

高 等 学 校 教 材

SPSS

统计分析基础教程 (第3版)

张文彤 编著

高等教育出版社

高 等 学 校 教 材

SPSS

统计分析基础教程

(第3版)

张文彤 编著

高等教育出版社·北京

内容提要

本书将统计理论与 SPSS 操作结合起来,是一本实用性很强的 SPSS 入门教程。本书共分 4 个部分。第一部分为软件入门与数据管理,主要内容包括 SPSS 入门、数据录入与数据获取、变量级别的数据管理、文件级别的数据管理、大型研究项目的数据管理、SPSS 编程;第二部分为统计描述与统计图表,主要内容包括连续变量的统计描述与参数估计、分类变量的统计描述与参数估计、数据的报表呈现、数据的图形展示;第三部分为常用假设检验方法,主要内容包括分布类型的检验、连续变量的统计推断、有序分类变量的统计推断、无序分类变量的统计推断、相关分析、线性回归模型入门;第四部分为统计实战案例集锦,主要内容包括 CCSA 项目的自动化生产、X 药物治疗原发性高血压的临床试验研究、咖啡屋需求调查、牙膏新品购买倾向研究、证券业市场绩效与市场结构关系的实证分析。本书基于 IBM SPSS Statistics 24 中文版,以真实案例贯穿全书,有助于读者提高实战能力。

本书可作为高等学校本科生和研究生统计分析课程教材,也可作为各行业中非统计专业背景、需要使用统计方法的人员,以及希望学习 SPSS 软件使用的人员的参考书。

图书在版编目(CIP)数据

SPSS 统计分析基础教程/张文彤编著.--3 版.--

北京:高等教育出版社,2017.4

ISBN 978-7-04-047460-2

I. ①S… II. ①张… III. ①统计分析-软件包-高等学校-教材 IV. ①C819

中国版本图书馆 CIP 数据核字(2017)第 027857 号

SPSS Tongji Fenxi Jichu Jiaocheng

策划编辑 刘艳
插图绘制 杜晓丹

责任编辑 唐德凯
责任校对 陈旭颖

封面设计 于文燕
责任印制 刘思涵

版式设计 杜微言

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 河北鹏盛贤印刷有限公司
开 本 787mm×1092mm 1/16
印 张 24.5
字 数 600 千字
购书热线 010-58581118
咨询电话 400-810-0598

网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.hepmall.com.cn>
<http://www.hepmall.com>
<http://www.hepmall.cn>
版 次 2004 年 10 月第 1 版
2017 年 4 月第 3 版
印 次 2017 年 8 月第 2 次印刷
定 价 46.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 47460-00

前 言

本书自 2004 年第 1 版出版以来,受到了广大读者的欢迎,被国内数百所高校选为本科生或研究生相关课程教材,在此感谢广大读者的支持与厚爱。

本书第 2 版出版于 2011 年,SPSS 在这几年间已经升级了 4 个版本,而且最新的版本易用性更强,软件功能也更丰富,因此需要对全书内容进行有针对性的修订。在第 2 版的基础上,结合 SPSS 版本的更新和读者的反馈,本版以 IBM SPSS Statistics 24 版为准,对内容做了如下调整。

1. 进一步实战化

仍然以中国消费者信心调研(CCSS)项目案例作为贯穿全书的主案例,并且每一章的编写都围绕该项目的实际分析展开,使全书的结构更贴近项目实际的分析流程。原有的“统计实战案例集锦”章节单列为第四部分,并进行了内容调整,以便于读者阅读。

2. 突出统计主线

在本书以前的版本中,考虑到当时用户的计算机应用水平,实际上一直采用软件+统计分析的双主线结构,这种两者兼顾的方式给读者带来了一些困扰。随着用户计算机应用水平的提高,本版不再需要两者兼顾,而能够突出统计分析的主线,软件操作步骤和技巧等均配合统计分析的主线展开介绍,供学有余力的读者学习。这样的结构更便于统计教学的开展。

3. 跟进软件新功能

SPSS 近几年来借助 Python 和 R 扩展增加了很多新的功能,对于这些新功能,没有必要全部进行介绍。本版采取授之以渔的思路,重点介绍在实际应用中价值的新功能,并在第 1 章提供了插件安装、设定的操作方式,这样有利于读者根据自身需求进行深入研究。

本书内容覆盖了目前高等学校大部分本科专业统计分析课程的教学范围,并结合 SPSS 的强大功能做了很好的扩展。每章后均附有思考与练习,涉及统计学理论的章节还提供了本章小结;书中带*的内容具有一定的深度,读者可以根据个人基础和兴趣选学。此外,本书和《SPSS 统计分析高级教程》构成了对 SPSS 的全面介绍。为便于读者交流和使用这套书,读者可以通过以下网址获得相关资源。

微信公众号:统计之星。

本书案例数据、内容修订、拓展阅读资料下载:<http://www.StatStar.com>。

希望广大读者能够一如既往地踊跃提出自己的宝贵意见和建议,使得本书再次改版时能够更上一层楼,更好地满足大家的学习和工作需求。

作 者

2016 年 11 月

目 录

第一部分 软件入门与数据管理

第 1 章 SPSS 入门 3	
1.1 软件概述 3	
1.1.1 SPSS 发展简史 3	
1.1.2 SPSS 的产品定位 4	
1.1.3 SPSS 的基本特点 5	
1.1.4 SPSS 的 Client/Server 结构与 模块化结构 5	
1.2 SPSS 操作入门 7	
1.2.1 SPSS 的安装与激活 7	
1.2.2 SPSS 的启动与退出 7	
1.2.3 SPSS 的操作方式 8	
1.2.4 SPSS 对话框操作基本规范 9	
1.3 SPSS 的窗口、菜单项和结果输出 11	
1.3.1 SPSS 的 4 种窗口 11	
1.3.2 SPSS 菜单项 12	
1.3.3 SPSS 的 4 种结果输出 13	
1.3.4 分析结果的保存和导出 14	
1.4 SPSS 的系统选项与扩展资源 16	
1.4.1 系统选项 16	
1.4.2 网站资源 17	
1.4.3 安装 Python/R 扩展 17	
1.5 SPSS 的帮助系统 18	
1.5.1 学习向导 18	
1.5.2 软件操作帮助 19	
1.5.3 针对高级用户的帮助功能 20	
1.6 数据分析方法论概述 21	
1.6.1 严格设计支持下的统计 方法论 21	
1.6.2 半试验研究支持下的统计 方法论 21	
1.6.3 偏智能化、自动化分析的 数据挖掘应用方法论 22	
思考与练习 23	
第 2 章 数据录入与数据获取 24	
2.1 CCSS 项目背景介绍 24	
2.1.1 项目背景 24	
2.1.2 项目问卷 24	
2.2 数据格式概述 26	
2.2.1 统计软件中数据的录入 格式 26	
2.2.2 变量属性 26	
2.3 在 SPSS 中直接建立数据集 31	
2.3.1 操作界面说明 31	
2.3.2 开放题和单选题的设定与 录入 31	
2.3.3 多选题的设定与录入 33	
2.4 读入外部数据 34	
2.4.1 读取电子表格数据文件 34	
2.4.2 读取文本数据文件 35	
2.4.3 用 ODBC 接口读取各种 数据库文件 38	
2.5 数据的保存 38	
2.5.1 保存为 SAV 格式 38	
2.5.2 保存为其他数据格式 39	
2.6 数据编辑窗口常用操作技巧 39	
2.6.1 数据录入技巧 40	
2.6.2 快速定位技巧 41	
2.6.3 窗口操作与切换技巧 41	
思考与练习 43	

第 3 章 变量级别的数据管理	44	第 5 章 大型研究项目的数据管理	71
3.1 变量赋值	45	5.1 数据字典	71
3.1.1 常用基本概念	45	5.1.1 数据字典简介	71
3.1.2 界面介绍	46	5.1.2 定义变量属性	72
3.1.3 案例:年龄变量 S3 的分组	46	5.1.3 复制数据属性	73
3.2 已有变量值的分组合并	47	5.1.4 新建自定义属性和设置新 变量的测量级别	74
3.2.1 对连续变量进行分组合并	47	5.2 数据核查	75
3.2.2 分类变量类别的合并	48	5.2.1 数据核查概述	75
3.3 连续变量的离散化	49	5.2.2 数据验证模块	76
3.3.1 可视离散化	49	5.2.3 标识重复个案	79
3.3.2 最优离散化	50	5.2.4 双录核查	80
3.4 自动重编码、编秩与数值计数	52	5.3 数据准备	82
3.4.1 自动重编码	52	5.3.1 标识异常个案	82
3.4.2 数据编秩	53	5.3.2 数据自动准备	84
3.4.3 指定数值的查找与计数	54	5.3.3 数据的匿名化	86
3.5 转换菜单中的其他功能	54	5.3.4 使用变量集	87
思考与练习	55	思考与练习	88
第 4 章 文件级别的数据管理	56	第 6 章 SPSS 编程	89
4.1 几个常用过程	57	6.1 CCSS 项目的数据处理需求	89
4.1.1 个案排序	57	6.2 SPSS 编程入门	90
4.1.2 拆分文件	57	6.2.1 基本语法规则	90
4.1.3 选择个案	59	6.2.2 SPSS 程序的创建方式	92
4.1.4 加权个案	59	6.2.3 结构化语句简介	93
4.1.5 分类汇总	61	6.2.4 简单程序示例	94
4.2 多个数据文件的合并	62	6.3 语法编辑窗口操作入门	95
4.2.1 一些基本概念	62	6.3.1 语法编辑窗口界面	95
4.2.2 数据文件的纵向拼接	63	6.3.2 程序的运行与调试	97
4.2.3 数据文件的横向合并	64	6.4 宏程序与 INSERT 命令	97
4.3 数据文件的重组与转置	66	6.4.1 宏程序	98
4.3.1 数据的长型与宽型格式	66	6.4.2 INSERT 命令	99
4.3.2 长型格式转换为宽型格式	67	6.5 OMS 与程序自动化	100
4.3.3 宽型格式转换为长型格式	68	6.5.1 OMS	100
4.3.4 数据转置	69	6.5.2 程序自动化	103
4.4 数据菜单中的其他功能	69	思考与练习	104
思考与练习	70		

第二部分 统计描述与统计图表

第 7 章 连续变量的统计描述与参数估计	107	描述	130
7.1 连续变量的统计描述指标体系	107	8.3.1 设定多选题变量集	130
7.1.1 集中趋势的描述指标	107	8.3.2 多选题的频数列表	131
7.1.2 离散趋势的描述指标	109	8.3.3 多选题的列联表分析	132
7.1.3 分布特征及其他趋势的描述指标	110	思考与练习	134
7.1.4 SPSS 中的相应功能	111	第 9 章 数据的报表呈现	135
7.2 连续变量的参数估计指标体系	112	9.1 统计表入门	135
7.2.1 正态分布	112	9.1.1 统计表的基本框架	135
7.2.2 参数的点估计	113	9.1.2 表头、数据区与汇总项	136
7.2.3 参数的区间估计	114	9.1.3 单元格的数据类型	137
7.2.4 SPSS 中的相应功能	115	9.1.4 几种基本表格类型	137
7.3 案例:信心指数的统计描述	115	9.1.5 SPSS 中的报表功能	139
7.3.1 使用频率过程进行分析	115	9.1.6 SPSS 中统计表的基本绘制步骤	139
7.3.2 使用描述过程进行分析	117	9.2 简单案例:题目 A3 的标准统计报表制作	140
7.3.3 使用探索过程进行分析	118	9.2.1 案例介绍	140
7.4 Bootstrap 方法	121	9.2.2 绘制表格基本框架	140
7.4.1 模型介绍	121	9.2.3 设置摘要统计量及格式	142
7.4.2 案例:对总指数进行 Bootstrap 估计	122	9.2.4 调整各种显示细节	144
思考与练习	124	9.3 复杂案例:题目 A3a 的标准统计报表制作	145
第 8 章 分类变量的统计描述与参数估计	125	9.3.1 案例介绍	145
8.1 指标体系概述	125	9.3.2 多选题、表格基本框架及汇总项的设定	146
8.1.1 单个分类变量的统计描述	125	9.3.3 设定分类变量小结和汇总项	147
8.1.2 多个分类变量的联合描述	126	9.3.4 “定制表”对话框的其他选项卡	147
8.1.3 多选题的统计描述	126	9.4 表格的编辑	150
8.1.4 分类变量的参数估计	127	9.4.1 基本编辑操作	150
8.1.5 SPSS 中的相应功能	127	9.4.2 主要编辑菜单功能介绍	152
8.2 案例:对学历等背景变量进行描述	128	9.4.3 表格属性的详细设置	153
8.2.1 使用频率过程进行描述	128	9.5 表格模板技术	154
8.2.2 使用交叉表过程进行描述	128	思考与练习	156
8.3 案例:对多选题 C0 还贷状况进行			

第 10 章 数据的图形展示	157	10.5.5 带误差线的条图与 误差图	185
10.1 统计图概述	157	10.6 线图、面积图、点图与垂线图	186
10.1.1 统计图的基本框架	157	10.6.1 多重线图案例:分城市 比较信心指数随时间的 变化趋势	186
10.1.2 统计图的种类	159	10.6.2 线图的编辑	187
10.1.3 SPSS 的统计绘图功能	162	10.6.3 面积图、点图与垂线图	189
10.2 直方图和茎叶图	163	10.7 散点图	189
10.2.1 案例:绘制消费者信心 值的直方图	163	10.7.1 简单散点图案例:年龄 S3 与消费者信心指数间的 关系	190
10.2.2 图形的基本编辑操作	164	10.7.2 散点图的编辑	190
10.2.3 直方图图形框架的修改	170	10.7.3 分组散点图案例:分性别 考察年龄对信心指数值的 影响	193
10.2.4 直方图的衍生图形	171	10.7.4 散点图矩阵案例:年龄 S3 与现状指数、预期指数的 关系	194
10.2.5 茎叶图	173	10.7.5 三维散点图	195
10.3 箱图	174	10.8 P-P 图和 Q-Q 图	196
10.3.1 案例:用箱图分月份考察 消费者信心的分布	174	10.8.1 P-P 图	196
10.3.2 箱图的编辑	175	10.8.2 Q-Q 图	198
10.4 饼图	177	10.9 控制图与 Pareto 图	199
10.4.1 案例:分城市、月份考察 样本性别比例	177	10.9.1 控制图	199
10.4.2 饼图的编辑	177	10.9.2 Pareto 图	201
10.5 条图与误差图	179	10.10 其他统计图	202
10.5.1 简单条图案例:比较不同 职业人群的消费者 信心值	180	10.10.1 ROC 曲线	202
10.5.2 复式条图案例:分职业 进一步比较不同人群的 现状和预期指数	181	10.10.2 高-低图	204
10.5.3 分段条图与百分条图 案例:比较不同月份的 A3a 选项比例分布	182	10.10.3 时间序列分析中使用的 图形	206
10.5.4 条图的编辑	184	思考与练习	206

第三部分 常用假设检验方法

第 11 章 分布类型的检验	209	反证法	210
11.1 假设检验的基本思想	209	11.1.3 假设检验的标准步骤	210
11.1.1 问题的提出	209	11.1.4 假设检验的两类错误	211
11.1.2 小概率事件与小概率		11.1.5 假设检验中的其他问题	211
		11.2 正态分布检验	212

11.2.1	K-S 检验的原理	212	信心值的比较	233
11.2.2	案例:考察信心指数分布 是否服从正态分布	213	12.6 本章小结	235
11.3	二项分布检验	216	思考与练习	235
11.3.1	二项分布检验的原理	216	第 13 章 连续变量的统计推断(二)	
11.3.2	案例:考察抽样数据的 性别分布是否平衡	216	——单因素方差分析	237
11.4	游程检验	217	13.1 方差分析概述	237
11.4.1	游程检验的原理	217	13.1.1 为什么要进行方差分析	237
11.4.2	案例:考察 GCSS 抽样 数据是否随机	218	13.1.2 方差分析的基本思想	237
11.5	本章小结	220	13.1.3 单因素方差分析的适用 条件	239
	思考与练习	220	13.2 案例:北京消费者不同时点信心 指数的比较	239
第 12 章 连续变量的统计推断(一)			13.3 均数间的多重比较	242
—— t 检验	221		13.3.1 直接校正检验水准	242
12.1 t 检验概述	221		13.3.2 专用的两两比较方法	243
12.1.1 t 检验的基本原理	221		13.3.3 两两比较方法的选择 策略	245
12.1.2 SPSS 中的相应功能	223		13.3.4 多重比较结果出现矛盾 时的解释	245
12.2 样本均数与总体均数的比较	224		13.3.5 案例:不同时点信心 指数的两两比较	246
12.2.1 案例:基期广州信心指数 均值与基准值的比较	224		13.3.6 制表模块的两两比较 输出	247
12.2.2 单样本 t 检验中的其他 问题	226		13.4 各组均数的精细比较	249
12.3 成组设计两样本均数的比较	226		13.4.1 方法原理	249
12.3.1 方法原理	226		13.4.2 案例:一个时点与另两个 时点均数的比较	250
12.3.2 案例:考察婚姻状况对 信心指数的影响	227		13.5 组间均数的趋势检验	252
12.3.3 两样本 t 检验的适用 条件	229		13.5.1 方法原理	252
12.4 正态性、方差齐性的考察与应对 策略	230		13.5.2 案例:对信心指数的变化 做趋势检验	252
12.4.1 正态性的考察方法	230		13.6 本章小结	253
12.4.2 方差齐性的考察方法	231		思考与练习	253
12.4.3 数据不符合适用条件时的 应对策略	231		第 14 章 有序分类变量的统计推断	
12.5 配对设计样本均数的比较	232		——非参数检验	255
12.5.1 方法原理	232		14.1 非参数检验概述	255
12.5.2 案例:同一受访者前后			14.1.1 非参数检验的特点	255

14.1.2 非参数检验预备知识	256	15.5.3 蒙特卡洛法的 SPSS 实现	281
14.2 两个配对样本的非参数检验	257	15.6 两分类变量间关联程度的度量 ...	282
14.2.1 方法原理	257	15.6.1 相对危险度与优势比	282
14.2.2 案例:同一受访者前后 单项指标的比较	258	15.6.2 案例:计算家庭收入级别 和轿车拥有情况的关联 程度	283
14.3 两个独立样本的非参数检验	260	15.7 一致性检验与配对卡方检验	284
14.3.1 方法原理	260	15.7.1 Kappa 一致性检验	284
14.3.2 案例:不同收入家庭经济 现状感受值的比较	261	15.7.2 配对卡方检验	286
14.4 多个独立样本的非参数检验	263	15.8 分层卡方检验	286
14.4.1 方法原理	263	15.9 本章小结	289
14.4.2 案例:北京消费者不同 时点的题目得分比较	264	思考与练习	289
14.5 多个相关样本的非参数检验	266	第 16 章 相关分析	291
14.5.1 Friedman 检验	266	16.1 相关分析简介	291
14.5.2 Kendall 协同系数检验与 Cochran 检验	267	16.1.1 相关分析的指标体系	291
14.6 秩变换分析方法	268	16.1.2 SPSS 中的相应功能	293
14.6.1 秩变换分析原理简介	268	16.2 简单相关分析	294
14.6.2 案例:用秩变换来比较 不同时点的家庭经济 感受值	268	16.2.1 方法原理	294
14.7 本章小结	269	16.2.2 案例:考察信心指数值和 年龄的相关性	297
思考与练习	270	16.2.3 秩相关系数	298
第 15 章 无序分类变量的统计推断 ——卡方检验	272	16.2.4 Kendall 等级相关系数	299
15.1 卡方检验概述	272	16.3 偏相关分析	300
15.1.1 卡方检验的基本原理	272	16.3.1 方法原理	300
15.1.2 卡方检验的用途	273	16.3.2 案例:控制家庭收入的 影响之后考察年龄 的作用	300
15.1.3 SPSS 中的相应功能	273	16.4 本章小结	302
15.2 单样本案例:考察抽样数据的 性别分布	274	思考与练习	302
15.3 两样本案例:不同收入级别 家庭的轿车拥有率比较	275	第 17 章 线性回归模型入门	303
15.4 卡方检验的事后两两比较	278	17.1 线性回归模型简介	303
15.5 确切概率法和蒙特卡洛法	279	17.1.1 相关分析与回归分析的 联系与区别	303
15.5.1 Yates 校正与确切概率法 ...	279	17.1.2 简单回归分析的原理和 要求	304
15.5.2 蒙特卡洛法简介	280	17.2 案例:建立用年龄预测总信心	

指数值的回归方程·····	306	17.3.4 SPSS 中与多重线性回归	
17.3 多重线性回归模型入门·····	310	模型相关的功能·····	313
17.3.1 模型简介·····	310	17.3.5 案例:建立自变量包括	
17.3.2 多重线性回归模型的标准		年龄、性别、家庭收入的	
分析步骤·····	310	信心指数回归方程·····	314
17.3.3 回归方程中的自变量筛选		17.4 本章小结·····	317
方法·····	312	思考与练习·····	318
第四部分 统计实战案例集锦			
第 18 章 CCSS 项目的自动化生产·····	321	20.2 数据预分析·····	338
18.1 项目背景·····	321	20.3 主体问卷分析·····	340
18.2 分析思路·····	321	20.4 项目总结与讨论·····	345
18.3 具体操作·····	323	思考与练习·····	345
18.4 项目总结与讨论·····	325	第 21 章 牙膏新品购买倾向研究·····	346
思考与练习·····	326	21.1 项目背景·····	346
第 19 章 X 药物治疗原发性高血压的		21.2 分析思路·····	346
临床试验研究·····	327	21.3 数据预分析·····	347
19.1 项目背景·····	327	21.4 数据建模·····	350
19.1.1 研究概况·····	327	21.5 项目总结与讨论·····	352
19.1.2 研究方法·····	327	思考与练习·····	353
19.2 数据准备·····	328	第 22 章 证券业市场绩效与市场结构	
19.3 基线情况比较·····	330	关系的实证分析·····	354
19.4 疗效评价·····	332	22.1 项目背景·····	354
19.5 安全性评价·····	334	22.2 数据的采集·····	354
19.6 结论与总结·····	336	22.3 数据预分析·····	355
思考与练习·····	336	22.4 数据建模·····	356
第 20 章 咖啡屋需求调查·····	337	22.5 项目总结与讨论·····	358
20.1 项目背景·····	337	思考与练习·····	358
附录·····	359		
附录 1 各种情形下最常用的统计检验方法索引·····	359		
附录 2 SPSS 函数一览表·····	360		
附录 3 统计术语英汉名词对照表·····	371		
参考文献·····	377		

第一部分

软件入门与数据管理

第1章 SPSS 入门

1.1 软件概述

SPSS 是世界上应用最广泛的专业统计软件之一,在全球拥有众多用户,分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业,全球 500 强企业中约有 80% 的公司使用 SPSS,而在市场研究和市场调查领域则有超过 80% 的市场占有率,和 SAS 并称为当今最权威的两大统计软件。



SPSS 实际上是该软件的简称,其全称则发生过几次变化,最早为 Statistical Package for Social Sciences,意为“社会科学统计软件包”;后来随着产品服务领域的扩大和服务深度的增加,SPSS 公司于 2002 年将英文全称更改为 Statistical Product and Service Solutions,意为“统计产品与服务解决方案”,以反映市场的新趋势;但是在 2009 年 4 月,SPSS 公司做了一个令广大用户费解的决定:将 SPSS 软件更名为 PASW (Predictive Analytics Software) Statistics! 幸好在当年 9 月 SPSS 公司就被 IBM 收购,而新东家则立即终止了更名计划,重新将软件命名为 IBM SPSS Statistics,算是给这一事件画上了句号,并一直沿用至今。但无论名称如何,SPSS 软件的风格和基本定位始终未变,用户都喜欢称其为 SPSS,它也一直是广大用户所喜爱的强大统计工具。

1.1.1 SPSS 发展简史

SPSS 的历史开始于 1968 年,斯坦福大学的 3 位不同专业的研究生(两位博士研究生,一位硕士研究生)编制出了世界上最早的统计软件系统,并将其命名为 SPSS。随后,该软件和后续成立的 SPSS 公司就走上了持续发展的道路。

1968~1974 年:SPSS 成为真正的产品。从一个雏形开始,通过不断的代码积累和修改,SPSS 最终形成了成熟的、可销售的产品。

1975~1983 年:SPSS 公司正式成立,并在一系列探索之后,最终确立了以统计软件和统计分析服务为主业的定位。

1984~1991 年:PC 时代。SPSS 公司在全球首家推出了 PC 版的统计分析软件 SPSS/PC+4.0,该版本为全球第一套以图形菜单为用户界面的统计软件,也是 DOS 时代的统计软件经典之作。

1992~1996 年:Windows 时代。在 1992 年,SPSS 公司在全球首家推出了 Windows 版的统计分析软件 SPSS 6,随着这一软件的成功,公司也走上了快速扩张之路,并购了诸如 SYSTAT (1994) 和 Jandel (1996) 等企业。

1997~2002 年:向大企业进化。期间 SPSS 不断推陈出新,经典的 11 版就是在这一期间推出

的。更重要的是,并购行动仍在继续,诸如 Quantime(市场研究应用软件)、ISL(数据挖掘软件)、ShowCase(商务智能中间件)、NetGenesis(网络数据分析应用)、LexiQuest(文本挖掘软件)和 netExs(OLAP 网络接口及界面)等一系列具有战略价值的公司被收入囊中,这也意味着公司开始形成完整的产品线。

2003~2008年:向预测分析转型。在完成上述并购后,SPSS 公司开始重新整合产品线,并开始统一向商务智能与预测分析转型。SPSS 被定位为产品线中的普及类工具,和其余产品形成高低搭配。然而这一过程并不顺利,显然市场的成熟速度落后于预期,SPSS 公司坚持了下来。SPSS 也仍然在不断更新,13 版堪称又一个经典的版本,自 17 版开始提供基本成熟的中文界面与结果输出。

2009 年至今:融入 IBM。随着并购的完成,SPSS 产品揭开了新的一页,以 IBM 商务智能(BI)方向的一个重要产品线的面貌出现。19~24 版的更新主要是对数据的自动化准备、自动化分析等智能化、易用性方面进行了增强,同时开始提供 Python、R 的扩展模块,以便在保持易用性的同时能够为用户提供尽可能多的先进统计分析方法。



SPSS 近年来始终以 1~1.5 年升级一个大版本号的速度进行升级,但显然并非每次更新都会有重大变化,基于十几年来对 SPSS 的使用经验,建议目前用户可以考虑使用 13 版、20 版和 24 版,前两个版本的质量很好,而 24 版则在新功能上更加完善。

1.1.2 SPSS 的产品定位

俗话说,尺有所短,寸有所长。每种工具都有其定位与特点,SPSS 虽然是一个很好的工具软件,但如果不能正确理解其行业定位,就无法最大限度地发挥其功用。实际上,在被 IBM 收购后,原 SPSS 公司的产品线被进一步整合为四大系列,具体如下。

(1) Data Collection Family:定位为中低端的数据采集与报告需求领域,是一个完整的技术平台,支持从创建调查到收集数据,再到报告的整个调查研究的生命周期。根据其应用领域,Data Collection Family 可以分为 6 个部分:在线调查(Online Surveys)、电话调查(Phone Surveys)、离线调查(Offline Surveys)、数据录入(Data Entry)、调查报告(Survey Reporting)、调查管理(Survey Management),其中每一部分都由数个产品组成。

(2) Statistics Family:定位为中端的统计分析服务需求领域,由原先的 SPSS 软件构成,但 Statistics Base 不再是必备模块,原先的每个附加模块现在都可以独立安装和运行,或者是几个模块组合在一起,每个模块都可以拥有数据访问、数据管理和绘图功能。

(3) Modeling Family:由原先的 Clementine 发展而来,现更名为 IBM SPSS Modeler,并进一步包括了 Text Analytics、Social Media Analytics 等相关产品。该产品线主攻高端的数据挖掘与商务智能需求领域,也是最有发展潜力的一块。

(4) Deployment Family:相对而言是对前 3 个产品系列的整合与后台支持,包括 SPSS Collaboration&Deployment Services、Analytic Decision Management 等产品,用于将市场调研、统计分析技术、数据挖掘技术以及报表技术整合到一个平台中,帮助企业建立统一的中央资产存储库,用完整的预测分析流程支持企业日常业务,方便数据分析人员分享资源。

1.1.3 SPSS 的基本特点

SPSS 受到用户的广泛欢迎并长盛不衰的原因在于其强大的统计分析与数据准备功能,方便的图表展示功能,以及良好的兼容性、界面的友好性满足了广大用户的需求,特别是得到了广大应用统计分析人员的喜爱。

1. 功能强大

SPSS 囊括了各种成熟的统计方法与模型,为用户提供了全方位的统计方法,如方差分析、回归分析、多元统计分析方法、生存分析方法等,方法体系覆盖全面。在数据准备方面,SPSS 提供了各种数据准备与数据整理技术。例如,利用值标签来快捷地录入数据、对连续型变量进行离散型转换、将几个小类别合并为一个大类别、重复记录的发现、异常数据的发现等。这些强大的数据整理技术可使数据更易于分析使用。

在结果报告方面,SPSS 提供了自由灵活的表格功能,使得制表变得更加简单、直接。同时,SPSS 可绘制各种常用的统计图形,如条图、线图、饼图、直方图、散点图等多种图形,以对数据进行全面直观的展示。

2. 兼容性好

在数据方面,不仅可在 SPSS 中直接进行数据录入工作,还可将日常工作中常用的 Excel 表格数据、文本格式数据导入 SPSS 进行分析,从而节省了相当的工作量,并且避免了因复制粘贴可能引起的错误。

在结果方面,SPSS 的表格、图形结果可直接导出为 Word、文本、网页、Excel 等格式,而且目前已彻底解决了中文兼容问题,用户不需要进行任何附加设置就可自由使用中文,并将中文结果输出到 Word 等软件中直接使用。

3. 易用性强

SPSS 之所以有广大的用户群,不仅因为它是一种权威的统计学工具,也因为它是一种非常简单易用的软件。人机界面友好、操作简单,使得统计分析人员对它“情有独钟”,事实上,不断地增强其易用性(而不是盲目追求方法的高精尖)几乎是近十几年来 SPSS 的核心改进方向。另外,SPSS 也向高级用户提供了编程功能,使分析工作变得更加节省时间和精力。

4. 扩展性高

SPSS 长期以来一直为竞争对手所诟病的问题主要是它对新方法、新功能的纳入速度很慢。这虽然与其市场定位有关,但毕竟是一个缺陷。对此,SPSS 提供了一个巧妙的解决办法,就是直接和 R 进行对接,通过调用 R 的各种统计模块来实现对最新统计方法的调用,从而彻底解决了这一问题。而这一扩展能力更是在 24 版中得到了重点强化,详见 1.4 节。

1.1.4 SPSS 的 Client/Server 结构与模块化结构

1. SPSS 的 Client/Server 结构

SPSS 自 10 版以来,一直是 Client/Server(客户-服务器)结构的体系。对于大数据量的分析,用户可以选择购买 SPSS Server,以利用 Server 的计算能力来解决数据量太大的问题。当然,对于数据量不大的客户,只用 SPSS Client 就可以了。现在国内绝大多数用户所使用的 SPSS,实际上就是指 SPSS Client。

2. SPSS 的模块结构

无论是 SPSS Client 还是 SPSS Server, 均是模块式结构, 即它把自己的所有功能划分为多个模块。用户可以根据分析中可能用到的数据处理和统计分析方法, 选择适当的模块进行购买, 而不必花更多的钱购买所有模块。

SPSS 的模块数量随版本的不同一直有所变化, 在 18 版以前的版本中, Statistics Base 是必需的, 软件的整个框架、基本的数据获取、数据准备等基本功能都被集中在这个模块上, 其他模块必须在 Statistics Base 搭建的平台上工作。从 18 版起, 其余模块也可以脱离 Statistics Base 单独存在并运行。但对于普通用户而言, 仍然是以 Base+其余模块的用法最为常见。这里列出 SPSS 主要模块的功能简介, 见表 1.1。

表 1.1 SPSS 常见模块与功能对应表

模块名称	功能
Statistics Base	提供最常用的数据管理和统计分析功能
Advanced Statistics	一般线性模型、混合线性模型、对数线性模型、生存分析等
Regression	Logistic 回归、非线性回归、Probit 回归等
Categories	对应分析、感知图、PROXSCAL 等
Missing Value	缺失数据的报告与填补等
Conjoint	正交设计、联合分析等, 适用于市场研究
Forecasting	Arima 模型、指数平滑、自回归等
Tables	交互式创建各种表格(如堆积表、嵌套表、分层表等)
Complex Samples	多阶段复杂抽样技术等
Bootstrap	提供计算统计学中的 Bootstrap 方法用于参数估计
Decision Trees	提供树结构模型分析方法
Neural Network	提供 BP 神经网络和 RBF 神经网络方法
Data Preparation	提供数据核查、自动清理等一系列数据准备工具
Statistic Adapter	实际上属于 SPSS 和 Deployment Family 产品的接口, 可以在企业应用程序、工具和解决方案环境中管理对象的生命周期
Direct Market	提供了一组用于改善直销活动效果的工具, 以针对特定目标群体最大限度地提高促销措施的响应率

SPSS 软件以前是通过 License 来控制相应的模块是否可被安装, 但是从 19 版起, 则不再限制模块的安装, 而是限制该模块是否可用。也就是说, 如果是完整的安装包, 那么在软件安装完毕之后, 在软件菜单中将会出现所有模块的菜单项, 但如果没有购买相应模块的 License, 则相应的模块是无法运行的。



有一点需要澄清, 国内一些 SPSS 书籍因对 SPSS 的功能模块介绍不全, 总是在前言中声明所使用的是 SPSS 标准版。实际上 SPSS 的区别只在于各模块在运行时是否有使用许可, 即便是为营销需求而打包销售的“标准版”, 其功能也远远超过了这些书籍所介绍的内容。