

商 业 统 计

江苏财经职业技术学院

二〇一三年元月

CONTENTS

作者序	3.3 分布形态、相对位置的度量以及异常值的检测 / 53
作者简介	3.4 探索性数据分析 / 57
前言	3.5 两变量间关系的度量 / 60
教学建议	3.6 加权平均数和使用分组数据 / 65
第1章 数据与统计资料 / 1	小结、关键术语、重要公式、补充练习
实践中的统计 《商业周刊》 / 1	案例 3-1 Pelican 商店 / 73
1.1 统计学在商务和经济中的应用 / 2	第4章 概率 / 74
1.2 数据 / 3	实践中的统计 Oceanwide 海产品公司 / 74
1.3 数据来源 / 8	4.1 试验、计数法则和概率分配 / 75
1.4 描述统计 / 11	4.2 事件及其概率 / 82
1.5 统计推断 / 12	4.3 概率的基本性质 / 84
1.6 计算机与统计分析 / 13	4.4 条件概率 / 87
1.7 数据挖掘 / 13	4.5 贝叶斯定理 / 92
1.8 统计实践的道德准则 / 14	小结、关键术语、重要公式、补充练习
小结、关键术语、练习	案例 4-1 Hamilton 县的法官 / 99
第2章 描述统计学 I：表格法和图形法 / 20	第5章 离散型概率分布 / 101
实践中的统计 高露洁 - 棕榄公司 / 20	实践中的统计 花旗银行 / 101
2.1 分类型数据汇总 / 21	5.1 随机变量 / 102
2.2 数量型数据汇总 / 24	5.2 离散型概率分布 / 103
2.3 探索性数据分析：茎叶显示 / 30	5.3 数学期望与方差 / 106
2.4 交叉分组表和散点图 / 33	5.4 二项概率分布 / 108
小结、关键术语、重要公式、补充练习	5.5 泊松概率分布 / 114
案例 2-1 Pelican 商店 / 42	5.6 超几何概率分布 / 117
第3章 描述统计学 II：数值方法 / 43	小结、关键术语、重要公式、补充练习
实践中的统计 Small Fry 设计公司 / 43	第6章 连续型概率分布 / 123
3.1 位置的度量 / 44	实践中的统计 宝洁公司 / 123
3.2 变异程度的度量 / 49	6.1 均匀概率分布 / 124
	6.2 正态概率分布 / 126

6.3 二项概率的正态近似 / 133

6.4 指数概率分布 / 134

小结、关键术语、重要公式、补充练习

案例 6-1 Specialty 玩具公司 / 138

第 7 章 抽样和抽样分布 / 140

实践中的统计 MeadWestvaco 公司 / 140

7.1 EAI 公司的抽样问题 / 141

7.2 抽样 / 142

7.3 点估计 / 146

7.4 抽样分布简介 / 148

7.5 $\bar{x}$ 的抽样分布 / 1507.6 $\bar{p}$ 的抽样分布 / 156

7.7 点估计的性质 / 159

7.8 其他抽样方法 / 160

小结、关键术语、重要公式、补充练习

第 8 章 区间估计 / 166

实践中的统计 Food Lion / 166

8.1 总体均值: σ 已知情形 / 1678.2 总体均值: σ 未知情形 / 170

8.3 样本容量的确定 / 176

8.4 总体比率 / 178

小结、关键术语、重要公式、补充练习

案例 8-1 Yong Professional 杂志 / 184

第 9 章 假设检验 / 186

实践中的统计 John Morrell 公司 / 186

9.1 原假设和备择假设的建立 / 187

9.2 第一类错误和第二类错误 / 189

9.3 总体均值: σ 已知 / 1919.4 总体均值: σ 未知 / 200

9.5 总体比率 / 204

9.6 假设检验与决策 / 206

9.7 计算第二类错误的概率 / 207

9.8 确定总体均值假设检验的样本容量 / 210

小结、关键术语、重要公式、补充练习

案例 9-1 Quality Associates 有限公司 / 215

第 10 章 两总体均值和比例的推断 / 217

实践中的统计 美国食品与药物管理局 / 217

10.1 两总体均值之差的推断:

 σ_1 和 σ_2 已知 / 218

10.2 两总体均值之差的推断:

 σ_1 和 σ_2 未知 / 222

10.3 两总体均值之差的推断: 匹配样本 / 227

10.4 两总体比例之差的推断 / 230

小结、关键术语、重要公式、补充练习

案例 10-1 Par 公司 / 236

第 11 章 关于总体方差的统计推断 / 238

实践中的统计 美国政府审计总署 / 238

11.1 一个总体方差的统计推断 / 239

11.2 两个总体方差的统计推断 / 245

小结、重要公式、补充练习

案例 11-1 空军训练计划 / 250

第 12 章 拟合优度检验和独立性检验 / 252

实践中的统计 联合劝募协会 / 252

12.1 拟合优度检验: 多项总体 / 253

12.2 独立性检验 / 256

12.3 拟合优度检验: 泊松分布与
正态分布 / 261

小结、关键术语、重要公式、补充练习

案例 12-1 两党议程变更 / 269

第 13 章 实验设计与方差分析 / 270

实践中的统计 Burke 市场营销服务公司 / 270

13.1 实验设计和方差分析简介 / 271

13.2 方差分析和完全随机化实验设计 / 275

13.3 多重比较方法 / 282

13.4 随机化区组设计 / 286

13.5 析因实验 / 291

小结、关键术语、重要公式、补充练习

案例 13-1 Wentworth 医疗中心 / 299

第14章 简单线性回归 / 300

实践中的统计 联盟数据系统 / 300

- 14.1 简单线性回归模型 / 301
- 14.2 最小二乘法 / 303
- 14.3 判定系数 / 308
- 14.4 模型的假定 / 313
- 14.5 显著性检验 / 313
- 14.6 应用估计的回归方程进行
估计和预测 / 320
- 14.7 计算机解法 / 324
- 14.8 残差分析: 证实模型假定 / 326
- 14.9 残差分析: 异常值和有影响的
观测值 / 331

小结、关键术语、重要公式、补充练习

案例 14-1 测量股票市场风险 / 340

第15章 多元回归 / 342

实践中的统计 dunnhumby / 342

- 15.1 多元回归模型 / 343
- 15.2 最小二乘法 / 344
- 15.3 多元判定系数 / 348
- 15.4 模型的假定 / 350
- 15.5 显著性检验 / 351
- 15.6 应用估计的回归方程进行
估计和预测 / 356
- 15.7 分类自变量 / 357
- 15.8 残差分析 / 362
- 15.9 logistic 回归 / 367

小结、关键术语、重要公式、补充练习

案例 15-1 消费者调查股份有限公司 / 379

第16章 回归分析: 建立模型 / 380

实践中的统计 Monsanto 公司 / 380

- 16.1 一般线性模型 / 381
- 16.2 确定什么时候增加或者删除变量 / 389
- 16.3 大型问题的分析 / 393
- 16.4 变量选择方法 / 395

16.5 实验设计的多元回归方法 / 400

16.6 自相关性和杜宾-瓦特森检验 / 403

小结、关键术语、重要公式、补充练习

案例 16-1 职业高尔夫球协会巡回赛的
统计分析 / 407

第17章 指数 / 408

实践中的统计 美国劳工部, 劳工统计局 / 408

- 17.1 价比 / 409
- 17.2 综合物价指数 / 409
- 17.3 根据价比计算综合物价指数 / 412
- 17.4 一些重要的价格指数 / 413
- 17.5 根据物价指数减缩一个数列 / 415
- 17.6 物价指数: 其他注意事项 / 417
- 17.7 物量指数 / 418

小结、关键术语、重要公式、补充练习

第18章 时间序列分析及预测 / 422

实践中的统计 内华达州职业健康诊所 / 422

- 18.1 时间序列的模式 / 423
- 18.2 预测精度 / 428
- 18.3 移动平均法和指数平滑法 / 432
- 18.4 趋势推测法 / 439
- 18.5 季节性和趋势 / 448
- 18.6 时间序列分解法 / 454

小结、关键术语、重要公式、补充练习

案例 18-1 预测食品和饮料的销售 / 465



第1章

数据与统计资料

实践中的统计

《商业周刊》^①

纽约州，纽约市

《商业周刊》是世界上拥有最广泛读者的商业类杂志，其全球发行量超过100万份。该杂志在全球的26个分社中有200多名忠于职守的记者和编辑，主要登载各类有关经济和贸易活动的饶有兴趣的文章，除了刊登当前热门话题的特色文章外，还有一些关于国际贸易、经济分析、信息处理和科技进展等常设栏目。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，以便评估其对商务和经济状况产生的影响。

几乎每一期《商业周刊》都会发表一篇有关当前热门话题的深度报告，通常这些深度报告包含统计论据和综述，以帮助读者了解商务和经济信息。例如，在2009年2月23日那一期，发表了一篇关于家庭丧失抵押品赎回权危机的专题文章；在2009年3月17日那一期，发表了一篇关于股票市场何时开始复苏的讨论；在2009年5月4日那一期，发表了一篇如何使减薪的痛苦少些的特别报告。另外，每一期《商业周刊投资者》都会发布关于经济状况的统计资料，包括产品指

数、股票价格、共同基金和利息率等。

《商业周刊》还利用统计报告和统计信息来帮助管理其自身的经营。例如，一份关于订户的年度综述报告可以帮助公司了解订户人口特征、阅读习惯、购物习惯、生活方式，等等。《商业周刊》的管理人员利用从调查中得出的统计结论更好地为订户和广告客户提供服务。最近对北美订户的调查显示，90%的《商业周刊》订户在家使用计算机，并且64%的订户因工作需要而购买了计算机。这一统计结果提醒《商业周刊》的管理者：订户会对个人计算机发展趋势方面的文章感兴趣，而且，调查的结论也会引起潜在的广告客户的兴趣。订户中在家使用计算机和因工作需要购买计算机有如此之高的百分比，将会刺激个人计算机厂商考虑在《商业周刊》上做广告。

本章我们讨论可用于统计分析的数据类型，并说明如何取得这些数据。我们将介绍描述统计和统计推断，它们是将数据转为有意义且易于解释的统计信息的方法。■

我们经常会在报纸和杂志的文章中看到下面的各种报道：

- 美国房地产协会的报告说，首次房屋购买者支付价格的中位数为165 000美元（*The Wall Street Journal*，

① 作者感谢为“实践中的统计”提供了本案例的《商业周刊》的调研管理人员 Charlene Trentham。

2009. 2. 11)。

- 美国大学生体育协会 (NCAA) 主席迈尔斯·布兰德 (Myles Brand) 报告说, 大学生运动员以创纪录的速度获得学位。最新数字显示, 有 79% 的男女学生运动员已毕业 (Associated Press, 2008. 10. 15)。
- 平均单程上班时间是 25.3 分钟 (U. S. Census Bureau, 2009. 3)。
- 高达 11% 的美国房屋是空置的, 这是房地产暴涨之后的暴跌造成的供大于求的现象 (UAS Today, 2009. 2. 13)。
- 历史上, 普通汽油全国平均价格首次达到每加仑 4.00 美元 (Cable News Network website, 2008. 6. 8)。
- 纽约扬基队在职业棒球大联盟中的薪酬最高。全队的总薪酬为 201 449 289 美元, 中位数为 5 000 000 美元 (UAS Today Salary Data Base, 2009. 4)。
- 道琼斯工业平均指数收于 8 721 点 (The Wall Street Journal, 2009. 6. 2)。

前面报道中的数值事实 (165 000 美元、79%、25.3 分钟、11%、4.00 美元、201 449 289 美元、5 000 000 美元和 8 721 点) 称为统计资料。在这种用法中, 术语统计资料是指数值事实, 如平均值、中位数、百分数和指数等, 它们可以帮助我们了解各种商务和经济情况。然而, 统计学的研究领域或对象的内涵更加丰富。广义上讲, 统计学 (statistics) 是搜集、分析、表述和解释数据的科学与艺术。特别是在商务和经济活动中, 搜集、分析、表述、解释数据旨在帮助管理者和决策者更好地理解商务和经济环境的变化, 从而使其能科学、正确地做出决策。在本书中, 我们强调统计学在商务和经济决策方面的用途。

本章以统计学在商务和经济中的一些应用实例为开端, 在 1.2 节我们定义数据这一术语的含义, 并介绍数据集的概念。在这一节中, 还介绍了变量和观测值等一些重要的术语, 讨论了数量型数据和分类型数据之间的区别, 并且说明了截面数据和时间序列数据的应用。在 1.3 节我们探讨了如何从现有的资料来源中获得数据, 或通过调查和设计的实验研究获得新数据。对如今在数据获取中起着重要作用的互联网, 也给予了高度的重视。1.4 节和 1.5 节论述了数据在描述统计和统计推断中的用途。在本章的最后 3 节还给出了计算机在统计分析中的角色, 简单介绍了数据挖掘这个比较新的领域, 并且讨论了统计实践的道德准则。本书配套光盘中的附录 1A 对 Excel 的插件 StatTools 做了介绍, 它扩展了 Microsoft Excel 用户的统计选项功能。

1.1 统计学在商务和经济中的应用

在如今贸易和经济全球化的背景下, 任何人都可以获取大量的统计信息。最成功的管理者和决策者是那些能够理解信息并有效利用信息的人。本节我们举例说明统计学在商务和经济中的应用。

1.1.1 会计

会计师事务所在对其客户进行审计时需要使用统计抽样程序。例如, 假设一个事务所想确定列示在客户资产负债表上的应收账款金额是不真实地反映了应收账款的实际金额。通常应收账款的数量是如此之大, 以致查看和验证每一账户将花费大量的时间和费用。在这种情况下, 一般的做法是: 审计人员从账户中选择一个子集作为样本, 在查看样本账户的准确性后, 审计师得出有关列示在客户资产负债表上的应收账款金额是不是可以接受的结论。

1.1.2 财务

财务顾问们利用各种各样的统计信息指导投资。在股票市场中,财务顾问们综合了包括市盈率和股息等方面的财务数据,通过对比单只股票和股票市场平均状况的信息,就可以得出某一只股票其价值是高估还是低估的结论。例如,《巴伦周刊》(2008.2.18)报道说,道琼斯工业平均指数的30只股票的平均股息收益率是2.45%,Altria集团的股息收益率是3.05%。在这种情况下,关于股息收益率的统计信息表明,Altria集团的股息收益率比道琼斯工业平均指数股票的平均收益率高。财务顾问从而可能得出Altria集团的股票被低估的结论。这一信息和其他有关Altria集团的信息将帮助财务顾问做出是买、是卖还是持股的建议。

1.1.3 市场营销

零售结账柜台的电子扫描仪正用于搜集各种市场调研用的数据。例如,数据供应商ACNielsen公司和Information Resources公司从商店购买POS扫描数据,经过加工处理,做出统计汇总后再出售给制造商。制造商为取得这些扫描数据,每一类产品要花费数十万美元。制造商也购买特价销售和利用店内陈列品等促销活动的数据及统计研究报告。品牌管理人员可以查看扫描和促销活动的统计资料,从而能更好地理解促销活动和销售额之间的关系。这样的分析对制定各种产品未来的市场营销战略大有裨益。

1.1.4 生产

由于当今进入了重视质量的时代,因此质量管理是统计学在生产中的一项重要应用。各种统计质量管理图用于监测生产过程的产出。特别地,用 $\bar{x}$ 控制图可以监测平均产出。例如,假定有一台12盎司的软饮料罐装机,定期地从产品中选择一些罐装饮料作为样本,计算出样本罐装量的平均值。这一平均值或 $\bar{x}$ 的值标在一张 $\bar{x}$ 控制图上。当该数值位于控制上限以上时,则表明产品罐装量过高;当该数值位于控制下限以下时,则表明产品罐装量不足;当 $\bar{x}$ 的值位于控制图的控制上限和控制下限之间时,这一过程称为处于“受控”状态,并且只要 $\bar{x}$ 的值落在控制图的控制上限和控制下限之间,就允许连续生产。更贴切的解释是, $\bar{x}$ 控制图能帮助确定何时要调整生产过程。

1.1.5 经济

人们经常要求经济学家对未来的经济或某一方面的发展做出预测。他们在预测时需要用到各种统计信息。例如,在预测通货膨胀率时,经济学家利用诸如生产者价格指数、失业率、制造业开工率等指标的统计信息。这些统计指标往往要输入到预测通货膨胀率的电脑预测模型中。

本节所介绍的统计学的这些应用是本书的组成部分。这些例子可以使我们对统计学的广泛应用有一个大致的了解。为了补充这些例子,我们请求商务和经济领域的专家提供每章开篇的“实践中的统计”,以便引出每章所要讲解的内容。“实践中的统计”充分显示了统计学在各种各样的商务和经济问题中的重要性。

1.2 数据

数据(data)是为描述和解释所搜集、分析和汇总的事实和数字。将用于特定研究而搜集的所有数据称为研究的数据集(data set)。表1-1是一个包含了25只共同基金信息的数据集,它们是2008年的Morningstar Funds 500的一部分。晨星(Morningstar)是一家公司,该公司追踪7 000多只共同基金并对其中2 000只做深度分析。

他们的建议受到财务分析家和个体投资者的密切关注。

表 1-1 25 只共同基金的数据集

基金名称	基金类型	净资产值 (美元)	5 年平均回报率 (%)	费用比率 (%)	晨星评级
American Century Intl. Disc	IE	14.37	30.53	1.41	3 星
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4 星
American Century Ultra	DE	24.94	10.88	0.99	3 星
Artisan Small Cap	DE	16.92	15.67	1.18	3 星
Brown Cap Small	DE	35.73	15.85	1.20	4 星
DFA U. S. Micro Cap	DE	13.47	17.23	0.53	3 星
Fidelity Contrafund	DE	73.11	17.99	0.89	5 星
Fidelity Overseas	IE	48.39	23.46	0.90	4 星
Fidelity Sel Electronics	DE	45.60	13.50	0.89	3 星
Fidelity Sh-Term Bond	FI	8.60	2.76	0.45	3 星
Gabelli Asset AAA	DE	49.81	16.70	1.36	4 星
Kalmar Gr Val Sm Cp	DE	15.30	15.31	1.32	3 星
Marsico 21st Century	DE	17.44	15.16	1.31	5 星
Mathews Pacific Tiger	IE	27.86	32.70	1.16	3 星
Oakmark I	DE	40.37	9.51	1.05	2 星
PIMCO Emerg Mkts Bd D	FI	10.68	13.57	1.25	3 星
RS Value A	DE	26.27	23.68	1.36	4 星
T. Rowe Price Latin Am.	IE	53.89	51.10	1.24	4 星
T. Rowe Price Mid Val	DE	22.46	16.91	0.80	4 星
Thornburg Value A	DE	37.53	15.46	1.27	4 星
USAA Income	FI	12.10	4.31	0.62	3 星
Vanguard Equity-Inc	DE	24.42	13.41	0.29	4 星
Vanguard Sht-Tm TE	FI	15.68	2.37	0.16	3 星
Vanguard Sm Cp Idx	DE	32.58	17.01	0.23	3 星
Wasatch Sm Cp Growth	DE	35.41	13.98	1.19	4 星

资料来源: Morningstar Funds 500 (2008)。

1.2.1 个体、变量和观测值

个体 (element) 是指搜集数据的实体。在表 1-1 的数据集中, 每一只可分割的共同基金是一个个体, 个体的名称列在表 1-1 中的第一列。有 25 只共同基金, 数据集中就有 25 个个体。

变量 (variable) 是个体中所感兴趣的那些特征。表 1-1 的数据集中有以下 5 个变量:

- 基金类型: 共同基金的类型, 标记为 DE (国内股本)、IE (国际股本) 和 FI (固定收益)。
- 净资产值 (美元): 2007 年 12 月 31 日的每股的收盘价。
- 5 年平均回报率 (%): 基金过去 5 年的平均年回报率。
- 费用比率 (%): 每财政年度扣除的基金费用占资产的百分比。
- 晨星评级: 每只基金整体风险调整星级, 晨星星级从最低的 1 星级到最高的 5 星级。

在一项研究中, 对每个个体的每个变量收集测量值, 从而得到了数据。对某一特定个体得到的测量值集合称为一个观测值 (observation)。如表 1-1 所示, 我们看到第一个观测值 (American Century Intl. Disc) 的测量值集合是 IE, 14.37, 30.53, 1.41 和 3 星。第二个观测值 (American Century Tax-Free Bond) 的测量值集合是 FI, 10.73,

3.34, 0.49 和 4 星。25 个个体的数据集有 25 个观测值。

1.2.2 测量尺度

搜集数据需要按下列几种测量尺度来度量：名义尺度、顺序尺度、间隔尺度和比率尺度。测量尺度决定了数据中蕴涵的信息量，并表明最适合的数据汇总和统计分析方法。

当一个变量的数据包含了用来识别个体属性的标记或名称时，测量尺度称为**名义尺度**（nominal scale）。例如表 1-1 中的数据，我们看到基金类型变量的测量尺度是名义尺度，因为 DE、IE 和 FI 是用来识别基金类别或类型的标记。在测量尺度是名义尺度的情形中，可以使用数字代码和非数字的标记。例如，为了便于数据搜集，并做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 代表国内股本；用 2 代表国际股本；用 3 代表固定收益。在这种情形下，数值 1, 2 和 3 代表基金类别。即使数据是数值型的，此时的测量尺度仍是名义尺度。

如果数据具有名义数据的性质，并且数据的顺序或等级的意义明确，那么这种变量的测量尺度是**顺序尺度**（ordinal scale）。例如，Eastside Automotive 向客户发放一份调查问卷，以征询客户对汽车维修服务质量方面的意见。每一位客户反馈一个维修服务质量的等级：优秀、好或差。得到的数据（优秀、好或差）是标记，具有名义数据的性质。另外，还可以按服务质量对数据进行排序或划分等级：优秀表示最好的服务质量，其次是好，然后是差。这样的测量尺度是顺序尺度。另一个例子，我们注意到，表 1-1 中的晨星评级数据也是顺序数据。它以晨星对基金风险调整后回报的评估为依据，给出了从 1 星到 5 星的等级。顺序尺度也可以用数字代码，例如你在学校中的班级排名。

如果数据具有顺序数据的所有性质，并且可以按一个固定的度量单位表示数值间的间隔，那么这种变量的测量尺度是**间隔尺度**（interval scale）。间隔数据永远是数值型的。学生能力测验（SAT）的分数是间隔尺度数据的一个例子。例如，三个学生的 SAT 分数分别为 620, 550 和 470，它们能够按最好到最差进行排序。另外，分数之间的差异是有意义的。比如，学生 1 的分数比学生 2 的分数多 $620 - 550 = 70$ （分），学生 2 的分数比学生 3 的分数多 $550 - 470 = 80$ （分）。

如果数据具有间隔数据的所有性质，并且两个数值之比是有意义的，则这种变量的测量尺度是**比率尺度**（ratio scale）。像距离、高度、重量和时间等变量都使用了测量的比率尺度。比率尺度需要有一个零值，变量取零值时表示什么也不存在。例如，让我们来考虑汽车的成本，零值意味着汽车没有成本或是免费的。另外，如果我们比较成本为 30 000 美元的汽车与成本为 15 000 美元的汽车，比率值 $30\,000/15\,000 = 2$ ，表示第一辆汽车的成本是第二辆汽车成本的 2 倍。

1.2.3 分类型数据和数量型数据

数据还可以进一步划分为分类型数据和数量型数据。归属于某一类别的数据称为**分类型数据**（categorical data）。分类型数据既可以用名义尺度度量也可以用顺序尺度度量。使用数量表示大小或多少的数据称为**数量型数据**（quantitative data）。数量型数据既可以用间隔尺度度量也可以用比率尺度度量。

分类变量（categorical variable）是用分类型数据表示的变量，**数量变量**（quantitative variable）是用数量型数据表示的变量。统计分析方法是否适合一个特定变量，取决于变量是分类变量还是数量变量^①。如果变量是分类变量，则统计分析方法是有限的。我们可以通过记录每一类别中观测值的数目，或计算每一类别中观测值的比例来汇总分类型数据。但是，即使分类型数据用数字代码表示，对其进行加、减、乘、除等数学运算也是没有意义

① 适合汇总数据的统计方法取决于数据是分类型数据还是数量型数据。

的。2.1 节将讨论汇总分类型数据的方法。

另一方面,对数量变量进行数学运算,可以得到有意义的结果。例如,对于数量型数据,可以先求和,然后除以观测值的个数,计算出平均数。这个平均数通常是有意义的并且易于解释。一般地,当数据是数量型时,有更多的统计方法可供选择。2.2 节和第3 章将介绍汇总数量型数据的方法。

1.2.4 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进行区分。**截面数据**(cross-sectional data)是在相同或近似相同的同一时点上搜集的数据。表1-1 中的数据是截面数据,因为它们描述了25 只共同基金的5 个变量在同一时点上的情况。**时间序列数据**(time series data)是在几个时期内搜集的数据。例如,图1-1 的时间序列是2006~2009 年美国常规普通汽油每加仑的平均价格曲线。我们注意到,夏季油价较高,在2008 年6 月创历史最高纪录,达到每加仑平均4.05 美元;2009 年1 月,油价急剧下降到每加仑1.65 美元,创3 年来的新低。

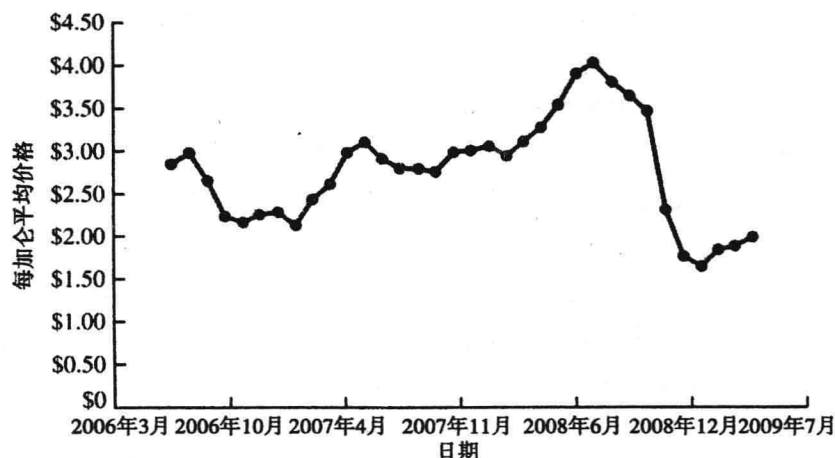


图1-1 美国常规普通汽油每加仑的平均价格

资料来源:美国能源部能源情报署,2009.5。

在商务与经济出版物中经常可以看到时间序列数据图形,这些图形可以帮助分析家们了解过去发生的情况,确定现象随时间变动的趋势,推测时间序列的未来水平。时间序列数据的图形有各种形态,如图1-2 所示。稍加研究,这些图形通常易于理解和解释。

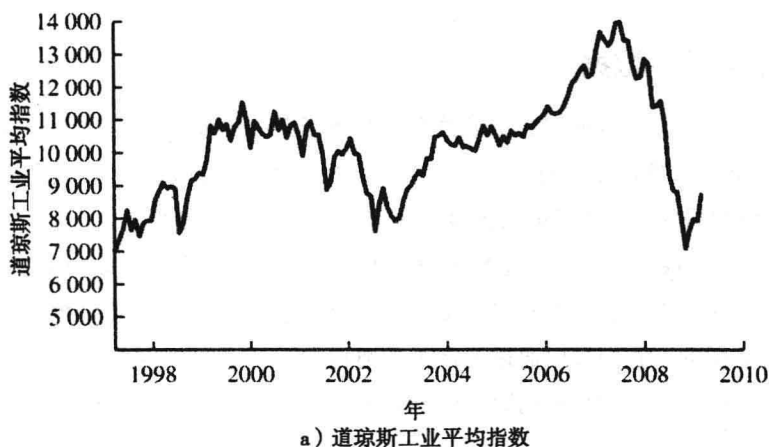
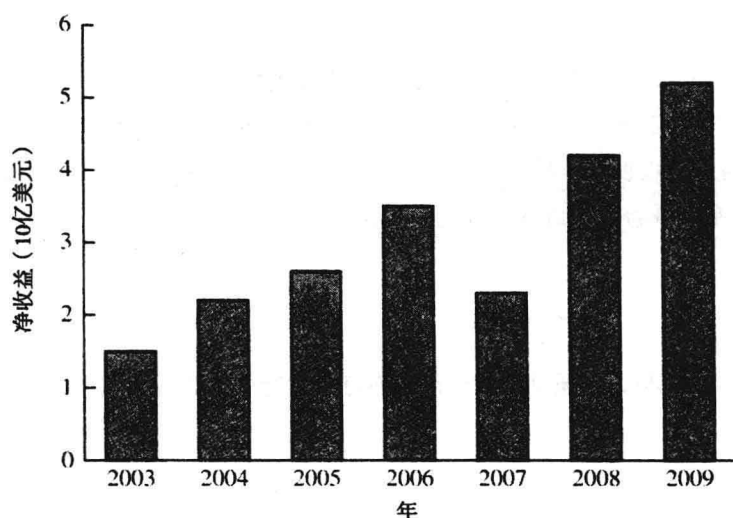
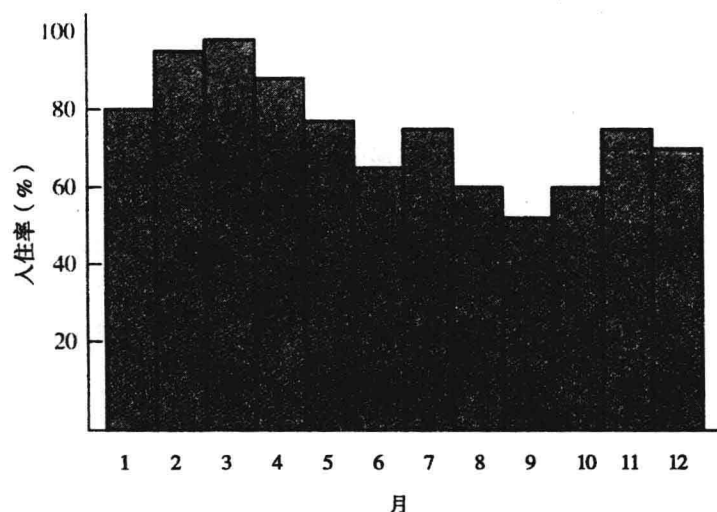


图1-2 时间序列数据的各种图形



b) 麦当劳公司的净收益



c) 佛罗里达州南部旅店的入住率

图 1-2 (续)

例如, 图 1-2a 是 1997 ~ 2009 年道琼斯工业平均指数的图形。1997 年 4 月普通股票市场指数接近 7 000 点。在随后的 10 年中指数不断攀升, 到 2007 年 6 月超过 14 000 点。然而, 在 2007 年创历史新高后, 股指急速下降。到 2009 年 3 月, 糟糕的经济条件迫使道琼斯工业平均指数重新回到 1997 年 7 000 点的水平。对投资者来说, 这是一个令人惊慌、失去信心的时期。2009 年 6 月, 指数有所复苏, 达到 8 700 点。

图 1-2b 是 2003 ~ 2009 年麦当劳公司的净收益图, 2008 年和 2009 年的经济衰退, 实际上对麦当劳公司而言是受益的, 其间其净收益创历史新高。麦当劳公司净收益的增长表明公司的繁荣是由于在经济衰退期人们减少了去相对昂贵的餐厅就餐, 转而寻求麦当劳提供更便宜的替代食品。

图 1-2c 是在一年期间佛罗里达州南部旅店入住率的一个时间序列图形。在 2 ~ 3 月, 当佛罗里达州南部的气候对游客有吸引力时, 出现 95% ~ 98% 的极高入住率。事实上, 1 ~ 4 月是佛罗里达州南部旅店典型的高入住率季节。另外, 8 ~ 10 月入住率较低, 最低的 50% 入住率出现在 9 月, 高温以及飓风季节是这段时间入住率下降的主要原因。

注释

1. 观测值是数据集中每个个体的测量值的集合。因此，观测值的个数总是与个体的个数一致。每个个体的测量值的个数等于变量的个数。因此，数据项的总数是个体的个数乘以变量的个数。
2. 数量型数据既可以是离散的也可以是连续的。度量可数事物多少的数量型数据是离散的（例如，5分钟内接到电话的个数），度量不可数事物多少的数量型数据是连续的（例如，体重或时间）。

1.3 数据来源

数据可以从现有的来源中搜集，或通过调查获得，亦可通过为了得到新数据而设计的实验性研究获得。

1.3.1 现有来源

在某些情况下，为某一特定应用所需要的数据已经存在。公司保存有关其雇员、客户和业务运作的各种数据库。关于雇员工薪、年龄和服务年限的数据通常可以从内部个人记录中获得；关于销售收入、广告支出、配送成本、库存水平以及产量的数据通常可以从其他内部记录中获得；绝大多数公司还保存其客户的详细资料。表1-2是一些从公司内部记录得到的常用的数据。

表 1-2 来自公司内部记录的可用数据的例子

数据来源	一些典型的可用数据
雇员记录	姓名、地址、社会保障号码、工资、休假天数、病假天数和奖金
生产记录	部件或产品号、生产数量、直接人工成本和原材料成本
存货记录	部件或产品号、库存单位数、重复订货水平、经济订货批量和折扣表
销售记录	产品号、销售量、区域销售量和分类客户销售量
信用记录	客户名称、地址、电话号码、信贷额度和应收账款余额
客户概况	年龄、性别、收入水平、家庭规模、地址和偏好

大量的商务和经济数据可以从专门搜集和保存数据的机构那里获得。公司可以通过租赁或购买的方式使用这些来自外部的数据。邓白氏、彭博社和道琼斯公司是三家能够为客户提供大量商业数据服务的公司。ACNielsen 和 Information Resources 公司也成功地为广告商和制造商提供商务数据搜集和加工服务。

数据也可以从各类行业协会和专门营利的机构中获得。美国旅游业协会保存了各州与旅游相关的信息，如游客数量和旅游费用等，旅游行业的公司和个人会对这些数据感兴趣。研究生管理咨询委员会保存了考试成绩、学生特长和研究生培养计划等数据。大多数来自这些类型的数据用户使用时需要支付适当的价格。

互联网的持续发展，已使其成为数据和统计信息的一个重要来源。几乎所有的公司都提供公司一般信息，如销售额、雇员人数、产量、产品价格和产品说明的网站。另外，大多数公司专门在互联网上获取有用的市场信息资料。因此，人们可以从网上查阅到股票报价、餐馆菜价、工资数据以及几乎无限的各种信息。

政府机构是现有数据的另一个重要来源。例如，美国劳工部保存了大量有关就业率、工资率、劳动力规模和工会会员等数据。表1-3列示了部分政府机构和它们能提供的一些数据。大部分政府机构也通过网站向社会提供其搜集和加工整理的数据。例如，美国人口普查局在其网站 www.census.gov 上有大量的数据。图1-3是人口普查局的主页。

表 1-3 来自选取政府机构的可用数据的例子

政府机构	一些可用的数据
人口普查局	人口数、家庭数和家庭收入的数据
联邦储备委员会	货币供应、信誉记录、汇率和贴现率的数据
管理和预算办公室	财政收入、支出和联邦政府债务的数据
商务部	商务活动数据、分行业总产值、分行业利润水平以及行业增长和下降的数据
劳工统计局	消费支出、小时工资、失业率、安全记录和国际统计资料

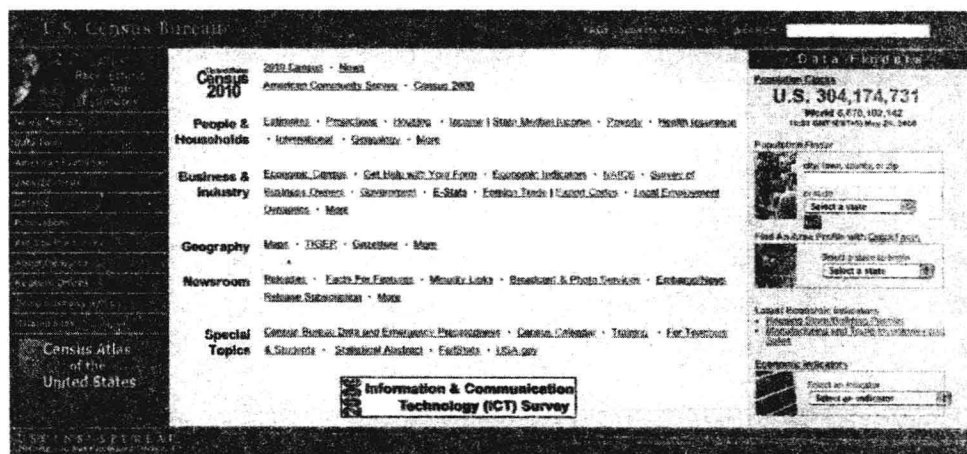


图 1-3 美国人口普查局主页

1.3.2 统计研究

有时一些特殊需要的数据并非总能从已有的来源中获得。在这种情况下，常常通过统计研究的方式获得数据。统计研究可划分为实验性的或观测性的两类。

在一项实验性的研究^①中，首先要确定感兴趣的主要变量，然后控制一个或多个其他变量，以便获得它们如何影响主要变量的数据。例如，一家制药公司可能会进行一项实验，以获得一种新药如何影响血压的情况。在研究中，血压是感兴趣的主要变量，新药的剂量是影响血压的另一个变量，实验希望能找出新药的剂量与血压之间的因果关系。为了获得有关新药疗效的数据，研究人员选择一些个体组成样本。控制新药的剂量，不同的个体组给予不同的药剂水平，然后搜集每组服药前后的血压数据。实验数据的统计分析将有助于我们了解新药如何影响血压。

在非实验性或观测性统计研究^②中，并不是有目的地去控制感兴趣的变量。调查也许是观测性研究中最常用的方法。例如，在一个征求顾客意见的调查中，首先要明确研究的问题，然后设计调查问卷，选择一些人作为样本。一些饭店利用观测研究来获取顾客对饭店的食品质量、服务和就餐环境等方面的数据。佛罗里达州那不勒斯的 Chops City Grill 饭店所使用的顾客意见问卷调查表如图 1-4 所示。注意，请顾客根据自己的感受对综合经验、接待问候、管理（餐桌巡视）、整体服务等 12 个变量按优秀、良好、平均、普通和较差的等级顺序打分，得到的数据能使 Chops City Grill 的管理人员保持高水平的饭店食品和服务的质量。

① 曾经进行过的最大规模的实验性统计研究是 1954 年美国公共卫生署的沙克脊髓灰质炎疫苗实验，该研究从全美 1~3 年级的小学生中选出了近 200 万名儿童。

② 对吸烟者和不吸烟者的研究是观测性研究，因为调查者并不确定或控制哪些人吸烟，哪些人不吸烟。

日期: _____

服务员姓名: _____

顾客是我们应予以最优先考虑的。请占用您一点点宝贵的时间填写这张调查表，这将使我们更好地为您提供服务。

您可将此表交到前台或邮寄给我们，谢谢！

调查项目	优秀	良好	平均	普通	较差
综合经验	☐	☐	☐	☐	☐
接待问候	☐	☐	☐	☐	☐
管理（餐桌巡视）	☐	☐	☐	☐	☐
整体服务	☐	☐	☐	☐	☐
职业素养	☐	☐	☐	☐	☐
菜单知识	☐	☐	☐	☐	☐
服务态度	☐	☐	☐	☐	☐
酒品选择	☐	☐	☐	☐	☐
菜单选择	☐	☐	☐	☐	☐
食品质量	☐	☐	☐	☐	☐
食品展示	☐	☐	☐	☐	☐
费用（美元）	☐	☐	☐	☐	☐

对改进我们的服务您还有何建议？

谢谢，非常感谢您的建议。——Chops City Grill 全体员工

图 1-4 佛罗里达州那不勒斯的 Chops City Grill 饭店的顾客意见问卷调查表

将数据和统计分析结果用于管理决策，管理人员必须清楚获得数据所必需的时间和成本。当必须在一个相对较短的时期内获得数据时，利用已有的数据来源是一个比较理想的方法；如果重要的数据不容易从已有的来源中获取，则必须考虑获取数据所要占用的时间和成本。在所有的情形中，决策者必须意识到统计分析对决策制定过程的贡献。数据搜集和统计分析所付出的成本不会超过利用这些信息制定一个好的决策所节省的费用。

1.3.3 数据采集误差

管理人员任何时候都应牢记：统计研究中的数据是可能产生误差的。使用错误的数据也许比不使用任何数据更糟糕。每当取得的数据值不等于真值或实际值时，就会产生数据搜集误差（真值或实际值可以通过一个正确的程序获得），这类错误可能在许多数据搜集方法中发生。例如，采访者可能发生记录错误，如将 24 岁写成 42 岁，或者被采访者回答问题时可能曲解了问题并做出不正确的回答。

实验数据分析人员在搜集和记录数据时要特别小心，以确保不发生错误。可以采取一些特殊措施以检查数据的内在一致性。例如，这种措施会提醒分析人员查看一下有关数据的准确性：如一个年龄为 22 岁的人却说有 20 年的工作经验。数据分析人员也应注意查看一些过大或过小的被称为异常值的数值，它们可能是错误的数

据。在第 3 章中我们将介绍一些统计学家用来识别异常值的方法。

在数据搜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搜集到的数据，将可能产生误差并导致做出错误的决策。因此，按照正确的步骤搜集准确的数据可以确保决策信息的可信度，提高数据的利用

价值。

1.4 描述统计

绝大多数的报纸、杂志、公司报告和其他出版物上的统计信息以读者易于理解的方式汇总和披露。将数据以表格、图形或数值的形式汇总的统计方法称为描述统计 (descriptive statistics)。

表 1-4 共同基金类型的频数和百分数

共同基金类型	频数	百分数 (%)	共同基金类型	频数	百分数 (%)
国内股本	16	64	固定收益	5	20
国际股本	4	16	总计	25	100

再次参见表 1-1 中 25 只共同基金的数据集。用描述统计的方法对这一数据集中的信息进行汇总。例如, 表 1-4 给出了基金类型这个分类变量数据的表格汇总, 图 1-5 给出了相同数据的图形汇总, 该类图形称为条形图。这些表格和图形汇总一般能使数据更容易解释。我们能很容易地从表 1-4 和图 1-5 看到, 绝大多数共同基金是国内股本类型。用百分数的形式表示, 64% 是国内股本类型, 16% 是国际股本类型, 20% 是固定收益类型。

图 1-6 是称为直方图的净资产值变量数据的图形汇总。从直方图上很容易看出, 净资产值的变动范围为 0 ~ 75 美元, 主要集中在 15 ~ 30 美元, 只有 1 只基金的净资产值超过 60 美元。

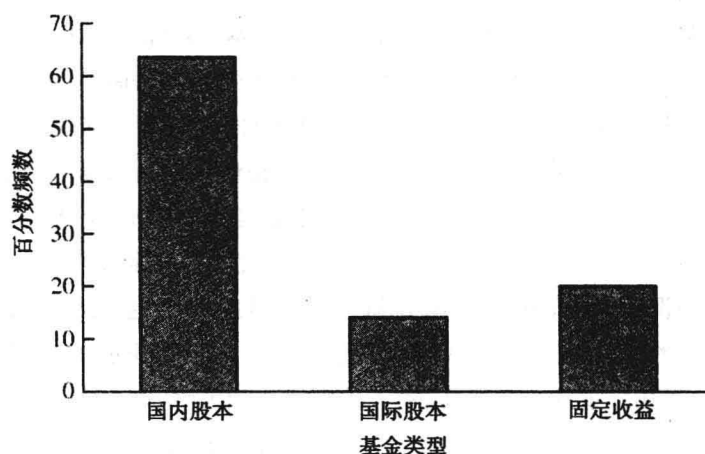


图 1-5 共同基金类型的条形图

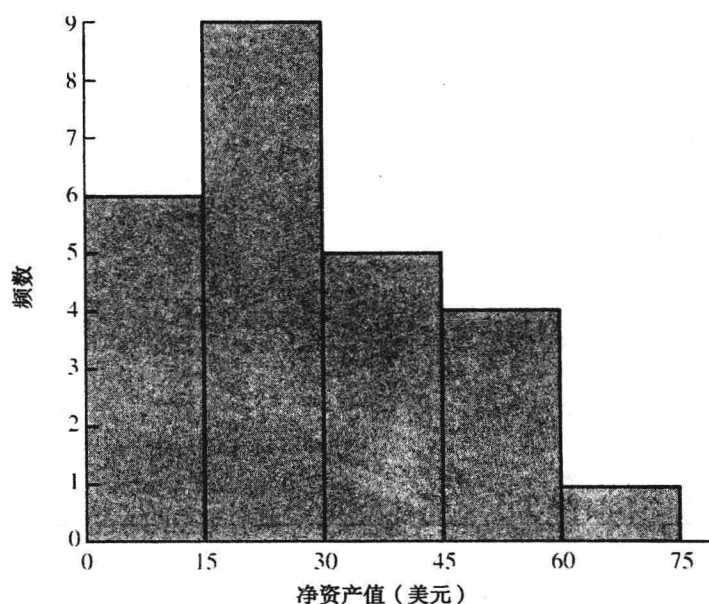


图 1-6 25 只共同基金净资产值的直方图

除了表格和图形表示方式以外,数值型描述统计也用于汇总数据。最常用的数值型描述统计是平均数或均值。利用表 1-1 的共同基金 5 年平均回报率数据,我们可以将所有 25 只共同基金的回报率相加再除以 25,计算出平均数。这样得到 5 年平均回报率的平均数为 16.50%。这个平均数是 5 年平均回报率变量数据的中心趋势或中心位置的度量。

近年来,人们关注描述统计的拓展和表述,对这类统计方法的兴趣在不断提高。在第 2 章和第 3 章,我们将集中讨论描述统计的表格、图形和数值方法。

1.5 统计推断

在很多情形下,我们需要有关大的个体组(个人、公司、选民、家庭、产品、客户,等等)的信息。出于对时间、成本和其他因素的考虑,只能从大的个体组的一小部分中搜集数据。在特定研究中,较大的个体组称为总体(population),较小的个体组称为样本(sample)。我们一般使用下列定义。

搜集总体全部数据的调查过程称为普查(census)^①,搜集样本数据的调查过程称为抽样调查(sample survey)。统计学的一个主要贡献就是利用样本数据对总体特征进行估计和假设检验,这一过程被称为统计推断(statistical inference)。

作为统计推断的一个例子,让我们考察由 Norris 电器公司所做的一项研究。Norris 电器公司生产一种能在各种电器产品上使用的高亮度灯泡。为了提高灯泡的使用寿命,产品设计部门开发出一种使用新型灯丝的灯泡。在这个例子中,总体定义为使用新灯丝生产的所有灯泡。为了估计新灯丝的优点,用生产出的 200 只新灯丝灯泡组成样本,并进行测试,记录每只灯泡在灯丝被烧断之前使用的小时数,样本数据见表 1-5。

总体

总体是在一个特定研究中所有个体组成的集合。

样本

样本是总体的一个子集。

表 1-5 Norris 电器公司 200 只灯泡样本的使用寿命

(单位:小时)

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

① 美国政府每 10 年进行一次人口普查,而市场调查机构每天进行抽样调查。

假设 Norris 电器公司希望利用样本数据对用新灯丝生产的所有灯泡总体的平均使用寿命进行推断。将表 1-5 中的 200 个数值相加再除以 200，可以计算出灯泡的样本平均使用寿命为 76 小时。我们可以用这一样本结果估计灯泡总体的平均使用寿命是 76 小时。图 1-7 是 Norris 电器公司统计推断过程的示意图。

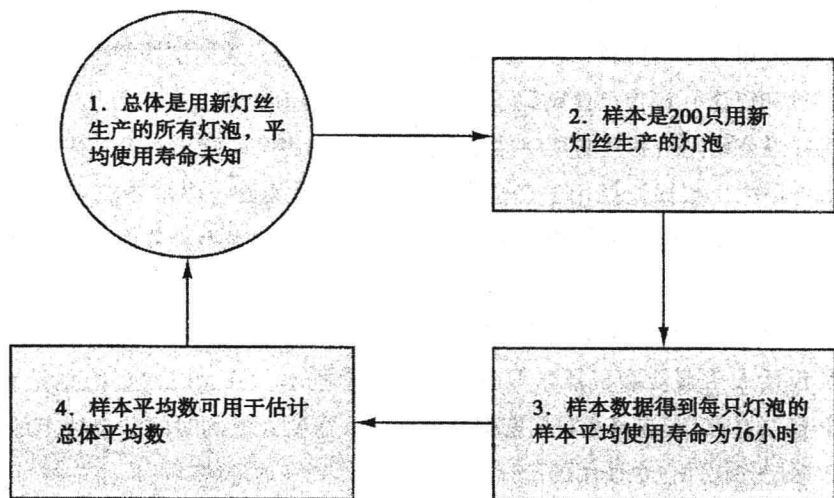


图 1-7 Norris 电器公司统计推断的过程

每当统计学家利用样本数据对我们感兴趣的总体特征进行估计时，他们通常都要注明估计的质量或精确度。以 Norris 电器公司为例，统计学家可能指出，新灯泡总体平均使用寿命的点估计值是 76 小时，误差边际是 ± 4 小时。这样，新灯泡总体平均使用寿命的区间估计是 72 ~ 80 小时。统计学家也可以指出，在 72 ~ 80 小时的区间内有多大的把握包含总体平均数。

1.6 计算机与统计分析

由于涉及大量的数据，因此统计人员经常使用计算机软件进行计算。例如，在 Norris 电器公司的例子中（见表 1-5），如果没有计算机，那么计算 200 只灯泡的平均使用寿命将相当乏味。为了方便计算机的使用，章末的附录还有使用 Minitab、Excel 和 Excel 的插件 StatTools 来完成本章所介绍的统计方法的详细步骤。

1.7 数据挖掘

由于有磁卡片输入机、条形码扫描仪和 POS 终端，许多机构每天可以获取大量的基础数据。即使一个当地的使用触摸屏显示器来登录订单和处理账单的小饭店，搜集大量的数据也是有意义的。对于大型零售公司，搜集数据的真值仅仅是概念性的，并且弄清楚如何有效地使用这些数据来提高效率也是一个难题。例如，大型零售商（如沃尔玛）每天获取 2 000 万 ~ 3 000 万个交易数据，电信公司（如法国电信和美国电话电报公司）每天有超过 3 亿个电话记录，Visa 每秒处理 6 800 次支付交易，或每天处理近 6 亿个交易。储存和管理这些交易数据是一件有意义的事情。

术语数据仓库是指获取、储存和维护数据的过程。计算能力和数据搜集工具可以在几秒钟内实现储存和检索非常庞大的数据。分析仓库中的数据可以对机构做出新的战略和较高收益的决策。

数据挖掘（data mining）问题是研究从非常大的数据库中开发出有用的决策信息的方法。利用统计学、数学和计算机科学的综合方法，分析人员在仓库中“挖掘数据”，并将其转化为有用的信息，取名为数据挖掘。数据