Statistics 21世纪统计学系列教材

Statistics with SPSS

统计学——基于SPSS

贾俊平 编著

中国人民大学出版社

Statistics 21世纪统计学系列教材

Statistics with SPSS

统计学——基于SPSS

贾俊平 编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计学：基于SPSS/贾俊平编著. —北京：中国人民大学出版社，2014.6
21世纪统计学系列教材
ISBN 978-7-300-19371-7

Ⅰ.①统… Ⅱ.①贾… Ⅲ.①统计学-高等学校-教材②统计分析-软件包-高等学校-教材 Ⅳ.①C8

中国版本图书馆CIP数据核字（2014）第122688号

21世纪统计学系列教材
统计学——基于SPSS
贾俊平 编著
Tongjixue：Jiyu SPSS

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京昌联印刷有限公司		
规　　格	185 mm×260 mm　16开本	**版　　次**	2014年7月第1版
印　　张	16.5插页1	**印　　次**	2014年7月第1次印刷
字　　数	352 000	**定　　价**	35.00元

前　言

在大数据时代，每天都会产生大量的数据，这些数据需要处理和分析。作为数据分析方法的统计学自然会受到越来越多的人关注，也会越来越广泛地应用于各个领域。难以想象，不使用计算机或统计软件如何处理和分析这些海量数据。

多数人都把统计学当作一门难学的课程来看待，原因之一就是对统计计算望而生畏，对复杂的统计公式望而却步。如果能从繁杂但属于简单劳动的计算中解脱出来，把它交给统计软件来完成，把统计计算统统“秒杀”，从而拿出更多的精力去理解统计方法的思想和原理，就会发现统计学不仅不像想象的那么难学，而且是一门非常有趣、非常有用的科学。

SPSS是最早引入到国内的优秀统计分析软件之一，以其视窗操作、易于使用和输出结果直观易懂等特点，被多数人广泛使用。目前，SPSS已有汉化版本，虽然汉化版本中许多术语和表述有不当之处，但还是会方便更多人使用。

本书是一本基于SPSS实现全部计算的统计学教材，书中例题的解答给出了SPSS的详细操作步骤。考虑到多数读者使用上的方便，本书使用的是SPSS 19.0中文版(建议有能力的读者使用英文版)。全书内容共11章，包括数据的描述性分析方法、推断方法以及实际中常用的一些统计方法等。每章均以一个实际问题引入该章要介绍的内容。在写法上完全立足于统计应用，避免统计公式的推导，力求通俗易懂。在形式上，本书给出了例题和练习题的数据文件，读者可通过扫描二维码下载。

本书可作为高等院校经济管理类专业本科生统计学课程的教材使用，也可作为其他文科专业及部分理、工、农、林、医、药专业的教材或参考书使用，对广大实际工作者也极具参考价值。由于作者水平有限，错误难免，希望读者在使用中对本书的不足之处多提宝贵意见，以便进一步修改和完善。

贾俊平

2014年1月于中国人民大学统计学院

教师教学服务说明

中国人民大学出版社工商管理分社以出版经典、高品质的工商管理、财务会计、统计、市场营销、人力资源管理、运营管理、物流管理、旅游管理等领域的各层次教材为宗旨。

为了更好地为一线教师服务，近年来工商管理分社着力建设了一批数字化、立体化的网络教学资源。教师可以通过以下方式获得免费下载教学资源的权限：

在“人大经管图书在线”（www. rdjg. com. cn）注册，下载“教师服务登记表”，或直接填写下面的“教师服务登记表”，加盖院系公章，然后邮寄或传真给我们。我们收到表格后将在一个工作日内为您开通相关资源的下载权限。

如您需要帮助，请随时与我们联络：

中国人民大学出版社工商管理分社

联系电话：010－62515735，62515749，82501704

传　　真：010－62515732，62514775　　电子邮箱：rdcbsjg@crup. com. cn

通讯地址：北京市海淀区中关村大街甲 59 号文化大厦 1501 室（100872）

教师服务登记表

姓 名		□先生 □女士	职　　称		
座机/手机			电子邮箱		
通讯地址			邮　　编		
任教学校			所在院系		
所授课程	课程名称	现用教材名称	出版社	对象（本科生/研究生/MBA/其他）	学生人数
需要哪本教材的配套资源					
人大经管图书在线用户名					

院/系领导（签字）：

院/系办公室盖章

目　录

第 1 章　数据与统计学 …… (1)
问题与思考：怎样理解统计结论？ …… (1)
1.1　统计学及其应用 …… (2)
1.1.1　什么是统计学 …… (2)
1.1.2　统计学的应用 …… (3)
1.2　数据及其来源 …… (5)
1.2.1　变量与数据 …… (6)
1.2.2　数据的来源 …… (7)
1.3　统计学与统计软件 …… (11)
主要术语 …… (13)
思考与练习 …… (14)

第 2 章　数据的描述性分析：图表展示 …… (16)
问题与思考：怎样用图表看数据？ …… (16)
2.1　类别数据的图表展示 …… (16)
2.1.1　用频数分布表观察类别数据 …… (17)
2.1.2　用图形展示类别数据 …… (20)
2.2　数值数据的图表展示 …… (22)
2.2.1　用频数分布表观察数据分布 …… (22)
2.2.2　用图形展示数值数据 …… (25)
2.3　使用图表的注意事项 …… (37)
主要术语 …… (38)
思考与练习 …… (38)

第 3 章　数据的描述性分析：概括性度量 …… (41)
问题与思考：怎样分析学生的考试成绩？ …… (41)
3.1　水平的描述 …… (42)
3.1.1　平均数 …… (42)
3.1.2　中位数和分位数 …… (42)
3.1.3　水平代表值的选择 …… (44)
3.2　差异的描述 …… (45)
3.2.1　极差和四分位差 …… (45)
3.2.2　方差和标准差 …… (46)
3.2.3　变异系数 …… (47)
3.2.4　标准得分 …… (49)
3.3　分布形状的描述 …… (51)
3.4　数据的综合描述 …… (51)
主要术语 …… (56)
思考与练习 …… (56)

第 4 章　随机变量的概率分布 …… (58)
问题与思考：彩票中奖的概率有多大？ …… (58)
4.1　什么是概率 …… (59)
4.2　随机变量的概率分布 …… (59)
4.2.1　随机变量及其概括性度量 …… (60)
4.2.2　随机变量的概率分布 …… (62)
4.2.3　其他几个重要的统计分布 …… (66)
4.3　样本统计量的概率分布 …… (70)
4.3.1　统计量及其分布 …… (70)
4.3.2　样本均值的分布 …… (71)
4.3.3　其他统计量的分布 …… (74)
4.3.4　统计量的标准误差 …… (74)
主要术语 …… (75)
思考与练习 …… (75)

第 5 章　参数估计 …… (77)
问题与思考：科学家做出重大贡献的最佳年龄是多少？ …… (77)
5.1　参数估计的基本原理 …… (78)
5.1.1　点估计与区间估计 …… (78)
5.1.2　评价估计量的标准 …… (81)
5.2　总体均值的区间估计 …… (83)

5.2.1 一个总体均值的估计 …… (83)
5.2.2 两个总体均值之差的估计 …… (86)
5.3 总体比例的区间估计 …… (91)
5.3.1 一个总体比例的估计 …… (91)
5.3.2 两个总体比例之差的估计 …… (93)
5.4 总体方差的区间估计 …… (95)
5.4.1 一个总体方差的估计 …… (95)
5.4.2 两个总体方差比的估计 …… (96)
5.5 样本量的确定 …… (97)
5.5.1 估计总体均值时样本量的确定 …… (97)
5.5.2 估计总体比例时样本量的确定 …… (99)
主要术语 …… (100)
思考与练习 …… (101)

第 6 章 假设检验 …… (104)
问题与思考：你相信饮用水瓶子标签上的说法吗？ …… (104)
6.1 假设检验的基本原理 …… (104)
6.1.1 怎样提出假设 …… (105)
6.1.2 怎样做出决策 …… (106)
6.1.3 怎样表述决策结果 …… (111)
6.2 总体均值的检验 …… (112)
6.2.1 一个总体均值的检验 …… (113)
6.2.2 两个总体均值之差的检验 …… (116)
6.3 总体比例的检验 …… (121)
6.3.1 一个总体比例的检验 …… (121)
6.3.2 两个总体比例之差的检验 …… (121)
6.4 总体方差的检验 …… (123)
6.4.1 一个总体方差的检验 …… (124)
6.4.2 两个总体方差比的检验 …… (125)
主要术语 …… (126)
思考与练习 …… (127)

第 7 章 类别变量分析 …… (130)
问题与思考：网购满意度与地区有关系吗？ …… (130)
7.1 一个类别变量的拟合优度检验 …… (130)
7.1.1 期望频数相等 …… (131)
7.1.2 期望频数不等 …… (133)

7.2 两个类别变量的独立性检验 …… (135)
7.2.1 列联表与 χ^2 独立性检验 …… (135)
7.2.2 应用 χ^2 检验的注意事项 …… (138)
7.3 两个类别变量的相关性度量 …… (138)
7.3.1 φ 系数和 Cramer's V 系数 …… (138)
7.3.2 列联系数 …… (139)
主要术语 …… (140)
思考与练习 …… (140)

第 8 章 方差分析 …… (143)
问题与思考：超市位置和竞争者数量对销售额有影响吗？ …… (143)
8.1 方差分析的基本原理 …… (144)
8.1.1 什么是方差分析 …… (144)
8.1.2 误差分解 …… (145)
8.1.3 方差分析的基本假定 …… (146)
8.2 单因子方差分析 …… (146)
8.2.1 数学模型 …… (146)
8.2.2 效应检验 …… (147)
8.2.3 多重比较 …… (151)
8.3 双因子方差分析 …… (153)
8.3.1 数学模型 …… (153)
8.3.2 主效应分析 …… (154)
8.3.3 交互效应分析 …… (162)
主要术语 …… (165)
思考与练习 …… (166)

第 9 章 一元线性回归 …… (169)
问题与思考：GDP 与消费水平有关系吗？ …… (169)
9.1 变量间的关系 …… (170)
9.1.1 确定变量之间的关系 …… (170)
9.1.2 相关关系的描述 …… (171)
9.1.3 关系强度的度量 …… (173)
9.2 一元线性回归模型的估计和检验 …… (175)
9.2.1 一元线性回归模型 …… (176)
9.2.2 参数的最小二乘估计 …… (177)
9.2.3 模型的拟合优度 …… (180)
9.2.4 模型的显著性检验 …… (183)

9.3 利用回归方程进行预测 …… (185)
9.3.1 平均值的置信区间 …… (185)
9.3.2 个别值的预测区间 …… (185)
9.4 用残差检验模型的假定 …… (188)
9.4.1 检验方差齐性 …… (188)
9.4.2 检验正态性 …… (189)
主要术语 …… (191)
思考与练习 …… (191)

第 10 章 多元线性回归 …… (195)
问题与思考：不良贷款受哪些因素影响？ …… (195)
10.1 多元线性回归模型 …… (196)
10.1.1 回归模型与回归方程 …… (196)
10.1.2 参数的最小二乘估计 …… (197)
10.2 拟合优度和显著性检验 …… (200)
10.2.1 模型的拟合优度 …… (200)
10.2.2 模型的显著性检验 …… (201)
10.3 多重共线性及其处理 …… (203)
10.3.1 多重共线性及其识别 …… (203)
10.3.2 变量选择与逐步回归 …… (205)
10.4 利用回归方程进行预测 …… (208)
10.5 哑变量回归 …… (210)
10.5.1 在模型中引入哑变量 …… (210)
10.5.2 含有一个哑变量的回归 …… (211)
主要术语 …… (216)
思考与练习 …… (217)

第 11 章 时间序列预测 …… (221)
问题与思考：如何预测社会消费品零售总额？ …… (221)
11.1 时间序列的成分和预测方法 …… (222)
11.1.1 时间序列的成分 …… (222)
11.1.2 预测方法的选择与评估 …… (225)
11.2 平稳序列的预测 …… (226)
11.3 趋势序列的预测 …… (229)
11.3.1 线性趋势预测 …… (229)
11.3.2 非线性趋势预测 …… (232)
11.4 多成分序列的预测 …… (236)

11.4.1 Winter 指数平滑预测 …… (237)
11.4.2 分解预测 …… (239)
主要术语 …… (243)
思考与练习 …… (243)

附录 SPSS 操作提示 …… (247)

参考文献 …… (252)

第 1 章

Chapter 1 数据与统计学

问题与思考：怎样理解统计结论？

每天我们都会看到各种统计数字或统计研究的某些结论。下面就是一些有趣的统计结论：

- 吸烟对健康是有害的，吸香烟的男性寿命减少 2 250 天。
- 不结婚的男性寿命会减少 3 500 天，不结婚的女性寿命会减少 1 600 天。
- 身体超重 30% 会使寿命减少 1 300 天。
- 每天摄取 500 毫升维生素 C，生命可延长 6 年。
- 身材高的父亲，其子女的身材也较高。
- 一项研究表明，杰出科学家做出重大贡献的最佳年龄在 25～45 岁之间，其最佳峰值年龄和首次贡献的最佳成名年龄随着时代的变化而逐渐增大。
- 学生们在听了 10 分钟莫扎特钢琴曲后做的推理，要比他们听 10 分钟其他娱乐性曲目后做得更好。
- 上课坐在前排的学生平均考试分数比坐在后排的高。
- 中国科学院空间环境研究预报中心的专家称，在神舟七号载人航天飞船飞行期间，遭遇空间碎片的概率在百万分之一以下。

这些结论是怎么得出的？你相信这些结论吗？你相信或不相信的理由是什么？要看懂这些结论似乎并不困难，但要合理解释这些结论就需要具备一定的统计学知识了。统计结论是一种归纳推理，这意味着不能肯定统计结论就一定正确。

在日常生活中，经常会接触到统计数据或一些统计研究结果。比如，在电视、报纸、网络等各种媒体中就会经常看到一些报道使用的统计数据、图表等。作为一门科学的统计学研究什么呢？怎样获得所需要的统计数据呢？这就是本章将要介绍的内容。

1.1 统计学及其应用

每个人都离不开统计，了解一些统计学知识对每个人来说都是必要的。比如，在外出旅游时，你需要关心一段时间内的详细天气预报；在投资股票时，你需要了解股票市场的价格信息，了解某只特定股票的有关财务信息；在观看足球比赛时，除了关心进球数之外，你还要知道各支球队的技术统计，等等。要正确阅读并理解统计数据或统计结论，需要具备一些统计学知识。

1.1.1 什么是统计学

在日常工作或管理中，总会面对各种各样的数据。如果不去分析这些数据，那它们也仅仅是一堆数据而已，没有太多的价值。如何分析这些数据，用什么方法分析数据，并从分析中得出某些结论以帮助我们做出决策，这正是统计学要解决的问题。简言之，**统计学**（statistics）是收集、处理、分析、解释数据并从数据中得出结论的原则和方法。统计学所提供的是一系列有关数据收集、处理和分析的方法。

数据收集就是取得所需要的数据。数据的收集方法可分为两大类：一是观察方法，二是实验方法。观察方法是通过调查或观测获得数据；实验方法是在控制实验对象条件下通过实验获得数据。

数据处理是对所获得的数据进行加工和处理，包括数据的计算机录入、筛选、分类和汇总等，以符合进一步分析的需要。

数据分析是利用统计方法对数据进行分析。数据分析所使用的方法大体上可分为**描述统计**（descriptive statistics）和**推断统计**（inferential statistics）两大类。描述统计主要是利用图表形式对数据进行展示，或通过计算一些简单的统计量（诸如比例、比率、平均数、标准差等）对数据进行分析。推断统计主要研究如何根据样本信息来推断总体的特征，内容包括参数估计和假设检验两大类。参数估计是利用样本信息推断所关心的总体特征，假设检验则是利用样本信息判断对总体的某个假设是否成立。比如，从一批灯泡中随机抽取少数几个作为样本，测出它们的使用寿命，然后根据样本灯泡的平均使用寿命估计这批灯泡的平均使用寿命，或者检验这批灯泡的使用寿命是否等于某个假定值，这就是推断统计要解决的问题。

数据解释是对分析结果进行的说明，包括结果的含义、从分析中得出的结论等。

统计学是一门关于数据的科学，它研究的是来自各领域的数据，提供的是一套通用于所有学科领域的获取数据、分析数据并从数据中得出结论的原则和方法。统计方法是通用于所有学科领域的，而不是为某个特定的问题或领域构造的。当然，统计方法和技术并不是一成不变的，使用者在给定的情况下必须根据所掌握的专业知识选择使用这些方法，而且如有需要还要进行必要的修正。

正如有的学者所指出的那样："统计学基本上是寄生的，靠研究其他领域内的工作而生存。这不是对统计学的轻视，这是因为对很多寄主来说，如果没有寄生虫就会死。对有的动物来说，如果没有寄生虫就不能消化它们的食物。因此，人类奋斗的很多领域，如果没有统计学，虽然不会死亡，但一定会变得很弱。"① 看上去统计似乎被边缘化了，但实际上正说明了统计在各学科领域的独特地位和作用，也表明了统计作为一门独立学科而具有的特点。

1.1.2　统计学的应用

说出哪些领域应用统计，这很困难，因为几乎所有的领域都用统计；说出哪些领域不用统计，同样也很困难，因为几乎找不到一个不用统计的领域。可以说，统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。只要有数据的地方就会用到统计方法。

1. 统计学的应用领域

统计学被广泛应用于各个学科领域，为各学科的发展做出了重要贡献。这里，我们不想列举统计学的应用领域，只想通过几个简单的例子说明统计学的应用。

例 1—1

用统计识别作者。1787—1788 年，三位作者亚历山大·汉密尔顿（Alexander Hamilton）、约翰·杰伊（John Jay）和詹姆斯·麦迪逊（James Madison）为了说服纽约人认可宪法，匿名发表了著名的 85 篇论文。这些论文中的大多数作者已经得到了确认，但是，其中的 12 篇论文的作者身份引起了争议。通过对这些论文不同单词的频数进行统计分析，得出的结论是，詹姆斯·麦迪逊最有可能是这 12 篇论文的作者。现在，对于这些存在争议的论文，认为詹姆斯·麦迪逊是原创作者的说法占主导地位，而且几乎可以肯定这种说法是正确的。

例 1—2

用简单的描述统计量得到一个重要发现。费希尔（R. A. Fisher）在 1952 年的一篇文章中举了一个例子，说明如何由基本的描述统计量知识引出一个重要的发现。20 世纪早期，哥本哈根卡尔堡实验室的施密特（J. Schmidt）发现在不同地区捕获的同种鱼类的脊椎骨和鳃线的数量有很大不同，甚至在同一海湾内不同地点所捕获的同种鱼类也有这样的倾向。然而，鳗鱼的脊椎骨数量变化不大。施密特在从欧洲各地、冰岛、亚速尔群岛以及尼罗河等几乎分离的海域里所捕获的鳗鱼的样本中，计算发现了几乎一样的均值和标准偏差值。由此，施密特推断所有各个不同海域内的鳗鱼是由海洋中某公共场所繁殖的。后来名为"戴纳"（Dana）的科学考察船在一次远征中发现

① C. R. 劳：《统计与真理——怎样运用偶然性》，北京，科学出版社，2004。

了这个场所。

例 1—3

挑战者号航天飞机失事预测。1986 年 1 月 28 日清晨，载有 7 名宇航员的挑战者号进入发射状态。就在发射前，有冰片牢附在机壳上。几分钟后，正当电视新闻报道它已进入轨道时，航天飞机在毁灭性的爆炸声中化为碎片，机上的宇航员全部遇难。推动航天飞机进入太空的两个固体燃料发动机是由莫顿·塞奥科公司（Morton Thiokol）制造的。失事前一天晚上，莫顿·塞奥科公司的经理们和美国宇航局（NASA）就飞机如期发射还是推迟发射产生了争执。天气预报预测发射时的气温为 31°F（华氏度）。争执的结果是采纳了莫顿·塞奥科公司经理们的建议：按计划发射航天飞机，因为他们觉得没有确凿证据表明低温会对固体燃料火箭推进器的性能产生影响。在此次失事前，该航天飞机已发射成功 24 次。将航天飞机送入太空的两个固体燃料推进器有 6 只 O 型项圈密封，在几次飞行中，曾发生过 O 型项圈被腐蚀或气体泄漏事故，这样的事故是极其危险的。前 24 次发射中有一次发动机遭到了永久性破坏。下面的表 1—1 是 23 次飞行中 O 型项圈损坏的个数（因变量 y）及发射时火箭连接处的温度（自变量 x）数据。

表 1—1　　挑战者号航天飞机 23 次飞行中损坏的 O 型项圈个数和发射时的温度

飞行次数	O 型项圈的损坏个数	温度（°F）	飞行次数	O 型项圈的损坏个数	温度（°F）
1	2	53	13	1	70
2	1	57	14	1	70
3	1	58	15	0	72
4	1	63	16	0	73
5	0	66	17	0	75
6	0	67	18	2	75
7	0	67	19	0	76
8	0	67	20	0	78
9	0	68	21	0	79
10	0	69	22	0	81
11	0	70	23	0	76
12	0	70			

根据表 1—1 中的数据进行线性回归，得到的回归方程为 $\hat{y}=3.698-0.047\,54x$。由此得到当温度为 31°F 时，O 型项圈发生事故的预计个数为 2.225 个。结果显示连接处的温度与 O 型项圈事故之间有一定的相关性。如果当时那些经理们看到了回归的预测结果，也许会推迟发射。

前两个是统计得以应用并取得成效的例子，后一个是统计结果未被采纳而酿成惨剧的例子。不管怎样，它们都表明统计在许多领域都有广泛的应用。

2. 统计的误用与滥用

大约在一个世纪以前，政治家本杰明·迪斯雷利（Benjamin Disraeli）曾有一个

著名的论断："谎言有三种：谎言、糟透的谎言和统计。"统计常常被人们有意或无意地滥用，比如错误的统计定义、错误的图表展示、不合理的样本、数据的篡改或造假，等等。这些误用有些是常识性的，有些是技术性的，有些则是故意的。作为从数据中寻找事实的统计，却被有些人变成了歪曲事实的工具。你也许常常看到这样的产品质检报告：某某产品的抽样合格率是 80%。乍看上去还可以，但如果实际上只抽查了 5 件产品，有 4 件合格，这样的合格率能说明什么问题呢？在马路上随便采访几个人，他们的看法能代表大多数人的观点吗？"调查结果表明……"调查了多少个人？是随机调查的吗？样本是怎样选取的？这看上去是在用事实说话，实际上成了统计陷阱。

在管理领域，统计也往往被作为两个极端使用：一个极端是复杂问题简单化，一些不懂或不太懂统计的人认为统计没什么用，他们因为不懂统计而看不起统计，他们不用或几乎不用统计方法分析数据，即使做些统计分析，也往往是表面上的。走入这一极端的人，他们决策的依据就是自己大脑中一些杂乱无章的信息组合出的某种直觉。如果他们的决策是正确的，更增加了他们的自信，更加感到不用统计也挺好；如果他们的决策出了毛病，便会找出一大堆推脱的理由：市场难测、环境突变、竞争激烈、需求疲软、价格下跌、管理不善、成本上升、出口下降……另一个极端是把简单问题复杂化，特别是在管理领域。一些管理者把本来可以用简单方法解决的问题故意复杂化，他们不用简单的分析方法，而用复杂的分析方法；他们为证明管理的科学性，建立一个别人看不懂的模型，编一大堆程序，输出一大堆数字和符号；他们得出用统计语言陈述的结论，提出一些似是而非的建议……这样的分析往往是既脱离了管理问题，对实际决策也未必有用。在管理中，这两个极端都是不可取的。管理决策中不用统计几乎不可想象，把简单问题复杂化对管理决策也未必有用。从统计的实际应用来看，简单的方法不一定没用，复杂的方法也不一定有用。统计应该被恰当地应用到它能起作用的地方。不能把统计神秘化，更不能歪曲统计，把统计作为掩盖事实的陷阱。

曲解统计是一种常见现象。在有些人看来，使用统计就是寻找支持：在他们的心目中可能早已有了某种"结论"性的东西，或者说他们希望看到符合他们需要的某种结论，便去找些数据来支持他们的结论。如果数据分析的结果与他们预期的结论一致，他们就会声张自己是用科学方法得到的结论；如果与预期的不一致，他们要么会篡改数据，要么对统计弃而不用。这恰恰背离了数据分析的本质。数据分析的真正目的是从数据中找出结论，从数据中寻找启发，而不是寻找支持。真正的数据分析事先是没有结论的，通过对数据的分析才得出结论。

1.2　数据及其来源

统计分析离不开数据，没有数据统计方法就成了无米之炊。数据是什么？怎样获

得所需的数据？这就是本节将要介绍的内容。

1.2.1 变量与数据

观察一个企业的销售额，你会发现这个月和上个月有所不同；观察股票市场上涨股票的家数，今天与昨天数量不一样；观察一个班学生的生活费支出，一个人和另一个人不一样；投掷一枚骰子观察其出现的点数，这次投掷的结果和下一次也不一样。这里的“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等就是变量。简言之，**变量**（variable）是描述观察对象某种特征的概念，其特点是从一次观察到下一次观察可能会出现不同结果。变量的观测结果就是**数据**（data）。

根据观测结果的特征，变量可以分为类别变量和数值变量两种。

类别变量（categorical variable）是取值为事物属性或类别以及区间值的变量，也称**分类变量**（classified variable）或**定性变量**（qualitative variable）。比如，观察人的性别、公司所属的行业、用户对商品的评价时，得到的结果就不是数字，而是事物的属性。例如，观测性别的结果是“男”或“女”，公司所属的行业为“建造业”、“零售业”、“旅游业”等；用户对商品的评价为“很好”、“好”、“一般”、“差”、“很差”。人的性别、公司所属的行业、用户对商品的评价等作为变量取的值不是数值，而是事物的属性或事物的类别。此外，考虑学生月生活费支出的档次可能分为1 000 元以下、1 000～1 500 元、1 500～2 000 元、2 000 元以上 4 档，变量“月生活费支出档次”的这 4 档取值也不是普通的数值，而是数值区间，因而变量也称为区间值类别变量。人的性别、公司所属的行业、用户对商品的评价、学生月生活费支出的档次等都是类别变量。

类别变量根据取值是否有序通常分为两种：**名义**（nominal）**值类别变量和顺序**（ordinal）**值类别变量**。名义值类别变量也称**无序类别变量**，其取值是不可以排序的。比如“公司所属的行业”这一变量的取值为“建造业”、“零售业”、“旅游业”等，这些取值之间不存在顺序关系。又如“商品的产地”这一变量的取值为甲、乙、丙、丁，这些取值之间也不存在顺序关系。顺序值类别变量也称**有序类别变量**，其取值可以排序。例如“对商品的评价”这一变量的取值为“很好”、“好”、“一般”、“差”、“很差”，这 5 个值之间是有顺序的。取区间值的变量当然是有序的类别变量。当类别变量只取两个值时也称**二值**（binary）**类别变量**，例如“性别”这一变量的取值为“男”和“女”。二值变量可以看做名义变量，也可以看做有序变量。

类别变量的观测结果称为**类别数据**（categorical data）。类别数据也称为**分类数据**或**定性数据**。与类别变量相对应，类别数据相应分为名义值类别数据和顺序值类别数据两种。其中只取两个值的类别数据也称为**二值类别数据**。

数值变量（metric variable）是取值为数字的变量，也称为**定量变量**（quantitative variable）。例如“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚