DVD-ROM

IBM SPSS

张文彤　钟云飞　编著

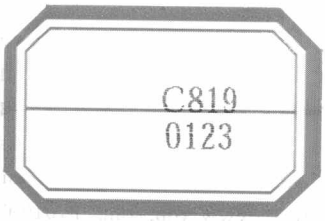
数据分析与挖掘实战案例精粹

- 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

IBM SPSS 数据分析与挖掘实战案例精粹

张文彤　钟云飞　编　著



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。

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

产品编号：044549-01

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一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

第二部分 影响因素发现与数值预测

第三部分　信息浓缩、分类与感知图呈现

第四部分　数据挖掘案例精选

附　录

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