



IBM SPSS

张文彤 钟云飞 编著

数据分析 与挖掘实战案例精粹

● 业内资深专家十余年实战经验总结，从上千个真实案例中精选出18个案例，帮助读者迅速成长为真正的数据分析与挖掘高手！

● 初学者的入门向导，进阶者的成才之路，实战专家的案例指南。

● “做中学”是学习统计分析的最优路径，翔实的案例，丰富的细节，助你快速上手数据分析！



清华大学出版社

1545210

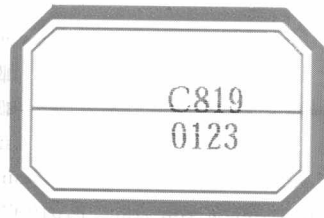


CS1706103

IBM SPSS 数据分析与 挖掘实战案例精粹

张文彤 钟云飞 编 著

C819
0123



重庆师大图书馆

清华大学出版社

北京

内 容 简 介

全书以 IBM SPSS Statistics 20.0 和 IBM SPSS Modeler 14.1 为工具, 提供了医疗、金融、保险、汽车、快速消费品、市场研究、互联网等多个行业的数据分析/挖掘案例, 基于实战需求, 详细讲解整个案例的完整分析过程, 并将模型和软件的介绍融于案例讲解之中, 使读者在阅读时能突破方法和工具的局限, 真正聚集于对数据分析精髓的领悟。本书所附光盘包括案例数据和分析程序/流文件, 读者可完整重现全部的分析内容。

本书适合从初学者到专家各个级别的数据分析人员阅读, 尤其适合于以下读者群: 需要提升实战能力的数据分析专业人员; 在市场营销、金融、财务、人力资源管理中需要应用数据分析的人士; 从事咨询、科研等工作的专业人士; 同时也适合于各专业的本科和研究生作为学习数据分析应用的参考书。

本书封面贴有清华大学出版社防伪标签, 无标签者不得销售。
版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

IBM SPSS 数据分析与挖掘实战案例精粹/张文彤, 钟云飞编著. --北京: 清华大学出版社, 2013
ISBN 978-7-302-29954-7

I. ①I… II. ①张… ②钟… III. ①统计分析—软件包 IV. ①C819

中国版本图书馆 CIP 数据核字(2012)第 209359 号

责任编辑: 李玉萍 桑任松

封面设计: 杨玉兰

责任校对: 周剑云

责任印制: 沈露

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社总机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

课 件 下 载: <http://www.tup.com.cn>, 010-62791865

印 装 者: 北京市清华园胶印厂

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 32.25 字 数: 783 千字

(附 DVD1 张)

版 次: 2013 年 2 月第 1 版 印 次: 2013 年 2 月第 1 次印刷

印 数: 1~4000

定 价: 64.00 元

前 言

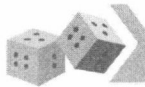
所谓艺术，就是指如果只靠系统地学习既有知识体系，但自身不具备相当的天赋，或者没有经过长期实践以积累经验和激发灵感，始终难以登堂入室成为大师的那些学科；音乐、舞蹈、绘画等就是如此。大英百科全书就把统计学定义为：一门收集数据、分析数据，并根据数据进行推断的艺术和科学。显然，作为一门应用学科，统计学非常强调实战能力。一名出色的统计师需要通过经历各种各样的实战分析项目来吸取经验、教训以便持续成长。光靠操作教科书上那些标准案例，他只能成为工匠，而不能成为大师。

近年来，随着计算机技术的飞速发展，统计工具出现了日新月异的变化，大大提高了其可用性。统计学和数据库技术、人工智能技术的融合，更是进一步催生了数据挖掘这个目前炙手可热，也更强调实战能力的领域。具体到 SPSS 系列产品，随着 IBM 的并购，原先的 SPSS 软件已经成为 IBM SPSS Statistics，它定位于标准的统计分析需求，而更贴近企业用户的数据分析与挖掘的需求则由 IBM SPSS Modeler 来满足。分析工具的高度易用性和实战需求的同步发展，使得各行各业对统计分析和数据挖掘人员的需求呈现爆炸性增长，远远超过了正常培养周期能够提供的数量，而广大统计分析人员也迫切希望能够得到的是本讲解提高实战操作技能的书，而不是单纯以介绍某一种统计软件为目的的参考书，以便帮助自己迅速提升实战能力。因此，笔者便有了编写这样一本书的打算。

笔者先后于 2000 年、2002 年和 2004 年编写过三轮 SPSS 教程/参考书，均获得了读者的好评。作为在数据分析领域从业十余年的统计专业人员，本书的作者深知在漫长的经验积累阶段所需要付出的努力和汗水，更能体会到编写一本实战案例书的市场价值。虽然作者从业以来经手的分析案例有上千个，但很多优秀案例都因涉及相应公司的业务机密而无法和读者分享。而且案例的复杂程度和代表性也颇费思量，过于复杂会牵扯太多的具体业务细节，影响案例的可读性，而案例过于简单，则无法展示实战分析中可能遇到的各种情况，参考价值不大。在反复讨论之后，笔者最终决定编写此书，因为这件事情有利于推动数据分析行业的发展，非常值得去做。

本书定位为实战类书籍，分为 4 个部分，共 20 章(不包括附录部分)，以 IBM SPSS Statistics 20.0 和 IBM SPSS Modeler 14.1 为准，完全从实际案例的分析需求出发，讲解各类方法的综合运用和实战操作，本书的具体特点如下。

- 行业实战：以案例集的方式提供医疗、电信、金融、零售、市场研究等行业的真实案例，完全从实际项目的分析需求出发，讲解各类方法的综合运用，使本书更贴近数据分析实战，更具参考价值。



- **内容全面**: 同样是从实战需求出发, 不再拘泥于常规统计方法, 也不再拘泥于 IBM SPSS Statistics 一个软件, 而是基于实际应用的需求, 随时使用各种 SPSS 软件中的新功能、新技巧, 必要时进一步引入 Modeler 来解决各种数据挖掘的具体应用, 从而在实际案例的背景下, 使读者充分了解 IBM SPSS 系列产品的强大功能。
- **易学易用**: 以实用性为唯一标准, 结合笔者多年的统计教学经验和现在的商业应用经验, 重点讲解实战分析应用, 案例的安排顺序从简到繁, 将软件操作的讲解自然融入案例分析过程中, 使读者的学习过程更加自然流畅。
- **案例重现**: 本书附带光盘中包含书中涉及的完整案例数据、案例实现程序和 Modeler 数据流, 并提供 IBM SPSS Statistics 和 IBM SPSS Modeler 试用软件的下载网址, 读者可以在学习时利用试用软件同步完整重现所有的分析过程和结果, 彻底避免纸上谈兵的尴尬。

对不同的读者群, 他们可以从书中学到以下知识和技能。

- **软件入门**: 对 IBM SPSS Statistics 和 IBM SPSS Modeler 新用户而言, 本书显然是最佳的学习软件操作和实战技能的教科书。本书采用相应软件的较新版本, 就统计分析和数据挖掘项目中的一些典型案例进行了深入浅出的介绍, 读者只需要按照讲解顺序操作, 就可以真正掌握相应的数据分析实战操作技能。
- **技能提升**: 对已经熟悉相应 SPSS 系列产品如何使用的老用户而言, 本书则是读者渴望多年的专家教程。笔者在案例中真正展示的并非简单的软件操作, 而是完整的统计思维和实战分析思路, 已有数据分析基础的读者通过对这些案例的学习, 能够更快地跨越从理论到实战的鸿沟, 从而使自身对软件工具的掌握和实战操作能力都得到真正的提升。
- **触类旁通**: 对资深的统计分析和数据挖掘人员而言, 其对分析工具的应用早已超越了个别产品的层面, 达到“不滞于物, 草木竹石皆可为剑”的地步, 但本书仍然具有很高的参考价值, 因为软件仅仅是实现工具, 其背后的统计思维、统计方法、基本原则等完全相同, 但不同的人在面对相同问题时所采用的分析流程、处理方法等各有千秋, 通过对书中案例的学习、参照和比较, 分析人员能够举一反三, 从而真正对实战操作达到“悟”的境界。

本书第 1 章由张文彤和钟云飞共同编写, 第 5 章和第 17~20 章由钟云飞编写, 第 4 章和第 16 章由王清华编写, 其余各章由张文彤编写。

作者新浪微博: @文彤老师、@数里寻道、@AllanVV。

读者交流微群: <http://q.weibo.com/749521>。

软件试用版下载: <http://peixun.pinggu.org/SPSSCaseBookDVD.zip>。

本书案例数据、内容更新下载: <http://www.StatStar.com>。



在本书的写作、出版、发行过程中，我们得到了 IBM 大中华区业务分析软件总经理 缪可延、IBM 大中华区业务分析软件技术经理周庆伟、IBM 大中华区商业智能及预测分析软件销售经理刘海亮、IBM 华西区市场经理邓宏等多位 IBM 领导与同事的鼓励、帮助与支持，人大经济论坛则为本书提供了试用软件的下载空间，这里一并表示由衷的感谢。

希望本书能够帮助读者更加深入地了解数据分析，进一步促进数据分析在国内的普及。也希望广大读者踊跃提出自己的宝贵意见和建议，使本书再版时能够更加完善。

编 者

目 录

第一部分 SPSS 数据分析基础

第 1 章 数据分析方法论简介3	2.3.2 单变量假设检验方法..... 26
1.1 三种数据分析方法论.....3	2.3.3 双变量假设检验方法..... 28
1.1.1 严格设计支持下的统计 方法论.....3	2.4 多变量模型..... 31
1.1.2 半试验研究支持下的统计 方法论.....4	2.4.1 方差分析/一般线性模型..... 31
1.1.3 偏智能化、自动化分析的 数据挖掘应用方法论.....5	2.4.2 广义线性模型和混合线性 模型..... 32
1.2 CRISP-DM 方法论介绍.....6	2.4.3 回归模型..... 34
1.2.1 概述.....6	2.4.4 其他常见模型..... 36
1.2.2 商业理解.....8	2.5 多元统计分析模型..... 38
1.2.3 数据理解.....8	2.5.1 信息浓缩..... 38
1.2.4 数据准备.....9	2.5.2 变量间内在关联结构的 探讨..... 38
1.2.5 建立模型.....9	2.5.3 数据分类..... 39
1.2.6 模型评价.....9	2.5.4 分析元素间的关联..... 41
1.2.7 结果部署.....10	2.6 智能统计分析/数据挖掘方法..... 42
第 2 章 数据分析方法体系简介11	2.6.1 树模型..... 42
2.1 统计软件中的数据存储格式.....11	2.6.2 神经网络..... 43
2.1.1 二维数据表.....11	2.6.3 支持向量机..... 43
2.1.2 变量的存储类型.....12	2.6.4 贝叶斯网络..... 44
2.1.3 变量的测量尺度.....12	2.6.5 最近邻元素分析..... 44
2.2 数据的统计描述与参数估计.....13	2.6.6 关联规则与序列分析..... 44
2.2.1 连续变量的统计描述.....13	第 3 章 IBM SPSS Statistics 操作
2.2.2 连续变量的参数估计.....16	入门 46
2.2.3 分类变量的统计描述和参数 估计.....18	3.1 案例背景..... 46
2.2.4 统计图形体系.....21	3.2 数据文件的读入与变量整理..... 47
2.3 常用假设检验方法.....24	3.2.1 SPSS 的基本操作界面..... 47
2.3.1 假设检验的基本原理.....25	3.2.2 数据准备..... 49
	3.3 问卷数据分析..... 53
	3.3.1 生成频数表..... 53
	3.3.2 计算均值..... 54



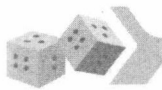
3.3.3 对多选题进行描述.....	55	5.2 IBM SPSS Modeler 相关操作	
3.4 项目总结和讨论.....	56	与技巧.....	77
第4章 IBM SPSS Statistics 操作进阶	57	5.2.1 IBM SPSS Modeler 的基本操作.....	77
4.1 案例背景.....	57	5.2.2 IBM SPSS Modeler 中的表达式.....	79
4.1.1 项目背景.....	57	5.2.3 IBM SPSS Modeler 的若干使用技巧.....	79
4.1.2 分析思路.....	59	5.3 IBM SPSS Modeler 功能介绍.....	81
4.2 问卷录入.....	59	5.3.1 数据整理案例.....	81
4.2.1 开放题的定义.....	59	5.3.2 探索性数据分析案例.....	82
4.2.2 单选题的定义.....	60	5.3.3 建立模型、模型检验与模型应用案例.....	83
4.2.3 多选题的定义.....	60	5.4 案例分析：药物选择决策支持.....	86
4.3 问卷质量校验.....	62	5.4.1 背景介绍.....	86
4.3.1 去除重复记录.....	62	5.4.2 数据说明.....	86
4.3.2 发现异常值.....	64	5.4.3 商业理解.....	87
4.3.3 逻辑校验.....	65	5.4.4 数据理解.....	87
4.4 问卷数据分析.....	67	5.4.5 数据准备.....	88
4.4.1 问卷加权.....	67	5.4.6 模型建立和评估.....	89
4.4.2 业务分析.....	70	5.4.7 模型发布.....	91
4.5 项目总结和讨论.....	71	5.5 如何进一步学习 IBM SPSS Modeler.....	93
第5章 IBM SPSS Modeler 操作入门	73		
5.1 IBM SPSS Modeler 概述.....	73		
5.1.1 IBM SPSS Modeler 的界面.....	73		
5.1.2 IBM SPSS Modeler 的架构与产品构成.....	76		

第二部分 影响因素发现与数值预测

第6章 酸奶饮料新产品口味测试研究案例	97	6.2.3 均值的图形描述.....	101
6.1 案例背景.....	97	6.3 不同品牌的评分差异分析.....	102
6.1.1 研究项目概况.....	97	6.3.1 单因素方差分析模型简介.....	103
6.1.2 分析思路与商业理解.....	98	6.3.2 品牌作用的总体检验.....	104
6.2 数据理解.....	98	6.3.3 组间两两比较.....	105
6.2.1 研究设计框架复查.....	98	6.3.4 方差齐性检验.....	108
6.2.2 均值的列表描述.....	99	6.4 两因素方差分析模型分析.....	108
		6.4.1 两因素方差分析模型简介.....	109



6.4.2	拟合包括交互项的饱和模型.....110	8.3.3	模型拟合效果的判断.....146
6.4.3	拟合只包含主效应的模型.....111	8.3.4	存储预测值和区间估计值.....148
6.4.4	组间两两比较.....112	8.4	曲线拟合.....148
6.4.5	随机因素分析.....114	8.4.1	用曲线估计过程同时拟合多个曲线模型.....149
6.5	分析结论与讨论.....116	8.4.2	模型拟合效果的判断.....151
6.5.1	分析结论.....116	8.4.3	模型的预测.....153
6.5.2	Benchmark: 用还是不用.....116	8.5	利用非线性回归进行拟合.....154
第7章	偏态分布的激素水平影响因素分析.....118	8.5.1	模型简介.....154
7.1	案例背景.....118	8.5.2	构建分段回归模型.....155
7.1.1	研究项目概况.....118	8.5.3	不同模型效果的比较.....157
7.1.2	分析思路与商业理解.....119	8.6	项目总结与讨论.....158
7.2	数据理解.....119	8.6.1	分析结论.....158
7.2.1	单变量描述.....119	8.6.2	行走在理想与现实之间.....158
7.2.2	变量关联探索.....122	第9章	脑外伤急救后迟发性颅脑损伤影响因素分析案例.....160
7.3	对因变量变换后的建模分析.....127	9.1	案例背景.....160
7.3.1	常见的变量变换方法.....127	9.1.1	研究项目概况.....160
7.3.2	本案例的具体操作.....128	9.1.2	分析思路和商业理解.....161
7.4	秩变换分析.....131	9.2	数据理解.....161
7.5	利用Cox模型进行分析.....132	9.2.1	变量关联的图表描述.....161
7.5.1	Cox回归模型的基本原理.....133	9.2.2	变量关联的单变量检验.....164
7.5.2	本案例的具体操作.....134	9.3	构建二分类Logistic回归模型.....167
7.6	项目总结与讨论.....136	9.3.1	模型简介.....167
7.6.1	分析结论.....136	9.3.2	初步尝试建模.....169
7.6.2	如何正确选择分析模型.....136	9.3.3	构建最终模型.....174
第8章	某车企汽车年销量预测案例.....138	9.4	利用树模型发现交互项.....175
8.1	案例背景.....138	9.4.1	模型简介.....176
8.1.1	研究项目概况.....138	9.4.2	进行树模型分析.....178
8.1.2	分析思路和商业理解.....139	9.5	使用广义线性过程进行分析.....181
8.2	数据理解.....140	9.5.1	模型简介.....181
8.3	变量变换后的线性回归.....142	9.5.2	构建仅包括主效应的模型.....182
8.3.1	线性回归模型简介.....142	9.5.3	在模型中加入交互项.....185
8.3.2	变量变换后拟合线性回归模型.....143	9.6	项目总结与讨论.....186
		9.6.1	分析结论.....186
		9.6.2	尺有所短,寸有所长.....187

**第 10 章 中国消费者信心指数影响**

因素分析	188
10.1 案例背景	188
10.1.1 项目背景	188
10.1.2 项目问卷	189
10.1.3 分析思路和商业理解	192
10.2 数据理解	193
10.2.1 考察时间、地域对信心指数的影响	193
10.2.2 考察性别、职业、婚姻状况对信心指数的影响	195
10.2.3 考察年龄对信心指数的影响	196
10.3 标准 GLM 框架下的建模分析	197
10.3.1 建立总模型	197

10.3.2 两两比较的结果	200
10.4 多元方差分析模型的结果	202
10.4.1 模型简介	202
10.4.2 拟合多元方差分析模型	203
10.5 最优尺度回归	209
10.5.1 方法简介	210
10.5.2 利用最优尺度回归进行分析	211
10.6 多水平模型框架下的建模分析	214
10.6.1 模型简介	215
10.6.2 针对时间拟合多水平模型	216
10.7 项目总结与讨论	221
10.7.1 分析结论	221
10.7.2 什么时候运用复杂模型来建模	222

第三部分 信息浓缩、分类与感知图呈现**第 11 章 探讨消费者购买保健品的**

动机	225
11.1 案例背景	225
11.1.1 研究项目概况	225
11.1.2 分析思路和商业理解	227
11.2 数据理解	227
11.2.1 单变量描述	227
11.2.2 变量关联探索	228
11.3 利用因子分析进行信息浓缩	229
11.3.1 模型简介	229
11.3.2 因子分析的具体操作	231
11.4 基于因子分析结果进行市场细分	238
11.4.1 不同婚姻状况受访者的差异	238
11.4.2 不同品牌保健品使用者的因子偏好差异	240
11.5 项目总结与讨论	241
11.5.1 研究结论	241

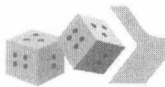
11.5.2 合理解读因子分析的结果	242
--------------------------	-----

第 12 章 1988 年汉城奥运会男子十项全能成绩分析

全能成绩分析	244
12.1 案例背景	244
12.1.1 项目概况	244
12.1.2 分析思路和商业理解	245
12.2 数据理解	246
12.2.1 单变量描述	246
12.2.2 变量关联性探索	246
12.2.3 尝试初步建模	247
12.3 利用因子分析进行信息浓缩	249
12.3.1 初步分析	249
12.3.2 因子旋转	252
12.3.3 继续寻找更好的分析结果	253
12.3.4 结果存储与发布	254
12.4 主成分回归	255
12.5 将主成分回归方程还原回原始变量的形式	257



12.6 项目总结与讨论.....	257	13.6.2 对多维偏好分析等信息浓缩 方法本质的讨论.....	297
12.6.1 研究结论.....	257		
12.6.2 正确诠释因子的方差解释 比例.....	258		
第 13 章 打败 SARS	259	第 14 章 住院费用影响因素挖掘	299
13.1 案例背景.....	259	14.1 案例背景.....	299
13.1.1 研究项目概况.....	259	14.1.1 项目概况.....	299
13.1.2 分析思路和商业理解.....	262	14.1.2 分析思路/商业理解.....	302
13.2 数据理解与数据准备.....	263	14.2 数据理解与数据准备.....	303
13.2.1 消费者关注的信息.....	263	14.2.1 费用数据分布.....	303
13.2.2 突发事件保险产品购买 倾向.....	265	14.2.2 变量合并.....	305
13.2.3 未来消费者生活方式的 变化.....	267	14.2.3 极端值清理.....	306
13.3 “非典”信息关注倾向的多维偏好 分析.....	269	14.2.4 病种分布考察.....	306
13.3.1 模型简介.....	269	14.2.5 变量变换.....	307
13.3.2 多维偏好分析的 SPSS 操作.....	270	14.3 采用聚类分析寻找费用类型.....	308
13.3.3 尝试初步建模.....	272	14.3.1 用因子分析汇总信息.....	308
13.3.4 引入更多的背景变量.....	275	14.3.2 聚类分析方法简介.....	310
13.4 突发事件险种购买倾向的多重 对应分析.....	278	14.3.3 对费用数据进行聚类分析.....	312
13.4.1 模型简介.....	278	14.4 住院费用影响因素的神经网络 分析.....	315
13.4.2 简单对应分析.....	280	14.4.1 模型简介.....	316
13.4.3 多重对应分析.....	284	14.4.2 初步尝试用神经网络建模.....	318
13.5 “非典”对未来生活方式的影响.....	289	14.4.3 对年龄离散化后重新建模.....	323
13.5.1 采用多维偏好分析进行 初步探索.....	289	14.4.4 构建双因变量神经网络.....	325
13.5.2 换用因子分析进行信息 汇总.....	291	14.4.5 进一步寻找更清晰的结果 解释.....	327
13.6 项目总结与讨论.....	295	14.5 不同疗法疗效与费用比较的神经 网络分析.....	328
13.6.1 研究结论.....	295	14.5.1 生成工作用数据集.....	329
		14.5.2 进行神经网络的建模预测.....	330
		14.5.3 模型预测值的比较.....	332
		14.6 项目总结与讨论.....	334
		14.6.1 研究结论.....	334
		14.6.2 数据挖掘方法和经典方法的 取舍.....	335



第四部分 数据挖掘案例精选

第 15 章 淘宝大卖家之营销数据分析339	17.2.4 如何从分析结果中获取实际收益..... 374
15.1 案例背景.....339	17.3 数据理解与数据准备..... 374
15.1.1 卖家张三.....339	17.3.1 分析的数据基础..... 374
15.1.2 分析思路和商业理解.....340	17.3.2 生成数据挖掘宽表..... 376
15.2 利用 RFM 模型定位促销名单.....341	17.3.3 数据探索性分析..... 382
15.2.1 RFM 模型简介.....341	17.4 建立模型与模型评估..... 390
15.2.2 对数据进行 RFM 模型分析.....343	17.4.1 模型的选择..... 390
15.3 寻找有重购行为买家的特征.....348	17.4.2 建模思路 1: 聚类..... 392
15.3.1 数据理解与数据准备.....348	17.4.3 建模思路 2: 用决策树生成规则集..... 394
15.3.2 利用直销模块寻找重购人群的特征.....354	17.4.4 建模思路 3: 用神经网络生成流失评分..... 395
15.4 总结与讨论.....356	17.5 模型的应用及营销预演..... 399
15.4.1 可使用的其他营销分析方法.....356	17.6 总结与讨论..... 401
15.4.2 研究总结.....357	17.6.1 研究总结..... 401
第 16 章 超市商品购买关联分析358	17.6.2 进一步阅读..... 402
16.1 案例背景.....358	第 18 章 信用风险评分方法 403
16.1.1 研究背景.....358	18.1 案例背景..... 403
16.1.2 分析思路和商业理解.....358	18.1.1 引言..... 403
16.2 数据准备.....359	18.1.2 信用评分的方法..... 405
16.3 商品购买关联分析.....362	18.2 商业理解..... 406
16.3.1 几种典型关联算法介绍.....362	18.3 数据理解与数据准备..... 409
16.3.2 商品购买关联分析.....364	18.4 建立模型与模型评估..... 410
16.4 结果应用.....369	18.4.1 对输入变量分箱..... 411
第 17 章 电信业客户流失分析370	18.4.2 用 Logistic 回归建立信用预测模型..... 415
17.1 案例背景.....370	18.4.3 生成信用评分模型..... 417
17.2 商业理解.....371	18.4.4 模型检验..... 420
17.2.1 如何定义流失.....372	18.5 对若干问题的说明..... 422
17.2.2 哪些变量可用于预测流失.....372	18.5.1 拒绝推断..... 422
17.2.3 如何定义分析用数据的时间窗口.....373	18.5.2 模型的监控..... 423
	18.5.3 进一步阅读..... 424



第 19 章 医疗保险业的欺诈发现425	
19.1 案例背景.....425	
19.2 商业理解.....426	
19.3 数据理解与数据准备.....427	
19.3.1 数据集概况.....427	
19.3.2 对数据进行描述.....429	
19.3.3 对数据源合并的考虑.....431	
19.4 建立模型.....432	
19.4.1 进行欺诈发现的若干技术 思路和方法.....432	
19.4.2 模型 1: 变量对比发现疑似 欺诈.....434	
19.4.3 模型 2: 通过 Benford 定律 发现疑似欺诈.....436	
19.4.4 模型 3: 通过对投保人细分 发现疑似欺诈.....439	
19.4.5 模型 4: 发现医疗保健机构 行为模式异常.....441	
19.4.6 模型 5: 使用关联规则发现 多个医保机构共用投保人 信息.....441	
19.4.7 模型 6: 发现异常诊断与 处理过程.....442	
	19.5 结果发布.....444
	19.6 进一步阅读.....445
	第 20 章 电子商务中的数据挖掘 应用446
	20.1 案例背景.....446
	20.1.1 引言.....446
	20.1.2 网络数据分析的分类.....447
	20.2 数据理解.....448
	20.2.1 分析的数据基础.....448
	20.2.2 网络数据的常见来源.....450
	20.3 数据准备.....452
	20.3.1 识别访问用户.....453
	20.3.2 从网络日志中提取有用 信息.....454
	20.3.3 合并网络日志与相关数据.....455
	20.4 建立模型与模型发布.....455
	20.4.1 对访问建立模型.....456
	20.4.2 自动选择模型功能及组合 模型的应用.....459
	20.4.3 对访问者建立模型.....462
	20.4.4 产品特征模型.....464
	20.5 进一步阅读.....465
附 录	
附录 A 本书光盘内容介绍.....469	附录 D IBM SPSS Statistics 函数 一览表.....474
附录 B SPSS 软件的安装与激活.....470	附录 E IBM SPSS Modeler 节点 功能简介.....485
附录 C 书中统计方法、模型与知识 索引.....472	
参考文献.....495	后记.....498



第一部分

SPSS 数据分析基础

第 1 章 数据分析方法论简介

1.1 三种数据分析方法论

所有的数据分析工作都需要在一定的方法论指导下才能正确进行。随着社会的进步，科学技术的发展，统计学的应用已经渗透到人们工作和生活的各个环节，但不同领域所需要的方法论体系有所差别，这些方法论体系大致可分为如下 3 种：

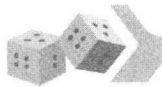
- 严格设计支持下的统计方法论。
- 半试验研究支持下的统计方法论。
- 偏智能化、自动化分析的数据挖掘应用方法论。

IBM SPSS Statistics 作为全球最为出色的统计软件之一，在功能上完全支持上述 3 种方法论体系，并满足绝大多数情况下的统计分析需求，Modeler 则倾向于数据挖掘方法论的具体实现需求。由于对方法论的理解比对分析方法体系的掌握更为重要，因此本章首先对此进行扼要介绍。用户在遇到实际分析需求时，需要首先判断在各自所属的领域中哪种方法论更为合适，并有针对性地加以学习和研究。

1.1.1 严格设计支持下的统计方法论

严格设计支持下的统计方法论也称为经典统计方法论，它之所以经典，不仅因为其发展较早，而且可使研究者在整个研究体系中尽量掌控一切，具体特征如下：

- 这类研究都具有非常严密的研究设计，并且严格遵循七大步骤，即试验设计、数据收集、数据获取、数据准备、数据分析、结果报告和模型发布。七大步骤中以试验设计步骤最为关键，它直接影响整个研究的成败。
- 在此类研究项目中，试验设计中会充分考虑需要控制的影响因素，并采用多种设计方案来对非研究因素的作用加以控制，比如配伍、完全随机抽样、随机分组等。
- 数据在设计完毕后开始采集，整个试验过程会在尽量理想的情况下进行，从而在试验及数据获取过程中对无关因素的作用加以严格控制。例如在毒理学实验中可以对小白鼠的种系、周龄、生活环境、进食等做出非常严格的设定。
- 原始数据往往需要从头采集，数据质量完全取决于试验过程是否严格依从设计要求，以及试验设计本身是否合理等因素。当然，这也意味着每个原始数据的成本都非常高。



- 在分析方法上，最终采用的统计模型应当基于相应的试验设计所定制的分析模型。由于在试验设计和试验实施过程中已经对非研究因素的影响做了充分考虑和控制，因此而在很多情况下往往可以只利用非常简单的统计方法(如 t 检验、卡方检验等)来得到最终结论。各种复杂高深的统计模型不是没有用武之地，但它们不是至关重要的工具。

此类统计方法论的应用在实验室研究、临床试验等领域最为常见，所使用的分析方法常常是单因素分析方法，或者针对一些复杂设计使用一般线性模型(方差分析模型)的定制框架。

1.1.2 半试验研究支持下的统计方法论

经典统计分析方法论对整个流程的控制和干预非常严格，但这在许多情况下是无法满足的，因此往往退而求其次，形成了所谓半实验研究支持下的统计分析方法论，其具体特征如下：

- 研究设计具有明显的向实际情况妥协的特征，所谓七大步骤可能不被严格遵循，例如在数据存在的情况下，数据收集过程就会被省略。总体而言，七大步骤中从数据准备开始的后三步的重要性比经典统计分析方法论高。
- 研究设计可能无法做到理想化，例如抽样与分组的完全随机性，试验组及对照组干预措施的严格控制都可能无法严格满足。举个最典型的例子，药物研究中理想状况应当设立安慰剂对照组，但是如果是治疗恶性肿瘤的药物，又怎么忍心让肿瘤病人吃安慰剂呢？此时往往设定标准治疗药物对照组，甚至在一些极端情形下不设对照组。虽然这样做在统计设计上并不理想，但更符合医疗道德的要求。
- 整个数据采集过程难以做到理想化，举一个简单的例子，定点调查(Central Location Test)是市场研究常用的样本采集方式，严格地说，调查地点、调查时间，甚至当天的天气都可能会对样本的代表性以及数据结果产生影响，但它们最终只能凭借访问者的责任心和运气来尽量加以保证，而从设计本身是很难控制的。
- 部分数据可能先于研究设计而存在，整个研究中需要在这些数据的基础上补充所需的其他部分信息。另一方面，这些数据可能不完全满足分析需求，但这种缺陷却无法得到修正。例如，利用全国各省的经济和人口数据进行省级综合发展程度排序，可以考虑使用因子分析来做，因子分析原则上要求至少有 50 个案例才能保证结果比较稳健，但全国只有 34 个省级行政区，难道为了这个统计分析再请有关部门弄出十几个新的省市来吗？这显然是不切实际的。
- 在分析方法上，由于试验设计难以做到完美，因此各种潜在影响因素的作用可能并不明确，需要以各种可能的影响因素中进行筛选和探索。可能用到的统计方法颇为繁杂，从简单的统计描述到复杂的广义线性模型都可能用到，因此对影响因素的筛选成为很多分析项目的重点任务之一。事实上，很多复杂的多因素分析模