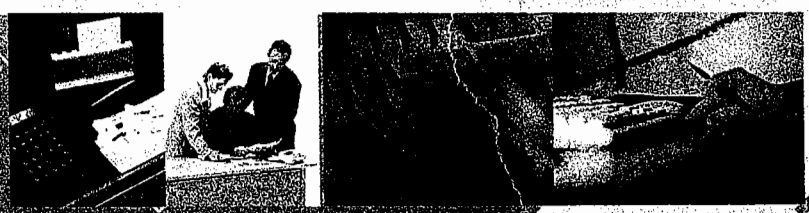
统计学系列

THOMSON

Statistics for Business and Economics

（第8版）

商务与经济统计



戴维·R·安德森（David R. Anderson）

[美] 丹尼斯·J·斯威尼（Dennis J. Sweeney）著

托马斯·A·威廉姆斯（Thomas A. Williams）

王·峰 卿前锋 袁振东 译

王国成 译校

中国出版社

CHINA PUBLISHING HOUSE

图书在版编目（CIP）数据

商务与经济统计/[美]安德森等著；王国成等译．—北京：中信出版社，2003.7
书名原文：Statistics for Business and Economics
ISBN 7-80073-875-2
Ⅰ.商… Ⅱ.①安… ②王… Ⅲ.经济统计－方法 Ⅳ.F222.1
中国版本图书馆 CIP 数据核字(2003)第 065516 号

Statistics for Business and Economics. 8th edition (ISBN: 0-324-06672-4) by Anderson, Sweeney, and Williams
Copyright © 2002 by South-Western, a division of Thomson Learning

Original language published by Thomson Learning (a division of Thomson Learning Asia Pte Ltd). All rights reserved. 本书原版由汤姆森学习出版集团出版。版权所有，盗版必究。

CITIC Publishing House is authorized by Thomson Learning to publish and distribute exclusively this simplified Chinese edition. This edition is authorized for sale in the People's Republic of China only (excluding Hong Kong, Macao SAR and Taiwan). Unauthorized export of this edition is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.
本书中文简体字翻译版由汤姆森学习出版集团授权中信出版社独家出版发行。此版本仅限在中华人民共和国境内(不包括中国香港、澳门特别行政区及中国台湾)销售。未经授权的本书出口将被视为违反版权法的行为。未经出版者预先书面许可，不得以任何方式复制或发行本书的任何部分。

版权所有，侵权必究。
981-240-760-X

商务与经济统计(第 8 版)

SHANGWU YU JINGJI TONGJI

著　　者：[美]安德森　斯威尼　威廉姆斯
译　　者：王　峰　卿前锋　袁晓东
译　　校：王国成
责任编辑：崔　璐　金　晔
出版发行：中信出版社(北京市朝阳区东外大街亮马河南路 14 号塔园外交办公大楼　邮编　100600)
经 销 者：中信联合发行有限公司
承 印 者：北京牛山世兴印刷厂
开　　本：880mm×1230mm　1/16　　印　　张：67.25　　字　　数：1103 千字
版　　次：2003 年 9 月第 1 版　　印　　次：2003 年 9 月第 1 次印刷
京权图字：01-2002-2043
书　　号：ISBN 7-80073-875-2/F・563
定　　价：125.00 元(含光盘)

版权所有・侵权必究

凡购本社图书，如有缺页、倒页、脱页，由发行公司负责退换。服务热线：010－85322521　85322522
E-mail: sales@citicpub.com

前　言

《商务与经济统计》是一本介绍统计学及其应用的教科书，是专门为那些在大学里学习经济学和工商管理的学生们编写的。本教材注重实用，并不需要很深的数学功底，只要有中学代数知识即可。

数据分析和统计方法的应用是本书内容组织和论述中的组成部分。每种分析技术的讨论和扩展都是以应用的方式表述，借助统计结果能加深对问题和答案的理解。

虽然本书是面向应用的，但我们还是要接触到一些稍微深点儿的方法，需要使用在统计领域里常用的一些符号。然而，学生们将会发现本教材为学习较高深的统计知识作了很好的准备，并将精选的一些参考书目作为附录列出以供进一步学习之用。

一、第 8 版中的改进

我们为前几版《商务与经济统计》能被读者接受和引起积极反响而由衷地感到高兴。相应地，在本书的修订中，我们仍然保持了原有的写作风格和可读性，现将新版中一些重要的改进列举如下。

1. 内容的改写

下面有选择地列出新版中经改写的内容。

- 对数据测度水平的论述（第 1 章）
- 有更多图表式描述性统计的内容（第 2 章）
- 用 Excel 软件进行随机抽样的方法（第 7 章）
- 对**边际误差**术语和当应用 z 分布和 t 分布进行区间估计时的附加信息作了初步的讨论
- 更加强调计算和解释 p – 值（第 9—12 章）
- 简化了回归计算公式，淡化概念程式，使其更加直观易懂（第 14 章）
- 修订了包含在 CPI 和 DJIA 中的指标数（第 17 章）

2. 基于实际数据的新的例题和练习

我们增添了近 200 个基于实际数据、源于最新参考文献和统计信息的新的例题和练习。利用

《华尔街杂志》、《今日美国》、《财富》、*Barron's* 和许多网站及其他一些信息源，结合我们的动态研究来扩展统计概念，并通过适当的练习来体会统计在经济学和商务中的多方面应用。我们相信，直接利用实际数据的做法能帮助更多的学生提高对所学内容的兴趣，既能够掌握统计方法又能够掌握它的应用。本教材第 8 版含有近 350 个基于实际数据的例题和练习。

3. 新的案例

这一版我们增加了 4 个案例研究，使本教材中的案例总数增至 26 个。新的案例基于被引用的数据集，出现在区间估计和回归分析两章中。这些案例研究为学生们提供了分析一些较大的数据集并在分析结果的基础上准备管理报告的训练机会。

4. 新的统计实例

每一章的开篇都引用一个统计实例，作为本章所要讲的统计方法的应用举例。这些实例由诸如宝洁、Mead、Dollar General、高露洁—棕榄、宝丽来、孟山都等公司里的实际统计人员提供。这一版又收入了两个新的统计实例：Small Fry Design（第 3 章）和花旗集团（第 5 章）。

5. Minitab 和 Excel 附录

每一章后面基本上都有 Minitab 和 Excel 电子数据表附录。这些附录有利于学生们循序渐进地利用 Minitab 和 Excel 电子表格软件演练每章所讲的统计分析方法。所有的附录都按最新版本的 Minitab 和 Excel 格式编写。目前这一版描述性统计（第 2 章）中的 Excel 附录包含了变幻莫测的函数、图表和关键报表的应用。这一做法能培养学生们运用 Excel 设制图表和绘制交互表格的能力。在描述统计、抽样、区间估计、假设检验和独立性检验等章后面共有 10 个新的或修改过的表格附录。

二、教材特色

我们沿用了前几版中的许多做法，其中一些重要特征如下。

1. 方法练习和应用练习

每节结束时的练习被分成方法和应用两部分：方法练习要求学生们利用公式做一些必要的计算；应用练习旨在使学生能在实际场合中运用所学的知识。因此，学生们应该首先掌握计算的核心和重点，然后再来体会统计应用和解释中的精妙之处。

2. 自测题

本书设置了一定量的自测题，这些练习题的全部解答作为附录 E 放在本书最后。读者可通过做自测题并直接与解答对比检查，估计自己对每章所学内容的理解程度。

3. 评注

在许多章节后面，我们都给出了一些评注（注解和评论），以加深学生们对统计方法及其应用的理解。注解和评论包括统计方法应用时的一些注意事项和限制、应用范围的扩展、相关技术的简述等内容。

4. 与教材相应的数据集

本教材附有一张包容 100 多个数据集的 CD 光盘。数据集可用于 Mintab 和 Excel 两种格式。教材中对 CD 光盘中所有的数据集都加注了标识符，所有案例研究的数据集以及较大的练习的数据集都包含于 CD 光盘中。

三、致谢

我们非常感谢一直关注本书的再版工作并为提高再版水平提出宝贵建议和评论的人们，他们是：

Lari Arjomand, Clayton College and State University

James Brannon, University of Wisconsin – Oshkosh

Donald Gren, Salt Lake Community College

Clifford Hawley, West Virginia University

Ronald Klimberg, St. Joseph's University

Bala Maniam, Sam Houston State University

Ceyhun Ozgur, Valparaiso University

H. V. Ramakrishna, Penn State University at Great Valley

我们还是要一如既往地感谢我们的同事和朋友，他们对本书的前几版和这一版都提出了许多建设性意见和评论，使我们受益匪浅：

Mohammad Ahmadi, University of Tennessee at Chattanooga

Robert Balough, Clarion University

Mike Bourke, Houston Baptist University

John Bryant, University of Pittsburgh

Peter Bryant, University of Colorado
Terri L. Byczkowski, University of Cincinnati
Robert Carver, Stonehill College
Robert Cochran, University of Wyoming
David W. Cravens, Texas Christian University
Robert Carver, Stonehill College
Robert Collins, Marquette University
Tom Dahlstrom, Eastern College
Gopal Dorai, William Patterson University
Nicholas Farnum, California State University – Fullerton
Paul Guy, California State University – Chico
Alan Humphrey, University of Rhode Island
Ann Hussein, Philadelphia College of Textiles and Science
C. Thomas Innis, University of Cincinnati
Ben Isslhardt, Rochester Institute of Technology
Jeffery Jarrett, University of Rhode Island
David Krueger, St. Cloud State University
Martin S. Levy, University of Cincinnati
David Lucking-Reiley, Vanderbilt University
Tom McCulloagh, University of California – Berkeley
Don Marx, University of Alaska, Anchorage
Glenn Nilligan, Ohio State University
Mitchell Muesham, Sam Houston State University
Roger Myerson, Northwestern University
Richard O' Connell, Miami University of Ohio
Alan Olinsky, Bryant college
Tom Pray, Rochester Instiute of Technology
Harold Rahmlow, St. Joseph's University
Tom Ryan, Case Western Reserve University
Alan Smith, Robert Morris College
Bill Seaver, University of Tennessee
Willbann Terpening, Gonzaga University
David Tufte, University of New Orleans
Ted Tsukahara, St. Mary's College of California
Hroki Tsurumi, Rutgers University
Victor Ukpolo, Austin Peay State University
Ebenge Usip, Youngstown State University

Cindy VanEs, Cornell University
Jack Vaughn, University of Texas – El Paso
Andrew Welki, John Carroll University
Ari Wijetunga, Morehead State University
J. E. Willis, Louisiana State University
Mustafa Yilmaz, Northeastern University
Gary Yoshimoto, St. Cloud State University
Charles Zimmerman, Robert Morris College

还要特别感谢为我们提供统计实例的来自工商界的合作者，虽然我们只是分别通过一篇篇报刊文章的署名栏中知道他们的。最后，我们也对在本教材再版过程中做了大量有益工作的，South-Western/Thomson Learning 出版公司的文稿征集部资深编辑 Charles McCormick、发展部资深编辑 Alice Denny、制作部资深编辑 Deanna Quinn 和资深的销售经理 Joe Sabatino 以及为本书出版付出辛勤劳动的其他编辑朋友们，表示深深的谢意。

戴维·R·安德森
丹尼斯·J·斯威尼
托马斯·A·威廉姆斯

作者简介

戴维·R·安德森 戴维·R·安德森是辛辛那提大学工商管理学院的数量分析教授。他出生于美国北达科他州的 Grand Forks，在普度大学获得学士、硕士和博士学位。安德森教授担任过数量分析和运筹管理系主任和工商管理学院的院长助理，还曾任第一届院务委员会的合作协调人。

在辛辛那提大学，安德森教授为工商管理专业的学生们讲授统计引论以及研究生水平的回归分析、多变量分析和管理科学等课程，也曾给华盛顿大学劳动系讲授过统计课程。他获得过优秀教学奖和先进工作者等荣誉奖励。

安德森教授已经合作编写出版了 9 本统计、管理科学、线性规划、生产和运营管理等方面的教材。他是抽样和统计方法领域内非常活跃的咨询专家。

丹尼斯·J·斯威尼 丹尼斯·J·斯威尼是辛辛那提大学数量分析教授和生产力进步研究中心主任。他出生于美国衣阿华州的 Des Moines，从德雷克大学获得工商管理学士学位，从印地安纳大学获得工商管理硕士和博士学位，并任该校的国防教育法学会会员。1978—1979 年，斯威尼教授在宝洁公司的管理科学小组工作；1981—1982 年在杜克大学做访问教授；还担任过辛辛那提大学的数量分析系主任和工商管理学院的院长助理。

斯威尼教授已经发表了管理科学和统计领域的 30 多篇论文和专题研究报告，得到过由美国国家科学基金、IBM、美国联邦储备局、宝洁公司、Kroger 和辛辛那提电气公司等方面提供的研究资助，在《管理科学》(*Management Science*)、《运筹学研究》(*Operations Research*)、《数学规划》(*Mathematical Programming*)和《决策科学》(*Decision Sciences*)等杂志上发表多篇论文。

斯威尼教授已经合作编写出版了 9 本统计、管理科学、线性规划、生产和运营管理等方面的教材。

托马斯·A·威廉姆斯 托马斯·A·威廉姆斯是罗切斯特技术研究院（RIT）工商学院的管理科学教授。他出生于美国纽约州的 Elmira，在 Clarkson 大学获得学士学位，然后到 Rensselaer 综合技术研究院做研究工作，并在那里获得硕士和博士学位。

在加盟 RIT 工商学院之前，威廉姆斯教授在辛辛那提大学的工商管理学院任教 7 年，并在那里进修了信息系统的研究生课程，同时还任协调人。在 RIT，他是管理科学系的第一任系主任，他讲授管理科学和统计类课程，也讲授回归分析和决策分析的研究生课程。

威廉姆斯教授在管理科学、统计、生产和运营管理以及数学等专业领域合作出版了 10 本教材，并对将数据分析运用到扩展大规模回归模型的方案排序问题做了专门研究。《财富》500 强中的许多企业都请威廉姆斯教授做过咨询。

要目

■第 1 章　数据和统计　/1
■第 2 章　描述性统计：表格法和图形法　/25
■第 3 章　描述性统计：数值方法　/81
■第 4 章　概率论简述　/147
■第 5 章　离散概率分布　/191
■第 6 章　连续概率分布　/231
■第 7 章　抽样和抽样分布　/263
■第 8 章　区间估计　/309
■第 9 章　假设检验　/353
■第 10 章　两个总体均值和比例的统计推断　/415
■第 11 章　关于总体方差的推断　/457
■第 12 章　拟合度检验和独立性检验　/485
■第 13 章　方差分析与试验设计　/515
■第 14 章　简单线性回归　/577
■第 15 章　多元回归　/657
■第 16 章　回归分析：建立模型　/715
■第 17 章　指数　/773
■第 18 章　预测　/797
■第 19 章　非参数方法　/849
■第 20 章　质量管理统计方法　/889
■第 21 章　抽样调查　/921
■附录 A　参考文献　/960
■附录 B　表格　/962
■附录 C　求和符号　/990
■附录 D　偶数习题的答案　/992
■附录 E　自测题解答　/1016
■译后记　/1056

目 录

第1章 数据和统计 /1

统计实例：商业周刊 /2

1.1 在商务和经济中的应用 /3
• 会计 • 财务 • 营销 • 生产 • 经济

1.2 数据 /5
• 元素、变量和观测值 • 度量尺度 • 定性和定量数据 • 截面数据和时间序列数据

1.3 数据来源 /8
• 现存数据源 • 统计研究 • 数据采集误差

1.4 描述性统计 /13

1.5 统计推断 /14
• 小结 • 术语释义 • 练习

第2章 描述性统计：表格法和图形法 /25

统计实例：高露洁—棕榄公司 /26

2.1 汇总定性数据 /27
• 频数分布 • 相对频数和百分比频数分布 • 柱状图和饼形图 • 练习

2.2 汇总定量数据 /34
• 频数分布 • 相对频数和百分比频数分布 • 点位图 • 直方图 • 累积分布 • 累积曲线 • 练习

2.3 探索性数据分析：茎叶显示 /44
• 练习

2.4 交叉分组列表和散点图 /49
• 交叉分组列表 • 散点图 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

案例研究 CONSOLIDATED 食品公司 /65

附录 2.1 在表格和图形方法中应用 Minitab 软件 /66

附录 2.2 在表格和图形方法中应用 Excel 软件 /69

第3章 描述性统计：数值方法 /81

统计实例：SMALL FRY DESIGN 公司 /82

3.1 位置的度量 /83
• 均值 • 中位数 • 众数 • 百分位数 • 四分位数 • 练习

3.2 变异程度的度量 /93
• 极差 • 四分位点内距 • 方差 • 标准差 • 变异系数 • 练习

3.3 相对位置的度量和异常值的检测 /100
• z-分数 • 切比雪夫定理 • 经验法则 • 异常值检测 • 练习

3.4 探索性数据分析 /107
• 五数概括法 • 箱形图 • 练习

3.5 两变量间关系的度量 /112
• 协方差 • 协方差的解释 • 相关系数 • 相关系数的解释 • 练习

3.6 加权平均值和使用分组数据 /121
• 加权平均值 • 分组数据 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

案例研究 1 CONSOLIDATED 食品公司 /137

案例研究 2 国家健康护理协会 /138

案例研究 3 亚太地区的商学院 /139

附录 3.1 利用 Minitab 计算描述性统计量 /141

附录 3.2 利用 Excel 计算描述性统计量 /144

第 4 章　概率论简述　147

统计实例：MORTON 国际公司　/148

4.1　试验、计数法则和概率指派　/149
• 计数法则、组合和排列 • 概率指派 • 肯塔基电力公司项目的概率 • 练习

4.2　事件和它们的概率　/159
• 练习

4.3　概率的一些基本关系　/163
• 事件的补 • 加法法则 • 练习

4.4　条件概率　/169
• 独立事件 • 乘法法则 • 练习

4.5　贝叶斯定理　/176
• 表格方法 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

案例研究　HAMILTON 县的法官　/188

第 5 章　离散概率分布　/191

统计实例：花旗银行　/192

5.1　随机变量　/193
• 离散随机变量 • 连续随机变量 • 练习

5.2　离散概率分布　/195
• 练习

5.3　数学期望和方差　/201
• 数学期望 • 方差 • 练习

5.4　二项概率分布　/206
• 二项试验 • 马丁服装商店问题 • 使用二项概率表 • 二项概率分布的数学期望和方差 • 练习

5.5　泊松概率分布　/217
• 一个涉及时间间隔的例子 • 一个涉及长度或距离间隔的例子 • 练习

5.6　超几何概率分布　/221
• 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

附录 5.1 利用 Minitab 计算离散随机概率　/228

附录 5.2 利用 Excel 计算离散概率分布　/229

第 6 章　连续概率分布　/231

统计实例：宝洁公司　/232

6.1　均匀概率分布　/233
• 作为概率度量的面积 • 练习

6.2　正态概率分布　/238
• 正态曲线 • 标准正态概率分布 • 计算任意正态概率分布的概率 • Grear 轮胎公司问题 • 练习

6.3　指数概率分布　/252
• 计算指数分布的概率 • 泊松分布与指数分布的关系 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

附录 6.1　利用 Minitab 计算连续概率分布　/259

附录 6.2　利用 Excel 计算连续概率分布　/260

第 7 章　抽样和抽样分布　/263

统计实例：MEAD 公司　/264

7.1　Electronics Associates 公司的抽样问题　/265

7.2　简单随机抽样　/266
• 从有限总体中抽样 • 从无限总体中抽样 • 练习

7.3　点估计　/272
• 练习

7.4　抽样分布介绍　/275

7.5　$\bar{x}$ 的抽样分布　/279
• $\bar{x}$ 的数学期望 • $\bar{x}$ 的标准差 • 中心极限定理 • EAI 抽样问题中 $\bar{x}$ 的抽样分布 • $\bar{x}$ 抽样分布的实际值 • $\bar{x}$ 抽样分布和样本容量之间的关系 • 练习

7.6　$\bar{p}$ 的抽样分布　/289
• $\bar{p}$ 的数学期望 • $\bar{p}$ 的标准差 • $\bar{p}$ 抽样分布的形式 • $\bar{p}$ 抽样分布的实际值 • 练习

7.7　点估计量的性质　/294
• 无偏性 • 有效性 • 一致性

7.8 其他抽样方法 /297
• 分层随机抽样 • 整群抽样 • 系统抽样 • 方便抽样 • 判断抽样 • 小结 • 术语释义 • 重要公式 • 补充练习
附录 7.1 $\bar{x}$ 的数学期望与标准差 /304
附录 7.2 利用 Minitab 进行随机抽样 /306
附录 7.3 利用 Excel 进行随机抽样 /307

第 8 章 区间估计 /309
统计实例：DOLLAR GENERAL 公司 /310
8.1 总体均值的区间估计：大样本情况 /311
• CJW 估计问题 • 抽样误差 • 假定 σ 已知的大样本情况 • 用 s 估计 σ 的大样本情况 • 练习
8.2 总体均值的区间估计：小样本情况 /320
• 假定 σ 已知的小样本情况 • 用 s 估计 σ 的小样本情况 • 总体分布的作用 • 练习
8.3 确定样本容量 /329
• 练习
8.4 总体比例的区间估计 /332
• 确定样本容量 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1 Bock 投资服务公司 /343
案例研究 2 GULF 房地产公司 /345
案例研究 3 大都会研究公司 /347
附录 8.1 用 Minitab 求总体均值的区间估计 /347
附录 8.2 用 Excel 进行总体均值的区间估计 /349

第 9 章 假设检验 /353
统计实例：HARRIS 公司 /354
9.1 原假设和备择假设的建立 /355
• 检验研究中的假设 • 检验声明的有效性 • 决策中的假设检验 • 对原假设和备择假设形式的总结 • 练习
9.2 第一类错误和第二类错误 /358
• 练习
9.3 总体均值的单侧检验：大样本 /360
• 检验统计量应用 • p - 值的应用 • 总体均值的单侧检验步骤总结 • 假设检验步骤 • 练习
9.4 总体均值的双侧检验：大样本 /371
• 双侧检验的 p - 值 • 总体均值的双侧检验步骤总结 • 区间估计与假设检验的关系 • 练习
9.5 总体均值的检验：小样本 /379
• p - 值和 t 分布 • 双侧假设检验 • 练习
9.6 总体比例的检验 /385
• 练习
9.7 假设检验和决策支持 /392
9.8 计算第二类错误的概率 /393
• 练习
9.9 确定总体均值假设检验的样本容量 /398
• 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1 Quality Associates 有限公司 /407
案例研究 2 失业研究 /408
附录 9.1 用 Minitab 进行假设检验 /409
附录 9.2 用 Excel 进行假设检验 /411

第 10 章 两个总体均值和比例的统计推断 /415
统计实例：FISONS 公司 /416
10.1 两个总体均值之差的估计：独立样本 /417
• $\bar{x}_1 - \bar{x}_2$ 的抽样分布 • $\mu_1 - \mu_2$ 的区间估计：大样本情形 • $\mu_1 - \mu_2$ 的区间估计：小样本情形 • 练习
10.2 两个总体均值之差的假设检验：独立样本 /426
• 大样本情形 • 小样本情形 • 练习

10.3　两个总体均值之差的统计推断:匹配样本 /435
• 练习
10.4　两个总体比例之差的统计推断 /441
• $\bar{p}_1-\bar{p}_2$ 的抽样分布 • p_1-p_2 的区间估计 • 关于 p_1-p_2 的假设检验 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究　Par 公司 /453
附录 10.1　用 Minitab 进行两个总体均值的假设检验 /453
附录 10.2　用 Excel 进行两个总体均值的假设检验 /455

第 11 章　关于总体方差的推断 /457
统计实例：美国会计总署 /458
11.1　关于一个总体方差的统计推断 /459
• σ^2 的区间估计 • 假设检验 • 练习
11.2　两个总体方差的统计推断 /469
• 练习 • 小结 • 重要公式 • 补充练习
案例研究　空军培训 /479
附录 11.1　用 Minitab 计算总体方差 /480
附录 11.2　用 Excel 计算总体方差 /481

第 12 章　拟合度检验和独立性检验 /485
统计实例：UNITED WAY /486
12.1　多样本总体的拟合度检验 /487
• 练习
12.2　独立性检验 /491
• 练习
12.3　拟合度检验：泊松分布与正态分布 /498
• 泊松分布 • 正态分布 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究　两党议程变更 /511
附录 12.1　用 Minitab 进行拟合度检验和独立性检验 /512
附录 12.2　用 Excel 进行拟合度检验和独立性检验 /513

第 13 章　方差分析与试验设计 /515
统计实例：BURKE 市场调研公司 /516
13.1　方差分析导论 /517
• 方差分析的假定 • 概念总览
13.2　方差分析：k 个总体均值相等性的检验 /520
• 总体方差组间估计 • 总体方差组内估计 • 方差估计量的比较：F 检验 • 方差分析表 • 方差分析的计算结果 • 练习
13.3　多重比较方法 /530
• 费舍尔 LSD 方法 • 第一类错误的概率 • 练习
13.4　试验设计介绍 / 535
• 数据收集
13.5　纯随机设计 / 537
• 总体方差组间估计 • 总体方差组内估计 • 方差估计量的比较：F 检验 • 方差分析表 • 配对比较 • 练习
13.6　随机区组设计 /543
• 空中交通管制员的工作压力测试 • 方差分析方法 • 计算与结论 • 练习
13.7　因子试验 /549
• 方差分析方法 • 计算与结论 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1　WENTWORTH 医疗中心 /568
案例研究 2　工业产品销售员的报酬 /569
附录 13.1　用 Minitab 进行方差分析和试验设计 /571
附录 13.2　用 Excel 进行方差分析和试验设计 /572

第 14 章　简单线性回归 /577
统计实例：宝丽来公司 /578

14.1 简单线性回归模型 /579
• 回归模型和回归方程 • 估计回归方程
14.2 最小二乘法 /582
• 练习
14.3 判定系数 /592
• 相关系数 • 练习
14.4 模型假设 /599
14.5 显著性检验 /601
• σ^2 的估计 • t 检验 • β_1 的置信区间 • F 检验 • 关于显著性检验解释的几点注意 • 练习
14.6 利用估计回归方程进行估计和预测 /610
• 点估计 • 区间估计 • y 的均值的置信区间估计 • 个别 y 值的预测区间估计 • 练习
14.7 计算机方法 /616
• 练习
14.8 残差分析：验证模型假定 /621
• 关于 x 的残差图 • 关于 $\hat{y}$ 的残差图 • 标准残差 • 正态概率图 • 练习
14.9 残差分析：异常值和有影响的观察值 /630
• 检测异常值 • 检测有影响的观察值 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1 教育支出和学生成绩 /646
案例研究 2 美国交通局 /648
案例研究 3 校友捐赠 /649
附录 14.1 最小二乘法的推导 /651
附录 14.2 相关系数的显著性检验 /652
附录 14.3 用 Minitab 进行回归分析 /653
附录 14.4 用 Excel 进行回归分析 /654

第 15 章 多元回归 /657
统计实例：CHAMPION 国际有限公司 /658
15.1 多元回归模型 /659
• 回归模型和回归方程 • 多元估计回归方程
15.2 最小二乘法 /660
• 示例：Butler 运输公司 • 回归系数解释 • 练习
15.3 多元判定系数 /670
• 练习
15.4 模型假设 /673
15.5 显著性检验 /674
• F 检验 • t 检验 • 多重共线性 • 练习
15.6 利用估计回归方程进行估计和预测 /682
• 练习
15.7 定性自变量 /684
• 示例：Johnson 过滤水公司 • 参数解释 • 更为复杂的定性变量 • 练习
15.8 残差分析 /693
• 检测异常值 • 按学生分布删除残差和异常值 • 有影响的观测值 • 利用库克距离测度识别有影响的观测值 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1 消费者调查有限公司 /709
案例研究 2 全美橄榄球联盟四分位等级排序 /710
案例研究 3 对学生测验成绩进行预测 /712
案例研究 4 Alumni 基金会 /713

第 16 章 回归分析：建立模型 /715
统计实例：孟山都公司 /716
16.1 一般线性模型 /717
• 模拟曲线关系 • 交互作用 • 包含因变量的转换 • 内涵线性的非线性模型 • 练习
16.2 增加或者删除变量的时间确定 /732
• 一般情形 • p – 值的应用 • 练习
16.3 大型问题分析 /739

16.4　变量选择方法　/743
• 逐步回归 • 前向选择 • 后向消元 • 最佳子集回归 • 最终选择 • 练习
16.5　残差分析　/751
• 自相关性和杜宾 – 瓦特森检验 • 练习
16.6　关于方差分析和试验设计的多元回归方法　/758
• 练习 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1　失业问题研究　/768
案例研究 2　职业高尔夫巡回赛统计量分析　/770
案例研究 3　预测大学的毕业率　/771

第 17 章　指数　/773
统计实例：美国劳工部劳工统计局　/774
17.1　比价　/775
17.2　综合物价指数　/775
• 练习
17.3　根据比价计算综合物价指数　/780
• 练习
17.4　一些重要的价格指数　/782
• 消费者价格指数 • 生产者价格指数 • 道琼斯股票平均价格指数
17.5　根据物价指数减缩序列　/785
• 练习
17.6　物价指数：其他一些相关问题　/788
• 商品项目的选择 • 基期的选择 • 质量变化
17.7　物量指数　/790
• 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

第 18 章　预测　/797
统计实例：内华达职业健康诊所　/798
18.1　时间序列成分　/799
• 趋势成分 • 循环成分 • 季节成分 • 不规则成分
18.2　平滑法　/802
• 移动平均法 • 加权移动平均法 • 指数平滑法 • 练习
18.3　趋势推测法　/813
• 练习
18.4　趋势和季节成分　/820
• 乘法模型 • 计算季节指数 • 消除季节影响的时间序列 • 利用消除了季节影响的时间序列来分析趋势 • 季节调整 • 根据月度资料建立模型 • 循环成分 • 练习
18.5　回归分析　/830
18.6　定性预测方法　/832
• 德尔菲法 • 专家判断法 • 远景规划法 • 直观预测法 • 小结 • 术语释义 • 重要公式 • 补充练习
案例研究 1　预测食品和饮料的销量　/841
案例研究 2　预测销量的损失　/842
附录 18.1　用 Minitab 进行预测　/844
附录 18.2　用 Excel 进行预测　/846

第 19 章　非参数方法　/849
统计实例：西壳房地产经纪公司　/850
19.1　符号检验　/852
• 小样本情形 • 大样本情形 • 中位数假设检验 • 练习
19.2　维尔克科森符号秩检验　/859
• 练习
19.3　曼 – 怀特尼 – 维尔克科森检验　/865
• 小样本情形 • 大样本情形 • 练习
19.4　克鲁斯卡尔 – 沃利斯检验　/874
• 练习
19.5　秩相关　/878
• 秩相关显著性检验 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

第 20 章　质量管理统计方法　/889

统计实例：美国道化学品公司　/890

20.1　统计过程管理　/891

• 控制图 • $\bar{x}$ 控制图：已知过程均值和标准差 • $\bar{x}$ 控制图：未知过程均值和标准差 • R 控制图 • p 控制图 • np 控制图 • 控制图的解释 • 练习

20.2　抽样验收　/906

• KALI 有限公司：抽样验收实例 • 计算批量验收的概率 • 选择抽样验收方案 • 多重抽样方案 • 练习 • 小结 • 术语释义 • 重要公式 • 补充练习

附录 20.1　用 Minitab 绘制控制图　/920

第 21 章　抽样调查　/921

统计实例：辛辛那提电气公司　/922

21.1　抽样调查所用术语　/922

21.2　调查种类与抽样方法　/923

21.3　调查误差　/925

• 非抽样误差 • 抽样误差

21.4　简单随机抽样　/926

• 总体均值 • 总体总量 • 总体比率 • 确定样本容量 • 练习

21.5　分层简单随机抽样　/933

• 总体均值 • 总体总量 • 总体比率 • 确定样本容量 • 练习

21.6　整群抽样　/942

• 总体均值 • 总体总量 • 总体比率 • 确定样本容量 • 练习

21.7　系统抽样　/950

• 小结 • 术语释义 • 重要公式 • 补充练习

附录 A　参考文献　/960

附录 B　表格　/962

附录 C　求和符号　/990

附录 D　偶数习题的答案　/992

附录 E　自测题解答　/1016

译后记　/1056

1 数据和统计

统计实例:商业周刊

1.1 在商务和经济中的应用

会计

财务

营销

生产

经济

1.2 数据

元素、变量和观测值

度量尺度

定性和定量数据

截面数据和时间序列数据

1.3 数据来源

现存数据源

统计研究

数据采集误差

1.4 描述性统计

1.5 统计推断

统计实例

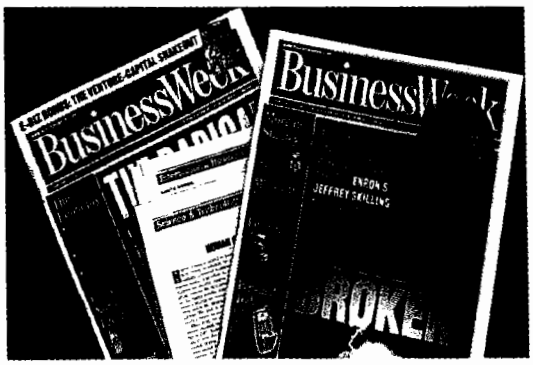
商业周刊*

纽约市，纽约州

《商业周刊》（*Business Week*）全球发行量超过100万份，是世界上拥有最广泛读者的商业类杂志，拥有200多名忠于职守的记者、编辑和26个部门分布在世界各国家，主要登载经济和商贸活动中饶有兴趣、内容丰富的文章。该杂志不仅刊登谈论热门话题的特色文章，还常设一些关于国际商贸、经济分析、信息处理和科技进展等栏目。特色文章和常设栏目所传递的信息能帮助读者及时透视时事风云变幻，以便估计其对经济商务活动环境产生的影响。

《商业周刊》几乎每一期都发表一份关于当前热门话题的深度报告，这样的深度报告常常包含一些统计论据和概述，以帮助读者理解商务和经济信息。例如，2000年1月17日那一期发表了一篇“提高利率谁受害最大”的报告；紧接着的一期（2000年1月24日）发表了一篇“共同基金何错之有”的报告；2000年2月28日那一期发表了一篇关于期权的报告。另外，每周一期的《商业周刊投资者》（*Business Week Investor*）都要发布关于产品指数、股票价格、共同基金和利率等方面的经济走势统计。

《商业周刊》中的统计报告和统计信息也能帮助经理们做好经营管理工作。例如，一份关于订阅者的年度综述报告可帮助公司了解订阅总人数、读者口味、可能达到的订阅量、生活风格，等等。《商业周刊》的经理们利用从综述报告中得出的统计结论更好地为订户和广告商服务。例如，有一个时期对北美订户的调查统计情况表明，有90%的订户愿意购买家庭电脑，64%的订户为了工作也需要购买电脑。这一统计结果提醒《商业周刊》的经理们注意到，订阅者会对关于电脑发展趋势方面的文章感兴趣，也会引起潜在的广告商的兴趣。这么高百分比的家庭电脑和工作用电脑的潜在购买者无疑对电脑制造商在《商业周刊》上刊登广告会产生极大的诱惑和激励。



《商业周刊》在许多文章中利用统计事例和汇总资料。© Joe Higgins/South-Western.

在本章中，我们讨论可用于统计分析的数据类型，并说明这些数据是如何获取的。我们介绍描述性统计和统计推断，它们是将数据转化为有用的而且容易被解释的统计信息的方法。

---

*作者衷心感谢《商业周刊》研究部经理 Charlene Trentham 提供这一统计实例。

我们经常能在报纸和杂志上的文章中看到如下的一些语句：

- 思科公司(Cisco)，互联网基础设备制造商，在纳斯达克星期一收盘时的市值为5 550亿美元（*USA Today*，March 28, 2000）。
- 50家最大公司的CEO的年平均收入是930万美元（*Forbes*，April 3, 2000）。

- 电子购物网站每招徕一位顾客的平均支出是 108 美元(*Business 2. 0,* March 2000)。
- 《华盛顿邮报》(*Washington Postreaches*)本地区订户比例平时工作日达到 46%，周末达到 61%，是大城市报纸发行量的顶尖大户（*Fortune*, January l0, 2000）。
- 股票平均能使 75% 的投资者盈利（*The Wall Street Journal*, March 27, 2000)。
- 1999 年赛季职业棒球队比赛门票的平均价格是 14. 91 美元(*USA Today*, April 15, 1999)。

上述语句中的量化数据（5 550 亿美元、930 万美元、108 美元、46% 、61% 、75% 和 14. 91 美元）就称为**统计**(statistics)。因此，在日常用法中，统计这一术语主要是指这些定量数据。然而，统计学的研究领域或对象的内涵更加丰富。广义上讲，统计是收集、分析、提炼和解释数据的科学与艺术。特别是在商务和经济中的统计应用，收集、分析、提炼和解释数据旨在帮助经理和决策者更好地把握商务和经济环境的变化，以便科学、正确地作出决策。在本教材中，我们强调统计在商务和经济决策方面的用途。

第 1 章先就统计在商务和经济方面的应用作了一些介绍。1. 2 节中明确界定了数据这一术语的含义，并给出数据集的概念。这一节也介绍了像**变量**(variables)和**观测值**(observations)等一些关键术语，讨论了定量数据和定性数据之间的区别，也说明了截面数据和时间序列数据之间的区别。1. 3 节探讨了如何从现实活动中获取数据或通过专门的实验性研究获得新数据。对如今在数据获取中起着重要作用的互联网，也给予了高度关注。数据在描述性统计和统计推断中的进一步用途将在 1. 4 节和 1. 5 节中予以论述。

1. 1　在商务和经济中的应用

在当今贸易和经济全球化的大背景下，需要应用大量的统计信息。最成功的经理和决策者就是能真正明白信息的作用并能有效地利用它们的人。这一节我们阐述了统计应用于商务和经济中的几个例子。

会　计

公共会计公司接受客户委托按统计抽样程序进行审计稽核。举例来说，一个会计公司要确定客户提供的账单上显示的收入账目是否真实地反映了实际收入。一般情况下，每一客户提供的账单要包括很多笔账目，一笔一笔地审核和确认太耗费时间并且代价太高，在这种情况下通常的做法是，由审计机构从账单上选择一些子集作为样本。在重新审核所抽出的样本的准确性后，审计师就可得出结论，以确定是否可以认为客户账单上的收入账目是可接受的。

财　务

财务顾问们利用多方面的统计信息进行投资指导。在股票市场上，顾问们综合了包括市盈率和股息等多方面的信息，在将个股信息与股市的平均信息比较之后，就可作出某一只股票是涨还

是跌的判定。例如，*Barron's*（January 10, 2000）报告，30 只股票的道琼斯工业平均指数的市盈率平均值是 24.7，菲利普·莫里斯(Philip Morris)的市盈率是 9。此时关于市盈率的统计信息表明，与道琼斯股票的平均情况相比，菲利普·莫里斯的股价偏低。于是，财务顾问就有可能得出菲利普·莫里斯股票当前正处于下跌时期的结论。这些和其他方面的信息帮助顾问们作出是买、是卖还是持股的建议。

营　销

零售付款机上的电子扫描仪是用来收集各种营销方案实际效果数据的。例如，像 ACNielsen 和信息资源公司(Information Resources, Inc.)那样的数据供应商从商店购买某一时点上的扫描数据，经过加工处理、作出数据统计汇总后再卖给制造商。制造商为获得每一产品目录的这种类型的扫描数据平均要花费 387 325 美元（Scanner Data User Survey, Mercer Management Consulting, Inc., April 1997）。制造商也购买诸如特殊定价和设立店内显示牌等促销活动效果的数据和统计研究报告。部门经理通过分析扫描统计和促销活动统计数据能明白促销措施与实际销售额之间的关系。这样的分析对制定各种产品未来的营销战略是大有裨益的。

生　产

由于已进入重视质量的时代，所以质量控制是统计在生产中非常重要的应用。大量的质量控制统计表用于监控生产过程的产出。特别是，x - 直方图用于监控产出的平均值。例如，假设有一台 12 盎司标准的软饮料灌装机，定期地从产品中选取一些软包装饮料作为样本，计算出样本灌装量的平均值。将平均值或 x - 直方图的值标绘在 x - 直方图上：如果标绘值在 x - 直方图上方控制线之上，就表明该批产品灌装过量；如果标绘值在 x - 直方图下方控制线之下，就表明该批产品灌装不足。这一过程处于“控制”之中，只要标绘值在 x - 直方图上方控制线和下方控制线之间，就允许继续生产。更贴切的解释是，x - 直方图能够帮助确定何时必须调整和校正生产过程。

经　济

人们会经常要求经济学家提供对未来经济走势或某些特殊领域的发展情况的预测。他们在预测时需要用到多方面的统计信息。举例来说，经济学家利用生产价格指数、失业率和生产能力利用情况等像指示器那样反映出来的统计信息，这些统计指示值常常被输入到计算机预测模型中用来预测通货膨胀率。

像这一节中介绍的描述性统计的应用是本教材的组成部分。这些例子可使我们对统计的广泛应用有一个大致的了解。为列举这些补充例子，我们已经请商务和经济领域的实际工作者提供统计应用实例作为每章的开篇，以便引出该章所要讲解的内容。统计应用实例充分显示出统计在各行各业制定决策时的重要性。

1.2　数　据

数据(Data)是为了表述和解释现实问题所收集、分析和汇总的事实依据与图表。为便于研究，将用于特定研究而采集的数据结集在一起，称之为**数据集**(data set)。表 1.1 列示出一个包含 25 只影子股票的数据集，这 25 只影子股票是由美国个体投资者协会跟踪调查的。影子股票是较小公司的普通股票，不会受到华尔街分析家们的密切关注。


CD 光盘数据
Shadow

表 1.1　25 只影子股票的数据集

公司名称	交易所	股票代号	年发售额（百万美元）	每股盈利（美元）	价格/盈利率（市盈率）
Advanced Comm. Systems	OTC	ACSC	75.10	0.32	39.10
Ag-Chem Equipment Co.	OTC	AGCH	321.10	0.48	23.40
Aztec Manufacturing Co.	NYSE	AZZ	79.70	1.18	7.80
Cal-Maine Foods, Inc.	OTC	CALM	314.10	0.38	11.70
Chesapeake Utilities	NYSE	CPK	174.50	1.13	16.20
Dataram Corporation	AMEX	DTM	73.10	0.86	11.00
Energy South, Inc.	OTC	ENSI	74.00	1.67	13.20
Gencor Industries, Inc.	AMEX	GX	263.30	1.96	4.70
Industrial Scientific	OTC	ISCX	43.50	2.03	11.50
Keystone Consolidated	NYSE	KES	365.70	0.86	9.40
LandCare USA, Inc.	NYSE	GRW	111.40	0.33	29.40
Market Facts, Inc.	OTC	MFAC	126.70	0.98	26.50
Meridian Diagnostics, Inc.	OTC	KITS	36.30	0.46	14.70
Merit Medical Systems	OTC	MMSI	67.20	0.27	24.50
Met-Pro Corporation	NYSE	MPR	61.90	1.01	12.40
Nobility Homes, Inc.	OTC	NOBH	45.80	0.87	14.70
Omega Research, Inc.	OTC	OMGA	27.60	0.11	27.30
Point of Sale Limited	OTC	POSIF	12.30	0.28	25.40
Psychemedics Corp.	AMEX	PMD	17.60	0.13	39.40
Roadhouse Grill, Inc.	OTC	GRLL	118.40	0.26	20.80
Selas Corp. of America	AMEX	SLS	97.10	0.77	10.70
Toymax International, Inc.	OTC	TMAX	104.50	1.08	4.70
VSI Holdings, Inc.	AMEX	VIS	166.8	0.25	21
Warrantech Corporation	OTC	WTEC	207.30	0.13	29.80
Webco Industries, Inc.	AMEX	WEB	153.50	0.88	7.50

资料来源: American Association of Individual Investors web site, March 1999。

元素、变量和观测值

元素(elements)是指作为数据来源的原有实物(体)。如在表 1.1 的数据集中，每一公司的股票就是一个元素，包含这 25 只股票的数据集就包含了 25 个元素。**变量**(variable)是元素中所感兴趣的那些特征。表 1.1 中的数据集具有下列五个变量：

- **交易所**：股票交易的场所——NYSE（纽约股票交易所），AMEX(美国股票交易所)和 OTC（场外交易）。
- **股票代号**：用来在交易表中确认股票的缩写。
- **年发售额**：公司最近 12 个月内发售股票的总额，以百万美元为单位计量。
- **每股盈利**：最近 12 个月的每股盈利，按美元计。
- **市盈率**：每股的市场价格除以最近 12 个月的每股盈利。

数据是在研究中收集与每一元素相应的任一变量的度量值而获取的。从某一特定元素获取的度量值集被称为一组**观测值**(observation)。参见表 1.1，我们看到第一组观测值（Advanced Comm. Systems）的度量值集是 OTC，ACSC，75.10，0.32 和 39.10。第二组观测值（Ag-Chem Equipment Co.）的度量值集是 OTC，AGCH，321.10，0.48 和 23.40，等等。因为数据集包含 25 个元素，所以它就有 25 组观测值。

度量尺度

数据按下列度量尺度其中之一的方式收集：名义尺度、序数尺度、区间尺度和比例尺度。度量尺度决定了数据中蕴涵的信息量，标示出非常适合于应用的数据综述和统计分析。

当数据是用来确认元素属性的标签或名称时，相应变量的度量尺度被称做**名义尺度**(nominal scale)。例如，还见表 1.1 中的数据，我们看到交易所变量的度量尺度就是名义尺度，因为 NYSE、AMEX 和 OTC 是用于标明公司股票进行交易的场所的。在度量尺度是名义尺度的情形中，数字代码以及非数字型的标识符都可使用。例如，为便于数据收集并作好将数据录入计算机数据库的准备，我们可以使用数字代码：用 1 表示纽约股票交易所；用 2 表示美国股票交易所；用 3 表示场外交易。在这一场合中，数字型的 1、2 和 3 等值都是用作标识股票交易场所的。即使数据是数字型的，此时的度量尺度仍是名义尺度。

如果数据既具有名义数据的性能又需要按先后顺序或赋予一定的顺序来排列，这种情况下称一个变量的度量尺度为**序数尺度**(ordinal scale)。例如，Eastside Automotive 向顾客发放调查表，征询汽车维修服务质量方面的意见，每一位顾客反馈一个维修服务质量的等级：优秀、良好或较差。因为获取的数据——优秀、良好或较差是标识型的，具有名义变量的性能；另外，还可以按与服务质量相应的顺序或赋予一定的顺序排列记录下的数据，如优秀表示服务质量最好，接下来是良好，然后是较差。如此的度量尺度是序数尺度。注意，序数数据也可以用数字代码记载。例如，我们可用 1 表示优秀，2 表示良好，3 表示较差，仍能保持序数数据的性质。由此可见，一个

序数数据既可以是非数字型的也可以是数字型的。

如果数据具有序数数据的性能，并且是以特指的度量单位表示的观测值之间的一段区间，则这种变量的度量尺度称为**区间尺度**(interval scale)。区间数据总是数字型的。学习能力倾向测试（SAT）分数是区间尺度数据的一个例子。例如，三个学生的 SAT 分数分别为 1 120、1 050 和 970，能够按最好的成绩和最差的成绩排序。另外，分数之间的差异是有一定含义的，比如，学生 1 比学生 2 多 1 120 - 1 050 = 70 分，学生 2 比学生 3 多 1 050 - 970 = 80 分。

如果数据具有区间数据的所有性能，并且有必要用两个值的比例来表示，则该变量的度量尺度称为**比例尺度**(ratio scale)。比例数据总是数字型的，诸如距离、高度、重量和时间等变量都要用到比例尺度。这一尺度必须要有一个零值，变量取零值时表示什么也不存在。例如，让我们来看汽车的成本。零值意味着汽车没有成本或是免费的。如果我们拿 30 000 美元成本的一辆汽车与 15 000 美元成本的另一辆汽车作比较，比例值 30 000/15 000 = 2 表示第一辆汽车的成本是第二辆汽车成本的两倍，或相当于两辆第二辆汽车的成本。

定性和定量数据

数据能进一步划分为定性数据（质量型数据）或者定量数据（数量型数据）。**定性数据**(qualitative data)用来确认每一元素属性的标签或名称，定性数据既可用名义尺度也可以用序数尺度表示，既可以是非数字型的也可以是数字型的。**定量数据**(quantitative data)是数字型取值的，表示大小和多少，定量数据既可用区间尺度也可用比例尺度。

用于汇总数据的统计方法取决于数据是定性的还是定量的。

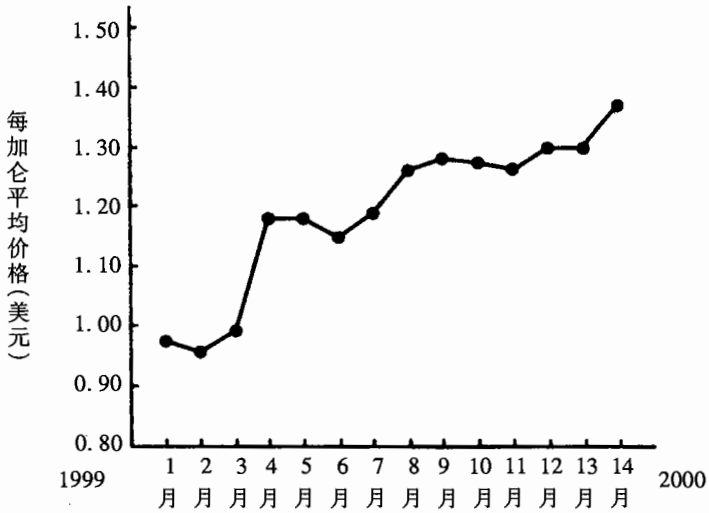
一个**定性变量**(qualitative variable)是用定性数据表示的变量，一个**定量变量**(quantitative variable)是用定量数据表示的变量。合适的统计方法的选择取决于变量是定性的或者是定量的。如果变量是定性的，统计分析是极其有限的。在每一统计类别中我们可以通过逐一列举观测值的方式概括出定性数据的特性或计算出定性数据在其中的比例。即使定性数据是以数字代码表示的，对其进行加、减、乘和除等算术运算也是没有意义的。2.1 节介绍汇总定性数据的方式。

另一方面，对定量数据进行算术运算，却经常能得出有意义的结果。例如，对一个定量变量，可先将数据求和然后再除以观测值的个数，从而求出平均值。这样的平均值有实际含义，并且很容易解释。一般情况下，当数据是定量的时候，有更多的统计方法可供选择。2.2 节和第 3 章介绍了概括定量数据的方法。

截面数据和时间序列数据

为了便于统计分析，有必要对截面数据和时间序列数据加以区分。**截面数据**(cross-sectional data)是在同一时点或近似同一时点上收集到的数据。表 1.1 中的数据是截面数据，因为它们这五个变量的取值，反映了在同一时刻 25 只影子股票的情况。**时间序列数据**(time series data)是在一

系列时间段采集的数据。例如，图 1.1 中的曲线标出了美国城市每加仑无铅标准汽油的平均价格，从 1999 年 1 月的每加仑 0.97 美元到 2000 年 2 月的每加仑 1.37 美元。本书所介绍的大部分统计方法可用于截面数据。



资料来源：U. S. Bureau of Labor Statistics。

图 1.1　美国城市每加仑无铅标准汽油的平均价格

评　注

1. 观测值是度量值的集合，相应地是为数据集中的每一元素赋值。因此，观测值的个数总是与元素的个数相同。每一元素的度量值的个数等于变量的个数。因此，数据项总数等于观测值个数乘以变量个数。
2. 定量数据可以是离散的或连续的，度量不可分事物的多少的定量数据就是连续的。

1.3　数据来源

数据可以从现有数据源中获取，或从观察和为得到新数据而设计的实验性研究中获取。

现存数据源

在某些场合，用于一些特殊需要的数据可能已存在于企业或组织中。所有的企业都要对它们的雇员、客户和经营状况等建立不同的数据库。关于雇员工资、年龄和工龄的数据通常可从内部个人记录中获取，关于销售、广告支出、配送成本、库存和产量的数据一般情况下都可以从其他

内部记录中获得。大部分企业也都保存有关于客户的详细数据。表 1.2 列举了一些数据，大部分企业通常都可从内部信息源获取。

表 1.2　可从企业内部获取的数据的例子

数据来源	一些可获得的典型数据
雇员记录	姓名、地址、身份证号码、工资、休假天数、病假天数和奖金
生产记录	部门或车间数量、产量、直接劳动成本和材料成本
库存记录	部门或仓库数量、实有库存量、重复订购率、经济订购率和折扣表
销售记录	生产部门、销售额、地区销售额和分类客户销售额
信誉记录	客户名称、地址、电话号码、信誉等级和财务收支状况
客户文件	年龄、性别、收入水平、家庭规模、地址和偏好

现在，实际中所需要的商务和经济数据可从专门收集和保存数据的机构那里获得，企业可通过租借或购买的方式从这些外部信息源获取数据。Dun & Bradstreet、Bloomberg 和 Dow Jones & Company 是三家能够为客户提供广泛的商务数据库服务的公司。ACNielsen 和信息资源公司也在成功地为广告商和制造商提供商务数据收集和加工服务。

数据也可从不同的行业协会和专门的营利机构获取。美国旅游行业协会保存了各周的诸如旅游者数量和旅游支出等与旅游有关的信息，旅游行业的公司和个人会对这类数据感兴趣。研究生管理委员会保存有关于考试成绩、学生特长和研究生培养计划等数据。从这些数据源公司获得的大部分数据所花费的成本都在用户的可接受范围之内。

近年来，互联网已成为数据来源的一个重要渠道。几乎所有的公司都有自己的互联网网站并提供公共访问端口。图 1.2 给出了部分公司的互联网网址，包括为每章开篇处提供统计实例的那些公司。进入这些网站，人们能很容易地获取许多关于产品和与这些公司有关的其他类型的信息。另外，大量的公司专门在互联网上提供可资利用的信息。结果使人们几乎能在网上获得像股票报价、餐馆饭菜的价格、工资数据等无限的、五花八门的信息。

政府代理机构是获取现存数据的另一重要渠道。举例来说，美国劳工部保存有关于就业率、工资水平、劳动力规模和联邦成员的数据信息。表 1.3 列举出部分政府机构和所能提供的几类数据。大部分政府机构也将数据收集和加工处理结果通过网站向社会提供。例如，美国人口普查局在网站 http: //www. census. gov 中有丰富的数据。图 1.3 显示出美国人口普查局的主页。

统计研究

有时一些特殊需要的数据并非总能从现存数据源得到。在这种场合，可通过统计研究的方式获得数据。统计研究可被划分为实验型和观测型两类。

在实验型统计研究中，首先要选准所感兴趣的主要变量。然后，找出一个或多个其他变量并加以控制，以便使获得的数据能反映出其他变量是如何影响主要变量的。例如，药厂可能会通过实验来掌握一种新药是如何影响血压的情况的。血压是该研究中最感兴趣的变量，新药的用量是另一个重要的变量，实验希望能找出其与血压之间的因果关系。为获得新药对血压影响效果的数

ACNielsen Corporation	**Delta Air Lines**	**Minnesota Mining and Manufacturing(3M)**
http://www.acnielsen.com	http://www.delta.com	http://www.mmm.com
Alcoa Inc.	**Disney**	**Morton International, Inc.**
http://www.alcoa.com	http://www.disney.com	http://www.morton.com
Amazon.com, Inc.	**Dollar General Corporation**	**Motorola, Inc.**
http://www.amazon.com	http://www.dollargeneral.com	http://www.motorola.com
American Airlines Inc.	**Dow Chemical Company**	**Novell, Inc.**
http://www.aa.com	http://www.dowchemical.com	http://www.novell.com
American Association of Individual Investors	**E. I. Dupont de Nemours and Company**	**Oracle Corporation**
http://www.aaii.com	http://www.dupont.com	http://www.oracle.com
American Express Company	**E*TRADE Securities, Inc.**	**Polaroid Corporation**
http://www.americanexpress.com	http://www.etrade.com	http://www.polaroid.com
Andersen Consulting	**Ford Motor Company**	**Procter & Gamble Co.**
http://www.ac.com	http://www.ford.com	http://www.pg.com
America Online, Inc.	**General Electric Company**	**The Prudential Insurance Company of America**
http://www.aol.com	http://www.ge.com	http://www.prudential.com
AT&T	**General Motors Corporation**	**Quicken.com**
http://www.att.com	http://www.gm.com	http://www.quicken.com
BMW of North America	**The Goodyear Tire & Rubber Company**	**SAP America, Inc.**
http://www.bmw.com	http://www.goodyear.com	http://www.sap.com
The Boeing Company	**Harris Corporation**	**Sears Roebuck and Company**
http://www.boeing.com	http://www.harris.com	http://www.sears.com
Business Week	**Hewlett-Packard Company**	**Siebel Systems, Inc.**
http://www.businessweek.com	http://www.hp.com	http://www.siebel.com
BUY.COM, Inc.	**The Home Depot, Inc.**	**Sprint Corporation**
http://www.buy.com	http://www.homedepot.com	http://www.sprint.com
Caterpillar, Inc.	**Honeywell**	**Toyota Motor Corporation**
http://www.caterpillar.com	http://www.honeywell.com	http://www.toyota.com
Charles Schwab & co.	**IBM Corporation**	**The Travelers Insurance Company**
http://www.charlesschwab.com	http://www.ibm.com	http://www.travelers.com
Chemdex Corporation	**International Paper**	**Union Carbide**
http://www.chemdex.com	http://www.ipaper.com	http://www.unioncarbide.com
Chevron Corporation	**Johnson & Johnson**	**United Way of America**
http://www.chevron.com	http://www.johnsonjohnson.com	http://www.unitedway.org
Cisco Systems, Inc.	**L. L. Bean, Inc.**	**US Airways, Inc.**
http://www.cisco.com	http://www.llbean.com	http://www.usair.com
Citibank	**Lucent Technologies**	**Visa International**
http://www.citibank.com	http://www.lucent.com	http://www.visa.com
The Coca Cola Company	**McDonald's Corporation**	**Wal-Mart Stores, Inc.**
http://www.cocacola.com	http://www.mcdonalds.com	http://www.walmart.com
Colgate-Palmolive Company	**Mead Corporation**	**Xerox Corporation**
http://www.colgate.com	http://www.mead.com	http://www.xerox.com
Dell Computer Corporation	**Microsoft Corporation**	**The Yankee Group**
http://www.dell.com	http://www.microsoft.com	http://www.yankeegroup.com

图 1.2　部分公司的互联网网址

表 1.3　部分政府机构和所能够提供的几类数据

政府机构	一些可供利用的数据
人口普查局 http://www.census.gov	人口数据和分布状况、家庭数量及其分布、家庭收入及其分布
联邦储备局 http://www.bog.frb.fed.us	关于货币供应、信誉记录、汇率和贴现率的数据
预算编制办公室 http://www.whitehouse.gov/omb	关于财政收入、支出和联邦政府发行债券的数据
商业部 http://www.doc.gov	关于商业活动、工业出口额、工业利润水平和产业的增长与衰退等方面的数据

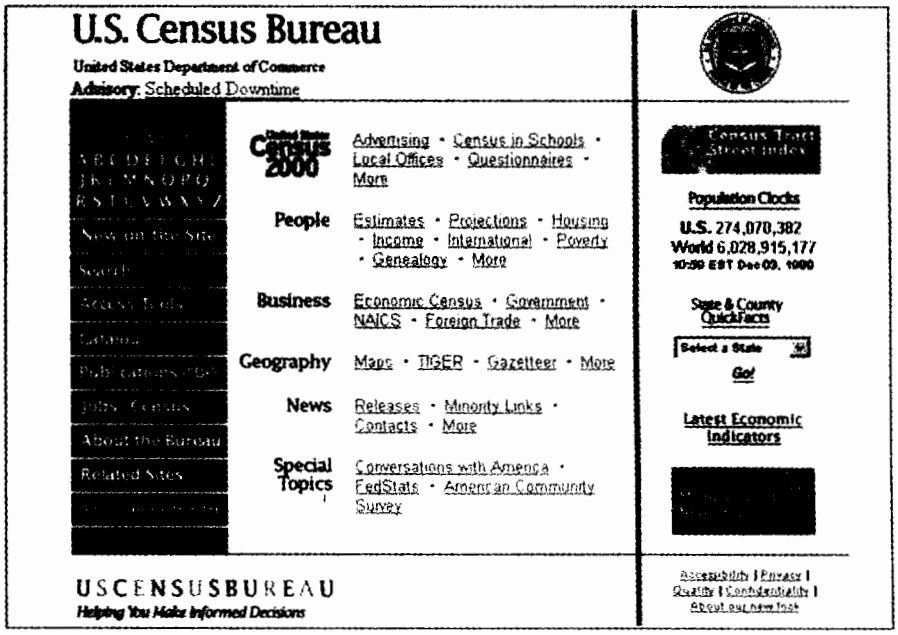


图 1.3　美国人口普查局的主页

据，可选择一些个体样本进行实验。控制新药的用量，对不同的组别给予不同的药量，然后收集每一组受试对象的血压情况的数据。实验数据的统计研究有助于弄清楚新药是如何影响血压的。

曾进行的最大规模的实验型统计研究应属 1954 年注射沙克疫苗的公共健康服务，该研究从全美国小学 1—3 年级的学生中选出了近 200 万个儿童作为注射对象。

非实验型或称为观测型统计研究并不是有目的地去控制所感兴趣的变量。观察也许是观测型研究中最常用的典型方法。举例来说，在对一次征求顾客意见活动的观察中，首先要明确需要研究的问题，然后设计调查问卷、选择一些调查对象作为样本。一些餐馆运用观测型研究获取顾客关于餐馆的饭菜质量、服务和就餐环境等方面的数据。由美国佛罗里达州 Redington 的 Lobster Pot 餐馆设计的调查问卷见图 1.4。注意，请顾客根据自己的感受对饭菜质量、服务态度、服务措施、清洁程度和管理水平这五个变量按优秀、良好、满意和不满意的等级顺序打分，由此获得的数据能帮助 Lobster Pot 的经理估计餐馆的经营状况。

对抽烟者和不抽烟者的研究是观测型研究，因为研究者不能确知和控制将来哪些人会抽烟，哪些人不抽烟。

The LOBSTER Pot RESTAURANT

很荣幸您能光临 Lobster Pot 餐馆，相信您还会再次光临。如果能占用您一点点时间，填写这张卡片，我们将对此表示由衷的赞赏。您的评价和建议对我们非常重要，谢谢！

顾客姓名_____

	优秀	良好	满意	不满意
饭菜质量	□	□	□	□
服务热情	□	□	□	□
服务速度	□	□	□	□
卫生状况	□	□	□	□
管理水平	□	□	□	□

总的评价_____

改进建议_____

请在入口处的意见箱留下意见，谢谢。

图 1.4　佛罗里达州 Redington 的 Lobster Pot 餐馆的顾客调查问卷（惠允使用）

将数据和统计分析结果用于管理决策，经理们必须了解获取数据所需要的时间和成本。当在相对较短的一段时期内需要获得数据时，利用现存数据源是一种理想的方法。如果一些重要的数据不便从现有数据中找到，就必须考虑到获取数据所要占用的时间和花费的成本。在各种情况中，决策者都会意识到统计分析对决策过程的积极影响。数据收集和进行统计分析所付出的成本是不会超过利用这些信息形成更好的决策所节省的成本的。

数据采集误差

经理们始终需要牢记在统计研究中数据是可能产生误差的。利用错误的数据也许会比根本不利用数据更糟糕。无论什么时候，获取的数据的值只要不等于实际值或不同于按正确的采集过程获得的数据的值，就会产生数据采集误差。许多数据采集方法都会造成这类误差。例如，访谈员可能会出现记录错误，像将 24 岁的年龄误记为 42 岁，或在向访谈对象提问题时可能会误导或作出不正确的反应。

实验数据分析更关注数据的收集和记录，以确保不出现错误。可采取一些特殊措施检查数据的内在一致性。举例来看，这样一种情况就表明要重新检查一下相应的数据记录：一个 22 岁的人却说有 20 年的工作经历。数据分析中一些过大或过小的数值，被称为异常值，也促使要复查数

据，圈出那些可能有错误的数据。第 3 章中介绍统计学家圈定异常值的一些方法。

数据采集过程中经常会发生误差。盲目地利用一切可以得到的或不认真地收集到的数据，将会产生误导和导致作出错误决策。因此，按照正确的步骤采集数据能确保决策信息的可信度，提高数据的利用价值。

1.4　描述性统计

绝大部分报纸、杂志、公司报告和其他大众化的统计信息都是以易于读者理解的方式披露的。将数据以图形、表格和数字的形式表示，这类统计方法被称为**描述性统计**(descriptive statistics)。

再来参见表 1.1 中列举的 25 只影子股票情况的数据集，可用描述性统计方法对这一数据集中的信息给予综述。例如，表 1.4 给出了定性变量“交易所”的数据的表格式汇总。同样，数据的图形式汇总在图 1.5 中显示，被称为柱状图。这些表格和图形式汇总的目的是为了使数据更容易解释。参见表 1.4 和图 1.5，我们能很容易地看出在数据集中，大部分股票是通过直接交易完成的。用百分比频数的形式表示，56% 是场外交易，24% 的交易在美国证券交易所完成，20% 的交易在纽约证券交易所完成。

表 1.4　交易所变量的频数和百分比频数

交　易　所	频　数	百分比频数
纽约证券交易所（NYSE）	5	20
美国证券交易所（AMEX）	6	24
场外交易（OTC）	14	56
总计	25	100

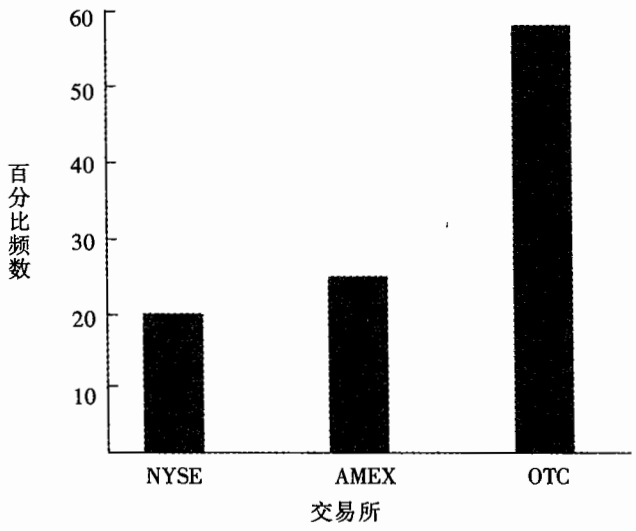


图 1.5　交易所变量的柱状图

这些股票的定量变量市盈率数据的图形式汇总由图 1.6 给出，被称为直方图。从这个直方图很容易看出，市盈率的值域范围从 0.0 到 39.9，在 8.0—15.9 之间达到最高。

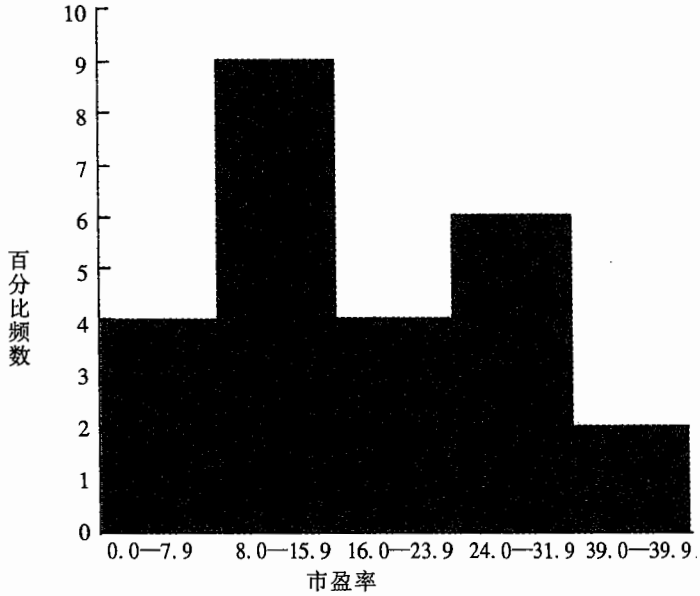


图 1.6　25 只股票的直方图

除表格和图形表示方式以外，数字型的描述性统计也用于汇总数据。最常用的数字型描述统计是平均值或均值。利用表 1.1 中 25 只股票年销售额的数据，我们可以将所有 25 只股票的年销售额相加再除以 25，算出年平均销售额。如此可得出年平均销售额为 1.255 4 亿美元。这一平均值被作为数据的中心趋势或中点定位的度量。

近几年，人们关注描述性统计的拓展和表述，对这类统计方法的兴趣在不断提高。第 2 章和第 3 章集中讨论描述性统计的表格式、图形式和数字式方法。

1.5　统计推断

在许多情形中，需要对其中的数据进行处理的元素群体（个人、公司、选民、家庭、产品、顾客，等等）非常之大。由于时间、成本和其他需要考虑的因素，只能从元素群体中收集一小部分数据。在特定研究中，元素的大群体被称为**总体**(population)，较小的一部分称为**样本**(sample)。我们给出下面的正式定义：

总体

作为研究对象的全部元素所组成的集合。

样本

总体的一个子集。

统计学的一个主要贡献就是能利用样本数据来估计总体的特征并进行假设检验。这一过程被称为**统计推断**(statistical inference)。

让我们来考察由 Norris 电器公司所做的一项研究，以此作为统计推断的一个例子。Norris 电器公司生产一种能在许多电器产品上使用的高密度灯管。为提高灯管的使用寿命，产品设计部门推出一种新灯丝灯管。这种情况下，能够生产出来的所有新灯丝灯管被定义为总体。为估计新灯丝对灯管性能的改进，生产出 200 只新灯丝灯管进行检测。收集每只灯管在灯丝被烧断之前的使用小时数的数据，由此样本所得到的数据记录见表 1.5。

表 1.5　Norris 电器公司作为样本的 200 只灯管的使用寿命(小时数)


CD 光盘数据
Norris

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73

假定 Norris 电器公司所感兴趣的是由样本作出统计推断，得知所有能用新灯丝制成的灯管的平均使用寿命。将表 1.5 中的 200 个值相加然后再除以 200，计算出灯管的平均使用寿命：76 小时。我们就能利用这一样本结果来估计灯管总体的平均使用寿命是 76 小时。图 1.7 是概括 Norris 电器公司统计推断过程的示意图。

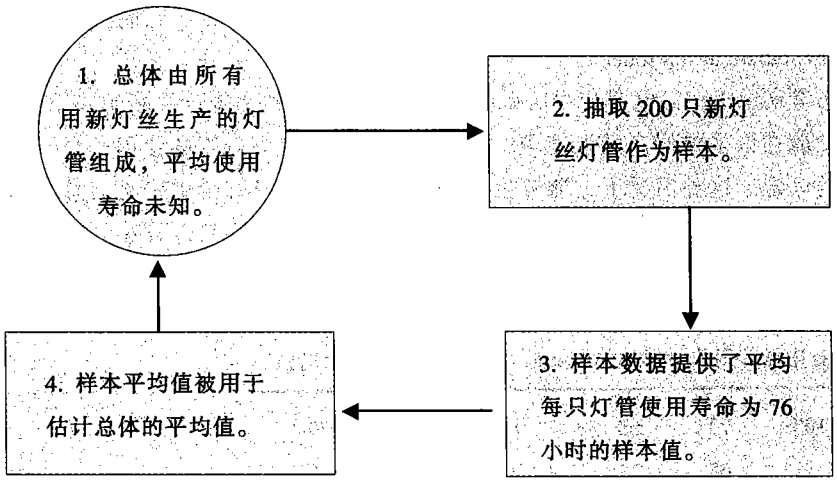


图 1.7 Norris 电器公司的统计推断过程

无论什么时候统计学家利用样本来估计总体的某些特征，他们通常都要相应地注明统计质量或精确度。以 Norris 电器公司为例，统计学家就会指出，新灯管总体的平均使用寿命 76 小时具有 4 小时的误差范围，那么，所有新灯丝灯管的平均使用寿命的区间估计就是 72 小时—80 小时。统计学家也可能会说明总体平均值是如何包含在 72 小时—80 小时之间的。

小 结

统计是收集、分析、提炼和解释数据的科学与艺术。几乎每一所院校商贸和经济学专业的大学生都要学习统计学课程。我们以介绍商务和经济中典型的统计应用案例的方式展开本章的论述。

数据是那些作为收集、分析、陈述和解释之用的事实依据和图表。四种度量尺度可用于获取关于特指变量的数据：名义尺度、序数尺度、区间尺度和比例尺度。当数据是用来确认元素属性的标签或名称时，相应变量的度量尺度被称为名义尺度。如果数据具有名义数据的性能，又有明确意义的先后顺序，该度量尺度称为序数尺度。如果数据具有序数数据的性能，并且是以特指的度量单位表示的观测值之间的一段区间，则称这种度量尺度为区间尺度。最后，如果数据具有区间数据的所有性能，并且有必要用两个值的比例来表示，则称该度量尺度为比例尺度。

为了统计分析之目的，数据可划分为定性数据和定量数据。定性数据用来确认每一元素属性的标签或名称，可以是非数字型的或数字型的。定量数据是数字型的，表示大小和多少。定量数据或者是用区间尺度或者是用比例尺度度量。只有当数据是定量的时候，通常的算术运算才有意义。因此，适合于定量数据的统计计算并非总是适合于定性数据。

在 1.4 和 1.5 节中，我们介绍了描述性统计和统计推断。描述性统计是以表格、图形和数字方式汇总数据的一种统计方法。统计推断是利用样本数据来估计总体的特征并进行假设检验的过程。

术语释义

统计： 收集、分析、提炼和解释数据的科学与艺术。

数据： 被收集、分析和解释的事实依据与图表。

数据集： 在特定研究中所有收集到的数据。

元素： 作为数据来源的原有实物(体)。

变量： 元素中所感兴趣的那些特征。

观测值： 对某一元素的一系列的度量。

名义尺度： 用标签或名称数据来确认某一元素属性的一种度量尺度。名义尺度可以是数字型的或者非数字型的。

序数尺度： 具有名义数据的性能、用来为数据排序的一种度量尺度。序数数据可以是数字型的或者非数字型的。

区间尺度： 具有序数数据的性能，并且是以特指的度量单位表示的观测值之间的区段变量的一种度量尺度。区间数据总是数字型的。

比例尺度： 具有区间数据的所有性能，并且有必要用两个值的比例来表示变量的一种度量尺度。比例数据总是数字型的。

定性数据： 用来确认某一元素属性的标签或名称的数据。定性数据以名义或序数尺度的形式表示，可以是数据型的或者非数据型的。

定量数据： 用来表示事物量的多少的数据。定量数据以区间或比例尺度的形式表示，总是数字型的。

定性变量： 用定性数据表示的变量。

定量变量： 用定量数据表示的变量。

截面数据： 在同一时点或近似同一时点上收集到的数据。

时间序列数据： 在一系列时间段采集的数据。

描述性统计： 将数据以图形、表格和数字形式表示的统计方法。

总体： 作为研究对象的全部元素所组成的集合。

样本： 总体的一个子集。

统计推断： 利用样本数据对总体进行估计或假设检验的过程。

练　习

1. 讨论作为数字整理的统计和作为一门学科或研究领域的统计学之间的区别。

2. *Condé Nast Traveler* 对顾客做了一次年度性的民意测验，以确定全世界什么地方最好玩。表

自测题 1.6是一个从最近的一次民意测验中抽出的9个欧式旅馆的样本。在旅游旺季旅馆的标准间的价格按从最低$到最高$$$$排列。总分包括顾客对每一个旅馆的房间、服务、餐饮、位置/氛围和公共面积等方面的评价,较高的分数表示较高的满意度。

a. 这一数据集中有多少个元素?

b. 这一数据集中有多少个变量?

c. 哪些变量是定性变量,哪些变量是定量变量?

d. 用于每一变量的度量尺度是什么类型的?

3. 参见表1.6。

a. 9个旅馆的平均房间数是多少?

b. 计算总分的平均值。

c. 位于英国的旅馆所占的百分比是多少?

d. 具有$$级房间的旅馆的百分比是多少?

表1.6 欧洲住宿地的等级

CD光盘数据 Hotel

旅馆名称	国家	客房等级	客房数量	总分
Graveteye Manor	英国	$$	18	83.6
Villa d'Este	意大利	$$$$	166	86.3
Hotel Prem	德国	$	54	77.8
Hotel d'Europe	法国	$$	47	76.8
Palace Luzern	瑞士	$$	326	80.9
Royal Crescent Hotel	英国	$$$	45	73.7
Hotel Sacher	澳大利亚	$$$	120	85.5
Duc de Bourgogne	比利时	$	10	76.9
Villa Gallici	法国	$$	22	90.6

资料来源: *Condé Nast Traveler*, January 2000。

4. 自测题 《财富》杂志发布全美500家最大的工业公司按年收入和利润排序情况的数据。这500家公司的样本数据由表1.7给出。

a. 这一数据集中有多少个元素?

b. 什么是总体?

c. 计算样本的平均收入。

d. 利用(c)所得的结果,估计出总体的平均收入是多少?

5. 考察表1.7中《财富》500强企业样本的数据集。

a. 这一数据集中有多少个变量?

b. 哪些变量是定性变量,哪些变量是定量变量?

c. 计算样本的平均利润。

d. 利润超过1亿美元的企业所占的百分比是多少?

e. 具有工业代码 3 的公司占多少百分比?

表 1.7　《财富》500 强企业中的一个 10 家企业的样本


CD 光盘数据
Fortune

公司名称	公司收入（百万美元）	利润（百万美元）	工业代码
US Airways Group	8 688.0	538.0	3
International Paper	19 500.0	213.0	23
Tyson Foods	7 414.1	25.1	20
Hewlett-Packard	47 061.0	2 945.0	13
Intel	26 273.0	6 068.0	49
Northrup Grumman	8 902.0	214.0	2
Seagate Technology	6 819.0	-530.0	11
Unisys	7 208.4	387.0	10
Westvaco	2 904.7	132.0	23
Campbell Soup	7 505.0	660.0	20

资料来源：*Fortune*, April 26, 1990。

6. 哥伦比亚出版社向它的邮购客户俱乐部成员提供 CD 光盘和磁带。哥伦比亚出版社的一份音乐调查表向俱乐部的新成员询问 11 个问题，其中所要问的一些问题是：
   a. 在最近 12 个月内您购买了多少 CD 光盘和磁带?
   b. 目前您是国家图书邮购俱乐部的成员吗（是或不是）?
   c. 您的年龄?
   d. 包括您自己在内，您家共有几口人(成人和小孩）?
   e. 您有兴趣购买哪种音乐?（列出 15 类音乐，包括疯狂摇滚、柔情摇滚、成人流行音乐、重金属乐、打击乐和田园音乐等）
   区分出每一问题提供的是定性数据还是定量数据。

7. 加利福尼亚州的一家中介机构将工作岗位划分为专家、白领和蓝领岗位。以 1 定义专家、2 定义白领和 3 定义蓝领的方式记录数据。
   a. 变量是工作岗位，它是定性变量还是定量变量?
   b. 这一变量所用的度量尺度是什么类型的?

8. 《华尔街日报》/NBC 向 2 013 名成人做了一次新闻调查："您对当前美国经济形势的满意程度如何？"(*The Wall Street Journal*, December 12, 1997)回答方式是从不满意、满意和不能肯定三者中选择其一。
   a. 这次调查的样本规模是多大?
   b. 数据是定性的还是定量的?
   c. 对这一问题适合用平均值或百分比进行数据汇总,哪种方法更好?
   d. 在收到的反馈表中，28% 的人对美国经济不满意。有多少人选择了这一项?

9. 美国商业部报告，收到申请美国 Malcolm Baldrige 国家质量奖的情况如下：23 家大的制造企业，18 家大的服务性企业和 30 家小的商行。

a. 行业类型是定性变量还是定量变量？

b. 小商行的申请者占多大百分比？

10. 指出下列每一变量是定性的还是定量的，并说明每一变量最适合的度量尺度。

a. 年龄

b. 性别

c. 社会地位

d. 汽车制造厂家

e. 倾向于设置死刑的人的数量

11. 指出下列每一变量是定性的还是定量的，并说明每一变量所用的度量尺度。

a. 年销售额

b. 软饮料规模(小、中、大)

c. 雇员级别(从 1 级到 18 级)

d. 每股盈利

e. 支付方式(现金、支票和信用卡)

12. 夏威夷旅游局收集了去夏威夷的旅游者的数据。下列问题是 6 月份在飞往夏威夷的航班上向乘客分发的问卷所征询的 16 个问题中的一部分。

- 这次夏威夷之旅是我的：第一次、第二次、第三次、第四次，等等
- 这次旅行最主要的理由是：（包括度假、惯例、蜜月等 10 项选择）
- 预期住所：（包括宾馆、公寓、亲戚、露营等 11 项选择）
- 在夏威夷逗留的总天数

a. 所要研究的总体是什么？

b. 进行问卷调查是研究飞往夏威夷航班上的乘客这一总体的好方法吗？

c. 将上述 4 个问题中的每一个以定性数据或定量数据的方式加以区分。

13. 图 1.8 是反映 1991—1997 年间河船娱乐场数量的柱状图。



a. 所给数据是定性的还是定量的？

b. 所给数据是截面数据还是时间序列数据？

c. 所感兴趣的变量是什么？

d. 分析发展趋势，您期望看到 1998 年是增长还是减弱？

14. 下列数据集提供了美国科技公司财务业绩的概况(*Barron's*, December 29, 1997)：

	1993 年	1994 年	1995 年	1996 年
每股盈利（美元）	2.78	2.13	3.41	3.83
总收入（10 亿美元）	11.87	12.57	13.43	14.92
净收入（10 亿美元）	1.51	1.17	1.89	2.12
每股售价（美元）	14.35	10.98	12.67	13.98

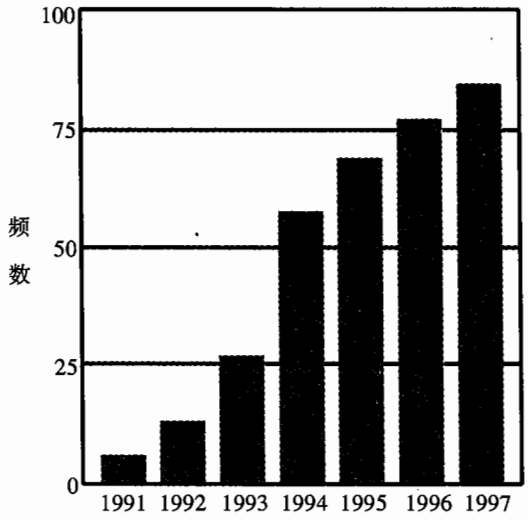


图 1.8　美国的河船娱乐场

资料来源：《华尔街日报》获准重新排印，1997 年道琼斯公司授权可对全世界公开。

a. 有多少变量？

b. 所给数据是定性的还是定量的？

c. 它们是截面数据还是时间序列数据，为什么？

15. 参见表 1.7 中《财富》500 强企业的数据。它们是截面数据还是时间序列数据，为什么？

16. 您所在公司的营销部门提出了一种新型软饮料的销售计划，它声称要占领年轻人市场的大部分份额。

a. 在作出将新产品投放市场、实际进行投资的决策之前，您需要看到什么数据？

b. 您认为上一问中的数据应如何获得？

17. 一家大公司的老总为有才干的下属加薪 10 000 美元，以防止他们跳槽到别的公司。需要哪些内部和外部数据源来确定这样的增加工资的标准是否合适？

18. 在最近的一次对 60 岁及以上年龄段的男性死亡原因的研究中，一个 120 人的样本中有 48 人死于心脏疾病。

a. 将此例扩充成一个描述性统计，以便用来估计 60 岁及以上年龄段的男性死于某种心脏疾病的百分比。

b. 死亡原因的数据是定性的还是定量的？

c. 讨论统计推断在这类医疗保健研究中的作用。

19. 《商业周刊》北美读者研究部 1996 年收集了 2 861 个读者的样本数据。根据反馈情况来看，有 59% 的读者的年收入是 75 000 美元，50% 以上的读者持有美国运通卡。

a. 该项研究中总体指的是什么？

b. 年收入是定性变量还是定量变量？

c. 美国运通卡的持有者是定性变量还是定量变量?

d. 这一研究中用的是截面数据还是时间序列数据?

e. 说出《商业周刊》基于这一调查可能作出的所有统计推断。

20. 对 50 家公司使用浏览数据情况的调查发现了下列现象 (Mercer Management Consulting, Inc., April 24, 1997):

- ACNielson 占有 56% 的市场份额。
- 对每一类顾客商品目录数据的年平均使用费是 387 325 美元。
- 按 1(非常不满意)到 5(非常满意)的等级标准,所有浏览数据用户的平均满意水平是 3.73。

a. 引用两种描述性统计方法表示此例。

b. 对浏览数据所有用户这一总体的总的满意水平作出推断。

c. 对浏览商品目录数据的平均支出水平作出大致的推断。

21. 一项为时 7 年的医药研究报告指出,母亲在怀孕期间服用过 DES 药物的那些妇女,她们的组织变异导致癌变的可能性要比母亲没有服用过此药的妇女高一倍。

a. 该项研究包含了哪两个总体?

b. 你认为这些数据是调查获得的还是实验获得的?

c. 对母亲在怀孕期间服用过 DES 的这一妇女总体中做了 3 980 位妇女的样本抽查,其中发现 63 例,她们的组织变异可能导致癌变。给出一个描述性统计,可用来估计在这一总体中 1 000 位妇女中患有组织变异的人数。

d. 对母亲在怀孕期间未服用过 DES 的这一妇女总体,1 000 人中可能患有组织变异的期望值是多少?

e. 医疗研究中经常使用相对较大的样本(此例中是 3 980),为什么?

22. 一企业有兴趣检验在一个新的商业电视活动中做广告的效果,因此在科罗拉多州府丹佛市下午 6:30 的当地新闻节目中播出此次活动。两天后,市场研究企业进行了电话询问,获取重复收看率的信息(观众重复收看该节目的百分比)和对这次活动的印象。

a. 这项研究中的总体是什么?

b. 这项研究中的样本是什么?

c. 为什么在这种情况中要用样本抽查,请解释。

23. ACNielsen 机构对全美国进行每周一次的电视台节目收看调查,ACNielsen 的统计数字显示每一主要电视节目的观众数目,对电视节目和观众收视率进行排序,每周发布电视台排行榜。

a. ACNielsen 机构想了解什么?

b. 总体是什么?

c. 为什么在这种情况中要用样本抽查?

d. 哪种决策和行动需要以 ACNielsen 的研究为基础?

24. 5 名学生的期中考试成绩的样本值如下:72, 65, 82, 90, 76。下列表述中哪一个是正确的,哪一个看起来太空泛?

a. 5 名学生这一样本的期中平均成绩是 77。
b. 参加这次考试的所有学生的平均成绩是 77。
c. 参加这次考试的所有学生的平均成绩的估计值是 77。
d. 一大半参加这次考试的学生的成绩是在 70—85 之间。
e. 如果这一样本还包括其他 5 名学生，他们的分数将在 65—90 之间。

2 描述性统计:表格法和图形法

统计实例:高露洁—棕榄公司

2.1 汇总定性数据

频数分布

相对频数和百分比频数分布

柱状图和饼形图

2.2 汇总定量数据

频数分布

相对频数和百分比频数分布

点位图

直方图

累积分布

累积曲线

2.3 探索性数据分析:茎叶显示

2.4 交叉分组列表和散点图

交叉分组列表

散点图

统计实例

高露洁—棕榄公司*

纽约市，纽约州

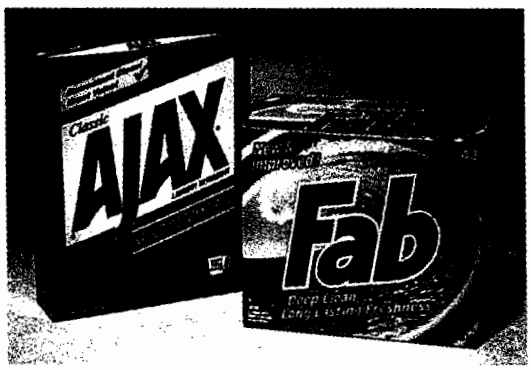
高露洁—棕榄公司(Colgate-Palmolive Company)于1806年在纽约城开张的时候，还只是一个小小的肥皂和蜡烛商店。但今天，它已经是一家产品遍布世界200多个国家和地区、年销售额90亿美元的大公司。该公司最著名的商标是高露洁、棕榄、Ajax和Fab，同时也经营希尔斯科学食品(Hill's Science Diet)和希尔斯传统食品(Hill's Prescription Diet)等产品。

高露洁—棕榄公司在其家用洗衣粉产品的生产过程质量保证程序中使用了统计学方法。质量保证中的一个关键问题是顾客对于盒子中洗衣粉数量的满意度。同一尺寸的每一个盒子都按重量装入同样数量的洗衣粉，但洗衣粉所占的体积受洗衣粉密度的影响。例如，如果洗衣粉密度较大的话，要达到所规定的包装重量就只需较小体积的洗衣粉。这样，当顾客打开包装盒时，盒子看上去显然未满。

为了控制洗衣粉密度过大的问题，需要对洗衣粉密度的可接受范围进行限定。首先由质检人员定期地抽取统计样本，并测量每个洗衣粉样本的密度。然后将数据汇总提供给操作人员，以便在需要将密度保持在规定的质量规格内时，操作人员可以采取正确的措施。

每周抽取一次得到的150个样本的密度值频数分布和直方图显示在所附的表和图中。超过0.40的密度水平即为不可接受的上限。频数分布和直方图显示出工艺操作符合质量标准，所有产品的密度都小于或等于0.40。检查统计汇总结果的经理将对洗衣粉产品的质量感到高兴。

在本章中，你将学习诸如频数分布、柱状图、直方图、茎叶显示、交叉分组列表以及其他一些属于描述性统计的表格和图形方法。这些方法的目的是汇总数据，以使得数据更易于人们理解。

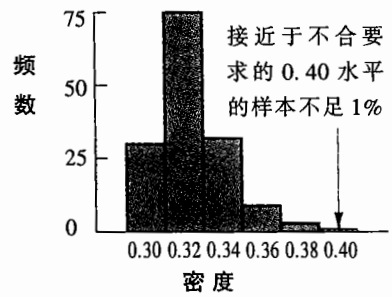


统计汇总有助于保持这些高露洁—棕榄产品的质量。© Joe Higgins/ South-Western.

密度数据的频数分布

密度	频数
0.29—0.30	30
0.31—0.32	75
0.33—0.34	32
0.35—0.36	9
0.37—0.38	3
0.39—0.40	1
总计	150

密度数据的直方图



*作者感谢高露洁—棕榄公司的质量保证经理William R. Fowle提供了该案例。

正如第 1 章所述，我们可以将数据分为定性数据和定量数据。**定性数据**(qualitative data)是一种标记或名称，用来标识诸如项目之类事物的类别；而**定量数据**(quantitative data)则是表示多少的数值。

本章的目的是介绍通常用来汇总定性和定量数据的表格和图形方法。人们能在年度报告、报纸文章以及研究报告中发现数据表格和图形汇总，每一个人都会面对这些类型的表述。因此，理解它们是怎样形成的以及应如何解释它们是至关重要的。我们从有关单变量汇总数据的图表方法开始，在最后一部分还将介绍当牵涉到两变量间的关系时，用来汇总数据的方法。

现代统计软件包拥有汇总数据和构建图形表述的强大功能，Minitab 和 Microsoft Excel 是其中应用最广泛的两个，在本章的附录中，我们将演示一些它们的具体功能。

2.1　汇总定性数据

频数分布

我们从定义**频数分布**(frequency distribution)开始，讨论如何使用图表方法来汇总定性数据这一问题。

> **频数分布**
>
> 频数分布是一种数据的表格汇总，表示在几个互不重叠组别中每一组包含的项目个数（即频数）。

我们引用下面的例子，来说明如何构建和解释定性数据的频数分布。根据 *Beverage Digest* 的资料，Coke Classic、Diet Coke、Dr. Pepper、Pepsi-Cola 和 Sprite 是软饮料销售量排行的前 5 名（*The Wall Street Journal Almanac*, 1998）。假设表 2.1 中的数据表示在只选择这 5 种软饮料的情况下，抽取的 50 次软饮料购买样本。

表 2.1　来自 50 次软饮料购买样本的数据



CD 光盘数据
Soft Drink

Coke Classic	Sprite	Pepsi-Cola
Diet Coke	Coke Classic	Coke Classic
Pepsi-Cola	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi-Cola
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi-Cola	Diet Coke
Pepsi-Cola	Coke Classic	Pepsi-Cola

（续表）

Pepsi-Cola	Coke Classic	Pepsi-Cola
Coke Classic	Coke Classic	Pepsi-Cola
Dr. Pepper	Pepsi-Cola	Pepsi-Cola
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi-Cola
Coke Classic	Pepsi-Cola	Sprite
Coke Classic	Diet Coke	

表 2.2 软饮料购买次数的频数分布

软饮料	频数
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5
总计	50

为把这些数据构成一个频数分布，我们将表 2.1 中每一种软饮料出现的次数累加起来。Coke Classic 共出现 19 次，Diet Coke 8 次，Dr. Pepper 出现 5 次，Pepsi-Cola 出现 13 次，而 Sprite 出现 5 次。这些加总值被概括在表 2.2 的频数分布中。

这个频数分布汇总说明了这 50 次软饮料购买是怎样分配给以上 5 种软饮料的，它提供了比表 2.1 中的原始数据更多的内涵。观察一下该频数分布，我们看到 Coke Classic 排在首位，Pepsi-Cola 随后，Diet Coke 位居第三，Dr. Pepper 和 Sprite 并列第四。频数分布还揭示了这 5 种销量最高的软饮料相对受欢迎程度的信息。

相对频数和百分比频数分布

频数分布表示在几个互不重叠组别中每一组所包含的项目个数(频数)。可是，我们往往对每一组的项目个数所占的比例或者是百分比更感兴趣。一组的相对频数分布就是一组的项目个数所占的比例。对于一个具有 n 个观察值的数据集，可计算每组的相对频数如下：

相对频数

$$\text{每一组的相对频数} = \frac{\text{组内项目的频数}}{n} \tag{2.1}$$

每一组的**百分比频数**是相对频数乘以 100。

相对频数分布(relative frequency distribution)是表示每一组的相对频数数据的表格汇总。**百分比频数分布**(percent frequency distribution)是表示每一组的百分比频数数据的表格汇总。表 2.3 列示了软饮料数据的相对频数分布和百分比频数分布。在表 2.3 中，我们看到 Coke Classic 的相对频数是 19/50 = 0.38，Diet Coke 的相对频数是 8/50 = 0.16，其他依此类推。从百分比频数分布中，我们看到购买的软饮料有 38% 是 Coke Classic，有 16% 是 Diet Coke，等等。我们还注意到前三名的软饮料占据了 38% + 26% + 16% = 80% 的份额。

表 2.3　软饮料购买的相对频数和百分比频数分布

软饮料	相对频数	百分比频数
Coke Classic	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi-Cola	0.26	26
Sprite	0.10	10
总计	1.00	100

柱状图和饼形图

柱状图(bar graph)是一种图形方法，用于描述已被汇总为频数分布、相对频数分布或者百分比频数分布的数据。在图的一个轴（通常是横轴）上，我们规定了用来对数据分组(类)的标记。在图的另一坐标轴（通常是纵轴）上标有频数、相对频数或者百分比频数的刻度。然后，利用一个位于每一组标记之上的固定宽度的柱，将这个柱的长度延伸，使之达到该组的频数、相对频数、百分比频数的数值。对于定性数据，应将这些柱互相隔开，以强调每组都是互相独立的这一事实。图 2.1 就是一个反映了 50 次软饮料购买情况的频数分布的柱状图。注意图形是怎样显示出 Coke Classic、Pepsi-Cola 和 Diet Coke 三种哪种是最受欢迎的。

在质量控制应用中，柱状图用于识别问题的最重要原因。当这些柱按照高度降低的顺序从左向右依次排列时，最经常出现的原因将排在第一位，这种柱状图被称为帕累托(Pareto)图。它是以它的创立者意大利经济学家维尔福来多·帕累托 (Vilfredo Pareto)的名字命名的。

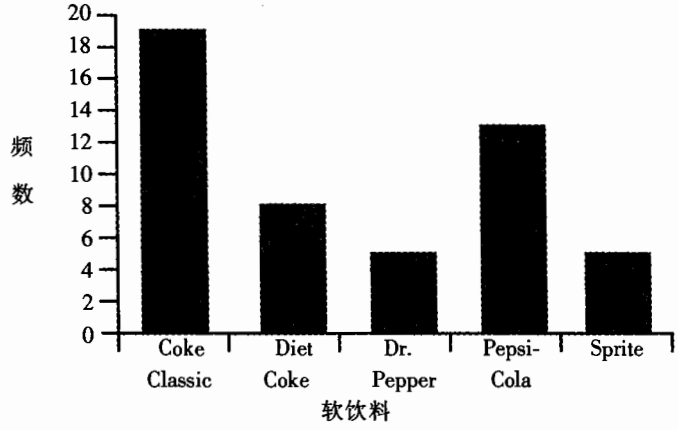


图 2.1　软饮料购买情况的柱状图

饼形图(pie chart)是另一种表示相对频数和百分比频数分布的图形方法。为构建饼形图，我们首先画一个圆来代表所有的数据，然后再利用相对频数把一个圆分割为几部分，使每部分对应于

每一组的相对频数。例如，因为一个圆有 360°，而 Coke Classic 的相对频数为 0.38，于是饼形图中代表 Coke Classic 的部分占有 $0.38 \times 360° = 136.8°$，而代表 Diet Coke 的部分占有 $0.16 \times 360° = 57.6°$。对其他组别进行同样计算，就产生了图 2.2 中的饼形图。饼形图中每一部分所显示的数值既可以是频数，也可以是相对频数，或者百分比频数。

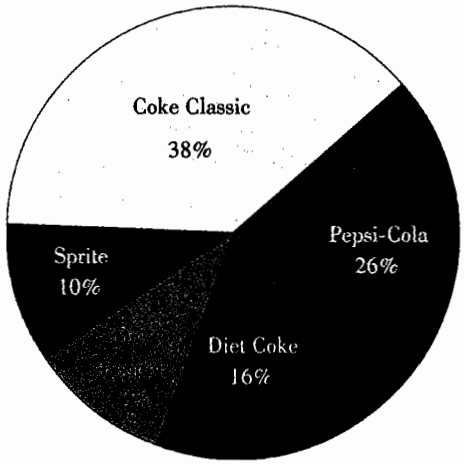


图 2.2　软饮料购买情况的饼形图

评　注

1. 在一个频数分布中，组别数往往与在数据中发现的类别数是相同的，就像在本章软饮料购买数据中的情况一样。数据类别只包括 5 种软饮料，一个独立的频数分布组别被定义为一种软饮料。包括所有软饮料的数据将需要很多类，其中的大部分类别会只有很少的购买次数。大多数统计学家建议把只有很少频数的组合并到一个叫“其他”的综合组中去。只有 5% 或更少频数的组常常被这样处理。
2. 在任何频数分布中，频数的总和总是等于观察值的数目。在任何相对频数分布中，相对频数的总和总是等于 1.00，在任何百分比频数分布中，百分数的总和总是等于 100。

练　习

方　法

1. 某问题的答案有三种选项：A、B 和 C。一个拥有 120 个答案的样本包括 60 个 A，24 个 B，36 个 C。求出它的频数和相对频数分布。

2. 已给出一部分相对频数分布如下：

组别	相对频数
A	0.22
B	0.18
C	0.40
D	

a. D 组的相对频数是多少？

b. 如果总的样本容量是 200，那么 D 组的频数是多少？

c. 求出频数分布。

d. 求出百分比频数分布。

3. 一份问卷调查表的答案中包括 58 个“是”，42 个“不”，20 个未表示意见。


自测题

a. 在构建饼形图时，表示“是”答案的部分占多少？

b. 表示“不”答案的部分占多少？

c. 作出饼形图。

d. 作出柱状图。

应　用

4. 根据 Nielsen 媒体研究的数据，2000 年 4 月 6 日晚 8:00 收视率最高的 4 个节目是 *Millionaire*、*Frasier*、*Chicago Hope* 和 *Charmed*（*USA Today*, April 13, 2000）。一个包括 50 名观众的样本数据如下：


CD 光盘数据
Nielsen

Millionaire	*Millionaire*	*Millionaire*	*Frasier*	*Charmed*
Frasier	*Frasier*	*Millionaire*	*Millionaire*	*Frasier*
Frasier	*Millionaire*	*Millionaire*	*Chicago Hope*	*Millionaire*
Charmed	*Millionaire*	*Frasier*	*Chicago Hope*	*Millionaire*
Chicago Hope	*Charmed*	*Frasier*	*Frasier*	*Millionaire*
Millionaire	*Frasier*	*Millionaire*	*Millionaire*	*Chicago Hope*
Frasier	*Millionaire*	*Millionaire*	*Charmed*	*Chicago Hope*
Chicago Hope	*Millionaire*	*Millionaire*	*Millionaire*	*Millionaire*
Frasier	*Frasier*	*Millionaire*	*Frasier*	*Frasier*
Millionaire	*Millionaire*	*Chicago Hope*	*Millionaire*	*Frasier*

a. 这些数据是定性的还是定量的？

b. 作出数据的频数分布和百分比频数分布。

c. 作出数据的柱状图和饼形图。

d. 从样本分析看，哪一个节目拥有最大市场份额？哪一个位居第二？

5. 进入东方大学商学院的新生被要求列出他们喜欢的专业，结果得到下列数据：

专业	数目
管理	55
会计	51
金融	28
营销	82

通过构建如下项目来汇总数据：

a. 相对频数分布和百分比频数分布。

b. 柱状图。

c. 饼形图。

6. 表 2.4 列出了八种 2000 年 2 月销售最好的精装商务图书（*Business Week*, April 3, 2000）。假设一个在科罗拉多丹佛地区取得的购书样本提供了这八种书的下列数据：



CD 光盘数据
BwBooks

7 Habits	Dad	7 Habits	Millionaire	Millionaire	WSJ Guide
Motley	Millionaire	Tax Guide	7 Habits	Dad	Dummies
Millionaire	Motley	Dad	Dad	Parachute	Dad
Dad	7 Habits	WSJ Guide	WSJ Guide	WSJ Guide	7 Habits
Motley	WSJ Guide	Millionaire	7 Habits	Millionaire	Millionaire
Millionaire	7 Habits	Millionaire	7 Habits	Motley	Motley
Motley	7 Habits	Dad	Dad	Dad	Dad
7 Habits	WSJ Guide	Tax Guide	Millionaire	Motley	Tax Guide
Motley	Motley	Millionaire	Millionaire	Dad	Dummies
Millionaire	Millionaire	Millionaire	Dad	Millionaire	Dad

a. 作出数据的频数分布和百分比频数分布。把所有频数在 5% 或者以下的图书归入一个“其他”组中。

b. 列出最畅销的图书。

c. *The Millionaire Next Door* 和 *Rich Dad, Poor Dad* 的销售百分比分别是多少？

表 2.4 销售最好的八种精装商务图书

- *The 7 Habits of Highly Effective People*
- *Investing for Dummies*
- *The Ernst & Young Tax Guide 2000*
- *The Millionaire Next Door*
- *The Motley Fool Investment Guide*
- *Rich Dad, Poor Dad*
- *The Wall Street Journal Guide to Understanding Money and Investing*
- *What Color is Your Parachute? 2000*

7. 

佛罗里达州 Maderia 海滨的 Leverock's Waterfront Steakhouse 使用一种问卷调查表来询问顾客对于餐馆的服务、食品质量、鸡尾酒、价格和气氛是如何评价的。对每一个属性分别以优异(O)、非常好(V)、好(G)、一般(A)和差(P)的标准来评价。利用描述性统计汇总有关食品质量的下列数据，你对该餐馆的食品质量如何评价？

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

8. 一个样本包括纽约 Cooperstown 棒球名人堂的 55 名成员，他们的场上位置数据如下。每一个观察值表示他们所担任的首要位置：投手（P）、接手（H）、一垒（1）、二垒（2）、三垒（3）、游击手（S）、左场（L）、中场（C）、右场（R）。

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

a. 使用频数分布和相对频数分布来汇总数据。

b. 什么位置造就的荣誉成员最多？

c. 什么位置造就的荣誉成员最少？

d. 哪一个外场位置(L, C 或 R)造就的荣誉成员最多？

e. 比较内场成员(1，2，3 和 S)和外场成员(L, C 和 R)。

9. 电子协会的雇员执行弹性工作时间，他们可以选择上午 7:00、7:30、8:00、8:30 或者 9:00 上班。下列数据来自一个雇员们选择的上班时间的样本：

7:00	8:30	9:00	8:00	7:30	7:30	8:30	8:30	7:30	7:00
8:30	8:30	8:00	8:00	7:30	8:30	7:00	9:00	8:30	8:00

构建下列项目以汇总数据：

a. 频数分布。

b. 百分比频数分布。

c. 柱状图。

d. 饼形图。

e. 关于在弹性工作时间制度下雇员的偏好，数据汇总显示了什么结果？

10. 辛辛那提大学工商管理学院的学生被要求在完成其课程后填写一个课程评估问卷调查表，它包括一系列问题，答案分为五个档次：

差　　一般　　好　　非常好　　优异

其中的一个问题为：与你已经完成的其他课程比较，你刚刚完成的这门课程的综合质量怎样？一个样本由60名刚刚完成2000年春季学期商务统计课程的学生组成，它给出了下列答案，为了便于进行计算机处理，使用了数值来表示调查结果，其中，1 = 差，2 = 一般，3 = 好，4 = 非常好，5 = 优异。

3	4	4	5	1	5	3	4	5	2	4	5	3	4	4
4	5	5	4	1	4	5	4	2	5	4	2	4	4	4
5	5	3	4	5	5	2	4	3	4	5	4	3	5	4
4	3	5	4	5	4	3	5	3	4	4	3	5	3	3

a. 说明这些数据为什么是定性数据。

b. 作出数据汇总的频数分布和相对频数分布。

c. 作出数据汇总的柱状图和饼形图。

d. 在你汇总的基础上，说明学生对这门课程的综合评价。

2.2 汇总定量数据

频数分布

正如2.1节所定义的那样，频数分布是表示在几个互不重叠组别中每一组所包含的项目个数(频数)的表格汇总。这个定义也适用于定量数据。不过对于定量数据，我们在定义用于频数分布的互不重叠分组时，不得不更加小心。

表2.5 年末审计时间(天)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13



CD光盘数据
Audit

例如，考虑表2.5中的定量数据。一个样本包括一个小会计师事务所Sanderson and Clifford的20个客户，数据表示完成样本的年末审计时间(天)。为定义定量数据的频数分布中组的概念，需要下列三个步骤：

1. 确定互不重叠分组的数目。
2. 确定每一组的宽度。
3. 确定组限。

我们通过对表2.5中的审计时间数据进行处理以构建频数分布，从而说明这些步骤。

组数(number of classes)。组是通过对数据规定范围而形成的。作为一般性原则，我们建议使用5—20个组。对于较少的数据项目，5或6组就可以汇总数据了。而对于较多的数据项目，通常需要更多的组数。最终的目的就是使用足够的组来显示数据的变异程度，而不是使用很多个组以致每个组内只含有很少的数据项目。因为表2.5中的数据项目的个数相对较少($n=20$)，我们

决定使用 5 个组来构建频数分布。

组宽(width of the classes)。构建定量数据频数分布的第二步是选择组宽。作为一般性原则，我们建议每一组的宽度应相同，因此对组数和组宽的选择就不再是互相独立的决定了。较大的组数意味着较小的组宽，反之亦然。为确定一个近似的组宽，我们从找到数据的最大值和最小值开始。然后，一旦确定了期望的组数，我们就能够使用如下表达式来确定近似组宽。

每组的宽度相同减少了使用者的不恰当解释。

$$\text{近似组宽} = \frac{\text{数据最大值} - \text{数据最小值}}{\text{组数}} \tag{2.2}$$

由公式(2.2)得出的近似组宽可以根据构建频数分布的那个人的意愿取整为更方便的值。例如，一个 9.28 的近似组宽可以被简单地取整为 10，因为 10 作为组宽在构建和表达频数分布时更方便。

对于年末审计时间数据，最大值是 33，最小值为 12。因为我们决定用 5 个组来汇总数据，由公式(2.2)，可计算近似组宽为(33 - 12)/5 = 4.2。因此我们决定在频数分布中使用 5 天作为组宽。

在实践应用中，组数和近似的组宽要用试误法确定。一旦选定了一个可能的组数，就根据公式(2.2)找出近似组宽。这个过程能对不同的组数反复进行。最终，分析者凭判断来确定组数和组宽的组合，以构建汇总数据的最佳频数分布。

对数据集来讲，不存在一个最佳频数分布。不同的人可以构建不同的频数分布，都能被接受。频数分布的目的是揭示出自然的分组和数据的差异。

对于表 2.5 中的审计时间数据，在决定了使用 5 个组别，每一组具有 5 天的组宽之后，下一步工作就是规定每一组的组限。

组限(class limits)。必须要选择组限以确保每个数据项目属于且只属于一个组别。**组下限**是被分到该组的数据的最小可能值，**组上限**是被分到该组的数据的最大可能值。在对定性数据构建频数分布的过程当中，我们不需要规定组限，因为每一个数据项都会自然地落入一个互不重叠的组(类)中。但是对于定量数据，比如表 2.5 中的审计时间，组限就是必要的，以确定每个数据值究竟属于哪一组。

对于表 2.5 中的审计时间数据，我们确定组限是 10—14，15—19，20—24，25—29，30—34。最小数据值 12 就包括在 10—14 组中，而最大数据值 33 包括在 30—34 组中。对于 10—14 组，10 是组下限，14 是组上限。相邻两组的组下限之差就是组宽。使用头两个组下限 10 和 15 来计算，组宽是 15 - 10 = 5。

一旦组数、组宽、组限都已确定，我们通过**清点**每一组中数据的个数就可以得到频数分布。例如，表 2.5 中的数据显示：4 个数据值——12，14，14，13 属于 10—14 组。因此，10—14 组的频数是 4。对 15—19，20—24，25—29，30—34 各组继续清点计数过程，就得到了表 2.6 中的

表 2.6 审计时间数据的频数分布

审计时间(天)	频数
10 — 14	4
15 — 19	8
20 — 24	5
25 — 29	2
30 — 34	1
总计	20

频数分布。利用这个频数分布，我们能够观察到：

1. 最频繁发生的审计时间处于 15—19 天这一组，在 20 个审计时间中有 8 个属于这一组。
2. 只有一次审计花了 30 天或更多时间。

得出其他结论也有可能，这取决于频数分布观察者的兴趣。频数分布的价值就在于它提供了对数据的深刻理解，而直接观察原始的未经组织的数据，是无法获得这种深刻理解的。

组中值(class midpoint)。在一些应用中，我们需要知道定量数据频数分布的组中点。**组中值**是组下限和组上限的中间值。对于审计时间数据，五个组中值分别是 12,17,22,27 和 32。

相对频数和百分比频数分布

我们以定义定性数据时同样的方式来定义定量数据的相对频数和百分比频数分布。首先，我们已知相对频数就是一个组的观察值个数占总数的比例。对于 n 个观察值，

$$\text{组的相对频数} = \frac{\text{组内频数}}{n}$$

组的百分比频数等于相对频数乘以 100。

根据表 2.6 中的频数且当 $n=20$ 时，表 2.7 显示了审计时间数据的相对频数和百分比频数分布。注意到对于所有的审计，有 0.40 或 40% 需要 15—19 天时间。只有 0.05 或 5% 的审计需要 30 天或更多时间。通过使用表 2.7，我们还能够得到其他的解释和理解。

表 2.7 审计时间数据的相对频数和百分比频数分布

审计天数	相对频数	百分比频数
10 — 14	0.20	20
15 — 19	0.40	40
20 — 24	0.25	25
25 — 29	0.10	10
30 — 34	0.05	5
总计	1.00	100

点位图

最简单的数据图形汇总之一就是**点位图**(dot plot)。它的横轴表示观察值的取值范围，每一个观察值由位于横轴之上的点表示。图 2.3 是表 2.5 中审计时间数据的点位图，位于横轴刻度 18 之上的 3 个点指出有 3 个观察值的数值等于 18。点位图能够显示数据的细节，它对于比较两个或更

多个变量的数据分布是有用的。

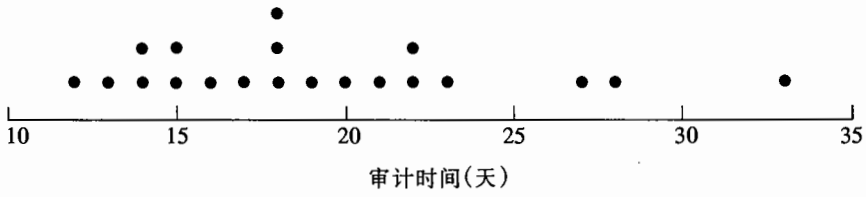


图 2.3　审计时间数据的点位图

直方图

一个常用的定量数据图形表达方式是**直方图**(histogram)。这种图形可以由任何形式的频数数据构建，而不论这些数据先前是被汇总为频数分布、相对频数分布还是百分比频数分布。把有关的变量设置为横轴，把频数、相对频数或百分比频数设置为纵轴，我们就可以作出一个直方图。在横轴之上画一个长方形，它以横轴上的组间距为底，以该组对应的频数、相对频数或百分比频数为高，就能够表示出每一组的频数、相对频数或百分比频数。

图 2.4 是审计时间数据的直方图。我们可以注意到具有最大频数的那个组由 15—19 天这一组上面的长方形表示，长方形的高度表示本组的频数是 8。本例中由相对频数或百分比频数分布表示的直方图与图 2.4 中的直方图看上去基本相同，只是纵轴换成了相对或百分比频数值。

如图 2.4 所示，直方图中的相邻长方形互相接触。与柱状图不同，直方图中相邻组的长方形之间没有自然的分隔，这种格式是直方图的惯例。因为审计时间数据的组限为 10—14，15—19，20—24，25—29，30—34，各组间从 14 到 15、19 到 20、24 到 25，以及 29 到 30 有一个单位的间隔，可以通过在组限间的中点位置画直方图的垂线而消除这个间隔。在图 2.4 中分隔各组的直方图垂线分别在 9.5，14.5，19.5，24.5，29.5 和 34.5 处。这个微调消除了直方图中组间的间隔，从而有助于表明：虽然数据经过了取整，但在第一组的下限和最后一组的上限之间数据取任何值都是可能的。

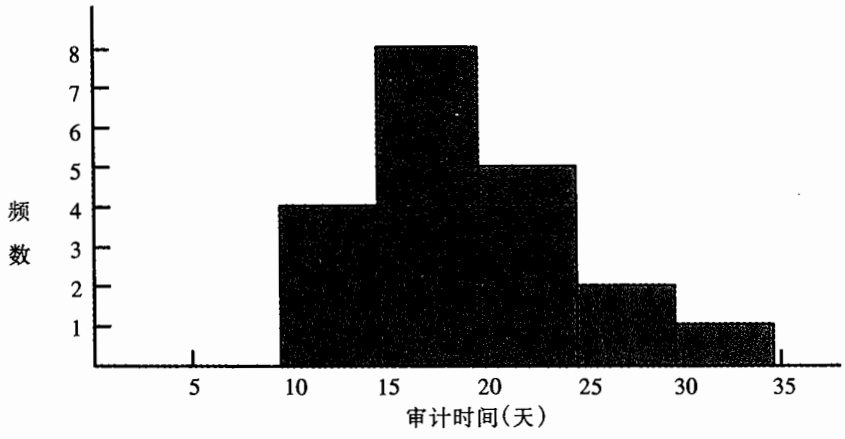


图 2.4　审计时间数据的直方图

累积分布

作为频数分布的一种变化形式，**累积频数分布**(cumulative frequency distribution)提供了定量数据的另一种表格汇总方法。累积频数分布也使用组数、组宽、组限这些要素来构建频数分布。但是，累积频数分布表示的是其数值**小于或等于每一组上限**的数据项的个数，而不是每一组的数据个数。表 2.8 的前两列给出了审计时间数据的累积频数分布。

表 2.8　审计时间数据的累积频数、累积相对频数和累积百分比频数分布

审计时间（天）	累积频数	累积相对频数	累积百分比频数
小于或等于 14	4	0.20	20
小于或等于 19	12	0.60	60
小于或等于 24	17	0.85	85
小于或等于 29	19	0.95	95
小于或等于 34	20	1.00	100

为了理解如何确定累积频数，考虑一个被描述为“小于或等于 24”的组。对数据值小于或等于 24 的所有组别的频数求和，即可得出该组的累积频数。对于表 2.6 中的频数分布，组 10—14，15—19 和 20—24 的频数之和表示共有 4 + 8 + 5 = 17 个数据值小于或等于 24。因此，该组的累积频数分布就是 17。另外，表 2.8 中的累积频数分布指出有 4 次审计在 14 天内完成，有 19 次审计在 29 天内结束。

最后，我们注意到**累积相对频数分布**(cumulative relative frequency distribution)表示的是数据个数占总数的比例，**累积百分比频数分布**(cumulative percent frequency distribution)表示数值小于或等于每组上限的数据个数所占的百分比。我们既可以对相对频数分布的相对频数求和，也可以将累积频数除以数据总数，从而计算得到累积相对频数分布。如果使用后一种方法，我们发现表 2.8 第 3 列的累积相对频数等于第 2 列的累积频数除以数据总数(n = 20)。累积百分比频数是由累积相对频数乘以 100 得出的。累积相对频数和百分比频数分布显示出有 0.85 或 85% 的审计在 24 天内完成，而有 0.95 或 95% 的审计在 29 天内结束，如此等等。

累积曲线

累积分布的图形中有一种被称为**累积曲线**(ogive)。数据值显示在横轴上，而累积频数、累积相对频数或累积百分比频数显示在纵轴上。图 2.5 就是表 2.8 中审计时间数据累积频数的累积曲线。

我们通过画出每个对应于每组累积频数的点，就可以得到累积曲线。因为审计时间数据的组限分别是 10—14，15—19，20—24 等，在 14—15，19—20 等之间有一个单位的间隔。正如我们在直方图中的处理方法一样，通过在两个组限的中点取点，可以消除这些间隔。因此，14.5 用于 10—14 组，19.5 用于 15—19 组，依此类推。“小于或等于 14”这一组的累积频数为 4，它在图2.5

的累积曲线上所对应的点的横坐标是 14.5，纵坐标是 4。“小于或等于 19”这一组的累积频数为 12，它所对应的点的横坐标是 19.5，纵坐标是 12。我们注意到还有一个额外的点位于累积曲线的最左端，这个点是曲线的起点，它表明没有数据值处于 10—14 组以下。它的横坐标是 9.5，纵坐标等于 0。将作出的点用直线连接起来，就完成了累积曲线。

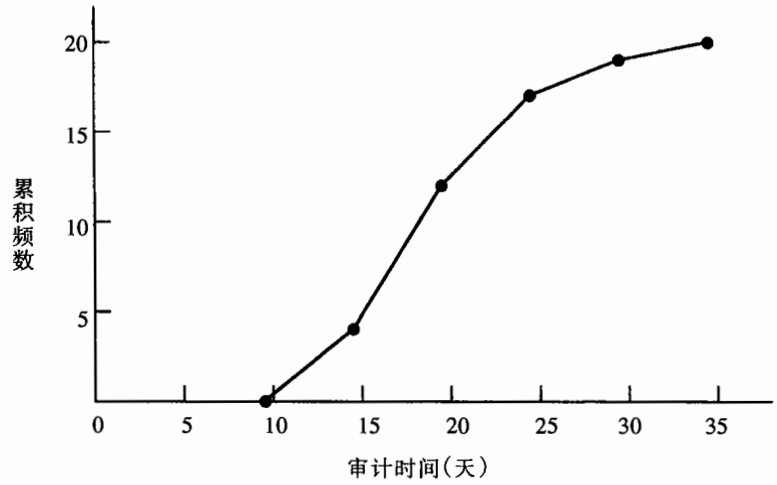


图 2.5　审计时间数据的累积曲线

评　注

1. 对于定量数据而言，组限的恰当值依赖于数据的精确水平。例如，对表 2.5 中的审计时间数据，因为数据已经被取整到最接近的天数，因此使用的组限是整数。如果数据被取到最接近的 1/10 天(如 12.3，14.4 等等)，那么组限也将以 1/10 天来表示，例如，第一组组限将会是 10.0—14.9。如果数据被取到最接近的 1/100 天(如 12.34，14.45 等等)，组限也会以 1/100 天来表示，例如，第一组的组限将会是 10.00—14.99。
2. **开口组**是指只有组下限或只有组上限的组。例如，在表 2.5 中的审计时间数据内，假设两次审计分别用了 58 和 65 天。以 5 为组宽把组区间延伸至 35—39，40—44，45—49 等等。我们能够通过使用一个“35 或更长时间”的开口组来简化频数分布，该组的频数为 2。开口组最经常出现在分布的上端，但有时也出现在下端，偶尔在两端都出现。
3. 累积频数分布的最后一个数据项总是观察值的合计数，累积相对频数分布的最后一个数据项总是 1.00，而累积百分比频数分布的最后一个数据项总是 100。
4. 柱状图和直方图基本上是同一事物，它们都是频数分布数据的图形表示。直方图就是各条之间没有间隔的柱状图。各柱之间的间隔适合于定性数据，因为定性数据是离散的，不存在中间值。而对于一些定量数据，各柱之间的间隔也是合适的。例如，在学院中注册的学生的班级数。也可以假设数据只取整数值，像 1.5，2.73 这样的中间值是不存在的。而对于连续的定量数据，例如表 2.5 中的审计时间，使用柱间的间隔就不合适。

练　习

方　法

11. 考虑下列数据：

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

a. 以 12—14，15—17，18—20，21—23 和 24—26 为组限构造频数分布。

b. 使用(a)中的组限构造相对频数和百分比频数分布。

12. 考虑下列频数分布：



组别	频数
10—19	10
20—29	14
30—39	17
40—49	7
50—59	2

构建累积频数和累积相对频数分布。

13. 用练习 12 中的数据构建直方图和累积曲线。

14. 考虑下列数据：

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

a. 作出点位图。

b. 作出频数分布。

c. 作出百分比频数分布。

应　用

15. 自测题

一名办公室职员医生研究了到办公室来急诊的病人等待时间。下面是在一个月的时间内收集到的数据(等待时间以分计)：

2　5　10　12　4　4　5　17　11　8　9　8　12　21　6　8　7　13　18　3

使用 0—4，5—9 等等来分组。

a. 作出频数分布。

b. 作出相对频数分布。

c. 作出累积频数分布。

d. 作出累积相对频数分布。

e. 需要急诊的病人中等待时间在 9 分钟以内的人数占多大比例?

16. CD 光盘数据 Stocks

表 2.9 显示了从 Stock Invester Pro 数据库中采集的一个由 25 个计算机硬件公司组成的样本。

a. 构建股票价格数据的表格汇总和直方图。

b. 构建每股收益数据的表格汇总和直方图，并分析你的结果。

表 2.9　25 个计算机硬件公司的数据集

公司	股价	机构持股 (%)	价格/账面价值	每股收益 (美元/年)
Amdahl	12.31	45.4	2.49	-2.49
Auspex Systems	11.00	66.1	2.22	0.85
Compaq Computer	65.50	83.0	6.84	2.01
Data General	35.94	91.5	4.25	1.15
Digi International	15.00	33.4	2.04	-0.89
Digital Equipment Corp.	43.00	58.8	1.92	-2.93
En Pointe Technologies	14.25	11.8	3.47	0.80
Equitrac	16.25	20.9	2.38	0.76
Franklin Electronic Pbls.	12.88	30.8	1.41	0.82
Gateway 2000	39.13	36.0	6.45	1.74
Hewlett-Packard	61.50	50.2	4.35	2.64
Ingram Micro	28.75	14.4	4.53	1.01
Maxwell Technologies	30.50	26.5	8.07	0.46
MicroAge	27.19	76.6	2.16	1.25
Micron Electronics	16.31	18.8	4.48	1.06

（续表）

公司	股价	机构持股（%）	价格/账面价值	每股收益（美元/年）
Network Computing Devices	11.88	39.8	3.34	0.15
Pomeroy Computer Resource	33.00	56.9	3.29	1.81
Sequent Computer Systems	28.19	57.0	2.65	0.36
Silicon Graphics	27.44	63.0	3.01	0.44
Southern Electronics	15.13	41.9	2.46	0.99
Stratus Computer	55.50	77.2	2.48	2.52
Sun Microsystems	48.00	59.3	7.50	1.67
Tandem Computers	34.25	61.3	3.61	1.02
Tech Data	38.94	82.3	3.80	1.50
Unisys	11.31	34.8	16.64	0.08

资料来源：*Stock Investor Pro*，American Association of Individual Investors, August 31, 1997。

17. 国民航空公司接受电话订票。下面的数据来自一个包含20次电话订票的样本，表示它们的通话时间（以分钟计）。构建这些数据的频数分布和相对频数分布，并作出直方图。

2.1	4.8	5.5	10.4
3.3	3.5	4.8	5.8
5.3	5.5	2.8	3.6
5.9	6.6	7.8	10.5
7.5	6.0	4.5	4.8

18. Wageweb对工资收入进行调查并在它的网站上公布了汇总结果。Wageweb根据2000年1月1日的工资收入数据，报道市场营销副总裁的年薪范围是85 090美元—190 054美元（*Wageweb. com*, April 12, 2000）。假定下列数据就来自一个包括50个营销副总裁年薪的样本，数据的单位是千美元。



CD光盘数据
Wageweb

145	95	148	112	132
140	162	118	170	144
145	127	148	165	138
173	113	104	141	142
116	178	123	141	138
127	143	134	136	137
155	93	102	154	142
134	165	123	124	124
138	160	157	138	131
114	135	151	138	157

a. 最低和最高年薪分别是多少？

b. 以 15 000 美元为组宽，作出年薪数据的表格汇总。

c. 年薪在 135 000 美元以内的副总裁占多大比例？

d. 年薪超过 150 000 美元的副总裁的百分比是多少？

e. 作出数据的直方图。

19. 最近 20 天来工人生产的产品数量数据如下：

160	170	181	156	176
148	198	179	162	150
162	156	179	178	151
157	154	179	148	156

构建下列项目以汇总数据：

a. 频数分布。

b. 相对频数分布。

c. 累积频数分布。

d. 累积相对频数分布。

e. 累积曲线。

20. 美国人口普查局公布了美国人口的变化信息。下表是 2000 年 7 月 1 日美国人口年龄的百分比频数分布（*The World Almanac and Book of Facts 2000*）：

年龄	百分比频数
0—13 岁	20.0
14—17 岁	5.7
18—24 岁	9.6
25—34 岁	13.6
35—44 岁	16.3
45—54 岁	13.5
55—64 岁	8.7
65 岁或以上	12.6
	100.0

a. 34 岁以下人口所占的百分比是多少？

b. 25 和 54 岁之间的人口百分比是多少？

c. 超过 34 岁的人口百分比是多少？

d. 已知总人口为 2.75 亿，那么有多少人年龄低于 25 岁？

e. 假设你相信在 55—64 岁这一组中有一半的人已退休，并且几乎所有超过 65 岁的人已经退休，估计退休人员的数目。

21. 《尼尔森家庭技术报告》（*Nielsen Home Technology Report*, February 20, 1996）专门报道家用技

术,它的用户年龄在 12 岁以上。下列数据是在一个 50 人的样本中一周内使用个人电脑的小时数:


CD 光盘数据
Computer

4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

构建下列项目以汇总数据:

a. 频数分布(组宽为 3 小时)。

b. 相对频数分布。

c. 直方图。

d. 累积曲线。

e. 分析家庭个人电脑的使用情况,这些数据显示出了什么问题?

2.3 探索性数据分析:茎叶显示

探索性数据分析(exploratory data analysis)技术包括简单的算术和易画的图形,它们可用于快速地汇总数据。其中有一种被称为**茎叶显示**(stem-and-leaf display)的技术就能够同时显示等级顺序和数据集的形状。

为了说明茎叶显示的使用,考虑表 2.10 的数据,它们是 50 名参加 Haskens 公司生产职位面试的人员对 150 个能力测试问题的回答结果,数据值显示的是被正确回答的问题个数。

表 2.10 在能力测试中被正确回答的问题个数


CD 光盘数据
ApTest

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

为构建茎叶显示,我们首先把每个数据值的高位数字排在竖线的左边。而在竖线的右边,我们以每个观察值原来的记录顺序来记录每个数据值的最后一位,并把它们放在对应的高位数字同一行中。示例如下:

6	9	8									
7	2	3	6	3	6	5					
8	6	2	3	1	1	0	4	5			
9	7	2	2	6	2	1	5	8	8	5	4
10	7	4	8	0	2	6	6	0	6		
11	2	8	5	9	3	5	9				
12	6	8	7	4							
13	2	4									
14	1										

经过这样组织数据以后，排列出每一行的数字等级顺序就相对简单了。排序后，产生了如下的茎叶显示：

6	8	9									
7	2	3	3	5	6	6					
8	0	1	1	2	3	4	5	6			
9	1	2	2	2	4	5	5	6	7	8	8
10	0	0	2	4	6	6	6	7	8		
11	2	3	5	5	8	9	9				
12	4	6	7	8							
13	2	4									
14	1										

于是，竖线左边的数字（6, 7, 8, 9, 10, 11, 12, 13, 14）形成了**茎**，而竖线右边的每个数字就是一个**叶**。例如，考虑第一行，6 是茎值，8 和 9 是两个叶值：

6 | 8 9

它表示有两个数据的值其首位数字是 6，叶值表示数据值是 68 和 69。同样，对于第二行：

7 | 2 3 3 5 6 6

它指出有 6 个数据的值首位数字是 7，叶值表示数据值是 72，73，73，75，76 和 76。

为了把注意力集中到茎叶显示所显示出的形状上来，我们使用长方形围住每个茎的叶值。这样做，可得到下图：

6	8	9									
7	2	3	3	5	6	6					
8	0	1	1	2	3	4	5	6			
9	1	2	2	2	4	5	5	6	7	8	8
10	0	0	2	4	6	6	6	7	8		
11	2	3	5	5	8	9	9				
12	4	6	7	8							
13	2	4									
14	1										

将该页逆时针旋转 90°，所得到的图形与组别为 60 — 69，70 — 79，80 — 89 等等的直方图很相似。

即使茎叶显示有可能与直方图提供了相同的信息，它还是具有两个主要的优点：

1. 茎叶显示更易于手画。
2. 在一个组内，由于茎和叶显示出了实际数据值，因此茎叶显示能比直方图提供更多的信息。

正如频数分布或直方图没有绝对的组数一样，茎叶显示也没有绝对的行或茎数。如果我们认为原始的茎叶显示把数据过分浓缩了，我们能够把每个高位数字拆分为两个或更多的茎，从而很容易地扩展这个显示。例如，对每一个高位数字使用两个茎，我们把所有以 0，1，2，3，4 结尾的数据放在一行，而把所有以 5，6，7，8，9 结尾的数据放在另一行。下列扩展的茎叶显示说明了这种方法：

6	8 9
7	2 3 3
7	5 6 6
8	0 1 1 2 3 4
8	5 6
9	1 2 2 2 4
9	5 5 6 7 8 8
10	0 0 2 4
10	6 6 6 7 8
11	2 3
11	5 5 8 9 9
12	4
12	6 7 8
13	2 4
13	
14	1

在扩展的茎叶显示中，每当茎值被重复两次，则第一个茎值对应于叶值 0— 4，第二个茎值对应于叶值 5— 9。

注意到值 72，73 和 73 的叶值在 0—4 范围内，并且被列在茎值 7 的第一行里。值 75，76 和 76 的叶值在 5—9 范围内，列在茎值 7 的第二行里。这个扩展的茎叶显示与组别为 65—69，70—74，75—79 等等的频数分布很类似。

前面的例子说明了三位数数据的茎叶显示。对于超过三位数的数据，其茎叶显示也是可能的。例如，考虑下列数据，它们表示在 15 个星期中每个星期快餐店销售的汉堡包数量。

1 565	1 852	1 644	1 766	1 888	1 912	2 044	1 812
1 790	1 679	2 008	1 852	1 967	1 954	1 733	

这些数据的茎叶显示如下：

叶单位 = 10

15	6			
16	4	7		
17	3	6	9	
18	1	5	5	8
19	1	5	6	
20	0	4		

我们注意到在此处使用一位数来表示每一个叶值，并且每个观察值只有前三位数字被用来构建显示。在该显示的顶部，我们规定了叶值的单位 = 10。为了说明如何解释该显示中的数值，考虑第一个茎值 15，以及它的叶值 6，把它们连起来，我们得到数 156。要得到原始观察值的一个近似值，我们须把它乘以**叶单位**的值 10。因此，$156 \times 10 = 1\,560$ 就是用来构建茎叶显示原始观察值的近似值。虽然从该茎叶显示得出精确数据是不可能的，但根据每个叶值使用一位数表示的惯例，可以保证我们用位数很多的数据也能构建出茎叶显示。对于叶单位没有给出的茎叶显示，我们就假设它等于 1。

在茎叶显示中使用一位数来表示每个叶。叶的单位表示为了得到原始数据的近似值，茎叶显示中的数值应乘上的数。叶单位可以是 100，10，1，0.1 等等。

练　习

方　法

22. 构建下列数据的茎叶显示：

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

23. 构建下列数据的茎叶显示：



11.3	9.6	10.4	7.5	8.3	10.5	10.0
9.3	8.1	7.7	7.5	8.4	6.3	8.8

24. 构建下列数据的茎叶显示，叶单位为 10。

1 161	1 206	1 478	1 300	1 604	1 725	1 361	1 422
1 221	1 378	1 623	1 426	1 557	1 730	1 706	1 689

应　用

25. 某心理学家发明了一种成人智力的新的测试方法。对 20 人进行了测试，取得如下数据：


自测题

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

构建数据的茎叶显示。

26. 由 20 家公司组成的样本的每股收益数据来自《商业周刊》的公司业绩版（*Business Week*, November 17, 1997），数据如下：

公司	每股收益（美元）	公司	每股收益（美元）
Barnes & Noble	0.78	Hershey Foods	1.97
Citicorp	7.10	Hewlett-Packard	2.82
Compaq Computer	2.16	Humana	0.89
Dana	3.42	Microsoft	2.66
Dell Computer	2.03	Procter & Gamble	2.53
Digital Equipment	1.28	Quaker State	0.41
General Dynamics	4.82	Sara Lee	2.08
Goodyear	0.94	Snap-On Tools	2.38
Harley-Davidson	1.11	Sunstrand	2.53
Heinz	0.98	Xerox	3.95

构建数据的茎叶显示，把叶单位设为 0.1。说明你从这些公司的每股收益数据中了解到了什么？

27. 在一个工作满意度研究中，对 50 个项目进行了一系列的测试。下面是取得的数据，其中较高的分数表示较大的不满意度。


CD 光盘数据
JobSat

87	76	67	58	92	59	41	50	90	75	80	81	70
73	69	61	88	46	85	97	50	47	81	87	75	60
65	92	77	71	70	74	53	43	61	89	84	83	70
46	84	76	78	64	69	76	78	67	74	64		

构建数据的茎叶显示。

28. *Barron's* 定期公布道琼斯工业股票平均指数成分股的盈利预测。下列数据就是由 *Barron's* 盈利预测给出的这些公司的 2000 年市盈率（P/E）预测（*Barron's*, February 14, 2000）。



CD 光盘数据
PEforcast

公司	2000 P/E 预测	公司	2000 P/E 预测
AT&T	23	Honeywell	13
Alcoa	15	IBM	28
American Express	25	Intel	37
Boeing	16	International Paper	14
Caterpillar	13	Johnson & Johnson	23
Citigroup	17	McDonald's	23
Coca-Cola	39	Merck	25
Disney	47	Microsoft	60
Dupont	18	Minnesota Mining	19
Eastman Kodak	11	J. P. Morgan	11
Exxon/Mobil	22	Philip Morris	5
General Electric	37	Procter & Gamble	26
General Motors	8	SBC Comm.	19
Hewlett-Packard	36	United Technologies	14
Home Depot	48	Wal-Mart	40

a. 构建数据的茎叶显示。

b. 用茎叶显示的结果来构建数据的频数分布和百分比频数分布。

2.4　交叉分组列表和散点图

本章到目前为止，我们已集中讨论了在**同一时间对一个变量**的数据进行汇总的表格和图形方法。但管理人员和决策者往往还对能够帮助其**理解两变量间关系**的表格和图形方法感兴趣。交叉分组列表和散点图就是两类这样的方法。

交叉分组列表和散点图被用来揭示变量间的关系，以此来汇总数据。

交叉分组列表

交叉分组列表(crosstabulation)是一种能够同时汇总两个变量数据的方法。我们通过下面的应用来说明交叉分组列表的用途。“Zagat 饭店评论”(Zagat 's Restaurant Review)是一个提供世界各地饭店数据的服务机构，它报告饭店各种变量的数据，例如饭店的质量等级、典型的食品价格等。

质量等级是定性数据，等级类型是好、非常好和优异。食品价格是定量变量，通常的变化范围是10美元—49美元。一个样本包括洛杉矶地区的300家饭店，收集了它们的质量等级和食品价格数据，表2.11列出了前10家饭店的数据。

表2.11　洛杉矶饭店的质量等级和食品价格



CD光盘数据
Restaurant

饭店	质量等级	食品价格(美元)
1	好	18
2	非常好	22
3	好	28
4	优异	38
5	非常好	33
6	好	28
7	非常好	19
8	非常好	11
9	非常好	23
10	好	13
…	…	…

该应用案例的数据交叉分组列表如表2.12所示。左边和上部的标志规定了两变量的组别。在左边，行标志(好、非常好、优异)对应于质量等级变量的三个组。在上部，列标志(10—19美元，20—29美元，30—39美元，40—49美元)对应于食品价格变量的四个组。样本中的每个饭店都给出了质量等级和食品价格。因此，它们都与交叉分组列表中某一行和某一列的交叉单元相联系。例如，饭店5被鉴定为具有非常好的质量等级和33美元的食品价格。它属于表2.12中第2行和第3列的交叉单元。在构建交叉分组列表时，我们只需把属于交叉分组列表中每个单元的饭店数累加起来即可。

表2.12　300家洛杉矶饭店的质量等级和食品价格的交叉分组列表

	食品价格				
质量等级	10—19美元	20—29美元	30—39美元	40—49美元	总计
好	42	40	2	0	84
非常好	34	64	46	6	150
优异	2	14	28	22	66
总计	78	118	76	28	300

观察表2.12，我们看到样本中饭店的绝大部分(64)具有非常好的等级和处于20—29美元范围的食品价格。只有两个饭店具有优异等级的同时食品价格在10—19美元之间，其他频数也可进行类似解释。另外，我们注意到交叉分组列表的右边和底部分别给出了质量等级和食品价格的频数分布。从右边的频数分布中，我们看到质量等级的数据显示出有84个好饭店，150个非常好饭店，以及66个优异饭店。同样，底部显示出食品价格变量的频数分布。

交叉分组列表的值提供了关于变量间关系的深刻含义。根据表 2.12 的结果，较高的食品价格与较高的质量等级相联系，而较低的食品价格对应于较低的质量等级。

把表中的项目转换成行百分比或列百分比，还能够提供关于变量间关系的其他内容。对行百分比，用表 2.12 中的每个频数除以对应的行总量，其结果显示在表 2.13 中。例如，第一行和第一列交叉单元的百分值是 50.0，它是由 42 除以 84 再乘以 100 得到的（42/84 × 100 = 50.0%）。在最低质量等级（好）的饭店中，我们看到最大百分比的饭店是较便宜的饭店（50.0% 具有 10 —19 美元的食品价格，47.6% 具有 20 — 29 美元的食品价格）。在最高质量等级（优异）的饭店中，我们看到最大比例的饭店是较昂贵的饭店（42.4% 具有 30 — 39 美元的食品价格，33.4% 具有 40 — 49 美元的食品价格）。因此，我们继续看到较贵的价格与较高的等级相联系。

表 2.13　每一类质量等级的行百分比

	食品价格				
质量等级	10 — 19 美元	20 — 29 美元	30— 39 美元	40 — 49 美元	总计
好	50.0	47.6	2.4	0.0	100
非常好	22.7	42.7	30.6	4.0	100
优异	3.0	21.2	42.4	33.4	100

交叉分组列表被广泛地用作显示两变量间关系的工具。在实践中，许多统计调查的最终报告包括大量的交叉分组列表。在洛杉矶饭店样本里面，交叉分组列表是基于一个定性变量（质量等级）和一个定量变量（食品价格）构建的。当所有的变量都是定性或定量变量时，同样也能够构建出交叉分组列表。

散点图

散点图（scatter diagram）是对两个定量变量间关系的图形表述。作为散点图的一个示例，考虑旧金山一个音像设备商店的情况。该商店在过去三个月内有 10 次利用周末电视广告来促销。经理们想要调查是否可以证实在广告播出次数和下周商店销售额之间存在相关关系。表 2.14 显示了 10 周的以百美元为单位的销售额样本数据。

图 2.6 是表 2.14 中数据的散点图。广告的次数（x）显示在横轴上，销售额（y）显示在纵轴上。在第 1 周，$x = 2$ 且 $y = 50$，可以在散点图上按此坐标画出该点。用同样方法可画出其他 9 周的点。我们注意到有两个星期做了一次广告，两个星期做了两次广告，依此类推。

图 2.6 完成后的散点图指出在广告数和销售额之间存在着正相关的关系，更高的销售额与更多的广告次数相联系。由于所有的点不在一条直线上，这种相关关系是不完全的。可是，各点的总的分布情况还是说明整个关系是正相关的。

图 2.7 显示了一些一般的散点图形式和它们所表现出的关系类型。左上图是与上个例子相类似的正相关关系。而右上图中，散点图显示出变量之间没有明显的关系。而在下方的图中，随着 x 的增加，y 趋向减小，显示出一种负相关关系。

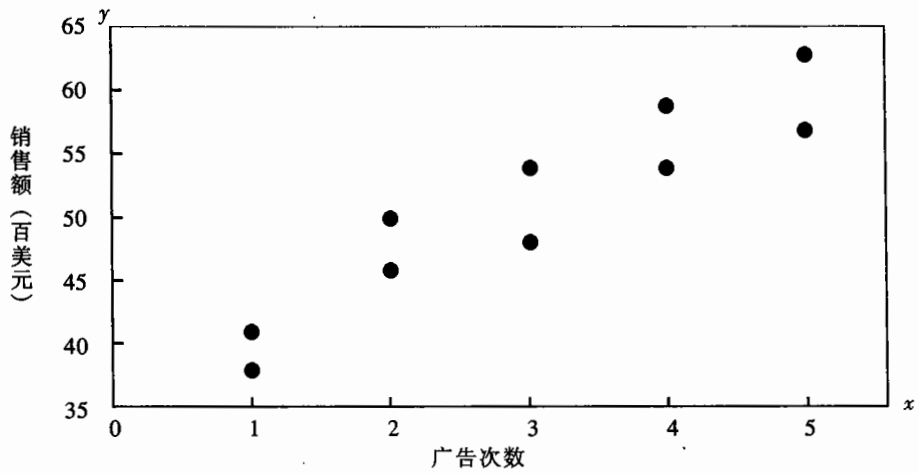


图 2.6　音像设备商店的散点图

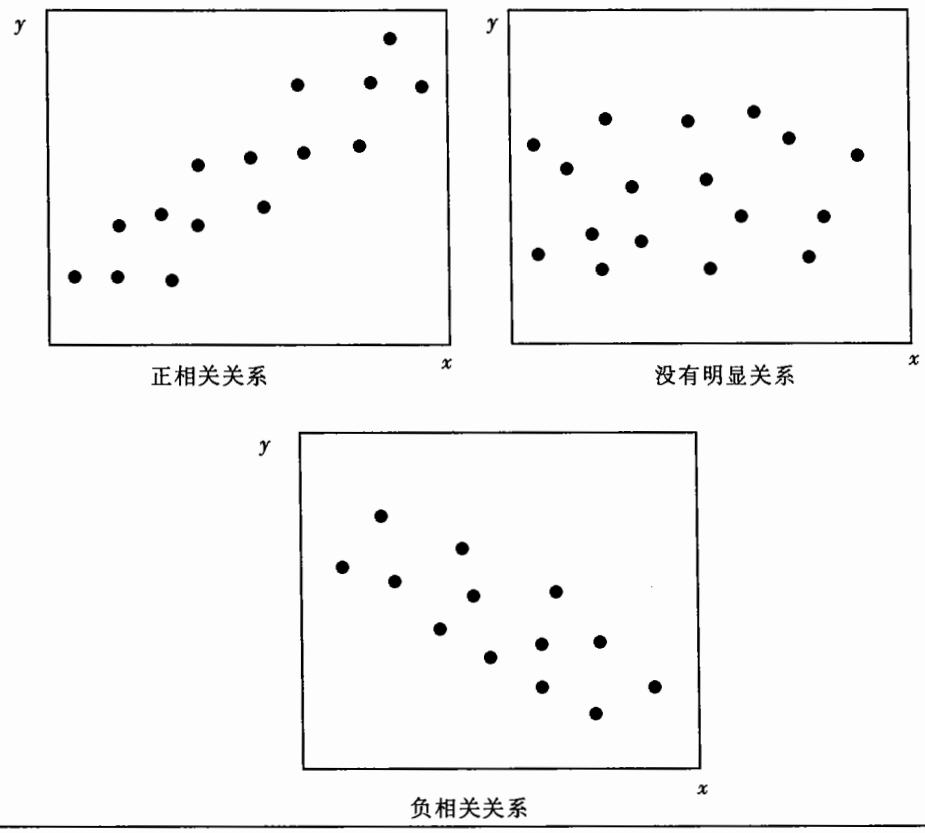


图 2.7　散点图显示出的关系类型

表 2.14　音像设备商店的样本数据


CD 光盘数据
Stereo

周	广告播出次数 x	销售额（百美元）y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

练　习

方　法

29. 下列数据是对两个定性变量 x 和 y 的 30 次观察值，x 的分类是 A，B 和 C，y 的分类是 1 和 2。


自测题


CD 光盘数据
Crosstab

观察次数	x	y	观察次数	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

a. 作出数据的交叉分组列表，以 x 为行，y 为列。

b. 计算行百分比。

c. 计算列百分比。

d. 如果有的话，x 和 y 之间是什么关系？

30. 下面是两个定量变量 x 和 y 的 20 对观察值：


CD 光盘数据
Scatter

观察次数	x	y	观察次数	x	y
1	-22	22	11	-37	48
2	-33	49	12	34	-29
3	2	8	13	9	-18
4	29	-16	14	-33	31
5	-13	10	15	20	-16
6	21	-28	16	-3	14
7	-13	27	17	-15	18
8	-23	35	18	12	17
9	14	-5	19	-20	-11
10	3	-3	20	-7	-22

a. 作出 x 和 y 之间关系的散点图。

b. 如果有的话，x 和 y 之间存在什么样的显著关系？

应　用

31. 计算表 2.12 中数据的列百分比。质量等级和食品价格之间的关系是什么？

32. 表 2.15 显示的是一个包括 36 家纽约证券交易所上市公司的样本的财务数据（*Investor's Business Daily*, April 7, 2000）。“销售额/边际收益/权益收益”栏的数据是一个基于公司的销售额增长率、边际利润以及所有者权益收益率的综合评分。每股收益评分是对公司每股收益增长的度量。


自测题

a. 构建销售额/边际收益/权益收益数据（行）和每股收益评分数据（列）的交叉分组列表。对每股收益评分数据按 0—19，20—39，40—59，60—79 和 80—99 分组。

b. 计算行百分比并对你所看到的变量间关系进行评论。

33. 参考表 2.15 中的数据。

a. 构建销售额/边际收益/投资收益率和产业群体相对实力数据的交叉分组列表。

b. 构建销售额/边际收益/投资收益率数据的频数分布。

c. 构建产业群体相对实力数据的频数分布。

d. 交叉分组列表是如何帮助构建(c)和(d)中的频数分布的？

表 2.15　36 家公司样本的财务数据


CD 光盘数据
IBD

公司	每股收益评分	相对价格实力	行业群体相对实力	销售额/边际收益/投资收益率
Advo	81	74	B	A
Alaska AirGp	58	17	C	B
Alliant Tech	84	22	B	B
Atmos Engy	21	9	C	E
Bank of Am.	87	38	C	A
Bowater PLC	14	46	C	D
Callaway Golf	46	62	B	E
Central Parking	76	18	B	C
Dean Foods	84	7	B	C
Dole Food	70	54	E	C
Elec. Data Sys	72	69	A	B
Fed. Dept. Stor.	79	21	D	B
Gateway	82	68	A	A
Goodyear	21	9	E	D
Hanson PLC	57	32	B	B
ICN Pharm.	76	56	A	D
Jefferson Plt	80	38	D	C
Kroger	84	24	D	A
Mattel	18	20	E	D
McDermott	6	6	A	C
Monaco	97	21	D	A
MurPhy Oil	80	62	B	B
Nordstrom	58	57	B	C
NYMAGIC	17	45	D	D
Office Depot	58	40	B	B
Payless Shoes	76	59	B	B
Praxair	62	32	C	B
Reebok	31	72	C	E
Safeway	91	61	D	A
Teco Energy	49	48	D	B
Texaco	80	31	D	C
US West	60	65	B	A
United Rental	98	12	C	A
Wachovia	69	36	E	B
Winnebago	83	49	D	A
York Intl.	28	14	D	B

资料来源：*Investor's Business Daily*, April 7, 2000。

34. 参考表 2.15 中的数据。

a. 构建每股收益评分和相对价格实力数据的散点图。

b. 评论变量间的关系，如果关系存在的话，每股收益评分的含义见习题 32 中的描述。相对价格实力是对过去 12 个月以来股票价格变化的度量，较高的值表示较强的实力。

35. 下列数据显示的是 10 个拉斯韦加斯赌场酒店的酒店收入和博彩收入，单位是百万美元（*Cornell Hotel And Restaurant Administration Quarterly*, October1997）。


CD 光盘数据
Casino

公司	酒店收入（百万美元）	博彩收入（百万美元）
Boyd Gaming	303.5	548.2
Circus Circus Enterprises	664.8	664.8
Grand Casinos	121.0	270.7
Hilton Corp. Gaming Div.	429.6	511.0
MGM Grand, Inc.	373.1	404.7
Mirage Resorts	670.9	782.8
Primadonna Resorts	66.4	130.7
Rio Hotel & Casino	105.8	105.5
Sahara Gaming	102.4	148.7
Station Casinos	135.8	358.5

a. 构建酒店收入和博彩收入数据的散点图。

b. 如果存在的话，评论变量间的关系。

小　结

对于数据集合，即使它们的规模适中，当它们处于刚被收集时的原始形式时，人们也往往很难对它们进行直接的解释。表格和图形方法提供了组织和汇总数据的方法，使得人们能够揭示出数据的模式并更加容易地解释它们。频数分布、相对频数分布、百分比频数分布、柱状图以及饼形图是汇总定性数据的表格和图形方法；而频数分布、相对频数分布、百分比频数分布、点位图、直方图、累积频数分布、累积相对频数分布、累积百分比频数分布以及累积曲线是汇总定量数据的方法；茎叶显示作为一种探索性的数据分析技术，能够用来汇总定量数据；交叉分组列表是汇总两个变量数据的表格方法；散点图是显示两个定量变量间关系的图形方法。图 2.8 是对本章介绍过的图形和表格方法的总结。

对于大型数据集合，在构建数据的图形和表格汇总时需要计算机软件包的帮助。在本章的两个附录中，我们将演示如何使用 Minitab 和 Microsoft Excel 来达到目的。

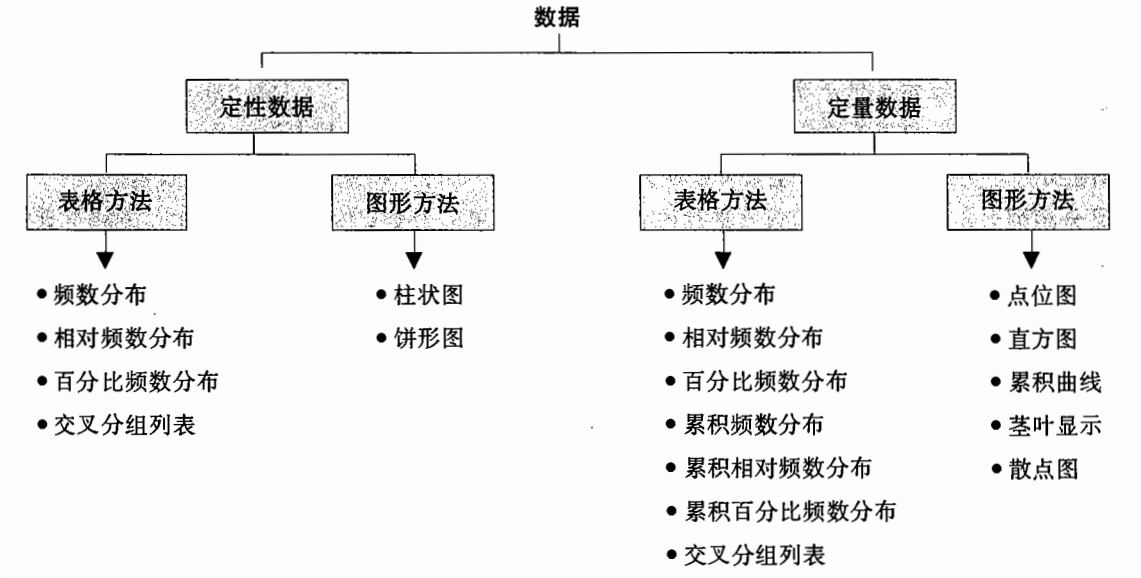


图 2.8　汇总数据的表格和图形方法

术语释义

定性数据： 用来识别项目等事物的种类，作为事物的标记或名称的数据。

定量数据： 表示数量多少的数据。

频数分布： 表示在几个互不重叠组别中每一组所包含的项目个数（频数）的数据表格汇总。

相对频数分布： 用来表示在几个互不重叠组别中每一组的项目个数所占比例（相对频数）的数据表格汇总。

百分比频数分布： 表示在几个互不重叠组别中每一组的项目个数所占百分比的数据表格汇总。

柱状图： 用于描述已被汇总为频数分布、相对频数分布和百分比频数分布的数据而使用的图形方法。

饼形图： 通过把一个圆分割为几部分，使每部分对应于每一组的相对频数，从而表示数据汇总结果的图形方法。

组中值： 每组中的组上限和组下限的平均值。

点位图： 一种简单的数据图形汇总方法，用位于横轴上面的点来代表观察值，以显示出数据值的分布情况。

直方图： 一种图形表示方法，它通过把横轴设为组间距，纵轴设为频数，从而构建出定量数据的频数分布、相对频数或百分比频数分布。

累积频数分布： 对定量数据进行的表格汇总，表示数值小于或等于每组上限的数据项个数。

累积相对频数分布： 对定量数据进行的表格汇总，表示数值小于或等于每组上限的数据项占总数的比例。

累积百分比频数分布： 对定量数据进行的表格汇总，表示数值小于或等于每组上限的数据个数所占的百分比。

累积曲线： 一种显示累积分布的图形。

探索性数据分析： 使用简单的算术和易画的图形来快速汇总数据的一种方法。

茎叶显示： 能够在同时既对定量数据排序又给出其分布形状的一种探索性数据分析技术。

交叉分组列表： 对两个变量的数据进行的表格汇总。其中一个变量的类别用行表示，另一变量的类别则用列表示。

散点图： 对两个定量变量间相互关系的图形表述。其中一个变量在横轴表示，另一变量在纵轴表示。

重要公式

相对频数

$$\frac{\text{组内频数}}{n} \tag{2.1}$$

近似组宽

$$\frac{\text{数据最大值} - \text{数据最小值}}{\text{组数}} \tag{2.2}$$

补充练习

36. *Autodata* 和《今日美国》的研究提供了 2000 年 3 月最畅销的汽车车型数据（*USA Today*, April 4, 2000）。其中在这里列出的有：Chevrolet Silverado/C/K 小货车、Ford F 系列小货车、Ford Taurus、Honda Accord 以及 Toyota Camry。数据来源是表 2.16 中的 50 次汽车购买样本。

表 2.16　50 次汽车购买的数据



CD 光盘数据 Vehicles

Silverado	Taurus	Accord	F 系列	Silverado
F 系列	Accord	Silverado	Camry	Taurus
Silverado	F 系列	Camry	Silverado	F 系列
Taurus	F 系列	Taurus	F 系列	Camry
Taurus	F 系列	F 系列	Accord	Camry
Silverado	Silverado	Silverado	Silverado	Taurus
Camry	Silverado	Accord	F 系列	F 系列
Taurus	F 系列	Accord	Camry	Accord
F 系列	Silverado	F 系列	F 系列	Taurus
Camry	Silverado	F 系列	F 系列	F 系列

a. 构建频数和百分比频数分布。

b. 最畅销的两款汽车是什么？

c. 作出饼形图。

37.《财富》1 000 强公司的每一家都可被归属于几种行业之一（*Fortune*, April 17, 2000）。以下的样本包括 20 家公司以及对应的行业分类。

公司	行业	公司	行业
IBP	食品	Borden	食品
Intel	电子	McDonnell Douglas	航天
Coca-Cola	饮料	Morton International	化学
Union Carbide	化学	Quaker Oats	食品
General Electric	电子	Pepsico	饮料
Motorola	电子	Maytag	电子
Kellogg	食品	Textron	航天
Dow Chemical	化学	Sara Lee	食品
Campbell Soup	食品	Harris	电子
Ralston Purina	食品	Eaton	电子

a. 作出显示每个行业公司个数的频数分布。

b. 作出百分比频数分布。

c. 作出数据的柱状图。

38. 1998 年和 1999 年期间，《时代》周刊采用了下列电影作为封面故事：*Blair Witch Project*, *Phantom Menace*, *Beloved*, *Primary Colors* 和 *Truman Show*。《时代》对收到的封面故事来信进行统计，关于这些电影的来信数目分别是 159，89，85，57 和 51 封（*Time*, September 27, 1999）。

a. 构建信件数据的百分比频数分布。

b. 作出数据的饼形图。

c. *Blair Witch Project* 和 *Phantom Menace* 的封面故事刊于 1999 年，其他的刊于 1998 年。关于 1999 年封面故事的信件占多大百分比？

39. 表 2.17 中的数据表示 1997 年 3 个季度内，健康护理服务行业 20 家公司的以百万美元为单位的销售额（*Business Week*, November 17, 1997）。

a. 构建频数分布汇总数据，组宽为 500。

b. 构建数据的相对频数分布。

c. 构建数据的累积频数分布。

d. 构建数据的累积相对频数分布。

e. 构建直方图以表示数据。

表 2.17 健康护理服务行业的公司 1997 年 3 个季度的销售额

Beverly Ent.	805	Novacare	357
Coventry	307	Phycor	284
Express Scripts	320	Quest Diag.	374
Healthsouth	748	Quorum Health	393
Horizon	445	Sun Healthcare	486
Humana	1 968	Tenet Healthcare	2 331
Int. Health	472	U. Wisconsin	389
Lab. Corp. Am.	377	Univ. Health	362
Manor Care	274	Vencor	845
Medpartners	1 614	Wellpoint	1 512

资料来源：*Business Week*, November 17, 1997。

40. 40 只普通股股票的收盘价如下（*The Wall Street Journal*, March 17, 2000）：


CD 光盘数据
Comstock

29 5/8	34	43 1/4	8 3/4	37 7/8	8 5/8	7 5/8	30 3/8	35 1/4	19 3/8
9 1/4	16 1/2	38	53 3/8	16 5/8	1 1/4	48 3/8	18	9 3/8	9 1/4
10	37	18	8	28 1/2	24 1/4	21 5/8	18 1/2	33 5/8	31 1/8
32 1/4	29 5/8	79 3/8	11 3/8	38 7/8	11 1/2	52	14	9	33 1/2

a. 构建数据的频数和相对频数分布。

b. 构建数据的累积频数和累积相对频数分布。

c. 作出数据的直方图。

d. 使用你的汇总结果，对普通股股票的价格进行评论和观察。

41. 美国个人投资者协会对 94 只新的影子股票进行了报道（*AAII Journal*, February 1997）。术语“影子”是指那些不被主要的经纪机构密切跟踪的中小型企业股票。有关股票交易地点——纽约股票交易所、美国股票交易所以及场外交易，每股收益和市盈率的信息由下列 20 只影子股票的样本给出：


CD 光盘数据
Shadow

股票	交易地点	每股收益(美元)	市盈率
Chemi-Trol	OTC	0.39	27.30
Candie's	OTC	0.07	36.20
TST/Impreso	OTC	0.65	12.70
Unimed Pharm.	OTC	0.12	59.30
Skyline Chili	AMEX	0.34	19.30
Cyanotech	OTC	0.22	29.30
Catalina Light.	NYSE	0.15	33.20
DDL Elect.	NYSE	0.10	10.20
Euphonix	OTC	0.09	49.70

股票	交易地点	每股收益(美元)	市盈率
Mesa Labs	OTC	0.37	14.40
RCM Tech.	OTC	0.47	18.60
Anuhco	AMEX	0.70	11.40
Hello Direct	OTC	0.23	21.10
Hilite Industries	OTC	0.61	7.80
Alpha Tech.	OTC	0.11	34.60
Wegener Group	OTC	0.16	24.50
U. S. Home & Garden	OTC	0.24	8.70
Chalone Wine	OTC	0.27	44.40
Eng. Support Sys.	OTC	0.89	16.70
Int. Remote Imaging	AMEX	0.86	4.70

a. 构建交易地点数据的频数和相对频数分布。在哪里交易的影子股票被列出的最多？

b. 构建每股收益和市盈率数据的频数和相对频数分布。对每股收益数据使用组限为 0.00—0.19，0.20—0.39，等等；对市盈率数据使用组限为 0.0—9.9，10.0—19.9，等等。你对影子股票的观察结果和评论是什么？

42. 1998 年各州的个人收入如下：


CD 光盘数据
Income

Ala.	21 500	Ga.	25 106	Md.	30 023
Alaska	25 771	Hawaii	26 210	Mass.	32 902
Ariz.	23 152	Idaho	21 080	Mich.	25 979
Ark.	20 393	Ill.	28 976	Minn.	27 667
Calif.	27 579	Ind.	24 302	Miss.	18 998
Colo.	28 821	Iowa	24 007	Mo.	24 447
Conn.	37 700	Kan.	25 049	Mont.	20 247
Del.	29 932	Ky.	21 551	Neb.	24 786
D. C.	37 325	La.	21 385	Nev.	27 360
Fla.	25 922	Maine	23 002	N. H.	29 219
N. J.	33 953	Ore.	24 775	Utah	21 096
N. M.	20 008	Penn.	26 889	Vt.	24 217
N. Y.	31 679	R. I.	26 924	Va.	27 489
N. C.	24 122	S. C.	21 387	Wash.	28 066
N. D.	21 708	S. D.	22 201	W. Va.	19 373
Ohio	25 239	Tenn.	23 615	Wis.	25 184
Okla.	21 056	Texas	25 028	Wyo.	23 225

资料来源：*Current Population Survey*, Bureau of Economic Analysis, March 2000。

构建数据的频数分布、相对频数分布和直方图。

43. 经济教育联合会对40个州的调查结论是学生们没有掌握足够的经济学知识，这个发现是基于对11和12年级学生所进行的一项测验结果。该测验包括46个问题，形式是考查诸如利润和供求定理这样的基础经济概念的多项选择题。下表给出了被正确回答的问题数的样本数据：

12	10	16	24	12	14	18	23
31	14	15	19	17	9	19	28
24	16	21	13	20	12	22	18
22	18	30	16	26	18	16	14
8	25	22	15	33	24	17	19

使用下列方法汇总数据：

a. 茎叶显示。

b. 频数分布。

c. 相对频数分布。

d. 累积频数分布。

e. 在这些数据的基础上，你同意学生们没有掌握足够经济学知识的结论吗？请解释。

44. 20个城市的每日高温和低温如下：



CD光盘数据
HighLow

城市	高温	低温	城市	高温	低温
雅典	75	54	墨尔本	66	50
曼谷	92	74	蒙特利尔	64	52
开罗	84	57	巴黎	77	55
哥本哈根	64	39	里约热内卢	80	61
都柏林	64	46	罗马	81	54
哈瓦那	86	68	汉城	64	50
香港	81	72	新加坡	90	75
约翰内斯堡	61	50	悉尼	68	55
伦敦	73	48	东京	79	59
马尼拉	93	75	温哥华	57	43

资料来源：*USA Today*, May 9, 2000。

a. 作出高温的茎叶显示。

b. 作出低温的茎叶显示。

c. 比较(a)和(b)的茎叶显示，并对每日高和低气温间的差异进行评论。

d. 使用(a)中的茎叶显示来确定高温为80°或80°以上的城市数目。

e. 作出高温和低温的气温频数分布。

45. 参考上题中20个城市的高温和低温的气温数据集。

a. 作出散点图以显示在高温和低温这两个变量之间的关系。

b. 评论高温和低温气温间的关系。

46. 有人对四种职业进行了工作满意度研究。工作满意度是通过一个包括 18 个问题的问卷调查表来测量的，每个问题对应着 1 到 5 分。18 个问题的得分总和就是样本中每个人的工作满意度分数。数据已给出如下，更高的分数代表更大的满意度。


CD 光盘数据
OccupSat

职业	满意度分数	职业	满意度分数	职业	满意度分数
律师	42	理疗师	78	系统分析师	60
理疗师	86	系统分析师	44	理疗师	59
律师	42	系统分析师	71	木工	78
系统分析师	55	律师	50	理疗师	60
律师	38	律师	48	理疗师	50
木工	79	木工	69	木工	79
律师	44	理疗师	80	系统分析师	62
系统分析师	41	系统分析师	64	律师	45
理疗师	55	理疗师	55	木工	84
系统分析师	66	木工	64	理疗师	62
律师	53	木工	59	系统分析师	73
木工	65	木工	54	木工	60
律师	74	系统分析师	76	律师	64
理疗师	52				

a. 作出职业和满意度分数的交叉分组列表。

b. 计算 a 中交叉分组列表的行百分比。

c. 关于这些职业的满意度水平，你能得出什么观察结果？

47. 是否更大的公司产生更多的收入？下列数据给出了样本的雇员人数和年收入，样本中包括了 20 家《财富》1 000 强公司（*Fortune*, April 17, 2000）。


CD 光盘数据
RevEmps

公司	雇员人数	收入（百万美元）	公司	雇员人数	收入（百万美元）
Sprint	77 600	19 930	American Financial	9 400	3 334
Chase Manhattan	74 801	33 710	Fluor	53 561	12 417
Computer Sciences	50 000	7 660	Phillips Petroleum	15 900	13 852
Wells Fargo	89 355	21 795	Cardinal Health	36 000	25 034
Sunbeam	12 200	2 398	Borders Group	23 500	2 999
CBS	29 000	7 510	MCI Worldcom	77 000	37 120
Time Warner	69 722	27 333	Consolidated Edison	14 269	7 491
Steelcase	16 200	2 743	IBP	45 000	14 075
Georgia-Pacific	57 000	17 796	Super Value	50 000	17 421
Toro	1 275	4 673	H&R Block	4 200	1 669

a. 作出散点图来显示收入和雇员人数两变量间的关系。

b. 对两变量间明显存在的关系进行评论。

48. 1992年对由辛辛那提煤气与电力公司(Cincinati Gas & Electric Company)提供服务的商业建筑进行了一项调查（CG&E Commercial Building Characteristics Survey, November 25, 1992）。问题之一是使用的主要燃料是什么，之二是商业建筑建设年份。调查结果的部分交叉分组列表如下：

	燃料类型				
建设年份	电力	天然气	石油	丙烷	其他
1973及以前	40	183	12	5	7
1974—1979	24	26	2	2	0
1980—1986	37	38	1	0	6
1987—1991	48	70	2	0	1

a. 通过作出行总计和列总计完成交叉分组列表。

b. 作出建筑年份和燃料类型的频数分布。

c. 作出显示列百分比的交叉分组列表。

49. 一份名为《CD ROM的财富》的档案列在书的封底。表2.18包括这些数据的一部分，它提供了50家《财富》500强公司样本的所有者权益、市场价值以及利润的数据（*Fortune*, April 26, 1999）。

a. 构建所有者权益和利润变量的交叉分组列表。对利润变量分组为0—200，200—400，…，1 000—1 200，对所有者权益分组为0—1 200，1 200—2 400，…，4 800—6 000。

b. 计算(a)中交叉分组列表的行百分比。

c. 你注意到利润和所有者权益之间存在着什么关系？

50. 参考表2.18中的数据集。

a. 构建市场价值和利润变量的交叉分组列表。

b. 计算(a)中交叉分组列表的行百分比。

c. 对两变量间存在的明显关系进行评论。

51. 参考表2.18中的数据集。

a. 作出散点图以显示利润和所有者权益变量间的关系。

b. 对两变量间存在的明显关系进行评论。

52. 参考表2.18中的数据集。

a. 作出散点图以显示市场价值和所有者权益变量间的关系。

b. 对两变量间存在的明显关系进行评论。

表 2.18　一个包括 50 家《财富》500 强公司的样本数据


CD 光盘数据
Fortune

公司	所有者权益（千美元）	市场价值（千美元）	利润（千美元）
AGCO	982.1	372.1	60.6
AMP	2 698.0	12 017.6	2.0
Apple Computer	1 642.0	4 605.0	309.0
Baxter International	2 839.0	21 743.0	315.0
Bergen Brunswick	629.1	2 787.5	3.1
Best Buy	557.7	10 376.5	94.5
Charles Schwab	1 429.0	35 340.6	348.5
…	…	…	…
Walgreen	2 849.0	30 324.7	511.0
Westvaco	2 246.4	2 225.6	132.0
Whirlpool	2 001.0	3 729.4	325.0
Xerox	5 544.0	35 603.7	395.0

案例研究　CONSOLIDATED 食品公司

Consolidated 食品公司在新墨西哥州、亚利桑那州和加利福尼亚州经营连锁超市。它举办了一个促销活动，宣传本公司新的信用卡政策，以使顾客在购物付款时除了使用现金和个人支票外，还能够选择使用诸如维萨卡和万事达卡这样的信用卡。新政策正在试验的基础上实行，公司希望信用卡支付方式能够鼓励顾客多消费。

在新政策实行一个月以后，工作人员选择了一周时间内的 100 位顾客作为随机样本，搜集了每位顾客的支付方式和消费金额数据。表 2.19 列出了样本数据的一部分。在实行新政策以前，该公司的顾客中大约有 50% 用现金付款，而用个人支票付款的也占大约 50%。

管理报告

使用描述统计的表格和图形方法来汇总表 2.19 中的样本数据。你的报告应包括下列形式的数据汇总：

1. 支付方式的频数和相对频数分布。
2. 支付方式的柱状图或饼形图。
3. 每种支付方式下消费金额的频数和相对频数分布。
4. 每种支付方式下消费金额的直方图和/或茎叶显示图。
5. 支付方式和消费金额变量的交叉分组列表。

表 2.19　100 位顾客样本的消费金额和支付方式


CD 光盘数据
Consolid

顾客	消费金额（美元）	支付方式
1	28.58	支票
2	52.04	支票
3	7.41	现金
4	11.17	现金
5	43.79	信用卡
6	48.95	支票
7	57.59	支票
8	27.60	支票
9	26.91	信用卡
10	9.00	现金
…	…	…
95	18.09	现金
96	54.84	支票
97	41.10	支票
98	43.14	支票
99	3.31	现金
100	69.77	信用卡

对于消费金额和支付方式，你能够得出什么样的初步结论？

附录 2.1　在表格和图形方法中应用 Minitab 软件

Minitab 软件在构建数据的表格和图形汇总方面具有广泛的功能。本附录中，我们将演示怎样应用 Minitab 构建几种图形汇总以及交叉分组列表的表格汇总。所演示的图形方法包括点位图、直方图、茎叶显示和散点图。

点位图

我们使用表 2.5 中的审计时间数据来说明。这些数据已经被输入 Minitab 工作表的 C1 列，通过下列步骤能够生成在图 2.3 中显示的点位图：


CD 光盘数据
Audit

步骤 1. 选择 **Graph** 下拉菜单；
步骤 2. 选择 **Dotplot** 选项；
步骤 3. 当 Dotplot 对话框出现时：
在 **Variables** 框中输入 C1；
点击 **OK**。

直方图

我们以表 2.5 中的审计时间数据作为纵轴来演示如何构建频数的直方图。这些数据已经被输入 Minitab 工作表的 C1 列，通过下列步骤将生成在图 2.4 中显示的直方图：



CD 光盘数据
Audit

步骤 1. 选择 **Graph** 下拉菜单；
步骤 2. 选择 **Histogram** 选项；
步骤 3. 当 Histogram 对话框出现时：
在 **Graph variables** 框的第 1 行输入 C1；
在 **Data display** 框的 **Display** 项目下选择 **Bar**，在 **For each** 项目下选择 **Graph;**
选择 **Options**；
步骤 4. 当 Histogram Options 对话框出现时：
在 **Type of Histogram** 项下选择 **Frequency**；
在 **Type of Intervals** 项下选择 **CutPoint**；
在 **Definition of Intervals** 项下选择 **Midpoint/cutpoint positions**，并在框中输入 10:35/5；*
点击 OK。
步骤 5. 当 Histogram 对话框出现时：
点击 **OK**。

茎叶显示



CD 光盘数据
ApTest

我们使用表 2.10 的能力测试数据来演示如何得出茎叶显示。这些数据已经被输入 Minitab 工作表的 C1 列，下列步骤能够生成在 2.3 节显示的扩展茎叶显示：

步骤 1. 选择 **Graph** 下拉菜单；
步骤 2. 选择 **Stem-and-Leaf** 选项；
步骤 3. 当 Stem-and-Leaf 对话框出现时：
在 **Variables** 框中输入 C1；
点击 **OK**。

散点图

我们使用表 2.14 的音像设备商店数据来演示散点图的构建。把星期数从 1 到 10 输入到 C1 列，

---

* 输入的 10:35/5 表示 10 是直方图的最小值，而 35 是最大值，组宽为 5。


CD 光盘数据
Stereo

把广告次数的数据输入 C2 列，销售额数据则输入 C3 列。通过下列步骤将生成图 2.6 中的散点图：

步骤 1. 选择 **Graph** 下拉菜单；

步骤 2. 选择 **Plot** 选项；

步骤 3. 当 Plot 对话框出现时：

在 **Graph variables** 部分 **Y** 项下输入 C3，**X** 项下输入 C2；

在 **Data display** 部分 **Display** 项下选择 **Symbol**，**For each** 项下选择 **Point**；

点击 **OK**。

交叉分组列表

我们使用来自“Zagat's 饭店评论”的数据来演示交叉分组列表的构建，数据的一部分见表 2.11。


CD 光盘数据
Restaurant

在 Minitab 工作表的 C1 列，饭店按 1 到 300 编号；质量等级输入 C2 列；食品价格输入 C3 列。

Minitab 只能创建定性变量的交叉分组列表，而食品价格是定量变量。因此，我们首先通过规定每个食品价格所归属的类，来对食品价格数据编码。下列步骤可以对食品价格数据编码，在 C4 列中产生四个类：10—19 美元，20—29 美元，30—39 美元以及 40—49 美元。

步骤 1. 选择 **Manip** 下拉菜单；

步骤 2. 选择 **Code** 选项；

步骤 3. 选择 **Numeric to Text** 选项；

步骤 4. 当 Code-Numeric to Text 对话框出现时：

在 **Code data from columns** 框中输入 C3；

在 **Into columns** 框中输入 C4；

在第一个 **Original values** 框中输入 10: 19，在邻近的 **New** 框中输入 10 - 19；

在第二个 **Original values** 框中输入 20: 29，在邻近的 **New** 框中输入 20 - 29；

在第三个 **Original values** 框中输入 30: 39，在邻近的 **New** 框中输入 30 - 39；

在第四个 **Original values** 框中输入 40: 49，在邻近的 **New** 框中输入 40 - 49；

点击 **OK**。

对于 C3 列的每个食品价格，对应的食品价格类别会出现在第 4 列中。我们现在能够使用第 2 列和第 4 列的数据来构建质量等级和食品价格类别的交叉分组列表。下列步骤将创建出与表 2.12 包含相同信息的交叉分组列表：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Tables** 选项；
步骤 3. 选择 **Cross Tabulation** 选项；
步骤 4. 当 Cross Tabulation 对话框出现时：
在 **Classification variables** 框中输入 C2 C4；
选择 **Counts**；
点击 **OK**。

附录 2.2　在表格和图形方法中应用 Excel 软件

Excel 软件在构建数据的表格和图形汇总时具有广泛的功能，其中三个功能最强大的工具是函数向导、图表向导和关键表报告。

函数和函数向导

Excel 提供了对统计分析有用的各种函数。如果你知道需要什么函数和如何使用的话，那么只需把函数直接输入 Excel 工作表的单元格中即可。如果你不知道的话，Excel 提供了函数向导来帮助你识别可用的函数并使用它们。

函数向导。为得到函数向导，点击标准工具条的 f_x，或者选择 **Insert** 下拉菜单并选择 f_x 函数，然后 **Paste Function** 对话框将会出现(见图 2.9)。在 **Function category** 框显示 Excel 中函数类别的列表；在图 2.9 中，我们已经选择了 **Statistical**。当 **Statistical** 选项高亮显示时，在 **Function name** 框中将显示出一个包括了所有可用统计函数的列表。这里，我们已经高亮显示了 COUNTIF

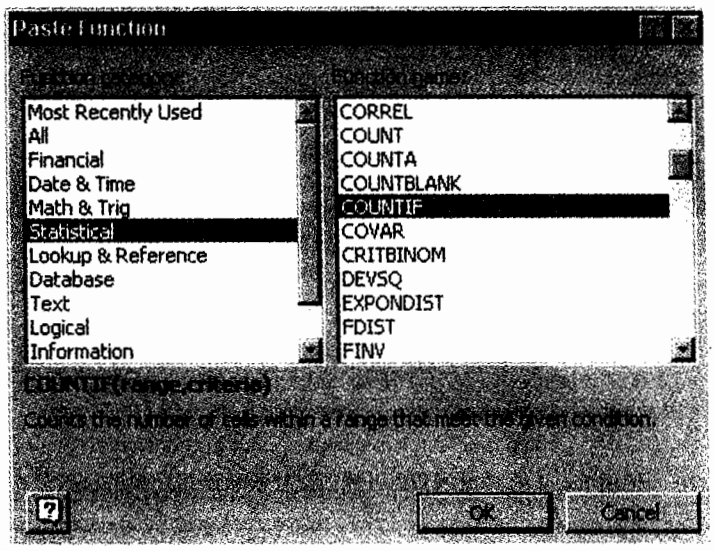


图 2.9　Excel 的 Paste Function 对话框

函数。而一旦某函数被高亮显示，它的正确形式以及简要介绍就会显示在 **Function category** 和 **Function name** 框下面。为了得到正确使用函数的帮助，点击 **OK**。

频数分布。我们来演示怎样利用 COUNTIF 函数来构建表 2.1 的软饮料购买次数数据的频数分布。首先参考图 2.10，公式工作表（显示使用的函数和公式）出现在底面，而数值工作表（显示用函数和公式得到的结果）则出现在表面。


CD 光盘数据
SoftDrink

把标志“购买的品牌”和 50 次软饮料购买的数据输入单元格 A1: A51，把一个标志和软饮料的品名输入单元格 C1: C6。Excel 的 COUNTIF 函数能够把显示在单元格 A2: A51 中的每种软饮料的购买次数加总起来。使用下列步骤就可以利用函数向导生成图 2.10 表面的频数分布：

步骤 1. 选择单元格 **D2**，得到函数向导，再从统计函数列表中选择 **COUNTIF**；

步骤 2. 点击 **OK**；

步骤 3. 当 COUNTIF 对话框出现时：

在 **Range** 框中输入 $A$2: $A$51；

在 **Criteria** 框中输入 C2；

点击 **OK**；

步骤 4. 把单元格 D2 复制到单元格 D3: D6 中。

Brand Purchased
Coke Classic
Diet Coke
Pepsi-Cola
Diet Coke
Coke Classic
Coke Classic
Dr. Pepper
Diet Coke
Pepsi-Cola
Pepsi-Cola
Pepsi-Cola
Pepsi-Cola
Coke Classic
Dr. Pepper
Pepsi-Cola
Sprite

Soft Drink	Frequency
Coke Classic	=COUNTIF($A$2:$A$51,C2)
Diet Coke	=COUNTIF($A$2:$A$51,C3)
Dr. Pepper	=COUNTIF($A$2:$A$51,C4)
Pepsi-Cola	=COUNTIF($A$2:$A$51,C5)
Sprite	=COUNTIF($A$2:$A$51,C6)

Brand Purchased
Coke Classic
Diet Coke
Pepsi-Cola
Diet Coke
Coke Classic
Coke Classic
Dr. Pepper
Diet Coke
Pepsi-Cola
Pepsi-Cola
Pepsi-Cola
Pepsi-Cola
Coke Classic
Dr. Pepper
Pepsi-Cola
Sprite

Soft Drink	Frequency
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi-Cola	13
Sprite	5

图 2.10　使用 Excel 的 COUNTIF 函数创建软饮料购买次数的频数分布

行 11—14 是隐藏的。

图 2.10 中底面的公式工作表表示应用这些步骤插入的单元格公式；图 2.10 中表面的数值工作表显示使用这些单元格公式计算出的数值。我们看到 Excel 工作表显示出了与表 2.2 相同的频数分布。

如果你对 COUNTIF 函数熟悉，并且不需要函数向导帮助的话，你能够直接向单元格 D2: D6 输入公式。例如，为加总 Coke Classic 出现的次数，我们把下面的公式输入单元格 D2:

$$=\text{COUNTIF}(\$A\$2{:}\$A\$51,\ C2)$$

在加总其他软饮料出现的次数时，只需把公式复制到单元格 D3: D6 中即可。

在以后的章节附录中，还要演示更多的 Excel 函数。我们将视函数的复杂程度，决定是直接把它输入适当的单元格还是利用函数向导。

图表向导

Excel 的图表向导在构建数据的图形表达方面具有广泛的功能。使用该工具，我们将能够超出仅仅利用函数和公式所做的工作范围。我们将演示如何应用它来构建柱状图、直方图和散点图。

柱状图和直方图。在此我们说明怎样利用图表向导来作出柱状图和直方图。让我们从构建软饮料数据的柱状图开始，首先我们要作出图 2.10 的频数分布，因为所要构建的图表是那个工作表的一个扩展。在我们描述要做的工作时参考图 2.11，图 2.10 的数值工作表在底面，而利用图表向导作出的图在表面。



CD 光盘数据
SoftDrink

下列步骤描述了怎样利用图表向导，从而在单元格 C1: D6 所显示的频数分布的基础上构建软饮料数据的柱状图。

步骤 1. 选择单元格 C1: D6;

步骤 2. 在标准工具条上选择 **Chart Wizard** 键(或者选择 **Insert** 下拉菜单，再选择 **Chart** 选项)；

步骤 3. 当 Chart Wizard – Step 1 of 4 – Chart Type 对话框出现时：

在 **Chart type** 列表中选择 **Column**；

从 **Chart sub-type** 显示中选择 **Clustered Column**；

选择 **Next**；

步骤 4. 当 Chart Wizard – Step 2 of 4 – Chart Source Data 对话框出现时：

选择 **Next>**；

步骤 5. 当 Chart Wizard – Step 3 of 4 – Chart Options 对话框出现时：

选择 **Titles** 选项条；

在 **Chart title** 框中输入 **Bar Graph of Soft Drink Purchases**；

在 **Category (X) axis** 框中输入 **Soft Drink**；

在 **Values(Y) axis** 框中输入 **Frequency**;

选择 **Legend** 选项条；

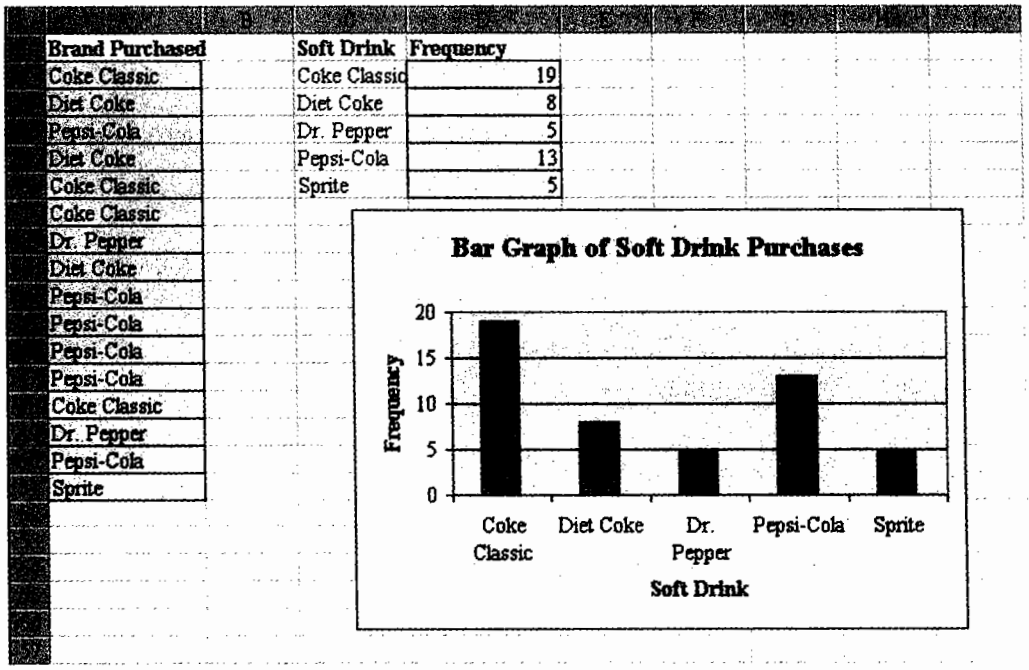


图 2.11　利用 Excel 图表向导作出的软饮料购买次数数据的柱状图

然后在 **Show legend** 框中去掉选择；

选择 **Next>**；

步骤 6. 当 Chart Wizard – Step 4 of 4 – Chart Location 对话框出现时：

指定新图表的位置（我们通过选择 **As object in** 还使用目前的工作表）；

选择 **Finish**。

作出的柱状图(表)显示在图 2.11 中。*

Excel 的图表向导能够以类似的方式生成软饮料数据的饼形图，主要的区别是在步骤 3，我们需要在 Chart type 列表中选择 Pie。

正如我们在 2.2 节末尾的注释中所述，直方图基本上与柱状图相同，只是在柱之间没有间隔。图 2.12 显示出审计时间数据的频数分布在底面，而利用图表向导构建的柱状图（使用与前述相同的步骤)显示在表面。因为在直方图中相邻的条必须接触，我们需要编辑列图表（柱状图）以便消除各柱之间的距离。下列步骤可以完成这个过程：

步骤 1. 右击柱状图的任何一根柱以生成选项列表；

步骤 2. 选择 **Format Data Series**；

步骤 3. 当 Format Data Series 对话框出现时：

选择 **Options** 选项条；

---

* 图 2.11 的柱状图比选择 **Finish** 以后 Excel 提供的图稍微大一些。要重新设置 Excel 图表的大小并不困难。首先，选择图表，则小黑方块——即称为尺寸控制器的东西就会出现在图表的边缘。点击尺寸控制器并拖放它们直到设置为需要的大小。

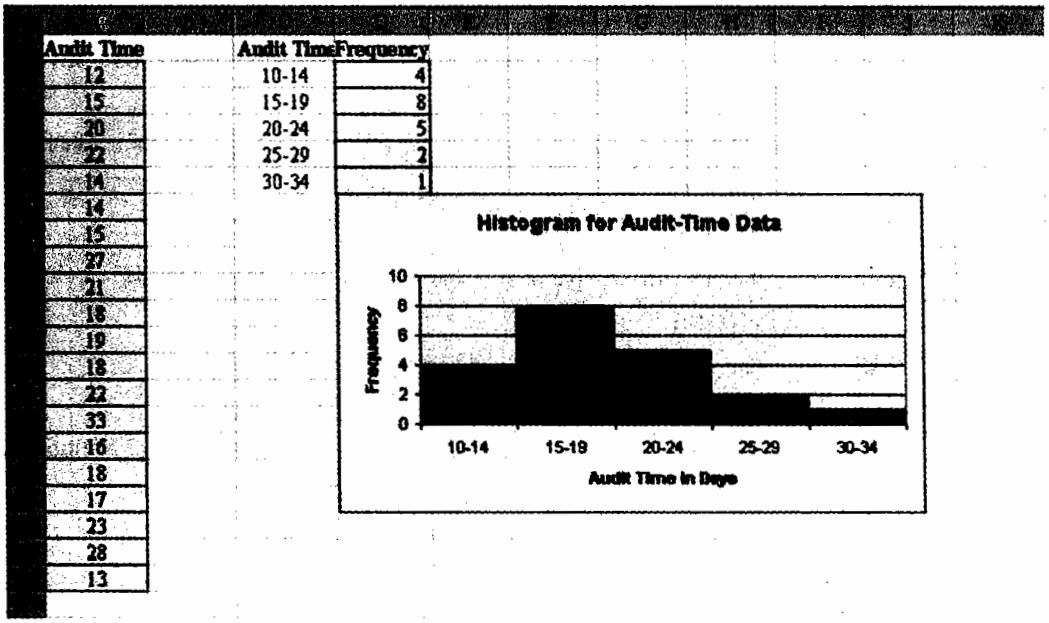


图 2.12　利用 Excel 作出的审计时间数据直方图

在 **Gap width** 框中输入 **0**；

点击 **OK**。

散点图。我们使用表 2.14 的音像设备商店数据来演示如何利用 Excel 的图表向导构建散点图。首先参考图 2.13，数据在图的底面而由图表向导生成的散点图显示在图的表面。下列步骤能够生成散点图：

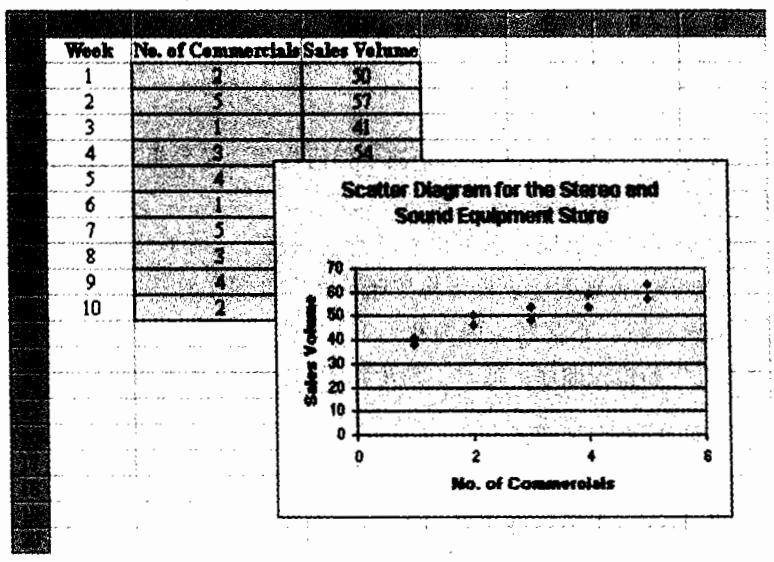


图 2.13　由 Excel 的图表向导生成的音像设备商店数据的散点图


CD 光盘数据
Stereo

步骤 1. 选择单元格 B1: C11;

步骤 2. 在标准工具条上选择 **Chart Wizard** 键（或者选择 **Insert** 下拉菜单并选择 **Chart** 选项）;

步骤 3. 当 Chart Wizard – Step 1 of 4 – Chart Type 对话框出现时：

在 **Chart type:** 显示中选择 **XY(Scatter)**;

选择 **Next>**;

步骤 4. 当 Chart Wizard – Step 2 of 4 – Chart Source Data 对话框出现时：

选择 **Next>**;

步骤 5. 当 Chart Wizard – Step 3 of 4 – Chart Options 对话框出现时：

选择 **Titles** 选项条;

在 **Chart title** 框中输入 **Scatter Diagram for the Stereo and Sound Equipment Store**;

在 **Value (X) axis** 框中输入 **No. of Commercials**;

在 **Value (Y) axis** 框中输入 **Sales Volume**;

选择 **Legend** 选项条;

去掉 **Show legend** 框的选择;

选择 **Next>**;

步骤 6. 当 Chart Wizard – Step 4 of 4 – Chart Location 对话框出现时：

指定新图表的位置(我们通过选择 **As object in** 还使用目前的工作表);

选择 **Finish**。

关键表报告

Excel 的关键表报告提供了管理超过一个变量的数据集的有效工具，我们通过构建交叉分组列表来说明它的应用。

交叉分组列表。我们使用图 2.14 的饭店数据来演示如何构建交叉分组列表。我们把标志输入第 1 行，把 300 家饭店中每一家的数据输入单元格 A2: C301。

表 2.12 的交叉分组列表在质量等级标志下面有三行，对应于三个质量等级类别：好、非常好、优异。除非我们另有规定，否则关键表报告将按照英文字母顺序对标志排序，这导致了质量等级的顺序被排列为：优异、好、非常好。因为我们需要的质量等级顺序为：好、非常好、优异，我们必须改变关键表报告的错误顺序。通过下列步骤我们能够做到这一点：


CD 光盘数据
Restaurant

步骤 1. 选择 **Tools** 下拉菜单;

步骤 2. 选择 **Options**;

步骤 3. 当 Options 对话框出现时：

选择 **Custom lists** 选项条（见图 2.15）;

在 **List entries** 框中，输入类型 **Good** 并按 Enter 键；输入类型 **Very Good**



CD 光盘数据
Restaurant

Restaurant	Quality Rating	Meal Price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
6	Good	28
7	Very Good	19
8	Very Good	11
9	Very Good	23
10	Good	13
291	Very Good	23
292	Very Good	24
293	Excellent	45
294	Good	14
295	Good	18
296	Good	17
297	Good	16
298	Good	15
299	Very Good	38
300	Very Good	31

图 2.14　包含饭店数据的 Excel 工作表

注意：隐藏了行 12 — 291。

并按 Enter 键；

输入类型 **Excellent**；

选择 **Add**；

点击 **OK**。

现在我们已经为使用关键表报告构建质量等级和食品价格数据的交叉分组列表做好了准备。以图 2.14 的工作表为开始，下列步骤是必需的：

步骤 1. 选择 **Data** 下拉菜单；

步骤 2. 选择 **PivotTable and PivotChart** 报告；

步骤 3. 当 PivotTable and PivotChart Wizard – Step 1 of 3 对话框出现时：

选择 **Microsoft Excel list or database**；

选择 **PivotTable**；

选择 **Next>**；

步骤 4. 当 PivotTable and PivotChart Wizard – Step 2 of 3 对话框出现时：

在 **Range** 框中输入 A1: C301；

选择 **Next>**；

步骤 5. 当 PivotTable and PivotChart Wizard – Step 3 of 3 对话框出现时：

选择 **New Worksheet**；

点击 **Layout** 键；

当 PivotTable and PivotChart Wizard – Layout 图出现时（见图 2.16）：

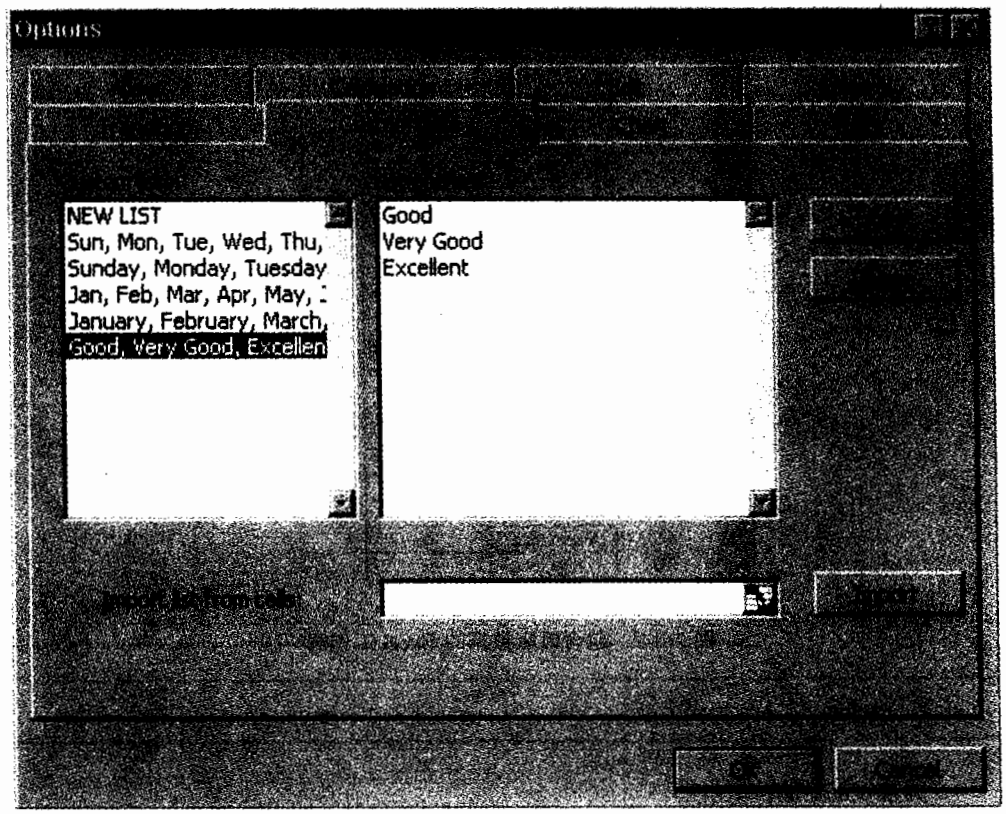


图 2.15　改变 Excel PivotTable 中种类顺序的对话框

把 **QualityRating** 键拖到图的 **ROW** 区域；
把 **Meal Price ($)**键拖到图的 **COLUMN** 区域；
把 **Restaurant** 键拖到图的 **DATA** 区域；
双击 DATA 区域的 **Sum of Restaurant** 键；
当 PivotTable Field 对话框出现时：
在 **Summarize by** 项下选择 **Count**；
点击 **OK**（图 2.17 显示了完成的 Layout 图）；
点击 **OK**；
当 PivotTable and PivotChart Wizard – Step 3 of 3 对话框再次出现时：
选择 **Finish>**。

由 Excel 生成的输出结果的一部分显示在图 2.18 中。我们注意到从 D 列到 AK 列的输出已被隐藏以使结果能够出现在合理尺寸的图形内。图 2.18 的行标志（Good, Very Good, and Excellent）和行总计（84, 150, 66, 300）与表 2.12 的行标志和行总计是相同的，但是，图 2.18 中的列被指定为食品价格的每个可能值。例如，B 列包含具有 10 美元食品价格的饭店加总数，C 列则包含具有 11 美元食品价格的饭店加总数，依此类推。为了以一种与表 2.12 相类似的方式观察关键表报告，我们必须把所有列合并为四个价格类别：10—19 美元, 20—29 美元, 30—39 美元, 以及 40—49

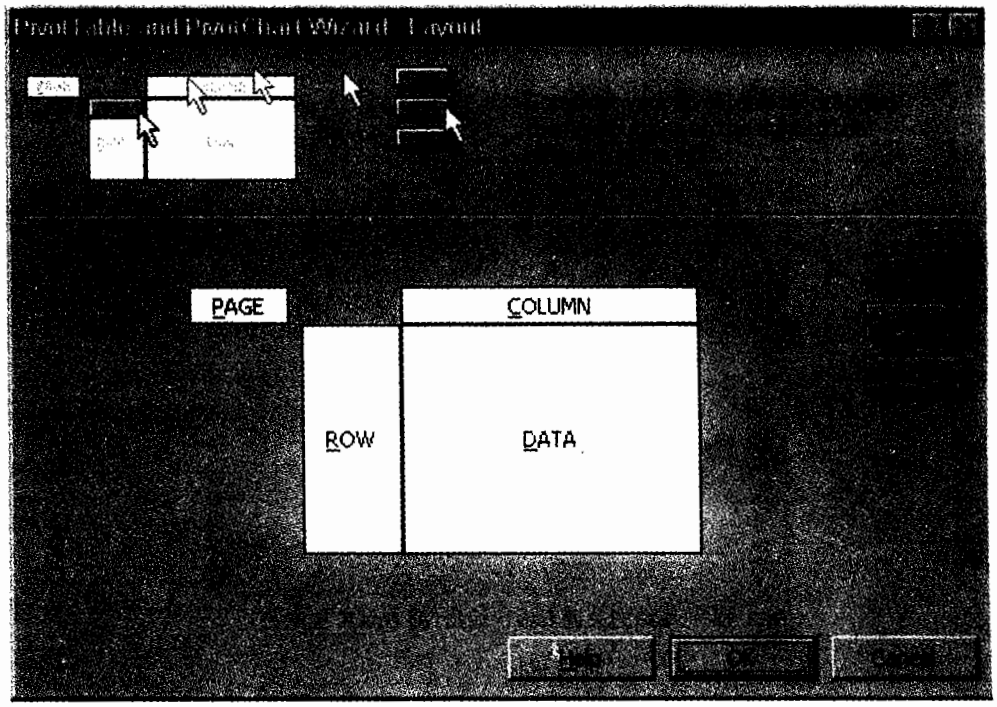


图 2.16　**PivotTable and PivotChart Wizard – Layout 图**

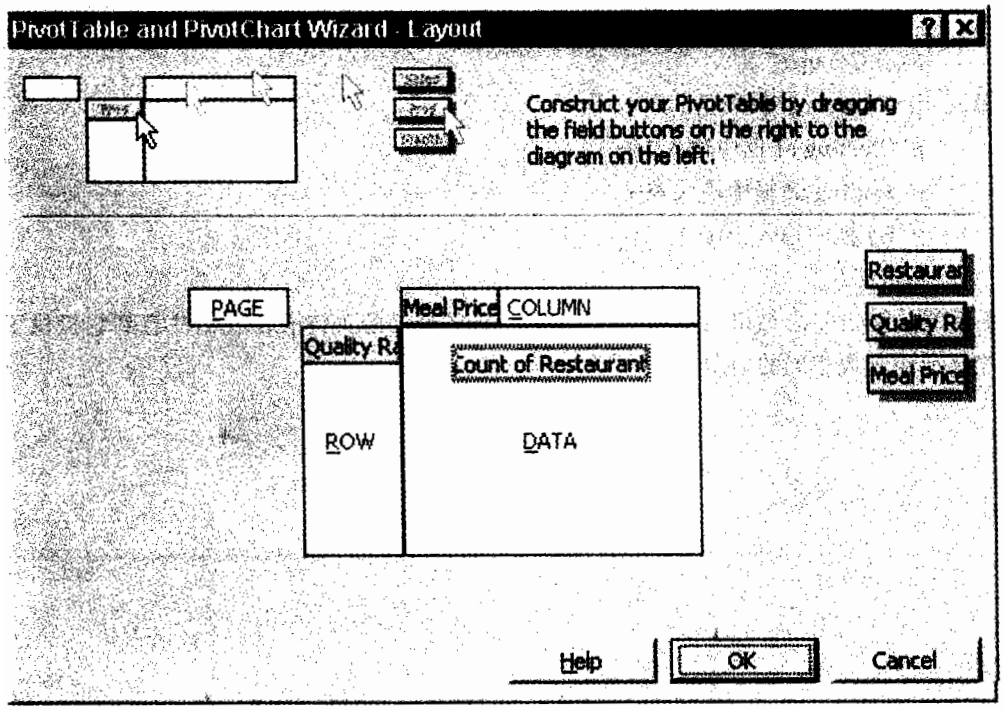
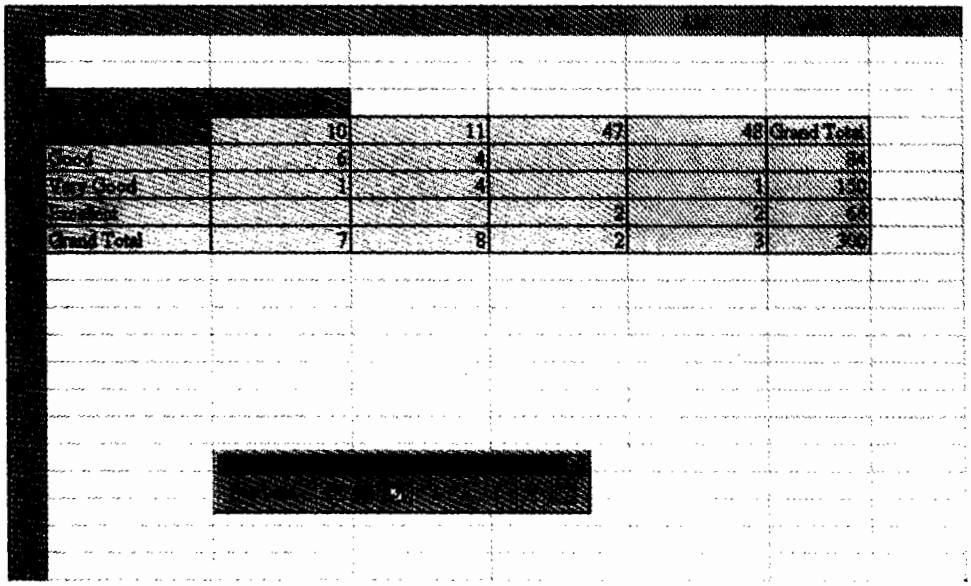


图 2.17　**完成的 Layout 图**



	10	11	47	48	Grand Total
Good	6	4			90
Very Good	1	4		1	150
[illegible]			2	2	60
Grand Total	7	8	2	3	300

图 2.18　原始的关键表报告输出(列 D:AK 被隐藏)

美元。合并图 2.18 工作表列的必需步骤如下:

步骤 1. 右击 PivotTable 的单元格 B3 内的 Meal Price ($);

步骤 2. 选择 **Group and Outline**;

再选 **Group**;

步骤 3. 当 **Grouping** 对话框出现时:

在 **Starting at** 框中输入 10;

在 **Ending at** 框中输入 49;

在 **By** 框中输入 10;

点击 **OK**。

修正后的关键表输出结果显示在图 2.19 中。它是最终的关键表,我们注意到它提供了与表 2.12 的交叉分组列表相同的信息。

	10-19	20-29	30-39	40-49	Grand Total
Good	42	40	2		84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Grand Total	78	118	76	28	300

图 2.19　最终的饭店数据关键表报告

3 描述性统计:数值方法

统计实例:SMALL FRY DESIGN 公司

3.1 位置的度量

均值

中位数

众数

百分位数

四分位数

3.2 变异程度的度量

极差

四分位点内距

方差

标准差

变异系数

3.3 相对位置的度量和异常值的检测

z-分数

切比雪夫定理

经验法则

异常值检测

3.4 探索性数据分析

五数概括法

箱形图

3.5 两变量间关系的度量

协方差

协方差的解释

相关系数

相关系数的解释

3.6 加权平均值和使用分组数据

加权平均值

分组数据

统计实例

SMALL FRY DESIGN* 公司

圣安娜，加利福尼亚州

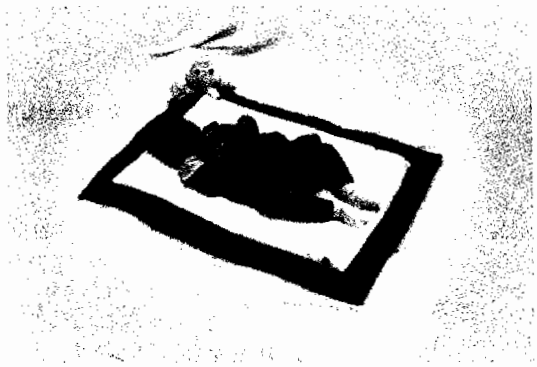
Small Fry Design 公司成立于 1997 年，是一家设计和进口婴儿产品的公司，主要经营玩具和附属用品。公司的产品包括泰德熊、玩具汽车、音乐玩具、拨浪鼓和安全床垫等，而公司的特长是设计强调颜色、材质和声音高质量的柔软玩具。公司的产品在美国设计，而在中国生产。

Small Fry Design 公司使用独立的销售代表把产品销售给婴儿用品零售商、儿童用品和服装商店、礼品店、大规模百货商场和主要的邮购公司。目前，公司的产品遍布美国的 1 000 多个零售终端。

在这个年轻公司的日常运营中，现金流量管理是最重要的经营活动之一。是否能够保证公司拥有足够的现金收入以满足目前和未来的偿债义务，决定着公司业务的成败。现金流量管理的一个关键因素是对应收账款的分析和控制。通过度量未付款发票的平均账龄和价值，经理能够预测可用现金和监视应收账款状态的变化。公司设置了下述目标：未付款发票的平均账龄不应超过 45 天，超过 60 天的未付款发票的价值不应超过所有应收账款价值的 5%。

在最近对应收账款的总结中，使用了下列描述性统计量来衡量未付款发票的账龄：

均值	40 天
中位数	35 天
众数	31 天



Small Fry Design 公司的一些产品。© Photo Courtesy of Small Fry Design.

对这些统计量的解释表明，一张发票的均值或者平均账龄是 40 天；而中位数显示一半的发票已经超过 35 天没有付款；31 天的众数表示最高频率的发票账龄，即一张未付款发票的最普通时间长度是 31 天。统计汇总还显示出应收账款总价值中只有 3% 超过了 60 天。基于这些统计信息，管理者可以感到满意，因为应收账款和收入现金流都处于控制之下。

在本章中，你将学习如何计算和理解 Small Fry Design 公司使用过的一些统计度量。除了均值、中位数和众数外，你还将学到其他的描述性统计量，例如：极差、方差、标准差、百分位数和相关系数。这些数值度量将有助于对数据的理解和解释。

* 作者衷心感谢 Small Fry Design 公司中的经理 John A. Mc-Carthy 提供了该案例。

在第 2 章里，我们讨论了用来汇总数据的表格和图形方法。这些方法在撰写报告时很有效，它们在向个人和团体表达时可作为一种直观的工具。在本章里，我们将给出几种描述性统计学的数值方法，它们提供了汇总数据的其他可选方法。

我们从考虑只包含一个变量的数据集开始对问题的讨论。可以使用 n 个数据值来计算位置和离散程度的度量数值，如果数据集含有超过一个的变量，我们能够分别计算每个单独变量的度量

数值。在两变量的例子中，我们将构建对变量之间关系的度量。

本章将介绍位置、离散程度和相关程度的几种数值度量。如果使用样本中的数据来计算度量值，则称这样的度量值为**样本统计量**(sample statistics)；如果它是用总体中的数据计算得出的，则称为**总体参数**(population parameters)。

3.1　位置的度量

均　值

也许位置的数值度量中最重要的就是变量的**均值**(mean)了，或者叫平均值。均值提供了对数据集中心位置的度量。如果数据来自样本，我们把均值记作$\bar{x}$；如果数据来自总体，我们则用希腊字母μ表示它。

在统计学公式中，我们习惯用x_1来表示变量x的第一个观察值，用x_2来表示变量x的第二个观察值，依此类推。作为一般，我们把数值的第i个观察值记作x_i。对于具有n个观察值的样本，计算样本均值的公式如下：

样本均值

$$\bar{x} = \frac{\sum x_i}{n} \tag{3.1}$$

在这个公式里，分子是n个观察值的数值之和，即

$$\sum x_i = x_1 + x_2 + \cdots + x_n$$

希腊字母$\sum$是求和符号。

为了说明样本均值的计算，让我们考虑下列班级人数数据，它来自一个由 5 个大学班级组成的样本。

46　54　42　46　32

我们使用符号x_1，x_2，x_3，x_4和x_5分别表示这 5 个班的学生人数。

$$x_1 = 46 \qquad x_2 = 54 \qquad x_3 = 42 \qquad x_4 = 46 \qquad x_5 = 32$$

因此，计算样本均值，我们有

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

班级人数的样本均值是 44 人。

计算样本均值的另一个例子如下：假设大学设置办公室向一个商学院毕业生样本发放问卷调查表，调查有关起始薪金的信息。表 3.1 显示了收集到的数据，则 12 个商学院毕业生样本的平均

月起薪可计算如下：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{x_1+x_2+\cdots+x_{12}}{12}$$
$$=\frac{2\,850+2\,950+\cdots+2\,880}{12}$$
$$=\frac{35\,280}{12}=2\,940$$

表 3.1　12 个商学院毕业生样本的月起薪


CD 光盘数据
Salaries

毕业生	月起薪（美元）	毕业生	月起薪（美元）
1	2 850	7	2 890
2	2 950	8	3 130
3	3 050	9	2 940
4	2 880	10	3 325
5	2 755	11	2 920
6	2 710	12	2 880

公式(3.1)说明了具有 n 个观察值的样本均值是怎样计算的，而计算总体均值的公式也基本相同，只是我们使用了与前面不同的符号来表示我们正在处理的完整总体。我们把总体观察值的个数记作 N，总体均值的符号记为 μ。

总体均值

$$\mu=\frac{\sum x_i}{N} \tag{3.2}$$

中位数

中位数(median)是对数据中心位置的另一种度量。中位数是当数据按升序排列时，排在中间的数据值。当有奇数个观察值时，中位数就是中间的数据值；而当有偶数个观察值时，因为不存在中间数据，所以在这种情况下，我们遵循惯例，定义中间两个观察值的平均数为中位数。为方便起见，我们重新表述中位数的定义如下：

中位数作为位置的度量，往往用在年度收入和资产价值数据的报告中，这是因为个别异常大的收入或资产价值能够使均值膨胀。在这样的例子中，中位数就是对于中心位置更好的度量。

中位数

把数据按升序排列（从最小值到最大值）。

(a)有奇数个观察值时，中位数就是中间数据值。

(b)有偶数个观察值时，中位数是中间两个数据值的平均值。

现在，让我们应用该定义来计算 5 个大学班级样本的班级人数中位数。将数据按升序排列如下：

32　42　46　46　54

因为 $n = 5$ 是奇数，中位数就是中间数据值。因此班级人数中位数是 46 人。即使该数据集中有两个值等于 46，当我们按升序排列数据时，也应该各自独立地处理每个观察值。

我们再来计算商学院毕业生的起薪中位数。首先，按升序排列表 3.1 的数据如下：

2 710　2 755　2 850　2 880　2 880　2 890　2 920　2 940　2 950　3 050　3 130　3 325

中间两个值

因为 $n = 12$ 为偶数，我们找到中间两个值：2 890 和 2 920。中位数是它们的平均数。

$$\text{中位数} = \frac{2\,890 + 2\,920}{2} = 2\,905$$

虽然我们在表示中心位置的度量时，更经常地使用均值，但在某些情况下，用中位数更合适。均值会受到异常大和异常小的数据值影响。例如，假设一位毕业生（见表 3.1）的起薪为每月 10 000 美元(也许他的家庭拥有这家公司)。这样我们把表 3.1 的最高起薪数据从 3 325 美元改变为 10 000 美元，再重新计算均值的话，样本均值也将从 2 940 美元变到 3 496 美元。但是，中位数仍将保持不变，为 2 905 美元，因为 2 890 美元和 2 920 美元还是中间的两个值不变。当存在异常高的起薪的情况下，中位数提供了比均值更好的中心位置度量。我们通常说，当数据集含有异常值时，中位数往往更适合于度量数据的中心位置。

众　数

第三种位置度量是**众数**(mode)。众数的定义如下：

> **众数**
> 众数是数据集中出现频率最高的数据值。

为了说明如何得到众数，让我们考虑一个由 5 个班级人数组成的样本。其中惟一超过一次出现的值是 46，它发生的频率是 2，具有最高的发生频率，因此它就是众数。作为另外一个例子，考虑商学院毕业生的起薪样本。只有 2 880 美元的月起薪出现超过了一次，它拥有最高的频率，因此它是众数。

我们也经常会遇到两个或更多不同数值拥有相同的最大出现频率的情况，这时存在着不止一个的众数。如果数据正好拥有两个众数，我们把数据称为**双峰**的。如果数据的众数超过两个，我们把这样的数据称为**多峰**的。在多峰的情况下，几乎不再报告众数。因为列出三个以上的众数对于描述数据的位置并不是特别有用。

百分位数

百分位数(percentile)提供了这样的信息：数据是怎样在从最小值到最大值的这个区间内分布

的。对于那些没有多个重复数值的数据，第 p 百分位数把数据分割为了两部分：大约有 $p\%$ 的观察值小于第 p 百分位数，而大约 $(100-p)\%$ 的观察值大于或等于第 p 百分位数。

百分位数

第 p 百分位数是满足下列条件的一个值：至少有 $p\%$ 的观察值小于或等于该值，至少有 $(100-p)\%$ 的观察值大于或等于该值。

大学或学院的入学考试成绩往往以百分位数的形式报告出来。例如，假设一个申请者在入学考试的语言部分取得了原始分 54 分，那么这个学生的表现与参加同样考试的其他学生相比怎样人们可能看不清楚，但是，如果原始的 54 分对应着百分位数为 70，我们就能够知道大约有 70% 的学生成绩低于他，而大约 30% 的学生成绩比他高。

下列程序可用来计算第 p 百分位数：

遵循这些步骤将使得计算百分位数相当容易。

计算第 p 百分位数

步骤 1. 把数据按升序排列(从最小值到最大值)。

步骤 2. 计算指数 i

$$i=\left(\frac{p}{100}\right)n$$

式中的 p 是所需的百分位数，n 是观察值的个数。

步骤 3. (a) 如果 i 不是整数的话，则向上取整。用比 i 大的下一个整数代表第 p 百分位数的位置。

(b) 如果 i 是整数的话，第 p 百分位数是第 i 项和第 $(i+1)$ 项数据的平均值。

作为该方法的一个示例，让我们来确定表 3.1 的起薪数据的第 85 百分位数。

步骤 1. 把数据按升序排列：

2 710　2 755　2 850　2 880　2 880　2 890　2 920　2 940　2 950　3 050　3 130　3 325

步骤 2.

$$i=\left(\frac{p}{100}\right)n=\left(\frac{85}{100}\right)12=10.2$$

步骤 3. 因为 i 不是整数，故向上取整。第 85 百分位数的位置是大于 10.2 的下个整数，即第 11 项。

现在回到数据中去，我们看到第 85 百分位数是第 11 项的值，即 3 130。

再举一个例子来说明该方法，我们考虑第 50 百分位数的计算。应用步骤 2，我们得到

$$i=\left(\frac{50}{100}\right)12=6$$

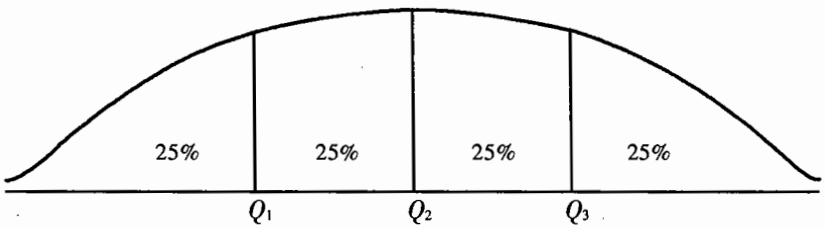
因为 i 是整数，步骤 3(b)规定第 50 百分位数是第 6 和第 7 项数据的平均值。因此第 50 百分位数是(2 890 + 2 920)/2 = 2 905。我们注意到**第 50 百分位数同时也是中位数。**

四分位数

我们往往需要把数据分成四部分，每部分都包括大约 1/4 或 25% 的观察值。图 3.1 显示了被分为四部分的数据集，其分割点就称为**四分位数**(quartiles)，它的定义为

Q_1 = 第一四分位数，或第 25 百分位数;
Q_2 = 第二四分位数，或第 50 百分位数(也是中位数);
Q_3 = 第三四分位数，或第 75 百分位数。

四分位数只是一种特殊的百分位数。因此，计算百分位数的步骤也能够直接应用于四分位数的计算。



第一四分位数　第二四分位数　第三四分位数
（第 25 百分位数）（第 50 百分位数）（第 75 百分位数）
（中位数）

图 3.1　四分位数的位置

把每月起薪数据按照升序重新排列。Q_2，即第二四分位数（中位数），已被明确为 2 905。

2 710　2 755　2 850　2 880　2 880　2 890　2 920　2 940　2 950　3 050　3 130　3 325

Q_1 和 Q_3 的计算需要使用计算第 25 百分位数和第 75 百分位数的方法。计算如下：

对 Q_1

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

因为 i 是整数，根据步骤 3(b)，第一四分位数，即第 25 百分位数，是第三和第四项值的平均数。因此，Q_1 = (2 850 + 2 880)/2 = 2 865

对 Q_3

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

同样，因为 i 是整数，根据步骤 3(b)，第三四分位数，或第 75 百分位数，是第九和第十项值的平均数。因此，Q_3 = (2 950 + 3 050)/2 = 3 000(美元)。

四分位数把数值分为四部分，每部分包括25%的观察值。

2 710　2 755　2 850 | 2 880　2 880　2 890 | 2 920　2 940　2 950 | 3 050　3 130　3 325

$Q_1 = 2\,865$　　$Q_2 = 2\,905$（中位数）　　$Q_3 = 3\,000$

我们已经把四分位数定义为第25、第50和第75百分位数。因此，我们用和计算百分位数相同的方法计算了四分位数。但是，有时也用其他方法来计算四分位数，而且根据所用方法不同，计算出的实际值可能会略有不同。不过，所有计算四分位数方法的目标都是把数据集平分为四部分。

评　注

1. 当数据集含有异常值时，使用中位数作为中心位置的度量比均值更合适。有时，我们使用另外一种度量方法，即**处理均值**。我们删掉数据集中最大的一些值和最小的一些值，然后计算剩余数据的均值，就能够得到处理均值。例如，我们从数据集删除5%最小的和5%最大的数据值，再计算剩下数据的均值，得到的就是5%处理均值。表3.1起薪数据的5%处理均值是2 924.50。
2. 探索性数据分析的支持者提议用其他方法替代四分位数，把数据集分为四个相等部分。该方法的**下临界点**对应于第一四分位数，上临界点对应于第三四分位数。因为计算方法不同，**临界点**的数值可能和四分位数略有差别。但是，它们都能够把数据集分成基本相等的四部分。对表3.1的起薪数据，临界点和四分位数的数值相同。

练　习

方　法

1. 考虑数据值为10，20，12，17和16的5容量样本，计算其均值和中位数。

2. 考虑容量为6的样本，其数据值分别是10，20，21，17，16和12，计算其均值和中位数。

3. 考虑容量为8的样本，其数据值分别是27，25，20，15，30，34，28和25，计算其第20、第25、第65和第75百分位数。

自测题

4. 考虑下述样本，数据值分别为53，55，70，58，64，57，53，69，57，68和53，计算其均值、中位数和众数。

应　用

5. 根据美国国家学院和雇主协会(National Association of Colleges and Employers)的薪金调查，获得会计专业学士学位的工作申请人收到的起始工资待遇在 1999 年平均为每年 34 500 美元(Bureau of Labor Statistics, *Occupational Outlook Handbook* 2000 年第一版)。30 名 2000 年毕业的会计学学士学位获得者组成了一个样本，其起薪数据如下，数据的单位是千美元。


CD 光盘数据
AcctSal

36.8	34.9	35.2	37.2	36.2
35.8	36.8	36.1	36.7	36.6
37.3	38.2	36.3	36.4	39.0
38.3	36.0	35.0	36.7	37.9
38.3	36.4	36.5	38.4	39.4
38.8	35.4	36.4	37.0	36.4

a. 起薪的均值是多少？
b. 起薪的中位数是多少？
c. 众数是多少？
d. 第一四分位数是多少？
e. 第三四分位数是多少？

6. 越来越多的投资者正在转向打折的经纪人以在买卖股票时节约资金。美国个人投资者协会对打折经纪人进行了一项年度调查。表 3.2 显示了 20 个打折经纪人收取的佣金样本，这些经纪人从事两类交易：以每股 50 美元的价格买卖 500 股，和以每股 5 美元的价格买卖 1 000 股。

表 3.2　打折经纪人收取的佣金


CD 光盘数据
Discount

经纪人	佣金(美元)	
	500 股，每股 50 美元	1 000 股，每股 5 美元
AcuTrade	38.00	48.00
Bank of San Francisco	140.00	79.50
Burke Christensen & Lewis	34.00	34.00
Bush Burns Securities	35.00	35.00
Charles Schwab	155.00	90.00
Downstate Discount	55.00	60.00
Dreyfus Lion Account	154.50	88.50
First Union Brokerage	140.00	90.00
Levitt & Levitt	35.00	70.00
Max Ule	195.00	70.00

（续表）

经纪人	佣金（美元）	
	500 股，每股 50 美元	1 000 股，每股 5 美元
Mongerson & Co	95.00	66.00
Quick & Reilly	119.50	60.50
Scottsdale Securities, Inc.	50.00	63.00
Seaport Securities Corp.	50.00	70.00
St. Louis Discount	66.00	64.00
Summit Financial Services	95.00	60.50
T. Rowe Price Brokerage	134.00	80.00
Unified Financial Services	154.00	90.00
Wall Street Access	45.00	45.00
Your Discount Broker	55.00	70.00

资料来源：*AAII Journal*, January 2000。

a. 计算以每股 50 美元买卖 500 股的交易收取佣金的均值、中位数、众数。

b. 计算以每股 5 美元买卖 1 000 股的交易收取佣金的均值、中位数、众数。

c. 哪种交易费用最高：是以每股 50 美元买卖 500 股，还是以每股 5 美元买卖 1 000 股？

d. 交易的费用是否与交易金额有关？举例来说，当以每股 50 美元买卖 500 股时交易金额就等于 25 000 美元。

7. 平均每人每天花费 45 分钟来欣赏录音音乐（*The Des Moines Register*, December 5, 1997）。下列数据是一个 30 人样本花费在欣赏录音音乐上的分钟数：


CD 光盘数据
Music

88.3	4.3	4.6	7.0	9.2
0.0	99.2	34.9	81.7	0.0
85.4	0.0	17.5	45.0	53.3
29.1	28.8	0.0	98.9	64.5
4.4	67.9	94.2	7.6	56.6
52.9	145.6	70.4	65.1	63.6

a. 计算均值和众数。

b. 这些数据是否显示出和报纸报道的平均数相一致？

c. 计算中位数。

d. 计算第一和第三四分位数。

e. 计算并解释第 40 百分位数。

8. 数以百万计的美国人每天早上起来后在家中办公。对个人电脑日渐增长的应用据说是使更多的人们能够在家处理业务的原因之一。下列是在家工作个人的年龄数据样本：


自测题

22	58	24	50	29	52	57	31	30	41
44	40	46	29	31	37	32	44	49	29

a. 计算均值和众数。

b. 所有成年人总体的中位数年龄是 35.1 岁(U. S. Census Bureau, November 1, 1997)。使用上述数据的中位数年龄，来评论在家办公人士的年龄比所有成年人总体是年轻还是年长？

c. 计算第一和第三四分位数。

d. 计算并解释第 32 百分位数。

9. Media Matrix 收集的数据显示了人们在家和工作中上网浏览时，最流行的网站名(*Business 2. 0*, January 2000)。下列数据表示人们在家浏览的前 25 位网站的个人访问者的人数(以 1 000 人为单位)：



CD 光盘数据
Websites

网站名	个人访问者（千人）
about. com	5 538
altavista. com	7 391
amazon. com	7 986
angelfire. com	8 917
aol. com	23 863
bluemountainarts. com	6 786
ebay. com	8 296
excite. com	10 479
geocities. com	15 321
go. com	14 330
hotbot. com	5 760
hotmail. com	11 791
icq. com	5 052
looksmart. com	5 984
lycos. com	9 950
microsoft. com	15 593
msn. com	23 505
netscape. com	14 470
passport. com	11 299
real. com	6 785
snap. com	5 730
tripod. com	7 970
xoom. com	5 652
yahoo. com	26 796
zdnet. com	5 133

a. 计算均值和中位数。

b. 你认为使用均值还是中位数来对这些数据的中心趋势进行度量更好一些？并解释。

c. 计算第一和第三四分位数。

d. 计算并解释第 85 百分位数。

10. 《洛杉矶时报》定期报道南加州各个地区的空气质量指数。0—50 的指数等级被认为属于良好，51—100 属于一般，101—200 属于不健康，201—275 属于非常不健康，超过 275 则为危险。Pomona 地区最近的空气质量指数是 28，42，58，48，45，55，60，49 和 50。

a. 计算数据的均值、中位数和众数，Pomona 的空气质量指数应该被认为是良好吗？

b. 计算 Pomona 空气质量数据的第 25 和第 75 百分位数。

11. 下列数据表示在 20 个时间段内——每段包括 10 分钟——到达收费站的汽车数，计算数据的均值、中位数、众数、第一四分位数以及第三四分位数。

26	26	58	24	22	22	15	33	19	27
21	18	16	20	34	24	27	30	31	33

12. 在汽车的汽油里程数和汽油消耗量测试中，13 辆汽车分别在城市和乡村行驶 300 英路的路况下接受了检测，下列数据记录了它们的每加仑英里数。

城市：16.2 16.7 15.9 14.4 13.2 15.3 16.8 16.0 16.1 15.3 15.2 15.3 16.2

乡村：19.4 20.6 18.3 18.6 19.2 17.4 17.2 18.6 19.0 21.1 19.4 18.5 18.7

运用均值、中位数和众数来评价在城市和乡村路况下汽车性能的差别。

13. 一个包括 15 名大学四年级学生的样本显示了第四年最后一学期学生们得到的学分数：

15 21 18 16 18 21 19 15 14 18 17 20 18 15 16

a. 学分数据的均值、中位数和众数分别是多少？计算并解释。

b. 计算第 1 和第 3 百分位数。

c. 计算并解释第 70 百分位数。

14. 由于最近的技术进步，今天的数字相机能够比几年前的同类产品拍出效果更好的图片。下列数据显示了 20 款最新型号数字相机样本的零售价格、最大图片容量以及电池的使用时间(分钟)(*PC World*, January 2000)。



CD 光盘数据
Cameras

相机	价格(美元)	最大图片容量	电池使用时间(分钟)
Agfa Ephoto CL30	349	36	25
Canon PowerShot A50	499	106	75
Canon PowerShot Pro70	999	96	118
Epson Photo PC 800	699	120	99
Fujifilm DX-10	299	30	229
Fujifilm MX-2700	699	141	124
Fujifilm MX-2900 Zoom	899	141	88
HP Photo Smart C200	299	80	68
Kodak DC215 Zoom	399	54	159
Kodak DC265 Zoom	899	180	186
Kodak DC280 Zoom	799	245	143

（续表）

相机	价格（美元）	最大图片容量	电池使用时间(分钟)
Minolta Dimage EX Zoom 1500	549	105	38
Nikon Coolpix 950	999	32	88
Olympus D-340R	299	122	161
Olympus D-450 Zoom	499	122	62
Richo RDC-500	699	99	56
Sony Cybershot DSC-F55	699	63	69
Sony Mavica MVC-FD73	599	40	186
Sony Mavica MVC-FD88	999	40	88
Toshiba PDR-M4	599	124	142

a. 计算平均价格。

b. 计算最大图片容量的均值。

c. 计算平均的电池使用时间。

d. 如果你必须从上面选择一款相机的话，你会选哪一款？请解释原因。

3.2　变异程度的度量

除了位置度量以外，人们往往还需要考虑变异程度，或者说是离散程度。例如，假设你是一个大型制造公司的采购代理，定期地向两个不同的供应商订货。经过几个月的运营，你发现两个供应商完成订单所需的时间大概都是 10 天。完成订单所需时间的直方图如图 3.2 所示。虽然两个供应商的平均交货日期都是 10 天左右，但他们在按期交货方面是否拥有相同的可信度？注意直方图中的离散程度或变异程度，你会选择哪一家供应商？

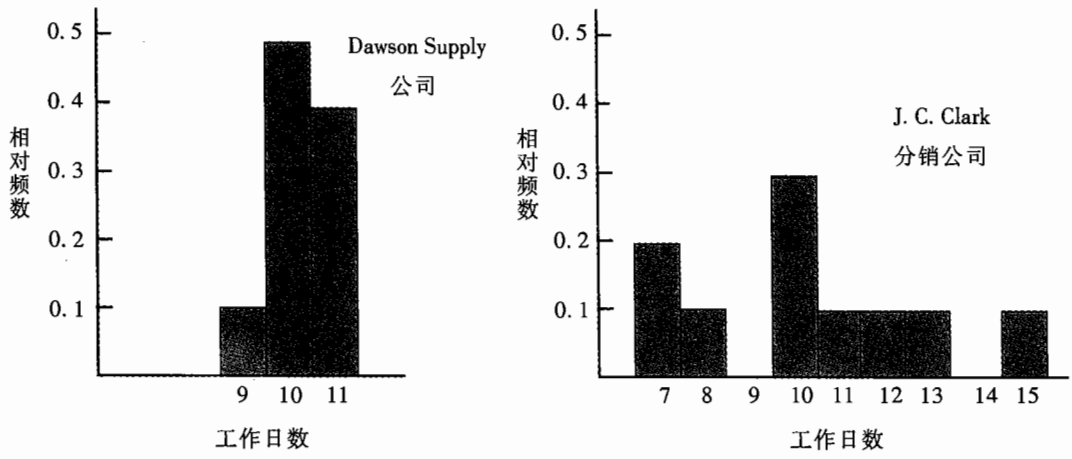


图 3.2　表示完成订单所需时间的直方图数据

对于大多数公司来说，按时收到原材料和货物很重要。从 J. C. Clark 分销公司的直方图来看，7 或 8 天的交货时间确实有，但那几个推迟到 13—15 天的交货时间对于保持公司的设备运转和按时生产将是灾难性的。这个例子说明，在某些情况下，交货时间的变异程度可能是选择供应商的关键考虑因素。对于大多数采购代理而言，只有像 Dawson Supply 公司这样变异程度较低的供应者，才是较理想的供货商。

现在我们来讨论一些常用的变异程度度量。

极　差

也许**极差**(range)是一种最简单的变异程度度量。

极差易于计算，但是只对两个数据值敏感：最大值和最小值。

极差

$$\text{极差} = \text{最大值} - \text{最小值}$$

让我们参见表 3.1 的商学院毕业生起始月薪数据，其最高月薪为 3 325 美元，最低月薪为 2 710。极差为 3 325 - 2 710 = 615。

虽然极差是最易于计算的对变异程度的衡量，但是它很少被单独用来度量变异程度。其原因是，极差的确定仅仅根据两个观察值，因此极易受到异常值的影响。假设某位毕业生得到了 10 000 美元的起始月薪，则极差将是 10 000 - 2 710 = 7 290 而不是 615。过大的极差将不能准确地描述数据的变异程度，因为在 12 个起薪中有 11 个集中在 2 710—3 130 之间。

四分位点内距

四分位点内距(interquartile range，IQR)作为变异程度的一种度量，能够克服异常值的影响。它是第三四分位数 Q_3 和第一四分位数 Q_1 的差值。换句话说，四分位点内距就是数据中间 50% 的距离。

四分位点内距

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

对于起始月薪数据，四分位数为 Q_3 = 3 000 和 Q_1 = 2 865。因此，四分位点内距等于 3 000 - 2 865 = 135。

方　差

方差(variance)是利用所有数据对变异程度进行的度量。它建立在每个观察值(x_i)和均值之差的基础上，每个 x_i 与均值（样本均值用$\bar{x}$表示，总体均值用 μ 表示）的差被称为**关于均值的离**

差。对于样本而言，关于均值的离差就记作($x_i - \bar{x}$)；对于总体，则记作($x_i - \mu$)。在方差的计算中，需要把关于均值的离差**平方**。

方差在比较两变量的变异程度时很有用。

如果数据来自总体，平方离差的平均值就被称为**总体方差**，总体方差用希腊符号 σ^2 表示。对于具有 N 个观察值的总体，用 μ 表示总体均值，可以定义总体方差如下：

总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3.4}$$

在大部分统计应用中，需要分析样本数据。当我们计算**样本方差**时，往往对于如何利用它来估计总体方差 σ^2 感兴趣。人们能够看到，如果关于样本均值的平方离差之和除以（$n-1$），而不是 n 的话，所得到的样本方差是总体方差的无偏估计，不过对这一问题的详细解释已经超出了本课程的范围。鉴于此，由 s^2 表示的**样本方差**定义如下：

样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{3.5}$$

为了说明样本方差的计算过程，我们使用 3.1 节的 5 个大学班级样本的班级人数数据。数据的有关汇总结果，包括关于均值的离差和均值的平方离差的计算过程，都显示在表 3.3 中。关于均值的平方离差之和 $\sum (x_i - \bar{x})^2 = 256$，并且 $n - 1 = 4$。因此，样本方差为

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{256}{4} = 64$$

表 3.3　对班级人数关于均值的离差和均值的平方离差的计算过程

班级学生数	班级人数均值	关于均值的离差	关于均值的平方离差
(x_i)	($\bar{x}$)	($x_i - \bar{x}$)	($x_i - \bar{x}$)^2
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\sum (x_i - \bar{x})$	$\sum (x_i - \bar{x})^2$

在转到其他问题以前，我们注意到样本方差的单位往往会引起人们的困惑。因为在方差计算

中需要求和的数值：$(x_i-\bar{x})^2$，是平方项，所以样本方差的单位也是平方项。例如，班级人数数据的样本方差是 $s^2=64$(人)^2。方差的平方单位使得人们很难对方差的数值进行直观的理解和解释。我们建议读者把方差作为比较两个以上变量变异程度的有用工具。在变量的比较中，具有更大方差的变量，其变异程度也更大。而对于方差数值的进一步解释可能就没有必要了。

作为说明样本方差计算方法的另外一个例子，我们考虑表 3.1 中 12 个商学院毕业生的起薪数据。在 3.1 节里，我们已经知道起薪样本的均值是 2 940。对样本方差的计算结果($s^2=27\,440.91$) 列示在表 3.4 中。

表 3.4　起薪数据样本方差的计算过程

月薪(x_i)	样本均值($\bar{x}$)	关于均值的离差($x_i-\bar{x}$)	关于均值的平方离差$(x_i-\bar{x})^2$
2 850	2 940	−90	8 100
2 950	2 940	10	100
3 050	2 940	110	12 100
2 880	2 940	−60	3 600
2 755	2 940	−185	34 225
2 710	2 940	−230	52 900
2 890	2 940	−50	2 500
3 130	2 940	190	36 100
2 940	2 940	0	0
3 325	2 940	385	148 225
2 920	2 940	−20	400
2 880	2 940	−60	3 600
		0	301 850
		$\sum(x_i-\bar{x})$	$\sum(x_i-\bar{x})^2$

使用公式(3.5)

$$s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{301\,850}{11}=27\,440.91$$

注意到在表 3.3 和表 3.4 中，我们既列出了关于均值的离差之和，也列出了关于均值的平方离差之和。对于任何数据集，关于均值的离差之和将**总是等于零**。我们可以看到，表 3.3 和表 3.4 中的 $\sum(x_i-\bar{x})=0$。由于正的离差和负的离差总是相互抵消，导致了关于均值的离差之和等于零。

标准差

我们定义**标准差**(standard deviation) 为方差的正平方根。按照以前采用的样本方差和总体方差符号，我们以 s 表示样本标准差，以 σ 表示总体标准差。标准差可以用以下方法从方差推导而来：

标准差

$$\text{样本标准差}=s=\sqrt{s^2} \tag{3.6}$$

$$\text{总体标准差}=\sigma=\sqrt{\sigma^2} \tag{3.7}$$

标准差比方差更易于解释，因为标准差与数据的单位相同。

我们已知 5 个大学班的班级人数的样本方差等于 $s^2 = 64$，因此样本标准差为 $s = \sqrt{64} = 8$。对于起薪数据集，样本标准差为 $s = \sqrt{27\,440.91} = 165.65$。

把方差转换为对应的标准差有什么好处？我们知道与方差相联系的单位是平方项，例如，商学院毕业生起薪数据的样本方差等于 $s^2 = 27\,440.91$(美元)^2。因为标准差是方差的平方根，则方差的单位美元平方就转换为标准差的单位美元。所以起薪数据的标准差是 165.65 美元。换句话说，标准差是以与原始数据相同的单位来度量的。因为这个原因，标准差更易于和单位也与原始数据相同的均值等其他统计量进行比较。

变异系数

在某些情况下，我们可能对表示标准差与均值相对大小的描述性统计量产生兴趣。它称为**变异系数**(coefficient of variation)，可计算如下：

变异系数

$$\frac{\text{标准差}}{\text{均值}} \times 100 \tag{3.8}$$

变异系数是对变异程度的相对度量。它衡量标准差相对于均值的大小。

对于班级人数数据，我们知道样本均值为 44，样本标准差为 8。变异系数等于 $(8/44) \times 100 = 18.2$。简而言之，变异系数说明样本标准差是样本均值的 18.2%。对于样本均值为 2 940，样本标准差为 165.65 的起薪数据，变异系数为 $(165.65/2\,940) \times 100 = 5.6$。它说明样本标准差只是样本均值的 5.6%。一般地，变异系数对于比较具有不同标准差和不同均值的变量变异程度，是很有用的统计量。

评　注

1. 我们能够使用统计软件包和分布表来计算本章的描述性统计量。把数据输入工作表以后，使用几个简单的命令就能够生成需要的结果。在附录 3.1 和 3.2 中，我们将演示怎样利用 Minitab 和 Excel 来计算描述性统计量。
2. 标准差通常用来度量与股票投资和股票基金投资相关的风险(*Business Week*, January 17, 2000)。它给出了每月收益如何围绕长期平均收益波动状况的度量。
3. 当使用计算器计算方差和标准差时，对样本

平均数$\bar{x}$和平方离差$(x_i-\bar{x})^2$进行四舍五入处理，可能会引起方差或标准差的计算误差。我们建议在计算的中间步骤中，保留6位有效数字，对得到的方差或标准差可以保留少一些的位数。

4. 另外一个可选的样本方差公式为

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

式中的$\sum x_i^2=x_1^2+x_2^2+\cdots+x_n^2$。

练　习

方　法

15. 考虑容量为5的样本，数据值分别为10，20，12，17和16，计算极差和四分位点内距。

16. 考虑容量为5的样本，数据值分别为10，20，12，17和16，计算方差和标准差。

17. 考虑容量为8的样本，数据值分别为27，25，20，15，30，34，28和25，计算极差、四分位点内距、方差和标准差。

自测题

应　用

18. 一位板球投球手在六局比赛中的得分分别为：182，168，184，190，170和174，使用这些数据作为样本，计算下列描述性统计量：

自测题

a. 极差。

b. 方差。

c. 标准差。

d. 变异系数。

19. 《PC世界》给出了销售最好的15种笔记本电脑的评分（*PC World*，February 2000）。每一种笔记本电脑的测试结果都以100为综合评分的满分。90分以上的分数是优异的，而70多分则表示在平均水平以上。被测试过的15种笔记本的综合评分如下所示：

CD光盘数据 Notebook

笔记本电脑	综合评分
AMS Tech Roadster 15CTA380	67
Compaq Armada M700	78
Compaq Prosignia Notebook 150	79
Dell Inspiron 3700 C466GT	80
Dell Inspiron 7500 R500VT	84
Dell Latitude Cpi A366XT	76
Enpower ENP-313 Pro	77
Gateway Solo 9300LS	92

（续表）

笔记本电脑	综合评分
HP Pavillion Notebook PC	83
IBM ThinkPad I Series 1480	78
Micro Express NP7400	77
Micron TransPort NX P Ⅱ-400	78
NEC Versa SX	78
Sceptre Soundx 5200	73
Sony VAIO PCG-F340	77

计算极差、四分位点内距、方差和标准差。

20.《洛杉矶时报》（*Los Angeles Times*）定期报告南加州不同地区的空气质量指数。Pomona 地区空气质量指数值的样本给出了下列数据：28，42，58，48，45，55，60，49 和 50。

a. 计算值域和四分位点内距。

b. 计算样本方差和样本标准差。

c. 从 Anaheim 地区得到的空气质量指数样本的样本均值为 48.5，样本方差为 136，样本标准差是 11.66。在这些描述性统计量的基础上，你对 Pomona 和 Anaheim 的空气质量进行比较，能够得到什么结果？

21. Davis 制造公司已经使用新工艺完成了 5 周运转，据认为这种新工艺能够提高劳动生产率。该公司每周生产的部件数是 410，420，390，400 和 380，计算样本方差和样本标准差。

22. 假设使用下列天数数据来构建直方图，这些天数是 Dawson Supply 公司和 J. C. Clark 分销公司为了完成客户订单所需要的（见图 3.2）。

Dawson Supply **公司的交货时间：**	11	10	9	10	11	11	10	11	10	10
Clark **分销公司的交货时间：**	8	10	13	7	10	11	10	7	15	12

使用极差和标准差来支持这样一种明显的观察结果：Dawson Supply 公司的交货时间更一致和可靠。

23. 警察记录显示了冬季样本和夏季样本的每日犯罪报告数：



CD 光盘数据
Crime

冬季	夏季
18	28
20	18
15	24
16	32
21	18
20	29
12	23
16	38
19	28
20	18

a. 计算每个季节的极差和四分位点内距。

b. 计算每个季节的方差和标准差。

c. 计算每个季节的变异系数。

d. 比较两个季节的变异程度。

24. 美国个人投资者协会进行了一个关于打折经纪人的年度调查(*AAII Journal*, January 1997)。表 3.2 表示了 20 个打折经纪人样本收取的佣金数，这些经纪人从事两类交易：以 50 美元买卖 500 股和以 5 美元买卖 1 000 股。



CD 光盘数据
Discount

a. 计算每种交易类型的极差和四分位点内距。

b. 计算每种交易类型的方差和标准差。

c. 计算每种交易的变异系数。

d. 比较两种交易费用的变异程度。

25. 生产部门使用抽样方法检验新产品的质量。该部门在检查站应用如下决策规则：如果 14 个产品的样本方差超过 0.005，则必须关闭生产线进行修理。假如刚收集到下列数据：

3.43	3.45	3.43	3.48	3.52	3.50	3.39
3.48	3.41	3.38	3.49	3.45	3.51	3.50

生产线应该关闭吗？为什么？

26. 下列数据是某大学田径队的队员跑 1/4 英里和 1 英里的时间（单位为分）：

1/4 英里跑时间：	0.92	0.98	1.04	0.90	0.99
1 英里跑时间：	4.52	4.35	4.60	4.70	4.50

一位教练在看到这个样本后认为，1/4 英里跑的时间更具有一致性。运用标准差和变异系数来表示数据的变异程度，使用变异系数来判断教练的说法是否成立？

3.3 相对位置的度量和异常值的检测

我们已经讲述了几种对数据位置和变异程度的度量方法。均值是应用最广泛的位置度量，而标准差和方差在度量变异程度时应用最广。只使用均值和标准差，我们也能够对数据在数据集里的相对位置了解很多。

z - 分数

通过运用均值和标准差，我们能够确定任何观察值的相对位置。假设我们有一个包括 n 个观察值的样本，其数值以 x_1，x_2，...，x_n 表示。另外，还假定样本均值和样本标准差 s 已知。与任一数据值 x_i 有关的另外一个值被称为该数据值的 **z - 分数**(z-score)。公式(3.9)显示了怎样计

算每个 x_i 的 z - 分数。

z - 分数

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

式中　z_i——x_i 的 z - 分数；
　　$\bar{x}$——样本均值；
　　s——样本标准差。

z - 分数往往被称为**标准化值**。标准化值或者说是 z - 分数 z_i，可以被认为是 **x_i 距均值的标准差个数**。例如，$z_1 = 1.2$ 就表示 x_1 比样本均值大 1.2 个标准差。类似地，$z_2 = -0.5$ 表示 x_2 比样本均值小 0.5 或 1/2 个标准差。我们还注意到当观察值大于均值时，z - 分数将大于零；而当观察值小于均值时，z - 分数将小于零。z - 分数为零则表示观察值等于均值。

任何观察值的 z - 分数都可看做是对数据集中观察值的相对位置度量。因此，两个不同数据集的观察值若具有相同的 z - 分数，则可以根据它们距离均值的标准差个数相同这个事实，而认为它们具有同样的相对位置。

班级人数的 z - 分数列于表 3.5 内。我们已知样本均值 $\bar{x} = 44$，样本标准差 $s = 8$，其计算过程参见有关章节。第五观察值的 z - 分数为 -1.50，显示出它离均值最远，位于均值之下 1.50 个标准差。

表 3.5　班级人数的 z - 分数

班级学生数 (x_i)	关于均值的离差 $(x_i - \bar{x})$	z - 分数 $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	-2	-2/8 = -0.25
46	2	2/8 = 0.25
32	-12	-12/8 = -1.50

切比雪夫定理

借助于**切比雪夫定理**(Chebyshev's theorem)，我们能够指出那些距离均值必定在某个特定标准差个数以内的数据项占数据总数的比例。

切比雪夫定理要求 $z>1$，但 z 不必是整数。练习 27 就涉及到了大于 1 的非整数 z 值。

切比雪夫定理

与均值的距离必定在 z 个标准差以内的数据比例至少为 $(1-1/z^2)$，这里的 z 是大于 1 的任何值。

z = 2, 3, 4 个标准差时，该定理的一些应用如下：

- 至少 0.75 或 75% 的数据项与均值的距离在 z = 2 个标准差以内。
- 至少 0.89 或 89% 的数据项与均值的距离在 z = 3 个标准差以内。
- 至少 0.94 或 94% 的数据项与均值的距离在 z = 4 个标准差以内。

作为切比雪夫定理的一个应用例子，假定在某学院商务统计课程的期中考试中，100 个学生的成绩均值为 70，标准差为 5。那么有多少学生的考试成绩在 60—80 之间？有多少学生的成绩在 58—82 之间？

对于 60—80 的成绩，我们注意到 60 低于均值 2 个标准差，而 80 高于均值 2 个标准差。运用切比雪夫定理，我们可以看到至少 0.75 或 75% 的观察值必定距离均值 2 个标准差以内。因此，100 个学生中至少有 75 个一定取得了 60—80 的分数。

对于 58 到 82 的考试分数，我们看到，$(58-70)/5 = -2.4$ 表示 58 低于均值 2.4 个标准差，$(82-70)/5 = +2.4$ 表示 82 高于均值 2.4 个标准差。应用切比雪夫定理且 $z=2.4$，我们有

$$\left(1-\frac{1}{z^2}\right) = \left[1-\frac{1}{(2.4)^2}\right] = 0.826$$

至少有 82.6% 的学生成绩必定在 58 到 82 之间。

经验法则

切比雪夫定理的优点之一就是它适用于任何数据集而无论数据是怎样分布的。但是在实际应用中，人们发现许多数据集具有类似图 3.3 所显示的峰形或钟形分布。当数据被认为具有类似这样的分布时，我们就能够运用**经验法则**(empirical rule)来确定与均值距离在某特定的标准差个数以内的数据项所占的百分比。

经验法则建立在正态概率分布的基础之上。对于正态分布，我们将在第 6 章予以介绍，它在本书中的应用非常广泛。

经验法则

对于具有钟形分布的数据：

- 大约 68% 的数据项与均值距离在 1 个标准差以内。
- 大约 95% 的数据项与均值距离在 2 个标准差以内。
- 几乎所有的数据项与均值距离在 3 个标准差以内。

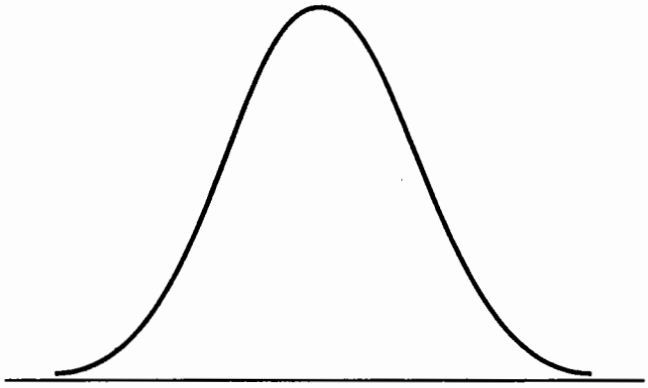


图 3.3　钟形或峰形分布

例如，在生产线上对去污液包装盒进行自动装填，填充重量具有钟形分布。如果填充重量的均值为 16 盎司，标准差为 0.25 盎司的话，我们能够使用经验法则得到下列结论：

- 大约 68% 的填充后的包装盒重量在 15.75 和 16.25 盎司之间（也就是说，距均值 1 个标准差之内）。
- 大约 95% 的填充后的包装盒重量在 15.50 和 16.50 盎司之间（也就是说，距均值 2 个标准差之内）。
- 几乎所有填充后的包装盒重量都在 15.25 和 16.75 盎司之间（也就是说，距均值 3 个标准差之内）。

异常值检测

有时数据集含有一个或多个数值异常大或者异常小的观察值，像这样的极端值被称为**异常值**(outliers)。有经验的统计人员会采取一定的步骤识别出异常值，并对每一个异常值进行仔细的检查。异常值可能是被不正确记录的数据值，则我们可以在进一步分析之前把它更正。异常值也可能是本不应当包括进数据集的观察值，那么我们可以把它删掉。也许，异常值只是数值不太正常，它并未记录错误，也确实属于数据集。在这种情况下，该数据项应该保留。

基于数据分析，做决策之前检测异常值是很有意义的。数据记录中经常会有误差并被输入计算机，异常值并非都要删除，但它们的精确度和适应性应予以确认。

标准化值（z－分数）可以帮助我们鉴别异常值。刚刚学过的经验法则使我们能够得出结论：对于具有钟形分布的数据而言，几乎所有的数据项都距离均值在 3 个标准差以内。所以，在应用 z－分数鉴别异常值时，我们建议把 z－分数小于 -3 或大于 3 的任何数据项都作为异常值。然后，对它们的准确性进行检查，以确定它们是否属于数据集。

参见表 3.5 班级人数数据的 z - 分数，-1.50 的 z - 分数表明第五项距均值最远，不过它的标准化值仍在判断值的标准限 -3—+3 以内。因此，根据 z - 分数，班级人数数据中没有异常值。

评 注

1. 切比雪夫定理可应用于任何数据集，并能够指出距离均值在某个标准差以内的数据项的最少个数。如果已知数据集近似呈钟形分布的话，则这样的数据项个数会比切比雪夫定理指出的更多。例如，经验法则让我们**确信**有 95% 的数据项距离均值在 2 个标准差以内，而切比雪夫定理指出在同一区间内的个数比例则至少为 75%。
2. 在对数据集进行分析之前，统计人员通常会作各种检查以确保数据的有效性。大型研究活动中，在记录数据或向计算机输入数据的过程中往往会出现误差。鉴别异常值是检查数据有效性的一个工具。

练 习

方 法

27. 自测题 考虑均值为 30 而标准差为 5 的样本，运用切比雪夫定理来确定处于下列范围内的数据比例或者百分比。

a. 20—40　　b. 15—45　　c. 22—38　　d. 18—42　　e. 12—48

28. 数据具有钟形分布，其均值为 30，标准差为 5，使用经验法则来确定处于下列范围内的数据比例或百分比。

a. 20—40　　b. 15—45　　c. 25—35

29. 考虑容量为 5 的样本，各项数据的值分别为 10，20，12，17，16，计算这 5 个数据值中每个 z - 分数。

30. 考虑均值为 500 而标准差为 100 的样本。对于数据值 520，650，500，450，280，每个值的 z - 分数是多少？

应 用

31. 对 1 154 个成人的美国国内调查结果显示，他们在工作日中每天平均睡眠 6.9 小时（2000 Omnibus Sleep in America Poll）。假定标准差是 1.2 小时。

a. 使用切比雪夫定理计算每天睡眠时间在 4.5 到 9.3 小时之间的人数百分比。

b. 使用切比雪夫定理计算每天睡眠时间在 3.3—9.9 小时之间的人数百分比。

c. 假定睡眠小时数是钟形的，使用经验法则计算每天睡眠时间处于 4.5—9.3 小时之间的人数百分比。这个结果与(a)中用切比雪夫定理计算出的值相比怎样？

32. 根据 ACNielsen 的调查，年龄在 12—17 岁之间的少年，在 8 月份结束的广播年度内平均每天看 3 个小时的电视（*Barron's*, November 8, 1999）。假定标准差是 1 小时，看电视的时间数据具有钟形分布。

a. 年龄在 12—17 岁的少年中，每天看 2—3 个小时电视的占多大百分比？

b. 年龄在 12—17 岁的少年中，每天看 1—4 个小时电视的占多大百分比？

c. 年龄在 12—17 的少年中，每天看 4 个小时以上电视的占多大百分比？

33. 假设 IQ 分数具有钟形分布，其均值为 100，标准差为 15。

a. IQ 分数在 85—115 之间的人士占多大百分比？

b. IQ 分数在 70—130 之间的人士占多大百分比？

c. IQ 分数超过 130 的人士占多大百分比？

d. IQ 分数超过 145 的人被认为是天才。经验法则支持这种说法吗？请解释。

34. 在芝加哥修理彩电的平均人力成本是 90.06 美元（*The Wall Street Journal*, January 2, 1998）。假设标准差是 20 美元。

a. 人力成本为 71 美元的修理工作的 z - 分数是多少？

b. 人力成本为 168 美元的修理工作的 z - 分数是多少？

c. 解释(a)和(b)中的 z - 分数。说明它们是否都应被认为是异常值。

35. 
CD 光盘数据
WageWeb

Wageweb 进行了一个关于薪资数据的调查，并在它的网站上公布了汇总结果。应用 2000 年 1 月 1 日的薪资数据，Wageweb 报告说捐赠基金的经理薪金在 50 935 美元—79 577 美元之间（Wageweb. com, April 12, 2000）。假定下列数据是来自由 30 位捐赠基金经理的年薪组成的样本(数据以千美元为单位)：

57.7	64.4	62.1	59.1	71.1
63.0	64.7	61.2	66.8	61.8
64.2	63.3	62.2	61.2	59.4
63.0	66.7	60.3	74.0	62.8
68.7	63.8	59.2	60.3	56.6
59.3	69.5	61.7	58.9	63.1

a. 计算样本数据的均值和标准差。

b. 使用(a)中计算出的均值和标准差作为捐赠基金经理总体薪金均值和标准差的估计值。运用切比雪夫定理来确定年薪在 53 000 美元—71 000 美元之间的经理所占的百分比。

c. 构建样本数据的直方图。它的形状是否说明年薪数据大致属于钟形分布的假定是合理的？

d. 假定年薪的分布是钟形分布。使用(a)中计算出的均值和标准差作为捐赠基金经理

总体薪金均值和标准差的估计值。运用经验法则来确定年薪在53 000美元—71 000美元之间的经理所占的百分比，并将答案与(b)中的计算结果相比较。

e. 样本数据中是否包含任何异常值?

36. 一个由10支NBA球队得分组成的样本提供了下列数据(*USA Today*, April 14, 2000):


CD光盘数据 NBA

获胜队	得分	失败队	得分	得分差距
费城	93	华盛顿	84	9
夏洛特	119	亚特兰大	87	32
密尔沃基	101	克利夫兰	100	1
印第安纳	77	多伦多	73	4
西雅图	110	明尼苏达	83	27
波士顿	95	奥兰多	91	4
底特律	90	迈阿密	73	17
纽约	91	新泽西	89	2
犹他	102	洛杉矶快艇	93	9
菲尼克斯	122	温哥华	116	6

a. 计算获胜队伍得分数的均值和标准差。

b. 假设所有NBA比赛获胜队伍的得分数呈钟形分布。使用(a)中计算出的均值和标准差作为所有NBA比赛总体得分均值和标准差的估计值,估计获胜队得分超过100的比赛场次所占的百分比。估计获胜队得分超过114的比赛所占的百分比。

c. 计算得分差距的均值和标准差。得分差距数据中是否包含任何奇异值?请解释。

37. 《消费者评论》(*Consumer Review*)邮寄对因特网上各种产品的评论和得分情况。下面是一个20种扬声器系统的样本和在1998年1月2号邮寄的得分情况(见http://www.audioreview.com)。得分的范围从1到5，5分为最好。


CD光盘数据 Speakers

扬声器	得分	扬声器	得分
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF 1000	4.17
Hsu Rsch. HRSW 12V	4.56	Joseph Audio RM 7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aerius	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics DIV	4.64	Sunfire True Subwoofer	4.17
Thiel CS 1.5	4.20	Yamaha NS-A636	2.17

a. 计算均值和中位数。

b. 计算第一和第三四分位数。

c. 计算标准差。

d. 与 Allison One 和 the Omni Audio SA 12.3 两种扬声器相联系的 z - 分数分别是多少？

e. 数据中是否含有异常值？请解释。

3.4　探索性数据分析

在第 2 章我们介绍了作为一种探索性数据分析技术的茎叶显示。我们知道探索性数据分析使我们能够运用简单的数学和易画的图形来汇总数据。在本节，我们将继续介绍探索性数据分析技术的两种方法：五数概括法和箱形图。

五数概括法

在**五数概括法**(five-number summary)中，使用下列五个数字来汇总数据:

1. 最小值;
2. 第一四分位数（Q_1）；
3. 中位数（Q_2）；
4. 第三四分位数（Q_3）；
5. 最大值。

构建五数概括法最容易的方法是首先按照升序排列数据，然后再确定最小值、第三四分位数和最大值。在这里把表 3.1 的 12 个商学院毕业生样本的起始月薪数据按照升序重复如下：

2 710　2 755　2 850 | 2 880　2 880　2 890 | 2 920　2 940　2 950 | 3 050　3 130　3 325

$Q_1 = 2\,865$　　$Q_2 = 2\,905$（中位数）　　$Q_3 = 3\,000$

在 3.1 节我们已经计算过中位数 2 905 和第一四分位数 $Q_1 = 2\,865$ 和第三四分位数 $Q_3 = 3\,000$。而最小值为 2 710，最大值为 3 325。因此对薪金数据的五数概括为 2 710，2 865，2 905，3 000，3 325。大约有 1/4 或 25% 的观察值在五数概括的邻近两个数之间。

箱形图

箱形图(box plot)是基于五数概括法的数据图形汇总方法。构建箱形图的一个关键是计算中位数及四分位数 Q_1 和 Q_3，还需要使用四分位点内距 $\text{IQR} = Q_3 - Q_1$。图 3.4 就是起始月薪数据的箱形图。具体的做图步骤如下：

1. 画箱形图时，把第一和第三四分位数作为箱体的边缘。起薪数据的 $Q_1 = 2\,865$，$Q_3 = 3\,000$。该箱体包含了数据中间的 50%。

2. 在中位数（对于起薪数据是 2 905）位置于箱体内画一条垂线，因此中位数线就把数据平分为两部分。
3. 通过使用四分位点内距 IQR = $Q_3 - Q_1$，定好界限的位置。箱形图的界限位置是在低于 Q_1 1.5 倍 IQR 处和高于 Q_3 1.5 倍 IQR 处。对于起薪数据，IQR = $Q_3 - Q_1$ = 3 000 − 2 865 = 135。因此，界限位置就是 2 865 − 1.5 × 135 = 2 662.5 和 3 000 + 1.5 × 135 = 3 202.5。界限外面的数据可被认为是异常值。
4. 图 3.4 中横的那根线被称做须线。须线从箱体边缘一直画到步骤 3 计算出的**界限以内**最小和最大值处。因此在本例中须线的尽头在 2 710 和 3 130 处。
5. 最后，每个异常值的位置用符号 * 来表示。在图 3.4 中，我们看到了一个异常值 3 325。

箱形图提供了另一种鉴别界限位置的方法，但它们不必要鉴别那些小于 −3 或大于 +3 的 z 值。可以选择一种或两种界限。

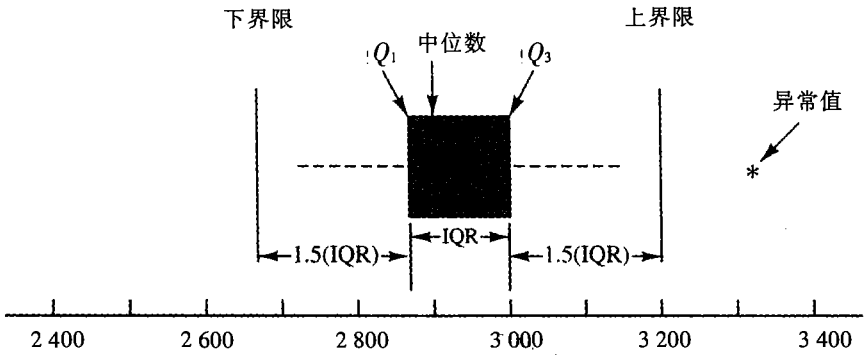


图 3.4　带有下界限和上界限的显示线条的起薪数据箱形图

在图 3.4 中，我们包括了显示界限位置的线条。画这些线是为了表示计算出的界限值是多少以及它们的位置怎样。虽然总是需要计算界限，但通常在箱形图中并不画出它们。图 3.5 是起薪数据的箱形图的一般形状。

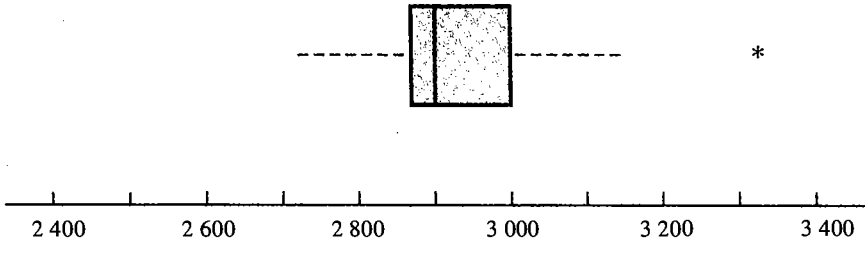


图 3.5　起薪数据的箱形图

评　注

1. 我们使用箱形图鉴别出的异常值与利用小于 -3 或大于 +3 的 z-分数挑选出的异常值可能相同也可能不同。不过，所有方法都是为了识别出那些应被再次检查以确保其有效性的数据项。对每种方法鉴别出的异常值都应该再进行检查。
2. 探索性数据分析的优点是它们易于使用，并且不需要很多数学计算。我们只需简单地把数据项按升序排列，并找到中位数和四分位数 Q_1 和 Q_3，就可以得到能够概括的五个数。我们能够很容易地确定界限和箱形图。该方法不需要计算数据的均值和标准差。
3. 在附录 3.1，我们将演示如何利用 Minitab 构建起薪数据的箱形图。这样得到的箱形图看起来很像图 3.5 中的那个，只是角度旋转了 90°。

练　习

方　法

38. 考虑容量为 8 的样本，其数据值分别为 27，25，20，15，30，34，28 和 25。作出数据的五数概括。

39. 作出练习 38 中数据的箱形图。

40. 给出以下数据：5，15，18，10，8，12，16，10 和 6。作出它们的五数概括和箱形图。

自测题

41. 数据集的第一四分位数是 42，第三四分位数是 50。计算其下界限和上界限。数据值 65 是否应被当做异常值？

应　用

42. 管理的目标是相对于投入的资本挣得尽可能多的利润。判断这种努力成功与否的一个尺度是权益收益率——即净收入与股东权益的比率。下面是 25 个公司的权益收益百分比（*Standard & Poor's Stock Reports*, November 1997）：

CD 光盘数据
ReEquity

9.0	19.6	22.9	41.6	11.4
15.8	52.7	17.3	12.3	5.1
17.3	31.1	9.6	8.6	11.2
12.8	12.2	14.5	9.2	16.6
5.0	30.3	14.7	19.2	6.2

a. 作出五数概括。

b. 计算下界限和上界限。

c. 有异常值显示出来吗？该信息对财务分析有什么帮助？

d. 作出箱形图。

43. 21 个制药公司的以百万美元为单位的年销售收入如下：



8 408	1 374	1 872	8 879	2 459	11 413
608	14 138	6 452	1 850	2 818	1 356
10 498	7 478	4 019	4 341	739	2 127
3 653	5 794	8 305			

a. 作出五数概括。

b. 计算下界限和上界限。

c. 有异常值显示出来吗？

d. 表中强生公司的销售额最大，为 14 138 百万美元。假设输入数据时发生误差（调换了数字位置），输入的数据是 41 138 百万美元。在(c)中用于检测异常值的方法能够发现问题并更正数据输入误差吗？

e. 作出箱形图。

44. 公司股票购买计划经常作为对股东的奖励而推出。但是，Robert Gabele——First Call/Thomson Financial 的内部研究主任，已经注意到在这些计划中有许多仅仅是作为对高级管理者的激励选择方案而执行的，目的是使他们能够获得股票。所有公司在 1998 年现存的股票期权占所有非流通普通股的 6.2%。下表数据显示了 15 个公司期权授予的股数和非流通股数。*Bloomberg* 期刊确认了这些公司需要重新购买的非流通股所占的百分比最高，才能满足授予期权的需要（*Bloomberg personal Finance*, January/February 2000）。



公司	期权授予的非流通股数（百万）	非流通普通股股数（百万）
Adobe Systems	20.3	61.8
Apple Computer	52.7	160.9
Applied Materials	109.1	375.4
Autodesk	15.7	58.9
Best Buy	44.2	203.8
Cendant	183.3	718.1
Dell Computer	720.8	2 540.9
Fruit of the Loom	14.2	66.9
ITT Industries	18.0	87.9
Merrill Lynch	89.9	365.5
Novell	120.2	335.0
Parametric Technology	78.3	269.3

（续表）

公司	期权授予的非流通股数（百万）	非流通普通股股数（百万）
Reebok International	12.8	56.1
Silicon Graphics	52.6	188.8
Toys R Us	54.8	247.6

a. 期权授予的非流通股数的均值和中位数是多少？

b. 期权授予的非流通股数的第一和第三四分位数是多少？

c. 对于期权授予的非流通股数，是否含有异常值？作出箱形图。

d. 计算期权授予的非流通股数与非流通的普通股数比例的百分比均值。该百分比与所报告的所有公司在 1998 年 6.2% 的百分比相比怎样？

45. 公路损失数据研究机构的伤害和碰撞损失实验报告根据事故后的保险索赔数字，对轿车型号进行评分。接近 100 的指数得分被认为是平均水平，较低的得分意味着更好、更安全的轿车型号。下列数据是 20 款中型轿车和 20 款小型轿车的得分：



CD 光盘数据 Injury

中型轿车：	81	91	93	127	68	81	60	51	58	75
	100	103	119	82	128	76	68	81	91	82
小型轿车：	73	100	127	100	124	103	119	108	109	113
	108	118	103	120	102	122	96	133	80	140

分别汇总中型和小型轿车的数据。

a. 分别作出中型和小型轿车的五数概括。

b. 作出箱形图。

c. 说明你的汇总里关于中型轿车和小型轿车安全性的比较结果。

46. Birinyi Associates 公司对全球股票市场进行了调查，以评估它们在 1997 年的表现。表 3.6 汇总了包括 30 个国家和地区的样本结果。



CD 光盘数据 World

a. 它们的变动百分比的均值和中位数是多少？

b. 第一和第三四分位数是多少？

c. 有异常值吗？作出箱形图。

d. 对美国市场，你报告的百分位数是多少？

表 3.6　全球股票市场 1997 年价值变动百分比

国家和地区	变动百分比	国家和地区	变动百分比
阿根廷	24.70	澳大利亚	7.91
巴林	49.67	巴巴多斯	48.29
百慕大	51.92	巴西	44.84
智利	12.80	哥伦比亚	69.60
克罗地亚	-1.07	捷克共和国	-3.25

（续表）

国家和地区	变动百分比	国家和地区	变动百分比
厄瓜多尔	5.37	爱沙尼亚	62.34
芬兰	32.31	德国	47.11
希腊	59.19	印度	18.60
以色列	27.91	日本	-21.19
立陶宛	16.82	墨西哥	54.92
纳米比亚	6.52	尼日利亚	-7.97
巴拿马	59.40	波兰	2.27
俄罗斯	125.89	斯洛文尼亚	18.71
斯里兰卡	15.49	台湾	18.08
土耳其	254.45	美国	22.64

资料来源：*The Wall Street Journal*, January 2, 1998。

3.5 两变量间关系的度量

到现在为止我们已经讲述了用来汇总**同一时间一个变量**数据的数值方法。但经理们或者决策者往往还对两变量间的关系感兴趣。在本节我们将讨论描述**两变量间关系**的参数：协方差和相关系数。

我们重新考虑2.4节中关于旧金山音像设备商店的例子。商店的经理想要调查在下列期间周末播出的电视广告次数和商店销售额之间的关系。表3.7中每星期有一个观察值($n=10$)，列出了以百美元为单位的销售额样本数据。图3.6的散点图显示出二者之间存在一种正的相关关系，较高的销售额(y)与较多的广告数(x)相联系。事实上，散点图提示我们可以用一条直线来大概描述这种线性关系。在下面的讨论中，我们将介绍**协方差**(covariance)，将其作为两变量间线性关系的描述性度量。

表3.7 音像设备商店的样本数据



CD光盘数据
Stereo

星期	广告数 x	销售额 y（100美元）
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

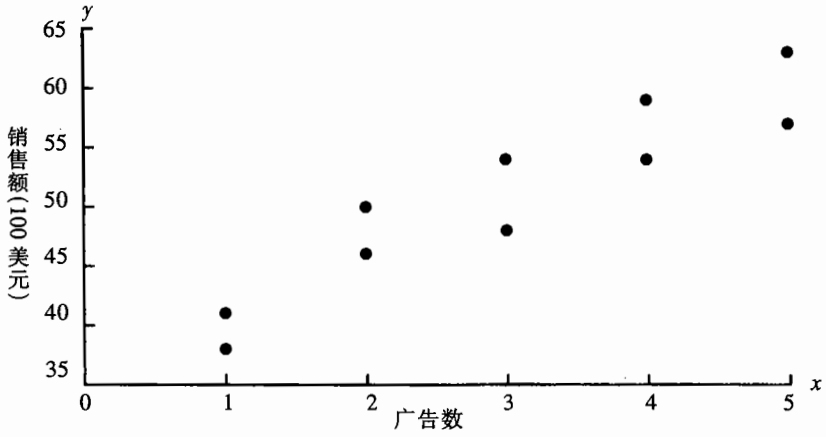


图 3.6　音像设备商店的散点图

协方差

对于容量为 n，观察值为 (x_1, y_1)，(x_2, y_2) 等的样本，**样本协方差**定义如下：

样本协方差

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.10}$$

该公式中，一个 x_i 与一个 y_i 相配对。我们把每个 x_i 与样本均值的离差乘以相对应的 y_i 与样本均值的离差，再把各个结果加总起来，然后除以 $n-1$。

为了度量在音像设备商店问题中广告次数 x 与销售额 y 之间线性关系的强度，我们用公式 (3.10) 来计算样本协方差。表 3.8 显示了 $\sum (x_i - \bar{x})(y_i - \bar{y})$ 的计算过程。注意到 $\bar{x} = 30/10 = 3$ 并

表 3.8　样本协方差的计算

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	-1	-1	1
	5	57	2	6	12
	1	41	-2	-10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	-2	-13	26
	5	63	2	12	24
	3	48	0	-3	0
	4	59	1	8	8
	2	46	-1	-5	5
总计	30	510	0	0	99

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

且$\bar{y}=510/10=51$。利用公式(3.10)，我们得到样本协方差为

$$s_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{99}{9}=11$$

计算容量为 N 的总体协方差的公式与公式(3.10)类似，但是我们使用不同的符号来表示我们正在处理的是完整的总体。

总体协方差

$$\sigma_{xy}=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{N} \tag{3.11}$$

在公式(3.11)中，我们使用符号 μ_x 表示变量 x 的总体均值，用符号 μ_y 表示变量 y 的总体均值。定义 σ_{xy} 为容量为 N 的总体的总体协方差。

协方差的解释

协方差是两变量间线性关系的度量。

为了有助于对样本协方差进行解释，我们参考图 3.7。它与图 3.6 的散点图相同，只是在$\bar{x}=3$ 处带有垂直虚线，在$\bar{y}=51$ 处带有水平虚线。这两条线把图分成四个部分：区域 I内的点对应着大于$\bar{x}$ 的 x_i 和大于$\bar{y}$ 的 y_i；区域 II内的点对应着小于$\bar{x}$ 的 x_i 和大于$\bar{y}$ 的 y_i，依此类推。因此，对于区域 I内的点，$(x_i-\bar{x})(y_i-\bar{y})$的值一定为正，区域 II内的点其值为负，区域 III 内的点其值又为正，而区域 IV 内的点其值为负。

如果 s_{xy} 的值是正值，那么对 s_{xy} 具有最大影响的点一定在区域 I和 III。因此，s_{xy} 为正值就表示 x 和 y 之间存在正线性关系。也就是说，当 x 的值增大时，y 的值也增大。如果 s_{xy} 的值为负，对它具有最大影响的点一定在区域 II和 IV。因此，S_{xy} 的负值就表明 x 和 y 之间存在负线性关系。也就是说，当 x 的值增大时，y 的值减小。最后，如果各点是均匀地分布在四个区域内，则 s_{xy} 的值将接近于零，表示 x 和 y 之间不存在线性关系。图 3.8 显示了与三种不同的散点图形式相

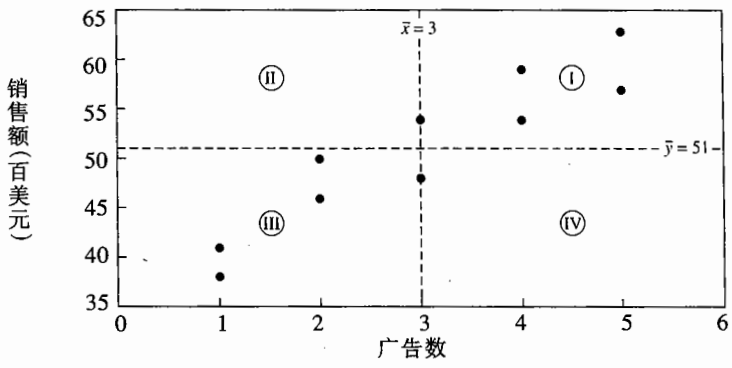


图 3.7 音像设备商店的分区散点图

对应的 s_{xy} 值。

再次参考图 3.7，我们看到音像设备的散点图属于图 3.8 顶部的那种类型，正如我们所期望的，样本协方差的值为正，等于 $s_{xy}=11$。

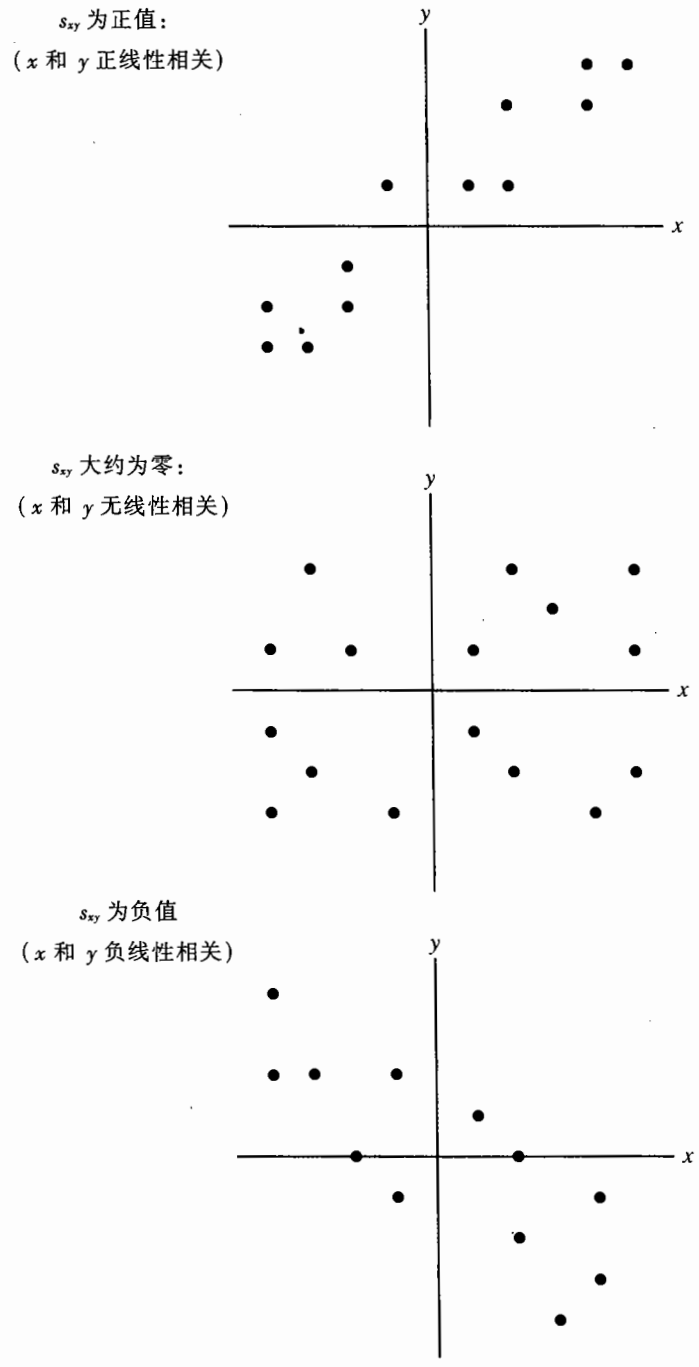


图 3.8　对样本协方差的解释

通过以上讨论，看起来如果协方差为大的正值，就表示存在强烈的正线性关系；如果协方差为大的负值，就表示存在强烈的负线性关系。不过，在使用协方差作为线性关系强度的度量时，存在的一个问题是我们得到的协方差值依赖于对 x 和 y 进行度量的单位。例如，假定我们对人们的身高 x 和体重 y 之间的关系感兴趣。很明显，不管我们使用英尺还是英寸来度量身高，二者关系的强度都应该相同。可是，当高度以英寸为单位时，对于 $(x_i-\bar{x})$ 的数值，我们的计算结果比采用英尺度量时要大得多。因此，高度单位用英寸时，公式(3.10)中的分子 $\sum(x_i-\bar{x})(y_i-\bar{y})$ 数值会更大，这样，协方差也就会更大。但实际上在这两种情况下二者关系并没有任何差别。使用**相关系数**(correlation coefficient)能够避免这种麻烦，并且也能够对两变量间的关系进行度量。

相关系数

对于样本数据，皮尔逊(Pearson)积矩相关系数的定义如下：

皮尔逊积矩相关系数：样本数据

$$r_{xy}=\frac{s_{xy}}{s_x s_y} \tag{3.12}$$

式中 r_{xy}——样本相关系数；

s_{xy}——样本协方差；

s_x——x 的样本标准差；

s_y——y 的样本标准差。

公式(3.12)表示样本数据的皮尔逊积矩相关系数（一般简称为**样本相关系数**）的计算方法是：x 的标准差和 y 的标准差的乘积与样本协方差相除。

现在让我们来计算音像设备商店的样本相关系数。使用表 3.7 的数据，我们能够得到两个变量的样本标准差

$$s_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{20}{9}}=1.490\ 7$$

$$s_y=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}=\sqrt{\frac{566}{9}}=7.930\ 3$$

现在，因为 $s_{xy}=11$，我们得到样本相关系数为

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{11}{(1.490\ 7)(7.930\ 3)}=0.93$$

总体相关系数用希腊字母表示，计算公式如下：

皮尔逊积矩相关系数：总体数据

$$\rho_{xy}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{3.13}$$

式中　ρ_{xy}——总体相关系数；
σ_{xy}——总体协方差；
σ_x——x 的总体标准差；
σ_y——y 的总体标准差。

样本相关系数 r_{xy} 是对总体相关系数 ρ_{xy} 的一个估计。

相关系数的解释

首先让我们考虑一个简单的样本，以说明完全正线性关系的概念。图 3.9 的散点图描述了下列样本数据 x 和 y 之间的关系：

x_i	y_i
5	10
10	30
15	50

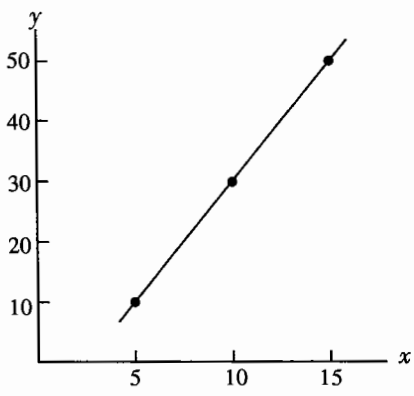


图 3.9　描述完全正线性关系的散点图

经过这三个点的直线显示了在 x 和 y 之间存在着完全线性关系。为了应用公式(3.12)计算样本相关系数，我们必须首先计算 s_{xy}，s_x 和 s_y。一些必要的计算过程列示在表 3.9 中。使用表 3.9 的结果，我们有

$$s_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{200}{2}=100$$

$$s_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{50}{2}}=5$$

$$s_y=\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}=\sqrt{\frac{800}{2}}=20$$

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{100}{5\times 20}=1$$

因此，我们看到样本相关系数的值为 1。

表 3.9 计算样本相关系数的计算过程

	x_i	y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
总计	30	90	0	50	0	800	200
	$\bar{x}=10$	$\bar{y}=30$					

一般地，我们发现如果数据集内所有的点都在一条正斜率直线上的话，则样本相关系数的值为 +1。也就是说，样本相关系数 +1 对应于 x 和 y 的完全正线性关系。如果数据集内所有的点都在一条负斜率直线上的话，则样本相关系数的值为 −1。即样本相关系数 −1 对应于 x 和 y 的完全负线性关系。

现在，我们假定某个数据集表明 x 和 y 之间存在着正线性关系，但不是完全线性。那么，r_{xy} 的值将小于 1，在散点图上就表现为所有的点不在一条直线上。当数据集内的点越来越偏离完全正线性关系时，r_{xy} 的值就变得越来越小。当 r_{xy} 等于零时就表示 x 和 y 之间不存在线性关系，r_{xy} 的值接近于零则表示弱线性关系。

相关系数的范围是从 −1 到 +1。接近于 −1 或 +1 的值表示强的线性关系，而相关系数与零越接近，线性关系也越弱。

对于音像设备商店数据集，我们已知 $r_{xy}=+0.93$。因此，广告数与销售额之间存在着强正线性关系。更具体地说，广告数的增长与销售额的增长相关联。

练 习

方 法

47. 两变量的五个观察值如下：



x_i	4	6	11	3	16
y_i	50	50	40	60	30

a. 构建以 x 为横轴的散点图。

b. (a)中作出的散点图表明两变量间存在着什么关系？

c. 计算并解释数据的样本协方差。

d. 计算并解释数据的样本相关系数。

48. 两变量的五个观察值如下：

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a. 构建这些数据的散点图。

b. 散点图显示 x 和 y 之间可能存在什么关系？

c. 计算并解释数据的样本协方差。

d. 计算并解释数据的样本相关系数。

应　用

49. 高中指导委员会收集了 6 个高年级学生平均等级点(GPA)和 SAT 数学考试成绩的数据如下：

GPA	2.7	3.5	3.7	3.3	3.6	3.0
SAT	450	560	700	620	640	570

a. 作出以 GPA 为横轴的散点图。

b. 从散点图来看，GPA 和 SAT 数学考试成绩之间存在关系吗？请解释。

c. 计算并解释数据的样本协方差。

d. 计算数据的样本相关系数。它说明两变量间的关系是什么？

50. 运输部门对于中型汽车的行驶速度和每英里油耗进行了研究，结果数据如下：

行驶速度	30	50	40	55	30	25	60	25	50	55
每英里油耗	28	25	25	23	30	32	21	35	26	25

计算并解释这些数据的样本相关系数。

51. 《PC 世界》提供了销量最大的 15 种笔记本电脑的评分（*PC World*, February 2000）。运行分数是对于该电脑在一系列普通商业应用中，与基准样机相比较其运转速度有多快的度量。例如，运行分数为 200 的电脑的运转速度是样机的两倍快。在研究中测试的每台电脑，其综合评分的满分为 100。90 分以上的成绩意味着优异，而 70 多分就表示高于平均水平。15 种笔记本的运行分数和综合评分如下：



CD 光盘数据 PCs

笔记本	运行分数	综合评分
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76

（续表）

笔记本	运行分数	综合评分
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavillion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. 计算样本相关系数。

b. 样本相关系数说明运行分数和综合评分之间存在着什么关系？

52. 道琼斯工业平均指数(DJIA)和标准普尔500指数（S&P）都用做对股市综合变动的度量。DJIA建立在30家巨型公司的价格变动基础上；S&P 500则是一个由500只股票构成的指数。有些人认为S&P 500对股市表现度量得更好，因为它的基础更广泛。下表列出了从2000年2月11日以来10个星期的DJIA和S&P 500的收盘价（*Barron's*, April 17, 2000）。


CD光盘数据
Dow S&P

日期	道琼斯指数	S&P500指数
2月11日	10 425	1 387
2月18日	10 220	1 346
2月25日	9 862	1 333
3月3日	10 367	1 409
3月10日	9 929	1 395
3月17日	10 595	1 464
3月24日	11 113	1 527
3月31日	10 922	1 499
4月 7日	11 111	1 516
4月14日	10 306	1 357

a. 计算这些数据的样本相关系数。

b. 两指数的相关程度是弱还是强？

53. 20个城市的日最高和最低温度如下（*USA Today*, May 9, 2000）：


CD光盘数据
HighLow

城市	最高温度	最低温度
雅典	75	54
曼谷	92	74
开罗	84	57
哥本哈根	64	39

（续表）

城市	最高温度	最低温度
都柏林	64	46
哈瓦那	86	68
香港	81	72
约翰内斯堡	61	50
伦敦	73	48
马尼拉	93	75
墨尔本	66	50
蒙特利尔	64	52
巴黎	77	55
里约热内卢	80	61
罗马	81	54
汉城	64	50
新加坡	90	75
悉尼	68	55
东京	79	59
温哥华	57	43

最高和最低温度间的关系是什么？

3.6　加权平均值和使用分组数据

在 3.1 节，我们对均值进行了介绍，它是最重要的描述性统计的度量方法之一。下面是具有 n 个观察值的样本均值公式：

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.14}$$

在该公式中，给予了每个 x_i 同等的重要性和权重。虽然在实践中这种形式最常见，但在某些情况下，需要给予每个观察值一个对应于其重要性的权重，来计算平均值。以这种方法计算出的平均值称为**加权平均值**(weighted mean)。

加权平均值

加权平均值可计算如下：

加权平均值

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3.15}$$

式中　x_i——观察值 i 的数值；

w_i——观察值 i 的权重。

当数据是来自样本时，公式(3.15) 算出的就是加权样本平均值；当数据来自总体，则用 μ 来代替$\bar{x}$，利用公式(3.15)算出的是加权总体平均值。

作为一个需要进行加权平均的样本，我们来考虑下面的一个过去 3 个月以来 5 次原材料购买的样本：

购买次数	每磅成本(美元)	磅数
1	3.00	1 200
2	3.40	500
3	2.80	2 500
4	2.90	1 000
5	3.25	800

注意到每磅成本从 2.80 美元—3.40 美元不等，购买数量在 500—2 500 磅之间。假定经理想要知道原材料的每磅成本平均值的有关信息。因为每次订购的数量不同，我们必须使用加权平均值公式。5 个每磅成本数据值分别是 $x_1 = 3.00$，$x_2 = 3.40$，$x_3 = 2.80$，$x_4 = 2.90$ 和 $x_5 = 3.25$。计算每磅成本的平均值时，需要根据每个成本对应的数量来确定每个成本的权数。在这个例子中，权数分别为 $w_1 = 1\,200$，$w_2 = 500$，$w_3 = 2\,500$，$w_4 = 1\,000$ 和 $w_5 = 800$。利用公式(3.15)，可计算加权平均值如下：

$$\bar{x} = \frac{1\,200 \times 3.00 + 500 \times 3.40 + 2\,500 \times 2.80 + 1\,000 \times 2.90 + 800 \times 3.25}{1\,200 + 500 + 2\,500 + 1\,000 + 800}$$

$$= \frac{17\,800}{6\,000} = 2.967$$

因此，加权平均值的计算结果表明，每磅原材料成本的平均值是 2.967 美元。如果我们使用公式(3.14)而不是加权平均公式的话，将会得到带有误导性的结果。本例中，5 个每磅成本的平均值是 $(3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = 3.07$ 美元，过高估计了实际的每磅成本平均值。

在特定的加权平均值计算过程中，对权重的选择依赖于具体的应用。一个被大学生们所熟知的例子是平均等级(GPA)的计算过程。在计算时，被普遍使用的数据值是：4 代表 A 级，3 代表 B 级，2 代表 C 级，1 代表 D 级，0 代表 F 级。权重就是每一级获得的学分数。本节后面的练习 56 提供了该加权平均值计算过程的示例。在其他加权平均值的计算中，经常使用诸如磅、美元数、体积等数量值作为权重。在任何情况下，当数据的重要性变化时，分析人员必须选择能够最好地反映每个数据值重要程度的权重，以确定平均值。

计算等级分平均值是一个很好的加权平均值应用例子。

分组数据

在大多数情况下，对位置和变异程度的度量是利用单个数据值来计算的。但有时我们只有分组的或频数分布形式的数据。在下面的讨论中，我们将说明如何利用加权平均值公式得到**分组数据**(grouped data)的均值、方差和标准差的大概值。

在 2.2 节，我们曾经介绍了 Sanderson and Clifford 会计师事物所完成年末审计所需要天数的频数分布。表 3.10 再次列出了该事物所的 20 家客户样本的审计时间频数分布。在频数分布的基础上，样本平均审计时间是多少？

表 3.10　审计时间的频数分布

审计时间（天）	频数
10—14	4
15—19	8
20—24	5
25—29	2
30—34	1
总计	20

为了只使用分组数据来计算均值，我们把每一组的组中值作为该组所有数据的代表。令 M_i 表示 i 组的组中值；令 f_i 表示 i 组的频数。然后我们就能够用 M_i 表示数据值，用频数 f_i 代表权重，从而可以使用加权平均值公式（3.15）。在本例中，公式（3.15）的分母是频数之和，即样本容量 n。也就是 $\sum f_i = n$。因此，分组数据的样本均值公式可表达如下：

分组数据的样本均值

$$\bar{x} = \frac{\sum f_i M_i}{n} \tag{3.16}$$

式中　M_i——i 组的组中值；

f_i——i 组的频数；

n——$\sum f_i$ = 样本容量。

对于组中值 M_i，两组限距离的一半，我们可以得到表 3.10 的第 1 组 10—14 的组中值为 $(10+14)/2 = 12$。5 个组的组中值以及审计时间的加权平均过程都汇总在表 3.11 中。我们可以看到，样本平均审计时间是 19 天。

表 3.11　分组数据样本加权平均审计时间的计算过程

审计时间（天）	组中值（M_i）	频数（f_i）	f_iM_i
10 — 14	12	4	48
15 — 19	17	8	136
20 — 24	22	5	110
25 — 29	27	2	54
30 — 34	32	1	32
		20	380

样本均值 $\bar{x} = \dfrac{\sum f_i M_i}{n} = \dfrac{380}{20} = 19$ 天

为了计算分组数据的方差，我们对方差公式（3.5）做一些轻微改动。在公式（3.5）中，关于样本均值$\bar{x}$的数据平方离差被写成$(x_i-\bar{x})^2$。但是在分组数据的情况下，该值是未知的。在本例中，我们把组中值M_i作为该组的x_i值的代表。因此关于样本均值的平方离差$(x_i-\bar{x})^2$，被$(M_i-\bar{x})^2$替换。然后，与我们在分组数据的样本均值计算中所做的处理相同，我们用组频数f_i作为每个值的权重。所有数据关于均值的平方离差之和就近似为$\sum f_i(M_i-\bar{x})^2$。分母使用$n-1$而不是n是为了以样本方差来估计总体方差σ^2。因此，利用下面公式可得到分组数据的样本方差：

分组数据的样本方差

$$s^2=\frac{\sum f_i(M_i-\bar{x})^2}{n-1} \tag{3.17}$$

基于表3.10分组数据审计时间样本的计算过程列示在表3.12中。

表3.12　分组数据审计时间样本方差的计算过程（样本均值$\bar{x}=19$）

审计时间（天）	组中值（M_i）	频数（f_i）	离差（$M_i-\bar{x}$）	平方离差（$M_i-\bar{x}$）^{2}	$f_i(M_i-\bar{x})^2$
10—14	12	4	-7	49	196
15—19	17	8	-2	4	32
20—24	22	5	3	9	45
25—29	27	2	8	64	128
30—34	32	1	13	169	169
		20			570
					$\sum f_i(M_i-\bar{x})^2$

样本方差 $s^2=\frac{\sum f_i(M_i-\bar{x})^2}{n-1}=\frac{570}{19}=30$

分组数据的标准差就是分组数据方差的正平方根。对于审计时间数据，样本标准差是$s=\sqrt{30}=5.48$。

在结束本节对分组数据的位置和分散程度的度量计算之前，我们注意到公式(3.16)和(3.17)是关于样本的。对总体的度量值计算方法也与此类似。分组数据总体均值和方差的分式如下：

分组数据的总体均值

$$\mu=\frac{\sum f_iM_i}{N} \tag{3.18}$$

分组数据的总体方差

$$\sigma^2=\frac{\sum f_i(M_i-\mu)^2}{N} \tag{3.19}$$

评　注

1. 在计算分组数据的样本方差时，另一个可选的公式为 $s^2=\dfrac{\sum f_i M_i^2-n\bar{x}^2}{n-1}$。式中，$\sum f_i M_i^2=f_1M_1^2+f_2M_2^2+\cdots+f_kM_k^2$，$k$ 是数据分组的组号。利用该公式可以稍微简化计算。
2. 在计算分组数据的描述性统计量时，使用组中值近似代替每组的数据值。结果导致了分组数据的描述性统计量只是直接利用原始数据得出的描述性统计量的近似值。因此，我们建议在可能的情况下，应直接使用原始数据而不是分组数据来计算描述性统计量。

练　习

方　法

54. 考虑下列数据和对应的权重：

x_i	权重（w_i）
3.2	6
2.0	3
2.5	2
5.0	8

a. 计算数据的加权平均值。

b. 计算不计权重的 4 个数据值的样本均值，注意这两种方法的结果差异。

55. 考虑下面频数分布的样本数据：



组别	组中值	频数
3 — 7	5	4
8 — 12	10	7
13 — 17	15	9
18 — 22	20	5

a. 计算样本均值。

b. 计算样本方差和样本标准差。

应　用

56. 

大学生的等级分平均值是通过计算加权平均值得到的。在大多数的大学中，用下列数据值来分别代表各个等级：A(4)，B(3)，C(2)，D(1)，F(0)。州立大学的某位学生在结束了60个学分的课程以后，有9个学分获得了A，15个学分获得了B，33个学分获得了C，3个学分得了D。

a. 计算这位学生的等级分平均值。

b. 州立大学的学生们必须在他们前60个学分的课程中，取得2.5个等级分平均值才能够被商学院录取。这名学生能被录取吗？

57. 红利收益率是作为年度红利支付给股东的收入占股票每股价值的百分比。一个样本由Innis投资公司拥有的8只股票组成，其红利收益率如下（*Barron's*, January 5, 1998）。Innis公司投资于每只股票上的金额也在下表列出，问该投资组合的平均红利收益率是多少？

公司	红利收益率	投资金额(美元)
Apple Computer	0.00	37 830
Chevron Corp.	2.98	27 667
Eastman Kodak	2.77	31 037
ExxonMobil	2.65	27 336
Merck & Co.	1.58	37 553
Franklin Resources	0.57	17 812
Sears	2.00	32 660
Woolworth	0.00	17 775

58. 一个样本由680辆汽车组成，某加油站记录了卖给其中每辆汽车的汽油数量的频数分布。

汽油（加仑）	频数
0—4	74
5—9	192
10—14	280
15—19	105
20—24	23
25—29	6
总计	680

计算这些分组数据的均值、方差和标准差。如果该加油站希望在某个给定的日期为120辆汽车加油，那么它能够销售的汽油总数的估计值是多少？

59. 在《财富》杂志的订户调查中，有这样一个问题："你阅读过或者浏览过最后四篇中的几篇？"假定下列频数分布汇总了500个答案：

阅读数目	频数
0	15
1	10
2	40
3	85
4	350
总计	500

a.《财富》订户阅读后四篇文章数的均值是多少？

b. 他们阅读后四篇文章数的标准差是多少？

小　结

本章我们介绍了几种能够概括数据的位置和变异程度的描述性统计量。与表格和图形方法不同，这几种度量方法是按照数值概括数据。当数值是由样本得来时，就称为样本统计量；而当数值来自于总体时，则称为总体参数。表示样本统计量和总体参数的一些符号如下所示：

	样本统计量	总体参数
均值	$\bar{x}$	μ
方差	s^2	σ^2
标准差	s	σ
协方差	s_{xy}	σ_{xy}
相关系数	r_{xy}	ρ_{xy}

作为对数据集中心位置的度量，我们定义了均值、中位数和众数。而百分位数的概念用来描述数据集的其他位置。接着，我们给出了极差、四分位点内距、方差、标准差和变异系数来度量变异程度或分散程度。我们还介绍了在应用经验法则和切比雪夫定理时，怎样一起使用均值和标准差，对数据的分布提供更多的信息，并鉴别出异常值。

在 3.4 节，我们演示了怎样构建五数概括法和箱形图。它们对数据的位置、变异程度和分布的形状提供了类似的信息。在 3.5 节，我们介绍了作为两变量间关系度量的协方差和相关系数。而在最后一节，我们说明了如何计算加权平均值和如何计算分组数据的均值、方差、标准差。

我们还能够利用统计软件包和分布表来构建本章讨论的描述性统计量。在附录 3.1 内，我们将演示怎样利用 Minitab 得到本章介绍过的大多数描述性统计量；在附录 3.2 内，我们将以 Excel 的应用为例说明同样的目的。

术语释义

样本统计量：用来汇总样本（例如，样本均值、样本方差 s^2 和样本标准差 s）度量结果的数值。

总体参数：用来汇总数据总体（例如，总体均值 μ、总体方差 σ^2 和总体标准差 σ）度量结果的数值。

均值：对一个数据集的中心位置的度量。它的计算方法是先加总所有数据值，再除以数据的个数。

中位数：对中心位置的度量。它是当数据按照升序排列时，处于数据集中间位置的数据的值。

众数：对中心位置的度量，它是数据集中出现频率最高的数据的值。

百分位数：它是这样一个数值，在数据集中至少有 $p\%$ 的观察值小于或等于该值，至少有 $(100-p)\%$ 的观察值大于或等于该值。而第 50 百分位数就是中位数。

四分位数：第 25、第 50 和第 75 百分位数分别是第一四分位数、第二四分位数和第三四分位数。四分位数能够把数据集分成四个部分，每部分都包含大约 25% 的数据。

临界值：下临界值大约等于第一四分位数，或第 25 百分位数；而上临界值大约等于第三四分位数，或第 75 百分位数。因为计算习惯不同，临界值可能与四分位数有轻微的差别，但它们的目标都是把数据集分割成相等的四部分。

极差：对数据集变异程度的度量，它的定义是最大值与最小值的差。

四分位点内距（IQR）：对数据集变异程度的度量，被定义为第三和第一四分位数之差。

方差：对数据集变异程度进行的度量，有关计算是建立在各个数据值关于均值的平方离差之和的基础上。

标准差：对方差取正平方根所得到的对数据集变异程度的度量。

变异系数：它是对数据集相对变异程度的度量，计算方法是把标准差除以均值再乘以 100。

z－分数：用关于均值的离差除以标准差 s 得到的数值。z－分数也被称为标准化值，表示 x_i 与平均值的距离相当于标准差的个数。

切比雪夫定理：该定理可应用于任何数据集，而无论它们的分布状况如何。该定理能够指出那些与平均值的距离必定处于某特定标准差个数以内的观察值占总数的比例。

经验法则：该法则适用于具有钟形分布的数据集，可用于计算那些距离数据集平均值必定在一个、两个和三个标准差以内的数据项占总数的百分比。

异常值：小得不正常或大得不正常的数据值。

五数概括法：一种探索性数据分析技术，它用五个数来概括数据：最小值、第一四分位数、中位数、第三四分位数和最大值。

箱形图：数据的一种图形概括。其箱体边缘从第一四分位数画到第三四分位数，表示了数据的中间 50% 位置。从箱体边缘发出的横虚线被称为须线，它表示大于第三四分位数和小于第一四分位数的数据值位置。在任何异常值所处的位置都做了标记。

协方差：对两变量间线性关系的数值度量。正值表示正的线性关系；负值表示负的线性关系。

相关系数：对两变量间线性关系的数值度量，其取值范围从 -1 到 $+1$。接近 $+1$ 的值表示强的正线性关系；接近 -1 的值表示强的负线性关系；而接近于零的值则表示缺乏线性关系。

加权平均值： 通过给每个观察值分配一个反映它在数据集内重要性的权重，从而计算得出的数据集平均值。

分组数据： 在该形式下，人们只能使用由频数分布汇总的组区间数据，而无法得到原始数据的单个值。

重要公式

样本均值

$$\bar{x} = \frac{\sum x_i}{n} \tag{3.1}$$

总体均值

$$\mu = \frac{\sum x_i}{N} \tag{3.2}$$

四分位点内距

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

总体方差

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3.4}$$

样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \tag{3.5}$$

标准差

$$\text{样本标准差} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{总体标准差} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

变异系数

$$\frac{\text{标准差}}{\text{均值}} \times 100 \tag{3.8}$$

z - 分数

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

样本协方差

$$s_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} \tag{3.10}$$

皮尔逊积矩相关系数：样本数据

$$r_{xy}=\frac{s_{xy}}{s_x s_y} \tag{3.12}$$

加权平均值

$$\bar{x}=\frac{\sum w_i x_i}{\sum w_i} \tag{3.15}$$

式中，w_i 为观察值 i 的权重。

分组数据的样本均值

$$\bar{x}=\frac{\sum f_i M_i}{n} \tag{3.16}$$

分组数据的样本方差

$$s^2=\frac{\sum f_i(M_i-\bar{x})^2}{n-1} \tag{3.17}$$

分组数据的总体均值

$$\mu=\frac{\sum f_i M_i}{N} \tag{3.18}$$

分组数据的总体方差

$$\sigma^2=\frac{\sum f_i(M_i-\mu)^2}{N} \tag{3.19}$$

补充练习

60. 美国人平均每月外出吃饭花去 65.88 美元（*The Des Moines Register*, December 5, 1997）。一个由年轻成年人组成的样本提供了过去几个月他们外出就餐的花费（美元）数据。



CD 光盘数据
Eat

253	101	245	467	131	0	225
80	113	69	198	95	129	124
11	178	104	161	0	118	151
55	152	134	169			

a. 计算均值、中位数和众数。

b. 考虑(a)中的结果，这些年轻人在外出就餐的花费上，是否与美国人的平均水平相同？

c. 计算第一和第三四分位数。

d. 计算极差和四分位点内距。

e. 计算方差和标准差。

f. 里面含有异常值吗？

61. 美国 100 家最大上市公司董事会成员的每年总收入中有一部分是现金收入，即他们为董事会服务的年收入。除了现金收入外，董事会成员还可能收到股票收入、股票赠与、股票期权和董事会会议津贴。即使每年的现金收入低至 15 000 美元，但他们的总收入仍能够轻易地超过 100 000 美元。下列数据显示了在一个由 20 家美国最大上市公司组成的样本中，董事会成员的现金收入(*USA Today*, April 17, 2000)。



CD 光盘数据
Retainer

公司	现金收入
American Express	64
Bank of America	36
Boeing	26
Chevron	35
Dell Computer	40
DuPont	35
ExxonMobil	40
Ford Motor	30
General Motors	60
International Paper	36
Kroger	28
Lucent Technologies	50
Motorola	20
Procter & Gamble	55
Raytheon	40
Sears Roebuck	30
Texaco	15
United Parcel Service	55
Wal-Mart Stores	25
Xerox	40

计算下列描述性统计量：

a. 均值、中位数和众数。

b. 第一和第三四分位数。

c. 极差和四分位点内距。

d. 方差和标准差。

e. 变异系数。

62. 有一项调查是关于计算机厂商迅速处理问题的能力的(*PC Computing*, November 1997)。其结果如下:

公司	解决问题的天数	公司	解决问题的天数
Compaq	13	Gateway	21
Packard Bell	27	Digital	27
Quantex	11	IBM	12
Dell	14	Hewlett-Packard	14
NEC	14	AT&T	20
AST	17	Toshiba	37
Acer	16	Micron	17

a. 解决问题所需天数的均值和中位数是多少?

b. 方差和标准差是多少?

c. 哪一个厂商拥有最好的记录?

d. Packard Bell 公司的 z - 分数是多少?

e. IBM 公司的 z - 分数是多少?

f. 有异常值吗?

63. 下面的数据是 Westwood 储蓄和贷款协会某贷款人员办理的家庭抵押贷款数额,数据以千美元为单位。



CD 光盘数据
Mortgage

52.0	68.5	63.0	57.5	64.0	42.5	55.9	73.2	67.5	66.2
55.2	60.9	53.8	58.4	43.0	61.0	63.5	55.4	63.5	50.2
69.0	68.1	60.5	75.5	60.5	82.0	70.5	81.6	72.5	74.8

a. 计算均值、中位数和众数。

b. 计算第一和第三四分位数。

64. 根据 Forrester 研究公司的调查,大约有 19% 的互联网用户进行在线游戏。下列数据表示了 10 家游戏网站 3 月份的个人访问者数量(以千为单位)(*The Wall Street Journal*, April 17, 2000)。

网站	个人访问者
AOLGames. aol	9 416
extremelotto. com	3 955
freelotto. com	12 901
gamesville. com	4 844
iwin. com	7 410
prizecentral. com	4 899
shockwave. com	5 582
speedyclick. com	6 628
uproar. com	8 821
webstakes. com	7 499

利用这些数据，计算均值、中位数、方差和标准差。

65. 一个包括 20 个城市的样本列出了这些城市典型的家庭收入(*Places Rated Almanac*, 2000)，数据的单位是千美元。


CD 光盘数据
Income

a. 计算样本数据的均值和标准差。

b. 利用(a)中计算出的均值和标准差，估计所有城市总体家庭收入的均值和标准差。使用切比雪夫定理来确定 75% 的城市总体家庭收入一定会落在哪个范围内。

c. 假定家庭收入的分布呈钟形。利用(a)中计算出的均值和标准差来估计城市总体家庭收入的均值和标准差。利用经验法则确定 95% 的城市总体家庭收入一定会落在哪个范围内，并把它和(b)中的结果进行比较。

d. 样本数据中是否包含异常值？

城市	收入
Akron, OH	74.1
Atlanta, GA	82.4
Birmingham, AL	71.2
Bismark, ND	62.8
Cleveland, OH	79.2
Columbia, SC	66.8
Danbury, CT	132.3
Denver, CO	82.6
Detroit, MI	85.3
Fort Lauderdale, FL	75.8
Hartford, CT	89.1
Lancaster, PA	75.2
Madison, WI	78.8
Naples, FL	100.0
Nashville, TN	77.3
Philadelphia, PA	87.0
Savannah, GA	67.8
Toledo, OH	71.2
Trenton, NJ	106.4
Washington, DC	97.4

66. 雇员们每天去上班时，能够使用公交工具和私人汽车两种方法。下面是一个记录了每种方法所花费时间的样本，时间的单位是分钟。

公交工具：	28	29	32	37	33	25	29	32	41	34
私人汽车：	29	31	33	32	34	30	31	32	35	33

a. 计算每种方法花费时间的样本平均值。

b. 计算每种方法花费时间的样本标准差。

c. 在前两问结果的基础上，人们应该选择哪种交通方法？请解释。

d. 构建每种方法的箱形图。你作出的箱形图支持你在(c)中的结论吗？

67. 25个统计学专业学生的期终考试成绩如下：


CD光盘数据
Exam

56	77	84	82	42	61	44	95	98	84
93	62	96	78	88	58	62	79	85	89
89	97	53	76	75					

a. 作出五数概括。

b. 作出箱形图。

68. 下面的数据是在NCAA大学橄榄球联赛的一个赛季中，由20个防守队员组成的样本的累计奔跑码数。

744	652	576	1 112	971	451	1 023	852	809	596
941	975	400	711	1 174	1 278	820	511	907	1 251

a. 作出五数概括。

b. 作出箱形图。

c. 识别异常值。

69. 20个城市样本的典型家庭收入和家庭物价水平如下（*Places Rated Almanac*, 2000），数据的单位是千美元。


CD光盘数据
Cities

城市	收入	家庭物价水平
Bismark, ND	62.8	92.8
Columbia, SC	66.8	116.7
Savannah, GA	67.8	108.1
Birmingham, AL	71.2	130.9
Toledo, OH	71.2	101.1
Akron, OH	74.1	114.9
Lancaster, PA	75.2	125.9
Fort Lauderdale, FL	75.8	145.3
Nashville, TN	77.3	125.9
Madison, WI	78.8	145.2
Cleveland, OH	79.2	135.8
Atlanta, GA	82.4	126.9
Denver, CO	82.6	161.9
Detroit, MI	85.3	145.0
Philadelphia, PA	87.0	151.5
Hartford, CT	89.1	162.1

（续表）

城市	收入	家庭物价水平
Washington, DC	97.4	191.9
Naples, FL	100.0	173.6
Trenton, NJ	106.4	168.1
Danbury, CT	132.3	234.1

a. 样本协方差是多少？它表示正的还是负的线性关系？

b. 样本相关系数是多少？

70. *Road & Track* 杂志提供了下列汽车轮胎的装载能力和轮胎得分样本：

轮胎得分	装载能力
75	853
82	1 047
85	1 135
87	1 201
88	1 235
91	1 356
92	1 389
93	1 433
105	2 039

a. 构建以轮胎得分为 x 轴的数据散点图。

b. 样本相关系数是多少？它说明在轮胎得分和装载能力之间存在什么关系？

71. 在练习 6 中，我们对于打折经纪人从事的两类交易：以每股 50 美元买卖 500 股和以每股 5 美元买卖 1 000 股，计算了有关的几种描述性统计量。表 3.2 列出了在 20 个打折经纪人的样本中，进行每种交易分别收取的佣金（*AAII Journal*, January 1997）。计算两类交易的协方差和相关系数。你对于两者的关系可以了解到什么？

72. 下面数据显示了 10 家公司报告的道琼斯股票连续追踪 52 周的优先股收益和账面价值（*The Wall Street Journal*, March 13, 2000）。

公司	账面价值	收益
Am Elec	25.21	2.69
Columbia En	23.2	3.01
Con Ed	25.19	3.13
Duke Energy	20.17	2.25
Edison Int'l	13.55	1.79
Enron Cp.	7.44	1.27
Peco	13.61	3.15
Pub Sv Ent	21.86	3.29

（续表）

公司	账面价值	收益
Southn Co.	8.77	1.86
Unicom	23.22	2.74

a. 构建以账面价值为 x 轴的数据散点图。

b. 样本相关系数是多少？它说明每股收益和账面价值之间存在什么关系？

73. 一个样本由5只货币市场基金组成，它们的到期天数如下所示。另外，还给出了投资于基金的美元数额。利用加权平均值来确定这5支基金的投资平均到期天数是多少？

到期天数	投资金额（百万美元）
20	20
12	30
7	10
5	15
6	10

74. 一种被称为移动平均的预测技术利用时间序列数据最近 n 期的平均值来预测下一期的值。3期移动平均预测值是以最近3期数据来做预测计算。考虑这样一种产品，它在本年度前3个月的需求量分别为：1月(800单位)，2月(750单位)，3月(900单位)。

a. 4月份的3个月移动平均预测值是多少？

b. 这种预测技术的一种变体被称为加权移动平均。该技术在预测计算过程中，对更近的时间序列数据给予了更大的权重。例如，加权3个月的移动平均可以把权重3给予1个月长度的数据，把权重2给予2个月长度的数据，把权重1给予3个月长度的数据。使用以上数据来计算4月份的3月期加权移动平均预测值。

75. 下面是20个长途电话以分计的通话时间频数分布，计算数据的均值、方差和标准差。

通话时间（分）	频数
4—7	4
8—11	5
12—15	7
16—19	2
20—23	1
24—27	1
总计	20

76. La Maison 法国餐馆晚餐账单金额的频数分布如下所示，计算数据的均值、方差和标准差。

账单金额（美元）	频数
25—34	2
35—44	6

（续表）

账单金额（美元）	频数
45—54	4
55—64	4
65—74	2
75—84	2
总计	20

77. 在纽约州高速公路行驶的汽车会受到该州警察雷达系统的速度检测。下面是速度的频数分布：

速度（英里/小时）	频数
45—49	10
50—54	40
55—59	150
60—64	175
65—69	75
70—74	15
75—79	10
总计	475

a. 在纽约州高速公路行驶的汽车平均速度为多少？

b. 计算速度的方差和标准差。

案例研究 1　CONSOLIDATED 食品公司

Consolidated 食品公司在新墨西哥州、亚利桑那州和加利福尼亚州经营连锁超市（见第 2 章案例）。表 3.13 列出了一个 100 位顾客样本的支付方式和付款金额数据的一部分。该公司的经理希望能够通过样本了解到商店顾客的支付情况，尤其很想知道新的信用卡支付的选择方式与顾客的购货金额之间有什么关系。

管理报告

使用第 3 章讲述的描述性统计方法来汇总样本数据。分别作出现金客户、个人支票客户和信用卡客户的购货金额汇总。你的报告应包括下列汇总和讨论：

1. 每种支付方式数据的均值和中位数的比较与理解。
2. 对每种支付方式数据的变异程度的不同度量方式——如极差和标准差的比较和理解。
3. 每种支付方式数据的五数概括，以及对它们的识别和理解。
4. 每种支付方式数据的箱形图。

利用报告的汇总结果，讨论你对于该公司客户的支付方式和付款金额能够得出什么结论。

表 3.13　100 位公司顾客样本的购货金额和支付方式


CD 光盘数据
Consolid

顾客	购货金额（美元）	支付方式
1	28.58	支票
2	52.04	支票
3	7.41	现金
4	11.17	现金
5	43.79	信用卡
6	48.95	支票
7	57.59	支票
8	27.60	支票
9	26.91	信用卡
10	9.00	现金
...	...	...
95	18.09	现金
96	54.84	支票
97	41.10	支票
98	43.14	支票
99	3.31	现金
100	69.77	信用卡

案例研究 2　国家健康护理协会

国家健康护理协会对健康护理专业培养的未来护士人员的缺乏情况很关注。为了了解护士们当前的工作满意度，协会发起了一项对全国医院护士的研究。作为该项研究的一部分，一个包括 50 名护士的样本被要求列出她们对工作、工资和晋升机会的满意度。满意度三方面中的每一个都用 0 — 100 分来衡量，较高的分值表示较大的满意度。收集的数据还显示护士所在医院的类型，医院的类型分别为私人医院(P)、退伍军人管理局下属医院(VA)、大学附属医院(U)。数据的一部分列在表 3.14(见下页)中，完整的数据储存在 CD 光盘的文件“Health”内。

管理报告

利用描述性统计方法来汇总数据。你的汇总结果应该有助于和其他人进行相互交流。讨论你的结果，尤其是对下列问题应作出评论：

1. 根据完整的数据集和三种工作满意度变量，哪个方面的工作最能使护士满意？哪方面的满意度最低？如果需要的话，你感觉在哪个方面应该进行改进？试讨论。
2. 根据对变异程度所做的描述性度量，护士们对哪一方面的工作满意度差别最大?解释原因。

表 3.14　50 名护士样本的满意度分数数据


CD 光盘数据
Health

护士	医院类型	工作	工资	晋升
1	Private	74	47	63
2	VA	72	76	37
3	University	75	53	92
4	Private	89	66	62
5	University	69	47	16
6	Private	85	56	64
7	University	89	80	64
8	Private	88	36	47
9	University	88	55	52
10	Private	84	42	66
...	...	...	...	...
45	University	79	59	41
46	University	84	53	63
47	University	87	66	49
48	VA	84	74	37
49	VA	95	66	52
50	Private	72	57	40

3. 关于医院的类型你可以了解到什么情况？是否存在某类医院，其护士满意度高于其他类型的医院？能否根据你的结果对了解和提高工作满意度提出建议？试讨论。
4. 你还能利用哪些描述性统计量及其结论来了解和改进工作满意度？

案例研究 3　亚太地区的商学院

当前，追求工商管理专业较高的学历已是一种世界性的潮流。有调查显示，越来越多的亚洲人选择攻读工商管理硕士（MBA）学位，把它作为通向成功的一种途径（*Asia, Inc.*, September 1997）。亚太地区商学院 MBA 课程申请者的数目在一年内上升了大约 30%。根据亚太地区 74 所商学院的报告，在 1997 年共有 170 000 个申请者竞争 11 000 个 1999 年授予的全日制 MBA 学位，创造了历史记录。需求高涨的一个主要原因是 MBA 学位能够极大地增加个人收入。


CD 光盘数据
Asian

在整个亚太地区，成千上万的亚洲人对于暂时放下他们的工作并花费两年时间来提高工商管理理论水平，正日益显示出增长的热情。这些商学院的课程明显很繁重，包括了经济学、金融学、市场营销、行为科学、劳动关系、决策学、运筹学和经济法等等。表 3.15 是 Asia 公司提供的数据，列出了亚太地区著名商学院的一些情况。

表 3.15　亚太地区 25 所商学院的数据

商学院名称	录取名额	每专业人数	本国学生学费（美元）	国外学生学费（美元）	平均年龄	国外学生百分比	是否要求 GMAT	是否要求英语测试	是否要求工作经验	起薪（美元）
墨尔本商学院	200	5	24 420	29 600	28	47	是	否	是	71 400
新南威尔士大学（悉尼）	228	4	19 993	32 582	29	28	是	否	是	65 200
印度管理学院（Ahmedabad）	392	5	4 300	4 300	22	0	否	否	否	7 100
香港中文大学	90	5	11 140	11 140	29	10	是	否	否	31 000
日本国际大学（Niigata）	126	4	33 060	33 060	28	60	是	是	否	87 000
亚洲管理学院（马尼拉）	389	5	7 562	9 000	25	50	是	否	是	22 800
印度管理学院（班加罗尔）	380	5	3 935	16 000	23	1	是	否	否	7 500
新加坡国立大学	147	6	6 146	7 170	29	51	是	是	是	43 300
印度管理学院（加尔各答）	463	8	2 880	16 000	23	0	否	否	否	7 400
澳大利亚国立大学（堪培拉）	42	2	20 300	20 300	30	80	是	是	是	46 600
南洋理工大学（新加坡）	50	5	8 500	8 500	32	20	是	否	是	49 300
昆士兰大学（布里斯班）	138	17	16 000	22 800	32	26	否	否	是	49 600
香港科技大学	60	2	11 513	11 513	26	37	是	否	是	34 000
麦夸里管理研究生院（悉尼）	12	8	17 172	19 778	34	27	否	否	是	60 100
Chulalongkorn 大学（曼谷）	200	7	17 355	17 355	25	6	是	否	是	17 600
Monash Mt. Eliza 商学院（墨尔本）	350	13	16 200	22 500	30	30	是	是	是	52 500
亚洲管理学院（曼谷）	300	10	18 200	18 200	29	90	否	是	是	25 000
Adelaide 大学	20	19	16 426	23 100	30	10	否	否	是	66 000
梅西大学（新西兰，Palmerston North）	30	15	13 106	21 625	37	35	否	是	是	41 400
皇家墨尔本技术工商学院	30	7	13 880	17 765	32	30	否	是	是	48 900
Jamnalal Bajaj 管理研究院（孟买）	240	9	1 000	1 000	24	0	否	否	是	7 000
柯廷理工学院（珀思）	98	15	9 475	19 097	29	43	是	否	是	55 000
拉合尔管理科学院	70	14	11 250	26 300	23	2.5	否	否	否	7 500
马来西亚 Sains 大学（槟城）	30	5	2 260	2 260	32	15	否	是	是	16 000
De La Salle 大学（马尼拉）	44	17	3 300	3 600	28	3.5	是	否	是	13 100

管理报告

利用描述性统计学方法来汇总表 3.15 的数据，并讨论你的结果。

1. 对数据集的每个变量进行汇总。根据最大值、最小值、适当的平均值和分位数，进行评论和解释。关于亚太地区的商学院，这些描述性统计量能够提供什么新的信息？
2. 汇总数据以比较：
   a. 本国和国外学生学费的差别。
   b. 要求工作经验和不要求工作经验的学校学生平均起薪的差别。
   c. 要求英语测试和不要求的学校学生平均起薪的差别。
3. 对表 3.15 的数据作出其他的图形和数值汇总，以便同他人交流。

附录 3.1　利用 Minitab 计算描述性统计量

在本附录中，我们将演示怎样利用 Minitab 计算描述性统计量。表 3.1 列出了 12 名商学院毕业生的起薪数据，图 3.10 的 Panel A 显示了利用 Minitab 汇总这些数据后得到的描述性统计量。Panel A 表头的文字含义为：

N	数据的个数
Mean	均值
Median	中位数
StDev	标准差
Min	最小数据值
Max	最大数据值
Q_1	第一四分位数
Q_3	第三四分位数

在注解中，我们已经讨论过标记为 TrMean 的数值度量，但还没有介绍 SEMean。TrMean 就是**处理均值**，它表示在消除了数据集里最小的一些值和最大的一些值以后的数据集中心位置。Minitab 提供了 5% 处理均值的计算工具，该方法消去了数据集里最小的 5% 数据值和最大的 5% 数据值。这样，通过计算中间 90% 的数据，我们可以得到 5% 处理均值。SEMean 被称为**均值标准误差**，我们用标准差除以 N 的平方根就能够把它计算出来。对它的解释和使用将在第 7 章中进行讨论，那时我们将会介绍抽样和抽样分布的问题。

虽然极差、四分位点内距、方差和变异系数不能出现在 Minitab 的输出中，但如果需要的话，可以利用下面的公式，根据图 3.10 的结果很容易地计算出这些值来。

$$\text{极差} = \text{最大值} - \text{最小值}$$

$$IQR = Q_3 - Q_1$$

$$方差 = (StDev)^2$$

$$变异系数 = (StDev/Mean) \times 100$$

Panel A: 描述性统计业务量

N	均值	中位数	TrMean	标准差	SEMean
12	2 940.0	2 905.0	2 924.5	165.7	47.8
最小值	最大值	Q_1	Q_3		
2 710.0	3 325.0	2 857.5	3 025.0		

Panel B: 箱形图

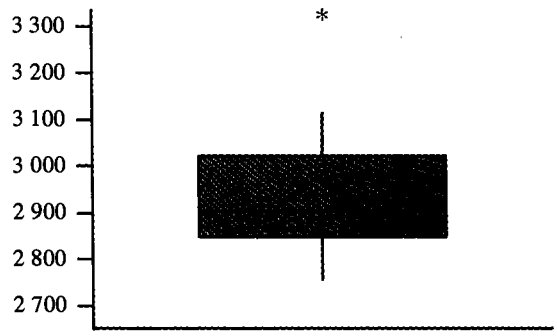


图 3.10 利用 Minitab 作出的描述性统计量和箱形图

最后，我们注意到 Minitab 计算出的四分位数 $Q_1 = 2\,857.5$, $Q_3 = 3\,025$，与 3.1 节计算出的结果 $Q_1 = 2\,865$, $Q_3 = 3\,000$ 不太一样。造成这种现象的原因*是在识别四分位数时，使用了不同的方法*。因此，通过某种方法得出的 Q_1 和 Q_3，不能等同于通过其他方法得到的 Q_1 和 Q_3。这些不同之处可以忽略，但是却不能忽视采用四分位数时所使用的方法。



CD 光盘数据
Salary

现在让我们来看一看图 3.10 中的统计量是如何产生的。把起薪数据输入 Minitab 工作表的 C2 列(第 2 例)，再利用下述步骤就能够产生这些描述性统计量。

步骤 1. 选择 **Stat** 下拉菜单；

步骤 2. 选择 **Basic Statistics** 选项；

步骤 3. 选择 **Display Descriptive Statistics**；

步骤 4. 当 Display Descriptive Statistics 对话框出现时：

在 **Variables** 框中输入 **C2**；

点击 **OK**。

图 3.10 的 Panel B 是用 Minitab 作出的箱形图。从第一到第三四分位数画出的箱体包含了中间 50% 的数据，箱体内的直线表示了中位数的位置，星号表示异常值 3 325。

---

* 对于按照升序排列（从最小值到最大值）的 n 个数据值，Minitab 利用公式 $(n+1)/4$ 和 $3(n+1)/4$ 来分别确定 Q_1 和 Q_3 的位置。当位置的数值是分数时，Minitab 在相邻的两个数据值之间内插，以确定对应的四分位数。

通过下列步骤能够产生这个箱形图：

步骤 1. 选择 **Graph** 下拉菜单；

步骤 2. 选择 **Boxplot** 选项；

步骤 3. 当 Boxplot 对话框出现时：

在 **Graph Variables** 框的 **Y** 项目下输入 C2；

点击 **OK**。



CD 光盘数据 Stereo

图 3.11 显示了用 Minitab 作出的表 3.7 内的音像设备商店数据的协方差和相关系数结果。在图中的协方差部分，Commerci 表示周末电视广告的次数，Sales 表示这些星期的销售额。行 Sales 和列 Commerci 交叉点的值为 11，就是在 3.5 节计算出的样本协方差；行 Commerci 和列 Commerci 交叉点的值 2.22 222 是广告次数的样本方差；行 Sales 和列 Sales 交叉点的值 62.88 889 是销售额的样本方差；而样本相关系数 0.930，则显示在输出结果的 Correlation 部分。注：对于相关系数结果中出现的 p – 值，我们将在第 14 章讨论其解释和应用。

现在，让我们来演示如何能够得到图 3.11 的信息。我们把广告次数数据输入 C2 列，把销售额数据输入 C3 列，再利用下列步骤即可生成图 3.11 前 3 行的协方差结果：

步骤 1. 选择 **Stat** 下拉菜单；

步骤 2. 选择 **Basic Statistics** 选项；

步骤 3. 选择 **Covariance**；

步骤 4. 当 Covariance 对话框出现时：

在 **Variables** 框内输入 C2 C3；

点击 **OK**。

要想得到图 3.11 的相关系数结果，只需在协方差步骤的基础上改动一个步骤即可。即在步骤 3 时选择 **Correlation** 选项。

```
Covariances: Commercials, Sales

          Commerci     Sales
Commerci   2.22222
Sales     11.00000  62.88889

Correlations: Commercials, Sales

Pearson correlation of Commercials and Sales = 0.930
P-Value = 0.000
```

图 3.11　利用 Minitab 作出的广告数和销售额数据的协方差和相关系数

附录 3.2 利用 Excel 计算描述性统计量

Excel 应用软件能够生成本章讨论过的描述性统计量。我们将演示怎样利用 Excel 做出几种单变量的位置和变异程度度量值，以及如何生成反映两变量间关系的协方差和相关系数结果。

使用 Excel 函数


CD 光盘数据
Salary

Excel 提供了计算均值、中位数、众数、样本方差和样本标准差的函数。我们以表 3.1 的起薪数据来说明使用这些 Excel 函数计算均值、中位数、众数、样本方差和样本标准差的具体方法。在我们描述所需要的步骤时，先参考图 3.12。图中，数据已被输入 B 列。

Graduate	Starting Salary
1	2850
2	2950
3	3050
4	2880
5	2755
6	2710
7	2890
8	3130
9	2940
10	3325
11	2920
12	2880

Mean	=AVERAGE(B2:B13)
Median	=MEDIAN(B2:B13)
Mode	=MODE(B2:B13)
Variance	=VAR(B2:B13)
Standard Deviation	=STDEV(B2:B13)

Graduate	Starting Salary
1	2850
2	2950
3	3050
4	2880
5	2755
6	2710
7	2890
8	3130
9	2940
10	3325
11	2920
12	2880

Mean	2940
Median	2905
Mode	2880
Variance	27440.91
Standard Deviation	165.65

图 3.12 使用 Excel 函数计算均值、中位数、众数、方差和标准差

Excel 中的 AVERAGE 函数能够计算均值，只需要把下面的公式输入单元格 E1：

= AVERAGE (B2: B13)

类似地，把公式“= MEDIAN(B2: B13)，= MODE(B2: B13)，=VAR(B2: B13)，=STDEV(B2: B13)”分别输入单元格 E2: E5，就能够计算出中位数、众数、方差和标准差。位于表面的工作表显示出：使用 Excel 函数计算出的值与我们在本章的有关部分计算出的结果是相同的。

Excel 还提供了计算方差和相关系数的函数。但是，在使用这些函数时，你必须十分小心，因为协方差函数把数据都作为总体数据来处理，相关系数函数则把数据都作为样本数据来处理。因

此，必须对使用 Excel 协方差函数得到的结果进行调整，以得到样本协方差。在这里，我们将演示如何使用这些函数来计算表 3.7 音像设备商店的样本协方差和样本相关系数。在讲述这些步骤时可以先参考图 3.13。

把公式“=COVAR(B2: B11, C2: C11)”输入单元格 F1，再利用 Excel 的协方差函数 COVAR 就能够计算出数据的总体协方差。类似地，把公式“=CORREL(B2: B11, C2: C11)”输入单元格 F2 就可计算样本相关系数。出现在底面的工作表显示了 Excel 函数计算出的数值结果，注意样本相关系数的值 0.93 与我们利用公式(3.12)得到的结果相同。不过，由 Excel COVAR 函数得到的结果 9.9，是把数据作为总体数据处理得出的结果。因此，我们必须对 Excel 的结果 9.9 进行调整，以得到样本协方差。调整过程其实相当简单，首先，注意总体协方差的公式——公式(3.11)需要除以数据集中观察值的总个数。但是，样本协方差的公式——公式(3.10)，则需要除以观察值的总个数减 1。所以，为了使用 Excel 的结果 9.9 来计算样本协方差，我们只需把 9.9 乘以 $n/(n-1)$。因为 $n=10$，我们有

$$s_{xy}=\left(\frac{10}{9}\right)9.9=11$$

因此，音像设备商店数据的样本协方差是 11。

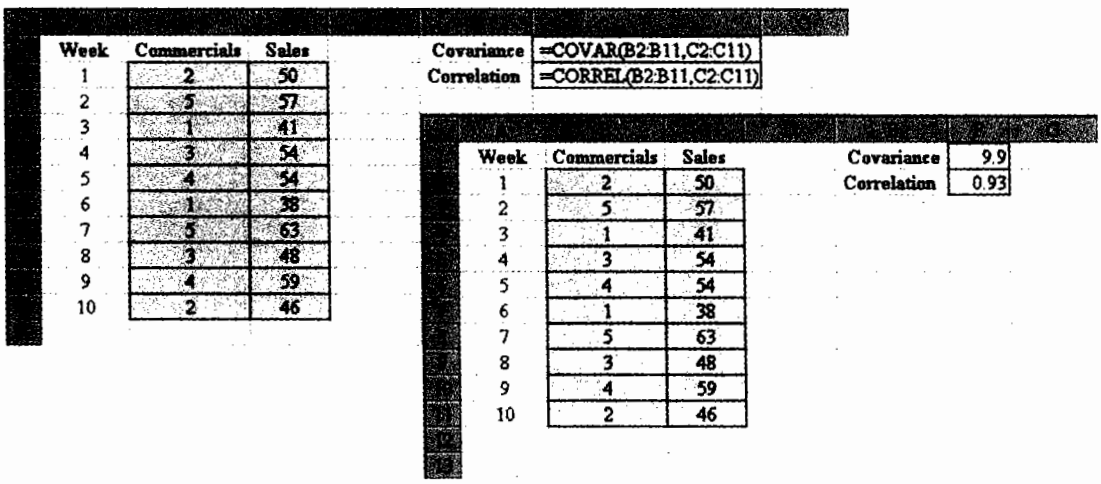


图 3.13　利用 Excel 函数计算协方差和相关系数

使用 Excel 的描述性统计量工具

我们已经举例说明过，Excel 能够提供统计函数计算数据集的描述性统计量，这些函数可以计算在同一时间的单个统计量（例如，均值、方差等）。Excel 还提供了各种数据分析工具，其中之一称为描述性统计量，它使得使用者可以马上计算出各种描述性统计量。我们将在这里演示如何利用它来计算表 3.1 中起薪数据的描述性统计量。在我们描述有关的步骤时需要参考图 3.14。

步骤 1. 选择 **Tools** 下拉菜单；

步骤 2. 选择 **Data Analysis**；

步骤 3. 从 Analysis Tools 列表中选择 **Descriptive Statistics** 选项：

步骤 4. 当 Descriptive Statistics 对话框出现时：

在 **Input Range** 框中输入 B1: B13；

选择 **Grouped By Columns**；

选择 **Labels in First Row**；

选择 **Output Range**；

在 **Output Range** 框中输入 D1（指定工作表部分的左上角来显示描述性统计量）；

点击 **OK**。

图 3.14 的单元格 D1: E15 列出了 Excel 作出的描述性统计量，其中黑体的项目已经在本章介绍过，非黑体的项目会在以后的课程或者更高级的课程中涉及到。

Graduate	Starting Salary		*Starting Salary*	
1	2850			
2	2950		Mean	2940
3	3050		Standard Error	47.8199
4	2880		Median	2905
5	2755		Mode	2880
6	2710		Standard Deviatio	165.653
7	2890		Sample Variance	27440.91
8	3130		Kurtosis	1.718884
9	2940		Skewness	1.091109
10	3325		Range	615
11	2920		Minimum	2710
12	2880		Maximum	3325
			Sum	35280
			Count	12

图 3.14　Excel 描述性统计量工具的应用

4 概率论简述

统计实例:MORTON 国际公司
4.1 试验、计数法则和概率指派
计数法则、组合和排列
概率指派
肯塔基电力公司项目的概率
4.2 事件和它们的概率
4.3 概率的一些基本关系
事件的补
加法法则
4.4 条件概率
独立事件
乘法法则
4.5 贝叶斯定理
表格方法

统计实例

MORTON 国际公司*

芝加哥,伊利诺伊州

Morton 国际公司是一家经营盐业、火箭推进器和特殊化学品的公司。它的一个子公司 Carstab 公司生产特殊化学品，并能够按照顾客要求的特殊规格设计并生产各种化学产品。Carstab 公司曾经为一家特殊的客户生产一种昂贵的催化剂，以满足客户的工艺需求。这批产品中有一部分，而非全部，符合该客户对产品的规格要求。

Carstab 的客户同意在收到产品以后对每批产品进行测试，以确定催化剂是否具有所需要的功能，没有通过客户测试的产品将被退回 Carstab 公司。一段时间以后，公司发现客户接受每个批量的 60%，退回 40%。按照概率术语来说，每一批发往客户的产品具有 0.60 的被接受概率和 0.40 的被拒绝概率。

无论是 Carstab 公司还是客户都不愿意看到这样的结果。为了改进服务，Carstab 公司尝试探索在发运之前重复客户测试方法的可能性。但是，由于测试所需的特殊检测设备成本高昂，使得这种办法行不通。该公司的化学工程师建议使用一种新的，相对低廉的测试方法，这种方法的目的是检测某批产品是否能够通过客户的测试。与此有关的概率问题是：如果某批产品通过了公司的这种测试的话，那么它们通过客户测试的概率(Probability)是多少？



Morton Salt："不雨则已，一雨倾盆。"© Joe Higgins/ South-Western.

从每批产品中取出一个样本，并使用新方法进行测试，只有通过测试的那批产品才发给客户。对数据的概率分析表明：如果一批产品通过了公司的测试，那么它们能够通过客户测试并被接受的概率为 0.909，相应地，被退回的概率就只有 0.091。概率分析提供了关键性的支持证据，支持采纳和执行新的测试方法。新方法使得对客户的服务得到了改善，并大大地减少了退货的运输和处理成本。

一批产品在通过公司测试后被客户接受的成本称为条件概率。在本章中，你将学习如何计算条件概率及其他一些对决策有帮助的概率。

*作者感谢 Morton 国际公司的 Michael Haskell 提供了该应用实例。

商务决策往往建立在对下面这些问题的不确定性进行分析的基础上：

1. 如果产品涨价的话，销售额下降的机会是多少？
2. 新的装配方法能够使生产率提高的可能性有多大？
3. 项目按期完工的概率是多少？
4. 新的投资能够盈利的可能性是多少？

概率(probability)是对一个事件发生可能性的数值度量。因此，概率可用来度量与上面所列四

类问题相关的不确定程度。如果我们知道概率的话，就能够确定每个事件发生的可能性。

关于概率的一些最早期的工作起源于 Pierre de Fermat 和 Blaise Pascal 之间在 17 世纪 50 年代来往的信件。最早的概率专著是 Pierre-Simon de Laplace 在 1812 年出版的 *Theorie Analytique des Probabilities*。

概率值总是分布在 0 — 1 之间。接近于 0 的概率表示某事件几乎不可能发生；接近于 1 的概率表示某事件几乎肯定要发生；处在 0 和 1 之间的其他概率则表示某事件发生的可能程度。例如，如果让我们考虑事件“明天下雨”，我们知道当天气预报报告说“降雨概率接近 0”时，就意味着几乎没有下雨的可能。但是，如果报告是“降雨概率为 0.90”的话，我们明白很可能会下雨。0.50 的概率则表示下雨的可能性和不下雨的可能性一样大。图 4.1 描述了概率作为对事件发生可能性的数值度量是怎样的一种情况。

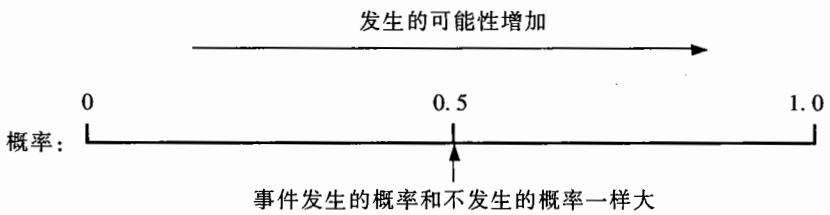


图 4.1　作为事件发生可能性数值度量的概率

4.1　试验、计数法则和概率指派

在讨论概率时，我们定义**试验**(experiment)就是能够产生明确结果的过程。如果对一个试验进行任何一次重复的话，所有可能的试验结果中有一个结果且只有一个结果会出现。下面是几个试验的例子和与试验相联系的结果：

试验	试验结果
抛掷一枚硬币	正面、反面
进行一次产品抽查	缺陷、无缺陷
打一次销售电话	购买、不购买
掷一个色子	1、2、3、4、5、6
进行一场橄榄球比赛	胜、负、平

在我们已经确定了所有可能的试验结果以后，我们也就确定了一个试验的**样本空间**(sample space)。

样本空间

一个试验的样本空间就是所有试验结果的集合。

一个试验结果也称为一个**样本点**(sample point)，表示它是样本空间的一个元素。

考虑上表的第一个试验——抛掷硬币，试验结果（样本点）由硬币向上的一面确定：正面或是反面。如果我们用 S 表示样本空间，我们可以使用下面的符号来描述它：

试验结果也称做样本点。

$$S = \{\text{正面，反面}\}$$

表中第二个试验——进行一次产品抽查的样本空间和样本点如下：

$$S = \{\text{缺陷，无缺陷}\}$$

刚才描述的试验都只有两个试验结果(样本点),但是如果我们考虑表中第四个试验——投掷色子的话，可能出现的试验结果就是色子朝上一面所显示的点数。这样，试验的样本空间就有 6 个样本点：

$$S = \{1, 2, 3, 4, 5, 6\}$$

计数法则、组合和排列

能够对试验结果进行确认和计数是指派概率的一个必要步骤。现在我们来讨论三种有用的计数法则。

多步骤试验(multiple-step experiments)。第一种计数法则是针对多步骤试验的。考虑投掷两枚硬币的试验，我们定义该试验的结果是按次序投掷两枚硬币，其朝上一面出现的情况。这个试验有多少个可能的试验结果？投掷两枚硬币的试验可被认为是一个两步骤试验，步骤 1 是投掷第一枚硬币，步骤 2 是投掷第二枚硬币。如果我们使用 H 来表示硬币的正面，使用 T 来表示硬币的反面,(H, H) 表示试验结果为第一枚硬币正面朝上,第二枚硬币也是正面朝上。沿用这种表示方法，我们可以描述该硬币投掷试验的样本空间 S 如下：

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

因此，我们看到有四个试验结果是可能的。在这个例子中，列出所有的试验结果并不困难。

多步骤试验的计数法则使得我们能够在不列出试验结果的情况下，也能够确定试验结果的数目。

多步骤试验的计数法则

如果一个试验依次进行 k 个步骤，第一步有 n_1 个可能的结果，第二步有 n_2 个可能的结果，依此类推，那么试验结果的总数就可由式 $n_1 \times n_2 \times \cdots \times n_k$ 给出。

回顾投掷两枚硬币的试验，按照顺序首先投掷一枚（$n_1 = 2$），然后再投掷另一枚（$n_2 = 2$），我们根据计数法则知道共有 $2 \times 2 = 4$ 个不同的试验结果。正如在上文指出的，它们是 $S = \{(H, H), (H, T), (T, H), (T, T)\}$。在一个投掷 6 枚硬币的试验中，试验结果的数目等于 $2 \times 2 \times 2 \times 2 \times 2 \times 2 = 64$ 个。

有一种图形表达方式有助于把试验直观地表现出来，并可以列举出多步骤试验的结果，它被称为**树形图**(tree diagram)。图 4.2 就是抛掷两枚硬币试验的树形图，试验中步骤的顺序为从树的左端移动到树的右端。步骤 1 对应于抛掷首枚硬币，它具有两个分支分别对应于两个可能的结果。步骤 2 对应于抛掷第二枚硬币，并且步骤 1 的每一个可能的结果都具有两个分支，分别对应于步骤 2 的两个可能结果。最终，树右端的每一点对应着一个试验结果。从树的左端结点通往树右端任一结点的每条路径都表示一个单独的结果次序。

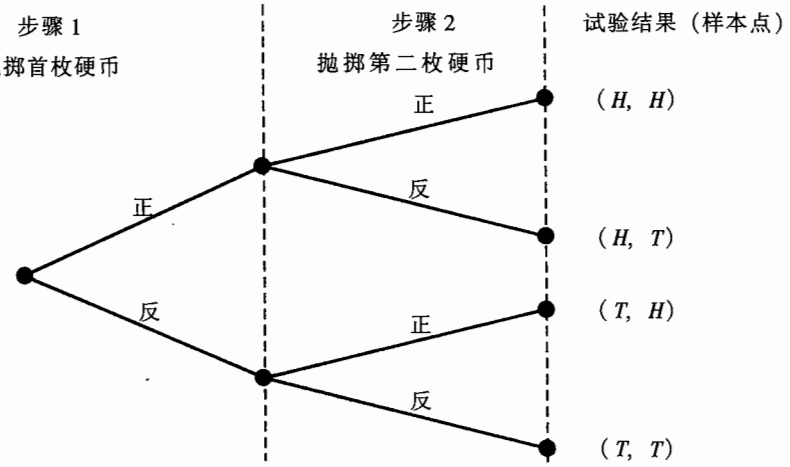


图 4.2　抛掷两枚硬币试验的树形图

不用树形图，人们可能认为对一枚硬币投掷两次的试验只有三个试验结果是可能的：0 次正面、1 次正面和 2 次正面。

现在让我们来看一看，怎样利用多步骤试验的计数法则对肯塔基电力公司（KP&L）所面临的扩大生产能力项目进行分析。KP&L 正在开始一个项目，其设计目的是增加它在北肯塔基一家工厂的发电能力。项目被分割为两个按顺序的阶段或步骤：阶段 1 是设计工作，阶段 2 是建设工作。由于对每一阶段都尽量严密地进行规划和控制，因此管理人员不能预测出完成项目每个阶段所需要的准确时间。对于以往类似建设项目的分析显示：完成设计阶段需要 2、3 或 4 个月，完成建设阶段则需要 6、7 或 8 个月。另外，由于用户对更多电力供应的迫切需要，管理者确定了用 10 个月完成整个项目的目标。

因为该项目在设计阶段（步骤 1）有 3 个可能的完成时间，在建设阶段（步骤 2）也有 3 个可能的完成时间，所以在这里我们能够应用多步骤试验的计数法则确定总的试验结果数为：$3 \times 3 = 9$ 个。为了描述试验结果，我们使用一个包括两个数字的符号。例如，(2, 6)表示设计在 2 个月内完成，建设阶段在 6 个月内完成。这个试验结果导致了完成整个项目需要总共 $2 + 6 = 8$ 个月。表

4.1 汇总了 KP&L 项目的全部 9 个试验结果，图 4.3 的树形图则显示了这 9 个结果(样本点)是怎样发生的。

表 4.1 **KP&L 项目的试验结果（样本点）**

完成时间（月）			
步骤 1（设计）	步骤 2（建设）	试验结果的符号	整个项目的完成时间（月）
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

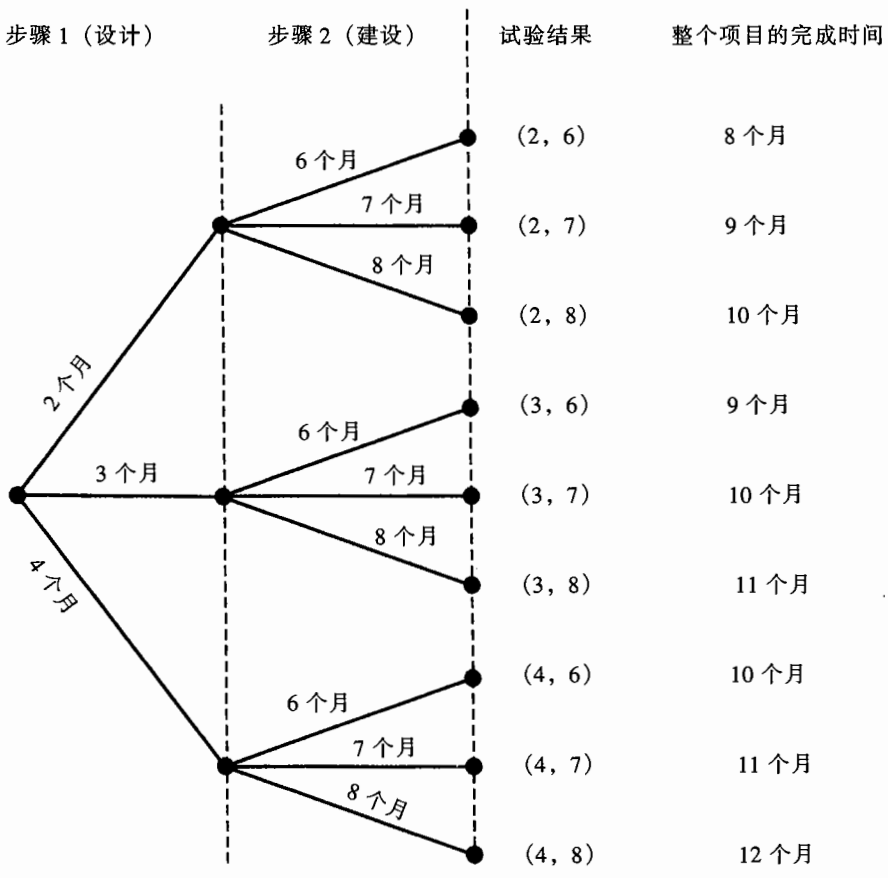


图 4.3 KP&L 项目的树形图

计数法则和树形图都可以被用来帮助项目经理确认试验结果和确定可能的项目完成时间。根据图 4.3 的信息，我们看到整个项目将在 8—12 个月内完工，并且 9 个试验结果中有 6 个能够满足所要求的 10 个月以内的完工时间。尽管确认试验结果是有用的，但我们还需要考虑怎样把概率值指派给各个试验结果，以估计项目将在要求的 10 个月内完工的概率。

组合(combinations)。当试验涉及到从 N 个项目的集合中选择其中 n 个（通常小于 N）的问题时，这第二种计数方法能够帮助我们计算试验结果的数目，它被称为组合计数法则。

组合计数法则

在同一时间从 N 个项目中抽取其中 n 个的组合数是

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

式中

$$N! = N \times (N-1) \times (N-2) \times \cdots \times 2 \times 1;$$

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1;$$

并且

$$0! = 1$$

符号“!”叫做阶乘，举例来说，5 的阶乘就是 $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$。而对于 0 的阶乘，就定义为 $0! = 1$。

从有限总体 N 中抽样，组合计数法则用来计算能够抽取的容量为 n 的不同样本数目。

作为组合计数法则的一个实例，考虑这样一个质量控制程序：检查者从 5 个零件中随机地抽取 2 件，以测试零件是否存在缺陷。问从 5 个零件中选择其中 2 件有多少个组合？根据公式(4.1)，对于 $N = 5$ 且 $n = 2$，我们有

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = \frac{120}{12} = 10$$

因此，该试验共有 10 个结果是可能的。如果我们把 5 个零件分别加上标记为 A、B、C、D 和 E，这 10 个组合或者说是 10 个试验结果就是 AB、AC、AD、AE、BC、BD、BE、CD、CE 和 DE。

再举一个例子，我们考虑俄亥俄州的博彩系统，该系统从 47 个数字中随机地抽取 6 个数字，来确定周末彩票的获奖者。根据组合计数法则公式(4.1)，我们知道从 47 个数字中抽取 6 个不同数字的方法总数为

$$\binom{47}{6} = \frac{47!}{6!(47-6)!} = \frac{47!}{6!41!} = \frac{47 \times 46 \times 45 \times 44 \times 43 \times 42}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 10\,737\,573$$

组合计数法则告诉我们，在彩票摇奖中共有超过 1 000 万个试验结果是可能的，也就是说购买了一张彩票的个人中奖机会是 1/10 737 573。

组合计算规则表明彩票中奖机会非常的小。

从有限总体 N 中抽样，组合方法用于计算容量为 n 的不同的样本的数目。

排列(permutations)。第三种计数法则是排列计数法则。在抽取的顺序也很重要的情况下，当我们从 N 个项目的集合中抽取其中 n 个时，使用它能够计算出试验结果的总数。在这类问题中，即使是同样的 n 个项目，如果取出次序不同的话，也被认为是不同的试验结果。

排列计数法则

在同一时间，从 N 个项目中抽取其中 n 个的排列数是

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

排列计数法则与组合计数法则有紧密的联系，但具有同样项目数的试验，其排列数比组合数要多。这是因为每选择 n 个项目就有 $n!$ 种不同方法去排列这 n 个项目。

作为一个例子，我们再次考虑这样一个质量控制程序：检查者从 5 个零件中抽取 2 个以检查是否存在缺陷。问抽取方法有多少个排列数？利用公式 (4.2) 的计数法则，对于 $N=5$ 且 $n=2$，我们有

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{5\times4\times3\times2\times1}{3\times2\times1} = 5\times4 = 20$$

因此，在必须考虑抽取次序时，该试验共有 20 个结果是可能的。如果我们把 5 个零件分别标记为 A、B、C、D 和 E，则 20 个排列结果就是 AB、BA、AC、CA、AD、DA、AE、EA、BC、CB、BD、DB、BE、EB、CD、DC、CE、EC、DE 和 ED。

概率指派

现在让我们来学习概率是怎样被指派给各个试验结果的。三种最常用的概率指派(赋值)方法是古典法、相对频数法和主观法。但不管采用什么方法，所指派的概率都必须满足概率的两个**基本要求**。

概率指派的基本要求

1. 指派给每个试验结果的概率必须介于 0 —1 之间。如果我们以 E_i 表示第 i 个试验结果，以 $P(E_i)$ 表示它的概率，那么该要求可以表示为

$$0 \leqslant P(E_i) \leqslant 1，\text{对于所有的 } i \text{ 都成立} \tag{4.3}$$

2. 所有试验结果的概率之和必定等于 1。对于 n 个试验结果，该要求可表示为

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{4.4}$$

当所有试验结果的可能性相等时，使用**古典法**(classical method)来指派概率是恰当的。如果有 n 个可能试验结果的话，那么指派给每个试验结果的概率就是 $1/n$。这种方法能够自动地满足概率指派的两个基本要求。

作为例子，考虑抛掷一枚硬币的试验。两个试验结果——正面和反面——是等可能的。因为两个等可能结果中的一个是正面，则观察到正面朝上结果的概率是 1/2 或 0.50。同样地，观察到

反面朝上结果的概率也是 1/2 或 0.50。

举另外一个例子，考虑投色子试验。我们知道 6 种可能结果出现的可能性相等，因此对每个结果指派的概率为 1/6。如果用 P(1)表示色子朝上一面是 1 点的话，那么 $P(1)=1/6$。类似地，$P(2)=1/6$、$P(3)=1/6$、$P(4)=1/6$、$P(5)=1/6$ 和 $P(6)=1/6$。我们还注意到：因为每个概率都大于或等于零且所有概率之和等于 1，它们满足公式(4.3)和(4.4)的要求。

当试验被重复很多次，并且我们能够得到足够的数据来估计各种试验结果出现的次数所占比例时，就可以利用**相对频数方法**(relative frequency method)进行概率指派。考虑这样一个例子：对某地方医院 X 射线部门病人的等候时间进行研究。下面的数据记录了连续 20 天在上午 9:00 等候服务的病人人数：

等候人数	该结果出现的天数
0	2
1	5
2	6
3	4
4	3
总计	20

这些数据表明：在这 20 天里有 2 天，等候服务的病人数为 0；有 5 天，等候的病人数为 1；其他依此类推。利用相对频数方法，我们把概率 2/20 = 0.10 指派给 0 病人的试验结果；把概率 5/20 = 0.25 指派给 1 个病人的试验结果；把概率 6/20 = 0.30 指派给 2 个病人的试验结果；把概率 4/20 = 0.20 指派给 3 个病人的试验结果；把概率 3/20 = 0.15 指派给 4 个病人的试验结果。

和使用古典法一样，当我们使用相对频数方法时，也自动地满足公式（4.3）和（4.4）的两个基本要求。

当试验结果发生的可能性相等这个假定不符合实际情况时，或者当只能得到很少的相关数据时，这时使用**主观方法**(subjective method)来指派概率最为合适。为了利用主观方法向试验结果指派概率，我们可以使用任何可以得到的信息，比如我们的经验或是直觉。在综合考虑了所有可用的信息之后，得出的概率值就表示我们对发生某个试验结果的**置信度**(范围为 0—1)。因为主观概率表示的是某个人的置信度，它是属于个人的。在使用主观方法时，对同一个试验结果，不同的人可能会指派不同的概率。

我们在使用主观概率指派方法时，要特别注意，确保主观概率要满足公式(4.3)和(4.4) 的要求。指派给每个试验结果的概率值都必须介于 0—1 之间，并且所有试验结果的概率之和必须等于 1。

考虑 Tom 和 Judy Elsbernd 申请购房的例子，有两个结果是可能的：

E_1 = 他们的申请被接受；

E_2 = 他们的申请被拒绝。

Judy 相信他们的申请被接受的概率是 0.8，因此 Judy 指派的概率为：$P(E_1)=0.8$，$P(E_2)=0.2$。可是，Tom 认为他们的申请被接受的概率只有 0.6，因此他指派的概率是：$P(E_1)=0.6$，

$P(E_2)=0.4$。我们注意到 Tom 对 E_1 的概率估计反映了这样一个事实：他对申请被接受可能性的看法比 Judy 悲观。

无论 Judy 还是 Tom，他们指派的概率都满足两个基本要求。而他们对概率的估计不同这一事实，则强调了主观方法的个人属性。

即使在能够应用古典方法和相对频数方法的商务环境中，经理们可能仍然愿意使用主观概率估计。在这种情况下，往往需要把从古典法和相对频数法得来的估计值与主观概率估计相结合，才能得到最好的概率估计。

贝叶斯理论(见 4.5 节)提供了一种计算均值的方法，可将先验主观概率与修正的或事后的概率结合起来。

肯塔基电力公司项目的概率

为了对 KP&L 项目进行深一步的分析，我们必须向表 4.1 列出的 9 个试验结果中的每一个赋予概率。在经验和判断的基础上，管理者得出了试验结果的可能性不等的结论。因此，不能使用古典法来指派概率。管理者于是决定，对 KP&L 公司过去三年来承担的类似项目的完工时间进行研究，表 4.2 列出了对 40 个类似项目的研究结果。

在检查完研究结果之后，管理者决定使用相对频数方法指派概率。虽然管理者能够作出主观概率估计，但觉得目前的项目与以前的 40 个项目非常近似，因此判断相对频数方法是最佳的。

表 4.2　40 个 KP&L 项目的完工情况

完工时间（月）			具有该完工时间的
阶段 1	阶段 2	样本点	过去项目数
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
		总计	40

在使用表 4.2 的数据计算概率的过程中，我们注意到：结果 (2, 6) —— 阶段 1 在 2 个月内完成，阶段 2 在 6 个月内完成——在过去的 40 个项目中发生过 6 次。因此我们能够使用相对频数方

法给这个结果指派概率为 6/40 = 0.15。类似地，结果(2, 7)也在 40 个项目中发生过 6 次，其概率为 6/40 = 0.15。继续以这种方式计算，我们把得到的 KP&L 项目样本点的概率指派结果列在表 4.3 中。注意 $P(2,6)$表示样本点(2, 6)的概率；$P(2,7)$表示样本点(2, 7)的概率，依此类推。

表 4.3　KP&L 项目基于相对频数方法的概率指派

样本点	项目完工时间	样本点的概率
(2, 6)	8 个月	$P(2,6)=6/40=0.15$
(2, 7)	9 个月	$P(2,7)=6/40=0.15$
(2, 8)	10 个月	$P(2,8)=2/40=0.05$
(3, 6)	9 个月	$P(3,6)=4/40=0.10$
(3, 7)	10 个月	$P(3,7)=8/40=0.20$
(3, 8)	11 个月	$P(3,8)=2/40=0.05$
(4, 6)	10 个月	$P(4,6)=2/40=0.05$
(4, 7)	11 个月	$P(4,7)=4/40=0.10$
(4, 8)	12 个月	$P(4,8)=6/40=0.15$
		总计　1.00

评　注

1. 在统计学中试验的概念与物理学中试验的概念很不一样。物理学的试验通常在试验室或可控的环境下进行，以便了解原因和结果。而统计学试验的结果则取决于概率，即使试验以完全相同的方式被重复，也可能会发生根本不同的结果。由于概率对试验结果有如此的影响，统计学试验有时又被称为**随机试验**。
2. 当我们从容量为 N 的总体里面无放回地抽取一个随机样本时，可以应用组合计数法则来计算能够选取的 n 容量不同样本的个数。

练　习

方　法

1. 一个试验要经过三个步骤，第一步有 3 个可能结果，第二步有 2 个可能结果，第三步有 4 个可能结果。问整个试验存在多少个试验结果？

2. 自测题　从 6 个项目的集合中抽取 3 个，共有多少种方法？分别用字母 A、B、C、D、E 和 F 来表示各个项目，列出 3 个项目的每一种组合。

3. 从6个项目的集合中选取3个的排列数共有多少？分别用字母A、B、C、D、E、F来表示各个项目，列出3个项目(B、D、F)的每一种排列。

4. 考虑抛掷一枚硬币3次的试验。
   a. 画出试验的树形图。
   b. 列出所有试验结果。
   c. 每种试验结果的概率分别是多少？

5. 假定某试验具有5个可能性相等的结果：E_1、E_2、E_3、E_4和E_5。给每个结果指派概率，使其能够满足公式(4.3)和(4.4)的条件。你使用了什么方法？

6. 

某个具有3个可能结果的试验被重复了50次，并且我们知道E_1出现了20次，E_2出现了13次，E_3出现了17次。给每个结果指派概率，你使用了什么方法？

7. 一位决策者对某试验的4个结果主观地指派了如下概率：$P(E_1) = 0.10$，$P(E_2) = 0.15$，$P(E_3) = 0.40$，$P(E_4) = 0.20$。这是有效的概率指派吗？检查其是否满足公式（4.3）和公式(4.4)。

应　用

8. 在Milford市，申请区域变更要经过一个两步骤的过程：先由规划委员会评议，再由市议会最终决定。在步骤1，规划委员会将对区域变更请求进行审查，给出肯定或否定的意见。在步骤2，市议会将审查规划委员会的意见，并投票表决是通过还是否决该区域变更。区域变更申请是由一家房地产公司申请的。把申请过程作为一个试验。
   a. 该试验有多少个样本点？列出所有样本点。
   b. 画出试验的树形图。

9. 

简单随机样本是从容量为N的总体中选取容量为n的样本，以获得能够推断总体特征的数据。假定我们有一个包括50个银行账户的总体，想要选取一个4账户的随机样本以了解总体特征。问有多少个4账户的不同随机样本是可能的？

10. 

近年来，风险资本投资在投向公司的所有资金中所占的比例有了很大增长。根据Venture Economics（*Investor's Business Daily*, April 28, 2000）提供的数据，1999年共发生了2 374起风险资本投资事件。在这些投资里面，1 434起投向了位于加利福尼亚州的公司，390起投向了位于马萨诸塞州的公司，217起投向纽约州的公司，112起投向科罗拉多州的公司。在接受资金的公司中，有22%处于发展的早期阶段，有55%处于扩张阶段。

假定你想要在这些公司里面随机地选取一家来了解它们是怎样使用资金的。
   a. 你选到的公司来自加利福尼亚州的概率是多少？
   b. 你选到的公司不是来自上面提到的这四州的概率是多少？
   c. 该公司不是处于发展早期阶段的概率是多少？

d. 假定处于发展早期阶段的公司平均分布于全国各地，有多少家接受风险投资的马萨诸塞州公司处于它们的早期发展阶段？

e. 投资的总额是 324 亿美元，估计投往科罗拉多州的数额。

11. Strom 建设公司向两份合同投标。公司老板已经确认了所有的可能结果，并主观地给它们指派概率如下：

试验结果	是否得到合同 1	是否得到合同 2	概率
1	是	是	0.15
2	是	否	0.15
3	否	是	0.30
4	否	否	0.25

a. 这些概率指派有效吗？为什么？

b. 如果需要的话，我们应该怎样使概率指派变得有效？

12. 州彩票已经成为基金增长的最普遍方式。Powerball 彩票对几个州开放，购买该彩票时，参与者从数字 1—49 中选择 5 个数字，然后再从数字 1—42 中选择一个幸运数字，全部选对者就可赢得 Powerball 大奖。1998 年某个星期，该彩票的大奖金额除了向正确选择了头 5 个数字的彩民支付的 10 万美元以外，已累积达到了 1 亿 5 千万美元。

a. 计算选择头 5 个数字的方法总数。

b. 正确选择头 5 个数字从而获得 10 万美元奖金的概率是多少？

c. 赢得 Powerball 大奖的概率是多少？

13. 生产牙膏的某公司正在研究 5 种不同的包装设计。假定每种设计被顾客选中的可能性是相等的，你将怎样给每种包装设计指派选中的概率？在一个实际的试验中，100 位顾客被要求选择他们喜欢的设计，并取得了下列数据。这些数据支持选中每种设计的可能性相等这一假设吗？请解释。

设计	被选中的次数
1	5
2	15
3	30
4	40
5	10

4.2　事件和它们的概率

到目前为止，我们还是在日常用语的意义上使用名词“事件”，现在我们来介绍与概率相关的术语**事件**(event)的正式定义。

事件

事件是由若干样本点组成的集合。

作为一个应用实例，让我们回到 KP&L 问题上。假定项目经理关心的是整个项目能在 10 个月以内完工这一事件，参见表 4.3，我们看到：有 6 个样本点(2,6),(2,7),(2,8),(3,6),(3,7)和(4,6)，能够使项目在 10 个月以内完工。用字母 C 来表示项目在 10 个月以内完成这一事件，我们记做：

$$C = \{(2,6),\ (2,7),\ (2,8),\ (3,6),\ (3,7),\ (4,6)\}$$

如果这 6 个样本点中的任何一个出现的话，我们就说事件 C 发生了。

KP&L 项目经理还可能关心的其他事件包括：

L = 项目完成时间小于 10 个月的事件

M = 项目完成时间超过 10 个月的事件

利用表 4.3 的信息，我们发现这些事件包括下列样本点：

$$L = \{(2,6),\ (2,7),\ (3,6)\}$$
$$M = \{(3,8),\ (4,7),\ (4,8)\}$$

KP&L 问题中的各个事件都能够定义，但无论是哪种情况，在确认事件时，都要注意它是否是试验样本点的集合。

给定如表 4.3 所示的样本点概率，我们就能够利用下面的定义计算 KP&L 经理需要考虑的任何一个事件的概率。

事件的概率

任何事件的概率就等于构成该事件的所有样本点的概率之和。

利用该定义，我们通过对构成事件的所有样本点的概率求和，就可以计算某个事件的概率。现在我们来计算项目将在 10 个月以内完成的概率。因为该事件可表示为 $C = \{(2,6),\ (2,7),\ (2,8),\ (3,6),\ (3,7),\ (4,6)\}$，所以事件 C 的概率 $P(C)$由以下公式给出：

$$P(C) = P(2,6) + P(2,7) + P(2,8) + P(3,6) + P(3,7) + P(4,6)$$

使用表 4.3 样本点的概率，我们有

$$P(C) = 0.15 + 0.15 + 0.05 + 0.10 + 0.20 + 0.05 = 0.70$$

类似地，由于项目完成时间小于 10 个月的事件就是 $L = \{(2,6),\ (2,7),\ (3,6)\}$，该事件的概率可计算如下：

$$P(L) = P(2,6) + P(2,7) + P(3,6)$$
$$= 0.15 + 0.15 + 0.10 = 0.40$$

最后，对于项目完成时间大于 10 个月的事件，我们有 $M = \{(3,8),\ (4,7),\ (4,8)\}$，因此

$$P(M) = P(3,8) + P(4,7) + P(4,8)$$
$$= 0.05 + 0.10 + 0.15 = 0.30$$

利用这些概率结果，现在我们能够告诉 KP&L 的经理，项目在 10 个月以内完工的概率是 0.70，完工时间少于 10 个月的概率是 0.40，完工时间超过 10 个月的概率是 0.30。这种计算事件概率的方法可以对 KP&L 经理关心的任何事件重复使用。

只要我们能够确认一个试验所有的样本点，并且能够向每个样本点指派概率，我们就可以利用定义来计算事件的概率。但是，有许多试验的样本点数目非常庞大，以至于要确认所有的样本点以及确定与之相关的概率，即使可能，也是非常繁琐的。在本章剩下的部分，我们将介绍一些基本的概率关系，利用这些基本关系，我们可以在不知道所有样本点概率的情况下，计算事件的概率。

评　注

1. 样本空间 S 也是一个事件。因为它包含了所有的试验结果，故样本空间的概率等于 1。即 $P(S) = 1$。
2. 在用古典方法指派概率时，其假设条件是每个试验结果发生的可能性相等。在这种情况下，通过计算事件中包括的试验结果数目，再除以试验结果的总数，就能够计算出事件的概率。

练　习

方　法

14. 某试验有 4 个等可能的结果：E_1、E_2、E_3 和 E_4。
    a. E_2 发生的概率是多少？
    b. 任意两个结果(如 E_1 或 E_3)发生的概率是多少？
    c. 任意三个结果(如 E_1 或 E_2 或 E_4)发生的概率是多少？

15. 考虑从一副 52 张牌中抽取 1 张牌的试验，每张牌对应的样本点的概率为 1/52。
    自测题
    a. 列出事件“抽中 A”的所有样本点。
    b. 列出事件“抽中草花”的所有样本点。
    c. 列出事件“抽中花牌(J、Q、K)”的所有样本点。
    d. 计算与(a)、(b)、(c)中每个事件相关的概率。

16. 考虑投掷一对色子的试验。假定我们关心的是两枚色子朝上一面所显示点数的总和。
    a. 有多少个样本点是可能的？(提示：利用多步骤试验的计数法则)
    b. 列出所有样本点。

c. 点数之和为 7 的概率是多少？

d. 点数之和为 9 或更大值的概率是多少？

e. 因为每次投掷能得到 6 个可能的偶数值 (2，4，6，8，10，12)和 5 个可能的奇数值 (3,5，7,9,11)，色子显示的点数之和为偶数的情况将比出现奇数的情况更经常地发生。你同意这种说法吗？请解释。

f. 你使用什么方法来指派需要的概率？

应　用

17. 参见表 4.3 KP&L 公司的样本点和样本点概率。



a. 如果在设计阶段(阶段 1)完成时间用去 4 个月的话，将超出预算。对于事件“设计阶段超出预算”，列出所有样本点。

b. 设计阶段超预算的概率是多少？

c. 如果建设阶段(阶段 2)完成时间用去 8 个月的话，将超出预算。对于事件“建设阶段超预算”，列出所有样本点。

d. 建设阶段超预算的概率是多少？

e. 每个阶段都超出预算的概率是多少？

18. 假设一个大型公寓的经理对于下个月存在的空房数量指派了以下的主观概率估计：

空房数	概率
0	0.05
1	0.15
2	0.35
3	0.25
4	0.10
5	0.10

列出下面每个事件的样本点，并给出事件的概率。

a. 无空房。

b. 至少有 4 套空房。

c. 两套或更少的空房。

19. 某家具店的经理每周销售 0—4 只中国柜子，在过去经验的基础上，他指派给各个销量的概率为：$P(0) = 0.08$，$P(1) = 0.18$，$P(2) = 0.32$，$P(3) = 0.30$，$P(4) = 0.12$。

a. 这些概率指派有效吗？为什么？

b. 以 A 代表一周销售小于等于 2 只的事件，计算 $P(A)$。

c. 以 B 代表一周销售大于等于 4 只的事件，计算 $P(B)$。

20. 《财富》杂志发行包括财富 500 强公司信息在内的年度特刊。下列数据显示了拥有财富 500 强

公司最多的 6 个州，以及位于这些州的公司总部的数量(*Fortune*, April 17, 2000)：

州名	公司数目
纽约	56
加利福尼亚	53
得克萨斯	43
伊利诺伊	37
俄亥俄	28
宾夕法尼亚	28

假设要选取一家 500 强公司进行问卷调查，下列事件的概率分别是多少？

a. 以 N 代表公司总部位于纽约的事件，计算 $P(N)$。

b. 以 T 代表公司总部位于得克萨斯的事件，计算 $P(T)$。

c. 以 B 代表公司总部位于这 6 个州中任何一个的事件，计算 $P(B)$。

21. 对 Tarpon Springs 学院的 50 名学生进行了课外活动情况调查，结果如下：

参加课外活动的次数	频数
0	8
1	20
2	12
3	6
4	3
5	1

a. 以 A 代表学生至少参加 1 次活动的事件，计算 $P(A)$。

b. 以 B 代表学生参加 3 次或以上活动的事件，计算 $P(B)$。

c. 学生恰好参加 2 次活动的概率是多少？

4.3　概率的一些基本关系

事件的补

给定一个事件 A，**A 的补**(complement of A)就定义为：是由所有不属于 A 的样本点构成的事件。A 的补记作 A^c，图 4.4 被称为**维恩图**(Venn diagram)，它图示说明了补的概念。图上的矩形区域代表试验的样本空间，因此包含了所有可能的样本点。圆圈代表事件 A，只包含属于 A 的样本点。矩形的阴影区域包含了所有不属于事件 A 的样本点，根据定义，它就表示 A 的补。

在任何有关概率的应用中，要么是事件 A，要么是它的补 A^c 必然会发生。因此，我们有

$$P(A) + P(A^c) = 1$$

为了解出 $P(A)$，我们得到下列公式：

利用补来计算概率

$$P(A) = 1 - P(A^C) \tag{4.5}$$

公式(4.5)表示：如果事件 A 的补的概率 $P(A^C)$已知的话，就能容易地计算出事件 A 的概率。

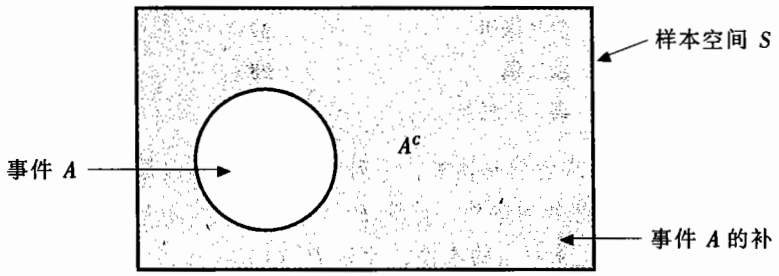


图 4.4　事件 A 的补

我们考虑下面的事例：销售经理在审查过销售报告以后，发现有 80% 的对新顾客的接触没有销售出产品。用 A 来表示销售事件，用 A^C 表示没有销售事件，经理发现 $P(A^C) = 0.80$。利用公式(4.5)，我们有

$$P(A) = 1 - P(A^C) = 1 - 0.80 = 0.20$$

我们可以得到结论，对新顾客的接触有 20% 的概率能够销售出产品。

再举另外一个例子，采购代理发现供应者发出无缺陷货物的概率为 0.90。利用补，我们能够得到货物中含有缺陷产品的概率是：$1 - 0.90 = 0.10$。

加法法则

当我们面对两个事件，并且想要知道两个事件中至少有一个发生的概率时，加法法则就能够发挥作用。也就是说，对于事件 A 和 B，我们想知道事件 A 或事件 B 或者两者都发生的概率。

在介绍加法法则之前，我们需要讨论两个与事件的组合有关的概念：事件的**并**和事件的**交**。给定两个事件 A 和 B，**A 和 B 的并**(union of A and B)定义如下：

两事件的并

A 和 B 的并就是这样一个事件，该事件包含属于 A 的或属于 B 的或同时属于两者的所有样本点，记作 $A \cup B$。

一个事件和它的补的并集就是整个样本空间。

图 4.5 的维恩图描述了事件 A 和 B 的并。注意到两个圆圈既包含了事件 A 的所有样本点，

也包含了事件 B 的所有样本点。两个圆圈有部分重叠的情况表明，一些样本点同时属于 A 和 B。

两事件 A 和 B 的交(intersection of two events A and B)定义如下：

两事件的交

给定两个事件 A 和 B，A 和 B 的交就是包含了所有**同时属于**事件 A 和 B 的样本点的事件，记作 $A \cap B$。

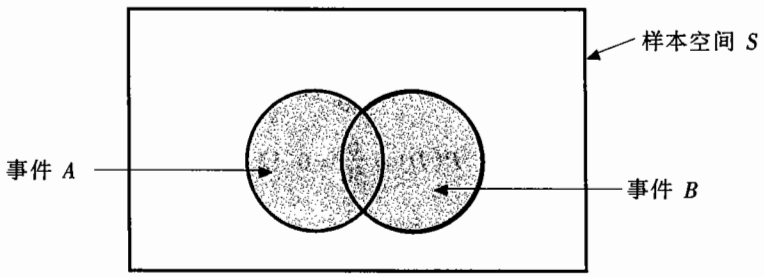


图 4.5　事件 A 和 B 的并

图 4.6 的维恩图描述了两事件的交。两个圆圈重叠的区域就是交，它包含了既属于 A 又属于 B 的样本点。

现在让我们继续进行加法法则的讨论。**加法法则**(addition law)提供了一种方法，可用来计算事件 A 发生或 B 发生或者两者同时发生的概率。换句话说，加法法则能够计算两个事件的并的概率。加法法则的数学表达如下：

加法法则

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.6}$$

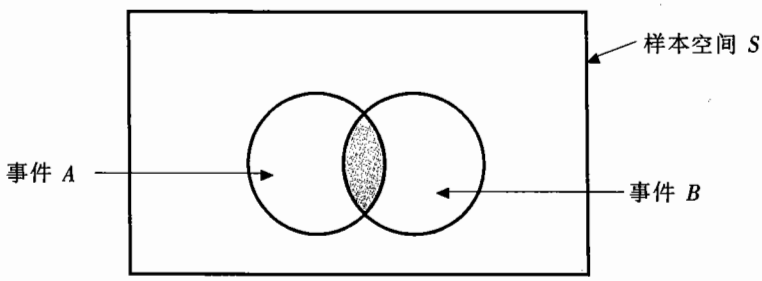


图 4.6　事件 A 和 B 的交

为了直观地理解加法法则，我们注意到加法法则公式的头两项：$P(A) + P(B)$，它包括了 $A \cup B$ 中的所有样本点。但是，因为交 $A \cap B$ 中的样本点同时属于 A 和 B，当我们计算 $P(A) + P(B)$时，实际上是把 $A \cap B$ 中的每一个样本点计算了两遍。于是，通过减去 $P(A \cap B)$，我们能够把重复计算给修正回来。

作为加法法则的一个应用例子，我们来考虑一个拥有 50 名雇员的小型加工厂案例。每个工人都被要求按时完成指派给他们的工作，并且所加工的产品还要通过最后的品质检查。但有时一些

工人因为没有按时完成工作和/或生产了次品，而没有达到操作标准。在操作情况评价阶段的末期，生产经理发现：50个工人中有5个未按时完成工作；有6个生产了次品；有2个工人既未按时完成工作又生产了次品。我们令

L = 未按时完成工作的事件；
D = 生产了次品的事件。

利用相对频数信息可以得到下列概率：

$$P(L) = \frac{5}{50} = 0.10$$

$$P(D) = \frac{6}{50} = 0.12$$

$$P(L \cap D) = \frac{2}{50} = 0.04$$

在检查完操作数据以后，生产经理决定把差的操作等级给予那些未按时完成工作或生产了次品的工人。与此相关的事件是 $L \cup D$。问生产经理给予雇员一个差操作等级的概率是多少？

我们注意到这个概率问题是关于两事件并的，需要求的是 $P(L \cup D)$。使用公式(4.6)，我们有

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

因为该公式右边的三个概率值已知，我们能够得出

$$P(L \cup D) = 0.10 + 0.12 - 0.04 = 0.18$$

这个结果表示随机选择到差等级工人的概率是0.18。

再举一个加法法则的例子，最近由一家主要计算机软件公司的人事经理进行了一项研究。结果发现在两年内离开公司的雇员当中，有30%是因为不满意薪金状况，有20%是因为对工作指派不满，有12%的上述雇员指出他们既对薪金待遇又对工作指派不满意。问在两年内离开公司的雇员当中，离职原因是对薪金不满，或对工作指派不满，或对两者都不满意的概率是多少？我们令

S = 雇员因为薪金离职的事件；
W = 雇员因为工作指派离职的事件。

有 $P(S) = 0.30$，$P(W) = 0.20$，$P(S \cap W) = 0.12$。使用公式(4.6)，我们得到

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38$$

即雇员因为薪金或工作指派原因离职的概率为0.38。

在结束加法法则的讨论之前，再来考虑一下由**互斥事件**(mutually exclusive events)引起的特殊情况。

互斥事件

如果两个事件没有共同的样本点，我们就称这两个事件是**互斥**的。

如果事件 A 和 B 中一个事件发生时，另一个事件就不会发生的话，它们就是互斥事件。因此，A 和 B 互斥要求它们的交必须不包含任何样本点。图 4.7 的维恩图描述了两个互斥事件。在这种情况下 $P(A \cap B) = 0$，并且加法法则能够表达为如下形式：

互斥事件的加法法则

$$P(A \cup B) = P(A) + P(B)$$

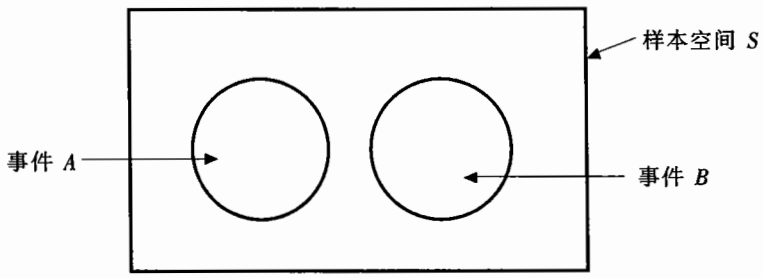


图 4.7　互斥事件

练　习

方　法

22. 假定样本空间具有 5 个可能性相等的试验结果：E_1，E_2，E_3，E_4，E_5。我们令

$$A = \{E_1, E_2\}$$
$$B = \{E_3, E_4\}$$
$$C = \{E_2, E_3, E_5\}$$

a. 计算 $P(A)$、$P(B)$和 $P(C)$。

b. 计算 $P(A \cup B)$。A 和 B 是互斥的吗？

c. 计算 A^c、C^c、$P(A^c)$、$P(C^c)$。

d. 计算 $A \cup B^c$ 和 $P(A \cup B^c)$。

e. 计算 $P(B \cup C)$。

23. 自测题　假定我们有样本空间 $S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$，其中 E_1，E_2，…，E_7 表示样本点。给定下面的概率指派：$P(E_1) = 0.05$，$P(E_2) = 0.20$，$P(E_3) = 0.20$，$P(E_4) = 0.25$，$P(E_5) = 0.15$，$P(E_6) = 0.10$，$P(E_7) = 0.05$。我们令

$$A = \{E_1, E_4, E_6\}$$
$$B = \{E_2, E_4, E_7\}$$
$$C = \{E_2, E_3, E_5, E_7\}$$

a. 计算 $P(A)$、$P(B)$和 $P(C)$。
b. 计算 $A \cup B$ 和 $P(A \cup B)$。
c. 计算 $A \cap B$ 和 $P(A \cap B)$。
d. 事件 A 和 C 是互斥的吗？
e. 计算 B^c 和 $P(B^c)$。

应　用

24. 根据《商业周刊》的 Harris 民意测验，有 14% 的成年人认为在 1998 年股市很有可能暴跌，而 43% 的成年人认为在 1998 年股市有可能暴跌（*Business Week*, December 29, 1997）。如果我们随机地选择一名成年人，则他或她认为股市不可能暴跌的概率是多少？

25. 30 家最大的债券基金提交了期间截止到 2000 年 3 月 31 日的 1 年和 5 年期投资百分比回报率数据（*The Wall Street Journal*, April 10, 2000）。假定我们认为 1 年期回报率超过 2% 属于高回报，5 年期回报率超过 44% 属于高回报。基金中有一半 1 年期回报率超过了 2%，有 12 家基金的 5 年期回报率超过了 44%，有 6 家基金 1 年期回报率超过了 2% 且 5 年期回报率超过了 44%。
   a. 计算基金取得 1 年期高回报率的概率，取得 5 年期高回报率的概率，以及取得 1 年期高回报率又取得 5 年期高回报率的概率。
   b. 基金取得 1 年期高回报率或者 5 年期高回报率或者同时取得两者的概率是多少？
   c. 基金既未取得 1 年期高回报率又未取得 5 年期高回报率的概率是多少？

26. 30 家最大的股票和权益基金提交了期间截止到 2000 年 3 月 31 日的 1 年和 5 年期百分比投资回报率数据（*The Wall Street Journal*, April 10, 2000）。假定我们认为 1 年期回报率超过 50% 属于高回报，5 年期回报率超过 300% 属于高回报。有 9 家基金的 1 年期回报率超过了 50%，7 家基金的 5 年期回报率超过了 300%，5 家基金的 1 年期回报率超过了 50% 且 5 年期回报率超过了 300%。
   a. 基金取得 1 年期高回报率的概率是多少？取得 5 年期高回报率的概率是多少？
   b. 基金既取得 1 年期高回报率又取得 5 年期高回报率的概率是多少？
   c. 基金既未取得 1 年期高回报率又未取得 5 年期高回报率的概率是多少？

27. 《商业周刊》对其消费者进行了一次关于他们平时酒类消费的调查，结果发现有 57% 的消费者消费国产酒，33% 的消费者消费进口酒，63% 的消费者消费过酒（*Business Week* 1996 Worldwide Subscriber Study）。问《商业周刊》的消费者平时既消费国产酒又消费进口酒的概率是多少？

28. 自测题 一个对杂志消费者的调查表明：过去 12 个月以来，他们中的 45.8% 因为商务原因而租车，54% 因为私人原因而租车，30% 既因商务又因私人原因租车。

a. 过去 12 个月以来，消费者因为商务或私人原因而租车的概率是多少？

b. 过去 12 个月以来，消费者既不为商务又不为私人原因而租车的概率是多少？

29. 根据人口普查署的数据，美国的人口死亡率为每年 2 425 000 人。国家健康统计中心报告称 1997年导致死亡的三个主要原因分别是：心脏病（725 790）、癌症（537 390）以及中风（159 877）。我们以 H、C、S 来分别代表某人死于心脏病、癌症和中风的事件。

a. 利用数据来估计 $P(H)$、$P(C)$和 $P(S)$。

b. 事件 H和 C是互斥的吗？计算 $P(H \cap C)$。

c. 某人死于心脏病或癌症的概率是多少？

d. 某人死于癌症或中风的概率是多少？

e. 计算除以上三种原因以外，其他原因导致死亡的概率是多少？

4.4　条件概率

一个事件的概率往往受到相关事件是否发生的影响。假定有一个事件 A，其概率为 $P(A)$。如果我们得到了新的信息，知道以 B 表示的相关事件已经发生，我们需要利用这个信息来计算事件 A 的新概率。事件 A 的新概率就被称为**条件概率**(conditional probability)，记做 $P(A|B)$。符号“|”表示这样一个事实：我们在给定事件 B 已经发生的条件下考虑事件 A 的概率。因此，符号 $P(A|B)$ 读做“给定 B 条件下 A 的概率”。

举一个条件概率的应用例子，考虑一个在美国东部某中心城市警察局的男性和女性职员晋升的情况。该警察局有 1 200 人，包括 960 位男职员和 240 位女职员。在过去两年里，有 324 人得到了提升。其中男性职员和女性职员的具体晋升情况显示在表 4.4 中。

表 4.4　过去两年来警察局职员的晋升情况

	男职员	女职员	总计
晋升	288	36	324
未晋升	672	204	876
总计	960	240	1 200

在检查过有关晋升记录以后，基于有 288 名男职员受到了提升，而只有 36 名女职员得到提升这一事实，女职员委员会对该警察局提出了性别歧视指控。警察局的管理层争辩说：女性职员相对较低的提升数不是由于性别歧视造成的，而是因为警察局里的女性人数本来就较少。现在让我们来说明如何利用条件概率来分析歧视指控。

令

M = 职员是男性事件；

W = 职员是女性事件；

A = 职员得到晋升事件；

A^c = 职员未被晋升事件。

把表 4.4 的数据值除以总职员数 1 200，能够使我们将可用的信息汇总为下列概率值：

$P(M \cap A) = 288/1\,200 = 0.24$ = 随机选择到的职员是男性并且得到晋升的概率
$P(M \cap A^C) = 672/1\,200 = 0.56$ = 随机选择到的职员是男性且未得到晋升的概率
$P(W \cap A) = 36/1\,200 = 0.03$ = 随机选择到的职员是女性并且得到晋升的概率
$P(W \cap A^C) = 204/1\,200 = 0.17$ = 随机选择到的职员是女性且未得到晋升的概率

因为这些值的每一个都给出了两个事件的交的概率，故被称为**联合概率**(joint probabilities)。表4.5汇总了警察局职员的晋升情况，它就是一个**联合概率表**。

表 4.5 晋升的联合概率表

显示在表中间的联合概率	男性（M）	女性（W）	总计
晋升（A）	0.24	0.03	0.27
未晋升（A^C）	0.56	0.17	0.73
总计	0.80	0.20	1.00

显示在表边缘的边际概率

联合概率表边缘的值提供了每个单独事件的概率，它们是：$P(M) = 0.80$，$P(W) = 0.20$，$P(A) = 0.27$，$P(A^C) = 0.73$。因为这些概率的位置处在联合概率表的边缘，所以被称为**边际概率**。我们注意到通过把联合概率表中对应的每行或每列的联合概率值加总起来，也能得到边际概率。例如，被提升的边际概率是：$P(A) = P(M \cap A) + P(W \cap A) = 0.24 + 0.03 = 0.27$。根据边际概率，我们看到 80% 的警察是男性，20% 的警察是女性；所有警察中的 27% 得到了提升，73% 未被提升。

我们通过计算职员在给定是男性的情况下得到提升的概率，来开始条件概率的分析。使用条件概率的符号，我们想要确定的是 $P(A \mid M)$。为了计算 $P(A \mid M)$，我们首先意识到该符号意味着：在给定的事件 M(职员是男性)已经存在的条件下考虑事件 A(晋升)的概率。因此 $P(A \mid M)$告诉我们：我们现在只关心 960 位男性职员的晋升情况。因为 960 位男性职员中有 288 位得到了晋升，所以职员在给定为男性条件下得到晋升的概率是 $288/960 = 0.30$。换句话说，给定某职员为男性，那么他有 30% 的机会在过去两年内获得晋升。

因为表 4.4 的值显示了每个类型的职员数，该方法较易于使用。但我们现在想要说明的是怎样利用相关的事件概率，而不是表 4.4 的频数数据，来计算像 $P(A \mid M)$这样的条件概率。我们已经知道 $P(A \mid M) = 288/960 = 0.30$。现在我们把分子和分母同除以 1 200，即本问题中的职员总数：

$$P(A \mid M) = \frac{288}{960} = \frac{288/1\,200}{960/1\,200} = \frac{0.24}{0.80} = 0.30$$

现在我们看到能够通过 $0.24/0.80$ 来计算条件概率 $P(A \mid M)$。参见联合概率表 4.5，尤其是注意到 0.24 是 A 和 M 的联合概率，即 $P(A \cap M) = 0.24$。我们还注意到 0.80 是随机选择到男性职员的边际概率，即 $P(M) = 0.80$。因此，条件概率 $P(A \mid M)$能够作为联合概率和边际概率 $P(A \cap M)$的比率 $P(M)$而得出。

$$P(A \mid M) = \frac{P(A \cap M)}{P(M)} = \frac{0.24}{0.80} = 0.30$$

条件概率能够用联合概率比上边际概率而计算出来的事实，使得我们可以得到下列计算两事件 A 和 B 条件概率的通用公式：

条件概率

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

或

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

图 4.8 的维恩图有助于对条件概率的直观理解。右边的圆圈表示事件 B 已经发生，与事件 A 重叠的圆圈部分代表事件$(A \cap B)$。我们知道一旦事件 B 发生，则能够观察到事件 A 的惟一途径是事件$(A \cap B)$发生。因此，比率 $P(A \cap B)/P(B)$给出了在给定事件 B 已经发生的情况下观察到事件 A 的条件概率。

让我们回到针对女性职员的歧视问题。表 4.5 第一行的边际概率显示：一个职员得到晋升的概率是 $P(A) = 0.27$(不管职员是男性还是女性)。但是，歧视问题上的批评意见涉及到两个条件概率：$P(A \mid M)$和 $P(A \mid W)$。也就是说，职员在给定为男性的条件下，得到晋升的概率是多少？职员在给定为女性的条件下，得到晋升的概率是多少？如果这两个概率是相等的，则说明男性职员和女性职员晋升机会相同，因此性别歧视的指责将失去依据。不过，如果两个条件概率不同，就将支持关于男性和女性职员在晋升问题上受到区别对待的指责。

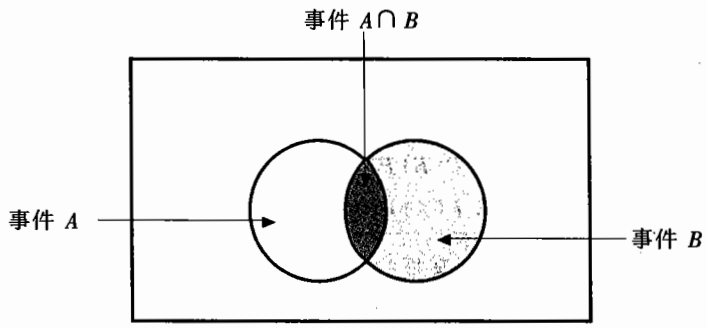


图 4.8　条件概率

我们已经确定了 $P(A \mid M) = 0.30$，现在让我们利用表 4.5 的概率值和公式(4.7)的条件概率基本关系，来计算一个职员在给定是女性的条件下得到晋升的概率，即 $P(A \mid W)$。利用公式(4.7)，我们得到

$$P(A \mid W) = \frac{P(A \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

根据这个结果，你得出了什么结论？一个职员在给定为男性的条件下得到晋升的概率是 0.30，两

倍于在给定为女性条件下的晋升概率 0.15。尽管使用条件概率本身不能证明在本例中存在性别歧视，但得出的条件概率值支持女性职员提出的指责。

独立事件

在上例中，$P(A)=0.27$，$P(A\mid M)=0.30$，$P(A\mid W)=0.15$。我们看到：晋升(事件 A)的概率受到职员是男性还是女性的影响。尤其是，因为 $P(A\mid M)\neq P(A)$，我们称事件 A 和 M 是相关事件。即事件 M(职员为男性)是否发生影响或改变了事件 A(晋升)的概率。类似地，因为，我们称事件 A 和 W 是相关事件。但是，如果事件 A 的概率不因发生事件 M 而改变的话，即 $P(A\mid M)=P(A)$，我们称事件 A 和 M 是**独立事件**(independent events)。两个事件互相独立的定义如下：

独立事件

如果 $$P(A\mid B)=P(A) \tag{4.9}$$

或 $$P(B\mid A)=P(B) \tag{4.10}$$

则称两事件 A 和 B 是独立的；否则称两事件是相关的。

乘法法则

加法法则用来计算两事件并的概率，而乘法法则用来计算两事件交的概率。乘法法则的基础是条件概率的定义，利用公式(4.7)和(4.8)解出，我们就可得到**乘法法则**(multiplication law)。

乘法法则

$$P(A\cap B)=P(B)P(A\mid B) \tag{4.11}$$

或 $$P(A\cap B)=P(A)P(B\mid A) \tag{4.12}$$

为了说明乘法法则的应用，考虑该问题：某报的发行部已经知道在某社区有 84% 的住户订阅了该报纸的日报。我们用 D 来代表事件“住户订阅了日报”，$P(D)=0.84$。另外，还知道已订阅了日报的住户订阅其周末刊的概率(事件 S)为 0.75，即 $P(S\mid D)=0.75$。则住户既订阅了日报，又订阅周末刊的概率是多少?利用乘法法则，我们可以算出所要的 $P(S\cap D)$，计算过程为

$$P(S\cap D)=P(D)P(S\mid D)=0.84\times 0.75=0.63$$

现在我们知道有 63% 的住户既订阅了日报又订阅了周末刊。

在结束本节之前，我们再来考虑一下涉及到独立事件这一特殊情况的乘法法则。我们已知当 $P(A\mid B)=P(A)$或 $P(B\mid A)=P(B)$时，事件 A 和 B 是独立的。因此利用公式(4.11)和(4.12)，针对独立事件的特殊情况，我们得到下面的乘法法则：

独立事件的乘法法则

$$P(A \cap B) = P(A)\,P(B) \tag{4.13}$$

为了计算两独立事件交的概率，我们只需把相应的概率相乘即可。注意到独立事件的乘法法则提供了确定 A 和 B 是否独立的另外一种方法，即如果 $P(A \cap B) = P(A)\,P(B)$，那么 A 和 B 是独立的；如果 $P(A \cap B) \neq P(A)\,P(B)$，那么 A 和 B 是相关的。

作为独立事件乘法法则的一个应用实例，考虑这样一种情况：一位加油站经理依据以往的经验知道，有 80% 的顾客在加油时使用信用卡。问接连两名顾客都使用信用卡加油的概率是多少？如果我们令

A = 第一个顾客使用信用卡的事件；
B = 第二个顾客使用信用卡的事件。

那么与问题有关的事件就是 $A \cap B$。在未给出其他信息的情况下，我们能够合理地假定 A 和 B 是独立事件。因此

$$P(A \cap B) = P(A)\,P(B) = 0.80 \times 0.80 = 0.64$$

对本节进行总结，我们注意到：由于事件往往是相关的，因此我们关心条件概率，并且在计算条件概率时必须使用公式(4.7)和(4.8)。如果两个事件不相关的话，则它们是相互独立的，任何一个事件都不受另一事件是否发生的影响。

评　注

不要混淆互斥事件和独立事件的概念。两个具有非零概率的事件不可能既是互斥的，同时又是互相独立的。如果已知互斥事件中有一个发生的话，另外一个事件就不可能发生。因此它们是相关事件。

练　习

方　法

30. 假定有两个事件 A 和 B，$P(A) = 0.50$，$P(B) = 0.60$，而 $P(A \cap B) = 0.40$。

自测题

a. 计算 $P(A \mid B)$。

b. 计算 $P(B \mid A)$。

c. A 和 B 相互独立吗？为什么？

31. 假定两事件 A 和 B 互斥，再进一步假定我们已知 $P(A) = 0.30$，$P(B) = 0.40$。

a. $P(A \cap B)$是多少？

b. $P(A \mid B)$是多少？

c. 一名统计学学生争论说：互斥事件和独立事件的概念实际上是相同的，如果事件互斥的话，那么它们必定是独立的。你赞同这种说法吗？利用本题中的概率信息来验证你的答案。

d. 根据本题的结果，关于互斥事件和独立事件你能够得到什么样的一般结论？

应　用

32. 关于 140 位顾客的年龄和婚姻状况，Daytona Beach 夜总会有如下数据：

		婚姻状况	
		单身	已婚
年龄	30 岁以下	77	14
	30 岁或以上	28	21

a. 构建这些数据的联合概率表。

b. 使用边际概率对夜总会顾客的年龄进行评论。

c. 使用边际概率对夜总会顾客的婚姻状况进行评论。

d. 顾客既是单身同时年龄又在 30 岁以下的概率是多少？

e. 如果某位顾客在 30 岁以下，则他或她是单身的概率为多少？

f. 婚姻状况是否独立于年龄？利用概率进行解释。

33. 在一项对 MBA 学生的调查中，对于学生申请学校的首要原因，获得了下列数据：

自测题

		申请的首要原因			
		学校质量	学费或方便性	其他	总计
注册状况	全日制	421	393	76	890
	非全日制	400	593	46	1 039
	总计	821	986	122	1 929

a. 构建这些数据的联合概率表。

b. 利用学校质量、学费/方便性、其他原因的边际概率，对选择学校的首要原因进行评论。

c. 如果某学生选择了全日制方式，学校质量是择校首要原因的概率是多少？

d. 如果某学生选择了非全日制方式，学校质量是择校首要原因的概率是多少？

e. 以 A 代表学生选择全日制的事件，以 B 代表将学校质量作为申请首要原因的事件。事件 A 和 B 相互独立吗？验证你的答案。

34. 下表显示了一般人群的血型分布(俄亥俄辛辛那提 Hoxworth 血液中心提供)：

	A	B	AB	O
Rh +	34%	9%	4%	38%
Rh –	6%	2%	1%	6%

a. 某人具有 O 型血的概率是多少？

b. 某人属于 Rh – 的概率是多少？

c. 夫妇两人都属于 Rh – 的概率是多少？

d. 夫妇两人血型都是 AB 的概率是多少？

e. 给定某人血型为 O 的条件下，他或她属于 Rh – 的概率是多少？

f. 给定某人属于 Rh + 的条件下，他或她具有 B 型血的概率是多少？

35. “从 1950 年以来，1 月的股市行情简直就是该年股市行情的晴雨表，其准确程度令人吃惊”（*1998Stock Trader's Almanac*）。从 1950—1997 的 48 年间，股市在 1 月上涨过 31 次，全年上涨过 36 次，既在 1 月上涨全年又上涨的有 29 次。

a. 估计股市将在 1 月上涨的概率。

b. 估计股市将全年上涨的概率。

c. 给定股市在 1 月上涨的条件下，它在整个年度上涨的概率是多少？

d. 概率值是否表明股市的 1 月份表现与它的全年表现是相互独立的事件？请解释。

36. 对四种职业进行了工作满意度研究：木工、律师、理疗师和系统分析师。满意度以 0—100 分来衡量，该研究得到的数据列示在下面的交叉分组列表中：

	满意度得分				
职业	50 分以下	50 — 59	60 — 69	70 — 79	80 — 89
木工	0	2	4	3	1
律师	6	2	1	1	0
理疗师	0	5	2	1	2
系统分析师	2	1	4	3	0

a. 构建联合概率表。

b. 某位研究对象满意度得分为 80 多分的概率是多少？

c. 给定研究对象为理疗师的条件下，其满意度得分为 80 多分的概率是多少？

d. 某位研究对象的职业为律师的概率是多少？

e. 某位研究对象的职业为律师并且其满意度得分在 50 以下的概率是多少？

f. 给定某人为律师的条件下，其满意度得分在 50 以下的概率是多少？

g. 研究对象满意度得分在 70 或以上的概率是多少？

37. 某采购代理向两个供应商 *A* 和 *B* 紧急订购了一批特殊原材料。如果两家供应商的货物均未在 4 日内到达，厂家将被迫关闭生产线直到货物运达为止。供应商 *A* 能够在 4 日内交货的概率是 0.55，供应商 *B* 能够在 4 日内交货的概率是 0.35。

a. 两家供应商都能够在 4 日以内交货的概率是多少？因为涉及到两家无关的厂商，我们假定

它们是互相独立的。

b. 至少有一家供应商在4日以内交货的概率是多少?

c. 在4日内由于原材料短缺而要关闭生产线(即两张订单都被延误)的概率是多少?

38. 由某机构对 the Future and the Gallup 组织的1 035名工作人员所做的调查发现:员工们正在被众多的消息淹没(*The Cincinnati Enquirer*, November 2, 1998)。研究还指出每个员工平均每天收到190条消息。下表显示了消息类型的分类数据:

消息来源	每日消息数量	消息来源	每日消息数量
电话	52	电子邮件	30
声音邮件	22	部门间邮件	18
美国邮政邮件	18	传真	15
邮政卡	11	电话消息纸条	10
活页	4	晚间急件	4
移动电话	3	美国快递邮件	3

a. 对任一员工而言,下一条消息来自电话的概率是多少?

b. 对相邻的两条消息,第一条来自电子邮件而第二条来自传真的概率是多少?

c. 下一条消息出自电话或者是部门间邮件的概率是多少?

4.5 贝叶斯定理

在对条件概率的讨论中,我们指出:当我们得到新的信息时,进行概率分析的一个重要方面就是要根据新的信息修正概率。我们往往要先对有关的具体事件进行原始的概率估计或者说是**先验概率**(prior probability)估计,以此来开始分析工作。然后,我们从一些诸如样品、特殊报告或产品检测等信息来源中获取有关该事件的其他信息。给定这些信息以后,我们就可以通过计算把先验概率修正为**后验概率**(posterior probabilities)。**贝叶斯定理**(Bayes' theorem)提供了进行修正概率计算的方法,该方法的具体步骤显示在图4.9中。

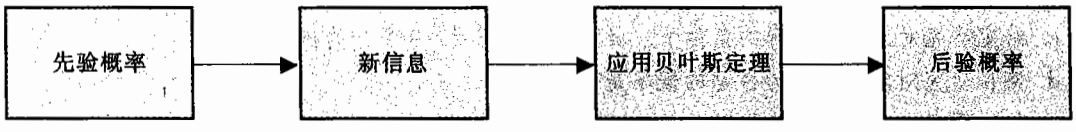


图4.9 利用贝叶斯定理进行概率修正

作为贝叶斯定理的一个应用例子,让我们考虑某个制造业公司,它从两个不同的供应商那里购买零件。令 A_1 代表零件来自供应商1的事件,A_2 代表零件来自供应商2的事件。目前该公司购买的零件有65%来自供应商1,剩下的35%则来自供应商2。因此,若随机地选择零件,我们指派的先验概率为:$P(A_1) = 0.65$,$P(A_2) = 0.35$。

该公司所采购零件的质量随供应商的不同而变化，两个供应商所提供的零件质量的历史数据见表 4.6。如果我们以 G 来代表零件质量好的事件，以 B 代表零件质量差的事件，表 4.6 的信息给出了下面的条件概率值：

$$P(G|A_1)=0.98 \qquad P(B|A_1)=0.02$$
$$P(G|A_2)=0.95 \qquad P(B|A_2)=0.05$$

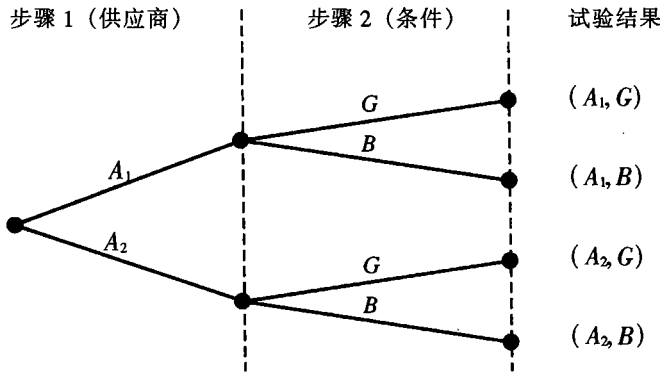
表 4.6　两供应商的历史质量水平

	好零件的百分比	差零件的百分比
供应商 1	98	2
供应商 2	95	5

图 4.10 的树形图描述了这个两步骤试验：公司首先从两供应商之一处购得零件，然后再检验某个零件是好的还是差的。我们看到共有四个可能的试验结果，两个试验结果对应于好零件的情况，两个对应于差零件的情况。

因为每个试验结果都是两事件的交，故我们能够利用乘法法则来计算概率。例如：

$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(G|A_1)$$



注：步骤 1 表示零件来自于两供应商之一；步骤 2 表示零件是好的还是差的。

图 4.10　两步骤树形图

我们还能够使用一种被称为概率树(见图 4.11)的方法来描述这些联合概率的计算过程。在步骤 1 中每个分支的概率是先验概率，而在步骤 2 中每个分支的概率是条件概率。为了得到每个试验结果的概率值，我们只需把通向各试验结果的那条分支上的两个概率值相乘即可。这些联合概率值和每个分支上已知的概率值一起显示在图 4.11 中。

现在我们假定该公司的生产过程使用两供应商的零件，当机器运转差零件时就会出现故障。给定已知零件是差的这样一个信息，那么它来自供应商 1 的概率是多少?来自供应商 2 的概率是多少？根据概率树提供的信息(见图 4.11)，我们可以利用贝叶斯定理来回答这个问题。

令 B 代表差零件事件，我们要找的是后验概率 $P(A_1|B)$和 $P(A_2|B)$。根据条件概率定理，

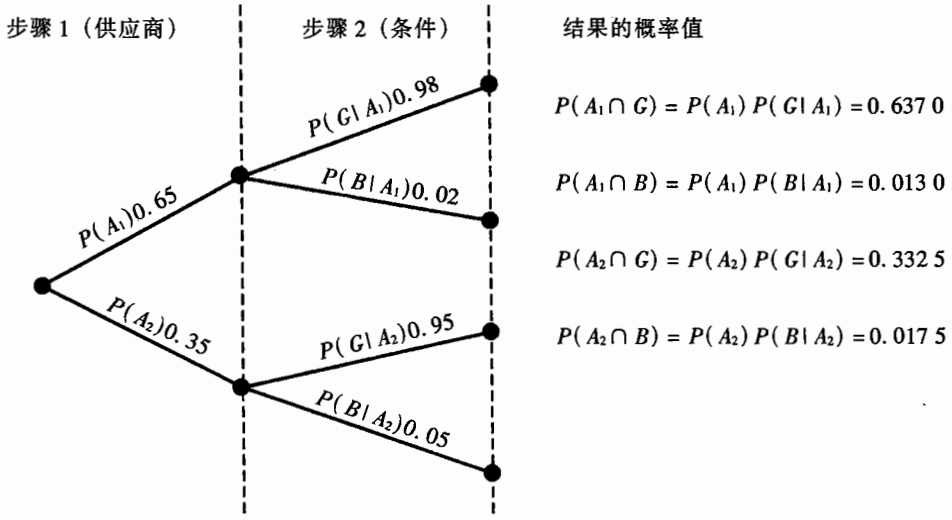


图 4.11　两供应商例子的概率树

我们知道

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} \tag{4.14}$$

参见概率树，我们看到

$$P(A_1 \cap B) = P(A_1)\,P(B \mid A_1) \tag{4.15}$$

为了得到 $P(B)$，我们注意到事件 B 只能以两种途径发生：$(A_1 \cap B)$和$(A_2 \cap B)$。因此，我们有

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(A_1)\,P(B \mid A_1) + P(A_2)\,P(B \mid A_2) \end{aligned} \tag{4.16}$$

将公式(4.15)和(4.16)代入公式(4.14)中，类似地写出 $P(A_2 \mid B)$的结果，我们得到两事件情况下的贝叶斯定理：

贝叶斯定理（两事件情况）

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.17}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.18}$$

我们至今沿用的这一定理源自基督教长老会牧师贝叶斯(Reverend Thomas Bayes)(1702 — 1761)的原创性工作。

利用公式(4.17)和本例的概率值，我们有

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)}$$

$$= \frac{0.65 \times 0.02}{0.65 \times 0.02 + 0.35 \times 0.05} = \frac{0.013\,0}{0.013\,0 + 0.017\,5} = \frac{0.013\,0}{0.030\,5} = 0.426\,2$$

另外，利用公式(4.18)，我们可以得到 $P(A_2 \mid B)$

$$P(A_2 \mid B) = \frac{0.35 \times 0.05}{0.65 \times 0.02 + 0.35 \times 0.05} = \frac{0.017\,5}{0.013\,0 + 0.017\,5} = \frac{0.017\,5}{0.030\,5} = 0.573\,8$$

注意到在这个例子中，开始时随机选择的零件有 0.65 的概率来自供应商 1。但是，给定零件为差的信息以后，零件来自供应商 1 的概率降至 0.426 2。事实上，如果是差零件，那么超过一半的可能是它来自供应商 2，因为 $P(A_2 \mid B) = 0.573\,8$。

当我们想要计算后验概率的事件是互斥的，并且它们的并就是整个样本空间时*，就可以应用贝叶斯定理。贝叶斯定理还能够扩展到包括 A_1，A_2，…，A_n 等 n 个互斥事件的情况，这 n 个事件的并构成了整个样本空间。在这种情况下，计算任一后验概率 $P(A_i \mid B)$的贝叶斯定理有下面的形式：

贝叶斯定理

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4.19}$$

如果已知先验概率 $P(A_1), P(A_2), \cdots, P(A_n)$和对应的条件概率 $P(B \mid A_1), P(B \mid A_2), \cdots, P(B \mid A_n)$，那么我们使用公式（4.19）就能够计算出事件 $A_1, A_2, \cdots, A_n$ 的后验概率。

表格方法

使用表格方法有助于贝叶斯定理的计算，表 4.7 显示了对零件供应商问题如何应用表格方法。其中的计算过程步骤如下：

步骤 1. 准备下面三列：

第 1 列——需要计算后验概率的互斥事件。

第 2 列——事件的先验概率。

第 3 列——每个事件在给定新信息以后的条件概率。

步骤 2. 在第 4 列利用乘法法则计算每个事件的联合概率，并计算新信息 B 的概率。把第 2 列的先验概率乘以第 3 列对应的条件概率就能够得到这些联合概率，即

$$P(A_i \cap B) = P(A_i)\,P(B \mid A_i)$$

---

* 如果事件的并就是整个样本空间，那么这些事件就被称为是完全的。

步骤 3. 对第 4 列的联合概率值求和，其值即为新信息的概率 $P(B)$。我们在表 4.7 中看到：给定供应商为 1 的条件下，差零件概率为 0.013 0；给定供应商为 2 的条件下，差零件的概率为 0.017 5。因为差零件只有这两个可能的来源，故 0.013 0 + 0.017 5 的和表示从来自两供应商的混合货物中发现差零件的综合概率为 0.030 5。

步骤 4. 在第 5 列利用条件概率的基本关系计算后验概率。

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)}$$

注意联合概率 $P(A_i \cap B)$就在第 4 列，而概率 $P(B)$是第 4 列的和。

表 4.7　两供应商问题的贝叶斯定理计算过程汇总

(1)	(2)	(3)	(4)	(5)
事件	先验概率	条件概率	联合概率	后验概率
A_i	$P(A_i)$	$P(B \mid A_i)$	$P(A_i \cap B)$	$P(A_i \mid B)$
A_1	0.65	0.02	0.013 0	0.013 0/0.030 5 = 0.426 2
A_2	0.35	0.05	0.017 5	0.017 5/0.030 5 = 0.573 8
	1.00		$P(B) = 0.030\ 5$	1.000 0

评　注

1. 贝叶斯定理在决策分析中有广泛的应用。先验概率往往由决策者提出，作为对概率的主观估计，在获得样本信息以后，再计算后验概率以选择最佳决策。
2. 事件和它的补是互斥事件，并且它们的并是整个样本空间。因此，贝叶斯定理总是能够应用于计算事件和补的后验概率。

练　习

方　法

39. 事件 A_1 和 A_2 的先验概率为：$P(A_1) = 0.40$，$P(A_2) = 0.60$，还已知 $P(A_1 \cap A_2) = 0$，假定 $P(B \mid A_1) = 0.20$ 和 $P(B \mid A_2) = 0.05$。

自测题

a. A_1 和 A_2 是互斥事件吗？请解释。

b. 计算 $P(A_1 \cap B)$和 $P(A_2 \cap B)$。

c. 计算 $P(B)$。

d. 应用贝叶斯定理计算 $P(A_1 \mid B)$和 $P(A_2 \mid B)$。

40. 事件 A_1、A_2 和 A_3 的先验概率为：$P(A_1) = 0.20$，$P(A_2) = 0.50$，$P(A_3) = 0.30$。给定 A_1，

A_2 和 A_3 下事件 B 的条件概率为：$P(B|A_1)=0.50$，$P(B|A_2)=0.40$ 和 $P(B|A_3)=0.30$。

a. 计算 $P(B\cap A_1)$，$P(B\cap A_2)$和 $P(B\cap A_3)$。

b. 应用贝叶斯定理公式(4.19)，计算后验概率 $P(A_2|B)$。

c. 使用贝叶斯定理的表格法，计算 $P(A_1|B)$，$P(A_2|B)$和 $P(A_3|B)$。

应　用

41. 某咨询公司向一个大型研究项目投标，公司管理者最初的感觉是公司有一半的机会拿到这个项目。但是接着，招标单位要求该公司提供更多的投标信息。根据以往的经验，有 75% 的中标公司和 40% 的失败公司被要求提供更多的信息。

a. 中标的先验概率(即在被要求提供更多信息以前)是多少？

b. 给定最终中标的条件下，公司被要求提供更多信息的条件概率是多少？

c. 给定公司被要求提供更多信息的条件下，计算中标的后验概率。

42. 自测题 某地方银行对其信用卡政策进行了审查，并考虑收回部分信用卡。在过去大约有 5% 的信用卡持有者不偿还债务从而造成了银行的坏账，因此管理层设定任一信用卡持有者不偿还债务的先验概率为 0.05。银行还进一步发现，最终偿还债务的信用卡客户拖欠一个月或几个月支付的概率是 0.20；当然，那些最终不偿还债务的客户拖欠一到几个月的概率为 1。

a. 在给定客户已经拖欠了一个月的条件下，计算其不偿还债务的后验概率。

b. 如果某客户不偿债的概率超过了 0.20，银行就将收回他的信用卡。问在客户已经拖欠了一个月的情况下，银行是否应该收回他的信用卡？为什么？

43. 小型轿车油耗省，但安全性不如大型轿车。小型轿车只占上路汽车的 18%，但最近一年来涉及到小型轿车的事故已经导致了 11 898 人丧生(*Reader's Digest*，May，2000)。涉及小型轿车的事故导致死亡的概率是 0.128，不涉及小型轿车的事故导致死亡的概率是 0.05。假定你正在研究造成人员死亡的事故，则该种事故涉及到小型轿车的概率是多少？

44. 某市有一支职业篮球队将在主场进行比赛，同天晚上，它的一支职业曲棍球队将到客场比赛。这支职业篮球队在主场获胜的概率是 0.641，而曲棍球队在客场获胜的概率是 0.462。以往当两支球队在同一晚比赛时，第二天早上的体育新闻报道篮球队的机会是 60%，报道曲棍球队的机会是 40%。假定比赛后的第二天早上，报纸体育报道的标题是“我们胜利了”，那么该报道是关于篮球队的概率为多少？

45. *M. D. Computing* (May,1991)描述了在医疗诊断中贝叶斯定理和条件概率的应用。疾病的先验概率是以医生对诸如地理位置、季节影响和传染病的发生等因素所作的评估为基础的。假定某病人被认为具有两种疾病中的一种，它们分别以 D_1 和 D_2 表示，其概率为：$P(D_1)=0.60$，$P(D_2)=0.40$。而该项医学研究的目的是确定伴随疾病的某种症状出现的概率。如果给定疾病 D_1 和 D_2，病人出现症状 S_1、S_2 或 S_3 的概率如下：

		疾病症状		
		S_1	S_2	S_3
疾病	D_1	0.15	0.10	0.15
	D_2	0.80	0.15	0.03

$P(S_3 \mid D_1)$

在发现某种症状出现以后，可以通过计算每种特定疾病的修正概率来帮助进行医疗诊断。在给定下列医疗检查结果的条件下，计算每种疾病的后验概率。

a. 病人出现症状 S_1。

b. 病人出现症状 S_2。

c. 病人出现症状 S_3。

d. 对于具有症状 S_1 的病人，假定我们还发现了症状 S_2，那么 D_1 和 D_2 的修正概率分别是多少？

小 结

在本章我们介绍了基本的概率概念并举例说明了如何利用概率分析来提供对决策有用的信息。我们描述了如何把概率解释为是事件发生可能性的数值度量。另外，我们看到既能够通过把某个事件所包括的每个试验结果（样本点）的概率求和，又能够利用概率的加法法则、条件概率法则和乘法法则所建立的概率关系来计算该事件的概率。在能够取得进一步信息的情况下，我们还演示了怎样利用贝叶斯定理来得到修正的或后验的概率。

术语释义

概率：事件发生可能性的数值度量。

试验：能够产生明确结果的过程。

样本空间：所有试验结果的集合。

样本点：样本空间的一个元素。一个样本点就表示一个试验结果。

树形图：一种图形表示方法，有助于在一个多步骤试验中确认样本点。

概率的基本要求：有两个要求制约了概率指派过程：(1)对每个试验结果 E_i，我们必须有 $0 \leqslant P(E_i) \leqslant 1$；(2)对所有的试验结果，我们必须有 $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$。

古典法：指派概率的一种方法，它的假设前提是所有的试验结果具有相等的发生可能性。

相对频数方法：一种根据试验或历史数据进行概率指派的方法。

主观方法：一种根据主观判断进行概率指派的方法。

事件：由若干样本点组成的集合。

事件 A 的补：由所有不属于事件 A 的样本点构成的事件。

维恩图：一种抽象地表示样本空间和事件运算的图形表示方法。它用矩形代表样本空间，用样本

空间内部的圆圈代表事件。

事件 A 和 B 的并： 包含了属于 A 或属于 B 或同时属于二者的所有样本点的事件，记作 $A \cup B$。

事件 A 和 B 的交： 包含了所有同时属于 A 和 B 的样本点的事件，记作 $A \cap B$。

加法法则： 该法则用于计算两事件的并的概率，公式为 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$。对于互斥事件，因为 $P(A \cap B) = 0$，该公式可简化为 $P(A \cup B) = P(A) + P(B)$。

互斥事件： 没有共同样本点的两个事件，即 $A \cap B$ 是空集 $P(A \cap B) = 0$。

条件概率： 在给定其他事件已经发生的情况下某事件发生的概率。在给定 B 事件下 A 的条件概率为：$P(A \mid B) = P(A \cap B) / P(B)$。

联合概率： 两个事件同时发生的概率，即两事件交的概率。

独立事件： 对于两事件 A 和 B，有 $P(A \mid B) = P(A)$ 或 $P(B \mid A) = P(B)$，即事件之间没有任何相互影响，则称它们为独立事件。

乘法法则： 该法则用于计算两事件的交的概率，公式为 $P(A \cap B) = P(A)\,P(B \mid A)$ 或 $P(A \cap B) = P(B)\,P(A \mid B)$。对于独立事件它可以简化为 $P(A \cap B) = P(A)\,P(B)$。

先验概率： 对事件概率的初始估计。

后验概率： 根据进一步的信息得到的事件修正概率。

贝叶斯定理： 一种用来计算后验概率的方法。

重要公式

组合计数法则

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{4.1}$$

排列计数法则

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

利用补来计算概率

$$P(A) = 1 - P(A^c) \tag{4.5}$$

加法法则

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.6}$$

条件概率

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{4.7}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{4.8}$$

乘法法则

$$P(A \cap B) = P(B)\,P(A \mid B) \tag{4.11}$$

$$P(A \cap B) = P(A)\,P(B \mid A) \tag{4.12}$$

独立事件的乘法法则

$$P(A \cap B) = P(A)\,P(B) \tag{4.13}$$

贝叶斯定理

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n)} \tag{4.19}$$

补充练习

46. 在《商业周刊》Harris 民意调查中，1 035 名成年人被问及他们对于商业的态度（*Business Week*, September 11, 2000），其中一个问题是："在生产好的产品和全球竞争力方面，你如何评价大型的美国公司？"回答是：优秀——18%，相当好——50%，一般——26%，差——5%，不知道/未回答——1%。

a. 某人评价美国公司为相当好或优秀的概率是多少？

b. 对美国公司评价为差的答案有多少个？

c. 不知道或未回答的答案有多少个？

47. 某财务经理刚刚进行了两项新投资——一项在石油工业，另一项投资于市政债券。在一年以后，就能够确定每项投资是成功还是失败。把进行这两项投资视做一个试验。

a. 该试验存在多少个样本点？

b. 作出树形图并列出样本点。

c. 令 O = 投资石油工业成功的事件，M = 投资市政债券成功的事件。列出 O 和 M 内的样本点。

d. 列出事件的并（$O \cup M$）的所有样本点。

e. 列出事件的交（$O \cap M$）的所有样本点。

f. 事件 O 和 M 是互斥的吗？请解释。

48. 一项对美国人的民意调查问道：你对今日美国的经济状况是否感到满意（*The Wall Street Journal*, June 27, 1997）？下表列出了所有参与者的答案和年龄组的分布：

	满意（%）	不满意（%）	其他（%）
所有参与者	**61**	**37**	**2**
18—34	64	35	1
35—49	58	41	1
50—64	57	40	3
65 +	70	26	4

a. 随机选取到的成年人感觉满意的概率是多少？

b. 哪一个年龄组的满意水平高于平均水平？

c. 65 岁以上的成年人未回答他们是否满意的概率是多少？

49. 由 Paul Weiler 领导的一个哈佛研究小组对纽约州的 31 000 例住院病例进行了研究（*Business Week*, March 27, 1995）。他们发现 4% 的住院病人会受到医疗事故的伤害，其中又有 1/7 导致了死亡。1/4 的医疗事故是由于疏忽引起的。每 7.5 起疏忽案例中只有 1 起被承认是工作失误，而只有 1/2 的工作失误能够向病人赔偿。

a. 某位住院病人遭受由医务人员疏忽引起的医疗事故伤害的概率是多少？

b. 某位住院病人死于医疗事故的概率是多少？

c. 在因为疏忽引起医疗事故伤害的情况下，病人的工作失误指控能获得赔偿的概率是多少？

50. 利用电话调查来确定观众对某个新电视节目的反映，得到了下列数据：

评价	频数
差	4
低于平均水平	8
平均水平	11
高于平均水平	14
优秀	13

a. 随机抽取到的观众对新电视节目评价为平均水平或更好的概率是多少？

b. 随机抽取到的观众对新电视节目评价为低于平均水平或更差的概率是多少？

51. 《商业周刊》对其消费者进行了一项关于他们拥有或者租赁的汽车数目的调查。消费者需要回答三个问题：你拥有几辆汽车？你租赁几辆汽车？你有（拥有或者租赁）几辆汽车？表 4.8 包括了从 932 户家庭取得的回答（*Business Week* 1996 Worldwide Subscriber Study）。在对该表解释时，我们注意到第 3 行表明有 401 户家庭拥有 2 辆汽车，47 户家庭租赁 2 辆汽车，447 户家庭有（拥有或/和租赁）2 辆汽车。

a. 家庭租赁一辆汽车的概率是多少？

b. 家庭拥有 2 辆或更少汽车的概率是多少？

c. 家庭有 3 辆或更多汽车的概率是多少？

d. 家庭既未拥有也未租赁汽车的概率是多少？

表 4.8 每个家庭拥有或租赁的汽车数以及有的汽车总数

	家庭户数		
汽车数	拥有	租赁	有
0	65	708	19
1	242	168	168
2	401	47	447
3	149	9	186
4 或以上	75	0	112
总计	932	932	932

52. 毕业生管理委员会对 MBA 注册新生的调查提供了以下 2 018 名学生的数据：

		申请学校是否超过一个	
		是	否
年龄分组	23 及以下	207	201
	24—26	299	379
	27—30	185	268
	31—35	66	193
	36 及以上	51	169

a. 随机选择一名 MBA 学生，进行观察该生的年龄和申请学校数目的试验，作出联合概率表。

b. 随机选择到的申请者年龄在 23 岁或以下的概率是多少？

c. 随机选择到的申请者年龄大于 26 岁的概率是多少？

d. 随机选择到的申请者申请学校超过一所的概率是多少？

53. 仍参见上题的新生调查数据。

a. 给定某学生申请学校超过一所的条件下，该生年龄在 24—26 岁的概率是多少？

b. 给定某学生在 36 岁及以上年龄组的条件下，该生申请学校超过一所的概率是多少？

c. 某学生年龄在 24 — 26 岁之间或者申请学校超过一所的概率是多少？

d. 假定已知某学生只申请了一所学校，该生年龄在 31 岁及以上的概率是多少？

e. 申请学校的数目与年龄独立吗？请解释。

54. 一项 IBD/TIPP 民意调查是为了了解人们对于投资和退休的态度（*Investor's Business Daily*, May 5, 2000）。向男性和女性受调查者询问的一个问题是："在选择退休投资时，你们认为风险水平的重要性如何？"下面的联合概率表是根据收集到的数据作出的，其中的重要是指调查对象的回答是重要或非常重要。

a. 调查对象认为风险水平重要的概率是多少？

b. 男性调查对象认为风险水平重要的概率是多少？

c. 女性调查对象认为风险水平重要的概率是多少？

d. 风险水平是否独立于被调查者的性别？为什么？

e. 男性和女性对待风险的态度是否相同？

	男性	女性	总计
重要	0.22	0.27	0.49
不重要	0.28	0.23	0.51
总计	0.50	0.50	1.00

55. 某大型消费品公司正在运作一个关于一种自产肥皂产品的电视广告项目，为此进行了一项调查。在该调查的基础上，对下面的事件指派了概率：

B = 消费者购买了产品；

S = 消费者能够回忆起广告；

$B \cap S$ = 消费者购买了产品且能够回忆起广告。

对它们指派的概率为：$P(B) = 0.20$，$P(S) = 0.40$，$P(B \cap S) = 0.12$。

a. 在给定消费者能够回忆起广告的条件下，他购买产品的概率是多少？消费者看过广告是否能够增加其购买产品的概率？假如你是决策者，你是否建议继续播放广告(假定广告费用是合理的)？

b. 假设未购买该公司肥皂产品的消费者从它的竞争对手那里购买产品，你对公司市场份额的估计是多少？你认为继续播放广告能够增加公司的市场份额吗？为什么？

c. 公司还检查了另一个广告，并对它指派概率值为：$P(S) = 0.30$ 且 $P(B \cap S) = 0.10$。该广告的 $P(B|S)$是多少？看起来哪一个广告对消费者的影响更大？

56. Cooper Realty 公司是一家位于纽约阿尔巴尼的小型房地产公司，主要进行居民住宅的销售。最近，这家公司想要确定在某特定时间段内一座居民住宅被售出的可能性。通过对去年公司销售的 800 座住宅进行分析，得到以下数据：

		销售时间			
		30 天以下	31—90	90 天以上	总计
初始价格	50 000 美元以下	50	40	10	100
	50 000—99 999 美元	20	150	80	250
	100 000—150 000 美元	20	280	100	400
	150 000 美元以上	10	30	10	50
	总计	100	500	200	800

a. 如果定义 A 是销售时间超过 90 天的事件，估计 A 的概率。

b. 如果定义 B 是初始价格低于 50 000 美元的事件，估计 B 的概率。

c. $A \cap B$ 的概率是多少？

d. 假定已经确定要出售一座初始价格低于 50 000 美元的住宅，则该公司在超过 90 天的时间才能将其售出的概率是多少？

e. 事件 A 和 B 独立吗？

57. 某公司对它设在得克萨斯州 Brownsville 的一个工厂的误工事件进行了研究。历史记录表明在去年有 6% 的员工发生过误工事件，管理人员相信，实行某项特殊的保障措施将会把今年的误工事件减少到 5%。此外，据估计在去年发生误工事件的员工中有 15% 在今年还会发生类似事件。
    a. 有多大百分比的员工在去年和今年都发生了误工事件？
    b. 有多大百分比的员工在去年和今年至少发生过一次误工事件？

58. 达拉斯的 IRS 审计人员正在关注如何确定可能的退税诈骗行为。根据过去的经验他们发现，如果一项退税项目中包含的特别税折扣超过了 IRS 标准，则存在退税诈骗行为的概率为 0.20；如果未超过 IRS 标准，则存在退税诈骗行为的概率为 0.02。假定共有 8% 的案例的特别税折扣超过了 IRS 标准，估计退税诈骗行为的百分比是多少？

59. 某石油公司购买了一块阿拉斯加的土地开采权，经初步的地质研究指派先验概率如下：

$$P(\text{含高质量石油}) = 0.50$$
$$P(\text{含中等质量石油}) = 0.20$$
$$P(\text{不含石油}) = 0.30$$

    a. 在这块土地上发现石油的概率是多少？
    b. 在第一口井钻至 200 英尺深时，进行了土壤测试。在测试中发现特殊土壤类型的概率如下：

$$P(\text{该土壤类型} \mid \text{含高质量石油}) = 0.20$$
$$P(\text{该土壤类型} \mid \text{含中等质量石油}) = 0.80$$
$$P(\text{该土壤类型} \mid \text{不含石油}) = 0.20$$

    该公司将如何解释土壤测试的结果？修正概率分别是多少？发现石油的新概率是多少？

60. 贝叶斯方法可以用来计算某勘探地点出产石油的修正概率（*Oil & Gas Journal*, January 11, 1988）。在一个例子中，地质评估显示某地出产石油的概率为 25%。并且在给定该地点出产石油的条件下，某个钻井能够勘探到石油的机会是 80%。
    a. 假设在该地点钻探了一口井后未发现石油，则该地点出产石油的概率是多少？
    b. 如果两口油井都没有发现石油，则该地点出产石油的概率是多少？
    c. 如果石油公司坚持只要出产石油的概率大于 1% 就会继续钻探，问至少要钻探多少口没有石油的井后，才会放弃这一地点？

案例研究　HAMILTON 县的法官

Hamilton 县的法官每年要处理几千个案件。对于绝大多数的案件，其裁决不会改动。但是也有一些案件提出上诉，并且确实有一些上诉案件的裁决被推翻。《辛辛那提调查》（*The Cincinnati Enquirer*）的 Kristen DelGuzzi 对 Hamilton 县的法官在 1994 年—1996 年中处理的案件进行了研究（*The Cincinnati Enquirer*, January 11, 1998）。表 4.9 显示了由普通诉讼庭、民事庭和地方庭的 38 位

法官处理过的 182 908 个案件的结果。有两位法官（Dinkelacker 和 Hogan）三年来从未在同一法庭共过事。

该报纸的研究目的是评估法官们的表现。上诉往往是法官错误判决的结果，而该报想要知道的是哪些法官工作得更好，哪些法官常犯错误。在该案例中，要求你帮助分析数据，并利用你学过的概率和条件概率知识帮助对法官进行评定。你还需要分析在不同法庭处理过的案件被上诉和推翻原判的可能性。



CD 光盘数据
Judge

表 4.9　Hamilton 县法庭在 1994—1996 年期间处理的全部案件和被上诉、推翻的案件

	普通诉讼庭		
法官	处理案件的总数	被上诉的案件数	被推翻的案件数
Fred Cartolano	3 037	137	12
Thomas Crush	3 372	119	10
Patrick Dinkelacker	1 258	44	8
Timothy Hogan	1 954	60	7
Robert Kraft	3 138	127	7
William Mathews	2 264	91	18
William Morrissey	3 032	121	22
Norbert Nadel	2 959	131	20
Arthur Ney, Jr.	3 219	125	14
Richard Niehaus	3 353	137	16
Thomas Nurre	3 000	121	6
John O'Connor	2 969	129	12
Robert Ruehlman	3 205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3 141	127	13
Ralph Winkler	3 089	88	6
总计	43 945	1 762	199
	民事庭		
法官	处理案件的总数	被上诉的案件数	被推翻的案件数
Penelope Cunningham	2 729	7	1
Patrick Dinkelacker	6 001	19	4
Deborah Gaines	8 799	48	9
Ronald Panioto	12 970	32	3
总计	30 499	106	17

（续表）

	地方庭		
法官	处理案件的总数	被上诉的案件数	被推翻的案件数
Mike Allen	6 149	43	4
Nadine Allen	7 812	34	6
Timothy Black	7 954	41	6
David Davis	7 736	43	5
Leslie Isaiah Gaines	5 282	35	13
Karla Grady	5 253	6	0
Deidra Hair	2 532	5	0
Dennis Helmick	7 900	29	5
Timothy Hogan	2 308	13	2
James Patrick Kenney	2 798	6	1
Joseph Luebbers	4 698	25	8
William Mallory	8 277	38	9
Melba Marsh	8 219	34	7
Beth Mattingly	2 971	13	1
Albert Mestemaker	4 975	28	9
Mark Painter	2 239	7	3
Jack Rosen	7 790	41	13
Mark Schweikert	5 403	33	6
David Stockdale	5 371	22	4
John A. West	2 797	4	2
总计	108 464	500	104

管理报告

准备一份评定法官的报告，包括对在三个法庭处理过的案件被上诉和推翻原判可能性的分析。你的报告最少应包括以下内容：

1. 三个法庭处理过的案件被上诉和推翻原判的概率。
2. 每个法官处理过的案件被上诉的概率。
3. 每个法官处理过的案件被推翻的概率。
4. 排列每个法庭法官的名次，写出你使用的标准并说明你选择该标准的理由。

5 离散概率分布

统计实例:花旗银行

5.1 随机变量

离散随机变量

连续随机变量

5.2 离散概率分布

5.3 数学期望和方差

数学期望

方差

5.4 二项概率分布

二项试验

马丁服装商店问题

使用二项概率表

二项概率分布的数学期望和方差

5.5 泊松概率分布

一个涉及时间间隔的例子

一个涉及长度或距离间隔的例子

5.6 超几何概率分布

统计实例

花旗银行*

长岛，纽约州

花旗银行作为花旗集团的一部分，提供范围广泛的金融服务，包括支票和储蓄账户、贷款和抵押、保险以及投资服务，所有这一切都处于一个被称为“Citibanking”的一体化战略框架之内。Citibanking在全世界树立了一个始终如一的商标标识、一贯的产品供应以及高质量的客户服务。无论何时、何地，花旗银行都使人们能够以其选择的任何方式管理自己的金钱。不管是需要为了明天而储蓄，还是需要为了今天而借贷，都能在花旗银行得以实现。



© PhotoDisc, Inc.

位于花旗卡业务中心（CBCs）的花旗银行业务自动处理机（ATMs），使顾客每周7天、每天24小时地在一个地方仅仅依靠自己手指的操作，就能够完成他们所有的银行业务，从储蓄到管理投资在内的超过150种不同的银行业务都能够方便地进行。花旗银行业务自动处理机远多于现在顾客所使用的自动提款机，而这些自动提款机经手的业务已占了顾客所有交易的80%。

每一个花旗银行的CBC对于随机地到达其任何一台ATM接受服务的顾客，运行了一个排队系统。如果所有的ATM都有人在使用，到达的顾客将排队等候。CBC容量的时间段研究被用来分析顾客的等待时间和确定是否需要增加ATM。

花旗银行收集的数据显示，随机到达的顾客遵循泊松概率分布。利用该分布，花旗银行能够计算在任何时间段内到达CBC的不同顾客数的概率，并据此决定所需要的ATM台数。例如，令 $x = 1$ 分钟内到达的顾客数，假定某个CBC每分钟的平均到达率为2名顾客，下表显示了在1分钟内到达顾客数的概率分布：

x	概率
0	0.135 3
1	0.270 7
2	0.270 7
3	0.180 4
4	0.090 2
5或以上	0.052 7

诸如花旗银行使用的这些离散概率分布，是本章将要讨论的主题。除了泊松概率分布以外，你还将学习到二项和超几何概率分布，以及如何利用它们来提供有用的概率信息。

* 作者感谢花旗银行的Stacey Karter先生提供了该统计实例。

在本章我们将继续研究概率问题，对随机变量的概念和概率分布进行介绍。本章的重点是离散概率分布，共包括三种不同的离散概率分布形式：二项分布、泊松分布和超几何分布。

5.1 随机变量

在第 4 章，我们定义了试验和试验结果的概念。随机变量则提供了一种用数值来描述试验结果的方法。

随机变量必须用数值表示。

随机变量

随机变量是对试验结果的数值性描述。

实际上，随机变量把数值与每个可能的试验结果联系起来，随机变量的取值依赖于试验结果。根据用来表示随机变量的数值特征，它可被分为**离散随机变量**和**连续随机变量**两类。

离散随机变量

使用有限个数值或者是像 0，1，2，……这样存在间隔的无穷数列表示的随机变量被称为**离散随机变量**(discrete random variable)。例如，考虑一名会计参加注册会计师（CPA）考试的试验。该考试有四门内容，我们能够定义随机变量为：x = 通过 CPA 考试的门数。因为它是用有限个数值：0，1，2，3，4 来表示的，所以属于离散随机变量。

再举一个离散随机变量的例子，考虑到达收费站的汽车数试验。有关的随机变量是 x = 一天内到达收费站的汽车数，x 的可能值是整数 0,1,2,……的无穷数列。因此，x 属于离散随机变量。

尽管许多试验结果可以自然而然地用数值表示，但有一些则不行。例如，某项调查要求调查对象回忆在最近的电视广告中出现的内容。该试验有两个可能的结果：调查对象回忆不起来和调查对象能够回忆起来。我们可以人为地规定离散变量 x 如下：如果调查对象不能回忆则令 $x=0$，如果调查对象能够回忆则令 $x=1$，这样我们就仍然可以使用数值来描述该试验的结果。该随机变量的数值是任意规定的(我们也可以使用 5 和 10)，但按照随机变量的定义，它们也是可接受的——因为 x 给出了对试验结果的数值描述，所以它就是随机变量。

表 5.1 列出了一些离散随机变量的例子。注意在每个例子中，离散随机变量都使用了有限个数值或像 0，1，2，……这样的无穷数列来表示。我们将在本章详细地讨论诸如此类的随机变量。

表 5.1　离散随机变量的例子

试验	随机变量（x）	随机变量的可能值
接触 5 位顾客	下订单的顾客数	0，1，2，3，4，5
检查一批 50 只收音机	次品收音机的数目	0，1，2，…，49，50
某饭店一天的经营情况	顾客数	0，1，2，3，…
销售一部汽车	顾客的性别	男性为 0，女性为 1

连续随机变量

可以用一个区间或区间集合内的任何数值表示的随机变量被称为**连续随机变量**(continuous random variable)。我们能够使用连续随机变量来描述建立在时间、重量、距离和温度等的度量值之上的试验结果。例如，考虑对打进某大型保险公司理赔办公室的电话进行监控这个试验。假定有关的随机变量为：x = 连续两个电话的间隔分钟数，该随机变量可以取区间 $0 \leqslant x \leqslant 90$ 内的任何值。实际上，x 可能取的值是无穷个，包括像 1.26 分钟、2.751 分钟和 4.333 3 分钟这样的数值。再举一个例子，佐治亚州亚特兰大以北有一段 90 英里长的州际公路 I-75。对于一个位于亚特兰大的事故救护机构而言，我们可以定义随机变量为：x = 在 I-75 的该路段发生的下一起交通事故的位置。在本例中，x 是一个连续随机变量，可以取区间内的任何值。表 5.2 列出了一些连续随机变量的例子。注意在每个例子中所描述的随机变量都可以取区间内的任何值。连续随机变量和它们的概率分布是第 6 章的主要内容。

表 5.2 连续随机变量的例子

试验	随机变量 (x)	随机变量的可能值
经营银行	两顾客到达时间间隔的分钟数	$x \geqslant 0$
填充饮料罐 (最大 = 12.1 盎司)	饮料罐的盎司数	$0 \leqslant x \leqslant 12.1$
建设新图书馆项目	项目在 6 个月以后完工的百分比	$0 \leqslant x \leqslant 100$
测试一个新化工工艺	所需要的反应发生温度 (最低 150°F，最高 212°F)	$150 \leqslant x \leqslant 212$

评 注

确定随机变量是离散还是连续的一个方法是：把随机变量的值当做线段上的点，选择两个点代表随机变量的两个值。如果两点间的整条线段也代表随机变量的值，那么该随机变量就是连续的。

练 习

方 法

1. 考虑抛掷硬币两次的试验。



a. 列出所有试验结果。

b. 定义在两次抛掷中表示正面出现次数的随机变量。
c. 对每个试验结果，列出随机变量的取值。
d. 该随机变量是连续的还是离散的？

2. 考虑某工人加工产品并记录花费时间的试验。
a. 定义表示加工产品所需时间的随机变量。
b. 随机变量可以取的值是多少？
c. 该随机变量是离散的还是连续的？

应　用

3. 

三个学生为了暑假在 Brookwood 学院工作而被安排进行面试。面试结果是录用或不录用，试验结果就以三个人的面试结果来定义。
a. 列出试验结果。
b. 定义表示录用人数的随机变量。它是离散的还是连续的？
c. 列出每个试验结果的随机变量值。

4. 假定我们知道 12 家佛罗里达信贷机构的住房抵押贷款利率，设有关的随机变量是在这些机构中 30 年固定利率为 8.5% 或以下的机构数目。该随机变量可以取的值是多少？

5. 为进行某种血液分析，实验员必须执行两道工序。第一道工序需要 1 或 2 个单独的步骤，第二道工序需要 1，2 或 3 个单独的步骤。
a. 列出与血液分析相关的试验结果。
b. 如果有关的随机变量是完成分析（全部工序）所需的总步骤数，列出每个试验结果对应的随机变量值。

6. 下表是一系列试验与相关的随机变量，确认每个例子中随机变量的可能值并说明该随机变量是离散的还是连续的。

试验	随机变量（x）
a. 答一份包括 20 道问题的试卷	答对的问题数
b. 观察在 1 小时内到达收费站的汽车	到达收费站的汽车数
c. 审计 50 例税款返还	错误的返还案例数
d. 观察一个雇员的工作	8 小时工作日内的非生产时间
e. 称一批货物的重量	重量的磅数

5.2　离散概率分布

随机变量的**概率分布**(probability distribution)描述了随机变量取不同值的概率。对于离散随机

变量 x，其概率分布由**概率函数**(probability function)来定义，用 $f(x)$表示。概率函数提供了随机变量取每个值时的概率。

作为离散随机变量及其概率分布的例子，考虑纽约萨拉托加 DiCarlo 汽车公司的汽车销售数量。在过去 300 个营业日中，有 54 天销量为 0，117 天销量为 1 辆，72 天为 2 辆，42 天为 3 辆，12 天为 4 辆，3 天为 5 辆。假定我们选择该公司一天的营业情况作为试验，定义有关的随机变量 x = 一天内售出的汽车数。根据历史数据，我们知道 x 是离散随机变量，可取的值为：0，1，2，3，4，5。在概率函数的符号中，$f(0)$表示销量为 0 的概率，$f(1)$表示销量为 1 的概率，依此类推。因为历史数据显示 300 天中有 54 天销量为 0，我们把数值 54/300 = 0.18 分配给 $f(0)$，表示一天内销售 0 辆汽车的概率是 0.18。类似地，因为 300 天中有 117 天销量为 1，我们把数值 117/300 = 0.39 分配给 $f(1)$，表示一天内销售 1 辆汽车的概率是 0.39。继续使用这种方法求得随机变量的其他值，可得出 $f(2)$、$f(3)$、(4)和 $f(5)$的值，表明 DiCarlo 汽车公司一天内汽车的销量，列在表 5.3 中。

表 5.3　Dicarlo 汽车公司每天内汽车销量的概率分布

x	$f(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01
总计	1.00

定义随机变量及其概率分布的主要好处在于，一旦知道了概率分布，对于各种事件有兴趣的决策者要确定事件的发生概率就相对简单了。例如，利用表 5.3 的 DiCarlo 汽车公司的概率分布，我们看到在 1 天内最可能的汽车销售数是 1 辆，其概率为 $f(1) = 0.39$。另外，1 天内销售 3 辆或以上汽车的概率是 $f(3) + f(4) + f(5) = 0.14 + 0.04 + 0.01 = 0.19$。这些概率加上决策者关心的其他因素，提供了有助于决策者了解该公司汽车销售情况的信息。

在得出任一离散随机变量的概率函数的过程中，必须满足以下两个条件：

离散概率函数的要求条件

$$f(x) \geqslant 0 \tag{5.1}$$

$$\sum f(x) = 1 \tag{5.2}$$

这些条件与第 4 章中对试验结果的两条基本要求相似。

表 5.3 显示出随机变量 x 的概率满足公式(5.1)，即对 x 的所有值，都大于或等于 0。另外，

全部概率之和为 1，故满足公式(5.2)。因此，DiCarlo 汽车公司的概率函数是有效的离散概率函数。

我们还可以用图形来表示概率分布。在图 5.1 中，DiCarlo 汽车公司随机变量 x 的值在横轴上表示，与 x 值对应的概率值则在纵轴上表示。

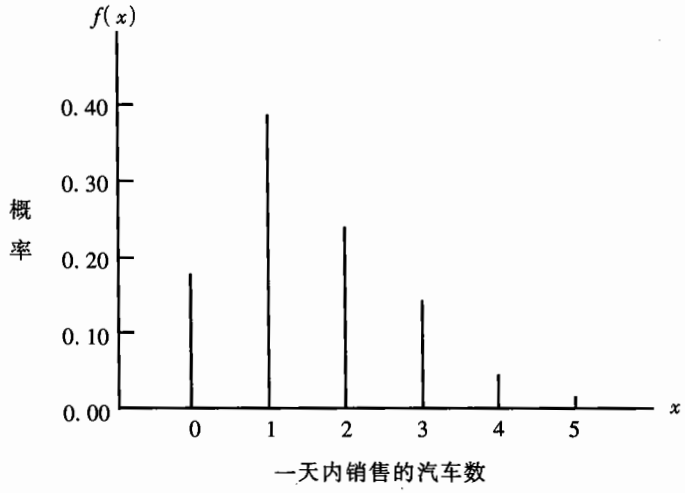


图 5.1　DiCarlo 汽车公司一天汽车销量的概率分布的图形表示

除了表格和图形以外，利用公式也能够对 x 的每个值给出概率函数 $f(x)$，所以往往利用公式来描述概率分布。以公式表示的离散概率函数的最简单例子是**离散均匀概率分布**(discrete uniform probability distribution)。它的概率函数如下所示：

离散均匀概率函数

$$f(x) = 1/n \tag{5.3}$$

式中　n——随机变量所有可能值的数目

举例说明该函数，考虑投掷色子的试验，定义随机变量 x 为朝上一面的点数。随机变量有 $n=6$ 个可能值，分别为 $x=1, 2, 3, 4, 5, 6$。因此该随机变量的概率函数是

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

随机变量的可能取值与对应的概率值列表如下：

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

注意在本例中随机变量取每个值的可能性相等。

再举一个例子，设某随机变量 x 具有如下的离散概率分布：

x	$f(x)$
1	1/10
2	2/10
3	3/10
4	4/10

该概率分布可由公式表示为

$$f(x)=\frac{x}{10},\quad x=1,\ 2,\ 3,\ 4$$

这样，给出随机变量的值就能够得到对应的概率值 $f(x)$。例如，对于上面的概率函数，我们看到：$f(2)=2/10$ 得出了随机变量取 2 时的概率。

由公式表示的离散概率分布通常应用更为广泛。三种最重要的离散概率分布公式是二项分布、泊松分布和超几何分布，我们将在本章的后面讨论它们。

练　习

方　法

7. 随机变量 x 的概率分布列示如下：



x	$f(x)$
20	0.20
25	0.15
30	0.25
35	0.40
总计	1.00

a. 它是适当的概率分布吗？检查它是否满足公式（5.1）和（5.2）。

b. $x=30$ 的概率是多少？

c. $x\leqslant 25$ 的概率是多少？

d. $x>30$ 的概率是多少？

应　用

8. 以下数据是 Tampa 总医院在 20 天的期间内每天使用手术室的数目：有 3 天使用 1 间，有 5 天使用 2 间，有 8 天使用 3 间，有 4 天医院的 4 间手术室全部都在使用。



a. 利用相对频数方法对任意一天使用手术室的数目建立概率分布。

b. 画出概率分布图。

c. 证实你的概率分布是否满足离散概率分布的条件。

9. 以下数据显示了联邦政府中 5 个行政级别各自的雇员数（*Pay Structure of the Federal Civil Service*, U. S. Office of Personnel Management, 1996）：

行政级别	雇员数
1	15
2	32
3	84
4	300
5	31
总计	462

假设我们为了调查雇员的工作条件而在各行政级别选择雇员作为样本，以随机变量 x 表示随机选择到的雇员的行政级别。

a. 使用数据建立 x 的概率分布。标明随机变量的取值及对应的概率函数的值。

b. 画出概率分布图。

c. 证实概率分布是否满足公式(5.1)和公式(5.2)。

10. 表 5.4 显示了一个信息系统(IS)的高级执行经理和中级经理工作满意度得分的百分比频数分布（*Computerworld*, May 26, 1997）。分数从最低的 1 分(非常不满意)到最高的 5 分(非常满意)。

a. 建立高级执行经理工作满意度分数的概率分布。

b. 建立中级经理工作满意度分数的概率分布。

c. 高级执行经理报告的工作满意度分数为 4 或 5 的概率是多少？

d. 中级经理非常满意的概率是多少？

e. 比较两种职位的工作满意度。

表 5.4　信息系统的高级执行经理和中级经理工作满意度得分的百分比频数分布

工作满意度分数	高级执行经理的百分比(%)	中级经理的百分比(%)
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28
总计	100	100

11. 一位技师负责为菲尼克斯地区的公司邮购机器提供售后服务。根据故障的类型，服务电话会

用去1，2，3或4个小时。设不同类型的故障发生频率相同。

a. 建立一个服务电话占用时间的概率分布。

b. 画出概率分布图。

c. 证实你的概率分布是否满足离散概率分布的条件。

d. 服务电话占用3个小时的概率是多少？

e. 有个服务电话已经打进，但故障类型未知。现在是下午3:00，服务技师通常在下午5:00下班，那么他今天超时工作修理机器的概率是多少？

12. Lakeville社区学院的招生主任客观地估计出入学人数 x 的概率分布如下：

x	$f(x)$
1 000	0.15
1 100	0.20
1 200	0.30
1 300	0.25
1 400	0.10

a. 这是有效的概率分布吗？

b. 入学人数是1 200或以下的概率是多少？

13. 某心理医生已经确定要获得一名新病人的信任需要1，2或3个小时。以随机变量 x 表示获得病人信任所需的小时数，他提出了以下的概率函数：

$$f(x)=\frac{x}{6},\quad x=1，2或3$$

a. 这是一个概率函数吗？请解释。

b. 取得病人信任恰好花费2小时的概率是多少？

c. 取得病人信任至少花费2小时的概率是多少？

14. 下表为MRA公司营业第一年计划利润（x=以千美元计的利润）的概率分布的一部分（负值代表亏损）。

x	$f(x)$
-100	0.10
0	0.20
50	0.30
100	0.25
150	0.10
200	

a. $f(200)$的值是多少？你对该值如何解释？

b. MRA公司盈利的概率是多少？

c. MRA公司至少盈利10万美元的概率是多少？

5.3　数学期望和方差

数学期望

随机变量的**数学期望**(expected value)，或者说是均值，是对随机变量中心位置的度量。离散随机变量 x 的数学期望的数学表达式如下：

离散随机变量的数学期望

$$E(x) = \mu = \sum xf(x) \tag{5.4}$$

数学期望是随机变量可取值的加权平均值，其权重就是概率。

符号 $E(x)$和 μ 都用来表示随机变量的数学期望。

公式（5.4）表明：为了计算离散随机变量的数学期望，我们必须把随机变量的每个值乘以相对应的概率 $f(x)$，并且将所得乘积相加。仍使用 5.2 节的 DiCarlo 汽车公司为例，在表 5.5 中我们说明了怎样计算一天内汽车销量的数学期望。这一列之和表示销量的数学期望为每天 1.50 辆汽车。因此我们了解到：尽管一天的销量可能是 0，1，2，3，4 或 5 辆，但 DiCarlo 公司仍可预期平均每天售出 1.50 辆汽车。设每月营业 30 天，我们可用数学期望 1.50 来预测每月的平均销量为 $30 \times 1.50 = 45$ 辆汽车。

表 5.5　DiCarlo 汽车公司一天内汽车销量的数学期望计算过程

x	$f(x)$	$xf(x)$
0	0.18	$0 \times 0.18 = 0.00$
1	0.39	$1 \times 0.39 = 0.39$
2	0.24	$2 \times 0.24 = 0.48$
3	0.14	$3 \times 0.14 = 0.42$
4	0.04	$4 \times 0.04 = 0.16$
5	0.01	$5 \times 0.01 = 0.05$
		1.50
	$E(x) = \mu = \sum xf(x)$	

方 差

尽管数学期望提供了随机变量的平均值，但我们还经常需要度量它的变异程度，或分散程度。正如我们在第 3 章中使用方差来概括数据的变异程度一样，现在我们使用**方差**(variance)来概括随机变量的变异程度。离散随机变量方差的数学表达式如下：

离散随机变量的方差

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x) \tag{5.5}$$

正如公式（5.5）所示，方差公式的基本部分是离差 $x - \mu$，它度量的是随机变量的某个特定值与数学期望或均值 μ 的距离。在计算随机变量的方差时，先将离差平方，再用对应的概率函数值加权，随机变量所有值的加权平方离差之和就被称为方差。符号 $\text{Var}(x)$ 和 σ^2 都表示随机变量的方差。

方差是随机变量与其均值的离差平方的加权平均值，概率就是权重。

表 5.6 汇总了计算 DiCarlo 汽车公司一天内汽车销量的概率分布方差的计算过程，我们看到方差的计算结果是 1.25。我们还定义**标准差**(standard deviation) σ 是方差的正平方根。因此，一天汽车销量的标准差为

$$\sigma = \sqrt{1.25} = 1.118$$

标准差的单位与随机变量的单位相同（$\sigma = 1.118$ 辆汽车），因此被更经常地用作对随机变量变异程度的描述。而方差 σ^2 的单位也是平方项，因此较难以解释。

表 5.6　DiCarlo 汽车公司一天内汽车销量的概率分布方差计算过程

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	$0 - 1.50 = -1.50$	2.25	0.18	$2.25(0.18) = 0.405\,0$
1	$1 - 1.50 = -0.50$	0.25	0.39	$0.25(0.39) = 0.097\,5$
2	$2 - 1.50 = 0.50$	0.25	0.24	$0.25(0.24) = 0.060\,0$
3	$3 - 1.50 = 1.50$	2.25	0.14	$2.25(0.14) = 0.315\,0$
4	$4 - 1.50 = 2.50$	6.25	0.04	$6.25(0.04) = 0.250\,0$
5	$5 - 1.50 = 3.50$	12.25	0.01	$12.25(0.01) = 0.122\,5$
				$1.250\,0$
				$\sigma^2 = \sum (x - \mu)^2 f(x)$

练　习

方　法

15. 下表是随机变量 x 的概率分布：

x		$f(x)$
3		0.25
6		0.50
9		0.25
	总计	1.00

a. 计算 x 的数学期望 $E(x)$。

b. 计算 x 的方差 σ^2。

c. 计算 x 的标准差 σ。

16. 下表是随机变量 y 的概率分布：

自测题

y		$f(y)$
2		0.20
4		0.30
7		0.40
8		0.10
	总计	1.00

a. 计算 $E(y)$。

b. 计算 Var(y)和 σ。

应　用

17. 一个志愿救护服务机构每天都要接 0 到 5 个服务电话，服务电话数的概率分布如下：

服务电话数	概率
0	0.10
1	0.15
2	0.30
3	0.20
4	0.15
5	0.10

a. 服务电话数的数学期望是多少？

b. 服务电话数的方差是多少？

18. 《1997 年美国统计摘要》显示，每户家庭平均拥有 2.3 台电视。假设下表是新奥尔良地区每户家庭拥有电视台数的概率分布：

自测题

x	$f(x)$
0	0.01
1	0.23
2	0.41
3	0.20
4	0.10
5	0.05

a. 计算该地区每户家庭电视拥有量的数学期望，并与《1997 年美国统计摘要》报道的平均拥有量比较。

b. 该地区每户电视拥有量的方差和标准差分别是多少？

19. 四支篮球队的实际投篮记录显示，投中 2 分球的概率是 0.50，而投中 3 分球的概率是 0.39。

a. 这些球队 2 分球得分的数学期望是多少？

b. 这些球队 3 分球得分的数学期望是多少？

c. 既然投中 2 分球的概率大于投中 3 分球的概率，教练为何还在有机会时要求队员投 3 分球？使用数学期望进行解释。

20. 牛顿汽车保险公司关于撞车保险的伤害赔付金额的概率分布如下表所示：

赔付金额（美元）	概率
0	0.90
400	0.04
1 000	0.03
2 000	0.01
赔付金额（美元）	**概率**
4 000	0.01
6 000	0.01

a. 利用撞车赔付金额的数学期望来确定使公司保本的撞车保险收费额。

b. 保险公司对撞车保险的收费为每年 260 美元。保险客户的撞车保险单的数学期望是多少？（提示：它等于从保险公司得到的期望赔付减去保险费用。）为什么客户会购买具有这一数学期望的撞车事故保险单？

21. 以下是一个信息系统高级执行经理和中级经理样本的工作满意度分数的概率分布（*Computerworld*, May 26, 1997），分数范围从最低的 1 分（非常不满意）到最高的 5 分（非常满意）。

	概率	
工作满意度	高级执行经理	中级经理
1	0.05	0.04
2	0.09	0.10
3	0.03	0.12
4	0.42	0.46
5	0.41	0.28
总计	1.00	1.00

a. 高级执行经理工作满意度分数的数学期望是多少？

b. 中级经理工作满意度分数的数学期望是多少？

c. 计算高级执行经理和中级经理工作满意度分数的方差。

d. 计算两者的工作满意度概率分布的标准差。

e. 比较高级执行经理和中级经理整体的工作满意度。

22. 各个月份对 Carolina 工业公司的产品需求有很大不同，下表的概率分布建立在过去两年的数据基础之上，显示了对公司产品的月需求量。

需求量	概率
300	0.20
400	0.30
500	0.35
600	0.15

a. 如果公司的每月订单数是根据月需求量来制定的，那么该公司的每月订单数是多少？

b. 假如每单位的需求量产生 70 美元收入，而每单位的订购成本为 50 美元。如果订购量就是你在(a)中的答案，并且实际的需求量为 300 单位时，该公司这个月的盈利或亏损是多少？

23. 根据调查，*The Wall Street Journal Interactive Edition* 的订户中有 95% 的家庭拥有个人电脑。下表列出了这些家庭的笔记本电脑和台式电脑数目的概率分布(*The Wall Street Journal Interactive Edition Subscriber Study*，1999)。

	概率	
电脑数目	笔记本	台式机
0	0.47	0.06
1	0.45	0.56
2	0.06	0.28
3	0.02	0.10

a. 每户家庭拥有每种类型电脑数目的数学期望分别是多少？

b. 每户家庭拥有每种类型电脑数目的方差分别是多少？

c. 对订户拥有的笔记本和台式机数目进行比较。

24. J. R. Ryland 计算机公司正在考虑一项厂房扩建计划，以生产一种新的计算机产品。公司总裁必须决定扩建项目是中型还是大型的，但又无法确定对新产品的需求量。需求量的预测可能为低、中或高，对应的概率估计为 0.20、0.50 和 0.30。令 x 代表以千美元计的年度利润，公司规划者已经作出了中型和大型扩建项目的利润预测。

		中型扩建项目的利润		大型扩建项目的利润	
		x	$f(x)$	y	$f(y)$
需求	低	50	0.20	0	0.20
	中	150	0.50	100	0.50
	高	200	0.30	300	0.30

a. 计算两种扩建方案利润的数学期望，哪个方案对实现期望利润最大化的目标更优？

b. 计算两种扩建方案利润的方差，哪个方案对实现风险或不确定性最小化的目标更优？

5.4 二项概率分布

二项概率分布是一种应用广泛的离散概率分布，它与一个被称为二项试验的多步骤试验有关。

二项试验

一个**二项试验**(binomial experiment)具有以下四个性质：

二项试验的性质

1. 二项试验是把相同的单次试验进行了 n 次所形成的一个序列。
2. 每一次单次试验都有两种可能的结果。我们把其中一个称为**成功**，另一个称为**失败**。
3. 单次试验的成功概率用 p 表示，它在各次试验中都相同。因此，单次试验的失败概率用 $1-p$ 表示，它在各次试验中也都相同。
4. 每一次的单次试验都独立进行。

如果出现性质 2、3 和 4 的话，我们就说试验是由柏努利过程产生的。另外，如果性质 1 也出现的话，我们就称其为**二项试验**。图 5.2 描述了一个包括 8 次试验的二项试验及一个可能的结果序列。在这个例子中，有 5 次成功和 3 次失败。

贾克伯·柏努利(Jakob Bernoulli)(1654—1705)，是瑞士数学家家族柏努利家族的第一代。他曾经发表过关于概率的论文，包括排列和组合理论，以及二项式定理。

性质 1：试验由 $n=8$ 次相同的试验构成。

性质 2：每次试验的结果是成功（S）或失败(F)。

试验 ⟶	1	2	3	4	5	6	7	8
结果 ⟶	S	F	F	S	S	F	S	S

图 5.2　一个包括 8 次试验的二项试验图

在一个二项试验中，我们关心的是在 ***n* 次试验中出现成功的次数**。如果以 x 表示 n 次试验中成功的次数，我们看到 x 可取的值为 0，1，2，3，…，n。因为值的个数是有限的，故 x 是离散随机变量。与该随机变量有关的概率分布被称为**二项概率分布**(binomial probability distribution)。例如，对于抛掷 5 次硬币的试验，每次都观察在硬币着地时是正面朝上还是反面朝上。假定我们关心的是 5 次抛掷中正面朝上的次数，该试验具备二项试验的性质吗？有关的随机变量是什么？我们注意到：

1. 该试验由 5 次相同的试验构成，每次试验就是抛一枚硬币。
2. 每次试验都有两个可能的结果：正面或反面。我们可以指定正面为成功，反面为失败。
3. 每次试验正面出现的概率都是相同的，为 $p=0.5$；每次试验反面出现的概率也是相同的，为 $1-p=0.5$。
4. 因为任意一次试验的结果都不影响其他试验，所以各次试验或抛掷都是独立的。

因此，该试验满足二项试验所有的性质。有关的随机变量是：x = 在 5 次试验中出现正面的次数，本例中，x 可取的值为 0，1，2，3，4 或 5。

再举一个例子，一个保险推销员随机地选择 10 户家庭进行访问。定义每次访问的结果为：如果该家庭购买了保险单则为成功，如果该家庭未购买则为失败。推销员根据以往的经验，知道随机选择的家庭购买保险单的概率为 0.10。与二项试验的性质相对照，我们观察到：

1. 该试验包括 10 次相同的试验，每次试验为访问一户家庭。
2. 每次试验都有两个结果是可能的：家庭购买保单（成功）和家庭未购买保单(失败)。
3. 每次推销中购买和未购买的概率都相同，分别为：$p=0.10$ 和 $1-p=0.90$。
4. 因为每户家庭都是随机选择的，故各次试验相互独立。

由于满足了四个假设，所以该例子也是一个二项试验，有关的随机变量是 10 户家庭中购买保单的户数。本例中 x 可取的值为：0，1，2，3，4，5，6，7，8，9，10。

二项试验的性质 3 称为**稳定性假设**，有时会与性质 4——试验的独立性相混淆。为了分辨它们的区别，再次考虑上面的保险推销员例子。如果随着时间的推移，推销员感觉到疲劳并失去了热情，例如，到第 10 次访问时，成功（售出保单）的概率降到了 0.05。在这种情况下，就不能满足性质 3（稳定性），也就构不成二项试验。即使性质 4——每户家庭购买保单的决定是独立的——满足时也是如此。

在涉及二项试验的应用中，有一个特殊的数学公式——**二项概率函数**(binomial probability

function)，可用来计算在 n 次试验中有 x 次成功的概率。利用第 4 章介绍的概率概念，我们将演示如何在一个实际问题中建立这个公式。

马丁服装商店问题

让我们考虑 3 个接连进入马丁服装商店的顾客的购买决定。根据过去的经验，商店经理估计任意一个顾客购买商品的概率为 0.30。那么 3 个顾客中有两个会购买的概率是多少?

利用树形图（图 5.3），我们可以看到：对 3 名顾客进行观察的试验有 8 个可能的结果。令 S 代表成功(购买)，F 代表失败(未购买)，我们想知道的是在 3 次试验(决定是否购买)中包含 2 次成功的试验结果。下一步，我们来证实这个包括 3 次购买决策的试验是一个二项试验。与二项试验的四条要求相对照，我们注意到：

1. 该试验可被描述为一个包括三次相同试验的序列，三个进店顾客中的每一个即为一次试验。
2. 两个结果——顾客购买(成功)和顾客未购买(失败)对每次试验都是可能的。
3. 顾客购买商品的概率(0.30)和顾客未购买商品的概率(0.70)被设定为对所有顾客都是相同的。
4. 每个顾客的购买决定都独立于其他顾客的购买决定。

因此，该试验满足二项试验的性质。

在 n 次试验中恰有 x 次成功的试验结果个数能够通过下面的公式计算：*

在 n 次试验中恰有 x 次成功的试验结果个数

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

式中 $n! = n(n-1)(n-2)\cdots(2)(1)$

并且 $0! = 1$

现在让我们回到涉及 3 位顾客的购买决定的马丁服装商店试验。可以使用公式(5.6)确定包含两次购买决定的试验结果个数，即在 $n = 3$ 次试验中获得 $x = 2$ 次成功的方法数。由公式(5.6)，我们有

$$\binom{n}{x} = \binom{3}{2} \frac{3!}{2!(3-2)!} = \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = \frac{6}{2} = 3$$

公式(5.6)的结果表明有 3 个试验结果包含了两次成功，从图 5.3 中我们看到这 3 个结果被表示为 *SSF*、*SFS* 和 *FSS*。

* 该公式曾在第 4 章介绍过，以确定在 n 个项目中一次抽取 x 个的组合数。对于二项试验，该组合公式给出了包含 x 次成功的试验结果(在 n 次试验的序列中)的数目。

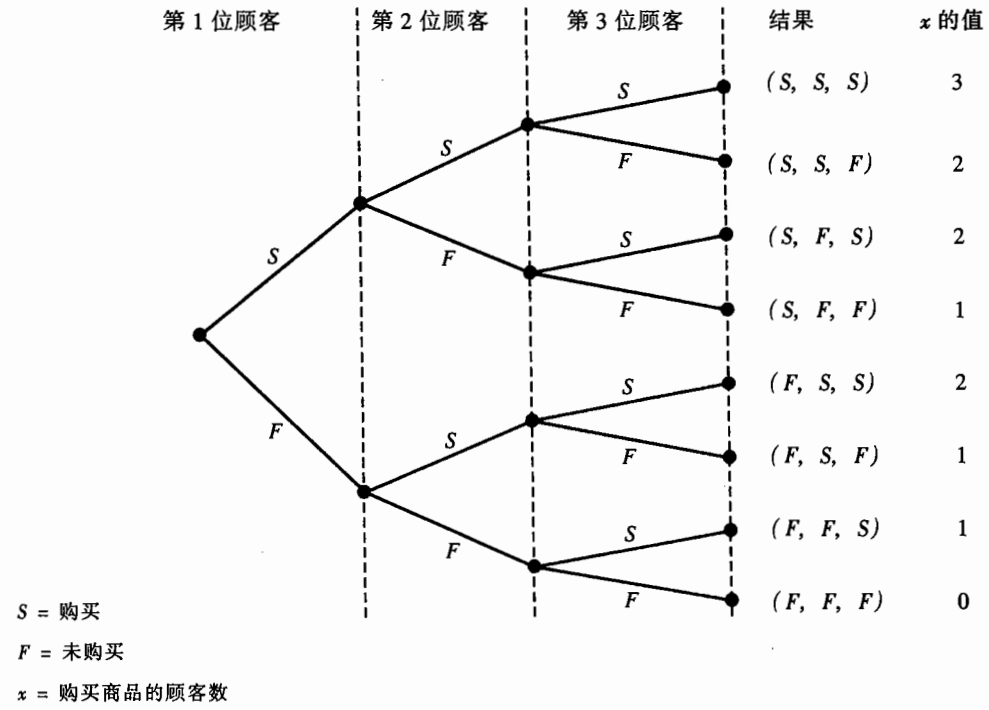


图 5.3　马丁服装商店问题的树形图

再次利用公式(5.6)来确定在 3 次试验中包含 3 次成功（购买）的试验结果个数，我们得到

$$\binom{n}{x} = \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = \frac{3\times 2\times 1}{3\times 2\times 1\times 1} = \frac{6}{6} = 1$$

从图 5.3 中我们看到有 1 个试验结果包含 3 次成功，它被表示为 *SSS*。

我们知道能够用公式(5.6)确定包含 x 次成功的试验结果个数，但是，如果我们想要确定在 n 次试验中包含 x 次成功的概率，还必须要知道每个试验结果的概率。因为二项试验中的每次试验都是独立的，我们只需把每个单次试验结果的概率相乘，就能够找到包括一系列成功和失败的二项试验结果的概率。

前两位顾客购买而第三位顾客未购买的概率可由以下公式得出：

$$pp(1-p)$$

因为在任何一次试验中购买的概率都是 0.30，所以上面问题的计算结果是

$$0.30\times 0.30\times 0.70 = 0.30^2\times 0.70 = 0.063$$

还有两个结果包含 2 次成功和 1 次失败。所有这三个包括 2 次成功的试验结果概率都列示在下表中：

单次试验结果			试验结果	试验结果的概率
第1位顾客	第2位顾客	第3位顾客	表示符号	
购买	购买	未购买	*SSF*	$pp(1-p)=p^2(1-p)=0.30^2\times0.70=0.063$
购买	未购买	购买	*SFS*	$p(1-p)p=p^2(1-p)=0.30^2\times0.70=0.063$
未购买	购买	购买	*FSS*	$(1-p)pp=p^2(1-p)=0.30^2\times0.70=0.063$

观察这三个包含2次成功的试验结果，它们都具有完全相同的**发生概率**，并且这一观察结果具有一般性。对于任何二项试验，所有在 n 次试验中取得 x 次成功的试验结果都具有相同的发生概率，其概率值如下：

$$\text{一个在 } n \text{ 次试验中包含 } x \text{ 次成功的特定试验结果序列出现的概率} = p^x(1-p)^{(n-x)} \tag{5.7}$$

对于马丁服装商店，该公式表明任何一个包含2次成功的试验结果具有概率为

$$p^2(1-p)^{(3-2)} = p^2(1-p)^1 = (0.30)^2(0.70)^1 = 0.063$$

公式(5.6)给出了在二项试验中恰有 x 次成功的试验结果个数，公式(5.7)则给出了每个包含 x 次成功的试验结果的发生概率，我们把这两个公式结合起来，得到了下面的二项概率函数：

二项概率函数

$$f(x)=\binom{n}{x}p^x(1-p)^{(n-x)} \tag{5.8}$$

式中 $f(x)$——在 n 次试验中取得 x 次成功的概率；

n——试验的次数；

$\binom{n}{x}$——$\dfrac{n!}{x!(n-x)!}$；

p——单次试验成功的概率；

$(1-p)$——单次试验失败的概率。

在马丁服装商店问题中，我们来计算没有顾客购买的概率、恰有1位顾客购买的概率、恰有2位顾客购买的概率以及3位顾客全都购买的概率。其计算过程汇总在表5.7内，该表还显示了购买商品的顾客数的概率分布。图5.4是概率分布图。

二项概率函数适用于**任何**二项试验。如果我们认为某试验具有二项试验的全部性质，并且知道 n、p 和$(1-p)$的值，就能够使用公式（5.8）来计算在 n 次试验中取得 x 次成功的概率。

表 5.7　购买商品的顾客数的概率分布

x	$f(x)$
0	$\frac{3!}{0!3!}\times 0.30^0\times 0.70^3 = 0.343$
1	$\frac{3!}{1!2!}\times 0.30^1\times 0.70^2 = 0.441$
2	$\frac{3!}{2!1!}\times 0.30^2\times 0.70^1 = 0.189$
3	$\frac{3!}{3!0!}\times 0.30^3\times 0.70^0 = 0.027$
	1.000

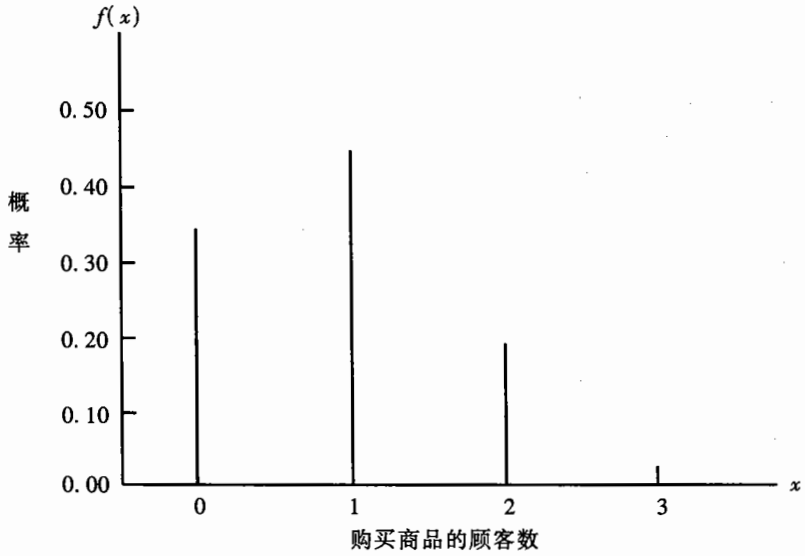


图 5.4　马丁服装商店问题的概率分布图形

用现有的计算方法制这样的表几乎是不可能的，公式(5.8)的值可以很容易直接估计出。

让我们再考虑对马丁试验作一些变化，比如有 10 位顾客而不是 3 位进入商店，二项概率函数公式（5.8）仍然适用。假定我们有二项试验的 $n=10$，$x=4$，$p=0.30$，则进入商店的 10 位顾客中恰有 4 人购买的概率为

$$f(4)=\frac{10!}{4!6!}\times 0.30^4\times 0.70^6 = 0.200\,1$$

使用二项概率表

人们已经用表格形式给出了在一个由 n 次试验构成的二项试验中取得 x 次成功的概率，该表格易于使用且比公式(5.8)更快捷。书后附录 B 的表 5 就是二项概率表，表 5.8 给出了这个概率表的一部分。为使用此表，我们必须确定二项试验的 n、p 和 x 的值。在表 5.8 上面的例子中，我们看到在 $n=10$，$p=0.40$ 时，二项试验取得 3 次成功的概率为 0.215 0。如果你直接使用二项概率函数计算的话，会得到与查表同样的答案。

现在我们使用表 5.8 来证实在马丁服装商店问题中 10 次试验恰有 4 次成功的概率。我们注意到对于 $n=10$，$x=4$ 且 $p=0.30$ 的二项试验，能够从二项概率表中直接读出概率值为：$f(4)=0.200\,1$。

表 5.8　从二项概率表选择数值的例子：$n=10$, $x=3$, $p=0.40$, $f(3)=0.215\,0$

		p									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.630 2	.387 4	.231 6	.134 2	.075 1	.040 4	.020 7	.010 1	.004 6	.002 0
	1	.298 5	.387 4	.367 9	.302 0	.225 3	.155 6	.100 4	.060 5	.033 9	.017 6
	2	.062 9	.172 2	.259 7	.302 0	.300 3	.266 8	.216 2	.161 2	.111 0	.070 3
	3	.007 7	.044 6	.106 9	.176 2	.233 6	.266 8	.271 6	.250 8	.211 9	.164 1
	4	.000 6	.007 4	.028 3	.066 1	.116 8	.171 5	.219 4	.250 8	.260 0	.246 1
	5	.000 0	.000 8	.005 0	.016 5	.038 9	.073 5	.118 1	.167 2	.212 8	.246 1
	6	.000 0	.000 1	.000 6	.002 8	.008 7	.021 0	.042 4	.074 3	.116 0	.164 1
	7	.000 0	.000 0	.000 0	.000 3	.001 2	.003 9	.009 8	.021 2	.040 7	.070 3
	8	.000 0	.000 0	.000 0	.000 0	.000 1	.000 4	.001 3	.003 5	.008 3	.017 6
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 8	.002 0
10	0	.598 7	.348 7	.196 9	.107 4	.056 3	.028 2	.013 5	.006 0	.002 5	.001 0
	1	.315 1	.387 4	.347 4	.268 4	.187 7	.121 1	.072 5	.040 3	.020 7	.009 8
	2	.074 6	.193 7	.275 9	.302 0	.281 6	.233 5	.175 7	.120 9	.076 3	.043 9
	3	.010 5	.057 4	.129 8	.201 3	.250 3	.266 8	.252 2	**.215 0**	.166 5	.117 2
	4	.001 0	.011 2	.040 1	.088 1	.146 0	.200 1	.237 7	.250 8	.238 4	.205 1
	5	.000 1	.001 5	.008 5	.026 4	.058 4	.102 9	.153 6	.200 7	.234 0	.246 1
	6	.000 0	.000 1	.001 2	.005 5	.016 2	.036 8	.068 9	.111 5	.159 6	.205 1
	7	.000 0	.000 0	.000 1	.000 8	.003 1	.009 0	.021 2	.042 5	.074 6	.117 2
	8	.000 0	.000 0	.000 0	.000 1	.000 4	.001 4	.004 3	.010 6	.022 9	.043 9
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 5	.001 6	.004 2	.009 8
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.001 0

尽管二项概率表相对容易使用，但不可能把二项试验中所有 n 和 p 的值都列出来。不过，使用现在的计算器根据公式(5.8)计算想要的概率并不困难，尤其是在试验的次数不很多时。在练习中，你应该尽量练习使用公式(5.8)来计算二项概率，除非题中特别要求你使用二项概率表。

像 Minitab 这样的统计软件包和像 Excel 这样的电子表格软件也有计算二项概率的功能。再来考虑马丁服装商店的例子，其中的 $n=10$，$p=0.30$。图 5.5 显示了由 Minitab 生成的 x 取所有可

能值时的二项概率。注意这些值与表 5.8 中 $p = 0.30$ 列的数值相同。在附录 5.1 中将演示如何利用 Minitab 产生图 5.5 的结果，附录 5.2 描述了怎样利用 Excel 来计算二项概率。

x	P(X = x)
0.00	0.0282
1.00	0.1211
2.00	0.2335
3.00	0.2668
4.00	0.2001
5.00	0.1029
6.00	0.0368
7.00	0.0090
8.00	0.0014
9.00	0.0001
10.00	0.0000

图 5.5　显示马丁服装商店问题二项概率的 Minitab 输出结果

二项概率分布的数学期望和方差

在 5.3 节，我们给出了计算离散随机变量的数学期望和方差的公式。在特定的情况下，随机变量可能具有二项概率分布，并且其试验次数已知为 n，成功概率已知为 p 时，计算数学期望和方差的通用公式就可以简化为

二项概率分布的数学期望和方差

$$E(x) = \mu = np \tag{5.9}$$

$$\text{Var}(x) = \sigma^2 = np(1-p) \tag{5.10}$$

对于有 3 位顾客的马丁服装商店问题，我们能够应用公式(5.9)计算购买商品顾客数的数学期望为

$$E(x) = np = 3 \times 0.30 = 0.9$$

假设下个月马丁服装商店预计有 1 000 个顾客会进入商店，那么购买商品顾客的期望数目是多少？答案是 $\mu = np = 1\,000 \times 0.3 = 300$。于是，为了增加销售的期望值，马丁服装商店必须吸引更多的顾客进入商店，和/或设法增加进店顾客购买商品的概率。对于有 3 位顾客的马丁服装商店问题，我们看到购物顾客数的方差和标准差分别为

$$\sigma^2 = np(1-p) = 3 \times 0.3 \times 0.7 = 0.63$$

$$\sigma = \sqrt{0.63} = 0.79$$

np 图和批次抽样验收作为质量管理的方法，都是二项概率分布应用的极好例子。这些方法将在第 20 章：质量管理统计方法中进行讨论。

在下个月，有 1 000 位顾客进入商店，则购物顾客数的方差和标准差分别为

$$\sigma^2 = np(1-p) = 1\,000 \times 0.3 \times 0.7 = 210$$

$$\sigma = \sqrt{210} = 14.49$$

评 注

1. 附录 B 的二项概率表只给出 p 小于等于 0.50 的值，似乎当成功的概率 p 大于 0.50 时我们就无法使用该表。然而，注意到 $n-x$ 次失败的概率也就是 x 次成功的概率，我们仍然能够使用该表。这样当成功的概率 p 超过 0.50 时，可转而计算 $n-x$ 次失败的概率。这时失败的概率 $1-p$ 将小于 0.50。
2. 某些二项概率表给出的是累积形式。在使用这种表时，必须做减法以得到 n 次试验中恰有 x 次成功的概率。例如，对于 $f(2) = p(x \leqslant 2) - p(x \leqslant 1)$，我们的表格直接提供的是概率。为了利用该表格计算累积概率，只需把单个概率值相加即可。例如，使用我们的表格计算 $p(x \leqslant 2)$ 时，应该用式子 $f(0) + f(1) + f(2)$。

练 习

方 法

25. 考虑一个由两次试验构成的二项试验，$p = 0.4$。

自测题

a. 画出树形图以显示出它是一个包括 2 次试验的二项试验(参见图 5.3)。

b. 计算 1 次成功的概率 $f(1)$。

c. 计算 $f(0)$。

d. 计算 $f(2)$。

e. 计算至少 1 次成功的概率。

f. 计算它的数学期望、方差和标准差。

26. 某二项试验的 $n=10$，$p=0.10$。
    a. 计算 $f(0)$。
    b. 计算 $f(2)$。
    c. 计算 $P(x\leqslant 2)$。
    d. 计算 $P(x\geqslant 1)$。
    e. 计算 $E(x)$。
    f. 计算 $\mathrm{Var}(x)$和 σ。

27. 某二项试验的 $n=20, p=0.70$。使用二项概率表(附录 B 的表 5)来回答从(a)到(d)的问题。
    a. 计算 $f(12)$。
    b. 计算 $f(16)$。
    c. 计算 $P(x\geqslant 16)$。
    d. 计算 $P(x\leqslant 15)$。
    e. 计算 $E(x)$。
    f. 计算 $\mathrm{Var}(x)$和 σ。

应　用

28. 由 American Savings Education Council, the Employee Benefit Research Institute 和 Mathew Greenwald & Associates 三家机构发起进行了 1999 年青年和金钱调查。通过与 1 000 名年龄在 16—22 岁的学生谈论他们的个人财务情况，调查发现 33% 的学生拥有自己的信用卡。
    a. 在包括 6 名学生的样本中，有 2 人拥有自己信用卡的概率是多少？
    b. 在包括 6 名学生的样本中，至少有 2 人拥有自己信用卡的概率是多少？
    c. 在包括 10 名学生的样本中，无人拥有自己信用卡的概率是多少？

29. 根据《商业周刊》Harris 民意测验对 1 035 名成年人所做的调查，有 40% 的被调查者强烈赞同商业对美国人生活影响过大的看法（*Business Week*, September 11, 2000）。假定该百分比代表了美国人的总体比例，我们从美国人总体的横截面中选取 20 人作为样本，了解他们对商业在生活中所起作用的看法。这些人中至少有 5 人认为商业对美国人生活影响过大的概率是多少？

30. 当一台新机器正常运转时，其生产的产品中只有 3% 是次品。假定随机抽取机器的两件产品，我们关心的是发现的次品数目。



    a. 描述该情况符合二项试验需满足的条件。
    b. 画出类似图 5.3 的树形图，以显示出该问题是一个包括 2 次试验的二项试验。
    c. 有多少个试验结果为恰好发现 1 件次品？
    d. 计算没有发现次品、恰好发现 1 件次品和恰好发现 2 件次品的概率。

31. 5% 的美国卡车司机是女性(*Statistical Abstract of the United States*, 1997)。假设随机选取了 10 名卡车司机，就工作环境的质量问题访问他们。

a. 选取 10 名司机是不是一个二项试验？请解释。
b. 抽到 2 名女司机的概率是多少？
c. 没有抽到女司机的概率是多少？
d. 至少抽到 1 位女司机的概率是多少？

32. 军事雷达和导弹探测系统的设计目的是在敌人攻击时向国家发出警报。它们的可靠性问题是指探测系统是否能够识别攻击并发出警报。假设某个探测系统能探测到导弹攻击的概率为 0.90。利用二项概率分布回答下列问题：
a. 一个单独的探测系统能够探测到攻击的概率是多少？
b. 如果在同一区域安装了两套探测系统，每一套都独立运转。至少有 1 套系统探测到攻击的概率是多少？
c. 如果安装 3 套系统，至少有 1 套系统探测到攻击的概率是多少？
d. 你建议使用多套系统吗？请解释。

33. 50% 的中型制造商计划利用北美自由贸易协定创造机会，派管理代表访问加拿大和墨西哥（*Grant Thornton Survey of American Manufacturers*, 1995）。位于加拿大多伦多的一家进出口集团邀请美国 20 家中型制造商参加一个开发贸易机会的会议。
a. 有 12 家或更多公司会派出代表的概率是多少？
b. 不超过 5 家公司派出代表的概率是多少？
c. 你预期有多少家公司会派出代表？
d. 派出代表的公司数的方差和标准差是多少？

34. 40% 的商务旅行者随身携带移动电话或笔记本电脑（*USA Today*, September 12, 2000）。在一个由 15 名商务旅行者构成的样本中：
a. 有 3 人携带移动电话或笔记本电脑的概率是多少？
b. 有 12 人既未携带移动电话也未携带笔记本电脑的概率是多少？
c. 至少 3 人携带移动电话或笔记本电脑的概率是多少？

35. 某大学发现有 20% 的学生没有完成基础统计学课程。假定本学期共有 20 名学生注册了这门课程。
a. 有 2 人或以下的学生放弃该课程的概率是多少？
b. 恰有 4 名学生放弃的概率是多少？
c. 超过 3 人放弃课程的概率是多少？
d. 放弃人数的期望值是多少？

36. 对于二项随机变量的特殊例子，我们发现其方差可用如下公式 $\sigma^2 = np(1-p)$ 进行计算。使用表 5.7 马丁服装商店问题的数据，我们发现 $\sigma^2 = np(1-p) = 0.63$。利用离散随机变量方差的通用定义公式（5.5）以及表 5.7 的数据，证实其方差的确是 0.63。

37. 目前有 29% 的律师和法官是女性（*Statistical Abstract of the United States*, 1997）。在一个由 30 名

法官和律师组成的法律小组中，女性人数的期望值是多少？方差和标准差是多少？

5.5　泊松概率分布

这一节我们将学习一种在估计特定时间或空间段内事件的发生次数方面十分有用的离散随机变量。例如，有关的随机变量可以是在一小时内到达汽车清洗站的汽车数，在 10 英里长的公路上需要修理的汽车数，或者在 100 英里长的管道上的泄露点个数。如果满足以下两个性质，则事件发生的次数就是一个可用**泊松概率分布**(poisson probability distribution)来描述的随机变量。

泊松分布经常被用来建立在排队情况下到达率的模型。

泊松试验的性质

1. 事件在任意两个等长度的区间内发生一次的概率相等。
2. 事件在任意区间内是否发生和在其他区间的发生情况相互独立。

泊松概率函数(poisson probability function)由公式（5.11）给出：

泊松概率函数

$$f(x)=\frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

式中　$f(x)$——事件在一个区间内发生 x 次的概率；

μ——在一个区间内事件发生次数的平均值或数学期望；

e——2.718 28。

西蒙·泊松（Siméon Poisson）于 1802—1808 年在巴黎的 Ecole Polytechnique 教授数学，1837 年，它发表了题目是《对于刑事和民事裁决概率问题的研究》的文章。该文包括了对后来成为泊松分布的问题所做的讨论。

在我们用一个具体例子演示怎样应用泊松分布以前，注意发生次数 x 是没有上限的，它是离散随机变量，其值可取无穷数列（$x = 0, 1, 2, \cdots$）。

一个涉及时间间隔的例子

假设我们想知道的是在周日早上 15 分钟内到达某银行出纳窗口的汽车数。如果我们能够假定在任意两个等长的时间段内汽车到达的概率相同，且在任意时段汽车到达与否和其他时段汽车到达与否相互独立的话，就可以应用泊松概率函数。假设这些条件都满足，对历史数据的分析表明，在 15 分钟内平均的到达车辆数为 10。该情况下，适用如下的概率函数：

$$f(x)=\frac{10^{x}e^{-10}}{x!}$$

贝尔实验室在建立打入电话的模型方面，已经对泊松分布有了实质性的应用。

这里的随机变量 x = 在任意 15 分钟内到达的汽车数。

如果管理者想要知道在 15 分钟内恰有 5 辆到达的概率，我们令 $x=5$ 然后得到

$$15\text{ 分钟内恰有 5 辆到达的概率}=f(5)=\frac{10^{5}e^{-10}}{5!}=0.037\,8$$

尽管通过计算 $\mu=10$ 和 $x=5$ 的概率函数可以确定这一概率，但查泊松概率分布表会更容易些。这些表给出了 x 和 μ 取各个值时的概率，我们在附录 B 的表 7 包括了这样一张表。为方便起见，我们在表 5.9 复制了该表的一部分。我们注意到为使用泊松概率表，仅需要知道 x 和 μ 的值即

表 5.9　从泊松概率表选择数值的例子：$\mu=10$，$x=5$，$f(5)=0.037\,8$

					μ					
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.8**	**10**
0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 0
1	.001 0	.000 9	.000 9	.000 8	.000 7	.000 7	.000 6	.000 5	.000 5	.000 5
2	.004 6	.004 3	.004 0	.003 7	.003 4	.003 1	.002 9	.002 7	.002 5	.002 3
3	.014 0	.013 1	.012 3	.011 5	.010 7	.010 0	.009 3	.008 7	.008 1	.007 6
4	.031 9	.030 2	.028 5	.026 9	.025 4	.024 0	.022 6	.021 3	.020 1	.018 9
5	.058 1	.055 5	.053 0	.050 6	.048 3	.046 0	.043 9	.041 8	.039 8	**.037 8**
6	.088 1	.085 1	.082 2	.079 3	.076 4	.073 6	.070 9	.068 2	.065 6	.063 1
7	.114 5	.111 8	.109 1	.106 4	.103 7	.101 0	.098 2	.095 5	.092 8	.090 1
8	.130 2	.128 6	.126 9	.125 1	.123 2	.121 2	.119 1	.117 0	.114 8	.112 6
9	.131 7	.131 5	.131 1	.130 6	.130 0	.129 3	.128 4	.127 4	.126 3	.125 1
10	.119 8	.121 0	.121 9	.122 8	.123 5	.124 1	.124 5	.124 9	.125 0	.125 1
11	.099 1	.101 2	.103 1	.104 9	.106 7	.108 3	.109 8	.111 2	.112 5	.113 7
12	.075 2	.077 6	.079 9	.082 2	.084 4	.086 6	.088 8	.090 8	.092 8	.094 8
13	.052 6	.054 9	.057 2	.059 4	.061 7	.064 0	.066 2	.068 5	.070 7	.072 9
14	.034 2	.036 1	.038 0	.039 9	.041 9	.043 9	.045 9	.047 9	.050 0	.052 1
15	.020 8	.022 1	.023 5	.025 0	.026 5	.028 1	.029 7	.031 3	.033 0	.034 7
16	.011 8	.012 7	.013 7	.014 7	.015 7	.016 8	.018 0	.019 2	.020 4	.021 7
17	.006 3	.006 9	.007 5	.008 1	.008 8	.009 5	.010 3	.011 1	.011 9	.012 8
18	.003 2	.003 5	.003 9	.004 2	.004 6	.005 1	.005 5	.006 0	.006 5	.007 1
19	.001 5	.001 7	.001 9	.002 1	.002 3	.002 6	.002 8	.003 1	.003 4	.003 7
20	.000 7	.000 8	.000 9	.001 0	.001 1	.001 2	.001 4	.001 5	.001 7	.001 9
21	.000 3	.000 3	.000 4	.000 4	.000 5	.000 6	.000 6	.000 7	.000 8	.000 9
22	.000 1	.000 1	.000 2	.000 2	.000 2	.000 2	.000 3	.000 3	.000 4	.000 4
23	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2
24	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1

可。从表 5.9 我们看到：通过查找对应于 $x=5$ 的行，和对应于 $\mu=10$ 的列，就能够得到在 15 分钟内恰有 5 辆汽车到达的概率。于是我们有 $f(5)=0.037\,8$。

这个例子涉及到一个 15 分钟的时段，但也可以使用其他的时段。假定我们需要计算在 3 分钟内恰有 1 辆汽车到达的概率。因为在 15 分钟内到达的期望数是 10 辆，则 10/15 = 2/3 辆就是在 1 分钟内到达的期望车数，从而 (2/3) ×3 分钟 = 2 是 3 分钟内到达的期望车数。因此，当 $\mu=2$ 时在 3 分钟时段内到达 x 辆的概率可由下列泊松概率函数计算：

$$f(x)=\frac{2^x e^{-2}}{x!}$$

假如现在我们想要计算在 3 分钟内恰有 1 辆到达的概率，可利用该函数得到

$$\text{在 3 分钟内恰有 1 辆到达的概率}=f(1)=\frac{2^1 e^{-2}}{1!}=0.270\,7$$

一个涉及长度或距离间隔的例子

让我们演示一个与时间间隔无关的泊松概率分布应用。假定我们需要知道的是公路在重新整修一个月以后存在的严重缺陷个数。我们假设在公路上任意两段等长度的距离内存在缺陷的概率相同，而在任意一段距离内是否存在缺陷与另一段内是否存在缺陷无关。于是，我们就能够对其应用泊松概率分布。

假设我们知道公路在重新整修一个月以后平均每英里存在两个严重缺陷，让我们来求出在3英里长的路段内没有严重缺陷的概率。由于我们关心的是 3 英里长的路段，所以 $\mu=2$ 个缺陷/英里 ×3 英里 = 6，表示在 3 英里长公路上的期望缺陷数。使用附录 B 中的表 7 或公式 (5.11)，我们看到没有缺陷的概率为 0.002 5。所以 3 英里长的路段内没有严重缺陷的概率几乎为零，事实上，由于 1 − 0.002 5 = 0.997 5，这段路面至少存在 1 个严重缺陷的概率为 0.997 5。

练　习

方　法

38. 考虑 $\mu=3$ 的泊松概率分布。
    a. 写出适当的泊松概率函数。
    b. 计算 $f(2)$。
    c. 计算 $f(1)$。
    d. 计算 $P(x\geq 2)$。

39. 考虑一个在每个时间段内事件平均发生 2 次的泊松概率分布。



    a. 给出适当的泊松概率函数。
    b. 在 3 个时间段内事件的平均发生次数是多少？

c. 求出适当的泊松概率函数以确定在3个时间段内事件发生 x 次的概率。

d. 计算在1个时间段内事件发生2次的概率。

e. 计算在3个时间段内事件发生6次的概率。

f. 计算在2个时间段内事件发生5次的概率。

应 用

40. Regional 航空公司的机票预订处平均每小时打进48个电话。



a. 计算在5分钟的时段内接到3个电话的概率。

b. 计算在15分钟内恰好接到10个电话的概率。

c. 假设目前有一个电话在线。如果话务员接听一个电话需要5分钟，在这段时间内你期望有多少个电话处于等待中？无人等待的概率是多少？

d. 如果目前没有电话需处理，话务员能够休息3分钟而不被电话打扰的概率是多少？

41. 在本地某大学的电话注册期间，平均每2分钟打进1个电话。

a. 在1小时内期望的电话数是多少？

b. 在5分钟内接到3个电话的概率是多少？

c. 在5分钟内没有接到电话的概率是多少？

42. 每年 *Barron's* 的订户因私人原因在国内飞行旅行的平均次数是4(*Barron's 1*995 Primary Reader Survey)。

a. 订户在1年内因私人原因进行2次国内飞行的概率是多少？

b. 订户在3个月内因私人原因进行国内飞行的平均次数是多少？

c. 订户在6个月内因私人原因进行1次或更多次国内飞行的概率是多少？

43. 在某大型国际机场，乘客们随机并且独立地到达乘客监视器。其平均到达率为每分钟10名乘客。

a. 在1分钟内无人到达的概率是多少？

b. 在1分钟内有3名或更少乘客到达的概率是多少？

c. 在15秒内无人到达的概率是多少？

d. 在15秒内至少有1人到达的概率是多少？

44. *The Wall Street Journal Interactive Edition* 订户的投资活动表明，他们每年进行证券交易的平均次数为15次（*The Wall Street Journal* Subscriber Study，1999）。假定某个投资者以该频率进行交易，此外，还假定在任意两个月内该投资者进行交易的概率相同，并且在一个月内交易与否与在其他任意月份的交易情况相互独立。回答下列问题：

a. 该投资者每月交易的平均次数是多少？

b. 该投资者在1个月内没有进行证券交易的概率是多少？

c. 该投资者在1个月内恰有1次证券交易的概率是多少？

d. 该投资者在 1 个月内证券交易超过 1 次的概率是多少？

45. 每年发生在 15—24 岁年龄组中的火器事故导致了 450 起死亡(*Accident Facts*，National Safety Council，1996)。

a. 每个星期由火器引起的死亡事故的平均次数是多少？

b. 在任意 1 个星期内，没有发生火器死亡事故的概率是多少？

c. 在任意 1 天内，由火器引起 2 次或以上死亡事故的概率是多少？

5.6　超几何概率分布

超几何概率分布与二项概率分布紧密相关。两种概率分布的主要不同之处在于：超几何分布的各次试验不是互相独立的，并且每次试验成功的概率各不相同。

在超几何概率分布的应用中，通常采用的标记是：令 r 代表在容量为 N 的总体中用成功表示的元素个数；令 $N-r$ 代表在总体中用失败表示的元素个数。**超几何概率函数**(hypergeometric probability function)可用来计算在一个包括 n 个元素的随机样本内进行无放回地选择，得到 x 次成功和 $n-x$ 次失败的概率。要取得这个结果，我们必须从总体的 r 次成功中抽到 x 次成功，从 $N-r$ 次失败中抽到 $n-x$ 次失败。下面的超几何概率函数给出了 $f(x)$ 即在容量为 n 的样本中获得 x 次成功的概率的计算方法：

超几何概率函数

$$f(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}\quad,\quad 0\leqslant x\leqslant r \tag{5.12}$$

式中　$f(x)$——在 n 次试验中获得 x 次成功的概率；

n——试验次数；

N——总体内的元素个数；

r——总体内用成功表示的元素个数。

注意 $\binom{N}{n}$ 表示从容量为 N 的总体中选择 n 容量样本的方法数；$\binom{r}{x}$ 表示从总体的总计 r 次成功中选择 x 次成功的方法数；$\binom{N-r}{n-x}$ 表示从总体内总计 $N-r$ 次失败中选择 $n-x$ 次失败的方法数。

为了说明使用公式(5.12)计算超几何概率的过程，让我们考虑这样一个问题：从一个五人委员会中选择两人参加在拉斯韦加斯的例会。假设五人委员会由三女二男组成。为确定随机选择到 2 名妇女的概率，我们能够对 $n=2$，$N=5$，$r=3$，$x=2$ 的情况，使用公式(5.12)

$$f(2)=\frac{\binom{3}{2}\binom{2}{0}}{\binom{5}{2}}=\frac{\left(\frac{3!}{2!1!}\right)\left(\frac{2!}{2!0!}\right)}{\left(\frac{5!}{2!3!}\right)}=\frac{3}{10}=0.30$$

假设我们后来知道将有 3 名委员参加这次旅行，则取 $n=3$，$N=5$，$r=3$，$x=2$，从而在 3 名委员中恰有 2 名女性的概率为

$$f(2)=\frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{2}}=\frac{\left(\frac{3!}{2!1!}\right)\left(\frac{2!}{1!1!}\right)}{\left(\frac{5!}{2!3!}\right)}=\frac{6}{10}=0.60$$

再举一个例子，假设某总体由 10 个项目组成，其中 4 项有缺陷，6 项没有缺陷。问在容量为 3 的随机样本中包括 2 项缺陷项目的概率是多少？对该问题，我们可以把抽到缺陷项目作为“成功”。这时 $n=3$，$N=10$，$r=4$，$x=2$，我们能够计算如下：

$$f(2)=\frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}}=\frac{\left(\frac{4!}{2!2!}\right)\left(\frac{6!}{1!5!}\right)}{\frac{10!}{3!7!}}=\frac{36}{120}=0.30$$

练　习

方　法

46. 假设 $N=10$ 且 $r=3$，计算具有下列 n 和 x 值的超几何概率。



a. $n=4$，$x=1$

b. $n=2$，$x=2$

c. $n=2$，$x=0$

d. $n=4$，$x=2$

47. 假设 $N=15$ 且 $r=4$，$x=3$ 和 $n=10$ 时的超几何概率是多少？

应　用

48. 根据《饮料文摘》的数据，可口可乐和百事可乐名列销售量的第一和第二位（*The Wall Street Journal Almanac*, 1998）。假设在一个 10 人小组中，有 6 人喜欢可口可乐，4 人喜欢百事可乐。从这 10 个人中选择一个 3 人随机样本。

a. 样本中恰有 2 人喜欢可口可乐的概率是多少？

b. 样本中的大多数人（2 人或 3 人）喜欢百事可乐的概率是多少？

49. 黑杰克，通常又被称为“21 点”，是一种在拉斯韦加斯赌场中流行的赌博游戏。在游戏中，每个玩家有两张牌，花牌（J、Q 和 K）和 10 的分值为 10，A 的分值为 1 或 11。在一副 52 张扑克中，共有 16 张牌分值为 10 以及 4 张 A。
   a. 两张牌都是 A 或 10 点牌的概率是多少？
   b. 两张牌都是 A 的概率是多少？
   c. 两张牌的分值都为 10 的概率是多少？
   d. 一张 10 点牌再加一张 A 的分值为 21 点，称为黑杰克。利用你在(a)、(b)、(c)中的答案确定一名玩家得到黑杰克的概率(提示：(d)不是超几何分布问题，你需要找到如何把(a)、(b)、(c)结合起来回答这个问题的逻辑关系的方法)。

50. 

Axline 计算机公司在两个工厂生产个人电脑，一个位于得克萨斯，另一个位于夏威夷。得克萨斯的工厂有 40 名员工，夏威夷工厂有 20 名员工。假设要从两工厂抽取 10 名雇员作为随机样本填写利益调查问卷。
   a. 样本没有抽到夏威夷工厂员工的概率是多少？
   b. 样本中有 1 名员工来自夏威夷工厂的概率是多少？
   c. 样本中有 2 名或更多员工来自夏威夷工厂的概率是多少？
   d. 样本中有 9 名员工来自得克萨斯工厂的概率是多少？

51. St. Andrew 学校的六年级班中有 25 名学生(14 个男生和 11 个女生)。周四有 5 名学生缺课。
   a. 缺课学生有 2 个是女生的概率是多少？
   b. 缺课学生有 2 个是男生的概率是多少？
   c. 所有缺课学生都是男生的概率是多少？
   d. 没有一个缺课学生是男生的概率是多少？

52. 在一批 10 件的货物中有 2 件次品，8 件正品。在检查货物时，将选择一个样本并对其进行测试。如果在样本中发现次品，该批 10 件货物将被全部拒收。
   a. 如果选择 3 件货物作为样本，那么这批货物被拒收的概率是多少？
   b. 如果选择 4 件货物作为样本，这批货物被拒收的概率是多少？
   c. 如果选择 5 件货物作为样本，这批货物被拒收的概率是多少？
   d. 如果管理者需要以 0.90 的概率拒收这批包括 2 件次品、8 件正品的货物，你建议应选择多大的样本？

小 结

本章介绍了随机变量的概念，以给出对试验结果的数值描述。我们看到，随机变量的概率分布显示了随机变量取不同值时的概率。对于任意的离散随机变量，都是通过概率函数来定义概率分布，概率函数用 $f(x)$表示，它给出了随机变量取每个值的概率。而一旦定义了概率函数以后，就可以计算随机变量的数学期望和方差。

二项概率分布可用来确定在 n 次试验中得到 x 次成功的概率，它具有下列性质：

1. 二项试验是把相同的单次试验进行了 n 次所形成的一个序列。
2. 每一单次试验都有两个可能结果，一个称为成功，另一个称为失败。
3. 每一单次试验的成功概率 p 都相等，从而每一单次试验的失败概率 $1-p$ 也相等。
4. 各次单次试验之间都是相互独立的。

当以上四个条件都满足时，就可以使用二项概率函数或二项概率表来确定在 n 次试验中得到 x 次成功的概率。本章还给出了二项概率分布的均值和方差公式。

如果需要确定在一段时间或空间内事件发生 x 次的概率时，可以使用泊松分布。但要适用泊松分布，还必须遵循以下的假设：

1. 在任意两个等长度的区间内，事件发生一次的概率相等。
2. 事件在任意的区间内是否发生与在其他区间内的发生情况互相独立。

第三种离散概率分布——超几何概率分布在5.6节作了介绍。像二项分布一样，它也被用来计算在 n 次试验中得到 x 次成功的概率。但与二项分布有所不同的是，它的各次单次试验的成功概率不同。

术语释义

随机变量：对试验结果的数值性描述。
离散随机变量：可取有限个值或者可取一个存在间隔的无穷数列值的随机变量。
连续随机变量：可取一个区间或区间集合内任意值的随机变量。
概率分布：描述了随机变量取各个可能值时所对应的不同概率。
概率函数：用 $f(x)$表示的函数，它提供了随机变量取每个值时的概率。
离散均匀概率分布：在这种概率分布形式下，随机变量取每个可能值的概率相同。
数学期望：对随机变量的平均值或中心位置的度量。
方差：对随机变量的变异程度或分散程度所进行的度量。
标准差：方差的正平方根。
二项试验：具有在5.4节开始部分所列出的四种性质的概率试验。

二项概率分布： 一种概率分布，用来显示在二项试验的 n 次单次试验中有 x 次成功的概率。

二项概率函数： 用来计算二项试验概率的函数。

泊松概率分布： 一种概率分布，显示了在某个特定的时间区间或空间区间内，一个事件发生 x 次的概率。

泊松概率函数： 用来计算泊松概率的函数。

超几何概率函数： 该函数是用于计算当每次的单次试验相互关联时，在 n 次单次试验中有 x 次成功的概率。

重要公式

离散均匀概率函数

$$f(x) = 1/n \tag{5.3}$$

式中，n—— 随机变量可取的数值个数

离散随机变量的数学期望

$$E(x) = \mu = \sum xf(x) \tag{5.4}$$

离散随机变量的方差

$$\mathrm{Var}(x) = \sigma^2 = \sum (x-\mu)^2 f(x) \tag{5.5}$$

在 n 次试验中恰有 x 次成功的试验结果个数

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{5.6}$$

二项概率函数

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)} \tag{5.8}$$

二项概率分布的数学期望

$$E(x) = \mu = np \tag{5.9}$$

二项概率分布的方差

$$\mathrm{Var}(x) = \sigma^2 = np(1-p) \tag{5.10}$$

泊松概率函数

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{5.11}$$

超几何概率函数

$$f(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \quad ,0\leqslant x\leqslant r \tag{5.12}$$

补充练习

53.《华尔街日报》/NBC 新闻调查向 2 004 名成年人询问了他们对于职业发展机会的看法（*The Wall Street Journal*, September19, 1997）。23% 的调查对象认为他们非常满意；38% 认为比较满意；3% 无法确定；18% 认为比较不满意；18% 认为非常不满意。以 x 作为表示满意水平的随机变量，我们令 $x=5$ 表示非常满意，$x=4$ 表示比较满意，$x=3$ 表示无法确定，$x=2$ 表示比较不满意，$x=1$ 表示非常不满意。

a. 利用上面的数据建立职业发展机会满意水平的概率分布。

b. 计算期望的满意水平。是否较高的分数表示更大的满意?

c. 计算满意分数的方差和标准差。

54. Louis Harris & Associates 公司对 1 009 名成年人进行了观点调查(*Business Week*, December 29, 1997)。表 5.10 显示了对有关股市价值问题观点的概率分布。

a. 说明表 5.10 的概率分布是否满足所有概率分布的性质。

b. 观点概率分布的数学期望和方差分别是多少?

c. 对于人们是否认为股市价值被高估了这一问题进行评论。

表 5.10　对有关股市价值观点的概率分布

股市价值	随机变量（x）	概率 $f(x)$
严重低估	1	0.02
有些低估	2	0.06
定价合理	3	0.28
有些高估	4	0.54
严重高估	5	0.10

55. Midwestern 学院的预算程序生成下一年的预测费用为(以百万美元为单位)：9，10，11，12 和 13。因为实际的费用未知，故分别给它们分配概率为：0.3，0.2，0.25，0.05 和 0.2。

a. 写出预测费用的概率分布。

b. 预测费用的数学期望是多少?

c. 预测费用的方差是多少?

d. 如果下年的计划收入是 1 200 万美元，评论该学院的财务状况。

56. Kristen DelGuzzi 对一个 3 年期 Hamilton 县的法官和法庭表现进行了研究(*The Cincinnati Enquirer*, January 11, 1998)。其中的一个发现是在普通诉讼法庭裁决的案件中有 4% 的案件被上诉。
    a. 如果该法庭某天审理了 20 个案件，有 3 件被上诉的概率是多少？
    b. 如果该法庭某天审理了 20 个案件，没有 1 件被上诉的概率是多少？
    c. 该法庭 1 个月审理 1 200 个案件很正常，问被上诉案件的期望数是多少？
    d. 如果该法庭 1 个月审理了 1 200 个案件，那么被上诉案件数的方差和标准差是多少？

57. 某公司正计划拜访互联网用户，以了解它计划开设的网站被不同年龄层次用户接受的情况。根据人口普查署的数据，年龄在 18—54 岁之间的居民有 40% 使用互联网，在 55 岁及以上年龄段的居民有 12% 使用互联网(*Statistical Abstract of the United States*, 2000)。
    a. 必须要接触多少个 18—54 岁年龄段的居民，才能使找到的互联网用户人数的期望值为至少 10 人？
    b. 必须要接触多少个 55 岁及以上年龄段的居民，才能使找到的互联网用户人数的期望值为至少 10 人？
    c. 如果你接触的 18—54 岁年龄段的居民人数为(a)中的答案数，那么其中互联网用户人数的标准差是多少？
    d. 如果你接触的 55 岁及以上年龄段的居民人数如(b)答案所示，那么其中互联网用户人数的标准差是多少？

58. 许多公司使用一种被称为**抽样验收**的质量控制技术，来检测运到的零部件、原材料等货物。在电子工业中，电子元件通常是以很大的批量从供应商那里发货。检测由 n 个元件组成的样本可被看做是二项试验的 n 次试验，根据每个元件的测试（即单次试验）结果可以把元件分为良好和不合格。如果某批货物中的不合格元件不超过 1% 的话，Reynolds 电子公司就将从供应商那里接受该批货物。假定从近期的一批货物中随机抽取 5 件作为样本进行测试。
    a. 假设该批货物有 1% 不合格，计算样本中的货物全部合格的概率。
    b. 假设该批货物有 1% 不合格，计算样本中恰有 1 件不合格品的概率。
    c. 如果该批货物有 1% 不合格的话，那么在样本中检测到 1 件或更多不合格品的概率是多少？
    d. 如果在样本中只发现 1 件不合格品的话，你对接受这批货物是否感到踏实？为什么？

59. 某个时期的失业率是 4.1%(*Barron's*, September 4, 2000)。假设随机地抽取了 100 名可就业人口。
    a. 其中失业人数的期望值是多少？
    b. 其中失业人数的方差和标准差是多少？

60. 在 1997 年 12 月，美国司法部提起了对微软公司的诉讼，控告该公司在它的 Windows 95 操作系统中捆绑了它的 IE 浏览器产品(*Fortune*, February 2, 1998)。社会公众对此事的意见可以根据是否认为微软公司构成垄断而分为两种。在《财富》杂志的民意调查中，有 41% 的调查对象赞同这样的说法："微软是一家垄断企业。"假设我们抽取 800 人作为样本。
    a. 其中你能够期望的赞同微软是垄断企业的人数是多少？

b. 其中相信微软是垄断企业的人数标准差是多少？

c. 其中不相信微软是垄断企业的人数标准差是多少？

61. 汽车随机且独立地到达汽车清洗站，在任意两个等长的时间段内，汽车到达的概率相同，其平均到达率为每小时 15 辆。在任意给定的 1 小时营业时间内，到达的汽车数大于等于 20 的概率是多少？

62. 一个新的自动生产线平均每天发生 1.5 起故障，由于与故障有关的成本因素，管理者想知道在 1 天内发生 3 起或更多故障的概率。假设故障是随机发生的，在任意两个等长的时间段内故障的发生概率相同，并且在一段时间内故障是否发生与在其他时段内的故障发生情况相互独立。问在 1 天内发生 3 起或更多故障的概率是多少？

63. 一个负责宾夕法尼亚州商业发展的地区官员关心小企业倒闭的数字。如果每月倒闭的小企业数平均为 10 的话，则在任意给定的 1 个月内恰有 4 家小企业倒闭的概率是多少？假设小企业在任何两个月的倒闭概率相同，并且在任何月份是否倒闭与在其他月份的倒闭情况相互独立。

64. 顾客随机且独立地到达银行。其在任意 1 分钟到达的概率与在其他 1 分钟时段内的到达概率相同。假设顾客的平均到达率为每分钟 3 人。

a. 1 分钟内恰有 3 人到达的概率是多少？

b. 1 分钟内至少有 3 人到达的概率是多少？

65. 大多数人都熟悉抽 5 张牌的扑克游戏。对于一副包括 4 张 A 的 52 张扑克，抽出的 5 张牌中有如下情况的概率是多少？

a. 1 对 A。

b. 恰有 1 张 A。

c. 没有 A。

d. 至少 1 张 A。

66. 根据 WTA 巡回赛和 ATP 巡回赛的数据，排名前 10 位的女网球选手中有 4 名使用 Wilson 球拍（*USA Today*, March 2, 1995）。假设前 10 名选手中有 2 位进入了锦标赛的决赛。

a. 恰有 1 人使用 Wilson 球拍的概率是多少？

b. 两人都使用 Wilson 球拍的概率是多少？

c. 无人使用 Wilson 球拍的概率是多少？

附录 5.1 利用 Minitab 计算离散随机概率

诸如 Minitab 这样的统计软件包提供了相对简单而有效的计算二项概率的方法。在本附录中，我们将演示如何确定 5.4 节马丁服装商店问题的二项概率。我们已知该二项概率的 $n = 10$ 和 $p = 0.30$。在开始 Minitab 程序之前，使用者必须先把所需的随机变量 x 的值输入工作表的一列中。我们把值 0，1，2，…，10 输入列 1(见图 5.5)，以生成整个二项概率分布。能够得到所需二项概

率的 Minitab 步骤如下：

步骤 1. 选择 **Calc** 下拉菜单；

步骤 2. 选择 **Probability Distributions** 选项；

步骤 3. 选择 **Binomial** 选项；

步骤 4. 当 Binomial Distribution 对话框出现时：

选择 **Probability**；

在 **Number of trials** 框输入 **10**；

在 **Probability of success** 框输入 0.3；

在 **Input column** 框输入 C1；

点击 **OK**。

Minitab 产生的二项概率输出结果如图 5.5 所示。

Minitab 还能够以类似的方法得到泊松概率。不同之处是在步骤 3 应选择 **Poisson** 选项，在步骤 4 应输入 **Mean** 的值而不是试验次数和成功概率。

附录 5.2　利用 Excel 计算离散概率分布

Excel 提供了能够计算本章介绍过的二项分布、泊松分布和超几何分布概率的函数。计算二项概率的 Excel 函数是 BINOMDIST，它有四个自变量：x(成功次数)，n(试验次数)，p(成功概率)和累积。如果我们需要求 x 次成功的概率，则第四个自变量（累积）取 FALSE；如果我们需要的是小于等于 x 次成功的累积概率，则第四个自变量取 TRUE。在这里我们演示一下对于 5.4 节的马丁服装商店问题，怎样计算得到 0—10 次成功的概率(参见图 5.5)。

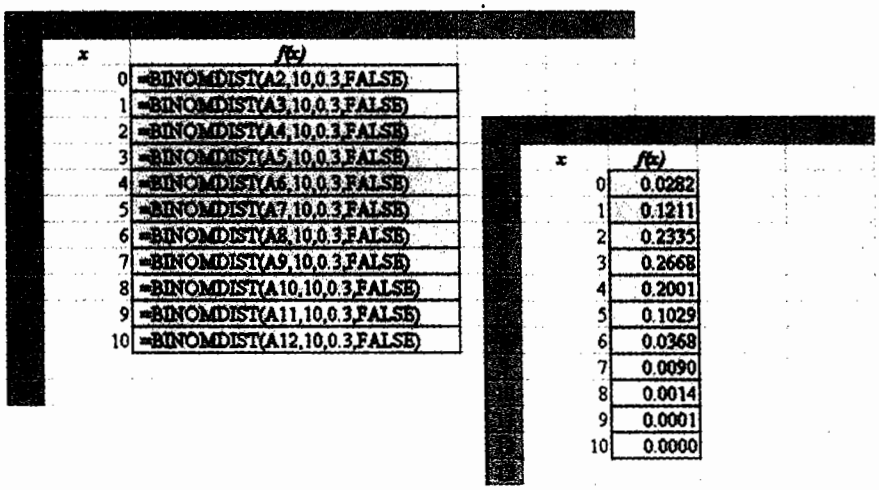
在我们描述如何创建工作表时，参考图 5.6。公式工作表处在屏幕显示的底面，而数值工作表处在屏幕显示的表面。通过下列步骤能够生成所需要的概率：

步骤 1. 使用 BINOMDIST 函数计算 $x=0$ 的概率向单元格 B2 输入下面的公式：

$$=\text{BINOMDIST(A2, 10, 0.3, FALSE)}$$

步骤 2. 把单元格 B2 中的公式复制到单元格 B3: B12。

图 5.6 表明的数值工作表得到的概率与图 5.5 得到的概率相同。用类似的方法也能够计算泊松和超几何概率，只是在计算时分别使用了 POISSON 和 HYPERGEOMDIST 函数。Excel 的函数向导能够帮助使用者向这些函数输入正确的自变量(见附录 2.2)。



x	f(x)
0	=BINOMDIST(A2,10,0.3,FALSE)
1	=BINOMDIST(A3,10,0.3,FALSE)
2	=BINOMDIST(A4,10,0.3,FALSE)
3	=BINOMDIST(A5,10,0.3,FALSE)
4	=BINOMDIST(A6,10,0.3,FALSE)
5	=BINOMDIST(A7,10,0.3,FALSE)
6	=BINOMDIST(A8,10,0.3,FALSE)
7	=BINOMDIST(A9,10,0.3,FALSE)
8	=BINOMDIST(A10,10,0.3,FALSE)
9	=BINOMDIST(A11,10,0.3,FALSE)
10	=BINOMDIST(A12,10,0.3,FALSE)

x	f(x)
0	0.0282
1	0.1211
2	0.2335
3	0.2668
4	0.2001
5	0.1029
6	0.0368
7	0.0090
8	0.0014
9	0.0001
10	0.0000

图 5.6　计算二项概率的 Excel 工作表

6 连续概率分布

统计实例:宝洁公司

6.1 均匀概率分布

作为概率度量的面积

6.2 正态概率分布

正态曲线

标准正态概率分布

计算任意正态概率分布的概率

Grear 轮胎公司问题

6.3 指数概率分布

计算指数分布的概率

泊松分布与指数分布的关系

统计实例

宝洁公司*

辛辛那提,俄亥俄州

宝洁公司(P&G)在全世界经营日用消费品业务。宝洁生产与销售种类繁多的产品，如去污剂、方便尿布、非处方药品、牙膏、肥皂、漱口剂和纸巾等产品。它比其他任何消费品公司拥有着更多种类的领先品牌。

作为一个在决策中引入统计方法的先行者，宝洁公司雇用了具有各种学术背景的人士，他们拥有如下专业：工程学、统计学、运筹学和商务。这些人提供的主要定量技术是：概率性决策和风险分析、预先模拟、质量改进和定量方法（如线性规划、回归分析）。

宝洁公司的工业化学部是脂肪乙醇的主要供应商，而脂肪乙醇是从椰子油等天然原料和石油衍生产品中提取的。该部门想知道扩建其脂肪乙醇生产设施的经济风险和机会，于是请了宝洁公司的概率决策和风险分析专家提供帮助。通过对问题建立模型并进行分析，专家们确定了获利能力的关键在于石油原料和椰子原料间的成本差异。虽然未来成本是未知的，但是借助于一些连续随机变量，分析人员能够对它进行描述。该问题需要的连续随机变量如下：

x = 每磅脂肪乙醇所需的椰子油价格；

和 y = 每磅脂肪乙醇所需的石油原料价格。

因为获利能力的关键就在于这两个随机变量的差，于是在分析中还使用了第三个随机变量：$d = x - y$。专家们被要求确定 x 和 y 的概率分布，以



宝洁公司众多知名产品中的一部分。© Joe Higgins/South-Western.

进一步使用这些概率分布的信息建立价差的概率分布。价差的连续概率分布表明价差小于等于 0.065 5 美元的概率为 0.90，价差小于等于 0.035 美元的概率为 0.50。另外，价差小于等于 0.004 5 美元的概率只有 0.10。**

工业化学部认为，是否能够量化原料价差的影响是达到一致意见的关键。以上分析取得的概率可用于对原料价差进行敏感性分析，使得分析人员有了充分的把握向管理当局提出建议。

使用连续随机变量及其概率分布有助于宝洁公司分析脂肪乙醇产品的经济风险。在本章，你将学习到连续随机变量和它们的概率分布，其中包括统计学中最重要的概率分布之一：正态分布。

---

* 作者感谢宝洁公司的 Joel Kahn 提供了该统计实例。

** 为了保护数据所有者的利益，对这里出现的价差已做了修改。

上一章我们讨论了离散随机变量和它们的概率分布，本章我们转向研究连续随机变量。我们重点讨论三种连续概率分布：均匀概率分布、正态概率分布和指数概率分布。

离散随机变量和连续随机变量间的基本区别在于如何计算它们的概率。对于离散随机变量，概率函数 $f(x)$给出了随机变量取某个特定值的概率；而对于连续随机变量，概率函数的对应者是**概率密度函数**(probability density function)，也记作 $f(x)$。它和概率函数的区别是概率密度函数不直接给出概率，而是通过在给定区间内 $f(x)$曲线下的面积，给出了连续随机变量 x 在该区间上取值的概率。所以当我们计算连续随机变量的概率时，就是计算随机变量在某个区间上取任意值的概率。

对连续随机变量定义的推论之一是随机变量取任意一个特定值的概率为零，这是因为 $f(x)$曲线在任何特定的点处，其下的面积为零。在 6.1 节，我们将举例说明具有均匀概率分布的连续随机变量的这些概念。

本章的大部分内容都用于描述和说明正态概率分布的应用。正态概率分布非常重要，它在统计推断中有着广泛的应用。本章的最后部分还将对指数概率分布进行讨论。

6.1　均匀概率分布

考虑一个表示从芝加哥到纽约的航班飞行时间的随机变量 x。假设飞行时间可以是 120—140 分钟区间内的任意值，由于随机变量 x 能够在该区间上取任何值，因此 x 是连续的而不是离散的随机变量。假定我们可以利用足够的实际飞行数据得出推断：在 120—140 分钟的区间内，单位为分钟的飞行时间数出现在任意 1 分钟时段内的概率与出现在其他任意 1 分钟时段内的概率相同。如果随机变量 x 在每个 1 分钟区间内具有相等的出现可能性，我们就称它具有**均匀概率分布**(uniform probability distribution)。对于飞行时间随机变量，我们定义均匀概率分布的概率密度函数为

$$f(x)=\begin{cases}1/20 & 120\leqslant x\leqslant 140\\ 0 & \text{其他}\end{cases}$$

只要概率与区间长度成比例，随机变量就是均匀分布。

图 6.1 是这个概率密度函数的图形。一般来说，通过下面的公式能够建立随机变量 x 的均匀概率密度函数:

均匀概率密度函数

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\leqslant x\leqslant b\\ 0 & \text{其他}\end{cases} \tag{6.1}$$

在飞行时间例子中，$a=120$，$b=140$。

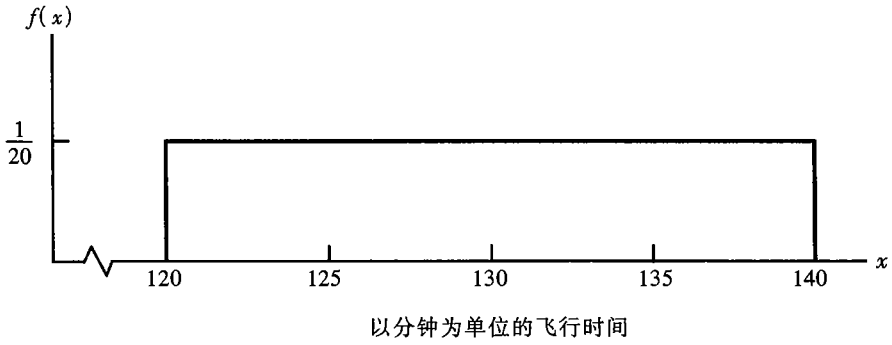


图 6.1　飞行时间的均均概率密度函数

正如在引言中所注意到的，对于连续随机变量，我们只是根据随机变量在某个特定区间内取值的可能性来考虑概率。在飞行时间例子中，一个合理的概率问题是：飞行时间介于 120—130 分钟之间的概率是多少？即 $P(120 \leqslant x \leqslant 130)$是多少？由于飞行时间必定处于 120—140 分钟之间，还由于概率在这个区间内是均匀分布的，我们可以放心地说 $P(120 \leqslant x \leqslant 130)=0.50$。在下面的内容中，我们将通过计算在 120—130 区间内 $f(x)$曲线下的面积，从而得到这个概率。

作为概率度量的面积

让我们对图 6.2 的曲线进行观察，考虑在 120—130 区间内 $f(x)$曲线下的面积。该面积区域是一个矩形，只需把长乘宽即可得出它的面积。由于区间的长度等于 $130-120=10$，而宽等于概率密度函数的值 $f(x)=1/20$，因此我们得到面积 = 长 × 宽 $=10 \times 1/20=0.50$。

你对于 $f(x)$曲线下的面积和概率观察到了什么？实际上它们是同一个东西！的确，所有的连续随机变量都是如此。一旦得出了概率密度函数 $f(x)$，通过计算在 x_1—x_2 区间内 $f(x)$曲线下的面积，我们就能够得到 x 取值介于较小的 x_1 和较大的 x_2 之间时的概率。

给定了飞行时间为均匀概率分布并且把面积作为概率，我们就可以回答任何有关的概率问

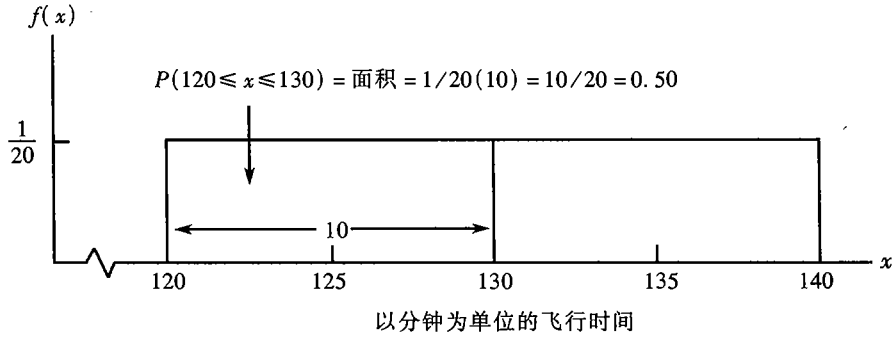


图 6.2　矩形的面积给出了飞行时间介于 120—130 分钟之间的概率

题。例如，飞行时间在 128—136 分钟之间的概率是多少？由于该区间的长是 136 - 128 = 8，宽等于概率密度函数值 1/20，因此我们看到 $P(128 \leqslant x \leqslant 136) = 8 \times 1/20 = 0.40$。

注意到 $P(120 \leqslant x \leqslant 140) = 20 \times 1/20 = 1$，也就是说，$f(x)$曲线下的总面积等于 1。这个性质适用于所有的连续概率分布，并且与离散概率函数的概率之和必定等于 1 的要求相对应。对于连续概率密度函数，我们还必须要求：对 x 的所有值，有 $f(x) \geqslant 0$。该项要求与离散概率函数$f(x) \geqslant 0$的要求相对应。

对连续随机变量和对应的离散随机变量的处理有两个主要区别：

1. 对于连续随机变量，不再讨论随机变量取某一特定值的概率。而是讨论随机变量在某一特定区间取值的概率。
2. 随机变量在从 x_1 到 x_2 的给定区间上取值的概率被定义为概率密度函数在 x_1 与 x_2 之间的图形面积。它暗示着连续随机变量取某一特定值的概率恰好为 0，因为 $f(x)$曲线在单点下的面积为 0。

为了看出任意单点的概率等于 0，参考图 6.2，并计算某单点的概率。举例来说，当 $x = 125$ 时，$P(x = 125) = P(125 \leqslant x \leqslant 125) = 0 \times 1/20 = 0$。

连续随机变量的方差和数学期望的计算过程与离散随机变量类似。不过，由于计算过程涉及积分计算，我们将把复杂的公式推导放在更高级的课程中。

对于本节介绍的均匀连续概率分布，其数学期望和方差公式分别为

$$E(x) = \frac{a+b}{2}$$

$$\mathrm{Var}(x) = \frac{(b-a)^2}{12}$$

在这两个公式里面，a 是随机变量的最小可能值，b 是随机变量的最大可能值。

把它们应用到芝加哥到纽约飞行时间的均匀概率分布中，可以得到

$$E(x) = \frac{120+140}{2} = 130$$

$$\mathrm{Var}(x) = \frac{(140-120)^2}{12} = 33.33$$

飞行时间的标准差取方差的正平方根即可，因此，$\sigma = 5.77$ 分钟。

评 注

1. 因为连续随机变量取任意的特定值时，其概率为0。所以，有 $P(a \leqslant x \leqslant b) = P(a < x < b)$。这个结果表明：随机变量在任一区间取值的概率，不管端点是否包括在内，都是相同的。
2. 为了更清楚地看出为什么概率密度函数的高度不是概率，考虑具有下列均匀概率分布的随机变量：

$$f(x) = \begin{cases} 2 & 0 \leqslant x \leqslant 0.5 \\ 0 & \text{其他} \end{cases}$$

当 x 在0到0.5之间时，概率密度函数 $f(x)$ 的高度为2。但是我们已经知道概率永远不可能大于1，因此 $f(x)$ 不能被解释为是 x 的概率。

练 习

方 法

1. 已知随机变量 x 在1.0—1.5之间服从均匀分布。



a. 作出它的概率密度函数曲线。
b. 计算 $P(x = 1.25)$。
c. 计算 $P(1.0 \leqslant x \leqslant 1.25)$。
d. 计算 $P(1.20 < x < 1.5)$。

2. 已知随机变量 x 在10—20之间服从均匀分布。
   a. 作出它的概率密度函数曲线。
   b. 计算 $P(x < 15)$。
   c. 计算 $P(12 \leqslant x \leqslant 18)$。
   d. 计算 $E(x)$。
   e. 计算 $\mathrm{Var}(x)$。

应 用

3. 三角航空公司宣称其从辛辛那提到坦帕的航班飞行时间为2小时零5分。假设我们认为实际的飞行时间服从2小时到2小时20分之间的均匀分布。
   a. 作出飞行时间的概率密度函数曲线。
   b. 飞行时间晚点不超过5分钟的概率是多少？

c. 飞行时间晚点不超过 10 分钟的概率是多少？

d. 飞行时间的期望值是多少？

4. 大部分计算机语言都有能够生成随机数的函数。Excel 应用程序使用 RAND 函数来生成 0 到 1 之间的随机数。如果我们以 x 表示生成的随机数，那么 x 就是一个具有如下概率密度函数的连续随机变量：



$$f(x)=\begin{cases}1 & 0\leqslant x\leqslant 1\\ 0 & \text{其他}\end{cases}$$

a. 画出它的概率密度函数。

b. 生成的随机数介于 0.25—0.75 之间的概率是多少？

c. 生成的随机数小于或等于 0.30 的概率是多少？

d. 生成的随机数大于 0.60 的概率是多少？

5. 在 LPGA 巡回赛中成绩最好的 60 名女高尔夫球手的击球距离在 238.9—261.2 码之间（*Golfweek*，December 6, 1997）。假设这些女选手的击球距离在该区间上服从均匀分布。

a. 给出击球距离的概率密度函数的数学表达式。

b. 一名女选手的击球距离小于 250 码的概率是多少？

c. 一名女选手的击球距离至少为 255 码的概率是多少？

d. 一名女选手的击球距离在 245 到 260 码之间的概率是多少？

e. 这些女选手中有多少人的击球距离至少为 250 码？

6. 一瓶液体清洁剂上的标签注明每瓶的容量为 12 盎司。生产线能够根据下面的概率密度函数均匀地填充瓶子：

$$f(x)=\begin{cases}8 & 11.975\leqslant x\leqslant 12.10\\ 0 & \text{其他}\end{cases}$$

a. 一个瓶子的填充量在 12—12.05 盎司之间的概率是多少？

b. 一个瓶子的填充量为 12.02 盎司或更多的概率是多少？

c. 该公司质量控制所要求的填充量为：在标签注明盎司数的上下 0.02 盎司范围内。问填充时达不到质量控制标准的概率是多少？

7. 假设我们有兴趣对一块土地投标，并且知道还有一位投标人*。卖方已经宣布超过 10 000 美元且最高的标价会被接受。假定竞争者的投标价格 x 是在 10 000—15 000 美元之间的均匀分布。

a. 假如你出价 12 000 美元，你中标的概率是多少？

b. 假如你出价 14 000 美元，你中标的概率是多少？

c. 为了使你得到土地的概率最大，你应出价多少？

d. 假设你知道某人愿意为这块土地向你支付 16 000 美元，你会考虑以小于(c)中的价格投标吗？为什么？

* 这一练习以西北大学的 Roger Myerson 教授向我们建议的问题为基础。

6.2 正态概率分布

最重要的描述连续随机变量的概率分布是**正态概率分布**(normal probability distribution)。正态概率分布有着广泛的实际应用，其中的随机变量可以是人的身高和体重、考试成绩、科学度量值和降雨量等等。它还普遍应用于统计推断方面，而这将是本书剩余部分的主要内容。在这些应用中，正态概率分布描述了从样本中得到的可能结果。

法国数学家 Abraham de Moivre 于 1733 年发表了 *The Doctrine of Chances*，他导出了正态概率分布。

正态曲线

正态概率分布的形状可以用图 6.3 的钟形曲线来表示。概率密度函数定义正态概率分布的钟形曲线如下：

正态概率密度函数

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \tag{6.2}$$

式中 μ = 均值；

σ = 标准差；

π = 3.141 59；

e = 2.718 28

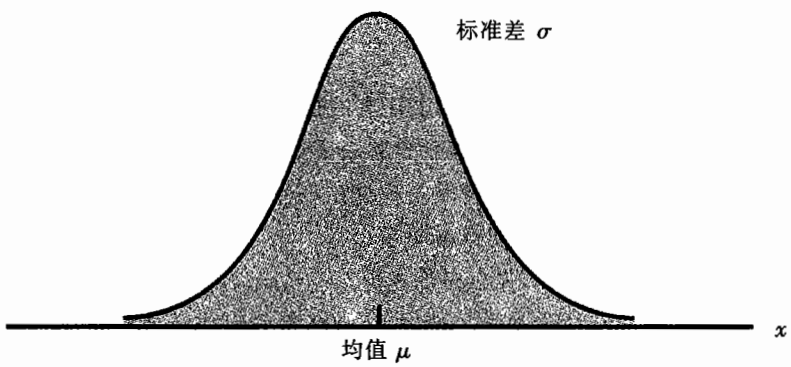
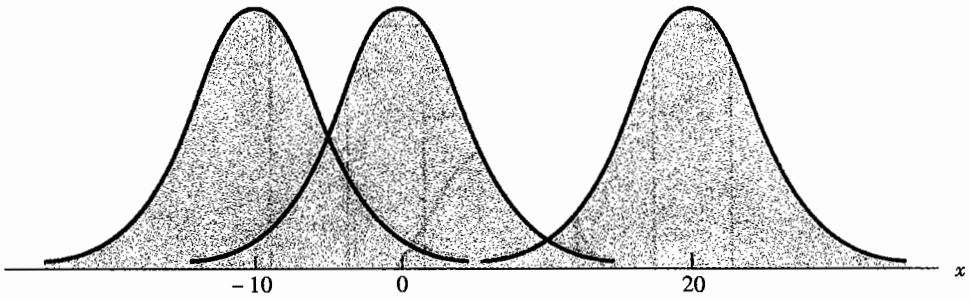


图 6.3 正态概率分布的钟形曲线

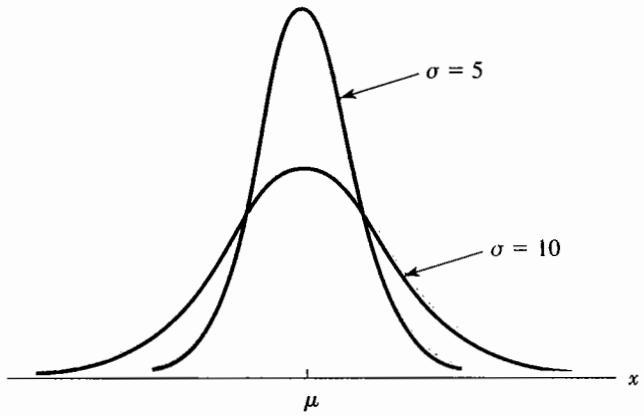
正态曲线有两个参数 μ 和 σ，它们确定了正态概率分布的位置和形状。

我们对正态概率分布的特征所做观察的结果如下：

1. 依靠均值 μ 和标准差 σ 可以区分不同的正态分布。
2. 正态曲线的最高点在均值位置，它同时也是正态分布的中位数和众数。
3. 正态分布的均值可以是任何数值：负数、零或者正数。三个标准差相同但均值分别为 -10，0 和 20 的正态分布曲线列示如下：



4. 正态概率分布是对称的，正态曲线在均值左边的形状与在均值右边的形状互为镜像。曲线的尾部向两个方向无限延伸，在理论上永远不会与横轴相交。
5. 标准差决定了正态曲线的宽度。更大的标准差导致了更宽、更扁的曲线形状，它表示数据有更大的变异性。两个均值相同但标准差不同的正态分布形状如下：



6. 正态随机变量的概率由正态曲线下的面积给出。正态概率分布曲线下的总面积为 1(对所有的连续概率分布都是如此)。因为分布是对称的，均值左边的曲线下总面积是 0.50，均值右边的曲线下总面积也是 0.50。

7. 随机变量在一些经常使用的区间内取值的百分比概率为

a. 正态随机变量有 68.26% 的值位于其均值加减 1 个标准差的范围内。
b. 正态随机变量有 95.44% 的值位于其均值加减 2 个标准差的范围内。
c. 正态随机变量有 99.72% 的值位于其均值加减 3 个标准差的范围内。

图 6.4 直观地显示了性质 a、b 和 c。

这些百分比是在 3.3 节介绍过的经验法则的基础。

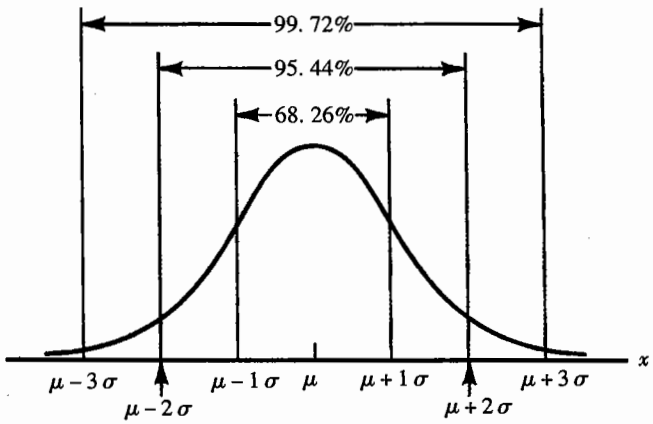


图 6.4 任意正态概率分布曲线下的面积

标准正态概率分布

如果随机变量服从均值为 0 且标准差为 1 的正态分布，则称它为具有**标准正态概率分布**(standard normal probability distribution)。通常用字母 z 表示这个特殊的正态随机变量。图 6.5 就是标准正态概率分布的曲线，它的外观与其他正态分布的图形大致相同，只是拥有 $\mu=0$ 和 $\sigma=1$ 的特殊性质。

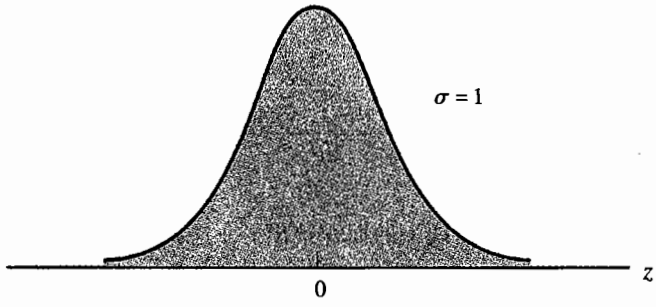


图 6.5 标准正态概率分布

与其他连续随机变量一样，任意正态概率分布的概率也是通过计算概率密度函数曲线下的面积得出的。因此，为了得到一个正态随机变量在某特定区间内的概率，我们必须计算在该区间内正态曲线下的面积。已经有人计算出了标准正态概率分布曲线下的面积，并且用表格的形式给出，这些表可用于计算正态随机变量的概率。表 6.1 就是这样一个表，附录 B 的表 1 也用于同样的目的。

对于正态概率密度函数，由于曲线的高是变化的，需要进行积分以计算代表概率的面积。

为了说明标准正态分布曲线下的面积表(表 6.1)是怎样计算概率的，让我们来考虑一些例子。稍后，我们还将看到如何利用同一张表计算任意正态分布的概率。首先，我们来学习怎样计算标准正态随机变量的 z 值在 0.00 和 1.00 之间取值的概率，即 $P(0.00 \leqslant z \leqslant 1.00)$。

给定一个 z 值，我们利用标准正态表查出对应的概率(曲线下面积)。

下图的阴影区域显示了这个概率值:

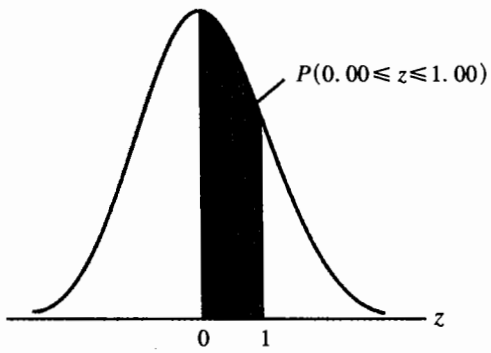


表 6.1 的各值给出了在均值 $z=0$ 和 z 的其他某个特定值之间标准正态曲线下的面积（见表顶部的图）。在本例中，我们感兴趣的面积处在 $z=0$ 和 $z=1.00$ 之间。因此，我们必须在表中找到与 $z=1.00$ 对应的值。首先，我们发现 1.0 在表的左列，然后在表上面的行中找到 0.00。通过在表的主体内查找，我们发现 1.0 行和 0.00 列交点处的值为 0.341 3。于是我们已经找到了所需要的概率：$P(0.00 \leqslant z \leqslant 1.00) = 0.341\,3$。表 6.1 的一部分说明了这些步骤。

利用同样的方法，我们还能够得到 $P(0.00 \leqslant z \leqslant 1.25)$。首先把位置定在 1.2 这一行，然后移动到与 0.05 列的交点，我们就可以得到 $P(0.00 \leqslant z \leqslant 1.25) = 0.394\,4$。再举一个例子，我们计算 z 在 $z=-1.00$ 和 $z=1.00$ 之间取值的概率，即 $P(-1.00 \leqslant z \leqslant 1.00)$。

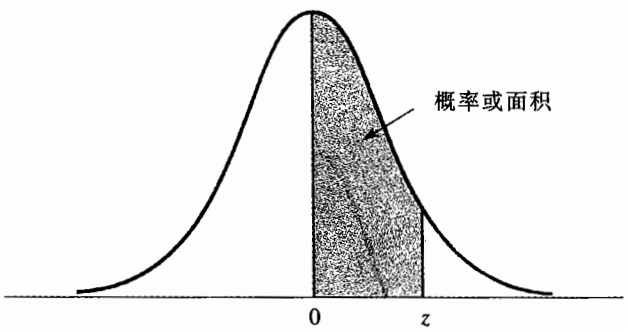
注意到我们已经在表 6.1 查到了 z 在 0.00 和 1.00 之间取值的概率是 0.341 3，并且我们知道正态概率分布是**对称的**。所以，z 在 0.00 和 −1.00 之间取值的概率与 z 在 0.00 和 +1.00 之间取值的概率相等。于是，z 值在 −1.00 和 +1.00 之间的概率是

$$P(-1.00 \leqslant z \leqslant 0.00) + P(0.00 \leqslant z \leqslant 1.00) = 0.3413 + 0.3413 = 0.6826$$

z	0.00	0.01	0.02
…			
0.9	0.3159	0.3186	0.3212
1.0	0.3413	0.3438	0.3461
1.1	0.3643	0.3665	0.3686
1.2	0.3849	0.3869	0.3888
…			

$P(0.00 \leqslant z \leqslant 1.00)$

表 6.1 标准正态分布的概率或面积

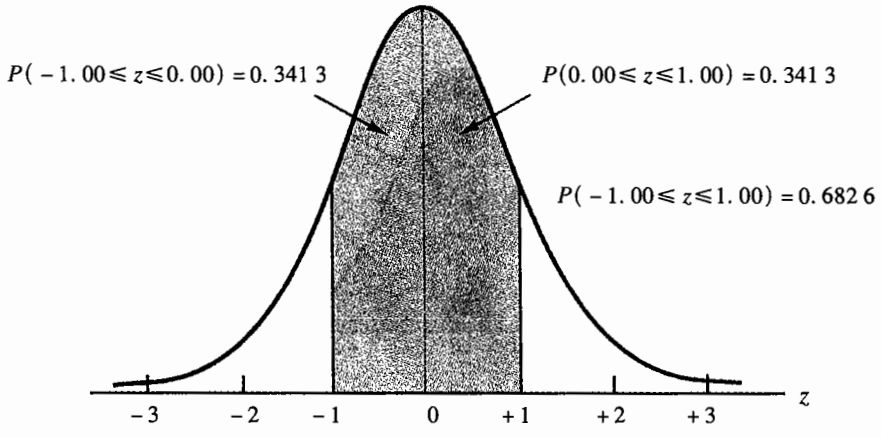


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015

（续表）

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.3	.403 2	.404 9	.406 6	.408 2	.409 9	.411 5	.413 1	.414 7	.416 2	.417 7
1.4	.419 2	.420 7	.422 2	.423 6	.425 1	.426 5	.427 9	.429 2	.430 6	.431 9
1.5	.433 2	.434 5	.435 7	.437 0	.438 2	.439 4	.440 6	.441 8	.442 9	.444 1
1.6	.445 2	.446 3	.447 4	.448 4	.449 5	.450 5	.451 5	.452 5	.453 5	.454 5
1.7	.455 4	.456 4	.457 3	.458 2	.459 1	.459 9	.460 8	.461 6	.462 5	.463 3
1.8	.464 1	.464 9	.465 6	.466 4	.467 1	.467 8	.468 6	.469 3	.469 9	.470 6
1.9	.471 3	.471 9	.472 6	.473 2	.473 8	.474 4	.475 0	.475 6	.476 1	.476 7
2.0	.477 2	.477 8	.478 3	.478 8	.479 3	.479 8	.480 3	.480 8	.481 2	.481 7
2.1	.482 1	.482 6	.483 0	.483 4	.483 8	.484 2	.484 6	.485 0	.485 4	.485 7
2.2	.486 1	.486 4	.486 8	.487 1	.487 5	.487 8	.488 1	.488 4	.488 7	.489 0
2.3	.489 3	.489 6	.489 8	.490 1	.490 4	.490 6	.490 9	.491 1	.491 3	.491 6
2.4	.491 8	.492 0	.492 2	.492 5	.492 7	.492 9	.493 1	.493 2	.493 4	.493 6
2.5	.493 8	.494 0	.494 1	.494 3	.494 5	.494 6	.494 8	.494 9	.495 1	.495 2
2.6	.495 3	.495 5	.495 6	.495 7	.495 9	.496 0	.496 1	.496 2	.496 3	.496 4
2.7	.496 5	.496 6	.496 7	.496 8	.496 9	.497 0	.497 1	.497 2	.497 3	.497 4
2.8	.497 4	.497 5	.497 6	.497 7	.497 7	.497 8	.497 9	.497 9	.498 0	.498 1
2.9	.498 1	.498 2	.498 2	.498 3	.498 4	.498 4	.498 5	.498 5	.498 6	.498 6
3.0	.498 6	.498 7	.498 7	.498 8	.498 8	.498 9	.498 9	.498 9	.499 0	.499 0

这个概率用图形表示如下：

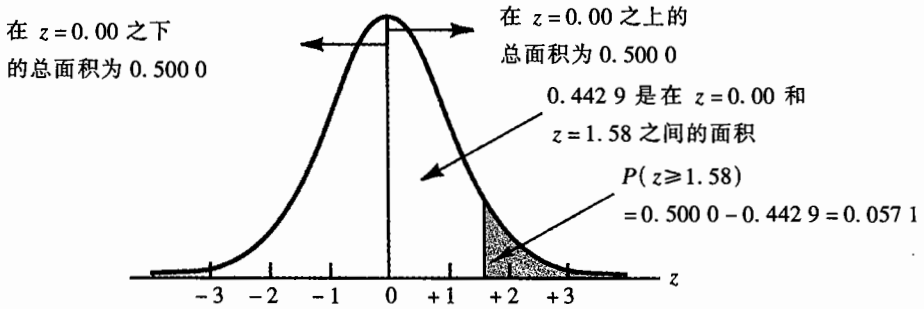


用类似的方式，我们可以用表 6.1 的值得出 z 值处在 −2.00 和 +2.00 之间的概率是 0.477 2 + 0.477 2 = 0.954 4，z 值处在 −3.00 和 +3.00 之间的概率是 0.498 6 + 0.498 6 = 0.997 2。因为我

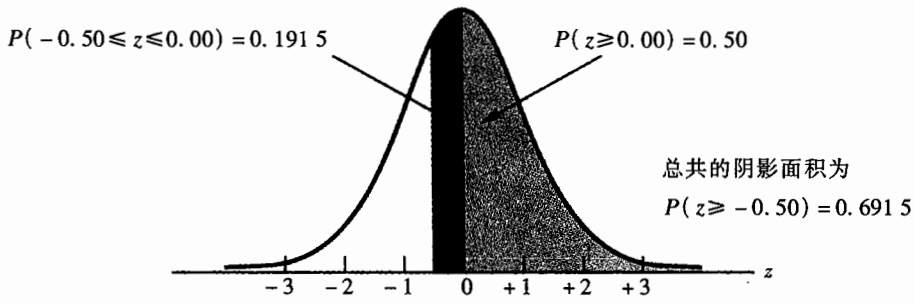
们知道任何连续随机变量曲线下的总概率或总面积必定等于 1.000 0，概率 0.997 2 说明 z 值几乎总是处在 -3.00 和 +3.00 之间。

这些概率计算是前面第 240 页观察结果 7 的计算基础。

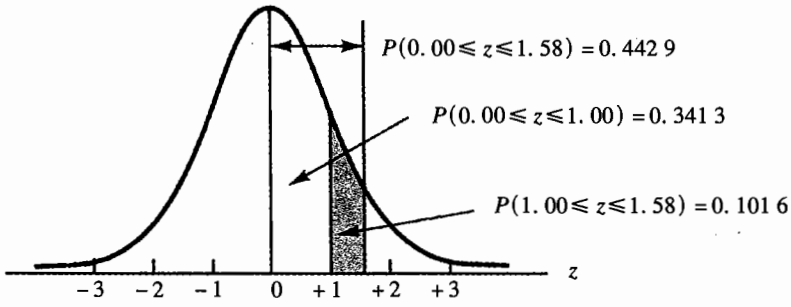
下面，我们来计算 z 值至少为 1.58 的概率，即 $P(z \geq 1.58)$。首先，我们在表 6.1 中的 $z=1.5$ 行和 0.08 列相交处找到 $P(0.00 \leq z \leq 1.58)=0.442\,9$。因为正态分布是对称的，且曲线下的总面积等于 1，所以我们知道曲线下面有 50% 的面积其 z 值大于均值（如 $z=0$），有 50% 的面积其 z 值小于均值。如果 0.442 9 是均值和 $z=1.58$ 之间的面积，那么对应于 $z \geq 1.58$ 的概率或面积必定为 0.500 0 - 0.442 9 = 0.057 1。这个概率显示在下图中：



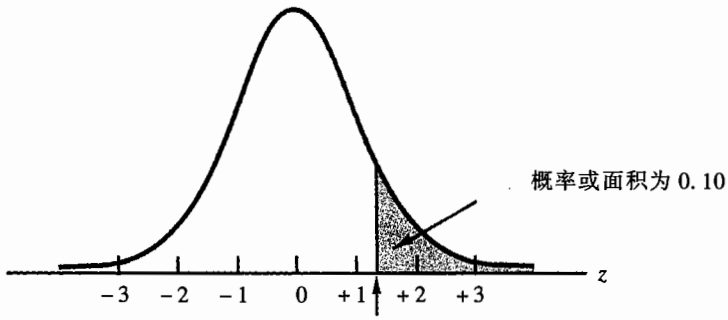
再举一个例子，考虑随机变量在 -0.50 以上取值的概率，即 $P(z \geq -0.50)$。为了得到所需的结果，我们注意到要找的概率能够被表达为两个概率之和的形式：$P(z \geq -0.50) = P(-0.50 \leq z \leq 0.00) + P(z \geq 0.00)$。我们已经知道 $P(z \geq 0.00)=0.50$。我们还知道：因为正态分布是对称的，所以 $P(-0.50 \leq z \leq 0.00) = P(0.00 \leq z \leq 0.50)$。参考表 6.1，我们得到 $P(0.00 \leq z \leq 0.50)=0.191\,5$，，因此 $P(z \geq -0.50)=0.191\,5+0.500\,0=0.691\,5$。下图显示了这个概率：



接下来，我们计算 z 值处在 1.00 和 1.58 之间的概率，即 $P(1.00 \leq z \leq 1.58)$。从上面的例子中，我们已知 z 值在 0.00 和 1.00 之间的概率是 0.341 3，并且 z 值在 0.00 和 1.58 之间的概率是 0.442 9。于是，z 值介于 1.00 和 1.58 之间的概率为 0.442 9 - 0.341 3 = 0.101 6，即 $P(1.00 \leq z \leq 1.58)=0.101\,6$。下图显示了这种情况：



最后，让我们来找到概率为 0.10 的那个较大 z 值。下图表现了这种情况：



z 值是多少?

这个问题是前面所举例子的反问题。在前面，我们设定了感兴趣的 z 值，然后找到对应的概率或面积。而本例则先给出了概率，要求我们找到对应的 z 值。为做到这一点，我们可以用不同的方式使用标准正态概率分布表(见表 6.1)。

给定概率，我们能够以一种相反的方式使用标准正态表，得到对应的 z 值。

我们知道表 6.1 给出了介于均值和某特定 z 值之间的标准正态曲线下面积。我们已经拥有的信息是已知上侧的曲线下面积为 0.10，于是，我们必须确定在均值和有关 z 值之间有多少面积。因为我们知道 0.500 0 的面积处在大于均值的位置，那么 0.500 0 − 0.100 0 = 0.400 0 必定是介于均值和所需 z 值之间的面积。观察表的主体，我们发现概率值 0.399 7 最接近于 0.400 0。表 6.1 的一部分提供了下面的结果：

z	0.06	0.07	0.08	0.09
…				
1.0	0.355 4	0.357 7	0.359 9	0.362 1
1.1	0.377 0	0.379 0	0.381 0	0.383 0
1.2	0.396 2	0.398 0	0.399 7	0.401 5
1.3	0.413 1	0.414 7	0.416 2	0.417 7
1.4	0.427 9	0.429 2	0.430 6	0.431 9
…				

表主体中最接近 0.400 0* 的面积值

*编者注：原文为 0.900 0，疑误。

读出表的最左列和顶行的 z 值，我们发现对应的 z 值为 1.28。因此，大约有 0.400 0（实际为 0.399 7）的面积介于均值和 $z=1.28$ 之间。*

就最初的问题而言，概率大约为 0.10，z 值比 1.28 稍大。该例子说明标准正态概率分布面积表还能够用来计算与标准正态随机变量 z 对应的概率值。在实际应用中，这两种类型的问题都可能被遇到，第一类问题设定了 z 的值，要求我们查表确定对应的面积或概率；第二类问题则给出了概率值，要求我们查表确定对应的 z 值。因此，我们需要灵活地运用标准正态概率表以回答所要求的概率问题。在大多数情况下，画出标准正态概率分布的草图并用适当的阴影表示概率或面积，可以帮助我们形象地了解问题并确定正确的答案。

计算任意正态概率分布的概率

我们之所以对标准正态分布进行如此广泛的讨论，原因是所有的正态分布概率都需要通过标准正态分布来计算。也就是说，当我们面对一个具有任意均值 μ 和任意标准差 σ 的正态分布时，首先要把它转换为标准正态分布，以回答分布的有关概率问题。然后我们利用表 6.1 和适当的 z 值，能够找到所求的概率。能够把具有均值 μ 和标准差 σ 的任意正态随机变量 x 转换为标准正态随机变量 z 的公式如下：

标准正态随机变量的转换公式与我们在第 3 章介绍过的计算数据集 z 分数的公式类似。

标准正态分布的转换公式

$$z=\frac{x-\mu}{\sigma} \tag{6.3}$$

当 x 的值等于它的均值 μ 时会导致 $z=(\mu-\mu)/\sigma=0$，因此，我们看到当 x 的值等于均值 μ 时对应的 z 值处在均值 0 处。现在假定 x 大于均值 1 个标准差，即 $x=\mu+\sigma$。应用公式(6.3)，我们看到对应的 z 值为 $z=[(\mu+\sigma)-\mu]/\sigma=6/6=1$。因此，大于均值 1 个标准差的 x 值对应于 $z=1$。换句话说，我们能够把 z 值解释为正态随机变量 x 距离均值 μ 的标准差个数。

下面我们来说明如何利用该转换计算任意正态分布的概率，假设我们有一个正态分布，$\mu=10$ 且 $\sigma=2$，问随机变量 x 位于 10 和 14 之间的概率是多少？应用公式(6.3)，我们看到当 $x=10$，$z=(x-\mu)/\sigma=(10-10)/2=0$；当 $x=14$，$z=(14-10)/2=4/2=2$。因此，x 介于 10—14 之间的概率问题等价于标准正态分布的 z 介于 0—2 之间的概率问题。换句话说，我们所要寻找的概率就是随机变量 x 介于均值和均值以上 2 个标准差区间内的概率。对于 $z=2.00$，运用表 6.1，我们得到概率为 0.477 2。于是 x 介于 10—14 的概率也就是 0.477 2。

---

* 我们能够使用内插法在表中得到对应于 0.400 0 面积的具有更好近似性的 z 值，这样做可以得到有更多小数位的更精确 z 值 1.282。不过，在大多数实际情况下，取得充分的精确性只需简单地利用最接近于所求概率的表内值获得的。

Grear 轮胎公司问题

现在我们转到正态概率分布的一个应用例子上来。假定 Grear 公司刚刚开发了一种新的钢丝子午线轮胎，并通过一家全国连锁的折扣商店出售。因为该轮胎是一种新产品，Grear 公司的经理们认为是否保证一定的行驶里程数将是该产品能否被顾客接受的重要因素。在制定这种轮胎的里程质保政策之前，经理们需要知道轮胎行驶里程数的概率信息。

根据对这种轮胎的实际路面测试，公司的工程师小组估计它们的平均行驶里程为 $\mu = 36\,500$ 英里，里程数的标准差为 $\sigma = 5\,000$。另外，收集到的数据显示，行驶里程数符合正态分布应该是一个合理的假设。问题是有多大百分比的轮胎能够行驶超过 40 000 英里？换句话说，轮胎行驶里程大于 40 000 英里的概率是多少？通过计算图 6.6 阴影区域的面积，我们能够回答这个问题。

当 $x = 40\,000$ 时，我们有

$$z = \frac{x - \mu}{\sigma} = \frac{40\,000 - 36\,500}{5\,000} = \frac{3\,500}{5\,000} = 0.70$$

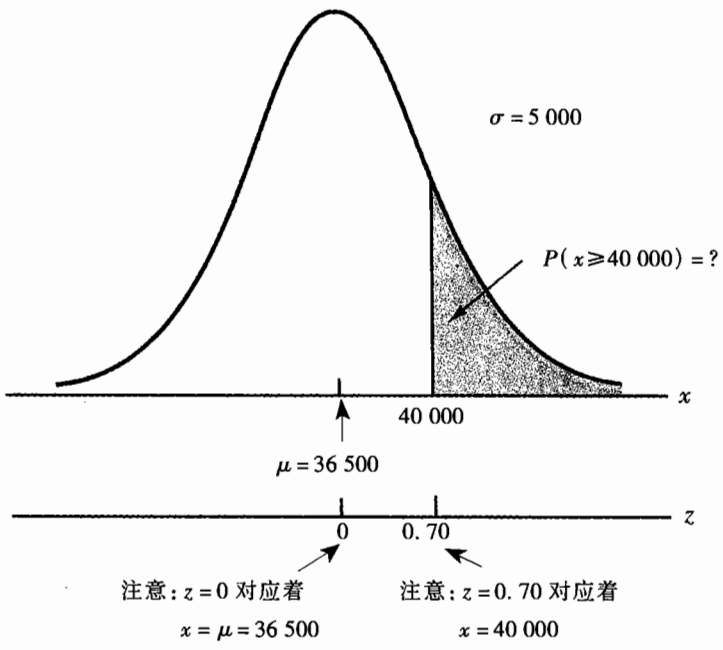


图 6.6　Grear 轮胎公司的行驶里程分布

现在参考图 6.6 的底部，我们看到在 Grear 轮胎正态分布中的 $x = 40\,000$ 对应着标准正态分布的 $z = 0.70$。利用表 6.1，我们看到均值和 $z = 0.70$ 之间的面积为 0.258 0。再次参考图 6.6，轮胎的正态分布中 $x = 36\,500$ 和 $x = 40\,000$ 之间的面积与上面的标准正态分布面积相同。因此，0.500 0 − 0.258 0 = 0.242 0 是 x 大于 40 000 的概率。于是我们能够得出结论：大约有 24.2% 的轮胎行驶里程会超过 40 000 英里。

现在我们假设公司正在考虑一项质量保证政策，如果初始购买的轮胎没有能够使用到保证的里程数，公司将以折扣价格为客户更换轮胎。如果公司希望符合折扣条件的轮胎不超过 10%，则保证的里程应为多少？这个问题在图 6.7 进行了说明。

根据图 6.7,处在均值和未知保证里程数之间的面积必须为 40%。我们在表 6.1 查找 0.400 0，看到该面积大约在均值与小于均值 1.28 个标准差处之间，即 $z=-1.28$ 是对应于公司在正态分布中保证里程数的标准正态随机变量值。为了得到对应于 $z=-1.28$ 的里程数 x，我们有

$$z=\frac{x-\mu}{\sigma}=-1.28$$

$$x-\mu=-1.28\sigma$$

$$x=\mu-1.28\sigma$$

由于 $\mu=36\,500$ 和 $\sigma=5\,000$

$$x=36\,500-1.28\times5\,000=30\,100$$

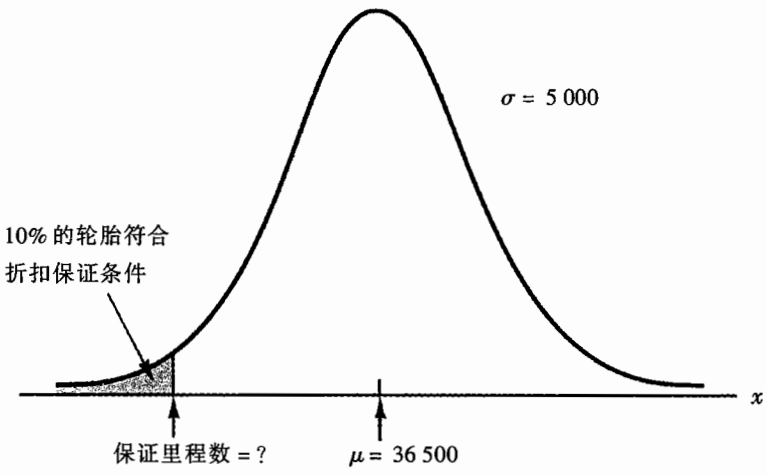


图 6.7　Grear 公司的折扣保证

当保证设定在 30 000 英里时，符合保证条件的实际百分比为 9.68%。

因此，30 100 英里的质量保证将满足只有大约 10% 的轮胎需要折价更换的要求。也许，根据这一信息，公司将把它的轮胎里程保证设在 30 000 英里。

我们再一次看到了概率分布在提供决策信息方面所起的重要作用。也就是说，只要对某一应用问题建立起了概率分布模型，就能够迅速而方便地取得有关问题的概率信息。虽然依据概率并不能直接提出决策建议，但它提供了可以帮助决策者更好地理解有关问题的风险和不确定性的有用信息。最终，这一信息能够帮助决策者制定出更好的决策。

练　习

方　法

8. 以图 6.4 为指导，画出均值为 $\mu = 100$ 且标准差为 $\sigma = 10$ 的正态随机变量 x 的曲线图。在横轴上标明数值 70、80、90、100、110、120 和 130。

9. 某随机变量服从正态分布，均值 $\mu = 50$，标准差 $\sigma = 5$。
   a. 画出概率密度函数的正态曲线。在横轴上标明数值 35、40、45、50、55、60 和 65，图 6.4 显示了正态曲线在均值前后 3 个标准差处(即 35 和 65 处)几乎与横轴相交。
   b. 随机变量取值介于 45 和 55 之间的概率是多少？
   c. 随机变量取值介于 40 和 60 的概率是多少？

10. 画出标准正态分布图，在横轴上标明数值 −3、−2、−1、0、1、2 和 3。然后利用标准正态分布概率表计算下列概率。
   a. $P(0 \leq z \leq 1)$。
   b. $P(0 \leq z \leq 1.5)$。
   c. $P(0 < z < 2)$。
   d. $P(0 < z < 2.5)$。

11. 给定 z 是标准正态随机变量，计算下列概率。
   a. $P(-1 \leq z \leq 0)$。
   b. $P(-1.5 \leq z \leq 0)$。
   c. $P(-2 < z < 0)$。
   d. $P(-2.5 \leq z \leq 0)$。
   e. $P(-3 \leq z \leq 0)$。

12. 给定 z 是标准正态随机变量，计算下列概率。
   a. $P(0 \leq z \leq 0.83)$。
   b. $P(-1.57 \leq z \leq 0)$。
   c. $P(z > 44)$。
   d. $P(z \geq -0.23)$。
   e. $P(z < 1.20)$。
   f. $P(z \leq -0.71)$。

13. 给定 z 是标准正态随机变量，计算下列概率。



   a. $P(-1.98 \leq z \leq 0.49)$。

b. $P(0.52 \leqslant z \leqslant 1.22)$。

c. $P(-1.75 \leqslant z \leqslant -1.04)$。

14. 给定 z 是标准正态随机变量，计算下面各种情况下的 z 值。

a. 0 和 z 之间的面积为 0.475 0。

b. 0 和 z 之间的面积为 0.229 1。

c. z 右侧的面积为 0.131 4。

d. z 左侧的面积为 0.670 0。

15. 给定 z 是标准正态随机变量，计算下面各种情况下的 z 值。



a. z 左侧的面积为 0.211 9。

b. $-z$ 和 z 之间的面积为 0.903 0。

c. $-z$ 和 z 之间的面积为 0.205 2。

d. z 左侧的面积为 0.994 8。

e. z 右侧的面积为 0.691 5。

16. 给定 z 是标准正态随机变量，计算下面各种情况下的 z 值。

a. z 右侧的面积为 0.01。

b. z 右侧的面积为 0.025。

c. z 右侧的面积为 0.05。

d. z 右侧的面积为 0.10。

应　用

17. 美国成年男性的平均身高为 5 英尺 9 英寸（*Astounding Averages*，1995）。假设标准差是 3 英寸，回答下列问题。

a. 成年男性高于 6 英尺的概率是多少？

b. 成年男性低于 5 英尺的概率是多少？

c. 成年男性身高在 5 英尺 6 英寸和 5 英尺 10 英寸之间的概率是多少？

d. 成年男性不高于 6 英尺的概率是多少？

18. 订户阅读一份《华尔街日报》的平均时间是 49 分钟（*The Wall Street Journal* Subscriber Study，1996）。假定标准差是 16 分钟，且阅读时间符合正态分布。



a. 订户至少花 1 个小时读报的概率是多少？

b. 订户读报时间不超过 30 分钟的概率是多少？

c. 对于 10% 的读报时间最长的人士，他们花费了多少时间？

19. 得克萨斯州达拉斯市 4 月的平均降雨量为 3.5 英寸（*The World Almanac*，2000）。假定降雨量服从正态分布，标准差为 0.8 英寸。

a. 4 月降雨量超过 5 英寸的概率是多少？

b. 4 月降雨量不足 3 英寸的概率是多少？

c. 如果某月的降雨量处于最高的 10% 概率范围内，则将该月归类为非常潮湿。要把某个 4 月份归类为非常潮湿的话，降雨量必须达到多少？

20. 根据调查，*The Wall Street Journal Interactive Edition* 的订户每周平均有 27 小时在工作中使用计算机（*WSJ. com* Subscriber Study, 1999）。假定正态分布适用于该问题，标准差为 8 小时。

a. 随机选择到的订户每周使用计算机工作少于 11 小时的概率是多少？

b. 多少百分比的订户每周使用计算机工作超过 40 小时？

c. 如果某人的使用时间在最高的 20% 概率范围内，则被归为频繁使用者行列。要把一个订户归为频繁使用者的话，他每周必须使用计算机多少小时？

21. 一个人必须在 IQ 测试中得分达到最高的 2% 范围内，才有资格加入 Mensa，即国际高智商协会（*US Airways Attache*, September , 2000）。如果人们的 IQ 得分服从正态分布，均值为 100，标准差为 15，要取得加入协会的资格必须得到多少分？

22. 根据劳工统计署的数据，1998 年美国产业工人的平均周工资为 441.84 美元（*The World Almanac*, 2000）。假设得到的数据表明他们的工资服从正态分布，其标准差为 90 美元。

a. 工人工资在 400—500 美元之间的概率是多少？

b. 工人收入如果处于最高的 20% 概率范围内的话，他必须有多少收入？

c. 对于一个随机选择的工人，他的每周收入少于 250 美元的概率是多少？

23. 学院某门课程的期末考试所需时间服从正态分布，其均值为 80 分钟，标准差为 10 分钟。回答下列问题：

a. 考生在 1 个小时或更短时间内完成期末考试的概率是多少？

b. 一名考生的完成时间超过 60 分钟但少于 75 分钟的概率是多少？

c. 假设一个班级有 60 名学生，考试时间为 90 分钟。你预期不能在规定时间内完成考试的学生有多少名？

24. 在 8 月和 9 月的 12 天当中，纽约股票交易所的股票日成交量（百万股）如下所示（*Barron's*, August 7, 2000, September 4, 2000 及 September 11, 2000）：

917	983	1 046
944	723	783
813	1 057	766
836	992	973

成交量的概率分布接近正态概率分布。

a. 计算日成交量的均值和标准差，作为总体均值和标准差的估计值。

b. 某日成交量小于 8 亿股的概率是多少？

c. 日成交量超过 10 亿股的概率是多少？

d. 如果交易所想在 5% 的成交量最大交易日发布新闻，那么引发新闻的成交量是多少？

25. 1998 年主要的联盟棒球比赛的平均票价为 11.98 美元(*USA Today*, November 11, 1998)。加上食品、停车与纪念品支出，一个四口之家观看比赛的平均花费大约是 110.00 美元。假设正态分布适用于该问题，且标准差为 20.00 美元。

a. 全家花费超过 100.00 美元的概率是多少？

b. 全家花费 90.00 美元或更低的概率是多少？

c. 全家花费在 80.00—130.00 美元的概率是多少？

6.3 指数概率分布

一种在描述完成任务所花费的时间方面十分有用的连续概率分布是**指数概率分布**(exponential probability distribution)。指数随机变量能够用来描述诸如汽车清洗站的车辆到达间隔时间、装运一辆卡车所需时间、公路上严重缺陷之间的距离等等问题。指数概率密度函数如下：

指数概率密度函数

$$f(x)=\frac{1}{\mu}e^{-x/\mu} \qquad x\geqslant 0,\ \mu>0 \tag{6.4}$$

作为指数概率分布的一个例子，假设 x = 在 Schips 装运码头装运一辆卡车所花费的时间，它服从指数分布。如果平均装车时间为 15 分钟（$\mu=15$），则恰当的概率密度函数为

$$f(x)=\frac{1}{15}e^{-x/15}$$

图 6.8 是这个概率密度函数的图形表示。

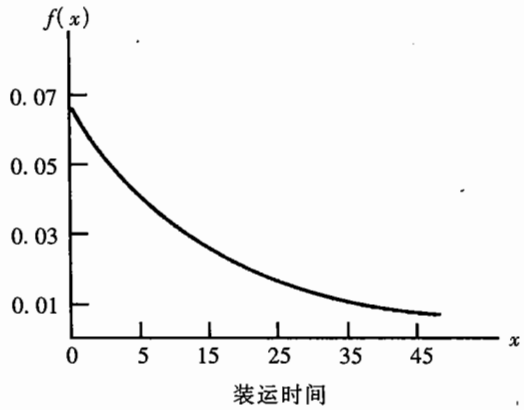


图 6.8　Schips 装运码头例子的指数概率分布

计算指数分布的概率

和任何连续概率分布一样，与某一区间相对应的曲线下面积给出了随机变量在该区间取值的

概率。在 Schips 装运码头例子中，装运一辆卡车花费 6 分钟或更短时间($x \leqslant 6$)的概率被规定为图 6.8 中从 $x=0$ 到 $x=6$ 区间内曲线下的面积。类似地，装运一辆卡车花费 18 分钟或更短时间 $x \leqslant 18$ 的概率是从 $x=0$ 到 $x=18$ 区间内曲线下的面积。而装运一辆卡车费时在 6 分钟到 18 分钟之间($6 \leqslant x \leqslant 18$)的概率就是从 $x=6$ 到 $x=18$ 区间内曲线下的面积。

为了计算这些问题中的指数概率，我们使用下面的公式，它给出了指数随机变量取值小于或等于 x 的某个特定值（记作 x_0）的累积概率。

指数分布：累积概率

$$P(x \leqslant x_0) = 1 - e^{-x_0/\mu} \tag{6.5}$$

对于 Schips 装运码头例子，x = 装运时间，公式(6.5)能够被表示为

$$P(x \leqslant x_0) = 1 - e^{-x_0/15}$$

于是，装运一辆卡车花费 6 分钟以内时间的概率 $P(x \leqslant 6)$为

$$P(x \leqslant 6) = 1 - e^{-6/15} = 0.329\,7$$

图 6.9 显示了装车时间为 6 分钟以内的面积或概率。而装车时间为 18 分钟以内的概率 $P(x \leqslant 18)$为

$$P(x \leqslant 18) = 1 - e^{-18/15} = 0.698\,8$$

因此，装车时间介于 6 到 18 分钟的概率等于 0.698 8 − 0.329 7 = 0.369 1。任意其他区间的概率也可以通过同样的方法进行计算。

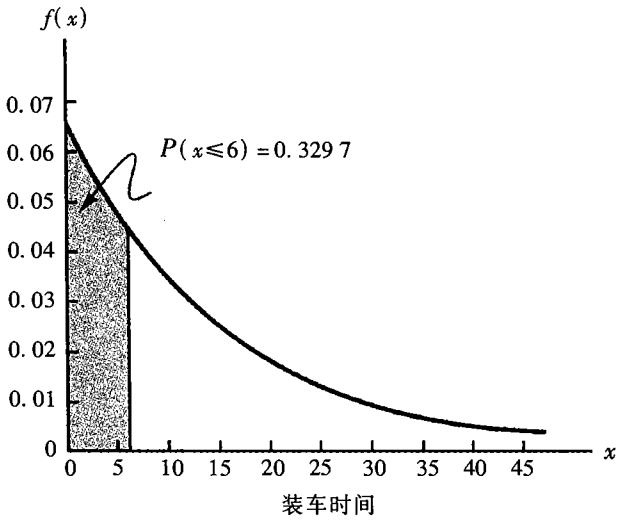


图 6.9　装车时间为 6 分钟以内的概率

泊松分布与指数分布的关系

在 5.5 节我们介绍过泊松分布是一种离散概率分布，它往往用于确定在一个特定的时间或空

间段内事件的发生次数。我们已知泊松概率函数为

$$f(x)=\frac{\mu^x e^{-\mu}}{x!}$$

式中 μ 为在一个区间内事件发生次数的期望值。

连续的指数概率分布与离散的泊松分布存在关系。如果说泊松分布给出了每个区间事件发生次数的恰当描述，那么指数分布则描述了事件的间隔区间长度。

为了说明这种关系，假设在 1 小时内到达清洗站的汽车数可以用泊松概率分布表示，其均值为每小时 10 辆汽车。于是泊松概率函数给出了每小时到达 x 辆汽车的概率为

$$f(x)=\frac{10^x e^{-10}}{x!}$$

因为平均到达数是每小时 10 辆汽车，则到达车辆的平均间隔时间为

$$\frac{1\text{小时}}{10\text{辆}}=0.1\text{ 小时/辆}$$

于是，描述到达车辆间隔平均时间的指数分布有均值 $\mu=0.1$ 小时/辆，故恰当的指数概率密度函数为

$$f(x)=\frac{1}{0.1}e^{-x/0.1}=10e^{-10x}$$

练 习

方 法

26. 考虑下面的指数概率密度函数：$f(x)=\frac{1}{8}e^{-x/8}$，$x\geqslant 0$。

a. 计算 $P(x\leqslant 6)$。

b. 计算 $P(x\leqslant 4)$。

c. 计算 $P(x\geqslant 6)$。

d. 计算 $P(4\leqslant x\leqslant 6)$。

27. 考虑下面的指数概率密度函数：$f(x)=\frac{1}{3}e^{-x/3}$，$x\geqslant 0$。

自测题

a. 给出 $P(x\leqslant x_0)$ 的公式。

b. 计算 $P(x\leqslant 2)$。

c. 计算 $P(x\geqslant 3)$。

d. 计算 $P(x\leqslant 5)$。

e. 计算 $P(2\leqslant x\leqslant 5)$。

应　用

28.《互联网杂志》监督互联网服务提供商(ISP)并对它们的业务表现提供统计数据。已知在免费的 ISP 那里下载一张欧洲网页的平均时间大约是 20 秒(*Internet Magazine*, January, 2000)。假定下载网页的时间服从指数分布。

a. 不到 10 秒下载一张网页的概率是多少?

b. 下载一张网页超过 30 秒的概率是多少?

c. 下载一张网页花费时间在 10 到 30 秒之间的概率是多少?

29. 车辆到达某特定集合点的时间间隔服从均值为 12 秒的指数分布。



a. 画出该指数概率分布的草图。

b. 车辆到达时间间隔在 12 秒以内的概率是多少?

c. 车辆到达时间间隔在 6 秒以内的概率是多少?

d. 车辆到达时间间隔在 30 秒以上的概率是多少?

30. 某种电子设备的寿命(小时)是一个服从下列指数概率密度函数的随机变量:

$$f(x)=\frac{1}{50}e^{-x/50},\ x\geqslant 0$$

a. 它的平均寿命是多少?

b. 该设备在运转的前 25 小时内损坏的概率是多少?

c. 该设备在损坏前能够运转 100 小时以上的概率是多少?

31. Sparagowski & Associates 公司对快餐店外卖窗口的服务时间进行了调查。在麦当劳餐厅从购买到取得食品花费的时间平均是 2.78 分钟(*The Cincinnati Enquirer*, July 9, 2000)。像这样的等待时间往往服从指数分布。

a. 等待时间少于 2 分钟的概率是多少?

b. 等待时间超过 5 分钟的概率是多少?

c. 等待时间超过 2.78 分钟的概率是多少?

32. 根据《Barron's 1998 年主要读者调查》的数据，一个订户平均每年的投资交易次数是 30 次(http://www.barronsmag.com, July 28, 2000)。假定投资者一年内的交易次数服从泊松概率分布。

a. 显示投资交易间隔时间的概率分布。

b. 某个订户在 1 月份没有进行交易的概率是多少?

c. 某个订户两次交易的间隔时间在半个月以内的概率是多少?

小 结

本章把对概率分布的讨论扩展到了连续随机变量。离散和连续概率分布的主要区别在于计算概率的方法不同。对离散分布，概率函数 $f(x)$ 给出了随机变量 x 取各个值的概率；对连续概率分布，我们则使用概率密度函数，也记作 $f(x)$。概率密度函数不直接给出连续随机变量的概率值，其概率是由概率密度函数 $f(x)$ 曲线下的面积给出。因为在单点上曲线下的面积为 0，所以我们观察到连续随机变量取任意特定值的概率为 0。

我们详细介绍了三种连续概率分布——均匀概率分布、正态概率分布和指数概率分布。正态概率分布广泛应用于统计推断，并且在本书的其他部分有着广泛的应用。

术语释义

概率密度函数： 用来计算连续随机变量概率的一种函数。某区间内概率密度函数曲线下的面积就表示连续随机变量在该区间取值的概率。

均匀概率分布： 如果随机变量在长度相等的每个区间内取值的概率相同，则它服从均匀概率分布。

正态概率分布： 一种连续概率分布，它的概率密度函数由均值 μ 和标准差 σ 确定，其形状呈钟形。

标准正态概率分布： 均值为 0 且标准差为 1 的一种正态分布。

指数概率分布： 一种连续概率分布，可用于计算完成任务所花费的时间等类问题。

重要公式

均匀概率密度函数

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\leqslant x\leqslant b\\ 0 & \text{其他}\end{cases} \tag{6.1}$$

正态概率密度函数

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \tag{6.2}$$

标准正态分布的转换公式

$$z=\frac{x-\mu}{\sigma} \tag{6.3}$$

指数概率密度函数

$$f(x)=\frac{1}{\mu}e^{-x/\mu} \qquad x\geq 0 \qquad \mu>0 \tag{6.4}$$

指数分布：累积概率

$$P(x\leq x_0)=1-e^{-x_0/\mu} \tag{6.5}$$

补充练习

33. 一个经理从芝加哥调动到了亚特兰大，因此急需出售她在芝加哥的住房。她的雇主报价 210 000 美元购买该房屋，不过这个报价将在周末到期。这位经理目前还没有得到更好的报价，但能够承受把房子留到下月出售。根据与经纪人的会谈，该经理认为她把房子保留到下个月出售的话，得到的价格将服从 200 000 美元—225 000 美元之间的均匀分布。
    a. 如果她把住房保留到下个月出售，售价的概率密度函数的数学表达式是什么？
    b. 如果她把住房保留到下个月出售，她至少得到 215 000 美元的概率是多少？
    c. 如果她把住房保留到下个月出售，得到的房款少于 210 000 美元的概率是多少？
    d. 她是否应该保留住房到下个月出售？为什么？

34. 美国家庭所欠的债务有 68% 是住房抵押贷款或产权信贷（*Federal Reserve Bulletin*, January 1997）。已知户主年龄在 35 岁以下的家庭，其抵押贷款额的中位数是 63 000 美元。假设这一年龄组的抵押债务额服从正态分布，其标准差为 15 000 美元。
    a. 该年龄组的抵押债务平均数额是多少？
    b. 10% 的最少债务家庭有多少抵押债务？
    c. 抵押债务超过 80 000 美元的家庭占多大百分比？
    d. 5% 的最高抵押债务超过了多少数额？

35. 摩托罗拉公司利用正态分布来确定一个生产工艺出产次品的概率以及次品数量的期望值（*APICS— The Performance Advantage*, July 1991）。假设某生产工艺的设计标准是产出 10 盎司重的产品，且工艺的均值为 10。计算在下列情况下，1 000 单位的产品中出现次品的概率以及次品数量的期望值：
    a. 工艺的标准差为 0.15，并且产品质量的控制范围设定在正负 1 个标准差以内。即如果该产品的单位重量低于 9.85 或高于 10.15 盎司的话，将被归为次品。
    b. 通过工艺设计改进，工艺标准差可以被减少到 0.05。假设产品质量控制范围仍与上面的情况相同，即产品单位重量低于 9.85 或高于 10.15 盎司的话，将被归为次品。

c. 减少工艺的变异程度，并把产品质量控制范围设定在距离均值更多标准差的范围内有什么好处？

36. 美国航空公司 737 客机的平均小时运营成本为 2 071 美元（*The Tampa Tribune*，February 17, 1995）。假定飞机的小时运营成本服从正态分布。
   a. 如果有 11% 的小时运营成本为 1 800 美元或更少，则运营成本的标准差是多少？
   b. 737 客机的小时运营成本在 2 000 和 2 500 美元之间的概率是多少？
   c. 对于 3% 具有最低运营成本的飞机，其小时运营成本是多少？

37. High-Brite 牙膏的销售量被认为近似服从正态分布，均值为每周 10 000 只，标准差为每周 1 500 只。
   a. 在任意 1 周内销售超过 12 000 只牙膏的概率是多少？
   b. 为了使公司的库存能够满足每周需求的概率达到 0.95，应生产多少只牙膏？

38. Ward Doering 自动销售公司正在考虑提供一项特殊的服务合同，以满足服务工作所要求的设备租赁总成本。根据经验，公司经理估计每年的服务成本近似服从正态分布，其均值为 150 美元，标准差为 25 美元。
   a. 如果公司提供给客户的服务合同每年收费 200 美元，那么任意一个客户的服务成本超过 200 美元合同价格的概率是多少？
   b. 公司对每一份服务合同的期望利润是多少？

39. 1997 年底和 1998 年初的亚洲金融危机被认为会导致大量进口廉价商品涌入美国市场，从而在美国引发严重的失业问题。据预测加利福尼亚州遭受的打击尤其严重。经济政策研究机构估计在加州的失业人数平均将达到 126 681（*St. Petersburg Times*，January 24, 1998）。假定加州的失业人数服从正态分布，其标准差是 30 000。
   a. 加州失业人数在 80 000—150 000 之间的概率是多少？
   b. 加州失业人数少于 50 000 人的概率是多少？
   c. 截止值应设定为多少，才能使得失业人数不超过该值的概率达到 0.95？

40. 假设某学院入学考试的成绩服从正态分布，其均值为 450，标准差为 100。
   a. 考试成绩在 400—500 之间的学生占多大百分比？
   b. 假定某人的成绩是 630 分，有多大百分比的学生取得了比他高的成绩？成绩比他低的占多大百分比？
   c. 如果某大学不招收成绩低于 480 分的学生，则参加考试的学生被该大学录取的百分比是多少？

41. 根据《广告时代》的数据，广告公司中女性文案作者的平均底薪要高于她们的男性同行。女性文案作者的平均底薪是 67 000 美元，而男性文案作者的平均底薪是 65 500 美元（*Working Woman*, July/August, 2000）。假设所有文案作者的薪金均服从正态分布，并且其标准差均为 7 000 美元。
   a. 女性作者的收入超过 75 000 美元的概率是多少？
   b. 男性作者的收入超过 75 000 美元的概率是多少？

c. 女性作者的收入低于 50 000 美元的概率是多少？

d. 女性作者要取得多少收入，才能高于 99% 的男性同行的收入？

42. 一台机器向容器填充某种特定产品。根据以往的数据，已知填充重量的标准差为 0.6 盎司。如果只有 2% 的容器填充重量低于 18 盎司，问该机器的平均填充重量是多少？也就是说，μ 必须等于多少？假设填充重量具有正态分布。

43. 一名学生在某大学的计算机中心使用一台计算机终端的时间(以分钟计)服从指数概率分布，其均值为 36 分钟。假定一名学生恰好在另一名学生刚开始在终端上工作时到达该终端。

a. 第二名学生等待时间为 15 分钟以内的概率是多少？

b. 第二名学生等待时间在 15 分钟和 45 分钟之间的概率是多少？

c. 第二名学生不得不等待 1 个小时以上的概率是多少？

44. 一条新的自动生产线平均每天发生 2 起故障，每天的故障数服从泊松概率分布。

a. 故障的平均间隔时间是多少？

b. 作出表示故障间隔时间的指数概率密度函数。

c. 两次故障间隔时间为 1 小时以上的概率是多少？

d. 该生产线能够无故障运转 8 个小时直至切换的概率是多少？

45. 某超市的结账柜台在两名结账顾客之间的空闲时间(以分钟计)服从指数概率分布，其均值为 1.2 分钟。

a. 作出该分布的概率密度函数。

b. 在一名顾客结账后 0.5—1.0 分钟内下一个顾客到达的概率是多少？

c. 结账柜台空闲时间超过 1 分钟的概率是多少？

46. 某保险理赔办公室两次打入电话的时间间隔(以分钟计)具有如下的指数概率分布：

$$f(x) = 0.50e^{-0.50x},\ x \geqslant 0$$

a. 打入电话的平均时间间隔是多少？

b. 时间间隔为 30 秒以内的概率是多少？

c. 时间间隔为 1 分钟以内的概率是多少？

d. 5 分钟以上没有电话打入的概率是多少？

附录 6.1　利用 Minitab 计算连续概率分布

我们以 Grear 轮胎公司问题为例，说明利用 Minitab 软件计算连续概率的方法。在 Grear 轮胎公司问题中，轮胎的行驶里程服从 $\mu = 36\,500$ 且 $\sigma = 5\,000$ 的正态概率分布。需要解决的一个问题是：轮胎行驶里程超过 40 000 英里的概率是多少？

对于连续概率分布，Minitab 能够给出累积概率，也就是说，Minitab 给出了随机变量取值小于或

等于某个特定常数的概率。在 Grear 轮胎行驶里程数问题中，Minitab 能够被用来确定行驶里程小于或等于 40 000 英里的累积概率(本例的特定常数是 40 000)。得到累积概率以后，我们必须用1 减去它，以确定行驶里程超过 40 000 英里的概率。

在使用 Minitab 计算概率之前，在工作表的一列中必须输入特定常数值。对 Grear 轮胎行驶里程数问题，我们把特定常数 40 000 输入 Minitab 工作表的列 1。使用 Minitab 计算正态随机变量取值小于或等于 40 000 的累积概率的步骤如下：

步骤 1. 选择 **Calc** 下拉菜单；
步骤 2. 选择 **Probability Distribution** 选项；
步骤 3. 选择 **Normal** 选项；
步骤 4. 当 Normal Distribution 对话框出现时：
选择 **Cumulative probability**；
在 **Mean** 框中输入 36 500；
在 **Standard deviation** 框中输入 5 000；
在 **Input column** 框中输入 C1（这一列含有 40 000）；
点击 **OK**。

在用户点击 **OK** 以后，Minitab 将显示出正态随机变量取值小于或等于 40 000 的累积概率为 0.758 0。由于我们感兴趣的是轮胎行驶里程超过 40 000 英里的概率，则所求的概率为：1 - 0.758 0 = 0.242 0。

提出的第二个问题是：Grear 公司应设置多大的里程保证，以确保符合保证条件的轮胎不超过 10%？在这里，给出了概率，我们需要找到随机变量的对应值。Minitab 使用一种相反的计算路线来得到与某个给定的累积概率相关的随机变量值。首先，我们必须在 Minitab 工作表的某一列中（如 C1）输入累积概率。在本例中，给出的累积概率为 0.10。然后，仍然使用上面列出的 Minitab 程序的前 3 个步骤；在步骤 4，我们选择 **Inverse cumulative probability** 来代替 **Cumulative probability**，再完成该步骤的剩下内容。Minitab 最后将显示出保证里程为 30 100 英里。

Minitab 还能够计算其他连续概率分布的概率，包括指数概率分布。为了计算指数概率分布，需要执行前面列出的正态概率分布的步骤，并在步骤 3 选择 **Exponential** 选项。步骤 4 也与前面所列的步骤相同，除了不需要输入标准差。累积概率与逆累积概率的输出也和正态概率分布的相同。

附录 6.2　利用 Excel 计算连续概率分布

Excel 能够计算几种连续概率分布的概率，包括正态概率分布与指数概率分布。在本附录中，我们描述怎样利用 Excel 来计算任意正态概率分布的概率。指数概率分布和其他连续概率分布的计算方法与我们对正态概率分布的描述相类似。

让我们回到 Grear 轮胎公司问题，其中的轮胎里程用 $\mu = 36\,500$ 和 $\sigma = 5\,000$ 的正态概率分布来描述。假设我们关心的是轮胎里程超过 40 000 英里的概率。

Excel 的 NORMDIST 函数提供了正态概率分布的累积概率。该函数的一般形式为：NORMDIST（x, μ, σ, cumulative）。对于第四个自变量，如果需要的是累积概率的话，取 TRUE。于是，为了计算轮胎里程小于或等于 40 000 英里的累积概率，我们把如下的公式输入 Excel 工作表的任意一个单元格内：

$$= \text{NORMDIST}(40000,36500,5000,\text{TRUE})$$

然后，0.758 0 将出现在输入公式的那个单元格内，表示轮胎里程小于或等于 40 000 英里的累积概率是 0.758 0。因此，轮胎里程超过 40 000 英里的概率为 1 - 0.758 0 = 0.242 0。

Excel 的 NORMINV 函数利用相反的计算过程来得到对应于给定的累积概率的 x 值。例如，假定我们想要知道公司应该提供多少里程保证，以使得符合保证条件的轮胎不超过 10%。我们把如下的公式输入 Excel 工作表的任意一个单元格内：

$$= \text{NORMINV}(0.1,36500,5000)$$

然后，30 092 将出现在输入公式的那个单元格内，表示某个轮胎不能达到 30 092 英里的概率是 0.10。

计算指数概率的 Excel 函数是 EXPONDIST，可以直接使用它来进行计算。但是，如果用户需要帮助以确定自变量的适当值，他可以利用 Excel 函数向导（参见附录 2.2）。

7 抽样和抽样分布

统计实例:MEAD 公司

7.1 Electronics Associates 公司的抽样问题

7.2 简单随机抽样

从有限总体中抽样

从无限总体中抽样

7.3 点估计

7.4 抽样分布介绍

7.5 $\bar{x}$ 的抽样分布

$\bar{x}$ 的数学期望

$\bar{x}$ 的标准差

中心极限定理

EAI 抽样问题中 $\bar{x}$ 的抽样分布

$\bar{x}$ 抽样分布的实际值

$\bar{x}$ 抽样分布和样本容量之间的关系

7.6 $\bar{p}$ 的抽样分布

$\bar{p}$ 的数学期望

$\bar{p}$ 的标准差

$\bar{p}$ 抽样分布的形式

$\bar{p}$ 抽样分布的实际值

7.7 点估计量的性质

无偏性

有效性

一致性

7.8 其他抽样方法

分层随机抽样

整群抽样

系统抽样

方便抽样

判断抽样

统计实例

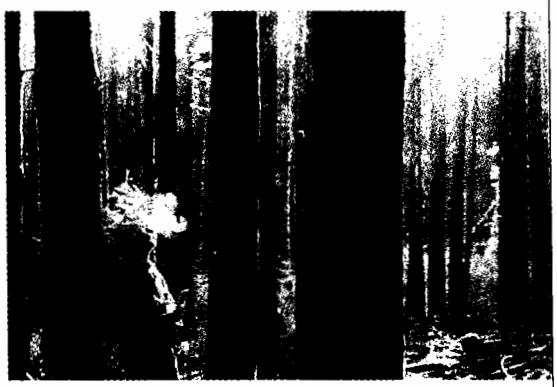
MEAD 公司*

代顿，俄亥俄州

Mead 公司拥有超过 16 000 名职工，生产品种众多的纸张和森林产品。它在 32 个国家进行跨国经营，公司的客户遍布 98 个国家。Mead 公司在纸张产品方面居于领先地位，年生产能力达到 180 万吨。公司的产品包括课本用纸、铜版纸、饮料包装纸和办公用纸。公司的内部顾问组利用抽样方法提供了各种信息，使得公司的生产率显著提高并保持了竞争力。

例如，公司保持了大量的林地储备，可以为公司的许多产品提供原料树木。经理们需要公司的木材林和森林的可靠而准确的信息，以评估公司满足未来原材料需求的能力。森林的现有储量如何？森林过去的生长速度是多少？未来的计划生长速度是多少？如果回答了这些重要问题，公司的经理们就能够制定包括树木的长期种植和采伐时间表在内的未来计划。

公司是怎样获取到它所需要的巨量森林储备信息的呢？答案是：从遍布森林的采样点收集到的数据为了解公司所拥有树木的总体特征打下了基础。为了确定采样点，首先要按照位置和树种把森林储备分成三部分。然后，利用地图和随机数，公司的分析人员在每部分森林中选取 1/5—1/7 英亩的树木作为随机样本。因此，采样点是公司林务员收集数据和了解森林总体特征的地方。



对其森林储备所做的随机抽样使得 Mead 公司能够满足未来的原材料需要。© Larry Goldstein / Tony Stone.

公司内的全体林务员都参与了各自区域的数据收集过程。他们两人一组定期地收集每个采样点中每棵树的信息，收集到的样本数据被输入公司的森林永续存货计算机系统。该计算机系统生成的报告包括了大量的频数分布信息，它们概括了树木种类、目前的森林储量、过去的森林生长速度以及计划的未来生长速度和储量。抽样和有关样本数据的统计汇总共同提供了对公司的木材林和森林进行有效管理所必需的报告。

在本章你将学习简单随机抽样和样本的选择方法。另外，你还将了解到像样本均值和样本比例这样的统计量是如何用于估计与其对应的总体均值和总体比例的。本章还会介绍抽样分布的重要概念。

* 作者感谢 Mead 公司的 Edward P. Winkofsky 博士提供了该统计实例。

在第 1 章，我们已经定义了统计研究中的两个重要方面：**总体**和**样本**。这里再复述一次：

1. **总体**是由研究对象全体构成的集合。
2. **样本**是总体的一个子集。

统计推断的目的是利用样本信息建立对总体性质的估计和检验对总体性质的假设。我们从引

用两个例子入手展开论述，在这两个例子中，所进行的抽样向决策者提供了关于总体的信息。

1. 某轮胎制造商开发了一种新型轮胎，它的设计里程数超过了公司目前的轮胎产品。为了估计新轮胎的平均里程数，制造商选取一个包括 120 条新轮胎的样本进行检测。检测结果显示样本均值为 36 500 英里。于是，新轮胎总体平均里程数的估计值为 36 500 英里。
2. 一个政党的成员正在考虑支持某位候选人竞选美国参议员，因此该党的领导人想要估计支持该候选人的登记选民比例。但要与登记选民中的每一位都进行接触，从时间和成本来看都是不现实的。于是，他们选取了一个由 400 名登记选民组成的样本，其中有 160 人支持该候选人。因此登记选民总体中支持该候选人的比例的估计值为 160/400 = 0.40。

上面的例子说明了如何利用抽样和样本结果估计总体特征。我们注意到在轮胎里程的例子中，要收集轮胎寿命的数据必须把每条检测轮胎使用至报废。很明显对总体中的每条轮胎都进行检测是不可行的，抽样检测是获取所需轮胎里程数据的惟一可行途径。而在选举例子中，接触总体中的每位选民在理论上是可能的，但是时间和成本都不允许这样做。于是，选取登记选民的样本进行调查是更好的办法。

这两个例子说明了人们使用样本的一些原因。但是应该注意，样本结果提供的只是总体特征值的估计值。也就是说，我们不能期望 36 500 英里的样本均值恰好等于总体所有轮胎的平均里程，我们也不能期望确实有 40% 的登记选民支持这位候选人。我们之所以这样说是因为样本只包含了一部分总体。尽管如此，如果使用了恰当的抽样方法，样本结果仍可以对总体特征给出一个“良好”的估计。但是我们究竟希望样本结果好到什么程度呢？幸运的是，统计方法能够回答这个问题。

本章我们将说明如何使用简单随机抽样从总体中选择样本，然后还将介绍如何利用从简单随机样本中取得的数据计算总体均值、总体标准差和总体比例的估计值。另外，我们还将描述抽样分布的重要概念。正如我们所说，拥有恰当的抽样分布知识能够使我们对样本结果的优良性作出度量。在本章最后一节，我们将讨论一些在实践中常用的其他简单随机抽样方法。

样本均值是总体均值的估计值，样本比例是总体比例的估计值。使用这些估计值时，会存在一些抽样误差或正、负值。本章的关键是介绍能够对抽样误差的大小作出概率描述的统计方法。

7.1　Electronics Associates 公司的抽样问题



CD 光盘数据
EAI

Electronics Associates 公司（EAI）的人事主管被分配了一项任务，制作公司 2 500 名经理的简况。在简况中需要确定的特征包括经理们的年薪以及完成了公司管理培训计划的经理所占的比例。

把 2 500 名经理作为该项研究的总体，我们通过查阅公司的人事记录就能够知道每个经理的年薪和培训情况。本书所附光盘储存的一个数据文件包含了总体所有 2 500 名经理的这些信息。

使用第 3 章给出的公式，我们能够计算出年薪数据的总体均值和总体标准差：

$$\text{总体均值：}\mu = 51\,800 \text{ 美元}$$
$$\text{总体标准差：}\sigma = 4\,000 \text{ 美元}$$

培训状况的数据显示，在 2 500 名经理中有 1 500 人已经完成了公司的培训计划。令 p 表示总体中已完成培训计划的经理所占的比例，我们看到 $p = 1\,500/2\,500 = 0.60$。

参数(parameter)是总体的数值特征。总体的平均年薪（$\mu = 51\,800$ 美元）、年薪的总体标准差（$\sigma = 4\,000$ 美元）和已完成培训计划的总体比例（$p = 0.60$）都是 EAI 经理总体的参数。

现在，假设我们无法在公司的数据库中方便地获取所有 EAI 经理的必要信息。这可能是由于新近雇用的员工记录尚未输入计算机，也可能由于员工记录中只有年薪信息（未记录培训状况），或者因为只有一部分岗位的雇员记录收入了计算机。不管具体是什么原因，事实往往是：要收集并且处理总体内所有元素的信息要么不可能，要么极其昂贵。这时，往往要使用来自样本的信息来估计有关的总体参数。

往往从样本中收集信息能够使得成本比从总体收集信息有大幅度的降低，尤其是当收集信息需要进行个人接触时更是如此。

我们现在需要考虑的问题是：公司的人事主管如何使用经理的样本而不是总体内的全部 2 500 名经理来取得有关总体参数的估计值。假设使用一个由 30 名经理组成的样本，显然，制作 30 名经理的简况所花费的时间与成本要比制作全部经理的简况少得多。如果人事主管能够保证由 30 名经理组成的样本可以提供 2 500 名经理总体的充分信息的话，使用样本将比使用整个总体更划算。现在我们从考虑如何确定包括 30 名经理的样本开始，探索使用样本进行 EAI 研究的可能性。

7.2 简单随机抽样

有若干种方法都可以用于从总体中选择样本，其中应用最普遍的一种是**简单随机抽样**(simple random sampling)。简单随机样本的定义和选取简单随机样本的方法都要依赖于总体是**有限的**还是**无限的**。由于 EAI 抽样问题研究的是一个包括 2 500 名经理的有限总体，故我们首先考虑的是从有限总体中抽样。

从有限总体中抽样

来自 N 容量有限总体的容量为 n 的简单随机样本定义如下：

简单随机样本（有限总体）

如果有限总体中每一个容量为 n 的可能样本都有相同的被选取概率，这样选择出的样本就称为容量为 n 的简单随机样本（有限总体）。

从有限总体中选取简单随机样本的一种方法是：每次只选择一个元素，并且每次选择时，要使留在总体内的每个元素都有相同的被选中概率。以这种方式抽取的 n 个元素将满足来自有限总体的简单随机样本定义。

为了从 EAI 经理的有限总体中选择一个简单随机样本，我们首先给每个经理分配一个号码。例如，我们按照经理们的姓名列在 EAI 人事档案中的顺序，能够把号码 1—2 500 分配给每个经理。接下来，我们查阅表 7.1 的随机数表，在表第一行中的每一个数字：6，3，2，…都是具有同样发生概率的随机数。由于 EAI 经理总体名单中的最大号码是 2 500，由 4 个数字组成，故我们将以 4 个数字为一组从表中选择随机数。虽然我们可以从表的任意位置，沿着任选的方向系统地移动来选择随机数，但这次我们从表 7.1 的第一行开始，自左向右地选择。这些 4 位随机数是

6 327　　1 599　　8 671　　7 445　　1 102　　1 514　　1 807

表 7.1　随机数表

63 271	59 986	71 744	51 102	15 141	80 714	58 683	93 108	13 554	79 945
88 547	09 896	95 436	79 115	08 303	01 041	20 030	63 754	08 459	28 364
55 957	57 243	83 865	09 911	19 761	66 535	40 102	26 646	60 147	15 702
46 276	87 453	44 790	67 122	45 573	84 358	21 625	16 999	13 385	22 782
55 363	07 449	34 835	15 290	76 616	67 191	12 777	21 861	68 689	03 263
69 393	92 785	49 902	58 447	42 048	30 378	87 618	26 933	40 640	16 281
13 186	29 431	88 190	04 588	38 733	81 290	89 541	70 290	40 113	08 243
17 726	28 652	56 836	78 351	47 327	18 518	92 222	55 201	27 340	10 493
36 520	64 465	05 550	30 157	82 242	29 520	69 753	72 602	23 756	54 935
81 628	36 100	39 254	56 835	37 636	02 421	98 063	89 641	64 953	99 337
84 649	48 968	75 215	75 498	49 539	74 240	03 466	49 292	36 401	45 525
63 291	11 618	12 613	75 055	43 915	26 488	41 116	64 531	56 827	30 825
70 502	53 225	03 655	05 915	37 140	57 051	48 393	91 322	25 653	06 543
06 426	24 771	59 935	49 801	11 082	66 762	94 477	02 494	88 215	27 191
20 711	55 609	29 430	70 165	45 406	78 484	31 639	52 009	18 873	96 927
41 990	70 538	77 191	25 860	55 204	73 417	83 920	69 468	74 972	38 712
72 452	36 618	76 298	26 678	89 334	33 938	95 567	29 380	75 906	91 807
37 042	40 318	57 099	10 528	09 925	89 773	41 335	96 244	29 002	46 453
53 766	52 875	15 987	46 962	67 342	77 592	57 651	95 508	80 033	69 828
90 585	58 955	53 122	16 025	84 299	53 310	67 380	84 249	25 348	04 332
32 001	96 293	37 203	64 516	51 530	37 069	40 261	61 374	058 15	06 714
62 606	64 324	46 354	72 157	67 248	20 135	49 804	09 226	64 419	29 457
10 078	28 073	85 389	50 324	14 500	15 562	64 165	06 125	71 353	77 669
91 561	46 145	24 177	15 294	10 061	98 124	75 732	00 815	83 452	97 355
13 091	98 112	53 959	79 607	52 244	63 303	10 413	63 839	74 762	50 289

也可以使用计算机生成的随机数来选择随机样本。Excel 在它的工作表中提供了一个能够生成随机数的函数。

因为表中的数字都是随机的，所以这些 4 位数字出现的概率或可能性相等。

现在我们能够使用这些 4 位随机数给总体内的每个经理提供一个被选入随机样本的同等机会。第一个数字6 327 大于 2 500，它不等于总体中任意一名经理的号码，于是放弃该数字。第二个数字 1 599 介于 1—2 500 之间，因此为简单随机样本选择的第一个经理在 EAI 经理名单上的编号是 1 599。继续进行这个选择过程，我们在确定选入编号为1 102，1 514 和 1 807 的经理之前，忽略数字8 671 和 7 445。该过程将一直持续下去，直到我们选出 30 名 EAI 经理组成一个简单随机样本为止。

在我们选出 30 名经理以前，已经被选入样本的随机数可能会在表中再次出现，由于我们不想把一个经理多次选入样本，故略过样本中已经出现过的随机数，因为与该数对应的经理已经进入了样本。这种选择样本的方法称为**无放回抽样**(sampling without replacement)。如果我们在选择样本时，可以多次把同一个随机数选入样本，某些经理可能会两次或多次出现在样本中，我们做的就是**放回抽样**(sampling with replacement)。放回抽样也是确定简单随机样本的一种有效方法，不过，还是无放回抽样的应用更为普遍。当我们提到简单随机抽样时，总是假定抽样是无放回的。

从无限总体中抽样

许多商务和经济中的抽样情况涉及到有限总体，但是在某些情况下，总体或者是无限的，或者特别大，以至于在实践中必须作为无限总体来处理。在从无限总体抽样时，我们必须使用新的简单随机样本定义。另外，因为无法将无限总体内的元素一一列出并编号，我们必须以不同的方法选择样本。

在实践中，如果所研究的总体涉及到一个正在进行的过程，则列出或统计总体的每个元素是不可能的，这时通常把总体视作无限的。

假设我们想要估计某个快餐店在上午 11:30 —下午 1:30 午餐时间内顾客从点餐到拿到食品所需的平均时间。如果我们把所有可能的顾客消费作为总体，那么列出所有可能的消费是不现实的。实际上，如果我们定义总体是在午餐时间有可能发生的所有顾客消费，我们可以认为总体是无限的。我们的任务是从这个总体中选择一个包括 n 名顾客的简单随机样本。来自无限总体的简单随机样本定义如下：

简单随机样本（无限总体）

如果样本的选择过程满足下面两个条件，这个样本就是来自无限总体的简单随机样本。

1. 选出的每个元素都来自同一个总体。
2. 对每个元素的选择都是独立的。

对于选择快餐店顾客消费的简单随机样本问题，由于任意一次顾客消费都发生在上午 11:30—下午 1:30 午餐时间内，这段时间快餐店都在正常的条件下营业，所以我们发现满足了定义的第一个条件。而通过确保对某位顾客的选择不会影响到对其他任意顾客的选择，定义的第二个条件也得到了满足，即对顾客的选择是独立的。

一家著名的快餐店就在这种条件下进行了一次简单随机抽样。设计该抽样方法的基础是这样一个事实：一些顾客持有三明治、饮料和法式炸薯条等食品的优惠购物券。当一位顾客出示优惠券时，他的下一位顾客将被选入样本。因为顾客出示优惠券的行为是随机且独立的，所以公司的抽样计划满足来自无限总体的简单随机样本的两个条件。

由于不可能列出无限总体，所以随机数选择方法不能用于无限总体。这时，必须对样本选择方法进行特殊的修改，以独立地选择每个元素，并且避免给予某种类型的元素以较高的选中概率，从而引起选择偏差。

评　注

1. 有限总体往往被定义为组织成员名册、学生注册记录、信用卡账户清单、存货产品编号等等。无限总体往往被定义为一个持续进行的过程，总体内的元素由在相同条件下无限运行的过程产生。在这种情况下，要列出所有的元素是不可能的。例如，由所有可能制造出的部件构成的总体，由所有可能的顾客消费构成的总体，由所有可能的银行交易构成的总体等等，都可以被归类为无限总体。
2. 从 N 容量有限总体中能够选择容量为 n 的不同简单随机样本的个数为 $\frac{N!}{n!(N-n)!}$。在该公式中，$N!$ 和 $n!$ 是在第 4 章讨论过的阶乘计算。对于 EAI 问题，$N=2\,500$，$n=30$，这个公式的计算结果显示出，大约有 2.75×10^{69} 个不同的简单随机样本由 30 名 EAI 经理组成。
3. 计算机也能够生成随机数以选择随机样本。例如，给定 EAI 总体的 2 500 名经理，我们就能够使用 Excel 的函数 = RANDBETWEEN(1,2 500) 来生成介于 1—2 500 之间的随机数。然后按照我们使用表 7.1 随机数的同样方法，可以使用这些计算机生成的随机数确定由 30 名经理组成的简单随机样本。

练 习

方 法

1. 考虑一个由标记分别为 A、B、C、D、E 的五个元素构成的有限总体。从中可以选择 10 个容量为 2 的简单随机样本。

自测题

a. 从 AB、AC 开始列出这 10 个样本。

b. 采用简单随机抽样时，容量为 2 的每个样本被选到的概率是多少？

c. 假定随机数 1 对应于 A，随机数 2 对应于 B，依此类推。列出使用随机数 8　0　5　7　5　3　2 选择的容量为 2 的简单随机样本。

2. 假定某个有限总体拥有 350 个元素。使用下面 5 位随机数的后三位数(601，022，448，…)，确定选入简单随机样本的前 4 个元素。

98 601　73 022　83 448　02 147　34 229　27 553　84 147　93 289　14 209

应 用

3. 《财富》公布了美国 500 家最大工业公司的销售额、利润、资产、股东权益、市场价值以及每股收益数据(*The Fortune* 500，2000)。假定你想从《财富》500 强名单中选择 10 家公司组成简单随机样本，使用表 7.1 第 9 行的后 3 位数，并且从 554 开始按列向下读取，确定所选入的 10 家公司的编号。

自测题

4. 其股票被广泛持有的 10 家公司为(*USA Today*，September 6, 2000)：

AT&T	IBM
America Online	Johnson & Johnson
Cisco Systems	Microsoft
General Electric	Motorola
Intel	Pfizer

a. 假设为了对大公司的公共商业实践进行深入研究，需要在这些公司中选择 5 家组成随机样本。从表 7.1 的第一个随机数开始，按列向下读取，使用一位的随机数来选择用于此项研究的包含 5 家公司的简单随机样本。

b. 根据评注第 2 条，从这 10 家公司的名单中能够选择多少个容量为 5 的简单随机样本？

5. 一个学生组织想要估计支持在选修课上实行强制“及格/不及格”记分制度的学生所占的比

例。从注册办公室可以得到本学期注册的 645 名学生的名单和住址，使用表 7.1 第 10 行的 3 位随机数，按行自左向右移动，利用简单随机抽样方法选择前面的 10 名学生。3 位随机数从 816，283 和 610 开始。

6. 由美国人口普查局出版的《县与市数据册》，列出了全美 3 139 个县的信息。假定有一项研究要随机选取 30 个县收集数据。使用表 7.1 最后一列的四位随机数确定对应于样本前 5 个县的编号。忽略表上的第一位数，从四位随机数 9 945，8 364 和 5 702 开始。

7. 假定我们想要从某城市的 372 名实习医生中选取 12 人作为简单随机样本。医生的姓名可以从一个地方医疗机构得到。使用表 7.1 第 8 列的 5 位随机数来选取样本中的 12 名医生。在每个五位随机数中忽略前两位数字。该过程从随机数 108 开始，按列向下选择。

8. 下表列出了 2000 赛季开始阶段 NCAA 橄榄球联赛的前 25 支球队（*Sports Illustrated*，August 14, 2000）。使用表 7.1 第 9 列的随机数并且从 13 554 开始，选择 6 支球队组成简单随机样本。在你的选择过程中从球队 13 开始，使用第 9 列每一行的前二位数字。问哪 6 支球队会被选为简单随机样本？

1. Nebraska	14. Tennessee
2. Florida State	15. TCU
3. Alabama	16. Purdue
4. Michigan	17. Mississippi
5. Wisconsin	18. USC
6. Kansas State	19. Penn State
7. Georgia	20. Southern Miss
8. Clemson	21. Illinois
9. Texas	22. Ohio State
10. Miami	23. Oklahoma
11. Florida	24. Colorado State
12. Virginia Tech	25. Colorado
13. Washington	

9. 《商业周刊》提供了 895 只共同基金的业绩数据和年度评分（*Business Week*，February 3, 1997）。假设为了对共同基金的业绩表现进行更多研究，需要从 895 只共同基金中选择 12 只组成简单随机样本。使用表 7.1 第 4 列的随机数，从 51 102 开始选择 12 只共同基金。你的选择过程从共同基金 511 开始，采用第 4 列每一行的前三位数。问简单随机样本中的 12 只共同基金的编号分别是多少？

10. 指出下列总体应该被认为是有限的还是无限的：
    a. 加利福尼亚州的全部登记选民。
    b. TV-M 公司的宾夕法尼亚 Allen 镇工厂生产的所有电视机。

c. 某邮购公司处理的所有订单。

d. 打入某地区警察局的所有紧急电话。

e. Fibercon 公司在 5 月 17 日第二班次生产的所有元件。

7.3 点估计

我们已经描述了如何选择简单随机样本，现在仍回到 EAI 问题中。假设已经选出 30 名经理组成了简单随机样本，其对应的年薪和培训状况数据列示在表 7.2 中。我们使用标记 x_1，x_2 等表示样本中第一位经理的年薪，其他依此类推。而参加过管理培训计划就在“是否完成管理培训计划”这一列里用“是”表示。

表 7.2 由 30 名 EAI 经理组成的简单随机样本的年薪与培训状况

年薪(美元)	是否完成管理培训计划	年薪(美元)	是否完成管理培训计划
$x_1 = 49\,094.30$	是	$x_{16} = 51\,766.00$	是
$x_2 = 53\,263.90$	是	$x_{17} = 52\,541.30$	否
$x_3 = 49\,643.50$	是	$x_{18} = 44\,980.00$	是
$x_4 = 49\,894.90$	是	$x_{19} = 51\,932.60$	是
$x_5 = 47\,621.60$	否	$x_{20} = 52\,973.00$	是
$x_6 = 55\,924.00$	是	$x_{21} = 45\,120.90$	是
$x_7 = 49\,092.30$	是	$x_{22} = 51\,753.00$	是
$x_8 = 51\,404.40$	是	$x_{23} = 54\,391.80$	否
$x_9 = 50\,957.70$	是	$x_{24} = 50\,164.20$	否
$x_{10} = 55\,109.70$	是	$x_{25} = 52\,973.60$	否
$x_{11} = 45\,922.60$	是	$x_{26} = 50\,241.30$	否
$x_{12} = 57\,268.40$	否	$x_{27} = 52\,793.90$	否
$x_{13} = 55\,688.80$	是	$x_{28} = 50\,979.40$	是
$x_{14} = 51\,564.70$	否	$x_{29} = 55\,860.90$	是
$x_{15} = 56\,188.20$	否	$x_{30} = 57\,309.10$	否

为了估计总体参数的值，我们需要计算对应的样本特征，即**样本统计量**(sample statistic)。例如，为了估计 EAI 经理年薪的总体均值 μ 和总体标准差 σ，我们使用表 7.2 的数据计算对应的样本统计量：样本均值 $\bar{x}$ 和样本标准差 s。利用第 3 章介绍的样本均值和样本标准差公式，得到样本均值为

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1\,554\,420}{30} = 51\,814.00 \text{ 美元}$$

样本标准差为

$$s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{325\ 009\ 260}{29}}=3\ 347.72\text{ 美元}$$

另外，通过计算样本中回答“是”的经理比例，我们还能够估计总体中已完成管理培训计划的经理所占的比例。表 7.2 显示在样本的 30 个经理中有 19 人已经完成了管理培训计划，因此，用$\bar{p}$表示的样本比例为

$$\bar{p}=\frac{19}{30}=0.63$$

它用做总体比例 p 的估计值。

通过上面的计算过程，我们进行了被称为**点估计**的统计过程。在点估计中，我们使用样本数据计算出样本统计量的值作为对总体参数的估计。利用点估计的术语，我们把$\bar{x}$作为总体均值 μ 的**点估计量**(point estimator)，把 s 作为总体标准差 σ 的点估计量，把$\bar{p}$作为总体比例 p 的**点估计量**。在一个特定样本中得到的$\bar{x}$、s 或$\bar{p}$的实际数值被称为参数的**点估计值**(point estimate)。于是对于 30 名 EAI 经理的样本，51 814.00 美元是 μ 的点估计值，3 347.72 美元是 σ 的点估计值，0.63 是 p 的点估计值。表 7.3 总结了样本的结果并且对点估计值与总体参数的实际值进行了比较。

表 7.3　从包含 30 名经理的简单随机样本取得的点估计值

总体参数	参数值	点估计量	点估计值
μ = 总体平均年薪	51 800.00 美元	$\bar{x}$ = 样本平均年薪	51 814.00 美元
σ = 年薪的总体标准差	4 000.00 美元	s = 年薪样本标准差	3 347.72 美元
p = 已经完成管理培训计划的总体比例	0.60	$\bar{p}$ = 已经完成管理培训计划的样本比例	0.63

正如表 7.3 所显示的，没有一个点估计值准确地等于对应的总体参数。因为在进行估计时，我们用样本数据代替了整个总体的统计数据，所以我们能够预期到这个偏差。总体参数与其对应的无偏点估计之差的绝对值被称为**抽样误差**(sampling error)。* 对于样本均值、样本标准差和样本比例，抽样误差分别为$|\bar{x}-\mu|$，$|s-\sigma|$，和$|\bar{p}-p|$。因此，对于 EAI 样本，本均值的抽样误差是 $|51\ 814-51\ 800|=14$ 美元，样本标准差的抽样误差是 $|3\ 347.72-4\ 000.00|=652.28$ 美元，样本比例的抽样误差是 $|0.63-0.60|=0.03$。

抽样误差是采用了总体的一个子集(样本)而没有采用整个总体的结果。

在这里由于总体参数已知，故我们能够计算抽样误差。但是在实际抽样应用中，因为总体参数的值是未知的，因此我们就无法准确地计算抽样误差。在后面的内容中将会说明为了对抽样误差的大小进行概率描述，统计学者将如何分析样本数据。

* 从技术上讲，只有当点估计量无偏时，点估计与总体参数之差的绝对值才被称为抽样误差。此处的点估计量($\bar{x}$, s, $\bar{p}$)是无偏的。

评 注

在对点估计量的讨论中，我们用$\bar{x}$表示样本均值，用$\bar{p}$表示样本比例。基于样本比例也是一种样本均值的事实，我们采用了$\bar{p}$的符号。例如，假设在容量为 n 的样本中数据值分别为 x_1，x_2，...，x_n，对于第 i 个观察值，当我们感兴趣的特征出现时，令 $x_i=1$；当我们感兴趣的特征未出现时，则令 $x_i=0$。然后用样本均值的公式 $\Sigma x_i/n$ 计算样本比例。就像用样本均值$\bar{x}$估计总体均值一样，我们在字母上加横线以提醒读者，$\bar{p}$是对总体比例的估计。也有一些教科书用$\hat{p}$代替$\bar{p}$来表示样本比例。

练 习

方 法

11. 下列数据来自一个简单随机样本：



5　　8　　10　　7　　10　　14

a. 总体均值的点估计是多少？

b. 总体标准差的点估计是多少？

12. 某项调查询问了一个由 150 人组成的样本，结果收到的回答中有 75 人为“是”，55 人为“否”，还有 20 人未发表意见。

a. 回答“是”的总体比例的点估计是多少？

b. 回答“否”的总体比例的点估计是多少？

应 用

13. 5 个月的销售数据给出了如下信息：



月份：	1	2	3	4	5
销售量：	94	100	85	94	92

a. 总体每月平均销售量的点估计是多少？

b. 总体标准差的点估计是多少？

14. 在一项调查中，一个由 784 名 9—14 岁的儿童所组成的样本被询问他们得到父母零花钱的原因（*Consumer Reports*，January 1997）。回答如下：

零花钱来源	频数
只有补贴	149
劳动奖励加上补贴	219
只有劳动奖励	251
没有零花钱	165
总计	784

a. 总体中零花钱来源只有补贴一项的儿童所占比例的点估计是多少？
b. 总体中零花钱来源只有劳动奖励一项的儿童所占比例的点估计是多少？
c. 考虑所有来源，总体中能够从父母那里得到零花钱的儿童所占比例的点估计是多少？

15. 《电器杂志》给出了家用电器预计使用寿命的估计值（*USA Today*，September 5, 2000）。由 10 个 VCR 组成的样本提供了如下的使用寿命数据（年）：

6.5　8.0　6.2　7.4　7.0　8.4　9.5　4.6　5.0　7.4

a. VCR 产品总体平均使用寿命的点估计是多少？
b. VCR 产品使用寿命的总体标准差的点估计是多少？

16. 美国交通部报告了主要航线准时或提前到达班次的统计结果（*Associated Press*，September 8, 2000）。假设要根据一个包括 1 400 个航班的样本数据来估计所有航班准时到达的比例。如果有 1 117 个航班准时抵达，那么在总体中准时抵达的航班所占比例的点估计是多少？

17. 一项 Louis Harris 民意测验调查了 1 008 名成年人，以了解人们对经济的看法（*Business Week*，August 7, 2000）。人们的回答如下：

595 人	经济正在增长
332 人	经济保持在原有水平
81 人	经济正在衰退

建立下列总体参数的点估计：
a. 所有成年人中认为经济正在增长的比例。
b. 所有成年人中认为经济保持在原有水平的比例。
c. 所有成年人中认为经济正在衰退的比例。

7.4　抽样分布介绍

在上一节，我们使用由 30 名 EAI 经理组成的简单随机样本，对所有 EAI 经理总体的年薪均值和标准差以及总体中已完成了公司管理培训计划的经理所占的比例分别建立了点估计。假设我

们另外选择一个包括 30 名 EAI 经理的简单随机样本，对第二个样本的数据进行分析，我们得到了下面的信息：

$$\text{样本均值} \quad \bar{x} = 52\,669.70 \text{ 美元}$$
$$\text{样本标准差} \quad s = 4\,239.07 \text{ 美元}$$
$$\text{样本比例} \quad \bar{p} = 0.70$$

这些结果说明由第二个样本得到的$\bar{x}$、s和$\bar{p}$的值与第一个样本不同。一般地，第二个简单随机样本与第一个简单随机样本包含的元素不会完全相同。让我们来想像一下，我们把选择简单随机样本的同样过程一再地进行下去，每次都计算$\bar{x}$、s和$\bar{p}$的值。以这种方式，我们能够得到这些点估计量的大量不同数值。为了说明这种情况，我们重复进行 EAI 问题中的简单随机抽样过程，直到取得 500 个包括 30 名经理的样本以及对应的$\bar{x}$、s和$\bar{p}$的值。表 7.4 列出了一部分结果，而表 7.5 则列出了 500 个$\bar{x}$值的频数和相对频数分布。图 7.1 是$\bar{x}$值的相对频数直方图。

表 7.4　从 500 个包括 30 名 EAI 经理的简单随机样本中得到的$\bar{x}$、s和$\bar{p}$值

样本编号	样本均值 ($\bar{x}$)	样本标准差 (s)	样本比例 ($\bar{p}$)
1	51 814.00	3 347.72	0.63
2	52 669.70	4 239.07	0.70
3	51 780.30	4 433.43	0.67
4	51 587.90	3 985.32	0.53
…	…	…	…
500	51 752.00	3 857.82	0.50

表 7.5　从 500 个包括 300 名 EAI 经理的简单随机样本中得到的频数分布

平均年薪(美元)	频数	相对频数
49 500.00—49 999.99	2	0.004
50 000.00—50 499.99	16	0.032
50 500.00—50 999.99	52	0.104
51 000.00—51 499.99	101	0.202
51 500.00—51 999.99	133	0.266
52 000.00—52 499.99	110	0.220
52 500.00—52 999.99	54	0.108
53 000.00—53 499.99	26	0.052
53 500.00—53 999.99	6	0.012
总计	500	1.000

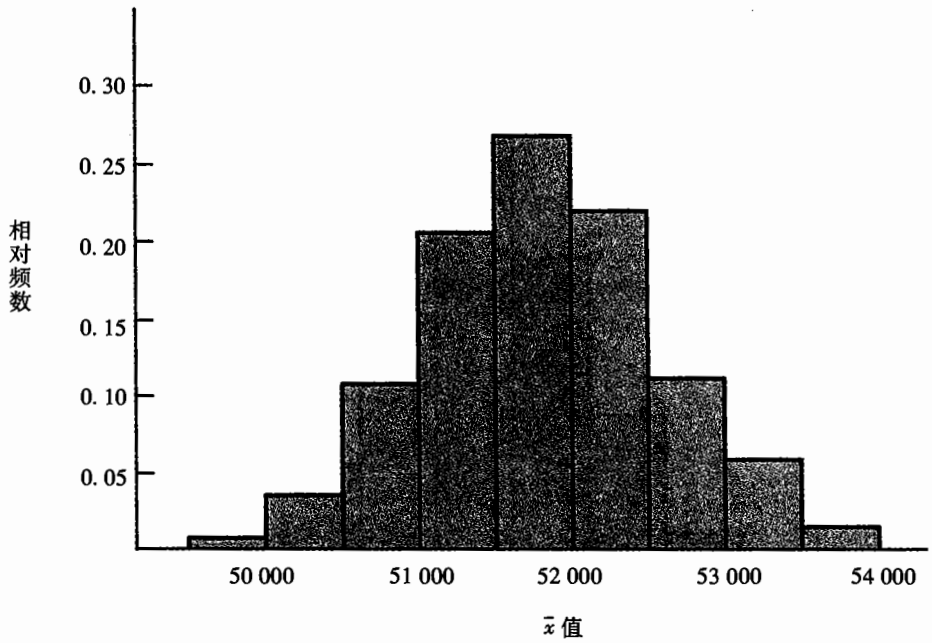


图 7.1　从 500 个容量为 30 的简单随机样本得到的 $\bar{x}$ 值的相对频数直方图

在第 5 章我们定义了随机变量是对试验结果的数值描述。如果我们把每次简单随机样本的选择过程都当做一次试验的话，样本均值$\bar{x}$ 就是对试验结果的数值描述。因此样本均值$\bar{x}$ 是一个随机变量。就像其他随机变量一样，$\bar{x}$ 具有均值或者说是数学期望、方差和概率分布。由于$\bar{x}$ 取不同的可能值是选择了不同的简单随机样本的结果，所以$\bar{x}$ 的概率分布被称为$\bar{x}$ 的**抽样分布**(sampling distribution)。通过抽样分布的有关知识和性质，我们能够对样本均值$\bar{x}$ 与总体均值 μ 的接近程度进行概率描述。

抽样分布的概念是本章最重要的内容之一，对以后各章节的理解，在很大程度上要依赖于对本章介绍的抽样分布的理解和应用能力。

让我们回到图 7.1 上来。我们需要列举出每一个可能的由 30 名经理组成的样本，并计算每个样本的均值以完全确定$\bar{x}$ 的抽样分布。500 个$\bar{x}$ 数据的直方图仅给出了该抽样分布的一个近似表示。根据这个近似表示，我们观察到该分布具有钟形的外观。我们注意到在各个$\bar{x}$ 数值之间的最大差距超过了 2 000 美元。但是我们也注意到$\bar{x}$ 值的最大集中区域就在总体均值 $\mu = 51\,800$ 美元附近，并且 500 个$\bar{x}$ 值的平均值也在总体均值附近。我们将在下一节更加全面地描述 $\bar{x}$ 抽样分布的这个性质。

500 个样本标准差 s 的数值和 500 个样本比例$\bar{p}$ 的数值也用相对频数直方图的形式汇总在图 7.2 和图 7.3 中。与$\bar{x}$ 的情形一样，s 和 $\bar{p}$ 都是对简单随机样本的结果提供数值描述的随机变量。

如果从总体中把每一个容量为 30 的样本都选择出来，并计算每个样本的 s 和 $\bar{p}$ 的值，所得到的概率分布将分别被称为 s 的抽样分布和 $\bar{p}$ 的抽样分布。500 个样本值的相对频数直方图，图 7.2 和图 7.3，分别给出了这两类抽样分布的一般外形特点。

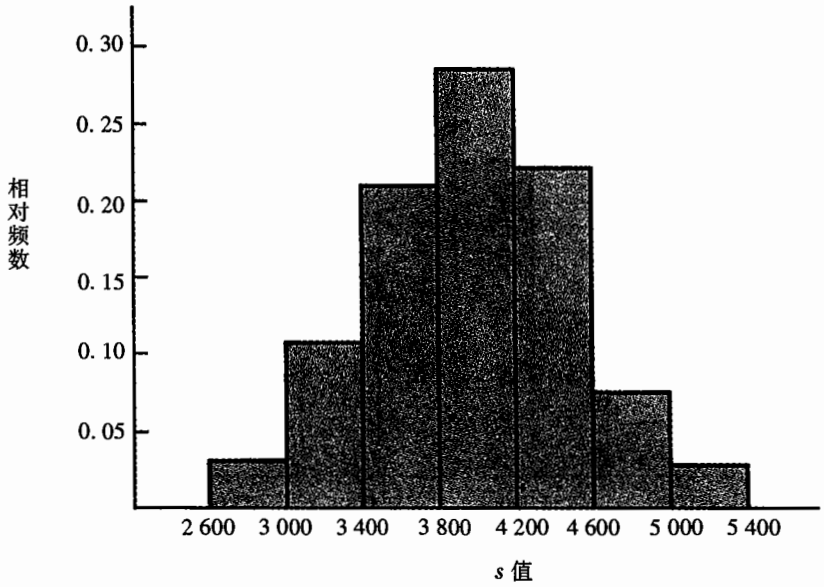


图 7.2　从 500 个容量为 30 的简单随机样本得到的 s 值的相对频数直方图

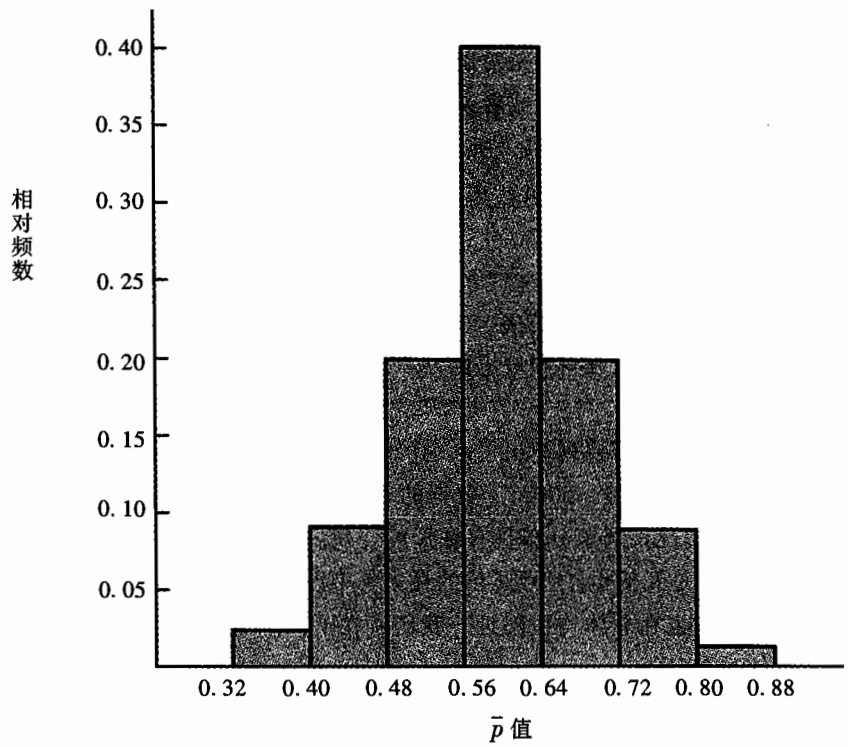


图 7.3　从 500 个容量为 30 的简单随机样本得到的 $\bar{p}$ 值的相对频数直方图

在实践中，我们通常只从总体中选择一个简单随机样本。在本节我们把抽样过程重复了 500 次，只是为了说明有许多不同的样本都是可能的，并且不同的样本产生了样本统计量 $\bar{x}$、s 和 $\bar{p}$ 的不同数值。任意一个样本统计量的概率分布被称为该统计量的抽样分布。在 7.5 节我们将介绍 $\bar{x}$ 抽样分布的性质。而在 7.6 节，我们将说明 $\bar{p}$ 抽样分布的性质。

7.5　$\bar{x}$ 的抽样分布

用样本 $\bar{x}$ 均值对总体均值 μ 进行推断是最常用的统计学方法之一。这个方法的过程如图 7.4 所示。每重复一次该过程，我们能够预期将得到不同的样本均值 $\bar{x}$ 数值。样本均值 $\bar{x}$ 所有可能值的概率分布被称为样本均值 $\bar{x}$ 的抽样分布。

$\bar{x}$ 的抽样分布

$\bar{x}$ 的抽样分布是样本均值 $\bar{x}$ 的所有可能值的概率分布。

本节的目的是描述 $\bar{x}$ 抽样分布的性质，包括 $\bar{x}$ 的均值或者说是数学期望、$\bar{x}$ 的标准差以及 $\bar{x}$ 抽样分布的形状或形式。利用 $\bar{x}$ 抽样分布的知识，我们在使用 $\bar{x}$ 估计 μ 时能够对涉及到的抽样误差进行概率描述。首先我们从考虑所有可能值的均值，即 $\bar{x}$ 的数学期望开始。

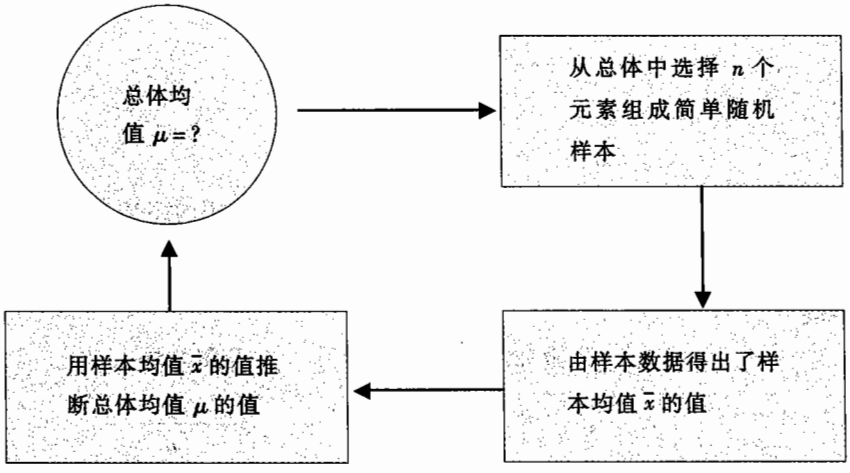


图 7.4　利用样本均值推断总体均值的统计学过程

$\bar{x}$ 的数学期望

在 EAI 抽样问题中我们看到，简单随机样本不同导致了样本均值 $\bar{x}$ 取不同的数值。由于随机变量 $\bar{x}$ 有许多可能数值，故我们往往想知道由不同简单随机样本所产生的 $\bar{x}$ 所有可能值的平均值。

随机变量$\bar{x}$的平均值也就是$\bar{x}$的数学期望。我们令 $E(\bar{x})$表示$\bar{x}$的数学期望，μ表示我们抽样所在的总体均值。能够证明对于简单随机抽样，$E(\bar{x})$和 μ是相等的。

$\bar{x}$的数学期望

$$E(\bar{x}) = \mu \tag{7.1}$$

式中 $E(\bar{x})$——$\bar{x}$的数学期望；

μ——总体均值。

$\bar{x}$的数学期望等于从总体中抽取样本的总体均值。

这个结果说明，对于简单随机抽样，$\bar{x}$的数学期望等于总体均值。在 7.1 节我们看到 EAI 经理的总体平均年薪为 $\mu = 51\,800$ 美元，于是根据公式(7.1)，EAI 研究中所有可能样本均值的数学期望也是 51 800 美元。

$\bar{x}$的标准差

我们首先来定义$\bar{x}$抽样分布的标准差。在讨论中我们使用下面的符号：

$\sigma_{\bar{x}}$ = $\bar{x}$抽样分布的标准差；
σ = 总体标准差；
n = 样本容量；
N = 总体容量。

能够证明，对于简单随机抽样，$\bar{x}$的标准差要依赖于总体是有限的还是无限的。$\bar{x}$标准差的两个数学表达式如下：

$\bar{x}$的标准差

有限总体	无限总体	
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	(7.2)

比较(7.2)中的这两个表达式，我们看到在有限总体情况下需要系数 $\sqrt{(N-n)/(N-1)}$，而无限总体则不需要该系数。该系数通常被称为**有限总体修正系数**(finite population correction factor)。在许多实际的抽样问题中，我们发现要研究的总体虽然是有限的，但容量很大，而样本容量则相对较小。这时，有限总体修正系数 $\sqrt{(N-n)/(N-1)}$ 接近于 1。结果是，在有限总体$\bar{x}$标准差的值与无限总体 $\bar{x}$标准差的值之间的差别变得可以忽略不计，这样，即使总体是有限的，$\sigma_{\bar{x}} = \sigma/\sqrt{n}$也是 $\bar{x}$标准差的良好近似。这个观察结果导出了计算 $\bar{x}$标准差的一般指导或经验法则：

使用下面的表达式$\bar{x}$ 计算的标准差

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{7.3}$$

该公式要适用，只有当：

1. 总体是无限的，或者
2. 总体虽然是有限的，但样本容量小于或等于总体容量的 5%，即 $n/N\leqslant 0.05$。

在 $n/N>0.05$ 的情况下，应该使用公式(7.2)中的有限总体表达式来计算$\sigma_{\bar{x}}$。除非特别指出，在本书中我们总是假定总体容量较大，有 $n/N\leqslant 0.05$，能够使用公式（7.3）来计算$\sigma_{\bar{x}}$。

现在，我们回到 EAI 研究上来，确定 30 名 EAI 经理样本产生的所有可能样本均值的标准差。在 7.1 节，我们看到年薪数据的总体标准差为 $\sigma=4\,000$。该问题中的总体是有限的，容量 $N=2\,500$。但是由于样本容量为 30，我们有 $n/N=30/2\,500=0.012$。根据公式(7.3)的经验法则，我们可以忽略有限总体修正系数并计算$\bar{x}$ 的标准差如下：

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}=\frac{4\,000}{\sqrt{30}}=730.30$$

稍后我们将看到的值有助于确定样本均值与总体均值的偏离程度。由于$\sigma_{\bar{x}}$ 在计算可能存在的抽样误差方面所起的作用，它被称为**均值的标准误差**(standard error of the mean)。

中心极限定理

确定$\bar{x}$ 抽样分布特征的最后一步是确定$\bar{x}$ 概率分布的形式。我们来考虑两种情况：总体分布未知和已知总体分布为正态分布。

当总体分布未知时，我们需要依靠统计学中最重要的定理之一：**中心极限定理**(central limit theorem)。中心极限定理在$\bar{x}$ 抽样分布中的应用如下：

中心极限定理适用于任何总体。因此，即使总体分布未知，它仍然能够描述$\bar{x}$ 的抽样分布。

中心极限定理

在从总体中选取容量为 n 的简单随机样本时，如果样本容量较大的话，能够用**正态概率分布**来近似样本均值$\bar{x}$ 的抽样分布。

图 7.5 显示了中心极限定理是怎样作用于三个不同总体的。图的第一格显示出每个总体都不是正态的。但是注意，当样本容量增加时，$\bar{x}$ 的抽样分布开始发生了什么变化。当样本容量为 2 时，我们看到$\bar{x}$ 的抽样分布开始呈现出与总体分布不同的外形。当样本容量为 5 时，我们看到这

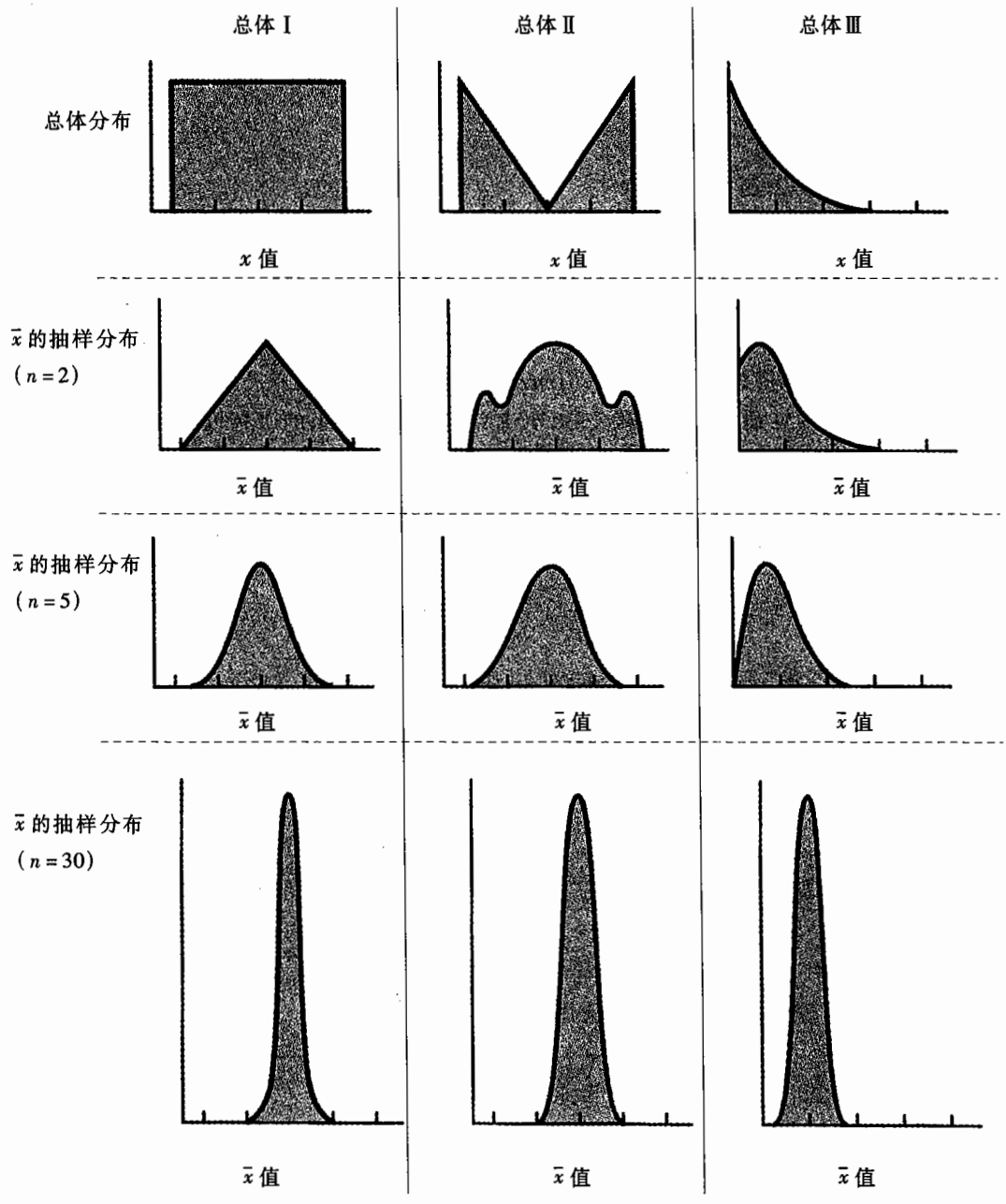


图 7.5　三个总体的中心极限定理图示

三个抽样分布都开始呈现出钟形外观。最后，容量为 30 的样本显示，这三个抽样分布都近似为正态分布。于是，对于容量足够大的样本，$\bar{x}$ 的抽样分布可以用正态概率分布近似表示。不过，样本容量必须要达到多大，我们才能够使用中心极限定理呢？统计学者通过研究各种总体在不同样本容量下的$\bar{x}$ 抽样分布，回答了这个问题。一般的统计实践假定，对于大多数的应用，当样本容量达到 30 或 30 以上时，就能够用正态概率分布来近似表示$\bar{x}$ 的抽样分布。实际上，在样本容量达到 30 或 30 以上时，就假定它满足了中心极限定理的大样本条件。这个结果非常重要，我们有必要重申一遍：

当样本容量较大时，$\bar{x}$ 的抽样分布可以由正态概率分布近似表示。我们规定简单随机样本的大样本条件为样本容量达到 30 或 30 以上。

当总体分布未知时，中心极限定理是确定 $\bar{x}$ 抽样分布形式的关键。但是，我们可能还会遇到一些总体被假定或认为具有正态概率分布的抽样情况。这时，下面的结果确定了 $\bar{x}$ 的抽样分布形式：

当总体具有正态概率分布时，对于任何样本容量，$\bar{x}$ 的抽样分布都具有正态概率分布。

总之，如果我们使用较大（$n \geqslant 30$）的简单随机样本，中心极限定理使我们能够用正态概率分布来近似 $\bar{x}$ 的抽样分布。当简单随机样本的容量较小（$n < 30$）时，只有当我们能够假定总体具有正态概率分布时，$\bar{x}$ 的抽样分布才能被认为是正态的。

EAI 抽样问题中 $\bar{x}$ 的抽样分布

在 EAI 研究中，我们已经得到 $E(\bar{x}) = 51\,800$，$\sigma_{\bar{x}} = 730.30$。因为我们使用的是一个包括 30 名经理的简单随机样本，中心极限定理使得我们能够用如图 7.6 所示的正态概率分布近似 $\bar{x}$ 的抽样分布。

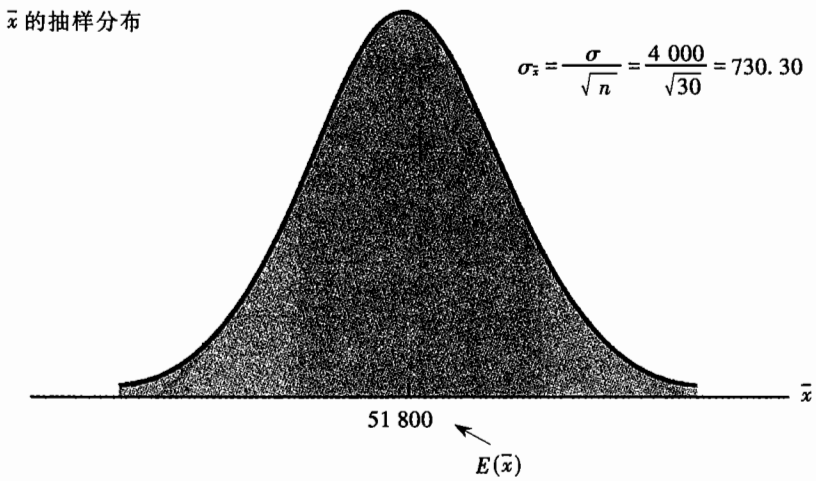


图 7.6　包括 30 名 EAI 经理的简单随机样本平均年薪的 $\bar{x}$ 抽样分布

$\bar{x}$ 抽样分布的实际值

当选取一个简单随机样本，并使用样本均值 $\bar{x}$ 的数值估计总体均值 μ 的数值时，我们不能期望样本均值准确地等于总体均值。正如前面所讲，样本均值 $\bar{x}$ 与总体均值 μ 的数值之差的绝对值：$|\bar{x} - \mu|$ 被称为**抽样误差**。我们对 $\bar{x}$ 的抽样分布之所以感兴趣，其实际原因是它能够对抽样误差的大小提供概率描述。为了说明 $\bar{x}$ 抽样分布的这种应用，我们再回到 EAI 问题中来。

假设人事主管认为，如果样本均值距离总体均值在 500 美元以内，那么就可以接受以样本均

值来估计总体均值。但是，要完全保证样本均值距总体均值在 500 美元以内又是不可能的。的确，表 7.5 和图 7.1 显示，在 500 个样本均值中有一部分与总体均值之间的差距超过了 2 000 美元。所以我们必须按照概率术语来考虑人事主管的要求，即人事主管关心的是如下问题：由 30 名 EAI 经理组成的简单随机样本的样本均值距总体均值在 500 美元以内的概率是多少？

由于我们已经确定了$\bar{x}$抽样分布的性质(见图 7.6)，我们就使用该分布来回答这个概率问题。参考图 7.7，人事主管询问的是样本均值介于 51 300 — 52 300 美元之间的概率。如果样本均值 $\bar{x}$ 的数值在该区间内，那么 $\bar{x}$ 的数值必定距总体均值在 500 美元以内。图 7.7 抽样分布中的阴影面积给出了本问题的恰当概率。因为抽样分布是正态的，均值为 51 800，标准差为 730.30，我们能够使用标准正态概率分布表来计算这个概率。在 $\bar{x}=51\,300$ 处，我们有

$$z=\frac{51\,300-51\,800}{730.30}=-0.68$$

查找标准正态概率分布表，我们得到 $z=0$ 与 $z=-0.68$ 之间的面积是 0.251 8。对于$\bar{x}=52\,300$，经过类似的计算，我们得到 $z=0$ 与 $z=+0.68$ 之间的面积也是 0.251 8。于是，样本均值的数值介于 51 300 — 52 300 之间的概率为 0.251 8 + 0.251 8 = 0.503 6。

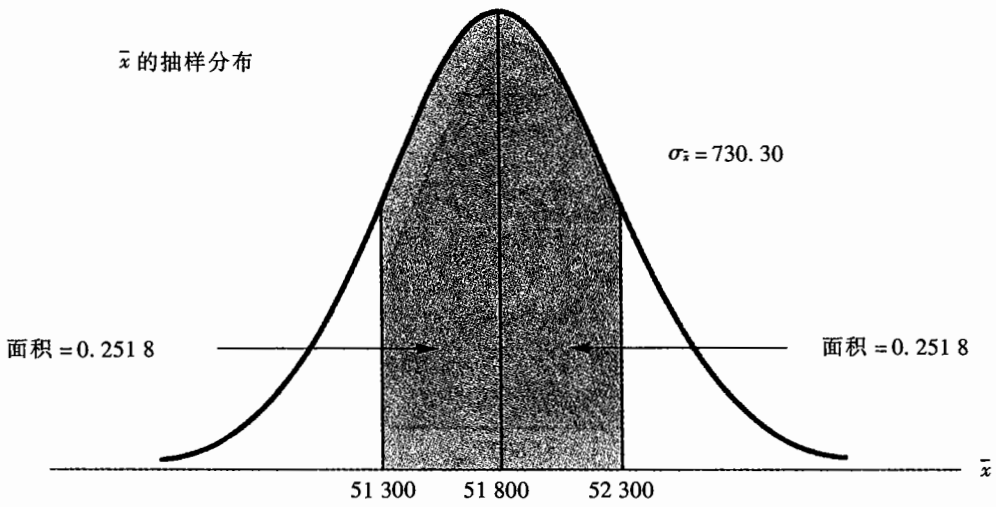


图 7.7　样本均值距总体均值在 500 美元以内的概率

以上的计算说明，30 名 EAI 经理组成的简单随机样本的样本均值 $\bar{x}$ 距总体均值在 500 美元以内的概率为 0.503 6。因此，样本均值距总体均值超过 500 美元的概率为 1 - 0.503 6 = 0.496 4。换句话说，由 30 名 EAI 经理组成的简单随机样本只有一半的可能使样本均值距离总体均值在可接受的 500 美元范围以内。也许应当考虑更大的样本容量，我们通过研究样本容量与 $\bar{x}$ 的抽样分布之间的关系来探索这个概率。

$\bar{x}$ 的抽样分布能够提供样本均值 $\bar{x}$ 与总体均值 μ 接近程度的概率信息。

$\bar{x}$ 抽样分布和样本容量之间的关系

假设在 EAI 抽样问题中，我们选取了 100 名经理而不是原来的 30 名，直观上，随着更大的样本容量提供了更多的数据，基于 $n=100$ 的样本均值对总体均值的估计应该能够比 30 人的样本均值所提供的总体均值估计更加优良。为了说明其优良的程度，我们来考虑样本容量与 $\bar{x}$ 的抽样分布之间的关系。

首先我们注意到无论样本容量如何，总有 $E(\bar{x})=\mu$ 。因此 $\bar{x}$ 所有可能值的平均值等于总体均值 μ，与样本容量 n 无关。但是，注意均值的标准误差 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$ 与样本容量的平方根有关。当样本容量增加时，均值的标准误差将减少。当 $n=30$ 时，EAI 问题的均值标准误差为 730.30。当样本容量增加到 $n=100$ 时，均值的标准误差将减少到

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}=\frac{4\,000}{\sqrt{100}}=400$$

$n=30$ 时和 $n=100$ 时的 $\bar{x}$ 抽样分布都显示在图 7.8 中。由于 $n=100$ 时的抽样分布具有更小的标准差，此时 $\bar{x}$ 的数值变异程度较小，因此比 $n=30$ 时的 $\bar{x}$ 值更接近总体均值。

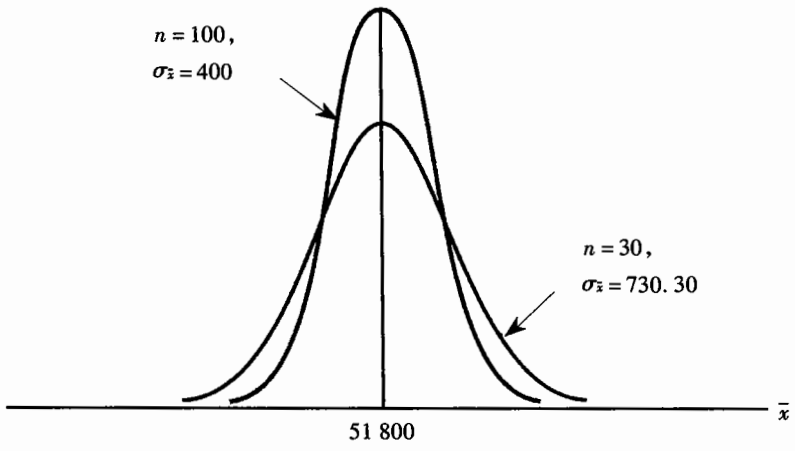


图 7.8　$n=30$ 名 EAI 经理与 $n=100$ 名 EAI 经理时简单随机样本的 $\bar{x}$ 抽样分布的相互比较

我们能够利用 $n=100$ 时的 $\bar{x}$ 抽样分布来计算由 100 名 EAI 经理组成的简单随机样本的样本均值距总体均值在 500 美元以内的概率。由于抽样分布是正态的，均值为 51 800，标准差为 400，我们可以使用标准正态概率分布表计算这个概率。在 $\bar{x}=51\,300$ 处(见图 7.9)，得到

$$z=\frac{51\,300-51\,800}{400}=-1.25$$

查找标准正态概率分布表，我们得到 $z=0$ 与 $z=-1.25$ 之间的面积是 0.394 4。对 $\bar{x}=52\,300$，经过类似的计算，我们看到样本均值的数值介于 51 300 — 52 300 之间的概率为 0.394 4 + 0.394 4 =

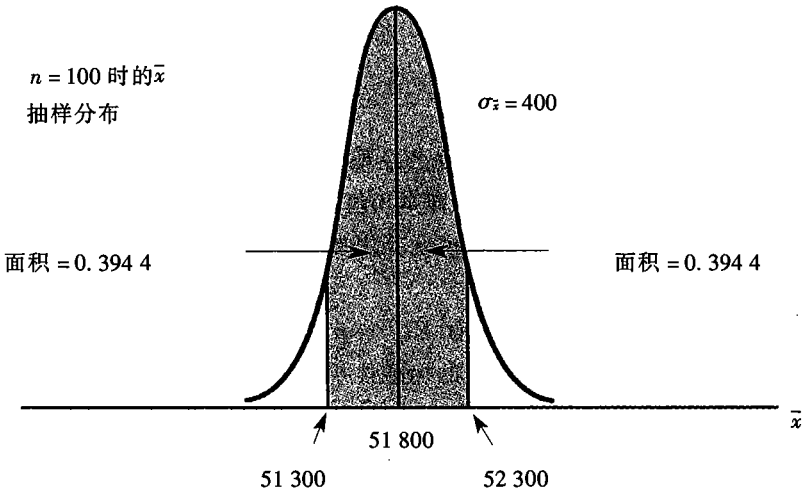


图 7.9 包括 100 名 EAI 经理的简单随机样本的样本均值距总体均值在 500 美元以内的概率图

0.788 8。因此，随着把样本容量从 30 增加到 100 人，样本均值距总体均值在 500 美元以内的概率也从 0.503 6 增加到 0.788 8。

该讨论的重点是当样本容量增加时，均值的标准误差减少。因此，随着样本容量的增加，样本均值距总体均值在某一特定距离内的概率也相应增加。

评 注

1. 在得出 EAI 问题的$\bar{x}$抽样分布的过程中，我们利用了已知总体均值 $\mu=51\,800$ 和总体标准差 $\sigma=4\,000$ 这一事实。但一般说来，确定$\bar{x}$抽样分布所需的总体均值 μ 和总体标准差 σ 的数值将是未知的。在第 8 章，我们将说明当 μ 和 σ 未知时，如何使用样本均值$\bar{x}$和样本标准差 s。
2. 对中心极限定理的理论证明要求样本中的各个观察值相互独立。对于无限总体和进行放回抽样的有限总体，所得到的样本满足该条件。虽然中心极限定理没有直接要求从有限总体的抽样是放回的，但在一般的统计应用中只有当总体容量较大时，才能够使用中心极限定理。

练 习

方 法

18. 某总体的均值为 200，标准差为 50。从中选取一个容量为 100 的简单随机样本，并且使用样本均值估计总体均值。

a. $\bar{x}$ 的数学期望是多少？

b. $\bar{x}$ 的标准差是多少？

c. 说明 $\bar{x}$ 的抽样分布。

d. $\bar{x}$ 的抽样分布表明了什么？

19. 某总体的均值为 200，标准差为 50。假设从中抽取一个容量为 100 的简单随机样本，并使用 $\bar{x}$ 估计 μ。



a. 样本均值距总体均值在 ±5 以内的概率是多少？

b. 样本均值距总体均值在 ±10 以内的概率是多少？

20. 假定总体标准差为 $\sigma = 25$。对于样本容量分别为 50，100，150 和 200 的样本，计算其均值的标准误差 $\sigma_{\bar{x}}$。当样本容量增加时，均值标准误差的大小将如何变化？

21. 假设某总体的标准差 $\sigma = 10$，从中抽取一个容量为 50 的简单随机样本。计算下面每种情况下的均值标准误差的数值(如果合适的话，使用有限总体修正系数)。

a. 总体容量无限。

b. 总体容量为 $N = 50\,000$。

c. 总体容量为 $N = 5\,000$。

d. 总体容量为 $N = 500$。

22. 某总体的均值为 400，标准差为 50，总体的概率分布未知。

a. 研究者使用分别包括 10，20，30 或 40 个元素的简单随机样本来收集有关总体的数据。在哪一个样本容量下，我们可以用正态概率分布来描述 $\bar{x}$ 的抽样分布？请解释。

b. 说明在适用正态概率分布的例子中 $\bar{x}$ 的抽样分布。

23. 某总体的均值为 100，标准差为 16。对于下面的每个样本容量，样本均值距总体均值在 ±2 以内的概率是多少？

a. $n = 50$。

b. $n = 100$。

c. $n = 200$。

d. $n = 400$。

e. 使用较大样本容量的优点是什么？

应　用

24. 参考 EAI 抽样问题。假设抽取的简单随机样本包括 60 名经理。

a. 当简单随机样本的样本容量为 60 时，画出 $\bar{x}$ 的抽样分布简图。

b. 如果简单随机样本的样本容量增加到 120，$\bar{x}$ 的抽样分布将发生什么变化？

c. 当样本容量增加时，你能够对 $\bar{x}$ 抽样分布的变化作出什么样的一般描述？这样的一般化概括符合逻辑吗？请解释。

25. 在EAI抽样问题中(见图7.7)，我们说明了对于 $n=30$，样本均值距总体均值在±500美元以内的概率为0.503 6。



a. 当样本容量增加到60时，$\bar{x}$ 距总体均值在500美元以内的概率是多少？

b. 当样本容量增加到120时，$\bar{x}$ 距总体均值在500美元以内的概率是多少？

26. 在美国出售的常规汽油的平均价格为每加仑1.20美元(*The Energy Information Administration*, March 3, 1997)。假定总体的每加仑平均价格为 $\mu=1.20$，总体标准差为 $\sigma=0.10$。假设选取50个加油站组成简单随机样本，使用从它们那里收集到的数据计算样本的平均每加仑汽油价格。

a. 以 $\bar{x}$ 表示50个加油站的样本平均每加仑汽油价格，说明样本均值 $\bar{x}$ 的抽样分布。

b. 由简单随机样本得出的样本均值距样本均值在0.02美元，总体均值在0.02以内的概率是多少？

c. 由简单随机样本得出的样本均值距样本均值在0.01美元，总体均值在0.01以内的概率是多少？

27. 根据美国大学考试程序委员会报告，SAT分数的总体均值为 $\mu=1\,017$ (*The New York Times*, 2000 Almanac)。假定总体标准差为 $\sigma=100$。

a. 由一个包括75名学生的简单随机样本得出的样本平均SAT分数距总体均值在10分以内的概率是多少？

b. 由一个包括75名学生的简单随机样本得出的样本平均SAT分数距总体均值在20分以内的概率是多少？

28. 市场营销专业毕业生的平均起始年薪为34 000美元(*Time*, May 8, 2000)。假定对于市场营销专业的毕业生总体，平均起始年薪为 $\mu=34\,000$ 美元，标准差为 $\sigma=2\,000$。

a. 对于样本容量分别为30，50，100，200和400的市场营销专业毕业生的简单随机样本，其样本均值距总体均值在±250美元以内的概率分别为多少？

b. 在试图估计总体均值时，较大样本容量的好处是什么？

29. 亚特兰大一套两卧室公寓的平均月租金为982美元(*Elle*, September 1998)。假定总体均值为982美元，总体标准差为210美元。

a. 由一个包括40套两卧室公寓的简单随机样本得出的样本平均月租金距总体均值在±100美元以内的概率是多少？

b. 由一个包括40套两卧室公寓的简单随机样本得出的样本平均月租金距总体均值在±25美元以内的概率是多少？

c. 讨论(a)和(b)中的结果。

30. 一种新式单身住宅的总体平均价格为166 500美元(*New One-Family Houses Sold*, U. S. Bureau of the Census, 1997)。假定总体标准差为42 000美元，样本中包括100套新式单身住宅。

a. 显示样本容量为100套新式单身住宅的样本平均价格抽样分布。

b. 该样本的平均价格距总体均值在10 000美元以内的100套新式单身住宅概率是多少？

c. 对于距离值5 000美元、2 500美元和1 000美元，重复(b)中的问题。

d. 为了在 ±2 500 美元或 ±1 000 美元的误差以内估计总体平均价格，你有什么建议？

31. 据《商业周刊》报道，它的订户为购买新车而计划花费的平均值为 27 100 美元（*Business Week*, Subscriber Profile, 1996）。假定《商业周刊》订户总体的平均计划购车价格为 $\mu=27\,100$ 美元，标准差为 $\sigma=5\,200$ 美元。
   a. 对于一个由 30 名订户组成的简单随机样本，其样本平均计划购车价格距总体均值在 1 000 美元以内的概率是多少？
   b. 对于一个由 50 名订户组成的简单随机样本，其样本平均计划购车价格距总体均值在 1 000 美元以内的概率是多少？
   c. 对于一个由 100 名订户组成的简单随机样本，其样本平均计划购车价格距总体均值在 1 000 美元以内的概率是多少？
   d. 如果要求样本均值距总体均值在 1 000 美元以内的概率至少为 0.90，你推荐的样本容量是 30，50 还是 100？

32. 为了估计 4 000 名员工总体的平均年龄，抽取 40 名员工组成简单随机样本。
   a. 在计算均值的标准误差时，你是否使用有限总体修正系数？请解释。
   b. 如果总体标准差为 $\sigma=8.2$ 岁，使用和不使用有限总体修正系数分别计算标准误差。当 $n/N\leqslant 0.05$ 时，忽略有限总体修正系数合理吗？
   c. 员工的样本平均年龄距总体平均年龄在 ±2 岁以内的概率是多少？

7.6　$\bar{p}$ 的抽样分布

在许多商务和经济应用中，我们使用样本比例 $\bar{p}$ 来对总体比例 p 进行统计推断。图 7.10 描绘了该过程。每重复一次该过程，我们可以预期会取得不同的样本比例 $\bar{p}$ 值。样本比例 $\bar{p}$ 的所有可能值的概率分布被称为样本比例 $\bar{p}$ 的抽样分布。

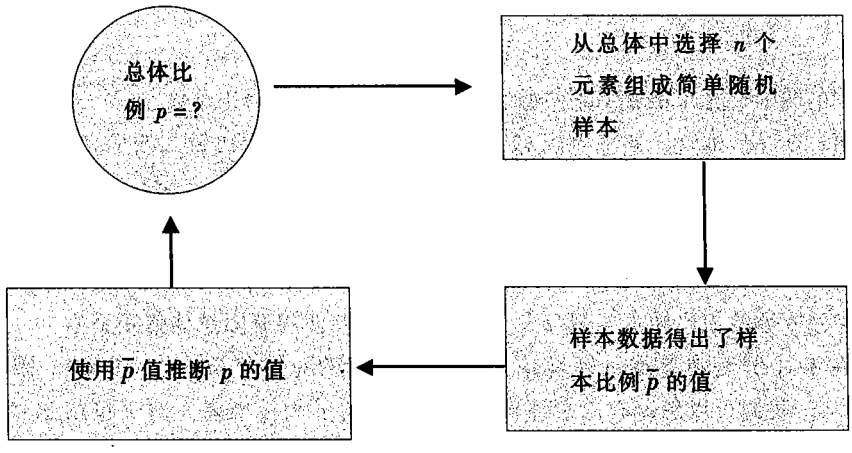


图 7.10　使用样本比例推断总体比例的统计过程

$\bar{p}$ 的抽样分布

$\bar{p}$ 的抽样分布是样本比例的所有可能值的概率分布。

为了确定样本比例$\bar{p}$与总体比例 p 的接近程度，我们需要理解$\bar{p}$抽样分布的特征：$\bar{p}$ 的数学期望、$\bar{p}$ 的标准差与$\bar{p}$ 抽样分布的形状。

$\bar{p}$ 的数学期望

$\bar{p}$ 的数学期望，即$\bar{p}$ 所有可能值的平均值，能够表达如下：

$\bar{p}$ 的数学期望

$$E(\bar{p}) = p \tag{7.4}$$

式中 $E(\bar{p})$——$\bar{p}$ 的数学期望；

p——总体比例。

公式(7.4)显示出$\bar{p}$ 所有可能值的平均值等于总体比例 p。回忆在 7.1 节，我们注意到 EAI 总体的 $p=0.60$，p 表示总体中参加过公司管理培训计划的经理所占的比例。于是，EAI 抽样问题中$\bar{p}$ 的数学期望为 0.60。

$\bar{p}$ 的标准差

$\bar{p}$ 的标准差又被称为**比例的标准误差**(standard error of the proportion)。和样本均值 $\bar{x}$ 一样，$\bar{p}$ 的标准差依赖于总体是有限的还是无限的。计算$\bar{p}$ 标准差的两种情况下的表达式如下：

$\bar{p}$ 的标准差

有限总体	无限总体	
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$	(7.5)

比较公式(7.5)中的两个表达式，我们看到两者惟一的不同是有限总体修正系数 $\sqrt{(N-n)/(N-1)}$。

和样本均值$\bar{x}$ 的情况相同，如果有限总体的容量与样本容量相比较大的话，那么有限总体和无限总体表达式之间的差别可以忽略。我们遵循在计算样本均值的标准误差时曾经使用过的经验法则，即如果总体是有限的且 $n/N \leqslant 0.05$，我们使用$\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$；但是如果总体是有限的且 $n/N > 0.05$ 的话，就应该使用有限总体修正系数，如公式(7.5)所示。除非特别注明，在本书中我们总是假定总体容量相比样本容量较大，因此没有必要使用有限总体修正系数。

我们知道在 EAI 研究中参加过管理培训计划的经理的总体比例为 $p=0.60$。因为 $n/N = 30/2\,500 = 0.012$，我们可以忽略有限总体修正系数来计算 $\bar{p}$ 的标准差。对于包括 30 名经理的简单随机样本，$\sigma_{\bar{p}}$ 为

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.089\ 4$$

$\bar{p}$ 抽样分布的形式

现在我们已知$\bar{p}$ 的均值和标准差，下一步就需要考虑$\bar{p}$ 抽样分布的形式。对于$\bar{p}$ 应用中心极限定理产生了下面的结果：

> 当样本容量较大时，$\bar{p}$ 的抽样分布可以用正态概率分布近似表示。

对于$\bar{p}$，在满足下列条件时，可以认为样本容量较大：

$$np \geqslant 5$$
$$n(1-p) \geqslant 5$$

在 EAI 抽样问题中，我们知道总体内参加过管理培训计划的经理所占比例为 $p = 0.60$。在简单随机样本的容量为 30 人时，我们有 $np = 30 \times 0.60 = 18$ 且 $n(1-p) = 30 \times 0.40 = 12$。因此，$\bar{p}$ 抽样分布可以用正态概率分布来近似，如图 7.11 所示。

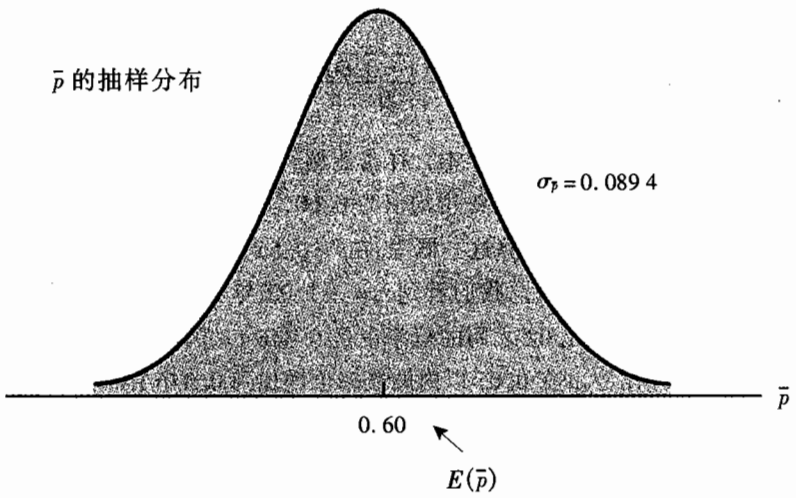


图 7.11　参加过管理培训计划的经理所占比例$\bar{p}$ 的抽样分布

$\bar{p}$ 抽样分布的实际值

当我们选择简单随机样本并且使用样本比例$\bar{p}$ 的值来估计总体比例 p 的值时，我们预期存在着一些抽样误差。这时，抽样误差是样本比例 $\bar{p}$ 与总体比例 p 之差的绝对值。我们可以利用 $\bar{p}$ 抽样分布的实际值对抽样误差进行概率描述。

假设在 EAI 问题中，人事主管想要知道 $\bar{p}$ 的值距离总体比例在 0.05 以内的概率是多少。即样本比例 $\bar{p}$ 介于 0.55—0.65 之间的概率是多少？图 7.12 阴影部分的面积显示了这个概率。利用

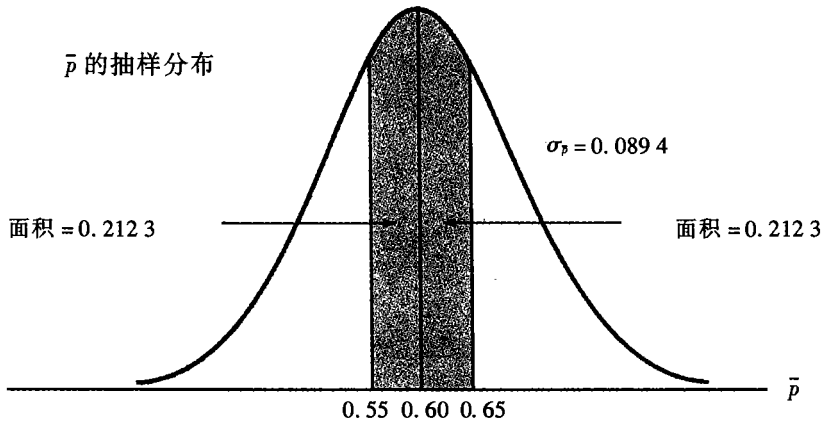


图 7.12　EAI 抽样问题的 $\bar{p}$ 抽样分布

$\bar{p}$ 的抽样分布能够被均值为 0.60、标准差为 $\sigma_{\bar{p}} = 0.089\,4$ 的正态概率分布近似表示这一事实，我们得到对应于 $\bar{p} = 0.55$ 的标准正态随机变量的 z 值为 $z = (0.55 - 0.60)/0.089\,4 = -0.56$。查找标准正态概率分布表，我们看到 $z = 0$ 与 $z = -0.56$ 之间的面积是 0.212 3。因此，样本比例 $\bar{p}$ 距总体比例 p 在 0.05 以内的概率为 0.212 3 + 0.212 3 = 0.424 6。

如果我们把样本容量增加到 100，比例的标准误差相应变为

$$\sigma_{\bar{p}} = \sqrt{\frac{0.60(1-0.60)}{100}} = 0.049\,0$$

对于样本容量为 100 名 EAI 经理的样本，样本比例 $\bar{p}$ 的值距总体比例在 0.05 以内的概率可以用同样方法计算出来。由于抽样分布能够用均值为 0.60，标准差为 0.049 0 的正态分布近似，故我们能够使用标准正态概率分布表计算这一概率。在 $\bar{p} = 0.55$ 处，可得 $z = (0.55 - 0.60)/0.049\,0 = -1.02$。查找标准正态概率分布表，我们看到 $z = -1.02$ 与 $z = 0$ 之间的面积是 0.346 1。类似地，在 $\bar{p} = 0.65$ 处，$z = 0$ 与 $z = 1.02$ 之间的概率也是 0.346 1。因此，如果样本容量从 30 增加到 100 的话，样本比例 $\bar{p}$ 总体比例 p 在 0.05 以内的概率也增加到 0.346 1 + 0.346 1 = 0.692 2。

练　习

方　法

33. 从比例为 $p = 0.40$ 的总体中抽取一个容量为 100 的简单随机样本。

a. $\bar{p}$ 的数学期望是多少？

b. $\bar{p}$ 的标准差是多少？

c. 显示 $\bar{p}$ 的抽样分布。

d. $\bar{p}$ 的抽样分布说明了什么？

34. 总体的比例为 0.40。从中抽取一个容量为 200 的简单随机样本并使用样本比例 $\bar{p}$ 来估计总体比例。



a. 样本比例距总体比例在 ±0.03 以内的概率是多少？

b. 样本比例距总体比例在 ±0.05 以内的概率是多少？

35. 假定总体的比例为 0.55。对于容量为 100、200、500 和 1 000 的样本，分别计算比例的标准误差。当样本增加时，比例标准误差的大小将如何变化？

36. 总体的比例为 0.30。对于容量分别等于下列数字的样本，其样本比例距总体比例在 ±0.04 以内的概率是多少？

a. $n = 100$

b. $n = 200$

c. $n = 500$

d. $n = 1\,000$

e. 取较大容量的样本有什么好处？

应　用

37. Doerman Distributors 公司的总裁认为，公司 30% 的订单来自新客户。使用一个包括 100 份订单的简单随机样本来估计其中新客户所占的比例。



a. 假定总裁的看法是正确的，即 $p = 0.30$。本研究中 $\bar{p}$ 的抽样分布是什么？

b. 样本比例 $\bar{p}$ 介于 0.20—0.40 之间的概率是多少？

c. 样本比例 $\bar{p}$ 介于 0.25—0.35 之间的概率是多少？

38. 根据美国百货制造商的报告，有 76% 的顾客会阅读产品标签所列示的配方。假定总体比例为 $p = 0.76$，从总体中抽取一个包括 400 名顾客的样本。

a. 以 $\bar{p}$ 表示样本中阅读过产品标签上配方的顾客比例，显示样本比例 $\bar{p}$ 的抽样分布。

b. 样本比例距总体比例在 ±0.03 以内的概率是多少？

c. 对于包括 750 名顾客的样本，回答(b)中的问题。

39. 《时代》/CNN 选民民意测验监测 2000 年美国总统大选期间公众对总统候选人的态度。由 Yankelovich Partners 公司进行的一项《时代》/CNN 民意测验选取 589 名可能的选民作为样本（*Time*, June 26, 2000）。假定在可能的选民中支持某位候选人的总体比例为 $p = 0.50$，令 $\bar{p}$ 表示支持该候选人的样本比例。

a. 显示 $\bar{p}$ 的抽样分布。

b. 《时代》/CNN 民意测验所提供的样本比例距总体比例在 ±0.04 以内的概率是多少？

c. 样本比例距总体比例在 ±0.03 以内的概率是多少？

d. 样本比例距总体比例在 ±0.02 以内的概率是多少？

40. 虽然大多数人都认为早餐是一天中最重要的一餐，但有25%的成年人却不吃早餐(*U. S. News & World Report*, November 10, 1997)。假定不吃早餐的总体比例为 $p=0.25$，$\bar{p}$ 是在一个包括200名成年人的样本中不吃早餐的样本比例。
    a. 显示 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ± 0.03 以内的概率是多少？
    c. 样本比例距总体比例在 ± 0.05 以内的概率是多少？

41. 根据妇女政策研究机构的报告，所有工会会员中妇女占了37%的比例。假设工会会员中妇女所占的总体比例为 $p=0.37$，选取1 000名工会会员组成简单随机样本。
    a. 显示样本中妇女所占比例 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ± 0.03 以内的概率是多少？
    c. 对于容量为500的简单随机样本，回答(b)中的问题。

42. 假定一个装配线的次品率为15%，但是该公司的生产经理尚不知道这个情况。假设公司的质量保证部门选取了50个部件进行检验以确定装配线的运行质量，令 $\bar{p}$ 表示通过质量检验在样本中发现的次品比例。
    a. 显示 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ± 0.03 以内的概率是多少？
    c. 如果检验显示 $\bar{p}=0.10$ 或更高，就应该关闭该装配线检查造成次品的原因。由50个部件组成的样本会导致作出关闭装配线决定的概率是多少？

43. 食品营销研究机构发现，17%的家庭每周购买杂货的支出超过100美元。假定该总体比例为 $p=0.17$，并且从总体中选取800户家庭组成简单随机样本。
    a. 以 $\bar{p}$ 表示每周购买杂货的支出超过100美元的样本比例，显示 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ± 0.02 以内的概率是多少？
    c. 对于容量为1 600户家庭的样本，回答(b)中的问题。

7.7 点估计量的性质

本章我们已经说明了如何把诸如样本均值 $\bar{x}$、样本标准差 s 和样本比例 $\bar{p}$ 这样的样本统计量作为与它们相对应的总体参数 μ、σ 和 p 的点估计量来使用。但是，在将样本统计量作为点估计量之前，应该先对它们进行统计检验，以确保这个样本统计量是否具有良好的点估计量所应具备的某些性质。本节我们将讨论良好的估计量所应具备的性质：无偏性、有效性和一致性。

因为存在着几种不同的样本统计量，我们将在本节使用下列通用符号：

θ = 有关的总体参数；

$\hat{\theta}$ = θ 的样本统计量或点估计量。

符号 θ 是希腊字母，读作“theta”，符号 $\hat{\theta}$ 读作“theta 尖”。通常，θ 代表任意的总体参数，比如总体均值、总体标准差和总体比例等等。$\hat{\theta}$ 代表对应的样本统计量，比如样本均值、样本标准差和样本比例等等。

无偏性

如果样本统计量的数学期望等于它所估计的总体参数，该样本统计量就被称为是总体参数的**无偏性**(unbiasedness)。无偏性的定义如下：

无偏性

如果　　$E(\hat{\theta}) = \theta$　　**(7.6)**

则样本统计量 $\hat{\theta}$ 是总体参数 θ 的无偏估计量。式中的 $E(\hat{\theta})$ 为样本统计量 $\hat{\theta}$ 的数学期望。

于是，无偏样本统计量的所有可能值的数学期望，或者说是均值等于它所估计的总体参数。

图 7.13 显示了无偏和有偏点估计量的情况。在说明无偏点估计量的图中，抽样分布的均值等于总体参数的值。在这种情形下，因为点估计量的值有时小于 $\hat{\theta}$，有时大于 θ，所以抽样误差相互抵消。而在有偏估计量的情况下，抽样分布的均值小于或大于总体参数的值。在图 7.13 的(b)图中 $E(\hat{\theta})$ 大于 θ，因此样本统计量过高估计总体参数值的概率很大，偏离的大小也显示在图中。

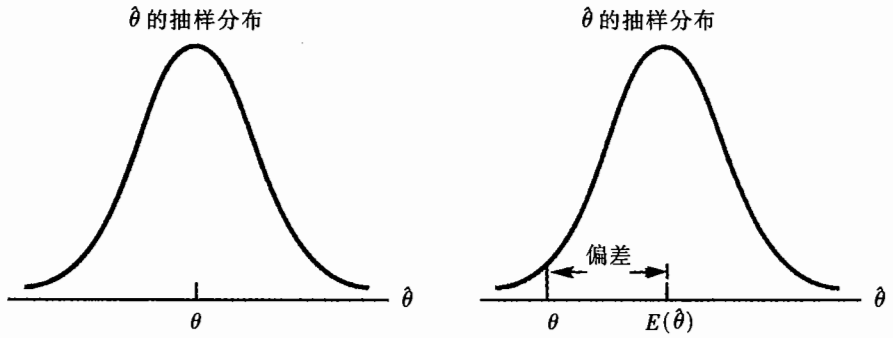


图 7.13　无偏和有偏点估计量的例子

在讨论样本均值和样本比例的抽样分布时，我们给出了 $E(\bar{x}) = \mu$ 和 $E(\bar{p}) = p$。于是，$\bar{x}$ 和 $\bar{p}$ 是相应的总体参数 μ 和 p 的无偏估计量。

我们在第 11 章再对样本标准差 s 和样本方差 s^2 的抽样分布进行更详细的讨论。不过我们可以证明 $E(s^2) = \sigma^2$，因此样本方差 s^2 是总体方差 σ^2 的无偏估计量。实际上，当我们在第 3 章首

次给出样本方差和样本标准差的公式时，分母是 $n-1$ 而不是 n，其原因正是为了使样本方差成为总体方差的无偏估计量。如果我们在分母中使用了 n，则样本方差为总体方差的有偏估计量，倾向于轻微地低估总体方差。

有效性

假设我们能够使用一个 n 元素的简单随机样本对同一个总体参数给出两个无偏的点估计量。在这种情况下，因为具有较小方差的点估计量的值会更加接近总体参数，故我们更愿意使用具有较小方差的点估计量。它被称为比其他点估计量具有更大的**相对有效性**(relative efficiency)。

在从正态总体抽样时，样本均值的标准差比样本中位数的标准差小，因此样本均值比样本中位数有效。

图 7.14 显示了两个无偏的点估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的抽样分布。注意到 $\hat{\theta}_1$ 的标准差小于 $\hat{\theta}_2$ 的标准差，因此 $\hat{\theta}_1$ 的值相比 $\hat{\theta}_2$ 的值具有更大的可能性接近于参数 θ。由于点估计量 $\hat{\theta}_1$ 的标准差小于点估计量 $\hat{\theta}_2$ 的标准差，$\hat{\theta}_1$ 相对 $\hat{\theta}_2$ 就更加有效，是更好的点估计量。

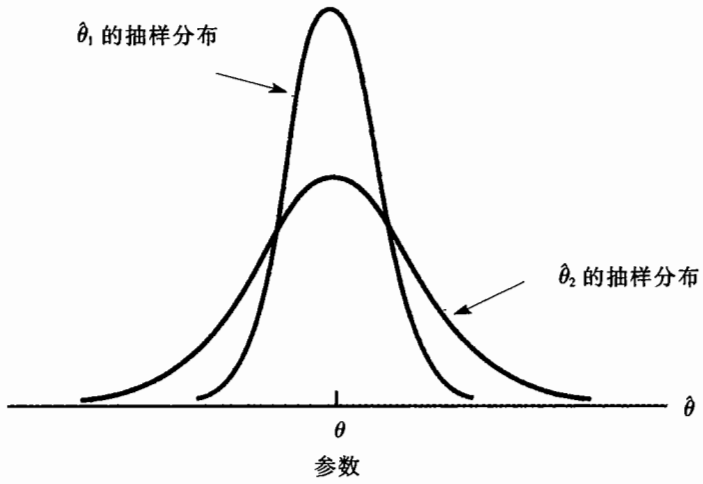


图 7.14　两个无偏点估计量的抽样分布

一致性

良好的估计量所需要的第三个性质是**一致性**(consistency)。粗略地说，如果当样本容量增加时，点估计量的值有接近于总体参数的趋势，则点估计量是一致的。换句话说，大样本比小样本更易于得出较好的点估计值。注意到对于样本均值 $\bar{x}$，它的标准差由 $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ 给出。由于 $\sigma_{\bar{x}}$ 与样本容量有关，故用较大的样本容量得到的 $\sigma_{\bar{x}}$ 值更小，我们能够得出结论，较大的样本容量趋向

于给出更接近于总体均值 μ 的点估计值。在这个意义上，我们可以说样本均值 $\bar{x}$ 是总体均值 μ 的一个一致估计量。同样道理，我们还可以得出样本比例 $\bar{p}$ 是总体比例 p 的一个一致估计量。

评　注

在第 3 章，我们介绍过均值和中位数是度量中心位置的两种方法。本章我们将只讨论均值，原因是当我们从正态总体抽样时，正态总体的均值与中位数相同，但中位数的标准误差大约比均值的标准误差大 25%。回忆在 EAI 问题中，$n=30$，均值的标准误差 $\sigma_{\bar{x}}=730.30$。而该问题的中位数标准误差大约为 $1.25\times 730.30=913$，因此样本均值更有效，且距离总体在某特定范围内的概率更大。

7.8　其他抽样方法

我们已经描述了简单随机抽样方法，并且讨论了在使用简单随机抽样方法时，$\bar{x}$ 和 $\bar{p}$ 抽样分布的特征。但是，简单随机抽样并不是惟一可用的抽样方法，我们还可以选择分层随机抽样、整群抽样和系统抽样，在某些情况下，它们要优于简单随机抽样。本节我们将对这些抽样方法加以简介。

本小节是对抽样方法的一个简介，而不仅是简单的随机抽样。

分层随机抽样

在**分层随机抽样**(stratified random sampling)中，首先把总体内的元素分隔为若干个被称为**层**的组，使得总体内的每个元素属于且只属于一个层。样本设计者可以按照部门、位置、年龄和产业种类等等来划分层。但是，只有当每一层内的元素都尽可能地相似时，取得的效果最佳。图 7.15 是一个把总体分成 H 层的图示。

当每一层内元素之间的方差相对较小时，分层随机抽样效果最佳。

在分层以后，从每一层选取一个简单随机样本。然后利用公式把单个层的样本结果联系起来，估计有关的总体参数。分层随机抽样的值依赖于层内元素的同质性。如果层内的元素是相似的(同质的)，则该层的方差较小，因此采用相对较小的样本容量就能够取得对该层特征的良好估计。如果各层是同质的，那么利用分层随机抽样方法能够以较小的样本容量获得与简单随机抽样同样准确的结果。

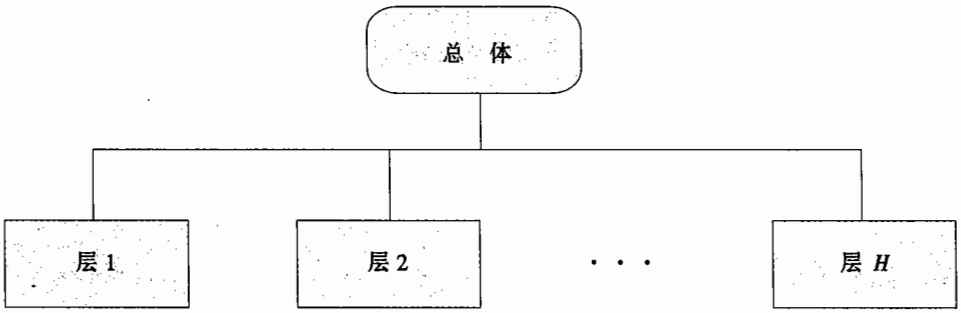


图 7.15 分层随机抽样的图示

整群抽样

在**整群抽样**(cluster sampling)中，首先把总体内的元素分隔为若干个叫做**群**的相互独立的组，总体内的每个元素属于且只属于一个群(见图 7.16)。然后选取一个以群为元素的简单随机样本，每个被选中的群内的所有元素共同组成了样本。当群内的元素不相似时，整群抽样的效果会提高。在理想的情况下，每一群都在较小的规模上代表总体。整群抽样的值依赖于每一群对整个总体的代表性。如果所有的群在代表总体的意义上是同质的，则抽取少量的群就可以得到总体参数的良好估计。

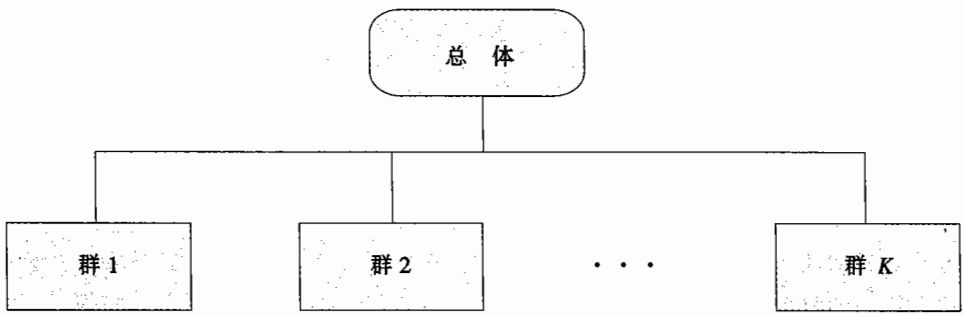


图 7.16 整群抽样的图示

当每一群都与总体相似时，整群抽样的效果最佳。

整群抽样最主要的应用之一是区域抽样，其中的群是街区或其他定义好的区域。整群抽样通常要比简单随机抽样和分层随机抽样所要求的样本容量大。但是，由于当派出一个调查员去一个样本群(如城市路口位置)时，可以在相对较短的时间内获得许多样本观察值，从而使用整群抽样节约了成本，因此，整群抽样虽然样本容量较大，但有时总成本却较低。

系统抽样

在一些抽样情况下，尤其是大容量总体的情况，如果通过先找到随机数，然后计数或在总体清单中寻找直到发现与随机数相对应的元素这一系列程序来选取简单随机样本，是很耗时间的。简单随机抽样的一个替代方法是**系统抽样**(systematic sampling)。例如，如果要求从一个包括 5 000 个元素的总体中选取一个容量为 50 的样本，我们从总体每 5 000/50 = 100 个元素中选出一个作为样本元素。在这种情况下，系统样本就是在总体清单的前 100 个元素中随机地选取一个，在确定其他样本元素时，从第一个已选出的元素开始，依次向后，选取已选出元素后面的第 100 个元素放入样本。实际上，我们可以通过在总体中系统地移动，从第一个已选出的元素开始，向后每隔 100 个元素选取一个作为样本点。以这种方法来确定一个容量为 50 的样本通常比用简单随机抽样方法容易些。由于选取的第一个元素是随机选择的，系统样本通常被假定具有简单随机样本的性质。尤其是当总体中的元素清单是按照随机顺序排列时，这个假设更适用。

方便抽样

迄今为止所讨论的抽样方法都被称为是**概率抽样**技术，即从总体中选出的元素都是以一个已知的概率被选入样本。概率抽样的优点在于通常能够确定出样本统计量的抽样分布。这样我们就能够使用本章介绍的简单随机抽样的公式来确定抽样分布的特征，然后利用抽样分布对结果的抽样误差进行概率描述。

方便抽样(Convenience sampling)是一种**非概率抽样**技术。就像它的名字一样，主要是通过简便的方法来确定样本。被选入样本的元素没有事先规定或已知的选取概率。例如，某教授在一所大学进行一 项研究，由于学生志愿者是现成的且加入的成本很低，故教授仅仅使用学生志愿者组成了样本。类似地，一个检验员可以从许多箱子中随意地选取橙子，这是因为给每个橙子贴上标签并且使用概率方法对橙子抽样是不现实的。像野生动物营地和消费者研究中的志愿者小组等也是方便样本。

方便样本具有相对易于选择样本和收集数据的优点，但是我们无法按照它对总体的代表程度评价方便样本的优良性。方便样本可能得到好的结果也可能得不到，没有统计上公认的方法对样本结果的质量进行分析和推断。某些研究者把为概率样本设计的统计方法应用于方便样本，他们认为这时能够把方便样本作为概率样本来处理。不过我们不支持这种说法，在使用方便样本的结果对总体进行推断和解释时应该保持谨慎。

判断抽样

另一个非概率抽样技术是**判断抽样**(judgment sampling)。在这个方法中，由对所研究的总体非常了解的人选择他或她认为最能代表总体的元素组成样本。往往这是一个相对容易的选择样本方法。例如，报告者可抽取两个或三个参议员，判断他们是否反映了所有参议员的一般意见。但

是，样本结果的质量依赖于选择样本的个人的判断能力。因此在判断抽样的基础上对总体的推断得出结论时，必须十分谨慎。

评 注

我们推荐使用以下概率抽样方法：简单随机抽样、分层随机抽样、整群抽样以及系统抽样。对于这些方法，在按照样本结果与它所估计的总体特征的接近程度这个标准评价样本结果的优良性时，可以运用公式进行量化研究。而在使用方便抽样和判断抽样时则无法对样本结果的优良性进行评价。因此对于非概率抽样方法，在解释它的抽样结果时必须十分谨慎。

小 结

本章我们介绍了简单随机抽样和抽样分布的概念。我们举例说明了如何选择简单随机样本和如何使用收集到的样本数据来作出总体参数的点估计。由于不同的简单随机样本给出了点估计量的不同值，因此诸如$\bar{x}$和$\bar{p}$这样的点估计量都是随机变量。这些随机变量的概率分布被称为抽样分布。特别地，我们描述了样本均值$\bar{x}$和样本比例$\bar{p}$的抽样分布。

在考虑$\bar{x}$和$\bar{p}$抽样分布的特征时，我们给出了$E(\bar{x})=\mu$和$E(\bar{p})=p$。在建立了这些统计量的标准差或者说是标准误差公式以后，我们给出了中心极限定理，该定理为在大样本情况下用正态概率分布来近似抽样分布提供了理论基础。我们还给出了满足大样本条件的经验法则。在本章的最后部分，我们讨论了包括分层随机抽样、整群抽样、系统抽样、方便抽样和判断抽样在内的其他抽样方法。

术语释义

参数：总体的数值特征，如总体均值μ、总体标准差σ和总体比例p等等。

简单随机抽样：在有限总体的情况下，选择样本时应使每一容量为n的可能样本被选中的概率相同；在无限总体的情况下，选择样本时应使样本内的每个元素都来自同一个样本并且对它们的选择是互相独立的。这样的抽样方法称为简单随机抽样。

无放回抽样：一旦把一个元素选入样本以后，就把它从总体中删除，使之不能被再次选入样本的抽样方法。

放回抽样：把一个元素选入样本后，再将它放回总体的抽样方法。这样一个以前选择过的元素可以被再次选入样本。

样本统计量：就是样本特征，比如样本均值$\bar{x}$、样本标准差s和样本比例$\bar{p}$等等。样本统计量的值可用于估计总体参数的值。

点估计值：用作总体参数估计值的数值。

点估计量：即样本统计量，如$\bar{x}$、s 或$\bar{p}$，它能够给出总体参数的点估计。

抽样误差：无偏点估计量的值与对应的总体参数值之差的绝对值。对于样本均值、样本标准差与样本比例，抽样误差分别为$|\bar{x}-\mu|$、$|s-\sigma|$与$|\bar{p}-p|$。

抽样分布：由样本统计量的所有可能值构成的概率分布。

有限总体修正系数：当从有限总体而不是无限总体抽样时，在计算$\sigma_{\bar{x}}$和$\sigma_{\bar{p}}$的公式中使用的修正系数 $\sqrt{(N-n)/(N-1)}$。当 $n/N \leqslant 0.05$ 时，一般公认的经验法则可以忽略有限总体修正系数。

标准误差：点估计量的标准差。

中心极限定理：该定理使人们在样本容量较大时，能够使用正态概率分布来近似$\bar{x}$和$\bar{p}$的抽样分布。

无偏性：如果点估计量的数学期望等于它所估计的总体参数，则该点估计量具有无偏性。

相对有效性：给定同一总体参数的两个无偏点估计量，具有较小标准差的点估计量就是更有效的。

一致性：如果一个点估计量随着样本容量的增加，得到的点估计值趋向于总体参数值，则该点估计量具有一致性。

分层随机抽样：一种概率抽样方法，总体首先被分隔为若干层，然后从每一层内选取简单随机样本。

整群抽样：一种概率抽样方法，首先把总体分隔为若干群，然后再选取一个以群为元素的简单随机样本。

系统抽样：一种概率抽样方法，首先在前 k 个元素中随机地选取一个，然后依次把选中元素后面的第 k 个元素选入样本。

方便抽样：一种非概率抽样方法，按照方便的原则选取元素组成样本。

判断抽样：一种非概率抽样方法，根据研究者的判断选择元素组成样本。

重要公式

$\bar{x}$ 的数学期望

$$E(\bar{x}) = \mu \tag{7.1}$$

$\bar{x}$ 的标准差

有限总体	无限总体	
$\sigma_{\bar{x}} = \sqrt{\dfrac{N-n}{N-1}}\left(\dfrac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = \left(\dfrac{\sigma}{\sqrt{n}}\right)$	(7.2)

$\bar{p}$ 的数学期望

$$E(\bar{p}) = p \tag{7.4}$$

$\bar{p}$ 的标准差

有限总体	无限总体	
$\sigma_{\bar{p}}=\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}}=\sqrt{\frac{p(1-p)}{n}}$	(7.5)

补充练习

44. 《商业周刊》的公司业绩公告版提供了899家公司以下指标的季度数据：销售额、利润、净收入、权益回报率、市盈率以及每股收益(*Business Week*, August 14, 2000)。按照它们出现在公司业绩公告版上的顺序把这些公司分别编号为从1到899。我们从表7.1第二列底部的随机数开始，忽略每组随机数的前两位，使用三位随机数并且从112开始，按列**向上**读取，确定前面的8家公司组成简单随机样本。

45. 青少年平均观看电视的时间是每天3小时(*Barron's*, November 8, 1999)。假定总体均值为 $\mu=3$，总体标准差为 $\sigma=1.2$ 小时。假设使用一个包括50名青少年的样本来监测青少年总体的观看电视时间。令 $\bar{x}$ 表示样本平均观看时间。
    a. 显示 $\bar{x}$ 的抽样分布。
    b. 样本均值距总体均值在 ± 0.25 小时以内的概率是多少？

46. 芝加哥居民上班途中所花费的平均时间为31.5分钟(*1998 Information Please Almanac*)。假定总体均值为 $\mu=31.5$ 分钟，总体标准差为 $\sigma=12$ 分钟。选取50名芝加哥居民组成简单随机样本。
    a. 令 $\bar{x}$ 表示50名芝加哥居民的平均上班途中时间，显示 $\bar{x}$ 的抽样分布。
    b. 样本均值距总体均值在 ± 1 分钟以内的概率是多少？
    c. 样本均值距总体均值在 ± 3 分钟以内的概率是多少？

47. 根据美国劳工统计局的报告，行政管理岗位的平均小时工资率为24.07美元(*The Wall Street Journal Almanac*, *1998*)。假定总体均值为 $\mu=24.07$ 美元，总体标准差为 $\sigma=4.80$ 美元。选取120名从事行政管理工作的人士组成样本。
    a. 样本均值距总体均值在 ± 0.50 美元以内的概率是多少？
    b. 样本均值距总体均值在 ± 1.00 美元以内的概率是多少？

48. 根据《今日美国》(April 11, 1995)的报道，商务旅行者每年因商务在外的平均天数为115天，标准差为每年60天。假定该结果适用于商务旅行者总体，从总体中选取50人组成样本。
    a. 均值标准误差的值是多少？
    b. 样本均值超过每年115天的概率是多少？
    c. 样本均值距总体均值在 ± 5 天以内的概率是多少？
    d. 如果样本容量增加到100，(c)中的概率将如何变化？

49. 三家公司的存货数量不同。A 公司的总体为 2 000 件，B 公司的总体为 5 000 件，C 公司的总体为 10 000 件。这些存货成本的总体标准差为 $\sigma = 144$。一个统计顾问建议每个公司都从它的总体中选取 50 件存货组成样本，以得出每件存货的平均成本在统计上有效的估计值。但小公司的经理表示，由于他的公司存货总体最小，因此其采集数据的样本应该比存货数量大的公司所要求的样本小得多。而顾问则认为，为了在样本结果中得到同样的标准误差和同样的精度，所有公司无论存货总体大小都应使用相同的样本容量。
    a. 使用有限总体修正系数，计算样本容量都是 50 的每一家公司存货样本的标准误差。
    b. 对于每一家公司，其样本均值 $\bar{x}$ 距总体均值 μ 在 ±25 以内的概率是多少？

50. 一个研究者报告的调查结果表明，均值的标准误差为 20。已知总体标准差为 500。
    a. 在该调查中使用了多大的样本？
    b. 估计值距总体均值在 ±25 以内的概率是多少？

51. 某质量控制检查员定期地检查一个生产工艺。该检查员选取 30 件已完工产品组成简单随机样本，并计算样本平均产品重量 $\bar{x}$。如果长期的检验结果显示有 5% 的 $\bar{x}$ 值高于 2.1 磅，有 5% 的 $\bar{x}$ 值低于 1.9 磅，问该工艺所生产产品总体的平均重量和标准差是多少？

52. “对于商务旅行者而言，当他们呆在宾馆时，最重要的因素是什么？”根据《今日美国》的报道，有 74% 的商务旅行者表示拥有可以自由吸烟的房间是最重要的因素（*USA Today*，April 11, 1995）。假定总体比例为 $p = 0.74$，并选取 200 名商务旅行者组成样本。
    a. 令 $\bar{p}$ 表示认为自由吸烟房间是他们呆在宾馆时最重要因素的商务旅行者的样本比例，显示 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ±0.04 以内的概率是多少？
    c. 样本比例距总体比例在 ±0.02 以内的概率是多少？

53. 某市场研究公司进行的电话调查在历史上的回复率为 40%，问在一个包括 400 个电话号码的新样本中，至少有 150 人回答问题的概率是多少？换句话说，样本比例至少为 150/400 = 0.375 的概率是多少？

54. 根据 ORC 国际公司的数据，有 71% 的互联网用户使用普通电话线联结计算机与互联网（*USA Today*，January 18, 2000）。假定总体比例为 $p = 0.71$。
    a. 由 350 名互联网用户组成的简单随机样本的样本比例距总体比例在 ±0.05 以内的概率是多少？
    b. 由 350 名互联网用户组成的简单随机样本比例大于等于 0.75 的概率是多少？

55. 在向 All-Driver 汽车保险公司投保的个人中，5 年内至少收到一张交通事故传票的比例为 0.15。
    a. 如果使用一个由 150 名投保人组成的简单随机样本来估计至少收到一张传票的总体比例，显示 $\bar{p}$ 的抽样分布。
    b. 样本比例距总体比例在 ±0.03 以内的概率是多少？

56. Lori Jeffrey 是一家大学课本主要出版商的成功销售代表。以前 Lori 打过的销售电话中，有 25% 获得了成功。观察她一个月的销售电话作为所有可能销售电话的样本。假定对数据进行的统计分析表明比例的标准误差为 0.062 5。
    a. 在该分析中采用了多大的样本？即在这个月内 Lori 打了多少次销售电话？
    b. 令 $\bar{p}$ 表示该月内取得成功的样本比例，显示 $\bar{p}$ 的抽样分布。
    c. 使用 $\bar{p}$ 的抽样分布，计算在一个月的期间内，Lori 的销售电话取得成功的比例大于等于 30% 的概率。

附录 7.1 $\bar{x}$ 的数学期望与标准差

在该附录中，我们将介绍 $E(\bar{x})$ 和 $\sigma_{\bar{x}}$ 表达式的数学证明。其中，$E(\bar{x})$ 由公式（7.1）给出，是 $\bar{x}$ 的数学期望；$\sigma_{\bar{x}}$ 由公式（7.2）给出，是 $\bar{x}$ 的标准差。

$\bar{x}$ 的数学期望

假定总体的均值为 μ，方差为 σ^2。从中抽取一个容量为 n 的简单随机样本，样本的各个观察值可表示为 $x_1, x_2, \ldots, x_n$。则样本均值 $\bar{x}$ 可计算如下：

$$\bar{x} = \frac{\sum x_i}{n}$$

重复选择容量为 n 的简单随机样本的过程，$\bar{x}$ 是一个随机变量，其取值依赖于选入样本的特定 n 个元素。随机变量 $\bar{x}$ 的数学期望就是 $\bar{x}$ 所有可能取值的平均值。

$$\bar{x} \text{ 的均值} = E(\bar{x}) = E\left(\frac{\sum x_i}{n}\right)$$

$$= \frac{1}{n}[E(x_1 + x_2 + \cdots + x_n)]$$

$$= \frac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)]$$

对于任意的 x_i，我们有 $E(x_i) = \mu$，因此

$$E(\bar{x}) = \frac{1}{n}(\mu + \mu + \cdots + \mu)$$

$$= \frac{1}{n}(n\mu) = \mu$$

这个结果说明，$\bar{x}$ 所有可能值的平均值等于总体均值 μ，即 $E(\bar{x}) = \mu$。

$\bar{x}$ 的标准差

我们再次假定总体的均值为 μ，方差为 σ^2，样本均值由下式给出：

$$\bar{x} = \frac{\sum x_i}{n}$$

重复选择容量为 n 的简单随机样本的过程，我们知道 $\bar{x}$ 是一个随机变量，其取值依赖于选入样本的特定 n 个元素。下面的内容推导了在无限总体情况下 $\bar{x}$ 值的标准差 $\sigma_{\bar{x}}$ 的表达式。而在有限总体无放回抽样情况下对 $\sigma_{\bar{x}}$ 表达式的推导过程则更加困难，超出了本书的范围。

回到无限总体情况，来自无限总体的简单随机样本由互相独立的观察值 $x_1, x_2, \ldots, x_n$ 构成。下面的两个表达式是计算随机变量方差的通用公式：

$$\mathrm{Var}(ax) = a^2\mathrm{Var}(x) \tag{7.7}$$

式中的 a 是常数，x 是随机变量。

$$\mathrm{Var}(x + y) = \mathrm{Var}(x) + \mathrm{Var}(y) \tag{7.8}$$

式中的 x 和 y 是独立的随机变量。利用公式(7.7)和(7.8)，我们能够得到随机变量的表达式如下：

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{\sum x_i}{n}\right) = \mathrm{Var}\left(\frac{1}{n}\sum x_i\right)$$

利用常数为 $1/n$ 的公式(7.7)，我们有

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)\mathrm{Var}\left(\sum x_i\right)$$

$$= \left(\frac{1}{n}\right)^2 \mathrm{Var}(x_1 + x_2 + \cdots + x_n)$$

在无限总体情况下，随机变量 $x_1, x_2, \ldots, x_n$ 是独立的，因此由公式(7.8)能够得到

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 [\mathrm{Var}(x_1) + \mathrm{Var}(x_2) + \cdots + \mathrm{Var}(x_n)]$$

对于任意的 x_i，我们有 $\mathrm{Var}(x_i) = \sigma^2$，因此

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

在这一表示中有 n 个 σ^2 值，我们就有

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \frac{\sigma^2}{n}$$

对上式取平方根，即可得出 $\bar{x}$ 的标准差公式

$$\sigma_{\bar{x}} = \sqrt{\mathrm{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

附录 7.2 利用 Minitab 进行随机抽样

如果在 Minitab 文件中含有一个总体的清单，那么我们就可以使用 Minitab 软件从该总体中选择简单随机样本。例如，在数据集 MetAreas 的第 1 列是美国和加拿大前 100 大的城市区域(*Places Rated Almanac — The Millennium Edition 2000*)。第 2 列是每个城市区域的综合评分。数据集中前 10 个城市区域以及对应的评分列示在表 7.6 中。

表 7.6 数据集 MetAreas 中前 10 个城市区域的综合评分



CD 光盘数据 MetAreas

城市区域	评分
Albany, NY	64.18
Albuquerque, NM	66.16
Appleton, WI	60.56
Atlanta, GA	69.97
Austin, TX	71.48
Baltimore, MD	69.75
Birmingham, AL	69.59
Boise City, ID	68.36
Boston, MA	68.99
Buffalo, NY	66.10

假设你需要选择一个由 30 个城市区域组成的简单随机样本，以对美国和加拿大的生活费用支出进行一个深入研究。我们可以通过下列步骤来选择样本：

步骤 1. 选择 **Calc** 下拉菜单；

步骤 2. 选择 **Random Data** 选项；

步骤 3. 选择 **Samples From Columns**；

步骤 4. 当 Samples From Columns 对话框出现时：

在 **Sample** _ 框中输入 30；

在 **Sample 30 rows from column(s)** 框中输入 C1 C2；

在 **Store samples in** 框中输入 C3 C4；

步骤 5. 点击 **OK**。

包含 30 个城市区域的随机样本出现在 C3 和 C4 列。

附录 7.3　利用 Excel 进行随机抽样


CD 光盘数据
MetAreas

如果在 Excel 文件中含有一个总体的清单，那么我们就可以使用 Excel 软件从该总体中选择简单随机样本。例如，在数据集 MetAreas 的 A 列是美国和加拿大前 100 大的城市区域（*Places Rated Almanac — The Millennium Edition 2000*）。B 列是每个城市区域的综合评分。数据集中前 10 个城市区域以及对应的评分列示在表 7.6 中。假设你需要选择一个由 30 个城市区域组成的简单随机样本，以对美国和加拿大的生活费用支出进行一个深入研究。

通过向数据集上加一个额外的列，并使用 = RAND()函数在该列填上随机数，我们能够把任意 Excel 数据集的各行按照随机顺序排列。然后对随机数那一列使用 Excel 的升序排列功能，数据集内的各行将被随机地重新排序。容量为 n 的随机样本就出现在重新排序后数据集的前 n 行。

在 MetAreas 数据集中，第 1 行是标志，100 个城市区域在第 2 行到第 101 行。我们可以通过下列步骤来选择包括 30 个城市区域的简单随机样本：

步骤 1. 在单元格 C2 中输入 =RAND()；

步骤 2. 把单元格 C2 复制到单元格 C3: C101；

步骤 3. 选择 C 列的任意一个单元格；

步骤 4. 点击工具条上的 **Sort Ascending** 按钮。

由 30 个城市区域组成的随机样本出现在重新排序后数据集的第 2 到第 31 行。这时我们就不再需要 C 列的随机数，如果愿意的话，可以删去它们。

8 区间估计

统计实例：DOLLAR GENERAL 公司

8.1 总体均值的区间估计：大样本情况

CJW 估计问题

抽样误差

假定 σ 已知的大样本情况

用 s 估计 σ 的大样本情况

8.2 总体均值的区间估计：小样本情况

假定 σ 已知的小样本情况

用 s 估计 σ 的小样本情况

总体分布的作用

8.3 确定样本容量

8.4 总体比例的区间估计

确定样本容量

统计实例

DOLLAR GENERAL 公司*

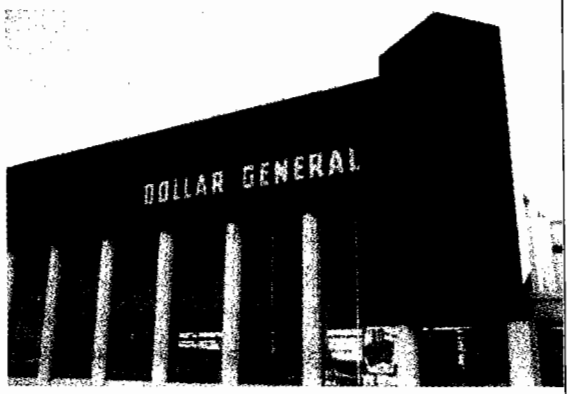
纳什维尔，田纳西州

Dollar General 公司是一家成立于 1939 年的干货批发公司。二战以后，公司开始在肯塔基州南部中心的乡村开设零售商店。今天 Dollar General 公司在美国的中部和东南部经营着超过 4 300 家商店。公司的重点是以小型便利店的模式，低价销售健康和美容用品、清洁用品、瓷器、文具、服装、鞋和生活必需品。

作为一个经营超过 20 000 种产品的存货密集型企业，Dollar General 公司决定采用 LIFO（后进先出）存货计价方法。该方法将当期的成本与当期的收入相配比，使价格变化对损益的影响最小化。另外，在通货膨胀时期，LIFO 方法减少了净收益，因此也减少了所得税。LIFO 方法还使得销售产生的自由现金流与收益相一致，并且考虑到以当期的成本对存货进行置换。

会计实践要求在 LIFO 计价方法下建立存货的 LIFO 指数。例如，LIFO 指数为 1.028，说明最近一年由于通货膨胀的原因，致使公司以当期成本计价的存货价值增加了 2.8%。

要建立 LIFO 指数需要对每种产品的年末存货按当年年末成本和上年年末成本分别计价。为了避免对超过 4 300 家地区零售商店的每种产品的存货都进行计量，公司从 100 个零售商店和 3 个仓库中抽取 800 件产品作为随机样本。首先在年末取得这 800 件样本产品的实物存货，然后由会计人员提供构建 LIFO 指数所需的当年和上年成本。



这家克利夫兰地区的商店是 4 300 多家 Dollar General 商店之一。© Jim Baron/ The Image Finders.

最近一年的 LIFO 指数是 1.030，不过由于该指数只是一个总体 LIFO 指数的样本估计值，所以需要对该估计的精度作出说明。基于样本结果和 95% 的置信水平，计算得到边际误差为 0.006。因此，总体 LIFO 指数 95% 置信水平的区间估计为 (1.024, 1.036)。据判断这一精度是良好的。

本章你将学习怎样计算与样本均值和样本比例相联系的边际误差，然后还会学习如何利用这些信息构建并解释总体均值和总体比例的置信区间估计。你也会了解为了保证边际误差在可接受的限度以内，应如何确定所需要的样本容量。

*作者感谢 Dollar General 公司的经理 Robert S. Knaul 先生提供该统计实例。

在第 7 章，我们讲述了点估计量是用来估计总体参数的样本统计量。例如，样本均值 $\bar{x}$ 是总体均值 μ 的点估计量，样本比例 $\bar{p}$ 是总体比例 p 的点估计量。由于点估计量不能提供关于估计值与总体参数接近程度的信息，统计学家更倾向于使用**区间估计**(interval estimate)，它能够提供有关估计精度的信息。

我们通过把点估计值减去和加上一个被称为**边际误差**(margin of error)的值，可以构建出总体参数的一个区间估计。我们在本章建立的所有区间估计都将采用以下的这种形式：

$$\text{点估计值} \pm \text{边际误差}$$

特别地，我们将说明如何建立总体均值 μ 和总体比例 p 的区间估计。总体均值的区间估计采用如下形式：

$$\bar{x} \pm \text{边际误差}$$

总体比例的区间估计形式如下：

$$\bar{p} \pm \text{边际误差}$$

边际误差提供了估计精度的信息。正如我们要说明的，第 7 章介绍的 $\bar{x}$ 抽样分布和 $\bar{p}$ 抽样分布在建立总体均值 μ 和总体比例 p 的区间估计方面起了重要的作用。

8.1　总体均值的区间估计：大样本情况

在这一节，我们将说明如何使用简单随机样本来建立总体均值的区间估计，重点是样本容量至少为 30 的大样本情况。我们从假定总体标准差 σ 已知的情况开始讨论，然后再考虑总体标准差未知的情况，以及怎样用样本标准差 s 估计 σ。

CJW 估计问题



CD 光盘数据
CJW

我们通过考虑 CJW 公司的一项每月客户服务调查，来举例说明如何建立总体均值的区间估计。CJW 公司专营体育设备及其附属产品，并提供网上订购服务。其客户服务的方式包括方便的在线订购，正确填写订单，及时传送订单，对客户的需求作出迅速反应等内容，它们是公司走向成功的关键。

CJW 公司的质量保证小组使用客户服务调查来度量客户对其在线订购系统的满意度。每个月该小组都把调查问卷送给由上月订过货的客户组成的随机样本，要求样本客户完成问卷，对公司的服务作出评价。每个客户的答案将被转换为满意度分数，范围从 0(最差的可能评价)到 100（最好的可能评价)。然后以每月的调查为基础，计算样本的平均客户满意度分数。然后，使用样本平均满意度分数作为所有客户平均满意度分数的点估计。样本均值向 CJW 公司的管理层提供了对公司在线订购服务质量的及时度量。这样当较低的客户满意度分数出现时，公司就可以迅速地采取有关措施。

以前的每月客户服务调查已经显示，满意度分数的标准差已经稳定在 20 附近。于是，总体标准差已知为 $\sigma = 20$ 这一假定得到了历史数据的支持。最近的客户满意度调查提供了包含 100 名客户（$n = 100$）的样本的满意度分数。样本的平均满意度分数 $\bar{x} = 82$，给出了总体均值 μ 的点估

计。在下面的讨论中，我们将计算与$\bar{x}=82$相联系的边际误差，然后建立起一个客户总体平均满意度分数的区间估计。

抽样误差

只要使用样本均值来估计总体均值，总有人会问：这个估计有多好？“有多好”其实是在询问以$\bar{x}$作为μ的点估计时的误差有多少。一般地，一个无偏点估计量与总体参数之差的绝对值，被称为**抽样误差**(sampling error)。在用样本均值估计总体均值的情况下，定义抽样误差如下：

$$\text{抽样误差} = |\bar{x}-\mu| \tag{8.1}$$

在实践中，由于总体均值μ未知而无法准确地确定抽样误差的值。但是，利用$\bar{x}$的抽样分布，我们能够对抽样误差进行概率描述，下面我们就以CJW公司为例来说明。

对于样本容量$n=100$，总体标准差$\sigma=20$的抽样分布，在第7章曾经介绍过的中心极限定理使得我们能够得出结论：可以通过具有均值μ和标准差$\sigma_{\bar{x}}=\sigma/\sqrt{n}=20/\sqrt{100}=2$的正态概率分布来近似$\bar{x}$的抽样分布。图8.1显示了该抽样分布。由于抽样分布说明了$\bar{x}$的值是如何围绕μ分布的，因此它提供了$\bar{x}$和μ之间可能的差值信息。而这个信息又是对抽样误差进行概率描述的基础。

中心极限定理可应用于任何总体，因此，即使总体分布未知，仍然可以使用本节介绍的方法。

利用标准正态概率分布表，我们发现具有正态分布的随机变量有95%的值处在距均值± 1.96个标准差的范围以内。由于$\bar{x}$的抽样分布能够被近似为正态分布，故所有的$\bar{x}$值中有95%必定距均值μ在$\pm 1.96\ \sigma_{\bar{x}}$以内。在CJW例子中，$1.96\ \sigma_{\bar{x}}=1.96\times 2=3.92$。于是，我们能够总结出有95%的样本均值$n=100$距总体均值$\mu$在$\pm 3.92$以内。

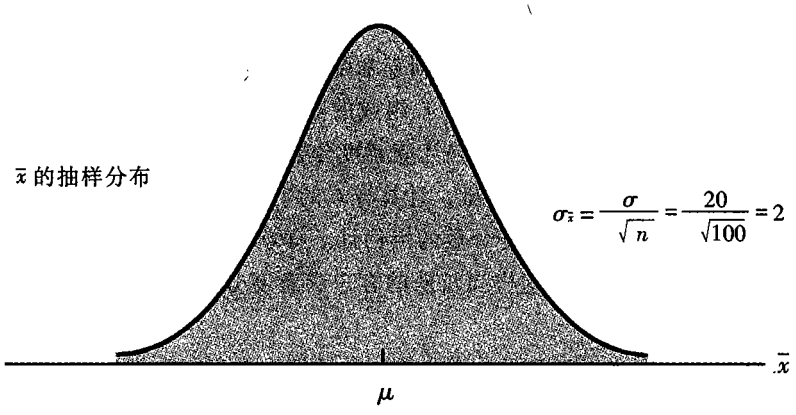


图8.1　来自一个由100名客户组成的样本的平均满意度分数的抽样分布

图 8.2 显示了与 μ 的差在 3.92 以内的样本均值的位置。注意到如果样本均值处在“所有 $\bar{x}$ 值的 95%”区域内的话，则抽样误差 $|\bar{x}-\mu|$ 小于等于 3.92。但是，如果样本均值位于抽样分布的下侧或者上侧的话，抽样误差 $|\bar{x}-\mu|$ 将大于 3.92。因此关于 CJW 问题的抽样误差，我们能够作出如下的概率描述：

样本均值的抽样误差小于等于 3.92 的概率是 0.95。

这个概率描述也就是当使用来自包含 100 名客户的简单随机样本的样本均值来估计总体均值时，对存在的抽样误差所做的**精度表述**(precision statement)。值 3.92 表示抽样误差的上限，被称为**边际误差**。

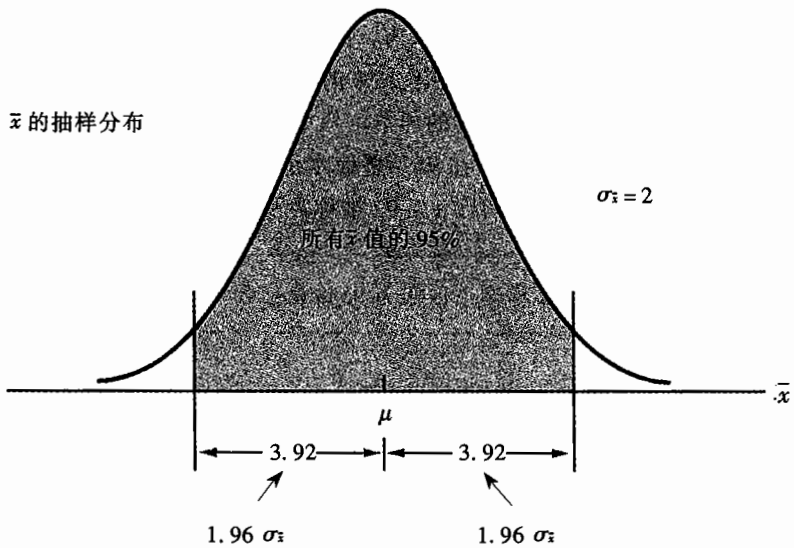


图 8.2　显示出抽样误差小于等于 3.92 的那部分样本均值位置的抽样分布

假定 σ 已知的大样本情况

$n=100$ 名客户的 CJW 样本满足大样本的条件 $n\geqslant 30$。另外，根据以前的客户调查，CJW 公司能够假定总体标准差已知为 $\sigma=20$。下面将说明我们如何利用精度表述来建立总体均值的置信区间估计。

正如我们在本章的开始所指出的那样，总体均值的区间估计采用了如下的形式：

$$\bar{x}\pm\text{边际误差}$$

对于大多数最近的 CJW 调查 $n=100$ 名的客户而言，样本均值为 $\bar{x}=82$。我们已经知道利用 0.95 的概率精度表述，其对应的边际误差是 3.92。于是，82 ± 3.92 就是总体均值的一个区间估计。由此我们能够得到该区间估计的下限为 $82-3.92=78.08$，该区间估计的上限为 $82+3.92=85.92$，78.08—85.92 是总体均值的一个区间估计。

为了说明如何解释总体均值 μ 的区间估计，假设有三个**不同的**随机样本，每个样本都由 100 名 CJW 公司的客户构成，让我们考虑从这三个样本得到的样本均值$\bar{x}$ 的所有可能值。假定第一个样本均值结果为图 8.3 中$\bar{x}_1$ 所示值。在这种情况下，图 8.3 显示了这个由$\bar{x}_1$ 加减 3.92 形成的区间包括了总体均值 μ。现在来考虑如果样本均值结果为图 8.3 中$\bar{x}_2$ 所示值的话，会发生什么情况。虽然该样本均值与第一个样本均值不同，我们看到$\bar{x}_2$ 加减 3.92 形成的区间也包括总体均值 μ。但是，当样本均值结果为图 8.3 中$\bar{x}_3$ 所示值时，会发生什么？这时，$\bar{x}_3$ 加减 3.92 形成的区间不包括总体均值 μ。原因是$\bar{x}_3$ 在分布的上侧，并且距离 μ 超过了 3.92。于是，$\bar{x}_3$ 加减 3.92 形成的区间就不会包括 μ。

图 8.3 阴影区域内的任意一个样本均值$\bar{x}$ 都能够构造出一个包含总体均值 μ 在内的区间。由于所有可能的样本均值中有 95% 处在阴影区域内，故$\bar{x}$ 加减 3.92 所形成的所有区间中也有 95% 的区间会包括总体均值 μ 在内。因此我们有 95% 的**置信度**认为所构建的区间 $(\bar{x}-3.92, \bar{x}+3.92)$ 将包括总体均值 μ。因为在所有可能的样本均值中有 95% 的样本均值能够构造出一个包含总体均值在内的区间，所以该区间是建立在 95% **置信水平**(confidence level)上的。值 0.95 被称为**置信系数**(confidence coefficient)，区间估计$\bar{x} \pm 3.92$ 被称为 95% 置信区间。

该讨论同时也说明了为什么称这个区间为 95% 置信区间。

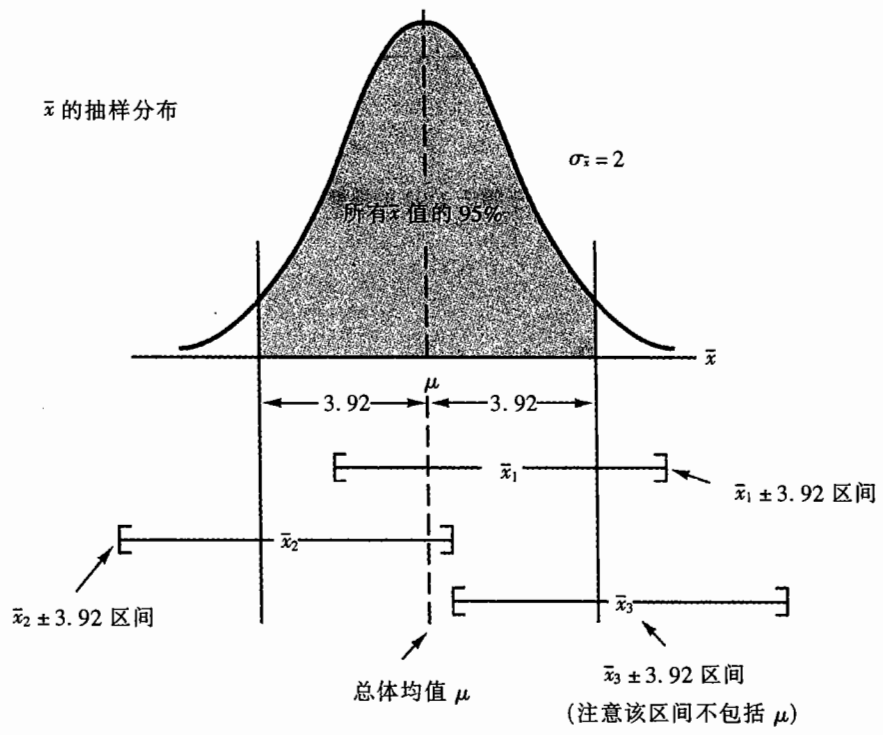


图 8.3　在所选样本均值分别为$\bar{x}_1$、$\bar{x}_2$ 和$\bar{x}_3$ 时形成的区间

当样本为大容量且假定总体标准差已知时，下面的公式说明了建立总体均值区间估计的通用方法：

总体均值的区间估计：假定 σ 已知的大样本情况（$n \geqslant 30$）

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

式中的 $1-\alpha$ 是置信系数，$z_{\alpha/2}$ 是当标准正态概率分布的上侧面积为 $\alpha/2$ 时的 z 值。

我们使用公式(8.2)来构建 CJW 问题的 95% 置信区间。对于一个 95% 的置信区间而言，置信系数是 $1-\alpha=0.95$，于是 $\alpha=0.05$。使用标准正态分布表，上侧面积 $\alpha/2=0.05/2=0.025$ 对应的 z 值为 $z_{0.025}=1.96$。由于 CJW 问题的样本均值 $\bar{x}=82$，假定已知 $\sigma=20$，样本容量 $n=100$，我们得到

$$82 \pm 1.96 \times \frac{20}{\sqrt{100}}$$

即

$$82 \pm 3.92$$

于是，使用公式(8.2)，得到边际误差为 3.92，置信区间为 $82-3.92=78.08$ 到 $82+3.92=85.92$。因此和前面的结果一样，CJW 公司能够有 95% 的把握相信总体平均满意度分数在 78.08 到 85.92 之间。

虽然 95% 的置信水平较为常用，但是也可以考虑使用其他一些置信水平，如 90% 和 99%。表 8.1 列出了最经常使用的一些置信水平所对应的 $z_{\alpha/2}$ 值*。利用这些值和公式(8.2)，CJW 问题 90% 的置信区间为

$$82 \pm 1.645 \times \frac{20}{\sqrt{100}}$$

即

$$82 \pm 3.29$$

因此在 90% 的置信水平上，边际误差为 3.29，置信区间为 $82-3.29=78.71$ 到 $82+3.29=85.29$。类似地，99% 的置信区间是

$$82 \pm 2.576 \times \frac{20}{\sqrt{100}}$$

即

$$82 \pm 5.15$$

因此在 99% 的置信水平上，边际误差为 5.15，置信区间为 $82-5.15=76.85$ 到 $82+5.15=87.15$。

比较 90%、95% 和 99% 置信水平的结果，我们看到，为了得到更高的确信程度，边际误差和

* 标准正态概率表给出的 z 值精确到小数点后两位。但是对于 90%、95%、99% 的置信水平，在实际应用中通常使用精确到小数点后三位的 z 值。

置信区间的宽度都必须更大。

表 8.1 最常用的一些置信水平所对应的值

置信水平	α	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

用 s 估计 σ 的大样本情况

使用公式(8.2)计算总体均值区间估计的困难之处在于，在许多实际应用中缺乏假定总体标准差为已知的基础。这时，只有使用样本标准差 s 来估计 σ。在大样本情况下($n \geq 30$)，当样本容量增大时，样本标准差 s 对 σ 作出了良好的估计这一事实以及中心极限定理，*都使得我们能够利用下面的方法来建立总体均值的区间估计。

总体均值的区间估计：用 s 估计 σ 的大样本情况 ($n \geq 30$)

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.3}$$

式中，s 为样本标准差，$1-\alpha$ 为置信系数，$z_{\alpha/2}$ 是标准正态概率分布的上侧面积为 $\alpha/2$ 时的 z 值。

由于总体标准差 σ 未知，使用样本标准差 s 来估计 σ。

为了说明这种区间估计方法，让我们来考虑一个抽样研究，它的设计目的是估计美国家庭的信用卡负债情况。一个包含 85 户家庭的样本提供了表 8.2 所示的信用卡余额。由于 $n=85$，这是一个大样本情况。另外，因为不能得到每户家庭信用卡余额的历史数据，只能使用样本标准差 s 来估计总体标准差 σ。现在我们来建立总体每户家庭的平均信用卡余额的 95% 置信区间估计。

首先，我们使用表 8.2 的数据计算出样本均值 $\bar{x}=5\,900$ 美元，样本标准差 $s=3\,058$ 美元。在 95% 的置信水平上，$z_{\alpha/2}=z_{0.025}=1.96$。对于样本容量 $n=85$，由公式(8.3)可得

$$5\,900 \pm 1.96 \times \frac{3\,058}{\sqrt{85}}$$

即

$$5\,900 \pm 650$$

于是，边际误差是 650 美元，总体均值的 95% 置信区间估计为 5 900 − 650 = 5 250 美元到 5 900 +

* 大样本理论显示，当样本容量增加时，样本方差 s^2 随机地收敛于总体方差 σ^2。这种收敛性使得我们能够使用 s 来估计 σ，并且使用公式 (8.3) 来计算总体均值的区间估计。

650 = 6 550 美元。因此，我们能够有 95% 的把握相信所有家庭总体平均的信用卡余额处在 5 250 美元和 6 550 美元之间。

表 8.2　85 户家庭样本的信用卡余额


CD 光盘数据
Balance

9 619	5 994	3 344	7 888	7 581	9 980
5 364	4 652	13 627	3 091	12 545	8 718
8 348	5 376	968	943	7 959	8 452
7 348	5 998	4 714	8 762	2 563	4 935
381	7 530	4 334	1 407	6 787	5 938
2 998	3 678	4 911	6 644	5 071	5 266
1 686	3 581	1 920	7 644	9 536	10 658
1 962	5 625	3 780	11 169	4 459	3 910
4 920	5 619	3 478	7 979	8 047	7 503
5 047	9 032	6 185	3 258	8 083	1 582
6 921	13 236	1 141	8 660	2 153	
5 759	4 447	7 577	7 511	8 003	
8 047	609	4 667	14 442	6 795	
3 924	414	5 219	4 447	5 915	
3 470	7 636	6 416	6 550	7 164	

Minitab、Excel 以及其他一些应用软件提供了在可以利用 s 估计 σ 时计算大样本情况置信区间的简便办法。在附录 8.2 和 8.3 我们描述了使用 Minitab 和 Excel 的步骤。Minitab 区间估计步骤的运行结果显示在图 8.4 中。由 85 户家庭的样本得到的样本平均信用卡余额为 5 900 美元，样本标准差为 3 058 美元，均值的标准误差为 332 美元，95% 的置信区间为 5 250 美元到 6 550 美元。

在第 7 章，我们曾指出 $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ 是均值的标准误差。Minitab 通过利用 s 估计 σ，给出了一个均值标准误差的估计值。

```
Variable   N   Mean   StDev   SE Mean        95.0% CI
Balance   85   5900    3058       332   (5250, 6550)
```

图 8.4　信用卡余额调查的置信区间的 Minitab 输出结果

评　注

1. 在建立总体均值的区间估计时，我们在选择样本之前已规定了所要求的置信系数 $1-\alpha$。因此在选取样本之前，我们已经知道最终计算的置信区间将包含总体均值 μ 在内的概率为

$1-\alpha$。但是，一旦选取了样本，并计算出样本均值 $\bar{x}$，确定了特定的区间估计以后，所得到的区间可能包括或者可能不包括总体均值 μ。如果 $1-\alpha$ 合理地选择了较大值的话，假设我们不断重复该过程，用这种方法建立的所有可能区间中有 $100(1-\alpha)$的区间会包括 μ 在内。

2. 在实践应用中很少会知道总体标准差 σ。但是，历史数据、理论和其他一些来源能够向使用者提供关于 σ 值的很多信息。如果使用者相信这些 σ 值要比使用样本标准差 s 估计出的 σ 值更好一些的话，就应该使用假定 σ 已知的大样本情况公式(8.2)来计算总体的区间估计。
3. 在区间估计的公式(8.2)和(8.3)中，我们注意到样本容量 n 出现在分母上。因而，当某个样本容量产生的区间太宽以至于失去使用价值时，我们可以考虑增大样本容量。由于 n 在分母上，增大样本容量能够使边际误差减小，估计区间变窄，精度提高。我们将在8.3节讨论为达到所需要的精度确定相应简单随机样本容量的方法。

练　习

方　法

1. 一个包含40项元素的简单随机样本的样本均值为25，总体标准差为 $\sigma=5$。
   a. 均值的标准误差 $\sigma_{\bar{x}}$ 是多少？
   b. 在95%的概率下，边际误差是多少？

2. 一个包含50项元素的简单随机样本的样本均值为32，样本标准差为6。
   a. 求总体均值的90%置信区间。
   b. 求总体均值的95%置信区间。
   c. 求总体均值的99%置信区间。

3. 包含60项元素的简单随机样本的样本均值为80，样本标准差为15。
   a. 计算总体均值的95%置信区间。
   b. 假定一个容量为120的样本具有相同的样本均值与样本标准差，求总体均值的95%置信区间。
   c. 样本容量增加对总体均值的区间估计有什么影响？

4. 已知某个总体均值的95%置信区间是(122,130)。如果样本均值为126，样本标准差为16.07，问该研究中使用的样本容量是多少？

应　用

5. 为了估计亚特兰大某个主要饭店每位客户午餐的平均消费额，从一个包含49名客户的样本中收集数据。假定总体标准差是5美元。
   a. 在95%的置信水平下，边际误差是多少？

b. 如果样本均值是 24.80 美元，总体均值的 95% 置信区间是什么？

6. 《纽约时报 1998 年年鉴》公布了各个行业从业人员的平均每周收入。其中服务业的从业人员平均每周收入 369 美元，假定该结果是建立在一个由 250 名服务业从业人员组成的样本基础上，并且样本标准差为 50 美元。计算服务业人员总体平均每周收入的 95% 区间估计。

7. 美国顶级的商学院所录取的研究生入学成绩平均学分达到 3.37（*Best Graduate Schools, U. S. News and World Report*, 2001 Edition）。假定该估计是建立在由 120 名被顶级商学院录取的学生组成的样本基础之上，利用以往年份的数据，可以假定总体标准差已知为 $\sigma = 0.28$。对于顶级商学院录取的研究生总体，其平均学分的 95% 置信区间估计是什么？

8. 美国国家独立学院与大学联合会报告说，毕业于公立高校的学生在毕业前平均负债额为 12 000 美元（*Kiplinger's Personal Finance Magazine*, November 1998）。假定该平均数额是建立在包括 245 名学生的样本基础上，并且根据以往的研究，负债的总体标准差已知为 2 200 美元。
   a. 建立总体均值的 90% 置信区间估计。
   b. 建立总体均值的 95% 置信区间估计。
   c. 建立总体均值的 99% 置信区间估计。
   d. 讨论当置信水平提高时，置信区间的宽度将发生什么变化。这个结果合理吗？请解释。

9. 《华尔街日报》于 1997 年 11 月 25 日总结报道了生产工人每小时的平均报酬。由 60 名生产工人的小时报酬率组成的样本包含在数据集 Workers 中。



CD 光盘数据
Workers

   a. 利用数据集 Workers 建立生产工人总体平均小时报酬率的点估计。
   b. 样本标准差是多少？
   c. 生产工人总体平均小时报酬率的 95% 置信区间是什么？

10. 根据尼尔森媒体研究的报告，在从晚上 8 点到晚上 11 点的时段内，家庭的平均收看电视时间为每周 7.75 小时（*The World Almanac 2000*）。假定样本容量为 180 户家庭，样本标准差为 3.45 小时，则在晚上 8 点到晚上 11 点时段内，总体的家庭每周平均收看电视时间的 95% 置信区间估计是什么？

11. 国际航空运输协会对商务旅行者进行了调查，以评定经营跨大西洋航线的机场的质量得分。最大的可能得分是 10。假定选取了 50 名商务旅行者组成简单随机样本，每个旅行者被要求给迈阿密国际机场评分。从样本中得到的评分如下：



CD 光盘数据
Miami

6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

建立商务旅行者总体对迈阿密机场平均打分的 95% 置信区间估计。

12. 在 2000 年暑假期间调查了包括 Wendy's、麦当劳和 Burger King 在内的 30 家快餐店（*The*


CD 光盘数据
FastFood

Cincinnati Enquirer, July 9, 2000)。在每次调查中，客户都来到外卖窗口，要一份诸如三明治、炸薯条之类的基本食品，然后记录快餐店从收到菜单到交付食品所花费的时间。在 30 次调查中所花费的分钟数如下：

0.9	1.0	1.2	2.2	1.9	3.6	2.8	5.2	1.8	2.1
6.8	1.3	3.0	4.5	2.8	2.3	2.7	5.7	4.8	3.5
2.6	3.3	5.0	4.0	7.2	9.1	2.8	3.6	7.3	9.0

a. 快餐店总体平均外卖花费时间的估计值是多少？

b. 在 95% 置信水平下，边际误差是多少？

c. 总体平均时间的 95% 置信区间估计是什么？

8.2 总体均值的区间估计：小样本情况

在 8.1 节介绍的区间估计方法建立在大样本理论和中心极限定理的基础上。这种情况下，无论总体的概率分布如何，都可以用正态概率分布来近似$\bar{x}$的抽样分布。因此，大样本方法不需要任何关于总体分布的假设。

在本节我们将说明如何在小样本（$n<30$）的情况下，建立总体均值的区间估计。小样本情况下$\bar{x}$的抽样分布依赖于总体的分布形式。因此下面的区间估计方法是基于总体具有正态分布的假设而提出的。如果该假设符合实际情况，则能够使用本节给出的方法来计算总体均值的区间估计。但是，如果该假设不符合实际，我们建议把样本容量增加到 $n\geq30$，改为使用 8.1 节的大样本方法。

在小样本情况下，我们从假设总体标准差 σ 已知的情况入手，然后考虑用样本标准差 s 估计总体标准差 σ 的情况。

本章的方法是基于假设总体呈正态分布。

假定 σ 已知的小样本情况

我们假设总体具有正态分布，并且假定总体标准差 σ 已知。这时$\bar{x}$的抽样分布对于任意的样本容量都是正态分布，其均值为 μ，标准差为$\sigma_{\bar{x}}=\sigma/\sqrt{n}$。因此，给定$\bar{x}$的抽样分布是正态分布并且假定 σ 已知的小样本区间估计方法与 8.1 节的大样本区间估计方法相同。该方法的表达式如下：

总体均值的区间估计：假定 σ 已知的小样本情况（$n<30$）

假设：总体具有正态概率分布。

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.4}$$

式中的 $1-\alpha$ 为置信系数，$z_{\alpha/2}$ 是在标准正态概率分布的上侧面积为 $\alpha/2$ 时对应的 z 值。

假定 σ 已知，并且已知对应于所要求置信水平的 $z_{\alpha/2}$ 值，样本容量 n，以及样本均值 $\bar{x}$，就能够使用公式(8.4)计算总体均值的区间估计。由于该计算过程与 8.1 节的计算过程相同，在此我们就不再举例说明了。

用 s 估计 σ 的小样本情况

我们首先假定总体具有正态分布。如果我们没有任何根据假定已知总体标准差 σ 的话，就要使用样本标准差 s 来估计 σ。这时的区间估计方法要依靠一种被称为 **t 分布**(t distribution)的概率分布。

William Sealy Gosset，笔名为 Student，创建了 t 分布。他是牛津大学数学系的一名研究生，曾为爱尔兰首都都柏林的 Guinness 啤酒厂工作。在此期间，他进行过小规模的材料和温度实验，在实验中他建立了这一新的小样本统计理论。

t 分布是一系列类似的概率分布，一个特定的分布依赖于一个被称为**自由度**(degrees of freedom)的参数。自由度为 1 的 t 分布是惟一的，自由度为 2 的 t 分布也是惟一的。实际上，对于每一个自由度，有且只有一个 t 分布与之相对应。随着自由度的增大，t 分布与正态概率分布之间的差别变得越来越小。图 8.5 显示了自由度分别为 10 和 20 时的 t 分布，以及它们与标准正态概率分布的关系。我们注意到，t 分布的自由度越大，它的变异程度越小，也越接近于标准正态概率分布。我们还注意到，t 分布的均值为 0。

我们给 t 加上下标以表示 t 分布的上侧面积。例如，就像我们用 $z_{0.025}$ 表示与标准正态概率分

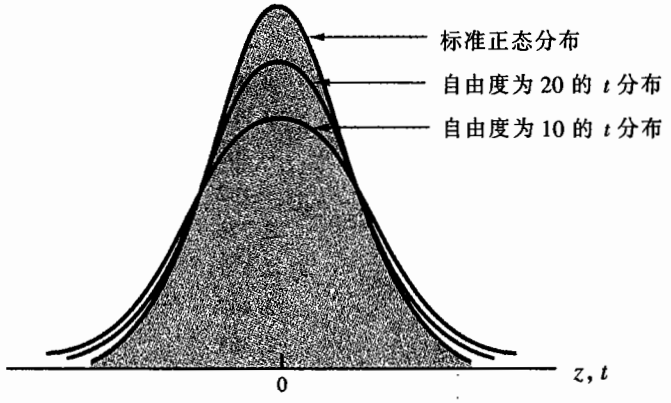


图 8.5　标准正态概率分布与自由度分别为 10 和 20 的两个 t 分布的比较

布的上侧面积 0.025 相对应的 z 值一样，我们使用 $t_{0.025}$ 表示与 t 分布的上侧面积 0.025 相对应的 t 值。一般地，我们使用符号 $t_{\alpha/2}$ 代表与 t 分布的上侧面积 $\alpha/2$ 对应的 t 值。见图 8.6。

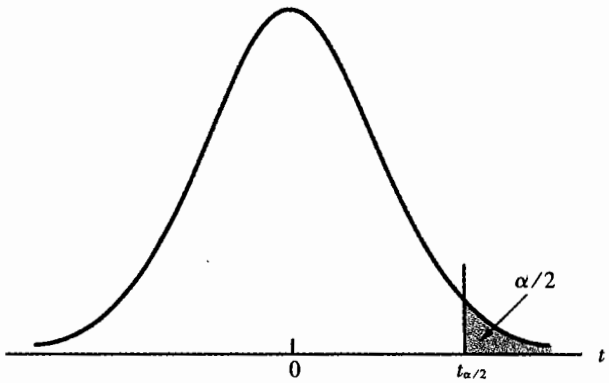
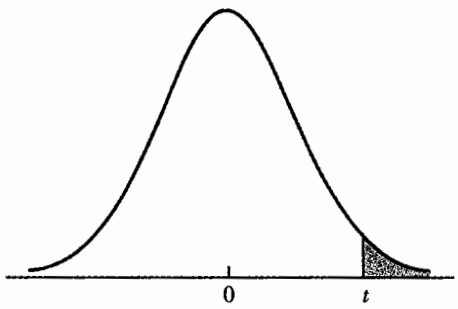


图 8.6　上侧面积或上侧概率为 $\alpha/2$ 的 t 分布

表 8.3 是 t 分布表。例如，对于自由度为 10 的 t 分布，$t_{0.025}=2.228$。类似地，对于自由度为 20 的 t 分布，$t_{0.025}=2.086$。当自由度持续增加时，$t_{0.025}$ 趋近于 $z_{0.025}=1.96$。

随着自由度的增加，t 分布趋近于标准正态分布。实际上，标准正态分布的 z 值可以在 t 分布表中自由度为无穷的那一行查到。

表 8.3　t 分布的上侧面积表（以自由度 10 为例：$t_{0.025}=2.228$）



自由度	上侧面积				
	0.10	**0.05**	**0.025**	**0.01**	**0.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250

（续表）

自由度	上侧面积				
	0.10	**0.05**	**0.025**	**0.01**	**0.005**
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

现在，我们已经知道了什么是 t 分布，接着我们来学习如何利用 t 分布对总体均值进行区间估计。假定总体具有正态概率分布，并且使用样本标准差 s 来估计总体标准差 σ。这时使用下面的区间估计方法：

总体均值的区间估计：用 s 估计 σ 的小样本情况（$n<30$）

假定：总体具有正态概率分布。

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.5}$$

式中的 s 是样本标准差，$1-\alpha$ 是置信系数，$t_{\alpha/2}$ 是与自由度为 $n-1$ 的 t 分布的上侧面积 $\alpha/2$ 相对应的 t 值。

与公式(8.5)中的 t 值相联系的自由度是 $n-1$，这是由于不得不采用样本标准差 s 来估计总体标准差 σ 的缘故。样本标准差 s 的表达式为

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

自由度是计算$\sum(x_i-\bar{x})^2$时所使用的独立信息个数。在计算$\sum(x_i-\bar{x})^2$时，涉及到的n个信息如下：$x_1-\bar{x}$，$x_2-\bar{x}$，…，$x_n-\bar{x}$。在3.2节，我们曾经指出：对于任意的数据集，$\sum(x_i-\bar{x})=0$。因此，只有$n-1$个$x_i-\bar{x}$值是独立的，也就是说，如果我们知道了$n-1$个值，利用$x_i-\bar{x}$值的总和必定等于0这一条件，就能够准确地确定剩下的那个值。因此，$n-1$是与$\sum(x_i-\bar{x})^2$相联系的自由度，也就是公式(8.5)中t分布的自由度。

我们来举例说明小样本区间的估计方法。考虑一项由Scheer工业公司进行的培训计划评估。该公司的生产主管对一套可用于培训公司维修工，提高他们设备维修操作水平的计算机辅助程序感兴趣，他希望计算机辅助方法能够减少培训员工所需的时间。为了评价该培训方法，生产主管需要估计这种计算机辅助程序带来的总体平均培训时间。

假设管理层同意利用计算机辅助程序培训15名员工。样本中每个员工所需的培训时间数据列在表8.4中。这些数据的样本均值和样本标准差如下：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{808}{15}=53.87\text{天}$$

$$s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{651.73}{14}}=6.82\text{天}$$

表8.4　Scheer工业公司计算机辅助培训计划需要的培训时间



CD光盘数据
Training

员工编号	时间（天）	员工编号	时间（天）	员工编号	时间（天）
1	52	6	59	11	54
2	44	7	50	12	58
3	55	8	54	13	60
4	44	9	62	14	62
5	45	10	46	15	63

总体平均培训时间的点估计是53.87天。通过建立总体均值的区间估计，我们能够得到关于该估计精度的信息。由于总体标准差未知，我们用样本标准差$s=6.82$天来估计σ。该问题中的样本容量为$n=15$，是小容量情况，故我们使用公式(8.5)来构建总体均值的95%置信区间估计。如果我们假定总体的培训时间具有正态概率分布，则自由度为$n-1=14$的t分布是适用于该区间估计方法的恰当概率分布。我们在表8.3看到当自由度等于14时，$t_{\alpha/2}=t_{0.025}=2.145$。使用公式(8.5)，我们有

$$\bar{x}\pm t_{0.025}\frac{s}{\sqrt{n}}$$

即

$$53.87\pm2.145\times\frac{6.82}{\sqrt{15}}$$

$$53.87\pm3.78$$

> t 分布是建立在总体具有正态概率分布的假设基础上，但是，只要总体分布与正态分布的差别不是很大，就能够使用基于 t 分布的置信区间。

因此，边际误差是 3.78 天，总体均值的 95% 置信区间估计是(50.09 天，57.65 天)。

Minitab 也能够计算本例中总体均值的区间估计，输出结果显示在图 8.7 中。结果表明：用 15 个员工的样本计算出的样本均值是 53.87 天，样本标准差是 6.82 天，均值的标准误差是 1.76 天，总体的 95% 置信区间是(50.09 天，57.65 天)。

Variable	N	Mean	StDev	SE Mean	95.0 % CI
Time	15	53.87	6.82	1.76	(50.09, 57.65)

图 8.7　Scheer 工业公司问题的置信区间的 Minitab 输出结果

总体分布的作用

在结束本节时，我们有必要指出，建立总体均值的区间估计时需要考虑的因素之一是要确定总体分布的形式。8.1 节的大样本区间估计方法不需要关于总体分布的假设，因此适用于具有任意分布的总体。而 8.2 节的区间估计方法是建立在总体具有正态分布的基础之上，故只能适用于具有正态分布的总体，或接近正态的总体。

在实践中，如果样本容量较大($n \geqslant 30$)并且总体分布未知的话，我们采用 8.1 节的方法。这时，可以构建总体均值的区间估计如下：

假设 σ 已知

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

用 s 估计 σ

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.3}$$

对于具有任意分布的总体，大样本理论支持使用公式(8.2)和(8.3)。

如果样本容量较小($n<30$)，并且总体具有正态分布的话，则采用 8.2 节的方法。这时，我们能够建立总体均值的区间估计如下：

假设 σ 已知

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.4}$$

用 s 估计 σ

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.5}$$

实际上，如果已知总体具有正态分布，无论是大样本情况还是小样本情况，都能够使用公式(8.4)和(8.5)来建立总体均值的区间估计。

在大多数的实践应用中，我们不知道总体的分布。由于公式(8.2)和(8.3)适用于具有任意分布的总体，故在给定样本容量较大（$n \geqslant 30$）的条件下，我们对总体分布不作任何假设就能够使用公式(8.2)或(8.3)。因此，需要假设总体具有正态分布并且考虑使用公式(8.4)和(8.5)的惟一场合是当样本容量较小($n<30$)时。图 8.8 总结了我们讲述过的所有区间估计方法，为我们计算总体均值的区间估计提供了实际指导。

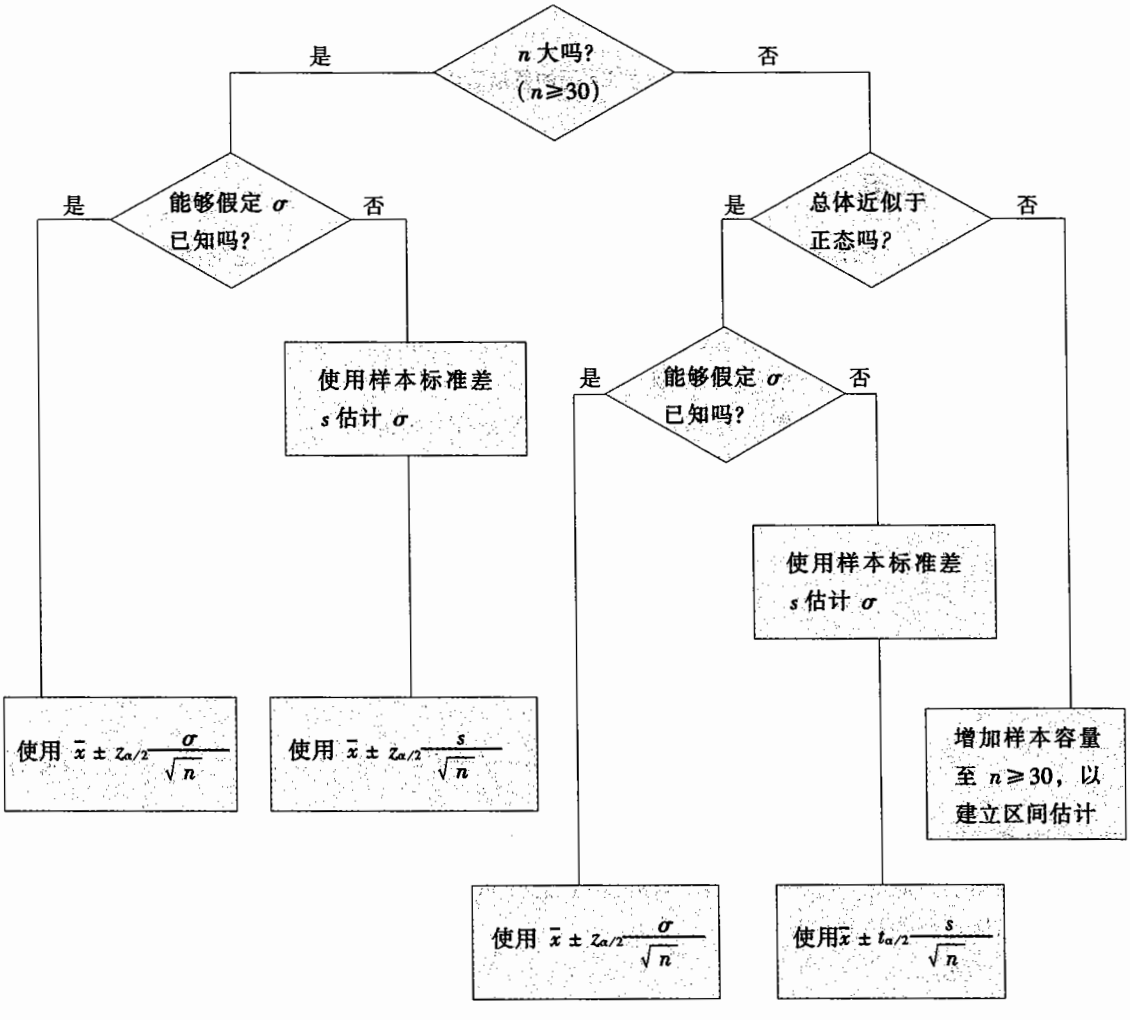


图 8.8　总体均值的区间估计计算方法总结

练　习

方　法

13. 对于自由度为 12 的 t 分布，计算下面每个区域内的面积或概率。
    a. 1.782 左边
    b. −1.356 右边
    c. 2.681 右边
    d. −1.782 左边
    e. −2.179 和 +2.179 之间
    f. −1.356 和 +1.782 之间

14. 计算下面每个例子的 t 值。
    a. 自由度为 18 的 t 分布的上侧面积为 0.05。
    b. 自由度为 22 的 t 分布的下侧面积为 0.10。
    c. 自由度为 5 的 t 分布的上侧面积为 0.01。
    d. 自由度为 14 的 t 分布在这两个 t 值之间的面积是 90%。
    e. 自由度为 28 的 t 分布在这两个 t 值之间的面积是 95%。

15. 自测题 下面的数据是从来自正态总体的样本中收集的：10，8，12，15，13，11，6，5。
    a. 总体均值的点估计是多少？
    b. 总体标准差的点估计是多少？
    c. 总体均值的 95% 置信区间是多少？

16. 包括 20 个观察值的简单随机样本来自正态总体，其样本均值为 17.25，样本标准差为 3.3。
    a. 建立总体均值的 90% 置信区间。
    b. 建立总体均值的 95% 置信区间。
    c. 建立总体均值的 99% 置信区间。

应　用

17. 自测题 在对一种新生产方法进行检验时，随机选取了 18 名雇员并要求他们试用该方法。这 18 个雇员的样本平均生产率为每小时 80 件，样本标准差为每小时 10 件。假定新方法下雇员总体的生产率具有正态分布，计算在新方法下总体平均生产率的 90% 和 95% 区间估计。

18. 《今日美国》对出租汽车的汽油价格进行了一项研究，在 12 个主要机场得到了下面的每加仑

汽油价格(*USA Today*, April 4, 2000):

1.58 1.53 1.60 1.55 1.80 1.75 1.58 1.62 1.69 1.21 1.50 1.55

a. 总体平均汽油价格的点估计是多少?

b. 总体汽油价格的标准差是多少?

c. 假定总体汽油价格具有正态分布,总体平均汽油价格的95%置信区间估计是什么?

19. 美国广告商协会记录了半小时最佳时段的电视节目中广告所占据的分钟数。选取20个主要广播网晚上8:30最佳时段播出的电视节目组成样本,从中得到广告时间(分钟)的代表性数据如下:



CD光盘数据
TVtime

6.0	6.6	5.8
7.0	6.3	6.2
7.2	5.7	6.4
7.0	6.5	6.2
6.0	6.5	7.2
7.3	7.6	6.8
6.0	6.2	

假定广告时间总体具有正态分布,计算晚8:30最佳半小时时段内平均广告时间的点估计和95%置信区间。

20. Skilling分销公司的销售人员被要求每周提交一份列有本周联系客户情况的报告。一个由61份每周联系报告组成的样本显示:销售人员每周平均联系客户22.4次,样本标准差为5次。

a. 利用大样本情况下的公式(8.3)计算销售人员总体每周联系客户的平均次数的95%置信区间。

b. 假定总体的每周联系次数服从正态分布,使用自由度为60的 t 分布来建立总体平均每周联系次数的95%置信区间。

c. 比较你在(a)、(b)中的答案。评论为什么在大样本的情况下,即使 t 分布也适用,仍然允许采用(a)中的方法进行区间估计。

21. 不断增长的处方药价格促使美国国会考虑制定新的法律,以迫使医药公司向老年公民提供减去利润的处方折扣。国会政府改革委员会提供了一些应用最广泛的药品的处方成本(*Newsweek*, May 8, 2000)。假设下面的数据取自一个由一种降胆固醇药物Zocor的处方成本构成的样本:

110 112 115 99 100 98 104 126

假定该药品总体的处方成本具有正态分布,则它的总体平均处方成本的95%置信区间估计是什么?

22. 美国人每晚的睡眠时间变化不大,总体中12%的人睡眠少于6小时,3%的人睡眠超过8小时

(*The Macmillan Visual Almanac*, 1996)。下面这个包含 25 人的样本报告了他们每晚睡眠时间的数据：


CD 光盘数据
Sleep

6.9	7.6	6.5	6.2	5.3
7.8	7.0	5.5	7.6	6.7
7.3	6.6	7.1	6.9	6.0
6.8	6.5	7.2	5.8	8.6
7.6	7.1	6.0	7.2	7.7

a. 美国人总体平均每晚睡眠时间的点估计是多少？

b. 假定总体服从正态分布，建立总体平均每晚睡眠时间的 95% 置信区间。

8.3　确定样本容量

在假定 σ 已知的大样本情况下，下式给出了总体均值的区间估计：

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

$z_{\alpha/2}(\sigma/\sqrt{n})$是边际误差。因此，我们看到 $z_{\alpha/2}$、总体标准差 σ 和样本容量 n 共同确定了边际误差。一旦我们选择了置信系数 $1-\alpha$，就能够确定 $z_{\alpha/2}$。然后，如果知道 σ 的值，我们就能够确定任意边际误差所需的样本容量 n。计算所需样本容量 n 的公式可以推导如下：

以 E 代表所希望的边际误差

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

解 $\sqrt{n}$，我们得到

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{E}$$

把等式的两边同时平方，我们得到样本容量的表达式为

对总体均值进行区间估计所需的样本容量

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \tag{8.6}$$

如果在抽样之前就选择了所希望的边际误差，则本节的方法可用于确定能够满足边际误差要求的样本容量。

该样本容量能够在给定的置信水平下，提供所希望的边际误差。

在公式(8.6)中，E 代表使用者可接受的边际误差，$z_{\alpha/2}$ 的值直接由用于建立区间估计的置信水平确定。虽然必须要考虑到使用者的因素，但 95% 的置信水平仍是最常选择的值($z_{0.025}=1.96$)。

最后，公式(8.6)要求已知总体标准差 σ 的值。但在大多数情况下，σ 都是未知的。不过，如果我们拥有 σ 的初步值或计划值的话，我们仍可以使用公式(8.6)。实践中，可选择下列方法之一来确定 σ 的值：

1. 使用以前由相同或类似元素构成的样本，用它的样本标准差代替 σ 的值。
2. 使用小规模试验研究选择一个初步样本，把来自初步样本的样本标准差作为 σ 的计划值使用。
3. 利用判断或者“最佳猜测”确定 σ 的值。例如，我们可以从估计总体的最大和最小数据值开始。把估计的最大和最小值之差作为数据极差的估计值，最后，我们往往建议用极差除以 4 作为标准差的粗略估计，也就是 σ 的一个可接受的计划值。

在确定样本容量之前必须先确定总体标准差 σ 的计划值。此处我们讨论了三种能够获得计划值的方法。

让我们考虑下面的例子，以说明如何使用公式(8.6)来确定样本容量。有一项研究调查了在美国租赁汽车的支出，该调查发现租赁一辆中型汽车的支出大约是每天 55 美元。假设进行该项研究的组织想要再进行一项新研究，以估计在美国租赁一辆中型汽车的总体平均每日支出。在设计这项新研究时，项目主管已经规定了边际误差为 2 美元，置信水平为 95%。

使用公式(8.6)，我们看到项目主管已经规定了边际误差 $E=2$，95% 的置信水平对应的 $z_{0.025}=1.96$。于是，为了计算所要求的样本容量，我们仅需要再找到总体标准差 σ 的计划值。为此，一个分析人员检查了来自原始研究的样本数据并且发现每日租赁支出的样本标准差是 9.65 美元。把该值作为 σ 的计划值，我们有

$$n=\frac{(z_{\alpha/2})^2\sigma^2}{E^2}=\frac{1.96^2\times 9.65^2}{2^2}=89.43$$

公式(8.6)提供了满足希望的边际误差要求的最小样本容量。如果计算出来的值不是整数，可取整到下一个整数值作为推荐的样本容量。

于是，为了满足项目主管要求的 2 美元边际误差，新研究需要的样本容量至少为 89.43 次中型汽车租赁。当得到的 n 不是整数时，我们将其**取整**到下一个整数值。于是，建议的样本容量为 90 次中型汽车租赁。

练　习

方　法

23. 为了得到边际误差为 5 的 95% 置信区间，应该选择多大的样本容量？假定总体标准差是 25。

24. 估计某数据集的极差为 36。



a. 总体标准差的计划值是多少？

b. 使用 95% 的置信度，为使边际误差为 3 应选择多大的样本容量？

c. 使用 95% 的置信度，为使边际误差为 2 应选择多大的样本容量？

应　用

25. 参见 8.2 节 Scheer 工业公司的例子。以 $\sigma = 6.82$ 天作为总体标准差的计划值。



a. 假定置信度为 95%，为使边际误差为 1.5 天，需要多大的样本容量？

b. 如果把精度设定为 90% 的置信水平，为使边际误差为 2 天，需要多大的样本容量？

26. 根据《新娘》杂志的报道，婚礼的平均支出为 19 000 美元（*USA Today*, April 17, 2000）。假定总体标准差为 9 400 美元，《新娘》杂志计划使用一个年度调查来监测婚礼的支出，以 95% 为置信度。

a. 如果需要的边际误差为 1 000 美元的话，则样本容量应为多少？

b. 如果需要的边际误差为 500 美元的话，则样本容量应为多少？

c. 如果需要的边际误差为 200 美元的话，则样本容量应为多少？

27. 拥有工商管理学位的大学毕业生起始年薪的标准差被认为大约是 2 000 美元。假设需要求得平均起始年薪的 95% 置信区间估计，当希望的边际误差分别为下列数值时，应选择多大的样本容量？

a. 500 美元。

b. 200 美元。

c. 100 美元。

28. RealFacts 是一家房地产研究公司，它提供了洛杉矶地区的公寓平均月租金（*Los Angeles Times*, August 1, 1999）。假定总体标准差为 220 美元，需要的边际误差为 50 美元。

a. 对于总体平均租金的 90% 置信区间估计，样本容量应为多少？

b. 对于总体平均租金的 95% 置信区间估计，样本容量应为多少？

c. 对于总体平均租金的 99% 置信区间估计，样本容量应为多少？

d. 如果需要的边际误差不变，当置信水平增大时，样本容量将如何变化？

29. *1998 Information Please Almanac* 报道了美国 15 个最大城市的居民上班途中花费时间的数据。假定利用由旧金山的居民组成的初步简单随机样本，得到总体标准差的计划值为 6.25 分钟。
   a. 如果我们想在边际误差为 2 分钟的情况下，估计旧金山居民总体的平均上班途中时间，应该使用多大的样本容量？假设置信度为 95%。
   b. 如果我们想在边际误差为 1 分钟的情况下，估计旧金山居民总体的平均上班途中时间，应该使用多大的样本容量？假设置信度为 95%。

30. 在纽约证交所上市股票的市盈率样本的标准差为 $s=7.8$（*The Wall Street Journal*，March 19, 1998）。假定我们对估计纽约证交所所有股票的总体平均市盈率感兴趣。如果我们需要的边际误差为 2 的话，样本中应包括多少只股票？使用 95% 的置信度。

8.4 总体比例的区间估计

在第 7 章我们介绍过样本比例 $\bar{p}$ 是总体比例 p 的无偏估计量，并且在大样本的情况下，$\bar{p}$ 的抽样分布可以用正态概率分布来近似，如图 8.9 所示。通过回忆我们知道在 np 和 $n(1-p)$ 都大于等于 5 的大样本条件下，正态分布才能作为 $\bar{p}$ 抽样分布的近似。当我们使用样本比例 $\bar{p}$ 估计总体比例 p 时，可以利用 $\bar{p}$ 的抽样分布对抽样误差进行概率描述。这时，抽样误差就定义为 $\bar{p}$ 和 p 之差的绝对值，记作 $|\bar{p}-p|$。

当样本容量较大时，对抽样误差可做如下的精度描述：

样本比例带来的抽样误差小于等于 $z_{\alpha/2}\sigma_{\bar{p}}$ 的概率为 $1-\alpha$。于是 $z_{\alpha/2}\sigma_{\bar{p}}$ 的值即为比例的标准误差。

一旦我们知道了边际误差 $z_{\alpha/2}\sigma_{\bar{p}}$，我们把 $\bar{p}$ 加减边际误差，就能够得到总体比例的区间估

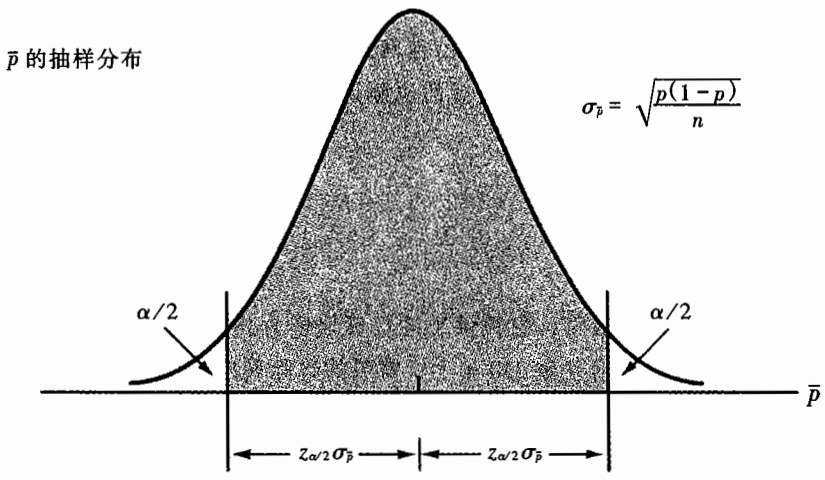


图 8.9　当 $np\geq 5$ 且 $n(1-p)\geq 5$ 时，$\bar{p}$ 抽样分布的正态近似的抽样分布

计。下面的公式给出了这样的区间估计：

$$\bar{p} \pm z_{\alpha/2}\sigma_{\bar{p}} \tag{8.7}$$

式中的 $1-\alpha$ 是置信系数。由于 $\sigma_{\bar{p}}=\sqrt{p(1-p)/n}$ ，我们能够把公式(8.7)重新表达为

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{8.8}$$

为了使用公式(8.8)建立总体比例 p 的区间估计，需要知道 p 的值。但由于 p 的值还需要我们估计，故只能用样本比例 $\bar{p}$ 代替 p。总体比例置信区间估计的一般表达式如下*：

当为样本比例扩展置信区间时，$\sigma_{\bar{p}}$ 可用作样本比例的标准 $z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})/n}$ 的值提供了误差的边际值。

总体比例的区间估计

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.9}$$

式中的 $1-\alpha$ 是置信系数，$z_{\alpha/2}$ 是与标准正态概率分布的上侧面积 $\alpha/2$ 相对应的 z 值。

我们使用下面的例子来说明边际误差和总体比例区间估计的计算过程。Ferrell Calvillo Communications 公司对 902 名国内高尔夫球女选手进行了一项调查，以了解女选手怎样看待自己在国内比赛的赛程安排。调查结果显示，有 397 名女选手对有下午茶时间感到满意。于是对取得下午茶时间感到满意的高尔夫女选手总体比例的点估计为 $397/902=0.44$。使用公式(8.9)和 95% 的置信水平，我们有

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

即

$$0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{902}}$$

$$0.44 \pm 0.0324$$

因此，边际误差为 0.0324，总体比例的 95% 置信区间估计是（0.4076，0.4724）。使用百分比表示，调查结果能够使我们以 95% 的置信度认为所有女选手中有 40.76% 到 47.24% 的人对取得下午茶时间感到满意。

---

* $\sigma_{\bar{p}}^2$ 的无偏估计是 $\bar{p}(1-\bar{p})/(n-1)$，则在公式(8.9)中应该用 $\sqrt{\bar{p}(1-\bar{p})/(n-1)}$ 代替 $\sqrt{\bar{p}(1-\bar{p})/n}$。但是在对总体比例进行估计时，我们通常使用的都是大样本，这时分母即使用 n，所产生的偏差也不会引起很大麻烦。所以对于使用 n 代替 $n-1$ 造成的结果数值上的差别，可以忽略不计。

确定样本容量

让我们来考虑这样一个问题：在给定的精度水平下，为了取得总体比例的估计值应该选择多大的样本容量？计算 p 的区间估计时确定样本容量的原理与 8.3 节估计总体均值时确定样本容量的原理相类似。

在本节的前面部分我们曾经指出，与总体比例的估计值相联系的边际误差是 $z_{\alpha/2}\sigma_{\bar{p}}$，由于 $\sigma_{\bar{p}}=\sqrt{p(1-p)/n}$，故边际误差由 $z_{\alpha/2}$ 的值、总体比例 p 和样本容量 n 共同决定。当置信系数 $1-\alpha$ 一定时，也就确定了 $z_{\alpha/2}$ 的值。然后，对于总体比例 p 的某特定值，边际误差就由样本容量 n 最后确定。样本容量越大，边际误差就越小，精度也就越好。

令 E 代表所希望的边际误差

$$E=z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

解出 n，我们得到下面的样本容量公式：

总体比例区间估计的样本容量

$$n=\frac{(z_{\alpha/2})^2p(1-p)}{E^2} \tag{8.10}$$

在公式(8.10)中，使用者已经规定了希望的边际误差 E，在大多数情况下，$E=0.10$ 或更小。使用者还规定了置信水平，于是也就确定了对应的 $z_{\alpha/2}$ 值。最后，由于总体比例 p 是未知的，使用公式(8.10)需要知道 p 的计划值。在实践中，通过下列方法之一，我们能够选定计划值。

1. 使用来自以前样本的样本比例，该样本要由与当前样本相同或类似的元素构成。
2. 使用小规模试验研究来选择初步样本。来自该样本的样本比例可以用作 p 的计划值。
3. 使用判断或“最佳猜测”作为 p 的值。
4. 如果以上三种方法均不适用，则以 $p=0.50$ 作为计划值。

因为总体比例 p 是需要我们从样本中估计的，为了使用公式(8.10)就必须得到 p 的计划值。此处介绍了四种取得计划值的方法。

让我们回到对女高尔夫球选手的调查上来，假设公司还想进行一项新的调查，以估计对取得下午茶时间感到满意的女选手的总体比例。如果调查主管需要在边际误差为 0.025，置信度为 95% 的条件下估计总体比例的话，他应该选择多大的样本容量？当 $E=0.025$ 且 $z_{\alpha/2}=1.96$ 时，我们需要 p 的计划值以回答样本容量问题。利用以前的调查结果 $\bar{p}=0.44$ 作为 p 的计划值，把数据代入公式(8.10)，我们有

$$n=\frac{(z_{\alpha/2})^2p(1-p)}{E^2}=\frac{1.96^2\times0.44\times(1-0.44)}{0.025^2}=1\,514.51$$

于是，样本容量至少必须达到 1 514.51 名高尔夫选手，才能满足边际误差的要求。取整到下一个整数值，则最终的样本容量为 1 515 人。

上面的第四种方法建议采用 $p=0.50$ 作为 p 的计划值。在无法取得其他信息的情况下，往往采用该值。为了理解这样做的理由，注意公式(8.10)的分子显示出样本容量与 $p(1-p)$的数值成正比，$p(1-p)$越大则样本容量越大。表 8.5 给出了 $p(1-p)$的一些可能值。注意到当 $p=0.50$ 时，$p(1-p)$的值达到最大。因此，在我们无法确定 p 的恰当计划值的情况下，$p=0.50$ 将产生最大的样本容量。所以出于安全或者保守的考虑，我们建议采用最大的可能样本容量。即使最终的比例不是 0.50，得到的精度也要好于预期。于是，当我们采用 $p=0.50$ 时，我们能够保证样本容量足够得到希望的边际误差。

如果在调查中询问许多关于比例的问题，我们往往采用 $p=0.50$ 作为样本容量计算的计划值。

在该例子中，如果采用 $p=0.50$ 的计划值，则我们得到样本容量为

$$n=\frac{(z_{\alpha/2})^2p(1-p)}{E^2}=\frac{1.96^2\times0.50\times(1-0.50)}{0.025^2}=1\,536.64$$

应该推荐比该数稍大一点的整数 1 537 个女高尔夫选手。

表 8.5　$p(1-p)$的一些可能值

p	$p(1-p)$	
0.10	$0.10\times0.90=0.09$	
0.30	$0.30\times0.70=0.21$	
0.40	$0.40\times0.60=0.24$	
0.50	$0.50\times0.50=0.25$	← $p(1-p)$的最大值
0.60	$0.60\times0.40=0.24$	
0.70	$0.70\times0.30=0.21$	
0.90	$0.90\times0.10=0.09$	

评　注

用来估计总体比例的边际误差几乎总是 0.10 或更小。在 Gallup 和 Harris 等组织进行的全国民意测验中，报告的边际误差通常为 0.03 或 0.04。根据这样的边际误差，公式(8.10)所计算出的样本容量几乎总是足够大的，能够满足 $np\geqslant5$ 和 $n(1-p)\geqslant5$ 的大样本要求。

练　习

方　法

31. 一个由400个元素组成的简单随机样本包含100个“是”的回答。



a. 总体中回答“是”的比例的点估计是多少？

b. 比例的标准误差$\sigma_{\bar{p}}$是多少？

c. 计算总体比例的95%置信区间。

32. 一个包含800个元素的简单随机样本的样本比例为$\bar{p}=0.70$。

a. 计算总体比例的90%置信区间。

b. 计算总体比例的95%置信区间。

33. 在一项调查中，总体比例p的计划值是0.35。则当边际误差为0.05时，为了得到总体比例的95%置信区间，应取多大的样本容量？

34. 在95%的置信度下，为了使总体比例估计值的边际误差为0.03，应选用多大的样本容量？假设无法得到过去的数据来建立p的计划值。

应　用

35. 一项《时代》/CNN民意测验向814名成年人询问他们对于美国国内事务状况的看法。提出的问题是：你认为目前美国的状况良好吗？总共有562人对问题回答了“是”。(*Time*, August 11, 1997)。

a. 认为目前状况良好的成年人总体比例的点估计是多少？

b. 在90%的置信水平下，边际误差是多少？

c. 认为目前状况良好的成年人总体比例的90%置信区间是什么？

36. 人力资源管理协会的一项调查询问了346名求职者，为什么员工如此频繁地变换工作(*The Wall Street Journal*, March 28, 2000)？受调查者选择最多（152次）的答案是“别处更高的补偿”。

a. 求职者把“别处更高的补偿”作为更换工作原因的总体比例的点估计是多少？

b. 总体比例的95%置信区间估计是什么？

37. Wirthlin Worldwide进行了一项调查，收集客户对零售商店服务质量的评价数据，结果发现有28%的美国人感到现在的服务比两年前要好(*USA Today*, January 20, 1998)。如果样本中包括650名成年人，建立成年人中认为现在的服务好于两年前的总体比例的95%置信区间。

38. ESPN Sports Zone 网站所收集的使用者简报数据显示，有 26% 的使用者为女性（*USA Today*，January 21, 1998）。假设该百分比是来自一个包括 400 名使用者的样本。
    a. 在 95% 置信水平下，女性使用者的估计比例的边际误差是多少？
    b. 女性使用者总体比例的 95% 置信区间是什么？
    c. 如果需要的边际误差为 3%，则应选取多大的样本容量？

39. 

雇员利益研究会进行的一项调查探索了小雇主向他们的雇员提供退休计划的原因（*USA Today*，April 4, 2000）。有 33% 的被调查小雇主认为是“为了具有竞争优势以招聘和留住好的员工”这个原因。
    a. 如果调查的目的是以 3% 的边际误差估计在提供退休计划的小雇主总体中，原因是“为了在招聘和留人上具有竞争优势”的人所占的比例，且置信水平为 95%，问应取多大的样本容量？
    b. 在 99% 的置信水平下，重复计算(a)中的问题。

40. 一项 Associated Press 民意测验调查了 1 018 名成年人，结果发现 255 人计划在 1998 年的假期中花费比往年同期更少的钱来购买礼品(ICR Media survey，November 13 – 17, 1998)。
    a. 所有成年人中，计划在 1998 年假期花费更少的钱用于购买礼品的人所占比例的点估计是多少？
    b. 在 95% 置信水平下，该估计的边际误差是多少？

41. 一项 American Express 零售调查发现，有 16% 的美国客户在 1999 年假期中曾经利用互联网购买礼品（*USA Today*，January 18, 2000）。如果有 1 285 名客户接受了调查，问边际误差是多少？在 95% 的置信水平下，计算总体中利用互联网购买礼品的客户所占比例的区间估计。

42. 一项关于 2000 年总统竞选的《今日美国》/CNN/Gallup 民意测验在 6 月选取了 491 名可能的投票人作为样本（*USA Today*，June 9, 2000）。民意测验的主要目的是估计在投票人的总体中，支持每个候选人的比例。假设总体比例的计划值为 $p = 0.50$，置信水平为 95%。
    a. 使用 $p = 0.50$，那么在 6 月的民意测验中计划的边际误差是多少？
    b. 随着日益临近 11 月的大选，需要更好的精度和更小的边际误差。假定下列边际误差是美国选举期间的调查所要求的，计算每一次调查的样本容量。

调查	边际误差
9 月	0.04
10 月	0.03
11 月	0.02
大选日前	0.01

43. 美国戏剧家与制作人联盟对百老汇剧院的观众进行了连续的追踪调查，以提供关于百老汇观众的最新信息（*Playbill*, Winter 1997）。每周该组织在百老汇轮流上演的剧目中随机地把一份一页的调查表分发到座位上去，该调查问卷仅用 5 分钟即可答完，使得观众能够与戏剧界进行交流。

a. 如果对任意总体比例要求的边际误差为0.04，则应选取多大的样本容量？设置信水平为95%，计划值为 $p=0.50$。

b. 假定已经使用了(a)中建议的样本容量，并且某一周的调查显示观众中有445名不住在纽约市。求百老汇的观众总体中，不住在纽约市的人所占比例的点估计。

c. 利用(b)中的数据，计算百老汇的观众中不住在纽约市的人所占比例的95%置信区间。

小 结

本章我们介绍了建立总体均值 μ 和总体比例 p 的置信区间的方法。建立置信区间的目的是提供估计精度的信息。置信区间越宽，表示精度越低，这时，可以增加样本容量以减少置信区间的宽度，提高估计的精度。

用来计算总体均值区间估计的表达式依赖于样本容量的大($n\geqslant30$)小($n<30$)，以及总体标准差是能够被假设为已知还是要用样本标准差 s 估计，有时还依赖于总体是否具有正态或近似正态的概率分布。如果样本容量较大的话，对总体的分布不需做任何假设，在区间估计的计算过程中使用 $z_{\alpha/2}$。如果样本容量较小的话，为了求得 μ 的区间估计，总体必须具有正态或近似正态的概率分布。这时如果总体具有正态分布的话，在假定 σ 已知的情况下，计算区间估计时使用 $z_{\alpha/2}$；而在用样本标准差 s 估计 σ 的情况下，则使用 $t_{\alpha/2}$。最后，当样本容量较小并且正态总体假设不符合实际情况时，我们建议把样本容量增加到 $n\geqslant30$，以便于使用大样本区间估计方法。

另外，我们还说明了如何确定样本容量，才能使 μ 和 p 的区间估计具有特定的边际误差。在实践中，进行总体比例的区间估计时所要求的样本容量通常较大。因此，我们给出了在 $np\geqslant5$ 且 $n(1-p)\geqslant5$ 的条件下，计算总体比例的大样本区间估计公式。

术语释义

区间估计： 一种总体参数的估计方法，它给出了一个被认为包含了参数值在内的区间。表达形式为点估计值 ± 边际误差。

边际误差： 为了建立总体参数的区间估计，点估计值需要加上和减去的那个值。

抽样误差： 无偏点估计量，比如样本均值 $\bar{x}$ 与对应的总体参数，比如总体均值 μ，这二者数值之差的绝对值被称为抽样误差。在参数为总体均值的情况下，抽样误差是 $|\bar{x}-\mu|$；在参数为总体比例的情况下，抽样误差是 $|\bar{p}-p|$。

精度表述： 对抽样误差所作的概率描述。

置信水平： 与区间估计相联系的置信度。例如，如果利用某区间估计方法得到的全部区间中有95%的区间包含总体参数，则该区间估计被称为是在95%的置信水平下构建的。

置信系数： 被表示为小数形式的置信水平。例如，0.95就是95%置信水平的置信系数。

***t* 分布：** 一系列的概率分布，当使用样本标准差 s 估计总体标准差 σ，并且总体具有正态或近似

正态的概率分布时，能够利用它来建立总体均值的区间估计。

自由度： t 分布的一个参数。我们在计算总体均值区间估计时使用的 t 分布具有 $n-1$ 的自由度，其中的 n 是简单随机样本的容量。

重要公式

用 μ 估计的抽样误差

$$|\bar{x}-\mu| \tag{8.1}$$

总体均值的区间估计：假定 σ 已知的大样本情况（$n\geqslant 30$）

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

总体均值的区间估计：用 s 估计 σ 的大样本情况（$n\geqslant 30$）

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.3}$$

总体均值的区间估计：假定 σ 已知的小样本情况（$n<30$）

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

总体均值的区间估计：用 s 估计 σ 的小样本情况（$n<30$）

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{8.5}$$

总体均值的区间估计所需的样本容量

$$n=\frac{(z_{\alpha/2})^2\sigma^2}{E^2} \tag{8.6}$$

总体比例的区间估计

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{8.9}$$

总体比例的区间估计所需的样本容量

$$n=\frac{(z_{\alpha/2})^2 p(1-p)}{E^2} \tag{8.10}$$

补充练习

44. 一项对首次购房者的调查发现，家庭年收入的平均值是 50 000 美元(*CNBC. com*, July 11, 2000)。假设该调查使用的样本包括 400 位首次购房者，收入的样本标准差为 20 500 美元。
    a. 在 95% 的置信水平下，该项研究的边际误差是多少？
    b. 首次购房者总体的平均家庭年收入的 95% 置信区间是什么？

45. 由美国汽车协会进行的一项调查显示，度假时一个四口之家平均日花费为 215.60 美元。假定选取去尼亚加拉大瀑布度假的 64 个四口之家为样本，其样本均值为 252.45 美元，样本标准差为 74.50 美元。
    a. 求去尼亚加拉大瀑布度假的四口之家总体平均日花费的 95% 置信区间。
    b. 使用(a)中求出的置信区间，去尼亚加拉大瀑布度假的四口之家日花费的均值与美国汽车协会所报告的均值是否存在差别？请解释。

46. 在 Accountemps 进行的一项调查中，要求一个由 200 名管理人员组成的样本提供每天办公室的职员们用于归位误标、误放项目所浪费的分钟数数据(*Astounding Averages*, by D. D. Dauphinais and K. Droste, 1995)。该项调查的数据包含在数据集 ActTemps 中。



CD 光盘数据
ActTemps

    a. 使用 ActTemps 中的数据，建立办公室职员每天用于归位误标、误放项目所浪费的总体平均时间的点估计。
    b. 样本标准差是多少？
    c. 总体每天平均浪费时间的 95% 置信区间是什么？

47. Arthur D. Little 公司估计一个家庭所收到的邮件中大约有 70% 是广告(*Time*, July 14, 1997)。一个包括 20 户家庭的样本显示了他们在一个星期内收到的邮件数和所收到的广告数的数据。



CD 光盘数据
Mail

家庭编号	广告数	邮件总数	家庭编号	广告数	邮件总数
1	24	35	11	13	19
2	9	14	12	16	28
3	18	30	13	20	27
4	9	12	14	17	22
5	15	28	15	21	24
6	23	33	16	21	33
7	13	20	17	15	25
8	17	20	18	15	24
9	20	23	19	18	24
10	20	25	20	12	16

    a. 家庭总体每周收到的广告平均份数的点估计是多少？该总体均值的 95% 置信区间是什么？

b. 家庭总体每周收到的邮件平均份数的点估计是多少？该总体平均值的 95% 置信区间是什么？

c. 使用(a)和(b)中求出的点估计值，这些估计值与大约 70% 的邮件是广告的说法相一致吗？

48. 《华尔街日报》的货币与投资部分总结了每天的股票交易、外汇、期权、贸易与期货等等的投资情况。在纽约证交所部分提供了每只股票的 52 周最高价、52 周最低价、红利率、收益、市盈率、每日成交量、每日最高价、每日最低价、每日收盘价和每日净变化的数据。每只股票的市盈率指标是用该股票的价格除以公司在最近四个季度所报告的每股收益而计算出来的。一个样本由 10 只从《华尔街日报》(September 29, 2000)选出的股票组成，它提供了下面的市盈率数据：5，7，9，10，14，23，20，15，3，26。

a. 在纽约证交所上市股票总体的平均市盈率的点估计是多少？

b. 在纽约证交所上市股票的总体市盈率标准差的点估计是多少？

c. 在 95% 的置信水平下，所有在纽约证交所上市股票的平均市盈率的区间估计是什么？假定总体具有正态分布。

d. 评价结果的精度。

49. 许多在大公司工作的美国人周末也要在家中工作或在办公室加班（*USA Today*, June 18, 1997)。在边际误差为 10 分钟的情况下，为了估计周末加班的平均时间应选取多大的样本容量？使用 95% 的置信区间并且假定总体标准差的计划值为 45 分钟。

50. 对某型号的汽车进行了里程测试，如果需要的精度是边际误差为每加仑 1 英里，置信区间为 98%，那么在测试中应该使用多少辆汽车？假设初步的里程测试显示标准差是每加仑 2.6 英里。

51. 在制定病人的预约时间表时，某医疗中心想要估计一个医生花费在每个病人身上的平均时间。如果在 95% 的置信水平下要求边际误差为 2 分钟，那么应该选取多大的样本容量？对于 99% 的置信水平，应选取多大的样本容量？设总体标准差的计划值为 8 分钟。

52. 《商业周刊》第 47 个年度薪金调查列出了首席执行官的年薪加奖金数据（*Business Week*, April 21, 1997）。一个初步的样本显示标准差为 675 美元，如果我们想在边际误差为 10 万美元的情况下估计首席执行官总体的平均年薪加奖金，那么样本中应包括多少个首席执行官？使用 95% 置信区间(注：如果数据是以千美元计，所希望的误差边际值应是 $E = 100$)。

53. 根据美国国家教育统计中心的报告，有 47% 的大学生为了赚取学费和生活费而工作(*The Tampa Tribune*, January 22, 1997)。假定该项研究选取了 450 名大学生作为样本。

a. 计算总体中为了支付学费和生活费而工作的大学生所占比例的 95% 置信区间。

b. 计算总体中为了支付学费和生活费而工作的大学生所占比例的 99% 置信区间。

c. 当置信水平从 95% 增加到 99% 时，边际误差将怎样变化？

54. 在《今日美国》/CNN/Gallup 对 369 对双职工父母进行的一项调查中，有 200 对父母承认，由于工作上的原因，他们花在子女身上的时间过少(*USA Today*, April 10, 1995)。

a. 求总体中认为花在子女身上时间过少的双职工父母所占比例的点估计。
b. 在95%的置信水平下，边际误差是多少？
c. 总体中认为他们花在子女身上的时间太少的双职工父母所占比例的95%置信区间估计是什么？

55. 一项《时代》/CNN的民意测验向1 400名美国成年人询问："你在闲暇时间愿意去哪里？"最多的回答是504人选择的大型商场。
a. 总体中在闲暇时间愿意去大型商场的成年人所占比例的点估计是多少？
b. 在95%的置信水平下，该估计的边际误差是多少？

56. 一家知名的银行信用卡公司想要估计月末余额不为零并且发生过利息支出的持卡人比例。假定在98%的置信水平下需要的边际误差为0.03。
a. 如果预期大约有70%的持卡人月末余额不为零，应选择多大的样本容量？
b. 如果无法得到总体比例的估计值，应该选择多大的样本容量？

57. 在一项调查中，有200人被要求确认他们主要的新信息来源。其中的110人认为他们主要的信息来源是电视新闻。
a. 建立总体中认为电视新闻是他们主要新信息来源的人所占比例的95%置信区间估计。
b. 当置信水平为95%，边际误差为0.05时，为了估计总体比例需要多大的样本容量？

58. 一项Roper Starch调查向18—29岁的雇员询问他们对于更好的健康保险和加薪两种选择，更喜欢哪一个(*USA Today*, September 5, 2000)。如果在500名雇员中有340人愿意选择更好的健康保险的话，请回答下列问题：
a. 18—29岁的雇员中愿意选择更好健康保险的雇员所占比例的点估计是多少？
b. 总体比例的95%置信区间估计是什么？

59. 《1997年美国统计摘要》公布了18岁及以上人群中吸烟者的百分比。假定要设计一项调查以收集吸烟者和非吸烟者新的数据，对吸烟者的总体比例最好的初步估计是30%。
a. 为了在边际误差为0.02时估计18岁及以上人群中吸烟者的总体比例，应选择多大的样本容量？使用95%的置信水平。
b. 假定该研究使用了你在(a)中建议的样本容量，并且在其中发现了520名吸烟者。吸烟者总体比例的点估计是多少？
c. 吸烟者总体比例的95%置信区间是什么？

60. 虽然航班时刻和费用是商务旅行者在选择航班时的重要考虑因素，但《今日美国》的一项调查发现，商务旅行者把航空公司的常客优惠计划列为最重要因素(*USA Today*, April 11, 1995)。在一个 $n = 1\,993$ 名商务旅行者的样本中，有618人把常客优惠计划作为首要因素。
a. 总体中在选择航班时认为常客优惠计划是首要因素的商务旅行者所占比例的点估计是多少？

b. 建立总体比例的 95% 置信区间估计。

c. 当置信水平为 95%，边际误差为 0.01 时，需要多大的样本容量？你是否会建议《今日美国》提供该精度水平？为什么？

案例研究 1　Bock 投资服务公司

Lisa Rae Bock 在 1994 年开办了 Bock 投资服务公司(BIS)，其目标是把 BIS 办成南加州地区领先的货币市场咨询服务公司。为了对目前的客户提供更好的服务并且吸引新的客户，她每周制作一份简讯。Lisa 一直在考虑向简讯中增加一项新内容，报道对基金经理的每周电话调查结果。为了调查该项服务的可行性，并确定在简讯中应包含哪些类型的信息，Lisa 选择了 45 家货币市场基金组成一个简单随机样本，所取得数据的一部分显示在表 8.6 中，它们报告了基金的净资产和过去 7 天与过去 30 天的收益率。在给货币市场基金经理打电话取得另外的数据以前，Lisa 决定对已经收集到的数据作一些初步分析。

管理报告

1. 使用恰当的描述性统计量来汇总货币市场基金的净资产和收益率数据。
2. 对于货币市场基金的总体，建立其平均净资产、平均 7 天收益率和平均 30 天收益率的 95% 置信区间估计。
3. 讨论 Lisa 在准备她的简报时，应该怎样使用这些类型的信息。
4. 为了向客户提供最有用的信息，你建议 Lisa 再收集什么其他信息？



CD 光盘数据
Bock

表 8.6　Bock 投资服务公司的数据

货币市场基金	净资产(百万美元)	7 天收益率(%)	30 天收益率 (%)
Amcore	103.9	4.10	4.08
Alger	156.7	4.79	4.73
Arch MM/Trust	496.5	4.17	4.13
BT Instit Treas	197.8	4.37	4.32
Benchmark Div	2 755.4	4.54	4.47
Bradford	707.6	3.88	3.83
Capital Cash	1.7	4.29	4.22
Cash Mgt Trust	2 707.8	4.14	4.04
Composite	122.8	4.03	3.91
Cowen Standby	694.7	4.25	4.19
Cortland	217.3	3.57	3.51

（续表）

货币市场基金	净资产(百万美元)	7天收益率(%)	30天收益率（%）
Declaration	38.4	2.67	2.61
Dreyfus	4 832.8	4.01	3.89
Elfun	81.7	4.51	4.41
FFB Cash	506.2	4.17	4.11
Federated Master	738.7	4.41	4.34
Fidelity Cash	13 272.8	4.51	4.42
Flex-fund	172.8	4.60	4.48
Fortis	105.6	3.87	3.85
Franklin Money	996.8	3.97	3.92
Freedom Cash	1 079.0	4.07	4.01
Galaxy Money	801.4	4.11	3.96
Government Cash	409.4	3.83	3.82
Hanover Cash	794.3	4.32	4.23
Heritage Cash	1 008.3	4.08	4.00
Infinity/Alpha	53.6	3.99	3.91
John Hancock	226.4	3.93	3.87
Landmark Funds	481.3	4.28	4.26
Liquid Cash	388.9	4.61	4.64
Market Watch	10.6	4.13	4.05
Merrill Lynch Money	27 005.6	4.24	4.18
NCC Funds	113.4	4.22	4.20
Nationwide	517.3	4.22	4.14
Overland	291.5	4.26	4.17
Pierpont Money	1 991.7	4.50	4.40
Portico Money	161.6	4.28	4.20
Prudential Money Mart	6835.1	4.20	4.16
Reserve Primary	1 408.8	3.91	3.86
Schwab Money	10 531.0	4.16	4.07
Smith Barney Cash	2 947.6	4.16	4.12
Stagecoach	1 502.2	4.18	4.13
Strong Money	470.2	4.37	4.29
Transamerica Cash	175.5	4.20	4.19
United Cash	323.7	3.96	3.89
Woodward Money	1 330.0	4.24	4.21

资料来源：*Barron's*, October 3, 1994。

案例研究 2　GULF 房地产公司

GULF 房地产公司位于佛罗里达西南部，该公司宣传自己是“房地产市场的专家”，它通过收集公寓住宅的位置、标价、售价和售出时间数据，来监测公寓住宅的销售情况。如果一套公寓直接坐落在墨西哥湾旁边的话，则被归类为**海景公寓**，如果一套公寓只是靠近海湾的话，则被归类为**非海景公寓**。从佛罗里达那不勒斯分支机构的销售服务目录中抽取样本，所取得的样本包括了 40 套海景公寓和 18 套非海景公寓的最新销售数据。*其中的价格以千美元为单位，表 8.7 显示了这些数据。

管理报告

1. 使用恰当的描述性统计量来汇总 40 套海景公寓三个变量中的每一个变量。
2. 使用恰当的描述性统计量来汇总 18 套非海景公寓三个变量中的每一个变量。

表 8.7　GULF 房地产公司的销售数据



CD 光盘数据
GulfProp

海景公寓			非海景公寓		
标价	售价	销售时间（天）	标价	售价	销售时间（天）
495.0	475.0	130	217.0	217.0	182
379.0	350.0	71	148.0	135.5	338
529.0	519.0	85	186.5	179.0	122
552.5	534.5	95	239.0	230.0	150
334.9	334.9	119	279.0	267.5	169
550.0	505.0	92	215.0	214.0	58
169.9	165.0	197	279.0	259.0	110
210.0	210.0	56	179.9	176.5	130
975.0	945.0	73	149.9	144.9	149
314.0	314.0	126	235.0	230.0	114
315.0	305.0	88	199.8	192.0	120
885.0	800.0	282	210.0	195.0	61
975.0	975.0	100	226.0	212.0	146
469.0	445.0	56	149.9	146.5	137
329.0	305.0	49	160.0	160.0	281
365.0	330.0	48	322.0	292.5	63
332.0	312.0	88	187.5	179.0	48
520.0	495.0	161	247.0	227.0	52

* 基于那不勒斯 MLS 的公寓销售数据(Coldwell Banker, June 2000)。

（续表）

海景公寓		
标价	售价	销售时间（天）
425.0	405.0	149
675.0	669.0	142
409.0	400.0	28
649.0	649.0	29
319.0	305.0	140
425.0	410.0	85
359.0	340.0	107
469.0	449.0	72
895.0	875.0	129
439.0	430.0	160
435.0	400.0	206
235.0	227.0	91
638.0	618.0	100
629.0	600.0	97
329.0	309.0	114
595.0	555.0	45
339.0	315.0	150
215.0	200.0	48
395.0	375.0	135
449.0	425.0	53
499.0	465.0	86
439.0	428.5	158

3. 比较你的汇总结果。对于任何能够帮助房地产公司更好理解公寓住宅市场的具体统计结果进行讨论。
4. 建立海景公寓总体平均售价和总体平均售出时间的95%置信区间估计，并解释你的结果。
5. 建立非海景公寓的总体平均售价和总体平均售出时间的95%置信区间估计，并解释你的结果。
6. 假定分支机构的经理要求在95%的置信水平下以40 000美元的边际误差估计海景公寓的总体平均售价，以15 000美元的边际误差估计非海景公寓的总体平均售价，则应该选取多大的样本容量？
7. 海湾房地产公司最近签署了两项新的代理销售合同：一套标价为589 000美元的海景公寓和一套标价为285 000美元的非海景公寓。你对最终售价以及所需销售时间的估计值是多少？

案例研究 3　大都会研究公司

大都会研究公司是一家消费者研究组织，它设计各种调查以评估消费者得到的产品和服务。在一项研究中，该公司希望了解消费者对底特律某个主要汽车厂商所产汽车性能的满意度。分发给该厂商的一种大型轿车用户的调查问卷发现了一些对该车型早期传动问题的抱怨。为了进一步了解传动故障的情况，大都会公司采用了底特律地区一家传动系统修理企业提供的传动系统实际维修记录为样本。下面的数据显示了 50 辆汽车在发生传动故障时已经行驶的英里数：


CD 光盘数据
Auto

85 092	32 609	59 465	77 437	32 534	64 090	32 464	59 902
39 323	89 641	94 219	116 803	92 857	63 436	65 605	85 861
64 342	61 978	67 998	59 817	101 769	95 774	121 352	69 568
74 276	66 998	40 001	72 069	25 066	77 098	69 922	35 662
74 425	67 202	118 444	53 500	79 294	64 544	86 813	116 269
37 831	89 341	73 341	85 288	138 114	53 402	85 586	82 256
77 539	88 798						

管理报告

1. 使用恰当的描述性统计量来汇总传动故障数据。
2. 对于有传动故障的汽车总体，建立其在发生传动故障时已经行驶的平均英里数的 95% 置信区间。
3. 根据一些汽车用户已经经历过早期传动故障的说法，讨论你得到的统计结果的含义。
4. 如果该研究公司想要在 95% 的置信区间下以 5 000 英里的边际误差估计总体在发生故障前行驶的平均英里数，那么应该选取多少维修记录组成样本？
5. 为了更加全面地评价传动故障问题，你还想收集其他什么信息？

附录 8.1　用 Minitab 求总体均值的区间估计

我们分成四种不同的情况说明如何使用 Minitab 建立总体均值的置信区间。

假定 σ 已知的大样本情况


CD 光盘数据
CJW

我们使用 8.1 节的 CJW 例子来说明这种情况。假设总体标准差已知为 $\sigma = 20$，100 位 CJW 客户的样本满意度分数已经输入 Minitab 工作表的 C1 列。我们能够使用下列步骤计算总体均值的 95% 置信区间估计：

步骤 1. 选择 **Stat** 下拉菜单；

步骤 2. 选择 **Basic Statistics** 选项；

步骤 3. 选择 **1-Sample Z**；

步骤 4. 当 1-Sample Z 对话框出现时：

在 **Variables** 框中输入 C1；

在 **Sigma** 框中输入 20；

步骤 5. 点击 **OK**。

Minitab 的默认值是 95% 的置信水平。为了指定其他的置信水平，以 90% 为例，应该在步骤 4 中加入下面内容：

选择 **Options**；

当 1-Sample Z 对话框出现时：

在 **Confidence level** 框中输入 90；

点击 OK。

用 s 估计 σ 的大样本情况



CD 光盘数据
Balance

我们使用表 8.2 中 85 户家庭样本的信用卡余额数据来说明这种情况。首先把数据输入 Minitab 工作表的 C1 列，在这种情况下，总体标准差 σ 是未知的，只能用样本标准差 s 估计。计算样本标准差的步骤如下：

步骤 1. 选择 **Calc** 下拉菜单；

步骤 2. 选择 **Column Statistics** 选项；

步骤 3. 当 Column Statistics 对话框出现时：

选择 **Standard deviation** 选项；

在 **Input variable** 框中输入 C1；

在 **Store result in** 框中输入 stdev；

点击 **OK**。

输出结果是样本标准差为 3 058，它作为 stdev 储存在 Minitab 中。

然后再按照假定 σ 已知的大样本情况下已经描述过的步骤，就能够得到总体均值的区间估计，惟一的变化在步骤 4，在这里用样本标准差 s 估计 σ。在 **Sigma** 框中输入 stdev 即可完成操作。

假定 σ 已知的小样本情况

假定 σ 已知的小样本情况使用与前面介绍过的假定 σ 已知的大样本情况相同的步骤。

用 s 估计 σ 的小样本情况

我们使用表 8.4 中 15 名雇员样本的培训时间来说明这种情况。首先把数据输入 Minitab 工作


CD 光盘数据
Training

表的 C1 列，在这个小样本情况下，由于总体标准差 σ 未知，只能用样本标准差 s 估计它。使用下面的步骤能够计算总体均值的 95% 置信区间估计：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Basic Statistics** 选项；
步骤 3. 选择 **1-Sample t**；
步骤 4. 当 1-Sample t 对话框出现时：
　　在 **Variables** 框中输入 C1；
步骤 5. 点击 **OK**。

Minitab 的默认值是 95% 置信水平。为了指定其他的置信水平，以 90% 为例，向步骤 4 中加入下列内容：

选择 **Options** 选项；
当 1-Sample t 对话框出现时：
在 **Confidence level** 框中输入 90；
点击 **OK**。

附录 8.2　用 Excel 进行总体均值的区间估计

我们分四种不同的情况说明如何使用 Excel 建立总体均值的置信区间。

假定 σ 已知的大样本情况


CD 光盘数据
CJW

我们以 8.1 节的 CJW 为例说明这种情况。假定已知总体标准差 $\sigma = 20$。首先把数据输入 Excel 工作表，即把标志 Score 输入单元格 A1，把 100 名 CJW 客户样本的满意度分数输入单元格 A2: A101。然后使用下面的步骤就能够计算总体均值的 95% 置信区间估计：

步骤 1. 在单元格 C2 计算样本均值；

= AVERAGE(A2: A101)

步骤 2. 在单元格 C3 计算边际误差*；

= CONFIDENCE(0.05, 20, 100)

步骤 3. 在单元格 C4 计算区间估计的下限；

= C2 – C3

* 边际误差函数的一般形式为 = CONFIDENCE(alpha, 标准差, 样本容量)，其中的 alpha = 1 – 置信系数 = 1 – 0.95 = 0.05。

步骤 4. 在单元格 C5 计算区间估计的上限;

$$= C2 + C3$$

加上标志后的格式化输出结果显示在图 8.10 中。

注:10 — 95 行隐藏了。

	Score			
	98	Sample Mean	82	
	79	Margin of Error	3.92	
	100	Lower Limit	78.08	
	86	Upper Limit	85.92	
	89			
	55			
	68			
	37			
	44			
	61			
	96			
	96			
	97			
	66			

图 8.10 CJW 总体均值区间估计的 Excel 输出结果

用 s 估计 σ 的大样本情况



CD 光盘数据
Balance

我们使用表 8.2 的信用卡余额数据来说明这种情况。首先把数据输入 Excel 工作表,即把标志 Balance 输入单元格 A1,把 85 户家庭样本的信用卡余额输入单元格 A2:A86。因为总体标准差 σ 未知,只能用样本标准差 s 估计它。在单元格 C1 计算样本标准差:

$$= STDEV(A2: A86)$$

然后使用与假定 σ 已知的大样本情况相同的步骤来建立区间估计。

步骤 1. 在单元格 C2 计算样本均值;

$$= AVERAGE(A2: A86)$$

步骤 2. 在单元格 C3 计算边际误差;

$$= CONFIDENCE(0.05, C1, 85)$$

步骤 2 的 Confidence 函数使用 C1 是由于用来估计 σ 的样本标准差储存在单元格 C1 中。

步骤 3. 在单元格 C4 计算区间估计的下限;

= C2 – C3

步骤 4. 在单元格 C5 计算区间估计的上限;

= C2 + C3

样本标准差显示在单元格 C1 中，其他的输出结果与图 8.10 的输出结果相似。

假定 σ 已知的小样本情况

如果总体具有正态分布的假设是恰当的，则假定 σ 已知的小样本情况与前面介绍过的假定 σ 已知的大样本情况使用同样的步骤。

用 s 估计 σ 的小样本情况


CD 光盘数据
Training

表 8.4 的数据显示了 15 名雇员样本的培训时间。首先把这些数据输入 Excel 工作表，即把标志 Time 输入单元格 A1，把培训时间输入单元格 A2: A16。在这种情况下，由于总体标准差 σ 未知，只能用样本标准差 s 估计它。我们能够使用下列步骤来计算总体均值的 95% 置信区间估计:

步骤 1. 选择 **Tools** 下拉菜单;

步骤 2. 选择 **Data Analysis** 选项;

步骤 3. 当 Data Analysis 对话框出现时:

选择 **Descriptive Statistics** 选项;

点击 **OK**。

步骤 4. 当 Descriptive Statistics 对话框出现时:

在 **Input Range** 框中输入 A1: A16;

选择 **Labels in First Row**;

选择 **Summary Statistics**;

选择 **Confidence Level for Mean** 并且在框中输入 95;

选择 **Output Range** 并且在框中输入 B1;

点击 **OK**。

样本均值 53.87 显示在单元格 C3 内，边际误差 3.78 显示在单元格 C16 中。注意 Excel 中边际误差的标志是 Confidence Level(95%)。通过向样本均值减去和加上边际误差，就能够得到置信区间。公式 = C3 – C16 能够用来在单元格 C17 中设置区间下限，而公式 = C3 + C16 则能够用来在单元格 C18 中设置区间的上限。

9 假设检验

统计实例:HARRIS 公司

9.1 原假设和备择假设的建立

检验研究中的假设

检验声明的有效性

决策中的假设检验

对原假设和备择假设形式的总结

9.2 第一类错误和第二类错误

9.3 总体均值的单侧检验:大样本

检验统计量应用

p - 值的应用

总体均值的单侧检验步骤总结

假设检验步骤

9.4 总体均值的双侧检验:大样本

双侧检验的 p - 值

总体均值的双侧检验步骤总结

区间估计与假设检验的关系

9.5 总体均值的检验:小样本

p - 值和 t 分布

双侧假设检验

9.6 总体比例的检验

9.7 假设检验和决策支持

9.8 计算第二类错误的概率

9.9 确定总体均值假设检验的样本容量

统计实例

HARRIS 公司*

墨尔本,佛罗里达州

Harris 公司是一家通信设备生产商,主要为全球市场提供无线电话、传播、管理和网络支持系统的产品和服务。这家公司在 80 多个国家进行销售和设备的售后服务。Harris 的大多数产品,包括印制电路装配线、最终产品装配线都需要中批量到大批量的制造和检验。

其中 RF 板装配线是公司大批量生产的产品之一。每个 RF 板由 16 个电子元件组成,它们都焊接在 RF 板的表层机器铸件上。在制造阶段的焊接过程中,发现 RF 板上的焊接流量没有达到产品质量标准。对可能影响焊接过程的大量因素进行考查,工程师初步确定,这一焊接问题极有可能是由 RF 板上的表层的疵点引起的。

工程师想进一步知道 Harris 存货中表层有疵点的 RF 板的比例是否超过供应商所规定的值。可令 p 代表 Harris 的 RF 存货中表层有疵点的板所占的比例,p_0 代表供应商所规定的值,建立如下假设:

$$H_0:\ p \leqslant p_0$$

$$H_a:\ p > p_0$$

H_0 代表 Harris 存货中表层有疵点的板所占的比例未超过供应商所规定的值。如果上述比例可以接受,则工程师将需要另外寻找焊接问题的原因。H_a 表示 Harris 存货中表层有疵点的板所占的比例超过供应商所规定的值。这时,表层的疵点很可能是引起焊接问题的原因,应当采取适当措施来找出造成存货中表层有疵点的 RF 板所占比例太高的原因。



Harris 公司运用了假设检验解决了电路板焊接过程中的问题。© CORRIS.

对 Harris 存货进行抽样检验的结果表明,应拒绝 H_0,而接受 H_a,这表明 Harris 存货中表层有疵点的板所占的比例超过了供应商所规定的值。通过对存货区域进一步调查发现,RF 板存储的架子受到污染是导致这一比例过高的原因。工程师改变了 RF 板存储的环境,从而解决了这个问题。

在本章中,我们将学习如何对总体均值和总体比例假设检验。通过对抽样样本数据分析,决定是否拒绝该假设。从而在研究中的假设检验、某项声明的有效性假设检验中以及决策检验中,可以得出合理解释并能采取适当的措施。

* 感谢 Harris 公司的 Richard A. Marshall 先生提供的材料。

在第 7 章和第 8 章中,我们解释了如何利用样本对总体参数进行点估计和区间估计。在本章中,我们将继续讨论在统计推断中如何使用假设检验来确定是否拒绝总体参数值的声明。

在假设检验中，我们首先对总体参数作一个试验性的假设，称该试验性的假设为**原假设**(null hypothesis)(零假设)，记作 H_0。然后，定义另一个与原假设的内容完全相反的假设，记作 H_a，称之**备择假设**(alternative hypothesis)。假设检验的过程是根据样本数据对这两个对立的假设 H_0 和 H_a 进行检验。

9.1　原假设和备择假设的建立

在一些应用中，如何建立原假设和备择假设并不是显而易见的。需要谨慎构建适当的假设，并且假设检验的结论是研究者或决策者所需的信息。在假设检验的过程中，建立原假设和备择假设的方法通常有三种情形。

正确表达假设需要通过实践学习。在开始选择适当的假设 H_0 和 H_a 时，会遇到一些困惑。本章将通过实际应用给出不同形式的 H_0 和 H_a。

检验研究中的假设

对于某一汽车模型，其汽油平均效能为每加仑 24 英里。某一产品研究小组专门开发一种新的燃料喷射器来提高每加仑油料的效能。为了评估这种新型的系统，生产出几个喷射器并安装在汽车上，通过驾驶对所研究的问题进行检验。产品研究小组通过寻找证据来证明新的系统确实能提高每加仑油料的效能。在这个例子中，研究的假设是新系统的平均每加仑油料效能将超过 24 英里，即 $\mu>24$。作为普遍的原则，研究的假设将被作为备择假设。因此，研究的原假设和备择假设如下：

$$H_0:\ \mu\leqslant 24$$
$$H_a:\ \mu>24$$

如果样本结果表明 H_0 没被拒绝，研究者也不能得出新的燃料喷射器效果更好的结论。这需要深入的研究调查和检验。然而，一旦数据结论表明 H_0 被拒绝，则研究者得出 H_a: $\mu>24$ 是正确的。在这个结论中，有统计数据证明新的系统提高了每加仑油料的效能。因此，可以投入生产这种新系统。在这样的研究中，原假设和备择假设的建立如下：拒绝 H_0 可得出研究结论是正确的，并可采取行动，所以研究中的假设将被作为备择假设。

研究中的假设一般是备择假设。如果原假设被拒绝，则研究中的假设为真。

检验声明的有效性

对于如何进行某一声明的有效性检验，我们以某一软饮料的生产商的声明为例。该生产商声称容量为2升的产品中至少有平均为67.6盎司的内含物。选取2升容量的产品为样本，对其内含物进行检测，以证明厂商的声明是否有效。在这种类型的假设检验中，除非样本能提供数据证明，我们将认为厂商的声明是正确的。对于软饮料这个例子我们将采用此方法，原假设和备择假设建立如下：

$$H_0: \mu \geq 67.6$$
$$H_a: \mu < 67.6$$

如果抽样结果表明 H_0 不被拒绝，就不能对生产商的声明表示异议。然而，如果抽样结果表明 H_0 被拒绝，则可推断 $H_a: \mu < 67.6$ 是真。有统计数据证明生产商的声明是错误的，软饮料的内含物平均含量小于67.6盎司。因此需对生产商采取适当措施。

对于任何关于检验某一声明的有效性的情形中，通常将假定为真的假设设为原假设。而将拒绝 H_0、统计数据证明原假设不正确的假设确定为备择假设。当拒绝 H_0 时，应采取措施纠正其声明。

生产商的声明一般被怀疑并将其设为原假设。如果原假设被拒绝，则该声明不正确。

决策中的假设检验

在检验研究中的假设或某一声明有效性时，如 H_0 被拒绝将会采取措施。然而，在一些情况下，不论 H_0 被拒绝与否，都需要采取相应措施。总体上，当决策者必须从这两种措施作出选择时，一种与原假设有关，另一种与备择假设有关，这时上述情况便发生了。例如，对一批收到的货物中的零件进行抽样检验，质量监督管理员决定是否收下这批货物，或者因为其质量不达标，将货物退还给供货商。假设这零件的规格要求是每个零件的平均长度是2英寸。如果平均长度大于或小于标准2英寸，则会在装配操作中出现质量问题。在这个例子中，原假设和备择假设建立如下：

$$H_0: \mu = 2$$
$$H_a: \mu \neq 2$$

如果抽样结果表明 H_0 不被拒绝，质量监督管理者可毫无疑问地认为这批货物达到规定标准，并可接受这批货物。然而，如果抽样结果表明被拒绝，这表明这批零件不符合规定标准。因此，质量监督管理者有足够的理由退货。在这里我们看到了不论 H_0 被拒绝还是没被拒绝都采取

措施的例子。

对原假设和备择假设形式的总结

令 μ_0 为原假设和备择假设中的指定值。总体上，关于总体平均值 μ 的假设检验的形式是下面三种之一：

$$H_0: \mu \geq \mu_0 \qquad H_0: \mu \leq \mu_0 \qquad H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \qquad H_a: \mu > \mu_0 \qquad H_a: \mu \neq \mu_0$$

这里给出了假设检验 H_0 和 H_a 的三种形式，三种形式中等号总是出现在原假设中。在特定的应用中，将你想证明的问题设为备择假设 H_a，则在选择合适假设时有帮助。

在许多情况下，H_0 和 H_a 的选择并非显而易见，需要通过判断来选择合适的形式。在上述的形式中表达式中的等号部分(不论≥、≤或=)总是出现在原假设中。在选择 H_0 和 H_a 适当形式时，要记住将试图建立的结果记为备择假设。因此，可以通过询问使用者寻找支持 $\mu<\mu_0$，$\mu>\mu_0$，$\mu\neq\mu_0$ 的证据来帮助确定 H_a。下面的习题是选择假设检验的适当形式的练习。

练　习

1. Danvers-Hilton 风景旅馆的管理者声称，周末顾客账单的均值小于或等于 600 美元。该旅馆的会计发现最近几个月客人的账单上的费用持续增加。因此，会计想根据周末账单所组成的样本来验证管理者的说法。

a. 验证管理者的说法，应采用下面哪种假设检验形式？为什么？

$$H_0: \mu \geq 600 \qquad H_0: \mu \leq 600 \qquad H_0: \mu = 600$$
$$H_a: \mu < 600 \qquad H_a: \mu > 600 \qquad H_a: \mu \neq 600$$

b. 如不能拒绝 H_0，则得到适当的结论是什么？

c. 如可以接受 H_0，则得到适当的结论是什么？

2. 自测题　一家汽车销售商的管理者正考虑实施一奖励计划来提高销售量。目前，平均销售量为每月 14 辆汽车。管理者想知道这项新型的奖励计划能否增加销售量。为了收集这项计划的数据，允许被选定的销售人员在一个月内采用这项新型的奖励计划进行销售。

a. 对于此项研究，建立合适的原假设和备择假设。

b. 如不能拒绝 H_0，试对此结论进行解释。

c. 如可拒绝 H_0，试对此结论进行解释。

3. 设计一条产品生产线，这条生产线是用于将洗衣用的洗涤剂装入纸盒，要求纸盒的重量的均值为 32 盎司。定期挑选一部分纸盒为样本，对其称量以便发现其是否分量不足或超重。如果样本数据得出纸盒重量不足或超重，则应该使该生产线停产并调整使其达到规定重量。
   a. 建立原假设和备择假设，以判断是否关闭和调整生产线。
   b. 如不能拒绝 H_0，试对此结论进行解释。
   c. 如拒绝 H_0，试对此结论进行解释。

4. 因为时间和成本对产量的影响很大，所以在采用某种新的工艺方法之前，生产负责人必须使管理层相信其新的工艺方法确实能降低成本。目前生产中所用的工艺方法所耗的成本均值为每小时 220 美元。为了检验该新的工艺方法，测量其一段生产期作为样本的成本。
   a. 对于此新的工艺方法，建立合适的原假设和备择假设。
   b. 如不能拒绝 H_0，试对此结论进行解释。
   c. 如可以拒绝 H_0，试对此结论进行解释。

9.2 第一类错误和第二类错误

原假设和备择假设是关于总体的两个对立的解释。要么原假设 H_0 是真，要么备择假设 H_a 是真，两者不能同时为真。理想的假设检验过程为当 H_0 为真时，接受 H_0，当 H_a 为真时，拒绝 H_0。然而，作出的结论不可能总是正确的。因为假设检验是建立在样本信息的基础之上，我们必须考虑错误发生的概率。表 9.1 表示了假设检验中可能发生的两种类型的错误。

表 9.1 的第一行表明当接受 H_0 时可能发生的情况。如果 H_0 为真，则得到的结论是正确的。然而，如果 H_a 为真，我们就犯了**第二类错误**(type Ⅱ error)，即原假设 H_0 为假而我们接受它时所犯的错误。表 9.1 的第二行表示拒绝 H_0 时可能发生的情况。如果 H_0 为真，我们犯了**第一类错误**(type Ⅰ error)，即原假设 H_0 为真而我们拒绝它时所犯的错误。如果 H_a 为真，则拒绝 H_0 是正确的。

表 9.1 假设检验中的错误和正确结论

		总体情况	
		H_0 为真	H_a 为真
结论	接受 H_0	正确结论	第二类错误
	拒绝 H_0	第一类错误	正确结论

在 9.1 节中，我们讨论了一个假设检验例子。一个汽车生产研究小组设计了一种能提高汽车每加仑油料效能的新型的燃料喷射器。目前所用的燃料喷射器的功效为平均每加仑 24 英里，其假设检验建立如下：

$$H_0: \mu \leqslant 24$$
$$H_a: \mu > 24$$

备择假设 H_a：$\mu>24$ 表明，研究者正试图找出能支持效率均值大于每加仑 24 英里的结论的样本数据。

在这个例子中，当 H_0 为真时而拒绝了 H_a，表示这种新型喷射器实际上不比目前所用的喷射器好，但却错误认为这种新型的喷射器提高了每加仑燃料的功效($\mu>24$)，则此时发生了第一类错误。反之，当 H_0 为假时而接受了 H_a，表示这种新型喷射器实际上比目前所用的喷射器好，但错误认为这种新型的喷射器没有提高每加仑燃料的功效($\mu\leqslant 24$)，则此时发生了第二类错误。

在实践中，人们将假设检验中所发生的第一类错误的概率最大允许值称为检验的**显著性水平**(level of significance)。一般显著性水平值为 0.05 和 0.01。参照表 9.1 第二行，发现如果**拒绝** H_0，要么这一结论正确，要么就发生第一类错误。因此，如果第一类错误发生的概率被限制在一个较小的显著性水平之内，则我们将有很高的置信度确定拒绝 H_0 是正确的。在这种情形下，从统计上我们可以得出 H_0 为假，而 H_a 为真的结论，从而可按照备择假设 H_a 建议的措施实施。

在绝大多数的假设检验运用中，一般都只对第一类错误进行控制，而通常不对第二类错误加以控制。因此，我们在决定接受 H_a 时，我们不能确定该决策有多大的置信度。由于第二类错误发生的不稳定性，统计学家建议我们采用“不能拒绝 H_0”而不采用“接受 H_0”这种说法。采用“不能拒绝 H_0”这种说法使我们避免了为判断或行动提供建议。实际上，不能直接接受 H_0，从而使统计学家避免了犯第二类错误的风险。但只要不对第二类错误控制，我们将不能得出接受 H_0 的结论。在这种情况下，只能得出两种可能的结论：**不能拒绝 H_0 或拒绝 H_0**。

如果样本数据与原假设 H_0 一致，则我们得出“不能拒绝 H_0”而不是“接受 H_0”的结论，因为接受 H_0 会发生第二类错误。

虽然通常不对第二类错误进行控制，但是事实我们做了。事实上，在 9.7 节和 9.8 节中，我们将介绍第二类错误概率和控制的程序。如果能对第二类错误进行适当控制，那么可以实施基于“不能拒绝 H_0”的结论所采取的措施。

评　注

在许多应用中假设检验都是为了决策。通过统计数据得出**拒绝** H_0，H_a 为真的结论，并采取适当的措施。虽然“不能拒绝 H_0”这种说法并未能作出结论，但使管理者认为似乎 H_0 是真的。在这种情形下，管理者需知道采取这一措施导致第二类错误的事实。

练　习

5. 自测题　美国人每天平均用于阅读报纸的时间为 8.6 分钟(*USA Today*，April 10, 1995)。某一研究者认为管理岗位上的人每天平均用于阅读报纸的时间比公众的平均值要高。该研究

者选取一部分管理岗位的人作为样本。根据他们每天用于阅读报纸时间的数据对下面的原假设和备择假设进行检验：

$$H_0: \mu \leqslant 8.6$$
$$H_a: \mu > 8.6$$

a. 在这种情况下，第一类错误发生的概率是什么？发生这类错误将导致什么后果？
b. 在这种情况下，第二类错误发生的概率是什么？发生这类错误将导致什么后果？

6. 容量为3夸脱的橙汁容器上的标签标明，该橙汁中所含脂肪的均值不超过1克。对标签上的说明进行检验，回答下面问题：
a. 建立原假设和备择假设。
b. 在这种情况下，第一类错误发生的概率是什么？发生这类错误将导致什么后果？
c. 在这种情况下，第二类错误发生的概率是什么？发生这类错误将导致什么后果？

7. Carpetland销售员每周的销售额为8 000美元，该公司的副总裁Steve Contois建议采取一种刺激销售的补偿方案。Steve想通过试行该计划的一个销售期间的数据，得出该方案能增加每名销售人员的平均销售量的结论。
a. 建立原假设和备择假设。
b. 在这种情况下，第一类错误发生的概率是什么？发生这类错误将导致什么后果？
c. 在这种情况下，第二类错误发生的概率是什么？发生这类错误将导致什么后果？

8. 如果假设检验支持某种生产方法能够降低每小时生产成本均值的结论，则这种新型生产方法可正式投入使用。
a. 目前该生产方法的平均成本为每小时220美元，试建立原假设和备择假设。
b. 在这种情况下，第一类错误发生的概率是什么？发生这类错误将导致什么后果？
c. 在这种情况下，第二类错误发生的概率是什么？发生这类错误将导致什么后果？

9.3 总体均值的单侧检验：大样本

美国联邦贸易委员会（FTC）定期设计调查对生产商的产品说明进行检验。例如，大瓶Hilltop咖啡的标签上标明其容量至少为3磅。假定我们利用假设检验对此项说明进行检验。

假设检验的第一步是建立适当的原假设和备择假设。在对Hilltop咖啡进行检验时，美国联邦贸易委员会主要是寻找消费者所消费的产品少于厂商所声明的产品。因此，如果Hilltop的罐装咖啡的总体平均重量为3磅**或多于**3磅，则其产品说明是正确的。但是，如果Hilltop的罐装咖啡的总体平均重量**小于**3磅，则其产品说明是不正确的。因此，联邦贸易委员会可对其产品不足量进行起诉，并对公司采取相应的法律措施。

首先我们假定Hilltop咖啡的标签的说明是正确的。令μ代表每罐重量的总体均值，则原假设和备择假设如下：

$$H_0:\ \mu \geqslant 3$$
$$H_a:\ \mu < 3$$

如果样本数据表明 H_0 没被拒绝，就不能对 Hilltop 采取任何措施。但是，如果样本数据表明 H_0 被拒绝，则美国联邦贸易委员会根据统计数据得出备择假设 H_a：$\mu<3$ 是正确的。在这种情况下，联邦贸易委员会可对其产品重量不足起诉。

假定随机抽取 36 听咖啡罐头作为样本，对样本的平均每听罐头的重量进行假设检验。如果样本平均值小于 3 磅，则对 Hilltop 咖啡的标签的说明产生怀疑。但是，当样本均值比 3 磅小多少的时候，我们才愿意拒绝 H_0 并承担犯第二类错误的风险控告该公司违背其说明？

为了回答这个问题，我们假定原假设 $\mu=3$ 为真。另外，从以前对 Hilltop 咖啡罐头检验的数据得出总体标准方差 $\sigma=0.18$ 磅。在第 7 章和第 8 章的抽样分布中了解到 Hilltop 咖啡罐头检验中的 $\bar{x}$ 的抽样分布是服从正态概率分布。图 9.1 给出了当原假设 $\mu=3$ 为真时，$\bar{x}$ 的抽样分布。现在我们利用图 9.1 中的样本平均值和抽样分布对假设 H_0：$\mu\geqslant 3$ 和 H_a：$\mu<3$ 检验。

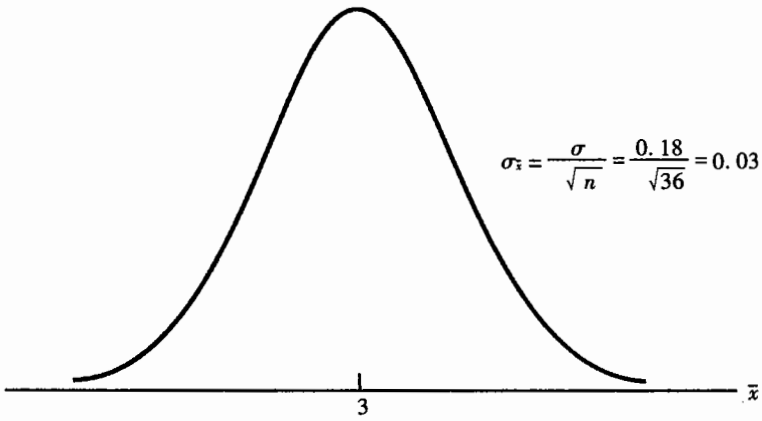


图 9.1　在 Hilltop 咖啡研究中，当原假设($\mu=3$)为真时 $\bar{x}$ 的抽样分布

在假设检验的方法中，我们事先指定犯第一类错误的最大允许概率，称此概率为检验的**显著性水平**(level of significance)。令 α 为显著性水平，α 指原假设为真时，第一类错误发生的概率。在 Hilltop 咖啡检验中，联邦贸易委员会的负责人作了如下说明："如果 Hilltop 产品重量达到规定要求($\mu=3$)，则在 99% 情况下不会对这家公司采取任何措施。而我只有 1% 的可能去犯对该公司的质量不足进行投诉的错误"。在这里，从联邦贸易委员会的负责人的说明了解到第一类错误的概率为 $\alpha=0.01$。图 9.2 表示了 $\bar{x}$ 的抽样分布和显著性水平 $\alpha=0.01$ 时的拒绝区域，这表明当 H_0：$\mu=3$ 为真时，$\bar{x}$ 只有 1% 的可能性落在**拒绝域**(rejection region)，即有 1% 的可能性犯第一类错误。

样本观察平均值 $\bar{x}$ 可通过计算**检验统计量**(test statistic)来得出是否落在拒绝域内。在大样本情形下，假定总体平均误差 σ 已知，检验统计量如下：

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

在这个表达式中，μ_0 为 H_0 中的总体平均值，n 为样本大小。统计量 z 的抽样分布是一个标准正态概率分布，因此可对 z 值估计判断样本平均值是否落在拒绝域内。在 Hilltop 咖啡例子中，$\mu_0=3$，$\sigma=0.18$，$n=36$。因此，一旦样本平均值$\bar{x}$ 确定，检验统计量为

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{\bar{x}-3}{0.18/\sqrt{36}}$$

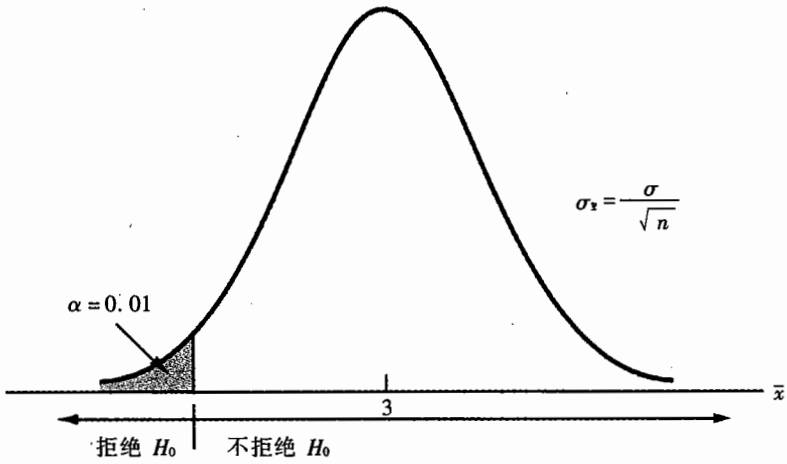


图 9.2　在 Hilltop 咖啡研究中，显著性水平 $\alpha=0.01$ 时的拒绝区域

现在我们可以通过样本结果得出假设检验的结论。我们介绍两种方法：一种是基于检验统计量的观察值，另一种是基于 p - 值准则。首先我们介绍检验统计量方法。

检验统计量应用

图 9.3 显示了$\bar{x}$ 的抽样分布和检验统计量 z 的抽样分布。每一个$\bar{x}$ 值对应一个 z 值。在前面，我们指定了 Hilltop 咖啡研究中显著性水平 $\alpha=0.01$。图 9.3 给出了 H_0 的拒绝域。

因为检验统计量 z 服从标准正态概率分布，我们可以通过标准正态分布表查找与分布在左侧的 $\alpha=0.01$ 区域相对应的 z 值。查表可得 $z=-2.33$，这在图 9.3 中表明。拒绝域的边界的检验统计值为检验**临界值**(critical value)。使用临界值 $z=-2.33$，从而我们建立 Hilltop 假设检验的拒绝规则

如果 $z<-2.33$，则拒绝 H_0

拒绝规则表明如样本平均值$\bar{x}$ 计算出的概率统计值小于 -2.33，则我们拒绝原假设 H_0：$\mu\geq 3$，得出备择假设 H_a：$\mu<3$ 是正确的。如果样本平均值$\bar{x}$ 计算出的检验统计量 z 值不在拒绝域中，$z<-2.33$，我们不能拒绝 H_0：$\mu\geq 3$。图 9.3 表示拒绝域在抽样分布的一侧，故这种情况叫**单侧**(one-tailed)检验。

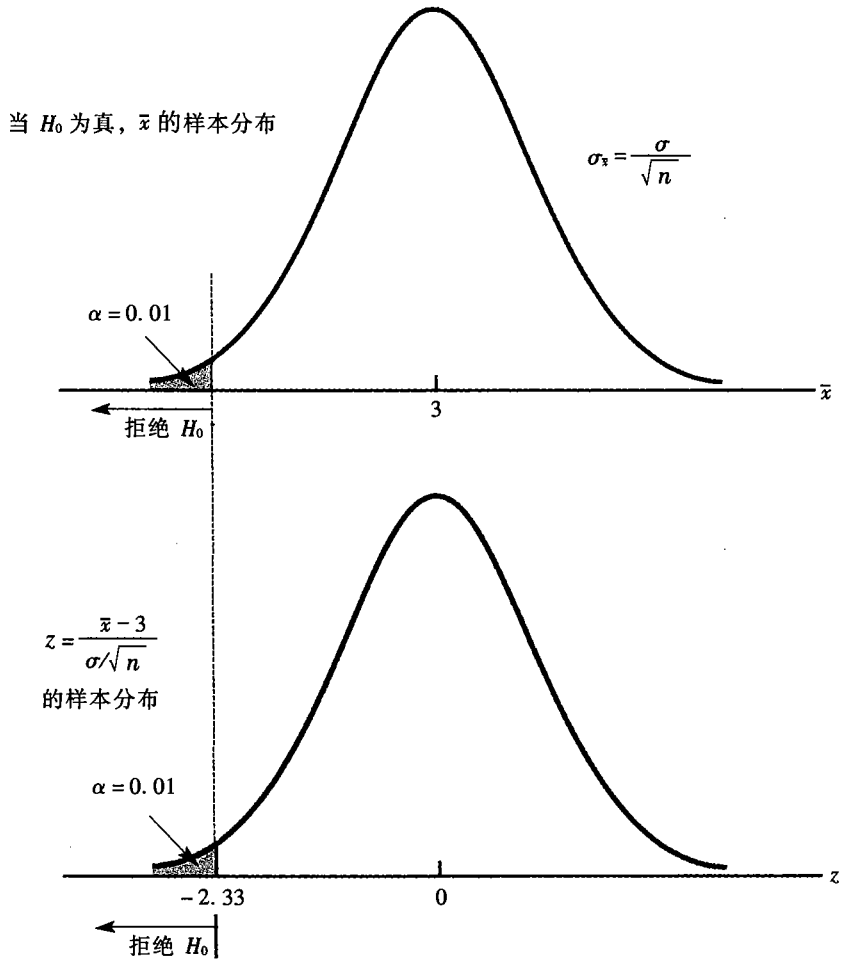


图 9.3　对于 Hilltop 咖啡，当 $\alpha=0.01$ 时$\bar{x}$ 的抽样分布和 z 的抽样分布

假定由 36 听罐头组成的样本平均值$\bar{x}=2.92$ 磅。是不是 $\bar{x}=2.92$ 落入拒绝域？为了回答这个问题，我们计算检验统计量 z

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{2.92-3}{0.18/\sqrt{36}}=-2.67$$

比较 $z=-2.67$ 与检验的临界值 $z=-2.33$，$z=-2.67<-2.33$。因而，$z=-2.67$ 落入拒绝域（如图 9.4 所示，见下页）。因此，我们拒绝 H_0：$\mu\geqslant 3$。所以，联邦贸易委员会得出 H_a：$\mu<3$ 正确，联邦贸易委员会可对 Hilltop 咖啡重量不足起诉。

p - 值的应用

p - 值（p-value）检验是另一种进行假设检验的方法。假定原假设为真，p - 值是样本结果与所

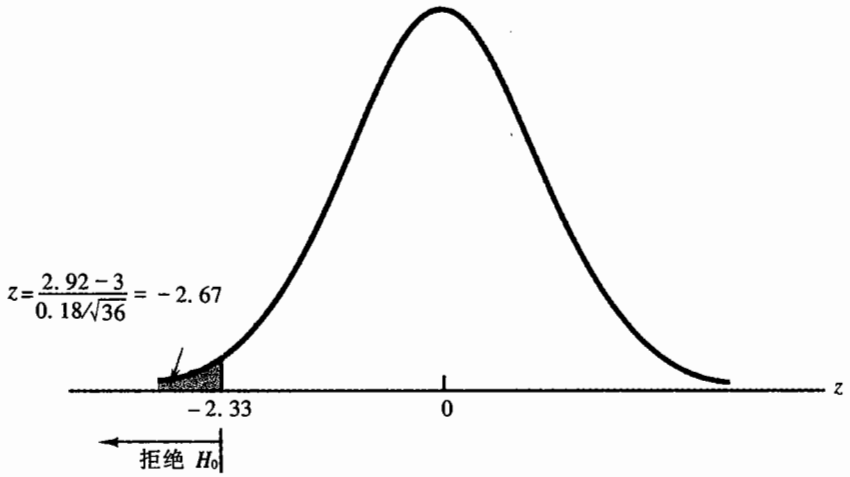


图 9.4　当$\bar{x}=2.92$时，其相应检验统计量的值落在拒绝域中

观察到的结果至少有一点不同的概率。在 Hilltop 咖啡例子中，拒绝域在左侧，所以 p－值是观测到的样本结果小于等于实测结果的概率。通常称 p－值为**观察显著性水平**。

在 Hilltop 咖啡例子中，样本平均值 $\bar{x}=2.92$ 磅对应着检验统计值 $z=-2.67$。因为拒绝域在分布的左侧，故 p－值是 z 值小于 -2.67 的概率。查找标准正态分布表，发现 $z=-2.67$ 时，概率为 0.496 2，所以如图 9.5 所示 p－值为 0.500 0－0.496 2＝0.003 8。p－值表明当 H_0 为真时，所得到的样本值 $\bar{x}=2.92$ 的概率很小。

单侧检验的 p－值总是这样一块面积：由检验统计值起向分布的拒绝域一侧扩展。

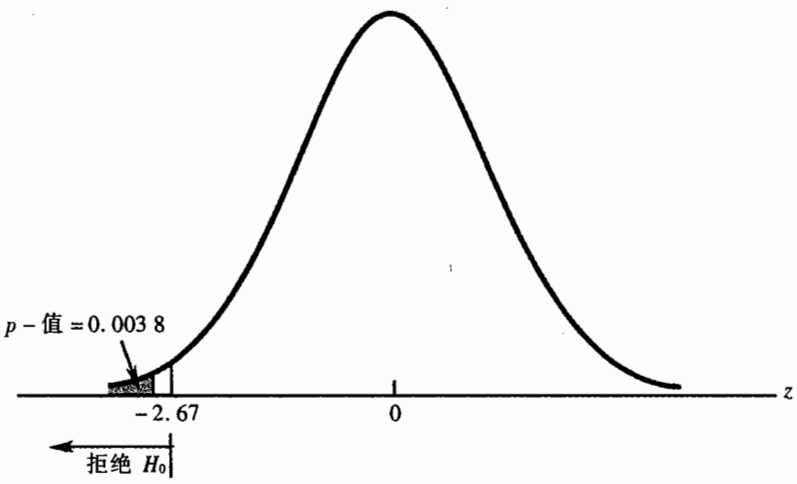


图 9.5　在 Hilltop 咖啡研究中，当 $\bar{x}=2.92$ 和 $z=-2.67$ 时的 p－值

p - 值拒绝规则如下：

如果 p - 值$<\alpha$，则拒绝 H_0

不论假设检验的形式，总是如果 p - 值$<\alpha$，则拒绝 H_0。

如果 p - 值$<\alpha$，p - 值在拒绝域内，则样本结果必定在拒绝区域内。在 Hilltop 咖啡例子中，p - 值 = 0.003 8$<\alpha$ = 0.01，因此拒绝原假设 H_0：$\mu \geqslant 3$。联邦贸易委员会可得出 H_a：$\mu < 3$ 为真的结论。因此，联邦贸易委员会可对 Hilltop 咖啡重量不足起诉。

检验统计量方法和 p - 值方法对同一假设检验应得出相同结果。选择这种方法而不选择那种方法完全是个人爱好。可选择其中一种，或两种都可以。计算机输出一般使用 p - 值方法。事实上，因为 p - 值准则易记易用，p - 值$<\alpha$ 经常被用作拒绝规则。

有时对某一特定的样本结果，研究者仅仅是设定 p - 值。研究结论的使用者可以选择任何能够接受的显著性水平 α。计算 p - 值和被选择的 α 值，可使使用者决定能否拒绝原假设。

在结束 Hilltop 咖啡例子之前，我们假设 36 听罐头组成的样本平均值 $\bar{x}$ = 2.97 磅。在这种情形中，检验统计值为

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.97 - 3}{0.18/\sqrt{36}} = -1.00$$

由于 $z = -1.00$ 远大于临界值 -2.33，$z = -1.00$ 不落在拒绝域内。因此，联邦贸易委员会不能拒绝 H_0。联邦贸易委员会不能根据统计数据来证明对 Hilltop 咖啡重量不足，故不能对其起诉。

查找标准正态分布表，当检验统计量 $z = -1.00$ 时，可得 $z = -1.00$ 与平均值之间的面积为 0.341 3。因此，拒绝侧面积即 p - 值 = 0.500 0 - 0.341 3 = 0.158 7。因为 p - 值 = 0.158 7 远大于 α = 0.01，则联邦贸易委员会不能拒绝 H_0。因此，没有统计数据来支持对 Hilltop 咖啡重量不足采取措施。

总体均值的单侧检验步骤总结

总体均值单侧检验的步骤如下。根据中心极限定理，在大样本情形下（$n \geqslant 30$），我们可以得出的抽样分布近似服从正态概率分布。对于大样本，我们可以利用历史数据值作为总体标准误差 σ 的值，或以样本标准差 s 代替 σ。在下面的单侧检验中，总体平均值 μ_0 已被指定。

大样本（$n \geq 30$）情形下总体均值单侧假设检验的形式

$$H_0: \mu \geq \mu_0$$
$$H_a: \mu < \mu_0$$

检验统计量：σ 已知

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \tag{9.1}$$

检验统计量：σ 未知

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \tag{9.2}$$

拒绝规则

使用检验统计量：如 $z < -z_\alpha$，则拒绝 H_0

使用 p - 值：如 p - 值 $< \alpha$，则拒绝 H_0

当检验统计量在抽样分布的右侧时，得出单侧假设检验拒绝原假设的第二种形式。单侧检验及拒绝规则总结如下(见图 9.6)。仍然考虑大样本情形。

在大多数应用中，因为总体标准差 σ 未知，在计算检验统计量时以样本标准差 s 代替。

大样本（$n \geq 30$）情形下总体均值单侧假设检验的形式

$$H_0: \mu \leq \mu_0$$
$$H_a: \mu > \mu_0$$

检验统计量：σ 已知

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \tag{9.1}$$

检验统计量：σ 未知

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \tag{9.2}$$

拒绝规则

使用检验统计量：如 $z > z_\alpha$，则拒绝 H_0

使用 p - 值：如 p - 值 $< \alpha$，则拒绝 H_0

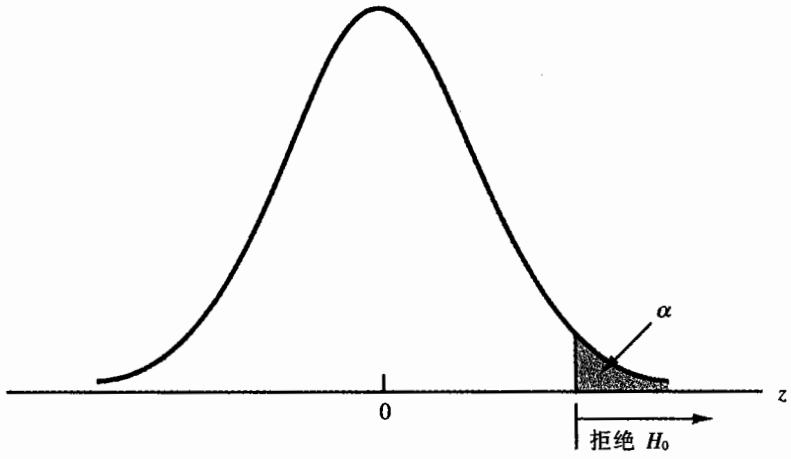


图 9.6　总体均值假设检验右侧的拒绝区域

假设检验步骤

在对 Hilltop 咖啡例子进行的假设检验步骤，同样适用于任何假设检验。其步骤总结如下，注意，无论检验统计量或 p – 值都可以得出假设检验结论。

假设检验步骤

1. 建立原假设和备择假设。
2. 指定显著水平 α。
3. 选择假设检验的检验统计量。

使用检验统计量

4. 利用显著性水平根据检验统计量的值建立拒绝 H_0。
5. 收集样本数据和计算检验统计值。
6. 使用检验统计量和拒绝规则，确定是否来拒绝 H_0。

使用 p – 值

4. 收集样本数据来计算检验统计量。
5. 使用检验统计量来计算 p – 值。
6. 如果 p – 值$<\alpha$，则拒绝 H_0。

评　注

p – 值是当原假设为真时，样本结果出现的概率。p – 值越小，则样本结果出现的概率越小。大多数统计软件包提供了与假设检验相对应的 p – 值，使用者可以将 p – 值与 α 相比较，而不必查找统计表便可得出假设检验的结论。

练　习

学生注意：一些假设检验的练习是使用检验统计量，另外一些练习是使用 p - 值。这两种方法都得出了相同的假设检验结论。下面提供了利用这两种方法的练习，可以使用这两种方法。在实际应用中，你可按自己的爱好选择其中一种方法。

方　法

9. 考虑下面假设检验：

$$H_0: \mu \geqslant 10$$
$$H_a: \mu < 10$$

一个样本大小为 50，样本平均值为 9.46，样本标准差为 2。

a. $\alpha = 0.05$，z 的临界值是多少？

b. 计算检验统计量 z 值，得出你的结论。

10. 考虑下面假设检验：

自测题

$$H_0: \mu \leqslant 15$$
$$H_a: \mu > 15$$

一个样本大小为 40，样本平均值为 16.5，样本标准差为 7。

a. $\alpha = 0.02$，z 的临界值是多少？拒绝规则是什么？

b. 计算检验统计量 z 的值。

c. p - 值为多少？

d. 得出你的结论。

11. 考虑下面假设检验：

$$H_0: \mu \geqslant 25$$
$$H_a: \mu < 25$$

样本大小为 100，总体标准差为 12，取 $\alpha = 0.05$。给出检验统计量 z 的值，并得出你的结论。

a. $\bar{x} = 22.0$　　b. $\bar{x} = 24.0$　　c. $\bar{x} = 23.5$　　d. $\bar{x} = 22.8$

12. 考虑下面假设检验：

$$H_0: \mu \leqslant 5$$
$$H_a: \mu > 5$$

假定检验统计量如下，计算相应的 p - 值，与 $\alpha=0.05$ 比较，得出结论。

a. $z=1.82$　　b. $z=0.45$　　c. $z=1.50$　　d. $z=3.30$　　e. $z=-1.00$

应　用

13. 在 3 月 31 号之前，美国个人缴纳的联邦所得税平均数为 1 056 美元。以所缴纳的税款在其所得税纳税期的最后 5 天(4 月 10 日—4 月 15 日)才汇出的纳税人所组成的所谓“最后一分钟”总体。



a. 一名研究人员认为，在最后 5 天才缴纳税款的原因之一是纳税人所缴纳的税款比在此之前那些人所缴的税款要少。建立适当的假设，以便得出当拒绝 H_0 时支持该研究人员说法的结论。

b. 当 $\alpha=0.05$，得出检验统计量的临界值和拒绝法则？

c. 由 400 名在 4 月 10 日—4 月 15 日之间缴纳个人所得税组成的样本，样本平均值为 910 美元，样本标准差为 1 600 美元。计算检验统计量。

d. 得出你的结论。

e. 检验 p - 值为多少？

14. Nielsen 一份调查报告估计，每天每户家庭看电视的平均时间是 7.25 小时(*New York Daily News*, November 2, 1997)。假定 Nielsen 调查涉及了 200 户家庭，其样本标准差为每元 2.5 小时。根据报道 10 年前每户家庭看电视时间的总体均值为 6.70 小时。令 μ 为 1997 年每户家庭看电视的时间的总体均值，检验假设 H_0：$\mu\leqslant 6.70$ 和 H_a：$\mu>6.70$。取 $\alpha=0.01$。

a. 得出检验统计量的临界值和拒绝规则。

b. 计算检验统计量。

c. 得出你的结论。

15. 据美国国家汽车交易联合会报告，二手汽车的平均价格是 10 192 美元。根据堪萨斯市二手车交易市场的管理人员回顾，由最近市场销售的 100 辆二手车价格的样本均值为 9 300 美元，样本标准误差为 4 500 美元。令 μ 代表堪萨城市场的二手车的总体平均价格，试在显著性水平 $\alpha=0.05$ 下检验 H_0：$\mu\geqslant 10\,192$ 和 H_a：$\mu<10\,192$。

a. 得出检验统计量和拒绝规则。

b. 计算检验统计量。

c. 得出你的结论。

16. Media Metrix 有限公司跟踪调查七个国家的上网者，这七国包括澳大利亚、英国、加拿大、法国、德国、日本和美国。根据最近调查数据，关于上网时间美国家庭上网者排在第一位，平均每月 13 个小时(*The Washington Post*, August 4, 2000)。假定由 145 名加拿大上网者组成的样本调查表明，样本平均值为每月 10.8 小时，样本标准误差为 9.2 小时。

a. 建立原假设和备择假设，以便能决定是否样本数据支持加拿大上网者的总体平均时间小于美国的平均时间 13 小时的结论。

b. 当 $\alpha=0.01$，得出检验统计量的临界值，并写出拒绝规则。

c. 得出检验统计量的值。

d. 得出你的结论。

17. Fowle 市场调查公司认为，电话调查可在平均 15 分钟以内结束，并据此向客户收费。如果调查表明所需时间超过该值，则需要加额外费用。假设由 35 个电话调查所组成的一个样本表明，样本平均时间为 17 分钟，样本标准差为 4 分钟。问费用比率是否合理？

a. 建立该应用的原假设和备择假设。

b. 计算统计检验量的值。

c. p－值大小是多少？

d. 当 $\alpha=0.01$，得出你的结论。

18. 第一套广播网研究表明，十几岁的年轻人每去一次快餐店如麦当劳、Berger King 和 Wendy's 的平均消费为 5.72 美元（*USA Today*, October 5, 1998）。在接下的研究中，由芝加哥的 102 名经常去快餐店的十几岁的年轻人组成一个样本，这一样本的平均值为 5.98 美元，样本平均误差为 1.24 美元。

a. 建立原假设和备择假设，以便于决定样本数据是否支持芝加哥十几岁的年轻人每去一次快餐店的消费总体均值为 5.72 美元的结论。

b. 检验统计量的大小。

c. p－值为多少？

d. 当 $\alpha=0.05$ 时，得出你的结论。

19. 美国全国新的单身家庭的平均消费量为 181 900 美元（*The New York Times Almanac 2000*）。由南部 40户单身家庭的消费量组成一个样本，样本平均值为166 400 美元，样本平均标准误差为 33 500 美元。

a. 建立原假设和备择假设，以便能决定是否样本数据支持南部单身家庭的总体平均销售量少于全国平均值 181 900 美元的结论。

b. 检验统计量的大小。

c. p－值的大小。

d. 当 $\alpha=0.01$ 时，得出你的结论。

20. 根据美国大学和就业国家联合会数据表明，2000 年毕业的会计专业的毕业生平均年薪为 37 000 美元（*Time*, May 8, 2000）。在 2001 年 6 月对此继续研究，由 48 名会计专业的大学生组成一个样本，这一样本的平均值为 38 100 美元，样本的标准误差为 5 200 美元。

a. 建立原假设和备择假设，以便于决定样本数据是否支持 2001 年 6 月会计专业毕业的大学平均年薪比 2000 年毕业的平均年薪 37 000 美元要高。

b. 检验统计量的大小。

c. p－值的大小。

d. 当 $\alpha=0.05$ 时，得出你的结论。

9.4　总体均值的双侧检验：大样本

双侧假设检验(two-tailed tests)与单侧检验不同，双侧检验的拒绝域既在抽样样本的左侧也在右侧。下面通过一个例子来介绍如何进行双侧检验和为什么采用双侧检验。

美国高尔夫球联合会（USGA）规定，如果高尔夫球生产商想让他们的产品被 USGA 接受用于赛事，就必须达到规定要求。其中高尔夫球的生产标准之一为："在 USGA 总部所特许的室外器械上，接受检验的高尔夫球射程的平均距离应不超过 280 码……"假定 Superflight 公司最近采用一种高技术的生产方法所生产的高尔夫球射程的平均距离为 280 码。

然而，Superflight 公司意识到，如果新的生产过程失控的话，则生产出来的高尔夫球的平均距离或者小于 280 码，或者大于 280 码。在第一种情况下，由于产品质量降低，可能会导致销售额下降；而在第二种情况下，由于超过了 280 码这一距离规则，USGA 可能会拒绝使用这些高尔夫球。

Superflight 公司的管理人员设计出一个质量控制程序来监控生产过程，并且定期地使用假设检验来决定这一生产过程是否操作不正确。首先，我们假定这一生产过程运行正常，所生产出的高尔夫球平均射程满足规定的 280 码。原假设和备择假设如下：

$$H_0:\ \mu = 280$$
$$H_a:\ \mu \neq 280$$

如果样本数据表明 H_0 没被拒绝，则不需要调整生产过程。但是，如样本数据表明 H_0 被拒绝，我们将断定所生产出的高尔夫球平均射程没有达到规定要求 280 码。在这种情况下应采取措施调整生产过程使其达到满足状态。

质量控制小组选定假设检验的显著性水平 $\alpha = 0.05$，每次进行质量控制检验时选定 36 个高尔夫球作为样本。检验统计量为

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

假设检验量跟 9.3 节一样。

通常，我们假定原假设是正确的。当检验统计量 z 表明样本平均射程明显少于或明显大于 280 码时，我们将拒绝 H_0：$\mu = 280$。因此，当 z 值在抽样分布的左侧或右侧时，H_0 将被拒绝。所以，称此检验假设为**双侧**(two-tailed)检验。

图 9.7 表明了 $\alpha = 0.05$ 时，抽样分布 z 的双侧拒绝域。在双侧假设检验中，通常将分布的双侧区域或概率分别为 $\alpha/2$ 时所对应的区域确定为拒绝域。查找标准正态分布表可以得出一侧面积为 0.025 所对应的 z 值。见图 9.7，$-z_{0.025} = -1.96$ 的左侧面积为 0.025，$z_{0.025} = +1.96$ 的右侧面积为 0.025。因而，$z = -1.96$ 和 $z = 1.96$ 是检验的临界值，拒绝规则如下：

如果 $z<-1.96$ 或 $z>1.96$，则拒绝 H_0

表 9.2 提供了 36 个高尔夫球所组成的简单随机样本数据。从这些数据得出样本平均值$\bar{x}=278.5$，样本标准差 $s=12$ 码。原假设中 $\mu_0=280$，样本标准差 $s=12$，估计总体标准误差 σ，检验统计量值为

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{278.5-280}{12/\sqrt{36}}=-0.75$$

根据拒绝规则，H_0 不能被拒绝。样本数据表明没有理由怀疑该生产过程生产出的高尔夫球的平均射程不是 280 码。

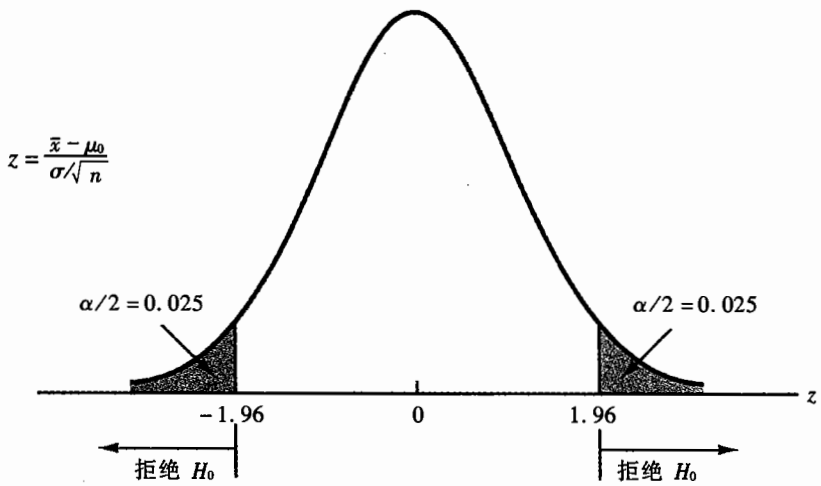


图 9.7 Superflight 公司双侧假设检验的拒绝域

表 9.2 36 个高尔夫球射程所组成的简单随机样本数据


CD 光盘数据
Distance

高尔夫球	码	高尔夫球	码	高尔夫球	码
1	269	13	296	25	272
2	300	14	265	26	285
3	268	15	271	27	293
4	278	16	279	28	281
5	282	17	284	29	269
6	263	18	260	30	299
7	301	19	275	31	263
8	295	20	282	32	264
9	288	21	260	33	273
10	278	22	266	34	291
11	276	23	270	35	274
12	286	24	293	36	277

双侧检验的 p – 值

我们计算一下 Superflight 高尔夫球例子中的 p – 值。样本平均值 $\bar{x}=278.5$ 相对应的 $z=-0.75$。查找标准正态分布表发现平均值与 $z=-0.75$ 之间的面积为 0.273 4。因此，左侧的面积为 0.500 0 – 0.273 4 = 0.226 6。图 9.7，拒绝域的左侧比例面积或概率为 $\alpha/2=0.05/2=0.025$。因而，0.226 6 > 0.025，检验统计量不在拒绝域内，原假设不能被拒绝。

在检验中，存在一个问题是：我们如何确定 p – 值？首先，你可能认为 p – 值为 0.226 6。如果这是你的答案，那么你必须记住两个不同的规则：一个是单侧检验的规则，如果 p – 值 $<\alpha$，则拒绝 H_0，另一个是双侧检验的规则，如果 p – 值 $<\alpha/2$，则拒绝 H_0。另外，我们假定双侧检验的 p – 值为分布一侧的**双倍**。这样在 Superflight 例子中，p – 值定义为 2（0.226 6）= 0.453 2。在双侧检验中，以这种方式定义的最大优点是 p – 值可以直接与显著性水平 α 相比较。因为 0.453 2>0.05，所以原假设不能被拒绝。双侧检验中 p – 值是分布一侧的面积的两倍。前面所定义的拒绝规则：如果 p – 值 $<\alpha$，则拒绝 H_0，适用于所有的假设检验。

在假设检验中，p – 值是单侧的两倍。将面积的双倍作为 p – 值可以直接与 α 比较。因此，无论是双侧检验还是单侧检验都使用相同的拒绝规则。

可以利用 Minitab 统计软件来求得总体均值的假设检验的信息。图 9.8 给出了 Superflight 进行假设检验的计算机的输出结果。计算结果如下：样本平均值 $\bar{x}=278.5$，统计检验量 $z=-0.75$，p – 值 = 0.453。获得 Minitab 输出的结果的程序将在附录 9.1 给出。Minitab 的输出结果给出了总体平均值的 95% 置信区间。

```
Test of mu = 280.00 vs mu not = 280.00
Variable    N      Mean    StDev    SE Mean       Z        P
Yards      36    278.50    12.00       2.00   -0.75    0.453
```

图 9.8　Superflight 高尔夫球的假设检验中 Minitab 的输出结果

总体均值的双侧检验步骤总结

令 μ_0 代表假设的平均值，则总体均值双侧假设检验的一般形式如下：

大样本（$n\geq30$）情形总体均值的双侧检验的形式

$$H_0:\ \mu=\mu_0$$
$$H_a:\ \mu\neq\mu_0$$

检验统计量 σ 已知

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

检验统计量以 s 代替 σ

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{9.2}$$

拒绝规则

使用检验统计量：如果 $z<-z_{\alpha/2}$ 或 $z>z_{\alpha/2}$，则拒绝 H_0

使用 p - 值：如果 p - 值 $<\alpha$，则拒绝 H_0

区间估计与假设检验的关系

在第 8 章中，我们介绍了如何对总体平均值进行置信区间估计。在大样本情形下，当置信系数为 $1-\alpha$，总体平均值的置信区间估计为

当 σ 已知时 $$\bar{x}\pm z_{\alpha/2}\frac{\alpha}{\sqrt{n}} \tag{9.3}$$

当 σ 未知时，以 s 代替 σ $$\bar{x}\pm z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{9.4}$$

进行假设检验时，我们首先对总体参数进行假设。就总体均值，双侧假设检验的形式如下：

$$H_0:\ \mu=\mu_0$$
$$H_a:\ \mu\neq\mu_0$$

式中，μ_0 为假设检验的总体平均值。利用双侧检验的拒绝规则，可以看到当样本平均值 $\bar{x}$ 落在 μ_0 附近 $-z_{\alpha/2}$ 和 $z_{\alpha/2}$ 的标准差内，不能拒绝 H_0。从而，在双侧检验中，给定显著性水平 α，样本平均值 $\bar{x}$ 的不能被拒绝域如下：

当 σ 已知时 $$\mu_0\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{9.5}$$

当 σ 未知时，以 s 代替 σ $$\mu_0\pm z_{\alpha/2}\frac{s}{\sqrt{n}} \tag{9.6}$$

观察公式(9.3)和(9.5)，可以发现统计推断的估计和假设检验的关系。特别注意的是，这两个程序都需要计算 $z_{\alpha/2}$ 和 $\frac{\sigma}{\sqrt{n}}$。假设检验中，对于给定的显著性水平 α，对应的区间估计的置信度系数为 $(1-\alpha)$。例如，假设检验的显著性水平 $\alpha=0.05$，则区间的置信度系数为 95%。而且，在公式(9.3)和(9.5)中，$z_{\alpha/2}$ $(\frac{\sigma}{\sqrt{n}})$ 是被加减的值，如果 $\bar{x}$ 落在公式 (9.5) 所定义的不能

被拒绝域中，那假设检验值就落在公式(9.3)所定义的置信区间中。相反，如果假设值 μ_0 落在公式(9.3)所定义的置信区间中，样本平均值 $\bar{x}$ 就落在公式(9.5)所定义的不能被拒绝假设 H_0：$\mu = \mu_0$ 的区域中。从这些观察中我们可以得出利用置信区间进行假设检验的程序。

使用置信区间方法进行假设检验的形式

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

1. 从总体中随机抽取一个简单样本，利用样本均值 $\bar{x}$ 求总体均值 μ 的置信区间。如果 σ 已知，计算估计区间

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

如果 s 代替 σ，计算估计区间

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

2. 如果置信区间包含假设值 μ_0，则不能拒绝 H_0。否则，拒绝 H_0。

对于双侧假设检验，如果置信区间没包括 μ_0，则原假设被拒绝。

我们仍回到 Superflight 高尔夫球的研究中，其双侧检验如下：

$$H_0: \mu = 280$$
$$H_a: \mu \neq 280$$

当显著性水平 $\alpha = 0.05$ 时，进行假设检验，我们抽取 36 个高尔夫球作为样本，样本平均射程为 $\bar{x} = 278.5$ 码，样本标准误差 $s = 12$ 码。其中 $z_{0.025} = 1.96$，总体均值的 95% 的置信区间为

$$\bar{x} \pm z_{0.025}\frac{s}{\sqrt{n}}$$
$$278.5 \pm 1.96\frac{12}{\sqrt{36}}$$
$$278.5 \pm 3.92$$

或

$$274.58 \text{ 到 } 282.42$$

这一结果使得质量控制管理人员可以以 95% 的置信水平来确定高尔夫球总体的平均射程在 274.58 到 282.42 之间。因为总体均值的假设值 $\mu_0 = 280$ 落在该区间内，所以假设检验的结论为不能拒绝原假设 H_0：$\mu = 280$。

上面所讨论的例子适合于总体均值的双侧检验。然而，对于其他总体参数，置信区间与假设

检验之间存在相同的关系。另外，对于单侧检验的其他总体参数，也同样适合。只不过将进行的区间估计设为相应的单侧置信区间。

评 注

1. p－值仅仅依赖于样本结果。但是，要了解所调查的假设检验是单侧还是双侧。给定样本中的$\bar{x}$，双侧检验的 p－值总是$\bar{x}$抽样分布的一侧面积的双倍。
2. 假设检验的区间估计方法强调样本大小的作用。从公式(9.3)中可以发现 n 值越大，则置信区间越小。对于给定的显著性水平 α，原假设为假时，样本容量越大则置信区间不易包含 μ_0。也就是说，当 H_0 为假时，样本容量越大，拒绝 H_0 的概率越大。

练 习

方 法

21. 考虑下面的假设检验：

$$H_0: \mu = 10$$
$$H_a: \mu \neq 10$$

一个样本由 36 项组成，样本平均值为 11，样本标准差为 2.5。

a. 当 $\alpha = 0.05$ 时，得出拒绝规则。

b. 计算统计检验量 z，得出你的结论。

22. 考虑下面的假设检验：



$$H_0: \mu = 15$$
$$H_a: \mu \neq 15$$

一个样本由 50 项组成，样本平均值为 14.2，样本标准差为 5。

a. 当 $\alpha = 0.02$ 时，得出拒绝规则。

b. 计算统计检验量 z。

c. p－值的大小是多少？

d. 得出你的结论。

23. 考虑下面的假设检验：

$$H_0: \mu = 25$$
$$H_a: \mu \neq 25$$

一个样本由 80 项组成，样本标准差为 10。当 $\alpha=0.05$，对于如下的样本结果，计算统计检验量 z，并描述你的结论。

a. $\bar{x}=22.0$　　b. $\bar{x}=27.0$　　c. $\bar{x}=23.5$　　d. $\bar{x}=28.0$

24. 考虑下面的假设检验：

$$H_0:\ \mu=5$$
$$H_a:\ \mu\neq5$$

假定假设检验统计量如下。计算相应的 p - 值，当 $\alpha=0.05$ 时，描述你的结论。

a. $z=1.80$　　b. $z=-0.45$　　c. $z=2.05$　　d. $z=-3.50$　　e. $z=-1.00$

应　用

25. 根据报告工人一周平均工作时间为 39.2 小时（*Investor's Business Daily*，September 11, 2000）。当 $\alpha=0.05$，对下述假设 H_0：$\mu=39.2$ 和 H_a：$\mu\neq39.2$ 进行检验。



a. 检验统计量的临界值和关于 H_0 的拒绝规则。

b. 假定样本由 112 名工人组成，样本的平均值为 38.5 小时，样本标准差为 4.8 小时。得出你的检验统计量值。

c. 原假设是否被拒绝？得出你的结论。

d. p - 值的大小是多少？

26. CNN 和 ActMedia 电视频道可用于播放新闻、广告和特别报道，以百货商店等待结账时排队的个人为播出对象。假定百货商店购物者排队等待结账时间的总体均值为 8 分钟，所以电视节目每 8 分钟循环放映一次（*Astounding Averages*，1995）。选定在某商店购物 120 名组成一个样本，其样本平均等待时间为 7.5 分钟，样本标准差为 3.2 分钟。当 $\alpha=0.05$ 时，对假设 H_0：$\mu=8$ 和 H_a：$\mu\neq8$ 检验。

a. 统计检验量的临界值的大小和拒绝规则分别是什么？

b. 检验统计量的大小是多少？

c. 得出你的结论。

27. 一条生产线生产的产品的平均充满重量为 16 盎司。超重或缺少会发生严重问题，导致生产线停产。从过去的数据中发现，假定 σ 为 0.8 盎司。质量监管人员每 2 个小时抽取 30 项样本进行检验，并决定是否需要停产作调整。

a. 当显著性水平 $\alpha=0.05$ 时，假设检验程序中的拒绝域规则是什么？

b. 如果样本平均值 $\bar{x}=16.32$ 盎司，那么你建议采取什么步骤？

c. 如果 $\bar{x}=15.82$ 盎司，那么你建议采取什么步骤？

d. 求(b)和(c)中的 p - 值。

28. Gallup 调查表明作为联邦税收返回的平均慈善捐赠为 1 075 美元(*USA Today*, April 10, 2000)。假定由 2001 年 4 月税收返回组成的样本进行假设检验，以便于决定是否有任何关于平均慈善捐赠额的变动。
    a. 建立原假设和备择假设。
    b. 假定由 200 项税收返回组成的一个样本，样本的平均值为 1 160 美元，样本标准差为 840 美元。统计检验量为多少?
    c. p - 值的大小为多少?
    d. 当 $\alpha=0.05$ 时，得出你的结论。

29. 历史数据表明，某一城市的晚上的长途电话的平均时间为 15.2 分钟。有由 35 个电话组成的一个随机样本，样本中每个电话平均时间为 14.3 分钟，样本标准差为 5 分钟。利用样本信息来检验是否长途电话的平均时间发生改变。
    a. 计算检验统计量。
    b. p - 值的大小是多少?
    c. 当 $\alpha=0.05$ 时，得出你的结论。

30. Florida 的劳动力和就业安全保障机构报道平均年薪为 26 133 美元(*The Naples Daily News*, February 13, 1999)。对全国的平均薪水检验看是否与一个州的平均薪水一样。
    a. 建立原假设和备择假设，以便于决定某一国家的样本数据支持这个国家的平均年薪与一个州的平均年薪 26 133 美元一致。
    b. 由 Collier 乡村的 550 名居民组成一个样本，样本的平均年薪为 25 457 美元，样本标准差为 7 600 美元。计算统计检验量。
    c. p - 值的大小为多少?
    d. 当 $\alpha=0.05$ 时，得出你的结论。

31. 历史数据表明，西部大学生入学文化考试成绩是 900，标准差为 180。每年对这些数据进行检验，看该年学生入学成绩与上年是否一样。检验中原假设 H_0: $\mu=900$。本年由 200 名学生组成的样本表明样本分数的平均值为 935。当显著性水平 $\alpha=0.05$。
    a. 用置信度的方法进行假设检验。
    b. 用统计检验进行假设检验。
    c. p - 值的大小是多少?

32. 某企业的每小时平均薪水为每小时 15.00 美元。从企业中选取 36 名工人组成样本，样本的平均薪水 $\bar{x}=14.50$ 美元，样本标准差 $s=0.60$ 美元。
    a. 单侧置信区间利用样本结果建立总体参数的上限或下限。本题中，建立企业所支付的每小时薪水的左侧 95% 置信限度。这种单侧置信区间的形式要求我们 95% 地确信总体均值小于等于该区间。对于此单侧置信区间，解释 95% 的置信的含义?
    b. 利用单侧置信区间结果来检验 H_0: $\mu\geqslant 15$。得出你的结论。

9.5　总体均值的检验：小样本

在 9.3 节和 9.4 节中总体平均的假设检验的程序是基于中心极限定理和大样本理论。在 9.5 节，我们考虑小样本情形下($n<30$)的总体均值检验。在这种情形下，$\bar{x}$ 的抽样分布主要依赖于总体的分布。事实上，小样本情形下的假设检验程序要求假定总体是一个标准正态分布。如果这个假设恰当，将使用本节中介绍的方法。但是，如果假设不成立，最好的替代方法是增加样本大小到 $n\geqslant 30$，依赖在 9.3 节和 9.4 节中介绍的大样本假设检验程序。

本节所提供的方法是基于总体是正态分布。

首先我们考虑样本容量为小样本($n<30$)时的情形，假设总体是正态分布，并根据历史数据、理论或其他原因，可得知总体标准差 σ。在这些条件下，对于任何大小的样本，$\bar{x}$ 的抽样分布以平均值 μ 为正态分布，标准差 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$。因为 $\bar{x}$ 的抽样分布是正态分布，小样本假设检验程序等同于 9.3 节和 9.4 节所介绍的大样本假设检验程序。检验统计量如下：

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \tag{9.7}$$

因为假设检验的计算与前面我们所计算的一样，所以我们不再需要举新的数字例子。

接下来我们考虑当样本容量为小样本($n<30$)时，假定总体是正态分布，用样本标准差 s 作为总体标准差 σ 的估计值。在这种情形下，利用 t 分布来对总体平均值作推断。利用假设检验的 t 分布，检验统计量为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{9.8}$$

检验统计量服从自由度为 $n-1$ 的 t 分布。

在小样本假设检验程序中，以 s 代替 σ，是基于 8.2 节中用于小样本区间估计理论。两者都使用总体正态分布假设和 t 分布。



CD 光盘数据
Heathrow

我们考虑一个小样本情形中关于总体均值单侧检验的例子。国际航空运输协会通过对商务旅行者调查，来评定跨太平洋通道的机场等级。最大可能的级别为 10。一本商务旅行方面的杂志决定按照他们收集的级别数据对各个机场分类。如果机场级别的总体均值超过 7，则被认为是高质量服务的机场。假设抽取 12 名商务旅行者组成一个简单随机样本，要求每个旅行者对伦敦的 Heathrow 机场评定级别，12 名旅行者的评定级别分别为 7，8，10，8，6，9，6，7，7，8，9 和 8。这些数据组成的样本均值 $\bar{x}=7.75$，样本标准差为 $s=1.215$。假定评定的总体级别近似服从正态概率分布，那么 Heathrow 能否被认为是高

质量服务的机场?

当显著性水平为 0.05 时，我们通过检验决定是否 Heathrow 机场的总体平均级别大于 7。原假设和备择假设如下：

$$H_0: \mu \leqslant 7$$
$$H_a: \mu > 7$$

如果 H_0 被拒绝，则 Heathrow 机场将被认为是高质量服务的机场。

拒绝域在抽样分布的右侧。自由度为 $n-1=12-1=11$，由前面的 t 分布表可查得临界值 $t_{0.05}=1.796$。因而，拒绝规则为

如果 $t>1.796$，则拒绝 H_0

当 $\bar{x}=7.75$ 和 $s=1.215$，利用公式(9.8)，得出检验统计量为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{7.75-7}{1.215/\sqrt{12}}=2.14$$

因为 2.14 远大于 1.796，原假设被拒绝。在显著性水平为 0.05 下，我们假定 Heathrow 机场的总体平均级别远大于 7。因此，Heathrow 可能认为是高质量服务机场，图 9.9 表示检验统计量落在拒绝域内。

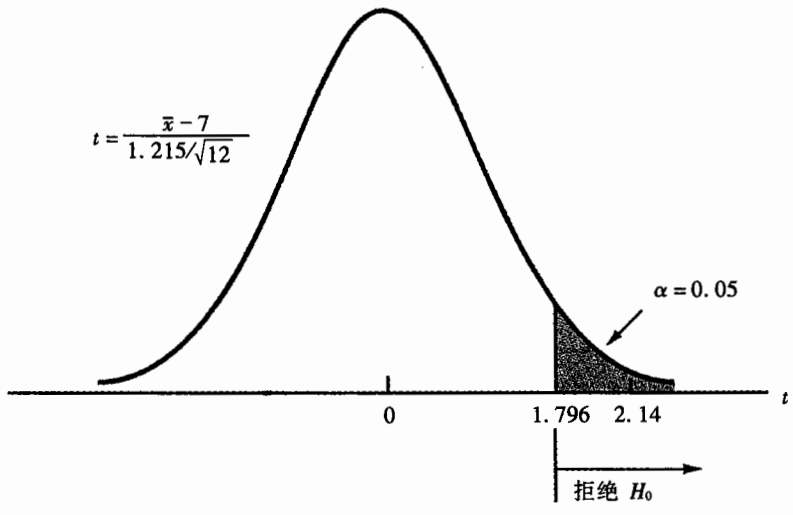


图 9.9 Heathrow 机场假设检验中检验统计量的值($t=2.14$)

p - 值和 t 分布

我们考虑用 p - 值对 Heathrow 机场进行假设检验。在 9.3 节和 9.4 节中，p - 值解释为检验

的实测显著性水平。一般所运用的规则为：如果 p – 值小于显著性水平 α，原假设将被拒绝。然而，在大多数统计教科书中所提供的 t 分布表不能满足检验中确定准确的 p – 值所要求的足够信息。但是，我们可利用 t 分布表来确定 p – 值分布范围。例如，Heathrow 机场假设检验的 t 分布自由度为 11。查 t 分布表，我们看到第 11 行给出了自由度为 11 的 t 分布的信息：

右侧面积	0.10	0.05	0.025	0.01	0.005
t 值	1.363	1.796	2.201	2.718	3.106

进行假设检验计算 $t = 2.14$，p – 值是 $t = 2.14$ 那一侧的面积。从上面表的信息可知，2.14 在 1.796 到 2.201 之间。虽然我们不能得到精确的 p – 值，但是我们知道 p – 值在 0.05 到 0.025 之间。当显著性 $\alpha = 0.05$ 时，我们得知 p – 值一定小于 0.05，因而原假设被拒绝。

---

p – 值很难从 t 分布表中直接获得。一般而言，最好的方法是像 0.05 和 0.025 一样指定其 p – 值区间。

我们可以利用 Minitab 统计软件来求小样本情形下总体均值检验的信息。图 9.10 表示 Heathrow 机场进行假设检验的计算机输出结果。计算机自动得出样本平均值 $\bar{x} = 7.75$，统计检验量 $t = 2.14$，p – 值 $= 0.028$。在附录 9.1 中详细描述了如何获得 Minitab 输出结果的程序。

```
Test of mu = 7.000 vs mu > 7.000
Variable     N     Mean    StDev   SE Mean      T      P
Rating      12    7.750    1.215     0.351   2.14  0.028
```

图 9.10　Heathrow 机场评定等级的假设检验中 Minitab 的输出结果

双侧假设检验

作为小样本情形下关于总体均值的双侧假设检验的例子，考虑下面的生产问题。设计一个生产过程来填充容器，容器的填充重量的平均值 $\mu = 16$ 盎司。消费者不愿意消费那些产品的重量少于产品标签下所标明的重量，所以该生产过程填充的重量不能少于此值。同样，如果生产过程所填充的容器重量超过要求，那么企业将会亏损，所以企业不希望该生产过程中填充的重量高于此值。质量保证人员定期抽取 8 只容器组成简单随机样本，进行如下的双侧假设检验。

$$H_0: \mu = 16$$

$$H_a: \mu \neq 16$$

如果被 H_0 拒绝，则生产管理者将要求停止生产，对负责调节填充重量的调节装置重新调整，确保容器的填充重量均值为 16 盎司。如果样本产生的数据为 16.02，16.22，15.82，15.92，16.22，16.32，16.12 和 15.92 盎司，假定填充重量总体分布服从正态分布，取显著性水平为 0.05。

由于数据没有整理，所以我们首先计算样本平均值和样本标准差，得到以下结果：

$$\bar{x}=\frac{\sum x_i}{n}=\frac{128.56}{8}=16.07 \text{ 盎司}$$

以及

$$s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{0.22}{7}}=0.177 \text{ 盎司}$$

在显著性水平 $\alpha=0.05$ 下的双侧检验中，$-t_{0.025}$ 和 $t_{0.025}$ 是检验的临界值。利用 t 分布，得出自由度 $n-1=8-1=7$，$-t_{0.025}=-2.365$ 和 $t_{0.025}=+2.365$。因而，拒绝规则为

如果 $t<-2.365$ 或 $t>2.365$，则拒绝 H_0

利用$\bar{x}=16.07$ 和 $s=0.18$，可得

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{16.07-16.00}{0.177/\sqrt{8}}=1.12$$

因为 $t=1.12$ 不在拒绝域中，原假设不能被拒绝，没有足够的证据来证明需要停止生产。

利用 t 分布表中的自由度为 7 的那一行，我们看到所计算的 t 值为 1.12，其右侧的面积**大于** 0.10。虽然 t 分布表不能得到更精确的值，但是我们至少可以得到双侧检验的 p - 值大于 $2(0.10)=0.20$。因为该值大于显著性水平 0.05，根据 p - 值我们能得到相同的结论，也就是不能拒绝 H_0。对于该假设检验通过计算机得出其精确 p - 值为 0.301。

练　习

方　法

33. 考虑以下的假设检验：

$$H_0:\ \mu\leqslant 10$$
$$H_a:\ \mu>10$$

一个样本是由 16 项组成，样本平均值为 11，样本标准差为 3。

a. 当 $\alpha=0.05$ 时，得出拒绝规则。

b. 计算检验统计量 t，得出你的结论。

34. 考虑下面的假设检验：



$$H_0:\ \mu=20$$
$$H_a:\ \mu\neq 20$$

一个样本的 6 项数据为：18，20，16，19，17，18。

a. 计算样本平均值。

b. 计算样本标准差。

c. 当 $\alpha=0.05$ 时，得出拒绝规则。

d. 计算统计检验量 t。

e. 得出你的结论。

35. 考虑下面的假设检验：

$$H_0:\ \mu\geqslant 15$$
$$H_a:\ \mu<15$$

样本由 22 项组成，其样本标准差 $s=8$。当 $\alpha=0.05$ 时，计算检验统计量 t 值，并得出你的结论。

a. $\bar{x}=13.0$　　b. $\bar{x}=11.5$　　c. $\bar{x}=15.0$　　d. $\bar{x}=19.0$

36. 考虑下面假设检验：

$$H_0:\ \mu\leqslant 50$$
$$H_a:\ \mu>50$$

假设一个样本由 16 项组成，当检验统计量取如下值时，计算 p - 值，并当 $\alpha=0.05$ 时，得出你的结论。

a. $t=2.602$　　b. $t=1.341$　　c. $t=1.960$　　d. $t=1.055$　　e. $t=3.261$

应　用

37. 包括美国运通、E*Trade Group、高盛和美林在内的金融服务公司每股的总体平均收益为 3 美元（*Business Week*，August 14, 2000）。在 2001 年，由 10 项金融服务公司组成的样本，其样本平均每股收益如下：

自测题

1.92　2.16　3.63　3.16　4.02　3.14　2.20　2.34　3.05　2.38

a. 建立原假设和备择假设，以便决定 2001 年总体每股平均收益与 2000 年每股 3 美元的收益不一样。

b. 当 $\alpha=0.05$ 时，假设检验的临界值是多少？拒绝规则是什么？

c. 计算样本平均值。

d. 计算样本标准差。

e. 计算检验统计量。

f. 得出你的结论。

g. p - 值是多少？

38. 在美国，平均每户家庭的消费为 90 美元（*American Demographics*，August 1997）。假定从纽约

Corning 选取 25 户家庭所组成一个样本，样本平均每天消费为 84.50 美元，样本标准差为 14.50 美元。当 $\alpha=0.05$ 时，对假设 H_0：$\mu=90$ 和 H_a：$\mu\neq90$ 检验。

a. 检验统计量的临界值大小，拒绝规则是什么？

b. 计算检验统计量。

c. 得出你的结论。

d. p－值是多少？

39. 平均而言，一个有丈夫和两个孩子的家庭主妇每周做家务活动的时间大概不超过 55 个小时。由 8 名家庭主妇所组成的一个样本一周做家务的时间是：58，52，64，63，59，62，62 和 55。当 $\alpha=0.05$ 时，检验假设 H_0：$\mu\leqslant55$ 和 H_a：$\alpha=0.05$ 时，$\mu>55$。

a. 检验的临界值是多少？拒绝规则是什么？

b. 计算样本平均值。

c. 计算样本标准差。

d. 计算统计检验量。

e. 得出你的结论。

f. 计算 p－值的大小。

40. 从美国钻石开采委员会了解到 VS2 清晰度、H 色彩的一克拉的钻石的成本是 4 000 美元（*www. diasource. com*，July 2000）。一名中西部珠宝制造商与纽约城珠宝地区珠宝商联系想了解相似的珠宝的平均价格是否是 4 000 美元。珠宝商从 14 名纽约城中珠宝商中收集关于珠宝成本的数据。

a. 建立原假设和备择假设，以便决定是否纽约城珠宝价格的平均价格不是 4 000 美元。

b. 当 $\alpha=0.05$，检验的临界值，得出拒绝规则。

c. 假定 14 名珠宝商组成的样本平均价格为 4 120 美元，样本标准差为 275 美元。计算检验统计量。

d. 得出你的结论。

e. 计算 p－值的大小。

41. Callaway Golf 公司生产的新钛合金 ERC 击球器械被认为不“合法”，因为 ERC 击球的距离超过了 USGA 的标准。Golf Digest 中将 ERC 击球器械和 USGA 特许击球器械的距离相比较，USGA 特许击球的距离的总体平均击球距离为 280 码。通过 9 次击球检验，ERC 击球器械平均击球距离为 286.9 码（*Golf World*, May 12, 2000）。假定样本击球距离的标准差为 10 码，问答下面问题。

a. 建立原假设和备择假设，以便于决定是否新的 ERC 击球器械击球的总体平均距离远大于 280 码。

b. 就平均而言，ERC*器械高尔夫球击出的距离是多远？

c. 当 $\alpha=0.05$ 时，检验的临界值是多少？拒绝规则是什么？

d. 计算检验统计量。

*编者注：疑原文有误。

e. 得出你的结论。

f. p – 值的大小为多少？

42. Joan 的苗圃公司是专门为居民区提供庭院布置方面的设计服务的。某一庭院布置计划的估计人力成本是基于计划所种植树木、灌溉等的数量计算出来的。为了估计成本的大小，管理者在 2 小时的人工时间内来种植中型树木。上个月 10 次种植时实际所用的时间组成一个样本（以小时为单位）。

1.9　1.7　2.8　2.4　2.6　2.5　2.8　3.2　1.6　2.5

当显著性水平为 0.05 时，检验平均种植时间是否超过 2 小时。

a. 建立原假设和备择假设。

b. 检验的临界值大小，拒绝规则？

c. 计算样本平均值。

d. 计算样本标准差。

e. 计算检验统计量。

f. 得出你的结论。

g. 计算 p – 值的大小。

9.6　总体比例的检验

令符号 p 代表总体比例，p_0 代表总体比例的某一特定的假设值，则关于总体比例的假设检验三种形式如下：

$$H_0:\ p \geq p_0 \qquad H_0:\ p \leq p_0 \qquad H_0:\ p = p_0$$
$$H_a:\ p < p_0 \qquad H_a:\ p > p_0 \qquad H_a:\ p \neq p_0$$

前面两种形式为单侧检验，第三种形式为双侧检验。具体使用哪种形式要看应用情况。

总体比例的假设检验是基于样本比例 $\bar{p}$ 和假设总体比例 p_0 之间的差进行的。检验使用的方法与对总体均值进行假设检验所使用的程序相似，惟一不同的是对总体比例检验时使用的是样本比例 $\bar{p}$ 和标准差 $\sigma_{\bar{p}}$。我们首先对总体比例的值建立原假设和备择假设。然后，利用样本比例 $\bar{p}$ 和标准差 $\sigma_{\bar{p}}$，计算检验统计量 z 值。将所得的检验统计量与临界值相比较或将 p – 值与 α 比较，从而能帮助我们决定原假设是否被拒绝。

检验统计量或者 p – 值都能被用来得出假设检验结论。

现在通过 Pine Creek 高尔夫球场所面临的问题来解释如何对总体比例进行假设检验。在过去的几个月里，在 Pine Creek 运动者中有 20% 的是女性。为了提高在 Pine Creek 打网球的女性的比例，决定以特殊推销的方式来吸引女性高尔夫球运动者。一周以后，随机抽取 400 名高尔夫球运动者组成的样本，这样本中有 300 名为男性，100 人为女性。球场管理人员想知道样本数据是否

能得出 Pine Creek 高尔夫球场女性运动者所占的比例上升了。

为了确定这种特殊推销的方法是否提高了女性高尔夫球运动者的比例，我们建立了下面的原假设和备择假设：

$$H_0: \ p \leqslant 0.20$$
$$H_a: \ p > 0.20$$

通常我们假定 H_0：$p = 0.20$ 为真，从而进行假设检验。接下来，利用样本比例 $\bar{p}$ 来估计 p，并考虑 $\bar{p}$ 的抽样分布。因为 $\bar{p}$ 是 p 的无偏估计，所以如果 $p = 0.20$，则 $\bar{p}$ 的抽样分布的平均值为 0.20。另外，从第 7 章我们了解到 $\bar{p}$ 的标准差为

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

假定 $p = 0.20$，样本大小 $n = 400$，$\bar{p}$ 的样本标准差为

$$\sigma_{\bar{p}} = \sqrt{\frac{0.20(1-0.20)}{400}} = 0.02$$

在第 7 章中，我们得出如果 np 和 $n(1-p)$都大于 5，则 $\bar{p}$ 的抽样分布近似为标准正态概率分布。在 Pine Creek 例子中，$np = 400(0.20) = 80$ 和 $n(1-p) = 400(0.80) = 320$，因此可以用正态概率分布来近似。图 9.11 表示出了 $\bar{p}$ 的抽样分布情况。

因为该抽样分布是近似正态，所以可以使用如下的检验统计量：

总体比例的假设检验中的检验统计量

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \tag{9.9}$$

式中

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} \tag{9.10}$$

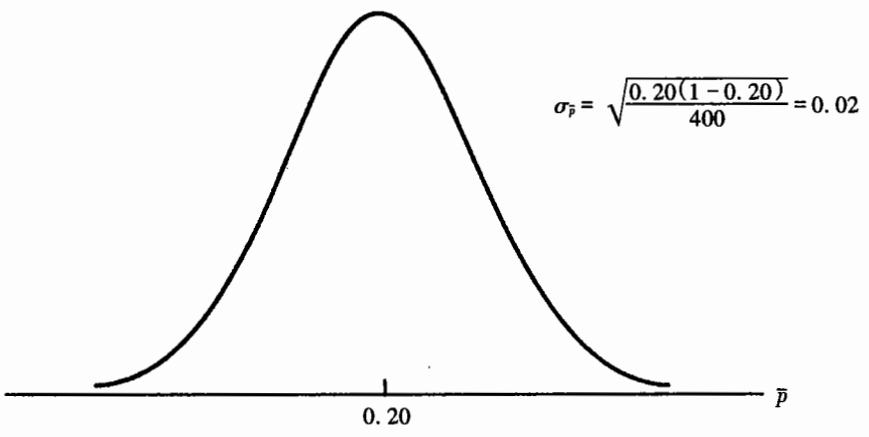


图 9.11　Pine Creek 女性高尔夫球员所占的比例的抽样分布

假设检验中，在 H_0 中假定总体比例 p_0。所以，当 H_0 为真时，在公式(9.10)中利用 p_0 可以计算 $\sigma_{\bar{p}}$。在计算 $\sigma_{\bar{p}}$ 时没有使用样本比例 $\bar{p}$。

假定检验的显著性水平 $\alpha = 0.05$。临界值 $z_{0.05} = 1.645$，假设检验的右侧拒绝区域给出了下面的拒绝规则（见图 9.12）：

如果 $z > 1.645$，则拒绝 H_0

拒绝域一经确定，我们通过收集数据，计算点估计 $\bar{p}$ 值及其相应的检验统计量 z。通过比较 z 值和临界值（$z_{0.05} = 1.645$），我们决定是否能拒绝原假设。

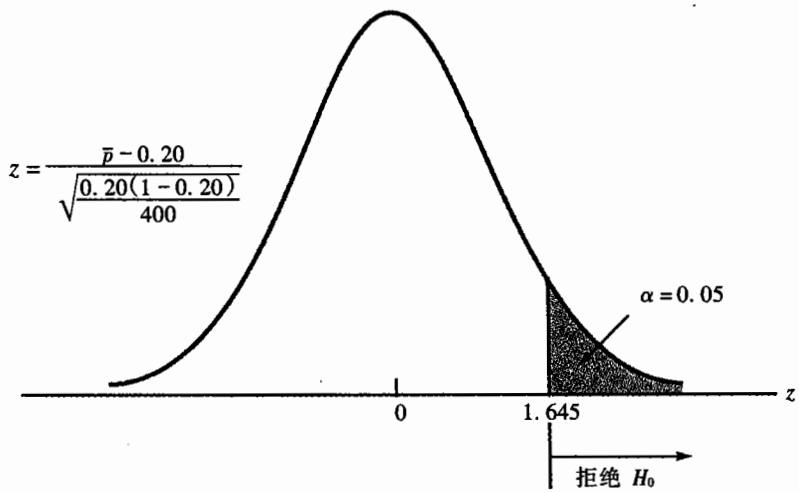


图 9.12　Pine Creek 高尔夫球场假设检验中的拒绝域

在推销期间由 400 名运动员组成的一个随机样本表明 100 名运动员是女性。因而，$\bar{p} = 100/400 = 0.25$，当 $\sigma_{\bar{p}} = 0.02$，检验统计量值为

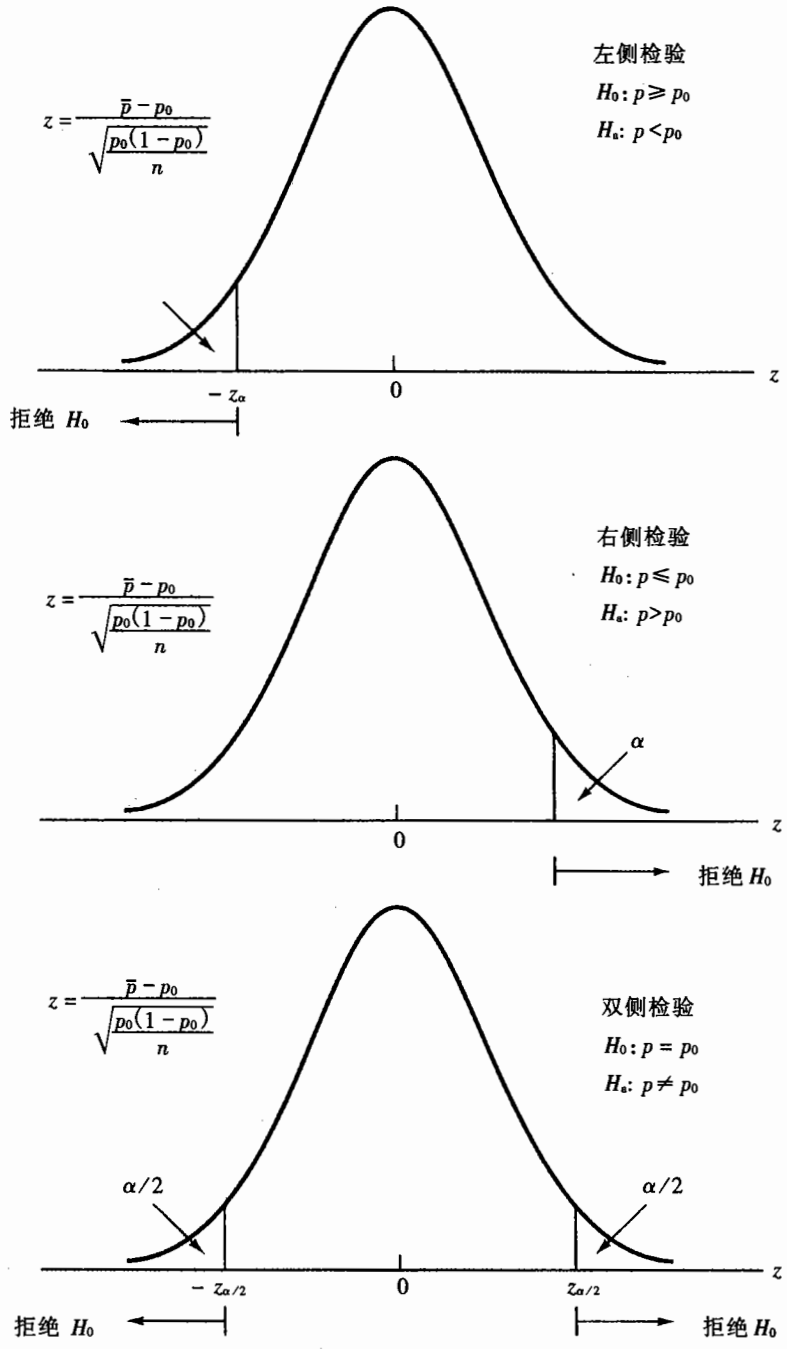
$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{0.25 - 0.20}{0.02} = 2.5$$

因为 $z = 2.5 > 1.645$，我们能拒绝 H_0。Pine Creek 管理人员得出在推销期间女性运动员的比例提高。

查标准正态概率分布的分布表，我们可以得到计算检验中的 p - 值。例如，对于检验统计量 $z = 2.50$，分布表表示介于平均值和 $z = 2.50$ 之间的面积或概率为 0.493 8。因而，检验的 p - 值为 0.500 0 - 0.493 8 = 0.006 2。当 p - 值小于 $\alpha = 0.051$，原假设被拒绝。

我们发现总体比例的假设检验与总体均值的假设检验相似。两者主要的区别在于当进行总体均值的假设检验时利用 $\bar{x}$ 的抽样分布构造检验统计量，而进行总体比例假设检验时利用 $\bar{p}$ 的抽样

分布来构造检验统计量。通过比较检验统计量和临界值或比较 p - 值和 α 来得出假设检验结论。图 9.13 总结了总体样本比例假设检验的拒绝规则。我们假定大样本情形下[$np \geqslant 5$ 和 $n(1-p) \geqslant 5$]，$\bar{p}$ 的抽样分布近似于正态概率分布。



注：在任何情况下，如果 p - 值<α，则拒绝 H_0。

图 9.13　关于总体样本比例假设检验的拒绝规则总结

评　注

我们没有给出小样本情形下总体比例的假设检验程序。在小样本情形下，$\bar{p}$ 的抽样分布服从二项分布，因此不能用正态概率分布。一些讲得深的课本将会对这种情况下如何进行假设检验进行讲述。但在实践中，很少对总体比例进行小样本检验。

练　习

方　法

43. 考虑下面的假设检验：

$$H_0:\ p \leqslant 0.50$$
$$H_a:\ p > 0.50$$

一个样本由 200 项组成，其样本比例为 $\bar{p} = 0.57$。

a. 当 $\alpha = 0.05$ 时，得出你的拒绝域。

b. 计算检验统计量 z。得出你的结论。

44. 考虑下面的假设检验：



$$H_0:\ p = 0.20$$
$$H_a:\ p \neq 0.20$$

一个样本由 400 项组成，其样本比例为 $\bar{p} = 0.175$。

a. 当 $\alpha = 0.05$ 时，得出你的拒绝域。

b. 计算检验统计量 z。

c. 计算 p - 值的大小。

d. 得出你的结论。

45. 考虑下面的假设检验：

$$H_0:\ p \geqslant 0.75$$
$$H_a:\ p < 0.75$$

一个样本由 300 项组成。当 $\alpha = 0.05$ 时，在下列情况下得出检验统计量 z、p - 值，并给出你的结论。

a. $\bar{p} = 0.68$　　b. $\bar{p} = 0.72$　　c. $\bar{p} = 0.70$　　d. $\bar{p} = 0.77$

应 用

46. Heldrich 工作效率发展研究中心调查发现 40% 的网络使用者每天收到超过 10 封 e-mail(*USA Today*, May 7, 2000)。在 2001 年，对 e-mail 的使用进行了相似的研究。研究的目的是看是否 e-mail 的使用增加了。
    a. 建立原假设和备择假设，以便决定网络使用者每天收到的 e-mail 超过 10 封的比例是否增加？
    b. 当 $\alpha=0.05$ 时，检验的临界值是多少？拒绝域是什么？
    c. 如果由 420 名网络使用者组成的样本中 188 人每天收到的 e-mail 超过 10 封，样本比例是多少？检验统计量是多少？
    d. 得出你的结论。

47. 自测题 《消费者报告》表明，64% 的购物者相信超市中的牌子与大众名牌在质量上是一样的。某一大众品牌的番茄酱的生产商为了调查这项结论对自己的产品也适用，特询问了 100 名超市购物者，看他们是否认为超市中的牌子与大众名牌在质量上是一样的。事实上，有 52 名购物者认为超市中的牌子与大众名牌在质量上是一样的。根据此数据检验 H_0: $p\geq 0.64$ 和 H_a: $p<0.64$，取显著性水平为 0.05。
    a. 临界值为多少？拒绝域是什么？
    b. 计算检验统计量。
    c. 得出你的结论。
    d. 计算 p - 值的大小。

48. Berger King 于 1998 年 1 月 2 日星期五发起名为“免费薯条日”促销活动，该活动被认为是目前快餐业历史上最大的一次产品赠送活动。在活动期间，驻留在 Berger King7 400 家快餐店前的任何人将会免费获得一份新鲜的 Berger King 烤薯条。在 Berger King 促销活动之前，需要将 Berger King 新的烤薯条与麦当劳的烤薯条比较来进行一次大众口味检验。在参加口味检验的 500 名消费者中，有 285 名表示喜欢吃这种口味的薯条(*USA Today*, December 10, 1997)。令 p 为喜欢 Berger King 薯条的总体比例。考虑下面的假设检验:

$$H_0: p\leq 0.50$$
$$H_a: p>0.50$$

    a. 喜欢 Berger King 薯条总体比例的点估计是什么？
    b. 计算检验统计量。
    c. 计算 p - 值的大小。
    d. 当 $\alpha=0.01$ 时，你的结论是什么？
    e. 你是否认为该大众口味检验有助于 Berger King 管理人员作出开展“免费薯条日”活动的决定？试讨论。

49. 对加拿大艾伯塔省的 Strathcona 县的司机进行调查，发现在乡村的道路上 48% 的司机在十字

路口的停车信号前不停车(*Edmonton Journal*, July 19, 2000)。假定在接下的调查中发现 800 名司机中有 360 名司机在乡村的道路上在十字路口的停车信号前不停车。

a. 建立原假设和备择假设，以便假定的调查研究是否能决定司机在十字路口的停车信号前不停车的比例发生变化?

b. 当 $\alpha=0.05$ 时，检验的临界点是什么?拒绝规则是什么?

c. 不停车的司机的样本比例是多少?

d. 计算检验统计量。

e. 得出你的结论。

50. 对壳牌石油公司办公人员进行调查,询问他们哪种工作安排更有吸引力:一天 8 小时,一周 5 个工作日,或一天 10 小时一周 4 个工作日(*USA Today*, September 11, 2000)。令 p 为办公人员喜欢一天 10 小时 4 个工作日的比例为一半。当 $\alpha=0.01$ 时,对假设 $H_0: p=0.50$ 和 $H_a: p\neq 0.50$ 检验。由 105 名办公人员组成的样本中有 67 名工作人员喜欢一天 10 小时的一周 4 个工作日的安排。

a. 喜欢一天 10 小时的一周 4 个工作日的安排办公人员的样本比例?

b. 检验统计量的大小?

c. p - 值的大小?

d. 得出你的结论。在这两个选择中是否有显著的偏好?

51. Drugstore. com 是第一家进行网上药品买卖的网络销售公司。Drugstore. com 为其消费者提供了在网上进行购买保健、美容、个人护理和可重复使用药品的机会。在开通的 10 个月后，公司报告 44% 的订单是来源于回头客(*Drugstore. com Annual Report*, January 2, 2000)。假定 Drugstore. com 每季度抽取消费者的订单,从订单中决定回头客的订单比例是否还是原来 $p=0.44$。

a. 建立原假设和备择假设。

b. 在第一季度中由 500 份订单组成的样本中有 205 名回头客。求出 p - 值? 当 $\alpha=0.05$ 时，得出你的结论。

c. 在第一季度中由 500 份订单组成的样本中有 245 名回头客。求出 p - 值? 当 $\alpha=0.05$ 时，得出你的结论。

52. 微软的 Outlook 被认为是最广泛使用的 e-mail 工具。一名微软执行官宣称微软的 Outlook 至少被 75% 的上网者使用。一份美林研究报告表明有 72% 的上网者使用微软的 Outlook(CNBC, June 2000)。当 $\alpha=0.05$ 时，对下面的假设 H_0: $p\geq 0.75$ 和 H_a: $p<0.75$ 进行检验。

a. 假设美林样本大小为 300 名使用者，计算检验统计量的大小。

b. p - 值的大小是多少?

c. 执行官的申明的至少 75% 是否被拒绝?

53. 一个快餐店决定计划实施一次特殊供应，使顾客能购买到专门印有著名卡通人物的杯装饮料。如果有 15% 的消费者购买了这种杯装饮料，则认为可以推行这种特殊供应。在某些地方已经进行的初步检验表明，500 名消费者有 88 名购买了这种杯装饮料。试问是否应推行这种特殊杯装饮料供应? 假设检验将会支持你的决定，当显著性水平为 0.01 时，得出你的建议。

54. 在垒球比赛中，就发生错误概率而言，第三垒是最难打的位置。考虑所有的第三垒选手，一般错误发生概率为 4.7%（*ESPN The Magazine*, July 12, 1999）。一名全明星第三垒选手 Brooks Robinson 在比赛中 1 182 次机会中只发生了 35 次失误。这个样本是否支持 Brooks Robinson 比一般的第三垒选手犯比较少的错误的结论？取 $\alpha=0.01$。

a. 建立适当的原假设和备择假设。

b. Brooks Robinson 发生错误概率的样本比例大小？

c. 计算检验统计量。

d. 计算 p－值的大小。

e. 得出你的结论。

55. 至少有 20% 的工人愿意采用减少工作时间降低收入来获得更多的时间进行个人休闲活动。一项 *USA Today*/CNN/Gallup 的民意调查表明，由 596 名应答者中有 83 人愿意减少工作时间降低收入来获得更多的时间进行个人休闲活动（*USA Today*, April 10, 1995）。当显著性水平为 0.05 时，对假设 H_0：$p\geqslant 0.20$ 和 H_a：$p<0.20$ 进行检验，得出你的结论。

9.7 假设检验和决策支持

在 9.1 节中我们注意到，假设检验所使用的三种形式：

1. 检验研究中的假设。
2. 检验某项声明的有效性。
3. 决策支持中的检验。

在前面两种情况下，只有当原假设 H_0 被拒绝且得出备择假设 H_a 为真的结论时，才需要采取相应措施。在第三种情况下——作出决策，不论是原假设拒绝与否，都有必要采取相应措施。

在进行决策时，因为假设检验程序中没有考虑当 H_0 为真时接受 H_0 的结论而采取相应的措施，所以它的应用受到很大的限制。当检验结果表明不能拒绝 H_0 时，然而决策者不愿接受 H_0，因为决策者不愿承担犯第二类错误的风险，也就是说，当 H_0 为假时才接受它。在前面的章节中所描述的假设检验程序中，第一类错误发生的概率能由检验中的显著性水平控制，但是，犯第二类错误的概率很难控制。

很明显，在某些特定的决策中，决策者可能甚至不得不在**不能在拒绝** H_0 和**拒绝** H_0 之间作出选择并采取相应的措施。对这种情况一个较好的解释是货物接收的例子，这个问题将在第 20 章深入讨论。例如，质量监管人员必须决定是从供应商处接受这船电池，还是因其质量差退还给供应商。假设供应商这批电池设计标准为使用寿命至少为 120 小时。为了评价这船电池的质量，选取 36 节电池组成一个样本对其检验，根据样本来决定是否接受这船电池或因其质量差的原因退还给供应商。令 μ 为这船电池的平均使用寿命时间。总体平均的原假设和备择假设如下：

$$H_0:\ \mu\geqslant 120$$

$$H_a:\ \mu<120$$

如果 H_0 被拒绝，得出备择假设正确。这结论表明应作出将货物退还给供应商的决定。但是，如果 H_0 没被拒绝，决策者仍需决定采取何种措施。因为，没有直接得出 H_0 为真，只是仅仅不能拒绝 H_0，决策者仍可以作出认为质量合格而接受这批货物的决定。

在这种决策的情况下，我们可以将假设检验程序扩展到对第二类错误发生的概率进行讨论。因为当我们不能拒绝 H_0 时，一旦作出决策并采取措施，知道第二类错误发生的概率是有用的。在第 9.8 节和第 9.9 节中，我们将介绍如何计算第二类错误发生的概率和如何通过样本大小来控制发生第二类错误的概率。

9.8　计算第二类错误的概率

在第一节我们将介绍如何计算对总体均值假设检验概率中发生第二类错误的概率。我们以 9.7 节中所描述的货物接收问题为例。关于这批电池的平均使用寿命的原假设和备择假设为 H_0：$\mu \geqslant 120$ 和 H_a：$\mu<120$。如果 H_0 被拒绝，因为其平均使用寿命小于规定的 120 小时，则这批货物将退还给供应商。如果 H_0 没被拒绝，将作出接收这批货物的决定。

假定假设检验中显著性水平 $\alpha=0.05$ 时，检验统计量为

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{\bar{x}-120}{\sigma/\sqrt{n}}$$

$z_{0.05}=1.645$，则左侧检验的拒绝规则为

如果 $z<-1.645$，则拒绝 H_0

假设由 36 个电池组成一个样本，根据以前的检验得知假设总体标准差 $\sigma=12$ 小时，则拒绝规则表明，当

$$z=\frac{\bar{x}-120}{12/\sqrt{36}}<-1.645$$

时，拒绝 H_0。上式的 $\bar{x}$ 的解表明，当

$$\bar{x}<120-1.645\left(\frac{12}{\sqrt{36}}\right)=116.71$$

时，拒绝 H_0。当 $\bar{x}<116.71$ 时表示拒绝 H_0。当 $\bar{x}\geqslant 116.71$ 时，我们作出接收这批货物的决策。

从上面的信息中了解到我们将计算与第二类错误相联系的概率。首先，当货物均值小于 120 小时而我们却作出接受 H_0：$\mu\geqslant 120$ 的决定时，我们就犯了第二类错误。因此，为了计算犯第二类错误的概率，我们必须选择小于 120 小时的 μ 值。例如，假定如果电池的平均使用寿命的时间 $\mu=112$ 小时，则这批电池被认为质量差。如果 $\mu=112$ 为真，则接受 H_0：$\mu\geqslant 120$ 而犯第二类的错误概率？注意到该概率是当 $\mu=112$ 时，样本平均值大于或等于 116.71 的概率。

图 9.14 表示当平均值 $\mu=112$ 时，μ 的抽样分布情况。在右侧的阴影部分表示 $\bar{x}\geqslant 116.71$ 时

的概率。根据图 9.14 查标准正态分布表计算得

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{116.71-112}{12/\sqrt{36}}=2.36$$

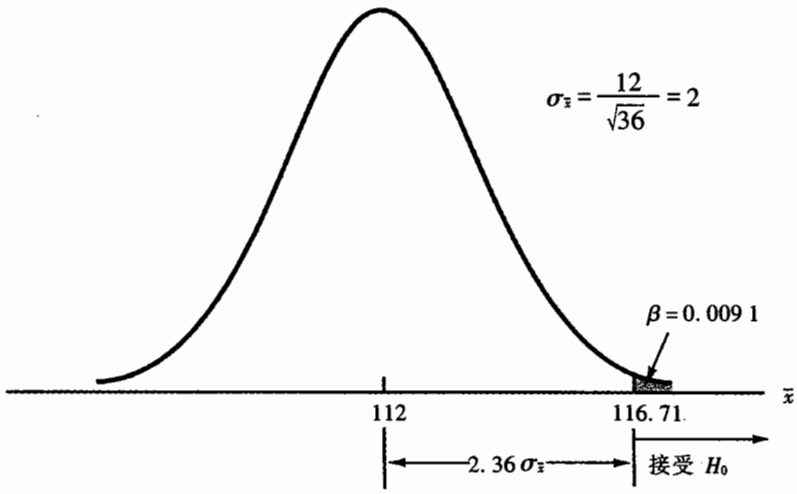


图 9.14　当 $\mu=112$ 时第二类错误的概率

由标准正态概率分布可知，当 $z=2.36$ 时，其右侧面积为 0.500 0 − 0.490 9 = 0.009 1。因而，0.0091 是当 $\mu=112$ 时发生第二类错误的概率。令 β 表示发生第二类错误的概率，我们得出当 $\mu=112$，则 $\beta=0.0091$。所以，我们得出当总体平均值为 112 小时，犯第二类错误的概率只有 0.009 1。

对其他小于 120 的 μ 值，我们可以重复计算过程，给出每一个 μ 值相应发生第二类错误的不同概率。例如，假定这批电池的平均使用寿命 $\mu=115$ 小时。当 $\bar{x}\geqslant 116.71$ 时，我们接受 H_0，计算 $\mu=115$ 时的 z 值

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{116.71-115}{12/\sqrt{36}}=0.86$$

查找标准正态概率分布表可得，当 $z=0.86$ 时，其标准正态概率分布的右侧面积为 0.500 0 − 0.305 1 = 0.194 9。因而，当均值的真值 $\mu=115$ 时，发生第二类错误的概率 $\beta=0.1949$。

在表 9.3 中给出了不同的小于 120 的 μ 值发生第二类错误的概率。注意到，μ 逐渐增加到 120 时，发生第二类错误的概率向右侧 0.95 增加。但是，当 μ 向小于 120 的值减少时，发生第二类错误的概率逐渐减少。这种形式是我们所希望的。当总体 μ 值越来越靠近原假设 $\mu=120$，发生第二类错误的概率越高。然而，当总体平均真值 μ 远小于原假设 $\mu=120$，我们发生第二类错误的概率越低。

表 9.3　货物接收的假设检验中发生第二类错误的概率

μ 值	$z=\frac{116.71-\mu}{12/\sqrt{36}}$	发生第二类错误的概率(β)	β 功效($1-\beta$)
112	2.36	0.009 1	0.990 9
114	1.36	0.086 9	0.913 1
115	0.86	0.194 9	0.805 1
116.71	0.00	0.500 0	0.500 0
117	-0.15	0.559 6	0.440 4
118	-0.65	0.742 2	0.257 8
119.999	-1.645	0.950 0	0.050 0

如表 9.3 所示，第二类错误的概率依赖于总体均值 μ 的值。对 μ_0 附近的 μ，发生第二类错误的概率更大。因为 μ 值未知，所以无法测量发生第二类错误的概率。

当 H_0 为假时，作出拒绝 H_0 的正确结论的概率称做检验的**功效**(power)。对于给定的 μ 值，功效为 $1-\beta$。也就是拒绝原假设的概率为 1 减去发生第二类错误的概率。表 9.3 列示了每个 μ 的功效。根据这些值，将与每一个 μ 值相对应的功效值绘制在图 9.15 中。这样的图被称为**功效曲线**(power curve)。注意，功能曲线渐近于原假设为假时的 μ 值。对于任意 μ 值，功效曲线的高度表示当 H_0 为假时作出拒绝 H_0 正确结论的概率*。

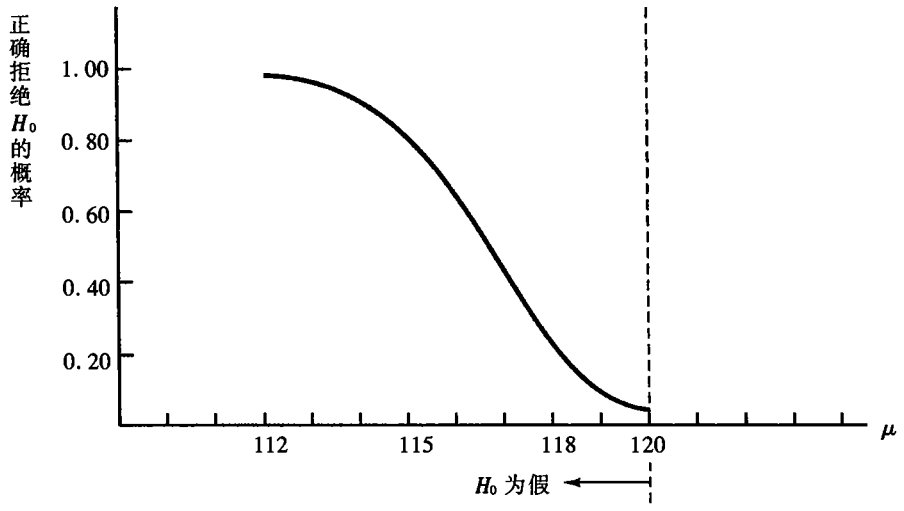


图 9.15　货物假设检验的功效曲线

*在另一张图中，有一条运营特征曲线，有时能提供造成第二类错误的概率的信息。运营特征曲线表明接受 H_0 的概率，因而提供原假设是错误的 μ 值的 β。发生第二类错误的概率也能直接从该图中读出。

总之，计算关于总体均值的假设检验中发生的第二类错误的概率的步骤如下：

1. 建立原假设和备择假设。
2. 在显著性水平 α 下，根据检验统计量来建立拒绝规则。
3. 利用拒绝规则，求检验落入拒绝域的样本均值的范围。
4. 利用步骤 3 中的结果，说明接受 H_0 时所对应的样本均值的值，这样也就同时定义了检验的接受域。
5. 对于备择假设中的 μ 值，利用的抽样分布$\bar{x}$检验从备样假设中的任意值 μ 和步骤 4，计算样本平均值落在接受域的概率。该概率值是当选择 μ 时发生第二类的概率。

练　习

方　法

56. 考虑下面的假设检验：

$$H_0: \mu \geq 10$$
$$H_a: \mu < 10$$

样本容量为 120，总体标准差已知，$\sigma = 5$，取 $\alpha = 0.05$。

a. 如果总体均值为 9，样本均值产生**不拒绝** H_0 的概率是多少？

b. 如果总体均值的真实值为 9 而我们得出 H_0：$\mu \geq 10$ 为真时，产生错误的类型？

c. 如果实际的总体均值为 8，产生第二类错误的概率？

57. 考虑下面假设检验：

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

样本由 200 项组成，总体标准差 $\sigma = 10$。取 $\alpha = 0.05$，当总体均值如下，计算发生第二类错误的概率。

a. $\mu = 18.0$　　b. $\mu = 22.5$　　c. $\mu = 21.0$

应　用

58. Fowle 市场调查公司假定电话调查可在 15 分钟以内结束，并据此向客户收费。如果调查需要更多的时间，则需要加额外费用。由 35 个调查电话组成的样本中，标准差为 4 分钟，显著性水平为 0.01，利用样本平均值来检验原假设 H_0：$\mu \leq 15$。

a. 请解释该问题中的第二类错误？它对公司有什么影响？

b. 当实际的平均时间为 $\mu=17$ 分钟时，发生第二类错误的概率？

c. 当实际的平均时间为 $\mu=18$ 分钟时，发生第二类错误的概率？

d. 画出该检验的功效曲线的一般形状。

59. 一个消费者调查小组对某一汽车生产商声明感兴趣，该汽车生产商称某种新的经济型汽车消耗加仑汽油至少行驶 25 英里（H_0：$\mu\geq 25$）。



a. 当显著性水平为 0.02 时，样本容量为 30 辆汽车，则检验中根据值来决定生产商的声明将被拒绝？假设 σ 为 3 英里/加仑。

b. 当实际的英里数为 23 英里/加仑，发生第二类错误的概率为多少？

c. 当实际的英里数为 24 英里/加仑，发生第二类错误的概率为多少？

d. 当实际的英里数为 25.5 英里/加仑，发生第二类错误的概率为多少？

60. *Young Adult* 杂志对订阅者的平均年龄作出如下假设：

$$H_0:\ \mu=28$$
$$H_a:\ \mu\neq 28$$

a. 在这种情况下发生第二类错误意味着什么？

b. 已知总体标准差 $\sigma=6$ 年，样本大小为 100。当 $\alpha=0.05$ 时，则当 μ 值分别等于 26，27，29 和 30 时，接受 H_0 的概率为多大？

c. 当 $\mu=26$ 时功能为多少？这说明了什么？

61. 对某一产品生产线生产的产品进行检验，对填充重量的精度作如下假设：

假设	结论和措施
H_0：$\mu=16$	填充合格：继续运作
H_a：$\mu\neq 16$	填充偏移标准：停产并调整机器

样本容量为 30，总体标准差 $\sigma=0.8$，取 $\alpha=0.05$。

a. 在这种情况发生第二类错误意味着什么？

b. 当机器填充重量超过 0.5 盎司时，发生第二类错误的概率是多少？

c. 当机器填充重量超过 0.5 盎司时，检验统计量的功效为多少？

d. 画出检验假设的功效曲线。对生产管理者来说，这包含什么信息？

62. 参见习题 58。假定公司选取 50 次调查组成一个样本，重复计算(b)和(c)。你可观察到，随着样本含量的增加，发生第二类错误的概率将如何变化？

63. Sparr 投资有限公司专门为其顾客提供如何进行投资来延缓纳税的机会。最近 Sparr 对某公司的员工提供了工资单扣减投资计划。Sparr 估计现在该公司员工每月延长纳税的投资额不超过 100 美元。Sparr 对由 40 名员工组成的样本对员工总体投资活动水平进行假设检验。假定员工每月延长纳税投资额的标准差为 75 美元。假设检验中取显著性水平为 0.05。

a. 在这种情况下，第二类错误表示什么意思？

b. 如果实际员工每份平均投资额为 120 美元，发生第二类错误的概率为多少？

c. 如果实际员工每份平均投资额为 130 美元，发生第二类错误的概率为多少？

d. 假定样本容量为 80 名员工，重复(b)和(c)。

9.9 确定总体均值假设检验的样本容量

假定对总体均值进行假设检验，由使用者事先指定的显著性水平决定了检验中发生第一类错误的概率。通过控制样本容量，使用者也能控制发生第二类错误的概率。下面我们将介绍如何在总体均值单侧检验中确定样本容量。

$$H_0:\ \mu \geqslant \mu_0$$
$$H_a:\ \mu < \mu_0$$

图 9.16 的右部表示当 H_0 为真时 $\mu=\mu_0$，$\bar{x}$ 的抽样分布。注意使用者实现指定的显著性水平 α 确定了检验的拒绝域。令 c 代表临界值，则 $\bar{x}<c$ 确定了检验中的拒绝域。以 z_α 表示标准正态概率分布左侧面积为 α 时所对应的 z 值，利用下面公式计算 c 值：

$$c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \tag{9.11}$$

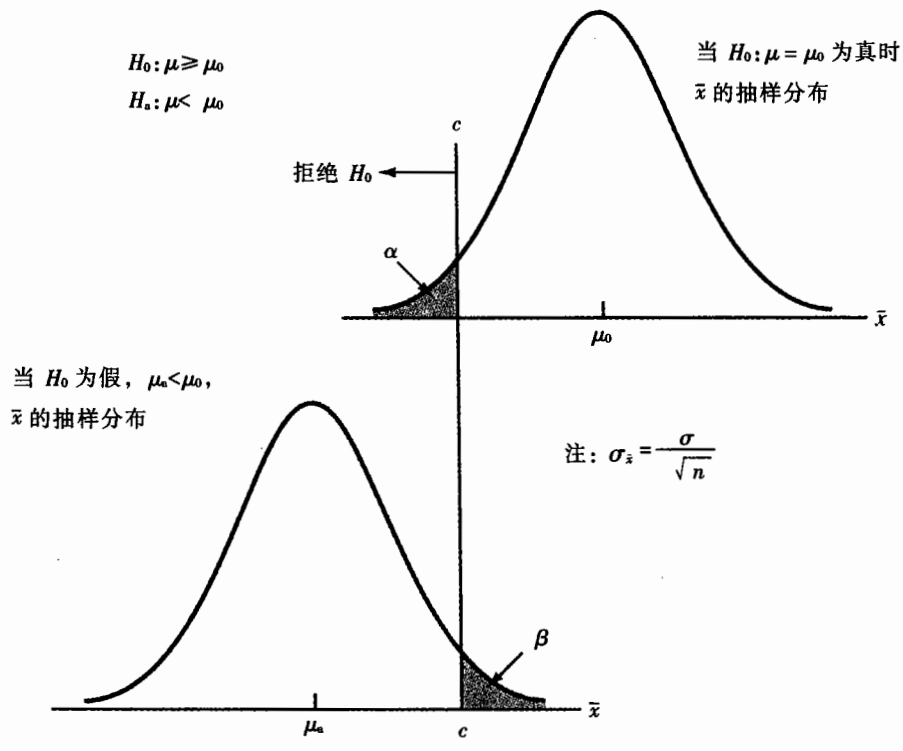


图 9.16　给定第一类错误水平(α)和第二类错误水平(β)的样本容量的确定

现在考虑图 9.16 左侧的抽样分布。当 H_0 为假而 H_a：$\mu_a<\mu_0$ 为真时，我们专门选取 μ_a 代表总体均值。我们假定当总体均值为 μ_a 时，使用者可以接受的第二类错误的概率。在图 9.16 中此概率用 β 表示。以 z_β 为标准正态概率分布左侧面积为 β 时所对应的 z 值，可利用下面公式计算临界值 c：

$$c=\mu_a+z_\beta\frac{\sigma}{\sqrt{n}} \tag{9.12}$$

因为公式(9.11)和(9.12)都是 c 的表达式，我们可知他们相等，因此下述的表达式正确：

$$\mu_0-z_\alpha\frac{\sigma}{\sqrt{n}}=\mu_a+z_\beta\frac{\sigma}{\sqrt{n}}$$

为了确定所要求的样本容量，我们能求 $\sqrt{n}$

$$\mu_0-\mu_a=z_\alpha\frac{\sigma}{\sqrt{n}}+z_\beta\frac{\sigma}{\sqrt{n}}$$

$$\mu_0-\mu_a=\frac{(z_\alpha+z_\beta)\sigma}{\sqrt{n}}$$

和

$$\sqrt{n}=\frac{(z_\alpha+z_\beta)\sigma}{(\mu_0-\mu_a)}$$

表达式两边开算术平方根，即得到如下所示的总体均值单侧检验的样本容量公式：

总体均值单侧假设检验中所需的样本容量

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\mu_0-\mu_a)^2} \tag{9.13}$$

式中　z_α——标准正态分布的一侧面积为 α 的 z 值；

z_β——标准正态分布的一侧面积为 β 的 z 值；

σ——总体标准差；

μ_0——原假设中总体均值；

μ_a——第二类错误所采用的样本均值。

注意：在双侧检验中，公式(9.13)中以 $z_{\alpha/2}$ 代替 z_α。

虽然公式(9.13)是由图 9.16 中显示的假设检验逻辑推演得到的，但其对一个总体均值的单侧检验也成立。注意在一个总体均值的双侧检验中，应采用 $z_{\alpha/2}$ 而不是 z_α。

我们仍回到 9.7 节和 9.8 节中货物接受的例子。这批电池设计的规格为电池的平均寿命至少为 120 小时。如果 H_0：$\mu\geq120$ 被拒绝，则这批货物将被拒绝。我们假定质量监管人员对第一类错误的概率和第二类错误的概率作出如下表达:

关于第一类错误的说明：如果这批电池的平均寿命为 $\mu=120$，那么我愿意冒 $\alpha=0.05$ 的风险概率拒绝这批货物。

关于第二类错误的说明：如果这批电池的平均寿命比规格要求少 5 个小时(即 $\mu=115$),那么我愿意冒 $\beta=0.10$ 的风险概率拒绝这批货物。

这些说明是建立在管理者的判断基础上的。不同的人对此概率作出不同的限制。但是，在样本容量被确定之前，必须对发生两类错误的概率作出说明。

在这个例子中，$\alpha=0.05$，$\beta=0.10$。查标准正态分布表的 $z_{0.05}=1.645$，$z_{0.10}=1.28$。从关于错误的概率中，我们注意到 $\mu_0=120$，$\mu_a=115$。最后，总体标准差 $\sigma=12$ 已知，利用公式(9.13)，我们得出货物接受例子中建议的样本容量为

$$n=\frac{(1.645+1.28)^2(12)^2}{(120-115)^2}=49.3$$

四舍五入，建议样本容量为 50。

当 $n=50$，因为第一类错误和第二类错误概率被控制在允许水平下，所以在假设检验中质量监管人员要在接受 H_0 和拒绝 H_0 之间作出判断，在所允许发生第一类错误和第二类错误的概率内作出相应的推断。

我们观察到 α、β 和样本容量的关系：

1. 如果三个中其中有两个已知，则可计算出第三者。
2. 对于给定的显著性水平 α，增加样本容量可减少 β。
3. 对于给定的样本容量，减少 α，则增加 β，反之，提高 α，则降低 β。

当未能对第二类错误控制，我们应该牢记第三条。这表明不能没必要选择太小的显著性水平。对于给定样本容量，选择较小的显著性水平意味着将使发生第二类错误的风险增大。缺乏经验的使用者，认为在假设检验中 α 越小越好。当我们只考虑第一类错误时，确实如此。然而，不利的是，较小的 α 值将增大发生第二类错误的概率。

练　习

方　法

64. 考虑下面的假设检验：

$$H_0:\ \mu\geq 10$$
$$H_a:\ \mu<10$$

样本容量为 120，总体标准差为 5。取 $\alpha=0.05$，如实际总体均值为 9，第二类错误的概率为 0.291 2。假定研究者想减少第二类错误的概率到 0.10。当实际总体均值为 9，建议的样本容量的大小为多少？

65. 考虑下面的假设检验：

$$H_0: \mu = 20$$
$$H_a: \mu \neq 20$$

总体标准差为 10。取 $\alpha = 0.05$。当实际总体均值为 22，研究者愿意接受发生第二类错误的概率为 0.05，建议的样本大小为多少？

应　用

66. 假定在 Hilltop 咖啡研究中，计划负责人要求在实际重量少于 1 盎司($\mu_a = 2.937\,5$ 磅)而我们却没对 Hilltop 提出上诉(见第 9.3 节)的概率为 0.10，则建议的样本容量为多少？

67. 某种工业电池的寿命至少为 400 小时。假设检验中显著性水平为 0.02。如果某生产过程中生产的电池的实际平均寿命为 385 小时，生产管理者想建立样本程序，保证错误接受这批产品的概率不大于 10%。假设检验中应该采用多大的样本？取总体标准差的估计值为 30 小时。

68. Young Adult 杂志对其订阅者的年龄的均值建立如下假设：

$$H_0: \mu = 28$$
$$H_a: \mu \neq 28$$

如果管理者进行检验时，当实际平均年龄为 29 时，允许发生第二类错误的概率为 0.15，选择样本的容量？假定 $\sigma = 6$，显著性水平为 0.05。

69. 对汽车里程研究者，对下面的假设检验进行检验：

假设	结论
H_0: $\mu \geq 25$mpg	支持生产商的声明
H_a: $\mu < 25$mpg	拒绝生产商的声明；每加仑的里程水平小于说明中的值

取 $\sigma = 3$，显著性水平为 0.02，如果研究者想以 80% 的可能性认为 μ 小于 25 英里(实际为 24)，那么应该选取多大的样本容量？

小　结

假设检验是利用样本数据来确定关于总体参数数值的某项说明应该或不应该被拒绝的统计程序。考虑到样本源的不同情况，必须有两个对立的假设：原假设 H_0 和备择假设 H_a。在某些应用中如何建立原假设和备择假设不是显而易见的。在 9.1 节我们建议谨慎构建三种通常形式的假设检验。

图 9.17*总结了总体均值假设检验的检验统计量，并提供了选择假设检验程序的实践规则。

* 在某些例子中，使用非参数统计方法进行假设检验。在第 19 章将讨论这些方法。

图中说明检验统计量依赖于样本是否是大样本，总体标准差是否已知，在某种情况下，总体是否是正态或近似正态概率分布。如果样本是大样本($n \geq 30$)，假设检验可利用 z 来检验。如果样本是小样本($n<30$)，总体必须是正态或近似正态分布时，才能利用对 μ 值进行假设检验。如果 σ 已知，则利用 z 检验统计量。如果 σ 被样本标准差替代，则利用 t 检验统计量。最后，注意到如果样本太小，而且关于总体是正态分布不合适，则我们建议将样本容量提高到 $n \geq 30$。

假设检验程序的拒绝规则都是将检验统计量的值与临界值相比较。对于左侧检验，如果检验统计量小于临界值，则拒绝原假设。对于右侧检验，如果检验统计量大于临界值，则拒绝原假设。对于双侧检验，如果检验统计量落在抽样分布的任一侧，则拒绝原假设。

我们同样可使用 p - 值进行假设检验。当原假设为真时，p - 值为所获得的样本结果至少与实测结果不同的概率。利用 p - 值进行假设检验时，当 p - 值小于 α 时，拒绝规则要求我们拒绝原假设。

将假设检验程序扩展，给出第二类错误的分析。在 9.8 节，计算发生第二类错误的概率。在 9.9 节中，我们介绍如何确定样本容量来保证我们能够控制发生第一类错误和第二类错误的概率。

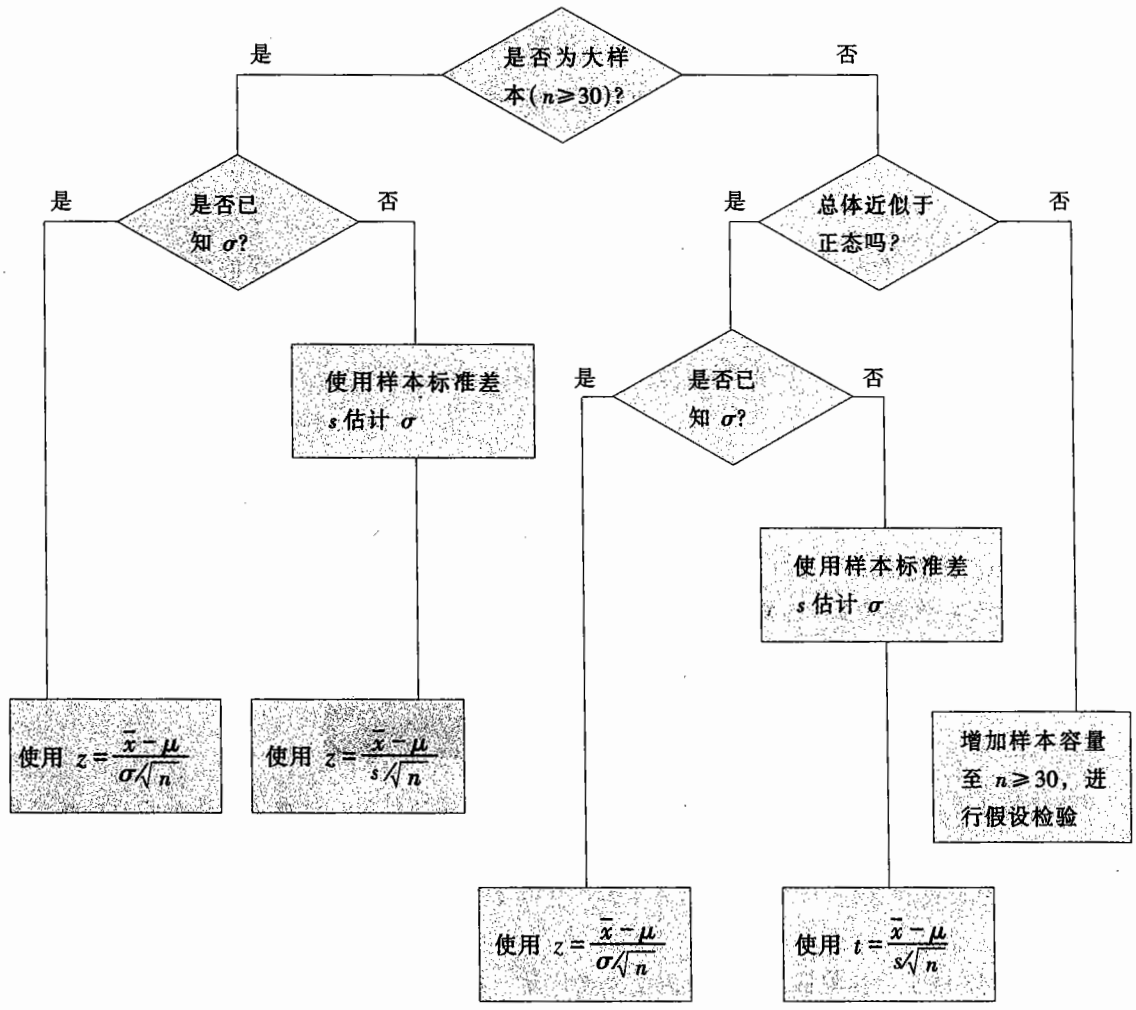


图 9.17　关于总体均值假设检验的检验统计量的总结

术语释义

原假设：在假设检验的程序中，首先假定为真的假设。

备择假设：在拒绝原假设时被认为是真的假设。

第一类错误：当 H_0 为真时却拒绝 H_0 所犯的错误。

第二类错误：当 H_0 为假时却接受 H_0 所犯的错误。

显著性水平：发生第一类错误最大允许的概率值。

拒绝域：产生拒绝原假设结论的值的范围。

检验统计量：决定是否拒绝原假设的统计量。

临界值：与检验统计量相比较，用于确定是否拒绝 H_0 的值。

单侧检验：假设检验的一种，当检验统计量的值在抽样分布的某一侧时，拒绝原假设。

双侧检验：假设检验的一种，当检验统计量的值在抽样分布的任一侧时，拒绝原假设。

***p*－值：**当原假设为真时，所获得的样本结果至少与实测结果不同的概率值，它通常被成为显著性水平。

功效：当 H_0 为假时，作出拒绝 H_0 的正确结论的概率。

功效曲线：是一种用于拒绝 H_0 的概率曲线图，由所有不满足总体参数原假设的值构成。

重要公式

大样本情形下（$n \geq 30$）总体均值单侧检验的检验统计量

σ 已知　$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$　　(9.1)

σ 未知　以 s 替代 σ　$z = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$　　(9.2)

小样本情形下（$n < 30$）总体均值单侧检验的检验统计量

σ 已知　$z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$　　(9.7)

σ 未知　以 s 替代 σ　$t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$　　(9.8)

总体比例假设检验中的检验统计量

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} \qquad (9.9)$$

式中 $\sigma_{\bar{p}}=\sqrt{\dfrac{p_0(1-p_0)}{n}}$ (9.10)

总体均值单侧检验中所需的样本容量

$$n=\frac{(z_\alpha+z_\beta)^2\sigma^2}{(\mu_0-\mu_a)^2} \tag{9.13}$$

在双侧检验中，以 z_α 替代 $z_{\alpha/2}$。

补充练习

70. 纽约州公立中学教师的总体年薪均值为 45 250 美元。纽约市公立中学老师组成的样本平均年薪为 47 000 美元(*Time*, April 3, 2000)。假设纽约市的数据是基于 95 名教师组成的样本。样本标准差为 6 300 美元。
    a. 建立原假设和备择假设，以便于决定样本数据是否支持纽约市公立中学的老师的平均年薪比纽约州公立中学老师平均年薪要高?
    b. 计算检验统计量的大小。
    c. 计算 p - 值的大小。
    d. 当 $\alpha=0.01$ 时，得出你的结论。

71. Ford Taurus 被列为在高速路上行驶每加仑汽油平均功效为 30 英里(1995 *Motor Trend* New Car Buyer's Guide)。一个保护消费者利益的小组对汽车进行检验，通过检验想搜索统计证据来证明汽车生产商夸大汽车汽油的每加仑英里的功效。在 Ford Taurus 例子中，假设检验如下：H_0: $\mu\geqslant 30$，H_a: $\mu<30$。从 Ford Taurus 的 50 次行驶检验组成的样本中，保护消费者利益小组获得在高速路上行驶的样本平均功效为 29.5 每加仑英里，样本标准差为每加仑 1.8 英里。从样本结果中得出你的结论，取显著性水平为 0.01。

72. 据佛罗里达州 Gulf Coast 社区的商业广告称，该地区住宅房产每块地皮的平均成本不超过 25 000 美元。取显著性水平为 0.05，检验这项申明的有效性。假定由 32 块房产组成的样本每块地皮为 26 000 美元，样本标准差为 2 500 美元，p - 值为多少?

73. 某种生产淋浴用的肥皂过程的设计规格为每批平均生产 120 块肥皂。超过或低于这个标准都是不合理的。由 10 批产品组成的样本中，每批生产的产品数量如下，假定总体为正态。

108	118	120	122	119	113	124	122	120	123

    取显著性水平为 0.05，检验该样本结果能否表示该生产过程运作正常与否。

74. 某个城市里两室公寓的月租金为平均 550 美元。假定我们想检验 H_0: $\mu=550$ 和 H_a: $\mu\neq 550$。选取 36 间两室公寓的租金作为样本，样本平均值为 $\bar{x}=562$ 美元，样本标准差 $s=40$ 美元。

a. 当 $\alpha=0.05$ 时，利用检验统计量进行假设检验。

b. 计算 p - 值。

75. Stout Electric 公司的一队卡车是专门为建筑公司提供电力服务的。每辆卡车平均维持费用为 75 美元。由 40 辆卡车组成一个样本，该样本平均维护成本为每月 82.50 美元，样本标准差为 30 美元。管理人员想通过检验来决定是否平均每月的维护费用增加。

a. 当显著性水平为 0.05 时，检验的拒绝规则是什么？

b. 当样本均值为 82.50 美元时，你的结论是什么？

c. 与样本结果相联系的 p - 值为多少？根据 p - 值得出你的结论。

76. 在建筑计划投标中，Sonneborn Builders 有限公司假设建筑工人有 15% 的空闲时间，因而在正常的 8 小时的工作制下，每位工人每天的平均空闲时间不超过 72 分钟。由 30 名建筑工人组成的样本中，建筑工人每月空闲时间为 80 分钟。样本标准差为 20 分钟。设计一个假设检验程序来检验公司假设的有效性。

a. 计算与样本结果相联系的 p - 值。

b. 你的结论是什么？

77. 60% 的美国人认为商业利润分配不合理(General Social Surveys，National Opinion Research Center，University of Chicago)。假定在由 40 名中西部人组成的样本中，有 27 人相信商业利润分配不合理？

a. 这个结果是否能推断大部分中西部人相信商业利润分配不合理？取 $\alpha=0.05$。

b. p - 值为多少？

78. 一份针对商务旅游者的促销计划是根据大约有 2/3 的商务旅游者在过夜的商务旅行中使用“膝上”电脑。取 $\alpha=0.05$，对这个假设进行检验 H_0：$p=0.67$，H_a：$p\neq0.67$。从美国快递在线调查中的样本结果发现，546 名商务旅行者有 355 名在过夜的商务旅行中使用“膝上”电脑(*The Cincinnati Enquirer*, August 31, 1998)。

a. 样本比例的大小为多少？

b. 计算检验统计量的大小。

c. p - 值的大小为什么？

d. 能否拒绝这项声明？试讨论。

79. 1993 年美国移民和归化局报道，79% 的外国旅游者参观美国主要目的是度假。在 2001 年对这项研究继续调查，由 500 名外国旅游者组成的样本中有 360 名声明他们参观美国的主要原因是度假。试问 2001 年来美国旅游度假的外国游客的比重比以前少了？取显著性水平为 0.05，利用检验统计量来支持你的结论？

80. Gallup 根据一个国会小组即国家职业信息协调委员会提高工作信息的使用率的要求对 1 350 人进行一次调查。关于这次调查的问题是：个人得到的工作是他计划想得到的，还是缺乏选择机会或者说是由于偶然原因而得到的？令 p 代表求职中能如愿以偿得到他们所计划得到的工作的

人所占的比例。

a. 如果假设为 H_0：$p \geq 0.50$ 和 H_a：$p < 0.50$，按照研究者的调查结果讨论所研究的假设 H_a。

b. Gallup 民意调查中发现，有 41% 的应答者认为他们得到了他们所计划得到的工作。在显著性水平为 0.01 下，你的结论是什么？试解释。

81. 一名著名的医生假设有 75% 的女性所穿的鞋过小。美国足部和踝部矫形协会对 356 名妇女调查发现，其中有 313 名妇女所穿鞋子的号码至少小了一号。取 $\alpha = 0.01$，检验 H_0：$p = 0.75$ 和 H_a：$p \neq 0.75$。你的结论是什么？

82. 运输部门报道 Amtrack 列车在过去的 12 个月里准点到达的记录为 78%（*USA Today*，November 23, 1998）。假定在 2001 年进行调查，发现 400 辆 Amtrack 列车有 330 辆准点到达。是否检验结果表明 Amtrack 列车准点到达率改变了？当 $\alpha = 0.05$ 时，检验 H_0: $p = 0.78$ 和 H_a: $p \neq 0.78$。

a. Amtrack 列车准点到达的样本比例为多少？

b. 计算检验统计量。

c. 计算 p - 值的大小。

d. 得出你的结论。

83. 据 Myrtle Beach 的电台报道，在纪念日的周末至少有 90% 的旅馆和汽车旅店会爆满。电台建议听众在周末想来风景点度假的话，请事先预定，星期六的晚上，一个由 58 家旅馆和汽车旅店会组成的一个样本，样本表明 49 家挂出无空房的招牌，有 9 家还有空房。在看了这些数据后，你对电台的报道有什么反应？取 $\alpha = 0.05$，进行统计检验，求出 p - 值？

84. 环境健康指标包括空气指标、水指标和食物指标。在 25 年前，美国食物样品中有 47% 存在杀虫剂残留物（*U. S. News & World Report*，April 17, 2000）。在最近的研究中，125 份样本食物有 44 份存在杀虫剂残留物。检验假设 H_0：$p \geq 0.47$ 和 H_a：$p < 0.47$。

a. 样本比例为多少？

b. 计算检验统计量的值。

c. p - 值为多少？

d. 当 $\alpha = 0.01$ 时，你的结论是什么？

85. 参见习题 76。

a. 当平均空闲时间为 80 分钟时，发生第二类错误的概率是多少？

b. 当平均空闲时间为 75 分钟时，发生第二类错误的概率是多少？

c. 当平均空闲时间为 70 分钟时，发生第二类错误的概率是多少？

d. 画出该问题的功效曲线。

86. 一个美国联邦救济方案仅适用于低收入地区。一个地区平均每年平均家庭收入小于 15 000 美元，该地区才能有资格得到这份补助。每年平均家庭收入大于或等于 15 000 美元的地区将没有资格得到补助。根据该地区居民所组成的样本，在显著性水平为 0.02 下进行假设检验，来决定该地区是否有资格得到这个资助。如果补助规则要求，一个地区平均每年平均家庭收入小于 14 000 美元不能得到补助的概率最大不超过 0.05 的话，那么在补助决策研究中应采用多

大的样本容量？取 $\sigma = 4\,000$ 美元作为计划值。

87. 用 H_0：$\mu = 120$ 和 H_a：$\mu \neq 120$ 来检验淋浴用的肥皂生产过程是否满足每批产量为 120 块的标准。取显著性水平 α 为 0.05，标准差的计划值为 5。
   a. 如果实际产量为每批 117 块，公司想有 98% 的机会得出该标准产品产量未达到的结论，应采用的样本容量是多少？
   b. 利用（a）的样本容量，当实际平均产量为每批 117、118、119、121、122 和 123 块肥皂？也就是说，在每种情况下，发生第二类错误的概率的大小是多少？

案例研究 1　Quality Associates 有限公司

Quality Associates 是一家咨询公司，是为其客户监控生产过程提供抽样和统计程序的建议。在某一应用中，客户为 Quality Associates 提供了其生产过程运行的 800 个观测值组成的样本。这些数据的样本标准差为 0.21，因此，总体标准差假定为 0.21。Quality Associates 建议抽取样本容量为 30 的样本，定期连续监控这一生产过程。通过分析这些样本，客户可以迅速了解到该生产过程是否运作正常。当生产过程运作不正常时，可采取适当措施来减少这些问题。该生产过程设计规格为 12。Quality Associates 建议采取下面的假设检验。

$$H_0: \mu = 12$$
$$H_a: \mu \neq 12$$

一旦 H_0 被拒绝，则应采取措施。

以下的样本是新的统计监控程序运行的第一天，每隔一个小时收集到的。它们被存在数据集 Quality 中。


CD 光盘数据
Quality

样本 1	样本 2	样本 3	样本 4
11.55	11.62	11.91	12.02
11.62	11.69	11.36	12.02
11.52	11.59	11.75	12.05
11.75	11.82	11.95	12.18
11.90	11.97	12.14	12.11
11.64	11.71	11.72	12.07
11.80	11.87	11.61	12.05
12.03	12.10	11.85	11.64
11.94	12.01	12.16	12.39
11.92	11.99	11.91	11.65
12.13	12.20	12.12	12.11
12.09	12.16	11.61	11.90
11.93	12.00	12.21	12.22

(续表)

样本 1	样本 2	样本 3	样本 4
12.21	12.28	11.56	11.88
12.32	12.39	11.95	12.03
11.93	12.00	12.01	12.35
11.85	11.92	12.06	12.09
11.76	11.83	11.76	11.77
12.16	12.23	11.82	12.20
11.77	11.84	12.12	11.79
12.00	12.07	11.60	12.30
12.04	12.11	11.95	12.27
11.98	12.05	11.96	12.29
12.30	12.37	12.22	12.47
12.18	12.25	11.75	12.03
11.97	12.04	11.96	12.17
12.17	12.24	11.95	11.94
11.85	11.92	11.89	11.97
12.30	12.37	11.88	12.23
12.15	12.22	11.93	12.25

管理报告

1. 对每个样本，取显著性水平为 0.01，进行假设检验，并在需要采取措施时，确定采取何种措施？给出每次检验的检验统计量和 p – 值。
2. 计算这四个样本的标准差。假定总体标准差 0.21 是否合理？
3. 当样本平均值在 $\mu = 12$ 附近的多大范围内，我们可以认为该生产运行正常？如果超过上限或低于下限，则采取求证措施。在质量控制中，这上限或下限被称作右侧或左侧的控制限。
4. 当显著性水平增大时，这表示什么意思？这时，如果这样做，哪种错误或误差将增加？

案例研究 2　失业研究

美国就业统计局每月公布失业统计数据，包括失业人数和平均失业时间。在 1998 年 11 月，美国就业统计局报告平均失业时间为 14.6 个星期。

费城市长对费城地区的失业状况进行调查。由费城的 50 名失业居民组成的样本中包括失业年龄和失业周期。在 1998 年 11 月收集的部分数据如下。完整数据集在光盘文件 BLS 中。


CD 光盘数据
BLS

年龄	周数	年龄	周数
56	22	25	5
35	19	40	20
22	7	25	12
57	37	25	1
40	18	59	33
22	11	49	26
48	6	33	13
48	22		

管理报告

1. 利用描述统计来总结数据。
2. 对费城地区的失业平均年龄建立 95% 置信区间估计。
3. 进行假设检验，以便于决定是否支持费城平均失业周期大于全国平均失业周期的 14.6 周？取显著性水平 $\alpha=0.01$，你的结论是什么？
4. 个人失业年龄和失业周期数之间是否有关系？请解释。

附录 9.1　用 Minitab 进行假设检验

在附录中，我们将介绍如何对大样本和小样本的总体进行检验。

大样本情形


CD 光盘数据
Distance

在表 9.2 中我们用高尔夫球射程例子来解释了大样本情形。将数据输入 Minitab 工作表的 C1 中。假设检验的显著性水平 $\alpha=0.05$，总体标准差 σ 用样本标准差 s 估计。采用下面步骤对假设 H_0：$\mu=280$ 和 H_a：$\mu\neq280$ 进行检验：

步骤 1. 选择下拉菜单 **Calc**；

步骤 2. 选择 **Column Statistics**；

步骤 3. 当 Column Statistics 对话框出现时：

选择 **Standard Deviation**；

在 **Input variable** 对话框中键入 C1；

在 **Store result in** 框内键入 stdev；

点击 **OK**；

步骤 4. 选择 **Stat** 下拉菜单；

步骤 5. 选择 **Basic Statistics** 下拉菜单；

步骤 6. 选择 **1-Sample Z**;

步骤 7. 当 1-Sample Z 出现时:

在 **Variables** 框中键入 C1;

在 **Sigma** 框中选择 stdev;

在 **Test mean** 框中键入 280;

选择 **Options**;

步骤 8. 在 **Confidence level** 框中键入 95*;

在 **Alternative** 框中选择不相等;

点击 **OK**;

步骤 9. 点击 **OK**。

图 9.8 中的假设检验结果中，Minitab 程序中提供了总体均值的 95% 置信区间。

在对 Superflight 高尔夫球问题所做的研究中，利用样本标准差 s 代替总体标准差 σ。在大样本的假设检验中如果 σ 已知，第一步到第三步的程序被略去。在这种情况下，从第四步开始，在第七步中在 **Sigma** 框中键入适当的 σ 值。最后，在对 Superflight 高尔夫球问题所做的研究中，所涉及到的假设检验是双侧检验。我们只需简单地在第八步的 **Alternative** 框中选择 less than 或 greater than 选项，便可以将该程序很容易转化为适用于单侧检验的情况。

小样本情形



CD 光盘数据
Heathrow

在第 9.5 节中研究的 Heathrow 机场评定级别的问题中，我们介绍其方法。假定将数据输入 Minitab 工作栏中的 C1 中，检验中显著性水平 $\alpha=0.05$，利用样本标准差 s 代替总体标准差 σ。采用下面的步骤来对假设 H_0: $\mu\leqslant 7$ 和 H_a: $\mu>7$ 检验:

步骤 1. 选择 **Stat** 下拉菜单;

步骤 2. 选择 **Basic Statistics** 下拉菜单;

步骤 3. 选择 **1-Sample t**;

步骤 4. 当 1-Sample t 出现时:

在 **Variables** 框中键入 C1;

在 **Test mean** 框中键入 7;

选择 **Options**;

步骤 5. 在 **Confidence level** 框中键入 95;

在 **Alternative** 框中选择 greater than;

点击 **OK**;

步骤 6. 点击 **OK**。

* Minitab 可同时进行假设检验和区间估计。使用者可选择关于总体均值的区间估计的任何显著性水平大小，这里的置信度为 95%。

在这个程序中利用样本标准差 s 代替总体标准差 σ。在小样本假设检验中，如果 σ 已知，可以使用大样本 1-Sample Z 程序的步骤 4 到步骤 9，在步骤 7 中将适当的 σ 值键入 **Sigma** 框中。最后，研究的 Heathrow 机场评定级别的问题所涉及的是单侧假设检验。我们只需简单的在步骤 5 的 **Alternative** 框中选择 less than 或 not equal 选项，便可以进行其他假设检验。

附录 9.2　用 Excel 进行假设检验

大样本情形



CD 光盘数据
Distance

在表 9.2 中我们用高尔夫球射程例子来解释了大样本情形。在 Excel 工作表中，变量 Yards 在单元格 A1 中，单元格 A2 到 A37 中存放着 36 个射程的数据。注意假设检验的显著性水平 $\alpha=0.05$，总体标准差 σ 用样本标准差 s 替代。

用 Excel 进行假设检验相对简单的方法是建立扩展的工作表，并利用 p - 值临界值得出你的结论。我们建立的工作表在图 9.18 中。采取下面的步骤检验 H_0：$\mu=280$ 和 H_a：$\mu\neq280$：

步骤 1. 在单元格 D2 中键入 280；

步骤 2. 在单元格 D6 中计算样本容量；

=COUNT(A2: A37)

步骤 3. 在单元格 D7 中计算样本平均值；

=AVERAGE(A2: A37)

步骤 4. 在单元格 D8 中计算样本标准差；

=STDEV(A2: A37)

步骤 5. 在单元格 D9 中计算检验统计量；

=（D7 - D2）/(D8/SQRT(D6))

步骤 6. 在单元格 D10 中计算 p - 值。

=2*(1 - NORMSDIST (ABS(D9)))

单元格 D9 中包含了检验统计量的公式：

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

单元格 D10 包含了计算 p - 值的公式，将 p - 值和单元格 D9 中的正态分布的双侧面积相关的检验统计量 z 比较。

在对 Superflight 高尔夫球研究中利用样本标准差 s 代替总体标准差 σ。在大样本假设检验中如果 σ 已知，因为没有必要计算，可直接将 σ 值键入单元格 D8 中。最后，对 Superflight 高

尔夫球研究涉及到双侧检验。只需要进行简单修改，可适用于单侧检验。当拒绝域在左侧时，单元格 D10 中公式 = NORMSDIST(D9)给出 p - 值；当拒绝域在右侧时，单元格公式 = 1 - NORMSDIST(D9)给出 p - 值。

Yards		
269	Hypothesis Test	
300	Test Mean	280
268		
278	Sample Results	
282	Sample Size	36
263	Mean	278.5
301	Standard Deviation	12
295	Test Statistic	-0.75
288	p-value	0.453
278		
276		
286		
274		
277		

图 9.18　Superflight 高尔夫球用 Excel 假设检验的扩展栏

小样本情形


CD 光盘数据
Heathrow

在第 9.5 节中研究的 Heathrow 机场评定级别的问题中，我们介绍其方法。在 Excel 工作表中，变量 Rating 键入 A1，单元格 A2：A13 中存放 12 个评定级别数据。取显著性水平 $\alpha = 0.05$，用样本标准差 s 代替总体标准差 σ。图 9.18 示出了扩展的工作表。采取下面的步骤对假设 H_0：$\mu \leqslant 7$ 和 H_a：$\mu > 7$ 检验：

步骤 1. 在单元格 D2 中键入 7；

步骤 2. 在单元格 D6 中计算样本容量；

= COUNT(A2: A13)

步骤 3. 在单元格 D7 中计算样本平均值；

= AVERAGE(A2: A13)

步骤 4. 在单元格 D8 中计算样本标准差；

= STDEV(A2: A13)

步骤 5. 在单元格 D9 中计算检验统计量；

= (D7-D2) /(D8/SQRT(D6))

步骤 6. 在单元格 D10 中计算 p - 值。

= IF(D9 > 0, TDIST(D9, D6-1, 1),1-TDIST (ABS(D9), D6-1,1))

单元格 D9 中包含了检验统计量的公式：

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

单元格 D10 包含了计算 p - 值的公式，将 p - 值和 t 分布面积相关的检验统计量 t 比较。因为计算 p - 值公式依赖于单元格 D9 中的检验统计量正负，因此这个公式相对复杂。

在对 Heathrow 机场研究中利用样本标准差 s 代替总体标准差 σ。在小样本假设检验中如果 σ 已知，因为没有必要计算，可直接将 σ 值键入单元格 D8 中。在这种情况下，检验统计量 z 和相关的 p - 值与大样本情形下的公式一样。

最后，对 Heathrow 机场研究涉及到右侧的单侧检验。只需要进行简单修改，可转化为适用其他假设检验。例如，当拒绝域在左侧时，单元格 D10 中公式 = IF(D9 > 0,1-TDIST(D9,D6-1,1), TDIST (ABS(D9),D6-1,1)) 给出 p - 值，当是双侧检验时，单元格公式 = TDIST(ABS(D9),D6-1, 2) 给出 p - 值。

10 两个总体均值和比例的统计推断

统计实例：FISONS 公司

10.1 两个总体均值之差的估计：独立样本

$\bar{x}_1 - \bar{x}_2$ 的抽样分布

$\mu_1 - \mu_2$ 的区间估计：大样本情形

$\mu_1 - \mu_2$ 的区间估计：小样本情形

10.2 两个总体均值之差的假设检验：独立样本

大样本情形

小样本情形

10.3 两个总体均值之差的统计推断：匹配样本

10.4 两个总体比例之差的统计推断

$\bar{p}_1 - \bar{p}_2$ 的抽样分布

$p_1 - p_2$ 的区间估计

关于 $p_1 - p_2$ 的假设检验

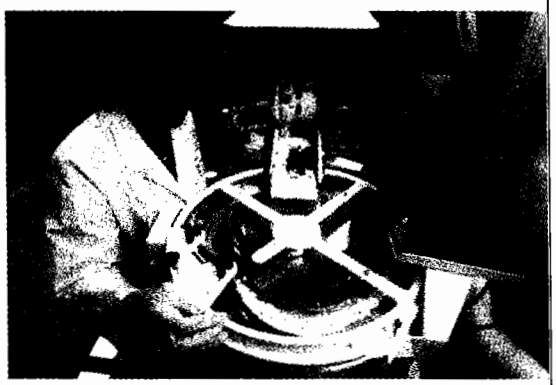
FISONS 公司

罗切斯特，纽约州

位于纽约州罗切斯特的 Fisons 公司，是英国 Fisons 跨国公司的子公司之一。1966 年，Fisons 公司在美国开始有业务了。

Fisons 公司的制药部门大量采用统计学方法检测、开发新药。制药业中的检测过程通常分为三个阶段：(1)临床前的使用；(2)长期服用安全性检测；(3)临床效果检测。在每个阶段，药物能够通过严格检测的机会很少。尽管如此，用于进一步检测的费用依然大量增加，行业调查表明，研究开发一种新药平均需要用 12 年时间，花费 2.5 亿美元。因此，在早期的检测过程中排除不成功的新药，以及识别可用于进一步研究的有开发前景的新药就显得相当重要了。

在药物研究中，统计学起了重要作用，政府部门对此有严格规定而且执行也很严格。在临床前的检测中，要决定一种新药能否继续用于长期使用和安全性检验，一般要对两个或三个总体进行统计研究，总体一般由新药、对照药和标准药组成，临床前检测过程是从将一种新药送交药理组评价，检验出药物产生药效开始。作为检验过程的一个环节是请一位统计学家来设计检测新药的实验。该实验必须详细说明样本容量和统计学分析方法。在两个总体的研究中，一个样本用于获得新药(总体 1)的药效数据，第二个样本用于获得标准药(总体 2)的药效数据。根据需要的不同，新药和标准药将在神经学、心脏病学及免疫学等一些学科进行检测。在多数研究中，统计学方法包括对有关新药总体和标准药总体均值之差的假设检验。如果一种新药同标准药相比缺乏药效或有副作用，那就要放弃对它进行进一步检测。只有那些同标准药相比，表现出广阔前景的新药，才能可以进行长期服用安全性检测。



统计方法应用于新药的研制开发。© Mark Richards/PhotoEdit.

研究者进一步收集数据，进行多种总体研究，并将其贯穿于长期使用。在安全性检测及临床检验的过程中，美国联邦食品药品管理局(FDA)要求，必须在这些检验之前确定统计方法，以避免与有关数据有关的偏差。另外，有些临床实验是双重或三重保密，为了避免人为因素的偏差。也就是说，无论被检者还是调查者都不知道哪种药分配给谁。如果一种新药达到了标准药的所有要求，FDA 将批发一份新药应用(NDA)文件。该文件将由主管部门的统计学家和科学家进行严格细致的审阅。

在本章中，你可以学到有关两个总体均值、比例之差和如何构造区间估计进行假设检验。这些方法将通过分析独立随机样本和匹配样本介绍给读者。

在第 8 章和第 9 章中，我们讲述了如何对一个总体均值及比例进行区间估计和假设检验。在本章中，我们将继续统计推断问题的讨论，讲述当两个总体均值或比例之差很重要的情况下，如何进行两个总体情形的区间估计和假设检验。比如，我们可以对男女两总体平均起始工资之差进行区间估计，或者，对供应者 A、B 生产的产品次品率的差异进行假设检验。我们将通过一个实例来讨论两总体均值和比例统计推断，该实例介绍了由 Greystone 百货公司进行一种抽样调查中，怎样对两个总体均值之差进行区间估计。

10.1　两个总体均值之差的估计：独立样本

Greystone 百货公司在纽约州的布法罗开了两家商店：一家在市中心，另一家在郊区购物中心。某地区的经理发现在一家商店畅销的东西不一定在另一家好卖。经理认为这种情况应归为两地顾客群体的差异，即顾客可能在年龄、受教育程度和收入等方面存在差异。假设经理要求调查两家商店顾客年龄上的差异。

我们定义总体 1 为市中心商店的所有顾客，总体 2 为郊区商店的所有顾客。

μ_1 = 总体 1 的均值（市中心商店的所有顾客平均年龄）

μ_2 = 总体 2 的均值（郊区商店的所有顾客平均年龄）

两个总体均值之差是 $\mu_1 - \mu_2$。

为了检验 $\mu_1 - \mu_2$，我们从总体 1 中抽取 n_1 个顾客，将其组成简单随机样本，从总体 2 中抽取 n_2 个顾客，将其组成简单随机样本。因为 n_1 个顾客和 n_2 个顾客的简单随机样本都是独立抽取的，所以我们将其称为**独立简单随机样本**(independent simple random samples)。

$\bar{x}_1$ = n_1 个市中心顾客组成的简单随机样本的样本平均年龄

$\bar{x}_2$ = n_2 个郊区中心顾客组成的简单随机样本的样本平均年龄

因为$\bar{x}_1$ 是 μ_1 的点估计量，$\bar{x}_2$ 是 μ_2 的点估计量，所以将两个总体均值之差的点估计量表示如下：

两个总体均值之差的点估计量

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

可知，总体均值之差的点估计量是两个独立简单随机样本的样本均值之差。图 10.1 是一个基于两个独立简单随机样本且用于估计两个总体均值之差的步骤示意图。

假设从 Greystone 顾客中的两个独立简单随机样本收集来的顾客平均年龄的数据如下：

商店	顾客抽样数	样本平均年龄	样本标准差
市区	36	$\bar{x}_1$ = 40 岁	s_1 = 9 岁
郊区	49	$\bar{x}_2$ = 35 岁	s_2 = 10 岁

利用公式(10.1)，我们求出两个总体平均年龄之差的点估计量为$\bar{x}_1 - \bar{x}_2 = 40 - 35 = 5$ 岁。因

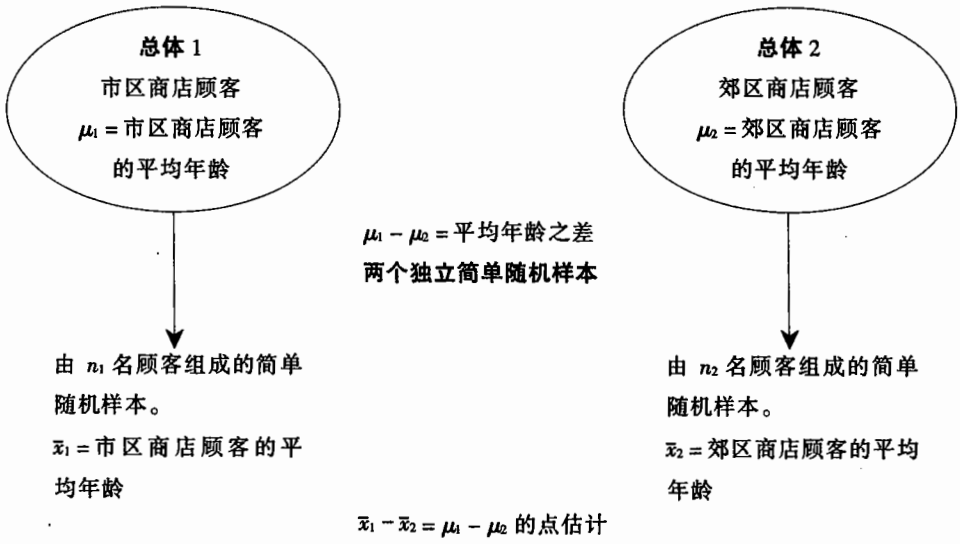


图 10.1　两个总体均值之差的估计

此，我们认为市区商店的顾客平均年龄比郊区大 5 岁。跟其他点估计量一样，5 岁只是两个总体平均年龄之差的很多可能的点估计量之一。如果 Greystone 选择了另一个由 36 名市区顾客组成的简单随机样本和另一个由 49 名郊区顾客组成的简单随机样本，这两个样本均值之差就可能不等于 5 岁，$\bar{x}_1 - \bar{x}_2$ 的抽样分布就是两个独立简单随机样本所有可能的样本均值之差的概率分布。

$\bar{x}_1 - \bar{x}_2$ 的抽样分布

我们利用$\bar{x}$ 的抽样分布对单个总体均值进行区间估计，同样，我们也可以利用$\bar{x}_1 - \bar{x}_2$ 的抽样分布进行两个总体均值之差的区间估计。$\bar{x}_1 - \bar{x}_2$ 的抽样分布有以下性质：

$\bar{x}_1 - \bar{x}_2$ 的抽样分布

期望值：$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$　　(10.2)

标准差：$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$　　(10.3)

式中　σ_1——总体 1 的标准差；

σ_2——总体 2 的标准差；

n_1——总体 1 的简单随机样本的样本容量；

n_2——总体 2 的简单随机样本的样本容量。

分布形式：若两个样本容量都很大（$n_1 \geqslant 30$ 且 $n_2 \geqslant 30$），则$\bar{x}_1 - \bar{x}_2$ 的抽样分布近似服从正态分布。

图 10.2 描述了$\bar{x}_1 - \bar{x}_2$ 的抽样分布及$\bar{x}_1$ 和 $\bar{x}_2$ 单个分布的关系。

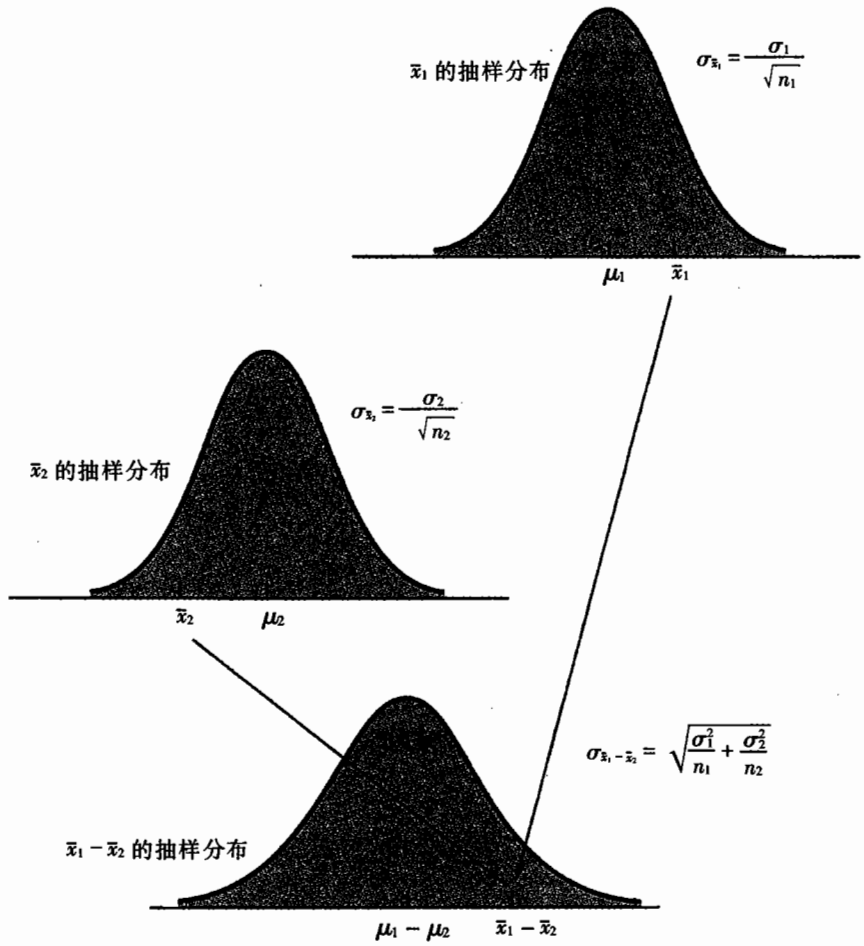


图 10.2　$\bar{x}_1-\bar{x}_2$ 的抽样分布及$\bar{x}_1$ 和$\bar{x}_2$ 单个分布的关系

现在我们进行对两个总体均值之差的区间估计。我们考虑两种情况：一种是样本容量较大的情形（$n_1\geqslant 30$ 且 $n_2\geqslant 30$），另一种是至少有一个样本容量较小的情形（$n_1<30$ 且/或 $n_2<30$）。首先，我们对大样本情形进行分析。

$\mu_1-\mu_2$ 的区间估计：大样本情形

在大样本情形下，$\bar{x}_1-\bar{x}_2$ 的抽样分布可以由正态概率分布来描述。因此，我们用下面的公式进行两个总体均值之差的区间估计：

两个总体均值之差的区间估计：大样本情形（$n_1\geqslant 30$ 且 $n_2\geqslant 30$），σ_1 和 σ_2 已知

$$\bar{x}_1-\bar{x}_2\pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2} \tag{10.4}$$

式中，$1-\alpha$ 为置信系数。

用公式(10.4)对两总体之差进行区间估计，就必须预先知道$\bar{x}_1-\bar{x}_2$的抽样分布的标准差$\sigma_{\bar{x}_1-\bar{x}_2}$的值。而公式(10.3)表明，$\sigma_{\bar{x}_1-\bar{x}_2}$的值取决于每个总体的标准差 σ_1、σ_2 的值。当总体标准差未知，我们用样本标准差作为总体标准差的估计，且将$\sigma_{\bar{x}_1-\bar{x}_2}$估计如下：

$\sigma_{\bar{x}_1-\bar{x}_2}$ 的点估计量

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{10.5}$$

因此，在大样本情形下，我们可以用$\sigma_{\bar{x}_1-\bar{x}_2}$的点估计量计算出如下的两个总体均值之差的值的新区间估计：

两个总体均值之差的置信区间估计

大样本情形（$n_1\geq 30$ 且 $n_2\geq 30$），σ_1、σ_2 由 s_1、s_2 估计

$$\bar{x}_1-\bar{x}_2\pm z_{\alpha/2}s_{\bar{x}_1-\bar{x}_2} \tag{10.6}$$

式中，$1-\alpha$ 是置信系数。

在对 Greystone 百货公司的研究中，我们可用公式(10.6)来进行两个顾客总体的平均年龄之差的置信区间估计。我们回顾一下，36 名市区顾客简单随机样本的均值和标准差分别为$\bar{x}_1=40$ 岁，$s_1=9$ 岁，49 名郊区顾客简单随机样本的均值和标准差分别为$\bar{x}_2=35$ 岁，$s_2=10$ 岁。利用公式(10.5)来估计$\sigma_{\bar{x}_1-\bar{x}_2}$，我们可得

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{(9)^2}{36}+\frac{(10)^2}{49}}=2.07$$

对于$z_{\alpha/2}=z_{0.025}=1.96$，由公式（10.6）可得下面的 95% 置信区间：

$$5\pm(1.96)(2.07)$$

或

$$5\pm 4.06$$

因此，在 95% 置信度下，Greystone 两个总体平均年龄之差的区间估计为 0.94 到 9.06 岁之间。

$\mu_1-\mu_2$ 的区间估计：小样本情形

我们现在考虑一下，小样本情形即当至少有一个样本容量小于 30 时，两个总体均值之差的区间估计方法，也就是，$n_1<30$ 且/或 $n_2<30$。

在第 8 章中，我们讲述了小样本情形下单个总体均值的区间估计方法。我们通过回顾知道，小样本情形需要总体服从正态分布的假设。假定总体标准差已知，通过标准正态分布 z 进行区间估计。如果用样本标准差 s 来对总体标准差 σ 估计，将通过 t 分布对其进行区间估计。

在进行小样本的两总体均值之差的区间估计时，我们假设两总体都是正态分布。若假设总体标准差 σ_1、σ_2 已知，那么大样本区间公式（10.4）可用于小样本情形。然而，如果总体标准差 σ_1、σ_2 被样本标准差 s_1、s_2 估计，那么用 t 分布对小样本两个总体均值之差进行区间估计。该过

程的步骤和举例如下，我们首先作出如下假设：

当用 s_1、s_2 估计 σ_1、σ_2 时，可用 t 分布对小样本的两个总体均值之差进行区间估计。

---

1. 两个总体服从正态分布。
2. 两个总体的方差相等($\sigma_1^2 = \sigma_2^2 = \sigma^2$)。

当样本容量相等，即使总体方差不等，也能在这部分的步骤中求得可接受的结果。因此，只要可能，研究者应考虑到的样本容量 $n_1 = n_2$。

在该假设下，无论样本容量的大小，$\bar{x}_1 - \bar{x}_2$ 的样本分布为正态分布。$\bar{x}_1 - \bar{x}_2$ 的期望值是 $\mu_1 - \mu_2$。由方差相等的假设，公式(10.3)可写为

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.7}$$

$\bar{x}_1 - \bar{x}_2$ 的抽样分布如图 10.3 所示。

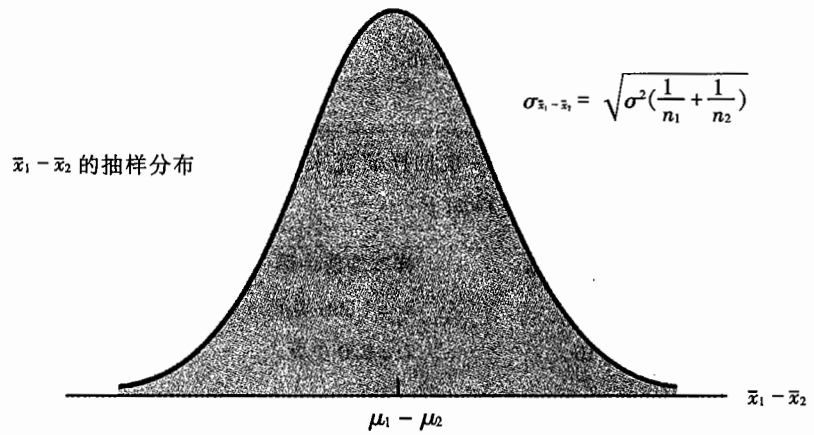


图 10.3　当总体服从正态分布时，$\bar{x}_1 - \bar{x}_2$ 的抽样分布

因为公式（10.7）是基于 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 的假设，所以我们不需要分别估计 σ_1^2 和 σ_2^2。实际上，我们可以将两个样本的数据组合，以得到 σ^2 的最优单个估计。将两个独立简单随机样本的数据结合为 σ^2 的一个估计量的过程为**合并**。σ^2 的**合并方差**(pooled variance)估计量是两个样本方差 s_1^2 和 s_2^2 的加权平均，记为 s^2。σ^2 的合并估计量的公式如下：

σ^2 的合并估计量

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{10.8}$$

σ^2 的合并方差估计量是两个样本方差的加权平均，其取值点介于 s_1^2 和 s_2^2 之间。

用 σ^2 的合并估计量 s^2，我们可由公式(10.7) 得 $\sigma_{\bar{x}_1 - \bar{x}_2}$ 的估计式如下：

当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 时，$\sigma_{\bar{x}_1 - \bar{x}_2}$ 的点估计量

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.9}$$

现在用 t 分布来计算两个总体均值之差的区间估计。总体 1 的样本有 $n_1 - 1$ 个自由度，总体 2 的样本有 $n_2 - 1$ 个自由度，所有该 t 分布的自由度为 $n_1 + n_2 - 2$。区间估计如下：

两个总体均值之差的区间估计

小样本情形（$n_1 < 30$ 且/或 $n_2 < 30$），σ_1、σ_2 由 s_1、s_2 估计

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2} \tag{10.10}$$

式中 t——自由度为 $n_1 + n_2 - 2$ 的 t 分布；

$1 - \alpha$——置信系数。

下面，我们用 Clearview 国家银行的一次抽样研究来解释区间估计方法。Clearview 两家支行的客户支票账户余额的独立随机样本的资料如下：

支行	支票账户数	样本均值余额	样本标准差
Cherry Grove	12	$\bar{x}_1 = 1\,000$ 美元	$s_1 = 150$ 美元
Beechmont	10	$\bar{x}_2 = 920$ 美元	$s_2 = 120$ 美元

我们利用这些数据来推出两家支行平均支票账户余额之差的 90% 置信区间。假定两个支行的支票账户余额都服从正态分布，并且方差相等。利用公式(10.8)，得到总体方差的合并估计值为

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)(150)^2 + (9)(120)^2}{12 + 10 - 2} = 18\,855$$

利用公式 (10.9)，$\sigma_{\bar{x}_1 - \bar{x}_2}$ 的相应估计值为

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{18\,855\left(\frac{1}{12} + \frac{1}{10}\right)} = 58.79$$

区间估计方法中 t 分布的自由度为 $n_1 + n_2 - 2 = 12 + 10 - 2 = 20$。对于 $\alpha = 0.10$，$t_{\alpha/2} = t_{0.05} = 1.725$。因此，利用公式(10.10)，可以得出区间估计为

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.05}s_{\bar{x}_1-\bar{x}_2}$$

$$1\,000 - 920 \pm (1.725)(58.79)$$

$$80 \pm 101.41$$

在 90% 置信水平下，两家支行的平均账户余额之差的区间估计为 -21.41 美元至 181.41 美元之间。该区间有负值意味着两个均值之差的实际值 $\mu_1 - \mu_2$ 可能为负。因此，μ_2 可能大于 μ_1，这意味着，尽管抽样调查资料表明，Cherry Grove 支行有较大的样本均值余额，Beechmont 支行的总体均值反而可能更大。置信区间含 0 的事实可解释为：我们没有足够的证据得出两家支行总体平均账户余额有差别的结论。

评　注

1. 本节介绍的 t 分布在小样本方法中的使用是基于两个总体都服从正态概率分布且 $\sigma_1^2 = \sigma_2^2$ 的假设。这种方法是稳健的统计方法，就是说它对上述假设相对并不敏感。例如，当 $\sigma_1^2 \neq \sigma_2^2$ 时，该方法在 n_1、n_2 比较接近时也能得到比较令人满意的结果。
2. t 分布并不局限于小样本情形，只要两个总体分布都服从正态分布且总体方差相等，它都是可行的。因而，公式(10.4)和(10.6)说明了当样本容量较大时，如何确定两总体均值之差的区间估计。因此，在大样本情形下，没必要使用 t 分布并作相应假设。所以，除了小样本外，我们不使用 t 分布。

练　习

方　法

1. 考虑来自两个总体的两个独立随机样本的如下资料：



样本 1	样本 2
$n_1 = 50$	$n_2 = 35$
$\bar{x}_1 = 13.6$	$\bar{x}_2 = 11.6$
$s_1 = 2.2$	$s_2 = 3.0$

a. 两个总体均值之差的点估计量是多少？

b. 求两个总体均值之差的 90% 置信区间。

c. 求两个总体均值之差的 95% 置信区间。

2. 考虑两个总体的两个独立随机样本的如下数据：

样本 1	样本 2
$n_1 = 10$	$n_2 = 8$
$\bar{x}_1 = 22.5$	$\bar{x}_2 = 20.1$
$s_1 = 2.5$	$s_2 = 2.0$

a. 两个总体均值之差的点估计量是多少?

b. 总体方差的合并估计量是多少?

c. 求两个总体均值之差的 95% 置信区间。

3. 考虑两个总体的两个独立随机样本的如下数据:

样本 1		样本 2	
10	7	8	7
12	7	8	4
9	9	6	9

a. 计算两个样本均值。

b. 计算两个样本的标准差。

c. 两个总体均值之差的点估计量是多少?

d. 总体方差的合并估计量是多少?

e. 求两个总体均值之差的 95% 置信区间。

应 用

4. 从 1999 年到 2000 年，石油价格有了增长。美国汽车协会提供了关于两年自助式销售的汽油每加仑平均价格的资料(*AAA Going Places*, May/June, 2000)。假设以下资料来自全国的独立样本:

2000 年的价格	1999 年的价格
$\bar{x}_1 = 1.58$ 美元	$\bar{x}_2 = 0.98$ 美元
$s_1 = 0.12$ 美元	$s_2 = 0.08$ 美元
$n_1 = 50$	$n_2 = 42$

a. 1999 年到 2000 年中，每加仑的平均价格上升值的点估计量是多少?

b. 求 1999 年到 2000 年中，每加仑的平均价格上升值的 95% 置信区间估计是多少?

5. 美国交通部收集的数据列举了 75 个最大城市的居民每天每辆汽车的行驶里程数。假定 50 个布法罗居民构成的样本平均值为 22.5 英里/天,标准差为 8.4 英里/天的简单随机样本。另外,一个与之独立的 100 位波士顿居民构成的平均值为 18.6 英里/天,标准差为 7.4 英里/天的简单随机样本。

a. 布法罗和波士顿居民平均每天行驶里程数之差的点估计量是多少?

b. 求两个总体均值之差的 95% 置信区间。

6. 国际航空运输协会对一些公务旅行者进行了调查以确定不同的国际机场的受欢迎程度。满分为10分。假设，有50名公务旅行者给迈阿密机场打分，另外有50名公务旅行者给洛杉矶机场打分。得分如下：


CD光盘数据
Airport

迈阿密

6	4	6	8	7	7	6	3	3	8	10	4	8
7	8	7	5	9	5	8	4	3	8	5	5	4
4	4	8	4	5	6	2	5	9	9	8	4	8
9	9	5	9	7	8	3	10	8	9	6		

洛杉矶

10	9	6	7	8	7	9	8	10	7	6	5	7
3	5	6	8	7	10	8	4	7	8	6	9	9
5	3	1	8	9	6	8	5	4	6	10	9	8
3	2	7	9	5	3	10	3	5	10	8		

求迈阿密和洛杉矶机场平均的分之差的95%置信区间。

7. 科隆大学对男女工资差别的研究指出，男士工资高于女士工资的原因之一是男士比女士有更多年的工作经验(*Business Week*, August 28, 2000)。假设如下的摘要样本表现了每组的工作年数：

男士	女士
$\bar{x}_1 = 14.9$ 年	$\bar{x}_2 = 10.3$ 年
$s_1 = 5.2$ 年	$s_2 = 3.8$ 年
$n_1 = 100$	$n_2 = 85$

a. 求两个总体均值之差的点估计量。

b. 在95%置信度下，误差范围有多大？

c. 求两个总体均值之差的95%置信区间估计。

8. 某城市规划小组想估算某大城市相邻两个地区平均收入之差，这两个地区家庭的独立随机样本的资料如下：


自测题

地区1	地区2
$n_1 = 8$	$n_2 = 12$
$\bar{x}_1 = 15\,700$ 美元	$\bar{x}_2 = 14\,500$ 美元
$s_1 = 700$ 美元	$s_2 = 850$ 美元

a. 对两个相邻地区平均收入之差进行点估计。

b. 对两个相邻地区平均收入之差进行95%置信区间估计。

c. 在(b)中进行区间估计需作什么假设？

9. 国家房屋建设者协会提供了有关最流行的家庭装修业务的数据(*USA Today*, June 17, 1997)。有关两种装修业务以千美元为单位计算的开支的样本数据如下：

厨房	主卧室
25.2	18.0
17.4	22.9
22.8	26.4
21.9	24.8
19.7	26.9
23.0	17.8
19.7	24.6
16.9	21.0
21.8	
23.6	

a. 求两种装修业务总体平均装修开支之差的点估计量。

b. 求两个总体均值之差的90%的置信区间。

10. 假设抽取15名女性工会会员和20名女性非工会会员的独立随机样本，以下是她们每小时的工资数：


CD光盘数据
Union

工会会员

22.40	18.90	16.70	14.05	16.20	20.00	16.10	16.30	19.10
16.50	18.50	19.80	17.00	14.30	17.20			

非工会会员

17.60	14.40	16.60	15.00	17.65	15.00	17.55	13.30	11.20
15.90	19.20	11.85	16.65	15.20	15.30	17.00	15.10	14.30
13.90	14.50							

a. 求两个总体小时工资数之差的点估计量。

b. 求总体方差的合并估计量。

c. 求两个总体均值之差的95%置信区间。

d. 这两个总体平均工资有差异吗？试解释。

10.2 两个总体均值之差的假设检验：独立样本

在这一节中，我们将介绍检验有关两个总体均值之差的假设方法。该方法仍分为大样本（$n_1 \geqslant 30$，$n_2 \geqslant 30$）和小样本（$n_1 < 30$/且/或 $n_2 < 30$）两种情形。

大样本情形

对两个培训中心教育质量差异进行研究，对两个中心学员进行一次标准化考试。考试分数是评价两个中心的教学质量差异的主要因素。两个中心的平均值如下：

μ_1 = 中心A学员总体的平均分

μ_2 = 中心B学员总体的平均分

我们从一个尝试性的假设开始讨论，假设两个中心的培训质量没有差异。于是，从平均分的角度来讲，原假设是 $\mu_1 - \mu_2 = 0$。如果抽样证据导致拒绝该假设，我们就能得出两个总体平均分有差异的结论。该结论表明，两个中心的教学质量有差异，继而，有可能批准某项研究以调查产生这种差异的原因。原假设及备择假设如下：

$$H_0: \ \mu_1 - \mu_2 = 0$$

$$H_a: \ \mu_1 - \mu_2 \neq 0$$

按照第 9 章的假设检验方法，我们认为尝试性假设 H_0 为真。利用样本均值之差作为总体均值之差的点估计量，并且考虑 H_0 为真时的 $\bar{x}_1 - \bar{x}_2$ 的抽样分布。对于大样本情形，该分布如图 10.4 所示，由于 $\bar{x}_1 - \bar{x}_2$ 的抽样分布可以由正态分布来逼近，因此，可以用下面的检验统计量：

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{10.11}$$

若 σ_1、σ_2 未知，我们可用抽样标准差 s_1、s_2 来计算检验统计量。

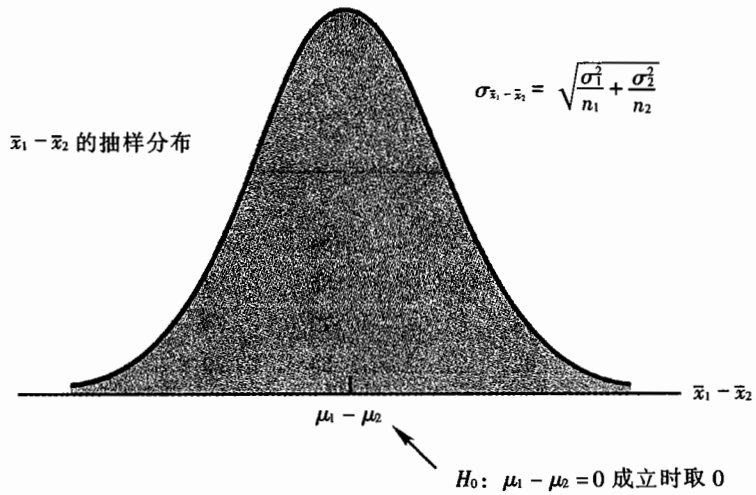


图 10.4　$H_0: \mu_1 - \mu_2 = 0$ 时的抽样分布

公式(10.11)给出的 z 值可解释为 $\bar{x}_1 - \bar{x}_2$ 与 H_0 中 $\mu_1 - \mu_2$ 的标准化偏差。对于 $\alpha = 0.05$，此时 $z_{\alpha/2} = z_{0.025} = 1.96$，双侧检验的拒绝域如图 10.5 所示。利用检验统计量，拒绝规则为

如果 $z < -1.96$ 或 $z > +1.96$，则拒绝 H_0

假设两个中心学员的独立随机样本所得考分由表 10.1 给出，汇总统计由表 10.2 给出。利用 s_1、s_2 估计 σ_1、σ_2，我们求得关于原假设 $H_0: \ \mu_1 - \mu_2 = 0$ 的假设统计量 z 为

$$z = \frac{(82.5 - 78) - 0}{\sqrt{(8)^2/30 + (10)^2/40}} = 2.09$$

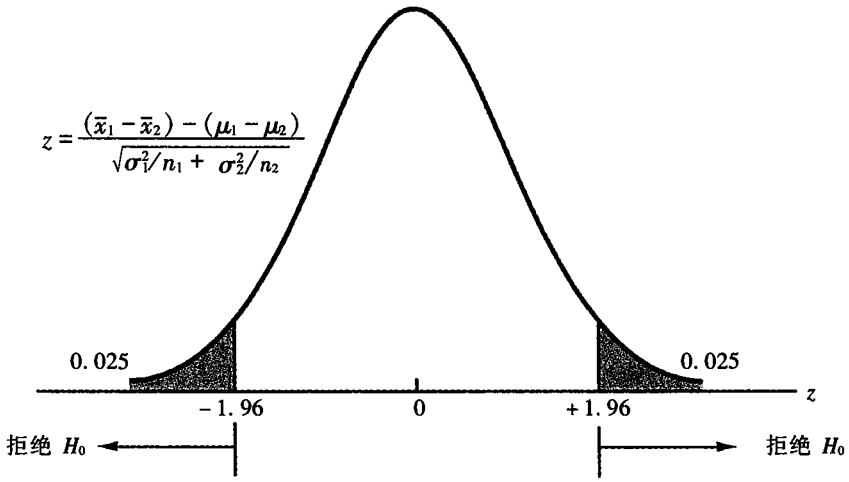


图 10.5　$\alpha=0.05$ 时双侧假设检验的拒绝域

因为 $z=2.09>1.96$，我们得出拒绝 H_0 的结论。于是，我们推论 μ_1、μ_2 不相等，即两个中心的教学质量有差异。

表 10.1　考试分数



CD 光盘数据
ExamData

培训中心 A			培训中心 B			
97	83	91	64	66	91	84
90	84	87	85	83	78	85
94	76	73	72	74	87	85
79	82	92	64	70	93	84
78	85	64	74	82	89	59
87	85	74	93	82	79	62
83	91	88	70	75	84	91
89	72	88	79	78	65	83
76	86	74	79	99	78	80
84	70	73	75	57	66	76

表 10.2　考试分数汇总结果

培训中心 A	培训中心 B
$n_1=30$	$n_2=40$
$\bar{x}_1=82.5$	$\bar{x}_2=78$
$s_1=8$	$s_2=10$

对于 $z=2.09$，可用标准正态概率分布表计算该双侧检验的 p－值。因为均值与 $z=2.09$ 之

间的面积是 0.481 7，p - 值为 $2\times(0.500\,0-0.481\,7)=0.036\,6$，因而该 p - 值小于 $\alpha=0.05$，由 p - 值法也能得出拒绝 H_0 结论。

---

p - 值也可用于假设检验。

在这个假设检验中，我们感兴趣的是确定两个总体均值是否有差异。由于我们事先不知道哪个均值更大或更小，故采用假设 H_0：$\mu_1-\mu_2=0$ 与 H_a：$\mu_1-\mu_2\neq0$。在有关两个总体均值之差的其他假设检验中，我们可能找出一个均值比另一个均值是大还是小？在这些情形下，单侧假设会更适合。有关两个总体均值之差的单侧检验的两种形式如下：

$$H_0:\ \mu_1-\mu_2\leqslant 0 \qquad H_0:\ \mu_1-\mu_2\geqslant 0$$
$$H_a:\ \mu_1-\mu_2>0 \qquad H_a:\ \mu_1-\mu_2<0$$

这些假设可用公式(10.11)给出的检验统计量 z 来检验。拒绝域可以仿造第 9 章介绍的单侧检验得出。

小样本情形

我们现在考虑，在小样本情形下，有关两个总体均值之差的假设检验，即 $n_1<30$ 且/或 $n_2<30$。我们使用的方法是基于自由度为 n_1+n_2-2 的 t 分布，就如在 10.1 节中讨论的那样，假设两个总体都服从正态分布且总体方差相等。

我们用一个例子来说明小样本情形，该例子涉及到一个计算机软件包，它是为协助系统分析员节约对信息系统设计、开发所需时间而开发的。为评价新软件包的优点，我们抽取由 24 名系统分析员组成的随机样本。给每位系统分析员一张信息系统的说明书。指定其中 12 名分析员使用当前技术开发的信息系统，另外 12 名分析员先接受新软件包的培训，然后用新软件包来开发该信息系统。

在本研究中有两个总体：使用当前技术的系统分析员总体和使用新软件包的系统分析员总体。从完成该信息系统设计项目所需时间的角度来看，总体均值如下：

μ_1 = 使用当前技术的系统分析员平均项目完成时间
μ_2 = 使用新软件包的系统分析员平均项目完成时间

新软件项目的研究人员希望证明新软件包使的平均项目完成时间缩短。因此，研究人员要寻找证据来推断 μ_2 小于 μ_1。此时，两个总体均值之差 $\mu_1-\mu_2>0$。将研究假设 $\mu_1-\mu_2>0$ 作为备择假设提出：

$$H_0:\ \mu_1-\mu_2\leqslant 0$$
$$H_a:\ \mu_1-\mu_2>0$$

研究人员要寻找到拒绝 H_0 的证据，才能得出新软件包能使平均完成时间缩短的结论。

假设 24 名分析员完成工作后得到表 10.3 所示资料。在总体方差相等的假设下，用公式（10.8）来计算的 σ^2 合并估计值

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=\frac{11\times(40)^2+11\times(44)^2}{12+12-2}=1\ 768$$

小样本情形的检验统计量为

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{10.12}$$

在两独立随机样本容量为 n_1、n_2 时，t 分布的自由度为 n_1+n_2-2。对于 $\alpha=0.05$，由 t 分布表可知自由度 $12+12-2=22$，$t_{0.05}=1.717$。因而，利用检验统计量值，单侧检验拒绝域为

若 $t>1.717$，则拒绝 H_0

表 10.3 软件包检验研究中的完成时间数据与汇总统计



CD 光盘数据
Software

	当前技术	新软件包
	300	276
	280	222
	344	310
	385	338
	372	200
	360	302
	288	317
	321	260
	376	320
	290	312
	301	334
	283	265
汇总统计		
样本容量	$n_1=12$	$n_2=12$
样本均值	$\bar{x}_1=325$	$\bar{x}_2=288$
样本标准差	$s_1=40$	$s_2=44$

由样本数据及公式(10.12)可知检验统计量的值如下：

$$t=\frac{(325-288)-0}{\sqrt{1\ 768\left(\frac{1}{12}+\frac{1}{12}\right)}}=2.16$$

观察拒绝域，我们可以看到当 $t=2.16$，显著性水平为 0.05 时，拒绝 H_0。因此，抽样资料使研

究者得出 $\mu_1-\mu_2>0$ 的结论，也就是新软件包使的项目平均完成时间缩短。

我们可用 Minitab 来检验关于两个总体均值之差的检验。比较当前技术和新软件包技术的输出结果如图 10.6 所示。输出结果的最后一行显示出 $t=2.16$ 和 $p-$值 $=0.021$。对于 $\alpha=0.05$，原假设被拒绝。因此，新软件包的平均完成时间比当前技术要短。

```
Two sample T for Current vs New
              N       Mean      StDev      SE Mean
Current      12      325.0       40.0           12
New          12      288.0       44.0           13
Difference = mu Current - mu New
Estimate for difference:  37.0
T-Test of difference = 0 (vs >): T = 2.16  P-value = 0.021  DF = 22
Both use Pooled StDev = 42.0
```

图 10.6　Minitab 关于当前技术和新软件技术的假设检验输出结果

评　注

1. 我们在上一节中指出，对于正态总体及等方差的假设，利用 t 分布进行关于两个总体均值的推断是不太敏感的。但是，如果读者认为在某些特殊应用场合，这些假设不合适，可采取下列方法之一：
   a. 考虑用第 19 章介绍的非参数维尔克科森秩检验。
   b. 如果是方差不相等($\sigma_1^2\neq\sigma_2^2$)的总体近似正态分布，利用公式（10.5）估算 $\sigma_{\bar{x}_1-\bar{x}_2}$，仍可用 t 分布，其自由度由下式给出：

$$\frac{\left[1/n_1+(s_2^2/s_1^2)/n_2\right]^2}{\left[\frac{1}{n_1^2(n_1-1)}\right]+\left[\frac{(s_2^2/s_1^2)^2}{n_2^2(n_2-1)}\right]}$$

   c. 把两个样本容量增加到 $n_1\geq30$ 且 $n_2\geq30$ 的大样本情形。

2. 在有关两个总体均值之差的假设检验中，原假设中几乎总有均值是无差异的情况。这样，下列原假设可供选择：

$$H_0:\ \mu_1-\mu_2=0$$
$$H_0:\ \mu_1-\mu_2\leq0$$
$$H_0:\ \mu_1-\mu_2\geq0$$

在某些例子中，我们可能要确定总体均值之间是否存在非零差距 D_0。特定值 D_0 的选择依赖所研究的应用情况。然而，在这种情形，原假设可以下列形式提出：

$$H_0:\ \mu_1-\mu_2=D_0$$
$$H_0:\ \mu_1-\mu_2\leq D_0$$
$$H_0:\ \mu_1-\mu_2\geq D_0$$

除了公式(10.11)和公式(10.12)中的 $\mu_1-\mu_2$ 的计算用 D_0 之外，其余计算完全相同。

练　习

方　法

11. 考虑下面的假设检验：



$$H_0: \mu_1 - \mu_2 \leqslant 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

下面是取自两个总体的两个独立样本资料。

样本 1	样本 2
$n_1 = 40$	$n_2 = 50$
$\bar{x}_1 = 25.2$	$\bar{x}_2 = 22.8$
$s_1 = 5.2$	$s_2 = 6.0$

a. 取 $\alpha = 0.05$，进行假设检验并给出结论。

b. p - 值是多少？

12. 考虑下面的假设检验：

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

下面是取自两个总体的两个独立样本资料：

样本 1	样本 2
$n_1 = 80$	$n_2 = 70$
$\bar{x}_1 = 104$	$\bar{x}_2 = 106$
$s_1 = 8.4$	$s_2 = 7.6$

a. 取 $\alpha = 0.05$，进行假设检验并给出结论。

b. p - 值是多少？

13. 考虑下面的假设检验：

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

下面是取自两个总体的两个独立样本资料。取 $\alpha = 0.05$，进行假设检验并给出结论。

样本 1	样本 2
$n_1 = 8$	$n_2 = 7$
$\bar{x}_1 = 1.4$	$\bar{x}_2 = 1.0$
$s_1 = 0.4$	$s_2 = 0.6$

应　用

14. 美国沿海地区如 Cape Cod，the Outer Banks，the Carolinas 以及 the Gulf Coast，自 20 世纪 90 年代以来有很高的人口增长率。对全国范围内居住在沿海地区和非沿海地区的居民进行收集整理数据（*USA Today*，July 21, 2000），假设下列样本取自两个总体中的个人的年龄：

沿海地区	非沿海地区
$\bar{x}_1 = 39.3$ 岁	$\bar{x}_2 = 35.4$ 岁
$s_1 = 16.8$ 岁	$s_2 = 15.2$ 岁
$n_1 = 150$	$n_2 = 175$

对这两个总体的平均年龄无差别进行检验。利用 $\alpha = 0.05$。

a. 用公式表示原假设和备择假设。

b. 拒绝规则是什么？

c. 检验统计量值是多少？

d. 得出你的结论。

e. p - 值是多少？

15. 在 10.1 节中对 Greystone 百货公司的研究，有两个商场地区独立随机样本的顾客年龄数据如下：

自测题

市中心商店	郊区商店
$n_1 = 36$	$n_2 = 49$
$\bar{x}_1 = 40$ 岁	$\bar{x}_2 = 35$ 岁
$s_1 = 9$ 岁	$s_2 = 10$ 岁

对于 $\alpha = 0.05$，检验 H_0：$\mu_1 - \mu_2 = 0$ 和备择假设 H_a：$\mu_1 - \mu_2 \neq 0$。你对这两个商店顾客总体的平均年龄有何结论？p - 值是多少？

16. 教育考试中心进行了一次有关学术潜质测试中男女考生差异的研究。研究选择了在数学考试中取得同样分数的 562 个女生及 852 个男生。也就是说，可以认为这些考生具有同样的数学能力。两个样本的 SAT 词汇部分的得分如下：

女生	男生
$\bar{x}_1 = 547$	$\bar{x}_2 = 525$
$s_1 = 83$	$s_2 = 78$

这些数据是支持这样一个结论——给定具有同样数学能力的一个女考生总体和一个男考生总体，女考生的词汇能力明显高于男生？在显著性水平为 0.02 下，进行假设检验。得出你的结论？

17. 某公司研究两家原材料供应商的送货时间。该公司对供应商 A 基本满意，如果其平均送货时间等于或小于供应商 B，则该公司将继续选择供应商 A。但若公司发现 B 的平均送货时间小于 A，则从 B 处购买原材料。

a. 这种情况下，原假设和备择假设分别是什么？

b. 假设独立样本给出了两个供应商的送货时间，数据如下：

供应商 A	供应商 B
$n_1 = 50$	$n_2 = 30$
$\bar{x}_1 = 14$ 天	$\bar{x}_2 = 12.5$ 天
$s_1 = 3$ 天	$s_2 = 2$ 天

若 $\alpha = 0.05$，你对（a）中的假设有何结论？你对选择供应商的建议如何？

18. Arnold Palmer 和 Tiger Woods 是本项运动中的两位最出色的打高尔夫球的人。如果两个都在最佳状态下，他们如何相比？下面的样本数据给出了在 PGA 锦标赛中 18 洞的比赛结果。Palmer 的分数取自 1960 年，而 Woods 的分数取自 1999 年（*Golf Magazine*，February 2000）。

Palmer，1960 年	Woods，1999 年
$\bar{x}_1 = 69.95$	$\bar{x}_2 = 69.56$
$n_1 = 112$	$n_2 = 84$

利用样本结果对两人 18 洞的平均分数总体无差别假设检验。

a. 假设两人 18 洞的平均分数都有 2.5 的总体标准差，检验统计量值是多少？

b. p - 值是多少？

c. 取 $\alpha = 0.01$，你的结论是什么？

19. 一些航空公司的互联网上报道，飞机着陆平均时间通常为 2—3 分钟（*USA Today*, September 11, 2000）。三角航空公司和西北航空公司飞机平均着陆时间的样本数据如下：

三角	西北
$\bar{x}_1 = 2.5$ 分	$\bar{x}_2 = 2.1$ 分
$s_1 = 0.8$ 分	$s_2 = 1.1$ 分
$n_1 = 22$	$n_2 = 20$

a. 若检验目的是检验一个相当两航空公司的飞机平均着陆时间有差异，请你建立假设。

b. 若 $\alpha = 0.05$，你的拒绝规则是什么？

c. 计算检验统计量值。

d. 得出你的结论？

e. 你对 p - 值有何评价？

20. 美林公司定期地邀请顾客来评估其财务顾问的服务(2000 美林 Client Satisfaction Survey)。在顾客满意度调查中，级别高表明服务越高。两个财务顾问的服务级别独立样本数据如下。财务顾问 A 有 10 年的经验，而财务顾问 B 只有 1 年的经验。若 $\alpha=0.05$，是否工作经验多的顾问会提供更高的服务级别?

顾问 A	顾问 B
$\bar{x}_1=6.82$	$\bar{x}_2=6.25$
$s_1=0.64$	$s_2=0.75$
$n_1=16$	$n_2=10$

a. 建立原假设和备择假设。
b. 你的拒绝规则是什么?
c. 计算检验统计量值。
d. 得出你的结论?
e. 你对 p - 值有何评价?

10.3　两个总体均值之差的统计推断：匹配样本

假设一家生产公司有两种方案供工人完成生产任务。为了使产出最大化，公司想确认哪种方法有最短完成时间。以 μ_1 记生产方法 1 的平均完成时间，μ_2 记生产方法 2 的平均完成时间。因为没有哪种生产方法更好的先验，我们从两种生产方法具有相同平均完成时间的尝试性假设开始。因此，原假设为 H_0：$\mu_1-\mu_2=0$，若该假设被拒绝，我们可得出平均完成时间不同。这种情形下，应采用那个完成时间较短的方案。原假设与备择假设如下：

$$H_0:\ \mu_1-\mu_2=0$$
$$H_a:\ \mu_1-\mu_2\neq 0$$

在进行收集生产时间数据以及假设检验的抽样步骤时，我们考虑使用两种可供选择的方案。一种是基于**独立样本**(independent samples)的，另一种是基于**匹配样本**(matched samples)的。

1. **独立样本方案**：抽取工人的一个简单随机样本，其中每个工人使用方法 1；抽取工人的另一个简单随机样本，其中每个工人使用方法 2。均值差的检验用 10.2 节介绍的方法。
2. **匹配样本方案**：抽取工人的一个简单随机样本，每个工人先使用一种方法，后用另一种方法。两种方法的次序是随机排列的，某些工人先执行方法 1，其他工人先执行方法 2。每个工人提供一组数据，一个是方法 1 的，另一个是方法 2 的。

在匹配样本方案中，两种生产方法是在相同条件下接受检验的(即由同一工人执行)；所以，该方案会比独立样本方案产生更小的误差。这主要是因为在匹配样本方案中，作为抽样误差来源之一的工人个体之间没有差别。

假设在匹配样本方案中，检验两种方法之间的差异，我们对这种抽样方案来进行分析。有

一个由 6 个工人组成的一个随机样本，6 个工人完成时间的数据由表 10.4 给出。注意每个工人提供一对有效数据，每种方法一个，最后一列为样本中每个工人完成时间之差 d_i。

样本均值和样本标准差的公式中的下标 d 是用于提示该数据体现了匹配样本数据值之差的差值。除此之外，样本均值与标准差的公式与上文中的一致。

分配匹配样本方案的关键在于，我们仅考虑差值这一列。因此，我们由 6 个数据值(0.6，-0.2，0.5，0.3，0.0 及 0.6)用来分析两生产方法均值之差。

令 μ_d = 两工人总体差值的平均值。原假设和备择假设可写为

$$H_0: \mu_d = 0$$
$$H_a: \mu_d \neq 0$$

若 H_0 被拒绝，我们可得出平均完成时间有差异。

下标 d 是提示匹配样本提供差值数据。表 10.4 中，6 个差值的样本均值与样本标准差如下：

$$\bar{d} = \frac{\sum d_i}{n} = \frac{1.8}{6} = 0.30$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{0.56}{5}} = 0.335$$

表 10.4　匹配样本方案的任务完成时间



CD 光盘数据
Matched

工人	方法 1 的完成时间（分钟）	方法 2 的完成时间（分钟）	完成时间的差值（d_i）
1	6.0	5.4	0.6
2	5.0	5.2	-0.2
3	7.0	6.5	0.5
4	6.2	5.9	0.3
5	6.0	6.0	0.0
6	6.4	5.8	0.6

在第 9 章中，我们曾指出，如果假定总体分布服从正态分布，则可使用自由度为 $n-1$ 的 t 分布来检验有关总体均值的原假设。假定表示差值的数据服从正态分布，检验统计量为

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \tag{10.13}$$

若 $\alpha = 0.05$，自由度为 $n-1=5$（$t_{0.025} = 2.571$），那么双侧检验拒绝规则为

若 $t < -2.571$ 或 $t > 2.571$，则拒绝 H_0

对于 $\bar{d} = 0.30$，$s_d = 0.335$ 和 $n = 6$，原假设检验统计量的值为

若 H_0：$\mu_d = 0$，

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0.30 - 0}{0.335/\sqrt{6}} = 2.20$$

所以 $t = 2.20$ 不在拒绝域中，样本数据没提供足够数据拒绝 H_0。

自由度为 5，t 分布表中指出了 $t = 2.20$ 是介于 $t_{0.025} = 2.571$ 和 $t_{0.025} = 2.015$ 之间。因此，与 $t = 2.20$ 相关的单侧域在 0.025 到 0.05 之间。对于双侧检验，我们知道 p - 值在 0.05 和 0.10 之间。因为此值大于 $\alpha = 0.05$，H_0 没被拒绝。利用 Minitab 和表 1014 中的数据表示 p - 值 = 0.08。

计算出差值数据后，匹配样本中的 t 分布方法与第 8 章及第 9 章中介绍的一个总体情形相同。

此外，我们可通过第 8 章中单个总量的方法得到两个总体均值差的区间估计。计算如下：

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

$$0.3 \pm 2.571\left(\frac{0.335}{\sqrt{6}}\right)$$

$$0.3 \pm 0.35$$

因此，两种生产方法均值差的 95% 置信区间估计为 -0.05 到 0.65 分钟。注意到置信区间含有零值，所以样本数据没有提供足够证据拒绝 H_0。

评　注

1. 在本节例子中，工人先用一种方法再用另一种方法来完成任务。本章介绍了一个匹配样本方案，其中每个样本个体提供了一对数据。也可用不同但“相似”的样本个体来提供一对数据值。例如，一个地方的工人可与另一个地方的工人相匹配（其年龄、受教育情况、性别、经验等方面相似）。每对工人可能提供用于匹配样本分析的不同数据。
2. 对于两个总体均值的推断，匹配样本方法通常对独立样本方法更精确，因此，前者便被推荐者使用。然而在某些应用中可能无法匹配，或者，可能时间和费用成本太高。这时，应采用独立样本方案。
3. 本章中的例子样本容量为 6 个工人，属于小样本情形。假设检验和区间估计中都采用了 t 分布。若样本容量较大（$n \geq 30$），就不必要使用 t 分布。这时，统计推断可基于标准正态概率分布的 z 值。

练　习

方　法

21. 考虑下面的假设检验：



$$H_0: \mu_d \leqslant 0$$

$$H_a: \mu_d > 0$$

下面是来自两个总体的匹配样本数据：

	总体	
个体	1	2
1	21	20
2	28	26
3	18	18
4	20	20
5	26	24

a. 计算每对个体的差值。

b. 计算$\bar{d}$。

c. 计算标准差 s_d。

d. 取 $\alpha = 0.05$ 进行假设检验，得出你的结论？

22. 下面是取自两个总体的匹配样本数据：

	总体	
个体	1	2
1	11	8
2	7	8
3	9	6
4	12	7
5	13	10
6	15	15
7	15	14

a. 计算每对个体的差值。

b. 计算$\bar{d}$。

c. 计算标准差 s_d。

d. 对两总体均值之差进行假设检验，得出你的结论。

e. 给出两个总体均值之差的 95% 置信区间。

应　用

23. 自测题

一个市场研究的机构通过一组被调查者的样本来给某特定商品的潜在购买力打分。在样本中，每个人都在看了该产品的电视广告之前和之后分别打分。潜在购买力的分值为 0—10 分，分值越高表示潜在购买力越高。原假设为："看后"平均得分小于或等于"看前"得分。拒绝该假设表明广告提高了平均潜在购买力得分。对于 $\alpha = 0.05$，用下列数据进行假设检验，并对广告给予评价。

	购买力得分			购买力得分	
个人	之后	之前	个人	之后	之前
1	6	5	5	3	5
2	6	4	6	9	8
3	7	7	7	7	5
4	4	3	8	6	6

24. 下面是 10 个国际长途的一个样本，给出了来自美国的 the Sprint and WorldCom 每分钟的电话费（*World Traveler*, July 2000）。

国家或地区	Sprint	Worldcom
澳大利亚	0.46	0.26
比利时	0.69	0.40
巴西	0.92	0.53
哥伦比亚	0.55	0.53
丹麦	0.50	0.26
法国	0.46	0.26
德国	0.46	0.26
香港	0.92	0.40
日本	0.69	0.40
英国	0.46	0.26

得出两个总体均值之差的 95% 置信区间估计。

25. 从机场到市中心的交通费用取决于交通工具。下面给出了 10 个大城市的出租车和班车的单程费用。求因坐出租车而引起的平均费用上升的 95% 置信区间。

城市	出租车（美元）	班车（美元）	城市	出租车（美元）	班车（美元）
亚特兰大	15.00	7.00	明尼阿波利斯	16.50	7.50
芝加哥	22.00	12.50	新奥尔良	18.00	7.00

（续表）

城市	出租车（美元）	班车（美元）	城市	出租车（美元）	班车（美元）
丹佛	11.00	5.00	纽约（LaGuardia）	16.00	8.50
休斯敦	15.00	4.50	费城	20.00	8.00
洛杉矶	26.00	11.00	华盛顿	10.00	5.00

26. 对于8个主要机场选取的出租车每加仑的汽油价格，Hertz公司和国家出租车提供的数据如下（*USA Today* April 4, 2000）：

机场	公司	国家公司
波士顿(Logan)	1.55	1.56
芝加哥(O'Hare)	1.62	1.59
洛杉矶	1.72	1.78
迈阿密	1.65	1.49
纽约（JFK）	1.72	1.51
纽约（LaGuardia）	1.67	1.50
奥兰治县	1.68	1.77
华盛顿（Dulles）	1.52	1.41

若 $\alpha=0.05$，检验两公司每加仑油价的总体均值相等的假设。

27. Book-of-the-Month-Club对其会员做了一次调查，用于确定其会员花在看电视上的时间是否比花在看书上时间多。假设15个回答者组成的样本给出了以下关于每周看电视和读书的小时数的数据。对显著性水平0.05，你能得出的该俱乐部会员用于看电视的时间比看书时间多的结论吗？



CD光盘数据
TVRead

回答者	看电视	读书	回答者	看电视	读书
1	10	6	9	4	7
2	14	16	10	8	8
3	16	8	11	16	5
4	18	10	12	5	10
5	15	10	13	8	3
6	14	8	14	19	10
7	10	14	15	11	6
8	12	14			

28. 下面是有关12种邮票组成的样本的1997年市盈率和估计的1998年市盈率的资料（*Kiplinger's Personal Finance Magazine*, November 1997）：



CD光盘数据
PERatio

股票	1997年市盈率	估计的1998年市盈率
可口可乐	40	32
沃特·迪斯尼	33	23

（续表）

股票	1997 年市盈率	估计的 1998 年市盈率
杜邦	24	16
柯达	21	13
通用电气	30	23
通用汽车	25	19
IBM	19	14
默克制药	29	21
麦当劳	20	17
摩托罗拉	35	20
飞利浦	17	13
惠普	20	17

a. 对 $\alpha=0.05$，检验平均市盈率在 2 年中是否有变化？

b. 给出总体平均市盈率在 2 年内变化的 95% 置信区间。

29. 一厂家生产豪华型和普通型的家用自动磨砂机，零售商提供的销售价格的样本如下：

	价格（美元）			价格（美元）	
零售商	豪华型	标准型	零售商	豪华型	标准型
1	39	27	5	40	30
2	39	28	6	39	34
3	45	35	7	35	29
4	38	30			

a. 厂商建议的两种型号的零售价有 10 美元的差价，利用显著性水平为 0.05，检验两种型号的平均差价是 10 美元。

b. 两种型号平均差价的 95% 置信区间是多少？

10.4　两个总体比例之差的统计推断

一税收机构想比较两家地区办事处的工作质量，随机抽取每个办事处的纳税申报单作为样本，并且想确认其中哪些是正确的，该机构可以通过估计每个办事处错误申报的比例，计算两者比例之差。

p_1 = 总体 1（办事处 1）的错误申报单的比例。

p_2 = 总体 2（办事处 2）的错误申报单的比例。

$\bar{p}_1$ = 来自总体 1 的简单随机样本的样本比例。

$\bar{p}_2$ = 来自总体 2 的简单随机样本的样本比例。

p_1-p_2 给出两个总体比例的差值，p_1-p_2 的点估计量如下：

两个总体比例差的点估计量

$$\bar{p}_1 - \bar{p}_2 \tag{10.14}$$

因此，两个总体比例之差的点估计量为两独立简单随机样本的样本比例之差。

$\bar{p}_1 - \bar{p}_2$ 的抽样分布

在两个总体比例差的研究中，$\bar{p}_1 - \bar{p}_2$ 是我们感兴趣的点估计量。我们在以往的例子中看到，点估计量的抽样分布是进行的待估参数区间估计和假设检验的关键因素。$\bar{p}_1 - \bar{p}_2$ 抽样分布的性质如下：

$\bar{p}_1 - \bar{p}_2$ 的抽样分布

期望值：
$$E(\bar{p}_1 - \bar{p}_2) = p_1 - p_2 \tag{10.15}$$

标准差：
$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{10.16}$$

式中 n_1——来自总体 1 的简单随机样本的样本容量；

n_2——来自总体 2 的简单随机样本的样本容量。

分布形式：若样本容量较大即 n_1p_1、$n_1(1-p_1)$、n_2p_2 和 $n_2(1-p_2)$都大于或等于 5，$\bar{p}_1 - \bar{p}_2$ 的抽样分布可由正态概率分布来逼近。

$\bar{p}_1 - \bar{p}_2$ 的抽样分布如图 10.7 所示。

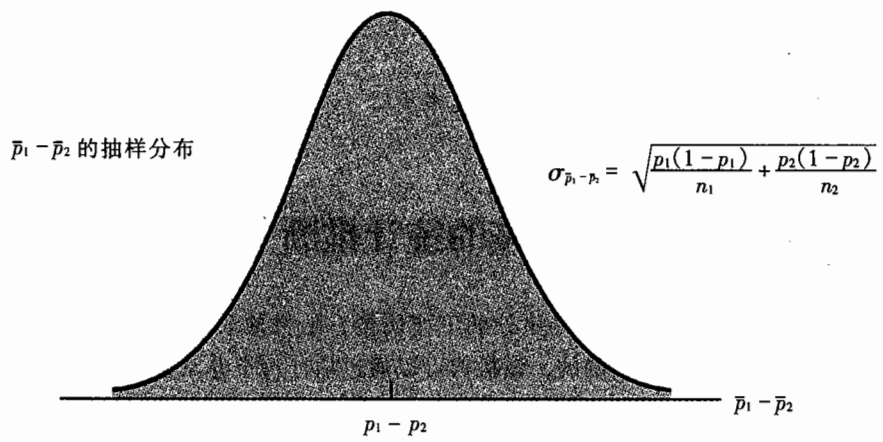


图 10.7 $\bar{p}_1 - \bar{p}_2$ 的抽样分布

$p_1 - p_2$ 的区间估计

我们假定两个办事处纳税申报单的独立简单随机样本提供以下信息：

办事处 1	办事处 2
$n_1 = 250$	$n_2 = 300$
错误申报数 = 35	错误申报数 = 27

两个办事处的样本比例如下：

$$\bar{p}_1 = \frac{35}{250} = 0.14$$

$$\bar{p}_2 = \frac{27}{300} = 0.09$$

两个总体错误申报比例之差的点估计量为$\bar{p}_1 - \bar{p}_2 = 0.14 - 0.09 = 0.05$。因此，我们估计办事处 1 的错误率比办事处 2 多 0.05 或 5%。

为推导两个总体之差的置信区间，我们需要如公式(10.16)所示的$\sigma_{\bar{p}_1 - \bar{p}_2}$的估计。然而，由于两个总体比例 p_1、p_2 未知，不能直接用公式(10.16)，此时，可用样本比例 $\bar{p}_1$ 估计 p_1，$\bar{p}_2$ 估计 p_2，于是得到$\sigma_{\bar{p}_1 - \bar{p}_2}$的点估计量如下：

$\sigma_{\bar{p}_1 - \bar{p}_2}$ 的点估计量

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2(1 - \bar{p}_2)}{n_2}} \qquad (10.17)$$

利用公式（10.17），下列表达式给出了两个总体比例差的一个置信区间估计，因为当样本容量较大时，$\bar{p}_1 - \bar{p}_2$ 的抽样分布可由正态分布逼近，所以公式中使用了 $z_{\alpha/2}$。

两个总体比例之差的区间估计：大样本情形 $n_1 p_1$、$n_1(1 - p_1)$、$n_2 p_2$ 和 $n_2(1 - p_2) \geqslant 5$。

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2} s_{\bar{p}_1 - \bar{p}_2} \qquad (10.18)$$

式中，$1 - \alpha$ 是置信系数。

用公式(10.17)，我们有

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{0.14(0.86)}{250} + \frac{0.09(0.91)}{300}} = 0.027\,5$$

在 90% 置信区间下，$z_{\alpha/2} = z_{0.05} = 1.645$，由公式(10.18)得出区间估计如下：

$$(0.14 - 0.09) \pm 1.645\ (0.027\,5)$$

$$0.05 \pm 0.045$$

因此，两个总体有错误率之差的 90% 置信区间为 0.005—0.095。

关于 $p_1 - p_2$ 的假设检验

考虑上例中的数据，作为有关两个总体比例差的假设检验的一个例子，并假设该机构试图确定两个办事处的有错误比例不同。我们通过检验下面的假设来介绍这种方法：

$$H_0: \ p_1 - p_2 = 0$$

$$H_a: \ p_1 - p_2 \neq 0$$

在两个总体均值之差相同(即 $p_1 - p_2 = 0$)的假设下，$\bar{p}_1 - \bar{p}_2$ 的抽样分布如图 10.8 所示。因抽样分布类似正态分布，两个总体比例之差的检验统计量可写为

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1 - \bar{p}_2}} \tag{10.19}$$

对 $\alpha = 0.10$ 和 $z_{\alpha/2} = z_{0.05} = 1.645$，拒绝规则为

如果 $z < -1.645$ 或 $z > 1.645$，则拒绝 H_0

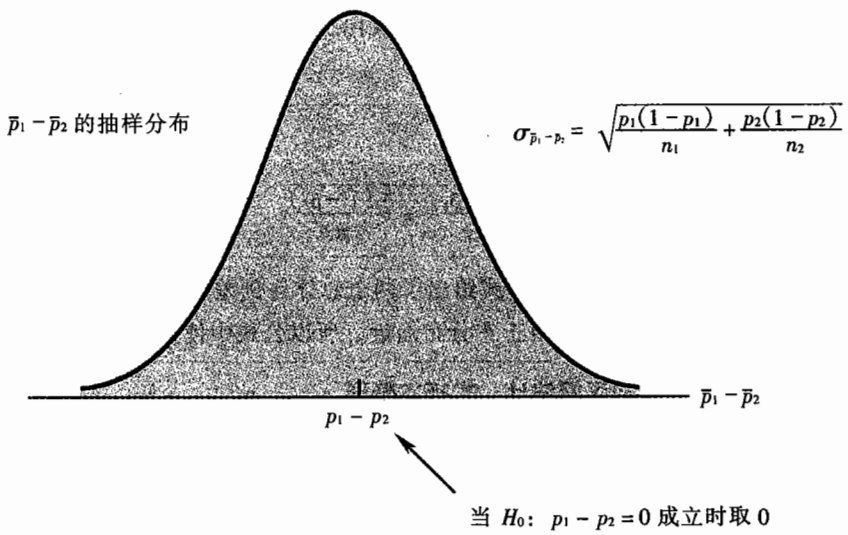


图 10.8 $H_0: \ p_1 - p_2 = 0$ 下，$\bar{p}_1 - \bar{p}_2$ 的抽样分布

公式(10.19)中的 z 的计算要用到两比例之差的标准差的值 $\sigma_{\bar{p}_1 - \bar{p}_2}$。因为该标准差未知，必须用样本数据来估计。类似于区间估计的步骤，我们可以尝试在公式(10.18)中用 $\bar{p}_1$ 和 $\bar{p}_2$。但在假设检验中，我们往往将公式（10.18）修正为略显不同的形式。在假设比例之差无差异的特殊情形(即 $H_0: \ p_1 - p_2 = 0$ 或 $H_0: \ p_1 - p_2 \leqslant 0$ 或 $H_0: \ p_1 - p_2 \geqslant 0$)，修正公式(10.18)是为了反映以下事实：当我们假设 H_0 中等号成立，相当于假设 $p_1 = p_2$。出现这种情形时，我们组合或合并两样本比例得到一个估计即合并估计量(记为 $\bar{p}$)。

在有关 H_0 为 $p_1 - p_2 = 0$ 或 H_0: $p_1 - p_2 \leqslant 0$ 或 H_0: $p_1 - p_2 \geqslant 0$ 的假设检验中，0 意味着两个总体的比例 p_1、p_2 将无差异。在这种情况下，应该将两个样本比例 $\bar{p}_1$、$\bar{p}_2$ 组合，得到总体比例的一个估计，见公式(10.20)。可按公式(10.21)计算 $s_{\bar{p}_1-\bar{p}_2}$。

两个总体比例的合并估计可写为

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2} \tag{10.20}$$

将公式(10.17)中的 $\bar{p}_1$、$\bar{p}_2$ 换成 $\bar{p}$，于是(10.17)变为

$$s_{\bar{p}_1-\bar{p}_2} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.21}$$

利用公式(10.20)和(10.21)，我们可进行以下计算：

$$\bar{p} = \frac{250(0.14) + 300(0.09)}{550} = \frac{62}{550} = 0.113$$

$$s_{\bar{p}_1-\bar{p}_2} = \sqrt{(0.113)(0.887)\left(\frac{1}{250} + \frac{1}{300}\right)} = 0.027\,1$$

利用公式(10.19)，检验统计量的值为

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{s_{\bar{p}_1-\bar{p}_2}} = \frac{(0.14 - 0.09) - 0}{0.027\,1} = 1.85$$

由于 1.85>1.645，所以原假设被拒绝。该侧区域与 $z = 1.85$ 相关的是 0.500 0 − 0.467 8 = 0.032 2。因此，双侧 p − 值是 2 × 0.032 2 = 0.064 4，它小于 $\alpha = 0.10$。样本资料表明，两个总体出错率差别明显。

就像我们见到的两个总体均值之差的假设检验，也可对两个总体比例之差进行单侧检验。单侧检验拒绝域的建立类似于单个总体比例的假设检验方法。

练　习

方　法

30. 考虑来自两个总体的两个独立样本的如下资料：

样本 1	样本 2
$n_1 = 400$	$n_2 = 300$
$\bar{p}_1 = 0.48$	$\bar{p}_2 = 0.36$

a. 两个总体比例之差的点估计量是多少？

b. 求两个总体比例之差的90%置信区间。

c. 求两个总体比例之差的95%置信区间。

31. 考虑假设检验：



$$H_0: p_1 - p_2 \leqslant 0$$
$$H_a: p_1 - p_2 > 0$$

下面的结果取自两个总体的两个独立样本：

样本1	样本2
$n_1 = 200$	$n_2 = 300$
$\bar{p}_1 = 0.22$	$\bar{p}_2 = 0.16$

a. 对 $\alpha = 0.05$，进行假设检验并得出你的结论。

b. p - 值是多少？

应　用

32. *Business Week*/Harris 对大公司的资深经理进行一次关于未来经济展望的问卷调查（*Business Week*, June 16, 1997）。调查的一个问题是“你认为你的公司在未来的12个月内，全日制雇员的数目是否会增加？”在1997年5月的调查中，400名经理中有220人的回答是“是”，而在1996年12月的调查中，400名经理中有192人回答“是”。求两个时点比例之差的95%置信区间。你对该区间估计有何解释？

33. Gallup 民意测验发现，在被调查的505名男子和496名女子中，分别有16%和25%的人赞同在全国范围内禁止销售啤酒、葡萄酒和烈性酒的立法。求出男子中赞同这种禁令的比例与女子中赞同这种禁令比例之差的95%置信区间。

34. *USA Today*/CNN 民意测验评估了在1998赛季 Mark McGwire 和 Sammy Sosa 创下驱逐对手本垒打记录，棒球联合总会的球迷对此的支持率。在1 082人组成的样本中，有682人对“你是棒球联盟的球迷吗”这个问题回答“是”（*USA Today*, September 17, 1998）。在1995年4月的棒球赛中，在1 008人组成的样本中，只有413人对同样的问题回答“是”。

a. 用样本率来评估1998年和1995年主要棒球联合总会的球迷支持情况。从1995年到1998年的棒球联盟支持率上升数的点估计量是多少？

b. 求球迷支持率上升数的区间估计的95%置信区间。

35. 《财富》杂志通过 Yankelovich Partners 对600名成年人做了一次民意调查，该调查是有关结婚、离婚和员工妻子对她当经理的丈夫的贡献（*Fortune*, February 2, 1998）。对于问题“丈夫在外工作、妻子无收入的长期婚姻，离婚后妻子应得结婚期间积累财产的一半”，300名女性中有279人同意，300名男性中有255名同意。

a. 对于男女两个总体，样本比例是多少？

b. 用 $\alpha=0.05$ 检验假设 H_0：$p_1-p_2=0$，得出你的结论？

c. 求两个总体比例之差的 95% 置信区间。

36. 由 1 545 名男性组成的样本和 1 691 名女性组成样本，来比较双职工婚姻中男女所做家务的数量。研究表明，67.5% 的男性和 60.8% 的女性认为自己做的家务是公平的（*American Journal of Sociology*，September 1994）。认为自己那份家务是公平的男性比例比女性比例大吗？取显著性水平 0.05，进行统计检验支持你的结论。

37. 在一次对两个电视广告质量的测试中，每个广告在一周内在某隔离的检验地区播放 6 次。下一周进行电话调查以识别已经看过广告的个人。对那些已看过广告的个人，要求他们陈述广告的主要内容。记录的资料如下：

广告	看过广告的人数	回忆起主要内容的人数
A	150	63
B	200	60

a. 用 $\alpha=0.05$，检验加设“两个广告的回忆比例无差别”。

b. 计算两个总体回忆比例之差的 95% 置信区间。

38. 2000 年早期，随着许多股票降至低于 1997 年的水平，股票市场随之低迷。事实上，纽约股票交易所的 81.5% 和纳斯达克股票的 72.4% 是在 1997 年高度以下交易的（*Barron's*, May 22, 2000）。若 232 只纽约股票交易所的股票和 210 个纳斯达克的数据已知，检验假设 H_0：$p_1-p_2=0$，$\alpha=0.05$。p-值是多少？得出你的结论？

39. 雅虎在线生活（*Yahoo Internet Life*）在一些大城市做了一些调查来评估成年人在工作中使用互联网的比例（*USA Taday*，May 7, 2000）。结果显示，40% 的华盛顿人和 32% 的旧金山人在工作中使用互联网。若各自的样本容量为 240 和 250，样本结果会显示出华盛顿成年人在工作中使用互联网的总体比例高于旧金山的吗？p-值是多少？若 $\alpha=0.05$，得出你的结论？

小　结

在本章中，我们讨论了进行两个总体的区间评估和假设检验的方法。首先，我们讲了如何在抽取了独立简单随机样本的情况下进行两个总体均值之差的推断。我们考虑大样本和小样本情形。当样本容量较大时，可用标准正态分布中的 z 值来推断两个总体均值之差。在小样本情形下，若总体为正态分布且方差相等时，用 t 分布进行推断。

接着，我们进行了两个总体均值之差的推断以用于讨论匹配样本方案。在匹配样本方案中，每个单元提供了一对来自各自总体的数据值。将该对数据值之差用于统计分析。由于匹配样本方案通常能减少抽样误差，提高估计的精确度，所以匹配样本方案优于独立样本方案。

最后，讨论了有关两总体比例之差的区间评估和假设检验。用于分析两总体比例之差的统计

学方法和用于分析两个总体均值之差的统计学方法是类似的。

术语释义

合并方差： 基于组合两个(或两个以上)样本资料的一个总体方差的估计；无论两个(或两个以上)总体方差是否相等，合并方差估计都正确。

独立样本： 组成其中一个样本的变量与组成其他样本的变量是相互独立选出的，即取自两个(或两个以上)总体的样本。

匹配样本： 一个样本中的每个数据值与另一个样本中对应的数据之相匹配的样本。

重要公式

两个总体均值之差的点估计量

$$\bar{x}_1 - \bar{x}_2 \tag{10.1}$$

$\bar{x}_1 - \bar{x}_2$ 的期望值

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2 \tag{10.2}$$

$\bar{x}_1 - \bar{x}_2$ 的标准差

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.3}$$

两个总体均值之差的区间估计：大样本情形（$n_1 \geqslant 30$ 且 $n_2 \geqslant 30$），σ_1 和 σ_2 已知

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2} \tag{10.4}$$

$\sigma_{\bar{x}_1 - \bar{x}_2}$ 的点估计量

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{10.5}$$

两个总体均值之差的置信区间估计：大样本情形下（$n_1 \geqslant 30$ 且 $n_2 \geqslant 30$），σ_1、σ_2 由 s_1、s_2 估计

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}s_{\bar{x}_1 - \bar{x}_2} \tag{10.6}$$

当 $\sigma_1^2 = \sigma_2^2$ 时，$\bar{x}_1 - \bar{x}_2$ 的标准差

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.7}$$

σ^2 的合并估计量

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \tag{10.8}$$

当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时，$\sigma_{\bar{x}_1-\bar{x}_2}$ 的点估计量

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{10.9}$$

两总体均值之差的区间估计：小样本情形（$n_1<30$ 且/或 $n_2<30$），σ_1、σ_2 由 s_1、s_2 估计

$$\bar{x}_1-\bar{x}_2\pm t_{\alpha/2}s_{\bar{x}_1-\bar{x}_2} \tag{10.10}$$

关于两个总体均值之差的假设检验统计量（大样本情形）

$$z=\frac{(\bar{x}_1-\bar{x}_1)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \tag{10.11}$$

两个总体均值在小样本情形的检验统计量

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{10.12}$$

匹配样本样本均值

$$\bar{d}=\frac{\sum d_i}{n}$$

匹配样本的样本标准差

$$s_d=\sqrt{\frac{\sum(d_i-\bar{d})^2}{n-1}}$$

匹配样本检验统计量为（小样本情形）

$$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}} \tag{10.13}$$

两个总体比例差的点估计量

$$\bar{p}_1-\bar{p}_2 \tag{10.14}$$

$\bar{p}_1-\bar{p}_2$ 的期望值

$$E(\bar{p}_1-\bar{p}_2)=p_1-p_2 \tag{10.15}$$

$\bar{p}_1-\bar{p}_2$ 的标准差

$$\sigma_{\bar{p}_1-\bar{p}_2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \tag{10.16}$$

$\sigma_{\bar{p}_1-\bar{p}_2}$ 的点估计量

$$s_{\bar{p}_1-\bar{p}_2}=\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{10.17}$$

两个总体比例之差的区间估计：大样本情形 n_1p_1、$n_1(1-p_1)$、n_2p_2 和 $n_2(1-p_2)\geqslant 5$

$$\bar{p}_1-\bar{p}_2\pm z_{\alpha/2}s_{\bar{p}_1-\bar{p}_2} \tag{10.18}$$

两个总体比例之差的假设检验统计量

$$z=\frac{(\bar{p}_1-\bar{p}_2)-(p_1-p_2)}{\sigma_{\bar{p}_1-\bar{p}_2}} \tag{10.19}$$

两个总体比例的合并估计

$$\bar{p}=\frac{n_1\bar{p}_1+n_2\bar{p}_2}{n_1+n_2} \tag{10.20}$$

当 $p_1=p_2$ 时，$\sigma_{\bar{p}_1-\bar{p}_2}$ 的点估计量

$$s_{\bar{p}_1-\bar{p}_2}=\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{10.21}$$

补充练习

40. 有两个关于商业专业的硕士和学士个人年薪的独立随机样本，利用下面数据来建立关于硕士薪水增长的置信度为90%置信区间。

硕士	学士
$n_1=60$	$n_2=80$
$\bar{x}_1=40\,000$ 美元	$\bar{x}_2=35\,000$ 美元
$s_1=2\,500$ 美元	$s_2=2\,000$ 美元

41. Safegate Foods 公司正在重新设计它在全国各地超市的安全监测通道。现有两种方案，他们在装有两种系统的两家商店分别进行有关顾客安检时间的检测。样本数据汇总如下：

系统 A	系统 B
$n_1=120$	$n_2=100$
$\bar{x}_1=4.1$ 分	$\bar{x}_2=3.3$ 分
$s_1=2.2$ 分	$s_2=1.5$ 分

在显著性水平为 0.05 下，检验两种系统的平均时间是否有差异？哪种系统更好？

42. 共同基金分为**有抵押金**的和**没有抵押金**的两种类型。有抵押金的要求投资者支付总投资额一定比例作为初始费用。没有抵押金就不要求支付此费用。一些金融顾问认为，有抵押金的基


CD 光盘数据
Mutual

金支付额外的费用是值得的，因为，这些基金比没有抵押金的有更高的平均收益率。从 1998 年 1 月 12 日出版的 *Barron's Lipper Mutual Funds Quarterly* 中选取 30 笔有抵押金和 30 笔没有抵押金作为样本。收集数据是关于 5 年期基金的年收入的前 5 个有抵押金及前 5 个没有抵押金的基金数据如下：

有抵押金的	平均收益率	没有抵押金的	平均收益率
American National Growth	15.51	Amana Income Fund	13.24
Arch Small Cap Equity	14.57	Berger One Hundred	12.13
Bartlett Cap Basic	17.73	Columbia International Stock	12.17
Calvert World International	10.31	Dodge & Cox Balanced	16.06
Colonial Fund A	16.23	Evergreen Fund	17.61

a. 用公式表示 H_0 和 H_a，其中拒绝 H_0 表示有抵押金的基金比没有抵押金的基金 5 年期间平均年收入高的结论。

b. 为进行共同基金假设检验使用 60 笔作为共同基金。使用 $\alpha = 0.05$，你的结论如何？

c. p - 值是多少？

43. 由两个老师教的两个班的统计学课程期末考试分数构成的样本如下所示，对于 $\alpha = 0.05$，检验得出的结果能否支持两个班级的平均分数不同的结论？

教师 A	教师 B
$n_1 = 12$	$n_2 = 15$
$\bar{x}_1 = 72$	$\bar{x}_2 = 78$
$s_1 = 8$	$s_2 = 10$

若 $\alpha = 0.05$，通过检验这些数据，是否可以推论出两个班的平均分不同？

44. Figure Perfect 公司是一家专门从事减肥项目的妇女健美沙龙。以下是客户样本体重在历时六周的引导性项目之前和之后的情况：

	体重	
客户	之前	之后
1	140	132
2	160	158
3	210	195
4	148	152
5	190	180
6	170	164

若 $\alpha = 0.05$，检验该项目是否导致统计上体重显著减轻，你的结论如何？

45. 1997 年的后几个月，亚洲经济出现了不景气。投资者预期亚洲经济的下滑应该对美国公司在 1997 年第四季度期间的收入有负面影响。下列样本数据给出了 1996 年第四季度和 1997 年第四季度一些公司股票的每股收入（*The Wall Street Journal*, January 28, 1998）。


CD 光盘数据
Asia

公司	1996 年收入	1997 年收入
Atlantic Richfield	1.16	1.17
Balchem Corp.	0.16	0.13
Black & Decker Corp.	0.97	1.02
Dial Corp.	0.18	0.23
DSC Communications	0.15	-0.32
Eastman Chemical	0.77	0.36
Excel Communications	0.28	-0.14
Federal Signal	0.40	0.29
Ford Motor Company	0.97	1.45
GTE Corp.	0.81	0.73
ITT Industries	0.59	0.60
Kimberly - Clark	0.61	-0.27
Minnesota Mining & Mfr.	0.91	0.89
Procter & Gamble	0.63	0.71

a. 计算 H_0 和 H_a，得出拒绝 H_0 的结论：1997 年第四季度每股的平均收入少于 1996 年第四季度的平均收入。

b. 利用光盘中文件 Asia 中的数据进行假设检验，若 $\alpha = 0.05$，你的结论如何？

46. *Business Week*/ Harris 对 1 035 名成年人进行“你认为美公司在全球的经济竞争力如何”的民意调查。其中 704 名回答者认为“好”/“极好”（*Business Week*, September 11, 2000）。在 1996 年类似的民意调查中，582 名回答者认为“好”/“极好”。这些样本结果能够说明从 1996 到 2000 年的 4 年中，成人回答“好”/“极好”的比例上升了吗？

a. 写出原假设和备择假设。

b. 计算 p - 值。

c. 若 $\alpha = 0.01$，你的结论如何？

47. 一家大的汽车保险公司抽取了未婚和已婚男性保险单持有人的样本，并记录了以往 3 年内要求索赔的次数。

未婚保险单持有人	已婚保险单持有人
$n_1 = 400$	$n_2 = 900$
要求索赔次数 = 76	要求索赔次数 = 90

a. 若 $\alpha = 0.05$，检验未婚及已婚保险单持有人的索赔率之间是否有差异？

b. 求两个总体比例之差的 95% 置信区间。

48. 为了研究结核耐药性，进行了医学检验。在新泽西州检验的 142 个病历中，有 9 例发现有耐药性。在得克萨斯州检验的 268 个病例中，有 5 例发现有耐药性。这些数据是否表明，两个州之间有耐药性病例的比例之间在统计上有明显的差异？在显著水平为 0.02 下，检验假设 H_0：$p_1 - p_2 = 0$。p - 值是多少？你的结论如何？

案例研究　Par 公司

Par 公司是一个较大的高尔夫球设施制造商。经理认为，引进某种耐磨持久的高尔夫球会使 Par 公司市场占有率上升。因此，Par 公司的研究小组开始调查一种为抵抗切割及使球更耐用而设计的高尔夫球外膜。对外膜的检验是很有前途的。

研究者之一关注新外膜对高尔夫球距离的影响。Par 公司希望新的耐磨球与目前型号的球的击球距离相同。为比较两种球的击球距离，用 40 只新型号球和 40 只目前型号球来作检验。为了能将两型号的平均距离的差异归因于方案的不同，检验是由一个机械击球装置进行的。检验结果如下，其中距离是按最接近的码数来测量的。该数据存储在 CD 光盘的 Golf 数据集中。



CD 光盘数据 Golf

型号		型号		型号		型号	
目前	新型	目前	新型	目前	新型	目前	新型
264	277	270	272	263	274	281	283
261	269	287	259	264	266	274	250
267	263	289	264	284	262	273	253
272	266	280	280	263	271	263	260
258	262	272	274	260	260	275	270
283	251	275	281	283	281	267	263
258	262	265	276	255	250	279	261
266	289	260	269	272	263	274	255
259	286	278	268	266	278	276	263
270	264	275	262	268	264	262	279

管理报告

1. 用公式表示并描述 Par 用于比较目前球与新球击球距离的假设检验基本原理。
2. 通过分析数据，得出假设检验的结论，检验 p – 值是多少？
3. 对每种型号的数据给出描述性的统一总结。
4. 每种型号的总体均值的 95% 置信区间是多少？两个总体均值之差的 95% 置信区间是多少？
5. 你认为高尔夫球需要更大的样本容量做进一步检验吗？讨论一下。

附录 10.1　用 Minitab 进行两个总体均值的假设检验

我们介绍了如何用 Minitab 对两个总体均值之间的差异进行三种假设的检验。

大样本情形

我们将用到表 10.1 中的两个培训中心考试分数的数据。培训中心 A 的分数出现在 C1 栏中，培训中心 B 的分数出现在 C2 栏中。Minitab 中设有基于 z 检验统计量的两个样本的假设检验。尽管如此，Minitab 两样本 t 检验程序在大样本情形下提供了几乎一样的结果。培训中心用 Minitab 两样本的 t 检验的必要步骤数据如下：


CD 光盘数据
ExamData

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Basic Statistics** 下拉菜单；
步骤 3. 选择 **2-Sample t** 选项；
步骤 4. 当出现 2-Sample t 对话框时：
选择 **Samples in different columns**；
在 **First** 框中输入 C1；
在 **Second** 框中输入 C2，确保 **Assume equal variances**（同方差假设）未被选择；
选择 **Options**
步骤 5. 输入 95 到 **Confidence level** 框中；*
在 **Test Mean** 框中输入 0，在 **Alternative** 框中输入非均值；
点击 **OK**；
步骤 6. 点击 **OK**。

Minitab 的输出结果说明 p - 值 = 0.04，在 0.05 意义水平下，我们拒绝 H_0。

在大样本情形下，我们已指出，z 检验统计量用于总体分布未知的情况下。然而，Minitab 设提供两样本 z 检验。这样，我们必须用两样本 t 检验来满足我们的需要。对于用 s 估计 σ 的大样本情形，t 值由 Minitab 提供，这同 z 值是 2.09 是相同的。因而，两样本 t 检验的惟一差别是 p - 值 = 0.04 稍大于 p - 值 = 0.036 6，这可以用 $z = 2.09$ 的标准正态可能分布得到。在多数样本中，p - 值的小差别不影响假设检验结论。

小样本情形

用于描述大样本情形的 Minitab，两样本 t 方法可用于两个总体均值之差的小样本假设检验。惟一不同的是 Assume equal variances 在步骤 4 选择。最后，在步骤 5 中的 Alternative 框中可能小于，不等于，或大于取决于备择假设 H_a。

匹配样本

我们用表 10.4 中的生产时间数据来说明匹配样本方法。方法 1 的完成时间输入 C1 栏，方法

* Minitab 提供了假设检验和区间估计结果，使用者应选择总体均值区间估计的任何一种置信水平，这里建议选择 95%。

2 的完成时间输入 C2 栏。匹配样本假设检验的 Minitab 步骤如下：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Basic Statistics;**
步骤 3. 选择 **Paired t;**
步骤 4. 当 Paired t 对话框出现时：
在 **First** 样本框中输入 **C1**，在 **Second** 样本框中输入 C2；
选择 **Options;**
步骤 5. 在 **Confidence level** 中输入 95，在 **Test Mean** 框中输入 0，在 **Alternative** 框中输入非均值，点击 **OK;**
步骤 6. 点击 **OK**。

附录 10.2　用 Excel 进行两个总体均值的假设检验

我们介绍了如何用 Excel 对两个总体均值之间的差异进行三种假设的检验。

大样本情形


CD 光盘数据
ExamData

我们用了表 10.1 中的考试分数，以及单元 A1 中的商标中心 A 和单元 B1 中的商标中心 B。A 中心的分数在单元 A2:A31 中，B 中心分数在单元 B2:B41 中，我们假定 VAR 功能用于计算两个总体变量，因而，我们有 $s^2_1 = (8)^2 = 64$ 和 $s^2_2 = (10)^2 = 100$。进行检验的步骤如下：

步骤 1. 选择 **Tools** 下拉菜单；
步骤 2. 选择 **Data Analysis;**
步骤 3. 当 Data Analysis 对话框出现时：
选择 ***z*-Test: Two samples for Means;**
点击 **OK;**
步骤 4. 当 *z*-Test: Two Sample for Means 对话框出现时：
在 **Variable 1 Range** 框中输入 A1:A31;
在 **Variable 2 Range** 框中输入 B1:B41;
在 **Hypothesized Mean Difference** 框中输入 0;
在 **Variable 1 Variance** 框中输入 64;
在 **Variable 2 Variance** 框中输入 100;
选择 **Labels**,
在 **Alpha** 框中输入 0.05;
选择 **Output Range** 和在框中输入 C1;

点击 **OK**。

检验统计量的值 $z=2.09$ 出现在单元格 D8 中，p - 值 $=0.0366$ 出现在单元格 D11 中。

小样本情形


CD 光盘数据
Software

我们利用表 10.3 中的软件研究数据。数据已输入到单元格 A1 中的当今 label 和单元格 B1 中的新 label 的 Excel 工作表中。当今技术的完成时间在单元格 A2:A13 中，新软件的完成时间在单元格 B2:B13 中，下列步骤用于计算有关两个总体均值差的小样本假设检验：

步骤 1. 选择 **Tools** 下拉菜单；
步骤 2. 选择 **Data Analysis**；
步骤 3. 当 Data Analysis 对话框出现时：
选择 **t-Test: Two Sample Assuming Equal Variances**；
点击 **OK**。
步骤 4. 当 t-Test: Two sample Assuming Equal 对话框出现时：
在 **Variable 1 Range** 框中输入 A1:A13；
在 **Variable 2 Range** 框中输入 B1:B13；
在 **Hypothesized Mean Difference** 框中输入 0，
选择 **Labels**；
在 **Alpha** 框中输入 0.05；
选择 **Output Range**；
在框中输入 C1；
点击 **OK**。

检验统计量的值 t = 2.16 出现在单元格 D10 中，单边 p - 值 $=0.021$ 出现在单元格 D11 中。

匹配样本


CD 光盘数据
Matched

我们利用表 10.4 中的匹配样本完成时间来说明：数据输入工作表中，label 方法 1 在单元格 A1 中，label 方法 2 在单元格 B1 中。单元格 A2:A7 中方法 1 的完成时间和单元格 B2:B7 中方法 2 的完成时间。Excel 方法使用步骤来精确描述除选取步骤 3 中的 t-Test: Paired Two Sample for Means 数据分析工具以外的小样本情形。Variable 1 range 是 A1:A7，variable 2 range 是 B1:B7。检验统计量的值 $t=2.196$ 在单元格 D10 中，双边 p - 值 $=0.08$ 在单元格 D13 中。

11 关于总体方差的推断

统计实例：美国会计总署

11.1 **关于一个总体方差的统计推断**

σ^2 的区间估计

假设检验

11.2 **两个总体方差的统计推断**

统计实例

美国会计总署

华盛顿*

美国会计总署(GAO)是联邦政府立法机构中的一个审计组织，该组织是独立非官方的。GAO的稽核员对联邦的日常计划和方案提议的有效性掌握生杀大权。为了履行他们的责任，稽核员必须精通复审和立法调查，并且需要熟练掌握统计分析技术。

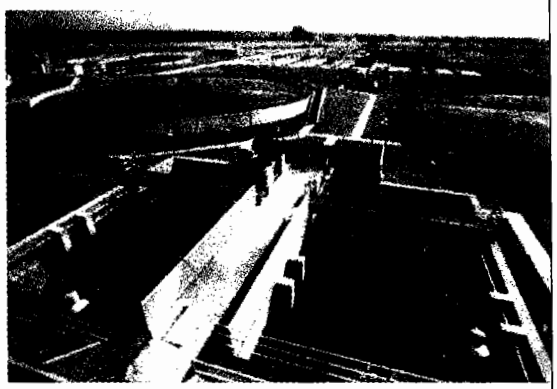
在这个案例中，GAO的稽核员对一个净化全国河流湖泊的内务部门进行研究。该项目其中一个任务是对全美的小城市进行审查。国会要求GAO提高判断项目的运作效率。为此，GAO就要求检查相关的记录并走访了几家废物处理厂。

GAO稽核员的目的之一是确认这些工厂的排泄物(治理后的污水)达到一定的标准。另外还包括其他的任务即样品资料的检验，如含氧量、pH值和排泄物中的悬浮固体物质。他们还必须在每个工厂进行日常的各种测验，收集资料定期送往各州的工程部门。GAO的调查情况还有排泄物的各项特性是否达标。

比如说，排泄物的平均pH必须经过仔细的检测，除此之外，上报的pH值的方差也需进行复查。下面就排泄物pH值的总体方差进行假设检验。

$$H_0:\ \sigma^2 = \sigma_0^2$$

$$H_a:\ \sigma^2 \neq \sigma_0^2$$



水处理厂的流出物需经过统计检验判断其pH值范围。
© John Boykin/The Stock Market.

在该检验中，σ_0^2代表运作良好的工厂必须达到的pH值总体方差。比如在一家工厂，如原假设被拒绝，进一步分析发现该厂的pH值方差明显低于正常状态。

稽核员到该厂检查测量设备并与该厂厂长讨论了有关统计数字。稽核员发现，操作员不知如何操作从未使用的测量设备。相反，操作员从工程师得知正常的pH值后，在没有做任何检验的情况下简单记录了几个类似的值。结果该厂因为其显著偏低的方差导致H_0被拒绝。GAO怀疑其他厂也有类似的问题，便建议对操作员培训提高收集治污资料的水平。

在本章中，将介绍如何进行单个或两个总体方差的统计推断。这里将出现两种新的用于进行总体方差的区间估值及假设检验的分布即χ^2分布和F分布。

* 作者感谢美国会计总署的Art Foreman和Dale Ledman提供统计应用的资料。

前 4 章我们介绍了总体的均值和比例方面的统计推断方法。本章将继续讨论总体方差的推断问题。为了解释一个方差如何提供决策性信息，我们以一个液体清洁剂的罐装生产过程为例。该生产过程的灌装机械装置经过调整，假定每个容器的重量为 16 盎司。规格设计量为 16 盎司，但灌入量仍然会发生偏差。也就是说，即使机械装置经过了很好的调整，我们也无法使容器装入量精确为 16 盎司。选择一些容量作为样品，我们可以计算出容器内液体的重量的方差。这个值可以作为整个灌装生产过程的方差。若样本方差适度，生产可以继续进行，若方差过大，即使均值恰好是 16 盎司也有可能发生过溢或不足的情况。在这种情况下，灌装机械装置就必须经过调整来减少容器的灌装方差。

在许多生产应用中，为了保证质量，控制方差极为重要。

下面我们先考虑单个总体方差的推断。接着讨论两个总体方差的推断步骤。

11.1　关于一个总体方差的统计推断

前几章我们用样本方差

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{11.1}$$

作为总体方差 σ^2 的估计值。使用样品方差作为推断单个总体方差的基础，$(n-1)s^2/\sigma^2$ 的样品抽样分布起着很大作用。该抽样分布如下所述：

$(n-1)s^2/\sigma^2$ 的抽样分布

从一个正态分布中任意抽取一个样本容量为 n 的样品，则

$$\frac{(n-1)s^2}{\sigma^2} \tag{11.2}$$

抽样分布为自由度为 $n-1$ 的 χ^2 分布。

χ^2 分布是以从正态总体取得样本为基础的。

图 11.1 列出了 $(n-1)s^2/\sigma^2$ 抽样分布的几种可能类型。

从概率表上我们很容易查到 χ^2 分布。因为只要样品取自一个正态分布，则 $(n-1)s^2/\sigma^2$ 抽样分布就是 χ^2 分布，我们可以利用 χ^2 分布计算一个总体方差的区间估计值并对其进行假设检验。

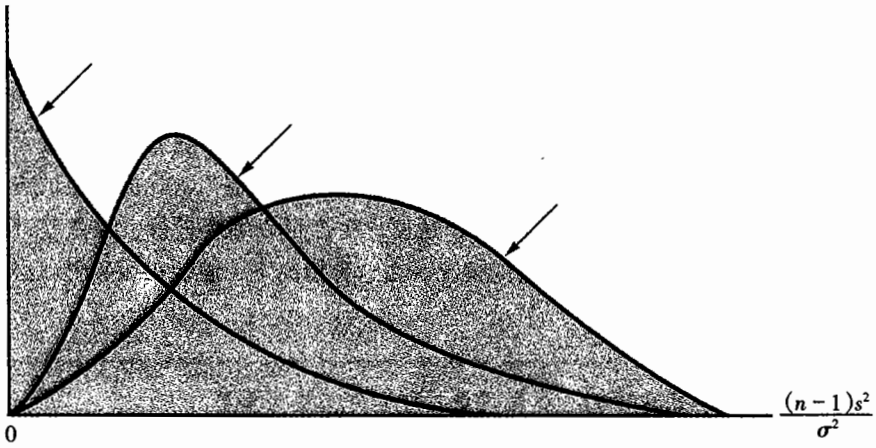


图 11.1 $(n-1)s^2/\sigma^2$ 的抽样分布（χ^2 分布）举例

σ^2 的区间估计

为了阐述如何利用 χ^2 分布来计算单个总体方差 σ^2 的精确区间估计值，回到刚才讨论的灌装生产过程，我们估算该过程的总体方差。取 20 个容器作为一个样本，可求得灌装量的样本方差为 $s^2=0.002\,5$。我们知道，20 个容器组成的样本无法提供全部容器的灌装量总体方差的精确值。因此，接下来，我们将研究如何计算总体方差的区间估计值。

记 χ^2_α 为 χ^2 分布的 α 水平**右侧**分位数。例如，图 11.2 所示的自由度为 19 的 χ^2 分布，$\chi^2_{0.025}=32.852\,3$ 表明有 2.5% 的 χ^2 分布值落在 32.852 3 右侧；$\chi^2_{0.975}=8.906\,55$ 表明有 97.5% 的 χ^2 分布值在 8.906 55 右侧。与表 11.1 对照可知自由度为 19 的 χ^2 分布值正确。附录 B 中的表 3 提供了更多的 χ^2 分布值。

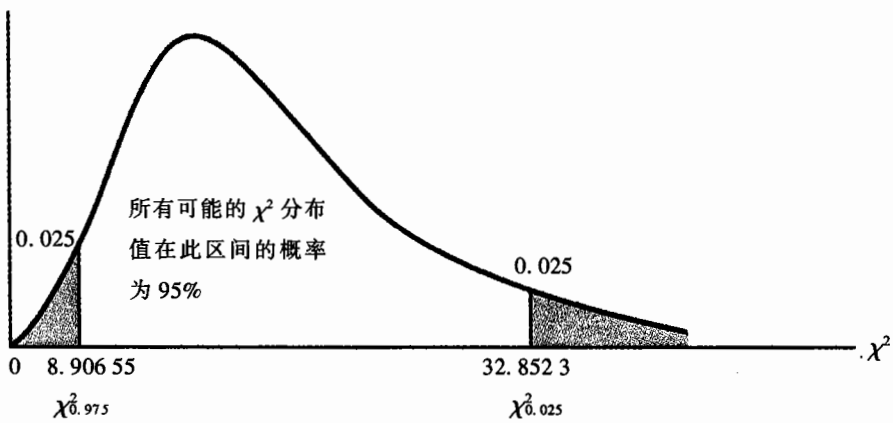
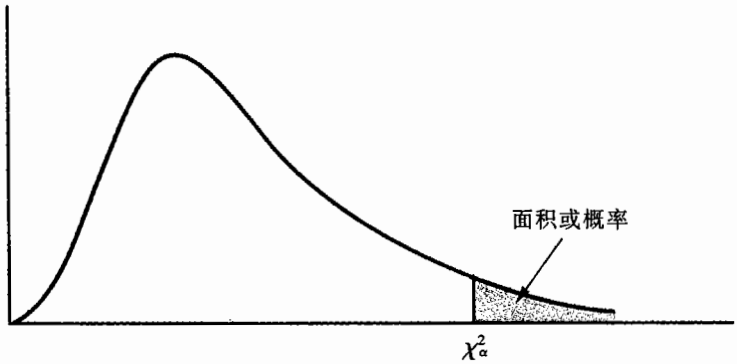


图 11.2 自由度为 19 的 χ^2 的分布图

从图 11.2 的曲线中我们可以看出，χ^2 分布值的 0.95 或 95% 在 $\chi^2_{0.975}$ 和 $\chi^2_{0.025}$ 之间，即获得 χ^2 值符合 $\chi^2_{0.975}\leqslant\chi^2\leqslant\chi^2_{0.025}$ 的概率为 0.95。

表 11.1　χ^2 分布表节选 *



自由度	右侧面积					
	0.99	**0.975**	**0.95**	**0.05**	**0.025**	**0.01**
1	$157\,088 \times 10^{-9}$	$982\,069 \times 10^{-9}$	$393\,214 \times 10^{-8}$	3.841 46	5.023 89	6.634 90
2	0.020 100 7	0.0 506 356	0.102 587	5.991 47	7.377 76	9.210 34
3	0.114 832	0.215 795	0.351 846	7.814 73	9.348 40	11.344 9
4	0.297 110	0.484 419	0.710 721	9.487 73	11.143 3	13.276 7
5	0.554 300	0.831 211	1.145 476	11.070 5	12.832 5	15.086 3
6	0.872 085	1.237 347	1.635 39	12.591 6	14.449 4	16.811 9
7	1.239 043	1.689 87	2.167 35	14.067 1	16.012 8	18.475 3
8	1.646 482	2.179 73	2.732 64	15.507 3	17.534 6	20.090 2
9	2.087 912	2.700 39	3.325 11	16.919 0	19.022 8	21.666 0
10	2.558 21	3.246 97	3.940 30	18.307 0	20.483 1	23.209 3
11	3.053 47	3.815 75	4.574 81	19.675 1	21.920 0	24.725 0
12	3.570 56	4.403 79	5.226 03	21.026 1	23.336 7	26.217 0
13	4.106 91	5.008 74	5.891 86	22.362 1	24.735 6	27.688 3
14	4.660 43	5.628 72	6.570 63	23.684 8	26.119 0	29.141 3
15	5.229 35	6.262 14	7.260 94	24.995 8	27.488 4	30.577 9
16	5.812 21	6.907 66	7.961 64	26.296 2	28.845 4	31.999 9
17	6.407 76	7.564 18	8.671 76	27.587 1	30.191 0	33.408 7
18	7.014 91	8.230 75	9.390 46	28.869 3	31.526 4	34.805 3
19	7.632 73	8.906 55	10.117 0	30.143 5	32.852 3	36.190 8
20	8.260 40	9.590 83	10.850 8	31.410 4	34.169 6	37.566 2
21	8.897 20	10.282 93	11.591 3	32.670 5	35.478 9	38.932 1
22	9.542 49	10.982 3	12.338 0	33.924 4	36.780 7	40.289 4
23	10.195 67	11.688 5	13.090 5	35.172 5	38.075 7	41.638 4
24	10.856 4	12.401 1	13.848 4	36.415 1	39.364 1	42.979 8
25	11.524 0	13.119 7	14.611 4	37.652 5	40.646 5	44.314 1
26	12.198 1	13.843 9	15.379 1	38.885 2	41.923 2	45.641 7

（续表）

自由度	右侧面积					
	0.99	0.975	0.95	0.05	0.025	0.01
27	12.878 6	14.573 3	16.151 3	40.113 3	43.194 4	46.963 0
28	13.564 8	15.307 9	16.927 9	41.337 2	44.460 7	48.278 2
29	14.256 5	16.047 1	17.708 3	42.556 9	45.722 2	49.587 9
30	14.953 5	16.790 8	18.492 6	43.772 9	46.979 2	50.892 2
40	22.164 3	24.433 1	26.509 3	55.758 5	59.341 7	63.690 7
50	29.706 7	32.357 4	34.764 2	67.504 8	71.420 2	76.153 9
60	37.484 8	40.481 7	43.187 9	79.081 9	83.297 6	88.379 4

* 其他 χ^2 分布值见附录 B 中表 3。

然而，如图 11.2 所述，$(n-1)s^2/\sigma^2$ 服从 χ^2 分布，因此可以将 $(n-1)s^2/\sigma^2$ 代入公式，记为

$$\chi^2_{0.975} \leqslant \frac{(n-1)s^2}{\sigma^2} \leqslant \chi^2_{0.025} \tag{11.3}$$

实际上，如公式(11.3)所示，$(n-1)s^2/\sigma^2$ 所求可能的取值落在区间 $(\chi^2_{0.975}, \chi^2_{0.025})$ 的概率为 0.95或 95%。现在我们对公式(11.3)进行代数运算以获得总体方差的 σ^2 区间估计值。由公式(11.3)左半部分可得

$$\chi^2_{0.975} \leqslant \frac{(n-1)s^2}{\sigma^2}$$

即

$$\sigma^2 \chi^2_{0.975} \leqslant (n-1)s^2$$

或

$$\sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11.4}$$

对右半部分同理，可得

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leqslant \sigma^2 \tag{11.5}$$

公式(11.4)和(11.5)结合为

$$\frac{(n-1)s^2}{\chi^2_{0.025}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{0.975}} \tag{11.6}$$

因为公式(11.3)对于 95% 的 $(n-1)s^2/\sigma^2$ 的值成立，故公式(11.6)给出了总体方差 σ^2 的 95% 的置信区间估值。

回到刚才那个计算罐装量总体方差的区间估值的问题上。已知由 20 个容器组成的样本可得样本方差 $s^2 = 0.002\ 5$。样本的容器量为 20，故其自由度为 19。如图 11.2 所示，我们已经知道 $\chi^2_{0.925} = 8.906\ 55$，$\chi^2_{0.025} = 32.852\ 3$，把这两个数值代入公式(11.6)可得到以下总体方差的区间估计：

$$\frac{(19)(0.002\,5)}{32.852\,3} \leqslant \sigma^2 \leqslant \frac{(19)(0.002\,5)}{8.906\,55}$$

或

$$0.001\,4 \leqslant \sigma^2 \leqslant 0.005\,3$$

开平方后可得如下所示的总体标准差的 95% 的置信区间：

$$0.037\,4 \leqslant \sigma \leqslant 0.072\,8$$

总体标准差的置信区间可通过计算总体方差的置信区间的下限和上限的平方根得出。

以上我们已经阐述了如何利用 χ^2 分布建立总体方差和总体标准差的区间估计。我们可以注意到，使用 $\chi^2_{0.975}$ 和 $\chi^2_{0.025}$ 时区间估计的置信系数为 0.95。推而广之，我们可以得到关于总体方差的区间估计公式如下：

单个总体方差的区间估计

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \tag{11.7}$$

式中，χ^2 值是以自由度为 $(n-1)$ 的 χ^2 分布为基础的，$1-\alpha$ 为置信系数。

假设检验

现在我们来介绍一个例子和对单个总体方差值进行假设检验所用的统计方法。St. Louis 汽车公司最近要求其员工恪守时间，以在公众面前树立一个良好的值得信赖的形象。该公司要求各个汽车的汽车到站时间变化不能太大，并将此作为一项标准的制度。就到站时间的方差来讲，公司要求的标准是到站时间方差不超过 4 分钟，公司定期地收集各汽车站的汽车到站时间，看各司机是否遵守时间。下面的假设检验就使用了样本资料：

$$H_0:\ \sigma^2 \leqslant 4$$
$$H_a:\ \sigma^2 > 4$$

若 H_0 为真，则到站时间的方差达到公司标准。只有当样本结果反映到站时间方差不符合公司标准时，我们才拒绝 H_0。从这一点上说，H_0 被拒绝就说明了必须采取进一步的措施以减少到站时间方差。



CD 光盘数据
MetroBus

假设在某市的汽车中转站抽取任意的 10 次汽车到站时间作为一个样本。若到站时间总体呈正态分布，我们可由公式(11.2)得知 $(n-1)s^2/\sigma^2$ 服从自由度为 $(n-1)$ 的 χ^2 分布。因此，检验统计数为

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{11.8}$$

服从自由度为 $n-1=9$ 的 χ^2 分布。根据原假设 $\sigma^2=4$、样本容量 $n=10$ 和样本 s^2，公式(11.8)给出了 χ^2 的下列观察值：$\chi^2=\frac{9s^2}{4}$

图 11.3 是 χ^2 的分布，显示出单侧检验的拒绝域。请注意只有当样本方差 s^2 导致 χ^2 分布值过大时，我们才拒绝 H_0。

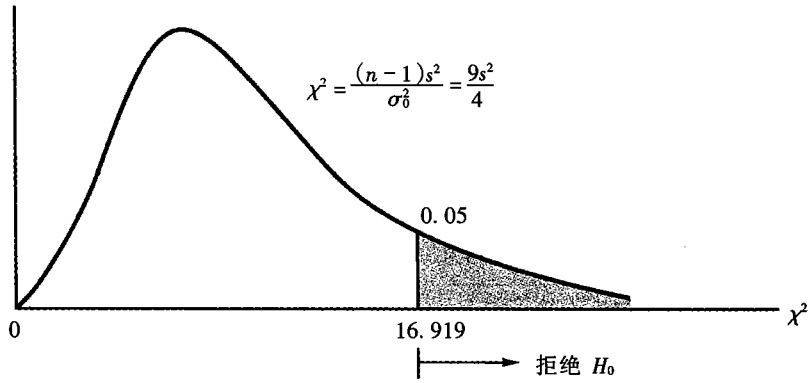


图 11.3　当 $\alpha=0.05$ 时，St. Louis 公司汽车检验的拒绝域

当 $\alpha=0.05$ 时，由表 11.1 可知，自由度为 9 时，$\chi^2_{0.05}=16.919$。以此作为检验的临界值，拒绝规则如下式：

$$\text{当 } \chi^2 > 16.919 \text{ 时，拒绝 } H_0$$

假设由 10 次到站时间组成的一样本，求得样本方差 $s^2=4.8$。这个值能否导致 H_0 被拒绝，因而得出汽车司机没有遵守制度的结论呢？我们可以将其代入公式得

$$\chi^2=\frac{9\times 4.8}{4}=10.8$$

因为 $\chi^2=10.8$ 小于 16.919，则不能拒绝 H_0。也就是，样本方差 $s^2=4.8$，不能说明汽车司机没有遵守时间制度。

假设检验还可以采用 p-值规则。常用的拒绝规则为：当 p-值$<\alpha$ 时拒绝 H_0。然而，从分布中很难查到 p-值，因此必须使用 Minitab 软件或 Excel 软件。附录 11.1 和 11.2 列出了相关的步骤。在前面的例子中，检验统计量 $\chi^2=10.8$ 相对应的 p-值为 0.29。因为 p-值 $0.29>\alpha=0.05$，所以不能拒绝原假设。

拒绝规则：如果 p-值$<\alpha$，则拒绝 H_0，被用于这一假设检验中。附录 11.1 或 11.2 说明了如何使用 Minitab 软件或 Excel 软件计算 p-值。

在实践应用中，检验统计量最经常使用的是单侧检验。在对汽车到站时间、生产时间、罐装重量、零件尺寸等检验时，一般都要求较小的方差，而较大的方差则不大可能接受。已知容许的最大总体方差，我们可以建立原假设，检验总体方差是否不超过容许的最大值，还是大于容许的最大值。

然而，我们除了对总体比例和均值检验时，还可对其他总体参数检验。单侧检验：H_0: $\sigma^2 \geqslant \sigma_0^2$ 与前面的检验相似，惟一不同的是，其拒绝域在 χ^2 分布临界值$\chi^2_{(1-\alpha)}$ 的左侧。现在总结单侧检验总体方差的步骤：

单侧检验总体方差

$$H_0: \sigma^2 \leqslant \sigma_0^2$$
$$H_a: \sigma^2 > \sigma_0^2$$

检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

拒绝规则

采用检验统计量：当 $\chi^2 > \chi^2_\alpha$ 时，拒绝 H_0

采用 p – 值：　　当 p – 值$<\alpha$ 时，拒绝 H_0

式中，σ_0^2 表示一个总体方差的假设值，χ^2_α 是自由度为 $n-1$ 的 χ^2 分布。

单侧检验总体方差

$$H_0: \sigma^2 \geqslant \sigma_0^2$$
$$H_a: \sigma^2 < \sigma_0^2$$

检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

拒绝规则

采用检验统计量：当 $\chi^2 < \chi^2_{(1-\alpha)}$ 时，拒绝 H_0

采用 p – 值：　　当 p – 值$<\alpha$ 时，拒绝 H_0

式中，σ_0^2 表示一个总体方差的假设值，$\chi^2_{(1-\alpha)}$ 是自由度为 $n-1$ 的 χ^2 分布。

上述中双侧检验与其他总体参数的双侧检验相似，分布两侧的面积占为 $\alpha/2$ 的面积是检验的临界值。下面介绍对一个总体方差双侧检验的拒绝规则：

一个总体方差的双侧检验

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_a: \sigma^2 \neq \sigma_0^2$$

检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

拒绝规则

采用检验统计量　当 $\chi^2 < \chi^2_{(1-\alpha/2)}$ 或 $\chi^2 > \chi^2_{\alpha/2}$ 时，拒绝 H_0

采用 p – 值：当 p – 值 $< \alpha$ 时，拒绝 H_0

式中，σ_0^2 表示一个总体方差的假设值，$\chi^2_{(1-\alpha/2)}$ 和 $\chi^2_{\alpha/2}$ 是自由度为 $n-1$ 的 χ^2 分布。

为了阐述对一个总体方差进行双侧检验时如何运用 χ^2 分布，我们介绍一个机动车辆管理处所遇到的问题。以前，申请驾照的个人考试分数方差为 $\sigma^2 = 100$。现在采用了新的考卷。该管理处负责人采用了新的考试试卷后个人考试分数保持在原有水平。为了估算新试卷考试分数的方差，他们采用了以下的双侧假设检验：

$$H_0: \sigma^2 = 100$$
$$H_a: \sigma^2 \neq 100$$

如果 H_0 被拒绝就说明了方差发生了变化，因此要使新的试卷分数方差保持原有水平必须对新试卷进行修订。取 30 份新版试卷的考试分数作为样本。

采用 χ^2 分布进行双侧检验。假设显著性水平为 0.05，则临界值为 $\chi^2_{0.975}$ 和 $\chi^2_{0.025}$。当自由度为 $n-1=29$ 时，表 11.1 中可知 $\chi^2_{0.975}{}^{*} = 16.047\,1$，$\chi^2_{0.025} = 45.722\,2$。双侧检验的拒绝规则为

如果 $\chi^2 < 16.047\,1$ 或 $\chi^2 > 45.722\,2$ 时，拒绝 H_0

若由 30 份新试卷考试的分数组成一个样本的方差 $s^2 = 64$ 时，我们可以得出什么样的结论呢？因为 $H_0: \sigma^2 = 100$ 为真，χ^2 的统计量计算为

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{29(64)}{100} = 18.56$$

因为 18.56 不在拒绝区域中，我们不能拒绝 H_0。因此我们可知，新试卷考试的分数与以往考试分数的方差没有什么变化。附录 11.1 和 11.2 中可看出该假设检验的 p – 值为 0.136。

练　习

方　法

1. 从表 11.1 或附录 B 的表 3 中找出下列 χ^2 分布的值：

* 编者注：疑原书误为 0.925。

a. 自由度 = 5，求 $\chi^2_{0.05}$

b. 自由度 = 15，求 $\chi^2_{0.025}$

c. 自由度 = 20，求 $\chi^2_{0.975}$

d. 自由度 = 10，求 $\chi^2_{0.01}$

e. 自由度 = 18，求 $\chi^2_{0.95}$

2. 由 20 个个体组成一个样本，样本标准差为 5：



a. 计算总体方差 90% 置信区间。

b. 计算总体方差 95% 置信区间。

c. 计算总体标准差 95% 置信区间。

3. 由 16 个个体组成一个样本，样本标准差为 8。取 $\alpha = 0.05$，检验下列假设，得出你的结论。

$$H_0: \sigma^2 \leq 50$$
$$H_a: \sigma^2 > 50$$

应　用

4. 药物重量的方差在制药工业中非常关键。以某一特定的药物为例，按克计量，由 18 个单位组成的样本求得样本方差为 $s^2 = 0.36$。

a. 建立一个该药物重量总体方差的 90% 置信区间。

b. 建立一个总体标准差的 90% 置信区间。

5. 有 8 个城市的租车的日费用如下所示（*The Wall Street Journal*, December 12, 1997）：

城市	租车日费用/美元	城市	租车日费用/美元
亚特兰大	47	菲尼克斯	40
芝加哥	50	匹兹堡	43
达拉斯	53	圣弗朗西斯科	39
新奥尔良	45	西雅图	37

a. 计算这些数据的方差和标准差。

b. 租车费用的总体方差的 95% 置信区间是多少？

c. 总体标准差的 95% 置信区间是多少？

6. Oppenheimer Capital Appreciation 互助基金 5 年中的平均收益率为 17.6%。假设下列各数分别代表这 5 年中每年的收益率：10.8，34.2，4.2，9.4 和 29.4。年收益率的标准差可反映出风险大小，标准差越大，方差越大，则年收益的不确定性也越大。一份报告中指出 10 只类似增长型基金的标准差为 18.2%。

a. 用以上数据作为 5 年收益的样本，由此求得的样本方差可以反映出 Oppenheimer Capital Appreciation 互助基金的风险，求出该值。

b. Oppenheimer Capital Appreciation 继今年收益的总体标准差的 95% 置信度区间是多少？

7. 作为一个样本，下面举出了 1998 年 8 份股票的利润估计值（*Barron's*, December 8, 1997）：

公司	每份股票的利润估计值
AT&T	2.92
Caterpillar	4.65
Eastman Kodak	4.27
Exxon	3.09
Hewlett-Packard	3.57
IBM	7.04
McDonald's	2.64
Wal-Mart	1.74

a. 计算以上数据的样本方差和样本标准差。

b. 总体方差的95%置信区间是多少?

c. 总体标准差的95%置信区间是多少?

8. 由12位证券分析员组成的小组对Qualcomm有限公司2001年每份股票的利润进行了估算，数据如下:

1.40 1.40 1.45 1.49 1.37 1.27 1.40 1.55 1.40 1.42 1.48 1.63

a. 计算每份股票利润估计值的样本方差。

b. 计算每份股票利润估计值的标准方差。

c. 给出总体方差和总体标准差的95%置信区间。

9. 

汽车零件必须经过细致加工才能使其公差小，顾客才会接受。生产技术要求零件长度最大公差为0.000 4。设由30个零件组成的样本求得样本方差 $s^2=0.000\,5$。取 $\alpha=0.05$，检验总体方差是否符合标准。

10. 一个VCR的平均寿命为6年，标准差为0.75年(*Consumer Reports 1995 Buying Guide*)。取30台电视机的寿命组成一个样本，其样本标准差为2年。进行假设检验电视机寿命的标准差是否显著大于VCR寿命的标准差，取 $\alpha=0.05$，得出你的结论。

11. 各地的30年定期贷款的抵押利息率各不相同。2000年夏天，全国各地的资料表明利息率标准差为0.096(*The Wall Street Journal*, September 8, 2000)，平方为$(0.096)^2=0.009\,216$。2001年对此继续研究。取20个贷款作为一个样本，其30年贷款的利息率的样本标准差为0.114。对下面的假设: H_0: $\sigma^2=0.009\,216$，检验样本资料是否表明利息率的方差有所改变。取 $\alpha=0.05$，得出你的结论。

12. 《财富》杂志作了一次调查，发现订阅该杂志的人拥有一辆或租用一辆车子数的方差为0.94。假设另一份杂志的订阅者拥有或租用的车辆数如下所示:2,1,2,0,3,2,2,1,2,1,0,1。

a. 计算这12位订阅者拥有或租用的车辆数的样本方差。

b. 对假设 H_0: $\sigma^2=0.94$ 检验，车辆数方差是否与《财富》的车辆数方差 $\sigma^2=0.94$ 不同。取 $\alpha=0.05$，得出你的结论。

11.2　两个总体方差的统计推断

在统计学的一些应用中，我们或许想比较两种生产线生产出来的产品质量的方差、两种装配方法所需装配时间的方差或者两种加热装置的温度的方差。对两个总体的方差进行比较时，我们可以使用两个独立的随机抽取的样本，设它们分别取自总体 1 和总体 2。两个样本方差s_1^2和s_2^2可以作为推断两个总体方差σ_1^2和σ_2^2的基础。当两个正态总体的方差相等时($\sigma_1^2=\sigma_2^2$),样本方差的比s_1^2/s_2^2的抽样分布如下所述：

当$\sigma_1^2=\sigma_2^2$时，s_1^2/s_2^2的抽样分布

设由两个具有相同方差的总体中抽取的两个独立的样本容量分别为 n_1 和 n_2，则

$$\frac{s_1^2}{s_2^2} \tag{11.9}$$

的抽样分布为分子自由度为 n_1-1 和分母自由度为 n_2-1 的 F 分布。s_1^2是由总体 1 中 n_1 个个体组成的样本方差，而s_2^2是由总体 2 中 n_2 个个体组成样本的样本方差。

F 分布表是以两个正态分布的抽样分布为基础的。

图 11.4 描述了分子和分母的自由度均为 20 时的 F 分布情况。由图可知，F 分布不对称，而且 F 值永远不取负值。任何 F 分布的形状取决于分子分母的自由度大小。

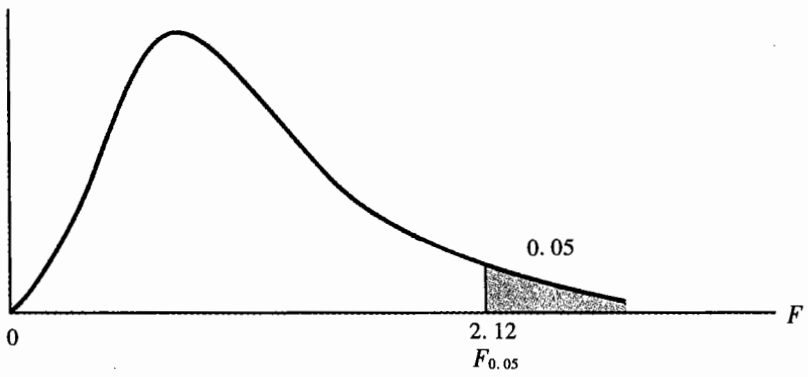
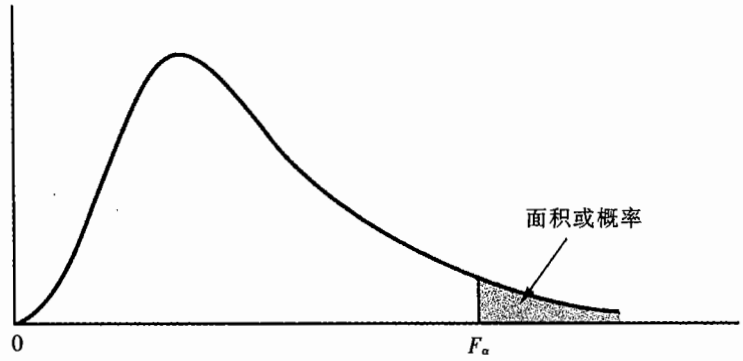


图 11.4　分子和分母自由度均为 20 的 F 分布图

我们用 F_α 表示其右侧的面积或概率值为 α 的 F 分布。比如说，由图 11.4 可见，$F_{0.05}$ 表示分子分母自由度为 20 时，右侧有 5% 的 F 值。表 11.2 中可见对于这种 F 分布，$F_{0.05}=2.12$。附录 B 的表 4 提供了更多的 F 分布值。现在介绍 F 分布在进行两个总体方差的假设检验中的应用。

Dullus 学校想更新一下来年的汽车服务的合同，但须在 Milbank 公司和 Gulf Park 公司中选择

表 11.2 F 分布表节选*



分母的自由度	$F_{0.05}$ 值的表							
	6	**7**	**8**	**9**	**10**	**12**	**15**	**20**
1	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0
2	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45
3	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66
4	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80
5	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56
6	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87
7	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44
8	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15
9	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94
10	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77
11	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65
12	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54
13	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46
14	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39
15	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33
16	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28
17	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23
18	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19
19	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16
20	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12
21	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10
22	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07
23	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05
24	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03

* 其他的 F 分布值见附录 B 的表 4。

一个。我们先用汽车到达的时间和运送时间作为衡量汽车服务质量的标准。方差越小，则表示服务质量稳定且水平高。如果两家公司的汽车到达时间的方差相等，Dullus 学校的负责人就会选择价格较低的那家公司。然而，如果两家公司的汽车到达时间方差迥异，这位负责人将考虑服务级别更好或方差更小的那家公司。对应的假设如下：

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

如果拒绝 H_0，则说明两家公司的服务质量不同。那么，这位负责人将选择方差较小的那家公司。

假设检验中取 $\alpha = 0.10$。假设我们从目前学校正在使用的两种汽车服务中收集汽车到达时间作为样本。其中样本 1 为 25 次 Milbank 公司汽车到达时间组成的样本，样本 2 为 16 次 Gulf Park 公司汽车到达时间组成的样本。图 11.5 中的曲线表示分子自由度为 $n_1 - 1 = 24$，分母自由度为 $n_2 - 1 = 15$ 的 F 分布。注意两侧检验的拒绝域两侧的临界值分别为 $F_{0.95}$ 和 $F_{0.05}$。

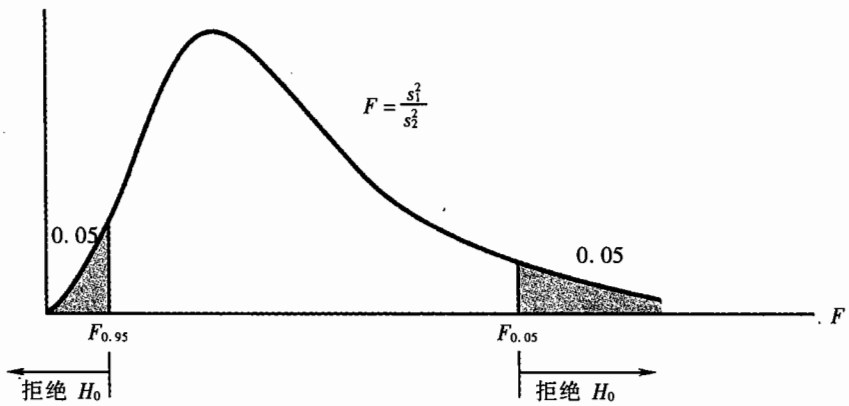


图 11.5　当 $\alpha = 0.10$ 时，Dullus 学校校车例子的拒绝域

设由两个样本求得 Milbank 公司服务级别的样本方差 $s_1^2 = 48$，而 Gulf Park 公司服务级别的样本方差 $s_2^2 = 20$。由此可得出什么恰当的结论？假设两个总体到达时间总体均值呈正态概率分布，并且假定 H_0：$\sigma_1^2 = \sigma_2^2$ 为真。我们可以用 F 分布得出结论。为此，我们参考图 11.5 所示的拒绝域进而计算出 $F = s_1^2/s_2^2$ 的值。具体如下所示：

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

参照附录 B 的表 4，我们可以查出分子自由度为 24，分母自由度为 15 的上侧临界值 $F_{0.05} = 2.29$。虽然表 4 没有给出 $F_{0.95}$ 的值，但注意到这里不需要确定下侧临界值，所以对本体没有影响。由上述可知，$F = 2.40$ 大于 $F_{0.05} = 2.29$，所以，在显著性水平为 0.10 时，H_0 被拒绝。由这个结果我们可知两家汽车公司运送时间方差不同。因此，Dullus 学校负责人就可以优先考虑服务更好或方差较低的 Gulf Park 公司。

对两个总体方差进行假设检验时也可以用 p - 值方法。常用的拒绝规则是：当 p - 值$<\alpha$ 时，拒绝 H_0。然而和 χ^2 分布一样，我们很难从 F 分布表中直接得到 p - 值。但是，我们可使用 Minitab 软件包或 Excel。附录 11.1 和 11.2 列出了这两种软件的应用方法步骤。在研究上面校车问题时，我们可以得知 $F=2.40$ 对应的 p - 值 $=0.082$。因为 p - 值 $=0.082<\alpha=0.10$，因此方差相等的假设被拒绝。

当 p - 值$<\alpha$ 时，拒绝 H_0 的拒绝规则对该检验同样适用。

在上面的校车的例子中，可能我们进行假设检验时是幸运的，因为 F 分布表中没有给出右侧临界值 $F_{0.95}$，而我们仍旧能够在不知道 $F_{0.95}$ 大小的情况下得出正确的结论。但如果需要知道右侧临界值 $F_{(1-\alpha)}$，可通过 F_α 和 $F_{(1-\alpha)}$的关系来求出 $F_{(1-\alpha)}$值，关系见下公式：

$$F_{(1-\alpha),\mathrm{df}_1,\mathrm{df}_2}=\frac{1}{F_{\alpha},\mathrm{df}_2,\mathrm{df}_1} \tag{11.10}$$

公式 11.10 说明了如何根据右侧临界值 F 值求出相应的左侧临界值 F 值。根据表 11.3 及附录 B 中表 4 所示的右侧临界值 F 值，我们可以计算出相应的左侧临界值 F 值。因此就没必要专门求出左侧临界值 F 值。

由上式可知，分子自由度为 24($\mathrm{df}_1=24$),分母自由度为 15($\mathrm{df}_2=15$)时，$F_{0.95}$ 是分子自由度为 15，分母自由度为 24 的 $F_{0.05}$ 的倒数。从 F 分布表中我们可以查得分子自由度为 15，分母自由度为 24 的值 $F_{0.05}$ 大小为 2.11，因此，分子自由度为 24，分母自由度为 15 的 $F_{0.95}$ 值的大小为：

$$F_{0.95}=\frac{1}{2.11}=0.47$$

虽然我们可以通过公式(11.10)求右侧 F 值，但常用的方法是进行假设检验计算，因此只需要知道左侧 F 值。进行 H_0：$\sigma_1^2=\sigma_2^2$假设检验时，我们只是简单的用总体 1 来代表**方差较大**的总体，就是说，可以随意规定哪个总体用总体 1 或总体 2。当我们设总体 1 为方差较大的总体时，只有在**右侧**才有可能出现 H_0 被拒绝的情况。虽然左侧临界值仍存在，但我们不需要知道它的值，因为使用方差较大的总体作为总体 1 的转换，s_1^2/s_2^2往往出现在右侧。上面校车的例子中，作为总体 1 的 Mibank 汽车公司的服务级别方差较大。因此，我们可以直接进行检验。如果 Gulf Park 公司的服务级别方差较大，我们就得设 Gulf Park 公司为总体 1，然后进行同样的统计检验。方差相等的两个总体的两侧检验步骤如下：

两个总体方差的两侧检验

$$H_0: \sigma_1^2=\sigma_2^2$$
$$H_a: \sigma_1^2\neq\sigma_2^2$$

设方差较大的总体为总体 1。

检验统计量

$$F=\frac{s_1^2}{s_2^2}$$

拒绝规则

用检验统计量：如果 $F>F_{\alpha/2}$ 时，则拒绝 H_0

用 p – 值：如果 p – 值$<\alpha$，则拒绝 H_0

式中，$F_{\alpha/2}$ 是基于分子自由度为 n_1-1、分母自由度为 n_2-1 的 F 分布。

我们还可以进行两个总体方差的单侧检验。同样还须利用 F 分布，其单侧拒绝区域可以表示出其中一个总体方差是否明显地大于或小于另一个总体方差。而且我们只须知道右侧临界值 F 值。进行单侧检验的时候，我们建立原假设。因此拒绝区域在右侧，我们只需将 H_a 方差较大的总体设为总体 1。其步骤如下所述：

两个总体方差的单侧检验已经形成定式，其拒绝域总是在 F 分布的右侧。这种方法省去了左侧 F 分布值的计算。

两个总体方差的单侧检验

$$H_0:\ \sigma_1^2\leqslant\sigma_2^2$$

$$H_a:\ \sigma_1^2>\sigma_2^2$$

检验统计量

$$F=\frac{s_1^2}{s_2^2}$$

拒绝规则

用检验统计量：如果 $F>F_{\alpha}$ 时，则拒绝 H_0

用 p – 值：如果 p – 值$<\alpha$，则拒绝 H_0

式中，F_{α} 是基于分子自由度为 n_1-1、分母自由度为 n_2-1 的 F 分布。

为了具体阐述如何利用 F 分别对两个总体方差进行单侧检验。以民意调查为例，取 31 位男性和 41 位女性对当前政治问题的看法作为样本，调查人员想看看样本结果是否表明对于同一政治问题女性偏差比男性大。根据上述的单侧检验形式，我们设女性为总体 1，男性为总体 2。具体检验步骤如下：

$$H_0:\ \sigma_{女}^2\leqslant\sigma_{男}^2$$

$$H_a:\ \sigma_{女}^2>\sigma_{男}^2$$

如果 H_0 被拒绝，则统计数据将证明对待同一政治问题上女性的看法比男性看法更不一致（方差更大）。

将女性的样本方差作为分子，男性的样本方差作为分母，故我们要采用分子自由度为 41 - 1 = 40，分母自由度为 31 - 1 = 30 的 F 分布进行单侧检验。取 $\alpha = 0.05$，其拒绝域基于 $F_{0.05}$。因此，拒绝规则为：如果 $F > 1.79$，则拒绝 H_0。F 为两个样本方差的比例 s_1^2/s_2^2。

假设调查结果表明 41 名女性的样本方差 $s_1^2 = 120$，31 名男性的样本方差 $s_2^2 = 80$，能得出什么适当的结论？F 统计量为

$$F = \frac{s_1^2}{s_2^2} = \frac{120}{80} = 1.50$$

因为 1.50 小于 1.79，H_0 不能被拒绝。就是说，样本结果不足以证明对待同一政治问题上女性的看法比男性看法更不一致(方差更大)。当 $F = 1.50$ 时，相应的 p - 值为 0.126。

评 注

1. 研究人员证实在总体为正态分布的前提下 F 分布的适用性，两个总体至少是近似正态分布，否则不能使用 F 分布。
2. 第 10 章中检验两个总体均值是否相等时，必须先假定 $\sigma_1^2 = \sigma_2^2$。有人建议，在 t 检验之前先进行基于 F 分布的 H_0：$\sigma_1^2 = \sigma_2^2$ 假设检验，才能验证假设 $\sigma_1^2 = \sigma_2^2$ 是否合理。但是，研究人员报道，t 检验发现不了的问题 F 检验同样无济于事，因此 t 检验之前进行 F 检验没有必要（*The American Statistician*, November 1990）。

练 习

方 法

13. 在附录 B 的表 4 中找出下列的 F 分布值：
    a. 自由度为 12 和 10 的 $F_{0.05}$。
    b. 自由度为 20 和 15 的 $F_{0.025}$。
    c. 自由度为 8 和 12 的 $F_{0.01}$。
    d. 自由度为 10 和 20 的 $F_{0.975}$。

14. 从总体 1 中取出 16 项组成一个样本。样本方差 $s_1^2 = 5.8$；从总体 2 中取出 20 项组成一个样本，样本方差为 $s_2^2 = 2.4$。取 $\alpha = 0.05$，进行下面的假设检验：

$$H_0:\ \sigma_1^2 \leqslant \sigma_2^2$$
$$H_a:\ \sigma_1^2 > \sigma_2^2$$

得出你的结论。

15. 考虑下面的假设检验：



$$H_0:\ \sigma_1^2 = \sigma_2^2$$
$$H_a:\ \sigma_1^2 \neq \sigma_2^2$$

设 $n_1 = 25$，$s_1^2 = 4.0$，$n_2 = 21$，$s_2^2 = 8.2$，取 $\alpha = 0.05$，得出你的结论。

应　用

16. Media Metrix 和 Jupiter Communications 收集了一些有关成人与小孩每日上网时间的资料。调查表明，一般情况下，成人上网时间比小孩多。假设以 25 个成人和 30 个小孩为样本进行调查，每月上网时间的标准差分别为 94 分钟和 58 分钟。样本结果是否表明成人上网时间方差比小孩大？取 $\alpha = 0.01$（*USA Today*, September 14, 2000）。

17. 大家清楚一辆汽车每年的平均维修费用取决于这辆车的使用的时间。一位研究人员想知道每年的维修费用的方差是否随着汽车的使用期的增加而增加。选择 25 辆 4 年的汽车和 25 辆两年的汽车作为样本，前后两种汽车的维修费用分别为 170 美元和 100 美元，并分别求得两个样本标准差。



a. 建立原假设和备择假设，检验使用期较长的汽车年均维修费用的方差较大。

b. 取 $\alpha = 0.01$，得出你的结论，讨论结果是否正确？

18. 10 家航天公司每份股票 12 个月利润的标准差为 4.27，而 7 家汽车公司每份股票 12 个月利润的标准差为 2.27（*Business Week*, August 14, 2000）。检验两个方差是否相等，取 $\alpha = 0.05$，得出你的结论。

19. 生产过程的方差可以衡量生产过程的质量优劣。方差过大，则必须采取措施改进生产过程以减小方差。下面的数据是两台机器上生产出来的茶包重量，茶包重量以克计量（*Quality Progress*, February 1995）。检验两类茶包重量的方差是否明显不同，取 $\alpha = 0.05$，得出你的结论。假如有差异，哪台机器必须改进呢？



机器 1	2.95	3.45	3.50	3.75	3.48	3.26	3.33	3.20
	3.16	3.20	3.22	3.38	3.90	3.36	3.25	3.28
	3.20	3.22	2.98	3.45	3.70	3.34	3.18	3.35
	3.12							
机器 2	3.22	3.30	3.34	3.28	3.29	3.25	3.30	3.27
	3.38	3.34	3.35	3.19	3.35	3.05	3.36	3.28
	3.30	3.28	3.30	3.20	3.16	3.33		

20. Romac 对薪水调查资料表明，国营会计公司的普通员工的年薪方差大约为 2.1，而老板的年薪

方差大约为 11.1，以 1 000 美元为单位。假设这份资料的调查对象是 25 位普通员工和 25 位老板，检验两者年薪方差是否相等。取 $\alpha=0.05$，得出你的结论。

21. 道琼斯工业股票的指数随着投资者买进买出 30 种股票而变化。该股票在 1997 年 11 月和 1997 年 12 月的前 5 天各抽取不同时间的指数，数据如下（*Barron's*, December 8, 1997）：



CD 光盘数据
DownJones

11 月	12 月
7 493	8 066
7 525	8 209
7 760	7 842
7 499	7 943
7 555	7 846
7 690	8 071
7 668	8 055
7 600	8 159
7 516	7 828
7 711	8 109

a. 计算道琼斯工业股票的指数在这两个时期的方差。

b. 取 $\alpha=0.05$，检验两个时期的总体方差是否相等，得出你的结论。

22. 一项研究发现在湿路上汽车的刹车距离方差比在干路上明显要大。16 辆速度相同的汽车先后在湿路和干路上进行检验。结果发现，在湿路上刹车距离的标准差为 32 英尺，而在干路上的标准差为 16 英尺。

a. 取 $\alpha=0.05$，样本资料是否表明在湿路上汽车刹车距离的方差要比在干路大？

b. 如果统计结论表明问题出在驾车安全上，应该注意什么？

小 结

本章介绍了总体方差的统计推断方法。其间我们介绍了两种新的概率分布：χ^2 分布和 F 分布。χ^2 分布可应用于正态总体方差的区间估计和假设检验。

我们还阐述了在两个总体方差的假设检验中 F 分布的运用。特别是，我们介绍了取自方差相等（$\sigma_1^2=\sigma_2^2$）的正态总体的随机样本（其容量分别为 n_1 和 n_2），样本方差比 s_1^2/s_2^2 呈 F 分布，其中分子自由度为 n_1-1，分母自由度为 n_2-1。

重要公式

单个总体方差的区间估计

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2)}} \tag{11.7}$$

单个总体方差的检验统计数

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{11.8}$$

两个总体方差的统计检验

$$F = \frac{s_1^2}{s_2^2} \tag{11.9}$$

补充练习

23. 由于人员分配问题，Gibson-Marimont 旅馆的经理想知道一年中特定季节里旅馆每天客满的房间数量的变化。为此，在旅馆正常运作的 20 天里，有关人员计算出每天客满房间的样本均值为 290，样本标准差为 30。
    a. 该总体方差的点估计是什么？
    b. 建立以该总体方差的 90% 置信区间估计。
    c. 建立总体标准差的 90% 置信区间估计。

24. 一般而言，股票的原始发行价格(IPOs)是折价的。标准差在折价—溢价指标中用于衡量其分布或变动情况。取 13 项在多伦多交易所交易的加拿大 IPOs 作为样本，其标准差为 14.95。请建立折价—溢价指标的总体标准差 95% 置信区间估计。

25. 下列是不同大城市公务旅行的日均生活费用(*Business Traveler*, June 1995)，下面的费用包括四星级酒店的单人间住宿费、饮料、早餐、出租车费和其他费用：



CD 光盘数据
Travel

城市	日均生活费用(美元)	城市	日均生活费用(美元)
曼谷	242.87	马德里	283.56
巴格达	260.93	墨西哥城	212.00
孟买	139.16	米兰	284.08
开罗	194.19	巴黎	436.72
都柏林	260.76	里约热内卢	240.87
法兰克福	355.36	汉城	310.41

（续表）

城市	日均生活费用(美元)	城市	日均生活费用(美元)
香港	346.32	特拉维夫	223.73
约翰内斯堡	165.37	多伦多	181.25
利马	250.08	华沙	238.20
伦敦	326.76	华盛顿	250.61

a. 计算样本均值。

b. 计算样本标准差。

c. 计算总体标准差的95%置信区间。

26. 零件方差在滚珠轴承的制造中至关重要。滚珠轴承尺寸方差过大会引起滚动失效或磨损过快。生产标准要求轴承尺寸以英寸计量，其最大方差为0.000 1。取15个轴承为样本，样本方差为0.014英尺。

a. 取 $\alpha=0.10$，判断样本方差是否超过生产要求的最大方差。

b. 计算球形轴承总体方差的90%置信区间估计。

27. 原计划粮食袋装量的方差必须小于或等于0.02。取41袋粮食作为样本求得其样本方差为0.16盎司。取 $\alpha=0.05$，判断该粮食袋装方法是否超过原计划的方差标准。

28. 市运输公司承诺严格恪守运输时间，及时为顾客服务。取22次卡车运输递送服务作为样本，求得其样本方差为1.5。检验是否可以拒绝 H_0：$\sigma^2\leqslant 1$，取 $\alpha=0.10$。

29. 取过去6个月中的9天作为样本，一位牙医先后接待了如下列人数的病人：22，25，20，18，15，22，24，19和26。如果每天接待的病人人数呈正态分布，分析以上的样本数据，是否可以拒绝“日均病人数的方差等于10”的假设？取 $\alpha=0.10$，得出你的结论。

30. 得到搭乘某一特定航班乘客的数量，其样本标准差为8。该总体标准差的95%置信区间估计为5.86位乘客—12.62位乘客。

a. 统计分析时是否可使用了容量为10或15的样本？

b. 如果取25次飞行作为样本求得样本标准差为 $s=8$，则总体标准差的置信区间将有什么变化？计算容量为25的95%的置信区间估计(σ)。

31. 在股票交易市场上每天都会有一组价格上涨最大的股票。某天10组纳斯达克股票的标准差为15.8个百分点，而10组纽约股票交易市场股票的标准差为7.9个百分比点（*USA Today*, September 14, 2000)。检验两者的总体方差，两者交易股票的方差是否一致。取 $\alpha=0.10$，得出你的结论。

32. 修完一门大学财会课程的352名学生，他们的总分的标准差为0.940。73名未修完的学生总分标准差为0.797。根据上述数据，判断修完与未修完的学生总分是否一样。取 $\alpha=0.05$(注：自由度 $F_{0.025}$ 为351和72的值为1.466)。

33. 财务部门分析两个生产部门上报的每周成本的方差。各取16份成本报告为样本，两个生产部

门的成本方差分别为 2.3 和 5.4。判断两个生产部门的成本方差是否一样?取 $\alpha=0.10$。

34. 两种装配方法经检验后其装配时间资料具体如下。取 $\alpha=0.10$,检验两个总体方差是否相等。

方法	样本容量	样本方差
A	31	25
B	25	12

案例研究　空军培训

空军电子学引导性课程采用了个人化教学方式。每位学员观看声像讲座后，自学声像教材。各学员独立学习直至培训结束参加考试，关心的问题是不同的学员完成该培训的不同速度。有些学员能够相当快地完成声像教材的学习，而其他学员需要花较长的时间甚至要加班加点才能修完这门课程。学得快的学员必须等学得慢学员学完这门课程后才能一起进行其他培训。

有人建议使用计算机辅助教学。这样的话，所有的学员看同样的声像讲座，然后全都被指派到一个计算机终端接受进一步的培训。这样计算机可以引导学员独立操作，直至完成该课程的自我训练。

为了比较这两种教学方法的优劣，刚入学的 122 名学员被随机安排到两种教学系统中去。61 位学员采用声像教学而另外 61 位采用计算机辅助教学。每位学员学习的时间都记录下来，具体数据如下：



CD 光盘数据
Training

采用声像教学完成教程的时间（小时）

76	76	77	74	76	74	74	77	72	78	73
78	75	80	79	72	69	79	72	70	70	81
76	78	72	82	72	73	71	70	77	78	73
79	82	65	77	79	73	76	81	69	75	75
77	79	76	78	76	76	73	77	84	74	74
69	79	66	70	74	72					

采用计算机教学完成教程的时间(小时)

74	75	77	78	74	80	73	73	78	76	76
74	77	69	76	75	72	75	72	76	72	77
73	77	69	77	75	76	74	77	75	78	72
77	78	78	76	75	76	76	75	76	80	77
76	75	73	77	77	77	79	75	75	72	82
76	76	74	72	78	71					

管理报告

1. 运用恰当的描述性统计方法，总结以下两种方法的培训时间数据，并检验上述样本数据有何异同?
2. 运用第 10 章的方法表述两种方法的总体均值有什么不同? 讨论你的结果。
3. 计算每种培训方法的标准差和方差。进行假设检验，判断两个总体方差是否相等，讨论你的结果。
4. 关于这两种方法的不同，得出你的结论，你认为采用那一种方法较好? 为什么?
5. 对于将来采用哪种培训方法，在决定之前还需哪些资料或者检验?

附录 11.1 用 Minitab 计算总体方差

我们描述了如何应用 Minitab 计算 χ^2 检验统计量和 F 检验统计量的 p - 值。这些 p - 值可用于单个总体方差与两个总体方差的假设检验。

单个总体

我们可以采用第 11.1 节的 St. Louis Metro Bus 公司例子中的数据。C1 列为汽车到达时间。在计算总体方差的假设检验所需的 p - 值之前，我们必须先算出样本方差和相应的 χ^2 检验统计量* $\chi^2 = 10.8$。进行对假设 $H_0: \sigma^2 \leq 4$ 和 $H_a: \sigma^2 > 4$ 的检验，我们可以用 Minitab 计算出如下所需的右侧 p - 值：



CD 光盘数据
MetroBus

步骤 1. 选择 **Calc** 下拉菜单；
步骤 2. 选择 **Probability Distributions**；
步骤 3. 选择 **Chi-Square**；
步骤 4. 弹出 Chi-Square Distribution 对话框时：
选择 **Cumulative probability**；
在 **Degrees of freedom** 中输入“9”；
选择 **Input constant** 并在框中输入“10.8”；
在 **Optional storage** 行中输入 CumProb；
点击 **OK**。
步骤 5. 选择 **Calc** 下拉菜单；
步骤 6. 选择 **Calculator**
对话框出现后：
在 **Store results in variable** 框内输入 p - 值；

*利用 Minitab 计算样本方差 S_2 在第 3 章中讨论，由公式(11.8)给出相应的 χ^2 检验统计可相对容易地由手工计算得出。

在 **Expression** 框中输入 1-CumProb;
点击 **OK**。

当拒绝域位于 χ^2 分布的右侧时，可遵循以上步骤。若假设检验是采用单侧检验（拒绝域位于左侧），则在步骤 6 中的 “Expression” 框中输入 “CumProb”。若假设检验采用双侧检验，当 χ^2 检验统计量位于右侧时，则在步骤 6 中“Expression”框中输入 “2 *（1-CumProb）”；位于左侧则输入 “2 * CumProb”两个总体。

两个总体


CD 光盘数据
Dullus

我们可以采用第 11.2 节中 Dullus 校车的例子中的数据。C1 列中是 Milbank 公司的汽车到达时间，C2 列中是 Gulf Park 公司的汽车到达时间。进行假设 $H_0: \sigma_1^2 = \sigma_2^2$ 和 $H_a: \sigma_1^2 \neq \sigma_2^2$ 检验时可遵循以下 Minitab 步骤：

步骤 1. 选择 **Stat** 下拉菜单;
步骤 2. 选择 **Basic Statistics**;
步骤 3. 选择 **2-Variances**;
步骤 4. 弹出 2-Variances 对话框时:
选择 **Samples in different columns;**
在 **First** 框中输入 C1;
在 **Second** 框中输入 C2;
点击 **OK**。

关于检验的信息可见于题为“F 检验（正态分布）”的部分，其中检验统计量 $F = 2.40$，p - 值 = 0.082。上述 Minitab 方法主要描述了如何双侧检验总体方差是否相等。如果应用 Minitab 方法进行单侧检验，必须牢记单侧区域是双侧 p - 值所涵盖区域的一半，因此计算单侧检验的 p - 值就显得相对容易多。

附录 11.2　用 Excel 计算总体方差

我们描述了如何应用 Excel 计算 χ^2 检验统计量和 F 检验统计量的 p - 值，这些 p - 值可用于单个总体方差与两个总体方差的假设检验。

单个总体

我们可以采用第 11.1 节的 St. Louis Metro Bus 公司例子中的数据。Excel 的 worksheet 中单元格 A1 中列其标题为“时间”，单元格 A2 至 A11 列出了 10 个到达时间。我们可以对 H_0：$\sigma^2 \leq 4$ 和

H_a：$\sigma^2 > 4$ 进行假设检验。使用 Excel 进行该检验的最简单的方法是扩展工作表，然后利用 p - 值得出该结论。图 11.6 表示建立的扩展工作表，每个单元格的内容如下所述：


CD 光盘数据
MetorBus

步骤 1. 在 D2 单元格中输入 4；

步骤 2. 在 D6 单元格中键入样本容量公式；

=COUNT(A2：A11)

步骤 3. 在 D7 单元格中键入样本方差公式；

=VAR(A2：A11)

步骤 4. 在 D8 单元格中键入检验统计量公式；

=(D6 - 1)*D7/D2

步骤 5. 在 D9 单元格中键入 p - 值计算公式。

=CHIDIST(D8，D6 - 1)

Time	Hypothesis Test	
15.2	Test Variance	4
17.5		
19.6		
16.6	Sample Results	
21.3	Sample Size	10
17.1	Sample Variance	4.8
15.0	Test Statistic	10.8
15.5	p-Value	0.290
20.0		
16.2		

图 11.6　St. Louis 汽车例子中假设检验的 Excel 的扩展工作表

D8 单元格包含了计算检验统计量的公式 $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$，

D9 单元格包含了计算 p - 值的公式，该 p - 值代表了基于检验统计量的右侧区域。

如果假设检验是拒绝域位于左侧的单侧检验，则 D9 单元格采用公式：=1 - CHIDIST(D8，D6 - 1)。如果是双侧检验，则计算 p - 值的公式为：=IF(CHIDIST(D8,D6 - 1)< 0.5,2 * (CHIDIST(D8,D6 - 1)),2 * (1 - CHIDIST(D8,D6 - 1)))。

两个总体


CD 光盘数据
Dullus

我们可以采用 11.2 小节中的 Dullus 学校校车的数据。Excel 工作表中的单元格 A1 键入“Milbank”，单元格 B1 键入“Gulf Park”。Milbank 的样本时间列于单元格 A2 和单元格 A26，Gulf Park 的样本时间列于单元格 B2 至单元格 B17。进行假设 H_0：$\sigma_1^2 = \sigma_2^2$ 和 H_a：$\sigma_1^2 \neq \sigma_2^2$ 检验的步骤如下所述：

步骤 1. 选择 **Tools** 下拉菜单;

步骤 2. 选择 **Data Analysis**;

步骤 3. Data Analysis 对话框出现后：选择 **F-Test Two-Sample for Variances**

点击 **OK**。

步骤 4. 弹出 F-Test Two Sample for Variances 对话框时：

在 **Variable 1 Range** 框中输入 A1：A26;

在 **Variable 2 Range** 框中输入 B1：A17;

选择 **Labels**;

在 **Alpha** 框中键入 0.05;

(注意：该 Excel 方法采用 alpha 作为单侧区域)

选择 **Output Range** 并在框中键入 C1;

点击 **OK**。

输出的 P(F≤f) = 0.041 是检验统计量 $F = 2.40$ 时的单侧区域的概率。因此,相应的双侧 p - 值为 2(0.041) = 0.082。如果假设检验是单侧检验，则根据标题为 P(F≤f)中各单元格中所示的单侧区域可以求得相应的 p - 值。

12 拟合度检验和独立性检验

统计实例:UNITED WAY
12.1 多样本总体的拟合度检验
12.2 独立性检验
12.3 拟合度检验:泊松分布与正态分布
泊松分布
正态分布

统计实例

UNITED WAY*

罗切斯特，纽约州

United Way of Greater Rochester 是一个非营利性组织，该组织致力于满足其服务区内 7 个县的居民的最重要的日常需求，以此来提高整个社区的生活质量。

每年春季，由 United Way 红十字会举行募捐活动，所得资金用于 200 个服务机构所提供的几百种服务项目。这些服务机构提供的服务种类繁多，如身体的、精神的、社会的，而且服务对象涵盖很广——包含年龄、背景和经济状况各异的人。

由于有大量自愿者参与，United Way of Greater Rochester 组织运作只花费所筹基金的 80%。

为了了解社区的慈善情况，该组织决定进行调查。他们安排与专业人员、服务人员和一般工作人员这些重点群体的谈话，从而得到关于慈善情况的第一手资料，根据这些资料，设计出调查问卷。问卷经过预审、修订以后分发给 440 个人，得到 323 份有效的问卷。

从搜集到的数据得到各种描述性统计，包括频数分布和交叉表。分析该数据有一项重要的内容是列联表和进行 χ^2 独立性检验。这样的统计量检验的好处之一是确定管理开支观念与职业无关。

独立性检验的假设为：

H_0：United Way 管理开支观念与职业无关

H_a：United Way 管理开支观念与职业有关



统计调查帮助 United Way 调整其方案，以便满足其客户的需求。© Tony Freeman/PhotoEdit.

调查中，两个问题的答案为以上的统计量提供了数据。其中一个问题得到了关于管理开支占资金百分比的数据（10% 以下，11%—20%，21% 以上）。另一个问题是应答者的职业。

进行 χ^2 检验后（取显著性水平为 5%）得到独立性假设被拒绝的结果，因此可以断定关于 United Way 管理开支的观念的确与应答者的职业有关。实际上管理费用小于 9%，但 35% 的反馈认为管理费用为 21% 或更多。其实管理开支的观念不准确，在这群人中，生产线工人、办事员、售货员和职业技术人员的管理开支观念要比其他人模糊。

该调查有助于 United Way 进行相应的项目调整以及筹集资金活动的变动。本章你将学习如何进行如上所述的独立性统计检验。

---

* 感谢 United Way 的市场顾问 Philip. R. Tyler 博士提供该统计应用的资料。

在第 11 章中我们介绍了 χ^2 分布在对单个总体方差进行估计和假设检验中的应用。这里，我们将介绍另外两种基于 χ^2 分布的假设检验方法。与其他假设检验方法一样，这两种检验也是将样本结果与所要求的结果(即当原假设为真时成立)进行比较，假设检验的结论便是确定样本结果与所要求的结果的接近程度。

下面的章节中我们介绍如何对多样本总体进行拟合度检验。接着我们将介绍利用列联表进行独立性检验，然后通过泊松分布和正态分布进行拟合度检验。

12.1　多样本总体的拟合度检验

这一节我们讨论一个总体中的每个个体被归类于几个类别中的一个组且仅被归类于这个组的情况。称该总体为**多样本总体**(multinomial population)。多样本概率分布可以看做是二项分布在三项或三项以上分布的推广。每次多样本检验有且只有一个结果，每次检验都视为独立的，且概率不变。

多样本检验假设的原理与双样本检验一致，惟一不同的是多样本检验每次检验都有三个或三个以上的结果。

比如说，Scott 市场调查公司进行市场份额的调查。在过去的一年里，公司 A 的市场份额保持在 30%，公司 B 保持在 50%，公司 C 保持在 20%。最近公司 C 开发了一种新产品，并取代当前市场上该公司所售产品。Scott 市场调查公司受公司 C 所托进行调查，想检验这种新产品的推广是否影响市场份额。

在这个例子中，检验总体为多样本总体，每位顾客有三种产品选择：公司 A 的产品、公司 B 的产品或公司 C 的产品。由此我们可以得到一个具有三个类别的多样本总体。我们可以使用下列符号来代表三种份额：

P_A = 公司 A 的市场份额；
P_B = 公司 B 的市场份额；
P_C = 公司 C 的市场份额。

Scott 市场调查公司进行了一次样本调查并计算出每个公司产品的销售比例，然后假设检验一下新产品是否改变了市场份额。假设公司 C 的新产品并未改变市场份额，则原假设如下：

H_0：$P_A = 0.30$，$P_B = 0.50$，$P_C = 0.20$
H_a：总体比例不变，即 $P_A = 0.30$，$P_B = 0.50$，$P_C = 0.20$

如果样本结果导致 H_0 被拒绝，则 Scott 市场调查公司就可以得出结论：新产品的引进影响了市场份额分布情况。

假设该市场调查公司对 200 名顾客进行调查。每位顾客在三种产品中挑选一种，结果如下：

频率调查结果		
公司 A 的产品	公司 B 的产品	公司 C 的产品
48	98	54

这 200 名顾客（每位顾客从三种产品中选择一种）相当于包含 200 次检验的多样本检验。

我们现在来进行**拟合度检验**(goodness of fit test)，看看由 200 名顾客组成的样本的**调查**结果是否与原假设一致。拟合度检验的目的是比较调查结果与原假设结果是否一致。因此，根据计算原假设（$P_A = 0.30, P_B = 0.50, P_C = 0.20$），这 200 名顾客应购买三种产品的数量的计算结果如下：

预期频数结果		
公司 A 的产品	公司 B 的产品	公司 C 的产品
$200 \times 0.30 = 60$	$200 \times 0.50 = 100$	$200 \times 0.20 = 40$

由此我们可以看到，把样本容量与原假设的比例相乘便得到每种产品的预期购买频数。

接下来，我们就要对调查所得的购买频数与预期的购买频数进行拟合度检验。如果两者相差过大，则原假设就无法成立。为了检验两者相差大小，我们可以进行如下所述的统计检验：

拟合度的检验统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{12.1}$$

式中 f_i——第 i 种的频数调查结果；

e_i——第 i 种的频数预期结果；

k——类别个数。

注意：该检验统计量是自由度为 $k-1$ 的 χ^2 分布(所有种类的频数期望值不小于 5)。

让我们回到三个公司的市场份额问题上来。由于所有种类的频数期望值大于等于 5，我们可以计算出 χ^2 检验统计量。表 12.1 列出了 Scott 市场调查公司计算 χ^2 检验统计量的情况，我们可以看到检验统计量 $\chi^2 = 7.34$。

表 12.1 **Scott 市场调查公司计算 χ^2 检验统计量的情况**

种类	假设比例	频数调查结果(f_i)	频数期望值(e_i)	差($f_i - e_i$)	差的平方 $(f_i - e_i)^2$	差的平方除以频数期望值 $(f_i - e_i)^2 / e_i$
公司 A	0.30	48	60	-12	144	2.40
公司 B	0.50	98	100	-2	4	0.04
公司 C	0.20	54	40	14	196	4.90
总计		200				7.34

我们原假设该多项总体的比例为 $P_A=0.30$，$P_B=0.50$，$P_C=0.20$，取显著性水平 $\alpha=0.05$。如果频数调查结果与频数期望值相差太大，则拒绝原假设，因而在 χ^2 分布右侧建立面积为 0.05。对照 χ^2 分布表(附录 B 的表 3)，我们可知自由度为 $k-1=3-1=2$ 时，$\chi^2_{0.05}=5.99$。由于 $7.34>5.99$，我们拒绝原假设 H_0。因此我们可以得出结论：公司 C 推出的新产品改变了当前市场份额结构。

使用计算机软件包(如 Minitab 和 Excel)我们采用 p－值进行拟合度计算和拟合度检验。这些方法的具体步骤见附录 12.1 和 12.2。通过使用 Minitab，我们可知 Scott 市场调查公司的假设检验的 p－值为 0.025。根据拒绝规则（当 p－值$<\alpha$ 时，拒绝 H_0），我们可知 $0.025<\alpha=0.05$，因此，H_0 被拒绝，我们得出结论，公司 C 推出的新产品改变了当前的市场份额结构。

---

p－值方法适用于拟合度检验。

拟合度检验一般是推断位于分布右侧的单侧检验。

虽然除此之外我们无法得出其他结论，但我们可以通过比较频数调查值和频数期望值来看看市场份额结构将发生什么样的变化。以公司 C 为例，我们可以发现调查值 54 比期望值 40 大，因为调查值是对当前市场份额进行调查的结果，可见新产品对公司 C 的市场份额发生了积极影响。从其他两家公司的两值比较可见，公司 C 市场份额的增加对公司 A 的影响比公司 B 大。

对任何多项分布进行拟合度检验的一般步骤总结如下：

总结：多项分布的拟合度检验

1. 建立原假设和备择假设。

H_0：对 K 类中的每一个特定概率，总体服从多元正态分布
H_a：对 K 类中的每一个特定概率，总体不服从多元正态分布

2. 选取一个任意样本，记录每个类别的频数调查值 f_i。
3. 假定原假设成立，把各个类别比例与样本容量相乘求得每个类别的频数期望值 e_i。
4. 计算检验统计量的值：

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}$$

5. 拒绝规则：　使用检验统计量：　当 $\chi^2>\chi^2_\alpha$ 时，拒绝 H_0；
　　使用 p－值：　　当 p－值$<\alpha$ 时，拒绝 H_0。

式中，α 为检验时的显著性水平，自由度为 $k-1$。

练 习

方 法

1. 用 χ^2 拟合度检验方法检验下面的假设：



H_0：$p_A=0.40$，$p_B=0.50$，$p_C=0.20$

H_a：总体比例不是 H_0，即 $p_A=0.30$，$p_B=0.50$，$p_C=0.20$

取样本容量为 200 的样本，可取的类别 A 的频数为 60，类别 B 的频数为 120，类别 C 的频数为 20。取 $\alpha=0.01$，检验 H_0 是否成立。

2. 假设一个多项总体有 4 个类别：A、B、C、D，原假设中各类别比例相同。

$$H_0: p_A=p_B=p_C=p_D=0.25$$

取容量为 300 的样本，频数分别为：A：85，B：95，C：50，D：70。取 $\alpha=0.05$，检验 H_0 是否成立。

应 用

3. 在前 13 周，有人调查了星期六晚上 8:00 到 9:00 各个电视节目的收视率，结果如下：ABC 29%，CBS 28%，NBC 25%，其他 18%。在接下的两个星期的星期六晚上，分别抽取 300 个家庭，得出收视情况：观看 ABC 的有 95 家，CBS 有 70 家，NBC 有 89 家，其他有 46 家。取 $\alpha=0.05$，检验各节目的收视率是否有变化？



4. M&M 巧克力饼干制造商 M&M/MARS 进行一次全国性民意调查，调查对象逾千万，调查内容为喜欢什么样的饼干颜色。调查结果为 M&Ms 古铜色饼干需换成蓝色饼干。M&M/MARS 设计出一本小册子《颜色》，上面列出了各种颜色饼干的喜爱程度。

棕色	黄色	红色	橙色	绿色	蓝色
30%	20%	20%	10%	10%	10%

杂志 Chance(no. 4, 1996)上有一份调查报告，调查目的是检验以上数据是否属实。下面是抽取 506 块饼干的调查结果：

棕色	黄色	红色	橙色	绿色	蓝色
177	135	79	41	36	38

取 $\alpha=0.05$，检验各种颜色的饼干喜爱程度是否如 M&M/MARS 公司所示？

5. 1996 年 *Business Week* 的订阅者中进行调查，其中一份问卷有一道题：“商业投资时，你选择全服

务性的中介公司还是打折的中介公司?"调查结果为:有 264 名答卷选择全服务性中介公司,255 名选择佣会的中介公司,还有 229 位两者都选。取 $\chi = 0.10$,检验三种服务在选择上是否有差异?

6. 广告中，反面宣传是一种很有效的广告方式。*The Journal of Advertising*(Summer 1997)上有一篇报道，介绍 24 份杂志上反面宣传广告的分析结果。下面是各杂志上反面广告的数目:

杂志类别	反面广告的数目
新闻评论	20
一般期刊	15
家庭类	30
商业/金融	22
妇女类	16
黑人杂志	12

取 $\alpha = 0.10$，检验以上 6 种杂志的反面广告所占的比重是否有差异?

7. 广大消费者对商店橱窗陈列方式喜好如下:

陈列方式 A	陈列方式 B	陈列方式 C
43	53	39

取 $\alpha = 0.05$，检验消费者对这三种陈列方式喜好是否有差异?

8. 航空公司的服务质量如何?某一份调查结果显示：非常好 3%，好 28%，一般 45%，差 24%(*Business Week*, September11, 2000)。在另一份对电信公司的调查报告中，他们选择 400 名成年人作为调查对象，结果为：非常好有 24 人，相当好有 124 人，一般有 172 人，差 80 人。航空公司的服务质量反馈结果与电信公司的结果是否有差异?取 $\alpha = 0.01$，检验两家公司的评价比例情况是否相同?

12.2　独立性检验

χ^2 分布的另一项重要应用是根据样本数据检验两个变量的独立性。以亚里桑那州的 Alber's Brewery of Tucson 公司为例，具体阐述如何进行独立性检验。Alber's 公司生产并销售三种啤酒：低度啤酒、中度啤酒和高度啤酒。为了分析这三种啤酒的市场份额，该公司的市场调查小组提出一个问题：男性饮酒者与女性饮酒者对三种啤酒的选择是否因性别而异?如果啤酒类别的选择与性别无关，则 Alber's 公司只需为所有的啤酒做同一广告；但是，如果啤酒类别与饮酒者性别有关，那么该公司就得针对不同的市场做不同的广告：

H_0：啤酒选择与饮酒者性别无关（相互独立）

H_a：啤酒选择与饮酒者性别有关（相互不独立）

表 12.2 可看出该例子的情形。把一个总体分为男性饮酒者和女性饮酒者两类，我们可以抽取一些人作为样本进行调查，让每个人在三种啤酒中进行选择。每个人被归入表格中的一个单元格中。比如说，一位男性选择中度酒[属于单元格(1, 2)]，一位女性选择低度酒[属于单元格(2, 1)]，

一位女性选择高度酒[属于单元格(2，3)]，等等。因为我们已经列出了所有可能的啤酒类型与性别的组合，话句话说，列出了所有列联。表 12.2 就称为**列联表**(contingency table)。独立性检验要用到列联表，因此有时也称为**列联表检验**。

假设我们任意选取 150 位饮酒者。当尝过每种酒后，这些人可以在其中选择。表 12.3 总结了调查结果。我们可以看到，为了进行独立性检验，我们必须收集每个单元格(或每个选择种类)的数目(或频数)。在这 150 人当中，有 20 位男性选择高度啤酒，有 40 位男性选择中度啤酒，等等。

表 12.2　啤酒选择与饮酒者性别的列联表

		啤酒选择		
		低度啤酒	中度啤酒	高度啤酒
性别	男	单元格 (1，1)	单元格 (1，2)	单元格 (1，3)
	女	单元格 (2，1)	单元格 (2，2)	单元格 (2，3)

表 12.3 中的数据是这 6 种类别的调查所得频数。如果我们可以假设当啤酒选择与饮酒者性别无关(相互独立)时的标准频数(预期频数)，我们就可以根据其 χ^2 分布得知调查结果与原假设是否有明显的差别。列联表中每个单元格的预期频数如下所述。首先我们假定啤酒选择与饮酒者性别无关(相互独立)的原假设成立。

表 12.3　啤酒选择与饮酒者性格检验的结果(预期频数)

		啤酒选择			
		低度啤酒	中度啤酒	高度啤酒	总计
性别	男	20	40	20	80
	女	30	30	10	70
	总计	50	70	30	150

为了检验两个方差是否相互独立，必须抽取一个样本，然后使用交叉列表总结两个方差的相关数据。

然后我们注意到这 150 位饮酒者中，总共有 50 位选择低度啤酒，70 位选择中度啤酒，30 位选择高度啤酒。通过分式我们可以得知有 50/150 = 1/3 的饮酒者选择低度啤酒，有 70/150 = 7/15 的饮酒者选择中度啤酒，有 30/150 = 1/5 的饮酒者选择高度啤酒。如果独立性假设成立，则我们可知这些分式对男性饮酒者和女性饮酒者都成立。这样的话，根据原假设，80 位男性饮酒者中有 (1/3)80 = 26.67 位选择低度啤酒，有 (7/15)80 = 37.33 位选择中度啤酒，还有 (1/5)80 = 16位选择高度啤酒，同理可得女性饮酒者的预期频数(如表 12.4)。

表 12.4　啤酒选择与饮酒者性别无关(相对独立)时的预期频数

		啤酒选择			
		低度啤酒	中度啤酒	高度啤酒	总计
性别	男	26.67	37.33	16.00	80
	女	23.33	32.67	14.00	70
	总计	50.00	70.00	30.00	150

我们设 e_{ij} 为列联表中第 i 行第 j 列的预期频数，据此，男性饮酒者(第一行)选择中度啤酒(第 2 列)，则其预期频数为 e_{12}。根据前面所述的预期频数的算法，我们可以得出

$$e_{12} = (7/15)80 = 37.33$$

请注意上式的 80 是所有男性饮酒者的总数(第 1 行的总数)，70 是所有选择中度啤酒者的总数(第 2 列的总数)，150 是整个样本的容量。由此我们可以看出

$$e_{12} = \frac{(\text{第1行总数})(\text{第2行总数})}{\text{样本容量}}$$

推广一下以上的公式我们可以得到计算独立性检验的列联表中的预期频数的公式：

独立性假设下列联表中预期频数

$$e_{ij} = \frac{(\text{第}i\text{行的总数})(\text{第}j\text{列的总数})}{\text{样本容量}} \tag{12.2}$$

根据以上公式计算男性饮酒者选择高度啤酒的频数，我们可得预期频数为 $e_{13} = (80)(30)/150 = 16.00$，如表 12.4 所示。运用公式(12.2)检验一下表 12.4 中的其他预期频数。

比较表 12.3 中的调查所得频数和表 12.4 中的预期频数的检验方法与 12.1 节中介绍的拟合度检验相似。基于调查所得频数和预期频数的 χ^2 计算公式如下：

独立性检验统计量

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{12.3}$$

式中　f_{ij}——(第 i 行第 j 列)列联表中第 i 行第 j 列所示种类的调查所得频数；

e_{ij}——独立性假设下列联表中第 i 行第 j 列所示种类的预期频数。

注意：由于列联表有 n 行 m 列，所以当所有种类的预期频数都大于等于 5 时，检验统计量 χ^2 为自由度为 $(n-1)(m-1)$ 的 χ^2 分布。

公式 12.3 中有两次累加，意思是列联表中所有的单元格都参与计算。

由表 12.4 中我们可以看到，每个种类的预期频数都大等于 5。因此我们可以计算检验统计量 χ^2，表 12.5 列出了检验统计量 χ^2 的计算过程，这个检验统计量 χ^2 可用于检验啤酒选择是否与饮酒者的性别相对独立(无关)，我们可以看到检验统计量 $\chi^2 = 6.13$。

表 12.5 检验啤酒选择是否与饮酒者性别相对独立的检验统计量 χ^2 的计算过程

性别	啤酒	调查所得	预期频数	两者之差	差的平方	差的平方除以预期
男	选择	频数(f_{ij})	(e_{ij})	($f_{ij}-e_{ij}$)	$(f_{ij}-e_{ij})^2$	频数$(f_{ij}-e_{ij})^2/e_{ij}$
男	低度	20	26.67	-6.67	44.49	1.67
男	中度	40	37.33	2.67	7.13	0.19
女	高度	20	16.00	4.00	16.00	1.00
女	低度	30	23.33	6.67	44.49	1.91
女	中度	30	32.67	-2.67	7.13	0.22
女	高度	10	14.00	-4.00	16.00	1.14
	总计	150				6.13

相应的 χ^2 分布的自由度可以通过行数减 1 乘以列数减 1 而得。因此，我们可以求得啤酒选择与饮酒者性别的独立性检验的自由度为$(2-1)(3-1)=1\times2=2$ 度。取 $\alpha=0.05$，附录 B 中的表 3 可查得右侧临界 χ^2 值为$\chi^2_{0.05}=5.99$。请注意我们再次用到了右侧临界值，因为我们只有在“调查所得频数”和“预期频数”的差别过大导致 χ^2 值过大时推翻原假设。在上面的例子里，$\chi^2=6.13$ 大于临界值$\chi^2_{0.05}=5.99$。因此，我们推翻原假设，并得到结论：啤酒选择与饮酒者性别有关(相对不独立)。

独立性检验常使用推翻域位于分布右侧的单侧检验。

Minitab 或 Excel 计算机软件包可简化独立性检验得计算过程，并且可以提供检验所需的 p 值 $=0.047$ 小于 $\alpha=0.05$，所以我们推翻原假设，得到正确结论：啤酒选择与饮酒者性别有关(相对不独立)。图 12.1 展示了 Alber's Brewery 独立性检验的 Minitab 输出结果。p-值 $=0.047$，小于 $\alpha=0.05$ 的区域。因此我们拒绝 H_0，并推断啤酒选择与性别有关。

```
Expected counts are printed below observed counts

          Light     Regular      Dark     Total
    1        20          40        20        80
          26.67       37.33     16.00

    2        30          30        10        70
          23.33       32.67     14.00

Total        50          70        30       150
DF = 2, P-Value = 0.047
```

图 12.1 Alber's Brewery 独立性检验的 Minitab 输出结果

虽然由检验我们无法得到进一步的结论，但我们可以比较一下调查所得频数由此得到一个关于啤酒选择与饮酒者性别独立性的概念。由表 12.3 和 12.4 我们可以看到，男性饮酒者选择中度啤酒和高度啤酒的调查所得频数比预期频数相对要高，而女性饮酒者只有选择低度啤酒时调查所

得频数比预期频数高。这些调查结果能让我们洞悉男女饮酒者在啤酒选择上的不同。让我们总结一下列联表独立性检验的步骤：

独立性检验：小结

1. 建立原假设和对立假设。

H_0：列方差与行方差不相对独立

H_a：列方差与行方差不相对独立

2. 选取样本，记录列联表每个单元格的调查所得频数。
3. 由公式(12.2)计算出每个单元格的预期频数。
4. 由公式(12.3)计算出检验统计量的值。
5. 拒绝法则：使用检验统计量：

当 $\chi^2>\chi_\alpha^2$ 时，推翻 H_0

使用 p－值：当 p 值$<\alpha$ 时，推翻 H_0

上式中 α 为显著偏差水平，列联表有 n 行 m 列，自由度为 $(n-1)(m-1)$。

评　注

本章的 χ^2 检验的检验统计量要求每个种类的预期频数为 5，如果有 5 个种类的预期频数小于 5，我们常常将相近的两个种类结合起来，以使得每个种类的预期频数都大于等于 5。

练　习

方　法

9. 下面的 2×3 列表列出了一个容量为 200 得样本的调查所得频数。取 $\alpha=0.025$，进行一下 χ^2 检验，检验行方差与列方差的独立性。

自测题

	列方差		
行方差	A	B	C
P	20	44	50
Q	30	26	30

10. 下面的 3×3 列联表列出了容量为 240 得样本的调查所得频数。取 $\alpha=0.05$，进行一下 χ^2 检验，检验行方差与列方差的独立性。

	列方差		
行方差	A	B	C
P	20	30	20
Q	30	60	25
R	10	15	30

应 用

11. 1996年《商业周刊》订阅者调查问卷中有一道题是:“在过去的12个月中,你出差时最常买的飞机票是哪一种?”下面的列联表列出了相关数据:



	航班类别	
飞机票种类	国内航班	国际航班
一等舱	29	22
商务行政类	95	121
全价经济舱	518	135

取 $\alpha=0.05$，检验一下航班种类与飞机票种类的独立性，结论如何?

12. 在一次汽车业的品牌效应调查中，购买新车的顾客们对“您的新车厂家是否与过去的车的产家相同?”这个问题做出回答。600位回答者的调查结果显示，他们对国产车、欧洲车和亚洲车的钟情情况。

	生产厂家		
所买的新车与旧车	国产车	欧洲车	亚洲车
同一厂家	125	55	68
不同厂家	140	105	107

a. 假设检验一下品牌选择与厂家是否相对独立。取 $\alpha=0.05$。

b. 若两者差别很大，哪个厂家的品牌最有吸引力?

13. 商业与工程专业的毕业生刚开始工作时所分配的岗位根据行业的不同分类如下:

	行业			
专业(本科)	石油	化学	电子	计算机
商业	30	15	15	40
工程	30	30	20	20

取 $\alpha=0.01$，检验一下专业与行业的独立性。

14. Marist民意调查机构进行了一次主题为“为谁买礼物最难”的调查，调查对象分为男女两类(*USA Today*, December 15, 1997)。假设调查100名男士和100名女士后的数据如下所示:

	性别	
为___买礼物最难	男	女
配偶	37	25
父母	28	31
孩子	7	19
兄弟姐妹	8	3
姻亲	4	10
其他亲属	16	12

取 $\alpha=0.05$，检验一下送礼对象与送礼者性别的独立性。

15. 反面宣传据说是一种有效的广告方式。*The Journal of Advertising*(Summer 1997)上报道了一项调查，调查内容是 24 份杂志上恐怖宣传广告的数量。调查结果如下所示：

	吸引力的类型	
杂志类型	反面宣传广告个数	恐怖宣传广告个数
新闻评论	20	10
一般评论	15	11
家庭类	30	19
商业/金融	22	17
妇女类	16	14
黑人杂志	12	15

取 $\alpha=0.01$，根据以上数据进行独立性检验。

16. 越来越多的商业组织开始采用网上订货的形式进行交易。Performance Measurement Group 调查了一下各行业电子订单填写正误情况，假定 700 份电子订单经调查所得结果如下所示：

	行业			
订单填写情况	制药业	消费品	计算机	电信
正确	207	136	151	178
错误	3	4	9	12

a. 假设检验订单填写情况与行业的独立性。取 $\alpha=0.05$，结论如何？

b. 哪个行业订单填写正确率最高？

17. 三家供应商提供了如下所示的零件质量数据：

	部件质量		
供应商	良好	小缺陷	大缺陷
A	90	3	7
B	170	18	7
C	135	6	9

取 $\alpha=0.05$，检验一下供应商与零件质量的独立性。你的分析结果可以为采购部提供什么信息？

18. 选民的受教育水平与他们所属党派有什么关系？下面是调查结果：

	党派关系		
受教育水平	民主党	共和党	独立党
高中以下	40	20	10
高中	30	35	15
大学	30	45	25

取 $\alpha=0.01$，检验一下选民的受教育水平与所属党派的独立性。

19. 在辛迪加电视节目中，两大影评人 Siskel 和 Ebert 对 160 部影片的评价情况[评价分为“Pro”(称赞)、“Con”(贬斥)和“Mixed”(褒贬皆有)]。

	Ebert 的评价		
Siskel 的评价	**贬斥**	**褒贬皆有**	**称赞**
贬斥	24	8	13
褒贬皆有	8	13	11
称赞	10	9	64

取 $\alpha=0.01$，根据以上数据进行独立性检验，得出结论。

12.3 拟合度检验：泊松分布与正态分布

在 12.1 小节中我们介绍了多项总体的拟合度检验。一般来说，拟合度检验适用于任何假设的概率分布。本节我们将具体阐述总体假设为泊松分布或正态概率分布时的拟合度检验。我们将看到，本节中的拟合度检验与检验过程中 χ^2 分布的运用和 12.1 小节中的拟合度检验步骤相同。

泊松分布

当总体服从泊松分布时，我们来阐述一下这时的拟合度检验。我们以佛罗里达州拉哈西的 Dubek 食品市场顾客光临情况作为例子。由于最近的员工分配问题，Dubek 的经理请当地的咨询公司协助进行结算通道人员配置规划。查访了结算通道的运营情况后，该咨询公司提供了一项人员配置规划的建议。该配置方法基于对待顾客队伍的数学分析，只有在一段特定时间内到达的顾客数服从泊松分布时才适用。就是说，决定这项人员配置方法是否采用之前，还得收集一下顾客到达情况的资料，然后进行一下统计检验看顾客到达数是否服从泊松分布。

我们用 5 分钟进入商店的顾客数来规定到达的**顾客数**，然后我们可以进行如下假设：

H_0：每 5 分钟内进入商店的顾客数呈泊松分布

H_a：每 5 分钟内进入商店的顾客数不呈泊松分布

如果取样检验后 H_0 不能被推翻，则 Dubek 就可以采用咨询公司提议的人员配置方法。如果 H_0 被推翻，则顾客光顾情况不呈泊松分布，Dubek 就必须考虑其他人员配置方法了。

为了检验工作日上午顾客到达数是否呈泊松分布，一名商店雇员随机抽取了 3 周，以每个工作日上午共 128 个 5 分钟时间段的顾客数为样本，该雇员记下了每 5 分钟到达的顾客数。总结数据时，他归结出没有顾客到达的 5 分钟时间段的个数，有一位顾客到达的 5 分钟时间段的个数，有 2 位顾客到达的 5 分钟时间段的个数……具体数据可见表 12.6。

表 12.6 128 个 5 分钟时间段顾客到达数调查结果

顾客到达数	调查所得频数
0	2
1	8
2	10
3	12
4	18
5	22
6	22
7	16
8	12
9	6
总计	128

表 12.6 给出了 10 个类别的调查频数。我们要用拟合度检验判断一下这个由 128 个时间段组成的样本是否说明了原假设成立。为了进行拟合优度检验，我们必须考虑原假设成立时 10 个类的期望频数。即如果顾客到达数呈泊松分布，这时候“没有顾客”、“1 个顾客”、“2 个顾客”……各类别的时间段个数的期望值是多少呢?

我们在第 5 章首次引入的泊松概率函数为

$$f(x)=\frac{\mu^x e^{-\mu}}{x!} \tag{12.4}$$

在该函数中，μ 代表每个 5 分钟时间段内到达的顾客数的调查值，$f(x)$代表在一个 5 分钟时间段内到达的顾客数为 x 的概率。

根据公式(12.4)计算泊松概率之前，我们必须知道 5 分钟时间段内到达顾客数的均值 μ 的估值。表 12.6 中得样本均值给出了该估值。我们可以看到有 2 个时间段没有顾客、8 个时间段有一个顾客等等，因此 128 个时间段的到达顾客总数为 $0\times2+1\times8+2\times10+\cdots+9\times6=640$。由时间段数为 128，到达顾客总数为 640，我们可得每个时间段的顾客平均数 μ 为 $640/128=5$。根据这个泊松概率分布均值，Dubek 食品市场的泊松概率函数的估值为

$$f(x)=\frac{5^x e^{-5}}{x!} \tag{12.5}$$

可代入不同的 x 值，计算出每个类别的概率。这些概率值可见表 12.7(也可在附录 B 的表 7中找到)。比如说，5 分钟时间段内没有顾客到达(顾客数为 0)的概率为 $f(0)=0.006\,7$，5 分钟时间段内有一个顾客到达的概率为 $f(1)=0.033\,7$，等等。如 12.1 小节所述，每个类别的期望频数等于概率乘以样本容量。举例说，没有顾客(0 个顾客)的时间段个数的期望值 $0.006\,7\times128=0.857\,6$。有一个顾客时间段个数的期望值为 $0.033\,7\times128=4.313\,6$。

在我们运用惯例的 χ^2 计算方法比较调查频数和期望频数之前，我们注意到表 12.7 有 4 个种

表 12.7　Dubek 顾客到达的期望频数(设为 =5 的泊松分布)

到达顾客数	泊松概率 $f(x)$	5 分钟时间段内顾客数为 x 的时段个数期望值，即 $128f(x)$
0	0.006 7	0.857 6
1	0.033 7	4.313 6
2	0.084 2	10.777 6
3	0.140 4	17.971 2
4	0.175 5	22.464 0
5	0.175 5	22.464 0
6	0.146 2	18.713 6
7	0.104 4	13.363 2
8	0.065 3	8.358 4
9	0.036 3	4.646 4
10 或以上	0.031 8	4.070 4
	总计	128.000 0

类的期望频数值小于5，这就违反了 χ^2 分布的使用条件。然而，我们可以合并相邻的种类以满足“期望频数值至少等于5”的条件。就是说，我们可以把“0”和“1”归为一类，“9”和“10”归为另一类。这样的话，每个种类都能满足“期望频数至少等于5”的条件。表12.8是合并以后的调查频数和期望频数。

与12.1小节介绍过的一样，拟合度检验重点在于检验调查频数与期望频数的差 $(f_i - e_i)$。因此根据表12.8的调查频数和期望频数，我们可以计算出 χ^2 检验统计量 $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$，计算 χ^2 检验统计量的值为 $\chi^2 = 10.976\,6$。

表12.8 合并后的Dubek顾客到达数的调查频数和期望频数

顾客到达数	调查频数(f_i)	期望频数(e_i)
0或1	10	5.171 2
2	10	10.777 6
3	12	17.971 2
4	18	22.464 0
5	22	22.464 0
6	22	18.713 6
7	16	13.363 2
8	12	8.358 4
9或大于9	6	8.716 8
总计	128	128.000 0

总的来说，拟合度检验的 χ^2 分布自由度为 $k - p - 1$，k 为种类个数，p 为由样本数据得到的总体参数。对于我们所研究的泊松分布拟合度检验来说由表12.9可知 $k = 9$，泊松分布的均值 $p = 1$。因此，Dubek食品市场例子中的 χ^2 分布的自由度为 $k - p - 1 = k - 2 = 9 - 2 = 7$。

表12.9 Dubek食品市场例子中的检验统计量 χ^2 的计算过程

顾客到达数(x)	调查频数(f_i)	期望频数(e_i)	差($f_i - e_i$)	差的平方$(f_i - e_i)^2$	差的平方除以期望频数$(f_i - e_i)^2 / e_i$
0或1	10	5.171 2	4.828 8	23.317 3	4.509 1
2	10	10.777 6	-0.777 6	0.604 7	0.056 1
3	12	17.971 2	-5.971 2	35.655 2	1.984 0
4	18	22.464 0	-4.464 0	19.927 3	0.887 1
5	22	22.464 0	-0.464 0	0.215 3	0.009 6
6	22	18.713 6	3.286 4	10.800 4	0.577 1
7	16	13.363 2	2.636 8	6.952 7	0.520 3
8	12	8.358 4	3.641 6	13.261 3	1.586 6
9或大于9	6	8.716 8	-2.716 8	7.381 0	0.846 8
总计	128	128.000 0			10.976 6

假定我们现在对“顾客到达数的概率分布呈泊松分布”的原假设进行检验，取 $\alpha=0.05$。由附录 B 的表 3 可知自由度为 7 时，$\chi^2_{0.05}=14.07$。如果进行的是单侧检验，当 $\chi^2>\chi^2_\alpha$ 时，我们将推翻 H_0。

回顾前面的计算结果，我们可以得知 $\chi^2=10.976\,6$。这个值小于临界值 14.07。因此我们不能推翻原假设。就是说，经分析，工作日上午顾客到达数呈泊松概率分布的假设不能被推翻。这个统计结果也表明 Dubek 的经理可以采用咨询公司提议的工作日早上的人员配置方法。

附录 12.1 和 12.2 是应用 Minitab 和 Excel 进行以上拟合度检验的详细步骤。应用 Minitab 时，p－值 $=0.14$。因为 p－值$>\alpha=0.05$，所以我们不能推翻总体呈泊松分布的原假设。泊松分布的拟合度检验的步骤总结如下：

泊松分布的拟合度检验：小结

1. 建立原假设和对立假设。

H_0：总体呈泊松概率分布

H_a：总体不呈泊松概率分布

2. 随机抽取一个样本，然后：
   a. 记录每个泊松随机变量的调查频数 f_i；
   b. 计算频数均值 μ。
3. 计算每个泊松随机变量的期望频数，将样本容量与泊松随机变量的泊松概率相乘。若有些变量值小于 5，合并相临的种类，必要的话可以减少种类个数。
4. 计算检验统计量的值

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}$$

5. 拒绝法则：

使用检验统计量：　当 $\chi^2>\chi^2_\alpha$ 时，推翻 H_0；

使用 p－值：　　当 p－值$<\alpha$ 时，推翻 H_a。

式中，α 为显著性水平，自由度为 $k-2$。

正态分布

正态概率分布的拟合度检验同样基于 χ^2 分布的运用，与我们讨论过的泊松分布的步骤相似。即假设总体呈正态分布，然后将样本数据中几个种类的调查频数与预期频数进行比较。因为正态概率分布是延续性的，因此我们必须改变种类的定义方式和预期频数的计算方法。为此，我们来看一下 Chemline 公司的求职检测数据(表 12.10)，接着我们将具体阐述正态概率分布的拟合度检验。

Chemline 公司每年为遍布全美的 4 家分工司招聘大约 400 名雇员。人事部的董事长想知道检测分数总体是否呈正态分布。如果是的话，就可以根据这样的分布评估出具体的分数。就是说可

表 12.10 Chemline 公司随机抽取 50 位求职者的资质检测分数

71	66	61	65	54	93
60	86	70	70	73	73
55	63	56	62	76	54
82	79	76	68	53	58
85	80	56	61	61	64
65	62	90	69	76	79
77	54	64	74	65	65
61	56	63	80	56	71
79	84				

以很快地算出前 20% 人的分数，后 40% 人的分数等等。为此，我们必须检测原假设，看看检测分数具体是否呈正态概率分布。

我们先运用表 12.10 中的数据算出原假设必需的均值、估值和正态分布的标准差估值。我们用样本均 $\bar{x}$ 值和样本标准差 s 作为正态分布的均值、估值和标准差估值。计算过程如下：

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3\ 421}{50} = 68.42$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5\ 310.036\ 9}{49}} = 10.41$$

运用以上数值，我们可以如下陈述求职者检测分数分布的假设：

H_0：检测分数总体呈正态分布(均值为 68.42，标准差为 10.41)

H_a：检测分数总体不呈正态分布(均值为 68.42，标准差为 10.41 的正态分布)

所假设的正态分布如图 12.2 所示。

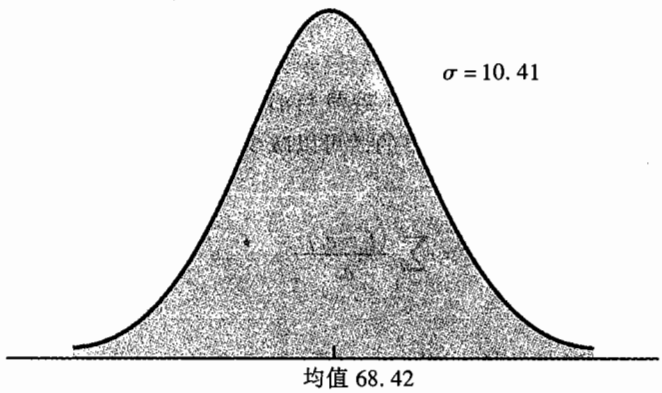


图 12.2 Chemline 公司求职者检测分数的假设正态分布图

现在让我们考虑一下正态分布拟合度检验中定义种类的方法。泊松分布检验时，在其离散概率分布中，各种类很容易通过顾客到达数来定义，如 0，1，2，等等。然而，在延续性正态分布中，我们要运用不同的方法定义种类。我们需要通过检测分数的**区间**来定义种类。

我们可以回想起“每个区间或种类的期望频数至少要等于 5”的前提，所以我们定义检测分数的种类时也必须使每个种类的期望频数至少为 5。取容量为 50 的样本，定义方法之一是把正态分布分为 10 个概率相等的区间(见图 12.3)。由于样本容量为 50，每个种类(或区间)可得 5 个结果，这样就满足了“期望频数”的前提条件。

根据连续概率分布，建立的区间应满足：每一区间都有一个期望频率为 5 或更多的范围。

我们再研究一下计算种类临界值的步骤。假设总体呈正态分布，我们可以根据标准正态概率表来确定这些临界值。首先看一下较低的 10% 检测分数。由附录 B 中的表 1 我们可以找到该检测分数的 z 值为 -1.28，因此检测分数 $x = 68.42 - 1.28 \times 10.41 = 55.10$ 就是较低的 10% 分数的切断临界值。求较低 20% 的分数，我们找到 $z = -0.84$，则 $x = 68.42 - 0.84 \times 10.41 = 59.68$，同理对整个正态分布，我们可以求得以下检测分数值：

较低 10%	$68.42 - 1.28(10.41) = 55.10$
较低 20%	$68.42 - 0.84(10.41) = 59.68$
较低 30%	$68.42 - 0.52(10.41) = 63.01$
较低 40%	$68.42 - 0.25(10.41) = 65.82$
中等分数	$68.42 + \quad 0\ (10.41) = 68.42$
较高 40%	$68.42 + 0.25(10.41) = 71.02$
较高 30%	$68.42 + 0.52(10.41) = 73.83$
较高 20%	$68.42 + 0.84(10.41) = 77.16$
较高 10%	$68.42 + 1.28(10.41) = 81.74$

这些区间切断临界值可见图 12.3。

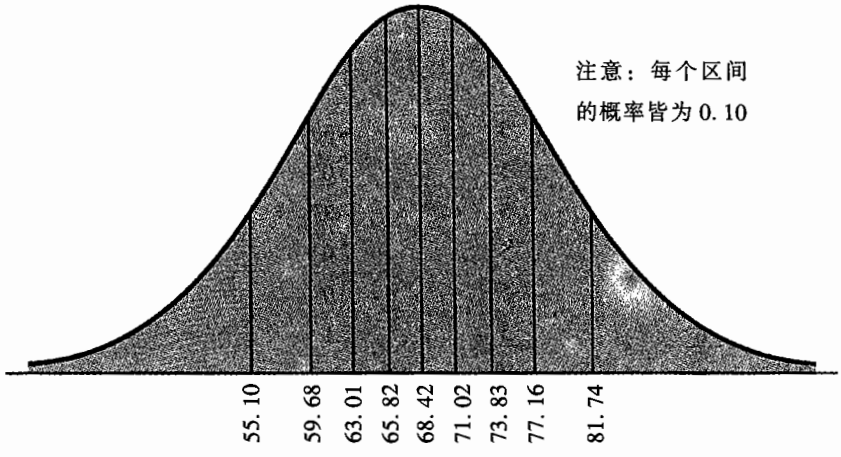


图 12.3　Chemline 公司例子中切断为 10 个概率相等的区间的正态概率分布

注意：每个区间的概率皆为 10 检测分数(考试分数)的种类或区间定义好了，每个种类的预期频数为 5 也知道了，现在我们可以回到表 12.10 的样本数据中确定一下各种类的调查频数。之后我们可得结果如表 12.11 所示。

有了表 12.11 的结果，接下来的拟合优度检验的计算过程与以前一模一样。就是说，我们计算出一个 χ^2 值然后比较调查频数和预期频数。χ^2 检验统计量的必要计算步骤见表 12.12，我们可

表 12.11 Chemline 公司求职者检测分数的调查频数和预期频数

检测分数区间	调查频数（f_i）	预期频数（e_i）
小于 55.10	5	5
55.10—59.68	5	5
59.68—63.01	9	5
63.01—65.82	6	5
65.82—68.42	2	5
68.42—71.02	5	5
71.02—73.83	2	5
73.83—77.16	5	5
77.16—81.74	5	5
81.74—结束	6	5
总计	50	50

表 12.12 Chemline 公司求职者的 χ^2 检验统计量的计算步骤

检测分数区间	调查频数(f_i)	预期频数(e_i)	差($f_i - e_i$)	差的平方 $(f_i - e_i)^2$	差的平方除以预期频数 $(f_i - e_i)^2 / e_i$
小于 55.10	5	5	0	0	0.0
55.10—59.68	5	5	0	0	0.0
59.68—63.01	9	5	4	16	3.2
63.01—65.82	6	5	1	1	0.2
65.82—68.42	2	5	−3	9	1.8
68.42—71.02	5	5	0	0	0.0
71.02—73.83	2	5	−3	9	1.8
73.83—77.16	5	5	0	0	0.0
77.16—81.74	5	5	0	0	0.0
81.74—结束	6	5	1	1	0.2
总计	50	50			7.2

以看到检验统计量 $\chi^2 = 7.2$。

为了确定 χ^2 值 7.2 是否是以推翻原假设 H_0，我们需要参照相应的 χ^2 概率分布表。根据拟合度检验自由度的计算规则，我们可求得自由度为 $k - p - 1 = 10 - 2 - 1 = 7$，[种类个数 $k = 10$，由样本数据估计而得参数个数 $p = 2$(均值和标准差)]。取 $\alpha = 0.10$ 进行假设检验，我们可得右侧推翻域的临界值 $\chi^2_{0.10} = 12.017$。因为 $7.2 < 12.017$，所以原假设不能被推翻。

估计正态分布的两个参数值将导致 χ^2 检验的自由度中的两类自由度值的减少。

对 Chemline 检测分数是否呈均值为 68.42 标准差为 10.41 的正态分布进行假设检验时，我们也可以使用 p - 值。通过 Minitab，我们可知 p - 值为 0.408。p - 值 $>\alpha=0.05$，所以我们不能推翻 H_0。正态分布拟合度检验过程总结如下：

正态分布拟合度检验：小结

1. 建立原假设和对立假设。

H_0：总体呈正态概率分布

H_a：总体不呈正态概率分布

2. 选取一个随机样本，然后：

a. 算样本均值和样本标准差；

b. 定义各值的区间使每个区间的预期频数至少等于 5，建议建立概率相等的区间；

c. 记录每个定义区间的调查频数。

3. 计算步骤 2(b)定义的每个区间的预期频数，将样本容量与正态随机变量的概率相乘。

4. 计算检验统计量的值 $\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}$

5. 拒绝法则：

使用检验统计量：　当 $\chi^2>\chi_\alpha^2$ 时，拒绝 H_0；

使用 p - 值：　　当 p - 值 $<\alpha$ 时，拒绝 H_0。

式中，α 为显著偏差水平，自由度为 $(k-3)$。

练　习

方　法

20. 每个时间段的发生次数与调查频数如下所示，取 $\alpha=0.05$，进行一下拟合度检验，看看以下数据是否符合泊松分布。

自测题

发生次数	调查频数
0	39
1	30
2	30
3	18
4	3

21. 下列数据来自正态分布，用 $\alpha=0.025$ 并做拟合度检验来检验这个假设。

自测题

17	23	22	24	19	23	18	22	20	13	11	21	18	20	21
21	18	15	24	23	23	43	29	27	26	30	28	33	23	29

应　用

22. 在一个指定的城市，每天的车祸发生次数服从泊松分布。去年中的80天的样本数据如下，这些数据是否能说明每天车祸发生次数服从泊松分布?取 $\alpha=0.05$。

车祸次数	观察频数(天)
0	34
1	25
2	11
3	7
4	3

23. 一家公司交换机在一分钟时间段内打入的电话数被认为服从泊松分布，利用 $\alpha=0.10$ 和下列数据检验打入电话数服从泊松分布的假设。

一分钟时间段内打入的电话数	观察频数
0	15
1	31
2	20
3	15
4	13
5	4
6	2
总计	100

24. 某产品的每周需求量服从正态分布。利用拟合度检验和下列数据检验假设，取 $\alpha=0.10$，样本均值为24.5，样本标准差为3。

18	20	22	27	22
25	22	27	25	24
26	23	20	24	26
27	25	19	21	25
26	25	31	29	25
25	28	26	28	24

25. 利用 $\alpha=0.01$，做拟合度检验，判断下列样本是否来自正态分布。

55	86	94	58	55	95	55	52	69	95	90	65	87	50	56
55	57	98	58	79	92	62	59	88	65					

在完成了拟合度计算后，建立数据的直方图，该直方图是否支持拟合度检验所得到的结论？(注：$\bar{x}=71$，$s=17$。)

小　结

本章中，我们介绍了拟合度检验和独立性检验，两者都基于 χ^2 分布。拟合度检验的目的是判断假设的概率分布是否可作为研究总体特征的模型。进行拟合度检验计算，包括当假设的概率分布为真时，将由样本观察到的频数与期望频数进行比较。χ^2 分布用来判断观察频数和期望频数之差是否大到可据此拒绝假设的概率分布。我们介绍了多项、泊松和正态分布的拟合度检验。

两个随机变量独立性检验是多项总体拟合度检验方法的推广。列联表确定观察频数和期望频数，然后计算 χ^2 值。若观察频数和期望频数之差较大，由 χ^2 值可导致拒绝独立性零假设。

术语释义

多样本总体：在几个类别中，每一元素被分配到且只能被分配到若干总体中的某一个。多项概率分布将由两个类别得到的二项分布推广到三个以上类别。

拟合度检验：为确定是否拒绝总体的假设概率分布所进行的统计检验。

列联表：在独立性检验中，用于汇总观察频数与期望频数的表格。

重要公式

拟合度的检验统计量

$$\chi^2=\sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i} \tag{12.1}$$

独立性假设下列联表中预期频数

$$e_{ij}=\frac{(\text{第}i\text{行的总数})(\text{第}j\text{列的总数})}{\text{样本容量}} \tag{12.2}$$

独立性检验统计量

$$\chi^2=\sum_i\sum_j\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \tag{12.3}$$

补充练习

26. 在配给销售定额时，市场经理做了每个销售地区潜在订货是相同的假设。由200次销售组成的样本如下，经理的假设能否被拒绝?取 $\alpha=0.05$。

销售地区			
Ⅰ	Ⅱ	Ⅲ	Ⅳ
60	45	59	36

27. 有7%的互助基金投资表认为公司股票“非常安全”，58%认为“比较安全”，24%的人认为“不太安全”，4%的认为“根本不安全”，7%的人“不确定”。《商业周刊》所做的一次哈里斯民意调查(Harris Poll)问到529名互助基金投资表关于他们如何评价公司债券的安全性问题。回答如下：

安全评价	频数
非常安全	48
比较安全	323
不太安全	79
根本不安全	16
不确定	63
总计	529

互助基金投资表对于公司债券的态度和他们对公司股票的态度是否不同?若 $\alpha=0.01$，用统计检验支持你的结论。

28. 一个社区公园将不久后开放。140人组成的样本说明他们最喜欢什么时候游园，结果如下：

工作日	周六	周日	节假日
20	20	40	60

在制定员工配给计划时，公园主管是否应基于每天的游园人数来制定计划?取 $\alpha=0.05$，用统计检验支持你的结论。

29. 某地区交通机构在关注一条公共汽车线路的乘车人数。在确定这条线路时，假设周一到周五每天下午乘车人数都相同。利用以下数据，取 $\alpha=0.05$，检验该交通机构的假设是否正确。

乘车日期	人数
周一	13
周二	16
周三	28
周四	17
周五	16

30. 《计算机世界》每年工作满意程度的调查结果表明，28% 的信息系统(IS)管理者对其工作非常满意，46% 比较满意，12% 无所谓满意与否，10% 有些不满，4% 非常不满。(*Computerworld*, May 26, 1997)假设 500 名计算机程序员组成的样本产生如下结果：

类别	回答人数
非常满意	105
比较满意	235
无所谓满意与否	55
有些不满	90
非常不满	15

取 $\alpha = 0.05$ 检验计算机程序员工作满意度与 IS 管理者的工作满意度是否不同。

31. 一个零件样本给出按生产班次划分的零件质量的数据如下：

班次	正品个数	次品个数
第一	368	32
第二	285	15
第三	176	24

利用 $\alpha = 0.05$ 检验零件质量与生产班次的独立性，你有何结论。

32. 1996 年 *Wall Street Journal Subscriber Study* 显示了关于订户职业的数据，东部版和西部版订户对应的样本结果如下所示：

	地区	
职业	东部版	西部版
全职	1105	574
临时工	31	15
自我雇用/咨询者	229	186
无职业	485	344

取 $\alpha = 0.05$，检验职业与地区是独立的假设，你有何结论?

33. 某贷款机构提供了关于四名贷款官员批准的贷款数据如下，利用 $\alpha = 0.05$ 检验批准贷款的决定与审核贷款申请的贷款官员是否独立。

	贷款批准决定	
贷款官员	批准	拒绝
Miller	24	16
McMahon	17	13
Games	35	15
Runk	11	9

34. 一项全国性调查提供了 20 至 29 岁之间的男女婚姻情况的统计数据。由 350 名男士和 400 名女士构成的样本结果如下，这些数据由《美国当今人口报告》出版(*The Statistical Abstract of the United States*, 1999)。

	婚姻状况		
性别	未婚	已婚	离异
男士	234	106	10
女士	216	168	16

a. 取 $\alpha=0.01$，检验婚姻状况和性别的独立性，你的结论如何?

b. 计算男、女士每个婚姻状况列中的比率。

35. 以下是一个关于消费品和银行业的100家公司的行业类型与市盈率表：

		市盈率					
		5—9	10—14	15—19	20—24	25—29	总计
行业	消费品	4	10	18	10	8	50
	银行业	14	14	12	6	4	50
	总 计	18	24	30	16	12	100

这些数据是否说明行业类型与市盈率有关?当 $\alpha=0.05$ 时，用统计检验支持你的结论。

36. 以下数据是关于弗吉尼亚州一个城市型县和一个乡村型县的紧急救护车呼叫次数的数据：

		星期							
		日	一	二	三	四	五	六	总计
	城市型	61	48	50	55	63	73	43	393
县	乡村型	7	9	16	13	9	14	10	78
	总计	68	57	66	68	72	87	53	471

用 $\alpha=0.05$ 进行独立性检验，你的结论如何?

37. 一大学课程期末考试的随机样本如下：

55	85	72	99	48	71	88	70	59	98	80	74	93	85	74
82	90	71	83	60	95	77	84	73	63	72	95	79	51	85
76	81	78	65	75	87	86	70	80	64					

用 $\alpha=0.05$ 检验分数的总体分布为正态分布的假设是否被拒绝。

38. 加利福尼亚州四个大都会地区的办公室使用率报告如下，下列数据能否说明办公室的空闲与地区是独立的?取 $\alpha=0.05$，你有何结论?

占用状态	洛杉矶	圣地亚哥	旧金山	圣何塞
占用	160	116	192	174
空闲	40	34	33	26

39. 一销售员每天打四个销售电话。由100天组成的一个样本，给出如下所示的销售量频数：

销售数量	观察频数(天)
0	30
1	32
2	25
3	10
4	3
总计	100

结果显示，销售额是销售电话的30%，假定独立销售电话，每天的销售量应符合二项分布，第5章给出的二项概率公式为：

$$f(x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

式中，假设总体有二项分布，$n=4$，$p=0.30$，且 $x=0$，1，2，3，4。

a. 利用二项概率函数，对 $x=0$，1，2，3 和 4，计算期望频数。若有必要可合并类别以使每类期望频数不少于5的条件成立。

b. 用拟合度检验决定二项分布的假设是否能被拒绝?取 $\alpha=0.05$。因为，二项分布从样本数据估出时没有限制，所以当 k 是表中数据时的自由度是 $k-1$。

案例研究　两党议程变更



CD 光盘数据
NYReform

在 Zogby 国际公司为“*Democrat and Chronicle*”进行的一次研究中，对700多位纽约人进行民意测验，判断纽约州政府是否正常运作。被调查者回答的问题涉及州立法机关经费缩减、党客限制和立法机关任期限制，及州公民是否能将事情直接交给州大选(*Democrat and Chronicle*, December 7, 1997)，涉及从人口到政治的几项改革提案获得广泛支持。

假定继而进行了一次对居住在纽约州西部地区的100个人的抽样调查。几位被调查的党派(民主党、独立党、共和党)人士，及他们对下列三个问题的回答被记录下来。

1. 因为每天的州预算都滞后，是否该缩减立法经费?
   是____否 ____
2. 是否对党客加以更多限制?
   是____否____
3. 是否该有要求立法机关服务固定年数的任期限制?
   是____否____

回答用1代表“是”，2代表“否”。所有数据都在CD光盘的NYReform文件中。

管理报告

1. 用描述统计汇总该研究中的数据。对调查中三个问题中的每一个问题，其答案(是与否)与党派类别之间的独立性，你有何初步结论?
2. 对问题 1，检验回答(是与否)与党派的独立性，取 $\alpha = 0.05$。
3. 对问题 2，检验回答(是与否)与党派的独立性，取 $\alpha = 0.05$。
4. 对问题 3，检验回答(是与否)与党派的独立性，取 $\alpha = 0.05$。
5. 对于改革所有政治方针，是否得到广泛支持?试解释之。

附录 12.1 用 Minitab 进行拟合度检验和独立性检验

我们讲述如何将 Minitab 用于拟合度检验和独立性检验。

拟合度检验

Minitab 方法可以用于 12.1 节中的多项分布的拟合度检验，泊松分布和 12.3 节中的正态分布。使用者要获得可观察频率，计算期望频率，并将可观察频率和期望频率输入到 Minitab 工作表中。C1 栏中录入观察频率，并包括观察频率，C2 栏中录入期望频数并包括期望频率。用 12.1 节中的 Scott 市场调查研究的例子计算，打开 Minitab 工作表，输入可观察频率 48.98 和 54 到 C1 栏中，输入期望频率 60, 100, 40 到 C2 栏中。拟合度检验的步骤如下：

步骤 1. 选择 **Calc** 下拉菜单；

步骤 2. 选择 **Calculator** 选项；

步骤 3. 出现 **Calculator** 对话框后，在 **Store result in variable** 栏输入 ChiSquare；
在 **Expression** 栏输入 Sum((C1 - C2)**2/C2)，点击 **OK**；

步骤 4. 选择 **Calc** 下拉菜单；

步骤 5. 选择 **Probability Distributions**；

步骤 6. 选择 **Chi-Square**；

步骤 7. 当 **Chi-Square Distribution** 对话框出现后，选择 **Cumulative Probability**，在 **Degrees of freedom** 栏中输入 CumProb，点击 **OK**；

步骤 8. 选择 **Calc** 下拉菜单；

步骤 9. 选择 **Calculator**，当 **Calculator** 对话框出现后，在 **Store results in variable box** 中输入 p - 值，在 **Expression** 框中输入 1-CumProb，点击 **OK**。

独立性检验

我们从 12.2 节中建一个新的 Minitab 工作表开始，将 Alber's 酿酒厂样本的观察频数数据分别是录入到第 1，2，3 列，于是，我们在 C1 中输入了对应于轻度啤酒偏好的观察频数(20 和 30)，在 C2 中输入了对应于中度啤酒偏好的观察频数(40 和 30)，在 C3 中输入了相对于高度啤酒偏好的观察频数(20 和 10)。独立性检验的 Minitab 步骤如下：

步骤 1. 选择 **Stat** 下拉菜单；

步骤 2. 选择 **Tables** 下拉菜单；

步骤 3. 选择 **Chi-Square Test**；

步骤 4. 当 **Chi-Square Test** 对话框出现后，在 **Columns containing the table** 栏输入 C1 - C3，点击 **OK**。

p - 值是 0.047，当 p - 值$<\alpha = 0.05$ 时，拒绝 H_0。

附录 12.2　用 Excel 进行拟合度检验和独立性检验

我们介绍如何用 Excel 表进行拟合度检验和独立性检验。

拟合度检验

Excel 方法可被用于 12.1 节中的多项分布的拟合度检验，如泊松分布和 12.3 节中的正态分布。使用者要获得观察频率，计算期望频率，在 Excel 工作表中输入观察和期望频率。用 12.1 节中的 Soctt 市场调查例子，打开 Excel 工作表，在单元格 A1 中输入观察频数，在单元格 B1 中输入期望频数。然后，在单元格 A2: A4 中输入观察频率 48, 98, 54，在单元格 B2: B4 中输入期望频率 60, 100, 40。

χ^2 检验统计和 p - 值可以计算如下：

步骤 1. 在单元格 C2 中输入下列公式；

$$= (A2 - B2)\text{^}2/B2$$

步骤 2. 把单元格 C2 复制到单元格 C3: C4；

步骤 3. 计算单元格 C5 中的 χ^2 的值；

$$= SUM\ (C2:\ C4)$$

步骤 4. 计算单元格 C6 中的 p 值；

$$= CHIDIST(C5, 2)$$

在步骤 4 中，输入的 2 是检验的自由度，在 Scott 市场调查举例中，$k=3$ 种类中的多项分布拟合度检验有 $k-1=2$ 的自由度。这个方法可扩展到 12.3 节中的泊松和正态分布的拟合度检验。步骤 4 中的 CHIDIST 功能包括了适合检验的自由度。

独立性检验

用 Excel 进行独立性检验要求使用者获得观察频数，并计算其期望的频数。Excel 工作表中有两个表，一个是观察频数，另一个是期望频数。图 12.4 表示了在 12.2 节中 Alber's 啤酒厂例子的 Excel 工作表。将观察值键入单元格 B4: D5，期望值在单元格 B11: D12。独立性检验的 p - 值可由单元格 C2 中的函数获得：=CHITEST(B4: D5, B11: D12)

Observed Frequencies				
	Beer Preference			
Gender	Light	Regular	Dark	Total
Male	20	40	20	80
Female	30	30	10	70
Total	50	70	30	150
Expected Frequencies				
	Beer Preference			
Gender	Light	Regular	Dark	Total
Male	26.67	37.33	16	80
Female	23.33	32.67	14	70
Total	50	70	30	150
			p-value=	0.047

图 12.4　Alber's 啤酒厂独立性检验的 Excel 扩展表

13 方差分析与试验设计

统计实例：BURKE 市场调研公司

13.1 **方差分析导论**

方差分析的假定

概念总览

13.2 **方差分析：k 个总体均值相等性的检验**

总体方差组间估计

总体方差组内估计

方差估计量的比较：F 检验

方差分析表

方差分析的计算结果

13.3 **多重比较方法**

费舍尔 LSD 方法

第一类错误的概率

13.4 **试验设计介绍**

数据收集

13.5 **纯随机设计**

总体方差组间估计

总体方差组内估计

方差估计量的比较：F 检验

方差分析表

配对比较

13.6 **随机区组设计**

空中交通管制员的工作压力测试

方差分析方法

计算与结论

13.7 **因子试验**

方差分析方法

计算与结论

统计实例

BURKE 市场调研公司*

辛辛那提,俄亥俄州

Burke 市场调研有限公司是在工业领域最富有经验的市场调研公司之一。Burke 每天都有比世界其他调研公司更多的建议或更多的项目。因为现代技术的支持，Burke 能够提供广泛的市场研究能力，可以对几乎所有市场问题提供答案。

在一项研究中，Burke 受一家公司雇用对儿童干谷类食品潜在需要的新品种进行评价。我们隐去真实名称,称该食品生产商为 Anon 公司。Anon 产品开发者认为提高谷类食品味道最主要的四个因素为：

1. 谷类食品的小薄片上小麦与玉米的比例。
2. 甜味剂的类型:白糖、蜂蜜或人工糖剂。
3. 果味香料的有无。
4. 烹调时间长短。

设计一个试验来决定这四种因素中哪种因素影响谷类食品的味道。例如：一种被用来检验的谷类食品是按特定的小麦与玉米的比例、糖作为甜味剂、有果味香料、烹调时间长短制作出来的。其他被检验的食品是在小麦与玉米比例不同而其他三个原料一样的情况下制作出来的。抽取几组小孩子品尝这些谷类食品，并要求他们说出他们对每种食品的感受。



Burke 的研究提供了有价值的统计信息，帮助了解消费者的需求。© Mary Kate Denny/PhotoEdit.

对品尝这些食品得来的数据研究进行统计的方法是方差分析。分析的结果如下：

- 小薄片上的成分和甜味剂类型对味道的影响最大。
- 果味香料其实破坏了谷类食品的味道。
- 烹调时间对食品味道没有影响。

这些信息用来判别出哪种因素是可能产生最佳口味的因素。

Burke 使用的试验设计和接下来的方差分析在推荐产品设计方案中很大帮助。本章我们将介绍如何实现这些统计程序。

*作者感谢 Burke 市场调研公司的 Dr. Ronald Tatham 所提供的统计应用材料。

在这一章中，我们将介绍一种叫**方差分析**（ANOVA）的统计程序。首先，我们将使用从某一观察研究中得来的数据来介绍如何用 ANOVA 来检验三个或三个以上的总体均值。接着，我们讨论使用 ANOVA 来分析数据的方法，这些数据来源于试验检验的三个类型：纯随机设计、随机区组设计和因素试验。在下一章将涉及到试验型数据和观测型数据的回归分析，我们将发现 ANOVA 对分析结果起很大的作用。

Ronald Alymer Fisher(1890—1962)创立了统计学分支——试验统计，他除了在统计领域有重大贡献外，还是一名基因学家。

13.1　方差分析导论

National Computer Products 有限公司(NCP)计划在亚特兰大、达拉斯和西雅图的工厂生产打印机和传真机。为了了解有多少员工知道全面质量管理知识，从每个工厂随机抽取 6 个员工作为样本，对他们进行质量意识考试。表 13.1 中给出 18 名员工的考试分数，同时给出每组的样本平均值、样本方差和样本标准差。管理人员利用每个工厂的数据，检验所有工厂的平均考试分数是否一样。

表 13.1　18 名员工的考试分数


CD 光盘数据
NCP

观察值	工厂 1 亚特兰大	工厂 2 达拉斯	工厂 3 西雅图
1	85	71	59
2	75	75	64
3	82	73	62
4	76	74	69
5	71	69	75
6	85	82	67
样本平均值	79	74	66
样本方差	34	20	32
样本标准差	5.83	4.47	5.66

我们定义在亚特兰大工厂的所有员工为总体 1，在达拉斯工厂的所有员工为总体 2，在西雅图工厂的所有员工为总体 3。令

μ_1 = 总体 1 的平均分数；
μ_2 = 总体 2 的平均分数；
μ_3 = 总体 3 的平均分数。

虽然我们不知道 μ_1、μ_2 和 μ_3 的真实值，但我们可用样本资料来检验下面的假设：

H_0：$\mu_1 = \mu_2 = \mu_3$
H_a：总体均值不全相等

若 H_0 被拒绝，则我们不能得出所有的平均值不同的结论。拒绝 H_0 表示至少有两个总体平均值不同。

简短的说，方差分析是一种判定三个样本的平均值是否大到能拒绝 H_0 的统计程序。

在介绍这一章时，我们曾说方差分析既可用来分析观察型数据，又可用来分析实验型数据。为了给出这两种类型的研究中方差分析所使用的一套相同的术语，我们将引进反应变量、因子和处理概念。

在 NCP 例子中，两个变量分别是工厂位置和质量意识考试分数。因为目的是为了判断亚特兰大、达拉斯和西雅图的工厂的分数是否相同，所以考分是因变量或**反应变量**。工厂的位置是自变量或**因子**。将用于研究的因子值称为因子水平或**处理**。这样，NCP 例子中三个处理分别为亚特兰大、达拉斯和西雅图。在 NCP 中这三个处理规定了研究总体，对于每个处理或总体，反应变量是考试分数。

方差分析的假定

在进行方差分析时需要三个假定：

1. **对于每个总体，反应变量服从正态分布。**这表示，在 NCP 例子中，每个工厂(反应变量)必须服从正态分布。
2. **反应变量的方差，记为 σ^2，对所有的总体都相同。**这表示，在 NCP 例子中，三个工厂考试分数的方差相同。
3. **观察值是独立的。**这表示，在 NCP 例子中，每个员工的考试分数与其他员工的考试分数是独立的。

如果样本容量相等，方差分析并不明显依赖于正态总体分布假定。

概念总览

如果三个总体的平均值相等，则得出三个样本平均值接近的结论。事实上，如果三个样本平均值越接近，我们得出总体均值相等的结论就越充分。相应地，如果样本均值越不同，我们得出总体均值不相等的结论就越充分。换句话说，如果样本平均值的差异“小”，则支持 H_0；如果样本平均值的差异“大”，则支持 H_a。

如果原假设 H_0：$\mu_1 = \mu_2 = \mu_3$ 为真，则我们可利用样本均值的差异来估计 σ^2。首先，满足方差分析的条件前提下，来自同一正态分布的每个样本的均值和方差 σ^2 相等。回顾第 7 章中讲述的，一个样本容量为 $\bar{x}$、平均值为 n 的简单随机样本来自正态分布总体，它的平均值为 μ，方差为 σ^2/n。图 13.1 描述了这种抽样分布。

因而，当原假设为真时，我们从表 13.1 中抽取的三个样本平均值为 $\bar{x}_1 = 79$，$\bar{x}_2 = 74$，$\bar{x}_3 = 66$，每一个都是从抽样分布中随机抽取的，图 13.1 上给出其值。这时，可用三个 $\bar{x}$ 的均值与方差来估计抽样分布的平均值和方差。在 NCP 例子中，$\bar{x}$ 的抽样分布均值的最好估计是三个样本均值

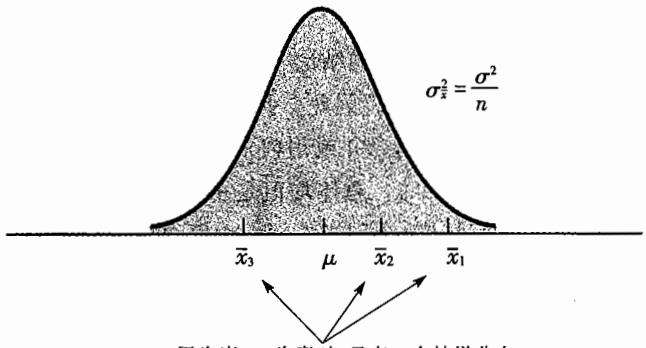


图 13.1　当 H_0 为真时，$\bar{x}$ 的抽样分布

的算术平均数。因此，在 NCP 例子中，$\bar{x}$ 的抽样分布的均值估计是(79 + 74 + 66) / 3 = 73。我们称该估计值为**总样本均值**。$\bar{x}$, $\sigma_{\bar{x}}^2$ 的抽样分布的方差的估计值是由样本均值的方差给出的。

$$s_{\bar{x}}^2 = \frac{(79-73)^2 + (74-73)^2 + (66-73)^2}{3-1} = \frac{86}{2} = 43$$

因为 $\sigma_{\bar{x}}^2 = \sigma^2 / n$，解：$\sigma^2$

因此

$$\sigma^2 = n\,\sigma_{\bar{x}}^2$$

$$\sigma^2 \text{ 的估计值} = n(\sigma_{\bar{x}}^2 \text{ 的估计值}) = n\,s_{\bar{x}}^2 = 6(43) = 258$$

结果为 $n\,s_{\bar{x}}^2 = 258$，称为 σ^2 的**组间估计**。

σ^2 的组间估计是建立在原假设为真的基础之上的。在这种情况下，每个样本来自同一样本，因此只有一个 $\bar{x}$ 的抽样分布。当 H_0 为假时，即假定总体均值全不相同，这是因为这三个样本来自不同均值的正态总体，他们将产生三个不同的抽样分布。图 13.2 表明在这种情况下，当 H_0 为真时，样本均值不接近。因此，$s_{\bar{x}}^2$ 将变大，从而 σ^2 的组间估计也将变大。一般地，若样本均值不等，总

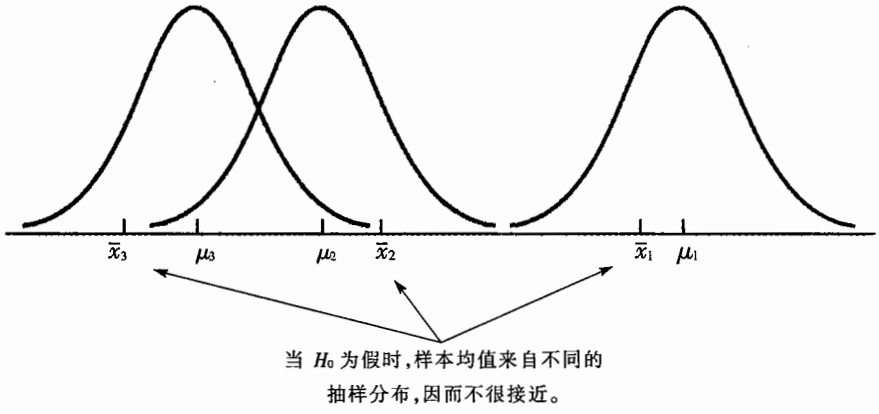


图 13.2　当 H_0 为假时，$\bar{x}$ 的抽样分布

体方差的组间估计将偏大。

每个样本的内部方差同样对方差分析得出的结论有影响。从每个总体中选取一个简单随机样本，则每个样本方差给出一个 σ^2 的无偏估计，因此，我们可将 σ^2 的个别估计组成或和并成一个总的估计。这样得出的 σ^2 估计被称为 σ^2 的合并或组内估计。因为每个样本方差所给出的 σ^2 估计值是建立在每个样本内部方差的基础上的，总体均值是否相等对 σ^2 的组内估计没有影响。当样本容量相等时，σ^2 的组内估计可通过计算个别样本的平均值得出。对于 NCP 例子，我们得出

$$\sigma^2 \text{ 的组内估计} = \frac{34 + 20 + 32}{3} = \frac{86}{3} = 28.67$$

在 NCP 例子中，σ^2 的组内估计(258)远大于 σ^2 的组内估计(28.67)。事实上，这两个估计 σ^2 值的比率为 258/28.67 = 9.00。但是，当原假设为真时，组间估计才是的最好估计；当原假设为假时，组间估计得到的 σ^2 估计偏大。在这两种情况下，组内估计都是 σ^2 的最好估计。因此，当原假设为真时，这两个估计值非常接近，它们的比率将接近 1。如果原假设为假时，组间估计将远大于组内估计，则它们的比率将很大。在下一节我们将介绍比率多大时，才需拒绝 H_0。

总之，ANOVA 隐含的逻辑是基于对两个具有共同方差 σ^2 总体的独立估计的推广。σ^2 的一个估计值是基于样本均值之间的差异，σ^2 的另一个估计值是基于每个样本内部数据的差异。通过比较这两个 σ^2 估计值，我们可以判断总体均值是否相等。

评 注

1. 在 10.1 节和 10.2 节我们介绍了检验两个总体的均值相等的统计方法。ANOVA 同样也能用于检验两个总体均值相等的假设。实际应用中，方差分析是用于三个或更多的总体的均值。
2. 在 10.2 节，我们介绍了当至少有一个样本的样本容量小于 30 时如何检验两个均值相等的假设。在这次讨论中，我们介绍了将两个独立随机样本的数据组合成的一个估计值的过程。这个过程称为合并，而得到的样本方差为 σ^2 的合并估计量。在方差分析中，σ^2 的组内估计量是这个概念在三个样本中的扩展，这就是为什么把组内的估计也称为 σ^2 的合并估计的原因。

13.2 方差分析：k 个总体均值相等性的检验

可用样本方差对 k 个总体均值进行检验。假设检验的一般形式如下：

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_a：总体均值不全相等

式中，μ_j = 第 j 个总体的均值。

假设从 k 个总体或处理中选择一个样本容量为 n 的简单随机样本。对于所得样本数据，令

x_{ij} = 第 j 个处理的观察值 i;

n_j = 第 j 个处理的观察值数目;

$\bar{x}_j$ = 第 j 个处理的样本平均值;

s_j^2 = 第 j 个处理的样本方差;

s_j = 第 j 个处理的样本标准差。

第 j 个处理的样本平均值和样本方差的公式如下:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

总样本平均值,记为$\bar{\bar{x}}$,等于所有观察值之和除以被观察值总个数。即

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

式中

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

如果每个样本的样本容量为 n, $n_T = kn$; 公式(13.3)可简化为

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{nk} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}/n}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \tag{13.5}$$

换句话说，只要样本容量相等，总体均值恰好等于 k 个样本均值的算术平均数。

因为在 NCP 例子中，每个样本都是由 $n = 6$ 个观察值组成。总样本均值可由公式(13.5)得出。对于表 13.1 的数据，我们得出下面的结论:

$$\bar{\bar{x}} = \frac{79 + 74 + 66}{3} = 73$$

如果原假设为真($\mu_1 = \mu_2 = \mu_3 = \mu$)，总样本均值 73 是总体均值 μ 的最好估计。

总体方差组间估计

在上节中，我们引进了 σ^2 的组间估计概念，并介绍当样本相等时如何计算 σ^2 的组间估计值。σ^2 的估计值称为**处理均方**，记为 MSTR。MSTR 的计算公式如下：

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{13.6}$$

公式(13.6)中的分子称为**处理平方和**，记为 SSTR，分母 k－1 表示 SSTR 相联系的自由度。因此，可按下列公式计算处理均方：

处理均方

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} \tag{13.7}$$

式中

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

如果 H_0 为真，则 MSTR 是 σ^2 的无偏估计。然而 k 个总体的均值不全相等，则 MSTR 不是 σ^2 的无偏估计。事实上，MSTR 估高了 σ^2。

利用表(13.1)中的 NCP 例子的数据，我们得出下面结论：

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 6(79-73)^2 + 6(74-73)^2 + 6(66-73)^2 = 516$$

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{516}{2} = 258$$

总体方差组内估计

在前面的章节中，我们介绍了 σ^2 的组内估计概念，并介绍当样本大小相等时的计算方法。σ^2 的估计值称为**误差均值**，记为 MSE。MSE 的计算公式如下：

$$\text{MSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 / n_T - k \tag{13.9}$$

公式(13.9)中的分子称为**误差平方和**，记为 SSE。MSE 的分母是与 SSE 相联系的自由度。因此，MSE 的公式如下表示：

误差均值

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \tag{13.10}$$

式中

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \tag{13.11}$$

注意到 MSE 是基于每个处理内部的差异，故它不受原假设是否为真的影响。因此，MSE 是 σ^2 的无偏估计。

利用表(13.1)中的 NCP 数据，我们得到下列结果：

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = (6-1)34 + (6-1)20 + (6-1)32 = 430$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{430}{18-3} = \frac{430}{15} = 28.67$$

方差估计量的比较：F 检验

我们先假定原假设为真。此时，MSTR 和 MSE 是 σ^2 的两个独立的无偏估计。第 11 章中指出，对于正态分布，σ^2 的两个独立估计量之比的抽样分布服从 F 分布。因此，如原假设为真，则 ANOVA 假设有效，MSTR/MSE 的抽样分布服从 F 分布，其分子自由度为 $k-1$，分母自由度为 $n_T - k$。

如果 k 个总体均值不相等，因为 MSTR 高估了 σ^2，则 MSTR/MSE 值将变大。因此，如果 MSTR/MSE 值太大，大得不像是从分子自由度为 $k-1$，分母自由度为 $n_T - k$ 的 F 分布中随机抽取的，我们将拒绝 H_0。导致拒绝 H_0 的 MSTR/MSE 值取决于显著性水平 α。α 一经确定，则拒绝值被确定。图 13.3 描述了 MSTR/MSE 的抽样分布以及显著性水平为 α 时的拒绝域，其中 F_α 表示临界值。

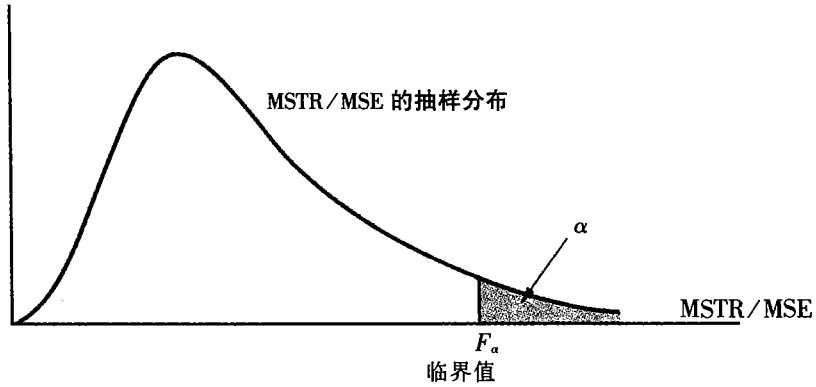


图 13.3　MSTR/MSE 的抽样分布；拒绝原假设均值相等的临界值为 F_α

假定在 NCP 中的决策的管理者愿意接受发生第一类错误的概率的 $\alpha=0.05$。从附录 B 中表 4 中我们可以查的 F 值的临界值，其中分子自由度为 $k-1=3-1=2$，分母自由度为 $n_T-k=18-3=15$。因此，我们得出 $F_{0.05}=3.68$。注意，这个数值表示如果我们从分子自由度为 2，分母自由度为 15 的 F 分布中随机选取一个值，观察到的这个数值大于 3.68 的可能性只有 5%。另外，方差分析理论告诉，我们如果假定原假设为真，MSTR/MSE 的比例是 F 分布中的一个值。因此，对于 NCP 例子中的拒绝域为

如果 $F=\text{MSTR}/\text{MES}>3.68$，则拒绝 H_0。

而 MSTR = 258，MSE = 28.67。因为 $F=\text{MSTR}/\text{MSE}=258/28.67=9.00$ 远大于临界值 $F_{0.05}=3.68$，我们有充分的证据拒绝三个总体均值相等的假设。换句话说，方差分布支持三个 NCP 工厂的总体平均考分不相等的结论。

p - 值也可用于假设检验。一般的拒绝规则为：如 p - 值$<\alpha$，则拒绝 H_0。但是，从 F 概率分布表中很难精确得出 P - 值，因此需要计算机统计软件包如 Minitab 或 Excel 的帮助。附录 13.1 和 13.2 介绍了使用方法，在下一节讨论通过计算机得出的方差分析结果。对于 NCP 例子，与检验统计量 $F=9.00$ 相关的 p - 值为 0.003。因为 p - 值 $=0.003<\alpha=0.05$，故原假设 H_0：三个总体均值相等将被拒绝。对 k 个总体均值相等性检验的程序总结如下：

k 个总体均值相等性检验

$$H_0:\ \mu_1=\mu_2=\cdots=\mu_k$$

H_a：总体均值不全相等

检验统计量

$$F=\frac{\text{MSTR}}{\text{MSE}} \tag{13.12}$$

拒绝规则

利用检验统计量：如果 $F>F_\alpha$，则拒绝 H_0

利用 p - 值：如果 p - 值$<\alpha$，则拒绝 H_0

式中，F_α 是基于分子自由度为 $k-1$，分母自由度为 n_T-k 的 F 分布。

方差分析表

上面的计算结果可以用方差分析表或称 **ANOVA 表**表示出来。表 13.2 是 NCP 例子的方差分析表。在方差来源一列中，平方和“合并”称为总平方和(SST)。注意到在 NCP 例子中，有 SST = SSTR + SSE，并且总的平方和对应的自由度是 σ^2 的组间估计所对应的自由度与 σ^2 的组内估计所对应的自由度之和。

应该指出，SST 被自由度 n_T-1 所除的结果只不过是将所有 18 个观察值视为一个数据集而得到的总样本方差。以整个数据集为一个样本，总的平方和 SST 的计算公式为

$$\text{SST}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{\bar{x}})^2 \tag{13.13}$$

NCP 例子中方差分布表观察得到的上述结果也可用于其他问题，即

$$SST = SSTR + SSE \tag{13.14}$$

方差分析可被视为总平方和被分解为不同成分的一种统计程序。

换句话说，SST 可划分为两个平方和之和：处理平方和与误差平方和。注意，SST 对应的自由度 $n_T - 1$ 也可划分为 SSTR 对应的自由度 $k-1$ 与 SSE 对应的自由度 $n_T - k$。方差分析就是将总的平方和及自由度**分解**(partitioning)为相应的来源：处理和误差。给适当的自由度划分总平方和提供了方差估计量以及用于检验总体均值相等性的值。

方差分析的计算结果

因为统计计算软件包的广泛使用，大样本容量和/或多个总体的方差分析计算很容易实现。图 13.4 表示用 Minitab 软件包得出的 NCP 例子的输出结果。计算机输出结果的第一部分是熟悉的 ANOVA 表形式。比较图 13.4 和表 13.2，可以看到虽然标题不同，其内容却是一样的。表头“Source”是表示方差来源这一列，“Factor”表示对应的处理的那一行。

```
Analysis of Variance
Source     DF        SS        MS        F        p
Factor      2     516.0     258.0     9.00    0.003
Error      15     430.0      28.7
Total      17     946.0

                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
 Level      N      Mean     StDev  ---+---------+---------+---------+---
Atlanta     6    79.000     5.831                      (------*------)
Dallas      6    74.000     4.472             (------*-----)
Seattle     6    66.000     5.657   (-----*------)
                                  ---+---------+---------+---------+---
Pooled StDev =     5.354           63.0      70.0      77.0      84.0
```

图 13.4　**Minitab 关于 NCP 方差分析的输出结果**

表 13.2　NCP 例子的方差分析表

方差来源	平方和	自由度	均方差	F
处理	516	2	258.00	9.00
误差	430	15	28.67	
总体	946	17		

平方和和自由度两列互换，p-值是为了检验。因而，在显著性水平 $\alpha = 0.05$ 下，因为 p-值 $= 0.003 < \alpha = 0.05$，拒绝 H_0。

注意 ANOVA 表下方，计算机输出结果中包含了各自样本容量、样本均值和样本标准差。另外，Minitab 给出了每个总体均值的 95% 置信区间估计。在进行置信区间估计时，Minitab 使用 MSE 为 σ^2 的估计值。因而，MSE 的平方根给出了总体标准差 σ 的最好估计。σ 的估计值在计算机的输出结果中对应“Pooled StDev”，它等于 5.354。为了解释这些区间估计是怎样得来的，我们计算一下亚特兰大工厂总体均值的 95% 置信区间估计。

第 8 章的区间估计方法中，我们知道总体均值区间估计的一般形式为

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{13.15}$$

式中，s 为总体标准差 σ 的估计值。在方差分析时，σ 的最好估计是由 MSE 的平方根或 Pooled StDev 给出，所以我们将公式(13.15)的 s 代为 5.354。t 值的自由度为 σ^2 的组内估计对应的自由度 15。因此，对于 $t_{0.025}=2.131$，我们得到

$$79 \pm 2.131\frac{5.354}{\sqrt{6}} = 79 \pm 4.66$$

这样，亚特兰大工厂的 95% 置信区间为 79 - 4.66 = 74.34 至 79 + 4.66 = 83.66。因为在 NCP 例子中样本容量相等，达拉斯和西雅图工厂各自的置信区间也可由各自样本均值加减 4.66 得出。因此，Minitab 得出的图中可以发现置信区间的宽度相等。

评 注

1. 总体均值也可由 k 个样本均值的加权平均得出。

$$\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_T}$$

当已知样本均值时，用这个公式计算总均值比公式(13.3)式简单。

2. 当每个样本由 n 个观察值组成时，公式(13.6)可写成

$$\text{MSTR} = \frac{n\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1} = n\left[\frac{\sum_{j=1}^{k}(\bar{x}_j - \bar{\bar{x}})^2}{k-1}\right] = ns_{\bar{x}}^2$$

注意，这个结论与 13.1 节中我们引入 σ^2 的组内估计概念时所介绍的结果一样。公式(13.6)是该结果在样本大小不相等时的简单推广。

3. 如果每个样本有 n 个观察值，$n_T = kn$；因此，$n_T - k = k(n-1)$，公式(13.9)可写成

$$\text{MSE} = \frac{\sum_{j=1}^{k}(n-1)s_j^2}{k(n-1)} = \frac{(n-1)\sum_{j=1}^{k}s_j^2}{k(n-1)} = \frac{\sum_{j=1}^{k}s_j^2}{k}$$

换句话说，如果样本大小相等，MSE 只是 k 个样本方差的算术平均值。注意这个结论与 13.1 节中我们引入 σ^2 的组内估计概念时所介绍的结果一样。

4. k 个总体每个均值的置信区间可按下式建立：$\bar{x} \pm t_{\alpha/2}\frac{\sqrt{\text{MSE}}}{\sqrt{n}}$

t 值的自由度为 σ^2 的组内估计相对应的自由度。

练　习

方　法

1. 从三个总体分别抽取五个观察值，得到如下数据：



观察值	样本 1	样本 2	样本 3
1	32	44	33
2	30	43	36
3	30	44	35
4	26	46	36
5	32	48	40
样本均值	30	45	36
样本方差	6.00	4.00	6.50

a. 计算 σ^2 的组间估计。

b. 计算 σ^2 的组内估计。

c. 在显著性水平 $\alpha=0.05$ 下，我们能否拒绝原假设：三个总体均值相等？

d. 建立 ANOVA 表。

2. 从三个样本分别抽取 4 个观察值，得到以下数据：

观察值	样本 1	样本 2	样本 3
1	165	174	169
2	149	164	154
3	156	180	161
4	142	158	148
样本均值	153	169	158
样本方差	96.67	97.33	82.00

a. 计算 σ^2 的组间估计。

b. 计算 σ^2 的组内估计。

c. 在显著性水平 $\alpha=0.05$ 下，我们能否拒绝原假设：三个总体均值相等？

d. 建立 ANOVA 表。

3. 从 3 个总体分别选取一个样本：

样本 1	样本 2	样本 3
93	77	88
98	87	75

（续表）

	样本 1	样本 2	样本 3
	107	84	73
	102	95	84
		85	75
		82	
$\bar{x}_j$	100	85	79
s_j^2	35.33	35.60	43.50

a. 计算 σ^2 的组间估计。

b. 计算 σ^2 的组内估计。

c. 在显著性水平 $=0.05$ 下，我们能否拒绝原假设：三个总体均值相等?试解释。

d. 建立 ANOVA 表。

4. 从 4 个样本中各随机选取 16 个观察值。ANOVA 表的一部分内容如下：

方差来源	平方和	自由度	均方差	F
处理			400	
误差				
总计	1 500			

a. 将 ANOVA 表中的残缺项补齐。

b. 在显著性水平 $\alpha=0.05$ 下，我们能否拒绝原假设：四个总体均值相等?

5. 从 3 个样本中各随机选取 25 个观察值。对于这些数据，SSTR = 120，SSE = 216。

a. 建立 ANOVA 表。

b. 在显著性水平 $\alpha=0.05$ 下，临界值 F 值的大小是多少?

c. 在显著性水平 $\alpha=0.05$ 下，我们能否拒绝原假设：三个总体均值相等?

应　用

6. （自测题）为了检验三个生产商生产的机器混合一批原材料所需的平均时间是否一致，Jocobs 化学公司得到关于混合原料所需时间的如下数据。利用这些数据检验三个生产商生产的机器混合一批原料总体均值是否相等，取 $\alpha=0.05$。

生产商		
1	2	3
20	28	20
26	26	19
24	31	23
22	27	22

7. 某组织的各级管理者需要适当的信息去完成各自的任务。一份研究调查信息来源对信息传播的影响。在这份研究中信息来源分为上级、同事和下属。在每个情形中，得到的传播程度数据的数值越高表示信息传播很广。取 $\alpha=0.05$，利用下面数据，检验信息来源对传播是否有明显的

影响。得出你的结论，对信息的使用和来源你有何建议？

上级	同事	下级
8	6	6
5	6	5
4	7	7
6	5	4
6	3	3
7	4	5
5	7	7
5	6	5

8. 一份研究调查市场专业人员的公司伦理观念。取 $\alpha=0.05$，利用下面数据(分值越高表明伦理价值观念程度越高)，检验三个专业人员群体的公司伦理理念是否存在明显的差异。

房地产代理商	建筑商	广告人员
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

9. 
CD 光盘数据
Stress

Journal of Small Business Management 报告的一个研究中称在该研究中为了度量一些模棱两可、含糊不清的方面，设计了 15 个问题，这 15 项问题按从强烈同意到强烈反对，分成 1—5 级打分。对每一个人，15 项问题打分总分在 15—75 之间。得分越高，表明工作压力越大(*Journal of Small Business Management*, October 1997)。假定在类似的研究中，用 20 个这样 1—5 回答选项的问题对随机抽取得的 15 名房地产代理商、15 名建筑商和 15 名股票经纪人度量各自的工作压力。结果如下：

房地产代理商	建筑商	股票经纪人
81	43	65
48	63	48
68	60	57
69	52	91
54	54	70
62	77	67
76	68	83
56	57	75
61	61	53
65	80	71
64	50	54
69	37	72
83	73	65
85	84	58
75	58	58

取 $\alpha=0.05$，检验在这三种职业中工作压力是否有显著差异。

10. 《商业周刊》的“全球1000强”栏目中按每家公司的市值排名(*Business Week*, July 7, 1997)。下面表中显示了29家金融行业公司的市盈率比。行业代码1代表银行业，行业代码2表示金融服务业，行业代码3表示保险公司。取 $\alpha=0.05$，检验这三种金融公司的市盈率比是否相同。


CD光盘数据
Mkt Value

公司	行业代码	市盈率	公司	行业代码	市盈率
Citicorp	1	15	MBNA	2	24
NationsBank	1	14	Cincinnati Financial	2	19
Wells Fargo	1	25	Franklin Resources	2	22
First Union	1	13	Fannie Mae	2	17
KeyCorp	1	14	American International	3	21
Chase Manhattan	1	12	Group		
Fifth Third Bancorp	1	23	Allstate	3	14
Bank of New York	1	17	Marsh & McLennan	3	20
First Chicago NBD	1	13	American General	3	16
Mellon Bank	1	16	Cigna	3	12
Fleet Financial	1	15	Lincoln National	3	13
Group			AFLAC	3	21
First Bank System	1	16	Equitable	3	11
American Express	2	19	Chubb	3	20
Travelers	2	15	General Re	3	15
Merrill Lynch	2	12			

13.3 多重比较方法

用方差分析检验 k 个总体均值是否相等时，拒绝原假设只能得出总体均值不全相等的结论。此时我们需要进一步检验并判定哪些均值之间存在差异。本节的目的是介绍两种**多重比较方法**(multiple comparison procedures)来对总体均值配对比较。

费舍尔 LSD 方法

假定通过方差分析拒绝总体均值相等的假设，此时，费舍尔的最小显著差异(LSD)方法可用于判定哪些均值之间有差异。在介绍用费舍尔LSD方法进行总体均值配对比较时，仍以13.1节的NCP为例子。通过方差分析，我们的三个工厂总体平均考试分数不等。接下的问题是：我们认为存在差异，但哪些工厂之间存在差异?也就是说，是总体1与总体2之间均值不同，还是总体1与总体3之间均值不同，还是总体2与总体3之间均值不同?

在第 10 章中我们介绍了对两个总体均值相等性检验的统计程序。费舍尔 LSD 方法是对总体方差估计稍微做以修改得出的，并且基于两个总体情形的 t 检验统计量。下表给出了费舍尔 LSD 方法的概要：

费舍尔 LSD 方法

$$H_0:\ \mu_i = \mu_j$$

$$H_a:\ \mu_i \neq \mu_j$$

检验统计量

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13.16}$$

拒绝规则

如果 $t < -t_{\alpha/2}$ 或 $t > t_{\alpha/2}$，则拒绝 H_0

式中，$t_{\alpha/2}$ 值是基于自由度为 $n_T - k$ 的 t 分布。

现在将此方法运用于判断总体 1(亚特兰大)和总体 2(达拉斯)的均值之间是否有明显的差异。表 13.1 表示亚特兰大工厂的样本均值为 79，达拉斯工厂的样本均值为 74。表 13.2 表示 MSE 值为 28.67，此值为 σ^2 的估计值，自由度为 15。在显著性水平为 0.05 下，查 t 分布表得，自由度为 $n_T - k = 18 - 3 = 15$ 时，$t_{0.025} = 2.131$。因此，如果 $t < -2.131$ 或 $t > 2.131$，则拒绝 H_0：$\mu_1 = \mu_2$。利用 NCP 例子中的数据，我们得出以下的 t 值：

$$t = \frac{79 - 74}{\sqrt{28.67\left(\frac{1}{6} + \frac{1}{6}\right)}} = 1.62$$

因为 $t = 1.62$，故我们没有充足的证据去拒绝原假设，因此，我们不能得出亚特兰大工厂与达拉斯工厂的总体平均分数不等的结论。

许多实际应用人员发现，容易确定样本均值之间的偏差大到何种程度时必须拒绝 H_0。此时检验统计量是 $\bar{x}_i - \bar{x}_j$，而且检验过程按如下步骤进行：

基于检验统计量的费舍尔 LSD 方法 $(\bar{x}_i - \bar{x}_j)$

$$H_0:\ \mu_i = \mu_j$$

$$H_a:\ \mu_i \neq \mu_j$$

检验统计量

$$\bar{x}_i - \bar{x}_j$$

在显著性水平为 α 时的拒绝规则

如果 $|\bar{x}_i - \bar{x}_j| > \text{LSD}$，则拒绝 H_0

式中

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{13.17}$$

对于 NCP 例子，LSD 值为

$$\text{LSD} = 2.131\sqrt{28.67\left(\frac{1}{6}+\frac{1}{6}\right)} = 6.59$$

注意当样本容量大小相等时，只需计算一个 LSD 值。此时，我们只需简单将任何两个样本均值之差与 LSD 值比较。例如，总体 1(亚特兰大)与总体 3（西雅图）样本均值之差为 79 - 66 = 13。差值大于 6.59，这表示我们能拒绝亚特兰大工厂与西雅图工厂的总体平均分数相等的原假设。类似的，总体 2 与总体 3 样本均值之差为 74 - 66 = 8 > 6.59，我们仍拒绝原假设即达拉斯工厂与西雅图工厂的总体平均分数相等。因此，我们可以得出结论，对于平均分数，亚特兰大工厂和达拉斯工厂都与西雅图工厂有差异。

费舍尔 LSD 也可用于两个总体均值之间的置信区间估计。一般方法如下：

利用费舍尔的 LSD 方法对两个总体均值之差进行置信区间估计

$$\bar{x}_i - \bar{x}_j \pm \text{LSD} \tag{13.18}$$

式中

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{13.19}$$

$t_{\alpha/2}$ 是基于自由度为 $n_T - k$ 的 t 分布。

如果公式(13.18)中的置信区间包含零值，则我们不能拒绝两个总体均值相等的假设。但是，如果置信区间没有包含零值，我们得出总体均值存在差异的结论。对于 NCP 例子，LSD = 6.59（相应的 $t_{0.025} = 2.131$）。因此，总体 1 与总体 2 之间差值的 95% 置信区间的估计为 79 - 74 ± 6.59 = 5 ± 6.59 = (- 1.59，11.59)。因为此区间包含 0，所以我们不能拒绝两个总体均值相等的假设。

第一类错误的概率

我们对费舍尔 LSD 方法进行讨论时，首先假定方差分析已经提供统计证据去拒绝总体均值相等的原假设。因此，我们介绍使用费舍尔 LSD 方法去判定哪些之间存在差异。从技术上讲，此方法为**保护性**或**限制性** LSD 检验，此方法必须通过方差分析找到一个显著的 F 值，才能使用。为了解释多重比较检验时，这种**区别**是十分重要的，因此，我们需要解释比较性第一类错误概率和试验性第一类错误概率的区别。

对于 NCP 例子，我们用费舍尔 LSD 方法做三个配对比较。

检验 1	检验 2	检验 3
H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_3$	H_0: $\mu_2 = \mu_3$
H_a: $\mu_1 \neq \mu_2$	H_a: $\mu_1 \neq \mu_3$	H_a: $\mu_2 \neq \mu_3$

对于每种情况，取显著性水平 $\alpha = 0.05$。所以，对于每个检验，如果原假设为真，我们犯第一类错误的概率为 $\alpha = 0.05$，而不犯第一类错误的概率为 1 - 0.05 = 0.95。在多重比较方法中我们称第一类错误的概率 $\alpha = 0.05$ 为**比较性第一类错误概率**(comparisonwise Type I error rate)。比较性第一类错误概率是指单一配对比较中的显著性水平。

先考虑一个稍有差异的问题。在进行三个配对比较检验时，此检验中至少有一次犯第一类错误的概率为多少？为了回答这个问题，注意三次检验都不犯第一类错误概率为 $0.95 \times 0.95 \times 0.95 = 0.857\,4$。* 所以，至少有一次犯第一次错误的概率为 $1 - 0.857\,4 = 0.142\,6$。因此，当我们用费舍尔 LSD 方法检验三个配对比例时，第一类错误的概率不是 0.05，而是 0.142 6；我们将此错误概率称为**总的**或是**试验性第一类错误概率**(experimentwise Type I error rate)。为了避免混淆，我们将试验性第一类错误记为 α_{EW}。

若样本个数越多，发生试验性第一类错误概率越大。例如，对于由 5 个总体的问题，将有 10 个可能的配对比较。当比较性错误概率 $\alpha = 0.05$ 时，如果我们用费舍尔 LSD 方法去检验所有可能配对比较，试验性第一类错误概率为 $1 - (1 - 0.05)^{10} = 0.40$。此时，使用者应去寻找更好的方法去控制试验性错误概率。

Bonferroni 修正方法是控制总的试验性错误概率的一种方法，这种方法采用对每个检验使用更小的比较错误的方法。例如，如果我们要检验 C 个配对比较，并且希望总的试验性检验中最大发生第一类错误的概率为 α_{EW}，我们简单将单个配对错误概率定为 α_{EW}/C。在 NCP 例子中，假定最大试验性概率 $\alpha_{EW} = 0.05$，如果我们用费舍尔 LSD 方法检验全部三个配对比较，则比较性错误概率为 $\alpha = 0.05/3 = 0.017$。对于有 5 个总体和 10 个可能配对比较的问题，Bonferroni 修正方法给定比较性错误概率为 $0.05/10 = 0.005$。回顾第 9 章中假设检验的讨论中，对于固定的样本容量，第一类错误发生的概率减少将导致第二类错误发生的概率增加。而第二类错误对应的是指当两个样本均值实际上相等时，而接受两个总体均值不等的假设。因此，由于犯第二类错误发生的概率增加，许多使用者情愿接受相对低的第一类错误发生的概率。

另外 Tukey 方法和 Duncan 多重比较方法也适用于这种情况。但是，对于哪种方法更好，在统计学界有很大争执。事实上，没有哪种方法能适用于所有的问题。

练　习

方　法

11. 自测题

在练习 1 中，从三个总体中分别抽取 5 个观察值。数据如下：$\bar{x}_1 = 30$，$\bar{x}_2 = 45$，$\bar{x}_3 = 36$，MSE = 5.5。在显著性水平为 $\alpha = 0.05$ 下，原假设样本均值相等被拒绝。在接下的计算中，取 $\alpha = 0.05$。

a. 用费舍尔 LSD 方法检验总体 1 和总体 2、总体 1 和总体 3、总体 2 和总体 3 平均值之间是否有差异？

b. 用费舍尔 LSD 方法对总体 1 和总体 2 之间的差值的 95% 置信区间估计。

---

* 假定三个检验是独立的，因此三个事件的联合概率为各自概率的简单相乘。事实上，因为在每个检验中都适用 MSE，三个检验不独立，所以发生错误的概率比上述大。

12. 从三个总体中分别抽取 4 个观察值。得到如下数据，在接下来的计算中，取 $\alpha=0.05$。

	样本 1	样本 2	样本 3
	63	82	69
	47	72	54
	54	88	61
	40	66	48
$\bar{x}_j$	51	77	58
s_j^2	96.67	97.34	81.99

a. 利用方差分析检验三个总体均值之间是否有明显的差异?

b. 用费舍尔 LSD 方法检验哪些均值之间有差异?

应　用

13. 参照练习 6。取显著性水平 $\alpha=0.05$，用费舍尔 LSD 方法检验生产商 1 和生产商 3 均值是否相等。检验后，得出你的结论。

自测题

14. 参照练习 13。利用费舍尔 LSD 方法对总体 1 和总体 2 的均值之差的 95% 置信区间进行估计。

自测题

15. 参照练习 8。假定显著性水平 $\alpha=0.05$，我们可以得出市场管理人员、市场研究人员与广告人员在伦理观念上有差异。利用本节介绍的方法判定哪些之间有差异，取 $\alpha=0.05$。

16. 为了检验 4 台机器的故障时间有任何明显差异，得到如下数据：

机器 1	机器 2	机器 3	机器 4
6.4	8.7	11.1	9.9
7.8	7.4	10.3	12.8
5.3	9.4	9.7	12.1
7.4	10.1	10.3	10.8
8.4	9.2	9.2	11.3
7.3	9.8	8.8	11.5

a. 取显著性水平 $\alpha=0.05$，4 部机器的总体平均值之间存在差异吗?

b. 用费舍尔 LSD 方法检验机器 2 和机器 4 之间均值相等性，取显著性水平 $\alpha=0.05$。

17. 参照练习 16。用 Bonferroni 修正方法检验所有成对均值之间有无明显差异，假定总体的试验性误差的最大概率为 0.05。

18. 参照练习 10。在显著性水平为 0.05 下，我们得出银行、金融服务公司和保险公司之间的平均价格/收入比率比有差异，利用本节介绍的方法检验哪些之间有差异，取 $\alpha=0.05$。

13.4　试验设计介绍

统计研究被分为试验型与观察型两类。在**试验型研究**中，对感兴趣的变量加以定义。因此，在研究一个或多个因素被控制，数据可从因素如何影响变量中获得。在**观察或非试验型研究**中，没必要对因素控制。调查（见第 21 章）是最常见的观察型研究的类型。

在观察例子中因果关系很难判断，但这些关系在试验设计中容易建立。

引入方差分析的 NCP 例子是观察型统计研究的一个例子。为了了解 NCP 员工对全面质量管理的理解程度，从 NCP 三个工厂中分别随机抽取 6 个员工，并对他们进行质量意识考试。这些员工考试分数通过方差分析检验三个工厂总体平均考分相等的假设。

我们以试验型研究的一个例子即 Chemitech 公司所面临的问题为例。Chemitech 开发了一种新型的城市供水过滤系统，这个过滤系统的组成部分是从各个供应商中购买的，Chemitech 只是在南加州哥伦比亚的工厂负责组装。工程部负责确定新过滤系统的最佳装配方法。在考虑了大量可能的方法后，负责人将其方法减少为三种：方法 A、方法 B 和方法 C。这三种方法在产品组装的步骤顺序上有不同。Chemitech 的管理人员希望能判断哪种装配方法每星期能生产的过滤系统最多。

在 Chemitech 试验中，装配方法是独立变量或**因子**(factor)。因为对应这些因子有三种装配方法，我们称这次试验中有三个处理，每个**处理**(treatment)对应三种装配方法中的一个。Chemitech 问题是涉及到一个定性因素（装配方法）的**单因子试验**(single-factor experiment)的例子。其他试验可能由多个因素组成，其中一些因素是定性的，而另外一些是定量的。

在 Chemitech 试验中根据三种装配方法或处理分类成三个总体。其中一个是使用装配方法 A 的全体 Chemitech 员工，一个是使用装配方法 B 的全体 Chemitech 员工，第三个是使用装配方法 C 的全体 Chemitech 员工。注意对于每个总体变量或反应变量是每个星期装配的过滤系统数目。试验的主要统计目的是判定三个总体每星期装配的平均数目是否相同。

假设从 Chemitech 生产车间中的所有装配工人中抽取三名员工。在试验设计术语中，三个随机选取的工人是**试验单元**(experimental units)。将 Chemitech 问题中使用的试验型设计称为**纯随机设计**(completely randomized design)。这种类型的设计要求将三种装配方式或处理中的其中一个随机分配给试验单元或工人。例如，方法 A 随机分配给第二名工人，方法 B 随机分配给第一名工人，方法 C 随机分配给第三名工人。这个例子介绍的**随机化**概念是一个重要的原理。

随机化是将处理随机地指派给试验单位的步骤。在 R. A. Fisher 的工作之前，处理是被系统或主观地指派的。

注意到对于每个处理，一次试验只有一个装配数目的度量。换句话说，每个处理的样本容量为 1。为了获得每种装配方法更多的数据，我们必须重复或复制基本试验步骤。例如，假设不是

从总体中随机抽取 3 个工人而是抽取 15 名工人，将这 15 名工人随机分配，每三个处理分配给 5 名工人。因为每种装配方法有 5 名工人，故可以得到 5 个重复结果。复制过程是试验设计另一个重要原理。图 13.5 显示了 Chemitech 装配方法试验的纯随机设计。

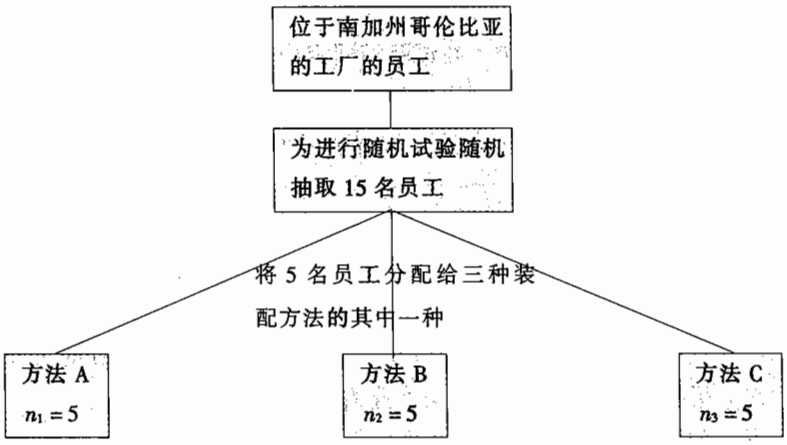


图 13.5　评估 Chemitech 装配方法试验的纯随机设计

数据收集

一旦对试验设计满意后，我们将收集数据并加以分析。在 Chemitech 例子中，首先将指导每个员工使用其分配的装配方法，然后利用该方法组装新的过滤系统。假设组装和培训完成，表 13.3 列出每个员工一个星期组装系统的数目。三种装配方法所生产的样本平均数见下表：

装配方法	平均生产数目
A	62
B	66
C	52

对照这些数据，发现方法 B 似乎比其他两种方法有更高的生产效率。

表 13.3　15 名工人生产的产品单位数目


CD 光盘数据
ChemTech

	方法		
观察值	A	B	C
1	58	58	48
2	64	69	57
3	55	71	59
4	66	64	47
5	67	68	49
样本均值	62	66	52
样本方差	27.5	26.5	31.0
样本标准差	5.24	5.15	5.57

现实问题是观察到的样本平均值差异是否足够大，以至于得出每种装配方法的总体均值不同。为了将此问题以统计术语表达，我们引入下列符号：

μ_1 = 方法 A 每周生产的新系统的数目；
μ_2 = 方法 B 每周生产的新系统的数目；
μ_3 = 方法 C 每周生产的新系统的数目。

虽然我们不知道 μ_1、μ_2 和 μ_3 的真实值，我们利用样本均值来检验下面的假设：

H_0：$\mu_1 = \mu_2 = \mu_3$
H_a：所有总体均值不全相等

对于纯随机试验性设计进行数据分析时所面临的问题，与我们开始引入方差分析检验两个以上总体均值是否相等时面临的问题一样。在下面几节中我们将介绍如何使用方差分析来处理类似 Chemitech 装配方法的问题。

评　注

1. 试验设计中的随机化是观察性研究中概率抽样的一个类比。
2. 在一些医药试验中，潜在的偏差可以通过双重未知的研究消除。在这样的研究中，无论是使用该处理的医师还是受治疗者，都不知道是用哪种处理方法。许多其他类型的试验也可借鉴这个类型的研究。

13.5　纯随机设计

在分析来自纯随机设计的数据时，我们进行的假设与 13.2 节中所介绍的假设的一般形式一样。

H_0：$\mu_1 = \mu_2 = \cdots = \mu_k$
H_a：总体均值不全相等

因此，从纯随机设计中收集数据并对其进行均值检验时，我们将使用 13.1 节和 13.2 节中介绍的方差分析。回顾知道进行方差分析时，需要计算总体方差 σ^2 的两个独立估计。

总体方差组间估计

σ^2 的总体估计是指**处理平方和**，记为 MSTR，MSTR 的计算公式如下：

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{13.20}$$

公式(13.20)中分子为**组间平方和**或**处理平方和**，记为 SSTR。分母 $k-1$ 表示与 SSTR 相关的自由度。

利用表 13.3 中的 Chemitech 数据，我们得出以下结论(注意：$\bar{\bar{x}} = 60$)：

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 = 5\ (62-60)^2 + 5(66-60)^2 + 5(52-60)^2 = 520$$

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{520}{3-1} = 260$$

总体方差组内估计

σ^2 的组内估计是指**误差均方**，记为 MSE，计算 MSE 的公式如下：

$$\text{MSE} = \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n_T - k} \tag{13.21}$$

公式(13.21)的分子为**组内平方和**或**误差平方和**，记为 SSE。MSE 的分母是指与组内方差估计有关的自由度。

利用表 13.3 中的 Chemitech 数据，我们得到如下结果：

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(27.5) + 4(26.5) + 4(31) = 340$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{340}{15-3} = 28.33$$

方差估计量的比较：*F* 检验

如果原假设为真，ANOVA 假设有效，MSTR/ MSE 的抽样分布是分子自由度为 $k-1$，分母自由度为 $n_T - k$ 的 F 分布。如果 k 个总体的均值不等，MSTR/ MSE 值将变大，因为 MSTR 高估了 σ^2。因此，如果 MSTR/ MSE 值太大，以至于不是分子自由度为 $k-1$，分母自由度为 $n_T - k$ 的 F 分布的值，我们将拒绝 H_0。

对于 Chemitech 问题，F 值 = MSTR/ MSE = 260/28.33 = 9.18。F 的临界值是基于分子自由度为 2，分母自由度为 12 的 F 值。在显著性水平为 0.05 下，附录 B 中的表 4 中查得 $F_{0.05} = 3.89$。如果 F 的观察值大于临界值，则我们拒绝原假设并得出总体均值不全等的结论。

方差分析表

现在我们可以写出总的平方和 SST 的分解形式：

$$\text{SST} = \text{SSTR} + \text{SSE} \tag{13.22}$$

同样这一结果也适用于与这些平方和对应的自由度，也就是说，总自由度等于 SSTR 和 SSE 对应的自由度之和。纯随机设计的 ANOVA 表的一般形式如表 13.4 所示，表 13.5 是 Chemitech 问题相对应的 ANOVA 表。

表 13.4　纯随机设计的 ANOVA 表一般形式

方差来源	平方和	自由度	均方差	F
处理	SSTR	$k-1$	$\text{MSTR} = \dfrac{\text{SSTR}}{k-1}$	$\dfrac{\text{MSTR}}{\text{MSE}}$
误差	SSE	$n_T - k$	$\text{MSE} = \dfrac{\text{SSE}}{n_T - k}$	
总计	SST	$n_T - 1$		

表 13.5　Chemitech 问题相对应的 ANOVA 表

方差来源	平方和	自由度	均方差	F
处理	520	2	260.00	9.18
误差	340	12	28.33	
总计	860	14		

配对比较

我们可以用费舍尔 LSD 方法来检验 Chemitech 问题中的所有可能配对比较。取显著性水平为 0.05，t 分布表明当自由度为 $n_T - k = 15 - 3 = 12$ 时，$t_{0.025} = 2.179$。将 MSE = 28.33 代入公式（13.17），我们得到费舍尔的最小显著差异

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = 2.179\sqrt{28.33\left(\frac{1}{5}+\frac{1}{5}\right)} = 7.34$$

如果任意两样本的差异大小超过 7.34，我们将拒绝相应的总体均值相等的假设。对于表 13.3 的 Chemitech 数据，我们得到以下结果：

样本差别	是否显著
方法 A − 方法 B = 62 − 66 = −4	否
方法 A − 方法 C = 62 − 52 = 10	是
方法 B − 方法 C = 66 − 52 = 14	是

因此，总体均值差异可归于方法 A 和方法 C 的均值之间的差异以及方法 B 和方法 C 的均值之

间的差异，故方法 A 和方法 B 比方法 C 要好。但是，对于方法 A 和方法 B，需要进一步检验，目前的研究没有给出足够证据得出这两种方法有差异。

练　习

方　法

19. 下面数据来自纯随机设计：



	处理		
观察值	A	B	C
1	162	142	126
2	142	156	122
3	165	124	138
4	145	142	140
5	148	136	150
6	174	152	128
$\bar{x}_j$	156	142	134
s_j^2	164.4	131.2	110.4

a. 计算处理平方和。

b. 计算处理均方。

c. 计算误差平方和。

d. 计算误差均方。

e. 在显著性水平 $\alpha=0.05$ 下，检验三个处理的均值是否相等。

20. 参照练习 19。

a. 建立 ANOVA 表。

b. 在显著性水平 $\alpha=0.05$ 下，利用费舍尔最小显著差异方法检验所由可能的配对比较。在进行检验后，得出你的结论。

21. 在一次纯随机试验设计中，5 个因子水平每个指派 7 个试验单元。试完成下列 ANOVA 表：

方差来源	平方和	自由度	均方差	*F*
处理	300			
误差				
总计	460			

22. 参照练习 21。

a. 本问题中蕴含着什么假设?

b. 在显著性水平 $\alpha=0.05$ 下，我们能否拒绝(a)中的原假设?试解释。

23. 在一次试验中为了检验三个不同处理的输出水平，得到以下数据：SST = 400，SSTR = 150，$n_T = 19$。建立 ANOVA 表，检验三个处理的平均输出水平之间是否有显著差异？取 $\alpha = 0.05$。

24. 在纯随机试验设计中，第 1 个处理指派 12 个试验单元，第 2 个处理指派 15 个试验单元，第 3 个处理指派 20 个试验单元。计算下列的方差分析。在显著性水平 $\alpha = 0.05$ 下，处理问题是否有显著差异？

方差来源	平方和	自由度	均方差	*F*
处理	120 0			
误差				
总计	180 0			

25. 对于下面的试验设计进行方差分析计算。在显著性水平 $\alpha = 0.05$ 下，处理均值之间是否有显著差异？



CD 光盘数据
Exer25

	处　理		
	A	**B**	**C**
	136	107	92
	120	114	82
	113	125	85
	107	104	101
	131	107	89
	114	109	117
	129	97	110
	102	114	120
		104	98
		89	106
$\bar{x}_j$	119	107	100
s_j^2	146.86	96.44	173.78

应　用

26. 一个工业小组对装配某种产品提出三种不同的方法。为了调查每种方法组装的产品单位的数量，随机选取 30 名工人，并随机分配给三种方法，每种方法分配 10 个。对组装的产品个体如实记录，对记录数据集进行方差分析。得到如下结果：SST = 10 800；SSTR = 4 560。

a. 针对此问题建立 ANOVA 表。

b. 取 $\alpha = 0.05$，三种装配方法组装的产品数目均值之间是否有显著差异？

27. 在一次对 4 种类型电缆的断裂强度试验设计中，得到如下数据：SST = 85.05，SSTR = 61.64，

$n_T=24$。建立 ANOVA 表，检验 4 种电缆的平均断裂强度是否有显著差异，取 $\alpha=0.05$。

28. 为了研究温度对某化学反应产生的影响，在三种不同的温度条件下分别产出 5 批物质。结果如下，建立方差分析表，取 $\alpha=0.05$，检验温度高低是否对某化学反应产出的影响有显著差异。

温度		
50℃	**60℃**	**70℃**
34	30	23
24	31	28
36	34	28
39	23	30
32	27	31

29. 审计员必须根据自己的直接经验、间接经验或两者组合来对某一个审计报告方方面面作出判断。在一次研究中，审计员被要求对审计报告中出错的概率作出判断，然后将审计员的判断与真实结果比较。假设从类似的研究中得到下面数据，分数越低判断越准确。


CD 光盘数据
AudJudg

直接	间接	组合
17.0	16.6	25.2
18.5	22.2	24.0
15.8	20.5	21.5
18.2	18.3	26.8
20.2	24.2	27.5
16.0	19.8	25.8
13.3	21.2	24.2

取 $\alpha=0.05$，检验判断的根据是否影响判断的质量?得出你的结论?

30. 广告中说四种不同涂料有相同的风干时间。为了检验生产商的广告，从每个涂料中检验 5 种样本。将每种涂料干到能涂第二层涂料的时间记录下来，得到如下数据：


CD 光盘数据
Paint

涂料 1	涂料 2	涂料 3	涂料 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

在显著性水平 $\alpha=0.05$ 下，检验每种类型的涂料平均风干时间是否相等?

31. 一家著名汽车杂志对美国生产的三种中等型号顶级汽车进行行驶检测，并根据一系列的标准进行比较。为了检验汽油里程性能，从每个牌子中各选取 5 部车子进行 500 英里行驶检测。取 $\alpha=0.05$，对其进行方差分析，判定三种类型汽车每加仑汽油平均行驶英里是否有显著差异。

汽车		
A	B	C
19	19	24
21	20	26
20	22	23
19	21	25
21	23	27

32. 参考练习 29。利用费舍尔最小显著性差异方法检验所有可能配对比较。在进行检验后，得出你的结论，取 $\alpha = 0.05$。

33. 参考练习 31。利用费舍尔最小显著性差异方法检验所有可能配对比较。在进行检验后，得出你的结论，取 $\alpha = 0.05$。

13.6　随机区组设计

迄今为止我们讨论了纯随机设计。回顾在检验处理均值之间的差异值时，我们用比率来计算 F 值：

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13.23}$$

当试验单元同质时，纯随机设计有用。如果试验单元异质，可用“区组划分”形成同质性小组。

当外在的因素(试验中没有考虑到的因素)引起的差异导致比率中的 MSE 项增加时，问题就出现了。此时，公式(13.23)中的 F 值减少。实际上差异是存在的，但 F 值给出了处理均值之间没有差异。

在本节中我们将介绍一种称为**随机区组设计**(randomized block design)。其目的是通过剔除 MSE 项中方差来控制方差的外部来源。这个设计给出了实际误差方差的最好估计，并且建立一个更好的假设来检验处理均值之间的差异。为了解释这个问题，我们以空中交通管制员工作压力研究为例。

空中交通管制员的工作压力测试

有一份衡量空中交通管制员疲劳和工作压力的研究报告，该报告建议修改和重新设计管制员的工作压力。在经过对这个工作站的几个设计考虑后，三种方案因其最有可能减少管制员压力而被选出。关键问题是：三种备选方案对于管制员压力的影响有多大差异?为了回答这一问题，我们需要设计一个试验来给出每个方案下空中管制员压力的度量。

在纯随机设计中，管制员的随机样本被分配给每个工作站。但是，一般认为管制员在处理压力局面时存在差异。一名管制员认为是高压力，可能另一名认为是中等，甚至又有另一名认为是小的压力。因此，考虑方差的组内来源(MSE)时，我们必须意识到方差包括随机误差，也可能包括个人管制员差异导致的误差。事实上，管制员希望个人差异是 MSE 项的最大来源。

企业的经验性研究经常是高度混合的，结果是常常用到随机区组设计。

使用随机区组设计方法将个人差异分离出来。这样的设计识别出来源于管制员个人差异的方差，并将其从 MSE 中剔除。随机区组设计需要一个简单的管制员样本。在试验设计的术语中，将工作站称为**兴趣因子**，管制员为**区组**，与工作站因子相关的三个处理或总体分别对应着三个工作站方案。为了简单起见，我们称工作站方案为系统 A、系统 B 和系统 C。

在实验设计中区组划分类似于样本分层。

随机区组设计中的**随机**是指将处理(系统)指派给管制员的顺序。如果每个管制员以相同的顺序检验这三种方案，任何观察到的系统差异都可能归因于检验的顺序，而不是系统中的真正差异。

为了提供有用的数据，将三种工作站方案运用于俄亥俄州奥柏林的克利夫兰管制中心。随机选出 6 名管制员，并且指派他们操作每个系统。对于参加研究的每个管制员，随后的见面交谈及对每个管制员做的医疗检验，提供了每个管制员在每个系统下的压力的变量。表 13.6 给出了数据。

表 13.6　空中交通管制员压力检验的随机区组设计

		处理		
		系统 A	系统 B	系统 C
区组	管制员 1	15	15	18
	管制员 2	14	14	14
	管制员 3	10	11	15
	管制员 4	13	12	17
	管制员 5	16	13	16
	管制员 6	13	13	13

表 13.7 是收集的压力数据汇总。在这个表中我们有汇总列（处理）和汇总行（区组），以及一些样本均值。这些数据对于 ANOVA 方法中的平方和计算将有用。因为较低的压力值表示较好，故样本数据表示采用系统 B，因为其平均压力为 13。但是，一般的问题仍未解决：样本结果能否判断三个系统的总体平均压力水平不同?类似于纯随机设计中方差分析的计算可以用来回答这个问题。

表 13.7　空中交通管制员工压力检验的压力数据汇总

		处理			行或区组总计	区组平均值
		系统 A	系统 B	系统 C		
区组	管制员 1	15	15	18	48	$\bar{x}_{1\cdot}=48/3=16.0$
	管制员 2	14	14	14	42	$\bar{x}_{2\cdot}=42/3=14.0$
	管制员 3	10	11	15	36	$\bar{x}_{3\cdot}=36/3=12.0$
	管制员 4	13	12	17	42	$\bar{x}_{4\cdot}=42/3=14.0$
	管制员 5	16	13	16	45	$\bar{x}_{5\cdot}=45/3=15.0$
	管制员 6	13	13	13	39	$\bar{x}_{6\cdot}=39/3=13.0$
行或处理总计		81	78	93	252	$\bar{\bar{x}}=\frac{252}{18}=14.0$
处理均值		$\bar{x}_{\cdot 1}=\frac{81}{6}=13.5$	$\bar{x}_{\cdot 2}=\frac{78}{6}=13.0$	$\bar{x}_{\cdot 3}=\frac{93}{6}=15.5$		

方差分析方法

随机区组设计 ANOVA 方法将总的平方和分解为三个部分：处理平方和、区组平方和和误差平方和。分解形式如下：

$$\text{SST}=\text{SSTR}+\text{SSBL}+\text{SSE} \tag{13.24}$$

表 13.8 所示的随机区组设计的 ANOVA 表汇总了这种平方和分解。表中记号分别表示：

k = 处理的个数；

b = 区组的个数；

n_T = 总的样本大小（$n_T=kb$）。

表 13.8　k 个处理、b 个区组的随机区组设计的 ANOVA 表

方差来源	平方和	自由度	均方差	F
处理	SSTP	$k-1$	$\text{MSTR}=\frac{\text{SSTR}}{k-1}$	MSTR/MSE
区组	SSBL	$b-1$	$\text{MSBL}=\frac{\text{SSBL}}{b-1}$	
误差	SSE	$(k-1)(b-1)$	$\text{MSE}=\frac{\text{SSE}}{(k-1)(b-1)}$	
总计	SST	n_T-1		

注意到 ANOVA 表说明了总的自由度 n_T-1 分解为处理自由度 $k-1$、区组自由度 $b-1$ 和误差自由度 $(k-1)(b-1)$ 之和，均方差列表示平方和被自由度除，F = MSTR/ MSE 为用于检验处

理均值之间差异显著性的 F 比。随机区组主要作用是通过区组从 MSE 项中剔除个人管制员差异，并得到一个关于三个工作站方案压力差别的更好的检验。

计算与结论

为了计算随机区组设计中处理均值差异的 F 统计量，我们需要计算 MSTR 和 MSE。对于这两个均方，我们首先计算 SSTR 和 SSE。为此，我们必须计算 SSBL 和 SST。为了简化这个表达式，我们分四步计算。除了先前定义的 k、b 和 n_T 外，另外引入下列记号：

x_{ij} = 对应于区组 i 中处理 j 的观察值；
$\bar{x}_{\cdot j}$ = 处理 j 的样本均值；
$\bar{x}_{i\cdot}$ = 区组 i 的样本均值；
$\bar{\bar{x}}$ = 总体样本均值。

步骤 1. 计算总的平方和(SST)；

$$\text{SST} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13.25}$$

步骤 2. 计算处理平方和(SSTR)；

$$\text{SSTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.26}$$

步骤 3. 计算区组平方和(SSBL)；

$$\text{SSBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.27}$$

步骤 4. 计算误差平方和(SSE)。

$$\text{SSE} = \text{SST} - \text{SSTR} - \text{SSBL} \tag{13.28}$$

对于表 13.7 中的空中交通管制员数据，上述计算得到如下平方和：

步骤 1. $\text{SST} = (15-14)^2 + (15-14)^2 + (18-14)^2 + \cdots + (13-14)^2 = 70$
步骤 2. $\text{SSTR} = 6[(13.5-14)^2 + (13.0-14)^2 + (15.5-14)^2] = 21$
步骤 3. $\text{SSBL} = 3[(16-14)^2 + (14-14)^2 + (12-14)^2 + (14-14)^2 + (15-14)^2 + (13-14)^2] = 30$
步骤 4. $\text{SSE} = 70 - 21 - 30 = 19$

表 13.9 为平方和除以相应的自由度得到相应的均方值。检验处理均值之间的差异的 F 比率 = MSTR/ MSE = 10.5/1.9 = 5.53。查附录 B 中表 4 的 F 值，我们得到 $\alpha = 0.05$，分子自由度为 2，分母自由度为 10 的 F 临界值为 4.10。而 $F = 5.53$，故我们拒绝原假设 H_0：$\mu_1 = \mu_2 = \mu_3$，得出在三个工作站方案下总体均值压力水平有差异的结论。

表 13.9　空中交通管制员的压力检测的 ANOVA 表

方差来源	平方和	自由度	均方差	F
处理	21	2	10.5	10.5/1.9 = 5.53
区组	30	5	6.0	
误差	19	10	1.9	
总计	70	17		

对于随机区组设计，需作一些说明。本节所讨论的试验设计是完全区组设计；**完全**一词是指每个区组都属于所有个处理，也就是说，所有管制员（区组）在所有三个系统（处理）下受到检测。**不完全**区组设计是一些但不是全部的处理被运用于每个区组的试验设计。不完全区组设计的讨论超出这本书的范围。

因为在空中交通管制员压力测试中，每个管制员要求使用所有三个系数，这种方法将保证是完全区组设计。在某些情况下，“区组划分”是根据区组中“相似”的试验单元划分的。例如，假定在空中交通管制员检测之前，管制员总体被分成从极高压力小组到极低压力小组。从每个压力小组中抽取 3 个参加这项研究，这个区组划分仍能实现。每个区组在同一压力小组中有三个管制员。

最后，表 13.8 中的 ANOVA 表提供了检验处理影响而不是检验区组的 F 值。其原因是设计的试验是为了检验单个因子——工作站而设计的。基于个人压力差异的区组划分是为了从 MSE 项中剔除来自个人差异的方差而进行的，但是，该研究设计不是为了特别检验个人压力的差别。

一些分析家计算 F = MSB/ MSE 并用该统计量检验区组的显著性。然后利用得到的结论去验证，同类型的区组分组是否可用在将来的试验中。但是，如果个人压力差异是研究中的一个因子，应该设计一个不同的试验。区组的显著性不应该将个人压力差异作为第二个因素结论的依据。

评　注

1. 第 10 章介绍的匹配样本 t 检验是两个区组的随机区组设计的一个例子。
2. 随机区组设计的自由度要小于纯随机设计的自由度，因为 b 个区组导致减少 $b-1$ 个自由度。如果 n 小，因为误差自由度的减少，区组的潜在影响将被掩盖；如果 n 很大，这种影响将被最小化。

练　习

方　法

34. 考虑下面随机区组设计的试验结果，并为建立方差分析表进行计算。



		处理		
		A	B	C
区组	1	10	9	8
	2	12	6	5
	3	18	15	14
	4	20	18	18
	5	8	7	8

取 $\alpha=0.05$，检验显著差异。

35. 一个随机区组设计涉及 5 个处理和三个区组，得到以下数据：SST = 430，SSTR = 310，SSBL = 85。建立 ANOVA 表并检验显著性差异，取 $\alpha=0.05$。

36. 对 4 个处理和 8 个区组进行试验，完成下面的方差分析表：

方差来源	平方和	自由度	均方差	*F*
处理	900			
区组	400			
误差				
总计	1800			

取 $\alpha=0.05$，检验显著差异。

应　用

37. 一个汽车交易商进行测试并判断一个小型引擎启动所需分钟数是否与使用计算机引擎分析器或电子分析器有关。因为微型、中型和大型汽车的启动时间是不同的，使用了三种类型的汽车作为试验中的区组，得到以下数据：

		分析器	
		计算机引擎分析器	电子分析器
汽车	微型	50	42
	中型	55	44
	大型	63	46

当 $\alpha=0.05$，检验任何显著差异?

38. 就总的审计时间采用 5 种不同的审计方法进行比较。为了控制属于进行审计的个人的可能的方差，随机选取了四名会计为试验的一个区组。由 ANOVA 方法得到以下数据：SST = 100，SSTR = 45，SSBL = 36。取 $\alpha=0.05$，检验 5 种审计方法平均总审计时间是否显著差异?

39. 选择字处理和数据管理系统的一个重要因素是需要多长时间学会这个系统。为了评价三个文件管理系统，一个公司设计一个涉及到 5 个字处理软件操作员的测试。因为操作员的不同被认为是一个显著的因素，对每位操作员都进行 3 种文件管理系统的培训，得到如下数据：

		系统		
		A	B	C
操作员	1	16	16	24
	2	19	17	22
	3	14	13	19
	4	13	12	18
	5	18	17	22

40. 《美国医药协会杂志》(*Journal of the American Medical Association*)研究的一个报告考察了重体力铲雪对心脏的要求。10 名健康男子进行了用脚踏车以及一种圆形曲柄测力计测试的运动。用一种轻型塑料雪铲和电动除雪器来铲除两堆重而湿的雪。把每个人铲雪时的心、血压、吸氧量以及感到的用力大小同用脚踏车和测力计测试时得到的数据进行比较。假定下表给出 10 个人的心率(每分钟跳动次数)的数据：


CD 光盘数据
Snow

测试人	脚踏车	曲柄测力计	雪铲	除雪器
1	177	205	180	98
2	151	177	164	120
3	184	166	167	111
4	161	152	173	122
5	192	142	179	151
6	193	172	205	158
7	164	191	156	117
8	207	170	160	123
9	177	181	175	127
10	174	154	191	109

用显著性水平 0.05 检验显著性差异。

13.7　因子试验

迄今为止的试验设计能得出一个因子的统计方法结论。但是，在一些试验中，需要得出关于多个变量或因子的结论。当要求同时得到关于两个或更多的因子的结论时，**因子试验**(factorial ex-

periments)以及相应的 ANOVA 计算都是有价值的设计。使用术语“**因子**”是因为试验中将所有可能的因子组合起来。例如，对于因子 A 的 a 水平和因子 B 的 b 水平，那么将收集 ab 处理组合的数据。在这节中我们将介绍两个因子的试验分析，这种方法可以扩展到两个以上的试验。

作为两个因子的试验分析，我们将研究管理类研究生入学考试(GMAT)。GMAT 是商学院为了评价申请人在该领域读研究生的能力的一种标准化考试，GMAT 的考分从 200 到 800，分数越高表示能力越强。

为了提高学生在 GMAT 考试中的分数，得克萨斯大学的一些学业招生考虑提供下面三种 GMAT 辅导课程：

1. 3 小时复习，内容覆盖了 GMAT 常考题型。
2. 1 小时练习，包括有关的考试资料，还有一次模拟考试并评分。
3. 10 个星期的强化班，包括发现考生个人的弱点并建立个人提高计划。

因为，研究中其中一个因素是 GMAT 辅导课程，它有三个处理：3 个小时复习，1 个小时练习和 10 周课程。在选择采纳哪种辅导课程之前，需要进一步研究判断建议的课程对 GMAT 有何影响?

GMAT 考生来自三种学院：商学院、工学院和艺术与科学学院。所以试验中感兴趣的第二因子是考生的毕业学院是否影响 GMAT 分数。第二因子即毕业学院，选择有三个处理是商学、工学和艺术与科学。对应因子 A，辅导课程有三种处理，对应因子 B，毕业三种处理，因此共有 $3 \times 3 = 9$ 处理组合。表 13.10 汇总了这些处理组合或试验条件。

表 13.10　两个因子 GMAT 试验的 9 个处理组合

		因子 B：毕业学院		
		商学	工学	艺术与科学
因子 A：辅导课程	3 小时复习	1	2	3
	1 天课程	4	5	6
	10 周强化班	7	8	9

假定从表 13.10 中相应的 9 个处理组合中每个随机选出 2 个学生组成样本。两个商学院学生参加 3 个小时复习班，两个参加 1 天练习，两个参加 10 周课程。另外，对于每种三个辅导课程，各有两个工学院学生和艺术与科学学院学生参加。用试验设计术语说，每个处理组合的样本容量为 2，则我们有 2 个**复制**(replications)。其他的复制以及更大的样本容量也比较容易得到，为了减少本例中的计算，我们选择 2 个复制。

试验设计要求从三个毕业学员中每一个随机地选取打算读研的 6 名学生。然后将每个学院的两个学生随机分配给每个辅导课程，这样在研究中共用了 18 名考生。

假定所有学生是随机选取的，参加辅导课程并且参加 GMAT 考试分数见表 13.11。

利用表 13.11 的数据进行方差分析可以回答下面的问题：

- **主影响(因子 A)**：辅导课程对 GMAT 考分的影响是否有不同?
- **主影响(因子 B)**：毕业学院对 GMAT 考分的影响是否有不同?

· **交互影响(因子 A 和 B)**：是否某些学院的考生参加一种辅导课程比参加另一种辅导课程的 GMAT 考分要高?

表 13.11　两个因子试验的 GMAT 分数


CD 光盘数据
GMAT

		因子 B：辅导课程		
		商学	工学	艺术与科学
因子 A：学院	3 小时复习	500	540	480
		580	460	400
	1 天课程	460	560	420
		540	620	480
	10 周强化班	560	600	480
		600	580	410

交互作用(interaction)是指一种新的影响。因为我们采用了因子试验，现在我们可以研究这种影响。如果交互影响对 GMAT 考分有明显作用，那么我们就能得出辅导课程效果依赖于毕业学院的结论。

方差分析方法

两个因子试验的 ANOVA 试验类似于纯随机试验和随机区组试验，因为我们还是将平方和与自由度按它们的来源分解。两个因子试验平方和分解的公式如下：

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE} \tag{13.29}$$

平方和与自由度的分解汇总于表 13.12。利用下列的记号：

a = 因子 A 的水平个数；

b = 因子 B 的水平个数；

r = 复制个数；

n_T = 试验中观察值的总的数目；$n_T = abr$。

表 13.12　有 r 个复制的两个因子试验的 ANOVA 表

方差来源	平方和	自由度	均方值	F
因子 A	SSA	$a-1$	$\text{MSA} = \dfrac{\text{SSA}}{a-1}$	$\dfrac{\text{MSA}}{\text{MSE}}$
因子 B	SSB	$b-1$	$\text{MSB} = \dfrac{\text{SSB}}{b-1}$	$\dfrac{\text{MSB}}{\text{MSE}}$
交互作用	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \dfrac{\text{SSAB}}{(a-1)(b-1)}$	$\dfrac{\text{MSAB}}{\text{MSE}}$
误差	SSE	$ab(r-1)$	$\text{MSE} = \dfrac{\text{SSE}}{ab(r-1)}$	
总计	SST	n_T-1		

计算与结论

为了计算检验因子 A、因子 B 和交互作用的显著性的 F 统计量，我们需要计算 MSA、MSB、MSAB 和 MSE。为了计算这四个均方，我们必须首先计算 SSA、SSB、SSAB 和 SSE。

这样我们也可计算 SST。为了简化这个表达式，我们分五步进行计算。除了先前定义的 a、b、r 和 n_T，将使用下面的符号：

x_{ijk} = 对应于来自因子 A 的第 i 个处理和因子 B 的第 j 个处理的第 k 个复制的观察值；

$\bar{x}_{i\cdot}$ = （因子 A）第 i 个处理样本均值；

$\bar{x}_{\cdot j}$ = （因子 B）第 j 个处理样本均值；

$\bar{x}_{ij}$ = 对应于（因子 A）第 i 个处理样本均值和（因子 B）第 j 个处理样本均值组合的样本均值；

$\bar{\bar{x}}$ = 所有 n_T 个观察值的总样本均值。

步骤 1. 计算总的平方和；

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk}-\bar{\bar{x}})^2 \tag{13.30}$$

步骤 2. 计算因子 A 的平方和；

$$\text{SSA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot}-\bar{\bar{x}})^2 \tag{13.31}$$

步骤 3. 计算因子 B 的平方和；

$$\text{SSB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j}-\bar{\bar{x}})^2 \tag{13.32}$$

步骤 4. 计算交互作用的平方和；

$$\text{SSAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{\bar{x}})^2 \tag{13.33}$$

步骤 5. 计算误差平方和。

$$\text{SSE} = \text{SST} - \text{SSA} - \text{SSB} - \text{SSAB} \tag{13.34}$$

表 13.13 列出了试验所收集的数据以及不同总和便于计算平方和。利用公式(13.30)到(13.34)，我们得到 GMAT 两个因子试验中下列的平方和：

步骤 1. $\text{SST} = (500-515)^2 + (580-515)^2 + (540-515)^2 + \cdots + (410-515)^2 = 82\,450$

步骤 2. $\text{SSA} = (3)(2)[(493.33-515)^2 + (513.33-515)^2 + (538.33-515)^2] = 6\,100$

步骤 3. $\text{SSB} = (3)(2)[(540-515)^2 + (560-515)^2 + (445-515)^2] = 45\,300$

步骤 4. $\text{SSAB} = 2[(540-493.33-540+515)^2 + (500-493.33-560+515)^2 + \cdots + (445-538.33-445+515)^2] = 11\,200$

步骤 5. $\text{SSE} = 82\,450 - 6\,100 - 45\,300 - 11\,200 = 19\,850$

表 13.13　GMAT 两个因子试验的汇总数据

处理交互作用总计		因子 B：毕业学院			行总计	因子 A 均值
		商学	工学	艺术与科学		
因子 A：辅导课程	3 小时的复习	500 580 1 080 $\bar{x}_{11}=\frac{1\,080}{2}=540$	540 460 1 000 $\bar{x}_{12}=\frac{1\,000}{2}=500$	480 400 880 $\bar{x}_{13}=\frac{880}{2}=440$	2 960	$\bar{x}_{1\cdot}=\frac{2\,960}{6}=493.33$
	1 天的课程	460 540 1 000 $\bar{x}_{21}=\frac{1\,000}{2}=500$	560 620 1 180 $\bar{x}_{22}=\frac{1\,180}{2}=590$	420 480 900 $\bar{x}_{23}=\frac{900}{2}=450$	3 080	$\bar{x}_{2\cdot}=\frac{3\,080}{6}=513.33$
	10 周强化班	560 600 1 160 $\bar{x}_{31}=\frac{1\,160}{2}=580$	600 580 1 180 $\bar{x}_{32}=\frac{1\,180}{2}=590$	480 410 890 $\bar{x}_{33}=\frac{890}{2}=445$	3 230	$\bar{x}_{3\cdot}=\frac{3\,230}{6}=538.33$
	列总计	3 240	3 360	2 670	9 270 ← 所有总计	
	因子 B 均值	$\bar{x}_{\cdot 1}=\frac{3\,240}{6}=540$	$\bar{x}_{\cdot 2}=\frac{3\,360}{6}=560$	$\bar{x}_{\cdot 3}=\frac{2\,670}{6}=445$	$\bar{x}=\frac{9\,270}{18}=515$	

表 13.14 列出这些平方和除以他们相应的自由度并给出检验两个主要因子（辅导课程和毕业学院)和交互作用的适当均值。用于检验辅导课程差异的比率为 1.38。当 $\alpha=0.05$，分子自由度为 2，分母自由度为 9 时 F 临界值为 4.26。而 $F=1.38$，我们不能拒绝原假设，得出参加三个辅导课程之间没有明显差异。但是，对于毕业学院的影响，$F=10.27$ 超过临界值 4.26。因此，方差分析的结果能帮助我们得出毕业于三个不同学院学生 GMAT 考试分数有差异，也就是说，三个毕业于不同学院的学生参加相同课程而 GMAT 考试不一样。最后，交互作用值为 $F=1.27$（$\alpha=0.05$ 的临界值 $F=3.63$)意味着我们不能判定显著的交互作用。所以，我们没有理由相信参加三种辅导课程对于不同学院的学生的 GMAT 考试影响不一样。

表 13.14　GMAT 的两个因子研究的 ANOVA 表

方差来源	平方和	自由度	均方差	F
因子 A	6 100	2	3 050	3 050/2 206 = 1.38
因子 B	45 300	2	22 650	22 650/2 206 = 10.27
交互作用	11 200	4	2 800	2 800/2 206 = 1.27
误差	19 850	9	2 206	
总计	82 450	17		

毕业学员是一个显著的因子。查看表 13.13 中的计算，我们得到样本均值：商学院的学生 $\bar{x}_{\cdot 1}=540$，工学院的学生$\bar{x}_{\cdot 2}=560$，文艺与科学学院的学生$\bar{x}_{\cdot 3}=445$。因此，可以进行个别处理均值的检验；通过观察这三个样本均值，我们期望商学院和工学院的学生之间没有差别。但是，艺术与科学学院的学生参加辅导课后的 GMAT 明显比其他学院低。也许这些观察值将使这所大学考虑其他方法来帮助这个学院的学生参加 GMAT 考试。

因为在中型和大型的因子试验中涉及到大量计算误差，计算机对方差分析计算的完成于汇总起重要作用。图 13.6 是 GMAT 两个因子试验方差分析的 Minitab 计算机输出的结果。

SOURCE	DF	SS	MS	F	P
Factor A	2	6100	3050	1.38	0.299
Factor B	2	45300	22650	10.27	0.005
Interaction	4	11200	2800	1.27	0.350
Error	9	19850	2206		
Total	17	82450			

图 13.6　GMAT 两个因子设计的 Minitab 输出结果

练　习

方　法

41. 一个涉及到因子 A 的两个水平和因子 B 的三个水平的因子试验产生下面数据：



		因素 B		
		水平 1	水平 2	水平 3
因子 A	水平 1	135	90	75
		165	66	93
	水平 2	125	127	120
		95	105	136

检验显著的主影响与交互作用，取 $\alpha=0.05$。

42. 对有四个水平的因子 A，三个水平的因子 B 和三个复制的因子试验进行计算，得到如下数据：SST = 280，SSA = 26，SSB = 23，SSAB = 175。建立 ANOVA 表并对主影响和交互影响进行检验，取 $\alpha=0.05$。

应　用

43. 一家邮购公司设计了一个因子试验以检验杂志广告大小以及广告方案对于收到邮购要求数目(单位 1 000 个)的影响。考察了三种广告方案和两种不同大小的广告，得到以下数据。利用因子设计的 ANOVA 方法检验广告设计类型、广告大小或交互作用的显著影响，取 $\alpha=0.05$。

		广告大小	
		小	大
设计	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

44. 一家游乐公园想通过使游客有效地上下游乐车来减少等待时间，提出了两种上下游乐车的方法。考虑各种游乐车类型的潜在差异以及上下方法与它们之间的交互作用，设计一个因子试验。利用下列数据，检验上下方法和游乐车类型以及交互作用的显著影响，取 $\alpha=0.05$。

	游乐车类型		
	木马	过山车	摩天轮
	41	52	50
方法 1	43	44	46
	49	50	48
方法 2	51	46	44

45. 美国劳工统计局收集不同职业的男性女性的收入情况。假定 *The Tampa Tribune* 的一名记者想调查作为财务管理、计算机程序员、药剂师等不同职业男女之间周薪是否有明显的差异，从每个职业中分别选取 5 名男子和 5 名女子组成样本，样本中每个人的周薪记录下来，得到以下数据：



CD 光盘数据
Salaries

周薪（美元）	职业	性别
872	财务管理	男
859	财务管理	男
1 028	财务管理	男
1 117	财务管理	男
1 019	财务管理	男
519	财务管理	女
702	财务管理	女
805	财务管理	女
558	财务管理	女
591	财务管理	女
747	计算机程序员	男
766	计算机程序员	男
901	计算机程序员	男
690	计算机程序员	男
881	计算机程序员	男
884	计算机程序员	女
765	计算机程序员	女
685	计算机程序员	女
700	计算机程序员	女
671	计算机程序员	女
1 105	药剂师	男
1 144	药剂师	男
1 085	药剂师	男
903	药剂师	男
998	药剂师	男
813	药剂师	女
985	药剂师	女
1006	药剂师	女
1034	药剂师	女
817	药剂师	女

取显著性水平 $\alpha=0.05$，检验职业、性别以及交互作用的显著影响。

46. *The Accounting Review* 的一份报告研究两种时间压力的水平(低和中)与三种知识水平(初级、一般、熟练)对税收的关键词的选择上的各自影响和交互作用。调查的内容是一个税收案例，该案例包含一系列事实、税收问题和有 1 336 个关键词组成的关键词目录。要求被询问者选择他们认为能解决该税收案例的关键词。在试验之前，税收专家判定在目录中有 19 个相近的关键字。在初级一组中所选择的关键词有很少或没有一般的和熟练的知识，而在知识水平一般的一组中选择的明显是一般的知识，没有或很少有熟练的知识，在知识水平熟练的一组中所选择明显是一般的知识和熟练的知识。一般的知识包括了解应用税收条款以及描述这些条款相应项，熟练知识是指了解指导税收研究者研究的相关关键字的条款。对于时间压力不紧的情况，调查者要求在 25 分钟完成这个问题，在这个时间内完成将绰绰有余；对于时间压力中等的情况下，调查者要求在 11 分钟完成这个案例(*The Accounting Review*, January 1995)。假定从 6 个处理组合中每个选取 25，每个组合中样本均值如下：

		知识水平		
		初级	一般	熟练
时间压力	低	1.13	1.56	2.00
		(1.12)	(1.33)	(1.54)
	中	0.48	1.68	2.86
		(0.80)	(1.36)	(1.80)

利用 ANOVA 方法检验时间压力、知识水平和交互作用的显著检验。取显著性水平为 0.05，该试验的平方和为 327.50。

小　结

本章中我们介绍了如何用方差分析检验几个总体或处理方法之间的差异。我们介绍了纯随机设计、随机区组设计和两个因子试验。纯随机设计和随机区组设计用于对单个因子均值之差做推断。随机区组设计的目的是将外来的方差来源剔除误差项。这样的随机区组划分给出真实误差的更好的估计，并给出判断该因子总体或处理均值之间是否有明显差异的更好的检验。

我们介绍了用于方差分析和试验设计的检验统计量的基础是建立总体方差 σ^2 的两个独立估计量。在单因子情形下，一个估计量是基于处理间的方差；该估计量仅当均值 μ_1，μ_2，…，μ_k 相等时，才给出 σ^2 的无偏估计量；σ^2 的第二个估计量是基于处理内部的方差，该估计量永远给出 σ^2 的无偏估计量。通过计算这两个估计量的比例(F 统计量)，我们建立了判断能否拒绝总体或处理均值相等的假设的拒绝域。在所有讨论过的试验设计中，平方和与自由度按其来源而进行分解，使我们能够计算方差分析计算和检验适当的数值。我们还介绍了费舍尔 LSD 方法和 Bonferroni 修正方法如何运用到对配对比较检验以决定哪些均值有差异。

术语释义

ANOVA 表：一种用来汇总方差分析计算和结果的表，它包括显示方差来源、平方和、自由度、均方和值。

分解：把总平方和与自由度划分成不同组成部分的过程。

多重比较方法：用于将进行匹配的总体均值加以比较的统计方法。

比较性第一类错误概率：对应于单个配对比较的第一类错误概率。

试验性第一类错误概率：几个配对比较中至少有 1 个犯第一类错误的概率。

因子：独立变量的另一说法。

处理：不同的因子水平。

单因子试验：只涉及到有 k 个总体或处理的一个因子的试验。

试验单元：试验中研究的对象。

纯随机设计：处理被随机地分配给试验单元的一种试验设计。

区组划分：将每个处理中相同或相近的化为一组的过程。区组划分的目的是从误差项中剔除外来方差，并因此给出总体或处理均值之间差异的更好的检验方法。

随机区组设计：采用区组划分的一种试验设计。

因子试验：对两个或两个以上因子统计推断的一种试验设计。

复制：在一个试验中每个试验的重复次数。

交互作用：当一个因子水平与另一个因子水平发生作用时，对反应变量的影响。

重要公式

检验 k 个总体均值的相等性

第 j 个处理的样本平均值

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \tag{13.1}$$

第 j 个处理的样本方差

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_j - 1} \tag{13.2}$$

总样本平均值

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \tag{13.3}$$

$$n_T = n_1 + n_2 + \cdots + n_k \tag{13.4}$$

处理均方

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} \tag{13.7}$$

处理平方和

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2 \tag{13.8}$$

误差均值

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} \tag{13.10}$$

误差平方和

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \tag{13.11}$$

总体均值相等的检验统计量

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13.12}$$

总的平方和

$$\text{SST} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \tag{13.13}$$

总的平方和的分解

$$\text{SST} = \text{SSTR} + \text{SSE} \tag{13.14}$$

多重比较方法

检验统计量

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{13.16}$$

费舍尔 LSD 方法

$$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \tag{13.17}$$

纯随机设计

处理平方和

$$\text{MSTR} = \frac{\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2}{k-1} \tag{13.20}$$

误差均方

$$\text{MSE} = \frac{\sum_{j=1}^{k} (n_j - 1)s_j^2}{n_T - k} \tag{13.21}$$

***F* 统计量**

$$F = \frac{\text{MSTR}}{\text{MSE}} \tag{13.23}$$

随机区组设计

总的平方和

$$\text{SST} = \sum_{i=1}^{b}\sum_{j=1}^{k}(x_{ij} - \bar{\bar{x}})^2 \tag{13.25}$$

处理平方和

$$\text{SSTR} = b\sum_{j=1}^{k}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.26}$$

区组平方和

$$\text{SSBL} = k\sum_{i=1}^{b}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.27}$$

误差平方和

$$\text{SSE} = \text{SST} - \text{SSTR} - \text{SSBL} \tag{13.28}$$

因子试验

总的平方和

$$\text{SST} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{\bar{x}})^2 \tag{13.30}$$

因子 A 的平方和

$$\text{SSA} = br\sum_{i=1}^{a}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2 \tag{13.31}$$

因子 B 的平方和

$$\text{SSB} = ar\sum_{j=1}^{b}(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \tag{13.32}$$

交互作用的平方和

$$\text{SSAB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2 \tag{13.33}$$

误差平方和

$$\text{SSE} = \text{SST} - \text{SSA} - \text{SSB} - \text{SSAB} \tag{13.34}$$

补充练习

47. 两个居住区 4 小时销售最新的报价(单元：千美元)的简单随机样本的数据如下：

地区 1	地区 2
92	90
89	102
98	96
105	88

a. 利用第 10 章介绍的方法检验这两个地区平均报价是否一样，取 $\alpha = 0.05$。

b. 利用 ANOVA 方法检验平均报价是否一样。将你的分析与(a)的结果比较，取 $\alpha = 0.05$。

c. 假设从另一居住地收集的数据。从第三个地区的报价简单随机样本是 81 000 美元，86 000 美元，75 000 美元和 90 000 美元。对于这三个地区平均报价是否一样?取 $\alpha = 0.05$。

48. 在目前的市场上体育比赛用的自行车以及一般载人的自行车的购买者发现有很大的选择余


CD 光盘数据
Trucks

地。对大多数购买者来说，其中一个也是最重要的一个因素是自行车的转售价格。下面的表中给出 10 种体育比赛用的自行车、10 种轻便型自行车以及 10 种重型自行车 2 年之后的转售价格(*Kilplinger's New Cars & Trucks 2000 Buyer's Guide*)。

比赛用车	转售价格	轻便车	转售价格
Chevrolet Blazer LS	55	Chevrolet S-10 Extended Cab	46
Ford Explorer Sport	57	Dodge Dakota Club Cab Sport	53
GMC Yukon XL 1500	67	Ford Ranger XLT Regular Cab	48
Honda CR-V	65	Fork Ranger XLT Supercab	55
Isuzu VehiCross	62	GMC Sonoma Regular Cab	44
Jeep Cherokee Limited	57	Isuzu Hombre Spacecab	41
Mercury Mountaineer	59	Mazda B4000 SE Cab Plus	51
Nissan Pathfinder XE	54	Nissan Frontier XE Regular Cab	51
Toyota 4Runner	55	Toyota Tacoma Xtracab	49
Toyota RAV4	55	Toyota Tacoma Xtracab V6	50

重型车	转售价格
Chevrolet K2500	60
Chevrolet Silverado 2500 Ext	64
Dodge Ram 1500	54
Dodge Ram Quad Cab 2500	63
Dodge Ram Regular Cab 2500	59
Ford F150 XL	58
Ford F-350 Super Duty Crew Cab XL	64
GMC New Sierra 1500 Ext Cab	68
Toyota Tundra Access Cab Limited	53
Toyota Tundra Regular Cab	58

取显著性水平 $\alpha=0.05$，检验三种类型的自行车的平均转售价格是否相同。

49. 下面是关于在食品、零售和个人护理行业的大型公司的 12 名董事长的年龄的数据(*Advertising Age*, December 1, 1997)：


CD 光盘数据
AgeExec

公司	行业	董事长	年龄
Campbell Soup Co.	食品	Dale F. Morrison	48
General Mills	食品	Stephen W. Sanger	51
Kellogg Co.	食品	Arnold G. Langbo	59
RJR Nabisco	食品	Stephen F. Goldstone	51
Estee Lauder Cos.	个人护理	Leonard A. Lauder	64
Gillette Co.	个人护理	Alfred M. Zeien	67
Procter & Gamble Co.	个人护理	John E. Pepper	59
Unilever	个人护理	Morris Tabaksblat	59
Federated Department Stores	零售	James W. Zimmerman	53
J. C. Penny Co.	零售	James E. Oesterreicher	55
Sears Roebuck & Co.	零售	Arthur C. Martinez	57
Kmart Corp.	零售	Floyd Hall	58

在显著性水平 $\alpha=0.05$ 下，检验三个行业的公司的董事长的平均年龄是否存在明显的差异?

50. 《小企业管理》(*Jorunal of Small Business Management*)报道的一个研究表明，自己当老板与被人雇用相比较，不能感受到高职位的满足感。在这个研究中，工作满足感被分为 18 个等级，并且每个项目都用 Likert 测量法评定。评定范围分为 1—5 个选项，从十分强烈到十分不满。在这次评定中，高分意味着较高的工作满足感。18 项评定的总和从 18—90，用这个总和作为评定工作满足感的尺度（*Journal of Small Business Management*, October 1997）。假定用这种方法来测量律师、理疗师、木工师和系统分析师职业的满足感，由 40 个人组成的样本中得到每种职业的结果资料如下：


CD 光盘数据
SatisJob

律师	理疗师	木工师	系统分析师
44	55	54	44
42	78	65	73
74	80	79	71
42	86	69	60
53	60	79	64
50	59	64	66
45	62	59	41
48	52	78	55
64	55	84	76
38	50	60	62

在显著性水平 $\alpha=0.05$ 下，检验四个职业的满足感是否存在差异。

51. Crown Plaza 旅馆和度假胜地对全国范围的风景居住区的旅馆提供特殊周末价格。全国三个不同地区的 30 家旅馆提供了房间价格的数据(*USA Today*, April 14, 2000)。


CD 光盘数据
Resorts

西部	价格（美元）	南部	价格(美元)	西北部	价格(美元)
Albuquerque	89	Atlanta	105	Albany	89
Irvine	79	Dallas	80	Boston	139
Las Vegas	119	Greenville	79	Hartford	85
Los Angeles	99	Houston	79	New York	159
Palo Alto	109	Jackson	69	Philadelphia	99
Phoenix	149	Macon	69	Pittsfield	99
Portland	79	Miami	89	Providence	149
San Francisco	139	Orlando	119	Washington	159
San Jose	99	Richmond	109	White Plains	109
Seattle	119	Tampa	119	Worchester	124

在显著性水平 $\alpha=0.05$ 下，检验三个地区的房间平均价格是否相同。

52. 由会计事务所 Altschuler，Melvoin &Glasser 对广告代理人的第六个年度调查表明，广告代理人可期待另一个有较好的广告牌回报年份（*Advertising Age*，December 1, 1997）。为了调查设计总监的每年的报酬是否存在差异，假设从四个地区：西、南、中北和东北各选出 10 个设计总监作为一个样本。每个人的基本年薪样本数据如下：



CD 光盘数据
AirDir

西	南	中北	东北
60.9	50.8	49.5	65.9
45.9	39.6	42.3	58.6
62.1	44.2	35.5	49.3
66.6	40.0	49.1	52.9
68.0	53.9	56.7	48.5
65.0	45.4	41.4	52.9
49.4	61.1	51.3	52.4
62.3	42.3	49.4	48.1
62.6	38.4	42.1	46.5
57.2	38.3	55.7	45.9

在显著性水平 $\alpha=0.05$ 下，检验四个地区的设计总监的基本年薪是否相同。

53. 美国橄榄球联赛对每个候选人按其位置评比打分，打分范围从 5 分到 9 分。评定分数如下：8—9 表示应该立即参加比赛；7.0—7.9 表示可以打比赛；6.0—6.9 表示应作为二线队员储备；5.0—5.9 表示应进入俱乐部进行培训。下面表是关于对三个位置上的 40 名新的候选人进行打分的资料（*USA Today*，April 14, 2000）。检验球员不同位置分值之间是否有差异？



CD 光盘数据
NFL

接球手		后卫		截球手	
姓名	分数	姓名	分数	姓名	分数
Peter Warrick	9.0	Cosey Coleman	7.4	Chris Samuels	8.5
Plaxico Burress	8.8	Travis Claridge	7.0	Stockar McDougle	8.0
Sylvester Morris	8.3	Kaulana Noa	6.8	Chris McIngosh	7.8
Travis Taylor	8.1	Leander Jordan	6.7	Adrian Klemm	7.6
Laveranues Coles	8.0	Chad Clifton	6.3	Todd Wade	7.3
Dez White	7.9	Manula Savea	6.1	Marvel Smith	7.1
Jerry Porter	7.4	Ryan Johanningmei	6.0	Michael Thompson	6.8
Ron Dugans	7.1	Mark Tauscher	6.0	Bobby Williams	6.8
Todd Pinkston	7.0	Blaine Saipaia	6.0	Darnell Alford	6.4
Dennis Northcutt	7.0	Richard Mercier	5.8	Terrance Beadles	6.3
Anthony Lucas	6.9	Damion McIntosh	5.3	Tutan Reyes	6.1
Darrell Jackson	6.6	Jeno James	5.5	Greg Robinson-Ran	6.0
Danny Farmer	6.5	Al Jackson	5.5		
Sherrod Gideon	6.4				
Trevor Gaylor	6.2				

54. 在纯随机试验设计中，对三种牌子的纸巾的吸水能力检验。用相同规格的纸巾，并将每种牌子的纸巾分为四部分对其检验。吸水能力检验如下，取显著性水平为 0.05，三种牌子的吸水能力是否有显著差异?

牌子		
x	y	z
91	99	83
100	96	88
88	94	89
89	99	76

55. 下面是道琼斯工业指数中的 6 个主要项的 4 年变动的百分比。问主要项对股票市场性能的影响是否显著?取 $\alpha=0.05$。



CD 光盘数据
MktPert

第一年	第二年	第三年	第四年
10.9	-18.9	15.2	4.3
-15.2	4.8	6.1	14.6
-16.7	-27.6	38.3	17.9
-17.3	-3.1	4.2	14.9
-9.2	19.6	20.3	-3.7
27.7	22.6	2.3	11.8
27.0	-4.3	20.3	4.2
13.7	2.1	33.5	26.0

56. 对一种新产品提出三种不同的装配方法。用纯随机试验设计来判断哪种方法每个小时组装产品数目最多。随机选择 30 个工人，并将其分别分配给其中一种方法。每个方法生产的产品数如下:



CD 光盘数据
Assembly

方法		
A	B	C
97	93	99
73	100	94
93	93	87
100	55	66
73	77	59
91	91	75
100	85	84
86	73	72
92	90	88
95	83	86

根据这些数据，检验每种方法生产的产品数量是否相等，取 $\alpha=0.05$。

57. Hargreaves 汽车零件有限公司想比较四种不同类型的刹车套管的平均使用里程数。每种类型生产 30 个套管，并且安装在出租车队中。每个套管一直使用到不能满足联邦安全标准的要求为止，并记录下其里程数，计算每种类型的套管的平均里程数。数据如下：

类型	样本容量	样本均值	样本标准差
A	30	32 000	1 450
B	30	27 500	1 525
C	30	34 200	1 650
D	30	30 300	1 400

检验相应的总体均值是否相等，取 $\alpha=0.05$。

58. 电子玩具和计算器的电池的生产商考虑三种新的电池设计。对这三种设计进行检验，判断每种电池的平均寿命是否相同？

设计 A	设计 B	设计 C
78	112	115
98	99	101
88	101	100
96	116	120

检验总体均值是否相等，取 $\alpha=0.05$。

59. 为了调查购买者的光顾行为，每个购物者被分为不光顾者、不经常光顾者或经常光顾者。在调查中，对于每个购物者进行测量，确定购物者在商场中的舒适程度，分值高表示舒适程度越高，假定从相关的研究中得到如下资料：



CD 光盘数据 Browsing

不光顾者	不经常光顾者	经常光顾者
4	5	5
5	6	7
6	5	5
3	4	7
3	7	4
4	4	6
5	6	5
4	5	7

a. 取 $\alpha=0.05$，检验三种类型的光顾者的舒适水平是否有差异。

b. 利用费舍尔的 LSD 方法比较不光顾者和不经常光顾者的舒适水平，取 $\alpha=0.05$，得出你的结论。

60. 一研究所检验三种牌子的汽油每加仑里程数的性能。因为对于不同牌子的汽车，不同汽油的

性能不一样，选取五种牌子的汽车，在试验中把每种牌子作为一个区组。也就是说，每个牌子的汽油用各种不同牌子汽车进行检验。试验结果(每加仑里程数)如下：

		汽油牌子		
		Ⅰ	Ⅱ	Ⅲ
汽车	A	18	21	20
	B	24	26	27
	C	30	29	34
	D	22	25	24
	E	20	23	24

取显著性水平 $\alpha=0.05$，这三种牌子的汽油的平均每加仑里程数性能是否存在差异？

61. 根据练习 60 中的试验数据，用纯随机设计 ANOVA 方法进行分析。将其结果与练习 60 比较，对于设法移除区组影响的优点是什么？

62. 每月 *Internet Magazine* 报告 100 多家网络设备供应商(Internet service providers, ISP)，其目的是调查 ISP 的利用率以及下载一定数量的网页的联接速度(以秒为计)。下面是关于 22 家 ISP 的网站在美国、英国和欧洲下载的时间的数据《网络杂志》(*Internet Magazine*, January 2000)：


CD 光盘数据
ISP

ISP	英国	美国	欧洲
Abel Gratis	10.62	14.64	17.08
Breathe	11.67	14.14	19.86
btclick.com	12.12	16.43	21.30
Bun	11.13	14.09	15.83
Cable & Wireless Life	9.99	13.07	18.43
con X	12.63	15.97	22.12
Freebeeb	11.71	15.52	19.57
Free-Online	13.77	13.98	23.35
Freeserve	10.65	13.62	25.56
FreeUK	12.20	14.96	18.95
Icom-Web	9.62	11.66	15.91
IPNet	13.82	16.70	22.86
I-way Soho	14.86	12.86	19.32
LineOne	12.01	17.82	21.88
Madasafish	13.38	15.59	19.61
NetDirect Online	11.71	15.52	19.57
Netscape Online	10.84	12.66	16.52
Screaming.net(BT Line)	13.23	15.91	23.08
Telinco Internet Services	12.83	15.34	18.76
UK Online	10.39	13.28	21.04
UKPeople	13.79	19.82	19.76
Virgin Net	12.17	15.47	21.94

取 $\alpha=0.05$，网站在不同国家平均下载时间是否有明显差异?

63. 设计一因子试验来检验两种计算机语言翻译外语所需时间是否有明显差异。在语言翻译中需要考虑一个重要因素：所翻译的语言种类，用两个系统翻译成三种不同语言：西班牙、法语和德语。翻译所需时间的资料如下：

	语言		
	西班牙语	法语	德语
系统 1	8	10	12
系统 2	12	14	16
	6	14	16
	10	16	22

检验语言翻译机、语言类型以及交互作用的差异是否显著?取 $\alpha=0.05$。

64. 一个制造商设计一个因子试验来判断两个机器生产出有缺陷的产品数目是否不同，并且判断是否有缺陷产品数量与每个机器所需的原材料是人工装载还是自动传输有关。下面数据给出生产有缺陷产品的数量。取 $\alpha=0.05$，机器类型、装载系统和交互作用的差异是否显著?

	装载系统	
	手工	自动
机器 1	30	30
	34	26
机器 2	20	24
	22	28

案例研究 1 WENTWORTH 医疗中心



CD 光盘数据 Medical1

作为对 65 岁以上的人长期研究的一部分，纽约北部地区的 Wentworth 医疗中心的社会学家和心理学家调查地理位置和患抑郁症之间的关系。由 60 个健康人组成一个样本，20 人居住在佛罗里达，20 人居住在纽约，20 人居住在北卡罗来纳。对每个人进行测量抑郁症标准化测验。收集到的数据如下:分值越高表示抑郁症程度越深。这些数据存储在 CD 光盘上的文件 Medical 1 中。



CD 光盘数据 Medical2

研究的第二部分是考察地理位置与 65 岁以上患有诸如关节炎、高血压和/或心脏失调慢性病之间的关系。对于这种身体状况的人也抽出 60 人。20 人居住在佛罗里达，20 人居住在纽约，20 人居住在北卡罗来纳。这次研究记录的抑郁症程度资料如下，这些数据在 CD 光盘上的文件 Medical 2 中。

Medical1 的数据			Medical2 的数据		
佛罗里达	纽约	北卡罗来纳	佛罗里达	纽约	北卡罗来纳
3	8	10	13	14	10
7	11	7	12	9	12
7	9	3	17	15	15
3	7	5	17	12	18
8	8	11	20	16	12
8	7	8	21	24	14
8	8	4	16	18	17
5	4	3	14	14	8
5	13	7	13	15	14
2	10	8	17	17	16
6	6	8	12	20	18
2	8	7	9	11	17
6	12	3	12	23	19
6	8	9	15	19	15
9	6	8	16	17	13
7	8	12	15	14	14
5	5	6	13	9	11
4	7	3	10	14	12
7	7	8	11	13	13
3	8	11	17	11	11

管理报告

1. 用描述统计方法汇总这两个研究的数据。对于抑郁症的得分，得出你初步观测结果。
2. 对这两个数据集使用方法分析，列出每种情况下的假设，得出你的结论。
3. 用推断法来说明单个处理均值的合理性。
4. 讨论这个研究的推广和你认为有用的其他分析。

案例研究 2　工业产品销售员的报酬



CD 光盘数据
IDSalary

在过去的 10 年里，《工业产品销售》（*Industrial Distribution*）一直在跟踪调查工业产品销售员的报酬。在 1997 年报酬调查中，有 358 名回答者的结果表明：有 27% 的回答者在销售额大于 4 000 万美元的销售公司工作，其中一般的工业销售员在 1 200 万美元的销售公司工作。在中小型公司（销售额在 600 万—2 000 万美元之间）工作的销售额的收入比在大公司工作要高。报酬最低的销售员在销售额小于 100 万美元的公司工

作。1996 年一般的户外销售员报酬为 50 000 美元，而一般的户内销售员报酬为 30 000 美元(*Industrial Distribution*, November, 1997)。假定在较大的旧金山地区工业产品销售员的一个分会，进行了关于会员资格的一次调查，以研究雇员资历与在户外或室内场所销售的年薪之间是否有关系。在这个调查中，被调查者分为三个资历水平：低(1—10 年)、中(11—20 年)和高(21 年及以上)。所采用的资料如下，完整的数据集包含 120 个观察值，存储在 CD 光盘的文件 IDSalary 中。

观察值	薪水(美元)	场所	资历
1	28 938	Inside	Medium
2	27 694	Inside	Medium
3	45 515	Outside	Low
4	27 031	Inside	Medium
5	37 283	Outside	Low
6	32 718	Inside	Low
7	54 081	Outside	High
8	23 621	Inside	Low
9	47 835	Outside	High
10	29 768	Inside	Medium
…	…	…	…
115	33 080	Inside	High
116	53 702	Outside	Medium
117	58 131	Outside	Medium
118	32 788	Inside	High
119	28 070	Inside	Medium
120	35 259	Outside	Low

管理报告

1. 用描述性统计汇总数据。
2. 不考虑销售员的资历，建立所有销售员平均年薪的 95% 置信区间。
3. 建立户外销售员平均年薪的 95% 置信区间，将你得到的结果与《工业产品销售》报告的全国值比较。
4. 建立户内销售员平均年薪的 95% 置信区间，将你得到的结果与《工业产品销售》报告的全国值比较。
5. 不考虑销售员的资历，建立户外销售员与户内销售员平均年薪的均值差的 95% 置信区间，得出你的结论是什么?
6. 用样本方差分析方法检验属于场所的显著差异,设显著性水平为 0.05,不考虑销售员的资历。
7. 用样本方差分析方法检验属于场所的显著差异,设显著性水平为 0.05,不考虑场所的影响。
8. 在显著性水平为 0.05 下，检验场所、资历和两者交互作用显著性差异并计算关于某个处理的均值，用推断法说明其合理性。

附录 13.1　用 Minitab 进行方差分析和试验设计

单因素观察研究和纯随机设计

在 13.2 节中我们介绍了利用观测得到的资料，用方差分析来检验个总体均值的相等性。在 13.5 节中我们介绍了在纯随机设计中搜集资料的情形下，如何用相同的方法来检验个总体均值的相等性。为了解释在上述情况下，如何用 Minitab 检验个总体均值的相等性，以 13.1 节中介绍的国家计算机产品公司为例，介绍了如何检验每个工厂的平均考试分数是否一致。将考试分数输入 Minitab 工作表的前三列：其中第一列为亚特兰大的数据，第二列为达拉斯的数据，第三列为西雅图的数据。利用 Minitab 按下列步骤得出输出结果，见图 13.4。


CD 光盘数据
NCP

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **ANOVA**；
步骤 3. 选择 **Oneway(Unstacked)**选项；
步骤 4. 出现 Oneway Analysis of Variance 对话框后：
在 **Responses(in separate columns)**框中输入 **C1 – C3**；
点击 **OK**。

随机区组设计

在 13.6 节中我们介绍了根据从随机区组设计得到的资料，如何用方差分析来检验个总体均值的相等性。为了说明对这类试验设计如何用 Minitab，我们介绍如何检验空中交通管制员在三个工作站的平均压力是否相等。将表 13.6 中的压力水平值输入 Minitab 工作表的第一列。对处理进行编号，1 表示系统 A，2 表示系统 B，3 表示系统 C，将处理编号输入工作表的第二列。最后，每个管制员对应的序号(1，2，3，4，5，6)输入工作表的第三列。因此，工作表的第一行值为 15，1，1；第二行为 15，2，1；第三行为 18，3，1；第四行为 14，1，2，以此类推。利用 Minitab 按下列步骤得出相应的 ANOVA 表的输出结果，见表 13.9。

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **ANOVA**；
步骤 3. 选择 **Two-way** 选项；
步骤 4. 出现 Two-way Analysis of Variance 对话框后：
在 **Response** 框中输入 C1；
在 **Row factor** 框中输入 C2；
在 **Column factor** 框中输入 C3；

选择 **Fit additive model**；
点击 **OK**。

因子试验

在 13.7 节中我们介绍了根据从因子试验中得来的数据，如何用方差分析来检验 k 个总体均值的相等性。为了解释如何用 Minitab 对这类试验设计来检验，我们以两因子 GMAT 试验为例来介绍如何分析这些数据。将表 13.11 中的 GMAT 分数输入 Minitab 工作表中的第一列；第二列为因子 A；第三列为因子 B。将因子 A 学员编号 1 为 3 小时的复习，2 为一天的程序，3 为 10 周的课程，将因子 A 编号值输入工作表的第二列。将因子 B 学员编号，1 为商学，2 为工学，3 为艺术与科学，将因子 B 的编号值输入第三列。因此，工作表的第一行值为 500，1，1；第二行值为 580，1，1；第三行值为 540，1，2；第四行值 460，1，2；依次类推。利用 Minitab 按下列步骤得出相应的 ANOVA 表的输出结果，见图 13.6。


CD 光盘数据
GMAT

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **ANOVA**；
步骤 3. 选择 **Two-way** 选项；
步骤 4. 出现 Two-way Analysis of Variance 对话框后：
在 **Response** 框中输入 C1；
在 **Row factor** 框中输入 C2；
在 **Column factor** 框中输入 C3；
点击 **OK**。

附录 13.2 用 Excel 进行方差分析和试验设计

单因素观察研究和纯随机设计

在 13.2 节中我们介绍了利用观测得到的资料，如何用方差分析来检验个总体均值的相等性。在 13.5 节中我们介绍了在纯随机设计中搜集资料的情形下，如何用相同的方法来检验个总体均值的相等性。为了解释在上述情况下，如何用 Excel 检验个总体均值的相等性，以 13.1 节中介绍的美国计算机产品公司为例，我们介绍了如何检验每个工厂的平均考试分数是否一致。将考试分数输入工作表 A、B、C 列中第二列到第七列，见图 13.7。注意工作表的 A、B、C 列第一行单元格为亚特兰大、达拉斯和西雅图。按照下面步骤得出的输出结果显示在单元 A10: G24；相应的 ANOVA 表的部分输出结果见表 13.2。


CD 光盘数据
NCP

步骤 1. 选择 **Tools** 下拉菜单；

步骤 2. 选择 **Data Analysis**；

步骤 3. 从 **AnalysisTools** 中选择 **Anova：Single-Factor**；

点击 OK。

步骤 4. 出现 Anova：Single-Factor 对话框后：

在 **Input Range** 框中输入 A1：C7；

选择 **Columns**；

选择 **Labels in First Row**；

选择 **Output Range**，并且在对话框中输入 A10；

点击 **OK**。

Atlanta	Dallas	Seattle
85	71	59
75	75	64
82	73	62
76	74	69
71	69	75
85	82	67

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Atlanta	6	474	79	34
Dallas	6	444	74	20
Seattle	6	398	66	32

ANOVA

Source of Variation	SS	df	MS	F	P-Value	F crit
Between Groups	516	2	258	9	0.002703	3.682317
Within Groups	430	15	28.66667			
Total	946	17				

图 13.7　利用 Excel 对 NCP 例子进行方差分析的输出结果

随机区组设计



CD 光盘数据 AirTrat

在 13.6 节中我们介绍了从随机区组设计得到的资料，如何用方差分析来检验个总体均值的相等性。为了说明对这类试验设计如何用 Excel，我们介绍如何检验空中交通管制员在三个工作站的平均压力是否相等。将表 13.6 的压力水平值输入工作表 B、C、D 列的第二行到第七行，见图 13.8。A 列第二行到第七行的单元格为每个管制员的序号(1,2,3,4,5,6)。利用 Excel 按下列步骤得到相应的 ANOVA 表的输出结果见表 13.9。

步骤 1. 选择 **Tools** 下拉菜单；

步骤 2. 选择 **Data Analysis**；

步骤 3. 从 **AnalysisTools** 中选择 **Anova**：**Two-Factor Without Replication**；
点击 **OK**；

步骤 4. 出现 Anova：Two-Factor Without Replication 对话框后：
在 **Input Range** 框中输入 A1: D7；
选择 **Labels**；
选择 **Output Range**，并且在对话框中输入 A10；
点击 **OK**。

Controller	System A	System B	System C
1	15	15	18
2	14	14	14
3	10	11	15
4	13	12	17
5	16	13	16
6	13	13	13

Anova: Two-Factor Without Replication

Summary	Count	Sum	Average	Variance
1	3	48	16	3
2	3	42	14	0
3	3	36	12	7
4	3	42	14	7
5	3	45	15	3
6	3	39	13	0
System A	6	81	13.5	4.3
System B	6	78	13	2
System C	6	93	15.5	3.5

ANOVA						
Source of Variation	SS	df	MS	F	P-Value	F crit
Rows	30	5	6	3.157895	0.057399	3.325837
Columns	21	2	10.5	5.526316	0.024181	4.102816
Error	19	10	1.9			
Total	70	17				

图 13.8　利用 Excel 对空中交通管制员压力测试的检验的输出结果

因子试验

在 13.7 节中我们介绍了根据从因子试验中得来的数据，如何用方差分析来检验个总体均值的相等性。为了解释如何用 Excel 对这类试验设计来检验，我们以两因子 GMAT 试验为例来介绍如何分析这些数据。将表 13.11 中的 GMAT 分数输入工作表的 B、C、D 列的第二行到第七行，见图 13.9。按下列步骤得到输出结果显示在单元格 A10: G45，相应的 ANOVA 表的部分输出结果显示在表 13.14 中。

	Business	Engineering	Arts and Sciences
3-hour review	500	540	480
	580	460	400
1-day program	460	560	420
	540	620	460
10-week course	560	600	480
	500	580	410

Anova: Two-Factor With Replication

SUMMARY	Business	Engineering	Arts and Sciences	Total
3-hour review				
Count	2	2	2	6
Sum	1080	1000	880	2960
Average	540	500	440	493.3333
Variance	3200	3200	3200	3946.667
1-day program				
Count	2	2	2	6
Sum	1000	1180	900	3080
Average	500	590	450	513.3333
Variance	3200	1800	1800	5386.667
10-week course				
Count	2	2	2	6
Sum	1160	1180	890	3230
Average	580	590	445	538.3333
Variance	800	200	2450	5936.667
Total				
Count	6	6	6	
Sum	3240	3360	2670	
Average	540	560	445	
Variance	2720	3200	1510	

ANOVA

Source of Variation	SS	df	MS	F	P-Value	F crit
Sample	6100	2	3050	1.382872	0.299436	4.256492
Columns	45300	2	22650	10.26952	0.004757	4.256492
Interaction	11200	4	2800	1.269521	0.350328	3.63309
Within	19850	9	2205.555556			
Total	82450	17				

图 13.9　利用 Excel 对两个因子 GMAT 试验检验的输出结果


CD光盘数据
GMAT

步骤 1. 选择 **Tools** 下拉菜单；

步骤 2. 选择 **Data Analysis**；

步骤 3. 从 **AnalysisTools** 中选择 **Anova：Two-Factor With Replication**；
点击 **OK**；

步骤 4. 出现 Anova：Two-Factor With Replication 对话框后：
在 **Input Range** 框中输入 A1:D7；
在 **Rows per sample** 框中输入 2；
选择 **Output Range**，并且在对话框中输入 A10；
点击 **OK**。

14 简单线性回归

统计实例:宝丽来公司

14.1 **简单线性回归模型**

回归模型和回归方程

估计回归方程

14.2 **最小二乘法**

14.3 **判定系数**

相关系数

14.4 **模型假设**

14.5 **显著性检验**

σ^2 的估计

t 检验

β_1 的置信区间

F 检验

关于显著性检验解释的几点注意

14.6 **利用估计回归方程进行估计和预测**

点估计

区间估计

y 的均值的置信区间估计

个别 y 值的预测区间估计

14.7 **计算机方法**

14.8 **残差分析:验证模型假定**

关于 x 的残差图

关于 $\hat{y}$ 的残差图

标准残差

正态概率图

14.9 **残差分析:异常值和有影响的观察值**

检测异常值

检测有影响的观察值

统计实例

宝丽来公司*

剑桥，马萨诸塞州

1947年，宝丽来公司创始人 Edwin Land 博士宣布一分钟成像技术已研发成功，公司已开展大众摄影的业务。宝丽来第一台照相机和第一卷胶卷诞生于1949年。在那以后，宝丽来公司在化学、光学和电子学方面进行试验和开发，生产出更高质量、更高可靠性和更为便利的摄影系统。

宝丽来公司的另一项主要业务是专门的和工业的照相技术，主要致力于宝丽来即时显像技术。这项技术在当今可视通讯环境下，逐渐成为日益增长的成像系统的主要组成部分。为此，宝丽来公司生产出大量不同种类的一次成像摄影系统、照相机、部件和胶片，以供专业摄影、工业、科学和医药应用。公司还在磁学、太阳镜、工业偏振镜、化工、传统涂料和全息摄影领域有业务。

用于衡量摄影材料感光度的测光计，可以提供胶片特性的信息，比如它的曝光时间范围。在宝丽来中心感光实验室中，科学家们将一次成像胶片置于适当的温度和湿度下，使之近似于消费者购买后的保存条件，然后再将其系统地抽样检验并进行分析。为了研究宝丽来彩色专业打印胶片感光速度和保存时间的关系，宝丽来中心感光实验室抽取保存时间从1—13个月（生产之后的时间）不等的胶片。数据表明胶片感光速度随着时间增加而递减，胶片感光速度和胶片保存时间的关系可以近似用一条直线或线性关系表示出。



回归分析帮助宝丽来提高胶片性能以满足顾客的需求。

© Joe Higgins/South-Western.

利用回归分析，宝丽来建立感光速度与保存时间之间的方程式。

$$\hat{y} = -19.8 - 7.6x$$

式中 $\hat{y}$——胶卷感光速度的变动；

x——胶卷的保存时间。

这个方程式表示每个月胶卷感光速度平均下降7.6个单元。把消费者购买和使用习惯与分析得来的信息相结合，将有助于宝丽来调整生产，提供满足顾客需要的胶卷。

在本章中，将介绍如何运用回归分析法建立起具有两个变量的方程式，如宝丽来例子中的胶片感光速度和保存时间。在下面几章中，我们将这一概念扩展到具有两个以上变量的情形。

---

*感谢摄影材料质量管理经理 Lawrence Friedman 提供统计应用资料。

管理决策经常是根据这两个或更多变量之间的关系作出的。例如，在考虑广告费用和销售额之间的关系后，销售经理想知道一定水平的广告费用能带来多少销售收入。又如，一家公用事业的公司想知道每月最高温度和用电量之间的关系，再根据下个月白天最高气温预报，来预测用电量。有时候，管理者要通过直觉来判断两个变量之间的关系。但是，如果能获得数据，可以利用**回归分析**统计方法建立一个方程来表示变量之间的关系。

在回归方法中，被预测变量称之为**因变量**(dependent variable)，进行预测的变量称之为**自变量**(independent variable)。例如，在分析广告费支出对销售收入的影响时，销售经理想预测销售收入，则销售收入为因变量，而用来预测收入的广告费是自变量。在统计符号上，y 代表因变量，x 代表自变量。

第一次用统计方法来研究两个个体间关系的是 Francis Galton(1822—1911)。Galton 对父亲和儿子的身高之间的关系感兴趣。Galton 的同事，Karl Pearson(1857—1936)利用该方法来研究 1 078 对父亲身高和儿子身高之间的关系。

在样本中我们只考虑两者的关系可以用一条直线近似表示出简单类型的回归分析，这种回归分析称之为**简单线性回归**(simple linear regression)。涉及两个或两个以上自变量的回归分析为多元回归分析。第 15 章和第 16 章将介绍包括曲线关系的多元回归分析和案例。

14.1　简单线性回归模型

Armand 比萨饼餐馆连锁店是经营意大利食品的餐馆，它在美国的 5 个州内都有分店。Armand 比萨餐饼馆连锁店的最佳位置是在大学校园附近。管理人员相信，这些饭店的季度销售收入(用 y 表示)与学生人数(用 x 表示)是正相关的。也就是说，在学生较多的校园附近的餐馆与在学生较少的校园附近的餐馆相比，季度销售收入更多。利用回归分析，我们能得出一个说明因变量 y 与自变量 x 之间关系的方程。

回归模型和回归方程

在 Armand 比萨饼餐馆连锁店例子中，每一个餐馆相应地都有一个 x 值(学生总数)和对应的 y 值(季节销售额)。描述 y 和 x 的关系的方程式被称为**回归模型**(regression model)。回归模型用于如下的简单线性回归中：

简单线性回归模型

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{14.1}$$

在简单线性模型中，y 是 x 的线性函数（$\beta_0 + \beta_1 x$ 部分）加上 ε。β_0 和 β_1 称为模型的参数，

ε 被称为误差项的随机变量。误差项说明了包含在 y 里面但不能被 x 和 y 之间的线性解释的变异性。

在 14.4 节中我们将讨论简单线性回归模型和 ε 假定。其中假定之一是 ε 的平均值或期望值为零。根据这一假定，y 的平均值或期望值 $E(y)$等于 $\beta_0 + \beta_1 x$，换句话说，y 的平均值是 x 的线性函数。这一方程式描述了 y 的平均值与 x 的关系，称此方程为**回归方程**(regression equation)。简单线性回归方程如下：

简单线性回归方程

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

简单线性回归方程是一条直线：β_0 是回归直线的 y 截距，β_1 是 y 斜率，$E(y)$是 y 的平均值或期望值。对于简单线性回归方程的几种回归直线见图 14.1。图 14.1 中，图 A 中的回归直线表示 y 的平均值与 x 正相关，较大的 x 值对应的 $E(y)$值也较大。图 B 中的回归直线表示 y 的平均值与 x 负相关，较大的 x 值对应的 $E(y)$值较小。图 C 中的回归直线表示 y 的平均值与 x 无关，即对于 x 的每个值，y 的平均值是相同的。

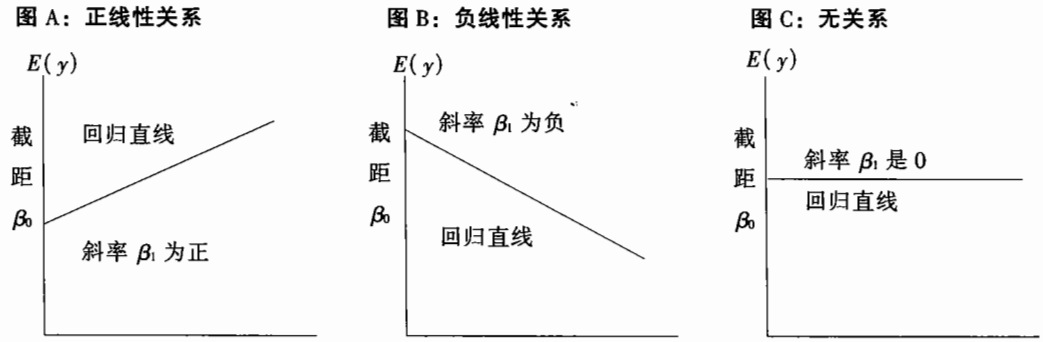


图 14.1　简单线性回归中的可能回归线

估计回归方程

如果参数 β_0 和 β_1 已知，对于一个已知的 x 值，我们能利用公式（14.2）来计算 y 的平均值。如果未知实际的参数值，可以通过样本数据进行估计。用样本统计量 b_0 和 b_1 来进行回归方程中的未知参数 β_0 和 β_1 的估计，我们将得到**估计回归方程**(estimated regression equation)。简单线性回归方程的估计如下：

估计简单线性回归方程

$$\hat{y} = b_0 + b_1 x \tag{14.3}$$

简单线性回归方程图被称为**估计的回归线**；b_0 是 y 的截距，b_1 是斜率，对于一个给定的 x，

$\hat{y}$ 是 y 的一个**估计值**。在下节中，我们介绍利用最小二乘法来计算估计回归方程中的 b_0 和 b_1。图 14.2 是简单线性回归的估计过程总结。

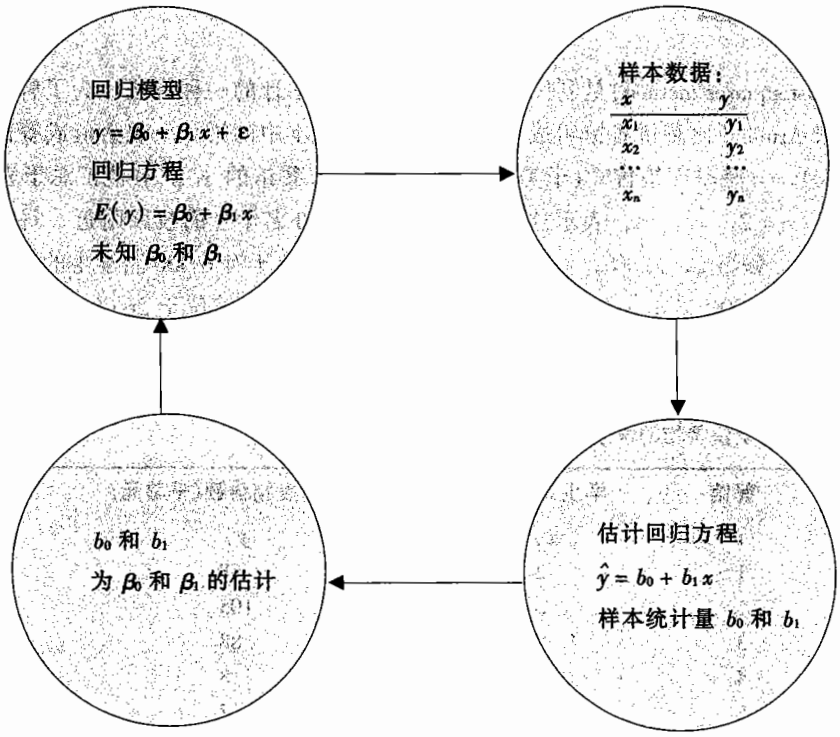


图 14.2　简单线性回归的估计过程

对于 β_0 和 β_1 的估计是一个统计过程，与第 7 章中的对 μ 的统计非常类似。β_0 和 β_1 是感兴趣的未知参数，用于估计未知参数的样本统计量。

评　注

回归分析不能被看做是在变量间建立一个因果关系的过程，回归分析只能表明变量是如何或者是以怎样的程度彼此联系在一起的。任何关于因果的结论必须建立在对个体或多个个体的有关应用的大量信息判断基础上的。

14.2 最小二乘法

最小二乘法(least squares method)是利用样本进行回归方程估计的一种方法。为了解释最小二乘法，假定由 10 家 Armand 比萨饼餐馆组成一个样本。对于样本中的第 i 个观察值或餐馆，x_i 是学生样本大小(千人)，y_i 是季度销售额(千美元)。样本中 10 家餐馆的 x_i 和 y_i 值汇总于表 14.1。从图中得出餐馆 1，$x_1=2$，$y_1=58$，表示这家餐馆位于有 2 000 名学生的校园附近，每个季度的销售额为 58 000 美元。餐馆 2，$x_2=6$，$y_2=105$，表示这家餐馆位于有 6 000 名学生的校园附近，每个季度的销售额为 105 000 美元。销售额最多的是餐馆 10，这家餐馆位于 26 000 名学生的校园附近，每季度的销售额为 202 000 美元。

表 14.1　10 家 Armand 比萨饼餐馆的学生人数和季度销售收入数据


CD 光盘数据
Armand's

餐馆	学生人数（千人）	季度销售额(千美元)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

图 14.3 是表 14.1 中的数据的散点图。学生人数的大小为横坐标，每季度销售额为纵坐标。根据横坐标自变量 x 和纵坐标因变量 y，可以得出回归分析的**散点图**(scatter diagrams)。可从散点图上观察数据，对变量之间可能存在的关系作出初步判断。

在简单线性回归中有两个变量：一个是自变量，另一个是因变量。

实际上，x 和 y 之间是正线性关系。所以我们可以选择简单线性回归模型来表示季度销售额和学生人数之间的关系。假如假定成立，我们利用表 14.1 中的样本数据得出简单线性回归方程中的和值。对于第 i 个餐馆，回归估计方程如下：

$$\hat{y}_i = b_0 + b_1 x_i \tag{14.4}$$

式中　$\hat{y}_i$——第 i 个餐馆季度销售额的估计值(千美元)；

b_0——估计回归直线的 y 的截距；

b_1——估计回归直线的斜率；

x_i——第 i 个餐馆的学生人数(千人)。

令 y_i 为第 i 家餐馆观察值(实际值)，公式(14.4)中的 $\hat{y}_i$ 为第 i 家餐馆的销售收入的估计值，样本中的每个餐馆有一个销售额 y_i 和 $\hat{y}_i$ 的估计值。为了使回归直线的估计与实际数据有很好的拟合，我们希望观察值与估计值之差要小。

最小二乘法利用样本数据，通过因变量 y_i 的**观察值**和**估计值**之间的离差平方和最小的方法来求得 b_0 和 b_1 值。最小二乘法的准则见公式(14.5)。

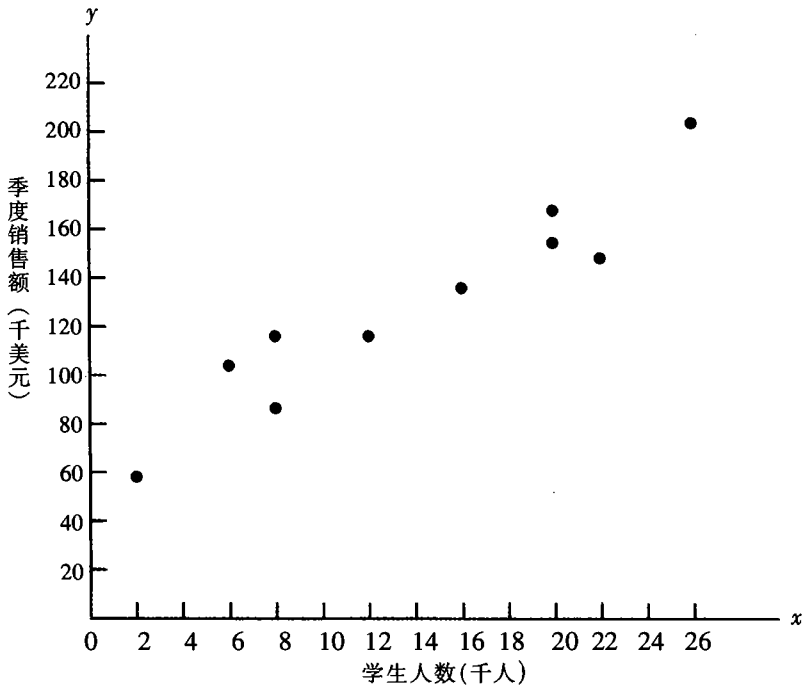


图 14.3　Armand 比萨饼餐馆的学生人数和季度销售收入的散点图

最小二乘准则

$$\min \sum (y_i - \hat{y}_i)^2 \tag{14.5}$$

式中　y_i——因变量的第 i 个观察值；

$\hat{y}_i$——因变量的第 i 个估计值。

Carl Friedrich Gauss(1777—1855)提出了最小二乘法。

通过微分学证明可知，要求得最小公式(14.5)的 b_0 和 b_1，可使用公式(14.6)和(14.7)。

回归方程估计的斜率和 y 截距*

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1\bar{x} \tag{14.7}$$

式中 x_i ——自变量的第 i 个观察值；

y_i ——因变量的第 i 个观察值；

$\bar{x}$ ——自变量的平均值；

$\bar{y}$ ——因变量的平均值；

n ——总的观察次数。

在用计算器计算时，在中间计算中尽可能多保留有效的数字，我们建议至少保留四位数。

对于 Armand 比萨饼餐馆例子，应用最小二乘法估计回归方程的必要计算步骤见表 14.2。对于由 10 家餐馆组成的样本，我们有 $n = 10$ 个观察值。因为公式(14.6)和(14.7)需要 $\bar{x}$ 和 $\bar{y}$，首先计算 $\bar{x}$ 和 $\bar{y}$。

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{1\,300}{10} = 130$$

利用公式(14.6)、(14.7)和表 14.2 中的数据，我们能计算出 Armand 比萨饼餐馆连锁店的估计回归方程中的斜率和截距。斜率(b_1)计算过程如下：

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_j - \bar{x})^2}$$

$$= \frac{2\,840}{568}$$

$$= 5$$

y 截距(b_0)的计算如下：

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 130 - 5(14)$$

$$= 60$$

* 计算可选择的公式为 $b_1 = \dfrac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$。在计算 b_1 时，如利用计算器可采用公式（14.6）。

于是，估计回归方程如下：

$$\hat{y} = 60 + 5x$$

图 14.4 给出在散点图中这个方程的直线形式。

表 14.2　Armand 比萨饼餐馆连锁店的最小二乘法估计回归方程计算

餐馆 i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
总计	140	1 300			2 840	568
	$\sum x_i$	$\sum y_i$			$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x_i - \bar{x})^2$

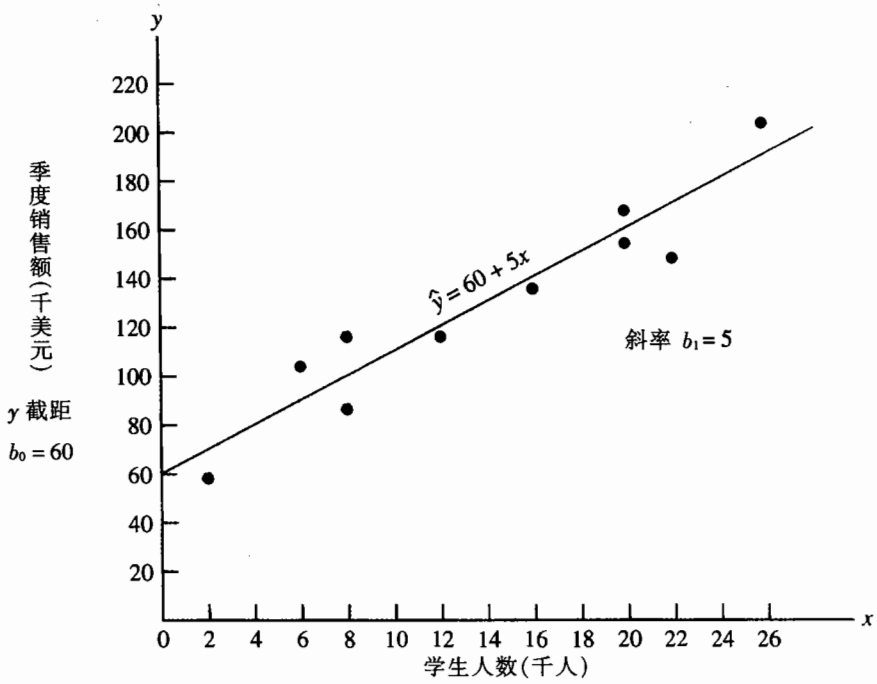


图 14.4　Armand 比萨饼餐馆估计回归方程图：$\hat{y} = 60 + 5x$

估计回归方程的斜率($b_1=5$)为正，表示学生人数增加，销售额增加，实际上，我们得出学生人数每增加 1 000 人，销售额将增加 5 000 美元的结论，也就是说，每增加一个学生，销售额将增加 5 美元。

如果我们确定最小二乘法估计回归方程描述了 x 和 y 之间的关系，我们可以在给定的 x 值下，利用回归方程去预测 y 值。例如，如果我们想预测一家位于有 16 000 名学生的校园附近的餐馆的销售额，我们可以这样计算：

$$\hat{y}=60+5(16)=140$$

利用回归方程进行预测应使外推预测值中独立变量的取值在合理范围之内，因为外推时，我们不能确保同一关系式是有效的。

因此，我们将预测这家餐馆的销售额为 140 000 美元。在下面章节中，我们将讨论如何利用回归方程进行估计和预测的合理性的评估方法。

评 注

最小二乘法通过使自变量 y_i 的观察值和估计值之间的离差平方和达到最小的方法，得到一个回归方程的估计。最小二乘法是选择能与样本数据有最好拟合的方程。如果利用其他准则，例如最小化 y_i 和 $\hat{y}_i$ 之间的绝对离差之和，就会得到不同的方程。事实上，最小二乘法是应用最广泛的方法。如果要预测超出自变量范围的值，利用回归方程的估计就要谨慎，因为我们不能保证变量之间存在此关系。

练 习

方 法

1. 自测题

已知两变量 x 和 y 的 5 组观察值如下：

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. 画出这些数据的散点图。

b. (a)中的散点图能表明这两个变量之间存在什么关系？

c. 试着画一条直线来近似表示 x 和 y 之间的关系。

d. 利用公式(14.6)和(14.7)，来计算 b_0 和 b_1 的值，求出估计回归方程。

e. 对于给定的 $x=4$，利用估计回归方程，预测 y 的值。

2. 对于两变量 x 和 y 的 5 组观察值如下：

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. 画出这些数据的散点图。

b. 根据(a)中的散点图来表明这二变量之间存在什么关系。

c. 试着画一条直线来近似表示 x 和 y 之间的关系。

d. 利用公式(14.6)和(14.7)，来计算 b_0 和 b_1 的值，求出估计回归方程。

e. 当 $x=6$ 时，利用估计回归方程，预测 y 的值。

3. 从两个变量中搜集 5 组观察值进行回归研究：

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. 画出这些数据的散点图。

b. 求出这些数据的回归方程。

c. 对于给定 $x=4$，利用估计回归方程，预测 y 的值。

应　用

4. 以下是搜集到的有关女子游泳运动员的身高(英寸)和体重(磅)的数据：



身高	68	64	62	65	66
体重	132	108	102	115	128

a. 以身高为自变量，画出这些数据的散点图。

b. 根据(a)中的散点图来表明这二变量之间存在什么关系。

c. 试着画一条直线来近似表示 x 和 y 之间的关系。

d. 利用公式(14.6)和(14.7)，来计算 b_0 和 b_1 的值，求出估计回归方程。

e. 如果运动员的身高为 63 英寸，则估计她的体重为多少？

5. 以下数据是 7 种牌子的软饮料的销售数量(百万箱)和广告费用(百万美元)(*Superbrands'98*, October 20, 1997)：

品牌	广告费用(百万美元)	销售数量(百万箱)
Coca-Cola Classic	131.3	1 929.2
Pepsi-Cola	92.4	1 384.6
Diet Coke	60.4	811.4
Sprite	55.7	541.5
Dr. Pepper	40.2	536.9
Mountain Dew	29.0	535.6
7-Up	11.6	219.5

a. 以广告费为自变量，画出这些数据的散点图。

b. 根据(a)中的散点图来表明这二变量之间存在什么关系。

c. 试着画一条直线来近似表示 x 和 y 之间的关系。

d. 利用最小二乘法，求出估计回归方程。

e. 对回归方程的斜率作出说明。

f. 如果一种品牌的软饮料的广告费用是7 000万美元，预测它的销售数量。

6. 美国各航空公司的业绩统计数据公布在《华尔街日报1998年鉴》(*The Wall Street Journal Almanac*，1998)上。有关航班正点到达率和每100 000名乘客中投诉的次数如下：

航空公司	航班正点到达率	投诉率
Southwest	81.8	0.21
Continental	76.6	0.58
Northwest	76.6	0.85
US Airways	75.7	0.68
United	73.8	0.74
American	72.2	0.93
Delta	71.2	0.72
America West	70.8	1.22
TWA	68.5	1.25

a. 以正点率为自变量，画出这些数据的散点图。

b. 根据(a)中的散点图来表明这二变量之间存在什么关系。

c. 求出描述每100 000名乘客中投诉的次数与航班正点到达率相关的估计回归方程。

d. 对回归方程的斜率作出说明。

e. 如果飞机正点到达率为80%，估计每100 000名乘客中投诉次数为多少？

7. 道琼斯指数(DJIA)和标准普尔500指数(S&P)是股票价格水平的测量手段。DJIA是基于30家大公司的股票价格得出的。标准普尔500是由500家股票组成的股票指数。有人认为标准普尔500所含股票数目多，因此是反映股票市场的最好方法。从2000年2月11日，由DJIA和标准普尔500指数10周的收盘价组成一个样本(*Barron's*, April 17, 2000)。



CD光盘数据
DowS&P

时间	道琼斯指数	标准普尔指数
2月11日	10 425	1 387
2月18日	10 220	1 346
2月25日	9 862	1 333
3月3日	10 367	1 409
3月10日	9 929	1 395
3月17日	10 595	1 464
3月24日	11 113	1 527
3月31日	10 922	1 499
4月7日	11 111	1 516
4月14日	10 306	1 357

a. 以道琼斯指数为自变量，画出这些数据的散点图。

b. 用最小二乘法建立估计回归方程。

c. 假定道琼斯指数的收盘价为 11 000，估计标准普尔 500 指数的收盘价。

8. 尼尔森媒体调查机构搜集在 ABC、CBS、NBC、Fox、UPN 和 WB 电视网的黄金时间各种广告播放的次数的数据(*USA Today*，May 5, 1997)。

广告品牌	广告播放次数	收看广告的家庭户数
Wendy's	28	191.7
Ford Escort	20	174.6
Austin Powers movie	14	161.3
Nissan	16	161.1
Pizza Hut	16	147.7
Saturn	16	146.3
Father's Day movie	11	138.2

a. 求出描述广告播放次数与收看广告的家庭户数关系的估计回归方程。

b. 对回归方程的斜率作出说明。

c. 如果播放广告次数为 15 次，估计收看广告的家庭户数为多少？

9. 一个销售经理收集关于销售员的年薪和工龄的数据如下：

销售员	工龄	年薪(千美元)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. 以工龄为自变量，画出这些数据的散点图。

b. 在给定的工龄下，求出预测年销售额的估计回归方程。

c. 利用估计的回归方程，预测有 9 年工龄的一位销售员的年销售额。

10. 《计算机世界》(*PC World*)对前 15 名的笔记本电脑作出评级(*PC World*，February 2000)。通过与标准机器运行同一商业应用程序的速度进行比较，可以对其性能分数作出评判。例如，性能打分为 200 是标准机器运行速度的 2 倍。在这次研究中每种笔记本总的评定分数为 100 分。90 分为优秀，而 70 分表示一般。15 种笔记本的性能分数



CD 光盘数据
PCs

的总的评定级别如下：

	性能打分	总的评定级别
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366 XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavilion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron Trans Port NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

a. 以性能分数为自变量，画出这些数据的散点图。

b. 用最小二乘法建立估计回归方程。

c. 如果一种新的 PC 机，它的性能打分为 225，估计其总的评定级别。

11. 下面是拉斯韦加斯 10 家博彩旅店的旅店收入和博彩业收入(百万美元)的统计数据(*Cornell Hotel and Restaurant Administration Quarterly*, October 1997)：



CD 光盘数据
Casino

公司	旅店收入(百万美元)	博彩业收入(百万美元)
Boyd Gaming	303.5	548.2
Circus Circus Enterprises	664.8	664.8
Grand Casinos	121.0	270.7
Hilton Corp. Gaming Div.	429.6	511.0
MGM Grand, Inc.	373.1	404.7
Mirage Resorts	670.9	782.8
Primadonna Resorts	66.4	130.7
Rio Hotel & Casino	105.8	105.5
Sahara Gaming	102.4	148.7
Station Casinos	135.8	358.5

a. 以旅店收入为自变量，画出这些数据的散点图。

b. 这两个变量之间是否存在线性关系？

c. 求出描述旅店收入和博彩业收入关系的估计回归方程。

d. 假定旅店收入为 5 亿美元，博彩业收入为多少？

12. 下面数据给出 20 家公司的雇员数量和收入(百万美元)的数据(*Fortune*, April 17, 2000)：


CD 光盘数据
EmpRev

	雇员	收入（百万美元）
Sprint	77 600	19 930
Chase Manhattan	74 801	33 710
Computer Sciences	50 000	7 660
Wells Fargo	89 355	21 795
Sunbeam	12 200	2 398
CBS	29 000	7 510
Time Warner	69 722	27 333
Steelcase	16 200	2 743
Georgia-Pacific	57 000	17 796
Toro	1 275	4 673
American Financial	9 400	3 334
Fluor	53 561	12 417
Phillips Petroleum	15 900	13 852
Cardinal Health	36 000	25 034
Borders Group	23 500	2 999
MCI Worldcom	77 000	37 120
Consolidated Edison	14 269	7 491
IBP	45 000	14 075
Super Value	50 000	17 421
H&R Block	4 200	1 669

a. 以雇员数量为自变量，画出这些数据的散点图。

b. 根据(a)中的散点图，来说明雇员的数量与收入之间的关系。

c. 通过这些数据得出估计回归方程。

d. 假定这家公司有 75 000 名员工，用回归方程估计来预测公司的收入。

13. 联邦国税局认为，总逐项列明的扣除减项依赖于纳税人调整后的总收入。对于大的减税，如慈善事业和医疗费用的大的扣除减项是比较合理的。如果一个纳税人自己申报的纳税的每项扣除超过已知收入水平的平均逐项列明的扣除减项，联邦国税局就需要对其进行审计。有关调整后的总收入和平均或合理的总逐项减税的数据如下(单位：千美元)：

调整后的总收入(千美元)	总逐项减税(千美元)
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

a. 以调整后的总收入为自变量，根据这些数据画出散点图。

b. 利用最小二乘法得出估计回归方程。

c. 假定一个纳税人调整后的总收入为 52 500 美元，估计其总逐项减税的合理水平。如果纳税人申称其总逐项减税为 20 400 美元，IRS 认为是否需要对其审计？试解释。

14. 下列数据是关于美国最大的旅馆客户居住率和房价的数据(*The Wall Street Journal Almanac, 1998*)：


CD 光盘数据
Hotel

旅馆	居住率(%)	房价(美元)
Los Angeles – Long Beach	67.9	75.91
Chicago	72.0	92.04
Washington	68.4	94.42
Atlanta	67.7	81.69
Dallas	69.5	74.76
San Diego	68.7	80.86
Anaheim – Santa Ana	69.5	70.04
San Francisco	78.7	106.47
Houston	62.0	66.11
Miami – Hialeah	71.2	85.83
Oahu Island	80.7	107.11
Phoenix	71.4	95.34
Boston	73.5	105.51
Tampa – St. Petersburg	63.4	67.45
Detroit	68.7	64.79
Philadelphia	70.1	83.56
Nashville	67.1	70.12
Seattle	73.4	82.60
Minneapolis – St. Paul	69.8	73.64
New Orleans	70.6	99.00

a. 以平均房价为自变量，画出这些数据的散点图。

b. 求描述客户居住率和平均房价关系的估计回归方程。

c. 假定平均房价为 80 美元，估计旅馆的居住率。

14.3 判定系数

对于 Armand 比萨饼连锁店例子，我们建立估计回归方程 $\hat{y}=60+5x$ 来近似表示学生人数 x 和季度销售额 y 的线性关系。现在有一个问题：我们如何使估计回归方程与样本数据更好地拟合？在本节中，我们介绍如何用**判定系数**(coefficient of determination)来判断估计回归方程的拟合程度。

对于第 i 个观察值，因变量观察值 y_i 与估计值 $\hat{y}_i$ 之差为**第 i 个残差**(ith residual)。第 i 个残差表示 $\hat{y}_i$ 估计 y_i 的误差。因而，对于第 i 个观察值，残差为 $y_i-\hat{y}_i$。这些残差或误差的平方和是利用最小二乘法计算的最小量。这个量也被称为**误差平方和**，记作 SSE。

误差平方和

$$SSE = \sum(y_i - \hat{y}_i)^2 \qquad (14.8)$$

SSE 值是用估计回归方程来估计样本中因变量而产生的误差。

在表 14.3 中我们列出了 Armand 比萨饼连锁店误差平方和计算的必要步骤。例如，对于旅店 1，自变量和因变量值为 $x_1 = 2$，$y_1 = 58$。利用估计回归方程，我们得出旅店 1 的估计的季度销售额为$\hat{y}_1 = 60 + 5(2) = 70$。因此，对于旅店 1，用$\hat{y}_1$ 估计 y_1 的误差为 $y_1 - \hat{y}_1 = 58 - 70 = -12$。表 14.3 的最后一列误差平方，$(-12)^2 = 144$。计算样本中的每一家饭店的残差并将其残差平方后，将其平方和汇总得到 SSE = 1 530。因此，SSE = 1 530 度量了利用估计回归方程$\hat{y} = 60 + 5x$ 预测销售收入所产生的误差。

表 14.3　Armand 比萨饼连锁店的 SSE 计算步骤

旅馆 (i)	x_i = 学生人数 (千人)	y_i = 季度销售额 (千美元)	$\hat{y}_i = 60 + 5x_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	58	70	-12	144
2	6	105	90	15	225
3	8	88	100	-12	144
4	8	118	100	18	324
5	12	117	120	-3	9
6	16	137	140	-3	9
7	20	157	160	-3	9
8	20	169	160	9	81
9	22	149	170	-21	441
10	26	202	190	12	144
					SSE = 1 530

现在我们假定在不知道学生人数的情况下，求出季度销售额的估计值。因为没有相关变量，我们用样本均值作为任何一家旅店的估计的季度销售额。表 14.2 中得到样本销售数据，$\sum y_i = 1\ 300$。因此，10 家 Armand 比萨饼连锁店的样本平均季度销售额为$\bar{y} = \sum y_i / n = 1\ 300/10 = 130$。表 14.4 表示利用样本平均值$\bar{y} = 130$ 来估计样本中每个旅店的季度销售额。对于样本中的第 i 个饭店，离差 $y_i - \bar{y}$得出用$\bar{y}$ 估计销售额的误差大小。其相应的平方和，称之为**总的平方和**，记为 SST。

总的平方和

$$SST = \sum(y_i - \bar{y})^2 \qquad (14.9)$$

在表 14.4 中最后一列底部是 Armand 比萨饼连锁店的总的平方和，即为 SST = 15 730。

在图 14.5 中我们画出$\hat{y} = 60 + 5x$ 回归方程直线和 $\bar{y} = 130$ 直线。注意到样本点散布在估计回

归线比在$\bar{y}=130$附近更密集。例如，对于样本中第10家餐馆，我们发现当用$\bar{y}=130$比用$\hat{y}_{10}=60+5(26)=190$估计$\hat{y}$时误差更大。因此，我们将SST看做是观测值在$\bar{y}=130$周围密集程度的度量，而将SSE看做是观测值在$\bar{y}$直线周围密集程度的度量。

表 14.4　Armand 比萨饼连锁店总的平方和的计算

旅馆 (i)	x_i = 学生人数 (千人)	y_i = 季度销售额 (千美元)	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
1	2	58	−72	5 184
2	6	105	−25	625
3	8	88	−42	1 764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1 521
9	22	149	19	361
10	26	202	72	5 184
				SST = 15 730

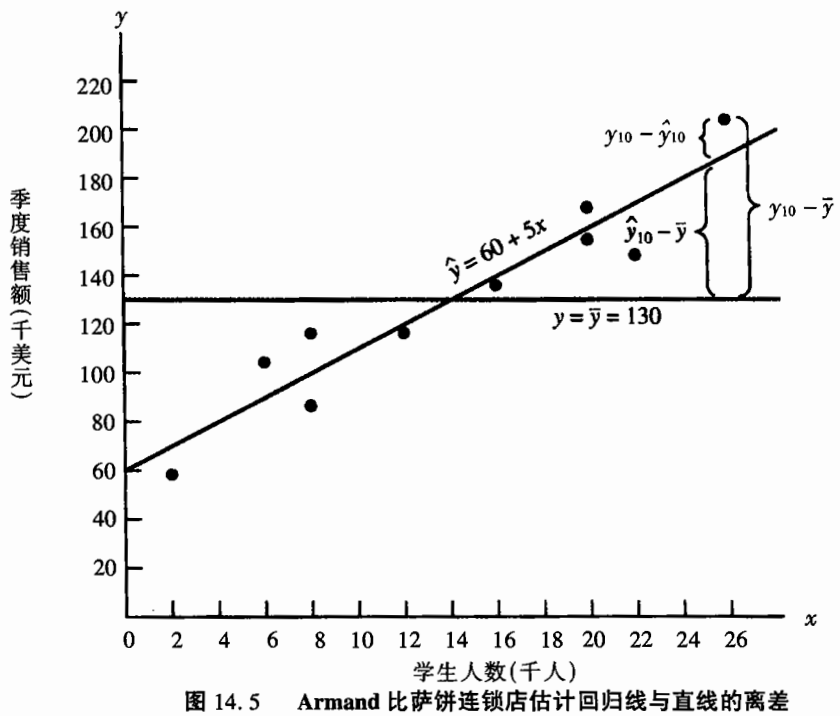


图 14.5　Armand 比萨饼连锁店估计回归线与直线的离差

为了判断估计回归线的$\hat{y}$与$\bar{y}$之间的离差的大小，需要计算另一个平方和，称该平方和为**回归平方和**，记为SSR。

回归平方和

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \tag{14.10}$$

从上述讨论中，我们可以发现 SST、SSR 和 SSE 三者之间的关系。实际上，这三个平方和之间的关系是统计中最重要的结论之一。

SST、SSR 和 SSE 之间的关系

$$SST = SSR + SSE \tag{14.11}$$

式中　SST——总的平方和；
　　SSR——回归平方和；
　　SSE——误差平方和。

公式(14.11)介绍了总的平方和能被分解为两个组成部分，回归平方和和误差平方和。因此，如果已知任意其中两个，就可求出第三个。例如，在 Armand 比萨饼连锁店中，我们已知 SSE = 1 530 和 SST = 15 730。所以，利用公式(14.11)可求得 SSR，我们得到回归总的平方和为

$$SSR = SST - SSE = 15\ 730 - 1\ 530 = 14\ 200$$

现在我们介绍如何利用这三个平方和 SST、SSR 和 SSE 来判断估计回归方程的拟合程度。如果因变量 y_i 值都恰巧落在估计回归线上，估计回归方程将给出一个很好的拟合。此时，$y_i - \hat{y}_i$ 为 0，因此 SSE = 0。因为 SST = SSR + SSE，故完全拟合时，SSR 一定等于 SST，比率（SSR/SST）一定为 1。拟合程度差，将会导致 SSE 变大，利用公式(14.11)，得 SSE = SST - SSR。因此，当 SSR = 0 时，SSE = SST，SSE 最大(拟合度最差)。

比率 SSR/SST 在 0—1 之间，我们可以利用此比值来评价估计回归方程的拟合程度。该比率称为**判定系数**(coefficient of determination)，记为 r^2。

判定系数

$$r^2 = SSR/SST \tag{14.12}$$

对 Armand 比萨饼连锁店的例子来说，判定系数的值是 $r^2 = \dfrac{SSR}{SST} = \dfrac{14\ 200}{15\ 730} = 0.902\ 7$

如果用一个百分比来表示判定系数，r^2 被解释为用估计回归方程来解释总的平方和的百分比。对于 Armand 比萨饼连锁店，用估计回归方程 $\hat{y} = 60 + 5x$ 来预测销售额时，我们得出有 90.27% 的总的平方和能被估计回归方程解释。换句话说，销售收入的变异性有 90.27% 能被学生人数和销售额之间的线性关系所解释。对此估计回归方程，我们得到有很好的拟合这一结论。

相关系数

在第 3 章中我们引入**相关系数**(correlation coefficient)来描述两个变量 x 和 y 之间的线性程度。相关系数在 -1—+1 之间。+1 表示两个变量 x 和 y 是完全线性正相关；-1 表示 x 和 y 是完全线

性负相关，所有在直线上的数据点都有一个负斜率。相关系数为零表示 x 和 y 不是线性相关。

在 3.5 节中我们给出计算样本相关系数方程。如果已经进行了回归分析并对相关系数进行计算，则样本相关系数 r^2 计算如下：

样本相关系数

$$r_{xy} = (b_1 \text{ 的符号}) \sqrt{\text{判定系数}}$$
$$= (b_1 \text{ 的符号}) \sqrt{r^2} \qquad (14.13)$$

式中，b_1 为估计回归方程 $\hat{y} = b_0 + b_1 x$ 的斜率。

如果估计回归方程中有一个正的斜率（$b_1 > 0$），则样本相关系数为正，反之，如果估计回归方程有一个负的斜率($b_1 < 0$)，则样本相关系数为负。

对于 Armand 比萨饼连锁店例子，估计回归方程 $\hat{y} = 60 + 5x$ 的判定系数为 0.902 7。因为估计回归方程的斜率为正，公式(14.13)表示样本相关系数为 $+\sqrt{0.902\ 7} = +0.950\ 1$。对于样本相关系数 $r_{xy} = +0.950\ 1$，得出 x 和 y 之间是强的正的线性关系的结论。

对于两个变量的线性关系，判定系数和样本相关系数都是其度量手段，判定系数值在 0—1 之间，而样本相关系数值在 −1—+1 之间。虽然样本相关系数是度量两个变量的线性关系的，但相关系数可用于度量非线性关系和两个或两个以上因变量的关系。从这个意义上，相关系数有很广阔的使用范围。

评 注

1. 在用最小二乘法估计回归方程和计算相关系数时，对 x 和 y 之间的显著性关系没有进行概率假设和统计检验。r^2 值越大意味着最小二乘法与数据拟合程度越好，也就是说，观察值越靠近最小二乘线。但是如果只有 r^2，我们不能从统计上得出 x 和 y 之间的关系是显著的结论。这样的结论必须是在考虑了样本容量和最小二乘估计的适当样本分布后得出的。
2. 在实际总的应用中，例如在社会科学中遇到的典型数据，经常要考虑 r^2 的数值低于 0.25 的情形。对于在自然科学和生命科学中的数据，经常遇到 r^2 的数值大于或等于 0.60 的情形。事实上，在某些实例中，有时还会遇到 r^2 的值大于 0.90 的情形。在商业应用中，r^2 值差异很大，主要是看每个应用的各自特性。

练 习

方 法

15. 自测题 练习 1 的数据如下：

x_i	1	2	3	4	5
y_i	3	7	5	11	14

这些数据的估计回归方程为 $\hat{y} = 0.20 + 2.60x$。

a. 利用公式(14.8)、(14.9)和(14.10)，计算 SSE、SST 和 SSR。

b. 计算判定系数 r^2，试评述其拟合程度。

c. 计算样本相关系数。

16. 练习 2 的数据如下：

x_i	2	3	5	1	8
y_i	25	25	20	30	16

这些数据的估计回归方程为 $\hat{y} = 30.33 - 1.88x$

a. 计算 SSE、SST 和 SSR。

b. 计算判定系数 r^2，试评述其拟合程度。

c. 计算样本相关系数。

17. 练习 3 的数据如下：

x_i	2	4	5	7	8
y_i	2	3	2	6	4

这些数据的回归方程为 $\hat{y} = 0.75 + 0.51x$。总的平方和中有多大的比例由估计回归方程解释?样本相关系数值为多少?

应　用

18. 

下面是关于商业管理专业信息系统方向的学士学位学生平均成绩 x 和毕业后的月薪 y 的数据。这些数据的估计回归方程为 $\hat{y} = 1\,790.5 + 581.1x$。

毕业平均成绩	月薪(美元)
2.6	3 300
3.4	3 600
3.6	4 000
3.2	3 500
3.5	3 900
2.9	3 600

a. 计算 SST、SSR 和 SSE。

b. 计算判定系数 r^2，试评述其拟合程度。

c. 计算样本相关系数。

19. 练习 7 的数据如下：



	x = DJIA	y = S&P
2 月 11 日	10 425	1 387
2 月 18 日	10 220	1 346

（续表）

	x = DJIA	y = S&P
2月25日	9 862	1 333
3月3日	10 367	1 409
3月10日	9 929	1 395
3月17日	10 595	1 464
3月24日	11 113	1 527
3月31日	10 922	1 499
4月7日	11 111	1 516
4月14日	10 306	1 357

这些数据的估计回归方程为$\hat{y} = -137.63 + 0.1489x$。总的平方和中有多大比例被估计回归方程所解释？试评述其拟合程度，样本相关系数为多大？

20. 由18个城市的居民家庭收入和房屋价格的数据（*Places Rated Almanac*，2000）如下：（数据以千美元为单位）



CD光盘数据
Cities

	收入	房屋价格
Akron, OH	74.1	114.9
Atlanta, GA	82.4	126.9
Birmingham, AL	71.2	130.9
Bismarck, ND	62.8	92.8
Cleveland, OH	79.2	135.8
Columbia, SC	66.8	116.7
Denver, CO	82.6	161.9
Detroit, MI	85.3	145.0
Fort Lauderdale, FL	75.8	145.3
Hartford, CT	89.1	162.1
Lancaster, PA	75.2	125.9
Madison, WI	78.8	145.2
Naples, FL	100.0	173.6
Nashville, TN	77.3	125.9
Philadelphia, PA	87.0	151.5
Savannah, GA	67.8	108.1
Toledo, OH	71.2	101.1
Washington, DC	97.4	191.9

a. 在给定的家庭收入下，利用这些数据建立估计一个典型的城市居民的房屋价格的回归方程。

b. 计算r^2。用此估计回归方程估计一个城市家庭的房屋价格是否觉得不妥？

c. 如果一个典型家庭收入为95 000美元，估计其房屋价格。

21. 回归分析的一个重要应用是成本估计。收集产品的数量和成本的数据，利用最小二乘法来建立关于产品产量和成本之间的估计回归方程，会计人员可以据此来进行某一个特定的生产操作中的成本估计。下面是某一生产操作中产品数量和总成本的数据：

产品数量（单位）	总成本(美元)
400	4 000
450	5 000
550	5 400
600	5 900
700	6 400
750	7 000

a. 在给定的产品数量下，利用这些数据建立估计的总成本的回归方程。

b. 每单位产品的可变成本或追加成本?

c. 计算判定系数。在总成本中有多大的比例能被产品数量解释?

d. 公司计划下个月生产 500 单位产品，估计其总成本?

22. 公司的董事长和执行总裁的年薪是否与公司的业绩有关？下表列出了两年时间的资产回报的变动率和两年后总经理和主要高级职员的年薪的变动率的数据（*Business Week*，April 21, 1997）：

公司	两年时间内资产的回报变动率(%)	执行总裁的薪水变动(%)
Dow Chemical	201.3	18
Rohm & Haas	146.5	28
Morton International	76.7	10
Union Carbide	158.2	28
Praxair	-34.9	15
Air Products & Chemicals	73.2	-9
Eastman Chemical	-7.9	-20

a. 以两年时间内资产的回报变动为自变量，建立估计回归方程。

b. 计算 r^2。利用两年资产的回报变动来预测总经理和主要高级职员的年薪的变动是否不妥？试讨论。

c. 样本相关系数多大？资产的回报变动与总经理和主要高级职员的薪水变动的线性关系是强还是弱?

14.4　模型假设

在进行回归分析时，首先我们对模型中自变量和因变量的关系做出适当的假定。对于简单线性回归，假定的回归模型为

$$y = \beta_0 + \beta_1 x + \varepsilon$$

接着，用最小二乘法分别计算 b_0 和 b_1 的值，即模型参数 β_0 和 β_1 的估计值。估计回归方程为

$$\hat{y} = b_0 + b_1 x$$

我们知道判定系数(r^2)是判断估计回归方程的拟合程度。然而，r^2 越大，那么在没进行适当的模型假设分析之前，就越不能随便使用估计回归方程。判断模型假设是否适当的一个重要步骤是检验关系是否显著。回归方程的显著性检验是基于误差项 ε 的假设。

回归模型误差项的假设

$$y = \beta_0 + \beta_1 x + \varepsilon$$

1. 误差项 ε 是一个平均值或期望值为零的随机变量，也就是说，$E(\varepsilon) = 0$。

   这表示：β_0 和 β_1 是常量，所以 $E(\beta_0) = \beta_0$，$E(\beta_1) = \beta_1$；因而，对于一确定的 x 值，y 的期望值为

   $$E(y) = \beta_0 + \beta_1 x \qquad \textbf{(14.14)}$$

   正如前面所写，公式(14.14)是指回归方程。
2. 对于所有的 x，误差项 ε 的方差一样，记为 σ^2。

   这表示：对于所有的 x，y 的方差也为 σ^2。
3. ε 值是独立的。

   这表示：对于某一 x 相对应的 ε 与其他 x 对应的 ε 值不相关，因而，对于某一 x 相对应的 y 与其他 x 对应的 y 值没有相关性。
4. 误差项 ε 是正态分布的随机变量。

   这表示：因为 y 是 ε 的线性函数，y 也是一个正态分布的随机变量。

图 14.6 是对模型假设及其涵义的一个说明。从这个图上，发现某一特定的 x 值变化，$E(y)$

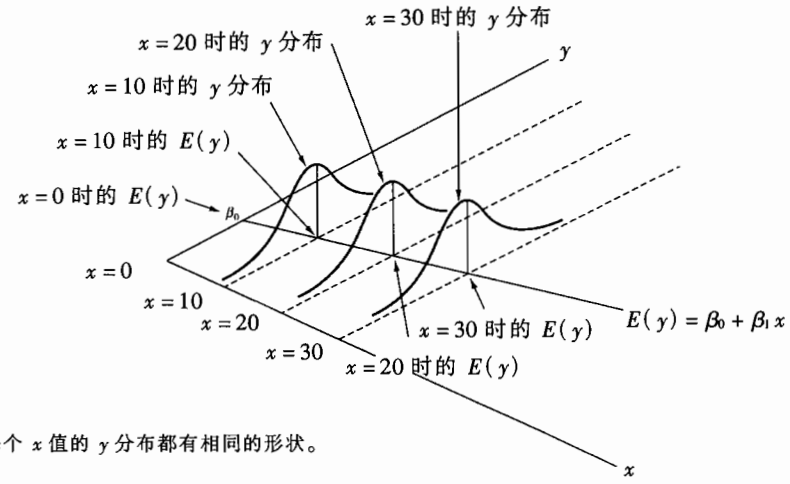


图 14.6　回归模型假设

值也随之变化。但是，不管 x 为多大，ε 的概率分布和 y 的概率分布都是正态分布，并且具有相同的方差。在任意特定点上，误差项 ε 某一特定值依赖于 y 的真实值大于或小于 $E(y)$。

此时我们要记住，要对变量 x 和 y 之间关系的形式作一个假定或假设。也就是说，假定直线 $\beta_0 + \beta_1 x$ 是变量关系的基础，我们不能忽略其他模型，例如，$y = \beta_0 + \beta_1 x^2 + \varepsilon$ 或许是描述变量之间关系的最佳模型。

14.5　显著性检验

在简单线性回归方程中，y 的平均值或期望值是 x 的线性函数：$E(y) = \beta_0 + \beta_1 x$。如果 β_1 值为零，$E(y) = \beta_0 + (0)x = \beta_0$。此时，$y$ 的均值不依赖于 x，因此得出 x 和 y 不是线性相关。另外，如果 β_1 不为零，则得出这两个变量相关，因此，检验回归关系是否显著时，我们可利用假设检验来判断 β_1 值是否为零。通常有两种检验方法，两者在回归模型中都使用了 σ^2 的估计和 ε 的方差。

σ^2 的估计

从回归模型和其假设中我们得出 σ^2(ε 的方差)代表回归线的 y 值的方差。将估计回归线中 y 的离差称之为残差。因此，残差平方和(SSE)是估计回归线的真实观察值变异的度量。**均方误差**（mean square error, MSE）是 σ^2 的一个估计，是 SSE 除以其自由度后得出的。

用 $\hat{y}_i = b_0 + b_i x_i$ 代入 SSE，则公式写为

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

每个平方和都有其相联系的自由度。统计学家已经证明 SSE 有 $n-2$ 个自由度，因为有两个参数(β_0 和 β_1)被用于进行 SSE 估计。故用 SSE 除以 $n-2$ 得到均方误差，MSE 是 σ^2 的一个无偏估计量。因为 MSE 是 σ^2 的估计，记为 s^2。

均方误差（σ^2 的估计量）

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \tag{14.15}$$

在 14.3 节的 Armand 比萨饼连锁店的例子中，得到 SSE = 1 530，因此，σ^2 的一个无偏估计量为

$$s^2 = \text{MSE} = \frac{1\,530}{8} = 191.25$$

为了估计 σ，对 s^2 开平方根，结果 s 为**估计量的标准误差**(standard error of the estimate)。

估计量的标准误差

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \tag{14.16}$$

在 Armand 比萨饼连锁店的例子中，$s = \sqrt{\text{MSE}} = \sqrt{191.25} = 13.829$。在下面的讨论中，我们用估计量的标准误差来对 x 和 y 之间的关系进行显著性检验。

t 检验

简单线性回归方程 $y = \beta_0 + \beta_1 x + \varepsilon$。如果 x 和 y 是线性相关，得出 $\beta_1 \neq 0$，检验的目的是检验是否能得出 $\beta_1 \neq 0$。我们用样本数据来对参数 β_1 进行假设检验：

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

如果拒绝 H_0，则得出 $\beta_1 \neq 0$，两个变量之间在统计上存在显著的关系。然而，如果 H_0 未被拒绝，我们没有充分理由得出存在显著关系的结论。b_1(β_1 的最小方差估计值)提供了假设检验的基础。

首先，让我们考察对同一个回归研究，进行不同的随机抽样将会出现的情况。例如，假设由 Armand 比萨饼连锁店的 10 家不同餐店组成的样本。对这个新的样本进行样本回归分析，得出的估计回归方程近似于先前的估计回归方程 $\hat{y} = 60 + 5x$，但是不可能得到相同的方程(截距恰好为 60，斜率恰好为 5)。实际上，最小二乘法估计量 b_0 和 b_1 是样本统计量，它们有各自的样本分布，估计量 b_1 的样本分布的性质如下：

b_1 的样本分布

期望值

$$E(b_1) = \beta_1$$

标准差

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{14.17}$$

正态分布形式

注意 b_1 的期望值等于 β_1，所以 b_1 是 β_1 的无偏估计量。

由于 σ 未知，我们类似于用 s 估计 σ 的方法，按公式(14.17)用 s_{b_1} 来定义 σ_{b_1} 的估计值，那么，我们就可以得到下列 σ_{b_1} 的估计值：

b_1 的估计标准差

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{14.18}$$

b_1 的标准离差也是 b_1 的标准差，因此，可以用 s_{b_1} 来作为 b_1 的标准差的估计量。

对于 Armand 比萨饼连锁店，$s=13.829$。见表 14.2，利用公式 $\sum(x_i-\bar{x})^2=568$，得到 b_1 的估计标准差为

$$s_{b_1}=\frac{13.829}{\sqrt{568}}=0.5803$$

利用检验进行显著性检验是因为统计检验量

$$\frac{b_1-\beta_1}{s_{b_1}}$$

是服从自由度为 $n-2$ 的 t 分布的。如果原假设为真，则 $\beta_1=0$ 并且 $t=\frac{b_1}{s_{b_1}}$。

当显著性水平 $\alpha=0.01$ 时，对于 Armand 比萨饼连锁店进行显著性检验，检验统计量为

$$t=\frac{b_1}{s_{b_1}}=\frac{5}{0.5803}=8.62$$

从附录 B 的表 2 中得到 $\alpha=0.01$，自由度为 $n-2=10-2=8$ 的分布双侧检验 t 值，即 $t_{0.005}=3.355$。因为 $8.62>3.355$，所以我们拒绝 H_0，并且认为在显著性水平 $\alpha=0.01$ 下，$\beta_1\neq0$。统计结果表明对于学生人数和销售额之间存在显著关系。

同样可用 p-值准则进行显著性检验。通常拒绝规则为：如果 p-值$<\alpha$，则拒绝 H_0。但是，因为从 t 概率分布表中直接判断得出 p-值比较困难，因此可以使用软件包 Minitab 或 Excel，附录 14.1 和 14.2 给出使用这两种方法的步骤，14.7 节中将介绍如何使用计算机进行回归分析。对于 Armand 比萨饼连锁店，检验统计量 $t=8.62$ 相对应的 p-值为 0.000(三位小数)。因为 p-值 $=0.000<\alpha=0.05$，所以拒绝 H_0，因此得出学生人数与销售额之间没有显著关系。利用检验进行显著性检验步骤如下：

简单线性回归显著性的 t 检验

$$H_0:\ \beta_1=0$$
$$H_a:\ \beta_1\neq0$$

检验统计量

$$t=\frac{b_1}{s_{b_1}} \tag{14.19}$$

拒绝规则

利用检验统计量：如果 $t<-t_{\alpha/2}$ 或 $t>t_{\alpha/2}$，则拒绝 H_0

利用 p-值：　如果 p-值$<\alpha$，则拒绝 H_0

$t_{\alpha/2}$ 是基于自由度 $n-2$ 时的 t 分布。

β_1 的置信区间

β_1 的置信区间形式如下：$b_1 \pm t_{\alpha/2} s_{b_1}$。点估计是 b_1，边际误差是 $t_{\alpha/2} s_{b_1}$。与这一区间相应的置信系数是 $1-\alpha$，$t_{\alpha/2}$ 是具有 $n-2$ 个自由度的 t 分布单侧面积为 $\alpha/2$ 时的 t 值。例如，我们要得出 Armand 比萨饼连锁店 β_1 的置信区间估计，最低水平应提高到 99%。从附录 B 的表 2 中我们可查到 $\alpha=0.01$、$n-2=10-2=8$ 自由度的 t 值是：$t_{0.005}=3.355$。那么，β_1 的 99% 的置信区间估计是

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

或 3.05 到 6.95。

在利用显著性 t 检验时，假设检验为

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

在显著性水平 $\alpha=0.01$ 下，我们用置信度为 99% 的区间来作为 Armand 比萨饼连锁店假设检验的结论。因为为零，假设检验值 β_1 不包括在置信区间(3.05，6.95)，我们拒绝 H_0，得出学生人数和销售额之间存在显著关系的结论。一般而言，置信区间可用来进行 β_1 双侧假设检验。如果 β_1 的假设检验值包含在置信区间内，则不能拒绝 H_0。

F 检验

建立在 F 概率分布基础上的 F 检验，同样也能对回归进行显著性检验。如果只有一个变量，F 值将得到与 t 检验相同的结果。也就是说，如果 t 检验表明 $\beta_1 \neq 0$，因此变量之间存在显著关系，那么 F 值也得出显著关系的结论。但是如果超过一个自变量，则只有 F 检验才能对所有的显著性关系进行检验。

利用 F 检验判断回归关系是否显著的基本原理是以 σ^2 的两个独立自变量为依据。我们得出 MSE 是 σ^2 的一个估计值。如果原假设 H_0：$\beta_1=0$ 为真，回归方程 SSR 除以其自由度得出 σ^2 的另一个独立估计量。这个估计量为**来自回归均方**，简单说为**均方回归**，记为 MSR。总体上

$$\text{MSR} = \frac{\text{SSR}}{\text{回归自由度}}$$

在本节中我们所考虑模型中，回归自由度一般等于自变量的个数。因此

$$\text{MSR} = \frac{\text{SSR}}{\text{自变量的个数}} \tag{14.20}$$

因为在本章中，我们只考虑只有一个自变量的回归模型，得 $\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR}$。因此，对于 Armand 比萨饼连锁店，MSR = SSR = 14 200。

如果原假设 H_0：$\beta_1=0$ 为真，则 MSR 和 MSE 是 σ^2 的两个独立的估计量，且 MSR/MSE 的样

本分布服从 F 分布，其分子的自由度为 1，分母自由度为 $n-2$。因此，$\beta_1=0$ 时 MSR/MSE 值为 1。但是，如果原假设为假($\beta_1\neq0$)，MSR 高估了 σ^2，MSR/MSE 值将变得无穷大。因此，MSR/MSE 值越大越有可能导致 H_0 得出 x 和 y 之间的关系在统计上是显著的。

以 Armand 比萨饼连锁店检验。检验统计量为

$$F=\frac{\text{MSR}}{\text{MSE}}=\frac{14\ 200}{191.25}=74.25$$

从附录 B 中的表 4 得到对于 $\alpha=0.01$，分子自由度为 1，分母自由度为 $n-2=10-2=8$ 的 $F_{0.01}=11.26$。因为 $74.25>11.26$，所以拒绝 H_0，得出当显著性水平为 0.01 时，β_1 不等于 0。F 检验提供了足够的统计数据，以便我们得出学生人数和销售额之间有显著关系。

F 检验和 *t* 检验对简单线性回归检验得出的结论一样。

可以用 p-值准则进行 F 检验。通常拒绝规则为：如果 p-值$<\alpha$，则拒绝 H_0。但是，从 F 概率分布表中直接判断得出 p-值比较困难，因此可以使用软件包 Minitab 或 Excel，附录 14.1 和 14.2 给出使用这两种方法的步骤，14.7 节中将介绍如何使用计算机进行回归分析。对于 Armand 比萨饼连锁店，检验统计量 $F=74.25$ 相对应的 p-值为 0.000(三位小数)。因为 p-值 $=0.000<\alpha=0.05$，所以拒绝 H_0，得出学生人数与销售额之间没有显著关系，利用 F 检验进行显著性检验步骤如下：

简单线性回归显著性水平的 *F* 检验

$$H_0:\ \beta_1=0$$
$$H_a:\ \beta_1\neq0$$

检验统计量

$$F=\frac{\text{MSR}}{\text{MSE}} \tag{14.21}$$

拒绝规则

利用检验统计量：如果 $F>F_\alpha$，则拒绝 H_0

利用 p-值：如果 p-值$<\alpha$，则拒绝 H_0

F_α 是基于分子自由度为 1，分母自由度为 $n-2$ 的 F 分布。

如果 H_0 为假，则 MSE 是 σ^2 的一个无偏估计。如果 H_0 为真，MSE 和 MSR 都是 σ^2 的一个无偏估计，此时 MSR/MSE 应该接近 1。

在第 13 章中我们介绍方差分析(ANOVA)以及如何利用**方差分析表**(ANOVA table)对方差分析计算的简单汇总。类似地，ANOVA 表可用于汇总回归显著性的 F 检验结果。表 14.5 就是将 ANOVA 用于简单线性回归的一般形式。表 14.6 是 Armand 比萨饼连锁店 F 检验计算的 ANOVA

表。第 1 列中的回归、残差和总计是方差的三个来源，第 2 列中的 SSR、SSE 和 SST 是对应的三个平方和，第 3 列是这些平方和的自由度，SSR 为 1，SSE 为 $n-2$，SST 为 $n-1$。第 4 列为 MSR 和 MSE 值，第 5 列为 $F=\frac{\text{MSR}}{\text{MSE}}$值。几乎所有回归分析的计算机输出结果都包含在显著性检验的 ANOVA 表中。

表 14.5　简单线形回归 ANOVA 表的一般形式

方差来源	平方和	自由度	均方差	F
回归方程	SSR	1	$\text{MSR}=\frac{\text{SSR}}{1}$	$F=\frac{\text{MSR}}{\text{MSE}}$
误差	SSE	$n-2$	$\text{MSE}=\frac{\text{SSE}}{n-2}$	
总计	SST	$n-1$		

表 14.6　Armand 比萨饼连锁店 F 检验计算的 ANOVA 表

方差来源	平方和	自由度	均方差	F
回归方程	14 200	1	$\frac{14\,200}{1}=14\,200$	$\frac{14\,200}{191.25}=74.25$
误差	1 530	8	$\frac{1\,530}{8}=191.25$	
总计	15 730	9		

在方差分析表中，每个分析的总的平方总和是回归平方和与误差平方和之和。另外，总的自由度为回归自由度和误差自由度之和。

关于显著性检验解释的几点注意

拒绝原假设 $H_0:\beta_1=0$，得出 x 和 y 之间的关系显著，仍不能得出 x 和 y 之间存在一个因果关系的结论。只有经过分析家从事实上得出这两个变量之间确实存在因果关系，才能认为存在这种因果关系。在 Armand 比萨饼连锁店例子中，我们得出学生人数 x 和销售额 y 之间存在显著关系。而且，估计回归方程 $\hat{y}=60+5x$ 给出了这个显著关系的最小二乘估计量。但是，我们仍然不能得出学生人数 x 变化**导致**销售额 y 的变化，因为我们仅仅从统计上得出其显著关系。这种因果关系的结论需要得到理论上的支持，另外需要分析家的准确判断。Armand 比萨饼连锁店管理者认为学生人数的增长是季度销售额增大的最大可能。因此，显著性检验的结论能得出存在这种回归关系。

回归分析只是被用来表明这个变量与另一个变量之间的关系，不能作为因果关系的证据。

另外，因为我们只不过拒绝 $H_0:\beta_1=0$，表明统计上存在显著性关系，但仍不能得出 x 和 y

之间的关系是线性的结论。我们只能得出在 x 的样本观察值范围内，x 和 y 是相关的，而且这线性关系是解释了 y 的变异性的显著部分的结论。图 14.7 说明了这种情形。虽然显著性检验拒绝了原假设 H_0：$\beta_1 = 0$，并且做出 x 和 y 之间存在显著性关系的结论。但是图中表示 x 和 y 之间的真正关系不是线性。虽然在 x 的样本观察值范围内，估计回归方程 $\hat{y} = b_0 + b_1 x$ 给出了一个近似线性关系，但对这个范围之外的 x 却很差。

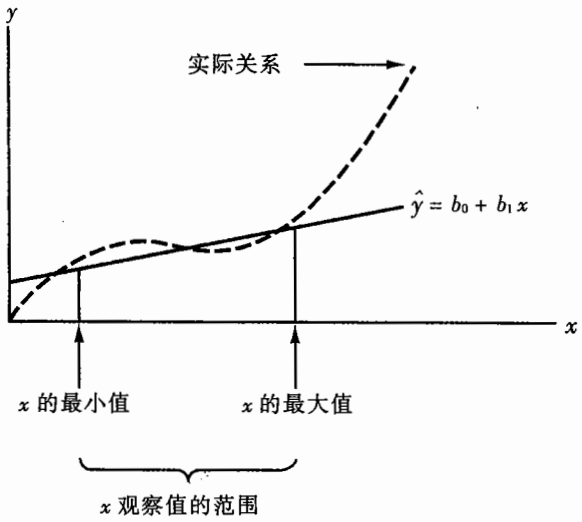


图 14.7　非线性关系的例子

对于已知的一个显著关系，利用估计的回归方程，对于 x 的样本观察值范围内 x 值进行预测，是可靠的。对于 Armand 比萨饼连锁店例子，x 的样本观察值取值范围在 2 到 26 之间。除非有理由表明这个模型超出这个范围仍适用，一般情况下，在自变量 x 的取值范围以外进行预测应十分谨慎。对于 Armand 比萨饼连锁店，因为在显著性水平为 0.01 下，回归关系显著，所以对于有 2 000 名至 26 000 名学生的校园附近的连锁店，可利用显著的回归关系 $\hat{y} = 60 + 5x$ 来预测连锁店的销售收入。

评　注

1. 在 14.4 节对误差项作了一些假定，才可能在本节中进行统计显著性检验。b_1 抽样分布的性质以及 t 检验和 F 检验都是从这些假定中直接得出的。
2. 不能把统计显著性与实际显著性混淆。当样本容量很大时，对于小的 b_1 值，我们仍能得出在统计上是显著的结论。而在这种情况下，要得出实际上是显著关系的结论，要特别谨慎。
3. x 和 y 之间的线性关系进行显著性检验，也可以利用样本相关系数 r_{xy} 进行检验。用 ρ_{xy} 来表示总体相关系数，建立如下假设：

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

如果 H_0 被拒绝，则得出存在一个显著性关系的结论，有关这种检验的细节在附录 14.2 中给出。但是，t 检验和 F 检验与先前利用相关系数进行显著性检验得出的**结论一样**。因此，如果已经进行 t 检验和 F 检验，就不必再利用相关系数进行显著性检验。

练　习

方　法

23. 练习 1 的数据如下：



x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. 利用公式(14.15)计算均方误差。

b. 利用公式(14.16)计算估计标准差。

c. 利用公式(14.18)计算 b_1 的估计标准差。

d. 利用 t 检验检验下面的假设($\alpha=0.05$)：

$$H_0:\ \beta_1=0$$
$$H_a:\ \beta_1\neq 0$$

e. 在显著性水平为 0.05 时，利用 F 检验检验(d)中的假设。请列出方差分析表中的结果。

24. 练习 2 的数据如下：

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. 利用公式(14.15)计算均方误差。

b. 利用公式(14.16)计算估计标准差。

c. 利用公式(14.18)计算 b_1 的估计标准差。

d. 利用 t 检验检验下面的假设($\alpha=0.05$)：

$$H_0:\ \beta_1=0$$
$$H_a:\ \beta_1\neq 0$$

e. 在显著性水平为 0.05 时，利用 F 检验检验(d)中的假设。请列出方差分析表中的结果。

25. 练习 3 的数据如下：

x_i	2	4	5	7	8
y_i	2	3	2	6	4

a. 估计标准误差值。

b. 利用 t 检验对显著性关系进行检验，取 $\alpha=0.05$。

c. 利用 F 检验检验变量的显著性关系，得出你的结论。

应　用

26. 练习 18 中平均分数和每月销售额如下：



总评分	月薪(美元)	总评分	月薪(美元)
2.6	3 300	3.2	3 500
3.4	3 600	3.5	3 900
3.6	4 000	2.9	3 600

a. 利用 t 检验来证明总评分与月薪之间是否有显著性关系，得出你的结论，取 $\alpha=0.05$。

b. 利用 F 检验检验显著性关系，得出你的结论，取 $\alpha=0.05$。

c. 列出 ANOVA 表。

27. 参照练习 9，建立工龄和年薪的关系的估计回归方程。在显著性水平为 0.05 下，判断工龄和年销售额是否相关。

28. 参照练习 10，建立一个关于评分和笔记本的总的等级关系的估计回归方程。在显著性水平为 0.05 下，检验两个变量是否相关。作出 ANOVA 表，得出你的结论。

29. 参照练习 21，利用制造产品的产量和总成本的数据，建立一个关于产量和总成本的估计回归方程。在显著性水平为 0.05 下，检验产量和总成本是否相关。作出 ANOVA 表，得出你的结论。

30. 参照练习 22，利用下面数据来确定公司总经理和主要高级职员的报酬是否根据公司的业绩为依据(*Business Week*，April 21, 1997)。

公司	两年股票收益变动率（%）	报酬变动率（%）
Dow Chemical	201.3	18
Rohm & Haas	146.5	28
Morton International	76.7	10
Union Carbide	158.2	28
Praxair	-34.9	15
Air Products & Chemicals	73.2	-9
Eastman Chemical	-7.9	-20

是否有证据表明这两个变量之间有显著性关系？进行适当的统计检验并陈述你的结论，取 $\alpha=0.05$。

31. 参照练习 20，建立关于一个家庭典型收入和房屋价格的估计回归方程。在显著性水平为 0.01 下，检验家庭典型收入和房屋价格是否相关。

14.6 利用估计回归方程进行估计和预测

简单线性模型是对 x 和 y 之间关系的假设，利用最小二乘法，得出估计简单线性回归方程。如果从统计中可得出 x 和 y 之间关系显著，则估计回归方程的拟合程度比较高，可利用该估计回归方程进行估计和预测。

点估计

在 Armand 比萨饼连锁店的例子中，估计回归方程$\hat{y}=60+5x$表示学生人数和季度销售额 y 之间的估计关系。我们可利用估计回归方程给出某一特定的 x 相对应的 y 的均值的点估计。例如，假定 Armand 管理者想得到所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额的点估计。利用估计回归方程$\hat{y}=60+5x$，得出对于 $x=10$（或 10 000 名学生），$\hat{y}=60+5(10)=110$。因此，所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额为 110 000 美元。

现在 Armand 管理者想得到位于有 10 000 名学生的 Talbot 大学附近的餐馆每季度销售额。此时，我们对所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额不感兴趣，我们只想预测个别餐馆的季度销售额。可以证明，个别餐馆的点估计与 y 的平均值的点估计一样。因此，我们可以用$\hat{y}=60+5(10)=110$ 或 110 000 美元作为个别餐馆的季度销售额的预测值。

区间估计

点估计没有提供相关估计值的任何准确信息，因此我们可以采用第 8 章、第 10 章和第 11 章中的区间估计。区间估计的第一种类型为**置信区间估计**(confidence interval eetimate)，是指对于给定的 x **值**，y **均值**的区间估计。区间估计的第二种类型为**预测区间估计**(prediction interval estimate)，此方法可得出对于给定的 x 值相对应的 y **个别值的区间估计**。均值的点估计与个别值的点估计一样。

置信区间和预测区间表示回归结果的正确程度，区间越小表示精确度越高。

y 的均值的置信区间估计

对于给定的 x 值估计回归方程给定了 y 均值的点估计，在下面的置信区间估计的方法中，我们可以使用下列符号：

x_p = 自变量 x 的一个特定值或给定值；

$E(y_p)$ = 对于给定的 x_p，因变量 y 的平均值或期望值；

当 $x = x_p$，$\hat{y}_p = b_0 + b_1 x_p$ 为 $E(y_p)$的点估计。

利用该符号来表示所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额的点估计，当 $x_p = 10$，$E(y_p)$表示所有餐馆的未知的平均销售额。$E(y_p)$的估计值为$\hat{y}_p = 60 + 5\ (10)\ = 110$。总体而言，$\hat{y}_p$ 不能完全等于 $E(y_p)$。如果想推断平均值 $E(y_p)$与$\hat{y}_p$ 的紧密程度，就需要估计$\hat{y}_p$ 的反差。已知的 x_p，估计$\hat{y}_p$ 的方差记为$s^2_{\hat{y}_p}$，计算公式如下：

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \tag{14.22}$$

$\hat{y}_p$ 的标准差估计为公式(14.22)的平方根：

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{14.23}$$

在 14.5 节中 Armand 比萨饼连锁店的计算结果得出 $s = 13.829$。$x_p = 10$，$\bar{x} = 14$，$\sum (x_i - \bar{x})^2 = 568$，利用公式(14.23)得出

$$s_{\hat{y}_p} = 13.829\sqrt{\frac{1}{10} + \frac{(10-14)^2}{568}}$$

$$= 13.829\ \sqrt{0.128\ 2} = 4.95$$

$E(y_p)$的置信区间估计的一般形式如下：

$E(y_p)$的置信区间估计

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{14.24}$$

式中，置信系数为 $1-\alpha$，$t_{\alpha/2}$ 是基于自由度为 $n-2$ 的 t 分布。

区间估计的误差边限为 $t_{\alpha/2}\, s_{\hat{y}_p}$。

对所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额，建立置信度为 95% 的区间估计，得出当 $\alpha/2 = 0.025$，自由度为 $n-2 = 10-2 = 8$ 时的值，查附录 B 中的表 2，得 $t_{0.025} = 2.306$。因此，当$\hat{y}_p = 110$ 时，极限误差 $t_{\alpha/2}\, s_{\hat{y}_p} = 2.306(4.95) = 11.415$，置信度为 95% 的区间估计为

$$110 \pm 11.415$$

故所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额的置信度为 95% 的置信区间为 110 000 ± 11 415(美元)。所以当学生人数为 10 000 时，平均销售额的置信区间为 98 585 美元到 121 415 美元。

注意由公式(14.23)给出的 x_p 的标准差的估计，当 $x_p=\bar{x}$和量化形式 $x_p-\bar{x}=0$ 时是最小的。在这种情况下，$\hat{y}_p$ 的标准差估计值为

$$s_{\hat{y}_p}=s\sqrt{\frac{1}{n}+\frac{(\bar{x}-\bar{x})^2}{\sum(x_i-\bar{x})^2}}=s\sqrt{\frac{1}{n}}$$

这一结果意味着当 $x_p=\bar{x}$时，我们能得出 y 的平均值的最好或最准确的估计。事实上，x_p 越远离$\bar{x}$，$x_p-\bar{x}$越大。故当远离$\bar{x}$ 时，y 的均值的置信区间将变大，图 14.8 画出此种情形。

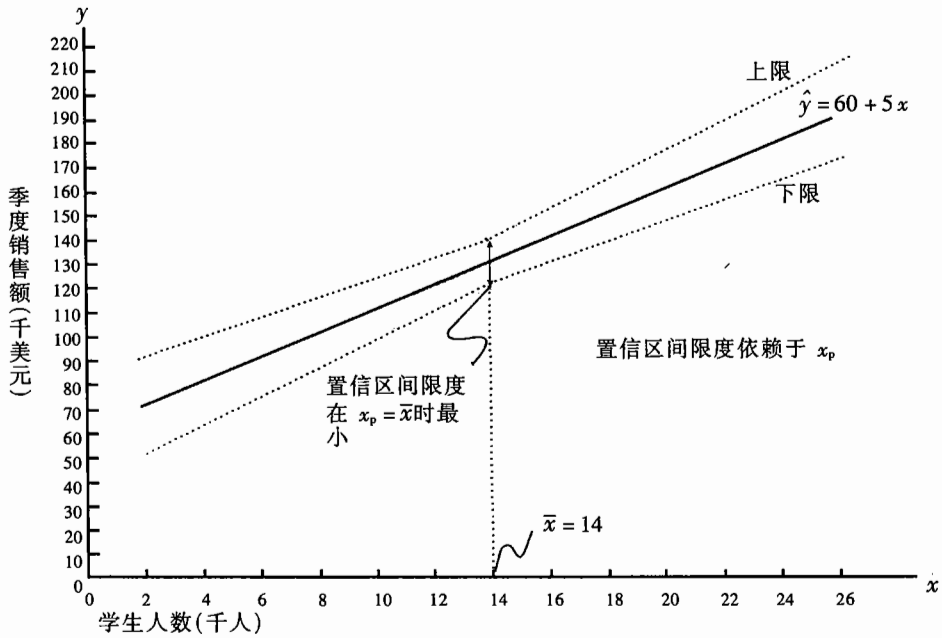


图 14.8　已知学生人数 x、平均销售额 y 的置信区间

个别 y 值的预测区间估计

假定所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额，然后估计有 10 000 名学生的 Talbot 大学附近的餐馆的销售额，正如前面表明的那样，假定 $x=x_p$，y 的个别值的点估计可利用估计的回归方程$\hat{y}_p=b_0+b_1x_p$ 求得。对于 Talbot 大学附近的餐馆，得 $x_p=10$，相应的估计的销售收入为$\hat{y}_p=60+5\ (10)\ =110$，或 110 000 美元。注意此值与所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额的点估计是一致的。

为了得出预测区间估计，首先我们必须确定当 $x=x_p$ 时，用$\hat{y}_p$ 来作为 y 的个别值的估计相联系的方差，这个方差由以下两部分组成：

1. 关于平均值 $E(y_p)$的个别 y 值的反差，它的估计量由 s^2 给出。
2. 与利用$\hat{y}_p$ 估计 $E(y_p)$相联系的反差，它的估计量由$s^2_{\hat{y}_p}$ 给出。

个别 y_p 值的估计方差，记为s^2_{ind}，它的计算形式如下：

$$\begin{aligned} s^2_{ind} &= s^2 + s^2_{\hat{y}_p} \\ &= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \\ &= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \end{aligned} \tag{14.25}$$

因此，个别 y_p 值的标准差的估计量为

$$s_{ind} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{14.26}$$

在 Armand 比萨饼连锁店例子中，对于位于有 10，000 名学生的校园附近的某一餐馆季度销售额与销售额预测相对应的估计标准差的计算如下：

$$\begin{aligned} s_{ind} &= 13.829\sqrt{1 + \frac{1}{10} + \frac{(10-14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

y 的个别值的预测区间估计的一般形式如下：

y_p 值的预测区间估计

$$y_p \pm t_{\alpha/2}\, s_{ind} \tag{14.27}$$

置信区间为 $1-\alpha$，$t_{\alpha/2}$ 是基于自由度为 $n-2$ 的 t 分布。

区间估计的误差极限为 $t_{\alpha/2}\, s_{ind}$。

利用 $t_{0.025} = 2.306$ 和 $s_{ind} = 14.69$，得出 Armand 中位于 Talbot 大学附近的餐馆的季度销售额的置信度为 95% 的预测区间。因此，$\hat{y}_p = 110$，误差极限 $t_{\alpha/2}\, s_{ind} = 2.306\ (14.69) = 33.875$，95% 预测区间的估计为

$$110 \pm 33.875$$

故预测区间为 110 000 ± 33 875 美元或 76 125 美元—143 875 美元。注意，某一个餐馆的预测区间比所有位于有 10 000 名学生的校园附近的餐馆平均每季度销售额的置信区间(98 585 美元—121 415 美元)要大，这表示估计 y 的平均值比预测 y 的个别值更准确。

当自变量的值为 $x_p = \bar{x}$时，置信区间估计和预测区间估计一样准确。置信区间和较宽的预测区间的一般形式见图 14.9。

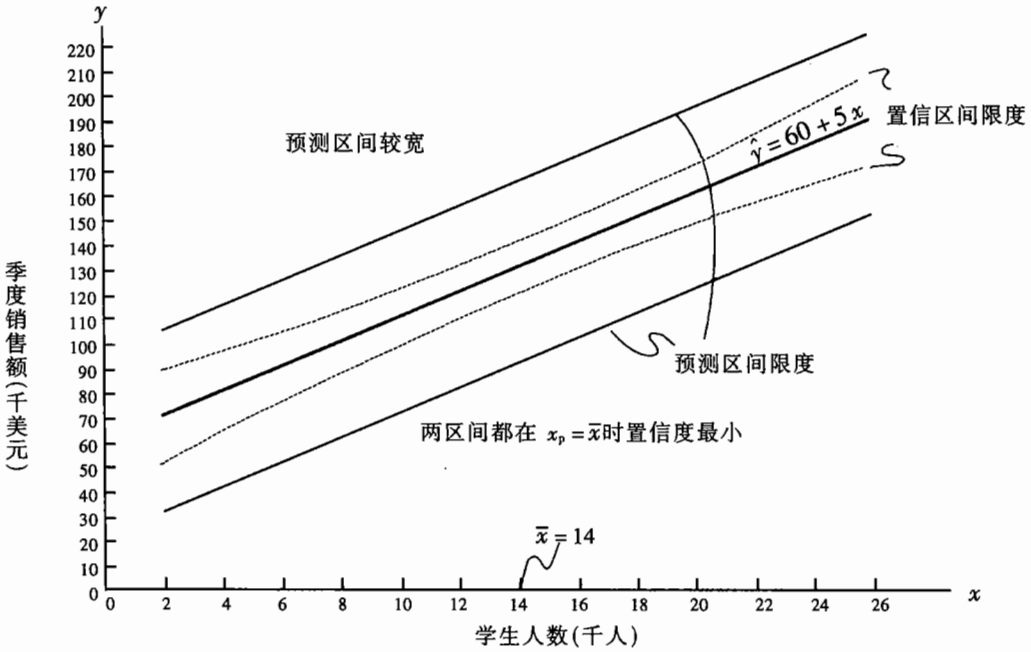


图 14.9　给定学生的人数及销售额的置信区间和预测区间

练　习

方　法

32.　练习 1 的数据如下：

自测题

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. 当 $x=4$ 时，利用公式(14.23)估计 $\hat{y}_p$ 的标准差。

b. 当 $x=4$ 时，利用公式(14.24)求出 y 的期望值的置信度为 95% 的置信区间估计。

c. $x=4$ 时，利用公式(14.26)估计 y 的个别值的标准差。

d. $x=4$ 时，利用公式(14.27)求出置信度为 95% 的预测区间。

33. 练习 2 的数据如下：

x_i	2	3	5	1	8
y_i	25	25	20	30	16

a. 当 $x=3$ 时，估计 $\hat{y}_p$ 的标准差。

b. 当 $x=3$ 时，求出 y 的期望值的置信度为 95% 的置信区间估计。

c. 当 $x=3$ 时，估计 y 的个别值的标准差。

d. 当 $x=3$ 时，求出置信度为 95% 的 y 的预测区间。

34. 练习 3 的数据如下：

x_i	2	4	5	7	8
y_i	2	3	2	6	4

当 $x=3$ 时，求出置信度为 95% 的置信区间和预测区间。试解释这两个区间有何不同。

应　用

35. 在练习 18 中，根据平均分数 x 和月薪 y 的数据得出回归方程 $\hat{y}=1\,790.5+581x$。


自测题

a. 对于总评分为 3.0 的学生，求出起点平均年薪置信度为 95% 的置信区间。

b. 如果 Joe Heller 是一名总评分为 3.0 的学生，求出起点月薪置信度为 95% 的预测区间估计。

36. 在练习 10 中，根据笔记本性能打分(x)和总的等级评分(y)的数据，得出估计回归方程 $\hat{y}=51.819+0.1452x$(*PC World*，February 2000)。


CD 光盘数据
PCS

a. 对于性能打分为 200 的笔记本，求出其总的等级评分的点估计？

b. 对于所有性能打分为 200 的笔记本，求出总的等级评分的置信度为 95% 的区间估计。

c. 假定戴尔公司生产的一款新的计算机性能分数为 200。求出这款新的计算机置信度为 95% 的预测区间估计。

d. 对比(b)和(c)中的答案，讨论它们的不同之处。

37. 在练习 13 中，给出了有关纳税人调整的总收入和逐项列明的扣减项的数据，数据以千美元为单位。如果纳税人调整后的总收入为 52 500 美元，利用回归估计方程 $\hat{y}=4.68+0.16x$，可以得出他的总的逐项列明的扣减项的合理水平的点估计为 13 080 美元。

a. 对于所有调整后的总收入为 52 500 美元的纳税人，求出平均总的逐项列明的扣减项的置信度为 95% 的置信区间估计。

b. 对一名调整后的总收入为 52 500 美元的纳税人，求出总的逐项列明的扣减项的置信度为 95% 的预测区间估计。

c. 假如(b)中的纳税人宣称其总的逐项列明的扣减项为 20 400 美元，国内税务局要求对该纳税人进行审计的要求是否合理？

d. 根据(b)中的答案，如果调整后的总收入为 52 500 美元，国内税务局对其纳税人进行审计之前，请给出一个有关总的逐项列明的扣减项的方案。

38. 参照练习 21，根据某一生产过程中的产量和其总成本 y 的数据，得出估计回归方程 $\hat{y}=1\,246.67+7.6x$。

a. 企业生产计划表明下个月将生产 500 个产品，求下个月的总成本的点估计。

b. 求下个月总成本置信度为 99% 的预测区间估计。

c. 如果下个月末的会计报告中称下个月实际生产成本为 6 000 美元，管理者是否认为这个月总成本过高？试讨论。

39. 尼尔森媒体研究搜集某一种广告收看的户数。这些数据对广告客户非常有用，因为这将使广告客户了解有多少消费者看到此广告。下面数据是关于收看家庭户数(百万)和 1997 年 10 月 27 日至 11 月 2 日各种广告播放次数(*USA Today*, November 17, 1997)的：



CD 光盘数据
Nielsen1

登广告品牌	广告播放次数	收看广告的家庭户数
McDonald's	49	359.6
Burger King	42	296.1
HBO	30	271.6
Red Corner movie	26	251.1
Pizza Hut	31	229.3
Sears	20	186.9
Isuzu Rodeo	21	186.3
MCI	24	172.7
Sprint	15	166.0
JC Penney	19	162.1

a. 对于给定的广告播放次数，根据这些数据，得出预测收看广告的家庭户数的估计回归方程。

b. 估计回归方程是否有很好的拟合程度？试解释。

c. 对于所有播放次数为 35 的广告，求出收看广告的家庭户数的置信度为 95% 的置信区间估计。

d. 假设 Wendy's 广告播放次数为 35。求出收看广告的家庭户数的置信度为 95% 的预测区间估计。

14.7 计算机方法

如果没有计算机，进行回归分析计算将是十分耗时间的。在本节中我们介绍利用 Minitab 软件包来减少计算负担。

首先我们将 Armand 例中的学生人数和销售额数据输入 Minitab 工作表中。为了便于理解计算机输出结果，我们将自变量命为 Pop，因变量为 Sales。利用 Minitab，得出 Armand 比萨饼连锁店的输出结果，见图 14.10*。输出结果的解释如下：

* Minitab 必要计算步骤的输出结果见附录 14.3。

```
The regression equation is
Sales = 60.0 + 5.00 Pop

Predictor       Coef      Stdev    t-ratio       p
Constant      60.000      9.226       6.50   0.000
Pop           5.0000     0.5803       8.62   0.000

s = 13.83       R-sq = 90.3%     R-sq(adj) = 89.1%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1       14200       14200      74.25   0.000
Error         8        1530         191
Total         9       15730

       Fit Stdev.Fit          95% C.I.             95% P.I.
    110.00      4.95   (  98.58, 121.42)   (  76.12, 143.88)
```

估计回归方程

ANOVA 表

区间估计

图 14.10　Armand 比萨饼连锁店的 Minitab 的输出结果

1. Minitab 打印出来的估计回归方程为 Sales = 60.0 + 5.00 Pop。
2. 打印出来的表中给出系数 b_0 和 b_1 值，每个系数的标准差，每个系数除以它的标准差得到的 t 值以及与 t 检验相联系的 p – 值。因此，为了检验 H_0：$\beta_1 = 0$ 和对立假设 H_a：$\beta_1 \neq 0$，我们将 8.62(位于 t 值列)和一个适当的临界值比较。这就是 14.5 节中所叙述的 t 检验步骤。同样，我们也可利用 Minitab 给出 p – 值进行相同的检验。因为在这个例子中 p – 值(精确到小数点后的 3 位有效数字)，样本结果表明应该拒绝原假设(H_0：$\beta_1 = 0$)。
3. Minitab 打印出的估计标准误差，$s = 13.83$ 和有关拟合程度的信息。注意“R-sq = 90.3%”是用百分比来表示的确定系数。输出“R-sq(adj) = 89.1%”将在第 15 章中介绍。
4. ANOVA 表打印在标题 Analysis of Variance 下面。注意 DF 是自由度的缩写，MSR = 14 200，MSE = 191。F 是这两个值之比为 74.25。在 14.5 节中，我们介绍了怎样利用 F 值来判断 Sales 和 Pop 之间是否存在显著关系。Minitab 同样还打印了 F 检验的 p – 值。因为 p – 值为零(精确到小数点后的 3 位有效数字)，从统计上可以判断其关系是显著的。
5. 在有 10 000 名学生的校园附近的餐馆的期望销售收入的 95% 的置信区间估计和 95% 的预测区间估计，打印在 ANOVA 表下面。与 14.6 节中介绍的相同，置信区间为 (98.58，121.42)，预测区间为(76.12，143.88)。

练 习

应 用

40. 自测题 一家房地产公司的销售部门想了解年租金 x(千美元)和销售价格 y(千美元)之间的关系进行回归分析。收集最近销售的公寓住宅的数据,下面是经过计算机处理得到的输出结果:

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor        Coef      Stdev    t-ratio
Constant       20.000     3.2213       6.21
X               7.210     1.3626       5.29

Analysis of Variance

SOURCE         DF           SS
Regression      1      41587.3
Error           7
Total           8      51984.1
```

a. 样本中有多少公寓住宅?

b. 写出估计回归方程。

c. 求出s_{b_1}值。

d. 在显著性水平为 0.05 下,利用 F 检验来进行显著性检验。

e. 如果总的年租金为 50 000 美元,估计公寓住宅的销售价格。

41. 下面是关于某种计算机终端的维护费用 y(英镑/月)和使用率 x(小时/周)的回归分析的计算机输出结果:

```
The regression equation is

Y = 6.1092 + .8951 X

Predictor        Coef      Stdev
Constant       6.1092     0.9361
X              0.8951     0.1490

Analysis of Variance

SOURCE         DF           SS          MS
Regression      1      1575.76     1575.76
Error           8       349.14       43.64
Total           9      1924.90
```

a. 写出估计回归方程。

b. 在显著性水平为 0.05 时，利用 t 检验判断每月维护费用是否与使用率有关？

c. 如果计算机终端每个星期使用 25 个小时，利用估计回归方程预测其每月维护费用。

42. 假定得出 x 在分店销售人员的数目和年销售额 y 的回归模型。对于有关数据进行回归分析的计算机输出结果如下：

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor       Coef      Stdev   t-ratio
Constant        80.0     11.333      7.06
X               50.0      5.482      9.12

Analysis of Variance

SOURCE        DF          SS            MS
Regression     1      6828.6        6828.6
Error         28      2298.8          82.1
Total         29      9127.4
```

a. 写出估计回归方程。

b. 在本次研究中有多少家分店？

c. 在显著性水平为 0.05 下，计算 F 统计量，并检验变量之间的关系是否显著？

d. 预测 Memphis 分店的年销售额，这家分店有 12 个销售人员。

43. 全国房屋建筑协会根据市场中等收入的家庭有能力购买住宅的比率，从最强到最弱的市场进行排序。从购买力最强的市场中抽取 12 个市场组成一个样本，中等收入（千美元）和房屋销售的中间价格（千美元）数据如下（*The Wall Street Journal Almanac 1998*）：



CD 光盘数据
HomeBldg

市场	收入（千美元）	价格（千美元）
Syracuse, NY	41.8	76
Springfield, IL	47.7	91
Lima, OH	40.0	65
Dayton, OH	44.3	88
Beaumont, TX	37.3	70
Lakeland, FL	35.9	73
Baton Rouge, LA	39.3	85
Nashua, NH	56.9	118
Racine, WI	46.7	81
Des Moines, IA	48.3	89
Minneapolis – St. Paul, MN	54.6	110
Wilmington, DE – MD	55.5	110

a. 对于给定的市场中等收入，根据这些数据求出能预测市场中房屋中间价格的估计回归方程。

b. 该估计回归方程是否提供了很好的拟合？试解释。

c. 当中等收入为 40 100 美元时，求市场中所有销售房屋的中间价格的 95% 的置信区间估计。

d. 据报道，纽约 Binghamton 的中等收入为 40 100 美元。求在 Binghamton 市场销售房屋的中间价格的 95% 的预测区间估计。

44. Cushman & Wakefield 股份有限公司，收集美国市场上办公用房的空闲率和租金率的数据。对于 18 个选取的销售地区，下面是这些地区的中心商业区的综合空置率(%)和平均租金(美元/平方英尺)的数据(*The Wall Street Journal Almanac 1998*)：



CD 光盘数据
OffRates

市场名称	空置率(%)	平均租金(美元)
Atlanta	21.9	18.54
Boston	6.0	33.70
Hartford	22.8	19.67
Baltimore	18.1	21.01
Washington	12.7	35.09
Philadelphia	14.5	19.41
Miami	20.0	25.28
Tampa	19.2	17.02
Chicago	16.0	24.04
San Francisco	6.6	31.42
Phoenix	15.9	18.74
San Jose	9.2	26.76
West Palm Beach	19.7	27.72
Detroit	20.0	18.20
Brooklyn	8.3	25.00
Downtown, NY	17.1	29.78
Midtown, NY	10.8	37.03
Midtown South, NY	11.1	28.64

a. 以空置率为水平轴，根据这些数据画出散点图。

b. 这两个变量之间呈现什么关系？

c. 求出在办公用房的综合空置率已知时，能用来预测平均租金率的估计回归方程。

d. 在显著性水平 0.05 下检验关系的显著性。

e. 估计回归方程提供了很好的拟合吗？请作出解释。

f. 在一个综合空置率为 25% 的中心商业区，预测该市场的期望租金。

g. 在劳德代尔堡的中心商业区，综合空置率是 11.3%，预测劳德代尔堡的期望租金。

14.8　残差分析：验证模型假定

我们在前面证明了，观察值 i 的**残差**是因变量的观察值（y_i）与估计值（$\hat{y}_i$）之差。

残差分析(residual analysis)是判断模型假定是否合适的主要模型。

第 i 个观察值的残差

$$y_i - \hat{y}_i \tag{14.28}$$

式中　y_i——因变量的观察值；

$\hat{y}_i$——因变量的估计值。

换句话说，第 i 个观察值的残差是利用估计回归方程预测 y_i 值产生的。对于 Armand 比萨饼连锁店，其残差计算结果见表 14.7。因变量的观察值在第二栏，利用估计值回归方程$\hat{y} = 60 + 5x$计算的因变量估计值在第三栏，对应的残差值在第四栏。残差分析将帮助确定回归模型的假设是否合适。

表 14.7　**Armand 比萨饼连锁店的残差计算结果**

学生人数	销售额	估计销售额	残差
x_i	y_i	$\hat{y}_i = 60 + 5x_i$	$y_i - \hat{y}_i$
2	58	70	-12
6	105	90	15
8	88	100	-12
8	118	100	18
12	117	120	-3
16	137	140	-3
20	157	160	-3
20	169	160	9
22	149	170	-21
26	202	190	12

回顾 Armand 比萨饼连锁店例子的回归假设，简单线形回归模型为

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{14.29}$$

这个模型假定季度销售额(y)是学生人数(x)加上一个误差项 ε 的线性函数。在 14.4 节中对误差项的假设如下：

1. $E(\varepsilon)=0$。
2. 对所有的 x，ε 的方差相同，记为 σ^2。
3. ε 值是相互独立的。
4. 误差项 ε 服从正态概率分布。

这些假设是用来判断 x 和 y 之间关系是否显著的 t 检验和 F 检验以及在 14.6 节中的置信区间估计和预测区间估计。如果误差项的假定显得有问题，则回归关系的显著性假设检验和区间估计结果可能不正确。

残差提供了有关误差项 ε 的最佳信息，因此误差分析是判断 ε 的假设是否适当的重要步骤。许多残差分析是基于残差图的观察得出的。在本节中，我们将介绍下面的残差图：

1. 关于自变量 x 的值的残差图。
2. 关于自变量的预测值 $\hat{y}$ 的残差图。
3. 标准残差图。
4. 正态概率图。

关于 x 的残差图

关于自变量 x 的值的**残差图**(residual plot)是用水平轴表示自变量的值，用纵轴表示相对应的残差值。残差图中每个残差用一点表示。每一点的第一个坐标由 x_i 值给出，第二个坐标由响应的 $y_i-\hat{y}_i$ 的残差给出。对于 Armand 比萨饼连锁店，根据表 14.7 的数据作出关于 x 的残差图，第一个点的坐标是(2，-12)，对应的 $x_1=2$ 和 $y_1-\hat{y}_1=-12$；第二个点的坐标是(6，15)，对应的 $x_2=6$ 和 $y_2-\hat{y}_2=15$，等等。图 14.11 是得到的残差图。

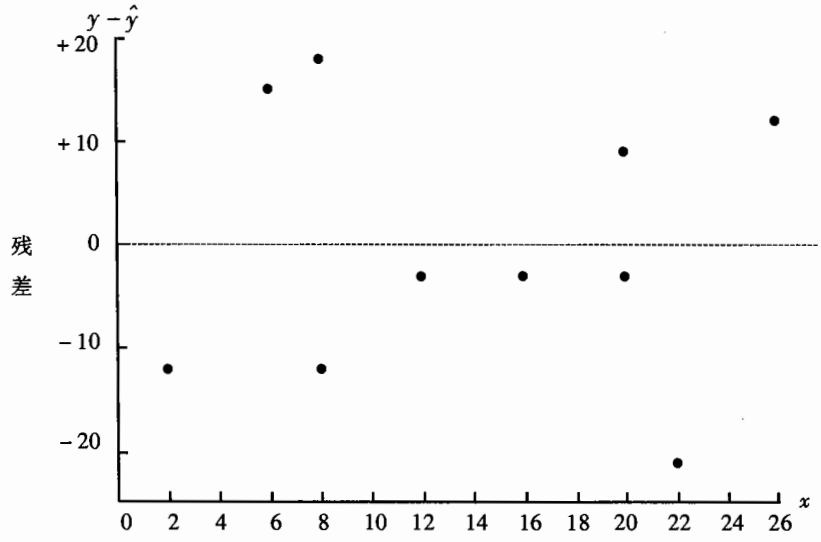


图 14.11　Armand 比萨饼连锁店的关于自变量 x 的残差图

在对此残差图的结果解释之前，首先考虑在任何残差图中观察到的形状。图 14.12 给出了三个例子。如果对于所有的 x 值 ε 的方差都是相同的，并且描述变量之间关系的回归模型是合理的，那么残差图如图 14.12A 中所示，所有的散点都落在一条水平带中间。但是，如果对于所有的 x 值，ε 的方差不相同，例如对于较大的 x 值，回归线的变异性也较大，如图 14.12B 所示，这种情形，就违背了只有一个常数方差的假设。另一种可能的残差图如图 14.12C 所示，

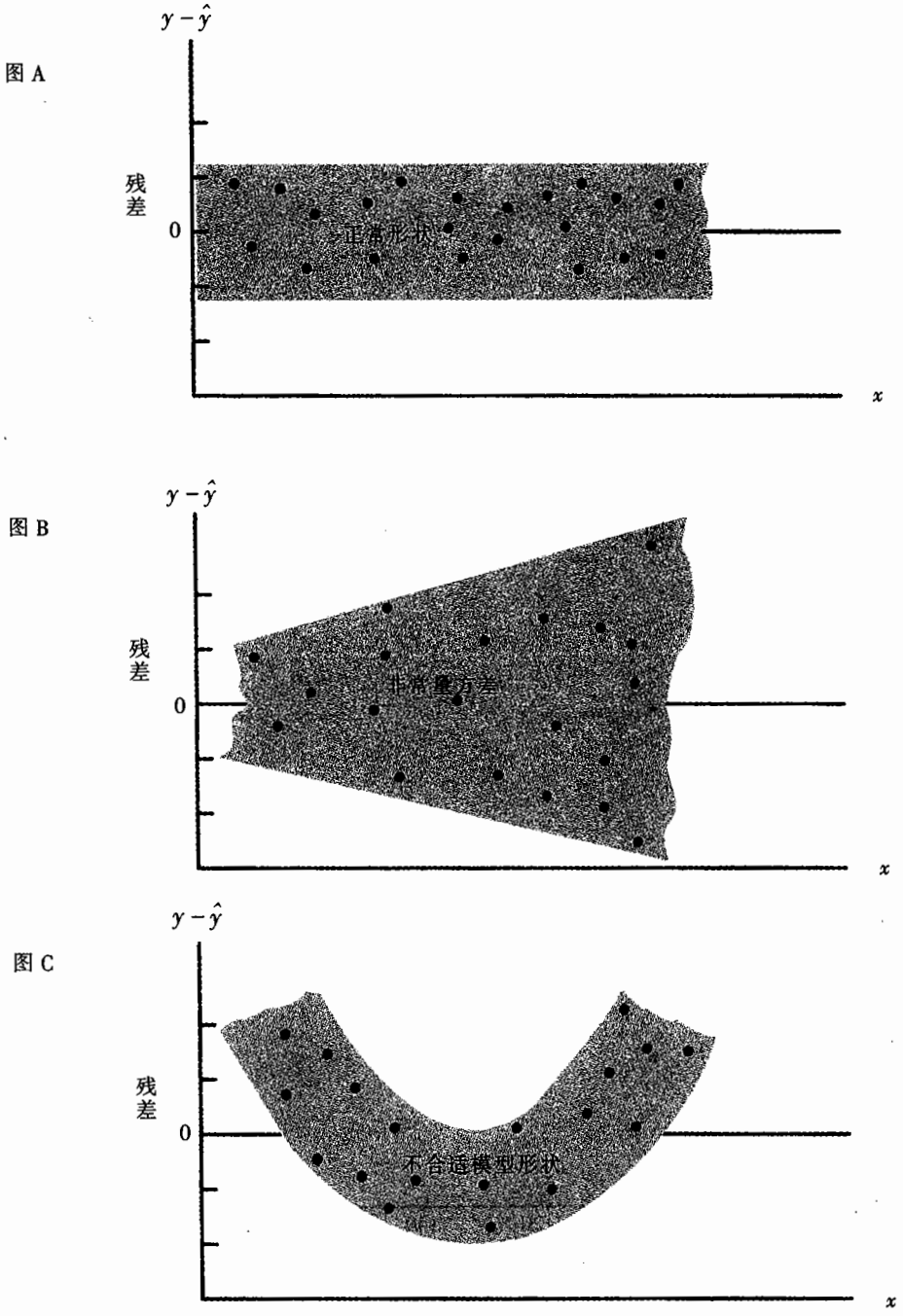


图 14.12　三种回归研究的残差图

在这种情况下，就得出了假定的回归模型不能合理地描述变量之间的关系的结论。这时应考虑曲线回归模型或者多元回归模型。

现在回到图 4.11 所示的 Armand 比萨饼连锁店的残差图，这些残差看来好像与图 14.2A 中的水平型图很相似。因此残差图并没有提供足够的证据使我们对 Armand 比萨饼连锁店回归模型所做的假定表示怀疑。此时，我们可以确信 Armand 简单线性回归模型是合理的。

在对残差图进行有效解释时，经验和好的判断永远是关键的。实际上残差图与图 14.12 中看到的三种类型之一恰好完全一致的情形是很少见的。经常进行回归研究和考察残差图的分析人员是否能清楚地理解合理的部分与模型假设部分的区别，是值得置疑的。如上所示，这里介绍的残差图是用来评价回归模型假设有效性的方法之一。

关于$\hat{y}$的残差图

另一种残差图是在水平轴上因变量的预测值$\hat{y}$和在纵轴上表示的残差值。每个残差都用一点表示。每点的第一个坐标为$\hat{y}_i$，第二个坐标为相对应的第 i 个残差 $y_i - \hat{y}_i$。从表 14.7 的 Armand 数据中可得，第一个点的坐标为(70，-12)，相应的$\hat{y}_1 = 70$，$y_1 - \hat{y}_1 = -12$；第二个点的坐标为(90，15)，等等。图 14.13 是得到的残差图，注意，这种形式的残差图与关于自变量 x 的残差图的类型是相同的。这种形式不会使我们怀疑模型假设。对于简单的线性回归，x 的残差图和$\hat{y}$的残差图的形状一样。对于多元回归分析，因为有多个自变量，所以许多的残差图有着更广泛的应用。

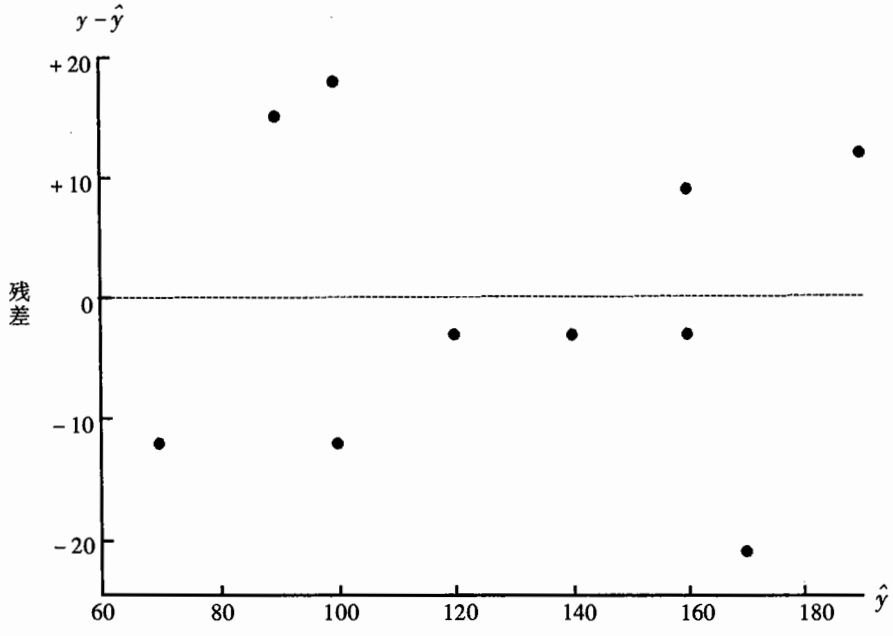


图 14.13 Armand 比萨饼连锁店的关于预测量$\hat{y}$的残差图

标准残差

计算机软件包作出的残差图利用了标准残差形式。就像在前面几章中我们看到的，一个随机变量可以通过减去均值再通过除以标准差得到。利用最小二重法，残差均值为 0。因此，只需简单的标准差，就得到了**标准残差**(standardized residual)。

可以证明，每 i 个残差的标准差取决于估计量的标准误差 s 和对应的自变量 x_i 的值。

第 i 个残差的标准差*

$$s_{y_i-\hat{y}_i} = s\sqrt{1-h_i} \tag{14.30}$$

式中　$s_{y_i-\hat{y}_i}$——第 i 个残差的标准差

s——估计量的标准误差

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum (x_i-\bar{x})^2} \tag{14.31}$$

注意公式(14.30)中有 h_i**，故第 i 个残差标准差依赖于 x_i。一旦计算出每个残差的标准差，则将每个残差除以它的标准差，得出标准残差。

表 14.8 给出了 Armand 比萨饼连锁店例子的标准残差计算过程。上面的计算给出了 $s=13.829$，图 14.14 是关于自变量 x 的标准残差图。

第 i 个观察值的标准残差

$$\frac{y_i-\hat{y}_i}{s_{y_i-\hat{y}_i}} \tag{14.32}$$

表 14.8　Armand 比萨饼连锁店标准残差的计算过程

旅店 i	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})_2$	$\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$	h_i	$s_{y_i-\hat{y}_i}$	$y_i-\hat{y}_i$	标准残差
1	2	-12	144	0.253 5	0.353 5	11.119 3	-12	-1.079 2
2	6	-8	64	0.112 7	0.212 7	12.270 9	15	1.222 4
3	8	-6	36	0.063 4	0.163 4	12.649 3	-12	-0.948 7
4	8	-6	36	0.063 4	0.163 4	12.649 3	18	1.423 0
5	12	-2	4	0.007 0	0.107 0	13.068 2	-3	-0.229 6
6	16	2	4	0.007 0	0.107 0	13.068 2	-3	-0.229 6
7	20	6	36	0.063 4	0.163 4	12.649 3	-3	-0.237 2
8	20	6	36	0.063 4	0.163 4	12.649 3	9	0.711 5
9	22	8	64	0.112 7	0.212 7	12.270 9	-21	-1.711 4
10	26	12	144	0.253 5	0.353 5	11.119 3	12	1.079 2
总计			568					

注意：残差值计算见表 14.7。

* 用 s 代替 σ，方程真实提供了第 i 个残差标准差的估计值。

** h_i 是指观察值 i 的杠杆率。在第 14.9 节中我们考虑有影响的观察值时，将讨论杠杆率。

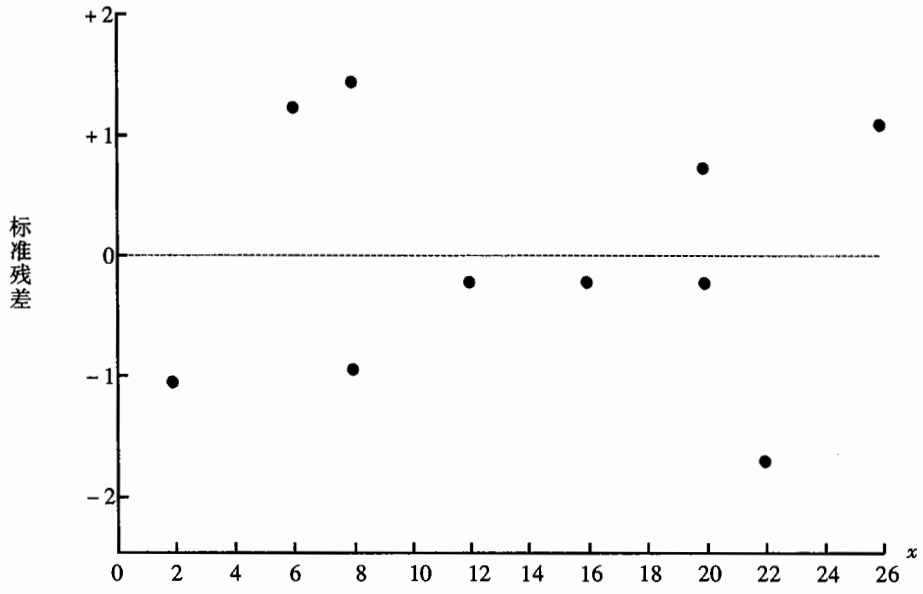


图 14.14　Armand 比萨饼连锁店自变量 x 的标准残差图

标准残差图提供了误差项 ε 服从正态分布假设的直观认识。如果满足这一假设，那么标准残差分布也应服从标准正态分布。于是，当查看标准残差图时，期望看到大约有 95% 的标准残差介于 −2 和 +2 之间。从图 14.14 中可以看出，在 Armand 例子中所有标准残差都在 −2 到 +2 之间。* 所以，根据标准残差的分布，我们没有理由怀疑 ε 服从标准正态分布的假定。

正态分布中的小的偏差对回归分析的统计检验没有大的影响。

因为需要计算估计值 $\hat{y}$，残差和标准残差，所以大多数统计软件包都将其作为可供选择的回归分析输出结果。因此，很容易得到残差图。对于较复杂的问题，为了作出本节所介绍的残差图，计算机软件包是惟一的实用工具。

正态概率图

另一种判断误差项服从正态分布的假设有效性的方法是**正态概率图**(normal probability plot)。为了介绍如何作出一张正态概率图，我们引入**正态分数**的概念。

假定从一个平均值为 0、标准差为 1 的正态概率分布中随机选取 10 个数值，并重复这一抽样过程，然后把每个样本中的 10 个数值按从小到大的顺序排列。现在我们考虑每个样本中的最小值。在反复进行的抽样过程中得到的最小值是一个随机变量，称之为一阶顺序统计量。

* 因为在(14.30)中用 s 代替 σ，从严格意义上说标准残差的概率分布并不是技术上的正态分布。但是，在大多数回归例子中，当样本大小足够大时，可近似于正态分布。

统计学家证明，对于来自标准正态概率分布的容量为 10 的样本，一阶统计量的期望值为 -1.55。这个期望值称为正态分数。对于样本容量 $n=10$ 的情形，有 10 个顺序统计量和 10 个正态分数(见表 14.9)。一般说，如果数据值由 n 个观察值组成，那么就有 n 个顺序统计量和 n 个正态分数。

表 14.9　$n=10$ 的正态分数

统计量顺序	正态分数
1	-1.55
2	-1.00
3	-0.65
4	-0.37
5	-0.12
6	0.12
7	0.37
8	0.65
9	1.00
10	1.55

对于 Armand 比萨饼连锁店例子，现在我们介绍如何利用 10 个正态分数去确定标准残差服从标准正态概率分布。首先将 10 个标准残差进行排序，见表 14.8。我们将 10 个正态分数和 10 个排好顺序的标准残差放在一起，如表 14.10 所示。如果满足正态性假设，那么最小的标准残差应接近最小的正态分数，第二小的标准差应接近第二小的正态分数，依此类推。用水平轴表示正态分数，纵轴表示标准残差，作一散点图，如果标准残差近似于正态分布，那么在图上标出的这些散点，应密集在通过坐标原点的 45°线附近。这张散点图为**正态概率图**。

表 14.10　Armand 比萨饼的正态分数和排序的标准残差

正态分数	排序的标准残差
-1.55	-1.711 4
-1.00	-1.079 2
-.65	-0.948 7
-0.37	-0.237 2
-0.12	-0.229 6
0.12	-0.229 6
0.37	0.711 5
0.65	1.079 2
1.00	1.222 4
1.55	1.423 0

图 14.15 是 Armand 比萨饼连锁店的正态概率分布图。根据这些散点偏离直线的程度可以判断标准残差是否来自一个标准正态概率分布。在图 14.15 中，可以看到这些散点紧密地聚集在 45 度线附近。所以我们得出误差项服从标准正态概率分布的假设是合理的。一般地说，越多的点越靠近 45°线，越可支持正态假设。在正态概率散点图上，与 45°线偏离越大，就越可证明残差不是来自一个标准正态概率分布。正态分数和与此相联系的正态概率图，可以很容易地利用统计软件包如 Minitab 得到。

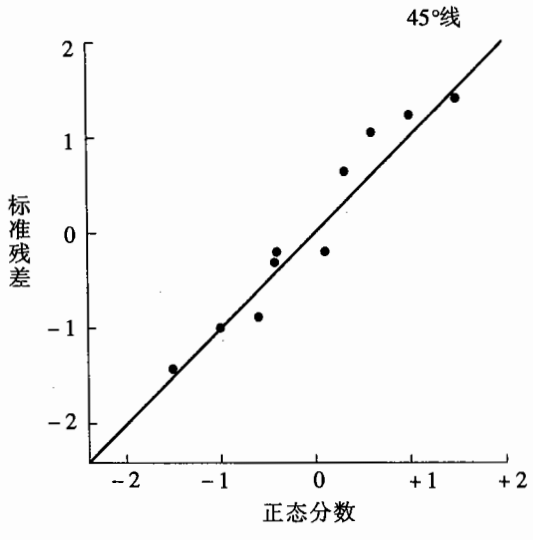


图 14.15　Armand 比萨饼连锁店的正态概率分布图

评 注

1. 我们利用残差和正态概率图去验证回归模型假设的有效性。如果验证表明一个或多个假设存在问题，就需要考虑一个不同的回归模型或进行数据转换。当违背假设时，一个适当的修正措施应依赖良好的判断，来自有经验的统计学家的建议可能有宝贵的价值，应对其重视。

2. 残差分析是用来证实回归模型假设成立的主要方法。即使没有发现任何被违背的假设，也不能得出这个模型给出好的预测的结论。但是，如果有其他的统计检验补充支持显著性结论，并且有较大的确定系数，那么可以用该统计的回归方程得到好的估计值和预测值。

练 习

方 法

45. 已知变量 x 和 y 的数据如下：

自测题

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. 根据这些数据，求出估计回归方程。

b. 计算残差。

c. 作出关于自变量 x 的残差图，关于误差项的假定是否得到满足？

d. 计算标准残差。

e. 作出关于 $\hat{y}$ 的标准残差图，从该图中得出你的结论。

46. 根据下列数据进行回归研究：

观察次数	x_i	y_i	观察次数	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. 根据这些数据，求出回归方程。

b. 作出残差图，关于误差项的假定是否得到满足？

应　用

47. 下面是关于四季餐厅的广告费支出（千美元）和收入（千美元）的数据：


自测题

广告费	收入
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a. 令 x 为广告费支出(千美元)，y 为收入(千美元)。利用最小二乘法，求出近似这两个变量之间关系的直线。

b. 在显著性水平为 0.05 下，检验广告费支出和收入是否相关。

c. 依据 $\hat{y}$ 准备出 $y-\hat{y}$ 的残差图，利用(a)的结论获得 $\hat{y}$ 的值。

d. 从残差分析中得出你的结论。是否使用该模型，还是寻找一个更好的模型？

48. 参照练习 9，已经得出关于销售员的工龄和年销售额的估计回归方程。

a. 计算残差并作出残差图。

b. 根据残差图，关于误差项的假定是否合理？

49. NYSE（纽约股票交易所）授权 10 家美国存托凭证公司(ADRs)代表外国公司的股票并将其存入美国国内的银行。下面表格是一家新的 ADRs 公司为 10 家印度公司进行承托凭证的市盈率(P/E)和投资回报率(ROE)(*Bloomberg Personal Finance*，April 2000)：


CD 光盘数据
ADRs

	ROE	P/E
Bharti Televentures	6.43	36.88
Gujarat Ambuja Cements	13.49	27.03
Hindalco Industries	14.04	10.83
ICICI	20.67	5.15
Mahanagar Telephone Nigam	22.74	13.35
NIIT	46.23	95.59
Pentamedia Graphics	28.90	54.85
Satyam Computer Services	54.01	189.21
Silverline Technologies	28.02	75.86
Videsh Sanchar Nigam	27.04	13.17

a. 利用计算机软件包来建立表达 y = P/E 和 x = ROE 之间关系的估计回归方程。

b. 作出关于自变量的标准残差的残差图。

c. 根据残差图能否得出误差项和模型是否合理的结论?

14.9 残差分析：异常值和有影响的观察值

在 14.8 节中我们介绍了当违背回归模型假设时，如何利用残差分析去判断。在本节中，我们将讨论在确定估计回归方程时，如何利用残差分析来识别异常值或有影响的观察值。这一节介绍当出现这些观察值时，我们将采取哪些步骤。

检测异常值

图 14.16 是数据集合中只有一个**异常值**(outlier)的散点图，这个异常值是一个数据点(观察值)，它与其余的点所显示的趋势不相吻合。异常值代表需要怀疑并且值得加以仔细检查的观察值，它们可能代表错误的数据。如果确实如此，需修正这些数据，它们也可能是违背模型假设的情形，倘若如此，就需要考虑其他模型。最后，它们可能完全是由于随机原因而产生的异常值，此时，应保留这些异常值。

为了解释检测异常值的过程，以表 14.11 中的数据集为例，图 14.17 是其数据集的散点图。

表 14.11 解释异常值的数据集

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

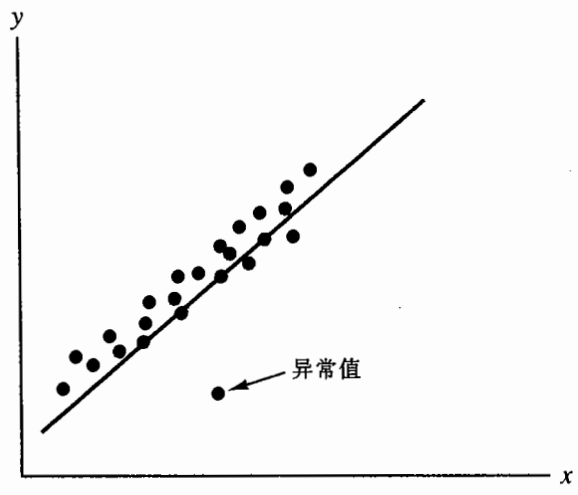


图 14.16 有一个异常值的数据集

除了第四个观察值（$x_4 = 3$，$y_4 = 75$）以外，散点图的形状明显表示存在一个负线性关系。事实上，根据散点图上其余数据点的形状，我们期望 y_4 更小，因此我们将其对应的观察值确认为异常值。对于简单性回归，人们只要简单观看散点图就能识别异常值。

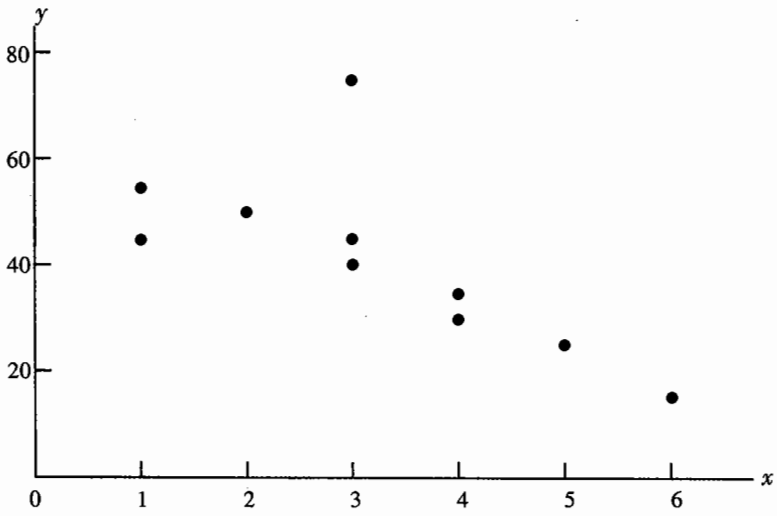


图 14.17　异常值数据集的散点图

标准残差也可用来识别异常值。如果一个观察值与其他数据的形状偏离很大(例如图 14.16 中的异常值),那么对应的标准残差绝对值将很大。许多计算机软件包可自动识别具有大的标准残差绝对值的观察值。在图 14.18 我们显示了对表 14.11 中的数据进行回归分析的 Minitab 输出结果。接下来输出结果的最后一行显示了观察值为 4 时的标准残差是 2.67。Minitab 指出标准差小于 -2 或大于 2 的任何观察值都被看做是不正常值。在这种情形下，这个观察值被单独打印在一行，并在标准残差后面打印一个字母 R，如图 14.18 所示。因为误差项服从正态分布，所以大约应有 5% 的标准残差超出这个范围。

```
The regression equation is
Y = 65.0 - 7.33 X

Predictor       Coef       Stdev    t-ratio        p
Constant      64.958       9.258       7.02    0.000
X             -7.331       2.608      -2.81    0.023

s = 12.67       R-sq = 49.7%     R-sq(adj) = 43.4%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      1268.2      1268.2      7.90    0.023
Error         8      1284.3       160.5
Total         9      2552.5

Unusual Observations
Obs.      X          Y       Fit Stdev.Fit  Residual   St.Resid
  4    3.00      75.00     42.97      4.04     32.03      2.67R

R denotes an obs. with a large st. resid.
```

图 14.18　异常值数据集回归分析的 Minitab 输出结果

在决定如何处理异常值时，首先我们检验这个异常值是否是一个有效的观察值。也许在最初记录时或将数据录入计算机文件时，就已经有误差了。例如，假定在核定表 14.17 中的异常值时，我们发现第四个观察值($x_4=3$，$y_4=75$) 出现误差，将其修正为 $x_4=3$，$y_4=30$。图 14.19 是修正 y_4的数值后得到的 Minitab 输出结果。我们看到，利用不经修正的数据会对拟合优度产生相当大的影响；而用修正数据，R-sq 的值从 49.7% 增加到 83.8%，b_0 的值从 64.958 减少到 59.237，回归线的斜率也从 -7.331 改变到 -6.949。识别出异常值使我们能够修正数据的误差并提高回归效果。

```
The regression equation is
Y = 59.2 - 6.95 X

Predictor       Coef      Stdev    t-ratio       p
Constant      59.237      3.835      15.45   0.000
X             -6.949      1.080      -6.43   0.000

s = 5.248       R-sq = 83.8%     R-sq(adj) = 81.8%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1      1139.7      1139.7     41.38   0.000
Error         8       220.3        27.5
Total         9      1360.0
```

图 14.19　异常值数据集修正后的 Minitab 输出结果

检测有影响的观察值

有时一个或几个观察值对回归结果有重要影响。图 14.20 介绍在简单线性回归情形下一个**有影响的观察值**(influential observation)的例子。在图中，估计回归线有一个负斜率，但是，如果将有影响的数据剔除数据集，估计回归线的斜率从负变成正，y 的截距变小。很明显，对于确定估计回归线，这个观察值的影响要比其他观察值的影响大得多，而从数据集中删除其他任一数据，对估计回归方程的影响较小。

当模型只存在一个自变量时，可从散点图识别出有影响的观察值。一个有影响的观察值可能是一个异常值(即有一个 y 的观察值远偏离散点图中的趋势)，它可能对应着远离平均值的 x 的观察值等等，(见图 14.20)，或者它也可能是由在某种程度上偏离 y 值的趋势同时在某种程度有趋近于 x 的极端值组合而引起的观察值。

因为有影响的观察值对估计的回归方程可能有如此明显的影响，所以需要对此观察值仔细检查。首先我们应该查明在采集或者录入数据时有没有出现错误。如果出现了错误，我们应修正并求出新的估计回归方程。如果这个观察值有效，我们应该为这一观察值感到幸运。因为如果这个观察值有效，就能帮助我们更好理解模型，并且使我们得到一个比较好的估计回归方程。在图

14.20 的数据集中，存在一个有影响的观察值，如果它是有效的观察值，那么将通过取得的中间值去判断 x 和 y 之间的一个更适合的关系。

自变量的值是极端值的观察值为**高杠杆率点**(high leverage points)。图 14.20 中的有影响的观察值是一个具有高杠杆率的点。可以根据自变量的值与它们的平均值距离的远近确定一个观察值的杠杆率。对于只有一个自变量的情形，用 h_i 表示第 i 个观察值的杠杆率，可利用公式(14.33)计算。

第 i 个观察值的杠杆率

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \tag{14.33}$$

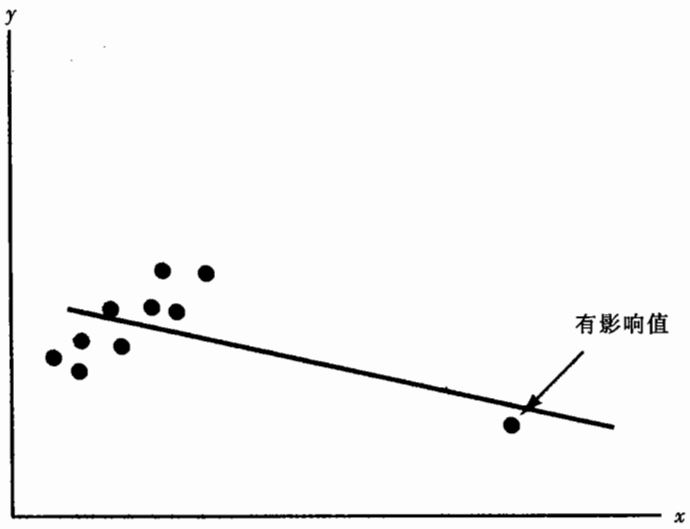


图 14.20　有影响的观察值的数据集

表 14.12　具有高杠杆率的观察值的数据集

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

从公式中可以看出，距离平均值愈远，第 i 个观察值的杠杆率愈高。

许多统计软件包能自动地识别高杠杆率的观察值作为标准回归分析输出的一部分。为了说明 Minitab 统计软件包如何识别具有高杠杆率的观察值，考虑表 14.12 中的数据集。

图 14.21 是表 14.12 中数据集的散点图，很清楚地看出第 7 个观察值($x = 70$, $y = 100$)是一个含有变量 x 的极端值的观察值。因此，我们期待高杠杆率的点被识别出来。对于这个观察值，利用公式(14.33)计算杠杆率。

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\sum (x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2\,621.43} = 0.94$$

对于简单线性回归，如果观察值的杠杆率 $h_i > \frac{6}{n}$，那么 Minitab 将把这个观察值识别为高杠杆率的观察值。对于表 14.12 中的数据集，

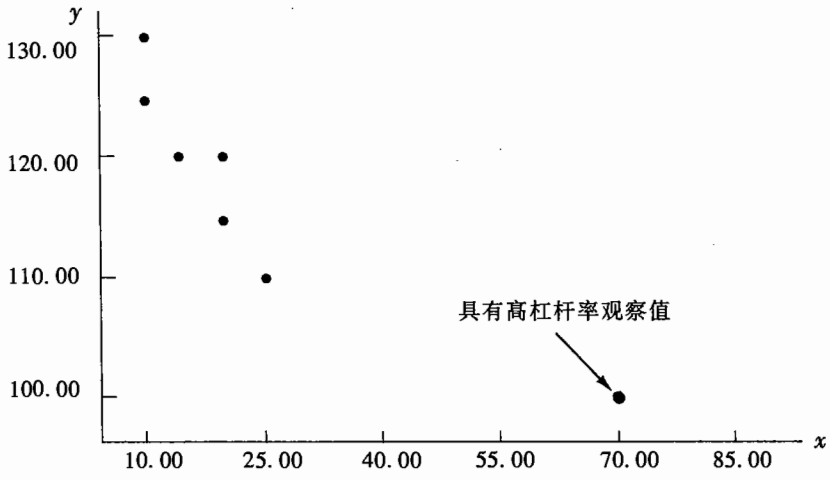


图 14.21 具有高杠杆率观察值的数据集的散点图

$\frac{6}{n}=\frac{6}{7}=0.86$。因为 $h_7=0.94>0.86$，所以 Minitab 将把第 7 个观察值识别为有高杠杆率的点。图 14.22 给出了这个数据集回归分析的 Minitab 输出结果。第 7 个观察值($x=70$, $y=100$)被识别为有重要影响的观察值，并在表的底部单独打印一行，在这行的右边缘处用 X 标出。

可以用计算机软件包识别有影响的观察值，Minitab 的选择规则讨论如下。

有影响的观察值是因为大的残差和高杠杆率的交互作用而产生的，所以一般很难检测出。在检测某一观察值的影响时，可利用两种检测方法，其中一种被称为库克(Cook) D 统计量，将在第 15 章中加以讨论。

```
The regression equation is
Y = 127 -0.425 X

Predictor       Coef       Stdev    t-ratio        p
Constant     127.466       2.961      43.04    0.000
X           -0.42507     0.09537      -4.46    0.007

s = 4.883       R-sq = 79.9%     R-sq(adj) = 75.9%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      473.65      473.65     19.87    0.007
Error         5      119.21       23.84
Total         6      592.86

Unusual Observations
Obs.      X          Y       Fit Stdev.Fit  Residual   St.Resid
  7    70.0     100.00     97.71      4.73      2.29       1.91 X

X denotes an obs. whose X value gives it large influence
```

图 14.22 具有高杠杆率观察值的数据集的 Minitab 输出结果

评　注

一旦一个观察值由于大的残差和高杠杆率被识别为是潜在的有影响的观察值，就需要这个观察值，对估计回归方程的影响进行评估。较高级的教科书将会对此进行讨论。但是，如果人们不熟悉教科书的内容，那么，可将一个简单的，不管包含或不包含有影响的观察值都适用的步骤用于回归分析。这个方法虽然费时间，但仍能得出观察值对分析结果的影响。

练　习

方　法

50. 考虑下列关于两个变量和的数据：

自测题

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. 计算这些数据的标准误差。你是否发现数据中存在任何异常值？试解释。

b. 作出关于$\hat{y}$的标准残差图，这个残差图显示出任何异常值吗？

c. 画出这些数据的散点图，散点图表明在这些数据中有任何异常值吗？通常，对于简单线性回归，异常值表示什么？

51. 考虑下面关于两个变量 x 和 y 的数据：

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

a. 计算这些数据的标准残差，这些数据中存在任何异常值吗？试解释。

b. 计算这些数据的杠杆率，在这些数据中是否有任何有影响的观察值？试解释。

c. 建立这些数据的散点图，此散点图中存在有影响的观察值吗？试解释。

应　用

52. 下面是 10 个主要啤酒品牌的广告费用(百万美元)和销售量(百万桶)的数据(*superbrands' 98*, October 20, 1997)：

自测题

CD 光盘数据
Beer

啤酒品牌名称	广告费(百万美元)	销售量(百万桶)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9

（续表）

啤酒品牌名称	广告费(百万美元)	销售量(百万桶)
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

a. 根据这些数据建立估计回归方程。

b. 应用残差分析来判断是否存在任何异常值和(或)有影响的观察值，简要概括你的发现和结论。

53. 尼尔森媒体研究搜集收看一个特定广告的家庭户数数据。这些信息对广告客户非常有益，因为它使广告客户了解有多少消费者看过这个广告。下面数据是1997年11月24日至30日一周内收看该广告的家庭户数(百万户)和广告播出次数(*USA Today*, December 15, 1997)：



CD光盘数据
Nielsen2

登广告的品牌	广告播出次数	收看广告的家庭户数(百万户)
Sears	95	758.8
JC Penney	46	323.0
Burger King	41	275.3
Polaroid One Step Express	38	241.8
Wendy's	29	219.9
McDonald's	32	198.5
Target	25	193.8
Kmart	21	189.7
Visa	21	161.9
Nissan Frontier	16	160.0

a. 在广告播出次数已知时，求出用来预测收看该广告家庭户数的估计回归方程。

b. 利用残差分析来判断是否存在任何异常值和（或）有影响的观察值，简要总结你的发现和结论。

54. 下表列出20家公司的市值和执行总裁(CEO)的年薪的数值(*the Wall Street Journal*, February 24, 2000, and April 6, 2000)：



CD光盘数据
CEO

	市值(百万美元)	CEO年薪(千美元)
Anheuser-Busch	32 977.4	1 130
AT&T	162 365.1	1 400
Charles Schwab	31 363.8	800
Chevron	56 849.0	1 350
DuPont	68 848.0	1 000
General Elec.	507 216.8	3 325
Gillette	44 180.1	978
IBM	194 455.9	2 000
Johnson & Johnson	143 131.0	1 365
Kimberly-Clark	35 377.5	950
Merrill Lynch	31 062.1	700
Motorola	92 923.7	1 275
Philip Morris	54 421.2	1 625
Procter & Gamble	144 152.9	1 318.3

（续表）

	市值（百万美元）	CEO 年薪（千美元）
Qualcomm	116 840.8	773
Schering-Plough	62 259.4	1 200
Sun Microsystems	120 966.5	116
Texaco	30 040.7	950
USWest	36 450.8	897
Walt Disney	61 288.1	750

a. 给定公司市值，给出用来预测 CEO 年薪的估计回归方程。

b. 利用残差分析来判断是否存在任何异常值和/或有影响的观察值。简要概括一下你的发现和结论。

小　结

在本章中我们介绍了如何用回归分析来推断一个因变量 y 是怎样依赖一个自变量 x 的方法。对于简单线性回归情形，回归模型 $y=\beta_0+\beta_1 x+\varepsilon$ 的简单线性回归方程为 $E(y)=\beta_0+\beta_1 x$，该方程描述了 y 的平均值或期望值是如何依赖 x 的。我们利用样本数据和最小二乘法求出估计回归方程 $\hat{y}=b_0+b_1 x$。实际上，b_0 和 b_1 是用来估计模型中未知参数 β_0 和 β_1 的样本统计量。

为了估计回归方程拟合度，我们引入确定系数的概念。确定系数是因变量 y 中的变异能被估计回归方程所解释的部分所占比例。我们还回顾了相关系数的概念，它是描述两变量之间线性相关强度的一个度量。

在这一章里，我们讨论了回归模型和它的误差项 ε 的假设，并且根据这些假设，建立了 t 检验和 F 检验，这两个检验都是判断二变量间的关系在统计上是否显著的方法。我们介绍了如何利用估计回归方程来得出 y 平均值的置信区间估计和 y 个别值的预测区间估计的方法。

在本章结尾，用一节介绍了回归问题的计算机解法，用两节介绍了利用残差分析来检验模型假设的有效性并且鉴别异常值和有影响的观察值。

术语释义

因变量：用来预测或被解释的变量，用 y 表示。

自变量：进行预测或进行解释的变量，用 x 表示。

简单线性回归：涉及一个自变量和一个因变量，并且变量之间的关系用一条直线近似表示的回归分析。

回归模型：表达因变量 y 如何依赖自变量 x 和误差项的方程。对于简单线性回归情形，回归模型是 $y=\beta_0+\beta_1 x+\varepsilon$。

回归方程：表达因变量的平均值或期望值如何依赖自变量的方程。对于简单线性回归情形，回归方程是 $E(y)=\beta_0+\beta_1 x$。

估计回归方程：利用最小二乘法，根据样本数据求出回归方程的估计。对于简单线性情形，估计

回归方程是$\hat{y} = b_0 + b_1 x$。

散点图：用水平轴表示自变量，用纵轴表示因变量的二维数据图。

最小二乘法：用来求出估计回归方程的过程，目标是使$\sum(y_i - \hat{y}_i)^2$最小。

判定系数：估计回归方程拟合优度的度量，它能被理解为因变量 y 的变异性能被估计回归方程所解释的部分所占的比例。

第 i 个残差：因变量的观察值与利用估计回归方程得到的预测值之间的差异；即对于第 i 个观察值，残差为 $y_i - \hat{y}_i$。

相关系数：两变量之间线性关系强弱的量度(在前面第 3 章中已讨论过)。

均方误差：误差项的方差 σ^2 的无偏估计量，用 MSE 或 s^2 表示。

估计量的标准误差：误差均方的平方根，用 s 表示，它是误差项 ε 的标准差 σ 的估计。

ANOVA 表：(方差分析表)经常用于概括与显著性 F 检验相联系的计算。

置信区间估计：对于一个给定的 x 值，y 的平均值的区间估计。

预测区间估计：对于一个给定的 x 值，y 的一个个别值的区间估计。

残差分析：用于查明关于回归模型所做的假定是否成立的分析方法，残差分析还被用于识别异常值和有影响的观察值。

残差图：用图表示残差，它是用于推断关于回归模型所做的假定是否成立的一种图示方法。

标准残差：残差除以它的标准差后得到的数值。

正态概率图：为标准残差值作出的正态分数图，正态概率图能帮助判断关于误差项是否服从正态分布的假定。

异常值：一个数据点或观察值，它与其余的数据点所显示的趋势不相吻合。

有影响的观察值：对回归结果有显著影响或重要作用的观察值。

高杠杆率点：含有自变量极端值的观察值。

重要公式

简单线性回归模型

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{14.1}$$

简单线性回归方程

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

估计简单线性回归方程

$$\hat{y} = b_0 + b_1 x \tag{14.3}$$

最小二乘准则

$$\min \sum(y_i - \hat{y}_i)^2 \tag{14.5}$$

估计回归方程的斜率和 y 截距

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{14.7}$$

误差平方和

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 \tag{14.8}$$

总的平方和

$$\text{SST} = \sum (y_i - \bar{y})^2 \tag{14.9}$$

回归平方和

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \tag{14.10}$$

SST、SSR 和 SSE 之间的关系

$$\text{SST} = \text{SSR} + \text{SSE} \tag{14.11}$$

判定系数

$$r^2 = \frac{\text{SSR}}{\text{SST}} \tag{14.12}$$

样本相关系数

$$r_{xy} = (b_1 \text{ 的计算式}) \sqrt{\text{判定系数}} \tag{14.13}$$

$$= (b_1 \text{ 的计算式}) \sqrt{r^2}$$

均方误差（σ^2 的估计量）

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \tag{14.15}$$

估计量的标准误差

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \tag{14.16}$$

b_1 的标准差

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{14.17}$$

b_1 的估计标准差

$$s_{b_1}=\frac{s}{\sqrt{\sum(x_i-\bar{x})^2}} \tag{14.18}$$

t 检验统计量

$$t=\frac{b_1}{s_{b_1}} \tag{14.19}$$

均方回归

$$\text{MSR}=\frac{\text{SSR}}{\text{自变量的个数}} \tag{14.20}$$

F 检验统计量

$$F=\frac{\text{MSR}}{\text{MSE}} \tag{14.21}$$

$\hat{y}_p$ 的估计的标准差

$$s_{\hat{y}_p}=s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \tag{14.23}$$

$E(\hat{y}_p)$ 的置信区间估计

$$\hat{y}_p\pm t_{\alpha/2}s_{\hat{y}_p} \tag{14.24}$$

个别值的估计的标准差

$$s_{\text{ind}}=s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \tag{14.26}$$

y_p 值的预测区间估计

$$\hat{y}_p\pm t_{\alpha/2}s_{\text{ind}} \tag{14.27}$$

第 i 个观察值的残差

$$y_i-\hat{y}_i \tag{14.28}$$

第 i 个残差的标准差

$$s_{y_i-\hat{y}_i}=s\sqrt{1-h_i} \tag{14.30}$$

第 i 个观察值的标准残差

$$\frac{y_i-\hat{y}_i}{s_{y_i-\hat{y}_i}} \tag{14.32}$$

第 i 个观察值的杠杆率

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \qquad (14.33)$$

补充练习

55. 一个较大值的 r^2 意味着两变量之间存在因果关系吗？试解释。

56. 请用你自己的话，对于一个给定的 x，对 y 的平均值的区间估计和 y 的一个个别值的区间估计之间的区别作出解释。

57. 检验 $\beta_1 = 0$ 是否成立的目的是什么？如果拒绝 $\beta_1 = 0$，这是否意味着一个很好的拟合？

58. 对 10 只选取出来的公开发售的原始股票，下表是它们的销售数量(百万)和期望价格(预计的低价格和预计的高价格的平均价格)的数据(*USA Today*, November17, 1997)：



CD 光盘数据
IPO

公司	销售数量	期望价格(美元)
American Physician	5.0	15
Apex Silver Mines	9.0	14
Dan River	6.7	15
Franchise Mortgage	8.75	17
Gene Logic	3.0	11
International Home Foods	13.6	19
PRT Group	4.6	13
Rayovac	6.7	14
RealNetworks	3.0	10
Software AG Systems	7.7	13

a. 以股票销售数量为自变量，期望价格作因变量，给出估计回归方程。
b. 在显著性水平为 0.05 下，这两个变量之间存在一个显著关系吗？
c. 估计回归方程对数据是否给出较好的拟合？试解释。
d. 利用估计回归方程对公开发售 600 万原始股的一个公司，估计该公司股票的期望价格。

59. 公司股票回购方案常被认为能保护股东的利益。但是，First Call/ Thomson 金融机构内部研究经理 Robert Gabele 发现股票回购方案只是作为公司高层管理人员期权股票的取得方法。在所有公司中，期权股票在 1998 年已存在了，占所有公开买卖普通股票的 6.2%。下面是 13 家公司期权股票所占的数量和公开买卖的股票数量(*Bloomberg Personal Finance*, January/February , 2000)：



CD 光盘数据
Options

	能公开买卖的期权股票(百万)	公开买卖的普通股票(百万)
Adobe Systems	20.3	61.8
Apple Computer	52.7	160.9
Applied Materials	109.1	375.4
Autodesk	15.7	58.9
Best Buy	44.2	203.8
Fruit of the Loom	14.2	66.9
ITT Industries	18.0	87.9
Merrill Lynch	89.9	365.5
Novell	120.2	335.0
Parametric Technology	78.3	269.3
Reebok International	12.8	56.1
Silicon Graphics	52.6	188.8
Toys R Us	54.8	247.6

a. 对于给定的普通股票数量，求出估计能公开买卖的期权股票数量的估计回归方程。

b. 假定一家公司有 15 000 万公开买卖的普通股票，利用估计回归方程估计能公开买卖的期权股票数量。

c. 你是否相信估计回归方程能很好地预测公开买卖的期权股票数量？用 r^2 来支持你的结论。

60. *Value Line* (February 24, 1995)报告 Woolworth 公司的市场 β 值。个别股票的市场 β 值能由简单线性回归确定。对于每一种股票，因变量是它的季度收益百分比(资本增值扣上股息)减去无风险投资的收益百分比(国债收益率为无风险贴现率)。自变量是证券市场(标准普尔 500)的季度收益百分比(资本增值加上股息)减无风险投资的收益百分比。利用季度数据求出估计回归方程。市场 β 值为估计回归方程(b_1)。市场 β 是股票风险的变量，市场 β 大于 1 表明这种股票比市场一般水平不稳定得多；如果市场 β 小于 1 表明股票比市场一般水平稳定得多。下表显示了 10 个季度的标准普尔 500 和 IBM 的收益百分比与无风险百分比之差：

标准普尔 500	IBM
1.2	-0.7
-2.5	-2.0
-3.0	-5.5
2.0	4.7
5.0	1.8
1.2	4.1
3.0	2.6
-1.0	2.0
0.5	-1.3
2.5	5.5

a. 求出能确定 IBM 的市场 β 值的估计回归方程。IBM 的市场 β 值是多少？

b. 在显著性水平为 0.05 下，进行显著性关系检验。

c. 估计的回归方程对观察数据给出了一个好的拟合吗？试解释。

d. 利用 Woolworth 和 IBM 的市场 β 值，比较这两种股票的风险。

61. 下面是 20 个城市的每天最高和最低温度的数据（*USA Today*, May 9, 2000）：



CD 光盘数据
HighLow

	最低	最高
雅典	54	75
曼谷	74	92
开罗	57	84
哥本哈根	39	64
都柏林	46	64
哈瓦那	68	86
香港	72	81
约翰内斯堡	50	61
伦敦	48	73
马尼拉	75	93
墨尔本	50	66
蒙特利尔	52	64
巴黎	55	77
里约热内卢	61	80
罗马	54	81
汉城	50	64
新加坡	75	90
悉尼	55	68
东京	59	79
温哥华	43	57

a. 以最低温度为横坐标，最高温度为纵坐标，画出散点图。

b. 利用（a）中的散点图来表示这两个变量之间的关系？

c. 对于给定的最低温度，求出估计最高温度的估计回归方程。

d. 在显著性水平为 0.05 下，进行显著性关系检验。

e. 估计回归方程提供了很好的拟合吗？试解释。

f. 求出样本相关系数。

62. PJH&D 公司正在决定是否为新的文字处理系统签订一份维修合同。管理者认为维修费用与使用次数有关，收集每周使用次数（小时）和年维修费用（百美元）的统计资料如下：

使用次数(小时)	年维修费用(百美元)
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

a. 求出表达年维修费用和每周使用次数的估计回归方程。

b. 在显著性水平为 0.05 下，对(a)中的关系进行显著性检验。

c. PJH&D 期望每个星期文字处理系统使用时间为每周 30 个小时。求出该公司的年维修费用的 95% 预测区间。

d. 如果维修费用合同的费用为每年 3 000 美元，你是否建议签订这份合同？阐述理由。

63. 在生产过程中，装配线的速度(英尺每分钟)认为是检验过程中次品数目的主要影响因素。为了验证这个结论，管理者设计了一种情形，在该情形下每批产品在不同的装配速度下接受检验，下表是收集的数据：

装配线装配速度	被发现的次品数量
20	21
20	19
40	15
30	16
60	14
40	17

a. 求出表达装配速度和被发现次品数量关系的估计回归方程。

b. 在显著性水平为 0.05 下，判断装配速度和发现次品数量之间是否相关？

c. 估计回归方程能为这些数据提供很好的拟合吗？

d. 如果装配速度为每分钟 50 英尺，给出次品个数的置信度为 95% 的置信区间。

64. 某一大城市医院聘请一名社会学家来调查每年员工无故缺工的天数与职工从家里到工作单位距离(英里)之间的关系。由 10 名员工组成一个样本，收集的数据如下：

到工作单位的距离	无故缺工的天数
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a. 画出这些数据的散点图。图中是否显现出线性关系？试解释。

b. 利用最小二乘法求出估计回归方程。

c. 这两个变量之间是否有显著关系？取 $\alpha = 0.05$。

d. 是否估计回归方程提供很好的拟合？试解释。

e. 对于住在离公司 5 英里的员工，利用 (b) 中的估计回归方程，求出该员工每年无故缺工的预期天数的 95% 置信区间估计。

65. 对于一个人口较多的地区，交通部门想了解公共汽车的使用时间和年维修费用之间是否存在

某种关系。由 10 辆公共汽车组成一个样本，收集的数据如下：

公共汽车使用时间(年)	维修费用(美元)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a. 利用最小二乘法求出估计回归方程。

b. 在显著性水平为 0.05 下，检验这两个变量是否显著相关？

c. 最小二乘回归线是否给出了观察数据一个好的拟合？试解释。

d. 如果有 1 辆公共汽车使用了 4 年，得出这辆公共汽车年维修费用的 95% 预测区间。

66. Givens 大学的一名市场营销学教授对学生用在学习上的时间和所取得的学习成绩之间的关系感兴趣。收集了 10 名学生最后一学期的学习成绩如下：

用在学习上的时间	学习成绩
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a. 求出表达学生取得的学习成绩是如何依赖用在学习上的时间的估计回归方程。

b. 在 $\alpha = 0.05$ 的显著性水平下，检验该模型的显著性。

c. 如果 Mark Sweeney 用在学习上的时间为 95 个小时，预测他的成绩。

d. 求出 Mark Sweeney 学习成绩总分数的 95% 预测区间。

67. 下面是由 Syracuse 大学的 Transactional Records Access Clearinghouse 记录的数据，该数据为国内税务局进行审计的比例。下表的数据是由国内税务局随机抽取的 20 个地区调整后的平均总收入和经审计的收入的比例(*the Wall Street Journal Almanac 1998*)：


CD光盘数据
IRSAudit

地区	调整后的总收入(美元)	经审计的总收入的比例(%)
洛杉矶	36 664	1.3
萨克拉门托	38 845	1.1
亚特兰大	34 886	1.1
博伊西	32 512	1.1
达拉斯	34 531	1.0
普罗维登斯	35 995	1.0
圣何塞	37 799	0.9
夏延	33 876	0.9
法戈	30 513	0.9
新奥尔良	30 174	0.9
俄克拉何马城	30 060	0.8
休斯敦	37 153	0.8
波特兰	34 918	0.7
菲尼克斯	33 291	0.7
奥古斯塔	31 504	0.7
阿尔伯克基	29 199	0.6
格林斯伯勒	33 072	0.6
哥伦比亚	30 859	0.5
纳什维尔	32 566	0.5
布法罗	34 296	0.5

a. 对于给定的调整后的收入，求出能用来预测经审计的收入比例的估计回归方程。

b. 在显著性水平为0.05下，判断调整后的总收入和经审计的收入比例是否相关？

c. 估计的回归方程是否给出很好的拟合？试解释。

d. 假定平均的调整后的总收入为35 000美元，利用(a)中的回归方程来计算该地区经审计的期望比例的95%置信区间估计？

案例研究1　教育支出和学生成绩

学习成绩与各州对教育的支出之间有多大的关系？在许多州，纳税人经常被州立学校要求提高在教育上的税收支出。因此需要通过分析支出和学生成绩的数据来判断在州立学校的学生成绩和教育支出之间是否有任何关系？

美国联邦政府的“国家教育进展评估”(National Assessment of Educational Progress, NAEP)方案经常对学生成绩进行评估。表14.13给出了对于参加NAEP方案的35个州，其每年花在每个学生身上总的费用和NAEP测试综合分数的统计数据，这些数据存储在文件NAEP中。综合分数是指数学、自然科学和阅读三门课程1996(1994年是阅读课)的NAEP考试的总和。对8年级的学生进行测试，对4年级学生进行阅读测试，满分为1 300。表14.14列出了未参加NAEP调查的

14 个州对每名学生每年的经常性教育支出。这些数据登载在《福布斯》(*Forbes*, November 3, 1997)的一篇关于教育支出和学生学习成绩水平的论文中。

表 14.13　参加 NAEP 方案的州每名学生的教育经费支出和综合分数


CD 光盘数据
NAEP

州名	教育经费支出(美元每名学生)	综合分数
路易斯安那	4 049	581
密西西比	3 423	582
加利福尼亚	4 917	580
夏威夷	5 532	580
南卡罗来纳	4 304	603
阿拉巴马	3 777	604
佐治亚	4 663	611
佛罗里达	4 934	611
新墨西哥	4 097	614
阿肯色	4 060	615
特拉华	6 208	615
田纳西	3 800	618
亚利桑那	4 041	618
西弗吉尼亚	5 247	625
马里兰	6 100	625
肯塔基	5 020	626
得克萨斯	4 520	627
纽约	8 162	628
北卡罗来纳	4 521	629
罗德岛	6 554	638
华盛顿	5 338	639
密苏里	4 483	641
科罗拉多	4 772	644
印第安纳	5 128	649
犹他	3 280	650
怀俄明	5 515	657
康涅狄格	7 629	657
马萨诸塞	6 413	658
内布拉斯加	5 410	660
明尼苏达	5 477	661
衣阿华	5 060	665
蒙大拿	4 985	667
威斯康辛	6 055	667
北达科他	4 374	671
缅因	5 561	675

管理报告

1. 对这些数据作出数值和图形的概述。
2. 利用回归分析来判断每个学生的教育经费支出和 NAEP 测试综合分数之间的关系,讨论你

的结论。

3. 根据这些数据求出估计回归方程，试问能否用该方程来估计未参加 NAEP 方案的州的学生分数吗？
4. 假定你只考虑每名学生的教育经费支出在 4 000 美元至 6 000 美元之间的州。对于这些州，两变量之间的关系与根据 35 个州的全部数据所得出的结论显现出任何不同吗？讨论你的结论。如果将教育经费支出少于 4 000 美元或者多于 6 000 美元的州删除，你认为这种删除是否合理？
5. 对未参加 NAEP 方案的州，估计其学生综合分数。
6. 根据上面的分析，你认为学生的教育水平与州教育经费支出之间有多少关系？

表 14.14　未参加 NAEP 方案的班每名学生的教育支出

州名	每名学生的支出(美元)
爱达荷	3 602
南达科他	4 067
俄克拉何马	4 265
内华达	4 658
堪萨斯	5 164
伊利诺伊	5 297
新罕布什尔	5 387
俄亥俄	5 438
俄勒冈	5 588
佛蒙特	6 269
密歇根	6 391
宾夕法尼亚	6 579
阿拉斯加	7 890

案例研究 2　美国交通局

为了研究交通安全，美国交通局收集了 42 个城市中每 1 000 个驾驶员中发生死亡事故的车祸次数和有驾驶执照的司机中 21 岁以下所占比例的数据。下面是一年内收集的数据，这些数据存储在光盘中的文件 Safety 中。


CD 光盘数据
Safety

21 岁以下所占比例(%)	每 1 000 个驾驶司机中发生车祸次数	21 岁以上所占比例(%)	每 1 000 个驾驶司机中发生车祸次数
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

管理报告

1. 对这些数据作出数值和图示的概述。
2. 利用回归分析去判断发生死亡事故的车祸次数和司机中 21 岁以下所占比例之间的关系。试讨论你的结论。
3. 从你的分析中，你能得出什么结论或得出什么建议吗?

案例研究 3　校友捐赠

校友捐赠现在成为大学重要的收入来源。如果管理者希望判断哪种因素能提高校友捐赠比例，他们可能相应采取措施来提高校友捐赠。研究表明那些对教师更满意的学生更可能顺利毕业。因此，有人怀疑班级越小，学生—全体教职员工比率越低，将会导致很高的顺利毕业率，从而提高校友捐赠的比例。表 14.15 列出了 48 家大学的数据(*America's Best Colleges*, 2000 年教育卷)。列小于

20 的班级的百分比表示为小于 20 名学生的班级百分比。列学生—全体教职员工比表示为学生总人数除以总教职员工数的比率。最后，列捐赠率表示为捐赠大学的人数比例。

表 14.15　48 个州立大学的数据


CD 光盘数据
Alumni

	小于 20 的班级的百分比	学生—全体教职员工比	捐赠率
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve Univ.	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State Univ.	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California – Berkeley	58	17	18
U. of California – Davis	32	19	7
U. of California – Irvine	42	20	9
U. of California – Los Angeles	41	18	13
U. of California – San Diego	48	19	8
U. of California – Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois – Urbana Champaign	29	15	23
U. of Michigan – Ann Arbor	51	15	13
U. of North Carolina – Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22

（续表）

	小于 20 的班级的百分比	学生—全体教职员工比	捐赠率
U. of Texas – Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin – Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University – St. Louis	73	7	33
Yale University	77	7	50

管理报告

1. 对这些数据作出数值和图示概述。
2. 对于给定的小于 20 人的班级的比例，利用回归分析求出能预测捐赠比例的估计回归方程。
3. 给于给定的学生比率，利用回归分析求出能预测捐赠比例的估计的回归方程。
4. 这两个回归方程哪个提供了更好的拟合程度？得出这个估计的回归方程，进行残差分析，并讨论你的结论。
5. 从你的分析中，得出你的结论和/或建议。

附录 14.1　最小二乘法的推导

正如本章所介绍，最小二乘法是通过确定 b_0 和 b_1 的值，而使残差平方和最小化的方法。计算残差平方和的公式如下：

$$\sum (y_i - \hat{y}_i)^2$$

代入 $\hat{y}_i = b_0 + b_1 x_i$，得

$$\sum (y_i - b_0 - b_1 x_i)^2 \tag{14.34}$$

求该表达式的极小值。

为了求公式(14.34)的极小值，应对 b_0 和 b_1 求偏导数，使其等于零。这样我们得到

$$\frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 \tag{14.35}$$

$$\frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\sum x_i (y_i - b_0 - b_1 x_i) = 0 \tag{14.36}$$

将公式(14.35)除以 2，并分别对每一项求和，得到

$$-\sum y_i + \sum b_0 + \sum b_1 x_i = 0$$

将 $\sum y_i$ 移到等号的右边并注意 $\sum b_0 = nb_0$，得到

$$nb_0 + (\sum x_i)b_1 = \sum y_i \tag{14.37}$$

对公式(14.36)作类似的代数化简为

$$(\sum x_i)b_0 + (\sum x_i^2)b_1 = \sum x_i y_i \tag{14.38}$$

公式(14.37)和(14.38)是**正态方程组**。从公式(14.37)中求出 b_0 为

$$b_0 = \frac{\sum y_i}{n} - b_1\frac{\sum x_i}{n} \tag{14.39}$$

将(14.39)代入公式(14.38)中的 b_0 得

$$\frac{\sum x_i \sum y_i}{n} - \frac{(\sum x_i)^2}{n}b_1 + (\sum x_i^2)b_1 = \sum x_i y_i \tag{14.40}$$

整理公式(14.40)，得

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{14.41}$$

因为 $\bar{y} = \sum y_i/n$ 并且 $\bar{x} = \sum x_i/n$，公式(14.39)可写为

$$b_0 = \bar{y} - b_1\bar{x} \tag{14.42}$$

公式(14.41)和公式(14.42)就是我们在本章中用来计算估计回归方程系数的公式(14.6)和公式(14.7)。

附录 14.2 相关系数的显著性检验

利用样本相关系数 r_{xy}，我们检验关于总体相关系数的假设

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

来判断 x 和 y 之间是否存在一个显著的线性关系。如果被 H_0 拒绝，则得出总体相关系数 ρ_{xy} 不为零，两个变量之间的线性关系显著的结论。显著性检验如下：

利用相关系数进行显著性检验

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

检验统计量

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \tag{14.43}$$

拒绝规则

如果 $t < -t_{\alpha/2}$ 或 $t > t_{\alpha/2}$，则拒绝 H_0。

式中，$t_{\alpha/2}$ 是基于自由度为 $n-2$ 的 t 分布。

在 14.4 节中，我们已经求出一个容量为 $n=10$ 的学生人数和季度销售收入的样本相关系数 $r_{xy}=0.9501$。检验的统计量是

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} = 0.9501\sqrt{\frac{10-2}{1-(0.9501)^2}} = 8.61$$

从附录 B 的表 2 中，查得 $\alpha=0.01$，自由度 t 为 $n-2=10-2=8$ 的分布的双侧分位数为 $t_{0.005}=3.355$。因为 $8.61>3.355$，故拒绝 H_0，得出在显著性水平 0.01 下，总体相关系数 ρ_{xy} 显著不为零。从而，这一结论提供了充分的统计证据表明学生人数和季度销售收入之间存在一个显著的线性关系。

注意，利用统计检验量得出的显著关系的结论与 14.5 节中利用 Armand 的例子估计回归方程 $\hat{y}=60+5x$ 进行检验得出的结论一样。回归分析得出和之间存在显著关系的结论，并且得到一个说明变量之间的关系的方程。由于大多数分析家是利用现代的计算机软件包去完成回归分析，因此利用相关系数进行显著性检验是没必要的。

附录 14.3　用 Minitab 进行回归分析



CD 光盘数据
Armand's

在 14.7 节中我们讨论了回归分析的计算机解法，并以 Armand 比萨饼连锁店问题的 Minitab 输出结果为例。在这个附录中，我们介绍了 Minitab 计算机解所需要的步骤。首先，将数据输入 Minitab 的工作表。学生人数的数据被输入到 C1 列，季度销售收入的数据被输入到 C2。输入变量名 Pop 和 Sales 作为列的标题。这样，我们就可以根据所使用的变量名 Pop 和 Sales 或者列 C1 和 C2 去查询数据。以下步骤介绍了如何使用 Minitab 进行回归分析，见图 14.10。

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Regression** 下拉菜单；
步骤 3. 选择 **Regression**；
步骤 4. 当 Regression 对话框出现后：
在 **Response** 框中输入 Sales；
在 **Predictors** 框中输入 Pop；
点击 **OK**。

Minitab 回归对话框可以通过选择可选按钮获得想得到的结果。例如，为了得到以因变量$\hat{y}$的预测值为横坐标，标准差为纵坐标的残差图，步骤 4 为：

步骤 4. 当 Regression 对话框出现后：
在 **Response** 框中输入 Sales；
在 **Predictors** 框中输入 Pop；
点击 **Graphs** 按钮；
当 **Regression-Graphs** 对话框出现后：
在 **Residuals for Plots** 下选择 **Standardized**；
在 **Residual Plots** 下选择 **Residuals versus fits**；
点击 **OK**；
当 **Regression** 对话框出现后：
点击 **OK**。

附录 14.4　用 Excel 进行回归分析

在本附录中我们介绍了如何用 Excel 回归工具对 Armand 比萨饼连锁店案例进行回归分析。图 14.23 给出了其步骤。将 Restaurant，Population，Sales 键入工作表的 A1: C1。为了区别这 10 个观察值，我们将数字 1: 10 输入单元格 A2: A11。样本数据输入单元格 B2 到 C11。下面步骤是描述如何用 Excel 进行回归分析：

步骤 1. 选择下拉菜单 **Tools**；
步骤 2. 选择 **Data Analysis**；
步骤 3. 从 **Analysis Tools** 中选择 **Regression**；
步骤 4. 点击 **OK**；
步骤 5. 当 Regression 对话框出现后：
在 **Input Y Range** 框中键入 C1: C11；
在 **Input X Range** 框中键入 B1: B11；
选择 **Labels**；
选择 **Confidence Level**；
在 **Confidence Level** 框中键入 99；
选择 **Output Range**；
在 **Output Range** 框中键入 A13
（此时可输入任一个作为输出开始的左上角单元格的序号）；
点击 **OK**。

输出结果的第一部分为 **Regression Statistics**，包含确定系数(R^2)等汇总信息。输出结果的第二部分为 ANOVA 即方差表分析。输出结果的最后一部分没有标题，是估计回归系数和相关信息。我

们进行回归分析输出的信息在单元格 A28: I30。

Restaurant	Population	Sales
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.9501
R Square	0.9027
Adjusted R Square	0.8906
Standard Error	13.8293
Observations	10

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14200	14200	74.2484	2.55E-05
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569
Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470

图 14.23　**Armand 比萨饼连锁店问题的 Excel 扩展表方法**

回归方程输出结果的说明

单元格 B29 为估计回归线 y 的截距 $b_0=60$，单元格 B30 为估计回归线的斜率 $b_1=5$。为了识别这两个变量，单元格 A29 标为 Intercept，单元格 A30 标为 Population。

在 14.5 节中我们得出 b_1 的估计标准差为 $s_{b_1}=0.580\,3$，注意单元格 C30 值为 0.580 3。C28 单元格的标签为 Standard Error，表明 C30 单元格的值为标准差或标准离差 b_1。回顾利用 t 检验显著性关系要求计算 t 统计量即为 $t=b_1/s_{b_1}$。对于 Armand 数据，计算 t 值为 $t=5/0.580\,3=8.62$。D28 单元格的标签为 *t Stat*，表明单元格 D30 包含 t 检验统计量。

在 14.5 节中我们介绍了在显著性水平 $\alpha=0.01$，自由度为 $n-2=10-2=8$ 时，$t_{0.005}=3.355$。因为 $t=8.62>3.355$，得出在显著性水平为 0.01 下，β_1 不等于 0。也就是说，有充分的统计证据来得出学生人数和销售额之间存在显著关系。单元格 E30 为与显著性检验相联系的 p-值。在单元格 E30 中 Excel 提供了其科学记数符号。为了获得十进制表示方法，我们将小数点左移 5 位，得到值为 0.000 025 5。因为 p-值 $=0.000\,025\,5<\alpha=0.01$，拒绝 H_0 则得出学生人数和季度销售额之间存在显著关系。

单元格 F28: I30 给出了 y 的截距和估计回归方程斜率的置信区间估计，另外 Excel 给出了置信度为 95% 的上限和下限。在前面的第 4 步中我们选择了 Confidence Level，并在 Confidence Level

框中键入 99。最后通过 Excel 回归工具提供了置信度为 99% 的置信区间的上下限。单元格 H30 是的置信度为 99% 的区间估计为 3.05 到 6.95。单元格 F30 和 G30 给出了 95% 置信区间的上限和下限。因此，95% 置信区间为 3.66 到 6.34。

ANOVA 输出结果的说明

单元格 A22: F26 是方差计算分析的汇总。方差的三个来源为 Regression，Residual 和 Total。单元格 B23 的 *df* 代表自由度，单元格 C23 标签为 *SS* 表示平方和，单元格 D23 标签为 *MS*，表示均方。

在 14.5 节中我们利用平方误差或残差和除以自由度得出平方差均值，其值为的一个估计量 σ^2。单元格 D25 中的值为 191.25，表示 Armand 回归输出的平方差均值，在 14.5 节中我们介绍了 *F* 检验来进行回归方程的显著性检验。单元格 F24 值为 0.000 025 5，表示显著性 *F* 检验相联系的 p - 值。因为 p - 值 = 0.000 025 5< α = 0.01，则我们拒绝 H_0，得出学生人数和季度销售额之间有显著关系。标签 Excel 是用来识别显著性 *F* 检验的 p - 值，见单元格 F23 即 *Significance F*。

如果认为单元格 F24 为 *F* 检验的观察值，则标签 Significance *F* 有更多的意义。

回归统计输出结果的说明

单元格 B17 为确定系数，值为 0.902 7，其相应的标签见 A17 为 R Square。单元格 B16 为确定系数的平方根，相应的样本相关系数为 0.950 1，注意 Excel 用标签 Multiple R(单元格 A16)来识别此值。在单元格 A19 中用标签 Standard Error 来识别单元格 B19 中的估计标准差的值，估计标准差的大小为 13.829 3。注意在 Excel 输出中，标签 Standard Error 出现在两个不同的位置，在回归统计的输出中，标签 Standard Error 表示 σ。在估计回归方程中，标签 Standard Error 表示样本分布 s_{b_1}——b_1 的标准离差。

15 多元回归

统计实例：CHAMPION 国际有限公司

15.1 多元回归模型

回归模型和回归方程

多元估计回归方程

15.2 最小二乘法

示例：Butler 运输公司

回归系数解释

15.3 多元判定系数

15.4 模型假设

15.5 显著性检验

F 检验

t 检验

多重共线性

15.6 利用估计回归方程进行估计和预测

15.7 定性自变量

示例：Johnson 过滤水公司

参数解释

更为复杂的定性变量

15.8 残差分析

检测异常值

按学生分布删除残差和异常值

有影响的观测值

利用库克距离测度识别有影响的观测值

统计实例

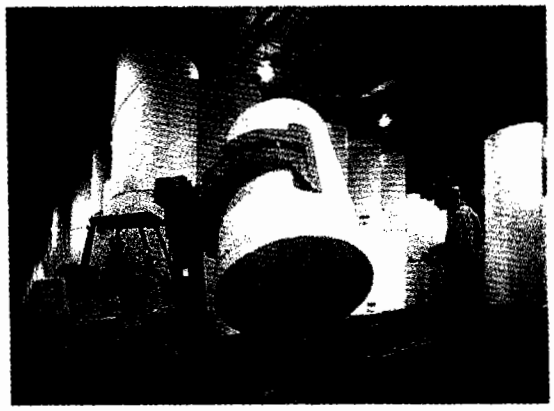
CHAMPION 国际有限公司*

斯坦福，康涅狄格州

Champion 国际有限公司是世界上最大的林业产品公司，它在美国拥有三百多万英亩的林业资源。该公司生产木料和复合板等建筑材料，还有诸如印刷和书写用的白纸制品，同时还生产比如做衬板和波纹容器用的棕色纸制品。为生产上述纸制品，Champion 的纸浆工厂将木屑和化学产品进行处理产出木质纸浆，然后利用这些纸浆在造纸厂里面生产出纸制品。

在白纸制品的生产过程中，必须对纸浆进行漂白去污处理，在此处理过程中，关键性的漂白剂是次氯酸，因为它具有易燃性，所以通常在 Champion 的纸浆工厂设备里制造出来，然后再通过管道输送到纸浆厂的漂白塔中。为了改进 Champion 生产次氯酸的一个工艺环节，公司对工艺的控制过程和效率进行了一项研究，其中的一方面就是研究次氯酸的生产过程中化学制剂的添加比例。

次氯酸的生产过程如下：将四种化学添加剂以一定的比例注入次氯酸发生器，该发生器产生的次氯酸气体流入一个吸收器，在吸收器里，冷却水吸收了次氯酸气体形成次氯酸溶液。然后，溶液被输送到造纸厂车间。这一控制过程的关键在于化学添加剂的比例，传统的做法是由富有经验的操作员来确定，然而这种做法会导致由于操作员失误而引起的失控。因此，工厂里的化学工程师需要开发一套控制程式，每种化学制剂均对应一个程式，以使操作员来确定比例。



利用多元回归分析，Champion 国际公司为他们的纸制品开发了更为先进的漂白工艺。© Lester Lefkowitz/The Stock Market.

利用多元回归分析，Champion 的统计分析师可以为生产过程中所使用的化学制剂各建立一个多元回归分析估值方程。每一个程式均将次氯酸的生产与化学添加剂的数量和浓度联系起来，这四个程式的结果被编入每个工厂的计算机上。在新系统中，操作员只需将次氯酸溶液的浓度和预计的生产率输入，计算机软件就会计算出要获得这一产出的化学添加剂量。由于操作员已经使用这一控制程式，次氯酸发生器的生产效率提高了，同时，溶液的浓度取值于允许范围内的次数也显著增加。

Champion 公司应用了多元回归分析来建立了它的控制程式。在本章中，我们将讨论诸如 Minitab 之类的统计计算软件包是怎样应用于上述问题的，第 14 章中介绍的大部分简单线性回归概念都可以扩展到多元线性回归中来。

* 作者感谢 Champion 国际有限公司的 Marian Williams 和 Bill Griggs 提供这一统计实例。

在第 14 章中，我们论述了简单线性回归问题以及它在建立描述两个变量之间关系的方程中的应用。让我们回忆一下，由回归方程预测或者解释的变量称之为因变量，用来预测或者解释因变量的变量称为自变量。在本章中，我们通过分析牵涉到两个或者更多自变量的情形来继续对回归分析进行研究，该主题领域即我们所说的**多元回归分析**(multiple regression analysis)。它使得我们可以考虑更多的因素，因而可以得到比简单线性回归更好的估计值。

15.1　多元回归模型

多元回归分析研究的是因变量 y 怎样依赖于两个或者更多的自变量的。在通常的情况下，我们将使用 p 来表示自变量个数。

回归模型和回归方程

我们在上一章中介绍的回归模型和回归方程的概念在本章多元回归中也是适用的。描述因变量 y 是怎样依赖于自变量 x_1，x_2，…，x_p 和误差项的方程称为**多元回归模型**(multiple regression model)。我们先假设多元回归模型有以下形式：

多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{15.1}$$

在多元回归模型中，β_0，β_1，β_2，…，β_p 是参数，ε（希腊字母 epsilon）是随机变量。仔细观察这个模型就会发现，y 是 x_1，x_2，…，x_p（$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 部分）的线性函数加上误差部分 ε。误差部分说明了 y 里面所不能被 p 个自变量之间的线性关系解释的差异。

在 15.4 节中我们将论述多元回归模型和 ε 的假设。其中的一个假设就是 ε 的均值或期望值为零。这一假设的推论是：y 的平均值或期望值，以 $E(y)$ 表示，等于 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$。描述 y 的平均值是如何依赖于 $x_1, x_2, \cdots, x_p$ 的方程即称为**多元回归方程**(multiple regression equation)。

多元回归方程

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.2}$$

多元估计回归方程

如果 β_0，β_1，$\beta_2\cdots$，β_p 值已知，且给定了 x_1，x_2，…，x_p 的值，则公式(15.2)可以用于计算 y 的平均值。不幸的是，这些参数值通常是未知的，必须通过样本数据去估计。我们可以利用一个简单随机样本去计算样本统计值 b_0, b_1, b_2, …, b_p 作为 β_0，β_1，β_2，…，β_p 的点估计。这些样本统计值提供了下述的**多元估计回归方程**(estimated multiple regression equation)：

多元估计回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15.3}$$

式中，b_0，b_1，b_2，…，b_p 是 β_0，β_1，β_2，…，β_p 的估计值，$\hat{y}$ 是因变量的估计值。

多元估计回归的步骤如图 15.1 所示。

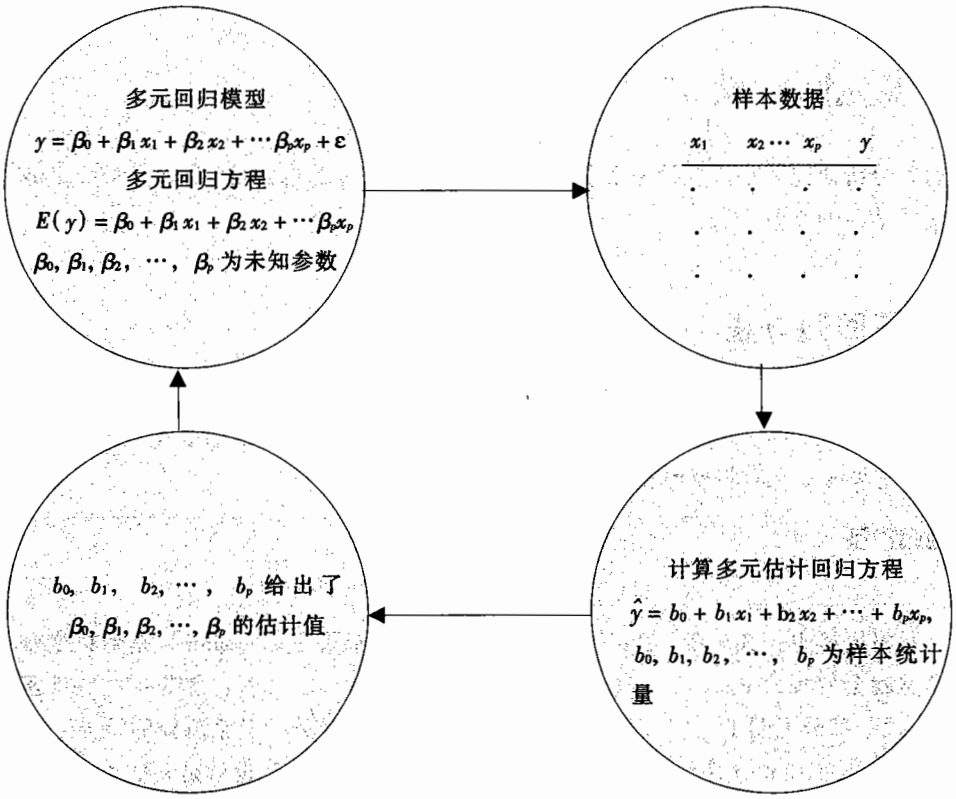


图 15.1 多元估计回归的步骤

在简单线性回归情况下，b_0 和 b_1 是用于估计未知参数 β_0，β_1 的样本统计量；同理，在多元回归情形下，b_0，b_1，b_2，…，b_p 是用于估计未知参数 β_0，β_1，β_2，…，β_p 的样本统计量。

15.2 最小二乘法

在第 14 章，我们利用**最小二乘法**(least squares method)建立了估计回归方程，该方程以最大可能的接近程度描述了因变量和自变量之间的线性关系。同样，我们可以利用该方法去建立多元

估计回归方程。最小二乘法的准则如下所述：

最小二乘法

$$\min \sum (y_i - \hat{y}_i)^2 \tag{15.4}$$

式中　y_i——因变量的第 i 次观测值；

$\hat{y}_i$——因变量的第 i 次估计值。

因变量的估计值可以利用多元估计回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

来计算。

如公式(15.4)所示，最小二乘法利用样本数据，求得残差(基因变量的观测值 y_i 与因变量 $\hat{y}_i$ 的估计值之间的离差)的最小平方和，从而得出 b_0, b_1, b_2, …, b_p 的值。

在第 14 章里，为了估计简单线性回归方程 $\hat{y} = b_0 + b_1 x$，我们介绍了计算最小二乘估计值 b_0 和 b_1 的公式，对于比较小的数集，我们通过手算就能利用这些公式计算出 b_0 和 b_1 的数值。但是，在多元回归中计算回归系数 b_0, b_1, b_2, …, b_p 的公式涉及到矩阵知识，这已经超出本书论述的范围，因此，在介绍多元回归时，我们将集中讨论计算机软件包是怎样被用来得到估计回归方程和其他信息的，重点将放在计算机输出结果的解释上而不是怎么进行多元回归的计算。

示例：Butler 运输公司

作为多元回归分析的一个示例，我们可以参考 Butler 运输公司所面临的一个问题，该公司是一个位于南加利福尼亚的独立运输公司。Butler 公司的主营业务地域为本地，为了建立更好的工作日期表，经理们计划为他们的驾驶员估计日常行驶时间。

一开始经理们认为日常驾驶时间应该与日常运输的里程数紧密相关。10 份行驶记录的简单随机样本提供了表 15.1 的数据和图 15.2 的散点图。经过对散点图的分析，经理们假设：$y = \beta_0 + \beta_1 x_1 + \varepsilon$ 可以被用于描述总驾驶时间和里程数之间的关系，为了估计参数 β_0 和 β_1，我们应用最小二乘法来建立估计的回归方程：

$$\hat{y} = b_0 + b_1 x_1 \tag{15.5}$$

在图 15.3 中，我们展示了对表 15.1 中的数据进行简单线性回归分析的 Minitab 计算机输出结果，回归估计方程如下：

$$\hat{y} = 1.27 + 0.067\,8 x_1$$

表 15.1 **Butler 运输公司初始数据**



CD 光盘数据
Butler

运送次数	x_1 = 行驶里程	y = 行驶时间(小时)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

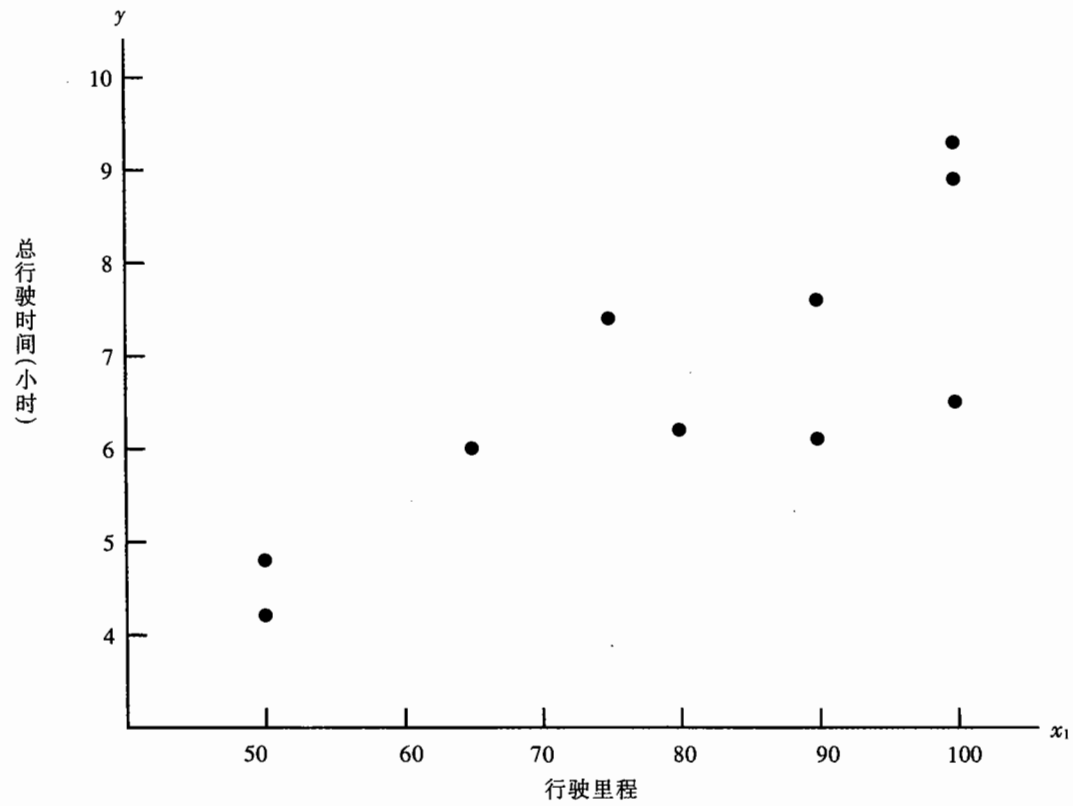


图 15.2 Butler 运输公司数据预处理的散点图

在 Minitab 输出结果中，变量名称 Miles(英里)和 Time(时间)作为工作单上的表头输入，那么，x_1 = 英里，x_2 = 运送次数，y = 时间。

```
The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor       Coef      Stdev    t-ratio       p
Constant       1.274      1.401       0.91   0.390
Miles        0.06783    0.01706       3.98   0.004

s = 1.002       R-sq = 66.4%     R-sq(adj) = 62.2%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      15.871      15.871      15.81   0.004
Error         8       8.029       1.004
Total         9      23.900
```

图 15.3　一个自变量的 Butler 运输公司的 Minitab 输出结果

在 Minitab 输出结果中，变量名称英里和时间作为每列的名称，于是 x_1 = 英里，x_2 = 时间。

在 0.05 的显著性水平下，F 统计量值为 15.81，它所对应的 p - 值为 0.004，这表明自变量与因变量之间的关系是显著的。这也就是说，因为 p - 值比 $\alpha = 0.05$ 要小，所以我们可以拒绝假设 H_0: $\beta_1 = 0$。注意，通过 $t = 3.98$ 以及与它相关的 p - 值为 0.004，我们也可以得出这一结论。由此可知，日常行驶时间与行驶里程显著相关，行驶时间越长，行驶里程也就越长，由于判定系数(以百分数表示)R - sq = 66.4%，我们可以看到行驶时间 66.4% 的变异可以由行驶里程的线性影响来解释。这一结果相当不错，但是管理者或许会考虑增加第二个自变量来解释因变量剩余的变异。

在确定另一个自变量时，管理者认为行驶时间也受到运送货物的次数的影响，表 15.2 显示了 Butler 公司添加了运输次数以后的数据，图 15.4 显示的是以行驶里程(x_1)和运送次数(x_2)作为自

表 15.2　以行驶里程和运送次数为自变量的 Butler 运输公司数据



CD 光盘数据
Butler

运送次数	x_1 = 行驶里程	x_2 = 运送次数	y = 行驶时间(小时)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

```
The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliv

Predictor       Coef        Stdev    t-ratio        p
Constant     -0.8687       0.9515      -0.91    0.392
Miles         0.061135     0.009888     6.18    0.000
Deliv         0.9234       0.2211       4.18    0.004

s = 0.5731      R-sq = 90.4%     R-sq(adj) = 87.6%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    2      21.601      10.800      32.88    0.000
Error         7       2.299       0.328
Total         9      23.900
```

图 15.4 设置两个自变量的 Butler 运输公司 Minitab 计算机输出结果

变量的 Minitab 计算机解决方案，估计的回归方程如下：

$$\hat{y} = -0.869 + 0.061\,1x_1 + 0.923x_2 \tag{15.6}$$

在下一章中，我们将讨论利用多元判定系数来判断估计回归方程是否提供了一个好的拟合，在此之前，让我们更仔细的检测一下公式(15.6)中的 $b_1 = 0.061\,1$ 和 $b_2 = 0.923$ 这两个值。

在 Minitab 输出结果中，变量名称 Miles(英里)、Deliv(运送)和 Time(时间)作为工作单上的表头输入，那么，x_1 = 英里，x_2 = 运送次数，y = 行驶时间。

回归系数解释

我们的检测应该是着眼于单自变量的估计回归方程和把运送次数作为第二自变量的方程之间的关系来进行，在这两种情形下，b_1 的值是不同的。简单线性回归情形下，b_1 被认为是当自变量改变一个单位时因变量改变的数值，多元回归情形下，这一解释必须做些修改，也就是说，多元回归情形下，我们对每一个回归系数解释如下：当所有其他自变量均保持不变时，b_i 是因变量 y 对应于自变量 x_i 改变一个单位时所作的改变的估计值。在包含两个自变量的 Butler 公司的事例中，$b_1 = 0.061\,1$。因此，当运送货物的次数保持不变时，行驶里程每增加一英里，行驶时间所增加的估计值是 0.061 1 小时，与此类似，由于 $b_2 = 0.923$，所以当运送货物行驶里程保持不变时，运输次数每增加一次时，与此相对应的行驶时间期望值的估计增加值为 0.923 小时。

练　习

请学生们注意，涉及本节和后面章节数据的练习均要求利用计算机软件包来完成。

方　法

1. 如下所示的估计回归方程模型包含两个自变量和 10 个观测值：

$$\hat{y} = 29.127\,0 + 0.590\,6x_1 + 0.498\,0x_2$$

a. 解释该估计回归方程中的 b_1 和 b_2。

b. 当 $x_1 = 180$ 和 $x_2 = 310$ 时，估计 y 值。

2. 对于因变量 y 和两个自变量 x_1 和 x_2，有如下数据：


自测题


CD 光盘数据
Exer2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

a. 利用这些数据，建立 y 关于 x_1 的估计回归方程，如果 $x_1 = 45$，求 y 的估计值。

b. 利用这些数据，建立 y 关于 x_2 的估计回归方程，如果 $x_2 = 15$，求 y 的估计值。

c. 利用这些数据，建立 y 关于 x_1 和 x_2 的估计回归方程，如果 $x_1 = 45$，$x_2 = 15$，求 y 的估计值。

3. 该回归分析包含 30 个观测值，有如下的估计方程：

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

a. 解释该方程中的 b_1，b_2，b_3，b_4。

b. 当 $x_1 = 10$，$x_2 = 5$，$x_3 = 1$，$x_4 = 2$ 时估计 y 值。

应 用

4. 一家制鞋厂建立了销售额和库存投入及广告投入的估计回归方程：

$$\hat{y} = 25 + 10x_1 + 8x_2$$

式中 x_1 = 库存投入（千美元）；
x_2 = 广告投入（千美元）；
y = 销售额（千美元）。

a. 当库存投入 x_1 为 15 000 美元以及广告投入为 10 000 美元时，估计销售额。

b. 解释该估计回归方程中的 b_1，b_2。

5. Showtime Movie Theaters 公司的老板打算将每周总营业收入作为广告支出的一个函数来估计，8 周的历史数据组成如下的样本：

自测题

CD 光盘数据 Showtime

每周总收入 千美元	电视广告费 千美元	报刊广告费 千美元
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. 将电视广告费作为自变量，建立估计回归方程。

b. 将电视广告费和报刊广告费作为自变量，建立估计回归方程。

c. 在(a)和(b)中建立的估计回归方程中，电视广告费的系数相同吗？解释每种情况的系数。

d. 如果电视广告费为 3 500 美元，报刊广告费为 1 800 美元，试估计该周总收入。

6. 下表显示了 16 种赛车和 GT 车的马力、自重、启动加速至 1/4 英里时的速度（*1998 Road & Track Sports & GT Cars*）：

CD 光盘数据 Auto1

赛车和 GT 车	自重(磅)	马力	启动加速至 1/4 英里时的速度(英里/小时)
Acura Integra Type R	2 577	195	90.7
Acura NSX-T	3 066	290	108.0
BMW Z3 2.8	2 844	189	93.2
Chevrolet Camaro Z28	3 439	305	103.2
Chevrolet Corvette Convertible	3 246	345	102.1

（续表）

赛车和 GT 车	自重(磅)	马力	启动加速至 1/4 英里时的速度(英里/小时)
Dodge Viper RT/10	3 319	450	116.2
Ford Mustang GT	3 227	225	91.7
Honda Prelude Type SH	3 042	195	89.7
Mercedes-Benz CLK320	3 240	215	93.0
Mercedes-Benz SLK230	3 025	185	92.3
Mitsubishi 3000GT VR-4	3 737	320	99.0
Nissan 240SX SE	2 862	155	84.6
Pontiac Firebird Trans Am	3 455	305	103.2
Porsche Boxster	2 822	201	93.2
Toyota Supra Turbo	3 505	320	105.0
Volvo C70	3 285	236	97.0

a. 以自重作为自变量，启动加速至 1/4 英里时的速度(Speed at 1/4 mile)作为因变量，求估计回归方程。

b. 以自重和马力作为自变量，启动加速至 1/4 英里时的速度作为因变量，求估计回归方程。

c. 1999 Porsche 911 Carrera 款式的赛车在广告中宣称其自重 2 910 磅，引擎马力 296，利用(b)中的结果预测该车启动加速至 1/4 英里时的速度。

7. Heller 公司生产割草机和相关草坪设备，管理者认为割草机销量取决于割草机的价格和竞争对手的价格，令



CD 光盘数据
Mower

y = 销量(千台)

x_1 = 竞争对手割草机价格(美元)

x_2 = Heller 公司割草机价格(美元)

Heller 公司的管理人员希望建立割草机的销售数量与 Heller 公司割草机价格和竞争对手割草机价格的估计回归方程，下表列出的是 10 个城市的价格资料：

竞争对手的价格 (x_1)	Heller 的价格 (x_2)	销售数量 (y)
120	100	102
140	110	100
190	90	120
130	150	77
155	210	46
175	150	93
125	250	26
145	270	69
180	300	65
150	250	85

a. 在竞争厂商价格和 Heller 公司的价格已知的情况下，建立估计回归方程预测 Heller 公司割草机的销售数量。

b. 解释该方程中的 b_1 和 b_2。

c. 如果在某个城市，Heller 公司的割草机价格为 160 美元，竞争对手的割草机价格为 170 美元，试预测 Heller 公司的销售数量。

8. 下表给出了 20 个国外基金的年回报率、安全系数(0 表示最危险，10 表示最安全)和年支出率(*Mutual Funds*, March 2000)：


CD 光盘数据 ForFunds

	安全系数	年支出率(%)	年回报率(%)
Accessor Int'l Equity "Adv"	7.1	1.59	49
Aetna "I" International	7.2	1.35	52
Amer Century Int'l Discovery "Inv"	6.8	1.68	89
Columbia International Stock	7.1	1.56	58
Concert Inv "A" Int'l Equity	6.2	2.16	131
Dreyfus Fourders Int'l Equity "F"	7.4	1.80	59
Driehaus International Growth	6.5	1.88	99
Excelsior "Inst" Int'l Equity	7.0	0.90	53
Julius Baer International Equity	6.9	1.79	77
Marshall International Stock "Y"	7.2	1.49	54
MassMutual Int'l Equity "S"	7.1	1.05	57
Morgan Grenfell Int'l Sm Cap "Inst"	7.7	1.25	61
New England "A" Int'l Equity	7.0	1.83	88
Pilgrim Int'l Small Cap "A"	7.0	1.94	122
Republic International Equity	7.2	1.09	71
Sit International Growth	6.9	1.50	51
Smith Barney "A" Int'l Equity	7.0	1.28	60
State St Research "S" Int'l Equity	7.1	1.65	50
Strong International Stock	6.5	1.61	93
Vontobel International Equity	7.0	1.50	47

a. 利用这些数据，以年回报率为因变量，年支出率和安全系数为自变量，建立一个估计回归方程。

b. 某公司安全系数为 7.5，年支出率为 2，试估计其年回报率。

9. 两个专家提供了他们认为是美国最好的学区名单，资料包括每一个学区的班级平均学生数，综合 SAT 分数，考入四年制大学的学生比例。


CD 光盘数据 Schools

学区	班级平均学生数	综合 SAT 分数	进入四年制大学百分比(%)
Blue Springs, MO	25	1 083	74
Garden City, NY	18	997	77
Indianapolis, IN	30	716	40
Newport Beach, CA	26	977	51
Novi, MI	20	980	53
Piedmont, CA	28	1 042	75
Pittsburg, PA	21	983	66
Scarsdale, NY	20	1 110	87

（续表）

学区	班级平均学生数	综合 SAT 分数	进入四年制大学百分比(%)
Wayne, PA	22	1 040	85
Weston, MA	21	1 031	89
Farmingdale, NY	22	947	81
Mamaroneck, NY	20	1 000	69
Mayfield, OH	24	1 003	48
Morristown, NJ	22	972	64
New Rochelle, NY	23	1 039	55
Newtown Square, PA	17	963	79
Omaha, NE	23	1 059	81
Shaker Heights, OH	23	940	82

a. 利用这些数据，以班级平均学生数和综合 SAT 分数为自变量，以进入四年制大学百分比为因变量，建立估计回归方程。

b. 如果班级平均学生数为 20，综合 SAT 分数为 1 000，试估计进入四年制大学的学生百分比。

10. 汽车租赁新闻报提供了如下数据：15 家汽车租赁公司的运营汽车数量(千辆)，运营场地数，租金收入(百万美元)(*The Wall Street Journal Almanac* 1998)。



CD 光盘数据
CarRent

公司	汽车数	场地	收入
Alamo	130	171	1 180
Avis	190	1 130	1 500
Budget	126	1 052	1 500
Dollar	63.5	450	560
Enterprise	315.1	2 636	2 060
FRCS(Ford)	55.25	1 784	312.5
Hertz	250	1 200	2 400
National	135	935	1 200
Payless	15	100	47
PROP(Chrysler)	27	1 500	160
Rent-A-Wreck	10.9	460	78
Snappy	15.5	259	85
Thrifty	34	480	340
U-Save	13.5	500	95
Value	18	45	150.1

a. 建立估计回归方程，在运营车辆数已知的情况下，预测租金收入。

b. 解释(a)中估计回归方程的斜率。

c. 建立估计回归方程，在运营车辆数和场地数已知的情况下可以用来预测租金收入。

15.3 多元判定系数

在简单线性回归情况下，我们已经说明总平方和可以被分成两部分：回归平方和、误差平方和。同样，多元回归平方和也适用该情形。

SST、SSR 和 SSE 之间的关系

$$SST = SSR + SSE \tag{15.7}$$

式中 SST——总平方和 $= \sum (y_i - \bar{y})^2$；

SSR——回归平方和 $= \sum (\hat{y}_i - \bar{y})^2$；

SSE——误差平方和 $= \sum (y_i - \hat{y}_i)^2$。

因为这三个平方和的计算难度很大，我们依靠计算机软件包来得到这些数值。对于两个自变量的 Butler 运输公司案例，图 15.4 的 Minitab 计算机输出结果里的方差分析部分显示了它的这三个数值，SST = 23.900, SSR = 21.601, SSE = 2.299。当只有行驶里程数一个自变量时，图 15.3 的 Minitab 计算机输出结果显示了 SST = 23.900, SSR = 15.871, SSE = 8.029。SST 值在两种情况下是一样的，因为它并不依赖于 $\hat{y}$。但是，在增加第二个自变量（运送次数）时，SSR 增加了且 SSE 减少了。这表明估计多元回归方程对所观测的数据给出了一个更好的拟合。

在第 14 章中，我们使用了判定系数 $r^2 = SSR/SST$ 来测度估计回归方程的拟合度，同样的概念适用于多元回归情形。术语**多元判定系数**(multiple coefficient of determination)代表着我们对估计多元回归方程拟合优度的测度，多元判定系数以 R^2 表示，其计算公式如下：

多元判定系数

$$R^2 = \frac{SSR}{SST} \tag{15.8}$$

增加自变量导致预测误差值变小，因而也减小了误差平方和 SSE。由于 SSR = SST - SSE, 当 SSE 变小时，SSR 就变大，因此 $R^2 = SSR/SST$ 变大。

多元判定系数可理解为因变量 y 的变异程度，y 能被估计多元回归方程所解释，因此，将此变异程度乘以 100，即可以理解为因变量 y 中的变异性能被估计多元回归方程解释的百分比。

对于带两个自变量的 Bulter 运输公司的例子，其中 SSR = 21.601, SST = 23.900，可得

$$R^2 = \frac{21.601}{23.900} = 0.904$$

因此，行驶时间 y 中 90.4% 的变异性可以被行驶里程数和运送次数作自变量的估计多元回归方程解释，在图 15.4 中，我们看到多元回归系数由 Minitab 的输出结果给出，表示为：R - sq =

90.4%。

图 15.3 显示了在仅有一个自变量——行驶里程 x_1 的估计多元回归方程里面，R-sq 的值为 66.4%，因此，当运送货物的次数作为第二个自变量被添加时，行驶时间 y 中的变异性能被估计多元回归方程解释的百分比由 66.4% 增加到 90.4%。通常说来，R^2 的值总是随着新的变量进入模型而增加的。

因为增加自变量的个数将影响到估计回归方程所解释的变异性，为了避免高估这种影响，很多统计分析师采用自变量的个数来修正 R^2 的值。观测值的数目用 n 来表示，自变量的数目用 p 来表示，**修正多元判定系数**(adjusted multiple coefficient of determination)的计算公式如下：

修正多元判定系数

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{15.9}$$

如果模型增加一个自变量，即使所增加的自变量在统计上并不显著，R^2 也会变大。修正多元判定系数补偿了模型中自变量个数的影响。

对于 Butler 运输公司的例子，$n = 10$, $p = 2$, 我们可以得到

$$R_a^2 = 1 - (1 - 0.904)\frac{10-1}{10-2-1} = 0.88$$

于是，在对这两个自变量进行修正后，我们得到修正后的多元判定系数 0.88。这个数值在图 15.4 中由 Minitab 输出结果表示为 R-sq(adj) = 87.6%，这和我们所计算的数值不一样，那是因为在计算中对 R^2 作了四舍五入的处理。

评　注

如果 R^2 的值比较小，且模型包含的自变量数目较多，修正多元判定系数会呈负值，在这种情况下，Minitab 将把修正多元判定系数的值取为 0。

练　习

方　法

11. 在练习 1 中，根据 10 次观测值有如下的估计回归方程：

$$\hat{y} = 29.127\,0 + 0.590\,6x_1 + 0.498\,0x_2$$

式中，SST = 6 724. 125, SSR = 6 216. 375。

a. 计算 SSE。　　b. 计算 R^2。　　c. 计算R_a^2。　　d. 评论拟合度。

12. 在练习 2 中，对因变量 y 和两个自变量 x_1 及 x_2 有 10 个观测值，根据这些数据，我们计算出 SST = 15 182. 9，SSR = 14 052. 2。



a. 计算 R^2。　　b. 计算R_a^2。

c. 估计回归方程是否解释了数据中的大部分变异性，请解释。

13. 在练习 3 中，根据 30 个观测值有如下的估计回归方程：

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

式中，SST = 1 805, SSR = 1 760。

a. 计算 R^2。　　b. 计算R_a^2。　　c. 评论拟合度。

应　用

14. 在练习 4 中给出了销售收入关于存货支出与广告支出的估计回归方程：

$$\hat{y} = 25 + 10x_1 + 8x_2$$

用来建立模型的数据来自于对 10 家鞋店的调查，式中，SST = 16 000, SSR = 12 000。

a. 在给定的估计回归方程条件下，计算 R^2。

b. 计算R_a^2。

c. 估计回归方程是否解释了数据中的大部分变异性，请解释。

15. 在练习 5 中，Showtime Movie Theaters 公司的老板利用多元回归分析来预测总收入（y），自变量是电视广告费用(x_1)和报刊广告费用(x_2)，估计回归方程如下：



$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

计算机输出结果给定：SST = 25. 5, SSR = 23. 435。

a. 计算并解释 R^2，R_a^2。

b. 当只有电视广告费用这一个自变量时，$R^2 = 0.653$，$R_a^2 = 0.595$，你是否愿意接受该多元回归结果？为什么？

16. 在练习 6 中，给出了 16 种型号的赛车和 GT 车的自重、马力和启动加速至 1/4 英里时的速度（*1998 Road & Track Sports & GT Cars*）。

a. 如果估计多元回归方程仅有自重这一个自变量，用它来预测启动加速至 1/4 英里时的速度，这样做是否提供了一个优良的拟合度？为什么？

b. 论述利用自重和马力这两个自变量来预测启动加速至 1/4 英里时的速度的好处。

17. 在练习 9 中，我们建立了估计多元回归方程，利用班级平均学生数和综合 SAT 成绩来预测进

入四年制大学的学生比例。

a. 计算并解释 R^2，R_a^2。

b. 估计回归方程是否提供了一个优良拟合？请解释。

18. 参考练习 10，它提供了 15 家汽车出租公司投入运营的车辆数(千辆)，营业场地的数量，租金收入(百万美元)等资料(The Wall Street Journal Almanac 1998)。

a. 在练习 10 的(c)中，我们已经建立了出租收入关于投入运营的汽车数量以及营业场地的数量的估计多元回归方程，试求其中的 R^2，R_a^2。

b. 估计回归方程是否对数据提供了一个优良拟合？请解释。

15.4　模型假设

在 15.1 节中，我们引入了以下形式的多元回归模型：

多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \qquad (15.10)$$

在多元回归模型中的误差项 ε 的假定类似于简单线性回归模型中的假定。

多元回归模型 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ 中误差项 ε 的假定

1. 误差项 ε 是一个随机变量，其均值或者期望值为 0，即 $E(\varepsilon) = 0$。

   这意味着：对于 $x_1, x_2, \cdots, x_p$ 的给定值，y 的期望值或均值取决于：

   $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.11)$$

   公式(15.11)即我们在 15.1 节中介绍的多元回归方程式，在这个方程式中，$E(y)$表示对于给定的 $x_1, x_2, \cdots, x_p$ 值，所有可能出现的 y 的期望值或者平均值。

2. 对所有的自变量值 $x_1, x_2, \cdots, x_p$，ε 的方差，以 σ^2 表示，均保持不变。

   这意味着：对 $x_1, x_2, \cdots, x_p$ 的所有值，y 的方差均等于 σ^2。

3. ε 的值是相互独立的。

   这意味着：自变量的一组特定值所对应的误差项与自变量的任意一组其他值所对应的误差项是不相关的。

4. 误差项 ε 是一个呈正态分布的随机变量，它反映了 y 值和由 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 所决定的 y 的期望值之间的离差。

   这意味着：对于给定的 $x_1, x_2, \cdots, x_p$ 值，由于 β_0，β_1，$\cdots$，β_p 均为常量，所以因变量 y 也呈正态分布。

为了对公式(15.11)式所给出的关系形式有更多的了解，我们可以考虑分析一下下面带有两个自变量的多元回归方程：

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

该方程的图形是三维空间的一个平面，如图 15.5 所示，注意，当 $x_1 = x_1^*$，$x_2 = x_2^*$ 时，ε 是 y 的实际值和期望值 $E(y)$之间的差。

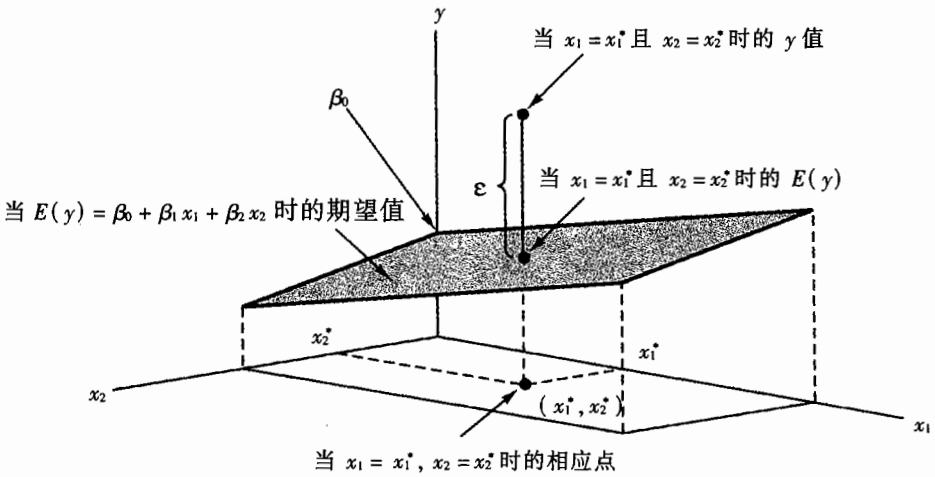


图 15.5　带两个自变量的多元回归方程的图形

在回归分析中，我们经常用术语**响应变量**来替代术语**因变量**，除此之外，由于多元回归方程所生成的图形是一个平面或者曲面，因此它的图形被称为**响应曲面**。

15.5　显著性检验

在本节我们将阐述如何对多元回归关系进行显著性检验。在简单线性回归情形中，我们使用的是 t 检验和 F 检验，对于简单线性回归情形，两种检验所提供的结果是相同的，即：如果零假设遭到拒绝，则我们认为 $\beta_1 \neq 0$。在多元回归情形中，t 检验和 F 检验的目的是不同的。

1. F 检验用于确定因变量和所有自变量之间是否存在一种显著性的关系，我们倾向于把 F 检验称为**总体的显著性检验**。
2. 如果 F 检验显示了总体的显著相关，则 t 检验用于确定每一个单独的自变量是否显著，模型中的每一个单独的自变量均进行 t 检验，我们倾向于将每一个这样的 t 检验称为**单独显著性检验**。

在接下来的内容中，我们将解释 F 检验和 t 检验，并将它们应用于 Butler 运输公司的示例。

F 检验

15.4 节中所定义的多元回归模型是：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

F 检验假设与多元回归模型中的参数是相关的：

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_a：参数中至少有一个不等于 0

如果 H_0 假设被拒绝，我们有充分的统计证据表明至少有一个参数不为 0，而且，因变量 y 和自变量系列 $x_1, x_2, \cdots, x_p$ 之间的总体关系是显著的。相反，如果 H_0 假设没有被拒绝，我们将缺乏足够的证据认为它们之间存在一个显著的关系。

在描述 F 检验步骤之前，我们有必要回顾一下**均方**的概念，均方是平方之和除以它所对应的自由度。在多元回归情形中，总平方和有 $n-1$ 个自由度，回归平方和(SSR)有 p 个自由度，误差平方和有 $n-p-1$ 个自由度，因此，回归所产生的均方(MSR)是 SSR/p，误差所产生的均方(MSE)是 SSE/$(n-p-1)$。

$$\text{MSR} = \frac{\text{SSR}}{p} \tag{15.12}$$

和

$$\text{MSE} = \frac{\text{SSE}}{n-p-1} \tag{15.13}$$

在第 14 章中讨论过，MSE 提供了误差项 ε 的方差 σ^2 的无偏估计量，如果假设 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 成立，MSR 也会提供 σ^2 的一个无偏估计量，并且，MSR/MSE 的值将会接近 1。相反，如果假设 H_0 被拒绝，MSR 将会过高估计 σ^2，此时 MSR/MSE 的值将会变大。为了确定 MSR/MSE 的值在多大的时候会拒绝 H_0，我们可以利用这样一个事实：如果 H_0 成立，并且多元回归模型的假定均有效，那么 MSR/MSE 的抽样分布是一个 F 分布，其分子自由度为 p，分母自由度为 $n-p-1$。对于多元回归情形，F 检验的主要步骤如下：

总体显著性 F 检验

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_a：至少有一个参数不等于 0

检验统计量

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{15.14}$$

拒绝规则

利用检验统计量：如果 $F > F_\alpha$，则拒绝 H_0；

利用 p-值：如果 p-值$< \alpha$，则拒绝 H_0。

式中，F_α 是基于分子自由度为 p、分母自由度为 $n-p-1$ 的 F 分布。

让我们将 F 检验应用于 Butler 运输公司的多元回归问题中来，它带有两个自变量，假设形式如下：

$$H_0: \beta_1 = \beta_2 = 0$$

H_a: β_1，β_2 至少有一个不等于 0

图 15.6 是以行驶里程(x_1)和运送次数(x_2)作为自变量的多元回归模型的 Minitab 计算机输出结果，在输出结果的方差分析部分，我们可以看到 MSR = 10.8, MSE = .328。利用公式(15.4)，我们可以得到检验统计量

$$F = \frac{10.8}{0.328} = 32.9$$

注意：Minitab 输出结果的 F 值为 32.88，这与我们所计算出的数值不同，原因在于我们在计算过程中使用了 MSR 和 MSE 的四舍五入近似值。在显著性水平 $\alpha = 0.01$ 下，从附录 B 表 4 中，我们可以查到分子自由度为 2，分母自由度为 7 的上侧分位数 $F_{0.01} = 9.55$。根据 32.9 > 9.55，我们可以拒绝 H_0: $\beta_1 = \beta_2 = 0$, 并且我们可以得出结论：在行驶时间 y 和两个自变量：行驶里程和运送次数之间存在显著相关关系。在方差分析表（图 15.6）的最后一列里，p - 值为 0.000，因为 p - 值比 $\alpha = 0.01$ 值要小，所以我们也可以根据这一点拒绝 H_0: $\beta_1 = \beta_2 = 0$。

正如前面所提到的，误差均方 MSE 提供了的方差 σ^2 的一个无偏估计量。参照图 15.6，我们可以看到 σ^2 的估计量是 MSE = 0.328。MSE 的平方根是误差项的标准差的估计。正如 14.5 节中

```
The regression equation is
Time = -0.869 + 0.0611 Miles + 0.923 Deliv

Predictor       Coef      Stdev    t-ratio       p
Constant     -0.8687     0.9515      -0.91   0.392
Miles       0.061135   0.009888       6.18   0.000
Deliv         0.9234     0.2211       4.18   0.004

s = 0.5731      R-sq = 90.4%     R-sq(adj) = 87.6%

Analysis of Variance

SOURCE        DF        SS        MS        F       p
Regression     2    21.601    10.800    32.88   0.000
Error          7     2.299     0.328
Total          9    23.900
```

图 15.6 带两个自变量：Butler 运输公司的行驶里程(x_1)和运送次数(x_2)的 Minitab 计算机输出结果

所定义的那样，这个标准差被称为估计量的标准误差，以 s 表示。因此，我们可以得到：$s = \sqrt{\text{MSE}} = \sqrt{0.328} = 0.573$。请注意，估计量的标准误差值出现在图 15.6 中的 Minitab 计算机输出结果中。

表 15.3 是一般性的误差分析表(ANOVA)，它提供了多元回归模型的 F 检验结果。检验统计

量 F 在表的最后一列，我们可以将它和分子自由度为 p，分母自由度为 $n-p-1$ 的 F_α 作个比较，从而得出假设检验的结果。通过回顾图 15.6 中的 Butler 运输公司的 Minitab 计算机输出结果，我们可以看到 Minitab 的方差分析表里面已经包含了上述信息。除此之外，Minitab 还提供了相对于 F 检验统计量的 p - 值。

表 15.3　p 个自变量的多元回归模型的 ANOVA 表

方差来源	平方和	自由度	均方差	F
回归方程	SSR	p	$\text{MSR}=\dfrac{\text{SSR}}{p}$	$F=\dfrac{\text{MSR}}{\text{MSE}}$
误差	SSE	$n-p-1$	$\text{MSE}=\dfrac{\text{SSE}}{n-p-1}$	
合计	SST	$n-1$		

t 检验

如果 F 检验显示多元回归关系在总体上显著，那么 t 检验则可用于确定每一个单独参数的显著性。对单独参数显著性的 t 检验步骤如下：

单独参数显著性的 t 检验

对任意参数 β_i

$$H_0:\ \beta_i=0$$
$$H_a:\ \beta_i\neq 0$$

统计量

$$t=\frac{b_i}{s_{b_i}} \tag{15.15}$$

拒绝法则

利用检验统计量：如果 $t<-t_{\alpha/2}$ 或者 $t>t_{\alpha/2}$，则拒绝 H_0

利用 p - 值：如果 p - 值 $<\alpha$，则拒绝 H_0

式中，$t_{\alpha/2}$ 是基于自由度为 $n-p-1$ 的 t 分布的双侧分位数。

在统计检验中，s_{b_i} 是 b_i 标准差的估计，s_{b_i} 之值由计算机软件包提供。

让我们将 t 检验应用于 Butler 运输公司的回归示例。请参考图 15.6 中对 t - ratio 计算的 Minitab 输出结果，b_1，b_2 和 s_{b_1}，s_{b_2} 的值如下：

$$b_1=0.061\,135 \qquad s_{b_1}=0.009\,888$$
$$b_2=0.9234 \qquad s_{b_2}=0.2211$$

利用公式(15.15)，我们可以得到有关参数 β_1，β_2 的假设的检验统计量：

$$t = 0.061\,135 / 0.009\,888 = 6.18$$
$$t = 0.923\,4 / 0.221\,1 = 4.18$$

请注意，这两个 t 值均在图 15.6 中由 Minitab 计算机输出结果提供，利用显著性水平 $\alpha = 0.01$，自由度 $n - p - 1 = 10 - 2 - 1 = 7$，我们可以通过附录 B 表 2 中得出双侧分位数 $t_{0.005} = 3.499$。根据 $6.18 > 3.499$，我们可以拒绝 H_0: $\beta_1 = 0$。同样，由于 $4.18 > 3.499$，我们拒绝 H_0: $\beta_2 = 0$。请注意，在 Minitab 的计算机输出结果中，p - 值为 0.000 和 0.004，这同样表明可以拒绝这些假设。因此，这两个参数均在统计上呈显著性。

多重共线性

我们在回归分析中使用**自变量**这个术语来表示用来预测或解释因变量的任何变量，但是这个术语并不意味着自变量本身在统计意义上是独立的。恰恰相反，在多元回归问题中大部分自变量在一定程度上都是彼此相关的。例如，在带两个自变量 x_1(行驶里程)和 x_2(运送次数)的 Butler 运输公司的示例中，我们可以将行驶里程作为因变量，把运送货物次数作为自变量，以此来确定这两个变量本身是否相关。然后我们可以计算样本相关系数 $r_{x_1x_2}$ 去确定这两个变量的相关程度。如此运算得出结果 $r_{x_1x_2} = 0.28$，因此，我们发现这两个自变量之间存在线性相关关系。在多元回归分析中，**多重共线性**(multicollinearity)指的就是自变量之间的这种关系。

为了对多重共线性的潜在问题进行更好的分析，我们可以考虑对 Butler 运输公司的示例做一些简单的修改，原来的 x_2 表示货物运送的次数，这次我们用它来表示汽油消耗的加仑数。很显然，x_1(行驶里程)和 x_2 是相关的，这意味着我们认为汽油消耗的加仑数取决于行驶里程。因此，我们可以从逻辑上推断 x_1 和 x_2 是高度相关的自变量。

假设我们已经建立了方程 $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$，并且 F 检验显示总体关系是显著的，然后，我们对 β_1 进行 t 检验以确定 $\beta_1 \neq 0$ 是否成立，如果我们不能拒绝 H_0: $\beta_1 = 0$，这是否意味着行驶时间与行驶里程是不相关的?并不一定，它很可能意味着由于模型里有 x_2，所以 x_1 并不对 y 值的确定有显著的作用，这种解释在我们的示例里是可行的。如果我们知道了汽油的消耗加仑数，我们在预测 y 值的过程中就不需要通过知道行驶里程数来获得有用信息。同样，因为 x_1 已经存在于模型之中，我们也就没有必要再去增加汽油的消耗加仑数信息，以此为基础，我们通过进行 t 检验即能推断 $\beta_2 = 0$。

总的说来，在对单个参数的显著性进行 t 检验时，由多重共线性而引起的困难在于：当 F 检验显示多元回归方程总体呈显著性时，有可能得出结论，即没有一个单独参数是显著地不等于零。这个问题只有在自变量之间相关性很小时才能回避。

为了确定多重共线性是否达到足够的程度以至于对模型估计产生问题，统计学家们已经开发出几种检验方法。对于带两个自变量的情形，如果它们的样本相关系数的绝对值超过 0.7，根据经验检验方法的规律，多重共线性将有可能成为潜在问题。其他的检验方法更为先进，已经超出本教材所讨论的范围。

对于有两个自变量情形，当它们的样本相关系数大于 +0.70 或者小于 -0.70 时，将有可能产生多重共线性潜在影响。

如果可能的话，我们应该尽一切努力去回避包含高度相关的自变量。然而在实际中，我们很难严格坚持这一规则，决策者们必须意识到，当他们有理由相信严重的多重共线性关系存在时，将每一个单独自变量对因变量的影响区分开将是一件很困难的事。

当自变量密切相关时，我们将无法确定任一自变量对因变量的单独影响。

评　注

一般说来，多重共线性并不影响我们进行回归分析的方法，或者对从研究中解释结果亦没有影响。然而，当多重共线性相当显著时，即两个或者更多的自变量彼此高度相关时，我们在对单独参数 t 检验的结果进行解释时将遇到困难。除了本节所说明的问题之外，严重的多重共线性将导致最小二乘估计出现错误信息，那意味着，在模拟研究中，研究人员设计了回归模型然后利用最小二乘法求出参数 β_0, β_1, β_2 等的估计量。研究已经表明：在高度多重共线性情形下，最小二乘估计量可能与被估参数有完全相反的符号。例如，β_2 的值实际上可能为 +10，但它的估计值却可能为 -2。因此，如果存在高度的多重共线性，我们应密切关注单个系数的符号。

练　习

方　法

19.　在练习 1 中，根据 10 次观测结果所建立的估计回归方程如下所示：

自测题

$$\hat{y} = 29.127\,0 + 0.590\,6x_1 + 0.498\,0x_2$$

式中，SST = 6 724.125，SSR = 6 216.375，$s_{b_1} = 0.081\,3$，$s_{b_2} = 0.056\,7$。

a. 计算 MSR 和 MSE。

b. 计算 F 统计量，并在 $\alpha = 0.05$ 的显著性水平下进行 F 检验。

c. 在 $\alpha = 0.05$ 的显著性水平下，对 β_1 的显著性进行 t 检验。

d. 在 $\alpha=0.05$ 的显著性水平下，对 β_2 的显著性进行 t 检验。

20. 参考在练习 2 中给出的数据，估计回归方程如下：

$$\hat{y}=-18.4+2.01x_1+4.74x_2$$

式中，SST = 15 182.9, SSR = 14 052.2, $s_{b_1}=0.247\,1$, $s_{b_2}=0.948\,4$。

a. 在 $\alpha=0.05$ 的显著性水平下，检验 x_1, x_2, 和 y 之间的显著性关系。

b. 在 $\alpha=0.05$ 的显著性水平下，β_1 是显著的吗？

c. 在 $\alpha=0.05$ 的显著性水平下，β_2 是显著的吗？

21. 以下的估计回归方程含有两个自变量：

$$\hat{y}=40.7+8.63x_1+2.71x_2$$

将 x_2 从模型中排除，再利用最小二乘法处理可得到以 x_1 为自变量的估计回归方程：

$$\hat{y}=42.0+9.01x_1$$

a. 请解释两个模型中自变量 x_1 的系数。

b. 能否用多重共线性来解释两个模型中 x_1 的系数为什么不同？如果能，怎么解释？

应　用

22. 在练习 4 中给出了销售额关于库存投资和广告支出费用的估计回归方程：

$$\hat{y}=25+10x_1+8x_2$$

建立该模型的数据来自于对 10 家鞋店进行的调查，这些数据有：SST = 16 000, SSR = 12 000

a. 请计算 SSE, MSE 和 MSR。

b. 在 $\alpha=0.05$ 的显著性水平下，利用 F 检验确定变量之间是否存在显著关系。

23. 参考练习 5。


自测题

a. 在 $\alpha=0.01$ 的显著性水平下，检验假设：

H_0: $\beta_1=\beta_2=0$

H_a: β_1, β_2 至少有一个不等于零

有模型 $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$

式中　x_1——电视广告费用(千美元)；

x_2——报纸广告费用(千美元)。

b. 在 $\alpha=0.05$ 的显著性水平下检验 β_1 的显著性，是否应该从模型中删除 x_1?

c. 在 $\alpha=0.05$ 的显著性水平下检验 β_2 的显著性，是否应该从模型中删除 x_2?

24. 参考练习 6 中的数据，利用自重和马力两个自变量预测赛车和 GT 车从启动加速至 1/4 英里时的速度。

a. 利用 F 检验确定这个关系的总体显著性，在 $\alpha=0.05$ 的显著性水平下，你的结论是什么？

b. 利用 t 检验来确定每个自变量的显著性，在 $\alpha=0.05$ 的显著性水平下，你的结论是什么？

25. 从《证券投资数据库》中抽取 16 个公司作为样本，每一个公司的市盈率(P/E)、总边际利润、每一个公司的销售量的增长率如下(*Stock Investor Pro*, American Association of Individual Investors, August 21, 1997)：



CD 光盘数据
Stocks

	市盈率	总边际利润(%)	销售量增长率(%)
Abbott Laboratories	22.3	23.7	10.0
American Home Products	22.6	21.1	5.3
Amoco	16.7	11.0	16.5
Bristol Meyers Squibb Co.	25.9	26.6	9.4
Chevron	18.3	11.6	18.4
Exxon	18.7	9.8	8.3
General Electric Company	13.1	13.4	13.1
Hewlett-Packard	23.3	9.7	21.9
IBM	17.3	11.5	5.6
Merck & Co., Inc.	26.2	25.6	18.9
Mobil	18.7	8.2	8.1
Pfizer	34.6	25.1	12.8
Pharmacia & Upjohn. Inc.	22.3	15.0	2.7
Procter & Gamble Co.	5.4	14.9	5.4
Texaco	12.3	7.3	23.7
Travelers Group, Inc.	28.7	17.8	28.7

a. 建立估计回归方程，该方程在已知总的边际利润和销售增长率的情况下，可以用于预测市盈率。

b. 利用 F 检验确定变量关系的总体显著性，在 $\alpha=0.05$ 的显著性水平下，你的结论是什么？

c. 利用 t 检验确定变量关系的总体显著性，在 $\alpha=0.05$ 的显著性水平下，你的结论是什么？

d. 从估计回归方程中消除不显著的自变量，你推荐的估计回归方程是什么？计算 R^2 并将其与(a)中的 R^2 进行比较，讨论它们之间的区别。

26. 在练习 10 中，我们建立了关于租金收入和投入运营的汽车数量以及运营场地之间关系的估计回归方程。

a. 用 $\alpha=0.05$ 来检测因变量和两个自变量之间的显著性关系。

b. 在 $\alpha=0.05$ 的显著性水平下，投入运营的汽车数量这个变量是显著的吗？

c. 在 $\alpha=0.05$ 的显著性水平下，运营场地这个变量是显著的吗？

15.6 利用估计回归方程进行估计和预测

在多元回归分析中，估计 y 的均值和某个特定值的步骤类似于包含一个自变量的回归分析。首先，回顾第 14 章，我们已经说明了对于 x 的给定值，y 的期望值的点估计和 y 的某个特定值的点估计是一样的。在这两种情况下，我们都利用$\hat{y}=b_0+b_1x$ 作为点估计。

在多元回归分析中，我们使用的步骤是一样的。即：将自变量 $x_1, x_2, \cdots, x_p$ 的给定值代入估计回归方程中，并利用相对应的值作为它们的点估计。考察 Bulter 运输公司的例子，我们可以利用包含两个自变量 x_1(行驶里程)和 x_2(运送的次数)的估计回归方程来建立下面的两个估计：

1. 对于所有行驶 100 英里和运送两次货物的卡车，建立这些汽车平均行驶时间的**置信区间估计**。
2. 对于某个给定的运货汽车，如果行驶 100 英里和运送两次货物，则建立该汽车行驶时间的**预测区间估计**。

利用估计回归方程$\hat{y}=-0.869+0.061\,1x_1+0.923x_2$，当 $x_1=100$ 和 $x_2=2$ 时，我们得到$\hat{y}$ 值如下：

$$\hat{y}=-0.869+0.061\,1\times100+0.923\times2=7.09$$

因此，在这两种情况下，汽车行驶时间的点估计近似为 7 个小时。

为了建立关于 y 的平均值和某个别值的区间估计，我们利用类似于包含一个自变量的回归分析步骤。虽然所需要的公式已经超出了本书的讨论范围，但是对于多元回归分析，一旦设定了自变量 $x_1, x_2, \cdots, x_p$ 的值，则计算机软件包往往能提供这些区间估计。在 Butler 运输公司的示例中，对于我们所选取的 x_1 和 x_2 的数值，我们在表 15.4 中给出了 95% 置信区间估计和 95% 预测区间估计；这些数据都可以利用 Minitab 得到。请注意，y 的某个特定值的区间估计比 y 的期望值的区间估计的区间范围更宽。这个区别仅仅反映了这样一个事实：对于给定的 x_1 和 x_2 值，我们对所有运货汽车的平均行驶时间的预测，它的精度比预测某一辆特定运货汽车的行驶时间的精度更高。

表 15.4 Butler 运输公司的 95% 置信区间估计和 95% 预测区间估计

		置信区间估计		预测区间估计	
x_1 的数值	x_2 的数值	下限	上限	下限	上限
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

练　习

方　法

27. 在练习 1 中，下面的估计回归方程是根据 10 个观测值所建立的：

$$\hat{y} = 29.127\,0 + 0.590\,6x_1 + 0.498\,0x_2$$

a. 当 $x_1 = 180$, $x_2 = 310$ 时，计算 y 的平均值的点估计。

b. 当 $x_1 = 180$, $x_2 = 310$ 时，计算 y 的个别值的点估计。

28. 参考练习 2 中的数据，估计回归方程如下所示：



$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

a. 当 $x_1 = 45$, $x_2 = 15$ 时，建立关于 y 的均值的 95% 置信区间估计。

b. 当 $x_1 = 45$, $x_2 = 15$ 时，建立关于 y 的 95% 预测区间估计。

应　用

29. 在练习 5 中，Showtime Movie Theatres 公司的老板利用多元回归分析来预测每周的总营业收入 y，每周的总营业收入 y 是电视广告费用(x_1)和报纸广告费用(x_2)的函数。估计回归方程如下：



$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

a. 当电视广告费用为 3 500 美元(即 $x_1 = 3.5$)，报纸广告费用为 1 800 美元(即 $x_2 = 1.8$)时，计算预期的每周总营业收入。

b. 当广告费用如(a)中所示时，建立一个每周总收入平均值的 95% 置信区间估计。

c. 假设两种广告费用如(a)中那样分配，建立一个下周总收入 95% 预测区间估计。

30. 在练习 6 中，我们已经给出了 16 种赛车和 GT 车的自重、马力和从开始启动加速至 1/4 英里时的速度(*1998 Road and Track Sports & GT Cars*)。

a. 对于 1999 Porsche 911 Carrera 型的赛车，它的自重为 2 910 磅，马力为 296，试估计这辆赛车加速至 1/4 英里时的速度。

b. 对所有具备(a)中特征的各种型号的赛车和 GT 车，建立关于加速至 1/4 英里时的速度的 95% 置信区间估计。

c. 对(a)中所示的 1999 Porsche 911 Carrera 型的赛车，建立关于加速至 1/4 英里时的速度的

95%预测区间估计。

31. 在练习9中，我们已经建立了关于进入4年制大学的学生比例的估计回归方程，以学区平均每班学生人数和综合SAT成绩作为自变量。
    a. 如果某个学区平均每班有25个学生，学生综合SAT成绩为1 000分，请建立一个能进入4年制大学的学生平均比例的95%置信区间估计。
    b. 假设南卡罗来纳州Conway的一个学区平均每班有25个学生，综合SAT成绩为950分，请建立一个该学区能够进入4年制大学的学生比例的95%预测区间估计。

15.7 定性自变量

到目前为止，我们所考察的示例都是关于定量自变量的，例如学生人数、运送货物行驶距离和运送货物次数等等。但是，在许多情形下，我们必须利用**定性自变量**(qualitative independent variables)来工作，例如性别(男、女)、付款方式(现金、信用卡、支票)，等等诸如此类的问题。本节的目的就在于说明在回归分析中如何应用定性变量。为了应用和解释定性变量，我们将考察Johnson过滤水公司所面临的一个问题。

自变量既可以是定性变量，也可以是定量变量。

示例：Johnson过滤水公司

Johnson过滤水公司给南加利福尼亚州的水过滤系统提供维修保养服务。当顾客的水过滤系统出现故障时，他们就会请求Johnson过滤水公司提供维修服务。为了估计维修服务的时间和成本，Johnson过滤水公司的管理者计划预测每一次维修服务所需要的时间。因此，将维修时间作为因变量，维修时间与两个因素相关：最后一次维修以来的月数、维修问题的类型(是机械的还是电子的)。10次维修服务的样本数据如表15.5所示。

表15.5 Johnson过滤水公司数据

服务序号	从最后一次维修以来的月数	维修类型	维修时间(小时)
1	2	电子	2.9
2	6	机械	3.0
3	8	电子	4.8
4	3	机械	1.8
5	2	电子	2.9
6	7	电子	4.9
7	9	机械	4.2
8	8	机械	4.8
9	4	电子	4.4
10	6	电子	4.5

令 y 表示维修时间，x_1 表示从最后一次维修以来的月数。仅用 x_1 来预测 y 的回归模型如下所示：

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

利用 Minitab 来建立估计回归方程，我们得到如图 15.7 所示的输出结果，估计回归方程如下所示：

$$\hat{y} = 2.15 + 0.304 x_1 \qquad (15.16)$$

在显著性水平 0.05 下，对于 t（或 F 检验），p - 值为 0.016，这表明最后一次维修以来的月数显著依赖于维修时间。R - sq = 53.4%，这表明 x_1 仅仅解释了维修时间 53.4% 的变异性。

```
The regression equation is
Time = 2.15 + 0.304 Months

Predictor       Coef      Stdev    t-ratio       p
Constant      2.1473     0.6050       3.55   0.008
Months        0.3041     0.1004       3.03   0.016

s = 0.7810      R-sq = 53.4%     R-sq(adj) = 47.6%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1      5.5960      5.5960      9.17   0.016
Error         8      4.8800      0.6100
Total         9     10.4760
```

图 15.7　Johnson 过滤水公司以从最后一次维修以来的月数作为自变量的 Minitab 输出结果

在 Minitab 输出结果中，变量名称 Months(月份)和 Time(时间)作为工作单上的表头输入，那么，x_1 = 月份，y = 时间。

为了将故障的维修类型导入回归模型，我们给出以下的变量定义：

$$x_2 = \begin{cases} 0 & \text{如果维修类型是机械} \\ 1 & \text{如果维修类型是电子} \end{cases}$$

在回归分析中，x_2 被称为**虚拟变量**(dummy)或**指标变量**(variable)。利用这个虚拟变量，我们可以得到如下的多元回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

表 15.6 是包含了虚拟变量数值的修正数据集。利用 Minitab 和表 15.6 中的数据，我们可以得到模型参数的估计值。图 15.8 中的 Minitab 输出结果表明估计多元回归方程如下所示：

$$\hat{y} = 0.93 + 0.388 x_1 + 1.26 x_2 \qquad (15.17)$$

表 15.6 用虚拟变量表示维修类型（$x_2=0$ 表示机械；$x_2=1$ 表示电子）的 Johnson 过滤水公司的示例数据


CD 光盘数据
Johnson

客户	最后一次维修以来的月数(x_1)	维修类型(x_2)	以小时计的维修时间(y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

在 0.05 的显著性水平下，与 F 检验($F=21.36$)相联系的 p-值为 0.001，这表明回归关系是显著的。在图 15.8 中 t 检验的打印输出部分表明，从最后一次维修服务以来的月数(p-值为 0.000)和维修的类型(p-值 =0.005)，这两个变量都是统计显著的。另外，R-sq = 85.9% 和 R-sq(adj) = 81.9%，这表明估计回归方程很好地解释了维修时间的变异性。因此，公式(15.17)在估计不同服务的必要维修时间时是很有用的。

```
The regression equation is
Time = 0.930 + 0.388 Months + 1.26 Type

Predictor        Coef      Stdev   t-ratio       p
Constant       0.9305     0.4670      1.99   0.087
Months        0.38762    0.06257      6.20   0.000
Type           1.2627     0.3141      4.02   0.005

s = 0.4590      R-sq = 85.9%     R-sq(adj) = 81.9%

Analysis of Variance

SOURCE        DF          SS          MS        F        p
Regression     2      9.0009      4.5005    21.36    0.001
Error          7      1.4751      0.2107
Total          9     10.4760
```

图 15.8 Johnson 过滤水公司例子的 Minitab 输出结果(以最后一次维修以来的月数 x_1 和维修类型 x_2 作为自变量)

在 Minitab 输出结果中，变量名称 Months(月份)、Type(类型)和 Time(时间)作为工作单上的表头输入，那么，x_1 = 月份，x_2 = 类型，y = 时间。

参数解释

Johnson 过滤水公司例子中的多元回归方程如下所示：

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2 \tag{15.18}$$

为了理解定性变量的参数 β_0、β_1、β_2 的意义，我们可以考虑当 $x_2=0$(即机械维修)时的情况。假设对于机械维修，我们用 $E(y|\text{机械})$来表示平均维修时间，可以得到如下公式：

$$E(y|\text{机械})=\beta_0+\beta_1x_1+\beta_2(0)=\beta_0+\beta_1x_1 \tag{15.19}$$

与此相似，对电子维修($x_2=1$)，我们可以得到

$$E(y|\text{电子})=\beta_0+\beta_1x_1+\beta_2(1)=\beta_0+\beta_1x_1+\beta_2=(\beta_0+\beta_2)+\beta_1x_1 \tag{15.20}$$

比较公式(15.19)和(15.20)，我们看到期望的机械和电子的维修时间是 x_1 的线性数。两个方程的斜率是 β_1，但 y 的截距不同。对机械维修来说，y 的截距是公式(15.19)中的 β_0，对公式(15.20)中的电子维修是$(\beta_0+\beta_2)$。β_2 说明了在机械的期望维修时间与电子的期望维修之间的差别。

如果 β_2 是正的，电子的期望维修时间要比机械的长；如果 β_2 是负的，电子的期望维修时间将比机械的短。最终，若 $\beta_2=0$，说明两者的期望维修时间没有差别。

利用多元估计回归方程$\hat{y}=0.93+0.388x_1+1.26x_2$，我们可以看到 0.93 是 β_0 的估计值，1.26 是 β_2 的估计值。因此，当 $x_2=0$（机械维修）时

$$\hat{y}=0.93+0.388x_1 \tag{15.21}$$

当 $x_2=1$(电子维修)时

$$\begin{aligned}\hat{y}&=0.93+0.388x_1+1.26(1)\\&=2.19+0.388x_1\end{aligned} \tag{15.22}$$

实际上，对维修故障类型引入虚拟变量，我们可以得到用于预测维修时间的两个方程，一个对应于机械类型故障的维修时间，另一个对应于电子类型故障的维修时间。除此之外，由于 $b_2=1.26$，我们可以计算出维修电子类型故障所需时间比维修机械类型故障平均多用 1.26 小时。

图 15.9 是 Johnson 过滤水公司数据散点图，用纵轴表示按小时计算的维修时间(y)，用水平

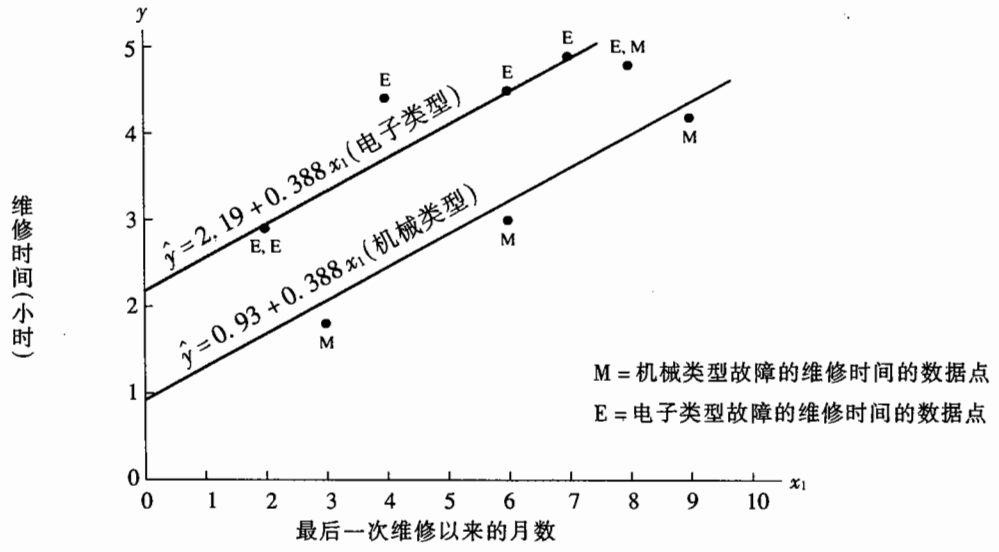


图 15.9　Johnson 过滤水公司数据散点图

轴表示最后一次维修以来的月数(x_1)。用 M 表示机械类型故障的维修时间的数据点，用 E 表示电子类型故障的维修时间的数据点。我们可以在散点图上绘出公式(15.21)和(15.22)所示的图形，这就从几何上证明了这两个方程可以用于预测维修时间，一个对应于机械维修，一个对应于电子维修。

更为复杂的定性变量

对于 Johnson 过滤水公司的例子，由于定性变量只有两个水平，即机械和电子，因此我们只需要很简单地定义一个虚拟变量就可以达到目的，用 0 表示机械类型故障维修，用 1 表示电子类型故障维修。但是如果定性变量多于两个水平时，则我们必须小心设置，既要考虑定义两个虚拟变量也要考虑对它们加以解释。就像我们所要说明的那样，如果某个定性变量有 k 个水平，则我们需要定义 $k-1$ 个虚拟变量，每一个虚拟变量的值可以取为 0 或者 1。

具有 k 个水平的定性变量必须用 $k-1$ 个虚拟变量来模拟，而且我们必须非常小心地对这些虚拟变量进行定义和解释。

举个例子，假设某个复印机制造商将其销售区域组织分为某个州的 A、B、C 三个区域。管理者希望能够利用回归分析方法来帮助他们预测每个星期复印机的销售数量。将复印机的销售数量作为因变量，同时考虑另外几个自变量(销售人员的数量、广告费用支出等等)。同时，管理者认为销售地区也是预测复印机销售数量的重要因素。由于销售地区是定性变量，它有 A，B，C 三个水平，所以我们需要 $3-1=2$ 个虚拟变量来表示销售地区。每个虚拟变量的定义如下所示：

$$x_1 = \begin{cases} 1 & \text{如果销售区域为 B} \\ 0 & \text{其他情况} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{如果销售区域为 C} \\ 0 & \text{其他情况} \end{cases}$$

根据这个定义，我们可以得到 x_1 和 x_2 的数值如下：

区域	x_1	x_2
A	0	0
B	1	0
C	0	1

对应于区域 A 的观测值被定义为 $x_1=0$，$x_2=0$；对应于区域 B 的观测值被定义为 $x_1=1$，$x_2=0$；对应于区域 C 的观测值被定义为 $x_1=0$，$x_2=1$。

复印机销售数量期望值 $E(y)$关于虚拟变量的回归方程如下所示：

$$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2$$

为了帮助解释 β_0、β_1 和 β_2 的含义，我们可以考察回归方程的三种变化：

$$E(y|\text{销售区域 A}) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 = \beta_0$$
$$E(y|\text{销售区域 B}) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 = \beta_0 + \beta_1$$
$$E(y|\text{销售区域 C}) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 = \beta_0 + \beta_2$$

因此，β_0 是销售区域 A 的销量平均值或者期望值，β_1 是销售区域 B 的销量平均值或者期望值，β_2 是销售区域 C 的销量平均值和销售区域 A 的销量平均值之间的差。

因为销售区域是有三个水平的定性变量，所以需要有 2 个虚拟变量来表示，但是，我们可以指定 $x_1 = 0$，$x_2 = 0$ 表示销售区域 A，指定 $x_1 = 1$，$x_2 = 0$ 表示销售区域 B，指定 $x_1 = 0$，$x_2 = 1$ 表示销售区域 C，这种指定是任意的。例如，我们同时也可以指定 $x_1 = 1$，$x_2 = 0$ 表示销售区域 A，指定 $x_1 = 0$，$x_2 = 0$ 表示销售区域 B，指定 $x_1 = 0$，$x_2 = 1$ 表示销售区域 C。在这种情形下，β_1 可以被解释为销售区域 A 的销量均值和销售区域 B 的销量均值之间的差，β_2 可以被解释为销售区域 C 的销量平均值和销售区域 B 的销量平均值之间的差。

我们需要重视的要点在于：在多元回归分析之中，如果某个定性变量有 k 个水平，则必须定义 $k-1$ 个虚拟变量。因此，如果在销售区域的例子中还有第四个销售区域，假设表示为 D，则我们需要定义三个虚拟变量，x_3 定义如下：

$$x_3 = \begin{cases} 1 & \text{如果销售区域为 D} \\ 0 & \text{其他情况} \end{cases}$$

练　习

方　法

32. 考察如下一个回归方程，该方程包含一个因变量 y，一个定量的自变量 x_1 以及一个具有两个水平的(水平 1 和水平 2)定性自变量。



a. 请写出因变量 y 关于自变量 x_1 和定性自变量的多元回归方程。

b. 对应于定性自变量的水平 1 的 y 的期望值是多少？

c. 对应于定性自变量的水平 2 的 y 的期望值是多少？

d. 解释回归方程中的参数。

33. 考察如下一个回归方程，该方程包含一个因变量 y，一个定量的自变量 x_1 以及一个具有两个水平的(水平 1、水平 2 和水平 3)定性自变量。

a. 为了表示定性变量，需要几个虚拟变量。

b. 请写出因变量 y 关于自变量 x_1 和定性自变量的多元回归方程。

c. 解释回归方程中的参数。

应　用

34. 下面的回归方程可以用于预测一家快餐店的销售收入：

自测题

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

其中　x_1——一英里之内的竞争对手数目；

x_2——一英里之内的人口数目(千人)；

x_3——1　如果为驾车者设置不需要下车的供应窗口；

　　0　其他情况；

y = 销量(千美元)。

在调查了 20 家快餐店以后，得到如下估计回归方程：

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a. 为驾车者设置不需要下车的供应窗口所带来的预期销量是多少？

b. 如果有一家快餐店，在一英里范围内有 2 个竞争者，有 8 000 人口，并且没有为驾车者设置不需要下车的供应窗口，则这家快餐店预测的销售收入是多少？

c. 如果有一家快餐店，在一英里范围内有 1 个竞争者，有 3 000 人口，并且为驾车者设置了不需要下车的供应窗口，则这家快餐店的预测销售收入是多少？

35. 考察本节所介绍的 Johnson 过滤水公司例子，假设管理者获得的资料既包括最后一次维修以来的月数和故障的类型是电子还是机械，同时还包括一张记录维修人员完成维修任务的表单，修正后的数据如下所示：

CD 光盘数据
Repair

维修时间(小时)	最后一次维修以来的月数	故障类型	维修人员
2.9	2	电子	Dave Newton
3.0	6	机械	Dave Newton
4.8	8	电子	Bob Jones
1.8	3	机械	Dave Newton
2.9	2	电子	Dave Newton
4.9	7	电子	Bob Jones
4.2	9	机械	Bob Jones
4.8	8	机械	Bob Jones
4.4	4	电子	Bob Jones
4.5	6	电子	Dave Newton

a. 暂时忽略从最后一次维修以来的月数(x_1)以及哪一个维修人员完成维修任务，建立估计简单线性方程，根据故障类型(x_2)来预测维修时间(y)。如果维修故障类型为机

械，则 $x_2=0$；如果维修故障类型为电子，则 $x_2=1$。

b. 怎么评价(a)中所建立的估计回归方程对观测数据所给出的拟合？请解释。

c. 暂时忽略从最后一次维修以来的月数以及故障类型，建立估计简单线性方程，在已知哪一个维修人员完成维修任务的情况下，预测维修时间。如果维修任务是由 Bob Jones 完成，则 $x_3=0$；如果维修任务是由 Dave Newton 完成，则 $x_3=1$。

d. 评论(c)中所建立的估计回归方程对观测数据所给出的拟合，并解释。

36. 这个问题是练习 35 的扩展。

a. 建立估计回归方程，在已知从最后一次维修以来的月数，维修故障类型和维修任务由哪一位维修人员完成的情况下，预测维修时间。

b. 在 0.05 的显著性水平下，检验(a)中所建立的估计回归方程的自变量和因变量之间是否存在显著性关系。

c. 自变量 x_3 表示维修任务由哪一位维修人员完成，取 $\alpha=0.05$ 这个自变量在统计上是否是显著的？对所观测的结果，如何解释？

37. 全国橄榄球联盟根据球员的场上位置将他们分成 5—9 个等级。等级的划分如下标准：8—9 级从第一年开始；7.0—7.9 级现在开始；6.0—6.9 级将作为球队后备力量；5.0—5.9 级可以组建俱乐部及其分部。下表所示的是 25 个球员的场上位置、体重、速度(40 码)以及等级(*USA Today*, April 14, 2000)：



CD 光盘数据
Football

	位置	体重(磅)	速度(秒)	等级
Cosey Coleman	守门员	322	5.38	7.4
Travis Claridge	守门员	303	5.18	7.0
Kaulana Noa	守门员	317	5.34	6.8
Leander Jordan	守门员	330	5.46	6.7
Chad Clifton	守门员	334	5.18	6.3
Manula Savea	守门员	308	5.32	6.1
Ryan Johanningmei	守门员	310	5.28	6.0
Mark Tauscher	守门员	318	5.37	6.0
Blaine Saipaia	守门员	321	5.25	6.0
Richard Mercier	守门员	295	5.34	5.8
Damion McIntosh	守门员	328	5.31	5.3
Jeno James	守门员	320	5.64	5.0
Al Jackson	守门员	304	5.20	5.0
Chris Samuels	攻球手	325	4.95	8.5
Stockar McDougle	攻球手	361	5.50	8.0
Chris McIngosh	攻球手	315	5.39	7.8
Adrian Klemm	攻球手	307	4.98	7.6
Todd Wade	攻球手	326	5.20	7.3
Marvel Smith	攻球手	320	5.36	7.1
Michael Thompson	攻球手	287	5.05	6.8
Bobby Williams	攻球手	332	5.26	6.8
Darnell Alford	攻球手	334	5.55	6.4
Terrance Beadles	攻球手	312	5.15	6.3
Tutan Reyes	攻球手	299	5.35	6.1
Greg Robinson-Ran	攻球手	333	5.59	6.0

a. 建立一个虚拟变量来表示球员的场上位置。

b. 建立一个估计回归方程来表示球员等级与场上位置、体重，以及速度之间的相关关系。

c. 在 0.05 的显著性水平下，检测(b)中所建立的估计回归方程里面自变量和因变量之间的关系是否显著？

d. 所建立的估计回归方程是否对所观测的数据给出了一个优良的拟合？

e. 球员的场上位置是否是决定球员等级的显著因素？利用 $\alpha = 0.05$，并解释原因。

f. 假设某个进攻球员体重 300 磅，40 码距离短跑时间为 5.1 秒。利用(b)中所建立的估计回归方程估计这个球员的级别。

38. 美国心脏协会进行了一次历时 10 年的研究，获得了和中风病发率有关的年龄、血压以及吸烟的统计资料。以下是这一研究的部分数据。中风风险指的是病人在今后 10 年内中风发作的概率(乘 100)。我们用一个虚拟变量来表示病人是否为吸烟者，1 表示病人是吸烟者，0 表示病人不是吸烟者。



CD 光盘数据
Stroke

发病概率	年龄	血压	是否吸烟
12	57	152	否
24	67	163	否
13	58	155	否
56	86	177	是
28	59	196	否
51	76	189	是
18	56	155	是
31	78	120	否
37	80	135	是
15	78	98	否
22	71	152	否
36	70	173	是
15	67	135	是
48	77	209	是
15	60	199	否
36	82	119	是
8	66	166	否
34	80	125	是
3	62	117	否
37	59	207	是

a. 利用这些数据，建立一个估计回归方程，因变量为发病概率，自变量是病人的年龄、血压，以及病人是否是吸烟者。

b. 吸烟是否是发病概率的一个显著因素？利用显著性水平 $\alpha = 0.05$，并解释原因。

c. Art Speen 是一个 68 岁的吸烟者，他的血压为 175，请计算他的发病概率。对这位病人，医生应该采取什么样的建议？

15.8　残差分析

在第 14 章中，我们阐述了在残差散点图和对异常值的识别中可以经常使用标准化残差。对于观测值 i 的标准化残差的一般性公式如下所示：

观测值 i 的标准化残差

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

式中，$s_{y_i-\hat{y}_i}$ 为残差 i 的标准差。

残差 i 的标准差的一般性公式定义如下；

残差 i 的标准差

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{15.24}$$

式中　s——估计值的标准误差；

h_i——第 i 次观测值的杠杆率。

正如我们在第 14 章所说的那样，观测值的**杠杆率**(leverage)取决于自变量的观测值与这些观测值的平均值之间的差距。对于多元回归情形，计算第 i 次观测的 h_i、$s_{y_i-\hat{y}_i}$ 和标准化残差，如果用手算将是极其复杂的。但是，如果利用统计软件包进行计算，我们可以很容易地从计算机输出结果里面得到标准化残差。表 15.7 列出了本章前面所介绍的 Butler 运输公司例子的预测值、残差和标准化残差，我们利用 Minitab 统计软件包可以得到这些数值。表中的预测值是根据估计回归方程 $\hat{y} = -0.869 + 0.061\,1x_1 + 0.923x_2$ 计算出来的。

表 15.7　Butler 运输公司回归分析的残差和标准化残差

行驶里程 (x_1)	运送次数 (x_2)	行驶时间 (y)	预测值 ($\hat{y}$)	残差 ($y-\hat{y}$)	标准化残差
100	4	9.3	8.938 46	0.361 541	0.783 44
50	3	4.8	4.958 30	−0.158 304	−0.349 62
100	4	8.9	8.938 46	−0.038 460	−0.083 34
100	2	6.5	7.091 61	−0.591 609	−1.309 29
50	2	4.2	4.034 88	0.165 121	0.381 67
80	2	6.2	5.868 92	0.331 083	0.654 31
75	3	7.4	6.486 67	0.913 331	1.689 17
65	4	6.0	6.798 75	−0.798 749	−1.773 72
90	3	7.6	7.403 69	0.196 311	0.367 03
90	2	6.1	6.480 26	−0.380 263	−0.776 39

图 15.10 所示的是 Butler 运输公司多元回归示例的标准化残差图，我们可以在表 15.7 中找到图中所使用的标准化残差和 y 的观测值。这个标准化残差图并没有表示出任何特殊反常的情况。与此同时，所有的标准化残差都位于 -2 和 +2 之间。因此，我们没有理由去质疑 ε 呈正态分布这一假定，我们可以得出结论：关于模型的假设是合理的。

我们还可以用正态概率分布图来确定 ε 是否呈正态分布这一假定。正态概率分布图的步骤和解释已经在 14.8 节中讨论过。同样的步骤也适用于多元回归分析，除此之外，我们仍然可以利用统计软件包来完成计算并绘出正态分布图。

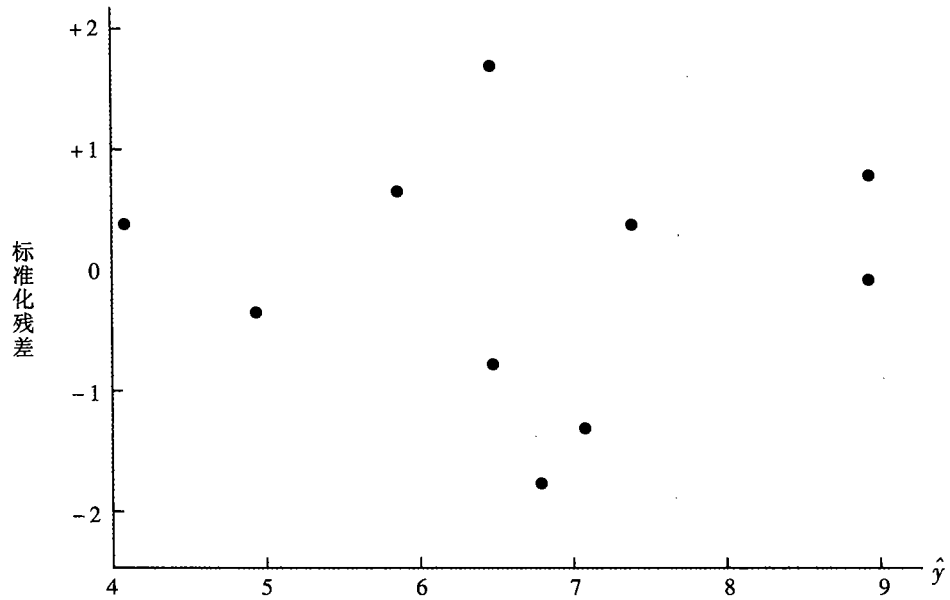


图 15.10 **Butler 运输公司的标准化残差图**

检测异常值

异常值(outlier)与其他数据相比较而言是一个特殊的观测值；换句话说，异常值与根据其他观测值所做的散点图所显示的趋势不相符。在第 14 章中，我们曾经给出了一个关于异常值的例子，并且讨论了如何利用标准化残差来检测异常值。如果某个观测值的标准化残差数值小于 -2 或者大于 +2，则 Minitab 能够识别出这个观测值是一个异常值。对 Butler 运输公司例子中的标准化残差(见表 15.7)，我们可以利用这一规则从数据集合中检测出所有的异常值。

总的说来，如果在数据集合中存在一个或者多个异常值使得 s 增大，则估计标准化残差 i 的数值$s_{y_i-\hat{y}_i}$将会增加。由于在标准化残差公式(15.23)中分母是，但是在计算时需要考虑 s，所以标准化残差的数值将会随着 s 的增加而减少。结果是残差虽然有可能很大，但是因为在公式(15.23)中存在一个大的分母，因此，有可能导致识别某个观测值为异常值的标准化残差规则无法

应用。我们可以通过应用一种被称为**按学生分布删除残差**(studentized deleted residuals)的标准化残差来避免这种情况的发生。

按学生分布删除残差和异常值

假设从数据集合中删除第 i 次观测值，并利用其他的 $n-1$ 次观测值建立一个新的估计回归方程。用 $s_{(i)}$表示删除了第 i 次观测值的数据集合所得到的估计标准误差，如果我们计算第 i 次观测残差的标准差，用 $s_{(i)}$代替公式(15.24)中的 s，则我们在计算第 i 次观测的标准化残差时公式(15.23)利用了$s_{y_i-\hat{y}_i}$的修正值，这样处理后所得到的标准化残差被称为按学生分布删除残差。如果第 i 次观测值是一个异常值，则 $s_{(i)}$会小于 s。所以，第 i 次观测的按学生分布删除残差的绝对值将大于标准化残差的绝对值。在这种意义下，按学生分布删除残差能够检测出标准化残差所不能检测出的异常值。

许多统计软件包提供了可以得到按学生分布删除残差的可选项。利用 Minitab，我们可以得到 Butler 运输公司的按学生分布删除残差，其结果如表 15.8 所示。我们可以利用 t 分布来确定按学生分布删除残差是否存在异常值。回顾前面章节，我们用 p 表示自变量的个数，用 n 表示观测值的个数。因此，如果我们将第 i 次观测值从数据集合中删除，那么数据集合中剩下 $n-1$ 个观测值；在这种情形下，误差平方和的自由度是$(n-1)-p-1$。对于 Butler 运输公司例子，$n=10$，$p=2$，删除第 i 次观测值以后，自由度是 $9-2-1=6$。在 0.05 显著性水平下，具有 6 个自由度的 t 分布双侧分位数 $t_{0.025}=2.447$ (见附录 B 表 2)。如果第 i 次观测的按学生分布删除残差的数值小于 -2.447 或者大于 $+2.447$，则我们可以得出结论：第 i 次观测值是异常值。因为在表 15.8 中，按学生分布删除残差没有超出这一界限，所以我们可以断定：在这个数据集合里，不可能存在有异常值。

表 15.8　Butler 运输公司按学生分布删除残差

行驶里程(x_1)	送货次数(x_2)	行驶时间(y)	标准化残差	学生化删除残差
100	4	9.3	0.783 44	0.759 39
50	3	4.8	−0.349 62	−0.326 54
100	4	8.9	−0.083 34	−0.077 20
100	2	6.5	−1.309 29	−1.394 94
50	2	4.2	0.381 67	0.357 09
80	2	6.2	0.654 31	0.625 19
75	3	7.4	1.689 17	2.031 87
65	4	6.0	−1.773 72	−2.213 14
90	3	7.6	0.367 03	0.343 12
90	2	6.1	−0.776 39	−0.751 90

有影响的观测值

在 14.9 节中，我们讨论了怎样利用观测值的杠杆率来识别具有较强影响的观测值，这些观测值对回归结果的影响比较强。正如我们在关于标准化残差的讨论中所述的那样，观测值的杠杆率，以 h_i 表示，它是关于自变量和观测值的平均值之间差距的度量。杠杆率的值可以很轻松地从统计软件包里面得到。Minitab 能够计算出杠杆率的数值，而且还能够利用经验法则：$h_i > 3(p+1)/n$ 来识别**有影响的观测值**(influential observations)。对于 Butler 运输公司的例子，$p=2$，$n=10$，杠杆率的临界值是：$3\times(2+1)/10=0.9$。利用 Minitab 可以得到 Butler 运输公司例子的杠杆率数值，如表 15.9 所示。由于 h_i 不超过 0.9，所以我们在这个数据集合里没有检测到有影响的观测值。

表 15.9 **Butler 运输公司的杠杆率和库克距离测度**

行驶里程(x_1)	运送次数(x_2)	行驶时间(y)	杠杆率(h_i)	库克距离测度(D_i)
100	4	9.3	0.351 704	0.110 994
50	3	4.8	0.375 863	0.024 536
100	4	8.9	0.351 704	0.001 256
100	2	6.5	0.378 451	0.347 923
50	2	4.2	0.430 220	0.036 663
80	2	6.2	0.220 557	0.040 381
75	3	7.4	0.110 009	0.117 562
65	4	6.0	0.382 657	0.650 029
90	3	7.6	0.129 098	0.006 656
90	2	6.1	0.269 737	0.074 217

利用库克距离测度识别有影响的观测值

利用杠杆率识别有影响的观测值时有可能出现一个问题：某个观测值被认为是一个有高杠杆率的观测值，但是根据所得的估计回归方程，这个观测值不一定有较强的影响。例如，表 15.10 是一个由 8 个观测值所组成的数据集合以及与它们相对应的杠杆率(由 Minitab 得出)。因为第 8 个观测值的杠杆率是 0.91>0.75(杠杆率的临界值)，所以我们可以确认这是一个有影响的观测值。但是，在得出任何最后的结论之前，让我们从不同角度来考虑这个问题。

表 15.10 **利用杠杆率标准说明潜在问题的数据集合**

x_i	y_i	杠杆率 h_i
1	18	0.204 170
1	21	0.204 170
2	22	0.164 205
3	21	0.138 141
4	23	0.125 977
4	24	0.125 977
5	26	0.127 715
15	39	0.909 644

图 15.11 显示的是表 15.10 数据集合的散点图和估计回归方程。我们利用 Minitab 建立这些数据的估计回归方程如下：

$$\hat{y} = 18.2 + 1.39x$$

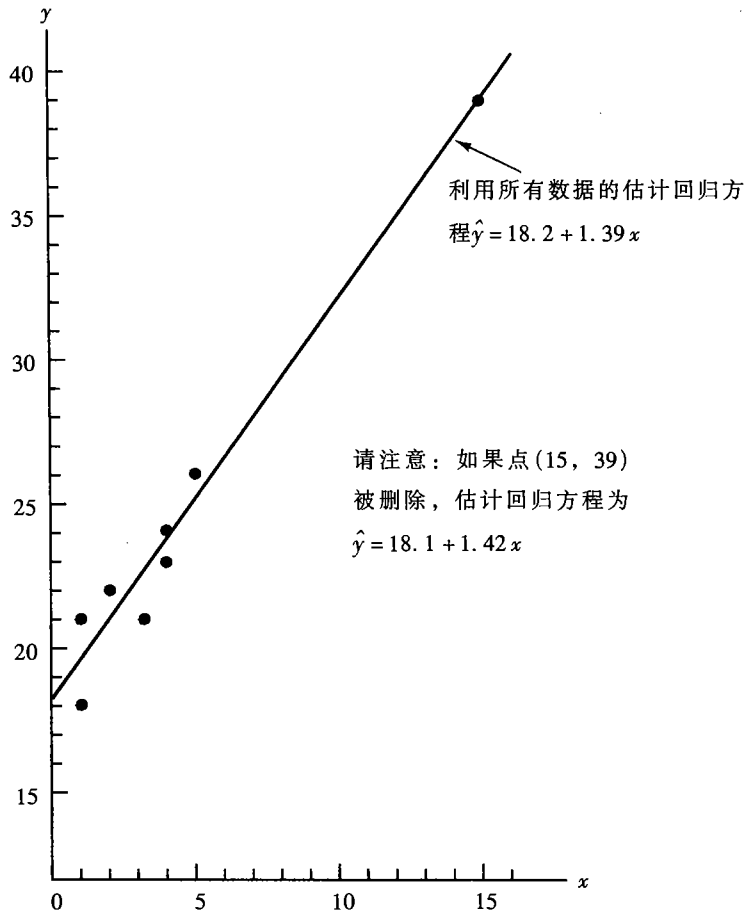


图 15.11　表 15.10 中的数据集合的散点图

图 15.11 中的直线即为这个方程的图形。现在，让我们将 $x = 15$，$y = 39$ 的观测值删除，并对剩余的 7 个观测值作出估计回归方程拟合。新的估计回归方程如下所示：

$$\hat{y} = 18.1 + 1.42x$$

请注意，新的估计回归方程的 y 轴截距和斜率与利用所有数据得到的数值并没有显著的不同。虽然，杠杆率法则确认第 8 个观测值为一个有影响的观测值，但是，很显然这个观测值对结果的影响微乎其微。因此，在某些情况下，仅仅使用杠杆率来确认有影响的观测值将导致错误的结论。

库克距离测度(Cook's distance measure)利用第 i 次观测值的杠杆率 h_i 和残差 $(y_i - \hat{y}_i)$ 来确定观测值是否有较大的影响。

库克距离测度区分有影响的观测值

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{15.25}$$

式中 D_i——对观测值 i 的库克距离测度；

$y_i - \hat{y}_i$——观测值 i 的残差；

h_i——观测值 i 的杠杆率；

p——自变量的数量；

s——估计值的标准差。

如果库克距离的测度随着观测值的残差和/或杠杆率的增大而增大，则表明这个观测值是一个有影响的观测值。作为经验法则，如果 $D_i>1$，则表明第 i 次观测值是一个有影响的观测值，因此应该对这个观测值作更深一步的考察。对于 Butler 运输公司的示例，表 15.9 最后一列是由 Minitab 给出的库克距离测度。第 8 个观测值的库克距离测度 $D_i = 0.650\,029$，这个观测值的影响最大。但是，应用经验法则 $D_i>1$，对于 Butler 运输公司的例子我们不应该考虑有影响的观测值的存在。

评 注

1. 确认异常值和有影响的观测值的步骤可以提前警告我们：某些观测值可能会对回归结果产生潜在的影响。我们应该对每一个异常值和有影响的观测值进行仔细的检查。如果发现数据误差，则应该修正数据误差并重新进行回归分析。总的说来，我们不应该将异常值和有影响的观测值从数据集合中删除，除非有明显的证据表明：它们并不是根据我们所研究的样本元素计算得来，因此也就不应该被纳入原始数据集合中。

2. 为了确定库克距离测度 D_i 的数值是否足够大，以使我们能够推断：第 i 次观测值是有影响的。我们也可以将 D_i 的数值和分子自由度为 $p+1$、分母自由度为 $n-p-1$ 的 F 分布的第 50 个百分位数（以 $F_{0.50}$）来进行比较。我们可以利用对应于 0.50 的显著性水平的 F 分布表来执行这一检验。我们所给出的经验法则（$D_i>1$）就是以下事实为根据的：对于很多广泛的不同情形，F 分布表的数值接近于 1。

练　习

方　法

39.　关于两个变量 x 和 y 的数据如下所示：

自测题

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. 根据这些数据建立估计回归方程。

b. 请做关于 $\hat{y}$ 的标准化残差图，在这些数据中是否存在异常值？并解释原因。

c. 计算这些数据的按学生分布删除残差。在 0.05 的显著性水平下，这些观测数据中的任何一个是否可以被认为是异常值？解释原因。

40. 关于两个变量 x 和 y 的数据如下所示：

x_i	22	24	26	28	40
y_i	12	21	31	35	70

a. 根据这些数据建立估计回归方程。

b. 计算这些数据的按学生分布删除残差。在 0.05 的显著性水平下，这些观测数据中的任何一个是否可以被认为是异常值？解释原因。

c. 计算这些数据的杠杆率。在这些数据中，是否存在有影响的观测值？解释原因。

d. 计算关于这些数据的库克距离测度。是否发现有影响的观测值？解释原因。

应　用

41.　练习 5 给出了 Showtime Movie Theaters 每周的营业总收入(千美元)，电视广告费用(千美元)，报纸广告费用(千美元)，如下所示：

自测题

CD 光盘数据
Showtime

每周的营业总收入（千美元）	电视广告费用（千美元）	报纸广告费用（千美元）
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

a. 建立每周的营业总收入关于电视广告费用和报纸广告费用的估计回归方程。

b. 画出关于$\hat{y}$的标准化残差图，这个残差图是否支持关于ε的假设？解释原因。

c. 检查这些数据中的每一个异常值，可以得出什么结论？

d. 是否存在有影响的观测值?解释原因。

42. 练习6给出了16种赛车和GT车的马力，自重，启动加速至1/4英里时的速度。假设每一种赛车和GT车的价格如下所示，完整的数据集合如下：



CD光盘数据 Auto2

赛车和GT车	价格(千美元)	自重(磅)	马力	启动加速至1/4英里时的速度
Accura Integra Type R	25.035	2 577	195	90.7
Accura NSX-T	93.758	3 066	290	108.0
BMW Z3 2.8	40.900	2 844	189	93.2
Chevrolet Camaro Z28	24.865	3 439	305	103.2
Chevrolet Corvette Convertible	50.144	3 246	345	102.1
Dodge Viper RT/10	69.742	3 319	450	116.2
Ford Mustang GT	23.200	3 227	225	91.7
Honda Prelude Type SH	26.382	3 042	195	89.7
Mercedes-Benz CLK320	44.988	3 240	215	93.0
Mercedes-Benz SLK230	42.762	3 025	185	92.3
Mitsubishi 3000GT VR-4	47.518	3 737	320	99.0
Nissan 240 SX SE	25.066	2 862	155	84.6
Pontiac Firebird Trans Am	27.770	3 455	305	103.2
Porsche Boxster	45.560	2 822	201	93.2
Toyota Supra Turbo	40.989	3 505	320	105.0
Volvo C70	41.120	3 285	236	97.0

a. 建立估计回归方程，利用价格和马力来预测从启动加速至1/4英里时的速度。

b. 画出关于$\hat{y}$的标准化残差图，这个残差图是否支持关于ε的假设？解释原因。

c. 检查这些数据中的每一个异常值，可以得出什么结论？

d. 是否存在有影响的观测值?解释原因。

43. 在练习9中，我们给出了学区的班级平均学生数、综合SAT分数和考入四年制大学的学生比例的数据。

a. 建立估计回归方程，在已知综合SAT分数的情况下，预测考入四年制大学的学生比例。

b. 以(a)中所建立的估计回归方程为基础，检查这些数据是否发现异常值和/或有影响的观测值？解释原因。

c. 建立估计回归方程，在已知班级平均学生数、综合SAT分数的情况下，预测考入四年制大学的学生比例。

d. 以(c)中所建立的估计回归方程为基础，检查这些数据，是否发现异常值和/或有影响的观测值？解释原因。

小　结

作为第 14 章中简单线性回归分析的延伸，我们在本章中介绍了多元回归分析。多元回归分析使得我们可以理解因变量是如何与两个或者多个自变量相关联的。回归方程 $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 描述了因变量 y 的期望值或者平均值是如何取决于自变量 x_1，x_2，…，x_p 的值。利用样本数据和最小二乘法我们建立了估计回归方程 $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$。事实上，$b_1$，$b_2$，…，$b_p$ 是用来估计模型未知参数 β_0，β_2，…，β_p 的样本统计量。在整个章节中，我们都使用了计算机输出结果，这样做的目的是为了强调：统计软件包是执行多元回归分析所需要的大量计算的惟一实用工具。

作为对估计回归方程拟合度的一个度量，我们引入了多元判定系数。多元判定系数测量了因变量 y 的变异程度，这可由估计回归方程来解释。与此相似，修正后的多元判定系数也是一个对拟合度的测量，它针对自变量的数目进行调整，因此可以避免过高估计由于自变量数目增加而带来的影响。

同时我们还介绍了 F 检验和 t 检验，这两个检验作为工具可以用于判断变量间的关系在统计上是否显著。F 检验可以用于确定因变量和所有自变量在总体上是否存在显著关系。t 检验可以用于确定在回归模型中的其他自变量已知的情况下，因变量和单个自变量之间是否存在显著关系。多重共线性描述的是自变量之间的相关性，前面已经讨论过。

定性自变量这一节描述了如何利用虚拟变量将定性数据用于多元回归分析。最后一节作为本章的结束部分，我们介绍了如何利用残差分析来证明模型的假设、检测异常值和确认有影响的观测值。同时，本章还讨论了标准化残差、杠杆率、按学生分布删除残差以及库克距离测度。

术语释义

多元回归分析：关于两个或者多个自变量的回归分析。

多元回归模型：描述因变量 y 是怎样依赖于自变量 x_1，x_2，…，x_p 和误差项 ε 的方程。

多元回归方程：描述关于因变量的期望值或均值与自变量的数值之间的关系的数学方程，即：$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$。

多元估计回归方程：以样本数据和最小二乘法方法为基础估计出的多元回归方程，即：$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$。

最小二乘法：用于建立估计回归方程的方法，该方法可以使残差(因变量观测值 y_i 和因变量估计值 $\hat{y}_i$ 之间的离差)平方和最小。

多元判定系数：估计多元回归方程拟合度的测量。它可以解释为：由估计回归方程所导致的因变量 y 的变异程度。

修正多元判定系数：对估计多元回归方程拟合度的测量。它随着模型中自变量个数的变化而进行调整，因此可以避免过高估计由于自变量增加所导致的影响。

多重共线性：用来描述自变量之间关系的术语。

定性自变量：含有定性数据的自变量。

虚拟变量：用来模拟定性自变量的影响的变量。虚拟变量的数值只能取为 1 或者 0。

杠杆率：反映自变量的数值与它们的平均值之间差距的度量。

异常值：与其他观测值轨迹不相吻合的观测值。

按学生分布删除残差：删除第 i 个观测值，然后进行回归分析和计算得到估计值的修正标准误差，再以此为基础所得到的标准化残差。

有影响的观测值：对回归结果有重大影响的观测值。

库克距离测度：根据第 i 次观测值的杠杆率和残差所得到的观测值影响的度量。

重要公式

多元回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{15.1}$$

多元回归方程

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.2}$$

多元估计回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{15.3}$$

最小二乘法

$$\min \sum (y_i - \hat{y}_i)^2 \tag{15.4}$$

SST、SSR 和 SSE 之间的关系

$$\text{SST} = \text{SSR} + \text{SSE} \tag{15.7}$$

多元判定系数

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15.8}$$

修正多元判定系数

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{15.9}$$

均方回归

$$\text{MSR} = \frac{\text{SSR}}{p} \tag{15.12}$$

均方误差

$$\text{MSE} = \frac{\text{SSE}}{n-p-1} \tag{15.13}$$

***F* 检验统计量**

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{15.14}$$

***t* 检验统计量**

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

观测值 *i* 的标准化残差

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

残差 *i* 的标准差

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{15.24}$$

库克距离测度

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(\text{p}-1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{15.25}$$

补充练习

44. Clearwater 大学的行政管理人员建立了如下的估计回归方程，因变量是学生最后的大学 GPA，自变量是学生的 SAT 数学分数和高中 GPA。方程如下所示：

$$\hat{y} = -1.41 + 0.0235x_1 + 0.004\,86x_2$$

式中　x_1——高中 GPA 的；

x_2——SAT 数学分数；

y——最后的 GPA。

a. 解释这个估计回归方程的系数。

b. 某个学生，他的高中 GPA 为 84，SAT 数学分数为 540，请估计他的最后大学 GPA。

45. 电子协会的人事主管建立了如下的估计回归方程，因变量是雇员关于工作满意程度测试的分

数，自变量是他（或她）的雇用时间和薪水。方程如下所示：

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

式中 x_1——雇用时间（年）；

x_2——薪水（美元）；

y——工作满意度测试分数(分数越高，表示工作满意度越高)。

a. 解释这个估计回归方程的系数。

b. 假设某个员工雇用年限为 4 年，每小时薪水为 6.50 美元，请计算出这位员工的工作满意度测试分数。

46. 关于回归分析的部分计算机输出结果如下所示：

The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor	Coef	Stdev	t-ratio
Constant	_____	2.667	_____
X1	_____	2.105	_____
X2	_____	0.613	_____

s = 3.35 R-sq = 92.3% R-sq(adj) = _____%

Analysis of Variance

SOURCE	DF	SS	MS	F
Regression	_____	1612	_____	_____
Error	12	_____	_____	
Total	_____	_____		

a. 计算合适的 t 值。

b. 在 $\alpha = 0.05$ 的显著性水平下，检测 β_1 和 β_2 的显著性。

c. 计算自由度(DF)，平方和(SS)以及均方(MS)列中的空白项。

d. 计算 R_a^2。

47. 回顾练习 44，Clearwater 大学的管理人员建立了如下的估计回归方程，因变量是学生最后的大学 GPA，自变量是学生的 SAT 数学分数和高中 GPA。方程如下所示：

$$\hat{y} = -1.41 + 0.023\,5x_1 + 0.004\,86x_2$$

式中，x_1 = 高中 GPA，x_2 = SAT 数学分数，y = 最后的 GPA。

Minitab 的一部分计算机输出结果如下所示：

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor              Coef          Stdev        t-ratio
Constant            -1.4053         0.4848          _____
X1                 0.023467       0.008666          _____
X2                    _____       0.001077          _____

s = 0.1298      R-sq = _____    R-sq(adj) = _____

Analysis of Variance

SOURCE          DF            SS           MS          F
Regression   _____       1.76209        _____      _____
Error        _____       _______        _____
Total            9       1.88000
```

a. 完成计算机输出结果中的空白项。

b. 计算 F 值，并在 0.05 的显著性水平下，检测是否存在一个统计显著关系。

c. 估计回归方程是否为数据提供了一个良好的拟合，解释原因。

d. 利用 t 检验，在 $\alpha=0.05$ 的显著性水平下，检验假设 H_0: $\beta_1=0$ 和假设 H_0: $\beta_2=0$。

48. 回顾练习 45，电子协会的人事主管建立了如下的估计回归方程，因变量是雇员关于工作满意程度测试的分数，自变量是他(或她)的雇用时间和薪水。方程如下所示：

$$\hat{y}=14.4-8.69x_1+13.5x_2$$

式中，x_1 = 雇用时间(年)，x_2 = 薪水(美元)，y = 工作满意度测试分数(分数越高，表示工作满意度越高)。

Minitab 的一部分计算机输出结果如下所示：

```
The regression equation is
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor              Coef          Stdev        t-ratio
Constant             14.448          8.191           1.76
X1                    _____          1.555          _____
X2                   13.517          2.085          _____

s = 3.773       R-sq = _____%   R-sq(adj) = _____%

Analysis of Variance

SOURCE          DF            SS           MS          F
Regression       2       _______        _____      _____
Error        _____          71.17        _____
Total            7          720.0
```

a. 完成计算机输出结果中的空白项。

b. 计算 F 值，并在 0.05 的显著性水平下，检测是否存在一个统计显著关系。

c. 估计回归方程是否为数据提供了一个良好的拟合，解释原因。

d. 利用 t 检验，在 $\alpha = 0.05$ 的显著性水平下，检验假设 H_0: $\beta_1 = 0$ 和假设 H_0: $\beta_2 = 0$。

49. 从 *Stock Investor Pro* 中摘录了 25 个计算机硬件公司的一些数据，包括每股价格、每股账面价值以及每股投资收益率，如下所示（*Stock Investor Pro*, American Association of Individual Investors, August 21, 1997）：



CD 光盘数据
Computer

	每股价格	每股账面价值	每股投资收益率(%)
Amdahl Corporation	12.31	4.94	-49.7
Apple Computer, Inc.	21.75	9.46	-71.8
Auspex Systems, Inc.	11.00	4.95	17.2
Capital Associates	3.25	4.33	5.1
Compaq Computer Corp.	65.50	9.58	20.8
Data General Corporation	35.94	8.46	13.3
Dell Computer Corporation	82.06	2.33	74.5
Digi International	15.00	7.35	-11.9
Digital Equipment Corp.	43.00	22.40	-12.9
En Pointe Technologies	14.25	4.11	18.8
Equitrac Corporation	16.25	6.83	10.7
Franklin Electronic Pbls.	12.88	9.13	9.0
Gateway 2000, Inc.	39.13	6.07	28.8
Hewlett-Packard Company	61.50	14.14	18.7
IBM	101.38	20.12	29.9
Ingram Micro, Inc.	28.75	6.35	15.1
Maxwell Technologies, Inc.	30.50	3.78	11.8
MicroAge, Inc.	27.19	12.59	9.8
Micron Electronics, Inc.	16.31	3.64	28.3
Network Computing Devices	11.88	3.56	4.0
Pomeroy Computer Resources	33.00	10.03	16.5
Sequent Computer Systems	28.19	10.64	3.3
Silicon Graphics, Inc.	27.44	9.12	-4.3
Southern Electronics Corp.	15.13	6.15	16.1
Stratus Computer, Inc.	55.50	22.38	11.1
Sun Microsystems, Inc.	48.00	6.40	26.2
Tandem Computers, Inc.	34.24	9.49	8.7
Tech Data Corporation	38.94	10.25	14.3
Unisys Corporation	11.31	0.68	1.6
Vitech America, Inc.	14.63	3.48	24.3

a. 建立估计回归方程，在已知每股账面价值的情况下，预测每股价格。在 0.05 的显著性水平下，检测是否存在统计显著关系。

b. (a)中所建立的估计回归方程是否对数据给出了良好的拟合？解释原因。

c. 建立估计回归方程，在已知每股账面价值和每股投资收益率的情况下，预测每股价格。在 0.05 的显著性水平下，检测总体显著性。

50. 以下的数据是关于 16 种赛车和 GT 车的价格、马力、自重、速度从 0 加速到每小时 60 英里所需的时间，启动加速至 1/4 英里时的速度(*1998 Road & Track Sports & GT Cars*)。


CD 光盘数据
Auto3

	价格(千美元)	自重(磅)	马力	0 到 60(秒)	启动加速至 1/4 英里时的速度(英里/小时)
Acura Integra Type R	25.035	2 577	195	7.0	90.7
Acura NSX-T	93.758	3 066	290	5.0	108.0
BMW Z 3 2.8	40.900	2 844	189	6.6	93.2
Chevrolet Camaro Z28	24.865	3 439	305	5.4	103.2
Chevrolet Corvette Convertible	50.144	3 246	345	5.2	102.1
Dodge Viper RT/10	69.742	3 319	450	4.4	116.2
Ford Mustang GT	23.200	3 227	225	6.8	91.7
Honda Prelude Type SH	26.382	3 042	195	7.7	89.7
Mercedes-Benz CLK320	44.988	3 240	215	7.2	93.0
Mercedes-Benz SLK230	42.762	3 025	185	6.6	92.3
Mitsubishi 3000GT VR-4	47.518	3 737	320	5.7	99.0
Nissan 240 SX SE	25.066	2 862	155	9.1	84.6
Pontiac Firebird Trans Am	27.770	3 455	305	5.4	103.2
Porsche Boxster	45.560	2 822	201	6.1	93.2
Toyota Supra Turbo	40.989	3 505	320	5.3	105.0
Volvo C 70	41.120	3 285	236	6.3	97.0

a. 建立一个估计回归方程，利用四个自变量：价格、自重、马力和速度从 0 加速到 60 英里每小时所需要的时间，来预测从启动加速到 1/4 英里时的速度。

b. 利用 F 检验来确定回归结果的显著性。在 0.05 的显著性水平下，你可以得出什么结论？

c. 利用 t 检验来确定每一个自变量的显著性，在 0.05 的显著性水平下，你可以得出什么结论？

d. 删除不显著的自变量，并给出你所推荐的估计回归方程。

e. 画出标准残差图，残差图的图形是否合理？

f. 数据里面是否含有异常值？

g. 数据里面是否含有有影响的观测值？

51. 尼尔森媒体研究搜集了关于在 ABC、CBS、NBC、Fox、UPN 和 WB 电视网的黄金时段播出的广告数据，数据显示了 1997 年 4 月 28 日到 5 月 4 日期间，观看广告的家庭户数(百万)和广告播出的次数(*USA Today*, May 5, 1997)。


CD 光盘数据
Nielsen

广告厂商	广告播出次数	观看广告的家庭户数
Burger King	86	616.7
McDonald's	54	439.2
Sears	33	338.0
Wendy's	28	191.7
Ford Escort	20	174.6
Austin Powers movie	14	161.3
Nissan	16	161.1
Pizza Hut	16	147.7
Saturn	16	146.3
Father's Day movie	11	138.2

a. 以广告播出次数作为自变量，观看广告播出的家庭户数作为因变量，绘制散点图。在这个散点图里面是否存在异常值？解释原因。

b. 建立一个估计回归方程，说明观看广告播出的家庭户数与广告播出的次数之间的联系。这两个变量之间的关系在统计上是否显著？利用 $\alpha=0.05$ 的显著性水平进行检测。

c. 考虑增加一个自变量 BigAds，其中如果广告播出次数大于 30，则 BigAds 数值为 1，反之则为 0。建立一个估计回归方程，在已知广告播出次数和虚拟变量 BigAds 的情况下，预测观看广告播出的家庭户数。

d. 利用 $\alpha=0.05$ 的显著性水平来检测(c)中所增加的虚拟变量的统计显著性。

e. 在模拟广告播出次数和观看广告的家庭户数之间的关系时，虚拟变量起了什么作用？

52. 市场为消费者购买轻便车和小型卡车提供了很大的选择余地，对很多消费者而言，一个重要的影响因素是车的转让销售情况。下表给出了 10 种轻便车，10 种小型卡车，以及 10 种大卡车两年后的转让销售率(%)，以及厂商所建议的零售价格(*Kiplinger's New Cars & Trucks 2000 Buyer's Guide*)。


CD 光盘数据
Trucks

	车型	建议零售价(美元)	转让销售率(%)
Chevrolet Blazer LS	轻便车	19 495	55
Ford Explorer Sport	轻便车	20 495	57
GMC Yukon XL 1500	轻便车	26 789	67
Honda CR-V	轻便车	18 965	65
Isuzu VehiCross	轻便车	30 186	62
Jeep Cherokee Limited	轻便车	25 745	57
Mercury Mountaineer Monterrey	轻便车	29 895	59
Nissan Pathfinder XE	轻便车	26 919	54
Toyota 4Runner	轻便车	22 418	55
Toyota RAV4	轻便车	17 148	55
Chevrolet S-10 Extended Cab	小型卡车	18 847	46
Dodge Dakota Club Cab Sport	小型卡车	16 870	53
Ford Ranger XLT Regular Cab	小型卡车	18 510	48
Ford Ranger XLT Supercab	小型卡车	20 225	55

（续表）

	车型	建议零售价（美元）	转让销售率（%）
GMC Sonoma Regular Cab	小型卡车	16 938	44
Isuzu Hombre Spacecab	小型卡车	18 820	41
Mazda B4000 SE Cab Plus	小型卡车	23 050	51
Nissan Frontier XE Regular Cab	小型卡车	12 110	51
Toyota Tacoma Xtracab	小型卡车	18 228	49
Toyota Tacoma Xtracab V6	小型卡车	19 318	50
Chevrolet K2500	大卡车	24 417	60
Chevrolet Silverado 2500 Ext	大卡车	24 140	64
Dodge Ram 1500	大卡车	17 460	54
Dodge Ram Quad Cab 2500	大卡车	32 770	63
Dodge Ram Regular Cab 2500	大卡车	23 140	59
Ford F150 XL	大卡车	22 875	58
Ford F-350 Super Duty Crew Cab XL	大卡车	34 295	64
GMC New Sierra 1500Ext Cab	大卡车	27 089	68
Toyota Tundra Access Cab Limited	大卡车	25 605	53
Toyota Tundra Regular Cab	大卡车	15 835	58

a. 建立一个估计回归方程，在已知建议零售价格的情况下预测转让销售率。在 $\alpha=0.05$ 的显著性水平下，检测统计显著性。

b. (a)中所建立的估计回归方程是否对数据给出了一个良好的拟合？解释原因。

c. 建立一个估计回归方程，在已知建议零售价格和车型的情况下预测转让销售率。

d. 利用 F 检验来确定回归结果的显著性，在 $\alpha=0.05$ 的显著性水平下，可以得出什么结论？

案例研究 1　消费者调查有限公司

消费者调查有限公司是一个独立的中介机构，它为很多不同的企业调查消费者意向和消费者行为。在一项调查中，某个客户为了预测信用卡用户的支付数额，要求调查消费者的行为特性。下面收集的数据显示了一个由 50 个消费者所组成的样本的年收入、家庭成员的人数以及信用卡年支付数额。统计数据如下页所示。

管理报告

1. 利用描述统计学的方法对这些数据进行整理、总结，并评论统计结果。
2. 建立估计回归方程，首先利用年收入作为自变量，然后再利用家庭人数作为自变量。哪一个变量能够更好地预测信用卡年支付额。
3. 建立一个估计回归方程，以年收入和家庭人数作为自变量，预测信用卡年支付额。对你的

发现进行讨论。

4. 某个家庭,其人数为3人,年收入40 000美元,则这个家庭信用卡年支付额的预测值是多少?
5. 讨论是否需要对这个模型添加其他的自变量。添加什么样的自变量会对模型的分析有所帮助?



CD光盘数据
Consumer

年收入(千美元)	家庭人数	信用卡年支付额(美元)	年收入(千美元)	家庭人数	信用卡年支付额(美元)
54	3	4 016	54	6	5 573
30	2	3 159	30	1	2 583
32	4	5 100	48	2	3 866
50	5	4 742	34	5	3 586
31	2	1 864	67	4	5 037
55	2	4 070	50	2	3 605
37	1	2 731	67	5	5 345
40	2	3 348	55	6	5 370
66	4	4 764	52	2	3 890
51	3	4 110	62	3	4 705
25	3	4 208	64	2	4 157
48	4	4 219	22	3	3 579
27	1	2 477	29	4	3 890
33	2	2 514	39	2	2 972
65	3	4 214	35	1	3 121
63	4	4 965	39	4	4 183
42	6	4 412	54	3	3 730
21	2	2 448	23	6	4 127
44	1	2 995	27	2	2 921
37	5	4171	26	7	4 603
62	6	5678	61	2	4 273
21	3	3623	30	2	3 067
55	7	5301	22	4	3 074
42	2	3020	46	5	4 820
41	7	4828	66	4	5 149

案例研究2　全美橄榄球联盟四分位等级排序

全美橄榄球联盟(NFL)对参赛个人和球队每周的比赛成绩进行记录。这些数据可以通过访问NFL的主页(www. nfl. com)获得。对很多球迷而言，最有意义的统计资料莫过于四分位传球比赛表现的排名。下面的四项指标可以用于编辑四分位传球比赛表现的排名:

1. 每次进攻传球触地的比例
2. 每次进攻传球到位的比例
3. 每次进攻抢断成功的比例

4. 每次进攻平均前进码数

为了说明怎么计算等级排名：我们可以考察 Steve Young，1997 年 NFL 最优秀的四分位球选手的比赛表现。在 1997 年赛季，Steve Young 组织进攻 356 次传球，完成 241 次传球，共计前进 3 029 码，并且 19 次触地，6 次抢断。通过以下 5 步可以计算 Steve Young 传球等级：

步骤 1. 计算触地次数(19)所占组织进攻传球(356)的比例，所得比 19/356 = 0.053 4。将这个结果除以 0.05 得到等级排序的触地分数，这个分数是 0.0534/0.05 = 1.068 0。

步骤 2. 计算传球成功的次数(241)占组织进攻传球次数(356)的比率，所得比率为 241/356 = 0.677 0。将所得到的结果减去 0.3，然后再除以 0.2，可得等级排序的第二个分值，即：(0.677 0 − 0.3)/0.2 = 1.885 0。

步骤 3. 计算抢断次数(6)所占组织进攻传球次数(356)的比率，所得比率为 6/356 = 0.016 9。用 0.095 减去所得到的结果，然后再除以 0.04，可得等级排序的抢断分值，即：(0.095 − 0.016 9)/0.04 = 1.952 5。

步骤 4. 计算传球前进的码数(3 029)占组织进攻传球次数(356)的比率，所得比率为 3 029/356 = 8.508 4。将所得到的结果减去 3，然后再除以 4，可得等级排序的传球前进码数分值，即：(8.508 4 − 3)/4 = 1.377 1。

步骤 5. 将第 1 步到第 4 步的分值加起来，再乘以 100 除以 6，所得总和为 1.068 0 + 1.885 0 + 1.952 5 + 1.377 1 = 6.282 6。

然后将所得结果乘以 100，再除以 6 即可得到 104.710 0 或者 104.7；这个数值就是 NFL 所记录的关于 Steve Young 传球到位率等级排序的成绩。


CD 光盘数据 NFL

由 NFL 所记录的关于 1997 年赛季的统计数据资料位于 NFL 数据集合中，每一列的标题定义如下：

Att	组织进攻传球的次数。
Comp	传球成功的次数。
Comp%	传球成功的次数除以组织进攻传球的次数，再乘以 100。
Yds	传球过程中前进的码数。
Yds/ Att	传球过程中前进的码数除以组织进攻传球的次数。
TD	传球触地的次数。
TD%	传球触地的次数除以组织进攻传球的次数，再乘以 100。
Long	传球成功的最长距离。
Int	抢断次数。
Int%	抢断次数除以组织进攻传球的次数。
Rating	四分位球传球到位等级排序。

管理报告

1. 利用描述统计学的方法对这些数据进行整理、总结，并对结果进行评论。
2. 建立一个估计回归方程，在已知 Comp% 数值的情况下预测 Int%。并对结果进行评论。
3. 建立一个估计回归方程来预测四分位球传球到位等级排序。对你的预测结果进行讨论，并评论你所建立的估计回归方程与前面所给出的计算四分位球等级排序方法有什么样的关系。
4. 如果 NFL 招聘你来计算明年的四分位传球排序等级，你将会使用什么样的方法呢？请解释原因。

案例研究 3　对学生测验成绩进行预测

为了预测一个学区学生的学习成绩和收入高低之间的关系，*The Cincinnati Enquirer* 收集了俄亥俄州教育管理服务局和俄亥俄州税务部门(*The Cincinnati Enquirer*, November 30, 1997)的相关数据。首先，他们从报纸上获得了 1996 年上半年 4 年级、6 年级、9 年级以及 12 年级的数学、阅读理解、自然科学、写作和公民考试的成绩。通过将这些数据进行综合分析，他们得出了每一个学区通过测验的学生比例。

他们同时也记录了学区的学生中，享受抚养儿童补助(*ADC*)的学生比例、享受免费午餐或者降价午餐的学生比例，以及学区中等家庭的收入等资料。所收集的关于 608 个学区的部分数据如下所示：


CD 光盘数据
Enquirer

排名	学区	县	通过比例(%)	ADC(%)	免费午餐(%)	中等家庭收入(美元)
1	Ottawa Hills Local	Lucas	93.85	0.11	0.00	4 8231
2	Wyoming City	Hamilton	93.08	2.95	4.59	4 2672
3	Oakwood City	Montgomery	92.92	0.20	0.38	4 2403
4	Madeira City	Hamilton	92.37	1.50	4.83	3 2889
5	Indian Hill Ex Vill	Hamilton	91.77	1.23	2.70	4 4135
6	Solon City	Cuyahoga	90.77	0.68	2.24	3 4993
7	Chagrin Falls Ex Vill	Cuyahoga	89.89	0.47	0.44	3 8921
8	Mariemont City	Hamilton	89.80	3.00	2.97	3 1823
9	Upper Arlington City	Franklin	89.77	0.24	0.92	3 8358
10	Granville Ex Vil	Licking	89.22	1.14	0.00	3 6235
…	…	…	…	…	…	…

数据的排名是根据“通过比例(%)”数字的大小来排列的，这一列的数据即为通过测验的学生的总体比例。以“ADC(%)”为列标题的数据是每个学区享受抚养儿童补助的学生比例，以“免费午餐(%)”为列标题的数据是每个学区享受免费午餐或者降价午餐的学生比例，以“中等家庭收入”为列标题的数据表示的是每个学区中等家庭的收入。数据还给出了每一个学区所属的县。请注意，在某些情况下，“免费午餐(%)”的数据为 0，这表明这个学区没有学生参与享受免费午餐或者降价午餐计划。

管理报告

利用本章和前面章节所介绍的方法来分析这个数据集合，请在管理报告中简要介绍一下你的分析结果，包括关键的统计指标结果、结论以及建议。你所需要的技术资料可以在附录里面查找。

案例研究 4　Alumni 基金会

Alumni 基金是学院和大学预算的一项重要来源。如果学校的管理人员能够按照该基金的要求提高某些办学指标，则这个基金将会加大对这所学校的捐赠力度，从而提高该校的预算。研究表明如果学生对他们的老师的教育很满意的话，学生就会更加容易顺利毕业。因此，如果班级人数减少，学生与老师的人数比例降低的话，学生的毕业率会更高，因而也会导致 Alumni 基金加大对这所学校的投入。在表 15.11 中，标题“毕业率”表示的是这所大学招收并毕业的学生人数。标题“人数在 20 以下的班级比例”表示的是班级人数在 20 以下的班级所占班级总数的比例。“学生和老师比”表示的是所招收的学生人数除以全体老师的人数。最后的“校友捐赠率”表示的是这个基金所捐赠的金额占学校预算的比例。

管理报告

1. 利用描述统计学的方法对这些数据进行整理、总结。
2. 建立一个估计回归方程，在已知毕业学生数量的情况下预测基金的捐赠率。并对结果进行评论。
3. 建立一个估计回归方程，根据这些数据预测基金的捐赠率。
4. 从你的分析中可以得出什么结论或者建议？

表 15.11 关于美国 48 所州立大学的数据


CD 光盘数据
Alumni

	州	毕业率	人数在 20 以下的班级比例(%)	学生和老师比	校友捐赠率(%)
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve Univ.	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	PA	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Inst. of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State Univ.	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
U. of California – Berkeley	CA	83	58	17	18
U. of California – Davis	CA	74	32	19	7
U. of California – Irvine	CA	74	42	20	9
U. of California – Los Angeles	CA	78	41	18	13
U. of California – San Diego	CA	80	48	19	8
U. of California – Santa Barbara	CA	70	45	20	12
U. of Chicago	IL	84	65	4	36
U. of Florida	FL	67	31	23	19
U. of Illinois – Urbana Champaign	IL	77	29	15	23
U. of Michigan – Ann Arbor	MI	83	51	15	13
U. of North Carolina – Chapel Hill	NC	82	40	16	26
U. of Notre Dame	IN	94	53	13	49
U. of Pennsylvania	PA	90	65	7	41
U. of Rochester	NY	76	63	10	23
U. of Southern California	CA	70	53	13	22
U. of Texas – Austin	TX	66	39	21	13
U. of Virginia	VA	92	44	13	28
U. of Washington	WA	70	37	12	12
U. of Wisconsin – Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University – St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

16 回归分析：建立模型

统计实例：孟山都公司

16.1 一般线性模型

模拟曲线关系

交互作用

包含因变量的转换

内涵线性的非线性模型

16.2 增加或者删除变量的时间确定

一般情形

p - 值的应用

16.3 大型问题分析

16.4 变量选择方法

逐步回归

前向选择

后向消元

最佳子集回归

最终选择

16.5 残差分析

自相关性和杜宾 - 瓦特森检验

16.6 关于方差分析和试验设计的多元回归方法

统计实例

孟山都公司*

圣路易斯，密苏里州

孟山都(Monsanto)公司的历史可以追溯到一位创业者500美元的投资和一间尘土飞扬的仓库，在那里，John F. Queeney 于1901年开始生产糖精，如今，孟山都公司已经成为全国最大的化学公司之一，生产从工业用化学品到现代体育馆用的合成运动地面在内的一千多种产品。孟山都公司作为一个跨国公司，其生产设备、实验室、技术中心和营销机构遍布65个国家。

孟山都公司的营养化工部生产和销售一种氨基酸补充物用于饲养禽类及猪牛等家畜，由于家禽饲养者的工作量大，但是边际效益低，因此他们需要高效而且尽可能富含养分的家禽饲料。对于一定水平的饲料消化水准，高效的养分组成能够使肉鸡生长迅速，体重迅速增加。为了优化家禽饲料，化工企业与家禽饲养者紧密合作，最终的成功取决于能否将家禽的价格保持在低于牛肉和其他肉类的价格水平。

孟山都公司利用回归分析将肉鸡体重 y 与家禽饲料中所添加的氨基酸数量 x 之间的关系建立了模型，最初所建立的简单线性估计回归方程式如下所示：

$$\hat{y} = 0.21 + 0.42x$$

这个估计回归方程式证明了统计的重要性，然而，对残差的分析表明曲线关系可以更好地模拟肉鸡体重和氨基酸之间的关系。



Monsanto 的科研工作者利用回归分析为家禽饲养开发最优化的饲料养分组成。© PhotoDisc，Inc.

进一步的研究表明，虽然少量的氨基酸能够增进家禽体重，但是超过某一限度时，体重将减小，多余的氨基酸量对于家禽体重的增加仅有微弱的或者根本没有作用，事实上，当氨基酸的增长超过营养需要的极限时，体重反而下降。下面的估计多元回归方程可以模拟家禽体重和氨基酸数量之间的曲线关系。

$$\hat{y} = -1.89 + 1.32x - 0.506x^2$$

回归分析的结果应用使得孟山都公司确定了家禽饲料产品中最理想的氨基酸含量水平。

在本章中，我们将通过介绍如何建立曲线模型，比如孟山都公司所应用的那种，来扩展回归分析的讨论，除此之外，我们还将描述一系列方法，这些方法有助于确定哪一个自变量会导致最优的估计回归方程。

* 作者衷心感谢 Monsanto 营养化工部的资深研究专家 James R. Ryland 和 Robert M. Schisla 提供这一统计实例。

建模是一个建立估计回归方程的过程，经过这一过程，我们可以得到描述一个因变量和一个或者多个自变量之间关系的估计回归方程。建模的主要结果应该是找到合适的函数形式来描述变

量之间的关系，并且选择该模型所应包含的自变量。在 16.1 节中，我们通过引入一般线性模型的概念确立了建模过程的基本框架。16.2 节则介绍了确定什么时候增加或者删除自变量的一般性方法，为更加复杂的以计算机为基础的处理方法提供了基础。在 16.3 节中，我们考察了一个涉及到 8 个自变量和 25 组观测值的更为大型的回归问题，该问题被用来解释 16.4 节中的变量选择步骤，包括逐步回归过程、前进选择过程、后向消元过程和最佳子集回归。在 16.5 节中，我们阐述了如何利用杜宾 - 瓦特森(Durbin-Watson)检验来检测序列相关或者自相关。在 16.6 节中，我们阐述了如何利用回归分析来进行方差分析和处理试验设计问题。

16.1　一般线性模型

假设我们已经收集了针对因变量 y 和 k 个自变量 x_1，x_2，…，x_k 的数据，我们的目标是利用这些数据建立一个估计回归方程，用它来给出因变量和自变量之间的最佳关系。我们引入了包含 p 个自变量的**一般线性模型**(general linear model)概念，作为在自变量之间建立更加复杂关系的一般性框架。

如果可以将回归模型写成公式(16.1)的形式，则我们可以应用第 15 章中所述的标准多元回归分析方法。

一般线性模型

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p + \varepsilon \qquad (16.1)$$

在公式(16.1)中，每一个因变量 z_j(其中 $j=1$，2，…，p)均为 x_1，x_2，…，x_k(所采集的数据即为这些变量)的函数。在某些情况下，每一个 z_j 有可能仅为一个 x 变量的函数。最简单的情形即只对一个变量 x_1 收集数据，并且利用一个直线线性关系去估计 y。这种情况下，$z_1 = x_1$ 并且公式(16.1)变成如下形式：

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \qquad (16.2)$$

公式(16.2)即为第 14 章中介绍的简单线性回归模型，不同的是自变量是以 x_1 表示而非 x。在统计模型文献中，这一模型被称为**含单个预测变量的简单一阶模型**。

模拟曲线关系



CD 光盘数据
Reynolds

我们可以用公式(16.1)来模拟形式更加复杂的关系。为了说明这一点，我们可以参考 Reynolds 公司(工业用天平和试验设备生产商)所面临的问题，该公司的管理者计划对公司的销售人员工作年限长短和电子试验设备的销售量之间的关系进行研究。表 16.1 给出了随机抽取的 15 名销售人员在最近销售期所销售的天平数以及每个销售人

员被公司所雇用的年限。图 16.1 给出的是这些数据点的散点图。该图表明，在雇佣年限的长短和销售量之间可能存在曲线关系。在考虑如何为 Reynolds 公司建立曲线关系模型之前，让我们首先考虑图 16.2 中相对应于简单一阶模型的 Minitab 计算机输出结果，估计回归方程如下所示：

$$\text{Sales} = 111 + 2.38\text{Months}$$

式中 Sales—— 电子试验天平的销售量(销售额)；

Months——销售人员被雇用的月数(雇佣月数)。

表 16.1 雷罗兹公司示例数据

雇佣月数	销售额
41	375
106	296
76	317
10	376
22	162
12	150
85	367
111	308
40	189
51	235
9	83
12	112
6	67
56	325
19	189

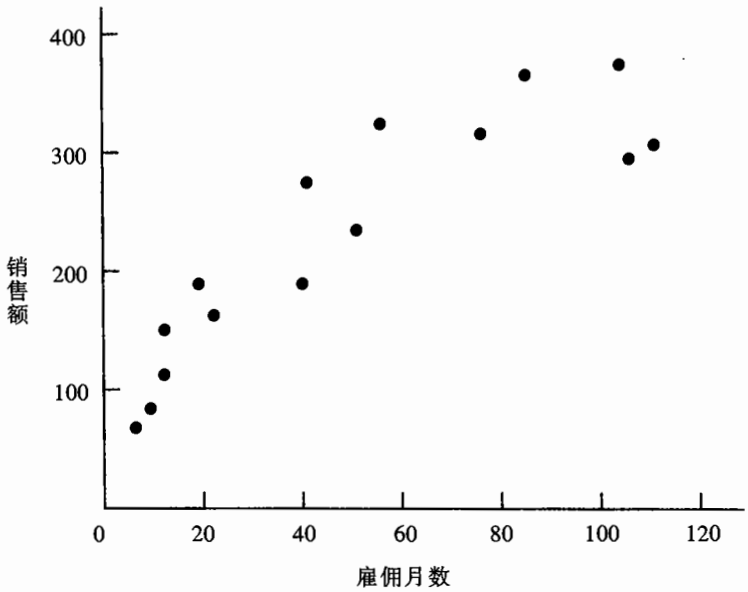


图 16.1 雷罗兹公司示例的散点图

```
The regression equation is
Sales = 111 + 2.38 Months

Predictor       Coef      Stdev    t-ratio       p
Constant      111.23      21.63       5.14   0.000
Months        2.3768     0.3489       6.81   0.000

s = 49.52       R-sq = 78.1%     R-sq(adj) = 76.4%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      113783      113783      46.41   0.000
Error        13       31874        2452
Total        14      145657
```

图 16.2 雷罗兹公司例子的 Minitab 计算机输出结果

图 16.3 是相应的标准化残差图。虽然计算机输出结果表明存在显著关系(p - 值 = 0.000)，而且线性关系解释了大部分的销售量变异性(R - sq = 78.1%)，但是标准残差图依然表明我们有必要研究曲线关系。

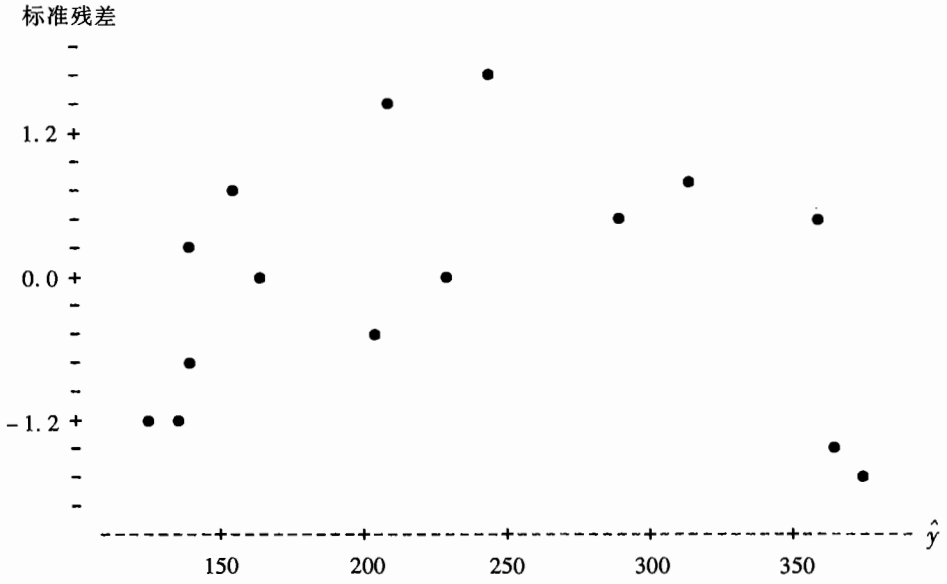


图 16.3　雷罗兹公司示例的标准残差图：一阶模型

为了表明曲线关系的存在，我们令公式(16.1)中的 $z_1 = x_1$, $z_2 = x_1^2$, 于是得到模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \qquad \textbf{(16.3)}$$

这个模型被称为**含一个预测变量的二阶模型**。为了建立与这个二阶模型所对应的估计回归方程，我们所使用的统计软件包需要表 16.1 中的原始数据，同时也需要添加第二个自变量所对应的数据，即销售人员被公司所雇用月数的平方。在图 16.4 中，我们给出了对于二阶模型的 Minitab 输出结果，估计回归方程如下：

$$\text{Sales} = 45.3 + 6.34\text{Months} - 0.034\,5\text{MonthsSq}$$

式中，MonthsSq 为销售人员被公司雇用的月数的平方。

MonthsSq 自变量的数据是将 Months 平方而得。

图 16.5 是相应的标准残差图，前面的曲线已经被移除。在 0.05 的显著性水平下，计算机输出结果表明总体模型是显著的(F 检验的 p - 值为 0.000)；同时我们还注意到对应于 MonthsSq (p - 值为 0.002)的 t 检验的 p - 值比 0.05 要小，因此，我们可以得出结论：在含变量 Months 的模型中所添加的变量 MonthsSq 是显著的。因为 R - sq(adj)的值为 88.6%，所以我们应该对这个估计回归方程所提供的拟合感到高兴。然而，更重要的是我们看到了在回归分析中处理曲线关系

```
The regression equation is
Sales = 45.3 + 6.34 Months - 0.0345 MonthsSq

Predictor       Coef        Stdev    t-ratio        p
Constant       45.35        22.77       1.99    0.070
Months         6.345        1.058       6.00    0.000
MonthsSq   -0.034486     0.008948      -3.85    0.002

s = 34.45       R-sq = 90.2%     R-sq(adj) = 88.6%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     2      131413       65707     55.36    0.000
Error         12       14244        1187
Total         14      145657
```

图 16.4 雷罗兹公司示例的 MINITAB 输出结果：二阶模型

是多么轻松的一件工作。

显然，许多类型的关系均可以通过公式(16.1)来模拟，我们在工作中所使用的回归方法并不仅仅局限于线性或者直线关系。在多元回归分析中，线性这个词在术语“一般线性模型”中仅指这样一种事实：β_0，β_1，…，β_p 均为一次幂，它并不意味着 y 和 x_i 之间呈线性关系。事实上，在本节中，我们已经看到公式(16.1)是怎样被用于模拟曲线关系的一个例子。

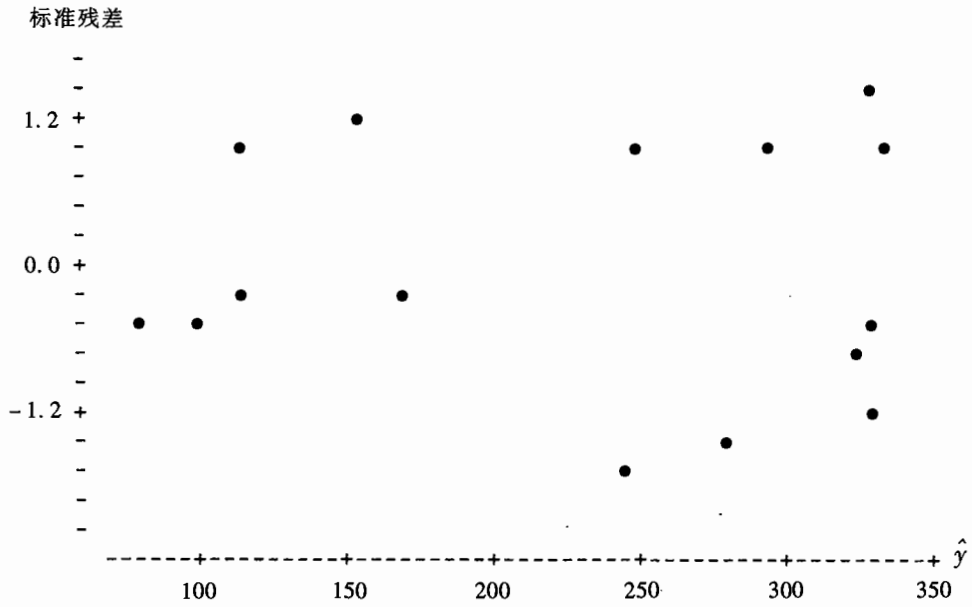


图 16.5 雷罗兹公司示例的标准残差散点图：二阶模型

交互作用

如果原始数据集合由因变量 y 和两个自变量 x_1、x_2 的观测值组成，在一般线性模型公式(16.1)中，我们令 $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$并且 $z_5 = x_1 x_2$, 这样就可以建立一个含两个自变量的二阶模型，所得到的模型如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon \qquad (16.4)$$

在这个二阶模型中，为了说明两个变量相互作用所产生的影响，我们添加了一个变量 $z_5 = x_1 x_2$。这种影响被称为**交互作用**(interaction)。

为了说明交互作用以及它的意义，让我们来考察一下 Tyler Personal Care 公司对它的一种新型洗发产品所作的回归研究。他们认为对销售量影响最大的两个因素是单位销售价格和广告支出。为了调查这两个变量对销售的影响，在 24 家试验商店中，价格为 2.00 美元、2.50 美元和 3.00 美元的产品广告费用支出分别为 50 000 美元和 100 000 美元，我们将所观测到的销售量(单位千瓶)记录于表 16.2 中。

表 16.2　Tyler Personal Care 公司例子的数据



CD 光盘数据
Tyler

价格(美元)	广告支出(千美元)	销售量(千瓶)	价格(美元)	广告支出(千美元)	销售量(千瓶)
2.00	50	478	2.00	100	810
2.50	50	373	2.50	100	653
3.00	50	335	3.00	100	345
2.00	50	473	2.00	100	832
2.50	50	358	2.50	100	641
3.00	50	329	3.00	100	372
2.00	50	456	2.00	100	800
2.50	50	360	2.50	100	620
3.00	50	322	3.00	100	390
2.00	50	437	2.00	100	790
2.50	50	365	2.50	100	670
3.00	50	342	3.00	100	393

表 16.3 是这些数据的总结。请注意：对应销售价格为 2.00 美元，广告费用为 50 000 美元的平均销售数量为 461 000 瓶，对应销售价格为 2.00 美元，广告费用为 100 000 美元的平均销售量为 808 000 瓶。因此，在销售价格保持 2.00 美元不变时，广告费用 50 000 美元和 100 000 美元所造成的平均销量差是 808 000 − 461 000 = 347 000 瓶。当产品销售价格为 2.50 美元时，平均销售量之差为 646 000 − 364 000 = 282 000 瓶。最后，当产品销售价格为 3.00 美元时，平均销售量之

差为 375 000 - 332 000 = 43 000 瓶。显然，广告费用为 50 000 美元和 100 000 美元时，平均销售量之差取决于产品销售价格。换句话说，当销售价格处于较高价位时，广告费用支出的增加所带来的销量增加的影响将减小。这些观察数据提供了价格和广告费用支出这两个变量交互作用的证据。

表 16.3 Tyler Personal Care 公司的平均销售量(千瓶)

	价格		
	2.00(美元)	**2.50(美元)**	**3.00(美元)**
广告支出 50 000 美元	461	364	332
广告支出 100 000 美元	808	646	375

当价格 = 2.00 美元、广告支出 = 100 000 美元时，平均销量为 808 000 瓶。

为了给出交互作用的另一方面影响，图 16.6 显示了 6 种不同的销售价格—广告费用组合的平均销售量。该图同时还表明了广告费用支出对平均销售量的影响依赖于价格水平；我们再一次看到了交互作用的影响。当两个变量之间存在交互作用时，我们不能排除其他变量而去单独考察某一个变量对因变量 y 的影响。换句话说，只有考虑两个变量对因变量的联合影响时，我们才能获得有意义的结论。

为了说明交互作用的影响，我们可以利用下面的回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{16.5}$$

式中 y——销售量(千瓶)；

x_1——销售价格(美元)；

x_2——广告费用支出(千美元)。

请注意，公式(16.5)反映了 Tyler 公司相信销售量和销售价格、广告支出线性相关(由 $\beta_1 x_1$，$\beta_2 x_2$ 项即可说明)，而且这两个变量之间存在交互作用(由 $\beta_3 x_1 x_2$ 项说明)。

为了建立估计回归方程，我们可以利用含 3 个自变量(z_1，z_2，z_3)的一般线性模型：

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \varepsilon \tag{16.6}$$

式中 $z_1 = x_1$；

$z_2 = x_2$；

$z_3 = x_1 x_2$。

图 16.7 是对应于 Tyler 个人日用品公司例子交互作用模型的 Minitab 输出结果，所得到的估计回归方程：Sales = -276 + 175 Price + 19.7 AdvExp - 6.08 PriceAdv

式中 Sales——销售量(千瓶)；

Price——产品销售价格(美元)；

AdvExp ——广告费用支出(千美元)；

PriceAdv——交互作用项(价格 × 广告支出)。

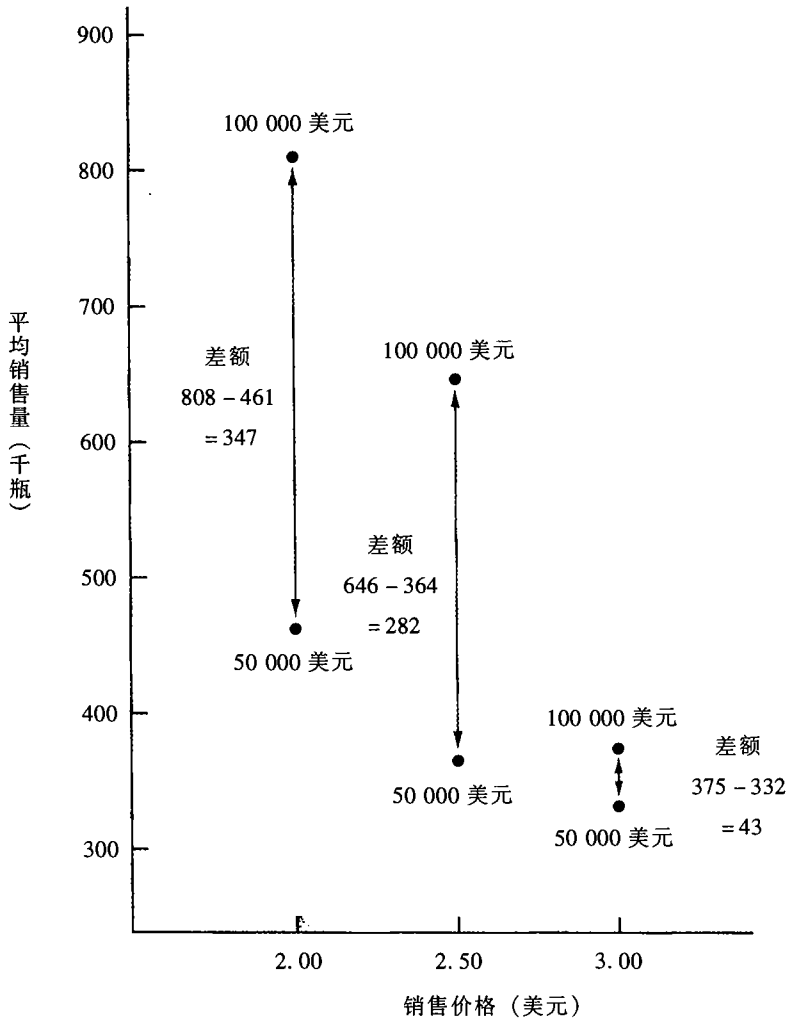


图 16.6　平均销量是销售价格和广告费用支出的函数

因为对应于交互作用项 PriceAdv 的 t 检验的 p - 值为 0.000，我们可以得出结论：在产品价格和广告支出的线性影响下交互作用是显著的。因此，回归结果显示广告费用支出对销售量的影响取决于价格。

包含因变量的转换

在说明怎样利用一般线性模型模拟自变量和因变量之间各种可能的不同类型的关系时，我们致力研究涉及到一个或多个自变量的转换，通常考虑包含因变量 y 的转换问题是有价值的。对于什么时候应该变换因变量的解释，我们可以考虑表 16.4 中关于 12 种汽车的英里/加仑定额和重量的数

```
The regression equation is
Sales = - 276 + 175 Price + 19.7 AdvExp - 6.08 PriceAdv

Predictor       Coef        Stdev     t-ratio        p
Constant      -275.8        112.8       -2.44    0.024
Price         175.00        44.55        3.93    0.001
Adver         19.680        1.427       13.79    0.000
PriceAdv     -6.0800       0.5635      -10.79    0.000

s = 28.17       R-sq = 97.8%      R-sq(adj) = 97.5%

Analysis of Variance

SOURCE        DF          SS          MS          F        p
Regression     3      709316      236439     297.87    0.000
Error         20       15875         794
Total         23      725191
```

图 16.7　Tyler Personal Care 公司例子的 Minitab 输出结果



CD 光盘数据
MPG

据。散点图 16.8 表明了在这两个变量之间存在负线性关系，所以我们利用一个简单一阶模型将这两个变量联系起来。图 16.9 是 Minitab 输出结果，所得到的估计回归方程如下：

$$\text{MPG} = 56.1 - 0.011\,6\ \text{Weight}$$

式中　MPG　——英里/加仑定额；

Weight——汽车的重量(磅)。

表 16.4　12 种汽车的英里/加仑定额和重量

重量	英里/加仑定额
2 289	28.7
2 113	29.2
2 180	34.2
2 448	27.9
2 026	33.3
2 702	26.4
2 657	23.9
2 106	30.5
3 226	18.1
3 213	19.5
3 607	14.3
2 888	20.9

这个模型是显著的(F 检验的 p - 值为 0.000)。而且拟合度也很优良(R - sq = 93.5%)。但是我们注意到图 16.9 中第 3 个观测值有较大标准残差。

图 16.10 是对应于一阶模型的标准残差图。如果关于误差项的假设成立，那我们所希望看到的应该是一条水平带状图，但实际所观测到的并非这样。相反，残差的变异性看起来随着 $\hat{y}$ 的增加而增加。换句话说，我们所看到是在第 14 章和第 15 章中被认为是存在非常数方差标志的楔型图。由于关于显著性检验的基本假设看来并未得到满足，因此我们缺乏足够的理由推出关于估计回归方程统计显著性的任何结论。

通常非常数方差问题是可以通过对因变量做一个不同比例变换来修正的。例如，如果我们用因变量的对数来替换原来的因变量，这样做的效果将是压缩因变量的值，从而减小了非常数方差的影响。大多数的统计软件包均提供对数转换的功能，或者以 10 为底(常用对数)，或者以 $e = 2.718\,28\cdots$(自然对数)为底，等等。我们对英里/加仑定额数据应用了自然对数转换，并且建立了英里/加仑定额的自然对数相关于重量的估计回归方程。图 16.11

显示了以英里/加仑定额的自然对数作自变量，以 LogeMPG 作输出标签的回归结果。图 16.12 为相对应的标准化残差图。

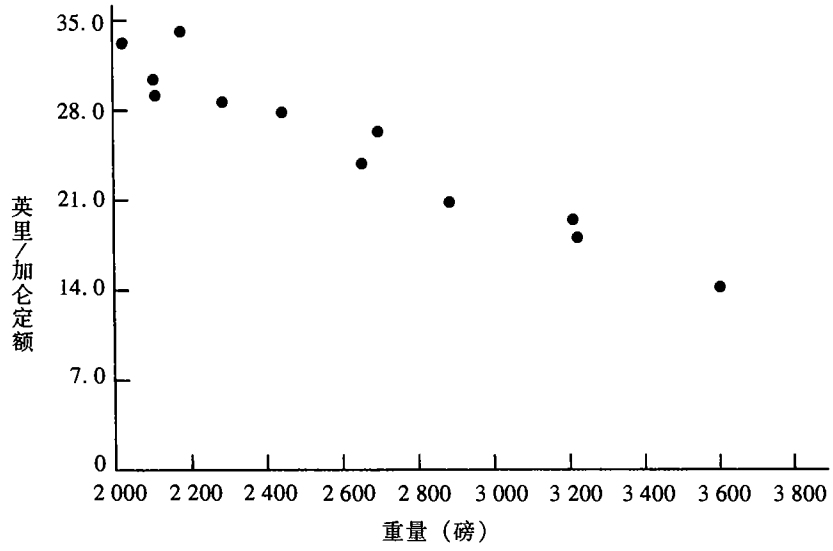


图 16.8　英里/加仑问题的散点图

```
The regression equation is
MPG = 56.1 - 0.0116 Weight

Predictor        Coef        Stdev    t-ratio        p
Constant       56.096        2.582      21.72    0.000
Weight    -0.0116436    0.0009677     -12.03    0.000

s = 1.671        R-sq = 93.5%     R-sq(adj) = 92.9%

Analysis of Variance

SOURCE        DF          SS          MS          F        p
Regression     1      403.98      403.98     144.76    0.000
Error         10       27.91        2.79
Total         11      431.88

Unusual Observations
Obs.   Weight        MPG        Fit Stdev.Fit   Residual   St.Resid
  3     2180     34.200     30.713     0.644      3.487      2.26R

R denotes an obs. with a large st. resid.
```

图 16.9　英里/加仑定额问题的 Minitab 输出结果

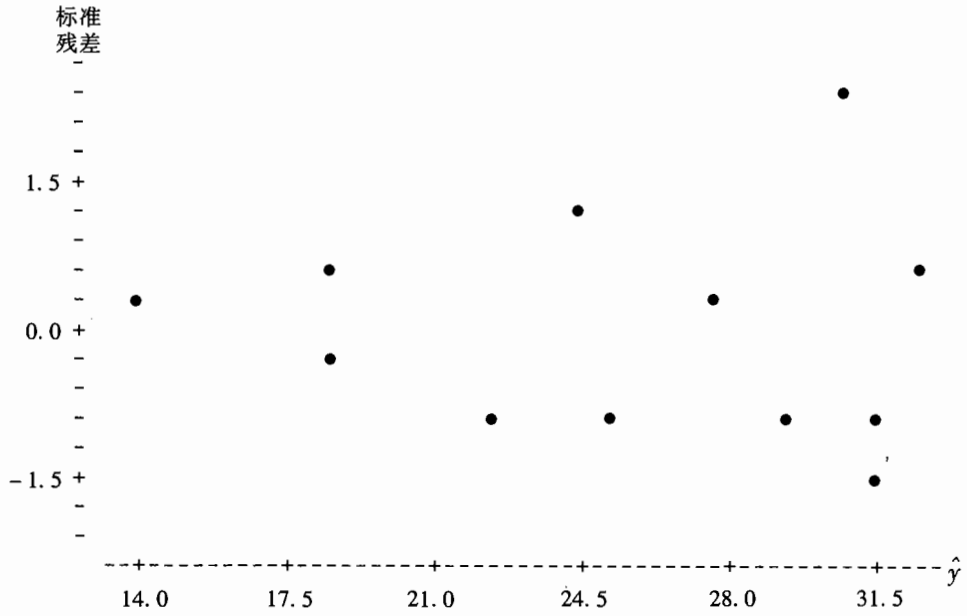


图 16.10 英里/加仑定额问题的标准残差图

```
The regression equation is
LogeMPG = 4.52 -0.000501 Weight

Predictor        Coef        Stdev    t-ratio        p
Constant      4.52423      0.09932      45.55    0.000
Weight    -0.00050110   0.00003722     -13.46    0.000

s = 0.06425      R-sq = 94.8%      R-sq(adj) = 94.2%

Analysis of Variance

SOURCE        DF          SS          MS          F        p
Regression     1     0.74822     0.74822     181.22    0.000
Error         10     0.04129     0.00413
Total         11     0.78950
```

图 16.11 英里/加仑定额问题的 Minitab 输出结果：对数转换

请看图 16.12 中的残差散点图，我们现在看到楔形图已经消失了。除此之外，没有一个观测值具备较大的标准化残差。以英里/加仑定额的对数作因变量的模型在统计上是显著的，并且对观测数据提供了一个优良拟合。因此，我们可以推荐使用估计回归方程：

$$\text{LogeMPG} = 4.52 - 0.000\,501\text{Weight}$$

为了估计重 2 500 磅的汽车英里/加仑的定额，我们首先求出一个英里/加仑定额的对数估计值：

$$\text{LogeMPG} = 4.52 - 0.000\,501(2\,500) = 3.267\,5$$

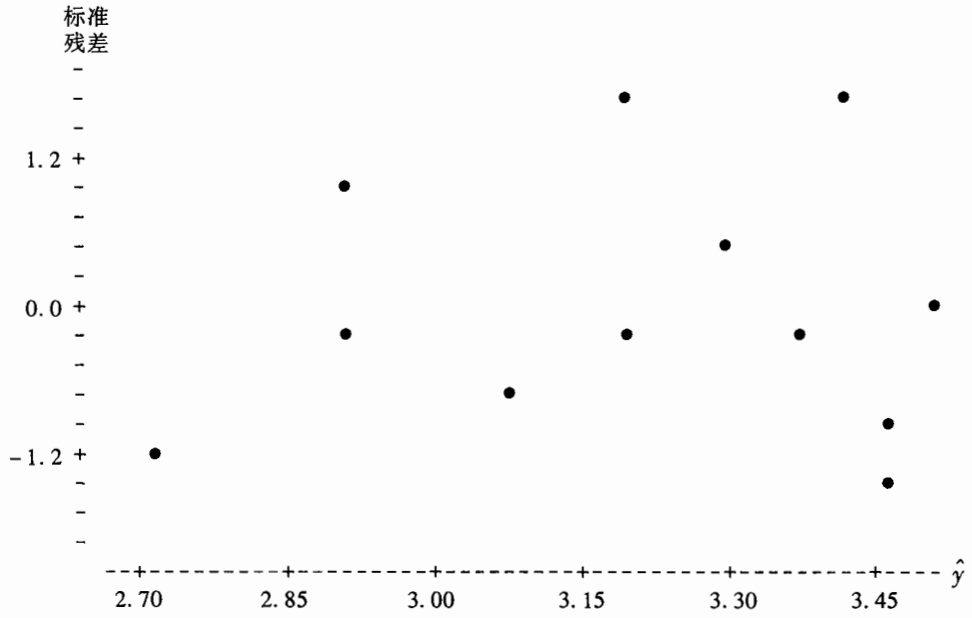


图 16.12　英里/加仑定额问题的标准残差图：对数转换

通过求自然对数值为 3.267 5 的数值，我们可以得到英里/加仑定额的估计。利用带指数功能的计算器，或者计算 e 的 3.267 5 次幂，我们得到定额为 26.2 英里每加仑。

解决非常数方差问题的另一个途径是用 $1/y$ 来代替 y 作因变量。这种类型的变换被称为**倒数变换**。例如，如果因变量的单位是英里/加仑，那么倒数变换将得到新的因变量，这个新的因变量的单位是 1/(英里/加仑)或加仑/英里。通常说来，如果不实际尝试一下，对数变换和倒数变换这两种方法是难以区分优劣的。

内涵线性的非线性模型

模型里面的参数(β_0，β_1，…，β_p)的幂如果超过一次的，则这种模型被称为非线性模型。然而，对指数模型而言，我们可以通过对变量进行变换将模型转换成公式(16.1)的一般线性模型形式，这样我们就可以对它进行回归分析。指数模型与下面的回归方程有关：

$$E(y) = \beta_0 \beta_1^x \tag{16.7}$$

这种模型适合如下情形：即因变量 y 随着 x 的增加以一个固定百分比增加或减少，而不是以一个固定数量增加或减少。

举个例子，假设某种产品的销售量 y 取决于广告费用 x(千美元)，该问题所对应的指数模型如下：

$$E(y) = 500(1.2)^x$$

于是，对 $x=1$，$E(y)=500(1.2)^1=600$；对 $x=2$，$E(y)=500\ (1.2)^2=720$；对 $x=3$，$E(y)=$

500 $(1.2)^3 = 864$。请注意，在这种情况下，$E(y)$不是按一个固定数量增加，而是按一个固定百分比增加：所增加的百分比是20%。

我们可以通过对公式(16.7)两边取对数，将这个非线性模型转化成一个线性模型

$$\log E(y) = \log\beta_0 + x\log\beta_1 \tag{16.8}$$

现在如果我们令 $y' = \log E(y)$，$\beta_0' = \log\beta_0$，$\beta_1' = \log\beta_1$，我们将公式(16.8)改写成

$$y' = \beta_0' + \beta_1' x$$

显然，我们可以利用简单线性回归公式得出 β_0'和 β_1'的估计量。以 b_0'和 b_1'表示估计量，可得下面估计的回归方程：

$$\hat{y}' = b_0' + b_1' x \tag{16.9}$$

为了在已知 x 值的情况下求得原来的因变量 y 的预测值，我们可以先将 x 的值代入公式(16.9)计算$\hat{y}'$，$\hat{y}'$的反对数即为我们所要求的 y 的观测值或 y 的期望值。

许多非线性模型无法被转换成一个等价的线性模型。这类模型已经限制了我们在商业和经济中的应用，而且研究此类模型所需的数学背景超出了本书的讨论范围。

练　习

方　法

1. 变量 x，y 有下列数据：



x	22	24	26	30	35	40
y	12	21	33	35	40	36

a. 建立一个估计回归方程，方程形式为：$\hat{y} = b_0 + b_1 x$。

b. 利用(a)中的结果，在 $\alpha = 0.05$ 的显著性水平下，检验 x 和 y 之间的显著性关系。

c. 画出关于这些数据的散点图，散点图所对应的方程形式是否为$\hat{y} = b_0 + b_1 x + b_2 x^2$ 形式？解释原因。

d. 建立一个估计回归方程，方程形式为：$\hat{y} = b_0 + b_1 x + b_2 x^2$。

e. 考察(d)中的方程，在 $\alpha = 0.05$ 的显著性水平下，x，x^2 和 y 之间的关系在统计上是否显著？

f. 当 $x = 25$ 时，预测 y 的数值。

2. 变量 x，y 有下列数据：

x	9	32	18	15	26
y	10	20	21	16	22

a. 建立一个估计回归方程，方程形式为：$\hat{y} = b_0 + b_1 x$。这个方程对 y 预测的精确性如何？

b. 建立一个估计回归方程，方程形式为：$\hat{y} = b_0 + b_1 x + b_2 x^2$。这个方程对 y 预测的精确性如何？

c. 当 $x = 20$ 时，预测 y 的数值。

3. 变量 x，y 有下列数据：

x	2	3	4	5	7	7	7	8	9
y	4	5	4	6	4	6	9	5	11

a. x 和 y 之间是否存在线性关系？解释原因。

b. 建立关于 x 和 y 的估计回归方程。

c. 画出(b)中方程关于 $\hat{y}$ 的标准残差图，模型的假设是否得到满足？解释原因。

d. 对因变量 y 作对数变换，利用变换以后的因变量建立估计回归方程，利用转换过的因变量是否能够满足模型假设？对于这种情形，倒数转换是否更为适合？解释原因。

应　用

4. 某公路管理部门对交通流量和车辆速度之间的关系进行了研究，所建立的假设模型如下所示：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

式中　y——每小时的车辆交通流量；

x——车速(英里/小时)。

下面的数据是在交通高峰时段进出城市的 6 个公路道口所收集的：

交通流量(y)	车辆速度(x)
1 256	35
1 329	40
1 226	30
1 335	45
1 349	50
1 124	25

a. 对这些数据建立估计回归方程。

b. 在 $\alpha = 0.01$ 的显著性水平下，检验显著性关系。

5. 自测题　对练习 4 中的问题作进一步研究，我们建议利用以下的曲线估计回归方程对统计数据进行分析。

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

a. 利用练习 4 中的数据，估计这个估计回归方程的参数。

b. 在 $\alpha = 0.01$ 的显著性水平下，检验显著性关系。

c. 如果车速为 38 英里每小时，请估计每小时的交通流量。

6. 在一项关于紧急服务交通工具的研究中，我们调查了交通工具的数量和提供紧急服务的平均距离(*Management Science*, July 1988)。下表给出了所收集的数据：

交通工具数量	平均距离(英里)
9	1.66
11	1.12
16	0.83
21	0.62
27	0.51
30	0.47

a. 建立关于这些数据的散点图，将平均距离作为因变量。

b. 简单线性模型是否适于这种情形？解释原因。

c. 对这些数据建立一个估计回归方程，尽可能准确地解释这两个变量之间的关系。

7. 下面的数据是 10 种主打啤酒品牌的媒体广告支出(百万美元)和销售量(百万桶)(*Superbrands' 98*, October 20, 1997)：



CD 光盘数据
Beer

品牌	媒体广告支出(百万美元)	销售量(百万桶)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

a. 建立关于这些数据的散点图，将媒体广告支出作为因变量。

b. 简单线性模型是否适于这种情形？

c. 对这些数据建立一个估计回归方程，尽可能准确地解释这两个变量之间的关系。

8. 在欧洲，每个国家互联网用户的数量都有很大的不同。在 1999 年，44.3% 的瑞典人使用互联网，但是在法国人数就减少了 10%。虽然互联网的应用预计在未来几年内将会有大大地增长，但是各国之间的用户数量差异依然会保持下去。下表显示的是欧洲国家 1999 年互联网用户数量和预计 2005 年的用户数量：



CD 光盘数据
Internet

	1999 年互联网用户数量(%)	预计 2005 年的用户数量(%)
奥地利	12.6	53.4
比利时	24.2	60.2
丹麦	40.4	71.2
芬兰	40.9	71.4
法国	9.7	53.1
德国	15.0	59.6
希腊	3.4	15.8
爱尔兰	12.1	46.7
意大利	8.4	34.4
荷兰	18.6	61.4
挪威	38.0	71.7
葡萄牙	4.63	36.6
西班牙	7.4	39.9
瑞典	44.3	71.9
瑞士	28.1	66.7
英国	23.6	66.8

a. 建立关于这些数据的散点图，将 1999 年互联网用户数量比率作为因变量。简单线性模型是否适于这种情形？对此进行讨论。

b. 建立估计多元回归方程，以 x = 1999 年互联网用户数量和 x^2 作为两个自变量。

c. 考察公式(16.7)所显示的非线性关系，利用对数变换建立关于该模型的估计回归方程。

d. 对于(b)和(c)中所建立的估计回归方程，你认为哪一个更为合适？解释原因。

9. 在一项关于顺风和逆风对高尔夫球飞行距离影响的研究中，Titleist 发现在顺风条件下高尔夫球飞行距离会略为增加；在逆风条件下，风速越大则高尔夫球飞行距离会越小(*Golf Magazine*, March 1997)。假设在持续不断的研究中，我们在各种风速中击打高尔夫球，所得数据如下。在这些数据中，风速以英里每小时测量；负值表示逆风，正值表示顺风，距离表示的是高尔夫球飞行滚动的码数。



CD 光盘数据
WindGolf

风速	距离	风速	距离
-30	202	0	269
-30	211	10	273
-30	206	10	286
-20	221	10	287
-20	239	20	283
-20	234	20	292
-10	240	20	298
-10	245	30	292
-10	260	30	301
0	260	30	299
0	278		

a. 建立关于高尔夫球飞行码数与风速之间关系的估计回归方程。
b. 在逆风 15 英里/小时的情况下，估计高尔夫球飞行和滚动距离。
c. 在顺风 25 英里/小时的情况下，估计高尔夫球飞行和滚动距离。
d. 你的计算结果是否证实了 *Golf Digest** 中所报告的内容？解释原因。

16.2 增加或者删除变量的时间确定

在本节中，我们将阐述如何利用 F 检验来确定将一个或者多个自变量添加到多元回归模型上的利弊。这种检验是根据多元回归模型添加一个或者多个自变量所得到的误差平方和的减少值来衡量的。首先让我们来说明怎样将这种检验应用于前一章所介绍的 Butler 运输公司的例子。

在第 15 章中，我们介绍了 Butler 运输公司的例子来描述多元回归分析的应用。在那个例子中，Butler 运输公司的管理者计划建立估计回归方程，利用两个自变量：行驶里程和运送次数来预测每天行驶的时间。将车辆的行驶里程 x_1 作为惟一的自变量，利用最小二乘法可以得到如下的估计回归方程：

$$\hat{y} = 1.27 + 0.067\,8x_1$$

在第 15 章中，我们已经说明了模型的误差平方和是 SSE = 8.029。如果将变量：运送货物的次数 x_2 作为第 2 个自变量加入模型之中，则我们可以得到如下的估计回归方程：

$$\hat{y} = -0.869 + 0.061\,1x_1 + 0.923x_2$$

这个模型的误差平方和是 SSE = 2.299，很明显，x_2 的添加使误差平方和减少了。我们想要研究的问题在于：将 x_2 加入模型是否会导致误差平方和显著减少？

我们用 SSE(x_1)来表示模型中仅仅含有 x_1 这一个自变量时的误差平方和，用 SSE(x_1，x_2)来表示模型中含有 x_1 和 x_2 两个自变量时的误差平方和，并依此类推。因此，相比仅仅含有 x_1 的模型的误差平方和，添加 x_2 所导致的误差平方和的减少值是

$$\text{SSE}(x_1) - \text{SSE}(x_1, x_2) = 8.029 - 2.299 = 5.730$$

我们可以利用 F 检验来确认这个减少是否显著。

F 统计量的分子是用 SSE 所减少的数值除以模型所增加的自变量个数，在这里仅仅只有 x_2 这一个自变量添加到模型中，所以，F 统计量的分子是

$$\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{1} = 5.730$$

这个结果是对 SSE 相应于模型中每添加一个自变量时的减少值的度量。F 统计量的分母是包含全部自变量的模型的均方误差。对于 Butler 运输公司这个例子，它的模型包含两个自变量 x_1 和 x_2，因此，$p = 2$，并且有

$$\text{MSE} = \frac{\text{SSE}(x_1, x_2)}{n - p - 1} = \frac{2.299}{7} = 0.328\,4$$

* 编者按：此处原书如此，疑为 *Golf Magazine*。

关于自变量 x_2 的增加是否统计显著，我们可以把下面的 F 统计量作为判断的根据：

$$F=\frac{\dfrac{\mathrm{SSE}(x_1)-\mathrm{SSE}(x_1,x_2)}{1}}{\dfrac{\mathrm{SSE}(x_1,x_2)}{n-p-1}} \tag{16.10}$$

F 检验分子的自由度等于模型所添加的自变量的个数，分母的自由度为 $n-p-1$。

对于 Butler 运输公司的例子，我们可以得到

$$F=\frac{\dfrac{5.730}{1}}{\dfrac{2.299}{7}}=\frac{5.730}{0.328\,4}=17.45$$

根据附录 B 的表 4，我们发现在 $\alpha=0.05$ 的显著性水平下，$F_{0.05}=5.59$。由于 $F=17.45>F_{0.05}=5.59$，因此我们可以拒绝零假设，即 x_2 在统计上是不显著的；换句话说，将 x_2 添加到仅仅含有 x_1 的模型中将会导致误差平方和显著减少。

如果希望检验仅仅添加一个自变量到模型上的显著性，则我们可以利用刚才所描述的 F 检验，也可以利用 15.4 节中所介绍的关于单个参数显著性的 t 检验。事实上，我们刚才所计算的 F 统计量是检验单个参数显著性的 t 统计量的平方。

因为在模型仅仅添加一个自变量时，t 检验和 F 检验是等价的，因此我们可以更进一步阐述检验单个参数显著性的 t 检验。如果单个参数不显著，则我们可以从模型中删除相对应的变量。但是，如果 t 检验表明存在两个或者多个参数不显著，则我们从模型中删除的变量永远不能多于 1 个；如果我们从模型中删除一个变量，则一开始并不显著的第二个变量将会变为显著变量。

现在我们开始讨论添加一个以上的自变量，或者一组变量，是否会导致误差平方和显著减少。

一般情形

考察下面含有 q 个自变量的多元回归模型，其中 $q<p$。

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_qx_q+\varepsilon \tag{16.11}$$

如果我们将变量 x_{q+1}，x_{q+2}，…，x_p 添加到这个模型中，即可得到一个含有 p 个自变量的多元回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_qx_q+\beta_{q+1}x_{q+1}+\beta_{q+2}x_{q+2}+\cdots+\beta_px_p+\varepsilon \tag{16.12}$$

为了检验变量 x_{q+1}，x_{q+2}，…，x_p 的添加在统计上是否显著，我们可以提出如下的零假设和备择假设：

H_0: $\beta_{q+1}=\beta_{q+2}=\cdots=\beta_p=0$

H_a：至少有一个参数不等于零

接下来的 F 统计量给出了检验自变量的添加在统计意义上是否显著的依据：

$$F=\frac{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q)-\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_p)}{p-q}}{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_p)}{n-p-1}} \tag{16.13}$$

许多诸如 Minitab 之类的计算软件包给出了按顺序进入模型的每个自变量所增加的平方和，在这种情况下，我们可以简化用来确定是否增加或者删除一组变量的 F 检验计算。

然后将所计算出的 F 统计量与分子自由度为 $p-q$、分母自由度为 $n-p-1$ 的 F 分布表的上侧分位数 F_α 进行比较。如果 $F>F_\alpha$，则我们可以拒绝 H_0，并得出结论：所添加的变量组在统计上是显著的。请注意，对于 $q=1$，$p=2$ 的特殊情形，公式(16.13)即简化为公式(16.10)。

很多学生感觉公式(16.13)过于复杂，为了给出一个关于 F 统计量的比较简单的表达方式，我们可以将自变量个数比较少的模型称为简化模型，将自变量个数比较多的模型称为完全模型。如果我们令 SSE(简化)表示简化模型的误差平方和，令 SSE(完全)表示完全模型的误差平方和，则我们可以把公式(16.13)的分子改写为

$$\frac{\mathrm{SSE}(\text{简化})-\mathrm{SSE}(\text{完全})}{\text{所添加项的个数}} \tag{16.14}$$

请注意，“所添加项的个数”作为分母表示的是完全模型的自变量个数和简化模型的自变量个数之间的差值。公式(16.13)的分母是完全模型误差平方和除以相对应的自由度，换句话说，分母是完全模型的均方误差。我们用 MSE(完全)来表示完全模型的均方误差，这使得我们可以将公式(16.13)改写成

$$F=\frac{\dfrac{\mathrm{SSE}(\text{简化})-\mathrm{SSE}(\text{完全})}{\text{所添加项的个数}}}{\mathrm{MSE}(\text{完全})} \tag{16.15}$$

为了描述 F 统计量的用途，假设我们面临一个包含 30 个观测值的回归问题。第一个模型，包含自变量 x_1，x_2，x_3，其误差平方和为 150；第二个模型，包含自变量 x_1，x_2，x_3，x_4 和 x_5，其误差平方和为 100。x_4 和 x_5 这两个自变量的添加是否导致误差平方和显著减少了？

首先，请注意 SST 的自由度为 $30-1=29$，完全模型回归平方和的自由度为 5(即完全模型中的自变量个数)。因此，完全模型的误差平方和的自由度是 $29-5=24$，MSE(完全) $=100/24=4.17$。由此 F 统计量为

$$F=\frac{\dfrac{150-100}{2}}{4.17}=6.00$$

将这个 F 计算值与分子自由度为 2，分母自由度为 24 的 F 分布表数值进行比较，在 $\alpha=0.05$ 的显著性水平下，附录 B 的表 4 显示：$F_{0.05}=3.40$。由于 $F=6.00>F_{0.05}=3.40$，因此我们可以得到结论：x_4 和 x_5 这两个自变量的添加在统计上是显著的。

p - 值的应用

我们也可以利用 p - 值标准来确定将一个或者多个自变量添加到多元回归模型中去的利弊问题。在前面的例子中，我们阐述了怎样利用 F 检验来确定，将 x_4 和 x_5 这两个自变量添加到含自变量 x_1，x_2，x_3 的模型中，这种添加在统计上是否显著的问题。对这个例子，F 统计量的计算值为 6.00，通过将 $F=6.00$ 与关键值 $F_{0.05}=3.40$ 进行比较，我们可以得出结论：x_4 和 x_5 这两个自变量的添加在统计上是显著的。与 $F=6.00$(分子自由度为 2，分母自由度为 24)相联系的 p - 值为 0.008。由于 p - 值 $=0.008<\alpha=0.05$，我们也可以得出结论：x_4 和 x_5 这两个自变量的添加在统计上是显著的。直接从 F 分布表上确定 p - 值是比较困难的，如果使用计算软件包，比如 Minitab 或者 Excel，则 p - 值的计算就会变得相当简单。

评　注

F 统计量的计算也可以根据回归平方和的差来完成，为了说明 F 统计量的这种计算形式，我们首先看到：

$$\text{SSE(简化)}=\text{SST}-\text{SSR(简化)}$$
$$\text{SSE(完全)}=\text{SST}-\text{SSR(完全)}$$

因此，SSE(简化) - SSE(完全) = [SST - SSR(简化)] - [SST - SSR(完全)] = SSR(完全) - SSR(简化)

$$\text{因此，}F=\frac{\dfrac{\text{SSR(完全)}-\text{SSR(简化)}}{\text{所添加项的个数}}}{\text{MSE(完全)}}$$

练　习

方　法

10. 某个回归分析包含 27 个观测值，有如下的估计回归方程：

$$\hat{y}=25.2+5.5x_1$$

对这个估计回归方程，SST = 1 550，SSE = 520。

a. 在 $\alpha=0.05$ 的显著性水平下，检测 x_1 的显著性。假设将 x_2 和 x_3 添加到模型中，得到如下的回归方程：$\hat{y}=16.3+2.3x_1+12.1x_2-5.8x_3$，对这个估计回归方程，有 SST = 1 550，

SSE = 100。

b. 在 $\alpha = 0.05$ 的显著性水平下，利用 F 检验来确定 x_2 和 x_3 是否对模型有显著作用。

11. 某个回归分析包含 30 个观测值，有如下的估计回归方程：



$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

对这个估计回归方程，SST = 1 805，SSR = 1 760。

a. 在 $\alpha = 0.05$ 的显著性水平下，检测变量之间关系的显著性。假设将 x_1 和 x_4 从模型中删除，得到如下的回归方程：$\hat{y} = 11.1 - 3.6x_2 + 8.1x_3$，对这个估计回归方程，有 SST = 1805，SSR = 1 705。

b. 计算 SSE(x_1，x_2，x_3，x_4)。

c. 计算 SSE(x_2，x_3)。

d. 在 $\alpha = 0.05$ 的显著性水平下，利用 F 检验来确定 x_1 和 x_4 是否对模型有显著作用。

应　用

12. 下表给出了全美橄榄球联赛中 14 支球队的 15 场比赛成绩。



a. 建立估计回归方程，在已知一支球队抢断次数的情况下预测这支球队的总得分。

b. 建立估计回归方程，在已知一支球队抢断次数、带球突破前进的码数、对方球队抢断次数的情况下预测这支球队的总得分。



c. 在 $\alpha = 0.05$ 的显著性水平下，检验带球突破前进的码数、对方球队抢断球次数这两个变量的添加是否对(a)中估计回归方程有显著作用？解释原因。

					交互作用	
球队	赢—输	总得分	带球突破前进的码数	传球码数	本球队抢断次数	对方球队抢断次数
Atlanta	5—10	305	1 907	2 473	19	23
Chicago	12—3	187	2 134	2 718	14	24
Dallas	3—12	358	1 858	3 386	24	10
Detroit	4—11	292	1 184	1 971	15	12
Green Bay	3—12	298	1 274	3 046	22	20
St. Louis Rams	9—6	277	1 882	3 604	17	22
Minnesota	10—5	206	1 744	3 633	16	35
New Orleans	9—6	274	1 843	2 963	15	17
N. Y. Giants	10—5	277	1 492	3 096	14	15
Philadelphia	9—6	312	1 812	3 247	17	29
Phoenix	7—8	372	1 909	3 633	19	14
San Francisco	10—5	256	2 453	3 131	14	21
Tampa Bay	4—11	340	1 650	3 169	33	18
Washington	7—8	367	1 377	3 930	24	14

13. 参考练习 12。

a. 建立估计回归方程，以球队总得分为因变量，以传球码数、本球队抢断次数、对方球队抢断次数为自变量。

b. 建立估计回归方程，以传球码数、本球队抢断次数、对方球队抢断次数、带球突破前进的码数作为自变量，以球队总得分为因变量。

c. 在 $\alpha = 0.05$ 的显著性水平下，带球突破前进的码数是否对(a)中所建立的估计回归方程有显著作用？解释原因。

14. 美国心脏协会进行了一次历时 10 年的研究，获得了和中风发病概率有关的年龄、血压以及吸烟的统计资料。以下是这一研究的部分数据，中风风险指的是病人在今后 10 年内中风发病概率(乘 100)。我们用一个虚拟变量来表示病人是否为吸烟者，1 表示病人是吸烟者，0 表示病人不是吸烟者。



CD 光盘数据
Stroke

发病概率	年龄	血压	是否吸烟
12	57	152	0
24	67	163	0
13	58	155	0
56	86	177	1
28	59	196	0
51	76	189	1
18	56	155	1
31	78	120	0
37	80	135	1
15	78	98	0
22	71	152	0
36	70	173	1
15	67	135	1
48	77	209	1
15	60	199	0
36	82	119	1
8	66	166	0
34	80	125	1
3	62	117	0
37	59	207	1

a. 利用这些数据，建立一个估计回归方程，因变量为发病概率，自变量是病人的年龄，血压。

b. 考虑对(a)中的估计回归方程添加两个自变量：一个是年龄和血压之间的相互作用，另一个

是此人是否吸烟。利用这4个自变量建立估计回归方程。

c. 在 $\alpha=0.05$ 的显著性水平下，检验年龄和血压之间的相互作用以及吸烟这两个自变量的添加是否对(a)中所建立的估计回归方程有显著作用。

15. 全国橄榄球联盟根据球员的场上位置将他们分成5—9个等级。等级的划分如下标准：8—9级从第一年开始；7.0—7.9级现在开始；6.0—6.9级将作为球队后备力量；5.0—5.9级可以组建俱乐部及其分部。下表所示的是40个球员的场上位置、体重、速度(40码)以及等级(*USA Today*, April 14, 2000)：



CD光盘数据
NFL

观测序号	姓名	位置	体重（磅）	速度（秒）	等级
1	Peter Warrick	接球手	194	4.53	9.0
2	Plaxico Burress	接球手	231	4.52	8.8
3	Sylvester Morris	接球手	216	4.59	8.3
4	Travis Taylor	接球手	199	4.36	8.1
5	Laveranues Coles	接球手	192	4.29	8.0
6	Dez White	接球手	218	4.49	7.9
7	Jerry Porter	接球手	221	4.55	7.4
8	Ron Dugans	接球手	206	4.47	7.1
9	Todd Pinkston	接球手	169	4.37	7.0
10	Dennis Northcutt	接球手	175	4.43	7.0
11	Athony Lucas	接球手	194	4.51	6.9
12	Darrell Jackson	接球手	197	4.56	6.6
13	Danny Farmer	接球手	217	4.60	6.5
14	Sherrod Gideon	接球手	173	4.57	6.4
15	Trevor Gaylor	接球手	199	4.57	6.2
16	Cosey Coleman	守门员	322	5.38	7.4
17	Travis Glaridge	守门员	303	5.18	7.0
18	Kaulana Noa	守门员	317	5.34	6.8
19	Leander Jordan	守门员	330	5.46	6.7
20	Chad Clifton	守门员	334	5.18	6.3
21	Manula Savea	守门员	308	5.32	6.1
22	Ryan Johanningmei	守门员	310	5.28	6.0
23	Mark Tauscher	守门员	318	5.37	6.0
24	Blaine Saipaia	守门员	321	5.25	6.0
25	Richard Mercier	守门员	295	5.34	5.8
26	Damion McIntosh	守门员	328	5.31	5.3
27	Jeno James	守门员	320	5.64	5.0
28	Al Jackson	守门员	304	5.20	5.0

（续表）

观测序号	姓名	位置	体重（磅）	速度（秒）	等级
29	Chris Samuels	攻球手	325	4.95	8.5
30	Stockar McDougle	攻球手	361	5.50	8.0
31	Chris McIngosh	攻球手	315	5.39	7.8
32	Adrian Klemm	攻球手	307	4.98	7.6
33	Todd Wade	攻球手	326	5.20	7.3
34	Marvel Smith	攻球手	320	5.36	7.1
35	Michael Thompson	攻球手	287	5.05	6.8
36	Bobby Williams	攻球手	332	5.26	6.8
37	Darnell Alford	攻球手	334	5.55	6.4
38	Terrance Beadles	攻球手	312	5.15	6.3
39	Tutan Reyes	攻球手	299	5.35	6.1
40	Greg Robinson-Ran	攻球手	333	5.59	6.0

a. 建立一个虚拟变量来表示球员的场上位置。

b. 建立一个估计回归方程来表示球员等级与场上位置、体重以及速度之间的相关关系。

c. 在 $\alpha = 0.05$ 的显著性水平下，检测(b)中所建立的估计回归方程里面自变量和因变量之间的关系是否显著。

d. 球员的场上位置是否是决定球员等级的显著因素？利用 $\alpha = 0.05$ 的显著性水平进行检测，并解释原因。

16.3　大型问题分析

在介绍多元回归分析的过程中，我们深入讨论过 Butler 运输公司的例子。这个例子所涉及的变量数目较少，因而便于引入概念，但是如果要说明在建模过程中我们应该选择哪些变量，则应用这个例子将会变得比较困难。为了描述下一节将要详细论述的变量选择过程，我们引入由 8 个自变量和 25 个观测值所组成的一个数据集合，这个数据集合是由得克萨斯州克里斯琴大学经济系的 David W. Cravens 博士提供给我们使用的，因此我们将这个数据集合称为 Cravens 数据。*

Cravens 数据是从一家在多个销售区域销售产品的公司采集的，这家公司在每一个销售区域都有独家代理商。我们对这些数据进行回归分析的目的在于确定每个预测(自)变量是否能够解释各个销售区域的销售情况。对 25 个销售区域随机抽取的样本组成了表 16.5 中的数据；表 16.6 给出了变量的定义。

---

* 详见 David W. Cravens, Robert B. Woodruff, 和 Joe C. Stamper, "An Analytical Approach for Evaluating Sales Territory Performance," *Journal of Marketing*, 第 36 期(January 1972): 31 - 37 页。Copyright © 1972 American Marketing Association.

表 16.5 Cravens 数据


CD 光盘数据
Cravens

Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work	Rating
3 669.88	43.10	74 065.1	4 582.9	2.51	0.34	74.86	15.05	4.9
3 473.95	108.13	58 117.3	5 539.8	5.51	0.15	107.32	19.97	5.1
2 295.10	13.82	21 118.5	2 950.4	10.91	-0.72	96.75	17.34	2.9
4 675.56	186.18	68 521.3	2 243.1	8.27	0.17	195.12	13.40	3.4
6 125.96	161.79	57 805.1	7 747.1	9.15	0.50	180.44	17.64	4.6
2 134.94	8.94	37 806.9	402.4	5.51	0.15	104.88	16.22	4.5
5 031.66	365.04	50 935.3	3 140.6	8.54	0.55	256.10	18.80	4.6
3 367.45	220.32	35 602.1	2 086.2	7.07	-0.49	126.83	19.86	2.3
6 519.45	127.64	46 176.8	8 846.2	12.54	1.24	203.25	17.42	4.9
4 876.37	105.69	42 053.2	5 673.1	8.85	0.31	119.51	21.41	2.8
2 468.27	57.72	36 829.7	2 761.8	5.38	0.37	116.26	16.32	3.1
2 533.31	23.58	33 612.7	1 991.8	5.43	-0.65	142.28	14.51	4.2
2 408.11	13.82	21 412.8	1 971.5	8.48	0.64	89.43	19.35	4.3
2 337.38	13.82	20 416.9	1 737.4	7.80	1.01	84.55	20.02	4.2
4 586.95	86.99	36 272.0	10 694.2	10.34	0.11	119.51	15.26	5.5
2 729.24	165.85	23 093.3	8 618.6	5.15	0.04	80.49	15.87	3.6
3 289.40	116.26	26 878.6	7 747.9	6.64	0.68	136.58	7.81	3.4
2 800.78	42.28	39 572.0	4 565.8	5.45	0.66	78.86	16.00	4.2
3 264.20	52.84	51 866.1	6 022.7	6.31	-0.10	136.58	17.44	3.6
3 453.62	165.04	58 749.8	3 721.1	6.35	-0.03	138.21	17.98	3.1
1 741.45	10.57	23 990.8	861.0	7.37	-1.63	75.61	20.99	1.6
2 035.75	13.82	25 694.9	3 571.5	8.39	-0.43	102.44	21.66	3.4
1 578.00	8.13	23 736.3	2 845.5	5.15	0.04	76.42	21.46	2.7
4 167.44	58.44	34 314.3	5 060.1	12.88	0.22	136.58	24.78	2.8
2 799.97	21.14	22 809.5	3 552.0	9.14	-0.74	88.62	24.96	3.9

表 16.6 Cravens 数据中的变量定义

变量	定义
Sales	销售代理商的总销售收入
Time	销售代理商为公司代理销售的时间
Poten	市场潜力：销售区域总销量
AdvExp	在销售区域所花费的广告费用
Share	所占市场份额：取前 4 年的加权平均值
Change	过去 4 年中市场份额的变动
Accounts	指定给销售代理商的商店数量*
Work	工作量：根据商店一年的零售和批发数额所计算的加权指数
Rating	根据 8 个方面的业绩表现对代理商进行排序：综合排序为 1—7 分

* 给这些数据加标记是为了观测置信度。

第一步，我们首先考虑每一对变量之间的样本相关系数。图 16.13 是利用 Minitab 的相关命令所得到的相关矩阵。请注意，Sales 和 Time 之间的样本相关系数为 0.623，Sales 和 Poten 之间的样本相关系数是 0.598，其他依次类推。

考察各个自变量之间的样本相关系数，我们可以看到 Time 和 Accounts 之间的相关系数为 0.758；因此，如果将 Accounts 作为自变量，则 Time 将不会对模型具有很大的解释能力。回顾 15.4 节中所介绍的关于多重共线性的经验检验法则：如果任意两个自变量之间的样本相关系数的绝对值大于 0.7，则多重共线性将会影响估计结果的精确性。如果可能的话，我们应该尽量避免在同一个回归模型中同时包含 Time 和 Accounts 这两个自变量。Change 和 Rating 之间的样本相关系数为 0.549，这也是比较高的，因此需要我们对此作进一步的考虑。

	Sales	Time	Poten	AdvExp	Share	Change	Accounts	Work
Time	0.623							
Poten	0.598	0.454						
AdvExp	0.596	0.249	0.174					
Share	0.484	0.106	-0.211	0.264				
Change	0.489	0.251	0.268	0.377	0.085			
Accounts	0.754	0.758	0.479	0.200	0.403	0.327		
Work	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277

图 16.13　Cravens 数据中的样本相关系数

考察一下 Sales 和每一个自变量之间的样本相关系数，我们很快就可以明白哪一个自变量能够作为最好的预测自变量。我们可以看到：Accounts 是 Sales 最好的预测变量，因为它们之间的样本相关系数高达 0.754，是所有关于 Sales 的样本相关系数里面最高的。回顾一个自变量的情形，样本相关系数的平方就是判定系数。因此，Accounts 可以解释 $(0.754)^2(100)$，或者 56.85% 的 sales 变异性。下一个最为重要的自变量是 Time、Poten 和 AdvExp，它们之中的每一个和 Sales 之间的样本相关系数都接近于 0.6。

虽然这里存在潜在的多重共线性问题，还是让我们来考察一下怎样利用所有 8 个自变量来建立估计回归方程。Minitab 计算软件包给出了图 16.14 中的计算结果。这个带有 8 个自变量的多元回归模型修正判定系数为 88.3%。但是，请注意，单个参数 t 检验的 p - 值显示：在 $\alpha = 0.05$ 的显著性水平下，如果其他所有变量的影响都是已知的，则只有 Poten、AdvExp 和 Share 这三个变量是显著的。因此，我们可能会倾向于仅仅利用这三个变量调查所得到的结果。图 16.15 显示了利用这三个自变量所得到的估计回归方程的 Minitab 计算结果。我们可以看到估计回归方程的修正判定系数是 82.7%，虽然它没有 8 个自变量的估计回归方程的系数那样好，但至少也是比较高的。

对于已知的可以使用的数据，我们怎样才能建立一个最为合适的估计回归方程呢？也就是说，我们可以建立 8 个只带一个变量的估计回归方程(每个回归方程对应一个自变量)，28 个带

```
The regression equation is
Sales = - 1508 + 2.01 Time + 0.0372 Poten + 0.151 AdvExp + 199 Share
        + 291 Change + 5.55 Accounts + 19.8 Work + 8 Rating

Predictor       Coef       Stdev    t-ratio       p
Constant      1507.8       778.6      -1.94   0.071
Time           2.010       1.931       1.04   0.313
Poten       0.037205    0.008202       4.54   0.000
AdvExp       0.15099     0.04711       3.21   0.006
Share         199.02       67.03       2.97   0.009
Change         290.9       186.8       1.56   0.139
Accounts       5.551       4.776       1.16   0.262
Work           19.79       33.68       0.59   0.565
Rating           8.2       128.5       0.06   0.950

s = 449.0       R-sq = 92.2%     R-sq(adj) = 88.3%

Analysis of Variance

SOURCE        DF          SS          MS         F       p
Regression     8    38153568     4769196     23.65   0.000
Error         16     3225984      201624
Total         24    41379552
```

图 16.14　包含所有 8 个自变量的模型的 Minitab 输出结果

```
The regression equation is
Sales = - 1604 + 0.0543 Poten + 0.167 AdvExp + 283 Share

Predictor       Coef       Stdev    t-ratio       p
Constant     -1603.6       505.6      -3.17   0.005
Poten       0.054286    0.007474       7.26   0.000
AdvExp       0.16748     0.04427       3.78   0.001
Share         282.75       48.76       5.80   0.000

s = 545.5       R-sq = 84.9%     R-sq(adj) = 82.7%

Analysis of Variance

SOURCE        DF          SS          MS         F       p
Regression     3    35130240    11710080     39.35   0.000
Error         21     6249310      297586
Total         24    41379552
```

图 16.15　包含 Poten、AdvExp 和 Share 三个变量的模型的 Minitab 输出结果

两个变量的估计回归方程(8 个变量每次取 2 个所组成的组合)，等等，依此类推。对于 Cravens 数据，共有 255 个包含一个或者多个自变量的互不相同的估计回归方程可以用于对数据进行拟合。

如今我们已经具有相当优秀的计算软件包，这使得我们可以计算所有可能的回归。但是这样做也会导致大量的计算任务，并且模型设计者不得不审查大量的计算机输出结果，其中大部分都与失败的模型相关。统计学家们更加倾向于用系统的方法从所有自变量中选取一部分自变量，利用这一部分自变量即可得到最适合的估计回归方程。在下一节中，我们将介绍一些更为通用的方法。

16.4　变量选择方法

在本节中，我们将讨论选择回归模型自变量所用的 4 种基本的变量选择方法：逐步回归、前向选择、后向消元和最佳子集回归。在已知所有可能的自变量数据集合时，我们可以利用这 4 种方法来确定哪些自变量可以给出最佳的模型。前 3 种方法都采用反复迭代；过程的每一步都添加或者删除一个自变量，并且对新建立的回归模型进行评估。这种迭代过程将会持续不断地进行下去，直到有一个终止标准可以表明：我们已经不可能再找到更好的模型。最后一种方法(最佳子集回归)并非单个变量、一次性方法，利用这种方法我们可以对包含不同自变量子集的回归模型进行评估。

变量选择方法在建模初期特别有用，但是这些方法还是无法代替分析师的经验和判断。

在逐步回归、前向选择和后向消元方法中，确定从模型中添加或者删除一个自变量的标准是基于 16.2 节中所介绍的 F 统计量。举个例子，我们正在考虑添加 x_2 到包含 x_1 的模型中，或者从包含 x_1 和 x_2 的模型中删除 x_2，为了检验 x_2 的添加或者删除是统计显著的，我们可以进行如下备择假设：

$$H_0:\ \beta_2 = 0$$
$$H_a:\ \beta_2 \neq 0$$

在 16.2 节中[见公式(16.10)]，我们已经阐述过：

$$F = \frac{\dfrac{\mathrm{SSE}(x_1) - \mathrm{SSE}(x_1, x_2)}{1}}{\dfrac{\mathrm{SSE}(x_1, x_2)}{n - p - 1}}$$

这个 F 值可以作为确定 x_2 的存在是否导致误差平方和显著减少的标准。对应于 F 统计量的 p -

值也可以作为标准，用于确定是否需要从模型中添加或者删除某个自变量。一般的拒绝法则是这样的：如果 p - 值< α，则我们可以拒绝 H_0。

逐步回归

逐步回归方法的第一步要确定**已经存在于模型中的**变量是否需要删除，它首先计算模型中每一个自变量的 F 统计量和相对应的 p - 值。用于确定某个自变量是否应该从模型中删除的 α 显著性水平可以参考 Minitab 中的 α **删除值**。如果任意自变量的 p - 值大于 α **删除值**，则我们可以从模型中删除 p - 值最大的自变量，然后逐步回归方法开始下一步新的计算。

如果没有自变量可以从模型中删除，则我们可以尝试将另一个新的自变量添加到模型中去。首先计算模型以外的每一个自变量的 F 统计量和相对应的 p - 值，用于确定某个自变量是否可以添加到模型中的 α 显著性水平可以参考 Minitab 中的“α 添加值”。如果变量的 p - 值小于 α **添加值**，则将 p - 值最小的自变量添加入模型之中。逐步回归法会如此不断重复操作，直至没有自变量可以从模型中删除，也没有自变量可以添加到模型中去。

图 16.16 显示了对 Cravens 数据应用 Minitab 逐项回归方法处理后所得的结果，α **删除值**为 0.05，α **添加值**也是 0.05。逐步回归法经过 4 个步骤后终止，应用 Minitab 逐步回归法处理后所得的估计回归方程如下所示：

$$\hat{y} = -1\,441.93 + 9.2\,\text{Accounts} + 0.175\,\text{AdvExp} + 0.038\,2\,\text{Poten} + 190\,\text{Share}$$

由于对一定数量的自变量而言，对变量进行每次单个处理的方法并没有考虑到每一个可能的子集，所以没有必要选择 R - sq 值最高的模型。

请注意图 16.16 中 $s = \sqrt{\text{MSE}}$ 经过 4 个步骤已经从带有最佳自变量模型(利用 Accounts)的 881 减少到了 454。R-sq 值已经从 56.85% 增加到了 90.04%，所推荐的估计回归方程的 R-sq（adj）值为 88.05%。

总的说来，逐步回归法的每一步都是首先考虑是否可以从当前模型中删除自变量。如果没有自变量从当前模型中删除，则逐步回归法会考虑是否将某个不在当前模型之中的变量添加到模型中来。由于逐步回归法的性质特点，某个自变量可以在某一步进入模型，也可以在某一步从模型中删除，然后又在下一步进入模型。如果没有自变量进入模型，或者没有自变量从模型中删除，逐步回归法就会终止。

前向选择

前向选择法在没有自变量的情况下开始，它每次将一个自变量添加到模型当中，确定自变量是否能够添加到模型中去的方法和逐步回归法一样。但是，不同的是，前向选择法将自变量添加到模型中以后，便不允许再将自变量从模型中删除。如果模型之外的每一个自变量的 p - 值都小

```
Alpha-to-Enter: 0.05     Alpha-to-Remove: 0.05

Response is Sales on 8 predictors, with N = 25

    Step         1        2         3          4
Constant    709.32    50.29   -327.24   -1441.93

Accounts      21.7     19.0      15.6        9.2
T-Value       5.50     6.41      5.19       3.22
P-Value      0.000    0.000     0.000      0.004

AdvExp                0.227     0.216      0.175
T-Value                4.50      4.77       4.74
P-Value               0.000     0.000      0.000

Poten                           0.0219     0.0382
T-Value                           2.53       4.79
P-Value                          0.019      0.000

Share                                        190
T-Value                                     3.82
P-Value                                    0.001

S             881       650       583        454
R-Sq        56.85     77.51     82.77      90.04
R-Sq(adj)   54.97     75.47     80.31      88.05
C-p          67.6      27.2      18.4        5.4
```

图 16.16　对 Cravens 数据应用逐步回归法处理所得的 Minitab 输出结果

于 α **添加值**，则前向选择法就会终止。

应用 Minitab 前向选择法处理后所得的估计回归方程如下所示：

$$\hat{y} = -1\,441.93 + 9.2\ \text{Accounts} + 0.175\ \text{AdvExp} + 0.038\,2\ \text{Poten} + 190\ \text{Share}$$

因此，对于 Cravens 数据，前向选择法（α **添加值**为 0.05）和逐步回归法所得到的估计回归方程是相同的。

后向消元

后向消元法的起点是模型中已经包含所有的自变量，然后它使用与逐步回归相同的方法每次删除一个自变量。但是，一旦一个自变量从模型中被删除，则后向消元法不允许这个自变量重新添加到模型中去。如果模型之内的每一个自变量的 p - 值都小于 α **删除值**，则后向消元法就会终止。

对 Cravens 数据应用 Minitab 后向消元法处理(α **删除值**为 0.05)后所得的估计回归方程如下所示:

$$\hat{y} = -1\,312 + 3.8\ \text{Time} + 0.044\,4\ \text{Poten} + 0.152\ \text{AdvExp} + 259\ \text{Share}$$

将利用后向消元法所得的估计回归方程与利用前向选择法所得的估计回归方程进行比较,我们可以看到两者都包含:AdvExp、Poten 和 Share。但是,后向消元法包含 Time,而前向选择法包含 Accounts。

前向选择和后向消元有可能得到不同的模型。

前向选择和后向消元是建模的两个极端,前向选择法起始于模型中没有自变量,它每次将一个自变量添加到模型当中,但是后向消元法起始于模型中的所有自变量,它每次从模型中删除一个变量。这两种方法有可能得到一个相同的估计回归方程,但是,就像我们在对 Cravens 数据的处理中所看到的,这两种方法也有可能得到两个不同的估计回归方程。究竟选取哪一个估计回归方程尚有待讨论,最后还是要利用统计分析师的判断能力。接下来我们要讨论的最佳子集回归建模给出了在最终决策之前所需的补充建模信息。

最佳子集回归

逐项回归、前向选择,后向消元都是通过每次从模型中添加或者删除一个自变量来选择回归模型。对于一组已知变量,它们都不能保证所建立的模型是最佳模型。因此,我们可以研究这些每次处理一个变量的方法,这样可以对我们选择一个最佳回归模型提供帮助。

一些软件包有执行最佳子集回归过程的功能,在给定一组自变量的情况下,最佳子集回归过程可以使用户获得最佳回归模型。Minitab 就具备这样的功能。图 16.17 是对 Cravens 数据集合进行最佳子集回归处理后所得的部分计算机输出结果。

这个输出结果给出了两个单变量的最佳估计回归方程,同时也给出了两个双变量的最佳估计回归方程,还有两个三变量的最佳估计回归方程,依此类推。对于任意数量的预测变量来说,用来确定哪一个估计回归方程是最佳回归方程的标准都是判定系数(R-sq)的数值。举个例子,变量 Accounts,它的 R-sq 值为 56.8%,给出了仅有一个自变量的最佳估计回归方程;AdvExp 和 Accounts,R-sq 值为 77.5%,给出了含两个自变量的最佳估计回归方程;Poten、AdvExp 和 Share,R-sq 值为 84.9%,给出了含三个自变量的最佳估计回归方程。对于 Cravens 数据,含有六个自变量:Time、Poten、AdvExp、Share、Change、Accounts 的模型的修正判定系数(Adj. R-sq = 89.4%)最大,但是,含有四个自变量(Poten、AdvExp、Share、Accounts)的模型的修正判定系数也比较高(88.1%)。在其他所有条件都相同的情况下,人们倾向于使用包含较少自变量的简单模型。

Vars	R-sq	Adj. R-sq	s	Time	Poten	AdvExp	Share	Accounts	WorkKing	Rating
1	56.8	55.0	881.09					X		
1	38.8	36.1	1049.3	X						
2	77.5	75.5	650.39			X		X		
2	74.6	72.3	691.11		X		X			
3	84.9	82.7	545.52		X	X	X			
3	82.8	80.3	582.64		X	X		X		
4	90.0	88.1	453.84		X	X	X	X		
4	89.6	87.5	463.93	X	X	X	X			
5	91.5	89.3	430.21	X	X	X	X	X		
5	91.2	88.9	436.75		X	X	X	X	X	
6	92.0	89.4	427.99	X	X	X	X	X	X	
6	91.6	88.9	438.20		X	X	X	X	X	X
7	92.2	89.0	435.66	X	X	X	X	X	X	X
7	92.0	88.8	440.29	X	X	X	X	X	X	X
8	92.2	88.3	449.02	X	X	X	X	X	X	X

图 16.17　最佳子集回归的部分 Minitab 输出结果

最终选择

对 Cravens 数据已经完成的统计分析可以给我们选择最后模型做一个很好的准备，但是在作出最终的选择之前，我们还需要对数据进行更多的分析。正如我们在第 14 和第 15 章所提过的那样，我们应该对残差进行仔细分析。对于所选择的模型，我们希望它的残差图应该呈一条水平带状。我们可以假设残差分析不会造成什么问题，并且可以利用最佳子集回归方法来帮助选择模型。

最佳子集过程已经告诉我们：最佳四变量模型包含 Poten、AdvExp、Share、Accounts 四个变量，利用逐项回归过程所得的四变量模型也是如此。表 16.7 可以帮助我们进行最后的选择，它给出了几个模型，这些模型包含全部或者一部分自变量。

表 16.7　包含 Poten、AdvExp、Share、Accounts 四个变量的模型

模型	自变量	Adj. R-sq
1	Accounts	55.0
2	AdvExp, Accounts	75.5
3	Poten, Share	72.3
4	Poten, AdvExp, Accounts	80.3
5	Poten, AdvExp, Share	82.7
6	Poten, AdvExp, Share, Accounts	88.1

从表 16.7 中我们可以看到：仅包含 AdvExp 和 Accounts 的模型是一个比较好的模型，其修正判定系数为 75.5%，而对于包含了全部四个自变量的模型，它的修正判定系数仅仅提高了 12.6 个百分点。举个例子，假如市场潜力(Poten)难以测量，则我们会倾向于选择只包含两个自变量的简单模型。但是，如果数据可以很轻松地获得，并且对销量的预测要求很精确，则模型的设计者很明显会倾向于使用包含所有四个自变量的模型。

评 注

1. 逐项回归法要求 α 删除值要大于或者等于 α 添加值。这项规定可以防止同一个变量在同一个步骤中被删除而又立即被添加到模型当中。
2. 对本章所介绍的任何一种方法，自变量的函数都可以用来生成新的自变量。举个例子，如果我们希望模型中 x_1x_2 表示这两个变量之间的相互作用，我们可以利用自变量 x_1 和 x_2 的数据生成新的变量 $z = x_1x_2$ 的数据。
3. 这些每次只能添加或者删除一个变量的过程，都不能保证确认最佳回归模型。但是它们是找到较好模型的优良方法，特别是多重共线性比较弱时。

练 习

应 用

16. 两个专家给出了他们认为在美国全国范围内最好的学区列表，对每一个学区，有如下的数据：每班平均人数，每个学生的教育经费(美元)，教师的平均工资(美元)，综合 SAT 分数，参加 SAT 考试的学生比例，考入四年制大学的毕业生比例。

自测题

CD 光盘数据 Schools

城市	每班平均人数	每个学生的教育经费(美元)	教师的平均工资(美元)	综合 SAT 分数(参加考试的百分比)	考入四年制大学的毕业生比例(%)
Blue Springs, MO	25	3 060	29 359	1 083/(8)	74
Garden City, NY	18	9 700	51 000	997/(99)	77
Indianapolis, IN	30	3 222	30 482	716/(42)	40
Newport Beach, CA (Newport – Mesa)	26	4 028	37 043	977/(46)	51
Novi, MI	20	3 067	39 797	980/(15)	53

（续表）

城市	每班平均人数	每个学生的教育经费(美元)	教师的平均工资(美元)	综合 SAT 分数(参加考试的百分比)	考入四年制大学的毕业生比例(%)
Piedmont, CA(Piedmont City)	28	4 208	37 274	1 042/(91)	75
Pittsburgh, PA(Fox Chapel area)	21	4 884	37 156	983/(80)	66
Scarsdale, NY(Edgemont)	20	9 853	31 555	1 110/(98)	87
Wayne, PA(Radnor Township)	22	5 022	40 406	1 040/(95)	85
Weston, MA	21	4 680	39 800	1 031/(99)	89
Farmingdale, NY	22	6 729	45 846	947/(75)	81
Mamaroneck, NY	20	10 405	49 625	1 000/(90)	69
Mayfield, OH	24	5 881	36 228	1 003/(25)	48
Morristown, NJ	22	6 300	37 000	972/(80)	64
New Rochelle, NY	23	8 875	41 650	1 039/(80)	55
Newtown Square, PA (Marple – Newtown)	17	5 313	38 000	963/(75)	79
Omaha, NB(Westside)	23	4 815	32 500	1 059/(31)	81
Shaker Heights, OH	23	4 370	38 639	940/(56)	82

设参加 4 年大学学习的毕业生数为因变量。

a. 建立最佳单变量估计回归方程。

b. 利用逐项回归法建立最佳估计回归方程。

c. 利用后向消元法建立最佳估计回归方程。

d. 利用最佳子集回归法建立最佳估计回归方程。

17. 参考练习 12 中的数据，取获胜数为自变量。

a. 建立最佳单变量估计回归方程。

b. 利用逐项回归法建立最佳估计回归方程。

c. 利用后向消元法建立最佳估计回归方程。

d. 利用最佳子集回归法建立最佳估计回归方程。

18.



CD 光盘数据 LPGATour

女子职业高尔夫球协会(LPGA)保留有关于它的成员在高尔夫球联盟巡回赛中的比赛成绩和收入的统计资料。以下是 1997 年底 LPGA 巡回赛的统计资料(*Golfweek*, December 6, 1997)。Scoring Avg. 指的是每 18 个高尔夫球的平均得分；Driving Distance 指的是每杆高尔夫球平均飞行的码数；Fairways 指的是高尔夫球落点位于球道之内的比例，Greens 指的是球手按比赛规则将高尔夫球击打在球洞场地之内的比例，Putts 指的是球手每场比赛将高尔夫球轻击入洞的比例；Sand Saves 指的是高尔夫球被打入球场旁边的沙地时，高尔夫球手可以克服地面起伏不平的困难将球救回的比例。如果高尔夫球的任何部分接触到球场的轻击区，并且球入洞杆数少于标准杆数两杆，则我们应该按

规则将球场的场地情况考虑在内。

球手 (Player)	平均得分 (Scoring Avg.)	码数 (Driving Distance)	入球道比 (Fairways)	入场内比 (Greens)	进洞比 (Putts)	救球比 (Sand Saves)
Annika Sorenstam	70.04	249.00	75.4	73.2	29.67	47.3
Karrie Webb	70.00	254.60	72.1	75.1	30.20	38.0
Kelly Robbins	70.35	256.00	69.3	78.6	29.96	44.9
Chris Johnson	70.84	249.40	66.9	70.4	30.17	39.6
Tammie Green	71.24	239.40	70.4	68.6	29.64	47.2
Juli Inkster	70.64	251.60	69.0	69.8	29.35	46.0
Liselotte Neumann	71.28	243.50	66.6	65.9	29.36	48.4
Laura Davies	70.86	258.40	56.4	68.8	29.90	40.0
Nancy Lopez	70.70	245.50	73.4	73.0	30.23	43.1
Betsy King	71.52	245.70	65.9	66.9	29.63	37.6
Lorie Kane	71.47	243.30	74.9	69.4	30.22	47.8
Michelle McGann	71.52	256.20	57.8	69.6	30.05	51.0
Donna Andrews	71.01	230.60	79.1	71.7	29.94	42.3
Colleen Walkeer	72.31	230.00	73.6	63.3	29.58	40.9
Rosie Jones	71.77	227.20	77.0	66.0	29.86	41.0
Lisa Hackney	71.34	246.00	69.8	65.1	29.91	36.4
Jane Geddes	71.34	261.20	64.5	70.0	30.15	44.2
Alison Nicholas	72.31	241.70	71.8	66.4	30.43	39.4
Pat Hurst	72.14	251.50	62.4	69.1	30.58	29.9
Cindy Figg-Currier	71.90	237.20	70.1	68.0	30.40	39.3

a. 如果将每18杆球的平均得分作为自变量，请建立单变量的最佳估计回归方程。根据这个估计回归方程，LPGA巡回赛的球手可以得到什么建议？

b. 利用本节所讲述的方法建立最佳估计回归方程，对参赛球手的平均得分作出估计。

c. (b)中所建立的估计回归方程是否能够合理解释比赛结果。

d. Michele Redman 的相关数据如下：Driving Distance 为 231.6 码，Fairways 为 75.7%，Greens 为 65.2%，Putts 为 30.69，Sand Saves 为 50.5%。请估计这个职业高尔夫球手的平均得分。

19. 从 *Stock Investor Pro* 数据库中我们选取了16家公司的资料组成样本，从这个样本中我们可以得到每一家公司的市盈率，净边际利润，以及销售量增长率(*Stock Investor Pro*，American Association of Individual Investors，August 21，1997)。“行业”栏指的是每一个公式所处的行业，1 = 能源—国际石油；2 = 医疗保健—医药；3 = 其他。建立一个估计回归方程，预测市盈率。并对建立所推荐的估计回归方程的过程进行简要的讨论。



CD 光盘数据
StkData

	市盈率	净边际利润(%)	销售量增长率(%)	行业
Abbott Laboratories	22.3	23.7	10.0	2
American Home Products	22.6	21.1	5.3	2
Amoco	16.7	11.0	16.5	1
Bristol Meyers Squibb Co.	25.9	26.6	9.4	2
Chevron	18.3	11.6	18.4	1
Exxon	18.7	9.8	8.3	1
General Electric Company	13.1	13.4	13.1	3
Hewlett-Packard	23.3	9.7	21.9	3
IBM	17.3	11.5	5.6	3
Merck & Co. Inc.	26.2	25.6	18.9	2
Mobil	18.7	8.2	8.1	1
Pfizer	34.6	25.1	12.8	2
Pharmacia & Upjohn, Inc.	22.3	15.0	2.7	2
Procter & Gamble Co.	5.4	14.9	5.4	3
Texaco	12.3	7.3	23.7	1
Travelers Group Inc.	28.7	17.8	28.7	3

20. 参考练习 14，利用年龄、血压、是否吸烟以及任何与这些变量相关的相互影响作为自变量，建立估计回归方程预测患病的风险。并对建立所推荐的估计回归方程的过程进行简要的讨论。

16.5　残差分析

在第 14 章和第 15 章中，我们阐述了如何利用残差分析对违背回归模型假设的一些情形进行检测。我们找出了与误差项 ε 和模型的假设函数形式相违背的情形，并在本章中讨论了如果检测到这种违背情形可以采取的措施。当我们需要一个不同的函数形式时，曲线和交互作用形式可以贯穿于一般的线性模型始终。当需要考虑几个自变量时，前面章节中的变量选择是比较适宜的。

在第 14 和第 15 章中，我们也讨论了在确定估计回归方程的过程中怎样用残差分析确认异常观测值或者有影响的观测值。我们同时论述了当发现这些观测值时所应该采取的一些步骤。在许多回归研究中，如果数据是随着时间推移而收集的，则误差项之间所存在的某种特殊的相关性可能会引起一些问题，这种特殊的相关性被称为**序列相关**(serial correlation)或者**自相关性**(autocorrelation)。在本节中，我们将阐述怎样利用**杜宾 - 瓦特森检验**(Durbin-Watson test)来检测显著的自相关性。

自相关性和杜宾 - 瓦特森检验

通常说来，商务和经济回归分析研究所需要的数据都是经过一段时间收集到的。一般情况

下，我们用 y_t 表示 y 在 t 时期的数值，而 y_t 也是和 t 时期以前的数值相关。在这种情况下，我们说数据中存在自相关性(又称为序列相关)。如果 y 在 t 时期的数值和 y 在 $t-1$ 时期的数值相关，则我们说数据中存在一阶自相关。如果 y 在 t 时期的数值和 y 在 $t-2$ 时期的数值相关，则我们说数据中存在二阶自相关，依此类推。

当数据中存在自相关时，回归模型的一个假设将无法得到满足：误差项将不再是独立的。如果是一阶自相关情形，t 时期的误差 ε_t 将和 $t-1$ 时期的误差 ε_{t-1} 相关。图 16.18 显示了一阶自相关的两种情形，其中图 A 显示的是正的自相关，图 B 显示的是负的自相关。如果存在正的自相关，某个时期残差为正，则下一个时期紧接着的残差也为正；某个时期残差为负，则下一个时期紧接着的残差也为负，依此类推。如果存在负的自相关，某个时期残差为正，则下一个时期紧接着的残差为负；某个时期残差为负，则下一个时期紧接着的残差为正，也是依此类推。

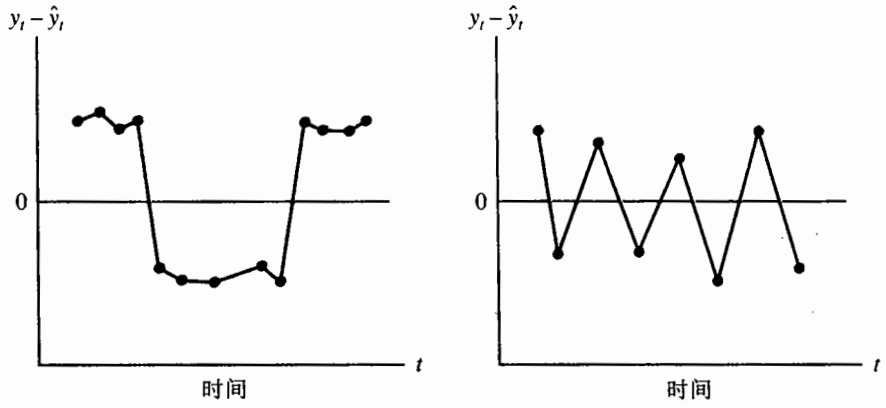


图 A　正的自相关　　图 B　负的自相关

图 16.18　关于一阶自相关的两个数据集合

当数据中存在自相关时，如果以假设的回归模型为基础进行统计显著性检验，则我们的检验会出现严重的误差。因此对自相关进行检验并采取正确的措施是相当重要的。接下来，我们将阐述如何利用杜宾 - 瓦特森统计量来检测一阶自相关。

假设误差项 ε 的值并不是独立的，而是以下面的形式相互作用：

$$\varepsilon_t = \rho\varepsilon_{t-1} + z_t \tag{16.16}$$

式中，ρ 是绝对值小于 1 的参数，z_t 是一个均值为 0，方差为 σ^2 的正态分布的独立变量。从公式(16.16)中我们可以看到，如果 $\rho=0$，则误差项之间互不相关，每一个误差项均值为 0，方差为 σ^2。在这种情况下，数据中不存在自相关性，回归假设也是可以满足的。如果 $\rho>0$，则我们有正的自相关性；如果 $\rho<0$，则自相关性为负。对于这两种情形的任意一种，关于误差项的回归假设均无法得到满足。

关于自相关性的杜宾 - 瓦特森检验利用残差来确定 ρ 是否等于 0。为了简化杜宾 - 瓦特森检验统计量的表示形式，我们用 $e_i = y_i - \hat{y}_i$ 来表示第 i 个残差。这样，杜宾 - 瓦特森检验统计量可

以通过下面的式子计算出来：

杜宾 - 瓦特森检验

$$d=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^2} \tag{16.17}$$

如果残差的相邻值之间差距不大(正自相关性)，杜宾 - 瓦特森检验统计量的数值将会比较小。如果残差的相邻值之间差距较大(负自相关性)，杜宾 - 瓦特森检验统计量的数值将会比较大。

杜宾 - 瓦特森检验统计量的数值位于 0 到 4 之间，并且有两个数值表明了不存在自相关性的范围。杜宾和瓦特森建立了一个临界值表，可以确定他们的统计量什么时候检测出了自相关性的存在。表 16.8 给出了 $\alpha=0.05$，$\alpha=0.025$，$\alpha=0.01$ 的自相关性假设检验的上下临界值(d_L 和 d_U)；n 表示观测值的数目。待检的零假设是：数据中总是不存在自相关性。

$$H_0:\ \rho=0$$

检测正自相关性的备择假设为　　$H_a:\ \rho>0$

检测负自相关性的备择假设为　　$H_a:\ \rho<0$

同时还可以进行双边检验，在这种情况下，备择假设形式为

$$H_a:\ \rho\neq 0$$

图 16.19 显示了如何利用表 16.8 中的 d_L 和 d_U 的数值来检测自相关性。其中图 A 说明了如何检测正自相关性，如果 $d<d_L$，我们可以得出结论：存在正的自相关性。如果 $d_L\leqslant d\leqslant d_U$，我们可以认为检验失效，无法检测自相关性。如果 $d>d_U$，则我们可以认为不存在正的自相关性。

图 B 显示了对负自相关性的检测，如果 $d>4-d_L$，我们可以得出结论：存在负的自相关性。如果 $4-d_U\leqslant d\leqslant 4-d_L$，则无法判断自相关性的存在与否。如果 $d<4-d_U$，我们可以得出结论：不存在负的自相关性。

图 C 显示了对自相关性的双边检测。如果 $d<d_L$，或者 $d>4-d_L$，我们可以拒绝 H_0，并得出结论：存在自相关性。如果 $d_L\leqslant d\leqslant d_U$，或者 $4-d_U\leqslant d\leqslant 4-d_L$，我们可以认为检测失效，无法判断自相关性的存在与否。如果 $d_U<d<4-d_U$，则我们可以得出结论：不存在自相关性。

如果已经确定了显著的自相关性，我们应该考察所假设的回归模型，看它是否忽略了一个或者几个关键自变量，这些关键自变量对因变量有着显著的时间排序影响。如果无法确认这些自变量，则我们可以在模型中引入一个度量观测次数的自变量(举个例子，对于第一次观测，这个变量可以取值为 1，对于第二次观测，这个变量可以取值为 2，依此类推)，这个自变量的引入有时可以消除或者减小自相关性。当这些消除或者减小自相关性的尝试都不奏效时，我们可以对因变量或者自变量进行一些转换，或许会有所帮助；关于这种转换，我们可以在有关回归分析的更高级的教材中进行讨论。

表 16.8　关于自相关性的杜宾 – 瓦特森检验临界值

注意：表中的数值是自相关性的单边杜宾 – 瓦特森检验临界值，如果是自相关性的双边杜宾 – 瓦特森检验，则显著性水平应该翻倍。

对于 $\alpha = 0.05$，d_L 和 d_U 显著点

	自变量个数									
	1		2		3		4		5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

（续表）

对于 $\alpha = 0.05$，d_L 和 d_U 显著点

	自变量个数									
	1		2		3		4		5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
16	0.98	1.24	0.86	1.40	0.75	1.59	0.64	1.80	0.53	2.03
17	1.01	1.25	0.90	1.40	0.79	1.58	0.68	1.77	0.57	1.98
18	1.03	1.26	0.93	1.40	0.82	1.56	0.72	1.74	0.62	1.93
19	1.06	1.28	0.96	1.41	0.86	1.55	0.76	1.72	0.66	1.90
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
21	1.10	1.30	1.01	1.41	0.92	1.54	0.83	1.69	0.73	1.84
22	1.12	1.31	1.04	1.42	0.95	1.54	0.86	1.68	0.77	1.82
23	1.14	1.32	1.06	1.42	0.97	1.54	0.89	1.67	0.80	1.80
24	1.16	1.33	1.08	1.43	1.00	1.54	0.91	1.66	0.83	1.79
25	1.18	1.34	1.10	1.43	1.02	1.54	0.94	1.65	0.86	1.77
26	1.19	1.35	1.12	1.44	1.04	1.54	0.96	1.65	0.88	1.76
27	1.21	1.36	1.13	1.44	1.06	1.54	0.99	1.64	0.91	1.75
28	1.22	1.37	1.15	1.45	1.08	1.54	1.01	1.64	0.93	1.74
29	1.24	1.38	1.17	1.45	1.10	1.54	1.03	1.63	0.96	1.73
30	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	0.98	1.73
31	1.26	1.39	1.20	1.47	1.13	1.55	1.07	1.63	1.00	1.72
32	1.27	1.40	1.21	1.47	1.15	1.55	1.08	1.63	1.02	1.71
33	1.28	1.41	1.22	1.48	1.16	1.55	1.10	1.63	1.04	1.71
34	1.29	1.41	1.24	1.48	1.17	1.55	1.12	1.63	1.06	1.70
35	1.30	1.42	1.25	1.48	1.19	1.55	1.13	1.63	1.07	1.70
36	1.31	1.43	1.26	1.49	1.20	1.56	1.15	1.63	1.09	1.70
37	1.32	1.43	1.27	1.49	1.21	1.56	1.16	1.62	1.10	1.70
38	1.33	1.44	1.28	1.50	1.23	1.56	1.17	1.62	1.12	1.70
39	1.34	1.44	1.29	1.50	1.24	1.56	1.19	1.63	1.13	1.69
40	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
45	1.39	1.48	1.34	1.53	1.30	1.58	1.25	1.63	1.21	1.69
50	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
55	1.45	1.52	1.41	1.56	1.37	1.60	1.33	1.64	1.30	1.69
60	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
65	1.49	1.55	1.46	1.59	1.43	1.62	1.40	1.66	1.36	1.69
70	1.51	1.57	1.48	1.60	1.45	1.63	1.42	1.66	1.39	1.70
75	1.53	1.58	1.50	1.61	1.47	1.64	1.45	1.67	1.42	1.70
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
85	1.56	1.60	1.53	1.63	1.51	1.65	1.49	1.68	1.46	1.71
90	1.57	1.61	1.55	1.64	1.53	1.66	1.50	1.69	1.48	1.71
95	1.58	1.62	1.56	1.65	1.54	1.67	1.52	1.69	1.50	1.71
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72

（续表）

对于 $\alpha=0.05$，d_L 和 d_U 显著点

	自变量个数									
	1		2		3		4		5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

资料来源：J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression Ⅱ," *Biometrika*, 38(1951), 159－178 页。

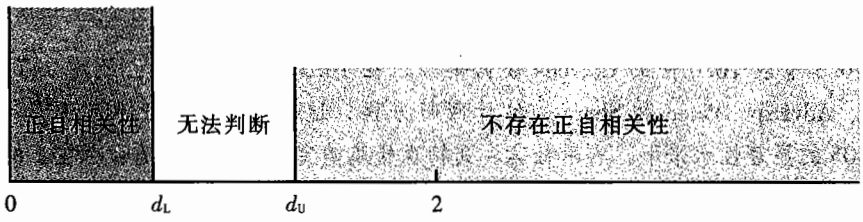


图 A：检测正自相关性

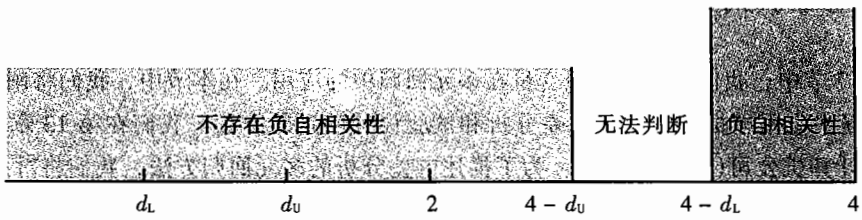


图 B：检测负自相关性

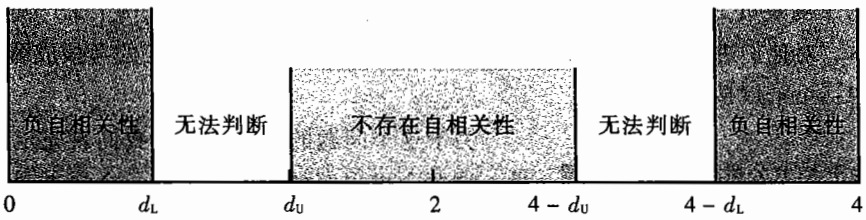


图 C：自相关性的双边检测

图 16.19　利用杜宾 – 瓦特森检验对自相关性假设进行检测

请注意，杜宾 – 瓦特森检验表所列出的最小样本容量为 15，因为当样本容量比较小时，检验通常是无效的。事实上，许多统计学家认为，如果要使检验结果比较有价值，样本容量至少应该为 50。

练　习

应　用

21.　考察练习 19 中的数据集合。



a. 建立一个估计回归方程，在已知边际利润的情况下，预测市盈率。

b. 根据(a)中所建立的估计回归方程，以观测数据出现的先后为顺序，画出残差散点图。在这些数据中是否存在自相关性？解释原因。

c. 在 $\alpha=0.05$ 的显著性水平下，检测数据中的正自相关性。

22. 参考 16.3 节表 16.6 中的 Cravens 数据集合，在 16.3 节中，我们已经建立了包含变量 Accounts、AdvExp、Poten 和 Share 的估计回归方程，这个方程的修正判定系数为 88.1%。在 $\alpha=0.05$ 的显著性水平下，利用杜宾 - 瓦特森检验确定这些数据中是否存在正的自相关性。

16.6 关于方差分析和试验设计的多元回归方法

在 15.7 节中，我们讨论了虚拟变量在多元回归中的应用。在本节中，我们将阐述另一种在多元回归方程中应用虚拟变量解决方差分析和试验设计问题的方法。我们在第 13 章中介绍了国民(National)计算机公司(NCP)的例子，为了描述方差分析的多元回归方法，我们将对 NCP 公司的示例应用这种方法。

回顾一下 NCP 公司的例子，这个公司在亚特兰大、达拉斯和西雅图建有工厂，生产打印机和传真机。为了测量员工对全面质量管理的理解程度，公司的管理人员从每个工厂随机抽取 6 名员工组成样本，对他们进行质量管理意识的考核。管理者计划用 18 名员工的考核成绩来确定平均考核成绩在每一个工厂是否相同。

表 16.9 关于 NCP 公司问题的虚拟变量

x_1	x_2	
0	0	观测值与亚特兰大的工厂相关
1	0	观测值与达拉斯的工厂相关
0	1	观测值与西雅图的工厂相关

关于这个问题的回归方法，我们首先定义两个虚拟变量，这两个虚拟变量用于表示每一个样本观测值是从哪一个工厂选取的。因为 NCP 公司有 3 个工厂，即 3 个总体，因此我们需要定义两个虚拟变量，一般情况下，如果正在研究的问题有 k 个不同的直观水平或者总体，则我们需要定义 $k-1$ 个虚拟变量。对于 NCP 公司的例子，我们定义了两个虚拟变量 x_1 和 x_2，如表 16.9 所示。

我们可以利用虚拟变量 x_1 和 x_2 将质量管理意识考核的成绩 y 与员工所在的工厂联系起来。

$$E(y)=\text{质量管理意识考核成绩的期望值}$$
$$=\beta_0+\beta_1x_1+\beta_2x_2$$

因此，如果我们对在亚特兰大工厂工作的员工的考核期望值感兴趣，则需要设虚拟变量 $x_1=x_2=0$。于是多元回归方程简化为

$$E(y)=\beta_0+\beta_1\times0+\beta_2\times0=\beta_0$$

这样我们就可以将 β_0 看做是亚特兰大工厂的员工考核成绩的期望值。

然后，我们考察其他工厂的多元回归方程形式，对于达拉斯的工厂，$x_1=1$，$x_2=0$，有

$$E(y)=\beta_0+\beta_1\times1+\beta_2\times0=\beta_0+\beta_1$$

对于西雅图的工厂，$x_1=0$，$x_2=1$，有

$$E(y)=\beta_0+\beta_1\times0+\beta_2\times1=\beta_0+\beta_2$$

我们可以看到 $\beta_0+\beta_1$ 表示的是达拉斯工厂的员工考核成绩的期望值，$\beta_0+\beta_2$ 表示的是西雅图工厂的员工考核成绩的期望值。

现在，我们估计系数 β_0，β_1，β_2 的数值，因此我们需要计算出每一个工厂考核成绩期望值的估计。将包含 x_1，x_2 和 y 的 18 个观测值的样本数据输入 Minitab。实际输入数据如表 16.10 所示，Minitab 的输出结果如图 16.20 所示。

表 16.10　NCP 问题的输入数据

亚特兰大			达拉斯			西雅图		
x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
0	0	85	1	0	71	0	1	59
0	0	75	1	0	75	0	1	64
0	0	82	1	0	73	0	1	62
0	0	76	1	0	74	0	1	69
0	0	71	1	0	69	0	1	75
0	0	85	1	0	82	0	1	67

```
The regression equation is
Y = 79.0 - 5.00 X1 - 13.0 X2

Predictor       Coef        Stdev    t-ratio        p
Constant      79.000        2.186      36.14    0.000
X1            -5.000        3.091      -1.62    0.127
X2           -13.000        3.091      -4.21    0.001

s = 5.354       R-sq = 54.5%     R-sq(adj) = 48.5%

Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     2      516.00      258.00      9.00    0.003
Error         15      430.00       28.67
Total         17      946.00
```

图 16.20　NCP 问题的多元回归输出结果

在图 16.20 中，我们可以看到 β_0，β_1，β_2 的估计值分别为 $b_0=79$，$b_1=-5$ 和 $b_2=-13$。因此，每一个工厂考核成绩的最佳估计期望值如下所示：

工厂	$E(y)$估计值
亚特兰大	$b_0=79$
达拉斯	$b_0+b_1=79-5=74$
西雅图	$b_0+b_2=79-13=66$

请注意，根据回归分析所得的每一个工厂考核成绩的最佳估计期望值与前面应用方差分析(ANOVA)所得的样本均值相同。

现在，让我们来考察一下怎样利用软件包的多元回归输出结果来对三个工厂的均值之间的差距进行 ANOVA 检测。首先，我们观察到，如果均值之间没有差距，则有

达拉斯工厂的期望值 $E(y)$ - 亚特兰大工厂的期望值 $E(y)=0$

西雅图工厂的期望值 $E(y)$ - 亚特兰大工厂的期望值 $E(y)=0$

由于亚特兰大工厂 β_0 等于 $E(y)$，达拉斯工厂 $\beta_0+\beta_1$ 等于 $E(y)$，所以第一个差距等于 $(\beta_0+\beta_1)-\beta_0=\beta_1$。除此之外，由于西雅图工厂 $\beta_0+\beta_2$ 等于 $E(y)$，所以第二个差距等于 $(\beta_0+\beta_2)-\beta_0=\beta_2$。如果 $\beta_1=0$，$\beta_2=0$ 则我们可以得出结论：三个工厂考核成绩的均值之间并没有任何差距。因此，对这三个工厂考核成绩期望值 $E(y)$ 之间的差距进行检验的零假设可以陈述如下：

$$H_0:\ \beta_1=\beta_2=0$$

回顾一下关于对回归关系的显著性进行检验的零假设，对于这种类型的假设，我们必须将 MSR/MSE 的数值与 F 分布的临界值进行比较，而且这个 F 分布的分子自由度应该等于回归平方和的自由度，分母自由度应该等于误差平方和的自由度。在当前的 NCP 示例中，回归平方和有两个自由度，误差平方和有 15 个自由度。因此，我们可以得到如下的 MSR 和 MSE 值：

$$\text{MSR}=\frac{\text{SSR}}{2}=\frac{516}{2}=258$$

$$\text{MSE}=\frac{\text{SSE}}{15}=\frac{430}{15}=28.67$$

于是，F 的计算值为

$$F=\frac{\text{MSR}}{\text{MSE}}=\frac{258}{28.67}=9.00$$

在 $\alpha=0.05$ 的显著性水平下，F 临界值为 3.68，它的分子自由度为 2，分母自由度为 15。因为 F 的观测值大于 3.68 的临界值，所以我们可以拒绝零假设 $H_0:\ \beta_1=\beta_2=0$，并得出结论：三个工厂的均值是不同的。除此之外，我们利用 p - 值方法也可以拒绝零假设 $H_0:\ \beta_1=\beta_2=0$，因为 p - 值 $=0.003<\alpha=0.05$。

练　习

方　法

23. 考察一个完全随机的设计，它包含四种情形：A、B、C 和 D。建立一个估计回归方


自测题

程，对这些数据进行分析，并定义所有的变量。

24. 一个随机群组设计包含 3 种情形，2 个群组，建立一个多元回归方程对这些数据进行分析，并定义所有的变量。

25. 某个双因素设计包含两个因素：A 和 B。其中，A 有两个水平，B 有三个水平，建立一个多元回归方程对这个设计进行分析，并定义所有的变量。

应　用

26. 
自测题

Jacobs 化学公司打算对原料在机器上搅拌的时间(分钟)进行估计，搅拌机器由 3 个不同的工厂生产。为了限制检测的成本，每一个生产厂商生产的机器都搅拌 4 批原料。搅拌时间如下所示：

生产厂商 1	生产厂商 2	生产厂商 3
20	28	20
26	26	19
24	31	23
22	27	22

a. 建立一个多元回归方程，对这些数据进行分析。

b. 回归模型中关于系数的最佳估计量是多少？

c. 根据回归模型中的系数，为了检测 3 家生产厂商所生产的机器搅拌时间是否一样，我们应该作出什么样的假设？

d. 对于 $\alpha = 0.05$ 的显著性水平，我们可以得出什么结论？

27. 有 4 种涂料产品的广告宣称它们的干燥时间一样，为了对生产厂商的宣传真实性作出检测，我们从每一种涂料产品里面都抽出 5 份作为样本进行检测。每一个样本从第一道工序开始干燥到可以应用第二道工序所需的平均时间记录如下：

涂料 1	涂料 2	涂料 3	涂料 4
128	144	133	150
137	133	143	142
135	142	137	135
124	146	136	140
141	130	131	153

a. 在 $\alpha = 0.05$ 的显著性水平下，检验 4 种涂料干燥所需的平均时间之间的差距。

b. 涂料 2 干燥所需的平均时间估计量是多少？这个估计量是怎样从计算机输出结果中得到的？

28. 一个汽车销售商对发动机进行调试，调试的设备有两种，一种是计算机控制的分析仪，另一

种是电子控制的分析仪，为了确定调试发动机所需要的时间是否取决于调试设备的种类，销售商做了一项检测。由于调试微型汽车、中型汽车和重型汽车发动机所需的时间都不一样，所以在检测中将汽车的类型作为分组根据，检测结果如下所示：

		车型		
		微型	中型	重型
分析仪	计算机控制	50	55	63
	电子控制	42	44	46

在 $\alpha=0.05$ 的显著性水平下，对调试所需时间的差别进行显著性检验。

29. 报刊上所刊登的广告，有尺寸设计大小和策划方式的不同，这种不同有可能影响到邮购商品的人数，为了检测这种不同，我们考察三种广告设计和两种尺寸大小，数据如下所示。在 $\alpha=0.05$ 的显著性水平下，检验因广告设计、尺寸大小以及相互作用所产生的显著影响。

		广告尺寸	
		小	大
设计	A	8	12
		12	8
	B	22	26
		14	30
	C	10	18
		18	14

小　结

在本章我们讨论了模型设计者用于确认最佳估计回归方程的一些概念。首先，我们介绍了一般线性模型的概念，并利用这个概念阐述了如何第将 14 章和第 15 章中所讨论的方法扩展到曲线关系和相互作用的处理上来。然后，我们讨论了如何利用因变量的变换来解决误差项中非常数方差的问题。

在回归分析的许多应用中，我们经常需要考虑大量的自变量。因此，我们介绍了基于 F 统计量的一般性方法，利用这些方法我们可以在回归模型中添加或者删除自变量。然后，我们介绍了包含 25 个观测值和 8 个自变量的大问题。可以看到，在自变量的数目很多时，迫切需要解决的问题是找出一个关于自变量的最佳子集。为了解决这个问题，我们讨论了几种变量选择的方法：逐步回归法、前向选择法、后向消元法和最佳子集回归法。

在 16.5 节中，我们将残差分析的应用扩展到了检测自相关性的杜宾－瓦特森检验。本章最后总结了如何利用多元回归模型来解决方差分析和实验设计问题的方法。

术语释义

一般线性模型：模型形式为 $y=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots+\beta_p z_p+\varepsilon$，对于其中每一个自变量 z_j，$j=1$，2，…，p 都是变量 x_1，x_2，…，x_k 的函数，所收集的数据都是关于 x_1，x_2，…，x_k 的。

交互作用：两个自变量共同作用所造成的影响。

变量选择方法：对一个回归模型选择自变量子集的方法。

自相关性：对于在时间上持续连贯的点，所出现的误差项之间的相关性。

序列相关：同自相关性。

杜宾－瓦特森检验：确定是否存在一阶自相关的检验。

重要公式

一般线性模型

$$y=\beta_0+\beta_1 z_1+\beta_2 z_2+\cdots+\beta_p z_p+\varepsilon \tag{16.1}$$

添加或者删除 $p-q$ 个变量的一般性 F 检验

$$F=\frac{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q)-\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_p)}{p-q}}{\dfrac{\mathrm{SSE}(x_1,x_2,\cdots,x_q,x_{q+1},\cdots,x_p)}{n-p-1}} \tag{16.13}$$

自相关误差项

$$\varepsilon_t=\rho\varepsilon_{t-1}+z_t \tag{16.16}$$

杜宾－瓦特森检验

$$d=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^2} \tag{16.17}$$

补充练习

30. 许多国际基金与美国的一些基金相比较而言投资价值取向更为理性。因为国际市场和美国国内市场的走势方向往往是不同的，在国际市场的投资往往能够降低投资者的整体风险。下表显示的是 20 家国际基金的基金类型(是否有抵押金)，管理费率(%)，安全等级(0＝最危险，10＝最安全)，以及从 1999 年 12 月 10 日以来的一年的业绩表现(*Mutual Funds*，


CD光盘数据
MutFunds

February 2000)。

	基金类型	管理费率(%)	安全等级	业绩(%)
ABN AMRO Int'l Equity "Com"	无抵押金	1.38	6.9	36
Accessor Int'l Equity "Adv"	无抵押金	1.59	7.1	42
Artisan International	无抵押金	1.45	6.8	72
Columbia Int'l Stock	无抵押金	1.56	7.1	54
Concert Inv. "A"Int'l Equity	有抵押金	2.16	6.3	116
Diversified Invstr Int'l Eqty	无抵押金	1.40	7.3	54
Driehaus Int'l Growth	无抵押金	1.88	6.5	92
Founders Passport	无抵押金	1.52	7.0	86
Guardian Baillie Fifford Int'l "A"	有抵押金	1.62	7.1	37
Jamestown Int'l Equity	无抵押金	1.56	7.1	35
Julius Baer Int'l Equity	无抵押金	1.79	6.9	71
Aetna "I"Int'l	无抵押金	1.35	7.3	46
Pilgrim Int'l Value "A"	有抵押金	1.80	7.1	42
Fidelity Diversified Int'l	无抵押金	1.48	7.5	42
Putnam "A"Int'l Growth	有抵押金	1.59	6.9	55
Sit Int'l Growth	无抵押金	1.50	6.9	49
Touchstone Int'l Equity "A"	有抵押金	1.60	7.5	35
United Int'l Growth "A"	有抵押金	1.28	7.1	47
Vontobel Int'l Equity	无抵押金	1.50	7.0	43
Waddell & Reed Int'l Growth "B"	有抵押金	2.46	7.0	75

a. 利用这一章所介绍的方法建立一个估计回归方程，根据所提供的数据估计基金的业绩。

b. (a)中所建立的估计回归方程是否给出了一个比较好的拟合？解释原因。

c. Acorn International 是一个免收发行费的基金，每年管理费率为 1.12%，安全等级为 7.6。利用(a)中所建立的估计回归方程估计 Acorn International 基金一年来的业绩。

31. 某项研究调查了审计延迟(从一个公司财务年度的末尾到审计报表发布的日期之间的时间)与描述客户和审计师的变量之间的关系。这项研究所涉及的一些自变量定义如下：

行业　　一个虚拟变量，如果值为 1，则代表该公司为实业公司；如果值为 0，则表示该公司为银行、存贷或者保险公司。

公开与否　　一个虚拟变量，如果值为 1，则代表该公司的证券交易可以通过证券交易所或者柜台进行；如果值为 0，则表示该公司不能通过上述方式交易。

质量　　内部全面质量控制的度量，由审计师裁定，有 5 个等级，从“实际空白”1 到“优秀”5。

完成与否　　由审计师裁定的一个度量，范围从 1 到 4，其中 1 代表“所有工作都是在年终以后才进行的”，4 代表“大部分工作在年终之前已经完成”。

从 40 个公司的样本中抽取了以下用于进行此项研究的数据：



CD 光盘数据
Audit

审计延迟	行业	公开与否	质量	完成与否
62	0	0	3	1
45	0	1	3	3
54	0	0	2	2
71	0	1	1	2
91	0	0	1	1
62	0	0	4	4
61	0	0	3	2
69	0	1	5	2
80	0	0	1	1
52	0	0	5	3
47	0	0	3	2
65	0	1	2	3
60	0	0	1	3
81	1	0	1	2
73	1	0	2	2
89	1	0	2	1
71	1	0	5	4
76	1	0	2	2
68	1	0	1	2
68	1	0	5	2
86	1	0	2	2
76	1	1	3	1
67	1	0	2	3
57	1	0	4	2
55	1	1	3	2
54	1	0	5	2
69	1	0	3	3
82	1	0	5	1
94	1	0	1	1
74	1	1	5	2
75	1	1	4	3

（续表）

审计延迟	行业	公开与否	质量	完成与否
69	1	0	2	2
71	1	0	4	4
79	1	0	5	2
80	1	0	1	4
91	1	0	4	1
92	1	0	1	4
46	1	1	4	3
72	1	0	5	2
85	1	0	5	1

a. 建立一个估计回归方程，包含所有的自变量。

b. (a)中所建立的估计回归方程是否给出了一个比较好的拟合？解释原因。

c. 将“审计延迟”作为“完成与否”的函数，画出关于它们的散点图。这个散点图表明“审计延迟”与“完成与否”之间存在什么关系？

d. 以“审计延迟”与“完成与否”之间的关系的观测值为基础，建立另一个可供选择的估计回归方程，与(a)中所建立估计回归方程相比，它能够解释“审计延迟”尽可能多的变异性。

32. 参考练习 31 中的数据，建立一个仅用“行业”来预测“审计延迟”的模型。在 $\alpha=0.01$ 的显著性水平下，对这些数据中的正自相关性进行检验。

33. 参考练习 31 中的数据。

a. 建立一个估计回归方程，以“行业”和“质量”为自变量，预测“审计延迟”。

b. 根据(a)中所建立的估计回归方程，将残差作为数据出现先后顺序的函数，画出残差图。在数据中是否存在自相关性？解释原因。

c. 在 $\alpha=0.05$ 的显著性水平下，对这些数据中的正自相关性进行检验。

34. 我们对购物者逛商店的行为进行了一项研究，可以把购物者分成几类：逛商店是有目的的，准备购买商品；目的性不是很明确，碰到合适的商品也会购买；仅仅是为了消遣而逛商店，不准备购买商品。对于研究中的每一类购物者，有一个度量标准反映了她在商店中感到舒适的程度。分值越高，则舒适程度越高。假设以下的数据即来自这项研究，在 $\alpha=0.05$ 的显著性水平下，检测三种类型购物者舒适程度之间的差异。



CD 光盘数据 Browsing

准备购物	有可能购物	不准备购物
4	5	5
5	6	7
6	5	5
3	4	7

（续表）

准备购物	有可能购物	不准备购物
3	7	4
4	4	6
5	6	5
4	5	7

35. 下面的数据显示了道琼斯工业平均指数(DJIA)在 8 任总统的每一个四年任期之内的变化(*1998 Stock Trader's Almanac*)，利用回归分析来确定政党和总统任期对道琼斯工业平均指数变化的影响。



CD 光盘数据
Election

总统	政党	当选年	任期年份	DJIA 每年的变化
约翰逊	民主党	1964	第一年	10.9
约翰逊	民主党		第二年	-18.9
约翰逊	民主党		第三年	15.2
约翰逊	民主党		第四年	4.3
尼克松	共和党	1968	第一年	-15.2
尼克松	共和党		第二年	4.8
尼克松	共和党		第三年	6.1
尼克松	共和党		第四年	14.6
尼克松	共和党	1972	第一年	-16.6
尼克松	共和党		第二年	-27.6
尼克松*	共和党		第三年	38.3
尼克松*	共和党		第四年	17.9
卡特	民主党	1976	第一年	-17.3
卡特	民主党		第二年	-3.1
卡特	民主党		第三年	4.2
卡特	民主党		第四年	14.9
里根	共和党	1980	第一年	-9.2
里根	共和党		第二年	19.6
里根	共和党		第三年	20.3
里根	共和党		第四年	-3.7
里根	共和党	1984	第一年	-27.7
里根	共和党		第二年	22.6
里根	共和党		第三年	2.3
里根	共和党		第四年	11.8
布什	共和党	1988	第一年	27.0
布什	共和党		第二年	-4.3
布什	共和党		第三年	20.3
布什	共和党		第四年	4.2

（续表）

总统	政党	当选年	任期年份	DJIA 每年的变化
克林顿	民主党	1992	第一年	13.7
克林顿	民主党		第二年	2.1
克林顿	民主党		第三年	33.5
克林顿	民主党		第四年	26.0

* 由于尼克松总统在 1974 年 8 月辞职，福特接任总统并完成了尼克松总统的任期。

案例研究 1　失业问题研究

一项研究提供了与制造业工人失业周数相关的一些变量的数据，在这项研究中，因变量(weeks)被定义为工人由于被解雇而失业的周数，同时在研究中应用到了以下的自变量：

Age	工人的年龄
Educ	受教育的年限
Married	一个虚拟变量：1 表示已婚，0 表示未婚
Head	一个虚拟变量：1 表示该工人为户主，0 表示该工人不是户主
Tenure	原来工作的工龄
Manager	一个虚拟变量：1 表示所做的工作为管理工作，0 表示不是管理工作
Sales	一个虚拟变量：1 表示所做的工作为销售工作，0 表示不是销售工作

假设下面的数据来自 50 个失业工人。

管理报告


Layoffs

利用这一章和前面章节中所阐述的方法，对这个数据集合进行分析。并在管理报告中进行总结，包括关键的统计结果、结论和建议。并且总结还应该在附录中给出必需的技术资料(计算机输出结果、残差图，等等)。

失业周数	年龄	受教育年限	婚姻状况	是否户主	原工龄	是否做管理工作	是否为销售人员
37	30	14	1	1	1	0	0
62	27	14	1	0	6	0	0
49	32	10	0	1	11	0	0
73	44	11	1	0	2	0	0
8	21	14	1	1	2	0	0
15	26	13	1	0	7	1	0
52	26	15	1	0	6	0	0
72	33	13	0	1	6	0	0
11	27	12	1	1	8	0	0

（续表）

失业周数	年龄	受教育年限	婚姻状况	是否户主	原工龄	是否做管理工作	是否为销售人员
13	33	12	0	1	2	0	0
39	20	11	1	0	1	0	0
59	35	7	1	1	6	0	0
39	36	17	0	1	9	1	0
44	26	12	1	1	8	0	0
56	36	15	0	1	8	0	0
31	38	16	1	1	11	0	1
62	34	13	0	1	13	0	0
25	27	19	1	0	8	0	0
72	44	13	1	0	22	0	0
65	45	15	1	1	6	0	0
44	28	17	0	1	3	0	1
49	25	10	1	1	1	0	0
80	31	15	1	0	12	0	0
7	23	15	1	0	2	0	0
14	24	13	1	1	7	0	0
94	62	13	0	1	8	0	0
48	31	16	1	0	11	0	0
82	48	18	0	1	30	0	0
50	35	18	1	1	5	0	0
37	33	14	0	1	6	0	1
62	46	15	0	1	6	0	0
37	35	8	0	1	6	0	0
40	32	9	1	1	13	0	0
16	40	17	1	0	8	1	0
34	23	12	1	1	1	0	0
4	36	16	0	1	8	0	1
55	33	12	1	0	10	0	1
39	32	16	0	1	11	0	0
80	62	15	1	0	16	0	1
19	29	14	1	1	12	0	0
98	45	12	1	0	17	0	0
30	38	15	0	1	6	0	1
22	40	8	1	1	16	0	1
57	42	13	1	0	2	1	0
64	45	16	1	1	22	0	0
22	39	11	1	1	4	0	0
27	27	15	1	0	10	0	1
20	42	14	1	1	6	1	0
30	31	10	1	1	8	0	0
23	33	13	1	1	8	0	0

案例研究 2　职业高尔夫巡回赛统计量分析


CD 光盘数据
PGATour

职业高尔夫球手联盟(PGA)保留有参加巡回赛的球手的比赛成绩和收入的统计资料。1997 年年底比赛成绩统计资料被取名为 PGA Tour 数据集合(*Golfweek*, November 15, 1997)。数据集合的每一行对应着一名职业高尔夫巡回赛球手，并且排名的顺序代表着选手参加 1997 年 PGA 巡回赛每一项赛事的收入高低，每一轮比赛包括 18 个高尔夫球洞。PGA 巡回赛的每一项赛事通常由安排在 4 天内的 4 场比赛组成。星期四和星期五两轮比赛的得分被用来限制星期六和星期日两轮比赛的场地规模。因此，选手只有在星期四和星期五两轮比赛中取得足够得分，才有资格参加星期六和星期日的这两轮比赛，即通过了资格赛，这种赛制被称为资格赛选拔。

参赛总收入	参加整个 PGA 巡回赛的总收入
参赛次数	参加 PGA 巡回赛赛事的数量
参赛轮数	参加 PGA 巡回赛所完成赛事的轮数
资格选拔次数	选手通过资格赛选拔的赛事数量
60 分赛事次数	参赛得分为 60 的赛事数量
低于标准杆数的赛事次数	参赛得分低于标准杆数的赛事数量
平均击球次数	在每一轮比赛中得分少于标准杆数的平均击球次数
平均科分	所有赛事的平均得分
码数	每次击球所达到的平均码数(击球长度是从击球点到球停止滚动之间的距离，不考虑高尔夫球是否落在高尔夫球道里。击球距离是在每次比赛的两个球洞之间测量)
入球道比	每轮比赛中高尔夫球手从发球处把球发出以后，球能够落在高尔夫球道里的次数
入场内比	高尔夫球手能够将高尔夫球按规则击入球洞场地的平均次数，(如果高尔夫球的任意部分接触到球洞场地的轻击区域，并且球入洞的杆数与标准杆数之间的差距多于 2 杆，则我们应该根据比赛规则将比赛场地列入考虑因素范围之内。)
进洞比	根据比赛规则，高尔夫球手将球洞场地上的高尔夫球轻击入洞的平均杆数
救球比	高尔夫球被打入球场旁边的沙地时，高尔夫球手可以克服地面起伏不平的困难将球救回的比例(克服地面起伏不平指的是：最多只能用两杆，即将球救出沙洞，并击入球洞中。)

管理报告

假设你被 PGA 巡回赛的行政管理人员聘用，你的职责是在每年一度的 PGA 巡回赛宴会上介绍赛事数据分析。PGA 巡回赛的管理人员会询问，是否可以利用这些数据，比如资格选拔次数、码数、入球道比等等，来确定高尔夫球手的比赛成绩，这些数据对高尔夫球手得分和比赛收入有什么样的影响？利用前面章节和本章所介绍的方法对这些数据进行分析，并为 PGA 巡回赛的管理人员准备一份报告，在报告中简明扼要地介绍你的分析，包括关键的统计结果、结论和建议，在附录中应该给出你认为比较合适的技术资料。

案例研究 3　预测大学的毕业率



CD 光盘数据
GradRate

对大学的管理部门来说，招收进入学院或者大学的学生能够顺利毕业的比率是一个很重要的统计量。与大学学生顺利毕业率相关的一些因素包括：人数少于 20 的班级比例，人数多于 50 的班级比例，师生比，申请这所大学的学生被录取的人数占总申请人数的比例，大一学生中高中成绩位于所在高中前 10% 的学生比例，以及这所大学的学术声誉。为了研究这些因素对毕业率的影响，我们收集了 48 所全国性大学的数据资料（*America's Best Colleges*，2000 年版）对这些数据资料的相关描述如下：

地区	该所大学位于所在区域的哪一部分(北部，南部，中西部，西部)
毕业率	招收进入这所大学的学生顺利毕业的比例
人数小于 20 的班级比例	班级人数少于 20 的班级比例
人数多于 50 的班级比例	班级人数多于 50 的班级比例
师生比	大学里面学生总人数除以教员的总人数
录取率	申请这所大学的学生被录取的人数占总申请人数的比例
前 10% 的学生所占比例	大一学生中高中成绩位于所在高中前 10% 的学生比例
学术声誉分数	由其他大学的调查员评估的关于这所大学的学术声誉，分值从 1(勉强够格)到 5(杰出)

管理报告

利用前面章节和本章所介绍的方法对这些数据进行分析，给出一份总结性的管理报告，包括关键的统计结果、结论和建议，在附录中给出合适的技术资料(计算机输出结果、残差图，等等)。

17 指 数

统计实例：美国劳工部劳工统计局

17.1 **比价**

17.2 **综合物价指数**

17.3 **根据比价计算综合物价指数**

17.4 **一些重要的价格指数**

消费者价格指数

生产者价格指数

道琼斯股票平均价格指数

17.5 **根据物价指数减缩序列**

17.6 **物价指数：其他一些相关问题**

商品项目的选择

基期的选择

质量变化

17.7 **物量指数**

统计实例

美国劳工部劳工统计局

华盛顿

美国劳工部通过其下属的劳工统计局编纂并发布各类指数和其他统计资料，这些指数和统计资料成为美国商业和经济的晴雨表。它们包括：消费者价格指数、生产者价格指数、各类工人的平均工作时间与收入等，其中引用的最为广泛的指数应该是消费者价格指数。它经常被认为是测量通货膨胀的尺度。

2000 年 9 月，劳工统计局报告显示：消费者价格指数(CPI)相比 7 月份的水平下降了 0.1%。这样的下降在 14 年内尚为首次，另一方面，劳工统计局同时指出，基础通货膨胀率在 8 月份却上涨了 0.2%。基础通货膨胀率排除了消费者价格指数中的挥发性食品和能源成分，有时被认为是对通货膨胀压力的较好的反映。能源指数下降了 2.9%，食品指数增加了 0.2%。

许多经济学家和分析师认为，消费者价格指数表明美国已经进入了非通货膨胀时代。他们指出，作为对消费者价格指数的补充，薪资水平上涨的趋势极其有限。劳工部也报告说：在前两个月保持不变的情况下，8 月份的每周平均收入经过通货膨胀调整后仅仅上升了 0.1%。

另外一个指数也表明在不久的将来通货膨胀压力将会降低，这个指数就是生产者价格指数。生产者价格指数可以测量批发市场的价格变化，并且通常被认为是消费者价格指数变动的最重要的反映。生产者价格指数在 8 月份下降了 0.2%。联邦储备委员会的决策者们从 1999 年以来已经 6 次提高银行贷款利率，但是，由于这些数据表明价格并没有迅速上升，因此联邦储备委员会估计将会保持近期利率不变。



美国劳工统计局根据所收集的商品和服务的销售价格计算消费者价格指数。© CORBIS.

本章中，我们将会讨论各种不同的指数，比如消费者价格指数和生产者价格指数，是怎么计算出来的以及怎样对它们作出解释。

美国政府每个月都会发布许多不同的指数，以帮助每一个公民了解当前商业和经济形势。这些指数中了解和应用最为广泛的应该是消费者价格指数(CPI)。正如它的名字所代表的，消费者价格指数反映的是消费者支付其所选购的商品项目的价格变动。特别地，消费者价格指数可以测量某一段时间内的价格变动。如果给定一个起始时期，也称为**基期**，其指数为 100，则消费者价格指数能够比较当前时期和基期的消费价格。例如，消费者价格指数为 125，则该指数表明当前期的消费价格在整体上比基期的消费价格提高了 25%。虽然只有很少的市民确切明白这个数字的含义，但是他们肯定知道这个指数的增加意味着价格的上升。

虽然消费者价格指数可能是最为著名的指数，但是还有许多其他的政府和私人机构的指数，它们也可以帮助我们测量和理解某个时期的经济状况和另外一个时期的经济状况的比较。本章的目的就在于阐述使用最为广泛的几种指数。我们可以构造一些简单指数，这样有助于理解指数的计算过程。

表 17.1　无铅汽油的价格

年份	每加仑价格(美元)
1984	1.21
1985	1.20
1986	0.93
1987	0.95
1988	0.95
1989	1.02
1990	1.16
1991	1.14
1992	1.13
1993	1.11
1994	1.11
1995	1.15
1996	1.23
1997	1.23
1998	1.06

资料来源：*Monthly Energy Review*, U. S. Energy Administration。

表 17.2　每加仑无铅汽油的比价(1984—1998)

年份	每加仑价格(美元)
1984	(1.21/1.21)100 = 100.0
1985	(1.20/1.21)100 = 99.2
1986	(0.93/1.21)100 = 76.9
1987	(0.95/1.21)100 = 78.5
1988	(0.95/1.21)100 = 78.5
1989	(1.02/1.21)100 = 84.3
1990	(1.16/1.21)100 = 95.9
1991	(1.14/1.21)100 = 94.2
1992	(1.13/1.21)100 = 93.4
1993	(1.11/1.21)100 = 91.7
1994	(1.11/1.21)100 = 91.7
1995	(1.15/1.21)100 = 95.0
1996	(1.23/1.21)100 = 101.7
1997	(1.23/1.21)100 = 101.7
1998	(1.06/1.21)100 = 87.6

17.1　比价

比价指数最简单的形式，即某个给定商品项目当前单价与基期同一商品项目单价的比值。例如，表 17.1 显示的是 1984 年到1998 年的每加仑无铅汽油的价格。为了与其他时期的价格进行比较，每加仑汽油的实际价格可以转化成**比价**(price relative)。比价表示的是当前时期单位产品价格与基期单位产品价格的百分比。

$$t\text{ 时期的比价} = \frac{t\text{ 时期的价格}}{\text{基期价格}}(100) \qquad (17.1)$$

根据表 17.1 中的汽油价格，以 1984 年作为基年，我们可以计算从 1984 年到 1998 年每加仑无铅汽油的比价。计算结果列于表 17.2 中。以比价为基础，我们可以很容易地将任何时期的价格与基期的价格进行比较。例如，1986 年的比价为 77，这表明 1986 年的汽油价格比 1984 年基期的价格低 23%。同样，1998 年的比价为 88，表明 1998 年的汽油价格比 1984 年基期的汽油价格下降了 12%。就像运用无铅汽油这样的比价，在理解和解释经济和商业情况随时间的推移而发生的变化方面是很有帮助的。

17.2　综合物价指数

虽然比价可以用于确定单个商品随着时间的推移而发生的价格变化，但是我们通常对商品总体价格的变化更感兴趣。例如，我们需要一个指数来测量生活费用随着时间的推移所发生的总体变化，我们希望这个指数能够以各种生活必需品为基础，包括食物、住房、服装、交通和医药等等。**综合物价指数**(aggregate price index)就是为了满足反映许多商品价格整体变动这一特定目的而产生的。

以正常的汽车消耗费用为例来考虑综合物价指数的编制。在这个产品分类项里，我们只考虑汽油、石油、轮胎和保险支出。

表 17.3 给出了在 1984 年到 1998 年期间上述 4 种汽车消耗费用的资料。以 1984 年作为基期，四种成分的综合物价指数可以反映出从 1984 年到 1998 年期间汽车正常消耗费用的变化。

未加权综合物价指数是对象期(例如 1998 年)的单价之和除以基期(1984 年)的单价之和。令

P_{it} = t 时期 i 商品的单价；

P_{i0} = 基期 i 商品的单价。

表 17.3 汽车消耗费用指数的数据资料

商品项目	单价(美元)	
	1984	**1998**
汽油（加仑）	1.21	1.06
石油（夸脱）	1.50	2.20
轮胎	80.00	145.00
保险费用	300.00	700.00

t 时期的未加权综合物价指数以 I_t 表示，计算公式为

$$I_t = \frac{\sum P_{it}}{\sum P_{i0}}(100) \tag{17.2}$$

式中，求和是针对该整体中所有项目而言的。

1998 年汽车正常消耗费用的未加权综合物价指数计算如下：

$$I_{1998} = \frac{1.06 + 2.20 + 145.00 + 700.00}{1.21 + 1.50 + 80.00 + 300.00}(100)$$

$$= \frac{848.26}{382.71}(100)$$

$$= 222$$

如果各个商品项的用量相同，未加权物价指数等于加权物价指数。但是在实际中，各个商品项的用量相同是很罕见的。

根据未加权的综合物价指数，我们可以得出结论：从 1984 年到 1998 年，汽车正常消耗费用的价格增加了 122%。但是请注意，汽车消耗费用的未加权综合物价指数近似于复合物价指数，它受单价高的商品项影响严重。因此，单价相对较低的商品项，比如汽油和石油，受轮胎和保险支出这样的单价较高的商品项所制约。汽车正常消耗费用的未加权综合物价指数受轮胎和保险支出价格变动的影响极为严重。

由于未加权指数受单价高的商品项影响，所以这种形式的综合物价指数应用并不广泛。当商

品项用量不同时，加权综合物价指数能够提供更好的对比。

表 17.4　汽车消耗费用指数的每年使用信息

商品项目	权值*
汽油（加仑）	1 000
石油（夸脱）	15
轮胎	2
保险费用	1

*根据每年行驶 15 000 英里确定，轮胎的使用根据每年行驶 30 000 英里确定。

加权综合物价指数(weighted aggregate price index)的本质在于：商品组合中的每一项应该根据它的重要性而赋予不同的权值。在大多数情况下，商品的使用数量是其重要性的最好的测量尺度。因此，我们必须得到商品组中各项不同商品的使用数量资料。表 17.4 是一辆中型汽车每年行驶 15 000 英里时所消耗费用的各项使用信息。表中所列的数量权值表明了此车的行驶情况的每年预计使用量。

令 $Q_i = i$ 项商品的使用数量。t 时期的加权综合物价指数计算如下：

$$I_t = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i}(100) \tag{17.3}$$

式中，求和是针对该整体中所有的商品项。对应于汽车消耗费用资料，加权综合物价指数等于 1998 年所有的消耗费用除以 1984 年所有的消耗费用。

令 $t = 1998$，并利用表 17.4 中的数量权值，我们可以得到 1998 年汽车消耗费用的加权综合物价指数如下：

$$\begin{aligned} I_{1998} &= \frac{1.06(1\,000) + 2.20(15) + 145.00(2) + 700.00(1)}{1.21(1\,000) + 1.50(15) + 80.00(2) + 300.00(1)}(100) \\ &= \frac{2\,083}{1\,692.5}(100) \\ &= 123 \end{aligned}$$

从这个加权综合物价指数，我们可以得出结论：从 1984 年到 1998 年，汽车消耗费用的价格已经上涨了 23%。

很明显，和未加权的综合物价指数相比，加权物价指数对汽车消耗费用在 1984 年到 1998 年期间的价格变动反映得更为精确。考虑汽油的使用数量有助于弥补保险支出的大量增加，加权指数所显示的汽车消耗费用增加比未加权指数所显示的要更为平缓。总的说来，加权综合物价指数以使用数量作为权值，因而它更适合于编制多个商品的价格指数。

在加权综合物价指数公式(17.3)里，数量 Q_i 没有用来表示时期的下标，原因在于数量 Q_i 被认为是固定的，而且不像价格那样随着时间变动而变动。固定的权值和数量是由指数的设计者指定的，其水平可以代表典型的使用情况。一旦确立下来，在整个指数使用期间它们都将保持固定不变。如果要计算 1998 年以外其他年份的指数，还需要收集新的价格数据 P_{it}，但是权值 Q_i 保持不变。

固定加权综合指数的一个特殊情形是：数量取决于基期的用量。在这种情况下，我们有：$Q_i = Q_{i0}$，其中下标 0 表示基期。于是公式(17.3)变为

$$I_t = \frac{\sum P_{it} Q_{i0}}{\sum P_{i0} Q_{i0}}(100) \tag{17.4}$$

任何时候只要根据基期的用量确定了固定的数量权值，则加权综合物价指数被称为**拉氏指数**(Laspeyres index)。

另外一个确定数量权值的项目是修正每一个时期的数量。每年计算指数时所用的数量以 Q_{it} 表示，则带有这些数量权值的 t 期加权综合物价指数为

$$I_t = \frac{\sum P_{it}Q_{it}}{\sum P_{i0}Q_{it}}(100) \qquad (17.5)$$

请注意对基期(0 时期)和 t 时期，所使用的数量权值是一样的。但是，这些权值是以 t 时期的使用情况为基础的，而不是以基期为基础。这个加权综合物价指数被称为**帕氏指数**(Paasche index)。它所具备的优点是：以当前的使用情况为基础。然而，这种计算加权综合物价指数的方法也有两个缺点：通常的使用数量 Q_{it} 必须每年重新确定，因此，增加了数据收集的时间和成本，并且每年必须重新计算前一年的指数，以反映新的数量的权重的影响。因为这些缺陷，拉氏指数的使用变得更为广泛。汽车消耗费用指数是以基期数量为基础计算的，因此，它也是拉氏指数。如果利用 1998 年的使用数据，我们就会得到帕氏指数。事实上，由于汽车的节油性能提高，汽油的耗量已经降低，这样帕氏指数就会不同于拉氏指数。

练　习

方　法

1. 下表是某两种商品在 1989 年和 2001 年的价格和用量：

自测题

	数量		单价(美元)	
商品项	1989	2001	1989	2001
A	1 500	1 800	7.50	7.75
B	2	1	630.00	1 500.00

a. 计算每一项商品在 2001 年的比价，以 1989 年为基期。

b. 计算这两项商品在 2001 年的未加权综合物价指数，以 1989 年作为基期。

c. 用拉氏指数计算两项商品的加权综合物价指数。

d. 用帕氏指数计算两项商品的加权综合物价指数。

2. 2001 年，某项商品比价为 132，价格为 10.75 美元，基期为 1990 年。

a. 在这 11 年期间，该商品价格涨幅或者跌幅的百分比是多少？

b. 1990 年该商品价格是多少？

应 用

3. 

某个大制造商从三个独立的供应商处采购完全一样的原料，三个供应商供应的单价和数量都不同，1999 年和 2001 年的相关数据如下：

		单价(美元)	
供应商	数量	1999	2001
A	150	5.45	6.00
B	200	5.60	5.95
C	120	5.50	6.20

a. 分开计算每一个原料供应商的比价，比较每一个供应商每两年的价格涨幅。

b. 计算 2001 年该原料的未加权综合物价指数。

c. 计算 2001 年该原料的加权综合物价指数，这个指数对制造商而言意味着什么？

4. R&B 饮料公司通过它设在衣阿华市中心的零售商店分销它的包括啤酒、白酒和软饮料产品在内的全部产品。1997 年和 2001 年的单价数据以及 1997 年成箱销售的数量如下所示：

	1997 年销售量	单价(美元)	
商品项	(箱)	1997	2001
啤酒	35 000	15.00	16.25
白酒	5 000	60.00	64.00
软饮料	60 000	9.80	10.00

计算 2001 年 R&B 饮料公司销量的加权综合物价指数，以 1997 年为基期。

5. 根据后进先出存货估计法(LIFO)，以纳税为目的建立存货的价格指数。数量权值以年末的存货水平为基础，以年初的单价作为基期价格，建立年底总库存额的加权综合物价指数。对于后进先出存货估计法，必须建立什么类型的加权综合物价指数？

		单价(美元)	
产品	年终库存量	年初	年终
A	500	0.15	0.19
B	50	1.60	1.80
C	100	4.50	4.20
D	40	12.00	13.20

17.3 根据比价计算综合物价指数

价格和数量必须是相对应的。例如，如果给定的价格是每箱的价格，则数量必须为箱数，而不是其他的单个单位。

在 17.1 节中，我们给出了比价的定义，并且阐述了如何根据当前时期的单价和基期的单价来计算比价。现在我们希望能够像 17.2 节建立综合物价指数那样，根据整体中每一项商品项的比价直接计算综合物价指数。由于未加权指数的用途有限，我们集中讨论加权综合物价指数。回顾前一节中的汽车消耗费用指数，表 17.5 给出了所需四种费用的必要信息。

令 w_i 为应用于第 i 项商品的比价的权值。则加权平均物价指数的一般性计算公式如下：

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)w_i}{\sum w_i} \tag{17.6}$$

在公式(17.6)中选择适当的权值可以使我们根据比价来计算加权综合物价指数。将用量与基期价格的乘积作为权值，即

$$w_i = P_{i0}Q_i \tag{17.7}$$

表 17.5 汽车消耗费用指数的比价

	单价			
商品项目	**1984(P_0)**	**1998(P_t)**	**比价(P_t/P_0)100**	**年用量**
汽油（加仑）	1.21	1.06	87.6	1 000
石油（夸脱）	1.50	2.20	146.7	15
轮胎	80.00	145.00	181.3	2
保险费用	300.00	700.00	233.3	1

将 $w_i = P_{i0}Q_i$ 代入公式(17.6)中，得到加权平均物价指数

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)(P_{i0}Q_i)}{\sum P_{i0}Q_i} \tag{17.8}$$

将分子中的 P_{i0} 项消掉，可得加权平均物价指数的另一个表达式

$$I_t = \frac{\sum P_{it}Q_i}{\sum P_{i0}Q_i}(100)$$

因此，我们可以得到以 $w_i = P_{i0}Q_i$ 为权数的加权平均物价指数，它和 17.2 节中公式(17.3)所表述的加权综合物价指数一样。在公式(17.7)中利用基期数量(即 $Q_i = Q_{i0}$)可以导出拉氏指数。在公式(17.7)中利用当前期数量(即 $Q_i = Q_{it}$)可以导出帕氏指数。

回顾汽车消耗费用的例子，我们可以利用表 17.5 中的比价和公式(17.6)来计算加权平均物价指数，用公式(17.7)所指定的权数计算出来的结果列于表 17.6 中。指数 123 表示汽车消耗费用增长了 23%，与 17.2 节所计算出来的加权综合指数结果相同。

表 17.6　根据加权比价计算的汽车消耗费用指数(1984—1998)

商品项目	比价 $(P_{it}/P_{i0})(100)$	基期价格(美元) P_{i0}	数量 Q_i	权值 $w_i = P_{i0}Q_i$	加权比价 $(P_{it}/P_{i0})(100)\,w_i$
汽油	87.6	1.21	1 000	1 210.0	105 996.00
石油	146.7	1.50	15	22.5	3 300.75
轮胎	181.3	80.00	2	160.0	29 008.00
保险费用	233.3	300.00	1	300.0	69 990.00
		$I_{1998} = \frac{208\,294.75}{1\,692.5} = 123$		总计 1 692.5	208 294.75

练　习

方　法

6. 下表显示的是三项商品的比价、基价和用量，计算当前时期的加权综合物价指数。

自测题

		基期	
商品项目	比价	基价	数量
A	150	22.00	20
B	90	5.00	50
C	120	14.00	40

应　用

7. Mitchell 化学公司生产一种特殊的工业化学品，它是由三种化学成分混合而成。年初和年底每磅产品的价格以及混合比例如下所示：

自测题

成分	每磅产品价格(美元) 年初	年底	每100磅产品的用量(磅)
A	2.50	3.95	25
B	8.75	9.90	15
C	0.99	0.95	60

a. 计算3种成分的比价。

b. 为了求得一年生产中所需要的原材料的价格指数，请计算加权平均物价指数。对这个指数值，你如何解释?

8. 某个证券投资组合由四种股票组成。购入时的价格、当前价格和持股数如下表所示:

股票	购入价格(美元)	当前价格(美元)	持股数
Holiday Trans	15.50	17.00	500
NY Electric	18.50	20.25	200
KY Gas	26.75	26.00	500
PQ Soaps	42.25	45.50	300

计算这个证券投资组合的加权平均价格指数，并解释它。

9. 计算练习4中的R&B饮料公司产品的比价。利用加权平均物价指数来证明：通过这种方法所计算出来的指数与加权综合法计算结果是一样的。

17.4 一些重要的价格指数

我们已经介绍了用于计算多个或单个商品项目物价指数的方法。现在，我们来考察一些价格指数，它们都是反映商业和经济状况的重要指标。重点考察的对象是消费者价格指数、生产者价格指数和道琼斯平均股票价格指数。

消费者价格指数

消费者价格指数包括服务费(例如医生和牙医的收费)，以及所有与购买和使用商品直接相关的税费。

消费者价格指数(consumer price index，CPI)，由美国劳工统计局每月定期发布，是美国生活消费价格水平首要的反映。用于计算这个指数的商品组合是一个由400种商品所组成的**购物篮**，里面包括食品、住房、服装、交通以及医药保健等等。消费者价格指数是一个加权综合物价指

数，它里面的权数都是固定的。*购物篮中每一项商品的权数都是通过对全美城市家庭的使用情况作出调查后确定的。

2000 年 8 月的消费者价格指数是 172.7，以 1982—1984 年的指数 100 作为基期指数。这个数值意味着从 1982—1984 年的基期到现在，购买购物篮中的商品和服务的价格已经上升了 72.7%。图 17.1 显示的是从 1950 年到 2000 年 50 年间的消费者价格指数时间序列。请注意，从图像上可以看出来，消费者价格指数反映出了 20 世纪 70 年代后期和 80 年代初期剧烈的通货膨胀现象。

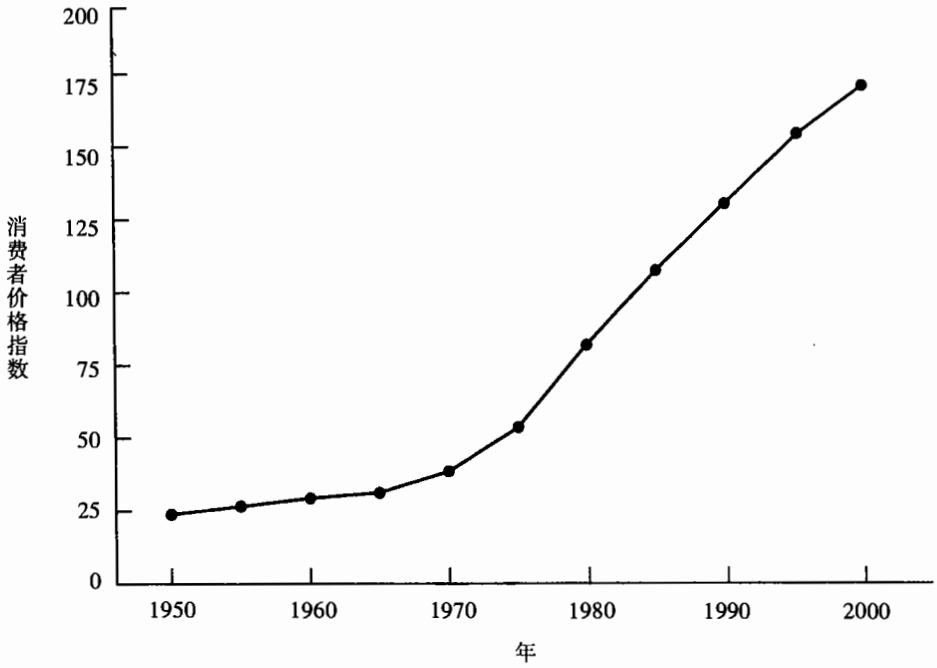


图 17.1　1950—2000 年消费者价格指数(以 1982—1984 年为基期，指数为 100)

生产者价格指数

生产者价格指数(producer price index，PPI)，也是由美国劳工统计局每月定期发布，它可以度量美国初级市场每个月价格的变动。生产者价格指数的计算以每种产品在非零售市场上的初次交易价格为基础。所有在这些市场上进行商品交换的商品都被包括在内。调查覆盖了原材料、经过制造的和在各个加工阶段上经过加工的商品，并且包括了制造业、农业、林业、渔业、矿业、煤气和电力以及公共事业的各类产出。这个指数常用于反映消费者价格指数和生活消费价格的未

* 实际上有两个消费者指数。美国劳工统计局公布的消费者价格指数里面，一个是所有城市消费者价格指数(CPI-U)，另一个是对城市工薪阶层和产业工人修正后的消费者价格指数(CPI-W)。其中城市消费者价格指数应用更为广泛，它由《华尔街日报》定期发布。

来趋势。生产者价格指数的增长表明了生产价格的上涨，而这种上涨最终要通过零售价格的上涨传递给消费者。

生产者价格指数中不同商品项的权数以货物价值为基础。加权平均价格指数由拉氏公式计算出来。2000 年 8 月的生产者价格指数是 138.1，以 1982 年的指数 100 为基期指数。

生产者价格指数只测量国内产品价格的变动；进口产品不包括在计算范围之内。

道琼斯股票平均价格指数

道琼斯股票平均价格指数（Dow Jones averages）反映的是纽约股票交易所股票交易价格的波动和变化趋势。道琼斯股票平均价格指数中最著名的是 30 种道琼斯工业股票平均价格指数(DJIA)，它是根据 30 个大公司的普通股票价格计算出来的。将这些股票价格的总和除以一个数，这个数因为拆股和公司之间的股票转换而需要一次次地重新计算。和我们所学过的其他价格指数不一样的是，它不是以基期价格的百分比来表示。表 17.7 显示的是 2000 年 9 月计算道琼斯工业股票平均价格指数所指定的 30 个公司。

查尔斯·亨利·道(Charles Henry Dow)于 1884 年 7 月 3 日在 *Customer's Afternoon Letter* 上发表了他的第一个股票平均价格指数。其中包括 11 种股票，里面有 9 种是铁路股票。1928 年 10 月 1 日首次发布了道琼斯 30 种工业股票平均价格指数。

表 17.7　道琼斯工业股票平均价格指数计算所用的 30 家公司(2000 年 9 月)

Alcoa	Exxon Mobil	J. P. Morgan Chase
American Express	General Electric	McDonald's
AT&T	General Motors	Merck
Boeing	Hewlett-Packard	Microsoft
Caterpillar	Home Depot	Minnesota Mining
Citigroup	Honeywell Int'l	Philip Morris
Coca Cola	IBM	Procter & Gamble
Disney	Intel	SBC Communications
DuPont	International Paper	United Technologies
Eastman Kodak	Johnson & Johnson	Wal-Mart Stores

资料来源：*Barron's*, March 12, 2001。

其他的道琼斯股票平均价格指数还包括 20 种交通运输股票指数和 15 种公共事业股票指数。道琼斯股票平均价格指数每天都会被计算出来，并在《华尔街日报》及其他金融出版物上发布。

17.5 根据物价指数减缩序列

许多商业和经济的时间序列，比如公司销售额、工业销售额和库存额，都是以美元来计量的。这些时间序列通常随着时间推移会呈现出上升的趋势，这种上升通常被解释成与这些活动相关的实物增长。例如，库存额的美元总数上升了10%，这完全可以被解释为库存总量增加了10%。如果一个时间序列使用美元计量，这样的解释有可能引起误解，因为价值总量的变动同时包含着价格的变动和数量的变动。因此，在某一段时期，当价格的变动成为主要因素时，如果我们不对时间序列进行调整以消除价格变动的影响，则价值总量的变动将无法反映数量的变动。

对时间序列进行缩减是为了消除通货膨胀的影响。

举个例子，从1976年到1980年，建筑业的支出总额增长了大约75%。这个数字表明建筑业的增长态势良好。但是，建筑商品的价格以同样的速度在增长，甚至在某些时候增长率比75%还要高。实际上，当建筑总支出增长时，虽然不断有新项目开工，但是建筑业的发展却保持着相对的稳定，甚至在某种程度上在倒退。为了正确剖析1976年到1980年间的建筑业发展状况，我们必须用价格指数来调整总支出序列以消除价格上涨所带来的影响。对一个时间序列进行调整，以消除价格上涨所带来的影响，我们把这种调整叫做**时间序列的减缩**。

关于个人收入和工资，我们经常听到比如“真实工资”或工资的“购买力”之类的说法，这些概念都是以对每小时的工资收入进行减缩为基础的。例如，图17.2显示了在1996年到2000年期间制造业工人每小时工资的变化情况。我们可以看到每小时工资由12.77美元上涨到了14.36美元，那么制造业工人是否会为这种工资的增长而感到欣慰呢？答案取决于他们的工资购买力是否增长。如果我们可以将1996年每小时工资12.77美元的购买力与2000年每小时工资14.36美元的购买力进行比较，我们就可以更清楚地判断工资的相对增长情况。

表17.8给出了1996年—2000年间的每小时工资率和消费者价格指数。利用这些数据，我们可以说明怎样利用消费者价格指数来对每小时工资指数进行减缩。减缩过的时间序列由每年的每小时工资除以相对应的消费者价格指数再乘以100构成。减缩后的制造业工人每小时工资指数列于表17.9中；图17.3显示的是减缩后的工资，或者叫真实工资的图像。

对于1996年到2000年间的工人实际工资或购买力，减缩后的工资序列能反映什么问题呢？根据基期的价格(1982年到1984年为100)，每小时工资率在这段时期的曲线是平坦的。在消除了通货膨胀的影响后，我们可以看到工人的购买力仅有轻微的变化。在图17.3中可以清晰地看到这一点。因此，用物价指数对序列进行减缩的好处在于：我们可以清楚地了解到实际购买力所发生的变化。

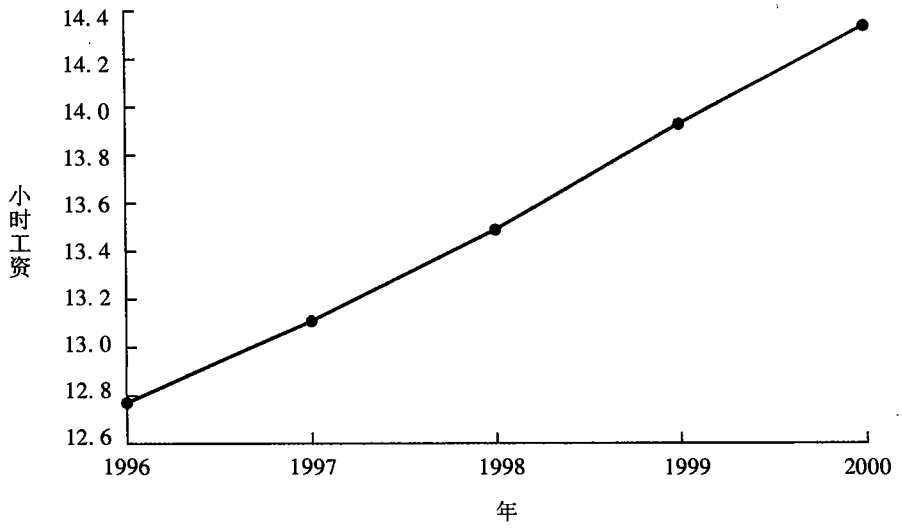


图 17.2 制造业工人的每小时实际工资

表 17.8 1996 年到 2000 年间制造业工人每小时工资和消费者价格指数

年份	每小时工资(美元)	消费者价格指数(以 1982—1984 年为基期)
1996	12.77	156.9
1997	13.11	160.5
1998	13.46	163.0
1999	13.93	166.6
2000	14.36	172.6

资料来源：Bureau of Labor Statistics。

表 17.9 经过减缩后的制造业工人每小时工资序列

年份	减缩后的每小时工资(美元)
1996	(12.77/156.9)(100) = 8.14
1997	(13.11/160.5)(100) = 8.17
1998	(13.46/163.0)(100) = 8.26
1999	(13.93/166.6)(100) = 8.36
2000	(14.36/172.6)(100) = 8.32

对时间序列进行减缩这种处理方法在计算国内生产总值(GDP)时有重要的应用。国内生产总值是某个国家所生产的产品和服务的总价值。很显然，如果不用物价指数来减缩国内生产总值，则国内生产总值将会随着物价的上涨而增长。因此，我们必须用物价指数来调整国内生产总值，以反映国内所生产和销售的产品和服务的总价值的实际变化。这种处理方法类似于在实际工资的计算中所用的方法。

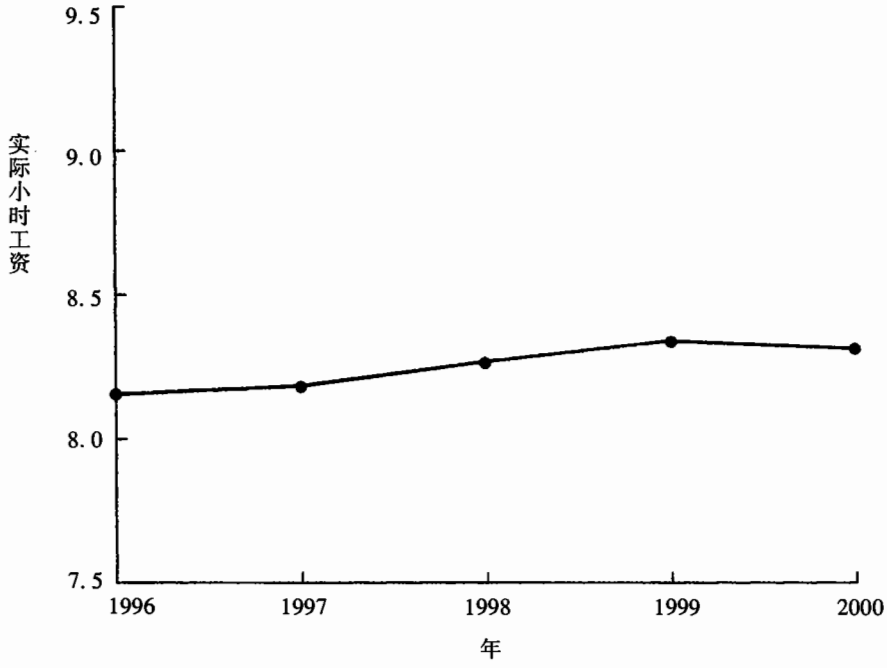


图 17.3　制造业工人每小时实际工资(1982—1984 = 100)

练　习

应　用

10. 制造业工人在 1980 年的每小时工资是 7.27 美元，在 2000 年，他们是 14.36 美元，1980 年的消费者价格指数是 82.4，2000 年的消费者价格指数是 172.6。

自测题

a. 对 1980 年和 2000 年的每小时工资进行减缩，计算实际工资。

b. 从 1980 年到 2000 年每小时实际工资的变化比例是多少？

c. 从 1980 年到 2000 年每小时真实工资的变化比例是多少？

11. 下表是从 1996 年到 2000 年 5 年期间，服务行业从业人员的每小时平均工资。利用表 17.8 中的消费者价格指数对工资序列进行减缩。从 1998 年到 2000 年，实际工资的增长幅度是多少？

年份	工资总额(10 亿美元)
1996	11.76
1997	12.23
1998	12.84

（续表）

年份	工资总额(10 亿美元)
1999	13.35
2000	13.82

资料来源：Bureau of Labor Statistics。

12. 美国人口普查局给出了 1997 年到 1999 年 3 年间制造业的总产值如下：

年份	制造业总产值(10 亿美元)
1997	3 929
1998	4 052
1999	4 260

a. 1997 年到 1999 年的消费者价格指数在表 17.8 中给出，利用这些信息对制造业总产值序列进行减缩，并评价制造业总产值变动情况。

b. 下面是 1997 年到 1999 年的生产者价格指数(不包括食品)，以 1982 年为基期。利用生产者价格指数对序列进行减缩。

年份	生产者价格指数(1982 年为 100)
1997	131.8
1998	130.7
1999	133.0

c. 利用消费者价格指数和生产者价格指数对制造业总产值进行减缩，您认为哪一个更为合适？

13. 下面是 Dooley 零售批发商店从 1982 年以来某些年份的销售额，还有以 1982 年—1984 年为基期的消费者价格指数。以 1982 年到 1984 年的美元不变价格来减缩销售额，并根据减缩后的资料评价商店的销售额。

年份	零售额(美元)	消费者价格指数(以 1982 年—1984 年为基期)
1982	380 000	96.5
1987	520 000	113.6
1992	700 000	140.3
1997	870 000	160.5
2000	940 000	172.6

17.6 物价指数：其他一些相关问题

在前面的章节中，我们描述了几种用于计算价格指数的方法，并讨论了一些重要指数的用途以及用价格指数对时间序列进行减缩的步骤。为了提高我们对物价指数的编制和用途的认识，我们有必要了解其他一些相关问题。接下来，本节将对它们进行讨论。

商品项目的选择

物价指数的首要目的在于测量某一类指定商品项目、产品等的价格随时间推移而发生的变动。由于同类商品的项目很多，所以我们不可能根据这一类型的所有商品项目来编制指数。但不管怎样，我们需要对这一类商品进行代表性的抽样。我们希望通过收集抽样项目的价格和数量信息，来获取指数所要描述的所有商品项目的价格情况。例如，对于消费者价格指数，一个消费者正常考虑所要购买的商品的数量超过 2 000 多种，但是消费者价格指数仅仅根据 400 项商品的价格资料来编制。对于编制指数所需项目的选择并不是一件简单的事情，可以通过对消费者购买商品进行调查来选择指定项目，但是不能通过简单随机抽样来选择这 400 种商品。

初步选定商品项以后，还必须随着采购意向的变化对指数中的项目进行定期修正。因此，在编制和修正指数之前，我们必须解决指数应该包含哪些商品项目这个问题。

基期的选择

大多数指数的编制都是以某个特定时期为基期，并且基期的数值设为 100。指数的所有将来值都是和基期数值相关的。但是对一个指数而言，什么样的基期才是合适的呢？这个问题并不简单，而且答案必须以指数编制者的判断为基础。

2000 年美国政府编制的许多指数都是以 1982 年为基期。作为一个总的指导思想，基期不应该离当前期太远。例如，如果消费者价格指数以 1945 年为基期，将会使公众难于理解，因为 1945 年的生活情形已经过于陌生。因此，大多数指数的基期都向着一个更近的时期定期调整。1988 年美国消费者价格指数基期从 1967 年调整为 1982—1984 年的平均值，当前的美国生产者价格指数以 1982 年作为基期(即 1982 年 = 100)。

质量变化

编制物价指数的目的在于测量价格随时间变化而发生的变化。理想的情况是，收集相同商品项目在不同时期的价格数据，然后再计算出价格指数。一个基本的假设是在每一个时期，相同商品项目的价格是可知的。当一个产品从一个时期进入下一个时期时，如果它的质量发生变化就有可能出现问题。举个例子，如果制造商使用廉价一点的原料，减少产品的特征等诸如此类的变动，一年一年累计，则有可能改变产品的质量。在接下来的年份里，价格反而会变得更高，但是与这个价格相对应的产品质量却下降了。因此，某商品价格的实际上升可能会比所显示的要高。即使我们能够针对商品质量的下降去调整指数，那也将会是一件相当困难的事。

如果对产品质量进行显著的改进，也会造成产品价格的上升。我们在进行指数计算时，应该排除产品质量提高所导致的价格上涨部分。但是，即使我们能够对指数进行调整以剔除质量提高所带来的价格影响，这样的调整也是一件很困难的事。

在编制指数时，虽然通常的做法是将微小的质量变化忽略不计，但是较大的质量变动必须被

考虑在内，因为随着时间的积累它会改变产品性质。如果产品性质发生了改变，则指数必须为此而修正；在某些情况下，该项产品甚至会被从指数计算中删掉。

在某些情况下，质量的显著改善反而伴随着价格的降低。这种例外的情形已经体现在20世纪90年代的个人电脑销售上。消费者价格指数的设计者们目前正在对消费者价格指数做适当的调整，以适应这种情形。

17.7 物量指数

除了前面章节中所描述的物价指数以外，其他类型的指数也是很有用的。特别要提到的是，指数的另外一个应用是测量数量随着时间变化所发生的变化。这种类型的指数被称为**物量指数**(quantity index)。

回顾17.2节中所讨论的加权综合物价指数的计算过程，为了计算 t 时期的指数，我们需要基期的单价(P_0)和 t 时期的单价(P_t)。公式(17.3)给出了加权综合物价指数的计算公式

$$I_t = \frac{\sum P_{it} Q_i}{\sum P_{i0} Q_i}(100)$$

分子 $\Sigma P_{it}Q_i$ 表示 t 时期指数项目固定数量的总值。分母 $\sum P_{i0}Q_i$ 表示基期指数项目同样固定数量的总值。

加权综合物量指数的计算类似于加权综合物价指数的计算。第 i 项基期和 t 期的数量分别记为 Q_{i0} 和 Q_{it}，然后，用同一固定价格、增加值或其他因素对数量进行加权。产品的“增加值”是销售额减去购入价格。计算 t 时期加权综合物量指数的公式为

$$I_t = \frac{\sum Q_{it} w_i}{\sum Q_{i0} w_i}(100) \tag{17.9}$$

在某些物量指数中，第 i 项的权数即为基期的价格(P_{i0})，在这种情况下，加权综合物量指数计算公式如下：

$$I_t = \frac{\sum Q_{it} P_{i0}}{\sum Q_{i0} P_{i0}}(100) \tag{17.10}$$

物量指数也可以根据加权物量比来计算，这种方法的计算公式如下：

$$I_t = \frac{\sum \frac{Q_{it}}{Q_{i0}}(Q_{i0}P_i)}{\sum Q_{i0}P_i}(100) \tag{17.11}$$

这个公式是17.3节中公式(17.8)所介绍的加权平均价格公式的数量表现形式。

由美国联邦储备委员会编制的**工业产品指数**(Index of Industrial Production)可能是最著名的物

量指数。它每月发布一次，以 1992 年作为基期。这个指数用于测量除了采矿业和公用事业之外的各种不同制造业的产品数量变动。2000 年 8 月的工业产品指数为 145.7。

练　习

方　法

14. 

以下是某三项产品在 1992 年和 2000 年的销量，以及 1992 年的产品单价。计算 2000 年的加权综合物量指数。

	销量		
产品	1992	2000	单价(1992 年)(美元)
A	350	300	18.00
B	220	400	4.90
C	730	850	15.00

应　用

15. 

一个卡车运输公司为某个指定销售商运输 4 种货物。下表是 1991 年和 2000 年的货物运输量和 1991 年的价格：

	货物运输量		1991 年运输价格
货物	1991	2000	(美元)
A	120	95	1 200
B	86	75	1 800
C	35	50	2 000
D	60	70	1 500

以 1991 年为基期，计算加权综合物量指数。评论 1991 年到 2000 年期间数量的增长或降低。

16. 下表是某汽车销售商 1989 年和 2000 年三种型号汽车的销量。请计算物量相对指数，并用它来计算 2000 年的加权综合物量指数。

	销量		销售的平均价格
型号	1989	2000	(1989 年)(美元)
Sedan	200	170	15 200
Sport	100	80	17 000
Wagon	75	60	16 800

小 结

在商业和经济环境中，价格指数和数量水平是衡量价格和数量变动的重要尺度。比价是某一商品的当前单价和基期单价的比值再乘以100。如果比价为100，则表示当前单价和基期单价没有差别。编制综合物价指数是为了对一组商品或产品的全部价格变动进行复合测量，一般情况下加权综合物价指数中的权数由其使用数量决定。根据商品的使用数量对指数中每一个商品的比价进行加权，也可以计算出加权综合物价指数。

消费者价格指数和生产者价格指数是两个应用非常广泛的指数，其中，消费者价格指数的计算是以1982—1984年为基期，生产者价格指数是以1982年为基期。道琼斯30种工业股票平均价格指数也是一个应用极为广泛的物价指数，它是纽约证券交易所上市的30种普通股票的加权和。和其他指数不一样的是，它不能表示为某一个基期数值的百分比。

物价指数通常可以用于对一些随时间变动的经济序列进行减缩，我们已经学习了如何用消费者价格指数减缩每小时的工资，从而得到实际工资指数。对指数计算中所涉及到的商品项目、基期进行选择以及因为商品品质的改变而进行调整，这些都是编制指数所应该考虑的重要问题。对物量指数只是进行了简短扼要的讨论。工业产品指数也是作为一个重要的物量指数而提及的。

术语释义

比价：商品的一种物价指数。计算方法是将当前期的单价除以基期的单价，然后再乘以100。

综合物价指数：根据一组商品的价格而得出的复合物价指数。

加权综合物价指数：对商品组中的商品，根据各自相对的重要性进行加权而得到的一种复合物价指数。

拉氏指数：以每一项商品的基期数量作为权数的加权综合物价指数。

帕氏指数：以每一项商品的当前期数量作为权数的加权综合物价指数。

消费者价格指数：一种每月定期发布的物价指数，根据由消费品和服务所组成的“购物篮”中价格的变动，来测量消费品价格随时间的变动。

生产者价格指数：一种每月定期发布的物价指数，它反映初级市场上销售商品(即在非零售市场上首次购买某种商品)的价格变动。

道琼斯股票平均价格指数：反映纽约证券交易所股票价格趋势和波动的综合物价指数。

物量指数：测量物量随时间变动的指数。

工业产品指数：测量实物数量或工业品产量随时间变动的一种物量指数。

重要公式

t 时期的比价

$$\frac{t\text{时期的价格}}{\text{基期价格}}(100) \tag{17.1}$$

t 时期的未加权综合物价指数

$$I_t = \frac{\sum P_{it}}{\sum P_{i0}}(100) \tag{17.2}$$

t 时期的加权综合物价指数

$$I_t = \frac{\sum P_{it}Q_i}{\sum P_{i0}Q_i}(100) \tag{17.3}$$

加权平均物价指数

$$I_t = \frac{\sum \frac{P_{it}}{P_{i0}}(100)w_i}{\sum w_i} \tag{17.6}$$

公式(17.6)的权数

$$w_i = P_{i0}Q_i \tag{17.7}$$

加权综合物量指数

$$I_t = \frac{\sum Q_{it}w_i}{\sum Q_{i0}w_i}(100) \tag{17.9}$$

补充练习

17. 1996—1999 年单亲家庭新住房的中等价格如下所示(*Statistical Abstract of the United States*, 2000):

年份	价格(千美元)
1996	140.0
1997	146.0
1998	152.5
1999	160.0

a. 以 1996 年作为基期，计算这 4 年来单亲家庭新住房的比价。

b. 以 1997 年作为基期，计算这 4 年来单亲家庭新住房的比价。

18. Nickerson 制造公司四种产品的装运数量和成本资料如下：

产品	基期数量(1998)	每单位平均装运成本(美元)	
		1998	2001
A	2 000	10.50	15.90
B	5 000	16.25	32.00
C	6 500	12.20	17.40
D	2 500	20.00	35.50

a. 计算每一种产品的比价。

b. 计算 4 年期间反映装运成本变动的加权综合物价指数。

19. 如果 2001 年 4 种产品的数量分别为：4 000，3 000，7 500，3 000。利用练习 18 中的价格数据计算装运成本的帕氏指数。

20. Boran 证券经纪人公司选择 4 种股票以编制指数反映股票市场行情。以 1999 年为基期，1999 年基期、2001 年 1 月和 2001 年 3 月每股价格如下，基期数量以 4 种股票的历史数量为基础。

股票	行业	1999 年数量	每股价格(美元)		
			1999 年基期	2001 年 1 月	2001 年 3 月
A	石油	100	31.50	32.75	32.50
B	计算机	150	65.00	59.00	57.50
C	钢铁	75	40.00	42.00	39.50
D	房地产	50	18.00	16.50	13.75

以 1999 年作为基期，计算 Boran 2001 年和 2001 年 3 月的指数。这个指数能说明股票市场什么行情？

21. 计算练习 20 中组成 Boran 指数的 4 种股票的比价。利用加权平均价格指数计算 2001 年 1 月、2001 年 3 月的 Boran 指数。

22. 以下是衣阿华地区谷物生产的数量和比价信息(*Statistical Abstract of the United States*，1997)：

产品	1991 年产量(百万蒲式耳)	基期价格(美元/蒲式耳)	1991—1996 年比价
玉米	1 427	2.30	113
黄豆	350	5.51	123

1996 年衣阿华地区谷物生产的加权综合物价指数是多少？

23. 1988 年和 1998 年新鲜水果价格以及数量如下(*Statistical Abstract of the United States*，1999)：

水果	1988 年人口平均消费量(磅)	1988 年价格(美元/磅)	1998 年价格(美元/磅)
香蕉	24.3	0.41	0.51
苹果	19.9	0.71	0.85
桔子	13.9	0.56	0.61
梨子	3.2	0.64	0.98

a. 计算每一种产品的比价。

b. 计算水果的加权综合物价指数，评论这 10 年期间水果价格的变动。

24. 以下是 Midwestern 大学企业管理助教的年薪起价(9 个月)。用消费者价格指数对年薪序列进行减缩，并据此讨论高等教育年薪的变化趋势。

年份	起薪(美元)	消费者价格指数(以 1982—1984 年为基期)
1970	14 000	38.8
1975	17 500	53.8
1980	23 000	82.4
1985	37 000	107.6
1990	53 000	130.7
1995	65 000	152.4
2000	80 000	172.6

25. 以下是某一股票 5 年来每股价格的历史数据，还有以 1982 年—1984 年为基期的消费者价格指数。

年份	每股价格(美元)	消费者价格指数(以 1982—1984 年为基期)
1996	51.00	156.9
1997	54.00	160.5
1998	58.00	163.0
1999	59.50	166.6
2000	59.00	172.6

对股票时间序列进行减缩，并评论该股票的投资价值。

26. 以下是某个大制造公司 1997 年—2001 年的产量和产品价格。根据这些数据计算加权综合物量指数，这个物量指数意味着什么？

	数量		
产品	1997	2001	价格(美元)
A	800	1 200	30.00
B	600	500	20.00
C	200	500	25.00

18 预 测

统计实例：内华达职业健康诊所

18.1 时间序列成分

趋势成分

循环成分

季节成分

不规则成分

18.2 平滑法

移动平均法

加权移动平均法

指数平滑法

18.3 趋势推测法

18.4 趋势和季节成分

乘法模型

计算季节指数

消除季节影响的时间序列

利用消除了季节影响的时间序列来分析趋势

季节调整

根据月度资料建立模型

循环成分

18.5 回归分析

18.6 定性预测方法

德尔菲法

专家判断法

远景规划法

直观预测法

统计实例

内华达职业健康诊所*

斯帕克斯，内华达州

内华达职业健康诊所(Nevada Occupational Health Clinic)是斯帕克斯的一个私人医疗诊所。它精于企业医疗，在同一个地方营业已经超过20年。在1991年初，该诊所开始进入快速增长期，在此后的26个月里，它的月收入从57 000美元一直增加到300 000美元。诊所业绩一直呈现快速的增长，而1993年4月6日它的主要建筑物却遭遇火灾而毁。

诊所的保险单包括固定资产和设备，还有因常规商业经营被打断而遭受的经济损失。计算固定资产和设备的损失，在受理保险索赔里面并不是一件很困难的事。但是，确定该诊所在7个月重建期间的经营损失则是一件相当复杂的事，它涉及到经营业主和保险公司之间的谈判。没有任何预先确立的原则可以用来计算诊所假如没有发生火灾将会有多少经营收入。为了估计损失额，诊所利用了预测方法来计算在7个月的停业期间可能实现的营业增长。正如本章所讨论的，在火灾之前的历史资料将会成为预测模型的基础，该模型考虑了线性趋势和季节成分的影响。这个预测模型使得诊所能精确估计所遭受的损失，并且最终被保险公司所接受。



1993年内华达职业健康诊所因火灾而被迫关闭7个月。© PhotoDisc, Inc.

*作者衷心感谢内华达职业健康诊所的经营总裁、执行管理助理Bard Betz提供这一统计实例。

管理任何组织的基本方法之一是制定将来的计划。事实上，一个组织长远的成功在很大程度上取决于管理部门是否能够预测未来并制定适宜的策略。准确的判断、直觉和对经济状况的了解将会使管理者对将来的态势有一个概略的认识和看法。但是，如何将这些看法和认识转量化，例如，将下一个销售季度的销量或者下一年的原材料价格进行预期，是困难的。本章的目的就是要介绍几种预测的方法以解决这一问题。

大多数的公司可以用小于5%的误差来预测所有产品的总需求。但是，对单个产品进行预测，则会有显著的误差。

假设我们需要提供即将到来的一年的某种特定产品的季度销量的预测值，生产进度安排、原材料采购、产品目录和销售定额都将受到这一预测值的影响。因此，较差的预测将导致计划的失败，并将使公司生产成本增加。我们应该怎样做才能提供销售量的季度预测值呢?

当然，我们需要观察该产品在过去时期的实际销售资料。利用这些历史资料，我们可以发现销量的一般水平和趋势，例如销量随时间而增加或减少。对这些资料进行更深入的观察，可以发现季节特征，例如销量峰值出现在每年第三季度，而在第一季度则是销量低谷。通过回顾历史数据，我们通常可以对销量有更好的了解，并对产品的未来销售有更好的预测。

销量的历史数据组成了一个时间序列。**时间序列**(time series)是一个变量在连续时点或持续时期上的观测值的集合。在本章中，我们将介绍几种分析时间序列的方法。分析时间序列的目的在于提供良好的**预测值**(forecasts)给未来的时间序列提供预测值。

预测仅仅是对将来可能发生情况的一种简单判断，管理人员必须接受这样一种事实：无论使用什么样的预测手段，他们都不可能得到完美的预测结果。

预测方法可以分为定性和定量两种方法。定性预测方法只有在以下几个条件都得到满足时才能应用：(1)被预测的变量的历史资料可用；(2)历史资料是可以定量的；(3)序列轨迹可以被顺推到将来。在这些条件都得到满足的情况下，我们可以使用时间序列分析方法或者因果分析方法来进行预测。

如果历史数据被可变的历史数值所制约，这种预测方法被称为**时间序列法**。时间序列法的研究目的是在历史数据中发现规律性的轨迹，并将它延伸到将来；这种预测方法只能以历史数据和/或历史预测误差为基础。在本章中，我们将讨论三种时间序列方法：平滑法(移动平均法、加权移动平均法和指数平滑法)、趋势推测法和调整季节影响后的趋势推测法。

因果预测法是建立在这样一个假定之上的：即我们所预测的变量与其他的一个或多个变量有因果关系，在本章中我们将讨论回归分析在因果预测法中的应用。例如，许多产品的销量受到广告支出的影响，因此可以建立回归分析方程来显示这两个变量是怎么相关的。然后，一旦确定了下一个时期的广告预算，即可将这个预算值代入方程得到下一个时期的销量预测或判断。但是请注意，如果我们使用时间序列预测法来预测，则不考虑广告支出，因为时间序列法仅仅依赖于历史销量。

定性预测法通常都涉及到专家判断，例如，一个专家组可以就今后的主要利率得到一个一致的预测结果。定性预测法的优点在于当所预测的变量无法定量的时候，或者无法利用或得到历史数据的时候，我们可以应用它。图 18.1 综合给出了各种类型的预测方法。

18.1　时间序列成分

时间序列数据的轨迹或行为包含几种成分。通常的假设认为有四种独立的成分：趋势、循环、季节和不规则，将这些成分综合起来可以得到时间序列的特定值。下面我们对每一种成分进行仔细深入的研究。

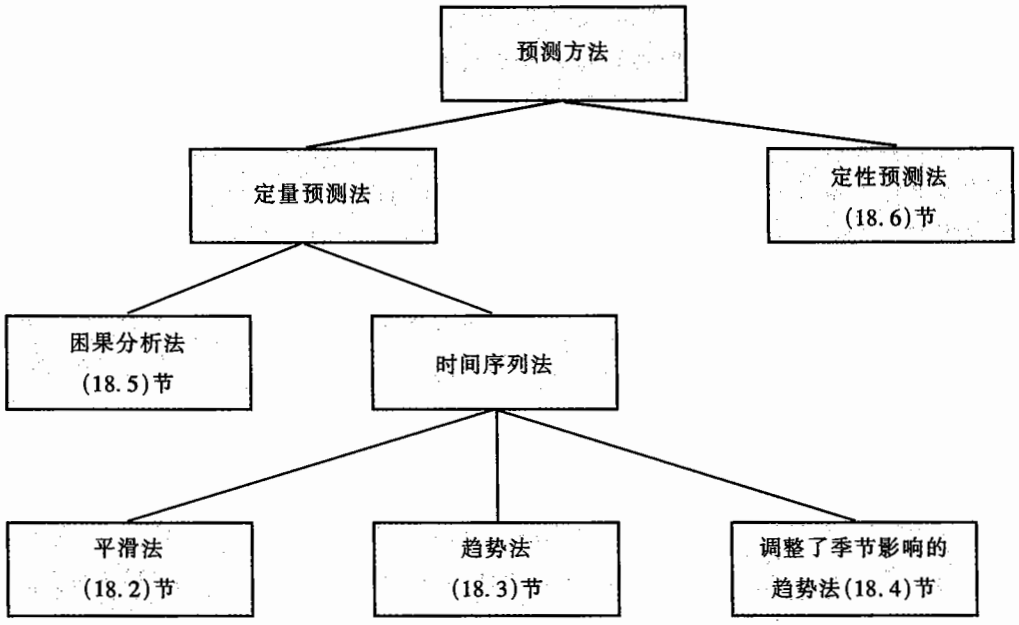


图 18.1 预测方法综合图

趋势成分

在时间序列分析中，测量间隔可以是每小时，每天，每周，每月，或者每年，还可以是某个特定的间隔周期。*虽然时间序列数据通常会呈现随机起伏的状态，但是它依然在一个较长的时期内逐渐呈现出相对升高或降低的运动趋势。这种时间序列的逐渐变化被称为时间序列的**趋势**(trend)；这种转变或趋势通常是长期因素影响的结果，例如人口数量的变化，人口构成特征的变化，技术的变化和/或消费者偏好的变化。

举个例子，一个摄影设备的制造商可以看到照相机销量每个月都在变化。但是，回顾过去 10 年到 15 年的销量，制造商会看到年销售量在逐渐增长。假设 1991 年销量接近 17 000 台，1996 年为 23 000 台，2001 年为 25 000 台。这种销量随着时间的逐渐增加表明了时间序列的增长趋势。图 18.2 所显示的直线可以被认为是照相机销量增长趋势的一个很好的近似。虽然照相机销量变化的趋势呈线性并随着时间的推移而增长，但是很多时候这种时间序列可以由其他图形得到更好的描述。

图 18.3 显示了另外一些可能的时间序列趋势图形。其中图 A 显示的是非线性趋势，在这种情况下，时间序列表明一开始增长缓慢，然后有一个迅速增长期，最后逐渐水平。这种趋势很好的模拟了一种产品从投入市场，经过销量增长，最后达到市场饱和的销售情形；图 B 可以应用于表示一个销量随着时间推移平稳减少的线性趋势；图 C 的水平线表明时间序列没有随着时间推移增加或减小，因此也就没有趋势。

* 我们对时间序列的讨论仅限于序列值是等间隔记录的情形，不等间隔记录观测值的情形超出了本书讨论的范畴。

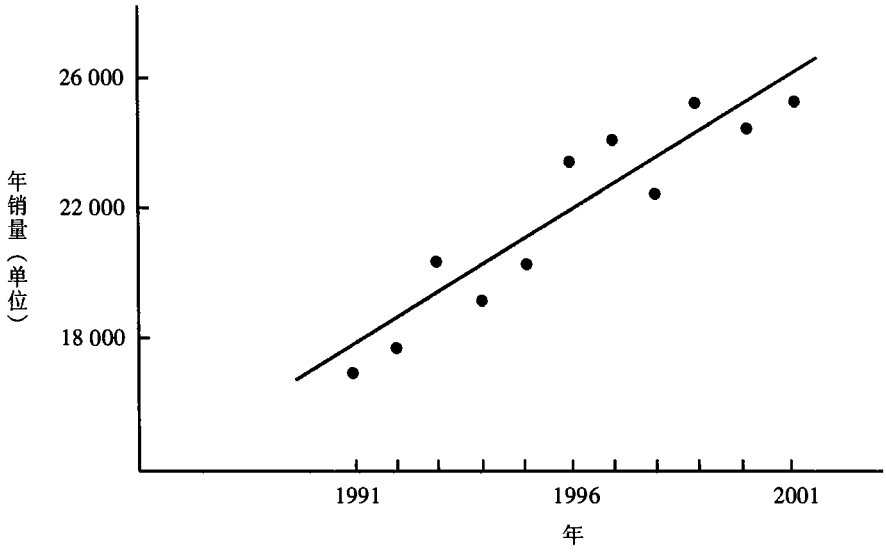


图 18.2　照相机销量的线性趋势

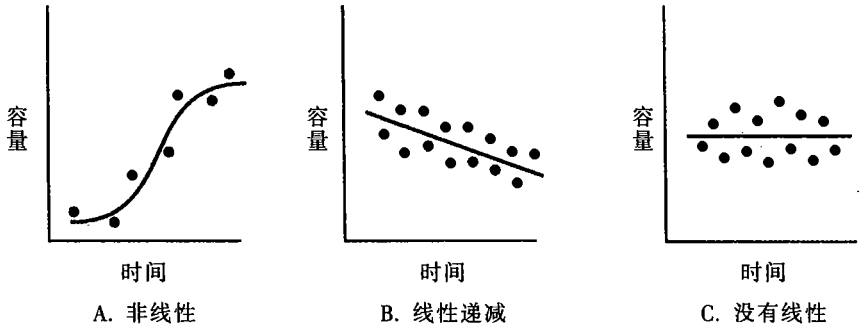


图 18.3　一些可能的时间序列趋势示例

循环成分

虽然时间序列可以显示长期变化趋势，但并不是所有的时间序列值都会准确地落在趋势线上。实际上，时间序列经常围绕着趋势线上下波动。任何围绕着趋势线上下波动时间超过一年的轨迹点都可以被称为时间序列的**循环成分**(cyclical component)。图 18.4 显示了一个有着明显循环成分的时间序列图形。观测值的读取间隔是一年。

许多时间序列的观测值环绕着趋势线有规则地上下波动，展示了明显的循环特性。通常，时间序列的循环成分归因于经济的多重循环运动。例如，轻度通货膨胀后紧接着高通货膨胀，这种经济现象会导致时间序列环绕着逐渐上升的趋势线(例如，关于房地产价格的时间序列)上下波动。20 世纪 80 年代早期的许多时间序列显示了这种类型。

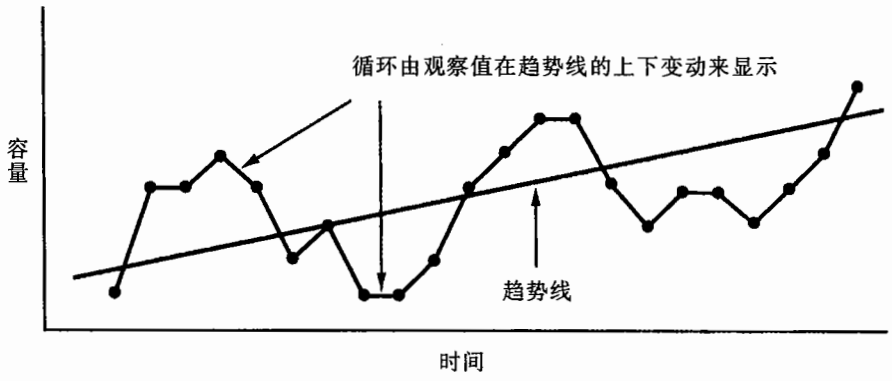


图 18.4 时间序列的趋势和循环成分(观测值的读取间隔为一年)

季节成分

虽然时间序列的趋势和循环成分是通过分析历史资料多年的运动来识别的，但是很多时候时间序列在一年的时间段内也会呈现出规则性的轨迹。例如，一个游泳工具的制造商可以预期它的产品销售在秋冬两季将会陷入低谷，但是在春季和夏季销售则会达到高峰。相反，铲雪设备和御寒衣服制造商对它们的产品销售则刚好有完全相反的预期。毫无疑问，时间序列中数据的变化应该归因于季节影响，因此我们把这种影响时间序列的成分称为**季节成分**(seasonal component)。尽管我们通常认为季节成分在时间序列中的运动是在一年之内出现，但是小于一年的规则性重复运动轨迹也可以用季节成分来表示。例如，每天的交通流量在上下班期间达到高峰，在其他的时间段内和傍晚流量比较适中，而在半夜到清晨则很微弱，这也是一种“季节”成分行为。

不规则成分

时间序列的**不规则成分**(irregular component)可以称为剩余或者包罗其他一切现象的解释因素，它被用来解释在剔除了趋势、循环和季节成分影响以后的时间序列实际变动现象。不规则成分是由那些短期的、无法预期的和不重复出现的因素影响时间序列而形成的。由于不规则成分所解释的是时间序列中的随机变动性，所以它是无法预测的。我们也不可能预测它对时间序列有什么样的影响。

18.2 平滑法

许多的生产环境需要每周或者每月预测成千上万的项目。因此，选择预测方法时简单性和易于使用性是一个很重要的标准。本章所论述的预测方法对数据量的要求都是最

少的，并易于使用和理解。

在本章中，我们讨论三种预测方法：移动平均法、加权移动平均法和指数平滑法。每一种方法的目的都是为了消除时间序列的不规则成分所造成的随机波动，因此它们都被称为平滑法。平滑法适于比较平稳的时间序列，即没有表示明显的趋势、循环或者季节影响，在这种情况下平滑法很适应时间序列的水平变化。但是，如果有明显的趋势、循环或者季节影响，则平滑法必须经过修正才能较好地反映时间序列变化。

平滑法易于使用，在通常情况下对短期预计，例如下一个时期的预计，精确度很高。平滑法里面的指数平滑法对数据的要求量最少，因此当我们需要对大量项目进行预测时。使用指数平滑法是一个很好的选择。

移动平均法

移动平均法(moving averages)使用时间序列中最近的 n 个数值的平均值作为下一时期时间序列的预测。移动平均法的数学计算公式如下：

移动平均值

$$\text{移动平均值} = \frac{\sum(\text{最近 } n \text{ 个数值})}{n} \tag{18.1}$$

“移动”这个词的含义在于：每次计算出时间序列的一个新观测值，就用它来代替公式(18.1)中最旧的观测值并计算出一个最新的平均值。因此，当新的观测值可以代入计算式中时，平均值将会变化或者移动。

为了阐述移动平均法，让我们考察表 18.1 中 12 个星期的数据和图 18.5。这些数据所记录的是位于佛蒙特州本宁顿的一个汽油销售商在过去 12 个星期中所销售的汽油加仑数。图 18.5 表明，虽然存在随机因素，但是时间序列随着时间的推移而发生的变化是平稳的。因此，可以应用本章所述的平滑法。

表 18.1　汽油销量时间序列

星期	销量(千加仑)
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

为了使用移动平均法来预测汽油销量，首先我们应该选择移动平均法里面所包含的数据个数。例如，我们使用 3 周的平均值来进行预测。汽油销量的时间序列中第一个 3 周的销量平均值计算如下：

$$\text{移动平均值(1—3 周)} = \frac{17 + 21 + 19}{3} = 19$$

然后我们利用这个移动平均值来预测第 4 周的汽油销量，由于第 4 周的实际销量是 23，所以第 4 周的预测误差是 23 − 19 = 4。通常说来，预测所产生的误差即为时间序列的实际观测值和预测值

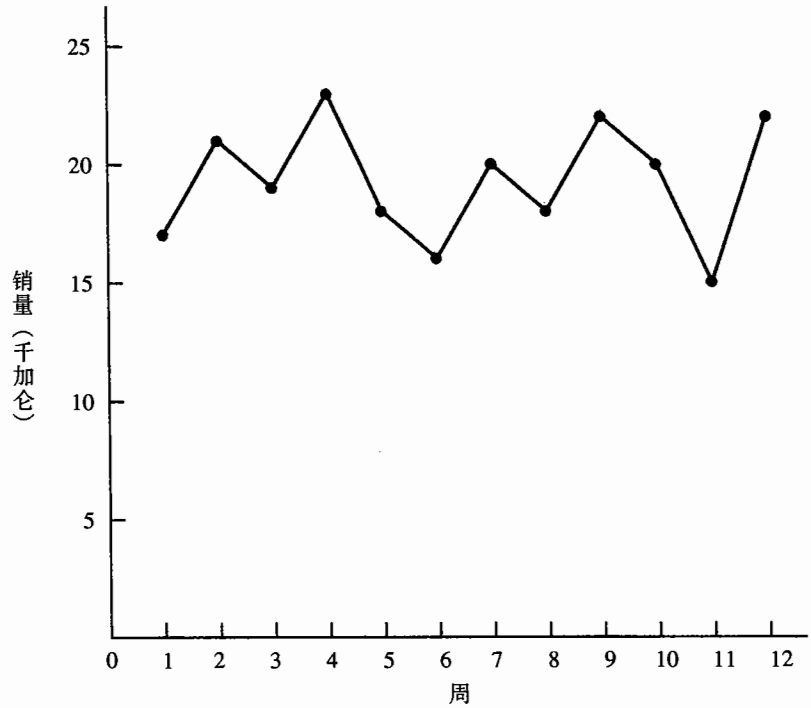


图 18.5 汽油销量时间序列

之间的差。

第 2 个 3 周移动平均值计算如下：

$$\text{移动平均值(2-4 周)} = \frac{21+19+23}{3} = 21$$

因此，第 5 周的预测值是 21。这次预测的误差是 18 - 21 = -3。这样的话，预测误差有可能为正也有可能为负，取决于预测值是偏高还是偏低。汽油销量时间序列的 3 周移动平均值计算结果总结于表 18.2 和图 18.6 中。

表 18.2 3 周移动平均值计算结果总结

周	时间序列值	移动平均预测值	预测误差	预测误差平方
1	17			
2	21			
3	19			
4	23	19	4	16
5	18	21	-3	9
6	16	20	-4	16
7	20	19	1	1
8	18	18	0	0
9	22	18	4	16
10	20	20	0	0
11	15	20	-5	25
12	22	19	3	9
			总计 0	92

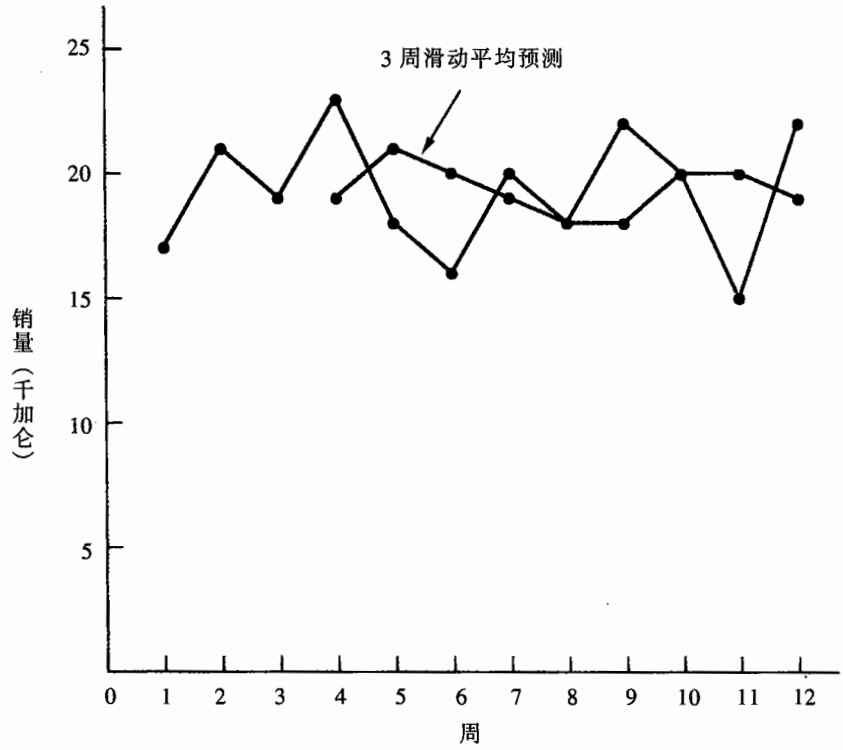


图 18.6　汽油销量的时间序列和 3 周移动平均预测值

预测精度并不是惟一的考虑因素。有时最精确的预测方法所需要的时间序列资料是很难获得的，或者要得到这些资料需付出很高的代价。因此，我们经常需要在费用和预测精度之间作出权衡取舍。

预测精度(forecast accuracy)。选择预测方法的一个重要考虑因素是预测精度。很明显，我们希望误差尽可能的小。表 18.2 中的最后两栏是预测误差和预测误差的平方，它们可以用于评估精度。

对于汽油销量时间曲线，我们可以使用表 18.2 中最后一栏来计算误差的平方和。计算可得

$$\text{误差平方和的平均值} = \frac{92}{9} = 10.22$$

误差平方和的平均值通常被称为**均方误差**(mean squared error，MSE)。MSE 均方误差经常用于度量预测方法的精度，也是我们本章所使用的。

正如我们前面所提到的，要使用移动平均法，我们首先必须选择移动平均值里面所应包含的数值个数。毫无疑问，对一个特定时间序列，不同步长的移动平均值对时间序列的预测精度是不同的。选择移动平均数中所包含的数值个数的一个可能的方法是通过实验和误差来确定，即所选定的步长应该使 MSE 的值最小。然后，如果我们假设历史数据的最佳步长同样适合于将来情况，

则我们可以用这个使历史事件序列的均方误差最小的最佳步长来预测时间序列的下一个数值。本章结尾的练习 2 将会要求计算 4 个星期或 5 个星期的汽油销量的移动平均值。通过比较均方误差，我们可以得到在计算移动平均数时最合适的周数。

加权移动平均法

在移动平均法里，移动平均值的计算中所用到的每一个观测值都是相同的权值。**加权移动平均值**(weighted moving averages)涉及到为每一个数据选定不同的权值，然后计算最近的 n 个数值的加权平均值作为预测值。在大多数情况下，离得最近的观测值所受到的权值最重，观测值离得越远则权值越小。例如，我们可以利用汽油销量的时间序列来阐明 3 周的加权平均值计算。离得最近的数据的权重是离得最远的数据权重的 3 倍，位于中间的那个观测值所受到的权重是离得最远的数据的 2 倍，第 4 周的数值计算如下：第 4 周数值 = 1/6(17) + 2/6(21) + 3/6(19) = 19.33。请注意加权移动平均法的权值总和等于 1。实际上移动平均法的权值总和也等于 1，而每一个观测值的权值都是 1/3。但是，简单或者未加权的移动平均法所得到的预测值是 19。

预测精度(forecast accuracy)。要使用加权移动平均法，我们首先必须选择计算加权移动平均值的数值个数，然后对每一个数值给出权重。总的说来，如果我们相信离得比较近的数据比离得远的数据能够更好地预测将来，则我们应该给离得比较近的数据赋予较大的权重。但是，当时间序列变动较大时，为每一个数据选择近似相等的权值也许是最佳选择。如果要在特定数量的数据和权值的组合之间比较所给出的预测值的精度，我们同样可以使用 MSE 值作为测量预测精度的标准。即，如果我们假设这个数据和权重的组合能够最准确地反映历史情况，也能够最准确地预测将来，则我们也应该能够用这个数据和权重的组合使得时间序列的 MSE 值最小。

指数平滑法

指数平滑法比较简单，数据量需求也较少，因此对每一时期均要进行很多预测的公司来说，它是一种比较经济的预测方法。

指数平滑法(exponential smoothing)将历史事件序列的加权平均值作为对将来的预测；它是加权移动平均法的一个特例，在这里我们只选择一个权值，即离得最近的观测值的权值。其他数据的权值可以自动计算并且随着时间的越往前推移将会变得越小。指数平滑法的基本模型如下：

指数平滑模型

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t \tag{18.2}$$

式中 F_{t+1}——$t+1$ 期的时间序列预测值；

Y_t ——t 时期的时间序列实际值；

F_t ——t 时期的时间序列预测值；

α ——平滑常数($0 \leqslant \alpha \leqslant 1$)。

公式(18.2)表明 $t+1$ 时期的预测值是 t 时期的实际值和 t 时期的预测值的加权平均值；请特别注意 t 时期实际值的权值是 α，t 时期预测值的权值是 $1-\alpha$。我们可以通过一个包含三个时期数据：Y_1，Y_2 和 Y_3 的时间序列，来证明任意时期的指数平滑法预测值也是时间序列的**所有历史实际数据**的加权平均值。为开始计算，我们可以令 F_1 等于时间序列第 1 时期的实际值；即 $F_1 = Y_1$。因此，第 2 时期的预测值为

$$\begin{aligned} F_2 &= \alpha Y_1 + (1-\alpha) F_1 \\ &= \alpha Y_1 + (1-\alpha) Y_1 \\ &= Y_1 \end{aligned}$$

所以第 2 期的指数平滑预测值等于时间序列第 1 期的实际值。

第 3 期的预测值为

$$F_3 = \alpha Y_2 + (1-\alpha) F_2 = \alpha Y_2 + (1-\alpha) Y_1$$

最后，将 F_3 的表达式代入 F_4 的表达式，我们可以得到

$$\begin{aligned} F_4 &= \alpha Y_3 + (1-\alpha) F_3 \\ &= \alpha Y_3 + (1-\alpha)[\alpha Y_2 + (1-\alpha) Y_1] \\ &= \alpha Y_3 + \alpha(1-\alpha) Y_2 + (1-\alpha)^2 Y_1 \end{aligned}$$

“指数平滑”这个术语来自于对历史数据进行加权的指数性质。

因此，F_4 是前三个时间序列值的加权平均值。Y_1，Y_2，Y_3 的系数，或者权值的总和等于 1。通常我们可以得到一个相似的结论：任何预测值 F_{t+1} 都是历史时间序列值的加权平均值。

尽管指数平滑法所提供的预测值是所有历史观测值的加权平均值，但是我们并不需要将所有历史数据保存到计算机上以进行下一时期的预测。事实上，一旦选择了**平滑常数**(smoothing constant) α，我们仅仅需要两项信息即可计算预测值。公式(18.2)表明：如果给定 α，我们只需要知道 t 时期的时间序列实际值和预测值，即 Y_t 和 F_t，就可以计算 $t+1$ 期的预测值。

为了说明指数平滑法的预测，我们可以考察表 18.1 和图 18.5 中的汽油销量时间序列。由于第 2 时期的指数平滑法预测值等于第 1 时期的时间序列实际值，而且有 $Y_1 = 17$，所以，我们在开始指数平滑计算时可令 $F_2 = 17$。参考表 18.1 中的时间序列，我们发现第 2 时期的时间序列实际值 $Y_2 = 21$。因此，第 2 时期的预测误差为 $21 - 17 = 4$。

继续利用指数平滑常数 $\alpha = 0.2$ 来进行指数平滑计算。我们可以得到第 3 时期的预测如下：

$$F_3 = 0.2 Y_2 + 0.8 F_2 = 0.2(21) + 0.8(17) = 17.8$$

一旦得知第 3 期的时间序列实际值 $Y_3 = 19$，我们可以得到第 4 期的预测如下：

$$F_4 = 0.2 Y_3 + 0.8 F_3 = 0.2(19) + 0.8(17.8) = 18.04$$

通过继续指数平滑计算，我们可以确定每周的预测值和相应的预测误差，如表 18.3 所示。请注意我们没有给出第 1 期的指数平滑预测和预测误差，因为对它进行预测是不可能的。对第12

周，我们有 $Y_{12}=22$，$F_{12}=18.48$。在知道第 13 周的实际值之前，我们是否能够利用这些信息来预测第 13 周呢？利用指数平滑模型，我们可以得到

$$F_{13}=0.2Y_{12}+0.8F_{12}=0.2(22)+0.8(18.48)=19.18$$

因此，第 13 周销量的指数平滑预测值为 19.18，即 19 180 加仑汽油。利用该预测，公司可以相应

表 18.3　汽油销量的指数平滑预测值和预测误差总结(平滑常数 $\alpha=0.2$)

周 (t)	时间序列值 (Y_t)	指数平滑预测值 (F_t)	预测误差 (Y_t-F_t)
1	17		
2	21	17.00	4.00
3	19	17.80	1.20
4	23	18.04	4.96
5	18	19.03	-1.03
6	16	18.83	-2.83
7	20	18.26	1.74
8	18	18.61	-0.61
9	22	18.49	3.51
10	20	19.19	0.81
11	15	19.35	-4.35
12	22	18.48	3.52

地制定计划和决策。预测的精度只有在第 13 周周末才能获知。

图 18.7 是时间序列的实际值和预测值的散点图。请特别注意预测值是如何将时间序列中的不规则波动平滑消除的。

预测精度(forecast accuracy)。在前面的指数平滑计算中，我们使用指数平滑常数 $\alpha=0.2$。虽然 α 在 0 和 1 之间的任意值都是可以接受的，但是有些 α 值所产生的预测值比别的 α 值更为精确。为了观察如何选择一个更为合适的 α 值，我们可以将基本指数平滑模型改写为如下形式：

$$\begin{aligned}F_{t+1}&=\alpha Y_t+(1-\alpha)F_t\\F_{t+1}&=\alpha Y_t+F_t-\alpha F_t\\F_{t+1}&=F_t+\alpha(Y_t-F_t)\end{aligned}\qquad(18.3)$$

F_t：t 时期的预测值；$\alpha(Y_t-F_t)$：t 时期的预测误差

因此，新的预测值 F_{t+1} 等于历史预测值 F_t 加上一个调整值，调整值等于 α 乘上最近的预测误差 Y_t-F_t。即通过调整 t 时期的预测值和一部分预测误差我们可以得到 $t+1$ 期的预测值。如果时间序列包含大量的随机变异，我们倾向于使用较小的平滑常数。这样选择的理由在于：许多预测误差归因于随机变异，我们不希望对预测作出过度的反映和过快的调整。对于随机变异较小的时间序列，可以选取较大的指数平滑常数，这样做的优点是在出现预测误差时可以迅速改变条件调整误差。

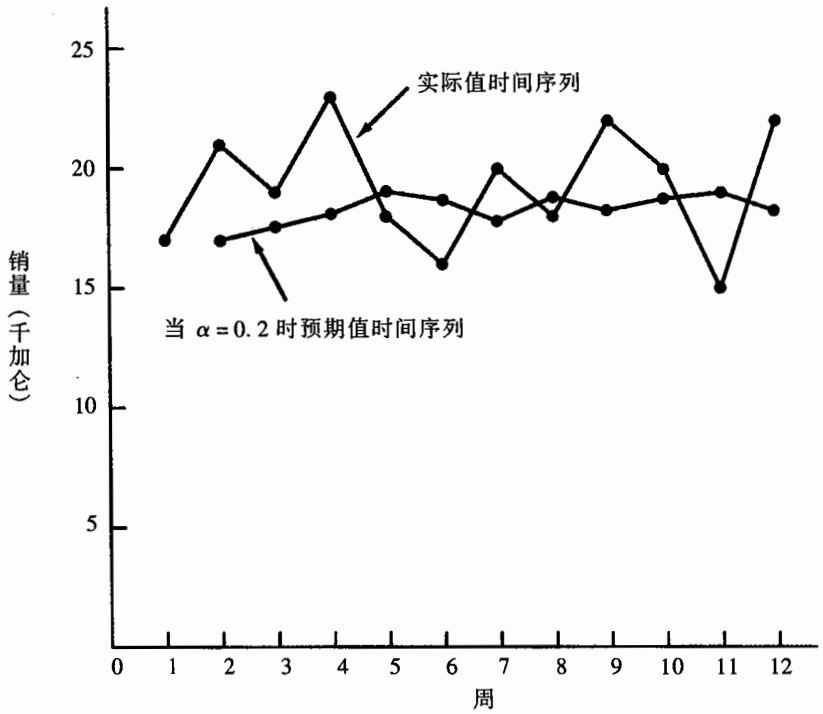


图 18.7　汽油销量时间序列的实际值和预测值的散点图(平滑常数 $\alpha=0.2$)

确定指数平滑法预测精度的标准依然是选择 α 值，这和我们在计算移动平均数中所提出的确定包含时期个数的准则相同，即所选择的 α 值应使均方误差(MSE)值达到最小。表 18.4 是汽油销量时间序列用 $\alpha=0.2$ 进行指数平滑预测的均方误差计算综述。请注意，误差平方项比时期项

表 18.4　汽油销量的均方误差(MSE)计算综述($\alpha=0.2$)

周 (t)	时间序列值 (Y_t)	预测 (F_t)	预测误差 (Y_t-F_t)	预测误差平方 $(Y_t-F_t)^2$
1	17			
2	21	17.00	4.00	16.00
3	19	17.80	1.20	1.44
4	23	18.04	4.96	24.60
5	18	19.03	-1.03	1.06
6	16	18.83	-2.83	8.01
7	20	18.26	1.74	3.03
8	18	18.61	-0.61	0.37
9	22	18.49	3.51	12.32
10	20	19.19	0.81	0.66
11	15	19.35	-4.35	18.92
12	22	18.48	3.52	12.39
			总计	98.80

$$\text{MSE}=\frac{98.80}{11}=8.98$$

数少一个，因为我们没有可以用来对第 1 时期进行预测的历史资料。不同的 α 值能否提供一个较小的，而且也更为合适的均方误差值呢？也许回答这个问题的最直接的办法是用另一个 α 值来进行测试，然后再将所得的均方误差和用平滑常数 $\alpha=0.2$ 所得的均方误差 8.98 来进行比较。

$\alpha=0.3$ 的指数平滑预测结果如表 18.5 所示，因为 MSE = 9.35，我们可以看到使用当前数据资料，指数平滑常数 $\alpha=0.3$ 的结果比指数平滑常数 $\alpha=0.2$ 的结果所得的精确度要差。因此，我们倾向于使用原来的平滑常数 $\alpha=0.2$。利用这种实验方法计算其他的 α 值，我们可以发现最为合适的指数平滑值，这个值可以应用于指数平滑模型中以预测将来。随着时间的往后推移，我们会得到新的时间序列观测值，然后我们可以分析新近收集的时间序列数据，确定是否需要修改指数平滑常数以得到更好的预测结果。

表 18.5　汽油销量的均方误差(MSE)计算综述($\alpha=0.3$)

周 (t)	时间序列值 (Y_t)	预测 (F_t)	预测误差 (Y_t-F_t)	预测误差平方 $(Y_t-F_t)^2$
1	17			
2	21	17.00	4.00	16.00
3	19	18.20	0.80	0.64
4	23	18.44	4.56	20.79
5	18	19.81	-1.81	3.28
6	16	19.27	-3.27	10.69
7	20	18.29	1.71	2.92
8	18	18.80	-0.80	0.64
9	22	18.56	3.44	11.83
10	20	19.59	0.41	0.17
11	15	19.71	-4.71	22.18
12	22	18.30	3.70	13.69
			总计	102.83

$$\text{MSE}=\frac{102.83}{11}=9.35$$

评　注

1. 衡量预测精度的另外一种尺度是**平均绝对偏差**(MAD)。它很简单，就是所有预测误差的绝对值的平均数。根据表 18.2 中的误差，我们可以得到：

$$\text{MAD}=\frac{4+3+4+1+0+4+0+5+3}{9}=2.67$$

均方误差和平均绝对偏差的一个主要区别是：均方误差受较大的误差影响比受较小误差的影响要大的多(因为均方误差所计算的是误差的平方)。选择衡量预测精度的方法并不是一件简单的事。实际上，即使是预测专家也经常为选择哪一种衡量方法产生分

歧，我们在本章使用 MSE 作为测量方法。

2. 棋盘式对照表是一种辅助工具，可以用于指数平滑法选择合适的 α 值，也可以用于加权移动平均法选择权值。在棋盘式对照表中，根据时间序列资料和预测公式，我们可以试验不同的 α 值(或者移动平均法的不同权值)，并对它们进行选择以使 MSE 或 MAD 的值最小。

练 习

方 法

1. 考察下列数据：



周	1	2	3	4	5	6
值	8	13	15	17	16	9

a. 求这个时间序列的 3 周移动平均值，第 7 周的预测值是多少？

b. 计算 3 周移动平均值的均方误差。

c. 用 $\alpha=0.2$ 计算这个时间序列的指数平滑值，第 7 周的预测值是多少？

d. 将 3 周的移动平均预测值和 $\alpha=0.2$ 的指数平滑预测值进行比较，哪一种方法所提供的预测较好？

e. 用指数平滑常数 $\alpha=0.4$ 来计算指数平滑预测值，$\alpha=0.2$ 和 $\alpha=0.4$ 所提供的指数平滑预测值哪一个更好？并解释原因。

2. 参考表 18.1 中的汽油销量时间序列数据。

a. 计算这个数列的 4 周和 5 周的移动平均值。

b. 计算这个数列的 4 周和 5 周的移动平均值的 MSE。

c. 在移动平均值计算中应该选择历史资料中几周的数据来计算最为合适？3 周的移动平均值的 MSE 值是 10.22。

3. 再参考表 18.1 中的汽油销量时间序列数据。

a. 令离得最近的数据的权值为 1/2，其次的数据的权值为 1/3，最远时期的数据权值为 1/6，计算这个时间序列的 3 周加权移动平均预测值。

b. 计算(a)中加权移动平均预测值的均方误差 MSE。在加权移动平均预测值和移动平均预测值中，你更倾向于使用哪一种预测方法？

c. 假设允许选择任意权值组合，只要满足所有权值总和为 1 的条件。你是否总是能够找到这样一组权值组合，它可以使加权移动平均值的 MSE 值比移动平均值的 MSE 值要小？解释原因。

4. 根据表 18.1 中的汽油销量时间序列数据计算指数平滑预测值，其中 $\alpha=0.1$。利用 MSE 值作为标准，对汽油销量时间序列，指数平滑常数 $\alpha=0.1$ 和 $\alpha=0.2$ 你倾向于使用哪一个？

5. 对表 18.1 中的汽油销量资料，平滑常数 $\alpha=0.2$，公式(18.2)表明第 13 周汽油销量的预测值为 $F_{13}=0.2Y_{12}+0.8F_{12}$，第 12 周的预测值为 $F_{12}=0.2Y_{11}+0.8F_{11}$。将这两个结果结合起来，第 13 周的预测值可以写成：

$$F_{13}=0.2Y_{12}+0.8\times(0.2Y_{11}+0.8F_{11})=0.2Y_{12}+0.16Y_{11}+0.64F_{11}$$

a. 根据 $F_{11}=0.2Y_{10}=0.8F_{10}$($F_9$ 和 F_{10} 类似)，继续扩展 F_{13} 的表达式，使它表示为历史数据 Y_{12}，Y_{11}，Y_{10}，Y_9，Y_8 和第 8 期预测值的组合。

b. 参考历史数据 Y_{12}，Y_{11}，Y_{10}，Y_9，Y_8 的系数或权值，对于指数平滑法在得出新的预测值时，是如何对历史数据进行权值分配的，你有什么样的观测结果？请将这个加权方式与移动平均法的加权方式进行比较。

应　用

6. 在过去 12 个月中，Hawkins 公司每个月及时收到的货物百分比为 80，82，84，83，83，84，85，84，82，83，84，83。

a. 利用指数平滑法预测 3 个月的移动平均值，其中 $\alpha=0.2$，并将结果和 3 个月的移动平均预测值作比较，哪一种预测方法更好？

b. 下一个月的预测值是多少？

7. 以下是某 3A 级公司的债券在连续 12 个月中的利息率：

9.5　9.3　9.4　9.6　9.8　9.7　9.8　10.5　9.9　9.7　9.6　9.6

a. 计算这个时间序列的 3 个月和 4 个月的移动平均值。在 3 个月和 4 个月的移动平均值中，哪一个提供了更好的预测结果？解释原因。

b. 下一个月的移动平均预测值是多少？

8. 在 12 个月中，阿拉巴马州的建筑合同值如下(百万美元)：



240　350　230　260　280　320　220　310　240　310　240　230

a. 利用指数平滑法预测 3 个月的移动平均值，其中 $\alpha=0.2$，并将结果和 3 个月的移动平均预测值作比较，哪一种预测方法更好？

b. 下一个月的移动平均预测值是多少？

9. 下面的时间序列是某种产品在过去 12 个月中的销量：

月	销量	月	销量
1	105	7	145
2	135	8	140

（续表）

月	销量	月	销量
3	120	9	100
4	105	10	80
5	90	11	100
6	120	12	110

a. 令 $\alpha=0.3$ 计算该时间序列的指数平滑预测值。

b. 用平滑常数 0.5 来计算指数平滑预测值，当平滑常数分别为 0.3 和 0.5 时，哪一个所提供的预测值更好？

10. 10 个星期的商品期货指数为：7.35，7.40，7.55，7.56，7.60，7.52，7.52，7.70，7.62 和 7.55。

a. 用 $\alpha=0.2$ 来计算指数平滑预测值。

b. 用 $\alpha=0.3$ 来计算指数平滑预测值。

c. 哪一个常数提供的平滑预测值更好？并预测第 11 周的期货指数。

11. 下面的数据所表示的是 15 个季度的生产能力利用情况(百分比)：

季度/年	利用情况(%)	季度/年	利用情况(%)
1/1998	82.5	1/2000	78.8
2/1998	81.3	2/2000	78.7
3/1998	81.3	3/2000	78.4
4/1998	79.0	4/2000	80.0
1/1999	76.6	1/2001	80.7
2/1999	78.0	2/2001	80.7
3/1999	78.4	3/2001	80.8
4/1999	78.0		

a. 计算这个时间序列的 3 个季度和 4 个季度的移动平均预测值。对 2001 年的第 4 季度，哪一个预测所提供的预测值更合适？

b. 用平滑常数 $\alpha=0.4$ 和 $\alpha=0.5$ 来计算 2001 年第 4 季度的预测值，哪一个平滑常数所提供的预测值更为精确？

c. 在(a)和(b)的分析基础上，哪一种方法：移动平均法或者指数平滑法，所提供的预测更为精确？解释原因。

18.3 趋势推测法

在本章中，我们说明了如何对拥有长期线性趋势的时间序列进行预测。这种类型的时间序列呈现连续的增长或降低趋势，因此可以应用趋势推测法。由于这种时间序列并不稳定，所以无法

应用前面章节中所讲述的平滑法。

表 18.6 自行车销量时间序列

年 (t)	销量(千辆) (Y_t)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4

考虑过去 10 年自行车销量的时间序列，表 18.6 和图 18.8 显示的是某个自行车制造商的自行车销量时间序列。第一年销售的自行车数量为 21 600 辆，第 2 年销售的自行车数量为 22 900 辆，等等。在第 10 年，销售了 31 400 辆自行车。虽然图 18.8 显示了在过去 10 年中有上下波动的运动轨迹，但是从总体看来时间序列还是呈增长或上升的趋势。

我们并不打算让时间序列的趋势成分去追溯图像的每一次上升或下降运动。但是，趋势成分应该反映时间序列的逐渐变动，在这个例子中，即逐渐的上升趋势。经过对表 18.6 中的时间序列数据和图 18.8 中的图像进行分析，我们认为图 18.9 所示的线性趋势对时间序列的长期运动描述是比较合理的。

我们利用自行车的销售数据，来说明应用回归分析判断线性趋势的计算。回顾第 14 章中关于简单线性回归分析的讨论，我们介绍了怎样利用最小二乘法来发现两个变量之间的最佳直线关系，同

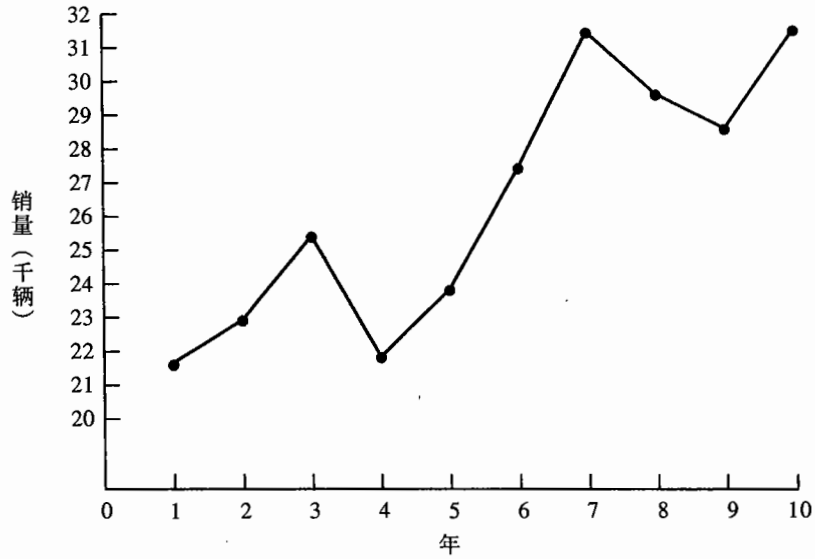


图 18.8 自行车销量时间序列图

样，最小二乘法也是我们用来建立关于自行车销量时间序列的趋势线的工具。特别地，我们可以利用回归分析来估计时间和自行车销量之间的关系。

在第 14 章中，描述自变量 x 和因变量 y 之间的直线关系的估计回归方程如下所示：

$$\hat{y} = b_0 + b_1 x \tag{18.4}$$

需要强调的是，在预测中，自变量是时间，我们将在公式(18.4)中用 t 来代替 x；除此之外，我们将使用 T_t 来代替 $\hat{y}$。因此，对一个线性趋势，销量的估计量作为时间的函数可写成如下形式：

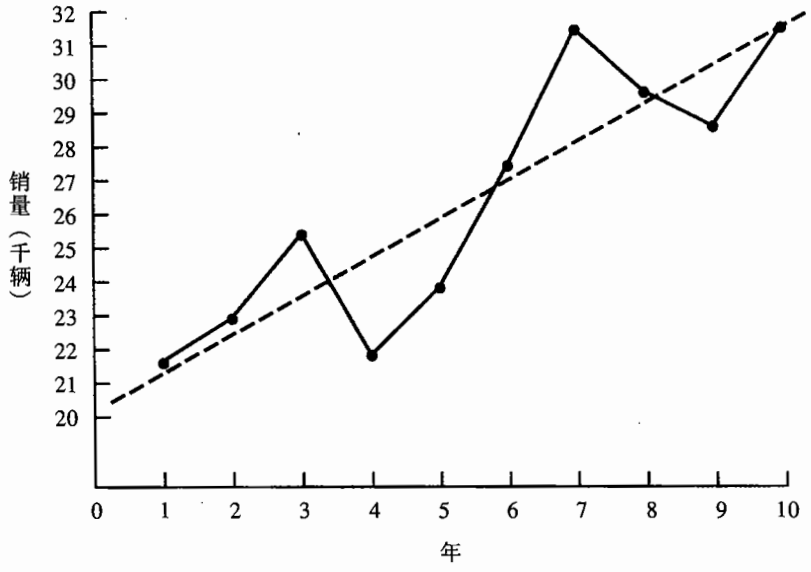


图 18.9　用线性函数描述自行车销量趋势

线性趋势方程

$$T_t = b_0 + b_1 t \tag{18.5}$$

式中　T_t——t 时期的时间序列趋势值；

b_0——趋势线的截距；

b_1——趋势线的斜率；

t ——时间。

在公式(18.5)中，我们令 $t=1$ 表示时间序列数据中第一个观测值所对应的时间，$t=2$ 表示第二个观测值所对应的时间，依此类推。请注意，对自行车销量的时间序列，$t=1$ 对应着最早的时间序列值，$t=10$ 对应着最近的数据。(18.5)中的估计回归系数(b_1 和 b_0)的计算公式如下所示：

计算斜率(b_1)和截距(b_0)

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n} \tag{18.6}$$

$$b_0 = \overline{Y} - b_1 \overline{t} \tag{18.7}$$

式中　Y_t——t 时期的时间序列值；

n——时期数；

$\overline{Y}$——时间序列的平均值；$\overline{Y} = \Sigma Y_t / n$；

$\overline{t}$——t 的平均值；$\overline{t} = \Sigma t / n$。

利用公式(18.6)和(18.7)以及表 18.6 中的自行车销量数据，我们可以计算 b_0 和 b_1 的值，如下所示：

	t	Y_t	tY_t	t^2
	1	21.6	21.6	1
	2	22.9	45.8	4
	3	25.5	76.5	9
	4	21.9	87.6	16
	5	23.9	119.5	25
	6	27.5	165.0	36
	7	31.5	220.5	49
	8	29.7	237.6	64
	9	28.6	257.4	81
	10	31.4	314.0	100
总计	55	264.5	1 545.5	385

$$\bar{t} = \frac{55}{10} = 5.5$$

$$\bar{Y} = \frac{264.5}{10} = 26.45$$

$$b_1 = \frac{1\,545.5 - (55)(264.5)/10}{385 - (55)^2/10} = 1.10$$

$$b_0 = 26.45 - 1.10(5.5) = 20.4$$

因此

$$T_t = 20.4 + 1.1t \qquad \textbf{(18.8)}$$

在用趋势方程进行预测之前，应该进行统计显著性检验(见第 14 章)。在实践中，这样的检验将成为拟合趋势线的程序之一。

公式(18.8)即为自行车销量时间序列的线性趋势成分的表达式。

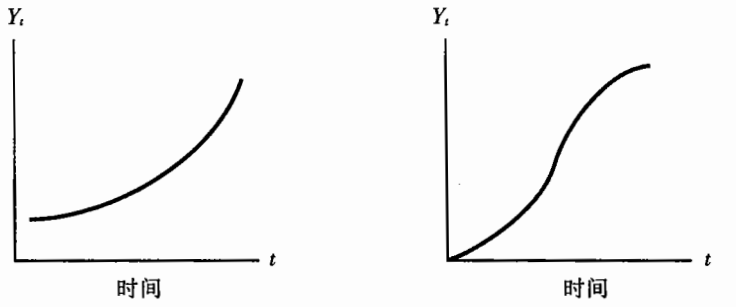
斜率为 1.1 表明在过去的 10 年中，公司每年的销量增长大约为 1 100 个单位。如果我们假设过去 10 年的销量趋势对未来也是适用的，则公式(18.8)可以用来预测时间序列的趋势成分。例如，将 $t=11$ 代入公式(18.8)中可以得到下一年的预测值 T_{11}。即

$$T_{11} = 20.4 + 1.1(11) = 32.5$$

因此，仅仅通过使用趋势成分，我们就可以预测出下一年的自行车销量为 32 500。

用线性函数来拟合趋势是比较常见的情形，但是，正如我们在前面所讨论过的，有时候时间序列会呈曲线或非线性，就像图 18.10 中所示。在第 16 章中，我们讨论了怎样用回归分析来拟合

图 18.10 A 中所示的曲线关系。关于图 18.10 B 中所示的如何建立更加复杂的关系回归模型，我们将在更高级的教材中详细讨论。



A. 指数趋势　　B. 冈伯茨(Gompertz)增长曲线

图 18.10　非线性趋势曲线形态的几种可能形式

练　习

方　法

12. 考察以下的时间序列：



t	1	2	3	4	5
Y_t	6	11	9	14	15

为该时间序列的线性趋势成分建立方程，$t=6$ 时的预测值是多少？

13. 考察以下的时间序列：

t	1	2	3	4	5	6
Y_t	205	202	195	190	191	188

为该时间序列的线性趋势成分建立方程，$t=7$ 时的预测值是多少？

应　用

14. 在过去 6 年中，某所州立大学的招生人数(千人)资料如下所示：



年	1	2	3	4	5	6
招生人数	20.5	20.2	19.5	19.0	19.1	18.8

建立该时间序列的线性趋势成分方程，评论这个学校的招生人数变化情况。

15. 下表提供的是过去 7 年中，在一所重点大学举办家庭足球赛事的平均参加人数，建立该时间序列的线性趋势成分方程。

年	参加人数
1	28 000
2	30 000
3	31 500
4	30 400
5	30 500
6	32 200
7	30 800

16. 下表是 B. J. Scott Motors 公司 10 年的汽车销量时间序列：

年	销量
1	400
2	390
3	320
4	340
5	270
6	260
7	300
8	320
9	340
10	370

请画出这个时间序列的散点图，并评论线性趋势成分的合理性。对于这个时间序列的趋势轨迹，你认为使用哪种类型的函数形式最为适合？

17. 某个小型制造公司的总经理很关注过去几年中制造成本的连续增加。下表所提供的资料，是该公司在过去 8 年中主打产品的单位成本时间序列：

年	单位成本(美元)
1	20.00
2	24.50
3	28.20
4	27.50
5	26.60
6	30.00
7	31.00
8	36.00

a. 画出这个时间序列的图形。是否出现线性趋势?

b. 建立这个时间序列的线性趋势成分方程，该公司意识到它的单位成本每年的平均增加额是多少?

18. 以下是 Walgreen 公司 10 年间的股票每股收益:

0.64，0.73，0.94，1.14，1.33，1.53，1.67，1.68，2.10，2.50

a. 利用这个时间序列的线性趋势来预测下一年的收益。

b. 通过分析这个时间序列，你认为 Walgreen 公司情况怎么样? 历史数据是否可以表明 Walgreen 公司是一个良好的投资对象?

19. 在 90 年代末，很多公司开始裁员以降低公司运营成本，这种成本削减的结果之一是私人企业中经理级的职位百分比下降，以下的数据是在 1990 年到 1995 年中，女性经理的百分比(*The Wall Street Journal Almanac*，1998)。

年	1990	1991	1992	1993	1994	1995
招生人数	7.45	7.53	7.52	7.65	7.62	7.73

a. 建立这个时间序列的线性趋势方程。

b. 利用这个线性趋势方程来估计 1996 年和 1997 年的女性经理的百分比数。

20. 以下是三角航空公司过去 10 年的总收入资料(单位: 百万美元):

年	总收入	年	总收入
1	2 428	6	4 264
2	2 951	7	4 738
3	3 533	8	4 460
4	3 618	9	5 318
5	3 616	10	6 915

a. 建立该时间序列的线性趋势方程。对三角航空公司过去 10 年间的总收入，这个方程能说明什么问题?

b. 预测第 11 年和 12 年的总收入。

21. ACT 网络有限公司开发、销售、制造并销售集成化的多领域网络产品。以下是从 1992 年到 1997 年的年销售收入资料(*Stock Investor Pro*，American Association of Individual Investors，August 31，1997):

年	销量(百万美元)
1992	5.4
1993	6.2
1994	12.7
1995	20.6
1996	28.4
1997	44.9

a. 建立该时间序列的线性趋势方程。

b. 该公司每年的平均销量增长额是多少?

c. 利用这个趋势方程来预测 1998 年的销量。

18.4 趋势和季节成分

我们已经阐述了如何预测含有趋势成分的时间序列。在本节中，我们将把讨论扩展到对同时有趋势和季节成分的时间序列进行预测。

商业和经济上的许多情形涉及到时期对时期的比较。例如，我们或许有兴趣想研究失业人数比上个月增加了 2%，钢产量比上个月增加了 5%，或者电力产量比上个月下降了 3% 等等这些问题。但是，使用这些资料时必须相当谨慎，因为只要有季节影响时，这些比较往往会误导人。例如，从 8 月份到 9 月份，电力的消耗量减少了 3%，这个现象有可能仅仅是因为空调使用减少这个季节因素所造成的，而并不是因为用电量将长期减少。事实上，在调整季节影响以后，我们甚至可能看到用电量反而增加了。

剔除了季节影响以后的时间序列被称为消除了季节影响的时间序列。经过这样处理后，时期和时期的对比将会更有意义，并有助于确定是否有趋势的存在。本章我们所介绍的方法适用于仅有季节影响的情形，也适用于既有季节影响又有趋势成分的情形。第一步是计算季节指数，并用它们来消除季节影响。然后，如果消除了季节影响后的数据有明显的趋势成分，我们可以对它们使用回归分析来估计趋势成分。

乘法模型

除了趋势成分(T)和季节成分(S)外，我们可以假设时间序列有不规则成分(I)。不规则成分用来解释不能被趋势成分和季节成分说明的其他任何随机影响。用 T_t，S_t，I_t 来代表时间序列 t 中的趋势成分、季节成分和不规则成分，我们假设时间序列值以 Y_t 表示，它可以表示成下面的**时间序列乘法模型**(multiplicative time series model)：

$$Y_t = T_t \times S_t \times I_t \tag{18.9}$$

表 18.7　电视机的季度销量数据

年	季度	销量(千台)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4

在这个模型中，T_t 是被预测的项目的趋势。但是，S_t 和 I_t 成分在相关项目中测量，如果其数值大于 1.00 则表明影响在趋势之上，数值小于 1.00 则表明其影响在趋势之下。

根据表 18.7 和图 18.11 中的季度资料，我们可以说明带趋势成分、季节成分和不规则成分的乘法模型的使用。这些资料是某电视机生产厂商在过去 4 年中的电视机销售数量(以千台为单位)。我们首先开始介绍时间序列的季节成分。

计算季节指数

图 18.11 表明在每年的第 2 个季度销量达到最低，第 3 个季度和第 4 个季度销量开始回升。因此，我们可以推断在电视机的销量时间序列中存在季节影响。我们首先通过计算移动平均数，将复合在一起的季节成分 S_t 和不规则成分 I_t 与趋势成分 T_t 区分开，然后开始计算程序来确定每个季度的季节影响。

我们在每一次计算中使用一年的数据。因为所使用的数据是季度时间序列，所以在每一次移动平均值计算中我们都使用 4 项数据。电视机在第一个 4 季度中的销量移动平均值计算如下：

$$\text{第一个移动平均值} = \frac{4.8+4.1+6.0+6.5}{4} = \frac{21.4}{4} = 5.35$$

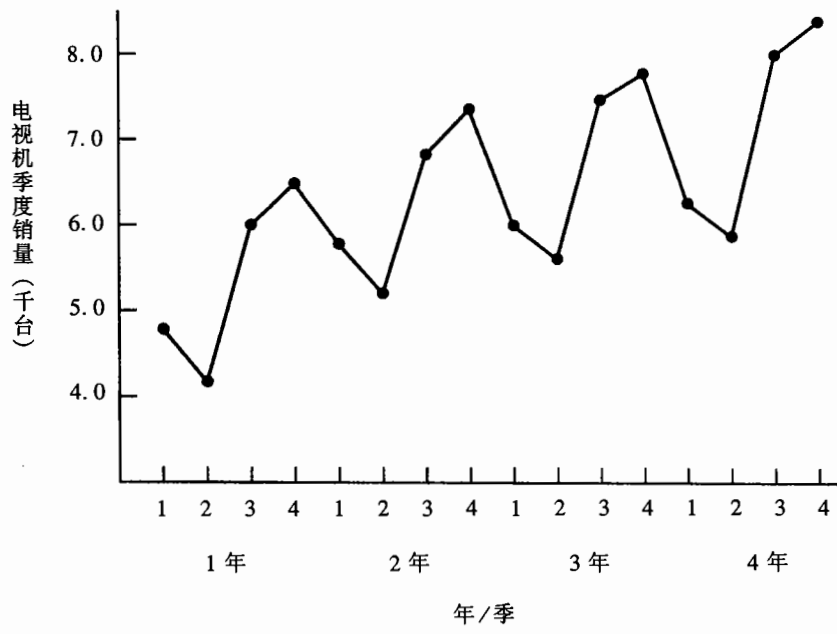


图 18.11　电视机季度销量的时间序列

请注意前 4 个季度的移动平均值正好等于时间序列第 1 年的季度销量的平均数。继续计算移动平均值，加上第 2 年第 1 季度的数值 5.8，去掉第 1 年第 1 季度的数值 4.8，得到第 2 个移动平均数为

$$\text{第二个移动平均值} = \frac{4.1 + 6.0 + 6.5 + 5.8}{4} = \frac{22.4}{4} = 5.60$$

同理，第三个移动平均值计算结果为$(6.0 + 6.5 + 5.8 + 5.2)/4 = 5.875$。

在我们对整个时间序列进行移动平均值计算之前，我们可以回顾第 1 个移动平均值计算结果 5.35。5.35 这个数字正好是第 1 年(包括各个季节)各个季度销量的平均数。我们回头分析 5.35 这个计算结果，发现它对应于所有移动平均数的中间季度。但是，请注意，在确定中间季度时我们会碰到困难，因为在移动平均值中只有 4 个季度，并不存在中间季度。5.35 的计算结果对应着第 2 季度的后半部分和第 3 季度的前半部分，同样，如果我们考察下一个移动平均值 5.60，则中间季度对应着第 3 季度的后半部和第 4 季度的前半部。

回顾计算移动平均数的目的，是为了将复合的季节成分和不规则成分分离。但是，我们所计算出来的移动平均值并不直接对应着时间序列的原始季度。我们可以解决这个困难：在这连续的移动平均值中取中间值。例如，如果 5.35 对应着第 3 季度的前半部，5.60 对应着第 3 季度的后半部，则我们可以取$(5.35 + 5.60)/2 = 5.475$作为第 3 季度的平均值。同样，我们可以用$(5.60 + 5.875)/2 = 5.738$来对应第 4 季度。这个结果被称为**移动平均值的中间值**。表 18.8 是关于电视机销量数据的移动平均值计算的完整总结。

如果在移动平均值计算过程中，数据点的数目是单数，则中间点将对应着时间序列中的某个时期。在这种情况下，我们没有必要像在表 18.8 中的计算那样，对移动平均值取中间值以对应于某个特定的时间序列。

表 18.8 中的移动平均值中间值能告诉我们时间序列的什么问题呢？图 18.12 是时间序列实际值和移动平均值中间值的散点图。请特别注意移动平均值的中间值是如何消除时间序列的季节和不规则波动的影响的。在所计算出的 4 个季度的移动平均数中，并不包括季节影响所产生的波动，因为季节成分的影响已经被剔除了。移动平均值中间值的每一个数据点所代表的时间序列都是没有季节成分和不规则成分影响的数值。

用时间序列的每一个观测值除以相对应的移动平均值中间值，我们可以确定时间序列的季节不规则影响值。例如，第 1 年第 3 季度的综合季节不规则影响值为 $6.0/5.475 = 1.096$。表 18.9 总结了整个时间序列的季节不规则影响值。

考察第 3 季度。第 1、2、3 年第 3 季度的季节影响值分别是 1.096，1.075 和 1.109。因此，在整个过程中，第 3 季度的季节不规则影响值比全年平均水平要高。由于季节不规则值每年的波动主要都是由于不规则成分所引起的，我们可以取其平均值，这样做可以消除不规则影响并得到第 3 季度的季节影响估计值。

$$\text{第 3 季度季节影响值} = \frac{1.096 + 1.075 + 1.109}{3} = 1.09$$

我们倾向于使用 1.09 作为第 3 季度的**季节指数**。在表 18.10 中，我们总结了关于电视机销量时间

表 18.8　关于电视机销量时间序列的移动平均值中间值的计算结果

年	季度	销量(千台)	第 4 季度移动平均值	移动平均值中间值
1	1	4.8		
	2	4.1		
	3	6.0	5.350	5.475
	4	6.5	5.600	5.738
2	1	5.8	5.875	5.975
	2	5.2	6.075	6.188
	3	6.8	6.300	6.325
	4	7.4	6.350	6.400
3	1	6.0	6.450	6.538
	2	5.6	6.625	6.675
	3	7.5	6.725	6.763
	4	7.8	6.800	6.838
4	1	6.3	6.875	6.938
	2	5.9	7.000	7.075
	3	8.0	7.150	
	4	8.4		

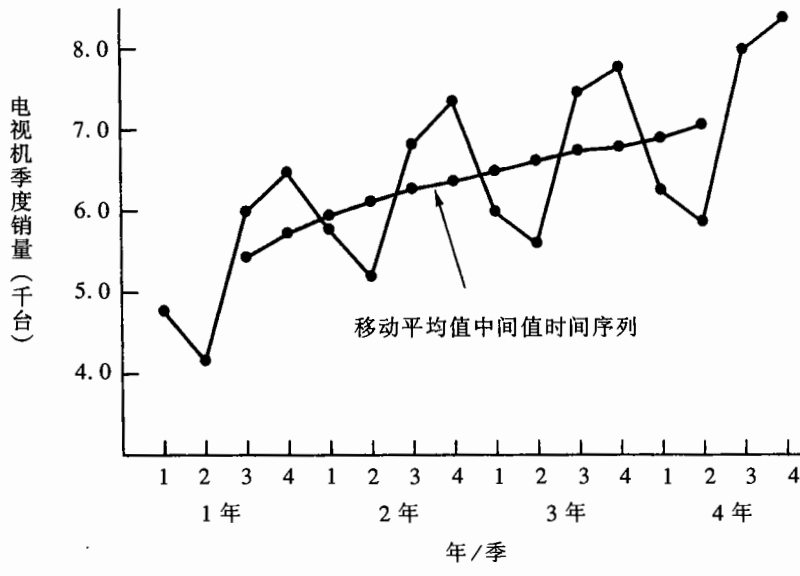


图 18.12　电视机季度销量时间序列和移动平均值中间值

序列季节指数的计算结果。四个季度的季节指数分别是：第 1 季度 0.93；第 2 季度 0.84；第 3 季度 1.09；第 4 季度 1.14。

表 18.10 中数值的评注提供了关于电视机销量的季节成分的一些观测值。最好的销售季度是第4 季度，销量比平均季度销量高出 14%。最差的，或者最少的销售季度是第 2 季度；它的季节指数 0.84表明其平均销量比季度平均销量值低 16%。季节成分很明显地对应着人们观看电视的兴趣，由于

表 18.9 电视机销量时间序列的季节不规则影响值

年	季度	销量(千台)	移动平均值	季节不规则成分值
1	1	4.8		
	2	4.1		
	3	6.0	5.475	1.096
	4	6.5	5.738	1.133
2	1	5.8	5.975	0.971
	2	5.2	6.188	0.840
	3	6.8	6.325	1.075
	4	7.4	6.400	1.156
3	1	6.0	6.538	0.918
	2	5.6	6.675	0.839
	3	7.5	6.763	1.109
	4	7.8	6.838	1.141
4	1	6.3	6.938	0.908
	2	5.9	7.075	0.834
	3	8.0		
	4	8.4		

表 18.10 电视机销量时间序列的季节指数计算

季度	季节不规则成分值(SI_t)	季节指数(S_t)
1	0.971, 0.918, 0.908	0.93
2	0.840, 0.839, 0.834	0.84
3	1.096, 1.075, 1.109	1.09
4	1.133, 1.156, 1.141	1.14

冬季到来人们的室外活动减少,因此观看电视的时间增加导致第 4 季度电视机销量达到高峰。第 2 季度的低销量反映了人们观看电视的兴趣减小,因为春季和初夏的活动吸引了很多电视的潜在消费者。

在计算季节指数时，有时候我们需要对它进行最后的调整。乘法模型需要平均季节指数等于 1.00，所以表 18.10 中的四个季节指数的总和必须等于 4.00。换句话说，季节影响在一年之内必须是持平的。在我们的例子中季节指数的平均值等于 1.00，因此，这种类型的调整就没有必要。在别的情形中，或许需要一个轻微的调整。季节指数进行调整是用每个季节指数乘以季度总和,然后再除以未调整的季节指数之和。举个例子,对于季度数据,每个季节指数乘上4再除以未调整的季节指数的和。为了获得合适的季节指数,有些练习需要进行这样的调整。

消除季节影响的时间序列

消除了季节影响后的资料可以用来比较连续时期的销量。没有消除季节影响的资料，相关的比较可在本期销售量和一年前同期销售量之间进行。

计算季节指数的目的在于从时间序列中消除季节指数的影响。这个过程被称为**消除**时间序列

的季节影响。调整季节差异后的经济时间序列**消除季节影响的时间序列**(deseasonalized time series)经常刊登在诸如《当前商务概览》(*Survey of Current Business*)、《华尔街日报》和《商业周刊》上。利用乘法模型的标记，我们有

$$Y_t = T_t \times S_t \times I_t$$

将每个时间序列观测值除以相对应的季节指数，我们可以将季节影响从时间序列中消除。表 18.11 总结了关于消除了季节影响的电视机销量时间序列，如图 18.13 所示。

表 18.11　消除了季节影响的电视机销量时间序列值

年	季度	销量(千台)(Y_t)	季节指数(S_t)	消除了季节影响的销量($Y_t / S_t = T_t I_t$)
1	1	4.8	0.93	5.16
	2	4.1	0.84	4.88
	3	6.0	1.09	5.50
	4	6.5	1.14	5.70
2	1	5.8	0.93	6.24
	2	5.2	0.84	6.19
	3	6.8	1.09	6.24
	4	7.4	1.14	6.49
3	1	6.0	0.93	6.45
	2	5.6	0.84	6.67
	3	7.5	1.09	6.88
	4	7.8	1.14	6.84
4	1	6.3	0.93	6.77
	2	5.9	0.84	7.02
	3	8.0	1.09	7.34
	4	8.4	1.14	7.37

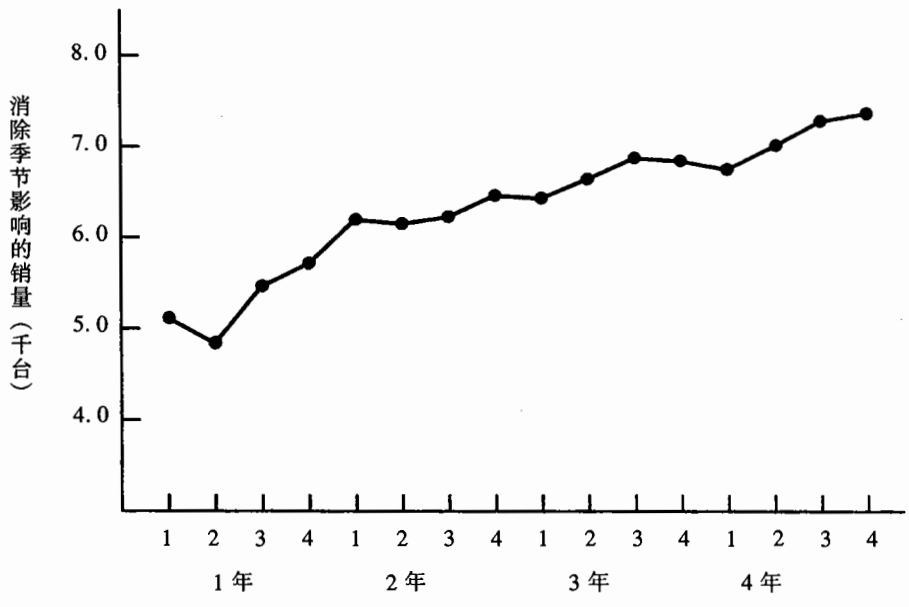


图 18.13　消除了季节影响的电视机销量时间序列

利用消除了季节影响的时间序列来分析趋势

虽然图 18.13 所显示的过去 16 个季度图像呈现出上下波动的运动轨迹，但是时间序列还是有明显的线性上升趋势。同样，我们可以使用前面章节所述的步骤来确定这个趋势；在这种情况下，数据是消除了季节影响后的季度销量。因此，对于这个线性趋势，销量的估计值可以表述成时间的函数，如下所示：

$$T_t = b_0 + b_1 t$$

式中 T_t——t 期电视机销量的趋势值；

b_0——趋势线的截距；

b_1——趋势线的斜率。

和前面一样，$t=1$ 对应于时间序列的第 1 个观测值的时间，$t=2$ 对应于时间序列的第 2 个观测值的时间，依此类推。因此，对消除了季节影响的电视机销量时间序列，$t=1$ 对应着第 1 个消除了季节影响的季度销量，$t=16$ 对应着最近的消除了季节影响的季度销量。b_0 和 b_1 的计算公式如下所示：

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n}$$

$$b_0 = \overline{Y} - b_1 \overline{t}$$

但是，请注意 Y_t 现在表示的是 t 时期的消除了季节影响的时间序列值，不是时间序列的实际值。利用 b_0 和 b_1 的给定关系和表 18.11 中的消除了季节影响的销量数据，我们可以计算如下：

t	Y_t(消除了季节影响)	tY_t	t^2
1	5.16	5.16	1
2	4.88	9.76	4
3	5.50	16.50	9
4	5.70	22.80	16
5	6.24	31.20	25
6	6.19	37.14	36
7	6.24	43.68	49
8	6.49	51.92	64
9	6.45	58.05	81
10	6.67	66.70	100
11	6.88	75.68	121
12	6.84	82.08	144
13	6.77	88.01	169
14	7.02	98.28	196
15	7.34	110.10	225
16	7.37	117.92	256
总计 136	101.74	914.98	1 496

式中　$\bar{t}=\frac{136}{16}=8.5$

$$\bar{Y}=\frac{101.74}{16}=6.359$$

$$b_1=\frac{914.98-(136)(101.74)/16}{1\,496-(136)^2/16}=0.148$$

$$b_0=6.359-0.148(8.5)=5.101$$

因此

$$T_t=5.101+0.148t$$

是这个时间序列的线性趋势成分的表达式。

斜率 0.148 表示在过去 16 个季度中，在消除了季节影响以后，公司的平均销量增长值为每个季度 148 台。如果我们假设过去 16 个季度销量趋势依然适用于未来，则这个方程可以用来推测未来季度的时间序列趋势成分。例如，将 $t=17$ 代入方程，可以得到下一个季度的趋势推测值 T_{17}：

$$T_{17}=5.101+0.148(17)=7.617$$

因此，趋势方程表明下一个季度的电视机销量预测为 7 617 台。同样，趋势方程对第 18，19，20 季度所作出的电视机销量预测分别为 7 765，7 913，8 061 台。

季节调整

当趋势和季节两种成分都存在时，进行预测的最后一步是利用季节指数调整趋势预测。再来看电视机销售的例子，我们对未来四个季度进行了趋势预测。现在我们必须用季节影响来调整预测值。第 5 年第 1 季度($t=17$)的季节指数是 0.93，所以，我们用季节指数 0.93 乘以趋势预测值($T_{17}=7\,617$)得到季节预测值。那么，该季度的预测值是 $7\,617(0.93)=7\,084$。表 18.12 给出了第 17 到第 20个季节的季节预测。第 4 季度的预测值最大，销售 9 190 台；最低销售量是第 2 季度，销售 6 523 台。

表 18.12　电视机销量时间序列的季度预测值

年	季度	趋势预测值	季节指数 (见表 18.11)	季度预测值
5	1	7 617	0.93	(7 617)(0.93) = 7 084
	2	7 765	0.84	(7 765)(0.84) = 6 523
	3	7 913	1.09	(7 913)(1.09) = 8 625
	4	8 061	1.14	(8 061)(1.14) = 9 190

根据月度资料建立模型

在前面电视机销售的例子中，我们利用季度数据说明季节指数的计算。然而，许多企业利用月度数据作预测而不是作季节预测。这种情况下，本节介绍的方法应用时应做些小的修正。第一

步，用 12 个月的滑动平均替代 4 个季度的滑动平均；第二步，必须计算出来 12 个月的季节指数，而不是 4 个季度的季节指数。除此之外，其他的计算和预测过程相同。

循环成分

公式(18.9)的乘法模型在数学上可以扩展到包含循环成分，即

$$Y_t = T_t \times C_t \times S_t \times I_t \qquad (18.10)$$

循环成分和季节成分一样也可以表示为趋势的百分比。正如 18.1 节中所提到的，循环成分可以归因于时间序列的多年循环。它和季节成分相似，只是所经历的时期更长。但是，由于所涉及的时期太长，所以要获得充分的相关数据资料来估计循环成分，通常是有一定困难的。另外一个困难在于，循环周期通常是变化的。我们将在关于预测方法的教材中对循环成分作更进一步的讨论。

练　习

方　法

22. 考察下面的时间序列数据：



季度	年		
	1	2	3
1	4	6	7
2	2	3	6
3	3	5	6
4	5	7	8

a. 计算这个时间序列 4 个季度的移动平均值和移动平均值的中间值。

b. 计算 4 个季度的季节指数。

应　用

23. 某大学过去 3 年的教材季度销量(所销售的册数)数据如下：

季度	第 1 年	第 2 年	第 3 年
1	1 690	1 800	1 850
2	940	900	1 100
3	2 625	2 900	2 930
4	2 500	2 360	2 615

a. 计算这个时间序列 4 个季度的移动平均值和移动平均值中间值。

b. 计算 4 个季度的季节指数。

c. 教材出版商什么时候的季节指数最大？这样的结果是否合理？请解释。

24. 以下是佛罗里达州南部一个 6 单元公寓 3 年中的支出资料，请计算每个月的季节指数。用 12 个月的移动平均值计算：

	支出		
	第 1 年	**第 2 年**	**第 3 年**
1 月	170	180	195
2 月	180	205	210
3 月	205	215	230
4 月	230	245	280
5 月	240	265	290
6 月	315	330	390
7 月	360	400	420
8 月	290	335	330
9 月	240	260	290
10 月	240	270	295
11 月	230	255	280
12 月	195	220	250

25. 在南加利福尼亚州，空气污染控制专家每小时检测空气中的臭氧、二氧化碳和二氧化氮的含量，每小时资料所组成的时间序列包含季节成分，它显示了一天中每个小时的污染物水平。以下是在 7 月 15 日，16，17 日，从早晨 6 点到下午 6 点的 12 个小时中，在城市中心区域观测到的二氧化氮含量水平。

7 月 15 日：25，28，35，50，60，60，40，35，30，25，25，20

7 月 16 日：28，30，35，48，60，65，50，40，35，25，20，20

7 月 17 日：35，42，45，70，72，75，60，45，40，25，25，25

a. 对每天的 12 个数据，确定每个小时的季节指数。

b. 根据(a)中的季节指数，可以得到消除了季节影响后的数据。由此而建立的趋势方程为 $T_t = 32.983 + 0.3922t$。仅仅考虑趋势成分，预测 7 月 18 日中 12 个小时的数据。

c. 利用(a)中的季节指数调整(b)中的趋势预测值。

26. 电力的消耗量是以千瓦每小时来测量的。某地公用事业公司有一个电力供应中断程序，需要商业用户的参与，如果用户参与该程序，则电力收费将会优惠，但是参与用户在该公司要求他们削减用电量时亦必须答应这一要求。Timko 公司在星期四上午 12 点削减用电量。为了评

估所节约的用电量，公用事业公司必须估计 Timko 公司在正常耗电情况下的用电量。停电时间段是从中午到晚上 8 点。Timko 公司前 72 个小时的用电量数据如下：

时间段	星期一	星期二	星期三	星期四
12 点— 4 点(上午)	—	19 281	31 209	27 330
4 点— 8 点(上午)	—	33 195	37 014	32 715
8 点—12 点(中午)	—	99 516	119 968	152 465
12 点— 4 点(下午)	124 299	123 666	156 033	
4 点— 8 点(下午)	113 545	111 717	128 889	
8 点—12 点(午夜)	41 300	48 112	73 923	

a. 24 小时时间段内是否有季节影响？对这 6 个 4 小时区间计算季节指数。

b. 利用调整了季节指数的趋势来估计 Timko 公司在停止供电服务期间的正常用电量。

18.5 回归分析

在第 14，15 和 16 章关于回归分析的讨论中，我们阐述了如何利用一个或多个自变量来预测单个因变量的值。现在采用回归分析来作为预测工具，我们可以把想要预测的时间序列值作为因变量。因此，如果能够确定一系列相应的自变量(或预测)和变量的集合，则我们可以建立估计回归方程来预测时间序列。

我们在 18.3 节中拟合自行车销量时间序列的线性趋势线时，所用的方法即是回归分析的一个特例。在那个例子中有两个变量，自行车销量和时间，呈线性相关。*在现实世界中，许多问题固有的复杂性需要考虑设置多个变量来预测目标变量，因此，多元回归分析的统计方法适用于这种情况。

回顾建立多元回归方程的情形，我们需要自变量和因变量观测值的样本。在时间序列分析中，n 期时间序列资料可以给每个变量提供 n 个观察值。对包含 k 个变量的函数，我们使用如下标记：

Y_t ——t 时期时间序列值

x_{1t} ——t 时期变量 1 的值

x_{2t} ——t 时期变量 2 的值

……

x_{kt} ——t 时期变量 k 的值

下表给出了建立估计回归方程所需的 n 期数据资料：

* 从纯技术的角度来看，自行车销量与时间并不相关；相反，如果把时间作为与自行车销量实际相关的一个变量，则它往往是未知的，或者很难测量或者测量费用太高。

时期	时间序列(Y_t)	自变量的数值						
		x_{1t}	x_{2t}	x_{3t}	…	…	…	x_{kt}
1	Y_1	x_{11}	x_{21}	x_{31}	…	…	…	x_{k1}
2	Y_2	x_{12}	x_{22}	x_{32}	…	…	…	x_{k2}
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
n	Y_n	x_{1n}	x_{2n}	x_{3n}	…	…	…	x_{kn}

正如你所设想的那样，预测模型中有几种可能的自变量选择。其中可能的一种是把时间作为自变量。在 18.3 节中即是这种情形，在那一节里，我们利用以时间作自变量的线性函数来预测时间序列的趋势。令 $x_{1t}=t$，我们可以得到估计回归方程的形式：

$$\hat{Y}_t = b_0 + b_1 t$$

式中，$\hat{y}_t$ 是时间序列值 Y_t 的估计值，b_0 和 b_1 是估计回归系数。在更为复杂的模型中，随着时间幂次的增加将会增加相对应的附属项目。例如，如果 $x_{2t}=t^2$，$x_{3t}=t^3$，则估计回归模型将会成为如下形式：

$$\begin{aligned}\hat{Y}_t &= b_0 + b_1 x_{1t} + b_2 x_{2t} + b_3 x_{3t} \\ &= b_0 + b_1 t + b_2 t^2 + b_3 t^3\end{aligned}$$

请注意，这个模型是用随着时间推移所产生的曲线特征来提供时间序列的预测。

其他以回归为基础的预测模型包含有经济和人口统计学相混合的自变量。例如，在预测冰箱的销量时，我们可以选择以下自变量：

x_{1t}——t 时期的价格；
x_{2t}——$t-1$ 时期整个行业的销量；
x_{3t}——$t-1$ 时期允许建造的建筑物数量；
x_{4t}——t 时期的人口预测值；
x_{5t}——t 时期的广告支出预算。

根据通常的多元回归方法，我们可以用包含 5 个自变量的估计回归方程来进行预测。

Spyros Makridakis，一个著名的预测专家所进行的研究表明：对于短期预测，简单的预测方法通常比更加复杂的预测方法更为有效。使用更为复杂和昂贵的预测方法不一定能保证更好的预测结果。

回归方法是否提供了合适的预测结果，主要取决于我们所确定的自变量是否与时间序列密切相关，以及我们是否能够获得这些与时间序列密切相关的数据。通常说来，在建立估计回归方程的时候，我们会考虑很多可能的自变量集合。因此，回归分析的部分步骤应该是对自变量的选

择，争取所选择的自变量能够提供最好的预测模型。

在本章的内容介绍部分，我们阐述了**因果预测模型**(causal forecasting models)，这种模型是用和被预测时间序列相关的其他时间序列，来解释它的行为特性的。回归分析是建立这种因果模型的最常用的工具。在这种模型里，相关的时间序列作为自变量，被预测的时间序列则作为因变量。

在以回归为基础的另外一种类型的预测模型中，将该时间序列以前的所有数值作为自变量。例如，假设时间序列值用 Y_1，Y_2，…，Y_n 来表示，如果因变量为 Y_t，则我们可以建立一个估计回归方程，将 Y_t 与最近的时间序列值 Y_{t-1}，Y_{t-2}，等等联系起来。将这三个最近的时期作为自变量，估计回归方程形式如下：

$$\hat{Y}_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3}$$

将时间序列以前的数值作为自变量的回归模型被称为**自回归模型**(autoregressive models)。

最后，另外一种以回归为基础的预测方法是将前面讨论过的各种自变量综合在一起。例如，我们可以选择时间变量的组合，包括一些经济的或人口统计的变量，以及时间序列变量本身以前的数值。

18.6 定性预测方法

如果历史资料不可用，管理人员必须使用定性预测方法来进行预测。但是使用定性方法的成本由于所涉及的人员时间的投入可能会变得很高。

在前面的章节中，我们讨论过几种类型的定量预测模型，其中大部分的技巧都需要研究历史资料，所以在历史资料不可用时这些模型也就无法应用。除此之外，即使那些历史资料可用，影响时间序列的环境如果发生显著的变化，这种使用历史资料来预测时间序列未来值的做法也就值得怀疑。例如，强制的汽油配售计划对根据历史资料来预测汽油销量预测的有效性提出了质疑。定性预测给这些情形以及其他情形提供了备用的预测方法。

德尔菲法

德尔菲法(Delphi method)是最常用的定性预测方法之一，最初是由兰德公司的一个研发小组开发出来的。它试图通过达成“小组意见的一致”来进行预测。在它通常的应用中，专家小组的成员被互相分开，并且彼此互不认识，这些成员需要回答一系列的咨询问题。第一轮问答的结果经过统计分析，被用于第二轮的提问调查，第二轮的提问包含了这个小组的答案信息和各种观点。因此，根据已经整理出来的小组信息和观点，每个专家在第一轮的答案都可能需要重新考虑和修正。这个过程一直持续，直到小组协调人感觉到整个小组在某种程度上已经达成一致意见为止。德尔菲法的目的不是提供单一的答案作为结果，而是给出一个比较小的范围，在这个范围中

有专家小组的主流意见。

专家判断法

经验证明和理论分析都建议在进行判断预测时应该选择 5 到 20 名专家。但是，在涉及到指数增长的情形时，判断预测法是不适用的。

定性预测通常以单个专家的判断或者一组专家的一致性意见为基础。例如，每年美林的专家汇集起来，预测下一年的道琼斯工业指数的平均水平和主要利率。在预测的过程中，专家们每人单独考虑他们认为将会影响股票市场和利率的信息；然后他们将结论结果总结成预测。在这些过程中，没有使用正式的模型，也不存在两个专家得出同一信息的情况。

当历史条件不适用于将来时，我们通常推荐使用专家判断预测方法。虽然在预测过程中没有使用正式的定量模型，但专家判断法在很多情形下所提供的预测都是相当精确的。

远景规划法

远景规划法(scenario writing)作为一种定性预测方法，包含根据一系列既定假设来建立一个概念性的远景规划。不同系列的假设导致不同的远景规划。决策者的工作就是决定每一个远景规划的相似性，然后据此作出决策。

直观预测法

主观或直观的定性预测方法根据的是人类处理不同信息的思维能力，而在大多数情况下，它是难以量化的。这些方法通常应用于团队工作，在团队中委员会或者工作小组通过“头脑风暴”之类的会议来寻求新的方法或者解决复杂问题。在这些讨论会中，每个人不像通常在小组中那样受到压力和批评的限制，因为他们阐述任何想法或者观点时不需要考虑它是否恰当，甚至更重要的是，不需要害怕遭到批评。

小 结

本章我们介绍了时间序列分析和预测的基本方法。首先，我们阐述了怎样解释时间序列的行为特性，将时间序列看成是包含四种成分：趋势、循环、季节和不规则成分，这样有助于加深我们对它的理解。将这些成分区分开来，并测量它们各自的明显影响，这样我们就能够预测时间序列的未来数值。

我们讨论了在没有显著的趋势、季节或循环成分影响时，如何用平滑法来预测时间序列。移

动平均法包括计算历史数据的平均值，然后将这个平均值作为下一时期的预测。在指数平滑法中，我们用时间序列历史数据的加权平均值来计算预测值。

对于只有长期趋势的时间序列，我们阐述了怎样利用回归分析来做趋势推测。对于趋势和季节影响都很显著的时间序列，我们讨论了怎样将这两个因素的影响分开以取得更好的预测结果。最后，我们说明了回归分析可以作为建立因果预测模型的方法。因果预测模型表明了时间序列值(因变量)与其他自变量的关系，利用这些自变量我们可以解释时间序列的行为特性。

当预测所需要的历史数据仅有部分可用或完全不可用时，我们讨论了使用定性预测方法。当时间序列的历史轨迹不会继续延伸到将来时，使用这些定性预测方法是最为合适的。

术语释义

时间序列： 时间序列是一个变量在连续时点或持续时期上的观测值的集合。

预测： 对一个时间序列未来值的推测和判断。

趋势： 在时间序列的几个时期内观测值的长期移动或者运动。

循环成分： 围绕着趋势线上下波动时间超过一年的轨迹点可以被称为时间序列的循环成分。

季节成分： 在一年或一年之内显示周期性运动的时间序列模型成分。

不规则成分： 时间序列的成分，被用来解释在剔除了趋势、循环和季节成分影响以后的时间序列实际变动现象。

移动平均法： 一种预测时间序列的方法，它通过求每一个连续数列值的平均值而得到预测结果。

均方误差： 测量预测模型预测精度的方法。它是预测值和时间序列实际值之差的平方和的平均值。

加权移动平均法： 预测时间序列的一种方法，计算历史数据的加权平均值。权数的总和必须为 1。

指数平滑法： 一种预测方法，利用历史时间序列值的加权平均值来得到可以用做预测的平滑的时间序列值。

平滑常数： 指数平滑模型的参数，即在计算预测值时，最近时间序列的给定权值。

时间序列乘法模型： 将时间序列的独立成分乘在一起来确定时间序列的预测值的模型。如果假设中有趋势、循环、季节和不规则四种成分，则模型为 $Y_t = T_t \times C_t \times S_t \times I_t$；如果不考虑循环成分，则模型为 $Y_t = T_t \times S_t \times I_t$。

消除季节影响的时间序列： 用时间序列的每一个原始观测值除以相对应的季节指数后所得到的时间序列。

因果预测模型： 一种说明时间序列与其他相关联变量的关系的预测方法，那些变量被认为可以解释或者产生时间序列的行为特性。

自回归模型： 一种时间序列模型，其回归关系仅仅依赖于历史时间序列值。用这种模型可以预测未来的时间序列值。

德尔菲法： 一种定性预测模型，它通过小组意见的一致性来进行预测。

远景规划法： 远景规划法作为一种定性预测方法，包含根据一系列既定假设来建立一个概念性的远景规划。

重要公式

移动平均值

$$\text{移动平均值} = \frac{\Sigma(\text{最近 } n \text{ 个数值})}{n} \tag{18.1}$$

指数平滑模型

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t \tag{18.2}$$

线性趋势方程

$$T_t = b_0 + b_1 t \tag{18.5}$$

带趋势、季节、不规则成分的时间序列乘法模型

$$Y_t = T_t \times S_t \times I_t \tag{18.9}$$

带趋势、循环、季节和不规则成分的时间序列乘法模型

$$Y_t = T_t \times C_t \times S_t \times I_t \tag{18.10}$$

补充练习

27. 移动平均值经常用于确定股价波动。下面是反斗城游戏公司 Toys R Us 股票在 1997 年 9 月 22 日—12 月 8 日的每周收盘价(Prudential Securities，Inc.)：

周	价格(美元)	周	价格(美元)
9 月 22 日	34 7/8	11 月 3 日	33 5/8
9 月 29 日	35 5/8	11 月 10 日	35 1/16
10 月 6 日	34 11/16	11 月 17 日	34 1/16
10 月 13 日	33 9/16	11 月 24 日	34 1/8
10 月 20 日	32 5/8	12 月 1 日	33 1/4
10 月 27 日	34	12 月 8 日	32 1/16

a. 计算这个时间序列 3 个月的移动平均值。并预测 1997 年 12 月 15 日的收盘价。

b. 计算这个时间序列 3 个月的加权移动平均值。最近的时期权值为 0.4，其次也为 0.4，第三个时期的权值为 0.2，预测 1997 年 12 月 15 日的收盘价。

c. 对这个时间序列应用指数平滑法，平滑常数 $\alpha = 0.35$，预测 1997 年 12 月 15 日的收盘价。

d. 以上三种方法你倾向于使用哪一种，为什么？

28. 美国联邦选举委员会保存有关于选举的一些数据：到达选举年龄的人口，登记选举的人口数，美国联邦选举集会人口数。下表显示的是从1972年到1996年美国联邦选举集会人口数占到达选举年龄人口数的百分比(*The Wall Street Journal Almanac*，1998)。

年	选举集会人数(%)	年	选举集会人数(%)	年	选举集会人数(%)
1972	55	1982	40	1990	37
1974	38	1984	53	1992	55
1976	54	1986	36	1994	39
1978	37	1988	50	1996	49
1980	53				

a. 用指数平滑法来预测这个时间序列，指数平滑常数 $\alpha=0.1$ 和 0.2，哪一个指数平滑常数提供的预测结果更好？

b. 1998年参加选举集会的人口比例预测值是多少？

29. 个人投资者所拥有的有价证券中股票所占的百分比取决于经济状况。1997年4月，一个典型的有价证券组合为：现金(19%)，股票(30%)，股票基金(37%)，债券(8%)，债券基金(6%)(*AAII Journal*，June 1997)。下表是1995年到1997年9个季度中，一个典型的有价证券组合中股票所占的百分比：

季度	股票所占百分比
1995年1季度	29.8
1995年2季度	31.0
1995年3季度	29.9
1995年4季度	30.1
1996年1季度	32.2
1996年2季度	31.5
1996年3季度	32.0
1996年4季度	31.9
1997年1季度	30.0

a. 利用指数平滑法来预测这个时间序列，指数平滑常数 $\alpha=0.2$，0.3 和 0.4。哪一个指数平滑常数所提供的预测结果更好？

b. 1997年第2季度有价证券中股票所占百分比的预测值是多少？

30. 某杂货连锁店注意观察某特定品牌的自动洗碗机清洁剂每周的需求量(单位：箱)。指数平滑常数 $\alpha=0.2$，用指数平滑法来预测第11周的需求量。

周	需求量
1	22
2	18
3	23
4	21
5	17
6	24
7	20
8	19
9	18
10	21

31. 联合乳业有限公司给佛罗里达州 Dade 县的几个独立的杂货商店供应牛奶，联合乳业有限公司的管理人员想建立半加仑包装的牛奶的销量预测。过去 12 周的销售数量数据如下所示：

周	销量	周	销量
1	2 750	7	3 300
2	3 100	8	3 100
3	3 250	9	2 950
4	2 800	10	3 000
5	2 900	11	3 200
6	3 050	12	3 150

用指数平滑常数 $\alpha=0.4$ 预测 13 周的牛奶需求量。

32. 第 7 花园大道销售音乐演出磁带，下表是过去 18 个月的销量数据。集团的管理人员需要对未来销量进行精确的预测。

月	销量	月	销量
1	293	10	433
2	283	11	470
3	322	12	481
4	355	13	549
5	346	14	544
6	379	15	601
7	381	16	587
8	431	17	644
9	424	18	660

a. 利用指数平滑常数 $\alpha=0.3$，0.4，0.5 进行预测。哪一个指数平滑常数提供的预测更为精确呢？

b. 利用趋势推测法进行预测。MSE 值是多少？

c. 你会向管理者推荐使用哪一种预测方法？为什么？

33. 衣阿华州 Davenport 的 Mayfair 杂货店由于密西西比河的水灾，被迫在 7 月和 8 月停业，该店计划确定它由于停业所遭受的销量损失。从 1 月到 6 月的销量资料如下所示：

月份	销量(千美元)
1	185.72
2	167.84
3	205.11
4	210.36
5	255.57
6	261.19

a. 利用指数平滑常数 $\alpha=0.4$，用指数平滑法预测 7 月和 8 月的销量。(提示：用 7 月份的预测值作为实际销量来预测 8 月份的销量)讨论对于将来不止一期的预测，指数平滑法的用处。

b. 利用趋势推测法来预测 7 月和 8 月的销量。

c. 以 7 月和 8 月所损失的销量为基础，Mayfair 所投保的保险公司提出了一个赔偿方案 240 000 美元。这个方案合理吗？如果不合理，你认为合理的赔偿额应该是多少？

34. Canton 公司是一个雇用了将近 100 名员工的服务性公司。Canton 公司的管理者很关注每月的现金流量，并计划预测每月的现金需求量。由于最近运营策略的变化，仅有过去 7 个月的历史资料是可用的，根据下面的历史资料，利用趋势推测法来预测今后两个月中每个月的现金需求量。

月份	1	2	3	4	5	6	7
现金需求量(千美元)	205	212	218	224	230	240	246

35. 下表是从 1980 年到 1994 年的有线电视用户的数量(*Television & Cable Factbook*，1994)。

年	数量
1980	16 000
1981	18 300
1982	21 000
1983	25 000
1984	30 000
1985	32 000
1986	37 500
1987	41 100
1988	44 000

（续表）

年	数量
1989	47 500
1990	50 000
1991	51 000
1992	53 000
1993	55 000
1994	57 000

a. 对这个时间序列建立线性趋势方程。讨论这个方程说明了 15 年来有线电视用户的数量发生了什么样的变化。

b. 预测 1995 年和 1996 年的有线电视用户数量。

36. Costello 音乐公司已经运营 5 年了。在这期间，电子琴的销量从第一年的 12 个单位增长到最近的 76 个单位。Fred Costello，公司的拥有者，计划预测来年的电子琴销量。历史数据如下所示：

年	1	2	3	4	5
销量	12	28	34	50	76

a. 画出这个时间序列的图像。图中是否出现线性趋势？

b. 建立这个时间序列的线性趋势方程。该公司每年实现的销量平均增长额是多少？

37. Hudson Marine 公司在过去 7 年中是 C&D 航海无线电设备的授权代理商。下表是每年无线电设备的销售数量：

年	1	2	3	4	5	6	7
销售数量	35	50	75	90	105	110	130

a. 画出这个时间序列的图像。图中是否出现线性趋势？

b. 建立这个时间序列的线性趋势方程。

c. 利用(b)中所建立的线性趋势来预测第 8 年的销量。

38. 美国电影协会(MPAA)收集了关于它的协会成员拍摄电影的成本资料，包括迪斯尼、派拉蒙、环球、华纳兄弟、米高梅、福克斯、索尼和特纳。下面的数据显示的是从 1985 年到 1996 年美国电影协会成员平均拍摄成本费用(*The Wall Street Journal Almanac*，1998)：

年	平均拍摄成本(千美元)
1985	16 779
1986	17 455
1987	20 051
1988	18 061
1989	23 454
1990	26 783

（续表）

年	平均拍摄成本(千美元)
1991	26 136
1992	28 858
1993	29 910
1994	34 288
1995	36 370
1996	39 836

a. 画出时间序列的散点图，并评论线性趋势的合理性。

b. 建立这个时间序列的线性趋势方程。

c. 美国电影协会成员拍摄成本费用每年的平均增长额是多少？

d. 利用趋势方程来预测 1997 年和 1998 年的拍摄成本费用。

39. 下表给出的是从 1986 年到 1994 年每年的移动电话用户数量(单位：千户)(*State of the Cellular Industry*，1994)。

年	数量	年	数量
1986	682	1990	5 283
1987	1 231	1991	7 557
1988	2 069	1992	11 033
1989	3 509	1993	16 009

画出这个时间序列的散点图，并评论线性趋势的合理性。对这个时间序列的函数轨迹，你认为什么样的函数形式最为适合？

40. 参考练习 37 中 Hudson Marine 的例子，假设 7 年来季度销量的历史数据如下：

年	第 1 季度	第 2 季度	第 3 季度	第 4 季度	总计年销量
1	6	15	10	4	35
2	10	18	15	7	50
3	14	26	23	12	75
4	19	28	25	18	90
5	22	34	28	21	105
6	24	36	30	20	110
7	28	40	35	27	130

a. 计算这个时间序列 4 个季度的移动平均值。并在同一个图上画出原来的时间序列图和移动平均值序列图。

b. 计算 4 个季度的季节指数。

c. 在什么时候 Hudson Marine 公司所经受的季节影响最大？这个季节影响是否合理？请解释。

41. 参练习 36 中 Costello 音乐公司的例子，季度销量数据如下：

年	第 1 季度	第 2 季度	第 3 季度	第 4 季度	总计年销量
1	4	2	1	5	12
2	6	4	4	14	28
3	10	3	5	16	34
4	12	9	7	22	50
5	18	10	13	35	76

a. 计算 4 个季度的季节指数。

b. 在什么时候 Costello 音乐公司所经受的季节影响最大？这个季节影响是否合理？请解释。

42. 参考练习 40 中 Hudson Marine 公司的数据。

a. 消除数据的季节影响，并利用消除了季节影响的时间序列来确定趋势。

b. 利用(a)中的结果，建立以趋势为基础的下一年季度预测。

c. 为了说明季节影响，利用练习 40 中的季节指数调节(b)中的预测值。

43. 参考练习 41 中 Costello 音乐公司的时间序列。

a. 消除数据的季节影响，并利用消除了季节影响的时间序列来确定趋势。

b. 利用(a)中的结果，建立以趋势为基础的下一年季度预测。

c. 为了说明季节影响，利用练习 41 中的季节指数调节(b)中的预测值。

案例研究 1　预测食品和饮料的销量

Vintage 饭店位于 Captiva 岛上，该岛靠近佛罗里达州的 Fort Myers，是一个人口稠密的社区。饭店由 Karen Payne 拥有并经营，已经持续了 3 年。在这期间，Karen 一直在寻求以建立精于海鲜的高档正餐的饭店声誉。Karen 和她的员工的努力是成功的，她的饭店已经成为岛上最好的且发展最快的饭店。

Karen 已经作出了关于饭店未来发展的计划，她需要建立一个系统以提前一年预测食品和饮料每月的销量。Karen 拥有 3 年经营中所有的食品和饮料的销售数据，数据如下页所示。

管理报告

对 Vintage 饭店的销量数据进行分析。请为 Karen 准备一份包括你的发现、预测和建议的总结报告。其中包括：

1. 时间序列的图形。

2. 数据的季节分析。计算出每个月的季节指数，并评论每月季节销量影响高低。季节指数是

月份	第一年	第二年	第三年
1月	242	263	282
2月	235	238	255
3月	232	247	265
4月	178	193	205
5月	184	193	210
6月	140	149	160
7月	145	157	166
8月	152	161	174
9月	110	122	126
10月	130	130	148
11月	152	167	173
12月	206	230	235

否有直观上的意义？并对它进行讨论。

3. 预测第四年从1月到12月的销量。
4. 提出建议，说明你的系统在面对新的销售数据时需要在什么时候进行更新。
5. 在你的报告附录中，给出你所分析的详细计算。

假设第四年的1月销量为295 000美元，则你的预测误差是什么？如果这个误差比较大，Karen或许会对你的预测值和实际值之间的差异表示疑惑。你怎样才能消除她关于预测方法的怀疑？

案例研究2　预测销量的损失

Carlson百货商店在2000年8月31日台风袭击时，遭受了巨大损失。商店关闭了四个月(2000年9月—2000年12月)，现在，Carlson和它的保险公司正在就停业期间所遭受的销售额损失而陷入争执。必须解决两个关键性的问题：(1)如果没有被台风袭击，Carlson将会有多少销售额；(2)Carlson是否会从台风过后明显增加的营业中，获得额外的销售额作为补偿。该县得到了超过80亿美元的联邦救灾补偿和保险金，导致百货商店销售额的增加和其他无数商业活动的繁荣。

表18.13给出了Carlson在遭受台风袭击之前48个月的销售额。表18.14给出了该县所有的杂货商店在遭受台风袭击之前48个月的总销售额，同时也给出了Carlson商店停业期间的总销售额。Carlson的管理者已经要求你分析这些数据，并预测Carlson商店在2000年9月—12月停业期间所损失的销售额。他们同时要求你确定在同期是否存在由于台风而产生的额外销售额。如果存在的话，Carlson除了得到正常销售额的补偿外，还将得到额外销售额作为补偿。

表 18.13　1996 年 9 月—2000 年 8 月 Carlson 商店的销售额（单位：百万美元）

月	1996 年	1997 年	1998 年	1999 年	2000 年
1		1.45	2.31	2.31	2.56
2		1.80	1.89	1.99	2.28
3		2.03	2.02	2.42	2.69
4		1.99	2.23	2.45	2.48
5		2.32	2.39	2.57	2.73
6		2.20	2.14	2.42	2.37
7		2.13	2.27	2.40	2.31
8		2.43	2.21	2.50	2.23
9	1.71	1.90	1.89	2.09	
10	1.90	2.13	2.29	2.54	
11	2.74	2.56	2.83	2.97	
12	4.20	4.16	4.04	4.35	

管理报告

请为 Carlson 的管理者准备一份报告，总结你的发现、预测和建议。其中包括：

1. 假如没有台风袭击，请估计销售额。
2. 假如没有台风袭击，请估计全县范围内所有的杂货商店的销售额。
3. 估计 Carlson 商店在 2000 年 9 月—12 月停业期间所损失的销售额。

除此之外，利用该县所有的杂货商店在 2000 年 9 月到 12 月之间的实际销售额和(2)中的估计，作出判断支持或者反对由于台风而产生的额外销售额。

表 18.14　该县所有的杂货商店在 1996 年 9 月—2000 年 12 月间的总销售额(单位：百万美元)

月	1996	1997	1998	1999	2000
1		46.8	46.8	43.8	48.0
2		48.0	48.6	45.6	51.6
3		60.0	59.4	57.6	57.6
4		57.6	58.2	53.4	58.2
5		61.8	60.6	56.4	60.0
6		58.2	55.2	52.8	57.0
7		56.4	51.0	54.0	57.6
8		63.0	58.8	60.6	61.8
9	55.8	57.6	49.8	47.4	69.0
10	56.4	53.4	54.6	54.6	75.0
11	71.4	71.4	65.4	67.8	85.2
12	117.6	114.0	102.0	100.2	121.8

附录 18.1 用 Minitab 进行预测

在本附录中，我们将介绍如何用 Minitab 来进行移动平均法、指数平滑法和趋势推测法预测。

移动平均法


CD 光盘数据
Gasoline

为了说明如何利用 Minitab 来进行移动平均法预测，我们将对表 18.1 和图 18.5 中自行车的销量时间序列进行预测。12 个星期的销量数据已经被输入工作表的第 2 列，以下的步骤可以产生第 13 周 3 个星期的移动平均值预测：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Time Series**；
步骤 3. 选择 **Moving Average**；
步骤 4. 当出现移动平均值对话框时：
在 **Variable** 框中输入 C2；
在 **MA length** 框中输入 3；
选择 **Generate forecasts**；
在 **Number of forecasts** 中输入 1；
在 **Starting from origin** 中输入 12；
点击 **OK**。

第 13 周 3 个星期的移动平均预测值将在任务窗口中显示。均方误差 10.22 在 Minitab 输出结果中被称为 MSD。我们还可以使用许多其他的输出项，包括与表 18.2 相似的总结表，以及和图 18.6 相似的输出结果图形。

指数平滑法


CD 光盘数据
Gasoline

为了说明怎样使用 Minitab 来进行指数平滑预测，我们将利用表 18.1 和图 18.5 中的汽油销量序列来预测第 13 周的销量。将 12 个星期的销量数据输入工作单中的第 2 列。用平滑常数 $\alpha=0.2$，通过以下步骤可以预测第 13 周的销量：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Time Series**；
步骤 3. 选择 **Single Exp Smoothing**；
步骤 4. 当出现 Single Exponential Smoothing 对话框时：
在 **Variable** 框中输入 C2；
对 Weight to Use in Smoothing 选择 **Use** 选项；

在 **Use** 框中输入 0.2；
选择 **Generate forecasts**；
在 **Number of forecasts** 中输入 1；
在 **Starting from origin** 中输入 12；
选择 **Options**；

步骤 5. 当出现 Single Exponential Smoothing 对话框时：
在 **Use average of first** 框中输入 1；
点击 **OK**；

步骤 6. 当出现 Single Exponential Smoothing 对话框时：
点击 **OK**。

第 13 周的指数预测值将在任务窗口中显示。均方误差在 Minitab 输出结果中被称为 MSD。* 我们还可以使用许多其他的输出项，包括与表 18.3 相似的总结表，以及和图 18.7 相似的输出结果图形。

趋势推测法


CD 光盘数据
Bicycle

为了说明怎样使用 Minitab 来进行趋势推测，我们将利用表 18.6 和图 18.8 中的自行车销量序列来预测第 13 周的销量。将年份输入工作单中的第 1 列，将销量数据输入工作单中的第 2 列。用趋势推测法，通过以下步骤可以预测第 13 周的销量：

步骤 1. 选择 **Stat** 下拉菜单；
步骤 2. 选择 **Time Series**；
步骤 3. 选择 **Trend Analysis**；
步骤 4. 当出现 Trend Analysis 对话框时：
在 **Variable** 框中输入 C2；
选择 **Linear** 作为模型类型；
选择 **Generate forecasts**；
在 **Number of forecasts** 中输入 1；
在 **Starting from origin** 中输入 10；
点击 **OK**。

下一期的线性趋势方程和预测值显示在任务窗口中。

---

* 由 Minitab 计算出来的 MSD 值与出现在表 18.4 中的 MSE 值不同。Minitab 用 17 作为第 1 周的预测值，计算 MSD 用的是所有 12 周的资料。在 18.2 节中计算均方误差时，我们只用从第 2 周到 12 周的数据来计算 MSE 值，因为缺乏可以用于预测第 1 期的历史数据。

附录 18.2 用 Excel 进行预测

在本附录中，我们将介绍如何用 Excel 来进行移动平均法、指数平滑法和趋势推测法预测。

移动平均法

为了说明如何利用 Excel 来进行移动平均法预测，我们将对表 18.1 和图 18.5 中汽油的销量时间序列进行预测。12 个星期的销量数据已经被输入工作表 B 列的第 2 到 13 行，以下的步骤可以产生 3 个星期的移动平均值预测：

步骤 1. 选择 **Tools** 下拉菜单；
步骤 2. 选择 **Data Analysis**；
步骤 3. 从 Analysis Tools 列表中选择 **Moving Average**；
点击 **OK**；
步骤 4. 当出现移动平均值对话框时：
在 **Input Range** 框中输入 B2:B13；
在 **Interval** 框中输入 3；
在 **Output Range** 框中输入 C2；
点击 **OK**。

3 周的移动平均值将会出现在工作单的 B 列。如果在 **Interval** 框中输入不同的数值可以很轻松的计算出其他步长的预测。

指数平滑法


CD 光盘数据
Gasoline

为了说明怎样使用 Excel 来进行指数平滑预测，我们将再次利用表 18.1 和图 18.5 中的汽油销量序列来进行预测。将 12 个星期的销量数据输入工作表 B 列的第 2 到 13 行。用平滑常数 $\alpha = 0.2$，通过以下步骤可以进行预测：

步骤 1. 选择 **Tools** 下拉菜单；
步骤 2. 选择 **Data Analysis**；
步骤 3. 从 Analysis Tools 列表中选择 **Exponential Smoothing**；
点击 **OK**；
步骤 4. 当出现 Exponential Smoothing 对话框时：
在 **Input Range** 框中输入 B2:B13；
在 **Damping factor** 框中输入 0.8；
在 **Output Range** 框中输入 C2；
点击 **OK**。

指数预测值将在工作单的 B 列中显示。请注意我们在 Damping factor 对话框中输入的是 $1-\alpha$；通过在 Damping factor 对话框中输入不同的 $1-\alpha$ 数值，我们可以轻松地计算出其他平滑常数的预测值。

趋势推测法



CD 光盘数据
Bicycle

为了说明怎样使用 Excel 来进行趋势推测，我们将利用表 18.6 和图 18.8 中的自行车销量序列来进行预测。在第 1 行以顺序标号，在工作表 A 列和 B 列中的第 1 行至第 11 行输入数据。用趋势推测法，通过以下步骤可以预测第 11 年的销量：

步骤 1. 在工作表中选择一个空单元格；

步骤 2. 选择“**Insert**”下拉菜单；

步骤 3. 选择“**Function**”；

步骤 4. 当出现 Paste Function 对话框时：
在 **Function Category** 框中选择“**Statistical**”；
在 **Function Name** 框中选择“**Forecast**”；
点击“**OK**”；

步骤 5. 当出现 Forecast dialog 对话框时：
在 **x** 框中输入 11；
在 **Known y's** 框中输入 B2：B11；
在 **Known x's** 框中输入 A2：A11；
点击“**OK**”。

第 11 年的预测值 32.5 将出现在第 1 步所选择的单元格中。

19 非参数方法

统计实例：西壳房地产经纪公司

19.1 符号检验

小样本情形

大样本情形

中位数假设检验

19.2 维尔克科森符号秩检验

19.3 曼－怀特尼－维尔克科森检验

小样本情形

大样本情形

19.4 克鲁斯卡尔－沃利斯检验

19.5 秩相关

秩相关显著性检验

统计实例

西壳房地产经纪公司*

辛辛那提，俄亥俄州

西壳房地产经纪公司(West Shell Realtors)成立于1958年，刚成立时只有一间办公室和三个销售人员。在1964年，公司开始走上长期扩张的发展道路，新的办事处每年都有增长。几年以后，现在的西壳公司已经是大辛辛那提地区最大的房地产经纪公司，在俄亥俄州西南部、印第安纳州东南部、肯塔基州北部设有办事处。

统计分析可以帮助诸如西壳这样的房地产经纪公司对自己的销售业绩进行监控。西壳公司所有办事处和总公司每个月都编制业绩报表，汇报有关销售总额、已售单元数量、每个单元平均售价的统计总结数据，这些数据对办事处经理和总公司高层领导把握公司的经营过程和难点所在是至关重要的。

除了对每月业绩进行总结外，西壳公司还利用统计来指导公司的发展计划和战略。西壳公司正在执行一项有计划的扩张战略，每当扩张计划需要公司成立新的销售办事处时，办事处的选址问题就会成为公司关心的焦点。房产的销售价格、资金的周转率、销量的预测都是在评价和选择办事处地点时需要考虑的重要数据。

例如，西壳公司选择了两个郊区：克利福顿和罗斯莱尔，计划将这两个地区作为一个新的办事处的首选地点。在对这两个地区做比较时需要考虑很多不同的因素，包括房产的售价。西壳公司利用小样本条件下的非参数方法来确定这两个地区销售情况的差异之处。



西壳房地产经纪公司利用统计手段对所销售的住宅数量进行分析，以确保其竞争力。© PhotoDisc, Inc.

他们从克利福顿选取了25笔销售业务作为样本，从罗斯莱尔选取了18笔业务作为样本，然后选择曼－怀特尼－维尔克科森检验对销售价格差异进行统计检验。在0.05的显著性水平下，曼－怀特尼－维尔克科森检验并没有拒绝这两个地区销售价格基本相同这一零假设。因此，西壳公司在地点选取过程中可以集中考察房产销售价格以外的选择标准。

在这一章中，我们将讨论诸如曼－怀特尼－维尔克科森检验之类的非参数检验是怎么应用的，同时，我们也将讨论怎样对这些检验作出正确的解释。

*作者衷心感谢西壳公司的Rodney Fightmaster提供这一统计实例。

到目前为止，本书所讨论过的统计方法都是通常所知的**参数方法**。在这一章中，我们将介绍几种**非参数方法**(nonparametric methods)。如果某些情形运用做前面章节中所介绍的参数方法不适用的话，则可用本章中的非参数方法。通常说来，非参数方法对数据测量的水平并没有很严格的假设，对样本数据所服从的概率分布形式假设也比较少。

选择应用参数方法还是非参数方法的一个考虑因素是数据生成的测量尺度。所有的数据都是由以下四种生成尺度之一生成的：名义、序数、区间和比例。因此，所有的统计分析都是根据这四种类型的数据：名义、序数、区间和比例进行的。

下面，让我们对这四种类型的测量尺度进行定义，并给出示例。

1. **名义尺度**。如果数据是简单的符号或者类别，而且这个数据是用于定义某个元素的属性，则其测量尺度是名义尺度，名义数据可以是数值，也可以是非数值的。

   举例：一只股票所在的上市交易所(纽约证券交易所、纳斯达克或者美国证券交易所)是非数值的名义数据。每个公民的社会保险号码是一个数值的名义数据。

2. **序数尺度**。如果数据用于对观测值进行排秩或者排序，则这种测量尺度是序数尺度，序数数据既可以是数值的，也可以是非数值的。

在第 1 章中，我们指出了名义尺度和序数尺度给出的是定性的数据，区间尺度和比例尺度给出的是定量数据。

   举例：在对某个个体项目大小进行测量时，可以将其分为小、中、大三种类型，这就是非数值的序数数据。个体级别秩为 1，2，3……这些就是数值序数数据。

3. **区间尺度**。如果一组数据具有序数数据的特征，并且数据观测区间可以用固定的测量尺度单位来表示，则这种测量尺度是区间尺度。区间数据必须是数值的。

   举例：温度的测量就是区间数据。假设某个地点是 70°，另外一个地点是 40°，我们可以根据温度对这两个地点进行排序；第一个地点比第二个地点更加温暖。测量尺度的固定单位——度，使得我们可以确定第二个地点比第一个地点暖和 30°。

4. **比例尺度**。如果数据具备区间数据的特征，而且测量尺度的比例是有意义的，则测量尺度是比例尺度。比例数据必须是数值的。

   举例：距离、高度、重量以及时间这样的变量都是以比例尺度的形式测量的。温度测量不是以比例尺度的形式进行，那是因为不存在固有的零点。例如，水的冰点是 32℉，即 0℃。温度数据的比值是没有任何意义的。例如，认为 80°的温度比 40°的温度暖和两倍，这种说法是没有意义的。

如果数据测量尺度是名义或者序数尺度，对方差、均值和标准差的计算都是没有意义的。因此，对于这种类型的数据，前面所讨论过的很多方法都是无法应用的。

大部分参数统计方法都需要利用区间尺度或者比例尺度的数据。在这些计量水平上进行数学运算是比较有意义的，而且均值、方差和标准差等等都是可以计算、解释并在分析中得到应用的。但是对名义或者序数尺度的数据计算，均值、方差和标准差都是不适用的。因此，通常情况下不能使用参数方法。非参数方法通常是对这些数据进行分析的惟一的方法，也是由此得出统计结论的惟一方法。

第 8、9、10 章所讨论过的以小样本和 t 分布为基础的统计方法都是根据数据来自正态总体这一假设。如果这个假设不适用，则应该使用非参数方法。

在确定是使用参数方法还是非参数方法的时候，另外一种考虑是关于数据来源的总体样本假设。举个例子，在第 10 章中，我们给出了一个对两个总体均值之间差异的假设进行检验的参数方法。在小样本情形中，如果我们能够假设这两个总体是方差相同的正态分布，则我们可以使用 t 分布进行检验。如果总体方差相同这一假设并不适用，则即使数据是区间尺度的或者比例尺度的，我们也不能利用以 t 分布为基础的参数方法进行检验。但是，由于非参数方法对总体概率分布的假设没有要求，因此可以用它对两个总体间的差异进行检验。正因为非参数方法对概率分布假设没有要求，因此常常被称为**自由分布检验**(distribution-free methods)。

总的说来，如果一种统计方法被归于非参数方法，则它至少满足以下条件中的一种：

1. 这种方法可以用于名义数据。
2. 这种方法可以用于序数数据。
3. 如果无法对总体概率分布作出假设时，这种方法可以用于区间或者比例数据。

如果数据测量水平是区间尺度或者比例尺度，并且总体概率分布的必要假设是合适的，则参数方法所提供的统计过程更为有效，区分识别能力也更强。在许多情形中，参数方法和非参数方法都是适用的，非参数方法和参数方法一样有效。对于某些情形，如果数据尺度类型为名义或者序数尺度，或者参数方法所必需的假设不成立，则我们只能利用非参数方法。由于非参数方法所需要的数据测量尺度要求并不是很严格，对总体分布的假设也比较宽松，因此我们认为通常情况下非参数方法比参数方法应用更为广泛。符号检验、维尔克科森符号秩检验、曼 - 怀特尼 - 维尔克科森检验、克鲁斯卡尔 - 沃利斯检验以及斯皮尔曼秩相关就是本章中所要阐述的非参数方法。

19.1 符号检验

在通常的市场调查研究中，**符号检验**(sign test)的应用涉及到由 n 名潜在顾客组成的样本的应用，由这 n 名顾客根据自己的偏好对某类产品的两种品牌进行选择，比如咖啡、软饮料或者清洁剂等等。由于顾客只需要指出或者标出所偏好的品牌，所以这 n 个选择是名义尺度的数据。利用这些数据，我们可以确定人们是否对所比较的两种品牌存在偏好上的差异。正如我们所看到的，符号检验是回答此类问题的非参数统计方法。

小样本情形

符号检验的小样本情形指的是 $n \leqslant 20$。让我们对太阳海岸农场进行一项研究，以说明小样本

* 见 W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. (John Wiley & Sons, 1998)。

情形符号检验的应用。太阳海岸农场生产一种 Citrus Valley 品牌的桔子汁。它的一个竞争对手已经开始生产一种新的桔子汁品牌：Tropical Orange。在对这两种品牌进行消费者偏好的研究中，12 名消费者所面对的产品都没有标出品牌，而且每一名消费者第一次品尝的产品品牌都是随机选择的。在消费者对这两种品牌进行品尝后，研究者要求消费者标明自己所偏好的产品。这项研究的目的在于确定消费者是否更加倾向于使用其中某一种品牌的产品。令 p 表示偏好 Citrus Valley 的消费者所占总体的比例，我们计划检验以下的假设：

$$H_0: \ p = 0.50$$
$$H_a: \ p \neq 0.50$$

如果我们不能拒绝 H_0，则对于这两种品牌的产品，没有证据表明存在消费者偏好的差异。但是，如果 H_0 被拒绝，我们就可以得出结论：这两种品牌存在消费者偏好差异。在这种情形下，由多数消费者所挑选出来的品牌可以被认为是消费者更加偏好的品牌。

在接下来的讨论中，我们将阐述如何使用小样本情形的符号检验来对关于消费者偏好的假设进行检验并得出结论。为了记录参加研究的 12 名消费者的偏好数据，我们作出如下规定：如果消费者标明所偏好的品牌是 Citrus Valley，则我们使用加号，如果消费者偏好的品牌是 Tropical Orange，则我们使用减号。由于数据是以加、减符号的形式表示，所以这种非参数检验被称为符号检验。

如果样本容量小于或等于 20，在附录 B 的表 5 中可以查到具体的二项概率。

表 19.1　$n = 12$，$p = 0.50$ 的二项概率分布

加号的个数	概率
0	0.000 2
1	0.002 9
2	0.016 1
3	0.053 7
4	0.120 8
5	0.193 4
6	0.225 6
7	0.193 4
8	0.120 8
9	0.053 7
10	0.016 1
11	0.002 9
12	0.000 2

在假设 H_0 为真的情况下(即 $p = 0.50$)，加号的个数服从 $p = 0.50$ 的二项概率分布。当样本容量 $n = 12$ 时，附录 B 的表 5 给出了如表 19.1 所示的 $p = 0.50$ 的二项分布的概率值。图 19.1 是这个二项概率分布的图形表示。在 H_0 假设成立的条件下，它表示了加号个数的概率，因此它也是假设检验的合适的抽样分布。我们利用这个抽样分布来确定一个拒绝 H_0 的规则；所使用的方法和我们在第 9 章中用于建立拒绝假设检验的规则的方法是相似的。举个例子，利用 $\alpha = 0.05$ 显著性水平，我们可以将图 19.1 中分布两端约 0.025 的部分定义为拒绝区间或者拒绝域。从这个分布的低端开始，我们发现加号个数为 0，1 或者 2 的概率为 0.000 2 + 0.002 9 + 0.016 1 = 0.019 2。请注意，我们在两个加号处停住，是因为如果加上 3 个加号的概率将导致这个低端区域等于 0.019 2 + 0.053 7 = 0.072 9，这个数值将明显超过了所要求的 0.025 区域。在这个分布的高端，我们发现相对应于 10、11、12 个加号的概率也是 0.019 2。因此，我们离 $\alpha = 0.05$ 最为靠近而又没有超出的概率即为 0.019 2 + 0.019 2 = 0.038 4。在这种情况下，我们可以采用如下的拒绝法则：

如果加号个数少于 3 个或者多于 9 个，则我们可以拒绝 H_0。

图 19.1 不能显示阴影区域，因为抽样分布是间断的，拒绝域即 0，1，2，10，11，12 处的点。

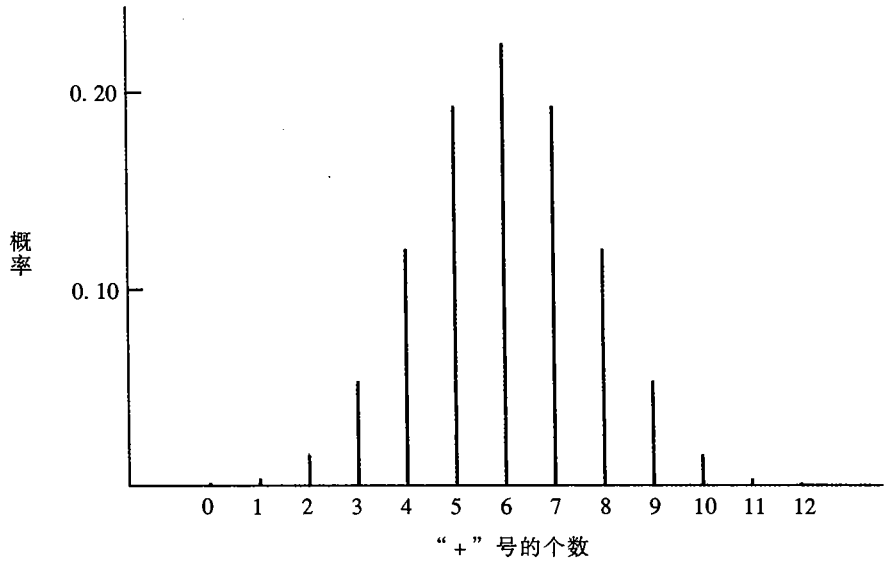


图 19.1　$n=12$，$p=0.50$ 的加号二项概率分布示意图

表 19.2 显示的是阳光海岸农场例子中反映消费者偏好的数据。由于仅仅观测到两个加号，所以我们可以拒绝零假设。这项研究证明了消费者对这两种品牌的桔子汁的偏好是存在差异的。我们有必要向阳光海岸农场提出建议，因为研究显示消费者更加倾向于消费其竞争对手的 Tropical Orange 品牌。

表 19.2　阳光海岸农场桔子汁风味检验中的消费者倾向性数据

消费者	品牌倾向	记录数据
1	Tropical Orange	−
2	Tropical Orange	−
3	Citrus Valley	+
4	Tropical Orange	−
5	Tropical Orange	−
6	Tropical Orange	−
7	Tropical Orange	−
8	Tropical Orange	−
9	Citrus Valley	+
10	Tropical Orange	−
11	Tropical Orange	−
12	Tropical Orange	−

对于非参数检验，如果 p－值$<\alpha$ 我们也可以拒绝 H_0。在有两个加号的情况下，这个双边检验的 p－值为 $2(0.016\,1+0.002\,9+0.000\,2)=0.038\,4$。对于单边检验，$p$－值是单边检验抽样分

布概率的总和。

在阳光农场示例中，进入调查中的所有 12 个消费者需要表明自己对产品的偏好。但是在很多情形下，样本中的一个或者多个消费者有可能无法明确表明对某一种品牌的偏好。在这种情况下，应该将这些没有明确喜好的消费者个体从总体中删除，分析过程可以在一个较小的样本中进行。

附录 B 表 5 所示的二项概率分布可以用于构造样本容量在 $n=20$ 以内的任何符号检验的决策规则。对于零假设 $p=0.50$ 和样本容量 n，我们可以建立对应于任何显著性水平的决策规则。而且，如果我们仅仅考虑二项概率分布的低端或者高端，则我们可以建立单边检验的拒绝规则。附录 B 没有给出样本容量大于 20 的二项概率分布表，在这种情形下，我们可以使用二项概率分布的大样本正态近似情形来确定适当的符号检验拒绝规则。

大样本情形

大样本符号检验和第 9 章中所讨论的 $p=0.5$ 总体比例的检验是相同的。

如果利用零假设 H_0：$p=0.50$ 和样本容量 $n>20$，则我们可以利用正态概率分布来近似加号个数的抽样分布。

如果不存在倾向性偏好，加号个数抽样分布的正态近似情形

$$\text{均值：}\ \mu=0.50n \tag{19.1}$$

$$\text{标准差：}\ \sigma=\sqrt{0.25n} \tag{19.2}$$

分布形式：当 $n>20$ 时，近似于正态分布。

让我们考虑将符号检验应用于政治民意测验。近期的总统选举进行了一项民意测验，调查询问了 200 名已登记的选民，调查的内容是关于民主党和共和党的总体对外政策优劣的问题。调查结果表明：有 72 人认为民主党候选人更好，有 103 人认为共和党候选人更好，25 人认为两党并无多大差别。这项民意测验是否表明了公众对民主党和共和党的对外政策存在显著的倾向性差异？

通过分析提炼出的问题要相应地予以解决。

利用符号检验，我们可以看到有 $n=200-25=175$ 人明确表明了他们对总体外交政策的倾向性态度。利用公式(19.1)和(19.2)，我们发现加号个数的抽样分布具有如下的特征：

$$\mu=0.50n=0.50(175)=87.5$$

$$\sigma=\sqrt{0.25n}=\sqrt{0.25(175)}=6.6$$

除此之外，当 $n=175$ 时我们可以假设抽样分布近似于正态分布。这个分布如图 19.2 所示。由于

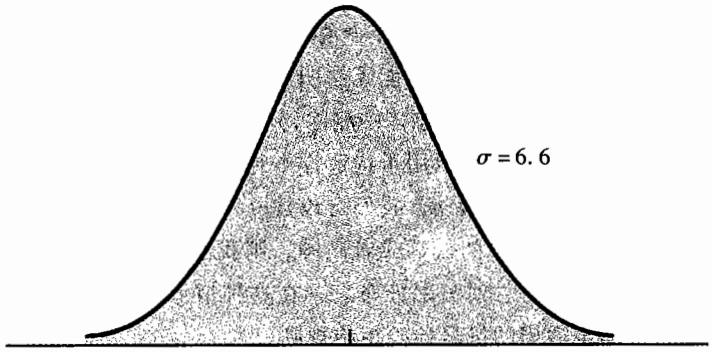


图 19.2　$n = 175$ 的符号检验中加号个数的概率分布

这个分布近似于正态分布，我们可以利用标准正态分布概率分布面积表来构造这个检验的拒绝法则。在 $\alpha = 0.05$ 显著性水平下，这个双边检验的拒绝法则如下：

如果 $z < -1.96$ 或者 $z > +1.96$，则拒绝 H_0。

将民主党候选人对外政策得到较高评价的次数作为加号的个数（$x = 72$），我们可以得到如下的检验统计量：

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 87.5}{6.6} = -2.35$$

如果分析师用对共和党候选人有较高评价的次数作为加号个数，$z = 2.35$ 也会使我们导出同样的结论。

因为 $z = -2.35$ 小于 -1.96，在 $\alpha = 0.05$ 显著性水平下，可以拒绝两党候选人外交政策没有差异的假设。在 $z = -2.35$ 情况下，我们可以用标准正态分布来表示 p - 值为 $2(0.500\,0 - 0.490\,6) = 0.018\,8$。这项研究表明公众对候选人的外交政策评价是有差异的。

中位数假设检验

在第 9 章中，我们描述了怎样利用假设检验来推断总体均值。现在，我们将阐述怎样用符号检验来对总体中位数进行假设检验。回顾前面的内容，中位数是将总体分成两部分，其中 50% 位于中位数之上，50% 位于中位数之下。我们可以利用符号检验，当样本中的数据大于所假设的中位数时，我们用加号来表示；当样本中的数据小于所假设的中位数时，我们用减号来表示。如果数据恰好等于中位数，则我们应该将它从样本中删除。符号检验的计算过程与前面所述的计算过程是完全一样的。

举个例子，关于密苏里州路易斯市的新建住宅楼的价格中位数的假设检验如下：

H_0：中位数 = 130 000 美元

H_a：中位数 ≠ 130 000 美元

某个样本包括 62 所房子，34 所房子价格在 130 000 美元之上，26 所价格在 130 000 美元之下，还有两所房子价格正好是 130 000 美元。

对于 $n=60$ 所，价格不为 130 000 美元的房子，利用公式(19.1)和(19.2)，我们可以得到

$$\mu = 0.50n = 0.50(60) = 30$$

$$\sigma = \sqrt{0.25n} = \sqrt{0.25(60)} = 3.87$$

由于加号个数 $x=34$，检验统计量为

$$z = \frac{x-\mu}{\sigma} = \frac{34-30}{3.87} = 1.03$$

利用双边检验和 $\alpha=0.05$ 显著性水平，如果 z 小于 -1.96 或者 z 大于 $+1.96$，则我们可以拒绝 H_0。检验统计量 $z=1.03$，因此我们不能拒绝 H_0。p - 值为 2(0.500 0 - 0.348 5) = 0.303。根据这些数据，我们无法拒绝路易斯新建住宅销售价格的中位数为 130 000 美元的零假设。

评　注

我们在计算中利用加号的个数来确定是否拒绝 p - 值为 0.5 的零假设。同样，我们也可以利用减号的个数进行计算，所得结果相同。

练　习

方　法

1. 下表所列的数据是关于 10 个消费者个体对于某产品两种品牌的风味检验结果：



消费者	品牌 A 比之品牌 B	消费者	品牌 A 比之品牌 B
1	+	6	+
2	+	7	−
3	+	8	+
4	−	9	−
5	+	10	+

在 $\alpha=0.05$ 的显著性水平下，对两种品牌的消费者偏好差异进行显著性检验。加号表

示消费者倾向于品牌 A。

2. 所做的假设检验如下所示：



H_0：中位数≤150

H_a：中位数>150

样本容量为 30，数值在 150 以上的有 22 例，正好等于 150 的有 3 例，小于 150 的有 5 例，利用 $\alpha=0.01$ 的显著性水平进行假设检验。

应　用

3. 股票分拆是否有利于股东？SNL 证券公司对银行业 18 个月来的股票分拆进行了一项研究，发现股票分拆会导致个人持股者的股票价值上升。假设某个样本由 20 个最近的股票分拆案例组成，其中，有 14 个案例导致股票价值上升，4 个案例导致股票价值下跌，2 个没有发生任何变化。利用符号检验确定股票分拆是否一直对银行股票持有者有利。

a. 零假设和备择假设是什么？

b. 在 $\alpha=0.05$ 的显著性水平下，给出拒绝法则。

c. 能得出什么结论？

4. 一项民意测验调查了 1 253 名成年人，就国家的经济状况和他们的孩子的未来进行了访谈。其中一个问题是这样的："你认为你的孩子能够过上比你现在更好的生活，还是更差的生活，或者就跟你现在所过的生活一样。"回答是：34% 的人认为会更好，29% 的人认为会更差，33% 的人认为会一样，4% 的人不是很确定。利用符号检验和 $\alpha=0.05$ 的显著性水平，确定相信他们的孩子会过上比他们自己更好的生活的成年人数目比认为孩子生活会更差的成年人数目要多，你的结论是什么？



5. 1996—1997 年的尼尔森媒介调查确证 *ER* 和 *Seinfeld* 是两个收视率最高的网络电视节目。假设在一项本地电视收视倾向调查中，要求 410 个被调查者说出他们最喜欢电视节目，其中 185 个被调查者选择了 *ER*，165 个被调查者选择了 *Seinfeld*，60 个选择了其他电视节目。利用 $\alpha=0.05$ 的显著性水平对 *ER* 和 *Seinfeld* 是否拥有同样的收视率进行检验，你能得出什么结论？

6. 在 1996 年，个人电脑的消费市场由惠普和康柏领导（*USA Today*，July 21，1997）。在一个由 500 名电脑采购者组成的样本中，202 名采购者购买的是惠普电脑，158 名采购者购买的是康柏电脑，140 名购买的是其他品牌的电脑，比如 IBM、苹果和 NEC。在 $\alpha=0.05$ 的显著性水平下，对惠普和康柏是否占有同样的市场进行检验，你的结论是什么？

7. Barron's 杂志的订阅者收入中位数为 131 000 美元（*www. barronsmag. com*，July 28，2000）。假设在一个由 300 名《华尔街日报》订阅者所组成的样本中，165 名订阅者的收入中位数在 131 000 美元之上，135 名订阅者的收入中位数在 131 000 美元之下。你能否得出结论：这两种杂志的订阅者群体收入中位数之间存在差异？p－值为多少？在 $\alpha=0.05$ 的显著性水平下，你的结论

是什么？

8. 在一个由 150 场大学篮球赛所组成的样本中，主队获胜 98 场比赛，利用 $\alpha=0.05$ 的显著性水平，对比赛数据进行检验，确定主队在大学篮球比赛中是否占有优势，你能得出什么结论？

9. 去年某个城市每个快餐店中的兼职雇员人数中位数为 15，市政府的官员认为兼职雇员的数目可能会增长。由 9 个快餐店所组成的样本显示，7 个快餐店有多于 15 个的兼职雇员在其中工作，有一个快餐店刚好有 15 个兼职雇员工作，还有一个快餐店兼职雇员人数少于 15 个。在 $\alpha=0.05$ 的显著性水平下，检测兼职雇员人数的中位数是否有所增加。

10. 根据一项全国调查，成年人所希望的能够使他们的梦想成真的年收入中位数为 152 000 美元。假设某个样本由俄亥俄州 225 个被调查者组成，122 名被调查者认为可以使他们梦想成真的年收入中位数在 152 000 美元之下，103 名被调查者认为必须在 152 000 美元之上。在 $\alpha=0.05$ 的显著性水平下，检验零假设：能够使俄亥俄州人们的梦想成真的年收入中位数为 152 000 美元，你能得出什么结论？

11. 美国劳工统计局的报告表明 1996 年在行政和管理岗位的妇女每周收入的中位数为 585 美元，一个由芝加哥地区行政和管理岗位的妇女所组成的样本提供了如下的数据。利用这个样本检验假设 H_0：中位数≤585，H_a：中位数>585，这个中位数指的是芝加哥行政和管理岗位妇女的每周收入。利用 $\alpha=0.05$ 的显著性水平，你能得出什么结论？



CD 光盘数据
Chicago

622	516	631	498	715
571	494	525	664	721
657	692	551	580	649
706	597	518	725	635
548	604	671	607	487
583	702	622	714	693
600	721	662	633	681
624	551	632	544	485
655	721	669	677	609
656	562	721	489	582

19.2　维尔克科森符号秩检验

维尔克科森符号秩检验(Wilcoxon signed-rank test)是第 10 章中所阐述的匹配样本参数检验的非参数替代方法。在匹配样本情况下，每个试验单元将会产生两个成对或者相匹配的观测值，其中一个来自总体 1，另一个来自总体 2。相匹配的观测值之间所存在的差异使得我们可以考察两个总体之间的差异。

匹配样本的参数分析方法(配对差异的 t 检验)需要区间数据并假设成对观测值的差异总体服从**正态分布**。在这个假定下，我们可以采用 t 分布来检验总体均值间不存在差异的零假设。如果差异服从正态分布的假设不成立，则我们可以使用维尔克科森符号秩检验。我们可以通过对两种不同生产方法的有效性进行比较来描述这种非参数检验。

某个制造企业正在尝试确定两种生产方法在任务完成时间上存在的差异。选择了 11 个工人组成一个样本，每一个工人均利用每一种方法完成生产任务。每一个工人首先使用的生产方法都是随机挑选的。因此，样本中的每一个工人均提供了一对观测值，如表 19.3 所示。任务完成时间的正差异表明方法 1 需要更多的时间，任务完成时间的负差异则表明方法 2 需要更多的时间。这些数据是否表示这些方法在任务完成时间上的差异是显著的?

表 19.3　生产任务完成时间(分钟)

	方法		
工人	1	2	差异
1	10.2	9.5	0.7
2	9.6	9.8	-0.2
3	9.2	8.8	0.4
4	10.6	10.1	0.5
5	9.9	10.3	-0.4
6	10.2	9.3	0.9
7	10.6	10.5	0.1
8	10.0	10.0	0.0
9	11.2	10.6	0.6
10	10.7	10.2	0.5
11	10.6	9.8	0.8

实际上，我们有两个任务完成时间的总体，每一个总体都对应于一种方法。我们可以对下面的检验进行检测：

H_0：总体相同

H_a：总体不相同

如果无法拒绝 H_0，则我们将没有证据推断两种方法的任务完成时间存在差异。但是，如果 H_0 被拒绝，则我们可以得出结论，这两种方法在任务完成时间上是存在差异的。

维尔克科森符号秩检验的第一步需要把两种方法差异的绝对值进行排序。我们将零差异剔除，然后将剩下的差异绝对值由低到高进行排序，相同的差异值将被赋予它们在混合数据集合中位置的平均秩。差异绝对值的排序情况如表 19.4 第 4 列所示。请注意工人 8 的零差异已经从排序中删除。因此，最小的差异绝对值 0.1 的秩被定为 1。这种差异绝对值的排序将一直持续到最大绝对值 0.9 的秩被赋予值 10 为止。工人 3 和工人 5 的相同绝对差异的秩被定义为平均值 3.5，与此同时，工人 4 和工人 10 的相同绝对差异的秩被定义为平均秩 5.5。

表 19.4　生产任务完成时间例子的差异绝对值排序

工人	差异	差异绝对值	秩	符号秩
1	0.7	0.7	8	+ 8
2	-0.2	0.2	2	- 2*
3	0.4	0.4	3.5	+ 3.5
4	0.5	0.5	5.5	+ 5.5
5	-0.4	0.4	3.5	- 3.5
6	0.9	0.9	10	+10
7	0.1	0.1	1	+ 1
8	0.0	0.0	—	—
9	0.6	0.6	7	+ 7
10	0.5	0.5	5.5	+ 5.5
11	0.8	0.8	9	+ 9
			符号秩的和	+44.0

一旦差异绝对值的秩已经确定，这些秩将被赋予数据中原始差异的符号。举个例子，工人 7 的 0.1 差异，其秩为 1，由于两种方法的观测值差异为正，所以这个秩被赋予值 +1。对于工人 2，由于两种方法之间的观测值差异为负，所以 0.2 的差异值，其秩为 2，被赋予值 -2。表 19.4 最后一列给出了完整的符号秩以及它们的和。

让我们回顾关于两种生产方法任务完成时间相同的原始假设。如果总体表明每一种方法任务完成时间是相同的，则我们可以认为正秩和负秩将会相互抵消，因此符号秩之和的数值将接近于零。于是，维尔克科森符号秩检验的显著性检验将会包括确定符号秩的总和(在我们的例子中是 +44)是否显著得不等于零。

用 T 表示维尔克科森符号秩检验中符号秩的总和。如果两个总体相同，并且相匹配的数据对个数为 10 或者更多时，T 的抽样分布可以用如下的正态概率分布来近似地表示：

总体相同的 T 抽样分布

均值：$\mu_T = 0$　　(19.3)

标准差：$\sigma_T = \sqrt{\dfrac{n(n+1)(2n+1)}{6}}$　　(19.4)

分布形式：当 $n \geqslant 10$ 时，近似于正态分布。

对于这个例子，我们在删除了差异为零的观测值(工人 8)以后得到 $n = 10$。因此，利用公式(19.4)我们有

$$\sigma_T = \sqrt{\frac{10(11)(21)}{6}} = 19.62$$

图 19.3 是在总体相同的假设下 T 的抽样分布。检验统计量 z 的数值为

* 编者按：此处原书为 +2，疑误。

$$z=\frac{T-\mu_T}{\sigma_T}=\frac{44-0}{19.62}=2.24$$

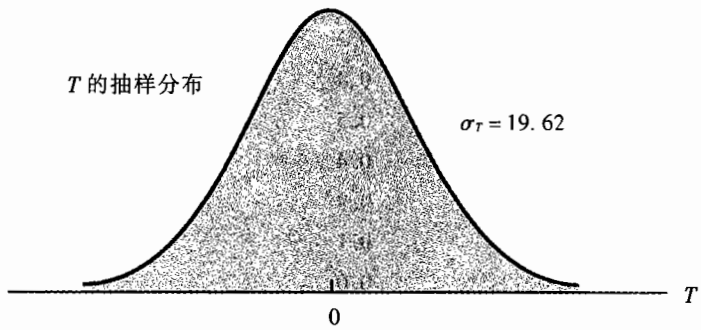


图 19.3 生产任务完成时间例子中的维尔克科森统计量 T 的抽样分布

利用 $\alpha=0.05$ 的显著性水平对无差异零假设进行检验。如果 $z<-1.96$ 或者 $z>1.96$，则我们可以拒绝 H_0。根据 $z=2.24$，我们可以拒绝 H_0 并得出结论：两个总体不相同，而且两种方法在任务完成时间上是存在差异的。利用 $z=2.24$，可以得出 p - 值为 $2(0.500\,0-0.487\,5)=0.025$。由于 11 个工人中有 8 个工人采用第二种方法较快地完成了工作任务，这一事实可以使我们得出结论：两个总体间的差异表明第二种方法是更为高效的生产方法。

练 习

应 用

12. 自测题 对两种燃料添加剂进行检验，以确定它们对客车每加仑行驶里数的影响。以下是对 12 辆车的检验结果。每一辆车都使用两种燃料添加剂进行检验。利用 $\alpha=0.05$ 的显著性水平和维尔克科森符号秩检验确定两种燃料添加剂之间是否存在显著差异。

	燃料添加剂			燃料添加剂	
客车	1	2	客车	1	2
1	20.12	18.05	7	16.16	17.20
2	23.56	21.77	8	18.55	14.98
3	22.03	22.57	9	21.87	20.03
4	19.15	17.06	10	24.23	21.15
5	21.23	21.22	11	23.21	22.78
6	24.77	23.80	12	25.02	23.70

13. 自测题

一项调查对男人入睡所需要的放松休息时间进行研究，随机挑选了 10 个人组成样本。10 个个体的数据显示了在有无放松条件下的入睡所需分钟数，使用 0.05 的显著性水平来确定放松是否会缩短入睡所需时间，你能得出什么结论？

调查对象	无放松	有放松	调查对象	无放松	有放松
1	15	10	6	7	5
2	12	10	7	8	10
3	22	12	8	10	7
4	8	11	9	14	11
5	10	9	10	9	6

14. 在 10 个主要的机场抽取了关于汽车每加仑汽油的价格数据。Avis 和 Budget 汽车租赁公司的数据如下所示（*USA Today*，April 4，2000）：

机场	Avis 汽车租赁公司	Budget 汽车租赁公司
Boston Logan	1.58	1.39
Chicago O'Hare	1.60	1.55
Chicago Midway	1.53	1.55
Denver	1.55	1.51
Fort Lauderdale	1.57	1.58
Los Angeles	1.80	1.74
Miami	1.62	1.60
New York(JFK)	1.69	1.60
Orange County, CA	1.75	1.59
Washington(Dulles)	1.55	1.54

在 $\alpha=0.05$ 的显著性水平下，对两个总体无差异的假设进行检验，你能得出什么结论？

15. 下面是一项有关两种隔夜邮件速递服务的检验，我们对同一个邮件都建立了两个相同的样本，以便评判这两种服务对同一个邮件的速递情况。每一次投递所需的时间如下所示。下面所示的数据是否表明这两种邮递服务存在差异?在 $\alpha=0.05$ 的显著性水平下，对此进行检验。

邮件	邮递服务	
	1	2
1	24.5	28.0
2	26.0	25.5
3	28.0	32.0
4	21.0	20.0
5	18.0	19.5
6	36.0	28.0

（续表）

邮件	邮递服务	
	1	2
7	25.0	29.0
8	21.0	22.0
9	24.0	23.5
10	26.0	29.5
11	31.0	30.0

16. 1997 年 12 种股票的市盈率如下表所示(*Barron's*, December 8, 1997)。假设财务分析给出了 1998 年的市盈率的估计值。在 $\alpha = 0.05$ 的显著性水平下，就 1997 年和 1998 年这两年的市盈率差异你能够得出什么结论？

股票	1997 年市盈率	1998 年市盈率
Coca-Cola	40	32
Du Pont	24	22
Eastman Kodak	21	23
General Electric	30	23
General Mills	25	19
IBM	19	19
McDonald's	20	17
Merck	29	19
Motorola	35	20
Philip Morris	17	18
Walt Disney	33	27
Xerox	20	16

17. 在一项为评价某一特定广告方案的有效性而设计的市场调查中，调查者挑选了 10 个城市的市场组成样本。每个城市在广告方案投入之前一周的销量(千美元)都已经被记录下来，然后收集了广告方案投入之后的每周销量(千美元)，两个销量(千美元)数据集合如下所示：

城市	广告活动前的销量	广告活动后的销量
Kansas City	130	160
Dayton	100	105
Cincinnati	120	140
Columbus	95	90
Cleveland	140	130
Indianapolis	80	82
Louisville	65	55
St. Louis	90	105
Pittsburgh	140	152
Peoria	125	140

在 $\alpha=0.05$ 的显著性水平下，关于这个广告方案的价值你能够得出什么结论？

19.3　曼 – 怀特尼 – 维尔克科森检验

在本节中，我们将讨论另一种用于确定两个主体间是否存在差异的非参数方法。这种检验和符号秩检验不同，它并不是基于匹配样本的检验。它所使用的两个独立样本来自不同的总体。由于这个检验是由曼、怀特尼和维尔克科森合作提出来的，因而它有时被称为曼 – 怀特尼检验，有时被称为维尔克科森秩和检验。这个检验的曼 – 怀特尼版本和维尔克科森版本是相同的；我们倾向于把这个检验称为**曼 – 怀特尼 – 维尔克科森检验**(Mann – Whitney – Wilcoxon test, MWW)。

回顾第 10 章，我们对两个总体间的均值差异进行了参数检验。我们曾经对下面的假设进行了检验：

$$H_0:\ \mu_1-\mu_2=0$$
$$H_a:\ \mu_1-\mu_2\neq 0$$

在小样本情形下，假设检验需要分散的数据，并且两个总体样本必须是正态分布的。在这些情况下，我们可以使用 t 分布来检验两个总体间均值的差异。

非参数的 MWW 检验并不需要分散数据，而且两个总体样本也不必呈正态分布。MWW 检验惟一的要求是数据的测量尺度至少必须是顺序的。除此之外，MWW 检验确定的是两个总体是否相同，而不是检验两个总体均值间的差异。MWW 检验的假设如下所示：

H_0：两个总体相同
H_a：两个总体不相同

首先我们通过对小样本情形进行检验来阐述如何应用 MWW 检验。

MWW 检验并不需要两个总体必须是正态分布。

小样本情形

如果两个样本的样本容量小于或者等于 10，则我们就应该使用 MWW 检验的小样本情形。我们通过考察约翰斯通高中学生的学习潜力这个例子来阐述 MWW 检验的小样本情形应用。约翰斯通高中大部分学习都在加非德初中或者码贝瑞初中就读，约翰斯通高中的管理者提出的问题是：来自加非德初中的入学学生是否在学习潜力上与来自码贝瑞初中的学生是一样的。假设如下所示：

H_0：两个样本总体的学生在学习潜力上是相同的
H_a：两个样本总体的学生在学习潜力上是不相同的

利用高中的记录，约翰斯通高中的管理人员挑选了 4 名加非德初中的学生和 5 名码贝瑞初中的学生组成随机样本。在本次研究中所涉及的学生如今在高中班级的名次记录如下，这 9 个学生在班级中的排名情况如表 19.5 所示。

表 19.5　高中班级排名数据

加非德初中学生		码贝瑞初中学生	
学生	班级排名	学生	班级排名
Fields	8	Hart	70
Clark	52	Phipps	202
Jones	112	Kirkwood	144
Tibbs	21	Abbott	175
		Guest	146

MWW 过程的第一步是将来自两个样本的混合数据按由低到高的顺序进行排列，最低值(班级排名第 8)的秩为 1，最高值(班级排名第 202)的秩为 9。9 名学生的排名情况如表 19.6 所示。

表 19.6　高中学生的排名情况

学生	班级排名	混合样本秩	学生	班级排名	混合样本秩
Fields	8	1	Kirkwood	144	6
Tibbs	21	2	Guest	146	7
Clark	52	3	Abbott	175	8
Hart	70	4	Phipps	202	9
Jones	112	5			

下一步是分别求出每一个样本的秩和，如表 19.7 所示。MWW 过程可以使用任何一个样本的秩和。在接下来的讨论中，我们使用 4 名来自加非德初中的学生所组成的样本的秩和，以 T 表示，因此，在我们的例子中有 $T=11$。

现在让我们来考察加非德初中样本秩和的特性。因为这个样本中有 4 名学生，加非德初中有可能在研究中有 4 名最好的学生，如果真的是这种情形，则 $T=1+2+3+4=10$ 将会是秩和的最小可能值。相反的情形是，加非德初中有可能有最差的学生，则 $T=6+7+8+9=30$ 将会是最大的可能值。因此，加非德初中样本的 T 值肯定位于 10 和 30 之间。

请注意，T 接近 10 就表示加非德初中的学生更为优秀或者名次更高，如果 T 接近 30，则表示加非德初中的学生学习潜力较差或者名次较低。如果两个学校的学生学习潜力相同的话，我们可以估计 T 值将会接近于平均值，即 $(10+30)/2=20$。

两个样本的容量都小于或者等于 10 的 MWW 检验的 T 临界值如附录 B 的表 9 所示，在这个表中，n_1 表示的是在这个检验中的秩和所使用的样本容量。我们可以直接从表中读出 T_L 的数值，T_U 的数值则可以从公式(19.5)中计算得出。

$$T_U = n_1(n_1 + n_2 + 1) - T_L \tag{19.5}$$

表 19.7　两所初中学生的秩和

	加非德初中			码贝瑞初中	
学生	班级排名	样本秩	学生	班级排名	样本秩
Fields	8	1	Hart	70	4
Clark	52	3	Phipps	202	9
Jones	112	5	Kirkwood	144	6
Tibbs	21	2	Abbott	175	8
		—	Guest	146	7
秩和		11			34

T_L 和 T_U 的值都不在拒绝域之中，只有在 T 严格小于 T_L 或者严格大于 T_U 时，我们才能拒绝相同总体的零假设。

举个例子，根据附录 B 的表 9 和 $\alpha = 0.05$ 的显著性水平，我们可以看到 $n_1 = 4$(加非德初中)、$n_2 = 5$(码贝瑞初中)的 MWW 统计量的低端临界值为 $T_L = 12$。根据公式(19.5)，我们计算 MWW 统计量的低端临界值为

$$T_U = 4(4 + 5 + 1) - 12 = 28$$

因此，MWW 决策法则可以表明：如果第一样本(加非德初中)的秩和小于 12 或者大于 28，则我们可以拒绝两个总体相同的零假设。拒绝法则如下：

如果 $T<12$ 或者 $T>28$，则我们可以拒绝 H_0。

如果我们用码贝瑞初中学生的秩和来进行检验，我们将得到：$n_1 = 5$，$n_2 = 4$，$T_L = 17$，$T_U = 33$，$T = 34$。根据 $T > T_U$，我们可以得到拒绝 H_0 的同样的结论。

参见表 19.7，可以看到 $T = 11$，因此，我们可以拒绝零假设，并且得出结论：在学习潜力方面，加非德初中的学生总体与码贝瑞初中的学生总体是不同的，加非德初中的学生所获得的班级排名比较高，这表明相比较于码贝瑞初中的学生，加非德初中的学生为高中学习所打下的基础要更好一些。

大样本情形

如果两个样本容量都大于或者等于 10 时，我们可以使用 T 分布的正态近似来进行 MWW 检验。以第三国民银行为例，我们来演示大样本情形。

第三国民银行有两个分行，表 19.8 列示的是从两个分行中的两个独立简单随机样本中挑选出来的数据。这些数据是否意味着两个分行的支票账户余额总体相同呢？

MWW 检验的第一步是将**混合**数据按由低到高的顺序进行排序，根据表 19.8 中的 22 个观测值，我们发现最低值(样本 2 的第 6 项)为 750 美元，我们把它的秩定为 1。继续排秩，可以得到

如下的表格：

余额(美元)	项目	秩
750	样本 2 的第 6 项	1
800	样本 2 的第 5 项	2
805	样本 1 的第 7 项	3
850	样本 2 的第 2 项	4
…	…	…
1 195	样本 1 的第 4 项	21
1 200	样本 1 的第 3 项	22

表 19.8　第三国民银行两个分行的账户余额

分行 1		分行 2	
账户	余额(美元)	账户	余额(美元)
1	1 095	1	885
2	955	2	850
3	1 200	3	915
4	1 195	4	950
5	925	5	800
6	950	6	750
7	805	7	865
8	945	8	1 000
9	875	9	1 050
10	1 055	10	935
11	1 025		
12	975		

在对混合数据进行排序时，我们可能会发现两个或者更多的相同数据。在这种情况下，我们可以将混合数据集合中所处位置的**平均秩**赋给它们。举个例子，(样本 1 的第 8 项)余额 945 美元，它的秩被定为 11。但是，数列中排在它后面的两个数值为 950 美元(样本 1 的第 6 项，样本 2 的第 4 项)，由于准备将这两个数值的秩定为 13 和 12，因此，根据前面的规则，将它们的秩定为 12.5。对于下面较高数值的 955 美元，我们可以继续排秩过程，将它的秩定为 14。表 19.9 即为全部数据集合和每一项观测值的秩。

在样本容量 $n_1 = 12$ 和 $n_2 = 10$ 的情况下，我们可以使用 T 的抽样分布正态近似。合适的抽样分布如下所示：

相同总体的 T 抽样分布

均值：$\mu_T = \frac{1}{2}\, n_1(n_1 + n_2 + 1)$ **(19.6)**

标准差：$\sigma_T = \sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}$ **(19.7)**

分布形式：如果 $n_1 \geqslant 10$ 且 $n_2 \geqslant 10$，则为正态分布。

表 19.9　第三国民银行两个样本数据的混合排秩

分行 1			分行 2		
账户	余额(美元)	秩	账户	余额(美元)	秩
1	1 095	20	1	885	7
2	955	14	2	850	4
3	1 200	22	3	915	8
4	1 195	21	4	950	12.5
5	925	9	5	800	2
6	950	12.5	6	750	1
7	805	3	7	865	5
8	945	11	8	1 000	16
9	875	6	9	1 050	18
10	1 055	19	10	935	10
11	1 025	17		秩和	83.5
12	975	15			
	秩和	169.5			

对于分行 1，我们可以得到

$$\mu_T = \tfrac{1}{2}\,12(12+10+1) = 138$$

$$\sigma_T = \sqrt{\tfrac{1}{12}12(10)(12+10+1)} = 15.17$$

T 的抽样分布如图 19.4 所示。紧接着是一般的假设检验过程，我们计算出检验统计量 z，用它来确定 T 观测值是否来自图 19.4 所示的抽样分布。如果 T 不是来自这个分布，则我们可以拒绝零假设，并得出结论两个总体是不同的。对统计量进行计算，我们可以得到

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{169.5 - 138}{15.17} = 2.08$$

在 0.05 的显著性水平下，我们知道如果要拒绝 H_0，z 必须要小于 -1.96 或者大于 $+1.96$。因为 $z = 2.08$，p - 值为 $2(0.500\,0 - 0.481\,2) = 0.037\,6 < \alpha = 0.05$，所以我们可以拒绝 H_0。因此，可以得出结论：两个总体是不同的。也就是说，两个分行的账户余额总体不同。

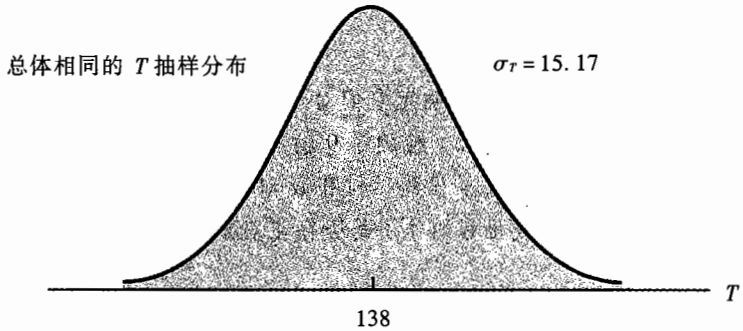


图 19.4　第三国民银行示例中的 T 抽样分布

如果我们使用的是分行 2 的秩和，则有 $n_1 = 10$，$n_2 = 12$，$\mu_T = 115$，$T = 83.5$。根据 $z = -2.08$，我们也可以拒绝 H_0 假设。

总的说来，曼 – 怀特尼 – 维尔克科森秩和检验确定两个独立随机样本是否来自相同总体的步骤如下：

1. 将混合样本观测值按由低到高的顺序进行排秩，如果值相同，则赋予相应秩的平均数。
2. 计算第一样本的秩和 T。
3. 在大样本情形下，将 T 的观测值与相同总体下的 T 抽样分布进行比较，以此来检验两个总体间的显著差异[如公式(19.6)和公式(19.7)所示]。标准化的检验统计量 z 的数值将决定是否拒绝 H_0。在小样本情形下，我们可以使用附录 B 的表 9 找到该检验的临界值。

评　注

本节中所讨论的非参数检验用于确定两个总体是否相同。如第 10 章中所讨论的 t 检验，参数统计检验是用于检验两个总体间均值是否相等。当我们可以拒绝均值相同的假设时，就可以得出结论：两个总体的均值是不相同的。如果我们利用 MWW 检验拒绝两个总体相同的假设，则无法证明它们是如何不同的。两个总体可能有不同的均值和不同的方差，以及(或者)不同的分布形式。因此，如果我们相信两个总体除了均值外其他方面都相同，则我们可以利用非参数方法拒绝 H_0，这表明其均值是不相同的。与参数 t 检验相比，MWW 检验的主要优点在于它对观测值来源的概率分布形式没有任何要求，并且它可以使用顺序数据。

练　习

应　用

18. 自测题 对两种燃料添加剂进行检验以确定它们对汽油消耗产生的作用，对于添加剂 1，我们检测了 7 辆汽车，对添加剂 2，检测了 9 辆汽车。下面的数据给出了使用两种添加剂所得到的每加仑汽油的行驶里程数。利用 $\alpha = 0.05$ 的显著性水平和 MWW 检验，检测这两种添加剂对汽油行驶里程的作用是否存在差异。

添加剂 1	添加剂 2
17.3	18.7
18.4	17.8
19.1	21.3
16.7	21.0
18.2	22.1
18.6	18.7
17.5	19.8
	20.7
	20.2

19. 注册会计师和财务分析师的个人起始年薪样本如下所示(*Fortune*, June 26, 1995)，起始年薪以千美元计。

自测题

注册会计师	财务分析师	注册会计师	财务分析师
25.2	24.0	30.0	28.6
33.8	24.2	25.9	24.7
31.3	28.1	34.5	28.9
33.2	30.9	31.7	26.8
29.2	26.9	26.9	23.9

a. 在 $\alpha=0.05$ 的显著性水平下，对注册会计师和财务分析师起始年薪没有差异的假设进行检验，能得出什么结论?

b. 两种职业年薪的样本均值是多少?

20. 拥有同等学历的男性和女性在职业收入上的差异正在缩小，但是收入并没有达到靠近的程度(*USA Today*, September 15, 2000)。具有学士学位的 7 位女性和男性的职业收入数据如下所示，收入以千美元计。

男性	30.6	75.5	45.2	62.2	38.2	49.9	55.3
女性	44.5	35.4	27.9	40.5	25.8	47.5	24.8

a. 男性的收入均值是多少? 女性呢?

b. 在 $\alpha=0.05$ 的显著性水平下，对相同总体进行假设检验，能得出什么结论?

21. 对两种型号的汽车进行行驶里程状况的检验，从每一种型号中随机挑选出 12 辆汽车，然后在高速公路上行驶 1 000 英里，得到各种型号的汽车每加仑汽油行驶里程数，如下所示：

型号 1		型号 2	
汽车	每加仑行驶里程	汽车	每加仑行驶里程
1	20.6	1	21.3
2	19.9	2	17.6
3	18.6	3	17.4
4	18.9	4	18.5
5	18.8	5	19.7
6	20.2	6	21.1
7	21.0	7	17.3
8	20.5	8	18.8
9	19.8	9	17.8
10	19.8	10	16.9
11	19.2	11	18.0
12	20.5	12	20.1

在 $\alpha=0.10$ 的显著性水平下，对这两种型号的汽车每加仑汽油行驶里程数总体间是否存在显著差异进行检验。

22. 《商业周刊》每年都公布有关世界最大的 1 000 家公司的统计数据。一个公司的市盈率(P/E)表示的是这个公司的当前股票价格除以最近 12 个月的每股收益。表 19.10 列出的是 10 家日本公司和 12 家美国公司的市盈率(*Business Week*, July 11, 1994)。这两个国家公司的市盈率是否存在显著差异？利用 MWW 检验和 $\alpha=0.01$ 的显著性水平来支持你的结论。

表 19.10 日本和美国公司的市盈率

日本		美国	
公司	市盈率	公司	市盈率
Sumitomo Corp.	153	Gannet	19
Kinden	21	Motorola	24
Heiwa	18	Schlumberger	24
NCR Japan	125	Oracle Systems	43
Suzuki Motor	31	Gap	22
Fuji Bank	213	Winn-Dixie	14
Sumitomo Chemical	64	Ingersoll-Rand	21
Seibu Railway	666	American Electric Power	14
Shiseido	33	Hercules	21
Toho Gas	68	Times Mirror	38
		WellPoint Health	15
		Northern States Power	14

23. 两个样本分别由冬季和夏季几个月中的一段时期组成，下面的警察记录显示了在这两个样本期间的每日犯罪数据。在 $\alpha = 0.05$ 的显著性水平下，请确定犯罪数量在冬季和夏季这两个样本期间是否存在显著差异。

冬季	夏季
18	28
20	18
15	24
16	32
21	18
20	29
12	23
16	38
19	28
20	18

24. 某一特定品牌的微波炉在达拉斯 10 家商场和圣·安东尼奥的 13 家商场的标价如下所示。在 $\alpha = 0.05$ 的显著性水平下，对这一品牌的微波炉在两个城市的销售价是否相同进行检验。

达拉斯	圣·安东尼奥
445	460
489	451
405	435
485	479
439	475
449	445
436	429
420	434
430	410
405	422
	425
	459
	430

25. 全美建筑协会给出了最为流行的家庭装修工程的成本数据（*USA Today*，June 17，1997），在 $\alpha = 0.05$ 的显著性水平下，利用曼－怀特尼－维尔克科森检验来检验厨房装修成本和主卧装修成本是否存在差异。

厨房	主卧
25 200	18 000
17 400	22 900
22 800	26 400
21 900	24 800
19 700	26 900
23 000	17 800
19 700	24 600
16 900	21 000
21 800	
23 600	

19.4 克鲁斯卡尔 - 沃利斯检验

19.3 节中的 MWW 检验是用于检验两个总体是否相同，而克鲁斯卡尔 - 沃利斯检验则将其扩展到检验 3 个或者更多总体是否相同的情形。对于 $k \geq 3$ 个总体的**克鲁斯卡尔 - 沃利斯检验**（Kruskal-Wallis test），其假设形式如下：

H_0：所有总体相同

H_a：并不是所有总体都相同

克鲁斯卡尔 - 沃利斯检验是基于对来自 k 个总体中的每一个总体独立随机样本进行分析的。

该检验是第 13 章中 ANOVA 的一个补充，它集中讨论总体为 k 的均值的相等性。

在第 13 章中，我们已经讨论了方差分析(ANOVA)可以用于检验 3 个或者更多总体的均值是否相等。ANOVA 过程要求数据类型为区间或者比例类型，所有总体均为正态分布，而且各个总体的方差必须相等。

表 19.11 20 名威廉姆斯员工的表现评估等级

大学 A	大学 B	大学 C
25	60	50
70	20	70
60	30	60
85	15	80
95	40	90
90	35	70
80		75

非参数的克鲁斯卡尔 - 沃利斯检验可以被用于序数、区间或者比例类型的数据。并且，克鲁斯卡尔 - 沃利斯检验不要求总体为正态分布的假设。因此，只要来自于 $k \geq 3$ 的独立随机样本的数据是序数的，或者正态假设以及方差相等的假设未能完全得到满足时，克鲁斯卡尔 - 沃利斯检验就可以作为一种替代统计检验方法用于检验总体是否相同。我们将克鲁斯卡尔 - 沃利斯检验应用于一个挑选雇员的例子，以此来阐述克鲁斯卡尔 - 沃利斯检验。

威廉姆斯制造公司从当地 3 所大学招聘管理工作人员。最近该公司人力资源部收集并评选出了员工的年度表现等级，希望能够通过这

种评选来确定从这 3 所大学招聘的管理人员是否存在工作表现差异。从 7 名来自 A 大学的员工、6 名来自 B 大学的员工、7 名来自 C 大学的员工所组成的样本中得到的评选数据如表 19.11 所示。每名管理人员的总体表现等级分值在 0 —100 之间,其中 100 是可能的最高表现等级。

假设我们计划检验这三个总体之间是否存在表现评价的差异，以每个样本的秩和为基础，克鲁斯卡尔 - 沃利斯检验统计量计算如下：

克鲁斯卡尔 - 沃利斯检验统计量

$$W = \left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n_T+1) \tag{19.8}$$

式中　k——总体个数；

n_i——样本 i 的个体数；

n_T——Σn_i = 所有样本的个体总数；

R_i——样本 i 的秩和。

克鲁斯卡尔和沃利斯已经证明在各个总体相同的零假设下，我们可以用一个自由度为 $k-1$ 的 χ^2 分布来近似 W 抽样分布。如果每个样本的容量都大于或者等于 5，则这种近似是可以被普遍接受的。

克鲁斯卡尔 - 沃利斯检验仅仅需要用到数据的秩。

为了计算例中的 W 统计量，我们首先必须对所有 20 个数据项进行排序。来源于 B 大学样本的 15 是最小的数据值，它的秩为 1；来源于 A 大学样本的 95 是最大数值，其秩为 20。表 19.12 给出了数据的数值、相应的秩以及三个样本的秩和。请注意对于同分值，我们把它的秩定义为秩的平均数，*例如，数值 60，70，80 和 90 都具有同分值。

表 19.12　20 名威廉姆斯公司的雇员混合秩

大学 A	秩	大学 B	秩	大学 C	秩
25	3	60	9	50	7
70	12	20	2	70	12
60	9	30	4	60	9
85	17	15	1	80	15.5
95	20	40	6	90	18.5
90	18.5	35	5	70	12
80	15.5		—	75	14
秩和	95		27		88

* 如果观测到了大量的同分秩，则我们必须对公式(19.8)进行修正，修正后的公式见 W. J. Conover 所著的 *Practical Nonparametric Statistics*。

样本容量为

$$n_1 = 7 \qquad n_2 = 6 \qquad n_3 = 7$$

并且

$$n_T = \Sigma n_i = 7 + 6 + 7 = 20$$

我们可以利用公式(19.8)计算 W 统计量

$$W = \frac{12}{20(21)}\left[\frac{(95)^2}{7} + \frac{(27)^2}{6} + \frac{(88)^2}{7}\right] - 3(20+1) = 8.92$$

附录 11.1 和 11.2 中的计算步骤说明了如何利用 Minitab 和 Excel 来计算 p-值 = 0.012。

卡方分布表(见附录 B 的表 3)显示出在自由度为 $k-1=2$ 并且在这个分布高端的 $\alpha=0.05$ 时，临界值为 $\chi^2 = 5.991\,47$。因为检验统计量 $W = 8.92 > 5.99\,147$，因此我们可以拒绝三个总体相同的零假设，并得出结论：来自不同大学的管理人员，他们的业绩表现是明显不同的。表现等级较低的为 B 大学，因此公司有理由减少从 B 大学招聘员工的数额，或者至少需要对毕业于这个大学的员工进行更加深入的考核。

评 注

本例中的克鲁斯卡尔-沃利斯检验过程开始于区间尺度数据的收集，而这些数据是用于表示员工的业绩评价等级。如果 20 名员工的数据为顺序数据，这个过程也是奏效的。在这种情况下，克鲁斯卡尔-沃利斯检验可以直接应用原始数据。关于员工业绩评价等级的排秩过程则是可以忽略的。

练 习

方 法

26. 自测题 由 15 名消费者所组成的一个评价组对下面三种商品性能给出了如下的评价：

商品		
A	B	C
50	80	60
62	95	45
75	98	30
48	87	58
65	90	57

利用克鲁斯卡尔－沃利斯检验和 $\alpha=0.05$ 的显著性水平来确定这三种商品性能评价是否存在显著差异。

27. 我们对三个接受测试的准备程序进行了评价，下面是 20 个曾经进行过这种测试准备程序的人所组成的样本得分，在 $\alpha=0.01$ 的显著性水平下，利用克鲁斯卡尔－沃利斯检验确定这三种准备程序是否存在显著差异。

程　序		
A	**B**	**C**
540	450	600
400	540	630
490	400	580
530	410	490
490	480	590
610	370	620
	550	570

应　用

28. 自测题 下面的活动如果每个星期进行三天、每次 40 分钟，这样的消耗会导致体重下降。以下的样本数据为三种不同的活动在 40 分钟内所消耗的卡路里。这些数据是否能够表明这三种活动所消耗的卡路里是不同的？在 $\alpha=0.05$ 的显著性水平下，你能得出什么结论？

游泳	网球	自行车
408	415	385
380	485	250
425	450	295
400	420	402
427	530	268

29. 以下是三种不同的汽车在测试中所得到的每加仑汽油行驶里程。在 $\alpha=0.05$ 的显著性水平下，利用克鲁斯卡尔－沃利斯检验来确定这三种汽车的汽油行驶里程是否有显著的不同。

汽车		
A	B	C
19	19	24
21	20	26
20	22	23
19	21	25
21	23	27

30. 一家大公司将它的许多第一流的经理送去进行脱产的管理技能学习，有四家不同的管理发展中心提供这种课程的培训，这家公司打算确定这四家管理中心所提供的课程质量是否有差异。抽出 20 名参加过课程培训的员工组成样本，然后将这些员工按管理技能进行排序。结果如下所示：

课程	管理技能秩				
1	3	14	10	12	13
2	2	7	1	5	11
3	19	16	9	18	17
4	20	4	15	6	8

请注意，秩最高的管理人员参加了课程 2，而秩最低的管理人员参加了课程 4。在 $\alpha=0.05$ 的显著性水平下，检验这四种课程是否存在显著的不同。

31. 畅销的糖果往往含有较高的卡路里。假设下列数据为 M&Ms、Kit Kat、Milky Way Ⅱ糖果样本中所含的卡路里。在 $\alpha=0.05$ 的显著性水平下检验这三种糖果所含有的卡路里，你能得出什么结论？

M&Ms	Kit Kat	Milky Way Ⅱ
230	225	200
210	205	208
240	245	202
250	235	190
230	220	180

19.5 秩相关

斯皮尔曼秩相关系数与应用于序数数据或排秩数据的皮尔逊相关系数是相同的。

相关系数是对两个可以得到区间或者比例数据的变量间的线性关系的度量。在本节中，我们

考察在只能得到顺序数据时两个变量间关系的度量。**斯皮尔曼秩相关系数**(Spearman rank-correlation coefficient)正是针对这一目的而提出来的。

斯皮尔曼秩相关系数

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{19.9}$$

式中　n——排秩的项目或个体的数量；

x_i——第 i 项对于第一个变量的秩；

y_i——第 i 项对于第二个变量的秩；

d_i——$x_i - y_i$。

下面我们用一个例子来阐述斯皮尔曼秩相关系数的应用。有一家公司打算确定那些在雇用的时候预计会成为优秀销售人员的人是否能够真正拥有良好的销售业绩。为了解决这个问题，负责人力资源的副经理仔细考察了公司目前在职的 10 名销售人员原来的求职面试总结、学习成绩和推荐信。经过对这些人员的检查，这位副经理根据这 10 名销售人员在面试时的信息将他们的预计销售业绩做了排秩。然后再根据他们在两年中的实际销售业绩进行第二次排秩。表 19.13 给出了相关的数据和排秩结果。统计上的问题是：预计销售业绩的排秩情况和两年来的实际销售业绩表现是否一致。

表 19.13　10 名销售人员两年来的预期和实际销售业绩的排秩

销售人员	预期销售排秩	两年来的销售额	根据两年来的销售额所做排秩
A	2	400	1
B	4	360	3
C	7	300	5
D	1	295	6
E	6	280	7
F	3	350	4
G	10	200	10
H	9	260	8
I	8	220	9
J	5	385	2

让我们来计算表 19.13 中数据的斯皮尔曼相关系数。计算过程如表 19.14 所示。我们可以看到秩相关系数为正的 0.73。斯皮尔曼秩相关系数的取值范围为 -1.0 到 +1.0 之间，而且斯皮尔曼秩相关系数的解释类似于样本相关系数，接近于 1.0 的正值表示两种排秩之间存在很强的正相关关系；其中一个秩上升，则另一个秩也会上升。秩相关系数接近于 -1.0 则意味着两种排秩之间存在很强的负相关关系，如果一个秩上升，则另一个秩将会下降。$r_s = 0.73$ 表示预计销售潜力和实际销售业绩之间存在正相关关系。预计销售潜力排秩较高的人在实际销售业绩表现上的排秩

也比较高。

表 19.14 预期销售和实际销售业绩的斯皮尔曼秩相关系数的计算过程

销售人员	x_i = 预期销售排秩	y_i = 根据两年来的销售额所做排秩	$d_i = x_i - y_i$	d_i^2
A	2	1	1	1
B	4	3	1	1
C	7	5	2	4
D	1	6	-5	25
E	6	7	-1	1
F	3	4	-1	1
G	10	10	0	0
H	9	8	1	1
I	8	9	-1	1
J	5	2	3	9
				$\Sigma d_i^2 = 44$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(44)}{10(100-1)} = 0.73$$

秩相关显著性检验

至此我们已经学习了怎样利用样本结果来计算样本秩相关系数。和许多其他的统计过程相似，我们也希望利用抽样结果来估计总体的秩相关系数 ρ_s。为了估计总体的秩相关系数，我们必须检验如下的假设：

$$H_0: \rho_s = 0$$
$$H_a: \rho_s \neq 0$$

在不存在秩相关($\rho_s = 0$)的零假设下，排秩过程是独立的，而且 r_s 的抽样分布如下所示：

r_s 的抽样分布

均值：$\mu_{r_s} = 0$ **(19.10)**

标准差：$\sigma_{r_s} = \sqrt{\dfrac{1}{n-1}}$

分布形式：如果 $n \geq 10$，则为正态分布。

预计销售潜力和销售表现之间的样本秩相关系数为 $r_s = 0.73$，根据这个数值，我们可以检验秩相关的显著性。从公式(19.10)我们可以看到 $\mu_{r_s} = 0$，而且根据公式(19.11)我们可以得到：$\sigma_{r_s} = \sqrt{1/(10-1)} = 0.33$。利用检验统计量，我们可以得到

$$z=\frac{r_s-\mu_{r_s}}{\sigma_{r_s}}=\frac{0.73-0}{0.33}=2.21$$

在 $\alpha=0.05$ 的显著性水平下，我们可以看到：如果 $z<-1.96$ 或者 $z>1.96$，则我们可以拒绝不相关的零假设。由于 $z=2.21$，并且 p - 值 $=2(0.500\,0-0.486\,4)=0.027\,2$ 小于 $\alpha=0.05$，所以我们可以拒绝不存在秩相关的假设。因此，我们可以得出结论：销售潜力与实际销售业绩之间存在显著的秩相关关系。

练　习

方　法

32.　考察下面的 10 个元素所组成的样本的排秩数据：



元素	x_i	y_i	元素	x_i	y_i
1	10	8	6	2	7
2	6	4	7	8	6
3	7	10	8	5	3
4	3	2	9	1	1
5	4	5	10	9	9

a. 计算这些数据的斯皮尔曼相关系数。

b. 在 $\alpha=0.05$ 的显著性水平下，检验秩相关的显著性，并陈述你的结论。

33. 考察下面 6 个项目的两组排秩：

情形 1			情形 2		
项目	第一次排秩	第二次排秩	项目	第一次排秩	第二次排秩
A	1	1	A	1	6
B	2	2	B	2	5
C	3	3	C	3	4
D	4	4	D	4	3
E	5	5	E	5	2
F	6	6	F	6	1

请注意第一种情形两次排秩是相同的，但是第二种情形两次排秩正好相反。对于这两种情形的每一种，你所预计的斯皮尔曼相关系数是多少？请解释原因，并计算每种情形下的秩相关系数。

应　用

34. 下面这个样本由 11 个州组成，下表给出了它们在学生教师数量的比值(1 = 最低，11 = 最高)以及每个学生费用(1 = 最低，11 = 最高)的排秩。



	秩			秩	
州	学生—教师比值	单位学生费用	州	学生—教师比值	单位学生费用
亚利桑那	10	9	马萨诸塞	1	1
科罗拉多	8	5	内布拉斯加	2	7
佛罗里达	6	4	北达科他	7	8
爱达荷	11	2	南达科他	5	10
衣阿华	4	6	华盛顿	9	3
路易斯安那	3	11			

在 $\alpha = 0.05$ 的显著性水平下，单位学生费用和学生—教师比值之间是否存在显著相关关系?

35. Harris 公司进行了一项国际性的调查，对一流跨国公司及其声誉进行评估(*The Wall Street Journal*, November 18, 1999)。下面的两个表格显示了 10 个跨国公司的声誉排秩以及被采访者愿意购买该公司股票的比例排秩情况。我们预计在这两者之间会有一个正的秩相关，因为一个公司的声誉越高，则公众会更加踊跃地购买它的股票。

	声誉	购买意愿
Microsoft	1	3
Intel	2	4
Dell	3	1
Lucent	4	2
Texas Instruments	5	9
Cisco Systems	6	5
Hewlett-Packard	7	10
IBM	8	6
Motorola	9	7
Yahoo	10	8

a. 计算声誉和购买意愿之间的秩相关系数。

b. 对正的秩相关系数进行显著性检验，p – 值是多少?

c. 在 $\alpha = 0.05$ 的显著性水平下，你能得出什么结论?

36. 下面是 1996 年职业高尔夫球手击球距离和击球杆数样本的排秩情况(*Golf Digest*，January 1997)。在 $\alpha=0.10$ 的显著性水平下，击球距离和击球杆数之间的秩相关系数是多少？

职业高尔夫球选手	击球距离	击球杆数
Fred Couples	1	5
David Duval	5	6
Ernie Els	4	10
Nick Faldo	9	2
Tom Lehman	6	7
Justin Leonard	10	3
Davis Love Ⅲ	2	8
Phil Mickelson	3	9
Greg Norman	7	4
Mark O'Meara	8	1

37. 一个学生组织对近期毕业的学生和在校生进行了一项调查，希望通过这次调查获得关于某所大学教学质量的信息。对调查结果进行分析，得到了下列教学能力的排秩。请问由在校生给出的排秩是否和近期毕业的学生给出的排秩情况相同？在 $\alpha=0.10$ 的显著性水平下，检验秩相关的显著性。

	秩	
教授	在校生	近期毕业的学生
1	4	6
2	6	8
3	8	5
4	3	1
5	1	2
6	2	3
7	5	7
8	10	9
9	7	4
10	9	10

小　结

在本章中，我们给出了几种非参数方法的统计过程。前面章节中的参数方法通常要求的是区间或者比例类型的数据，而且一般都是以总体的各种假设为基础的(比如说，假设概率分布是正态分布)。因为非参数方法不仅可以应用于区间或者比例类型的数据，而且可以应用于名

义和序数类型的数据，而且对总体分布没有要求，所以它们扩展了统计分析研究对象问题的范围。

符号检验是一种在只能够得到名义数据时确认两个总体之间差异的非参数过程。在小样本情形下，我们可以用二项概率分布来确定符号检验的临界值；在大样本情形下，我们可以使用正态近似。维尔克科森符号秩检验是在能够得到匹配对象的区间或者比例数据时分析匹配样本的过程。它对总体的分布并没有假设要求。维尔克科森符号过程检验的是被考察的两个总体是否相同的假设。

曼－怀特尼－维尔克科森检验是一种基于两个独立随机样本，用于检验两个总体间差异的非参数方法。对于小样本情形，我们给出了临界值表。对于大样本情形，我们则使用正态近似。克鲁斯卡尔－沃利斯检验将曼－怀特尼－维尔克科森检验扩展到了三个或者更多的总体，它是和检验总体均值间差异的 ANOVA 参数检验相对应的非参数方法。

在本章的最后一节中，作为一种对考察对象两种顺序或者排秩间的相关性的度量，我们介绍了斯皮尔曼秩相关系数。

术语释义

非参数方法： 对总体的概率分布假设和测量尺度要求很少的参数方法。如果能够得到名义或者顺序数据时，我们可以应用这些方法。

自由分布检验： 另一种非参数统计方法的名称，它表明不需要对总体概率分布作出假设。

符号检验： 一种非参数统计检验方法，它以对名义数据的分析为基础，可以用于对两个总体间的差异进行判断。

维尔克科森符号秩检验： 一种非参数统计检验，以对相互匹配或成对样本的分析为基础，用于判断两个总体间的差异。

曼－怀特尼－维尔克科森检验： 一种非参数统计检验，以对两个独立样本进行分析为基础，用于判断两个总体间的差异。

克鲁斯卡尔－沃利斯检验： 一种非参数检验，用于确定三个或者多个总体间的差异。

斯皮尔曼秩相关系数： 一种基于两个变量之间的排秩相互关系的测量。

重要公式

符号检验(大样本情形)

$$\text{均值：}\ \mu = 0.50n \tag{19.1}$$

$$\text{标准差：}\ \sigma = \sqrt{0.25n} \tag{19.2}$$

维尔克科森符号秩检验

$$\text{均值：}\ \mu_T = 0 \tag{19.3}$$

$$\text{标准差：}\ \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{6}} \tag{19.4}$$

曼－怀特尼－维尔克科森检验(大样本情形)

$$\text{均值：}\ \mu_T = 1/2\ n_1(n_1 + n_2 + 1) \tag{19.6}$$

$$\text{标准差：}\ \sigma_T = \sqrt{1/12\ n_1 n_2(n_1 + n_2 + 1)} \tag{19.7}$$

克鲁斯卡尔－沃利斯检验统计量

$$W = \left[\frac{12}{n_T(n_T+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(n_T+1) \tag{19.8}$$

斯皮尔曼秩相关系数

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{19.9}$$

补充练习

38. 《美国民意调查》(*The American Opinion Survey*)(*The Wall Street Journal*, March 4, 1997) 提出了下面的一些问题：你是赞同还是反对为将子女送入私立学校的家庭提供来自国家税收的代金券或者减税？在接受调查的 2 010 人中，有 905 人赞同这种资助政策，1 045 人反对这种资助，60 人没有表态。这些数据是否能够表明人们对为将子女送入私立学校的家庭进行资助存在分歧？利用 $\alpha = 0.05$ 进行显著性检验。

39. 一居室住宅的全国平均销售价格是 118 000 美元(*The Wall Street Journal Almanac*, 1998)。假设下面的数据来自休斯敦和波士顿现有的一居室住宅销售情况：

	高于 118 000 美元	等于 118 000 美元	低于 118 000 美元
休斯敦	11	2	32
波士顿	27	1	13

a. 休斯敦的平均销售价格是否低于 118 000 美元的全国平均价格？在 $\alpha = 0.05$ 的显著性水平下，利用统计检验量来支持你的结论。

b. 波士顿的平均销售价格是否高于 118 000 美元的全国平均价格？在 $\alpha = 0.05$ 的显著性水平下，利用统计检验量来支持你的结论。

40. 有 12 名家务人员对两种品牌的冰箱进行零售价格的评估。他们估计的零售价格如下，根据这

些数据评价两种品牌冰箱零售价之间是否存在差异，利用 $\alpha=0.05$ 的显著性水平。

家务人员	品牌1(美元)	品牌2(美元)	家务人员	品牌1(美元)	品牌2(美元)
1	650	900	7	700	890
2	760	720	8	690	920
3	740	690	9	900	1 000
4	700	850	10	500	690
5	590	920	11	610	700
6	620	800	12	720	700

41. 有一项研究用于评价某种新饲料对家禽的增重能力，这项研究利用了由12只小鸡组成的样本，研究历时六星期。每只鸡在六周前后的重量均被记录下来。12只鸡的重量在六周前后的差值分别为：1.5，1.2，-0.2，0.0，0.5，0.7，0.8，1.0，0.0，0.6，0.2，-0.01。其中负值表示在测试期间鸡的体重减少，而0.0则表示体重没有什么变化。利用 $\alpha=0.05$ 的显著性水平，确定这种新饲料是否能使鸡的体重增加。

42. 以下的数据是关于两条生产线上的产品重量，利用 $\alpha=0.10$ 的显著性水平，检验两条生产线所生产的产品重量是否存在差异。

生产线1	生产线2
13.6	13.7
13.8	14.1
14.0	14.2
13.9	14.0
13.4	14.6
13.2	13.5
13.3	14.4
13.6	14.8
12.9	14.5
14.4	14.3
	15.0
	14.9

43. 某个经纪人计划确定对一个过程进行评估所需要的时间，他有三种不同的方法可以用于对这个过程进行评估。下面有18名评估者利用这三种方法对过程进行了评估，所需时间(小时)分别如下所示：

第一种方法	第二种方法	第三种方法
68	62	58
74	73	67
65	75	69
76	68	57

第一种方法	第二种方法	第三种方法
77	72	59
72	70	62

利用 $\alpha=0.05$ 的显著性水平对此进行检验，以确定这三种方法所需要的时间是否存在显著差异。

44. 某个公司有 20 名工程师，这 20 名工程师都已经在该公司工作超过三年以上，由他们组成一个样本，就这些工程师的管理能力对这个样本进行排秩。其中有一些工程师已经参加过公司所举办的管理课程培训，一些参加了当地大学的脱产培训，剩下的则没有参加任何培训科目。利用下面的排秩结果和 $\alpha=0.025$ 的显著性水平，检验他们的管理能力是否存在显著的差异。

未参加培训	参加公司培训	参加脱产培训
16	12	7
9	20	1
10	17	4
15	19	2
11	6	3
13	18	8
	14	5

45. 以下是四位教师的课程评估排序，利用 $\alpha=0.05$ 的显著性水平和克鲁斯卡尔－沃利斯检验来检验这四位教师的教学能力是否存在显著差异。

教师	课程评估排序								
Black	88	80	79	68	96	69			
Jennings	87	78	82	85	99	99	85	94	
Swanson	88	76	68	82	85	82	84	83	81
Wilson	80	85	56	71	89	87			

46. 在一项统计课程中，有 15 名学生期中和期末的考试成绩如下所示：

秩		秩		秩	
期中	期末	期中	期末	期中	期末
1	4	6	2	11	14
2	7	7	5	12	15
3	1	8	12	13	11
4	3	9	6	14	10
5	8	10	9	15	13

计算这一组数据的斯皮尔曼秩相关系数，并利用 $\alpha=0.10$ 的显著性水平检验其显著相关性。

20 质量管理统计方法

统计实例:美国道化学品公司

20.1 **统计过程管理**

控制图

$\bar{x}$ 控制图:已知过程均值和标准差

$\bar{x}$ 控制图:未知过程均值和标准差

R 控制图

p 控制图

np 控制图

控制图的解释

20.2 **抽样验收**

KALI 有限公司:抽样验收实例

计算批量验收的概率

选择抽样验收方案

多重抽样方案

统计实例

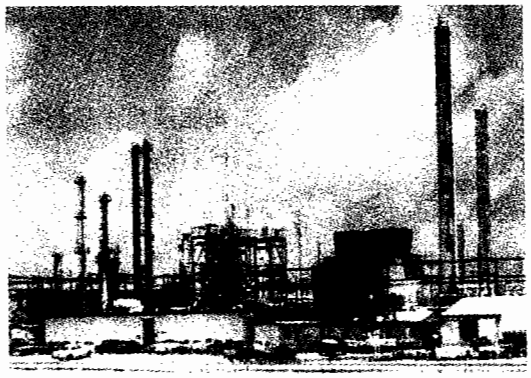
美国道化学品公司*

弗里波特，得克萨斯州

美国得克萨斯州道化学品公司（Dow Chemical U. S. A.，Texas Operations）始建于1940年，当年道化学品公司在得克萨斯州的Gulf海湾购买了800英亩的土地用以建造镁生产设施。如今那片厂址已经扩展到超过5 000英亩，该公司亦成为世界上最大的石油化工复合型企业之一。得克萨斯州道化学品公司的产品包括镁、苯乙烯、塑料、胶合剂、溶剂、乙二苯和氯气等。一些产品制造出来仅仅是为了用于其他产品的生产过程，但是很多其他终端产品则成为医药、牙膏、狗食、水管、冰箱、奶桶、垃圾袋、洗发水和家具等产品的必要的基础性原料。

得克萨斯州道化学品公司所生产的镁产量超过世界总产量的30%。镁是一种很轻的金属，它可以用于生产从网球拍到衣物箱等各类的产品。该公司的镁事业部是首先利用质量统计管理方法培训其技术人员和管理人员的部门，统计质量管理的最初的成功应用是在化学处理的过程之中。

其中的一个应用即为干燥剂的生产。在这个应用中，他们每隔一定周期抽取产品作为样本，计算出每个样本的平均值，并且将它们记录在一种被称为$\bar{x}$控制图的控制图上。利用这种控制图，道化学品公司的分析专家可以监控产品生产过程，随时预警非正常操作。在某个例子中，分析专家开始观测样本均值的数值，该数值显示有一个操作已经超出了它的设计范围。经过对控制图和该操作本身的更深层次的观察，分析专家发现误差可以追溯到某一操作人员所引起的问题上来。在那个操作员重新培训后所记录的$\bar{x}$控制图上，我们可以看到操作质量已经有了明显的改进。



统计质量管理使得道化学品公司大幅改进它的生产工艺，提高了产量。© Dan Guravich / CORBIS.

道化学品公司在它任何应用过统计质量管理的地方均取得了质量的大幅改进。他们实现了每年数十万美元的成本节约，并且不断地取得新的应用发现。

在本章中我们将阐述诸如道化学品公司所使用的$\bar{x}$控制图是如何生成的。控制图是统计质量管理的一部分，它被称为统计过程管理。我们将针对某个样本是否决定接受或拒绝一组项目的情形来讨论质量管理的方法。

* 作者衷心感谢道化学公司的技术经理Clifford B. Wilson，提供这一统计实例。

美国质量管理协会（ASQC）对**质量**是这样定义的：质量指的是产品的所有性质和特征，这些性质和特征使得该产品能够满足特定的需要。换句话说，质量可以测度一个产品或服务满足顾客需求的程度。很多企业意识到在全球经济竞争的今天，他们必须追求高水平的质量。因此，对质量的检验和维护有必要在监控和维护方法上加以重视。

质量保证指的是为了达到和维护质量，由某个组织制定的政策、步骤和指导规范的整套系统。质量保证有两个重要组成部分：质量工程和质量管理。**质量工程**的目标包括了产品设计和生产过程，还包括在生产之前识别潜在的质量问题。**质量管理**(quality control，亦称做质量控制)包括进行一系列的检验和测量以确定质量是否达到相关标准。如果质量标准没有被满足，则可以通过纠正或者预防活动来达到和维护质量的一致性。就如我们在本章中所要讲述的，统计方法在质量管理中是非常有用的。

传统的生产方法在质量管理中已经被认为是不令人满意的，并且正在被改进过的管理工具和技术取代。颇具讽刺意味的是，两个美国顾问 W·爱德华·德明博士 (W. Edwards Deming) 和约瑟福·朱南博士 (Dr. Joseph Juran) 在质量管理方面曾帮助过日本人的培训。

第二次世界大战后，W·爱德华·德明博士成为日本工业顾问。他被尊称为说服日本高层管理者使用统计质量管理方法的人。

虽然质量是每个人的工作，但德明强调质量必须有管理者来领导。他建立了一个具有 14 个条款的表格，其内容是他认为对管理者至关重要的职责。例如，德明认为管理者必须结束对大量检验的依赖；必须终止仅凭价格即作出商业行为的做法；必须寻求对所有产品生产过程和服务的不间断的改良；必须培养团队合作精神的环境；以及必须消除数量化的目标、口号和确定数量定额的工作规范。也许最重要的是管理者必须创造一个工作环境，在这个环境中，质量和生产率的标准在任何时候均能得到维护。

在 1987 年，美国国会颁布了 107 公共法：Malcolm Baldrige 国家质量改进法令(Malcolm Baldrige National Quality Improvement Act)。Malcolm Baldrige 国家质量奖每年授予那些在质量方面有突出成绩的美国公司。这个奖项，以及那些个人，比如爱德华博士和约瑟福博士的观点已经对高层管理者很有帮助，他们意识到了提高服务质量和产品质量对他们的企业而言是至关重要的挑战。Malcolm Baldrige 国家质量奖的获得者包括摩托罗拉、IBM、施乐和联邦快递等公司。在本章中，我们阐述在质量管理中所用的两种统计方法。第一种方法，**统计过程管理**，利用了图像显示的方法，比如**控制图**，来监控生产过程。应用的目标在于确定生产过程是否能继续或是否需要调整以达到期望的质量水平。第二种方法，**抽样验收**，该种方法用于根据在样本中所观测的质量水平来决定是否接受或拒绝一组项目时的情形。

20.1　统计过程管理

在本节中，我们通过观测一个货物的连续生产过程来研究质量管理步骤。在抽样验收和对产品生产进行监控的基础上，生产者将作出决定：或者继续生产过程或者调整生产过程，以使所生产的产品或生产项目达到能够接受的质量标准。

虽然在制造和生产过程中会执行严格的高标准，但是机械设备会不可避免地磨损，振动将会使机械设备的设置难以调整，所采购的材料亦会有缺陷，人工操作员也会犯操作上的错误。任何

一个或其中的全部因素均会导致低质量的产品生产出来。幸运的是，监控产品生产的程序是行之有效的，生产的过程可因此而得到调整或修正。

不断的改进是全面质量管理活动的最重要的概念之一，控制图的最重要的应用即为改进过程。

如果所生产的产品在质量上的差异来源于**特定因素**(assignable causes)，比如工具的磨损、错误的机器设置、质量低下的原材料或者操作员的操作失误，那么生产过程应该立即调整或纠正。除此之外，如果差异来源于所谓的**一般因素**(common causes)，即发生在材料质量上的随机化差异、温度、湿度等诸如此类的因素，这些因素生产者可能无法控制，生产过程也不需要调整。统计质量管理的主要目标就在于确定生产产品的质量差异是来源于特定因素还是一般因素。

无论何时我们发现特定因素，我们均可以断定生产过程处于**失控状态**。在那种情况下，应该采取纠正措施将生产过程调整，重新回到可以接受的质量水平。然而，如果生产过程中的产品差异仅仅来源于一般因素，我们可以断定生产过程正处于**统计控制状态**，或者简单地说处于**控制状态**；在这种情况下，我们没有必要进行任何改变或者调整。

过程控制的统计程序是建立在第 9 章介绍的假设检验方法基础之上。零假设 H_0 代表生产过程处于控制状态，备择假设 H_a 代表生产过程处于失控状态。表 20.1 显示我们有可能得出一个正确的结论：继续进行一个处于控制状态的过程和调整一个处于失控状态的过程。但是，和其他假设检验程序一样，我们有可能发生两类错误：第一类错误（调整一个已经处于控制状态的过程）和第二类错误(允许一个处于失控状态的过程继续）。

过程控制程序与本书较早部分所讨论过的假设检验程序密切相关。在本质上，控制图为处于控制状态过程这个假设提供了一个不断发展的检验。

表 20.1　统计管理控制的结果

		生产过程的状态	
		假设 H_0 为真 过程在控制状态	**假设 H_0 为假** 过程在失控状态
决定	**继续生产过程**	正确的决定	第二类错误 （允许失控的生产过程得以继续）
	调整生产过程	第一类错误 （调整控制状态下的过程）	纠正决定

控制图

控制图(control chart)对确定产品中的质量差异是来源于特定因素（失控状态）还是来源于一般因素（控制状态）奠定了基础。无论任何时候检测到失控状态，我们都可以进行调整和/或采取其他纠正措施，使过程重新回到控制状态。

根据用连续尺度测量所得的数据而构造的控制图称为变量控制图，$\bar{x}$ 控制图即为一个变量控制图。

控制图可以根据它所包含的数据类型进行分类。如果根据一个比如长度、重量或温度等的变量来测量产品质量，那我们可以使用**$\bar{x}$ 控制图**($\bar{x}$ chart)。在这种情况下，根据在产品样本中所发现的均值来确定是继续还是调整生产过程。为了对所有控制图介绍一些一般性的概念，我们可以先考察$\bar{x}$ 控制图的一些特征。

图 20.1 显示的是$\bar{x}$ 控制图的一般结构。控制图的中线代表的是**控制**过程的均值，垂直线代表的是所感兴趣的变量的测量尺度。每次我们可以从生产过程中抽取一个样本，计算出样本的均值$\bar{x}$，然后将代表$\bar{x}$ 值的数据点标在控制图上。

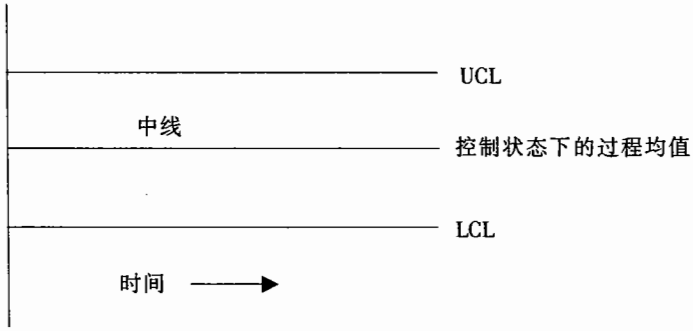


图 20.1　$\bar{x}$ 控制图的一般结构

以 UCL 和 LCL 为标志的两条线对确定过程是处于控制中还是失控是很重要的，这两条线分别被称为**控制上限**和**控制下限**。选择它们可以使得当过程处于控制状态时，$\bar{x}$ 的数值位于上下控制限之间的概率很大。位于控制限之外的数值可以给出明显的统计证据来表明过程已经失控，我们应该采取纠正措施。

随着时间的积累，越来越多的数据点将被添加到控制图上。数据点的顺序为从左向右，相同于样本过程抽取的顺序。在本质上，每当有一个点被添加到控制图上，都代表着我们在进行一次假设检验以确定过程是否处于控制状态。

除了$\bar{x}$ 控制图之外，可供使用的其他控制图还有：监测样本中测量范围的 ***R* 控制图**(*R* chart)，监测样本中缺陷比率的 ***p* 控制图**(*p* chart)，检测样本中缺陷项数目的 ***np* 控制图**(*np*

chart)。在每种情况下，控制图的基本结构都类似于图 20.1 $\bar{x}$ 控制图的格式，都有一条 LCL，一条中线，一条 UCL。这些控制图的主要区别在于所使用的垂直测量尺度不一样，比如在 p 控制图中，垂直测量尺度为样本中有缺陷项目的比率，而不是样本均值。在接下来的讨论中，我们将阐述 $\bar{x}$ 控制图、R 控制图、p 控制图以及 np 控制图的结构和使用。

$\bar{x}$ 控制图：已知过程均值和标准差

为了阐述 $\bar{x}$ 控制图的结构，让我们来参考 KJW 包装公司的情形。该公司经营一条填装谷物的流水生产线。首先假设 KJW 公司知道生产过程运转正常，即系统处于控制状态，平均填装重量为 $\mu = 16.05$ 盎司，并且过程标准差为 $\sigma = 0.10$ 盎司。除此之外，假设填装重量是正态分布的。分布图如图 20.2 所示。

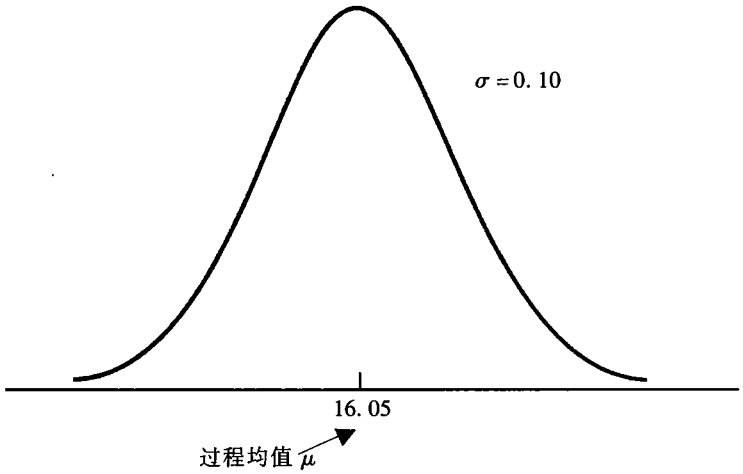


图 20.2　谷物填装重量的分布图

在第 7 章里我们阐述过，$\bar{x}$ 的抽样分布可以用于确定一个在控制状态下的过程的 $\bar{x}$ 值所期望的变异。让我们首先简要复习一下 $\bar{x}$ 的抽样分布的性质。首先，$\bar{x}$ 的均值或期望值等于 μ，μ 是在生产线处于控制状态下的时候的平均装填量。对容量为 n 的样本，$\bar{x}$ 的标准差，也称为均值的标准误差，它的计算公式为

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{20.1}$$

除此之外，因为装填重量呈正态分布，$\bar{x}$ 的抽样分布对任何样本容量而言也呈正态分布。因此，$\bar{x}$ 的抽样分布是均值为 μ、标准差为 $\sigma_{\bar{x}}$ 的正态分布。图 20.3 显示了这个分布。

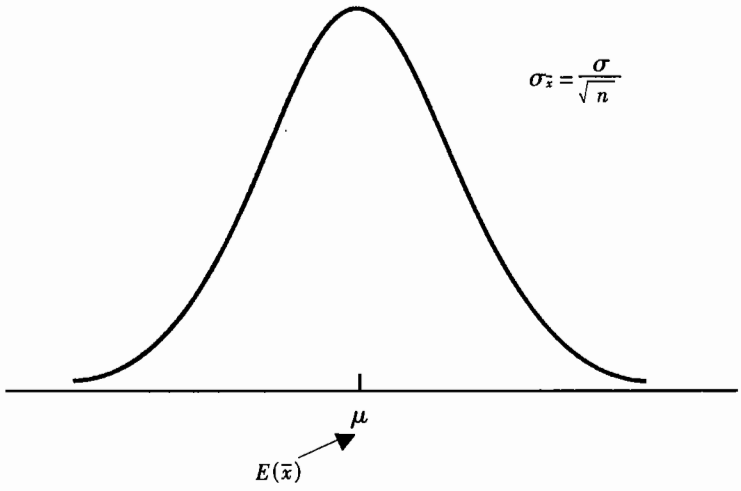


图 20.3　$\bar{x}$ 的抽样分布

如果过程处于控制状态，$\bar{x}$ 的抽样分布可用于确定什么样的$\bar{x}$ 值是合适的，质量管理的通常做法是将位于均值上下 3 个标准差之间的$\bar{x}$ 值都认为是合适的。回顾对正态分布的研究：它的近似 99.7% 的数值位于正态分布均值上下 ±3 个标准差之内，因此，如果$\bar{x}$ 的数值位于 $\mu - 3\,\sigma_{\bar{x}}$和 $\mu + 3\,\sigma_{\bar{x}}$之间，我们可以假设过程处于控制状态。作为总结，我们可以认为的控制图的控制上下限如下所示：

$\bar{x}$ 控制图的控制上下限：已知过程均值和标准差

$$\text{UCL} = \mu + 3\,\sigma_{\bar{x}} \tag{20.2}$$

$$\text{LCL} = \mu - 3\,\sigma_{\bar{x}} \tag{20.3}$$

重新考虑 KJW 包装公司的例子，图 20.2 显示了装填重量的过程分布，图 20.3 显示了的 $\bar{x}$ 抽样分布。假设有一个质量检察员定期抽取 6 箱作为样本，并且用填装重量的样本均值来确定生产过程是处于控制状态还是处于失控状态。根据公式(20.1)我们发现均值的标准差 $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 0.10/\sqrt{6} = 0.04$。因此，在均值等于 16.05 的情况下，控制限为 UCL = 16.05 + 3(0.04) = 16.17，LCL = 16.05 − 3(0.04) = 15.93。图 20.4 是以每 10 小时为观测周期所取的 10 个样本组成的控制图。为了方便阅读，控制图的下方列有样本标号 1—10。

请注意图 20.4，其中第 5 个样本的均值表明过程正处于失控状态。换句话说，第 5 个样本的均值在 LCL 下方，显示生产过程存在产品差异的可认定原因以及未填装现象的出现。作为结论，在这个点上应该进行纠正活动以使生产过程重新回到控制状态。$\bar{x}$ 控制图上的其他点均位于上下控制限之间，这个事实表明纠正活动是成功的。

$\bar{x}$ 控制图：未知过程均值和标准差

在 KJW 包装公司的例子中，我们介绍了在过程均值和标准差已知时怎样构造 $\bar{x}$ 控制图。在大

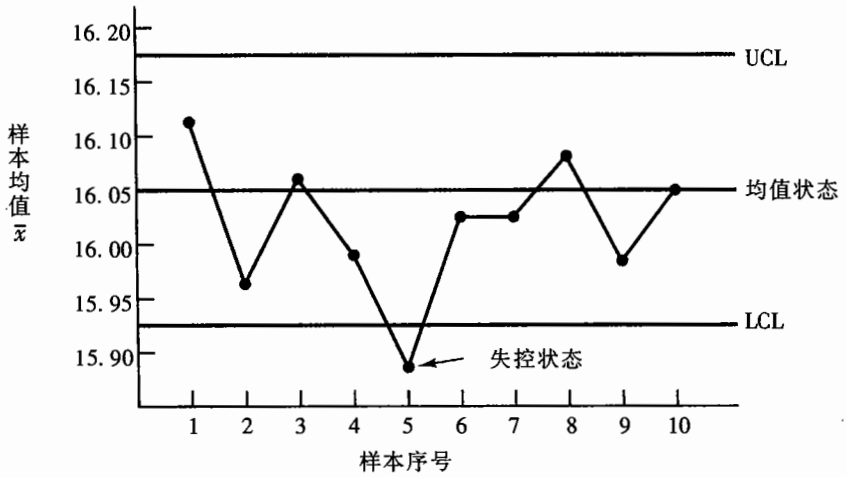


图 20.4 谷物填装过程的$\bar{x}$控制图

多数情况下，我们需要从已处于控制状态下的过程中抽取样本来估计这些数值。比如，KJW 公司在处于控制状态下的 10 天里每天上午和每天下午各选择 5 箱组成一个随即样本。对每一个子组或样本，计算它们的样本均值和标准差。然后再用样本均值和样本标准差的总体平均数构造过程均值和过程标准差的控制图。

对过程的均值和差异均保持控制是很重要的。

在实践中，我们一般使用全距而不是使用标准差来检验过程差异，这是因为全距更加易于计算。全距可以应用于构建 $\bar{x}$ 控制图的控制上下限，并且计算量很少。为了阐述控制图的构建，我们可以参考 Jensen 计算机有限公司所面临的问题。

Jensen 计算机用品供应有限公司（JCS）生产直径为 3.5 英寸的微型计算机软盘。假设在生产过程开始之后，第一个小时、第二个小时 …… 分别抽取由 5 张软盘所组成的随机样本，直到抽取了 20 个样本为止。表 20.2 给出了每一个样本中磁盘的直径和样本均值 $\bar{x}_j$ 以及全距 R_j。

表 20.2 JENSEN 计算机用品供应公司示例数据


CD 光盘数据
JENSEN

样本序号	观测值					样本均值 $\bar{x}_j$	样本全距 R_j
1	3.505 6	3.508 6	3.514 4	3.500 9	3.503 0	3.506 5	0.013 5
2	3.488 2	3.508 5	3.488 4	3.525 0	3.503 1	3.502 6	0.036 8
3	3.489 7	3.489 8	3.499 5	3.513 0	3.496 9	3.497 8	0.023 3
4	3.515 3	3.512 0	3.498 9	3.490 0	3.483 7	3.500 0	0.031 6
5	3.505 9	3.511 3	3.501 1	3.477 3	3.480 1	3.495 1	0.034 0
6	3.497 7	3.496 1	3.505 0	3.501 4	3.506 0	3.501 2	0.009 9
7	3.491 0	3.491 3	3.497 6	3.483 1	3.504 4	3.493 5	0.021 3
8	3.499 1	3.485 3	3.483 0	3.508 3	3.509 4	3.497 0	0.026 4
9	3.509 9	3.516 2	3.522 8	3.495 8	3.500 4	3.509 0	0.027 0

（续表）

样本序号	观测值					样本均值 $\bar{x}_j$	样本全距 R_j
10	3.488 0	3.501 5	3.509 4	3.510 2	3.514 6	3.504 7	0.026 6
11	3.488 1	3.488 7	3.514 1	3.517 5	3.486 3	3.498 9	0.031 2
12	3.504 3	3.486 7	3.494 6	3.501 8	3.478 4	3.493 2	0.025 9
13	3.504 3	3.476 9	3.494 4	3.501 4	3.490 4	3.493 5	0.027 4
14	3.500 4	3.503 0	3.508 2	3.504 5	3.523 4	3.507 9	0.023 0
15	3.484 6	3.493 8	3.506 5	3.508 9	3.501 1	3.499 0	0.024 3
16	3.514 5	3.483 2	3.518 8	3.493 5	3.498 9	3.501 8	0.035 6
17	3.500 4	3.504 2	3.495 4	3.502 0	3.488 9	3.498 2	0.015 3
18	3.495 9	3.482 3	3.496 4	3.508 2	3.487 1	3.494 0	0.025 9
19	3.487 8	3.486 4	3.496 0	3.507 0	3.498 4	3.495 1	0.020 6
20	3.496 9	3.514 4	3.505 3	3.498 5	3.488 5	3.500 7	0.025 9

过程均值的估计值 μ 由总体样本均值给出。

总体样本均值

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \tag{20.4}$$

式中　$\bar{x}_j$——第 j 个样本的均值，$j = 1, 2, \cdots, k$；

k——样本序号。

对表 20.2 中的 JCS 数据，总体样本均值 $\bar{\bar{x}} = 3.499\ 5$。这个值即为 $\bar{x}$ 控制图的中线。每个样本的全距，以 R_j 表示，即每一个样本里面的最大值和最小值之间的差。平均全距如下所示：

平均全距

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \tag{20.5}$$

式中　R_j——第 j 个样本的全距，$j = 1, 2, \ldots, k$；

k——样本序号。

对表 20.2 中的 JCS 数据，平均全距 $\bar{R} = 0.025\ 3$。

在前面的章节中，我们阐述了控制图的上下限，它们为

$$\bar{x} \pm 3\frac{\sigma}{\sqrt{n}} \tag{20.6}$$

因此，为了构造 $\bar{x}$ 控制图的控制限，我们需要估计过程的标准差 σ 的值，σ 的估计值可以通过分析全距数据得出。

已经证明：过程标准差 σ 的一个估计量为平均全距除以 d_2，d_2 为一个常数，仅取决于样本大小 n。即

$$\sigma \text{ 的估计值} = \frac{\bar{R}}{d_2} \tag{20.7}$$

美国材料检验学会（ASTM）数据图表述及控制图分析的工具手册（Manual on Presentation of Data and Control Chart Analysis）给出了 d_2 的数值，见附录 B 中的表 11。例如，当 $n=5$ 时，$d_2=2.326$，因此 σ 的估计值为平均全距除以 2.326。如果我们将 $\sigma=\dfrac{\overline{R}}{d_2}$ 代入公式(20.6)，则 $\bar{x}$ 控制图的控制限可以写成如下形式：

$$\bar{\bar{x}} \pm 3\frac{\overline{R}/d_2}{\sqrt{n}} = \bar{\bar{x}} \pm \frac{3}{d_2\sqrt{n}}\overline{R} = \bar{\bar{x}} \pm A_2\overline{R} \qquad (20.8)$$

请注意，$A_2=3/(d_2\sqrt{n})$ 是一个仅仅取决于样本大小的常数。A_2 的数值在附录 B 的表 11 中给出。对 $n=5$，$A_2=0.577$。因此，控制图的控制限为

$$3.499\,5 \pm (0.577)(0.025\,3) = 3.499\,5 \pm 0.146$$

因此，UCL = 3.514，LCL = 3.485。

图 20.5 是 Jensen 计算机用品供应公司示例的 $\bar{x}$ 控制图。我们用表 20.2 中的数据和 Minitab 控制图来构建它。中心线为总体样本均值 $\bar{\bar{x}}=3.499$。控制上限(UCL)为 3.514。Minitab 用标记 3.0SL 来表示 UCL，即 $\bar{\bar{x}}$ 上方的 3 个 σ 限或 3 个标准差(SL)。控制下限（LCL）为 3.485，记为 −3.0SL 或 $\bar{\bar{x}}$ 下方 3 个 σ 限。$\bar{x}$ 控制图显示了 20 个样本均值随时间变动的散布情况。由于所有 20 个样本均值均在控制限范围之内，这意味着 Jensen 公司的生产过程是处于控制中的。这个控制图现在可以用来在不断发展的基础上监控过程均值。

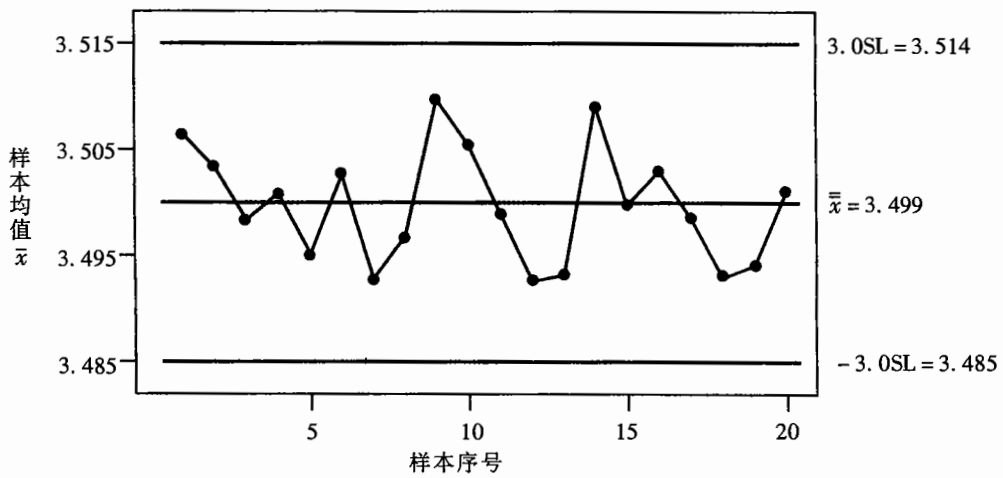


图 20.5　Jensen 计算机用品供应公司示例的 $\bar{x}$ 控制图

R 控制图

我们现在考察全距控制图（R 控制图），用它可以控制一个过程的差异。为了建立 R 控制图，我们有必要将样本的全距作为一个自带均值和标准差的随机变量来考察。平均全距 $\overline{R}$ 给出了这个随机变量的均值的一个估计值。除此之外，可以证明全距标准差的估计值为

$$\hat{\sigma}_R = d_3 \frac{\overline{R}}{d_2} \tag{20.9}$$

式中，d_2 和 d_3 是仅取决于样本容量的常数；在附录 B 的表 11 中我们给出了 d_2, d_3 的值。因此，R 控制图的 UCL 由下式给出

$$\overline{R} + 3\hat{\sigma}_R = \overline{R}(1 + 3\frac{d_3}{d_2}) \tag{20.10}$$

LCL 为

$$\overline{R} - 3\hat{\sigma}_R = \overline{R}(1 - 3\frac{d_3}{d_2}) \tag{20.11}$$

如果我们令

$$D_4 = 1 + 3\frac{d_3}{d_2} \tag{20.12}$$

$$D_3 = 1 - 3\frac{d_3}{d_2} \tag{20.13}$$

对 R 控制图，我们可以将其上下限表示为

$$\text{UCL} = \overline{R} D_4 \tag{20.14}$$

$$\text{LCL} = \overline{R} D_3 \tag{20.15}$$

D_3 和 D_4 的数值也在附录 B 的表 11 中给出。$n = 5$, $D_3 = 0$, $D_4 = 2.115$，因此，根据 $\overline{R} = 0.0253$，控制限为

$$\text{UCL} = 0.025\,3(2.115) = 0.053\,5$$

$$\text{LCL} = 0.025\,3(0) = 0$$

图 20.6 显示的是 Jensen 计算机用品供应公司示例的 R 控制图。我们利用表 20.2 中的数据和 Minitab 的控制图来构建它。控制图的中心线为 20 个样本全距的总体平均数 $\overline{R} = 0.025\,27$。UCL 为 0.053 44 或者 $\overline{R}$ 上方 3 个 σ 限(3.0SL)。LCL 为 0.0 或者 $\overline{R}$ 下方 3 个 σ 限。R 控制图显示了 20 个

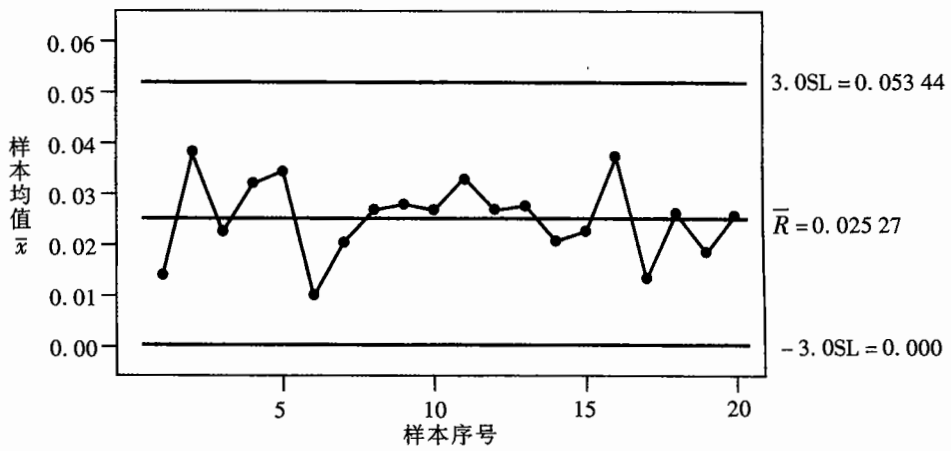


图 20.6　Jensen 计算机用品供应公司示例的 R 控制图

样本的全距随时间变动而变动的分布情形。因为所有 20 个样本全距均在控制限内部，我们可以确定这个生产过程在抽样期间保持在控制状态下。

如果 R 控制图显示过程处于失控状态，那么控制图将无法得到解释，除非 R 控制图表明过程差异处于控制状态。

---

p 控制图

现在让我们来考察怎样通过检测产品是否有缺陷来确定产品质量的情形。是继续生产过程还是调整生产过程取决于$\bar{p}$ 的值，$\bar{p}$ 值指的是在样本中所发现的缺陷产品的比例。根据缺陷产品比例构造的控制图称为 p 控制图。

根据显示有缺陷的数据或有缺陷的数量的比例而构造的控制图称为属性控制图，p 控制图即为属性控制图。

为了阐述 p 控制图的结构，让我们来考察邮局使用邮件自动分类处理机的情况。这些自动化机器扫描邮件上的邮政编码，然后将它们分发到各自合适的邮递路线中去。但是即使机器运转正常，依然有一些邮件被投入错误的邮递线路中。假设当机器运转正常时，或机器处于控制状态时，有 3% 的邮件被误投。因此，在控制状态下错误投递的信件比例 p 为 0.03。

为检验过程差异，我们可以构造样本标准差控制图（s 控制图）替代 R 控制图。如果样本容量为 10 或更小，R 控制图和 s 控制图将给出相似的结果；如果样本容量大于 10，我们倾向于使用 s 控制图。

$\bar{p}$ 的抽样分布，正如第 7 章所述，可以被用于确定在一个处于控制状态的过程中$\bar{p}$ 值所期望的变异。回顾$\bar{p}$ 的期望值或均值，它等于 p，即处于控制状态下的过程的缺陷产品比例。已知样本容量为 n，则 $\bar{p}$ 的标准差，又称为比例的标准误差，可由下式计算：

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{20.16}$$

在第 7 章中我们提到，当样本容量足够大的时候，$\bar{p}$ 的抽样分布可以近似于正态分布。对于$\bar{p}$，当以下两个条件得到满足时，我们可以认为样本容量足够大：

$$np \geqslant 5$$
$$n(1-p) \geqslant 5$$

总的来说，当样本容量足够大时，$\bar{p}$ 的抽样分布近似于均值为 p，标准差为$\sigma_{\bar{p}}$ 的正态分布。图 20.7 显示了这一分布。

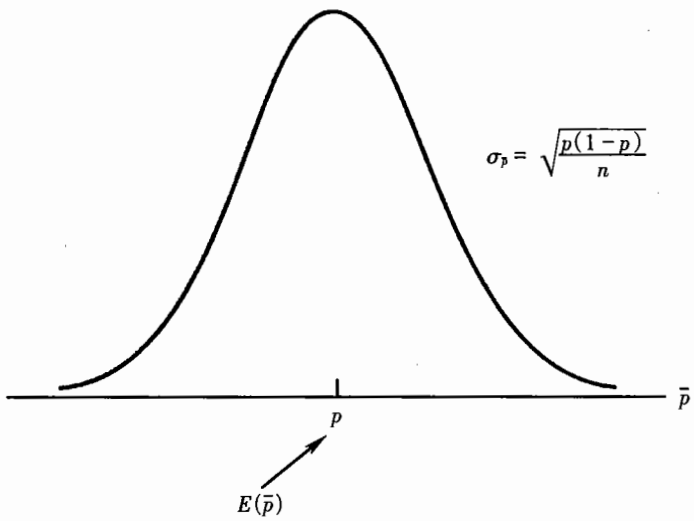


图 20.7　$\bar{p}$ 的抽样分布图

为了给 p 控制图建立控制限，我们遵循给$\bar{x}$ 控制图建立控制限的同样步骤。即，当过程处于控制状态时，控制图的控制限为缺陷比例的上下 3 个标准差或标准误差，因此，我们可以得到下面的控制限：

p 控制图的控制限

$$\text{UCL} = p + 3\sigma_{\bar{p}} \tag{20.17}$$

$$\text{LCL} = p - 3\sigma_{\bar{p}} \tag{20.18}$$

已知 $p = 0.03$ 和样本容量 $n = 200$，由公式(20.16)可得标准误差为

$$\sigma_{\bar{p}} = \sqrt{\frac{0.03(1-0.03)}{200}} = 0.012\,1$$

因此，控制限为 UCL = 0.03 + 3(0.012 1) = 0.066 3，LCL = 0.03 - 3(0.012 1) = -0.006 3。由于 LCL 为负数，因此在控制图中被设为零。

图 20.8 是邮件分拣过程的控制图。所标出的点为从过程中抽取的信件样本中发现的有缺陷的样本比率，由于所有的点都在控制限之内，因此没有任何证据表明分拣过程处于失控状态。事实上，p 控制图表明该过程处于控制状态下并且应该继续运转。

如果无法知道一个处于控制状态下的过程的缺陷产品比例，那么，我们应该通过样本数据来估计它。例如，假设从一个处于控制状态的过程中抽取 M 个不同的样本，每一个样本容量为 n。然后确定出每个样本中缺陷产品的比例。将收集的所有资料看成是一个大样本，我们可以确定出所有数据的平均缺陷产品比例，这个数值即可用于确定 p 的估计值，p 的估计值指的是在控制状态下的过程的缺陷产品比例。请注意，p 的估计值同时也有助于我们去估计比率的标准误差。然后我们可以建立控制上下限。

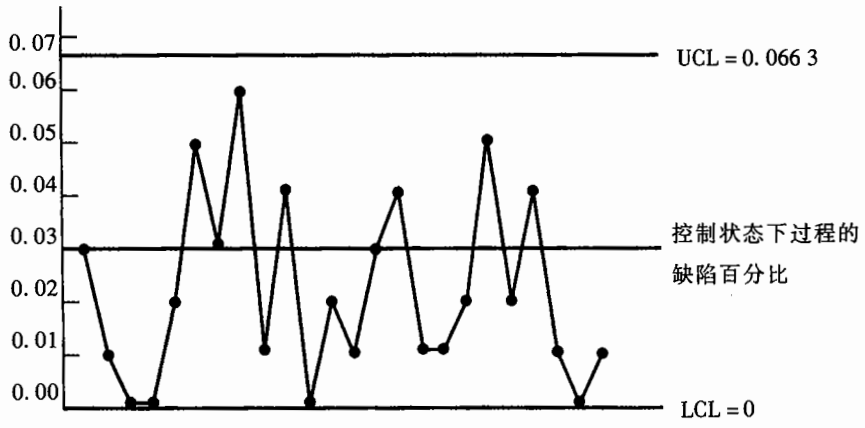


图 20.8 邮件分拣过程的缺陷产品比率的 p 控制图

np 控制图

np 控制图是对样本中缺陷产品数量而构造的控制图。在这里面，n 是样本容量，p 是当过程处于控制状态下时所发现的缺陷产品的概率。不管样本容量有多大，当 $np \geqslant 5$、$n(1-p) \geqslant 5$ 时，在一个样本容量为 n 的样本中所发现的缺陷产品数量的分布可以近似于均值为 np，标准方差为 $\sqrt{np(1-p)}$ 的正态分布。因此，对邮件分拣的例子，$n=200$，$p=0.03$，在一个 200 封信的样本中所发现的错误投递信件的数量可以近似于一个均值为 $200(0.03)=6$，标准方差为 $\sqrt{200(0.03)(0.97)}=2.412\,5$ 的正态分布。

当过程处于控制状态下时，np 控制图控制限为所观测到的缺陷产品平均数量的上下 3 个标准差。因此，我们可以得到如下所示的控制限：

np 控制图的控制限

$$\text{UCL} = np + 3\sqrt{np(1-p)} \tag{20.19}$$

$$\text{LCL} = np - 3\sqrt{np(1-p)} \tag{20.20}$$

对邮件分拣过程的例子，$p=0.03$，$n=200$，控制限为 $\text{UCL}=6+3\,(2.412\,5)=13.237\,5$，$\text{LCL}=6-3(2.412\,5)=-1.237\,5$。当 LCL 为负时，LCL 在控制图中设为零。因此，如果被错误投递到别的线路的信件数量多于 13 时，我们可以断定过程已经失控。

np 控制图所提供的信息等同于 p 控制图所提供的信息；惟一的区别在于 np 控制图是所观测到的缺陷产品数量的分布图，而 p 控制图是所观测到的缺陷产品比率的分布图。因此，如果我们在 p 控制图的基础上可以断定某个特定过程处于失控状态，则我们也可以在 np 控制图的基础上断定该过程失控。

控制图的解释

根据控制图中的点的位置和分布轨迹，我们可以确定一个过程是否处于统计控制状态之中，而且判断错误概率很小。过程有可能处于失控状态的最初迹象是数据点落到控制限之外，例如图 20.4 中的点 5。发现这样的数据点是在统计上证明过程失控的证据；在这些情况下，我们应该立即采取纠正措施。

控制图可以识别可指明原因的变化。管理人员必须授权进行消除可指明原因的纠错行动，以将生产过程带回控制状态。

除了数据点在控制限之外这种迹象，某些在控制限之内的数据点的分布轨迹也可以是质量控制错误的预警信号。例如，假设所有的数据点均在控制限之内，但是有大量的数据点处于控制中心线的一侧，这种分布就有可能表明设备出问题了，或者原材料变化，或者出现了其他质量上可指出的变化。我们应该对生产过程进行仔细的调查以确定质量是否发生了变化。

另一个观测控制图的分布轨迹是某些因素随时间推移而发生的逐渐变化或趋势。例如工具的磨损、制造零件的尺寸逐渐偏离设计水平、温度或湿度的逐渐变化、设备的整体性老化、污垢的堆积、操作员的疲劳，这些都有可能反映控制图中的变化趋势。如果一行中有六七个数据点呈现增长或下降趋势，就应该引起关注，即使这些数据点均处于控制限之内。当这种情况发生时，我们应该检查生产过程以防备质量上可能的变化或转变，同时有必要采取纠正措施将过程带回到控制状态下。

即使所有的数据点都在上下控制限之内，过程仍然有可能处于失控状态。样本数据点的趋势或在中心线上下的长期趋势也可以显示过程处于失控状态。

评　注

1. $\bar{x}$ 控制图的控制限取决于平均全距的数值，这些控制限没有很多意义，除非过程变异处于控制状态之中。在实践中，R 控制图通常在 $\bar{x}$ 控制图前构造；如果 R 控制图表明控制变异处于控制下，然后再去构造 $\bar{x}$ 控制图。Minitab 的 X－R 选项可以同时给出 $\bar{x}$ 控制图和 R 控制图。在附录 20.1 中描述了这些步骤的过程。
2. np 控制图可用于在过程中监控有缺陷产品的数量。摩托罗拉的 6 个 σ 质量水平法则确立了缺陷产品数量的比率不超过 3.4×10^{-6} 的目标（*American Production and Inventory Control*, July 1991）。这个目标意味着 $p=0.000\,003\,4$。

练　习

方　法

1. 某过程处于控制状态下，均值 $\mu=12.5$，标准差为 $\sigma=0.8$。
   a. 如果样本容量为 4，试构造 $\bar{x}$ 控制图。
   b. 如果样本容量为 8 和 16，试构造 $\bar{x}$ 控制图。
   c. 当样本容量增加时，控制图的控制限将会有什么变化？试讨论其合理性。

2. 从一个处于控制状态下的过程中抽取 25 个样本，每个样本容量均为 5，所有收集的数据总计为 677.5 磅。
   a. 当该过程处于控制状态时，试求过程均值的估计值(以磅为单位)。
   b. 假设在该过程处于控制状态时，过程标准差为 0.5，过程均值为(a)中所估计的结果，如果样本容量为 5，试建立该过程的控制图。

3. 某过程目前运转比较令人满意，检查其中 25 个样本，每个样本有 100 项，在这 25 个样本中，总共有 135 项有缺陷。
   a. 当该过程处于控制状态时，试求缺陷产品项的估计值。
   b. 如果样本容量为 100，试求该统计过程的比率的标准误差。
   c. 计算该控制图的上下控制限。

4. 自测题 从一个过程中抽取 20 个样本，每个样本容量为 8，结果为 $\bar{\bar{x}}=28.5$，$\bar{R}=1.6$，试计算该过程的 $\bar{x}$ 控制图和 R 控制图的上下控制限。

应　用

5. 温度测量某个生产过程的产品，当该过程处于控制状态时，过程均值 $\mu=128.5$，标准差 $\sigma=0.4$。
   a. 如果样本容量为 6，试构造 $\bar{x}$ 控制图。
   b. 利用下面给出的样本值，判断该过程是否处于控制状态。

   128.8　128.2　129.1　128.7　128.4　129.2

   c. 利用下面给出的样本值，判断该过程是否处于控制状态。

   129.3　128.7　128.6　129.2　129.5　129.0

6. 一个质量管理过程监控每箱的洗涤剂重量，控制限设为 UCL = 20.12 盎司，LCL = 19.90 盎司，对于抽样和抽验过程，样本容量为 5。试求该制造过程的过程均值和过程标准差。

7. Goodman 轮胎和橡胶公司定期在模拟道路情形下检验它的轮胎磨损情况，为了研究和控制生产过程，从不同班组中抽取 20 个样本，每个样本包含 3 个轮胎，抽样数据如下。假设这些数据是在生产过程处于控制状态下时抽取的，试建立 R 和 $\bar{x}$ 控制图。


CD 光盘数据
Tires

样本	轮胎磨损*		
1	31	42	28
2	26	18	35
3	25	30	34
4	17	25	21
5	38	29	35
6	41	42	36
7	21	17	29
8	32	26	28
9	41	34	33
10	29	17	30
11	26	31	40
12	23	19	25
13	17	24	32
14	43	35	17
15	18	25	29
16	30	42	31
17	28	36	32
18	40	29	31
19	18	29	28
20	22	34	26

* 单位：百分之一英寸。

8. 在某个生产过程正常运转或出于控制状态下的几个星期中，抽取了 20 个样本检验断裂强度，每个样本包含 150 包合成材料网球线，在对这 3 000 包产品检验中，总共有 141 包没有达到制造商的要求。
   a. 当该系统处于控制状态下，缺陷产品比率的估计值是多少？
   b. 计算 p 控制图的上下限。
   c. 如果检验 150 包产品，发现其中 12 包产品有缺陷，根据(b)中的计算结果，对这个过程应该得出什么推论？
   d. 计算 np 控制图的上下控制限。
   e. 利用（d）中的结果回答（c）中的问题。
   f. 在这种情况下应该倾向于使用什么样的控制图？解释原因。

9. 汽车厂商为几种型号的汽车提供活塞产品，在生产过程处于控制状态下时，选择了 20 个样本，每个样本包含 200 个活塞，在样本中所发现的缺陷产品的数量如下所示：

8	10	6	4	5	7	8	12	8	15
14	10	10	7	5	8	6	10	4	8

a. 活塞的生产过程处于控制状态时，缺陷产品的比率的估计值是多少？

b. 假设每个样本有200个活塞，建立活塞生产过程的 p 控制图。

c. 利用(b)中的结果，假如在一个有200个活塞的样本中，发现有20个产品有缺陷，可以得出什么结论？

d. 计算 np 控制图的上下控制限。

e. 利用(d)中的结果回答（c）中的问题。

20.2 抽样验收

在抽样验收中，我们需要研究的项目可能是尚在装运的原材料或采购的产品部件，或者是从最后生产线上下来的成品。假设我们需要在特定产品质量特性的基础上决定是否接受或拒绝一组产品项目。在质量管理术语上，一组项目称为**批**(lot)，**抽样验收**(acceptance sampling)就是在对一批项目的样本进行检查的基础上，然后再作出接受或拒绝决定的统计方法。

抽样验收的一般步骤如图20.9所示。接受一批产品后，从中抽取样本以供检查。将抽样的结果与特定质量特性进行比较。如果这些质量特性得到满足，这批产品就被接受，并送往生产线或装船运送给客户。如果该批产品被拒绝，管理人员必须决定对它作出合理的处置：在一些情况下，可能决定保留该批产品，但是应该将无法接受的或者失效的产品剔除掉；在另一些情况下，这批产品被退回供货商处，且所有费用均由供货商支付，这批额外的工作量和支出将促使供货商

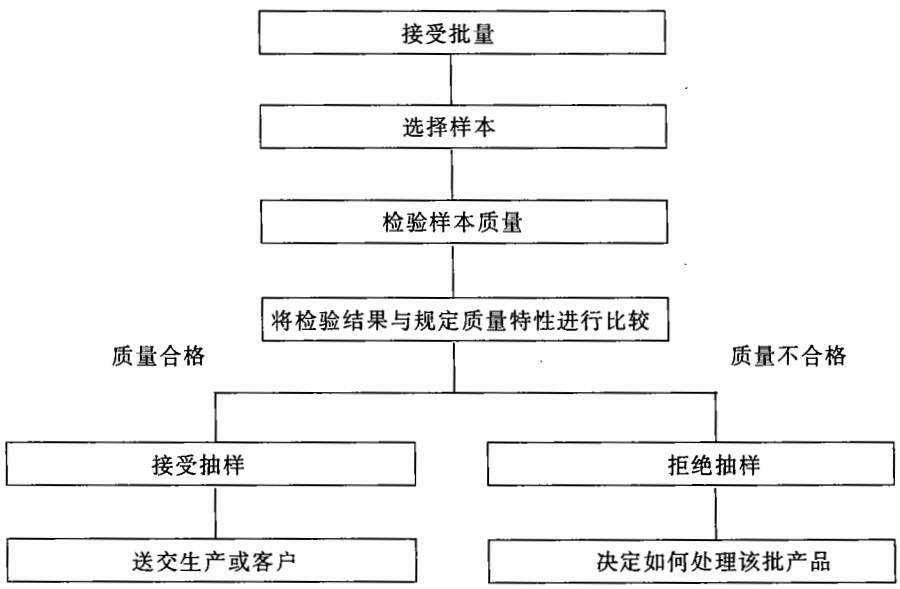


图20.9 抽样验收步骤

供应高质量的批量产品；最后，如果这些批量产品中包含产成品，那么这些产品必须废弃或重新加工以达到可以接受的质量标准。

抽样验收在 100% 监督时具有如下优势：

1. 通常费用较小；
2. 由于较少地监督和处理，因而产品损坏较少；
3. 需要较少的监督员；
4. 如果必须要利用破坏性检验时，这是惟一可行的办法。

抽样验收的统计程序以第 9 章所介绍的假设检验方法为基础。零假设与备择假设表述如下：

H_0：高质量批

H_a：低质量批

表 20.3 是假设检验程序的结果，请注意正确的结论意味着接受一批高质量产品和拒绝一批低质量产品。然而，同其他假设检验程序一样，我们应该知道有可能存在第一类错误（拒绝一批高质量产品）和第二类错误(接受一批低质量产品)。

表 20.3　抽样验收结果

		批量产品状态	
		H_0 成立	H_0 不成立
结论	接受该批产品	正确结论	第二类错误（接受不合格质量产品）
	拒绝该批产品	第一类错误（拒绝合格质量产品）	正确结论

第一类错误对批量产品的生产者造成一定的风险，因此称为**生产者风险**(producer 's risk)。例如，若生产者风险为 0.05，那就意味着质量合格的批量产品被拒收的概率为 5%。另一方面，第二类错误的概率对批量产品的消费者造成了风险，因此被称为**消费者风险**(consumer 's risk)。例如，若消费者风险为 0.10，则表明错误接受一批低质量产品的概率为 10%，并且这批低质量产品有可能被用于生产或装运给消费者。抽样验收程序的设计者可以控制生产者风险和消费者风险的规定值。为了阐述具体做法，我们可以考察 KALI 有限公司所面临的问题。

KALI 有限公司：抽样验收实例

KALI 有限公司是一家家庭用品制造商，拥有一系列产品商标。然而，KALI 有限公司并不生产其产品所用的每一个部件。一些产品部件直接从供货商处采购。例如，KALI 有限公司所生产的家庭用空调的一个部件是超载保护器，它可以在空调压缩器过热时自动关闭它。如果过载保护器运转不正常，空调压缩器有可能被严重损坏，因此，KALI 有限公司很关注过载保护器的质量。确保质量的一个方法是测试所接受的每一个产品，这种方法被称为 100% 检验法。但是，如果要检测过载保护器的功能，必须对设备进行长时间的运行检验，检验代价昂贵。所以，KALI 有限公司不可能检测它所接受的每一个过载保护器。

KALI 有限公司采用抽样验收的方法来监控过载保护器的质量，抽样验收的方法需要 KALI 有限公司的质量监控员从每一批接受的产品中抽取样本并检验。如果在抽样产品中仅仅检测到极少的缺陷产品，则该批量产品质量合格可以验收。相反，如果在样本中发现大量的缺陷产品，则该批量产品有可能质量低下而被拒收。

抽样验收方案包含样本容量 n 和验收准则 c。**验收准则**(acceptance criterion)指的是在样本中所发现的缺陷产品的最大数量，在该数量下样本依然是可以接受的。例如，对 KALI 有限公司，假设我们从刚刚运抵的产品中每次抽取 15 件组成样本，除此之外，质量管理人员声明仅在没有发现缺陷产品的情况下才接受该批产品。在此种情形中，质量管理人员所建立的抽样验收方案即为 $n=15$, $c=0$。

这个抽样验收方案对质量监控人员来说是很容易执行的。质检员只需要抽取 15 件产品组成样本，对每一件产品进行检测，并根据以下的决策法则得出结论即可：

- 如果没有发现缺陷产品则**接受该批**产品。
- 如果发现一个或多个缺陷产品则**拒绝接受该批**产品。

在执行抽样验收方案之前，质量管理人员需要估计该方案的风险或出错概率。只有在生产者风险（第一类风险）和消费者风险（第二类风险）均控制在合理的水平时，这个方案才能得到执行。

计算批量验收的概率

分析生产者风险和消费者风险的关键之处在于“What - if ?”类型的分析，即我们假设一批产品中缺陷产品的百分比，并在给定样本验收方案下计算接受该批产品的概率。通过改变缺陷产品的假设百分比，我们可以检验抽样方案关于两类风险的效果。

假设我们已经接受到大批量的过载保护器，而且其中有 5% 的过载保护器质量上有缺陷。在 $n=15$，$c=0$ 的抽样验收方案下，我们接受该批产品的概率有多大？由于每个过载保护器被检测的结果要么是有缺陷，要么是没有缺陷，而且产品批量的容量很大，因此，对于容量为 15 的样本，其有缺陷产品的数量服从**二项概率分布**。二项概率函数在第 5 章中介绍过，形式如下所示：

抽样验收的二项概率函数

$$f(x)=\frac{n!}{x!(n-x)!}p^{x}(1-p)^{(n-x)} \tag{20.21}$$

式中　n——样本容量；

p——批量中缺陷产品的比率；

x——样本中缺陷产品的数量；

$f(x)$——样本中有 x 个缺陷产品的概率。

对于 KALI 有限公司的抽样验收方案，$n=15$; 因此，对有 5% 缺陷产品的批量（$p=0.05$），我们可以得到

$$f(x)=\frac{15!}{x!(15-x)!}(0.05)^{x}(1-0.05)^{(15-x)} \tag{20.22}$$

利用公式(20.22)，$f(0)$将给出过载保护器有缺陷产品数量为零并被接受的概率。在使用公式(20.22)中，有 0! = 1，因此

$$\begin{aligned}f(0)&=\frac{15!}{0!(15-0)!}(0.05)^{0}(1-0.05)^{(15-0)}\\&=\frac{15!}{0!(15)!}(0.05)^{0}(0.95)^{15}\\&=(0.95)^{15}\\&=0.463\,3\end{aligned}$$

我们现在知道，在 $n=15$, $c=0$ 的抽样方案中，对有缺陷产品的概率为 5% 的批量的接受概率为 0.463 3。因此，对有 5% 缺陷产品的批量的拒绝概率必定为相应的 1 − 0.463 3 = 0.536 7。

利用二项概率的计算表(见附录 B 的表 5)，可以在确定批量接受的概率时减少很多计算工作量。表 20.4 中列出了 $n=15$ 和 $n=20$ 的部分二项概率，利用这个表，如果对缺陷产品为 10% 的批量和 $n=15$, $c=0$ 的抽样验收方案，接受批量的概率为 0.205 9。在表 20.5 中我们给出了根据 $n=15$，$c=0$ 的抽样验收方案，接受缺陷产品比例分别为 1%，2%，3%，…… 的批量的概率。

根据表 20.5 中的概率数据，将批量接受的概率和有缺陷产品的比例作为因变量和自变量绘成图 20.10。这个图形，或曲线，被称为 $n=15$，$c=0$ 抽样验收方案的**抽样验收特性曲线**[(operating characteristic (OC) curve]。

也许我们应该考察其他的样本验收方案，它们有不同的样本容量 n 和/或不同的验收准则。首先考察样本容量保持 15 不变，但是验收准则从 $c=0$ 上升到 $c=1$，即我们在样本中发现的缺陷产品数为零或一个时都可以接受该批产品。对缺陷产品比例为 5% 的批量（$p=0.05$），表 20.4 显示 $n=15$, $p=0.05$, $f(0)=0.463\,3$, $f(1)=0.365\,8$。因此，对 $n=15$，$c=1$ 的方案，缺陷比例为 5% 的批量被接受的概率为 0.463 3 + 0.365 8 = 0.829 1。

继续我们在图 20.11 中的计算，图 20.11 描述了对 KALI 有限公司例子的其他 4 种抽样验收特性曲线。分别考察样本容量为 15 和 20 的样本，我们发现，不论批量中有缺陷产品的比例如

表 20.4 样本容量为 15 和 20 的部分二项概率

		p								
n	x	.01	.02	.03	.04	.05	.10	.15	.20	.25
15	0	.860 1	.738 6	.633 3	.542 1	.463 3	.205 9	.087 4	.035 2	.013 4
	1	.130 3	.226 1	.293 8	.338 8	.365 8	.343 2	.231 2	.131 9	.066 8
	2	.009 2	.032 3	.063 6	.098 8	.134 8	.266 9	.285 6	.230 9	.155 9
	3	.000 4	.002 9	.008 5	.017 8	.030 7	.128 5	.218 4	.250 1	.225 2
	4	.000 0	.000 2	.000 8	.002 2	.004 9	.042 8	.115 6	.187 6	.225 2
	5	.000 0	.000 0	.000 1	.000 2	.000 6	.010 5	.044 9	.103 2	.165 1
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.001 9	.013 2	.043 0	.091 7
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 3	.003 0	.013 8	.039 3
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 5	.003 5	.013 1
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 7	.003 4
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 7
20	0	.817 9	.667 6	.543 8	.442 0	.358 5	.121 6	.038 8	.011 5	.003 2
	1	.165 2	.272 5	.336 4	.368 3	.377 4	.270 2	.136 8	.057 6	.021 1
	2	.015 9	.052 8	.098 8	.145 8	.188 7	.285 2	.229 3	.136 9	.066 9
	3	.001 0	.006 5	.018 3	.036 4	.059 6	.190 1	.242 8	.205 4	.133 9
	4	.000 0	.000 6	.002 4	.006 5	.013 3	.089 8	.182 1	.218 2	.189 7
	5	.000 0	.000 0	.000 2	.000 9	.002 2	.031 9	.102 8	.174 6	.202 3
	6	.000 0	.000 0	.000 0	.000 1	.000 3	.008 9	.045 4	.109 1	.168 6
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.002 0	.016 0	.054 5	.112 4
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 4	.004 6	.022 2	.060 9
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.001 1	.007 4	.027 1
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 2	.002 0	.009 9
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 5	.003 0
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 8

何，$n=15$, $c=1$ 抽样方案给出的接受这批产品的概率最大；$n=20$, $c=0$ 的抽样方案给出的接受概率最小，但该方案却给出了拒绝这批产品的最大概率。

表 20.5 $n=15$, $c=0$ 的抽样验收方案对 KALI 问题的概率

批量中的缺陷百分比	接受批量的概率
1	0.860 1
2	0.738 6
3	0.633 3
4	0.542 1
5	0.463 3
10	0.205 9
15	0.087 4
20	0.035 2
25	0.013 4

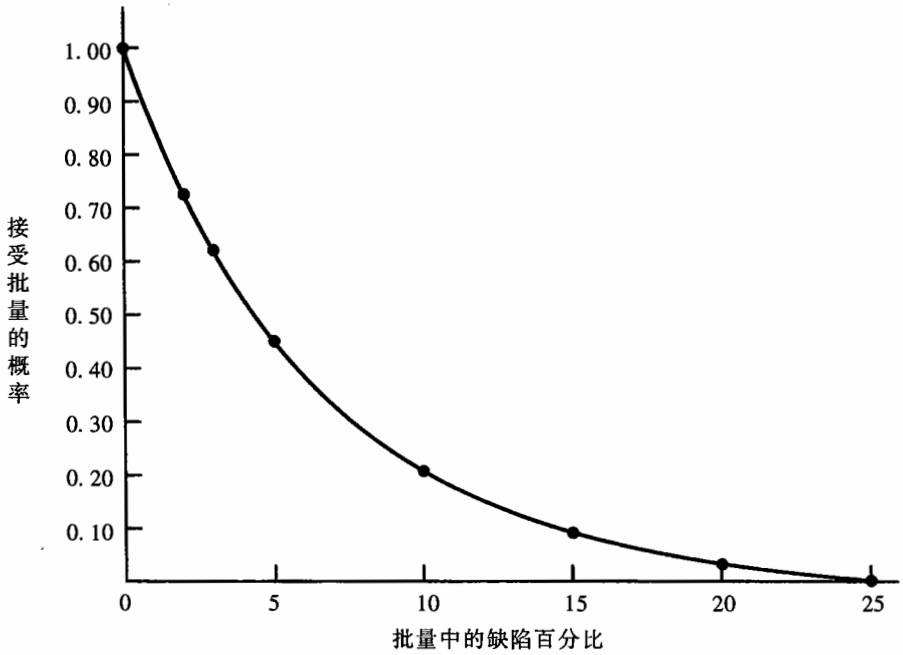


图 20.10　$n = 15, c = 0$ 的抽样验收方案的特性曲线

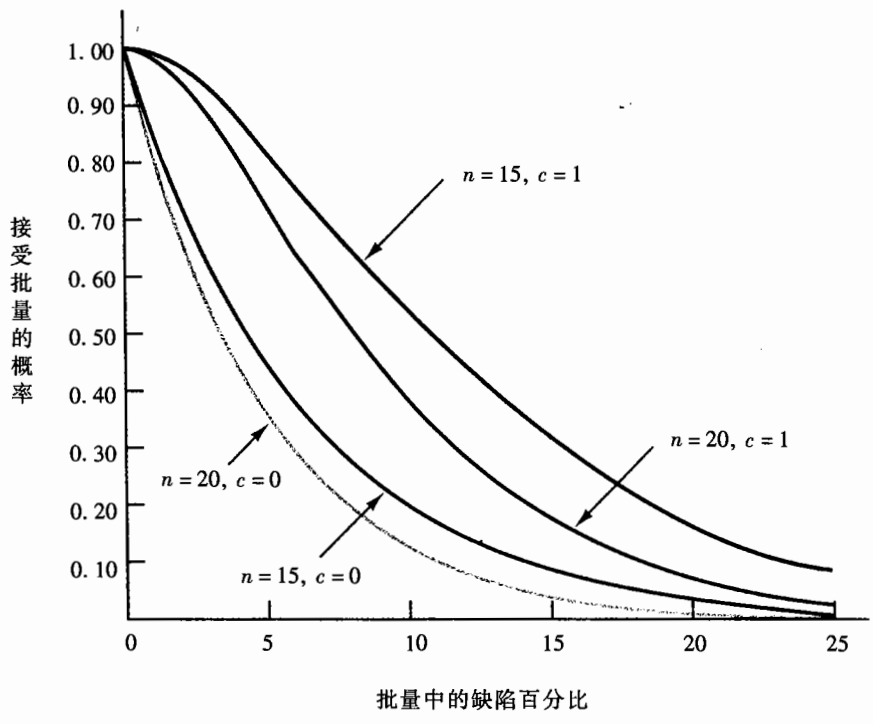


图 20.11　4 种抽样验收方案的特性曲线

选择抽样验收方案

现在我们已经掌握如何使用二项概率分布，根据一个给定的缺陷产品比例来计算接受批量的概率，那么我们可以选择 n 和 c 的数值来为实践应用确定抽样验收方案。为了做到这点，管理人员必须为批量中缺陷产品的比率指定两个变量：一个变量以 p_0 表示，用来控制生产者风险；另一个变量，以 p_1 表示，将被用于控制消费者风险。

为了阐明具体做法，我们可以使用如下标记：

α ——生产者风险；缺陷产品率为 p_0 的批量被拒收的概率；

β ——消费者风险；缺陷产品率为 p_1 的批量被拒收的概率。

对 KALI 问题，假设管理人员指定 $p_0 = 0.03$, $p_1 = 0.15$。由图 20.12 中 $n = 15$, $c = 0$ 的抽样特性曲线我们可以看到 $p_0 = 0.03$，这意味着生产者的风险近似于 $1 - 0.63 = 0.37$，$p_1 = 0.15$ 表明消费者风险近似于 0.09。因此，如果管理人员愿意接受缺陷产品率为 3% 的批量的概率为 0.37（生产者风险）和接受缺陷产品率为 15% 的批量的概率为 0.09(消费者风险)，则应该可以接受 $n = 15$ 和 $c = 0$ 的抽样验收方案。

但是，假如管理人员要求生产者风险 $\alpha = 0.10$，消费者风险 $\beta = 0.20$，我们可以看到现在 $n = 15$, $c = 0$ 的抽样验收方案在消费者风险方面比所希望的要小得多，但是在生产者风险方面却比所希望的要大得多。$\alpha = 0.37$ 的事实表明，拒收缺陷产品率为 3% 的批量的概率为 37%。由此可见生产者风险过高，我们应该考虑制定一个新的抽样验收方案。

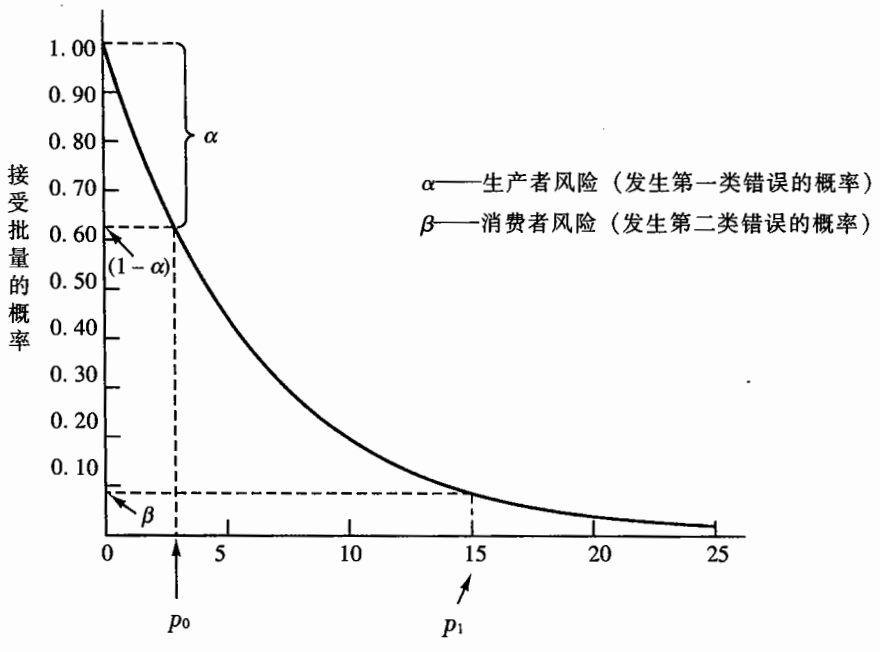


图 20.12　$n = 15$, $c = 0$, $p_0 = 0.03$, $p_1 = 0.15$ 的抽样特性曲线

根据图 20.11 中的 $p_0 = 0.03$, $\alpha = 0.10$, $p_1 = 0.15$, $\beta = 0.20$，可以表明 $n = 20$, $c = 1$ 的抽样验收方案最接近于同时满足生产者风险和消费者风险的需要。本节末的练习 13 将会要求你计算 $n = 20$, $c = 1$ 的抽样验收方案的生产者风险和消费者风险。

正如本节所介绍的，在决定抽样验收方案时，我们应该考虑一些计算或抽样验收特性曲线，并结合考虑生产者风险和消费者风险。幸运的是，已经出版了一些抽样验收表。例如美国军用标准表，MIL－STD－105D, 在设计抽样验收方案时提供了很多有用信息。在有关质量管理的更为高级的教材中描述了这些表的使用。更为高级的教材也讨论了抽样成本在确定最为适合的抽样方案中的作用。

多重抽样方案

我们对 KALI 问题所介绍的抽样验收程序是**一次抽样**方案，它被称为一次抽样方案是因为仅仅使用了一个样本或仅有一个抽样阶段。在确定了样本中缺陷产品数量后，必须作出接受或者拒绝该批产品的结论。另一种可供选择的抽样方案是**多重抽样方案**(multiple sampling plan)，其中涉及到两次或多次抽样阶段。在每一个阶段都有三种可能的结论：停止抽样并接受该批产品，停止抽样并拒收该批产品，或者继续抽样。虽然这样操作比较复杂，但是根据相同的 α、β 值，多次抽样方案所得结果的样本总容量比一次抽样方案要小。

图 20.13 是二阶段抽样方案，或者双重抽样方案的流程图。一开始选择 n_1 个产品项组成样本，如果缺陷产品数量 x_1 少于或者等于 c_1，则接受该批产品。如果 x_1 大于或者等于 c_2，则拒收该批产品。如果 x_1 位于 c_1 和 c_2 之间即 $c_1 < x_1 < c_2$，则选择 n_2 个产品项重新组成样本。确定第一份样本缺陷产品数量 x_1 和第二份样本缺陷产品数量 x_2 的总和，如果 $x_1 + x_2 \leqslant c_3$，则接受该批产品；否则拒绝该批产品。双重抽样验收方案的建立显然更加复杂，因为样本容量 n_1 和 n_2 以及接受准则的数值 c_1、c_2 和 c_3 必须同时满足生产者风险和消费者风险的双重要求。

评　注

1. 抽样验收中所使用的二项概率分布是以大批量的假设为基础的。如果样本容量较小，则应用超几何分布较为适宜。质量管理研究领域的专家认为，当样本容量 n 至少为 16，所抽样的产品总数至少为样本容量的 10 倍，并且 p 值小于 0.1 时，泊松分布是抽样验收的较为合适的分布。对于大样本容量，我们可以使用正态分布近似模拟二项概率分布。
2. 在 MIL－ST－105D 的抽样表中，p_0 被称为可接受质量水平(AQL)。在一些抽样表中，p_1 被称为极限产品缺陷率（LTPD）或拒收质量水平(RQL)。许多已经出版的抽样验收方案亦采用比如无差异质量水平（IQL）和平均交货质量限(AOQL)等。在参考文献中列出的更为高级的教材里有这些质量指数的比较深入的讨论。
3. 在本节中我们讨论了质量**抽样方案**。在这些方案中，被抽样的产品项被区分为无缺陷和有缺陷两种。在**变量抽样方案**中，我们可以

抽取一个样本并测量它的质量特性。例如，对黄金珠宝质量的测验可以是它的含金量。计算样本中黄金的平均含量并将它与一个允许的标准值进行比较，即可确定是接受还是拒绝这批产品。

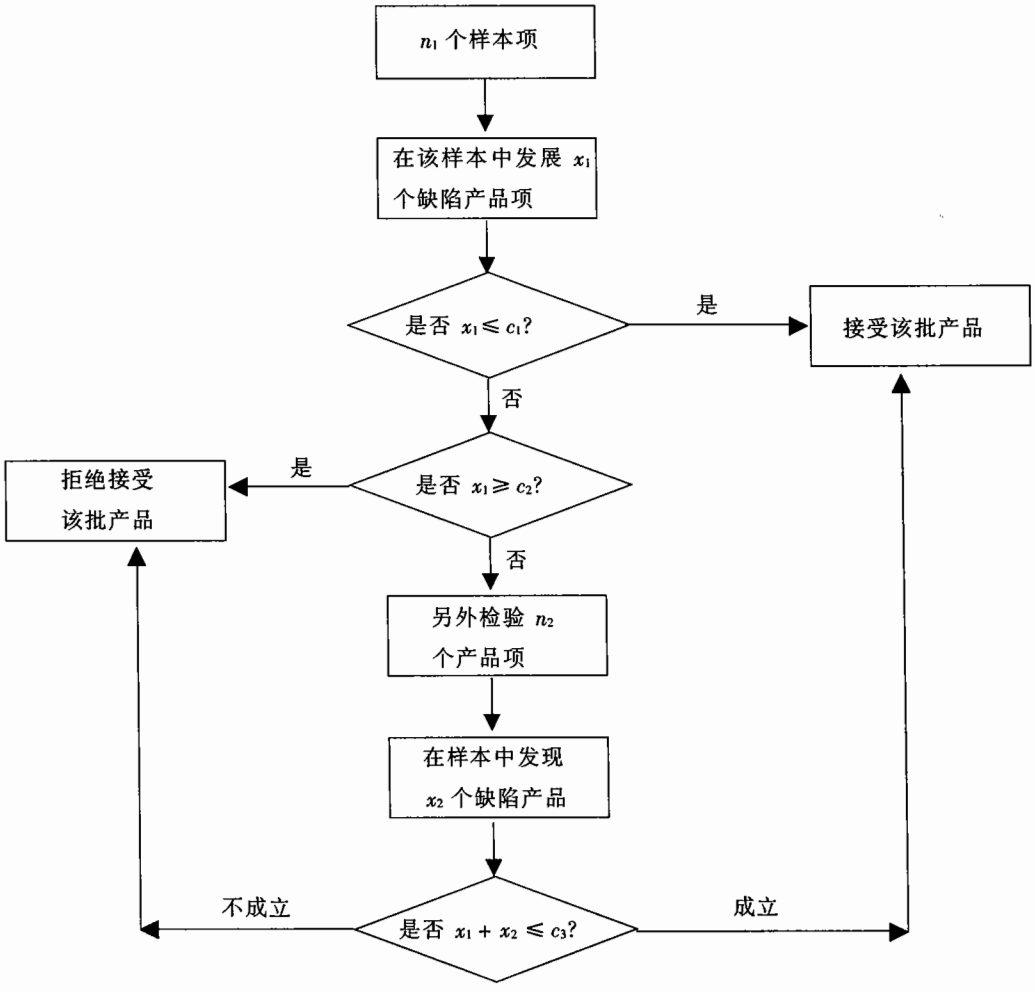


图 20.13　二次抽样验收方案

练　习

方　法

10. 自测题　对 $n=25$, $c=0$ 的抽样验收方案，试求当它的产品缺陷率为 2% 时的接受概率。如果产品缺陷率为 6% 时，接受概率是多少？

11. 对 $n=20, c=0$ 的抽样验收方案，试计算下面每种情形的生产者风险：
    a. 该批产品的缺陷率为 2%。
    b. 该批产品的缺陷率为 6%。

12. 对 $n=20, c=1$ 的抽样验收方案，重复练习 11 的运算。如果验收准则的数值 c 增加的话，生产者风险将会有什么变化？试解释原因。

应　用

13. 对本节所介绍的 KALI 问题，质量管理人员要求：当 $p_0=0.03$ 时，生产者风险控制为 0.10；当 $p_1=0.15$ 时，消费者风险控制为 0.20。对于样本容量为 20、接受准则为 1 的抽样验收方案，试回答下列问题。
    a. 对 $n=20, c=1$ 的抽样验收方案，生产者风险是多少？
    b. 对 $n=20, c=1$ 的抽样验收方案，消费者风险是多少？
    c. $n=20, c=1$ 的抽样验收方案是否满足了质量管理人员的风险要求？对此进行讨论。

14. 为了检验刚刚卸货的原材料，制造商考虑抽取样本 10，15 和 20。利用附录 B 中表 5 的二项概率，选择一个抽样方案，要求：当 $p_1=0.30$ 时，消费者风险 $\beta=0.12$；当 $p_0=0.05$ 时，生产者风险 $\alpha=0.03$。

15. 某国内手表制造商从一个瑞士公司采购石英玻璃，石英玻璃装船批量为 1 000，抽样验收方案为从中随机选择 20 个进行验收。
    a. 假如验收准则分别为 0，1，2，试构造验收特性曲线图。
    b. 如果 $p_0=0.01$，$p_1=0.08$，请计算（a）中每个抽样验收方案的生产者风险和消费者风险。

小　结

本章我们讨论了统计方法在质量管理中的应用。首先我们介绍了$\bar{x}$、R、p 控制图以及 np 控制图作为图形控制手段在监控质量过程中的应用。控制图均有各自的控制限，定期选取样本，并在数据图上描绘数据点。数据点落在控制限之外则表明过程处于失控状态，需要采取纠正措施；数据点落在控制限之内的某些轨迹也能表明存在潜在的质量问题，需要采取纠正措施。

我们同时考察了抽样验收的技巧，根据抽样验收的程序选取样本并检测它们。样本中所发现的缺陷产品的数量是决定接受产品或拒收产品的依据，可以利用对样本容量和验收准则的调整来控制生产者风险（第一类错误）和消费者风险（第二类错误）。

术语释义

质量管理： 确定是否达到质量标准的一系列检验和测量行为。

特定因素： 所生产的产品在质量上的差异来源于如工具的磨损、错误的机器设置、质量低下的原材料或者操作员的操作失误等因素，这些因素被称为特定因素。一旦发现特定因素，应该立即调整生产过程。

一般因素： 产品质量差异来源于随机化差异的因素被称为一般因素。如果发现产品质量差异来源于这些因素，生产过程不需要进行调整。

控制图： 用于确定某个生产过程是否处于控制状态中或失控状态下的图形工具。

$\bar{x}$ 控制图： 当生产过程以长度、重量和温度等参数的均值来测量时所用的控制图。

R 控制图： 当生产过程以某个变量的全距来测量时所用的控制图。

p 控制图： 当生产过程以缺陷产品的比率来测量时所用的控制图。

np 控制图： 当生产过程以缺陷产品的数量来测量时所用的控制图。

批： 一组产品项目，比如正在装运的原材料、采购的产品部件或者生产线上最后组装的成品。

抽样验收： 通过在样本中发现缺陷产品的数量来确定接受还是拒绝一批产品的统计程序。

生产者风险： 拒收一批质量合格产品的风险，即第一类错误。

消费者风险： 接受一批质量不合格产品的风险，即第二类错误。

验收准则： 在仍然能够接受这批产品的前提下，样本中所发现的缺陷产品的最大数量。

抽样验收特性曲线： 以缺陷产品的百分比为自变量、接受这批产品的概率为因变量而建立的函数图像。这条曲线可以用于帮助确定某个特定的抽样验收方案是否同时满足生产者风险和消费者风险的需求。

多重抽样方案： 包含不止一个样本或一个样本阶段的抽样验收方案。以在样本中所发现的缺陷产品的数量为基础来决定是接受该批产品还是拒绝该批产品，或者是继续抽样检查。

重要公式

均值的标准差

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{20.1}$$

$\bar{x}$ 控制图的控制上下限：已知过程均值和标准差

$$\text{UCL} = \mu + 3\,\sigma_{\bar{x}} \tag{20.2}$$

$$\text{LCL} = \mu - 3\,\sigma_{\bar{x}} \tag{20.3}$$

总体样本均值

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} \tag{20.4}$$

平均全距

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k} \tag{20.5}$$

$\bar{x}$ 控制图的控制限：过程均值和标准差未知

$$\bar{\bar{x}} \pm A_2 \bar{R} \tag{20.8}$$

R 控制图的上下限

$$\mathrm{UCL} = \bar{R} D_4 \tag{20.14}$$

$$\mathrm{LCL} = \bar{R} D_3 \tag{20.15}$$

比例的标准误差

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{20.16}$$

***p* 控制图的控制限**

$$\mathrm{UCL} = p + 3\sigma_{\bar{p}} \tag{20.17}$$

$$\mathrm{LCL} = p - 3\sigma_{\bar{p}} \tag{20.18}$$

***np* 控制图的控制限**

$$\mathrm{UCL} = np + 3\sqrt{np(1-p)} \tag{20.19}$$

$$\mathrm{LCL} = np - 3\sqrt{np(1-p)} \tag{20.20}$$

抽样验收的二项概率函数

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \tag{20.21}$$

补充练习

16. 从某一处于控制状态下的生产过程抽取样本，样本容量为 5，并给出如下 20 个均值：

95.72	95.24	95.18
95.44	95.46	95.32
95.40	95.44	95.08

（续表）

95.50	95.80	95.22
95.56	95.22	95.04
95.72	94.82	95.46
95.60	95.78	

a. 以这些数据为基础，当过程处于控制状态下时，试求样本均值的估计值。

b. 假设过程标准差 $\sigma = 0.50$，试建立该过程的控制图，假设过程均值即为(a)中所求。

c. 20个样本均值里面是否有表明过程处于失控状态的值？

17. 产品填装重量呈正态分布，均值为350克，标准差为15克。

a. 建立样本容量为10，20，30的样本的控制限。

b. 当样本容量增加时，控制限有什么变化？

c. 如果出现第一类错误，将会出现什么情况？

d. 如果出现第二类错误，将会出现什么情况？

e. 当样本容量为10，20，30时，出现第一类错误的概率是多少？

f. 对控制图而言，增加样本容量会有什么优点？当样本容量增加时，哪一类错误的概率将会减少？

18. 有25个样本，样本容量为5，结果为 $\overline{\overline{x}} = 5.42$，$\overline{R} = 2.0$，请计算 $\overline{x}$ 控制图和 R 控制图的控制限，并估计过程标准差。

19. 下面是Kensport化学公司的生产过程质量管理数据。数据显示了一个生产周期内5个观测点的摄氏温度。该公司计划利用控制图来监控它的生产过程的温度。请构造 $\overline{x}$ 控制图和 R 控制图。对该公司的生产过程质量可以得出什么结论？

样本	$\overline{x}$	R	样本	$\overline{x}$	R
1	95.72	1.0	11	95.80	0.6
2	95.24	0.9	12	95.22	0.2
3	95.18	0.8	13	95.56	1.3
4	95.44	0.4	14	95.22	0.5
5	95.46	0.5	15	95.04	0.8
6	95.32	1.1	16	95.72	1.1
7	95.40	0.9	17	94.82	0.6
8	95.44	0.3	18	95.46	0.5
9	95.08	0.2	19	95.60	0.4
10	95.50	0.6	20	95.74	0.6

20. 以下是Master Blend咖啡的生产过程数据，数据显示的是3磅重的罐装咖啡的填装重量，利用这些数据构造 $\overline{x}$ 和 R 控制图，对生产过程的质量可以得出什么结论？


CD 光盘数据
Coffee

样本	观测值				
	1	2	3	4	5
1	3.05	3.08	3.07	3.11	3.11
2	3.13	3.07	3.05	3.10	3.10
3	3.06	3.04	3.12	3.11	3.10
4	3.09	3.08	3.09	3.09	3.07
5	3.10	3.06	3.06	3.07	3.08
6	3.08	3.10	3.13	3.03	3.06
7	3.06	3.06	3.08	3.10	3.08
8	3.11	3.08	3.07	3.07	3.07
9	3.09	3.09	3.08	3.07	3.09
10	3.06	3.11	3.07	3.09	3.07

21. 考察以下情形，试评论该情形是否对生产过程质量产生影响。

a. p 控制图的 LCL = 0，UCL = 0.068，当过程处于控制状态下时，产品缺陷率是 0.033，请在控制图上描绘出以下 7 个样本结果：0.035，0.062，0.055，0.049，0.058，0.066，0.055，并对它们进行讨论。

b. $\bar{x}$ 控制图的 LCL = 22.2，UCL = 24.5，当过程处于控制状态下时均值 $\mu = 23.35$，请在控制图上描绘出以下 7 个样本：22.4，22.6，22.65，23.2，23.4，23.85，24.1；并对它们进行讨论。

22. 1 200 个零售商店从中心供应仓库每月进货两次，以往的经验表明 4% 的订单会有这样或那样的错误，例如产品装运错误、装运数量错误，还有已经预定的产品被漏装等。从中每个月随机抽取 200 份订单组成样本，以检查其精确性。

a. 构造该情形的控制图。

b. 6 个月中发生一个或多个错误的订单数目如下：10，15，6，13，8，17。请在控制图上描绘出这些数据点。对这个订单过程，描绘出来的点能表明什么？

23. 抽样验收方案 $n = 10$，$c = 2$，假设 $p_0 = 0.05$，$p_1 = 0.20$。

a. 计算这个抽样验收方案的生产者风险和消费者风险。

b. 生产者或者消费者，或者双方都对所拟定的抽样验收方案不满意吗？

c. 是否应该对抽样验收方案作出改进？如果需要，怎么改进？

24. 有抽样验收方案 $n = 15$，$c = 1$，生产者风险为 0.075。

a. 当 p_0 的值为 0.01，0.02，0.03，0.04，0.05 时，抽样验收方案情形如何？这个值意味着什么？

b. 如果 p_1 为 0.25，则抽样验收方案中的消费者风险是多少？

25. 某个食品生产商大量生产罐装食品，令 p 表示没有达到质量标准的产品比例，使用的是 $n = 25$，$c = 0$ 的抽样验收方案。

a. 当 $p = 0.01$，0.03，0.10，0.20 时，计算抽样特性曲线上的数据点。

b. 描绘抽样特性曲线。

c. 抽样验收方案拒绝缺陷产品率为 0.01 的产品批的概率是多少？

26. 某些时候抽样验收方案将会以很大的样本为基础。在这种情况下，可以用正态分布近似二项概率分布来计算该方案的生产者风险和消费者风险。参考第 6 章，我们知道正态分布近似二项概率分布时有均值 np 和标准差 $\sqrt{np(1-p)}$，假设抽样验收方案 $n=250$, $c=10$。

a. 如果 $p_0=0.02$，生产者风险是多少？正如第 6 章所讨论的，在这种情况下应该使用连续纠正因子，因此，验收概率是根据随机变量的正态分布概率小于或等于 10.5 来计算的。

b. 如果 $p_1=0.08$，消费者风险是多少？

c. 抽样验收中采用大容量样本的优点是什么？缺点是什么？

附录 20.1　用 Minitab 绘制控制图

在这个附录中，我们描述了利用表 20.2 中所示的 Jensen 计算机公司样本数据来生成 Minitab 控制图的步骤。样本序号在 C1 列，第一个观测值在 C2 中，第 2 个观测值在 C3 列中，依此类推。以下的步骤描述了怎样利用 Minitab 来同时构造$\bar{x}$ 控制图和 R 控制图：



CD 光盘数据
Jensen

步骤 1. 选择 **Stat** 下拉菜单；

步骤 2. 选择 **Control Charts**；

步骤 3. 选择 **Xbar-R**；

步骤 4. 出现 Xbar-R 控制图对话框时：
选择 **Subgroups across rows of** 在 **Subgroups across rows of** 中输入 C2-C6
选择 **Tests**；

步骤 5. 当 Tests 对话框出现时：
选择 **One point more than 3 sigmas from center line**；*
点击 **OK**；

步骤 6. 当 Xbar-R 控制图对话框出现时，
点击 **OK**。

$\bar{x}$ 控制图和 R 控制图将在 Minitab 输出结果中一起出现。在前面的 Minitab 程序里的第 3 步中有一些可选项，提供了对不同控制图的访问。例如，$\bar{x}$ 控制图和 R 控制图可以分开选取。备用选项包括 p 控制图，np 控制图和其他的一些控制图。

* Minitab 提供了另外几种检验方法以检验质量差异和过程失控的起因，用户可以同时选择其中几种检验方法。

21 抽样调查

统计实例：辛辛那提电气公司

21.1 抽样调查所用术语

21.2 调查种类与抽样方法

21.3 调查误差

非抽样误差

抽样误差

21.4 简单随机抽样

总体均值

总体总量

总体比率

确定样本容量

21.5 分层简单随机抽样

总体均值

总体总量

总体比率

确定样本容量

21.6 整群抽样

总体均值

总体总量

总体比率

确定样本容量

21.7 系统抽样

统计实例

辛辛那提电气公司*

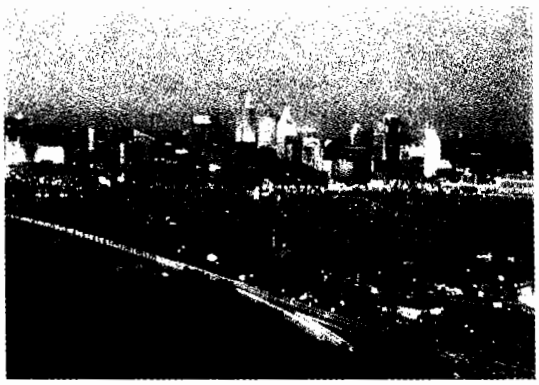
辛辛那提,俄亥俄州

Cinergy,前辛辛那提电气公司(Cincinnati Gas & Electric),是一个公用事业型公司,它为大辛辛那提地区的居民提供煤气和电力服务。为了向它的顾客提供更好的服务,该公司不断努力满足顾客最新的需要。该公司进行了一次关于建筑物特征的抽样调查,以了解在其服务范围内的商务建筑的能源需求量。

此项调查需要搜集有关商业建筑物的大量资料,诸如楼面面积、雇员数量、能源最终使用量、建筑物寿命、建筑材料类型及能源节约标准等等。在调查准备期间,该公司的分析家们发现,在该公司服务的范围内有大约 27 000 个商业建筑物。根据调查经费和精度的要求,他们建议从中选择 616 个商业建筑物作为调查样本。

利用分层简单随机抽样方法选择样本,从公司的记录中可以得到在其服务范围内过去一年每个商业建筑物的总用电量。由于许多建筑物要研究的特征(如规模、雇员数量等)都与其用电量有关,因此他们选择用电量这一标准将建筑物总体划分为 6 层。

第一层包含 100 个商业建筑物,他们都是前 100 名的用电大户,将这些建筑物中的每一个都包含在样本中。尽管它们的数量仅占总数的 0.2%,但是它们的用电量却占总用电量的 14.4%。对于其他层,建筑物的数量是取决于单位成本获得最大精度的基本条件。



Cinergy 公司利用统计调查确定其用户的用电需求量。
© PhotoDisc, Inc.

他们仔细设计了调查表,并且在正式调查之前还做了试验性调查,采用个人采访法搜集资料。最后,搜集到 616 个商业建筑物中的 526 个建筑物的资料,85.4% 的答复率是很优秀的。目前,辛辛那提电气公司正在利用调查所得到的结果,进行能源需求的预测并改进对商业顾客的服务。

就像辛辛那提电气公司进行调查所要考虑的问题一样,本章我们将学习有关抽样调查的设计与实施等问题,这些都是统计学家经常需要考虑的问题。抽样调查常常被用来树立公司的形象,政府和其他机构也常常利用抽样调查来了解总体各个部分的情况。

* 作者衷心感谢 Jim Riddle of Cinergy 提供这一统计实例。

21.1 抽样调查所用术语

在第 1 章中,我们给出了个体、总体和样本的定义,如下所示:

- **个体**(element)是作为数据来源的一个实体。
- **总体**(population)是有待研究的所有个体的总和。
- **样本**(sample)是总体的一个子集。

为了说明这些概念，我们以 Dunning 微系统有限公司（DMI）的情况作为示例，Dunning 微系统有限公司是一个生产微型计算机及其终端的公司，该公司想搜集购买 DMI 微型计算机的用户的特征。为了得到这些资料，该公司对 DMI 微型计算机的拥有者进行了一次抽样调查。这个调查的**个体**是购买 DMI 微型计算机的每个人，**总体**是购买 DMI 微型计算机的所有人组成的集合，**样本**是要调查的 DMI 微型计算机拥有者的一个子集。

在抽样调查中，我们有必要区分目标总体与抽样总体。**目标总体**(target population)是我们要推断的总体，**抽样总体**(sampled population)是实际抽取样本的总体，明确这两个总体不总是一致的很重要。在 DMI 例子中，目标总体是购买了 DMI 微型计算机的所有人，抽样总体是将保修登记卡寄回 DMI 公司的所有计算机拥有者。既然存在购买 DMI 微型计算机但没有寄回保修卡的人，可见抽样总体和目标总体是不一致的。

如果由样本得出的推断是正确的，则抽样总体必须可以代表目标总体。

抽样调查获得的结果仅仅适用于抽样总体，这些结果是否能扩展到目标总体取决于分析家的判断。关键的问题在于：在有待研究的个体上，抽样总体与目标总体是否一致，以此来决定是否可以扩展。

在抽样之前，我们必须将总体划分为**抽样单位**(sampling units)。在某些情形下，抽样单位可以是一个简单的个体，而在另外一些情形，抽样单位也可以是一组个体。例如，假设我们要调查持有证书的专业工程师，这些工程师从事商业建筑物的供暖和空调系统的设计工作。如果无法利用这个名册，则我们必须寻找其他的方法。商业电话号码簿或许可以提供所有工程公司的名册，根据获得的工程公司的名册，我们能够确定工程公司的一个调查样本。然后，对每一个公司，我们应该采访该公司所有的专业工程师。在这种情况下，工程公司是抽样单位，被采访的工程师是个体。

对于某一项特定的研究，抽样单位的列表被称为**抽样框**(frame)。在专业工程师的调查中，抽样框是商业电话号码簿上的所有工程公司的列表，而不是所有专业工程师的列表，这是因为专业工程师的列表是无法得到的。调查所用的特殊抽样框常常取决于可以得到的和确定的名册，然后再确定抽样单位。在实践中，抽样框的编制是进行抽样调查的一个最为困难的而又相当重要的步骤。

21.2　调查种类与抽样方法

最常用的三种调查是邮寄调查、电话调查和私人访谈；每一种类型都涉及到调查问卷的设计

与管理。用于搜集资料的其他调查种类则不需要调查问卷。例如，雇用会计事务所抽取一个公司货物库存的样本，以估计该公司资产负债表上的库存额，这种调查只需简单统计各类项目并记录它们的数量。

> 邮寄和电话调查的费用较为低廉，但是如果访谈员受过良好训练，则对个人调查来说，比较高的回复率和较长的问卷调查表也是可行的。如果需要对每一个个体搜集大量的资料，则个人访谈是惟一的途径，本章统计实践中的示例：辛辛那提电气公司即是如此。

在使用问卷调查表的调查中，最关键的是问卷调查表的设计。设计者必须抵御涵盖所有要研究问题的诱惑，因为每个问题都会增加问卷表的长度，长的问卷调查表不仅会导致回答者的疲劳，访谈员也一样会疲劳，邮寄和电话调查尤其如此。然而，如果采用私人访谈调查，长而复杂的调查问卷却是可行的。对于调查问卷，有关措辞、排序和问题组合都存在大量的知识。这些问题会在关于抽样调查的更为全面的书籍中进行讨论；有关这一类的有关资料见参考文献。

样本的抽样调查可以根据抽样方法进行分类。利用**概率抽样方法**(probabilistic sampling)，我们可以计算出每一个可能样本的概率；但如果利用**非概率抽样方法**(nonprobabilistic sampling)，我们则无法知道可能样本的概率。如果研究者计划对估计的精度作出陈述，则不能采用非概率抽样方法。与之相反，利用概率抽样方法可以构造出置信区间，而正是这个置信区间可以给出抽样误差的界限。在接下来的章节中，我们将讨论四种最常用的概率抽样方法：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。

虽然统计学家更倾向于使用概率抽样方法，但是非概率抽样方法通常也是很必要的。非概率抽样方法的优点在于它的低成本和易于执行，缺点在于无法就估计精度作出有效的统计上的陈述。两种比较常用的非概率方法是方便抽样和判断抽样。

利用**方便抽样**(convenience sampling)的时候，我们根据方便性对包含在样本中的单位进行选择。举个例子，一个教授在大学里进行一项研究，他挑选学生志愿者参加他的研究的原因可能仅仅是因为这些学生恰好在他的班级里。在这种情形下，这个学生志愿者样本被认为是方便抽样。在某些情况下，方便抽样是惟一可行的途径。举个例子，对一批桔子进行抽样，检查人员为了得到桔子样本可以随意从几箱桔子中进行挑选，因为对整批桔子中的每一个桔子都贴上标签以构造样本框，或者使用概率抽样分布方法都将是不现实的。野生动物捕猎和消费者研究的志愿者也是方便抽样的例子。

虽然方便抽样是样本选择和资料收集的一种比较简单的方法，但是对于通过这种方法所取得的样本统计量，我们无法对有待研究的总体参数的优良性进行评估。方便抽样有可能得到比较好的结果，也有可能导致不良结果。这种方法不存在根据样本结果进行统计推断的统计证明过程。但是很多时候研究人员利用方便样本来搜集根据统计方法设计的概率样本资料，对于这个过程，研究人员有必要强调方便样本是作为一个随机样本进行对待的，这样，它就可以成为总体的代表。但是通过这种方法处理得出的结论却经常受到质疑。因此，我们在使用方便样本对总体参数进行推断时，必须相当谨慎。

在使用非概率抽样技术的过程中，我们将根据个人主观意识来选择总体中具备代表性的抽样单位的方法称为**判断抽样**(judgment sampling)。尽管判断抽样通常是一种比较简单的挑选样本的方法，但是调查结果的使用者必须意识到这些结果的质量取决于个人在选择样本时所进行的判断。因此，利用判断样本对总体参数进行统计推断也是需要相当谨慎的。一般说来，我们没有必要对利用判断样本所得出的结果精度作出统计说明。

对于非概率抽样方法，如果所使用的方法能够得到一个具有代表性的样本，则基于这个样本的点估计是有用的。但是，即使那样，结果的精度仍然是不可知的。

在选择样本时，概率抽样方法和非概率抽样方法都是可行的。非概率抽样方法的优点在于成本较低而且易于执行。但是，如果必须要对估计精度作出说明时，则我们必须使用概率抽样方法。几乎所有的大型抽样调查使用的都是概率抽样方法。

21.3　调查误差

在进行调查的过程中有可能产生两种类型的误差，一种是**抽样误差**(sampling error)，它是所得到的样本点估计值与总体参数之间的差异。换句话说，抽样调查是因为没有对总体的所有个体进行调查而产生的误差。另外一种类型的误差是**非抽样误差**(nonsampling error)，它包括抽样调查过程中可能出现的其他所有类型的误差，如测量误差、访谈员误差和处理误差等等。抽样误差仅仅出现在抽样调查中，而非抽样调查则可能出现在普查中，也有可能出现在抽样调查中。

非抽样误差

非抽样误差最常见的形式即我们无法准确地测量要研究的特征。测量误差可以出现在普查或者样本抽样中。对于调查的任何一种形式，调查者都必须相当谨慎，保证测量工具（例如调查问卷表）足够的精确，并且从事测量的人员也必须要经过充分的培训。在大多数情况下，对于细节的足够重视是一种最好的错误防范措施。

缺乏反应所导致的误差是负责设计调查的统计人员和使用调查结果的管理人员所关心的问题。如果无法或者仅仅只能得到部分被调查个体的资料，就有可能产生这种非抽样的误差。如果出现偏差的话将会是一个很严重的问题。举个例子，对妇女走出家庭参加工作的观点进行调查，如果仅仅在白天进行家庭采访，很明显会出现偏差，因为在社会上工作的妇女并没有被包含在调查样本中。

在 1990 年美国的一项抽样调查中，25.9% 的家庭妇女没有接受调查。2000 年的调查专门访谈了没有接受调查的人群，以此来估计这部分总体的特性。

在技术性的调查中通常会出现一些由于缺乏回答问题所需要的知识而产生的非抽样误差。举个例子，对大厦的管理人员进行调查以获得大厦所使用的通风系统的详细资料，大型商务楼的管理人员对这些系统可能会有充分的知识，因为他们可能参加过相关的培训研讨班，而且他们会有专业同行的支持。相比之下，小型商务楼的管理人员则可能对这些通风系统缺乏足够的知识，因为他们经常需要处理大量各种类型的事务。这种知识上的差异可能会对调查结果产生明显的影响。

另外两种类型的非抽样误差是选择误差和处理误差。如果调查中包含有不适当的项目，则有可能产生选择误差。假设设计一个抽样调查，对留胡须的男人进行描述，关于留胡须的男人，有些访谈员的理解认为应该包括留小胡子的男人，另外一些访谈员则有可能不这么认为，这样就会导致调查结果产生缺陷。如果出现登记错误或者输入错误时，就有可能出现处理误差。如果将调查表中的资料输入计算机产生错误，即为输入错误。

尽管在大多数的调查中有可能会出现一些非抽样误差，但是我们通过严密的计划可以使这些误差达到最小。例如保证抽样总体与目标总体一致，遵循良好的问卷调查表设计原则，对访谈员进行良好的培训等等。调查的最后结论应该包括关于非抽样误差对调查结果可能产生的影响的讨论。

抽样误差

回顾 21.1 节中介绍的 DMI 抽样调查问题，假设 DMI 公司计划对购买 DMI 计算机的客户的平均年龄作出估计，如果可以调查 DMI 计算机购买者的总体（即普查），就不会存在非抽样误差，同时我们也能够准确地计算出他们的平均年龄。但是如果不能调查 DMI 购买者的全部总体，将会出现什么样的调查结果呢？在这种情形下，样本均值和总体均值之间可能存在差异，差异的绝对值即为抽样误差。在实践中，对于任何一个特定样本，我们都无法知道抽样误差，因为总体均值总是未知的；但是，对于抽样误差的大小我们却可以给出概率说明。

选择合适的样本设计，可以使抽样误差达到最小。

正如前面所讲的，抽样误差出现的原因在于调查对象仅仅是一个样本，而不是整个总体。尽管抽样误差无法避免，但是它却是可以控制的。选择适当的抽样方法或者抽样设计是对这种误差进行控制的一个重要方法。在接下来的几节中，我们将讨论四种概率抽样方法：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。

21.4 简单随机抽样

回顾在第 7 章中简单随机样本的定义：

从一个容量为 N 的有限总体中抽取一个容量为 n 的简单随机样本，每一个容量为 n 的样本被选中的概率均等。

利用**简单随机抽样**(simple random sampling)进行抽样调查，我们首先建立一个抽样框，即抽样总体中每个个体的列表。然后根据随机数表进行抽样，利用随机数表可以确保抽样总体中每一个个体都有均等的概率被选中。在这一节中，我们将介绍利用简单随机抽样对总体均值、总体总量以及总体比率进行估计。

总体均值

在大多数抽样调查中，总体概率的分布形式是未知的。举个例子，在 DMI 调查中，管理人员想估计 μ，即购买了 DMI 计算机的顾客的平均年龄的数值。但是，DMI 并不知道他所有的顾客总体平均年龄概率的分布形式。不知道总体概率分布形式并不是什么问题，因为 μ 的点估计$\bar{x}$ 的抽样分布特性只依赖于样本设计的选取。

在第 7 章中我们曾经指出，如果选择大样本（$n \geqslant 30$），则应用中心极限定理可以保证$\bar{x}$ 的抽样分布近似服从正态概率分布。在第 8 章中我们有：如果$\bar{x}$ 的抽样分布近似服从于正态概率分布，则 μ 的区间估计为

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} \tag{21.1}$$

式中，$\sigma_{\bar{x}}$ 为均值的标准误差。

$1-\alpha$ 被称为置信系数，$z_{\alpha/2}$ 为标准正态概率分布水平 $\alpha/2$ 的上侧分位数。举个例子，如果置信系数为 95%，则 $z_{0.025}=1.96$，同时也应该注意到，均值的标准误差$\sigma_{\bar{x}}$ 正好是$\bar{x}$ 的抽样分布的标准差。一般情况下，本章中不论任何时候使用**标准误差**，都是指需要考虑点估计的抽样分布的标准差。

从一个容量为 N 的有限总体中，抽取一个容量为 n 的简单随机样本，均值标准误差的估计值为

$$s_{\bar{x}}=\sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{21.2}$$

此时用$s_{\bar{x}}$ 作为$\sigma_{\bar{x}}$ 的估计值，总体均值的区间估计为

$$\bar{x} \pm z_{\alpha/2}s_{\bar{x}} \tag{21.3}$$

在抽样调查中，一般取 $z=2$ 构造置信区间。因此，在使用简单随机样本时，总体均值近似 95%的置信区间表达式如下所示：

利用 $z=2$ 来构造置信区间是以抽样分布为近似正态分布的假设为基础的；几乎 95% 的数据落在均值的两个标准离差的范围内。

总体均值近似 95% 的置信区间表达式

$$\bar{x} \pm 2\,s_{\bar{x}} \tag{21.4}$$

我们以 *Great Lakes Recreation* 的出版商的情形作为例子对此加以说明，*Great Lakes Recreation* 是一本关于划船以及垂钓类主题的地区性杂志。杂志目前拥有 8 000 家订户，从中抽取了 484 户作为简单随机样本，这 484 户的年平均收入为 30 500 美元，标准差为 7 040 美元。因此，所有订户年平均收入的无偏估计为$\bar{x}$ = 30 500 美元。根据这个结果和公式(21.2)，我们可以得到均值标准误差的估计值

$$s_{\bar{x}} = \sqrt{\frac{8\,000-484}{8\,000}}\left(\frac{7\,000}{\sqrt{484}}\right) = 310$$

因此，根据公式(21.4)，我们发现这本杂志订户年平均收入的近似 95% 的置信区间为

$$30\,500 \pm 2(310) = 30\,500 \pm 620，或\ 29\,880—31\,120$$

以上的过程也适用于对其他总体参数，比如总体总量和总体比率进行计算。对于点估计的抽样分布近似服从正态概率分布的所有情形，近似 95% 的置信区间都为

$$点估计值 \pm 2 \times 点估计量的标准误差的估计值$$

在抽样调查的术语中，±的数值通常被称为抽样误差界限值，这与第 8 章所讨论的边际误差是相同的。

例如，在 *Great Lakes Recreation* 的抽样调查中，点估计量的标准误差的估计值为$s_{\bar{x}}$ = 310 美元，**允许误差**(bound on the sampling error)为 2 × 310 美元 = 620 美元。

总体总量

考察东北电力电气公司（NEG）的例子，作为能源使用研究的一部分，NEG 需要对它所服务的区域里面的 500 家公立学校的总面积作出估计。用 X 表示这 500 所公立学校的总面积；换句话说，X 表示总体总量，如果 μ，即 500 所公立学校的总面积已知的话，X 的数值可以计算出来，即 X 等于 N 乘以 μ。但是，因为 μ 是未知的，因此可以用 $N\bar{x}$ 作为 X 的点估计。用$\hat{X}$ 表示 X 的点估计，有

总体总量的点估计

$$\hat{X} = N\bar{x} \tag{21.5}$$

点估计的标准误差的估计量为

$$s_{\hat{x}} = Ns_{\bar{x}} \tag{21.6}$$

式中

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{21.7}$$

请注意公式(21.7)为估计均值的标准误差公式。根据这个标准误差和公式(21.6)，我们可以得到总体总量近似 95% 的置信区间表达式：

总体总量近似 95% 的置信区间

$$N\bar{x} \pm 2s_{\hat{X}} \tag{21.8}$$

假设在 NEG 公司的研究中，从 $N=500$ 所公立学校中挑选了 $n=50$ 所公立学校作为简单随机样本；样本均值 $\bar{x}=22\,000$ 平方英尺，样本标准差 $s=4\,000$ 平方英尺。利用公式(21.5)，我们可以得到总体总量的点估计值

$$\hat{X} = (500)(22\,000) = 11\,000\,000$$

公式(21.7)可以用于计算均值的标准误差

$$s_{\bar{x}} = \sqrt{\frac{500-50}{500}}\left(\frac{4\,000}{\sqrt{50}}\right) = 536.66$$

然后根据公式(21.6)，可以得到 $\hat{X}$ 的标准误差估计值

$$s_{\hat{X}} = (500)(536.66) = 268\,330$$

因此，利用公式(21.8)，我们可以得到 NEG 公司服务区域内的 500 所公立学校总面积近似 95% 的置信区间为

$$11\,000\,000 \pm 2(268\,330) = 11\,000\,000 \pm 536\,660$$

或者 10 463 340 至 11 536 660 平方英尺。

总体比率

总体比率 p 指的是总体中具备感兴趣特性的个体所占的比重。举个例子，在市场调查研究中，人们希望了解喜好某一特定品牌的消费者比重，这就是总体比率。样本比率 $\bar{p}$ 是总体比率的无偏点估计。总体比率标准误差的估计值为

$$s_{\bar{p}} = \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{\bar{p}(1-\bar{p})}{n-1}\right)} \tag{21.9}$$

总体比率近似 95% 的置信区间表达式如下所示：

总体比率近似 95% 的置信区间

$$\bar{p} \pm 2s_{\bar{p}} \tag{21.10}$$

以 NEG 公司的抽样调查作为示例，NEG 公司计划对它的服务区域内使用天然气作为燃料供热 500 所的学校的比率作出估计。如果在所抽出的 50 所学校的样本中，有 35 所学校使用天然气，则在

总体 500 所学校中使用天然气比率的点估计值$\bar{p}=35/50=0.70$，根据式(21.9)我们可以得到比率标准误差的估计值为

$$s_{\bar{p}}=\sqrt{\left(\frac{500-50}{500}\right)\left(\frac{0.7-(1-0.7)}{50-1}\right)}=0.062\ 1$$

因此，利用公式(21.10)，我们可以得到总体比率的95%的置信区间为

$$0.7\pm 2(0.062\ 1)=0.7\pm 0.124\ 2$$

或者 0.575 8 至 0.824 2。

从这个例子我们可以看到，估计总体比率的时候置信区间的宽度可能很宽。一般说来，我们需要大样本容量来保证总体比率的估计精度。Louis Harris & Associates 进行了一项关于 529 个共同基金投资者的抽样调查，其结果表明：调查结果应该可以精确到 4.3%（*Business Week*, August 15, 1994)。这意味着近似 95% 的置信区间宽度为 0.086。对于样本容量很大的总体，我们经常要使用 $n=1\ 200$ 或者更大的样本。

确定样本容量

在抽样设计中样本容量的选取是一个很重要的问题。最好的选择通常是在费用和精度之间进行权衡。比较大的样本可以给出较高的精度（允许误差比较少），但是这样做会导致调查更加昂贵。通常情况下，研究预算将会决定样本容量的大小。对于其他的情形，样本容量应该足够的大，以确保调查能够达到指定的精度。

选择样本容量的一般性方法是首先规定所要求达到的精度，然后确定能够满足精度要求的最小样本容量。在这里，精度指的是近似置信区间的大小；比较小的置信区间可以得到较高的精度。因为近似置信区间的大小取决于样本误差 B，这意味着选择精度水平相当于选择 B 的数值。下面我们介绍估计总体均值时选择样本容量需要用到的方法。

公式（21.2）为均值的标准误差估计公式

$$s_{\bar{x}}=\sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

回顾前面所提到的允许误差为点估计标准误差估计值的 2 倍，因此我们有

$$B=2\sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{21.11}$$

对于允许误差 B，根据公式（21.11）我们可以得到

$$n=\frac{Ns^2}{N\left(\frac{B^2}{4}\right)+s^2} \tag{21.12}$$

一旦选择了所需要达到的精度水平（通过选择 B 的数值来进行选择），根据公式(21.12)我们即可以得到为了达到指定精度水平所需要的 n 的数值。在利用公式(21.12)确定 n 的数值时，存在一个实际问题，即除了规定所需要的允许误差 B 之外，我们还必须知道样本方差 s^2。但是我们却只有在得到实际样本时才可能获知 s^2。

科克伦*提出了在实践中确定 s^2 估计值的各种方法，其中三种方法如下所示：

1. 分两个阶段进行抽样，根据第一个阶段公式(21.12)中所得到的 s^2 估计值，确定出全部样本容量 n，然后再抽取第二阶段所需要的其余单位数。
2. 利用试点调查或者预先检验的结果对 s^2 进行估计。
3. 利用以前的样本资料信息进行估计。

现在让我们以某一所大学的毕业生起始年薪为例，来考察总体均值的估计。假设这所大学有 $N=5\,000$ 名毕业生，我们计划建立近似 95% 的置信区间，宽度在 1 000 美元之内。对这个置信区间有 $B=500$。利用公式(21.12)确定 n 之前我们需要估计 s^2 的数值，假设以去年同样的研究为基础，我们得到 $s=3\,000$ 美元。然后我们可以利用这个数值来估计 s^2，根据 $B=500$、$s=3\,000$ 以及 $N=5\,000$，以及公式(21.12)，我们可以确定样本容量为

$$n=\frac{5\,000(3\,000)^2}{5\,000\left(\frac{(500)^2}{4}\right)+(3\,000)^2}=139.97$$

综上所述，对于规定宽度为 1 000 美元的近似 95% 置信区间，所需要的样本容量为 140。但是必须牢记这个计算结果取决于 s 的最初估计值 3 000 美元。如果对于今年的抽样调查，s 的数值增大，则近似置信区间宽度也将超过 1 000 美元。因此，如果调查预算许可，调查设计者可将样本容量确定为 150，以确保近似 95% 的置信区间宽度小于 1 000 美元。

对于给定的允许误差 B，估计总体总量所必需的样本容量计算公式为

$$n=\frac{Ns^2}{\left(\frac{B^2}{4N}\right)+s^2} \tag{21.13}$$

在前面的示例中，我们用允许误差 $B=500$ 来估计起始年薪。现在假设我们将允许误差设为 2 000 000 美元，以此来估计 5 000 名毕业生的全部年薪。则将 $B=2\,000\,000$ 代入公式(21.13)，得到所需样本容量

$$n=\frac{5\,000(3\,000)^2}{\frac{(2\,000\,000)^2}{4(5\,000)}+(3\,000)^2}=215.31$$

综上所述，对于给定允许误差为 2 000 000 美元的近似 95% 的置信区间，所需要的样本容量为 216。在此我们应该注意到，如果要同时满足总体均值的允许误差为 500 美元和总体总量允许

* 见 William G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, 1977。

误差为 2 000 000 美元的调查，则样本容量必须为 216。该容量为总体均值所提供的允许误差比所需要的更小，是满足总体总量所需精度的最小值。

在对总体比率进行估计时，选择样本容量的公式类似于估计总体均值的公式。我们所需要做的仅仅是将公式(21.12)中的 s^2 替换为 $\bar{p}(1-\bar{p})$，即可得到

$$n=\frac{N\bar{p}(1-\bar{p})}{N\left(\frac{B^2}{4}\right)+\bar{p}(1-\bar{p})} \tag{21.14}$$

为了使用公式(21.14)，我们必须规定允许误差 B 和 $\bar{p}$ 的估计值。如果缺乏 $\bar{p}$ 的合适的估计值，我们可以用 $\bar{p}=0.5$ 代替，这样将会确保近似置信区间的允许误差会在所要求的误差范围之内。

练　习

方　法

1. 自测题 利用简单随机抽样，从总体 $N=800$ 中抽取容量为 50 的样本，样本均值 $\bar{x}=215$，样本标准差 $s=20$。
   a. 请估计总体均值。
   b. 估计均值的标准误差。
   c. 建立总体均值近似 95% 的置信区间。

2. 利用简单随机抽样，从总体 $N=400$ 中抽取容量 $n=80$ 的样本，样本均值 $\bar{x}=75$，样本标准差 $s=8$。
   a. 请估计总体均值。
   b. 估计均值的标准误差。
   c. 建立总体均值近似 95% 的置信区间。

3. 利用简单随机抽样，从总体 $N=1\,000$ 中抽取容量 $n=100$ 的样本，样本比率 $\bar{p}=0.30$。
   a. 请估计总体比率。
   b. 建立总体比率的标准误差。
   c. 建立总体比率近似 95% 的置信区间。

4. 选择一个样本建立总体均值近似 95% 的置信区间，总体包含 450 个个体，试点研究得出 $s=70$。如果计划建立宽度为 30 的近似 95% 的置信区间，则应该抽取多大容量的样本？

应　用

5. 自测题 1996 年北卡罗来纳州小型企业管理委员会为本州的小企业批准了 771 项由政府担保的企业贷款（*The Wall Street Journal Almanac*, 1998）。假设根据一个由 50 个小型企业组成的样本得知平均贷款额为 149 670 美元，标准差为 73 420 美元，样本中有 18 家企业为制造型企业。

   a. 建立总体平均贷款额近似 95% 的置信区间。
   b. 建立北卡罗来纳州 771 家小型企业贷款总额近似 95% 的置信区间。
   c. 在样本中，50 家企业中的 18 家为制造企业。建立制造企业贷款比率近似 95% 的置信区间。

6. 根据加利福尼亚州某县 724 家企业的税务申报表，得到这些企业年平均收入为 161 220 美元，标准差为 31 300 美元。如果计划在下一年建立这些公司年平均收入近似 95% 的置信区间，需要抽取多大容量的样本？精度要求为区间宽度小于 5 000 美元。

21.5　分层简单随机抽样

在**分层简单随机抽样**(stratified simple random sampling)中，我们首先将总体分成 H 组（即层），然后从第 h 层抽取一个容量为 n_h 的简单随机样本。根据这 H 个简单随机样本的综合资料我们可以得到关于总体均值、总体总量和总体比率等等各种总体参数的估计。

如果每个层内的差异比层间差异小，则分层简单随机样本可以获得更高的精度（总体参数的区间估计将更窄），各层之间的划分应该根据样本设计者的判断。以应用为基础，总体可以按部门、地域、年龄、产品类型、行业类型和销售水平等等进行分层。

举个例子，假设 Lakeside 大学的商学院计划对今年毕业生进行调查，目的是了解他们的起始年薪。这个学院有 5 个专业：会计、金融、信息系统、市场和运营管理。在今年的 1 500 名毕业生中，会计专业有 $N_1 = 500$ 名，金融专业有 $N_2 = 350$ 名，信息系统专业有 $N_3 = 200$ 名，市场营销专业 $N_4 = 300$ 名，运营管理专业 $N_5 = 150$ 名。往年的起始年薪资料表明：起始年薪在专业之间的差异要大于专业内部的差异。因此，抽取了 $n = 180$ 名学生组成简单随机样本，180 名学生中，其中会计专业 45 名（$n_1 = 45$），金融专业 40 名（$n_2 = 40$），信息系统专业 30 名（$n_3 = 30$），市场营销专业 35 名（$n_4 = 35$），运营管理专业 30 名（$n_5 = 30$）。

总体均值

在分层抽样中，总体均值的无偏估计是各层样本均值的加权平均数，权数即总体在各层的比重。我们用 $\bar{x}_{st}$ 表示总体均值的点估计，其定义如下：

总体均值的点估计

$$\bar{x}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{x}_h \tag{21.15}$$

式中 H——层数；

$\bar{x}_h$——第 h 层的样本均值；

N_h——第 h 层的单位数；

N——总体所有的个体数：$N = N_1 + N_2 + \cdots + N_H$。

对分层简单随机样本，均值的标准误差计算如下：

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{s_h^2}{n_h}} \tag{21.16}$$

利用这个结果，我们可以得到总体均值近似 95% 的置信区间表达式：

总体均值近似 95% 的置信区间

$$\bar{x}_{st} \pm 2 s_{\bar{x}_{st}} \tag{21.17}$$

假设表 21.1 为 Lakeland 大学商学院 180 名毕业生的样本调查结果。各个专业，或者层的样本均值分别为：会计 35 000 美元、金融 33 500 美元、信息系统 41 500 美元、市场营销 32 000 美元、运营管理 36 000 美元，利用这些结果，根据公式(21.15)我们可以得到总体均值点估计为

$$\bar{x}_{st} = \left(\frac{500}{1\,500}\right)(35\,000) + \left(\frac{350}{1\,000}\right)(33\,500) + \left(\frac{200}{1\,500}\right)(41\,500) + \left(\frac{300}{1\,500}\right)(32\,000) + \left(\frac{150}{1\,500}\right)(36\,000) = 35\,017$$

表 21.1 Lakeland 大学商学院毕业生起始年薪抽样调查

专业(h)	$\bar{x}_h$(美元)	s_h	N_h	n_h
会计	35 000	2 000	500	45
金融	33 500	1 700	350	40
信息系统	41 500	2 300	200	30
市场营销	32 000	1 600	300	35
运营管理	36 000	2 250	150	30

表 21.2 给出了估计标准误差所需要的一部分计算结果，其中有

$$\sum_{h=1}^{5} N_h (N_h - n_h) \frac{s_h^2}{n_h} = 42\,909\,037\,698$$

因此有

$$s_{\bar{x}_{st}} = \sqrt{\left(\frac{1}{(1\,500)^2}\right)(42\,909\,037\,698)} = \sqrt{19\,070.68} = 138$$

所以利用公式(21.17)我们可以得到总体均值近似 95% 的置信区间为：35 017 ± 2(138) = 35 017 ± 276，或者 34 741 至 35 293 美元。

表 21.2　Lakeland 大学商学院毕业生起始年薪抽样调查估计均值标准误差的部分计算结果

专业	h	$N_h(N_h - n_h)\frac{s_h^2}{n_h}$
会计	1	$500(500-45)\frac{(2\,000)^2}{45} = 20\,222\,222\,222$
金融	2	$350(350-40)\frac{(1\,700)^2}{40} = 7\,839\,125\,000$
信息系统	3	$200(200-30)\frac{(2\,300)^2}{30} = 5\,995\,333\,333$
市场营销	4	$300(300-35)\frac{(1\,600)^2}{35} = 5\,814\,857\,143$
运营管理	5	$150(150-30)\frac{(2\,250)^2}{30} = 3\,037\,500\,000$
	$\sum_{h=1}^{5} N_h(N_h - n_h)\frac{s_h^2}{n_h}$	$= 42\,909\,037\,698$

总体总量

总体总量（X）的点估计由 N 乘以 $\bar{x}_{st}$ 得到。

总体总量的点估计

$$\hat{X} = N\bar{x}_{st} \qquad (21.18)$$

这个点估计量的标准误差估计值为

$$s_{\hat{x}} = Ns_{\bar{x}_{st}} \qquad (21.19)$$

因此，总体总量 95% 的近似置信区间表达式如下所示：

总体总量 95% 的近似置信区间

$$N\bar{x}_{st} \pm 2\,s_{\hat{x}} \qquad (21.20)$$

现在假设 Lakeland 大学商学院计划对其 1 500 名商科毕业生的总收入作出估计，目的是为了

评估他们对经济的影响。利用公式(21.18)我们可以得到总收入的无偏估计量:

$$\hat{X} = (1\,500)35\,017 = 52\,525\,500$$

利用公式(21.19)我们得到总体总量标准误差的估计值:

$$s_{\hat{X}} = 1\,500(138) = 207\,000$$

因此，利用公式(21.20)，我们发现1 500名毕业生总收入近似95%的置信区间为52 525 500 ± 2(207 000) = 52 525 500 ± 414 000，或者52 111 500至52 939 500美元。

总体比率

分层简单随机抽样总体比率 p 的无偏估计为各层比率的加权平均数，权数为总体在每一层的比重。我们用 $\bar{p}_{st}$ 表示总体比率的点估计，定义如下所示：

总体比率的点估计

$$\bar{p}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)\bar{p}_h \tag{21.21}$$

式中 H——层数；

$\bar{p}_h$——第 h 层的样本比率；

N_h——第 h 层的个体数；

N——总体所有的个体数：$N = N_1 + N_2 + \cdots + N_H$。

标准误差的估计值计算如下：

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]} \tag{21.22}$$

因此，总体比率近似95%的置信区间表达式如下所示：

总体比率近似95%的置信区间

$$\bar{p}_{st} \pm 2\, s_{\bar{p}_{st}} \tag{21.23}$$

在Lakeland大学抽样调查中，该大学想了解它的毕业生起始年薪不低于36 000美元的比率。180名毕业生抽样调查的结果显示表明63名毕业生的起始年薪在36 000美元之上，63名学生中有16名专业为会计，3名为金融，29名为信息系统，没有市场营销专业，15名为运营管理。

利用公式(21.21)我们可以得到工作的起始年薪不低于36 000美元的比率点估计为

$$\bar{p}_{st} = \left(\frac{500}{1\,500}\right)\left(\frac{16}{45}\right) + \left(\frac{350}{1\,500}\right)\left(\frac{3}{40}\right) + \left(\frac{200}{1\,500}\right)\left(\frac{29}{30}\right) + \left(\frac{300}{1\,500}\right)\left(\frac{0}{35}\right) + \left(\frac{150}{1\,500}\right)\left(\frac{15}{30}\right)$$

$$= 0.314\,9$$

在表 21.3 中，我们给出了估计标准误差所需要的部分计算结果：

$$\sum_{h=1}^{5} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right] = 1\,570.691\,3$$

因此有

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{(1\,500)^2}(1\,570.691\,3)} = 0.026\,4$$

利用公式(21.23)我们可以得到毕业生起始年薪不低于 36 000 美元的近似 95% 置信区间为：0.314 9 ± 2(0.026 4) = 0.314 9 ± 0.052 8，或者 0.262 1 至 0.367 7。

表 21.3　Lakeland 大学抽样调查中估计的标准误差所需要的部分调查结果

专业	h	$N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right]$
会计	1	$500(500-45)\left[\frac{(16/45)(29/45)}{45-1}\right] = 1\,184.736\,3$
财务	2	$350(350-40)\left[\frac{(3/40)(37/40)}{40-1}\right] = 193.004\,8$
信息系统	3	$200(200-30)\left[\frac{(29/30)(1/30)}{30-1}\right] = 37.777\,8$
营销	4	$300(300-35)\left[\frac{(0/35)(35/35)}{35-1}\right] = 0.000\,0$
运营管理	5	$150(150-30)\left[\frac{(15/30)(15/30)}{30-1}\right] = 155.172\,4$
		$\sum_{h=1}^{5} N_h(N_h - n_h)\left[\frac{\bar{p}_h(1-\bar{p}_h)}{n_h - 1}\right] = 1\,570.691\,3$

确定样本容量

对分层简单随机抽样我们可以用两阶段过程来选取样本容量，首先确定总的样本容量 n；然后再确定每一层应该分配的样本单位数。或者首先确定每一层应选择的样本个体数，然后将这些加起来得到总样本容量。因为通常需要估计每一层的均值、总量和比率，所以这两种组合方法都是经常使用的。确定总的样本容量 n 以及怎样对它进行分配，这样可以给所有有待研究的总体参数提供必要的精度。然后，如果某些层的样本单位数没有达到满足层内估计量所要求的必要精度，我们就需要将这些层的样本单位数向上调整，在本节中，我们将讨论如何恰当地将总样本容量分配到不同的层，并给出选取总样本容量以及对它进行分配的方法。

分配工作指的是确定总样本分配到每一层的部分，这些部分将确定每一层的简单随机样本数量大小，在进行分配时需要考虑的较为重要的因素如下：

1. 每一层的个体数。
2. 每层内的个体方差。
3. 每一层选择个体所需的费用。

一般说来，较大的样本应分配给个体数目较多的层和方差较大的层。与此相反，在给定费用的情形下，为了获得更多的信息，单位抽样成本最大的层应该分配较少的样本数目。

层与层之间的差异通常是很大的。举个例子，假设在某一项特定研究中，我们想确定每一栋建筑物中员工的数量。因为大型建筑物每层之间的差异比小型建筑每层之间的差异要大，因此应该给大型建筑物的层赋予较大的样本。如果抽样个体仅位于某些层，而非其他层，则采访者必须在抽样个体之间往返，因此选择费用也是一个很重要的方面。某些层涉及到农村地区，而另外一些层则包含城市地区，这种情况是经常出现的。

在很多调查中，关于各层的抽样单位成本近似相等(比如邮寄和电话调查)。在这种情况下，进行分配可以忽略抽样成本。对于这种情形，我们可以给出确定样本容量和进行分配的近似公式。在有关抽样调查的较为高级的教材中，我们给出了抽样成本层间差异很显著的情形的适用公式。本节我们所给出的公式，能够满足在给定的精度水平下使总抽样成本最小的要求。这种方法就是著名的 **Neyman 分配法**，它将总样本容量 n 分配到各层，结果如下所示：

$$n_h = n\left(\frac{N_h s_h}{\sum_{h=1}^{H} N_h s_h}\right) \tag{21.24}$$

公式(21.24)表明分配到各层的个体数受每一层容量和标准差的影响。在进行分配之前我们必须首先确定总样本容量 n。对于给定的精度水平 B，在估计总体均值和总体总量时，我们可以利用下面的公式来确定总样本容量：

估计总体均值时的样本容量

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h s_h^2} \tag{21.25}$$

估计总体总量时的样本容量

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h s_h^2} \tag{21.26}$$

举个例子，假设 Chevrolet 销售商计划调查购买 Corvette、Geo Prizm 或者 Cavalier 商品的消费者，以获得确定广告时所需要的一些有用信息。尤其需要指出的是，假设允许误差为 100 美元，销售商想估计这些消费者的月平均收入，它将 600 名 Corvette、Geo Prizm 和 Cavalier 的消费者分为 3 层，100 名购买了 Corvette，200 名购买了 Geo Prizm，300 名购买了 Cavalier。用试点调查对每一层的标准差进行估计，所得结果如下所示：Corvette 的消费者 $s_1 = 1\,300$ 美元，Geo Prizm 的消费者 $s_2 = 900$ 美元，Cavalier 的消费者 $s_3 = 500$ 美元。

这个调查的第一步是选择样本容量，在给定允许误差 $B = 100$ 美元的情况下，根据公式（21.25）确定估计总体均值时所需要的总样本容量。首先我们计算

$$\sum_{h=1}^{3} N_h s_h = 100(1\,300) + 200(900) + 300(500) = 460\,000$$

然后，我们计算

$$\sum_{h=1}^{3} N_h s_h^2 = 100(1\,300)^2 + 200(900)^2 + 300(500)^2 = 406\,000\,000$$

将这些结果代入公式(21.25)中，在给定允许误差 $B = 100$ 美元的条件下，我们可以确定所需要的总样本容量为

$$n = \frac{(460\,000)^2}{\dfrac{(600)^2(100)^2}{4} + 406\,000\,000} = 162$$

因此，总样本容量为 162 即能够满足所要求的精度。为了将总样本分配到 3 层中去，我们可以利用公式(21.24)

$$n_1 = 162\left(\frac{100(1\,300)}{460\,000}\right) = 46$$

$$n_2 = 162\left(\frac{200(900)}{460\,000}\right) = 63$$

$$n_3 = 162\left(\frac{300(500)}{460\,000}\right) = 53$$

因此对于 162 个消费者的总样本容量，我们建议按如下方式进行抽样：46 名为 Corvette 购买者，63 名为 Geo Prizm 购买者，53 名为 Cavalier 购买者。

在估计总体比率的时候为了确定样本容量，我们只需简单地将公式（21.25）中的 s_h 用 $\sqrt{\bar{p}_h(1-\bar{p}_h)}$ 来代替，结果如下：

$$n = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{\bar{p}_h(1-\bar{p}_h)}\right)^2}{N^2\left(\dfrac{B^2}{4}\right) + \sum_{h=1}^{H} N_h \bar{p}_h(1-\bar{p}_h)} \tag{21.27}$$

一旦估计总体比率时所需的总样本容量被确定，我们就可以用 $\sqrt{\bar{p}_h(1-\bar{p}_h)}$ 代替 s_h 后的公式（21.24）对各层进行分配。

评 注

1. 分层简单随机抽样的优点在于各层总体参数的估计能够自动成为抽样过程的派生产品。举个例子，在 Lakeland 大学的抽样调查中，我们除了能够获得所有毕业生起始平均年薪的估计外，还可以获得每个专业的学生起始年薪的估计值。因为每个专业工作的起始年薪估计值仅仅取决于从每一层所得的简单随机样本，因此选取一个简单随机样本建立近似置信区间的过程[见公式(21.4)]也适用于对每一层建立均值近似 95% 的置信区间。同样地，利用公式(21.8)和公式（21.10）也可以分别建立每一层的总体总量和总体比率的区间估计。

2. 有时候也可以用分层简单随机抽样分配的另一种形式，即**比例分配**，利用这种方法分配给每一层的样本单位数的公式如下：

$$n_h = n\left(\frac{N_h}{N}\right) \qquad (21.28)$$

如果各层方差近似相等而且抽样单位成本在每个层间几乎完全相同时，比较合适的方式是进行比例分配。在各层方差相等的情形下，比例分配和 Neyman 分配过程的结果是相同的。

练 习

方 法

7. 一个分层简单随机抽样有如下的结果：

自测题

层（h）	$\bar{x}_h$	s_h	$\bar{p}_h$	N_h	n_h
1	138	30	0.50	200	20
2	103	25	0.78	250	30
3	210	50	0.21	100	25

a. 计算每一层总体均值的估计值。

b. 建立每一层总体均值近似 95% 的置信区间。

c. 建立整个总体均值近似 95% 的置信区间。

8. 重新考察练习 7 的样本抽样结果。

a. 建立每一层总体总量的估计值。

b. 对于总体中的 550 个个体，计算总体总量的一个点估计。

c. 建立总体总量近似 95% 的置信区间。

9. 重新考察练习 7 的样本抽样结果。
   a. 建立每一层总体比率近似 95% 的置信区间。
   b. 对于总体中的 550 个个体，计算总体比率的一个点估计。
   c. 估计总体比率的标准误差。
   d. 建立总体比率近似 95% 的置信区间。

10. 某个总体被划分为 3 层，其中 $N_1 = 300$，$N_2 = 600$，$N_3 = 500$。根据过去的调查，每一层的标准差估计值如下：$s_1 = 150$，$s_2 = 75$，$s_3 = 100$。
    a. 如果总体均值的估计值允许误差为 $B = 20$，则样本容量应该为多少？每一层应该分配多少个体？
    b. 如果总体均值的估计值允许误差为 $B = 10$，则样本容量应该为多少？每一层应该分配多少个体？
    c. 如果总体均值的估计值允许误差为 $B = 15\,000$，则样本容量应该为多少？每一层应该分配多少个体？

应　用

11. 某个药品连锁店在 4 个城市设有经营网点：38 个分店位于印第安诺波利斯，45 个分店位于路易斯维尔，80 个分店在圣路易斯，70 个分店位于孟菲斯。由于竞争的缘故，4 个城市的分店销售额存在很大的差异，以下是根据抽样调查所得的销售额（千美元）资料，我们将每一个城市作为一个单独的层，利用分层简单随机抽样抽取样本。

印第安诺波利斯	路易斯维尔	圣路易斯	孟菲斯
50.3	48.7	16.7	14.7
41.2	59.8	38.4	88.3
15.7	28.9	51.6	94.2
22.5	36.5	42.7	76.8
26.7	89.8	45.0	35.1
20.8	96.0	59.7	48.2
	77.2	80.0	57.9
	81.3	27.6	18.8
			22.0
			74.3

   a. 估计每一个城市（层）的平均销售额。
   b. 建立每一个城市平均销售额的近似 95% 的置信区间。
   c. 估计销售额在 50 000 美元之上的分店的比率。
   d. 关于销售额在 50 000 美元之上的分店的比率，建立近似 95% 的置信区间。

12. 重新考察练习 11 中的样本调查结果。
    a. 估计圣路易斯的总体总量。
    b. 估计印第安诺波利斯的总体总量。
    c. 建立药品连锁店平均销售额的近似 95% 的置信区间。
    d. 建立药品连锁店总体销售额的近似 95% 的置信区间。

13. 一个会计事务所在银行金融业、保险业和经纪人行业具有大量的客户，其中金融业 $N_1 = 50$，保险业 $N_2 = 38$，经纪人行业 $N_3 = 35$。这个会计事务所雇用了一个市场调查公司对这三个行业的客户进行调查，该市场调查公司关于客户的业务和这些客户对该会计事务所提供服务的满意程度进行了调查，假设在允许误差 $B = 30$ 的情况下，估计 123 名客户的平均雇员数量近似 95% 的置信区间。
    a. 试点调查得到如下结果：$s_1 = 80$，$s_2 = 150$ 和 $s_3 = 45$。选择样本总容量，确定分配到 3 层中的样本单位数。
    b. 假设在进行试点检验时，选择样本容量的每一层标准差都等于 100。请选择样本总容量并确定每一层的抽样单位数。

21.6 整群抽样

整群抽样(cluster sampling)需要将总体的各个个体分为 N 组(也称为群)，并使总体中每一个个体只属于一群。例如，我们想调查俄亥俄州的登记选民，一种方法是建立包含俄亥俄州所有登记选民的抽样框，然后根据抽样框选取选民的一个简单随机样本；另一种方法是整群抽样，将俄亥俄州 $N = 88$ 个县的表单作为抽样框(见图 21.1)，对于这种方法，每个县（或群）都将包含一组登记选民，而这个州的每个登记选民都仅属于一个群。

假设我们从 88 个县中挑选一个 $n = 12$ 的简单随机样本，如果我们搜集 12 个抽样群体中**全部**登记选民的资料，这种方法被称为**单阶段整群抽样**。如果我们从 12 个抽样群体中选取登记选民的一个简单随机样本，这种方法被称为**二阶段整群抽样**。对于这两种情形，我们都可以利用样本结果得到比如总体均值、总体总量或者总体比率等总体参数的点估计和区间估计公式。在这一章中，我们只讨论单阶段整群抽样，二阶段整群抽样在有关抽样调查的更为高级的教材中进行讨论。

分层抽样和整群抽样都是将总体划分为组，因此从这个角度上讲这两种抽样过程是相似的。但是，选取整群抽样的原因和选取分层抽样的原因却是不同的。如果群内的个体存在差异时，整群抽样将会给出比较好的结果。理想化的情形是每一个群都是总体的一个缩影，在这种情形下，我们只需要抽取很少的群即可获得整个总体的特性信息。

整群抽样的一个最初的应用就是区域抽样，这种情形的群可以是县、区、城市社区或者其他已经有所规定的地理区域。由于资料仅仅从整个地理区域（或者群）中的一个样本中进行搜集，而且群内个体特性很相近，因此资料搜集人员或者调查人员去采访一个抽样个体时，可以有效地



图 21.1　作为登记选民群的俄亥俄州各县

节约时间和经费。因此，如果需要较大的样本总容量，整群抽样比简单随机抽样或者分层简单随机抽样都会节省费用。除此之外，建立抽样框或者选中个体的表单，整群抽样亦可以使所耗费的时间和费用最小化。原因在于整群抽样不需要建立总体中每一个个体的表单，而仅仅需要建立群中个体的表单。

并不需要将总体中的每一个个体列出表单，对于整群抽样，我们只需要列出群中个体的表单。

为了阐述整群抽样，我们以某个州的注册会计师协会的一项调查作为例子，这个注册会计师协会拥有 12 000 名职业注册会计师。作为一部分调查内容，注册会计师协会搜集和注册会计师的收入、性别以及生活相关的资料，因为信息搜集采用的是个人采访的方法，所以注册会计师协会采用整群抽样，目的在于使总的采访差旅费和采访费用最小。抽样框中包含所有在该州登记注册

的会计事务所。假设由 $N=1\,000$ 群，即所有在该州登记注册的会计事务所为 1 000 所，选取 $n=10$ 个会计事务所作为简单随机样本。

在介绍整群抽样中，建立总体均值、总体总量和总体比率近似 95% 的置信区间需要用到一些公式，公式中的记号意义如下：

N——总体群数；

n——样本中所选取的群数；

M_i——第 i 群个体数；

M——总体中的单位数，$M=M_1+M_2+\cdots+M_N$；

$\overline{M}=M/N$——每一群的平均单位数；

x_i——第 i 群所有观察值的总量；

a_i——第 i 群具有某种特性的观察值的数量。

关于这个注册会计师协会的抽样调查，我们有以下的已知条件：

$$N=1\,000$$
$$n=10$$
$$M=12\,000$$
$$\overline{M}=12\,000/1\,000=12$$

表 21.4 给出了每个所挑选的群中 M_i 和 x_i 的数值，还有女性注册会计师（a_i）的资料。

表 21.4　注册会计师抽样调查结果

事务所(i)	注册会计师数(M_i)	第 i 个事务所年薪总额(千美元)(x_i)	女性注册会计师人数(a_i)
1	8	384	2
2	25	1 350	8
3	4	148	0
4	17	857	6
5	7	296	1
6	3	131	2
7	15	761	2
8	4	176	0
9	12	577	5
10	33	1 880	9
	总计 128	6 560	35

总体均值

根据整群抽样所得到的总体均值点估计公式如下所示：

总体均值点估计

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \tag{21.29}$$

这个点估计量的标准误差估计值为

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n-1}} \tag{21.30}$$

因此，总体均值近似 95% 的置信区间表达式为

总体均值近似 95% 的置信区间

$$\bar{x}_c \pm 2\, s_{\bar{x}_c} \tag{21.31}$$

利用表 21.4 中的资料我们可以得到注册会计师平均年薪的点估计

$$\bar{x}_c = \frac{6\,560}{128} = 51.250$$

利用表 21.4 中的年薪资料，以千美元计算，我们可以得到注册会计师平均年薪估计值为 51 250 美元。

表 21.5 中列出了估计标准误差所需的部分计算结果

$$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2 = 51\,281.378$$

因此，有

$$s_{\bar{x}_c} = \sqrt{\left[\frac{1\,000-10}{(1\,000)(10)(12)^2}\right]\frac{51\,281.378}{10-1}} = 1.979$$

所以，我们可以得到标准误差为 1 979 美元，利用公式(21.31)我们计算出平均年薪近似 95% 的置信区间为：51 250 ± 2(1 979) = 51 250 ± 3 958，或者从 47 292 美元至 55 208 美元。

表 21.5 注册会计师抽样调查中估计均值标准误差所需的部分计算结果，其中：$\bar{x}_c = 51.250$

事务所(i)		M_i	x_i	$(x_i - 51.250M_i)^2$
1		8	384	$[384 - 51.250(8)]^2 = 676.000$
2		25	1 350	$[1\,350 - 51.250(25)]^2 = 4\,726.563$
3		4	148	$[148 - 51.250(4)]^2 = 3\,249.000$
4		17	857	$[857 - 51.250(17)]^2 = 203.063$
5		7	296	$[296 - 51.250(7)]^2 = 3\,937.563$
6		3	131	$[131 - 51.250(3)]^2 = 517.563$
7		15	761	$[761 - 51.250(15)]^2 = 60.063$
8		4	176	$[176 - 51.250(4)]^2 = 841.000$
9		12	577	$[577 - 51.250(12)]^2 = 1\,444.000$
10		33	1 880	$[1\,880 - 51.250(33)]^2 = 35\,626.563$
	总计	128	6 560	51 281.378
			$\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2$	

总体总量

总体总量的点估计表达式如下：

总体总量的点估计

$$\hat{X} = M\bar{x}_c \tag{21.32}$$

这个点估计的标准误差估计量为

$$s_{\hat{x}} = Ms_{\bar{x}} \tag{21.33}$$

因此总体总量近似 95% 的置信区间表达式如下所示：

总体总量近似 95% 的置信区间

$$M\bar{x}_c \pm 2s_{\hat{x}} \tag{21.34}$$

对于注册会计师示例的抽样调查，有

$$\hat{X} = M\bar{x}_c = 12\,000(51\,250) = 615\,000\,000(\text{美元})$$

$$s_{\hat{x}} = Ms_{\bar{x}} = 12\,000(1\,979) = 23\,748\,000(\text{美元})$$

因此，根据公式(21.34)我们可以得到近似 95% 的置信区间为：615 000 000 美元 ±2(23 748 000 美元) = 615 000 000 美元 ±47 496 000 美元，或者 567 504 000 美元至 662 496 000 美元。

总体比率

整群抽样总体比率点估计表达式如下：

总体比率点估计

$$\bar{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \tag{21.35}$$

式中　a_i——第 i 群中具备某种特性的个体数量。

这个点估计的标准误差估计量为

$$s_{\bar{p}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2}{n-1}} \tag{21.36}$$

因此，总体比率近似 95% 的近似置信区间表达式为

总体比率近似 95% 的近似置信区间

$$\bar{p}_c \pm 2\, s_{\bar{p}_c} \tag{21.37}$$

对于注册会计师抽样调查，我们可以利用公式(21.35)和表 21.4 中的资料，计算得出女性注册会计师比例的估计量为

$$\bar{p}_c = \frac{2+8+\cdots+9}{8+25+\cdots+33} = \frac{35}{128} = 0.273\,4$$

在表 21.6 中我们给出了估计标准误差所需要的部分计算结果：

表 21.6　注册会计师抽样调查中估计标准误差所需要的部分计算结果，其中 $\bar{p}_c = 0.273\,4$

事务所(i)		M_i	a_i	$(a_i - 0.273\,4M_i)^2$
1		8	2	$[2-0.273\,4(8)]^2 = 0.035\,0$
2		25	8	$[8-0.273\,4(25)]^2 = 1.357\,2$
3		4	0	$[0-0.273\,4(4)]^2 = 1.196\,0$
4		17	6	$[6-0.273\,4(17)]^2 = 1.828\,4$
5		7	1	$[1-0.273\,4(7)]^2 = 0.835\,0$
6		3	2	$[2-0.273\,4(3)]^2 = 1.391\,9$
7		15	2	$[2-0.273\,4(15)]^2 = 4.414\,2$
8		4	0	$[0-0.273\,4(4)]^2 = 1.196\,0$
9		12	5	$[5-0.273\,4(12)]^2 = 2.955\,6$
10		33	9	$[9-0.273\,4(33)]^2 = 0.000\,5$
	总计	128	35	15.209 8
			$\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2$	

$$\sum_{i=1}^{n}(a_i-\bar{p}_c M_i)^2 = 15.2098$$

因此，有

$$s_{\bar{p}_c} = \sqrt{\left[\frac{1\,000-10}{(1\,000)(10)(12)^2}\right]\frac{15.2098}{10-1}} = 0.0341$$

所以利用公式(21.37)我们可以得到女性注册会计师比例近似95%的置信区间为：0.2734±2(0.0341)=0.2734±0.0682，或者0.2052—0.3416。

确定样本容量

群一旦形成以后，样本容量确定的基本问题即为选取群的数量 n。群即抽样的过程类似于其他的抽样方法。首先选择 B 的数值，即允许误差的数值，确定一个可以接受的精度水平，然后建立满足需要的精度 n 数值的计算公式。

每群平均个体数量和群与群之间的方差是确定样本中包含群的数量多少的关键因素。如果整群比较相似，则群与群之间的方差比较小，因此抽样群数将会减少。除此之外，如果每个群的平均个体数量较大，则抽样群数将会减少。精确确定样本容量的公式将会在有关抽样的比较高级的教材中进行论述。

练　习

方　法

14. 自测题 某个总体包含N=25个群，M=300个个体，从中选择4个群。每个群的 M_i，x_i，和 a_i 数值如下所示：

整群（i）	M_i	x_i	a_i
1	7	95	1
2	18	325	6
3	15	190	6
4	10	140	2
总计	50	750	15

a. 计算总体均值、总体总量和总体比率的点估计值。

b. 估计(a)中每个估计量的标准误差。

c. 建立总体均值近似95%的置信区间。

d. 建立总体总量近似95%的置信区间。

e. 建立总体比率近似95%的置信区间。

15. 某个总体包含 $N=30$ 个群，$M=600$ 个个体，从中选择 6 个群。每个群的 M_i，x_i 和 a_i 数值如下所示：

整群(i)	M_i	x_i	a_i
1	35	3 500	3
2	15	965	0
3	12	960	1
4	23	2 070	4
5	20	1 100	3
6	25	1 805	2
总计	130	10 400	13

a. 计算总体均值、总体总量和总体比率的点估计值。

b. 建立总体均值近似 95% 的置信区间。

c. 建立总体总量近似 95% 的置信区间。

d. 建立总体比率近似 95% 的置信区间。

应　用

16. 某个公共事业公司正在进行一项调查，以获得影响机械工程师在新的商务建筑物中选择供暖、通风和空调系统的因素。在公共事业公司的服务区域内，从事供暖、通风和空调系统设计施工的公司总共有 120 家。该公共实业公司计划采用整群抽样的调查方式，每个公司为一个群，对于抽样的公司中的每个机械工程师均进行调查采访。这 120 个公司雇用了大约 500 名机械工程师，从其中挑选出 10 个公司，将每个受采访者的年龄和是否参加地方大学培训记录下来。

整群(i)	M_i	被采访者年龄的总和	参与地方大学培训的次数
1	12	520	8
2	1	33	0
3	2	70	1
4	1	29	1
5	6	270	3
6	3	129	2
7	2	102	0
8	1	48	1
9	9	337	7
10	13	462	12
总计	50	2 000	35

a. 估计从事这种工作的机械工程师的平均年龄。

b. 在公共事业公司的服务区域内，估计参加过地方大学培训的工程师比率。

c. 建立机械工程师平均年龄近似95%的置信区间。

d. 在公共事业公司的服务区域内，对参加过地方大学培训的机械工程师的比率建立近似95%的置信区间。

17. 某家全国性的房地产公司近期收购了一家小型公司，这家小型公司在洛杉矶和南加利福尼亚拥有150个办事处和6 000名代理人，这家全国性的房地产公司为了获悉它的新雇员的态度看法和其他特性，进行了一项抽样调查。他们从新公司中抽取了8个办事处，对这些办事处所有的代理人进行了问卷调查，调查结果如下所示：

办事处	代理人	平均年龄	大学毕业生数	男代理人数
1	17	37	3	4
2	35	32	14	12
3	26	36	8	7
4	66	30	38	28
5	43	41	18	12
6	12	52	2	6
7	48	35	20	17
8	57	44	25	26

a. 估计代理人的平均年龄。

b. 估计其中大学毕业生以及男性代理人的比率。

c. 建立代理人平均年龄近似95%的置信区间。

d. 建立大学毕业生比率近似95%的置信区间。

e. 建立男性代理人比率近似95%的置信区间。

21.7 系统抽样

我们经常用**系统抽样**(systematic sampling)，对于某些抽样情况，尤其是大型的总体，通过先确定随机数，然后根据抽样框寻找与随机数相对应的个体的方法选取一个简单随机样本，这种方法是需要耗费大量时间的。在这种情形下，我们可以用系统抽样代替简单随机抽样。例如，从容量为5 000的总体中抽取一个容量为50的样本，我们可以从总体每100个个体中选取一个个体。对于这种情形的系统样本，我们从抽样框前100个个体中随机选择一个；然后根据所选的第一个个体位置，在后面的抽样框中每隔100个再选择一个，这样就可以得到样本中其余的个体。事实上，我们通过对总体进行系统的排列，在随机抽取第一个个体之后，每隔100个选择一个个体，就可以得到一个容量为50的样本，利用这种方式选择容量为50的样本常常比利用简单随机抽样方式更加容易。由于第一个个体的选择是随机的，因此系统样本常常假定具有简单随机样本的性质。如果抽样框是由总体中的个体随机排列而成的，则这种假设通常是可行的。

小　结

本章我们扼要介绍了有关调查抽样的一些知识，抽样调查的目的在于搜集比如总体均值、总体总量和总体比率等总体参数的资料。抽样调查作为搜集资料的方法与进行实验所产生的数据是相应的。在进行抽样调查的过程中，确定资料搜集的抽样计划设计是非常重要的。而在进行实验时，决定产生哪些数据的实验设计问题同样也是非常重要的。

抽样调查有可能会产生两种类型的误差：抽样误差和非抽样误差。抽样误差是由于仅仅使用同一个样本而不是整个总体估计总体参数而产生的误差。非抽样误差则包含所有可能出现的其他类型的误差，比如测量误差、访谈误差、缺乏回复误差和处理误差等等。我们可以通过设计合适的问卷调查表、对访谈员进行全面培训和仔细审核数据等手段，控制非抽样误差。而且通过选取合适的样本设计和样本容量，可以使抽样误差最小化。

本章我们讨论了四种常用的样本设计：简单随机抽样、分层简单随机抽样、整群抽样和系统抽样。样本设计的目的是花费最少的经费得到最为准确的估计。如果将总体划分为若干层，并使各层的个体相对一致，则分层简单随机抽样所获得的精度将比简单随机抽样所获得的精度更高(近似置信区间范围更窄)。如果将个体纳入每一个群中，并使每个群的所有个体在地理上连成一体，则整群抽样就能减少调查采访的费用。对于这种情形，利用整群抽样我们可以使每个单位成本均获得较高的精度。我们经常利用系统随机抽样来代替简单随机抽样。

术语释义

个体：作为数据来源的一个实体。
总体：有待研究的所有个体集合。
样本：总体的子集。
目标总体：有待进行统计推断的总体。
抽样总体：样本来源的总体。
抽样单位：供抽样选取的单位，一个抽样单位可以包含很多个体。
抽样框：有待研究的抽样单位的表单，根据抽样框选择单位以获得样本。
概率抽样方法：一种可以计算每个可能样本的概率的抽样方法。
非概率抽样方法：一种抽样方法，这种方法无法计算每个可能样本的概率。
方便抽样：一种基于方便性原则来选择个体的非概率抽样方法。
判断抽样：一种根据对所研究问题的判断来选择个体的非概率抽样方法。
抽样误差：由于使用一个样本，而不是整个总体来估计总体参数而产生的误差。
非抽样误差：除了抽样误差以外所有的误差，比如测量误差、调查误差和处理误差。
简单随机样本：按一种方法选取的样本，这种方法可以使每一个容量为 n 的样本被选中的概率相

同。

允许误差： 一个数值，点估计值加上或者减去它可以建立近似95%的置信区间。它是点估计标准误差的2倍。

分层简单随机抽样： 一种选择样本的概率方法，它首先将总体划分为层，然后从每个层抽取一个简单随机样本。

整群抽样： 一种选择样本的概率方法，它首先将总体划分为群，然后从其中选取一个或者多个群，在单阶段整群抽样中，中选群的每一个个体均被选中；在二阶段整群抽样中，从已经中选的个体中再选取一个样本。

系统抽样： 一种选择样本的方法，这种方法首先随机选择第一个个体，然后每隔 k 个距离再选择一个个体。

重要公式

简单随机抽样

总体均值的区间估计

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} \tag{21.1}$$

总体均值的标准误差估计值

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right) \tag{21.2}$$

总体均值的区间估计

$$\bar{x} \pm z_{\alpha/2}s_{\bar{x}} \tag{21.3}$$

总体均值近似95%的置信区间

$$\bar{x} \pm 2\,s_{\bar{x}} \tag{21.4}$$

总体总量的点估计

$$\hat{X} = N\bar{x} \tag{21.5}$$

标准误差的估计量

$$s_{\hat{X}} = Ns_{\bar{x}} \tag{21.6}$$

总体总量近似95%的置信区间

$$N\bar{x} \pm 2\,s_{\hat{X}} \tag{21.8}$$

总体比率标准误差的估计值

$$s_{\bar{p}} = \sqrt{\left(\frac{N-n}{N}\right)\left(\frac{\bar{p}(1-\bar{p})}{n-1}\right)} \tag{21.9}$$

总体比率近似 95% 的置信区间

$$\bar{p} \pm 2s_{\bar{p}} \tag{21.10}$$

估计总体均值的样本容量

$$n = \frac{Ns^2}{N\left(\frac{B^2}{4}\right) + s^2} \tag{21.12}$$

估计总体总量的样本容量

$$n = \frac{Ns^2}{\left(\frac{B^2}{4N}\right) + s^2} \tag{21.13}$$

估计总体比率的样本容量

$$n = \frac{N\bar{p}(1-\bar{p})}{N\left(\frac{B^2}{4}\right) + \bar{p}(1-\bar{p})} \tag{21.14}$$

分层简单随机抽样

总体均值的点估计

$$\bar{x}_{st} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\bar{x}_h \tag{21.15}$$

总体均值标准误差的估计量

$$s_{\bar{x}_{st}} = \sqrt{\frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{s_h^2}{n_h}} \tag{21.16}$$

总体均值近似 95% 的置信区间

$$\bar{x}_{st} \pm 2s_{\bar{x}_{st}} \tag{21.17}$$

总体总量的点估计

$$\hat{X} = N\bar{x}_{st} \tag{21.18}$$

$\hat{X}$ 的标准误差估计值

$$s_{\hat{X}} = N s_{\bar{x}_{st}} \tag{21.19}$$

总体总量 95% 的近似置信区间

$$N\bar{x}_{st} \pm 2\, s_{\hat{X}} \tag{21.20}$$

总体比率的点估计

$$\bar{p}_{st} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{p}_h \tag{21.21}$$

标准误差的估计值

$$s_{\bar{p}_{st}} = \sqrt{\frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \left[\frac{\bar{p}_h (1 - \bar{p}_h)}{n_h - 1}\right]} \tag{21.22}$$

总体比率近似 95% 的置信区间

$$\bar{p}_{st} \pm 2\, s_{\bar{p}_{st}} \tag{21.23}$$

总样本容量 n 在层间的分配：Neyman 分配

$$n_h = n\left(\frac{N_h\, s_h}{\sum_{h=1}^{H} N_h\, s_h}\right) \tag{21.24}$$

估计总体均值时的样本容量

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h\, s_h^2} \tag{21.25}$$

估计总体总量时的样本容量

$$n = \frac{\left(\sum_{h=1}^{H} N_h s_h\right)^2}{\frac{B^2}{4} + \sum_{h=1}^{H} N_h\, s_h^2} \tag{21.26}$$

估计总体比率时的样本容量

$$n = \frac{\left(\sum_{h=1}^{H} N_h \sqrt{\bar{p}_h (1 - \bar{p}_h)}\right)^2}{N^2\left(\frac{B^2}{4}\right) + \sum_{h=1}^{H} N_h \bar{p}_h (1 - \bar{p}_h)} \tag{21.27}$$

总样本容量 n 在层间的比例分配

$$n_h = n\left(\frac{N_h}{N}\right) \tag{21.28}$$

整群抽样

总体均值点估计

$$\bar{x}_c = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i} \tag{21.29}$$

总体均值的标准误差估计值

$$s_{\bar{x}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(x_i - \bar{x}_c M_i)^2}{n-1}} \tag{21.30}$$

总体均值近似 95% 的置信区间

$$\bar{x}_c \pm 2\, s_{\bar{x}_c} \tag{21.31}$$

总体总量的点估计

$$\hat{X} = M\bar{x}_c \tag{21.32}$$

$\hat{X}$ 的标准误差估计量

$$s_{\hat{X}} = M s_{\bar{x}_c} \tag{21.33}$$

总体总量近似 95% 的置信区间

$$M\bar{x}_c \pm 2\, s_{\hat{X}} \tag{21.34}$$

总体比率点估计

$$\bar{p}_c = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M_i} \tag{21.35}$$

$\bar{p}_c$ 的标准误差估计量

$$s_{\bar{p}_c} = \sqrt{\left(\frac{N-n}{Nn\bar{M}^2}\right)\frac{\sum_{i=1}^{n}(a_i - \bar{p}_c M_i)^2}{n-1}} \tag{21.36}$$

总体比率近似 95% 的近似置信区间

$$\bar{p}_c \pm 2 s_{\bar{p}_c} \tag{21.37}$$

补充练习

18. 为了确定 Miller Lite 啤酒的系列广告的消费者认可程度，Louis Harris 在全国范围内对 363 名收看过 Miller Lite 广告的成年人进行了一项民意测验（*USA Today*, November 17, 1997），以下的回复是根据调查结果得出的(注：由于仅仅调查了很小一部分观众，因此在任何包含标准误差的公式中，可令 $(N-n)/N=1$)。
    a. 有 19% 的受访者表示他们对这个新广告感觉不错，请对此建立总体比率 95% 的置信区间。
    b. 有 31% 的受访者表示他们不喜欢这个新广告，请建立总体比率 95% 的置信区间。
    c. 有 17% 的受访者认为这个新广告很有效，对认为广告很有效的成年人比率建立 95% 的置信区间。
    d. Louis Harris 的调查报告显示允许误差为 5%，这个结果代表着什么？对他们得到这个结果你怎样理解？
    e. 对于这样的抽样调查，怎样才能不产生抽样偏差？

19. 《华尔街日报》就它的关联期刊面向它的订阅者进行了一项调查，第一个问题调查了 504 名受访者，询问他们在旅行时是否携带便携式电脑，55% 的受访者回答说携带。另一个问题调查的是他们在旅行时是否使用特快专递或者包裹邮寄服务，31% 的被采访者给出了肯定的回答（*The Wall Street Journal Interactive Edition Subscriber Survey*, 2000）。
    a. 建立使用便携式电脑的受访者比率的标准差估计。
    b. 建立使用特快专递或者包裹邮寄服务的受访者比率的标准差估计。
    c. (a)和(b)中的标准差估计是否一样？如果它们不同，请解释原因。
    d. 建立使用便携式电脑的受访者比率的 95% 的置信区间。
    e. 建立使用特快专递或者邮寄包裹服务的受访者比率的 95% 的置信区间。

20. 对一家制造公司的职员进行了一项关于生活质量的调查，在这家制造公司的 3 000 名雇员中，抽取了 300 名进行问卷调查，获得了 200 份有效调查问卷，回复率为 67%。
    a. 样本中年薪平均值 = 23 200 美元，标准差 s = 3 000 美元，对总体的平均年薪建立近似 95% 的置信区间。
    b. 利用(a)中的资料，对 3 000 名雇员的年薪总量建立近似 95% 的置信区间。
    c. 73% 的被调查者认为他们对工作总体上比较满意，对这一总体比率建立近似 95% 的置信区间。
    d. (c)中的结果是否存在误差？如果你知道受访者是匿名的，你的结论是否会发生改变？

21. 美国参议院司法委员会的报告显示了 1991 年每个州的杀人犯数量，在印第安纳州、俄亥俄州、肯塔基州杀人犯的数量分别为 380，760 和 260 人。为了更好地了解受害者和死亡原因，他们采用分层随机抽样得到了以下的调查结果：

层	样本容量	枪击死亡	殴打死亡	黑人受害者
印第安纳州	30	10	9	21
俄亥俄州	45	19	12	34
肯塔基州	25	7	11	15

a. 建立印第安纳州枪击死亡比率的近似 95% 的置信区间。

b. 计算俄亥俄州枪击死亡总量的估计值。

c. 建立俄亥俄州枪击死亡比率的近似 95% 的置信区间。

d. 建立三个州枪击死亡比率的近似 95% 的置信区间。

22. 重新参考练习 21 中的数据。

a. 计算三个州殴打死亡总量的估计值。

b. 建立三个州殴打死亡比率近似 95% 的置信区间。

c. 建立黑人受害者比率的近似 95% 的置信区间。

d. 计算黑人受害者总量的估计值。

23. 利用分层简单随机抽样获得银行顾客的一个样本，以了解银行顾客的某些意见以及一些统计上的问题。分层是以 2001 年 6 月 30 日账户的存款余额为根据的，各层账户余额的分布以及标准差结果如下所示：

层（美元）	账户数量	账户余额的标准差
0.00—1 000.00	3 000	80
1 000.01—2 000.00	600	150
2 000.01—5 000.00	250	220
5 000.01—10 000.00	100	700
超过 10 000.00	50	3 000

a. 假设每一层的抽样单位成本是近似相等的，请确定样本中所应包含的人数。并假设估计账户的平均余额允许误差为 $B=20$ 美元。

b. 利用 Neyman 分配确定每一层的样本个体数量。

24. 某个公共事业代理机构打算了解某个特定城市小型私人护理医院的住院病人情况。在这个城市共有 100 家私人护理医院，总共照顾着 4 800 名病人，从中抽取了 6 家作为样本，对这 6 家医院中的每一个病人均进行了访谈，抽样调查的部分结果如下所示：

护理医院	住院人数	病人的平均年龄	残疾病人数量
1	14	61	12
2	7	74	2

护理医院	住院人数	病人的平均年龄	残疾病人数量
3	96	78	30
4	23	69	8
5	71	73	10
6	29	84	22

a. 建立这个城市私人护理医院住院病人的平均年龄的估计值。

b. 建立这个城市私人护理医院中的残疾病人比率的近似95%的置信区间。

c. 估计这个城市私人护理医院所住病人的总数。

附　录

附录 A　参考文献

附录 B　表格

附录 C　求和符号

附录 D　偶数习题的答案

附录 E　自测题解答

附录 A 参考文献

通用文献

Bowerman, B. L., and R. T. O'Connell, *Applied Statistics: Improving Business Processes*, Irwin, 1996.

Freedman, D., R. Pisani, and R. Purves, Statistics, 3rd ed., W. W. Norton, 1997.

Hogg, R. V., and A. T. Craig, *Introduction to Mathematical Statistics*, 5th ed., Prentice-Hall, 1994.

Hogg, R. V., and E. A. Tanis, *Probability and Statistical Inference*, 6th ed., Prentice Hall, 2001.

Joiner, B. L., and B. F. Ryan, *Minitab Handbook*, Brooks/Cole, 2000.

Miller, I., and M. Miller, *John E. Freund's Mathematical Statistics*, Prentice Hall. 1998.

Roberts, H., *Data Analysis for Managers with Minitab*, Scientific Press, 1991.

Tanur, J. M., *Statistics: A Guide to the Unknown*, 4thed., Brooks/Cole, 2002.

Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

实验设计

Cochran, W. G., and G. M. Cox, *Experimental Designs*, 2d ed., Wiley, 1992.

Hicks, C. R., and K. V. Turner, *Fundamental Concepts in the Design of Experiments*, 5th ed., Oxford University Press, 1999.

Montgomery, D. C., *Design and Analysis of Experiments*, 5thed., Wiley, 2000.

Winer, B. J., K. M. Michels, and D. R. Brown, *Statistical Principles in Experinental Design*, 3rd ed., McGraw-Hill, 1991.

预 测

Bowerman, B. L., and R. T. O' Connell, *Forecasting and Time Series: An Applied Approach*, 3rd ed., Duxbury, 1993.

Box, G. E. P., G. C. Reinsel, and G. Jenkins, *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice-Hall, 1994.

Makridakis, S., S. C. Wheelwright, and R. J. Hyudman, *Forecasting: Methods and Applications*, 3rd ed., Wiley, 1997.

指标值

U. S. Department of Commerce, *Survey of Current Business.*

U. S. Department of Labor, Bureau of Labor Statistics, *CPIDe-tailed Report.*

U. S. Department of Labor, *Producer Price Indexes.*

非参数方法

Conover, W. J., *Practical Nonparametric Statistics*, 3rd ed., Wiley, 1998.

Gibbons, J. D., and S. Chakraborti, *Nonparametric Statistical Inference*, 3rd ed., Marcel Dekker, 1992.

Siegel, S., and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*, 2d ed., McGraw-Hill, 1988.

Sprent, P., *Applied Non-Parametric Statistical Methods*, CRC, 1993.

概 率

Feller, W., *An Introduction to Probability Theory and Its Application*, Vol. I, 3rd ed., Wiley, 1968.

Hogg, R. V., and E. A. Tanis, *Probability and Statistical Inference*, 6th ed., Prentice-Hall, 2001.

Ross, S. M., *Introduction to Probability Models*, 7th ed., Academic Press, 2000.

Wackerly, D. D., W. Mendenhall, and R. L. Scheaffer, *Mathematical Statistics with Applications*, 5th ed., PWS, 1996.

质量控制

Deming, W. E., *Quality, Productivity, and Competitive Position*, MIT, 1982.

Duncan, A. J., *Quality Control and Industrial Statistics*, 5th ed., Irwin, 1986.

Evans, J. R., and W. M. Lindsay, *The Management and Control of Quality*, 4th ed., South-Western, 1998.

Gryna, F. M., and I. M. Juran, *Quality Planning and Analysis: From Product Development Through Use*, 3rd ed., McGraw-Hill, 1993.

Ishikawa, K., *Introduction to Quality Control*, Kluwer Academic, 1991.

Montgomery, D. C., *Introduction to Statistical Quality Control*, 3rd ed., Wiley, 1996.

回归分析

Belsley, D. A., *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, Wiley, 1991.

Chatterjee, S., and B. Price, *Regression Analysis by Example*, 3rd ed., Wiley, 1999.

Draper, N. R., and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, 1998.

Graybill, F. A., and H. Iyer, *Regression Analysis: Concepts and Applications*, Duxbury Press, 1994.

Kleinbaum, D. G., L. L. Kupper, and K. E. Muller, *Applied Regression Analysis and Other Multivariate Methods*, 3rd ed., Duxbury Press, 1997.

Kutner, M. H., C. J. Nachtschiem, W. Wasserman, and J. Neter, *Applied Linear Statistical Models*, 4th ed., Irwin, 1996.

Mendenhall, M., and T. Sincich, *A Second Course in Statistics: Regression Analysis*, 5th ed., Prentice Hall, 1996.

Myers, R. H., *Classical and Modern Regression with Applications*, 2d ed., PWS, 1990.

Wonnacott, T. H., and R. J. Wonnacott, *Regression: A Second Course in Statistics*, Krieger, 1986.

抽 样

Cochran, W. G., *Sampling Techniques*, 3rded., Wiley, 1977. Deming, W. E., *Some Theory of Sampling*, Dover, 1984.

Hansen, M. H., W. N. Hurwitz, W. G. Madow, and M. N. Hanson, *Sample Survey Methods and Theory*, Wiley, 1993.

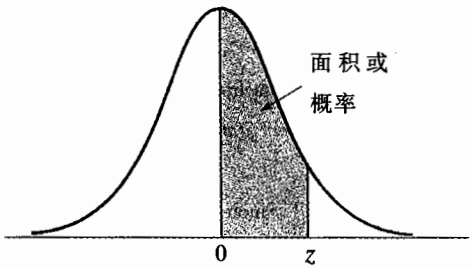
Kish, L., *Survey Sampling*, Wiley, 1995.

Levy, P. S., and S. Lemeshow, *Sampling of Populations: Methods and Applications*, 3rd ed., Wiley, 1999.

Scheaffer, R. L., W. Mendenhall, and L. Ott, *Elementary Survey Sampling*, 5th ed., Duxbury, 1996.

附录 B 表格

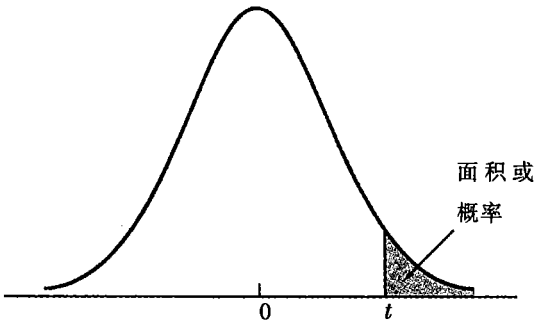
表格 1 标准正态分布



表中的值给出了曲线下方、水平线上方和均值与标准差 z 之间的面积。例如，$z=1.25$，在均值和 z 之间、曲线之下的面积是 0.394 4。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.000 0	.004 0	.008 0	.012 0	.016 0	.019 9	.023 9	.027 9	.031 9	.035 9
.1	.039 8	.043 8	.047 8	.051 7	.055 7	.059 6	.063 6	.067 5	.071 4	.075 3
.2	.079 3	.083 2	.087 1	.091 0	.094 8	.098 7	.102 6	.106 4	.110 3	.114 1
.3	.117 9	.121 7	.125 5	.129 3	.133 1	.136 8	.140 6	.144 3	.148 0	.151 7
.4	.155 4	.159 1	.162 8	.166 4	.170 0	.173 6	.177 2	.180 8	.184 4	.187 9
.5	.191 5	.195 0	.198 5	.201 9	.205 4	.208 8	.212 3	.215 7	.219 0	.222 4
.6	.225 7	.229 1	.232 4	.235 7	.238 9	.242 2	.245 4	.248 6	.251 8	.254 9
.7	.258 0	.261 2	.264 2	.267 3	.270 4	.273 4	.276 4	.279 4	.282 3	.285 2
.8	.288 1	.291 0	.293 9	.296 7	.299 5	.302 3	.305 1	.307 8	.310 6	.313 3
.9	.315 9	.318 6	.321 2	.323 8	.326 4	.328 9	.331 5	.334 0	.336 5	.338 9
1.0	.341 3	.343 8	.346 1	.348 5	.350 8	.353 1	.355 4	.357 7	.359 9	.362 1
1.1	.364 3	.366 5	.368 6	.370 8	.372 9	.374 9	.377 0	.379 0	.381 0	.383 0
1.2	.384 9	.386 9	.388 8	.390 7	.392 5	.394 4	.396 2	.398 0	.399 7	.401 5
1.3	.403 2	.404 9	.406 6	.408 2	.409 9	.411 5	.413 1	.414 7	.416 2	.417 7
1.4	.419 2	.420 7	.422 2	.423 6	.425 1	.426 5	.427 9	.429 2	.430 6	.431 9
1.5	.433 2	.434 5	.435 7	.437 0	.438 2	.439 4	.440 6	.441 8	.442 9	.444 1
1.6	.445 2	.446 3	.447 4	.448 4	.449 5	.450 5	.451 5	.452 5	.453 5	.454 5
1.7	.455 4	.456 4	.457 3	.458 2	.459 1	.459 9	.460 8	.461 6	.462 5	.463 3
1.8	.464 1	.464 9	.465 6	.466 4	.467 1	.467 8	.468 6	.469 3	.469 9	.470 6
1.9	.471 3	.471 9	.472 6	.473 2	.473 8	.474 4	.475 0	.475 6	.476 1	.476 7
2.0	.477 2	.477 8	.478 3	.478 8	.479 3	.479 8	.480 3	.480 8	.481 2	.481 7
2.1	.482 1	.482 6	.483 0	.483 4	.483 8	.484 2	.484 6	.485 0	.485 4	.485 7
2.2	.486 1	.486 4	.486 8	.487 1	.487 5	.487 8	.488 1	.488 4	.488 7	.489 0
2.3	.489 3	.489 6	.489 8	.490 1	.490 4	.490 6	.490 9	.491 1	.491 3	.491 6
2.4	.491 8	.492 0	.492 2	.492 5	.492 7	.492 9	.493 1	.493 2	.493 4	.493 6
2.5	.493 8	.494 0	.494 1	.494 3	.494 5	.494 6	.494 8	.494 9	.495 1	.495 2
2.6	.495 3	.495 5	.495 6	.495 7	.495 9	.496 0	.496 1	.496 2	.496 3	.496 4
2.7	.496 5	.496 6	.496 7	.496 8	.496 9	.497 0	.497 1	.497 2	.497 3	.497 4
2.8	.497 4	.497 5	.497 6	.497 7	.497 7	.497 8	.497 9	.497 9	.498 0	.498 1
2.9	.498 1	.498 2	.498 2	.498 3	.498 4	.498 4	.498 5	.498 5	.498 6	.498 6
3.0	.498 6	.498 7	.498 7	.498 8	.498 8	.498 9	.498 9	.498 9	.499 0	.499 0

表格 2 t 分布

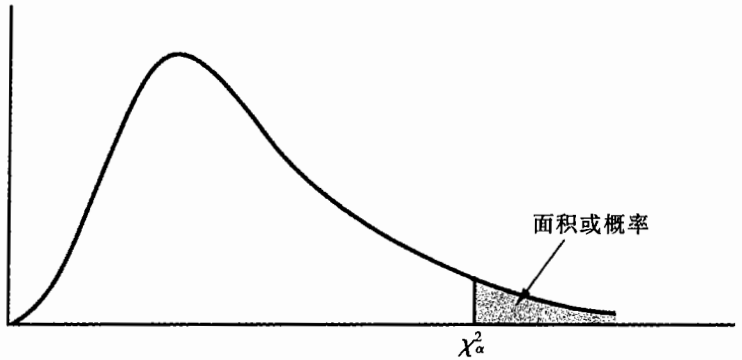


表中数字给出与 t 分布上侧的面积或概率相应的 t 值。例如，自由度为 10 和上侧面积为 0.05 时，$t_{0.05} = 1.812$。

自由度	上侧的面积				
	.10	.05	.025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

该表经牛津大学出版社许可重印，基于 Biometrika Trustees 的表 12，t 分布的百分点见 E. S. Pearson 和 H. O. Hartley, *Biometrika* 统计表，1966 年第 1 卷第 3 版。

表格 3　χ^2 分布



表中数字给出χ^2_α的值，其中 α 是 χ^2 分布上侧的面积或概率。例如，自由度为 10 和上侧面积是 0.01 时，$\chi^2_{0.01} = 23.209\ 3$.

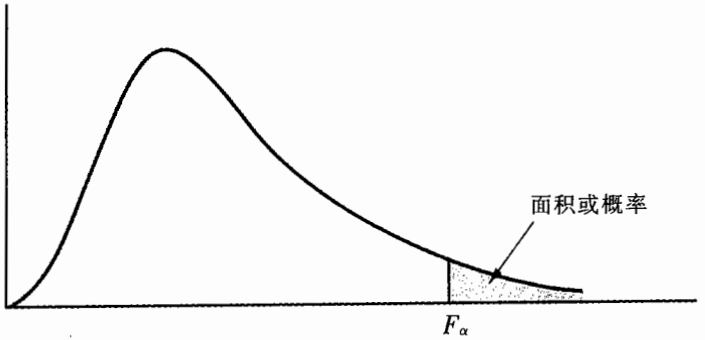
	上侧分布									
自由度	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	$392\ 704 \times 10^{-10}$	$157\ 088 \times 10^{-9}$	$982\ 069 \times 10^{-9}$	$393\ 214 \times 10^{-8}$	.015 790 8	2.705 54	3.841 46	5.023 89	6.634 90	7.879 44
2	.010 025 1	.020 100 7	.050 635 6	.102 587	.210 720	4.605 17	5.991 47	7.377 76	9.210 34	10.596 6
3	.071 721 2	.114 832	.215 795	.351 846	.584 375	6.251 39	7.814 73	9.348 40	11.344 9	12.838 1
4	.206 990	.297 110	.484 419	.710 721	1.063 623	7.779 44	9.487 73	11.143 3	13.276 7	14.860 2
5	.411 740	.554 300	.831 211	1.145 476	1.610 31	9.236 35	11.070 5	12.832 5	15.086 3	16.749 6
6	.675 727	.872 085	1.237 347	1.635 39	2.204 13	10.644 6	12.591 6	14.449 4	16.811 9	18.547 6
7	.989 265	1.239 043	1.689 87	2.167 35	2.833 11	12.017 0	14.067 1	16.012 8	18.475 3	20.277 7
8	1.344 419	1.646 482	2.179 73	2.732 64	3.489 54	13.361 6	15.507 3	17.534 6	20.090 2	21.955 0
9	1.734 926	2.087 912	2.700 39	3.325 11	4.168 16	14.683 7	16.919 0	19.022 8	21.666 0	23.589 3
10	2.155 85	2.558 21	3.246 97	3.940 30	4.865 18	15.987 1	18.307 0	20.483 1	23.209 3	25.188 2
11	2.603 21	3.053 47	3.815 75	4.574 81	5.577 79	17.275 0	19.675 1	21.920 0	24.725 0	26.756 9
12	3.073 82	3.570 56	4.403 79	5.226 03	6.303 80	18.549 4	21.026 1	23.336 7	26.217 0	28.299 5
13	3.565 03	4.106 91	5.008 74	5.891 86	7.041 50	19.811 9	22.362 1	24.735 6	27.688 3	29.819 4
14	4.074 68	4.660 43	5.628 72	6.570 63	7.789 53	21.064 2	23.684 8	26.119 0	29.141 3	31.319 3

（续表）

自由度	上侧分布									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
15	4.600 94	5.229 35	6.262 14	7.260 94	8.546 75	22.307 2	24.995 8	27.488 4	30.577 9	32.801 3
16	5.142 24	5.812 21	6.907 66	7.961 64	9.312 23	23.541 8	26.296 2	28.845 4	31.999 9	34.267 2
17	5.697 24	6.407 76	7.564 18	8.671 76	10.085 2	24.769 0	27.587 1	30.191 0	33.408 7	35.718 5
18	6.264 81	7.014 91	8.230 75	9.390 46	10.864 9	25.989 4	28.869 3	31.526 4	34.805 3	37.156 4
19	6.843 98	7.632 73	8.906 55	10.11 70	11.650 9	27.203 6	30.143 5	32.852 3	36.190 8	38.582 2
20	7.433 86	8.260 40	9.590 83	10.850 8	12.442 6	28.412 0	31.410 4	34.169 6	37.566 2	39.996 8
21	8.033 66	8.897 20	10.282 93	11.591 3	13.239 6	29.615 1	32.670 5	35.478 9	38.932 1	41.401 0
22	8.642 72	9.542 49	10.982 3	12.338 0	14.041 5	30.813 3	33.924 4	36.780 7	40.289 4	42.795 8
23	9.260 42	10.195 67	11.688 5	13.090 5	14.847 9	32.006 9	35.172 5	38.075 7	41.638 4	44.181 3
24	9.886 23	10.856 4	12.401 1	13.848 4	15.658 7	33.196 3	36.415 1	39.364 1	42.979 8	45.558 5
25	10.519 7	11.524 0	13.119 7	14.611 4	16.473 4	34.381 6	37.652 5	40.646 5	44.314 1	46.927 8
26	11.160 3	12.198 1	13.843 9	15.379 1	17.291 9	35.563 1	38.885 2	41.923 2	45.641 7	48.289 9
27	11.807 6	12.878 6	14.573 3	16.151 3	18.113 8	36.741 2	40.113 3	43.194 4	46.963 0	49.644 9
28	12.461 3	13.564 8	15.307 9	16.927 9	18.939 2	37.915 9	41.337 2	44.460 7	48.278 2	50.993 3
29	13.121 1	14.256 5	16.047 1	17.708 3	19.767 7	39.087 5	42.556 9	45.722 2	49.587 9	52.335 6
30	13.786 7	14.953 5	16.790 8	18.492 6	20.599 2	40.256 0	43.772 9	46.979 2	50.892 2	53.672 0
40	20.706 5	22.164 3	24.433 1	26.509 3	29.050 5	51.805 0	55.758 5	59.341 7	63.690 7	66.765 9
50	27.990 7	29.706 7	32.357 4	34.764 2	37.688 6	63.167 1	67.504 8	71.420 2	76.153 9	79.490 0
60	35.534 6	37.484 8	40.481 7	43.187 9	46.458 9	74.397 0	79.081 9	83.297 6	88.379 4	91.951 7
70	43.275 2	45.441 8	48.757 6	51.739 3	55.329 0	85.527 1	90.531 2	95.023 1	100.425	104.215
80	51.172 0	53.540 0	57.153 2	60.391 5	64.277 8	96.578 2	101.879	106.629	112.329	116.321
90	59.196 3	61.754 1	65.646 6	69.126 0	73.291 2	107.565	113.145	118.136	124.116	128.299
100	67.327 6	70.064 8	74.221 9	77.929 5	82.358 1	118.498	124.342	129.561	135.807	140.169

该表经牛津大学出版社允许重印，基于 Biometrika Trustees 的表 8，χ^2 分布的百分点见 E. S. Pearson 和 H. O. Hartley, *Biometrika* 统计表，1996 年第 1 卷第 3 版。

表格 4　F 分布



表中的数字给出 F_α 的值，其中 α 是 F 分布上侧的面积或概率。例如，第一自由度为 12、第二自由度为 15 和上侧面积是 0.05 时，$F_{0.05}=2.48$。

$F_{0.05}$ 上侧临界值表

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13

（续表）

$F_{0.05}$ 上侧临界值表

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

该表经牛津大学出版社许可重印，基于 Biometrika Trustees 的表 18，*F* 分布的百分点见 E. S. Pearson 和 H. O. Hartley, *Biometrika* 统计表，1966 年第 1 卷第 3 版。

（续表）

$F_{0.025}$ 上侧临界值表

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1 001	1 006	1 010	1 014	1 018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.21	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.20	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

（续表）

$F_{0.01}$ 上侧临界值表

第二自由度 \ 第一自由度	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4 052	4 999.5	5 403	5 625	5 764	5 859	5 928	5 982	6 022	6 056	6 106	6 157	6 209	6 235	6 261	6 287	6 313	6 339	6 366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.06
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

表格 5　二项分布概率

表中的数字给出 n 次二项分布实验中，出现 x 次成功的概率，其中 p 是一次实验中成功的概率。例如，具有成功率为 0.05 的六次实验中，两次成功的概率是 0.030 5。

		p								
n	x	.01	.02	.03	.04	.05	.06	.07	.08	.09
2	0	.980 1	.960 4	.940 9	.921 6	.902 5	.883 6	.864 9	.846 4	.828 1
	1	.019 8	.039 2	.058 2	.076 8	.095 0	.112 8	.130 2	.147 2	.163 8
	2	.000 1	.000 4	.000 9	.001 6	.002 5	.003 6	.004 9	.006 4	.008 1
3	0	.970 3	.941 2	.912 7	.884 7	.857 4	.830 6	.804 4	.778 7	.753 6
	1	.029 4	.057 6	.084 7	.110 6	.135 4	.159 0	.181 6	.203 1	.223 6
	2	.000 3	.001 2	.002 3	.004 6	.007 1	.010 2	.013 7	.017 7	.022 1
	3	.000 0	.000 0	.000 0	.000 1	.000 1	.000 2	.000 3	.000 5	.000 7
4	0	.960 6	.922 4	.885 3	.849 3	.814 5	.780 7	.748 1	.716 4	.685 7
	1	.038 8	.075 3	.109 5	.141 6	.171 5	.199 3	.225 2	.249 2	.271 3
	2	.000 6	.002 3	.005 1	.008 8	.013 5	.019 1	.025 4	.032 5	.040 2
	3	.000 0	.000 0	.000 1	.000 2	.000 5	.000 8	.001 3	.001 9	.002 7
	4	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
5	0	.951 0	.903 9	.858 7	.815 4	.773 8	.733 9	.695 7	.659 1	.624 0
	1	.048 0	.092 2	.132 8	.169 9	.203 6	.234 2	.261 8	.286 6	.308 6
	2	.001 0	.003 8	.008 2	.014 2	.021 4	.029 9	.039 4	.049 8	.061 0
	3	.000 0	.000 1	.000 3	.000 6	.001 1	.001 9	.003 0	.004 3	.006 0
	4	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 2	.000 3
	5	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
6	0	.941 5	.885 8	.833 0	.782 8	.735 1	.689 9	.647 0	.606 4	.567 9
	1	.057 1	.108 5	.154 6	.195 7	.232 1	.264 2	.292 2	.316 4	.337 0
	2	.001 4	.005 5	.012 0	.020 4	.030 5	.042 2	.055 0	.068 8	.083 3
	3	.000 0	.000 2	.000 5	.001 1	.002 1	.003 6	.005 5	.008 0	.011 0
	4	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 3	.000 5	.000 8
	5	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
7	0	.932 1	.868 1	.808 0	.751 4	.698 3	.648 5	.601 7	.557 8	.516 8
	1	.065 9	.124 0	.174 9	.219 2	.257 3	.289 7	.317 0	.339 6	.357 8
	2	.002 0	.007 6	.016 2	.027 4	.040 6	.055 5	.071 6	.088 6	.106 1
	3	.000 0	.000 3	.000 8	.001 9	.003 6	.005 9	.009 0	.012 8	.017 5
	4	.000 0	.000 0	.000 0	.000 1	.000 2	.000 4	.000 7	.001 1	.001 7
	5	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
8	0	.922 7	.850 8	.783 7	.721 4	.663 4	.609 6	.559 6	.513 2	.470 3
	1	.074 6	.138 9	.193 9	.240 5	.279 3	.311 3	.337 0	.357 0	.372 1
	2	.002 6	.009 9	.021 0	.035 1	.051 5	.069 5	.088 8	.108 7	.128 8
	3	.000 1	.000 4	.001 3	.002 9	.005 4	.008 9	.013 4	.018 9	.025 5
	4	.000 0	.000 0	.000 1	.000 2	.000 4	.000 7	.001 3	.002 1	.003 1
	5	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 2
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0

（续表）

		p								
n	x	.01	.02	.03	.04	.05	.06	.07	.08	.09
9	0	.913 5	.833 7	.760 2	.692 5	.630 2	.573 0	.520 4	.472 2	.427 9
	1	.083 0	.153 1	.211 6	.259 7	.298 5	.329 2	.352 5	.369 5	.380 9
	2	.003 4	.012 5	.026 2	.043 3	.062 9	.084 0	.106 1	.128 5	.150 7
	3	.000 1	.000 6	.001 9	.004 2	.007 7	.012 5	.018 6	.026 1	.034 8
	4	.000 0	.000 0	.000 1	.000 3	.000 6	.001 2	.002 1	.003 4	.005 2
	5	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 3	.000 5
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
10	0	.904 4	.817 1	.737 4	.664 8	.598 7	.538 6	.484 0	.434 4	.389 4
	1	.091 4	.166 7	.228 1	.277 0	.315 1	.343 8	.364 3	.377 7	.385 1
	2	.004 2	.015 3	.031 7	.051 9	.074 6	.098 8	.123 4	.147 8	.171 4
	3	.000 1	.000 8	.002 6	.005 8	.010 5	.016 8	.024 8	.034 3	.045 2
	4	.000 0	.000 0	.000 1	.000 4	.001 0	.001 9	.003 3	.005 2	.007 8
	5	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 3	.000 5	.000 9
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
12	0	.886 4	.784 7	.693 8	.612 7	.540 4	.475 9	.418 6	.367 7	.322 5
	1	.107 4	.192 2	.257 5	.306 4	.341 3	.364 5	.378 1	.383 7	.382 7
	2	.006 0	.021 6	.043 8	.070 2	.098 8	.128 0	.156 5	.183 5	.208 2
	3	.000 2	.001 5	.004 5	.009 8	.017 3	.027 2	.039 3	.053 2	.068 6
	4	.000 0	.000 1	.000 3	.000 9	.002 1	.003 9	.006 7	.010 4	.015 3
	5	.000 0	.000 0	.000 0	.000 1	.000 2	.000 4	.000 8	.001 4	.002 4
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 3
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
15	0	.860 1	.738 6	.633 3	.542 1	.463 3	.395 3	.336 7	.286 3	.243 0
	1	.130 3	.226 1	.293 8	.338 8	.365 8	.378 5	.380 1	.373 4	.360 5
	2	.009 2	.032 3	.063 6	.098 8	.134 8	.169 1	.200 3	.227 3	.249 6
	3	.000 4	.002 9	.008 5	.017 8	.030 7	.046 8	.065 3	.085 7	.107 0
	4	.000 0	.000 2	.000 8	.002 2	.004 9	.009 0	.014 8	.022 3	.031 7
	5	.000 0	.000 0	.000 1	.000 2	.000 6	.001 3	.002 4	.004 3	.006 9
	6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 6	.001 1
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	13	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0

（续表）

		p								
n	x	.01	.02	.03	.04	.05	.06	.07	.08	.09
18	0	.834 5	.695 1	.578 0	.479 6	.397 2	.328 3	.270 8	.222 9	.183 1
	1	.151 7	.255 4	.321 7	.359 7	.376 3	.377 2	.366 9	.348 9	.326 0
	2	.013 0	.044 3	.084 6	.127 4	.168 3	.204 7	.234 8	.257 9	.274 1
	3	.000 7	.004 8	.014 0	.028 3	.047 3	.069 7	.094 2	.119 6	.144 6
	4	.000 0	.000 4	.001 6	.004 4	.009 3	.016 7	.026 6	.039 0	.053 6
	5	.000 0	.000 0	.000 1	.000 5	.001 4	.003 0	.005 6	.009 5	.014 8
	6	.000 0	.000 0	.000 0	.000 0	.000 2	.000 4	.000 9	.001 8	.003 2
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 5
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	13	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	16	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	17	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	18	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
20	0	.817 9	.667 6	.543 8	.442 0	.358 5	.290 1	.234 2	.188 7	.151 6
	1	.165 2	.272 5	.336 4	.368 3	.377 4	.370 3	.352 6	.328 2	.300 0
	2	.015 9	.052 8	.098 8	.145 8	.188 7	.224 6	.252 1	.271 1	.281 8
	3	.001 0	.006 5	.018 3	.036 4	.059 6	.086 0	.113 9	.141 4	.167 2
	4	.000 0	.000 6	.002 4	.006 5	.013 3	.023 3	.036 4	.052 3	.070 3
	5	.000 0	.000 0	.000 2	.000 9	.002 2	.004 8	.008 8	.014 5	.022 2
	6	.000 0	.000 0	.000 0	.000 1	.000 3	.000 8	.001 7	.003 2	.005 5
	7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 5	.001 1
	8	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	13	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
15	14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	16	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	17	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	18	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	19	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	20	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0

（续表）

		p								
n	x	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.810 0	.722 5	.640 0	.562 5	.490 0	.422 5	.360 0	.302 5	.250 0
	1	.180 0	.255 0	.320 0	.375 0	.420 0	.455 0	.480 0	.495 0	.500 0
	2	.010 0	.022 5	.040 0	.062 5	.090 0	.122 5	.160 0	.202 5	.250 0
3	0	.729 0	.614 1	.512 0	.421 9	.343 0	.274 6	.216 0	.166 4	.125 0
	1	.243 0	.325 1	.384 0	.421 9	.441 0	.443 6	.432 0	.408 4	.375 0
	2	.027 0	.057 4	.096 0	.140 6	.189 0	.238 9	.288 0	.334 1	.375 0
	3	.001 0	.003 4	.008 0	.015 6	.027 0	.042 9	.064 0	.091 1	.125 0
4	0	.656 1	.522 0	.409 6	.316 4	.240 1	.178 5	.129 6	.091 5	.062 5
	1	.291 6	.368 5	.409 6	.421 9	.411 6	.384 5	.345 6	.299 5	.250 0
	2	.048 6	.097 5	.153 6	.210 9	.264 6	.310 5	.345 6	.367 5	.375 0
	3	.003 6	.011 5	.025 6	.046 9	.075 6	.111 5	.153 6	.200 5	.250 0
	4	.000 1	.000 5	.001 6	.003 9	.008 1	.015 0	.025 6	.041 0	.062 5
5	0	.590 5	.443 7	.327 7	.237 3	.168 1	.116 0	.077 8	.050 3	.031 2
	1	.328 0	.391 5	.409 6	.395 5	.360 2	.312 4	.259 2	.205 9	.156 2
	2	.072 9	.138 2	.204 8	.263 7	.308 7	.336 4	.345 6	.336 9	.312 5
	3	.008 1	.024 4	.051 2	.087 9	.132 3	.181 1	.230 4	.275 7	.312 5
	4	.000 4	.002 2	.006 4	.014 6	.028 4	.048 8	.076 8	.112 8	.156 2
	5	.000 0	.000 1	.000 3	.001 0	.002 4	.005 3	.010 2	.018 5	.031 2
6	0	.531 4	.377 1	.262 1	.178 0	.117 6	.075 4	.046 7	.027 7	.015 6
	1	.354 3	.399 3	.393 2	.356 0	.302 5	.243 7	.186 6	.135 9	.093 8
	2	.098 4	.176 2	.245 8	.296 6	.324 1	.328 0	.311 0	.278 0	.234 4
	3	.014 6	.041 5	.081 9	.131 8	.185 2	.235 5	.276 5	.303 2	.312 5
	4	.001 2	.005 5	.015 4	.033 0	.059 5	.095 1	.138 2	.186 1	.234 4
	5	.000 1	.000 4	.001 5	.004 4	.010 2	.020 5	.036 9	.060 9	.093 8
	6	.000 0	.000 0	.000 1	.000 2	.000 7	.001 8	.004 1	.008 3	.015 6
7	0	.478 3	.320 6	.209 7	.133 5	.082 4	.049 0	.028 0	.015 2	.007 8
	1	.372 0	.396 0	.367 0	.311 5	.247 1	.184 8	.130 6	.087 2	.054 7
	2	.124 0	.209 7	.275 3	.311 5	.317 7	.298 5	.261 3	.214 0	.164 1
	3	.023 0	.061 7	.114 7	.173 0	.226 9	.267 9	.290 3	.291 8	.273 4
	4	.002 6	.010 9	.028 7	.057 7	.097 2	.144 2	.193 5	.238 8	.273 4
	5	.000 2	.001 2	.004 3	.011 5	.025 0	.046 6	.077 4	.117 2	.164 1
	6	.000 0	.000 1	.000 4	.001 3	.003 6	.008 4	.017 2	.032 0	.054 7
	7	.000 0	.000 0	.000 0	.000 1	.000 2	.000 6	.001 6	.003 7	.007 8
8	0	.430 5	.272 5	.167 8	.100 1	.057 6	.031 9	.016 8	.008 4	.003 9
	1	.382 6	.384 7	.335 5	.267 0	.197 7	.137 3	.089 6	.054 8	.031 2
	2	.148 8	.237 6	.293 6	.311 5	.296 5	.258 7	.209 0	.156 9	.109 4
	3	.033 1	.083 9	.146 8	.207 6	.254 1	.278 6	.278 7	.256 8	.218 8
	4	.004 6	.018 5	.045 9	.086 5	.136 1	.187 5	.232 2	.262 7	.273 4
	5	.000 4	.002 6	.009 2	.023 1	.046 7	.080 8	.123 9	.171 9	.218 8
	6	.000 0	.000 2	.001 1	.003 8	.010 0	.021 7	.041 3	.070 3	.109 4
	7	.000 0	.000 0	.000 1	.000 4	.001 2	.003 3	.007 9	.016 4	.031 2
	8	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 7	.001 7	.003 9

（续表）

		p								
n	x	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.387 4	.231 6	.134 2	.075 1	.040 4	.020 7	.010 1	.004 6	.002 0
	1	.387 4	.367 9	.302 0	.225 3	.155 6	.100 4	.060 5	.033 9	.017 6
	2	.172 2	.259 7	.302 0	.300 3	.266 8	.216 2	.161 2	.111 0	.070 3
	3	.044 6	.106 9	.176 2	.233 6	.266 8	.271 6	.250 8	.211 9	.164 1
	4	.007 4	.028 3	.066 1	.116 8	.171 5	.219 4	.250 8	.260 0	.246 1
	5	.000 8	.005 0	.016 5	.038 9	.073 5	.118 1	.167 2	.212 8	.246 1
	6	.000 1	.000 6	.002 8	.008 7	.021 0	.042 4	.074 3	.116 0	.164 1
	7	.000 0	.000 0	.000 3	.001 2	.003 9	.009 8	.021 2	.040 7	.070 3
	8	.000 0	.000 0	.000 0	.000 1	.000 4	.001 3	.003 5	.008 3	.017 6
	9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 8	.002 0
10	0	.348 7	.196 9	.107 4	.056 3	.028 2	.013 5	.006 0	.002 5	.001 0
	1	.387 4	.347 4	.268 4	.187 7	.121 1	.072 5	.040 3	.020 7	.009 8
	2	.193 7	.275 9	.302 0	.281 6	.233 5	.175 7	.120 9	.076 3	.043 9
	3	.057 4	.129 8	.201 3	.250 3	.266 8	.252 2	.215 0	.166 5	.117 2
	4	.011 2	.040 1	.088 1	.146 0	.200 1	.237 7	.250 8	.238 4	.205 1
	5	.001 5	.008 5	.026 4	.058 4	.102 9	.153 6	.200 7	.234 0	.246 1
	6	.000 1	.001 2	.005 5	.016 2	.036 8	.068 9	.111 5	.159 6	.205 1
	7	.000 0	.000 1	.000 8	.003 1	.009 0	.021 2	.042 5	.074 6	.117 2
	8	.000 0	.000 0	.000 1	.000 4	.001 4	.004 3	.010 6	.022 9	.043 9
	9	.000 0	.000 0	.000 0	.000 0	.000 1	.000 5	.001 6	.004 2	.009 8
	10	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.001 0
12	0	.282 4	.142 2	.068 7	.031 7	.013 8	.005 7	.002 2	.000 8	.000 2
	1	.376 6	.301 2	.206 2	.126 7	.071 2	.036 8	.017 4	.007 5	.002 9
	2	.230 1	.292 4	.283 5	.232 3	.167 8	.108 8	.063 9	.033 9	.016 1
	3	.085 3	.172 0	.236 2	.258 1	.239 7	.195 4	.141 9	.092 3	.053 7
	4	.021 3	.068 3	.132 9	.193 6	.231 1	.236 7	.212 8	.170 0	.120 8
	5	.003 8	.019 3	.053 2	.103 2	.158 5	.203 9	.227 0	.222 5	.193 4
	6	.000 5	.004 0	.015 5	.040 1	.079 2	.128 1	.176 6	.212 4	.225 6
	7	.000 0	.000 6	.003 3	.011 5	.029 1	.059 1	.100 9	.148 9	.193 4
	8	.000 0	.000 1	.000 5	.002 4	.007 8	.019 9	.042 0	.076 2	.120 8
	9	.000 0	.000 0	.000 1	.000 4	.001 5	.004 8	.012 5	.027 7	.053 7
	10	.000 0	.000 0	.000 0	.000 0	.000 2	.000 8	.002 5	.006 8	.016 1
	11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.001 0	.002 9
	12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2
15	0	.205 9	.087 4	.035 2	.013 4	.004 7	.001 6	.000 5	.000 1	.000 0
	1	.343 2	.231 2	.131 9	.066 8	.030 5	.012 6	.004 7	.001 6	.000 5
	2	.266 9	.285 6	.230 9	.155 9	.091 6	.047 6	.021 9	.009 0	.003 2
	3	.128 5	.218 4	.250 1	.225 2	.170 0	.111 0	.063 4	.031 8	.013 9
	4	.042 8	.115 6	.187 6	.225 2	.218 6	.179 2	.126 8	.078 0	.041 7
	5	.010 5	.044 9	.103 2	.165 1	.206 1	.212 3	.185 9	.140 4	.091 6
	6	.001 9	.013 2	.043 0	.091 7	.147 2	.190 6	.206 6	.191 4	.152 7
	7	.000 3	.003 0	.013 8	.039 3	.081 1	.131 9	.177 1	.201 3	.196 4
	8	.000 0	.000 5	.003 5	.013 1	.034 8	.071 0	.118 1	.164 7	.196 4
	9	.000 0	.000 1	.000 7	.003 4	.001 6	.029 8	.061 2	.104 8	.152 7
	10	.000 0	.000 0	.000 1	.000 7	.003 0	.009 6	.024 5	.051 5	.091 6
	11	.000 0	.000 0	.000 0	.000 1	.000 6	.002 4	.007 4	.019 1	.041 7
	12	.000 0	.000 0	.000 0	.000 0	.000 1	.000 4	.001 6	.005 2	.013 9
	13	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.001 0	.003 2
	14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 5
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0

（续表）

		p								
n	x	.10	.15	.20	.25	.30	.35	.40	.45	.50
18	0	.150 1	.053 6	.018 0	.005 6	.001 6	.000 4	.000 1	.000 0	.000 0
	1	.300 2	.170 4	.081 1	.033 8	.012 6	.004 2	.001 2	.000 3	.000 1
	2	.283 5	.255 6	.172 3	.095 8	.045 8	.019 0	.006 9	.002 2	.000 6
	3	.168 0	.240 6	.229 7	.170 4	.104 6	.054 7	.024 6	.009 5	.003 1
	4	.070 0	.159 2	.215 3	.213 0	.168 1	.110 4	.061 4	.029 1	.011 7
	5	.021 8	.078 7	.150 7	.198 8	.201 7	.166 4	.114 6	.066 6	.032 7
	6	.005 2	.030 1	.081 6	.143 6	.187 3	.194 1	.165 5	.118 1	.070 8
	7	.001 0	.009 1	.035 0	.082 0	.137 6	.179 2	.189 2	.165 7	.121 4
	8	.000 2	.002 2	.012 0	.037 6	.081 1	.132 7	.173 4	.186 4	.166 9
	9	.000 0	.000 4	.003 3	.013 9	.038 6	.079 4	.128 4	.169 4	.185 5
	10	.000 0	.000 1	.000 8	.004 2	.014 9	.038 5	.077 1	.124 8	.166 9
	11	.000 0	.000 0	.000 1	.001 0	.004 6	.015 1	.037 4	.074 2	.121 4
	12	.000 0	.000 0	.000 0	.000 2	.001 2	.004 7	.014 5	.035 4	.070 8
	13	.000 0	.000 0	.000 0	.000 0	.000 2	.001 2	.004 5	.013 4	.032 7
	14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 2	.001 1	.003 9	.011 7
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 2	.000 9	.003 1
	16	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 6
	17	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
	18	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
20	0	.121 6	.038 8	.011 5	.003 2	.000 8	.000 2	.000 0	.000 0	.000 0
	1	.270 2	.136 8	.057 6	.021 1	.006 8	.002 0	.000 5	.000 1	.000 0
	2	.285 2	.229 3	.136 9	.066 9	.027 8	.010 0	.003 1	.000 8	.000 2
	3	.190 1	.242 8	.205 4	.133 9	.071 6	.032 3	.012 3	.004 0	.001 1
	4	.089 8	.182 1	.218 2	.189 7	.130 4	.073 8	.035 0	.013 9	.004 6
	5	.031 9	.102 8	.174 6	.202 3	.178 9	.127 2	.074 6	.036 5	.014 8
	6	.008 9	.045 4	.109 1	.168 6	.191 6	.171 2	.124 4	.074 6	.037 0
	7	.002 0	.016 0	.054 5	.112 4	.164 3	.184 4	.165 9	.122 1	.073 9
	8	.000 4	.004 6	.022 2	.060 9	.114 4	.161 4	.179 7	.162 3	.120 1
	9	.000 1	.001 1	.007 4	.027 1	.065 4	.115 8	.159 7	.177 1	.160 2
	10	.000 0	.000 2	.002 0	.009 9	.030 8	.068 6	.117 1	.159 3	.176 2
	11	.000 0	.000 0	.000 5	.003 0	.012 0	.033 6	.071 0	.118 5	.160 2
	12	.000 0	.000 0	.000 1	.000 8	.003 9	.013 6	.035 5	.072 7	.120 1
	13	.000 0	.000 0	.000 0	.000 2	.001 0	.004 5	.014 6	.036 6	.073 9
	14	.000 0	.000 0	.000 0	.000 0	.000 2	.001 2	.004 9	.015 0	.037 0
	15	.000 0	.000 0	.000 0	.000 0	.000 0	.000 3	.001 3	.004 9	.014 8
	16	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 3	.001 3	.004 6
	17	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 2	.001 1
	18	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 2
	19	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
	20	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0

表格 6　$e^{-\mu}$ 值

μ	$e^{-\mu}$	μ	$e^{-\mu}$	μ	$e^{-\mu}$
.00	1.000 0				
.05	.951 2	2.05	.128 7	4.05	.017 4
.10	.904 8	2.10	.122 5	4.10	.016 6
.15	.860 7	2.15	.116 5	4.15	.015 8
.20	.818 7	2.20	.110 8	4.20	.015 0
.25	.778 8	2.25	.105 4	4.25	.014 3
.30	.740 8	2.30	.100 3	4.30	.013 6
.35	.704 7	2.35	.095 4	4.35	.012 9
.40	.670 3	2.40	.090 7	4.40	.012 3
.45	.637 6	2.45	.086 3	4.45	.011 7
.50	.606 5	2.50	.082 1	4.50	.011 1
.55	.576 9	2.55	.078 1	4.55	.010 6
.60	.548 8	2.60	.074 3	4.60	.010 1
.65	.522 0	2.65	.070 7	4.65	.009 6
.70	.496 6	2.70	.067 2	4.70	.009 1
.75	.472 4	2.75	.063 9	4.75	.008 7
.80	.449 3	2.80	.060 8	4.80	.008 2
.85	.427 4	2.85	.057 8	4.85	.007 8
.90	.406 6	2.90	.055 0	4.90	.007 4
.95	.386 7	2.95	.052 3	4.95	.007 1
1.00	.367 9	3.00	.049 8	5.00	.006 7
1.05	.349 9	3.05	.047 4	6.00	.002 5
1.10	.332 9	3.10	.045 0	7.00	.000 9
1.15	.316 6	3.15	.042 9	8.00	.000 335
1.20	.301 2	3.20	.040 8	9.00	.000 123
				10.00	.000 045
1.25	.286 5	3.25	.038 8		
1.30	.272 5	3.30	.036 9		
1.35	.259 2	3.35	.0351		
1.40	.246 6	3.40	.033 4		
1.45	.234 6	3.45	.031 7		
1.50	.223 1	3.50	.030 2		
1.55	.212 2	3.55	.028 7		
1.60	.201 9	3.60	.027 3		
1.65	.192 0	3.65	.026 0		
1.70	.182 7	3.70	.024 7		
1.75	.173 8	3.75	.023 5		
1.80	.165 3	3.80	.022 4		
1.85	.157 2	3.85	.021 3		
1.90	.149 6	3.90	.020 2		
1.95	.142 3	3.95	.019 3		
2.00	.135 3	4.00	.018 3		

表格 7 泊松概率

表中数字给出具有均值为 μ 的泊松分布中发生 x 次的概率。例如，当 $\mu=2.5$ 时，发生 4 次的概率是 0.1336。

	μ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.904 8	.818 7	.740 8	.670 3	.606 5	.548 8	.496 6	.449 3	.406 6	.367 9
1	.090 5	.163 7	.222 2	.268 1	.303 3	.329 3	.347 6	.359 5	.365 9	.367 9
2	.004 5	.016 4	.033 3	.053 6	.075 8	.098 8	.121 7	.143 8	.164 7	.183 9
3	.000 2	.001 1	.003 3	.007 2	.012 6	.019 8	.028 4	.038 3	.049 4	.061 3
4	.000 0	.000 1	.000 2	.000 7	.001 6	.003 0	.005 0	.007 7	.011 1	.015 3
5	.000 0	.000 0	.000 0	.000 1	.000 2	.000 4	.000 7	.001 2	.002 0	.003 1
6	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 3	.000 5
7	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1

	μ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.332 9	.301 2	.272 5	.246 6	.223 1	.201 9	.182 7	.165 3	.149 6	.135 3
1	.366 2	.361 4	.354 3	.345 2	.334 7	.323 0	.310 6	.297 5	.284 2	.270 7
2	.201 4	.216 9	.230 3	.241 7	.251 0	.258 4	.264 0	.267 8	.270 0	.270 7
3	.073 8	.086 7	.099 8	.112 8	.125 5	.137 8	.149 6	.160 7	.171 0	.180 4
4	.020 3	.026 0	.032 4	.039 5	.047 1	.055 1	.063 6	.072 3	.081 2	.090 2
5	.004 5	.006 2	.008 4	.011 1	.014 1	.017 6	.021 6	.026 0	.030 9	.036 1
6	.000 8	.001 2	.001 8	.002 6	.003 5	.004 7	.006 1	.007 8	.009 8	.012 0
7	.000 1	.000 2	.000 3	.000 5	.000 8	.001 1	.001 5	.002 0	.002 7	.003 4
8	.000 0	.000 0	.000 1	.000 1	.000 1	.000 2	.000 3	.000 5	.000 6	.000 9
9	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 2

	μ									
x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.122 5	.110 8	.100 3	.090 7	.082 1	.074 3	.067 2	.060 8	.055 0	.049 8
1	.257 2	.243 8	.230 6	.217 7	.205 2	.193 1	.181 5	.170 3	.159 6	.149 4
2	.270 0	.268 1	.265 2	.261 3	.256 5	.251 0	.245 0	.238 4	.231 4	.224 0
3	.189 0	.196 6	.203 3	.209 0	.213 8	.217 6	.220 5	.222 5	.223 7	.224 0
4	.099 2	.108 2	.116 9	.125 4	.133 6	.141 4	.148 8	.155 7	.162 2	.168 0
5	.041 7	.047 6	.053 8	.060 2	.066 8	.073 5	.080 4	.087 2	.094 0	.100 8
6	.014 6	.017 4	.020 6	.024 1	.027 8	.031 9	.036 2	.040 7	.045 5	.050 4
7	.004 4	.005 5	.006 8	.008 3	.009 9	.011 8	.013 9	.016 3	.018 8	.021 6
8	.001 1	.001 5	.001 9	.002 5	.003 1	.003 8	.004 7	.005 7	.006 8	.008 1
9	.000 3	.000 4	.000 5	.000 7	.000 9	.001 1	.001 4	.001 8	.002 2	.002 7
10	.000 1	.000 1	.000 1	.000 2	.000 2	.000 3	.000 4	.000 5	.000 6	.000 8
11	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 2	.000 2
12	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1

（续表）

	μ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.045 0	.040 8	.036 9	.034 4	.030 2	.027 3	.024 7	.022 4	.020 2	.018 3
1	.139 7	.130 4	.121 7	.113 5	.105 7	.098 4	.091 5	.085 0	.078 9	.073 3
2	.216 5	.208 7	.200 8	.192 9	.185 0	.177 1	.169 2	.161 5	.153 9	.146 5
3	.223 7	.222 6	.220 9	.218 6	.215 8	.212 5	.208 7	.204 6	.200 1	.195 4
4	.173 4	.178 1	.182 3	.185 8	.188 8	.191 2	.193 1	.194 4	.195 1	.195 4
5	.107 5	.114 0	.120 3	.126 4	.132 2	.137 7	.142 9	.147 7	.152 2	.156 3
6	.055 5	.060 8	.066 2	.071 6	.077 1	.082 6	.088 1	.093 6	.098 9	.104 2
7	.024 6	.027 8	.031 2	.034 8	.038 5	.042 5	.046 6	.050 8	.055 1	.059 5
8	.009 5	.011 1	.012 9	.014 8	.016 9	.019 1	.021 5	.024 1	.026 9	.029 8
9	.003 3	.004 0	.004 7	.005 6	.006 6	.007 6	.008 9	.010 2	.011 6	.013 2
10	.001 0	.001 3	.001 6	.001 9	.002 3	.002 8	.003 3	.003 9	.004 5	.005 3
11	.000 3	.000 4	.000 5	.000 6	.000 7	.000 9	.001 1	.001 3	.001 6	.001 9
12	.000 1	.000 1	.000 1	.000 2	.000 2	.000 3	.000 3	.000 4	.000 5	.000 6
13	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2
14	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1

	μ									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.016 6	.015 0	.013 6	.012 3	.011 1	.010 1	.009 1	.008 2	.007 4	.006 7
1	.067 9	.063 0	.058 3	.054 0	.050 0	.046 2	.042 7	.039 5	.036 5	.033 7
2	.139 3	.132 3	.125 4	.118 8	.112 5	.106 3	.100 5	.094 8	.089 4	.084 2
3	.190 4	.185 2	.179 8	.174 3	.168 7	.163 1	.157 4	.151 7	.146 0	.140 4
4	.195 1	.194 4	.193 3	.191 7	.189 8	.187 5	.184 9	.182 0	.178 9	.175 5
5	.160 0	.163 3	.166 2	.168 7	.170 8	.172 5	.173 8	.174 7	.175 3	.175 5
6	.109 3	.114 3	.119 1	.123 7	.128 1	.132 3	.136 2	.139 8	.143 2	.164 2
7	.064 0	.068 6	.073 2	.077 8	.082 4	.086 9	.091 4	.095 9	.100 2	.104 4
8	.032 8	.036 0	.039 3	.042 8	.046 3	.050 0	.053 7	.057 5	.061 4	.065 3
9	.015 0	.016 8	.018 8	.020 9	.023 2	.025 5	.028 0	.030 7	.033 4	.036 3
10	.006 1	.007 1	.008 1	.009 2	.010 4	.011 8	.013 2	.014 7	.016 4	.018 1
11	.002 3	.002 7	.003 2	.003 7	.004 3	.004 9	.005 6	.006 4	.007 3	.008 2
12	.000 8	.000 9	.001 1	.001 4	.001 6	.001 9	.002 2	.002 6	.003 0	.003 4
13	.000 2	.000 3	.000 4	.000 5	.000 6	.000 7	.000 8	.000 9	.001 1	.001 3
14	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2	.000 3	.000 3	.000 4	.000 5
15	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2

	μ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.006 1	.005 5	.005 0	.004 5	.004 1	.003 7	.003 3	.003 0	.002 7	.002 5
1	.031 1	.028 7	.026 5	.024 4	.022 5	.020 7	.019 1	.017 6	.016 2	.014 9
2	.079 3	.074 6	.070 1	.065 9	.061 8	.058 0	.054 4	.050 9	.047 7	.044 6
3	.134 8	.129 3	.123 9	.118 5	.113 3	.108 2	.103 3	.098 5	.093 8	.089 2
4	.171 9	.168 1	.164 1	.160 0	.155 8	.151 5	.147 2	.142 8	.138 3	.133 9

（续表）

	μ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
5	.175 3	.174 8	.174 0	.172 8	.171 4	.169 7	.167 8	.165 6	.163 2	.160 6
6	.149 0	.151 5	.153 7	.155 5	.157 1	.158 7	.159 4	.160 1	.160 5	.160 6
7	.108 6	.112 5	.116 3	.120 0	.123 4	.126 7	.129 8	.132 6	.135 3	.137 7
8	.069 2	.073 1	.077 1	.081 0	.084 9	.088 7	.092 5	.096 2	.099 8	.103 3
9	.039 2	.042 3	.045 4	.048 6	.051 9	.055 2	.058 6	.062 0	.065 4	.068 8
10	.020 0	.022 0	.024 1	.026 2	.028 5	.030 9	.033 4	.035 9	.038 6	.041 3
11	.009 3	.010 4	.011 6	.012 9	.014 3	.015 7	.017 3	.019 0	.020 7	.022 5
12	.003 9	.004 5	.005 1	.005 8	.006 5	.007 3	.008 2	.009 2	.010 2	.011 3
13	.001 5	.001 8	.002 1	.002 4	.002 8	.003 2	.003 6	.004 1	.004 6	.005 2
14	.000 6	.000 7	.000 8	.000 9	.001 1	.001 3	.001 5	.001 7	.001 9	.002 2
15	.000 2	.000 2	.000 3	.000 3	.000 4	.000 5	.000 6	.000 7	.000 8	.000 9
16	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2	.000 2	.000 3	.000 3
17	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1

	μ									
x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.002 2	.002 0	.001 8	.001 7	.001 5	.001 4	.001 2	.001 1	.001 0	.000 9
1	.013 7	.012 6	.011 6	.010 6	.009 8	.009 0	.008 2	.007 6	.007 0	.006 4
2	.041 7	.039 0	.036 4	.034 0	.031 8	.029 6	.027 6	.025 8	.024 0	.022 3
3	.084 8	.080 6	.076 5	.072 6	.068 8	.065 2	.061 7	.058 4	.055 2	.052 1
4	.129 4	.124 9	.120 5	.116 2	.111 8	.107 6	.103 4	.099 2	.095 2	.091 2
5	.157 9	.154 9	.151 9	.148 7	.145 4	.142 0	.138 5	.134 9	.131 4	.127 7
6	.160 5	.160 1	.159 5	.158 6	.157 5	.156 2	.154 6	.152 9	.151 1	.149 0
7	.139 9	.141 8	.143 5	.145 0	.146 2	.147 2	.148 0	.148 6	.148 9	.149 0
8	.106 6	.109 9	.113 0	.116 0	.118 8	.121 5	.124 0	.126 3	.128 4	.130 4
9	.072 3	.075 7	.079 1	.082 5	.085 8	.089 1	.092 3	.095 4	.098 5	.101 4
10	.044 1	.046 9	.049 8	.052 8	.055 8	.058 8	.061 8	.064 9	.067 9	.071 0
11	.024 5	.026 5	.028 5	.030 7	.033 0	.035 3	.037 7	.040 1	.042 6	.045 2
12	.012 4	.013 7	.015 0	.016 4	.017 9	.019 4	.021 0	.022 7	.024 5	.026 4
13	.005 8	.006 5	.007 3	.008 1	.008 9	.009 8	.010 8	.011 9	.013 0	.014 2
14	.002 5	.002 9	.003 3	.003 7	.004 1	.004 6	.005 2	.005 8	.006 4	.007 1
15	.001 0	.001 2	.001 4	.001 6	.001 8	.002 0	.002 3	.002 6	.002 9	.003 3
16	.000 4	.000 5	.000 5	.000 6	.000 7	.000 8	.001 0	.001 1	.001 3	.001 4
17	.000 1	.000 2	.000 2	.000 2	.000 3	.000 3	.000 4	.000 4	.000 5	.000 6
18	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2	.000 2
19	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1

	μ									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.000 8	.000 7	.000 7	.000 6	.000 6	.000 5	.000 5	.000 4	.000 4	.000 3
1	.005 9	.005 4	.004 9	.004 5	.004 1	.003 8	.003 5	.003 2	.002 9	.002 7
2	.020 8	.019 4	.018 0	.016 7	.015 6	.014 5	.013 4	.012 5	.011 6	.010 7
3	.049 2	.046 4	.043 8	.041 3	.038 9	.036 6	.034 5	.032 4	.030 5	.028 6

（续表）

	μ									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
4	.087 4	.083 6	.079 9	.076 4	.072 9	.069 6	.066 3	.063 2	.060 2	.057 3
5	.124 1	.120 4	.116 7	.113 0	.109 4	.105 7	.102 1	.098 6	.095 1	.091 6
6	.146 8	.144 5	.142 0	.139 4	.136 7	.133 9	.131 1	.128 2	.125 2	.122 1
7	.148 9	.148 6	.148 1	.147 4	.146 5	.145 4	.144 2	.142 8	.141 3	.139 6
8	.132 1	.133 7	.135 1	.136 3	.137 3	.138 2	.138 8	.139 2	.139 5	.139 6
9	.104 2	.107 0	.109 6	.112 1	.114 4	.116 7	.118 7	.120 7	.122 4	.124 1
10	.074 0	.077 0	.080 0	.082 9	.085 8	.088 7	.091 4	.094 1	.096 7	.099 3
11	.047 8	.050 4	.053 1	.055 8	.058 5	.061 3	.064 0	.066 7	.069 5	.072 2
12	.028 3	.030 3	.032 3	.034 4	.036 6	.038 8	.041 1	.043 4	.045 7	.048 1
13	.015 4	.016 8	.018 1	.019 6	.021 1	.022 7	.024 3	.026 0	.027 8	.029 6
14	.007 8	.008 6	.009 5	.010 4	.011 3	.012 3	.013 4	.014 5	.015 7	.016 9
15	.003 7	.004 1	.004 6	.005 1	.005 7	.006 2	.006 9	.007 5	.008 3	.009 0
16	.001 6	.001 9	.002 1	.002 4	.002 6	.003 0	.003 3	.003 7	.004 1	.004 5
17	.000 7	.000 8	.000 9	.001 0	.001 2	.001 3	.001 5	.001 7	.001 9	.002 1
18	.000 3	.000 3	.000 4	.000 4	.000 5	.000 6	.000 6	.000 7	.000 8	.000 9
19	.000 1	.000 1	.000 1	.000 2	.000 2	.000 2	.000 3	.000 3	.000 3	.000 4
20	.000 0	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2
21	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1

	μ									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.000 3	.000 3	.000 2	.000 2	.000 2	.000 2	.000 2	.000 2	.000 1	.000 1
1	.002 5	.002 3	.002 1	.001 9	.001 7	.001 6	.001 4	.001 3	.001 2	.001 1
2	.010 0	.009 2	.008 6	.007 9	.007 4	.006 8	.006 3	.005 8	.005 4	.005 0
3	.026 9	.025 2	.023 7	.022 2	.020 8	.019 5	.018 3	.017 1	.016 0	.015 0
4	.054 4	.051 7	.049 1	.046 6	.044 3	.042 0	.039 8	.037 7	.035 7	.033 7
5	.088 2	.084 9	.081 6	.078 4	.075 2	.072 2	.069 2	.066 3	.063 5	.060 7
6	.119 1	.116 0	.112 8	.109 7	.106 6	.103 4	.100 3	.097 2	.094 1	.091 1
7	.137 8	.135 8	.133 8	.131 7	.129 4	.127 1	.124 7	.122 2	.119 7	.117 1
8	.139 5	.139 2	.138 8	.138 2	.137 5	.136 6	.135 6	.134 4	.133 2	.131 8
9	.125 6	.126 9	.128 0	.129 0	.129 9	.130 6	.131 1	.131 5	.131 7	.131 8
10	.101 7	.104 0	.106 3	.108 4	.110 4	.112 3	.114 0	.115 7	.117 2	.118 6
11	.074 9	.077 6	.080 2	.082 8	.085 3	.087 8	.090 2	.092 5	.094 8	.097 0
12	.050 5	.053 0	.055 5	.057 9	.060 4	.062 9	.065 4	.067 9	.070 3	.072 8
13	.031 5	.033 4	.035 4	.037 4	.039 5	.041 6	.043 8	.045 9	.048 1	.050 4
14	.018 2	.019 6	.021 0	.022 5	.024 0	.025 6	.027 2	.028 9	.030 6	.032 4
15	.009 8	.010 7	.011 6	.012 6	.013 6	.014 7	.015 8	.016 9	.018 2	.109 4
16	.005 0	.005 5	.006 0	.006 6	.007 2	.007 9	.008 6	.009 3	.010 1	.010 9
17	.002 4	.002 6	.002 9	.003 3	.003 6	.004 0	.004 4	.004 8	.005 3	.005 8
18	.001 1	.001 2	.001 4	.001 5	.001 7	.001 9	.002 1	.002 4	.002 6	.002 9
19	.000 5	.000 5	.000 6	.000 7	.000 8	.000 9	.001 0	.001 1	.001 2	.001 4

（续表）

	μ									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
20	.000 2	.000 2	.000 2	.000 3	.000 3	.000 4	.000 4	.000 5	.000 5	.000 6
21	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2	.000 2	.000 2	.000 3
22	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1

	μ									
x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 0
1	.001 0	.000 9	.000 9	.000 8	.000 7	.000 7	.000 6	.000 5	.000 5	.000 5
2	.004 6	.004 3	.004 0	.003 7	.003 4	.003 1	.002 9	.002 7	.002 5	.002 3
3	.014 0	.013 1	.012 3	.011 5	.010 7	.010 0	.009 3	.008 7	.008 1	.007 6
4	.031 9	.030 2	.028 5	.026 9	.025 4	.024 0	.022 6	.021 3	.020 1	.018 9
5	.058 1	.055 5	.053 0	.050 6	.048 3	.046 0	.043 9	.041 8	.039 8	.037 8
6	.088 1	.085 1	.082 2	.079 3	.076 4	.073 6	.070 9	.068 2	.065 6	.063 1
7	.114 5	.111 8	.109 1	.106 4	.103 7	.101 0	.098 2	.095 5	.092 8	.090 1
8	.130 2	.128 6	.126 9	.125 1	.123 2	.121 2	.119 1	.117 0	.114 8	.112 6
9	.131 7	.131 5	.131 1	.130 6	.130 0	.129 3	.128 4	.127 4	.126 3	.125 1
10	.119 8	.121 0	.121 9	.122 8	.123 5	.124 1	.124 5	.124 9	.125 0	.125 1
11	.099 1	.101 2	.103 1	.104 9	.106 7	.108 3	.109 8	.111 2	.112 5	.113 7
12	.075 2	.077 6	.079 9	.082 2	.084 4	.086 6	.088 8	.090 8	.092 8	.094 8
13	.052 6	.054 9	.057 2	.059 4	.061 7	.064 0	.066 2	.068 5	.070 7	.072 9
14	.034 2	.036 1	.038 0	.039 9	.041 9	.043 9	.045 9	.047 9	.050 0	.052 1
15	.020 8	.022 1	.023 5	.025 0	.026 5	.028 1	.029 7	.031 3	.033 0	.034 7
16	.011 8	.012 7	.013 7	.014 7	.015 7	.016 8	.018 0	.019 2	.020 4	.021 7
17	.006 3	.006 9	.007 5	.008 1	.008 8	.009 5	.010 3	.011 1	.011 9	.012 8
18	.003 2	.003 5	.003 9	.004 2	.004 6	.005 1	.005 5	.006 0	.006 5	.007 1
19	.001 5	.001 7	.001 9	.002 1	.002 3	.002 6	.002 8	.003 1	.003 4	.003 7
20	.000 7	.000 8	.000 9	.001 0	.001 1	.001 2	.001 4	.001 5	.001 7	.001 9
21	.000 3	.000 3	.000 4	.000 4	.000 5	.000 6	.000 6	.000 7	.000 8	.000 9
22	.000 1	.000 1	.000 2	.000 2	.000 2	.000 2	.000 3	.000 3	.000 4	.000 4
23	.000 0	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 1	.000 2	.000 2
24	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 1

	μ									
x	11	12	13	14	15	16	17	18	19	20
0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
1	.000 2	.000 1	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
2	.001 0	.000 4	.000 2	.000 1	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0
3	.003 7	.001 8	.000 8	.000 4	.000 2	.000 1	.000 0	.000 0	.000 0	.000 0
4	.010 2	.005 3	.002 7	.001 3	.000 6	.000 3	.000 1	.000 1	.000 0	.000 0
5	.022 4	.012 7	.007 0	.003 7	.001 9	.001 0	.000 5	.000 2	.000 1	.000 1
6	.041 1	.025 5	.015 2	.008 7	.004 8	.002 6	.001 4	.000 7	.000 4	.000 2
7	.064 6	.043 7	.028 1	.017 4	.010 4	.006 0	.003 4	.001 8	.001 0	.000 5

（续表）

	μ									
x	11	12	13	14	15	16	17	18	19	20
7	.064 6	.043 7	.028 1	.017 4	.010 4	.006 0	.003 4	.001 8	.001 0	.000 5
8	.088 8	.065 5	.045 7	.030 4	.019 4	.012 0	.007 2	.004 2	.002 4	.001 3
9	.108 5	.087 4	.066 1	.047 3	.032 4	.021 3	.013 5	.008 3	.005 0	.002 9
10	.119 4	.104 8	.085 9	.066 3	.048 6	.034 1	.023 0	.015 0	.009 5	.005 8
11	.119 4	.114 4	.101 5	.084 4	.066 3	.049 6	.035 5	.024 5	.016 4	.010 6
12	.109 4	.114 4	.109 9	.098 4	.082 9	.066 1	.050 4	.036 8	.025 9	.017 6
13	.092 6	.105 6	.109 9	.106 0	.095 6	.081 4	.065 8	.050 9	.037 8	.027 1
14	.072 8	.090 5	.102 1	.106 0	.102 4	.093 0	.080 0	.065 5	.051 4	.038 7
15	.053 4	.072 4	.088 5	.098 9	.102 4	.099 2	.090 6	.078 6	.065 0	.051 6
16	.036 7	.054 3	.071 9	.086 6	.096 0	.099 2	.096 3	.088 4	.077 2	.064 6
17	.023 7	.038 3	.055 0	.071 3	.084 7	.093 4	.096 3	.093 6	.086 3	.076 0
18	.014 5	.025 6	.039 7	.055 4	.070 6	.083 0	.090 9	.093 6	.091 1	.084 4
19	.008 4	.016 1	.027 2	.040 9	.055 7	.069 9	.081 4	.088 7	.091 1	.088 8
20	.004 6	.009 7	.017 7	.028 6	.041 8	.055 9	.069 2	.079 8	.086 6	.088 8
21	.002 4	.005 5	.010 9	.019 1	.029 9	.042 6	.056 0	.068 4	.078 3	.084 6
22	.001 2	.003 0	.006 5	.012 1	.020 4	.031 0	.043 3	.056 0	.067 6	.076 9
23	.000 6	.001 6	.003 7	.007 4	.013 3	.021 6	.032 0	.043 8	.055 9	.066 9
24	.000 3	.000 8	.002 0	.004 3	.008 3	.014 4	.022 6	.032 8	.044 2	.055 7
25	.000 1	.000 4	.001 0	.002 4	.005 0	.009 2	.015 4	.023 7	.033 6	.044 6
26	.000 0	.000 2	.000 5	.001 3	.002 9	.005 7	.010 1	.016 4	.024 6	.034 3
27	.000 0	.000 1	.000 2	.000 7	.001 6	.003 4	.006 3	.010 9	.017 3	.025 4
28	.000 0	.000 0	.000 1	.000 3	.000 9	.001 9	.003 8	.007 0	.011 7	.018 1
29	.000 0	.000 0	.000 1	.000 2	.000 4	.001 1	.002 3	.004 4	.007 7	.012 5
30	.000 0	.000 0	.000 0	.000 1	.000 2	.000 6	.001 3	.002 6	.004 9	.008 3
31	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 7	.001 5	.003 0	.005 4
32	.000 0	.000 0	.000 0	.000 0	.000 1	.000 1	.000 4	.000 9	.001 8	.003 4
33	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 5	.001 0	.002 0
34	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 6	.001 2
35	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 3	.000 7
36	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2	.000 4
37	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1	.000 2
38	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1
39	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 0	.000 1

表格 8　自相关的 D－W 检验的临界值

表中数字给出自相关分析中单边 D－W 检验的临界值。对于双边检验，重要性水平应是表中数字的两倍。

$\alpha=0.05$ 时 d_L 和 d_U 的重要程度独立变量的个数										
k	1		2		3		4		5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

（续表）

$\alpha=0.025$ 时 d_L 和 d_U 的重要程度独立变量的个数										
k	1		2		3		4		5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.95	1.23	0.83	1.40	0.71	1.61	0.59	1.84	0.48	2.09
16	0.98	1.24	0.86	1.40	0.75	1.59	0.64	1.80	0.53	2.03
17	1.01	1.25	0.90	1.40	0.79	1.58	0.68	1.77	0.57	1.98
18	1.03	1.26	0.93	1.40	0.82	1.56	0.72	1.74	0.62	1.93
19	1.06	2.28	0.96	1.41	0.86	1.55	0.76	1.72	0.66	1.90
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.70	0.70	1.87
21	1.10	1.30	1.01	1.41	0.92	1.54	0.83	1.69	0.73	1.84
22	1.12	1.31	1.04	1.42	0.95	1.54	0.86	1.68	0.77	1.82
23	1.14	1.32	1.06	1.42	0.97	1.54	0.89	1.67	0.80	1.80
24	1.16	1.33	1.08	1.43	1.00	1.54	0.91	1.66	0.83	1.79
25	1.18	1.34	1.10	1.43	1.02	1.54	0.94	1.65	0.86	1.77
26	1.19	1.35	1.12	1.44	1.04	1.54	0.96	1.65	0.88	1.76
27	1.21	1.36	1.13	1.44	1.06	1.54	0.99	1.64	0.91	1.75
28	1.22	1.37	1.15	1.45	1.08	1.54	1.01	1.64	0.93	1.74
29	1.24	1.38	1.17	1.45	1.10	1.54	1.03	1.63	0.96	1.73
30	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	0.98	1.73
31	1.26	1.39	1.20	1.47	1.13	1.55	1.07	1.63	1.00	1.72
32	1.27	1.40	1.21	1.47	1.15	1.55	1.08	1.63	1.02	1.71
33	1.28	1.41	1.22	1.48	1.16	1.55	1.10	1.63	1.04	1.71
34	1.29	1.41	1.24	1.48	1.17	1.55	1.12	1.63	1.06	1.70
35	1.30	1.42	1.25	1.48	1.19	1.55	1.13	1.63	1.07	1.70
36	1.31	1.43	1.26	1.49	1.20	1.56	1.15	1.63	1.09	1.70
37	1.32	1.43	1.27	1.49	1.21	1.56	1.16	1.62	1.10	1.70
38	1.33	1.44	1.28	1.50	1.23	1.56	1.17	1.62	1.12	1.70
39	1.34	1.44	1.29	1.50	1.24	1.56	1.19	1.63	1.13	1.69
40	1.35	1.45	1.30	1.51	1.25	1.57	1.20	1.63	1.15	1.69
45	1.39	1.48	1.34	1.53	1.30	1.58	1.25	1.63	1.21	1.69
50	1.42	1.50	1.38	1.54	1.34	1.59	1.30	1.64	1.26	1.69
55	1.45	1.52	1.41	1.56	1.37	1.60	1.33	1.64	1.30	1.69
60	1.47	1.54	1.44	1.57	1.40	1.61	1.37	1.65	1.33	1.69
65	1.49	1.55	1.46	1.59	1.43	1.62	1.40	1.66	1.36	1.69
70	1.51	1.57	1.48	1.60	1.45	1.63	1.42	1.66	1.39	1.70
75	1.53	1.58	1.50	1.61	1.47	1.64	1.45	1.67	1.42	1.70
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.70
85	1.56	1.60	1.53	1.63	1.51	1.65	1.49	1.68	1.46	1.71
90	1.57	1.61	1.55	1.64	1.53	1.66	1.50	1.69	1.48	1.71
95	1.58	1.62	1.56	1.65	1.54	1.67	1.52	1.69	1.50	1.71
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.70	1.51	1.72

（续表）

$\alpha = 0.01$ 时 d_L 和 d_U 的重要程度独立变量的个数

	k	1		2		3		4		5	
n		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15		0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16		0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17		0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18		0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19		0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20		0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21		0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22		1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23		1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24		1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25		1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26		1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27		1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28		1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29		1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30		1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31		1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32		1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33		1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34		1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35		1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36		1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37		1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38		1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39		1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40		1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45		1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50		1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55		1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60		1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65		1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70		1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75		1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80		1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85		1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90		1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95		1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100		1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

该表经牛津大学出版社许可重印，基于 Biometrika Trustees，见 J. Durbin 和 G. S. Watson，“最小方差回归分析 II 中的序列相关检验”，*Biometrika* 38(1951)，159 - 178。

表格9　MANN-WHITNEY-WILCOXON 检验的 T_L 值

如果项目 n_1 的秩之和比下列表中的 T_L 值小，或者项目 n_1 的秩之和比 T_U 值大，则拒绝识别总体的假设。式中，$T_U = n_1(n_1 + n_2 + 1) - T_L$。

$\alpha = 0.01$		n_2								
		2	3	4	5	6	7	8	9	10
n_1	2	3	3	3	4	4	4	5	5	5
	3	6	7	7	8	9	9	10	11	11
	4	10	11	12	13	14	15	16	17	18
	5	16	17	18	20	21	22	24	25	27
	6	22	24	25	27	29	30	32	34	36
	7	29	31	33	35	37	40	42	44	46
	8	38	40	42	45	47	50	52	55	57
	9	47	50	52	55	58	61	64	67	70
	10	57	60	63	67	70	73	76	80	83

$\alpha = 0.05$		n_2								
		2	3	4	5	6	7	8	9	10
n_1	2	3	3	3	3	3	3	4	4	4
	3	6	6	6	7	8	8	9	9	10
	4	10	10	11	12	13	14	15	15	16
	5	15	16	17	18	19	21	22	23	24
	6	21	23	24	25	27	28	30	32	33
	7	28	30	32	34	35	37	39	41	43
	8	37	39	41	43	45	47	50	52	54
	9	46	48	50	53	56	58	61	63	66
	10	56	59	61	64	67	70	73	76	79

表格 10　学生分布的临界值

	$\alpha = 0.05$																		
自由度	总体数目																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	50.6	52.0	53.2	54.3	55.4	56.3	57.2	58.0	58.8	59.6
2	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.4	14.7	15.1	15.4	15.7	15.9	16.1	16.4	16.6	16.8
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72	9.95	10.2	10.3	10.5	10.7	10.8	11.0	11.1	11.2
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03	8.21	8.37	8.52	8.66	8.79	8.91	9.03	9.13	9.23
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

（续表）

$\alpha = 0.01$																			
自由度	总体数目																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	90.0	135.0	164.0	186.0	202.0	216.0	227.0	237.0	246.0	253.0	260.0	266.0	272.0	277.0	282.0	286.0	290.0	294.0	298.0
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	32.6	33.4	34.1	34.8	35.4	36.0	36.5	37.0	37.5	37.9
3	8.26	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	17.1	17.5	17.9	18.2	18.5	18.8	19.1	19.3	19.5	19.8
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	12.6	12.8	13.1	13.3	13.5	13.7	13.9	14.1	14.2	14.4
5	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.2	10.5	10.7	10.9	11.1	11.2	11.4	11.6	11.7	11.8	11.9
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.49	9.65	9.81	9.95	10.1	10.2	10.3	10.4	10.5
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22
11	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95
12	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39
15	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26
16	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82
24	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
40	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69	5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21
60	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38	5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65

该表经牛津大学出版社许可重印，基于 Biometrika Trustees 的统计学家表，见 E. S. Pearson 和 H. O. Hartley, 1966 年第 1 卷第 3 版。

表格 11 $\bar{x}$ 因子和 R 控制图

样本观测值 n	d_2	A_2	d_3	D_3	D_4
2	1.128	1.880	0.853	0	3.267
3	1.693	1.023	0.888	0	2.574
4	2.059	0.729	0.880	0	2.282
5	2.326	0.577	0.864	0	2.114
6	2.534	0.483	0.848	0	2.004
7	2.704	0.419	0.833	0.076	1.924
8	2.847	0.373	0.820	0.136	1.864
9	2.970	0.337	0.808	0.184	1.816
10	3.078	0.308	0.797	0.223	1.777
11	3.173	0.285	0.787	0.256	1.744
12	3.258	0.266	0.778	0.283	1.717
13	3.336	0.249	0.770	0.307	1.693
14	3.407	0.235	0.763	0.328	1.672
15	3.472	0.223	0.756	0.347	1.653
16	3.532	0.212	0.750	0.363	1.637
17	3.588	0.203	0.744	0.378	1.622
18	3.640	0.194	0.739	0.391	1.608
19	3.689	0.187	0.734	0.403	1.597
20	3.735	0.180	0.729	0.415	1.585
21	3.778	0.173	0.724	0.425	1.575
22	3.819	0.167	0.720	0.434	1.566
23	3.858	0.162	0.716	0.443	1.557
24	3.895	0.157	0.712	0.451	1.548
25	3.931	0.153	0.708	0.459	1.541

节选自《关于数据和控制图使用的 ASTM 手册》(*ASTM Manual on Presentation of Data and Control Chart Analysis*)(美国检验和材料学会，1976 年版)中的 ASTM STP 15D 的表 27，经许可重印。

附录 C　求和符号

求　和

定义

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{C.1}$$

例子：如果 $x_1 = 5$, $x_2 = 8$, $x_3 = 14$：

$$\begin{aligned}\sum_{i=1}^{3} x_i &= x_1 + x_2 + x_3 \\ &= 5 + 8 + 14 \\ &= 27\end{aligned}$$

结果 1

对一个常数 c：

$$\sum_{i=1}^{n} c = (\underbrace{c + c + \cdots + c}_{n\text{倍}}) = nc \tag{C.2}$$

例子：如果 $c = 5$, $n = 10$：

$$\sum_{i=1}^{10} 5 = 10(5) = 50$$

例子：如果 $c = \bar{x}$：

$$\sum_{i=1}^{n} \bar{x} = \mathrm{n}\,\bar{x}$$

结果 2

$$\begin{aligned}\sum_{i=1}^{n} cx_i &= cx_1 + cx_2 + \cdots + cx_n \\ &= c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i\end{aligned} \tag{C.3}$$

例子：如果 $x_1 = 5$, $x_2 = 8$, $x_3 = 14$, $c = 2$:

$$\sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 2(27) = 54$$

结果 3

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i \tag{C.4}$$

例子：如果 $x_1 = 5$, $x_2 = 8$, $x_3 = 14$, $a = 2$, $y_1 = 7$, $y_2 = 3$, $y_3 = 8$, $b = 4$:

$$\begin{aligned}\sum_{i=1}^{3}(2x_i+4y_i) &= 2\sum_{i=1}^{3}x_i+4\sum_{i=1}^{3}y_i\\ &= 2(27)+4(18)\\ &= 54+72\\ &= 126\end{aligned}$$

双求和

考虑如下的包含变量 x_{ij} 的数据，式中 i 是标明行位置的下标，j 是标明列位置的下标。

		列		
		1	**2**	**3**
行	**1**	$x_{11}=10$	$x_{12}=8$	$x_{13}=6$
	2	$x_{21}=7$	$x_{22}=4$	$x_{23}=12$

定义

$$\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}=(x_{11}+x_{12}+\cdots+x_{1m})+(x_{21}+x_{22}+\cdots+x_{2m})+(x_{31}+x_{32}+\cdots+x_{3m})+\cdots+(x_{n1}+x_{n2}+\cdots+x_{nm}) \quad \textbf{(C.5)}$$

例子：

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3}x_{ij} &= x_{11}+x_{12}+x_{13}+x_{21}+x_{22}+x_{23}\\ &= 10+8+6+7+4+12\\ &= 47\end{aligned}$$

定义

$$\sum_{i=1}^{n}x_{ij}=x_{1j}+x_{2j}+\cdots+x_{nj} \quad \textbf{(C.6)}$$

例子：

$$\begin{aligned}\sum_{i=1}^{2}x_{i2} &= x_{12}+x_{22}\\ &= 8+4\\ &= 12\end{aligned}$$

缩写符号

有时是对所有的下标求和时，我们使用下列缩写符号：

$$\sum_{i=1}^{n}x_i=\sum x_i \quad \textbf{(C.7)}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}=\sum\sum x_{ij} \quad \textbf{(C.8)}$$

$$\sum_{i=1}^{n}x_{ij}=\sum_{i}x_{ij} \quad \textbf{(C.9)}$$

附录 D 偶数习题的答案

第 1 章

2. a. 9

b. 4

c. 定性型：国家和房间等级

定量型：房间的数量和总分

d. 国家是名义尺度：房间等级是序数尺度；房间数量和总分是比例尺度。

4. a. 10

b. 《财富》美国 500 强公司

c. 14 227.59 百万美元

d. 14 227.59 百万美元

6. 问题 a、c 和 d 是定量型的

问题 b 和 e 是定性型的

8. a. 2 013

b. 定性型

c. 百分比

d. 563 或 564

10. a. 定量型；比例

b. 定性型；名义

c. 定性型；序数

d. 定性型；名义

e. 定性型；比例

12. a. 所有到夏威夷的游客

b. 是

c. 问题 1 和 4 提供定量型数据

问题 2 和 3 提供定性型数据

14. a. 4

b. 全部是定量型的

c. 时间序列

16. a. 产品口味检验和市场分析

b. 专项设计的统计研究

18. a. 40%

b. 定性型的

20. a. 56% 和 387 325 美元

b. 3.73

c. 387 325 美元

22. a. 所有参加活动的成人观众

b. 电话询问的观众

c. 样本

24. a. 正确

b. 不正确

c. 正确

d. 不正确

e. 不正确

第 2 章

2. a. 0.20

b. 40

c/d.

组别	频数	百分比频数
A	44	22
B	36	18
C	80	40
D	40	20
总计	200	100

4. a. 定性型

b.

电视节目	频数	百分比频数
Millionaire	24	48
Frasier	15	30
Chicago Hope	7	14
Charmed	4	8
总计	50	100

d. Millionaire 占有最大的市场份额；Frasier 位居第二。

6. a.

图书	频数	百分比频数
7Habits	10	16.66
Millionaire	16	26.67
Motley	9	15.00
Dad	13	21.67
WSJ Guide	6	10.00
Other	6	10.00
总计	60	100.00

b. 前 5 名：*Millionaire*、*Dad*、*Motley*、*7Habits*、*WSJ Guide*

c. 48.33%

8. a.

位置	频数	百分比频数
P	17	0.309
H	4	0.073
1	5	0.091
2	4	0.073
3	2	0.036
S	5	0.091
L	6	0.109
C	5	0.091
R	7	0.127
总计	55	1.000

b. Pitcher

c. 经三场棒球

d. 右场

e. 内场成员 16 比外场成员 18

10. a. 按定性划分

b.

档次	频数	百分比频数
差	2	0.03
一般	4	0.07
好	12	0.20
非常好	24	0.40
优异	18	0.30
总计	60	1.00

12.

组别	累积频数	累积相对频数
≤19	10	0.20
≤29	24	0.48
≤39	41	0.82
≤49	48	0.96
≤59	50	1.00

14. b/c.

组别	频数	百分比频数
6.0—7.9	4	20
8.0—9.9	2	10
10.0—11.9	8	40
12.0—13.9	3	15
14.0—15.9	3	15
总计	20	100

16. a.

股价(美元)	频数	相对频数	百分比频数
10.00—19.99	10	0.40	40
20.00—29.99	4	0.16	16
30.00—39.99	6	0.24	24
40.00—49.99	2	0.08	8
50.00—59.99	1	0.04	4
60.00—69.99	2	0.08	8
总计	25	1.00	100

b.

每股收益(美元)	频数	相对频数	百分比频数
-3.00—-2.01	2	0.08	8
-2.00—-1.01	0	0.00	0
-1.00—-0.01	2	0.08	8
0.00—0.99	9	0.36	36
1.00—1.99	9	0.36	36
2.00—2.99	3	0.12	12
总计	25	1.00	100

18. a. 最低年薪：93 000 美元

最高年薪：178 000 美元

b.

年薪(千美元)	频数	相对频数	百分比频数
91—105	4	0.08	8
106—120	5	0.10	10
121—135	11	0.22	22
136—150	18	0.36	36
151—165	9	0.18	18
166—180	3	0.06	6
总计	50	1.00	100

c. 20/50

d. 24%

20. a. 48.9%

b. 43.4%

c. 51.1%

d. 9 707.5 万

e. 4 661.25 万

22.

5	7	8					
6	4	5	8				
7	0	2	2	5	5	6	8
8	0	2	3	5			

24. 叶单位 = 10

11	6		
12	0	2	
13	0	6	7
14	2	2	7
15	5		
16	0	2	8
17	0	2	3

26. 叶单位 = 0.1

0	4	7	8	9	9			
1	1	2	9					
2	0	0	1	3	5	5	6	8
3	4	9						
4	8							
5								
6								
7	1							

28. a.

0	5	8				
1	1	1	3	3	4	4
1	5	6	7	8	9	9
2	2	3	3	3	5	5
2	6	8				
3						
3	6	7	7	9		
4	0					
4	7	8				
5						
5						
6	0					

b.

2000P/E 预测	频数	相对频数
5 — 9	2	6.7
10 — 14	6	20.0
15 — 19	6	20.0
20 — 24	6	20.0
25 — 29	2	6.7
30 — 34	0	0.0
35 — 39	4	13.3
40 — 44	1	3.3
45 — 49	2	6.7
50 — 54	0	0.0
55 — 59	0	0.0
60 — 64	1	3.3
总计	30	100.0

30. b. 负相关

32. a.

销售额/ 边际收益/ 权益收益	每股收益评分 0 — 19	20 — 39	40 — 59	60 — 79	80 — 100	总计
A				1	8	9
B		1	4	5	2	12
C	1		1	2	3	7
D	3	1		1		5
E		2	1			3
总计	4	4	6	9	13	36

b.

销售额/边际收益/权益收益	0—19	20—39	40—59	60—79	80—100	总计
A				11.11	88.89	100
B		8.33	33.33	41.67	16.67	100
C	14.29		14.29	28.57	42.86	100
D	60.00	20.00		20.00		100
E		66.67	33.33			100

较高的 EPS 比例似乎是与较高的销售/边际收益/投资收益回报率相一致

34. 无明显相关

36. a.

车辆	频数	百分比频数
F-Series	17	34
Silverado	12	24
Taurus	8	16
Camry	7	14
Accord	6	12
	50	100

b. 福特 F 系列小货车和 Chevrolet Silverado

38. a.

电影	频数	百分比频数
Blair Witch Project	159	36.0
Phantom Menace	89	20.2
Beloved	85	19.3
Primary Colors	57	12.9
Truman Show	51	11.6
总计	441	100.0

c. 56.2%

40.

收盘价	频数	相对频数	累积频数	累积相对频数
0—9⅞	9	0.225	9	0.225
10—19⅞	10	0.250	19	0.475
20—29⅞	5	0.125	24	0.600
30—39⅞	11	0.275	35	0.875
40—49⅞	2	0.050	37	0.925
50—59⅞	2	0.050	39	0.975
60—69⅞	0	0.000	39	0.975
70—79⅞	1	0.025	40	1.000
总计	40	1.000		

42.

收入(美元)	频数	相对频数
18 000—21 999	13	0.255
22 000—25 999	20	0.392
26 000—29 999	12	0.235
30 000—33 999	4	0.078
34 000—37 999	2	0.039
总计	51	1.000

44. a. 高温

3	
4	
5	7
6	1 4 4 4 4 6 8
7	3 5 7 9
8	0 1 1 4 6
9	0 2 3

b. 低温

3	9
4	3 6 8
5	0 0 0 2 4 4 5 5 7 9
6	1 8
7	2 4 5 5
8	
9	

c. 低温的值域在高温的值域之下

d. 8 个城市

e.

	频数	
温度	高温	低温
30—39	0	1
40—49	0	3
50—59	1	10
60—69	7	2
70—79	4	4
80—89	5	0
90—99	3	0
总计	20	20

46. a.

	满意度分数						
职业	30—39	40—49	50—59	60—69	70—79	80—89	总计
木工			2	4	3	1	10
律师	1	5	2	1	1		10
理疗师			5	2	1	2	10
系统分析师		2	1	4	3		10
总计	1	7	10	11	8	3	40

b.

	满意度分数						
职业	30—39	40—49	50—59	60—69	70—79	80—89	总计
木工			20	40	30	10	100
律师	10	50	20	10	10		100
理疗师			50	20	10	20	100
系统分析师		20	10	40	30		100

c. 木工似乎具有最高的职业满意分数；律师似乎最低

48. b.

年份	频数	相对频数	百分比频数
1973 以前	247	Elect.	149
1974—1979	54	Nat. Gas	317
1980—1986	82	Oil	17
1987—1991	121	Propane	7
总计	504	Other	14
		总计	504

50. a. 市场价值和盈利的交叉分组表

市场价值	利润(千美元)				
(千美元)	0—300	300—600	600—900	900—1 200	总计
0—8000	23	4			27
8 000—16 000	4	4	2	2	12
16 000—24 000		2	1	1	4
24 000—32 000		1	2	1	4
32 000—40 000		2	1		3
总计	27	13	6	4	50

b. 行百分比的交叉分组表

市场价值	利润(千美元)				
(千美元)	0—300	300—600	600—900	900—1 200	总计
0—8 000	85.19	14.81	0.00	0.00	100
8 000—16 000	33.33	33.33	16.67	16.67	100
16 000—24 000	0.00	50.00	25.00	25.00	100
24 000—32 000	0.00	25.00	50.00	25.00	100
32 000—40 000	0.00	66.67	33.33	0.00	100

c. 表明在市场价值和利润之间是正相关；当盈利增加时，市场价值也随之升高

52. b. 市场价值和所有者权益之间的关系是正相关

第 3 章

2. 16，16.5

4. 59.727，57，53

6. a. 91.45，87.5，120

b. 66.3，69.5，70

c. 500@ 50 美元

d. 是

8. a. 38.75，29

b. 38.5

c. 29.5，47.5

d. 31

10. a. 48.33, 49; 不报告众数

b. 45，55

12. **城市**：均值 = 15.58，中位数 = 15.9，众数 = 15.3

乡村：均值 = 18.92，中位数 = 18.7，众数 = 18.6 和 19.4

14. a. 639 美元

b. 98.8 图片

c. 110.2 分钟

16. 16，4

18. a. 22

b. 75.2

c. 8.67

d. 4.87

20. a. 值域 = 32, 四分位点内距 = 10

b. 92.75, 9.63

22. ***Dawson:*** 值域 = 2, $s = 0.67$

Clark: 值域 = 8, $s = 2.58$

24. a. 161，92.5；56，19.5

b. 2 705.38；52.01；290.85，17.05

c. 56.38，25.56

d. 以 500@50(美元)时变异程度较大

26. 1/4 英里: $s = 0.056$，变异系数 = 5.8

英里: $s = 0.130$，变异系数 = 2.9

28. a. 95%

b. 几乎所有

c. 68%

30. 0.20，1.50，0，−0.50，−2.20

32. a. 34%

b. 81.5%

c. 16%

34. a. −0.95

b. 3.90

c. (b)部分中的劳动成本是异常值

36. a. 100；13.88 或近似为 14

b. 16%

c. 11.1 和 10.77；无异常值

38. 15，22.5，26，29，34

40. 5，8，10，15，18

42. a. 5，9.6，14.5，19.2，52.7

b. 界限：−4.8，33.6

c. 41.6 是异常值

52.7 是异常值

44. a. 105.79，52.7

b. 15.7，78.3

c. Silicon Graphics, 反斗城玩具公司(疑有误，原文如此——编者注)

d. 26.73；大得多

46. a. 37.48，23.67

b. 7.91，51.92

c. 界限：−58.11，117.94

俄罗斯和土耳其是异常值

48. b. 在 x 和 y 之间存在线性关系

c. $s_{xy} = 26.5$

d. $r_{xy} = 0.69$

50. −0.91；负相关

52. a. 0.92

b. 强正线性相关

54. a. 3.69

b. 3.175

56. a. 2.50

b. 能

58. 10.74，25.63，5.06

60. a. 138.52，129，0

b. 不同，高得多

c. 95 169

d. 467 74

e. 9 271.01，96.29

f. 有，467 美元的值

62. a. 18.57，16.5

b. 53.49，7.31

c. Quantex

d. 1.15

e. −0.90

f. 没有，根据 z 分数值

64. 7 195.5；7 019；7 165 941；2 676.93

66. a. 公交工具：32；私人汽车：32

b. 公交工具：4.64；私人汽车：1.83

c. 私人汽车有较小的变异性

68. a. 400，624，836，999，1 278

c. **界限**：61.5，1 561.5

无异常值

70. 51.50，227.37，15.08

72. b. 0.75

74. a. 817

b. 833

76. 51.5，227.37，15.08

第 4 章

2. 20 种方法

4. b. (H,H,H),(H,H,T),(H,T,H),(H,T,T),
(T,H,H),(T,H,T),(T,T,H),(T,T,T)
c. 1/8

6. 0.40，0.26，0.34；相对频数法

8. a. 4: 规划委员会肯定——议会通过，
规划委员会肯定——议会不通过，
规划委员会否定——议会通过，
规划委员会肯定——议会不通过

10. a. 0.60
b. 0.09
c. 0.78
d. 86
e. 15.3 亿美元

12. a. 1 906 884
b. 1/1 906 884
c. 1/80 089 128

14. a. 1/4
b. 1/2
c. 3/4

16. a. 36
c. 1/6
d. 5/18
e. 不同意；P(奇数值) = P(偶数值) = $\frac{1}{2}$
f. 经典方法

18. a. $P(0) = 0.05$
b. P(4 或 5) = 0.20
c. P(0, 1, 或 2) = 0.55

20. a. 0.112
b. 0.086
c. 0.49

22. a. 0.40, 0.40, 0.60
b. 0.80, 是互斥
c. $A^c = (E_3, E_4, E_5)$; $C^c = (E_1, E_4)$;
$P(A^c) = 0.60$; $P(C^c) = 0.40$
d. (E_1, E_2, E_5); 0.60
e. 0.80

24. 0.43

26. a. 0.30, 0.23
b. 0.17
c. 0.64

28. a. 0.698
b. 0.302

30. a. 0.67
b. 0.80
c. 不独立

32. a.

	单身	已婚	总计
30 岁以下	0.55	0.10	0.65
30 岁以上	0.20	0.15	0.35
总计	0.75	0.25	1.0

b. 30 岁以下的来夜总会的概率大
c. 单身来夜总会的概率大
d. 0.55
e. 0.846 2
f. 不独立

34. a. 0.44
b. 0.15
c. 0.022 5
d. 0.002 5
e. 0.136
f. 0.106

36. a.

	满意度分数					
职业	50 以下	50—59	60—69	70—79	80—89	总计
木工	0.000	0.050	0.100	0.075	0.025	0.250
律师	0.150	0.050	0.025	0.025	0.000	0.250
理疗师	0.000	0.125	0.050	0.025	0.050	0.250
系统分析师	0.050	0.025	0.100	0.075	0.000	0.250
总计	0.200	0.250	0.275	0.200	0.075	1.000

b. 0.075
c. 0.20
d. 0.25
e. 0.15
f. 0.60
g. 0.275

38. a. 52/190 = 0.273 7
b. 0.012 5
c. 0.368 4

40. a. 0.10，0.20，0.09
b. 0.51
c. 0.26，0.51，0.23

42. a. 0.21
b. 应该收回

44. 0.675 4

46. a. 0.68
b. 52
c. 10

48. a. 0.61
b. 18—34 和 65 +
c. 0.30

50. a. 76
b. 0.24

52. b. 0.202 2
c. 0.461 8
d. 0.400 5

54. a. 0.49
b. 0.44
c. 0.54
d. 不独立
e. 相同

56. a. 0.25
b. 0.125
c. 0.012 5
d. 0.10
e. 不独立

58. 3.44%

60. a. 0.062 5
b. 0.013 2
c. 3 口井

第 5 章

2. a. x = 以分钟计加工产品所需时间
b. 所有正数；$x>0$
c. 连续的

4. $x=0,1,2,\cdots,12$

6. a. 0, 1, 2, …, 20；离散的
b. 0, 1, 2, …；离散的
c. 0, 1, 2, …, 50；离散的
d. $0\leqslant x\leqslant 8$；连续的
e. $x>0$；连续的

8. a.

x	1	2	3	4
$f(x)$	0.15	0.25	0.40	0.20

c. $f(x)\geqslant 0$，$\sum f(x)=1$

10. a.

x	1	2	3	4	5
$f(x)$	0.05	0.09	0.03	0.42	0.41

b.

x	1	2	3	4	5
$f(x)$	0.04	0.10	0.12	0.46	0.28

c. 0.83
d. 0.28
e. 高级执行经理有更高的满意度

12. a. 是有效的
b. 0.65

14. a. 0.05
b. 0.70
c. 0.40

16. a. 5.20
b. 4.56, 2.14

18. a. $E(x)=2.3$, 相同
b. $\mathrm{Var}(x)=1.23$, $\sigma=1.11$

20. a. 166

b. −94; 关注大事故的赔付额

22. a. 445

b. 1 250 亏损

24. a. 中：145; 高：140

b. 中：2 725; 高：12 400

26. a. $f(0) = 0.348\ 7$

b. $f(2) = 0.193\ 7$

c. 0.929 8

d. 0.651 3

e. 1

f. $\sigma^2 = 0.900\ 0$, $\sigma = 0.948\ 7$

28. a. 0.329 2

b. 0.642 2

c. 0.018 2

30. a. 每一次试验中发现次品的概率必须是 0.03；试验必须是独立的

c. 2

d.

次品数目	0	1	2
概率	0.940 9	0.058 2	0.000 9

32. a. 0.90

b. 0.99

c. 0.999

d. 建议安装

34. a. 0.063 4

b. 0.063 4

c. 0.972 9

38. a. $f(x) = \dfrac{3^x e^{-3}}{x!}$

b. 0.224 1

c. 0.149 4

d. 0.800 8

40. a. 0.195 2

b. 0.104 8

c. 0.018 3

d. 0.090 7

42. a. 0.146 5

b. 1

c. 0.864 7

44. a. $\mu = 1.25$

b. 0.286 5

c. 0.358 1

d. 0.355 4

46. a. 50

b. 0.067

c. 0.466 7

d. 0.30

48. a. 0.50

b. 0.333 3

50. a. 0.01

b. 0.07

c. 0.92

d. 0.07

52. a. 0.533 3

b. 0.666 7

c. 0.777 8

d. $n = 7$

54. a. $f(x) \geqslant 0$ 和 $\sum f(x) = 1$

b. 3.64, 0.670 4

c. 股市价值被高估

56. a. 0.036 4

b. 0.442 0

c. 48

d. 6.788 2

58. a. 0.951 0

b. 0.048 0

c. 0.049 0

60. a. 328

b. 13.91

c. 13.91

62. 0.191 2

64. a. 0.224 0

b. 0.576 7

66. a. 0.533 3

b. 0.133 3

c. 0.3333

第6章

2. b. 0.50
c. 0.60
d. 15
e. 8.33

4. b. 0.50
c. 0.30
d. 0.40

6. a. 0.40
b. 0.64
c. 0.68

10. a. 0.341 3
b. 0.433 2
c. 0.477 2
d. 0.493 8

12. a. 0.296 7
b. 0.441 8
c. 0.330 0
d. 0.591 0
e. 0.884 9
f. 0.238 8

14. a. $z = 1.96$
b. $z = 0.61$
c. $z = 1.12$
d. $z = 0.44$

16. a. $z = 2.33$
b. $z = 1.96$
c. $z = 1.645$
d. $z = 1.28$

18. a. 0.245 1
b. 0.117 0
c. 69.48 分钟或更多

20. a. 0.025
b. 5.16%
c. 33.72 或更长时间

22. a. 0.419 4
b. 517.44 美元或更多
c. 0.0166

24. a. 902.75, 114.185
b. 0.184 1
c. 0.197 7
d. 10.91 亿

26. a. 0.527 6
b. 0.393 5
c. 0.472 4
d. 0.134 1

28. a. 0.3935
b. 0.223 1
c. 0.383 4

30. a. 50 小时
b. 0.393 5
c. 0.135 3

32. a. $f(x) = 30e^{-30x}$
b. 0.082 1
c. 0.713 5

34. a. 63 000 美元
b. 43 800 美元或更少
c. 12.92%
d. 87 675 美元

36. a. 220.33 美元
b. 0.599 9
c. 1 656.78 美元

38. a. 0.022 8
b. 50 美元

40. a. 38.3%
b. 3.59% 更高，96.41% 更低
d. 38.21%

42. $\mu = 19.23$ 盎司

44. a. 4 小时
b. $1/4e^{-x/4}$
c. 0.778 8

d. 0.135 3

46. a. 2 分钟

b. 0.221 2

c. 0.393 5

d. 0.082 1

第 7 章

2. 22，147，229，289

4. a. IBM, Microsoft, Intel, GE, AT&T

b. 252

6. 2 782, 493, 825, 1 807, 289

8. Washington, Clemson, Oklahoma, Colorado, USC, 和 Wisconsin

10. 有限，无限，无限，无限，有限

12. a. 0.50

b. 0.366 7

14. a. 0.19

b. 0.32

c. 0.79

16. 0.80

18. a. 200

b. 5

c. 正态分布具有 $E(\bar{x}) = 200$ 和 $\sigma_{\bar{x}} = 5$

d. $\bar{x}$ 的概率分布

20. 3.54，2.50，2.04，1.77

当 n 增加时，$\sigma_{\bar{x}}$ 减少

22. a. 只对 $n = 30$ 和 $n = 40$ 可用

b. $n = 30$; $E(\bar{x}) = 400$ 和 $\sigma_{\bar{x}} = 9.13$

$n = 40$; 正态分布中, $E(\bar{x}) = 400$

和 $\sigma_{\bar{x}} = 7.91$

24. a. 正态分布中 $E(\bar{x}) = 51\,800$ 和 $\sigma_{\bar{x}} = 516.40$

b. $\sigma_{\bar{x}}$ 递减至 365.15

c. 当 n 增加时，$\sigma_{\bar{x}}$ 减少

26. a. 正态分布中 $E(\bar{x}) = 1.20$ 和 $\sigma_{\bar{x}} = 0.014$

b. 0.841 4

e. 0.522 4

28. a. 0.503 6, 0.621 2, 0.788 8, 0.923 2, 0.987 6

b. 以较高的概率在 ±250 美元内

30. a. 正态分布其 $E(\bar{x}) = 166\,500$ 和 $\sigma_{\bar{x}} = 420\,0$

b. 0.982 6

c. 0.766 0, 0.451 4, 0.189 6

d. 增加样本容量

32. a. $n/N = 0.01$; 否

b. 1.29, 1.30; 忽略误差

c. 0.876 4

34. a. 0.615 6

b. 0.853 0

36. a. 0.615 6

b. 0.781 4

c. 0.948 8

d. 0.994 2

e. 对较大的 n 有较高的概率

38. a. 正态分布中 $E(\bar{p}) = 0.76$ 和 $\sigma_{\bar{p}} = 0.021\,4$

b. 0.838 4

c. 0.945 2

40. a. 正态分布中 $E(\bar{p}) = 0.25$ 和 $\sigma_{\bar{p}} = 0.030\,6$

b. 0.673 0

c. 0.896 8

42. a., 正态分布中 $E(\bar{p}) = 0.15$ 和 $\sigma_{\bar{p}} = 0.050\,5$

b. 0.444 8

c. 0.838 9

44. 112, 145, 73, 324, 293, 875, 318, 618

46. a. 正态分布中 $E(\bar{x}) = 31.5$ 和 $\sigma_{\bar{x}} = 1.70$

b. 0.444 8

c. 0.923 2

48. a. 8.49

b. 0.500 0

c. 0.444 8

d. 0.593 4

50. a. 625

b. 0.788 8

52. a. 正态分布中 $E(\bar{p}) = 0.74$ 和 $\sigma_{\bar{p}} = 0.031$

b. 0.803 0

c. 0.477 8

54. a. 0.960 6

b. 0.049 5

56. a. 48

b. 正态分布，$E(\bar{p}) = 0.25$ 和 $\sigma_{\bar{p}} = 0.062\,5$

c. 0.211 9

第 8 章

2. a. 30.60—33.40

b. 30.34—33.66

c. 29.81—34.19

4. 62

6. 362.80—375.20

8. a. 11 769—12 231

b. 11 725—12 275

c. 11 638—12 362

d. 宽度增加，更加可信

10. 7.25—8.25

12. a. 3.8

b. 0.81

c. 2.99—4.61

14. a. 1.734

b. −1.321

c. 3.365

d. −1.761 和 1.761

e. −2.048 和 2.048

16. a. 15.97—18.53

b. 15.71—18.79

c. 15.14—19.36

18. a. 1.58

b. 0.147 4

c. 1.49—1.67

20. a. 21.15—23.65

b. 21.12—23.68

c. 区间基本相同

22. a. 6.86

b. 6.54 到 7.18

24. a. 9

b. 35

c. 78

26. a. 340

b. 1 358

c. 8 487

28. a. 53

b. 75

c. 129

d. 必须增大 n

30. 59

32. a. 0.673 3—0.726 7

b. 0.668 2—0.731 8

34. 1 068

36. a. 0.439 3

b. 0.387 0—0.491 6

38. a. 0.043 0

b. 0.217 0—0.303 0

c. 822

40. a. 0.250 5

b. 0.026 6

42. a. 0.044 2

b. 601, 1 068, 2 401, 9 604

44. a. 2 009

b. 47 991—52 009

46. a. 49.8

b. 15.99

c. 47.58—52.02

48. a. 13.2

b. 7.8

c. 7.62—18.78

d. 区间增宽；n 应该再大些

50. 37

52. 176

54. a. 0.542 0

b. 0.050 8

c. 0.491 2—0.592 8

56. a. 1 267

b. 1 509

58. a. 0.68

b. 0.639 1—0.720 9

60. a. 0.310 1

b. 0.289 8—0.330 4

c. 8 219；不提供，样本容量不够大

第9章

2. a. H_0: $\mu \leqslant 14$

H_a: $\mu > 14$

4. a. H_0: $\mu \geqslant 220$

H_a: $\mu < 220$

6. a. H_0: $\mu \leqslant 1$

H_a: $\mu > 1$

b. 索赔，当 $\mu > 1$ 不真时

c. 索赔，当 $\mu \leqslant 1$ 不真时

8. a. H_0: $\mu \geqslant 220$

H_a: $\mu < 220$

b. 索赔当 $\mu < 220$ 不真时

c. 索赔当 $\mu \geqslant 220$ 不真时

10. a. 拒绝 H_0 如果 $z > 2.05$

b. 1.36

c. 0.086 9

d. 不拒绝 H_0

12. a. 0.034 4; 拒绝 H_0

b. 0.326 4; 不拒绝 H_0

c. 0.066 8; 不拒绝 H_0

d. 近似为 0；拒绝 H_0

14. a. 拒绝如果 $z > 2.33$

b. 3.11

c. 拒绝 H_0

16. a. H_0: $\mu \geqslant 13$

H_a: $\mu < 13$

b. 拒绝如果 $z < -2.33$

c. -2.88

d. 拒绝 H_0

18. a. H_0: $\mu \leqslant 5.72$

H_a: $\mu > 5.72$

b. 2.12

c. 0.017 0

d. 拒绝 H_0

20. a. H_0: $\mu \leqslant 37\,000$

H_a: $\mu > 37\,000$

b. 1.47

c. 0.070 8

d. 不拒绝 H_0

22. a. 拒绝 H_0 如果 $z < -2.33$ 或 $z > 2.33$

b. 1.13

c. 0.258 4

d. 不拒绝 H_0

24. a. 0.071 8; 不拒绝 H_0

b. 0.652 8; 不拒绝 H_0

c. 0.040 4; 拒绝 H_0

d. 近似为 0；拒绝 H_0

e. 0.317 4; 不拒绝 H_0

26. a. 拒绝 H_0 如果 $z < -1.96$ 或如果 $z > 1.96$

b. -1.71

c. 不拒绝 H_0

28. a. H_0: $\mu = 1\,075$

H_a: $\mu \neq 1\,075$

b. $z = 1.43$

c. 0.152 8

d. 不拒绝 H_0

30. a. H_0: $\mu = 26\,133$

H_a: $\mu \neq 26\,133$

b. -2.09

c. 0.036 6

d. 拒绝 H_0

32. a. 14.66 或更少

b. 拒绝 H_0

34. a. 18

b. 1.41

c. 拒绝 H_0 如果 $t<-2.571$ 或 $t>2.571$

d. -3.47

e. 拒绝 H_0

36. a. 0.01；拒绝 H_0

b. 0.10；接受 H_0

c. 在 0.025 和 0.05 之间；拒绝 H_0

d. 大于 0.10；不拒绝 H_0

e. 近似 0；拒绝 H_0

38. a. 拒绝 H_0 如果 $t<-2.064$ 或 $t>2.064$

b. -1.90

c. 不拒绝 H_0

40. a. H_0: $\mu=4\,000$

H_a: $\mu\neq4\,000$

b. 拒绝 H_0 如果 $t<-2.160$ 或如果 $t>2.160$

c. 1.63

d. 不拒绝 H_0

e. 在 0.10 和 0.20 之间

42. a. H_0: $\mu\leqslant2$

H_a: $\mu>2$

b. 拒绝 H_0 如果 $t>1.833$

c. 2.4

d. 0.516

e. 2.45

f. 拒绝 H_0

g. 在 0.01 和 0.025 之间

44. a. 拒绝 H_0 如果 $z<-1.96$ 或 $z>1.96$

b. -1.25

c. 0.211 2

d. 不拒绝 H_0

46. a. H_0: $p\leqslant0.40$

H_a: $p>0.40$

b. 拒绝 H_0 如果 $z>1.645$

c. 1.99

d. 拒绝 H_0

48. a. 0.57

b. 3.13

c. 小于 0.001

d. 拒绝 H_0

e. 是

50. a. 0.638 1

b. 2.83

c. 0.004 6

d. 拒绝 H_0

52. a. -1.20

b. 0.115 1

c. 不拒绝 H_0

54. a. H_0: $p\geqslant0.047$

H_a: $p<0.047$

b. 0.029 6

c. -2.82

d. 0.002 4

e. 拒绝 H_0

56. a. 0.291 2

b. 第二类错误

c. 0.003 1

58. a. 得出 $\mu\leqslant15$，当它不真时

b. 0.267 6

c. 0.017 9

60. a. 得出 $\mu=28$，当它不真时

b. 0.085 3，0.617 9，0.617 9，0.085 3

c. 0.914 7

62. 0.115 1，0.001 5

增加 n 减少 β

64. 214

66. 109

68. 324

70. a. H_0: $\mu\leqslant45\,250$

H_a: $\mu>45\,250$

b. 2.71

c. 0.003 4

d. 拒绝 H_0

72. $z=2.26$. p-值 $=0.011\ 9$
拒绝 H_0

76. a. 0.014 3
b. 拒绝 H_0

78. a. 0.650 2
b. -0.98
c. 0.327 0
d. 不拒绝 H_0

80. a. 显示 $p<0.50$
b. $z=-6.62$; 拒绝 H_0

82. a. 0.825
b. 2.17
c. 0.03
d. 拒绝 H_0

84. a. 352
b. -2.64
c. 0.004 1
d. 拒绝 H_0

86. 219

第 10 章

2. a. 2.4
b. 5.27
c. 0.09—4.71

4. a. 0.60
b. 0.56—0.64

6. -0.51—1.27

8. a. 1 200
b. 438—1 962
c. 两总体是正态分布且方差相等

10. a. 2.18
b. 4.41
c. 0.71—3.65

12. a. $z=-1.53$; 不拒绝 H_0
b. 0.126 0

14. a. H_0: $\mu_1-\mu_2=0$
H_a: $\mu_1-\mu_2\neq 0$
b. 拒绝如果 $z<-1.96$ 或 $z>1.96$
c. 2.18
d. 拒绝 H_0
e. 0.029 2

16. $z=4.49$, p-值 ≈ 0
拒绝 H_0

18. a. 1.08
b. 0.280 2
c. 不拒绝 H_0

20. a. H_0: $\mu_1-\mu_2\leqslant 0$
H_a: $\mu_1-\mu_2>0$
b. 拒绝 H_0 如果 $t>1.711$
c. 2.07
d. 拒绝 H_0
e. 近似 0.025

22. a. 3, -1, 3, 5, 3, 0, 1
b. 2
c. 2.028 2
d. 2
e. 0.07—3.93

24. 0.16—0.35

26. $t=1.63$; 不拒绝 H_0

28. a. $t=7.34$; 拒绝 H_0
b. 4.96—9.21

30. a. 0.12
b. 0.058 6—0.181 4
c. 0.046 9—0.193 1

32. 0.000 9—0.139 1

34. a. 0.220 6
b. 0.178 8—0.262 4

36. $z=3.94$; 拒绝 H_0

38. $z=2.28$; p-值 $=0.022\ 6$
拒绝 H_0

40. a. 4 354—5 646

42. a. H_0: $\mu_1-\mu_2\leqslant 0$

H_a: $\mu_1 - \mu_2 > 0$

b. $z = 0.59$; 不拒绝 H_0

c. 0.277 6

44. $t = 2.29$; 拒绝 H_0

46. a. H_0: $P_1 - P_2 \leq 0$

H_a: $P_1 - P_2 > 0$

b. $z = 4.70$; p - 值 ≈ 0

c. 拒绝 H_0

48. 0.017 4; 拒绝 H_0

第11章

2. a. 15.76—46.95

b. 14.46—53.33

c. 3.8—7.3

4. a. 0.22—0.71

b. 0.47—0.84

6. a. 13.30

b. 7.97—38.23

8. a. 0.008 45

b. 0.091 9

c. 0.004 2—0.024 4

d. 0.065 1—0.156 1

10. $\chi^2 = 206.22$; 拒绝 H_0

12. a. 0.810 6

b. $\chi^2 = 9.49$; 不拒绝 H_0

14. $F = 2.42$; 拒绝 H_0

16. $F = 2.63$; 拒绝 H_0

18. $F = 3.54$; 不拒绝 H_0

20. $F = 5.29$; 拒绝 H_0

22. a. $F = 4$; 拒绝 H_0

b. 在湿路上要谨慎驾驶

24. 10.72—24.68

26. a. $\chi^2 = 27.44$; 拒绝 H_0

b. 0.000 12—0.000 42

28. $\chi^2 = 31.5$; 拒绝 H_0

30. a. 15

b. 6.25—11.13

32. $F = 1.39$; 不拒绝 H_0

34. $F = 2.08$; 拒绝 H_0

第12章

2. $\chi^2 = 15.33$, $\chi^2_{0.05} = 7.8147\,3$; 拒绝 H_0

4. $\chi^2 = 29.51$, $\chi^2_{0.05} = 11.07$; 拒绝 H_0

数据百分比已发生变化

6. $\chi^2 = 10.69$, $\chi^2_{0.10} = 9.24$; 拒绝 H_0

8. $\chi^2 = 16.31$, $\chi^2_{0.01} = 11.34$; 拒绝 H_0

10. $\chi^2 = 19.78$, $\chi^2_{0.05} = 9.49$; 拒绝 H_0

12. a. $\chi^2 = 7.36$, $\chi^2_{0.05} = 5.99$; 拒绝 H_0

b. 国内的为 47.2%

14. $\chi^2 = 13.43$, $\chi^2_{0.05} = 12.59$; 拒绝 H_0

16. a. $\chi^2 = 7.85$, $\chi^2_{0.05} = 7.81$; 拒绝 H_0

b. 参数, 98.6%

18. $\chi^2 = 13.42$, $\chi^2_{0.01} = 13.28$; 拒绝 H_0

20. $\chi^2 = 9.03$, $\chi^2_{0.05} = 7.81$; 拒绝 H_0

22. $\chi^2 = 4.30$, $\chi^2_{0.05} = 5.99$; 不拒绝 H_0

24. $\chi^2 = 2.8$, $\chi^2_{0.10} = 6.25$; 不拒绝 H_0

26. $\chi^2 = 8.04$, $\chi^2_{0.05} = 7.81$; 拒绝 H_0

28. $\chi^2 = 31.43$, $\chi^2_{0.05} = 7.81$; 拒绝 H_0

30. $\chi^2 = 42.53$, $\chi^2_{0.05} = 9.49$; 拒绝 H_0

32. $\chi^2 = 23.37$, $\chi^2_{0.05} = 7.81$; 拒绝 H_0

34. a. $\chi^2 = 12.86$, $\chi^2_{0.01} = 9.21$; 拒绝 H_0

b. 66.7, 30.3, 2.9

54.0, 42.0, 4.0

36. $\chi^2 = 6.20$, $\chi^2_{0.05} = 12.59$; 不拒绝 H_0

38. $\chi^2 = 7.78$, $\chi^2_{0.05} = 7.81$; 不拒绝 H_0

第13章

2. a. MSTR = 268

b. MSE = 92

c. 不能拒绝 H_0, 因为 $F = 2.91 < F_{0.05} = 4.26$

d.

方差来源	平方和	自由度	均方差	F
处理	536	2	268	2.91
误差	828	9	92	
总计	1364	11		

4. b. 拒绝 H_0，因为 $F=80>F_{0.05}=2.76$

6. 拒绝 H_0，因为 $F=10.63>F_{0.05}=4.26$

8. 显著差异；

$F=7.00>F_{0.05}=3.68$

10. 显著差异；

p-值 $=0.015$

12. 无显著差异；p-值 $=0.403>\alpha=0.05$

14. -8.54 到 -1.46

16. a. 显著差异；$F=19.86>F_{0.05}=3.10$

b. 显著差异；$2.3>\text{LSD}=1.19$

18. 1 和 2：显著差异(LSD = 3.38)

1 和 3：显著差异(LSD = 3.17)

2 和 3：无显著差异(LSD = 3.51)

20. a.

方差来源	平方和	自由度	均方差	F
处理	1 488	2	744	5.50
误差	2 030	15	135.3	
总计	3 518	17		

b. 在 A 和 C 之间有显著差异

22. a. H_0: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$

H_a: 并非所有的总体均值都相等

b. 拒绝 H_0 因为 $F=14.07>2.69$

24. 显著差异；$F=43.99$ 超过临界值

在 3.15 和 3.23 之间

26. b. 显著差异；$F=9.87>F_{0.05}=3.35$

28. 不显著；$F=1.78<F_{0.05}=3.89$

30. 不显著；$F=2.54<F_{0.05}=3.24$

32. 均值都不相等(LSD = 2.53)

34. 显著；$F=6.60>F_{0.05}=4.46$

36. 显著；$F=12.60>F_{0.05}=3.07$

38. 显著；$F=7.12>F_{0.05}=3.26$

40. 显著差异；$F=22.46>F_{0.05}=2.96$

42. 因子 A 是显著的，因为 $F=3.72>F_{0.05}=3.01$

因子 B 是显著的，因为 $F=4.94>F_{0.05}=3.40$

交互影响是显著的，因为 $F=12.52>F_{0.05}=2.51$

44. 上下方法、游乐车类型或交互作用的影响无显著效果

46. 因子 A 无显著影响

因子 B 影响显著

交互影响显著

48. 显著差异；p-值 $=0.000<\alpha=0.05$

50. 显著差异

52. 显著差异

54. 显著：$F=7.23>F_{0.05}=4.26$

56. 不显著；$F=1.48<F_{0.05}=3.35$

58. 显著；$F=5.19>F_{0.05}=4.26$

60. 显著；$F=6.99>F_{0.05}=4.46$

62. 显著差异；p-值 $=0.000<\alpha=0.05$

64. 机器类型有显著影响；装载系统和交互作用无显著影响

第 14 章

2. b. x 和 y 之间存在线性相关关系

d. $\hat{y}=30.33-1.88x$

e. 19.05

4. b. x 和 y 之间存在线性相关关系

d. $\hat{y}=-240.5+5.5x$

e. 106 磅

6. c. $\hat{y}=6.02-0.07x$

e. 0.42

8. a. $\hat{y}=107.13+3.07x$

c. 153.2

10. b. $\hat{y}=51.82+0.145x$

c. 84.4

12. c. $\hat{y}=1\,293+0.316\,5x$

d. 25 031

14. b. $\hat{y}=49.63+2.455x$

c. 69.3%

16. a. SSE = 6.332 5, SST = 114.80, SSR = 108.47

b. $r^2 = 0.945$

c. $r = -0.972\ 1$

18. a. SSE = 85 135.14, SST = 335 000,
SSR = 249 864.86

b. $r^2 = 0.746$

c. $r = +0.863\ 7$

20. a. $\hat{y} = -48.11 + 2.332\ 5x$

b. $r^2 = 0.82$

c. 173 500 美元

22. a. $\hat{y} = -1.02 + 0.13x$

b. $r^2 = 0.363\ 1$

c. $r = +0.602\ 6$

24. a. 2.11

b. 1.453

c. 0.262

d. 显著

$t = -7.18 < -t_{0.05} = -3.182$

e. 显著

$F = 51.41 > F_{0.05} = 10.13$

26. a. 显著

$t = 3.43 > t_{0.025} = 2.776$

b. 显著

$F = 11.74 > F_{0.05} = 7.71$

28. 它们是相关的因为 $F = 20.17 > F_{0.05} = 4.67$

30. 不显著相关

32. a. 1.11

b. 7.07—14.13

c. 2.32

d. 3.22—17.98

34. 置信区间：−0.4—4.98
预测区间：−2.27—7.31

36. a. 80.859

b. 78.58—83.14

c. 72.92—88.80

38. a. 5 046.67 美元

b. 3 815.10 美元—6 278.24 美元

c. 不算过高

40. a. 9

b. $\hat{y} = 20.0 + 7.21x$

c. 1.362 6

d. 显著相关，因为 $F = 28 > F_{0.05} = 5.59$

e. 380 500 美元

42. a. $\hat{y} = 80.0 + 50.0x$

b. 30

c. 显著相关，因为 $F = 83.17 > F_{0.05} = 4.20$

d. 680 000 美元

44. b. 呈现线性关系

c. $\hat{y} = 37.1 - 0.779x$

d. 显著相关；p − 值 = 0.003

e. $r^2 = 0.434$; 拟合度不高

f. 12.28 美元—22.91 美元.

g. 17.49 美元—39.05 美元

46. a. $\hat{y} = 2.32 + 0.64x$

b. 不满足；对所有同样的 x 值，方差并不出现

48. b. 合理

50. a. 存在；$x = 135$，$y = 145$ 可能是异常值

b. 存在

c. 存在

52. a. $\hat{y} = 4.09 + 0.196x$

b. Minitab 确认出观测值 1 有较大的标准残差，我们应将其看做异常值。

54. a. $\hat{y} = 707 + 0.004\ 82x$

b. 观测值 6 是有影响的观测值

58. a. $\hat{y} = 9.26 + 0.711x$

b. 显著相关

c. $r^2 = 0.744$；拟合度好

d. 13.53 美元

60. a. 市场 $\beta = 0.95$

b. 显著相关

c. $r^2 = 0.470$; 拟合度不好

d. 有较高的风险

62. a. $\hat{y} = 10.5 + 0.953x$

b. 显著相关；p - 值 = 0.000

c. 2 874 美元—4 952 美元

d. 建议签

64. a. 负线性相关

b. $\hat{y} = 8.10 - 0.344x$

c. 显著相关；p - 值 = 0.002

d. $r^2 = 0.711$; 合理的拟合度

e. 5.2 到 7.6 天

66. a. $\hat{y} = 5.85 + 0.830x$

b. 显著相关；p - 值 = 0.000

c. 84.65 点

d. 65.35—103.96

第 15 章

2. a. $\hat{y} = 45.06 + 1.94x_1$；$\hat{y} = 132.36$

b. $\hat{y} = 85.22 + 4.32x_2$；$\hat{y} = 150.02$

c. $\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$;
$\hat{y} = 143.18$

4. a. 255 000 美元

6. a. 速度 = 49.8 + 0.015 自重

b. 速度 = 80.5 - 0.00312 自重 + 0.105 马力

8. 年回报率 = 247 - 32.8 安全系数 + 34.6 年支出率

b. 70.2

10. a. 收入 = 33.3 + 7.98 汽车数

b. 每增加 1 000 辆汽车，将增加 798 万美元收入

c. 收入 = 106 + 8.94 汽车数 - 0.191 场地

12. a. 0.926

b. 0.905

c. 是

14. a. 0.75

b. 0.68

16. a. 否，$R^2 = 0.311$

b. 多元回归分析

18. a. $R^2 = 0.942, R_a^2 = 0.932$

b. 拟合度是优

20. a. 显著的；p - 值 = 0.000

b. 显著的；p - 值 = 0.000

c. 显著的；p - 值 = 0.000

22. a. SSE = 4 000, $s^2 = 571.43$,
MSR = 6 000

b. 显著；$F = 10.50 > F_{0.05} = 4.74$

24. a. 拒绝 H_0: $\beta_1 = \beta_2 = 0$; p - 值 = 0.000

b. 自重：不拒绝 H_0: $\beta_1 = 0$; p - 值 = 0.386
马力: 拒绝 H_0: $\beta_2 = 0$; p - 值 = 0.000

26. a. 显著；p - 值 = 0.000

b. 显著；p - 值 = 0.000

c. 不显著；p - 值 = 0.087

28. a. 132.16 到 154.15

b. 111.15 到 175.17

30. a. 速度 = 72.6 + 0.096 8 马力；速度估计值是 101.29

b. 99.49 到 103.09

c. 94.594 到 107.986

32. a. $x_2 = 0$ 如果水平 1；$x_2 = 1$ 如果水平 2
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

b. $E(y) = \beta_0 + \beta_1 x_1$

c. $E(y) = \beta_0 + \beta_1 x_1 + \beta_2$

d. $\beta_2 = E(y|\text{水平 2}) - E(y|\text{水平 1})$

34. a. 15 300 美元

b. 56 100 美元

c. 41 600 美元

36. a. $\hat{y} = 1.86 + 0.291$ 月数 $+ 1.10$ 故障类型 $- 0.609$ 维修人员

b. 显著；p - 值 = 0.002 < $\alpha = 0.05$

c. 维修人员是不显著的

38. a. $\hat{y} = -91.8 + 1.08$ 年龄 $+ 0.252$ 血压 $+$ 8.74 吸烟者

b. 显著；p - 值 = 0.01 < $\alpha = 0.05$

c. 95% 的预测区间是 21.35 到 47.19，或是 0.213 5 到 0.471 9 的概率；停止抽烟或采取一些降低血压的措施

40. a. $\hat{y} = -53.3 + 3.11x$

b. −1.40, −0.15, 1.36, 0.47, −1.39; 不支持

c. 0.38, 0.28, 0.22, 0.20, 0.98; 无

d. 0.60, 0.00, 0.26, 0.03, 11.09; 有，第 5 个观测值

42. b. 不支持

c. 无异常值

d. 观测值 2 是有影响 的观测值

44. b. 3.19

46. a. 3.04, 3.61, 5.08

b. 两者都显著

d. 0.91

48. b. 显著； $F = 22.79 > F_{0.05} = 5.79$

c. $R^2_a = 0.861$; 拟合度好

d. 两者都显著

50. a. 速度 = 97.6 + 0.069 3 价格 − 0.000 82 自重 + 0.059 0 马力 − 2.48 加速到 60

b. 显著相关

c. 价格和重量不显著相关

d. 速度 = 103 + 0.055 8 马力 − 3.190 加速到 60

e. 不合理

f. 观测值 2 是异常值

g. 观测值 12 是有影响的观测值

52. a. 转让销售率 = 38.8 + 0.000 766 建议零售价

b. 拟合度不高； R-Sq = 36.7%

c. 转让销售率 = 42.6 + 9.09 车型 1 + 7.92 车型 2 + 0.000 341 建议零售价

其中 1 = 1，如果是轻便车；车型 2 = 1，如果是赛车

d. 显著相关； p − 值相应于 $F = 14.79 = 0.000 < \alpha = 0.05$

第 16 章

2. a. $\hat{y} = 9.32 + 0.424x$; p − 值 = 0.117 表明 x 和 y 之间的相关性不显著

b. $\hat{y} = -8.10 + 2.41x - 0.048\,0x^2$

$R^2_a = 0.932$; 拟合度好

c. 20.965

4. a. $\hat{y} = 943 + 8.71x$

b. 显著； p − 值 = 0.005 < α = 0.01

6. b. 不运用，适宜用曲线相关。

c. 几种可能的模型

$\hat{y} = 2.90 - 0.185x + 0.003\,51x^2$

8. a. 不适用简单线性回归模型

b. 2 005% = 17.1 + 3.15 × 1 999% − 0.044 5 × 1 999% Sq

c. Log2 000% = 1.17 + 0.449Log1 999%

d. 部分(b); 较高的变动百分比

10. a. 显著； $F = 49.52 > 4.24$

b. 显著； $F = 48.3 > 3.42$

c. 显著； $t = -4.46 < -2.069$(此答案 C、D 与习题不符，疑有误。——编者注)

d. x_2 可能被减弱

12. a. $\hat{y} = 170 + 6.61$ 抢断次数

b. $\hat{y} = 280 + 5.18$ 抢断次数 − 0.003 7 带球 − 3.92 对手抢断次数

c. 两个独立变量之和不显著

14. a. $\hat{y} = -111 + 1.32$ 年龄 + 0.296 血压

b. $\hat{y} = -123 + 1.51$ 年龄 + 0.448 血压 + 8.87 吸烟 − 0.002 76 年龄血压

c. 显著

16. a. 考入大学百分比 = −26.6 + 0.097 0 SAT 分数

b. 考入大学百分比 = −26.93 + 0.084 SAT 分数 + 0.204% 综合 SAT

c. 同(b)

18. a. Greens 占的百分比

b. 平均分数 = 58.2 − 0.009 96 Distanle − 0.152 Greens % + 0.869Putts

c. 能

d. 72.65

20. $\hat{y} = -91.8 + 1.08$ 年龄 + 0.252 血压 + 8.74 吸烟者

22. $d = 1.60$; 检验未能得出结论

24.

x_1	x_2	处理
0	0	1
1	0	2
0	1	3

$x_3 = 0$ 如果群组 1；$x_3 = 1$ 如果群组 2

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

26. a.

D_1	D_2	生产厂商
0	0	1
1	0	2
0	1	3

$E(y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$

b. $\hat{y} = 23.0 + 5.00 D_1 - 2.00 D_2$

c. H_0: $\beta_1 = \beta_2 = 0$

d. 对每个生产厂商来说，平均时间并不相等；$p-$ 值 $= 0.004$

28. 两个分析仪之间有显著差异

30. a. 令 ExS 为费用比例和安全比例之间的交互作用

业绩百分比 $= 23.3 + 222$ 费用百分比 $- 28.9$ExS

b. R-Sq(adj) $= 65.3\%$

c. 25.8 或近似为 26%

32. a. AUDELAY $= 63.0 + 11.1$ INDUS；正的自相关性不显著

34. 三种类型购物者的舒适度有显著差异

第 17 章

2. a. 32%

b. 8.14 美元

4. $I_{2001} = 105$

6. $I = 125$

8. $I = 105$; 证券组合高于 5%

10. a. 1980 工资：8.82 美元

2000 工资：8.32 美元

b. 97.5% 增加

c. 5.7% 减少

12. a. 2 448, 2 486, 2 557

按美元计算，制造业每年略有增长

b. 2 981，3 100，3 203

c. PPI

14. $I = 110$

16. $I = 83$

18. a. 151, 197, 143, 178

b. $I = 170$

20. $I_{Jan} = 96$, $I_{Mar} = 92$

22. $I = 117$

24. 36 082 美元; 32 528 美元; 27 913 美元; 34 387 美元; 40 551 美元; 42 651 美元; 46 350 美元

26. $I = 143$; 数量高于 43%

第 18 章

2. a.

周	4 周	5 周
10	19.00	18.80
11	20.00	19.20
12	18.75	19.00

b. 9.65, 7.41

c. 5 周

4. 第 10，11 和 12 周的：18.48, 18.63, 18.27

MSE = 9.25; $\alpha = 0.2$ 更好一些

6. a. MSE(3 月) = 1.24

MSE($\alpha = 0.2$) = 3.55

使用 3 月移动平均

b. 83.3

8. a.

月	3 月	$\alpha = 0.2$
10	256.67	265.51
11	286.67	274.41
12	263.33	267.53

对两者仅使用第 4 到第 12 个月的误差，

$\alpha=0.2$ 更好些

b. 260

10. c. 用 $\alpha=0.3$; $F_{11}=7.57$

12. $T_t=4.7-2.1t$; 17.3

14. $T_t=20.7466-0.3514t$

16. 考虑一个非线性趋势

18. a. $T_t=0.365+0.193t$; 2.49 美元

b. EPS 按平均 0.193 美元的比例增长

20. a. $T_t=1\,997.6+397.545t$

b. $T_{11}=6\,371$, $T_{12}=6\,768$

22. a. 四个季度的滑动平均值：3.50, 4.00, 4.25, 4.75, 5.25, 5.50, 6.25, 6.50, 6.75

移动平均中间值：3.750, 4.125, 4.500, 5.000, 5.375, 5.875, 6.375, 6.625

b. 调整季节指数：1.205 0, 0.746 3, 0.8675, 1.191 2

注：调整 = 0.991 2

24. 调整季节指数：0.707, 0.777, 0.827, 0.966, 1.016, 1.305, 1.494, 1.225, 0.976, 0.986, 0.936, 0.787

注：调整 = 0.996

26. a. 是

b. 12—4：166 761.13

4—8：146 052.99

28. a. 0.2 更好一些

b. 46.1

30. 20.26

32. a. $\alpha=0.5$

b. $T_t=244.778+22.088t$

c. 趋势推测法；较小的 MSE

34. $T_8=252.28$, $T_9=259.10$

36. a. 是

b. $T_t=25+15t$

38. a. 出现线性趋势是合理的

b. $T_t=12\,899.98+2\,092.066t$

c. 2 092 066 美元

d. 1997: 40 096 838 美元

1998: 42 188 904 美元

40. b. 调整季节指数：0.899, 1.362, 1.118, 0.621

注：调整 = 1.010 1

c. 第 2 个季节；似乎合理

42. a. $T_t=6.329+1.055t$

b. 36.92, 37.98, 39.03, 40.09

c. 33.23, 51.65, 43.71, 24.86

第 19 章

2. $z=3.27$; 拒绝 H_0

4. $z=3.15$; 拒绝 H_0

6. $z=2.32$; 拒绝 H_0

8. $z=3.76$; 拒绝 H_0

10. $z=1.27$; 不拒绝 H_0

12. $z=2.43$; 拒绝 H_0

14. $z=2.29$; 拒绝 H_0

价格有差异

16. $z=2.62$; 拒绝 H_0

18. $T=34$; 拒绝 H_0

20. $T=36$; 拒绝 H_0

22. $z=2.77$; 拒绝 H_0

P/E 比有差异

24. $z=-0.25$; 不拒绝 H_0

26. $W=10.22$; 拒绝 H_0

28. $W=9.26$; 拒绝 H_0

30. $W=8.03$; 拒绝 H_0

32. a. 0.68

b. $z=2.06$; 拒绝 H_0

34. $z=0.72$; 不拒绝 H_0

36. $r_s=-0.709$; $z=-2.13$; 拒绝 H_0

38. $z=3.17$; 拒绝 H_0

40. (此答案与练习不符，疑有误。—编者注)

a. $z=-3.20$; 拒绝 H_0

休斯敦在全国平均标准之下

b. $z=2.21$; 拒绝 H_0

高于全国平均标准

42. $z = -2.97$; 拒绝 H_0

44. $W = 12.61$; 拒绝 H_0

46. $r_s = 0.76$; $z = 2.84$; 拒绝 H_0

第 20 章

2. a. 5.42

b. UCL = 6.09, LCL = 4.75

4.

	R 控制图	$\bar{x}$ 控制图
UCL	2.98	29.10
LCL	0.22	27.90

6. 20.01, 0.082

8. a. 0.047 0

b. UCL = 0.098 9, LCL = 2.004 9(用 LCL = 0)

c. $\bar{p} = 0.08$; 处于控制状态

d. UCL = 14.826, LCL = −0.726

如果有缺限的产品大于 14，则处于过程失控状态

e. 12 个有缺陷的，在控制限以内

f. np 控制图

10. $p = 0.02$; $f(0) = 0.603\ 5$

$p = 0.06$; $f(0) = 0.212\ 9$

12. $p_0 = 0.02$; 生产者的风险 = 0.059 9

$p_0 = 0.06$; 生产者的风险 = 0.339 6

当可接受值 c 增加时，生产者的风险减少。

14. $n = 20$, $c = 3$

16. a. 95.4

b. UCL = 96.07, LCL = 94.73

c. 否

18.

	R 控制图	$\bar{x}$ 控制图
UCL	4.23	6.57
LCL	0	4.27

标准差的估计值 = 0.86

20.

	R 控制图	$\bar{x}$ 控制图
UCL	0.112 1	3.112
LCL	0	3.051

22. a. UCL = 0.081 7, LCL = −0.001 7
(用 LCL = 0)

24. a. 0.03

b. $\beta = 0.080\ 2$

26. a. 生产者风险 = 0.006 4

b. 消费者风险 = 0.013 6

c. **优点**：控制效果好

缺点：成本高

第 21 章

2. a. 30 000

b. 320

c. 29 360—30 640

4. 73

6. 337

8. a. 分层 1: 27 600

分层 2: 25 750

分层 3: 21 000

b. 74 350

c. 70 599.88—78 100.12

10. a. $n = 93$, $n_1 = 30$, $n_2 = 30$, $n_3 = 33$

b. $n = 306$, $n_1 = 98$, $n_2 = 98$, $n_3 = 109$

c. $n = 275$, $n_1 = 88$, $n_2 = 88$, $n_3 = 98$

12. a. 3 617 000 美元

b. 1 122 265 美元

c. 41 066 美元—56 499 美元

d. 9 568 261 美元—13 164 197 美元

14. a. 15, 4 500, 0.30 美元

b. 1.470 8，441.24，0.048 4

c. 12.058 4—17.941 6

d. 3 617.52—5 382.48

e. 0.203 2—0.396 8

16. a. 40

b. 0.70

c. 35.863 4—44.136 6

d. 0.523 4—0.876 6

18. a. 0.148 8—0.231 2

b. 0.261 5—0.358 5

c. 0.130 6—0.209 4

20. a. 22 790 美元—23 610 美元

b. 68 370 366 美元—70 829 634 美元

c. 0.669 2—0.790 8

22. a. 431

b. 0.217 5—0.398 3

c. 0.623 0—0.800 2

d. 996

24. a. 75.275

b. 0.198—0.502

c. 1 680

附录 E 自测题解答

第 1 章

2. a. 9

b. 4

c. 国家和房间等级是定性变量；房间数量和总分是定量变量

d. 国家是名义尺度；房间等级是序数尺度；房间数量和总分是比例尺度

3. a. 平均房间数 = 808/9 = 89.78 或近似为 90 间

b. 2/9 在英国约占 22%

c. 4/9 需要 $$ 级的客房，约占 44%

4. a. 10

b. 《财富》美国 500 家最大的工业公司

c. 平均收入 = 142 275.9/10 美元 = 14 227.59 百万美元

d. 利用样本平均值，统计推断使我们估计出 500 家公司总体的平均收入为 14 277.59 百万美元

13. a. 定量的

b. 具有 7 个观测值的时间序列

c. 河船娱乐场的数量

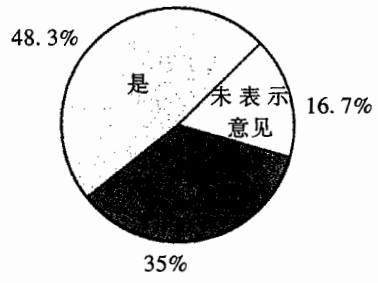
d. 时间序列显示了快速增长；1998 年预期增长，但增速变缓

第 2 章

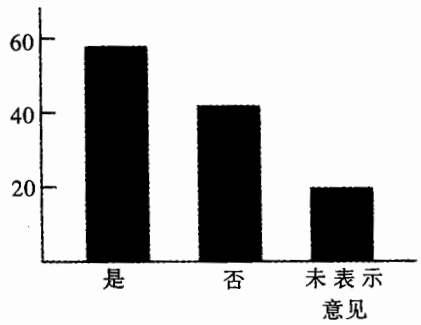
3. a. $360° \times 58/120 = 174°$

b. $360° \times 42/120 = 126°$

c.



d.



7.

等级	频数	百分比频数
优异	19	0.38
非常好	13	0.26
好	10	0.20
一般	6	0.12
差	2	0.04

经理对此结果会高兴的：64% 等级是非常好到优异，84% 的等级是好或者较好；将这些等级与以前的结果比较，可表明顾客对饭菜质量的评价是否有所提高。

12.

组别	累积频数	累积相对频数
≤19	10	0.20
≤29	24	0.48
≤39	41	0.82
≤49	48	0.96
≤59	50	1.00

15. a/b.

等待时间	频数	相对频数
0 — 4	4	0.20
5 — 9	8	0.40
10 — 14	5	0.25
15 — 19	2	0.10
20 — 24	1	0.05
总计	20	1.00

c/d.

组别	累积频数	累积相对频数
≤4	4	0.20
≤9	12	0.60
≤14	17	0.85
≤19	19	0.95
≤24	20	1.00

e. 12/20 = 0.60

23. 叶单位 = 0.1

6	3
7	5 5 7
8	1 3 4 8
9	3 6
10	0 4 5
11	3

25.

9	8 9
10	2 4 6 6
11	4 5 7 8 8 9
12	2 4 5 7
13	1 2
14	4
15	1

29.

a.

x \ y	1	2	总计
A	5	0	5
B	11	2	13
C	2	10	12
总计	18	12	30

b.

x \ y	1	2	总计
A	100.0	0.0	100.0
B	84.6	15.4	100.0
C	16.7	83.3	100.0

c.

x \ y	1	2
A	27.8	0.0
B	61.1	16.7
C	11.1	83.3
总计	100.0	100.0

d. A 值总是处于 $y=1$

B 值几乎总是处于 $y=1$

C 值几乎总是处于 $y=2$

32. a.

销售额/边际收益/权益收益	0—19	20—39	40—59	60—79	80—100	总计
A				1	8	9
B		1	4	5	2	12
C	1		1	2	3	7
D	3	1		1		5
E		2	1			3
总计	4	4	6	9	13	36

b.

销售额/边际收益/权益收益	每股收益评分 0—19	20—39	40—59	60—79	80—100	总计
A				11.11	88.89	100
B		8.33	33.33	41.67	16.67	100
C	14.29		14.29	28.57	42.86	100
D	60.00	20.00		20.00		100
E		66.67	33.33			100

较高的每股收益评分似乎与较高的销售额/边际收益/权益收益一致；它们中的销售额/边际收益/权益收益为 88.89% 的 A 级公司中有 80 或更高的每股收益评分；8 个 D 或 E 级的公司仅有 1 个的每股收益评分超过 60

第 3 章

3. 将数据排序为:15,20,25,25,27,28,30,34

$i=\frac{20}{100}(8)=1.6$; 靠近第 2 位数

第 20 百分位数 = 20

$i=\frac{25}{100}(8)=2$; 用第 2 和第 3 位数

第 25 百分位数 $=\frac{20+25}{2}=22.5$

$i=\frac{65}{100}(8)=5.2$; 靠近第 6 位数

第 65 百分位数 = 28

$i=\frac{75}{100}(8)=6$; 用第 6 和第 7 位数

第 75 百分位数 $=\frac{28+30}{2}=29$

8. a. $\bar{x}=\frac{\sum x_i}{n}=\frac{775}{20}=38.75$

众数 = 29(出现 3 次)

b. 数据排序：22,24,29,29,29,30,31,31,32,37,40,41,44,44,46,49,50,52,57,58

中位数(第 10 第 11 个位置)

$\frac{37+40}{2}=38.5$

在家中办公的人年轻一些

c. $i=\frac{25}{100}(20)=5$; 用第 5 和第 6 位数

$Q_1=\frac{29+30}{2}=29.5$

$i=\frac{75}{100}(20)=15$; 用第 15 和第 16 位数

$Q_3=\frac{46+49}{2}=47.5$

d. $i=\frac{32}{100}(20)=6.4$; 靠近第 7 位数

第 32 百分位数 = 31

至少 32% 的人的年龄是 31 岁或更小

17. 根差 = 34 − 15 = 19

将数据排序为:15,20,25,25,27,28,30,34

$i=\frac{25}{100}(8)=2$; $Q_1=\frac{20+25}{2}=22.5$

$i=\frac{75}{100}(8)=6$; $Q_3=\frac{28+30}{2}=29$

IQR = $Q_3-Q_1=29-22.5=6.5$

$\bar{x}=\frac{\sum x_i}{n}=\frac{204}{8}=25.5$

x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
27	1.5	2.25
25	−0.5	0.25
20	−5.5	30.25
15	−10.5	110.25
30	4.5	20.25
34	8.5	72.25
28	2.5	6.25
25	−0.5	0.25
		242.00

$s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{242}{8-1}=34.57$

$s=\sqrt{34.57}=5.88$

18. a. 极差 = 190 − 168 = 22

b. $\bar{x}=\frac{\sum x_i}{n}=\frac{1\,068}{6}=178$

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

$$=\frac{4^2+(-10)^2+6^2+12^2+(-8)^2+(-4)^2}{6-1}$$

$$=\frac{376}{5}=75.2$$

c. $s=\sqrt{75.2}=8.67$

d. $\frac{s}{\bar{x}}(100)=\frac{8.67}{178}(100)=4.87$

27. 切比雪夫定理：至少$(1-1/z^2)$

a. $z=\frac{40-30}{5}=2;\ (1-1/2^2)=0.75$

b. $z=\frac{45-30}{5}=3;\ (1-1/3^2)=0.89$

c. $z=\frac{38-30}{5}=1.6;\ (1-1/1.6^2)=0.61$

d. $z=\frac{42-30}{5}=2.4;\ (1-1/2.4^2)=0.83$

e. $z=\frac{48-30}{5}=3.6;\ (1-1/3.6^2)=0.92$

40. 将数据排序为:5,6,8,10,10,12,15,16,18,

$i=\frac{25}{100}(9)=2.25$; 靠近第 3 位数

$Q_1=8$

中位数(第 5 个位置) = 10

$i=\frac{75}{100}(9)=6.75$; 靠近第 7 位数

$Q_3=15$

五数概括：5，8，10，15，18



43. a. 将数据由低到高排序

$i=\frac{25}{100}(21)=5.25$; 靠近第 6 位数

$Q_1=1\ 872$

中位数(第 11 个位置) = 4 019

$i=\frac{75}{100}(21)=15.75$; 靠近第 16 位数

$Q_3=830\ 5$

五数概括:608,1 872,4 019,8 305,14 138

b. IQR = $Q_3-Q_1=8\ 305-1\ 872=6\ 433$

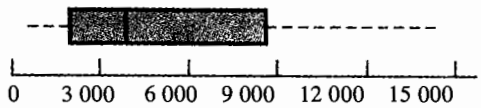
下限值：$1\ 872-1.5(6\ 433)=-7\ 777$

上限值：$8\ 305+1.5(6\ 433)=17\ 955$

c. 没有；数据在限值之内

d. 41 138>27 604; 41 138 应该算异常值；数据值应重新核对和更正

e.



47. b. 在 x 和 y 之间呈负线性相关关系

c.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
4	50	−4	4	−16
6	50	−2	4	−8
11	40	3	−6	−18
3	60	−5	14	−70
16	30	8	−16	−128
40	230	0	0	−240

$\bar{x}=8;\ \bar{y}=46$

$$s_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{-240}{4}=-60$$

样本协方差表明 x 和 y 之间存在着负线性关系

d. $r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{-60}{(5.43)(11.40)}=-0.97$

样本协方差系数 −0.97 表明存在很强的负线性相关关系

55. a.

f_i	M_i	$f_i M_i$
4	5	20
7	10	70
9	15	135
5	20	100
25		325

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{325}{25} = 13$$

b.

f_i	M_i	$(M_i - \bar{x})$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
4	5	−8	64	256
7	10	−3	9	63
9	15	2	4	36
5	20	7	49	245
25				600

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1} = \frac{600}{25-1} = 25$$

$$s = \sqrt{25} = 5$$

56. a.

等级分 x_i	权重 w_i
4(A)	9
3(B)	15
2(C)	33
1(D)	3
0(F)	0
	60 学时

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{9(4) + 15(3) + 33(2) + 3(1)}{9 + 15 + 33 + 3}$$

$$= \frac{150}{60} = 2.5$$

b. 能

第 4 章

2. $\binom{6}{3} = \frac{6!}{3!3!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(3 \cdot 2 \cdot 1)} = 20$

ABC	ACE	BCD	BEF
ABD	ACF	BCE	CDE
ABE	ADE	BCF	CDF
ABF	ADF	BDE	CEF
ACD	AEF	BDF	DEF

6. $P(E_1) = 0.40$, $P(E_2) = 0.26$, $P(E_3) = 0.34$

利用相对频数法

9. $\binom{50}{4} = \frac{50!}{4!46!} = \frac{50 \cdot 49 \cdot 48 \cdot 47}{4 \cdot 3 \cdot 2 \cdot 1} = 230\ 300$

10. a. 利用相对频数法

P(加利福尼亚) = 1 434/2 374 = 0.60

b. 不来自这四个州的数

= 2 374 − 1 434 − 390 − 217 − 112

= 221

P(不来自这四个州) = 221/2 374 = 0.09

P(不是处于早期阶段) = 1 − 0.22 = 0.78

d. 马萨诸塞的公司处于早期发展阶段的估计数 = (0.22)390 ≈ 86

e. 如果假设各州接受投资的规模无差别，我们可用风险投资总支付额乘以投向科罗拉多州的概率进行估计

科罗拉多基金的估计 = (112/2 374)(32.4 美元) = 15.3 亿美元

作者注：实际到科罗拉多的资金额是 17.4 亿美元

15. a. S = (草花点数，方格点数，红桃点数，黑桃点数)

b. S = (草花点数)

c. 四种花色其中之一的 J、Q 或 K

d. (a): 4/52 = 1/13 = 0.08

(b): 13/52 = 1/4 = 0.25

(c): 12/52 = 0.23

17. a. (4, 6), (4, 7), (4, 8)

b. 0.05 + 0.10 + 0.15 = 0.30

c. (2, 8), (3, 8), (4, 8)

d. 0.05 + 0.05 + 0.15 = 0.25

e. 0.15

23. a. $P(A) = P(E_1) + P(E_4) + P(E_6)$

$$= 0.05 + 0.25 + 0.10 = 0.40$$

$$P(B) = P(E_2) + P(E_4) + P(E_7)$$
$$= 0.20 + 0.25 + 0.05 = 0.50$$

$$P(C) = P(E_2) + P(E_3) + P(E_5) + P(E_7)$$
$$= 0.20 + 0.20 + 0.15 + 0.05 = 0.60$$

b. $A \cup B = \{E_1, E_2, E_4, E_6, E_7\}$

$$P(A \cup B) = P(E_1) + P(E_2) + P(E_4) + P(E_6) + P(E_7)$$
$$= 0.05 + 0.20 + 0.25 + 0.10 + 0.05$$
$$= 0.65$$

c. $A \cap B = \{E_4\}$, $P(A \cap B) = P(E_4) = 0.25$

d. 是的，它们互相排斥

e. $B^c = \{E_1, E_3, E_5, E_6\}$

$$P(B^c) = P(E_1) + P(E_3) + P(E_5) + P(E_6)$$
$$= 0.05 + 0.20 + 0.15 + 0.10$$
$$= 0.50$$

28. 令 B = 因商务而租车

P = 因私事而租车

a. $P(B \cup P) = P(B) + P(P) - P(B \cap P)$
$$= 0.540 + 0.458 - 0.300$$
$$= 0.698$$

b. P(不租车) $= 1 - 0.698 = 0.302$

30. a. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.40}{0.60} = 0.666\,7$

b. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.40}{0.50} = 0.80$

c. 不独立，因为 $P(A \mid B) \neq P(A)$

33. a.

申请理由

	学校质量	学费或方便性	其他	总计
全日制	0.218	0.204	0.039	0.461
非全日制	0.208	0.307	0.024	0.539
总计	0.426	0.511	0.063	1.000

b. 学生最有可能是以学费或方便为首要理由(概率 = 0.511)；以学校质量为首要理由的学生第二多(概率 = 0.426)

c. (学校质量 | 全日制) $= 0.218/0.461 = 0.473$

d. P(学校质量 | 非全日制) $= 0.208/0.539 = 0.386$

e. 由于独立性，我们有 $P(A)P(B) = P(A \cap B)$；从表中得出 $P(A \cap B) = 0.218$，$P(A) = 0.461$, $P(B) = 0.426$

$$P(A)P(B) = (0.461)(0.426) = 0.196$$

因为 $P(A)P(B) \neq P(A \cap B)$，事件不独立

39. a. 互斥，因为 $P(A_1 \cap A_2) = 0$

b. $P(A_1 \cap B) = P(A_1)P(B \mid A_1) = 0.40(0.20) = 0.08$

$P(A_2 \cap B) = P(A_2)P(B \mid A_2) = 0.60(0.05) = 0.03$

c. $P(B) = P(A_1 \cap B) + P(A_2 \cap B)$
$$= 0.08 + 0.03 = 0.11$$

d. $P(A_1 \mid B) = \dfrac{0.08}{0.11} = 0.727\,3$

$$P(A_2 \mid B) = \frac{0.03}{0.11} = 0.272\,7$$

42. M = 坏账

D_1 = 客户拖欠

D_2 = 客户未拖欠

$P(D_1) = 0.05$, $P(D_2) = 0.95$, $P(M \mid D_2) = 0.2$, $P(M \mid D_1) = 1$

a. $P(D_1 \mid M) =$

$$\frac{P(D_1)P(M \mid D_1)}{P(D_1)P(M \mid D_1) + P(D_2)P(M \mid D_2)}$$

$$= \frac{(0.05)(1)}{(0.05)(1) + (0.95)(0.2)}$$

$$= \frac{0.05}{0.24} = 0.21$$

b. 应该收回，因造成坏账的概率已超过 0.20

第 5 章

1. a. 正，正(H, H)

正，反(H, T)

反，正(T, H)

反，反(T, T)

b. x = 掷两次硬币正面朝上的次数

c.

结果	x 的值
(H, H)	2
(H, T)	1
(T, H)	1
(T, T)	0

d. 离散的；可定为 3 个值：0，1 和 2

3. 令：Y = 录用

N = 不录用

a. $S = \{(Y, Y, Y), (Y, Y, N), (Y, N, Y), (Y, N, N), (N, Y, Y), (N, Y, N), (N, N, Y), (N, N, N)\}$

b. 令 N = 录用的数目；N 是离散随机变量

c.

面试结果	(Y, Y, Y)	(Y, Y, N)	(Y, N, Y)	(Y, N, N)	(N, Y, Y)	(N, Y, N)	(N, N, Y)	(N, N, N)
N 的值	3	2	2	1	2	1	1	0

7. a. $f(x) \geqslant 0$ 对所有 x 值

$\sum f(x) = 1$；于是，它可以为概率分布

b. 概率 $x = 30$ 是 $f(30) = 0.25$

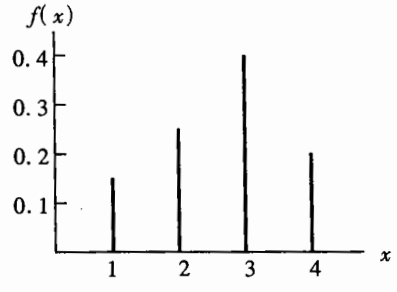
c. 概率 $x \leqslant 25$ 是 $f(20) + f(25) = 0.20 + 0.15 = 0.35$

d. 概率 $x > 30$ 是 $f(35) = 0.40$

8. a.

x	$f(x)$
1	3/20 = 0.15
2	5/20 = 0.25
3	8/20 = 0.40
4	4/20 = 0.20
	总计 1.00

b.



c. $f(x) \geqslant 0$ 对 $x = 1, 2, 3, 4$

$\sum f(x) = 1$

16. a.

y	$f(y)$	$yf(y)$
2	0.20	0.40
4	0.30	1.20
7	0.40	2.80
8	0.10	0.80
总计	1.00	5.20

$E(y) = \mu = 5.20$

b.

y	$y-\mu$	$(y-\mu)^2$	$f(y)$	$(y-\mu)^2 f(y)$
2	−3.20	10.24	0.20	2.048
4	−1.20	1.44	0.30	0.432
7	1.80	3.24	0.40	1.296
8	2.80	7.84	0.10	0.784
			总计	4.560

$\text{Var}(y) = 4.56$

$\sigma = \sqrt{4.56} = 2.14$

18. a/b.

x	$f(x)$	$xf(x)$	$(x-\mu)$	$(x-\mu)^2$	$(y-\mu)^2 f(y)$
0	.01	.00	−2.3	5.29	0.052 9
1	.23	.23	−1.3	1.69	0.388 7
2	.41	.82	−0.3	0.09	0.036 9
3	.20	.60	0.7	0.49	0.098 0
4	.10	.40	1.7	2.89	0.289 0
5	.05	.25	2.7	7.29	0.364 5

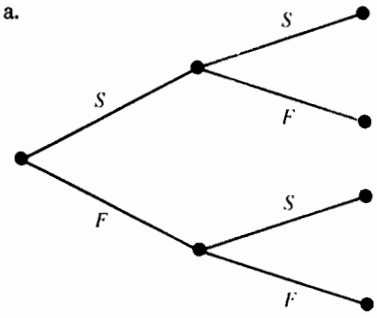
$E(x) = 2.30$ $\text{Var}(x) = 1.230\,0$

$\sigma = 1.11$

概率分布的数学期望 $E(x) = 2.3$，与 1997 年《美国统计摘要》(*Statistical Abstract of the United States*)报道的电视机的平均拥有量 $Var(x) = 1.23$ 相同，标准差

$\sigma = \sqrt{1.23} = 1.11$

25. a.



b. $f(1)=\binom{2}{1}(0.4)^1(0.6)^1=\frac{2!}{1!1!}(0.4)(0.6)=0.48$

c. $f(0)=\binom{2}{0}(0.4)^0(0.6)^2=\frac{2!}{0!2!}(1)(0.36)=0.36$

d. $f(2)=\binom{2}{2}(0.4)^2(0.6)^0=\frac{2!}{2!0!}(0.16)(0.1)=0.16$

e. $P(x\geqslant 1)=f(1)+f(2)=0.48+0.16=0.64$

f. $E(x)=np=2(0.4)=0.8$

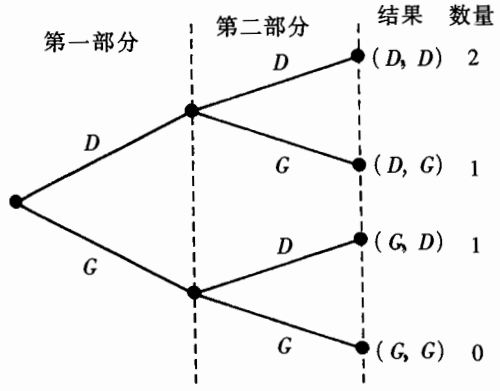
$\mathrm{Var}(x)=np(1-p)=2(0.4)(0.6)=0.48$

$\sigma=\sqrt{0.48}=0.6928$

30. a. 每次试验中发现次品的概率必须是 0.03；试验必须是相互独立的

b. 令 D = 次品

G = 无次品



c. 两种结果正好发现一件次品

d. $P(\text{无次品})=(0.97)(0.97)=0.9409$

$P(1\text{ 件次品})=2(0.03)(0.97)=0.0582$

$P(2\text{ 件次品})=(0.03)(0.03)=0.0009$

39. a. $f(x)=\frac{2^xe^{-2}}{x!}$

b. $\mu=6$ 在 3 个时段内

c. $f(x)=\frac{6^xe^{-6}}{x!}$

d. $f(2)=\frac{2^2e^{-2}}{2!}=\frac{4(0.1353)}{2}=0.2706$

e. $f(6)=\frac{6^6e^{-6}}{6!}=0.1606$

f. $f(5)=\frac{4^5e^{-4}}{5!}=0.1563$

40. a. $\mu=48(5/60)=4$

$f(3)=\frac{4^3e^{-4}}{3!}=\frac{(64)(0.0183)}{6}=0.1952$

b. $\mu=48(15/60)=12$

$f(10)=\frac{12^{10}e^{-12}}{10!}=0.1048$

c. $\mu=48(5/60)=4$; 期望有 4 个电话等待 5 分钟以上

$f(0)=\frac{4^0e^{-4}}{0!}=0.0183$; 无人等待 5 分钟以上的概率是 0.0183

d. $\mu=48(3/60)=2.4$

$f(0)=\frac{2.4^0e^{-2.4}}{0!}=0.0907$; 3 分钟内未被电话打扰的概率是 0.0907

46. a. $f(1)=\frac{\binom{3}{1}\binom{10-3}{4-1}}{\binom{10}{4}}=\frac{\left(\frac{3!}{1!2!}\right)\left(\frac{7!}{3!4!}\right)}{\frac{10!}{4!6!}}=\frac{(3)(35)}{210}=0.50$

b. $f(2)=\frac{\binom{3}{2}\binom{10-3}{2-2}}{\binom{10}{2}}=\frac{(3)(1)}{45}=0.067$

50. $N=60,\ n=10$

a. $r=20,\ x=0$

$$f(0)=\frac{\binom{20}{0}\binom{40}{10}}{\binom{60}{10}}=\frac{(1)\left(\frac{40!}{10!30!}\right)}{\frac{60!}{10!50!}}$$

$$=\left(\frac{40!}{10!30!}\right)\left(\frac{10!50!}{60!}\right)$$

$$=\frac{40\cdot 39\cdot 38\cdot 37\cdot 36\cdot 35\cdot 34\cdot 33\cdot 32\cdot 31}{60\cdot 59\cdot 58\cdot 57\cdot 56\cdot 55\cdot 54\cdot 53\cdot 52\cdot 51}$$

≈ 0.01

b. $r=20,\ x=1$

$$f(1)=\frac{\binom{20}{1}\binom{40}{9}}{\binom{60}{10}}=20\left(\frac{40!}{9!31!}\right)\left(\frac{10!50!}{60!}\right)$$

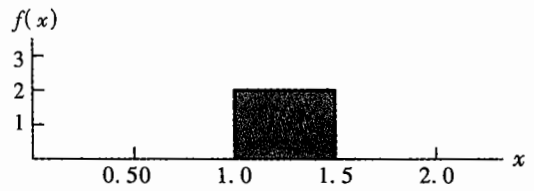
≈ 0.07

c. $1-f(0)-f(1)=1-0.08=0.92$

d. 与洛来自夏威夷的概率相等；在(b)中近似等于0.07

第6章

1. a.

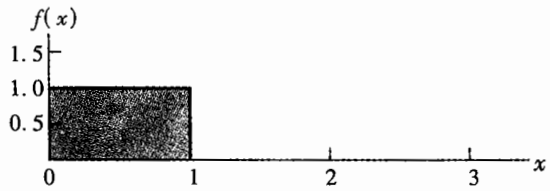


b. $P(x=1.25)=0$；任何单点的概率均为零，因为在任一单点上方的面积为零

c. $P(1.0\leqslant x\leqslant 1.25)=2(0.25)=0.50$

d. $P(1.20<x<1.5)=2(0.30)=0.60$

4. a.



b. $P(0.25<x<0.75)=1(0.50)=0.50$

c. $P(x\leqslant 0.30)=1(0.30)=0.30$

d. $P(x>0.60)=1(0.40)=0.40$

13. a. $0.6879-0.0239=0.6640$

b. $0.8888-0.6985=0.1903$

c. $0.9599-0.8508=0.1091$

15. a. 查表可知面积为 $0.5000-0.2119=0.2881$；

$z=0.80$ 切断 0.2119 右侧的面积；因而，0.2119 左侧的面积，$z=-0.80$

b. 查表可知 $0.9030/2=0.4515$；$z=1.66$

c. 查表可知 $0.2052/2=0.1026$；$z=0.26$

d. 查表可知 0.4948；$z=2.56$

e. 查表可知 0.1915；因为要查的值小于均值，所以 z 值肯定为负；于是 $z=-0.50$

18. a. 查出 $P(x\geqslant 60)$

在 $x=60,\ z=\frac{60-49}{16}=\frac{11}{16}=0.69$

$P(x<60)=0.7549$

$P(x\geqslant 60)=1-P(x<60)=1-0.7549$

$=0.2451$

b. 查出 $P(x\leqslant 30)$

在 $x=30,\ z=\frac{30-49}{16}=-1.19$

$P(x\leqslant 30)=0.5000-0.3830$

$=0.1170$

c. 查出 $z-$值以便 $P(z\geqslant z-\text{值})=0.10$

$z-$值为 1.28，使右侧减少 10%；由此计算相应的 x 值

$$1.28=\frac{x-49}{16}$$

$$x=49+(16)(1.28)$$

$$=69.48$$

于是 10% 的读者花 69.48 分钟或更多的时间阅读《华尔街日报》

27. a. $P(x\leqslant x_0)=1-e^{-x_0/3}$

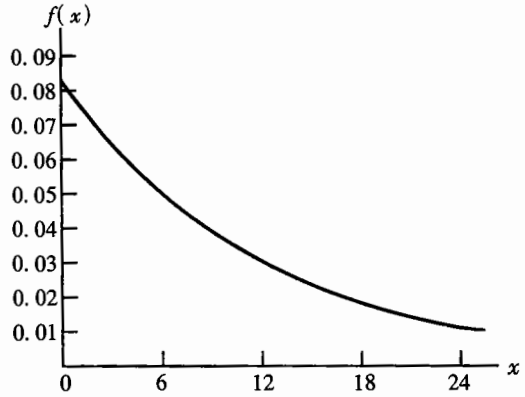
b. $P(x\leqslant 2)=1-e^{-2/3}=1-0.5134=0.4866$

c. $P(x\geqslant 3)=1-P(x\leqslant 3)=1-(1-e^{-3/3})$
$=e^{-1}=0.3679$

d. $P(x\leqslant 5)=1-e^{-5/3}=1-0.1889=0.8111$

e. $P(2\leqslant x\leqslant 5)=P(x\leqslant 5)-P(x\leqslant 2)$
$=0.8111-0.4866=0.3245$

29. a.



b. $P(x\leqslant 12)=1-e^{-12/12}=1-0.3679=0.6321$

c. $P(x\leqslant 6)=1-e^{-6/12}=1-0.6065=0.3935$

d. $P(x\geqslant 30)=1-P(x<30)$
$=1-(1-e^{-30/12})$
$=0.0821$

第 7 章

1. a. AB, AC, AD, AE, BC, BD, BE, CD, CE, DE

b. 每 10 个样本，每个样本被选中的概率是 1/10

c. E 和 C 因为 8 和 0 不用；5 定义为 E；7 不用；5 跳过去因样本中已有；3 定义为 C；2 不需要因为样本容量为 2 的条件已满足

3. 459, 147, 385, 113, 340, 401, 215, 2, 33, 348

11. a. $\bar{x}=\dfrac{\sum x_i}{n}=\dfrac{54}{6}=9$

b. $s=\sqrt{\dfrac{\sum(x_i-\bar{x})^2}{n-1}}$

$\sum(x_i-\bar{x})^2=(-4)^2+(-1)^2+1^2+(-2)^2+1^2+5^2=48$

$s=\sqrt{\dfrac{48}{6-1}}=3.1$

13. a. $\bar{x}=\dfrac{\sum x_i}{n}=\dfrac{465}{5}=93$

b.

x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
94	+1	1
100	+7	49
85	−8	64
94	+1	1
92	−1	1
总计 465	0	116

$s=\sqrt{\dfrac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\dfrac{116}{4}}=5.39$

19. a. 样本服从正态分布，其中：

$E(\bar{x})=\mu=200$

$\sigma_{\bar{x}}=\dfrac{\sigma}{\sqrt{n}}=\dfrac{50}{\sqrt{100}}=5$

对于 +5，$(\bar{x}-\mu)=5$，

$z=\dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\dfrac{5}{5}=1$

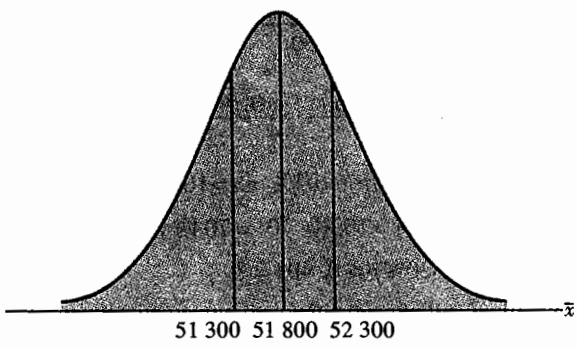
面积 $=0.3413\times 2=0.6826$

b. 对于 ±10，$(\bar{x}-\mu)=10$，

$z=\dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\dfrac{10}{5}=2$

面积 $=0.4772\times 2=0.9544$

25. a.



$\sigma_{\bar{x}}=\dfrac{\sigma}{\sqrt{n}}=\dfrac{4\,000}{\sqrt{60}}=516.40$

$$z=\frac{52\,300-51\,800}{516.40}=+0.97$$

面积 = 0.334 0 × 2 = 0.668 0

b. $\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}=\frac{4\,000}{\sqrt{120}}=365.15$

$$z=\frac{52\,300-51\,800}{365.15}=+1.37$$

面积=0.4147×2=0.8294

34. a. $E(\bar{p})=0.40$

$$\sigma_{\bar{p}}=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{(0.40)(0.60)}{200}}$$

=0.034 6

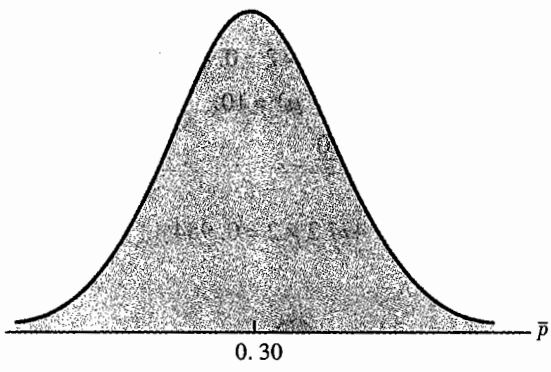
$$z=\frac{\bar{p}-p}{\sigma_{\bar{p}}}=\frac{0.03}{0.0346}=0.87$$

面积 = 0.307 8 × 2 = 0.615 6

b. $z=\frac{\bar{p}-p}{\sigma_{\bar{p}}}=\frac{0.05}{0.034\,6}=1.45$

面积 = 0.426 5 × 2 = 0.853 0

37. a.



$$\sigma_{\bar{p}}=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.30(0.70)}{100}}=0.045\,8$$

服从正态分布,因为 $np=100(0.30)=30$ 和 $n(1-p)=100(0.70)=70$ 两个值都大于 5

b. $P(0.20\leqslant\bar{p}\leqslant0.40)=?$

$$z=\frac{0.40-0.30}{0.0458}=2.18$$

面积 = 0.485 4 × 2 = 0.970 8

c. $P(0.25\leqslant\bar{p}\leqslant0.35)=?$

$$z=\frac{0.35-0.30}{0.045\,8}=1.09$$

面积 = 0.362 1 × 2 = .724 2

第 8 章

2. 用 $\bar{x}\pm z_{\alpha/2}(\sigma/\sqrt{n})$ 具有样本标准差 s 用来估计 σ

a. $32\pm1.645(6/\sqrt{50})$
32 ± 1.4; 30.6—33.4

b. $32\pm1.96(6/\sqrt{50})$
32 ± 1.66; 30.34—33.66

c. $32\pm2.76(6/\sqrt{50})$
32 ± 2.19; 29.81—34.19

5. a. $1.96\sigma/\sqrt{n}=1.96(5.00/\sqrt{49})=1.40$

b. 24.80 ± 1.40; 23.40—26.20

15. a. $\bar{x}=\frac{\sum x_i}{n}=\frac{80}{8}=10$

b. $s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{84}{8-1}}=3.46$

c. 自由度为 7, $t_{0.025}=2.365$

$$\bar{x}\pm t_{0.025}\frac{s}{\sqrt{n}}$$

$$10\pm2.365\frac{3.46}{\sqrt{8}}$$

10 ± 2.90; 7.10 到 12.90

17. 90% 时, $80\pm t_{0.05}(s/\sqrt{n})$ 自由度 = 17

$t_{0.05}=1.740$

$80\pm1.740(10/\sqrt{18})$
80 ± 4.10; 75.90—84.10

95% 时, $80\pm t_{0.025}(10/\sqrt{18})$ 自由度 = 17

$t_{0.025}=2.110$

$80\pm2.110(10/\sqrt{18})$
80 ± 4.97; 75.03—84.97

24. a. σ 的计划值 $=\frac{\text{极差}}{4}=\frac{36}{4}=9$

b. $n=\frac{z_{0.025}^2\sigma^2}{E^2}=\frac{(1.96)^2(9)^2}{(3)^2}=34.57$;用 $n=35$

c. $n=\frac{(1.96)^2(9)^2}{(2)^2}=77.79$;用 $n=78$

25. a. 用 $n=\frac{z_{\alpha/2}^2\sigma^2}{E^2}$

$n=\frac{(1.96)^2(6.82)^2}{(1.5)^2}=79.41$;用 $n=80$

b. $n=\frac{(1.645)^2(6.82)^2}{(2)^2}=31.47$;用 $n=32$

31. a. $\bar{p}=\frac{100}{400}=0.25$

b. $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}=\sqrt{\frac{0.25(0.75)}{400}}=0.021\ 7$

c. $\bar{p}\pm z_{0.025}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

$0.25\pm1.96(0.021\ 7)$

$0.25\pm0.042\ 4$; $0.207\ 6$—$0.292\ 4$

35. a. $\bar{p}=562/814=0.690\ 4$

b. $1.645\sqrt{\frac{0.690\ 4(1-0.690\ 4)}{814}}=0.026\ 7$

c. $0.690\ 4\pm0.026\ 7$; $0.663\ 7$—$0.717\ 1$

39. a. $n=\frac{1.96^2p(1-p)}{E^2}$

$n=\frac{1.96^2(0.33)(0.67)}{(0.03)^2}=943.75$;

用 $n=944$

b. $n=\frac{2.576^2(0.33)(0.67)}{(0.03)^2}=1\ 630.19$;

用 $n=1\ 631$

第 9 章

2. a. H_0: $\mu\leqslant14$

H_a: $\mu>14$

b. 没有新计划能增加销量的证据

c. 支持备选假设 $\mu>14$; 新计划增加销量

5. a. 拒绝 H_0: $\mu\leqslant8.6$ 它是真的

b. 接受 H_0: $\mu\leqslant8.6$ 它是假的

10. a. $z=2.05$

拒绝 H_0 如果 $z>2.05$

b. $z=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{16.5-15}{7/\sqrt{40}}=1.36$

c. 当 $z=1.36$ 时,面积 $=0.4131$

p – 值 $=0.500\ 0-0.413\ 1=0.086\ 9$

d. 不拒绝 H_0

13. a. H_0: $\mu\geqslant105\ 6$

H_a: $\mu<105\ 6$

b. 如果 $z<-1.64\ 5$ 拒绝 H_0

c. $z=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{910-1\ 056}{1\ 600/\sqrt{400}}=-1.83$

d. 拒绝 H_0;得出 $\mu<1\ 056$

e. p – 值 $=0.500\ 0-0.466\ 4=-0.033\ 6$

22. a. 如果 $z<-2.33$ 或 $z>2.33$,拒绝 H_0

b. $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{14.2-15}{5/\sqrt{50}}=1.13$

c. p – 值 $=2(0.500\ 0-0.370\ 8)=0.258\ 4$

d. 不拒绝 H_0

25. a. 如果 $z<-1.96$ 或 $z>1.96$,拒绝 H_0

b. $z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{38.5-39.2}{4.8/\sqrt{112}}=-1.54$

c. 不拒绝 H_0

d. p – 值 $=2(0.500\ 0-0.438\ 2)=0.123\ 6$

34. a. $\bar{x}=\frac{\sum x_i}{n}=\frac{108}{6}=18$

b. $s=\sqrt{\frac{\sum(x_i-\bar{x})}{n-1}}=\sqrt{\frac{10}{6-1}}=1.41$

c. 拒绝 H_0 如果 $t<-2.571$ 或 $t>2.571$

d. $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{18-20}{1.41/\sqrt{6}}=-3.47$

e. 拒绝 H_0; 得出 H_a 或是真的结论

37. a. H_0: $\mu=3.00$

H_a: $\mu\neq3.00$

b. 拒绝 H_0 如果 $t<-2.262$ 或如果 $t>2.262$

c. $\bar{x}=\sum x_i/n=\frac{28}{10}=2.80$

d. $s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}=0.70$

e. $t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{2.80-3.00}{0.70/\sqrt{10}}=-0.90$

f. 不拒绝 H_0

g. $t_{0.10}=1.383$; p - 值大于 $2(0.10)=0.20$

44. a. 拒绝 H_0; 如果 $z<-1.96$ 或 $z>1.96$

b. $\sigma_{\bar{p}}=\sqrt{\frac{0.20(0.80)}{400}}=0.02$

$$z=\frac{\bar{p}-p}{\sigma_{\bar{p}}}=\frac{0.175-0.20}{0.20}=-1.25$$

c. p - 值 $=2(0.5000-0.3944)=0.2112$

d. 不拒绝 H_0

47. a. 拒绝 H_0; 如果 $z<-1.645$

b. $\bar{p}=52/100=0.52$

$$\sigma_{\bar{p}}=\sqrt{\frac{p(1-p)}{n}}=\sqrt{\frac{0.64(1-0.64)}{100}}$$

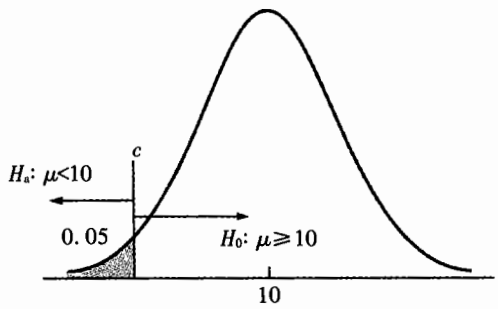
$$=0.0480$$

$$z=\frac{\bar{p}-p}{\sigma_{\bar{p}}}=\frac{0.52-0.64}{0.0480}=-2.50$$

c. 拒绝 H_0; 得出小于 64% 是一致的结论

d. p - 值 $=0.5000-0.4938=0.0062$

56.



$c=10-1.645(5/\sqrt{120})=9.25$

如果$\bar{x}<9.25$, 拒绝 H_0

a. 当 $\mu=9$,

$$z=\frac{9.25-9}{5/\sqrt{120}}=0.55$$

$P(H_0)=(0.5000-0.2088)=0.2912$

b. 第二类错误

c. 当 $\mu=8$ 时,

$$z=\frac{9.25-8}{5/\sqrt{120}}=2.74$$

$\beta=(0.5000-0.4969)=0.0031$

59. a. H_0: $\mu\geq 25$

H_a: $\mu<25$

若 $z<-2.05$, 拒绝 H_0

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{\bar{x}-25}{3/\sqrt{30}}=-2.05$$

解得 $\bar{x}=23.88$

决策规划: 接受 H_0 若 $\bar{x}\geqslant 23.88$

拒绝 H_0 若 $\bar{x}<23.88$

b. 对 $\mu=23$,

$$z=\frac{23.88-23}{3/\sqrt{30}}=1.61$$

$\beta=0.5000-0.4463=0.0537$

c. 对 $\mu=24$,

$$z=\frac{23.88-24}{3/\sqrt{30}}=-0.22$$

$\beta=0.5000-0.0871=0.5871$

d. 这种场合不可能发生第二类错误; 注意当 $\mu=25.5$ 时, H_0 是真: 第二类错误只可能在 H_0 是假时发生

64. $n=\frac{(z_\alpha-z_\beta)\sigma^2}{(\mu_0-\mu_a)^2}=\frac{(1.645+1.28)^2(5)^2}{(10-9)^2}=$

214

67. 在 $\mu_0=400$, $\alpha=0.02$; $z_{0.02}=2.05$

在 $\mu_a=385$, $\beta=0.10$; $z_{0.10}=1.28$

具有 $\sigma+30$,

$$n=\frac{(z_\alpha+z_\beta)\sigma^2}{(\mu_0-\mu_a)^2}=\frac{(2.05+1.28)^2(30)^2}{(400-385)^2}$$

$=44.4$ 或 45

第 10 章

1. a. $\bar{x}_1-\bar{x}_2=13.6-11.6=2$

b. $s_{\bar{x}_1-\bar{x}_2}=\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}=\sqrt{\dfrac{(2.2)^2}{50}+\dfrac{3^2}{35}}=0.595$

$2\pm1.645(0.595)$

2 ± 0.98 或 1.02—2.98

c. $2\pm1.96(0.595)$

2 ± 1.17, 或 0.83—3.17

8. a. $\bar{x}_1-\bar{x}_2=15\,700-14\,500=1\,200$

b. 合并方差

$$s^2=\frac{7(700)^2+11(850)^2}{18}=632\,083$$

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{632\,083\left(\frac{1}{8}+\frac{1}{12}\right)}=362.88$$

自由度为 18　$t_{0.025}=2.101$,

$1\,200\pm2.101(362.88)$

$1\,200\pm762$, 或 438—1 962

c. 两个总体是正态分布且方差相等

11. a. $s_{\bar{x}_1-\bar{x}_2}=\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}=\sqrt{\dfrac{(5.2)^2}{40}+\dfrac{6^2}{50}}=1.18$

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_{\bar{x}_1-\bar{x}_2}}$$

$$=\frac{(25.2-22.8)}{1.18}=2.03$$

若 $z>1.645$, 那么拒绝 H_0; 得出 H_a 是真和 $\mu_1>\mu_2$

b. p – 值 $=0.500\,0-0.478\,8=0.021\,2$

15. H_0: $\mu_1-\mu_2=0$

H_a: $\mu_1-\mu_2\neq0$

若 $z<-1.96$ 或若 $z>1.96$, 拒绝 H_0

$$z=\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}=\frac{40-35}{\sqrt{(9)^2/36+(10)^2/49}}$$

$=2.41$

p – 值 $=2(0.500\,0-0.492\,0)=0.016\,0$

拒绝 H_0; 两个百货公司的顾客的平均年龄不同

21. a. 1, 2, 0, 0, 2

b. $\bar{d}=\dfrac{\sum d_i}{n}=\dfrac{5}{5}=1$

c. $s_d=\sqrt{\dfrac{\sum(d_i-\bar{d})^2}{n-1}}=\sqrt{\dfrac{4}{5-1}}=1$

d. 自由度为 4, $t_{0.05}=2.132$; 拒绝 H_0, 若

$t>2.132$

$$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}=\frac{1-0}{1/\sqrt{5}}=2.24$$

拒绝 H_0; 得出 $\mu_d>0$

23. d = 之后比例 – 之前比例

H_0: $\mu_d\leqslant0$

H_a: $\mu_d>0$

自由度为 7, 拒绝 H_0 若 $t>1.895$; 当 $\bar{d}=0.63$ 和 $s_d=1.302\,5$ 时,

$$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}=\frac{0.63-0}{1.302\,5/\sqrt{8}}=1.36$$

不拒绝 H_0; 我们能得出看广告能提高潜在购买力的结论

31. a. $\bar{p}=\dfrac{n_1\bar{p}_1+n_2\bar{p}_2}{n_1+n_2}=\dfrac{200(0.22)+300(0.16)}{200+300}$

$=0.184$

$$s_{\bar{p}_1-\bar{p}_2}=\sqrt{(0.184)(0.816)\left(\frac{1}{200}+\frac{1}{300}\right)}$$

$=0.035\,4$

若 $z>1.645$, 拒绝 H_0

$$z=\frac{(0.22-0.16)-0}{0.035\,4}=1.69$$

拒绝 H_0

b. p – 值 $=(0.500\,0-0.454\,5)=0.045\,5$

第 11 章

2. $s^2=25$

a. 自由度为 19 $\chi^2_{0.05}=30.143\,5$ 和 $\chi^2_{0.95}=10.117\,0$

$$\frac{19(25)}{30.143\,5}\leqslant\sigma^2\leqslant\frac{19(25)}{10.117\,0}$$

$$15.76\leqslant\sigma^2\leqslant46.95$$

b. 自由度为 19 $\chi^2_{0.025}=32.852\,3$ 和 $\chi^2_{0.975}=$

8.906 55

$$\frac{19(25)}{32.852\ 3} \leqslant \sigma^2 \leqslant \frac{19(25)}{8.906\ 55}$$

$$14.46 \leqslant \sigma^2 \leqslant 53.33$$

c. $3.8 \leqslant \sigma \leqslant 7.3$

9. H_0: $\sigma^2 \leqslant 0.000\ 4$

H_a: $\sigma^2 > 0.000\ 4$

$n = 30$

$\chi^2_{0.05} = 42.556\ 9$(29 自由度)

$$\chi^2 = \frac{(29)(0.000\ 5)}{0.000\ 4} = 36.25$$

不拒绝 H_0:细致加工产品未超过标准

15. 我们建议分子中用较大的样本方差;具有 $\alpha = 0.05$, $F_{0.025,20,24} = 2.33$: 拒绝 H_0 若 $F > 2.33$

$$F = \frac{8.2}{4.0} = 2.05;\ \text{不拒绝 } H_0$$

或者,我们采用左侧 F 值,

$$F_{0.025,20,24} = \frac{1}{F_{0.025,24,20}} = \frac{1}{2.46} = 0.41$$

$$F = \frac{4.0}{8.2} = 0.49$$

$F > 0.41$; 不拒绝 H_0

17. a. 令:σ_1^2 = 修理成本方差(4 年的汽车)

σ_2^2 = 修理成本方差(2 年的汽车)

H_0: $\sigma_1^2 \leqslant \sigma_2^2$

H_a: $\sigma_1^2 > \sigma_2^2$

b. $s_1^2 = (170)^2 = 28\ 900$

$s_2^2 = (100)^2 = 10\ 000$

$$F = \frac{s_1^2}{s_2^2} = \frac{28\ 900}{10\ 000} = 2.89$$

$F_{0.01,24,24} = 2.66$

拒绝 H_0;得出 4 年的汽车的年平均修理费高于 2 年汽车的平均修理费的结论;得出这一结论是由于较旧的汽车更有可能花昂贵的修理费,因而导致更大的年均修理费方差

第 12 章

1. 期望频数:$e_1 = 200(0.40) = 80$

$e_2 = 200(0.40) = 80$

$e_3 = 200(0.20) = 40$

实际频数:$f_1 = 60$, $f_2 = 120$, $f_3 = 20$

$$\chi^2 = \frac{(60-80)^2}{80} + \frac{(120-80)^2}{80} + \frac{(20-40)^2}{40}$$

$$= \frac{400}{80} + \frac{1\ 600}{80} + \frac{400}{40}$$

$$= 5 + 20 + 10 = 35$$

$\chi^2_{0.01} = 9.210\ 34$, 具有 $k - 1 = 3 - 1 = 2$ 自由度因为 $\chi^2 = 35 > 9.210\ 34$, 拒绝原假设;而总体比例与原假设不同

3. H_0: $p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$

H_a: 比例不是

$p_{ABC} = 0.29$, $p_{CBS} = 0.28$, $p_{NBC} = 0.25$, $p_{IND} = 0.18$

期望频数:$300(0.29) = 87$ $300(0.28) = 84$

$300(0.25) = 75$ $300(0.18) = 54$

$e_1 = 87$, $e_2 = 84$, $e_3 = 75$, $e_4 = 54$

实际频数:$f_1 = 95$, $f_2 = 70$, $f_3 = 89$, $f_4 = 46$

$\chi^2_{0.05} = 7.81$(3 自由度)

$$\chi^2 = \frac{(95-87)^2}{87} + \frac{(70-84)^2}{84} + \frac{(89-75)^2}{75}$$

$$+ \frac{(46-54)^2}{54} = 6.87$$

不拒绝 H_0;在观众比例中无显著变化

9. H_0: 列变量独立于行变量

H_a: 列变量与行变量不独立

期望频数:

	A	B	C
P	28.5	39.9	45.6
Q	21.5	30.1	34.4

$$\chi^2 = \frac{(20-28.5)^2}{28.5} + \frac{(44-39.9)^2}{39.9} + \frac{(50-46.5)^2}{45.6}$$

$$+\frac{(30-21.5)^2}{21.5}+\frac{(26-30.1)^2}{30.1}+\frac{(30-34.4)^2}{34.4}$$

$=7.86$

$\chi^2_{0.025}=7.377\,76$, $(2-1)(3-1)=2$ 自由度

因为 $\chi^2=7.86>7.377\,76$, 拒绝 H_0; 即得出行变量与列变量不独立的结论

11. H_0: 购票种类独立于航班类别

H_a: 购票种类与航班类别不独立

期望频数:

$e_{11}=35.59$　　$e_{12}=15.41$

$e_{21}=150.73$　　$e_{22}=65.27$

$e_{31}=455.68$　　$e_{32}=197.32$

机票	航班	观测频数 (f_i)	期望频数 (e_i)	$(f_i-e_i)^2/e_i$
一等舱	国内	29	35.59	1.22
一等舱	国际	22	15.41	2.82
商务舱	国内	95	150.73	20.61
商务舱	国际	121	65.27	47.59
全价经济舱	国内	518	455.68	8.52
全价经济舱	国际	135	197.32	19.68
总计		920		100.43

$\chi^2_{0.05}=5.99$ 具有 $(3-1)(2-1)=2$ 自由度

因为 $100.43>5.99$ 拒绝 H_0; 机票类型与航班类别不独立

20. 首先利用样本数据估计(样本容量 = 120)

$$\mu=\frac{0(39)+1(30)+2(30)+3(18)+4(3)}{120}$$

$$=\frac{156}{120}=1.3$$

因此, 我们用 $\mu=1.3$ 的泊松概率计算期望频数

x	观测频数	泊松概率	期望频数	差异 (f_i-e_i)
0	39	0.272 5	32.700	6.300
1	30	0.354 3	42.516	−12.516
2	30	0.230 3	27.636	2.364
3	18	0.099 8	11.976	6.024
≥4	3	0.043 0	5.160	−2.160

$$\chi^2=\frac{(6.300)^2}{32.700}+\frac{(-12.516)^2}{42.516}+\frac{(2.364)^2}{27.636}$$

$$+\frac{(6.024)^2}{11.976}+\frac{(-2.160)^2}{5.160}=9.034\,8$$

$\chi^2_{0.05}=7.814\,73$ 具有 $5-1-1=3$ 的自由度

因为 $\chi^2=9.034\,8>7.814\,73$, 拒绝 H_0; 即得出数据不服从泊松概率分布的结论。

21. $n=30$ 我们将按 $16\frac{2}{3}\%$ 的概率分为 6 组; 每组具有

$\bar{x}=22.80$, $s=6.266\,5$

分成 6 个区间的 z 值, 每个具有概率是 0.166 7, −0.98, −0.43, 0, 0.43, 0.98。

z	x 的截止值
−0.98	$22.8-0.98(6.266\,5)=16.66$
−0.43	$22.8-0.43(6.266\,5)=20.11$
0	$22.8+0.00(6.266\,5)=22.80$
0.43	$22.8+0.43(6.266\,5)=25.49$
0.98	$22.8+0.98(6.266\,5)=28.94$

区间	观测频数	期望频数	差异
小于 16.66	3	5	−2
16.66—20.11	7	5	2
20.11—22.80	5	5	0
22.80—25.49	7	5	2
25.49—28.94	3	5	−2
28.94 及以上	5	5	0

$$\chi^2=\frac{(-2)^2}{5}+\frac{(2)^2}{5}+\frac{(0)^2}{5}+\frac{(2)^2}{5}+\frac{(-2)^2}{5}$$

$$+\frac{(0)^2}{5}=\frac{16}{5}=3.20$$

$\chi^2_{0.025}=9.348\,40$ 具有 $6-2-1=3$ 自由度

因为 $\chi^2=3.20\leq 9.348\,40$, 不拒绝 H_0;

数据来自正态总体的声明不能被拒绝

第 13 章

1. a. $\bar{\bar{x}} = (30 + 45 + 36)/3 = 37$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$$= 5(30-37)^2 + 5(45-37)^2 + 5(36-37)^2$$

$$= 570$$

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{570}{2} = 285$$

b. $$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

$$= 4(6) + 4(4) + 4(6.5) = 66$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{66}{15-3} = 5.5$$

c. $$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{285}{5.5} = 51.82$$

$F_{0.05} = 3.89$(分子自由度为 2 和分母自由度为 12)

因为 $F = 51.82 > F_{0.05} = 3.89$, 我们拒绝意味着三个总体的均值相等的原假设

d.

方差来源	平方和	自由度	均方差	F
Treatments	570	2	285	51.82
Error	66	12	5.5	
总计	636	14		

6.

	生产商 1	生产商 2	生产商 3
样本均值	23	28	21
样本方差	6.67	4.67	3.33

$\bar{\bar{x}} = (23 + 28 + 21)/3 = 24$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$$= 4(23-24)^2 + 4(28-24)^2 + 4(21-24)^2$$

$$= 104$$

$$\text{MSTR} = \frac{\text{SSTR}}{k-1} = \frac{104}{2} = 52$$

$$\text{SSE} = \sum_{j=1}^{k} (n_j - 1)s_j^2$$

$$= 3(6.67) + 3(4.67) + 3(3.33) = 44.01$$

$$\text{MSE} = \frac{\text{SSE}}{n_T - k} = \frac{44.01}{12-3} = 4.89$$

$$F = \frac{\text{MSTR}}{\text{MSE}} = \frac{52}{4.89} = 10.63$$

$F_{0.05} = 4.26$(2 分子自由度和 9 分母自由度)

因为 $F = 10.63 > F_{0.05} = 4.26$, 我们拒绝对每一厂商来说混合一批原材料所需时间相等这一原假设

11. a. $$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$= t_{0.025}\sqrt{5.5\left(\frac{1}{5} + \frac{1}{5}\right)}$$

$$= 2.179\sqrt{2.2} = 3.23$$

$|\bar{x}_1 - \bar{x}_2| = |30 - 45| = 15 > \text{LSD}$; 显著差异

$|\bar{x}_1 - \bar{x}_3| = |30 - 36| = 6 > \text{LSD}$; 显著差异

$|\bar{x}_2 - \bar{x}_3| = |45 - 36| = 9 > \text{LSD}$; 显著差异

b. $$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$(30 - 45) \pm 2.179\sqrt{5.5\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$-15 \pm 3.23 = -18.23 \text{ 到 } -11.77$$

13. $$\text{LSD} = t_{\alpha/2}\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}$$

$$= t_{0.025}\sqrt{4.89\left(\frac{1}{4} + \frac{1}{4}\right)}$$

$$= 2.262\sqrt{2.45} = 3.54$$

因为 $|\bar{x}_1 - \bar{x}_3| = |23 - 21| = 2 < 3.54$, 在总体 1 和 3 的均值之间不存在任何显著差异

14. $\bar{x}_1 - \bar{x}_2 \pm \text{LSD}$

$23 - 28 \pm 3.54$

$-5 \pm 3.54 = -8.54$ 到 -1.46

19. a. $\bar{\bar{x}} = (156 + 142 + 134)/3 = 144$

$$\text{SSTR} = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

$=6(156-144)^2+6(142-144)^2+6(134-144)^2$

$=1\,488$

b. $\text{MSTR}=\dfrac{\text{SSTR}}{k-1}=\dfrac{1\,488}{2}=744$

c. $s_1^2=164.4,\ s_2^2=131.2,\ s_3^2=110.4$

$$\text{SSE}=\sum_{j=1}^{k}(n_j-1)s_j^2$$

$=5(164.4)+5(131.2)+5(110.4)$

$=2030$

d. $\text{MSE}=\dfrac{\text{SSE}}{n_T-k}=\dfrac{2\,030}{18-3}=135.3$

e. $F=\dfrac{\text{MSTR}}{\text{MSE}}=\dfrac{744}{135.3}=5.50$

$F_{0.05}=3.68$(分子自由度 2 和分母自由度 15)

因为 $F=5.50>F_{0.05}=3.68$, 我们拒绝三个处理均值是相等的这一原假设

34. 处理均值

$\bar{x}_{.1}=13.6,\ \bar{x}_{.2}=11.0,\ \bar{x}_{.3}=10.6$

组内均值

$\bar{x}_{1.}=9,\ \bar{x}_{2.}=7.67,\ \bar{x}_{3.}=15.67,\ \bar{x}_{4.}=18.67,\ \bar{x}_{5.}=7.67$

总均值

$\bar{\bar{x}}=176/15=11.73$

步骤 1

$$\text{SST}=\sum_i\sum_j(x_{ij}-\bar{\bar{x}})^2$$

$=(10-11.73)^2+(9-11.73)^2+\cdots+(8-11.73)^2$

$=354.93$

步骤 2

$$\text{SSTR}=b\sum_j(\bar{x}_{.j}-\bar{\bar{x}})^2$$

$=5[(13.6-11.73)^2+(11.0-11.73)^2+(10.6-11.73)^2]=26.53$

步骤 3

$$\text{SSBL}=k\sum_j(\bar{x}_{i.}-\bar{\bar{x}})^2$$

$=3[(9-11.73)^2+(7.67-11.73)^2+(15.67-11.73)^2+(18.67-11.73)^2+(7.67-11.73)^2]=312.32$

步骤 4

$\text{SSE}=\text{SST}-\text{SSTR}-\text{SSBL}$

$=354.93-26.53-312.32=16.08$

方差来源	平方和	自由度	均方差	F
处理	26.53	2	13.27	6.60
区组	312.32	4	78.08	
误差	16.08	8	2.01	
总计	354.93	14		

$F_{0.05}=4.46$(分子自由度 2 和分母自由度 8)

因为 $F=6.60>F_{0.05}=4.46$, 我们拒绝三个处理均值相等的原假设

41. 见表 E13.41

步骤 1

$$\text{SST}=\sum_i\sum_j\sum_k(x_{ijk}-\bar{\bar{x}})^2$$

$=(135-111)^2+(165-111)^2+\cdots+(136-111)^2=9\,028$

表 E13.41

		因子 B			
		水平 1	水平 2	水平 3	因子 A 均值
因子 A	水平 1	$\bar{x}_{11}=150$	$\bar{x}_{12}=78$	$\bar{x}_{13}=84$	$\bar{x}_{1.}=104$
	水平 2	$\bar{x}_{21}=110$	$\bar{x}_{22}=116$	$\bar{x}_{23}=128$	$\bar{x}_{2.}=118$
因子 B 均值		$\bar{x}_{.1}=130$	$\bar{x}_{.2}=97$	$\bar{x}_{.3}=106$	$\bar{\bar{x}}=111$

步骤 2

$$SSA = br \sum_i (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$

$$= 3(2)[(104-111)^2 + (118-111)^2] = 588$$

步骤 3

$$SSB = ar \sum_j (\bar{x}_{\cdot j} - \bar{\bar{x}})^2$$

$$= 2(2)[(130-111)^2 + (97-111)^2 + (106-111)^2] = 2328$$

步骤 4

$$SSAB = r\sum_i \sum_j (\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

$$= 2[(150-104-130+111)^2 + (78-104-97+111)^2 + \cdots + (128-118-106+111)^2] = 4\,392$$

步骤 5

$$SSE = SST - SSA - SSB - SSAB$$

$$= 9\,028 - 588 - 2\,328 - 4\,392 = 1\,720$$

方差来源	平方和	自由度	均方差	F
因子 A	588	1	588	2.05
因子 B	2 328	2	1 164	4.06
交互作用	4 392	2	2 196	7.66
误差	1 720	6	286.67	
总计	9 028	11		

$F_{0.05} = 5.99$(分子自由度 1 和分母自由度 6)

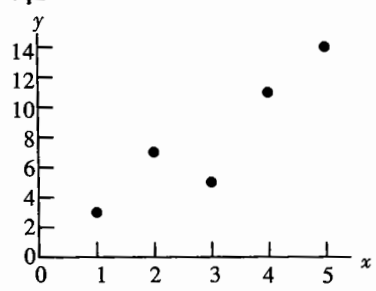
$F_{0.05} = 5.14$(分子自由度 2 和分母自由度 6)

因为 $F = 2.05 < F_{0.05} = 5.99$, 因子 A 不显著;

因为 $F = 4.06 < F_{0.05} = 5.14$, 因子 B 不显著;

因为 $F = 7.66 > F_{0.05} = 5.14$, 相互作用显著。

第 14 章



1. a.

b. 在 x 和 y 之间存在线性相关关系

c. 可画出多条不同的直线作为 x 和 y 之间相关关系的线性近似;在(d)中我们确定一条按最小二乘标准最能代表相关关系的一条直线方程

d. 共需计算出斜率和 y 的截距:

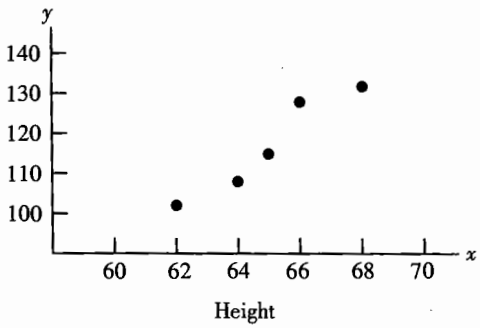
$\sum x_i = 15$, $\sum y_i = 40$, $\sum (x_i - \bar{x})(y_i - \bar{y}) = 26$, $\sum (x_i - \bar{x})^2 = 10$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{26}{10} = 2.6$$

$$b_0 = \bar{y} - b_1 \bar{x} = 8 - (2.6)(3) = 0.2$$

e. $\hat{y} = 0.2 - 2.6x$

$\hat{y} = 0.2 + 2.6x = 0.2 + 2.6(4) = 10.6$



4. a.

b. 它表明变量之间可能存在线性相关关系

c. 可画出多条不同的直线作为 x 和 y 之间关系的线性近似;在(d)中我们将按最小二乘法定出“最”能代表它们之间关系的一条直线方程

d. 共需计算出斜率和 y 的截距:

$\sum x_i = 325$, $\sum y_i = 585$, $\sum (x_i - \bar{x})(y_i - \bar{y}) = 110$, $\sum (x_i - \bar{x})^2 = 20$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{110}{20} = 5.5$$

$$b_0 = \bar{y} - b_1 \bar{x} = 117 - (5.5)(65) = -240.5$$

e. $\hat{y} = -240.5 + 5.5x$

$\hat{y} = -240.5 + 5.5(63) = 106$

重量的估计是 106 磅

x_i	y_i	$\hat{y}_i$	$y_i-\hat{y}_i$	$(y_i-\hat{y}_i)^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$
1	3	2.8	0.2	0.04	−5	25
2	7	5.4	1.6	2.56	−1	1
3	5	8.0	−3.0	9.00	−3	9
4	11	10.6	0.4	0.16	3	9
5	14	13.2	0.8	0.64	6	36
				SSE = 12.40		SST = 80

SSR = SST − SSE = 80 − 12.4 = 67.6

15. a. $\hat{y}_i = 0.2 + 2.6x_i$ 和 $\bar{y} = 8$

b. $r^2 = \dfrac{\text{SSR}}{\text{SST}} = \dfrac{67.6}{80} = 0.845$

最小二乘直线的拟合度好；变量 y 中的 84.5% 可由最小二乘直线解释

c. $r = \sqrt{0.845} = +0.919\,2$

18. a. 估计回归方程和因变量的均值为：

$\hat{y} = 1\,790.5 + 581.1x$, $\bar{y} = 3\,650$

由误差引起的平方和和总平方和：

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 85\,135.14$

$\text{SST} = \sum(y_i - \bar{y})^2 = 335\,000$

于是，SSR = SST − SSE

= 335 000 − 85 135.14 = 249 864.86

b. $r^2 = \dfrac{\text{SSR}}{\text{SST}} = \dfrac{249\,864.86}{335\,000} = 0.746$

最小二乘直线计算出总平方的 74.6%

c. $r = \sqrt{0.746} = +0.863\,7$

23. a. $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-2} = \dfrac{12.4}{3} = 4.133$

b. $s = \sqrt{\text{MSE}} = \sqrt{4.133} = 2.033$

c. $\sum(x_i - \bar{x})^2 = 10$

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{2.033}{\sqrt{10}} = 0.643$$

d. $t = \dfrac{b_1 - \beta_1}{s_{b_1}} = \dfrac{2.6-0}{0.643} = 4.04$

$t_{0.025} = 3.182$(3 自由度)

因为 $t = 4.04 > t_{0.05} = 3.182$, 拒绝 H_0: $\beta_1 = 0$

e. $\text{MSR} = \dfrac{\text{SSR}}{1} = 67.6$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{67.6}{4.133} = 16.36$$

$F_{0.05} = 10.13$(分子自由度 1 和分母自由度 3)

因为 $F = 16.36 > F_{0.05} = 10.13$,

拒绝 H_0: $\beta_1 = 0$

方差来源	平方和	自由度	均方差	F
回归	67.6	1	67.6	16.36
误差	12.4	3	4.133	
总计	80	4		

26. a. $s^2 = \text{MSE} = \dfrac{\text{SSE}}{n-2} = \dfrac{85\,135.14}{4} = 21\,283.79$

$s = \sqrt{\text{MSE}} = \sqrt{21\,283.79} = 145.89$

$\sum(x_i - \bar{x})^2 = 0.74$

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{145.89}{\sqrt{0.74}} = 169.59$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{581.08-0}{169.59} = 3.43$$

$t_{0.025} = 2.776$(4 自由度)

因为 $t = 3.43 > t_{0.025} = 2.776$, 拒绝 H_0: $\beta_1 = 0$

b. $\text{MSR} = \dfrac{\text{SSR}}{1} = \dfrac{249\,864.86}{1} = 249\,864.86$

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{249\,864.86}{21\,283.79} = 11.74$$

$F_{0.05} = 7.71$(分子自由度 1 和分母自由度 4)

因为 $F = 11.74 > F_{0.05} = 7.71$, 拒绝 H_0: $\beta_1 = 0$

c.

方差来源	平方和	自由度	均方差	F
回归	29 864.86	1	29 864.86	11.74
误差	85 135.14	4	21 283.79	
总计	335 000	5		

32. a. $s = 2.033$

$\bar{x} = 3$, $\sum(x_i - \bar{x})^2 = 10$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

$$=2.033\sqrt{\frac{1}{5}+\frac{(4-3)^2}{10}}=1.11$$

b. $\hat{y}=0.2+2.6x=0.2+2.6(4)=10.6$

$\hat{y}_p \pm t_{\alpha/2}s_{\hat{y}_p}$

$10.6\pm3.182(1.11)$

10.6 ± 3.53, 或 7.07—14.13

c. $s_{ind}=s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

$$=2.033\sqrt{1+\frac{1}{5}+\frac{(4-3)^2}{10}}=2.32$$

d. $\hat{y}_p \pm t_{\alpha/2}s_{ind}$

$10.6\pm3.182(2.32)$

10.6 ± 7.38, 或 3.22—17.98

35. a. $s=145.89, \bar{x}=3.2, \sum(x_i-\bar{x})^2=0.74$

$\hat{y}=290.54+581.08x=290.54+581.08(3)$

$=2\,033.78$

$$s_{\hat{y}_p}=s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

$$=145.89\sqrt{\frac{1}{6}+\frac{(3-3.2)^2}{0.74}}=68.54$$

$\hat{y}_p \pm t_{\alpha/2}s_{\hat{y}_p}$

$2\,033.78\pm2.776(68.54)$

$2\,033.78\pm190.27$, 或 1 843.51 美元—2 224.05 美元

b. $\hat{y}=290.54+581.08x=290.54+581.08(3)$

$=2\,033.78$

$$s_{ind}=s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

$$=145.89\sqrt{1+\frac{1}{6}+\frac{(3-3.2)^2}{0.74}}$$

$=161.19$

$\hat{y}_p \pm t_{\alpha/2}s_{ind}$

$2\,033.78\pm2.776(161.19)$

$2\,033.78\pm447.46$, 或 1 586.32 美元—2 481.24 美元

40. a. 9

b. $\hat{y}=20.0+7.21x$

c. 1.362 6

d. $SSE=SST-SSR=51\,984.1-41\,587.3$

$=10\,396.8$

$MSE=10\,396.8/7=1\,485.3$

$$F=\frac{MSR}{MSE}=\frac{41\,587.3}{1\,485.3}=28.0$$

$F_{0.05}=5.59$(分子自由度 1 和分母自由度 7)

因为 $F=28>F_{0.05}=5.59$, 拒绝 $H_0: \beta_1=0$

e. $\hat{y}=20.0+7.21(50)=380.5$, 或 380 500 美元

45. a. $\sum x_i=14, \sum y_i=76, \sum(x_i-\bar{x})(y_i-\bar{y})=200, \sum(x_i-\bar{x})^2=126$

$$b_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{200}{126}=1.587\,3$$

$b_0=\bar{y}-b_1\bar{x}=15.2-(1.587\,3)(14)$

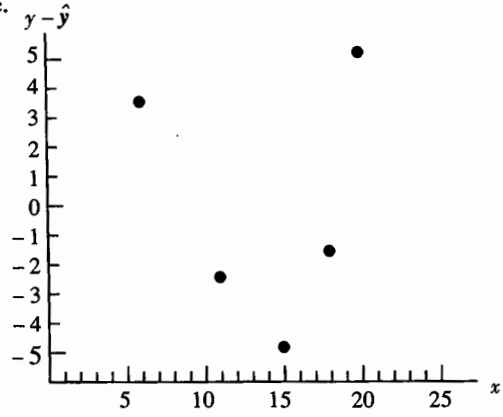
$=-7.022\,2$

$\hat{y}=-7.02+1.59x$

b.

x_i	y_i	$\hat{y}_i$	$y_i-\hat{y}_i$
6	6	2.52	3.48
11	8	10.47	−2.47
15	12	16.83	−4.83
18	20	21.60	−1.60
20	30	24.78	5.22

c.



由于只有 5 个观测值，很难确定假设是否成立；然而，这些点确定表明残差中的曲线关系，进而说明误差项假设不能满足；这些数据的散点图也指出 x 和 y 之间最可能的关系是曲线相关

d. $s^2 = 23.78$

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

$$= \frac{1}{5} + \frac{(x_i - 14)^2}{126}$$

x_i	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	标准残差
6	0.707 9	2.64	3.48	1.32
11	0.271 4	4.16	-2.47	-0.59
15	0.207 9	4.34	-4.83	-1.11
18	0.327 0	4.00	-1.60	-0.40
20	0.485 7	3.50	5.22	1.49

e. 对应 $\hat{y}$ 的标准残差的点与初始的残差的点相同；如(c)中所述观测到的曲线表明对误差项的假设可能不会满足

47. a. 令 x = 广告支出，y = 收入

$\hat{y} = 29.4 + 1.55x$

b. SST = 1 002, SSE = 310.28, SSR = 691.72

$$\text{MSR} = \frac{\text{SSR}}{1} = 691.72$$

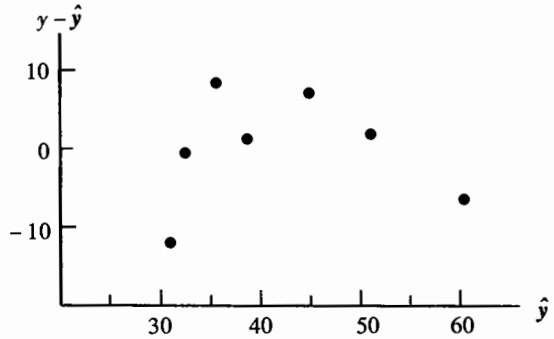
$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{310.28}{5} = 62.055\,4$$

$$\text{F} = \frac{\text{MSR}}{\text{MSE}} = \frac{691.72}{62.0554} = 11.15$$

$F_{0.05} = 6.61$(分子自由度 1 和分母自由度 5)因为 $F = 11.15 > F_{0.05} = 6.61$，可得出两个变量相关的结论

c.

x_i	y_i	$\hat{y}_i = 29.40 + 1.55x_i$	$y_i - \hat{y}_i$
1	19	30.95	-11.95
2	32	32.50	-0.50
4	44	35.60	8.40
6	40	38.70	1.30
10	52	44.90	7.10
14	53	51.10	1.90
20	54	60.40	-6.40

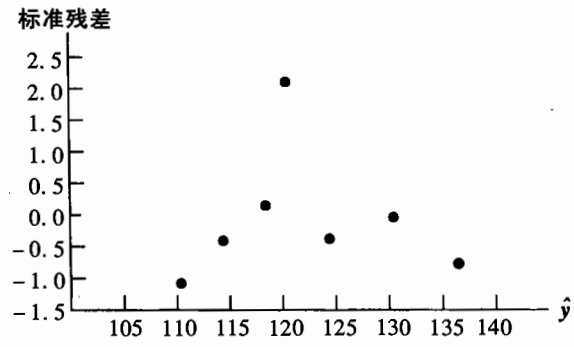


d. 散点图让我们对 x 和 y 之间线性关系的假设产生疑问；在 $\alpha = 0.05$ 水平上相关关系似乎是显著的，超出数据极差进行推断可能是很不确切的

50. a. 利用 Minitab 软件，可得到估计回归方程 $\hat{y} = 66.1 + 0.402\,3x$；一部分 Minitab 的输出结果如图 E14.50 所示；拟合值和标准残差如下：

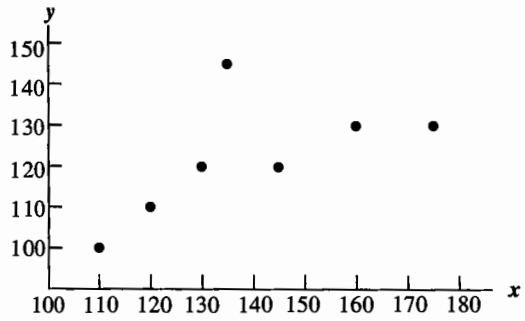
x_i	y_i	$\hat{y}_i$	标准残差
135	145	120.41	2.11
110	100	110.35	-1.08
130	120	118.40	0.14
145	120	124.43	-0.38
175	130	136.50	-0.78
160	130	130.47	-0.04
120	110	114.38	-0.41

b.



标准残差点显示观测值 $x = 135$, $y = 145$ 可能是异常值；注意，这一观测值有一标准残差 2.11

c. 散点图如下：



散点图显示观测值 $x=135$, $y=145$ 可能是异常值;其含义是对简单的线性回归异常值可由散点图识别出

52. a. 一部分 Minitab 的输出结果如图 E14.52 所示

b. Minitab 将观测值 1 定为有大的标准残差的值;因而,我们可视其为异常值

图 E14.50

```
The regression equation is
Y = 66.1 + 0.402 X

Predictor       Coef      Stdev    t-ratio       p
Constant       66.10      32.06       2.06   0.094
X             0.4023     0.2276       1.77   0.137

s = 12.62      R-sq = 38.5%     R-sq(adj) = 26.1%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1       497.2       497.2      3.12   0.137
Error         5       795.7       159.1
Total         6      1292.9

Unusual Observations
Obs.        X         Y        Fit Stdev.Fit Residual   St.Resid
  1       135    145.00     120.42      4.87    24.58      2.11R
```

图 E14.52

```
The regression equation is
Shipment = 4.09 + 0.196 Media$

Predictor       Coef       StDev          T       P
Constant       4.089       2.168       1.89   0.096
Media$       0.19552     0.03635       5.38   0.000

S = 5.044      R-Sq = 78.3%      R-Sq(adj) = 75.6%

Analysis of Variance
Source       DF          SS          MS         F       P
Regression    1      735.84      735.84     28.93   0.000
Error         8      203.51       25.44
Total         9      939.35

Unusual Observations
Obs    Media$   Shipment        Fit  StDev Fit   Residual     St Resid
  1       120      36.30      27.55       3.30       8.75        2.30R

R denotes an observation with a large standardized residual
```

第 15 章

2. a. 估计回归方程是 $\hat{y}=45.06+1.94x_1$

y 的一个估计，当 $x_1=45$ 时是

$\hat{y}=45.06+1.94(45)=132.36$

b. 估计回归方程 $\hat{y}=85.22+4.32x_2$

当 $x_2=15$ 时的估计值是

$\hat{y}=85.22+4.32(15)=150.02$

c. 估计回归方程是

$\hat{y}=-18.37+2.01x_1+4.74x_2$

当 $x_1=45$ 和 $x_2=15$ 时 y 的一个估计是

$\hat{y}=-18.37+2.01(45)+4.74(15)=143.18$

5. a. Minitab 的输出结果如图 E15.5a 所示

b. Minitab 的输出结果如图 E15.5b 所示

c. 在(a)中系数是 1.60，(b)中是 2.29；在(a)中系数是根据一单位电视广告支出的变化对收入进行的估计；而在(b)中系数代表了根据在报纸广告保持不变时，电视广告支出变化一单位所引起收入变化的估计

d. 收入 $=83.2+2.29(3.5)+1.30(1.8)=93.56$ 或 93 560 美元

12. a. $R^2=\dfrac{\text{SSR}}{\text{SST}}=\dfrac{14\,052.2}{15\,182.9}=0.926$

b. $R_a^2=1-(1-R^2)\dfrac{n-1}{n-p-1}$

$=1-(1-0.926)\dfrac{10-1}{10-2-1}=0.905$

图 E15.5a

```
The regression equation is
Revenue = 88.6 + 1.60 TVAdv

Predictor       Coef      Stdev    t-ratio       p
Constant      88.638      1.582      56.02   0.000
TVAdv         1.6039     0.4778       3.36   0.015

s = 1.215      R-sq = 65.3%     R-sq(adj) = 59.5%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1      16.640      16.640     11.27   0.015
Error         6       8.860       1.477
Total         7      25.500
```

图 E15.5b

```
The regression equation is
Revenue = 83.2 + 2.29 TVAdv + 1.30 NewsAdv

Predictor       Coef      Stdev    t-ratio       p
Constant      83.230      1.574      52.88   0.000
TVAdv         2.2902     0.3041       7.53   0.001
NewsAdv       1.3010     0.3207       4.06   0.010

s = 0.6426     R-sq = 91.9%    R-sq(adj) = 88.7%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    2      23.435      11.718     28.38   0.002
Error         5       2.065       0.413
Total         7      25.500
```

c. 是的;在修正于模型中独立变量的数目之后,我们可看到 y 中 90.5% 的变异性已得到解释

15. a. $R^2=\dfrac{\text{SSR}}{\text{SST}}=\dfrac{23.435}{25.5}=0.919$

$$R_a^2=1-(1-R^2)\frac{n-1}{n-p-1}$$

$$=1-(1-0.919)\frac{8-1}{8-2-1}=0.887$$

b. 可按受多元回归结果,因为 R^2 和 R_a^2 两者都显示出:当用到两个相互独立的变量时,解释变量 y 的变异性的百分比在增加

19. a. $\text{MSR}=\dfrac{\text{SSR}}{p}=\dfrac{6\,216.375}{2}=3\,108.188$

$$\text{MSE}=\frac{\text{SSE}}{n-p-1}=\frac{507.75}{10-2-1}=72.536$$

b. $F=\dfrac{\text{MSR}}{\text{MSE}}=\dfrac{3\,108.188}{72.536}=42.85$

$F_{0.05}=4.74$(分子自由度 2 和分母自由度 7)

因为 $F=42.85>F_{0.05}=4.74$, 总模型是显著的

c. $t=\dfrac{b_1}{s_{b_1}}=\dfrac{0.590\,6}{0.081\,3}=7.26$

$t_{0.025}=2.365$(自由度 7)

由于 $t=7.26>t_{0.025}=2.365$, β_1 是显著的

d. $t=\dfrac{b_2}{s_{b_2}}=\dfrac{0.498\,0}{0.056\,7}=8.78$

由于 $t=8.78>t_{0.025}=2.365$, β_2 是显著的

23. a. $F=28.38$

$F_{0.01}=13.27$(分子自由度 2 和分母自由度 1)

因为 $F>F_{0.01}=13.27$, 拒绝 H_0

另一种情况,p - 值是 0.002 导出同数的结论

b. $t=7.53$

$t_{0.025}=2.571$

因为 $t>t_{0.025}=2.571$, β_1 是显著的,x_1 不应从模型中剔除

c. $t=4.06$

$t_{0.025}=2.571$

因为 $t>t_{0.025}=2.571$, β_2 是显著的,x_2 不应从模型中剔除

28. a. 利用 Minitab 软件,95% 置信区间是 132.16 到 154.15

b. 利用 Minitab 软件,95% 预测区间是 111.15 到 175.17

29. a. 见图 E15.5b 中 Minitab 的输出结果

$\hat{y}=83.230+2.290\,2(3.5)+1.301\,0(1.8)$

$=93.588$ 或 93 588 美元

b. 利用 Minitab 软件:92.840—94.335,或 92 840 美元—94 335 美元

c. 利用 Minitab 软件:91.774—95.401,或 91 774 美元—95 401 美元

32. a. $E(y)=\beta_0+\beta_1x_1+\beta_2x_2$

式中 $x_2=\begin{cases}0 & \text{若水平1}\\ 1 & \text{若水平2}\end{cases}$

b. $E(y)=\beta_0+\beta_1x_1+\beta_2(0)=\beta_0+\beta_1x_1$

c. $E(y)=\beta_0+\beta_1x_1+\beta_2(1)=\beta_0+\beta_1x_1+\beta_2$

d. $\beta_2=E(y\mid\text{水平 2})-E(y\mid\text{水平 1})$

β_1 是当 x_2 保持不变 x_1 不变 1 单位时,$E(y)$所发生的变化

34. a. 15 300 美元, 因为 $b_3=15.3$

b. $\hat{y}=10.1-4.2(2)+6.8(8)+15.3(0)$

$=10.1-8.4+54.4$

$=56.1$

销售预测:56 100 美元

c. $\hat{y}=10.1-4.2(1)+6.8(3)+15.3(1)$

$=10.1-4.2+20.4+15.3$

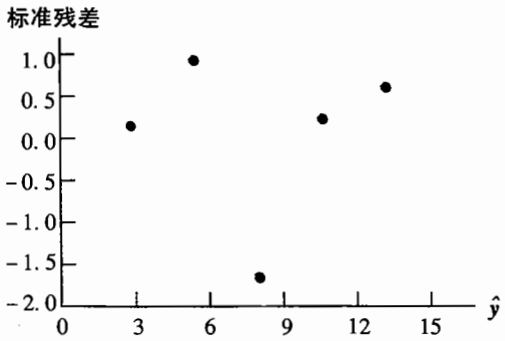
$=41.6$

销售预测:41 600 美元

39. a. Minitab 的输出结果如图 E15.39 所示

b. 利用 Minitab 软件,我们得出下列值:

x_i	y_i	$\hat{y}_i$	标准残差
1	3	2.8	0.16
2	7	5.4	0.94
3	5	8.0	−1.65
4	11	10.6	0.24
5	14	13.2	0.62



点(3,5)不符合其他点的趋势;然而,这一点的标准残差,−1.65,并不大到足以让我们得出该点是异常值这一结论

c. 利用 Minitab,我们得出下列值:

x_i	y_i	按学生分布删除残差
1	3	0.13
2	7	0.92
3	5	−4.42
4	11	0.19
5	14	0.54

$t_{0.025}=4.303(n-p-2=5-1-2=2$ 自由度)因为按学生分布删除异常值,对点(3,5)来说是 −4.42<−4.303,我们就可得出第3个观测值是异常值的结论

图 E15.39

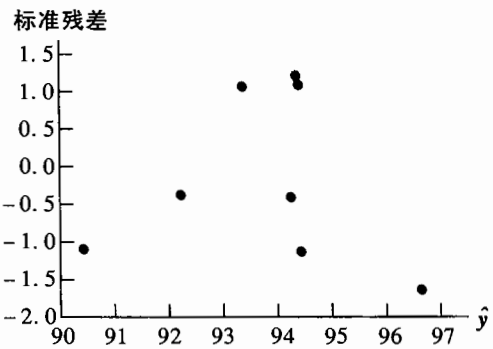
```
The regression equation is
Y = 0.20 + 2.60 X

Predictor       Coef      Stdev    t-ratio       p
Constant       0.200      2.132       0.09   0.931
X             2.6000     0.6429       4.04   0.027

s = 2.033      R-sq = 84.5%     R-sq(adj) = 79.3%

Analysis of Variance
SOURCE       DF          SS          MS        F       p
Regression    1      67.600      67.600    16.35   0.027
Error         3      12.400       4.133
Total         4      80.000
```

41. a. Minitab 的输出结果如图 E15.5b 所示;估计回归方程是

Revenue=83.2+2.29TVAdv+1.30NewsAdv

b. 利用 Minitab,我们得到下列值:

$\hat{y}_i$	标准残差	$\hat{y}_i$	标准残差
96.63	−1.62	94.39	1.10
90.41	−1.08	94.24	−0.40
94.34	1.22	94.42	−1.12
92.21	−0.37	93.35	1.08



由于观测值相对太少;很难确定关于 ε 的任何假设已背离;例如,既可得出这些点不存在

任何关系的结论，也可得出这些点是曲线相关的结论

c. 标准残差值大于 -2 和小于 +2；那么利用这一检验，没有异常值

进一步检测异常值，我们用 Minitab 软件计算出按学生分布删除的残差如下：

观测值	按学生分布删除残差	观测值	按学生分布删除残差
1	-2.11	5	1.13
2	-1.10	6	-0.36
3	1.31	7	-1.16
4	-0.33	8	1.10

$t_{0.025}=2.776(n-p-2=8-2-2=4$ 自由度)

因为没有一个按学生分布删除的残差小于 -2.776 或大于 2.776，我们得出数据中没有异常值的结果论

d. 利用 Minitab，我们得出下列值：

观测值	h_i	D_i
1	0.63	1.52
2	0.65	0.70
3	0.30	0.22
4	0.23	0.01
5	0.26	0.14
6	0.14	0.01
7	0.66	0.81
8	0.13	0.06

临界值是

$$\frac{3(p+1)}{n}=\frac{3(2+1)}{8}=1.125$$

因为没有一个值超过 1.125，我们就可得出没有有影响的值的结论，然而，利用库克距离测度我们看到 $D_1>1$(拇指临界值规则)，那么，可得出第 1 个观测值是有影响的

最终结论：观测值 1 是一个有影响的值

第 16 章

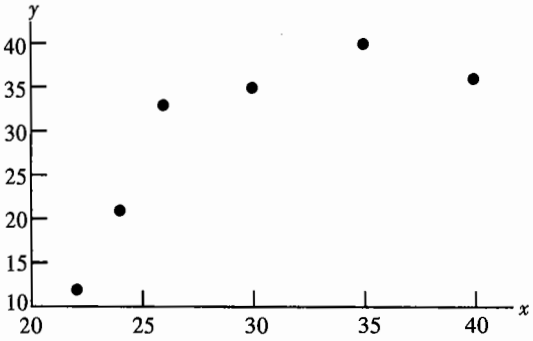
1. a. Minitab 的输出结果如图 E16.1a 所示

b. $F=6.85$ 的 p- 值是 $0.059>\alpha=0.05$，因此相关性不显著

图 E16.1a

```
The regression equation is
Y = - 6.8 + 1.23 X

Predictor       Coef      Stdev    t-ratio       p
Constant       -6.77      14.17      -0.48   0.658
X             1.2296     0.4697       2.62   0.059

s = 7.269       R-sq = 63.1%     R-sq(adj) = 53.9%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    1      362.13      362.13      6.85   0.059
Error         4      211.37       52.84
Total         5      573.50
```

c.



散点图表明曲线相关可能是合适的

d. Minitab 的输出结果如图 E16. 1d 所示

e. 相应于 $F = 25.68$ 的值 p- 值是 $0.013 < \alpha = 0.05$, 因此相关性显著

f. $\hat{y} = -168.88 + 12.187(25) - 0.177\,04(25)^2 = 25.145$

5. a. Minitab 的输出结果如图 E16. 5a 所示

b. 相应于 $F = 73.15$ 的值 p- 值是 $0.003 < \alpha = 0.01$, 因此相关性显著; 应该拒绝 $H_0: \beta_1 = \beta_2 = 0$

c. 见图 E16. 5c

图 E16. 1d

```
The regression equation is
Y = - 169 + 12.2 X - 0.177 XSQ

Predictor       Coef      Stdev    t-ratio       p
Constant     -168.88      39.79      -4.74   0.024
X             12.187      2.663       4.58   0.020
XSQ          -0.17704    0.04290      -4.13   0.026

s = 3.248       R-sq = 94.5%     R-sq(adj) = 90.8%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    2      541.85      270.92     25.68   0.013
Error         3       31.65       10.55
Total         5      573.50
```

图 E16. 5a

```
The regression equation is
Y = 433 + 37.4 X -0.383 XSQ

Predictor       Coef      Stdev    t-ratio       p
Constant       432.6      141.2       3.06   0.055
X             37.429      7.807       4.79   0.017
XSQ          -0.3829     0.1036      -3.70   0.034

s = 15.83       R-sq = 98.0%     R-sq(adj) = 96.7%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    2       36643       18322     73.15   0.003
Error         3         751         250
Total         5       37395
```

图 E16.5c

Fit	Stdev.Fit	95% C.I.	95% P.I.
1302.01	9.93	(1270.41, 1333.61)	(1242.55, 1361.47)

11. a. SSE = 1 805 − 1 760 = 45

$$F = \frac{\text{MSR}}{\text{MSE}} = \left(\frac{1\,760/4}{45/25}\right) = 244.44$$

$F_{0.05} = 2.76$(分子自由度 4 和分母自由度 25)

因为 244.44>2.76 拒绝 H_0;相关性显著

b. SSE(x_1, x_2, x_3, x_4) = 45

c. SSE(x_2, x_3) = 1 805 − 1 705 = 100

d. $F = \dfrac{(100-45)/2}{1.8} = 15.28$ $F_{0.05} = 3.39$

因为 $F = 15.28 > 3.39$, x_1 和 x_2 是显著的

12. a. Minitabr 的输出结果如图 E16.12a 所示

Minitab 的输出结果如图 E16.12b 所示

c. $F =$

$$\frac{[\text{SSE(简化)} - \text{SSE(完全)}]/(\#\text{额外方式})}{\text{MSE(完全)}}$$

$$= \frac{(23\,157 - 14\,317)/2}{1\,432} = 3.09$$

$F_{0.05} = 4.10$(分子自由度 2 和分母自由度 10)

因为 $F = 3.09 < F_{0.05} = 4.10$, 两个独立变量的可能性不显著

注意:假设我们考虑只将对手拦截的次数相加;相应的 Minitab 结果如图 E16.12c 所示;在此情况下

$$F = \frac{(23\,157 - 14\,335)/1}{1\,303} = 6.77$$

$F_{0.05} = 4.84$(分子自由度 1 和分母自由度 11)

因为 $F = 6.77 > F_{0.05} = 4.84$, 对手拦截次数相加结果是显著的

16. a. Minitab 的输出结果如图 E16.16a 所示

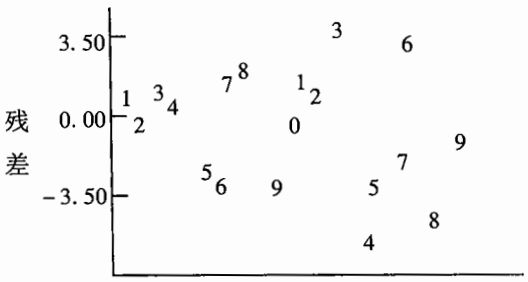
b. 逐步回归(见图 E16.16b)

c. 后向删除法(见图 E16.16c)

d. 最佳子集回归(见图 E16.16d)

21. a. Minitab 的输出结果如图 E16.21a 所示

b. 将数据按顺序列出的函数的残差点如下:未显示出任何正的负相关性



c. 杜宾 − 瓦特森统计(由 Minitab 得出)是 $d = 2.34$; 在 $\alpha = 0.05$, $d_L = 1.18$, $d_U = 1.39$; 因为 $d > d_U$, 不存在显著的正负相关

23.

x_1	x_2	x_3	处理
0	0	0	A
1	0	0	B
0	1	0	C
0	0	1	D

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

26. a.

D_1	D_2	厂商
0	0	1
1	0	2
0	1	3

$$E(y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

b. 见图 E16.26b

c. H_0: $\beta_1 = \beta_2 = 0$

d. p − 值是 0.004<$\alpha = 0.05$, 因此我们得出混合一批原材料的平均时间对每一个商来说并不相等的结果

图 E16.12a

The regression equation is
Points = 170 + 6.61 TeamInt

Predictor	Coef	Stdev	t-ratio	p
Constant	170.13	44.02	3.86	0.002
TeamInt	6.613	2.258	2.93	0.013

s = 43.93 R-sq = 41.7% R-sq(adj) = 36.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	16546	16546	8.57	0.013
Error	12	23157	1930		
Total	13	39703			

Unusual Observations

Obs.	TeamInt	Points	Fit	Stdev.Fit	Residual	St.Resid
13	33.0	340.0	388.4	34.2	-48.4	-1.75 x

X denotes an obs. whose X value gives it large influence.

图E16.12b

The regression equation is
Points = 280 + 5.18 TeamInt - 0.0037 Rushing - 3.92 OpponInt

Predictor	Coef	Stdev	t-ratio	p
Constant	280.34	81.42	3.44	0.006
TeamInt	5.176	2.073	2.50	0.032
Rushing	-0.00373	0.03336	-0.11	0.913
OpponInt	-3.918	1.651	-2.37	0.039

s = 37.84 R-sq = 63.9% R-sq(adj) = 53.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	25386	8462	5.91	0.014
Error	10	14317	1432		
Total	13	39703			

SOURCE	DF	SEQ SS
Regression	1	16546
Error	1	776
Total	1	8064

图 E16.12c

```
The regression equation is
Points = 274 + 5.23 TeamInt - 3.96 OpponInt

Predictor       Coef      Stdev    t-ratio       p
Constant      273.77      53.81       5.09   0.000
TeamInt        5.227      1.931       2.71   0.020
OpponInt      -3.965      1.524      -2.60   0.025

s = 36.10       R-sq = 63.9%     R-sq(adj) = 57.3%

Analysis of Variance

SOURCE        DF          SS          MS         F       p
Regression     2       25386       12684      9.73   0.004
Error         11       14335        1303
Total         13       39703

SOURCE        DF      SEQ SS
TeamInt        1       16546
OpponInt       1        8822
```

图 E16.16a

```
The regression equation is
%College = -26.6 + 0.0970 SATScore

Predictor       Coef      Stdev    t-ratio       p
Constant      -26.61      37.22      -0.72   0.485
SATScore     0.09703    0.03734       2.60   0.019

s = 12.83       R-sq = 29.7%     R-sq(adj) = 25.3%

Analysis of Variance

SOURCE        DF          SS          MS         F       p
Regression     1       1110.8      1110.8      6.75   0.019
Error         16       2632.3       164.5
Total         17       3743.1
```

图 E16.16b

```
STEP             1        2
CONSTANT     -26.61   -26.93

SATScore     0.097    0.084
t-RATIO       2.60     2.46

%TakeSAT              0.204
t-RATIO               2.21

s             12.8     11.5
R-sq         29.68    46.93
```

图 E16. 16c

STEP	1	2	3	4
CONSTANT	33.71	17.46	-32.47	-26.93
Size	-1.56	-1.39		
t-RATIO	-1.43	-1.42		
Spending	-0.0024	-0.0026	-0.0019	
t-RATIO	-1.47	-1.75	-1.31	
Salary	-0.00026			
t-RATIO	-0.40			
SATScore	0.077	0.081	0.095	0.084
t-RATIO	2.06	2.36	2.77	2.46
%TakeSAT	0.285	0.274	0.291	0.204
t-RATIO	2.47	2.53	2.60	2.21
S	11.2	10.9	11.2	11.5
R-sq	59.65	59.10	52.71	46.93

图 E16. 16d

Vars	R-sq	Adj. R-sq	S	Spend	SAT	%Take SAT
1	29.7	25.3	12.826		X	
1	25.5	20.8	13.203			X
2	46.9	39.9	11.508		X	X
2	38.2	30.0	12.417	X	X	
3	52.7	42.6	11.244	X	X	X
3	49.5	38.7	11.618	X	X	X
4	59.1	46.5	10.852	X	X	X
4	52.8	38.3	11.660	X	X	X
5	59.6	42.8	11.219	X	X	X

图 E16.21a

The regression equation is
P/E = 6.51 + 0.569 %Profit

Predictor	Coef	Stdev	t-ratio	p
Constant	6.507	1.509	4.31	0.000
%Profit	0.5691	0.1281	4.44	0.000

s = 2.580 R-sq = 53.7% R-sq(adj) = 51.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	131.40	131.40	19.74	0.000
Error	17	113.14	6.66		
Total	18	244.54			

图 E16.26b

The regression equation is
Time = 23.0 + 5.00 D1 - 2.00 D2

Predictor	Coef	Stdev	t-ratio	p
Constant	23.000	1.106	20.80	0.000
D1	5.000	1.563	3.20	0.011
D2	-2.000	1.563	-1.28	0.233

s = 2.211 R-sq = 70.3% R-sq(adj) = 63.7%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	104.000	52.000	10.64	0.004
Error	9	44.000	4.889		
Total	11	148.000			

第 17 章

1. a.

项目	比价
A	103 = (7.75/7.50)(100)
B	238 = (1 500/630)(100)

b. $I_{2001} = \dfrac{7.75 + 1\,500.00}{7.50 + 630.00}(100) = \dfrac{1\,507.75}{637.50}(100) = 237$

c. $I_{2001} = \dfrac{7.75(1\,500) + 1\,500.00(2)}{7.50(1\,500) + 630.00(2)}(100) = \dfrac{14\,625.00}{12\,510.00}(100) = 117$

d. $I_{2001} = \dfrac{7.75(1\,800) + 1\,500.00(1)}{7.50(1\,800) + 630.00(1)}(100) = \dfrac{15\,450.00}{14\,130.00}(100) = 109$

3. a. 比价 A = (6.00/5.45)100 = 110
B = (5.95/5.60)100 = 106
C = (6.20/5.50)100 = 113

b. $I_{2001} = \dfrac{6.00 + 5.95 + 6.20}{5.45 + 5.60 + 5.50}(100) = 110$

c. $I_{2001}=\dfrac{6.00(150)+5.95(200)+6.20(120)}{5.45(150)+5.60(200)+5.50(120)}$

$(100)=109$

在两年期间速增 9%

6.

		基期			
商品项目	比价	价格	用途	重量	加权比价
A	150	22.00	20	440	66 000
B	90	5.00	50	250	22 500
C	120	14.00	40	560	67 200
			总计	1 250	155 700

$$I=\frac{155\,700}{1\,250}=125$$

7. a. 比价 A = (3.95/2.50)100 = 158

B = (9.90/8.75)100 = 113

C = (0.95/0.99)100 = 96

b.

商品项目	比价	基价	数量	重量 $P_{i0}Q_i$	加权比价
A	158	2.50	25	62.5	9 875
B	113	8.75	15	131.3	14 837
C	96	0.99	60	59.4	5 702
			总计	253.2	30 414

$$I=\frac{30\,414}{253.2}=120$$

化学品原材料的成本增加 20%

10. a. 1980 工资减缩：

$$\frac{7.27(\text{美元})}{82.4}(100)=8.82(\text{美元})$$

1996 工资减缩：

$$\frac{14.36(\text{美元})}{172.6}(100)=8.32(\text{美元})$$

b. $\dfrac{14.36}{7.27}(100)=197.5$; 实际工资增加百分比是 97.5%

c. $\dfrac{8.32}{8.82}(100)=94.3$; 真实工资的变化是减少 5.7%

14. $I=\dfrac{300(18.00)+400(4.90)+850(15.00)}{350(18.00)+220(4.90)+730(15.00)}(100)$

$=\dfrac{20\,110}{18\,328}(100)=110$

15. $I=$

$$\frac{95(1\,200)+75(1\,800)+50(2\,000)+70(1\,500)}{120(1\,200)+86(1\,800)+35(2\,000)+60(1\,500)}$$

$(100)=99$

数量略有下降

第 18 章

1. a.

周	时间序列值	预测	预测误差	预测平方差
1	8			
2	13			
3	15			
4	17	12	5	25
5	16	15	1	1
6	9	16	−7	49
			总计	75

第 7 周的预测是(17 + 16 + 9)/3 = 14

b. MSE = 75/3 = 25

c.

周	时间序列值	预测	预测误差	预测平方差
(t)	(Y_t)	F_t	Y_t-F_t	$(Y_t-F_t)^2$
1	8			
2	13	8.00	5.00	25.00
3	15	9.00	6.00	36.00
4	17	10.20	6.80	46.24
5	16	11.56	4.44	19.71
6	9	12.45	−3.45	11.90
总计				138.85

第 7 周的预测值是 0.2(9) + 0.8(12.45) = 11.76

d. 对 $\alpha=0.2$ 指数平滑预测

$$\text{MSE}=\frac{138.85}{5}=27.77$$

因为 3 周的滑动平均有较小的 MSE，它能提供更好的预测

e.

周 (t)	时间序列值 (Y_t)	预测 F_t	预测误差 $Y_t - F_t$	预测平方差 $(Y_t - F_t)^2$
1	8			
2	13	8.0	5.0	25.00
3	15	10.0	5.0	25.00
4	17	12.0	5.0	25.00
5	16	14.0	2.0	4.00
6	9	14.8	-5.8	33.64
			总计	112.64

$$\text{MSE} = \frac{112.64}{5} = 22.52$$

平滑常数 0.4 能提供更好的预测；用 = 0.4 得出第 7 周的预测值为 0.4(9) + 0.6(14.8) = 12.48

8. a.

月份	时间序列值	3 个月的滑动平均值	平方差	$\alpha = 0.2$ 预测	平方差
1	240				
2	350			240.00	12 100.00
3	230			262.00	1 024.00
4	260	273.33	177.69	255.60	19.36
5	280	280.00	0.00	256.48	553.19
6	320	256.67	4 010.69	261.18	3 459.79
7	220	286.67	4 444.89	272.95	2 803.70
8	310	273.33	1 344.69	262.36	2 269.57
9	240	283.33	1 877.49	271.89	1 016.97
10	310	256.67	2 844.09	265.51	1 979.36
11	240	286.67	2 178.09	274.41	1 184.05
12	230	263.33	1 110.89	267.53	1 408.50
总计			17 988.52		27 818.49

MSE(3 个月) = 17 988.52/9 = 1 998.72

MSE($\alpha = 0.2$) = 27 818.49/11 = 2 528.95

基于下面的 MSE 值，3 个月的滑动平均更好些；然而，指数平滑法则要求包含两个月的数据，这使任何方法都难以预测。仅用第 4—12 月的误差，指数平滑的 MSE 被修正为

MSE($\alpha = 0.2$) = 14 694.49/9 = 1 632.72

那么，指数平滑法更适用第 4—12 个月的预测

b. 运用指数，

$$F_{13} = \alpha Y_{12} + (1 - \alpha) F_{12}$$
$$= 0.20(230) + 0.80(267.53) = 260$$

12. $\sum t = 15, \sum t^2 = 55, \sum Y_t = 55, \sum tY_t = 186$

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n}$$

$$= \frac{186 - (15)(55)/5}{55 - (15)^2/5} = 2.1$$

$$b_0 = \bar{Y} - b_1 \bar{t} = 11 - 2.1(3) = 4.7$$
$$T_t = 4.7 + 2.1t$$
$$T_6 = 4.7 + 2.1(6) = 17.3$$

14. $\sum t = 21, \sum t^2 = 91, \sum Y_t = 117.1, \sum tY_t = 403.7$

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n}$$

$$= \frac{403.7 - (21)(117.1)/6}{91 - (21)^2/6} = -0.3514$$

$$b_0 = \bar{Y} - b_1 \bar{t} = 19.5167 - (-0.3514)(3.5)$$
$$= 20.7466$$
$$T_t = 20.7466 - 0.3514t$$

学生注册人数每年大约减少 351 人

22. a.

年份	季度	Y_t	4 个季度滑动平均	滑动平均中心值
1	1	4		
	2	2		
			3.50	
	3	3		3.750
			4.00	
	4	5		4.125
			4.25	
2	1	6		4.500
			4.75	
	2	3		5.000
			5.25	
	3	5		5.375
			5.50	
	4	7		5.875
			6.25	
3	1	7		6.375
			6.50	
	2	6		6.625
			6.75	
	3	6		
	4	8		

b.

年份	季度	Y_t	滑动平均中心值	不规则季节成份
1	1	4		
	2	2		
	3	3	3.750	0.800 0
	4	5	4.125	1.212 1
2	1	6	4.500	1.333 3
	2	3	5.000	0.600 0
	3	5	5.375	0.930 2
	4	7	5.875	1.191 5
3	1	7	6.375	1.098 0
	2	6	6.625	0.905 7
	3	6		
	4	8		

季度	不规则季节成分值	季节指数
1	1.333 3, 1.098 0	1.215 7
2	0.600 0, 0.905 7	0.752 9
3	0.800 0, 0.930 2	0.865 1
4	1.212 1, 1.191 5	1.201 8
	总计	4.035 5

$$\text{调节季节指数} = \frac{4}{4.035\ 5} = 0.991\ 2$$

季度	调节季节指数
1	1.205 0
2	0.746 3
3	0.857 5
4	1.191 2

第 19 章

1. $n=10$, $p=0.50$ 的二项分布

$P(0)+P(1)=0.0108$; 加上 $P(2)$ 后尾数超过 0.025; 因此, 若"+"号的个数小于 2 或大于 8 则拒绝 H_0

加号的数目是 7

x	频率	x	频率
0	0.001 0	6	0.205 1
1	0.009 8	7	0.117 2
2	0.043 9	8	0.043 9
3	0.117 2	9	0.009 8
4	0.205 1	10	0.001 0
5	0.246 1		

不拒绝 H_0; 得出未指示出差异存在的结论

2. 用 $n=27$ 得出不等于 150 的值; 利用正态分布近似, 其中 $\mu=np=0.5(27)=13.5$ 和 $\sigma=\sqrt{0.25n}=\sqrt{0.25(27)}=2.6$

用 $x=22$ 作为加号的个数并得到如下的检验统计:

$$z=\frac{x-\mu}{\sigma}=\frac{22-13.5}{2.6}=3.27$$

由于 $\alpha=0.01$, 拒绝若 $z>2.33$; 因为 $z=3.27>2.33$, 拒绝 H_0 并得到中位数大于 150

4. 我们需要确定"更好"反应的数目和"更差"反应的数目; 两者的和是研究所用的样本规模

$n=0.34(1\ 253)+0.29(1\ 253)=789.4$

用大样本正态分布检验; 这意味差 n 的值 ($n=789.4$)。利用

$\mu=0.5n=0.5(789.4)=394.7$

$\sigma=\sqrt{0.25n}=\sqrt{0.25(789.4)}=14.05$

令 $p=$ 认为孩子们将会有更好的未来的成年人的比例

H_0: $p\leqslant 0.50$

H_a: $p>0.50$

$$x=0.34(1\ 253)=426.0$$

$$z=\frac{x-\mu}{\sigma}=\frac{426.0-394.7}{14.05}=2.23$$

由于 $\alpha=0.05$, 拒绝若 $z>1.645$; 因为 $z=2.23>1.645$, 拒绝 H_0 并得出认为孩子们的未来会更好的成年人超过半数的结论

12. H_0: 总体已知

H_a: 总体未知

燃料添加剂					
1	2	差	绝对值	秩	符号秩
20.12	18.05	2.07	2.07	9	+9
23.56	21.77	1.79	1.79	7	+7
22.03	22.57	-0.54	0.54	3	-3
19.15	17.06	2.09	2.09	10	+10
21.23	21.22	0.01	0.01	1	+1
24.77	23.80	0.97	0.97	4	+4
16.16	17.20	-1.04	1.04	5	-5
18.55	14.98	3.57	3.57	12	+12
21.87	20.03	1.84	1.84	8	+8
24.23	21.15	3.08	3.08	11	+11
23.21	22.78	0.43	0.43	2	+2
25.02	23.70	1.32	1.32	6	+6
					$T=62$

$\mu_T=0$

$$\sigma_T=\sqrt{\frac{n(n+1)(2n+1)}{6}}=\sqrt{\frac{12(13)(25)}{6}}$$

$=25.5$

$$z=\frac{T-\mu_T}{\sigma_T}=\frac{62-0}{25.5}=1.83$$

双侧检验；拒绝 H_0 若 $z<-1.96$ 或 $z>1.96$；因为 $z=2.43>1.96$，拒绝 H_0 并得出两种添加剂之间有显著差异的结论

13.

无放松	有放松	差异	绝对差的秩	符号秩
15	10	5	9	+9
12	10	2	3	+3
22	12	10	10	+10
8	11	-3	6.5	-6.5
10	9	1	1	+1
7	5	2	3	+3
8	10	-2	3	-3
10	7	3	6.5	+6.5
14	11	3	6.5	+6.5
9	6	3	6.5	+6.5
				$T=36$

$\mu_T=0$

$$\sigma_T=\sqrt{\frac{n(n+1)(2n+1)}{6}}=\sqrt{\frac{10(11)(21)}{6}}$$

$=19.62$

$$z=\frac{T-\mu_T}{\sigma_T}=\frac{36}{19.62}=1.83$$

单侧检验；若 $z>1.645$ 拒绝 H_0

拒绝 H_0 在放松偏好方面有显著差异

18. 求出组合样本的秩和每一样本的秩的和；这是一个小样本检验因为 $n_1=7$ 和 $n_2=9$

添加剂 1		添加剂 2	
MPG	秩	MPG	秩
17.3	2	18.7	8.5
18.4	6	17.8	4
19.1	10	21.3	15
16.7	1	21.0	14
18.2	5	22.1	16
18.6	7	18.7	8.5
17.5	3	19.8	11
	34	20.7	13
		20.2	12
			102

$T=34$

由于 $\alpha=0.05$，$n_1=7$，和 $n_2=9$，

$T_L=41$ 和 $T_U=7(7+9+1)-41=78$

因为 $T=34<41$，拒绝 H_0，并得出在汽油里程方面有显著差异的结论

19. a.

注册会计师	秩	财务分析师	秩
25.2	5	24.0	2
33.8	19	24.2	3
31.3	16	28.1	10
33.2	18	30.9	15
29.2	13	26.9	8.5
30.0	14	28.6	11
25.9	6	24.7	4
34.5	20	28.9	12
31.7	17	26.8	7
26.9	8.5	23.9	1
	136.5		73.5

$$\mu_T = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}(10)(10+10+1)$$

$$= 105$$

$$\sigma_T = \sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}$$

$$= \sqrt{\frac{1}{12}(10)(10)(10+10+1)}$$

$$= 13.23$$

$T = 136.5$

拒绝 H_0 若 $z < -1.645$ 或若 $z > 1.645$

$$z = \frac{136.5 - 105}{13.23} = 2.38$$

拒绝 H_0；这两个职业薪水明显不同

b. 注册会计师　32 300 美元

财务分析师　26 700 美元

26. 次序：

商品 A	商品 B	商品 C
4	11	7
8	14	2
10	15	1
3	12	6
9	13	5
34	65	21

$$W = \frac{12}{(15)(16)}\left[\frac{(34)^2}{5} + \frac{(65)^2}{5} + \frac{(21)^5}{5}\right] - 3(15+1) = 58.22 - 48 = 10.22$$

$\chi^2_{0.05} = 5.991\,47$（自由度 2）

拒绝 H_0 并得出产品比例不等的结论

28. 次序：

游泳	网球	自行车
8	9	5
4	14	1
11	13	3
6	10	7
12	15	2
41	61	18

$$W = \frac{12}{15(15+1)}\left[\frac{41^2}{5} + \frac{61^2}{5} + \frac{18^2}{5}\right] - 3(15+1)$$

$$= 9.26$$

$\chi^2_{0.05} = 5.991\,47$（自由度 2）

因为 $9.26 > 5.991\,47$，拒绝 H_0，并且得出活动不同的结论

32. a. $\sum d_i^2 = 52$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(52)}{10(99)} = 0.68$$

b. $\sigma_{r_s} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{9}} = 0.33$

$$z = \frac{r_s - 0}{\sigma r_s} = \frac{0.68}{0.33} = 2.06$$

拒绝 H_0 若 $z < -1.96$ 或 $z > 1.96$；因为 $z = 2.06 > 1.96$，拒绝 H_0，并得出存在秩显著相关的结论

34. $\sum d_i^2 = 250$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(250)}{11(120)} = -0.136$$

$$\sigma_{r_s} = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{10}} = 0.32$$

$$z = \frac{r_s - 0}{\sigma_{r_s}} = -\frac{0.136}{0.32} = -0.425$$

拒绝 H_0 若 $z < -1.96$ 或 $z > 1.96$；因为 $z = -0.425$，不拒绝 H_0；我们不能得出次序之间显著相关的结论

第 20 章

4. R 控制图：

$$UCL = \bar{R}D_4 = 1.6(1.864) = 2.98$$

$$LCL = \bar{R}D_3 = 1.6(0.136) = 0.22$$

$\bar{x}$ 控制图：

$$UCL = \bar{\bar{x}} + A_2\bar{R} = 28.5 + 0.373(1.6) = 29.10$$

$$LCL = \bar{\bar{x}} - A_2\bar{R} = 28.5 - 0.373(1.6) = 27.90$$

10. $f(0)=\dfrac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$

当 $p=0.02$ 时，接受多数的概率是

$$f(0)=\frac{25!}{0!(25-0)!}(0.02)^0(1-0.02)^{25}$$

$$=0.6035$$

当 $p=0.06$ 时，接受多数的概率是

$$f(0)=\frac{25!}{0!(25-0)!}(0.06)^0(1-0.06)^{25}$$

$$=0.2129$$

第21章

1. a. $\bar{x}=215$ 是总体均值的估计值

b. $s_{\bar{x}}=\dfrac{20}{\sqrt{50}}\sqrt{\dfrac{800-50}{800}}=2.7386$

c. $215\pm2(2.7386)$ 或 209.5228—220.4772

5. a. $\bar{x}=149\,670$ 和 $s=73\,420$

$$s_{\bar{x}}=\sqrt{\left(\frac{771-50}{771}\right)}\frac{73\,420}{\sqrt{50}}=10\,040.83$$

近似 95% 的置信区间

$149\,760\pm2(10\,040.83)$，或

129 588.34 美元—169 751.66 美元

b. $\hat{X}=N\bar{x}=771(149\,670)=115\,395\,570$

$s_{\hat{x}}=Ns_{\bar{x}}=771(10\,040.83)=7\,741\,479.93$

近似 95% 的置信区间

$115\,395\,570\pm2(7\,741\,479.93)$，或

99 912 810.14 美元—130 878 729.86 美元

c. $\bar{p}=18/50=0.36$ 和

$$s_{\bar{p}}=\sqrt{\left(\frac{771-50}{771}\right)\frac{(0.36)(0.64)}{49}}=0.0663$$

近似 95% 的置信区间

$0.36\pm2(0.0663)$，或

0.2274—0.4926

这一区间相当大；样本容量也必须大，以便得出关于总体比例较窄的置信区间

7. a. 第 1 层：$\bar{x}_1=138$

第 2 层：$\bar{x}_2=103$

第 3 层：$\bar{x}_3=210$

b. 第 1 层

$$\bar{x}_1=138;\ s_{\bar{x}_1}=\left(\frac{30}{\sqrt{20}}\right)\sqrt{\frac{200-20}{200}}=6.3640$$

近似 95% 的置信区间

$138\pm2(6.3640)$，或 125.272—150.728

第 2 层

$$\bar{x}_2=103;\ s_{\bar{x}_2}=\left(\frac{25}{\sqrt{30}}\right)\sqrt{\frac{250-30}{250}}=4.2817$$

近似 95% 的置信区间

$103\pm2(4.2817)$，或 94.4366—111.5634

第 3 层

$$\bar{x}_3=210;\ s_{\bar{x}_3}=\left(\frac{50}{\sqrt{25}}\right)\sqrt{\frac{100-25}{100}}=8.6603$$

近似 95% 的置信区间

$210\pm2(8.6603)$，或 192.6794—227.3206

c. $\bar{x}_{st}=\left(\dfrac{200}{550}\right)138+\left(\dfrac{250}{550}\right)103+\left(\dfrac{100}{550}\right)210$

$$=50.1818+46.8182+38.1818$$

$$=135.1818$$

$$s_{\bar{x}_{st}}=$$

$$\sqrt{\left(\frac{1}{(550)^2}\right)\left(200(180)\frac{(30)^2}{20}+250(220)\frac{(25)^2}{30}+100(75)\frac{(50)^2}{25}\right)}$$

$$=\sqrt{\left(\frac{1}{(550)^2}\right)3\,515\,833.3}=3.4092$$

近似 95% 的置信区间

$135.1818\pm2(3.4092)$，或 128.3634—142.0002

14. a. $\bar{x}_c=\dfrac{\sum x_i}{\sum M_i}=\dfrac{750}{50}=15$

$$\hat{X}=M\bar{x}_c=300(15)=4\,500$$

$$\bar{p}_c=\frac{\sum a_i}{\sum M_i}=\frac{15}{50}=0.30$$

b. $\sum(x_i-\bar{x}_cM_i)^2=[95-15(7)]^2+[325-15(18)]^2+[190-15(15)]^2+[140-15(10)]^2$

$$=(-10)^2+(55)^2+(-35)^2+(-10)^2$$

$$=4\,450$$

$$s_{\bar{x}_c} = \sqrt{\left(\frac{25-4}{(25)(4)(12)^2}\right)\left(\frac{4\,450}{3}\right)} = 1.470\,8$$

$$s_{\hat{X}} = M s_{\bar{x}_c} = 300(1.470\,8) = 441.24$$

$$\sum (a_i - \bar{p}_c M_i)^2 = [1 - 0.3(7)]^2 + [6 - 0.3(18)]^2 + [6 - 0.3(15)]^2 + [2 - 0.3(10)]^2$$

$$= (-1.1)^2 + (0.6)^2 + (1.5)^2 + (-1)^2$$

$$= 4.82$$

$$s_{\bar{p}_c} = \sqrt{\left(\frac{25-4}{(25)(4)(12)^2}\right)\left(\frac{4.82}{3}\right)} = 0.048\,4$$

c. 总体均值近似 95% 的置信区间

$15 \pm 2(1.470\,8)$，或 12.058 4—17.941 6

d. 总体均值近似 95% 的置信区间

$4\,500 \pm 2(441.24)$，或 3 617.52—5 382.48

e. 总体比例近似 95% 的置信区间

$0.30 \pm 2(0.048\,4)$，或 0.203 2—0.396 8

译后记

由中信出版社推出的这本译作《商务与经济统计》，可谓恰逢其时和恰到好处。

在经济全球化进程加速、我国加入WTO的大背景下，各项商务和经贸活动都在尽快地与国际全面、深度接轨，在实践层面有许多繁杂琐碎的事情要做，因而需要更加具体和详细地了解世界上经贸实务的操作程序及规范等事项。该译作的出版，在一定程度上满足了商务统计教学和实际工作人员对这方面的迫切需求，此乃恰逢其时。

说它恰到好处，是因为我国原来实务方面的教材就比较少，本书的翻译出版，正好起弥补的作用。本书原著有一些显著特色：定位鲜明准确、内容精选得当；编写体例系统规范、通俗易懂、教学方便、突出实用；案例选择颇具代表性、生动鲜活；国际流行风格，注重习题训练，并附有以两种格式录制的数据光盘。本书中统计术语的翻译，以现行大学教材中的用法为基本依据，而一些不常见的地名、公司名称等专用词汇则是直接引用原文。译者分工为：王国成全面组织协调，负责全书的统一校译和最终定稿工作，并承担第1章、附录等零散辅助内容的翻译；王峰翻译第2—8章，并承担了许多繁杂的事务性工作；袁晓东翻译第9—14章；卿前锋翻译第15—21章；虽然我们力图保持原著生动活泼、文笔流畅的原有风格，但由于该书取材极其广泛，翻译工作量大，时间相对紧迫，译本难免存在一些不足之处，欢迎批评指正。

在本书翻译过程中，郭金龙博士提供了一些有益建议，李玉丰硕士和王秀玲女士在联系沟通、文稿打印整理等方面给予了大量帮助，在此一并对他们表示衷心的感谢。