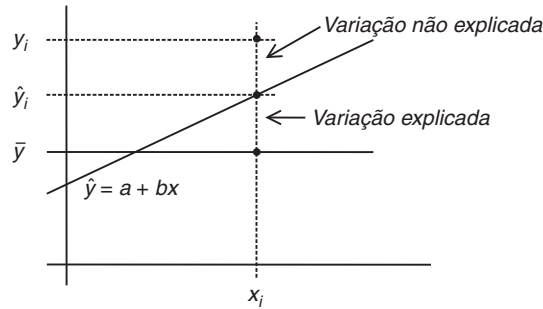


FIGURA 15.7 Variação explicada e variação não explicada.

- *Variação explicada* é o resultado da soma dos quadrados dos desvios dos valores estimados \hat{y} com relação à média \bar{y} : $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.
- *Variação não explicada* é o resultado da soma dos quadrados dos desvios de y com relação aos valores projetados \bar{y} : $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.

A Figura 15.7 mostra que a variação total é o resultado da soma da variação não explicada, mais a variação explicada. Demonstra-se que:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Coeficiente de determinação

O *coeficiente de determinação* r^2 é definido como a relação que mede a proporção da variação total da variável dependente, que é explicada pela variação da variável independente.

$$r^2 = \frac{\text{Variação explicada}}{\text{Variação total}}$$

Substituindo as expressões matemáticas na expressão anterior, temos:

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Essa expressão mostra que o coeficiente de determinação r^2 é sempre um número positivo entre zero e um. Da própria fórmula, pode-se deduzir que, quanto maior for r^2 , melhor será o poder de explicação da reta de regressão.

EXEMPLO 15.5

Calcule o coeficiente de determinação do Exemplo 15.1.

Solução. Na planilha **Coeficiente de Determinação**, incluída na pasta **Capítulo 15**, foi calculado o coeficiente de determinação do Exemplo 15.1, como mostra a figura seguinte. Os resultados parciais são os seguintes:

- *Variação total*, $SST = \sum_{i=1}^n (y_i - \bar{y})^2 = 129.950$.
- *Variação explicada*, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 95.969,39$.
- *Variação não explicada*, $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 33.980,61$

	A	B	C	D	E	F	G	H
1	Cálculo do Coeficiente de Determinação							
2								
3								
4		x	y					
5		30	430					
6		21	335					
7		35	520					
8		42	490					
9		37	470					
10		20	210					
11		8	195					
12		17	270					
13		35	400					
14		25	480					
15		Média y	380					
16		b	9,74					
17		a	117,07					
18								
19								
20								

Variação			
Projeção	Explicada	Não explicada	Total
409,21	853,48	432,04	2.500
321,57	3.413,93	180,33	2.025
457,91	6.069,21	3.855,77	19.600
526,07	21.337,07	1.301,20	12.100
477,38	9.483,14	54,49	8.100
311,83	4.646,74	10.369,96	28.900
194,98	34.234,14	0,00	34.225
282,62	9.483,14	159,23	12.100
457,91	6.069,21	3.353,01	400
360,52	379,33	14.274,58	10.000
Soma	95.969,39	33.980,61	129.950

Coeficiente de Determinação	
Com a fórmula	0,7385
Com a função RQUAD	0,7385

Substituindo na expressão do coeficiente de determinação, da forma como foi realizado na célula H18 da na planilha **Coeficiente de Determinação**:

$$r^2 = \frac{95.969,39}{129.950} = 0,7385$$

Em vez de realizar todo esse procedimento de cálculo, o coeficiente de determinação pode ser calculado com a função estatística RQUAD do Excel, como foi feito na planilha **Coeficiente de Determinação**, registrando na célula H19 a fórmula =RQUAD(C5:C14;B5:B14).

• RQUAD(val_conhecidos_y; val_conhecidos_x)

A função estatística RQUAD¹⁰ retorna o coeficiente de determinação r^2 da reta de regressão $\hat{y} = a + bx$, considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar a função RQUAD, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente *y* e o argumento *val_conhecidos_x* aos valores da variável independente *x*. Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

Não querendo registrar os valores das amostras, o cálculo do coeficiente de determinação pode ser realizado em uma única célula da planilha, registrando, por exemplo, a fórmula:

=RQUAD({430;335;520;490;470;210;195;270;400;480};
{30;21;35;42;37;20;8;17;35;25})

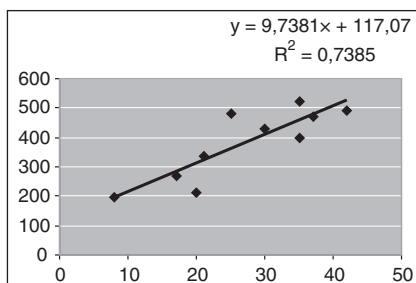
na célula H21.

Utilizando o comando *Linha de tendência* do Excel, também é possível obter o valor do coeficiente de determinação dentro do gráfico de dispersão. Na planilha **Coeficiente de Determinação**, foi construído o gráfico de dispersão a partir da coluna J. O procedimento é o mesmo apresentado no início do capítulo.

- Selecione a trajetória dos pontos do gráfico de dispersão clicando uma vez em um dos pontos do gráfico. Os pontos do gráfico mudarão de cor.

¹⁰ Em inglês, a função estatística RQUAD é RSQ.

- Depois de escolher **Adicionar linha de tendência** no menu **Gráfico**, será exibida a caixa de diálogo **Adicionar linha de tendência** com duas folhas, **Tipo** e **Opções**.
- No quadro **Tendência/tipo de regressão** da folha **Tipo**, selecione *Linear*.
- Na folha **Opções** da caixa de diálogo, selecione **Exibir equação no gráfico** e **Exibir valor de R-quadrado no gráfico**.
- Depois de clicar no botão **OK**, o Excel construirá a reta ajustada e registrará no mesmo quadro sua equação e o R^2 , como mostra a figura seguinte, depois de realizar ajustes de formatação.



O coeficiente de determinação r^2 , também denominado *r-quadrado*, é sempre um número positivo dentro do intervalo (0; 1) e deve ser interpretado como a proporção da variação total da variável dependente y , que é explicada pela variação da variável independente x . Observe que o coeficiente de correlação mede as variações dos dados da amostra y com relação aos valores projetados da reta, sempre na direção de y . No caso do Exemplo 15.5, pode-se dizer que 73,85% das variações das vendas podem ser explicadas pela variabilidade do investimento em propaganda, ficando 26,15% sem explicação.

Embora na determinação do coeficiente de correlação não seja necessário separar as variáveis entre independente e dependente, há uma relação importante entre correlação e regressão. Uma delas é a declividade da reta de regressão, que é função do coeficiente de correlação. Demonstra-se, também, que o coeficiente de determinação é igual ao quadrado do coeficiente de correlação, e vice-versa, $r^2 = (r)^2$. Partindo do coeficiente de correlação $r=0,859366$, obtido na planilha do Exemplo 15.4, temos o valor do coeficiente de determinação $r^2 = (0,859366)^2 = 0,7385$ que o mesmo valor já determinado. O coeficiente de correlação é mais indicado para medir a força da relação linear entre as variáveis, e o coeficiente de determinação é mais apropriado para medir a explicação da reta de regressão. Dessa maneira, para apreciar o ajuste de uma reta, é melhor utilizar o coeficiente de determinação que mede o sucesso da regressão em explicar y .

O coeficiente de correlação também pode ser calculado a partir do coeficiente de determinação, pois $\sqrt{r^2} = \pm r$. Contudo, como o coeficiente de determinação é sempre positivo, o sinal de r será o mesmo que o sinal do coeficiente b da reta de regressão. No caso do Exemplo 15.5, o coeficiente de correlação 0,8594 é determinado, na célula H22 da mesma planilha, com a fórmula =SINAL(C16)*RAIZ(RQUAD(C5:C14;B5:B14)).

Erro padrão da estimativa

Ao ajustar uma reta, espera-se que ela explique o grupo de valores amostrados. Embora a reta de regressão tenha sido obtida minimizando a soma dos quadrados dos desvios, sempre haverá uma variabilidade dos dados ao redor da reta, exceto se os dados fizerem parte da própria reta de regressão. O desvio padrão dos dados ao redor da reta de regressão¹¹ é denominado erro padrão da estimativa S_e , cuja medida é obtida da variância com $(n-2)$ graus de liberdade, definida com a fórmula, onde SSE mede a parte não explicada pela regressão:

¹¹ O conceito do erro padrão da estimativa é equivalente ao do desvio padrão, que mede a variabilidade dos valores da amostra ao redor da média aritmética desses valores.

$$S_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

O erro padrão da estimativa é também uma medida da qualidade do ajuste da reta, pois, atendidas as premissas da regressão linear, por exemplo, espera-se que aproximadamente 95% dos valores da amostra y se encontrem no intervalo $\pm 2 \times S_e$ de seus respectivos valores projetados pela reta de regressão \hat{y} .

EXEMPLO 15.6

Calcule o erro padrão da estimativa do Exemplo 15.1.

Solução. Na planilha **Erro padrão**, incluída na pasta **Capítulo 15**, foi calculado o erro padrão da estimativa das amostras do Exemplo 15.1. Na célula F4, foi registrada uma fórmula que utiliza relações matriciais que evitam a necessidade de construir a coluna de projeções e a coluna de variações não explicadas. Depois de registrar a seguinte fórmula, sem pressionar a tecla **Enter**.

$$= \text{RAIZ}(\text{SOMA}((\text{C4:C13} - \text{PREVISÃO}(\text{B4:B13}; \text{C4:C13}; \text{B4:B13}))^2) / (\text{CONT.NÚM}(\text{B4:B13}) - 2))$$

Para inserir essa fórmula como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**; mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, as fórmulas receberam as chaves { }.

Em vez de utilizar essa fórmula, o erro padrão da estimativa pode ser calculado com a função estatística EPADYX do Excel, como feito na planilha **Erro padrão**, registrando na célula F5 a fórmula =EPADYX(C4:C13;B4:B13).

	A	B	C	D	E	F
1	Cálculo do Erro Padrão da Estimativa					
2						
3		x	y		Erro padrão da estimativa	
4		30	430		Com a fórmula	65,17
5		21	335		Com a função EPADYX	65,17
6		35	520			
7		42	490			
8		37	470			
9		20	210			
10		8	195			
11		17	270			
12		35	400			
13		25	480			
14						

• EPADYX(val_conhecidos_y; val_conhecidos_x)

A função estatística EPADYX¹² retorna o erro padrão da estimativa S_e da reta de regressão $\hat{y} = a + bx$, considerando os valores das amostras informados nos argumentos *val_conhecidos_y* e *val_conhecidos_x*. Ao utilizar a função EPADYX, deve-se tomar o cuidado de fornecer os valores na ordem correta, o primeiro argumento *val_conhecidos_y* se refere aos valores da variável dependente y e o argumento *val_conhecidos_x*, aos valores da variável independente x . Os dois argumentos desta função devem ser números ou nomes, matrizes ou referências que contenham números.

Não querendo registrar os valores das amostras, o cálculo do erro padrão da estimativa pode ser realizado em uma única célula da planilha, registrando, por exemplo, a fórmula:

$$= \text{EPADYX}(\{430;335;520;490;470;210;195;270;400;480\}; \{30;21;35;42;37;20;8;17;35;25\})$$

na célula E17.

¹² Em inglês, a função estatística EPADYX é STEYX.

O resultado do erro padrão da estimativa significa que o valor real de vendas é diferente do valor estimado no valor igual a \$65,17 milhões. Embora a reta de regressão possa ajudar a estimar valores de vendas, não podemos esperar uma diferença menor do que \$65,17 milhões com relação aos valores das amostras. Espera-se que aproximadamente 95% dos valores da amostra y se encontrem no intervalo $\pm 2 \times S_{e^*}$ ou \$134,30 milhões, de seus respectivos valores projetados pela reta de regressão \hat{y} .

As premissas do modelo de regressão linear

A amostragem aleatória utilizada para obter a reta de regressão captura alguns pontos da população. A regressão linear realizada é uma estimativa da relação entre as variáveis, relação que é desconhecida. Portanto, os coeficientes de regressão a e b são estimativas pontuais dos correspondentes parâmetros da população α e β .

$$\begin{aligned}\hat{y} &= a + bx \\ \hat{y} &= \alpha + \beta x + e\end{aligned}$$

O valor e^{13} representa a *dispersão* na população, pois não há um relacionamento perfeito entre as duas variáveis na população. De outra maneira, há outras variáveis não consideradas na regressão que também influem na relação, pois a regressão foi realizada com apenas duas variáveis do experimento. Devido à variabilidade amostral, deve-se aceitar que, cada amostra aleatória gerará uma equação de regressão diferente. Portanto, o coeficiente a é um estimador de α e b é um estimador de β . Se toda a população fosse amostrada, o coeficiente a seria igual a α , e b igual a β .

A dispersão na população significa que há diversos valores de y para cada valor de x . Portanto, para cada valor de x há uma distribuição de frequências de y que o modelo de regressão linear supõe que seja uma distribuição normal, denominada *distribuição condicional*, pois depende da *condição* x . Todas as distribuições condicionais de y têm o mesmo desvio padrão, denominado *desvio padrão condicional*. Resumindo, as premissas do modelo de regressão linear são:

- Para cada valor de x , há um grupo de valores de y , e todos os grupos de y têm distribuição normal com o mesmo desvio padrão.
- As médias das distribuições normais de y pertencem à reta de regressão.
- A média dos desvios ou erros é nula, pois a variância é mínima.
- A variância dos desvios é constante e igual à variância da população, pois se supõe que todos os desvios têm a mesma variância.
- Os desvios são variáveis aleatórias independentes e têm distribuição normal. Portanto, o coeficiente de correlação entre os desvios tomados dois a dois é nulo, e os desvios e a variável independente x não têm nenhuma correlação.

Observe que se os dados amostrais disponíveis não forem apropriados, então as inferências da regressão linear poderão ser incorretas.

Intervalo de projeção

Com a reta de regressão das vendas em função do investimento em propaganda, vimos que para um investimento em propaganda de $x=30$, a projeção das vendas é 409,21, resultado obtido no Exemplo 15.3. Cabe perguntar: qual é a variação do valor projetado y para um determinado x , considerando as

¹³ Também denominado *resíduo*.

possíveis amostras que podem ser obtidas da mesma população? A resposta dependerá do objetivo da projeção:

- A média de todas as projeções y para um determinado x . Projeção denominada *média* y .
- A projeção de um único valor y para um determinado x . Projeção denominada *específico* y .

O intervalo de confiança de uma projeção *média* y para um determinado x_i é obtido com a expressão:

$$\hat{y}_i \pm t_c \times S_e \times \sqrt{h_i}$$

E o intervalo de confiança de uma projeção *específico* y para um dado x_i é obtido com a expressão:

$$\hat{y}_i \pm t_c \times S_e \times \sqrt{1+h_i}$$

Em ambos os casos,
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

A primeira parcela dessas expressões, a projeção \hat{y} corresponde a um ponto da reta de regressão e é a mesma para a projeção *média* y e para a projeção *específico* y , correspondente a um determinado x_i . A utilização de um ou outro intervalo de regressão dependerá do objetivo da análise ou do analista. Analisemos a segunda parcela:

- O t crítico t_c é definido com gl igual ao tamanho da amostra menos dois e nível de significância α . Para as duas caudas da distribuição, o valor crítico pode ser obtido com a fórmula =INVT($\alpha;gl$). Quanto maior for o nível de significância, maior será o erro de estimativa.
- Quanto maior for o erro padrão da estimativa S_e , maior será a margem de erro.
- O valor de h_i depende:
 - Da inversa do tamanho da amostra n , quanto maior for n , menor será o intervalo, tendendo a zero.
 - Diretamente do quadrado do desvio de x_i , numerador da segunda parcela da fórmula. Quanto mais afastado de sua média o valor x_i estiver, maior será o erro de estimativa, e vice-versa, quanto mais próximo da média, menor será o erro. Para o próprio valor da média de x , essa parcela será igual a zero.
 - Inversamente da variância de x cuja influência dependerá do valor do numerador.

Vejamos alguns resultados do *modelo* que será explicado a seguir, utilizando os valores do Exemplo 15.1. Para um investimento em propaganda de $x=30$, a projeção das vendas é 409,21 em qualquer um dos casos.

- Para $x=30$, o intervalo da média de projeções das vendas é $409,21 \pm 49,59$. De outra maneira, a média de vendas de todas as lojas para $x=30$ é um valor entre 359,62 e 458,81.
- O intervalo de projeção de um único y é $409,21 \pm 158,26$. De outra maneira, a projeção de vendas de uma única loja para $x=30$ é um valor entre 250,95 e 567,48.

Na planilha **Modelo Intervalo de projeção**, incluída na pasta **Capítulo 15**, foram realizados os cálculos para determinar o intervalo da estimativa dos dois casos apresentados, como mostra a Figura 15.8. Na caixa de grupo **Intervalo de projeção**, pode-se escolher *Média* y ou *Específico* y , clicando no botão de opção correspondente. A Figura 15.8 mostra o gráfico de dispersão dos pontos amostrados, a reta de regressão, o limite inferior e o limite superior do intervalo de confiança da projeção selecionada

e dentro do intervalo de variação da amostra x , e as linhas tracejadas demarcando as médias das amostras x e y . Ainda, o *modelo* conta com dois grupos de informações:

- Informando o valor do nível de significância α na célula C22, na célula C23 é informado o t crítico correspondente, considerando o número de graus de liberdade das amostras, neste caso $8=10-2$.
- Informando um valor de x qualquer na célula F18, o *modelo* calcula todos os resultados relevantes, que também são mostrados no gráfico em uma linha de cor vermelha.

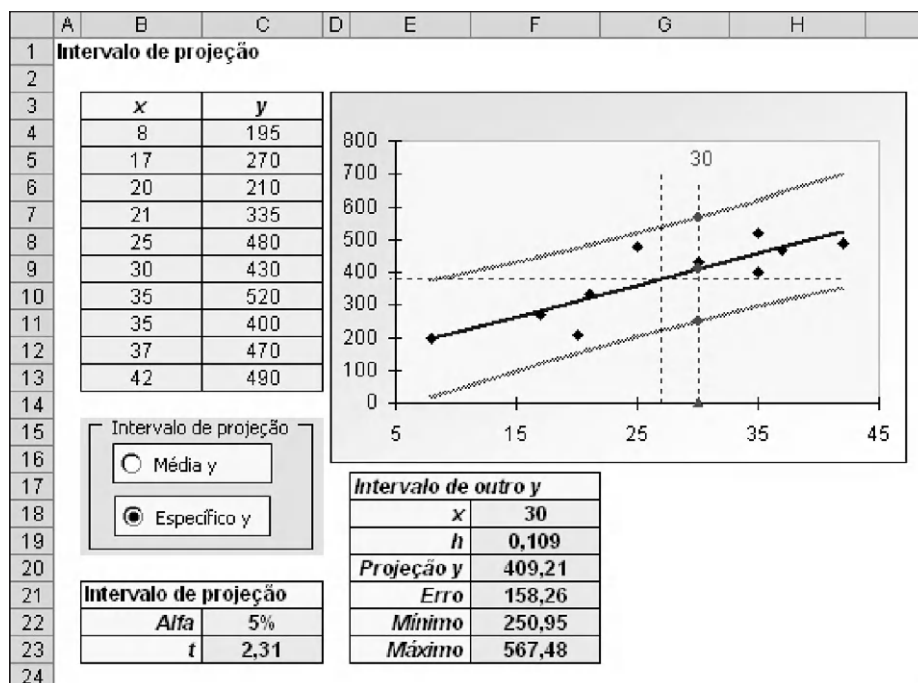


FIGURA 15.8 Intervalo de confiança da projeção, Exemplo 15.1.

Operando com o *modelo*, podem ser observadas as seguintes características:

- O ponto com coordenadas iguais às médias das amostras x e y pertence à reta de regressão.
- O intervalo de confiança da projeção da *Média y* é menor do que da projeção *Específico y*.
- A estimativa correspondente à média da amostra x tem o menor intervalo de todas as estimativas dentro do intervalo de amostragem de x .
- Quanto mais afastado de sua média o valor x estiver, maior será o erro de estimativa. Pela forma das curvas do intervalo, as estimativas fora do intervalo de amostragem de x não devem ser realizadas.
- Variando o nível de significância α na célula C22, pode-se verificar o comportamento do intervalo de estimativa.

Ferramenta de análise Regressão

A ferramenta *Regressão* realiza a análise da regressão linear múltipla incluindo a regressão linear simples. Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel exibirá a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 do livro. Ao escolher a ferramenta **Regressão** e depois de clicar no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostrado na Figura 15.9, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel apresentará a página

Sobre a caixa de diálogo *Regressão* pertencente à *Ajuda do Excel*. Essa ferramenta será apresentada em duas partes, a primeira contendo os resultados gerais e a segunda com os resultados dos resíduos.

A Figura 15.10 mostra o primeiro grupo de resultados dessa ferramenta aplicada no Exemplo 15.1 na planilha **Ferramenta Regressão**, da pasta **Capítulo 15**, a partir da célula E2. As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo dessa ferramenta são:

- **Intervalo Y de entrada:** Informe o intervalo de células no qual os dados da variável dependente estão registrados, incluindo o título.
- **Intervalo X de entrada:** Informe o intervalo de células da planilha no qual os dados da variável independente estão registrados, incluindo o título. O número máximo de amostras independentes é 16.
- **Constante é zero:** Selecione esta opção quando desejar que a linha de regressão passe pela origem.
- **Rótulos:** Selecione este item, pois os intervalos incluem os nomes das amostras.
- **Nível de confiança:** Selecione a caixa e depois informe o intervalo de confiança desejado, por exemplo, neste caso 95% que é o valor *default*.

FIGURA 15.9 Caixa de diálogo da ferramenta *Regressão*.



Na primeira parte do quadro **Opções de saída**, deve ser obrigatoriamente informado um endereço a partir do qual a ferramenta de análise registrará os resultados. Há três alternativas excludentes de informar esse endereço, identificadas por três botões de opção que aceitam a escolha de uma única alternativa:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, neste caso E2, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel definirá automaticamente o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes. Mais informações podem ser obtidas no Capítulo 4 ou na *Ajuda do Excel*.
- **Nova planilha.** Os resultados serão apresentados a partir da célula A1 de uma nova planilha da mesma pasta.
- **Nova pasta de trabalho.** Os resultados serão apresentados em uma nova pasta e a partir da célula A1 da planilha **Plan1**.

As demais seleções disponíveis do quadro **Opções de saída**, da caixa de diálogo da ferramenta **Regressão**, serão apresentadas mais adiante. Depois de completar as informações e clicar em **OK** na caixa de diálogo, o Excel apresentará a partir da célula E2 os resultados divididos em três partes denominadas *Estatística de regressão*, *ANOVA* e uma terceira sem nome. A Figura 15.10 mostra a primeira parte.

	A	B	C	D	E	F
1	Regressão Linear				Ferramenta de Análise	
2					RESUMO DOS RESULTADOS	
3					<i>Estatística de regressão</i>	
4		y	x		R múltiplo	0,85936613
5		430	30		R-Quadrado	0,73851014
6		335	21		R-quadrado ajustado	0,7058239
7		520	35		Erro padrão	65,1734299
8		490	42		Observações	10
9		470	37			
10		210	20			
11		195	8			
12		270	17			
13		400	35			
14		480	25			

FIGURA 15.10
Estatísticas de regressão
do Exemplo 15.1.

A partir da célula E4, no grupo *Estatística de regressão* estão registrados os seguintes resultados:

- **R múltiplo.** É o coeficiente de correlação $r=0,859366$.
- **R-Quadrado.** É o coeficiente de determinação $r^2=0,73851$ da regressão.
- **R-quadrado ajustado.** É o coeficiente de determinação ajustado \bar{r}^2 , medida utilizada em regressão linear múltipla. Partindo da regressão linear simples, com uma única variável independente, o significado do coeficiente de determinação é a porcentagem de explicação dessa regressão. Ao adicionar uma ou mais variáveis independentes, demonstra-se que o r^2 não deverá diminuir, devendo aumentar em alguns casos. O \bar{r}^2 tenta compensar o aumento natural de explicação provocado pelo aumento do número de variáveis independentes e o tamanho da amostra, sendo calculado com a expressão:

$$\bar{r}^2 = r^2 - \frac{k}{n - k - 1} \times (1 - r^2)$$

Nessa expressão, n é o número de valores das amostras, e k é o número de variáveis independentes. Substituindo os dados do exemplo, teremos:

$$\bar{r}^2 = 0,73851 - \frac{1}{10 - 1 - 1} \times (1 - 0,73851) = 0,705824$$

Observe que à medida que n aumenta, \bar{r}^2 se aproxima de r^2 . Esse ajuste pode ser útil para comparar projeções de uma mesma variável dependente realizada com número diferente de variáveis independentes.

- **Erro padrão.** É o erro padrão da estimativa $S_e=65,17$, que já foi calculado no Exemplo 15.6 utilizando a fórmula correspondente e a função estatística EPADYX. Atendidas as premissas da regressão linear, espera-se que aproximadamente 95% das observações y se encontrem dentro do intervalo $\pm 2S_e$ de seus respectivos valores projetados \hat{y} da reta de regressão.
- **Observações.** É o número de valores das amostras que devem ter o mesmo tamanho.

FIGURA 15.11

ANOVA do Exemplo
15.1, ferramenta de
análise *Regressão*.

	D	E	F	G	H	I	J
11		ANOVA					
12			<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
13		Regressão	1	95969,3923	95969,392	22,59392	0,001439122
14		Resíduo	8	33980,6077	4247,576		
15		Total	9	129950			
16							

A partir da célula E11, a ferramenta *Regressão* apresenta o grupo ANOVA de resultados, comentados a seguir, apresentados na Figura 15.11.

- **Coluna *gl*.** São registrados os graus de liberdade. A célula F13 registra o número de variáveis independentes. A célula F14 registra o resultado de $n-k-1=10-1-1=8$. A célula F15 registra $n-1=10-1=9$.
- **Coluna *SQ*.** Todos os resultados seguintes já foram obtidos no Exemplo 15.5. Assim temos que na célula G13 foi registrada a soma dos quadrados das variações explicadas pela regressão, $SSR=95.969,39$; na célula G14, foi registrada a soma dos quadrados das variações não explicadas pela regressão $SSE=33.980,61$ e, na célula G15, a soma dos quadrados das variações totais $SST=129.950$.
- **Coluna *MQ*.** Cada um dos dois valores registrados nessa coluna é o resultado da divisão do valor do *SQ* pelo correspondente número de graus de liberdade da coluna *gl*.
- **Coluna *F*.** O valor do *F* observado, registrado na célula I13, é o resultado da divisão do *MQ* da linha *Regressão* pelo *MQ* da linha *Resíduo*, resultando no valor 22,59.

Outra forma de obter o resultado do *F* observado é o seguinte. Enquanto a distribuição *t* é utilizada para realizar testes de hipóteses dos coeficientes da reta de regressão, a distribuição *F* é utilizada para realizar testes de hipóteses da equação da reta de regressão. A distribuição *F* testa a hipótese de que nenhum dos coeficientes de regressão tenha significado. Para isso, o *F* observado é:

$$F_o = \frac{\text{Variância explicada}}{\text{Variância não explicada}}$$

Para operar com variâncias, a variação explicada deve ser dividida pelo *gl* do numerador ($k-1$), e a variação não explicada deve ser dividida pelo *gl* do denominador, sendo k o número de amostras e n o tamanho das amostras. Portanto, o *F* observado é igual a:

$$F_o = \frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k-1}}{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-k}}$$

Considerando o coeficiente de determinação, demonstra-se que:

$$F_o = \frac{\frac{r^2}{k-1}}{\frac{1-r^2}{n-k}}$$

Nesse exemplo, o *F* observado é igual a 22,59, valor obtido com a última fórmula, considerando o coeficiente de determinação 0,7385:

$$F_o = \frac{\frac{0,7385}{2-1}}{\frac{1-0,7385}{10-2}} = 22,59$$

Observe que ao aplicar o teste F na regressão linear simples, o número de graus de liberdade do numerador é sempre igual a um, e a distribuição F é igual à distribuição t ao quadrado, isto é $F = t^2$.

- **Coluna F de significação.** É o p -value do F crítico correspondente, registrado em I13. O p -value=0,00143912 ou 0,144%, registrado na célula J13, pode ser obtido com a fórmula =DISTF(22,59392;1;8), utilizando os graus de liberdade do numerador e do denominador registrados nessa fórmula. O procedimento do teste de hipóteses é:

$$H_0 : \beta=0$$

$$H_1 : \beta \neq 0$$

Como o p -value 0,144% é menor do que o nível de significância 5%, valor equivalente ao intervalo de confiança 95% definido, a hipótese nula deve ser rejeitada. De outra maneira, há evidências do, que o β seja diferente de zero e, conseqüentemente, a regressão deve ser aceita.

	D	E	F	G	H	I	J	K
17								
18			<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
19		Interseção	117,070158	59,0298508	1,9832366	0,082634	-19,05300997	253,1933262
20		x	9,73814229	2,04870922	4,7533062	0,001439	5,013807314	14,46247727
21								

FIGURA 15.12

Resultados do Exemplo 15.1 com a ferramenta de análise Regressão.

A partir da célula E18, a ferramenta *Regressão* apresenta o último grupo de resultados, como mostra a Figura 15.12. Para compreender os resultados, a partir da célula E24 são registrados os mesmos resultados baseados nos conceitos desenvolvidos no livro e utilizando as fórmulas conhecidas.

- **Coluna Coeficientes.** Na célula F19, é registrado o valor do coeficiente a , e na célula F20, o do coeficiente b . No caso de regressão linear múltipla, os outros coeficientes b s serão apresentados em sequência, a partir da célula F21.
- **Coluna Erro padrão.** São os erros padrões dos coeficientes a e b .
 - **Erro padrão do coeficiente a .** O erro padrão S_a do coeficiente a indica aproximadamente quão distante o coeficiente a está do coeficiente da população devido à variabilidade amostral. A fórmula utilizada é:

$$S_a = S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1) \times S_x^2}}$$

A fórmula mostra que o erro padrão do coeficiente a é proporcional ao erro padrão da estimativa S_e . Nesse exemplo, o erro padrão do coeficiente a é igual a 59,03, resultado obtido com a fórmula:

$$S_a = 65,17 \sqrt{\frac{1}{10} + \frac{27^2}{(10-1) \times \frac{10}{9} \times 101,20}} = 59,03$$

- **Erro padrão do coeficiente b .** O erro padrão S_b do coeficiente b indica, aproximadamente, quanto distante o coeficiente b está do coeficiente da população β devido à variabilidade amostral. A fórmula utilizada é:

$$S_b = \frac{S_e}{\sqrt{(n-1) \times S_x^2}}$$

A fórmula mostra que o erro padrão do coeficiente b é diretamente proporcional ao erro padrão da estimativa S_e , e inversamente proporcional à variância de x e o tamanho da amostra menos um. O erro padrão do coeficiente b do Exemplo 15.1 é igual a 2,05, resultado obtido com a fórmula:

$$S_b = \frac{65,1734}{\sqrt{(10-1) \times \frac{10}{9} \times 101,20}} = 2,0487$$

- **Coluna Stat t:** É a estatística t ou t observado dos coeficientes a e b . Supondo que as variáveis x e y não sejam relacionadas, o que se pode dizer dos coeficientes da reta da população?

- **Stat t do coeficiente a .** Se as variáveis x e y não são relacionadas, então $\alpha=0$ e o teste de hipóteses é:

$$H_0 : \alpha=0$$

$$H_1 : \alpha \neq 0$$

Utilizando a distribuição t , o t observado é $t = \frac{a - \alpha}{S_a}$, e como o coeficiente α é zero, o t observado

será $t = \frac{a - 0}{S_a}$. Neste exemplo, o t observado é $t = \frac{117,07}{59,0298} = 1,983237$.

- **Stat t do coeficiente b .** Se as variáveis x e y não são relacionadas, então $\beta=0$ ¹⁴ e o teste de hipóteses é:

$$H_0 : \beta=0$$

$$H_1 : \beta \neq 0$$

Utilizando a distribuição t , o t observado é $t = \frac{b - \beta}{S_b}$, e como o coeficiente β é zero, o t observado

será $t = \frac{b - 0}{S_b}$. Neste exemplo, o t observado é $t = \frac{9,7381}{2,0487} = 4,753$.

- **Coluna Valor-P.** É a probabilidade $P(t \geq t \text{ observado})$, ou p -value correspondente. Com a função estatística DISTT do Excel, obtém-se os seguintes resultados:

- **Coeficiente a .** Com a fórmula =DISTT(1,98323656;8;2), obtém-se o resultado do p -value igual a 0,0826337, ou 8,26%.
- **Coeficiente b .** Com a fórmula =DISTT(4,75330624;8;2), obtém-se o resultado do p -value igual a 0,0014391, ou 0,144%.

Devido à variabilidade amostral, a reta de regressão obtida da amostra extraída da população é uma das muitas retas possíveis. Da mesma maneira, supondo que seja possível utilizar os valores da própria população, teremos de aceitar a diferença entre os valores reais e os valores projetados pela reta de regressão da população $\hat{y} = \alpha + \beta x + e$, sendo e o erro cometido na projeção. Essa diferença é devida às limitações do modelo linear em conseguir representar a realidade com apenas uma variável aleatória independente. Portanto, os coeficientes a e b obtidos de uma amostragem aleatória não se-

¹⁴ Na ausência de melhores informações, a melhor estimativa de uma variável aleatória é sua própria média.

rão iguais, em geral, aos coeficientes a e b da população. Entretanto, demonstra-se que a e b são os melhores estimadores não tendenciosos de α e β , respectivamente.

- **Coluna 95% inferiores.** É o valor do limite inferior do intervalo de confiança de cada coeficiente de regressão.¹⁵
- **Coeficiente a .** O t crítico da distribuição t é $\text{INVT}(0,05;8)=2,306$, com $(10-2)=8$ graus de liberdade e considerando o nível de significância 5% nas duas caudas. E a estimativa do coeficiente de regressão a , com nível de significância de 5%, é $a = a \pm t \times S_a$. O limite inferior do coeficiente de regressão a é -19,05, resultado obtido com a fórmula:

$$\begin{aligned} a_{\min} &= a - t \times S_a \\ a_{\min} &= 117,070158 - 2,3060 \times 59,0298 = -19,05 \end{aligned}$$

- **Coeficiente b .** Na seção **Intervalo da Projeção**, foi analisado o intervalo de confiança do valor projetado \hat{y} , considerando como média ou valor de y . O mesmo pode ser feito com o coeficiente b , que é um estimador pontual do coeficiente da população β . Considerando o nível de significância α , a estimativa do coeficiente de regressão b é $b = b \pm t \times S_b$. Utilizando o t crítico da distribuição t igual a 2,306, resultado obtido anteriormente, o limite inferior do coeficiente de regressão b é 5,01, resultado obtido com a fórmula:

$$\begin{aligned} b_{\min} &= b - t \times S_b \\ b_{\min} &= 9,738142 - 2,3060 \times 2,04871 = 5,01 \end{aligned}$$

- **Coluna 95% superiores.** É o valor do limite superior de cada coeficiente de regressão.
- **Coeficiente a .** O limite superior do coeficiente de regressão a é -253,19, resultado obtido com a fórmula:

$$\begin{aligned} a_{\max} &= a + t \times S_a \\ a_{\max} &= 117,070158 + 2,3060 \times 59,0298 = 253,19 \end{aligned}$$

- **Coeficiente b .** O limite superior do coeficiente de regressão b é 14,46, resultado obtido com a fórmula:

$$\begin{aligned} b_{\max} &= b + t \times S_b \\ b_{\max} &= 9,738142 + 2,3060 \times 2,04871 = 14,46 \end{aligned}$$

A reta de regressão passa pela origem

Se a reta de regressão da população passar pela origem ($x=0, y=0$), a equação dessa reta será $\hat{y} = bx$. É a mesma equação utilizada até este momento, porém com o intercepto a igual a zero. A reta de regressão do Exemplo 15.1, que relaciona as vendas e o investimento em propaganda, não passa pela origem, pois o intercepto $a=117,07$. Embora essas duas variáveis tenham sido relacionadas, na realidade nem todas as vendas são provocadas pelas campanhas de propaganda, pois o modelo mostra que para propaganda igual a zero as vendas serão iguais a 117,07. Como exercício, suponhamos que seja possível separar as vendas em dois grupos, as vendas provocadas pela propaganda e todas as demais vendas não provocadas pela propaganda, divisão difícil de realizar na prática. Nesse caso, a reta de regressão das vendas provocadas pela propaganda passará pela origem ($x=0, y=0$), pois se num determinado período não houver investimento em propaganda, também não ocorrerão essas vendas.

¹⁵ No Excel 2002 em português, versão (10.5815.4219) SP-2, este resultado e o seguinte são repetidos nas duas colunas seguintes.

Embora não seja frequente, há casos em que o analista pode saber antecipadamente que a reta de regressão passa pela origem. Por exemplo, um provedor de Internet cujo faturamento depende somente da venda de contratos mensais de fornecimento de acessos, se não tiver contratos, não terá faturamento. Outro caso, a reta de regressão entre os custos mensais de envio de produtos vendidos mensalmente pela TV com despesas de envio incluídas também passa pela origem, pois se num mês não for vendido nenhum produto, as despesas de envio serão igual a zero.

A fórmula da declividade b da equação da reta de regressão $\hat{y} = bx$ que passa pela origem é:

$$b = \frac{\sum_{i=1}^n x_i \times y_i}{\sum_{i=1}^n x_i^2}$$

As fórmulas para calcular o erro de estimativa, erros padrões etc. também são diferentes¹⁶ e não serão apresentadas neste livro. Entretanto, a ferramenta de análise *Regressão* aceita retornar as respostas, considerando que a reta passa pela origem. Para isso, no quadro **Entrada** da caixa de diálogo da Figura 15.9, deve-se selecionar a caixa **Constante é zero**. Conhecendo o significado das respostas do caso geral da ferramenta *Regressão*, você não terá dificuldade em compreendê-las quando a reta de regressão passa pela origem. No Apêndice 3 deste capítulo, são apresentadas as funções PROJ.LIN, PROJ.LOG e CRESCIMENTO, que retornam respostas da reta de regressão e também aceitam retornar as respostas, considerando que a reta passa pela origem.

Completando os resultados da ferramenta de análise regressão

No restante do quadro **Opções de saída**, dividido em **Resíduos** e **Probabilidade normal**, a ferramenta de análise *Regressão* retorna outras respostas úteis para a análise dos resultados da regressão linear. Na caixa de diálogo *Regressão* da Figura 15.13, foram selecionadas todas as demais respostas disponíveis.



FIGURA 15.13
Ferramenta Regressão,
incluindo resíduos.

Para facilitar a compreensão dos resultados, foi realizada uma nova regressão, agora completa, a partir da célula R2. Depois de clicar em OK, a ferramenta apresentará os resultados numéricos da Figura 15.14.

- **Resíduos.** A análise dos resíduos é um procedimento gráfico que permite analisar o ajuste da reta de regressão. As medidas de variação que definiram o coeficiente de determinação e o erro de estimativa partiram do conceito de desvios medidos a partir da média da amostra y dos desvios medidos ao redor da reta de regressão. O desvio do valor projetado com relação ao valor observado é denominado *resíduo*, como mostra a fórmula $e = y_i - \hat{y}_i$. Lembrando que uma das premissas do modelo de regressão linear estabelece que a média dos desvios ou erros é nula, pois a variância é mínima, a forma do gráfico dos desvios em função dos valores x da amostra ajudará a verificar o acerto da reta de regressão. Se os desvios não mantêm nenhum padrão com os valores x , pode-se deduzir que a reta de regressão é uma boa representação dos dados observados, como mostra a Figura 15.15, construída pela ferramenta por ter sido selecionada a caixa **Plotar resíduos**.

	Q	R	S	T	U	V	W	X
23		RESULTADOS DE RESÍDUOS					RESULTADOS DE PROBABILIDADE	
24								
25		<i>Observação</i>	<i>Previsto(a) y</i>	<i>Resíduos</i>	<i>Resíduos padrão</i>		<i>Percentil</i>	<i>y</i>
26		1	409,214427	20,7855731	0,338273275		5	195
27		2	321,571146	13,4288538	0,218546889		15	210
28		3	457,905138	62,0948617	1,01055824		25	270
29		4	526,072134	-36,0721344	-0,587053287		35	335
30		5	477,381423	-7,38142292	-0,120128422		45	400
31		6	311,833004	-101,833004	-1,657273703		55	430
32		7	194,975296	0,02470356	0,000402036		65	470
33		8	282,618577	-12,6185771	-0,205360101		75	480
34		9	457,905138	-57,9051383	-0,942372898		85	490
35		10	360,523715	119,476285	1,94440797		95	520
36								

FIGURA 15.14
Resultados da
Ferramenta
Regressão,
incluindo resíduos.

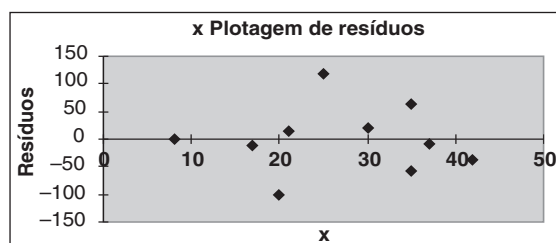


FIGURA 15.15 Gráfico dos
resíduos em função dos
valores amostrados x .

- **Resíduos padronizados.** As últimas duas premissas do modelo de regressão linear estabelecem que a variância dos desvios é constante e igual à variância da população, pois se supõe que todos os desvios têm a mesma variância. Contudo, os resíduos não são independentes e possuem variâncias diferentes, que dependem do valor de x correspondente. Os resíduos padronizados são resíduos transformados. O resíduo padronizado para cada x_i pode ser obtido com:¹⁷

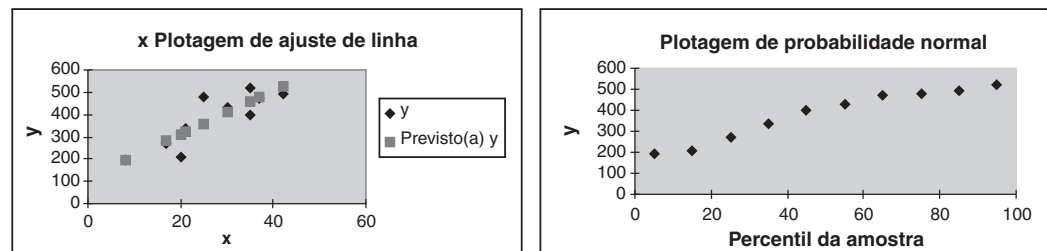
$$R_i = \frac{e_i}{S_e \times \sqrt{1 - h_i}}$$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

¹⁷ *Análise de Modelos de Regressão Linear com Aplicações* – Charnet R. et al – Editora da Unicamp, 1999.

- **Plotar resíduos.** É o gráfico dos resíduos para cada valor de x .
- **Plotar ajuste de linha.** É o gráfico de dispersão contendo os pares amostrados e a reta de regressão.
- **Plotagem de probabilidade normal.** É o gráfico de cada valor da amostra y em função de seu correspondente percentil em uma escala de 5 a 95%. O percentil de cada valor da amostra y ordenada de forma crescente deste exemplo é obtido com a expressão $p\% = \frac{90}{10 - 1} \times (d - 1) + 5$. Nessa fórmula, d é a ordem de um valor da série, e p é o percentil em porcentagem dessa ordem, em uma escala de 5 a 95%. Essa expressão é equivalente à expressão apresentada ao estudar as medidas de ordenamento no Capítulo 3, porém em uma escala de 5 a 95%.

FIGURA 15.16
Outros gráficos da
ferramenta *Regressão*,
Exemplo 15.1.



Regressão linear múltipla

O modelo de regressão linear apresentado é o mais simples dos modelos de regressão, que nem sempre atende à modelagem mais complexa. Como vimos no Exemplo 15.1, as vendas não dependem somente do investimento em propaganda, pois há uma parte da variação das vendas que não é explicada pela propaganda. Da mesma maneira, a demanda de um determinado produto pode ser explicada pela combinação do preço unitário e do investimento em propaganda. Em geral, a aplicação quantitativa de conceitos econômicos requer a estimação de funções de oferta, demanda, custo etc. Os modelos lineares com mais de uma variável independente se denominam modelos de regressão linear múltipla. O desenvolvimento da equação de regressão linear múltipla é similar ao da equação de regressão linear simples incluindo a dependência de duas ou mais variáveis independentes. Tanto a ferramenta de análise *Regressão* quanto as funções PROJ.LIN, PROJ.LOG e CRESCIMENTO, apresentadas no Apêndice 3, realizam análises de regressão múltipla.

Dispondo de um grupo de amostras do mesmo tamanho, sendo uma variável dependente y e n variáveis independentes x_i , o objetivo é determinar os coeficientes da equação da reta $\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$, cujos coeficientes minimizam a soma dos quadrados dos desvios da variável \hat{y} com relação a y . A análise de regressão múltipla será apresentada através do Exemplo 15.7, utilizando a ferramenta de análise *Regressão*.

EXEMPLO 15.7

O analista de marketing de uma rede de varejo acredita que um modelo que relacione a quantidade (y em milhares) de peças de roupa íntima vendidas por trimestre com o preço médio de (x_1 em \$) e o investimento em propaganda (x_2 em \$milhares) poderá ser útil para projetar a quantidade de peças do trimestre seguinte e reduzir o custo unitário ocasionado pelo menor risco de encalhe. Para encontrar essa relação linear, foi extraída a amostra de valores trimestrais registrada na tabela seguinte. Determine a equação de regressão e analise quão bem ela se ajusta às necessidades do analista.

y	x_1	x_2
252	32	655
339	26	616
358	26	678
327	31	676
414	27	501
353	27	636
281	34	632
265	39	712
260	39	523
413	36	474

Solução. Este exemplo está resolvido na planilha **Regressão múltipla** da pasta **Capítulo 15**. Depois de selecionar no menu **Ferramentas – Análise de dados – Regressão**, o Excel exibirá a caixa de diálogo **Regressão**, mostrada na figura a seguir, depois de selecionadas as opções do exemplo.

Regressão ? X

Entrada

Intervalo Y de entrada:

Intervalo X de entrada:

☒ Rótulos ☐ Constante é zero

☒ Nível de confiança %

Opções de saída

☒ Intervalo de saída:

☐ Nova planilha:

☐ Nova pasta de trabalho

Resíduos

☒ Resíduos ☒ Plotar resíduos

☒ Resíduos padronizados ☒ Plotar ajuste de linha

Probabilidade normal

☒ Plotagem de probabilidade normal

OK Cancelar Ajuda

A próxima figura mostra os resultados da ferramenta de análise *Regressão*.

	F	G	H	I	J	K	L
1	Ferramenta de Análise						
2	RESUMO DOS RESULTADOS						
3							
4	<i>Estatística de regressão</i>						
5	R múltiplo	0,792597168					
6	R-Quadrado	0,628210271					
7	R-quadrado ajustado	0,521984634					
8	Erro padrão	41,747129					
9	Observações	10					
10							
11	ANOVA						
12		<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>	
13	Regressão	2	20613,84054	10308,92027	5,913923314	0,031335805	
14	Resíduo	7	12199,75946	1742,82278			
15	Total	9	32813,6				
16							
17		<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
18	Interseção	818,1449356	144,2127051	5,673182089	0,000756056	477,1363197	1159,153552
19	x1	-7,045635658	2,716647789	-2,59350354	0,035761838	-13,46948231	-0,621789008
20	x2	-0,440108611	0,171069263	-2,572692509	0,036862686	-0,844622849	-0,035594373
21							

Dessa tabela, tiramos os seguintes resultados:

- A equação da reta de regressão é $\hat{y} = 818,145 - 7,046x_1 - 0,44x_2$. Os dois coeficientes de regressão são negativos. Um aumento do preço médio reduzirá a quantidade vendida e vice-versa, uma redução do preço médio aumentará a quantidade de peças vendidas. O comportamento da propaganda é parecido com o preço médio, o aumento no investimento em propaganda reduzirá a quantidade vendida, porém com menor força do que a redução do preço médio.
- O coeficiente de determinação r^2 é igual a 0,6282, resultado obtido dividindo a soma dos quadrados dos desvios explicados 20.613,84 (célula H13) pela soma dos quadrados dos desvios totais 32.813,6 (célula H15).
- O coeficiente de determinação ajustado \bar{r}^2 tenta compensar o aumento natural de explicação provocado pelo aumento do número de variáveis independentes e o tamanho da amostra. O resultado 0,521985, apresentado na célula G7, foi obtido com a fórmula $\bar{r}^2 = r^2 - \frac{k}{n-k-1} \times (1-r^2)$, onde k é o número de variáveis independentes e n é o número de observações das amostras. O coeficiente de determinação ajustado mostra que somente 52,20% da variação da quantidade das peças vendidas podem ser explicadas pelas duas variáveis independentes.
- O erro padrão da estimativa $S_e = 41,7471$, apresentado na célula G8, foi obtido com a fórmula

$$S_e = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1}}$$

- Atendidas às premissas da regressão linear, espera-se que aproximadamente 95% das observações y se encontrem dentro do intervalo $\pm 2S_e$ de seus respectivos valores projetados \hat{y} da reta de regressão.
- Com o F observado 5,9139 (célula J13), obtém-se o p -value igual a 0,03134 (célula K13) ou 3,13%, resultado obtido com a fórmula =DISTF(5,9139;2;7), utilizando os graus de liberdade do numerador e do denominador registrados nessa fórmula. O procedimento do teste de hipóteses é:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \beta_1 \neq 0 \text{ e } \beta_2 \neq 0$$

Como o p -value 3,13% é menor do que o nível de significância 5%, valor equivalente ao intervalo de confiança 95% definido, a hipótese nula deve ser rejeitada. De outra maneira, há evidências de que a quantidade vendida seja explicada pelo preço médio das vendas unitárias e o investimento em propaganda.

A planilha Solver RLM, incluída na pasta Capítulo 15, foi preparada para calcular os coeficientes de regressão das amostras do Exemplo 15.7, utilizando o comando Solver do Excel. Sugerimos que você calcule os coeficientes de regressão seguindo e adaptando as instruções do Apêndice 1 deste capítulo.

Problemas

Problema 1

O programador de produção gostaria de utilizar um modelo de regressão linear para realizar previsões de demanda e conseguir estabelecer as quantidades de produção requeridas para atender a essas previsões. Os dados históricos disponíveis na empresa relacionam as vendas com o dispêndio em propaganda, como mostra a tabela seguinte, ambos os valores monetários na mesma escala. Determine a equação da reta de regressão.

Vendas	26	11,6	18	13,5	19	23,5
Propaganda	2,5	1,3	1,4	1,1	1,3	2,2

$$R: \hat{y} = 4,35 + 8,72 \times x$$

Problema 2

Analise os resultados da regressão do Problema 1 e verifique se esse modelo deve ser utilizado para representar os dados históricos.

Problema 3

Continuando com o Problema 1. Se o gerente de marketing informou que no próximo mês espera investir 2,35 em propaganda, qual deve ser o valor de vendas esperado?

R: Vendas = 24,85

Problema 4

Os resultados da regressão linear realizada foram os seguintes: $a=-7,98$ – $b=111,23$ – $r=0,965$ – $F_{\text{observado}}=6,73$. Calcule o coeficiente de determinação, analise os resultados e verifique se esse modelo representa a relação das variáveis dependente e independente.

Problema 5

Os custos totais do produto mais importante da empresa durante os últimos seis meses estão registrados na tabela seguinte junto com as quantidades produzidas nos mesmos meses. Determine a equação da reta de regressão.

R: $\hat{y} = 500 + 2 \times x$

Problema 6

Analise os resultados da regressão do Problema 5 e verifique se esse modelo deve ser utilizado para representar os dados históricos.

Problema 7

Construa a reta de regressão linear dos retornos das ações PN em função dos retornos das ações ON, a partir das amostras registradas na tabela seguinte.

ON%	37,5	-45	0	31,5	-1	20,1	212,5	46,3	11,1	43	67	9,4
PN%	20,9	5,4	49,4	31,1	30	28	367,1	6,9	45,4	27,8	43,1	13,4

R: $PN = 4,7014 + 1,4155 \times ON$

Problema 8

Continuando com o Problema 7. Verifique se a regressão deve ser aceita aplicando a distribuição F e considerando o nível de significância de 5%.

R: Aceitar a regressão linear. Porém, você deve analisar o diagrama de dispersão das rentabilidades das ações tipo PN em função das ações ON.

Problema 9

Continuando com o Problema 7. Retirando as observações suspeitas 212,50% e 367,10%, obtenha a nova reta de regressão linear.

R: $PN = 25,0771 + 0,1162 \times ON$

Problema 10

Continuando com o Problema 9. Verifique se a regressão deve ser aceita aplicando a distribuição F com nível de significância de 5%.

R: Rejeitar a regressão linear.

Problema 11

O gerente de vendas está sempre insistindo com os vendedores que a venda dos seus produtos tem forte relação com as visitas realizadas pelos vendedores aos seus clientes. A empresa tem onze vendedores e, como regra, eles visitam seus clientes uma vez por mês. Para confirmar a *crença* do gerente de vendas, foi preparada a tabela a seguir com as visitas realizadas e as vendas de cada vendedor durante o mês passado. Obtenha a reta de regressão das vendas em função das visitas.

	Visitas do mês	Vendas do mês
Samuel	42	140
Ricardo	105	330
Suely	66	190
Manoel	87	350
Ivany	50	110
Rafaela	55	135
Carlos	51	140
João	60	235
Susana	40	70
Marcos	87	320
Andréa	78	220

R: $a=-72,2062$ e $b=4,2084$

Problema 12

Continuando com o Problema 11. Calcule o coeficiente de determinação e o erro padrão da estimativa.

R: $r^2=0,8555$ e $S_e=38,3871$

Problema 13

Continuando com o Problema 11. Considerando o nível de significância 5%, verifique se a regressão deve ser aceita aplicando a distribuição *F*.

R: *Aceitar regressão linear.*

Problema 14

Refça alguns dos Problemas anteriores utilizando a ferramenta de análise *Regressão*.

Problema 15

O sistema de TV a *cabo* cobra uma mensalidade de \$150 mais \$15 por cada pagamento por evento que o usuário solicitar. Considerando que a variável *x* represente o número de pagamentos por evento por mês e a variável *y* o pagamento total por mês, qual é o tipo de relação entre as duas variáveis? Defina a equação entre as duas variáveis.

Problema 16

Continuando com o Problema 15, analise os resultados e verifique se essa reta de regressão deve ser aceita, considerando o intervalo de confiança de 95%.

Problema 17

Os *prêmios* e *preços de exercícios* de cinco séries de opções de compra com mesmo vencimento estão registrados na tabela seguinte. Realize uma análise de regressão dos *Prêmios* em função do *Preço de Exercício*.

Prêmios	Preços de Exercício
\$257,52	\$2.100
\$99,25	\$2.200
\$38,17	\$2.300
\$14,65	\$2.400
\$5,61	\$2.500

Apêndice 1

Determinação dos coeficientes de regressão com o Solver

O objetivo é determinar os coeficientes a e b da reta de regressão $y = a + bx$, que minimizam a soma dos quadrados dos desvios ou, de outra forma, encontrar a e b tal que a soma dos quadrados dos desvios seja um *mínimo*. De forma matemática:

$$\text{minimizar} \Rightarrow \sum_{i=1}^n (y_i - a - bx_i)^2$$

Utilizando o Exemplo 15.1, será mostrado como utilizar o comando Solver para encontrar os valores dos coeficientes de regressão que cumprem com a condição de mínima soma dos quadrados dos desvios. Começamos por preparar a planilha Solver,¹⁸ como mostra a Figura 15.17.

- No intervalo B3:C13, foram registradas as amostras conhecidas.
- No intervalo G4:G5, foram registrados os títulos dos coeficientes de regressão, a e b , e, no intervalo H4:H5, o comando Solver registrará os resultados procurados.
- Na célula D4, foi registrada a fórmula $=\$H\$4+\$H\$5*B4$ que projeta o valor \hat{y} para $x=30$, utilizando os coeficientes de regressão do intervalo H4:H5. Depois, essa fórmula foi copiada até a célula D13.
- Na célula E4, foi registrada a fórmula $=(C4-D4)^2$, que calcula o quadrado do desvio da projeção da célula D4. Depois, essa fórmula foi copiada até a célula E13. Na célula D14, foi registrada a fórmula que calcula a soma dos quadrados dos desvios $=\text{SOMA}(E4:E13)$. Neste momento, o modelo está preparado para utilizar o Solver.

FIGURA 15.17
Preparação da planilha para utilizar o Solver.

	A	B	C	D	E	F	G	H
1	Coeficientes de regressão do Exemplo 15.1 utilizando o Solver							
2								
3		x	y	Projeção	Erro		Coeficientes de regressão	
4		30	430	0	184.900		a	0,00
5		21	335	0	112.225		b	0,00
6		35	520	0	270.400			
7		42	490	0	240.100			
8		37	470	0	220.900			
9		20	210	0	44.100			
10		8	195	0	38.025			
11		17	270	0	72.900			
12		35	400	0	160.000			
13		25	480	0	230.400			
14				Soma	1.573.950			
15								

No menu **Ferramentas**, selecione **Solver**¹⁹ e depois preencha as opções como mostrado na Figura 15.18.

¹⁸ O Solver é um *Suplemento* que nem sempre é incorporado ao iniciar o Excel. Para obter mais informações, veja o Apêndice 1 do Capítulo 1, ou a *Ajuda* do Excel.

¹⁹ Se o comando Solver não estiver incluído no menu **Ferramentas**, então verifique se o Solver aparece no menu **Ferramentas – Suplementos**, onde deve ser selecionado. Se em **Suplementos** não aparecer o Solver, então esse suplemento não foi instalado.

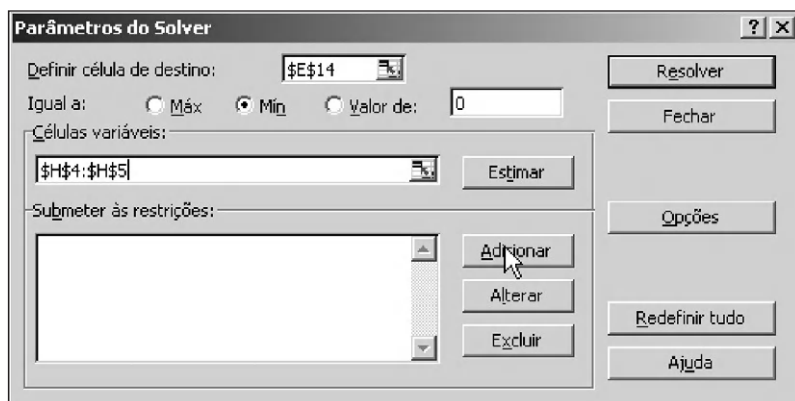


FIGURA 15.18 Caixa de diálogo do comando Solver.

Depois de clicar no botão **Resolver**, o comando Solver registrará a solução encontrada nas células H4:H5, neste caso $a=117,07$ e $b=9,74$, e exibirá a caixa de diálogo da Figura 15.19. Clicando no botão **OK**, os resultados serão mantidos no intervalo H4:H5. Se clicar **Cancelar**, serão mantidos os valores iniciais registrados nesse intervalo, da mesma forma se selecionar a caixa **Restaurar valores originais** e depois clicar em **OK**.

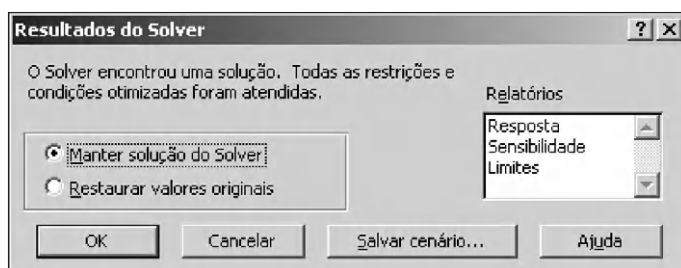


FIGURA 15.19 Caixa de diálogo Resultados do Solver.

Apêndice 2

Fórmulas dos coeficientes de regressão

O objetivo é obter as expressões dos coeficientes a e b da reta de regressão $y = a + bx$, que minimizam a soma dos quadrados dos desvios ou, de outra forma, encontrar a e b tal que a soma dos quadrados dos desvios seja um *mínimo*. De forma matemática, onde D é utilizado para facilitar o desenvolvimento das fórmulas:

$$D = \sum_{i=1}^n (y_i - a - bx_i)^2$$

Começamos por calcular as derivadas parciais da soma dos quadrados dos desvios com relação, primeiro, ao coeficiente a e, depois, ao coeficiente b .

$$\frac{\partial D}{\partial a} = -2 \times \sum_{i=1}^n (y_i - a - bx_i)$$

$$\frac{\partial D}{\partial b} = -2 \times \sum_{i=1}^n (y_i - a - bx_i) \times x_i$$

Para encontrar os valores mínimos, as duas derivadas são igualadas a zero.

$$\begin{aligned} -2 \times \sum_{i=1}^n (y_i - a - bx_i) &= 0 \\ -2 \times \sum_{i=1}^n (y_i - a - bx_i) \times x_i &= 0 \end{aligned}$$

Desenvolvendo essas fórmulas:

$$\begin{aligned} \sum_{i=1}^n y_i - \sum_{i=1}^n a - \sum_{i=1}^n bx_i &= 0 \\ \sum_{i=1}^n x_i y_i - \sum_{i=1}^n ax_i - \sum_{i=1}^n bx_i^2 &= 0 \end{aligned}$$

Simplificando as parcelas das duas fórmulas.

$$\begin{aligned} \sum_{i=1}^n y_i &= na + b \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i &= a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \end{aligned}$$

Resolvendo esse sistema de duas equações lineares, obtém-se as seguintes expressões, que permitem calcular os coeficientes de regressão a e b . Da primeira equação do sistema anterior, temos a fórmula de a .

$$\begin{aligned} \sum_{i=1}^n y_i &= na + b \sum_{i=1}^n x_i \\ a &= \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n} \end{aligned}$$

Depois, substituindo a expressão de a na segunda equação.

$$\begin{aligned} \sum_{i=1}^n x_i y_i &= \frac{1}{n} \left(\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \right) \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i y_i &= \frac{1}{n} \left(\sum_{i=1}^n y_i \times \sum_{i=1}^n x_i \right) - b \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 + b \sum_{i=1}^n x_i^2 \\ b &= \frac{n \sum_{i=1}^n x_i \times y_i - \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \end{aligned}$$

Por último, agrupando as fórmulas dos coeficientes de regressão a e b .

$$\begin{cases} a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n} \\ b = \frac{n \sum_{i=1}^n x_i \times y_i - \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \end{cases}$$

As expressões encontradas correspondem a um mínimo, pois não há um máximo para a função da soma dos quadrados dos desvios. Isso pode ser provado analiticamente; entretanto, com o **Modelo Ajuste da reta**, mostrado na Figura 15.1, pode-se constatar que para qualquer reta definida no espaço dos pontos do gráfico de dispersão, sempre haverá uma reta por cima ou por baixo desses pontos, cuja soma dos quadrados dos desvios será maior.

Vamos um passo adiante para mostrar como as expressões dos coeficientes de regressão são formadas pelas medidas estatísticas das séries de dados x e y , como mostrado a seguir.

A fórmula do coeficiente a que repetimos de forma diferente:

$$a = \frac{1}{n} \sum_{i=1}^n y_i - b \frac{1}{n} \sum_{i=1}^n x_i$$

mostra que é o resultado da soma algébrica de duas médias. A fórmula final em função das médias é $a = \bar{y} - b\bar{x}$

A fórmula do coeficiente b , que repetimos de forma diferente:

$$b = \frac{n \left(\sum_{i=1}^n x_i \times y_i - \frac{1}{n} \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i \right)}{n \left(n \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)}$$

Simplificando as parcelas dessa fórmula.

$$b = \frac{\frac{1}{n} \left(\sum_{i=1}^n x_i \times y_i - \bar{x} \times \bar{y} \right)}{\frac{1}{n} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)}$$

O numerador dessa expressão é a covariância σ_{xy} , como apresentado no Capítulo 6, e o denominador é a variância de σ_x^2 , como mostrado no Apêndice 2 do Capítulo 4. Resumindo, os coeficientes de regressão da função $y = a + bx$ são:

$$\begin{cases} a = \bar{Y} - b\bar{X} \\ b = \frac{\sigma_{xy}}{\sigma_x^2} \end{cases}$$

Como a covariância está relacionada com o coeficiente de correlação pela expressão $\sigma_{xy} = r_{xy} \sigma_x \sigma_y$, as expressões anteriores passam a ser as seguintes.

$$\begin{cases} a = \bar{Y} - b\bar{X} \\ b = \frac{r_{xy} \sigma_x \sigma_y}{\sigma_x^2} = r \frac{\sigma_y}{\sigma_x} \end{cases}$$

Apêndice 3

Outras funções estatísticas

Nos exemplos apresentados neste capítulo, foi obtida a maioria dos resultados de uma regressão linear. Neste apêndice, são apresentadas as funções estatísticas PROJ.LIN, PROJ.LOG e CRESCIMENTO do Excel, que fornecem todos os resultados em uma única célula de onde se podem extrair os resultados de interesse.

PROJ.LIN(*val_conhecidos_y; val_conhecidos_x; constante; estatística*)

A função estatística PROJ.LIN²⁰ retorna uma *matriz* com os resultados da reta de regressão linear múltipla $\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$ pelo método dos quadrados mínimos. O significado dos argumentos é:

- No argumento *val_conhecidos_y* devem ser informados os valores da amostra *y*, variável dependente.
- No argumento *val_conhecidos_x* devem ser informados os valores de uma ou mais amostras *x*, variáveis independentes, tendo em consideração que:
 - Se há apenas uma variável independente *x*, os intervalos das duas únicas variáveis *y* e *x* podem ter qualquer forma.
 - Se há mais de uma variável independente, o intervalo deve ser informado abrangendo todas as variáveis independentes juntas.
 - Se o intervalo da variável independente for omitido, a função assumirá que *x* é a *matriz* de números {1, 2, 3, ..., *n*}, com *n* igual ao número de valores da variável *y*.
- No argumento *constante* deve ser informado um dos dois valores lógicos seguintes:
 - VERDADEIRO (ou omitido), a função retornará todos os coeficientes *a* e *bs* da reta de regressão linear múltipla completa.
 - FALSO: a função retornará apenas os coeficientes *bs* da reta de regressão que passa pela origem $\hat{y} = b_1x_1 + b_2x_2 + \dots + b_nx_n$, com *a*=0.
- No argumento *estatística*, deve ser informado um dos dois valores lógicos seguintes:
 - FALSO: a função retornará somente os coeficientes *a* e *bs*.
 - VERDADEIRO (ou omitido), a função retornará os coeficientes *a* e *bs* e as seguintes estatísticas: erros padrões dos coeficientes *a* e *bs*; o coeficiente de determinação r^2 ; o erro padrão da estimativa S_e ; o *F observado*; o número de graus de liberdade *gl* da regressão; a soma dos quadrados dos desvios explicados SSR; a soma dos quadrados dos desvios não explicados SSE.

Neste argumento, como no anterior, em vez de VERDADEIRO, pode-se utilizar o valor 1 (um) e, em vez de FALSO, o valor 0 (zero).

Para compreender a utilização da função PROJ.LIN, serão apresentadas aplicações gradativas da regressão linear simples $\hat{y} = a + bx$ do Exemplo 15.1, utilizando a planilha **Função PROJ.LIN**, incluída na pasta **Capítulo 15** e apresentada na Figura 15.20.

²⁰ Em inglês, a função estatística PROJ.LIN é *LINEST*.

Obtenção de dois resultados

Informando VERDADEIRO, ou 1, no argumento *constante*, a função retornará os resultados da equação de regressão $\hat{y} = a + bx$, e informando FALSO, ou 0, no argumento *estatística*, a função retornará uma *matriz* com dois resultados, os coeficientes *a* e *b*.

- A fórmula =PROJ.LIN(B4:B13;C4:C13;VERDADEIRO;FALSO) registrada na célula F4 retornou o resultado 9,7381. Contudo, a função PROJ.LIN registrou uma *matriz* com dois resultados, mostrando na célula F4 o primeiro deles. Os dois resultados podem ser vistos procedendo como segue:
 - Selecione a célula F4.
 - Pressionando primeiro a tecla F2 e depois a tecla F9, será mostrada a matriz ={9,73814229249012.117,070158102767} com os dois resultados. O primeiro resultado é o coeficiente *b*, e o segundo resultado separado por um ponto, símbolo (.), é o coeficiente *a*.
 - Pressione a tecla Esc para manter a fórmula na célula F4.

Para separar o resultado retornado pela função em duas células, um resultado em cada célula, procede-se desta forma:

- Selecione a célula F5 e depois, mantendo pressionado o botão esquerdo do mouse, arraste o mouse até a célula G5, definindo o intervalo F5:G5.
- Insira a fórmula =PROJ.LIN(B4:B13;C4:C13;1;0), seja por digitação ou utilizando o assistente de função do Excel, sem pressionar a tecla Enter. Para inserir essa função como matriz, pressione simultaneamente as três teclas Ctrl + Shift + Enter; mantendo pressionada a tecla Ctrl, pressione e mantenha pressionada a tecla Shift e, por último, pressione a tecla Enter. Depois de pressionar as três teclas simultaneamente, as duas fórmulas receberam as chaves { }.

	A	B	C	D	E	F	G
1	Função PROJ.LIN						
2							
3		y	x		Dois resultados		
4		430	30		Função PROJ.LIN	9,7381	
5		335	21		Função PROJ.LIN	9,7381	117,0702
6		520	35				
7		490	42		Dez resultados		
8		470	37		Função PROJ.LIN	9,7381	
9		210	20		Função PROJ.LIN	9,7381	117,07
10		195	8			2,05	59,03
11		270	17			0,74	65,17
12		400	35			22,59	8,00
13		480	25			95,969	33,981
14							

FIGURA 15.20

A função PROJ.LIN resolvendo o Exemplo 15.1.

Obtenção de todos os resultados

Informando VERDADEIRO, ou 1, no argumento *constante*, a função retornará os resultados da equação de regressão $\hat{y} = a + bx$, e informando VERDADEIRO, ou 1, no argumento *estatística*, a função retornará uma *matriz* com dez resultados.

- A fórmula =PROJ.LIN(B4:B13;C4:C13;1;1), registrada na célula F8, retornou o resultado 9,7381. Entretanto, a função PROJ.LIN registrou uma *matriz* com dez resultados, mostrando na célula F8 o primeiro deles. Os dois resultados podem ser vistos procedendo como segue:
 - Selecione a célula F8.

- Pressionando primeiro a tecla F2 e depois a tecla F9, será mostrada a matriz

$$= \{9,73814229249012.117,070158102767;$$

$$2,04870921558875.59,0298507995893;$$

$$0,738510136917969.65,1734298885576;$$

$$22,593920183783.8;$$

$$95969,3922924901.33980,6077075099\}$$

com os dois resultados. O primeiro resultado é o coeficiente b , e o segundo resultado separado por um ponto, símbolo ($.$), é o coeficiente a . Os demais quatro pares de valores estão separados por ponto e vírgula, símbolo ($;$)

- Pressione a tecla Esc para manter a fórmula na célula F8.

Para separar o resultado retornado pela função em duas células, um resultado em cada célula, proceda-se como segue:

- Selecione a célula F9 e depois, mantendo pressionado o botão esquerdo do mouse, arraste o mouse até a célula G13, definindo o intervalo F9:G13.
- Insira a fórmula $=\text{PROJ.LIN}(B4:B13;C4:C13;1;1)$, seja por digitação ou utilizando o assistente de função do Excel, sem pressionar a tecla Enter. Para inserir essa função como matriz, pressione simultaneamente as três teclas **Ctrl + Shift + Enter**; mantendo pressionada a tecla **Ctrl**, pressione e mantenha pressionada a tecla **Shift** e, por último, pressione a tecla **Enter**. Depois de pressionar as três teclas simultaneamente, as dez fórmulas receberam as chaves $\{ \}$. As dez células da tabela, separadas em cinco grupos, têm o seguinte significado:

Coeficiente $b=9,7381$	Coeficiente $a=117,0702$
Erro padrão do coeficiente b $S_b=2,05$	Erro padrão do coeficiente a $S_a=59,03$
Coeficiente de determinação $r^2=0,7385$	Erro padrão da estimativa $S_e=65,17$
F observado 22,59	Graus de liberdade da regressão 8
Soma dos quadrados dos desvios explicados $SSR=95.969$	Soma dos quadrados dos desvios não explicados $SSE=33.981$

Se o argumento *constante* for FALSO, ou zero, a função PROJ.LIN retornará os resultados da reta de regressão $\hat{y} = b_x$, aplicando o método dos quadrados mínimos. Mudando o valor do argumento *constante*, você pode analisar o comportamento dessa função. Para completar, a partir da linha 16 da planilha **Função PROJ.LIN**, foram repetidos os cálculos anteriores informando as amostras como matrizes.

PROJ.LOG(*conhecidos_y; val_conhecidos_x; constante; estatística*)

A função estatística PROJ.LOG²¹ retorna uma *matriz* com os resultados da reta de regressão linear múltipla $\hat{y} = b \times m_1^{x_1} \times m_2^{x_2} \times \dots \times m_n^{x_n}$, pelo método dos quadrados mínimos. O significado dos argumentos é o seguinte:

²¹ Em inglês, a função estatística PROJ.LOG é LOGEST.

- No argumento *val_conhecidos_y*, devem ser informados os valores da amostra *y*, variável dependente.
- No argumento *val_conhecidos_x*, devem ser informados os valores de uma ou mais amostras *x*, variáveis independentes, levando em consideração que:
 - Se há apenas uma variável independente *x*, os intervalos das duas únicas variáveis *y* e *x* podem ter qualquer forma.
 - Se há mais de uma variável independente, o intervalo deve ser informado abrangendo todas as variáveis independentes juntas.
 - Se o intervalo da variável independente for omitido, a função assumirá que *x* é a *matriz* de números {1, 2, 3, ..., *n*}, com *n* igual ao número de valores da variável *y*.
- No argumento *constante*, deve ser informado um dos dois valores lógicos seguintes:
 - VERDADEIRO (ou omitido): a função retornará todos os coeficientes *a* e *bs* da reta de regressão linear múltipla completa.
 - FALSO: a função retornará apenas os coeficientes *bs* da reta de regressão que passa pela origem $\hat{y} = m_1^{x_1} \times m_2^{x_2} \times \dots \times m_n^{x_n}$, com *b*=1.
- No argumento *estatística*, deve ser informado um dos dois valores lógicos seguintes:
 - FALSO: a função retornará somente os coeficientes *a* e *bs*.
 - VERDADEIRO (ou omitido): a função retornará os coeficientes *a* e *bs* e as seguintes estatísticas: erros padrão dos coeficientes *a* e *bs*; o coeficiente de determinação r^2 ; o erro padrão da estimativa S_e ; o *F observado*; o número de graus de liberdade *gl* da regressão; a soma dos quadrados dos desvios explicados SSR; a soma dos quadrados dos desvios não explicados SSE.

Neste argumento, como no anterior, em vez de VERDADEIRO, pode-se utilizar o valor 1 (um) e, em vez de FALSO, o valor 0 (zero).

Como a forma de utilizar e os tipos de resultados da função PROJ.LOG são os mesmos que os da função PROJ.LIN, os detalhes da função PROJ.LOG não serão mostrados. Na planilha **Função PROJ.LOG**, incluída na pasta **Capítulo 15**, foi aplicada esta função no Exemplo 15.1, como mostra a Figura 15.21. A partir da linha 16 da planilha **Função PROJ.LOG**, foram repetidos os mesmos cálculos, informando as amostras como matrizes.

	A	B	C	D	E	F	G
1	Função PROJ.LOG						
2							
3		y	x		Dois resultados		
4		430	30		Função PROJ.LOG	1,0298	
5		335	21		Função PROJ.LOG	1,0298	162,7431
6		520	35				
7		490	42		Dez resultados		
8		470	37		Função PROJ.LOG	1,0298	
9		210	20		Função PROJ.LOG	1,0298	162,74
10		195	8			0,0061	0,1771
11		270	17			0,7411	0,1955
12		400	35			22,9029	8,0000
13		480	25			0,8754	0,3058
14							

FIGURA 15.21
A função PROJ.LOG resolvendo o Exemplo 15.1.

CRESCIMENTO(*val_y;val_x;novos_val_x;constante*)

A função estatística CRESCIMENTO²² retorna o valor projetado \hat{y} da curva exponencial de regressão, para um único ou um grupo de valores de *x*, denominado *xs*, quando são conhecidos valores das amostras *val_y* e *val_x*. Deve-se tomar o cuidado de fornecer os dados na ordem correta, o primeiro argu-

²² Em inglês, a função CRESCIMENTO é GROWTH.

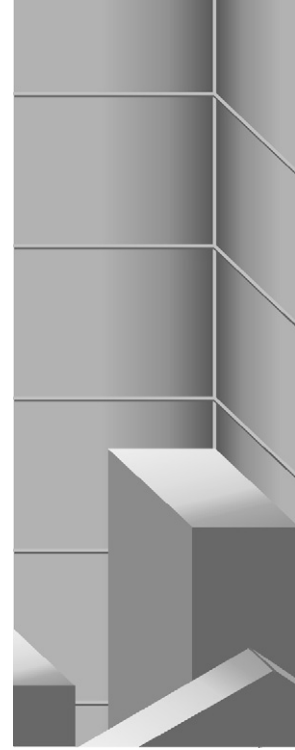
mento *val_y*, correspondente aos valores da variável dependente *y*, e o argumento *val_x*, correspondente aos valores da variável independente *x*. Se o argumento *constante* for:

- VERDADEIRO (ou omitido), a função fornecerá um único ou um grupo de valores da curva exponencial de regressão $\hat{y} = b \times m^x$.
- FALSO, a função fornecerá apenas o coeficiente *m* de regressão $\hat{y} = m^x$, e *b*=1.

A função CRESCIMENTO é equivalente à função TENDÊNCIA, apresentada no Exemplo 15.3. A partir da coluna I da planilha **Função PROJ.LOG**, incluída na pasta **Capítulo 15**, você encontra uma aplicação dessa função.

Capítulo 16

AJUSTE NÃO LINEAR



Realizar previsões ou projeções é uma das preocupações das atividades de negócios e governamentais. Nas empresas, é necessário prever as vendas, os estoques, os custos, o fluxo de caixa etc. para um determinado período, como é o orçamento anual do próximo ano. Na administração pública, é necessário prever o número de habitantes, a arrecadação, os custos dos serviços prestados etc. Essas previsões implicam estabelecer relações entre duas ou mais variáveis que tenham a habilidade de prever uma ou mais delas em função das demais. As previsões podem ser realizadas a partir do conhecimento dos dados de um corte transversal da população; por exemplo, amostras da quantidade produzida e do preço médio dos produtos, ou das vendas e do investimento em propaganda etc. Também essa relação pode ocorrer entre uma variável e o tempo, como o consumo de energia mensal, as vendas semanais de uma empresa, as exportações e importações mensais do país etc.

A análise da reta de regressão linear mostrou que nem todos os pares de valores das amostras estão incluídos na própria reta e, em alguns casos, esse afastamento pode insinuar um tipo de curva diferente de uma linha reta; por exemplo, o gráfico de dispersão dos pares de valores das amostras pode exibir a forma de uma curva exponencial ou de um polinômio de segundo grau. Neste capítulo, trataremos das previsões realizadas com o ajuste de funções não lineares transformadas em retas e, depois, das previsões futuras de observações coletadas periodicamente, ou em função do tempo.

Transformação de funções não lineares em lineares

Como muitos processos econômicos são mais bem explicados com funções matemáticas não lineares, foram desenvolvidos modelos não lineares que se tornam lineares depois de uma transformação com logaritmos¹, como mostrado na tabela da Figura 16.1. Na primeira linha dessa tabela, foi registrada a equação da regressão linear simples conhecida. Nas outras três linhas da tabela, estão registradas três funções não lineares e as transformações das variáveis x e y para torná-las funções lineares semelhantes à da primeira linha da tabela. Nas duas últimas colunas da tabela da Figura 16.1, são mostradas as transformações com logaritmos das variáveis x e y , esgotando as quatro combinações de logaritmos

¹ Nas fórmulas, foi aplicado o logaritmo natural; entretanto, em alguns casos o logaritmo pode ter base. O importante é aplicar corretamente as propriedades dos logaritmos.

com duas variáveis, incluindo a alternativa de não aplicar logaritmos. Para cada uma dessas equações, será apresentado o procedimento de ajuste de cada curva.

Tipo	Equação	Transformação	Variável x	Variável y
Linear	$\hat{y} = a + bx$	$\hat{y} = a + bx$	x	y
Exponencial	$\hat{y} = a \cdot e^{bx}$	$\ln \hat{y} = \ln a + bx$	x	$\ln y$
Logarítmica	$\hat{y} = a + b \cdot \ln x$	$\hat{y} = a + b \cdot \ln x$	$\ln x$	y
Potência	$\hat{y} = a \cdot x^b$	$\ln \hat{y} = \ln a + b \cdot \ln x$	$\ln x$	$\ln y$

FIGURA 16.1 Transformação de função não linear em linear.

Função exponencial

A função exponencial $\hat{y} = a \cdot e^{bx}$ é muito útil para os casos em que a variável dependente varia com uma taxa percentual constante. Aplicando logaritmos nos dois membros dessa função exponencial, temos a expressão linear $\ln \hat{y} = \ln a + bx$. Para realizar essa transformação, deve-se proceder como segue:

- Os valores da amostra y devem ser transformados em $\ln y$, formando a nova amostra com valores $\ln y$. Os valores da amostra x permanecem sem transformação.
- Com os valores das novas variáveis x e $\ln y$:
 - Calcule os coeficientes de regressão, intercepto h e declividade k. Foram adicionadas as novas constantes h e k para distingui-las dos coeficientes da função exponencial a e b.
 - Calcule o coeficiente de determinação r^2 .
- Calcule os coeficientes da função exponencial a e b, tendo presente:
 - Como o intercepto h da reta é $\ln a$, o coeficiente $a = e^h$.
 - A declividade k é o próprio coeficiente $b = k$.

EXEMPLO 16.1

O departamento de vendas da rede de varejo relacionou as vendas anuais y com o investimento anual x em propaganda, ambos em milhões, cujos valores estão registrados na tabela seguinte. Ajuste a curva da função exponencial $\hat{y} = a \cdot e^{bx}$.

x	30	21	35	42	37	20	8	17	35	25
y	430	335	520	490	470	210	195	270	400	480

Solução. Este exemplo foi resolvido na planilha **Função exponencial**, incluída na pasta **Capítulo 16**, como mostra a próxima figura.

- No intervalo B4:C14, foram registrados os valores das variáveis x e y, incluindo as duas células de títulos.
- Na célula D5, foi registrada a fórmula =B5 que, depois, foi copiada até a célula D14.
- Na célula E5, foi registrada a fórmula =LN(C5), e, depois que foi copiada até a célula E14.

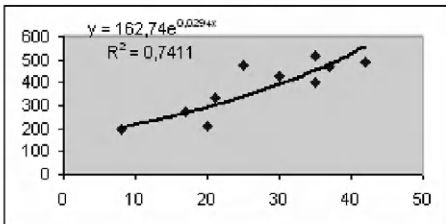
• LN(número)

A função LN retorna o logaritmo natural do argumento *número*, que pode ser um número, fórmula ou referência, com a condição de que seu resultado seja um número real positivo. A função LN é a inversa da função EXP. A base do logaritmo natural é a constante $e=2,71828...$

	A	B	C	D	E	F	G	H	I	J															
1	Ajuste Exponencial																								
2																									
3	Amostras				Ajuste Exponencial				<div>Resultados</div> <table><tr><td><i>h</i></td><td>5,0922</td><td>=INTERCEPÇÃO(E5:E14;D5:D14)</td></tr><tr><td><i>k</i></td><td>0,0294</td><td>=INCLINAÇÃO(E5:E14;D5:D14)</td></tr><tr><td><i>r-quadrado</i></td><td>0,7411</td><td>=RQUAD(E5:E14;D5:D14)</td></tr><tr><td><i>a</i></td><td>162,7431</td><td>=EXP(H4)</td></tr><tr><td><i>b</i></td><td>0,0294</td><td>=H5</td></tr></table>		<i>h</i>	5,0922	=INTERCEPÇÃO(E5:E14;D5:D14)	<i>k</i>	0,0294	=INCLINAÇÃO(E5:E14;D5:D14)	<i>r-quadrado</i>	0,7411	=RQUAD(E5:E14;D5:D14)	<i>a</i>	162,7431	=EXP(H4)	<i>b</i>	0,0294	=H5
<i>h</i>	5,0922	=INTERCEPÇÃO(E5:E14;D5:D14)																							
<i>k</i>	0,0294	=INCLINAÇÃO(E5:E14;D5:D14)																							
<i>r-quadrado</i>	0,7411	=RQUAD(E5:E14;D5:D14)																							
<i>a</i>	162,7431	=EXP(H4)																							
<i>b</i>	0,0294	=H5																							
4	<i>x</i>	<i>y</i>	<i>x</i>	<i>ln y</i>																					
5	30	430	30	6,06																					
6	21	335	21	5,81																					
7	35	520	35	6,25																					
8	42	490	42	6,19																					
9	37	470	37	6,15																					
10	20	210	20	5,35																					
11	8	195	8	5,27																					
12	17	270	17	5,60																					
13	35	400	35	5,99																					
14	25	480	25	6,17																					
15																									
16																									
17																									
18																									

$y = 162,74e^{0,0294x}$

$R^2 = 0,7411$



Os valores do intervalo D5:E14 são os valores das novas variáveis utilizados para realizar a regressão linear. Para isso:

- O intercepto h da reta de regressão é calculado na célula H4 com a fórmula =INTERCEPÇÃO(E5:E14;D5:D14).
- A declividade k da reta de regressão é calculada na célula H5 com a fórmula =INCLINAÇÃO(E5:E14;D5:D14).
- O coeficiente de determinação r -quadrado é calculado na célula H6 com a fórmula =RQUAD(E5:E14;D5:D14).

Com os coeficientes h e k da reta de regressão, o próximo passo é calcular os coeficientes a e b da função exponencial $\hat{y} = a \cdot e^{bx}$, procedendo assim:

- Pela definição de logaritmo, se $h = \ln a$, então o coeficiente $a = e^h$. Na célula H7, foi registrada a fórmula =EXP(E16), que retornou o valor $a = 162,74$.
- **EXP(número)**
A função EXP retorna a constante $e = 2,71828...$ elevada ao argumento *número*, que pode ser um número, uma fórmula ou uma referência. A função EXP é a inversa da função LN.
- Como o coeficiente b é a própria declividade k , na célula H8 foi registrada a fórmula =H5, que retornou o valor do coeficiente $b = 0,0294$.

Para visualizar o ajuste da curva exponencial, foram adicionados os gráficos de dispersão dos pares de valores das variáveis e a curva exponencial, com seus resultados obtidos com o comando *Linha de tendência*.

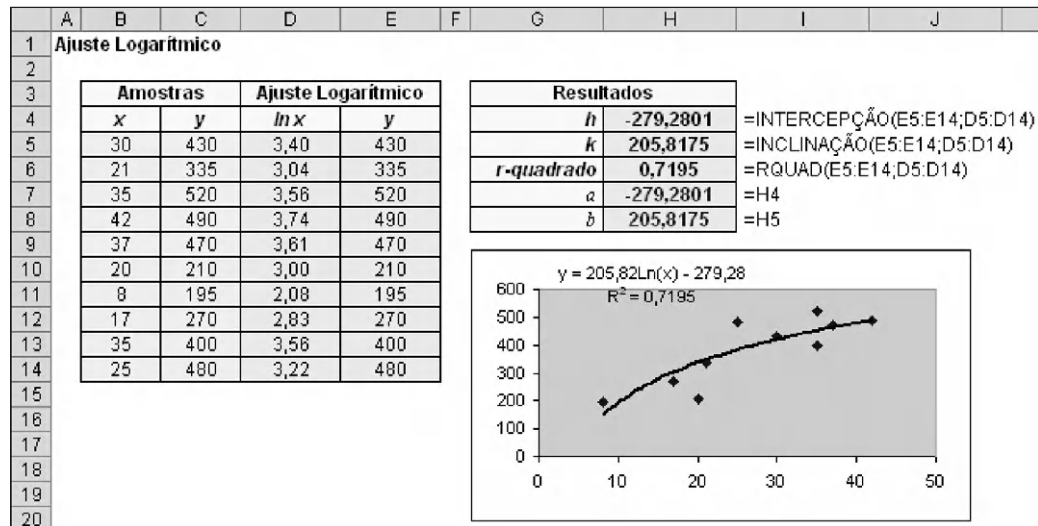
Função logarítmica

A função logarítmica $\hat{y} = a + b \cdot \ln x$ já é uma expressão linear. Entretanto, os valores da variável x devem ser transformados:

- Os valores da amostra y permanecem sem transformação, e os valores da amostra x devem ser transformados em $\ln x$, formando a nova amostra com valores $\ln x$.
- Com os valores das novas variáveis y e $\ln x$:
 - Calcular os coeficientes de regressão, intercepto h e declividade k para distingui-los dos coeficientes da função exponencial a e b .
 - Calcular o coeficiente de determinação r^2 .
- Calcular os coeficientes da função logarítmica a e b , considerando que:
 - O intercepto h é o próprio coeficiente $a = h$.
 - A declividade k é o próprio coeficiente $b = k$.

EXEMPLO 16.2

Continuando com o Exemplo 16.1. Ajuste a curva da função logarítmica $\hat{y} = a + b \cdot \ln x$.



Solução. Este exemplo foi resolvido na planilha **Função logarítmica**, incluída na pasta **Capítulo 16**, como mostra a figura acima.

- No intervalo B4:C14, foram registrados os valores de x e y , incluindo as duas células de títulos.
- Na célula D5, foi registrada a fórmula $=LN(B5)$, que retorna o logaritmo natural do valor da célula B5. Essa fórmula foi copiada até a célula D14.
- Na célula E5, foi registrada a fórmula $=C5$, que depois foi copiada até a célula E14.

Os valores do intervalo D5:E14 são os valores das novas variáveis, que serão utilizados para realizar a regressão linear. Para isso:

- O intercepto h da reta de regressão é calculado na célula H4 com a fórmula $=INTERCEPÇÃO(E5:E14;D5:D14)$.
- A declividade k da reta de regressão é calculada na célula H5 com a fórmula $=INCLINAÇÃO(E5:E14;D5:D14)$.
- O coeficiente de determinação r -quadrado é calculado na célula H6 com a fórmula $=RQUAD(E5:E14;D5:D14)$.

Com os coeficientes h e k da reta de regressão, o próximo passo é calcular os coeficientes a e b da função logarítmica $\hat{y} = a + b \cdot \ln x$, procedendo desta forma:

- Como $h=a$, na célula H7, foi registrada a fórmula $=H4$, que retorna o valor $a=-279,2801$.
- Como $k=b$, na célula H8, foi registrada a fórmula $=H5$, que retorna o valor $b=205,8175$.

Para visualizar o ajuste da curva logarítmica, foi adicionado o gráfico de dispersão dos pares de valores das variáveis e a curva exponencial, com seus resultados obtidos, com o comando *Linha de tendência*.

Função potência

A função potência $\hat{y} = a \cdot x^b$ é muito útil para negócios, principalmente a *curva de aprendizado*. Aplicando logaritmos nos dois membros da função potência, temos a expressão linear $\ln \hat{y} = \ln a + b \cdot \ln x$. Para realizar essa transformação, deve-se proceder assim:

- Os valores da amostra y devem ser transformados em $\ln y$, formando a nova amostra com valores $\ln y$.
- Os valores da amostra x devem ser transformados em $\ln x$, formando a nova amostra com valores $\ln x$.
- Com os valores das novas variáveis $\ln y$ e $\ln x$:
 - Calcule os coeficientes de regressão, intercepto h e declividade k para distingui-los dos coeficientes da função potência a e b .
 - Calcule o coeficiente de determinação r^2 .
- Calcule os coeficientes da função potência a e b , considerando que:
 - Se o intercepto h da reta é $\ln a$, então o coeficiente $a=e^h$.
 - A declividade k é o próprio coeficiente $b=k$.

Na planilha **Função potência**, incluída na pasta **Capítulo 16**, você encontra a transformação potência referente às amostras do Exemplo 16.1. Para visualizar o ajuste da curva logarítmica, foi adicionado o gráfico de dispersão dos pares de valores das variáveis e a curva exponencial com seus resultados obtidos com o comando **Linha de tendência**.

Resumo das transformações

A tabela da Figura 16.2 foi copiada da planilha **Resumo das transformações**, incluída na pasta **Capítulo 16**.

	Linear	Exponencial	Logarítmica	Potência
<i>Intercepto h</i>	117,0702	5,0922	-279,2801	3,8569
<i>Declividade k</i>	9,7381	0,0294	205,8175	0,6336
<i>Coefficiente a</i>	117,0702	162,7431	-279,2801	47,3165
<i>Coefficiente b</i>	9,7381	0,0294	205,8175	0,6336
<i>r-quadrado</i>	0,7385	0,7411	0,7195	0,7501

FIGURA 16.2 Resumo dos resultados das transformações lineares.

Substituindo os coeficientes a e b nas respectivas funções matemáticas temos as equações das funções ajustadas.

- Função Linear: $\hat{y} = a + bx = 117,0702 + 9,7381x$.
- Função Exponencial: $\hat{y} = a \times e^{bx} = 162,7431 \times e^{0,0294x}$.
- Função Logarítmica: $\hat{y} = a + b \times \ln x = -279,2801 + 205,8175 \times \ln x$.
- Função Potência: $\hat{y} = ax^b = 47,3165x^{0,6336}$.

Qual dessas curvas deve ser escolhida? Se as premissas da regressão linear foram atendidas pelas quatro transformações, deve-se escolher a curva com maior coeficiente de determinação. No exemplo que está sendo desenvolvido, a curva que melhor explica é a *potência*, pois seu coeficiente de determinação é o maior das quatro regressões analisadas. De forma geral, os exemplos apresentados mostraram que a transformação das variáveis relacionados de forma não linear cria novas variáveis relacionadas de forma linear, que podem ser analisadas dentro do modelo de regressão linear. Foi visto que a transformação das funções exponencial, logarítmica e potência permite utilizar o modelo de regressão linear simples, apesar de não ser linear a relação entre as variáveis originais. Essa ideia é estendida para o modelo de regressão linear múltipla; por exemplo, transformando a relação não linear de mais de duas variáveis num polinômio de grau n . Neste livro, será mostrado o comando *linha de tendência* para ajustar um polinômio.

Linha de tendência do Excel

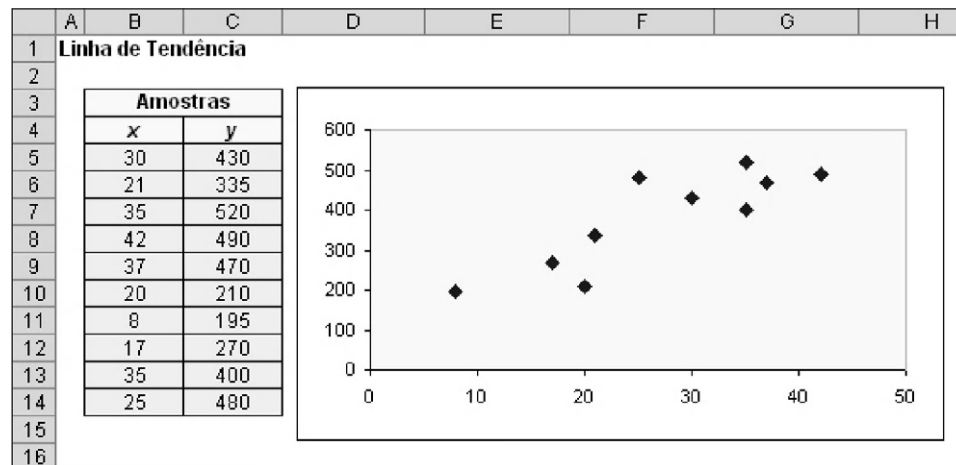
As transformações anteriores foram realizadas utilizando os recursos das funções estatísticas e o registro de fórmulas na planilha Excel. Com o comando *Linha de tendência*² do Excel, é possível realizar essas e outras transformações dentro do ambiente de gráficos do Excel, tais como gráficos de áreas 2-D não empilhadas, barras, colunas, linhas, ações, dispersão (xy) e bolhas. Para construir a linha de ten-

² Em inglês, o comando *Linha de Tendência* é *Trendline*.

dência em uma planilha Excel, deve-se registrar a tabela com os dados das duas amostras e depois construir o gráfico de dispersão, procedimento realizado na planilha **Linha de tendência** da pasta **Capítulo 16**, mostrado na Figura 16.3.

Para construir a linha de tendência, siga o procedimento a seguir. Selecione a trajetória dos pontos do gráfico clicando uma vez num dos pontos do gráfico, verificando que os pontos do gráfico mudaram de cor.

FIGURA 16.3 Gráfico de dispersão do Exemplo 16.1.



- No menu **Gráfico**, escolha **Adicionar linha de tendência**.
- Em vez de utilizar o menu, com o cursor dentro do gráfico, clique com o botão direito do mouse para selecionar **Adicionar linha de tendência**.

Em ambos os casos, o Excel exibirá a caixa de diálogo **Adicionar linha de tendência** contendo duas guias, **Tipo** e **Opções**, Figura 16.4.

No quadro **Tendência/tipo de regressão** da guia **Tipo**, selecione o tipo de curva de ajuste clicando no desenho da curva desejada. Há seis tipos de ajustes, cinco tipos de curvas de ajuste e uma curva de média móvel. Os tipos de ajuste são:

- **Linear**. É a reta de regressão linear simples.
- **Logarítmica**. É a função logarítmica já apresentada.
- **Exponencial**. É a função exponencial já apresentada.
- **Potência**. É a função potência já apresentada e selecionada neste caso.
- **Polinomial**. O ajuste é realizado com um polinômio cujo grau é escolhido pelo usuário, de um a seis.
- **Média móvel**. Ajusta uma curva de média móvel, na qual o usuário define a quantidade de valores da variável para calcular a média móvel.

Na guia **Opções** da caixa de diálogo, são completadas as informações para construção da linha de tendência desejada, conforme mostra a Figura 16.5. No quadro **Linha de tendência**, é possível registrar um nome para a linha de tendência construída. Há duas alternativas de escolha:

- **Automática**: A linha de tendência construída recebe o nome do tipo de regressão escolhido na guia **Tipo**. Se no menu **Gráfico-Opções de gráfico-Legenda** for escolhido **Mostrar legenda**, o gráfico mostrará o nome registrado junto com a sequência de valores e o nome do tipo de linha de tendência. Como neste exemplo não foi registrado nenhum nome durante a construção do gráfico, o nome registrado será **Potência** para a linha de tendência e **Sequência1** para os pontos do gráfico de dispersão.

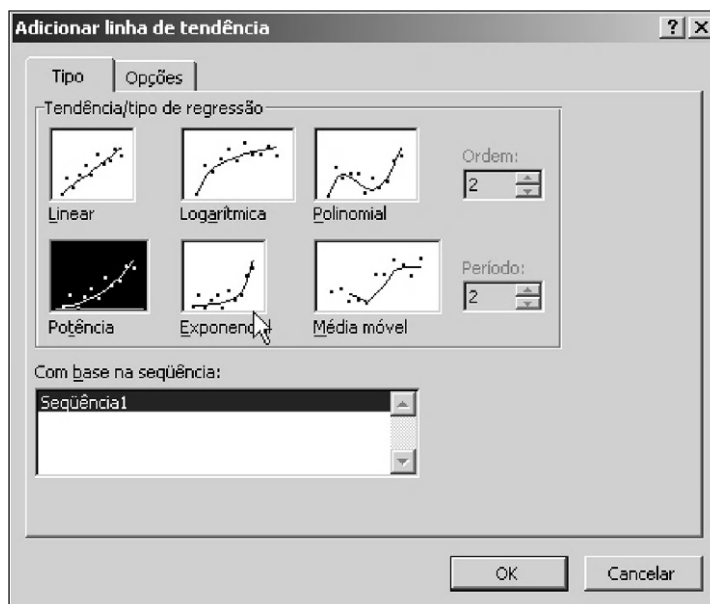


FIGURA 16.4 Caixa de diálogo da **Linha de tendência, Tipo**.

- **Personalizar:** A linha de tendência construída recebe o nome registrado pelo leitor, por exemplo, neste caso *Exemplo 16.1*.

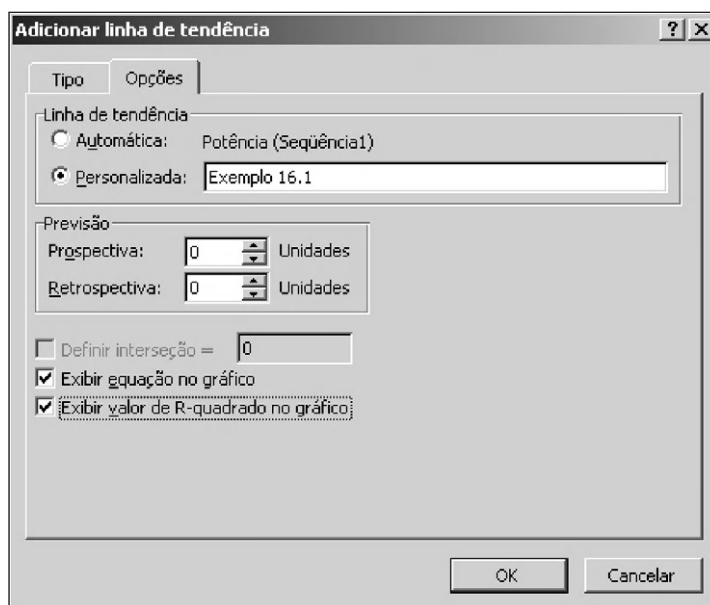


FIGURA 16.5 Caixa de diálogo do comando **Linha de tendência, Opções**.

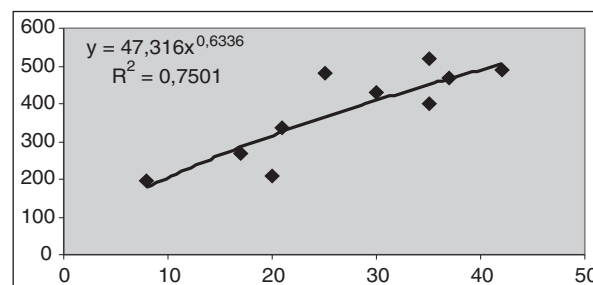
Como a linha de tendência é construída no intervalo dos pontos desenhados no gráfico de dispersão, no quadro **Previsão** será possível incluir mais pontos, antes e depois do intervalo dos dados. Essa alternativa está disponível somente para as curvas de ajustes de regressão e não se aplica ao ajuste da média móvel.

- **Prospecção:** Deve-se informar a quantidade de períodos, ou unidades do gráfico de dispersão xy , que o comando linha de tendência incluirá depois do limite superior do intervalo de dados. Na regressão linear, recomenda-se que a projeção da variável dependente y seja limitada ao intervalo dos valores da variável independente x .

- **Retrospectiva:** É equivalente a **Prospectiva**, porém antes do limite inferior do intervalo de dados.
- **Definir interseção:** Pode ser definido o ponto no qual a curva ajustada interceptará o eixo y. Está disponível apenas para alguns tipos de regressão.
- **Exibir equação no gráfico:** Exibe a equação da reta ajustada no gráfico de dispersão.
- **Exibir valor de R-quadrado no gráfico:** Exibe o valor do coeficiente de determinação no gráfico de dispersão.

A Figura 16.5 apresenta as escolhas realizadas na guia **Opções** da caixa de diálogo do comando linha de tendência. Depois de pressionar **OK**, o Excel constrói a curva ajustada e registra no mesmo quadro sua equação e o coeficiente de determinação. Esses valores estão registrados em um bloco que pode ser mudado de posição, como foi feito na Figura 16.6. Sugerimos que você construa as outras três curvas de ajuste e compare com os resultados obtidos na tabela da Figura 16.2.

FIGURA 16.6 Curva de ajuste potência, Exemplo 16.1.

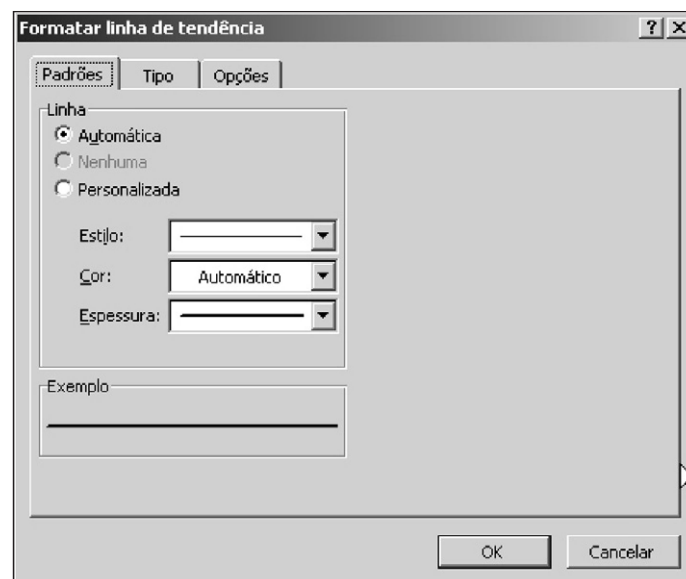


É possível modificar as definições da linha de tendência depois de construída, procedendo assim:

- Clique em qualquer ponto da linha de tendência construída e depois, mantendo o cursor dentro do gráfico, clique com o botão direito do mouse e selecione **Formatar linha de tendência**.
- A alternativa mais rápida é clicar duas vezes seguidas com o botão esquerdo do mouse em qualquer ponto da linha de tendência.

Nos dois casos, o Excel exibirá a caixa de diálogo **Formatar linha de tendência** da Figura 16.7. Essa caixa de diálogo contém três guias, as duas guias **Tipo** e **Opções** conhecidas mais a guia **Padrões**, onde será possível alterar o estilo, a cor e a espessura da linha de tendência.

FIGURA 16.7 Caixa de diálogo de **Formatar linha de tendência**.



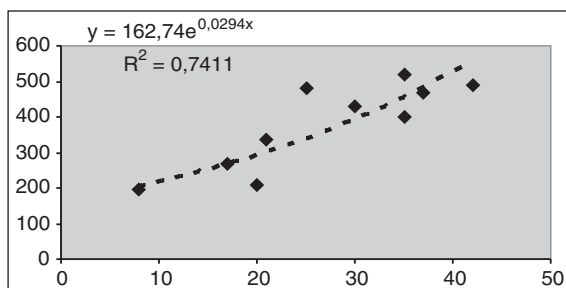


FIGURA 16.8 Linha de tendência do Exemplo 16.1, função exponencial.

Como exemplo, a Figura 16.8 mostra modificações no estilo e na espessura da linha de tendência da Figura 16.6, mudando, também, o tipo de regressão, de potência para exponencial.

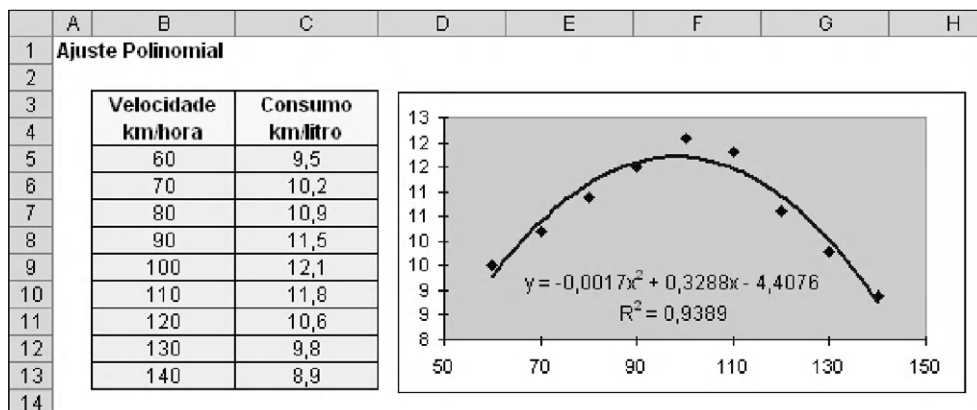
Ajuste polinomial

Um polinômio é uma função do tipo $\hat{y} = a + b_1x + b_2x^2 + \dots + b_nx^n$. Uma linha de tendência polinomial pode ajustar uma curva quando os dados têm muitas variações. A escolha da ordem da polinomial pode ser determinada pelo próprio perfil que os dados sugerem em um gráfico de dispersão. Por exemplo, uma linha de tendência polinomial de segundo grau possui apenas um máximo ou um mínimo relativo, pois se trata de uma parábola. Um polinômio de terceiro grau geralmente possui um ou dois máximos ou mínimos relativos. Um polinômio de quarto grau pode possuir até três máximos ou mínimos relativos. No entanto, é importante lembrar que, em geral, o ajuste será realizado com um ramo da curva do polinômio. Uma função polinomial de segundo grau $\hat{y} = a + b_1x + b_2x^2$ é muito útil para modelar curvas de custos, como mostra o Exemplo 16.3.

EXEMPLO 16.3

O gerente de projeto do novo motor realizou testes de consumo de combustível em função da velocidade do protótipo de carro que utilizará esse motor. Ajuste a curva polinomial adequada aos dados registrados na planilha **Exemplo 16.4**, incluída na pasta **Capítulo 16**.

Solução. O ajuste polinomial foi realizado na planilha **Ajuste polinomial**, incluída na pasta **Capítulo 16**, considerando a linha de tendência polinomial de segundo grau, como mostra a figura a seguir.



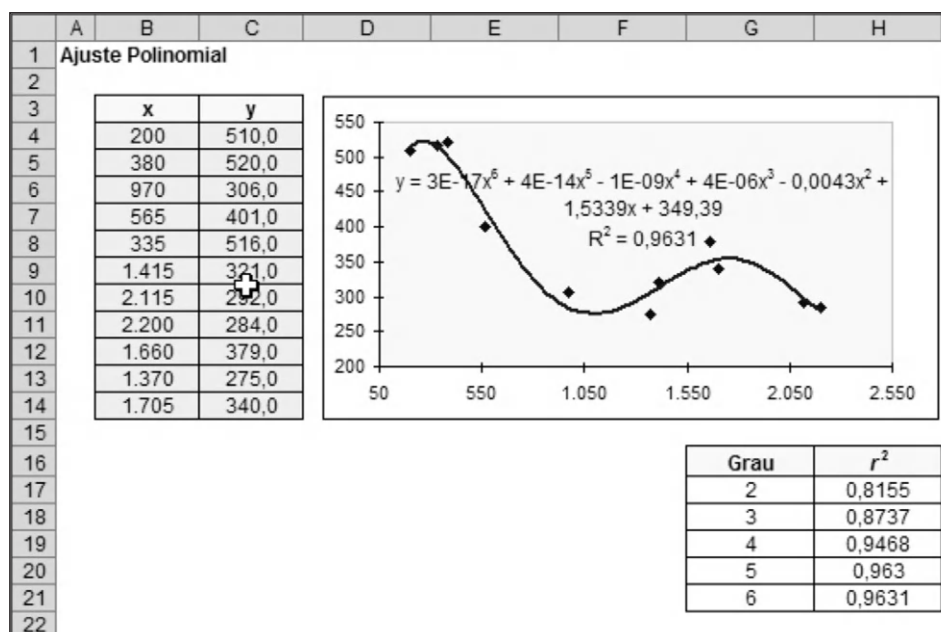
O polinômio de ajuste é $\hat{y} = -0,0017 \cdot x^2 + 0,3288 \cdot x - 4,4076$, que corresponde à equação de polinômio de segundo grau denominada parábola. O *R-quadrado* dessa curva ajustada é 93,89%. Então:

- Descreva o procedimento de ajuste da polinomial de segundo grau utilizando o comando linha de tendência.
- Repita o ajuste com curvas polinomiais com graus maiores do que dois, verificando que os melhores ajustes são realizados com curvas polinomiais com graus *pares* maiores do que dois.

EXEMPLO 16.4

Ajuste a curva polinomial adequada aos dados registrados na planilha **Ajuste polinomial II**, incluída na pasta **Capítulo 16**.

Solução. O gráfico de dispersão da figura seguinte mostra o ajuste com um polinômio de sexto grau que apresentou o maior coeficiente de determinação $R^2=0,9631$. Esse resultado foi conseguido depois de tentar manualmente as cinco alternativas disponíveis, do polinômio grau dois até o polinômio de grau seis, utilizando a caixa de diálogo **Formatar linha de tendência**. Os resultados das cinco tentativas estão registrados no intervalo G16:H21 da planilha **Ajuste polinomial II**, onde se pode verificar que o ajuste com o polinômio de quinto grau tem praticamente o mesmo valor de R^2 .



Séries temporais

Iniciamos este capítulo dizendo que realizar previsões ou projeções é uma das preocupações das atividades de negócios e governamentais. Em geral, as previsões são realizadas com dois tipos de observações. No primeiro grupo, estão incluídos os dados coletados em um determinado período, por exemplo, durante uma hora, um dia, uma semana, um mês, três anos etc. Embora não tenham sido coletados no mesmo instante, esses dados não sofrerão influência do tempo decorrido entre eles, aceitando-se que o prazo da coleta de informações é adequado para atender ao objetivo da pesquisa. Esses dados serão utilizados para realizar previsões que não dependerão do tempo. Por exemplo, a previsão do consumo de combustível para uma velocidade de 105 km por hora do novo motor do Exemplo 16.3. Outro exemplo, na previsão dos resultados de uma pesquisa de opinião, a demora de uma semana para coletar os dados não influirá nas inferências que serão realizadas a partir dos resultados da pesquisa; entretanto, um prazo de seis meses poderá não ser adequado. O outro grupo de observações inclui os dados co-

letados periodicamente, por exemplo, as vendas diárias da loja, a taxa de inflação mensal, as cotações da bolsa, cada trinta minutos etc. Esses dados formam uma série temporal, pois são periodicamente coletados, e a variável de interesse y está associada à variável tempo t ou à variável dependente y e à variável independente t . Dessa maneira, y_t é o valor da variável y no tempo t , como mostrado a seguir.

Valores coletados				Valores projetados			
...	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$
...	y_{t-3}	y_{t-2}	y_{t-1}	y_t	y_{t+1}	y_{t+2}	y_{t+3}

O objetivo é projetar o valor \hat{y}_{t+1} a partir do conhecimento dos valores coletados y até o tempo t , descrito com a função geral $\hat{y}_{t+1} = f(y_t, y_{t-1}, y_{t-2}, \dots)$. Como realizar as projeções ou que função utilizar para realizar a melhor projeção? Há diversas formas de realizar projeções, das quais destacamos três grupos que serão apresentados a seguir: Taxa média de crescimento, Regressão e Média móvel.

Procedimento inicial

O procedimento de projeção *simples*³ considera que o valor do próximo período $t+1$ é o do período anterior t utilizando a função $\hat{y}_{t+1} = y_t$. Na planilha **Modelo simples**, incluída na pasta **Capítulo 16**, foi construído o modelo de projeção das vendas diárias de uma empresa utilizando o procedimento *simples*. No intervalo de células B4:C16 dessa planilha, estão registradas as vendas diárias da empresa em milhares durante os últimos doze meses, como mostra a Figura 16.9. Por exemplo, a projeção das vendas em $t=13$ é 308. O erro de previsão do procedimento *simples* pode ser medido realizando as previsões dos dados conhecidos, como mostrado na própria planilha:

- No intervalo D6:D16, foram realizadas as projeções diárias partindo de $t=2$, por exemplo, a projeção $\hat{y}_2 = y_1 = 295$. Na célula D6, foi registrada a fórmula =C5, que depois foi copiada até a célula D17.
- No intervalo E6:E16, foram calculados os erros das projeções com a fórmula $e_2 = y_2 - \hat{y}_2 = 305 - 295 = 10$. Na célula E6, foi registrada a fórmula =C6-D6, que depois foi copiada até a célula E16.

	A	B	C	D	E	F	G
1	Projeções iniciais						
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

t	Vendas	Projeção		Projeção	
		Simples	Erro	Tendência	Erro
1	295				
2	305	295	10		
3	316	305	11	315	1
4	298	316	-18	327	-29
5	305	298	7	280	25
6	310	305	5	312	-2
7	316	310	6	315	1
8	306	316	-10	322	-16
9	301	306	-5	296	5
10	295	301	-6	296	-1
11	312	295	17	289	23
12	308	312	-4	329	-21
$t+1=13$		308		304	

FIGURA 16.9 Modelo simples de projeção, incluindo tendência.

3 Em inglês, denominado *Naive*.

Embora seja fácil de calcular, a projeção simples de $t+1$ não leva em consideração nenhuma possível tendência das vendas, a variável y , pois utiliza somente o valor observado no período anterior t . Tentando incluir a tendência dos valores da série antes da data de projeção, a projeção pode ser melhorada considerando os valores dos dois períodos anteriores ao da projeção $t+1$, utilizando a fórmula $\hat{y}_{t+1} = y_t + (y_t - y_{t-1})$. O resultado dessa projeção é o valor de t mais o resultado da diferença do valor de t menos o valor de $t-1$. Essa projeção foi realizada nas colunas F e G da planilha a partir da projeção em $t=3$, como você pode ver na Figura 16.9.

- No intervalo F7:F17, foram realizadas as projeções diárias partindo de $t=3$, por exemplo, $\hat{y}_3 = 305 + (305 - 295) = 315$. Na célula F7, foi registrada a fórmula $=C6+(C6-C5)$, que depois foi copiada até a célula F17.
- Na coluna ao lado, foram calculados os erros das projeções utilizando a mesma fórmula da projeção anterior.

Taxa média

Mesmo que seja fácil de calcular, a projeção simples de $t+1$ incluindo tendência utiliza somente os valores observados em $t-1$ e t . A projeção pela taxa média é um procedimento que utiliza todos os dados disponíveis, ou parte desses dados. O cálculo da taxa média é fácil; entretanto, esse procedimento necessita de atenção para ser aplicado, como mostra o Exemplo 16.5.

EXEMPLO 16.5

A planilha **Exemplo 16.5**, incluída na pasta **Capítulo 16**, registra a rentabilidade de uma carteira de investimento durante dez meses, do mês $t-9$ até o mês t , medida com a taxa mensal de juros i . O objetivo é projetar a taxa de juros para o mês $t+1$.

Solução. Conhecidas as taxas mensais de juros de dez meses, parece sensato calcular a taxa média mensal utilizando a média aritmética das dez taxas de juros mensais. No entanto, quando as variações são significativas, a média da taxa de juros retornará um resultado superior ao que seria obtida na prática financeira, como é o valor da média aritmética 0,44% ao mês calculada na célula G4 da planilha. O procedimento recomendado é utilizar juros compostos, que calculam a taxa equivalente de juros i utilizando a fórmula seguinte, onde Mg é o resultado da média geométrica das taxas de juros mais um:

$$Mg = \left(\prod_{j=1}^{10} (1+i_j) \right)^{\frac{1}{10}}$$

$$Mg = ((1+i_1) \times (1+i_2) \times \dots \times (1+i_{10}))^{\frac{1}{10}}$$

$$i = Mg - 1$$

	A	B	C	D	E	F	G	H	I	J
1	Projeção da taxa de juro									
2										
3										
4		Mês	i	$i+1$	Produto	Outros resultados				
5		t-9	1,10%	1,0110	1,0110	0,44%	=MÉDIA(C4:C13)			
6		t-8	2,00%	1,0200	1,0312	0,42%	=MÉDIA.GEOMÉTRICA(D4:D13)-1			
7		t-7	-1,45%	0,9855	1,0163	0,42%	=VFPLANO(1;C4:C13)^(1/10)-1			
8		t-6	2,00%	1,0200	1,0366					
9		t-5	-2,50%	0,9750	1,0107					
10		t-4	1,36%	1,0136	1,0244					
11		t-3	-1,50%	0,9850	1,0091					
12		t-2	2,87%	1,0287	1,0380					
13		t-1	2,50%	1,0250	1,0640					
14		t	-2,00%	0,9800	1,0427					
15		t+1	0,42%							

A figura anterior mostra o resultado da taxa de juros para o mês $t+1$ igual a 0,42% ao mês, célula C14. Observe que, para conseguir esse resultado, foram obtidos os seguintes resultados intermediários:

- Na célula D4, foi registrada a fórmula $=1+C4$, que retorna o valor da taxa de juros mais um. Depois, essa fórmula foi copiada até a célula D13.
- Na célula E4, foi registrada a fórmula $=D4$. Na célula E5, foi registrada a fórmula $=E4*D5$, que acumula o produto da soma da taxa mais um. Depois, essa fórmula foi copiada até a célula E13. O valor retornado pela célula E13 é o resultado da fórmula $(1 + i_1) \times (1 + i_2) \times \dots \times (1 + i_{10})$.
- Na célula C14, foi registrada a fórmula $=E13^{(1/10)}-1$, que retorna a taxa mensal de juros 0,42%.

O mesmo resultado pode ser obtido com funções do Excel. Por exemplo:

- A fórmula $=MÉDIA.GEOMÉTRICA(D4:D13)-1$ registrada na célula G5 utiliza a função estatística MÉDIA.GEOMÉTRICA, apresentada no Capítulo 3. Observe que essa função não utiliza os valores das taxas de juros, e sim os valores do intervalo D4:D13, que correspondem ao resultado de somar um ao valor de cada taxa de juro. Como essa função retorna o valor da taxa de juros mais um, deve-se subtrair o valor um para obter a projeção da taxa.
- A fórmula $=VFPLANO(1;C4:C13)^{(1/10)}-1$ registrada na célula G6.⁴

• VFPLANO(*capital;plano*)

A função financeira VFPLANO retorna o valor futuro de um *capital* inicial, sujeito a capitalizações periódicas com valores de taxas de juros definidas no argumento *plano*. A função financeira VFPLANO calcula o futuro F da fórmula conhecida $F = P \times \prod_{j=1}^n (1 + i_j)$, sendo conhecidos o capital inicial P e o plano das taxas

de juros das n operações elementares. Se *capital* for igual a 1, a função VFPLANO retornará o valor $(1+i)$. Nesse caso, a taxa total de juros da operação poderá ser obtida com a fórmula $=VFPLANO(1;plano)-1$. Observe que essa função utiliza os valores das taxas de juros do intervalo C4:C13.

A projeção utilizando taxa média de uma série de dados temporais é de fácil aplicação, não requerendo cálculos complexos, e pode ser útil para obter de forma rápida uma estimativa aproximada. Esse procedimento de projeção pode ser aplicado a qualquer tipo de série; entretanto, é mais recomendado para séries que apresentem tendência de crescimento positivo ou negativo ou cíclico, mas com pouca volatilidade. A projeção deve ser aceita como tentativa e não deve ser utilizada para mais de um período.

EXEMPLO 16.6

Com o procedimento de taxa média, projete as vendas em $t+1$ das vendas registradas na Figura 16.9.

Solução. Na planilha **Exemplo 16.6**, incluída na pasta **Capítulo 16**, foi resolvido este exemplo, como mostra a figura seguinte. Na resolução deste exemplo, foi utilizado o procedimento de taxa média apresentado no Exemplo 16.5. A projeção das vendas em $t+1$ é 309,2, resultado obtido da seguinte forma:

- Na célula D5, foi registrada a fórmula $=C5/C4-1$, que retorna o resultado da taxa de crescimento do dia 2 com relação ao dia 1. Depois, essa fórmula foi copiada até a célula D15.
- Na célula E5, foi registrada a fórmula $=1+D5$, que retorna o resultado de somar um ao valor de cada taxa de juro calculada em D5. Depois, essa fórmula foi copiada até a célula E15.
- A fórmula $=C15*(1+(MÉDIA.GEOMÉTRICA(E5:E15)-1))$, registrada na célula C16, retorna a projeção das vendas em $t+1$ igual a 309,2.

	A	B	C	D	E
1	Exemplo 16.6				
2					
3		t	Vendas	Taxa	1+Taxa
4		1	295		
5		2	305	3,39%	1,0339
6		3	316	3,61%	1,0361
7		4	298	-5,70%	0,9430
8		5	305	2,35%	1,0235
9		6	310	1,64%	1,0164
10		7	316	1,94%	1,0194
11		8	306	-3,16%	0,9684
12		9	301	-1,63%	0,9837
13		10	295	-1,99%	0,9801
14		11	312	5,76%	1,0576
15		12	308	-1,28%	0,9872
16		t+1=13	309,2		
17					

Projeção média móvel

A projeção \hat{y}_{t+1} pela média móvel é o resultado da média dos k últimos valores coletados $t, t-1, t-2, \dots, t-k+1$ e é calculada com $\hat{y}_{t+1} = \frac{1}{k} \times \sum_{i=t-k+1}^t y_i$, mantendo constante o número de valores k utilizados no cálculo da média. Pode-se dizer que o futuro é projetado pela média do passado.

EXEMPLO 16.7

Com os dados registrados na Figura 16.9, projete as vendas diárias da empresa pelo modelo da média móvel, considerando a média dos três últimos meses. Depois, repita a projeção com a média dos seis últimos meses.

Solução. Os dados e a solução estão na planilha **Média móvel** do **Capítulo 16**. A projeção de vendas em $t+1$ com média móvel dos três últimos é 305, resultado registrado na célula D17. Apenas para explicar o resultado dessa célula, o valor projetado em $t+1$ pode ser obtido com a fórmula:

$$\hat{y}_{t+1} = \frac{y_{12} + y_{11} + y_{10}}{3}$$

$$\hat{y}_{t+1} = \frac{308 + 312 + 295}{3} = 305,0$$

	A	B	C	D	E	F	G
1	Média Móvel						
2							
3				Projeção			
4		t	Vendas	k=3	Erro	k=6	Erro
5		1	295				
6		2	305				
7		3	316				
8		4	298	305,3	-7,3		
9		5	305	306,3	-1,3		
10		6	310	306,3	3,7		
11		7	316	304,3	11,7	304,8	11,2
12		8	306	310,3	-4,3	308,3	-2,3
13		9	301	310,7	-9,7	308,5	-7,5
14		10	295	307,7	-12,7	306,0	-11,0
15		11	312	300,7	11,3	305,5	6,5
16		12	308	302,7	5,3	306,7	1,3
17		t+1=13		305,0		306,3	
18							

Para analisar o comportamento dos resultados das projeções pela média móvel, a fórmula =MÉDIA(C5:C7) foi registrada na célula D8 e depois copiada até a célula D17. Na coluna E, foi medido o erro da projeção da forma conhecida.

O procedimento de projeção considerando a média dos seis últimos meses foi construído nas colunas F e G da planilha, como mostra a figura anterior.

A venda da empresa do Exemplo 16.7 foi projetada pela média móvel considerando 3 meses, coluna D, e 6 meses, coluna F, anteriores à data de projeção. Qual das duas projeções é a melhor? Deve-se escolher a projeção que apresentar menor erro de projeção. O procedimento de média móvel não é prático quando o número de valores coletados é grande e são necessárias atualizações frequentes, ou quando apenas os últimos valores são relevantes. Para realizar projeções com média móvel, o Excel dispõe dos recursos naturais da planilha, como apresentado acima, da ferramenta de análise *Média móvel* e do comando *Linha de tendência*, utilizado no Capítulo 15 e apresentado de forma completa no início deste capítulo.

Ferramenta de análise média móvel

A ferramenta *Média móvel*⁵ realiza projeções pelo procedimento de *média móvel*. Para compreender a utilização dessa ferramenta, será utilizado o Exemplo 16.7, como registrado na planilha *Ferramenta Média móvel*. Depois de selecionar *Análise de dados* dentro do menu *Ferramentas*, o Excel apresentará a caixa de diálogo *Análise de dados* com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 deste livro. Ao escolher a ferramenta *Média móvel* e depois clicar no botão OK, será exibida a caixa de diálogo com o mesmo nome, conforme mostra a Figura 16.10, depois de selecionadas as opções do exemplo. Clicando no botão *Ajuda* dessa caixa de diálogo, o Excel apresentará a página *Sobre a caixa de diálogo Média móvel* pertencente à *Ajuda do Excel*. As informações que devem ser registradas no quadro *Entrada* da caixa de diálogo dessa ferramenta são:

- **Intervalo de entrada:** Informe o intervalo de células da planilha no qual os dados estão registrados, incluindo o título.
- **Rótulos da primeira coluna:** Selecione este item, pois o intervalo inclui o nome da amostra.
- **Intervalo:** Informe o número de dados que será utilizado no cálculo da média móvel; no nosso caso o valor $k=3$.

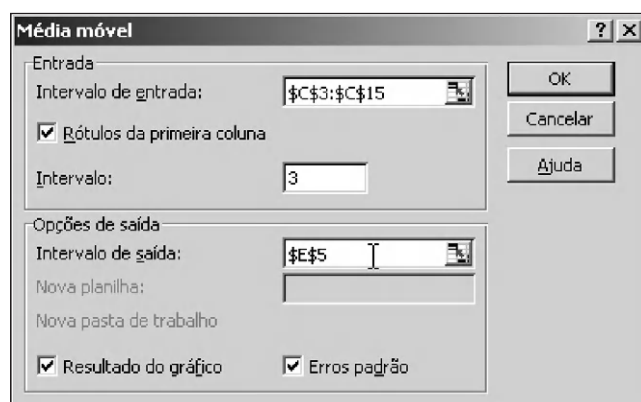


FIGURA 16.10
Caixa de diálogo da ferramenta *Média móvel*.

⁵ Em inglês, a ferramenta de análise Média móvel é *Moving average*.

No quadro **Opções de saída**, deve ser informado:

- **Intervalo de saída:** Os resultados serão apresentados na mesma planilha a partir da célula informada, neste caso E5, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes. Mais informações podem ser obtidas no Capítulo 4 ou na Ajuda do Excel.
- **Resultado do gráfico:** Selecione essa opção se for necessário que a ferramenta construa o gráfico dos valores coletados e das projeções.
- **Erros padrão:** Escolhendo esta alternativa, a ferramenta calculará o desvio da projeção baseado na média dos k erros ao quadrado. O resultado é o erro padrão da estimativa ou projeção S , cujo valor é obtido com a fórmula:

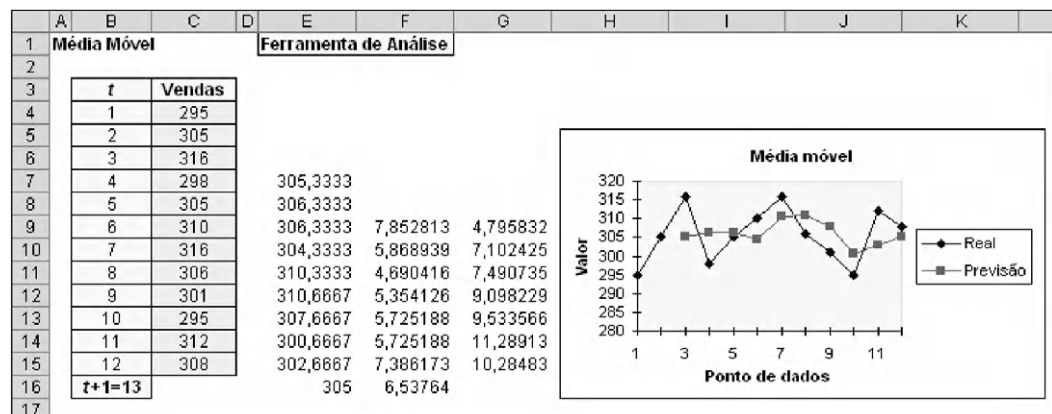
$$S = \sqrt{\frac{\sum_{i=t-k+1}^t (y_i - \hat{y}_i)^2}{k}}$$

Clicando em OK, a ferramenta registrará os resultados e construirá o gráfico na planilha **Ferramenta Média Móvel**, mostrado na Figura 16.11, depois de receber alguns ajustes de formatação. No cálculo da soma dos quadrados das diferenças, a ferramenta de análise *Média móvel* utiliza a função matemática SOMAXMY2.

- **SOMAXMY2(matriz_x; matriz_y)**

A função matemática SOMAXMY2⁶ retorna a soma dos quadrados das diferenças dos valores correspondentes de *matriz_x* e *matriz_y*. Essa função retorna o resultado de $\sum (y_i - \hat{y}_i)^2$, a soma dos quadrados dos erros de projeção.

FIGURA 16.11
Resultados do Exemplo 16.7 com a ferramenta *Média móvel*.



A ferramenta *Média móvel* registrou fórmulas nas células do intervalo E7:F16 da planilha, de forma que se forem mudados os valores das vendas, as projeções e os erros serão automaticamente atualizados, sem necessidade de ativar novamente a ferramenta *Média móvel*. Entretanto, os resultados dessa ferramenta apresentam alguns desvios; por exemplo, as fórmulas do erro padrão não foram registradas de forma correta, sendo os resultados corretos registrados na coluna G; ainda, o gráfico de projeção começa no tempo $t=3$, quando deveria começar em $t=4$. Deixamos por sua conta a construção das projeções com média móvel utilizando o comando *Linha de tendência*. Como ajuda, na planilha **Ferramenta Média móvel** foi construído esse gráfico.

⁶ Em inglês, a função SOMAXMY2 é SUMXMY2.

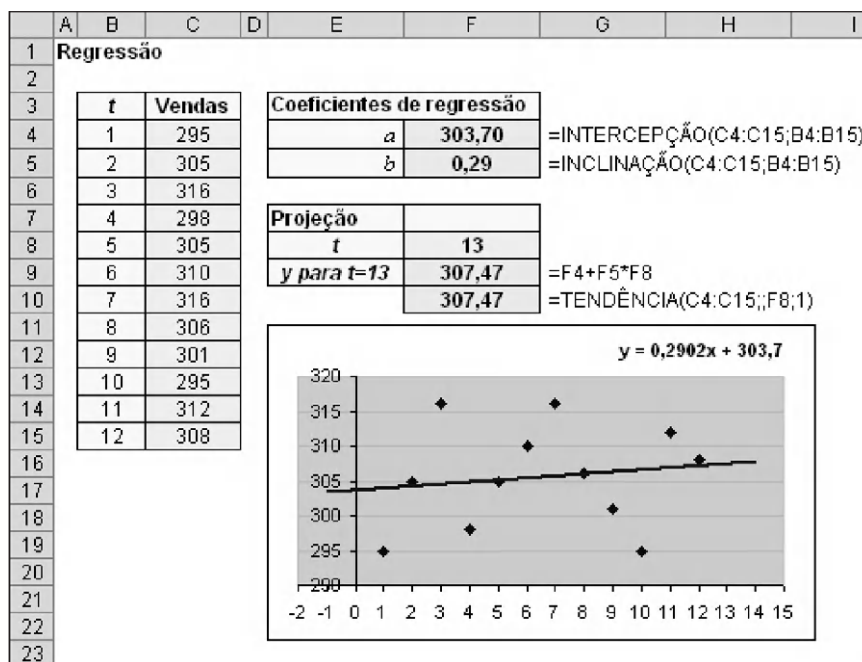
Projeção regressão linear

O ajuste de uma reta de regressão é um modelo linear que relaciona a variável dependente y e a variável independente x por meio da equação de uma reta do tipo $y = a + bx$, que resume a relação linear entre duas variáveis, onde as variações de y são provocadas pelas variações de x . Agora a variável independente é o tempo x_t , que varia de forma periódica e provoca as variações da variável dependente y_t através da função $y_t = a + bx_t$. Lembrando que a melhor reta é aquela cuja soma dos quadrados dos desvios é mínima, será possível ajustar uma reta em uma variável y que varia com o tempo utilizando os conceitos apresentados no Capítulo 15.

EXEMPLO 16.8

Com os dados registrados na Figura 16.9, determinar os coeficientes de regressão das vendas da empresa em função do tempo.

Solução. O exemplo foi resolvido na planilha **Regressão**, incluída na pasta **Capítulo 16**.



Para tomar conhecimento do que ocorre com a série das vendas diárias, foi construído o gráfico de dispersão, que mostra, enganosamente, uma grande variabilidade de vendas diárias. Na coluna D da planilha **Exemplo 16.6**, pode-se verificar que a taxa de variação das vendas não é grande, é a escala do eixo de ordenadas do gráfico de dispersão que amplifica essas pequenas variações. Continuando, os coeficientes de regressão foram obtidos no intervalo F4:F5 da planilha utilizando as funções estatísticas registradas na figura anterior. Dessa maneira, a equação da reta de regressão é $y_t = 303,07 + 0,29 x_t$ e será utilizada para realizar projeções. Conhecida a equação da reta, podem ser realizadas projeções como a que foi feita na célula F9 para $t=13$, registrando a fórmula $=F4+F5*F8$, que retorna a projeção 307,47 em $t=13$. Verifique o leitor que, na realidade, não seria necessário conhecer a equação da reta de regressão para realizar projeções, pois, com a função estatística TENDÊNCIA, a projeção pode ser realizada diretamente a partir dos dados das séries e o valor da variável independente, neste caso 13. A fórmula $=TENDÊNCIA(C4:C15;F8;1)$ registrada na célula F10 retorna a projeção 307,47 em $t=13$.

Outra forma de obter a equação da reta é utilizando o comando *Linha de tendência*, como mostra o gráfico de dispersão anterior, depois de receber alguns ajustes de formatação. Observe que a reta de regressão foi construída incluindo dois-pontos antes e dois-pontos depois dos valores da variável tempo. Lembrando o que foi apresentado neste capítulo, como a linha de tendência é construída no intervalo dos pontos desenhados

no gráfico de dispersão, no quadro **Previsão** da guia **Opções** da caixa de diálogo **Linha de tendência** é possível incluir mais pontos, antes e depois do intervalo dos dados. Essa alternativa está disponível somente para as curvas de ajustes de regressão e não se aplica ao ajuste da média móvel. Na caixa **Prospectiva**, pode-se escolher a quantidade de períodos, ou unidades do gráfico de dispersão, que serão incluídas depois do limite superior do intervalo de dados. Na caixa **Retrospectiva**, pode-se escolher a quantidade de períodos que será incluída antes do limite inferior do intervalo de dados.

A projeção utilizando a regressão linear simples é muito fácil de utilizar devido às facilidades operacionais do Excel, que resume todo o conteúdo das informações nos coeficientes da reta de regressão. Contudo, não se deve esquecer o tratamento linear da solução recebida, que pode ser melhorada utilizando os ajustes não lineares apresentados utilizando os recursos do comando *Linha de tendência* do Excel.

Projeção alisamento exponencial

Embora seja fácil de aplicar, a projeção pela *média móvel* requer que uma considerável quantidade de dados se mantenha armazenada. Outra desvantagem é que todos os dados da série têm o mesmo peso, sendo que em muitos casos os dados mais recentes são mais relevantes que os anteriores. A primeira desvantagem poderia ser eliminada calculando médias ponderadas; por exemplo, na média de três dados, o mais próximo teria mais peso do que os dois restantes, mantendo a soma dos pesos igual a um. Esse procedimento eliminaria a primeira desvantagem, mas manteria a necessidade de manter muitos dados armazenados, adicionando complexidade ao procedimento de cálculo.

Essas duas desvantagens da projeção com média móvel são atenuadas com o *alisamento exponencial*, realizando a projeção de \hat{y}_{t+1} em $(t+1)$ com a expressão $\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \hat{y}_t$, sendo α a *constante de alisamento* com valores entre zero e um. Analisando essa fórmula, podemos ver que o valor projetado de y em $(t+1)$ é a média ponderada do dado coletado y_t no período anterior t e da projeção \hat{y}_t no mesmo período t . Para compreender melhor a forma de operar do modelo, analisemos algumas características de sua expressão matemática:

- A expressão da projeção pode ser reescrita como $\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t)$. Ou seja, a projeção em $(t+1)$ é igual à projeção em t mais uma parte ($0 < \alpha < 1$) do erro de projeção na data t , diferença entre o valor coletado e o valor projetado.
- A projeção \hat{y}_{t+2} é obtida com a expressão, $\hat{y}_{t+2} = \alpha y_{t+1} + (1 - \alpha) \hat{y}_{t+1}$. Substituindo \hat{y}_{t+1} nessa expressão $\hat{y}_{t+2} = \alpha y_{t+1} + (1 - \alpha) \times (\alpha y_t + (1 - \alpha) \hat{y}_t)$ e depois de reescrevê-la $\hat{y}_{t+2} = \alpha y_{t+1} + \alpha \times (1 - \alpha) y_t + (1 - \alpha)^2 \times \hat{y}_t$, pode-se ver de onde vem o nome exponencial.
- Para iniciar as projeções, deve-se definir o primeiro valor da série de projeções. Uma forma de fazer isso é repetindo o valor observado na data t . Outra forma é iniciar com a média das primeiras observações coletadas. O procedimento da ferramenta de análise *Ajuste exponencial* do Excel começa no período seguinte à primeira observação, repetindo a observação do primeiro período. Dessa maneira, a primeira projeção será realizada em $t=2$.

EXEMPLO 16.9

Continuando com o Exemplo 16.8, projete as vendas da empresa pelo modelo de alisamento exponencial, considerando o coeficiente de alisamento $\alpha=0,10$.

Solução. Este exemplo foi resolvido na planilha **Alisamento exponencial**, incluída na pasta **Capítulo 16**. Com os valores coletados registrados na coluna C:

	A	B	C	D	E
1	Alisamento Exponencial				
2					
3			Constante alisamento	0,100	
4		t	Vendas	Projeção	Erro
5		1	295		
6		2	305	295,0	10,0
7		3	316	296,0	20,0
8		4	298	298,0	0,0
9		5	305	298,0	7,0
10		6	310	298,7	11,3
11		7	316	299,8	16,2
12		8	306	301,4	4,6
13		9	301	301,9	-0,9
14		10	295	301,8	-6,8
15		11	312	301,1	10,9
16		12	308	302,2	5,8
17		t+1=13		302,8	
18					

- Em $t=1$, não é realizada nenhuma projeção.
- Em $t=2$, a projeção é o valor da amostra em $t=1$.
- Em $t=3$, é realizada a primeira projeção com alisamento exponencial $\alpha=0,1$. Apenas para compreender a utilização da fórmula, o valor projetado em t_3 é 296, resultado obtido com:

$$\hat{y}_3 = \alpha y_2 + (1 - \alpha) \hat{y}_2$$

$$\hat{y}_3 = 0,10 \times 305 + (1 - 0,10) \times 295 = 296$$

Esse resultado foi obtido com a fórmula `=D$3*C6+(1-D$3)*D6`, registrada na célula D7 da planilha **Alisamento Exponencial**. Depois, essa fórmula foi copiada até a célula D17. A projeção para $(t+1)$ é 302,8, resultado obtido na célula D17. No intervalo E6:E16, foi calculado e registrado o erro da projeção da forma conhecida, como mostra a figura a acima.

Ferramenta de análise *ajuste exponencial*

A ferramenta de análise *Ajuste exponencial*⁷ realiza projeções com alisamento exponencial. Para compreender a utilização dessa ferramenta, será utilizado o Exemplo 16.9, como registrado na planilha **Ferramenta Ajuste exponencial**. Depois de selecionar **Análise de dados** dentro do menu **Ferramentas**, o Excel apresentará a caixa de diálogo **Análise de dados** com todas as ferramentas de análise disponíveis, como mostrado na Figura 1.7 do Capítulo 1 deste livro. Ao escolher a ferramenta **Ajuste Exponencial** e depois clicar no botão **OK**, será exibida a caixa de diálogo com o mesmo nome, conforme mostra a Figura 16.12, depois de selecionadas as opções do exemplo. Clicando no botão **Ajuda** dessa caixa de diálogo, o Excel exibirá a página *Sobre a caixa de diálogo Ajuste exponencial* pertencente à *Ajuda do Excel*. As informações que devem ser registradas no quadro **Entrada** da caixa de diálogo dessa ferramenta são:

- **Intervalo de entrada:** Informamos o intervalo dos valores coletados y , incluindo a célula de seu título.
- **Fator de amortecimento:** A informação esperada pelo Excel é o resultado de um menos o fator de alisamento α . Neste exemplo informamos o valor 0,90.
- **Rótulos:** Selecione, pois o intervalo inclui o nome da amostra.

No quadro **Opções de saída** deve ser informado:

- **Intervalo de saída.** Os resultados serão apresentados na mesma planilha a partir da célula informada, neste caso E4, que é o endereço da célula superior esquerda da tabela de respostas que a ferramenta construirá. Também, o Excel automaticamente definirá o tamanho da área dos resultados e

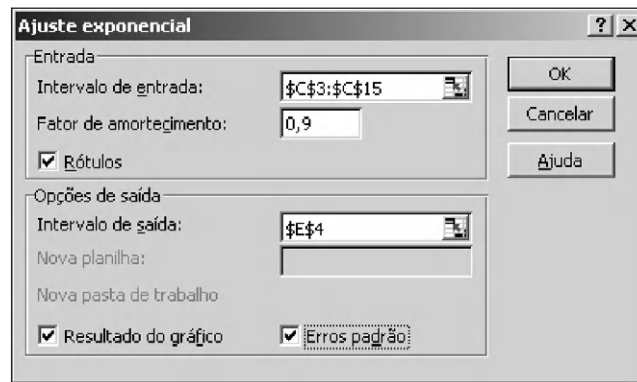
⁷ Em inglês, a ferramenta de análise Ajuste exponencial é *Exponential smoothing*.

exibirá uma mensagem se a tabela de saída estiver prestes a substituir dados existentes. Mais informações podem ser obtidas no Capítulo 4 ou na Ajuda do Excel.

- **Resultado do gráfico:** Selecionar este item, pois queremos que a ferramenta construa o gráfico dos valores observados e das projeções.
- **Erros padrão:** Escolhendo essa alternativa, a ferramenta calculará o desvio da projeção baseado na média dos k erros ao quadrado. O resultado é o erro padrão da estimativa ou projeção S , cujo valor é obtido com a fórmula:

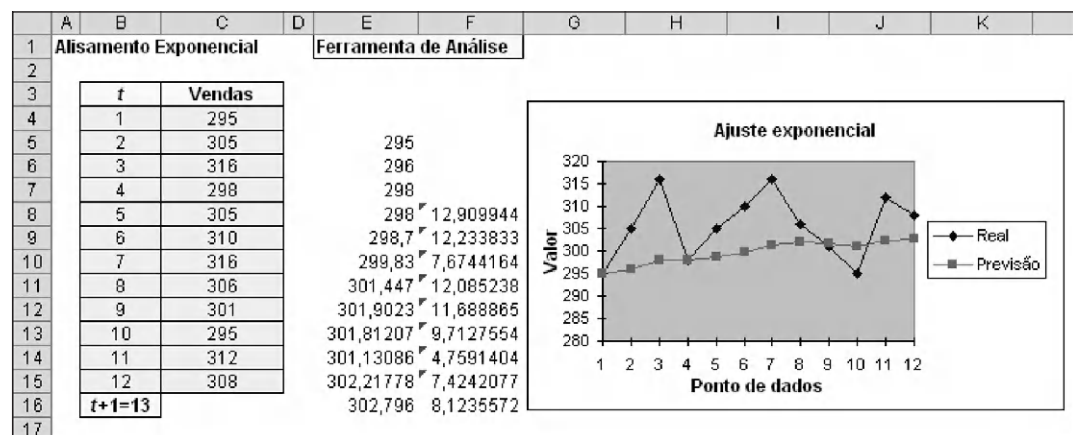
$$S = \sqrt{\frac{\sum_{i=t-3+1}^t (y_i - \hat{y}_i)^2}{3}}$$

FIGURA 16.12 Caixa de diálogo da ferramenta *Ajuste exponencial*.



Clicando no botão OK, a ferramenta registrará os resultados e construirá o gráfico com a série de dados e as projeções realizadas na planilha **Ferramenta Ajuste exponencial**, conforme mostra a Figura 16.13. Para calcular a soma dos quadrados das diferenças, a ferramenta aplica a função matemática SOMAXMY2, registrando fórmulas nas células da planilha, de forma que se forem mudados os valores da série de dados, os valores das projeções e do *erro* serão automaticamente recalculados sem necessidade de ativar novamente a ferramenta *Ajuste exponencial*. Na planilha, foram adicionados dois resultados pintados de cor azul, copiando as fórmulas anteriores, para projetar y na data $t=13$.

FIGURA 16.13 Resultados da ferramenta *Ajuste exponencial*, Exemplo 16.9.



Ajuste da constante de alisamento

No Exemplo 16.9, foram projetadas as vendas da empresa utilizando alisamento exponencial com $\alpha=0,10$. Qual a constante de alisamento que deve ser utilizada para obter uma boa projeção? Deve-se escolher a constante de alisamento que gere o menor erro de projeção. Como se deve medir o erro de projeção? Deve-se minimizar o erro padrão, que considera a soma dos quadrados dos desvios de todas as projeções, utilizando a fórmula seguinte:

$$S = \sqrt{\frac{\sum_{i=2}^n (y_i - \hat{y}_i)^2}{n-1}}$$

Na célula E18 da planilha **Ajuste da constante**, foi registrada a fórmula do erro padrão utilizando a função estatística SOMAXMY2 e $12-1=11$. O objetivo é determinar o valor registrado na célula D3 que minimiza o resultado da célula E18, procedimento de otimização que pode ser resolvido com o comando Solver do Excel, que foi apresentado no Apêndice 1 do Capítulo 15. No menu **Ferramentas**, selecione **Solver**⁸ e depois preencha as opções, como mostrado na Figura 16.14.

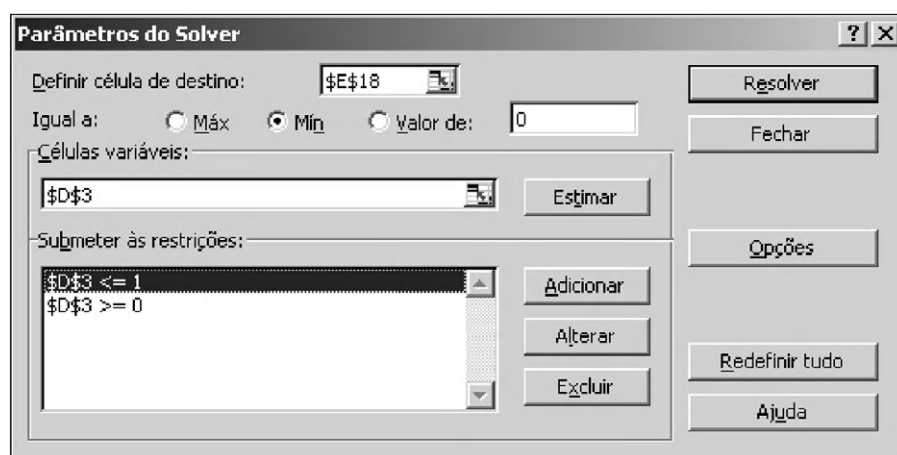


FIGURA 16.14 Caixa de diálogo do comando Solver.

Depois de clicar no botão **Resolver**, o comando Solver registrará a solução encontrada na célula D3, neste caso o valor 0,356, e exibirá a caixa de diálogo da Figura 16.15. Clicando no botão **OK**, o resultado será mantido na célula D3. Se clicar em **Cancelar**, será mantido o valor inicial registrado nessa célula, da mesma forma se selecionar a caixa **Restaurar valores originais** e depois clicar em **OK**. A Figura 16.15 também mostra a planilha **Ajuste da constante**, depois de o Solver registrar na célula D3 o valor da constante de alisamento igual a 0,356, que minimiza o resultado da célula E18, ou o valor da soma dos quadrados dos desvios de todas as projeções é mínimo.

⁸ Se o comando Solver não estiver incluído no menu **Ferramentas**, então verifique se o Solver aparece no menu **Ferramentas – Suplementos**, onde deve ser selecionado. Se em **Suplementos** não aparecer o Solver, então significa que não foi instalado.

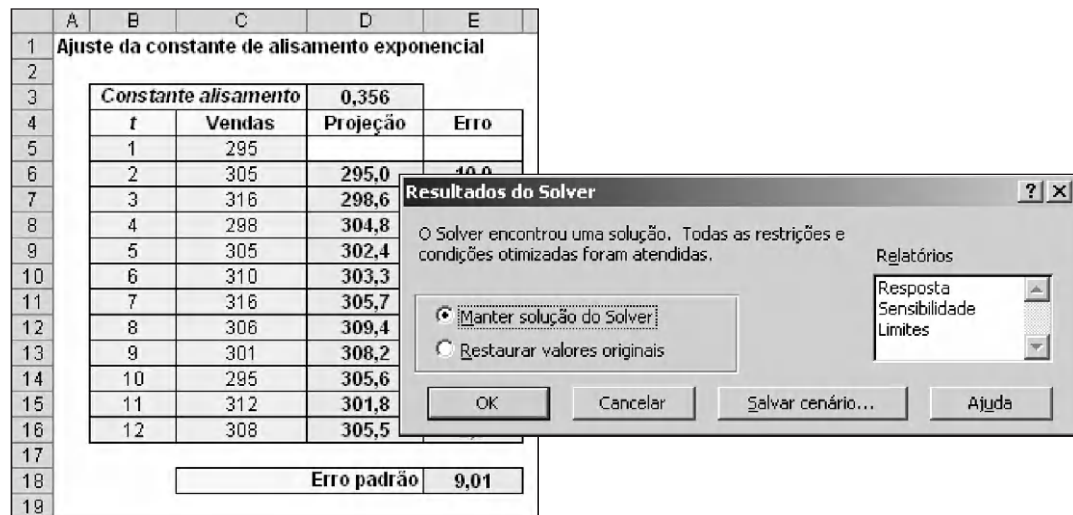


FIGURA 16.15 Determinação da constante de alisamento utilizando o Solver.

Problemas

Problema 1

As variáveis x e y do experimento foram medidas e registradas na tabela seguinte.

x	1,5	3,5	4,3	5	3,8	4	2,9	4,8	1,9	2,7
y	55	153	215	300	174	220	120	230	63	103

Construa o gráfico de dispersão, analise a relação entre as variáveis e recomende uma função para projetar valores de y .

Problema 2

Continuando com o Problema 1. Calcule os coeficientes de uma função exponencial para projetar y utilizando as fórmulas.

Problema 3

Repita o Problema 1, projetando os valores de y com uma função potência cujos coeficientes não são conhecidos e devem ser determinados utilizando as fórmulas.

Problema 4

Repita os Problemas 1 e 2, utilizando o comando *Linha de tendência* do Excel.

Problema 5

Analisar a relação entre as variáveis x e y registradas na tabela seguinte, construindo o gráfico de dispersão correspondente.

x	20	25	32	39	47	51	54	58	64	75
y	45	77	95	77	61	52	66	105	131	142

Problema 6

O objetivo da empresa é dispor de um modelo que projete valores de y baseados nos dados históricos da tabela do Problema 5. Selecione a função que melhor explique a relação entre as variáveis.

Problema 7

O analista de custos da empresa preparou a tabela a seguir, que registra o custo médio para diversas quantidades produzidas. O objetivo do analista é dispor de um modelo que projete valores do custo médio em função da quantidade produzida, ficando ao seu encargo:

- Desenhar o gráfico de dispersão.
- Ajustar a melhor função de projeção utilizando o comando *linha de tendência*.

q	1.00	2.00	3.20	4.20	5.20	6.20	7.20	8.20	9.50	1.050	1.200	1.400
CM	5.735	4.900	4.500	4.110	4.135	4.035	4.305	4.650	5.143	5.765	6.582	8.500

Problema 8

O gerente de marketing considera que as vendas durante os próximos seis anos serão as registradas na tabela seguinte.

t	1	2	3	4	5	6
$Venda$	1	4	10	18	30	43

Construa o gráfico de dispersão e analise a relação entre o tempo e as vendas.

Problema 9

Continuando com o Problema 8. O objetivo do gerente é dispor de um modelo que projete vendas em função do tempo. Ajuste a melhor função de projeção utilizando o comando *linha de tendência*.

Problema 10

O custo do produto é fortemente dependente do preço do petróleo. A tabela registra os custos trimestrais durante os últimos 10 meses.

t	1	2	3	4	5	6	7	8	9	10
y	140	134	112	117	119	140	115	133	124	114

Construa o gráfico de dispersão e analise a relação entre o tempo, os custos trimestrais e o preço do petróleo.

Problema 11

Continuando com o Problema 11. Ajuste uma reta de regressão e analise seu comportamento comparando com o resultado da análise do Problema 10.

Problema 12

Continuando com o Problema 11. Determine a taxa média de crescimento e projete o custo total em $t=11$.

Problema 13

Repita o Problema 12, considerando a projeção com média móvel de dois trimestrais. Depois repita a projeção com média móvel de três trimestrais e compare os resultados.

Problema 14

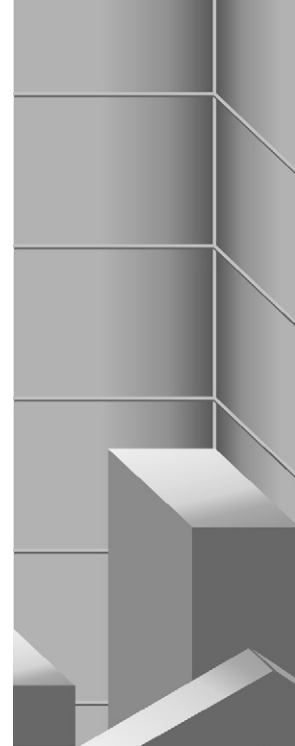
Repita o Problema 12, considerando a projeção com alisamento exponencial adotando a constante de alisamento igual a 0,10. Depois, determine a constante de alisamento que minimiza o erro padrão da projeção.

Problema 15

Repita do Problema 10 até o Problema 14, considerando a série de dados anuais da tabela a seguir.

<i>t</i>	1	2	3	4	5	6	7	8	9	10
<i>y</i>	55	62	58	62	58	63	65	63	61	73

TABELAS



Na pasta TABELAS incluída na página do livro, no site da Editora, foram construídas as seguintes sete tabelas estatísticas utilizando as funções matemáticas e estatísticas do Excel.

- Números Aleatórios.
- Distribuição Binomial.
- Distribuição de Poisson.
- Distribuição Z.
- Distribuição t .
- Distribuição F .
- Distribuição *Qui-quadrado*.

Neste capítulo, foram reproduzidas as seguintes seis tabelas estatísticas copiadas das planilhas da pasta TABELAS na página do livro, no site da Editora.

- Números Aleatórios.
- Distribuição Z.
- Distribuição t .
- Distribuição F , para $\alpha=0,01$.
- Distribuição F , para $\alpha=0,05$.
- Distribuição *Qui-quadrado*.

Tabela de números aleatórios

2445	8615	2895	6331	5698	8294	1935	9192	4277	6365	1461	4693
8737	5112	3148	3470	1180	3662	5837	7458	7096	8545	2559	8704
8074	0700	8866	4050	5611	9691	7283	0279	0882	5464	2218	3014
8241	9290	3101	4657	8337	8247	1492	2507	1209	6216	3784	5015
6059	4324	0055	3590	7708	1107	8633	4402	8571	9892	9181	0602
0283	4899	2450	5647	7008	5411	5915	7467	4815	6311	4542	2468
6462	2135	7113	8994	2328	6156	7084	8395	4463	6345	9409	3804
2360	1613	4347	2364	9811	4581	5611	5835	2148	4565	0956	3918
0580	3417	6611	8927	3229	9247	4785	1877	5262	0646	8966	7341
6915	3167	5548	7352	3761	7086	0636	0079	0506	0718	1759	2979
8395	0617	4946	5390	8008	2785	7629	3176	5114	1410	0569	7877
3069	5769	3617	1149	0276	5783	2837	7487	8159	3478	8152	1191
1859	8790	3106	7156	5673	6967	0812	1603	1330	5588	3706	6479
9645	7574	2954	5940	6263	6559	9450	2281	1362	3000	0482	8066
1136	6008	0598	8617	2380	0960	4412	7829	2840	8729	4840	1130
4220	5296	9960	1179	9882	3223	7574	3009	0586	8087	9234	0536
2745	0643	5915	7618	4488	8871	4909	2972	6106	7307	5255	5101
2887	8586	0033	6146	6995	7415	9267	3306	4876	9378	0709	1284
1308	5453	4265	2823	3980	9271	7984	9418	1928	7429	9430	3280
5688	8902	6741	1182	5137	6712	3235	1171	7707	2947	3106	4346
7095	2239	2388	1595	4608	0700	2123	9659	1199	3279	5117	4105
4278	1820	8244	9860	1660	9044	8928	4588	1803	8097	9058	6465
1395	1223	7100	5349	2947	9933	9883	6823	5558	6412	5570	1611
3180	0778	4992	5550	0392	6390	7495	6931	7169	7232	3336	3652
3069	1381	0722	5843	1771	5534	5498	1546	5629	0224	6874	6951
4494	2202	2245	5608	3987	1528	3547	5980	9320	5533	8915	9216
8431	5616	5772	7644	9890	3449	5349	0389	7361	6234	8730	8171
1688	3956	1452	2480	3272	7092	8004	8737	2321	7437	7276	3751
4353	9117	1514	9291	4073	4805	9291	7661	4127	4496	6437	4170
9467	1614	3055	8553	0962	3077	6939	8316	1557	9434	7754	0610
5324	4023	3288	1114	6548	8542	3484	1978	8413	4244	9551	7006
8894	2447	8149	0216	1971	1131	8226	5815	3688	3426	1527	0414
9156	0602	9590	0068	1887	0380	3409	7273	7315	8982	0491	2657
2324	3096	0098	6216	3899	9344	4450	4551	7003	3423	3451	8903
2437	6537	7403	7848	1668	2026	5659	0189	1405	0699	0836	0036

Z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,10	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,20	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,30	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,40	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,50	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,60	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,70	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,80	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,90	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4,00	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Distribuição t

	Intervalo de Confiança					
Duas caudas	80%	90%	95%	98%	99,00%	99,90%
Uma cauda	90%	95%	98%	99%	99,50%	99,95%
	Nível de significância					
Duas caudas	0,20	0,10	0,05	0,02	0,01	0,001
Uma cauda	0,10	0,05	0,025	0,01	0,005	0,0005
g.l. = 1	3,078	6,314	12,706	31,821	63,656	636,578
2	1,886	2,920	4,303	6,965	9,925	31,600
3	1,638	2,353	3,182	4,541	5,841	12,924
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,768
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,689
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,660
30	1,310	1,697	2,042	2,457	2,750	3,646

Distribuição F

Nível de significância: 0,01

Graus de liberdade: Numerador (colunas) e Denominador (linhas)

	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6157	6209	6240
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,43	99,45	99,46
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,87	26,69	26,58
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,20	14,02	13,91
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,72	9,55	9,45
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,56	7,40	7,30
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,31	6,16	6,06
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,52	5,36	5,26
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	4,96	4,81	4,71
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,56	4,41	4,31
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,25	4,10	4,01
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,01	3,86	3,76
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,82	3,66	3,57
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,66	3,51	3,41
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,52	3,37	3,28
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,41	3,26	3,16
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,31	3,16	3,07
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,23	3,08	2,98
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,15	3,00	2,91
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,09	2,94	2,84
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,03	2,88	2,79
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	2,98	2,83	2,73
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	2,93	2,78	2,69
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,89	2,74	2,64
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,85	2,70	2,60
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,81	2,66	2,57
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,78	2,63	2,54
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,75	2,60	2,51
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,73	2,57	2,48
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,70	2,55	2,45
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,52	2,37	2,27
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,42	2,27	2,17
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,35	2,20	2,10
70	7,01	4,92	4,07	3,60	3,29	3,07	2,91	2,78	2,67	2,59	2,51	2,45	2,31	2,15	2,05
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,48	2,42	2,27	2,12	2,01
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52	2,45	2,39	2,24	2,09	1,99
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,22	2,07	1,97
500	6,69	4,65	3,82	3,36	3,05	2,84	2,68	2,55	2,44	2,36	2,28	2,22	2,07	1,92	1,81
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,06	1,90	1,79

Distribuição F

Nível de significância: 0,05

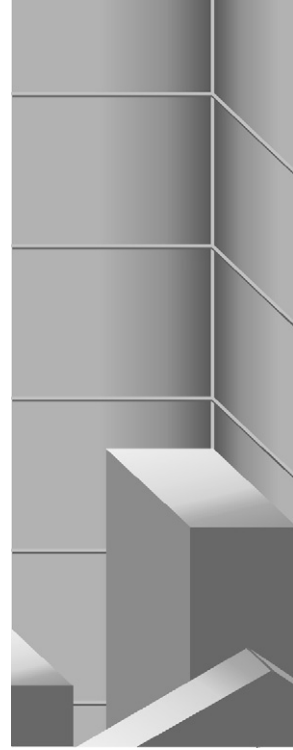
Graus de liberdade: Numerador (colunas) e Denominador (linhas)

	1	2	3	4	5	6	7	8	9	10	11	12	15	20	25
1	161	199	216	225	230	234	237	239	241	242	243	244	246	248	249
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,43	19,45	19,46
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,63
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,86	5,80	5,77
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,62	4,56	4,52
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,94	3,87	3,83
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,51	3,44	3,40
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,22	3,15	3,11
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,01	2,94	2,89
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,85	2,77	2,73
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,72	2,65	2,60
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,62	2,54	2,50
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,53	2,46	2,41
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,46	2,39	2,34
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,40	2,33	2,28
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,35	2,28	2,23
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,31	2,23	2,18
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,27	2,19	2,14
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,23	2,16	2,11
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,20	2,12	2,07
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,18	2,10	2,05
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,15	2,07	2,02
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,13	2,05	2,00
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,11	2,03	1,97
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,09	2,01	1,96
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,07	1,99	1,94
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,06	1,97	1,92
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,04	1,96	1,91
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,03	1,94	1,89
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,01	1,93	1,88
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,92	1,84	1,78
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,87	1,78	1,73
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,84	1,75	1,69
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,93	1,89	1,81	1,72	1,66
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,91	1,88	1,79	1,70	1,64
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94	1,90	1,86	1,78	1,69	1,63
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,77	1,68	1,62
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81	1,77	1,69	1,59	1,53
1000	3,85	3,00	2,61	2,38	2,22	2,11	2,02	1,95	1,89	1,84	1,80	1,76	1,68	1,58	1,52

Distribuição Qui-quadrado

gl	Nível de significância				
	0,10	0,05	0,025	0,010	0,005
1	2,71	3,84	5,02	6,63	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,81	9,35	11,34	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,64	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,53	20,09	21,95
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19
11	17,28	19,68	21,92	24,73	26,76
12	18,55	21,03	23,34	26,22	28,30
13	19,81	22,36	24,74	27,69	29,82
14	21,06	23,68	26,12	29,14	31,32
15	22,31	25,00	27,49	30,58	32,80
16	23,54	26,30	28,85	32,00	34,27
17	24,77	27,59	30,19	33,41	35,72
18	25,99	28,87	31,53	34,81	37,16
19	27,20	30,14	32,85	36,19	38,58
20	28,41	31,41	34,17	37,57	40,00
21	29,62	32,67	35,48	38,93	41,40
22	30,81	33,92	36,78	40,29	42,80
23	32,01	35,17	38,08	41,64	44,18
24	33,20	36,42	39,36	42,98	45,56
25	34,38	37,65	40,65	44,31	46,93
26	35,56	38,89	41,92	45,64	48,29
27	36,74	40,11	43,19	46,96	49,65
28	37,92	41,34	44,46	48,28	50,99
29	39,09	42,56	45,72	49,59	52,34
30	40,26	43,77	46,98	50,89	53,67
40	51,81	55,76	59,34	63,69	66,77
50	63,17	67,50	71,42	76,15	79,49
60	74,40	79,08	83,30	88,38	91,95
70	85,53	90,53	95,02	100,43	104,21
80	96,58	101,88	106,63	112,33	116,32
90	107,57	113,15	118,14	124,12	128,30
100	118,50	124,34	129,56	135,81	140,17

BIBLIOGRAFIA



- Banks J et al. *Discrete-Event System Simulation* – Prentice Hall, 2ª edição, 1996.
- Berenson M.L. and Levine D.M. *Basic Business Statistics – Concepts and Applications*. Prentice Hall, 1996.
- Bernstein P. *Desafio aos Deuses – A Fascinante História do Risco* – Editora Campus, 1997.
- Charnet R. et al. *Análise de Modelos de Regressão Linear com Aplicações*– Editora da Unicamp, 1999.
- Copeland T. *Opções Reais* – Editora Campus, 2001.
- Daniel W. W. *Biostatistics - A Foundation for Analysis in the Health Sciences* – John Wiley & Sons, 6ª edição, 1995.
- Freund J. E. *Mathematical Statistics* – Prentice Hall, 5ª edição, 1992.
- Kume H. *Métodos Estatísticos para Melhoria da Qualidade* – Editora Gente, 1993.
- Lapponi, J.C. *Matemática Financeira Usando Excel – Como Medir Criação de Valor* – Editora Lapponi, 2002.
- Lapponi, J.C. *Modelagem Financeira com Excel* – Editora Campus Elsevier, 2004.
- Lewis E. E. *Introduction to Reliability Engineering* – John Wiley, 2ª edição, 1996.
- Mendenhall W. e Sincich T. *A Second Course in Statistics* – Prentice Hall, 5ª edição, 1996.
- Mason R.D., Lind D.A. and Marchal W.G. *Statistical Techniques in Business and Economics*. Irwin – McGraw-Hill, 1999.
- Moore D.S and McCabe G.P. *Introduction to the Practice of Statistics*. Freeman, 1998.
- Pineda O. L. *Técnicas de Pronósticos para la Toma de Decisiones Empresariales* – Alfaomega, edição 2002.
- Sheskin D.J. *Handbook of Parametric and Nonparametric Statistical Procedures* – Editora Chapman & Hall/CRC. 2ª edição, 2000.
- Siegel A.F. and Morgan C.J. *Statistics and Data Analysis – An Introduction*. John Wiley, 1996.
- Zimmerman S.M. and Icenogle M.L. *Statistical Quality Control Using Excel*. ASQ Quality Press, 1999.

ÍNDICE

A

Ajuste de uma reta, *veja* Regressão linear
Ajuste linear, *veja* regressão linear
Ajuste não linear, 437
 linha de tendência do Excel, 439
 resumo das transformações, 439
 transformação de funções, 435
 função exponencial, 436
 função logarítmica, 437
 polinomial, 443
 função potência, 438
Ajuste polinomial, 443
Alfa, *veja* erro tolerado
Alisamento exponencial, 452
 ajuste da constante alisamento, 455
 ferramenta de análise, 453
Amostra, 1, 9
 aleatória, 11
 escolha do tamanho da, 309, 320
 ordenada, 70
 representativa, 10
Amostragem, 15
 Outros tipos de, 26
 Tabela de números aleatórios, 16
Amostragem com reposição, 15
Amostragem sem reposição, 15, 25, 33
Amostra probabilística, 15
 geração de, 15
 ferramenta de análise *Amostragem*, 22
 modelo em Excel para, 21
Amostras estratificadas, 26
 como são feitas as pesquisas, 27
Amostragem sistemática, 26
Análise da forma da distribuição, 117
Análise da Variância, 379
 conceituação da, 380
 ferramenta de análise,
 Anova: fator único, 386
 Anova: fator duplo com repetição, 390
 Anova: fator duplo sem repetição, 392

premissas da, 381

tabela ANOVA,
 um fator, 384

Análise de carteira de investimento, 277
Análise do VPL de um investimento, 275
Anova, *veja* Análise da Variância
Árvore de possibilidades, 155
Assistente de gráfico do Excel, 48
Atingir meta, comando do Excel, 85, 296

B – C

Banco de dados, 139
 funções do Excel para, 141
Benford, lei de, 155
Beta, *veja* teste de hipóteses
Binomial, *veja* Distribuição binomial
Boxplot, 125
 construção de um boxplot com Excel, 127
 intervalo entre quartis, 124
 recursos do Excel, 127
Cálculo inverso com a DN, 236-237
Carteira de investimento,
 Veja Análise de
Cenários com VA discreta, 195
Censo 2000, IBGE, 1
Classes, dados contínuos, 44
Classificar dados, comando do Excel, 71
Coeficiente de correlação, 175
 anomalias do, 182
 características do, 177
 e causalidade, 181
 interpretação dos valores do, 177-178
 simulador, 179
 tabela de, 183
 variáveis não correlacionadas, 179
 perfeitamente correlacionadas, 177
Coeficiente de determinação, 405
 ajustado, 413

- Coeficientes de regressão, 398
 - cálculo dos, 399-400
 - com as medidas estatísticas, 403
 - com o Solver, 426
 - Coeficiente de variação, 117
 - Combinação linear de VA's, 259
 - com uma variável aleatória, 264
 - combinando as medidas estatísticas, 262
 - combinação de duas VA's, 262
 - análise dos resultados importantes, 263-264
 - combinação de n VA's, 265
 - VA's independentes, 267
 - conceituação, 259
 - distribuição da VA resultante, 267
 - formação carteira de investimento, 277
 - utilizando o Solver, 278
 - modelo da, 268
 - transformação linear de uma VA, 257
 - VPL de um projeto de investimento, 275
 - Como registrar uma função do Excel, 30
 - Complemento de um evento, 149
 - Combinações, 163
 - Confiança da estimativa, 301
 - Construção de gráficos com Excel, 48
 - Construção de um boxplot com Excel, 127
 - construção de dois ou mais boxplot, 128
 - Contagem, técnicas de, 161
 - combinações, 163
 - permutações, 162
 - Correlação, 169
 - coeficiente de, 175
 - covariância, 171
 - Corte transversal numa data ou período, 9
 - Covariância, 171
 - características da, 173
 - coeficiente de correlação, 175
 - como valor esperado, 219
 - outra forma da, 220
 - tabela de contingências, 219
 - tabela de, 183
 - Curtose de uma distribuição, 119
- D**
- Dados
 - boxplot, 125
 - construção de um ou mais com Excel, 127
 - classificação dos, 7
 - contínuos, 8
 - discretos, 7
 - e variáveis, 6
 - escala de medição dos, 8
 - nominais, 8
 - ordenamento de, 70
 - ordinais, 8
 - origem dos, 5
 - qualitativos, 8
 - quantitativos, 7
 - contínuos, 44
 - discretos, 36
 - suspeitos, 126
 - unidade elementar, 6
 - Dados suspeitos, 126
 - Desvio, 84
 - quadrado do, 85
 - soma dos quadrados dos, 85, 109
 - Desvio médio absoluto, 108
 - Desvio padrão, 112
 - da distribuição amostral, 288
 - da variável aleatória, 199, 222
 - erro padrão, 288
 - fator de correção finita, 291
 - funções estatísticas do Excel para, 113
 - normalizado Z, 231
 - regra prática, 115
 - relação entre os desvios padrões, 114
 - significado do, 114
 - teorema de Chebyshev, 114
 - Desvio padrão normalizado Z, 231
 - Diagrama de Venn, 149
 - eventos no, 148
 - coletivamente exaustivos, 150
 - complemento de um evento, 149
 - elementar, 149
 - mutuamente excludentes, 150
 - operações com eventos, 149
 - interseção, 149
 - união, 149
 - Dígitos e números aleatórios, 11
 - Dispersão, 107
 - Distribuição,
 - inclinação, 118
 - curtose, 119
 - Distribuição amostral, 285
 - desvio padrão da, 288
 - erro padrão, 288
 - fator de correção finita, 292
 - forma da, 289
 - formação da, 286
 - média amostral, 287, 288
 - variabilidade da, 285
 - procedimento com a distribuição Z, 292
 - simulador teorema central do limite, 290
 - teorema central do limite, 289
 - Distribuição binomial, 201
 - fórmula da, 202
 - modelo da, 203
 - parâmetros, média e variância, 205
 - premissas da, 201
 - probabilidade da, 203
 - probabilidade acumulada da,
 - 204-205
 - tabela da, 208
 - Distribuição binomial negativa, 213
 - Distribuição de frequências absolutas, 36
 - Distribuição de frequências acumuladas, 38
 - Distribuição de frequências relativas, 37
 - Distribuição de frequências com classes,
 - dados contínuos, 44
 - agrupamentos em classes, 45
 - utilizando o Excel, 46
 - dados discretos, 58
 - utilizando Excel, 58-59

Distribuição de Poisson, 210
 modelo da, 211
 premissas da, 209
 probabilidade da, 209
 tabela da, 212

Distribuição exponencial, 240
 modelo da, 242
 premissas da, 227
 probabilidade da, 241

Distribuição F, 365
 ferramenta de análise, 372
 funções estatísticas, 367
 função estatística TESTEF, 372
 tabela da, 65, 414, 463, 464

Distribuição hipergeométrica, 214

Distribuição lognormal, 243
 modelo da, 244-245
 premissas da, 244

Distribuição normal, 224
 cálculo de probabilidade, 226
 com o modelo DN, 222
 utilizando a função do Excel, 226
 cálculo dos parâmetros da, 238
 cálculo inverso com a, 236
 família da, 225
 influência dos parâmetros na, 225
 modelo completo, 230
 outros cálculos com a, 235
 propriedades da, 225
 resultados importantes, 229

Distribuição normal padronizada, 231
 cálculo de probabilidade, 233
 com a tabela Z, 233
 com o modelo DN, 235
 utilizando as funções do Excel, 232
 desvio padrão normalizado Z, 231
 outros cálculos com a, 223
 propriedades da, 231
 tabela da, 233

Distribuição qui-quadrado, 373-374
 funções estatísticas da, 374, 376
 tabela da, 376, 465

Distribuição *t* de Student, 310
 graus de liberdade, 310
 modelo da, 310
 tabela da, 311, 462

Distribuição uniforme, 222
 média, 223
 modelo da, 223
 variância, 223

Distribuições discretas, 194

Distribuições contínuas, 221

E

Erro de estimativa, 302

Erro padrão, 288
 da estimativa na regressão linear, 407

Erro tolerado, 306

Erros no teste de hipóteses, 326

Erro *tipo I*, veja *teste de hipóteses*

Erro *tipo II*, veja *teste de hipóteses*

Escala de medição dos dados, 8

Espaço amostral, 148

Estatística, 286

Estatísticas, 88
 parâmetros, 88

Estatística descritiva, 5
 inferência estatística, 5, 301

Estimação, 301

Estimativa da média da população, 301
 com distribuição padronizada Z, 302
 com a distribuição *t*, 310
 graus de liberdade, 310
 tabela da distribuição *t*, 311
 confiança da estimativa, 301
 erro de estimativa, 302
 erro tolerado, *alfa*, 306
 desvio padrão desconhecido, 308
 amostra adequadamente grande, 306
 cálculo do tamanho da amostra, 309
 erro tolerado, 306
 intervalo de confiança, 302
 margem de erro, 301
 probabilidade de acerto, 302
 probabilidade de erro α , 306
 simulação do intervalo da, 304
 tamanho da amostra, escolha do, 309

Estimativa intervalar, veja *estimativa da média*

Estimativa pontual, 301

Evento, 148
 coletivamente exaustivos, 150
 complemento de um evento, 149
 elementar, 149
 mutuamente excludentes, 150
 operações com eventos, 149
 interseção, 149
 união, 149

Excel,
 assistente de gráfico, 48, 54
 classificação de uma lista, 70
 colar função, 30
 comando *Atingir meta*, 85, 296
 comando Dados – Classificar, 70
 comando *Linha de tendência*,
 média móvel, 426
 polinomial, 443
 regressão linear, 396
 comando Solver, 29, 426, 455
 como registrar uma função, 30
 construção de histograma com Excel, 47-48
 construção de um histograma combinado, 54
 construção de um ou mais boxplots, 127
 cópia de uma planilha, 66
 ferramentas de análise, veja
ferramentas de análise
 fixando o endereço de células, 66
 funções, veja *Funções*
 gráficos,
 barras verticais, 47
 boxplot, 127

combinados, 54
dispersão, 170
ogiva, 53
pizza, 63
poligonal, veja ogiva
inserir uma fórmula como matriz, 41, 201
preencher sequência, 73
preparando antes de começar, 29
registro de uma função, 30
solver, 29, 278, 426, 455
suplementos, 29
VBA, na maioria dos capítulos
Experimento aleatório, 148

F

Fator de correção finita, 291
Ferramentas de análise do Excel, 21-22
 Ajuste exponencial, 453
 Amostragem, 22
 Anova: fator único, 386
 Anova: fator duplo com repetição, 392
 Anova: fator duplo sem repetição, 390
 Correlação, 186
 Covariância, 184
 Estatística descritiva, 120
 Geração de número aleatório, 254
 Histograma, 57
 como escolher o intervalo de seleção, 59
 sem intervalo de seleção, 60
 gráfico de Pareto, 61
 Média móvel, 449
 Ordem e percentil, 77
 Regressão, 411
 Teste f: duas amostras para variâncias, 372
 Teste t: duas amostras em par para médias, 363
 Teste t: duas amostras presumindo variâncias diferentes, 360
 Teste t: duas amostras presumindo variâncias equivalentes, 356
 Teste Z: duas amostras para média, 351
Formatação condicional, 18, 313
Frequência, 36
Frequências absolutas, distribuição de, 36
Frequências acumuladas, distribuição de, 39
Frequências relativas, distribuição de, 37
Funções estatísticas do Excel,
 ALEATÓRIOENTRE, 12
 COMBIN, 164
 CONT.NÚM, 65
 CONT.SE, 144
 CONT.VALORES, 66
 CONTAR.VAZIO, 66
 CORREL, 176
 COVAR, 172
 CRESCIMENTO, 433
 CRIT.BINOM, 209
 CURT, 119, 137
 DESV.MÉDIO, 109, 134
 DESV.PAD, 114, 135
 DESV.PADA, 136
 DESV.PADP, 114, 135
 DESV.PADPA, 136
 DESVQ, 134
 DISTEXP, 242
 DISTRORÇÃO, 118, 136
 DISTF, 319
 DISTRBINOM, 204
 DISTT, 319
 DIST.BIN.NEG, 214
 DIST.HIPERGEOM, 215
 DIST.LOGNORMAL, 245
 DIST.NORM, 226, 293
 DIST.NORMP, 233, 293
 DIST.QUI, 374
 EPADYX, 408
 FREQUÊNCIA, 40
 INCLINAÇÃO, 400
 INTERCEPÇÃO, 400
 INVF, 367
 INVT, 312, 319
 INT.CONFIANÇA, 307
 INVLOG, 245
 INV.NORM, 237
 INV.NORMP, 237
 INV.QUI, 376
 MAIOR, 100
 MÁXIMO, 64
 MÁXIMO.A, 65
 MED, 80, 104
 MÉDIA, 83, 104
 MÉDIA.A, 104
 MÉDIA.GEOMÉTRICA, 105
 MÉDIA.HARMÔNICA, 105
 MÉDIA.INTERNA, 105
 MENOR, 99
 MODO, 82, 104
 MÍNIMO, 65
 MÍNIMO.A, 65
 ORDEM, 34, 98
 ORDEM.PORCENTUAL, 74, 98
 PADRONIZAR, 233, 292-293
 PEARSON, 176
 PERCENTIL, 75, 100
 PERMUT, 162
 POISSON, 213
 PREVISÃO, 401
 PROB, 209
 PROJ.LIN, 430
 PROJ.LOG, 432
 QUARTIL, 77, 99
 RQUAD, 406
 TENDÊNCIA, 402
 TESTEF, 372
 TESTE.QUI, 377
 TESTET, 355
 TESTEZ, 340
 VAR, 111, 135
 VARA, 135
 VARP, 111, 135

VARPA, 135
 Funções para Banco de Dados do Excel, 139
 resumo das funções, 141
 BDCONTAR, 142
 BDCONTARA, 142
 BDEST, 142
 BDESVPA, 142
 BDEXTRAIR, 143
 BDMÁX, 142
 BDMÉDIA, 142
 BDMÍN, 142
 BDMULTIPL, 142
 BDSOMA, 142
 BDVAREST, 143
 BDVARP, 143
 Funções para procura e referência do Excel,
 CORRESP, 98
 ÍNDICE, 17
 PROCH, 33
 PROCV, 32
 Funções matemáticas do Excel,
 ABS, 108
 ALEATÓRIO, 12
 ARRED, 13
 ARREDONDAR.PARA.BAIXO, 13
 ARREDONDAR.PARA.CIMA, 13
 ARREDMULTB, 13
 EXP, 437
 FATORIAL, 163
 INT, 13
 LN, 436
 MATRIZ.MULT, 92
 RAIZ, 113, 293
 SOMA, 83, 103
 somando produtos, 92
 SOMARPRODUTO, 92, 197
 SOMASE, 145
 SOMAXMY2, 450
 SUBTOTAL, 143
 VFPLANO (financeira), 447
 TRUNCAR, 13

G – H – I

Geração de números aleatórios com Excel, 22
 a partir de distribuições, 253
 com fórmulas, 12
 entre dois números, 12
 com a função ALEATÓRIOENTRE, 12
 Gráfico de barras, *veja* Histograma
 Gráfico de Pareto, 61
 Grande média, *veja* Análise da Variância
 Graus de liberdade, 310
 Hipóteses nula e alternativa, 324
 Histograma, 47
 com dados contínuos, 56
 combinado, 54
 construção com Excel, 47
 dados qualitativos, 61
 ferramenta de análise, 57

 outras representações gráficas, 54
 Pareto, 61
 IBGE, censo 2000, 1
 projeção de indicadores sociais, 3
 IC, *veja* intervalo de confiança
 Inclinação de uma distribuição, 118
 Inferência estatística, 5, 301
 Interseção de eventos, 149
 Intervalo de confiança, 302
 na regressão linear, 409
 simulação do, 304-305
 Intervalo entre quartis, 124
 Boxplot, 176

L – M

Lei de Benford, 155
 Lei dos grandes números, 155
 Linha de tendência do Excel, 439
 média móvel, 449
 polinomial, 443
 regressão linear, 396, 451
 transformação de funções,
 exponencial, 436
 linha de tendência do Excel, 439
 logarítmica, 437
 potência, 438
 resumo das transformações, 439
 Margem de erro numa estimativa, 301
 Média, 82-83
 análise do resultado da, 88
 confiança da estimativa da média, 301
 primeira propriedade da, 84
 comando *Atingir meta*, 85
 soma dos desvios, 85
 segunda propriedade da, 86
 soma dos quadrados dos desvios, 85
 prova do mínimo da, 102
 símbolo somatório, 101
 propriedades operacionais, 101
 visualização das propriedades, 86
 Média amostral, 288
 Média ponderada, 90-91
 Média de longo prazo, 196
 simulador, 197
 Média móvel, projeção pela, 448
 ferramenta de análise, 449
 Mediana, 79
 Medida relativa de dispersão, 116
 coeficiente de variação, 117
 Medidas de dispersão, 116
 coeficiente de variação, 117
 curtose de uma distribuição, 118
 dados suspeitos, 126
 desvio médio absoluto, 108
 desvio padrão, 112
 regra prática, 115
 significado do, 114
 relação entre os desvios padrões, 114
 teorema de Chebyshev, 114

utilizando o Excel, 113
 variância, 109
 funções estatísticas do Excel para, 134
 inclinação de uma distribuição, 118
 outra forma de analisar dispersão, 116
 boxplot, 125
 relativa, 116
 variância, 109
 características da, 112
 outra forma de calcular a, 137
 regras operacionais da, 112
 relação entre as variâncias, 111
 Medidas de ordenamento, 69
 boxplot, 125
 ordem, 72
 percentil, 72
 quartil, 76
 Medidas de tendência central, 69, 79
 média, 82
 análise do resultado da, 88
 da população, 88
 propriedades da, 84
 média ponderada, 90
 mediana, 79
 moda, 81
 vantagens e desvantagens das, 88
 Métodos gráficos, *veja Histograma*
 Moda, 81
 Modelo,
 Ajuste de uma reta, 394
 Amostragem sem Reposição, 33
 Análise Numérica, 119
 Cálculo com DN, 235
 Cálculo dos parâmetros da DN, 238
 Cálculo Inverso com a DN, 237
 Combinação Linear de VA's, 268
 Determinação do tamanho da amostra, 309
 Distribuição Amostral, 294
 Distribuição Binomial, 203
 Distribuição de Poisson, 211
 Distribuição Exponencial, 242
 Distribuição F, 366
 Distribuição Lognormal, 245
 Distribuição Normal, 229
 Distribuição Qui-Quadrado, 374
 DN, 230
 DN Padronizada, 232
 Distribuição t , 310
 Distribuição uniforme, 223
 Estimativa da Média com t , 312
 Estimativa da Média com Z , 307
 F crítico nas duas caudas, 368
 Geral para Estimativas de Médias, 313
 Histogramas, 61
 Intervalo de Projeção, 410
 Poder do Teste de Hipótese, 343
 Probabilidade de Sucesso, 205
 t – Comparação de duas médias, 357-358
 Tamanho da amostra, 309
 Tamanho da amostra população finita, 321
 Teste F – Diferença entre variâncias, 371

Teste de Hipóteses - Distribuição Qui-Quadrado, 375
 TH com Intervalo de Confiança, 328
 TH com p -value, 338
 TH com Z e t , 333
 Visualização propriedades da média, 86

N – O – P

Nível de significância, 326
 Normal, *veja Distribuição normal*
 Normal padronizada, *veja Distribuição normal padronizada*
 Números e dígitos aleatórios, 11
 funções do Excel, 11
 geração com distribuições, 22, 253
 geração com fórmulas, 12
 tabela de, 20
 Ordem de um dado, 70
 Ogiva, ou gráfico poligonal, 54
 Parâmetro, 88, 286
 estatísticas, 88
 Pareto, gráfico de, 61
 Percentil, 71
 Permutações, 162
 Poder de explicação, regressão linear, 405
 Poder do Teste de Hipótese, 342
 Planilha em Excel, 343
 Poisson, *veja Distribuição de Poisson*
 Ponderada, média, 90
 População, 9, 88
 Posição de um dado, 71
 Preencher uma sequência com Excel, 73
 Preparando o Excel antes de começar, 29
 Primeiro quartil, 76
 Probabilidade, 147, 150
 árvore de possibilidades, 155
 definição de, 150
 diagrama de Venn, 148
 espaço amostral, 148
 evento, 148
 coletivamente exaustivos, 150
 complemento de um evento, 149
 elementar, 149
 mutuamente excludentes, 150
 operações com eventos, 149
 interseção, 149
 união, 149
 experimento aleatório, 148
 frequência relativa, 152
 lei dos grandes números, 154
 regra da soma, 156
 regra do produto, 160
 eventos independentes, 160
 simulador lançamento de uma moeda, 152
 análise dos resultados, 153
 tabela de, 158
 técnicas de contagem, 161
 Probabilidade condicional, 157
 Probabilidade conjunta, 158

Probabilidade de acerto, 306
 Probabilidade de erro α , 306
 Probabilidade marginal, 158
 Probabilidade teórica, 151
 Probabilidade total, 158
 Processo de Bernoulli, 201
 Projeção,
 alisamento exponencial, 452
 ajuste da constante alisamento, 455
 ferramenta de análise, 453
 média móvel, 448
 ferramenta de análise, 449
 procedimento inicial, 445
 regressão linear, 451
 taxa média, 446
p-value no teste de hipóteses, 334-335
 cálculo do *p-value*, 336
 definição do *p-value*, 335

Q – R

Quartil, 76
 intervalo entre quartis, 124
 boxplot, 125
 primeiro, segundo e terceiro, 76
 Qui-quadrado, 374
 distribuição, 374
 funções estatísticas da, 374, 376
 Registrando uma função no Excel, 30
 Regra da soma, 156
 Regra do produto, 160
 eventos independentes, 160
 Regressão linear múltipla, 420
 Ferramenta de análise, 420
 Regressão linear simples, 393
 ajuste de uma reta, 394
 linha de tendência para, 396
 modelo de, 394
 método dos quadrados mínimos, 398
 procedimento manual, 394
 coeficiente de determinação, 405
 ajustado, 413
 coeficientes de regressão, 398, 427
 cálculo dos, 399
 com as medidas estatísticas, 403
 com o Solver, 426
 desvios, 398-399
 erro padrão,
 da estimativa, 407
 do coeficiente a , 415
 do coeficiente b , 416
 ferramenta de análise, 411
 função estatística PROJ.LIN, 430
 outras funções estatísticas, 432
 intervalo de projeção, 409
 da média y , 410
 do específico y , 410
 linha de tendência do Excel, 396
 medidas de variação, 404
 outra forma os coeficientes da reta, 403

poder de explicação, 405
 projeção, 401
 premissas do modelo de regressão, 409
 reta de regressão, 398-399
 reta passa pela origem, 418
 resíduos, 419
 teste de hipóteses com a distribuição F , 414
 teste de hipóteses do coeficiente a , 416
 teste de hipóteses do coeficiente b , 416
 variação da estimativa, 409
 Relação entre as variâncias, 111
 Resíduos, *veja* Regressão linear simples
 Reta de regressão, *veja* Regressão linear
 Retirada de um número de uma urna, 13

S

Segundo quartil, 76
 Séries temporais, 9, 444
 projeção,
 alisamento exponencial, 452
 ajuste da constante alisamento, 455
 ferramenta de análise, 453
 média móvel, 448
 ferramenta de análise, 449
 procedimento inicial, 445
 regressão linear, 451
 taxa média, 446
 Símbolo somatório, propriedades, 101
 Simulador,
 Coeficiente de correlação, 179
 Do intervalo de estimação, 304
 Lançamento de uma moeda, 152
 Média de Longo Prazo, 197
 Retirada de um número de uma urna, 13
 Teorema Central do Limite, 289
 Solver, 29, 278, 279, 426, 455
 Somatório, propriedades, 101
 Student, *veja* distribuição t
 Suspeitos, *veja* dados suspeitos

T

Tabela Anova, *veja* Análise da variância
 Tabela de coeficientes de correlação, 183
 Tabela de contingências, 158
 Tabela de covariâncias, 183
 Tabela de números aleatórios, 15
 Tabela de frequências absolutas, 36
 Tabela de frequências acumuladas, 38
 Tabela de frequências relativas, 37
 Tabela de probabilidades,
 conjuntas e marginais, 158
 Tabelas estatísticas, 459
 distribuição F , $\alpha=0,01$, 463
 distribuição F , $\alpha=0,05$, 438
 distribuição *Qui-quadrado*, 465
 distribuição t , 462
 distribuição Z , 461
 números aleatórios, 460

Taxa média, projeção pela, 2, 446

Técnicas de contagem, 161

- combinações, 163
- permutações, 162

Tendência central, *veja* Medidas de

Teorema central do limite, 289

Teorema de Chebyshev, 114

Terceiro quartil, 76

Teste qui-quadrado, 375

- simulação, 375

Testes de hipóteses, regressão linear,

- com a distribuição F , 414
- do coeficiente a , 416
- do coeficiente b , 416

Testes de hipóteses, com duas amostras, 349

- amostras emparelhadas, 362
 - ferramenta de análise para, 363
- amostras grandes e independentes, 350
 - ferramenta de análise para, 351
- amostras pequenas e
 - populações com variâncias diferentes, 358
 - ferramenta de análise para, 360
- amostras pequenas e
 - populações com variâncias iguais, 354
 - ferramenta de análise para, 356
 - função estatística TESTET, 355
- para diferenças entre médias, 349

Teste de hipóteses para médias, 323

- com as distribuições Z ou t , 330
 - com Z , 332
 - com t , 333
 - modelo TH , 333
- com o intervalo de confiança, 326
 - com a distribuição Z , 326
 - com a distribuição t , 328
 - modelo TH , 328
- com o p -value, 334
 - cálculo do, 336
 - com t , 338
 - com Z , 336
 - definição do p -value, 335
 - modelo TH , 338
- erros nos, 326
- função estatística ZTEST para, 339
- hipóteses nula e alternativa, 324
- nível de significância, 326
- numa e nas duas extremidades, 325
- poder do teste, 342
- para diferenças entre médias, 349

Transformação de funções,

- exponencial, 436
- linha de tendência do Excel, 439
- logarítmica, 437
- potência, 438
- resumo das transformações, 439

Transformação linear de uma VA , 257

- combinação linear, 259

U – V

União de eventos, 149

Unidade elementar, 6

VA , *veja*: *Variável aleatória*

Valor esperado da variável aleatória, 196

- covariância como, 219
- desvio padrão como, 199
- simulador de média de longo prazo, 197
- variância como, 199

Valores suspeitos, *veja dados suspeitos*, 126

- boxplot, 125

Variabilidade amostral, 289, 304-305

Variabilidade dos dados, 4

Variação da estimativa, regressão linear, 404

Variância, 109

- análise da, 379
- características da, 112
- da variável aleatória, 199
- outra forma de calcular a , 137
- outra fórmula da, 218
- regras operacionais da, 112
- relação entre as variâncias, 111
- como valor esperado, 200
- outra fórmula da, 218

Variância entre e dentro, 381

Variável, 1, 7

- corte transversal numa data, período, 9
- definição, 6
- número de, 6
- série temporal, 9
- tipo de, 9
- combinação linear, 257

Variável aleatória, 194

- contínua, 221
- discreta, 194
- transformação linear, 257

Variável aleatória discreta, 195-196

- cenários, 195
- definição da, 196
- desvio padrão da, 199
- simulador de média de longo prazo, 197
- valor esperado da, 196
- variância da, 199

Variável aleatória contínua, 221

- premissas, 221
- valor esperado, 222
- variância e desvio padrão, 222

VBA, em todos os modelos em Excel

Visual Basic for Applications, *veja* VBA

VPL de um projeto de investimento,

- Veja*, *Análise do*