

高 等 学 校 教 材

SPSS

统计分析基础教程 (第2版)

张文彤 邝春伟 编著



高等教育出版社
HIGHER EDUCATION PRESS

本书特色:

1. 本书是作者多年使用SPSS进行教学、科研与项目实战工作的经验结晶。
2. 本书介绍最新的IBM SPSS Statistics 20版本, 结合实战需求全面介绍软件的新功能和使用技巧, 保留了第1版对统计理论深入浅出的讲解风格, 大幅降低了初学者的入门难度。
3. 本书采用真实的数据案例进行结构安排和讲解, 同时增加“统计实战案例集锦”, 提供医疗、经济、市场研究等多行业真实案例, 使本书更贴近统计分析实战。
4. 本书经SPSS中国公司审定并加以推荐, 既可作为高等学校统计分析课程教材, 也可作为培训类教材使用。

ISBN 978-7-04-033241-4



9 787040 332414 >

定价 39.00元

高等学校教材

SPSS 统计分析基础教程

SPSS Tongji Fenxi Jichu Jiaocheng

(第2版)

张文彤 邝春伟 编著



高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

内容提要

本书采用的 IBM SPSS Statistics 20 中文版,以真实案例贯穿全书,从统计分析实战的角度出发详细介绍 SPSS 的界面操作、数据管理、统计图表制作、统计描述和常用单因素统计分析方法的原理与实际操作,并结合 SPSS 的强大功能进行很好地扩展。书中还提供医疗、经济、市场研究等行业的综合案例,完全从实际案例出发讲解各类方法的综合运用,以更好地协助读者提高实战能力。

本书对第 1 版内容进行了全面改写,以一种全新的实战案例风格出现,是一本难得的统计理论与 SPSS 操作相结合的参考书。

本书可作为统计学、社会学、教育学等专业本科生和研究生课程教材,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从头学习 SPSS 软件使用方法的人员的参考书。

图书在版编目(CIP)数据

SPSS 统计分析基础教程/张文彤, 邝春伟编著. —2 版. —北京: 高等教育出版社, 2011.11
ISBN 978-7-04-033241-4
I. ①S… II. ①张… ②邝… III. ①统计分析-软件包, SPSS-高等学校-教材 IV. ①C819
中国版本图书馆 CIP 数据核字(2011)第 211085 号

策划编辑 耿 芳	责任编辑 耿 芳	封面设计 于文燕	版式设计 杜微言
插图绘制 宗小梅	责任校对 杨凤玲	责任印制 田 甜	

出版发行 高等教育出版社	网 址 http://www.hep.edu.cn
社 址 北京市西城区德外大街 4 号	http://www.hep.com.cn
邮政编码 100120	网上订购 http://www.landradio.com
印 刷 廊坊市科通印业有限公司	http://www.landradio.com.cn
开 本 787mm × 1092mm 1/16	
印 张 27	版 次 2004 年 9 月第 1 版
字 数 660 千字	2011 年 11 月第 2 版
购书热线 010-58581118	印 次 2011 年 11 月第 1 次印刷
咨询电话 400-810-0598	定 价 39.00 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 33241-00

诸多帮助,并同意将本书作为 SPSS China 官方推荐教材使用,特在此一并致谢。

为便于读者交流和使用本书,这里特公布相关网址如下:

作者微博:<http://weibo.com/wintone>

读者交流微群:<http://q.weibo.com/749521>

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

希望本书能够帮助读者更好地了解统计分析方法,从而进一步促进统计分析方法在国内的普及。也希望广大读者能一如既往地踊跃提出自己使用中的宝贵意见和建议,使得本书推出第 3 版时能够更上一层楼,更好地满足大家的学习和工作需求。

编 者
2011 年 8 月

目 录

第一部分 数据管理与软件入门

第1章 SPSS 入门 3	1.6.2 半试验研究支持下的统计 方法论 23
1.1 SPSS 概述 3	1.6.3 偏智能化、自动化分析的数据 挖掘应用方法论 24
1.1.1 SPSS 发展简史与版本选择 3	思考与练习 25
1.1.2 SPSS 的产品定位 5	第2章 数据录入与数据获取 26
1.1.3 SPSS 的基本特点 6	2.1 CCSS 案例项目背景 26
1.1.4 SPSS 的客户机/服务器结构 与模块化结构 6	2.1.1 项目背景 26
1.2 SPSS 操作入门 7	2.1.2 项目问卷 27
1.2.1 SPSS 的安装与激活 7	2.2 数据格式概述 28
1.2.2 SPSS 的启动与退出 8	2.2.1 统计软件中数据的录入 格式 28
1.2.3 SPSS 的操作方式 9	2.2.2 变量属性 29
1.2.4 SPSS 对话框操作基本规范 9	2.3 数据的直接录入 33
1.3 SPSS 的窗口、菜单和结果输出 11	2.3.1 操作界面说明 33
1.3.1 SPSS 的4种窗口 11	2.3.2 开放题和简单单选题的 录入 34
1.3.2 SPSS 的菜单 12	2.3.3 多选题的录入 37
1.3.3 SPSS 的4种结果输出 14	2.4 外部数据的获取 39
1.3.4 分析结果的保存和导出 16	2.4.1 读取电子表格数据文件 39
1.4 SPSS 的系统选项、中文化设置与 附加安装包 17	2.4.2 读取文本数据文件 41
1.4.1 SPSS 的系统选项与中文化 设置 17	2.4.3 用 ODBC 接口读取各种数据 库文件 43
1.4.2 SPSS 网站提供的附加安 装包 18	2.5 数据的保存 44
1.5 SPSS 的帮助系统 19	2.6 数据编辑窗口常用操作技巧集锦 45
1.5.1 学习向导 19	思考与练习 48
1.5.2 帮助菜单 21	第3章 变量级别的数据管理 49
1.5.3 针对高级用户的帮助功能 22	3.1 变量赋值 50
1.6 数据分析方法论概述 22	3.1.1 常用基本概念 50
1.6.1 严格设计支持下的统计 方法论 23	

3.1.2 “计算变量”过程对话框	51	4.4.3 复制数据属性	80
3.1.3 案例:年龄变量 S3 的分组	51	4.4.4 新建自定义属性和设置未知 测量属性	81
3.2 已有变量值的分组合并	52	4.5 与数据准备有关的功能	82
3.2.1 对连续性变量进行分组 合并	52	4.5.1 SPSS 中与数据准备相关的 功能	82
3.2.2 分类变量类别的合并	53	4.5.2 数据验证模块	83
3.3 连续性变量的离散化	54	4.5.3 标识重复个案	85
3.3.1 可视离散化过程	54	4.5.4 标识异常个案	87
3.3.2 最优离散化过程	55	思考与练习	89
3.4 变量的自动重编码与数值移动	57	第 5 章 SPSS 编程与扩展	90
3.4.1 变量的自动重编码	57	5.1 SPSS 编程入门	90
3.4.2 变量值的移动	58	5.1.1 基本语法规则	90
3.5 转换菜单中的其他功能	59	5.1.2 SPSS 程序的创建方式	92
3.5.1 指定数值的查找与计数	59	5.1.3 结构化语句简介	93
3.5.2 变量的编秩	59	5.1.4 一个简单程序示例	95
3.5.3 自动准备建模数据	60	5.2 语法编辑窗口操作入门	96
3.5.4 随机数字生成器	61	5.2.1 语法编辑窗口界面	96
思考与练习	62	5.2.2 程序的运行与调试	97
第 4 章 文件级别的数据管理	63	5.3 INCLUDE 命令与宏程序	98
4.1 几个常用过程	63	5.3.1 INCLUDE 命令	98
4.1.1 排序个案	63	5.3.2 宏程序	99
4.1.2 分割文件	65	5.4 OMS 系统与程序自动化	100
4.1.3 选择个案	65	5.4.1 OMS 系统	100
4.1.4 加权个案	67	5.4.2 程序自动化	103
4.1.5 分类汇总	68	思考与练习	104
4.2 数据文件的重组与转置	70	第 6 章 统计实战案例集锦(一)	105
4.2.1 数据的长型与宽型格式	70	6.1 数据异常值的自动核查与报告	105
4.2.2 长型格式转换为宽型格式	71	6.1.1 项目背景	105
4.2.3 宽型格式转换为长型格式	73	6.1.2 分析思路	106
4.2.4 数据转置	74	6.1.3 利用数据验证模块实现 查错	106
4.3 多个数据文件的合并	75	6.1.4 利用函数功能实现查错	108
4.3.1 一些基本概念	75	6.1.5 项目总结与讨论	110
4.3.2 数据文件的纵向拼接	75	6.2 CCSS 项目数据的自动计算与 处理	110
4.3.3 数据文件的横向合并	76		
4.4 与数据字典有关的功能	78		
4.4.1 数据字典的基本概念	78		
4.4.2 定义变量属性	79		

6.2.1 项目背景	110	6.2.4 项目总结与讨论	114
6.2.2 分析思路	111	思考与练习	114
6.2.3 具体操作	112		

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

117

7.1 连续变量的统计描述指标体系

117

7.1.1 集中趋势的描述指标

118

7.1.2 离散趋势的描述指标

119

7.1.3 分布特征、其他趋势的

描述指标

120

7.1.4 SPSS 中的相应功能

121

7.2 连续变量的参数估计指标体系

122

7.2.1 正态分布

122

7.2.2 参数的点估计

123

7.2.3 参数的区间估计

124

7.2.4 SPSS 中的相应功能

125

7.3 案例:信心指数的统计描述

125

7.3.1 使用频率过程进行分析

125

7.3.2 使用描述过程进行分析

127

7.3.3 使用探索过程进行分析

128

7.4 Bootstrap 方法

131

7.4.1 模型

131

7.4.2 案例:对总指数进行 Bootstrap

估计

132

思考与练习

134

第8章 分类变量的统计描述与参数估计

135

8.1 指标体系概述

135

8.1.1 单个分类变量的统计

描述

135

8.1.2 多个分类变量的联合

描述

136

8.1.3 多选题的统计描述

136

8.1.4 分类变量的参数估计

137

8.1.5 SPSS 中的相应功能

137

8.2 案例:对学历等背景变量进行

描述

138

8.2.1 使用频率过程进行描述

138

8.2.2 使用交叉表过程进行

描述

138

8.3 案例:对多选题 C0 还贷状况进行

描述

140

8.3.1 多选题的频数列表

140

8.3.2 多选题的列联表分析

141

思考与练习

143

第9章 数据的报表呈现

144

9.1 统计表入门

144

9.1.1 统计表的基本框架

144

9.1.2 表头、数据区与汇总项

145

9.1.3 单元格的数据类型

146

9.1.4 几种基本表格类型

146

9.1.5 SPSS 中的报表功能

148

9.1.6 SPSS 中统计表的基本绘制

步骤

149

9.2 简单案例:题目 A3 的标准统计

报表制作

149

9.2.1 案例简介

149

9.2.2 绘制表格基本框架

150

9.2.3 设置摘要统计量及格式

152

9.2.4 调整各种显示细节

153

9.3 复杂案例:题目 A3a 的标准统计

报表制作

154

9.3.1 案例简介

154

9.3.2 多选题、表格基本框架及

汇总项的设定

155

9.3.3 设定分类变量小结和汇

总项	155	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	192
9.3.4 对话框的其他选项卡	157	10.5.4 条图的编辑	194
9.4 表格的编辑	158	10.5.5 带误差线的条图与误 差图	194
9.4.1 基本编辑操作	159	10.6 线图、面积图、点图与垂线图	197
9.4.2 主要编辑菜单功能	160	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	197
9.4.3 表格属性的详细设置	161	10.6.2 线图的编辑	198
9.5 表格模板技术	163	10.6.3 面积图、点图与垂线图	199
9.5.1 模板技术简介	163	10.7 散点图	200
9.5.2 表格的中文兼容问题的 解决	165	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	200
思考与练习	165	10.7.2 散点图的编辑	201
第 10 章 数据的图形展示	166	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	203
10.1 统计图概述	166	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	204
10.1.1 统计图的基本框架	166	10.7.5 三维散点图	205
10.1.2 统计图的种类	168	10.8 P-P 图和 Q-Q 图	206
10.1.3 SPSS 的统计绘图功能	171	10.8.1 P-P 图	206
10.2 直方图与茎叶图	171	10.8.2 Q-Q 图	208
10.2.1 案例:绘制消费者信心值 的直方图	172	10.9 控制图与 Pareto 图	208
10.2.2 图形的基本编辑操作	174	10.9.1 控制图	208
10.2.3 直方图图形框架的修改	178	10.9.2 Pareto 图	211
10.2.4 直方图的衍生图形	180	10.10 其他统计图	212
10.2.5 茎叶图	182	10.10.1 高低图	212
10.3 箱图	183	10.10.2 ROC 曲线	213
10.3.1 案例:用箱图分月份考察消 费者信心的分布	183	10.10.3 时间序列分析中使用的 图形	215
10.3.2 箱图的编辑	184	思考与练习	216
10.4 饼图	186	第 11 章 统计实战案例集锦(二)	217
10.4.1 案例:分城市、月份考察 样本性别比例	186	11.1 探索消费者信心指数随背景资料的	
10.4.2 饼图的编辑	187		
10.5 条图与误差图	188		
10.5.1 简单条图案例:比较不同职 业人群的消费者信心值	189		
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	190		

变化规律	217	生产	225
11.1.1 项目背景	217	11.2.1 项目背景	225
11.1.2 分析思路	217	11.2.2 分析思路	225
11.1.3 具体操作	218	11.2.3 具体操作	227
11.1.4 项目总结与讨论	224	11.2.4 项目总结与讨论	230
11.2 CCSS 项目分析报告的自动化		思考与练习	230

第三部分 常用假设检验方法

第 12 章 分布类型的检验	233	第 13 章 连续变量的统计推断(一)—— t 检验	251
12.1 假设检验的基本思想	233	13.1 t 检验概述	251
12.1.1 问题的提出	233	13.1.1 t 检验的基本原理	251
12.1.2 假设检验的标准步骤	234	13.1.2 SPSS 中的相应功能	253
12.1.3 假设检验的两类错误	235	13.2 样本均数与总体均数的比较	253
12.1.4 假设检验中的其他问题	235	13.2.1 单样本案例:基期一线 城市信心指数与基准 值的比较	253
12.2 正态分布检验	236	13.2.2 单样本 t 检验中的其他 问题	255
12.2.1 K-S 检验的原理	236	13.3 成组设计两样本均数的比较	256
12.2.2 案例:考察信心指数分布是 否服从正态分布	236	13.3.1 方法原理	256
12.2.3 使用旧对话框分析案例	240	13.3.2 案例:不同收入水平家庭的 信心指数比较	257
12.3 二项分布检验	241	13.3.3 适用条件与方差齐性 检验	259
12.3.1 二项分布检验的原理	241	13.4 配对设计样本均数的比较	260
12.3.2 案例:考察抽样数据的性别 分布是否平衡	241	13.4.1 方法原理	261
12.3.3 使用旧对话框分析案例	242	13.4.2 案例:治疗前后舒张压均 数的比较	261
12.4 游程检验	243	13.5 本章小结	263
12.4.1 游程检验的原理	243	思考与练习	263
12.4.2 案例:考察 CCSS 抽样数据 是否随机	244	第 14 章 连续变量的统计推断(二)—— 单因素方差分析	265
12.4.3 使用旧对话框分析案例	245	14.1 方差分析简介	265
12.5 蒙特卡罗方法	247		
12.5.1 蒙特卡罗方法简介	247		
12.5.2 蒙特卡罗方法的 SPSS 实现	247		
12.6 本章小结	249		
思考与练习	250		

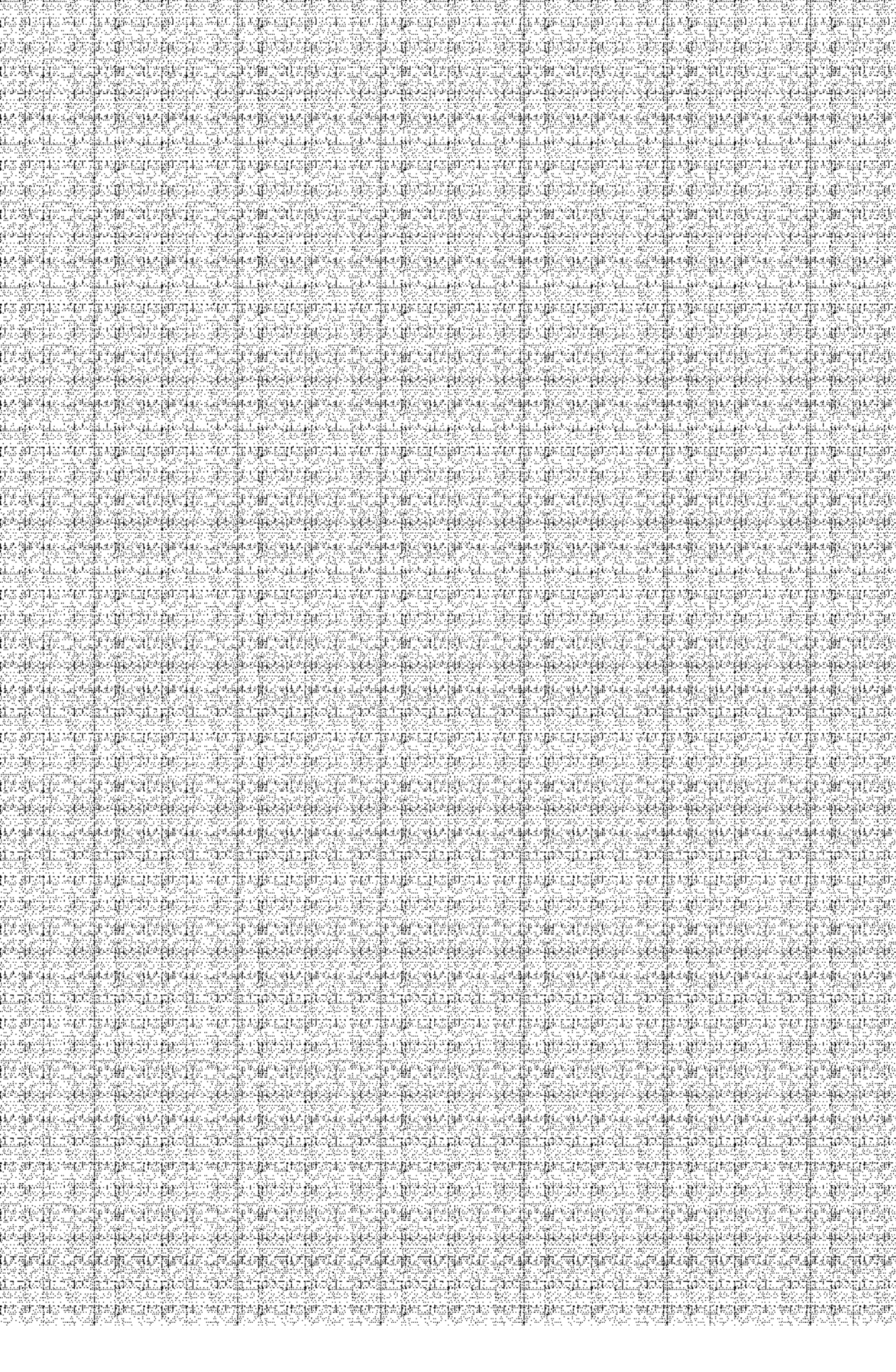
14.1.1 进行方差分析的原因	265	15.3 两个独立样本的非参数检验	291
14.1.2 方差分析的基本思想	265	15.3.1 方法原理	291
14.1.3 单因素方差分析的应用 条件	267	15.3.2 案例:不同收入家庭经济 现状感受值的比较	293
14.2 案例:不同时间消费者信心 指数的比较	269	15.3.3 使用旧对话框分析案例 ...	294
14.3 均数间的多重比较	272	15.4 多个独立样本的非参数检验	295
14.3.1 直接校正检验水准	272	15.4.1 方法原理	296
14.3.2 专用的两两比较方法	273	15.4.2 案例:不同时间上的家庭经 济现状感受值比较	296
14.3.3 两两比较方法的选择 策略	274	15.4.3 使用旧对话框分析案例 ...	299
14.3.4 多重比较结果出现矛盾 时的解释	275	15.5 多个相关样本的非参数检验	300
14.3.5 案例:不同时间信心指数 的两两比较	275	15.5.1 Friedman 检验	300
14.4 各组均数的精细比较	277	15.5.2 案例:不同时间的世博会入 园人数比较	301
14.4.1 方法原理*	277	15.5.3 使用旧对话框分析案例 ...	303
14.4.2 案例:事先计划的两时间 均数比较	278	15.5.4 Kendall 协和系数检验与 Cochran 检验	303
14.5 组间均数的趋势检验	279	15.6 秩变换分析方法	306
14.5.1 方法原理	279	15.6.1 秩变换分析原理简介	306
14.5.2 案例:前3个时间的信心 指数线性趋势检验	280	15.6.2 案例:用秩变换来比较不同 时间的家庭经济感受值	306
14.6 本章小结	281	15.7 本章小结	307
思考与练习	281	思考与练习	308
第15章 有序分类变量的统计推断—— 非参数检验	283	第16章 无序分类变量的统计推断—— 卡方检验	310
15.1 非参数检验概述	283	16.1 卡方检验概述	310
15.1.1 非参数检验的意义	283	16.1.1 卡方检验的基本原理	310
15.1.2 非参数检验预备知识	284	16.1.2 卡方检验的用途	311
15.2 两个配对样本的非参数检验	285	16.1.3 SPSS 中的相应功能	311
15.2.1 方法原理	285	16.2 单样本案例:考察抽样数据的 性别分布	312
15.2.2 案例:北京大学与清华 大学2002年高考录取 分数比较	286	16.2.1 用新对话框界面分析本 案例	312
15.2.3 使用旧对话框分析案例 ...	289	16.2.2 使用旧对话框分析案例 ...	314
		16.3 两样本案例:不同收入级别家庭的 轿车拥有率比较	315

16.4 两分类变量间关联程度的度量	318
16.4.1 相对危险度与优势比	318
16.4.2 案例:计算家庭收入级别和轿车拥有情况的关联程度	319
16.5 一致性检验与配对卡方检验	320
16.5.1 Kappa 一致性检验	320
16.5.2 配对卡方检验	322
16.6 分层卡方检验	322
16.7 本章小结	325
思考与练习	325
第 17 章 相关分析	327
17.1 相关分析简介	327
17.1.1 相关分析的指标体系	327
17.1.2 SPSS 中的相应功能	329
17.2 简单相关分析	331
17.2.1 方法原理	331
17.2.2 案例:考察信心指数值和年龄的相关性	333
17.2.3 秩相关系数	335
17.2.4 Kendall 等级相关系数	336
17.3 偏相关分析	336
17.3.1 方法原理	336
17.3.2 案例:控制家庭收入的影响之后考察年龄的作用	337
17.4 Distance 过程	338
17.4.1 距离测量与相似性测量的指标体系	339
17.4.2 案例:基因间距离的计算	340
17.5 本章小结	342
思考与练习	342
第 18 章 线性回归模型入门	343
18.1 线性回归模型简介	343
18.1.1 相关分析与回归分析的联系与区别	343
18.1.2 简单回归分析的原理和要求	344
18.2 案例:建立用年龄预测总信心指数值的回归方程	346
18.3 多重线性回归模型入门	350
18.3.1 模型简介	350
18.3.2 多重线性回归模型的标准分析步骤	350
18.3.3 回归方程中的自变量筛选方法	353
18.3.4 SPSS 中与多重线性回归模型相关的功能	354
18.3.5 案例:建立自变量包括年龄、家庭收入的信心指数回归方程	355
18.4 本章小结	359
思考与练习	359
第 19 章 统计实战案例集锦(三)	360
19.1 X 药物对原发性高血压治疗的临床试验研究	360
19.1.1 项目背景	360
19.1.2 研究方法	360
19.1.3 数据准备	361
19.1.4 基线情况比较	363
19.1.5 疗效比较	366
19.1.6 安全性评价	367
19.1.7 分析结论与总结	369
19.2 咖啡屋需求调查案例	370
19.2.1 项目背景	370
19.2.2 数据预分析	372
19.2.3 主体问卷分析	374
19.2.4 项目总结与讨论	379
19.3 牙膏新品购买倾向研究案例	379
19.3.1 研究背景	379

19.3.2 分析思路	380	19.4.1 项目背景	388
19.3.3 数据预分析	381	19.4.2 数据的采集	388
19.3.4 数据建模	384	19.4.3 数据预分析	389
19.3.5 项目总结与讨论	387	19.4.4 数据建模	390
19.4 证券业市场绩效与市场结构关系 的实证分析	388	19.4.5 项目总结与讨论	392
		思考与练习	393
附录			394
附录 1 SPSS 函数一览表			394
附录 2 各种情形下最常用统计检验方法索引			405
附录 3 统计术语英汉名词对照表			407
附录 4 IBM SPSS Statistics 19/20 介绍			413
参考文献			416

第一部分

数据管理与软件入门



第1章 SPSS 入门

1.1 SPSS 概述

SPSS 软件是世界上应用最广泛的专业统计软件之一,在全球约有 25 万用户,分布于通信、医疗、银行、证券、保险、制造、商业、市场研究和科研教育等多个领域和行业,全球 500 强中约有 80% 的公司使用 SPSS,而在市场研究和市场调查领域则拥有超过 80% 的市场占有率,和 SAS 被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 SPSS 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,基于一系列原因,SPSS 公司做出了一个令广大用户无法接受的决定: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。但无论名称如何更改,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 发展简史与版本选择

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(两位博士生、一位硕士生)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,该软件和相应成立的 SPSS 公司走上了持续发展的创新之路。

1. 公司与软件简史

(1) 1968—1975 年:SPSS 成为真正的产品。从一个雏形开始,经过不断的代码积累和修改,SPSS 终于成为成熟的、可销售的产品。

(2) 1975—1984 年:SPSS 公司成为真正的公司。在一系列探索之后,SPSS 公司终于确立了以统计软件和统计分析服务为主业的方向。

(3) 1984—1992 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,该版本为全球第一套以图形菜单为驱动界面的统计软件,也是 DOS 时代的统计软件经典之作。

(4) 1992—1996 年:Windows 时代。在 1992 年,SPSS 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,公司也走上了快速发展之路,并收购了诸如 SYSTAT (1994) 和 Jandel (1996) 等一系列同行企业。

(5) 1997—2002 年:向大企业进化。SPSS 软件在不断推陈出新,经典的 11 版就在这一期间推出,更重要的是并购行动在继续,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并购,这也意味着公司开始形成完整的产品线,SPSS 这一产品的定位及重要性开始下降。

(6) 2003—2008 年:向预测分析转型。在完成上述并购后,SPSS 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 软件被定位为产品线中的普及类工具,和其余产品形成高低端搭配。但这一过程并不顺利,显然市场的成熟速度落后于预期,但公司坚持了下来,并走完了这一段路。SPSS 软件也仍然在不断更新,13 版堪称又一经典,从 17 版开始则提供了基本成熟的中文界面与结果输出。

(7) 2009—现在:新的一页。随着 IBM 的收购,SPSS 产品揭开了新的一页,已站在了更高的平台之上,未来表现令人期待。而最新的 SPSS 19 及 SPSS 20 则为其并购之后的作品,其软件界面已经彻底改变为 IBM 的蓝色风格。

2. 软件版本比较

由于 SPSS 大约 1 年时间就会推出一个新版本,导致用户使用的版本可能很多,并不一定都是最新版。为了便于读者选择,这里列出从 11 版至今各版本的基本情况 & 笔者的评价。

(1) 11 版:重新设计了软件界面,加入了混合线性模型等新方法,随后的 11.5 版又新增了 Custom Tables 模块,并提供多语言输出,并首次能将结果直接导出为 XLS 文件。

(2) 12 版:图形输出更改为目前使用的系统,加入了复杂抽样模块,也是首个提供简体中文版(单独销售)的版本。由于 12 版是软件开发转向 Java 之后的第一个作品,产品质量实在不能算成功,当时 SAS 的新版本也因为采用了 Java 语言而未能成功,两种软件相当。

(3) 13 版:真正的经典之作,很多用户目前仍在使⤵用。开始加入树模型等智能统计分析方法,复杂抽样模块等有了较大更新,输出接口加入了 OMS 系统。对于配置较低(内存存在 1 GB 以下,CPU 仍在迅驰之前的级别)的旧机器,笔者强烈推荐使用该版本。

(4) 14 版:首次可以同时打开多个数据文件,提供了现在已成为主要绘图界面的新的“图表生成器”界面,新增了 Data Validation 模块,此外还提供了很多算法、数据管理、统计方法等细节的更新。

(5) 15 版:提供了 Programmability Extension 功能,可直接调用 Python 等语言编写的代码。加入了 GEE 等统计模型。该版本也较为经典,但代价是和 13 版相比,对系统的硬件要求明显提高。

(6) 16 版:用 Java 重写了整个用户界面,操作更加灵活。加入了神经网络和 PLS(偏最小二乘法),提供了对 R 语言的支持。结果文件也改成了新的 spv 格式。

(7) 17 版:首次提供了包括简体中文的多语言界面,至此 SPSS 中文版才开始在国内得到广泛使用。引进了 SPSS EZ RFM 模块、最近邻分析等一系列比较特殊的分析方法/模块,对语法窗口做了大幅升级,增加了最受 SPSS Statistics 专业用户欢迎的新功能,例如,自动完成、Syntax 代码字符颜色标记、代码行数和断点展示等功能。

(8) 18 版:增加了 Bootstrapping 和 Direct Marketing 两大模块,以及一系列统计方法、数据管理、用户界面、第三方接口等方面的更新,而且每个模块均可独立存在并运行,不再依赖于 Base

模块。应当说该版本做得较好,但 PASW 的 Logo 让老用户很不习惯。

(9) 19 版:作为公司并购后推出的第一个版本,IBM 暂时并未对软件做太大改动,除了更换了 Logo 和配色体系外,主要进行了一些细节上的更新,如提供了广义线性混合模型,对语法编辑器做了较大的易用性增强等,此外开始对 IBM 的服务器硬件开始提供支持。

(10) 20 版:该版本实际上更像是对上一版本的问题修正,除重新纳入老版本中就有,但后被取消的统计地图功能外,还对结果输出、服务器功能等方面进行改进,但统计功能则未做明显提升,仅在广义线性模型上增加了有序分类因变量的建模功能。

3. 软件版本的选择

从上述介绍可以看出,SPSS 软件并非每次更新都会有重大变化,也并非每个新版本都值得尝试,基于笔者十几年对 SPSS 的使用经验,这里给出版本选择的建议如下:

- (1) 低配置机器,比如 2007 年以前购置的计算机,建议使用 13 版。
- (2) 对统计术语不熟悉的用户,选择 17、19、20 版皆可,这样可以使用中文界面。
- (3) 对统计术语熟悉的用户,根据机器硬件配置,选择 15~20 版皆可。
- (4) 需要 R、Python 等第三方扩展功能的用户,应尽量使用最新版本。

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,就无法在最大程度上发挥其功用。前面已经提到,SPSS 软件在公司产品线中的地位近年来在不断下降,实际上,目前 SPSS 公司原产品已经形成了由 4 大系列产品构成的完整的产品线,具体如下:

(1) Data Collection Family:定位为中低端的数据采集与报告需求,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 Survey)、电话调查(Phone Survey)、离线调查(Offline Survey)、数据录入(Data Entry)、调查报告(Survey Reporting)和调查管理(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,由原先的 SPSS 软件构成,但 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,或者将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,主攻高端的数据挖掘与商务智能需求领域,也是原公司资产中最有价值的一块。随着 IBM 的收购,该系列必将发生令人惊讶的变化。

(4) Deployment Family:相对而言是对前 3 个产品系列的整合与后台支持,可以把市场调研、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。作者认为该产品被 IBM 收购之后前景不明,未来可能会被逐渐整合。



由上可知,SPSS 的定位就是针对常规统计分析应用的统计软件,用它来进行数据管理或者数据挖掘都不合适。虽然它也提供神经网络等数据挖掘方法,但也仅仅是提供方法,并不适合作为一个标准的数据挖掘平台来使用。

1.1.3 SPSS 的基本特点

SPSS 得到用户广泛欢迎并长盛不衰的原因在于其强大的统计分析 with 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钟爱。

(1) 功能强大: SPSS 囊括了各种成熟的统计方法与模型, 为统计分析用户提供了全方位的统计学算法, 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。如利用值标签来快捷录入数据、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结构、内容更易于分析。

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、更加直接。同时利用 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, 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到的 Excel 表格数据、文本格式数据导入 SPSS 中进行分析, 从而节省了相当大的工作量, 并且避免了因复制和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 格式等, 且目前已彻底解决了中文兼容问题, 用户不需任何附加设定就可以自由使用中文, 并且可以在 Word 等软件中直接使用中文输出结果。

(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, 提供了强大的统计功能, 也因为它是一种非常简单易用的软件。人机界面的友好、操作的简单, 使得各位统计分析人员对它青睐不已。另外, SPSS 也向高级用户提供了编程功能, 使其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 终于找到了一个很巧妙的解决办法, 就是直接和强大的 R 语言进行对接, 通过直接调用 R 语言的各种统计模块, 直接实现了对最新统计方法的调用, 从而彻底解决了这一问题。

1.1.4 SPSS 的客户机/服务器结构与模块化结构

1. SPSS 的客户机/服务器结构

SPSS 软件自 10 版本以来, 已发展为 Client/Server 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 的计算能力来解决速度慢、网络阻塞等由于数据量大而引起的问题。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说的 SPSS, 实际上指的单机版。

2. SPSS 的模块结构

无论是 SPSS 客户机还是 SPSS 服务器, 均是模块式结构, 即它把自己的所有功能分散为多个模块。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, 自己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而有所变化, 在 18 版以前的版本中, SPSS 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 SPSS Base 搭建的平台上才能工作。从 18 版起,其余模块也可以脱离 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,如表 1.1 所示。

表 1.1 SPSS 常见模块与功能对应表

模块名称	功能
Statistic 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
Custom 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的,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,虽然在软件安装完毕之后,在软件菜单中可能会出现所有模块的菜单项,但如果没有购买相应模块的许可证,则相应的模块是无法运行的。



有一点需要澄清:国内许多 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 软件只有一个版本,不存在所谓的标准版和专业版之分,差异只在于各模块是否均可使用而已。

1.2 SPSS 操作入门

1.2.1 SPSS 的安装与激活

1. SPSS 的安装

以 IBM SPSS Statistics 20 版为例,其安装文件分为 32 位和 64 位两种,被集中放置在容量超过 4GB 的一张 DVD 中,光盘中除 Windows 版的 SPSS 和 AMOS 安装文件之外,还包括了 Mac OS 版 SPSS 的安装文件。本书只讲解 Windows 版 SPSS 软件的情况。

SPSS 在 Windows 系统下的安装与其他软件并无太大差异,同样是在安装光盘上启动安装程序,然后按照界面的说明进行操作。在较老的版本中,可能需要选择希望安装的模块;而在较新的版本中,模块都是默认全部安装的,反而是需要在安装时选择所需的语言种类。显然对于本书的绝大多数读者而言,英文和简体中文语言包是必选的内容。



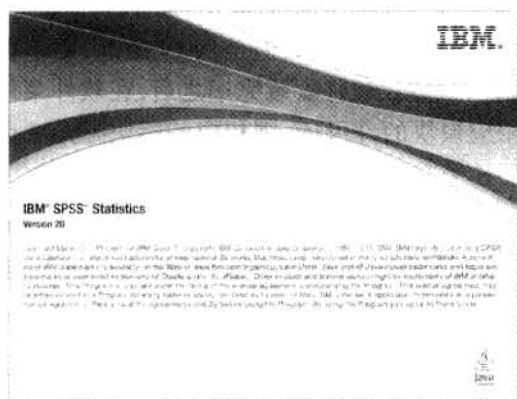
如果用户希望在同一台机器上安装不同版本的 SPSS 软件,只需要按照先安装低版本,再安装高版本的顺序,并将其分别安装在不同目录中即可,这些不同版本的 SPSS 可以并存,一般不会相互冲突。

2. SPSS 的激活

SPSS 在刚安装完毕时,尚未进行软件授权确认,此时至多只能获得一定的试用期,过了试用期软件将会被自动锁闭。用户需要在“开始”菜单中找到 IBM SPSS Statistics 组,然后运行其中的“IBM SPSS Statistics 20 许可证授权向导”,在连网的状态下输入授权码以将软件激活,激活完毕后所购买的模块就可以正常使用了。

1.2.2 SPSS 的启动与退出

以 Windows 系统为例,在“开始”菜单中找到 IBM SPSS Statistics 组(老版本可能为 SPSS Inc 组),选择其中的启动项 IBM SPSS Statistics 20(老版本则可能为 SPSS for Windows),就会启动 IBM SPSS Statistics 20,其闪屏如图 1.1(a)所示,之后就会打开 SPSS 的数据编辑窗口。对于第一次使用 SPSS 的用户,系统会弹出使用向导(如图 1.1(b)所示),用户可在其中选择所需的操作,如果不希望该向导再出现,则选中右侧的“输入数据”单选按钮,然后选中左下角的“以后不再显示此对话框”复选框并单击“确定”按钮即可。



(a)



(b)

图 1.1 SPSS 20 启动时的闪屏和使用向导

如果要关闭该软件,则选择“文件”→“退出”菜单项,或者直接关闭窗口,即可退出 SPSS。



本书编写时所采用的软件环境为:Windows 7 32/64 位版、SPSS 20 32/64 位版,如果采用其他 Windows 版本或 SPSS 版本,可能操作界面和对话框会略有差异,特在此提醒用户。

1.2.3 SPSS 的操作方式

1. 统计软件的常见操作方式

初学者对 SPSS 存在的一个广泛误解是 SPSS 只使用菜单对话框方式来操作。实际上,在经历了几十年的发展之后,现今任何一个成熟的统计软件都会提供从初学者到专家各个层面所需要使用的操作方式。具体而言,统计软件常见的操作方式有如下几种。

(1) 命令行方式:即用户一条一条地提交命令,软件系统直接对命令进行解释执行,用户再根据执行结果提交下一条命令。这是出现最早的一种操作方式,也是目前大型统计软件中 Stata 的主要操作方式。

(2) 程序方式:由于命令行方式无法实现一些复杂功能,因此随后就出现了将命令组合成程序,用户批量提交,系统按程序要求执行,批量输出结果的程序方式。程序方式不仅可以提高运行效率,还可以利用程序结构中的分支、循环等语句来实现更复杂的统计功能。目前统计软件中的 SAS 就是以程序方式作为主要执行方式的。

(3) 菜单对话框方式:由于程序方式需要用户首先学习编程语法规则,对于初学者来说学习门槛较高,因此 SPSS 针对初学者的需求在全球首家推出了以菜单对话框为主要操作方式的软件版本,在随后的几十年里这种操作方式成为 SPSS 的主要操作方式。这一方式随着 Windows 系统的普及而得到了极大的发展,已经成为了最受各阶层统计软件用户欢迎的方式。

需要指出的是,各种大型统计软件实际上都支持上述 3 类操作方式,只是有其首选方式和最佳操作方式的区别而已。例如,上面提到的 Stata 和 SAS 都支持菜单对话框方式,而 SPSS 也支持程序方式。

2. SPSS 对各种操作方式的支持和扩展

具体而言,作为全球知名的统计软件,SPSS 对上述几种操作方式都是支持的,还进一步提供了一些扩展,其重点在于提高软件的易用性。

(1) 菜单对话框方式:用户可以自定义对话框,以便将工作中一些常用的程序直接以对话框方式实现,其实质就是对软件进行二次开发。对菜单对话框操作方式基本规则的介绍可参见下文。

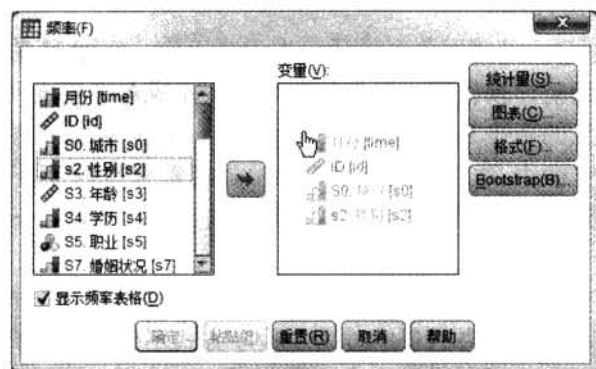
(2) 命令行/程序方式:作为最易学、易用的统计软件,SPSS 对程序方式做了很大的改进。首先可以利用对话框中的“粘贴”按钮自动生成程序,其次可以通过宏、Include 命令等方法使得已有代码段得到充分利用;最后,SPSS 还提供了程序全自动执行的“生产工作(Production Facility)”方式,进一步简化了操作。对程序方式操作细节的详细介绍参见本书第 5 章。

1.2.4 SPSS 对话框操作基本规范

SPSS 的对话框都遵循统一的操作规范,因此这里首先介绍一下相应的规范,以便于后续学习。

1. 对话框元素构成

这里以“频率”对话框为例,选择“分析”→“频率”菜单项,打开“频率”对话框,如图 1.2 所示。




(a)



(b)

图 1.2 “频率”对话框及其子对话框

(1) 变量列表框:图 1.2(a)中共有两个列表框,左边列表为候选变量(包含当前数据文件中的可分析变量或指定变量集)列表,右边列表为需分析变量列表。每个变量均按照“测量尺度+变量名标签+[变量名]”这种 3 段式结构来显示,例如图 1.2(a)中的变量 time,可以看到其测量尺度为定序,变量名标签为“月份”,这里提到的这些变量属性的详细介绍参见第 2 章。

(2) 变量移动按钮:即图 1.2(a)中的,用于在待选和分析列表中移动变量。在选中变量时,该移动按钮会变黑,表示可用,同时箭头方向会指向可移动的方向。

(3) 标准按钮:位于图 1.2(a)所示对话框的下部,几乎在所有的 SPSS 对话框中均可见到,由“确定”、“粘贴”、“重置”、“取消”、“帮助”5 个按钮组成。其中“粘贴”指的是将当前的对话框设定转换为 SPSS 程序,详见第 5 章的讲解。

(4) 其他按钮及选项:根据具体功能,不同的对话框还会出现一些特殊的按钮,单击后一般会弹出子对话框,对相应的操作做进一步的设定。如图 1.2(a)所示对话框最右侧有 4 个按钮,分别对本次分析中的某些细节做进一步的设定,如单击“统计量”按钮会打开有关“统计量”的子对话框,如图 1.2(b)所示。

(5) 二级对话框:由于统计功能的选项较多,许多对话框会将一类功能放在一起,做成一个二级对话框,在一级对话框上用一个按钮来调用,如同图 1.2(b)所示的“统计量”二级对话框一样。二级对话框中常见的元素有单选按钮、复选框、下拉列表框、文本框等,对于熟悉 Windows 操作的读者而言,这些元素的功能不言自明,这里只需要指出一点:在 SPSS 中各个对话框元素往往按照其功能被分成若干个组,每个组都执行某一方面的操作,如图 1.2(b)中共出现了 4 个框组,即“百分位值”、“集中趋势”、“离散”和“分布”,使用户清楚这些框组中元素所具有的功能,本对话框的具体功能解释可参见第 7 章,这里不再展开讲述。

2. 对话框基本操作规律

(1) 按钮颜色:注意当按钮为灰色时,表示当前对话框设定尚不满足使用条件,只有满足条件后相应按钮才会变黑可用。

(2) 变量的选中:单击列表中的变量名即可选中某个变量,按住 Shift 键可以选中多个连续变量,而按住 Ctrl 键可以选中多个不连续的变量。如果变量列表较长,除了可以使用滚动条拖列表至相应位置外,也可以先单击该列表,然后直接用键盘输入希望查找的变量的变量名标签首字母(或汉字),变量列表就会快速跳跃到标签为相应字母的变量处。

(3) 变量的移动:在选中变量后,即可以单击相应的变量移动按钮将选中的变量移动到新的框组中,也可以直接用鼠标左键进行拖放操作。另外,对于已选入需分析变量列表框中的变量,直接双击就可使其退回候选变量列表框中。

(4) 更改变量的显示与排序方式:在候选变量列表框中单击右键,可以更改变量显示方式(变量名或者标签)、排序方式(字母顺序、文件顺序或者测量尺度),或者显示具体的变量信息。在其余的变量列表框中,则只能显示变量信息。

(5) 更改变量测量尺度:对于图表构建程序对话框等对变量测量尺度有严格要求的对话框,在候选变量列表中选中相应变量并单击右键,就可以在此处直接更改其测量尺度。

1.3 SPSS 的窗口、菜单和结果输出

1.3.1 SPSS 的 4 种窗口

SPSS 是多窗口软件,运行时使用的窗口种类最多有 4 种:数据窗口、输出窗口、语法窗口和脚本窗口,如图 1.3、图 1.4 所示,其中数据窗口和输出窗口是最常用的两个。

time	id	s0	s1	s2	s3
200704	1	100	1	20	4
200704	2	100	1	24	2
200704	3	200	1	20	2
200704	4	100	2	65	3
200704	5	200	2	40	2
200704	6	100	1	50	3
200704	7	100	2	53	1
200704	8	300	1	44	2
200704	9	200	2	35	3
200704	10	200	1	21	4

(a)

统计量	求和	平均值	数量百分比	累积百分比
月份				
200704	300	26.2	26.2	26.2
200712	304	26.5	26.5	52.7
200812	304	26.5	26.5	78.2
200912	239	20.9	20.9	100.0
总计	1147	100.0	100.0	100.0

(b)

图 1.3 SPSS 的数据窗口和输出窗口

(1) 数据窗口(SPSS Data Editor):也称为数据编辑器,此窗口类似于 Excel 窗口,SPSS 处理数据的主要工作全在此窗口中进行。又分为两个视图:数据视图用于显示具体的数据,一行代表一个观测个体(在 SPSS 中称为 Case),一列代表一个属性(在 SPSS 中称为 Variable);变量视图则专门显示有关变量的信息:变量名称、类型、格式等,关于这些术语的详细解释,可参阅第 2 章。

注意在 14 版以上的 SPSS 中,是可以同时打开多个数据文件的,每个数据文件独占一个不同的数据窗口,系统会对这些数据窗口自动按照“数据集 0”、“数据集 1”这样的工作名称来加以区分,详见后续章节的讲解。



(a)



(b)

图 1.4 SPSS 的语法窗口和脚本窗口



老版本的 SPSS 用户需要非常注意数据集工作名称这个新概念,以图 1.3 为例,CCSS Samples. sav 是硬盘上该数据文件的存储名,但当该文件被 SPSS 读入后,SPSS 用于识别该数据集的却是后面中括号里的“数据集 1”,这一点在使用 SPSS 程序进行数据分析时特别重要!

(2) 输出窗口 (SPSS Output Viewer): 也称为结果查看器,此窗口用于输出分析结果。在窗口中进行的操作非常类似于资源管理器。整个窗口分两个区:左边为目录区,是 SPSS 分析结果的一个目录;右边是内容区,是与目录一一对应的内容。

(3) 语法窗口 (SPSS Syntax Editor): 也称为语法编辑器。SPSS 最大的优势在于其简单易用性,即菜单对话框式的操作。除此之外,SPSS 还提供了语法方式或程序方式进行分析。该方法既是对菜单功能的一个补充,也可以使烦琐的工作得到简化,尤其适用于高级分析人员。

(4) 脚本窗口 (SPSS Script Editor): SPSS 脚本是用 Sax Basic 语言编写的程序,具体可以使用的基本版本有 VBA 和 Visual Basic. NET 两种,在脚本中可以像 SPSS 宏一样构建和运行 SPSS 命令,而且可以在命令中利用当前数据文件的变量信息;还可以对结果进行编辑;或者构建一些新的自定义的对话框。脚本可用于使 SPSS 内部操作自动化、使结果格式自定义化、实现 SPSS 新功能、将 SPSS 与 VB 和 VBA 兼容应用程序连接起来。

启动 SPSS 时,默认打开数据编辑窗口。其他窗口可以通过选择“文件”→“新建”/“打开”→相应的窗口名称而打开。



需要指出的是,在目前的 SPSS 版本中,上述 4 类窗口都可以同时打开多个,比如同时打开多个数据文件,或者多个结果文件,也就是说实际工作中使用的窗口数可以远远多于 4 个。而此时 SPSS 系统对数据窗口、输出窗口都是使用工作名称来进行定位的。

1.3.2 SPSS 的菜单

SPSS 的每种窗口都有 10 个以上的菜单,这里以数据窗口为例,简单介绍一下各个菜单项的具体功能。

1. 文件

该菜单用于对文件进行管理,除了常见的“新建”、“打开”、“保存”、“打印”菜单项外,比较特殊的菜单项有以下几个。

(1) 将文件标记为只读:用于锁定当前数据文件为只读状态,如果之后保存文件,则只能重命名并另存。

(2) 重新命名数据集:对当前文件的工作名称进行更改,读者一定注意修改的是工作名称而不是文件名。

(3) 显示数据文件信息:在输出窗口中以表格的形式列出当前文件或指定外部数据文件的信息,包括变量列表信息,以及变量值标签信息等。对于较复杂的数据文件,该功能可以用来查错。

(4) 停止处理程序:用于停止执行当前的 SPSS 命令。如果正在对一个大型的数据执行非常复杂的分析时,中途发现选项设定有误,则可以用此命令让系统停止运算。但并非所有命令的执行都可以中断,许多数据库操作命令(计算变量、合并等)因为涉及数据文件自身的修改,因此是无法中断的。

(5) 缓存数据、开关服务器与存储库:这 3 个菜单项实际上涉及前述 SPSS 的客户机/服务器结构,分别表示将服务器方的数据缓存到本地、切换到新的服务器,以及访问服务器方的数据文件,普通用户可以无视这些功能。

2. 编辑

该菜单用于对当前窗口进行复制、粘贴、剪切等操作,大部分功能望文知义,无须解释。但是最下方提供的是 SPSS 系统的一些选项设定,下面将详细讲解。

3. 视图

该菜单用于对当前窗口视图进行显示切换,也可进行自定义,特别是可以自行设定快捷工具栏和菜单项。

4. 数据与转换

这两个菜单提供数据管理相关的功能,第 3、4 章中将对其进行详细讲解。

5. 分析

该菜单提供了 90% 以上的统计分析功能,以及少数与分析功能紧密相关的统计绘图功能,如质控图、ROC 曲线、时间序列模型相关图形等。可能有的读者会追问:另外 10% 的统计分析功能到哪里去了? 答案是需要用程序方式来实现,如岭回归、典型相关分析等。

6. 直销

该菜单提供了一组用于改善直销活动效果的工具,它可以标识那些用于定义不同消费者群体的人口统计学、购买和其他特征,针对特定目标群体最大限度地提高正面响应率。具体方法包括 RFM 分析、聚类分析和邮政编码响应率等方法,由于该菜单项更多的是基于应用分析需求来划分而不是基于统计方法分类来划分,本书将不对其进行介绍。

7. 图形

该菜单提供了 90% 左右的统计绘图功能,另外 10% 的绘图功能由于和统计分析结合得较为紧密,因此在分析菜单中提供。第 10 章将对统计图的操作进行讲解。

8. 实用程序

该菜单为用户提供了一些比较方便的数据文件管理功能和界面编辑功能,熟悉这些操作有时可以大大简化工作。

(1) 变量:用于显示各个变量的基本信息,包括变量名标签、值标签、存储类型、测量尺度等。

(2) OMS 控制面板与 OMS 标识符:用于对 OMS,即输出管理系统(Output Management System)进行设定,或给出 OMS 系统的标识符列表。该系统为用户提供了提取和控制结果分析窗口中输出内容的功能,属于较为高级和复杂的功能,第9章中将对其进行介绍。

(3) 定义变量集/使用变量集:这两个菜单项是联合使用的,用于将某些变量定义为一个集合(Set),便于分析时调用。该功能主要是在变量相当多的时候有用,比如说数据文件中有300个变量,而现在要进行的分析只涉及其中的20个,那么不妨将这20个变量设定为一个集合,然后在使用的变量集中指定只使用这个新的变量集,这样设定以后,所有的对话框中将只出现相应的20个变量,其余变量则均会被屏蔽掉。

(4) 指定窗口:该菜单项只存在于结果窗口的相应菜单中,并且只有同时存在两个以上的结果窗口时才可用,目的是指定系统输出分析结果时所用的结果窗口。当存在两个以上结果窗口时,SPSS 默认使用当前/最后一个当前结果窗口来输出结果,但这样有时会带来不便,使用指定窗口菜单项就可以将当前结果窗口指定为结果输出窗口,而无论将来分析时它是否仍为当前窗口。

(5) 自动执行:即 SPSS 提供的程序全自动执行方式入口,详情参见本书第5章的讲解。

(6) 其余菜单项:主要是一些较为专业的菜单项,如“运行 Basic 脚本”、“自定义对话框”、“添加扩展束”等,普通用户很难用到,因此这里不再解释。

9. 窗口

用于对各个窗口进行切换和管理,需要指出的是数据窗口中该菜单的第一项“拆分”,用于将整个窗口拆分为4部分,详细介绍参见第2章“冻结行或列”中的相应内容。

10. 帮助

为不同层次的用户提供完整而系统的帮助功能,后面将对其进行详细介绍。

1.3.3 SPSS 的4种结果输出

作为功能强大的统计分析工具,为了使得分析结果能更好地满足用户的需求,SPSS 一共提供了4种格式的统计分析结果:枢轴表、文本格式、统计图表和模型。

1. 枢轴表/轻量表

在 SPSS 中,绝大部分分析结果都以专用的枢轴表格式展示,如图1.5(a)所示。这些表可以是二维表,也可以是多维表,并且都可以直接粘贴到其他应用程序(如 Word、PowerPoint、Excel)中使用。SPSS 的制表功能非常强大,可以很好地满足用户各种情况下的需求,详见报表呈现的相关章节。

由于枢轴表较耗费系统资源,对于低配置机器而言负担较重,因此从19版起 SPSS 又提供了所谓的“轻量表(Lightweight Tables)”格式,和枢轴表相比,主要是关闭了表格编辑状态下的透视、旋转、分层以及大部分格式编辑功能,但显示速度要比枢轴表快很多。随后在20版中,所有的表格都已改用增加了编辑功能的轻量表兼容格式进行输出,这使得结果输出速度大大提高,但相应的表格无法在18及更早版本中进行编辑。

教程的读者而言,此类输出几乎不会被用到。

1.3.4 分析结果的保存和导出

1. 直接保存

SPSS 的分析结果可以保存为 SPSS 自身的格式,即“. spv”格式,只需要选择结果窗口中的“文件”→“保存”菜单项即可。但是这种方式只能将结果文件保存为这种特殊格式。如果希望保存为其余常用格式,则需要使用导出功能。



更早版本的 SPSS 输出文件格式为“. spo”,两者完全不兼容,但是 SPSS 新版本的用户可以在官方网站上下载免费的 SmartViewer 软件来读取 . spo 格式的输出文件。

2. 导出


导出功能可以将结果文件保存为另外几种常用的格式,包括 HTML 格式、Word 格式、Excel 格式和 Text 格式等。其具体操作是:在结果窗口中选择“文件”→“导出”菜单项,或者在工具栏上直接单击按钮,打开如图 1.7 所示的对话框。



图 1.7 SPSS 结果的导出选项

(1) “导出的对象”框组:用于选择希望导出的内容,需要指出的是,由于在结果输出中默认会隐藏运行记录等次要项目。而在导出操作中,默认设置会将这些不常用的内容全部输出,因此这里一般均需更改为“所有可见”以简化导出内容。

(2) “文档”框组:左侧的“类型”下拉列表框用于设置导出格式(Export Format),右侧的“选项”列表框则用于进行格式的具体设定,如需更改,可单击下方的“更改选项”按钮进行修改,对格式做进一步的设定。

(3) “文件名”文本框:指定希望保存的文件名称。

(4) “图形”框组:该下拉列表在指定只导出图形时有效,可用于进一步设定图形格式的细节,如 . bmp、. jpg 等存储格式,以及 24 位、16 位等颜色深度等。

3. 直接复制和粘贴

除了可以保存结果之外,还可以将结果内容直接通过“复制”、“粘贴”命令应用到其他软件中,在默认情况下,枢轴表会自动转换为 Word 或 Excel 中的表格,而统计图则会被转换为图片。

1.4 SPSS 的系统选项、中文化设置与附加安装包

1.4.1 SPSS 的系统选项与中文化设置

从 SPSS 19 版起,如果中文 Windows 系统的用户在安装时选择了中文安装包,则软件启动后会自动使用中文界面,几乎不需要用户另行设定。但如果用户使用的仍然是老版本,或者曾经更改过相应设定,则可能会遇到中文兼容问题,此时只要选择“编辑”→“选项”菜单项,就会打开 SPSS 软件的系统“选项”对话框,如图 1.8 所示。

如果用户已经将软件调整为中文界面和中文输出,那么所有内容不言自明。但是老版本的 SPSS 在安装后一般默认是英文界面,因此这里重点介绍如何将相应的设定更改为简体中文,以及其他一些建议的设定。

(1) 界面语言:见图 1.8 中“常规”选项卡右下侧的“用户界面”框组,“语言”下拉列表框,将其中的设定改为“简体中文”(Simplified Chinese)即可。

(2) 结果输出语言:见图 1.8 中“常规”选项卡右上侧的“输出”框组,“语言”下拉列表框,将其中的设定改为“简体中文”(Simplified Chinese)即可。

(3) 枢轴表默认格式:目前枢轴表的模板设定已不存在中文兼容问题,但统计表格要求没有竖线,默认模板是带有竖线的,因此建议在“枢轴表”选项卡左上角的“表格外观”框组中,将表格模板更改为 Academic,本书随后的结果输出将都使用该表格模板。在 19 版中,建议将右侧的“表呈现”框组修改为“快速呈现表格”,即默认按照轻量表输出。而在 20 版中,如果希望表格格式和 18 及以前版本兼容,则需要选择右侧的“呈现为旧表格”复选框,但代价是结果输出速度明显变慢。

(4) 查看器字体设定:在一般情况下不需要更改,但如果遇到文本输出列对齐混乱的情况,则可以将相应的文本字体设定为 MingLiu,则相应的新输出就会自动列对齐了。在更早一些的 SPSS 版本中,需要将字体设定为“宋体”方可对齐,但在较新的版本中则必须是 MingLiu。



图 1.8 SPSS 的系统“选项”对话框

经过上述设定之后,SPSS 软件的界面和输出就完全中文化了,但在 19 版以前的版本中,默认是不安装中文帮助文件的,其帮助系统默认使用英文进行说明,需要用户在官网下载中文安装包自行安装,详见下节。

1.4.2 SPSS 网站提供的附加安装包

SPSS 的产品支持网址原本是 <http://support.spss.com/>,其中提供了大量附加工具的安装包,但是随着公司合并和整合的进行,该网页的内容将会被整体移动,目前登录该支持网址已经出现相应的说明,提示用户按照说明链接至 IBM 的支持网页进行注册登录。由于该页面还会出现变化,因此这里不再详述具体页面内容及下载操作,仅给出产品支持页面中应当会提供的内容。

1. 简体中文版帮助

如果用户在安装时未选择安装中文内容,则在使用时即使将 SPSS 界面切换为中文,帮助系统也仍然是英文版。另一方面,18 版及以前版本默认是不安装中文帮助的,对此可直接在支持页面中下载相应版本所对应的中文帮助安装包单独安装即可。

2. 结果文件阅读器

上文提到过的 spo 文件阅读器 SPSS SmartViewer,也是在本页面中下载。

3. ODBC 数据驱动包

为 SPSS 的 OEM 版数据驱动,提供了多种数据库的 ODBC 驱动接口。

4. R/Python 语言插件

用于提供在 SPSS 中调用 R 或者 Python 语言的支持,注意下载时需要选择和所用 SPSS 版本相匹配的安装包,否则可能无法正常使用。

5. 用户手册

包含所有模块的手册,用户选择希望阅读的手册 PDF 链接直接下载即可,但目前仅提供英文版下载,这不能不说是一个小遗憾。

6. 系统补丁

SPSS 一般会在 3~6 个月的期间内提供系统升级补丁,以修正所发现的软件错误。相应的软件补丁现在已经被统一放置在 IBM Fix Central Site(<http://www.ibm.com/support/fixcentral>)中提供,读者直接至该页面下载即可。

更多的下载内容请读者自行在上述页面中浏览,这里不再详述。

1.5 SPSS 的帮助系统

SPSS 提供了无处不在的“帮助”功能,可以随时随地为不同层次的用户提供帮助。其帮助功能主要包括学习向导、帮助菜单和高级用户相关的帮助功能 3 大类。事实上,国内有相当一部分 SPSS 教材都是在翻译或引用 SPSS 完整而详细的帮助内容,这里将直接给出原版教材。



从 SPSS 19 版起,帮助系统从原先的 CHM 格式更改成了 IBM 系软件通用的网页格式,但基本框架并无大的更动,因此这里不再单独说明其用法的差异。

1.5.1 学习向导

SPSS 为初学者提供了非常完整和系统的自学向导,它相当于一个手把手的教练,会浅显易懂地告诉用户各种基本的统计分析问题在 SPSS 中是如何实现的。SPSS 中的学习向导有如下几种。

1. 统计辅导

对于需要新手紧急完成的一些常用统计分析操作,SPSS 提供了统计辅导(Statistics Coach)功能,可以告诉用户为达到分析目的应选择什么统计方法,并一步步地指导用户如何进行统计分析。该模块实际上是一个编译好的交互式网页,使用起来非常方便,如图 1.9 所示。该功能通过选择“帮助”→“统计辅导”菜单项即可使用。

2. 教程

教程(Tutorial)同样是为初学者提供的,是关于某个主题的详细指导,以图例化的方式告诉初学者如何使用这个软件。初学者可以通过该教程掌握 SPSS 的几乎全部常用操作(数据的输入、分析和绘图)。选择“帮助”→“教程”菜单项即可使用该功能,其初始界面为一个目录列表,即所有教程内容的索引,用户可在里面选择需要阅读的主题,如图 1.10 所示,如果对 SPSS 完全不熟悉,则可以从最上面的简介开始,里面提供了使用 SPSS 的一些最基本的操作教程。

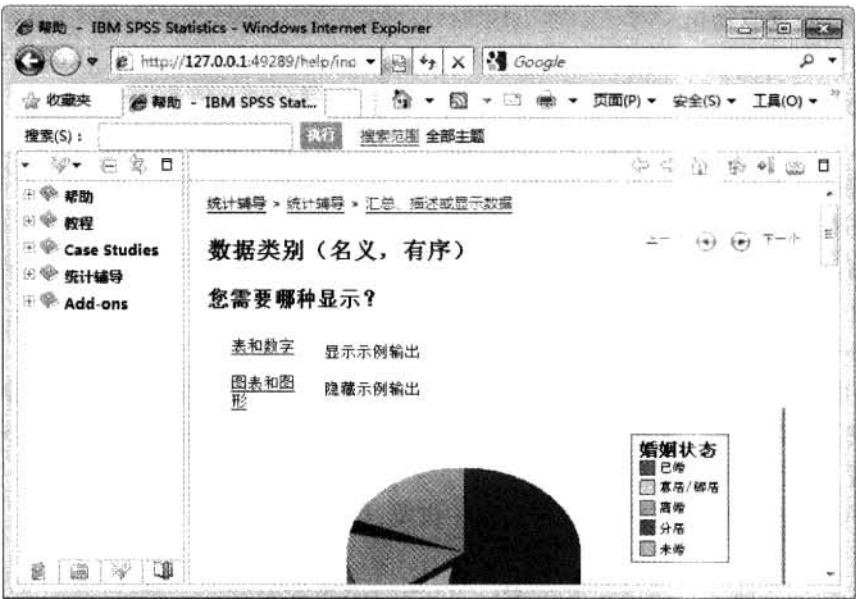


图 1.9 统计辅导的界面示意

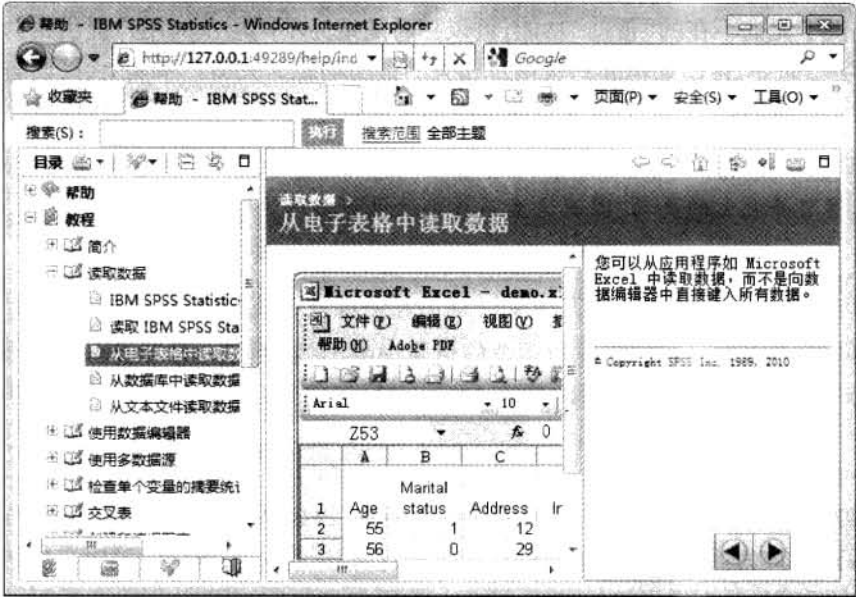


图 1.10 教程向导演示如何读取 XLS 数据

3. 个案研究

上述两个帮助功能或多或少都涉及一些入门和救急方面的内容,对于希望系统学习 SPSS 中统计功能的用户而言,就可以使用个案研究(Case Study)这一详细的案例向导,如图 1.11 所示。选择“帮助”→“个案研究”菜单项即可进入,它提供了 SPSS 各模块的主要分析方法的基本操作和结果解释。其讲解方式也是示例化、图形化的。实际上这一模块的讲解要优于市面上出版的绝大多数 SPSS 教材,但遗憾的是这一模块只有英文版本,目前尚未汉化。

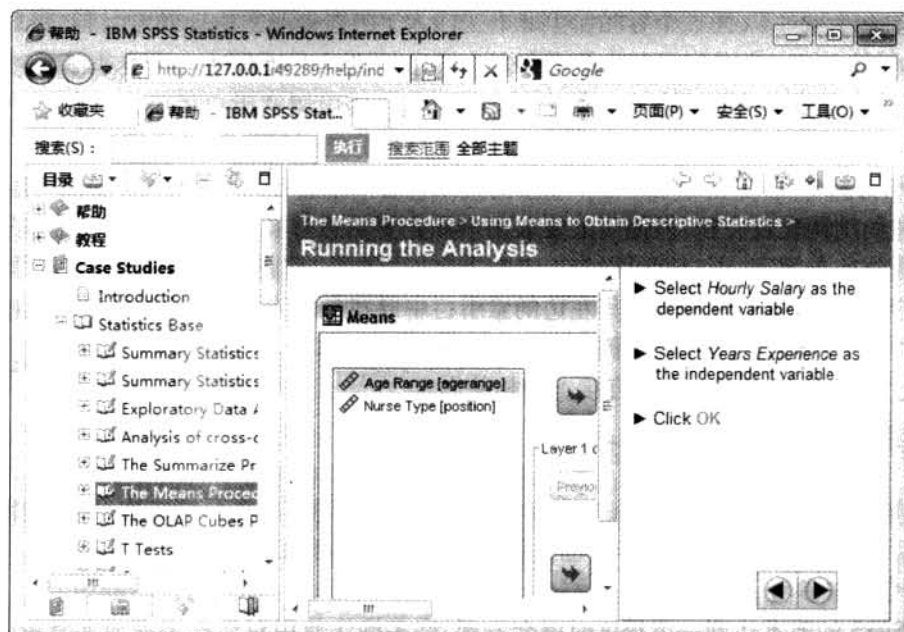


图 1.11 个案研究向导演示 Means 过程

1.5.2 帮助菜单

SPSS 的帮助文件从 19 版起已经被设计为网页格式,在菜单栏中选择“帮助”→“主题”菜单项,就会打开帮助网页,如图 1.12 所示,该网页在使用上没有太多特殊的地方,主要也是通过目录树和索引两种方式查找所需的内容的。

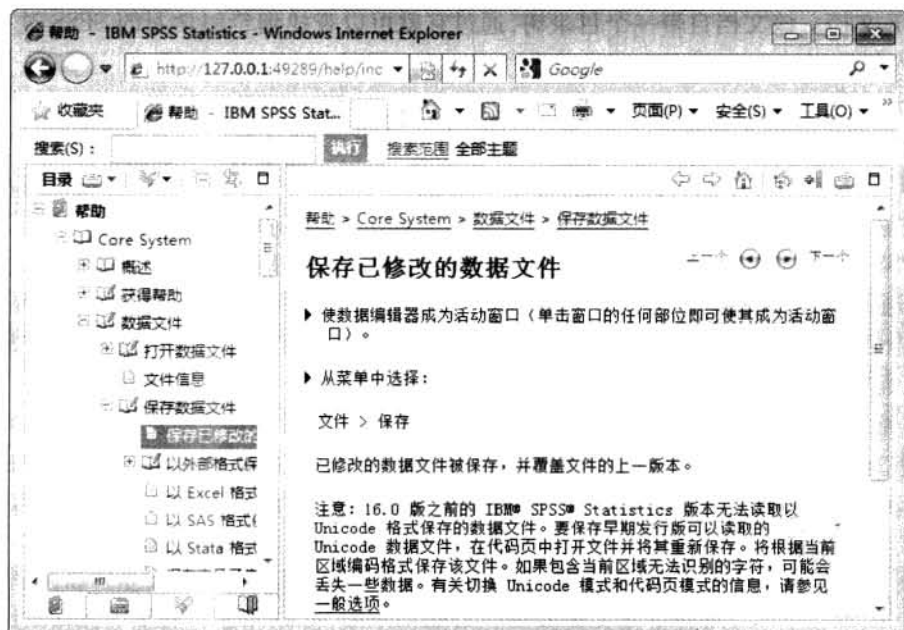


图 1.12 SPSS 帮助主题

1. 目录树方式

目录树像一本电子书的目录一样,将所有主题组织成一个树状结构。只要循着该目录的各级分支,最终总能找到所需的内容。用户可以在“目录”表中浏览用户手册,从而学习 SPSS 的使用方法。从左边选择一个主题,右边内容区即显示此部分的详细内容。

2. 索引方式

目录树的结构比较完整,但使用上要求用户首先要熟悉分类,而且要一层层地找下去,如果知道希望查找的关键字,用户就可以先单击左下方的“索引”按钮将左侧切换为“索引”表,然后在搜索框中输入关键词,系统会在其左边的索引栏中寻找与输入的关键词完全匹配的内容,双击并选择其中感兴趣的一个分项,右侧即可出现相关的详细解释。当然,如果关键词是不确定的,也可以通过上方的“搜索”文本框直接查询相关内容。

1.5.3 针对高级用户的帮助功能

对于高级用户而言,编程帮助、扩展包和插件的相关信息,甚至系统二次开发的相关信息就变得必不可少,SPSS 针对此类需求也提供了非常全面的帮助功能。

1. 指令语法参考

当对 SPSS 的熟悉达到一定程度时就会发现,许多操作使用对话框来完成非常麻烦,甚至于无法用对话框来实现。实际上,至少有 10% 的高级分析功能是必须要使用程序方式才能实现的,而且使用编程方式来完成相同的工作时,操作效率也要高得多。

由于目前国内几乎没有对 SPSS 编程加以深入讲解的资料,此时可以直接参考 SPSS 附带的语法指南。在 SPSS 的安装文件中都附送了所有模块语法指南书的 PDF 格式文档,这是 SPSS 官方提供的最为权威的使用指导,学会如何使用它,比将市面上所有的 SPSS 书都买两册起的作用还要大。语法指南的调用非常简单,只要选择“帮助”→“指令语法参考”菜单项,就会自动打开相应的 PDF 文档。该文档自带一个目录树,通过它就可以查找到希望学习的 SPSS 过程名称,从而进行深入学习。

2. 算法

在有的科研或商业问题中,用户可能会希望明确分析模型的某些算法细节究竟是如何实现的,为此 SPSS 公开了几乎所有方法的数学算法,选择“帮助”→“算法”菜单项,就可以打开相应内容,注意此内容仍然是英文的,但对于希望阅读算法文档的高级用户而言,这并不是问题。

3. SPSS 社区

SPSS 通过与 Python、R 等编程语言的对接,大大提升了其可扩展性。但是 R 等毕竟是完全不同的编程语言,有必要为用户提供相关的帮助文档,此外,进行 SPSS 的深度应用也需要许多复杂的知识,而这些资源都可以在 SPSS 社区获得。

选择“帮助”→“SPSS Community”菜单项,打开的是一个非常偏重于开发人员需求的社区网页,其中提供了对 SPSS 进行二次开发所需的文档和资源,包括扩展命令的说明与教程、Python 插件的教程、R 资源、相关的 DLL 库等资源。对 SPSS 二次开发有兴趣的用户尽可在此处进行深入钻研。

1.6 数据分析方法论概述

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。而随着社会的进步,科

学技术的发展,统计学的应用已经渗透到了人们工作和生活的各个方面,不同的领域所需要的方
方法论体系也会有所差异。根据作者的理解,这些方法论体系大致可被分为如下3种情形。

- (1) 严格设计支持下的统计方法论。
- (2) 半试验研究支持下的统计方法论。
- (3) 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,在功能上可以非常好地支持上述3
种方法论体系,并满足绝大多数情况下的统计分析需求,但读者需要自行判断在各自所从事的领
域中究竟哪一种方法论更为合适,并有针对性地加以学习和钻研。

1.6.1 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可称为经典统计方法论,作者觉得它之所以经典,不仅是因为
发展较早,还因为研究者在整个研究体系中可以掌控一切,具体特征如下:

(1) 这些研究都具有非常严密的研究设计,且往往严格遵循所谓的7大步骤: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。在这7大步骤中以试验设计步骤
最为关键,直接影响整个研究的成败。

(2) 在此类研究项目中,试验设计过程中会充分考虑需要控制的影响因素,并采用各种精巧
的设计方案来对非研究因素的作用加以控制,如配伍、完全随机抽样、随机分组等。

(3) 整个试验过程会在尽量理想的情况下进行,从而在试验/数据获取过程中也会对无关因
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、进食等作
出非常严格的设定。

(4) 原始数据往往需要从头加以采集,数据质量完全取决于试验过程是否严格依从设计的
要求,以及试验设计是否合理。当然,这也意味着每个原始数据的成本都非常高昂。

(5) 在分析方法上,最终所采用的统计模型应当是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的考虑和控制,因此在很
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。

此类统计方法论的应用在实验室研究、临床试验等领域中最为常见,而所使用的分析方法以
常用的单因素分析方法,或者针对一些复杂设计的一般线性模型(方差分析模型)最为常见。

1.6.2 半试验研究支持下的统计方法论

上述这种经典的统计分析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得
到满足的,因此往往退而求其次,形成了所谓的半试验研究支持下的统计分析,具体特征如下:

(1) 研究设计具有明显的向实际情况妥协的特征,因此所谓的7大步骤可能不被严格遵循,
例如,在数据本来就存在的情况下,数据收集过程就可能被省去。总体而言,在这7大步骤中从
数据准备开始的后3步的重要性要比经典统计分析高。

(2) 研究设计可能无法做到理想化,例如,抽样/分组的完全随机性,试验组/对照组干预措
施的严格控制都可能无法严格满足。最典型的例子,在药物研究中理想状况是设立安慰剂对照
组,但是如果是治疗恶性肿瘤的药物,则不可能让肿瘤病人去吃安慰剂。

(3) 整个数据采集过程难以做到理想化,举一个简单的例子,街头拦截(Central Location

Test) 是市场研究中非常常用的样本采集方式,但如果细究起来,拦截地点、拦截时间、拦截的星期,甚至于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,但这些最终只能凭借访问者的责任心和运气来尽量加以保证,从设计本身是无法进行控制的。

(4) 部分数据可能先于研究设计而存在,在整个研究中需要在这些数据的基础上去补充所需的其他部分信息。而另一方面,数据有可能不完全满足分析需求,但这种缺陷却无法得到补正。例如,利用全国各省的经济和人口数据进行各省的综合发展程度排序,可以考虑使用因子分析来做,但因因子分析原则上要求至少有 50 个案例,中国有 34 个省市自治区(包括香港、澳门、台湾),不可能为了进行这个统计分析再请有关部门划分出十几个新的省出来。

(5) 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也并不明确,需要在各种可能的影响因素中进行筛选和探索。而相应的可能用到的统计方法也颇为繁杂,从简单的统计描述,到复杂的广义线性模型都可能用到,而影响因素的筛选则成为很多分析项目的重点任务之一。但无论如何,使用的方法仍然以经典统计分析方法为主。

此类统计方法论的应用范围目前应当是最广的,在社会学、经济学研究中特别常见。

1.6.3 偏智能化、自动化分析的数据挖掘应用方法论

此类分析方法是随着近年来计算机技术的飞速发展而诞生的,一方面数据库技术使得许多行业出现了业务系统,有了自动积累的海量业务数据库,相应的也诞生了大批新的分析需求,但其数据量却使得传统方法论很难有效满足这些需求。另一方面,人工智能和计算能力的发展也诞生了一批全新的分析方法,如 Bootstrap、Bayes 方法与 MCMC、神经网络、树模型与随机森林等,赋予了分析人员全新的能力。在这些因素的相互作用之下,一种新的分析方法论:数据挖掘方法论就应运而生了。

数据挖掘是近年来由计算机人工智能、统计学和数据仓库技术交叉发展而产生的一种新方法体系,它通过采用各种自动或半自动地分析技术,在海量数据中发现有意义的行为和规则,迅速找到大量资料间的关联与趋势。其最大的特点是自动化、智能化,即充分利用计算机人工智能技术,自动/半自动地分析数据间的复杂联系,探寻一种独特的,通过其他方法可能难以发现的模式,快速发现有价值的信息。整个分析框架是动态、可更新的,并且在分析结果的验证上提供了许多新的思路。

和上面两种较为传统的分析方法论相比,数据挖掘方法论的特点如下:

(1) 完全以商业应用的需求为导向,或许可以认为传统方法论和数据挖掘方法论的最大区别在于:前者需要方法体系/逻辑正确,后者由于所处理的问题的数据量大、时间要求高,只需要结果正确,分析方法的理论正确性并不重要,而算法细节也可以是灰箱甚至黑箱。



数据挖掘所需要解决的问题往往具有很强的时间要求,例如消费者在网上购物时,页面上会出现“购买此商品的顾客也同时购买”之类的推荐栏目。其中的商品就是利用快速的数据挖掘算法筛选出来的。虽然这类分析的准确率能高一些,但相比之下,网站更愿意选择 2s 就能反馈给浏览者的弱关联算法,而不是采用 10 min 才能计算出更准确结果的强关联算法。

(2) 分析的流程出现了很大的变化,不再是线性的 7 大步骤,而转换成了周而复始的循环结

构,且非常强调前期的商业理解,以及后期的模型发布/应用。在几种常见的数据挖掘方法论中,以 CRISP-DM 最具代表性。

(3) 由于数据往往来源于业务系统,如超市的 POS 机、银行的 ATM 机、电信公司的业务数据库,因此数据采集过程是全自动进行的,完全先于整个研究项目存在。但这也意味着这些数据根本就不是为数据分析准备的,从而难以做到理想化。例如,对 POS 机数据进行分析,如果知道购物者的年龄、性别、家庭收入状况等,将会得到更有价值的分析结果,但即使关联了会员卡数据,这些背景资料也几乎是不可能补全的。

(4) 由于业务系统的数据是动态增加的,因此几乎不可能考虑另行人工收集希望补足的数据,只能结合实际能力收集,否则整个项目将永无止境。

(5) 在分析方法上,由于极端强调商业应用,因此分析方法的选择其实并不重要,往往采取多种方法并行、从中择优的分析思路。例如,对于一个客户流失预测项目,完全可以同时采用判别分析、Logistic 回归、神经网络、支持向量机(SVM)、Bayes 分析、树模型等多种方法平行分析,然后采用投票或者优选的方式得到最终的预测模型/结果。

在完全以满足商业需求为目标的背景之下,很多被认为是非常经典和基础的统计方法,比如参数估计和常规的假设检验,在数据挖掘中反而很难用到。另一方面,由于上述海量数据库、动态增量、平行分析等特点的存在,意味着数据挖掘中非常强调自动化,即使在项目期间会有很多人脑的智力投入,但最终项目结束时提交的一定是自动化的业务流,即以硅脑代替人脑。

能否满足商业需求,或者说模型是否能够在业务系统中得到真正的发布/应用,是判断整个数据挖掘项目是否成功的唯一标准,这一点和传统方法论有非常鲜明的区别。

思考与练习

1. 检查 SPSS 软件共有几个模块,其中包括了哪些功能,并思考平时进行统计分析时究竟需要哪些模块才能够满足需求。

2. 浏览 SPSS 产品支持网址或 SPSS 社区网址,熟悉其提供的各项内容,从中寻找符合自己工作需求的附加安装包和文档。

第2章 数据录入与数据获取

数据是统计研究的基础,没有数据,分析也就无从谈起。在 SPSS 中建立数据文件大致有两种情况:一种是非电子化的原始数据资料,需要直接将调查问卷中的数据录入进 SPSS 软件,建立数据文件;另一种是已经被录入为其他数据格式的资料,需要将其内容直接读入 SPSS 中。

针对上述两种情况,本章将主要说明两个问题,即如何将数据录入到 SPSS 中,以及如何将其其他格式的数据读进 SPSS 中。对于第一个问题,根据问题类型的不同,将会介绍开放题、单选题和多选题的录入方式;对于第二个问题,则重点介绍如何用 SPSS 直接读取 Excel 类型和文本格式的数据,以及如何通过 ODBC 接口读取数据库文件。

2.1 CCSS 案例项目背景

为了使本书内容更贴近实战,全书将尽量使用中国消费者信心调研项目的数据作为教学案例,通过该项目数据的实际运用对 SPSS 的各项功能进行讲解。本节将首先介绍该项目的背景,以方便读者的后续阅读。

2.1.1 项目背景

消费者信心是指消费者根据国家或地区的经济发展形势,对就业、收入、物价、利率等问题进行综合判断后得出的一种看法和预期,消费者信心指数则是对消费者整体所表现出来的信心程度及其变动的一种测度。消费者信心指数的概念和方法最早是由美国密歇根大学调查研究中心的乔治·卡通纳在 20 世纪 40 年代后期提出的,随后在美联储的委托之下开展了相应调研直至今日。60 余年的历史已经证明了这一指标体系在预测未来宏观经济走向方面具有不可替代的价值,目前已成为各市场经济国家非常重要的经济风向标之一。

联恒市场研究看到了这一指标体系潜在的市场价值,于 2007 年启动了中国消费者信心调研(CCSS)项目,这一项目是联恒与美国密歇根大学社会研究所消费者信心调查课题组负责人 Richard Curtin 博士共同设计开发完成的,整个方法体系与密歇根大学的消费者信心调查基本相同,同时也根据中国的具体国情进行了补充和完善,使之更贴近中国的实际情况。

CCSS 的调查始于 2007 年 4 月,每月在东部与中西部 30 个具有代表性的中国城市中抽取 1 000 个左右的家庭,通过计算机辅助电话访问(CATI)取得,目前已累计了 3 年多近 4 万个样本的历史数据。为化繁为简,这里将只截取北京、上海、广州 3 个城市在 2007 年 4 月、2007 年 12 月、2008 年 12 月和 2009 年 12 月的 1 147 个样本用于随后的讲解,具体数据参见文件 CCSS_Sample.sav。



CCSS 产品目前为英德知联恒市场咨询(上海)有限公司所有,本书所涉及的只是完整历史数据库的一小部分,且出于产品保密需要,在数据文件中删除了对指数计算至关重要的权重值,因此分析结果仅用于案例教学,所计算出的指数值会和真实指数值有一定偏差,不代表真实情况。

2.1.2 项目问卷

CCSS 项目的问卷是标准化的,每月固定执行。由于问卷内容较长,这里选择了其中部分题目作为教学案例,具体如下(注意:为了便于讲解,下列题目顺序和内容均进行过调整,并非访问时的原始状况)。

中国消费者信心指数研究问卷

S0 受访者所在城市:

100 北京 200 上海 300 广州

S1 请问您贵姓是? ____

S2 记录被访者性别:

1 男性 2 女性

S3 请问您的年龄是? ____

S4 请问您的学历是?

1 初中/技校或以下 2 高中/中专 3 大专 4 本科 5 硕士或以上

S5 请问您的职业是?

1 企/事业管理人员 2 工人/体力工作者(蓝领) 3 公司普通职员(白领)

4 国家公务员 5 个体经营者/私营业主 6 教师

7 学生 8 专业人士(医生、律师等) 9 无/待/失业、家庭主妇

10 退休 11 其他职业

S7 请问您的婚姻状况是?

1 已婚 2 未婚 3 离异/分居/丧偶

S9 请问您的家庭月收入(包括工资、奖金和各种外快收入)大约在什么范围呢?

1 999 元或以下 2 1 000 ~1 499 元 3 1 500 ~1 999 元

4 2 000 ~2 999 元 5 3 000 ~3 999 元 6 4 000 ~4 999 元

7 5 000 ~5 999 元 8 6 000 ~7 999 元 9 8 000 ~9 999 元

10 10 000 ~14 999 元 11 15 000 ~19 999 元 12 20 000 ~29 999 元

13 30 000 以上 98 无收入 99 拒答

C0 请问您的家庭目前有下列还贷支出吗?

C0_1 房贷 1 有 2 无 99 拒答

C0_2 车贷 1 有 2 无 99 拒答

C0_3 其他一般消费还贷 1 有 2 无 99 拒答

O1 请问您家里有家用轿车吗?

1 有 2 没有

A3 首先,请问与一年前相比,您的家庭现在的经济状况怎么样呢?是变好、基本不变还是变差?

1 明显好转 2 略有好转 3 基本不变 4 略有变差 5 明显变差 9 说不清/拒答

A3a 为什么您这样说呢?(最有限选两项)____

0 中性原因 90 不知道/拒答

10 改善:收入相关 110 恶化:收入相关

20 改善:就业状况相关 120 恶化:就业状况相关

30 改善:投资相关 130 恶化:投资相关

40 改善:家庭开支相关 140 恶化:家庭开支相关

50 改善:政策/宏观经济 150 恶化:政策/宏观经济相关

A4 那么与现在相比,您觉得一年以后您的家庭经济状况将会发生什么变化?

1 明显好转 2 略有好转 3 基本不变 4 略有变差 5 明显变差 9 说不清/拒答

A8 那么与现在相比,您认为一年以后本地区的经济发展状况将会如何?

1 非常好 2 比较好 3 保持现状 4 比较差 5 非常差 9 说不清/拒答

A9 您认为一年之后本地区的就业状况将会如何变化?

1 明显改善 2 略有改善 3 保持现状 4 略有变差 5 明显变差 9 说不清/拒答

A10 那么与现在相比,您认为5年之后,本地区的经济将会出现怎样的变化?

1 明显繁荣 2 略有改善 3 保持现状 4 略有衰退 5 明显衰退 9 说不清/拒答

A16 对于大宗耐用消费品的购买,如家用电器、家用计算机以及高档家具之类的,您认为当前是购买的好时机吗?

1 很好的时机 2 较好时机 3 很难说,看具体情况而定 4 较差时机 5 很差的时机 9 不知道/拒答

2.2 数据格式概述

2.2.1 统计软件中数据的录入格式

统计软件中数据的录入格式和大家平时记录数据用的格式不太相同,SPSS 所使用的数据格式也需要遵守相应的格式要求,其基本原则如下。

(1) 不同个案(Case)的数据不能在同一条记录中出现,即同一个案的数据应当独占一行。

(2) 每一个测量指标/影响因素只能占据一列的位置,即同一个指标的测量数值都应当录入到同一个变量中去。



但有时分析方法会对数据格式有特别的要求,此时可能会违反“一个个案占一行,一个变量占一列”的原则,这种情况在配对数据中和重复测量数据中最多见。这是因为根据分析模型的要求,需要将同一个观察对象某个观察指标的不同次测量看成是不同的指标,因此被录入成了不同的变量,这是允许的。但对于统计的初学者而言,最好能够严格遵守以上规则,而且无论表现格式怎样,最终的数据集都应当能够包含原始数据的所有信息。

2.2.2 变量属性

数据录入就是要把每个被访者的每个指标值录入到软件中。在录入数据时,大致可归纳为“数据录入三部曲”:定义各变量名,即给每个指标起个名字;指定每个变量的各种属性,即对每个指标的一些统计特性做出指定;录入数据,即把每个被访者的各指标取值录入为电子格式。因此这里首先介绍一下变量的各种属性问题。

任何一个变量显然都应当有变量名与之对应,但为了进一步满足统计分析的需要,除变量名外,在统计软件中还往往对每一个变量进一步定义许多附加的变量属性,如变量类型(Type)、变量宽度(Width)、小数位(Decimals)等。在第1章所讲解的数据管理窗口的变量视图中,可以看到SPSS会为每一个变量指定11种变量属性,但这里将重点介绍变量类型和测量尺度这两个属性,对于其他的一些属性,比如变量标签和缺失值等,会给出简单介绍,至于像变量列格式、变量对齐方式这样的属性,根据字面意思,就能理解其内涵。

1. 变量的存储类型

SPSS中的变量有3种基本类型,分别是数值型、字符串和日期型。根据不同的显示方式,数值型又被细分为5种(在20版中则分为6种),所以SPSS中的变量类型共有8种(在20版中则为9种)。在变量视图选择“类型”单元格时,右侧会出现形如的省略号按钮,单击会打开“变量类型”对话框,如图2.1所示。左侧为具体的存储类型,右侧则用于进一步定义变量宽度、小数位数等。

(1) 数值型(Numeric):在以上3大类变量类型中,数值型是SPSS最常用的变量类型。数值型的数据是由0~9的阿拉伯数字和其他特殊符号,如美元符号、逗号或圆点组成的。如工资、年龄、成绩等变量都可定义为数值型数据。数值型数据根据内容和显示方式的不同,又可分为标准数值型(Numeric)、每3位用逗号分隔的逗号数值型(Comma)、每3位用圆点分隔的圆点数值型(Dot)、科学计数型(Scientific Notation)、显示时带美元符号的美元数值型(Dollar)、用户自定义型(Custom Currency)等6种不同的表示方法。实际上上述表示方法中只有标准数值型最为常用,关于其余几种表示方法的详情读者如果有兴趣可以直接查阅软件帮助信息,这里不再赘述。



图 2.1 “变量类型”对话框

关于其余几种表示方法的详情读者如果有兴趣可以直接查阅软件帮助信息,这里不再赘述。

(2) 字符型(String):字符型也是 SPSS 较常用的数据类型,字符型数据的默认显示宽度为 8 个字符位,它区分大小写字母,并且不能进行数学运算。字符型数据在 SPSS 的数据处理过程(如在计算生成新变量时)中是用一对引号引起来的。需要注意的是,在输入数据时不应输入引号,否则,双引号将会作为字符型数据的一部分。

(3) 日期型(Date):这种类型的数据是用来表示日期或时间的。日期型数据的显示格式有很多,SPSS 在对话框右侧会以列表框的方式列出各种显示格式以供用户选择。如果此处选择的是 mm/dd/yy 或类似的两位数年份记录方式,则需要系统选项的“数据”选项卡中确定具体的世纪范围,目前系统默认为 1941—2040 年区间。

事实上,SPSS 中的日期型变量存储的是该时间与 1582 年 10 月 14 日零点相差的秒数,如 1582 年 10 月 15 日存储的就是 $60 \times 60 \times 24 = 86\,400$,将变量类型变换为数值型就可以看到。但是这里只能存储正数,即 1582 年 10 月 14 日及更早的时间在 SPSS 中是无效的。日期型数据主要在时间序列分析中比较有用,在较为简单的分析问题中完全可以用普通数值型数据来代替。

2. 变量的测量尺度

如果只使用变量类型,很多时候并不能准确地说明变量的含义和属性。比如 CCSS 数据中的以下几个变量。

(1) 变量 S2“性别”:用 1 代表男,2 代表女。在这里 1 和 2 只是一个符号,没有任何数字意义。2 并不比 1 大,1 也并不比 2 小。

(2) 变量 S4“学历”:用 1 表示“初中”,2 表示“高中”,3 表示“本科”等,1 和 2 虽然也是符号,但这里有一个顺序之分,1 就比 2 的学历低。但是究竟低多少,是本科和高中的差距更大,还是高中和初中的差距更大,不知道,各级别之间的差距大小无法衡量,更无法进行比较。


(3) 变量 S3“年龄”:20 和 21 就是有区别的,差 1。而且这个差距大小,和 39 与 40 之间的差距是相等的,都是 1,也等于 50 和 55 之间差距的 1/5。

由上可知,上述 3 个变量的存储类型同样都是数值型,但数值的具体含义不同,所携带的信息量不同,适用的统计方法也就不同。如果只以存储类型来说明这个变量的属性,就不能反映上述区别。为此,就有必要给变量增加测量尺度这一属性。

在统计学中,按照对事物描述的精确程度,将所采用的测量尺度从低到高分 4 个层次:定类尺度、定序尺度、定距尺度和定比尺度。在这 4 种测量尺度之间,按照信息量的高低,可将高层次测量尺度的测量结果转换为低层次测量尺度的测量结果,但这样会损失一部分信息,但不能将低层次的测量尺度转换为高层次测量尺度的结果,这样可能会引入错误的信息。

1) 定类尺度


定类尺度(Nominal Measurement)是对事物的类别或属性的一种测度,按照事物的某种属性对其进行分类或分组。定类变量的特点是其值仅代表了事物的类别和属性,仅能测定类别差,不能比较各类之间的大小,所以各类之间没有顺序或等级,如变量 S0“城市”就是一个定类尺度的变量。对于定类尺度的变量只能计算频数和频率,如在所有个案中,北京有多少人,占总人数的百分率是多少等。对于 S2“性别”这种两分类变量,一般仍然将其归为定类尺度变量。但是两分类变量较为特殊,即使将其归为其他类型,一般也不会影响后续分析。

在 SPSS 中使用度量标准(Measure)属性对变量的测量尺度进行定义,其中定类尺度变量用“名义(N)”来表示。能使用的定类尺度的数据可以是数值型变量,也可以是字符型变量。使

用定类变量对事物进行分类时,必须符合穷尽和互斥的原则。穷尽的原则就是指“每个个体都必须能归为一个类别”,互斥的原则是指“每个个体都只能归为一个类别”。

2) 定序尺度

定序尺度(Ordinal Measurement)是对事物之间等级或顺序差别的一种测度,可以比较优劣或排序。定序变量比定类变量的信息量多一些,不仅含有类别的信息,还包含了次序的信息;但是由于定序变量只测度类别之间的顺序,无法测出类别之间的准确差值,即测量数值不代表绝对的数量大小,所以其计量结果只能排序,不能进行算术运算。CCSS 数据中的变量 S4“学历”就是一个典型的定序变量。

在 SPSS 的度量标准属性对话框中,定序尺度变量用“序号(0)”来表示。定序变量同定类变量一样,其数据可以是数值型变量,也可以是字符型变量。对于定序变量除了可以计算频率之外,还可以计算累计频率。如足球喜欢程度这一变量的取值有:1—非常喜欢,2—喜欢,3—无所谓,4—不喜欢,5—非常不喜欢,这是一个定序尺度的变量,因而可以计算累计频数和累计频率。如对于“足球喜欢程度”,不仅可以计算喜欢的人数和比例,还可以计算喜欢及非常喜欢的累计人数和比例。


3) 定距尺度

定距尺度(Interval Measurement)是对事物类别或次序之间间距的测度。定距变量的特点是其不仅能将事物区分为不同类型并进行排序,而且可准确指出类别之间的差距是多少;定距变量通常以自然或物理单位为计量尺度,因此测量结果往往表现为数值,所以计量结果可以进行加减运算,生活中最典型的定距尺度变量就是温度。

4) 定比尺度

定比尺度(Scale Measurement)是能够测算两个测度值之间比值的一种计量尺度,它的测量结果同定距变量一样也表现为数值,如职工月收入、企业销售额等。其与定距变量的差别在于有一固定的绝对“零点”,而定距变量则没有,定距变量中的“0”并不表示“没有”,仅仅是一个测量值,而定比变量中的“0”则真正表示“没有”,比如温度,0℃只是一个普通的温度(水的冰点),并非没有温度,因此它只是定距变量,而重量则是真正的定比变量,0 kg 就意味着没有重量可言。上面提到的变量 S2“年龄”就是一个典型的定比变量。

定比变量是测量尺度的最高水平,它除了具有其他 3 种测量尺度的全部特点外,还具有可计算两个测度值之间比值的特点,因此可进行加、减、乘、除运算,而定距变量严格来说只可进行加减运算。

SPSS 中默认的变量测量尺度就是定比尺度。但由于后两种测量尺度在绝大多数统计分析中没有本质上的差别,在 SPSS 中就将其合并为一类,统称为“度量(S)”。



这 3 种尺度在许多统计书籍中会有更为通俗的名称:无序分类变量、有序分类变量和连续性变量。从实用的角度出发,本书将同时采用这两种命名体系。

3. 变量名与变量值标签

除了上边介绍的变量类型和测量尺度外,变量的其他属性也很重要,比如说,标签(Label)属性用于定义变量名标签,对变量名的含义进行进一步解释说明,该标签会在结果中输出以方便阅读,增强变量名的可视性和统计分析结果的可读性。另外,值(Values)属性也是一个不得不提的

选项,用于定义变量值标签(如图2.2所示),变量值标签是对变量取值含义的解释说明信息。例如,对于性别数据,假设用1表示男,用2表示女,如果在录入数据时数据集中没有设定变量值标签,其他人就很难弄清楚是1表示男还是2表示男。因此,变量值标签对于定序变量(如职称)和定类变量(如民族、性别)来说是必不可少的,它不但使定类和定序变量的数据录入变得更加方便,并且明确了数据的含义,也同样增强了分析结果的可读性。

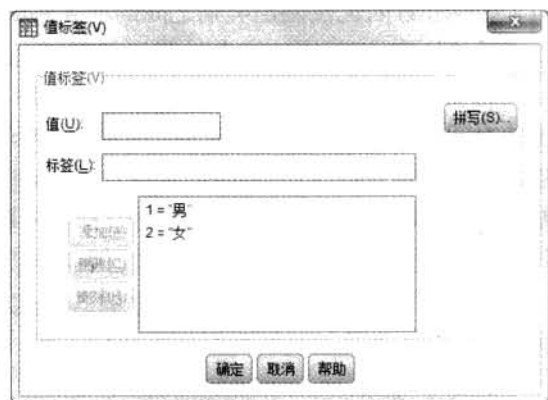



图 2.2 “值标签”对话框

变量“值标签”对话框上部的两个文本框分别为变量值输入框和变量值标签输入框,分别在其中输入“1”和“男”,此时下方的“添加”按钮变黑,单击它,该变量值标签就会被加入下方的“标签”列表框内。与此类似定义变量值“2”为“女”,最后单击“确定”按钮,变量值标签就设置完成。此时做任何分析,在结果中都有相应的标签出现。如果现在就想看显示效果,可切换回数据视图,然后选择“视图”→“值标签”菜单项。

另外,SPSS 在 12 版本以前,对变量名有一个限制,即要求变量名在 8 个字符之内。但令人欣喜的是,从 12 版本开始,此限制已经被取消,变量名最多可以有 64 个字符。当然,出于兼容性的考虑,变量名的定义还有一些限制,即不能以数字开头,中间不能有空格,一个数据文件中不能有相同的变量名等。读者只要在使用中尝试即可,不必记那么多规则。

4. 缺失值

缺失(Missing)属性是一个重要而且容易被忽视的变量属性,它用于定义变量缺失值。SPSS 中的缺失值有用户自定义缺失值和系统缺失值两大类。对于数值型变量的数据,系统缺失值用一个圆点“.”表示,而字符型变量默认就是空字符串。如果在问卷调查中有些数据项漏填了,则数据录入时只能跳过,相应的数据单元格就会被系统自动当成缺失值来处理。

另外一类缺失值是用户自定义缺失值,这往往出现在一些设计较严格的大型调查中,在一些题项处会给出一个选项:不知道/拒答。相应的代码可以用 9 或者 99 来表示,例如,CCSS 项目中的 S9“家庭月收入”中就是以 99 来表示拒答的。显然,这里的 99 不是一个真实的答案,仅仅是缺失值代码,需要告知 SPSS 这个特定的标记数据,以在进行统计分析时区别对待缺失值和正常的分析数据。具体做法为单击相应变量缺失值属性对话框右侧的按钮,会打开“缺失值”对话框,如图 2.3 所示,利用该对话框,用户可以自定义缺失值。界面上有 3 个单选按钮,默认值为最上方的“没有缺失值”;第二项可指定离散的缺失值(Discrete Missing Values),最多可以定义 3 个

值;最后一项指定缺失值所在的区间范围,并可同时指定一个离散值。

5. 角色

该属性是较新的 SPSS 版本中新增的,实际上来源于数据挖掘方法体系的要求,某些对话框支持可用于预先选择分析变量的预定义角色。当打开其中一个对话框时,满足角色要求的变量将自动显示在目标列表中。可用角色包括以下几个。

- (1) 输入:变量将用做输入(例如,预测变量、自变量)。
- (2) 目标:变量将用做输出或目标(例如,因变量)。
- (3) 两者:变量将同时用做输入和输出。
- (4) 无:变量没有角色分配(将不纳入分析)。
- (5) 分区:变量用于将数据划分为单独的训练、检验和验证样本。



图 2.3 “缺失值”对话框

- (6) 拆分:该项的存在主要是为了能够和 Clementine(即现在的 IBM SPSS Modeler)相互兼容。具有此角色的变量不会在 SPSS 中自动成为拆分文件变量。

在默认情况下,SPSS 将为所有变量分配输入角色,需要指出的是,角色分配只影响支持角色分配的对话框。而此类对话框在现有版本的 SPSS 中很少,因此一般用户可以直接无视这一属性。

其他的变量属性,即使不讲解,大家也可以根据 SPSS 界面的提示做出正确的选择,所以这里就不再赘述了。但是有一点要强调的是,就数据录入这部分内容来说,变量属性的设置是最重要的一部分工作,属性的设置不仅涉及对错,而且还有一个设置好坏的问题,属性设置得好,会简化后边的数据分析工作,所以不能忽略这部分工作。

2.3 数据的直接录入

在 SPSS 中,新建一个数据文件非常容易。只要打开 SPSS,系统就已经生成了一个空数据文件,用户只要按自己的需要在其中定义变量、输入数据,然后保存即可。



对于这个空数据文件,还要注意窗口左上角的文字是“未标题 1[数据集 0]”,其含义是该数据暂时未被存储为数据文件,所以没有文件名称(未标题);但是 SPSS 系统内部在使用该数据文件时,将会按照“数据集 0”这个名称来标识该文件,这就是所谓的工作名称。

2.3.1 操作界面说明

数据窗口是一个典型的 Windows 软件界面,如图 2.4 所示,第一次使用 SPSS 会使人觉得很亲切,从中可以看到菜单栏、工具栏,在 SPSS 的工具栏下方是数据栏,数据栏下方则是数据编辑窗口的主界面。该界面由若干行和列组成,每行对应一条记录,每列对应一个变量。由于现在没有输入任何数据,所以行、列的标号都是灰色的。注意第一行第一列的单元格边框为深色,表明该数据单元格为当前单元格。

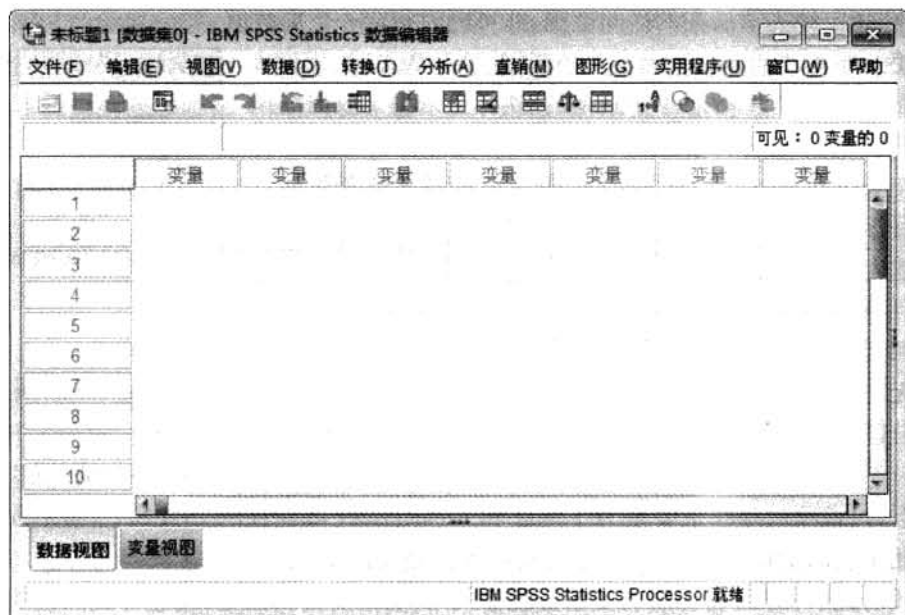


图 2.4 SPSS 的数据编辑窗口

在界面的左下角可以看到“数据视图”和“变量视图”标签,图 2.4 中显示的是数据视图,如果单击右边的“变量视图”按钮,则会切换到变量视图。前面提到的变量属性的设置都在变量视图图中进行,而数据的录入工作则应当在数据视图中直接通过键盘完成。



初学者往往会关心 SPSS 的数据容量问题,实际上,作为一个功能完善的统计软件,只要相应机器的内存和硬盘足够大,SPSS 理论上可以加载的变量数和案例数是无限大的,笔者亲自处理过的数据文件变量数最多达到上千个,案例数最多达到千万条的级别。而且达到此数据量时,SPSS 用现在的主流硬件配置完成常用的统计分析工作耗时也是非常少的。

2.3.2 开放题和简单单选题的录入

根据在调查问卷中设计问题的类型的不同,定义变量的方式也不同。通常调查问卷中的问题包括单选题、多选题和开放题等几种,以 CCSS 问卷为例,可以发现在这份问卷中,ID 是数值型开放题,S1“姓名”是字符型开放题,S2“性别”是单选题,c0 系列、a3a 系列均为开放题。下面将分别就这几种类型题目的录入方式加以介绍。

1. 在 SPSS 中定义变量

前边已经说过,录入数据的第一步是定义变量属性,随后才能进行数据录入。虽然在空白的变量列中直接输入数据,SPSS 会自动给该列设定一个变量名,但是这样往往不能完全满足用户的需要,所以首先要定义需要使用的变量。

定义变量属性,首先要定义变量名,如图 2.5 所示,变量名是变量的唯一标识,前边已经讨论过相关的知识,这里就不再重复,在前 3 行的“名称”属性列中直接输入变量名——ID、S1、S2,可以看到 SPSS 会在变量类型等列自动填入默认值。

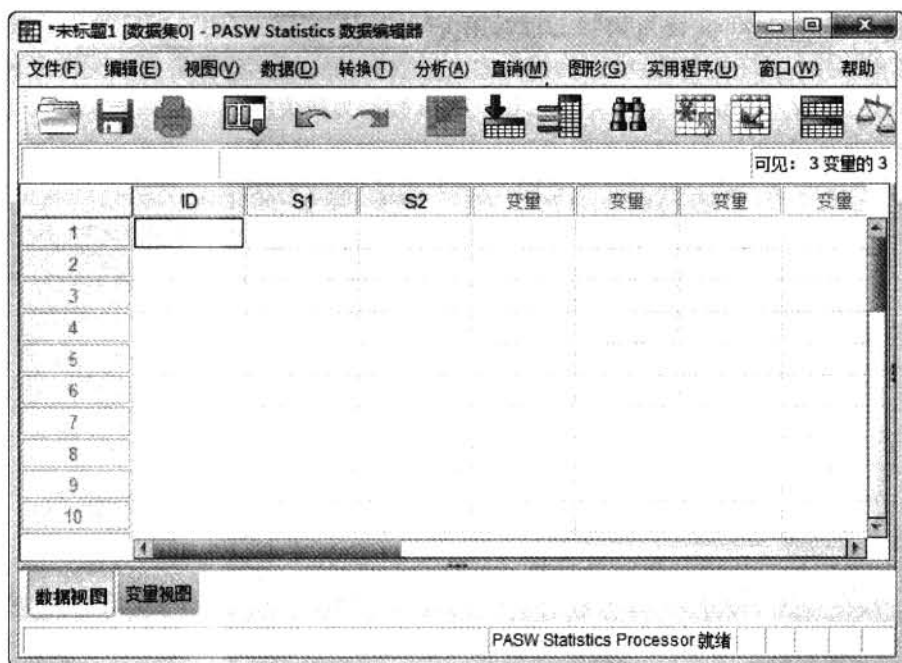


图 2.5 变量定义

在绝大多数情况下,SPSS 给出的默认数据类型和数据精度可以满足需要,如果默认值满足分析的需要,变量定义到此就可以结束了,否则就需要对不满足条件的选项进行进一步的设置。在本例中:

(1) 变量 ID 是被访者的记录号,它的测量尺度应该是定类尺度。但需要指出的是,因为变量 ID 只是方便检查和核对问卷,不参与后边的数据分析工作,所以,在要求不严格的情况下,此处的变量类型可采用默认形式不进行修改。

(2) 变量 S1 是被访者姓名,应是字符型变量,这里应当将“类型”中的“数值”改成“字符串”,如图 2.6 所示,并在必要时放大默认的 8 位宽度以满足需要,因为默认的 8 个字符的宽度只能存放 4 个汉字,要根据该变量可能出现的最大字符长度来确定宽度,只要最大不超过 256 个字符即可。

现在切换回数据视图,数据编辑窗口如图 2.7(a)所示。可见前 4 列的名称均为深色显示,就是刚才定义的内容,表明这 4 列已经被定义为变量,其余各列的名称仍为灰色的“变量”,表示尚未使用。同样地,各行的标号也为灰色,表明现在还未输入过数据,即该数据集内没有记录。在变量定义完毕后,就可以向这个文件中录入数据了。

2. 开放题的录入

现在开始录入数据,首先输入变量 ID 的值,确认 1 行 1 列单元格为当前单元格,放弃鼠标而用键盘,输入第一个数据 1,此时界面显示如图 2.7(b)所示。

注意:在按回车键之前,输入的数据在数据单元格内左对齐显示,表示该单元格为第一次录入数据,同时数据栏内同步显示出输入的数值。现在按回车键,界面如图 2.7(c)所示。和图 2.7(b)相比发生了以下变化。首先,当前单元格下移,变成了 2 行 1 列单元格,而 1 行 1 列单元格的内容则被替换成了 1.00,出现两位小数是因为数值型(num)变量默认为两位小数(由于序号只

可能是整数,可以将 Decimal 设为“0”);其次,第 1 行的标号变黑,表明该行已输入了数据;第三,在 1 行 2 列单元格(字符型变量)中因为没有输入过数据,显示为空,在 1 行 3 列单元格(数值型变量)中因为没有输入过数据,显示为“.”,这代表该数据为缺失值。

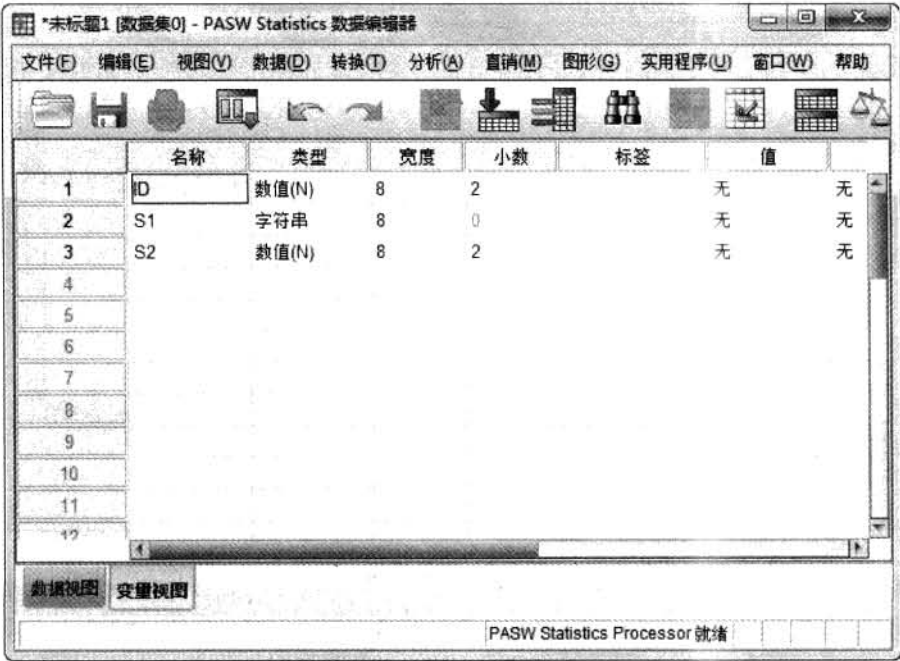


图 2.6 定义好变量的数据编辑窗口

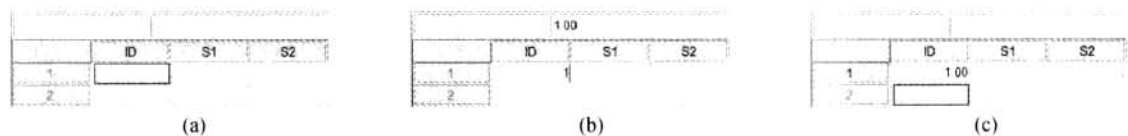


图 2.7 录入数据过程

如果要继续录入,则用类似的输入方式将数据录入完毕即可,但有一点不得不提醒大家,在数据录入过程中,要随时注意保存,如果突然断电或者死机,辛苦工作的成果就会丢失。

3. 单选题的录入

单选题的录入方式与开放题类似,不同的是,在单选题中可以定义变量值标签,通过这种方式既可以减少数据录入的工作量,而且还可以方便后边的数据分析工作。具体而言,单选题的录入可以采用字符直接录入、字符代码 + 值标签、数值代码 + 值标签 3 种方式。对应这 3 种录入方式,变量 gender 定义后的界面如图 2.8 所示。

	名称	类型	宽度	小数	标签	值
1	S2_1	字符串	8	0	性别	无
2	S2_2	字符串	8	0	性别	{1. 男}
3	S2_3	数值(N)	8	2	性别	{1 00. 男}

图 2.8 单选题的 3 种录入方式说明

对于这3种录入方式,原则上都是可以的,但是第3种录入方式“数值代码+值标签”能够方便后边的分析工作,推荐读者使用第3种录入方式。

4. 半开放题的录入

半开放题指的是问卷数据中有含“其他,请指出”选项的单选题,此类题目在录入时可以使用两个变量对其进行定义,在第1个变量中,“其他,请指出”作为选项中的一个可进行选择;第2个变量将“其他,请指出”的具体内容看做一个独立的开放题,按照开放题的录入方式进行数据录入,将没有选择该选项的被访者作为缺失值处理。

为使得变量名之间具有一定的逻辑联系,可以考虑将第2个变量的名称设置为由第1个变量名称后直接加“a”之类的字符,另外在数据录入完毕后,可能会在数据预处理阶段对第2个变量中的数据进行编码处理,以便进行后续分析,在SPSS中的相关功能可参见第3、4章的相应讲解,此处不再赘述。

2.3.3 多选题的录入

多选题,又被称为多重响应(Multiple Response),是在社会调查和市场调研中极为常见的一种数据记录类型。通常,对于问卷中的一个单选题一个被访者只能取一个值。而多选题,比如CCSS项目中的c0和a3a题目均为多选题,被访者可以选择一个选项,也可以选择两个或者多个。这样一来,由于在多选题中每道题都可能有一个以上的答案,多选题就不能用一个变量来直接编码(否则无法进行分析),而需使用几个变量来进行记录。在统计软件中用于多选题的常见方法有两种:多重二分法(Multiple Dichotomy Method)和多重分类法(Multiple Category Method)。下面将进行详细说明。

1. 多重二分法

所谓多重二分法,是指在编码时,对应每一个选项都要定义一个变量,有几个选项就有几个变量,这些变量分别代表对其中一个选项的选择结果,一般均为二分类,而其中必然有一个类别代表选中了这一选项。

在SPSS中对多选题进行数据录入与单选题的录入程序相同,均是首先在变量视图进行变量定义,然后直接录入数据,多选题的不同之处是变量的定义方式不同,而且,数据录入完毕,在分析之前,还需要定义多选题集。

首先来定义变量。每个选项对应一个变量,如CCSS项目中的c0题目,对应所需选择的3种选项,分别设定了c0_1、c0_2、c0_3这3个变量,且均以1表示选中,2表示未选中,如图2.9所示,可见第2个个案每月有房贷支出,但没有车贷和其他消费还贷支出。而第3个个案则每月只有其他消费还贷支出。

显然,在多重二分法中无论有多少个变量,其变量值标签的定义应该一致,否则将会引起混乱。还有一点要说明的是,在c0题目中还增设了代码99代表拒答,这主要是根据访问的实际需求增设的,在后续分析中可以将99和2合并成一类,即按未选中该选项来进行分析。

2. 多重分类法

多重二分法实际上是多选题的标准数据格式,但这种数

c0_1	c0_2	c0_3
2	2	2
1	2	2
2	2	1
1	1	2
2	2	2
2	2	2
2	2	2

图 2.9 多重二分法数据录入格式

据格式有时也会给数据录入带来麻烦,以 CCSS 项目中的 a3a 题目为例,每个受访者被限制只能回答最多两个选项,但总选项数量多达 12 个,显然,如果使用多重二分法录入,则大部分数据都需要录入为“未选中”,徒增许多数据录入的工作。对于此类多选题,则使用多重分类法进行记录更为便捷。

多重分类法也是利用多个变量来对一个多选题的答案进行定义,应该用多少个变量由受访者实际可能给出的最多答案数而定。而且,这些变量必须为数值型变量,利用值标签将答案标出,所有变量采用一套值标签。之所以称其为多重分类法,是因为每个变量都是多分类的,每个变量代表被访者的一次选择。

多重分类法适用于问题的选项较多的情况,尤其适用于“请在下列选项中选出您最喜欢的几个选项”一类的问题。以 a3a 为例,由于限定最多回答两个选项,因此只需要设定 a3a_1 和 a3a_2 两个变量即可,从图 2.10 中可见个案 1 选择了 120 和 140 两个选项,而个案 4 只回答了 130 这一个选项,随后的 a3a_2 则为缺失值。显然,这种“数据缺失”的现象在多重分类法中其实是一种正常情况。

a3a_1	a3a_2
120	140
0	140
0	.
130	.
30	.
0	0
90	.

图 2.10 多重分类法
数据录入格式

3. 设定多选题变量集

在进行多选题录入时,只需要将相应的变量设定好即可进行操作,但是录入完毕后 SPSS 只会默认它们是若干个分散的变量,并不明白它们代表的是一道多选题,只有将其设定为多选题变量集(也称为多重响应集),SPSS 才能对其进行正确识别,从而将多选题的全部变量当成一整道题目来进行分析。

在 SPSS 中提供了专门的菜单用来处理多选题,Tables 模块和多重响应(Multiple Response)菜单都可以用来设定多选题变量集。所不同的是,多重响应菜单中的定义变量集(Define Sets)项定义的多选题变量集信息不能在 SPSS 数据文件中保存,关闭数据文件后相应信息就会丢失,如果再次使用,则必须重新加以定义;而 Tables 模块可以保存所定义的信息。这两个过程的操作基本相同,现在就以 Define Sets 过程为例来介绍一下多选题集是如何定义的。

在 SPSS 中选择“分析”→“多重响应”→“定义变量集”菜单项,打开“定义多重响应集”对话框,如图 2.11 所示。关于该对话框有如下几个需要注意的地方。

(1) “集合中的变量”(Variables in Sets)列表框:选择需要加入同一个多选题变量集中的变量列表,对于采用多重二分法录入的多选题,这些变量必须为二分类,并按照相同的方式来编码(如都用 1 代表选中)。对于采用多重多分类法录入的多选题,这些变量必须为多分类,并共用一套值和值标签。

(2) “将变量编码为”(Variables Are Coded As)单选按钮组:选择变量的编码方式。在多重二分法方式中,需要在右侧的“计数值”文本框中指定用哪个数值表示选中。在多重分类法方式中此时则需要设定取值范围,在该范围内的记录值将纳入分析,注意在 Tables 模块中是不需要设定取值范围的。

(3) “名称”(Name)文本框:输入多选题变量集的名称,在此定义的变量集名称为 c0,在下方的“标签”文本框中可以为相应的多选题变量集定义一个名称标签。

所有设定均完成后单击右侧的“添加”按钮,相应的多选题变量集就会被加入最右侧的“多重响应集”列表框中。

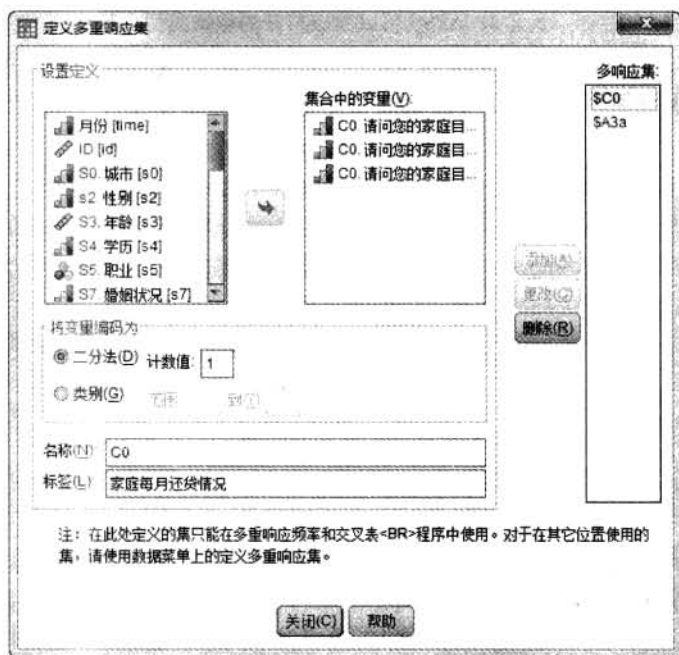


图 2.11 定义多选题变量集

4. 半开放多选题的处理方式

对于含有“其他,请指出”答案的附加内容的多选题,基本处理思路和半开放单选题非常相似,即首先将“其他”当成一个答案选项,而用另一个变量来表示其他的内容。在数据录入完毕后再对附加内容根据频次高低进行二次编码,以进行更为深入的分析。

2.4 外部数据的获取

对于 SPSS 格式的数据,只要选择“文件”→“打开”→“数据”菜单项,然后选择文件路径和文件名打开即可。如果数据不是 SPSS 格式的,也可以直接读入 SPSS,用 SPSS 进行分析。SPSS 可以读入许多非 SPSS 默认类型的数据文件,方式主要有 3 种,包括直接打开、利用文本向导读入文本数据,以及利用数据库 ODBC 接口读取数据。对于这 3 种方法,下面将以常见的 Excel 格式的数据、文本数据和 Access 数据为例,介绍 SPSS 获取数据的功能。

2.4.1 读取电子表格数据文件

1. 可支持的文件类型


在 SPSS 中可以直接读入许多常用格式的数据文件,选择“文件”→“打开”→“数据”菜单项,或直接单击快捷工具栏上的快捷按钮,系统就会弹出“打开数据”对话框,在“文件类型”列表框中可以看到直接打开的数据文件格式,SPSS 在这方面的兼容性做得非常出色,和所有常见的数据格式都有直接读取的接口,具体如表 2.1 所示。

表 2.1 SPSS 可以直接打开的数据类型

数据标识	数据类型
SPSS Statistics (*. sav)	SPSS 各版本的数据文件
SPSS/PC + (*. sys)	SPSS/PC + 版本的数据文件
Systat(*. syd, *. sys)	Systat 数据文件
便携(*. por)	SPSS 便携格式的数据文件
Excel(*. xls, *. xlsx, *. xlsxm)	Excel 各版本的数据文件
Lotus(*. w *)	Lotus 各版本的数据文件
SYLK(*. slk)	以 SYLK(符号链接) 格式保存的数据文件
dBASE(*. dbf)	dBASE 系列数据文件(从 dBASE II ~ IV)
SAS(*. sas7bdat, (*. sd7, ...)	SAS 各版本的数据文件
Stata(*. dat)	Stata 4 ~ 8 版的数据文件
文本格式(*. txt, *. dat)	纯文本格式的数据文件

和老版本的菜单相比,文件类型列表框做了大幅的合并,显得更为简洁。在其中选择所需的文件类型,然后选中需要打开的文件,SPSS 就会按照要求打开相应的数据文件,并自动转换为 SPSS 格式。

2. 操作实例

下面以 SPSS 自带的文件 demo. xls 为例说明 SPSS 如何直接读取这个文件,该文件位于 SPSS 安装目录下的 Samples 子目录中。首先在 Excel 中打开 demo. xls,了解一下这个文件的结构,重点需要了解这样几项内容:第一,该文件中包含几个数据表,具体应当打开哪个表;第二,如果不需要该表的所有数据,而只需读入一部分,这时需要了解要读入数据的精确位置,如单元格 A2:F5;第三,此部分数据的第 1 行是否是变量名。从这个文件中可以看出,第 1 行是变量名,该文件只有一个表,要读取的是该表中的全部数据。

第一步,在打开文件对话框中,选择路径(此例中为 Samples\English),选择文件类型 Excel (. xls),文件列表中出现所有的 Excel 文件,单击文件 demo. xls;第二步,打开如图 2. 12 所示的对话框:在“工作表”下拉列表框中选择一个表;在“范围”文本框中指定读取的数据的具体位置,用单元格的起(左上角单元格名称,如 A2)止(右下角单元格名称,如 F5)位置来表示,中间用冒号“:”隔开;上方的复选框用于确定单元格范围的第 1 行是否为变量名。指定完毕,单击“确定”按钮,数据就会被顺利地读入 SPSS 中。



图 2. 12 “打开 Excel 数据源”对话框

这种直接读取的方法要优于“复制 + 粘贴”，采用这种方法不仅可以顺利地进行变量名的转化，最重要的是可以直接读取字符型变量，若采用“复制 + 粘贴”的方法，字符型变量就全部变成缺失值了，并且操作简单，不容易出错，就和读取 SPSS 自己的文件一样方便。

在上面的实例中只需要读取一个表单的数据，如果需要将两个或者多个 Sheet 放在一个数据文件中，仍然可以像读取单个 Sheet 文件那样轻松方便。有两种方式可以实现这一要求，第一种方式是打开两个 SPSS 窗口，分别读取两个 Sheet，然后使用 merge 命令（详见第 4 章）对两个文件进行合并；第二种方式是使用前面所讲的方式，首先读取其中的一个 Sheet 并保存，然后直接从该文件读取另一个 Sheet，实现 SPSS 和 Excel 的合并。

2.4.2 读取文本数据文件

SPSS 可以通过两种菜单操作方式读取文本数据：一种是选择“文件”→“打开文本数据”菜单项；另一种是选择“文件”→“打开”→“数据”菜单项，这两种情况是一样的，系统会弹出打开数据对话框，只是前者的文件类型自动跳到了 Text(*.txt)，后者需要在“文件类型”下拉列表框中进行选择。之所以在其中保留“打开文本数据”选项有两个原因：① 读入纯文本的情况非常普遍，放在这里更加醒目；② 为了和 SPSS 老版本在菜单上保持兼容。

这里以系统自带的文件 demo.txt 为例来说明如何将文本数据导入 SPSS 中。与读取 Excel 数据一样，首先打开该数据，观察这个数据的基本结构，如变量间是固定宽度，还是用某种分隔符区分，第 1 行是否为变量名等。然后关掉这个文本文件，打开 SPSS 软件。首先，在“打开文件”对话框选择相应的文件并单击“确定”按钮，系统会自动打开“文本导入向导”对话框，如图 2.13 所示，从对话框标题可以看到该向导共分 6 步，下面一步步地来讲解。

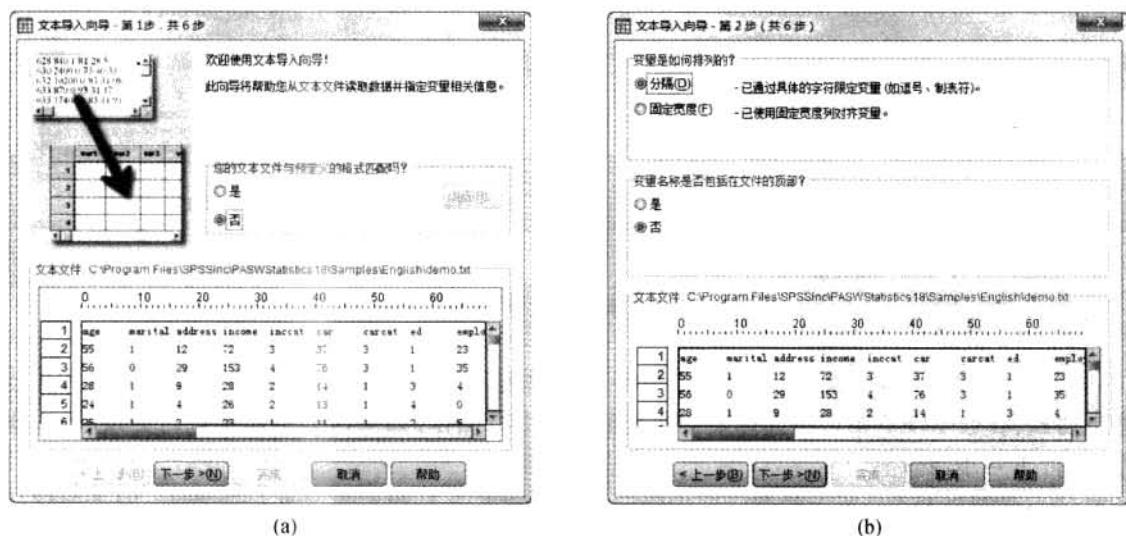


图 2.13 “文本导入向导”的第 1、2 步

(1) 系统首先会询问有无预定义格式,如图 2.13(a)所示,如果有则在此处选择相应文件,下方则为按预定义格式读入的数据文件的预览效果(后面的各个向导界面也会随时更新预览状况)。显然,在未给定预定义格式文件时,SPSS 基本上是不可能正确识别该文件的。因此保持默认的选择“否”并直接单击“下一步”按钮。

(2) 在图 2.13(b)所示的对话框中设定变量排列方式和变量名行,如果文件中有变量名,则需要将“变量名称是否包括在文件的顶部?”单选按钮组改为“是”,然后单击“下一步”按钮。

(3) 在图 2.14(a)所示的对话框中确定数据开始行、每个个案所占的行数、希望导入的个案数量,一般前两者的默认设定就是最常见的情况,第 3 个功能则可以用于对个案进行随机抽样。

(4) 对变量分隔符以及文本限定符进行设定,如图 2.14(b)所示,根据相应选项的设定情况,下方会动态显示出数据的预览情况。本数据采用的是 Tab 键,可见系统已经自动识别并选择了 Tab 键,而下方的数据预览窗口也已经显示出了正确的数据读入情况。右侧的文本限定符单选按钮组提供了“无”、“单引号”、“双引号”和“自定义”4 种选择。如果数据中的字符串变量使用限定符进行了分隔,则需要在此处指定。



(a)



(b)

图 2.14 “文本导入向导”的第 3、4 步

(5) 在图 2.15(a)所示的对话框中对各变量做进一步的属性设定,包括更改变量名和更改数据格式,在数据预览窗口中选择某一列变量即可进行操作,如果这里不需要进行更改,可以直接单击“下一步”按钮。

(6) 在图 2.15(b)所示的对话框中确定是否希望重复利用本次操作的选择,可以考虑将这次的文件设定保存为预定义格式文件,或者将本次操作粘贴为 SPSS 语句。如果直接单击“完成”按钮,则向导结束,随后就可以看到 SPSS 成功地读入了该文本数据。



(a)



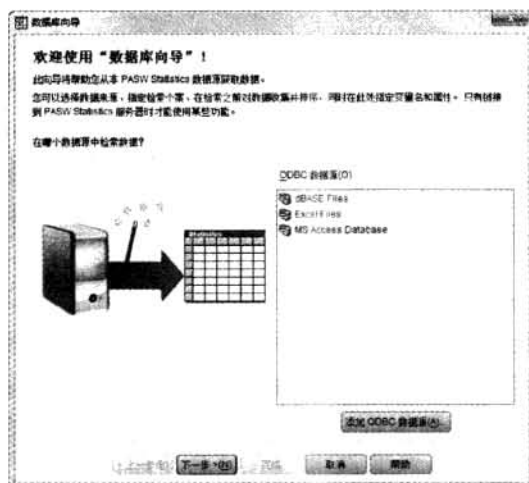
(b)

图 2.15 “文本导入向导”的第 5、6 步

2.4.3 用 ODBC 接口读取各种数据库文件

SPSS 可以直接读取很多类型的数据文件,对于不能直接打开的数据格式,SPSS 提供了利用通用的数据库 ODBC 接口读取数据的方法。这里以 SPSS 系统自带的文件 demo.mdb 为例来介绍如何使用数据库查询方法读取这个文件。

首先,选择“文件”→“打开数据库”→“新建查询”菜单项,系统会打开数据库向导的第一个对话框,其中会列出本机上已安装的所有数据源,如图 2.16 所示,其中列出了需要的 MS Access DataBase 数据源。选中并单击“下一步”按钮,则打开 ODBC 驱动程序登录对话框,要求选择数



(a)



(b)

图 2.16 “数据库向导”初始对话框中的数据源列表和 ODBC 数据源管理器

数据库文件。在该对话框中选 `demo.mdb` 文件并单击“确定”按钮,系统就会进入数据库向导的第2个对话框,采用拖放式操作将所需变量拖入右侧列表框中。数据库向导的第3、4步用于进行数据的选择性读入、字符值到数值与值标签的转换等操作。第5步则提供了将生成的SQL语句保存为文件以供再次使用,以及将前面的指定粘贴成Syntax语句等功能。如果不需要这些设置,则可在第2步时直接单击“完成”按钮,数据就被成功读入了。



在有的SPSS版本中,ODBC驱动可能无法直接使用,需要先进行定义:单击“数据库向导”对话框中的“添加ODBC数据源”按钮,系统会弹出操作系统的ODBC数据管理器窗口,在其中单击“配置”或者“新增”按钮,按照相应的提示进行操作即可。

由于SPSS现在可以直接打开许多常用格式的数据文件,因此数据库查询接口的用处不是很大。但是使用ODBC接口可以直接和绝大多数流行的数据库进行数据交换,如SQL Server、DB2、Oracle等,这是采用直接打开的方式无法做到的。其次,在例行工作中,比如每月都要读入相同的数据库,可以将所使用的SQL语句存储起来,每次只需调用SQL语句即可,这一方法也可用来解决一些需要对动态数据库进行统计分析的问题。数据是在需要分析时才临时读入的,从而可以保证数据始终是最新的。建议读者到SPSS官网(<http://support.spss.com>)下载SPSS的OEM版ODBC驱动程序,其中提供了大量数据库格式的ODBC驱动程序。

2.5 数据的保存

在数据录入过程中,要注意随时保存,以防出现意外情况,导致信息丢失。SPSS不仅能将数据保存为自己的数据格式(*.sav文件),而且还可以将数据保存为其他类型,基本上对于能够读入的文件格式,目前SPSS都可以做到反向回写。

1. 保存为SAV格式

无论是数据录入过程还是对数据做了修改,随时保存数据文件都是必不可少的工作之一。选择“文件”→“保存”菜单项,如果数据文件曾经存储过,则系统会自动按原文件名保存数据;否则会弹出“将数据保存为”对话框,只需为所要保存的文件指定文件名和保存的路径即可。

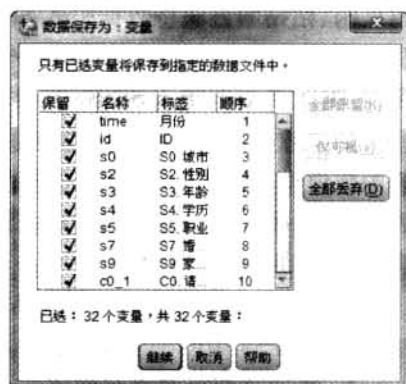
另外,有时分析者会在分析过程中生成一些临时变量,如果不希望保存全部变量,那么就可以单击图2.17(a)所示对话框中的“变量”按钮来指定需要保存的变量。如图2.17(b)所示,在每个变量的最左侧都有一个复选框,表明它们是否会被保存在文件中。对于不需要的变量,单击相应复选框取消选中,则该变量就不会出现在新保存的数据文件中了。

2. 保存为其他数据格式

SPSS软件的开放和友好之处不仅在于可以读取非SPSS类型的数据,而且它还允许将数据保存为很多种非SPSS格式的数据。从图2.17(a)所示的对话框中可以看到,最下方有一个“保存类型”下拉列表框,单击后可以看到SPSS能够保存的各种数据类型,有DBF、Excel、SAS版本的各种数据格式、纯文本格式等,用户只需要选择合适的类型,然后单击“确定”按钮即可。不过,将数据保存为SPSS以外的其他类型,有些设置可能会丢失,如标签和缺失值等,虽然在保存为SAS等数据格式时SPSS会提示将标签等另存为一个SAS程序文件,但这样毕竟不太方便,因此除非确实需要和其他软件交换数据;否则在决定保存为其他类型的数据时,一定要慎重从事。



(a)



(b)

图 2.17 “将数据保存为”对话框及“数据保存为:变量”对话框

2.6 数据编辑窗口常用操作技巧集锦

在本章的最后将回归到 SPSS 软件本身,介绍一下数据编辑器中有哪些常用技巧可以方便人们的日常工作。和其他常用统计软件相比,SPSS 数据界面最大的优势就是支持鼠标的拖放操作,以及复制、粘贴等命令,下面的数据录入技巧就是对这些功能的利用,它们都是作者在这些年使用 SPSS 的过程中总结出来的,希望对读者能有所帮助。

1. 连续输入多个相同值

如果需要在数据窗口的许多连续单元格中输入相同数值,则可以首先在其中任意一个单元格内输入相应的数值,比如“1”,按回车键后右击该单元格,在弹出的快捷菜单中选择“复制”菜单项,然后用鼠标左键拖动选择所有希望填入该数值的单元格区域,再单击右键,在弹出的快捷菜单中选择“粘贴”菜单项,则所有被选中的单元格都会被自动填入该数值。

需要指出的是,该操作在数据视图和变量视图中均可进行,大家可以自行尝试一下。

2. 快速定义成批变量

在变量视图中定义新变量时,按回车键后当前单元格默认向右侧单元格移动,直到移动了 10 个定义框后,才开始定义下一个变量,实际上其中绝大部分都可以采用默认值,如果需要同时定义大批变量,这样将非常浪费时间。可以在输入变量名后使用方向键而不是回车键让当前单元格向下移动,直到将所有新变量名称都定义完毕之后再使用 Label 栏定义批量变量名标签,Values 栏定义变量值标签,这样可以成倍地提高工作速度。

另外一种快捷的方式是,如果需要定义很多变量,同时对变量名要求不严,SPSS 自定义的变量名就可以满足需求,则可以在变量视图中直接跳到最后一行变量设定处,例如,需要定义 50 个变量,就直接跳到第 50 行,在此处输入变量名,按回车键后就可以看到 1~49 行将会自动填充好相应的变量。然后只需要修改不合适的设定,就可以使用了。

3. 将 Excel 或 Word 中的数据直接导入 SPSS

对于 Excel 数据文件而言,如果在 Excel 中已经打开原数据文件,并且数据量较少,可以直接用复制、粘贴的方法将数据引入 SPSS 中。先在 Excel 中选中所有的数据(不包括变量名),然后执行“复制”命令;然后切换到 SPSS,最好使行 1 列 1 单元格成为当前单元格,然后执行“粘贴”命令,数据就会全部转入 SPSS 中,再定义相应的变量即可。

如果数据中含有文本,则不能直接粘贴;否则会丢失数据。这是因为 SPSS 默认的数据格式均为数值型,这样将文本粘贴过来就会变为缺失值。解决办法是先在 SPSS 中设定好相应的变量列表,包括数值型、字符串型这些属性,然后再对相应的列进行粘贴,此时字符串数据就不会丢失了。

对于 Word 文档中的数据表格,其操作方式和 XLS 文件基本相同,粘贴后原来的单元格会自动对应为 SPSS 中的一个单元格。


4. 快速改变变量排列次序


在数据视图中选中列首的相应变量名,松开左键后再按下左键不放,就可以将该列数据拖动到任何希望的地方去了。选择时可以选中连续的多个变量,此时这些变量会同时发生变化;但如果选中不连续的多个变量,拖动时只对居中的一个起作用。

该操作也可以在变量视图中进行,此时应当选中变量的相应行号,其余操作相同。

5. 快速定位记录

记录的快速定位在大型数据集操作中非常有用,具体可以分成以下两种情况。

(1) 快速定位到第 N 条记录,此时可选择“编辑”→“转至个案”菜单项,或者直接单击工具栏上的按钮,在打开的对话框中输入相应的记录号,单击“确定”按钮后即可。


(2) 定位到变量值等于某个取值的记录(如 $ID = 34980$),此时需要先使相应的变量成为当前列,然后单击按钮,在打开的“查找”对话框中输入相应的数值,单击“确定”按钮后系统就会查找到符合条件的第 1 条记录,单击“查找下一个”按钮则会继续查找第 2 条记录。当然,如果事先排好序,则查找一次就可以了。

6. 利用排序功能快速查找异常值、极端值

对于异常值、极端值的发现,标准的做法应当是做出频数表看看有无异常值,但这样做过于麻烦,而且往往无法马上知道是哪一条记录出错了。其实最简单的做法是在数据视图中选中列首的相应变量名,然后右击,根据需要选择快捷菜单下方的“升序排列”或“降序排列”菜单项,相应的最小值(或缺失值)、最大值就会成为第 1 条记录。现在数据中有无异常值、极端值即可一目了然。



7. 利用变量值标签检查录入错误

前面曾经提到过对于单选题,最好采用“数值代码+值标签”的方式进行录入。这里再详细解释一下用法:实际上除了注释变量外,绝大多数字符型变量都只有少数的几种取值,因此可以将这些变量一律按照数值型变量来设置,录入时只需要输入编制的代码 1,2,3,...,然后利用变量值标签将实际含义一一写入标签中,这样可以大大加快速度。

下面来进行最重要的一步:在菜单栏上选择“视图”→“值标签”菜单项,或者直接单击工具栏上的按钮,在该按钮被按下后,数据编辑器中所有设定了值标签的变量值均会切换成相应的值标签。该按钮弹起后,则仍然按照录入的数值来显示,这不仅仅只是一个好看的问题,当单

击数据单元格时,相应的变量值标签会以下拉列表的形式呈现供用户选择,以免出现录入错误,同时通过排序,可以很快地发现缺失值和无标签的数值,而后者往往就是错误的数值。

8. 冻结行或列

对于熟悉 Excel 操作的用户而言,对右侧的若干列,或者上方的若干行进行冻结是非常便捷的,SPSS 实际上也可以实现类似的功能。仔细观察可以发现,数据编辑器电子表格的右侧、下方分界线中部有类似的标记,将鼠标移动到该界线处,可以发现鼠标会变成形这种双向调整符号。此时按住左键不放,就可以上下/左右拖动分界线,至合适的位置松开左键,就会发现电子表格将被该界线分为两半。对右侧和下侧同时进行该操作,则最多可将数据编辑器分为 4 部分,如图 2.18 所示。这 4 部分均有独立的上下和左右滚动条,从而可以实现某些行/列滚动、其余行/列固定的效果。

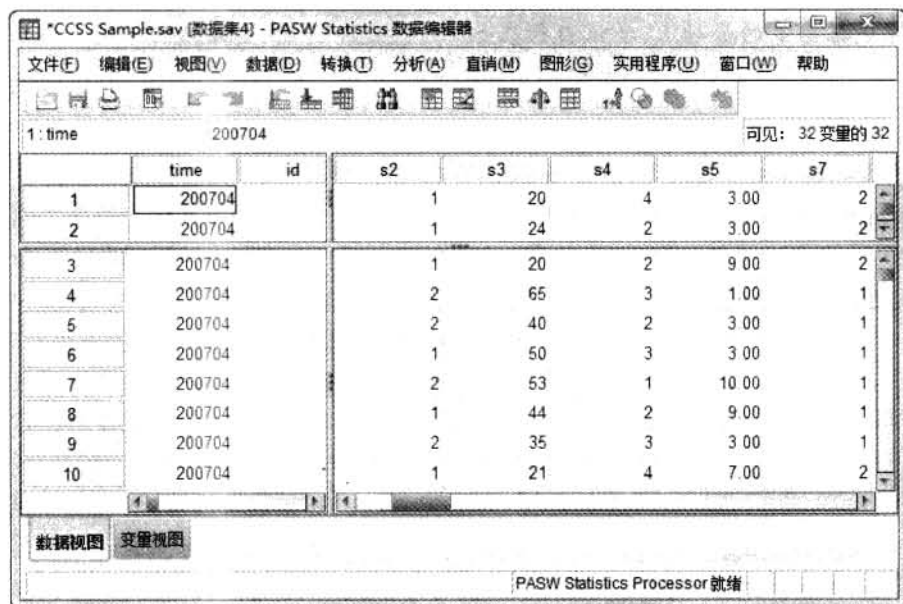



图 2.18 拆分状态的数据窗口

对于不习惯鼠标拖放操作的用户,也可以选择“窗口”→“拆分”菜单项,则数据窗口就会被直接拆分成 4 部分,然后再将分界线拖动至合适的位置即可。


如果希望取消冻结,则将分界线重新拖动至右侧/下侧即可,相应的分区就会消失。

9. 快速重复调用对话框

一般进行分析都是从菜单上依次选择相应的菜单项,这样比较麻烦,该问题在需要重复做几次相同的分析时尤为突出。实际上只要单击工具栏上的按钮,所弹出的下拉列表中就会依次列出最近几次使用的一些对话框。可以直接从中选择所需要的功能,比选择菜单方便多了。

该对话框中可以列出最多 9 个最近使用过的对话框,并且对话框中的相应选项设定都会得到保留(在该数据集关闭前均有效),用户如果需要重复使用这些对话框,就不需要再去选择菜单了。

10. 从其他窗口中快速切换回数据窗口

数据编辑窗口是 SPSS 的核心窗口,当需要从其他 SPSS 窗口中切换回去时,如果从系统任务栏上选择显得比较麻烦,实际上所有其他类型的 SPSS 窗口在工具栏上都有一个按钮,只要单击这个按钮,系统就会立刻切换回数据编辑窗口。如果有多个数据窗口同时存在,则切换回最后一次打开的那个数据窗口。

思考与练习

1. 针对 SPSS 自带文件 demo. xls,进行以下练习。
 - (1) 将该文件读入 SPSS 中,仅包含以下变量:年龄、婚姻状况、家庭住址、收入。
 - (2) 对变量 MARITAL(婚姻状况)设置值标签,1 代表已婚,0 代表未婚。
2. 在完成练习 1 的基础上,尝试自行在 SPSS 中按照 CCSS 项目的问卷建立相应的数据集结构。

第3章 变量级别的数据管理

通过第2章的学习,读者应该已经掌握如何将CCSS项目的原始数据在SPSS中录入为数据文件。但是,原始数据库往往不能直接用于最终的统计分析,这不仅是因为数据库可能因录入错误或原始问卷记录错误等情况包括不正确的数据,还因为针对同一个研究目的,往往要从各种不同的侧面对数据进行研究,采取多种统计方法进行分析,而不同的统计方法对数据文件结构的要求不尽相同,这就需要对数据文件的结构进行重新调整或转换,以便适合于相应的统计方法使用,上述这些工作被统称为数据管理。数据管理是统计分析工作的一个非常重要的环节,直接关系到数据分析的结果,是统计分析工作中不可缺少的一个关键步骤。

在SPSS中,数据文件的管理功能基本上都集中在“转换”和“数据”两个菜单中,其中前者主要实现变量级别的数据管理,如计算新变量、变量取值重编码等,主要与变量数值的转换有关;而后者功能主要是实现文件级别的数据管理,如变量排序、文件合并、拆分等。本章和下一章将分别介绍这两个菜单的相应功能。

“转换”菜单中提供了较多的变量转换功能,如图3.1所示,初学者往往弄不清楚孰轻孰重,实际上,该菜单中的项目大致可分为以下几类。

(1) 计算新变量:是菜单最上方的“计算变量”过程,这是该菜单中最为常用和重要的过程。

(2) 变量转换:包括从菜单第2项开始的多个计数过程、重编码过程和离散化过程,它们实际上都可以被看成是“计算变量”过程某一方面功能的强化和打包。

(3) 时间序列模型专用过程:包括最下方的“日期和时间向导”、“创建时间序列”、“替换缺失值”3个过程,由于其均专用于时间序列模型,对它们的讲解可参见本丛书的高级教程,本基础教程将不对其进行介绍。

(4) 自动数据准备(ADP):准备分析数据是项目中最重要的一步骤之一,而从传统来说也是最耗时的步骤之一。自动数据准备是指基于统计算法来协助用户自动分析数据并识别、修正、筛选出存在问题或可能无用的字段,并在适当的情况下派生新的变量,以简化用户的分析操作。由于该模块过于智能化,并不适用于普通用户,因此这里只做简单介绍。

(5) 其他:包括用于设定伪随机函数种子的随机数字生成器,以及继续执行编程中被挂起(Pending)的转换操作的“运行挂起的转换”。本书也将只做简单介绍。

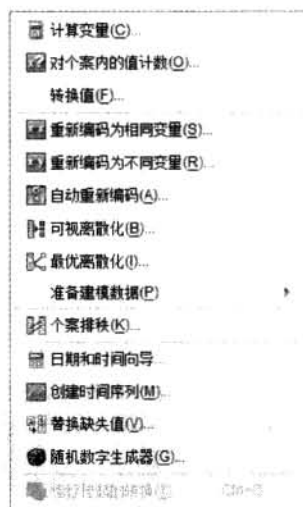


图3.1 “转换”菜单

3.1 变量赋值

所谓变量赋值,就是指在原有数据的基础之上,根据用户的要求,使用 SPSS 算术表达式及函数,对所有记录或满足 SPSS 条件表达式的某些记录进行四则运算,并将结果存入一个用户指定的变量中。该指定变量可以是一个新变量,也可以是一个已经存在的变量。变量赋值是极为常用的操作,大致可占数据管理操作的一半以上。

3.1.1 常用基本概念

在变量赋值中会涉及 SPSS 算术表达式、SPSS 函数、SPSS 条件表达式等基本概念,这里首先简单介绍一下这些概念。

1. 算术表达式

在变量转换的过程中,应根据实际需要,指出按照什么方法进行变量转换。这里的方法一般以 SPSS 算术表达式的形式给出。SPSS 算术表达式(Numeric Expression)是由常量、SPSS 变量名、SPSS 的算术运算符、圆括号等组成的式子,参与运算的数据类型和最终结果均为数值型,字符型和日期型变量/常量则要先进行函数转换然后才能参与运算。

算术表达式中的运算符由加(+)、减(-)、乘(*)、除(/)、乘方(**)构成,运算顺序以及括号的使用均遵循四则运算法则。

2. 函数

数据处理中仅有算术表达式显然是不够的,为此 SPSS 提供了多达百余种的系统函数。根据函数功能和处理对象的不同,可以将 SPSS 函数分成 8 类,分别是算术函数、统计函数、分布函数、逻辑函数、字符串函数、日期时间函数、缺失值函数和其他函数。

函数的具体书写形式为:函数名(参数)。这里,函数名是系统已经规定好的。圆括号中的参数有时是一个,也可以是多个;而参数的类型有时是常量(字符型常量应用单引号引起来),也可以是变量名或 SPSS 的算术表达式。此外,函数中如果有多个参数,各参数之间要用单字符逗号“,”隔开。

SPSS 函数一般也会与 SPSS 的算术表达式混合出现,用于完成更加复杂的计算。各种函数的释义可参考本书附录。

3. 条件表达式与逻辑表达式

通过 SPSS 的算术表达式和函数对所有记录进行计算时会得到一个结果,如果仅希望对部分记录进行计算,则应当利用 SPSS 的条件表达式指定对哪些记录进行计算。根据实际需要构造条件表达式之后,SPSS 将对条件表达式进行计算得到一个逻辑常量(真或非真),然后从所有记录中自动挑选出满足该条件的记录,然后再对它们进行计算处理。

在 SPSS 中条件表达式中常用的关系运算符有以下几种: <、>、<=、>=、=、~=,其中最后一个符号的含义为“不等于”,在 SPSS 中也可以使用英文缩写,例如,“~= ”也可写为“NE”,但对国内读者而言,建议还是使用数学运算符来书写。

除了条件表达式外,在 SPSS 中还会使用到逻辑表达式;其作用和赋值类型均类似于条件表达式,只是会用到以下 3 个逻辑运算符:&、|、~,分别表示 AND、OR 和 NOT。

3.1.2 “计算变量”过程对话框

在 SPSS 中,变量赋值主要是通过“计算变量”过程来实现的,选择“转换”→“计算变量”菜单项,即可打开如图 3.2 所示的对话框。

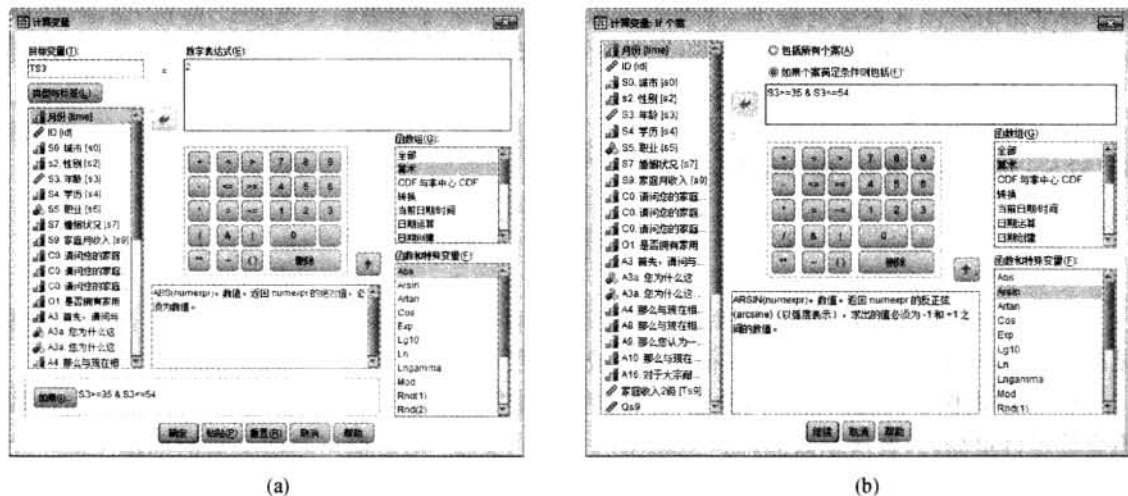


图 3.2 “计算变量”对话框

(1) “目标变量”文本框:用于输入需要赋值的变量名,在输入变量名后,下方的“类型与标签”按钮就会变黑,喜欢精确设定的朋友可以在这里对变量进行具体的定义,但在大多数情况下都是不需要更改的。

(2) 候选变量列表:位于“目标变量”文本框下方,可以用鼠标和右侧的变量移动按钮将选中变量移入右侧的“数字表达式”文本框中。

(3) “数字表达式”文本框:其实翻译成数值表达式更为妥当,用于给目标变量赋值。

(4) 软键盘:位于对话框中部,是类似计算器的软键盘,可以用鼠标按键输入数字和符号。

(5) 函数列表:位于软键盘右侧和下侧,分为“函数组”列表框、“函数和特殊变量”列表框、函数解释文字文本框 3 部分,可以在这里找到并使用所需的 SPSS 函数,图 3.2(a) 中所示的是选择了“算术”函数组、abs 函数之后,解释文字文本框中出现 abs() 相应解释的情形。

(6) “如果”按钮:用于对个案筛选条件进行设定,单击后打开如图 3.2(b) 所示的对话框,默认选中“包括所有个案”单选按钮,如果需要进行个案筛选,则更改为“如果个案满足条件则包括”,然后在下方的表达式文本框中输入相应筛选条件即可。完成之后单击“继续”按钮,会看到“如果”按钮右侧显示出相应的筛选条件表达式。

3.1.3 案例:年龄变量 S3 的分组

例 3.1 CCSS 项目中的受访者年龄为 18~64 岁,分析时将其分为 18~34、35~54、55~64 三组。为了便于使用,年龄变量 S3 被重新赋值后将会保存为新变量 TS3,其取值 1、2、3 分别代表上述 3 种情况。

本例实际上属于变量重编码的情形,但这里首先利用数值计算过程的条件筛选方式来实现。也就是说,如果希望对全部个案生成一个新变量,但不同人群采用不同的算术表达式,可以通过设定不同的筛选条件多次调用“计算变量”过程实现。

(1) 打开“计算变量”对话框,设定目标变量名为 TS3,数字表达式为“1”,确认后即建立该新变量,取值为 1。

(2) 再次打开“计算变量”对话框,更改数字表达式为“2”,单击“如果”按钮,设定筛选条件为“S3 >= 35 & S3 <= 54”,如图 3.2(b)所示,依次确认。

(3) 再次打开“计算变量”对话框,更改数字表达式为“3”,单击“如果”按钮,设定筛选条件为“S3 >= 55”,依次确认,操作完成。

3.2 已有变量值的分组合并

在数据分析中,将连续变量转换为等级变量,或者将分类变量不同的变量等级进行合并是常见的工作。而通过变量重编码可以很好地完成这一类任务。

SPSS 中提供了功能类似的两种重编码过程,其中“重新编码为相同变量”是对原始变量的取值直接进行重编码,替换原数值;而“重新编码为不同变量”则是根据原始变量的取值生成一个新变量来记录重编码结果。两者除了输出目标不同之外,其余功能是非常类似的,因此本节将以功能更强的后者为主进行讲解。

3.2.1 对连续性变量进行分组合并

在 SPSS 中可以将连续变量转换为离散(等级/定序)变量,按照某种一一对应的关系生成新变量值,可以将新值赋给原变量,也可以生成一个新变量。通过重编码过程和下一节讲解的可视化分段过程都可以完成这一任务,但前者更为简单和常用。

现在仍以例 3.1 的分析为例来讲解重编码的具体操作,前面用变量赋值的方式完成了相应操作,但需要重复调用 3 次对话框。如果使用重编码过程,则一次就可以完成。选择“转换”→“重新编码为其他变量”菜单项,打开如图 3.3 所示的对话框。将 S3 年龄选入“数字变量→输出

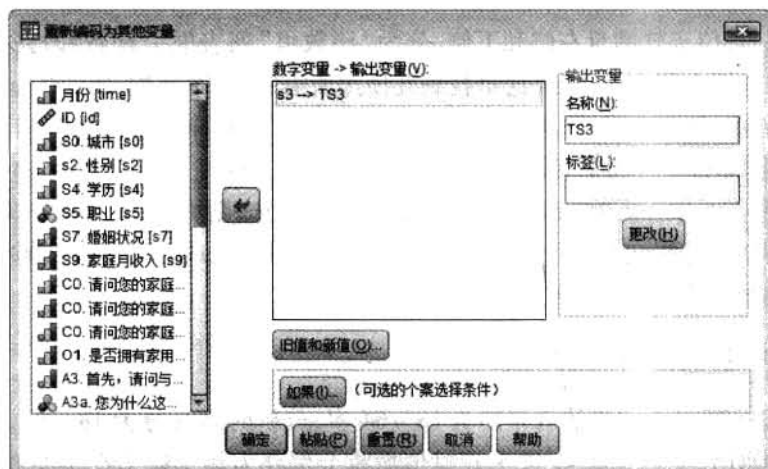


图 3.3 “重新编码为其他变量”对话框

变量”列表框中,此时“输出变量”框组变黑,在“名称”文本框中输入新变量名 TS3 并单击“更改”按钮,原来的 S3 -> ? 就会变成 S3 -> TS3,即新老变量名间已经建立了对应关系。

现在单击“旧值和新值”按钮,系统打开“重新编码到其他变量:旧值和新值”对话框,如图 3.4 所示。对话框左侧为原有变量的取值情形,右侧为新变量的赋值设定。两边设定完毕后单击“添加”按钮,相应的对应规则就会被加入规则列表中去。但要注意所有的范围都是包含了端点的,虽然此时前面设定的变换会优于后面的变换,但为了避免误解,这里将不包括端点数值的情形均设定为小数数值(已知 S3 为整数),读者可仔细体会这一技巧。

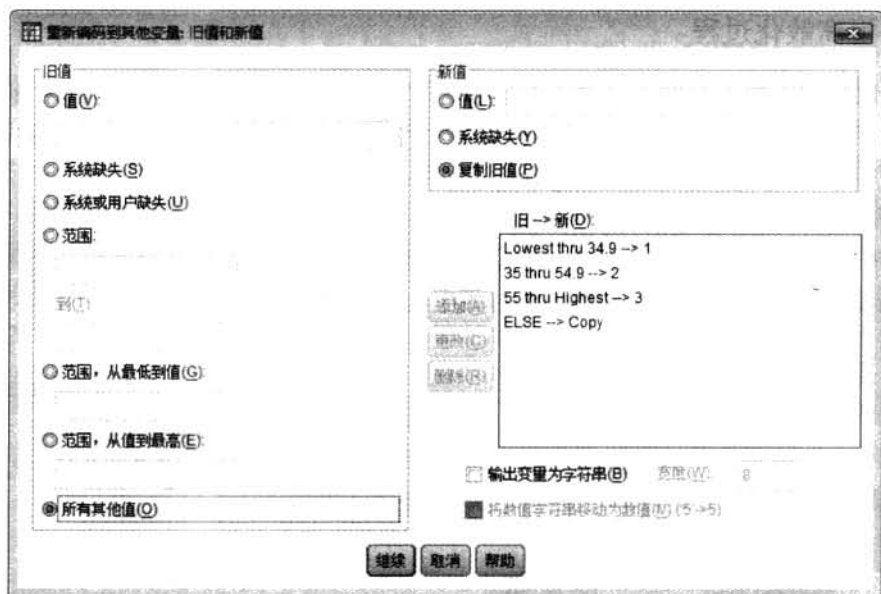



图 3.4 “重新编码到其他变量:旧值和新值”对话框

 此处设定的转换规则为 4 条,很多初学者会觉得 3 条就够了,比如说第 3、4 条规则完全可以用一条“ELSE -> 3”来替代。但需要指出的是,在数据整理工作中一定要考虑到数据出错等例外情形,如果变量 S3 中存在缺失值、错误的异常数值等情形,只采用 3 条转换规则就会将错就错无法发现,而上述设定可以将异常情况保留在新变量中,从而能够在后续分析中将其发现。一言以蔽之,数值处理程序不能随意简化,而应当充分考虑各种极端情形下的需求。

上述重编码过程既可以将连续变量转化成数值型或者字符型离散变量,也可将数值型字符变量转化成数值型变量,只需选中图 3.4 中的复选框“将数值字符串移动为数值”即可。

3.2.2 分类变量类别的合并

重编码过程也常用于合并某个分类变量的几个水平为一个水平,如果分类变量的记录格式为数值型,则操作与例 3.1 基本上没有区别。但如果其存储格式为字符型,则需要注意默认的转换格式为数值型,如果仍希望将其转换为字符型,则需要选中复选框“输出变量为字符串”。

3.3 连续性变量的离散化

重编码过程提供了精确分组的功能,但是如果希望进行的分组是较有规律的,比如说等距分组,或者等样本量分组,使用重编码过程进行操作就显得非常麻烦,且可视化程度不高,此时可以考虑使用可视化过程进行分段。SPSS 中提供了两种可视化分段过程,分别为需要用户自行判断设定的可视离散化过程,以及基本全自动的最优离散化过程。

3.3.1 可视离散化过程

可视离散化过程用于在可视界面下将连续变量进行分段,在该过程中可以使用百分位数、标准差范围或者等间距方式将连续变量划分为若干组段,并采用图形化操作的方式,非常直观好用。

1. 对话框

选择“转换”→“可视离散化”菜单项之后,打开的对话框要求用户选择希望进行离散化的变量,选择完毕后单击“继续”按钮,则系统会对相应的变量进行数值扫描,并打开如图 3.5 所示的对话框。

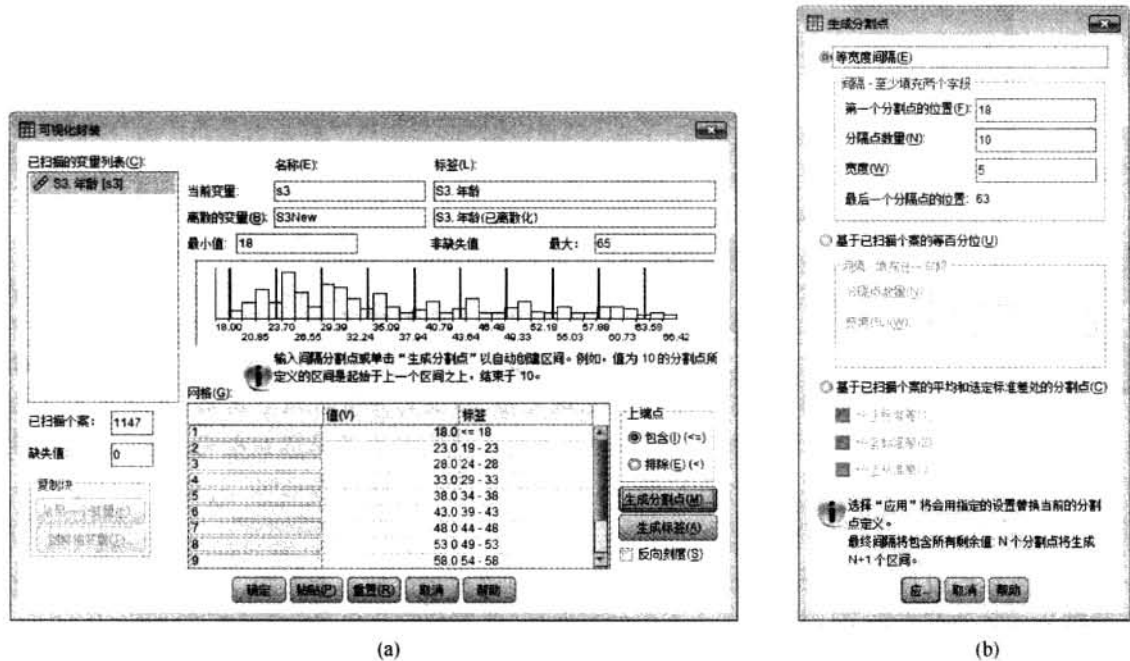


图 3.5 “可视化封装”对话框和“生成分割点”对话框

(1) 已扫描的变量列表:列出了在前一个对话框中所选择的所有变量,更改此处的变量选择,则对话框中所有其余部分的内容均会按照所选中的变量状况进行更新。

(2) 左下侧“复制块”框组:当选择了多个变量且其中部分变量已设定完离散化选项时可用,可以将设定好的属性复制“到其他变量”,也可以“从另一个变量”(即已设定好的变量)读取

相应的设定。

(3) 上部变量属性:列出新老变量的名称和标签,注意其中新变量名称是必填的;否则离散化完毕后不会生成任何新变量。

(4) 中部直方图:扫描完原变量取值情况后在此处绘制该变量的直方图,如果已设定完毕分割点,也会一并显示。

(5) 下部数值标签网格:在本网格处显示所设定的分割点数值位置和相应的标签。

(6) 右下侧“上端点”框组:用于设定端点是否被包括在上侧区间内。

(7) “生成分割点”按钮:单击后打开如图 3.5(b) 所示的对话框,其中可以选择使用等间距(Equal Width Inter)、等比例(等样本量, Equal Percentiles based on Scanned Cases)或者按照指定的标准差范围(Cutpoints at Mean and Selected Standard Deviations based on Scanned Cases)3 种方式进行分段,其中第 3 种方式显然可以用来在数据分析或质量控制中筛选异常值。

(8) “生成标签”按钮:在分割点数值设定完毕后,单击该按钮可以自动生成相应的值标签。

(9) “反向刻度”复选框:在默认情况下,新的离散化变量的值是从 1 到 n 的升序整数。反向刻度会使得这些值成为从 n 到 1 的降序整数。

2. 实例分析

例 3.2 将 S3 年龄变量分为 10 组,要求等间距。

本例实际上是要求对连续变量进行统计描述中的直方图分组,由于已知年龄范围为 18 ~ 65 岁,全距为 48,因此在分为 10 组的情况下,组距为 5 即可覆盖全部取值范围。当然组数、组距和第一组段下限三者是相互联系的,在对话框中一般只需要定义其中两者即可自动确定第 3 个因素的取值。

(1) 选择“转换”→“可视离散化”菜单项,将 S3 年龄选入“要离散的变量”列表框中,单击“继续”按钮进入主对话框。

(2) 单击“生成分割点”按钮,设定分割点数量为 10,宽度为 5,可见系统会自动填充第一个分割点的位置为 18,单击“应用”返回到主对话框。

(3) 此时会看到下部数值标签网格的“值”列已被自动填充,单击“生成标签”按钮,使标签列也得到自动填充。

(4) 将离散的变量的名称设定为 S3New,单击“确定”按钮,系统会提示“封装规范将创建一个变量”,确认后就会在数据集中生成新变量 S3New。

如果注意一下结果窗口中的 LOG 输出就会发现,可视化分段过程实际上运行的是记录重编码所对应的 Recode 过程,也就是说,两者在代码级别上实际是一回事,只不过可视化分段过程在对话框界面上进行了进一步的开发而已。

3.3.2 最优离散化过程

“最优离散化”过程是对前述可视化离散过程的进一步自动化,根据某些作为“关键指示变量”的分类变量,将原有的一个或多个连续性变量按照该分类变量类间差异最大化的优化原则离散化为分类变量,然后就可以使用离散化变量而非原始数据值进行后续的分析了。

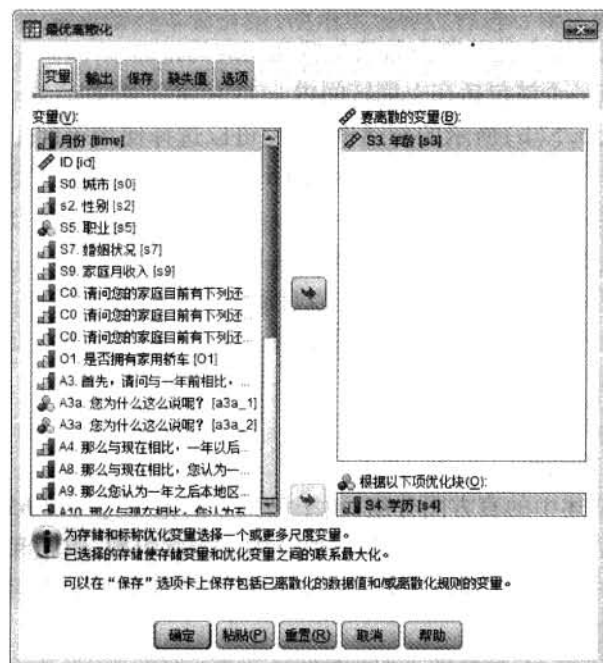
初学者可能对该过程的用途感到疑惑,实际上这主要和建模分析有关:当模型中的因变量为分类变量时,在分析中往往会对自变量进行离散化(分类化),此时就可以使用该过程。如果最

终目标是生成预测模型,则“最优离散化”的效果一般会优于可视离散化。

由于该过程涉及统计建模,这里不再进行详细讨论,只对其进行简单介绍。

1. 对话框

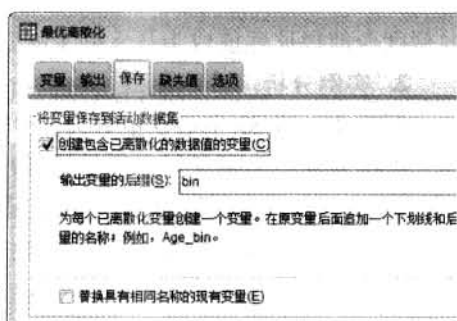
最优离散化过程涉及的对话框如图 3.6 所示。



(a)



(b)



(c)

图 3.6 最优离散化过程涉及的对话框

(1) “变量”选项卡:在图 3.6(a)将需要离散化的一个或多个连续性变量选入右上方的“要离散的变量”列表框中,右下方则用于选入作为关键指示变量(一般即为模型中的因变量)的分类变量,注意这里只能选入一个分类变量。

(2) “输出”选项卡:在图 3.6(b)设定在离散化结束后输出哪些统计结果,需要解释的是第 3 项“熵”,对于每个离散化输入变量,此选项显示相对于关键指示变量的预测准确性的改善情况,可作为离散化效果的测量指标,因为进行最优离散化一般就是为了改善预测效果。

(3) “保存”选项卡:在通过阅读输出结果确认离散化效果可以接受后,可在图 3.6(c)这里选择保存离散结果为新变量用于后续分析。同时也可以将相应的 recode 语句(还是 recode 命令!)存为程序文件以便重复利用。

(4) “缺失值”选项卡:定义当数据中存在缺失值时系统的处理方式,一般不用更改。

(5) “选项”选项卡:设定在要处理的是海量数据集、关键指示变量存在罕见类别(稀疏块)等情况下的处理选项,块的端点设定等细节一般不用更改。

2. 实例分析

例 3.3 利用 S3 年龄变量对 S4 学历进行预测建模,要求基于此构思对 S3 进行最优离散化。

本例相应的设定如图 3.6 所示,单击“确定”按钮后相应的结果输出如图 3.7 所示。

	N	极小值	极大值	相异值数	块个数
S3. 年龄	1 147	18	65	48	2

图 3.7 描述统计

图 3.7 给出的是对 S3 的描述结果,可见在 1 147 例个案中,S3 共有 48 种不同取值,在离散化结束后它们被分为两类(即图 3.7 中的块)。

图 3.8 为离散化结果的熵评价,这里不再详细介绍熵值的计算公式,只是明确一点:熵值只能在同一模型的不同离散化结果间进行大小比较,而不能在不同模型间进行比较。在同一模型框架下,熵值越低的离散化方式,将其用于预测时效果就越好。

模型熵	
S3. 年龄	2. 048

模型熵越小表示参照变量 S4. 学历上的离散化变量的预测准确性越高。

图 3.8 模型熵

图 3.9 为离散化后的分类变量和 S4 学历变量的交叉表,从中可以看出,40 岁以下组学历偏高,而 40 岁以上组则学历偏低。

块	端点		水平 S4. 学历的个案数					总计
	下限	上限	初中/技校 或以下	高中/中专	大专	本科	硕士 或以上	
1	a	40	56	149	236	234	48	723
2	40	a	98	164	95	58	9	424
总计			154	313	331	292	57	1 147

每个块的计算方法为:下限≤S3. 年龄<上限。

a. 无限制。

图 3.9 S3. 年龄

在后面学习了相关分析、单因素方差分析等方法之后,对本例感兴趣的读者可以对上述问题做深入分析,这里不再详述。

3.4 变量的自动重编码与数值移动

3.4.1 变量的自动重编码

在数据分析中,将字符变量转换为数值变量,或者将数值变量重编码是非常实用的功能。除了使用前面介绍的重编码过程手工设定转换规则外,还可以使用自动重编码过程自动按原变量值的大小或者字母排序生成新变量,而变量值就是原值的大小次序。

例 3.4 在 CCSS_Sample. sav 数据中,S0 城市的数值分别为 100、200 和 300,现将其自动重编码为 S0New。

选择“转换”→“自动重新编码”菜单项,打开如图 3. 10(a)所示的对话框。

由于图 3. 10(a)所示的对话框非常简单,这里就不再详细介绍,直接给出相应的结果如下。

s0 into S0New (S0. 城市)

Old Value New Value Value Label

100	1	100 北京
200	2	200 上海
300	3	300 广州

该结果输出列出了原变量数值和新变量数值的对应关系,可见原先的 100、200、300 现在分别重新编码为 1、2、3 了。

3.4.2 变量值的移动

在时间序列模型以及一些特殊方法中,个案是需要按照时间顺序排列的,而在分析中可能需要将相应的变量值前移或者后移,该操作在 SPSS 中以前可以利用 Lag() 函数来实现,现在则将相应的功能编制成了对话框,可以在菜单上直接调用。

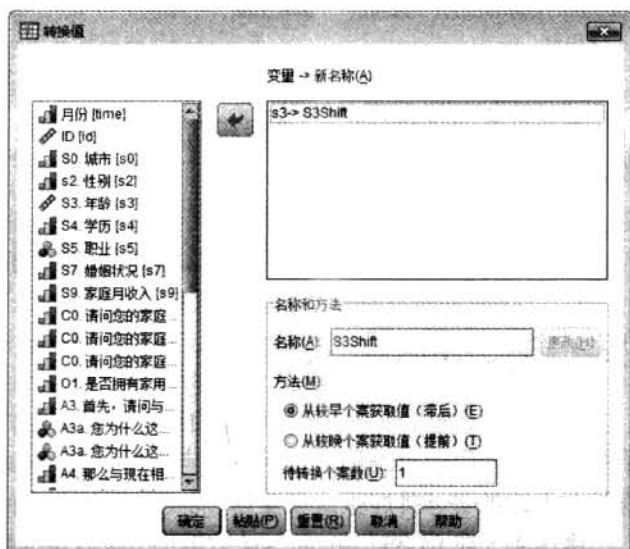
选择“转换”→“转换值”菜单项(注意该菜单实际上翻译有误,将 Shift Values 翻译成“数值平移”或者“数值移动”更为合适,而不是现在的“转换值”),打开如图 3.10(b)所示的对话框。该对话框同样非常简单,只说明以下内容。

(1) 滞后或提前:该单选按钮组用于确定相应变量列的数据单元格是按照案例顺序向前移动还是向后移动。

(2) 待转换个案数:该处同样翻译有误,“Number of cases to shift”翻译成“移动的个案数”更为合适,意思是将数据列向前/后移动的行数,默认为一行。



(a)



(b)

图 3.10 “自动重新编码”和“转换值”对话框

3.5 转换菜单中的其他功能

3.5.1 指定数值的查找与计数

对个案内的值计数(Count)的过程用于标识某个变量的取值中是否出现某个值,可以是单个数值,也可以指定区间,并且可以给出条件,从而不必对整个数据集进行操作。

选择“转换”→“对个案内的值计数”菜单项,打开如图 3.11 所示的对话框。

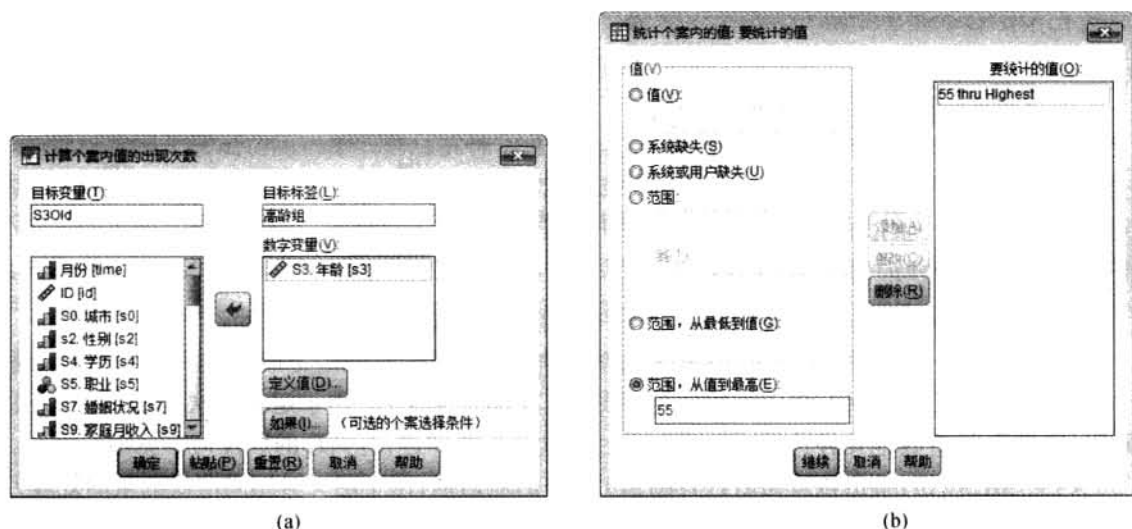


图 3.11 “计算个案内值的出现次数”和“统计个案内的值:要统计的值”对话框

(1) “目标变量”文本框:在图 3.11(a)中用于输入希望生成的计数变量名称,SPSS 在此处非常奇怪没有进行变量名的自动生成,这显然不太符合其一贯风格。

(2) “数字变量”列表框:名称显然是“数值变量”的误译,用于选入希望进行计数的数值型变量。

(3) “定义值”按钮:用于定义希望进行查找/计数的变量值范围,单击此按钮打开的对话框设定非常类似于重编码处的子对话框,因此不再重复解释。

例 3.5 生成新变量 S3Old,用于标识出 $S3 \geq 55$ 的个案。

相应的操作如图 3.11 所示,单击“确定”按钮后即会在数据集中生成新变量 S3Old,对于 $S3 \geq 55$ 的个案取值为 1;否则为 0。

3.5.2 变量的编秩

实际上,这里遇到的就是一个排次序的问题。个案排秩过程就是用来排次序的一个专用过程。具体来说,它就是根据某变量的数值大小来排出次序(秩次),然后将秩次结果存储到一个

新变量中去的过程。

例 3.6 根据 S2 性别分组计算 S3 年龄的秩次。

选择“转换”→“个案排序”菜单项,打开如图 3.12 所示的对话框。

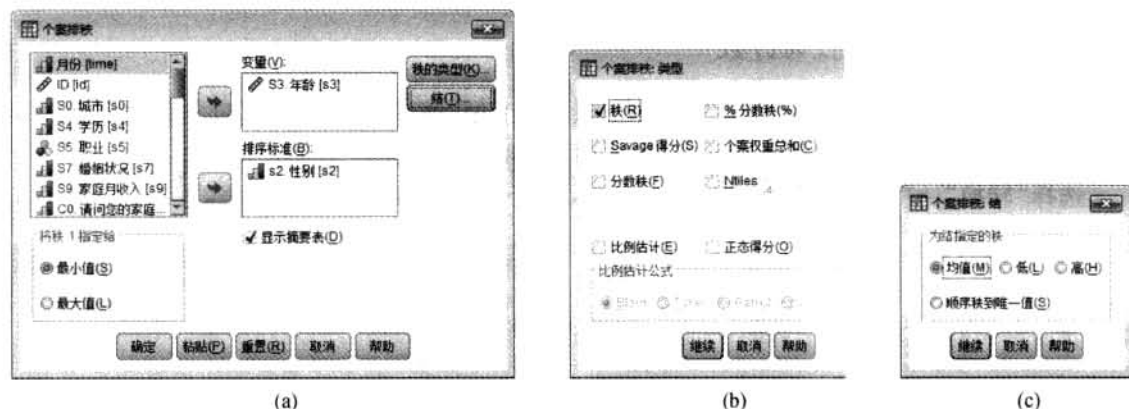


图 3.12 “个案排序”、“个案排序:类型”和“个案排序:结”对话框

在图 3.12 对话框中有以下内容。

(1) “排序标准”列表框:在图 3.12(a)中指的是分组编秩时的分组变量。

(2) “将秩 1 指定给”单选框组:用于选择将秩次 1 赋给最小值还是最大值。

(3) “秩的类型”按钮:单击该按钮打开图 3.12(b)的对话框,用于定义秩次类型,默认为最常用的“秩”,另有其他几种选择,因为很少用到,这里不再详述,有兴趣的朋友可参见用户手册。

(4) “结”按钮:单击该按钮后打开图 3.12(c)的对话框,用于定义对相同值观测量的处理方式,可以是“均值”、“低”、“高”或“顺序秩到唯一值”,默认值为取平均秩次。

这里将变量 S3 选入“变量”列表框中,分组变量 S2 选入“排序标准”列表框中,其他设置使用默认值,然后确认即可,此时系统会建立一个新变量 Rs3(即原变量名前加 R,表示“秩”),其取值为按照 S2 分组的 S3 秩次,同时在结果窗口中会给出汇总报表,如图 3.13 所示。

源变量	函数	新变量	标签
s3 ^a	秩	Rs3	按照 S2 分组的 S3 秩次

a. 秩按升序排列。

b. 值相同的平均秩用于结。

图 3.13 已创建的变量^b

许多时候参数检验的条件无法满足,需要使用非参数方法,而稍微复杂些的非参数方法就无法直接用对话框来完成了,所以需要先计算秩次再进行分析。这方面的内容详见非参数检验一章,这里不再详述。

3.5.3 自动准备建模数据

准备分析数据是分析项目中最重要的一步,而从传统来说也是最耗时的步骤之一。为此 SPSS 也开发了许多工具,前面介绍的最优离散化过程就是其中之一,这里的自动数据准备

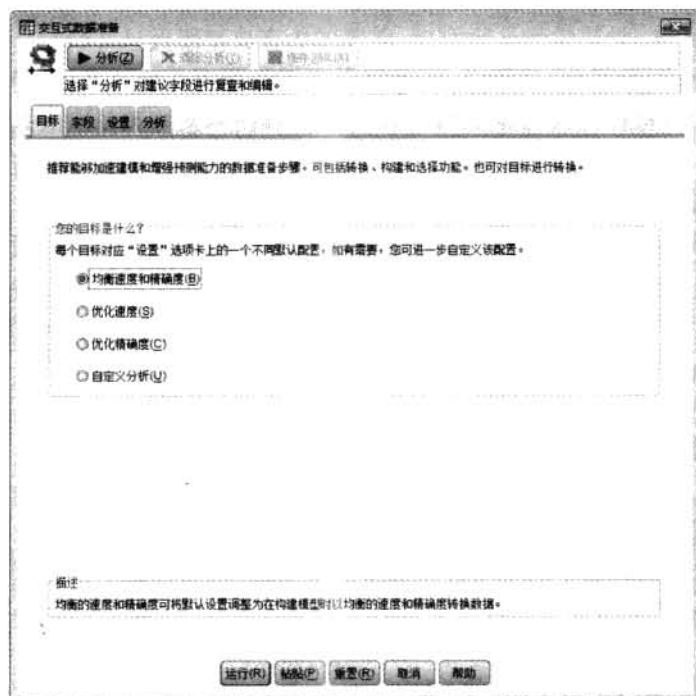
(ADP)则更进一步,在最终建立预测模型之前,使用该模块可以自动分析数据,对其中的异常值进行识别修正,筛选出存在问题或可能无用的字段,并在适当的情况下派生新的变量,并通过智能筛选技术改进性能。

自动准备数据过程中的因变量可以是连续、有序、无序等任何一种测量尺度,系统会自动选择相应的算法加以分析。用户可以采用几乎完全自动的方式使用该模块(这种方式允许选择并应用修正);也可以通过交互式方式(如图 3.14(a)所示)使用算法,这种方式可以在做出更改前对其进行预览,并根据需要选择接受或拒绝。

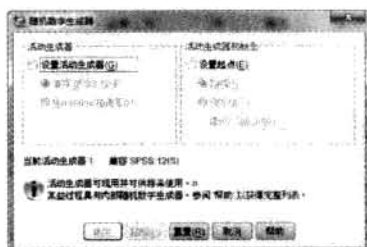
由于自动准备数据过程的自动化程度较高,并不适合于初学者使用,而其功能又从属于数据准备模块体系之下,因此这里将不对其进行深入介绍,而是在下一章从整个数据准备工作的层面上来说明数据准备模块的情况。

3.5.4 随机数字生成器

该过程用于设定伪随机函数的随机种子,但它对真随机函数没有任何影响。在默认情况下伪随机种子会随着时间不停改变,这样所计算出的随机数值无法重复,这在临床试验等场合中是不符合要求的。此时可用随机数字生成器(如图 3.14(b)所示)事先人为指定一个种子,之后所有的伪随机函数都会从该种子开始计算,本书第 5 章的程序示例中即使用了该生成器,只是以程序代码方式实现而已,感兴趣的读者可以参阅相关内容。



(a)



(b)

图 3.14 “交互式数据准备”和“随机数字生成器”对话框

思考与练习

1. 自行完成本章中涉及的对 CCSS 案例数据的数据管理操作。
2. 针对 SPSS 自带数据 Employee data. sav 进行以下练习。
 - (1) 根据变量 bdate 生成一个新变量“年龄”(提示:可以使用函数 XDATE. YEAR())。
 - (2) 根据 jobcat 分组计算 salary 的秩次。
 - (3) 根据雇员的性别变量对 salary 的平均值进行汇总。
 - (4) 生成新变量 grade, 当 salary < 20 000 时取值为 d, 在 20 000 ~ 50 000 范围内时取值为 c, 在 50 000 ~ 100 000 范围内时取值为 b, 大于等于 100 000 时取值为 a。

第4章 文件级别的数据管理

第3章主要介绍了如何对变量进行转换,可以满足数据管理中的许多基本需求。但是在数据管理中还会遇到许多文件级别的数据管理操作,如变量排序、文件合并、拆分等,在SPSS中,这些功能基本上都被集中在“数据”菜单中,如图4.1所示,根据各自的功能特点,该菜单中的所有项目可分为以下几类。

(1) 简单命令:包括插入变量、插入个案、到达某条个案、复制数据集等,它们的功能不言自明,且大多都可以使用鼠标在数据表界面上直接使用,很少使用菜单来调用,本书将不再对其进行讲解。

(2) 常用的简单过程:包括排序、拆分文件、个案筛选和个案加权,这几个过程并不复杂,但使用得极为频繁,是必须要掌握的内容。

(3) 数据重组向导:用于进行数据转置,或者对重复测量数据进行长型、宽型记录格式间的转换,后面会具体介绍。

(4) 文件合并向导:将几个数据文件合并为一个大的SPSS数据文件,含横向合并和纵向合并两种情况,后面会具体介绍。

(5) 与数据字典有关的功能:包括定义变量属性、复制变量属性,以及新建设定属性3个向导界面,以及在20版中新增的设置未知测量级别这一向导界面。对于较复杂的数据管理项目而言,这些都是非常有用的功能。

(6) 与数据准备有关的功能:同样是针对复杂数据管理项目的需求而开发的,用于简化数据管理工作。包括用于数据自动查错的数据验证模块,用于快速查找异常记录的重复个案与异常个案查找向导等。

(7) 与统计模型密切相关的过程:正交设计过程实际上是结合分析模块的一部分,用于生成结合分析所需的设计,它的讲解参见本丛书高级教程中的结合分析一章;定义日期变量过程用于时间序列数据的分析,将在高级教程的时间序列章节中讲解。

(8) 其他过程:包括定义多重响应集、数据汇总过程等,将在本章最后一节加以介绍。

由于内容庞杂,本章将着重介绍日常工作中最为常用和重要的一些过程,而对于在大型数据管理项目中才需要应用的一些向导和过程则以简单入门介绍为主。

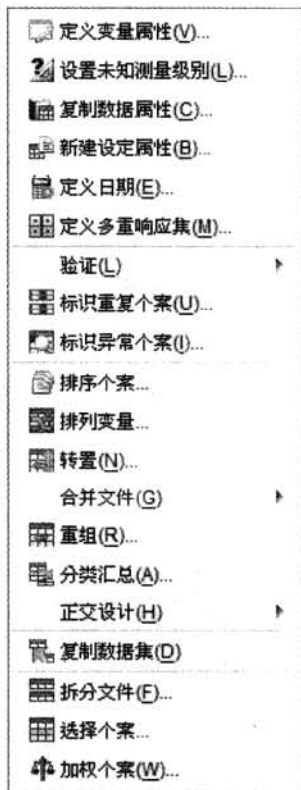


图4.1 “数据”菜单

4.1 几个常用过程

4.1.1 排序个案

数据编辑窗口中的记录的前后次序在默认情况下由录入时的先后顺序决定。在实际工作

中,有时希望按某种顺序来观察一批数据,例如,在 CCSS 数据存档中,将数据设定为首先按照月份升序,同月份数据按 ID 号升序的规则来排列,以便于随时检索和浏览。类似的情形还有很多:在销售报表中,希望按销售额从低到高的顺序,或者按销售时间从早到晚的顺序来浏览数据。观察排序后的记录数据,便于对数据获得更好的了解。

SPSS 中的个案排序就是将数据编辑窗口中的数据,按照用户指定的某一个或多个变量的变量值的升序或降序重新排列,这里用户所指定的变量称为排序变量。当对所有记录进行排序时,可以按照排序变量取值的大小次序对记录数据进行重新整理后显示出来。当对记录进行分组排序时,在每个组内,按照排序变量取值的大小次序对记录数据进行排序。

对于单变量排序,SPSS 提供了一种简易操作方法,就是在数据表格的变量名处右击,弹出的快捷菜单的最后两项就是“升序排列”和“降序排列”。但是,对于多变量排序,则需要使用这里讲述的“排序个案”对话框(如图 4.2 所示)来进行操作。由于该对话框并不复杂,因此这里不再详细讲解。

图 4.3 所示,将 CCSS 数据首先按照月份升序排列,月份相同时再按照 ID 号进行升序排列,注意在每个变量名后面都会跟有升序或者降序的说明。如果要改变升、降序,则选中相应的变量,然后直接在“排列顺序”单选按钮组中修改选择即可。在 20 版中,还可以在对话框中直接要求将排序后的文件存储为一个外部数据文件。



图 4.2 “排序个案”对话框



图 4.3 “分割文件”对话框

最后还需要说明以下几点。

(1) 在多重排序中,指定排序变量名的次序是很关键的,先指定的变量在排序时必然优先于后指定的变量。即记录首先按第 1 个变量进行排序,对于第 1 个变量的取值相同的记录考虑按第 2 个变量排序,依次类推。

(2) 可以指定按某变量值升序排序的同时按另一变量值降序排序,或相反。

(3) 排序以后,原来记录数据的排列次序将被打乱。因此,在时间序列的数据中,如果数据中没有存放记录标志的变量,如年份等,则应注意保存原数据的排列顺序,以免造成数据混乱。

4.1.2 分割文件

由于 CCSS 项目数据是逐月采集的,在对历史数据进行分析的过程中,经常会遇到希望将某种分析结果进行逐月对比的情形。对于此类需求可以有两种解决方式:将数据按月份进行拆分,然后同时完成各月数据的分析;或者将数据按月份进行筛选,然后依次加以分析。显然前者效率要更高一些,下面介绍对数据的拆分是如何实现的。

“分割文件”(Split File)对话框如图 4.3 所示,各个元素的用途说明如下:

(1) 右上部单选按钮组:用于设定如何拆分文件,默认为不拆分文件;第 2 项为按所选变量拆分文件,各组的分析结果会尽量放在一起输出(甚至于放在同一张表格里)以便于相互比较;第 3 种方式则为按所选变量拆分文件后,各组分析结果单独放置。

(2) 中部变量选择框:用于选入进行数据拆分的变量,可以选入多个。

(3) 右下部单选按钮组:设定文件的排序操作。默认为拆分时将数据按所用的拆分变量排序。但如果数据集很大,而所用的拆分变量已经排过序了,可选中该单选按钮以节省运行时间,但该功能其实较少用到。

按照如图 4.3 所示的设定对数据集进行拆分后,可以看到状态栏右侧会出现“拆分条件 time”的提示,表明按照变量 time 所做的拆分正在生效,此时如果进行 S3 年龄的统计描述,则看到的结果如图 4.4 所示。

月份		N	极小值	极大值	均值	标准差
200704	S3. 年龄	300	20	65	38.65	12.876
	有效的 N (列表状态)	300				
200712	S3. 年龄	304	20	64	38.54	13.028
	有效的 N (列表状态)	304				
200812	S3. 年龄	304	20	64	37.73	13.381
	有效的 N (列表状态)	304				
200912	S3. 年龄	239	18	59	28.96	8.599
	有效的 N (列表状态)	239				

图 4.4 描述统计量

显然数据已经按照不同月份进行了同一种分析,且结果被输出到了同一张表格中以便于进行比较。

需要指出的是,分割文件的设定一旦完成,就将在之后的分析中一直有效,而且会被存储在数据集中,直到再次进行设定为止。

4.1.3 选择个案

很多时候并不需要分析全部的数据,而是按要求分析其中的一部分,比如只分析 2009 年 12 月的数据,或者只对男性受访者的数据进行分析,这时就可以使用“选择个案”对话框来操作。

“选择个案”对话框主要由“选择”框组和“输出”框组构成,图 4.5 中图形右上侧的“选择”

单选按钮组用于确定个案的筛选方式。除默认的不做筛选(使用全部个案)外,还可以只分析满足条件的记录、从原数据中按某种条件抽样、基于时间或记录序号来选择记录,或者使用筛选指示变量来选择记录。

(1) 如果条件满足:此时将只分析满足所指定条件的记录,单击下方的“如果”按钮会打开“如果”对话框,用于定义筛选条件,该对话框几乎和图 3.2 所示的变量赋值过程对话框完全相同,因此不再重复解释。

(2) 随机个案样本:从原数据中按某种条件抽样,使用下方的“样本”按钮进行具体设定,可以按百分比抽取记录,或者精确设定从前若干个记录中抽取多少条记录。

(3) 基于时间或个案全距:基于时间或记录序号来选择记录,使用下方的“范围”按钮设定记录序号范围。

(4) 使用筛选器变量:此时需要在下面选入一个筛选指示变量,该变量取值为非 0 的记录将被选中,进行之后的分析。

“输出”单选按钮组则用于选择对没有选中的记录的处理方式,可以选择以下可选项之一来处理未选定个案。

(1) 过滤掉未选定的个案:未选定的个案不包括在分析中,但保留在数据集中,使用该选项则会在数据文件中生成名为 filter_ \$ 的变量,对于选定个案该变量的值为 1,对于未选定个案该变量的值为 0。而相应的未被选中的个案 ID 号处也会以反斜杠加以标记,如图 4.6 所示。

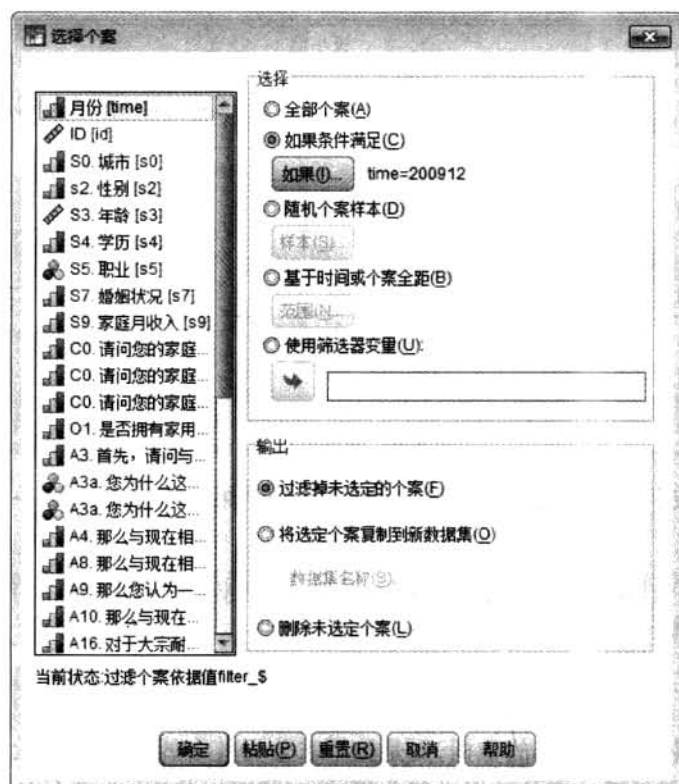


图 4.5 “选择个案”对话框

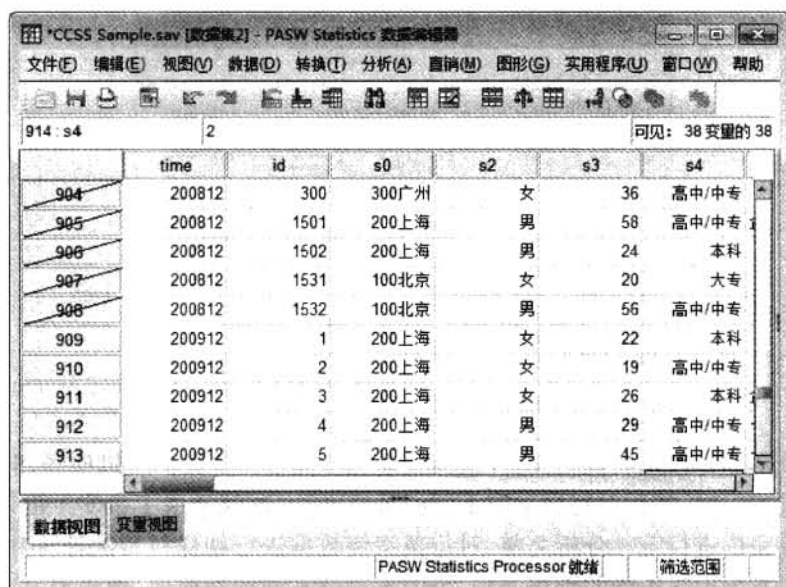


图 4.6 选择生效后的数据界面

(2) 将选定个案复制到新数据集:将选定的个案复制到新数据集时,原始数据集不会受到影响。未选中个案不包括在新数据集中,而在初始数据集中保持其初始状态。

(3) 删除未选定个案:直接从数据集中删除未选定个案。只有退出文件而不保存任何更改,然后重新打开文件,才能恢复删除的个案。如果保存了对数据文件的更改,则会永久删除个案。因此该选项一般不要使用,以免无法恢复造成损失。

当对数据集做出筛选后,可以看到状态栏右侧会出现“筛选范围”的提示,表明所做的筛选正在生效。和分割文件操作相类似,筛选功能将在之后的分析中一直有效,而且会被存储在数据集中,直到再次改变选择条件为止。

4.1.4 加权个案

加权个案会给不同个案赋以不同的权重,以改变个案在统计分析中的重要性。一般而言,在如下两种情形下需要进行该操作。

(1) 以频数格式录入的数据:在默认情况下,数据集中的每一行就是一条原始记录,这在多数情况下没有什么问题,但有时却非常麻烦。例如,图 4.7 所示的数据,如果每一行就是一条原始记录,需要输入 121 行!这时候一般使用频数格式录入数据,即相同取值的个案只录入一次,另加一个频数变量用于记录该数值共出现了多少次,这样就需要在分析时用“加权个案”对话框(图 4.8)将数据指定为频数格式。

	gender	group	count
1	1.00	1.00	34.00
2	2.00	1.00	23.00
3	1.00	2.00	45.00
4	2.00	2.00	19.00

图 4.7 以频数格式录入的数据

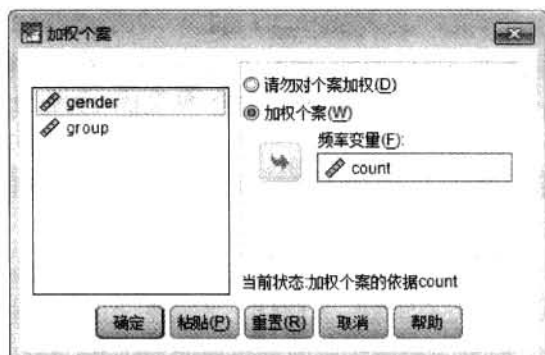


图 4.8 “加权个案”对话框

(2) 案例数据抽样权重的调整:统计抽样在理想情况下是等概率随机抽样,但许多时候是将整个总体拆分成若干层,然后对每层采取不同的抽样方法,这就造成了事实上的不等概率抽样,需要在数据采集完毕进行统计分析之前,对每条案例数据进行抽样权重的计算和调整。抽样权重可以理解成一系列因素影响的乘积,每一个因素对应某种抽样概率、覆盖率、应答率等方面的差异所导致的偏倚的调整。CCSS 项目数据就是如此,每月的原始数据采集完毕后,根据人口分布特征、应答率等因素进行权重计算都是重要的工作内容。但出于数据保密的需要,本书所附的 CCSS 案例数据中删除了相应的权重变量,这里特别说明一下。

对以上两种情形而言,具体的对话框操作是相同的,在对话框中有两个单选按钮,分别是“请勿对个案加权”和“加权个案”,如果选择后者,则需要选中一个加权变量。进行加权以后,SPSS 界面右下角会出现“加权范围”的字样,并且可以被存储到数据集中,直到取消加权;否则一直加权对数据进行处理。



目前在 SPSS 中权重变量可以为小数,但是有的过程会将小数权重简单地四舍五入为最接近的整数,某些过程甚至会完全忽略权重变量,具体情况需读者注意相应过程的文档说明。

一旦应用了一个权重变量,该权重变量始终保持有效,直到选择另一个权重变量或关闭加权。如果保存了加权后的数据文件,加权信息会随数据文件一起保存。可以随时关闭加权,即使在文件以加权形式保存之后也可以。

4.1.5 分类汇总

所谓分类汇总就是按指定的分类变量对个案进行分组,并按分组对变量求指定的描述统计量,结果可以存入新数据文件中,也可以替换当前数据文件。对数据文件进行分类汇总是实际工作中经常遇到的事情。例如,对于学生基本情况的数据,现希望了解不同性别学生的平均分数情况,这时就需要首先对数据按不同性别分类,然后再分别求出各类学生的分数平均值,这个过程本质上就是一个数据的分类汇总过程。

1. 界面简介

分类汇总过程涉及的对话框如图 4.9 所示。

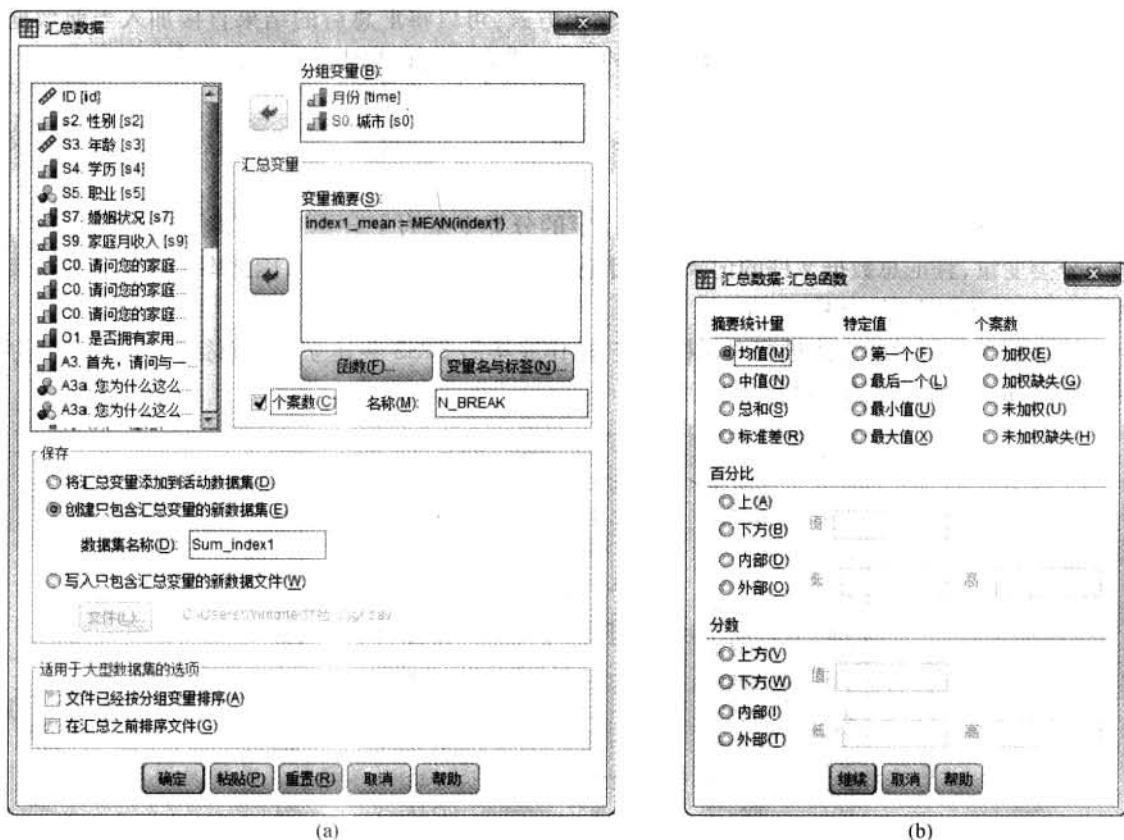


图 4.9 分类汇总过程涉及的对话框

在 SPSS 中,数据文件的分类汇总是经过以下 3 个步骤完成的。

(1) 指定分组变量(Break Variable)和汇总变量(Aggregate Variable)。

(2) SPSS 自动根据分类变量的取值将记录数据分成若干类,并对每类记录分别计算汇总变量的描述统计量。

(3) 将分类汇总的计算结果加以输出。

但是为了方便用户,整个操作过程都在一个统一的对话框中进行设定,具体如下:

(1) “分组变量”列表框:在图 4.9(a)中用于选择分组变量,可以有多个。

(2) “变量摘要”列表框:用于选择被汇总的变量,可以有多个,包括对同一个变量的多种不同汇总方式。

(3) “函数”按钮:单击该按钮会打开定义汇总函数的对话框,此处共提供了 5 组函数,分别为“摘要统计量”、“特定值”、“个案数”、“百分比”和“分数”(Fraction)。以最常用的第一组为例,可选的函数有“均值”、“中值”、“总和”、“标准差”4 种。SPSS 默认对各类记录分别计算汇总变量的均值,如图 4.9(b)所示。

(4) “变量名与标签”按钮:单击该按钮打开的对话框用于定义新产生的汇总变量的名称和标签。

(5) “个案数”复选框:用于定义一个新变量以存储同组的个案数。

(6) “保存”框组:设定汇总结果的具体输出方式,可以将汇总后的结果直接加入当前数据文件中,也可以定义一个新文件以存储汇总的结果,或者用汇总的结果替换原来的数据。

2. 实例分析

例 4.1 按 time 月份和 s0 城市对 CCSS 案例数据中的变量 index 进行均数汇总,并将结果输出到新数据文件 Sum_index1 中。

本例的分组变量不止一个,此时第一个指定的分类变量为主分类变量,其他的依次为第 2、第 3 分类变量,且汇总数据文件的记录数等于各分类变量类别数的乘积,因此本例的汇总数据文件中会有 $4 \times 3 = 12$ 条记录。

按照图 4.9 所示的方式进行对话框设定,操作完毕后 SPSS 会建立一个新数据文件,其中存储的就是相应的汇总结果,如图 4.10 所示。

	time	s0	index1_mean	N_BREAK
1	200704	100北京	100.05	100
2	200704	200上海	97.79	100
3	200704	300广州	97.16	100
4	200712	100北京	97.13	101
5	200712	200上海	92.10	101
6	200712	300广州	93.19	102
7	200812	100北京	91.97	102
8	200812	200上海	87.68	102
9	200812	300广州	91.70	100
10	200912	100北京	102.58	75
11	200912	200上海	102.56	84
12	200912	300广州	100.86	80

图 4.10 数据 Sum_index1 的内容

4.2 数据文件的重组与转置

数据文件的重新排列,是数据分析中经常用到的一个功能。数据录入的格式未必能一步到位地满足分析的要求,很多时候要根据分析的要求改变数据的排列格式,数据重组是一个图形化的向导,直观地实现了这一功能。

4.2.1 数据的长型与宽型格式

长型格式和宽型格式指的是重复测量数据的两种不同的排列方式,由于重复测量模型可以使用不同的统计模型加以分析,根据模型的要求进行长型格式和宽型格式之间的互转实际上是分析中经常会遇到的问题。

这里以 SPSS 的自带文件 anxiety.sav 和 anxiety2.sav 来说明这两种数据排列格式的特点。这两个文件记录的都是 12 名精神病患者在接受治疗后的 4 个时间点的精神状态评分,其中“科目”(该名称又是误译 Subject 的结果)为病人的 ID 号,“评分”(Score)为评估的分数,“跟踪”(Trial)为测量时的时间点编号,焦虑(Anxiety)和紧张(Tension)记录了病人在治疗前是否焦虑和

紧张,如图 4.11 所示。

anxiety. sav 文件是长型格式,以每次测量作为一条记录,用“科目”和“跟踪”来区分是哪位病人的第几次测量,“焦虑”和“紧张”作为携带变量在相同病人的记录中重复出现,这样 12 个病人共形成了 48 条记录;而 anxiety2. sav 是宽型格式,每位病人各有一条记录,4 次测量分别用测量 1~测量 4 这 4 个变量来分别记录,原先用于区分测量次数的变量测量不再需要,同一个病人的“跟踪”、“焦虑”和“紧张”也只出现一次。通过图 4.11 可以更好地理解这两种数据格式的特点。


	科目	焦虑	紧张	分数	跟踪
1	1	1	1	18	1
2	1	1	1	14	2
3	1	1	1	12	3
4	1	1	1	6	4
5	2	1	1	19	1
6	2	1	1	12	2
7	2	1	1	8	3
8	2	1	1	4	4
9	3	1	1	14	1
10	3	1	1	10	2

(a)

	科目	焦虑	紧张	跟踪1	跟踪2	跟踪3	跟踪4
1	1	1	1	18	14	12	6
2	2	1	1	19	12	8	4
3	3	1	1	14	10	6	2
4	4	1	2	16	12	10	4
5	5	1	2	12	8	6	2
6	6	1	2	18	10	5	1
7	7	2	1	16	10	8	4
8	8	2	1	18	8	4	1
9	9	2	1	16	12	6	2
10	10	2	2	19	16	10	8

(b)

图 4.11 数据集 anxiety. sav 和 anxiety2. sav 的内容

事实上,在学习了第 2 章后,大家应当能够明白长型格式才是符合统计分析要求的标准记录格式,但是由于重复测量数据会使用特殊的重复测量模型来进行分析,此时就需要将数据变换为宽型格式,该模型的详情参见本丛书高级教程的相关章节。

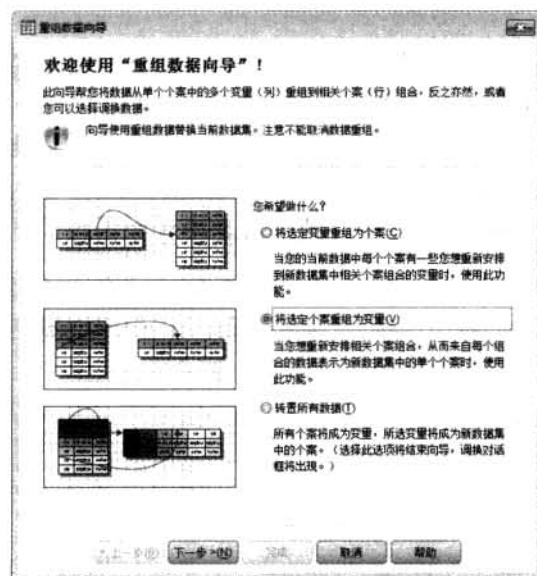
4.2.2 长型格式转换为宽型格式

本节将介绍如何使用数据重组向导实现数据结构的重建。

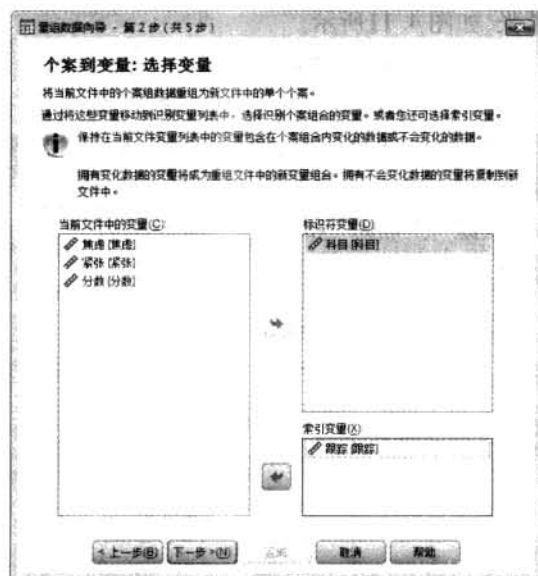
例 4.2 将 SPSS 自带文件 anxiety. sav 转换为 anxiety2. sav 的格式。

选择“数据”→“重组”菜单项,打开重组数据向导的第 1 个界面,如图 4.12 所示,从图 4.12 (a)可以看出,在向导中共提供了 3 种数据重组功能,分别是长型格式转换为宽型格式,宽型格式转换为长型格式,以及行列转置。根据要求,在这个例子中要使用的是第 2 种功能,选中“将选定个案重组为变量”单选按钮,单击“下一步”按钮后打开向导的第 2 个界面,如图 4.12 (b) 所示。

根据要求可知,用户需要指定被重复测量个体的 ID 标识变量和用于反映测量次别的 Index 变量,此处分别为“科目”和“跟踪”,将它们分别选入相应变量列表框中后单击“下一步”按钮,打开向导的第 3 个界面,如图 4.13 所示,此时“完成”按钮已经可用,也就是说后续界面的选项都有默认值填充,可以直接运行相应的过程了。如果希望更改,则会对是否需要排序、重组后数据文件的结构、给出产生一条新记录的原记录的数目、是否需要标识变量等进行设定,最后单击“完成”按钮,就可以得到相应的转换后的数据集,将该结果与数据文件 anxiety2. sav 的内容进行比较,可以看出除变量名和标签不同外,两个文件的内容实际上是一致的。另外,也可以看一下系统在结果窗口中的汇总输出,这常被用来检查操作是否有误。

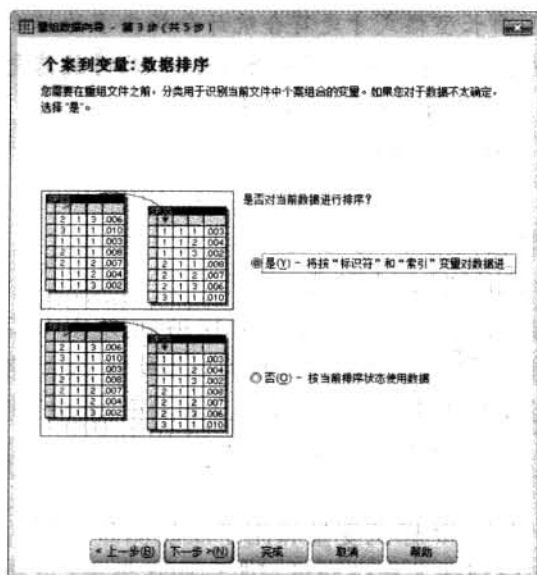


(a)

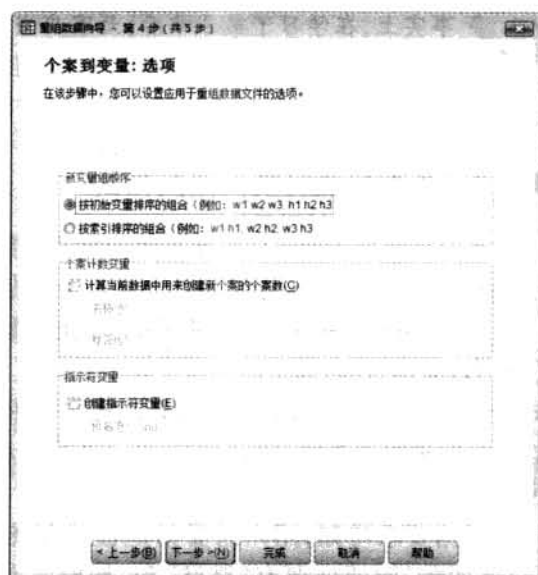


(b)

图 4.12 数据重组向导的第1、2个界面



(a)



(b)

图 4.13 数据重组向导的第3、4个界面



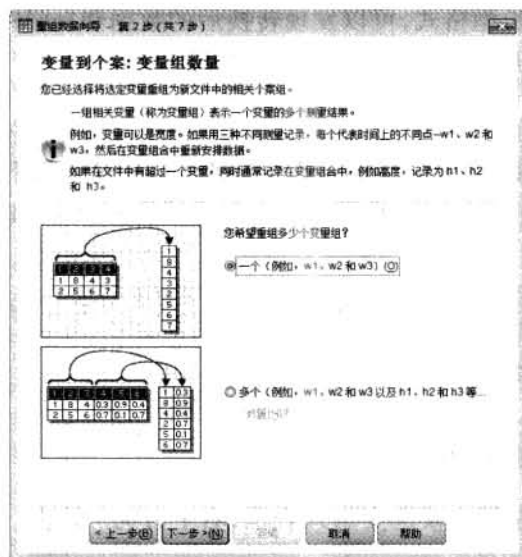
这里有一个非常有趣的问题:本例中没有说明哪个变量需要转换,但最后程序只将“分数”转换为了宽型格式,“焦虑”和“紧张”变量则保持不变,未加转换。这是因为程序会自动扫描需要转换的变量,如果该变量在相同个体内取值均恒定,则会被自动携带过来而不加转换,本例中的“焦虑”和“紧张”变量正属于这种情况。显然,SPSS的这种设计大

大方便了用户的使用。

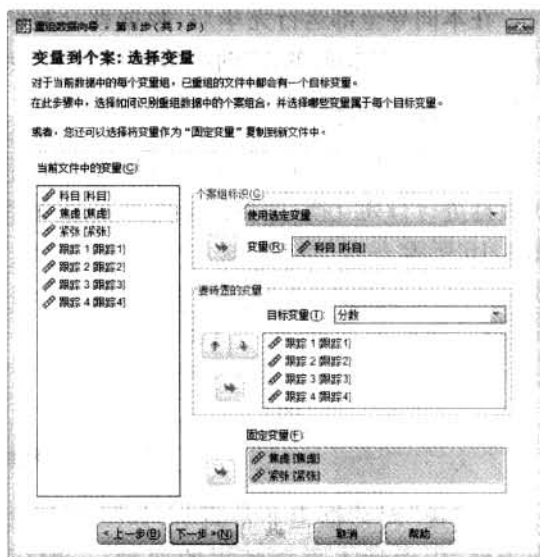
不过,这种自动化识别虽然给用户提供了便利,但也有可能犯错。例如,如果在本例中实际上每次跟踪都测量了一次“焦虑程度”,只是“焦虑”的值每次正好都相同,这时自动识别系统显然就无法正确识别数据格式。

4.2.3 宽型格式转换为长型格式

下面说明如何将宽型格式的数据转换为长型格式,有了前面的基础,这一部分内容应当很容易理解了。假设此处的任务是将 anxiety2. sav 转换为如 anxiety. sav 的长型格式,则在第 1 个向导界面上选择第 1 项,单击“下一步”按钮后打开如图 4.14(a)所示的对话框,询问共有几组重复测量变量需要转换,此处只有一个,单击“下一步”按钮后打开最重要的“选择变量”对话框(图 4.14(b)):“个案组标识”框组用于设定重复测量个体的 ID 标识变量,此处设定为变量“科目”;中部的“要转置的变量”框组则用于设定被转换的变量组,首先将变量组名称改为分数,随后在下方的列表框中将跟踪 1~跟踪 4 选入。如果有多组变量需要转换,则依次设定即可;最下方的“固定变量”列表框则用于选入数值恒定的固定变量,此处为“焦虑”和“紧张”变量。



(a)



(b)

图 4.14 数据重组向导的第 2、3 个界面

在正确设定了变量之后,下面的操作就非常简单了,随后的创建索引界面(如图 4.15(a)所示)用于设定重复测量指示变量(如同本例中的变量测量),而创建索引变量界面(如图 4.15(b)所示)则用于具体设定该变量的数值,此处将索引变量命名为“跟踪”。现在就可以直接单击“完成”按钮结束本向导了,如果希望进行更详细的设定,则最后还有两个界面用于选择缺失值、未选中变量的处理方式,以及是直接执行,还是生成相应的程序。

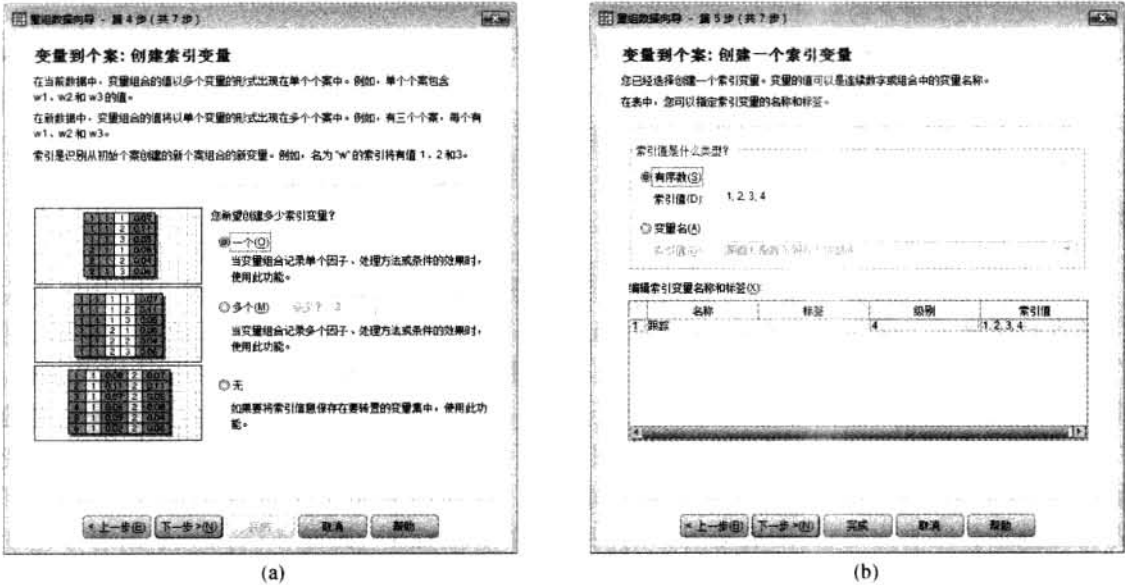


图 4.15 数据重组向导的第 4、5 个界面

在本向导全部运行完毕后,数据就会被转换成长型格式,可以发现转换后的数据和 anxiety.sav 基本上是相同的,同时结果窗口中也会给出相应操作的汇总表格用于查错。

4.2.4 数据转置

下面介绍转置 (Transpose) 过程,这实际上也是数据重构向导的第 3 个功能。转置过程用于对数据进行行列互换,即将记录转为变量,将变量转为记录后,重新显示在数据编辑窗口中,如图 4.16 所示。

	varname	v1	v2
1	A	1.00	2.00
2	B	3.00	4.00

(a)

	CASE_LBL	A	B
1	v1	1.00	3.00
2	v2	2.00	4.00

(b)

图 4.16 数据转置前后的数据文件

“转置”对话框也非常简单,如图 4.17 所示,左侧为候选变量列表框;右上方为“变量”列表框,用于选入需要转置的变量,一般应选入除名称变量外的所有其他变量,如果有变量未选入,则转置时会被自动丢弃;右下方为“名称变量”列表框,用于指定原数据文件中记录转置后变量名的字符变量,但不是必需的,此时系统会将新变量自动按 var001, var002, … 的顺序命名。

对统计分析的初学者而言,可能无法想象这个功能有什么用处。实际上,数据转置主要是用于编程进行矩

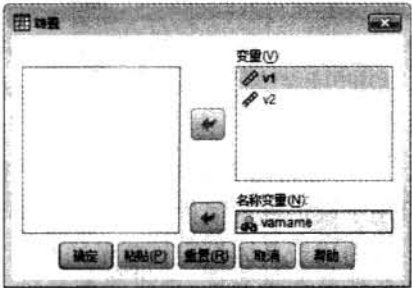


图 4.17 “转置”对话框

阵运算时的矩阵转置操作的,对于只需要调用现成的分析程序、不需要自行编写算法的用户而言,转置功能的确没有多少实际用途。

4.3 多个数据文件的合并

4.3.1 一些基本概念

1. 纵向拼接和横向合并

进行统计分析的第一步工作就是将待分析的数据录入到 SPSS 中。在数据量较大时,经常需要分别由不同的录入人员进行录入,这样就会出现一份大数据被分别存储在几个不同的数据文件中的现象。除此以外,如果数据有多个来源,则可能会使变量被分散存储在几个文件中,需要按照某种规则加以合并后才能进行分析。因此,将这若干个小的数据文件合并成一个大的数据文件是进行数据分析的前提。

SPSS 数据文件的合并方式有两种:纵向拼接和横向合并,分别对应上述的两种情况。

(1) 纵向拼接:指的是几个数据集中的数据纵向相连,组成一个新的数据集,新数据集中的记录数是原来几个数据集中记录数的总和。其实质就是将两个数据文件的变量列按照各个变量名的含义一一对应进行首尾连接。

(2) 横向合并:指的是按照记录的次序,或者某个关键变量的数值,将不同数据集中的不同变量拼接为一个数据集,新数据集中的变量数是所有原数据集中不重名变量的总和。横向合并的实质就是将两个数据文件的记录按照某种对应关系一一进行左右对接。

2. 案例文件解释

这里使用 3 个简单的案例文件来演示相关的合并操作:文件 a. sav 包括了 id 号为偶数的 5 位受访者的性别、年龄和身高,而 b. sav 则包括 id 为奇数的 4 位受访者的性别、年龄身高和体重, c. sav 则提供了 4 位受访者的体重(如图 4.18 所示)。注意在这 3 个文件中,相同的变量可能采用不同的变量名称。

	id	sex	age	height
1	2	1	16	158
2	4	1	34	164
3	6	2	56	170
4	8	1	68	172
5	10	2	25	178

(a)

	id	sex	age	h	w
1	1	2	19	165	53
2	3	1	30	175	70
3	5	2	28	162	48
4	7	1	44	169	68

(b)

	id	weight
1	2	46
2	6	92
3	8	51
4	12	70

(c)

图 4.18 数据文件 a. sav、b. sav 和 c. sav 内容示意

4.3.2 数据文件的纵向拼接

例 4.3 将数据 b. sav 中的记录添加到 a. sav 中,注意 b. sav 中的变量 h 对应 a. sav 中的 height。

在数据窗口中分别打开数据文件 a. sav 和 b. sav, 然后选择“数据”→“合并文件”→“添加个案”菜单项, 并在第 1 个对话框中选择待合并的文件 b. sav, 则打开如图 4.19(a) 所示的对话框。

(1) “非成对变量”列表框: 该列表框中的变量名后面都跟有 * 或 + 号, * 表示该变量名是当前活动数据集中的变量, + 表示该变量名是外部待合并数据文件中的变量。在默认情况下, 如果一个变量名没有两个文件中同时出现, 则 SPSS 认为这些变量不是待合并的两个文件所共有的, 无法被自动对应匹配, 因此它们不会自动成为合并后新数据文件中的变量。

(2) “新的活动数据集中的变量”列表框: 在该列表框中, 两个待合并的数据文件中共有的变量名会被自动对应匹配, 并出现在本变量列表框中。SPSS 默认它们具有相同的数据含义, 自动成为合并后新数据文件中的变量。如果需要修改默认设置, 可以将它们剔除到“非成对变量”列表框中。

(3) 强行配对: 在本例中显然 h 和 height 应当是同一个变量, 因此可以将其同时选中, 然后单击中部的“对”按钮强行配对, 表示它们具有相同的数据含义, 从而将其选入新数据集变量列表框中。此时新变量默认会按照当前文件中相应变量的名称来设定。

(4) “重命名”按钮: 如果希望新数据集中的变量名与先前不同, 则可以先单击“重命名”按钮改名后再选入。

(5) “将个案源表示为变量”: 如果希望在合并后的数据文件中看出哪些记录来自合并前的哪个 SPSS 数据文件, 可以选中该复选框, 此时合并后的数据文件中将自动出现名为“源 01”的变量, 取值为 0 或 1。0 表示该记录来自第 1 个数据文件, 1 表示该记录来自第 2 个数据文件。

按图 4.19(b) 设定完毕后, 生成的新数据集将有 9 条个案, 如图 4.18 所示。



图 4.19 文件纵向拼接涉及的对话框

4.3.3 数据文件的横向合并

数据文件的横向合并由于较为复杂, 因此应遵循如下 3 个条件。

(1) 如果不是按照记录号对应的规则进行合并的, 则两个数据文件必须至少有一个变量名

相同的关键变量,该变量是数据文件横向对应拼接的依据,如学号、贵宾卡号等,关键变量可以是多个,且关键变量的取值在不同个案间最好具有唯一性。

(2) 如果是使用关键变量进行合并的,则两个数据文件都必须事先按关键变量进行升序排列,否则系统将报错。

(3) 为了方便 SPSS 数据文件的合并,在不同的数据文件中,数据含义不同的列,变量名尽量不要相同。

例 4.4 将数据 c. sav 中的变量添加到 a. sav 中,并尽量保留数据。

在数据窗口中分别打开数据文件 a. sav 和 c. sav,然后选择“数据”→“合并文件”→“添加变量”菜单项,并在打开的对话框中选择待合并的文件 c. sav,则打开如图 4.20 所示的对话框。

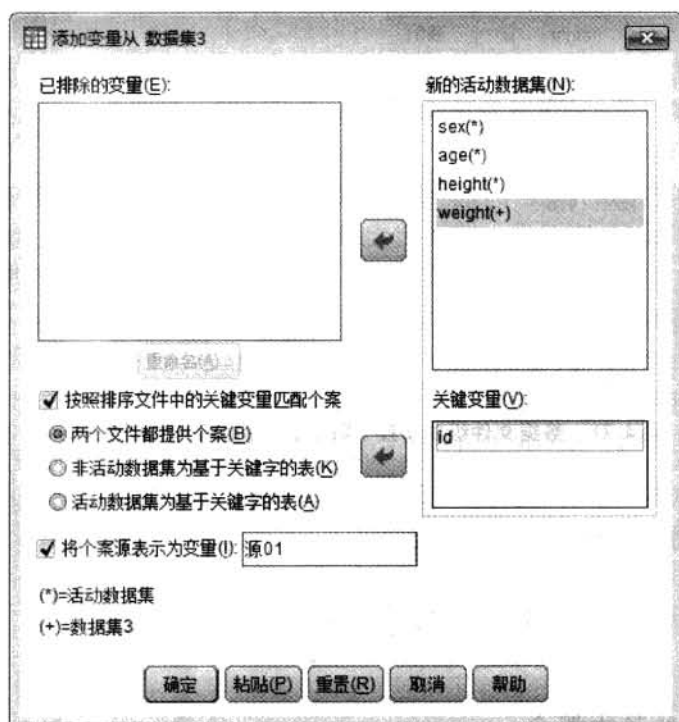


图 4.20 文件横向合并的对话框

(1) “新的活动数据集”列表框:该列表框中的变量名后面都跟有 * 或 + 号,* 表示该变量名是当前活动数据集中的变量,+ 表示该变量名是外部待合并数据文件中的变量。在默认情况下,如果变量名没有两个数据集中同时出现,则 SPSS 会自动将其列入新数据文件的变量列表中。

(2) “已排除的变量”列表框:与当前数据集变量同名的外部数据集变量,为了避免重复而被列在此处。

(3) “关键变量”列表框:如果两个待合并的数据文件中的记录数据是按照记录号横向一一对应的,则可直接确认完成合并工作,否则必然按照关键变量进行匹配,此时外部文件中的关键变量名必然因重名而出现在“已排除的变量”列表框中。将其选中,并选入“关键变量”列表框中,本例中为变量 id。

(4) 按照排序文件中的关键变量匹配个案:对于合并后文件中的数据按哪种方式提供,SPSS中有3个选项可供选择,默认是数据由原来的两个数据文件共同提供以尽量保留信息,这也是最常用的选项;后两个选项的翻译不易理解,简单地说第2个选项(External file is keyed table)指在合并时保留当前文件的所有数据,但丢弃只在外部数据文件中才有的个案,当外部数据根据关键变量是无重复记录,而当前数据根据关键变量是有重复记录时使用此选项;第3个选项(Working data file is keyed table)指在合并时保留外部文件的所有数据,但丢弃只在当前文件中才有的个案,如果当前数据根据关键变量是无重复记录,而外部数据根据关键变量是有重复记录时使用此选项。

其余对话框元素前面已经出现过,不再重复解释。操作成功后可以看到数据集的格式,如图4.21所示。

	id	sex	age	height	w	源01
1	2	1	16	158		0
2	4	1	34	164		0
3	6	2	56	170		0
4	8	1	68	172		0
5	10	2	25	178		0
6	1	2	19	165	53	1
7	3	1	30	175	70	1
8	5	2	28	162	48	1
9	7	1	44	169	68	1

(a)

	id	sex	age	height	weight	源01
1	2	1	16	158	46	1
2	4	1	34	164		0
3	6	2	56	170	92	1
4	8	1	68	172	51	1
5	10	2	25	178		0
6	12				70	1

(b)

图 4.21 数据文件纵向拼接、横向合并完成后的数据文件示意

最后再次提醒大家,使用关键变量进行横向合并前,数据文件必须要按照关键变量排序,否则相应的合并操作将会失败。

4.4 与数据字典有关的功能

4.4.1 数据字典的基本概念

在大型的数据分析项目中,数据管理是非常重要的一个环节,为了保证工作质量,数据处理人员往往会事先定义好一个非常详细的数据格式,包括变量格式、变量标签、值标签、缺失值定义等,将其称为数据字典,它将成为使用者定义具体数据文件格式的标准模板。在SPSS中,数据字典其实就是一个数据文件,它可以是一个只有结构没有数据的空数据文件,也可以是有预实验数据存储在外的一个实际数据文件,但无论如何,对其都只限于使用其中的数据结构定义。

SPSS 19 版中共提供了3个与数据字典相关的对话框,专门用于定义数据字典,或者将预定义的数据字典直接引入当前数据文件中。在20版中又进一步新增了设置未知测量级别对话框,对于大型或者连续性的数据分析项目而言,这些都是非常有用的功能,可以大大减轻数据处理人员的工作负担。下面具体总结一下如何应用这些向导对话框来完成数据管理任务。

(1) 如果有事先定义的数据字典格式,则可以先生成一个没有记录的空数据文件,将全部的

数据字典设定好,将来在数据录入完毕后使用复制文件属性向导即可套用字典。

(2) 如果没有事先定义的数据字典格式,则可以在录入工作进行了一段时间以后先使用变量属性定义向导完成数据字典的设定工作,然后随着录入工作的进行扫描数据的情况,及时更新字典,最后在录入工作完毕后使用复制文件属性向导应用字典的最终版本。

(3) 如果数据管理任务不太复杂,也可以直接在数据字典中录入数据,或者直接在变量视图中修改属性,或者直接在 SPSS 中录入/导入数据,然后利用设置未知测量级别向导来快速设定数据字典。但是在真正的大型数据管理项目中,单独建立和维护数据字典是非常关键的一环,建议有志于从事大型数据管理与分析工作的读者尽早养成这一良好的维护习惯。

4.4.2 定义变量属性

定义变量属性(Define Variable Property)指的是对于数据集中已存在的变量进一步定义其属性。具体说来,可以列出所选变量的所有取值;分辨没有值标签的值,并且提供自动给出值标签的功能;可以将另一个变量的属性复制给所选的变量,也可以将所选变量的属性复制给其他变量。从表面上看,该向导的绝大多数功能都可以在变量视图中实现,似乎有些多余,但对于复杂的数据管理项目而言,首先,它的可视化能力可以大大提高工作效率;其次,对初学者而言,使用该向导进行变量的设置也是非常好的选择。

这里仍以 CCSS 案例数据为例对该向导加以说明。选择“数据”→“定义变量属性”菜单项,则首先打开预定义对话框,要求选择希望进行设定的变量,可以选择多个,单击“继续”按钮后 SPSS 将会对选入的变量都进行扫描,随后进入向导的主界面,如图 4.22 所示。

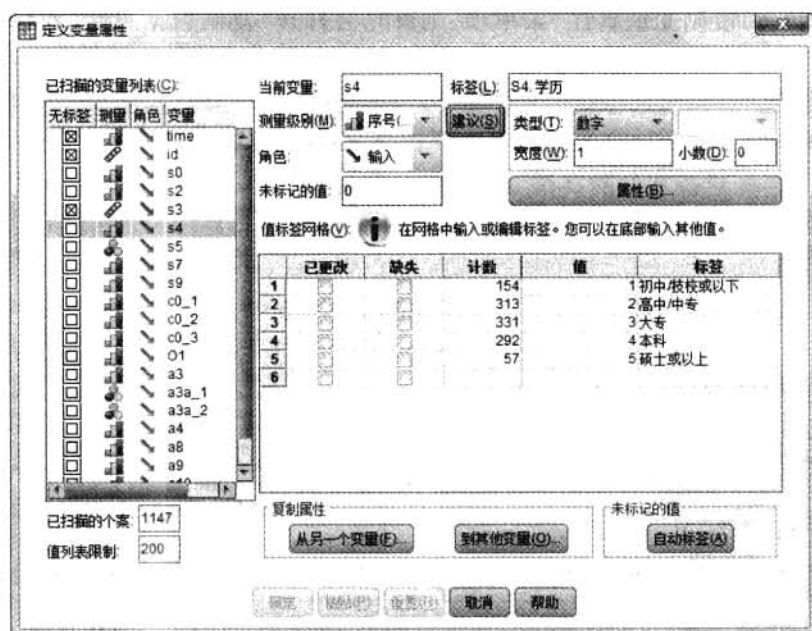


图 4.22 “定义变量属性”对话框

(1) 左侧变量列表:其中会列出所有被选择/扫描的变量,具体又分为 4 列显示,分别表示该变量有无标签定义、测量尺度、角色和变量名。选中相应的变量名称,则右侧会显示出相应的设

定,并且允许用户更改。

(2) 右上侧变量属性组:用于设定测量尺度、存储格式、变量名标签等,如果单击“建议”按钮,则系统会根据扫描到的数据给出建议的测量尺度;注意其中的“属性”按钮用于新建自定义属性,详见4.5.3小节介绍。

(3) 中部值标签网格:用于列出该变量所有取值的频数、当前值标签和缺失值设定等,双击后即可更改标签和缺失值设定(将当前数值设定为自定义缺失值)。

(4) 下部“复制属性”框组:用于将另一个被扫描变量的属性复制给所选的变量,也可以将所选变量的属性复制给其他被扫描变量,该框组实际上已经在可视化离散过程中介绍过了。

(5) “自动标签”按钮:用于自动生成值标签,实际上就是将所有的变量值均赋给空白值标签。

图4.21显示了S4的属性定义情况,由此可以看出在这一个界面中就可以完成对变量的所有属性定义,而且可以一次性定义多个变量,并且由系统帮助扫描出全部取值范围,这显然要比在变量视图中进行操作容易得多,可以大大方便数据字典的定义工作。

4.4.3 复制数据属性

复制数据属性(Copy Data Property)过程用于将定义好的数据字典直接应用到当前文件中,操作时不仅可以将一个外部数据文件的相关属性复制到当前数据文件中,还可以进行自定义,只选择某些变量,或者某些属性进行复制,这无疑大大提高了连续性项目对原有资源的利用程度。对于一些特殊的文件属性,如多选题变量集、普通变量集、权重变量设定等,使用该向导进行复制还会减少许多重复工作。

选择“数据”→“复制变量属性”菜单项,则首先会打开“复制数据属性”对话框,如图4.23所示。

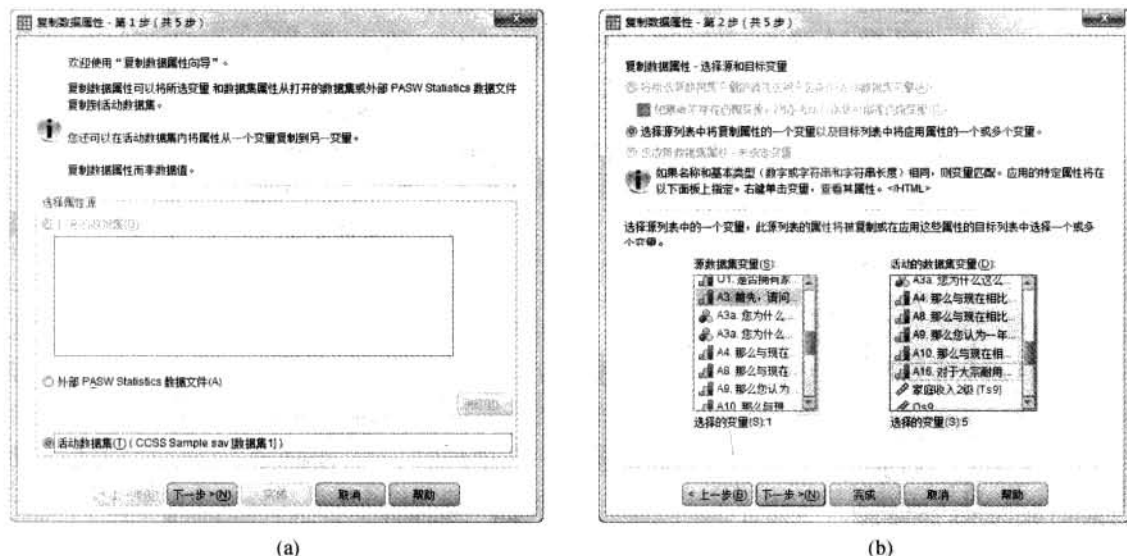


图 4.23 “复制数据属性”对话框

(1) 第1个对话框:在该对话框中可以选择希望复制的数据属性来源,可以是已打开的另一个数据文件,也可以是未打开的另一个数据文件,还可以是当前文件本身。这里选择了后者,然

后单击“下一步”按钮。

(2) 第2个对话框:第2个对话框用于设定希望复制的属性种类,有3种选择,分别为选择同名、同类型、同长度变量的属性进行复制(Apply properties from selected source file variables to matching working file),选择一个变量的属性进行复制(Apply properties from a single source variable to selected working file variable)和仅复制文件属性——多选题集定义、权重设定等(Apply dataset properties only—no variable selection)。在图4.23中选择的是第2项,具体设定是将变量A3的属性定义复制到A4~A16这5个变量上去,单击“下一步”按钮。

(3) 第3个对话框:该对话框要求详细指定希望复制的变量属性,共有7种,并且可以选择是替换原有属性,还是和原属性进行合并。

(4) 在第3个对话框出现时,用户其实就可以单击“完成”按钮结束向导了,第4、5个界面分别用于选择希望复制的文件属性,以及是否生成相应的SPSS程序。运行完毕后就会看到,A4~A16的变量属性全部更改成了与A3相同的设定。

4.4.4 新建自定义属性和设置未知测量属性

在默认情况下,SPSS将为每个变量设定名称、类型等共11个属性,这在绝大多数情况下都是足够的,但是在一些大型数据管理项目中,可能需要用户自行设定一些特殊的变量属性,例如,可以创建识别调查问题类型的变量属性(单选、多选、开放)或存储计算变量使用的公式等。和标准变量属性一样,这些定制变量属性也将随数据文件一同保存。

在SPSS中建立自定义属性的操作非常简单,首先将数据窗口切换到变量视图,然后选择“数据”→“新建设定属性”(Create Custom Attribute)菜单项,打开的对话框如图4.23所示。在“属性名称”文本框中输入希望建立的属性名称,然后在“属性值”文本框中输入希望默认设定的属性值,并将希望进行默认属性值设定的变量选入右上方的“选择的变量”列表框中,例如,图4.24中选入的是index1,单击“确定”按钮以后会看到变量视图最右侧新增一列“基准值”,其中index1对应的一行已经填充了数值100,用户可以单击单元格右侧的省略号来增删该列的可能取值列表,对于空单元格也可以双击后直接输入信息。

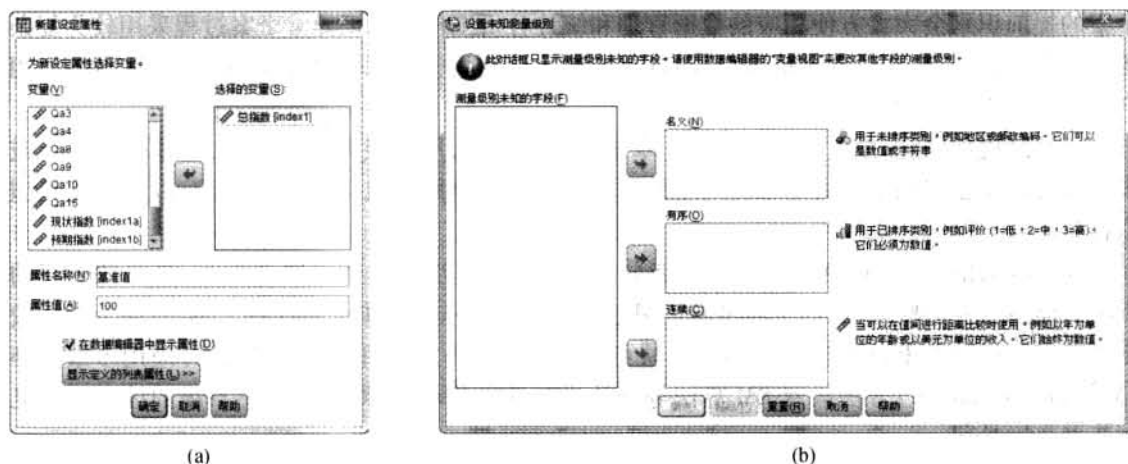


图 4.24 “新建设定属性”对话框和“设置未知测量级别”对话框

在如图 4.24 所示的对话框中如果单击下方的“显示定义的列表属性”按钮,则可以列出所有已经定义好的属性,事实上,SPSS 目前提供的这一功能更类似于变量注解,以和“实用程序”菜单中的“数据文件注释”功能相对应。

4.5 与数据准备有关的功能

4.5.1 SPSS 中与数据准备相关的功能

在进行正式的统计分析前,将数据集中的数据错误清理完毕,并按分析需求转换其格式是必要前提,准备分析数据是分析项目中最重要的一步,从传统上来说也是最耗时的步骤之一。而在大型的数据分析项目中,往往会涉及数据的多点录入、合并、重复更新、修改等操作,使得数据清理和数据准备工作更加复杂。在较早的 SPSS 版本中,这些操作都需要采用程序和函数方式来实现,而随着 SPSS 产品定位的下移,如何能快速、便捷地协助用户做好数据准备工作就成为 SPSS 考虑解决的问题之一,而最终所新增的一系列和数据准备相关的自动/半自动模块就是为用户提供相应的解决方案。

目前 SPSS 中,与数据准备有关的功能主要有如下几个。

(1) 数据验证模块:便于用户自行定义数据验证规则,并运行这些规则对数据进行检查,以标识无效个案、变量和数据值。当找到无效数据时,可以进一步分析原因并加以更正,从而实现原先必须使用 DATAENTRY 等产品才能实现的数据核查功能。

(2) 自动数据准备过程:该过程在第 3 章已经做过简单介绍,可以针对所设定的建模需求自动分析数据,对其中的异常值进行识别修正,筛选出存在问题或可能无用的字段,并在适当的情况下派生新的变量,并通过智能筛选技术改进性能。

(3) 标识重复个案过程:将相应的程序功能整合到一个对话框中,用户只需通过简单的菜单操作,就可以迅速地发现变量值重复的记录。

(4) 标识异常个案过程:在数据建模中出现异常数值往往是非常令人头痛的问题,因此异常个案的提前识别会大大方便相应的数据管理和统计分析和统计工作。标识异常个案过程采用较为复杂的统计算法,可以在探索性数据分析步骤中,快速检测到用于数据审核的异常个案,从而协助用户提前对其进行处理。

(5) 最优离散化过程:该过程在第 3 章也已做过介绍,用于根据建模目的,将原有的一个或多个连续性变量按照该分类变量类间差异最大化的优化原则离散化为分类变量,以最优化模型拟合效果。

(6) 缺失值分析(MVA)模块:在功能分类上,MVA 模块也和数据准备有关,具有缺失值的个案可能会引发一些问题,因为典型的建模过程会简单地从分析中丢弃这些个案,在缺失数据较多的情况下就会严重损失信息。MVA 模块可以帮助用户确定数值的缺失模式,并应用多重插补(Multiple Imputation)方法或者 EM 算法等方法进行缺失值填充,对该模块感兴趣的读者可参见本丛书高级教程。

4.5.2 数据验证模块

数据验证模块用于实现数据核查功能,用户通过自行定义数据验证规则,并运行这些规则对数据进行检查,以确定个案取值是否有效。验证规则主要有以下两种。

(1) 单变量规则:单变量规则包含一组应用于单个变量的数值检查规则,例如,范围外值的检查。对于单变量规则,有效值可以表示为一个范围,也可以表示为一个有效值列表。

(2) 交叉变量规则:交叉变量规则是用户定义的涉及多个变量间逻辑关系的规则,由标记无效值的逻辑表达式定义,可以应用于单个变量,也可以应用于变量组合。

在验证规则验证完毕后,用户可以将其保存在数据文件的数据字典中,这样指定一次规则后就可以反复使用。

CCSS 案例数据在问卷设定上有许多规则,例如,有下列几种情形。

(1) 年龄 S3:取值应当在 18 ~ 65 岁之间。

(2) 性别 S2:只有 1、2 两种取值编码。

(3) 关键题目取值逻辑:A3、A4、A8 不应当同时选择 9;否则应作为废卷处理。

上述 3 种情形正好分别对应了取值范围、取值列表和交叉规则这 3 种情形,下面就以它们为例介绍如何在数据验证模块中定义这些验证规则。

1. 定义验证规则

选择“数据”→“验证”→“定义规则”菜单项,打开“定义验证规则”对话框,如图 4.25 所示,可见“单变量规则”和“交叉变量规则”分别作为一个选项卡出现,首先在“单变量规则”选项卡中对 S3 和 S2 的规则进行定义,具体操作如下。



图 4.25 “定义验证规则”对话框

(1) S3:将规则定义名称设定为“RuleS3”,类型为默认的“数字”,有效值核查方式为默认的“在范围内”,最小值和最大值分别设定为 18 和 65。

(2) S2:将规则定义名称设定为“RuleS2”,类型为默认的“数字”,有效值核查方式更改为“在列表中”,然后在下方的“值”列表中依次添加 1、2 作为有效值。

下面说明如何设定交叉规则,切换到“交叉变量规则”选项卡,然后进行如下操作。

(1) 为交叉规则设定适当的名称,此处仍采用默认的 CrossVarRule1。

(2) 在下方的“逻辑表达式”文本框中输入使得个案无效的条件表达式“ $A3 = 9 \ \& \ A4 = 9 \ \& \ A8 = 9$ ”。

操作界面下方的软键盘、变量列表、函数列表等和第3章介绍过的“计算变量”过程对话框完全相同,这里不再重复解释。完成上述操作后,只要单击“确定”按钮,就可以生成相应的规则,并且这些规则可以直接保存在数据集中,便于重复使用。

2. 进行数据验证

在规则定义完毕后,下一步自然是使用这些规则来进行数据验证。选择“数据”→“验证”→“验证数据”菜单项,则会打开“验证数据”对话框,如图 4.26 所示,有了上面的基础,相应的操作就非常容易理解了。这里简单介绍该对话框的设定。

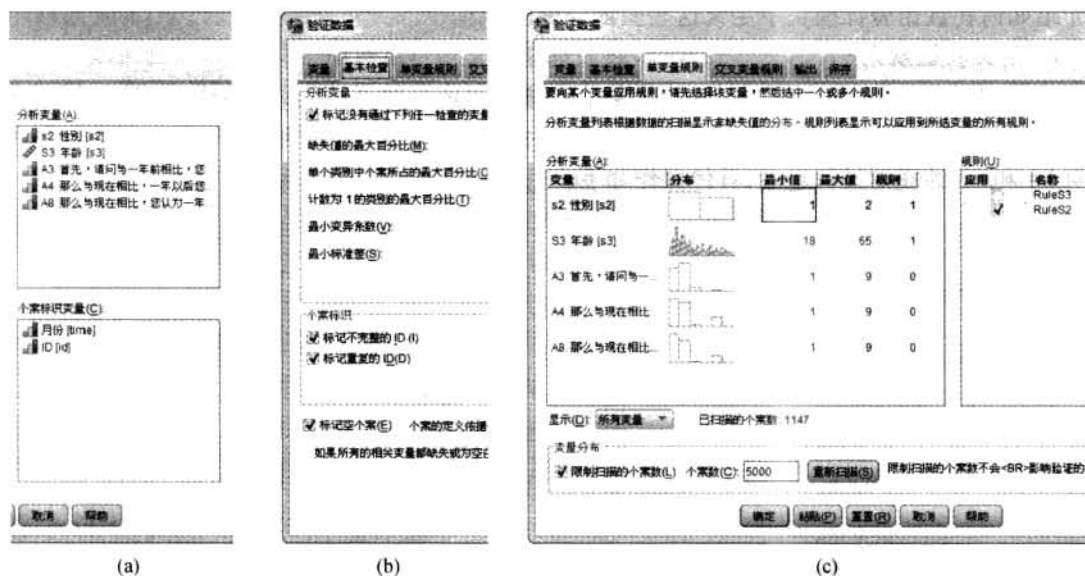


图 4.26 “验证数据”对话框

(1) “变量”选项卡:用于选入分析变量和表示个案的 ID 变量,为了节省核查时间,建议分析变量中只选入确实需要核查的变量,如本例中就只选入所涉及的 5 个变量。在本例中需要联合使用月份和 ID 才能唯一标识个案,因此在标识变量中需要将两者均选入,否则在结果中会报告个案标识符重复的错误。

(2) “基本检查”选项卡:进行数据核查时会对所有入选变量/个案进行分析,并报告明显表现异常的变量/个案,本选项卡用于对变量/个案的核查/报告标准进行设定,但一般使用默认值即可,无须更改。

(3) “单变量规则”选项卡:本选项卡用于将前面定义的单变量规则应用到具体变量上,左侧会列出所有分析变量,右侧使用复选框列表将定义好的规则和变量对应起来。本例应当在 S2 变量处选择 RuleS2,在 S3 处选择 RuleS3。此外如果发现规则还不完善,还可以单击右下方的“定义规则”按钮新增或者修改单变量规则。

(4) “交叉变量规则”选项卡:以复选框列表的形式列出所有的交叉规则,使用时将希望应用的规则选中即可。同样,如果发现规则还不完善,则可以单击右下方的“定义规则”按钮新增或者修改交叉变量规则。

(5) “输出”选项卡:设定数据核查在结果窗口中的错误报告输出形式。

(6) “保存”选项卡:可以将数据核查的情况以标记变量的形式保存在数据集中,以便直接对应原始案例进行修改,这些标记变量所反映的问题包括空变量、ID 变量异常、验证违规总数等。

在设定完毕后单击“确定”按钮,SPSS 就会按照要求对数据进行核查,并在结果窗口中提交相应的报告,因输出内容非常简单易懂,对此感兴趣的读者可以自行修改原始数据以观察核查结果,这里不再详述。

3. 加载预定义规则

为了方便用户使用,SPSS 默认在 Predefined Validation Rules. sav 文件中设定了一些常用的单变量规则,如非负整数、月份、星期等,用户只需要选择“数据”→“验证”→“加载预定义规则”菜单项即可将其载入加以使用。当然,对于自己常用的规则,用户也完全可以将其保存在该文件中形成自己的规则库,以方便自己使用。



实际上,用户也可以将自定义的规则存储为任何一个 SPSS 数据文件,然后在随后的工作中进行加载,但这种操作只能在程序级别实现,详见第 6 章的案例,加载预定义规则的对话框目前只能指定加载默认路径下 Predefined Validation Rules. sav 文件中的规则。

4.5.3 标识重复个案

标识重复个案(Identifying Duplicate Cases)的相应功能被整合到一个对话框中实现,只需通过简单的菜单操作,用户就可以迅速地发现个别变量值重复,或者所有数值完全重复的记录。

例 4.5 将 CCSS 案例数据第 2、4 条个案的 ID 变量值更改为 1,然后按照 time、id 均相同的标准查找重复记录。

由于 CCSS 案例数据是清理干净的文件,因此这里人工构造了重复个案用于演示。选择“数据”→“标识重复个案”菜单项,打开如图 4.27 所示的对话框。

(1) “定义匹配个案的依据”:用于确认重复个案的变量列表。如果某个个案的所有这些变量值与另一个个案的均相同,则将其视为重复个案。

(2) “在匹配组内的排序标准”:对于所发现的重复个案,将按照所选择的变量值对个案排序。

(3) “基本个案指示符”(Indicator):对于重复个案,可以指定其中一个为主个案,其余为多余的“重复”个案。可以将第一个或者最后一个个案设为主个案,主个案标识变量取值为 1,对于重复个案组中其余的非主要重复个案该变量取值为 0。

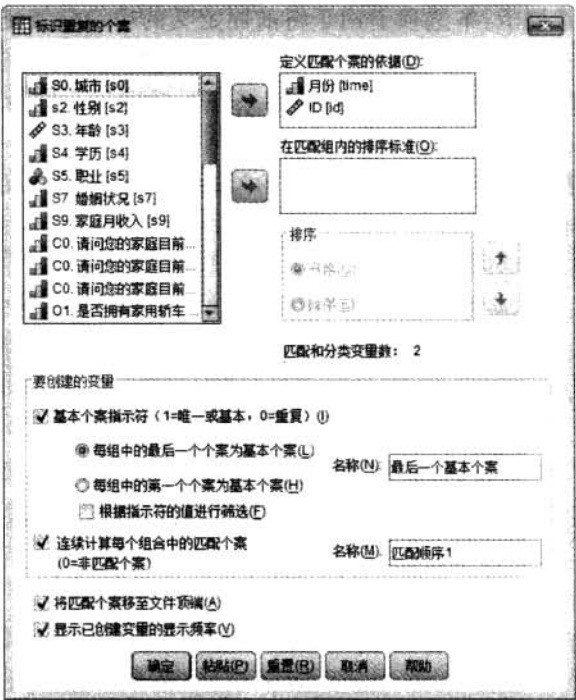



图 4.27 “标识重复的个案”对话框

(4) “连续计算每个组合中的匹配个案”:在每一个匹配组合中为个案创建序列值为 1 到 n 的变量。该序列基于每一组中当前个案的顺序,可以是原文件顺序,也可以是由任何指定的排序变量决定的顺序。

 在后续分析中可利用上述生成的指示变量作为过滤变量进行个案选择,从而在报告和分
析中排除重复个案,同时又无须从数据文件中删除这些个案。

操作完毕后,得到的结果如图 4.28 所示,可见变量“最后一个基本个案”等于 0 表示相应的
记录为重复记录,本例中共发现 2 条,它们均与第 3 条个案记录重复,这正是前面所设定的情形。

	time	id	最后一个基本个案	匹配顺序
1	200704	1	重复个案	1
2	200704	1	重复个案	2
3	200704	1	主个案	3
4	200704	4	主个案	0
5	200704	5	主个案	0
6	200704	6	主个案	0
7	200704	7	主个案	0
8	200704	8	主个案	0
9	200704	9	主个案	0
10	200704	10	主个案	0
11	200704	11	主个案	0
12	200704	12	主个案	0

图 4.28 操作结束后的数据界面

最后,结果窗口中还会给出本次操作的汇总信息,如图 4.29、图 4.30 所示。

		频率	百分比	有效百分比	累积百分比
有效	重复个案	2	.2	.2	.2
	主个案	1145	99.8	99.8	100.0
	合计	1147	100.0	100.0	

图 4.29 所有最后一个匹配个案的指示符为主个案

		频率	百分比	有效百分比	累积百分比
有效	0	1144	99.7	99.7	99.7
	1	1	.1	.1	99.8
	2	1	.1	.1	99.9
	3	1	.1	.1	100.0
	合计	1147	100.0	100.0	

图 4.30 匹配个案的连续计数

4.5.4 标识异常个案

异常个案往往是统计分析中非常令人头痛的问题,这些个案的出现有可能是因为录入错误所致,这种情况比较好处理,找到并更正即可;更麻烦的情形是数据无误,但变量值的确异常,此类个案往往就会成为分析者,特别是统计初学者的一大难题,因为最常用、最正统的分析模型可能会因其存在而无法使用,必须换用更合适的分析方法。但无论怎样,异常个案的提前识别显然会大大方便相应的数据管理和统计分析工作。有鉴于此,SPSS 提供了标识异常个案过程,该过程采用较为复杂的统计算法,可以在探索性数据分析步骤中,快速检测到用于数据审核的异常个案,从而协助用户提前对其进行处理。

1. 算法原理

由于在标识异常个案的过程中使用了较为复杂的算法,这里只是简单地介绍其原理,希望深入了解其内容的读者可以在学习完相应统计模型后再参阅本过程的算法文档。

首先,由于所用聚类方法的原因,模型的分析结果会受到个案顺序的影响,因此理论上要求个案完全随机排序以将其影响降至最低。为了保证结果的稳定性,在文件较大时可使以随机顺序排序的个案样本运行多次来对结果进行验证。

本算法假设所有变量相互独立,连续性变量均服从正态分布,分类变量均服从多项分布。但实际上当变量违反上述独立性假设和分布假设时,该过程的分析结果相当稳健,用户无须严格拘泥于此。

整个计算过程可分为如下 3 步。

(1) 建模:采用两步聚类方法(本方法详情参见本丛书高级教程),将所有个案按照其相似性自动分为若干类(称为对等组)。所建立的聚类模型以及相应的计算统计量均被存储起来供后续分析使用。

(2) 评分:使用该聚类模型对每一个案例进行其相对于所属类别的异常度评估,并计算出相应的异常索引(Anomaly Indices,综合各变量值的偏差度指标而得,具体算法与对数似然值有关)。计算完毕后所有案例将按该指标降序排列,索引值最高的一部分(具体比例在对话框选项

案例	异常索引
397	10.651
577	7.311
703	6.215
622	6.020
588	5.706

图 4.32 异常个案索引列表

图 4.32 给出了异常索引值最大的 5 个案例名单,注意个案是按照异常索引值降序排序的,其中索引值最大的是 397 号个案,索引值达到了 10.651。

案例	对等 ID	对等大小	对等大小百分比
397	1	598	52.1%
577	1	598	52.1%
703	1	598	52.1%
622	1	598	52.1%
588	1	598	52.1%

图 4.33 异常个案对等 ID 列表

图 4.33 给出的是与聚类分析相关的异常个案报告,可见这 5 个异常案例在聚类分析中均被分入第一个对等组,该对等组共有 598 条个案,占总样本量的 52.1%。

图 4.34 给出的是这些案例被标识为异常个案的原因,以索引值最高的 397 号个案为例,系统是根据 index1 的取值将其标识出来的,index1 在第一个对等组中的范数(其实就是本组均数)为 80.42,但 397 号的变量值则为 0,其影响度衡量指标达到了 0.463。

原因:1

案例	原因变量	变量影响	变量值	变量范数
397	index1	.463	.00	80.4185
577	index1b	.461	.00	78.6153
703	index1b	.542	.00	78.6153
622	index1	.426	23.43	80.4185
588	index1	.449	23.43	80.4185

图 4.34 异常个案原因列表



需要指出的是,在作者看来,标识异常个案中使用的算法对初学者而言过于复杂,且分析结果也不一定符合需求,因此各位读者不要对此方法过于迷信,只要将其作为一个强有力的辅助工具来加以使用即可。

思考与练习

1. 自行完成本章中涉及的数据管理操作。
2. 尝试为 CCSS 案例建立相应的数据字典文件,并思考数据字典文件在此类大型项目中的使用价值究竟有多大。

第5章 SPSS 编程与扩展

如同第1章中所提到的:SPSS 中 90% 的功能都可以通过对话框实现。但是,正是因为这 10% 的功能较为复杂,才无法编制成对话框来操作,因此它们也最终决定了高手和初学者的区别在哪里。遗憾的是,由于 SPSS 的大多数用户已经习惯了在图形化的对话框中进行操作,也由于国内大多数教材的导向,许多使用者已经不再熟悉 SPSS 程序,也不再了解程序编辑窗口等专用模块的用途。本章的目的就是打通各位读者成为高手之路,为读者彻底掌握 SPSS 的高级操作打下基础。

除介绍编程知识以外,本章还将进一步介绍结果输出系统、扩展功能接口等,以将对 SPSS 的使用从该软件本身延伸到所有可以利用的外部资源上,了解这一点非常重要,因为这正是 SPSS 软件的战略发展方向之一。

5.1 SPSS 编程入门

注意,为了节省版面,提高本书性价比,本节将基于 SPSS 的实际使用需求,对用户必需的编程语言知识进行介绍。对于希望迅速成长为 SPSS 编程高手的读者,可以在学习完本节后自行阅读帮助系统中的语法参考手册来进一步提高自己的 SPSS 编程技能。

5.1.1 基本语法规则

通俗地讲,SPSS 程序是由若干条 SPSS 语句构成的,这些 SPSS 语句基本上以易于识别的英语单词作为命令关键词,同时遵循一定的基本规则。

1. 主命令格式

每条 SPSS 命令必须从新行开始,但可以在该行的任何列中开始,并持续所需任意数量的行。为了保证兼容性,单行长度最好不要超过 254 个字符。每条命令应该以句点为命令终止符。如果没有句点作为命令终止符,也可以将空行解释为命令终止符。下面是一个典型的 SPSS 命令:

```
COMPUTE NEWVAR = OLDVAR * 2.
```

这一程序命令执行的就是第3章介绍的计算新变量,其组成如下:

(1) 命令动词:最前面的“COMPUTE”为命令动词,不分大小写,在 SPSS 中,所有的命令动词均由相应功能的英文单词或者词组构成,且都可以缩写为前 4 个字母,例如,在本例中书写为“COMP”也是可以的。

(2) 分隔符:命令动词后的空格用于分割命令动词和表达式。空格是最常见的分隔符,但在特殊情况下,斜杠和逗号也有可能作为分隔符,后面会详细说明。

(3) 命令表达式:在空格后紧跟的就是具体的命令表达式,根据所执行命令的不同,该表达

式可以是变量列表,也可以是数学表达式,本例中就是一个数学表达式。

(4) 终止符:整个命令的最后会以一个句号终止,因此一个 SPSS 语句完全可以占多行,系统只有在读取了句号结束符的时候才会认为该语句已经结束。

根据上述要求,下面的命令虽然可读性不强,但却是合法的语句。

```
Var1 LABE var1
```

```
'label of var1'
```

```
/ var2 'label of var2'.
```

提示:应尽量养成良好的编程习惯。

2. 子命令格式

对于较为复杂的 SPSS 命令,还需要对主命令之下的各种选项细节加以设定,此时就会用到子命令(Subcommand)。

子命令是对命令的进一步说明,必须要依附于某个命令动词而存在,大多数统计分析命令都需要进行自命令的定义。当然,很多非关键的子命令都会有其默认设定,因此书写时只需要进行少数几个关键子命令的设定即可。

下面就是一个典型的带子命令的 SPSS 命令行:

```
FREQUENCIES VARIABLES = var1 var2
```

```
/STATISTICS = MEAN
```

```
/ORDER = ANALYSIS.
```

该命令要求对变量 var1、var2 做频数分析,同时输出均数,对其中出现的子命令说明如下:

(1) 子命令名:所有的子命令动词均由相应功能的英文单词或者词组构成,且都可以缩写为前 4 个字母。

(2) 分隔符:在同一命令中有多个子命令时,需要用“/”分隔。但是对于第一个子命令,“/”是可以省略的,本例中就是如此。

(3) 子命令顺序:在有的命令语句中,子命令的先后顺序是有规定的,颠倒可能会报错。不过读者无须记忆严格的规定,可以利用对话框中的“粘贴”按钮来自动生成合法的命令程序,详见本书 5.1.2 小节。

3. 关键字与保留字

关键字(Keywords)用于识别命令、子命令、函数以及其他指令。除了上述指明特定的命令/子命令/函数名称的关键字外,还包括一些保留字,它们均不能用做自定义变量名;否则会造成错误,比较常见的一些保留字如下:

(1) 逻辑运算符:AND、OR、NOT。

(2) 关系运算符:EQ、GE、GT、LE、LT、NE。

(3) 变量关系指定符:ALL、BY、TO、WITH。

(4) 数值定义符:LOWEST、LO、HIGHEST、HI、THRU、MISSING、SYSMIS。

下面对上述保留字中较为有用的 ALL 和 TO 解释如下:

ALL:用于指代全部变量,例如,下列语句会对数据集中的全部变量进行频数分析。

FREQUENCIES VARIABLES = ALL.

TO:当变量为有规律的流水编号时,可以用 TO 来代替依次书写相应的变量名,例如,下面两个语句的执行效果是等价的:

FREQUENCIES VARIABLES = v1 v2 v3 v4 v5.

FREQUENCIES VARIABLES = v1 TO v5.

4. 临时变量与系统变量

当在程序中需要定义一些临时变量,但又不准备将其写入数据集中时,可以将这些变量设置为以“#”开头的名称,系统就会自动识别其为临时变量(Scratch Variable),在程序运行期间存储在内存中,而程序运行结束后自动丢弃,不再写入数据文件中。

系统变量则是由 SPSS 系统预定义好的一些特殊变量,它们均以“\$”符号开头,在数据转换命令中可以直接调用,就如同一个普通的变量一样,但用户不能更改其数值。

(1) \$CASENUM:返回个案的顺序号,除非程序的编写非常特殊,在绝大多数情况下该顺序号就等于个案的相应行号。

(2) \$SYSMIS:返回系统缺失值。

(3) \$JDATE:返回当前日期距离 1582 年 10 月 14 日的天数。

(4) \$DATE:返回以“dd-mmm-yy”方式记录的字符串格式日期。

(5) \$DATE11:返回以“dd-mmm-yyyy”方式记录的字符串格式日期。

(6) \$TIME:返回当前时间距离 1582 年 10 月 14 日午夜的秒数。

(7) \$LENGTH:返回当前页面长度。

(8) \$WIDTH:返回当前页面宽度。

5. 几个特殊命令

SPSS 命令的数量和种类都很多,这里没有一一进行详细介绍,只是结合实际使用中的需求给出如下两个特殊命令的解释。

(1) EXECUTE 命令:SPSS 的命令大致可以分成数据转换命令(Transformation Command)和统计分析过程(Procedure Command)两大类。对于后者,提交后就会直接运行;但是对于数据转换命令,由于可能涉及数据结构和数据内容的变化,因此提交后只是进入缓存,不会立即执行,而是要等到 EXECUTE 语句提交后才会一并解释执行,因此通常可以在数据整理程序段的末尾看到该语句。

(2) COMMENT 命令:为了增强程序的可读性,几乎所有的程序设计语言都有注释命令,本命令也可以简化为以“*”开头。需要指出的是,该命令仍然以“.”结束;否则系统会误以为随后的新命令仍然是注释的一部分,导致程序出错。

5.1.2 SPSS 程序的创建方式

在 SPSS 中创建程序的最基本和原始的方法就是在语法编辑器中直接编写程序,但这样显然会事倍功半,为了提升用户的工作效率,在 SPSS 中真正常用的是如下 3 种方式:对话框粘贴程序、输出 LOG 粘贴程序和日志文件编辑程序。

1. 对话框粘贴程序

“粘贴”按钮在几乎所有 SPSS 对话框中均存在,它是专门为编程准备的。当选择菜单和对话框进行操作时,只要在设置完毕后单击对话框中的“粘贴”按钮,与对话框设定相对应的命令语句就会被完整地粘贴到语法编辑窗口中去,这是 SPSS 编程中最常见,也是最轻松的一种方式。

2. 输出 LOG 粘贴程序

有的时候,用户在分析工作基本结束后,会希望将刚才的操作保存为程序以便下次发生类似情况时使用,这时可以利用输出 LOG 来生成相应的程序:新版的 SPSS 默认会将所有操作所对应的程序命令以 LOG 文本的形式输出到结果窗口中去,用户只需要找到相应操作所对应的那些程序段,然后将其依顺序粘贴到语法窗口中,并加以编辑和保存即可。

如果在输出窗口中没有看到 LOG 文本,则选择“编辑”→“选项”菜单项,打开“选项”对话框,在“查看器”选项卡中选中“在日志中显示命令”复选框即可。

3. 日志文件编辑程序

在 SPSS 系统中,几乎所有的操作都会以程序代码的形式保存在系统日志文件中,这样就为用户重复利用已有分析操作提供了便利条件。首先进入“编辑”→“选项”菜单项,打开“选项”对话框,在“文件位置”选项卡中部的日志文件栏找到日志文件,日志文件在默认情况下为用户文档路径下的文本文件 statistics.jnl。然后利用任何一种文字编辑软件打开该文件,即可看到从该文件建立起至今的全部命令代码。可以选择相应时间段的命令代码,将其编辑为所需要的命令程序并另存为程序文件。

5.1.3 结构化语句简介*

采用每一种完善的结构化语言编写的程序都由顺序、分支、循环 3 种结构构成,SPSS 程序也不例外。本节就来简要介绍一下分支和循环语句的语法,以使读者对 SPSS 编程有一个全面的了解。

1. 分支(条件)语句

(1) IF 语句。分支语句就是大家非常熟悉的判断语句,SPSS 中最简单的判断语句是 IF 语句,其格式如下:

IF 逻辑表达式 目标表达式

逻辑表达式用于给出逻辑判断条件,而目标表达式则是当逻辑条件被满足时需要进行的操作。最常见的情况是给一个变量赋值,如 compute 语句。比如下面的语句

```
IF (AGE > 20 AND SEX = 1) GROUP = 2.
```

```
EXECUTE.
```

其含义就是当 AGE > 20 并且 SEX = 1 时,变量 GROUP 被赋值为 2。注意最后的 EXECUTE 语句不能省略,否则程序被存在缓冲区里,没有真正执行。

(2) DO IF & END IF 语句。IF 语句适合于比较简单的情况,只能进行一种后续操作,如果

* 本节适合于希望在编程知识上有所深入的读者阅读,读者如对此不感兴趣可以跳过,不影响对后续内容的理解。

需要多重分支,或者进行多种后续操作,则可以使用这里要介绍的 DO IF/END IF 语句,其格式如下:

DO IF 逻辑表达式

程序段

ELSE

程序段

END IF

DO IF/END IF 语句的作用主要是生成多重分支的判断结构,举例如下:

DO IF (age < 20).

COMPUTE ageclass = 1.

COMPUTE younger = 1.

ELSE IF (age < 30).

COMPUTE ageclass = 2.

ELSE IF (age < 50).

COMPUTE ageclass = 3.

ELSE.

COMPUTE ageclass = 4.

END IF.

EXECUTE.

当然,对于比较简单的情况,以上类似的工作也可使用 recode 语句完成,但在 DO IF/END IF 语句中可以进行复杂的条件判断,功能要更强些。

2. 循环语句

在 SPSS 中提供了多个循环语句,有 DO REPEAT/END REPEAT、LOOP/END LOOP 等,这里只介绍后者,LOOP/END LOOP 语句的语法格式如下:

LOOP 控制变量名 = 起始值 TO 终止值 [BY 步长]

程序段

END LOOP

该语句主要用于建立数据集和进行数据变换操作,举例如下:

SET MXLOOPS = 10.

设置最大允许循环次数为 10

LOOP.

开始无限循环,直到达到最大次数

COMPUTE X = X + 1.

将变量 X 累加 1

END LOOP.

结束循环

EXECUTE.

开始执行以上程序

该程序会将数据文件中的 X 重复加 10 次 1,即加 10。但如果文件中没有变量 X,则执行后 X 为缺失值。再看下面的程序

LOOP #lop = 1 TO 5.	开始循环,要求循环 5 次
COMPUTE X = X + 1.	将变量 X 累加 1
END LOOP.	结束循环
EXECUTE.	开始执行以上程序

该程序会将数据文件中的 X 重复加 5 次 1,其中变量 lop 前带有#号,表明为临时变量,不写入数据集中;否则将会在数据集中建立一个新变量 lop,其大小等于循环结束后 lop 的取值 6。

5.1.4 一个简单程序示例

下面给出一个数据集生成程序,里面用到了许多前面介绍过的知识,同时还用到了建立数据文件所需的一些语句,希望读者通过这个示例能对 SPSS 程序有一个更深入的了解。

SET SEED 1.	将伪随机种子设为 1
INPUT PROGRAM.	开始数据录入程序段
LOOP #LOP = 1 TO 50.	一共循环 50 次,变量 LOP 不写入文件中
COMPUTE A = RV. NORMAL(0,1).	新变量 A 服从标准正态分布
END CASE.	结束一条记录的定义
END LOOP.	结束循环
END FILE.	结束数据文件
END INPUT PROGRAM.	结束数据录入程序
EXECUTE.	开始执行以上程序
DO IF (A >= 0).	
COMPUTE B = A.	如果 $A \geq 0$,则新变量 $B = A$
ELSE.	
COMPUTE B = A * 2.	否则, $B = A * 2$
END IF.	
EXECUTE.	开始执行以上程序
LIST.	在结果窗口中输出数据列表

在程序运行完毕后,就会生成一个有 50 条记录的新数据集,其中变量 A 服从均数为 0、标准差为 1 的标准正态分布,而变量 B 的取值在变量 A 大于等于 0 时和 A 相等;否则等于 A 的两倍。同时在结果窗口中会将所有记录打印输出。由于采用的是伪随机数,以上程序重复运行时得到的结果都是相同的。



如果是简单的数据录入,则使用下面的程序框架即可,这里给出的程序示例演示的是比较复杂的分支、循环等操作。

```
DATA LIST FREE / 变量名称及格式列表.
BEGIN DATA
数据块列表
END DATA.
```

5.2 语法编辑窗口操作入门

5.2.1 语法编辑窗口界面

语法编辑器或称为语法编辑窗口,是 SPSS 中专为创建、编辑和运行命令语法而设计的窗口环境。在现有版本的 SPSS 中,语法编辑器有以下特色。

(1) 自动完成。随着输入,可以从上下文敏感列表中选择命令、子命令、关键字和关键字值。可以选择自动提示列表或按需要显示列表。

(2) 颜色编码。命令语法(命令、子命令、关键字和关键字值)的识别元素是颜色编码,因此浏览一下即可定位未识别项。另外,一些常见语法错误,如未匹配的引号,都经过颜色编码以供快速识别。

(3) 分界点。可以在指定点停止执行命令语法,从而可以在查看数据或输出后再继续。

(4) 书签。可以设置书签,从而可以快速导航大型命令语法文件。

(5) 逐步执行。可以一次一个命令逐步执行命令语法,单击“前进”按钮到下一个命令。

图 5.1 所示的是一个典型的语法编辑器窗口,可见其被分为 4 个区域:编辑器窗格、装订线、导航窗格和错误窗格,下面就对其进行介绍。

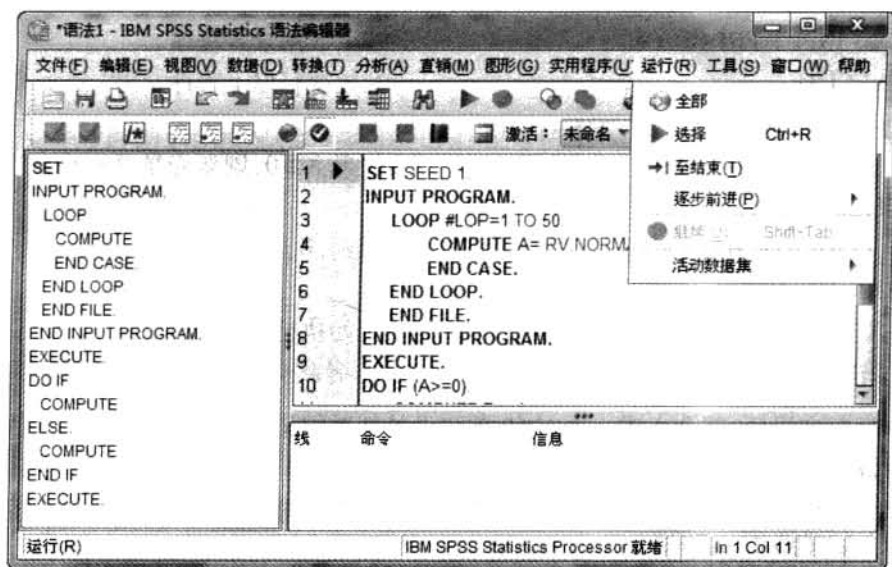


图 5.1 程序语法编辑器窗口

1. 编辑器窗格

位于窗口右侧,是语法编辑器窗口的主要部分,用于输入和编辑程序命令。实际上,对于熟悉现代编程环境的读者而言,该窗格的功能是不需要做任何解释的。但考虑到初学者的需求,这里还是简要列出该窗格的特点如下。

(1) 自动填充:对于系统可以自动识别的命令关键词,在输入过程中会自动弹出下拉列表用于选择,直接按回车键即自动填充完整的关键词进入窗格。

(2) 彩色标注:窗格中的程序会自动按照所识别的内容被标记为蓝色、黑色、灰色、红色,以及加粗等,易于用户识别,特别指出其中红色代表已确认的错误代码,需要用户加以修改。

2. 装订线

装订线实际上位于编辑器窗格内部左侧,用于显示行号、分界点、书签、命令跨度和进度指示等信息。

(1) 行号:可以通过选择“视图”→“显示行号”菜单项以显示或隐藏行号。

(2) 分界点:在指定点停止执行,显示为一个与设置分界点的命令相邻的红圈。

(3) 书签:在命令语法文件中标记特定行,显示为包含分配到书签的数字(1~9)的正方形。悬停在书签图标上将显示分配到书签的书签编号以及名称(如果有)。

(4) 命令跨度:是提供命令开始和结束的可视指示符的图标。可以通过选择“视图”→“显示命令跨度”菜单项以显示或隐藏命令跨度。

(5) 进度指示:给定语法运行的进度,在装订线中使用向下箭头表示,从第一个命令运行扩展到最后一个命令运行。这在运行包含分界点的命令语法和逐步执行命令语法时最有用。

3. 导航窗格

导航窗格位于窗口最右侧,列出所有已识别命令的列表,且自动按照缩进格式以它们在窗口中出现的顺序显示。

(1) 操作:单击导航窗格中的命令会将编辑器窗格中的光标置于相应命令开始位置,也可以使用向上和向下箭头键移动通过命令列表或单击命令以导航到该命令。双击将选择命令。

(2) 颜色标识:检查无误的命令标识为黑色,发现语法错误的命令名称在默认情况下显示为红色加粗文本,未识别文本的每行第一个单词显示为灰色。

(3) 显示方式:可以通过选择“视图”→“显示导航窗格”菜单项以显示或隐藏导航窗格。

4. 错误窗格

错误窗格显示最近运行过程中发生的运行时间错误,包括每个错误的信息包含错误发生的行号。可以使用向上和向下箭头移动通过错误列表,单击列表中的一个条目会将光标置于生成该错误的行上。



可以通过从菜单栏中选择“视图”→“显示错误窗格”菜单项以显示或隐藏错误窗格。

5.2.2 程序的运行与调试

无论以何种方式生成程序,最终都是在语法窗口中加以运行的,这里以 5.1.4 小节中列出的程序为例来介绍程序的调试方法。首先对程序做一点修改,将语句“COMPUTE B = A.”句末的英文句点删除,人为造成一个错误。然后选择“运行”→“全部”菜单项,程序会立即执行,但由于代码有误,此时在结果窗口会报错,执行完毕后语法窗口也会出现变化,如图 5.2 所示。

可以注意到错误窗格中显示第 11 行的命令出错,详细信息为表达式意外结束。再看装订线处,可以看到有一个单向箭头跨越了整个程序段的范围,表示整个程序段的开始和结束位置。但

同时在 11~12 行有一个线段出现,标识这两行为一条命令。显然,这是因为删除了一个句点所致。根据上述信息将程序修改为正确内容,重新运行,即可得到正确的结果输出。

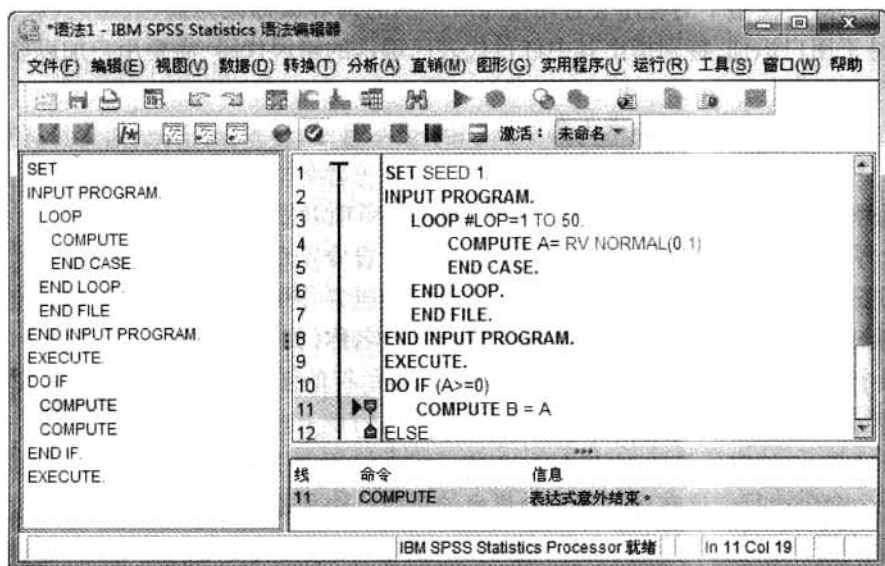


图 5.2 发现程序错误后的语法编辑窗口

除了第一项运行全部程序代码外,在“运行”菜单中还可以选择只运行所选择的程序语句、运行当前光标所在行的语句、从当前语句起一直运行到程序结束,以及逐句以用于调试等,读者可以自行尝试,这里不再详述。

5.3 INCLUDE 命令与宏程序

通过前面几节的学习,读者应该已经掌握了在 SPSS 中编写和运行程序的基本知识,本节将在此基础上更进一步,讲解程序代码段的重复利用方法。

5.3.1 INCLUDE 命令

当编写 Syntax 程序时,如果发现需要编写的程序语句正好是另一个 Syntax 文件的内容;或者当发现需要编写的程序语句其实是几个 Syntax 文件的总和时,除了可以通过复制、粘贴的方法来利用原有资源,生成一个新的 Syntax 文件外,还有一种更简单的办法,那就是使用 INCLUDE 命令。以笔者的 CCSS 出表程序为例,由于相同框架的表格需要针对不同地区的数据重复编制,因此笔者将出表的程序段单独存储为“CCSS 出表核心程序.sps”,然后在主程序中以如下方式调用:

```
* *****全国.
USE ALL.
EXEC.
INSERT FILE ='CCSS 出表核心程序.sps'.
```

```
* *****华北.
```

```
COMPUTE FILTER_$ = (TSO_1 = 1).
```

```
FILTER BY FILTER_$.
```

```
EXEC.
```

```
INSERT FILE = 'CCSS 出表核心程序.sps'.
```

采用这种方式,出表核心程序实际上有 500 余行,由于需要重复调用 8 次,如果将其全部写入主程序,则主程序将有 4 000 余行!但现在采用 INCLUDE 方式来调用,使得主程序只有数十行,且结构非常简单明了,大大减少了程序编写、调试和查错的工作量。

5.3.2 宏程序

宏技术对于很多人而言可能已经不是什么新鲜事物了,在 Word 中就有宏功能。但 SPSS 中的宏可能大家还不太了解,实际上,SPSS 很早就嵌入了宏功能,用于实现已有程序的重复利用,以提高工作效率,满足大量类似分析任务的需求。

1. 宏的基本格式定义

下面给出一个非常简单的宏示例:

```
DEFINE ! M_SAMPLE() 'ABC'
```

```
* 任何有效的 SPSS 程序段.
```

```
! ENDDEFINE.
```

```
IF VARX = 1 VARY = ! M_SAMPLE.
```

```
EXECUTE.
```

这个宏的名称是 M_sample,其作用是将字符串“ABC”赋值给宏名称本身,在随后的 IF 语句中,直接使用了宏名称来代替对字符串“ABC”的使用。从上述程序段中可以看出宏的基本格式定义如下。

(1) 一个宏应当以 DEFINE 命令开头,在其后指定宏程序的名称,宏名称需要以“!”开头,以保证不会和程序中的其他变量重名。

(2) 宏名称后的括号用于定义宏参数,这些参数在宏被调用的时候会被一同读入。即使没有宏参数,也应保留此括号。

(3) 宏的主体部分可以是 SPSS 命令,也可以是一些专门定义的宏语句,如条件语句、循环语句等。

(4) 一个宏必须以! ENDDEFINE 语句结束。

(5) 一个宏可以使用其宏名进行调用,在不会引起歧义的情况下可以省略“!”号,但建议读者尽量保留该符号以增强程序可读性。如果宏定义中含有参数,则在调用时需要对每个参数进行赋值,或每个参数均有默认值可以对应,否则将会报错。

2. 宏参数

在 CCSS 项目数值计算中使用的一小段代码如下。

```

DEFINE M_COMP ( INVAR1 = ! CHAREND('/ ')).
RECODE
  ! INVAR1
  (1 THRU 5 = COPY) (ELSE =9) INTO ! CONCAT('T',! INVAR1) .
EXEC.
! ENDDDEFINE.

M_COMP INVAR1 = A3 .

```

在宏 M_comp 的定义中涉及了宏变量 invar1,因此在随后的宏调用中,就需要同时对对应的宏变量 invar1 赋值。

当有多个宏变量需要定义时,只需要依次书写,并用“/”分隔即可,例如,下列 CCSS 出表程序段中的代码:

```

DEFINE M_Tb02 ( invar1 = ! charend('/ ') / strcat1 = ! charend('/ ') / strcat2 =
  ! charend('/ ') ).
* 宏代码段主体程序 .
! ENDDDEFINE.

M_tb02 invar1 = a3a
/ strcat1 = subtotal, 10, 20, 30, subtotal, 110, 120, 130, hsubtotal, othernm
/ strcat2 = hsubtotal, 10, 20, 30, 110, 120, 130, hsubtotal, othernm .

```

多个宏变量的定义格式有多种,这里只列出最简单的一种,大家只需要确保定义格式和调用格式相一致,就不会出现错误。

当需要运行大量类似的分析程序时,宏程序的优势是非常明显的。读者可以仔细阅读第6章 CCSS 分析实例中的相应内容,就能深刻地体会到这一优势。

5.4 OMS 系统与程序自动化

5.4.1 OMS 系统

随着统计分析知识的逐渐普及,用户对统计分析报表的要求越来越高,SPSS 默认输出的格式已经不一定能够满足需求,而输出重定向可以将分析结果指定输出到相应的文件格式中,使得制作特定格式的输出报表成为可能。

OMS(Output Management System,输出管理系统)为用户提供了提取和控制结果分析窗口中输出内容的功能。OMS 系统在 12 版开始提供,目前已经非常成熟,现在不仅可以将输出结果存储为 SPSS 数据格式(SAV)、XML 格式、HTML 格式、TXT 格式、PDF 格式等多种常见格式,还可以指定输出内容是分析结果中的表格、文本、图形中的一部分,如只输出回归分析中回归系数的检验结果,或者全部分析中的直方图等。而相应的 OMS 设定已经实现了对话框操作,这更是大

大方便了 OMS 的使用。

1. 操作界面

选择“实用程序”→“OMS 控制面板”菜单项,即可打开 OMS 控制面板对话框,如图 5.3 所示。

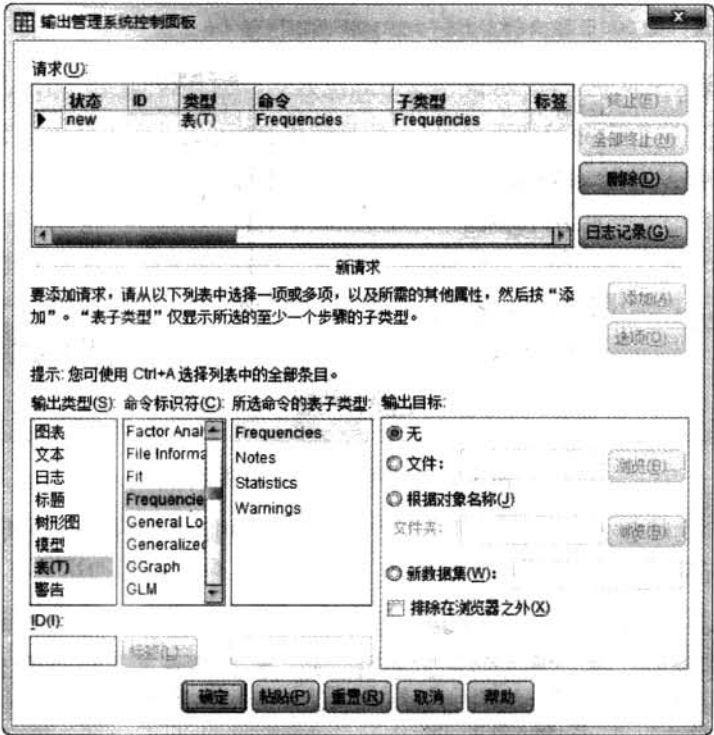


图 5.3 OMS 控制面板对话框

图 5.3 所示的对话框上方会列出所有已经设定完毕的输出控制条目,单击右侧的按钮可以对这些输出控制进行启动、终止或者删除操作。下方则给出了所有可被控制的输出项,可见几乎所有的输出种类均可被重定向输出。在每一类里面又会具体到命令标识,以及该命令的所有输出中的某一个具体表格,以做到精确控制。右侧的“输出目标”框组则用于指定希望输出的文件名称和格式(具体文件格式通过右侧中部的“选项”按钮设定)。

具体操作时,用户先在下部选择好希望重定向输出的元素,然后在右侧设定好输出格式和文件名称,单击中部的“添加”按钮即可将该条目加入请求列表中。此时如果单击“确定”按钮,就可以启动相应的条目。

除 OMS 控制面板外,“程序”菜单栏内还有一个栏目为“OMS 标识符”,所打开的对话框用于显示各类输出元素所对应的 OMS 标识符,如图 5.4 所示,由于该对话框的实质就是一个 OMS 字典,主要用于编程参考,因此这里不再详细解释。

2. 实例分析

这里假设希望将 Frequencies 过程中的所有频数表重定向输出至数据文件 freq.sav 中,则首先按照图 5.4 中的设定选择输出类型,即“表”→Frequencies→ Frequencies,然后在右侧将输出组

件设定为新数据集:freq. sav,单击“添加”按钮后单击“确定”按钮,此时 OMS 系统就会开启相应的输出条目重定向,并将所有符合要求的表格内容写入指定的数据文件中,当然,此时被写入的目标数据文件 freq 在内存中一直处于锁定状态,无法在前台显示,也无法使用。



图 5.4 “OMS 标识符”对话框

现在来运行一次频数分析,例如,对 CCSS 文件中的 S0 城市进行频数分析,结果窗口中的相应输出如图 5.5 所示。

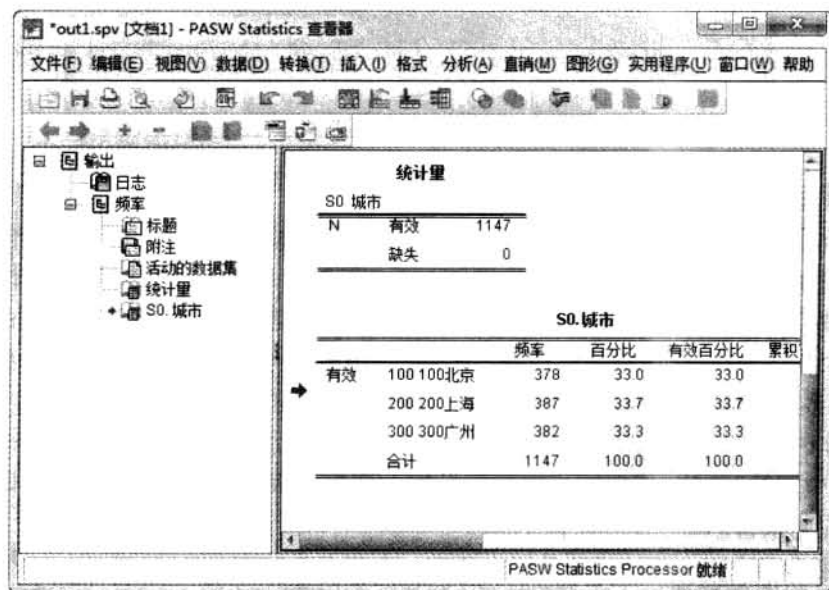


图 5.5 结果窗口中的输出

还可以运行其他一些分析过程,会发现一切结果输出如常,没有任何特殊之处。但如果此时重新进入 OMS 控制面板,选择刚才的 Frequencies 表格所对应的请求并单击“终止”按钮和“确认”按钮,此时就会打开一个新的数据窗口,该数据文件自动被命名为 freq. sav,其内容如图 5.6 所示。

	Command_	Subtype_	Label_	Var1	Var2	频率	百分比
1	Frequencies	Frequencies	S0 城市	有效	100北京	378	33.0
2	Frequencies	Frequencies	S0 城市	有效	200上海	387	33.7
3	Frequencies	Frequencies	S0 城市	有效	300广州	382	33.3
4	Frequencies	Frequencies	S0 城市	有效	合计	1147	100.0

图 5.6 新生成的 freq. sav 数据文件内容

可见除了命令索引、亚类索引、变量名称等必要的变量外,数据文件右侧的各变量实际上正好和相应的频数表格中的内容一一对应,也就是说,上述数据文件的内容实质上就是频数表输出的精确重定向,现在就可以利用该数据完成随后希望进行的工作了,这就是 OMS 系统的实质。

3. 程序实现

OMS 系统属于非常高级和复杂的功能,实际上以程序方式实现更为多见,因此这里简单介绍一下 OMS 的程序实现,以上面的分析为例,单击“粘贴”按钮,可以看到对应的 OMS 程序如下:

```
* OMS.
DATASET DECLARE freq. sav.
OMS
/SELECT TABLES
/IF COMMANDS = ['Frequencies'] SUBTYPES = ['Frequencies']
/DESTINATION FORMAT = SAV NUMBERED = TableNumber_
OUTFILE = 'freq. sav'.
```

上面就是一个非常简单的 OMS 程序,用于在特定的情况下打开 OMS 输出系统。如果用文字对内容加以解释,指的就是监视所有的表格输出,当运行的过程命令为 Frequencies,且所生成的结果表格为 Frequencies 时,将相应的表格内容输出到数据文件 freq. sav 中。

以上程序运行完毕后,OMS 系统就会一直保持打开状态,直到新的 OMS 命令对其加以更改,或者 SPSS 关闭为止。在此期间,OMS 系统会将所有符合要求的表格内容写入指定的数据文件中,而相应的目标数据文件也一直处于锁定状态,无法使用。如果希望将其关闭,则可以使用如下命令:

```
OMSEND.
```

这时 OMS 系统会关闭,并将所有数据写入目标文件中,并将其释放。

5.4.2 程序自动化

作为国际一流的统计分析软件,SPSS 不仅可以完成简单的分析操作,也可以针对海量数据完成大规模的统计运算。但是,针对海量数据进行的分析一般都较为耗时,如何实现程序运行的批处理、自动化就变得十分重要。下面介绍如何在 SPSS 中实现程序的自动化运行。

1. 界面说明

生产设施(Production Facility)模块原先是独立于 SPSS 的一个单独软件,现在已经被整合到 SPSS 软件内,选择“实用程序”→“生产设施”菜单项,系统就会启动生产设施界面,如图 5.7 所示。

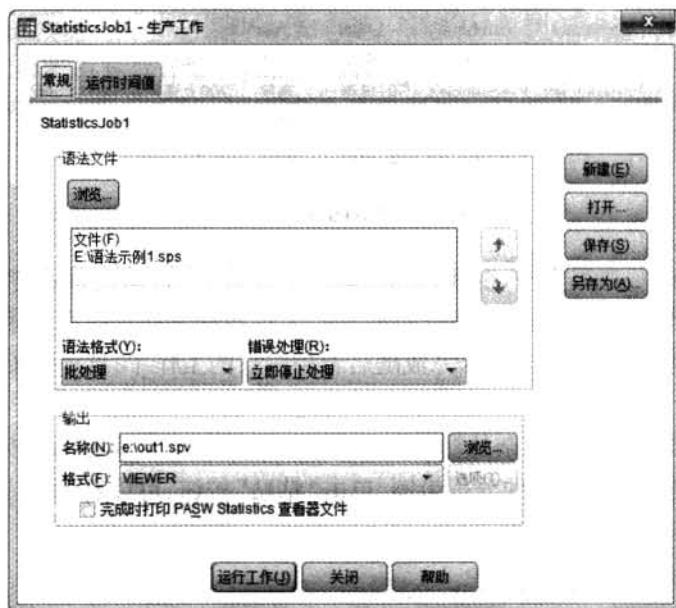


图 5.7 生产设施的操作界面

在生产设施的主界面中,主要是用于定义相应工作需要调用的程序段、具体的执行方式、结果输出方式等参数。

(1) 语法文件:用于指定工作中需要调用执行的程序名称,可以调用多个程序文件,系统将会依次执行。

(2) 语法格式与错误处理:默认为交互式执行,并且出错后继续执行,图 5.7 中已经修改为“批处理”,并且出错后“立即停止处理”。

(3) 输出:指定希望输出的文件名称和文件类型,输出格式除了默认的 SPV 格式以外,还支持导出功能可以实现的 DOC、XLS、PDF 等多种格式。

2. 操作实例

这里仍然以 5.1.4 小节的程序为例讲解生产设施的操作方法。首先将上述程序另存为“e:\语法示例.sps”,然后在生产设施界面上单击右侧的“新建”按钮,进行一个新的工作文件的设定,此时就可以将程序文件和结果文件设定为图 5.7 中所示意的情形,在全部设定完毕后单击右侧的“运行”按钮,会弹出对话框提示正在运行相应的工作,此时系统就可以进入无人值守方式自动运行,用户可以离开,也可以在机器上进行其他工作。待工作项目中指定的程序运行完毕后,系统就会弹出对话框提示“Job1 已完成”,此时可以发现所指定的结果文件 E:\out1.spv 已经生成,打开后就是相应的数据列表输出结果。

思考与练习

1. 自行尝试编写一个非常简单的数据集生成程序(提示:可以利用 5.1.4 小节最后提供的程序模板)。

2. 自行练习本章中涉及的各种操作,特别是宏程序和 OMS 程序部分。

第6章 统计实战案例集锦(一)

6.1 数据异常值的自动核查与报告

6.1.1 项目背景

CCSS 每月的数据均由计算机辅助电话访问系统(CATI)收集而来,该系统的 WINCATI 软件自带数据核查功能,但由于数据采集完毕后还需要进行开放题的重编码、废卷清理等工作,为了保证数据质量,在提交分析数据之前重新按照问卷设定要求进行查错是必备步骤,本节就将介绍如何利用 SPSS 来完成相应的数据核查工作。

1. 数据核查的主要工作内容

一般而言,数据核查可分为如下几种情形。

(1) 配额检查:对于有配额限制的项目,需要检查项目设计中所规定的配额要求是否被满足。

(2) 数值检查——封闭题:具体选项有限,数值中不应当出现选项以外的取值,如变量 A3 ~ A16 只能取值为 1、2、3、4、9。

(3) 数值检查——数值开放题:相应的连续变量应在有效范围内取值,如变量 S3 年龄的取值应当在 18 ~ 65 之间。

(4) 数值检查——多选题:如果采用多重分类法记录数据,则同一个选项代码不应当在不同列中重复出现。例如,A3A_1 和 A3A_2 两个变量就不应当取相同的数值;否则就意味着同一个选项出现了重复选择。

(5) 逻辑查错:出于质量控制的要求,问卷中对数值题目的取值进行了逻辑控制,例如,A3、A4、A8 不应当同时选择 9,否则按废卷处理。

在 CCSS 项目中需要进行上述所有种类的数据核查,配额检查由于比较简单,这里不再进行详细讲述,而是主要针对数值检查和逻辑查错进行讲解。在 CCSS 问卷中对各题项的取值与逻辑关系都有着严格的规定,为了简化叙述,对这几种数据核查类型只给出上文中提到的几个变量的具体操作加以演示。

2. 数据核查的技术路线

无论采取哪种软件或者技术方式来实现,数据核查工作的技术路线是基本相同的,可大致分解为如下几个步骤。

(1) 内容分解:将各种查错工作归类为若干个基本独立的种类,实际上,上面所讲的就是在完成这项工作。

(2) 查错实现:对于每个分解出的类别给出适当的错误识别规则定义,并采用适当的技术手段来实现。

(3) 结果反馈:采用适当的技术手段作为查错结果的输出接口,从而使得反馈给用户的查错结果清楚并且格式统一。

实际上对于学习过软件工程的读者而言,上述基本思路 and 软件编制思路是完全一致的,有兴趣的读者可以按此直接编制一套查错软件,当然也可以用 SPSS 来实现相应的功能。

6.1.2 分析思路

由于 SPSS 并不是一个专业的数据录入和管理软件,所以在早期的 SPSS 版本中只提供了统计分析软件所需的完整数据管理功能,但专业的数据管理功能则没有提供,如有效值核查、逻辑查错、双份录入等(此类功能会在 DATAENTRY 等产品中提供)。但是随着市场竞争的日益激烈,以及产品线的分化重组,SPSS 也开始渐次下移,近年来推出的版本越来越向数据管理方向进行强化,从加入与数据字典相关的数个向导,到 18 版之后加入的数据验证模块均是这一发展趋势的延续。因此对于普通用户而言,使用 SPSS 独立完成数据管理工作已经变得切实可行。

具体而言,在 SPSS 中,要实现上述数据查错功能可以采用以下两种方式:购买了数据验证模块的用户可以直接利用该模块的相关功能来实现,而只拥有 BASE 模块的用户则可以利用 SPSS 函数和程序来加以实现。

1. 使用数据验证模块实现

由于数据验证模块提供了较为完善的验证与查错功能,因此利用该模块进行查错显得一点“技术含量”都没有,只要按相应要求设定好规则,然后在数据文件中应用相应的规则即可。但实际上,对于大多数用户而言,真正有价值的不是干活的过程是否有“技术含量”,而在于能否高效而精确地完成相应的工作任务,舍本逐末地去追求技术快感,那是统计软件爱好者才应当去做的事情。

2. 使用函数功能实现

实际上,认真研究就会发现,通过将数据查错分解为如上所示的封闭题、多选题等多种情形,就可以有针对性地找到每种情形下的实现方法,具体如下。

(1) 查错实现:在 SPSS 中提供了上百种函数,完全可以利用一些特殊的函数来对该个案的某个变量值是否违反查错规则做出逻辑判断,而当逻辑判断结果为真时,即意味着该个案的这一变量值可能存在错误。

(2) 结果反馈:可以考虑按照上述逻辑判断结果形成有特定含义的字符串,每一种错误都用相应的字符串表示,可以将该字符串直接输出到结果窗口中,也可以生成一个或数个专用的指示变量,当出现相应的错误时,就将所对应的字符串加入到错误指示变量中去。这里显然推荐后者,因为这样做之后,查错完成后只需要检查错误指示变量就可以得知相应案例的错误。

6.1.3 利用数据验证模块实现查错

借助于功能完善的数据验证模块,可以对数据查错的工作流程进行非常清晰的描述,具体如下。

1. 清空原有查错规则

这一步骤的目的是清除数据文件中原有的查错规则,避免因重名等原因影响随后的工作。本工作在使用定义验证规则的过程进行规则定义时会自动执行,也可以使用程序方式,操作如下。

删除现有的单变量验证规则:

```
DATAFILE ATTRIBUTE DELETE = $VD. SRule.
```

删除现有的交叉变量验证规则:

```
DATAFILE ATTRIBUTE DELETE = $VD. CRule.
```

2. 定义新的查错规则

在定义验证规则过程的协助下,按照需求给出所需的查错规则定义。初学者可以完全依赖于对话框完成此项工作,对于操作熟练的用户而言,这部分工作也可以用程序来实现。以第4章中涉及的规则 RuleS2、RuleS3 和交叉规则 CrossVarRule1 为例,其相应的程序代码如下。

定义(重新定义)单变量验证规则:

```
DATAFILE ATTRIBUTE ATTRIBUTE =
$VD. SRule[2] ("Label = RuleS2', Type = 'Numeric', Domain = 'List',
FlagUserMissing = 'No', "FlagSystemMissing = 'No', FlagBlank = 'No',
CaseSensitive = 'No', List = '1' '2' ").
$VD. SRule[1] ("Label = RuleS3', Type = 'Numeric', Domain = 'Range', Minimum = '18',
"Maximum = '65', FlagUserMissing = 'No', FlagSystemMissing = 'No',
FlagBlank = 'No', FlagNoninteger = 'No', FlagUnlabeled = 'No' ")
```

定义(重新定义)交叉变量验证规则:

```
DATAFILE ATTRIBUTE ATTRIBUTE =
$VD. CRule[1] ("Label = 'CrossVarRule 1', OutcomeVar = 'CrossVarRule1 _',
Expression = 'A3 = 9 & A4 = 9 & A8 = 9' ").
```

上述语句看似复杂,但实际上需要定义或修改的关键字只有 Label、Domain、List、Minimum、Maximum、Expression 几个,只需将对话框操作和程序内容相对应,就可以理解其含义。

3. 将规则存储为数据字典

规则定义完毕后,用户应当首先将其保存为一个数据字典文件,以便将来需要时使用。虽然确实可以将工作用的数据文件直接作为字典文件,但这并不符合数据管理的安全原则,因此不建议读者在较复杂的数据管理项目中这样做。

4. 加载定义完毕的规则

当需要使用时,只需要加载相应的字典文件即可,SPSS 加载预定义规则的对话框目前只能指定加载默认路径下 Predefined Validation Rules. sav 文件中的规则。如果用户的数据字典文件并非该数据文件,则只需使用该过程所对应的程序即可。例如,数据字典文件为 C:\Data-Dict. sav,则程序的具体格式如下:

```
APPLY DICTIONARY FROM 'C:\DataDict. sav'
/FILEINFO ATTRIBUTES = MERGE
/VARINFO.
```

5. 进行数据验证并报告结果

在规则加载完毕后,下一步工作自然是使用这些规则来进行数据验证。通过第4章中的验证数据对话框可以很好地完成验证、输出报告、存储指示变量等工作,操作也非常容易理解,这里不再赘述。

6.1.4 利用函数功能实现查错

数据验证模块虽好,但也是需要付费的,对于只拥有 Base 模块的用户而言,直接利用函数功能实际上就可以完成数据查错的绝大部分功能,甚至于可以完成数据验证模块也无法完成的查错任务。为了简化描述,这里直接给出每种情形所对应的函数解决方案。

1. 数值检查——封闭题

封闭题由于只有若干个特定取值,因此只需要判断相应取值是否有效即可,这可以使用 IF、RECORD 等命令来实现,但最方便的方式是使用专门的 ANY 函数,以变量 A3 为例,程序如下:

```
IF A3 ~=1 & A3 ~=2 & A3 ~=3 & A3 ~=4 & A3 ~=5 & A3 ~=9 ERROR =1.
RECODE A3 (1=0) (2=0) (3=0) (4=0) (5=0) (9=0) (ELSE=1) INTO ERROR.
COMP ERROR =1 - ANY(A3, 1, 2, 3, 4, 5, 9).
```

上面的变量 ERROR 为查错结果变量名,ANY 函数用于检查变量 A3 数值是否在给定的数值列表范围之内,返回值为 1 的时候恰恰就是数据正确的情形,从而相应的 ERROR 其变量值计算方式就需要将数值重新颠倒回来,以使得 1 能够代表案例出错的情形。

2. 数值检查——开放题

(1) 任意取值的连续变量取值范围查错:此类变量一般会存在一个合理的上界和下界,超出此范围的就可以作为可疑数据加以核对。以变量 S3 为例,可以采用如下 3 种方式来查错。

```
IF S3 <18 | S3 >65 ERROR =1.
RECODE S3 (18 THRU 65 =0) (ELSE =1) INTO ERROR.
COMP ERROR =1 - RANGE(S3, 18, 65)
```

但显然,按照上述查错方法,对于落在正常范围内的错误录入是无法核查的。

(2) 取值方式有限制的连续变量:此类变量除了上界和下界之外,只能取整数,或者某些特别的小数,这时可以使用下面的函数来实现查错功能。

① 为整数

```
IF RND(VAR) ~= VAR ERROR =1.
```

② 为特定的小数(如只能是*.3)

```
IF MOD( RND(VAR * 10) ) ~=3 ERROR =1.
```

③ 为某个数的倍数(如 3 的倍数)

```
IF MOD(VAR, 3) ~=0 ERROR =1.
```

当然,也可以使用 RECODE 和 COMP 语句实现相同的功能,具体方式读者可自行设计。

3. 多选题查错

多选题是一种特殊的题型,其记录方式在本质上是由多个相互关联的分类变量组成的,因此较为基本的查错方法为分别检查相应分类变量的取值。但是,这样做过于烦琐,完全可以结合其自身特点简化操作。由于多选题有两种记录方式,其查错技巧也不同,分述如下。

(1) 多重二分法:一般规定某种取值表示该题项被选中,其余取值均代表未被选中。因此可以检查上述题项所对应的变量是否均为相同的取值情形,以多选题 C0 为例,程序如下。

同时检查:

```
IF NOT ( ANY(C0_1, 1, 2, 99) & ANY(C0_2, 1, 2, 99) & ANY(C0_3, 1, 2, 99) )
  ERROR = 1.
```

分别检查:

```
IF ANY(C0_1, 1, 2, 99) ERROR = 1.
```

```
IF ANY(C0_2, 1, 2, 99) ERROR = 2.
```

```
IF ANY(C0_3, 1, 2, 99) ERROR = 3.
```

上述第一种方式无法区分具体是哪个变量出错,第二种方式可以直接标记出具体的出错变量,但需要事先定义好错误代码,如 2 代表的是 C0_2 出错。

(2) 多重分类法:除了进行类似于上面的取值范围检查外,多重分类法还有可能出现的错误是对选项进行了重复选择。这种情况常常是由于对“其他”选项进行重编码后,没有检查编码是否已经选中就将其加入了数据集中所致。CCSS 数据的 A3A 题目就是采用多重分类法加以记录的,虽然 A3A 的题目设定允许重复选择的情形出现,但也可以借用该题目演示相应的查错方式,程序如下:

```
IF MISSING(A3A_1) = 0 & (A3A_1 = A3A_2) ERROR = 1.
```

上面的 MISSING 函数用于过滤掉本题完全未答的个案,因为此时 A3A_1 和 A3A_2 相等(均缺失)是正常情况。

4. 逻辑关系查错

在正常的问卷中,各变量间往往存在某种逻辑关系,变量的取值受到其他变量取值的约束,这时可以利用该联系编制逻辑查错程序以减少错误的发生。例如,身高 152 cm,体重 140 kg 这两个变量值分别观察并无异常,但放在同一个受访者身上显然就需要引起注意了。

逻辑错误又可以被分为严格逻辑错误和可疑逻辑错误两种,前者有明确的错误界限,后者则没有,有可能的确是正确数值,但这两种逻辑错误在核查方法上是没有区别的。

逻辑关系的查错方式是利用已知的逻辑关系,直接编制相应的程序,主要使用 IF 和 COMP 实现,例如,对于 CCSS 问卷中的 A3、A4、A8 不应当同时选择 9 这一逻辑设定,可直接按如下方式设定。

```
IF A3 = 9 & A4 = 9 & A8 = 9 ERROR = 1.
```

5. 查错结果的报告

前面已经提到,查错结果虽然可以直接在结果窗口中进行输出,但最佳方式还是生成相应的

查错变量,具体而言有如下几种方式。

(1) 简单标识变量:只给出一个查错结果变量,用1或者某个数值表示该个案数据有错,但变量太多时,按此查找具体的错误显然非常费时。

(2) 单独重编码:比如共有12个查错条件组合,则为每个组合分别给出ERR1~ERR12这些变量,分别代表各自的查错结果。

(3) 错误代码字符串:可以将其看成是上面这种查错结果单独重编码的进一步改进,每种错误都用相应的字符串表示,当出现相应错误时,就将所对应的字符串加入到错误指示变量中去,最后只需要检查错误指示变量,就可以得知相应案例的错误情况。

STRING ERRSTR (A20).

IF S3 < 18 | S3 > 65 ERRSTR = CONCAT(RTRIM(ERRSTR), "S3").

注意上面程序中的第一句用于生成新的字符串变量,不可省略,且字符串的长度定义要足够长(在本例中设定为20)。在上述程序运行完毕后,凡是S3数值有问题的个案,其变量ERRSTR的取值终究都会包含字符串S3。

6.1.5 项目总结与讨论

SPSS的初学者经常会不自觉地陷入两个误区:希望SPSS的所有功能都可以在工作中找到用武之地,或者能用SPSS一步到位地解决工作中遇到的问题,这些显然是无法做到的。

实际上,不仅是SPSS,所有工具软件都无法完全满足上述要求。作为一个商业化的统计软件,SPSS必须提供较为完备的统计相关功能,用户只需从中发现有用的部分并加以妥善利用即可。而另一方面,对于相同的任务,可能会有许多种解决方案,用户也只需从中找到最为切实可行的一种加以实施即可。

具体到本项目,数据查错显然不是SPSS的强项,但是在没有其他更好的工具可用的情况下(无论是出于版权的原因还是熟悉程度的原因),都可以用该软件工具完成查错任务。本节一开始就理清了整个数据查错流程的技术路线,然后采用数据验证模块或者Base模块分别独立地完成这一工作。细心的读者会发现,只要遵循基本的技术路线,实际上工具的选择有时候并不重要,Base模块完成查错的效率并不见得会比数据验证模块低。这是因为工具的使用需要在正确的方法论指导下进行,或者说方法论正确与否要远比工具先进与否重要得多,否则只能是事倍功半。

6.2 CCSS项目数据的自动计算与处理

6.2.1 项目背景

在清理完CCSS项目数据之后,就可以进入数值分析阶段了。由于CCSS系统引进了美国密歇根大学的消费者信心指数项目体系,因此所需的分析处理工作内容非常明确。具体而言,首先需要根据样本的人口统计学背景资料分布进行个案权重的计算,然后将相应的数值题转化为题目得分(信心值),并将相应题目的信心值汇总为信心指数值,最后则进行出表汇总操作。下面

首先介绍数值计算部分的工作是如何完成的。

1. 计算题目得分

CCSS 问卷中的大多数主干题目均为五级得分,类似于非常好(VF)、比较好(F)、一般、比较差(U)、非常差(VU),以及不知道/拒答。此类题目都需要转换为相应的题目得分,以反映消费者的乐观/悲观程度。具体方式为针对每一道题目,计算每个选项被选中的百分比(包括“不知道/拒答”),随后使用以下公式计算其相对得分:

$$\text{题目得分} = 100\% + 1.0 \times \text{VF}\% + 0.5 \times \text{F}\% - 0.5 \times \text{U}\% - 1.0 \times \text{VU}\%$$

因此,这一数值反映的是答案偏向乐观的人群和偏向悲观的人群的比例之差,当人群中这两者的比例基本平衡时,得分接近于 100(100%);如果乐观人群比例偏高,则得分大于 100;反之,则小于 100。

2. 计算信心指数

消费者信心指数的计算是基于下面 5 道题目的回答进行的。

A3: 首先,请问与一年前相比,您的家庭现在的经济状况怎么样?

A4: 那么与现在相比,一年以后您的家庭经济状况将会如何变化?

A8: 那么与现在相比,您认为一年以后本地区的发展状况将会如何?

A10: 那么与现在相比,您认为 5 年之后,本地区的经济将会出现怎样的变化?

A16: 对于大宗耐用消费品的购买,如家用电器、家用计算机,以及高档家具之类的,您认为当前是购买的好时机吗?

首先计算出上述 5 题的题目得分,然后将其直接相加,再除以“基线”调查时的这一数值,即为当期的信心指数值。因此,所计算出的指数代表的是当期数值相对于“基线”调查数值的变动比例。如果乐观人群的比例高于“基线”,则指数大于 100,反之,则小于 100。目前作为基线水平的是 2007 年 4 月的数值。

实际上,上述指数算法和美国密歇根大学消费者信心指数的计算方法完全相同。

3. 其他数值题目的转换

除了上述信心指数相关题目的计算外,问卷中还有其他类型的数值封闭题,如家庭收入 S9,对于此类题目也需要进行重编码以进行均数汇总等操作,因操作比较简单,这里不再详细解释。

6.2.2 分析思路

首先,由于 CCSS 项目以月度为周期,所有相应的计算分析操作每月都要进行一次,因此采用程序实现相应的功能然后进行调用是必然的选择。

其次,由于项目的报表格式、个案权重大小等每月均不尽相同,需要人工加以干预,因此不考虑全自动的“生产工作”调用方式。而是采用半自动化的程序段方式,每一个程序段用于完成可以全自动执行的部分工作任务,然后人工完成无法自动执行的工作,然后再调用下一个工作程序段继续执行。

第三,由于大量的数值题目均采用相似的选项列表以及题目得分计算方式,因此完全可以利用宏代码的方式简化程序结构,使整个程序变得易于维护。

6.2.3 具体操作

1. 提取基本程序框架

以 A3 为例,相应的数值计算程序段如下:

```
RECODE A3
  (1 THRU 5 = COPY) (ELSE = 9) INTO TA3.
RECODE A3
  (1 = 200) (2 = 150) (3 = 100) (4 = 50) (5 = 0) (ELSE = 100) INTO QA3.
EXEC.
```

为了使程序清晰可读,这里规定以“T”开头的是出表用中间变量,“Q”开头的是数值计算用中间变量,因此 A3 会生成 TA3 和 QA3 两个中间变量用于进行后续分析。这种对中间变量的统一命名规范是增强程序可读性、减少出错概率的重要措施,应重视。

2. 宏代码编写

由于大部分问卷主干题目均需要执行类似于上述 A3 所需的数值计算程序,因此完全可以将其泛化为一个可调用的宏程序,最终使用的程序如下:

```
* 基本宏代码,计算 QSCORE.
DEFINE M_COMP ( INVAR1 = ! CHAREND('/') ).
RECODE ! INVAR1
  (1 THRU 5 = COPY) (ELSE = 9) INTO ! CONCAT('T',! INVAR1).
RECODE ! INVAR1
  (1 = 200) (2 = 150) (3 = 100) (4 = 50) (5 = 0) (ELSE = 100)
  INTO ! CONCAT('Q',! INVAR1).
EXEC.
! ENDDEFINE.

M_COMP INVAR1 = A3.
```

上述程序的运行结果完全等同于原始代码,但是已经具备了被重复调用的可能。注意为了使程序清晰可读,上述宏代码定义和调用程序写得比较啰嗦,熟悉其应用的读者完全可以采用更加简单的编写方式来完成同样的任务。

3. 在程序中实现信心指数的计算

计算消费者信心指数时需要用当前数值除以 2007 年 4 月时的“基线”数值,可能许多初学者在认真学习了本书第一部分的内容之后,对本问题详加考虑,并最终决定采取如下操作方式来实现。

- (1) 采用添加个案(Merge)过程,将当月数据与 2007 年 4 月的历史数据合并起来。
- (2) 采用分类汇总(Aggregate)过程,计算出当月数值与基线时期的 5 道题目的得分之和。
- (3) 采用数值平移(Shift Value)或者 LAG 函数,用当月数值除以基线数值,得出信心值。

当然,上述方式是最为严谨,也的确能得到所需结果的方式,但显然过于技术化了。分析者只需要注意到其中使用的“基线”数值其实是一个恒定值,不需要每次都重新计算,就可以将上

述操作彻底简化为如下两步。

(1) 用任何一种分析过程计算出基线时期的 5 道题目得分之和。

(2) 在程序中直接用当月数值除以基线时期的数值。

基于上述分析,最终实现程序如下:

```
COMP INDEX1 = (QA3 + QA4 + QA8 + QA10 + QA16)/5/1.2803.
```

```
EXEC.
```

其中的 1.2803 就是基线时期这 5 道题目的得分平均数。

4. 完成主程序

最终完成的数据计算程序段如下:

* 基本宏代码,计算 QSCORE.

```
DEFINE M_COMP ( INVAR1 = ! CHAREND('/') ).
```

```
RECODE ! INVAR1
```

```
(1 THRU 5 = COPY) (ELSE = 9) INTO ! CONCAT('T', ! INVAR1).
```

```
RECODE ! INVAR1
```

```
(1 = 200) (2 = 150) (3 = 100) (4 = 50) (5 = 0) (ELSE = 100)
```

```
INTO ! CONCAT('Q', ! INVAR1).
```

```
EXEC.
```

```
! ENDDDEFINE.
```

***** ----- 定型指数计算 ----- *****

```
M_COMP INVAR1 = A3.
```

```
M_COMP INVAR1 = A4.
```

```
M_COMP INVAR1 = A8.
```

```
M_COMP INVAR1 = A10.
```

```
M_COMP INVAR1 = A16.
```

* 计算信心指数.

```
COMP INDEX1 = (QA3 + QA4 + QA8 + QA10 + QA16)/5/1.2803.
```

```
VARIABLE LABELS INDEX1 "总指数".
```

```
EXEC.
```

***** ----- 定型指数计算结束 ----- *****

* 其他数值题目的计算.

```
RECODE S3
```

```
(15 THRU 34 = 1) (35 THRU 54 = 2) (55 THRU 70 = 3) (ELSE = 9) INTO TS3.
```

```
RECODE S9
```

```
(1 THRU 5 = 1) (6 THRU 13 = 2) (ELSE = 9) INTO TS9 .
```

```
RECODE S9
```

```
(1 = 500) (2 = 1250) (3 = 1750) (4 = 2500) (5 = 3500) (6 = 4500) (7 = 5500) (8  
= 7000)
```

```
(9 = 9000) (10 = 12500) (11 = 17500) (12 = 25000) (13 = 35000) INTO QS9 .
```

```
EXECUTE .
```

```
VALUE LABELS TS3 1 '18 - 34' 2 '35 - 54' 3 '55 +'
```

```
VALUE LABELS TS9 1 'BELOW 48,000' 2 'OVER 48,000' 9 'NA'.
```

在主程序中涉及的 S3、S9 等题目的计算程序由于只是 RECODE 等语句的依次书写,因此这里不再详细解释。

6.2.4 项目总结与讨论

菜单对话框是轻松愉快的操作方式,有对话框可用时没人会喜欢写程序,作者也不例外。但是在涉及具体的业务需求时,就必须加以认真权衡:显然在需要重复执行相同分析任务的背景下,程序方式就成了较好的选择。

在本项目中,由于每个变量的数值计算程序并不长,因此也可以直接编写而不使用宏程序,但是宏程序的好处在于:首先,宏程序的简洁性会很好地提高程序的可读性和可查错性;其次,宏程序由于会重复调用相同的程序段,因此出错概率很低,只要宏本身正确,则所有的调用程序都不会出错,而重复编写程序段就无法做到这一点;最后,如果将来需要对变量的标准计算方法加以修改,则只需要更改宏代码部分即可实现批量修改,显然效率会更高。

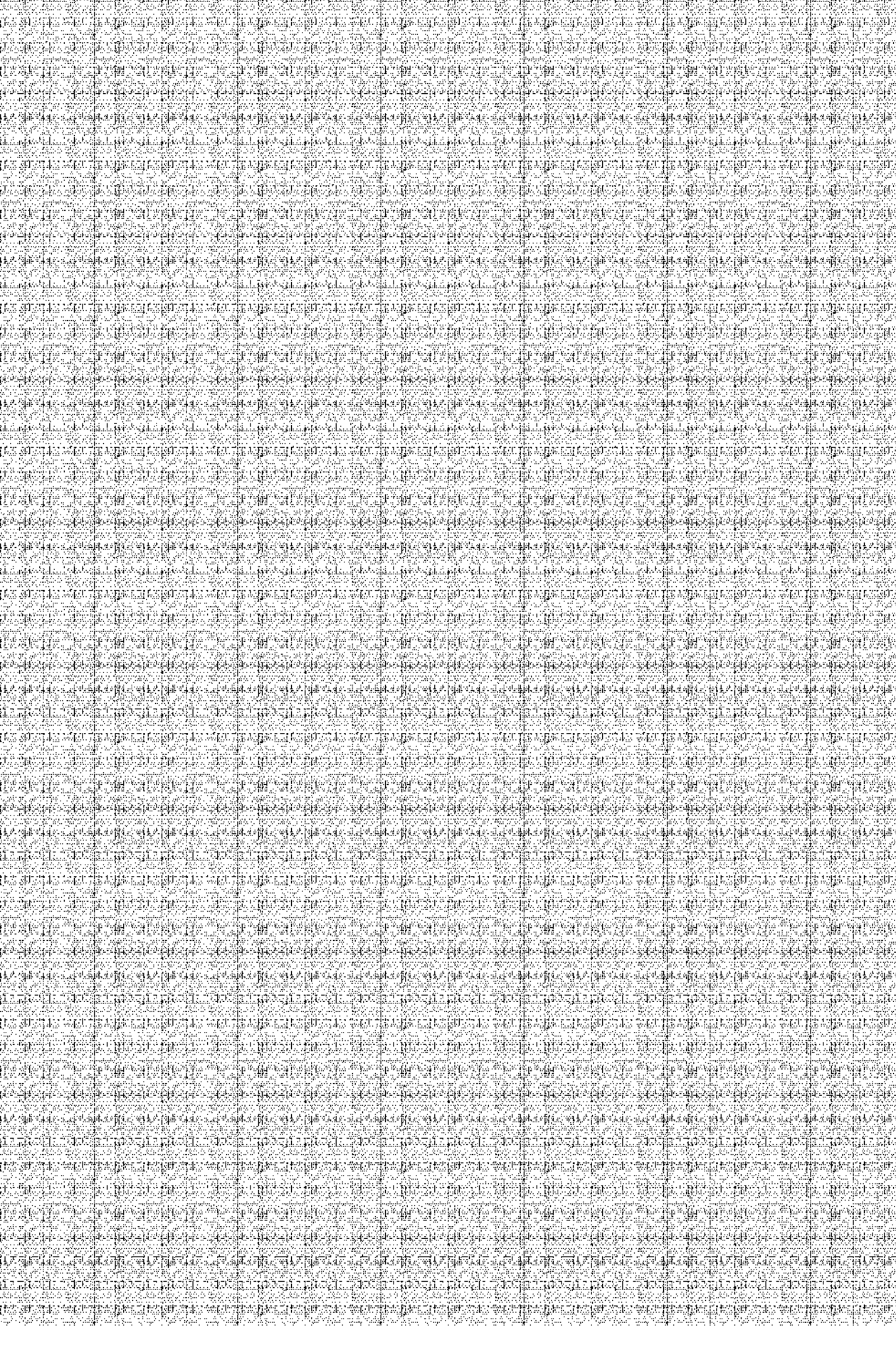
实际上,对于宏的使用完全可以灵活一些,例如在 CCSS 项目中,对于不具有普遍性的一些变量计算操作,或者说虽然具有普遍性,但相应的宏编制较为复杂的一些操作,作者就并未采用宏来执行。总而言之,工具是为业务服务的,工具的使用原则应当是使得业务操作更加便捷,而不是将工具不分场合、背景的尽量加以应用。

思考与练习

自行练习本章中涉及的案例数据操作。

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

在第一部分的章节中已经对 CCSS 项目数据完成了编码、录入、查错、合并存档等工作,那么,每个月具体的指数值在 SPSS 中究竟是如何计算出来的?这就涉及对相关变量进行统计描述以及参数估计的问题。

本章将介绍连续变量的统计描述与参数估计,而第 8 章将介绍无序分类变量、有序分类变量和多选题变量集的统计描述与参数估计。

7.1 连续变量的统计描述指标体系

当数据量较少,如只有 5 个人的身高,或者 7 个人的性别资料时,研究者可以通过直接观察原始数据来了解几乎所有的信息。但是,在实际工作中所接触到的数据量往往要远大于人脑可以直接处理、记忆的容量,此时最直接的方法是将原始数据按照其大小分组汇总,计算各组段的频数大小,最终汇总成相应的分组频数表(或直方图),以反映数据的大致趋势。

图 7.1 是对 CCSS 案例数据的年龄 S3 绘制的直方图,通过对这张图形的观察可以发现,如果要使用统计指标对该年龄变量加以描述,则主要是表现以下几个方面:集中趋势(Central Tendency)、离散趋势(Dispersion Tendency)、分布特征(Distribution Tendency),以及其他趋势,下面就分别介绍其所用的统计指标。

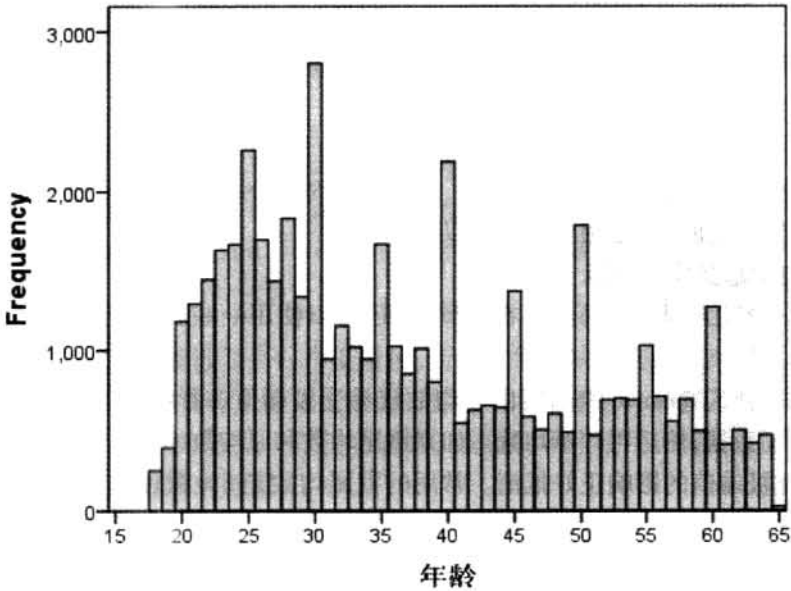


图 7.1 年龄的直方图

7.1.1 集中趋势的描述指标

人群的平均年龄可能是人们希望了解的最基本的汇总信息,在统计学中用于描述集中趋势,或者数据分布的中心位置的统计量就被称为位置统计量(Location Statistic)。针对不同的数据分布状况,统计学家提供了多种统计量来代表原始数据的中心趋势,如平均值、中位数和众数等。

1. 算术均数

算术均数(Arithmetic Mean)是最常用的描述数据分布的集中趋势的统计指标,因此也往往将其直接简称为均数。总体均数(Population Mean)用希腊字母 μ 表示,样本均数常用 \bar{X} 表示。对一组数据 X_1, \dots, X_n 而言,均数的算法为各数据直接相加,再除以例数 n ,即

$$\bar{X} = \sum X/n$$

均数是最常用的集中趋势描述指标,但它不适用于对严重偏态分布的变量进行描述,只有分布资料单峰和基本对称时使用均数作为集中趋势描述的统计量才是合理的。



均数误用最常见的实例就是平均工资,假设某单位有6个人,其中5个员工,1个经理。员工的月收入分别是360元、380元、400元、420元、440元,经理的月收入为40000元,这样他们的月收入均数为7000元。显然此时用均数并不能准确地反映其收入的一般水平,中位数才是更妥当的指标。

2. 中位数

中位数(Median)是将全体数据按大小顺序排列,在整个数列中处于中间位置的那个值。它把全部数值分成两部分,比它小和比它大的数值个数正好相等,具体而言:

(1) 当 n 为奇数时, $M = X_{(n+1)/2}$;当 n 为偶数时, $M = (X_{n/2} + X_{n/2+1})/2$ 。

(2) 由于中位数是位置平均数,因此不受极端值的影响,在具有个别极大值或极小值的分布数列中,中位数比算术平均数更具有代表性。例如上面员工收入的例子,其中位数是410元,显然要比均数更能够代表数据的集中趋势。

(3) 中位数适用于任意分布类型的资料,不过,由于中位数只考虑居中位置,对信息的利用不充分,当样本量较小时数值会不太稳定。因此对于对称分布的资料,分析者会优先考虑使用均数,只有当均数不能使用时才用中位数加以描述。

3. 其他集中趋势描述指标

除了上述最常用的两种指标外,在SPSS中还可以使用一些较为复杂和专业的统计描述指标,简介如下。

(1) 截尾均数(Trimmed Mean):由于均数较易受极端值的影响,因此可以考虑按照一定比例去掉最两端的数据,然后再计算均数。如果截尾均数和原均数相差不大,则说明数据不存在极端值,或者两侧极端值的影响正好抵消。常用的截尾均数是5%截尾均数,即两端各去掉5%的数据。在SPSS中Explore过程可以自动计算5%截尾均数。

(2) 几何均数(Geometric Mean):几何均数用 G 表示,适用于原始数据分布不对称,但经对数转换后呈对称分布的资料。例如医学中的血清滴度资料就常用几何均数描述其分布的集中趋势。其计算公式是: $G = \sqrt[n]{X_1 X_2 \cdots X_n}$,或者 $G = \lg^{-1}(\sum \lg X/n)$ 。可以发现,几何均数实际上就

是进行对数转换后的数据 $\lg X$ 的算术均数的反对数。在 SPSS 中,几何均数可以在 Report 子菜单的报表过程中计算输出。

(3) 众数 (Mode): 众数指的是样本数据中出现频次最大的那个数,众数容易理解,也不受极端值影响,但不易确定,且没有太明确的统计特性,一般很少使用该指标。在 SPSS 中,众数可以在 Report 子菜单和 Tables 子菜单的全部报表过程和制表过程中计算输出。

(4) 调和均数 (Harmonic Mean): 调和均数用符号 H 表示,现在已经很少使用,它实际上是观察值 X 的倒数均数的倒数,常用于完成的工作量相等而所用的时间不同的情况,主要用来求平均速度。实际上,中学物理课程中讲过的并联电路的总电阻就是各分电路电阻的调和均数,各原始数据的大小相差越悬殊,该均数的“调和”作用就越明显。在 SPSS 中,调和均数可以在 Report 子菜单的报表过程中计算输出。

7.1.2 离散趋势的描述指标

显然,仅仅反映数据的集中趋势是远远不够的,图 7.1 还反映出年龄的波动范围为 18 ~ 65 岁,这被称为数据的离散趋势。描述该趋势的统计量称为尺度统计量 (Scale Statistic),常用的尺度统计量有全距、方差、标准差、四分位数间距等。

1. 全距

全距 (Range) 又称为极差,是一组数据中最大值与最小值之差,是最简单的变异指标,但显然过于简单了,因此全距一般只用于预备性检查。

2. 方差和标准差

对于每个数据而言,其离散程度的大小就是和均数的差值,简称离均差,而总体方差就是用离均差平方和除以观察例数 n :

$$\sigma^2 = \sum (X - \mu)^2 / n$$

对于样本数据而言,方差 (Variance) 的计算公式有所不同:

$$S^2 = \sum (X - \bar{X})^2 / (n - 1)$$

其中的 $n - 1$ 称为自由度 (Degree of Freedom),用符号 ν 表示。

但是,方差在使用上存在不便,就是量纲不合常理,是原始指标量纲的平方,为此又将方差开平方,这就是所谓的标准差 (Standard Deviation),总体和样本的标准差分别用 σ 和 s 来表示:

$$\text{总体标准差 } \sigma = \sqrt{\sum (X - \mu)^2 / n}, \text{ 样本标准差 } S = \sqrt{\sum (X - \bar{X})^2 / (n - 1)}$$

由于标准差和方差的计算涉及每一个变量值,所以它们反映的信息在离散指标中是最全的,是最理想、最可靠的变异描述指标。但也正是由于标准差和方差的计算涉及每一个变量值,所以它们也会受到极端值的影响,当数据中有较明显的极端值时不宜使用。实际上,方差和标准差的适用范围应当是服从正态分布的数据。

3. 百分位数、四分位数与四分位间距

百分位数 (Percentile) 是一种位置指标,用 P_x 表示。一个百分位数 P_x 将一组观察值分为两部分,理论上有 $x\%$ 的观察值比它小,有 $(100 - x)\%$ 的观察值比它大。前面所讲的中位数实际上就是一个特定的百分位数,即 P_{50} 。

除中位数外,常用的百分位数还有四分位数,即 P_{25} 、 P_{50} 和 P_{75} 分位数的总称。这 3 个分位数正好是能够将全部总体单位按标志值的大小等分为 4 部分的 3 个数值,且 P_{25} 和 P_{75} 这两个分位数间包括了中间 50% 的观察值,因此四分位间距既排除了两侧极端值的影响,又能够反映较多数据的离散程度,是当方差、标准差不适用时较好的离散程度描述指标。



严格地讲,百分位数并不应当被限于只描述离散程度,显然,也可以对数据的集中趋势等其他特征进行描述,而将多个百分位数联合起来,实际上就可以完整地反映整个数据的分布规律。

4. 变异系数

当需要比较两组数据离散程度大小的时候,往往直接使用标准差来进行比较并不合适,存在以下两种情况。

(1) 测量尺度相差太大:例如希望比较蚂蚁和大象的体重变异,直接比较其标准差显然是不合理的。

(2) 数据量纲不同:例如希望比较身高和体重的变异,两者的量纲分别是米和千克,那么,究竟是 1 m 大,还是 2 kg 大? 根本没法比较。

在以上情形中,就应当消除测量尺度和量纲的影响,而变异系数(Coefficient of Variation)就可以做到这一点,它是标准差与其平均数的比:

$$CV = S/\bar{X}$$

CV 显然没有量纲,同时又按照其均数大小进行了标准化,这样就可以进行客观比较了。

7.1.3 分布特征、其他趋势的描述指标

除了以上两大基本趋势外,随着对数据特征了解的逐渐深入,研究者常常会提出假设,认为该数据所在的总体应当是服从某种分布的。那么,针对每一种分布类型,都可以采用一系列的指标来描述数据偏离分布的程度。例如对于正态分布而言,偏度系数、峰度系数就可以用来反映当前数据偏离正态分布的程度。当然,相对而言,这些分布指标使用得较少。

由于所假定的分布不同,所使用的分布特征描述指标也会有所差异,这里只简单介绍和正态分布有关的偏度系数和峰度系数的概念。

1. 偏度

偏度(Skewness)是用来描述变量取值分布形态的统计量,指分布不对称的方向和程度。样本的偏度系数记为 g_1 :

$$g_1 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3 / s^3$$

这是根据矩法(详情见后面)测定分布偏度的计算公式。测定分布偏度的其他方法还有分位数法和 Pearson 规则等,这里不做介绍,读者可以参考有关专业书籍。偏度是与正态分布相比较而言的统计量。当 $g_1 > 0$ 时分布为正偏或右偏,即长尾在右,峰尖偏左; $g_1 < 0$ 时分布为负偏或左偏,即长尾在左,峰尖偏右; $g_1 = 0$ 时分布为对称分布。



需要特别提醒的是,偏态的方向指的应是长尾的方向,而不是高峰的位置。国内的不少统计书籍对左/右偏态的理解有误,正好弄颠倒了。

2. 峰度

峰度(Kurtosis)是用来描述变量取值分布形态陡缓程度的统计量,是指分布图形的尖峭程度或峰凸程度。样本的峰度系数记为 g_2 :

$$g_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4 / s^4 - 3$$

这也是根据矩法测定分布峰度的计算公式,测定分布峰度的方法还有分位数法(略)。峰度也是与正态分布相比较而言的统计量,当 $g_2 > 0$ 时峰的形状比较尖,比正态分布峰要陡峭;当 $g_2 < 0$ 时峰的形状比正态分布要平坦;当 $g_2 = 0$ 时分布为正态峰。

3. 其他趋势的描述指标

在统计描述中还可能需要描述一些上面未提到的数据趋势,如数据是呈单峰还是双峰分布,数据是否存在极端值等,常用的有专门针对异常值数据进行描述的极端值(Outlier)列表等,详见后面介绍。

7.1.4 SPSS 中的相应功能

SPSS 的许多模块均可完成统计描述的任务,除了各种用于统计推断的过程会附带进行相关的统计描述外,还专门提供了几个用于连续变量统计描述的过程,它们均集中在“描述统计”(Descriptive Statistics)子菜单中。

1. 频率过程

频率过程的特色是产生原始数据的频数表,并能计算各种百分位数。由图 7.2 可见,它所提供的统计描述功能非常全面,且对话框布置很有规律,基本上按照数据的集中趋势、离散趋势、百分位数和分布指标四大块将各描述指标进行了归类。有了上面的基础,读者使用它应当不存在任何的困难。

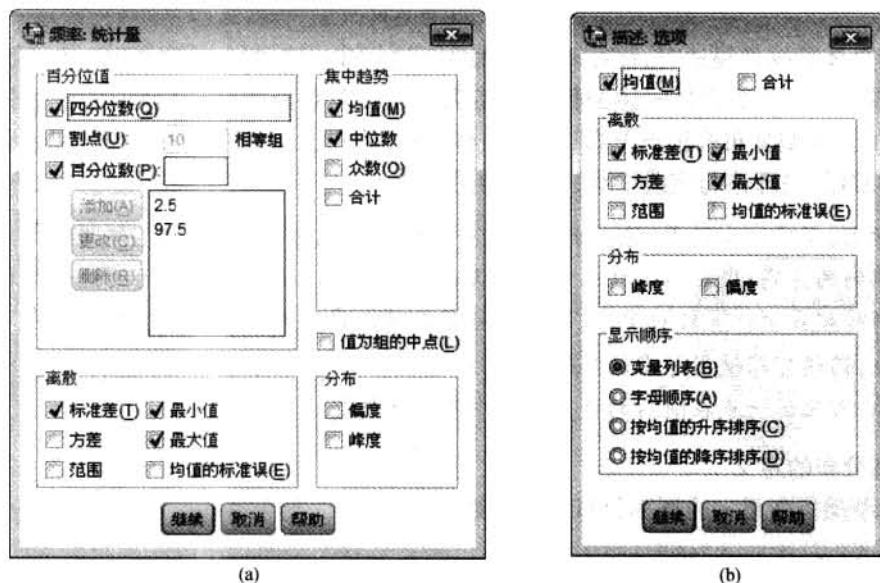


图 7.2 “频率:统计量”和“描述:选项”对话框

除了统计指标外,频率过程还可以为数据直接绘制相应的统计图,如用于连续型变量的直方图,用于分类变量的饼图和条图等。

2. 描述过程

描述(Descriptive)过程用于进行一般性的统计描述,相对于频率过程而言,它不能绘制统计图,所能计算的统计量也较少,但使用频率却是最高的。实际上从统计选项可以看出,该过程适用于对服从正态分布的连续性变量进行描述。

3. 探索过程

探索(Explore)过程用于在连续性资料分布状况不清时进行探索性分析,它可以计算许多描述统计量,除常见的均数、百分位数之外,还可以给出截尾均数、极端值列表等,并绘制出各种统计图,是功能最为强大的一个描述过程。

4. 比率过程

比率(Ratio)过程的功能比较特殊,用于对两个连续性变量计算相对比指标,除中位数、均值、加权均值等常见指标外,还可以计算出一系列专业指标,如离差系数(COD)、以中位数为中心的变异系数、以均值为中心的变异系数、价格相关微分(PRD)、平均绝对偏差(AAD)等。但由于这些指标在实际工作中应用较少,因此本书将不对它做过多介绍,对此感兴趣的读者可参见《SPSS11 统计分析教程》(基础篇)。

7.2 连续变量的参数估计指标体系

通过统计描述,研究者已经可以对样本数据的情况有了详细的了解。但研究的真正目的是考察样本所代表的总体情况如何,下面就来介绍如何进行连续变量的参数估计。

7.2.1 正态分布

在进行总体数据的描述时,往往首先对该总体的分布规律进行一定的假定,如假定年龄服从正态分布,这样就可以将总体描述的任务归结为对几个参数值的估计(此即参数估计名称的由来)。常见的连续型分布有正态分布、均匀分布、卡方分布、 t 分布和 F 分布等,其中以正态分布最为重要和常用,在理论与实践中都占有重要的地位。



实际上,在现实生活中,绝对服从正态分布的变量几乎是不存在的,包括统计书中最常用来举例的身高,现在其实也已经不服从正态分布了。由于许多常用的统计指标和统计方法都对此具有一定的耐受力(统计上称其为结果稳健的),因此只要偏离程度不影响分析结论,仍然可以使用原有的“正统”方法,否则必须采用其他方法。本书因为非常突出实战性,因此这一点在随后的许多章节中都会反复提及。

1. 正态分布的定义

若连续型随机变量 x 的概率分布密度函数为

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

其中, μ 为平均数, σ^2 为方差,则称随机变量 x 服从正态分布(Normal Distribution),记为 $x \sim N(\mu,$

σ^2)。不同的 μ 、不同的 σ ,对应于不同的正态分布。

正态分布曲线是一条对称曲线,关于均数($x=\mu$)对称,因此均数被称为正态分布的位置参数,而该曲线的高矮形状则与标准差有关。标准差越大,个体差异越大,正态曲线也越矮阔;反之,标准差越小,个体差异越小,正态曲线也越尖峭。因此标准差被称为正态分布的尺度参数。除此以外,正态曲线下的面积也有一定的分布规律,例如约有95%的个体的取值与平均数的距离在1.96个标准差($\mu \pm 1.96\sigma$)之内,据此可以做出一些相应的总体推断。

2. 标准正态分布

均数为0、标准差为1的正态分布称为标准正态分布(Standard Normal Distribution, SND),对于其他的正态分布,则可以通过使用以下变换将其转换为SND:

$$u = \frac{X - \mu}{\sigma}$$

该变换称为标准正态变换。在国外,标准正态分布被称为 u 分布或者 z 分布,因此变换也被称为 u 变换或者 z 变换。

标准化变换和标准正态分布的意义非常重大,因为在统计分析中经常要求曲线下面积,有了上面的变换方法,则只需要知道标准正态曲线下面积的分布规律,就可以解决所有正态分布的曲线下面积计算问题了。

7.2.2 参数的点估计

参数的点估计就是选定一个适当的样本统计量值作为参数的估计值,如将样本均数作为总体均数的点估计值。对于所选统计量是否适于作为参数估计量,有无偏性、一致性和有效性3个原则。

(1) 无偏性:虽然估计量的值不全等于参数,但应当在真实值附近摆动。

(2) 一致性:样本量越大,估计值离真实值的差异应当越小。

(3) 有效性:如果有两个统计量都符合上述要求,则应当选取误差更小的一个作为估计值。例如均数和中位数,实际上两者在反映正态分布的集中趋势时,在无偏性和一致性上是一样好的,但中位数误差更大,所以应当尽量使用样本均数来反映正态分布集中趋势。

参数点估计可用的方法有矩法和极大似然法两种,下面分别进行介绍,Bootstrap方法由于不属于经典统计学的方法体系,因此将在后面单独介绍。

1. 矩法

矩法的名称比较专业,实际上含义非常简单,它指的是在许多情况下,样本统计量本身往往就是相应的总体参数的最佳估计值,此时就可以直接取相应的样本统计量作为总体参数的点估计值。例如样本均数、方差、标准差都是相应总体均数、方差、标准差的矩估计量。对于常用的正态分布,矩法几乎可以满足全部参数的点估计需求,所以平常书上所说的点估计实际上用的就是矩法。

2. 极大似然法

极大似然法是另一种更好的参数估计方法,其优点在于估计量通常能满足一致性、有效性等要求,且具有不变性。不变性是指当原始数据进行某种函数变换后,相应估计量的同一函数变换值仍是新样本的极大似然估计量。

该方法的原理是在已知总体分布,但未知其参数值时,在待估参数的可能取值范围内进行探索,使似然函数值(在参数所确定的总体中获得现有样本的概率)最大的那个数值即为极大似然估计值。

因极大似然法已超过本书读者需要了解的范畴,这里将不再深入讨论,读者只需要知道还有这样一个点估计的方法即可。

3. 稳健估计值

矩法和极大似然法虽然能够很好地满足点估计的需要,但也有很明显的缺陷,就是估计值受异常值的影响十分显著,或因数据分布的偏离而使估计值产生较大变化。稳健估计方法就是针对这种情况的解决方案之一,即当观测数据不符合假定模型,与假定模型有偏离时,分析结论仍然保持稳定并正确的统计方法。而稳健估计指的就是该统计量受数据异常值的影响较小,而且对大部分的分布而言都很好(当然,这种特征意味着它不会对每个分布都是最佳的)。

稳健估计有 M 估计、R 估计等不同方法,前者是稳健估计常用的方法。M 估计最早是由尤伯提出的,其实是“极大似然型估计”的简称,即该方法的核心仍然是极大似然法,但是在估计时首先要构造一个 ψ 函数,该函数能够减小异常值的影响,而且对所考虑的分布集合中的每个分布都是好的估计量。随后再对 ψ 函数的集中趋势进行参数的极大似然估计,因此相应的估计值受异常值的影响要小得多。

7.2.3 参数的区间估计

显然,仅仅有参数的点估计是不够的,比如打靶,打了 2 枪,平均 9 环;打了 100 枪,平均也是 9 环,显然人们更相信后者确是一个好的枪手,而对前者的水平会产生很大的怀疑。这就涉及参数的估计值究竟有多大的误差的问题。

1. 标准误

虽然原始数据可能服从各种各样的分布,但是根据中心极限定理,当样本量 n 足够大(如 $n > 50$)时,其抽样均数都会近似服从正态分布,而此正态分布所对应的标准差就可用来表示抽样误差的大小,此即标准误(Standardized Error)。



标准误是最常见的用来描述参数估计值和真实值相差多大的统计量,注意其英文原文和标准差的区别,标准差的 Deviation 说明该指标表示的是“偏差”,而标准误的 Error 说明该指标表示的是“误差”,即进行参数估计时可能的错误大小,标准误越大,则说明相应参数的点估计值越不可信。

2. 区间估计的计算

结合样本统计量和标准误可以确定一个具有较大的可信度(如 95% 或 99%)包含总体参数的区间,该区间称为总体参数的 $1 - \alpha$ 可信区间或置信区间(Confidence Interval, CI)。

下面来看一下可信区间是如何求得的,以最常用的 95% 双侧可信区间为例,其计算公式为

$$\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}$$

上述公式看起来很完美,但有一个大问题,就是 σ 也是未知总体参数,计算时必须使用样本标准差 s 来代替,因此必须对公式加以修正,统计学家发现此时样本均数 \bar{X} 按照前述标准化公式变换后服从的是 t 分布而不是 u 分布,相应的可信区间公式修改为

$$\bar{X} - t_{\alpha, \nu} s / \sqrt{n} < \mu < \bar{X} + t_{\alpha, \nu} s / \sqrt{n}$$

上述公式就是最常用的可信区间计算公式,显然在使用中 t 分布的界值需要根据自由度 ν 来确定,非常麻烦,用 SPSS 进行分析时会直接完成这些工作,使用者只需要了解如何阅读结果即可。



必须指出,可信度的概念往往会引起误解,它仅仅是进行大量重复抽样时的一个渐近概念。认为“95%的可信区间包括真实参数值的概率为 0.95”是错误的。这里得到的区间是固定的,而总体参数值也是固定的。因此只有两种可能:包含或者不包含,这当中没有任何概率可言。95%的可信度只是说如果能够进行大量重复试验,则平均下来在所计算的每 100 个可信区间中,会有大约 95 个覆盖真实值。

7.2.4 SPSS 中的相应功能

SPSS 的许多过程均可完成连续变量参数估计的任务,如 7.2.3 小节介绍的几个过程均可计算标准误。但针对性较强的是描述统计子菜单中的以下几个过程。

1. 描述过程

描述(Descriptive)过程较为特殊的一个功能是将原变量变换为标准正态分布下的得分,只需要选中主对话框左下角的“将标准化得分另存为变量”复选框即可。

2. 探索过程

探索(Explore)过程不仅会计算标准误,还可以直接给出均数 95% 可信区间,而对于均数的点估计,还可直接提供稳健估计值,显然要更为专业。

3. P-P 图和 Q-Q 图

这两个过程用图形方式来直接观察样本数据分布是否服从所假设的理论分布,详见第 10 章的介绍。

7.3 案例:信心指数的统计描述

在系统介绍了连续变量的统计描述指标体系后,下面将用 CCSS 的实际数据来说明各种描述指标在 SPSS 中的实现方法。

7.3.1 使用频率过程进行分析

例 7.1 对 CCSS 数据中的消费者信心总指数 index1、现状指数 index1a 和预期指数 index1b 进行统计描述,并计算出 95% 个体参考值范围。

本例要求计算出 95% 个体参考值范围,个体参考值范围可以用百分位数法和正态分布法两种方法加以计算。由于目前尚不了解 index1 是否服从正态分布,且样本量较大,因此可以考虑使用频率过程计算出 P2.5 和 P97.5 的数值,这就是百分位数法得出的 95% 个体参考值范围的上下界。

1. 界面说明

选择“分析”→“描述统计”→“频率”菜单项,打开“频率”对话框,如图 7.3 所示,该对话框

上面的内容非常容易理解,下面简要介绍各部分的功能。

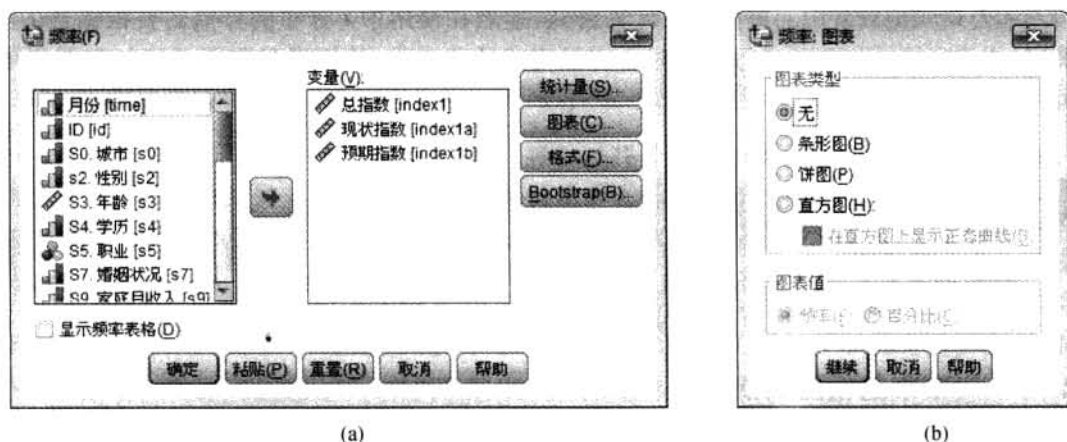


图 7.3 “频率”对话框和“频率:图表”子对话框

(1) 主对话框:“变量”列表框用于选入需要进行描述的变量,如果选入多个,系统会依次对其进行分析。左下角的“显示频率表格”复选框用于输出频数表,默认选中。

(2) “统计量”按钮:单击后打开的对话框用于定义需要计算的描述统计量,包括集中趋势、离散趋势、分布特征和百分位数 4 组,比较特殊的是右侧的“值为组的中点”复选框,当输入的数据是分组频数数据,并且具体数值是组中值时,需要选中该复选框,这样 SPSS 在计算各种百分位数的时候会将数据按频数表对待,而不会认为同一组内的数据取值都是组中值的大小。

(3) “图表”按钮:单击后打开的对话框用于设定所做的统计图,可以绘制分类数据描述用的条图和饼图,也可以绘制连续变量描述用的直方图,相关的图形知识参见第 10 章。

(4) “格式”按钮:单击后打开的对话框用于定义输出频数表的格式,不过一般不用更改,使用默认设置即可。

(5) “Bootstrap”按钮:单击后打开的对话框可使用 Bootstrap 这种计算统计学方法进行任意总体参数的估计,详见 7.4 节的介绍。

2. 操作说明与结果解释

根据题目要求,本例操作如下。

(1) 将 index1、index1a 和 index1b 选入“变量”列表框,取消左下方的“显示频率表格”复选框(因为本例不需要)。

(2) 单击“统计量”按钮进入“统计量”子对话框,选中所需的常用统计量,并且在百分位数中设定输出 P2.5 和 P97.5,最终对话框界面应当如图 7.2 所示。

本例的输出结果如图 7.4 所示,可见总信心指数的均数和中位数非常接近,而根据百分位数法计算出的 95% 个体参考值范围为 46.86 ~ 132.78。大家如果有兴趣,可以利用均数和标准差计算出正态分布下的 95% 个体参考值范围是 54.74 ~ 137.05,显然和百分位数法的结果差异并不太大,这些信息都在暗示 index1 的分布可能是大致对称的。而用同样的方式可以发现现状指数的分布可能略呈偏态分布。

		总指数	现状指数	预期指数
N	有效	1147	1147	1147
	缺失	0	0	0
均值		95.8935	99.2227	94.0598
中值		93.7280	88.0359	96.8570
标准差		20.99710	28.43333	23.11645
极小值		.00	.00	.00
极大值		156.21	176.07	145.29
百分位数	2.5	46.8640	44.0180	48.4285
	25	85.9174	88.0359	84.7499
	50	93.7280	88.0359	96.8570
	75	109.3494	110.0449	108.9641
	97.5	132.7814	154.0629	133.1784

图 7.4 统计量

7.3.2 使用描述过程进行分析

下面使用描述过程来对 index1、index1a 和 index1b 这 3 个变量进行分析,来看看题目的要求是否能完全得到满足,以及两个过程的输出形式有何不同。

1. 界面说明

选择“分析”→“描述统计”→“描述”菜单项,就会打开描述过程的主对话框,如图 7.5 所示。

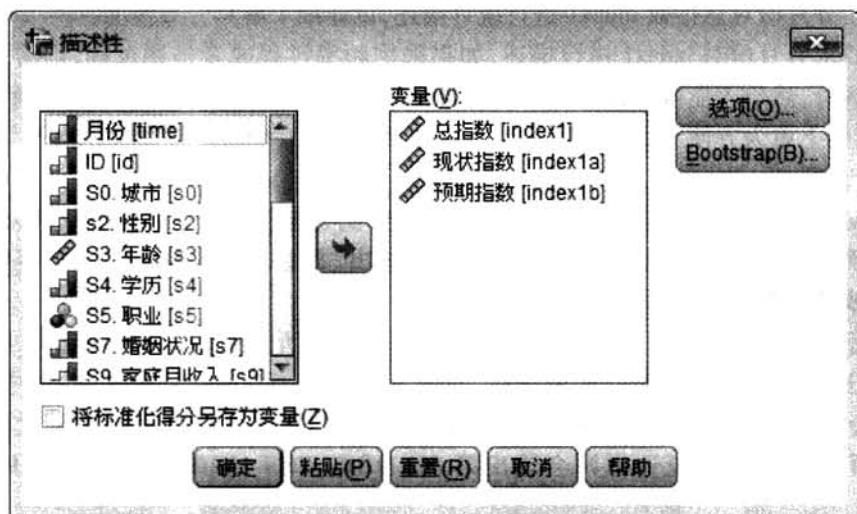


图 7.5 描述过程的主对话框

(1) 主对话框:“变量”列表框用于选入需要进行描述的变量,如果选入多个,系统会在同一张表格内输出描述结果。选中下方的“将标准化得分另存为变量”复选框会在数据集中生成一

个新的变量,该变量自动命名为“Z + 原变量名”,大小即为原变量的标准正态变换结果,详见标准正态分布的内容。

(2) “选项”按钮:单击后打开的子对话框用于设定描述统计量,显然其功能要比频率过程中的相应子对话框少很多,实际上这些统计量均只适用于正态分布资料。

(3) “Bootstrap”按钮:单击后打开的子对话框和频率过程中的子对话框完全相同,不再重复说明。

2. 操作说明与结果解释

该过程的操作非常简单,只需要将希望描述的变量选入即可,由于描述过程无法输出百分位数,因此个体参考值范围无法计算,最终本例的分析结果如图 7.6 所示。

	N	极小值	极大值	均值	标准差
总指数	1147	.00	156.21	95.8935	20.99710
现状指数	1147	.00	176.07	99.2227	28.43333
预期指数	1147	.00	145.29	94.0598	23.11645
有效的 N (列表状态)	1147				

图 7.6 描述统计量

由图 7.6 可见,这里的大部分内容都在上面见过,因此就不再多解释了。但是很显然,在同时描述多个变量时,描述过程会以一种紧凑的表格形式将正态分布资料常用的统计量一并输出,显然非常简洁。

7.3.3 使用探索过程进行分析

例 7.2 分月份对总指数 index1 进行统计描述,以详细了解其分布情况。

本例中要求分月份对 index1 进行描述,如果采用频率过程或者描述过程,则需要首先对数据文件进行拆分,然后才能得到相应的分析结果。而使用探索过程时可以直接得到这种分组的分析结果,使用上更为方便。

1. 界面说明

(1) 主对话框:“因变量列表”用于选入需要分析的变量,“因子列表”用于选入分组变量,“标注个案”列表框用于选入标签变量,而下方的“输出”框组用于选择结果中是否包含统计描述、统计图,或者两者均包括,如图 7.7(a)所示。

(2) “统计量”按钮:单击后打开的子对话框用于选择所需要的描述统计量。默认选中的“描述性”复选框可以输出一系列常用指标,如图 7.7(b)所示,详见分析实例;选中“M - 估计量”复选框会给出集中趋势的最大稳健估计值;“界外值”复选框会输出 5 个最大值与 5 个最小值备查;而“百分位数”复选框则会输出第 5%、10%、25%、50%、75%、90%、95% 分位数备查。

(3) “绘制”按钮:单击后打开的子对话框用于选择所需要的统计图。“箱图”框组用于设置绘制分组箱图或者单一箱图;“描述性”框组用于设置绘制茎叶图和直方图;选中中部“带检验的正态图”复选框则可以绘制正态分布的 QQ 图,并进行变量是否符合正态分布的 K - S 检验;而最下方的“伸展与级别 Levene 检验”(Spread vs. Level with Levene Test)框组则用于设置当存在分



(a)



(b)

图 7.7 “探索”对话框和“探索:图”子对话框

组变量时,可自动判断各组间的离散程度是否相同,并为此寻求一个比较合适的变量变换方法。具体会输出分布——水平图,给出回归直线斜率,并进行稳健的 Levene 方差齐性检验。茎叶图、直方图和 QQ 图等的介绍参见第 10 章,而分布——水平图及相关功能的实用价值不大,因此本书将不再对其进行深入介绍。

(4) “选项”按钮:单击后打开的对话框主要用于控制存在缺失值时的处理方式,一般不用更改。

2. 基本输出结果

按照图 7.7 中的变量选择方式,SPSS 会给出如图 7.8 所示的分析结果。

月份		统计量		标准误
总指数	200704	均值	98.3363	1.09239
		均值的 95% 置信区间	下限	96.1866
			上限	100.4861
		5% 修整均值	98.9930	
		中值	101.5387	
		方差	357.994	
		标准差	18.92074	
		极小值	31.24	
		极大值	140.59	
		范围	109.35	
		四分位距	23.43	
		偏度	-.535	.141
		峰度	.768	.281

图 7.8 变量的统计描述表格

图 7.8 给出的就是身高的统计描述表格,因本例中的结果输出较长,为了便于解释,这里仅给出 2007 年 4 月数据的分析结果。因内容较多,依次解释如下。

(1) 集中趋势指标:可见 2007 年 4 月的总指数均值为 98.3,而 5% 截尾均数为 99.0,中位数为 101.5,三者相差不明显,说明数据基本上对称分布。

(2) 离散趋势指标:总指数方差为 358.0,其平方根即标准差为 18.9,样本中总指数极小值为 31.2,极大值为 140.6,两者之差为全距(范围)109.35,中间一半样本的全距为四分位间距 23.43。

(3) 参数估计:可见总指数均数的标准误为 1.09,相应的总体均数 95% 可信区间为 96.2 ~ 100.5。

(4) 分布特征指标:图 7.8 最下方还给出了表示数据偏离正态分布程度的偏度系数和峰度系数,及其各自的标准误,这里不再详述。

在统计描述表格之后,探索过程还会给出身高分性别的茎叶图和箱图,从图形分布上可以看出,分月份的总指数的确基本呈对称分布。对这两种图形的介绍参见第 10 章,这里不再详述。

3. M - 统计量

如果选择了“统计量”子对话框中的 M - 统计量,则会给出如图 7.9 所示的结果。

	月份	Huber 的 M - 估计器 ^a	Tukey 的双权重 ^b	Hampel 的 M - 估计器 ^c	Andrews 波 ^d
总指数	200704	99.6194	100.3020	99.5448	100.3332
	200712	95.7921	96.5184	95.7521	96.5143
	200812	91.0241	91.2941	91.0482	91.2996
	200912	100.3076	100.0637	100.6882	100.0618

- a. 加权常量为 1.339。
- b. 加权常量为 4.685。
- c. 加权常量为 1.700、3.400 和 8.500。
- d. 加权常量为 1.340 * pi。

图 7.9 M - 估计器

图 7.9 中一共会输出 Huber、Andrews、Hampel 和 Tukey 共 4 种 M 统计量,其中 Huber 法适用于数据接近正态分布的情况,另 3 种则适用于数据中有过多异常值时。同样以 2007 年 4 月为例,可以发现上述 4 种统计量的估计值和原始均数相类似,同样说明数据分布应当是接近对称的。

4. 极端值列表

当选中“统计量”子对话框中的“界外值”复选框后,即可输出极端值列表,如图 7.10 所示。

这里同样只给出了 2007 年 4 月的情况,图 7.10 中输出了 5 个最高值与 5 个最低值,以及这些数值所对应的记录号,从两侧极值的大小可见,在最高、最低两个方向上并没有特别明显的异常值,该结果同样支持前面得出的数据分布基本对称的结论。

月份			案例号	ID	值
总指数	200704	最高	1	105	105
			2	158	158
			3	184	184
			4	194	194
			5	288	288
		最低	1	258	258
			2	230	230
			3	248	248
			4	140	140
			5	72	72

图 7.10 极值

5. 百分位数

如果选中“百分位数”复选框,则会输出百分位数表,如图 7.11 所示。

			百 分 位 数						
月份			5	10	25	50	75	90	95
加权平均 (定义 1)	总指数	200704	62.4854	78.1067	85.9174	101.5387	109.3494	117.1600	124.9707
		200712	54.6747	62.4854	85.9174	93.7280	109.3494	117.1600	124.9707
		200812	54.6747	62.4854	78.1067	93.7280	101.5387	117.1600	117.1600
		200912	78.1067	78.1067	85.9174	101.5387	109.3494	132.7814	140.5920
Tukey 的枢纽	总指数	200704			85.9174	101.5387	109.3494		
		200712			85.9174	93.7280	109.3494		
		200812			78.1067	93.7280	101.5387		
		200912			85.9174	101.5387	109.3494		

图 7.11 百分位数

图 7.11 输出了第 5%、10%、25%、50%、75%、90%、95% 分位数,并分别采用了两种算法,当数据量较大,且基本无重复值时,两法的结果相同;反之,则加权平均法会对数据进行内插,此时其结果应当比 Tukey 法更加准确。

7.4 Bootstrap 方 法

7.4.1 模型

前面已经对经典统计学的参数估计方法进行了介绍,但从中也可以看出,这些方法无一例外的需要先对变量的分布进行假定,然后才能够进行相应的计算;另一方面,经典统计学对均数的参数估计,特别是区间估计的研究比较完善,但对于其他一些分布参数,例如中位数、四分位数、标准差、变异系数等的区间估计的研究则较少,这无疑是在方法体系上的一大缺憾。

20 世纪 80 年代以来,随着计算机技术的飞速发展,借助于日益强大的机器计算能力,计算

统计学这一新的统计学分支得到了飞速发展,而 Bootstrap 方法就是发展较早且较为实用的一种计算统计学方法,可以很好地解决经典统计学所无法解决的难题。

1. 基本原理

Bootstrap 方法由 Efron 于 1979 年提出,是基于大量计算的一种模拟抽样统计推断方法,它的使用主要出于两种目的:① 判断原参数估计值是否准确;② 计算出更准确的可信区间,判断得出的统计学结论是否正确。

Bootstrap 方法的基本思想为:在原始数据的范围内做有放回的抽样,样本含量仍为 n ,原始数据中每个观察单位每次被抽到的概率相等,为 $1/n$,所得样本称为 Bootstrap 样本。于是可得到任何一个参数 θ 的一个估计值 $\theta^{(b)}$,重复抽取这样的样本若干次,记为 B 。例如 $B=1\,000$,就得到该参数的 1 000 个估计值,则参数 θ 的标准误的 Bootstrap 估计为

$$se_B = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2 / (B-1) \right\}^{1/2}$$

其中, $\hat{\theta}^*(.) = \sum_{b=1}^B \hat{\theta}^*(b) / B$,根据其性质可以估计得到 θ 的一些性质,如 $\hat{\theta}^{(b)}$ 的分布是否为正态, $\theta^{(b)}$ 的均数及标准差(误), θ 的可信区间等。

2. 参数法和非参数法

Bootstrap 方法有参数法和非参数法两种,前者需要假定 $\hat{\theta}^{(b)}$ 的分布状况,而后者则无任何限制。以可信区间的估计方法为例,其基本原理为当 $\hat{\theta}^{(b)}$ 的分布近似正态时,可以其均数 $\hat{\theta}^{(-)}$ 做点估计,用正态原理估计 Bootstrap 可信区间;而当 $\hat{\theta}^{(b)}$ 的频数分布为偏态时,以其中位数做点估计,用上、下 2.5% 分位数估计 95% 可信区间。

和经典统计学中的情况类似,在一般情况下参数法的效率高于非参数法。但是,正是因为参数法需要实现假定分布类型,导致当数据违反假定时分析结果可能不准确。另外,如果数据存在明确的层次结构,则进行分层抽样而不是完全随机抽样也可以有效地提高分析效率。SPSS 默认为非参数 Bootstrap 方法,并采用完全随机抽样,但也可以根据需求改为分层抽样方法。

3. 抽样次数的确定

在使用 Bootstrap 方法时需要确定的一个基本参数是计算中的抽样次数 B 。显然, B 取值越大,则计算结果越准确,但需要花费的计算时间也越长。从经验值上讲,一般取 50 ~ 200 即可保证参数估计值的相对误差不大于 5%,但如果采用百分位数法来计算可信区间,则显然此时可用于计算区间的数据量太少,最好能增加到 1 000 次上下。高于 1 000 次在多数情况下带来的精度改善非常有限,且过于耗时。因此在多数情况下抽样次数定为 1 000 次最为常见。

7.4.2 案例:对总指数进行 Bootstrap 估计

例 7.3 对总指数的均数、标准差进行 Bootstrap 方法的参数点估计和区间估计。

按照经典统计学的思路,对任何参数进行点估计都是比较容易的,但是如果希望求得标准差的可信区间就非常困难了。而利用 Bootstrap 方法就可以轻松地解决这一问题。

1. 界面说明

SPSS 目前在许多过程的对话框中均纳入了 Bootstrap 模块,在其中以一个子对话框的方式出

现,如图 7.12 所示。

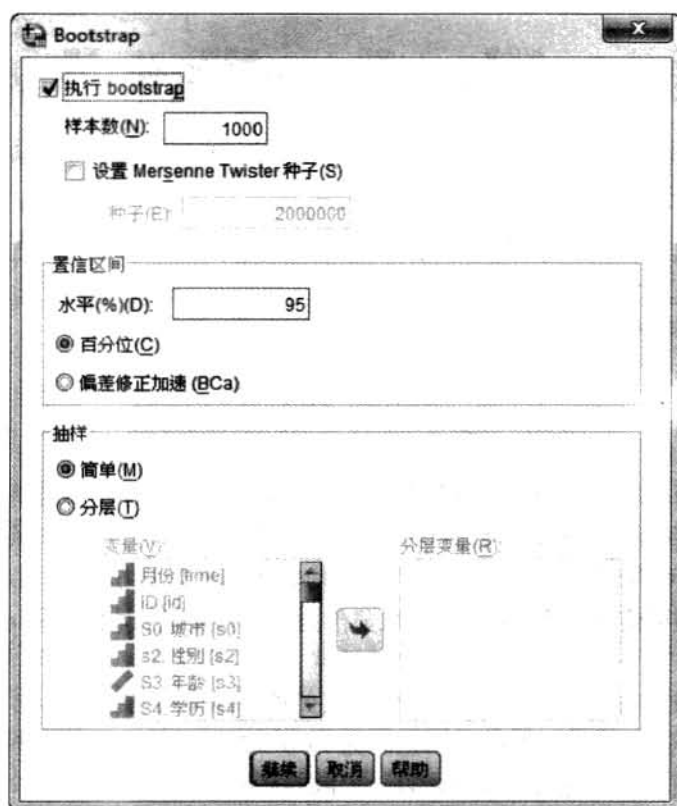


图 7.12 Bootstrap 子对话框

(1) 执行 bootstrap:要求进行 Bootstrap 抽样,下方的“样本数”文本框则用于指定抽样次数,默认为 1 000 次,该设定适用于大多数情形,一般不需要修改。

(2) 设置 Mersenne Twister 种子:作为一种计算统计学方法,在默认情况下 Bootstrap 每次的抽样计算结果都是随机出现的,不能重复。使用该选项就可以在下方的文本框中自行指定随机种子,从而在设定相同随机种子的情况下得到完全相同的分析结果。

(3) “置信区间”框组:默认采用百分位数法计算出 95% 的可信区间,如果希望得到更为精确的结果,则可以使用偏差修正加速 (BCa) 算法来调整区间,它更加准确,但代价是需要更长的计算时间。

(4) “抽样”框组:SPSS 默认为不分层的完全随机抽样,如果确认数据存在层次结构,则可以通过指定分层变量来实现分层抽样,以得到更为准确的分析结果,例如对于 CCSS 数据就可以指定为按照月份、城市来分层抽样以改善分析结果。

2. 结果解释

这里以描述过程为例来解释 Bootstrap 方法的输出,如果是对 index1 进行描述,则在笔记本电脑的 I3 CPU 上,整个计算过程用时小于 10 s,得到的分析结果如图 7.13 所示。

		Bootstrap ^a				
		统计量	偏差	标准误	95% 置信区间	
					下限	上限
总指数	N	1147	0	0	1147	1147
	极小值	.00				
	极大值	156.21				
	均值	95.8935	-.0077	.6158	94.7364	97.1594
	标准差	20.99710	-.00153	.55104	19.91674	22.10868
有效的 N (列表状态)	N	1147	0	0	1147	1147

a. 除非特别说明,否则 Bootstrap 结果都是基于 1 000 个 Bootstrap 样本的。

图 7.13 描述统计量

图 7.13 所示的统计量大家应当非常熟悉,就是普通的描述分析结果,但从其右侧的偏差列起则全部是和 Bootstrap 相关的输出。以“均值”行为例,“偏差”列显示采用 Bootstrap 方法计算出的点估计值要比直接计算出的均数低 0.007,显然该误差几乎可以忽略;采用 Bootstrap 方法计算出的 95% 置信区间为 94.7~97.2,读者可以用均数和标准误差算出(也可以用探索过程直接得到结果)传统方法的可信区间为 94.7~97.1,显然两者非常接近,这说明 index1 整体而言并未明显呈偏态分布。

“均值”行下方是“标准差”的统计结果,显然,此处 Bootstrap 方法显示出了其独特的能力,由结果可知,index1 总体标准差的 95% 置信区间为 19.9~22.1,而经典的标准差点估计值 21.0 也基本上接近 Bootstrap 点估计值,可以用来表示离散趋势的大小。

当采用 Bootstrap 抽样得到的结果与经典统计学结果明显不同时,则说明变量分布很可能违反了经典统计学的前提假设,例如呈偏态分布,或者存在明显的极端值,此时基本上应当以 Bootstrap 方法计算出的点估计和区间估计值为准来加以使用。

思考与练习

- 1. 根据 CCSS_Sample.sav 数据,分析受访者的年龄分布情况,尝试分城市/合并描述。
- 2. 使用描述过程,对 CCSS_Sample.sav 中的总指数、现状指数和预期指数进行标准正态变换,对变换后的变量进行统计描述。

第8章 分类变量的统计描述与参数估计

第7章中主要介绍了如何对CCSS案例中的连续变量进行统计描述和参数估计,本章将继续介绍如何对分类变量完成这些工作,包括其中的无序分类变量、有序分类变量和多选题变量集。

8.1 指标体系概述

8.1.1 单个分类变量的统计描述

相对于连续变量而言,分类变量的统计描述指标体系非常简单,主要是对各个类别取值分别进行频数和比例计算,再进一步计算所需的一些相对数指标。

1. 频数分布

对于分类变量,分析时首先应当了解各类别的样本数有多少,以及各类别占总样本量的百分比各为多少。这些信息往往会被整理在同一张频数表中加以呈现,稍后将在案例中介绍。

对于有序分类变量,除给出各类别的频数和百分比外,研究者往往还对累积频数和累积百分比感兴趣,即低于/高于某类别取值的案例所占的次数和百分比。当然,出于一些特殊的分析目的,累积频数和累积百分比也可能被用于无序分类变量,如希望知道各少数民族占总人数的比例情况等。但需要注意的是,统计软件一般都只按类别编码从小到大进行频数和百分比的累计,如果编码不符合要求,则研究者只能手工加以统计。

2. 集中趋势

除原始频数外,研究者如果希望知道哪一个类别的频数最大,还可以使用众数(Mode)来描述它的集中趋势。显然,众数只反映频数最大的类别的情况,而浪费了所有其他信息,因此只有集中趋势显著时,众数才较有价值。而当变量的类别数不多时,原始频数表的观察并不复杂,此时众数的使用价值并不高。

可能有人会觉得奇怪,为什么在本章中对于分类数据只描述其集中趋势,而忽略掉了离散趋势呢?这是因为对于分类数据而言,其数据的离散程度实际上是和集中趋势有关联的,它们受同一个参数的控制,因此不需要分别描述,后面会详细说明。

3. 相对数指标

除了以上比较简单的频数、比例外,研究者还经常为分类数据计算一些原始频数的相对指标用于统计描述,这些指标称为相对数。下面简单介绍常用的3种相对数。

(1) 比(Ratio):指的是两个有关指标之比 A/B ,用于反映这两个指标在数量/频数上的大小关系。事实上,比也可以被拓展到连续变量的范畴内,如本月销售额/销售人员数。

(2) 构成比(Proportion):用于描述某个事物内部各构成部分所占的比重,其取值在0~100%之间。事实上,前面提到的百分比就是一个标准的构成比,而累积百分比则是构成比概念

的直接延伸。

(3) 率(Rate):率是一个具有时间概念,或者说具有速度、强度含义的指标,用于说明某个时期内某个事件发生的频率或强度,其计算公式为

某事件的发生率 = $\frac{\text{观察期内发生某事件的对象数}}{\text{该时期开始时的观察对象数}}$

准确地讲,率应当是一个时间点上的强度测量值,但这在实际工作中很难做到,因此一般都按一个时段来进行测量,它的分子往往是一个时期的累计数。

以上相对数在使用时应当注意适用条件,如样本量较大时相对数才会比较稳定,基数不同的相对数不能直接相加求和等。

8.1.2 多个分类变量的联合描述

在工作中,往往需要对两个甚至多个分类变量的频数分布进行联合观察,此时就涉及到了多个分类变量的联合描述。以两个变量为例,假设有 n 个个体根据两个属性 A 和 B 进行分类。属性 A 有 r 类: A_1, A_2, \dots, A_r , 属性 B 有 c 类: B_1, B_2, \dots, B_c 。 n 个个体中既属于 A_i 类又属于 B_j 类的有 n_{ij} 个,那么就构成如表 8.1 所示的一个二维的 $r \times c$ 列联表。

表 8.1 二维的 $r \times c$ 列联表

	B_1	B_2	\dots	B_c	合 计
A_1	n_{11}	n_{12}	\dots	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	\dots	n_{2c}	$n_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_r	n_{r1}	n_{r2}	\dots	n_{rc}	$n_{r\cdot}$
合 计	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot c}$	n

在表 8.1 中,除“合计”栏外的每一个单元格反映了 A 、 B 两个变量某种类别交叉时的频数情况,而“合计”栏则分别反映了 A 、 B 两个变量各自的类别频数情况,且表格中的数据有如下的换算关系

$$n_{i\cdot} = \sum_j n_{ij}, n_{\cdot j} = \sum_i n_{ij}, n = \sum_i n_{i\cdot} = \sum_j n_{\cdot j}$$

除给出原始频数外,各单元格内还可能给出行百分比、列百分比和总百分比等,分别用于反映该单元格频数占所在行、列、总样本的比例情况。

8.1.3 多选题的统计描述

多选题是调查问卷中极为常见的调查题目类型,第 2 章中已经对其录入方式进行了讲解,由于它所收集的数据也属于分类数据,因此本章将继续讲解对于这类多选题如何进行描述分析。

以标准的多重二分法为例,多选题会使用多个变量加以记录,当然,可以对每一个单独的题项/变量进行统计描述,但这样做是不全面的,因为这些变量实际上回答的是一个大问题,将问题割裂开来可能会导致不正确的分析结果,而且无法计算一些汇总指标。在多选题分析中比较特别的描述指标有以下 4 个。

(1) 应答人数(Count):是指选择各选项的人数,或者说原始频数。

(2) 应答人数百分比(Percent of Case):选择该项的人占总人数的比例,应答人数百分比可以反映该选项在人群中的受欢迎程度。

(3) 应答人次(Response):是指选择各选项的人次,对于单个选项,应答人次和应答人数是相同的,但是对整个问题而言,应答人次可能远远大于应答人数,因为如果一个受访者选择了两个选项,则将会被计为1个人数,2个人次。

(4) 应答次数百分比(Percent of Response):在做出的所有选择中,选择该项的人次占总人次数的比例。应答次数百分比可以用于比较不同选项的受欢迎程度。

8.1.4 分类变量的参数估计

对于分类变量而言,由于只能取若干个离散的值,因此参数估计关心的就是各类别在总体中的比例是多少,或者当从中进行一次抽样时,抽得相应类别的概率是多少。在各种分类变量的分布中,二项分布最为常见,本书将以其为准加以介绍。

1. 二项分布的定义

假设存在一个随机变量 X , 它的可能取值是 $0, 1, \dots, n$, 且相应的取值概率为

$$P(X=k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

由于 $\binom{n}{k} \pi^k (1-\pi)^{n-k}$ 是二项式 $[\pi + (1-\pi)]^n$ 展开式中的各项, 故称随机变量 X 服从以 n 、 π 为参数的二项分布, 记为 $X \sim B(n, \pi)$ 。对于该变量而言, 有均数 $\mu_X = n\pi$, 方差 $\sigma_X^2 = n\pi(1-\pi)$, 标准差 $\sigma_X = \sqrt{n\pi(1-\pi)}$ 。显然, 对于样本量 n 确定的情形, 均数和标准差间存在着明确的换算关系, 它们都只受 π 的影响, 这也是前面不对离散趋势加以描述的理论依据。

2. 二项分布的参数估计

在实际问题中, 对于一个二项分布的总体而言, 其试验次数 n 是可以人为确定和控制的, 因此只需要对参数 π 加以估计, 就可以明确整个分布的情况。由中心极限定理可知, 当 n 较大、 π 不接近 0 也不接近 1 时 (一般认为这个界限是 $n > 40$, 且 $n\pi$ 和 nq 均大于 5), 二项分布 $B(n, \pi)$ 近似正态分布, 这样就可以利用正态分布中的相应成果来进行参数估计, 相应的 $100(1-\alpha)\%$ 可信区间为 $P \pm 1.96 \sqrt{P(1-P)/n}$ 。

当不满足正态近似的条件时, 则可以直接利用二项分布的概率分布规律计算相应的可信区间, 此处略。

8.1.5 SPSS 中的相应功能

作为比较基本的功能, SPSS 的许多分析过程均可完成分类变量统计描述的任务, 但常用的有位于“描述统计”子菜单中的频率过程和交叉表过程, 以及另外两个用于多选题描述的制表过程/菜单项。

1. 频率过程

第 7 章中已经介绍过频率过程了, 显然, 针对单个分类变量输出频数表是其基本功能, 从中可以得到“频数”、“百分比”和“累积百分比”统计量。除了原始频数表外, 该过程还可给出描述集中趋势的众数, 以及直接绘制用于分类变量的条图和饼图等。

2. 交叉表过程

其强项在于两个/多个分类变量的联合描述,可以产生二维至 n 维列联表,并计算相应的行/列/合计百分比、行/列汇总指标等。

3. 多重响应子菜单

多重响应 (Multiple Response) 子菜单专门用于对多选题变量集进行设定和统计描述,包括多选题的频数表和交叉表均可制作,可以满足基本的多选题分析需求。

4. 表格模块

表格模块提供了非常强大的制表功能,自然也可以使用多选题进行统计描述,详见第 9 章的介绍。

8.2 案例:对学历等背景变量进行描述

这里以 CCSS 案例的背景变量为例,来演示分类变量的统计描述在 SPSS 中的具体实现方法。

8.2.1 使用频率过程进行描述

如果希望了解 CCSS 项目中受访者的学历分布情况,则可以使用频率过程输出相应的频数表,操作非常简单,将变量 S4 学历选入“变量”列表中,单击“确定”按钮后,相应的结果如图 8.1 所示。

		频数	百分比	有效百分比	累积百分比
有效	初中/技校或以下	154	13.4	13.4	13.4
	高中/中专	313	27.3	27.3	40.7
	大专	331	28.9	28.9	69.6
	本科	292	25.5	25.5	95.0
	硕士或以上	57	5.0	5.0	100.0
	合计	1147	100.0	100.0	

图 8.1 S4. 学历

图 8.1 无须进行过多解释,依次为频数、百分比、有效百分比、累积百分比的数值。这里的有效百分比指的是去除缺失样本后,各类别在有效样本中所占的比例,在本例中由于学历没有缺失值,因此数值等同于其左侧的百分比。

读者可自行对性别、职业、婚姻状况等背景变量进行分析,这里不再详述。

8.2.2 使用交叉表过程进行描述

如果研究者希望知道性别和学历的交叉频数分布,以及各种百分比的情况,就需要用到交叉表过程了。

1. 界面说明

选择“分析”→“描述统计”→“交叉表”菜单项,就会打开交叉表过程的对话框,如图 8.2 所

示,下面简要介绍各部分的功能。



图 8.2 交叉表过程的对话框

(1) 主对话框:中部依次排列的“行”列表框、“列”列表框分别用于选择交叉表中的行、列变量。而下方的“层”框组则用于选入更多的分类变量作为层变量(详见第9章中关于表格结构的介绍),注意此处最多可进行多达10层的嵌套,同时行、列、层变量也是可以同时选择多个分类变量的。在左下角可以指定绘制复式条形图来呈现数据,而当交叉表太大的时候,也可以禁止表格输出。

(2) “精确”按钮:单击后打开的子对话框用于设定对行 * 列表是否进行确切概率的计算,以及具体的计算方法。本部分内容的介绍详见12.5节中关于蒙特卡罗方法的介绍。

(3) “统计量”按钮:单击后打开的子对话框中提供了一整套用于计算行/列变量关联性的统计指标和检验方法,详见卡方检验和相关分析两章中的介绍。

(4) “单元格”按钮:单击后打开的子对话框用于定义列联表单元格中需要显示的指标,这些指标被分为计数、百分比和残差三大类,实际上以前两类较常用。此外在现在的新版本中还提供了列于列之间进行对比的Z检验结果输出。

(5) “格式”按钮:单击后打开的对话框用于设定单元格的排序方式,使用价值不大。

(6) “Bootstrap”按钮:要求在相应的参数估计和假设检验中使用Bootstrap方法进行估计,该方法的原理和使用方式在第7章中已经介绍过了。

2. 操作说明与结果解释

根据分析目的,只需要分别将“性别”和“学历”选入“行”、“列”列表框中,然后在“单元显示”子对话框中选择列百分比输出,即可得到所需的结果,如图8.3所示。

图8.3很清楚地给出了性别和学历的交叉分布情况,从中可以看出,随着学历的上升,男性所占的比例从初中/技校或以下的48%,逐渐上升至硕士或以上的63%。当然,由于这只是样本数据的描述情况,这究竟是因为抽样误差所致,还是总体中也的确存在此趋势,还需要通过假设

检验来加以确认。

			S4. 学历					
			初中/技校或以下	高中/中专	大专	本科	硕士或以上	合计
S2. 性别	男	计数	74	167	191	169	36	637
		S4. 学历 中的 %	48.1%	53.4%	57.7%	57.9%	63.2%	55.5%
	女	计数	80	146	140	123	21	510
		S4. 学历 中的 %	51.9%	46.6%	42.3%	42.1%	36.8%	44.5%
合计	计数		154	313	331	292	57	1147
	S4. 学历 中的 %		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

图 8.3 S2. 性别 * S4. 学历交叉制表

8.3 案例:对多选题 C0 还贷状况进行描述

这里以 CCSS 案例中的 C0 还贷状况这一多选题为例来说明如何使用 SPSS 的多重响应 (Multiple Response) 子菜单对其进行描述。首先需要明确,该多选题在数据集中按照多重二分法的记录格式存储为 C0_1、C0_2、C0_3 三个变量,并且在分析前需要将其成功设定为多选题变量集 C0,对上述这些定义及操作有疑问的读者可参见第 2 章的相关内容。

8.3.1 多选题的频数列表

如果希望给出各选项的频数分布情况,则需要通过多选题的频数分析过程来完成。

1. 界面说明

选择“分析”→“多重响应”→“频率”菜单项,打开的对话框如图 8.4 所示。

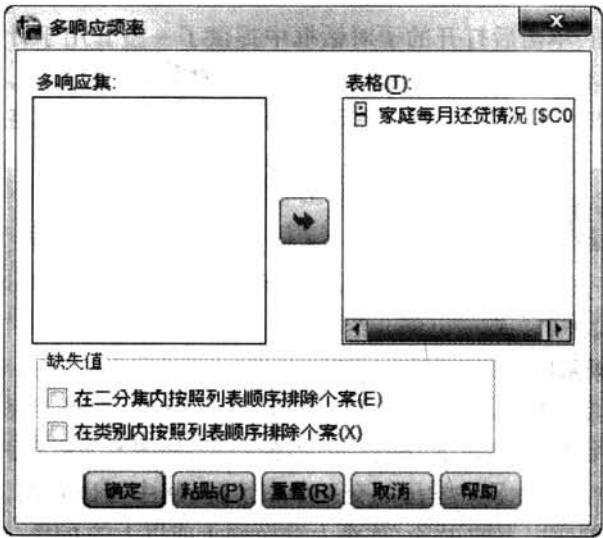


图 8.4 “多响应频率”对话框

该对话框内容非常简单,没有多余的选项,只有下方的“缺失值”框组用于选择对缺失值的处理方式,两个复选框实际上分别对应了多重二分法和多重分类法两种多选题编码方式,读者应注意正确选择,不能交错使用。

2. 操作说明与结果解释

本例的操作非常简单,将变量集 C0 选入即可,相应的结果输出如图 8.5 所示。

	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
\$ C0 ^a	163	14.2%	984	85.8%	1147	100.0%

a. 值为 1 时制表的二分组。

图 8.5 个案摘要

图 8.5 提供了数据的基本信息,在所有的这 1 147 人次中,有 163 人选择了至少一个贷款种类。随后的分析将基于这 163 人的情况进行。

图 8.6 提供的信息解释如下。

(1) 在 199 个有效的回答中,各种贷款种类一共被选择了 199 次,其中“房贷”118 次,“车贷”33 次,“其他消费还贷”48 次。

(2) 右边的响应百分比指的是每个选项被选中的次数占总选择次数的比例,即应答人次百分比。比如这 118 人次选择了房贷,占总选择次数的比例为 $118/199 = 59.3\%$ 。

(3) 最右侧的个案百分比指的是选择某选项的人数占总人数的比例,即应答人数百分比。仍然以房贷为例,这 118 个人占总应答人数的比例为 $118/163 = 72.4\%$,而最下方的比例 122.1% 则说明这 163 人平均而言每人选择了 1.22 个贷款种类。

		响应		个案百分比
		N	百分比	
家庭每月还贷情况 ^a	C0. 请问您的家庭目前有下列还贷支出吗:房贷	118	59.3%	72.4%
	C0. 请问您的家庭目前有下列还贷支出吗:车贷	33	16.6%	20.2%
	C0. 请问您的家庭目前有下列还贷支出吗:其他一般消费还贷	48	24.1%	29.4%
总计		199	100.0%	122.1%

a. 值为 1 时制表的二分组。

图 8.6 \$C0 频率

8.3.2 多选题的列联表分析

前面直接给出了多选题的频数表,但有的时候还希望能够对不同的人群分别进行描述,即对多选题变量集和其他分类变量进行交叉描述。例如在本例中希望分婚姻状况考察贷款状况,则需要用到多响应交叉表过程来完成相应的分析。

1. 界面说明

选择“分析”→“多重响应”→“交叉表”菜单项,打开的对话框如图 8.7 所示。

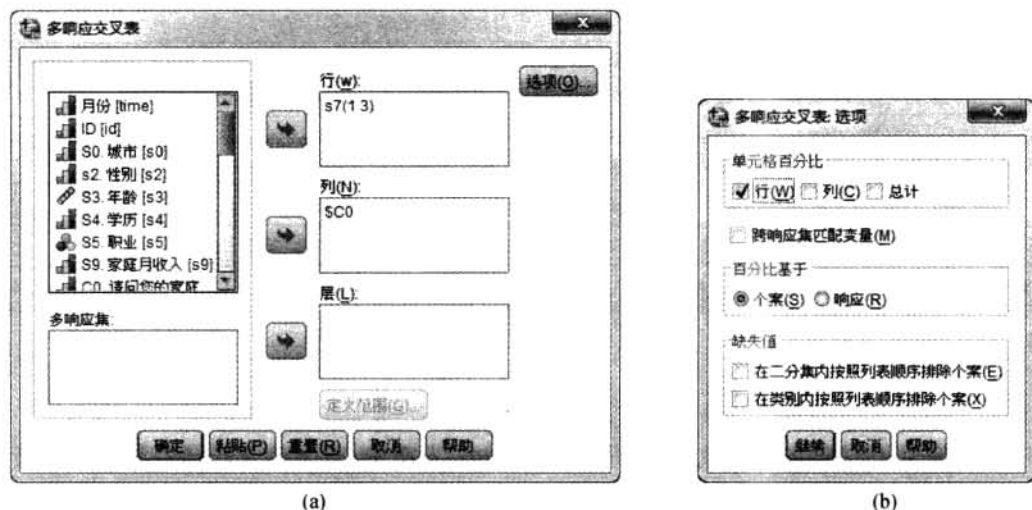


图 8.7 多响应交叉表过程的对话框

(1) 主对话框:由图 8.7 可见该主对话框和普通交叉表过程的主对话框非常相似,只是左下方会单独列出多响应集供选择。需要指出的是,多响应变量集在使用上没有任何限制,可以被任意选入“行”/“列”/“层”列表框中,只是不合适的选择可能会使得表格过于复杂。此外,对于选入“行”/“列”/“层”列表框中的分类变量,还需要使用最下方的“定义范围”按钮,为该变量设置取值范围。

(2) “选项”按钮:单击后打开如图 8.7(b)所示的子对话框,最上方的“单元格百分比”框组用于定义输出行百分比、列百分比和总百分比指标;当选中下方的“跨响应集匹配变量”复选框后当行/列变量均为多重分类法记录的多选题变量集时,可以要求结果表格按两个变量集取值一一对应的方式来生成,但实际应用价值不大;下方的“百分比基于”框组则可以定义交叉表中的比例计算是基于应答人数,还是应答人次;最下方的“缺失值”框组则用于控制缺失值的处理方式,前面已经介绍过了。

2. 操作说明与结果分析

根据分析要求,这里只需要分别将 S7 和 C0 选入“行”/“列”列表框中,并在“选项”子对话框中设置输出行百分比即可,分析结果如图 8.8 所示。

图 8.8 中婚姻状况给出了家庭的还贷情况,为了便于输出,图 8.8 中的列标签做了一定的删减,可以发现已婚人群的房贷比例高于未婚受访者,而未婚人群的车贷和其他消费还贷比例则均高于已婚人群,贷款的范围的确要更广一些。但对于这一结论有两点需要指出:首先,上述比例是基于 163 位有贷款的受访者计算的,而不是基于全部的 1 147 人计算的,因此结论可能有一定的偏差;其次,上述趋势仍然只是样本情况,未经过假设检验的验证,因此仅仅是一种可能存在的趋势,还不能下最终结论。

			家庭每月还贷情况 ^a			
			房贷	车贷	其他一般消费还贷	总计
S7. 婚姻状况	已婚	计数	91	23	30	120
		S7 内的 %	75.8%	19.2%	25.0%	
	未婚	计数	27	10	17	42
		S7 内的 %	64.3%	23.8%	40.5%	
	离异/分居/丧偶	计数	0	0	1	1
		S7 内的 %	.0%	.0%	100.0%	
总计		计数	118	33	48	163

百分比和总计以响应者为基础。

a. 值为 1 时制表的二分组。

图 8.8 S7 * \$ C0 交叉制表

思考与练习

1. 根据 SPSS 自带数据 Employee data. sav,分析员工的性别、受教育程度、少数民族、职位类别的分布情况,并尝试分析这些属性之间的关系以及这些属性和工资之间的关系。
2. 根据 SPSS 自带数据 1991 U. S. General Social Survey. sav,分析健康问题(对应的变量为 hlth1 ~ hlth9,为多选题)的分布情况。

第9章 数据的报表呈现

通过前面几章的学习,大家已经能够对任意类型的资料自由地进行汇总描述了,并且此时 SPSS 会自动地将相应的指标用表格呈现出来。但是,这些标准的表格格式可能无法满足千变万化的实际工作需求。现代社会瞬息万变,如何能高效、快捷地将数据内涵呈现出来已经变得非常重要。本章将介绍如何用 SPSS 中的自定义表格(Custom Tables)模块生成更为专业和复杂的统计报表。

9.1 统计表入门

SPSS 的分析结果现在已经主要以表格形式出现。但是,SPSS 的结果表格并非如表面上看到的那样是一个简单的二维表格,而是一种拥有数据透视、数据旋转、格式变换等多种强大功能的交互式表格,也正因为如此,将其称为枢轴表(Pivot Table)。




从 SPSS 19 版起,为了加快输出速度,表格输出开始提供轻量表和枢轴表两个格式选择,此时的轻量表还只能浏览不能编辑。但从 20 版起,全部的表格都已改用轻量表兼容格式输出,使得结果输出速度大大加快,但该格式不能在 18 及以前版本中进行编辑。

9.1.1 统计表的基本框架

在 SPSS 的表格操作中,行、列、层是非常重要又经常用到的 3 个概念。它们实际上都是表格的一个维度,所谓行(Row)指的是形成表格横行的元素,而列(Column)指的是形成表格纵列的元素。行、列元素相交就会形成一个最简单的二维表,由行、列元素不同取值的组合就确定了一个单元格(Cell)。

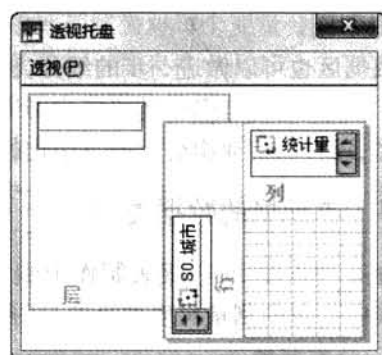
与行、列的概念相比,层(Layer)的概念稍微复杂一些,它指的是表格中的第三个维度,不妨把此时的表格想象成一个立方体,行、列、层就对应了该立方体的长、宽和高。由于在屏幕上能够直接展示的只能是二维表,因此在三维表中,使用者能够直接观察到的只能是三维表中的一层,而其余各层被隐藏在所观察到的层背后,无法同时看到。

需要注意的是,表格中的元素和人们所说的变量并不相同,它既可能是一个分类变量的不同取值,也有可能是一个变量组,还可能是一个统计量组。也就是说,表格中的一个维度可以由多个变量联合构成。以 CCSS 数据中的 S0 城市为例,其频数表输入如图 9.1 所示。

在结果窗口中双击激活该表格,则会进入表格编辑模式,此时会打开表格托盘,如图 9.1(b)所示,该托盘显示的就是当前表格的框架结构,每一个表格都有行、列、层三维,注意形如的图标,它代表的就是一个表格元素。这时在行、列上都有元素出现,分别对应着变量 S0 和统计量(具体内容为表格中见到的频数、百分比等)。而在层上无表格元素出现,说明该表格为一个简单的二维表。将表格结构和前面的表格输出进行对应,应当更容易理解相应的这些概念。


S0. 城市					
		频率	百分比	有效百分比	累积百分比
有效	100北京	378	33.0	33.0	33.0
	200上海	387	33.7	33.7	66.7
	300广州	382	33.3	33.3	100.0
	合计	1147	100.0	100.0	

(a)



(b)

图 9.1 频数表与其相应的表格托盘

 如果进入编辑模式后没有看到托盘出现,则可以选择“透视”→“透视托盘”菜单项进行显示。

表格托盘除了显示表格结构外,还可以直接进行表格透视方向的旋转,例如选中列元素“统计量”的图标,再将其拖动到层元素位置上,则表格会立刻发生相应的变化,原先的分列取消,在表格最上方出现“统计量”下拉列表,这实际上就对应了层元素的设置。默认显示的是统计量组的第一项“频率”所对应的结果。使用者也可以在“统计量”下拉列表框中选择所需要的统计量层,如图 9.2 所示。



(a)

S0. 城市		
统计量	频率	
	频率	京 378
	百分比	海 387
	有效百分比	州 382
	累积百分比	1147

(b)

图 9.2 使用托盘将列元素转换为层元素

9.1.2 表头、数据区与汇总项

在了解了表格基本框架后,现在将基本框架和具体的表格内容对应起来。任何一个二维表格的第 1 行、列就对应了托盘中行、列元素的具体取值,就是前述的表格框架,因此第 1 行、列也被称为表头。由于在 SPSS 的表格中行、列实际上是没有本质区别的,因此这里的表头和包括第 1 行的表头概念不同,需要注意区别。

除了表头之外,剩余表格部分均是由行、列元素相交而成的,用于给出相应的数值,这些部分被统称为数据区。区分表头和数据区非常重要,因为它们的格式设置、操控方式等

均完全不同。

数据区也可以做进一步的细分,例如在上述产地的频数表中,除了各类别以外,行元素中还出现了汇总项。在 SPSS 的表格中可以出现行合计、列合计、层合计项,对于叠加表、嵌套表等表格类型,还可以有亚组合计等更细化的合计方式出现。

9.1.3 单元格的数据类型

在某种程度上,在报表制作中能对变量所进行的呈现方式完全取决于该变量的测量尺度。在报表中变量的测量尺度被简单而明确地分为两大类:分类变量和连续变量。

1. 分类变量

分类变量包括名义和有序尺度两大类,虽然在制表的对话框中会将这两类变量用不同图标标识出来,但实际上在报表制作中几乎并未对它们加以区分。对于分类变量,原始类别频数和构成百分比是最常用的描述指标。将其中的百分比和具体的计算方向相结合,又形成了许多更细化的指标,如行百分比、列百分比、层百分比、总表格百分比。在存在缺失值的情况下,又可按照合计数中是否包括缺失值而出现了有效例数、行有效例数百分比、列有效例数百分比、层有效例数百分比、表格有效例数百分比等新的组合。



在制表时,多选题变量集是作为一类特殊的“分类变量”来处理的,还使用一组较为特殊的百分比、频数指标等来对其进行描述。

2. 连续变量

连续变量包括间距尺度和比率尺度两大类,同样在报表制作中不进行区分。相对而言,连续变量在报表中可供使用的统计指标要比分类变量丰富得多,包括了大家在前面学习过的各种集中趋势、离散趋势指标,这里分述如下。

- (1) 集中趋势指标:均数、中位数、众数、最大值、最小值。
- (2) 离散趋势指标:全距、标准误、标准差、方差。
- (3) 百分位数:第 5、25、75、95、99 百分位数及任意指定的百分位数。
- (4) 百分比:按相应合计方向当前变量的行、列、层、表格合计百分比。
- (5) 其他:例数、有效例数、总和等。

3. 汇总项

汇总项的情况类似于普通单元格,其数据类型仍然只有以上两种。但是除默认使用被汇总单元格的统计指标外,还可以自定义不同的汇总项统计指标,例如各分项列出频数,而汇总项则使用某一个指标的均数,在后面的分析实例中会见到这种输出。

9.1.4 几种基本表格类型

在熟悉了表格的基本结构和常用术语后,下面来了解一下几种常见的表格类型。需要指出的是,虽然下面的例子中几乎都是类别频数的描述,但在这些表格中也完全可以给出其他连续变量的描述指标。

1. 叠加表

叠加表(Stacking)指的是在同一张表格中对两个变量进行描述,或者说表格中有一个维度

的元素是由两个以上的变量构成的。叠加表其实可以被简单地理解为对每个变量分别绘制两个简单报表,然后将它们拼接到一起,如图 9.3 所示的叠加表就是在一张表格中同时给出了城市和性别的频数。连续变量也可被放在叠加表中,例如前面介绍过的 Descriptive 过程,如果同时计算多个变量,则实际上其结果就是一个叠加表。

S0. 城市			S2. 性别	
100 北京	200 上海	300 广州	男	女
计数	计数	计数	计数	计数
378	387	382	637	510

图 9.3 横向叠加表示意

虽然“叠加”在字面上是纵向拼接的意思,但也存在横向拼接的叠加表,在学习了表格基本框架后,这并不难理解。

2. 交叉表

交叉表(Crosstabulation)十分常见,是观察两个分类变量间联系时最常用的表格,它的两个维度都是由分类变量的各类别(及汇总)构成的,图 9.4 显示了性别和城市的关联,显然广州的男性比例要高一些。

		S0. 城市			
		100 北京	200 上海	300 广州	合计
S2. 性别	男	188	221	228	637
	女	190	166	154	510
合计		378	387	382	1147

图 9.4 交叉表示意

3. 嵌套表

类似于交叉表,嵌套表(Nesting)也可以用于显示两个分类变量间的联系,但是在嵌套表中,这两个变量被放置在同一个表格维度中,即该维度是由两个变量的各种类别组合构成的,如图 9.5 所示。例如仍然是显示城市和性别不同组合下的频数,但此时这两个变量都被放置在行上。显然,一般而言,嵌套表并不如交叉表直观。但是当在每个单元格内需要呈现的统计指标非常多时,嵌套表则更为美观和紧凑。

4. 多层表

如果指定了层元素,则表格就由二维扩展到了三维,即多层表(Layers)。事实上,多层表和嵌套表也非常相似,只是现在每次只能观察到其中一层的数据而已。在数据仓库技术中,多层表也被称为数据立方体(Club),在前面介绍表格的基本框架时已经给出了几张多层表,因此这里不再给出具体的实例。

5. 复合表格

以上给出的只是最简单的几种表格类型,在实际的工作中,这些表格类型还有可能互相组合,以更好地达到相应的分析目的。比如叠加-交叉表(一个维度是分类变量,另一个维度则是

两个分类变量的叠加)、嵌套 - 交叉表(一个维度是分类变量,另一个维度则是两个分类变量的嵌套)等。

				计数
S0. 城市	100 北京	S2. 性别	男	188
			女	190
	200 上海	S2. 性别	男	221
			女	166
	300 广州	S2. 性别	男	228
			女	154

图 9.5 嵌套表示意

9.1.5 SPSS 中的报表功能

作为功能非常完善的统计软件,SPSS 提供了非常强大的统计报表功能,除 Base 已具有非常完善的统计报表功能外,SPSS 还提供了专门的 Custom Tables 模块用于生成更为专业的统计报表。在较早版本的 SPSS 中还有 Original Tables 模块用于制表,但目前该模块已经取消。

1. Base 模块

SPSS 的 Base 模块已经为用户提供了非常完善的统计报表功能,除涉及统计描述的多个过程可以生成各种描述统计量的基本报表外,还在“分析”主菜单的“报告”和“多重响应”子菜单中提供了专用的统计报表功能。

(1) “报告”子菜单:提供了从最基本的变量值标签代码本、对原始数据进行列表显示,到将原始数据汇总为数据立方体进行数据透视、对数据计算一些常用的描述统计量并制作精细定义的输出表格等多种统计报表功能,可以满足用户的各种苛刻要求。相比之下,该子菜单中各过程的操作都较为简单,用户可自行学习掌握,因此本书不再详述。如果读者希望进一步了解这些过程的功能细节,可以参见《SPSS 11 统计分析教程》(基础篇)中的相关章节,或参考 SPSS 的用户手册。

(2) “多重响应”子菜单:专门为多选题数据的描述而设计,提供了设置多选题变量集、生成多选题频数表和交叉表的全部功能。其相应的功能和操作已在前面的相应章节中介绍过了。

2. Custom Tables 模块

Custom Tables 模块是从 SPSS 11.5 版起新增加的一个功能非常强大的专业制表模块,和 Base 模块中的相应功能相比,它不仅功能更为强大和灵活,而且还提供了完全交互式的操作界面,使用上更为方便、快捷,该模块也是本章随后介绍的重点。

3. Original Tables 模块

Original Tables 模块在 SPSS 11 版中被称为 Tables 模块,后来为了和上面提到的 Custom Tables 模块进行区分而改为现在的名称。它是 SPSS 专门为生成出版级报表而设计的模块,可以针

对各种要求产生复杂的多层/嵌套表格。由于后来新增的 Tables 模块指标功能更强,且全交互的操作界面也更为灵活,因此现在的 SPSS 版本中已不再提供原有的 Original Tables 模块。如果读者对该模块感兴趣,希望进一步了解,可以参见《SPSS 统计分析基础教程》的相关章节,也可以参考 SPSS 的用户手册。

9.1.6 SPSS 中统计表的基本绘制步骤

如果只是绘制一两个比较简单的报表,则在操作上并无太多要点需要注意,只需要找到能够满足相应需求的过程,然后将表格设置正确即可完成。但是,大多数实际任务要比这复杂得多,有可能有数十张甚至上百张特定格式的表格需要绘制,而表格的复杂程度又超出常见的范围。此时使用 SPSS 制表时一般不会一次到位,而是一个由简入繁、循序渐进的过程。初学者往往希望通过对话框的设置一次将全部选项所需的设定完毕,但这恰恰会导致事倍功半。为此,有必要给出常用的制表步骤,如下。

(1) 确定所需绘制表格的基本结构,如行、列元素都由什么构成,是否会在表格中出现多个元素的嵌套,有多少种汇总,是否出现了嵌套汇总等。

(2) 使用对话框绘制表格的基本结构。这里不要拘泥于单元格的格式设置或者统计量是否选择完全这些细节,也不要考虑标题、脚注等次要问题,而是要将注意力集中在是否已经得到了所需的表格结构上。如果结构还不相同,则继续修改直至完成。

(3) 对细节进行完善,包括每个具体统计量的输出格式、汇总项的输出位置等,使得其中至少有一部分单元格的输出格式已符合要求。

(4) 添加其余变量、统计量到表格中来,使表格中的内容满足相应问题的需求。

(5) 对表格中的文本进行修饰,包括标题、统计量标签、变量名和变量值标签等。

(6) 最后一次审核所绘制的表格,考虑有无需要改进之处。

(7) 生成相应的表格,并将其格式保存为模板,供后续任务使用。

本章随后的分析实例就会按照上述结构安排,以利于读者养成良好的制表习惯。

9.2 简单案例:题目 A3 的标准统计报表制作

下面结合 CCSS 项目中的实际制表案例来说明自定义报表模块制表的操作步骤。

9.2.1 案例简介

例 9.1 CCSS 项目每月都会生成固定格式的统计表格,图 9.6 所示为对题目 A3 的固定表格格式,行标题首先为 A3 选项的占比,随后为题目感受值的均数,列标题则为受访月份。要求用 SPSS 的制表模块实现该表格。

图 9.6 所示的表格结构并不复杂,首先它是一个二维表,其列元素就是访问时间变量 time,而行元素则由两部分构成:首先是 A3 选项的构成比,其次是 A3 的题目得分均数,后者可以用数据集中已经生成的中间变量 QA3 来计算,下面几小节将介绍如何在 SPSS 中进行具体的操作。

	2009.9	2009.10	2009.11	2009.12
明显改善	12.3	10.3	11.7	12.2
略有改善	20.1	22.7	34.2	31.1
基本不变	46.3	53.1	41.3	50.6
略有恶化	8.8	7.6	6.9	3.7
明显恶化	11.9	4.4	3.4	0.2
不知道/无回答	0.5	1.8	2.5	2.2
感受值	106.0	113.5	121.9	125.7

图 9.6 CCSS 报告中的 A3 结果表格

9.2.2 绘制表格基本框架

1. 界面说明

选择“分析”→“表”→“设定表格”菜单项,就会打开报表生成器的操作界面,如图 9.7 所示。和 SPSS 中的其他过程不同,自定义表格过程是多层选项卡界面。其中最常用的就是图 9.7 所示的“表格”选项卡,用于对表格框架进行定义。



图 9.7 报表生成器的主对话框

(1) “变量”列表框:位于左上角,会列出所有可用的变量,如果有多选题设定,则会显示在列表最后。用户可采用拖放操作将相应变量/多选题变量集拖入右侧的画布区域。

(2) 制表画布(Canvas):在界面中部占据绝大部分空间,类似于画家绘画时的空白画布,用

户在制表时就是在这张空白画布上进行拖放操作,最终得到合适的表格的。该画布有两种显示界面:正常视图和紧凑视图,分别在画布上方用“普通”和“压缩”按钮加以控制。对于多层表,右上方还提供了“层”按钮,单击后会出现“层”列表框,用于选入层变量。

(3) “类别”列表框:会在“变量”列表框中选中分类变量时自动列出所有的类别取值/标签。例如如图 9.7 中为选中变量 A3 时,下方自动列出了该变量的各类取值标签。

(4) “定义”按钮组:用于对制表变量的统计指标、汇总方式等进行设定。

(5) “摘要统计量”框组:用于控制不同类统计量的排列方向和变量标签显示方向。

(6) 类别位置:用于设定类别标签的显示和排列方向。

上述各框组的详细功能,以及其余几个选项卡的功能将在后面介绍,下面首先介绍制表案例的具体操作。

2. 具体拖放操作

首先以放置在行上的变量 A3 为例来说明基本的拖放操作要点:选中变量列表中 A3 的图标,将其拖动入画布区内。当鼠标接近画布的行区域时,相应的行区域边框变红,同时鼠标图标还原为手形,表明该已找到泊留位置。此时松开左键,则变量 A3 会被放置在列框中,而相应的变量名标签、变量值标签会立刻在画布上显示出来,如图 9.8 所示。

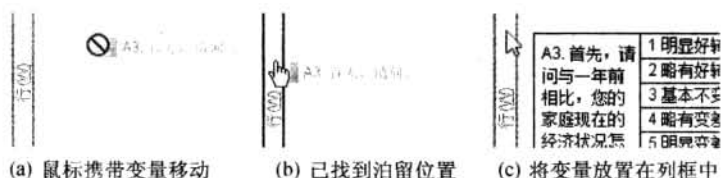


图 9.8 拖放操作示意

对行变量 QA3 以及列变量 time 的操作方式与 A3 基本相同,不再详述。但是在拖放 QA3 的时候,由于此时行上已有变量 A3 存在,放置位置不同时可以得到完全不同的 5 种结果:上叠加、下叠加、左嵌套、右嵌套和替代。读者参照前面介绍的基本表格类型就会理解。显然本例中应当为下叠加,即拖放完毕后,表格框架应当基本上如图 9.9 所示。

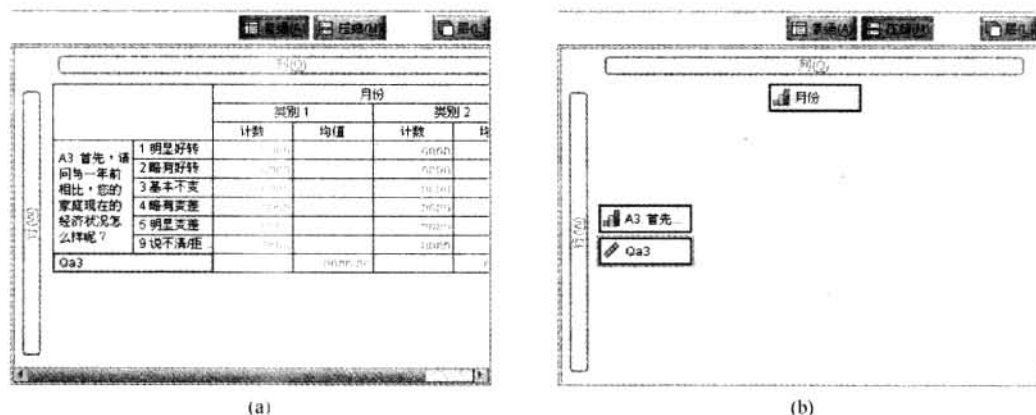


图 9.9 表格框架的普通视图和紧凑视图

最后,如果所绘制的表格太大,则可以切换到紧凑视图,此时画布上将只显示变量的设定位置,而不再给出具体的单元格设定等,这样表格框架会更为清晰,但是对标签等的精细设定在此视图中无法完成。

9.2.3 设置摘要统计量及格式

在图 9.9 所示的表格画布普通视图中可以看到,对 A3 的每个类别默认输出的是频数,而对 QA3 的均值默认输出的则是两位小数,这些都和题目的要求不符,因此需要再对连续变量的统计量加以设定,这些操作都在“摘要统计量”子对话框中完成,具体如下。

1. 对分类变量的摘要统计量进行设定

单击画布上变量 A3 的图标,此时“摘要统计量”按钮变黑可用,单击该按钮后打开如图 9.10 所示的对话框。可见在默认情况下右上侧的显示列表中只有频数(计数),将其移除,然后选入“行 N%”,注意此处需要修改显示格式为不带%的 nnnn.n,操作完毕后单击下方的“应用选择”按钮即可使设定生效。

从图 9.10 所示的对话框中可以看出,在默认情况下汇总项的统计量是和单元格相同的,但如果希望采用不同的设定,则选中左中侧的“设定关于总计和小计的摘要统计量”复选框,即可激活下部的列表框,用于对汇总项的统计量进行单独设定。



图 9.10 分类变量摘要统计量设定对话框

2. 对连续变量的摘要统计量进行设定

单击画布上变量 QA3 的图标,再单击“摘要统计量”按钮,此时打开的就是针对连续变量的摘要统计量设定子对话框,如图 9.11 所示,此处默认显示的均值就是所需要的指标,但是相应的标签、格式和小数位数均需要修改:双击标签,就会进入编辑状态;格式和小数位数使用下拉列表和计数器修改,操作更为简单。

“定义”按钮组中的“分类和总计”子对话框在本例中没有使用,因此暂不介绍,随后会在 9.3 节的复杂案例中加以介绍。

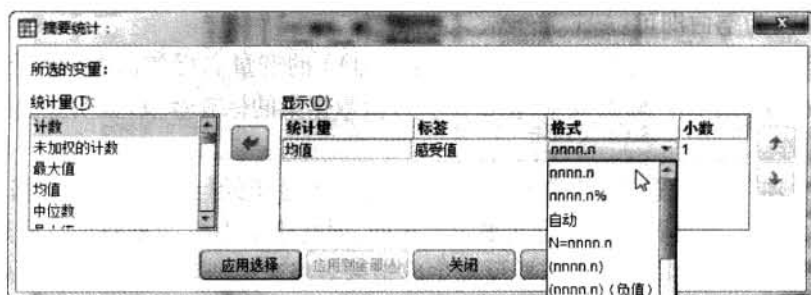


图 9.11 连续变量摘要统计量设定对话框

9.2.4 调整各种显示细节

现在已经基本上完成了所需的表格框架,如图 9.12 所示,但是从所显示的框架示意图可以看出,表格在显示细节上还有很多问题需要修改,依次解决如下。

		月份			
		类别 1		类别 2	
		行 N %	感受值	行 N %	感受值
A3. 首先, 请问与一年前相比, 您的家庭现在的经济状况怎么样呢?	1 明显好转	nnnn.n		nnnn.n	
	2 略有好转	nnnn.n		nnnn.n	
	3 基本不变	nnnn.n		nnnn.n	
	4 略有变差	nnnn.n		nnnn.n	
	5 明显变差	nnnn.n		nnnn.n	
	6 说不清楚/拒	nnnn.n		nnnn.n	
Qa3			nnnn.n		nnnn.n

图 9.12 汇总项设定完毕后的 A3 表格框架示意图

1. 隐藏变量名标签

目前变量名/变量名标签仍然被显示在表格中,只需要分别在画布上 A3、QA3、Time 的变量名处右击,弹出的快捷菜单如图 9.13 所示。取消选中其最下方的“显示变量标签”复选框,即可使变量名标签在表格中被隐藏起来。

2. 使百分比和均数同列显示

在“摘要统计量”框组的“位置”下拉列表框中将默认的“列”选项改为“行”选项即可。

3. 隐藏统计量标签“行 N%”

在“摘要统计量”框组的“位置”下拉列表框右侧有一个“隐藏”复选框,选中它可以将所有统计量标签全部隐藏起来,但在本例中需要将 QA3 的标签“感受值”显示出来。对此有以下两种解决方式。

(1) 将 QA3 的均数统计量标签改为“感受值”,然后再将 A3 的百分比标签改为空白。前一步前面已经完成,现在只需要

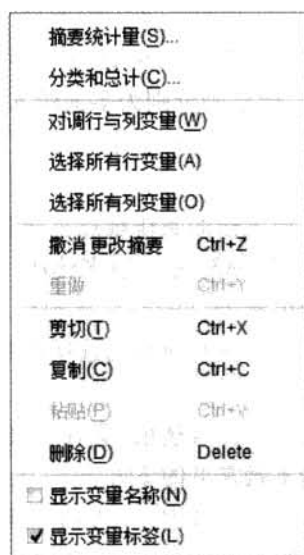


图 9.13 变量名的右键菜单

将 A3 的标签修改为空白即可。

(2) 仍然隐藏所有的统计量标签,然后将变量 QA3 的变量名标签改为“感受值”并加以显示。这样操作也可以达到同样的效果,但由于会涉及数据集的标签修改,且明显会导致变量含义混淆,因此不推荐。

在进行了上述各项设置后,从表格框架就可看出所希望绘制的表格已经基本完成,如图 9.14 所示。因最终的输出和框架示意几乎完全相同,因此这里不再重复列出。在这里也可以看到自定义制表模块的一大特色是不需要反复生成结果表格来检查制表过程,只需要考察画布上的表格框架,就可以很好地控制最终表格的质量。

		月份	
		类别 1	类别 2
A3. 首先, 请问与一年前相比, 您的家庭现在的经济状况怎么样呢?	1 明显好转	00000.0	00000.0
	2 略有好转	00000.0	00000.0
	3 基本不变	00000.0	00000.0
	4 略有变差	00000.0	00000.0
	5 明显变差	00000.0	00000.0
	9 说不清楚/拒...	00000.0	00000.0
Qa3	感受值	00000.0	00000.0

图 9.14 最终设定完成后的 A3 表格框架

和最终的表格相比,这里生成的表格其变量值标签还略有不同,可以在数据集中修改好相应的标签属性,然后重新生成表格,也可以直接在表格中进行编辑,修改相应变量的值标签以达到同样的效果,详见 9.4 节。

9.3 复杂案例:题目 A3a 的标准统计报表制作

9.3.1 案例简介

例 9.2 图 9.15 所示为 CCSS 项目报告中题目 A3a 的固定表格格式,列标题仍然为受访月,行标题则分别为多选题 A3a 的乐观与悲观答案的响应百分比,在其上方则分别对乐观与悲观应答比例进行了小计,注意小计的汇总指标为应答人数百分比。要求用 SPSS 的制表模块实现该表格。

该表格的制作难点主要有以下几个。

- (1) A3a 是一道多选题,因此首先需要将相应的变量 A3a_1、A3a_2 设定为多选题变量集才能进行制表。
 - (2) 并非所有的选项都需要在表格中出现,中性原因、拒答/不知道这两个选项是需要隐藏起来的。
 - (3) 乐观选项、悲观选项需要分别进行小计,而不是对全部选项进行合计,且小计位置在上方而不是常见的下方。
 - (4) 小计项采用的是应答人数百分比,而选项采用的是应答人次百分比,指标不同,需要分别设定。
- 在随后的软件操作中将依次解决上述问题。

	2009.9	2009.10	2009.11	2009.12
导致家庭经济状况改善的原因	25.6	20.4	30.7	25.0
与收入相关的原因	17.3	15.8	24.6	22.8
与就业状况相关的原因	2.2	1.8	1.9	0.6
与投资相关的原因	1.9	0.5	1.2	0.0
与家庭开支相关的原因	4.2	2.3	2.2	1.9
与政策/宏观经济相关的原因	1.6	0.8	1.7	2.3
导致家庭经济状况恶化的原因	38.4	20.9	24.2	12.3
与收入相关的原因	11.7	7.0	5.6	6.2
与就业状况相关的原因	4.8	3.1	6.6	2.2
与投资相关的原因	1.5	0.5	0.2	0.0
与家庭开支相关的原因	24.1	12.0	14.4	4.1
与政策/宏观经济相关的原因	2.2	1.5	0.3	1.4

图 9.15 CCSS 报告中的 A3a 结果表格

9.3.2 多选题、表格基本框架及汇总项的设定

1. 设定多选题变量集

首先检查数据集中变量 A3a_1、A3a_2 的测量尺度是否正确(如果错误地设定为“度量(S)”,后面的操作将会出错),然后按照第 2 章中的讲解将其设定为变量集 TA3a,这里不再赘述。

2. 绘制表格基本框架

有了 9.3.1 小节的基础,这里只需要简单地列出如下所需的操作。

- (1) 将变量 time 拖放至列框内。
- (2) 将变量集 TA3a 拖放至行框内。
- (3) 通过快捷菜单设定 time 和 TA3a 的变量名标签为隐藏。
- (4) 在“摘要统计量”框组中,将“位置”下拉列表框中的选项更改为“行”,选中右侧的“隐藏”复选框以隐藏统计量标签输出。

3. 设定摘要统计量

在画布上选中变量集 TA3a 后,单击右下方的“摘要统计量”按钮,在打开的子对话框中进行如下操作。

- (1) 变量统计量,在右侧的“显示”列表中删除“计数”选项,选入“列响应%”选项,并将其格式改为“nnnn. n”,1 位小数。
- (2) 选中右侧的“设定关于总计和小计的摘要统计量”复选框,清除下方显示列表中已有的选择,选入“列 N%”选项,同样将其格式改为“nnnn. n”,1 位小数。
- (3) 单击“应用选择”按钮退出子对话框。

9.3.3 设定分类变量小结和汇总项

下面重点进行多选题选项小结的设定,这需要在“定义”按钮组的“分类和总计”子对话框中实现,如图 9.16 所示。

1. “分类和总计”子对话框界面说明

- (1) “值”框组:直观地显示该分类变量各类的显示方式、顺序、汇总等。上部显示的是各类



图 9.16 “分类和总计”子对话框

的取值和值标签,其排列顺序与表格输出中的顺序相对应。

(2) “小计和计算的类别”框组:用于在类别中插入子汇总项,并可插入多个。

(3) “对类别排序”框组:用于设定各类别的排序方式,可按照数值、标签、频数进行升、降序的排列。但是,如果有类别被剔除,或者加入了子汇总项,则排序功能不可用。

(4) “排除”列表框:如果使用者不希望在列表中出现某些类,则将相应的取值选入该列表框中即可。

(5) “显示”框组:用于设定某些项目是否显示,包括合计项、空类、未提供值标签的类别。

(6) “显示总计和小计”框组:用于设定汇总和子汇总项的标签是在左/上部显示还是在右/下部显示。在许多项目中,客户习惯于汇总项位于左/上部,显然,这一功能将非常有用。

2. 具体操作

本例中所需的操作如下。

(1) 将“中性原因”、“不知道/拒答”两个选项移入“排除”列表框中。

(2) 在右下角的“显示总计和小计”框组中,选中“上述类别中它们适用的类别”单选按钮。此处翻译有误,实际含义是将小计、汇总在相应类别的左/上侧显示出来。

(3) 在“值”框组中选中 10(改善:收入相关),然后单击下方的“添加小计”按钮,在打开的“定义小计”子对话框中将小计名称改为“导致家庭经济状况改善的原因”。

(4) 按照和上述方式类似的操作,在 110(恶化:收入相关)上方插入名称为“导致家庭经济状况恶化的原因”的小计。



在本例的操作中使用了汇总子对话框中的“添加小计”按钮来实现项目的汇总,实际上也可以使用“添加类别”按钮实现完全相同的结果。采用添加类别方式可以对已有的类别、汇总项按照四则运算方式组合成新的类别,并在结果中加以呈现,功能上要比添加小计更加灵活,不仅可以实现减法、除法、乘方等运算,而且生成的新类别不需要和原有类别相邻,而采用添加小计方式得到的汇总项就必须要和原有类别相邻。

设定完毕后子对话框应当如图 9.16 所示,单击“应用”按钮退出后,画布上的表格框架应当如图 9.17 所示。最后单击“确认”按钮,就可以得到所需要的数据表了。当然,该数据表在列宽、小计黑体显示等方面还有问题,这些修改的具体操作将在下节加以讲述。

		月份	
		类别 1	类别 2
\$TA3a	导致家庭经...	nnnnn.n	nnnnn.n
	改善:收入...	nnnnn.n	nnnnn.n
	改善:就业...	nnnnn.n	nnnnn.n
	改善:投资...	nnnnn.n	nnnnn.n
	改善:家庭...	nnnnn.n	nnnnn.n
	改善:政策...	nnnnn.n	nnnnn.n
	导致家庭经...	nnnnn.n	nnnnn.n
	恶化:收入...	nnnnn.n	nnnnn.n
	恶化:就业...	nnnnn.n	nnnnn.n
	恶化:投资...	nnnnn.n	nnnnn.n
	恶化:家庭...	nnnnn.n	nnnnn.n
	恶化:政策...	nnnnn.n	nnnnn.n

(a)

	200704	200712
导致家庭经济状况改善的原因	81.1	51.0
改善:收入相关	45.2	31.4
改善:就业状况相关	7.9	2.6
改善:投资相关	15.8	8.3
改善:家庭开支相关	5.1	3.9
改善:政策宏观经济	4.0	.9
导致家庭经济状况恶化的原因	18.9	54.6
恶化:收入相关	7.9	6.1
恶化:就业状况相关	5.1	5.2
恶化:投资相关	.6	.4
恶化:家庭开支相关	8.5	40.6
恶化:政策宏观经济相关	.0	.4

(b)

图 9.17 设定完毕后的 A3a 结果表格框架及最终输出的数据表(部分)

9.3.4 对话框的其他选项卡

前面主要介绍了对话框的“表格”选项卡,该对话框中还有另外 3 个选项卡,分别用于完成制表工作中的一些任务,使得最终得到的表格更为完美。

1. “标题”选项卡

“标题”选项卡用于设定标题、脚注(对话框中翻译为题注)、角注等,并且将日期、时间、表格框架表达式这 3 个可用的系统变量做成按钮放在最上方,用户直接单击相应按钮,即可将相应的宏代码写入相应框中,使用非常便捷。

2. “检验统计量”选项卡

“检验统计量”选项卡为所制作的表格提供了检验相应变量间关联的能力,如图 9.18(a)所示。具体的检验方式有以下 3 种。

(1) 比较列的平均值:当表格的列维度上有分类变量,而行维度上有连续变量时,则按列上分类变量的取值进行该连续变量各组均数的两两比较,具体为 t 检验。如果表格为叠加表,则分别进行叠加维度上每个变量类别间的两两比较。如果为嵌套表,则按照嵌套外层分类变量的各种取值,依次进行被嵌套在内部的分类变量各类别间的两两比较。用户可以自行设定检验中使用的 Alpha 水准。由于当类别较多时比较次数会很多,为了控制一类错误的大小,用户还可以选择使用 Bonferroni 方法进行 p 值的校正。

(2) 独立性检验:考察被配置在表各行、列上的分类变量是否独立,具体采用的是卡方检验。如果表格为叠加表,则分别进行叠加维度上每个变量和另一个维度上分类变量间的卡方检验。如果为嵌套表,则按照嵌套外层分类变量的各种取值,依次进行被嵌套在内部的分类变量和另一个维度上分类变量间的卡方检验。用户可以自行设定检验中使用的 Alpha 水准。

(3) 比较列的比例:当表格的行、列维度上都有分类变量时,则按照行维度的不同取值分别进行各列间构成比是否均衡的检验,具体方法为比较中的近似 z 检验。对叠加表和嵌套表的处理方式同前。用户可以进行 Alpha 水准的设置,也可以选择使用 Bonferroni 方法进行 p 值的校正。

因以上提到的各种检验方法大家尚未学习,因此这里不再列举相应的分析实例,仅指出一点:这里的结果输出比较特殊,为非常紧凑的组间差异结果输出格式。并且由于上述分析结果都是单独列表输出的,不能加入到主统计表中进行格式的自定义,因此实用性相对较差。

3. “选项”选项卡

“选项”选项卡如图 9.18(b)所示。



(a)



(b)

图 9.18 “检验统计量”与“选项”选项卡

“数据单元格外观”框组:用于进行空单元格和缺失统计量显示方式的设定。

“数据列宽度”框组:为该模块特有的功能,用于自定义数据列的宽度,如果数据较为特殊,或制表的要求较为特殊,则可以在此自定义列宽。

“尺度变量缺失值”框组:用于设定当连续型变量存在缺失值时对数据的利用方式,功能和统计描述过程中的相应框组完全相同,这里不再重复。

9.4 表格的编辑

前面已经基本完成了 A3 和 A3a 表格的制作,但和最终格式相比,还存在如下问题。

- (1) A3 表格中的类别标签文字还需要修改。
- (2) A3、A3a 表格中均存在加粗显示的行。
- (3) A3a 表格中有过多的横线需要删除。
- (4) A3a 表格中小计标签因为默认列宽不足被折行显示。

上述问题都可以通过对表格进行手工编辑来解决,下面来说明具体的操作。





可能有的读者会对表格编辑有其他建议:为什么不将表格复制到 Word 里面进行编辑, Word 表格编辑起来不是更容易吗?是的,在 Word 里面操作的确更加容易,但并非结果表格最终都会放置在 Word 里面使用,粘贴到不同的软件里,就需要用不同的编辑方式来操作,因此学习 SPSS 的表格编辑操作更为稳妥。另一方面,许多编辑操作可以存储为表格模板以达到自动化出表的效果,这在实际工作中是非常有用的。

9.4.1 基本编辑操作

1. 两种不同的编辑窗口

在对结果表格进行编辑前,显然需要首先进入它的编辑模式。相应的操作非常简单,只需双击选中的表格,就会进入编辑状态。但由于 SPSS 的系统设置不同,可能是在新窗口中进入编辑模式,也可能是在结果浏览器中嵌套进入编辑模式。一般而言,对于较大的表格,单个窗口的编辑模式在操作上要更方便一些。如果希望能控制相应的编辑方式,除了可以在系统选项中加以设定外,还可以在选中相应表格后右击,在弹出的快捷菜单中选择“编辑内容”→“在阅读器中”或者“在单个窗口中”菜单项,前者使用嵌套模式,而后者将使用打开新窗口的方式进入表格编辑状态。

进入编辑状态后,在默认情况下窗口中会同时出现浮动的编辑工具栏和透视托盘,用于方便用户进行编辑操作。透视托盘在 9.1 节已经介绍过,用于控制和修改表格框架。通过工具栏则可以对选定单元格进行文字格式、对齐方式等的设定,单击其左侧的按钮可以切换透视托盘是否出现,单击右侧的按钮则可以根据表格数据生成统计图形,包括线图、饼图等,如图 9.19 所示。

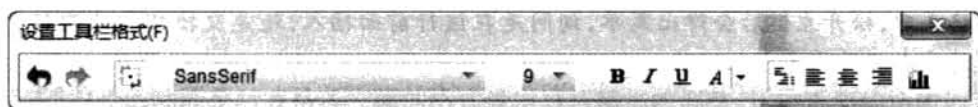


图 9.19 枢轴表的工具栏

工具栏自身是否出现是通过“视图”→“工具栏”菜单项切换的。

2. 表格元素的选择方式

在表格编辑中,单元格是基本的操作单位,包括表格标题和脚注均被看成特殊的单元格来处理。虽然根据所使用的表格模板设定不同,有的单元格间的分界线并未绘制出来,但它们在编辑操作中并不会被合并在一起,仍然是相互独立的编辑单位。

在对表格中的具体内容进行编辑操作时,显然应当首先将具体的元素选中,以使得系统得知相应操作是针对什么的。最常见的情形就是对单元格的选择,只需单击即可。不仅可以选中某个单元格,还可以选中其中的 1 行或 1 列,但首先要选中最上侧或左侧的标题单元格,然后选择“编辑”→“选择”菜单项,有 4 个选项:表格、表格主体、数据单元格、数据和标签单元格。在选中相应的单元格后,用户就可以对它们同时进行删除、复制、更改格式等操作,显然会方便得多。

3. 单元格内容的编辑

在题目 A3 生成的表格中,变量值标签还需要进行进一步修改。单击可以选中单元格,双击则进入单元格内数据的编辑状态,此时单元格内如果是数值,不仅会显示相应数据的全精度确切

值,还可以直接加以修改。图 9.20 演示了对变量的均数单元格进行编辑的全过程,显然在编辑中用户可以随意修改其中的内容,甚至于将数字改为无关的纯文本。

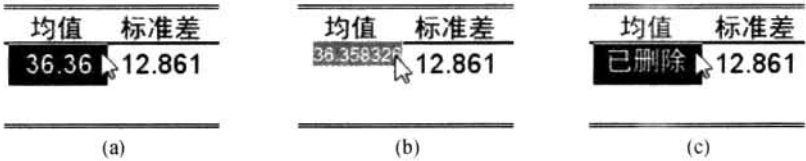


图 9.20 单元格编辑示意

4. 单元格位置的移动

单元格在表格中的位置并非固定不变,而是可以进行移动的。但是,为了保证表格内容不发生混乱,移动需要以行、列为基本单位进行,图 9.21 演示了如何进行行间的位置交换,首先选中行标题单元格,然后按下左键移动鼠标,可以看到鼠标携带着交换符号在移动。在到达合适的位置后松开左键,则该列会插入到示意的位置,最终操作结果见图 9.21 中最右侧的图。

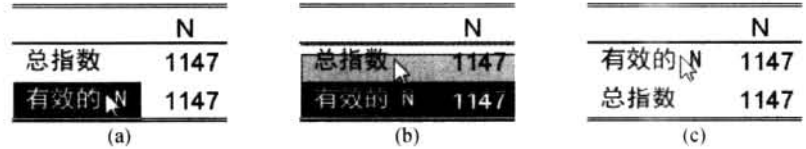


图 9.21 行交换操作示意图



这里图示的是 SPSS 20 版的轻置表输出中的交换操作。在 SPSS 19 及以前所使用的表格格式中,松开左键后会弹出菜单,询问是在该行前面插入,还是交换相应的两行,但两者的基本操作是完全相同的。

5. 列宽的更改

在题目 A3a 的表格中,小计标签因为默认列宽不足被折行显示。其实表格中的列宽也并非完全固定,而是可以自由拖动的。为了方便操作,可以首先选择“视图”→“网格线”菜单项,这样可以将单元格的分界线用虚线精确地表示出来。然后就可以用鼠标直接对列宽进行拖放操作了,具体的操作方式和在 Word 表格中一样,如图 9.22 所示。

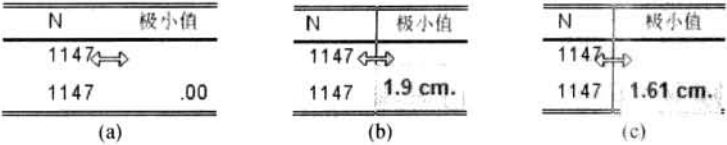


图 9.22 拖动列宽操作示意

除内容、位置和宽度外,单元格的其他属性也可以加以更改,后面将详细介绍。

9.4.2 主要编辑菜单功能

在用惯了 Windows 软件后,许多菜单功能都会无师自通,因此这里主要是对编辑过程中常用


菜单的功能进行讲解,除了非常复杂的操作外,不再进行具体的讲解。

1. “编辑”菜单

“编辑”菜单提供了复制、粘贴、删除、选择等常用的编辑操作,比较特殊的功能有以下几个。

(1) 分组/取消分组:用于给标题单元格加上、去掉亚组的标签,选中标题单元格这两个菜单项才会变黑,用户可以将相应的组标签改为自己想要的名称。

(2) 拖放复制:相当于一个切换按钮,选中该选项会使对单元格的拖动成为复制操作,反之,则会弹出关联菜单,确认是和当前单元格交换还是插入。

(3) 创建图形:该菜单项可以将统计表中的内容以图形的方式立体地呈现在面前,功能等同于前面提到的工具栏上的按钮。

2. “视图”菜单和“插入”菜单

“视图”菜单用于切换表格中各元素的显示/隐藏,几个菜单项分别控制了编辑工具栏、表格维度标签、类别标签、脚注和单元格网格线的显示。“插入”菜单用于插入新的标题、说明、脚注等。

3. “透视”菜单

“透视”菜单的功能是改变结果表格的显示方式。

(1) 重新排序类别:可以对行、列标签重新排序,首先在表头区域选中希望移动/插入的位置,然后在该菜单中选择希望移动/插入的类别。当然,熟悉操作的读者完全可以用上面提到的拖放操作来达到相同的效果。

(2) 行列转置:用于实现表格的行列转置操作,该菜单项在表格太宽时非常有用。

(3) 透视托盘:透视托盘的显示开关。

4. “格式”菜单

“格式”菜单的功能是对表格各方面的格式设定进行精细的调整,比较重要的功能如下。

(1) 单元格属性:对选中单元格的字体、阴影、颜色等属性加以更改。

(2) 表格属性:对表格进行各个选项的精细设置,如字符格式、边框样式等。

(3) 表格外观:可以在这里直接更换表格模板,但所做的选择只对当前表格生效。

(4) 自动调整:表格的行、列宽会自动按内容的多少调整为最小。

除以上功能外,其余各菜单项的含义均非常明确,这里不再详述。

9.4.3 表格属性的详细设置

在各种编辑功能中,相对比较复杂,但是又非常常用的是选项卡格式的“表格属性”对话框,因此下面将对其功能进行专门介绍。

1. “常规”选项卡

(1) “常规”框组:显示或隐藏空的行/列,以及控制在长表中显示的默认行数。

(2) “行维度标签”框组:用于控制行维度标签的显示格式,可以位于左上角或嵌套。

(3) “列宽度”框组:用于控制最大、最小行/列标签宽度。

图 9.23 所示的右侧“样本”框组会即时显示相应更改的效果。

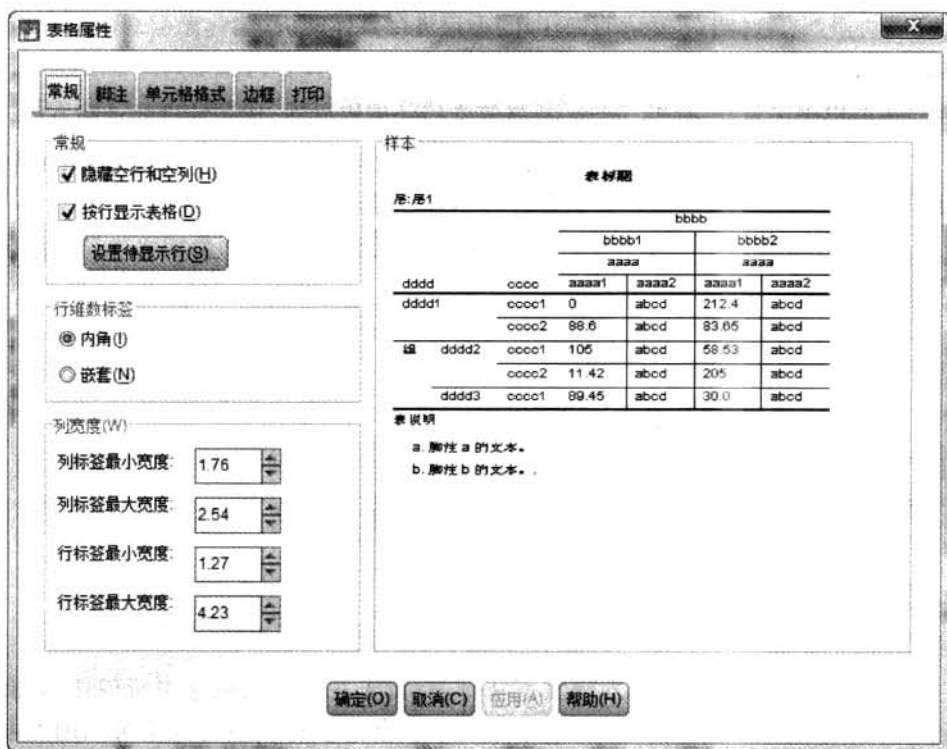


图 9.23 “表格属性”对话框的“常规”选项卡

2. “脚注”选项卡

设定表格中脚注的显示格式,可以将脚注序号设定为字母顺序或者数值顺序,具体位置可以是右上角或右下角,右侧会即时显示相应更改的效果。

3. “单元格格式”选项卡

“单元格格式”选项卡可以设定表格中单元格的基本显示格式。左半侧从上到下依次用于设定单元格的字体、对齐方式、阴影及颜色、边距。右侧则用于选择具体的单元格区域,并显示出相应的格式设定。注意 SPSS 表格将单元格分成了若干组,每组单元格只能使用相同的格式设定。在使用该选项卡时,首先应当在右侧的“区域”下拉列表框中选中相应的单元格区域,然后才能进行相应的设定,如图 9.24(a)所示。

4. “边框”选项卡

“边框”选项卡用于进行表格中各种框线的格式设定,左侧的“边框”列表框中列出了表格中全部框线的名称,右侧则为相应的示意图,在左侧选择名称和在右侧示意图中单击均可选中相应框线,选中后在左下角的两个下拉列表选择线型和颜色,如图 9.24(b)所示。

5. “打印”选项卡

“打印”选项卡用于进行表格打印时的设定,因目前在国内较少使用其中的高级功能,这里不再详述。

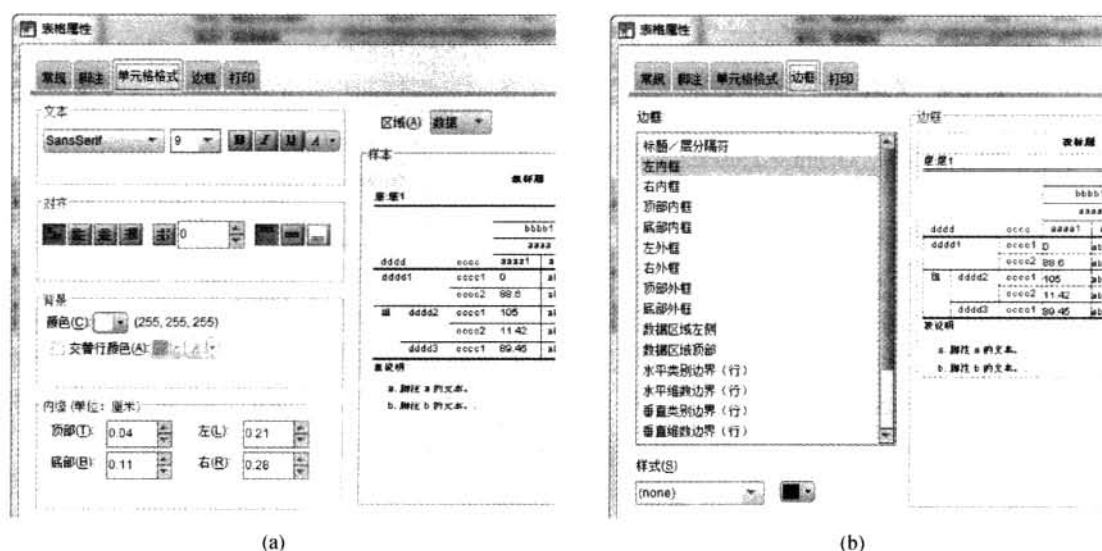


图 9.24 “表格属性”对话框的“单元格格式”选项卡和“边框”选项卡

9.5 表格模板技术

9.5.1 模板技术简介

上面已经详细讲解了如何对表格进行编辑,但是,所有的编辑操作都只是针对当前表格进行的,对于一个新绘制的表格,SPSS 仍然会使用默认设定的表格格式进行输出。大家可以设想这样的一种场景:该项目中共需要绘制 1 000 个表格,具体的格式都是统一的,但是和 SPSS 默认的格式不相同。如果进行这样一张表格的格式编辑需要 5 min,那么,1 000 张表格就需要 5 000 min,合计 80 多个小时!显然,如果能够有一种方法将所需设定保存下来,并且使得 SPSS 输出的全部表格均自动使用该设定绘制,将会大大减轻相应的工作量。

使用模板技术就可以达到上述目的。所谓表格模板指的是存储了表格框线、单元格字体、颜色等设定的一种特殊格式的文件,SPSS 可以读取其中的设定值,并将其应用于当前表格。

1. 为当前表格应用、存储不同的表格模板

除默认的表格格式外,在 SPSS 中还预制了一大批其他样式的表格模板,如果希望为当前表格更换一个新的模板,则选择“格式”→“表格外观”菜单项,打开的对话框如图 9.25 所示,左侧列出的就是所有可用的表格模板,右侧则为相应格式的示意图。用户只需要在左侧列表中选中合适的模板名称,然后确认即可,此时就可以看到当前表格已经被更改为相应模板的设定格式。

“表格外观”对话框还可以用于将当前表格的格式设定存储为一个新的模板,供其余表格使用。注意对话框下方一排的 3 个按钮,“保存外观”按钮用于将格式的更改存储到当前使用的模板文件中,“另存为”按钮用于将当前格式存储为一个新的模板文件。“编辑外观”按钮则用于继续对现有表格的格式设定进行更改,单击后会打开“表格属性”对话框。

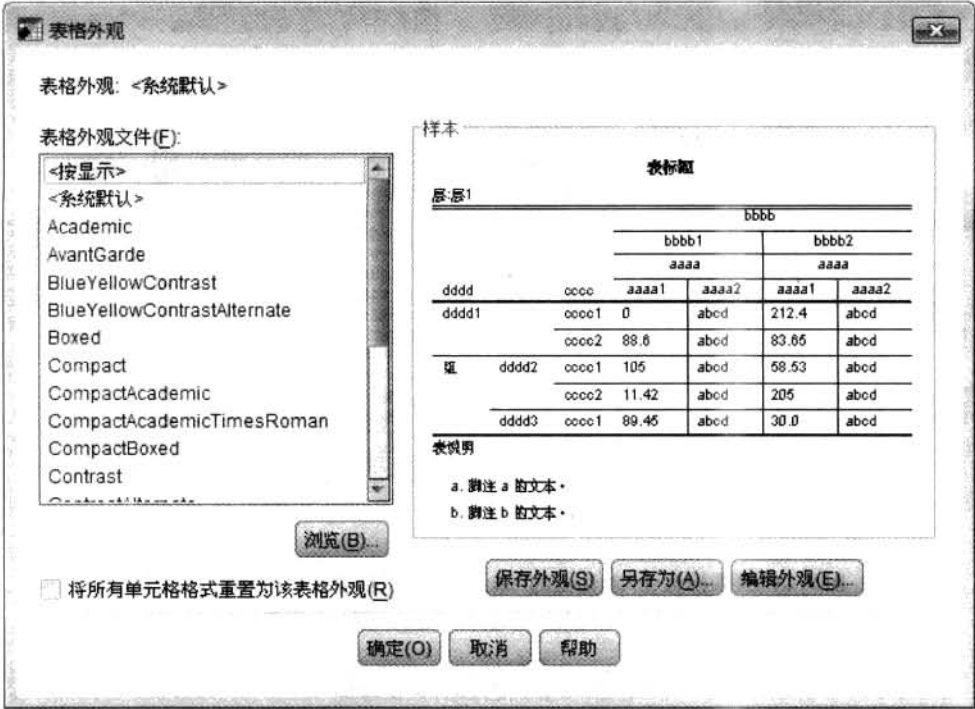


图 9.25 “表格外观”对话框

2. 将表格模板设定为系统默认值

通过上面的操作,已经可以将需要的格式设定保存为表格模板,然后再将其应用到别的表格上去,这虽然大大节省了工作时间,但是当需要操作的表格数量较多时,仍然非常麻烦,此时可以将相应的模板设定为系统默认的表格格式,从而在指标程序执行时就完成相应的表格格式设定工作。具体的操作在“系统”选项卡中进行,在 SPSS 中选择“编辑”→“选项”菜单项,在打开的对话框中选择“枢轴表”选项卡,在该选项卡中首先使用“浏览”按钮找到希望使用的表格模板文件,然后将该模板名称设定为“默认表格外观”即可,确定后 SPSS 输出的所有表格将均使用该模板的格式设置。

在各种 SPSS 预设的模板中,以 Academic 模板和 Report 模板最为有用,如图 9.26 所示,Academic 模板只保留了主要的横线,完全符合统计学中的统计表格要求,实际上就是统计学中最常用的三线表模板。而 Report 模板更进一步,只保留了分隔表头和表格正文的横线,是调研报告中最常见的表格格式。笔者在此建议大家尽量使用这两种模板,以养成良好的表格格式习惯。

	N	极小值
总指数	1147	.00
有效的N	1147	

(a)

	N	极小值
总指数	1147	.00
有效的N	1147	

(b)

	N	极小值
总指数	1147	.00
有效的N	1147	

(c)

图 9.26 枢轴表的默认模板、Academic 模板和 Report 模板格式

9.5.2 表格的中文兼容问题的解决

SPSS 目前的最新版本已经完全兼容中文,将枢轴表直接复制并粘贴到 Word、Excel 等软件中,相应的中文字符都不会变成乱码。但是对于仍然在使用老版本的用户而言,可能还是会遇到乱码问题。对此可以采用以下 3 种解决方案。

(1) 在粘贴过来后重新输入全部的中文,显然,这是最简单,也是最麻烦的办法。

(2) 去除表格的全部格式,以纯文本格式进行表格内容的粘贴。在 Word 中选择“编辑”→“选择性粘贴”菜单项,然后选择其中的无格式文本,这样整个表格就会按照 Tab 键分隔的纯文本形式粘贴到 Word 中,里面的中文也完好无损。再将文本选中,选择“表格”→“转换”→“文字转换到表格”菜单项即可。

(3) 在模板中加以设定。这种方法较为复杂,但一劳永逸。上面曾经提到可以在 SPSS 中自行设定默认的表格模板,因此只需要将相应模板中可能出现中文的区域的字体设定为中文,在保存后将其设定为默认表格模板,之后在使用 SPSS 表格的时候就再也不会出现中文乱码的问题了。

思考与练习

自行完成本章中涉及的对 CCSS 案例数据的制表操作。

第 10 章 数据的图形展示

第 9 章介绍了如何制作复杂和精细的统计报表。报表可以对数据细节做出精确呈现,但其缺点在于不够直观,阅读者很难立刻抓住主要的数据特征。统计图的特点则正好和报表相反,图形可以直观地反映数据的主要特征,但对数据细节的呈现却会很困难。只有将图表结合起来,才能使得呈现的数据最为全面和清晰。本章将介绍数据的图形展示技术。



由于统计图对数值的呈现稍显粗略,因此当所展示的数据大小较接近时最好考虑采用统计表。如果一定要用图形来呈现,则可在图中标出具体数值备查。

对于不同地域、月份、人口背景特征的消费者,要初步了解其信心指数存在着怎样的差异,虽然完全可以用其他统计描述方法完成,但本章将首先来说明利用统计图可以达到的效果。

10.1 统计图概述

统计图能够简洁、直观地对主要的信息数据进行呈现。针对这一特点,制作统计图有两个基本要求:一是正确,二是简洁,以反映事物内在的规律和关联。

10.1.1 统计图的基本框架

一个完整的统计图大致可以被分解为标题区、图例区、数据区等多个部分,如图 10.1 所示,下面就按此一一进行介绍。

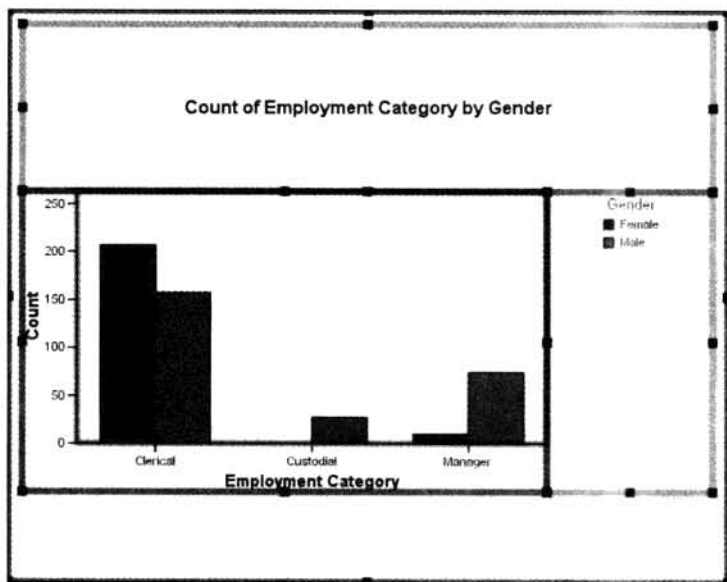


图 10.1 统计图结构示意图

1. 标题区和注解区

标题区和注解区这两个区域分别位于图形的最上方和最下方,位于图形最上方用于列出图形名称的就是标题区,如图 10.1 中写着“Count of ...”的部分。标题中一般应注明图的编号,标题内容则简明扼要,用于说明资料的内容、地点、时间等。图 10.1 中最下方空白的区域即为注解区,主要用于添加对图形内容的简单说明,一般文字不宜过多,讲清楚即可。

需要指出的是,由于习惯不同,国内出版物中的统计图一般要求在图形的正下方给出标题,本书也是如此。在这种情况下,如果再添加注解,就会使图形显得不太对称,因此往往会将注解改为正文中的一段文字叙述。因此读者在实际绘图时,往往不会使用标题区和注解区的功能,而是自己另行添加。本书的实例也基本上不会用到这两个区域。

2. 坐标轴

坐标轴、图形本身(绘图区)在内的区域一般被统称为数据区,是统计图的主要部分,这里将分开讲述。坐标轴用于表示相应变量的取值情况,由于二维统计图最为常用,相应的两个坐标轴往往被直接称为横轴和纵轴。实际上应当按照所表示的数据类型将坐标轴分为连续轴和分类轴两大类,如图 10.1 中所示的横轴就是分类轴,其数轴刻度间无大小之分,仅代表不同的类别。其纵轴则为连续轴,刻度严格而准确地表示了数量上的差异。连续轴和分类轴的编辑功能相差极大,但与其位于横轴还是纵轴则完全无关。

坐标轴一般都应注有标目,用于说明其表示的具体含义。对于连续轴而言,往往还需要注明单位,如年份、克、% 等。连续轴的刻度设定应该是等距的,而且在一般情况下为算术等距,但必要时也可以是几何等距,以满足特殊的分析需求,如图 10.2 所示。纵横尺度一般从 0 开始(对数线图、点图例外),以免曲解统计图所表示的指标关系。

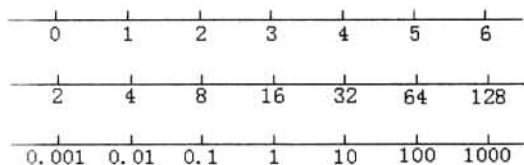


图 10.2 算术尺度、几何尺度和对数尺度的连续轴



3. 绘图区

绘图区指的是被坐标轴包围,直接使用图形元素来对数据进行呈现的区域,在 SPSS 中也被称为内框区,以和表示整个图形范围的外框区相对应。绘图区中主要有表示变量数值情况的直条、区块、点、线等图形元素,使用者在阅读图形时需要首先注意相应的各坐标轴的具体含义,以明确各图形元素的坐标究竟表示的是数量大小,还是类别。如图 10.1 中所示的不同直条(横轴为分类轴)表示的是汽车的不同产地,而直条的高低(纵轴为连续轴)则表示了具体指标(这里为 MPG、ACCEL)的算术均数。

除了基本的图形元素外,绘图区中还可能出现各种文字注解、辅助坐标线等用于方便图形阅读的元素。

4. 图例区

图例区位于整个图形的右侧,当在图形中需要使用不同的颜色、线形等将图形元素分组以表

示不同类别时,就需要在图例中对此加以说明了。以图 10.1 为例,图例中填充格式为“”的直条表示变量 MPG 的均数,则表示变量 ACCEL 的均数。当然,出于美观和使用习惯上的考虑,使用者往往会将图例加以移动,最常见的位置是右上方。

以上介绍的一个完整的统计图中可能被划分出的各种结构,实际上,这些结构并非在所有的统计图中都会出现,例如标题区和注解区就往往不会用到,而如果不存在图形元素分组的问题,则图例区也不会出现。一般而言,由坐标轴和绘图区所组成的数据区是一个统计图的核心部分,一般都会出现,其余部分则都是根据需要而有选择地加以使用的。

10.1.2 统计图的种类

统计图的分类方法有许多种,但和统计学体系最为贴近的分类方法是首先按照其呈现变量的数量将其大致分为单变量图、双变量图、多变量图等,然后再根据相应变量的测量尺度进行更细的区分。本节就将按此进行讲述,毕竟大家是在学习一种统计软件而不是绘图软件。虽然这种分类方法会将许多图形分成更细的小类,但是这样更有利于将来正确使用。



在 SPSS 中创建图形时,变量的测量尺度很重要,如果对变量的测量尺度定义有误,则可能无法生成相应的图形。目前 SPSS 将绘图用变量主要分为如下 3 类:无序、有序和连续性变量。但同时又将多选题变量集作为一类特殊的无序变量进行处理。

1. 单变量图:连续性变量

单变量图指的是通过图形元素的位置高低、范围大小等对某一个变量的数值/类别分布情况进行呈现,常用于描述、考察变量的分布类型。绘制这类图形时只需一个变量。

一个连续性变量的分布特征描述最常用的图形工具就是直方图,如图 10.3(a)所示,它通过直条在各个取值区段的分布范围和长度来直观地显示连续变量的数量分布规律,图形中的横轴代表不同的取值区段,而纵轴则表示相应区段的频数。对于样本量较小的情形,直方图会损失一部分信息,此时可以使用茎叶图来进行更精确的描述。

除直方图外,箱图也常用于连续性变量的描述,如图 10.3(b)所示,它主要使用百分位数指标,如中位数、四分位数等对该变量的分布规律进行呈现,还可帮助进行对称性、极值判定。

对于更为深入的统计分析,研究者往往还希望考察该连续性变量是否服从某种理论分布,例如考察其是否服从正态分布。除了进行假设检验外,利用 P-P 图(如图 10.3(c)所示)和 Q-Q 图也可以达到这一目的,实际上这种图形在前面几章中已经出现过。

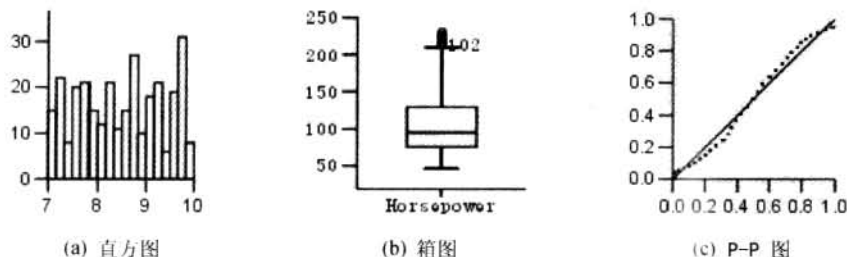


图 10.3 用于描述连续变量的几种常用单变量图示意

2. 单变量图:分类变量

对分类变量的描述可以分为两种情况:展示分类变量各类别的频数,或者表示各部分占总体的构成比例。对于前者,最常用的工具是简单条图(如图 10.4(b)所示),它使用等宽直条的长度来表示相互独立的各类别的频数高低,换言之,横轴表示不同的类别,而纵轴则和直方图一样,也用于表示频数的多少。

在表示各部分的构成情况时,饼图是最常用的工具,如图 10.4(a)所示,它使用饼块的大小来表示各类别的百分比构成情况。

对于一些特殊的问题,研究者可能希望在一幅图中同时表示该变量各类别的原始频数和百分构成,Pareto 图(如图 10.4(c)所示)就可以满足这一要求,它在图形中使用直条代表频数高低,同时又使用折线来表示累计百分比的变化情况。

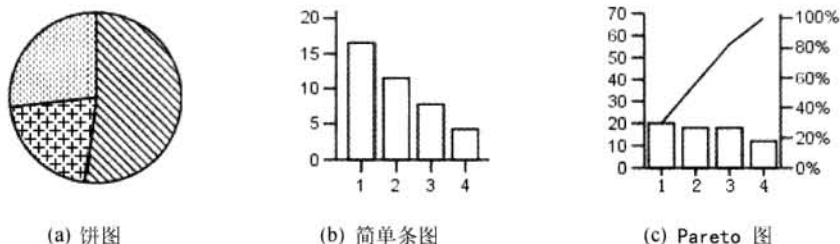


图 10.4 用于描述分类变量的几种常用单变量图示意

3. 双变量图:连续应变量

顾名思义,绘制这类图形时需要两个变量,而图形也主要用于呈现这两个变量在数量上的联系方式,或者说当一个变量改变时,另一个变量会如何变化。该图形常用于对不同亚群(Sub-group)的研究对象进行比较。

为了方便起见,这里首先考虑应变量为连续变量的情形。此时应变量一般会使用纵轴刻度的高度加以呈现,而人们实际关心的指标可能是其均数,或者标准差等。当另一个主动变化的变量(自变量)为无序分类变量时,所用的图形工具实际上还是简单条图,只是此时每一个直条的高度代表的是相应类别的该应变量统计指标的高低。

当自变量为有序分类变量,特别是代表年代或时间时,统计学中习惯上用线图来对其进行关联呈现,用于直观地表现随着有序变量的变化,相应的应变量指标是如何上升或下降的。显然,这一问题用条图似乎也是可行的,但这主要是一个使用习惯的问题。最后,如果自变量也是连续性变量,则所用的工具就是大家所熟悉的散点图。它使用散点的疏密程度和变化趋势来对两个连续变量间的数量联系进行呈现。

4. 双变量图:分类应变量

当应变量为分类变量而自变量为连续变量时,目前尚没有很好的图形工具可以利用,常见的处理方式是将自/应变量交换后使用条图来进行呈现。当自变量也是分类变量时,实际上所使用的图形工具是比较单一的,基本上以条图为主。但是,按照其具体呈现方式,又可分为复式条图、分段条图和马赛克图 3 种,复式条图重点呈现两个分类变量各个类别组合情况下的频数情况,分段条图则主要突出一个分类变量各类别的频数,并在此基础上表现两个类别的组合频数情况。

马赛克图也是以一个分类变量为主的,它呈现的是在一个变量的不同类别下,另一个变量各类别的百分比变化情况。10.5节将会对这些图形进行详细的讲解。

事实上,以上所介绍的仅仅是最为正规和常见的双变量统计图,实际上,在掌握了单变量图的特性后,完全可以将其加以充分利用,在自变量为分类变量时,分类别绘制相应的单变量图进行数值特征的呈现,以达到对数据更为充分和深入的展示。最常见的情况有分组箱图、复式饼图、直方图组等,对此感兴趣的读者可参见相应图形的详细介绍,这里不再详述。

5. 多变量图

当在一幅图形中需要呈现出3个甚至3个以上变量的数量关联时,所构成的图形就称为多变量图。一般而言,由于一个坐标轴只用于呈现一个变量的数值特征,因此用最常见的二维平面统计图表示两个变量的特征是比较合适的。如果要表现3个变量的关联,最好的办法是采用三维坐标的立体统计图。但是,由于实际上还是在纸平面或者显示器平面上对三维图进行呈现,立体图在使用上并不方便。因此,当其中有变量为分类变量时,统计学家往往采用图例这一方式对二维图进行扩充,使二维图能够表现出更多的信息。例如在散点图(如图10.5(b)所示)中用点的形状或者颜色区分不同的类别,这样实际上就在一幅带图例的散点图中同时呈现了两个连续变量和一个分类变量的数量关联信息。类似的图形还有多线图等。当然,如果所有变量均为连续变量,则图例并不能解决问题,仍然需要使用高维的散点图才能对其关系加以呈现。为了方便分析对高位散点图的观察,SPSS中也提供了一系列的功能,如散点图矩阵、立体散点图的动态旋转等,详见10.7节。

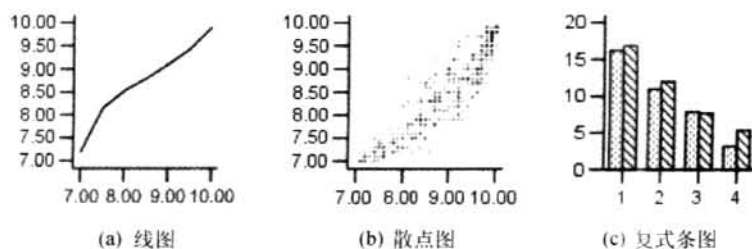


图 10.5 几种常见的多变量图示意



读者在具体应用多变量图时要注意“度”的问题,切勿将统计图做得太复杂,因为这样会丢失统计图“直观明了”的优点,那样将得不偿失。

6. 其他特殊用途的统计图

除了以上可按照统计原则加以归类的图形外,针对一些特殊的应用领域和分析目的,SPSS还提供了一系列的专用统计图,它们或者用于满足某一个行业的特殊需求,或者用于解决某种专门的统计分析问题。前者如用于将统计数据与地域分布相结合的统计地图、用于工业质量控制的控制图(如图10.6(a)所示)、用于股票分析的高低图,后者的例子有用于描述样本指标可信区间或分布范围的误差条图、用于诊断性试验效果分析的ROC曲线(如图10.6(b)所示)、用于时间序列数据预分析的序列图等。对于这些工具本书将会有选择地在相应章节加以介绍。

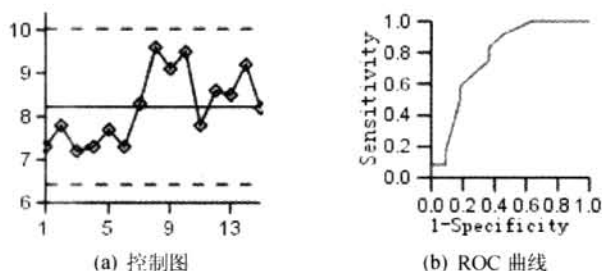


图 10.6 特殊用途的统计图示例

10.1.3 SPSS 的统计绘图功能

1. SPSS 统计图的 3 种版本

近 10 年来,SPSS 的统计绘图系统一直处于不断的演进之中,先后出现过 3 种版本:12 版以前的统计图系统、8 版之后新增的交互图系统,以及目前采用的统计图系统。这 3 种绘图系统在图形种类、操作、编辑等各方面都各不相同,给用户(也给教材编写者)带来了较大的困扰。好在目前 SPSS 已经彻底将统计绘图系统统一为一个版本,本书也将只针对现在的绘图系统加以讲解。如果读者对老版本统计绘图系统,以及交互图的详细操作感兴趣,希望进一步了解,可以参见《SPSS 11 统计分析教程》(基础篇)以及《SPSS 统计分析基础教程》中的相关章节,也可参考 SPSS 的用户手册。

2. 统计图的 3 种对话框操作方式

虽然统计图系统已被统一为一种,但目前 SPSS 在绘制统计图时还是保留了新老两种操作对话框,而新对话框更有两种界面:

(1) 可视化的图形生成器:类似于第 9 章介绍的画布式的全交互对话框,可以采用非常舒服的拖放方式操作,并且每一个对话框元素的可操作性都大大强于普通对话框,以前需要两至三层对话框才能完成的工作,现在只要在一层对话框中就可以完成了。不仅如此,由于几乎全部统计图的操作都被统一在这一界面之中,因此用户的学习和操作效率也会大大提高。

(2) 图形的可视化模板:是一个类似于绘图向导的可视化界面,会根据用户所选择的变量数量和测量尺度自动给出可供绘制的图形供用户选择。实际上,该界面的很多操作细节非常类似于交互图。

(3) 继承自老版本的传统对话框:属于标准的 SPSS 对话框,每一种或者每一类图形均提供不同的操作界面。对于 SPSS 的老用户而言,操作该系列对话框是不需要重新学习的。

出于控制篇幅的考虑,本书将只介绍可视化图形生成器的操作,可视化模板因对话框较为简单,用户可自行学习。至于老版本对话框的操作方式,对此感兴趣的读者可以参见《SPSS 统计分析基础教程》中的相应内容。

10.2 直方图与茎叶图

直方图(Histogram),用于表示连续性变量的频数分布,在实际应用中常用于考察变量的分

布是否服从某种分布类型,如正态分布。在直方图中以各矩形(直条)的面积表示各组段的频数(或频率),各矩形的面积总和为总频数(或等于1)。若各组段组距不等,则以各组段组距除该组段频数之商为矩形的高度,以该组段的组距为矩形的宽度,以保证矩形的面积等于该组的频数。

10.2.1 案例:绘制消费者信心值的直方图

例 10.1 对总样本的消费者信心值绘制直方图,以考察其是否服从正态分布。

本例明确了所需绘制的图形种类,因此操作上并无歧义,选择简单直方图即可。但是由于目的是考察变量 index1 是否服从正态分布,因此最好能在绘图时一并加绘正态曲线以便于比较。

1. 界面说明

选择“图形”→“图表构建程序”菜单项,就会打开“图表构建程序”对话框主操作界面,如图 10.7 所示。该界面实际上也是一个多层选项卡界面,只是选项卡被压缩到了绘图画布区下方而已。

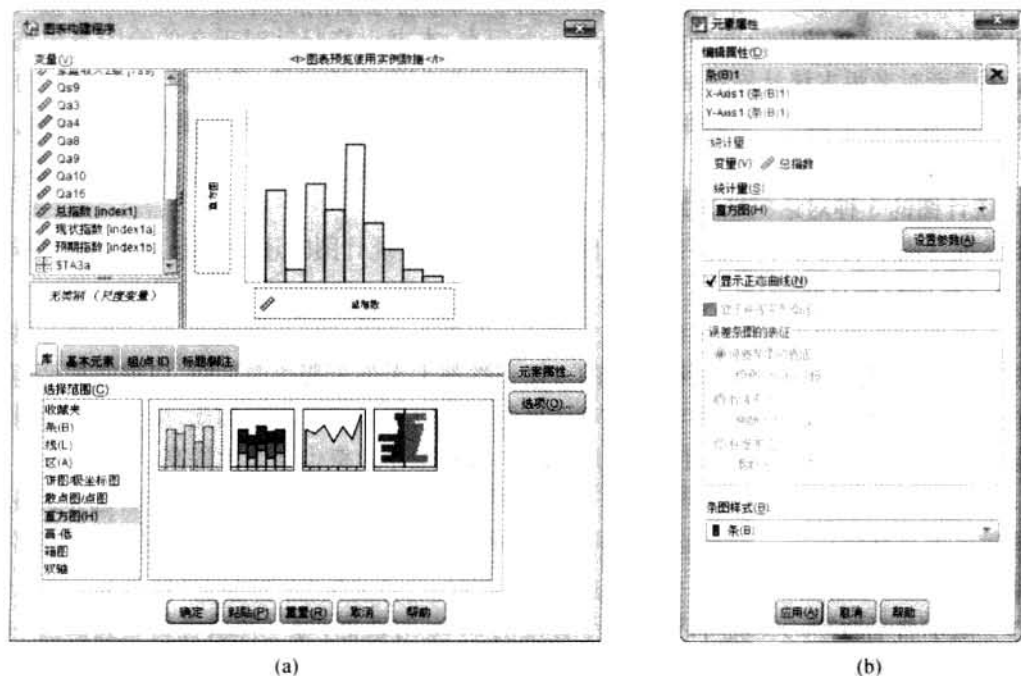


图 10.7 “图表构建程序”对话框与“元素属性”对话框

(1) “变量”列表框:位于左上角,会列出绘图中可用的所有变量,如果有多选题设定,则会显示在列表最后。用户可采用拖放操作将相应变量/多选题变量集拖入右侧的画布区域。

(2) 绘图画布(Canvas):在界面中部占据绝大部分空间,类似于画家绘画时的空白画布,用户在制图时就是在这张空白画布上进行拖放操作的,以最终得到合适的图形。注意在画布上有一些用虚线标出的放置区,变量只能被拖放入这些区域中。目前图中显示的是两个数轴放置区。


根据绘制图形的种类不同,还会有分组放置区(如复式条图或堆积条图)、面板放置区和点标签放置区等出现。

(3) 类别列表框:会在“变量”列表框中选中分类变量时自动列出所有的类别取值/标签,图中因选中的是连续变量 index1,所以没有任何显示。

(4) “库”选项卡:用于列出图库中的候选图形。图库中将图形按照基本特征分成了若干组(范围),用户可在左侧先选择图形范围,然后在右侧列出的图标中选择所需的图形。

(5) “元素属性”对话框:会在画布中选入变量后自动打开,也可使用界面右侧的“元素属性”按钮切换其显示和隐藏。该对话框用于对图形元素的种类、统计量设定、元素显示格式等进行详细设定,在最上方的“编辑属性”列表框中所选中的图形元素不同,其下方所显示的选项也会有很大差异。

(6) “选项”按钮:用于对缺失值、图形模板等进行设定,一般较少使用。

 图 10.7(b) 所示的“元素属性”对话框显示的是对直条属性的设定界面,该界面的选项在以直条为基本图形元素的图形,如直方图、箱图、条图中相似,因此读者应仔细观察该界面中的内容,下文将不再对其进行重复讲解。

2. 具体操作

本例操作非常简单,具体如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“直方图”组,将右侧出现的简单直方图图标拖入画布中。
- (3) 在变量列表中找到 index1,将其拖入画布的横轴框中。
- (4) 在“元素属性”对话框中选中“显示正态曲线”复选框,注意随后一定要单击下方的“应用”按钮,否则相应的操作不会生效。

最终生成的图形如图 10.8 所示,可见 index1 的分布还是非常接近正态曲线的,只是左侧稍

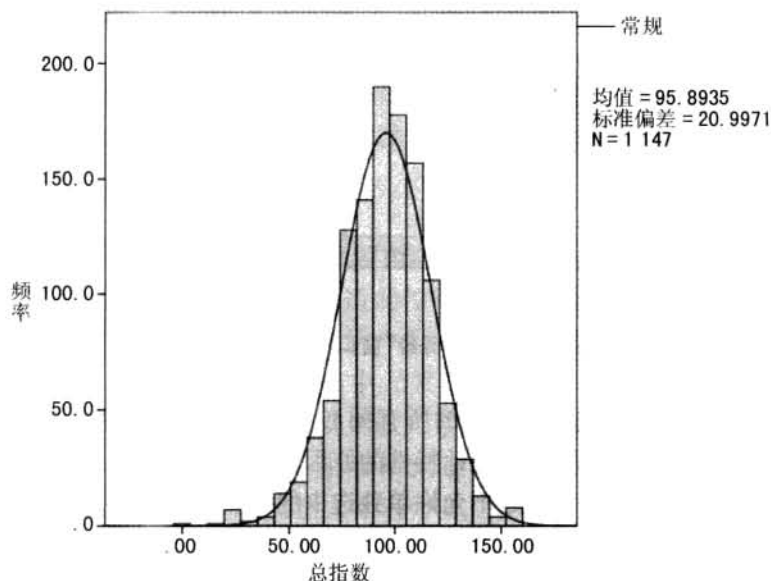


图 10.8 绘制完成的信心指数直方图

有拖尾,也就是有几个偏低的极端值存在。图形右侧还自动给出了样本的均数、标准差和样本量,以便于使用者全面了解样本情况。



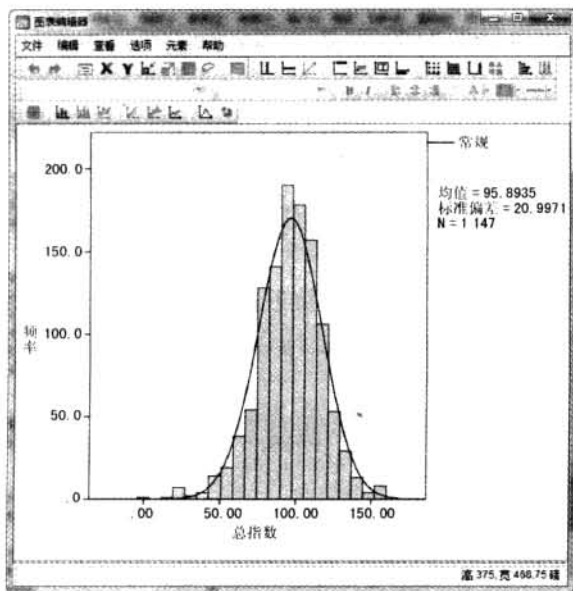
本章基本上都是按照先绘制出默认图形,然后再进行图形编辑的顺序加以讲解的,如果希望在绘制图形时就能够控制连续轴选项、直条分段方式等非图形特征选项,则可以在“元素属性”对话框上部的列表中先选中相应的数轴、直条栏目,然后在其中部和下部进行相应的参数设定,此处不再详述。

10.2.2 图形的基本编辑操作

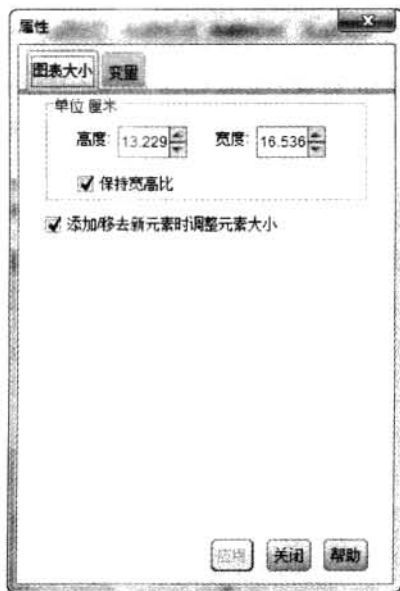
10.2.1 小节输出的图形无论是从统计学的要求上讲,还是从美观程度上讲都存在着很大的问题,好在 SPSS 赋予了使用者相当大的自主权,可以对图形进行全方位的编辑。

1. “图形编辑器”窗口

在结果窗口中双击欲进行编辑的统计图,就会打开一个独立的“图形编辑器”窗口,如图 10.9 所示。一般与之相配套出现的还有一个“属性”子对话框。



(a)



(b)

图 10.9 “图形编辑器”窗口与“属性”子对话框

(1) “图形编辑器”窗口:在该窗口中的所有图形元素都可以被单独选中或者成组选中,如果该元素只能被成组选中,则说明由于统计特性的原因不允许对该元素进行单独编辑,例如直方图中的直条组、数轴的刻度标签等都是不允许单独选中的。

(2) “属性”子对话框:为多选项卡界面。对应着“图形编辑器”窗口中被选中的元素种类,该子对话框中出现的选项卡种类也会发生变化。大多数编辑操作都要在此对话框中进行。如果

该对话框被关闭,则可以选择“编辑”→“属性”菜单项将其重新打开。


2. 图形编辑的基本操作要点

这里首先列出图形编辑的基本操作要点如下。

(1) 选择图形元素:统计图中的各种图形元素,如散点、直条、数轴等,都会按照其统计特征进行编组,在选择这些元素时,基本规律是第一次单击会选中图中的所有同类/组元素;在原位置上两次单击,则会变为只选中该图形元素本身(在使用图例时,两次单击会选中同组元素,3次单击才会选中元素本身)。如果希望选择不同的多个图形元素,则按住 Ctrl 键分别选择即可。对于所选中的一个或者一组图形元素,用户可同时对其进行相同的格式编辑操作,如颜色、填充样式,甚至于单独标出具体的数值、ID 号等。

(2) 进行文本编辑:首先用鼠标单击选中文本,对于可编辑的文本元素,再次单击则进入编辑状态,可对其自由进行内容、格式、字体等的编辑。

(3) 移动图形元素或改变其大小:同样需要先用鼠标选中相应元素,然后视选中框四周是否出现控制柄来确定该元素能否被移动或改变大小,对于无控制柄的选中框,相应元素是无法移动的,否则就可以移动或改变大小,详见下面介绍。

 文本的编辑设定一般分为 3 种情况:对于数轴刻度等移动/更改可能导致图形误读的文本元素,既不能移动,也不能编辑;对于图例文字等,则可以编辑,但不能在图例内部进行相对位置的移动;其他不太重要的文本元素则既可以随意移动,也可以进行内容编辑。

图形元素的位置设定也大致分为 3 种情况:数轴刻度等重要图形元素的位置完全固定,以保证基本的图形特征;图形中的坐标参考线等的位置为半固定,不能使用鼠标随意拖动,而只能在选项卡中输入坐标进行精确定位,以保证位置的准确;而对于一般性的文字注解、标签等,用户可以在图形中进行随意拖动。

3. 更改图形长宽比例

默认图形是按照 1:2 的长宽比绘制的,相对而言有点窄。如果希望更改为更宽的 3:4 等比例,则在“属性”子对话框的“图表大小”选项卡中取消选中“保持宽高比”复选框,然后在上方的“高度”、“宽度”数值框中自由输入希望设定的图形大小,再单击“应用”按钮即可,本例更改为 12:16。

4. 图例元素的位置移动和改变大小

除长宽比不合适外,默认绘制的直方图由于右侧完全被统计量图例占据,导致图形看起来更窄,完全可以将该图例拖动到更合适的位置上去:用鼠标将整个图例元素选中,则会出现如图 10.10(a)所示的带 8 个控制柄的方框,将光标移动到框线上,则光标会变为十字形,此时按下鼠标左键即可随意移动所选元素的位置。如果将光标移动到控制柄上,则光标变为双向箭头形,此时按下左键可以更改元素的大小,如图 10.10(b)所示。在移动位置或改变大小时相应区域内的文本大小和格式设置不会改变,只会随着区域的形状“流动”。而其中的图形元素则会自动调整大小和形状,如改变直条的长度、宽度等,以达到最佳的显示效果。



图 10.10 图形元素的移动和改变大小

至于图例中的文字大小、种类、颜色等属性,则可以在窗口上方的格式工具栏中更改,这里不再详述。

5. 更改背景色、直条颜色、边框等图形元素属性

如果希望更改如颜色、线型等图形元素属性,则首先需要在图形中选中相应的图形元素。当选中不同的图形元素时,“属性”对话框会同步发生改变,及时给出可用于编辑该元素的各种选项卡,如更改填充格式、线型、颜色的选项卡等,如图 10.11 所示,可见其共同特征是可以进行颜色的更改,此外还各自有一些特征性的编辑选项,如图形区块可以更改填充样式、边框样式,线条可以更改线型和粗细,而文本则可以更改字体、大小和对齐方向等。用户可以在对话框上自行选择所需的格式,最上方的“预览”栏中可以直接显示出相应的效果。读者可以自行操作,去掉背景色,并且将直条颜色更改为更醒目的红色,这里不再详述。

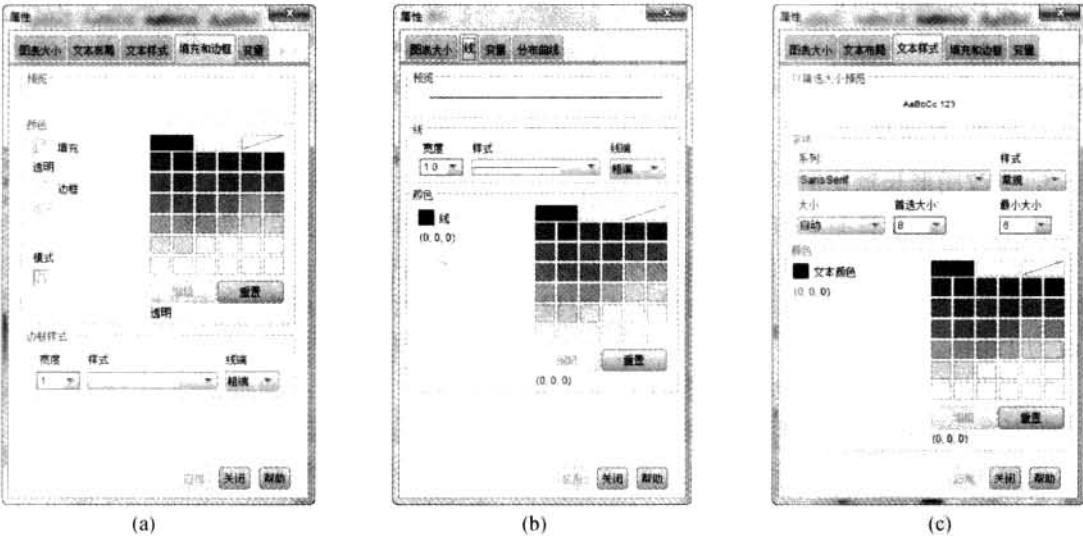


图 10.11 图形区域、线条、文本的选项卡

6. 更改连续轴选项

下面介绍一下对连续轴可以进行哪些修改。由于直方图中的两个数轴都是连续轴,因此其可用的选项卡也是完全相同的。这里以纵轴为例,当选中纵轴的任意部分时(刻度、轴线、标题文字均可),“属性”对话框中都会出现连续轴适用的选项卡,如图 10.12 所示。

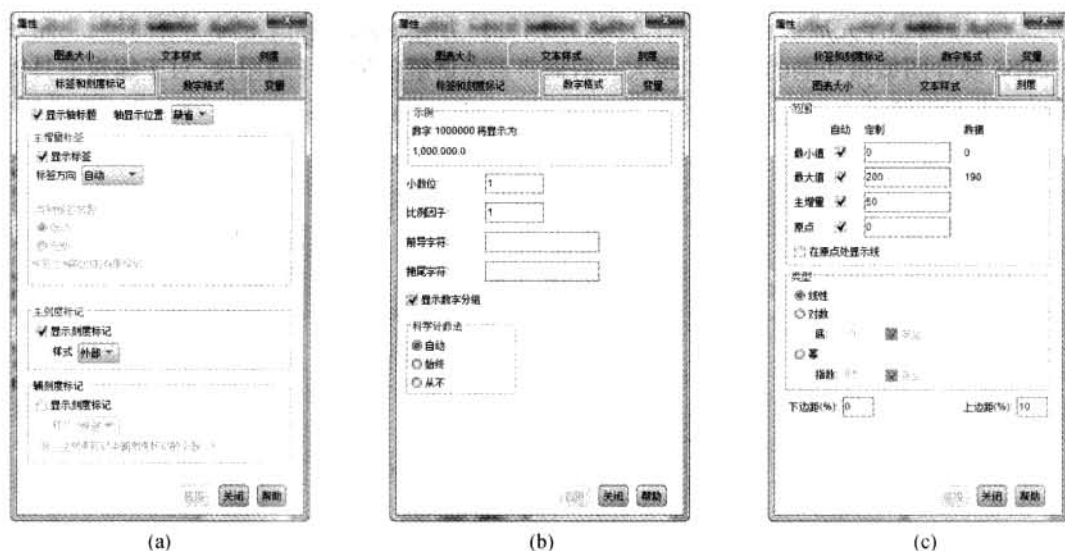


图 10.12 连续轴的“标签和刻度标记”、“数字格式”和“刻度”选项卡

(1) “标签和刻度标记”选项卡:可在其中控制轴标题、主刻度、次要刻度的显示,并控制标签显示方向。

(2) “数字格式”选项卡:主要用于设定数值显示格式,包括小数位、比例因子、前导字符和拖尾字符,本例可以将小数位数均设为 0 以简化输出;在数值较大时,数轴刻度将会按照原始数值除以比例因子加以显示;前导和后置字符则主要用于为数字显示加入说明文字以方便阅读。

(3) “刻度”选项卡:用于设定数轴的起、止数值,间距大小和原点所在位置,本例可以缩小显得过大的上方和左侧边距。选项卡下方则用于更改连续轴的刻度方式,默认为算术等距,也可更改为对数等距或指数等距尺度。对于需要绘制对数线图的朋友,现在就可以知道在 SPSS 中应当如何操作了。

除以上选项卡外,连续轴的选项卡还有“文本样式”等,因为比较简单,这里不再详述。此外,数轴的标签也是可以修改的,只需要在标签上连续单击两次,即可进入编辑状态,在本例中可以将纵轴标签改为“频数”,横轴标签改为“总消费者信心指数”。

7. 增删图形元素

显然,信心指数低于 50 的受访者是比较悲观的,这时可以在图形上添加一条参考线以突出哪些组段达到了这一标准,只需要在图形窗口中右击,然后在弹出的快捷菜单中选择“添加 X 轴参考线”菜单项即可。只是操作完毕后该参考线默认被放置在连续轴正中间,即 80 的位置,因此需要再次选中该参考线,然后在“参考线”选项卡中将其位置由 80 改为 50。

如果要删除某些图形元素,则首先选中该元素,然后右击,如果该元素可被删除,则在菜单上会出现“删除”菜单项。

因编辑需求不同,各种图形元素的右键弹出菜单也各不相同。但其功能和图形需求是完全对应的,因此这里不再详述,感兴趣的读者可以自行尝试。

在以上编辑操作完成后,最终的直方图如图 10.13 所示,通过这个实例,读者应当能充分体会到 SPSS 统计图编辑功能的便捷和强大。后面将重点讲解各种图形的用途,对于编辑的具体操作则不再详细讲解。

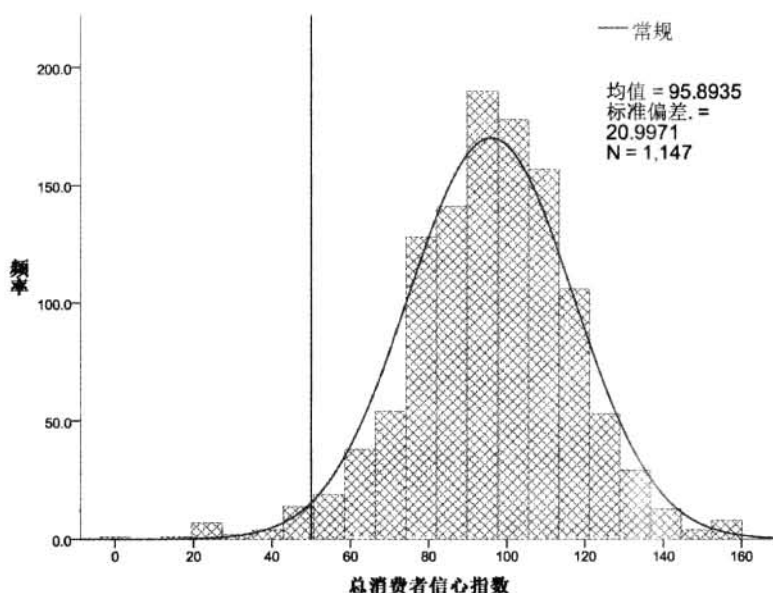


图 10.13 编辑完毕的直方图

10.2.3 直方图图形框架的修改

本节将进一步介绍涉及图形框架变化的一些操作,有的可以直接通过图形编辑来完成,有的则需要重新绘图。

1. 直方图组的绘制

如果希望比较北京、上海、广州三地受访者的信心分布有无差异,除了可以分别绘制 3 张直方图外,还可以用图组的方式来实现。除了和上面完全相同的对话框操作外,绘图时新增的操作如下。

(1) 切换至“组/点 ID”选项卡:选中“行嵌板变量”复选框。

(2) 将 SO 城市选入画布上新增的嵌板框中。

(3) 单击“选项”按钮,在打开的“选项”对话框中确认未选中下方的“换行嵌板”复选框。

最终绘制的图组如图 10.14(a)所示(注意该图形已经进行过编辑),从中可以看出三地的信心指数基本上都服从正态分布,但均数的差异则相对并不明显。

2. 累积直方图的绘制

累积直方图主要用于描述连续型变量的累积分布,其基本绘制原理和普通直方图是一样的,只是要从小到大将各直条的频数累积起来。和普通直方图的操作相比,其新增的操作如下。

“元素属性”对话框:在“统计量”下拉列表框中将默认的“直方图”选项修改为“累计计数”选项。

绘制完成的相应的累积直方图如图 10.14(b)所示。

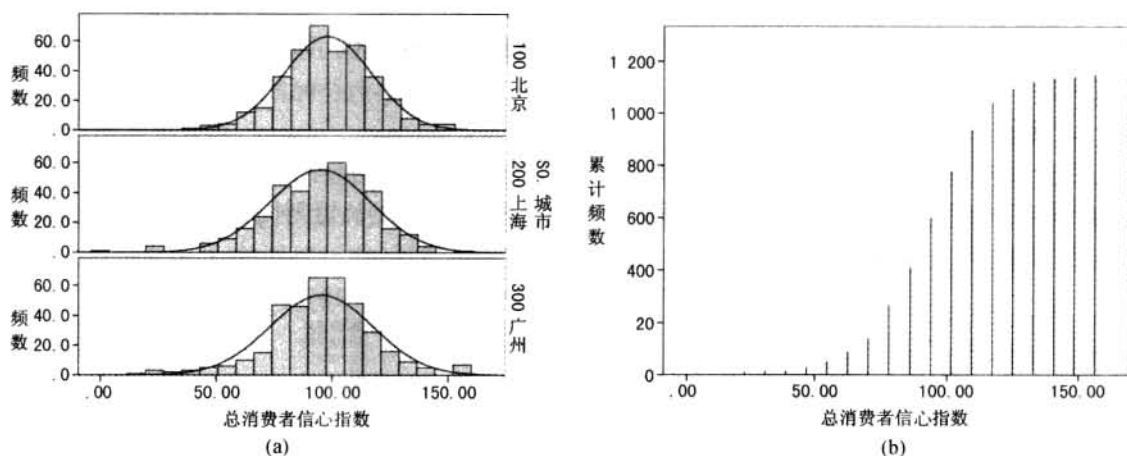


图 10.14 信心指数的直方图图组与累积直方图

3. 直方图选项的修改

直方图默认绘制的是正态分布曲线,如果怀疑数据实际上服从其他种类的分布,则可以在选中分布曲线后,在新出现的“分布曲线”选项卡(如图 10.15(a)所示)中选择相应的分布及参数,确认后图形就会出现相应的变化。

除更改期望分布外,对直方图进行编辑的另一个主要方面是对直方图中的直条数数目进行修改。为此首先应当在图形编辑窗口中选中直条,并在“属性”对话框的“分箱”选项卡(如图 10.15(b)所示)中对直条(组段)的起始位置(Anchor First Bin)和直条数(Bin Sizes)等进行设定。

最后,在“变量”选项卡(如图 10.15(c)所示)中,可以对图形框架进行最为详细的修改设定,包括各数轴上出现的元素,统计量等。但由于直方图在此处可做的选择不多,因此这里只做功能提示,详细的功能讲解将在后续图形的编辑功能中展开。

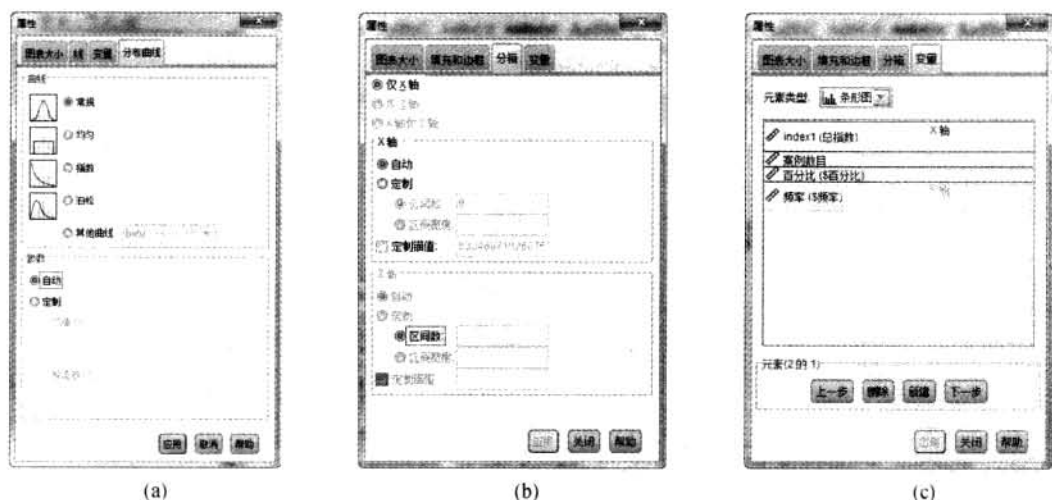


图 10.15 “分布曲线”、“分箱”和“变量”选项卡

10.2.4 直方图的衍生图形

在图形生成器图表库的直方图组中,除了基本的直方图,还提供了以下几种衍生图形,下面进行简单介绍。

1. 分段直方图

在普通直方图的基础上会增加一个分段变量,图形中的直条将会按分段变量的不同取值被分为若干段,直条全长仍然代表某个变量的组段总频数,各分段的长短则代表该组段各组成部分的频数,如图 10.16(a) 所示。

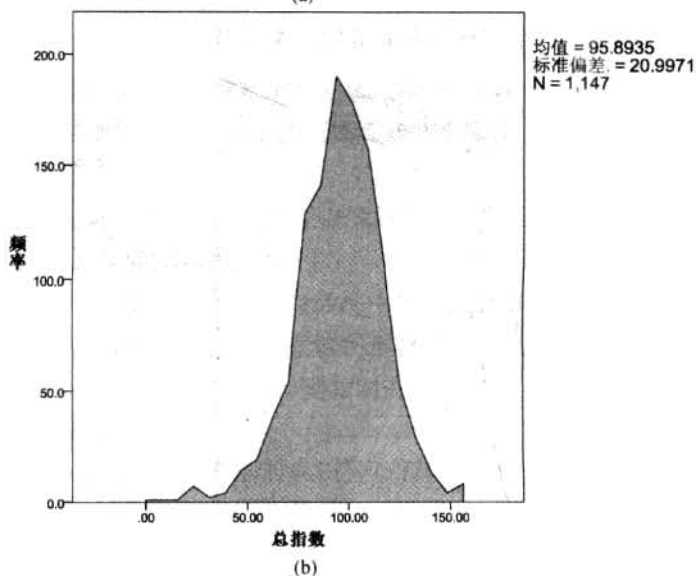
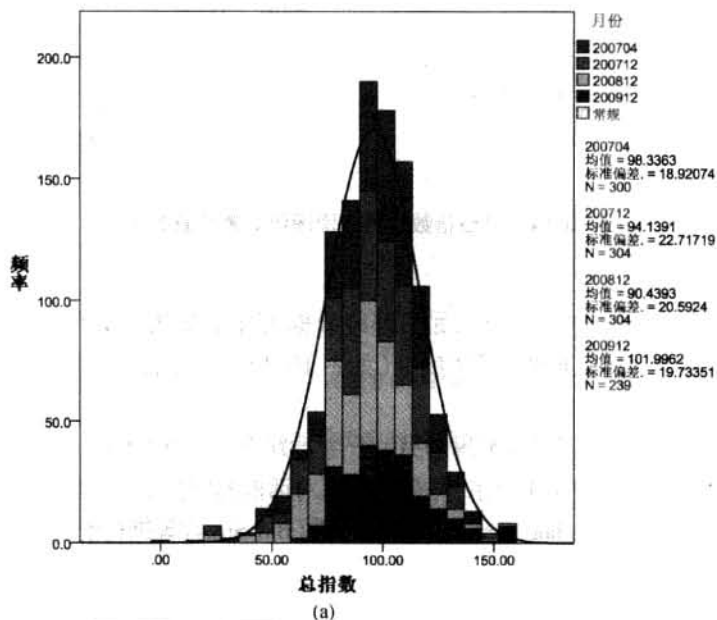


图 10.16 分段直方图与面积直方图示例

绘制分段条图时并无特殊之处,只是画布上多了一个分段变量框,将相应的分段变量拖入该框即可。


因分段直方图(Stacked Histogram)相对较少使用,因此这里不展开讲解,关于分段变量的详细介绍参见 10.5.3 小节的分段条图部分。

2. 面积直方图

也称为频数多边形(Frequency Polygon),也属于很少使用的一类图形,其实就是将原直方图各直条的顶点连接起来形成的特殊用途的面积图,如图 10.16(b)所示。绘图时的操作几乎和普通直方图完全相同,只是可以在“元素属性”对话框中的“插值”下拉列表框中选择不同的顶点连接方式,默认为“直连左连接”,如果希望曲线圆滑一些,可以改为“样条光滑”。

3. 人口金字塔

人口金字塔(Population Pyramid)是直方图的另一种变体,最常用于汇总人口数据,一般是按性别分割的,提供两个紧挨着的有关年龄数据的水平直方图且左右对称,该图形在人口统计学中有重要的分析价值。因其在人口为年轻型的国家/地区中,所产生的图形呈现金字塔形状,故此得名。

 该图形在 SPSS 对话框中被称为总体锥形图,实际上,该图形也可以用来绘制营销分析中非常常用的漏斗图(也称为销售漏斗),因操作和人口金字塔的绘制非常类似,这里不再单独讲述。

在 SPSS 中绘制人口金字塔需要指定两个变量:首先是用于拆分金字塔的分类变量,将其拖入画布中的拆分变量放置框中即可。其次为用于绘制水平直方图的变量,将其拖入“分布变量”区放置框中即可。

需要指出的是,虽然用于绘制水平直方图的分布变量从理论上讲应当是连续变量,但人口金字塔在绘制时往往会使用汇总后的频数数据。SPSS 对两种数据都可以进行正确的绘制,但如果是后者,则一定要先正确设定相应的频数变量,然后再创建该图表。

图 10.17 为基于 CCSS 样本数据绘制出的人口金字塔示意图,读者也可以自行寻找全国人

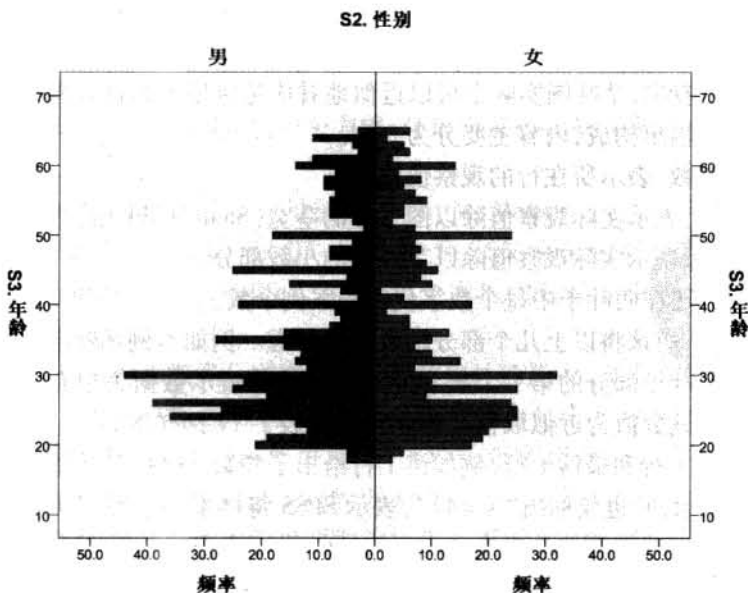


图 10.17 基于 CCSS 数据绘制的人口金字塔示意图



和直方图相比,茎叶图在反映数据整体分布趋势的同时还能够精确地反映出具体的数值大小,因此样本少时优势非常明显,该图形在国外非常流行。

10.3 箱 图

箱图(Box Plot)也称为箱线图,和直方图一样都用于描述连续变量的分布情况,但两者的功能并不重叠,直方图侧重于对一个连续变量的分布情况进行详细考察,而箱图更注重基于百分位数指标勾勒出统计上的主要信息。由于使用箱图便于对多个连续变量同时进行考察,或者对一个变量分组进行考察,因此在使用上要比直方图更为灵活,用途也更加广泛。

10.3.1 案例:用箱图分月份考察消费者信心的分布

例 10.2 利用箱图分月份对样本的消费者信心指数进行描述,以考察指数随时间的基本变化趋势。

由于箱图可以将不同月份的多个箱体绘制在同一张图中,因此比较起来非常方便。本例的操作步骤如下。

(1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(2) 在图库中选择“箱图”组,右侧出现的图标组中的第 1 个即是简单箱图,将该图标拖入画布中。

(3) 在变量列表中找到 index1,将其拖入画布的纵轴框中。

(4) 将月份 time 拖入横轴框中。

(5) 单击“确定”按钮。

通过上面的操作,就由 index1 的取值范围控制了连续轴的尺度范围,并最终生成图中的箱体,而 time 的不同取值则用于形成分类轴中的类别。

绘制的箱图如图 10.19 所示,显然整个样本按访问月份的不同被分成了 4 组,从而在图中一共绘制了 4 个箱形。

(1) 每个箱形都由最中间的粗线、一个方框、外延出来的两条细线和最外端可能有的单独散点组成。

(2) 箱体中间的粗线表示当前变量的中位数(M, Median, 注意不是算术均数),方框的两端分别表示上、下四分位数(Q1 和 Q3, 即 25% 和 75% 百分位数),两者之间的距离为四分位数间距(Interquartile Range, IQR)。显然,整个方框内包括了中间 50% 样本的数值分布范围。

(3) 方框外的上、下两个细线分别表示除去异常值外的最大、最小值。

(4) 在箱图中,凡是与四分位数值(图 10.19 中即为方框上下界)的距离超过 1.5 倍四分位间距的都会被定义为异常值,其中离方框上/下界的距离超过四分位数间距 1.5 倍的为离群值,在图 10.19 中以“O”表示;超过 3 倍的则为极值,用“*”表示。散点旁边默认标出相应案例号备查。从图 10.19 中可见 397 号案例被标识为极端值,而 258 号等多个案例均被标识为离群值。

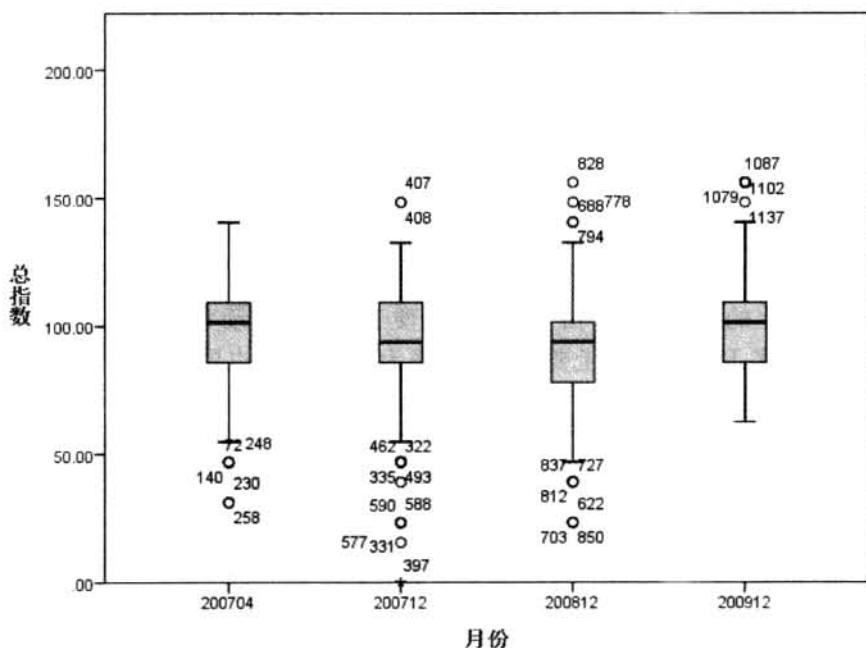


图 10.19 箱图的分析结果

在箱图基本结构介绍完毕后,现在从整体上观察图 10.19,可以得到如下信息。

(1) 从中位数角度看,2007、2008 年末的消费者信心较低,而 2007 年初和 2009 年末的信心都较高。

(2) 从箱体宽度角度看,4 个月份的信心指数离散程度相差不明显,未发现有(相应总体)方差不齐的迹象。

(3) 从离群值和极端值分布情况角度看,可以发现样本中有一些离散程度较大的数值,但情况并不十分严重,进一步分析时特别关注一下其影响即可。

显然,和直方图相比,箱图更为简明清晰地突出了数据分布的主要趋势,且用于组间比较时有优势,因此箱图往往被作为数据预分析时的有力工具加以使用。



需要指出的是,由于箱图主要用于对以百分位数为基础的信息进行呈现,因此当百分位数不稳定时,箱图并不适用。由此可知,当样本量太少,或者相同数值过多时,不宜使用箱图进行呈现,此时茎叶图或者条图是更好的选择。

10.3.2 箱图的编辑

由于箱图是由方框、线段和散点构成的,因此前面对区块、线条等进行的编辑操作,如对填充样式、颜色、线型等进行的修饰操作在箱图中也完全相同,唯一新增的是当在图中选中异常值散点后就可以使用标记选项卡更改散点的样式、颜色等,这里不再对一般的图形编辑操作进行讲解,而是重点对一些新出现的编辑功能和箱图中的特色功能加以介绍。

1. 分类轴选项的修改

分类变量所包含的信息量是低于连续变量的,与此相对应,分类轴中可供修改的选项也明显少于连续轴。最主要的是“类别”选项卡,如图 10.20(a)所示,用于设定各类别在数轴中的排列顺序,以及该类别是否在图中显示。在选中变量名称后,其右侧的▲和▼两个按钮就用于更改变量在分类轴上的排列次序,而✕和➡两个按钮则用于将变量移出或重新移入显示列表中。除此以外,选中选项卡最上方的汇总复选框可以将各小类加以合并,默认将构成比小于5%的各类合并成一个“其他”类(注意合并后的总构成比是可以大于5%的)。

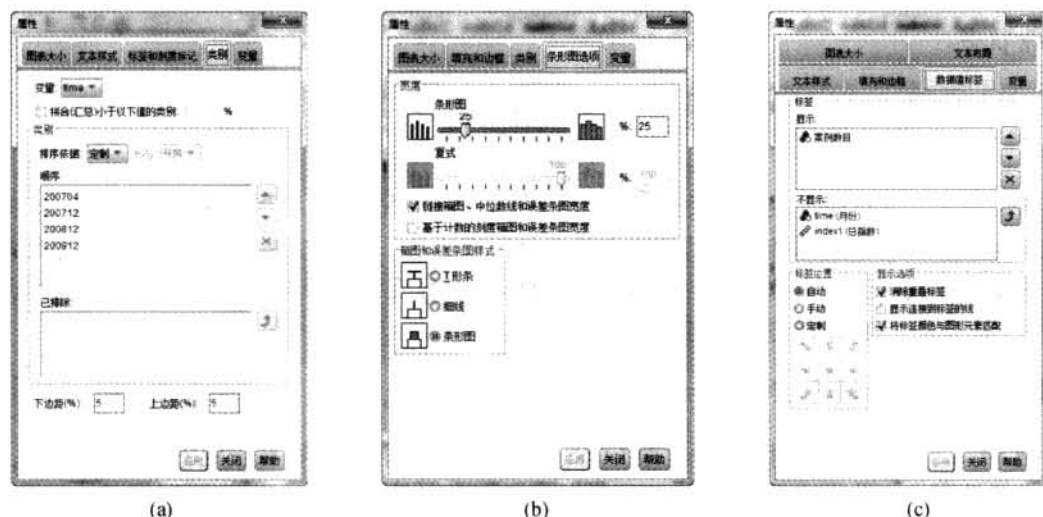


图 10.20 “类别”、“条形图选项”和“数据值标签”选项卡

在分类轴编辑中另一个可能用到的选项卡是“标签和刻度标记”,它用于控制主、次刻度的显示方式。该选项卡在前面连续轴编辑时已经出现过,这里就不再重复讲解了。

2. 箱图主体格式的编辑



箱图主体也可以进行一定的修改,当选中箱形时,就会打开“条形图选项”选项卡,如图 10.20(b)所示,其上部用于设定中间方框的宽度,通过选中“基于计数的刻度箱图和误差条图宽度”复选框可以按照各组样本量多少来设定宽度。当选中外侧的细线时,则可用下方的选项设定细线的显示格式。可更改为无两端的细线,或者以细直条方式加以显示。

3. 设定异常值散点的标签

在默认情况下异常值旁边会显示相应的案例号作为标签,对此也可以进行更改,选中图形中的标签,就可以在“数据值标签”选项卡(如图 10.20(c)所示)中更改用做标签的变量名称和显示位置等。注意如果在绘图操作中不指定标签变量,则此处只有“Case Label”这一个标签变量可供选择,即要么在散点旁显示记录号,要么什么也不显示。

在编辑状态下,“数据值标签”选项卡中只会出现在绘图操作中被引入图形的变量,如果希望新增其他变量到候选列表中,则需要重新绘图。具体操作是在画布区右击,在弹出的快捷菜单选择“点 ID 区域”→“添加”菜单项,或者在下方的“组/点 ID”选项卡中选中下方

“指定 ID 标签”复选框,都会在画布区域中增加一个点标签变量框,将相应的标签变量拖入该框中即可。

此外,如果图中的散点太多,默认将其标签号都显示出来就会将统计图变成一张抹布。这时通过 SPSS 提供的功能可以只显示某些标签,选择“元素”→“数据标签模式”菜单项,或者直接在工具栏上单击按钮,则系统进入数据 ID 模式,光标也会变成,此时只需要在相应的散点上单击,它所对应的标签就会在显示/隐藏间进行切换。而如果因散点过于重叠而同时选中了多个散点,则系统会首先弹出选择对话框,要求指明对哪些散点进行操作,这时只需要选出希望更改的散点即可。当更改完毕后,只需要再次选择“元素”→“数据标签模式”菜单项,系统就会切换回正常状态。

10.4 饼图

饼图(Pie Chart)用于表示各类别某种特征的构成比情况,它以圆形的总面积为 100%,扇形面积的大小表示事物内部各组成部分所占的百分构成比。一般以时钟 12 点处为起点,各组成部分按习惯顺序或数值大小依次顺时针排列,“其他”类别放在最后。当同时绘制多个圆图并进行比较时,图例应一致,以便进行比较。

10.4.1 案例:分城市、月份考察样本性别比例

例 10.3 现希望分城市、月份考察 CCSS 数据的性别比例是否存在一定的变化趋势。

由于性别为两分类变量,因此本例既可以用饼图来表现性别构成比,也可以简单地用男性或者女性一方的比例采用条图直接呈现。这里还是采用饼图来实现。由于需要分城市、月份进行考察,因此可以考虑将这两个变量分别设定为行面板和列面板变量,且从使用习惯上讲,月份这一有序分类变量设置为列面板变量应当更为妥当。

本例的具体操作如下:

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“饼图”组,将右侧出现的饼图图标拖入画布中。
- (3) 切换至“组/点 ID”选项卡,选中“行嵌板变量”和“列嵌板变量”复选框。
- (4) 将性别 S2 拖入“分区依据”列表框中。
- (5) 将月份 time 拖入列嵌板变量框中,城市 S0 拖入行嵌板变量框中。
- (6) 将统计量下拉列表由合计改为计数,单击“应用”按钮。
- (7) 单击“确定”按钮。

最终完成的饼图如图 10.21 所示,可见:

- (1) 在绝大多数的月份*城市组合中,男性受访者的比例都要高于女性。
- (2) 北京、上海、广州三地相比,广州的男性受访者比例明显更高。
- (3) 随着时间的推移,男性受访者在三地似乎都有一定的上升趋势,广州的这一趋势似乎更为明显。

通过观察上述图形可以发现,性别比例在不同时间、城市间的样本是存在波动的,且随时间推移男性比例似乎存在上升的趋势。显然性别比例的变化可能会影响到最终计算出的信心指数

值。好在 CCSS 项目在实际计算各项指标之前都会对样本进行人口特征资料的加权调整,因此并不会受到影响。

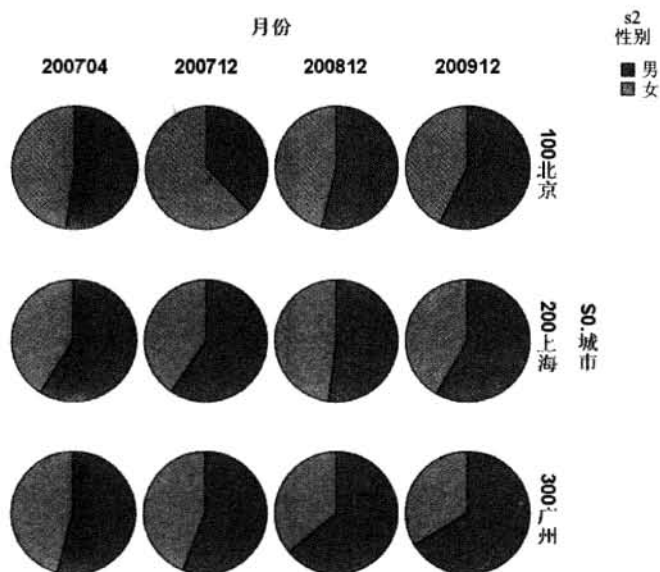


图 10.21 绘制的饼图

10.4.2 饼图的编辑

因其对数据特征的表现形式比较特殊,饼图是没有横、纵轴的,所以也不存在对数轴的设定问题。但饼图中有一些前面各节中未出现过的编辑功能,如图 10.22 所示,下面就将对此进行介绍。


1. 行、列嵌板的编辑

嵌板格式的编辑不单属于饼图,但这里一并讲解。当选中国形区域主体时,就会打开“嵌板”选项卡如图 10.22(b)所示。可见在其中可以设置图形水平或者垂直翻转显示,或者允许嵌板变量换行显示。

2. 饼图主体的编辑

选中饼图主体后“属性”对话框会切换到“深度和角度”选项卡,可用于饼块的格式设置,如阴影效果、三维效果等。在选项卡中部还可以定义第一个饼块起始于时钟的哪个方向,以及整个饼图是沿顺时针还是逆时针方向排列。这些功能都非常简明,这里不再详述。

3. 设定饼块标签

绘制的饼图默认不显示数据标签,如希望显示,则首先选中希望显示标签的饼块或饼块组,然后“元素”→“显示数据标签”菜单项,或者直接单击快捷工具栏上的按钮 ,则相应的饼块就会出现数值标签,用于给出相应的统计指标,如比例或频数等。

如果希望改变标签显示内容,则单击选中标签,可在相应的“数据值标签”选项卡中设定标签的位置、内容等,也可以使用数据标签模式选择个别标签加以显示,这些操作和箱图中的操作

完全相同,这里不再重复。

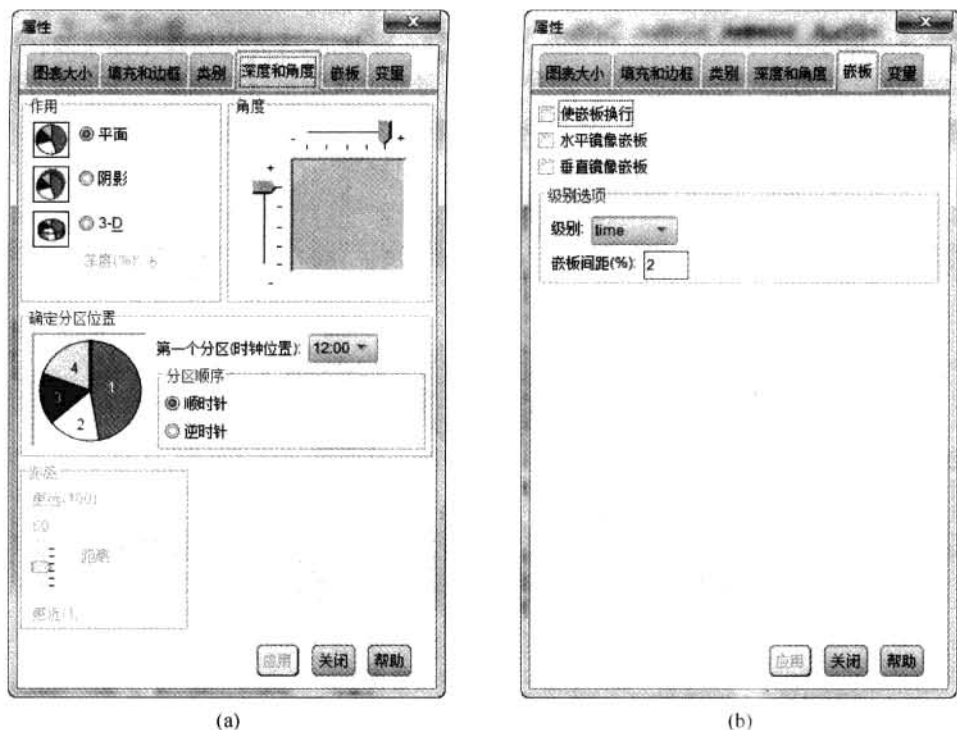



图 10.22 饼图的编辑选项卡

4. 饼块的突出显示与合并

有时候为了特别关注饼图的某一部分,希望突出显示该部分,则首先用鼠标选中想突出显示的那一部分(注意要连续单击两次才能做到),然后选择“元素”→“分解分区”(Explode Slices)菜单项,或直接单击工具栏上的按钮,则相应的饼块就会被突出显示,再次选择后饼块就会回复原位。

在实际应用中,往往不需要将所有部分都单独显示,对于那些所占比例很小(比如小于5%)的部分,常常不再逐一图示,而是合并为“其他”一类,这样图形显得更简洁清晰。这一功能实际上在箱图中已经遇到过了,就是 Categories 选项卡最上方的 Collapse 复选框,只是这里针对的是饼图而不是分类轴。除了合并显示以外,该选项卡也可以用于调整各饼块的排列顺序、隐藏某些类等,操作和前面相同。

10.5 条图与误差图

条图(Bar Chart)用等宽直条的长短来表示相互独立的各指标数值大小,该指标可以是连续性变量的某汇总指标,也可以是分类变量的频数或构成比。各(组)直条间的间距应相等,其宽度一般与直条的宽度相等或为直条宽度的一半。为了便于比较,一般将被比较的指标按大小顺序排序或者按某种自然顺序排列。

绘制条图时纵轴尺度应当尽量从0开始,中间不宜折断,否则将给人以错误的印象。如图10.23中甲组某观察指标值为8,是乙组的两倍。若纵轴从“2”开始,则给人以甲组该观察指标值是乙组的3倍的错觉,需进一步对照坐标轴尺度才能得出正确结论。而这恰恰是现在许多IT硬件评测文章都喜欢的做法,把原本性能差别甚微的两种芯片表现成相差很大。虽然从吸引眼球和图形美观的角度来讲这样做无可厚非,但在明白了这一点后,大家就可以更为冷静、客观地阅读图形,避免一时冲动之下做出错误的购买决策了。

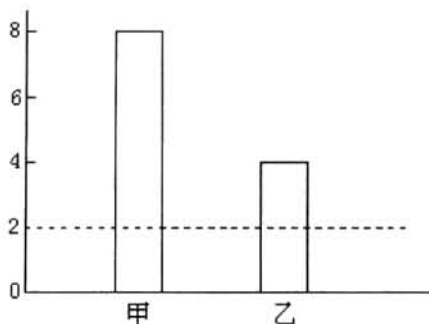


图 10.23 条图的纵轴尺度起点必须为零的示意图

虽然条图的结构非常简单,但由于它可以灵活反映各种各样的信息组合,因此在操作上反而比较复杂。本节将由浅入深地对各种条图加以介绍。

10.5.1 简单条图案例:比较不同职业人群的消费者信心值

例 10.4 用条图展示不同职业人群的消费者信心均数有无差异。

根据题目要求,直条类别需要用职业 S5 来定义,但直条的长度则需要用 index1 的均数来定义。同时为了使图形的展示更为清晰,可以在绘图完成后再对直条进行排序,本例的具体操作如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“条”图组,将右侧出现的简单条图图标拖入画布中。
- (3) 将职业 S5 拖入横轴框中。
- (4) 将 index1 拖入纵轴框中。

(5) 单击“确定”按钮绘制出图形,然后双击图形进入编辑状态,选中“类别”分类轴,在“属性”对话框的“分类”选项卡中,在“排序依据”下拉列表框中选择“统计”选项,在“方向”下拉列表框中选择“降序”选项,单击“应用”按钮。



可能有反应快的读者会想到:为什么不在绘图的时候就在“元素属性”对话框中,将排序依据改为“统计”,而一定要绘图完毕后再行编辑呢?可以试试看,结果就是在绘图界面中的排序依据中根本找不到“统计”这个选项!

对此不用感到莫名其妙,SPSS 在这个问题上的逻辑是这样的:在图形绘制完成之前,相关的统计量均未进行计算,因此不可能用其排序。当绘图完毕之后,相关的统计量已经被存储存储在图形中,因此可以在编辑状态下进行调用。不仅在条图中,大家在随后学习的

带误差线的条图等多种图形中都会遇到这种情况。

最终所绘图形如图 10.24 所示,可见由于对未来充满希望,平均而言学生的信心值最高,紧随其后的是经济地位相对不错的私营业主,而衣食无忧的公务员信心值排在第三,其排名甚至还高于医生、律师,以及企业管理者。而蓝领工作者、退休人员以及失业人员的信心值则分列最后 3 位,显然,上述统计结果是非常符合逻辑的。

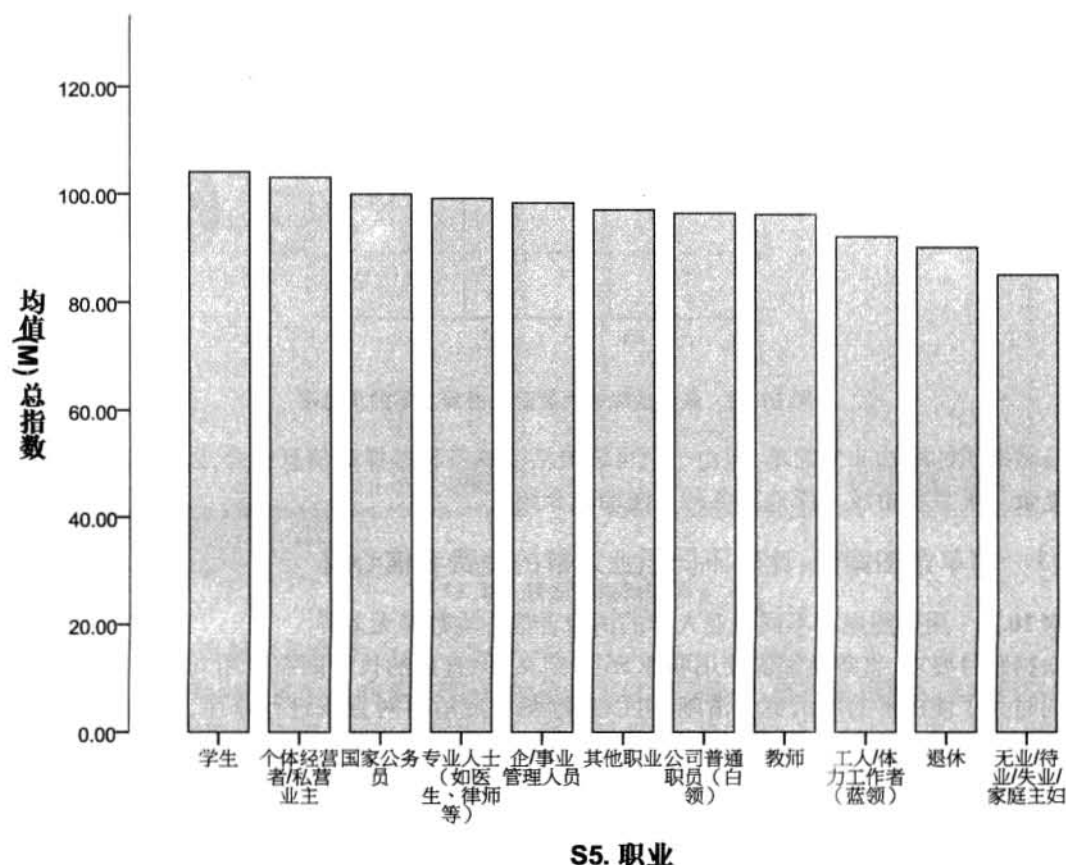


图 10.24 简单条图示例

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

例 10.5 在 10.5.1 节的分析基础上,进一步用条图展示不同职业人群的现状指数和预期指数均数的差异情况。

根据题目要求,对于每一个职业类别,需要同时展示总信心、现状指数、预期指数 3 个均数。显然一个直条就不够用了,需要用到复式条图。复式条图(Clustered Bar Chart)是指由两个或两个以上小直条组成条组的条图,各条组之间有间隙,组内小条之间无间隙。在本例中,直条类别仍然需要用职业 S5 来定义,但直条组则需要同时使用 3 个变量的均数来定义,相应的操作是第一次遇到,具体如下。

(1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(2) 在图库中选择“条”图组,将右侧出现的复式条图图标拖入画布中。

(3) 将职业 S5 拖入横轴框中。

(4) 按住 Shift 键,在左侧“变量”列表框中同时选中 index1、index1a 和 index1b 三个变量。将其拖入纵轴框中。此时 SPSS 会弹出创建摘要组确认框,单击“确定”按钮。

(5) 单击“确定”按钮绘制出图形,然后双击图形进入编辑状态,选中“类别”分类轴,在“属性”对话框的“类别”选项卡中,用手工方式将顺序框中的类别排序方式更改为和 10.5.1 小节简单条图中相同的顺序,然后单击“应用”按钮。

(6) 将均值连续轴刻度范围修改为 0~110,小数位数更改为 0。拖放调整图例位置和绘图区大小至合适的比例。



本例中由于构建了多个汇总变量的摘要组,因此不能按照上例的方式用“分类”选项卡进行排序。

如果未像本例中一样构建摘要组,则进行绘图操作时需要再指定一个分类变量作为条图组的分组因素。

图 10.25 即为最终绘制出的复式条图,从中可以看到如下信息。

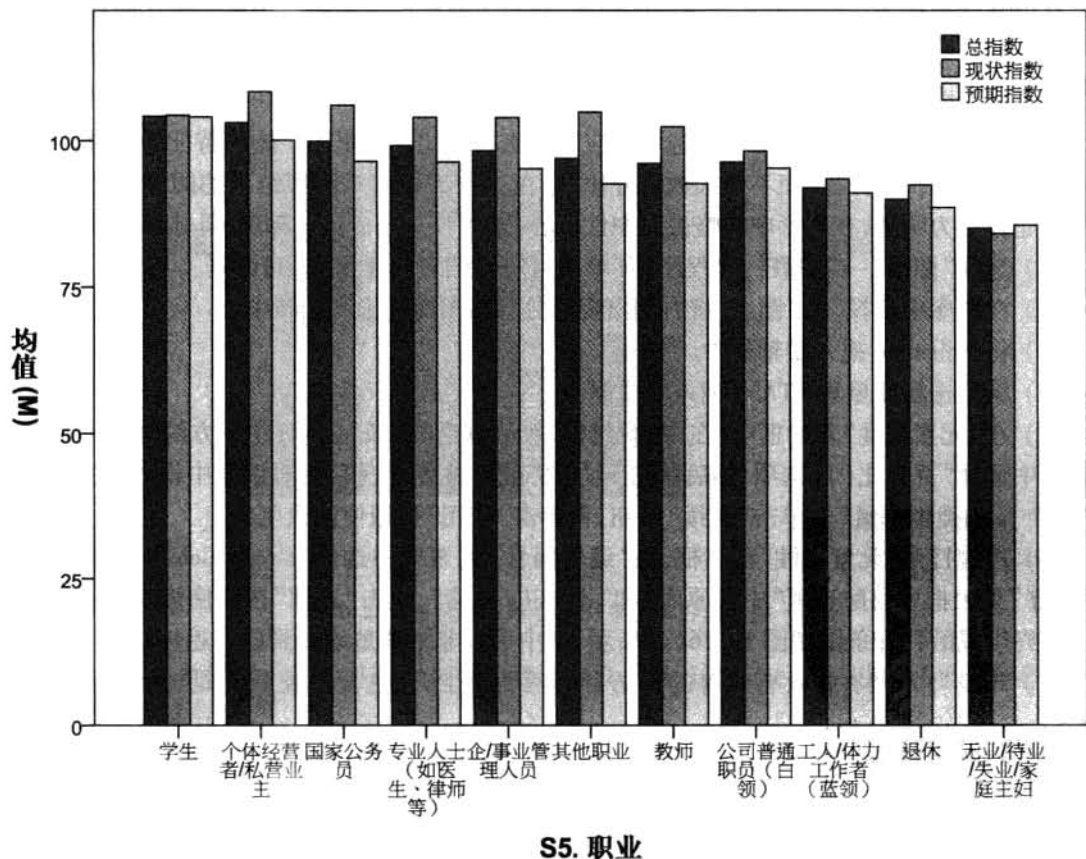


图 10.25 复式条图示例

(1) 收入较高的私营业主、专业人士、企事业管理人员,以及收入虽然偏低但很稳定的国家公务员、教师的现状指数明显较高。

(2) 除失业人群外,其余所有职业人群的预期指数均低于现状指数。

(3) 教师群体的预期指数明显低于其现状指数,也明显低于其他大多数职业群体的预期指数,这说明教育体制是需要改革的。

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

1. 分段条图与百分条图的定义

分段条图(Stacked Bar Chart)和复式条图一样,也需要多考察一个分组因素,但在分段条图中以直条全长代表某个变量的总量,而用其中的分段长度表示不同亚群对总量的贡献(构成比或数量大小)。

与分段条图密切相关的是百分条图(Percent Bar Chart),也称为马赛克图,是用直条内部各部分面积的大小表示事物内部各组成部分所占的百分构成比的。在目前的 SPSS 版本中,百分条图既可以直接绘制,也可以在图形编辑时由分段条图加以转换。

2. 百分条图的绘制

例 10.6 多选题 A3a 记录的是受访者做出当前家庭经济状况判断的依据,现希望考察不同月份各选项回答比例的变化有无某种内在趋势,以提供对信心指数变化趋势的辅助解读信息。

根据题目要求,需要对 A3a 各选项的提及率进行逐月的对比,显然如果用各选项的应答人次百分比来绘图,整个直条的长度应当正好都是 100%,而不同的选项将会将直条切分成若干段,定义每段长度的指标就应当是各选项的应答人次百分比。这里由于图形特征的限制,只能在图中显示应答人次百分比,而不是人数百分比,好在这并不影响对图形结果的阅读。此外,为了使变化趋势更为明显,在图形中将略去对中性回答、拒答等选项的比例输出,具体操作如下。

(1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(2) 在图库中选择“条”图组,将右侧出现的分段条图图标拖入画布中。

(3) 将月份 time 拖入横轴框中。

(4) 将多选题变量集 \$TA3a 拖入堆栈框中。

(5) 在“元素属性”对话框中,在“编辑属性”列表框选择“条”选项,在下方的“统计量”下拉列表框中选择“百分比(%)”选项,然后单击下方的“设置参数”按钮,在对话框中将百分比分母设为“每个 X 轴类别总量”。单击“继续”按钮,再单击“应用”按钮使修改生效。

(6) 再次打开“元素属性”对话框,在“编辑属性”列表框中选择“Group Color”选项,在下方的“顺序”列表框中移除选项“中性原因”和“不知道/拒答”,单击“应用”按钮使修改生效。

最终生成的百分条图如图 10.26(a)所示,从中可以非常清楚地看到以下趋势。

(1) 2009 年末回答收入有改善的受访者比例明显上升,这显然反映的是经济刺激计划的效果。

(2) 2008 年末,次贷危机全面恶化的时间段,回答收入变差的比例大幅增加。

(3) 回答投资收益有改善的受访者比例一直在持续减少,正好对应了 2007 年以来一溃千里的股市。

(4) 回答家庭开支恶化的比例在 2007 年末达到高峰,随后逐渐降低,这正好对应了当时一

路飞涨的物价。

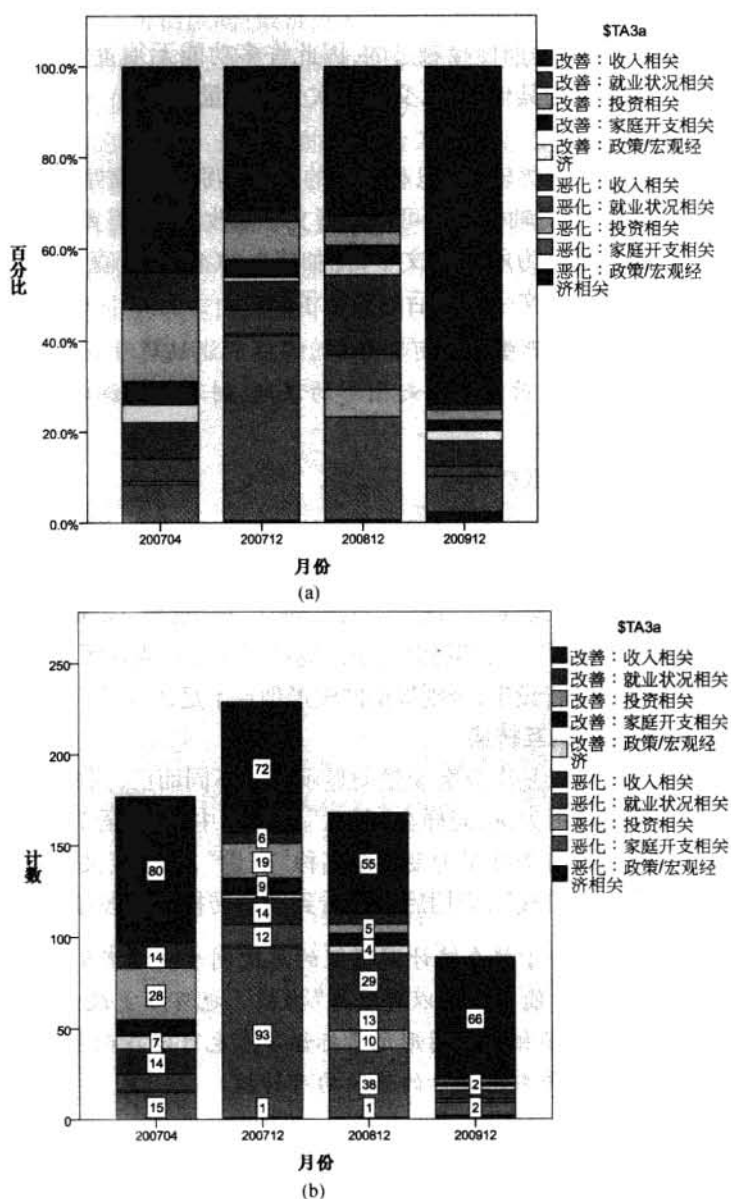


图 10.26 多选题 A3a 的百分条图与分段条图

显然,上述图形提供了非常丰富而直观的数据解读信息,这对随后进一步分析和理解信心指数的变化趋势非常有帮助。

3. 分段条图的绘制

如果在上例中希望考察的是各选项的回答频数,则只要将直条的统计量仍然保留为默认的“计数”即可,所绘制出的分段条图如图 10.26(b)所示,此时各直条的总长度不再相等,而是反映了当月相应所有选项的应答总频次。注意图形被编辑了,标出了数值标签,标签上给出的就是应答频次数。

10.5.4 条图的编辑

条图中的绘图元素基本上都是前面接触过的,因此许多功能无须重复讲述。这里只讨论一些条图的特色编辑功能,如转换为其他图形,交换主次分类变量等。

1. 分类轴标签的编辑

在前面绘制的简单条图中,分类轴为“职业”,有的职业类别标签文字很长,可能会影响图形效果。实际上,SPSS统计图的分轴标签是可以进行文字修改的,只需要通过几次单击,就可进入相应标签的编辑状态,将其修改为所需的文字,比如将“...(白领)”直接简化为“白领”,编辑完毕后直接按回车键,就可以看到文字修改后的效果了。



初学者对于究竟图形中的哪些文字可以编辑,哪些不能编辑往往感到无所适从,实际上其基本规律是:如果文字修改可能引起图形的误读,则不允许编辑;否则,就可以进行文字内容的编辑修改。

2. 条图与其他统计图形的相互转换

由于饼图、条图、线图、面积图的基本结构都可以用于反映一个分类变量的数据分布情况,区别仅在于用于表达统计量的图形元素不同,因此这些图形是可以相互转换的。以上面绘制的简单条图为例,只需要选中图形元素主体,随后在“属性”对话框中打开“元素”选项卡,可见“元素类型”下拉列表框中目前的选项为“条形图”。将其修改为所需的饼图、线图,或者面积图种类即可。注意有些转换需要同时对数轴变量等进行设定,否则提供的图形信息不足,无法进行转换。

3. 复式条图和分段条图的相互转换

复式条图和分段条图相比,只是次分类变量的显示方式不同而已,因此完全可以实现相互转换。以例10.5中绘制的复式条图为例,同样在“元素”选项卡中,可以看到该图的主分类变量为“职业 S5”,用于定义 X 轴;而次分类变量为变量组名称“变量”,用于定义 X 轴聚类,且采用颜色标识。这里只要将其功能改为“堆栈”,应用后就会看到图形转换为分段条图了。




实际上“元素”选项卡相当于整个统计图框架的总控制台。读者们可能已经发现,在“元素”选项卡中,上述变量功能不仅可以更改为“堆栈”,也可以更改为“Y 轴”或“Z 轴”,甚至于行嵌板、列嵌板等。而相应的图形元素标识方式也可以在颜色、模式、大小等之间自由转换,只要是黑色而不是灰色显示的选项均可转换。

4. 复式条图/分段条图中主、次分类变量的相互转换

有了上面的基础,这里的操作不言自明,只需要在“元素”选项卡中将主、次分类变量的功能对换一下即可。

5. 分段条图和百分条图的相互转换

分段条图和百分条图在编辑状态下也可以相互转换,但这两种图形的框架完全相同,因此无法在“元素”选项卡中加以操作,而是选择“选项”→“缩放至 100%”菜单项,或者直接单击快捷工具栏上的按钮,图形就会在分段条图和百分条图之间相互切换显示了。

10.5.5 带误差线的条图与误差图

1. 绘制带误差线的条图

在许多分析问题中,研究者希望用条图来表示各类某指标均数的高低,并同时给出其区间估

计的范围。具体而言,所希望标识出的范围有以下3种情况。

(1) 指定可信度的均数可信区间:最常见的情形为95%可信区间。

(2) 均值 \pm 指定倍数的标准差:最常见的情形为2倍标准差,此时计算出的区间实际上是正态分布下的95%个体参考值范围。

(3) 均值 \pm 指定倍数的标准误:最常见的情形为2倍标准误,此时计算出的区间基本上等价于正态分布下的95%可信区间。

上述这种带误差线的条图在SPSS中也是可以绘制的,具体操作是在绘图时,在“元素属性”选项卡的“编辑属性”列表框中选择图形元素“条”,并在其下方选中“显示误差条形图”复选框,此时可以将误差线范围指定为确定比例的可信区间(默认为95%,可修改),或者2倍标准差/标准误,此处倍数也可以修改。图10.27给出的就是误差线范围分别表示均数可信区间和个体参

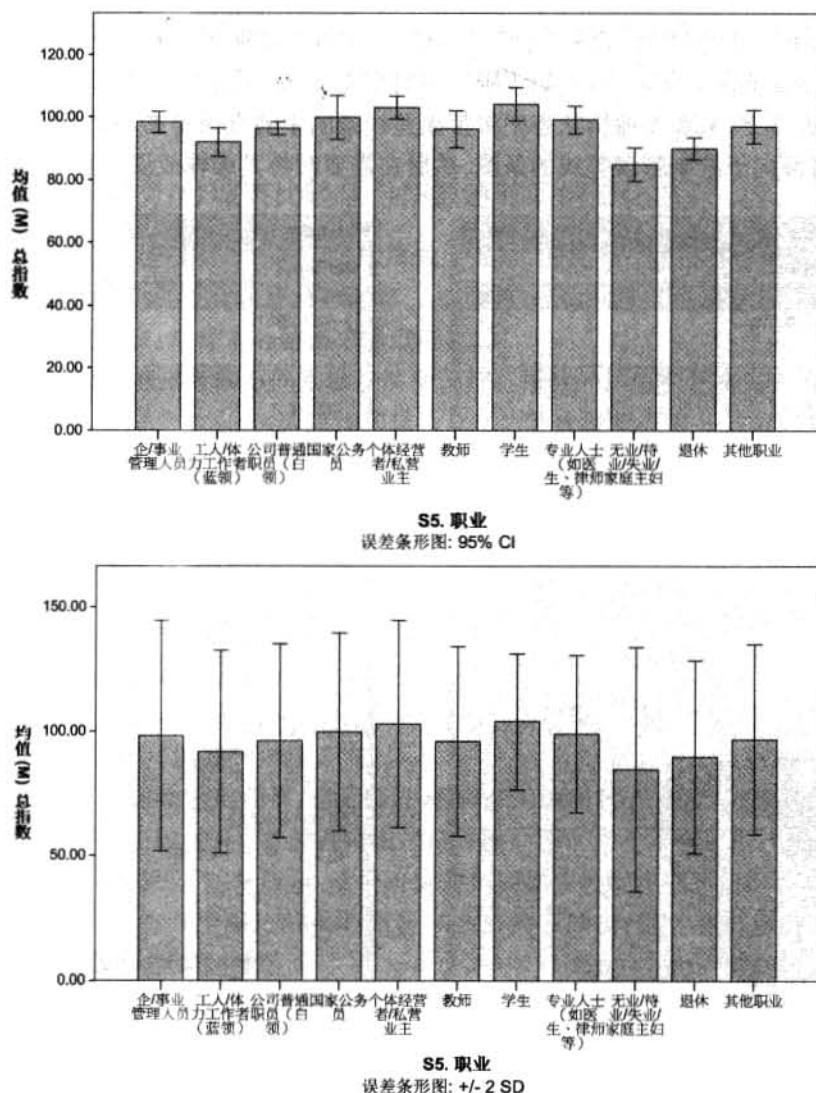


图 10.27 误差线范围分别表示均数 95% 可信区间和个体值 95% 参考值范围的条图

考值范围的条图。注意由于 index1 基本上服从正态分布,因此采用 95% 可信区间和采用 2 倍标准误差绘制的图形基本相同,此处只给出了前者的示意图。

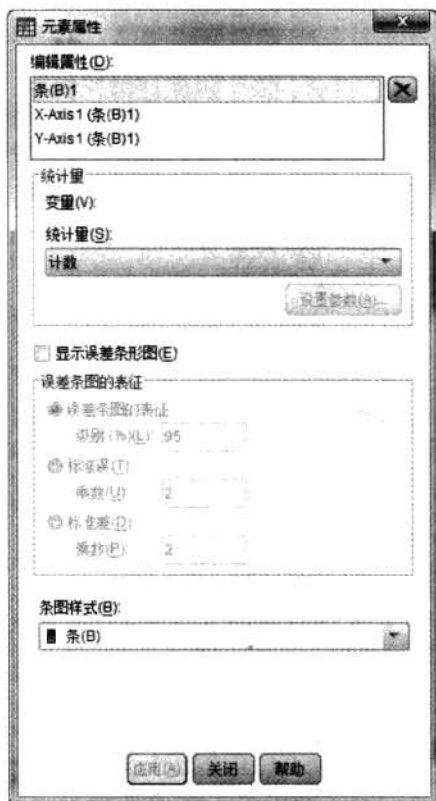


上述带误差线的条图可以在编辑状态中通过隐藏误差线转换为普通条图,当然也可以再重新转换回来。但是普通条图则无法通过编辑方式在图形中增加误差线。原因很简单,图形中并未存储相应的统计信息!这和前面普通条图在绘制时不能直接按统计量排序的逻辑是完全一致的。

2. 误差图

误差图目前也被归入条图组中,用于显示数据的可信区间、标准差或均值标准误的范围,从而估计其离散度。

SPSS 中的误差图也可以绘制为单式或复式,其绘图操作界面和条图没有明显的差异,图 10.28 给出了条图和误差图的“元素属性”对话框,从中可以发现除最下方隐去了直条样式以外,两者几乎是完全相同的。实际上,上面刚刚介绍的带误差线的条图就可以看成是普通条图和误差图的组合。换言之,只要在编辑状态中将带误差线条图中的直条隐去,相应的图形就变成了误差图,因此作者倾向于绘制带误差线条图,然后在需要时将其编辑成误差图。



(a)



(b)

图 10.28 条图与误差图的“元素属性”对话框界面比较

10.6 线图、面积图、点图与垂线图

线图用线段的升降表示一个事物随另一个事物(如时间)的变化趋势,一般而言,它所反映的指标类型和条图完全相同,可以是频数、构成比等分类变量描述指标,也可以是均数、标准差等连续变量的汇总指标。区别在于线图更倾向于反映连续变量的汇总指标,同时线图的另一个数轴应当代表一个有序分类变量的取值情况(最常见的例子就是年代),从而通过连线的走向变化来考察相应指标的变化趋势。因此,线图的两个坐标轴和条图一样,一般是一个分类轴,一个连续轴,只是分类轴代表的是一个有序变量而已。



从绘图原理上讲,线图实际上是先将有序分类变量各类别上相应指标的散点绘制出来,然后将各散点连接起来(一般使用直线)。因此虽然线图往往是由一条或多条折线构成的,但图形的骨架实际上是由多个隐藏起来的散点构成的。明白了这一点,会对理解线图的编辑功能大有帮助。

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

10.5节中对A3a的选项随月份的变化规律做了探讨,得到了一些很有价值的线索。本节将进一步对信心指数随月份的变化规律进行考察。显然,完成这一任务的最佳图形工具是简单线图。不过这里将更进一步:将样本分城市来加以观察。

例 10.7 现希望分城市考察不同月份总消费者信心指数的均值变化有无内在趋势。

首先由于月份为有序分类变量,因此线图最为合适。其次题目要求分城市观察,因此需要在图形中绘制多条折线,即使用多重线图以分别呈现不同城市的数据变化规律。另外为了使得图形显示更为清晰,这里还会稍做编辑,具体操作如下。

(1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(2) 在图库中选择“线”图组,将右侧出现的多重线图图标拖入画布中。

(3) 将月份 time 拖入横轴框中。

(4) 将总指数 index1 拖入纵轴框中。

(5) 将城市 s0 拖入分组(设置颜色)框中,然后再双击该框,在打开的“分组区域”对话框中将分组依据由“颜色”改为“图案”。

(6) 单击“确定”按钮绘制出图形,然后双击图形进入编辑状态,将均值连续轴刻度范围修改为85~105,小数位数更改为0。拖放调整图例位置和绘图区大小至合适的比例。

最终所绘制的线图如图10.29所示,从中可以观察到如下数据特征。

(1) 在2008年底之前,3个城市的信心指数都是持续下跌的,随后在经济刺激计划的作用下开始上升,且在2009年底超过初值。

(2) 北京、上海、广州三地的信心指数变化规律不一,广州相对而言变化较平缓,而上海则涨跌幅度最大。

(3) 就平均水平而言,北京消费者的信心指数最高,其次为广州,上海消费者的信心指数最低。

(4) 在2008年之前,3地消费者的信心指数存在较大差异,但2009年末的指数差异则大为缩小。

将上述数据特征和A3a题目的分析结果相结合,就可以对信心指数随月份的变化规律及原因有一个全面的了解了。

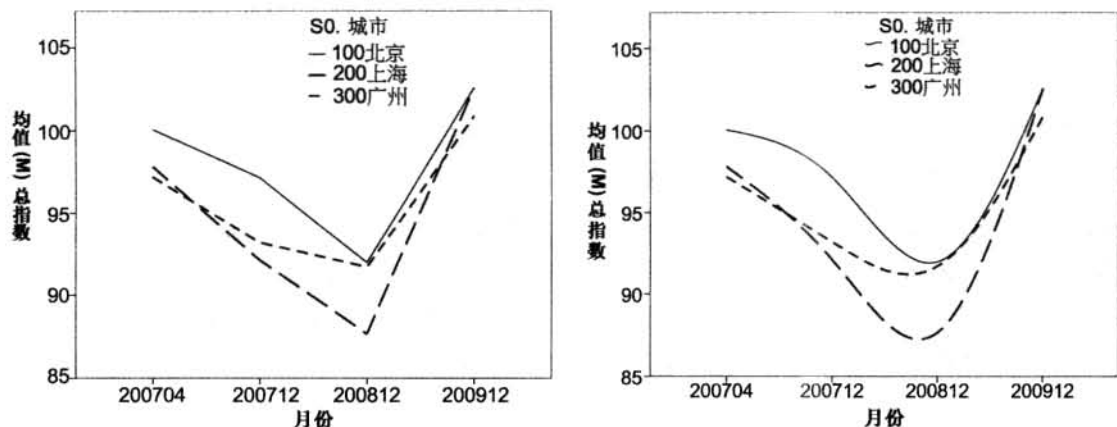



图 10.29 绘制出的多重线图及修饰后的多重线图

10.6.2 线图的编辑

如前所述,线图的图形框架实际上是由一些隐藏的散点所构成的,因此线图的编辑功能也会围绕着这一图形框架展开。

1. 更改数据点的显示方式

简单线图和多重线图默认不会显示各数据散点,但有时希望将其显示出来,此时可以选择“元素”→“添加标记”菜单项,或直接单击工具栏上的按钮,图形中所有的散点就会被显示出来。随后还可以使用针对散点的各种编辑功能进行修改,使之更为突出。

2. 更改数据点间的连接方式

在默认情况下,各散点间是采用直线方式连接的,因而整个线图呈折线形式。如果希望更改连接方式,则可以选中线图主体,之后在“内插线”选项卡(如图10.30(a)所示)中更改各数据点之间的连接方式,具体连接方式有4种。直线(Straight)、步长(Step)、跳跃(Jump)和样条光滑(Spline)。在本例中可以将连接方式改为“样条光滑”,以使得指数的变化趋势显得更为连贯。

3. 突出显示某一段连线

对于选中的线图主体,还可使用“线选项”选项卡(如图10.30(b)所示)进行一些修饰。上部的“显示类别范围条”复选框是要求将同分类下的散点垂直连接起来,即加绘垂线图;下方的“显示投影线”复选框则用于突出显示线的某一段。可以在下方的“类别”下拉列表框中选择一个具体的分类,该分类会将线图一分为二,其中一部分正常显示,另一部分则会突出显示。具体突出哪一部分将由最下方的“方向”下拉列表框确定。



为了在图形中更突出这一界值,可以使用加入横轴参照线的方法在相应位置添加一条参考线,使观察更为容易。

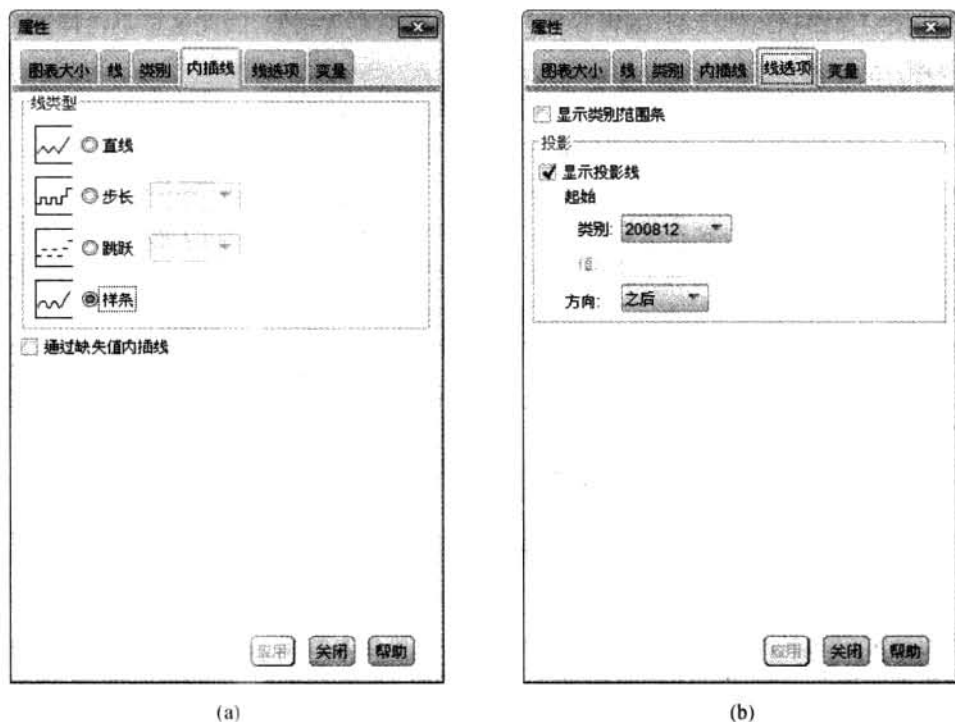


图 10.30 “内插线”选项卡和“线选项”选项卡

4. 半对数线图

半对数线图(Semi-logarithmic Line Graph)用于比较两种或两种以上事物的发展速度而不是绝对数量,当事物数量间差别较大时,普通线图往往难以客观地表达或相互比较发展速度,这时可以绘制半对数线图。由于0和负数均不能取对数,所以半对数线图的纵轴尺度起点为0.01,0.1,1,10……

在SPSS中绘制半对数线图非常容易,只需要在制图时在“元素属性”对话框中将相应数轴的刻度类型更改为对数即可。对于绘制好的图形,也可以将其编辑为半对数尺度:选中相应的连续轴,在“刻度”选项卡中将连续轴刻度更改为对数刻度即可。

10.6.3 面积图、点图与垂线图

1. 面积图

面积图(Area Chart)在绘图对话框中也被归在“线”图组中,是指用面积区块的大小来对不同类别情况下某指标的大小加以呈现的图形。实际上,面积图和条图、线图反映的是同类信息,之间没有本质性的区别。对于简单图形而言,只需要将条图中直条的顶点相连,就构成了线图,而将线图的折线下方全部涂黑,就变成了相应的面积图。

多重面积图和另外两种图形间的对应关系要略为复杂一些,分段条图和分段面积图可直接相互对应,它们可直接反映主分类变量各类别的情况,而多重线图实际上是和复式条图相对应的,可以确切地表示各分类组合下的情况。

上述3种图形可以在编辑状态下相互转换,具体的变换操作在条图的编辑中已经介绍了,

这里不再重复。而面积图的绘制、编辑等也与另两种图形非常相似,因此这里不再详述。

2. 点图

点图(Dot Chart)在绘图对话框中被分组在“散点”图组中,但这只是从图形元素上进行的分组,从统计特征上讲它和线图或条图的关系更近:如果在线图中将图形框架中隐藏的散点显示出来,同时将连线隐去,则图形就变换成了点图。

点图的适用范围与线图较为接近,也用于反映某指标随另一个指标的发展变化趋势,也最常用于反映数据本身或数据的变化速度随着时间变化的趋势,当数据点比较多时,点图比较有用。

3. 垂线图

垂线图(Drop-Line Chart)同样被分组在“散点”图组中,前面在多重线图的编辑中已经提到过,也需要多个变量或者多个分类的信息,但是不是绘制出多条折线,而是将属于同一类别的各散点连接起来,因而垂线的长短就可以反映出随着时间的变化数值的差异大小变动情况。因此,与多重线图相比,垂线图更加强调几个变量值随另一变量变化情况的差别所在。

10.7 散点图

散点图是常用的表现两个变量或多个连续性变量间有无数量关联的统计图,它用点的密集程度和趋势表示两个变量之间的相关关系与变化趋势。在进行相关/回归分析之前,绘制合适的散点图考察两个或多个变量间的相关关系及变化趋势是必须的。

在SPSS中有4种散点图,即用于描述两变量间关系的简单散点图、多个变量之间两两关系的散点图矩阵、多个自变量与一个应变量或多个应变量与一个自变量之间关系的重叠散点图以及3个变量之间综合关系的三维散点图,下面就来依次进行介绍。

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

例 10.8 利用简单散点图考察年龄S3与总消费者信心指数间的数量关联趋势。

由于在本例中实际上考察的是index1如何随着年龄的变化而变化,也就是说,在这两个变量中,S3相当于影响因素(自变量),而index1则是被影响的指标(因变量),在这种情况下习惯上将因变量index1置于纵轴,具体操作如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“散点”图组,将右侧出现的简单散点图图标拖入画布中。
- (3) 将年龄S3拖入横轴框中,将总指数index1拖入纵轴框中。
- (4) 单击“确定”按钮。

最终所绘制的散点图如图10.31所示,从中可以观察到如下数据特征。

- (1) 随着年龄的上升,消费者信心指数的平均水平有缓慢的下降趋势,且两者间关联基本上呈线性趋势。
- (2) 消费者信心指数在不同年龄段上的离散程度相差不明显。
- (3) 消费者信心指数存在若干偏小的数值,其中在30~40岁间的一位消费者其信心指数居然为0。

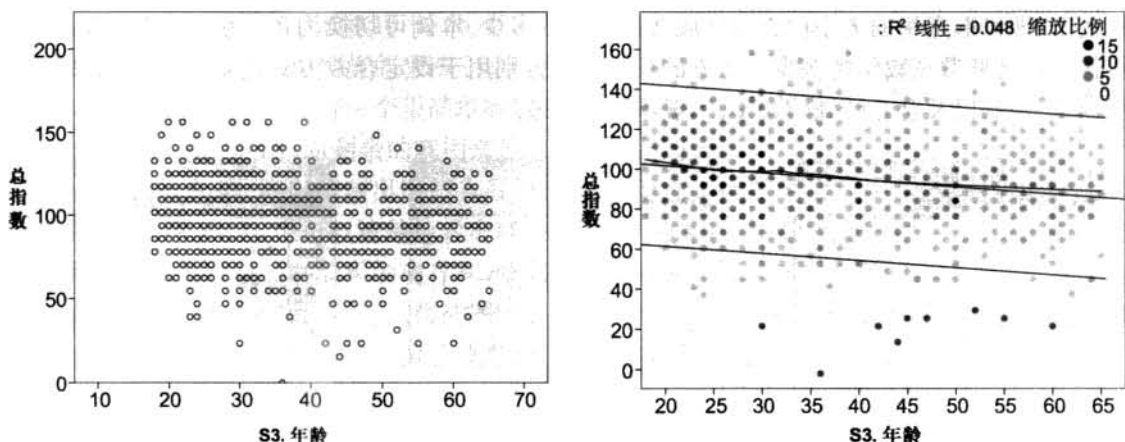



图 10.31 默认绘制生成及编辑完毕后的简单散点图

如果进一步进行两变量间的回归或相关分析,上述信息将对分析工作起到重要的方向指导作用。

10.7.2 散点图的编辑

散点图中的图形元素以散点为主,因此前面讲述过的各种散点编辑功能,如更改散点样式、大小、只显示某些散点的标签等均可以加以应用。除此以外,散点图中还有一些独特的编辑功能,如更改散点密度的显示方式、加入回归趋势线等,这里将结合 10.7.1 小节的实例来加以讲解。

1. 用套索模式选择离群散点


在图 10.31 所示的散点图中可以看到下方有一些偏离稍远的点,可以将其显示方式编辑得更为醒目一些。但由于数量太多,如果采取按住 Ctrl 键依次选中的方式显然太麻烦,这里可以采用套索模式来实现。选择“编辑”→“套索选择标记”菜单项,或者直接单击工具栏上的按钮 , 可以看到鼠标变成套索形状,此时在散点图中按下鼠标左键,将希望选中的散点圈入一个闭合曲线环中,松开左键后就会看到这些散点已被同时选中,下面就可以对这些散点同时进行各种编辑操作了。



套索模式是 19 版新增的,老版本无此选项。

2. 改变过密散点图的显示方式

CCSS 案例数据的样本量虽然上千,但由于年龄只能取整数,信心值也并非任意取值,导致所绘制的散点图中有大量散点是重叠显示的,各部分的疏密无法分清,这会严重影响对散点图趋势的观察。

事实上,在大样本数据,或者变量取值范围有限的情况下,散点重叠的情形非常常见,对此问题可以采用散点合并的显示技巧加以解决。在编辑状态下选择“选项”→“块元素”菜单项,或者在快捷工具栏上直接单击按钮 , 就可以将散点转化为合并显示。同时在“元素”对话框中会增加一个“分箱”选项卡(如图 10.32(a)所示),用于设定具体的合并选项。默认的显示方式为标

记大小,即以散点块的大小代表该区域散点数量的多少,本例可切换为色彩强度,即以颜色的深浅代表该区域散点数量的多少。下方的几个框组分别用于设定合并方式的显示位置、合并区域的计算方式,以及合并区域的大小,一般不需要更改。

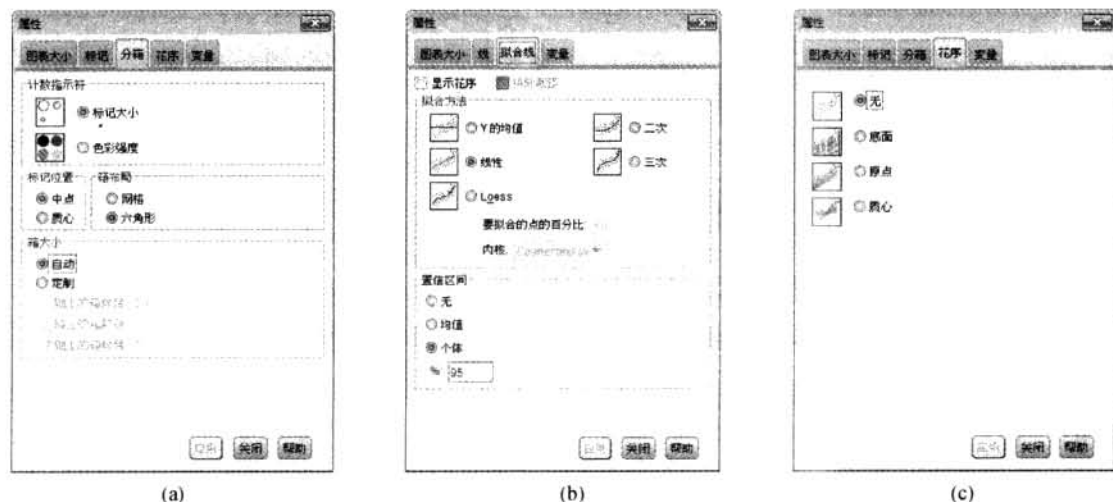



图 10.32 “分箱”选项卡、“拟合线”选项卡和“花序”选项卡

通过将本例的散点图转化为合并显示,就可以清楚地看出在 20~30 岁范围内,信心值在 100 上下的散点密度是比较高的,而整个散点图也遵循了中部较密、两边较稀疏的基本规律,并未发现明显违背常规的情形。

3. 添加回归趋势线和可信区间

作为回归问题预分析中的重要工具,如果能够在散点图中直接添加各种回归线,就能够提供更为丰富的信息,利用 SPSS 的散点图可以轻松地完成这一任务。在选中散点图主体后,选择“元素”→“总计拟合线”菜单项,或者直接单击工具栏上的按钮,就可以在图形中添加一条总样本的回归趋势线。

但是,上述操作在默认情况下添加的是线性趋势线,如果希望加以更改,则可以在选中趋势线后使用“拟合线”选项卡来进行操作。除给出应变量的均值(水平线)外,回归线的拟合方式还有以下 4 种。

- (1) 线性:就是根据最小二乘法确定的线性回归方程直线。这也是系统默认的方式。
- (2) 二次:根据最小二乘法,用二次方回归曲线对散点图中的数据点进行最佳拟合。
- (3) 三次:根据最小二乘法,用三次方回归曲线对散点图中的数据点进行最佳拟合。

(4) Loess:即局部加权回归光滑曲线(Locally Weighted Regression Smoother),该方法是根据数据局部的点拟合一条曲线。也就是说对于任何一点,该点的曲线仅依赖于这点以及指定范围内的临近点的观察值来确定,因此可以将曲线拟合得非常光滑,与实际点吻合得很好。下方的百分比文本框用于指定拟合曲线时利用样本中多少比例的散点。拟合的点越多,则曲线越临近于直线。拟合的点越少,则充分利用临近点的信息,所得曲线越圆滑,与散点越吻合。在多数情况下默认设定就可满足要求,不需要更改。

在“拟合线”选项卡(如图 10.32(b)所示)的下方还有一个“置信区间”框组,用于绘制相应回归曲线的均数或个体预测值的 95%(或其他指定的可信度)可信区间。当要求绘制区间时,回归线本身将会消失。对此有一个很简单的解决办法,即多绘制几条相同的回归线,将其中某几条变换为希望显示的区间,而剩余的就用来显示原有的回归线。

对例 10.8 而言,可以绘制出回归直线所对应的 95% 个体参考值范围,并同时加绘出 Loess 样条曲线,从图 10.31 中可以清楚地看出样条曲线和回归直线的趋势非常近似,也就是说,年龄和总指数之间的数量关联如果的确存在,那么应当基本上是服从线性趋势的。

除添加回归趋势线外,还可以像线图中一样在散点图中添加散点间的连接线,并进行相应的编辑操作,如突出显示某一段等,留给读者自行操作,这里不再详述。

4. 添加钉线

钉线即 Spikes,其原意是钉子,或细而长的线,这里指的是在散点图上添加辅助线。可以是数据点到某一点,到轴线或平面的线,向下到 X 轴的线常称为垂线。


散点图的“花序”选项卡(如图 10.32(c)所示)用于添加钉线,钉线可以是数据点到原点(Origin),到两个轴线,或者到数据中心(Centroid)的线。一般而言,钉线主要用在一些有特殊用途的散点图中,如市场研究中多维偏好分析的结果图形,因此这里不再详述。

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

有时出于研究需要,需将两个或多组两个变量的散点图绘制在同一个图中,这样可以更好地比较它们之间的相关关系。此时可以考虑绘制分组散点图,也称为重叠散点图。在绘制分组散点图时要注意的是用于绘制统计图的变量取值大小应比较接近,否则有的变量组的相关关系表现很清楚,而有的变量组的相关关系则缩小成一堆,难以分辨。

例 10.9 进一步分性别考察年龄 S_3 与总指数 $index1$ 的关系,以判断不同性别间年龄对信心指数的影响趋势是否不同。

由于本例重点在于比较不同性别间的散点图趋势是否不同,因此重叠散点图是较好的选择,而且绘图重点应当是散点分布与回归线并重,因此操作如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“散点”图组,将右侧出现的分组散点图图标拖入画布中。
- (3) 将年龄 S_3 拖入横轴框中,将总指数 $index1$ 拖入纵轴框中。
- (4) 将性别拖入分组框中。
- (5) 单击“确定”按钮绘制出图形,随后双击进入编辑状态,对坐标轴尺度、图例位置等进行适当的调整。
- (6) 将散点图更改为按标记大小区分的合并方式。
- (7) 选择“元素”→“子组拟合线”菜单项,或直接单击工具栏上的按钮,在图形中添加分组回归线,并在“拟合线”选项卡中将回归线种类更改为 Loess。

最终绘制完毕的分组散点图如图 10.33(a)所示,从中可以看出在不同的性别人群中,无论是散点分布范围还是回归曲线的趋势都不存在明显的区别,也就是说不同性别间年龄对信心指数的影响趋势是基本相同的。

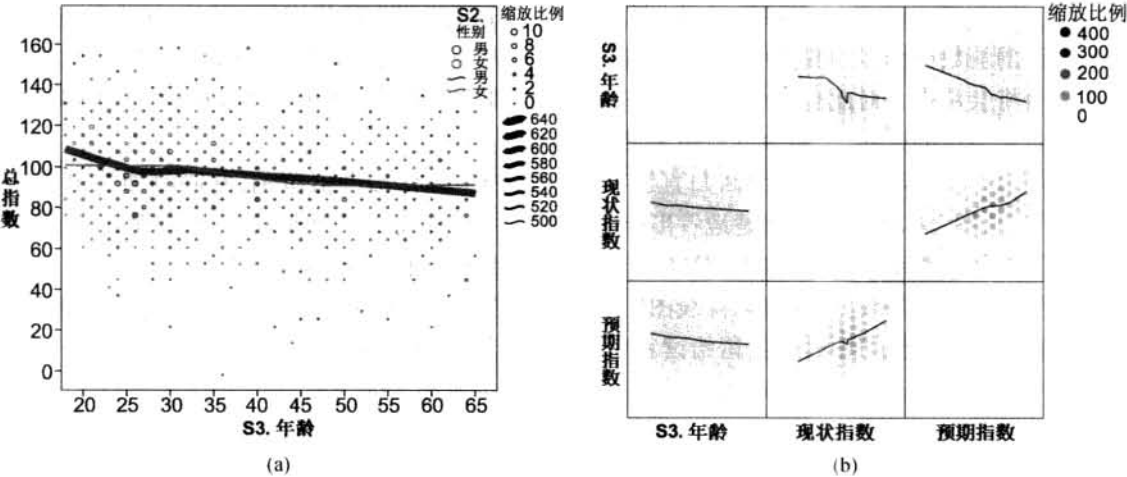


图 10.33 分组散点图及散点图矩阵示例



在分组散点图的横轴和纵轴中也可以直接选入变量组,此时所绘制的散点图按照横、纵变量的排列组合加以确定,例如在横轴中选入变量组 X、Y,在纵轴中选入变量组 A、B,则所绘制的分组散点图中会包含以下几对散点组合: $X * A$ 、 $X * B$ 、 $Y * A$ 和 $Y * B$ 。

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

当欲同时考察多个变量间的相关关系时,若一一绘制它们间的简单散点图,十分麻烦。此时可利用散点图矩阵来同时绘制各自变量间的散点图,这样可以快速发现多个变量间的主要相关性,这一点在进行多元线性回归时显得尤为重要。

例 10.10 在前面分析的基础上,进一步考察年龄 S3 对现状指数、预期指数的影响。本例实际上也可以采用分组散点图来考察,但可能会遇到散点范围严重重叠的问题,因此改用矩阵观察,操作步骤如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“散点”图组,将右侧出现的散点图矩阵图标拖入画布中。
- (3) 按住 Ctrl 键选中年龄 S3、现状指数 index1a、预期指数 index1b,将其一起拖入画布上的矩阵框中。
- (4) 单击“确定”按钮绘制出图形,随后双击进入编辑状态,对坐标轴尺度、图例位置等进行适当的调整。
- (5) 将散点图更改为按颜色区分的合并方式。
- (6) 选择“元素”→“总计拟合线”菜单项,在图形中添加回归线,并在“拟合线”选项卡中将回归线种类更改为 Loess。

最终绘制出的矩阵如图 10.33 (b) 所示,整个图形类似于一个 3×3 矩阵,不同的是此处矩阵的元素是一个一个的散点图。三个变量两两交叉,就形成了 9 个格子。每个变量所在的横行的图形,其纵轴都是该变量所在的那一列的图形,其横轴也为该变量,对角线处则为空白。



散点图矩阵的对角线处实际上会显示该变量的直方图,只是在默认情况下是隐藏的,如果希望显示,则选择“选项”→“显示沿对角线绘制的图表”菜单项即可。

从散点图矩阵中可见,年龄与现状指数和预期指数均呈负相关关系,年龄越大,指数值越低,但似乎年龄与现状指数之间存在一定的曲线关联,在后续分析中要加以注意。

10.7.5 三维散点图

在散点图矩阵中虽然可以同时观察多个变量间的联系,但是两两进行平面散点图的观察的,有可能漏掉一些重要的信息。三维散点图就是由3个变量确定的三维空间中研究变量之间的关系,由于同时考虑了3个变量,常常可以发现在两维图形中发现不了的信息。

1. 三维散点图的绘制

仍以上面的问题为例,如果希望直接做出 S3、index1a、index1b 的三维散点图,则只要在对话框中将它们依次定义为 X、Y、Z 轴即可,所绘制出的三维散点图如图 10.34 所示。该图形将3个变量间的关系在同一个坐标空间中立体地表现了出来,使用它可以更加清晰和直观地对应/自变量间的关系进行观察,发现在二维空间中可能无法看到的信息,如曲线关系、异常值等。但是,由于实际上只能在二维平面上观察三维散点图,所以在观察时必须要结合旋转功能,这将在下面加以讲述。

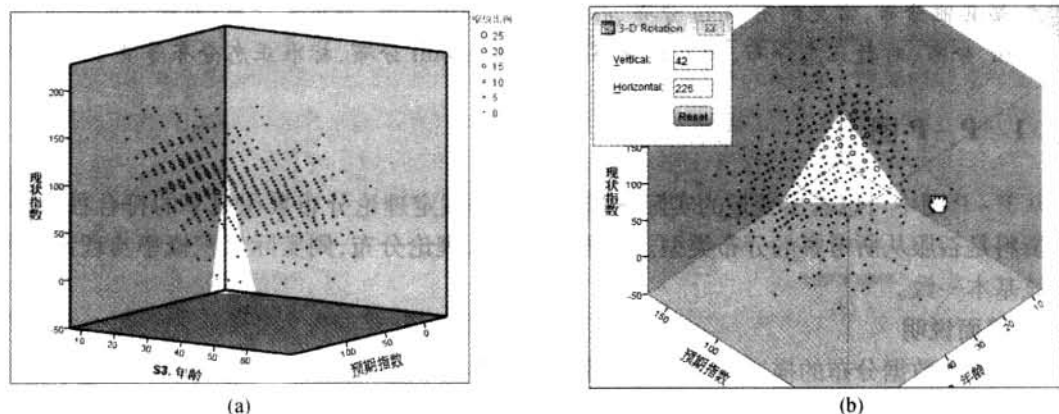



图 10.34 三维散点图及其自由旋转操作示例

2. 三维散点图的旋转观察

对三维散点图的旋转进行观察需要进入编辑状态,然后选择“编辑”→“3D-Rotation”菜单项,或直接单击工具栏上的按钮,就会看到出现一个3D旋转托盘,如图10.34所示,用它进行旋转,可马上看到旋转的效果。具体操作时既可以在对话框中更改纵横坐标,也可以按下鼠标左键对三维图形做各个方向的直接拖动,即压住左键,然后将鼠标向需要的方向移动,图形就会向相应方向转动,直至松开鼠标左键为止。

3. 三维散点图的缩放

在默认情况下,三维散点图主体只占据了图形的一部分区间,这是由于默认对散点图的观察距离较远所致。如果希望散点图占据主要显示区间,则可以选中散点图主体,然后在“3-D 旋转”选项卡中将距离改得较近一些即可,默认为70,一般改为30即可使图形占据大部分显示面积。


4. 三维散点图与其他散点图的相互转换

由于散点图都是以散点形式来表现各连续性变量间的数量关联的,因此它们之间是可以相互转换的。具体而言,仍然是首先在“属性”对话框中找到“元素”选项卡,然后将其中所列的各图形元素赋值修改为所需的种类即可。

在理论上,三维散点图和散点图矩阵、分组散点图都是可以相互转换的,它们也都可以转换为简单散点图。但目前的 SPSS 版本似乎不太推荐这样做,其绘图引擎对这些转换的支持也不是非常便捷,因此这里也不再展开论述。

10.8 P-P 图和 Q-Q 图

大多数假设检验方法都假定研究数据总体服从某种特定分布,如正态分布、二项分布等,除使用专门的检验方法加以考察外,更常用的方法是用图形来直接观察。直方图和茎叶图都是评估数据分布的常用图形,但它们不能直观给出数据分布与理论值相差多少,P-P 图和 Q-Q 图则可以给出上述信息,是非常有用的观察工具。

 P-P 图和 Q-Q 图最常应用于判断变量是否服从正态分布,但实际上它们还可以用于考察其他分布,常见的有 Beta 分布、指数分布、伽玛分布、半对数分布、拉普拉斯分布、logistic 分布、对数正态分布、帕累托分布、t 分布、weibull 分布、标准正态分布等共 13 种分布。

10.8.1 P-P 图

从 P-P 图中可以看出变量的实际累积概率与其假定理论分布累积概率的符合程度,从而判断资料是否服从所考察的分布类型。如果变量服从理论分布,则实际累积概率与理论累积概率应该基本一致。

1. 界面说明

由于涉及数据分布的描述,P-P 图在几个版本的 SPSS 中均被放在“分析”→“描述统计”菜单中,其对话框如图 10.35(a)所示,内容看似庞杂,实际上非常简单。



(a)



(b)

图 10.35 “P-P 图”和“Q-Q 图”对话框

(1) “变量”列表框:用于选入希望考察的变量,可一次性选入多个变量同时绘制多个 P-P 图。

(2) “检验分布”下拉列表框:用于指定希望考察的理论分布,默认为正态分布,在下方可以进一步指定相应分布的自由度、位置参数、形状参数等。

(3) “转换”框组:提供了“自然对数变换”、“标准值”、“差分”以及“季节差分”这 4 种数据变换方法,以考查变换后的数据分布情况。

(4) “比例估计公式”框组:实际上应当翻译成“概率估计公式”,即用于估计样本累计概率分布的具体算法,一般不需要更改。

(5) “为结指定的秩”框组:指定样本中出现重复数值时的处理方式,默认的均值就非常合适,不需要更改。

2. 结果解释

这里以 CCSS 项目中的总指数为例来说明如何阅读 P-P 图,具体图形如图 10.36 所示。

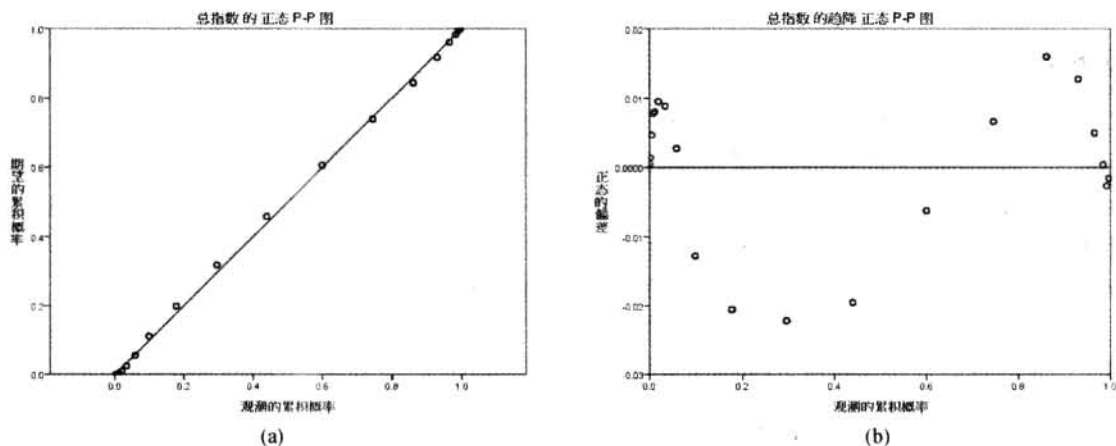


图 10.36 总指数的 P-P 图分析结果

图 10.36 所示的两个图分别为 P-P 图和去势 P-P 图 (De-trend, 软件中翻译为“趋降”并不恰当), 图 10.36(a) 的两个坐标轴分别表示理论累积概率和实际累积概率, 如果数据服从正态分布, 则其中的数据点应和理论直线(对角线)基本重合。可见 index1 的实际分布和理论分布基本接近。为了进行更仔细的观察, 可以继续观察右侧的去势 P-P 图, 该图反映的是按正态分布计算的理论值和实际值之差的分布情况, 即分布的残差图。如果数据服从正态分布, 则数据点应较均匀地分布在 $Y=0$ 这条直线上下。从图 10.36 中可见残差虽然有一定的上下波动, 但绝对差异均小于 0.05, 这在绝大多数研究中都是可以忽略的分布概率差异。由此可以看出, 变量 index1 的原始数据与正态分布的理论数据相差很小, 可以认为其基本服从正态分布。

下面来看一个不服从正态分布的例子, 如图 10.37 所示, 变量为年龄 S3, 从其 P-P 图和去势 P-P 图可见, 年龄的实际分布和理论分布有明显的差异, 其残差绝对值最高时超过了 0.1, 因此可以判断年龄并不服从正态分布。

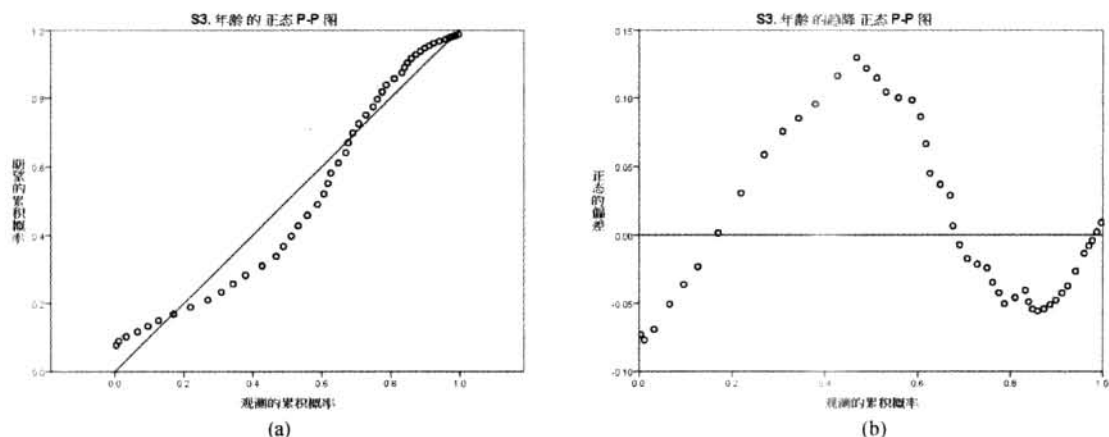


图 10.37 年龄 S3 的 P-P 图和去势 P-P 图

10.8.2 Q-Q 图

Q-Q 图的基本原理与 P-P 图非常类似,也用于比较变量的实际分布与其所假定的理论分布是否一致。但 P-P 图比较的是两者的累计概率分布,而 Q-Q 图则是根据变量的实际百分位数与理论百分位数进行绘制的,或者说得更通俗一点,相比之下 Q-Q 图的适用条件更宽松,结果也更稳健一些。但是对 Q-Q 图进行考察时存在一个很大问题,即不像 P-P 图可以用经验界值来判断样本是否和理论分布存在明显差异,因此应用相对较少。

Q-Q 图的对话框界面、操作方式和结果阅读方式和 P-P 图几乎完全相同,读者参照 P-P 图的相关内容进行操作即可,这里不再重复说明。

10.9 控制图与 Pareto 图

10.9.1 控制图

1. 图形简介

任何自然过程都有随机变异,产品的生产线也不例外,在生产过程中,产品质量一方面会出现随机波动,但另一方面也可能是由可辨识的、作用明显的原因,如误操作、设备故障等所引起,后一种情况显然通过采取适当措施可以被及时发现并排除。而控制图就是用于分析和判断生产工序是否处于稳定状态的一种统计图。

控制图的作用原理非常简单,当生产过程仅受随机因素的影响,产品的质量特征的平均值和变异都基本保持稳定时,称之为处于受控状态。此时产品的质量特征服从某种确定概率分布的随机变量,因此可以每隔一定时间在生产线进行抽样,若其数值符合原分布规律,就认为生产过程正常,否则就认为生产中出现某种系统性变化,或者说过程失控。此时就需要考虑采取包括停产检查在内的各种措施,以期查明原因并将其排除,以恢复正常生产,不使失控状态延续下去。

2. 控制图的种类

控制图的类型可以分得非常细,多达十几种,但如果按照数据特征,则首先可分为计量控制图和计数控制图两大类。此外再根据是对个体还是均数的变动情况进行监测,以及具体是用全距、百分位数还是标准差作为控制范围来做进一步的细分。

从“预定义”对话框就可以看出,SPSS 提供了比较全面的控制图种类,其具体用法如下。

(1) X 条形图、R 图和 s 图:均数、全距和标准差控制图,本选项包括两种组合控制图,即均数—全距组合控制图和均数—标准差组合控制图。前者将在图中显示每个亚组测量值的均值和全距。当亚组内例数比较少(比如少于 10 个),不宜计算标准差时,选用这种图。而当例数较多时,由于采用标准差的效率更高,也更稳定,因此推荐使用后者。

(2) 个体,移动全距:均数的计算要求每个亚组中的案例数大于 1,当各亚组中均只有一个案例时,就只能采用这里的个体值移动全距图,在图中显示个体测量值,图中个体值的顺序与数据的顺序一样。移动全距图用于显示每个所选间隔段里的数值全距,也就是说,如果间隔段是 3,移动范围图用于显示目前记录、其前一条记录和前两条记录之间的数值全距,反映数据波动情况的变化。

(3) p、np:不合格品率、不合格品数控制图。p 显示每个亚组里不一致的记录所占的比例,用于控制对象为不合格品率或合格品率等计数值质量指标的场合。但是由于计算不合格品率需要进行除法运算,比较麻烦,所以在样本量大小相同的情况下,用 np 图比较方便,后者显示的是每个亚组内不一致记录的数量。

(4) c、u:缺陷数、单位缺陷数控制图。u 显示指定单位范围里所出现的缺陷数目。当样品的大小保持不变时可用 c 控制图,而当样品的大小变化时则应换算为平均每单位的缺陷数后再使用 u 控制图。

3. 界面说明

“控制图”菜单项在近几个版本的 SPSS 中被放置在菜单“分析”→“质量控制”中,选择后首先打开预定义对话框,用于选择具体的图形种类,随后会进一步打开各种控制图的具体操作界面,如图 10.38 所示,这里以个体值控制图为例说明如下。

(1) “过程度量”文本框:选入用于质量控制的变量。

(2) “图表”框组:选择只绘制个体值控制图,还是同时绘制移动全距控制图,下方的“跨度”文本框用于输入计算移动全距时的指定范围。

(3) “选项”子对话框:选择控制限和均数线间包括的标准差数,默认为 3 倍标准差。

(4) “控制规则”子对话框:指定一个或多个规则。如果某个点违反规则,则它在图表中具有与受控点不同的形状和颜色。该功能允许用户快速识别不受控制的点。

(5) “统计量”子对话框:在其中可人为规定控制限,并可选择控制图中使用的一些统计指标。

4. 实例分析

使用控制图来考察 2007 年 4 月的 CCSS 数据采集是否正常,如果数据是随机采集的,则按照样本 ID 号采集的数据应当上下随机波动,即使出现异常值,也不应当有聚集性,反之则可能存在数据造假或错误可能。为了简化讲述,这里只考察总指数 index1 这一个变量,操作步骤如下。

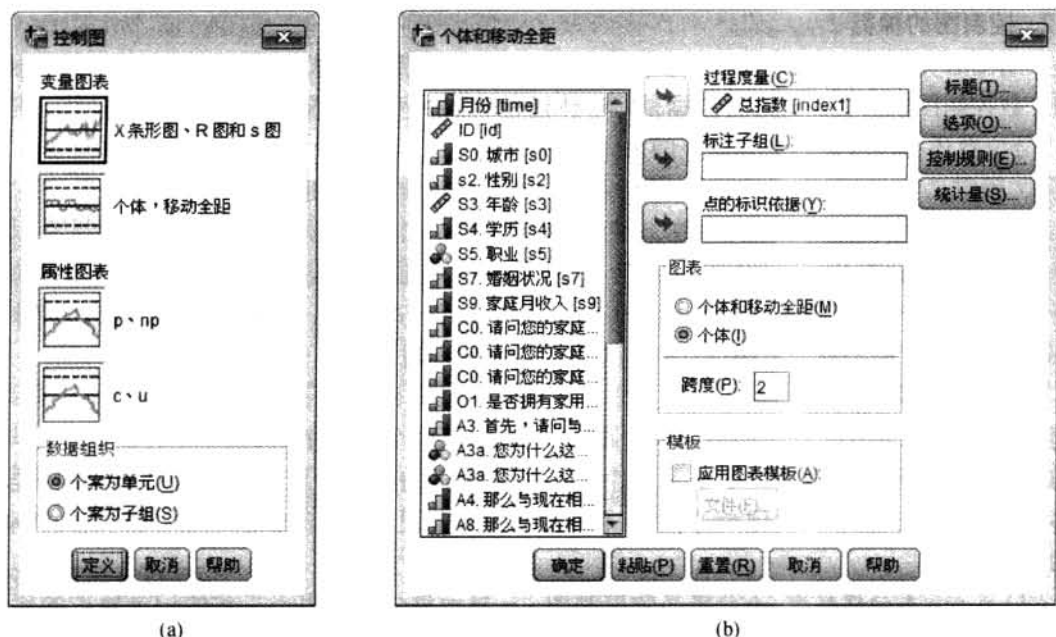


图 10.38 控制图的预定义对话框和主对话框

- (1) 选择 2007 年 4 月的所有案例进行研究。
- (2) 选择“分析”→“质量控制”→“控制图”菜单项。
- (3) 在打开的控制图预定义对话框中选择“个体,移动全距”选项。
- (4) 将 index1 选入“过程度量”文本框,图表设定为“个体”。
- (5) 在“控制规则”子对话框中选中“在 $+3\sigma$ 以上”和“在 -3σ 以下”复选框。
- (6) 单击“确定”按钮。

按照上述选择,本例绘制出的控制图如图 10.39 所示,可见有两个案例的 index1 数值偏低,

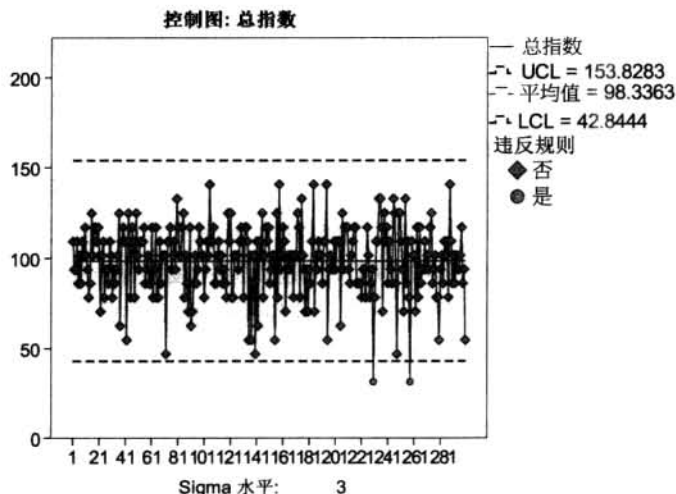


图 10.39 个体值控制图示例

且超过了控制线下界,但由于按照上下 3 倍的标准差范围,2007 年 4 月的 300 例样本应当大致出现 3 个异常值,因此两个异常值是可以接受的。此外从曲线的走势来看,整个样本的 index1 数值是在均数上下随机波动的,并未出现上升或下降的趋势,这说明 CCSS 项目一切运转正常,数据的质量是可以得到保证的。

10.9.2 Pareto 图

1. 图形简介

Pareto 图(Pareto Chart)又称为排列图,来自于 Pareto 定律,该定律认为绝大多数的问题或缺陷产生于相对有限的起因,实际上也就是常说的 80/20 定律,即 20% 的原因造成 80% 的问题。

Pareto 图属于双纵轴图,本质上是条图和线图的组合,管理者或研究者常常会面对许多选择类别,需要用较快的视觉方式评估每类的相对重要性。而 Pareto 图以条图方式将各因素按降序排列,其条形的长短表示各组绝对数的大小;在条图上方加绘直条累计百分比的曲线(称为 Pareto 曲线),线段的上升表示累计百分比的增加情况,可直观地找出主要、次要因素。

Pareto 图的典型应用是显示由于各种原因引起的缺陷数量或不一致的频数分布,按照 Pareto 图的一般阅读习惯,影响质量的主要因素通常分为 3 类:A 类为累计百分数在 70%~80% 范围内的因素,是主要影响因素;B 类是除 A 类之外累计百分数在 80%~90% 范围内的因素,是次要因素;C 类为除 A、B 两类之外百分比在 90%~100% 范围的因素。按此原则,使用者就可以根据条图顶端生成的曲线快速确定项目实施失败的主要原因。

2. 实例分析

“Pareto 图”菜单项也被放置在菜单“分析”→“质量控制”中,有简单和堆积两大类,分别对应了简单条图和分段条图。其主操作界面均非常简单,直接将希望分析的变量选入类别轴框中即可。然后在上方设定直条高度所指示的指标,默认为类别频数。

这里以职业 S5 为例解读一下 Pareto 图的输出。从图 10.40 中可看出受访者职业以白领最



图 10.40 Pareto 图对话框界面及示例

多,其次为管理人员、退休和私营业主。在所分的一共11个类别中,前5类大约占80%,并没有出现很高的聚集性,这说明项目的样本分布还是比较分散的。

10.10 其他统计图

10.10.1 高低图

1. 图形简介

股票、商品、货币及其他市场数据每周、每日甚至每时的波动都相当大。为了图示长期变动趋势,同时又能知道短期的变化,必须采用相应的专用图形工具来分析。高一低图就是为此而设计的。

SPSS 在“图表构建程序”对话框中专门提供了一组高-低图组,其中包括以下几种图形。

(1) 高-低-收盘图:表示单位时间内某现象的最高数值、最低数值和最后数值。这种图形适用于股票、期货和外汇金融等,它可以说明每天的最高价格、最低价格和收盘时的价格。

(2) 简单全距图:也称为单式全距图,表明单位时间内某现象的最高数值和最低数值。单式全距图与单式高低收盘图的区别是省去了最后数值。

(3) 分组全距图:也称复式全距图,它表示在单位时间内两个或以上现象的最高数值和最低数值。

(4) 差异面积图:是说明两个现象在同一时间内相互变化对比关系的线性统计图。

2. 实例分析

下面以上证指数(SH999999)2010年上半年的实际走势数据来说明高低图的绘制方法,数据见文件 SH999999.txt,具体操作步骤如下。

(1) 在 SPSS 中用文本数据向导读入数据文件,注意日期变量应当设置为日期型。

(2) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(3) 在图库中选择“高-低”图组,将右侧出现的高-低-收盘图图标拖入画布中。

(4) 将最高、最低、收盘3个变量依次拖入高变量、低变量、关闭变量3个框中,将日期变量拖入横轴框中。

(5) 单击“确定”按钮。

相应的简单高低图如图 10.41 所示,每个直条代表一天的交易数据,直条的上、下范围分别代表当天的指数最高、最低值,圆圈则代表当天的收盘指数值。显然,上证指数在上半年呈现的是先下跌,然后横盘反弹,最后一路大跌的走势。



从图 10.41 中可以发现序列有若干间断点,这实际上代表了股市停盘的日期,比如春节以及五一假期。在股票软件中,这些停盘日期都是自动略去的,而此处指定了日期为日期型格式,因此 SPSS 不会将其省略。在本例中如果将日期指定为字符型变量,则这些时间点也不会出现。

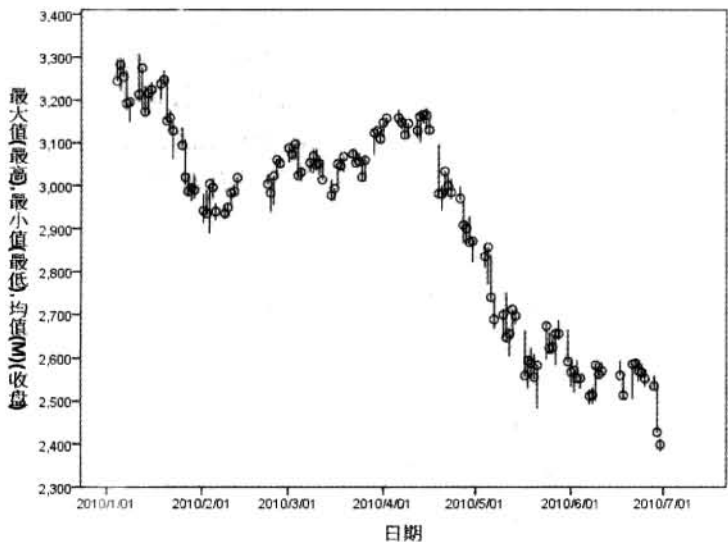


图 10.41 上证 A 股指数 2010 年上半年走势的高低图

10.10.2 ROC 曲线

ROC(Receiver Operating Characteristic, 受试者工作特征)曲线,也称为“接收者操作特征曲线”。它是一种得到广泛应用的数据统计方法,最早于 1950 年应用于雷达信号检测的分析,用于区别“噪声”与“信号”,后来应用于心理学研究。1960 年 Lee Lusted 首先认识到 ROC 分析法在医学判别疾病方面可能会有作用,从而开拓了其新的应用领域。

随着医学的发展,新的检测检验方法层出不穷。ROC 曲线及 ROC 曲线面积可作为评价某一诊断方法准确性的指标。通过对同一疾病的多种诊断试验进行分析比较,可帮助临床医生筛选出最佳诊断方案。

1. ROC 曲线的基础知识

对于一组经金标准诊断的病人和正常人,进行某项新的诊断试验,其结果汇总如表 10.1 所示。

表 10.1 诊断试验结果汇总表

试验	病人	正常人	合计
阳性	a	b	$a + b$
阴性	c	d	$c + d$
合计	$a + c$	$b + d$	$a + b + c + d$

真阳性率(灵敏度) = $\frac{a}{a + c} \times 100\%$

真阴性率(特异度) = $\frac{d}{b + d} \times 100\%$

假阳性率(误诊率) = $\frac{b}{b + d} \times 100\%$

假阴性率(漏诊率) = $\frac{c}{a + c} \times 100\%$

若检测结果为定量资料(或等级资料),以不同的检测值作为判断阳性、阴性结果的阈值时可分别计算出相对应的特异度和灵敏度,以 $1 - \text{特异度}$ 为横轴、灵敏度为纵轴,将坐标为 $(1 - \text{特异度}, \text{灵敏度})$ 的数据点在平面直角坐标系上绘制出来,所得曲线即为 ROC 曲线。

由 ROC 曲线的原理可知,一个优良的诊断试验其 ROC 曲线应该是从左下角垂直上升至顶线,然后水平方向向右延伸到右上角的。如果 ROC 曲线沿着对角线方向分布,表示分类是机遇造成的,正确分类和错分的概率各为 50%,此时该诊断方法完全无效。

如果两条曲线不交叉,那么可以根据它们的表现形态比较两个试验的优劣:更外面的、离对角线更远的曲线,其灵敏度和特异度均高于里面的、离对角线更近的曲线。

2. 实例分析

例 10.11 某医师对经金标准诊断的 55 名病人、45 名正常人分别进行两种诊断试验检查,结果分别为 test1、test2。试对其绘制 ROC 曲线,数据见 roc.sav。

选择“分析”→“ROC 曲线图”菜单项,打开如图 10.42(a)所示的对话框,由于界面非常简单,这里不再详细解释,直接给出操作步骤如下。

- (1) 将 test1 和 test2 选入“检验变量(T)”列表框中。
- (2) 将 diag 选入“状态变量(S)”文本框,在下方指定 $\text{diag} = 1$ 表示研究对象为病人。
- (3) 在下方的复选框组中设置输出对角线、标准误和可信区间。

相应的 ROC 曲线如图 10.42(b)所示,可见 test1 的效果是远远好于 test2 的。SPSS 还进一步输出了两条 ROC 曲线下面积的标准误及各自的可信区间。

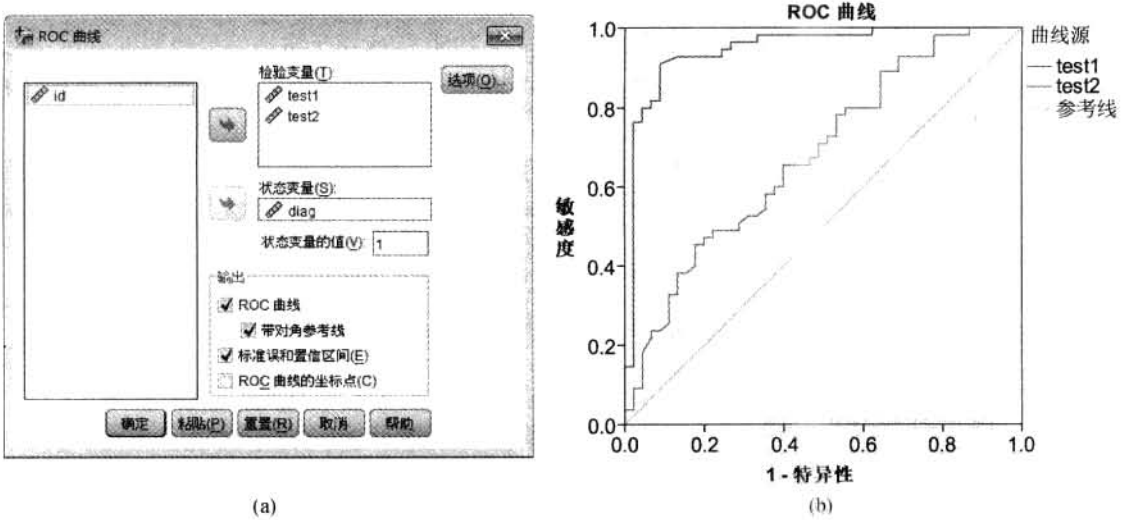


图 10.42 “ROC 曲线”对话框和图形

由图 10.43 可见,检测一(test1)的 ROC 曲线下面积为 0.947,标准误为 0.024,其 95% 可信区间为 0.900 ~ 0.994;检测二(test2)的 ROC 曲线下面积为 0.679,标准误为 0.053,其 95% 可信区间为 0.574 ~ 0.784。ROC 曲线下面积取值范围为 0.5 ~ 1.0。一般的,ROC 曲线下面积在 0.5 ~ 0.7 之间表示诊断价值较低,在 0.7 ~ 0.9 之间表示诊断价值中等,在 0.9 以上表示诊断价值较高。

检验结果变量	面积	标准误 ^a	渐进 Sig. ^b	渐近 95% 置信区间	
				下限	上限
test1	.947	.024	.000	.900	.994
test2	.679	.053	.002	.574	.784

检验结果变量: test1, test2 在正的和负的实际状态组之间至少有一个结。统计量可能会出现偏差。

a. 在非参数假设下。

b. 零假设: 实面积 = 0.5。

图 10.43 ROC 曲线下的面积

图 10.43 最后输出的是近似 P 值 (Asymptotic Sig.), 该检验的无效假设是检测方法总体 ROC 曲线下面积是否为 0.5。

SPSS 未提供两条或多条 ROC 曲线下面积的比较, 可以近似地根据它们的 95% 可信区间是否交叉来判断各总体 ROC 曲线下的面积是否相等。这里说近似是因为假设检验计算统计量时是根据无效假设 (各总体 ROC 曲线下的面积相等) 出发的, 此时各样本 ROC 曲线下面积的标准误也相等, 而表 10.2 中输出的 95% 可信区间是根据各自的标准误进行计算的, 而不是两条 ROC 曲线下面积的合并标准误计算的 95% 可信区间, 因此只能说是近似判断。



有的读者在绘制 ROC 曲线时可能会得到面积小于 0.5 的 ROC 曲线, 这一般发生在该检测方法的检测结果值越小, 该研究对象是病人的可能性越大的情况下。SPSS 默认检测结果值越大, 研究对象是病人的可能性越大, 因此得出的结果正好相反。对于这种类型的资料, 可以通过绘制 ROC 曲线时使用“选项”子对话框中的“检验方向”框组予以纠正。

10.10.3 时间序列分析中使用的图形

SPSS 提供了强大的时间序列分析功能, 其图形工具也比较全面, 除最简单的线图等以外, 还有以下几种专用图形。

(1) 序列图: 实际上就是一种特殊的线图, 但比一般的线图有着更多适合时间序列特点的功能, 用于对时间序列的直观描述。与普通线图一样, 它也把时间坐标轴变量当成分类变量来处理, 所以在数据时间序列存在间断的情况下要小心应用。

(2) 自相关图: 做单个序列, 任意滞后 (包括负的滞后, 也就是超前) 的自相关和偏相关图。

(3) 互相关图: 交叉相关图, 做两个或两个以上的时间序列, 任意滞后的交叉相关图。

(4) 频谱图: 在进行频谱分析时给出一个或多个序列的周期图和谱密度图。

由于上述这些图形的使用和解读均与时间序列分析密切相关, 对于选择模型参数及进行模型残差分析有着重要意义, 因此它们和时间序列模型一起被统一放置在“分析”→“预测”子菜单中。这些专用工具也将和时间序列模型一起在高级教程中加以介绍, 对此感兴趣的朋友可参见高级教程的相应章节。

思考与练习

- 1. 简述本章所介绍的各个统计图的特点及适合的资料类型。
- 2. 自行练习本章所介绍的对 CCSS 案例的各种图形绘制以及编辑操作。
- 3. 自行练习复式条图、线图、面积图间的转换功能,并从图形的本质考虑为什么这些图形可以互相自由转换。
- 4. SPSS 输出 ROC 曲线下面积时输出的是近似 P 值(Asymptotic Sig.),为什么该检验的无效假设是检测方法总体 ROC 曲线下的面积是否为 0.5?
- 5. 为研究工人矽肺患病率与工龄的关系,某市疾病控制中心收集了一些资料,见题表。

题 表

工龄	甲矿			乙矿		
	检查人数	矽肺人数	患病率	检查人数	矽肺人数	患病率
<5 年	5 406	39	0.007 2	1 856	11	0.005 9
5 年 -	2 537	77	0.030 4	2 734	84	0.030 7
10 年 -	2 169	265	0.122 2	3 185	347	0.108 9
合计	10 112	381	0.037 7	7 775	442	0.056 8

对于以上资料,可以选用何种统计图进行统计描述,为什么?还可以选用其他类型的统计图吗?为什么?

第 11 章 统计实战案例集锦(二)

11.1 探索消费者信心指数随背景资料的变化规律

11.1.1 项目背景

在统计绘图等章节中已经对总信心指数等指标进行了初步描述,并从中得到如下几点线索。

(1) 总信心指数在各城市间的差异相对较小,但是在不同月份间的数值差异相对较大。

(2) 收入较高的私营业主、专业人士、企事业管理人员,以及收入虽然偏低但很稳定的公务员、教师的现状指数明显较高。除了失业人群外,其余所有职业人群的预期指数均低于现状指数。

(3) 随着年龄的上升,消费者信心指数的平均水平有缓慢的下降趋势。

但是,由于篇幅限制等原因,前面对总信心指数在不同背景资料人群间的变化情况进行的描述并不完善;另一方面,总信心指数可以被分为现状指数和预期指数,这两个分指数在不同人群之间的变化规律也是非常值得加以研究的,因此在进入随后的假设检验相关章节之前,下面将对总指数、现状指数、预期指数在月份、城市、人群背景资料间的变化规律进行描述,以便后续分析能够更加有的放矢。

11.1.2 分析思路

有了前面对 CCSS 案例的反复使用,本分析项目看起来需要完成的工作非常明确,但动手之前也需要注意以下几点。

(1) 在数据分析中,任何已经获取的数据信息都应当被加以有效利用,以避免后续分析误入歧途。在前面的分析中已经发现总信心指数在月份、城市间存在差异,而且这两个分类指标显然是整个指数体系所监测的重点变量:城市间的指数差异用于揭示不同地区受访者对宏观经济感受和预期有何不同,而月份间的变化更是整个指数体系监测的重点。因此后续分析应当重点关注现状指数、预期指数等是否也存在差异,甚至于人口背景资料的影响和月份、城市变量之间是否存在交互。

(2) 图形和统计表都可以用于数据描述,图形可以提供更为直观的信息,但操作较为繁复;统计表阅读起来稍显枯燥,但操作起来比较容易,因此两者应当搭配使用。根据笔者的经验,可以考虑初步分析用统计表,发现一些线索之后可以换用统计图来直观刻画。

(3) 虽然数据描述往往以单一影响因素的分析为主,但也需要考虑不同影响因素的作用可能是重叠的,例如在年龄和婚姻状况之间,以及年龄和家庭收入之间就可能存在关联,有些变量的作用更可能存在交互。这些问题在分析中都应当加以探索并注意。



注意此处分析中使用的 SPSS 过程将不简单限于前面各章节曾经介绍过的内容,而是将根据需求灵活使用,因为最终解决问题才是最重要的,工具和方法,就是工具和方法,没必要对它们的使用过于拘泥。

11.1.3 具体操作

由于第 10 章中已经对职业的影响做了比较深入的分析,因此这里不再重复,重点放在其他因素的作用分析上。

1. 对月份、城市的影响进行分析

首先可以使用均值过程对各指数在月份和城市间的变化进行简单描述。

- (1) 选择“分析”→“比较均值”→“均值”菜单项。
- (2) 在打开的对话框中将“总指数”、“现状指数”、“预期指数”选入“因变量”列表框中。
- (3) 将月份、城市选入“自变量”列表框中。
- (4) 单击“选项”按钮,打开“选项”子对话框,在单元格统计量框组中只保留“均值”选项。
- (5) 单击“确定”按钮。

图 11.1、图 11.2 给出了如下信息。

- (1) 总指数、现状指数和预期指数在 2007 年 4 月—2008 年 12 月之间均呈现下降趋势,然后在 2009 年 12 月出现反弹。仔细观察可以发现,现状指数实际上在 2007 年 12 月就跌至和 2008 年 12 月相近的低点,也就是说其下跌要明显早于预期指数。
- (2) 总体而言北京的总指数要高于上海、广州,从分指标可以看出,主要是预期指数明显高于后两者,其现状指数则优势不明显。

均值			
月份	总指数	现状指数	预期指数
200704	98.3363	100.5810	97.0991
200712	94.1391	95.1309	93.5913
200812	90.4393	94.4069	88.2546
200912	101.9962	108.8478	98.2247
总计	95.8935	99.2227	94.0598

图 11.1 总指数 现状指数 预期指数 * 月份

均值			
S0. 城市	总指数	现状指数	预期指数
100 北京	97.5920	100.3796	96.0563
200 上海	94.6766	98.1589	92.7587
300 广州	95.4456	99.1556	93.4023
总计	95.8935	99.2227	94.0598

图 11.2 总指数 现状指数 预期指数 * S0. 城市

上面的均值过程重点显示出了均数的变化趋势,对于上述信息,完全可以继续用探索过程加

以深入刻画,操作步骤如下。

- (1) 选择“分析”→“描述统计”→“探索”菜单项。
- (2) 在打开的对话框中将总指数、现状指数、预期指数选入“因变量”列表框中。
- (3) 将月份、城市选入“因子”列表框中。
- (4) 单击“确定”按钮。

图 11.3 给出了现状指数和预期指数随月份变化的箱图,从中可以很清楚地看出,现状指数在 2007 年 12 月就已经下跌,其中位数实际上已经和 2008 年 12 月时相同。另外一个有趣的信息是:2009 年 12 月的现状指数的离散度似乎明显高于另外 3 个月,而在预期指数中,该情况则发生在 2008 年 12 月。该数据似乎暗示在 2008 年 12 月次贷危机第一波爆发,政府的经济刺激政策发布之后,消费者对未来宏观经济走势的判断发生了较大分歧;而在 2009 年 12 月,经济刺激政策实施一年之后,不同消费者所感受到的宏观经济走势也出现了较大分歧。一个有趣的问题是原先悲观的人一年之后是否仍然悲观。但由于本研究没有进行 100% 的回访,因此只进行简单的数据描述是无法回答上述研究假设的。

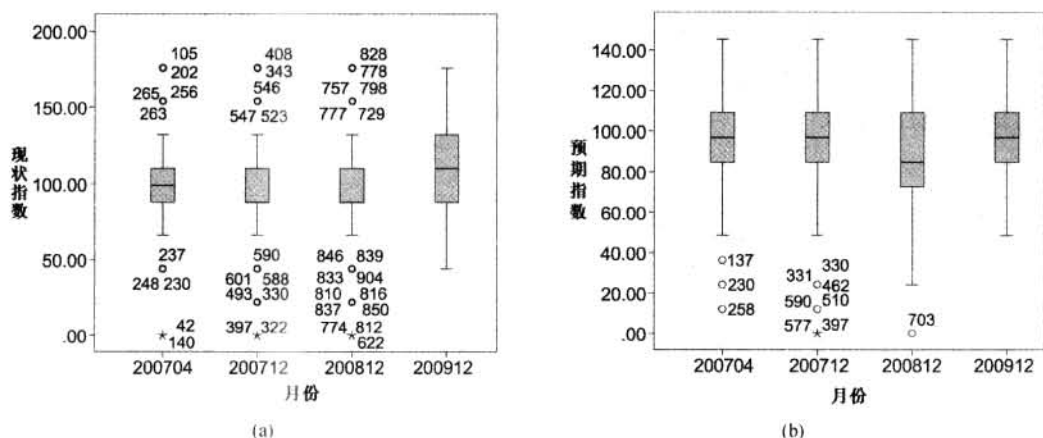


图 11.3 现状指数和预期指数随月份的变化趋势

图 11.4 则进一步给出了两个分指数在不同城市间的变化趋势,可以看出北京的现状指数平均而言的确是高于上海和广州的,而在预期指数方面,每个城市都会有个别极端悲观者。

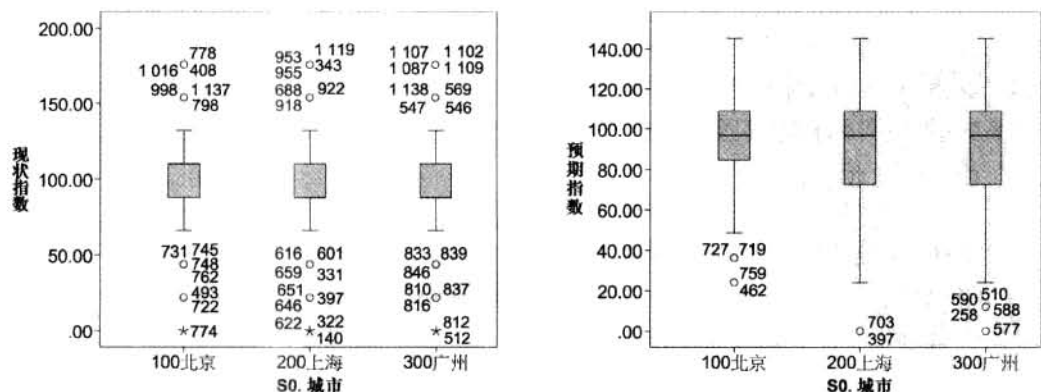


图 11.4 现状指数和预期指数在不同城市间的变化趋势

2. 对性别和职业的影响进行分析

下面对性别和职业的影响进行分析,这里仍然首先采用均数过程加以描述,因操作与前面基本相同,此处省略相应的说明。

图 11.5 暗示不同性别的受访者其信心指数似乎没有太大差异,该结果应当符合整个项目的设计,毕竟所询问的是整个家庭的情况,从逻辑上讲不同性别的受访者其感受是不应当有太大差异的。

均值			
S2. 性别	总指数	现状指数	预期指数
男	95. 9719	99. 8869	93. 8160
女	95. 7956	98. 3931	94. 3643
总计	95. 8935	99. 2227	94. 0598

图 11.5 总指数 现状指数 预期指数 * S2. 性别

图 11.6 显示出在本科学历之前,随着学历的上升,受访者的总信心指数是上升的,随后在硕士级别开始下降。分指标的描述结果表明,实际上现状指数基本上是随着学历的上升而上升的,但预期指数则在硕士及以上级别出现明显下跌。

均值			
S4. 学历	总指数	现状指数	预期指数
初中/技校或以下	93. 6773	94. 6100	93. 1620
高中/中专	94. 6264	97. 6692	92. 9502
大专	96. 6305	101. 6669	93. 8576
本科	97. 6066	100. 1710	96. 1936
硕士或以上	95. 7835	101. 1641	92. 8213
总计	95. 8935	99. 2227	94. 0598

图 11.6 总指数 现状指数 预期指数 * S4. 学历

可以进一步考虑将性别和学历进行联合描述,这里考虑使用多重线图来完成此任务,操作步骤如下。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 选择“线”图组,将多重线图拖入画布中。
- (3) 将“学历”拖入横轴框中。
- (4) 同时选中“总指数”、“现状指数”、“预期指数”,将其拖入纵轴框中。
- (5) 选择“组/点 ID”选项卡,选中列嵌板变量。
- (6) 将“性别”拖入画布中新增的“列嵌板变量”框中。
- (7) 单击“确定”按钮。

从图 11.7 中可以发现,虽然总体而言不同性别受访者的信心水平没有明显差异,但是当性别和学历交叉之后,性别×学历对信心指数的反映并不相同,大致有如下趋势。

(1) 男性受访者,其信心指数在专科学历时最高,学历较高或者较低时信心均下降。

(2) 女性受访者,其总信心指数大致呈随着学历上升而上升的趋势,但是如果考察分指数,就会发现大专-本科是一个明显的分水岭,本科及以上学历的受访者其现状、预期信心均明显高于大专及以下受访者。

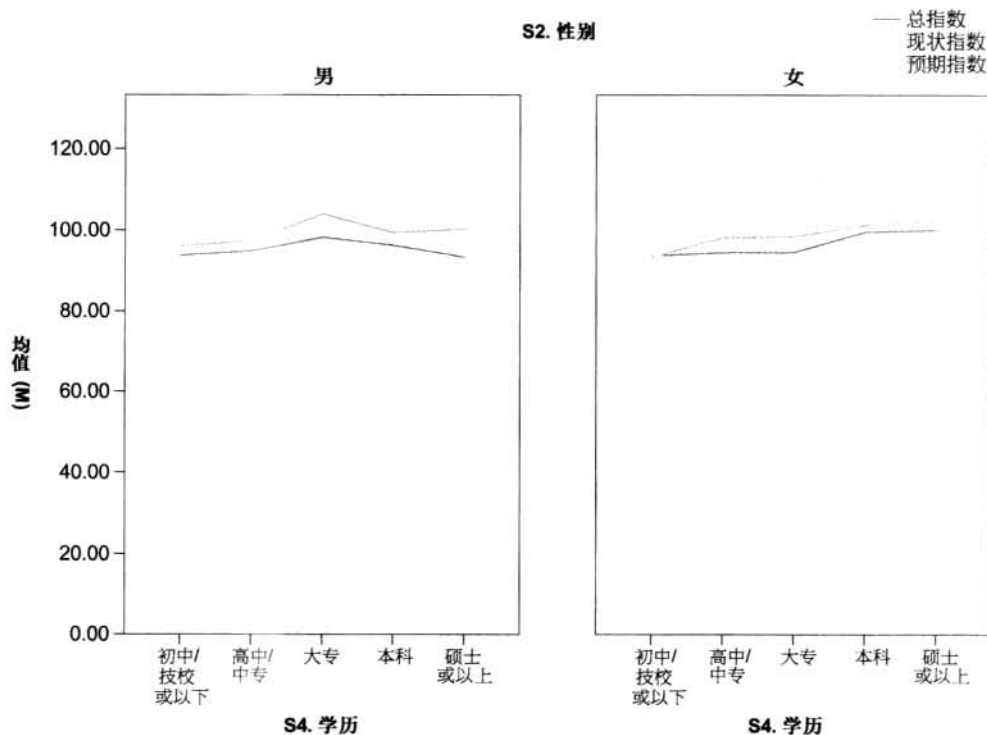


图 11.7 将学历、性别联合对信心指数进行描述的多重线图

对于这里所发现的这些特征暂时还不能下任何结论,因为这些趋势有可能混杂着收入、婚姻状况、年龄等因素的作用,因此目前的分析任务仅仅是发现可能的数据趋势并加以记录,而不是立刻去寻找合理的解释。

3. 对婚姻状况的影响进行分析

此处虽然婚姻状况被分为未婚、已婚和离异/分居/丧偶 3 类,但第 3 类只有 14 例,因此重点仍然放在未婚人群和已婚人群的比较上。另一方面,考虑到男性和女性可能在婚姻方面的看法、婚龄有所差异,或许会影响到对宏观经济的看法,因此同时按性别分组进行描述,这里采用制表过程来进行输出。

(1) 选择“分析”→“表”→“设定表”菜单项。

(2) 在打开的对话框中同时选中“总指数”、“现状指数”、“预期指数”,将其选入“行变量”列表框中。

(3) 保持对行变量的选中状态,进入“摘要统计量”子对话框,在显示列表中增加计数,单击“应用选择”按钮。

(4) 将“性别”选入“列变量”列表框中,然后将婚姻状况以嵌套在性别下方的方式选入“列

变量”列表框中。

(5) 单击“确定”按钮。

从结果图 11.8 中可以看出,无论男女,已婚人群的总信心、现状信心和预期信心值都低于未婚人群,但总体而言男性受访者在已婚、未婚人群上的信心值差异要更大一些。当然,这种差异究竟反映的是婚姻状况的影响,还是反映的年龄甚或学历等因素的影响目前还难以下结论。

	S2. 性别											
	男						女					
	S7. 婚姻状况						S7. 婚姻状况					
	已婚		未婚		离异/分居/丧偶		已婚		未婚		离异/分居/丧偶	
	均值	计数	均值	计数	均值	计数	均值	计数	均值	计数	均值	计数
总指数	94.70	426	98.67	204	94.84	7	95.42	364	97.72	139	76.99	7
现状指数	98.47	426	103.03	204	94.32	7	98.31	364	99.75	139	75.46	7
预期指数	92.62	426	96.26	204	95.13	7	93.83	364	96.60	139	77.83	7

图 11.8 性别×婚姻状况的信心指数描述表格

从信心指数的表现上看,似乎不结婚会更幸福一些,确实有学者做过此类研究:苏联的拉里科夫跟踪研究了 15 000 名调查对象,按照结婚动机:

- (1) 70%~80% 回答说是因为爱情而结婚,这些受访者 100% 觉得不幸福。
- (2) 有 3%~10% 的受访者是因为个人利益而结婚的,结果有 70% 觉得不幸福。
- (3) 15%~20% 是因为人人结婚才结婚,他们有 55% 觉得不幸福。

从这个研究中可以得出结论如下:如果希望婚姻幸福,只要这个婚姻不是基于爱情,那就还有指望。

4. 对年龄的影响进行分析

由于年龄也是连续性变量,因此最适合进行年龄分析的工具应当是散点图,由于在第 10 章中已做过初步分析,因此重点可以放在给出回归趋势线上。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“散点”图组,将右侧出现的简单散点图图标拖入画布中。
- (3) 将年龄 S3 拖入横轴框中,同时选中“总指数”、“现状指数”和“预期指数”,将其拖入纵轴框中。
- (4) 单击“确定”按钮绘制出图形,然后双击图形进入编辑状态。
- (5) 在图形上右击,在弹出的快捷菜单中选择“添加总计拟合线”菜单项。
- (6) 在打开的对话框中保持对总计拟合线的选中状态,在“拟合线”选项卡中更改拟合方法为 Loess,单击“应用”按钮。
- (7) 对横轴、纵轴和图形宽度进行编辑,以使得趋势线更为清晰。

最终得到的趋势线(如图 11.9 所示)非常清晰地显示出,随着年龄的上升,总指数、现状指数和预期指数基本上呈持续下降趋势,且 3 个指标的走势接近。唯一出现波动的是 25~35 岁区间段,在该区间内似乎指数值保持了一定程度的稳定。

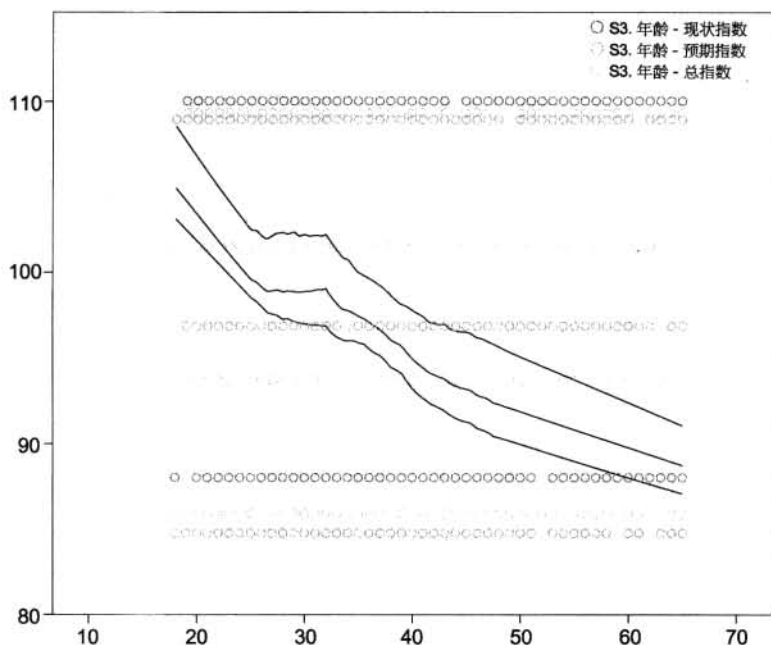


图 11.9 年龄和信心指数间的回归趋势线

5. 对家庭收入的影响进行分析

最后再对家庭收入的影响进行分析,虽然家庭收入的选项分档较细,但仍然没有细化到可以绘制散点图的地步,因此这里将绘制条图以进行考察。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
 - (2) 在图库中选择“条”图组,将右侧出现的简单条图图标拖入画布中。
 - (3) 同时选中总指数、现状指数和预期指数,将其拖入纵轴框中,对弹出的创建摘要组对话框进行确认。
 - (4) 在“组/点 ID”选项卡中,选中“列嵌板变量”复选框,将画布上横轴框中的“INDEX”拖入列嵌板框中。
 - (5) 将“家庭收入”S9 拖入横轴框中。
 - (6) 在“元素属性”对话框中的“编辑属性”列表框中选中“条”选项,然后选中下方的“显示误差条形图”复选框。
 - (7) 单击“确定”按钮绘制出图形,然后双击图形进入编辑状态。
 - (8) 选中横轴标签,在“标签和刻度标记”选项卡中将其标签显示方向由“自动”更改为“水平”,单击“应用”按钮。
 - (9) 编辑纵轴的数字格式和刻度范围,使图形更易于观察。
- 最终得到的条图如图 11.10 所示,在图中可以观察到如下信息:
- (1) 当家庭月收入在 2 000 元以下时信心指数很低,随着收入的上升,信心指数呈快速上升趋势。
 - (2) 现状指数大致始终呈随着家庭收入上升而上升的趋势,月收入 30 000 元以上的样本因为只有 31 例,因此该群体现状指数的下降趋势尚不足以被确认。

(3) 和现状指数的上升趋势不同,当家庭收入达到 2 000 元以上之后,预期指数就基本上不再和家庭收入的上升有什么关联了。

(4) 随着家庭收入的上升,总指数、现状指数和预期指数的离散程度都有所增大。

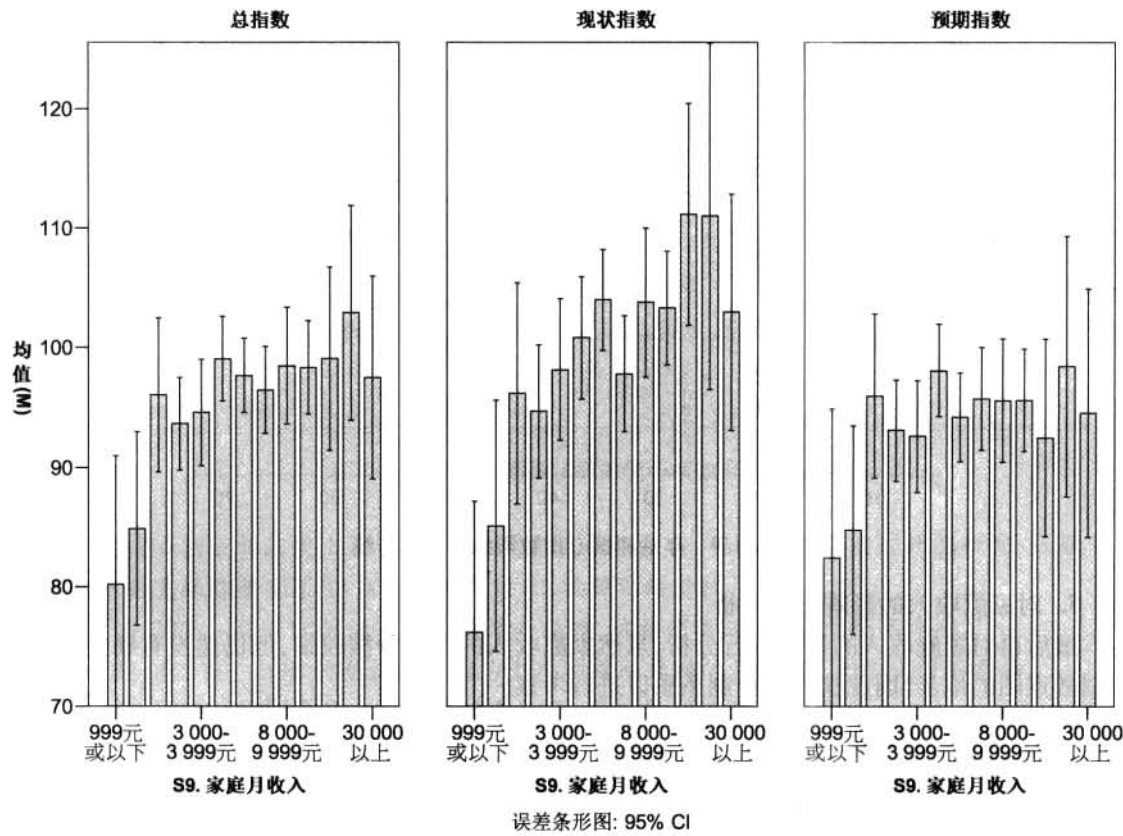


图 11.10 信心指数随家庭收入变化的条图

综上所述,当前家庭收入的高低会比较明显地影响现状的感受,但对未来的预期则影响较弱,这究竟反映的是真实的关联,还是其他背景变量所导致的假象,目前还难以做出判断,需要做进一步的深入分析方能得出结论。

11.1.4 项目总结与讨论

在对各种背景变量可能产生的影响做了深入的探讨后,得到的数据信息总结如下。

- (1) 时间的确对指数有影响,总指数和两个分项指数在 2007 年 4 月—2008 年 12 月之间均呈现下降趋势,然后在 2009 年 12 月出现反弹。但是现状指数的下跌似乎要早于预期指数。
- (2) 总体而言,北京的总指数要高于上海、广州,从分指标可以看出,主要是预期指数明显高于后两者,其现状指数则优势不明显。
- (3) 男性和女性的信心指数相差不大。
- (4) 男性受访者的信心指数在专科学历时最高,学历较高或者较低时信心均下降;而女性受访者其总信心指数大致呈随着学历上升而上升的趋势,但是在分指数方面,本科及以上学历的受

访者其现状、预期信心均明显高于大专及以下受访者。

(5) 已婚人群的总信心、现状信心和预期信心值都低于未婚人群。

(6) 随着年龄的上升,总指数、现状指数和预期指数基本上呈持续下降趋势。

(7) 当家庭月收入在 2 000 元以下时信心指数很低,随着收入的上升,信心指数呈快速上升趋势;如果考察分指数,则现状指数大致始终呈随着家庭收入上升而上升的趋势,但当家庭收入达到 2 000 元以上之后,预期指数和家庭收入的上升就基本上没有什么关联了。

上述这些信息将会被用于指导后续的假设检验和多变量建模,以保证进行复杂建模时分析方向的正确性。

在上面的分析中可以发现,笔者并未使用某一个或几个常用的过程,而是尽可能地展示了多种分析过程的结果,以期读者能够对这些过程的功能有所了解。事实上,在实际的分析工作中,最常见的操作仍然是首先使用“描述”子菜单中的几个常用过程如频率、描述、探索、交叉表等进行分析,这些过程虽然输出比较简单,但操作方便。在得到了一定的趋势性信息之后,分析者就可以考虑换用更为准确和复杂的过程,如制表过程,或者统计图,来对发现的数据趋势进行重点刻画。而最终如果需要将这些信息反映到分析报告中去,则上述制表过程、绘图过程所给出的就是所需的分析结果了。

11.2 CCSS 项目分析报告的自动化生产

11.2.1 项目背景

在第 6 章中已经介绍了如何对 CCSS 项目的原始数据进行清理和变量计算,也得到了分析所需的中间变量。但是这仅仅意味着数据分析和报告撰写所需的数据原料已经备好,后续的工作应当是对其进行分析,并将结果呈现为所需的数据报告或分析报告。

在 CCSS 项目中,相应的数据报告和分析报告需求如下。

(1) 每月最后一个工作日,向付费用户提交当月 CCSS 项目的分析结果。

(2) 数据报告以 Excel 制作,最终封装为 PDF 格式,主要提供比较详细的数据汇总表格。

(3) 分析报告以 PowerPoint 制作,最终封装为 PDF 格式,提供对当月数据的深入解读。

由于上述分析和报告工作每月进行,工作内容相近,且有比较严格的时间要求,因此如何尽量减少人工干预,提高工作的自动化程度就显得非常重要,这里就来探讨一下如何充分利用 SPSS 的相关功能来达到此目的。

11.2.2 分析思路

简单地讲,这里需要完成的工作就是将整个工作流程分解为若干个环节,然后将其中可被计算机自动/半自动执行的环节尽量自动化,以同时满足提高工作效率、减少出错概率两个目的。具体而言,项目的分析思路如下。

1. 工作流程的分解

首先应当对工作流程进行分解,以便分别研究对策。这里涉及的流程可以分解为如下步骤。

(1) 利用 SPSS 进行数据分析,计算出报告所需的汇总统计量/汇总结果、进行所需的假设检

验、绘制需要的统计图。

(2) 在 Excel 中将计算出的统计量/汇总结果按标准输出格式的要求进行表格呈现。

(3) 利用得到的统计表、统计图,在 PPT 中完成分析报告的撰写工作,并对 PPT 中相应的统计图进行更新。

(4) 将 XLS 文件和 PPT 文件转换成 PDF,完成报告封装。

2. 分析需求的整理

可见在上面的4步中,最后一步和 SPSS 完全无关;第1步完全是 SPSS 自己的事情;第2、3步都会利用到 SPSS 的结果输出,因此相应的分析需求可以归纳为如下几点。

(1) 在第1步尽量做到程序化、自动化。

(2) 在第2、3步中,重点研究 SPSS 给出的输出结果能否被直接应用/转化为所需的表格/图形,如果不能直接使用,则尽量减少转化所需的人工干预工作量。



实际上,CCSS 项目报告产生时的绝大部分工作都已经实现自动化,但其中涉及的工具不仅限于 SPSS,本书因为是 SPSS 教程,只涉及了利用 SPSS 所完成的部分。

3. 输出格式需求的整理

在理清分析需求之后马上会发现,第1步的程序化、自动化是比较容易的,毕竟每个月所需要做的分析非常相似,完全可以考虑将程序代码固定。但是在第2、3步中,就必须考虑到其格式是否能满足最终报告的需求。在仔细分析之后,发现大致有如下几种情况。

(1) 统计报表:可以基本满足报告需求。如同第9章中所讲到的,诸如 A3、A3a 等题目在数据报告中所需的格式,完全可以在 SPSS 中利用制表过程实现,只是个别细节需要进行调整。

(2) 统计表:只能计算出所需的指标,难以直接得到设定格式的表格输出。此类表格数量较少,但确实存在,最典型的情况是用于显示统计量和检验结果的表格,SPSS 几乎没有过程可以将统计量和 P 值直接进行设定格式的输出。

(3) 统计图:虽然 SPSS 可以实现全部的图形需求,但由于 PPT 下 MS Office 风格的图形更容易为客户所接受,因此基本上不能考虑直接使用 SPSS 制作的图形。

4. 需求实现方式的确定

根据上面整理出来的输出格式需求,最终确定的实现方式如下。

(1) 可基本满足需求的统计表:在 SPSS 中进行表格制作,然后将结果导出为 .xls 格式完成剩余的编辑工作。

(2) 难以直接满足需求的统计表/统计图:用 SPSS 计算出表格/图形中所需的汇总数据,然后再考虑使用 SPSS 以外的工具完成剩余的工作。

5. 业务流程的技术实现

在需求实现方式确定之后,剩下的只是相应方式的技术实现问题了。对于同一个需求,往往有多种可能的技术实现方式,而确定最终采用哪种技术路线的原则是:简单、高效、低故障率。在本项目中,最终选择的是如下方式。

(1) 可基本满足需求的统计表:由于相应的格式比较规范,大致可以被归纳为 3~4 种表格模板,因此直接在 SPSS 中进行模板设定,然后采用宏代码方式编写程序。

(2) 难以直接满足需求的统计表/统计图:利用 SPSS 的 OMS 系统,将表格/图形中所需的数

据输出为 .sav 数据文件,然后根据表格要求,或者进一步采用 SPSS 绘制出表格雏形,或者直接导出至 .xls 文件中以便在 Excel 中完成剩余的制表工作。

11.2.3 具体操作

下面就来具体说明业务流程的具体技术实现方式。

1. 表格模板设计

从第 9 章的案例讲解可以知道,如 A3、A3a 这样的表格框架实际上就是最终产品表格中被广泛使用的格式,因此可以考虑将其制作成标准模板加以应用,包括对其字体、行高、网格线等都按照最终产品所需的格式来精心设定,以尽可能地减少出表后的人工编辑操作。

因篇幅所限,这里不再详细讲解具体的模板设定操作,具体的技术细节可以参考 9.5 节中的相应内容。

2. 基本制表程序框架提取

下面考虑将表格框架所对应的代码提取出来,并改写为宏代码,以大大简化编程工作。这里以 A3 对应的表格为例来加以说明。按照第 9 章中例 9.1 的操作,绘制 A3 表格所对应的 SPSS 代码如下:

```
CTABLES
  /VLABELS VARIABLES = a3 Qa3 time DISPLAY = NONE
  /TABLEa3 [C][ROWPCT. COUNT "F40.1"] + Qa3 [S][MEAN '感受值'F40.1]
  BY time [C]
  /SLABELS POSITION = ROW
  /CATEGORIES VARIABLES = a3 ORDER = A KEY = VALUE EMPTY = INCLUDE
  /CATEGORIES VARIABLES = time ORDER = A KEY = VALUE EMPTY = EXCLUDE.
```

3. 制表宏代码编写

考虑到 A4、A5 等题目实际上都使用相同的表格框架,因此只需要将代码中与具体题目有关的部分替换为宏变量即可,此处为 A3、QA3(为了便于阅读,代码中已加粗显示)。相应的宏变量替换方式有以下两种。

(1) 将 A3、QA3 分别用两个宏变量替换,调用宏程序时则需要同时指定好这两个宏变量。这种方式稍显麻烦,但易于理解。

(2) 利用宏函数,只指定一个宏变量,然后用宏函数生成所需的另一个宏变量。

这里介绍更为简洁的后者,相应的宏代码如下:

```
* 表框架一:频数 + 均数的组合输出,Q 变量用于计算均数.
DEFINE M_Tb01 ( invar1 = !charend('/') ) .
```

```
CTABLES
  /VLABELS VARIABLES = !invar1 !concat("Q",!invar1) time DISPLAY = NONE
  /TABLE !invar1 [C][ROWPCT. COUNT "F40.1"]
```

```

+ !concat("Q", !invar1)[S][MEAN '感受值'F40.1] BY time [C]
/SLABELS POSITION = ROW
/CATEGORIES VARIABLES = !invar1 ORDER = A
KEY = VALUE EMPTY = INCLUDE
/CATEGORIES VARIABLES = time ORDER = A KEY = VALUE EMPTY = EXCLUDE.

!ENDDEFINE.

```

```
M_tb01 invar1 = a3.
```

其中的!concat()就是用于合并字符串的宏变量,利用!concat,只需要指定一个宏变量,就可以生成所需的另外一个宏变量。而最后一句调用 M_tb01 宏所得到的分析结果完全等同于原先的程序段。



编程爱好者或许会想到:这里的宏变量指定显然还可以写得更简洁,例如只需将 invar 指定为数字3,然后将 A3 用!concat("A", !invar1)来加以实现,QA3 用!concat("QA", !invar1)来加以实现,这样代码不是更漂亮吗?

很遗憾,有这种想法,并且希望能够将其付诸实践的读者恐怕已经中了计算机编程病毒,而且目前没有杀毒软件可用。静下心来考虑,如果这样写:

(1) 代码更简单了吗? 显然可读性随着函数用量的增多而变差了。

(2) 代码执行效率提高了吗? 多了一个!concat()函数,无论如何程序的执行效率都降低了。

(3) 代码的使用范围扩展了吗? 只要是输出相同的表格框架,原先的代码无论是 A3 还是 C3 都可以使用,而新的代码只能用于 A 开头的变量,适用范围明显缩小了。

简言之,这样将得到一个更难阅读、速度更慢、适用范围更窄的代码,唯一的报答是获得了一点将简单问题复杂化的快感。对于不需要发表学术论文/获取学位,而是需要解决实际问题的读者,这里要慎重提醒一句:简单、实用的方法才是最好的方法。

4. 结果表格的导出

对于上述生成的可以直接用于最终报告产品的结果表格,可以考虑直接将其导出为 .xls 格式使用。

```

SAVE TRANSLATE OUTFILE = 'D:\OutTbl1.xls'
/TYPE = XLS
/VERSION = 8
/MAP
/REPLACE
/FIELDNAMES
/CELLS = VALUES.

```




在导出之前,如果能够使表格生成顺序与报告产品中的顺序相一致,则可以大大简化后续操作。对 Excel 中的 VB 宏比较熟悉的读者更可以利用 VB 宏代码来尽量做到自动化。至少在 CCSS 项目中,这些工作都是不需要人工干预的。

5. OMS 代码编写

现在来考虑如何处理无法直接生成的统计表/统计图。如前所述,考虑利用 OMS 系统来输出相应的汇总数据。这部分工作在很多方面都类似于上面详细讲解的内容,因此不再展开讨论,这里只指出几个关键点。

(1) OMS 系统对于不同格式的数据框架无法做到同时输出,必须分别加以指定,因此必须先对数据需求加以整理,将其统一为少数几个数据模板。

(2) 数据模板的设定必须要考虑到后续步骤的制表/绘图需求,如果输出的数据格式还要求随后进行大量的手工编辑工作才能用于制表/制图,则相应的格式显然还有很大的改进空间。

(3) 在 OMS 系统中同样可以利用宏代码,特别是将需求统一为几个数据模板之后,就可以针对每种模板需求进行宏代码的编写工作了。

下面是 CCSS 项目中所用到的一个 OMS 程序实例:

* OMS 例程开始,此处的 D:\temp. sav 是临时文件,对 OMS 熟悉的用户可以不生成临时文件,而是将信息存储在内存工作区中。

OMS

/SELECT TABLES

/IF COMMANDS = ['CTables '] SUBTYPES = ['Custom Table ']

/DESTINATION FORMAT = SAV NUMBERED = TableID

VIEWER = yes OUTFILE = 'D:\temp. sav '.

* 相应的汇总数据指标宏代码段,此处省略。

* OMS 例程结束。

OMSEND.

* 读取临时文件,准备导出。

GET

FILE = 'D:\temp. sav '.

* 指定变量 var5 的输出格式,以方便随后的工作。

Formats var5 (f5.1).

* 将数据导出为 . xls 格式文件。

SAVE TRANSLATE OUTFILE = 'D:\InSheet. xls '

/TYPE = XLS

/VERSION = 8

/MAP

/REPLACE

/FIELDNAMES

/CELLS = VALUES.

对于上述 OMS 代码所涉及的详细知识,可参见 5.4 节的相应内容,这里不再详述。

11.2.4 项目总结与讨论

在 6.2 节的案例中介绍过如何利用代码方式来完成 CCSS 项目数据自动计算的任务。本案例中也使用了大量的宏代码等编程操作,看上去和 6.2 节的案例非常相似,但并不是完全相同的。除了用到制表代码、OMS 系统等新增的内容之外,本案例非常引人注目的一点是:并非所有的工作都是在 SPSS 系统内部完成的,而是涉及了和 MS Office 软件交互的问题,在多个软件平台的协作下才能满足项目的全部业务需求。

对于绝大多数真实世界中的业务项目而言,仅依靠一种软件平台就能够 100% 满足业务需求几乎是不可能出现的情况,多平台/多系统协作是普遍的现象(当然,那种在 500 强企业中被强制使用公司内部业务系统来完成工作的情况例外)。因此,在公司政策/版权许可的情况下,充分地发掘可用工具的功能,并对其进行优化搭配是业务人员所需要思考的问题。例如在笔者近期做的一个项目中,整个业务流程中共使用过两种统计软件,一种数据挖掘软件——并非不能只用其中一个完成,事实上三者均可独立完成相应的工作,但笔者考虑的是利用这 3 个软件各自的优势,以形成一个最为便捷的业务流程。只有紧紧抓住业务的核心需求,不局限于个人的使用习惯与好恶,去充分发掘可用的资源,才能最大程度地去切合需求,并最终得到客观真实的分析结果。

思考与练习

自行练习本章中涉及的案例数据操作。

第三部分

常用假设检验方法

第12章 分布类型的检验

本章将涉及统计学分析中最为主要的理论之一:假设检验,它是分析统计数据、构建统计模型进行决策支持的基石。本章将首先介绍假设检验的相关思想、理论基础、分析步骤等,然后介绍几个比较重要的分布类型检验——正态分布检验、二项分布检验以及游程检验在 SPSS 中的实现方法,并借此使读者进一步了解假设检验基本思想的具体应用。

12.1 假设检验的基本思想

12.1.1 问题的提出

下面以一个假设的场景来引出后面的讨论:为了纪念葡式蛋挞诞生 $x \times$ 周年(当然只是个借口),你决定参加港澳游,并顺便去澳门的赌场试试运气。具体的博彩方式为最简单的掷单颗骰子,猜到点数为胜。那么如果这时你参加下注,会下多少注,结果又会怎么样呢?相信大家在下注之前都相信,对于每个人来讲,在掷骰子时 6 个点都是以同等机会出现的,关键就在于谁的运气好,所以一般都是随机选择一个点进行投注。其实,在做出下注决策的时候就已经做了相应的假设:假设这个骰子是均匀的,因此每个点出现的几率是相等的,可以随机地选择点数进行下注。其实大家都知道,如果反复下注,大概平均每下 6 次注会赢一次。当然,这只是平均的情形,如果每次都这样,也就不会有人去玩这种游戏了。每次博彩时猜中的比例可能会多一些,也可能少一些。参与者都是冲着可能出现的高猜中率来的,这也算是人性的弱点吧。但是如果把多次参与的猜中率进行平均,则仍然应当在 $1/6$ 左右。

现在来讨论一种不太走运的情形,假设今天一共下了 600 次注,由于假设这颗骰子是均匀的,因此平均应当赢大约 100 次。但是最终竟然一共只猜中了一次(别激动,仅仅是假设)! 这有两种解释:①皇历上显示,今天不宜博彩,运气实在太差;②骰子有鬼,掷骰子的人可以人为控制结局,使得每种点数出现的概率不均匀,从而利用这种能力使自己得到了更多的收益。虽然第一种解释是可能的,但是理论上的 100 次胜利和实际的仅仅 1 次胜利实在相差太远了,这种解释很难让人接受,因此,大多数赌徒都会立刻选择第二种解释,认为骰子均匀的假设实际上不成立,这一切根本就是一个骗局。

事实上,上面的讨论所展示出来的整个思路就是一个标准的假设检验流程,如果将上述过程用标准的统计学流程复述一遍,则会表述为下列内容。

- (1) 建立假设, H_0 : 骰子均匀, $\pi = 1/6$; H_1 : 骰子不均匀。
- (2) 确定检验水准, 一类错误 $\alpha = 0.05$ 。
- (3) 下赌场, 在假设 H_0 成立的前提下亲力亲为, 进行样本量为 600 次的掷骰子试验, 得到 $1/600$ 的样本率。
- (4) 基于样本数据计算 P 值, 发现如果 H_0 成立, 得到现有样本率(以及更极端情况)的可能

性微乎其微,远远小于可以容忍的一类错误 $\alpha = 0.05$ 。

(5) 得出推断结论,由于基于 H_0 出现了小概率事件,因此认为 H_0 假设不成立。

显然,上面这种描述方式会使得赌博听上去令人索然无味,但却是继续下面的内容所必须了解的,下面就将详细介绍假设检验的基本思想。

12.1.2 假设检验的标准步骤

1. 小概率事件

在讨论假设检验的基本思想之前,首先需要明确小概率事件这一概念。衡量一个事件发生与否可能性的标准是概率大小,通常概率大的事件容易发生,概率小的事件不容易发生。习惯上将发生概率很小,如 $P \leq 0.05$ 的事件称为小概率事件,表示在一次实验或观察中该事件发生的可能性很小,因此如果只进行一次试验,可以视为不会发生。

这里需要澄清一个事实:注意上面的表述是“一次试验中小概率事件不应当发生”,这并不表示小概率事件不可能发生,也就是说,这里有一个前提:只进行一次试验,结果应当不会是小概率事件。如果进行多次(可能无穷多)试验,那么小概率事件就肯定会发生,或者说,小概率事件在一次试验中不大可能发生,然而在大量试验中几乎是必然发生的。

2. 小概率反证法

假设检验的基本思想是统计学的“小概率反证法”原理:对于一个小概率事件而言,其对立面发生的可能性显然要大大高于这一小概率事件,可以认为小概率事件在一次试验中不应当发生。因此可以首先假定需要考察的假设是成立的,然后基于此进行推导,来计算一下在该假设所代表的总体中进行抽样研究得到当前样本(及更极端样本)的概率是多少。如果结果显示这是一个小概率事件,则意味着如果假设是成立的,则在一次抽样研究中竟然就发生了小概率事件!这显然违反了小概率原理,因此可以按照反证法的思路推翻所给出的假设,认为它们实际上是不成立的,这就是小概率反证法原理。

3. 假设检验的标准步骤

根据大量的实践经验,假设检验的过程一般可以被归纳为如下步骤。

(1) 建立假设:根据问题的需要提出原假设 H_0 , 以及其对立面备择假设 H_1 。前面例子中的无效假设为“骰子均匀”,而备择假设为“骰子不均匀”。

(2) 确立检验水准:即设立小概率事件的界值,称之为 α 水准,一般这一步非常简单,习惯上会使用 0.05 作为该界值。

(3) 进行试验:即得到用于统计分析的样本,以该试验的结果作为假设检验的根据。在本例中即下注 600 次。

(4) 选定检验方法,计算检验统计量:本例的问题比较简单,可以直接利用二项分布计算出相应的 P 值,因此这一步基本上是被省略掉了。

(5) 确定 P 值,给出推断结论:这里的 P 值对应的是当原假设 H_0 成立时,进行试验得到现有样本这种情况,以及比现有样本情况更极端的情形的累积概率。在本例中,这就意味着下注 600 次只赢一次,和甚至于一次也没有赢这两种情形的概率之和。由于获胜的比例太低,小概率事件 A 在一次试验中发生了,这与小概率事件实际不可能发生的原理相矛盾,从而推翻原假设 H_0 , 接受其对立面 H_1 , 认为骰子不均匀;反之,若获胜比例大约在 $1/6$ 上下,则在 H_0 成立的情况下这只

是一个很普通的非小概率事件,则找不到任何的理由来推翻原假设,因此最终的结论只能是不能拒绝无效假设,这等于什么也没说!当然,从实用的角度出发,在检验所得到的概率值非常大的时候,研究者往往会将结果引申为接受 H_0 ,但注意这仅仅是一个引申,和统计学已经无关了。

12.1.3 假设检验的两类错误

显然,在经过假设检验后,得到的结论并不可能绝对正确,存在着一定的犯错概率,那么这一概率是多大呢?为了回答这个问题,首先需要了解假设检验中的两类错误。

假设检验的依据是“小概率事件在一次试验中不会发生”这一原理,然而小概率事件并非不可能发生的事件(只是它不是经常发生的),分析者并不能完全排斥它发生的可能性,因而假设检验的结果就有可能出现错误,可以按照错误发生的不同情境将其分为两类,如表 12.1 所示。

表 12.1 推断结论和两类错误

实际情况	检验结果	
	拒绝 H_0	不拒绝 H_0
H_0 真	I 类错误 (α)	结论正确 ($1 - \alpha$)
H_0 不真	结论正确 ($1 - \beta$)	II 类错误 (β)

(1) 第一类错误:无效假设 H_0 实际上是正确的,但由于抽样误差的原因,或者说恰好发生了小概率事件的原因,使得人们错误地拒绝了它,从而犯了“弃真”的错误,统计学上称它为“第一类错误”。犯第一类错误的概率是人为指定的,就等于检验水准 α 。

(2) 第二类错误:无效假设 H_0 实际上是不正确的,但由于抽样误差的原因,检验中得到的 P 值大于检验水准,使得人们未能拒绝 H_0 ,从而犯了“存伪”的错误,统计学中称它为“第二类错误”,用字母 β 表示。和第一类错误不同,犯第二类错误的概率大小在进行假设检验时一般并不知道,但可以根据相关信息进行估计。

12.1.4 假设检验中的其他问题

还有一个需要说明的问题就是检验的方向问题,这里涉及两个概念:单侧检验(One-sided Test)以及双侧检验(Two-sided Test)。如果备择假设是以单方向形式表述的,则对零假设的检验称为单侧检验。如果研究者需要检验假设是否发生了变化,但是并不是非常清楚地了解发生变化的方向,就要用双侧检验,这也是占绝大多数的情形。

单双侧检验首先应根据专业知识来确定,同时也应考虑解决问题的目的。如果研究的背景比较明确,从专业知识判断一种方法的结果不可能低于或高于另一种方法的结果,则可以考虑使用单侧检验。但是在尚不能从专业知识给出结论方向的判断时,则最好使用双侧检验,一般认为双侧检验要更加保守和稳妥一些。

除分为单/双侧检验两类外,正如最初建立假设时所提到的,假设检验还可以被分为参数检验以及非参数检验。通常参数检验是在已经知道了相关数据的分布形式,只是不了解相应参数取值时采用的检验形式。而如果对相关数据的分布形式也并不了解,就必须先确定数据的分布形式,这样才可以进一步对分布做出更为具体的说明以及解释。本章随后的主要内容就是介绍

几种常用分布的假设检验,并借此使读者进一步掌握假设检验的基本思想。

12.2 正态分布检验

在第4章中已经给出了正态分布的定义以及特征,并且介绍了它是统计分析中最为重要的分布。因此在许多时候,研究者希望能够确认数据是服从该分布的。在SPSS中,正态分布的考察方法有:通过计算偏度系数和峰度系数加以考察;通过绘制直方图、PP图等图形工具来考察;也可以进行各种假设检验。而最常用的对正态分布进行的检验就是K-S单样本检验。

12.2.1 K-S 检验的原理

Kolmogorov-Smirnov(K-S)单样本检验(Kolmogorov-Smirnov One-Sample Test)是一种分布拟合优度的检验,其方法是将一个变量的累积分布函数与特定分布进行比较。用 A_i 表示理论(假设)分布每个类别的累积相对频数, O_i 表示样本频数的相应值,K-S检验是以 A_i 和 O_i 的绝对差异为基础的,其检验统计量为

$$K = \max |A_i - O_i|$$

显然,如果无效假设成立,则每次抽样研究中所得到的 K 值应当不会偏离0太远, K 值越大,说明基于无效假设得到当前样本的可能性就越小,就越有可能判断 H_0 为错误。当基于无效假设成立的前提得到当前样本的 K 值,以及比当前样本更大的 K 值的概率小于检验水准时,研究者就可以根据小概率反证法原理,认为一次抽样中不应当出现这样的结果,从而拒绝 H_0 ,接受 H_1 ,认为样本实际上并不服从所假设的理论分布。

以上给出的是K-S检验的基本思路,为了方便计算出各种情况下 K 值所对应的概率大小,统计软件还往往将 K 值进一步转化为 Z 值(注意此处的 Z 值不是标准正态得分):

$$Z = \sqrt{N}K$$

随后再利用Smirnov于1948年提出的相应公式来计算出相应的 P 值。因公式较繁,这里不再列出。但这种变换只是为了便于求出 P 值,并不会改变K-S检验的本质。

通常分析者可以直接应用K-S检验来对样本数据进行正态分布的检验。但是,值得推荐的第一步是对样本数据进行图形描述,图形可以给分析者一个直观的印象:该数据可能服从什么样的分布类型。

12.2.2 案例:考察信心指数分布是否服从正态分布

例 12.1 采用假设检验方法对消费者信心指数进行分布特征的检验,以便为随后的深入分析做准备。

这里考察分布的目的是对指数的分布情况做大致估计,以便后续的分析方法能更有针对性。考虑到不同月份的指数分布可能会存在差异,这里只选取2007年4月的数据进行分析。这里要检验的假设如下。

H_0 :2007年4月的指数样本来自于一个正态分布的总体,理论分布与实际数据间的差异完全是抽样误差造成的。

H_1 : 样本并非来自一个正态总体, 理论分布与实际数据间的差异除了由抽样误差造成外, 确实也反映了这种偏差。

下面将利用小概率反证法原理来推断出上述两种假设中哪一种成立的可能性更大。

1. 界面说明

SPSS 为上述分布检验提供了多种实现方式, 既可以使用第 7 章中介绍过的探索过程在绘制正态图的同时进行检验, 也可以使用“非参数检验”菜单中的相应菜单项来实现, 而后者同时提供了新老两种对话框, 这里首先介绍较新对话框的实现方式。

选择“分析”→“非参数检验”→“单样本”菜单项, 就会打开“单样本非参数检验”对话框, 如图 12.1、图 12.2 所示, 该对话框中的内容非常容易理解, 下面进行简要介绍。

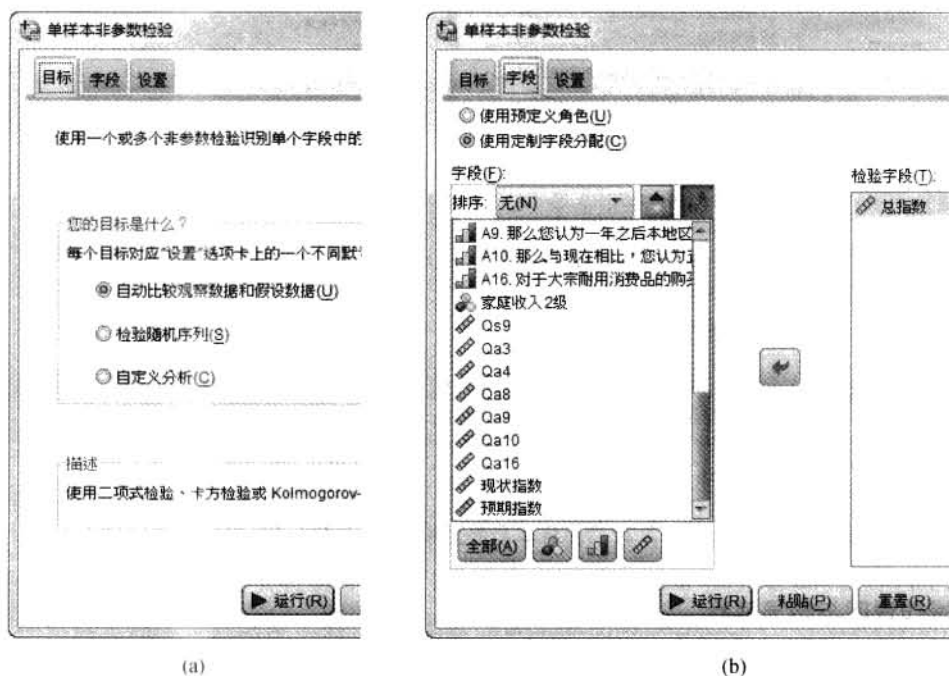


图 12.1 “单样本非参数检验”对话框(一)

(1) “目标”选项卡: 包括“自动比较观察数据和假设数据”、“检验随机序列”和“自定义分析”3 种。此处更改选择实际上会采用不同的分析方法, 本例就是默认的第 1 种, 因此不需要修改。

(2) “字段”选项卡: 指定需要分析的变量, 如果按照预定义角色分配, 则软件会默认将全部可供分析的变量纳入。虽然把所有变量都遍历一遍的想法的确更好, 但出于时间考虑, 建议本例只选入需要的 index1。

(3) “设置”选项卡: 这里默认会按照所选变量的测量尺度自动选择检验方法, 比如对二分类变量自动进行二项分布检验、多分类变量自动进行卡方检验、连续性变量自动进行正态分布检验等。但由于很多用户都没有事先正确设定变量测量尺度的习惯, 因此最佳操作方式还是“自

定义检验”。本例随后将选择进行分布的 K-S 检验,并且在相应的选项中选择进行正态分布检验。另外初学者不要忽略,该选项卡中还有“检验选项”和“用户缺失值”两个选项,可以对 Alpha 水准等做进一步的设定。

虽然这里给出的例子是关于正态分布的检验,但从对话框中就可以看出,实际上 K-S 检验还可以检验数据是否服从均匀分布、泊松分布以及指数分布。只是从分布的使用广泛程度而言,正态分布的检验显然是最为常用的。

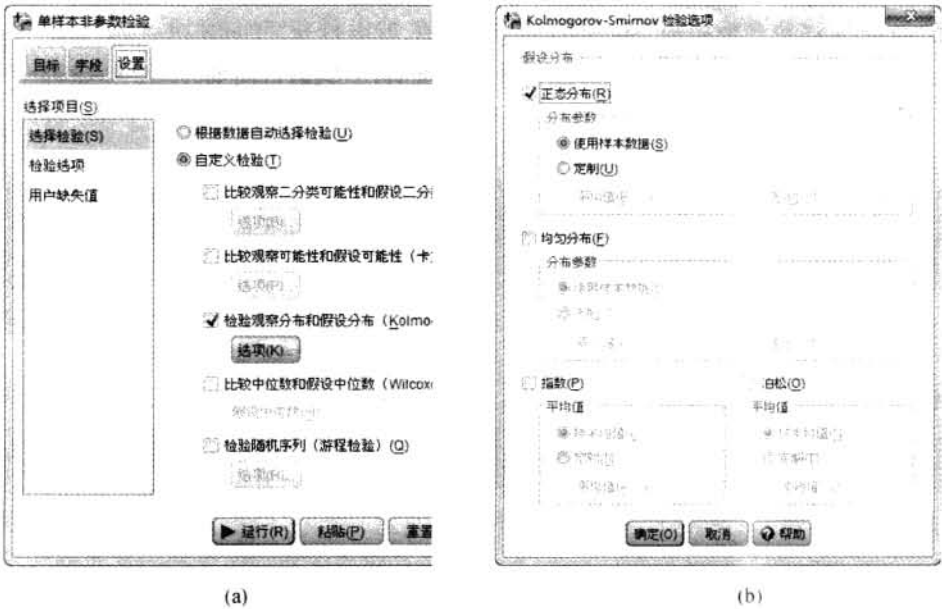


图 12.2 “单样本非参数检验”对话框(二)

2. 操作说明与结果解释

按照本例的分析目的,只需在筛选出 2007 年 4 月的案例之后,按上面所述将 index1 选入,然后选择用 K-S 方法进行正态分布检验即可,最后单击“运行”按钮,系统会给出一个非常简洁的图形结果,如图 12.3 所示。

假设检验汇总			
	原假设	测试	Sig.
1	总指数的分布为正态分布, 平均值为 98.34, 标准差为 18.92。	单样本 Kolmogorov-Smirnov 检验	.000

显示渐进显著性。显著性水平是 .05。

图 12.3 K-S 方法的模型输出

初次接触这种结果类型的读者可能会被弄得一头雾水,实际上,这就是第 1 章中曾经介绍过

的 SPSS 结果输出的种类之一:模型。上述输出只是整个模型的一个简报,双击模型则可进入编辑状态,从而得到更为详尽的结果描述。

在图 12.4 中可以看到,模型输出被分为左侧的主视图和右侧的辅助视图两部分,前者给出分析结果的汇总信息,而后者则给出更详细的信息。

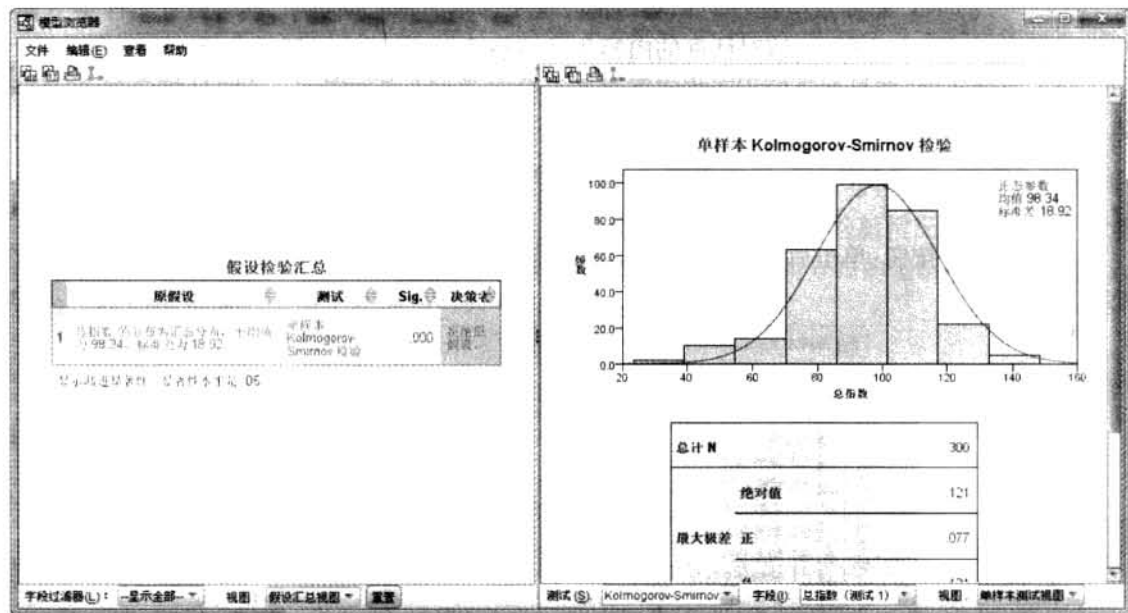


图 12.4 K-S 方法的详细模型输出

在辅助视图中可以看到 2007 年 4 月信心指数的均数为 98.34,标准差为 18.92,而在 300 例样本中,实际分布与假设分布之间的正向最大频数差为 0.077,负向最大频数差为 -0.121,因此用于计算统计量的绝对值最大频数差为 0.121。随后的统计量 Z 值为 2.088,相应的 P 值小于 0.001。根据这个标准得到的结论为:如果无效假设是成立的,则从这样一个正态分布的总体中按照现有样本量进行抽样,平均每 100 次中会有少于一次得到实际数据和理论分布之间的差值 K 等于甚至大于现有样本的 K 值 0.121,这显然是一个小概率事件,因此拒绝无效假设,可以在统计意义上认为信心指数不服从正态分布。

而作为对检验结果的汇总,左侧的主视图会给出最为精简的结论,具体内容实际上前面都已经介绍过了,这里不再重复。



这里有一个对初学者而言会颇为纠结的问题:既然此处拒绝了信心指数的正态分布假设,那么后面分析时还可以使用诸如 t 检验等对变量分布有要求的方法吗?事实上, $K-S$ 检验从实用性角度来说远不如图形工具,因为在样本量少的时候它不够敏感,而样本量大时又总是过于敏感。本例就属于敏感过头的情况,实际上读者们只需要绘制 $P-P$ 图就可以发现,该数据实际上是基本符合正态分布趋势的,进行后续数据分析时遵循正态数据的分析思路应当不会有任何问题。

12.2.3 使用旧对话框分析案例

对于仍在使用 SPSS 老版本,或者因许可证问题无法使用上述分析界面的用户而言,还可以使用旧对话框来完成上述案例的分析。

选择“分析”→“非参数检验”→“旧对话框”→“1 样本 K-S”菜单项,就会打开单样本 K-S 检验对话框,如图 12.5 所示,该对话框的内容非常容易理解,下面进行简要介绍。

(1) 主对话框:变量列表框用于指定需要进行分布类型分析的变量,可同时指定多个,下方用于指定具体要检验的分布。

(2) “精确”按钮:单击后打开的对话框用于对相应的检验提供确切概率法的结果,详见 12.5 节的介绍。

(3) “选项”按钮:单击后打开的对话框用于提供四分位数等常用统计量,以及对缺失值的处理方式。



图 12.5 单样本 K-S 检验对话框

按照本例的设定,相应的分析结果如图 12.6 所示,该结果和前面的输出内容实际上是完全相同的,因此不再重复解释。

		总指数
N		300
正态参数 ^{a,b}	均值	98.3363
	标准差	18.92074
最极端差别	绝对值	.121
	正	.077
	负	-.121
Kolmogorov - Smirnov Z		2.088
渐近显著性(双侧)		.000

a. 检验分布为正态分布。
b. 根据数据计算得到。

图 12.6 单样本 Kolmogorov - Smirnov 检验分析结果

12.3 二项分布检验

对于两分类变量而言,二项分布是最常见的分布类型,本节就来介绍一下二项分布检验方法。

12.3.1 二项分布检验的原理

二项分布检验(Binomial Test)是对二分类变量的拟合优度检验,用于考察每个类别中观察值的频数与特定二项分布下的预期频数间是否存在统计学差异。二项分布检验的原理实际上和 K-S 检验的原理相同,只是这里使用的是二分变量,是一个离散分布的检验情况。

在第 8 章中已经介绍了二项分布的基本知识,对于一个服从二项分布的随机变量而言,在 n 次试验中结局 A 出现的次数 X 的概率分布为

$$P(X=k) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \quad k=0,1,\dots,n$$

使用上述公式就可以算出基于无效假设时各发生次数的出现概率,利用小概率反证法,按照和 K-S 检验中类似的逻辑给出相应的检验结论。

12.3.2 案例:考察抽样数据的性别分布是否平衡

例 12.2 在人群中性别比例基本上是 1:1,考察 CCSS 样本中的数据是否仍然符合此规律。

本例所对应的检验假设如下。

H_0 : 男性(或女性)比例 $\pi=0.5$, 样本所对应的总体男女比例一致。

H_1 : $\pi \neq 0.5$, 男女性比例不一致。

注意此处仍然只使用 2007 年 4 月的样本数据进行分析,具体的分析操作也仍然使用“单样本非参数检验”对话框来完成。

- (1) 选择“数据”→“选择个案”菜单项。
- (2) 在打开的对话框中选择框组:如果条件满足。
- (3) “如果”子对话框:time = 200704。
- (4) 选择“分析”→“非参数检验”→“单样本”菜单项,打开“单样本非参数检验”对话框。
- (5) “目标”选项卡:自动比较观察数据和假想数据。
- (6) “字段”选项卡:使用定制字段分配,将 S2 性别选入“检验字段”列表框中。
- (7) “设置”选项卡:选择“自定义检验”选项组中的第一项“二项式检验”,相应选项中“假设比例”已经是所需的 0.5 了,因此不需要更改,如图 12.7 所示。
- (8) 确定。

最终的模型输出如图 12.8 所示,右侧辅助视图中同时给出了 165 例的男性样本频数,以及标准化统计量 1.674 的数值,最终基于样本和无效假设推导出的 P 值为 0.094,因此当无效假设成立时,100 次中平均有 9 次可以得到偏离理论值 150 例和现有样本一样远甚或更远(包括大于等于 165 例,或者小于等于 135 例两种情况)的样本,按照默认的 0.05 水准,这并非小概率事件,因

此不能拒绝无效假设,尚不能认为 CCSS 抽样数据的性别分布有差异。

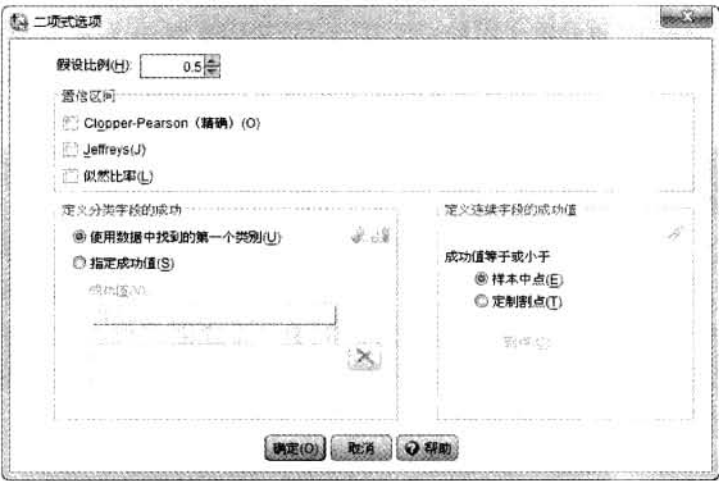


图 12.7 “二项式选项”对话框

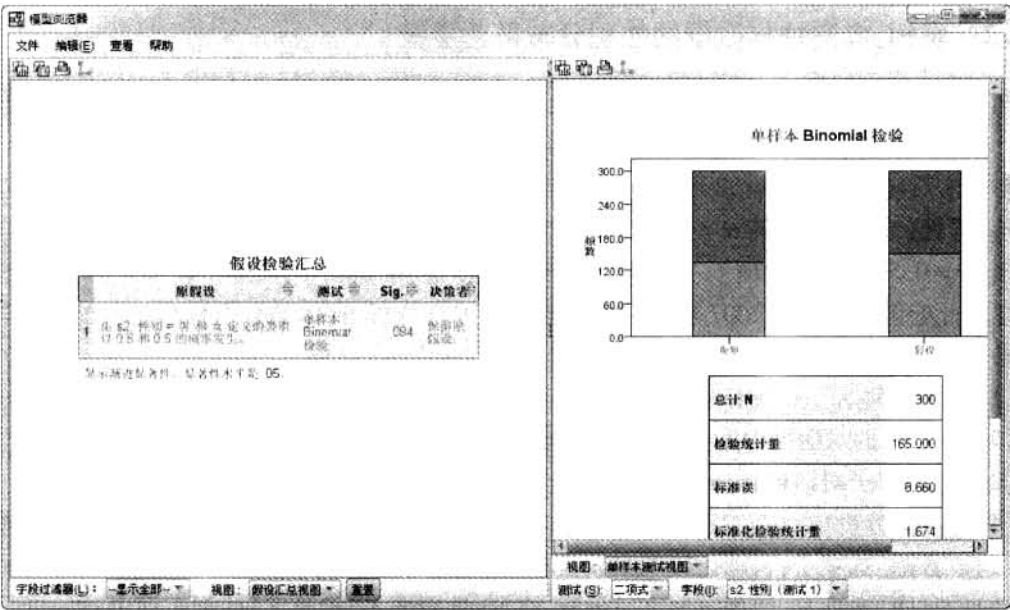


图 12.8 二项分布检验的详细模型输出

12.3.3 使用旧对话框分析案例

选择“分析”→“非参数检验”→“旧对话框”→“二项式”菜单项,就会打开“二项式检验”对话框,如图 12.9 所示,该中文界面非常简明,因此不再做重复解释。

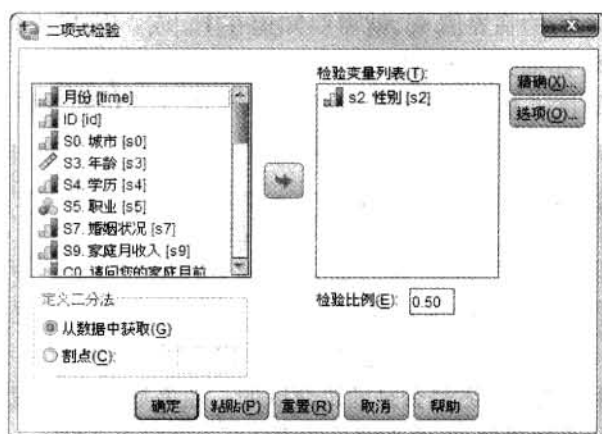


图 12.9 “二项式检验”对话框

分析结果的输出内容实际上和前面相同,如图 12.10 所示,因此也不再进行解释。

		类别	N	观察比例	检验比例	精确显著性(双侧)
S2. 性别	组 1	男	165	.55	.50	.094
	组 2	女	135	.45		
	总数		300	1.00		

图 12.10 二项式检验

12.4 游程检验

12.4.1 游程检验的原理

许多时候,研究者关心的不仅仅是分布的位置或者形状,也希望考察样本的随机性如何。因为如果样本不是从总体中随机抽取出的,那么所做的任何推断都将变得没有价值。而游程检验就是满足此类分析需求的一种基本的检验方法。

游程检验(Runs Test)是对二分变量的随机检验,它可用于判断观察值的顺序是否为随机的。对于两分类变量,连续数个相同取值的记录称为一个游程,比如下面这个序列

0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0 0 1 0

它有 6 个 0 的游程,其长度为 1、2、3 的各 2 个,并有 5 个 1 的游程,其中 3 个长度为 1,1 个长度为 2,1 个长度为 3。上面的序列总共有 11 个游程。如用 U 表示序列总的游程数,那么对于上面的序列来讲, $U=11$ 。

根据游程检验的假设,如果序列是真随机序列,那么游程的总数应当不太多也不太少,比较适中。如果游程的总数极少,就意味着样本由于缺乏独立性,内部存在着一定的趋势或结构,这可能是由于观察值间不独立(如传染病的发病),或者是来自不同总体;若样本中存在大量的游程,则可能存在系统的短周期波动影响着观察结果,同样不能认为序列是随机的。

为了确定游程检验所需的临界值 u_α , 就需要知道在 H_0 成立时 U 的概率分布。这一点比较复杂, 这里不做介绍, 读者如果感兴趣可以参考相应的文献。

SPSS 的 Runs 过程提供了基于游程个数的检验方法, 对于连续性变量, 该过程首先要将变量值进行分类, 然后进行检验。另外还有一种游程长度检验, 在 SPSS 中没有提供。

12.4.2 案例: 考察 CCSS 抽样数据是否随机

例 12.3 利用游程检验考察 CCSS 项目 2007 年 4 月样本的采集是否为随机的。

CCSS 项目反映的是抽样城市的普通城市常住居民对宏观经济的感受和预期, 因此要求抽样样本对总体有很好的代表性。该项目在质量控制上有很多措施和指标可供使用, 其中可用的一种指标就是不同性别、年龄的受访者是否是随机获取的。如果基本随机, 则理论上性别、年龄等背景变量的游程就应当属于真随机序列, 否则就可能说明某些人群进入样本的时间可能存在聚集性。

为了便于说明, 这里仍然只考察 2007 年 4 月的样本情况, 而数据集中的变量 ID 的大小顺序就代表了每个样本的进入顺序。

注意此处仍然只使用 2007 年 4 月的样本数据进行分析, 具体的分析操作也仍然使用“单样本非参数检验”对话框来完成。

(1) 选择“分析”→“非参数检验”→“单样本”菜单项, 打开“单样本非参数检验”对话框, 进行如下设置。

① “目标”选项卡: 检验随机序列。

② “字段”选项卡: 使用定制字段分配, 将“S2. 性别”、“S3. 年龄”选入“检验字段”列表框中。

③ “设置”选项卡: 选择“自定义检验”选项组中的最后一项“检验随机序列(游程检验)”, 相应的检验选项已经设定好了, 如图 12.11 所示, 因此不需要更改。

(2) 单击“确定”按钮。

最终的分析结果如图 12.12 所示, 首先来看性别, 模型输出说明 H_0 假设在理想情况下样本中应当有 150 个游程, 但现在样本中共有 146 个游程, 和 H_0 的理想状况差了 4 个, 相应的标化后统计量为 -0.409 , 但这个数值其实没人关心, 真正关心的是后面的 P 值: 0.683 , 之所以称为“渐进显著性”, 因为这里是按照正态近似方法计算的近似 P 值, 详见下一节的解释。

无论如何, 该检验结果说明在 H_0 成立的总体中, 平均 100 次抽样中有高达 68 次可以得到游程数大于等于 154, 或者小于等于 146 的情况, 显然属于大概率事件, 因此不能拒绝 H_0 , 尚无法认为性别序列是非随机的。

图 12.13(b) 给出的是年龄的游程检验结果, 这里只需要指出此处的 147 个游程是按照年龄中位数 35 岁作为分割点计算出来的, 而按此分割点, 游程检验的 P 值为 0.808 , 同样不能拒绝年龄序列随机的假设。对于连

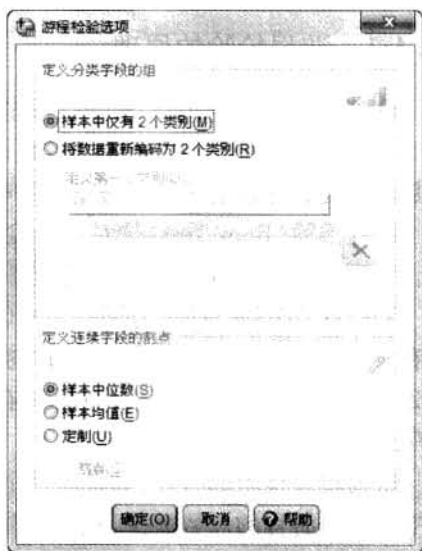


图 12.11 “游程检验选项”对话框

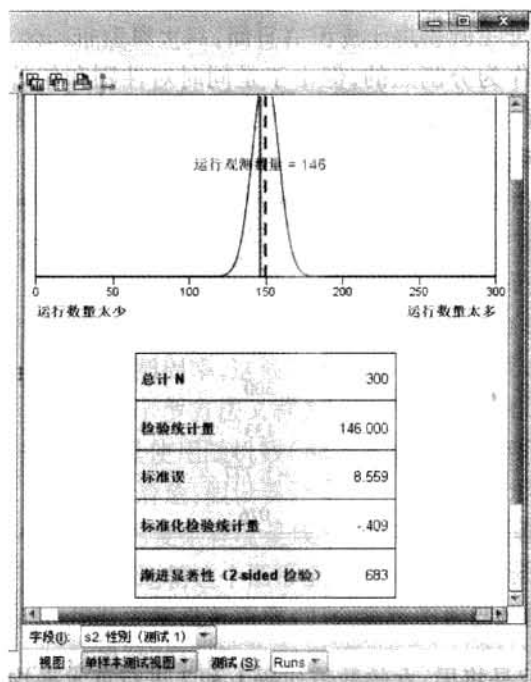
续性变量,分割点不同,则游程数量,以及游程分析的结果就会不同,这一点在阅读结果时至关重要。

假设检验汇总

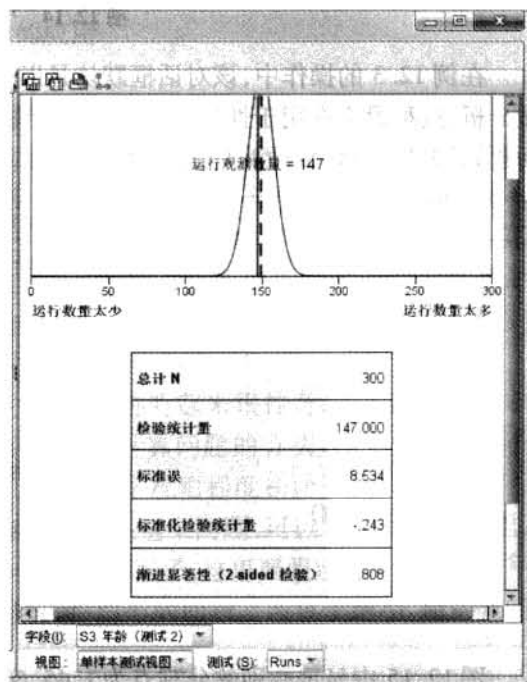
	原假设	测试	Sig.	决策者
1	由 S2. 性别 = (男) 和 (女) 定义的值的序列是随机序列。	单样本运行检验	.683	保留原假设。
2	由 S3. 年龄 ≤ 35.00 和 >35.00 定义的值的序列是随机序列。	单样本运行检验	.808	保留原假设。

显示渐进显著性。显著性水平是 .05。

图 12.12 游程检验的模型输出



(a)



(b)

图 12.13 游程检验的详细模型输出

12.4.3 使用旧对话框分析案例

如果使用旧对话框来完成游程检验,则选择“分析”→“非参数检验”→“旧对话框”→“游程”菜单项,就会打开相应的检验对话框,如图 12.14 所示,该中文界面非常简明,因此不再做重复解释。

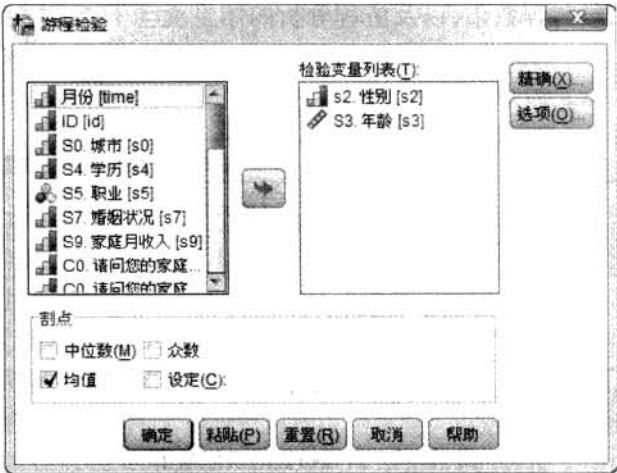


图 12.14 “游程检验”对话框

在例 12.3 的操作中,该对话框默认是以中位数作为分割点的,但由于是同时对性别和年龄做分析,这种设定在用于性别这种二分类变量时会有问题(读者可以自行尝试一下看看是什么问题),因此将设定修改为均值,最终得到相应的结果,如图 12.15 所示。

	S2. 性别	S3. 年龄
检验值 ^a	1.45	38.65
案例 < 检验值	165	171
案例 ≥ 检验值	135	129
案例总数	300	300
Runs 数	146	133
Z	-.409	-1.777
渐近显著性(双侧)	.683	.076

a. 均值

图 12.15 游程检验

图 12.15 中显示性别的分割点为 1.45,实际上就是将男、女的数值代码 1 和 2 进行算术平均后的均数值。SPSS 计算出游程数共有 146 个,对应的渐进 P 值为 0.683,和上面的输出结果完全相同。但是年龄的分析结果则让人意外,此处给出的 P 值竟然是 0.076,和原先的 0.808 大不相同!仔细阅读结果可以得知,这是因为该结果表格中年龄所使用的分割点为均数 38.65,而不是原先的中位数 35,因此导致游程数量减少为 133, P 值为 0.076,虽然仍然没有统计学差异,但已经比较接近检验水准了。

从上面的分析中可以看出,在对连续性变量进行游程检验时,采用不同的分割点,就可能得到截然不同的分析结果,因此在实际分析中,应当尽可能多取几个在专业背景上有实际意义的数值作为分割点,比较其游程检验的结果,以得到对序列随机性更为稳健和客观的结论。实际上就本例而言,如果取偏离均值 38.65 的数值作为分割点,就会发现 P 值会迅速升高并远离 0.05 的

界值,因此刚才这个偏小的 P 值可能多半是个特例,年龄序列的随机性应当是可以被确认的。

12.5 蒙特卡罗方法

在上面介绍旧对话框操作的时候,读者可能已经注意到有一个“精确”子对话框,顾名思义,该子对话框就是用于对相应的检验提供更为精确的 P 值计算结果的,本节就将其功能进行简单介绍。

在许多统计检验中,出于计算效率和计算难度的考虑,往往会在样本量充足等条件被满足时,采用对统计量的正态近似方法来得到一个近似的 P 值,该近似 P 值可能会略微偏离精确的 P 值,但计算速度更高,例如在 12.2 节的 K-S 旧对话框输出中,相应的 P 值就是“渐近显著性” P 值,即近似 P 值,这是当假设检验的计算工作必须要手工完成时非常重要的一种近似技术。

但是,近似结果总归还是有偏差的,科学家们总有力求完美的倾向,因此一直有学者在致力于更快速、更高效地取得更确切的计算结果。而这里的蒙特卡罗 (Monte Carlo) 方法就是最早期的一批成果之一,而且在历史上也得到了非常成功的应用。

12.5.1 蒙特卡罗方法简介

说起来,蒙特卡罗方法在起源和赌博有关的统计学中颇为高贵,一般认为是由 20 世纪 40 年代美国在第二次世界大战中研制原子弹的“曼哈顿计划”计划的成员 S. M. 乌拉姆和 J. 冯·诺依曼在核武器研究中产生的方法学的副产品。至于这个充满神秘色彩的名字,则是因其计算原理和赌博有点关联,因此数学家冯·诺依曼用驰名世界的赌城——摩纳哥的 Monte Carlo 来命名这种方法。实际上,在这之前,蒙特卡罗方法就已经存在。1777 年,法国 Buffon 提出用投针实验的方法求圆周率,这被认为是蒙特卡罗方法的起源。

蒙特卡罗方法又称为统计模拟法、随机抽样技术,其名称听起来很神奇,但实际上原理非常简单,就是使用随机数(或更常见的伪随机数)来解决很多计算问题的方法,用下面这个例子就可以解释清楚:假设要计算一个不规则图形的面积,图形的不规则程度和计算方法(比如是否使用积分)的复杂程度是成正比的。要使用蒙特卡罗方法解决这个问题,可以假设有一袋豆子,把豆子均匀地朝这个图形上撒,然后数这个图形之中有多少颗豆子,这里要假定豆子都在一个平面上,相互之间没有重叠,那么最终图形内豆子的数目就是图形的面积。当豆子越小,撒的量越多时,结果就越精确。

显然,该方法的原理很简单,但要在手工计算的条件下真正使用却并不容易,这也是为什么传统蒙特卡罗方法长期得不到推广的主要原因。但是 20 世纪下半叶以来随着计算机技术的发展,数据模拟和数据抽样变得非常容易,使得蒙特卡罗方法在最近这些年得到快速的普及。

12.5.2 蒙特卡罗方法的 SPSS 实现

1. 界面说明

在 SPSS 的一些操作对话框,特别是涉及有序分类、无序分类资料推断的非参数分析和卡方检验方法对话框中,都会提供如图 12.16 所示的“精确检验”子对话框,其中提供了以下 3 种选项。

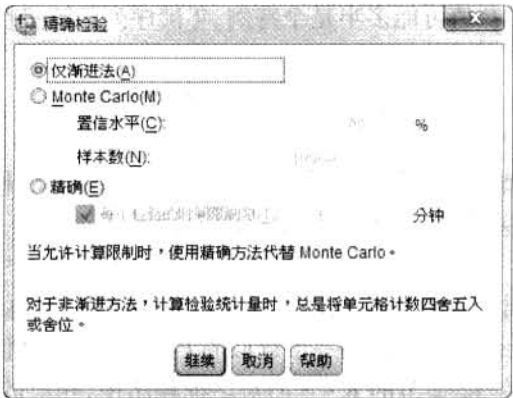


图 12.16 “精确检验”子对话框

(1) 仅渐进法:只计算基于检验统计量的渐近分布的近似概率值,而不计算确切概率。此为默认选项,当数据集样本量较大,且真实 P 值远离检验水准时,这样做可以节省计算时间。但是如果 P 值接近 Alpha 水准,或者样本量并不足够大/分布太偏,则计算出的结果可能偏差较大。

(2) Monte Carlo:即利用模拟抽样方法来求得对确切 P 值的无偏估计,当样本量较大时,和严格计算确切概率相比,蒙特卡罗方法可以在节约计算时间的情况下得到足够精确的结果。在下方可以进一步设定抽样中的细节,一般而言,采用默认的 10 000 次抽样计算出真实 P 值的 99% 可信区间就可以了,而其耗时一般在 10s 左右,远低于确切概率法。

(3) 精确:计算出确切的概率值,由于样本量较大时这样做有可能非常耗时,因此默认计算时间限制在 5 min 内,超过此时限则自动停止。该默认值可以更改。根据个人经验,除非是非常大的数据集,否则现在的计算机在大多数情况下都不会超过此时限。

2. 操作说明与结果解释

这里采用 12.4 节游程检验的案例来说明蒙特卡罗方法的输出,如果在上面的案例中选择使用蒙特卡罗抽样来计算 P 值,则相应的结果表格如图 12.17 所示。

	S2. 性别	S3. 年龄
检验值 ^a	1.45	38.65
案例 < 检验值	165	171
案例 ≥ 检验值	135	129
案例总数	300	300
Runs 数	146	133
Z	-.409	-1.777
渐近显著性(双侧)	.683	.076
Monte Carlo 显著性(双侧)	显著性	.728 ^b
	99% 置信区间	
	下限	.717
	上限	.740
		.086

a. 均值。
b. 基于 10 000 个具有起始种子 299 883 525 的采样表。

图 12.17 游程检验

图 12.17 中最下方会多出 3 行,分别是根据蒙特卡罗抽样计算出的 P 值估计值,以及相应的

P 值的 99% 可信区间上下界。以性别为例,其蒙特卡罗 P 值为 0.728,99% CI 为 0.717 ~ 0.740,显然和渐进性方法得到的 P 值 0.683 存在差异。年龄的情况也与此类似,但无论如何,其检验结果并未发生变化,仍然是无统计学差异的。

看到这里,可能有的读者会想到两个问题:首先,这两个 P 值究竟哪个更接近精确 P 值?其次,精确的 P 值究竟应当是多少呢?对于第一个问题,答案是在绝大多数情况下,蒙特卡罗的结果都应当是更精确的(但也会出现相反的情况,这是典型的小概率事件)。而对于第二个问题,只需要将上述选项更改为要求计算精确概率,即可得到结果表格,如图 12.18 所示。

	S2. 性别	S3. 年龄
检验值 ^a	1.45	38.65
案例 < 检验值	165	171
案例 ≥ 检验值	135	129
案例总数	300	300
Runs 数	146	133
Z	-.409	-1.777
渐近显著性(双侧)	.683	.076
精确显著性(双侧)	.726	.077
点概率	.043	.010

a. 均值。

图 12.8 游程检验

从图 12.18 中可见,性别游程检验的精确 P 值为 0.726,非常接近于蒙特卡罗方法的 0.728,而和近似方法的 0.683 相差较远,在年龄的检验中两种近似结果的精度相似。但如果比较两种计算方法所需的时间,则由 SPSS 输出中的附注信息可以知道蒙特卡罗抽样用时 0.94 秒,确切概率用时 22 秒,竟然相差 20 余倍!

12.6 本章小结

本章介绍了假设检验中最基本,也是最核心理论基础——小概率反证法。在进行假设检验时首先应当明确相应的假设是什么,随后围绕该假设来构建相应的统计量,并进行检验。如果给出了错误的假设,那么就可能选择错误的统计分析方法,最终将得到毫无意义的检验结果。相信读者在学习了本章的例子之后会对这一点有比较深的感受。

通过本章的学习,希望读者能够掌握下面涉及的知识 and 内容。

(1) 假设检验的理论基础是“小概率反证法”原理,无论多复杂的检验方法,其分析的逻辑基础都是该原理。

(2) 假设检验分析的基本步骤。

(3) 假设检验涉及的几个概念:原假设,备择假设;第一类错误,第二类错误;显著性水平;单尾检验,双尾检验。

(4) 参数检验以及非参数检验的概念。

(5) 几种常用的非参数检验:正态分布检验、二项分布检验、游程检验,熟悉使用 SPSS 进行分析的过程,懂得如何理解所获得的结果。

思考与练习

1. 假设检验的基本分析思路与基本理论基础是什么?
2. 如何衡量第一类错误与第二类错误? 它们之间的关系是什么?
3. 分析者可以接受原假设吗? 为什么?
4. 分析一个崭新的数据分析问题时,应该首先考虑哪些因素?
5. 正态分布检验的理论基础是什么? 找一个合适的例子加以练习。
6. 二项分布检验的理论基础是什么? 找一个合适的例子加以练习。
7. 什么是游程? 如何进行游程检验? 找一个合适的例子加以练习。

第 13 章 连续变量的统计推断(一)—— t 检验

通过第 12 章中对几种分布类型检验方法的学习,读者应该已经掌握了假设检验的基本原理——小概率反证法。但是,针对不同的数据类型,研究者还需要使用不同的方法和统计量来实现具体的检验问题。从本章开始就将针对各种数据类型进行相应检验方法的介绍。

13.1 t 检验概述

13.1.1 t 检验的基本原理

在针对连续变量的统计推断方法中,最常用的有 t 检验和方差分析两种,其中 t 检验是最基本的检验方法,也是统计学中跨里程碑的一个杰作。它最初是由 W. S. Gosset 在 1908 年以笔名“Student”发表的一篇关于 t 分布的论文中提出的,并从此开创了利用小样本计量资料进行统计推断的先河,迎来了统计学的新纪元。

1. 均数比较的一个实例

这里用一个典型的均数比较实例来引入 t 检验。

例 13.1 在 CCSS 项目中,以项目启动时的 2007 年 4 月的数据作为指数基线(具体方法学介绍参见本书第 6 章),基线期指数值为 100,随后各期所计算出的指数则代表当期数值相对于“基线”调查数值的变动比例。CCSS_Sample.sav 中提供了北京、上海、广州 3 个一线城市的调查数据,现希望考察 2007 年 4 月北京、上海、广州 3 个一线城市的消费者信心指数值是否和基准值 100 存在差异。

如果从统计学的角度来看,这是一个典型的对总体均数进行假设检验的问题。在这种问题中研究者所关心的变量为定距变量,因此可以使用均数来代表该定距变量的集中趋势。研究者对该样本所在总体的均数有一个事先的假设(本例中为指数值 100),而研究目的就是推断实际上该样本所在总体的均数是否等于这一已知总体均数。根据第 12 章中介绍的假设检验知识可以给出两种可能的假设如下:

$H_0: \mu = \mu_0$, 样本均数与假定总体均数的差异完全是由抽样误差造成的。

$H_1: \mu \neq \mu_0$, 样本均数与假定总体均数的差异除了由抽样误差造成外,确实也反映了实际的总体均数与假定的总体均数间的差异。

那么,究竟哪一种假设才是正确的呢? 根据假设检验的步骤,可以首先假定 H_0 是成立的。那么该样本就真的是从均数为 100 的总体中随机抽样而来的。但是如果考察该样本的实际数据,则其具体的统计描述指标如图 13.1 所示。

	N	极小值	极大值	均值	标准差
总指数	300	31. 24	140. 59	98. 3363	18. 92074
有效的 N (列表状态)	300				

a. 月份 = 200704。

图 13.1 描述统计量*

显然,2007 年 4 月北京、上海、广州 3 地的总样本均数不等于 100,而是 98. 34,两者间存在着差异。如果用公式来表示,就是 $\bar{X} - \mu = -1. 66$ 。仅看这一个数字很难判断出这种差异究竟是大还是小。因为这还和数据的离散程度有关,如果消费者的信心值差异较大,本身信心指数的离散程度就比较大,那么这一差值可能并不起眼。反之,则这一差值可能相对比较明显。为此需要找到某种方式对这一差值进行标准化。

2. U 检验

显然,标准化的基本思路应当是将该差值除以某种表示离散程度的指标。在第 7 章中曾经讨论过样本均数的抽样分布规律,这里再来复习一下:假设有一个已知服从正态分布的总体 $N(\mu, \sigma^2)$,现对其进行抽样研究,每次抽样的样本量固定为 n ,这样对每一个样本均可以计算出其均数 \bar{X} 。由于这种抽样可以进行无限多次,这些样本均数就会构成一个分布。统计学家发现,该分布正好就是正态分布 $N(\mu, \sigma^2/n)$ 。也就是说,样本均数所在分布的中心位置和原数据分布的中心位置相同,而其标准差(记为 $\sigma_{\bar{X}}$)则为 $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ 。为了区分样本所在总体的标准差,通常称样本均数的标准差为样本均数的标准误(简称均数标准误,有的书上也称之为标准误差);而且,即使是从偏态总体随机抽样,当 n 足够大时(如 $n > 50$), \bar{X} 也近似正态分布。这一规律就是数理统计中的中心极限定理(Central Limit Theorem)。显然,由于样本均数 \bar{X} 的分布规律为正态分布 $n(\mu, \sigma^2/N)$,只需要进行如下的标准化变换:

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

则 U 服从标准正态分布 $N(0, 1)$ 。也就是说,若资料服从正态分布 $N(\mu, \sigma^2)$,样本含量为 n 的样本均数 \bar{X} 出现在 $(\mu - 1. 96 \frac{\sigma}{\sqrt{n}}, \mu + 1. 96 \frac{\sigma}{\sqrt{n}})$ 之中的概率为 0. 95,这样就完成了对差值的标准化工作,可以具体计算出在相应的 H_0 总体中抽得当前样本(即更极端情况)的概率大小,从而做出统计推断结论了。该方法就是所谓的 U 检验。

3. 从 U 检验到 t 检验

但是, U 检验看上去虽然很好,却实际上毫无用处,因为 $\sigma_{\bar{X}}$ 在计算中需要使用总体标准差,但在实际工作中和总体均数一样也常常未知,能够使用的仅仅是样本标准差 s 。W. S. Gosset 的贡献正在于此,他发现如果用样本标准差来代替总体标准差进行计算,即 $s_{\bar{X}} = s/\sqrt{n}$,则由于样本标准差 s 会随样本而变。相应的标化统计量的变异程度要大于 U ,它的密度曲线看上去有些像标准正态分布,但是尖一些,而且尾巴长一些,这种分布称为 t 分布,如图 13. 2 所示。而相应的标化后统计量也就被称为 t 统计量。显然, t 统计量的分布规律是和样本量有关的,更准确地说

是和自由度有关。自由度(Degree of Freedom, 一般用 ν 或者英文缩写 df 来表示)这个概念还出现在其他分布之中, 它基本上是信息量大小的一个度量, 描述了样本数据能自由取值的个数, 在 t 分布中由于有给定的样本均数这一限定, 所以自由度为 $\nu = n - 1$ 。从图 13.2 中可以看出, 当自由度增加时, 它的分布就逐渐接近标准正态分布了。因此, 在样本量较大时, 可以用标准正态分布来近似 t 分布。

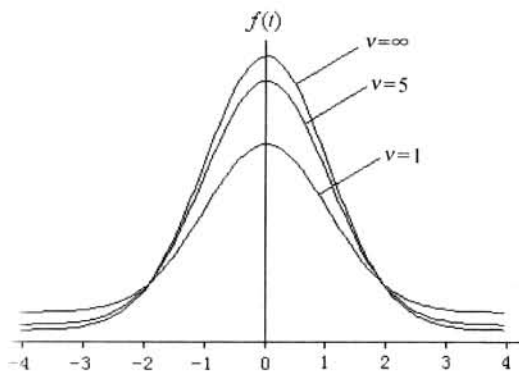


图 13.2 t 分布示意图

t 检验即是应用 t 分布的特征, 将 t 作为检验的统计量来进行的检验, 由于 W. S. Gosset 已经对不同自由度时 t 分布下面积的概率分布规律进行了很好的总结, 所以就可以利用 t 统计量来回答上述关于均数的假设检验问题了。具体的统计量计算公式为

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{X}}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \text{ 自由度 } df = n - 1$$

13.1.2 SPSS 中的相应功能

t 检验在 SPSS 中基本上被集中在“比较均值”子菜单中, 具体如下。

- (1) 单样本 t 检验过程: 进行样本均数与已知总体均数的比较。
- (2) 独立样本 t 检验过程: 进行两样本均数差别的比较, 即通常所说的两组资料的 t 检验。
- (3) 配对样本 t 检验过程: 进行配对资料的均数比较, 即配对 t 检验。



“比较均值”子菜单中的第一项均值过程实际上更倾向于对样本进行描述, 它可以对需要比较的各组计算描述指标进行检验前的预分析。当然如果愿意也可以直接进行比较。

13.2 样本均数与总体均数的比较

13.2.1 单样本案例: 基期一线城市信心指数与基准值的比较

单个样本均数检验问题是一种关于总体均数的假设检验问题。这种问题中只有一个随机抽取的样本, 研究目的是推断这个样本的总体均数是否等于(或大于, 或小于)某个已知总体均数。以例 13.1 为例, 首先应当建立相应的假设。

$H_0: \mu = \mu_0$, 2007 年 4 月一线城市的总信心指数均值为 100。

$H_1: \mu \neq \mu_0$, 2007 年 4 月一线城市的总指数均值不是 100。

$\alpha = 0.05$ 。

数据见文件 CCSS_Sample.sav, 其中变量 index1 为 2007 年 4 月的总指数, 这是一个典型的单样本总体均数检验问题。

1. 界面说明

选择“分析”→“比较均值”→“单样本 t 检验”菜单项, 即可打开“单样本 t 检验”对话框, 如图 13.3 所示, 该对话框非常简单, 界面简介如下。

(1) 主对话框: 检验变量框用于选入需要分析的变量, 下方的“检验值”文本框则用于输入已知的总体均数, 默认值为 0。

(2) “选项”子对话框: “置信区间百分比”文本框用于设定需要计算的均数差值可信区间范围, 默认为 95%。如果是和总体均数为 0 相比, 则此处计算的就是样本所在总体均数的可信区间。而“缺失值”单选框组则用于对缺失值的处理方法加以定义, 一般不用更改。

(3) “Bootstrap”子对话框: 要求对相应的单样本 t 检验进行指定的 Bootstrap 抽样估计。该方法的详情已经在第 7 章中进行了介绍, 因此这里不再重复。

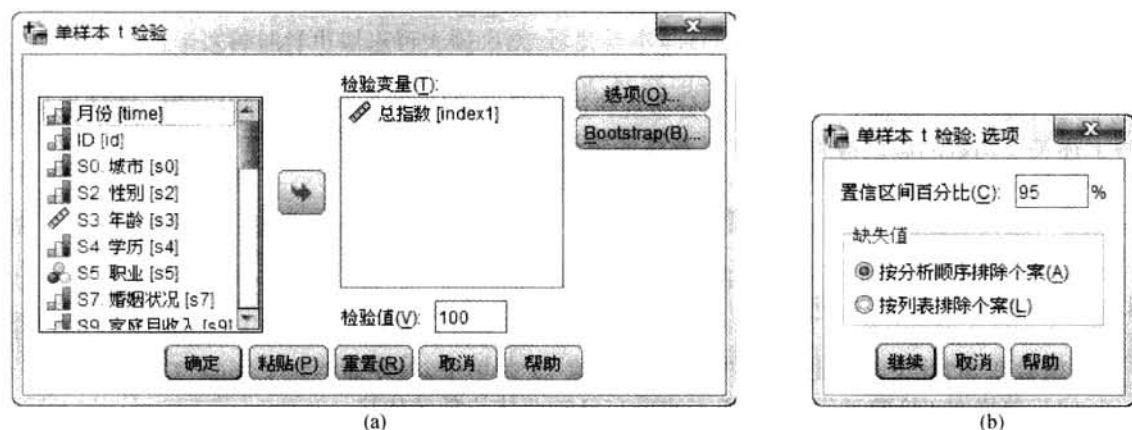


图 13.3 “单样本 t 检验”对话框

2. 操作说明与结果解释

用 SPSS 完成本案例的具体操作如下。

- (1) 选择“数据”→“选择个案”菜单项。
- (2) 在打开的对话框中选择框组: 如果条件满足。
- (3) 单击“如果”按钮, 设置 time = 200704, 继续。
- (4) 单击“确定”按钮。
- (5) 选择“分析”→“比较均值”→“单样本 t 检验”菜单项, 打开“单样本 t 检验”对话框。
- (6) “检验变量”列表框: 总指数[index1]。
- (7) “检验值”文本框: 输入 100。
- (8) 单击“确定”按钮。

本例的输出如图 13.4 所示。

	N	均值	标准差	均值的标准误
总指数	300	98.3363	18.92074	1.09239

图 13.4 单个样本统计量

首先给出的是对当前样本进行的统计描述,可见 2010 年 4 月北京、上海、广州 3 个一线城市总样本的信心指数均值为 98.3,低于基线水平 100。注意最右侧给出的均值的标准误,是对样本均数抽样误差大小的描述指标。

图 13.5 即为单样本 t 检验的分析结果,第 1 行注明了用于比较的假设总体均数为 100,下面从左到右依次为 t 值(t)、自由度(df)、 P 值(Sig.(双侧))、两均数的差值(均值差值)、差值(差分的 95% 置信区间)。根据上面的检验结果 $t = -1.523$, $P = 0.129$,由于 P 值大于检验水准 0.05,因此不能拒绝 H_0 ,尚不能认为样本所在总体的均数与假设的总体均数不同。

检验值 = 100						
	t	df	Sig. (双侧)	均值差值	差分的 95% 置信区间	
					下限	上限
总指数	-1.523	299	.129	-1.66367	-3.8134	.4861

图 13.5 单个样本检验

13.2.2 单样本 t 检验中的其他问题

1. 总体均数置信区间与 t 检验的一致性

图 13.5 中同时给出了总体均数的置信区间和 t 检验的结果,两者的结论实际上是完全一致的,置信区间可用于回答假设检验的问题,同时这两者又是互为补充的关系:置信区间回答“量”的问题,即总体均数的范围在哪里,而假设检验回答“质”的问题,即总体均数之间是否存在差异,以及在统计上确认这种差异的把握有多大。

置信区间在回答有无统计学意义的同时,还可进一步回答这种差异有无实际意义,如在 13.2.1 节中的案例中,2007 年 4 月份的总指数与 100 相差在一定范围内都是正常的,则即使差异具有统计学意义,如果差值的可信区间并未超过范围,这个差值也可以认为正常。

2. 单样本 t 检验的应用条件

由中心极限定理可知,即使原数据不服从正态分布,只要样本量足够大,其样本均数的抽样分布仍然是正态的。因此当样本量较大时,研究者很少去考虑单样本 t 检验的适用条件,此时真正会限制该方法使用的是均数是否能够代表相应数据的集中趋势。也就是说,只要数据分布不是强烈的偏态,一般而言单样本 t 检验都是适用的。

当样本例数 n 较小时,一般要求样本取自正态总体,这可以通过第 12 章所介绍过的正态性检验($K-S$ 检验)来考察,该方法适用于大样本(SPSS 规定样本大于 5 000 个),也可以用更直观的作图方法来判断,详见相应章节。但是一般而言,单样本 t 检验是一个非常稳健的统计方法,只要没有明显的极端值,其分析结果都是稳定的。

13.3 成组设计两样本均数的比较

在实际问题中,除了一个总体的检验问题外,还常碰到两个总体均数的比较问题,此时可以考虑使用成组设计的 t 检验来进行分析。

13.3.1 方法原理

两样本 t 检验和单样本 t 检验的基本原理实际上非常相似,设两组样本量分别为 n_1 和 n_2 ,且均来自两个正态分布的总体: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$,则两样本 t 检验所建立的假设为

$H_0: \mu_1 = \mu_2$, 两样本均数的差异完全是抽样误差造成的,两总体均数相同。

$H_1: \mu_1 \neq \mu_2$, 两样本均数的差异除由抽样误差造成外,也确实反映了两总体均数存在的差异。

1. 两样本 t 检验的基本思想

显然,无效假设等价于认为 $\mu_1 - \mu_2 = 0$,而当前样本信息和这一假设情况的差异为

$$(\bar{X}_1 - \bar{X}_2) - 0 = \bar{X}_1 - \bar{X}_2$$

和单样本 t 检验时的情形相同,上述数值虽然可以代表与 H_0 假设情形的差异大小,但该数值的大小还和数据的离散程度有关,同样需要找到某种方式对这一差值进行标准化。统计学家发现,如果这两个总体的方差完全相同,即 $\sigma_1^2 = \sigma_2^2$,即这两个总体实际上同正态分布是一个总体时,从该总体中分别进行样本量为 n_1 和 n_2 的随机抽样,则样本均数差值 $\bar{X}_1 - \bar{X}_2$ 也服从正态分布,其均数为 0,标准差(标准误)则为

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2(1/n_1 + 1/n_2)}$$

但是,和单样本 t 检验时的情况相似, $\sigma_{\bar{X}_1 - \bar{X}_2}$ 在计算中也需要使用总体标准差 σ ,但在实际工作中它常常未知,能够使用的仅仅是两个样本的标准差 s_1 和 s_2 而已,此时相应的合并标准误计算公式为

$$s_c^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

将该总体方差估计值代入公式中,即可解出相应的样本均数差值标准误的估计值 $s_{\bar{X}_1 - \bar{X}_2}$ 。统计学家发现,如果这两个样本所在总体的标准差的确是完全相同的,则标化后的差值应当服从自由度为 $(n_1 - 1) + (n_2 - 1)$ 的 t 分布,即

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_c^2(1/n_1 + 1/n_2)}}, v = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

在上面自由度计算中减去的两个限制条件其实就对应了两个样本均数。由以上推导可知,进行两样本均数比较的 t 检验要求两样本来自的总体方差相等,即方差齐性。总体方差是否相等,可通过方差齐性检验来进行统计推断,本章后面将对此做专门讲解。

2. 校正的 t' 检验

当两样本所在总体的方差不同,即方差不齐时,根据上式计算出的“ t ”值并不服从相应的 t 分布,此时需要对结果进行一定的校正,其中对 t 统计量和自由度的校正计算公式分别为

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, v = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

按相应的 t 值和自由度,即可计算出相对应的 P 值来,这就是所谓的当方差不齐时用于比较两个样本的 t' 检验。

13.3.2 案例:不同收入水平家庭的信心指数比较

例 13.2 研究者认为家庭收入的高低可能会影响消费者信心的平均水平,收入较高的家庭其消费者信心应当较低收入家庭更高。根据前期研究的结果,CCSS 项目中将受访家庭按照年收入是否大于 4.8 万元人民币分为两组,这里以 2007 年 4 月的数据为例,比较这两组家庭的消费者信心均值有无差异。



统计初学者可能对本案例的做法略有异议:家庭收入在原始问卷上明明至少是一个有序分类的变量,为什么硬要拆分成两分类变量来进行比较!是的,这一质问是有道理的,这样转换的确会损失有效信息,但是在数据分析中同时也需要考虑结果的可理解性和可应用性。将家庭收入分为高、中、低三级,或者中高、中低两级是数据分析中常见的做法,其分析结果也容易为使用者所理解,这可以看做是在精确性和易用性上所做的折中,当然底线是不能因此导致错误的分析结论。

本案例的数据见文件 CCSS_Sample.sav,其中变量 index1 为总指数,Ts9 为家庭收入 2 级。这是一个典型的两样本 t 检验问题,建立的假设如下:

$H_0: \mu_1 = \mu_2$, 两个家庭收入级别在总指数上没有差别。

$H_1: \mu_1 \neq \mu_2$, 两个家庭收入级别在总指数上有差别。

$\alpha = 0.05$ 。

1. 界面说明

选择“分析”→“比较均值”→“独立样本 t 检验”菜单项,即可打开“独立样本 t 检验”对话框,如图 13.6 所示,界面简介如下。

(1) “检验变量”列表框:用于选入需要分析的变量。

(2) “分组变量”列表框:用于选入分组变量。注意选入后还要定义需比较的组别。

(3) “定义组”按钮:单击后打开的对话框用于定义需要相互比较的两组的分组变量值。可以直接指定给分组变量两个取值,相应的两组将进行比较。也可以使用割点按照分组变量的某个取值将样本分为两组来进行比较(如分为小于 30 岁的和大于等于 30 岁的进行比较)。

(4) “选项”按钮和“Bootstrap”按钮:相应功能和“单样本 t 检验”对话框中完全相同,此处不再重复。

2. 操作说明与结果解释

本例的具体操作如下。

(1) 选择“数据”→“选择个案”菜单项。

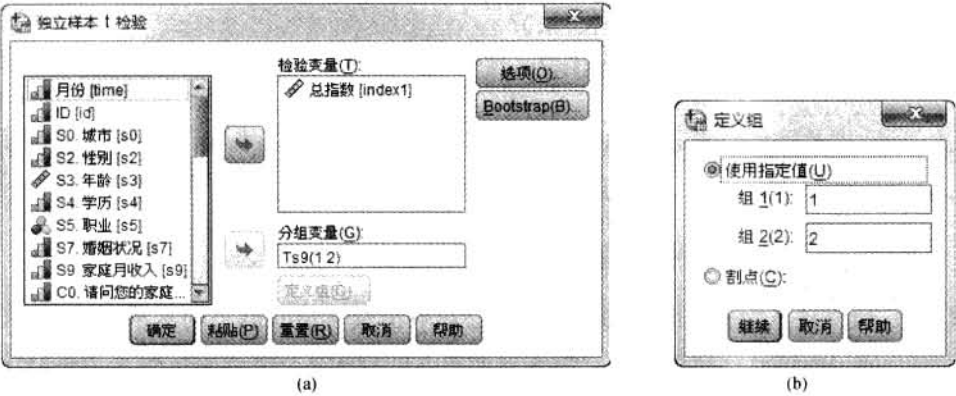


图 13.6 “独立样本 t 检验”对话框

- (2) 在打开的框组中选择框组:如果条件满足。
 - (3) 单击“如果”按钮:设置 time = 200704 ,继续。
 - (4) 单击“确定”按钮。
 - (5) 选择“分析”→“比较均值”→“独立样本 t 检验”菜单项,打开“独立样本 t 检验”对话框。
 - (6) “检验变量”列表框:选入总指数[index1]。
 - (7) “分组变量”列表框:选入 Ts9。
 - (8) “定义组”子对话框:组 1:1;组 2:2,单击“继续”按钮。
 - (9) 单击“确定”按钮。
- 本例的分析结果如图 13.7 所示。

	家庭收入 2 级	N	均值	标准差	均值的标准误
总指数	Below 48,000	110	90.7458	21.23893	2.02505
	Over 48,000	145	104.4475	14.92637	1.23957

图 13.7 组统计量

首先给出的是两组需检验变量的基本情况描述,不再详述。

随后结果中会给出最重要的方差齐性检验和 t 检验分析结果,由于内容较多,为了便于讲解,下面使用表格编辑功能将其拆分为两部分分别加以说明,如图 13.8、图 13.9 所示。

		方差方程的 Levene 检验	
		F	Sig.
总指数	假设方差相等	11.930	.001
	假设方差不相等		

图 13.8 独立样本检验(1)

分析结果的第一部分为 Levene's 方差齐性检验,用于判断两总体方差是否为齐性方差,这里的检验结果为 $F = 11.930, P = 0.001$,因此拒绝 H_0 ,认为本例中的两个样本所在总体的方差是不齐的。

		均值方程的 t 检验					差值的 95% 置信区间	
		t	df	Sig. (双侧)	均值差值	标准误差值	下限	上限
总指数	假设方差相等	-6.047	253	.000	-13.70173	2.26593	-18.16421	-9.23924
	假设方差不相等	-5.771	186.197	.000	-13.70173	2.37431	-18.38574	-9.01771

图 13.9 独立样本检验(2)

结果表格的第二部分会分别给出两组所在总体方差为齐性方差和非齐性方差时的 t 检验结果,当假设两总体方差为齐性方差时,就直接进行标准的两样本 t 检验;否则,就根据两样本的方差情况对标准差进行校正,得到的是校正 t 检验的结果。具体应当看这两种结果中的哪一种需要根据方差齐性检验的结果加以判断。在本例中由于前面的方差齐性检验结果为方差不相等(不齐),因此应选用方差不相等(不齐)时的 t 检验结果,即图 13.9 中第二行列出的 $t = -5.771, df = 186.197, P$ 值(Sig.)显示为 0.000,小于 0.05,从而最终得到的统计结论为按 $\alpha = 0.05$ 水准,拒绝 H_0 ,接受 H_1 ,可以认为两个家庭收入级别在总指数上存在统计学差异。

图 13.9 的最后面还附有两组均数差值的置信区间等其他指标,此处不再详细解释。

13.3.3 适用条件与方差齐性检验

在应用 t 检验进行两样本均数的比较时,要求数据满足以下 3 个条件。

- (1) 独立性,各观察值之间是相互独立的,不能相互影响。
- (2) 正态性,各个样本均来自于正态分布的总体。
- (3) 方差齐性,各个样本所在总体的方差相等。

在实际应用中,独立性对结果的影响较大,但检验数据独立性的方法比较复杂,一般都是根据资料的性质来加以判断的。例如遗传性疾病、传染病的数据可能就存在非独立的问题。如果从专业背景上可以肯定数据不存在这些问题,则一般独立性总是能够满足的。

t 检验对于资料的正态性有一定的耐受能力,如果资料只是稍微偏离正态,则结果仍然是很稳定的。当然,如果数据分布偏离正态很远,可知此时均数不能很好地代表数据的集中趋势,在这种情况下最好考虑采用变量变换或者非参数方法加以分析,详见相关章节。一般对正态性的考察可以通过直方图等工具进行,当数据量较少时甚至可以直接观察数据。但是要注意应当分组考察正态性,而不是合并进行。

和总体的正态性相比,方差齐性对结论的影响较大。在进行均数比较时进行方差齐性检验就显得更为重要。在 SPSS 中方差齐性检验可通过 Levene's 检验来进行,其假设为

$$H_0:\sigma_1^2 = \sigma_2^2, \text{两总体方差相同。}$$

$$H_1:\sigma_1^2 \neq \sigma_2^2, \text{两总体方差不同。}$$

Levene's 检验的实质是将两组数据的方差进行比较,其统计量的计算公式为

$$F = s_1^2 / s_2^2, v_1 = n_1 - 1, v_2 = n_2 - 1$$

其中分子为较大的方差,如果两组方差的比值较大,其所对应的 P 值小于设定的检验水准,则按照小概率反证法原理拒绝 H_0 ,认为两组所在总体的方差不齐。

在上面两样本 t 检验的结果中已经提供了 Levene's 检验的结果,实际上在 SPSS 的探索过程中可以进行更为详细的 Levene's 方差齐性检验,对于本例相应的操作如下(探索过程的对话框解释请参见第 7 章相应内容)。

(1) 选择“分析”→“描述统计”→“探索”菜单项,打开“探索”对话框,进行如下设置。

- ① “因变量”列表框:总指数[index1]。
- ② “因子”列表框:家庭收入 2 级[Ts9]。

(2) 单击“绘制”按钮,伸展与级别 Levene 检验框组,选择未转换,继续。

(3) 单击“确定”按钮。

结果如图 13.10 所示,其中包括了 4 种水平的 Levene's 检验结果,分别为基于均值(Based on Mean)、基于中值(Based on Median)、基于调整自由度的中位数(Based on Median and with Adjusted df)和基于修整(截尾)均数(Based on Trimmed Mean)的 Levene 检验,后面是相应的统计量 F 值(Levene 统计量)、两个自由度值(df1、df2),以及 P 值(Sig.)。以上结果分别适用于不同的数据情况,如果数据为对称分布,则可以使用基于均数的结果;偏态数据则使用基于中位数的结果。如果存在极端值,则可以考虑使用基于截尾均数的结果。这样,Levene's 检验的结果就可以适用于任意分布类型的资料,适用范围更广。

		Levene 统计量	df1	df2	Sig.
总指数	基于均值	11.930	1	253	.001
	基于中值	11.274	1	253	.001
	基于中值和调整后的 df	11.274	1	229.017	.001
	基于修整均值	11.781	1	253	.001

图 13.10 方差齐性检验

13.4 配对设计样本均数的比较

在很多科学研究中,常采用配对设计来提高研究效率,常见的配对设计有 4 种情况:① 同一受试对象处理前后的数据;② 同一受试对象两个部位的数据;③ 同一样品用两种方法(仪器等)检验的结果;④ 配对的两个受试对象分别接受进行两种处理后的数据。情况①的目的是推断其处理有无作用;情况②、③、④的目的是推断两种处理(方法等)的结果有无差别。在进行配对设计得到的样本数据中,每对数据之间都有一定的相关,如果忽略这种关系就会浪费大量的统计信息,因此在分析中应当采用和配对设计相对应的分析方法。当进行配对设计所测量到的数据为定距变量时,配对 t 检验就是最常用的分析方法。

13.4.1 方法原理

配对 t 检验的基本原理是为每对数据求差值:如果两种处理实际上没有差异,则差值的总体均数应当为 0,从该总体中抽出的样本其均数也应当在 0 附近波动;反之,如果两种处理有差异,差值的总体均数就应当远离 0,其样本均数也应当远离 0。这样,通过检验该差值总体均数是否为 0,就可以得知两种处理有无差异。

配对 t 检验相应的假设为

$H_0: \mu_d = 0$, 两种处理没有差别。

$H_1: \mu_d \neq 0$, 两种处理存在差别。

其统计量的计算公式为

$$t = \frac{\bar{d} - 0}{s_{\bar{d}}} = \frac{\bar{d}}{s/\sqrt{n}}, \quad df = n - 1 \quad (n \text{ 为对子数})$$

有了前面的基础,其实可以看出,配对样本 t 检验过程的功能实际上是和单样本 t 检验过程相重复的(等价于已知总体均数为 0 的情况),但配对样本 t 检验过程使用的数据输入格式和前者不同,因此它仍然有存在的价值。



由于配对 t 检验的本质就是单样本 t 检验,因此其适用条件的考察也和单样本 t 检验近似(注意应当考察差值而不是原始数据),这里不再重复。

13.4.2 案例:治疗前后舒张压均数的比较

例 13.3 用某药治疗 10 名高血压病人,对每一病人治疗前、后的舒张压(mmHg)进行了测量,结果如图 13.11 所示,问该药有无降压作用?

病例编号	1	2	3	4	5	6	7	8	9	10
治疗前	120	127	141	107	110	114	115	138	127	122
治疗后	123	108	120	107	100	98	102	152	104	107

图 13.11 治疗前后的舒张压(mmHg)测量结果

这是一个典型的个体自身治疗前后的配对设计,应当采用配对设计差值的 t 检验来进行分析。按照配对 t 检验对数据格式的要求,这里在输入数据时应当用每个变量(一列)代表一个组,而每条记录(一行)代表一对数据。最终数据见文件 pairedt.sav。本例建立的假设为

$H_0: \mu_d = 0$, 同一病人治疗前后的舒张压差值总体均数为 0。

$H_1: \mu_d \neq 0$, 同一病人治疗前后的舒张压差值总体均数不为 0。

$\alpha = 0.05$ 。

1. 界面说明

配对 t 检验所使用的对话框非常简单,如图 13.12 所示,简介如下。

(1) “成对变量”表:用于选入希望进行比较的一对或几对变量——注意这里的量词是对而不是个。变量需要成对选入,即按住 Ctrl 键,用鼠标依次选中一对变量,再将其选入。如果只选

中一个变量,则变量移动按钮为灰色,不可用。

(2) “选项”子对话框、“Bootstrap”子对话框:功能和前面介绍的完全相同,此处不再重复。

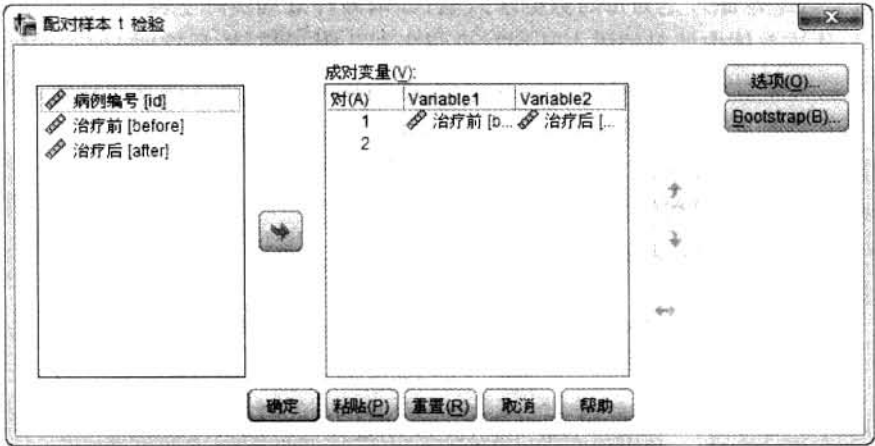


图 13.12 “配对样本 t 检验”对话框

2. 操作说明与结果解释

在 SPSS 中完成本例的具体操作如下。

- (1) 选择“分析”→“比较均值”→“配对样本 t 检验”菜单项,打开“配对样本 t 检验”对话框。
 - (2) 在“成对变量”表中同时选中“治疗前[before]”、“治疗后[after]”选项并将其添加到成对变量列表框中。
 - (3) 单击“确定”按钮。
- 本例的分析结果如图 13.13 所示。

		均值	N	标准差	均值的标准误
对 1	治疗前	122.1000	10	11.31813	3.57911
	治疗后	112.1000	10	16.17577	5.11523

图 13.13 成对样本统计量

首先给出的是配对变量各自的统计描述,因此处只有 1 对,故列表中只有对 1 出现。

随后给出的是成对变量间的相关性分析,其结果实际上就是两变量的积矩相关系数及其检验结果,如图 13.14 所示,详见第 17 章的讲解。

		N	相关系数	Sig.
对 1	治疗前 & 治疗后	10	.674	.033

图 13.14 成对样本相关系数

最后输出的才是配对 t 检验的结果,如图 13.15 所示,其中给出了对差值的统计描述。注意

上面的均值、标准差、标准误和可信区间等都是针对差值的统计量。随后给出的是对差值的检验结果,由图 13.11 可见,差值均数为 10,相应的 $P=0.027$,故可以认为使用该药会影响病人的血压,由于样本中治疗前-治疗后的差值均数为正,因此可推断出是使得病人血压下降,即有降压作用。

		成对差分		差分的 95% 置信区间		t	df	Sig. (双侧)
		均值	标准差	均值的标准误	下限	上限		
对 1	治疗前 - 治疗后	10.00000	11.95361	3.78006	1.44890	18.55110	2.645	.027

图 13.15 成对样本检验

13.5 本章小结

(1) 本章介绍的是假设检验中非常基础和重要的 t 检验, t 检验仍然采用小概率反证法原理,其基本思想是:首先假设 H_0 成立,然后考察在 H_0 成立的条件下,按照现有样本量做随机抽样,在相应的总体中抽到现有样本,以及比现有样本与总体的差异更大的样本的累积概率,如果相应的概率 $P \leq \alpha$ (检验水准),则拒绝 H_0 假设,接受对立的 H_1 假设,认为现有样本并非来自于所假定的总体。

(2) 在整个推断过程中,由于利用了 t 分布求得 t 值,并据此而得到相应的概率值,因此检验方法被称为 t 检验。而根据具体的设计方案和希望解决的问题不同,又可以将其分为单样本 t 检验、两样本 t 检验和配对 t 检验等,但它们的基本原理都是相同的。

(3) 作为参数方法, t 检验也有适用的条件,但相对而言它比较稳健,对使用条件的违反有一定的耐受性。但如果使用条件被严重违反,则可以采用校正的 t 检验,或者换用非参数方法来进行分析。

思考与练习

1. 从一批木头里抽取 5 根,测得直径如下(单位:cm),是否能认为这批木头的平均直径是 12.3 cm?

12.3 12.8 12.4 12.1 12.7

2. 为研究女性服用某避孕新药后是否影响其血清总胆固醇,将 20 名女性按年龄配成 10 对。从每对中随机抽取一人服用新药,另一人服用安慰剂。经过一定时间后,测得血清总胆固醇含量(mmol/L),结果如题表 1 所示。问该新药是否影响女性血清总胆固醇?

题表 1

配对号	1	2	3	4	5	6	7	8	9	10
新药组	4.4	5	5.8	4.6	4.9	4.8	6	5.9	4.3	5.1
安慰剂组	6.2	5.2	5.5	5	4.4	5.4	5	6.4	5.8	6.2

3. 比较两批电子器材的电阻,随机抽取的样本测量电阻如题表 2 所示,试比较两批电子器材的电阻是否相同(提示:需考虑方差齐性问题)。

题表 2

A 批	0.140	0.138	0.143	0.142	0.144	0.148	0.137
B 批	0.135	0.140	0.142	0.136	0.138	0.140	0.141

4. 配对 t 检验的实质就是对差值进行单样本 t 检验,要求按此思路对例 13.3 进行重新分析,比较其结果和配对 t 检验的结果有什么异同。

第 14 章 连续变量的统计推断(二)——单因素方差分析

14.1 方差分析简介

14.1.1 进行方差分析的原因

第 13 章中介绍的 t 检验可以解决单样本、两样本时的均数比较问题,但真实的世界不可能总是如此简单,如例 14.1 就是比较复杂的情形。

例 14.1 CCSS 案例中提供了 2007 年 4 月,以及 2007 年、2008 年、2009 年 12 月 4 个时间点的消费者信心监测数据,现希望考察这 4 个时间点的消费者信心指数平均水平是否存在差异。

在本例中,所涉及的问题其实就是在单一处理因素之下,多个不同水平(或简单地理解为多组)之间的连续性观察值的比较,目的是通过对多个样本的研究来判断这些样本是否来自于同一总体。如果假设检验拒绝了多个样本来自于同一总体的 H_0 假设,研究者将更加关心这几个样本到底来自于几个不同的总体,而通过传统的 t 检验已经无法做到。

那么,能否使用两两 t 检验,例如在本例中做 4 组比较,则分别进行 6 次 t 检验来解决此问题?这样做在统计上是不妥的。因为统计学的结论都是概率性的,存在犯错误的可能。比如说,要用 6 次 t 检验来考察 4 个时点的信心指数均值是否相同,对于某一次比较,其犯一类错误的概率是 α ,则连续比较 6 次,其犯一类错误的概率不是 α^6 ,而是 $1 - (1 - \alpha)^6$ 。也就是说,如果检验水准取 0.05,那么在连续 6 次 t 检验中,犯一类错误的概率将上升为 0.264 9! 就好像考试及格线原本是 60 分,现在被降到了 20 分,导致考试的权威性大打折扣一样。因此,进行多个均数比较时不宜采用 t 检验做两两比较。

感谢 R. A. Fisher 爵士,他在 Rothamsted 试验站“下放”期间,为后人奠定了方差分析(Analysis of Variance, ANOVA)的理论基础:将总变异分解为由研究因素所造成的部分和由抽样误差所造成的部分,通过比较来自于不同部分的变异,借助 F 分布做出统计推断。后人又将线性模型的思想引入方差分析,更是为这一方法提供了近乎无穷的发展空间。

本章主要介绍单因素方差分析的基本原理及其在 SPSS 中的实现方式。在此基础上,给出方差分析的一些引申内容,包括多重比较、精细比较和趋势分析等。

14.1.2 方差分析的基本思想

方差分析是基于变异分解的思想进行的,在单因素方差分析中,整个样本的变异可以看成由如下两个部分构成:

总变异 = 随机变异 + 处理因素导致的变异

其中随机变异是永远存在的,确定处理因素导致的变异是否存在就是所要达到的研究目标,即只要能证明它不等于 0,就等同于证明了处理因素的确存在影响。

在方差分析中,代表变异大小,并用来进行变异分解的指标就是离均差平方和,代表总的变异程度,记为 SS_T 。可以发现在实际样本数据中,该总变异可以被分解为两项,第1项是各组内部的变异(组内变异),该变异只反映随机变异的大小,其大小可以用各组的离均差平方和之和,或称为组内平方和(Sum of Squares within Groups)来表示,记为 SS_W ;第2项为各组均数的差异(组间变异),它反映了随机变异的影响与可能存在的处理因素的影响之和,其大小可以用组间平方和(Sum of Squares Between Groups)来表示,记为 SS_B ,即

$$\text{总变异} = \text{组内变异} + \text{组间变异}$$

并且该等式和上面的等式存在着如下的对应关系:

$$\begin{array}{ccc} \text{总变异} = \text{随机变异} + \text{处理因素导致的变异} & & \\ \downarrow \quad \quad \quad \downarrow \quad \quad \quad \downarrow & & \\ \text{总变异} = \text{组内变异} & + & \text{组间变异} \end{array}$$

这样,可采用一定的方法来比较组内变异和组间变异的大小(用均方 MS 来比较),如果后者远远大于前者,则说明处理因素的影响的确存在,如果两者相差无几,则说明该影响不存在,以上就是方差分析法的基本思想。

方差分析的检验统计量可以简单地理解为利用随机误差作为尺度来衡量各组间的变异,即

$$F = \frac{\text{组间变异测量指标}}{\text{组内变异测量指标}}$$

可以想象,在 H_0 成立时,处理所造成的各组间均数的差异应为 0(理论上应为 0,但由于抽样误差不可能恰好为 0),即

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

于是,组间变异将主要由随机误差构成,即组间变异的值应当接近组内变异。于是检验统计量 F 值应当不会太大,且接近于 1;否则, F 值将会偏离 1,并且各组间的不一致程度越强, F 值越大。

下面介绍单因素方差分析的假设检验过程。

方差分析的零假设和备择假设分别为

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k;$$

$$H_1: k \text{ 个总体均数不同或者不全相同。}$$

沿用上面的变量标记方式,有检验统计量为

$$F_{k-1, N-k} = \frac{MS_B}{MS_W} = \frac{SS_B/(k-1)}{SS_W/(N-k)}$$

在上式中,检验统计量 F 的分子和分母上的平方和都除以一个数字:分子上除以了 $k-1$,而分母上除以了 $N-k$,这两个数字分别称为组间自由度和组内自由度,记作 v_B 和 v_W ,两者之和为 $N-1$,称为总自由度,记作 v_T 。分子上的组间平方和除以自由度后得到的数值称为组间均方(Mean Square Between Groups, MS_B),分母上组内平方和除以自由度后得到的数值称为组内均方(Mean Square Within Groups, MS_W)。进行分子、分母上的除法的出发点与为什么用标准差而非离均差平方和来描述资料离散程度的道理相同,即变异程度不当受样本含量的影响。显然样本含量越大 SS 就会越大,故需要扣除样本含量的影响,这样得到的比值才真正有可比性。

在零假设成立时, F 值应该服从自由度为 $k-1, N-k$ 的中心 F 分布 (Central F Distribution)。自由度为 1, 5 的 F 分布如图 14.1 所示。而若检验统计量落在相应检验水准所确定的拒绝域内 (即 F 值大于或等于相应自由度下的检验界值), 意味着在一次抽样研究中在假设总体内得到了小概率事件, 则有理由拒绝 H_0 , 其风险为相应 F 值所对应的 P 值。

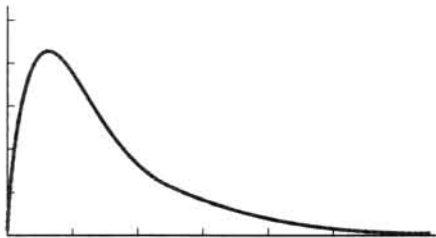


图 14.1 自由度为 1, 5 的 F 分布

在方差分析中常常将所计算出来的一些指标列成一张表格, 称为方差分析表 (Analysis of Variance Table), 如表 14.1 所示。

表 14.1 方差分析表

变异来源	离差平方和	自由度	均方	F	P
组间变异	SS_B	$k-1$	MS_B	MS_B/MS_W	$P = F_{k-1, k(n-1)} \geq F $
组内变异	SS_W	$k(n_i-1)$	MS_W		
总变异	SS_T	$N-1$	MS_T		

后面还会讲到, SPSS 的输出结果中将包含一张与此几乎完全相同的表格。
如果假设检验拒绝了 H_0 , 可以得出多个样本不是来自同一总体的结论。但是到底这些样本来自于几个不同的总体, 这次假设检验还不能回答这个问题, 而需要进一步进行单因素不同水平间的多重比较 (Multiple-Comparison), 详见后面讲解。

14.1.3 单因素方差分析的应用条件

1. 基本的应用条件

统计学中也许有成千上万的方法, 但没有哪种方法是通用的, 自然这里的方差分析也不例外。一般而言, 要应用方差分析, 数据应当满足以下几个条件, 或者说以下的假设应当成立。

- (1) 观察对象是来自于所研究因素的各个水平之下的独立随机抽样 (Independence)。
- (2) 每个水平下的应变量应当服从正态分布 (Normality)。
- (3) 各水平下的总体具有相同的方差 (Homoscedascity)。

其实, 与 t 检验的应用条件大同小异, 概括起来就是独立性、正态性和方差齐性。

2. 应用条件的检查与变量变换

以上适用条件可以使用统计描述进行观察, 或者绘制相应的统计图形, 当然也可以使用相应的检验方法。正态性检验的方法不再介绍, 这里简要介绍 3 个或 3 个以上样本的方差齐性检验方法。

(1) Bartlett 法: 其基本思想是比较各组方差的加权算术均数与几何均数, 若两者差异过大, 可以认为各组间的方差不齐。当各组样本含量均大于 5 时, 其检验统计量近似服从自由度为

$k-1$ 的卡方分布。

(2) Hartley 法:统计量 $H = \max(s_i^2) / \min(s_i^2)$, 当各组样本含量相同时可以使用此法。

(3) Cochran 法:统计量 $C = \max(s_i^2) / \sum_{i=1}^k s_i^2$, 该方法同样用于各组样本量相同的情况。

以上3种方法都要求所检验的样本来自于正态总体,而SPSS中所使用的是Levene法,这种方法对于正态性假设是稳定的。Levene法的基本思想是将各组变量值中心化后,利用 F 检验来检验各组间的差别。有兴趣的读者可以参考Levene(1960)的著作。

有时原始资料并不满足方差分析的要求,这时除了可以使用非参数检验方法外,也可以考虑进行变量变换(Transformation):通过对原始数据的数学变换,使其满足或者近似满足方差分析的要求。一般认为,对于通过变量变换达到方差齐性要求的资料,其正态性问题也会有所改善。常用的变量变换有以下几种。

(1) 对数转换(Logarithmic Transformation):将原始数据的自然对数值作为分析数据,其最常用的形式为 $y = \lg X$,也可选用 $y = \lg(X+k)$ 或 $y = \lg(k-X)$,当原始数据有0时,可用 $\lg(X+k)$ 进行数据转换,其中 k 为一小值。对数转换可用于:服从对数正态分布的资料;部分正偏态资料、等比资料,特别是各组的 S 与 \bar{X} 的比值相差不大(各组 CV 相近)的资料。

(2) 平方根转换(Square Root Transformation):可用于服从Poisson分布的资料、轻度偏态资料、样本的方差与均数呈正相关的资料以及观察变量为率、取值在0~20%或80%~100%范围内的资料。

(3) 平方根反正弦转换(Arcsine Transformation):将原始资料的平方根反正弦变换值 $y = \sin^{-1}\sqrt{X}$ 作为分析数据。平方根反正弦函数转换可用于原始数据为率且取值广泛的资料。

(4) 平方变换(Square Transformation):即将原始资料的平方作为分析数据。常用于方差与均数呈反比或资料呈左偏的情况。

(5) 倒数变换(Reciprocal Transformation):将原始资料的倒数作为分析数据。用于方差与均数的平方呈正比的情况,并且往往要求资料中没有接近或小于0的数据。

(6) Box-Cox变换:有时并不能很容易地找到一种合适的变换方式,Box和Cox于1964年提出如下的一类变换:

$$f(y) = \begin{cases} y^\lambda & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

研究者需要根据原始资料来尝试不同的 λ 值。实际上 λ 分别为-1、0、0.5、2时,Box-Cox变换分别等价于倒数变换、对数变换、平方根变换和平方变换。

此外,当观察指标为率且取值在30%~70%之间时,一般不考虑变量变换。

3. 应用条件得不到对方差分析结果的影响

(1) 独立性:举例来说,对于田间试验,两个区块中庄稼的产量差别应当仅仅与处理有关,而与两块地是否邻近无关;对于实验室研究,应当尽量避免由于试验者主观的系统误差而导致相关性。然而测量误差或者进行试验设计时的失误往往均会导致独立性的要求得不到满足,此时原始资料存在着信息“重叠”的现象,方差分析的结果往往会受到相当大的影响。因此在试验设计阶段就应当保证随机化真正得到实施。

(2) 正态性:Box和Anderson等人的研究表明,正态性得不到满足时,方差分析的结论并不

会受到太大的影响。也就是说,方差分析对于正态性的要求是稳健的。

(3) 方差齐性:在每组间样本含量相差不太大时,方差轻微不齐仅会对方差分析的结论有少许影响。一般而言,只要最大/最小方差之比小于3,分析结果都是稳定的。

应当注意的是,在方差分析中,各组在样本含量上的均衡性将会为分析计算提供极大的便利,也能在一定程度上弥补正态性或方差齐性得不到满足时对检验效能所产生的影响,这一点在多因素时体现得尤为明显。因此,在进行试验设计时就应当注意到均衡性的问题。

14.2 案例:不同时点消费者信心指数的比较

这里以例 14.1 为例演示在 SPSS 中进行方差分析的具体操作。

1. 预分析

注意,在进行方差分析之前,一定要注意其应用条件。利用均值过程可以得到各血型身高的一般描述。

(1) 选择“分析”→“比较均值”→“均值”菜单项,打开“均值”对话框。

(2) “因变量”列表框:index1。

(3) “自变量”列表框:time。

(4) 单击“确定”按钮。

从图 14.2 中可见,4 组的标准差相差不大,即方差可能是齐性的。

总指数			
月份	均值	N	标准差
200704	98.3363	300	18.92074
200712	94.1391	304	22.71719
200812	90.4393	304	20.59240
200912	101.9962	239	19.73351
总计	95.8935	1147	20.99710

图 14.2 报告

同时还可以使用箱图、直方图等工具考察数据的正态性、方差齐性,这里可以直接使用第 10 章图 10.19 所示的箱图分析结果,从中可知各组资料的正态性比较理想,未发现有明显方差不齐的迹象,虽然可以发现样本中有一些离散程度较大的数值,但情况并不是十分严重,因此可以考虑应用原始资料进行分析。

2. 界面说明

选择“分析”→“比较均值”→“单因素方差分析”菜单项,就可以打开“单因素方差分析”对话框,如图 14.3 所示,对其中的内容解释如下。

(1) “因变量列表”框:选入需要分析的变量,如果选入多个结果变量(应变变量),则系统会依次对其进行单因素方差分析。

(2) “因子”文本框:选入需要比较的分组因素,只能选入一个。

(3) “对比”按钮:单击后打开的对话框有两个用途,分别是对均数的变动趋势进行趋势检

验,以及定义根据研究目的需要进行的某些精确两两比较。该对话框太专业,也较少用,将在14.4节中加以介绍。

(4) “两两比较”按钮:单击后打开的对话框用于选择进行各组间两两比较的方法,详见14.3节的介绍。

(5) “选项”按钮:“统计量”框组提供了所需的一些统计量输出,“描述性”复选框用于指定输出描述性统计量;“固定和随机效果”复选框对于固定效应模型,输出其标准差、标准误和95%可信区间,对于随机效应模型,输出其标准误、95%可信区间及方差成分;“方差同质性检验”复选框用于指定进行方差齐性检验;“Brown-Forsythe”复选框用于指定输出用 Brown-Forsythe 法比较各组均数的统计量,适用于各组方差不齐时;“Welch”复选框用于指定输出用 Welch 法比较各组均数的统计量,适用于各组方差不齐时;“均值图”复选框用于指定输出各组均数的线图,以直观地显示它们的差异,同时可辅助对均数间的趋势做出判断。“缺失值”框组用于定义分析中对缺失值的处理方法,内容与前面介绍过的很多过程相同,不再赘述。

(6) “Bootstrap”按钮:单击后打开的对话框用于对相应输出的统计分析进行指定的 Bootstrap 抽样估计。该方法的详情已经在第7章中进行了介绍,因此这里不再重复。



图 14.3 “单因素方差分析”对话框

3. 操作步骤与结果解释

下面开始进行方差分析,操作步骤如下。

(1) 选择“分析”→“比较均值”→“单因素方差分析”菜单项,打开“单因素方差分析”对话框,进行如下设置。

- ① “因变量”列表框:总指数[index1]。
- ② “因子”文本框:月份[time]。
- (2) 单击“选项”按钮。
- (3) 在打开的对话框中选中“方差同质性检验”和“均值图”复选框。
- (4) 单击“继续”按钮。
- (5) 单击“确定”按钮。

图 14.4 给出的是方差齐性检验结果,Levene 法检验统计量为 1.929,在当前自由度下对应的 P 值为 0.123,可以认为样本所来自的总体满足方差齐性的要求。

总指数			
Levene 统计量	df1	df2	显著性
1.929	3	1143	.123

图 14.4 方差齐性检验

图 14.5 即为单因素方差分析的结果,第 1 列为变异的来源,分别是组间变异、组内变异和总变异。第 2、3、4 列分别为离均差平方和、自由度、均方,检验统计量 F 为 16.252,显著性 (Sig.) $P<0.001$ 。由此可以认为 4 个时点的消费者信心指数总体均值存在差异。

总指数					
	平方和	df	均方	F	显著性
组间	20670.426	3	6890.142	16.252	.000
组内	484575.873	1143	423.951		
总数	505246.298	1146			

图 14.5 ANOVA

各组间样本均数的折线图如图 14.6 所示,它可以更直观地展现各组样本的大小关系及其与相应的分组变量间的关系。从图 14.6 中可以很清楚地看出,在 2008 年年底之前,信心指数的平均水平是持续下跌的,随后在经济刺激计划的作用下开始上升,且在 2009 年年底超过初值。分析师随后所需要做的工作就是结合当时的宏观经济信息、政策背景等情况,尽力展开自己的想象能力,对该结果进行尽量合理和完美的诠释,该工作已经脱离了统计的范畴,此处不再赘述。

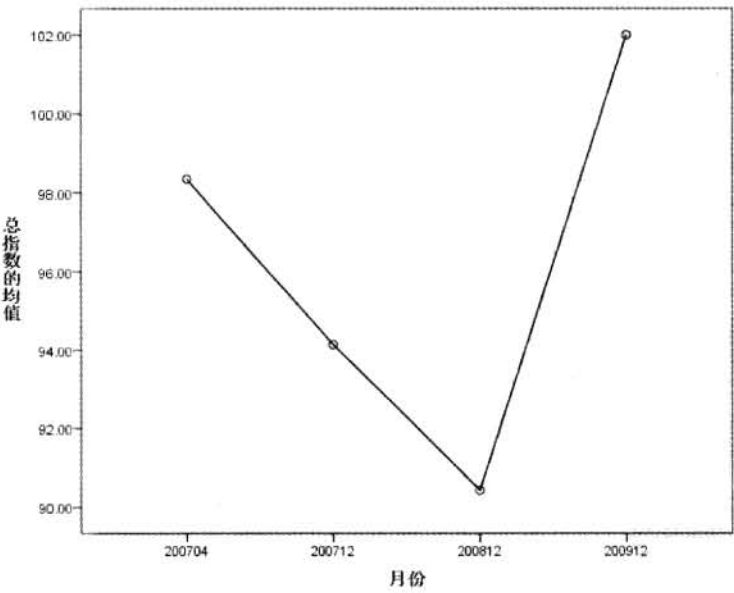


图 14.6 各组间样本均数的折线图

14.3 均数间的多重比较

在前面的分析中已经得到了拒绝 H_0 的结论,但实际上单因素方差分析并不这样简单,任务还没有最终完成:在解决实际问题时,往往仍需要确定多个均数中究竟哪些间存在差异。虽然结论提示各时点的信心指数不同,但研究者并不知道到底是4个时点之间均有差别,还是某一月份与其他月份之间有差别。尽管 Means Plot 可以让人们对大体的形式做到心中有数,但若是没有相应的假设检验结果用于说明这种样本均数的差别是否有推广至总体的意义,总归让人心里不太踏实,这时应当通过两两比较(或者说多重比较, Multiple Comparison)来对其进行考察。



从上面的叙述可以得知,如果方差分析的检验结果没有拒绝 H_0 ,则除非在研究设计中事先有计划,否则就不存在随后的两两比较问题,这一点一定要加以注意。

14.3.1 直接校正检验水准

现在问题又回到了两两比较上。显然,在进行两两比较时关键的问题就是如何控制好一类错误的大小,首先,对于两两比较中所遇到的一类错误,有以下几个概念需要了解。

(1) CER:比较误差,即每进行一次比较犯一类错误的概率。

(2) EERC:在完全无效假设下的试验误差率,即在 H_0 成立时做完全部比较所犯一类错误的概率。

(3) MEER:最大试验误差率,即在任何完全或部分无效假设下做完全部比较所犯一类错误的最大概率值。

如前所述,当无效假设实际上成立,各组均数无差别时, k 组完全两两比较的次数 $c = k(k-1)/2$,做完所有这些比较犯第一类错误的概率为 $1 - (1 - \alpha_{ij})^c$,此即 EERC,所做的方差分析的实质也就是控制 EERC 为所设定的水准。因此,进行一类错误控制时最直接的想法就是将总的 α 水准控制到 0.05,从而由上述公式反推得出每一个检验所使用的 $\alpha_{ij} = 1 - (1 - \alpha)^{1/c}$,这种校正方式称为 Sidak 校正。但是,这只是在无效假设成立的情况下才成立的校正方式,在多数实际问题中,都是有些组的均数相同,而有些组的均数不同,因此控制 MEER 更为合适。Bonferroni 不等式被广泛地用于此目的,它通过控制 CER,使得 MEER 被控制在所设定的水准以内,其公式为

$$CER = \alpha/c$$

只要 CER,即每次比较时使用的一类错误水准 α_{ij} 小于 α/c ,就可以保证 MEER 不会大于 α 。



Bonferroni 提出,如果在 α 水准上进行 c 次假设检验,当无效假设为真时,检验结果为至少有一次拒绝无效假设的累积 I 类错误概率 α' 不超过 $c \times \alpha$,即有不等式 $\alpha' < c \times \alpha$ 。因此可以重新选择 I 类错误概率水准 α ,以便使累积 I 类错误概率 $\alpha' = 0.05$,此即所谓的 Bonferroni 不等式。

实际上,可以简单地理解成 Sidak 校正认为各次比较的一类错误与总的一类错误概率间的关系为累乘,而 Bonferroni 校正则认为它们之间的关系是累加的,从而分别按照累乘和累加的方式对总的概率进行了解。

Bonferroni 校正等直接校正方法虽然可以解决两两比较的问题,但首先各次比较是分别进行的,使用上比较麻烦;其次,它保证的是 MEER 不会大于 α ,这显然意味着多数比较的检验水准实际上是小于 α 的,因而结论仍然比较保守。

14.3.2 专用的两两比较方法

除了相对粗糙的直接校正法外,针对不同的分析需求,统计学上还发展出了一系列专用的两两比较方法。一般而言,可以把多重比较分为两种类型:计划好的和非计划的。所谓计划好的多重比较(Planned Comparisons),即在收集数据之前便决定了要通过多重比较来考察多个组与某个特定组间的差别或者某几个特定组间彼此的差别;而非计划的多重比较(Unplanned Comparisons, Post-Hoc Comparisons)只有在方差分析得到有统计学意义的 F 值后才有必要进行,是一种探索性的分析。前者需要通过“对比”按钮的有关内容来进行,而后者则要借助于“两两比较”按钮了。

单击“两两比较”按钮,将打开如图 14.7 所示的对话框。



图 14.7 多重比较的选择对话框

从图 14.7 所示的对话框中可以看出,在“假定方差齐性”框组内有 14 种两两比较的方法!这并不是说两两比较的方法如百花齐放般衬托了统计学的欣欣向荣,相反却说明目前为止仍然没有令人完全信服的方法或者没有统一的解决之道。便如流感,治疗的药物很多,却没有一种真正有效的药物,而大叶性肺炎却是普通的青霉素就可以治疗好的。

对于非计划的多重比较,随着比较目的和应用条件的不同,各种多重比较方法也有其不同的侧重点,以下简要介绍常用的几种多重比较的方法。

(1) LSD 法:即最小显著差法(Least-Significance-Difference Method),是最简单的比较方法之一。它其实只是 t 检验的一个简单变形,并未对检验水准做出任何校正,只是在标准误的计算中充分利用了样本信息,为所有组的均数统一估计出了一个更为稳健的标准误,因此它一般用于计划好的多重比较。由于单次比较的检验水准仍为 α ,因此可以认为 LSD 法是最灵敏的。

(2) Sidak 法:它实际上就是 Sidak 校正 LSD 法上的应用,也即通过 Sidak 校正降低每次两

两比较的一类错误率,以达到最终整个比较的一类错误率为 α 的目的。但是,由于在统计分析中习惯上将每次比较的水准都定为0.05,为符合阅读习惯,统计软件往往采用倒乘的方式,即固定检验水准,将检验的 P 值进行反向放大。例如当需要进行 C 次比较时,对于相同的比较,Sidak法的 P 值和LSD法的 P 值间的关系为 $P_{\text{Sidak}} = 1 - (1 - P_{\text{LSD}})^C$ 。显然,Sidak法要比LSD法保守得多。

(3) Bonferroni法:和Sidak法类似,它的每一次比较实际上是Bonferroni校正在LSD法上的应用,对于相同的比较,Bonferroni法的 P 值和LSD法的 P 值间的关系为 $P_{\text{Bonferroni}} = P_{\text{LSD}} \times C$ 。一般而言,Bonferroni要比Sidak法更为保守一些。它也是列联表检验中SPSS使用的检验方法。

(4) Scheffe法:与一般的多重比较不同,Scheffe法的实质是对多组均数间的线性组合是否为0进行假设检验(也即所谓的Contrast)。多用于进行比较的两组样本含量不等的情况,详见后面介绍。

(5) Dunnett法:常用于多个试验组与一个对照组间的比较。因此在指定Dunnett法时,还应当指定对照组。

以上几种方法的排列顺序大致是从最灵敏到最保守,除了这几种方法以外,还有另外一大类用于寻找同质亚组的检验方法,常用的有以下几个。

(1) S-N-K法:经常在有关统计学教材中出现,全称为Student-Newman-Keuls法。它实质上是根据预先指定的准则将各组均数分为多个子集,利用studentized range分布来进行假设检验,并根据所要检验的均数的个数调整总的一类错误概率不超过 α 。

(2) Tukey法:即Tukey's Honestly Significant Difference法,应用这种方法要求各组样本含量相同。它也是利用studentized range分布来进行各组均数间的比较的,与SNK法不同的是,它用于控制所有比较中最大的一类错误的概率,即MEER不超过 α 。

(3) Duncan法:其思路与SNK法相类似,只不过检验统计量服从的是Duncan's Multiple Range分布。

剩下的一些方法并不常用,本书中不再阐述。此外,在每组方差不齐时,SPSS在“未假定方差齐性”框组中也给出了4种方法。但从方法的接受程度和结果的稳健性角度讲,建议尽量不要在方差不齐时进行方差分析甚至两两比较,进行变量变换或者非参数检验往往更可靠。

在图14.4所示对话框的“显著性水平”文本框中还可以定义多重比较的检验水准,一般而言,默认的0.05足以满足要求。

14.3.3 两两比较方法的选择策略

由于两两比较方法非常多,很多统计学家对进行方差分析后两两比较的策略均提出了自己的看法,国内也有多篇文献对不同的方法进行了比较。以下是笔者参考多本参考书后的心得,仅供参考。

(1) 如两个均数间的比较是独立的,或者虽有多个样本均数,但事先已计划好要进行某几对均数的比较,则不管方差分析的结果如何,均应进行比较。一般采用LSD法或Bonferroni法。


(2) 如果事先未计划进行多重比较,在进行方差分析得到有统计学意义的 F 值之后,可以利用多重比较进行探索性数据分析。此时两两比较方法要根据研究的目的和样本的性质选择。比如说,需要进行多个试验组和一个对照组的比较时,可以采用Dunnett法;需要进行任意两组之

间的比较而各组样本含量又相同时,可以选用 Tukey 法;当样本含量彼此不同时,可以采用 Scheffe 法。而若是事先未计划进行多重比较,且进行方差分析未检出差别,此时不应当进行多重比较。

(3) 绘制均值图,或者进行详细的统计描述有利无弊。


(4) 事先未计划的多重比较,各组间的差别只是一种提示,要确认这种差别最好重新设计实验。

有的时候,研究者在进行试验设计之初就考虑了比较特定的几组均数,这种比较往往不像 Post Hoc 那样需要对几乎所有的组合进行比较,所以在进行相应的统计分析时不需要对检验水准或统计量进行太多修正。计划好的比较(Planned Comparison),或者称为事前比较(Prior Comparison),主要是通过“单因素方差分析”对话框中的“对比”按钮所对应的功能来实现的。14.4 节中将会详细介绍 Planned Comparison 的实现方法。

 需要提醒的是,如果组数较少,如 3 组、4 组,比较方法的选择可能结果差异不大,但如果组数很多,则一定要慎重选择两两比较方法。

14.3.4 多重比较结果出现矛盾时的解释

多重比较可能会出现看上去似乎存在矛盾的结论,即样本 1 与样本 2 差异无统计学意义,样本 2 与样本 3 差异无统计学意义,但样本 1 与样本 3 差异却有统计学意义。对于这种情形,只能说两两比较还不能判明样本 2 来自何总体,而以下两种解释都是错误的:①“样本 2 所代表的总体介于总体 1 和总体 3 之间。”这种结论实际上已经默认了 3 个样本分别来自 3 个不同的总体;②“既然样本 1 与样本 2 差异无统计学意义,样本 2 与样本 3 差异无统计学意义,所以样本 1 与样本 3 差异也没有统计学意义。”须知抽样误差是不能递推的,否则将引导出荒唐的结论。

 一个经典案例可以恰如其分地说明上述推理逻辑的荒唐性:头上一根头发都没有的人毫无疑问是秃子,头上有一根头发的人和一根头发都没有的人之间看不出什么差别(差别无统计学意义),所以也是秃子,以此类推,最后会得到一个满头黑发的人也是一个秃子的荒谬结论!

有时,方差分析拒绝 H_0 ,但方差分析后的两两比较却找不到有差异的任何两个样本。在 14.4 节中引入对比的概念后,方差分析中的这一个特殊现象就可以很容易地解释了。这是因为方差分析的差别有统计学意义时仅仅保证诸多对比中的某一个或某几个不为 0,但这些对比却不一定是人们所关心的。此时下结论应当极为谨慎,最好的方法是增加样本含量重新进行试验。

14.3.5 案例:不同时点信心指数的两两比较

这里继续对例 14.1 进行分析,考察在 0.05 的显著性水平下,究竟这 4 个时点的总指数均值之间存在怎样的差异。这是一个非计划的多重比较(Post Hoc),由于各组样本含量不同,因此在多重比较的对话框中选择“Scheffe”选项,相应的分析结果如图 14.8 所示。

总指数
Scheffe

(I) 月份	(J) 月份	均值差 (I - J)	标准误	显著性	95% 置信区间	
					下限	上限
200704	200712	4. 19721	1. 67563	. 100	-. 4940	8. 8884
	200812	7. 89700 *	1. 67563	. 000	3. 2058	12. 5882
	200912	- 3. 65990	1. 78522	. 241	- 8. 6579	1. 3381
200712	200704	- 4. 19721	1. 67563	. 100	- 8. 8884	. 4940
	200812	3. 69979	1. 67008	. 179	-. 9758	8. 3754
	200912	- 7. 85711 *	1. 78001	. 000	- 12. 8405	- 2. 8737
200812	200704	- 7. 89700 *	1. 67563	. 000	- 12. 5882	- 3. 2058
	200712	- 3. 69979	1. 67008	. 179	- 8. 3754	. 9758
	200912	- 11. 55690 *	1. 78001	. 000	- 16. 5403	- 6. 5735
200912	200704	3. 65990	1. 78522	. 241	- 1. 3381	8. 6579
	200712	7. 85711 *	1. 78001	. 000	2. 8737	12. 8405
	200812	11. 55690 *	1. 78001	. 000	6. 5735	16. 5403

*. 均值差的显著性水平为 0.05。

图 14.8 多重比较

由于这些多重比较方法都需要有一个对照组,在分析结果中就将所有组依次作为对照组,和其余各组进行比较。图 14.8 中依次给出的是两组间均数差值、差值的标准误、P 值以及差值的可信区间。其中,如果均数差别有统计学意义,则自动在后面加上“*”作为标记。

显然,上述两两比较的输出虽然详细,但阅读起来令人头晕,因此 Scheffe 方法也提供了类似于 S-N-K 等方法的输出格式,如图 14.9 所示。

Scheffe^{a,b}

月份	N	alpha = 0.05 的子集		
		1	2	3
200812	304	90. 4393		
200712	304	94. 1391	94. 1391	
200704	300		98. 3363	98. 3363
200912	239			101. 9962
显著性		. 206	. 117	. 214

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 283. 761。

b. 组大小不相等。将使用组大小的调和均值。不保证 I 类错误级别。

图 14.9 类似于 S-N-K 等方法的输出格式

S-N-K 这一类方法的目的是寻找同质子集 (Homogeneous Subsets), 简单地说, 各组首先在表

格的纵向上按均数大小排序,然后即根据多重比较的结果将所有的组分为若干个子集,子集之间的各组间有差别(P 值小于0.05),子集之内的各组间无差别。Scheffe 方法的输出结果如果采取这种输出方式,则可以很清楚地看出,4 个月份之间的总指数大致可以被分为3组,但这3组之间存在重叠。此时需要利用表格最后一行的输出,这里会给出子集内部各组比较的 P 值,可见第2组,即2007年12月和4月数值比较的 P 值最小,为0.117,因此出于实际应用的考虑,可以将结论设定为上述4个时点可以被分为两个层次,2007年12月~2008年12月为谷底,而2009年12月以及2007年4月则为高峰。也就是说,信心指数在3年的时间里走出了—个U型,而在2009年12月已经恢复到基线水平。

14.4 各组均数的精细比较

14.4.1 方法原理*

前面所讲的多重比较实际上都可以归结为对均数的线性组合 $L = a_1\mu_1 + a_2\mu_2 + a_3\mu_3$ 的假设检验,其中 a_1, a_2, a_3 是研究者指定的常数。于是,若 a_1, a_2, a_3 分别为1、-1、0,则 $L = \mu_1 - \mu_2$ 。若对假设 $L=0$ 进行假设检验,则等价于前面所述的第1组和第2组均数是否相等的两两比较。同样,要比较第1组和第3组是否相等,只需要对 a_1, a_2, a_3 分别为1、0、-1时的线性组合是否为0进行检验就可以了。

不失一般性,如果将现有的样本分为 k 组,则表达式

$$L = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k$$

称为 k 个均数的对比(Contrast),其中 a_1, a_2, \cdots, a_k 为任意指定的常数。两个对比如下:

$$L = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k$$

$$L' = a'_1\mu_1 + a'_2\mu_2 + \cdots + a'_k\mu_k$$

如果满足 $a_1a'_1 + a_2a'_2 + \cdots + a_ka'_k = 0$,则称之为正交的(Orthogonal),对于样本均数,其线性组合为

$$\hat{L} = a_1y_{1.} + a_2y_{2.} + \cdots + a_ky_{k.}$$

是总体均数相应的线性组合的无偏估计(Unbiased Estimator)。

在引入正交和对比的概念后,便可以不再被束缚于简单的两两比较,而是可以通过指定 a_i 的值完成更多、更复杂的比较。根据方差分解的有关原理,组间变异可以分解为由 $k-1$ 个正交对比所能解释的部分,即总变异就可以分解为由 $k-1$ 个正交对比所能解释的变异和一个组内变异,即

$$SS_T = SS_{L_1} + SS_{L_2} + \cdots + SS_{L_{k-1}} + SS_W$$

例如,对于4组样本,对比 $\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$,此时 $a_1 = a_2 = \frac{1}{2}, a_3 = a_4 = -\frac{1}{2}$;如果对比 $\frac{\mu_1 + \mu_3}{2} =$

$\frac{\mu_2 + \mu_4}{2}$,此时, $a'_1 = a'_3 = \frac{1}{2}, a'_2 = a'_4 = -\frac{1}{2}$,且有

* 本小节理论深度较高,基础较差的读者可跳过此部分,不影响后续内容的理解。

$$a_1 a'_1 + a_2 a'_2 + a_3 a'_3 + a_4 a'_4 = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \left(-\frac{1}{2}\right) + \left(-\frac{1}{2}\right) \times \frac{1}{2} + \left(-\frac{1}{2}\right) \times \left(-\frac{1}{2}\right) = 0$$

因此对比 $L_1 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$ 和 $L_2 = \frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$ 间是正交的。此

时便可以对诸如 $H_0: \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$, 或者 $H_0: \frac{\mu_1 + \mu_3}{2} = \frac{\mu_2 + \mu_4}{2}$ 之类的假设进行检验, 甚至连 $H_0:$

$\mu_1 = \frac{\mu_2 + \mu_3 + \mu_4}{3}$ 这样的假设检验也可以完成!

14.4.2 案例: 事先计划的两时点均数比较

14.4.1 节中提到了计划好的比较, 本节将通过一个例子来说明它的实现方法。对于 CCSS_Sample.sav 文件中的总指数与月份之间关系的案例, 假设将事前计划好了 2007 年 4 月的总指数与 2009 年 12 月的总指数进行比较, 则若以 $\mu_1, \mu_2, \mu_3, \mu_4$ 分别表示 200704、200712、200812、200912 的总指数, 计划好的比较实质上是用于检验下列等式是否成立。

$$a_1\mu_1 + a_2\mu_2 + a_3\mu_3 + a_4\mu_4 = 0, a_1 = 1, a_2 = 0, a_3 = 0, a_4 = -1$$

1. 界面说明

对比子对话框(图 14.10)中的各选项功能介绍如下。

(1) “多项式”复选框: 定义是否在方差分析中进行趋势检验, 即随着组别的变化, 各组均数是否呈现某种变化趋势。

(2) “度”(Degree)下拉列表框: 和“多项式”复选框配合使用, 用于定义需检验的趋势曲线的最高次方项, 可选择从线性趋势一直到 5 次方曲线。如果选择了高次方曲线, 系统会给出所有相应的各低次方曲线的拟合优度检验结果(比如选择 3 次方曲线时, 系统会给出线性、二次方、三次方 3 个结果)以供选择。

(3) “系数”文本框: 精确定义某些组间均数的比较。这里按照分组变量升序给每组一个系数值, 注意最终所有系数值相加应为 0。例如在上例中要对第 1、3 组进行单独比较, 则在这里给 3 组分别分配系数为 1、0、-1, 就会在结果中给出相应的检验内容。这里可以同时进行多组比较系数的设定, 只需要用“上一张”和“下一张”按钮翻页即可。

(4) “系数总计”信息栏: 动态提供输入系数的总和, 以免因用户疏忽而导致系数总和不



所有系数值相加不为 0 时仍可以得到检验结果, 但 SPSS 不推荐这样做! 因为此时该检验的适用条件已被违反, 其结果的准确性可疑, 分析结论仅供参考。在 SPSS 的帮助文件中对此有明确的说明。

2. 结果解释

按照图 14.10 中的设定方式, 可以得到如图 14.11 所示的分析结果。

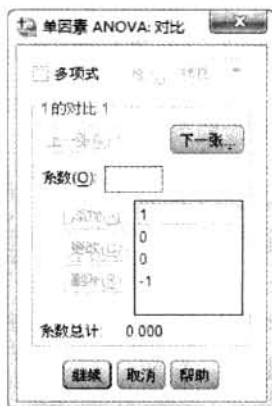


图 14.10 对比子对话框

对比	月份			
	200704	200712	200812	200912
1	1	0	0	-1

图 14.11 对比系数

首先输出的是对比系数的列表,该列表可用于查错,从图 14.10 中的系数设定就可得知,随后将会进行的是 2007 年 4 月和 2009 年 12 月数值的比较。

图 14.12 分别针对相比较的两组间方差齐和不齐的情形给出了比较的结果。其中“对比”列中指明了所比较的对子,“对比值”列给出了所要检验的对比的实际值(这里就是相比较的两组间的均数之差),“标准误”列中给出了均数之差的标准误,“t”、“df”、“显著性(双侧)”列中分别列出了检验统计量值、自由度和双侧 P 值。本例中按照方差齐性时的比较结果,2007 年 4 月(200704)的总指数与 2009 年 12 月(200912)的总指数的 $P=0.041$,按 0.05 的检验水准均拒绝 H_0 ,得到了有差别的结论。

		对比	对比值	标准误	t	df	显著性(双侧)
总指数	假设方差相等	1	-3.6599	1.78522	-2.050	1143	.041
	不假设等方差	1	-3.6599	1.68007	-2.178	500.560	.030

图 14.12 对比检验



细心的读者会发现,这里得到了和前面似乎矛盾的结果。在 Scheffe 法中,这两个时点是无统计学差异的,但是在这里的对比中却有差异!这是因为比较方法的适用条件不同,Scheffe 适用于无事先计划的比较,对一类错误控制得比较严格。而这里的对比方法实质上等价于 LSD 法,本质上并未对一类错误做任何控制,因此只适用于有事先计划的比较。简言之,统计软件只提供所需的各种分析结果,而从中正确选择所需的内容则应当是使用者自己的责任。

14.5 组间均数的趋势检验

14.5.1 方法原理

从理论上讲,方差分析所对应的分组变量应该是一个无序分类的变量。但实际上,分组变量的取值往往也可以体现顺序的意义,比如,多个时间点上的某个指标的比较;不同 pH 值下某些化学物质转化率的比较等。这一类型的资料并不少见。

对于这类资料,既然是多组间计量资料的比较,当然是优先考虑单因素方差分析。但是在各位得到各组间有差异的结论时,不得不提醒大家:单纯的方差分析并未利用到分组变量中蕴涵的次序信息。

在线性模型的方法被引入方差分析之前,对于有序分组信息的问题有一些折衷的解决方法,

如 Spearman 相关、Kendall τ 等。但当自变量各取值间间隔相等时,现在除了对此进行方差分析之外,还可以利用线性模型的有关原理对数据进行进一步的分析,以考察应变变量与处理之间是否存在某种依存关系,统计学上称为趋势检验(Trend Analysis)。这种趋势并非仅仅指线性的,也可能为一种多项式关系。

一般而言,对于趋势检验,首先考虑的是应变变量和分组变量之间的线性关系,即检验模型 $Y = b_0 + b_1X$ 是否成立。然而,从本例中也可以看到,应变变量与分组变量间并不呈现线性的趋势,有可能呈二项式关系甚至三项式关系,即 $Y = b_0 + b_1X + b_2X^2$, 或者是 $Y = b_0 + b_1X + b_2X^2 + b_3X^3$ 。对于这类模型,要选择相对比较合适的模型,利用失拟合检验(Lack of Fit Test)可以达到这样的目的。然而可以想象,一次项、二次项、三次甚至更高次项之间肯定存在着相关性,这对最后的结果解释是不利的。因此,一般通过建立正交多项式(Orthogonal Polynomials)模型的方法来进行趋势检验。关于正交多项式模型本章中不进行过多阐述,感兴趣的读者可以参考有关的著作。但是需要指出的是,趋势检验的目的并不是拟合线性或者非线性模型,而是希望知道当因素的水平改变时,均数以什么样的形式(线性、二次性或者其他)随之改变。

对于趋势分析,可以利用正交多项式的方法得到 $k-1$ 个正交的 Contrast,分别对应于一次多项式(线性),二次多项式,三次多项式, ..., $k-1$ 次多项式,然后再将总变异分解为由这 $k-1$ 个 Contrast 所能解释的部分和一个剩余变异(Lack of Fit Test 中常称为纯误差, Pure Error),再利用方差分析得到相应的结论。

14.5.2 案例:前3个时点的信心指数线性趋势检验

在例 14.1 分析结果的均数图中可以发现,前3个时点的均数似乎处在一条直线上,可能存在线性关系。下面就通过“对比”子对话框来考察这一假设是否成立,注意这里首先需要选择所需的3个时点的记录进行分析,操作步骤如下。



注意这里由于只使用了一部分样本进行分析,因此方差分析的输出并无分析意义,分析者只需要关注趋势检验的结果即可。

- (1) 选择“数据”→“选择个案”菜单项。
- (2) 在打开的对话框中选择框组:如果条件满足。
- (3) 单击“如果”按钮,在打开的对话框中设置 $\text{time} < 200912$,单击“继续”按钮。
- (4) 单击“确定”按钮。
- (5) 选择“分析”→“比较均值”→“单因素方差分析”菜单项,打开“单因素方差分析”对话框。
- (6) 在“因变量列表”框中添加“总指数[index1]”选项。
- (7) 在“因子”列表框中添加“月份[time]”选项。
- (8) 单击“对比”按钮。
- (9) 在打开的对话框中选中“多项式”复选框。
- (10) 在“度”下拉列表框中选择“二次项”选项。
- (11) 单击“继续”按钮。
- (12) 单击“确定”按钮。

图 14.13 所示的方差分析表中分别对均数变化趋势是否服从线性、二次多项式的方程进行了检验,可见线性项、二次项检验所对应的 P 值均小于 0.05,因此,从统计检验的角度来看,总指数和月份之间的关系不能用线性关系来描述。但这里需要说明的是,虽然此处的检验结果拒绝了相应的趋势假设。但是多项式形式的判定并不是完全靠 P 值来决定的,许多时候还要依靠图形、专业知识以及经验来给出结论。

总指数			平方和	df	均方	F	显著性
组间	(组合)		9426.352	2	4713.176	10.884	.000
	线性项	加权的	7309.948	1	7309.948	16.881	.000
		偏差	2116.404	1	2116.404	4.887	.027
	二次项	加权的	2116.407	1	2116.407	4.887	.027
组内			391895.938	905	433.034		
总数			401322.290	907			

图 14.13 ANOVA



表 14.9 中的“加权的”输出指的是对样本是否符合当前模型假设进行的检验,而“偏差”则指的是当前模型和含有最高次项的模型相比是否有区别,换言之,就是当前模型是否还需要继续增加高次项。

14.6 本章小结

(1) 单因素方差分析所针对的是多组均数间的比较。它的基本思想是变异分解,即将总变异分解为组间变异和组内变异,再利用 F 分布做出有关的统计推断。

(2) 单因素方差分析要求资料满足正态性、独立性和方差齐性的要求。

(3) 方差分析拒绝 H_0 只能说明各组之间存在差异,但并不足以说明各组之间的关系。利用多重比较可以初步判断各组间的关系。

(4) 多重比较可以分为事前计划好的比较和事后比较。前者往往借助于 Contrast,而后者有很多种不同的方法,这些方法的核心问题都是如何控制总的一类错误的大小。

(5) 在分组变量包含次序信息时,如果方差分析给出了各组间差异有统计学意义的结论,并且 Means-Plot 提示各组均数的某种趋势时,可以利用趋势分析探讨观察值与分组变量取值间的数量依存关系。

思考与练习

1. 一家汽车厂设计出 3 种新型号的手刹,现欲比较它们与传统手刹的寿命。分别在传统手刹,型号 I、II 和型号 III 中随机选取了 5 只样品,在相同的试验条件下,测量其使用寿命(单位:月),结果如下。

传统手刹: 21.2 13.4 17.0 15.2 12.0

型号 I : 21.4 12.0 15.0 18.9 24.5

型号 II : 15.2 19.1 14.2 16.5 20.3

型号 III : 38.7 35.8 39.0 32.2 29.6

(1) 各种型号间寿命有无差别?

(2) 厂家的研究人员在研究设计阶段,便关心型号 III 与传统手刹寿命的比较结果。此时应当考虑什么样的分析方法? 如何利用 SPSS 实现?

(3) 如果方差分析拒绝了 H_0 , 要考虑多重比较吗? 利用 SPSS 尝试一些多重比较, 并解释结果。

2. 研究者要比较 4 种新型避孕药对雌激素分泌水平的影响。试验对象为相同品系的雌性大鼠, 将 20 只大鼠随机分入 4 组中, 给予相应的药物, 两周后通过测量大鼠的子宫重量来衡量其雌激素水平。试验数据如下:

药物 1: 89.8 93.8 88.4 110.2 95.6

药物 2: 84.4 116.0 84.0 68.0 88.5

药物 3: 65.6 79.4 65.6 70.2 82.0

药物 4: 88.4 90.2 73.2 87.7 85.6

(1) 该数据是否满足方差分析的要求?

(2) 4 种药物对雌激素水平的影响是否相同?

(3) 是否要考虑一些多重比较? 利用 SPSS 尝试一些多重比较, 并解释结果。

第 15 章 有序分类变量的统计推断——非参数检验

通过第 14 章的学习可知,如果想要检验两个正态总体是否具有相同的均数,做一个 t 检验即可,这是一个典型的参数统计方法。参数统计方法往往假设统计总体的分布形态已知,但是在更多的实际场合,常常由于缺乏足够信息,无法合理地去假设一个总体具有某种分布形式,此时就不能去使用相应的参数方法了。推而广之,不能使用参数方法的情形可能是:当不知道所研究样本来自总体的具体分布时,或已知总体分布与检验所要求的条件不符;数据的测量尺度是名义和顺序尺度,甚至某些变量可能无法精确测量,均值、方差的计算已经没有意义时……但是,此时有的人却忽略参数统计方法的前提,仍然牵强地使用参数方法,面对由此得到的不合理结果却不知问题何在。实际上,正确的思路应当是放弃对总体分布参数的依赖,转而寻求更多的纯粹来自数据的信息,这就是所谓的非参数统计方法。

在第 12 章中其实已经讲到二项分布检验、单样本 $K-S$ 检验等简单的非参数方法,事实上非参数检验的方法层出不穷,其根本的技术核心在于针对简单的数据样本,充分挖掘利用样本信息构造的别出心裁的检验统计量。熟悉并体会这些变化,对于理解统计要素意义非凡,对于培养统计的直观能力也是一个很好的训练。

本章将针对不同的设计类型,以秩统计量为基础着重介绍采用秩和检验对样本分布位置进行检验的非参数方法。



出于方法讲解的目的,本章采用的个别案例的样本量较小,且并不一定明显违反参数方法适用的条件。

15.1 非参数检验概述

15.1.1 非参数检验的意义

在现实生活中,从生活经验到经济活动乃至政策制定和评价,很多时候需要选择、比较、决策,小至柴米油盐品牌的不同偏好,百姓对未来生活的预期,公司在决定是否给雇员加薪时对雇员能力进行的考核,企业扩张对于新销售处的选址,……大至政治竞选中对候选人的民意调查等问题都可以借助统计方法对样本数据进行有益的判断分析,但是任何方法都是有前提的,17 世纪犹太籍哲学家史宾诺莎强调整理解是自由之道(他有句广为传颂的格言:“不要哭,不要笑,要理解”)。各种数据资料中隐藏的信息是有助于人们的理解的,当所熟悉的方法失效时,应该转而使用新的方法代替!

非参数统计方法主要用于那些总体分布不能用有限个实参数来刻画,或者不考虑被研究的对象为何种分布以及分布是否已知的情形,它对总体分布几乎没有什么假定,只是有时对分布的形状做一些诸如连续、对称等的简单假设。顾名思义,这种检验方法的着眼点不是总体的有关参

数的比较,其推断方法和总体分布无关(Distribution Free),它们进行的并非是参数间的比较,而是分布位置、分布形状之间的比较,研究目标总体与理论总体分布之间的比较,或者各样本所在总体的分布位置之间的比较等,因此不受总体分布的限制,适用范围广,故而称为非参数检验。但这个名称很容易让人产生误解,它指的是推断过程和结论均与原总体参数无关,并非说在推断中什么分布参数都不利用,事实上,最常用的秩和检验就是基于秩次的分布特征推导出来的,即可能会用到秩分布的参数。所以有学者提出将中文名称改为分布自由检验可能更为妥当。

非参数检验依然遵循于假设检验的基本思想和基本准则,在缺乏总体分布信息的支撑下,利用统计思想、数学方法和技巧构造相应的统计量进行检验,拓宽了人们的分析领域,将统计方法的魅力施展到一个更广阔的空间。

和参数方法相比,非参数检验方法的优势如下。

(1) 稳健性。因为对总体分布的约束条件大大放宽,不至于因为统计中的假设过分理想化而无法切合实际情况,不至于对个别偏离较大的数据太敏感。

(2) 对数据的测量尺度无约束,对数据的要求也不严格,什么数据类型都可以做。

(3) 适用于小样本、无分布样本、数据污染样本、混杂样本等。



由于非参数统计推断对总体的要求和假设较少,也许有人会问:为什么不一直使用它,而忘记参数检验呢?当掌握了这些检验方法,领悟了它们的统计思想后,就会给出答案。

这里举一个非参数方法的实际应用案例:在股票市场上存在周末效应,即股市中周一的收益率比其他交易日的收益率低,且风险较大;周五的收益率比其他交易日高,且风险相对较小。但是,国内对周末效应的研究存在一些缺陷,如研究的样本区间较短;未考虑我国股市收益率的分布特征,从而忽视许多检验模型的正态分布假设前提,对非正态分布的数据进行了正态分布下的研究;对收益率的风险分析不足等。因此有学者利用非参数方法对我国沪市的周末效应进行了验证:首先采用K-S检验得出我国股市收益率的非正态性,再利用K-W检验股票指数收益率周末效应的存在性得出股市一周内各天的收益率存在显著差异,可是周末效应的模式如何呢?也就是异常收益率存在于一周中的周几?于是利用Mann-Whitney检验两两比较来分析发现周二与周五的收益率差异最为显著,认为沪市存在“二、五”效应。这里提到的方法就是本章中要重点介绍的,感兴趣的读者可以在更深入地学习了经典的非参数方法后进行验证!

15.1.2 非参数检验预备知识

(1) 心中有数:当获取了数据后,首先要对它进行充分、直观的了解,使用直方图、茎叶图、箱图、QQ图等可以对数据的分布形状进行探索,避免因对数据的特性缺乏了解而盲目使用一些方法得出错误的或不合理的结论。记住,在统计分析中数据的预处理很重要!

(2) 顺序统计量:因为非参数统计方法并不假定总体分布,因此往往把观察值的顺序及其性质作为研究对象,只利用大小间的次序关系,而不利用具体的数值信息。正是由于这一特点,非参数方法中的秩和检验实际上成为有序分类资料的标准分析方法。对于样本数据 X_1, \dots, X_n ,如果将其按升幂排列,则可以得到

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(i)} \leq \dots \leq X_{(n)}$$

以上次序就是顺序统计量,其中 $X_{(i)}$ 为第 i 个顺序统计量,对它的性质的研究就构成非参数统计的理论基础之一。

(3) 秩(Rank)及秩统计量:对于样本 X_1, \dots, X_n , 按由小到大的顺序排成一列,若 X_i 在这一列中占据第 R_i 位,称 X_i 的秩为 R_i , $R_i = \sum_{j=1}^n I(X_j \leq X_i)$, 即小于或等于 X_i 的样本点个数,称 $R = (R_1, \dots, R_n)$ 是原样本的秩统计量。实际上考试成绩的排名就是一个最简单的秩,只是倒了过来,最大的被排在了第一位。而这里所讲的秩应当对应着倒数的名次,如倒数第一、倒数第二……

(4) 结(Ties)和结统计量:在许多情况下,数据中会有相同的值出现,此时如果排秩就会出现同秩的现象,就像考试排名中的并列第5、并列第7,这种情况称为数据中的结。结中数值的秩为它们按大小顺序排列后所处位置的平均值。结统计量用 τ_i 表示,为第 i 个结中的观察值数量。例如,数据 2, 2, 5, 7, 7, 7, 10, 该数据序列一共有两个结: $\tau_1 = 2, \tau_2 = 3$ 。相应数据的秩分别为 1.5, 1.5, 3, 5, 5, 5, 7。

结的修正与否将影响到检验的结果,但对于这一点不用过分担心,因为统计软件会自动完成这些工作。

15.2 两个配对样本的非参数检验

出于难度由浅入深的考虑,本章将首先介绍配对样本的非参数检验方法。

15.2.1 方法原理

事实上,配对样本的非参数检验方法的基本逻辑和参数检验并无区别,也是首先求出配对数据的差值,然后考察差值总体的中心位置是否为0。只是由于不再涉及分布类型,因此不能使用均数这一与总体分布有关的参数加以检验。一般而言,相应的假设都被归结为考察总体中位数是否为0。

H_0 : 差值的总体中位数 $M_d = 0$; H_1 : 差值的总体中位数 $M_d \neq 0$ 。

但是,仅有假设是不够的,还需要找到一个合适的统计量。为了构建统计量,统计学家们想出了各种各样的方法,下面就来依次介绍。

1. Sign 符号检验

符号检验可以说是最早被提出来的非参数统计方法,其原理是如果两个配对样本实际上无区别,则将样本数据相减所得的差值应当大致有一半为正,一般为负,数量基本平衡。用数学符号来表示,就是将差值为正的个数记为 S^+ , 差值为负的个数记为 S^- , 按照中位数的意义,若 $H_0: M = M_0$ 成立,那么 S^+, S^- 应大体相等, S^+, S^- 都服从二项分布 $B(n, 0.5)$ 。当 S^+, S^- 过大或过小,或者 $\min(S^+, S^-)$ 过小时,就有理由拒绝 H_0 。

显然,符号检验只利用了这些数据对的差值在正的一侧还是负的一侧更多这一信息,但并没有利用这些差值的大小所包含的信息,因此它虽然简单易行,但检验效能较低,精度较差。一般而言,这种方法更适用于对无法用数字计量的情况进行比较,比如资料本身就是两分类,对于连续性的资料则最好不要使用。



需要指出的是,SPSS 在利用二项分布进行符号检验时一律都会给出确切概率值,因此符号检验的结果给出的也是确切的概率,和手工查表的结果会有所差异,即更为准确。

2. Wilcoxon 符号秩检验

由于符号检验方法的功效较低,因此 Wilcoxon 符号秩检验又按此思路进行了改进,既考虑了样本差数的符号,同时又考虑到差数的顺序。不同的符号代表了在中心位置的哪一边,而差的绝对值代表了距离中心的远近,两者结合会更有效(注意该秩和检验利用的是样本差数的顺序,而不是样本差数数值本身,在这方面又比参数检验利用样本数值本身的信息逊色)。

Wilcoxon 符号秩检验的假设和符号检验是相同的,也是考察均值差值所在总体的中间位置是否为 0,这一般被归结为考察总体中位数是否为 0。

H_0 : 差值的总体中位数 $M_d = 0$; H_1 : 差值的总体中位数 $M_d \neq 0$ 。

进行检验时,对于配对样本 $(x_1, y_1), \dots, (x_n, y_n)$, 计算出每对数据之差,用 d_i 表示。若 d_i 为连续变量并且服从正态分布,一般可以用 t 检验,但若 d_i 不是正态分布,就只能采用非参数分析方法。对 $|d_i|$ 由低到高进行排序,相同的差异将被赋予平均秩,若 X, Y 具有相同的分布,那么 $P(d_i > 0) = P(d_i < 0)$ 。把 $|d_i|$ 看成单样本,令 W^+ 表示 $|d_i| > 0$ 的秩和, W^- 表示 $|d_i| < 0$ 的秩和。检验统计量取 $W = \min(W^+, W^-)$, 在文献中也记为统计量 T ; 当 H_0 (差值的总体中位数 $M_d = 0$) 成立时,任一配对的差值出现正号与出现负号的机会均等,因此它们的秩和 W^+ 与 W^- 的理论数(期望值)也应相等,可以证明:当 H_0 为真时,秩统计量 T 是对称分布的,对称轴为 $T = n(n+1)/4$; 当 H_0 非真时,统计量 T 呈偏态分布,并且在大多数情况下 T 远离 $n(n+1)/4$ 。因此在 H_0 成立的情况下 T 远离 $n(n+1)/4$ 为小概率事件,可以认为在一次抽样中是不会发生的,故当出现这种情况时将得出拒绝 H_0 的结论。

在样本量较大的情形下, W 的抽样分布近似于正态概率分布, $Z = \frac{W - \mu_w}{\sigma_w}$, $\mu_w = \frac{n(n+1)}{4}$, $\sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{24}}$, n 为配对值的总数。

3. 其他检验方法

在 SPSS 中共给出了 4 种可以用来进行配对样本间非参数检验的方法。除了以上两种方法以外,还提供了以下两种方法。

(1) McNemar: 实际上就是常用的配对卡方检验,因此只适用于两分类资料,它考察的重点是两组间分类的差异,对于相同的分类则忽略不计。该检验特别适合于自身对照设计,用于分析处理前后的变化情况,详见第 16 章。

(2) Marginal Homogeneity: 是 McNemar 法向多分类情形的扩展,适用于资料为有序分类的情况。

15.2.2 案例:北京大学与清华大学 2002 年高考录取分数比较

例 15.1 北京大学与清华大学是国内一流的两所大学,每年高考录取时都是分数最高的学校。这两个学校都声称自己的录取分数是全国最高的,这里就用统计方法来判断他们的录取分数是否有差别。数据见北大清华 2002 理科录取比较.sav。

此处的检验假设为 $H_0: M_x = M_y$; $H_1: M_x \neq M_y$, 图 15.1 中同时列出了 Wilcoxon 符号秩检验的简单计算过程。计算得到 $W^+ = 280.0$, $W^- = 216.0$, 两个符号秩和的值相差不太大。下面进一步计算统计量, 这里取 $W = \min(W^+, W^-)$, 即 $W = \sum S^- = 216.0$ 为检验统计量。用正态分布表查到近似 P 值为 0.531, 大于给定检验水平 0.05, 因此保留零假设, 认为北京大学与清华大学两所高校 2002 年高招理科录取平均分的差别没有统计学意义。

地区	北大(b)录取 平均分	清华(q)录取 平均分	$d_i = b_i - q_i$	$ d_i $	秩	S^+	S^-
北京	637.8	638.7	-.9	.9	3.0		3.0
天津	671.7	663.8	7.9	7.9	22.0	22.0	
河北	665.3	671.7	-6.4	6.4	20.0		20.0
山西	658.6	654.3	4.3	4.3	17.0	17.0	
内蒙古	660.7	664.8	-4.1	4.1	15.5		15.5
辽宁	667.1	669.7	-2.6	2.6	9.0		9.0
吉林	665.4	660.1	5.3	5.3	19.0	19.0	
黑龙江	668.1	667.8	.3	.3	1.5	1.5	
上海	568.1	556.0	12.1	12.1	27.0	27.0	
江苏	652.7	650.0	2.7	2.7	10.0	10.0	
浙江	682.9	681.0	1.9	1.9	8.0	8.0	
安徽	656.0	669.5	-13.5	13.5	29.0		29.0
福建	664.7	660.1	4.6	4.6	18.0	18.0	
江西	673.2	674.2	-1.0	1.0	4.0		4.0
山东	683.6	670.2	13.4	13.4	28.0	28.0	
河南	651.5	647.6	3.9	3.9	14.0	14.0	
湖北	670.8	671.1	-.3	.3	1.5		1.5
湖南	665.2	668.7	-3.5	3.5	12.0		12.0
广东	839.3	835.2	4.1	4.1	15.5	15.5	
广西	809.5	808.2	1.3	1.3	7.0	7.0	
海南	798.4	827.9	-29.5	29.5	31.0		31.0
重庆	673.9	664.4	9.5	9.5	25.0	25.0	
四川	668.7	667.5	1.2	1.2	6.0	6.0	
贵州	660.5	649.7	10.8	10.8	26.0	26.0	
云南	628.2	636.3	-8.1	8.1	24.0		24.0
西藏	634.0	626.0	8.0	8.0	23.0	23.0	
陕西	669.8	672.6	-2.8	2.8	11.0		11.0
甘肃	642.9	656.6	-13.7	13.7	30.0		30.0
青海	624.7	631.8	-7.1	7.1	21.0		21.0
宁夏	649.8	646.1	3.7	3.7	13.0	13.0	
新疆	649.1	650.2	-1.1	1.1	5.0		5.0

图 15.1 北京大学与清华大学 2002 年各省、直辖市、自治区招生理科录取平均分

1. 界面说明

这里还是以新对话框为主介绍 SPSS 对配对符号秩检验的实现方式, 选择“分析”→“非参数检验”→“相关样本”菜单项, 就会打开相应的对话框, 如图 15.2 所示, 该对话框中的内容非常容易理解, 下面进行简要介绍。

(1) “目标”选项卡: 包括“自动比较观察数据和假设数据(总体)”、“自定义分析”两种情况, 最下方的“描述”框会给出相应分析方法的简单说明。实际上, 后面对选项卡的更改会使得此处的设定被自动调整, 因此一般不需要专门设定。

(2) “字段”选项卡: 指定需要分析的变量, 将需要进行比较的一对变量选入即可。需要说明的是, 这里选入的变量数量应当和随后指定的分析方法相一致, 如果是两样本比较方法, 则此

处只能选入一对变量;否则系统将拒绝执行。

(3) “设置”选项卡:这里列出了可供使用的各种两相关样本、K 相关样本比较方法,在变量的测量尺度设定正确的情况下,可以选中“根据数据自动选择检验”单选框,此时系统会按照所选变量的测量尺度自动选择检验方法,否则就需要用户进行检验方法的自定义。其中的“检验选项”和“用户缺失值”选择项目和第12章中介绍的基本相同,这里不再重复介绍。

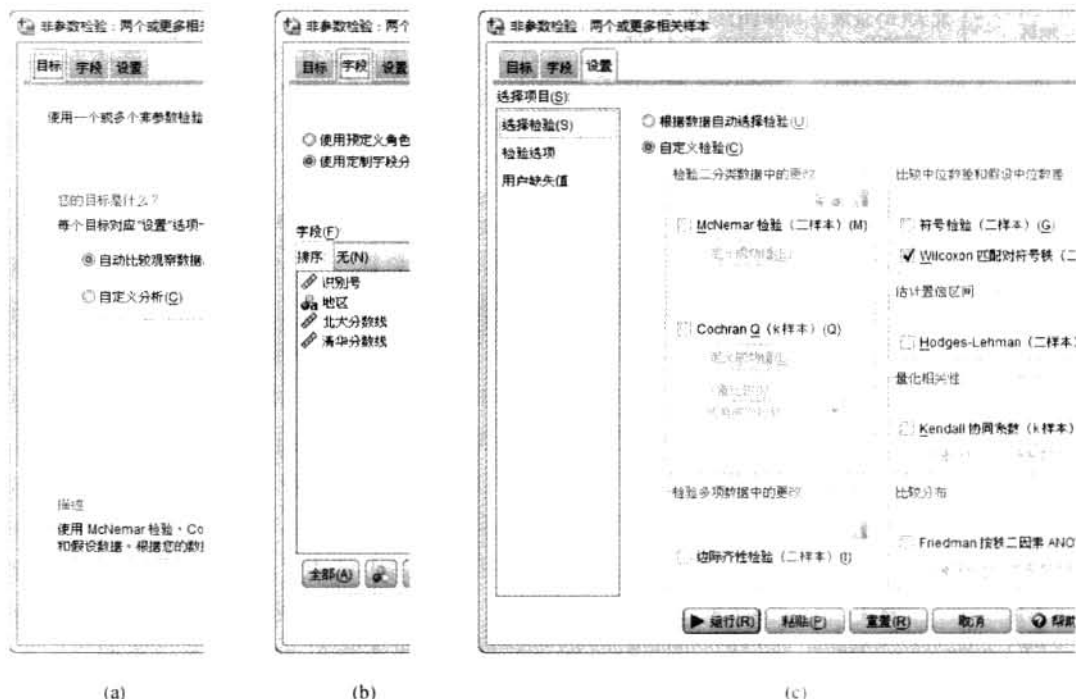


图 15.2 “非参数检验:两个或更多相关样本”对话框

2. 操作说明与结果解释

本例的操作步骤如下。

(1) 选择“分析”→“非参数检验”→“相关样本”菜单项,打开“非参数检验:两个或更多相关样本”对话框。

(2) “字段”选项卡:在检验字段列表框中选入北京大学、清华大学录取平均分。

(3) 可保留默认的“根据数据自动选择检验”选项,也可切换为“自定义检验”选项,选中“Wilcoxon 匹配对符号秩和检验”复选框。

(4) 单击“运行”按钮。

本例的分析结果为模型输出,如图 15.3 所示,在进入编辑状态后,可见在清华-北大正差一侧有一个较大的数值,对照原始数据可知为海南省分数。最终该方法给出的近似概率(Asymp. Sig.,即近似 P 值)为 0.531,大于 0.05 的显著性水平,所以保留 H_0 假设,就此数据来说,北京大学与清华大学两所高校 2002 年高招理科录取的平均分的差别没有统计学意义。显然,计算机处理的结果与通过计算 Wilcoxon 符号秩检验的秩和统计量再查表得到的结论是一致的。



该模型输出无法得到精确 P 值,如果希望得到该数值,需要使用旧对话框来实现,后面将详细介绍。

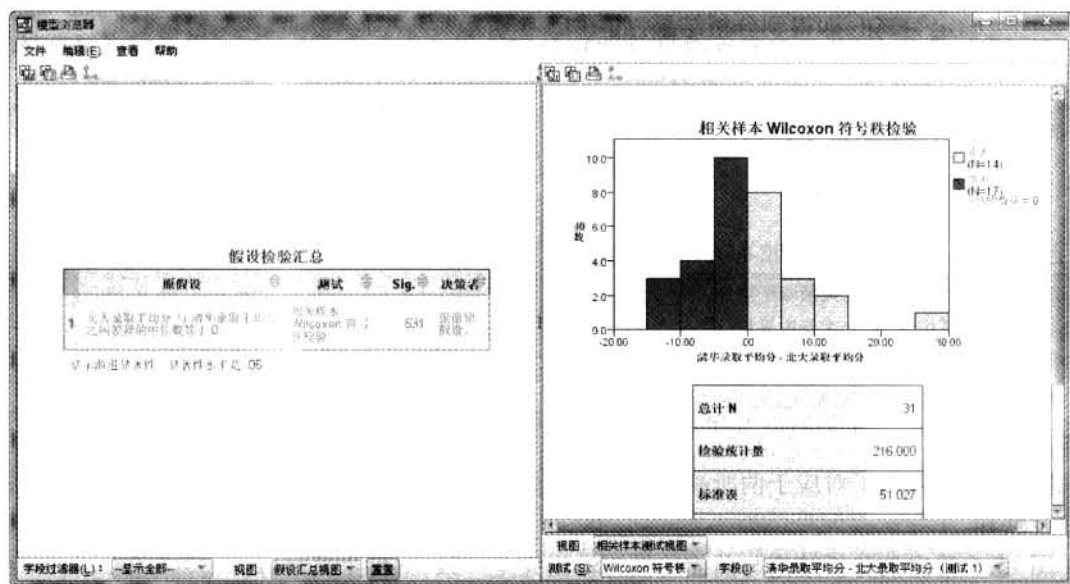


图 15.3 配对符号秩和检验的分析结果

15.2.3 使用旧对话框分析案例

本例使用 SPSS 完成的操作非常简单,需要设置的对话框和“配对 t 检验”对话框非常相似,如图 15.4 所示,操作也基本相同,因此这里不再解释。

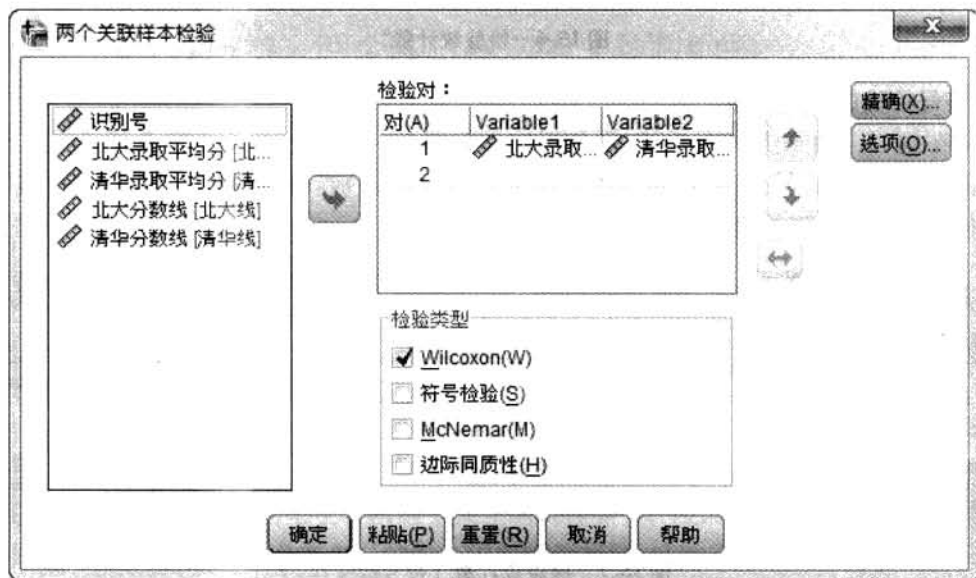


图 15.4 “两个关联样本检验”对话框



和图 15.4 所示的新对话框相比,旧对话框有两个优势:可以计算确切概率值,以及可以同时 进行多对变量的检验。

单击“确定”按钮,即得到分析结果,如图 15.5 所示。

		N	秩均值	秩和
清华录取平均分 - 北大录取平均分	负秩	17 ^a	16.47	280.00
	正秩	14 ^b	15.43	216.00
	结	0 ^c		
	总数	31		

- a. 清华录取平均分 < 北大录取平均分。
- b. 清华录取平均分 > 北大录取平均分。
- c. 清华录取平均分 = 北大录取平均分。

图 15.5 秩

图 15.5 中列出了对应于两所高校录取平均分数差值为正、负、相等 3 种情况时的秩频数、秩均值及秩和。

图 15.6 给出的分析结论与前面的模型输出完全一致,不再重复解释。

		清华录取平均分 - 北大录取平均分
Z		-.627 ^a
渐近显著性(双侧)		.531

- a. 基于正秩。
- b. Wilcoxon 带符号秩检验。

图 15.6 检验统计量^b

1. 符号检验的分析结果

如果在例 15.1 中使用符号检验,则分析结果如图 15.7 所示,可见近似 P 值等于 0.719,虽然在本例中结论相同,但是,两者的近似 P 值有比较大的差别。这主要是由于符号检验对信息的利用程度不如符号秩检验,检验效能不够充分所致。



在绝大多数情况下,对信息利用程度较差的方法所得到的 P 值会更大,但也的确会存在相反的情形。

		清华录取平均分 - 北大录取平均分
Z		-.359
渐近显著性(双侧)		.719

- a. 符号检验。

图 15.7 检验统计量^a(符号检验)

另两种方法因在本例中不满足使用条件,此处不再给出结果示例。

2. 确切概率的计算

在前面的分析结果中可以看到,在概率一项中,显示的是近似概率(Asymp. Sig)。这是因为上面使用的是秩统计量的正态近似法计算的概率值。如果同时安装了 SPSS Exact Tests 模块,还可以计算精确概率(Exact Sig.)。当然这项工作花费的时间相对要长一些,尤其是在数据量大时。操作方法如下。

- (1) 单击“精确”按钮。
- (2) 在打开的对话框中选中“精确”单选框。
- (3) 单击“继续”按钮。

由图 15.8 可见,在 H_0 假设之下,获得这样差别或更大差别样本的精确概率为 0.539,和近似的 0.531 有所差异。显然,精确与渐进法相比,会给出更为准确的概率值。尤其是当近似概率接近显著性水平时,精确概率就显得更为重要。

清华录取平均分 - 北大录取平均分	
Z	-.627 ^a
渐近显著性(双侧)	.531
精确显著性(双侧)	.539
精确显著性(单侧)	.269
点概率	.003

a. 基于正秩。
b. Wilcoxon 带符号秩检验。

图 15.8 检验统计量^b

15.3 两个独立样本的非参数检验

在两个独立样本的非参数检验方法中,Mann-Whitney U 检验,即两样本秩和检验是应用最广的一种,本节就将以它为主加以讲解,并对其余几种方法加以介绍。

15.3.1 方法原理

1. Mann-Whitney U 检验

简单地讲它是和参数 t 检验相对应的一种非参数检验方法,就是人们最常用的两样本秩和检验方法,它在检验时利用了大小次序,即检验 A 样本中的数值是否多数都大于 B 样本。这种方法是由 H. B. Mann 和 D. R. Whitney 在秩和的基础上改进而来的,用来检验两个独立样本是否取自同一总体。前面介绍过两个总体均值间差异的参数检验,是基于两个总体均为正态分布、两个总体方差相同的假设的,而这里仅要求两个独立随机样本中产生的数据的测量尺度是顺序的,而具体所检验的就是两个总体分布各自的中心位置是否相同,这就是建立零假设和备择假设的基础。

设有 X_1, \dots, X_m 和 Y_1, \dots, Y_n 两个总体具有连续分布,建立的假设为: H_0 : 两总体分布的中心

位置相同, H_1 : 两总体分布的中心位置不相同。将 m 个 x , n 个 y 数据混合排序, 这样可以计算出每个数值在混合样本中所在位置的次序, 即等级或秩 R 。在有结的情况下, 每个结得到平均秩。

分别计算出样本 X 和 Y 的秩和, 即令 $W_X = \sum_{i=1}^m R_i$, $W_Y = \sum_{j=1}^n R_j$ 。显然, 如果这两个总体分布的中心位置相同, 则两个样本中各数据的秩次都应当围绕着平均秩次 $(N+1)/2$ 均匀分布, 样本 X 的秩和应当接近于 $m(N+1)/2$, Y 的秩和接近于 $n(N+1)/2$, 如果和该理论值差别较大, 则可推断总体的中心位置是有差异的。为了进行检验, 可以计算每个样本的 U -统计量:

$$U_{XY} = mn + m(m+1)/2 - \sum_{i=1}^m R_i, U_{YX} = mn + n(n+1)/2 - \sum_{j=1}^n R_j$$

U_{XY} 表示 Y 的观察值大于 X 观察值的个数, U_{YX} 表示 X 的观察值大于 Y 观察值的个数。注意有 $mn = U_{XY} + U_{YX}$, $m+n = N$ 。因此以上两式简化为 $U_{XY} = W_Y - n(n+1)/2$, $U_{YX} = W_X - m(m+1)/2$ 。当 m, n 均大于 10 时, U 近似服从正态分布, 此时可以进一步计算标准正态分布的统计量 $Z = \frac{U - \mu}{\sigma} = \frac{U - mn/2}{\sqrt{mn(m+n+1)/12}}$ 。在 X, Y 的样本有相同的值, 即混合样本有结时, 可以用结统计

量对 Z 值进行修正, 由于公式较复杂, 这里不再给出。在 SPSS 中相应的校正是自动进行的, 并可以直接给出精确计算的概率值, 因此不需要用户对此做特别关注。

除了 Mann-Whitney U 检验外, 在统计教材中更为常见的是 Wilcoxon 秩和检验, 这两种方法是独立提出的, 但仅仅是统计量的构造略有不同, 其原理和检验结果完全等价, 因此不再单独解释, 而 SPSS 在分析时也会同时给出这两种统计量。

2. Kolmogorov-Smirnov Z 检验

和单样本检验中的 $K-S$ 检验是一类的, 可以对连续性资料的分布情况加以考察。 $K-S$ 检验的原理如下: 分别做出已知理论分布下的累积频数分布以及观察的累积频数分布, 然后对两者进行比较, 从中确定两种分布的最大差异点。如果样本确实服从理论分布, 则最大差异值不应太高; 否则就应当拒绝该假设。不过这次是检验两个独立样本是否取自同一总体, 操作原理是做出两个样本的累积频数分布曲线, 然后观察两条曲线究竟差了多少。显然, 这种方法检验的是总体分布情况是否相同, 而不仅仅是考察所在总体的中心位置是否相同。因此, 如果只是要检验中心位置是否相同, 最好不要选择这种方法。

3. Moses Extreme Reactions 检验

该检验有其特定用途, 注意给出的结果均为单侧检验。顾名思义, 如果施加的处理使得某些个体出现正向效应, 而另一些个体出现负向效应, 就应当采用该检验方法。比如说要研究人民群众对电信资费下调的反应, 多数人当然会很高兴, 但是在电信工作的人就会比较沮丧了, 因此如果研究的目标人群中电信职工较多, 不妨考虑采用此法。

4. Wald-Wolfowitz Runs 检验

从名字就可以看出它属于游程检验的一种, 即检验的是总体分布情况是否相同。更准确地说, 只要两样本各自所在总体有任何一点分布上的差别, 无论是集中趋势、离散趋势、偏度还是波动情况, 统统都可以检验出来。因此如果只是要检验中心位置是否相同, 最好不要选择这种方法。该方法同样给出的是单侧检验的结果。

15.3.2 案例:不同收入家庭经济现状感受值的比较

例 15.2 在 CCSS 案例中,Qa3 记录的是根据题目 A3(过去一年家庭经济现状感受)计算出的感受值,现希望考察 2007 年 4 月的中高收入家庭和中低收入家庭相比的家庭经济状况感受值有无差异。

Qa3 虽然取值为连续性变量,但其特殊性在于只有 0、50、100、150、200 这 5 种取值方式,因此直接使用 t 检验来比较似乎不大妥当,这里考虑使用秩和检验来分析。

1. 界面说明

选择“分析”→“非参数检验”→“独立样本”菜单项,就会打开相应的对话框,如图 15.9 所示,该对话框中的内容非常容易理解,下面进行简要介绍。

(1) “目标”选项卡:包括“自动比较不同组间的分布”、“比较不同组间的中位数”,以及“自定义分析”3 个选项,最下方的“描述”框会给出相应分析方法的简单说明。实际上,后面对选项卡的更改会使得此处的设定自动被调整,因此一般不需要专门设定。

(2) “字段”选项卡:指定需要分析的变量,包括希望检验的字段,以及相应的分组变量。

(3) “设置”选项卡:这里列出了可供使用的各种两组、多组独立样本的比较方法,在变量的测量尺度设定正确的情况下,可以选中“根据数据自动选择检验”复选框,此时系统会按照所选变量的测量尺度自动选择检验方法,否则就需要用户进行检验方法的自定义。其中的“检验选项”和“用户缺失值”选择项目和第 12 章中介绍的基本相同,这里不再重复介绍。



图 15.9 “非参数检验:两个或更多独立样本”对话框

2. 操作说明与结果解释

- (1) 选择“数据”→“选择个案”菜单项。
- (2) 在打开的对话框中选择框组:如果条件满足。
- (3) “如果”按钮:在单击后打开的对话框中设置 time = 200 704,继续。
- (4) 单击“确定”按钮。
- (5) 选择“分析”→“非参数检验”→“独立样本”菜单项,打开如图 15.4 所示的对话框。
- (6) 在“字段”选项卡中设置使用定制字段分配,在检验字段列表框中选出“Qa3”,在组变量列表框中选出“家庭收入两级[ts9]”。
- (7) 可使用默认的“根据数据自动选择检验”选项,也可切换为“自定义检验”选项,选择 M-W U 检验。
- (8) 单击“运行”按钮。

相应的分析结果如图 15.10 所示,从平均秩次可以粗略看出不同收入人群在 Qa3 上的秩和相差较大。但究竟有无统计学意义还要看后面的结果。秩次分布图下方的表格给出了最终的检验结果,包括 Mann-Whitney U 统计量、Wilcoxon W 统计量和 Z 值(即常用的 u 值),近似 P 值显示为 0.000,小于给定水平 0.05,所以拒绝原假设,说明不同收入人群在家庭收入状况感受值上的差异有统计学意义。

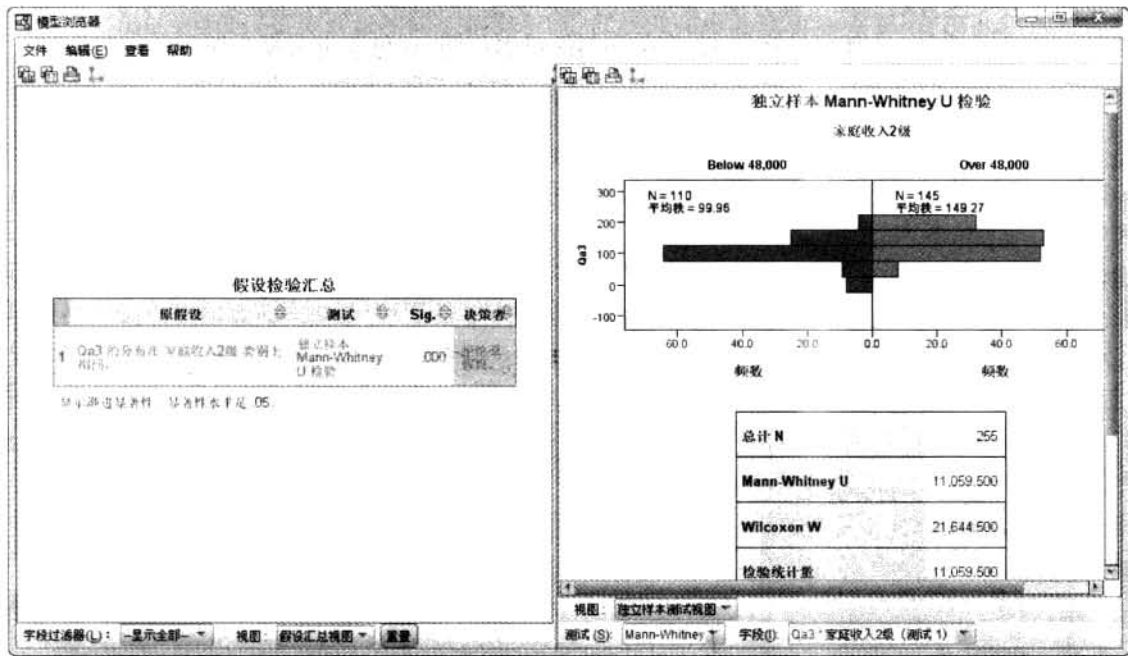


图 15.10 M-W U 检验的分析结果

15.3.3 使用旧对话框分析案例

由于此处的操作非常简单,和两样本 t 检验几乎完全相同,如图 15.11 所示,因此不再重复解释。



图 15.11 “两个独立样本检验”对话框

相应的操作如下。

(1) 选择“分析”→“非参数检验”→“旧对话框”→“两个独立样本”菜单项,打开“两个独立样本检验”对话框,进行如下设置。

“检验变量列表”框:Qa3。

“分组变量”列表框:Ts9。

(2) 单击“定义组”按钮,在打开的对话框中设置组 1 为 1,组 2 为 2,单击“继续”按钮。

(3) 单击“确定”按钮。

相应的分析结果如图 15.12、图 15.13 所示。

家庭收入 2 级		N	秩均值	秩和
Qs9	Below 48,000	110	55.50	6105.00
	Over 48,000	145	183.00	26535.00
总数		255		

图 15.12 秩

		Qs9
Mann-Whitney U		.000
Wilcoxon W		6105.000
Z		-13.738
渐近显著性(双侧)		.000

a. 分组变量: 家庭收入 2 级。

图 15.13 检验统计量^a

15.4 多个独立样本的非参数检验

多样本问题主要涉及如何检验几种不同的方法、决策或处理所产生的结果是否一样。比如生活中不同的消费者对不同的产品偏好是否有显著差异;不同的运动方式或饮食习惯对减肥效果是否一样;商业活动中采取不同的决策方案风险的大小是否有区别;不同的销售方式购买率是否相同。在第 14 章中进行多组均数的比较时,利用了方差分析来推断 3 个或 3 个以上总体的均值的相等性。但是该过程需要若干条件,如要求间隔或比例数据,所有总体服从正态分布,且各总体的方差均相等。但是有时候所采集的数据常常不能满足这些条件,事实上假使有一个条件不满足都会使人陷入尴尬之中。当不满足这些条件时, F 检验就受到了限制。

15.4.1 方法原理

1. Kruskal-Wallis H 检验

克罗斯考尔和瓦里斯于1952年设计了一种类似于 Wilcoxon 秩和检验的方法,用来解决此类问题。于是在进行 $k \geq 3$ 个独立随机连续分布样本的比较后,正态性假设及等方差假设存在问题时, K-W 检验就提供了一种可用于检验总体是否相同的替代统计方法。

解决多样本问题的思路与前面两样本的 Wilcoxon 秩和检验一样。实际上, Kruskal-Wallis H 检验可以被简单地看成是两样本的 Wilcoxon 方法在多样本时的推广:将数据转化为秩统计量。因为秩统计量的分布与总体分布无关,可以摆脱总体分布的束缚。具体而言,就是把大小为 n_1, n_2, \dots, n_k 的样本混合成为一个单样本,将数据按大小顺序排秩,每一个观测值在新样本中都有自己的秩,如果有相同的数据,则和以前一样取秩的平均值,记观测值 x_{ij} 的秩为 R_{ij} ,对每一个样本的观测值的秩求秩和 R_i ,再找到它们在每组中的平均值 $\bar{R}_i = R_i/n_i$,此处的检验假设仍然针对分布的中心位置, $H_0: m_1 = m_2 = \dots = m_k$; H_1 : 至少有一个 m_j 不同。如果零假设为真,秩应该在 k 个样本之间均匀分布,也就是说多样本实际的秩和与期望秩和的偏差应该很小, K-W 检验便建立在这一基础上。若这些 \bar{R}_i 相差太大,就可以怀疑零假设。基于上述原理, K-W 检验构造的检验统计量为

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2 = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

其中 $N = \sum_{i=1}^k n_i$, $\bar{R} = \sum_{i=1}^k R_i / N = \frac{N+1}{2}$ 。 R_i 是样本 i 的秩和; k 是总体个数; N 是所有样本个体总数; n_i 是样本 i 的个体数(样本大小可以不一样!)。可以验证 Mann-Whitney 统计量 U_{xy} 就是 Kruskal-Wallis 统计量 H 在两样本时的特例。存在打结时,检验统计量 H 同样可以修正为

$$H_c = \frac{H}{1 - \sum_{i=1}^g (\tau_i^3 - \tau_i) / (N^3 - N)}$$

在样本量较大的情形下,当 $\min(n_1, \dots, n_k) \rightarrow \infty$ 时,在 H_0 下,有 H 近似于 $\chi^2(k-1)$ 分布。

2. SPSS 中的其他检验方法

除上述 Kruskal-Wallis H 检验外, SPSS 还为多组比较提供了如下两种非参数方法。

(1) 中位数: 中位数检验, 检验各个样本是否来自具有相同中位数的总体, 3 种方法中它的检验效能最低。但对于厚尾的对称分布该方法倒是很有有效的检验。

(2) Jonckheere-Terpstra: 该检验对连续性资料或有序分类资料都适用, 并且当分组变量为有序分类资料时, 此法的检验效能要高于 Kruskal-Wallis 法。

15.4.2 案例: 不同时点上的家庭经济现状感受值比较

例 15.3 继续对 CCSS 案例数据进行分析, 现希望考察在 4 个不同时点上的受访者家庭经济状况感受值有无差异。

本例是比较典型的 4 组平均水平比较问题, 同样由于 Qa3 取值范围的问题, 这里考虑使用秩和检验来分析。由于使用的对话框和 15.4.1 节中完全相同, 因此这里不再重复介绍, 直接给出

操作步骤如下。

- (1) 选择“分析”→“非参数检验”→“独立样本”菜单项,打开如图 15.4 所示的对话框。
- (2) 在“字段”选项卡设置使用定制字段分配,在检验字段列表框中选取 Qa3,在组变量列表框中选取月份[time]。
- (3) 可保留默认的“根据数据自动选择检验”选项,也可切换为“自定义检验”选项,然后选择 K-W ANOVA 检验,多重比较采用默认的“所有成对比较”。
- (4) 单击“运行”按钮。

1. 总体检验结果

分析结果如图 15.14 所示,由 Kruskal-Wallis H 检验的总体检验结果可见其近似显著性概率小于 0.001,当然也就小于 0.05,故应拒绝原假设,可以认为不同时间点的家庭经济现状信心值差异有统计学意义,在右侧上方的箱图中还直接给出了 4 个时点的 Qa3 分布情况,从中可见似乎 2009 年 12 月的 Qa3 平均水平更高一些。

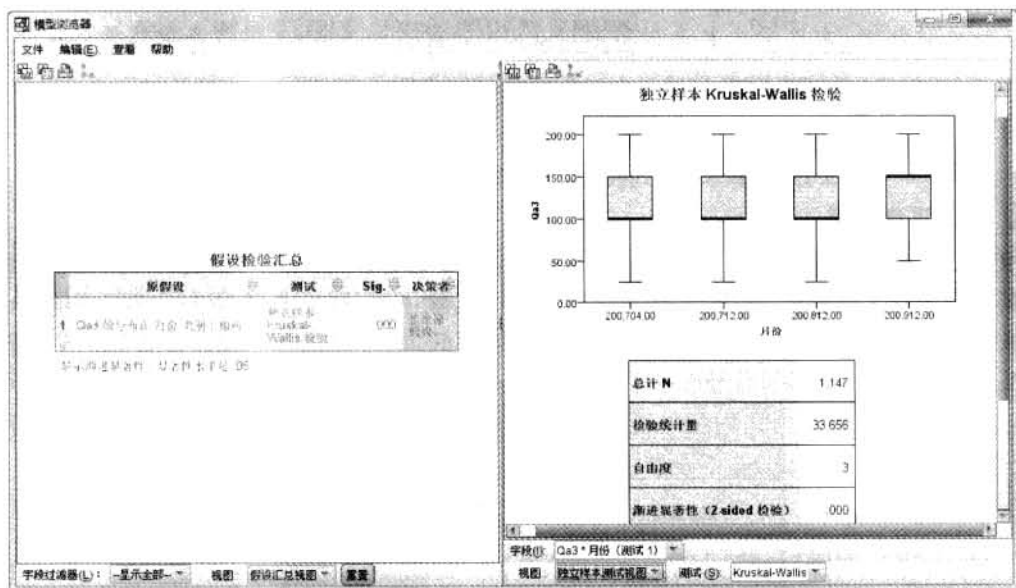


图 15.14 4 样本组 K-W 检验的分析结果

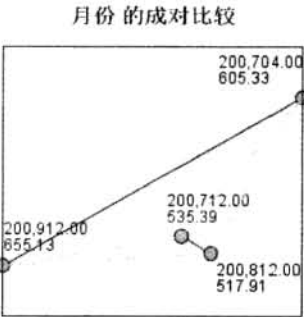
2. 两两比较结果

显然,仅仅考察 4 组在总体上有无差异是远远不够的,还应当进行随后的两两比较,如图 15.15 所示,在右侧底部的“视图”下拉列表框中将视图从默认的“独立样本测试视图”更改为“成对比较”,则可以给出具体的两两比较结果。

在“成对比较”视图的上半部,是以网络图形式给出的 4 个时间点的样本平均秩差异,其节点的距离远近形象地反映了平均秩差异的大小,从图 15.8 中可见似乎 2007 年 12 月和 2008 年 12 月的平均秩非常接近,而它们和 2007 年 4 月和 2009 年 12 月则存在较大的差异。

在“成对比较”视图的下半部,则会给出两两比较的具体检验结果,从检验结果可以看出,2007 年 12 月确实和 2008 年 12 月无统计学差异,而和 2007 年 4 月存在差异。从实际含义角度看,这说明在 2007 年年底时,消费者的家庭经济现状感受值就已经跌入谷底,其后虽然总信心指

数仍在下跌,但家庭经济现状感受值已经持稳,直至 2009 年 12 月重新反弹回 2007 年 4 月基期的水平附近。将 Qa3 的两两比较结果和总指数 index1 的两两比较结果加以对比,研究者就可以发现更多值得解释的深层次信息,这一工作读者可自行完成,此处不再继续展开。



每个节点显示月份的样本平均秩。
(a)

样本1-样本2	检验统计量	标准误	标准检验统计量	Sig.	调整显著性
200,812.00-200,712.00	17.485	25.310	.691	.490	1.000
200,812.00-200,704.00	87.420	25.394	3.443	.001	.003
200,812.00-200,912.00	-137.222	26.976	-5.087	.000	.000
200,712.00-200,704.00	69.935	25.394	2.754	.006	.035
200,712.00-200,912.00	-119.737	26.976	-4.439	.000	.000
200,704.00-200,912.00	-49.801	27.055	-1.841	.066	.394

每行检验原假设：样本1和样本2分布相同。
显示渐进显著性(2-sided检验)。显著性水平是 .05。
(b)

图 15.15 4 样本组平均秩次差异的图形化显示及两两比较的结果表格

3. 齐性子集结果

和方差分析中两两比较的情形类似,在秩和检验中也存在寻找同质亚组的两两比较方法,在前述 K-W ANOVA 检验的“多重比较”下拉列表框中,除了默认的“所有成对比较”选项外,还可以选择“逐步降低”选项,这时就会按照寻找同质亚组的方式进行结果输出,如图 15.16 所示。

基于 Qa3 的齐性子集

	子集	
	1	2
200,812.00	517.908	
200,712.00	535.393	
200,704.00		605.328
200,912.00		655.130
检验统计量	284	3.487
Sig. (2-sided 检验)	.594	.062
调整后的显著性 (2-sided 检验)	.835	.120

齐性子集是基于渐进显著性。显著性水平是 .05。
¹每个单元格显示Qa3的样本平均秩。

图 15.16 齐性子集的划分结果

图 15.16 所示表格的阅读方式完全类似于第 14 章的 SNK 等同质方法的输出,可以看出 4 个时点被明确地分为了两个同质组,分析结论则和上面完全相同。但是需要注意,其检验的 P 值并不相同,且还提供了调整后的 P 值,该调整类似于第 14 章所讲解的控制一类错误 α 大小的概念,因此相应的 P 值会有所放大。



在多个总体比较中,如果拒绝了无效假设,则随后必然会遇到和单因素方差分析中相同的组间两两比较问题,以进一步来判断到底哪些总体之间有差异,甚至于准确衡量差异的程度,但由于这方面在方法学上还有一定争议,各位统计学家的意见也并不太统一,这里建议采用以下两种基本策略。

(1) 直接使用两组比较的方法进行两组间的非参数检验,此时和参数的两两比较方法一样,也会涉及控制一类错误的问题。但是,由于非参数方法相对而言检验效能会略低一些,因此对于是否一定要调整 α 水准尚有争议。一般而言,现在比较常见的看法是如果样本量较小,则不一定需要调整 α 水准,直接比较即可,这样可以补偿非参数方法检验效能不足所带来的损失;如果样本量较大,比如每组均在几十例以上,则必须要调整 α 水准,否则就会犯和进行多组均数比较时采用两两 t 检验性质相同的错误。

(2) 当各组例数较多时,可以采用秩变换分析,操作更加方便,而结论也更加准确,后面会详细介绍。

15.4.3 使用旧对话框分析案例

本例用 SPSS 完成的操作非常简单,打开的对话框和 15.3 节的两样本检验对话框非常相似,如图 15.17 所示,操作也基本相同,因此这里不再解释。但是由于月份变量 time 值的位数超出组变量允许的定义范围,因此需要转换生成一个月份的新变量 time1 (1 代表 200704, 2 代表 200712, 3 代表 200812, 4 代表 200912),相应的程序如下(当然也可以使用对话框来完成)。



图 15.17 “多个独立样本检验”对话框

```
RECODE TIME (200704 = 1) (200712 = 2) (200812 = 3) (200912 = 4) INTO TIME1.  
EXEC.
```

随后的检验操作如下。

(1) 选择“分析”→“非参数检验”→“旧对话框”→“K 个独立样本”菜单项,打开“多个独立样本检验”对话框,进行如下设置。

“检验变量列表”框:Qa3。

“分组变量”列表框:time1。

(2) 单击“定义范围”按钮,在打开的对话框中设置最小值为 1,最大值为 4,单击“继续”按钮。

(3) 单击“确定”按钮。

分析结果如图 15.18、图 15.19 所示。

	time1	N	秩均值
Qa3	1. 00	300	605. 33
	2. 00	304	535. 39
	3. 00	304	517. 91
	4. 00	239	655. 13
总数		1147	

图 15.18 秩

	Qa3
卡方	33. 656
df	3
渐近显著性	.000

a. Kruskal Wallis 检验。
b. 分组变量: time1。

图 15.19 检验统计量^{a,b}

图 15.18、图 15.19 所示的分析结果和由新对话框得到的基本一致,但是,采用旧对话框无法直接给出两两比较的分析结果,这不能不说是一个操作上的缺憾。

15.5 多个相关样本的非参数检验

除配对案例外,前面的问题相当于一种没有区组(Block)影响的单因子试验设计的分析:样本之间是独立的,每一个样本中的观测值也是相互独立的。每一个样本代表了一个“处理”(Treatment)。可是在实际生活中,除了“处理”之外,还有别的因素起作用。比如在一个新口味的食品或饮料的推广中,在不同的地区针对不同的人群进行测试,对测试者按年龄分组,或者按收入分组。这里不同的地区(假定为 3 个)代表了 3 种不同的处理($k=3$),如果将收入分成 5 等,则表示有 5 个区组($b=5$)。当区组存在时,代表处理的样本的独立性就不再成立了。一般来说,对于 k 个处理及 b 个区组,就形成 $b \times k$ 的交叉表, X_{ij} 表示表中位于第 i 个区组和第 j 个处理那一格的观察值。

15.5.1 Friedman 检验

Friedman 检验也称为弗里德曼双向评秩方差分析,在 1937 年由 Friedman 提出,也是关于位置参数的检验。该方法的基本思想是:由于区组间的差异是各式各样的,只有同区组的处理值的比较才有意义,一个观察值的秩是在某一区组中的秩,而不是对所有数据而言的。因此应当独立地在每一个区组内分别对数据进行排秩,这样就可以消除区组间的差异以检验各种处理之间是

否存在差异。该检验的假设如下:

$H_0: M_1 = \dots = M_k$ (所有的位置参数都相等)。

H_1 : 至少有一个 M_i 与其他不同 (不是所有的位置参数都相等)。

从假设上看似乎和前面的 Kruskal-Wallis H 检验一样,但是由于区组的影响,需要首先在分区组单独计算各个处理的秩,再把每一个处理在各区组中的秩相加,最后再对各处理进行比较。倘若 k 种处理不存在差异 (原假设 H_0),那么无论从哪一个区组去观察,每一种处理所得到的数据在该区组内可能地排秩都为 1 至 k 中的任何一个数。因此,对于每一种处理,它关于各区组内所取秩的总和应该等于其他任何一种处理关于各区组内所排秩的总和,或者这两种处理的秩平均数相等。1937 年 Friedman 提出的检验统计量为

$$Q = \frac{12}{bk(k+1)} \sum_{i=1}^k \left(R_i - \frac{b(k+1)}{2} \right)^2 = \frac{12}{bk(k+1)} \sum_{i=1}^k R_i^2 - 3b(k+1)$$

对于有限的 b 和 k 有零假设下的分布表可查 (要做变换 $W = Q/(b(k-1))$)。样本量较大时 Q 近似服从自由度为 $(k-1)$ 的 χ^2 分布 (当某区组存在结时, Q 可以修正为 Q_c , $Q_c = Q/(1-C)$, 其中 $C = \sum_{i,j} (\tau_{ij}^3 - \tau_{ij}) / (bk(k^2 - 1))$, τ_{ij} 是第 j 区组中第 i 个结统计量)。



Friedman 检验的最大缺陷在于检验效能太低——其独特的按区组分别编秩的做法,虽然有效利用了区组信息,但同时也使得秩次的可取值范围大大缩减,这意味着后续分析中可资利用的信息实际上非常有限,因此当样本量有限的时候,该方法的实际应用价值不大。

15.5.2 案例:不同时段的世博会入园人数比较

例 15.4 在 2010 年上海世博会期间,每天有几十万人入园参观。这些人从上午 9 点到晚上 21 点陆续进入园区。本例从官方公布的入园数据中整理出了几个时段的入园人数,变量 a 表示的是在 12~14 点之间入园的人数,变量 b 表示的是在 14~16 点之间入园的人数,变量 c 表示的是在 16~18 点之间入园的人数,变量 d 表示的是在 18~20 点之间入园的人数,现希望检验这几个时段的入园人数有没有差别,数据见 EXPO2010.sav。

本例是比较典型的配对设计 4 组平均水平比较问题,由于入园人数波动较大,存在极端值,这里考虑使用秩和检验来分析。由于使用的对话框和 15.2 节中完全相同,因此不再重复介绍,直接给出操作步骤如下。

(1) 选择“分析”→“非参数检验”→“相关样本”菜单项。

(2) 在打开的对话框中选择“字段”选项卡,在“检验字段”列表框中选中 $a-d$ 这 4 个时段的人数变量。

(3) 可使用默认的“根据数据自动选择检验”,也可切换为“自定义检验”,选择“Friedman 检验”方式,多重比较更改为“逐步降低”。

(4) 单击“运行”按钮。

检验结果 (如图 15.20 所示) 中的概率值小于给定水平 0.05,故拒绝原假设,认为 4 个时段的入园人数是有差异的。而从图形 15.20 给出的平均秩分布状态可知,似乎 12~14 点和 16~18 点这两个时段的入园人数较多。

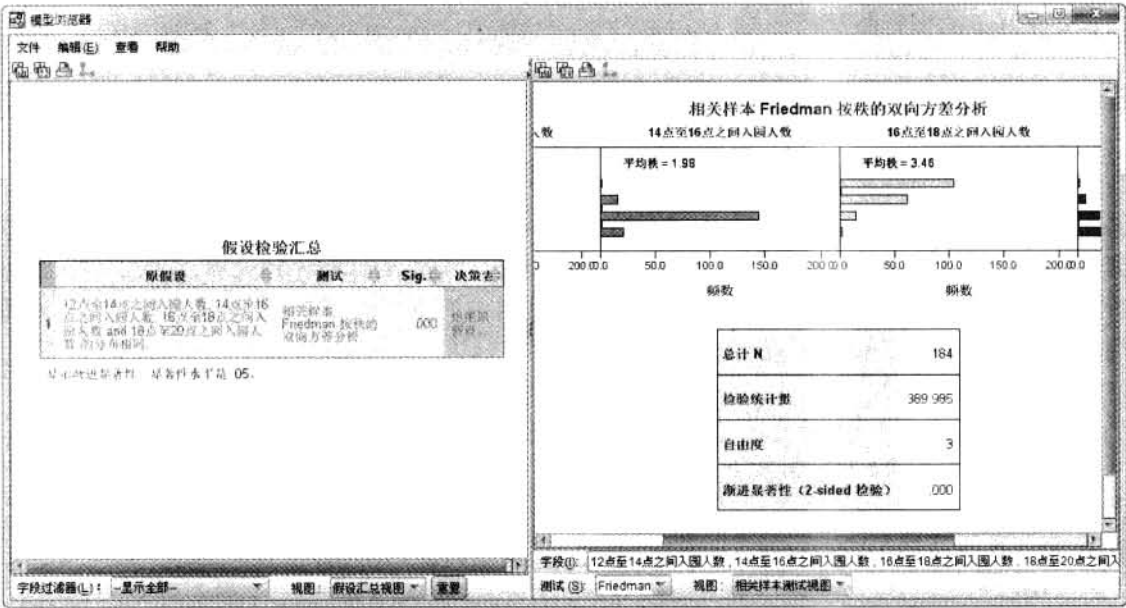


图 15.20 Friedman 秩和检验的分析结果

齐性子集的分析结果如图 15.21 所示,其中给出了更加准确的两两比较信息,可见 4 个时段被分为 3 个子集,可以发现 12~14 点以及 16~18 点之间的入园人数的是最多的,前者应当是避开早上入园高峰,在外面解决了午饭问题之后专攻下午场和夜场的游客,而后者则是因为世



齐性子集是基于渐进显著性。显著性水平是 .05。

¹每个单元格显示样本平均秩。

²无法计算,因为子集仅包含单个样本。

图 15.21 齐性子集的分析结果

博会组委会规定,从下午4点开始夜场票可以进场,因此主要应当由夜场游客入园构成;而14~16点的入园人群应当是承接了12~14点人群的特征而来的,最少的18~20点也很好解释,毕竟花了同样的夜票价格,却少看了几个小时,晚饭也无法解决,去的人自然会变少。

15.5.3 使用旧对话框分析案例

本例的对话框(如图15.22所示)比较简单,不再详细讨论,在SPSS中的操作如下。

- (1) 选择“分析”→“非参数检验”→“旧对话框”→“K个相关样本”菜单项,打开“多个关联样本检验”对话框。
- (2) 在“检验变量”列表框中添加a、b、c、d这4个变量。
- (3) 设置检验类型为Friedman。
- (4) 单击“确定”按钮。

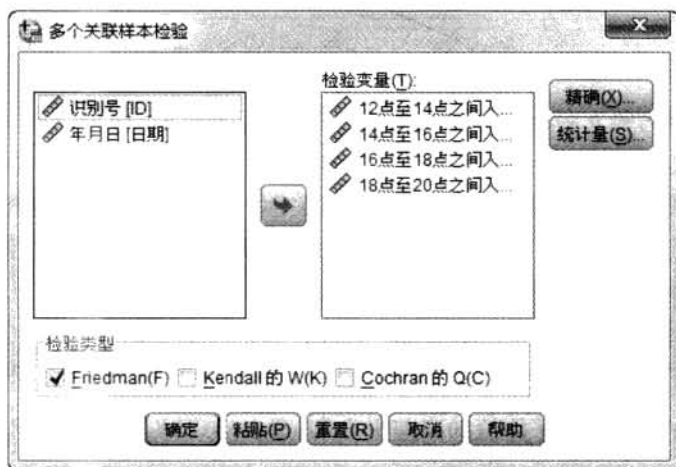


图 15.22 “多个关联样本检验”对话框

得到的分析结果如图15.23、图15.24所示。

	秩均值
12点至14点之间入园人数	3.33
14点至16点之间入园人数	1.98
16点至18点之间入园人数	3.46
18点至20点之间入园人数	1.22

图 15.23 秩

N	184
卡方	389.985
df	3
渐近显著性	.000

a. Friedman 检验。

图 15.24 检验统计量^a

15.5.4 Kendall 协和系数检验与 Cochran 检验

1. Kendall 协和系数检验

在实际生活中,经常需要按照某些特别的性质来多次对多个个体进行评估或排序,比如消费者对品牌商品的偏好、选民对候选人的评价、咨询机构对一系列企业的评估以及裁判对参赛人的

打分等。人们往往想知道,这多个评价结果是否一致。如果很不一致,则这些评估多少有些随机,没有多大意义。令零假设为 H_0 : 这些评估(对于不同个体)是不相关的或者是随机的;而备择假设为 H_1 : 评估是正相关的或者是一致的。这里完全有理由用前面的 Friedman 方法来检验。但是,在 Friedman 检验结果中如果 P 值大于 0.05,仅仅说明尚不能认为有差异,并不能明确究竟一致程度怎样,显然这离真正分析的目的还有一段距离。

例 15.5 3 名电影评论家对目前上映的一系列电影评级打分,评判等级范围从 1 到 10 共有 10 级,“1 = 很差”,…,“10 = 很好”,数据见影评家.sav,如图 15.25 所示。若在 $\alpha = 0.05$ 水平下比较 3 组评论,问他们在评级时是否依赖于相同的价值评判体系,即他们的评判是否一致。

评论者	电影 1	电影 2	电影 3	电影 4	电影 5	电影 6	电影 7	电影 8
影评家 1	9.0	8.0	7.0	8.0	9.0	9.0	7.0	8.0
影评家 2	7.0	8.0	9.0	6.0	10.0	9.0	8.0	9.0
影评家 3	6.0	6.0	5.0	7.0	8.0	10.0	9.0	6.0

图 15.25 不同影评家的评分结果

该问题可以理解为有 b 名评论家对 k 部电影打分, X_{ij} 表示第 j 个评论家对第 i 部电影打的分数,这样得到样本 $(X_{1j}, \cdots, X_{bj}) (j=1, \cdots, b)$ 。以 R_{ij} 表示 X_{ij} 在 (X_{1j}, \cdots, X_{bj}) 中的秩。如果评判是不相关的,则由任一部电影所得的秩应该也没有相关性,每部电影的秩和应相差不大。但如果评论家的评判是一致的(正相关的),则会有一些电影的秩和较大,而另一些电影的秩和较小,这时就可以采用 Friedman 检验来判断他们在评级时的判断取向是否相同。但是,为了得到对相关性的具体数量评价,还必须在此基础上进行进一步的扩展。每个评估者(共 b 个)对于所有参加排序的 k 个个体有一个从 1 到 k 的排列(秩),而每个个体有 b 个打分(秩),则用 T 表示个体的总秩 $R_i = \sum_{j=1}^b R_{ij} (i=1, \cdots, k)$ 与其平均值的离差平方和: $T = \sum_i (R_i - \frac{1}{k} \sum_{i=1}^k R_i)^2$ 。如果评判是不相关的,则 T 的值应当较小;否则,则 T 值应较大,所以 T 就可以用来刻画多个变量的相关性。因为 $\sum_{i=1}^k R_i, (i=1, \cdots, k)$ 是所有秩的和,于是 $\sum_{i=1}^k R_i = b(1 + \cdots + k) = \frac{bk(k+1)}{2}$,从而 $T = \sum_i \left(R_i - \frac{b(k+1)}{2} \right)^2$ 。当第 1 部电影的秩全取 1,第 2 部电影的秩全取 2, …, 第 k 部电影的秩全取 k 时,那么这 b 名评论家的评判是完全一致的,此时 T 取得最大值 $\frac{b^2k(k^2-1)}{12}$ 。为了与习惯一致,取一个在 0 与 1 之间的数来刻画多个变量的相关性,所以考虑用以下统计量来度量

$$W = T / \left(\frac{b^2k(k^2-1)}{12} \right)$$

该统计量就是 Kendall 协和系数。 W 愈接近 1, b 个变量间的正相关性愈好,即表现的一致性愈强;反之, W 愈接近 0, 变量间的正相关性愈差,一致性愈弱。因此与 Friedman 检验相比, Kendall 协和系数不仅可以检验 k 个相关样本是否来自同一总体,还能检验 b 个变量间的相关性。它表示的是 k 个指标间相互关联的程度(一致性程度),取值在 0 ~ 1 之间。

从 SPSS 的对话框中可见,相关样本的新对话框中直接提供了 Kendall 协和系数这一方法,操作如下。

(1) 选择“分析”→“非参数检验”→“相关样本”菜单项。

(2) 在打开的对话框中选择“字段”选项卡:在“检验变量”列表框中选入 K_1 ~ K_8 这 8 个评分变量。

(3) 切换为自定义检验,选择 Kendall 协同系数检验。

(4) 单击“运行”按钮。

结果如图 15.26 所示。

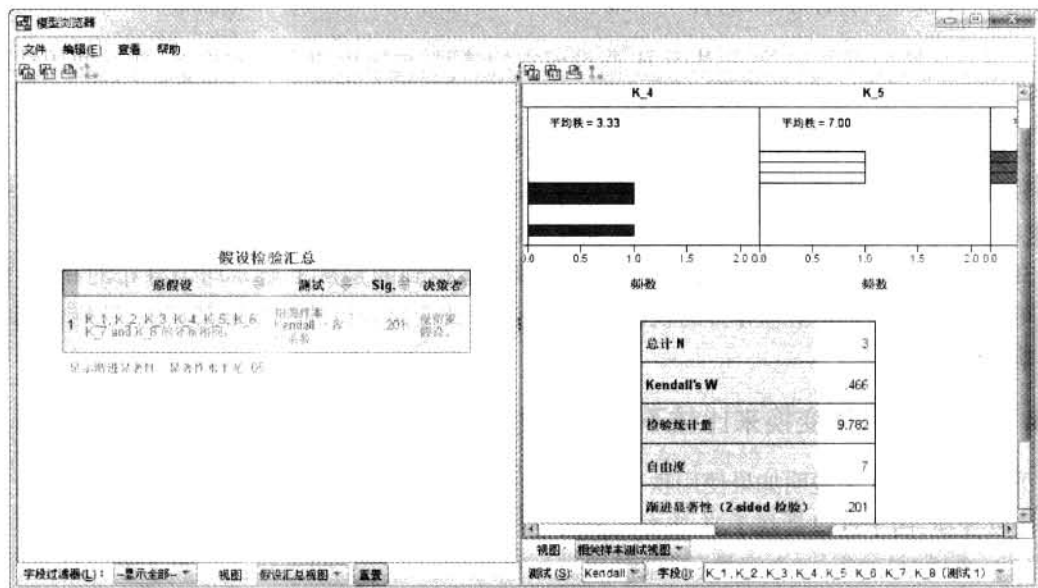


图 15.26 Kendall 协和系数检验的分析结果

从图 15.26 中可以看出检验的 P 值为 0.201, 该检验说明的是这 8 部电影的评价是否一致, 显然目前并未拒绝该假设, 也就是说, 尚无法认为这些电影的水平有差异。有兴趣的读者可以自行执行 Friedman 检验, 会发现两种方法的 P 值完全一致。同时结果中给出协和系数为 0.466, 说明 3 个影评家对 8 部电影的评判有一定的一致性, 但程度并不高。结果输出窗口的右上侧给出的是每部电影所得评分的平均秩和, 以及具体每一个评分的秩次, 可用于仔细观察影评家评分以及电影平均评价之间的差异。



这里的检验所回答的问题和希望解决的问题之间其实不太一致, 此处检验的实际上是 8 部电影的评价是否存在差异 (注意数据中变量的含义!), 而希望知道的是影评家的评价标准是否不同。因此在进行 Kendall 协和系数检验时, 应将注意力放在 W 系数值的大小, 而不是检验结果是否有统计学意义上。

2. Cochran 检验

还有很多时候在经济生活中比如民意调查或者市场调查中顾客对商品的信息反馈或满意度评价, 观察值是定性数据和二元 (0 ~ 1) 数据, 通常以“好”或“差”, “有效”或“无效”, “成功”或“失败”, “是”或“否”等形式出现, 如果用 Friedman 检验将会有很多打结现象, 即有许多相同的秩, 这时可以使用菜单中提供的 Cochran 检验。它是两个配对样本 McNemar 方法的推广, 只适用于二分类变量。

15.6 秩变换分析方法

本章前面已经介绍了很多非参数分析方法,但这些还远远不够,还有更多的问题无法解决,这里介绍一种通用的非参数分析原理,希望能对大家有所帮助。

15.6.1 秩变换分析原理简介

所谓秩变换分析方法,就是基于 H_0 假设成立的情况,先求出原变量的秩次,然后使用秩次代替原变量进行参数分析。当样本含量较大时,该方法的分析结果和相应的非参数方法基本一致,但该方法可以充分利用已知的参数方法,如多组样本的两两比较、多元回归等,从而大大扩展了非参数分析方法的范围。事实上,如果充分理解了前面讲述的各种秩和检验方法的原理就会发现,这些方法的实质都是秩变换方法的不同应用。

SPSS 中的秩过程可以用来求出秩次,该过程默认得到的是从 1 至 n 均匀分布的秩次,使用者也可以自行指定生成正态分布的秩次,但由于进行秩变换分析的样本量都较大,这样做基本上不影响分析结果。

15.6.2 案例:用秩变换来比较不同时点的家庭经济感受值

下面用例 15.3 来说明如果使用秩变换,相应的分析流程和结果是怎样的。首先应当进行原始变量的秩变换,由于这里是基于 H_0 成立的假设进行秩变换的,因此不需要分组进行,操作如下。

- (1) 选择“转换”→“个案排序”菜单项。
- (2) 在打开的对话框中的“变量”列表框中选入 Qa3。
- (3) 单击“确定”按钮。

上述操作会在数据集中生成新变量 RQa3,数值为 Qa3 的不分组秩次。

随后使用该变量进行标准的单因素方差分析,操作如下。

- (1) 选择“分析”→“比较均值”→“单因素方差分析”,打开“单因素方差分析”对话框。
- (2) 在“因变量”列表框中选入“RQa3”。
- (3) 在“因子”列表框中选入“月份[time]”。
- (4) 单击“两两比较”按钮:选中 S-N-K 方法;继续。
- (5) 单击“确定”按钮。

分析结果如图 15.27 所示。

Rank of Qa3					
	平方和	df	均方	F	显著性
组间	3277136.915	3	1092378.972	11.528	.000
组内	1.083E8	1143	94759.156		
总数	1.116E8	1146			

图 15.27 ANOVA

图 15.27 为对秩次进行方差分析的结果,可见家庭经济状况感受值的秩次在不同时点的差别是有统计学意义的。

图 15.28 为使用 S-N-K 法进行的不同时间点秩次的同质子集划分,如果和 15.4 节中的齐性子集分析结果相对比就会发现,子集划分结果是完全相同的,但子集内的检验 P 值略有差异,秩变换分析方法的 P 值更低一些。


	月份	N	alpha = 0.05 的子集	
			1	2
Student-Newman-Keuls ^{a,b}	200812	304	517. 90789	
	200712	304	535. 39309	
	200704	300		605. 32833
	200912	239		655. 12971
	显著性		. 499	. 054

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 283. 761。

b. 组大小不相等。将使用组大小的调和均值。不保证 I 类错误级别。

图 15.28 Qa3 的秩

 虽然这里由于两两比较方法不同,不宜严格进行 P 值大小的比较,但一般而言,秩变换方法的检验效能应当是不低于(即等同于或者高于)秩和检验的。

为了提高分析效率,还可以采用更复杂的变换方式,如要求生成的秩次服从正态分布,在随机区组设计数据中要求分组生成秩次等。因篇幅所限,本书不再深入,对此感兴趣的朋友可参见相关统计专业书籍。

15.7 本章小结

本章给出了几种常用的非参数方法的统计过程,在多数情况下,如果非参数检验结论为有统计学意义,相应的正确应用的参数检验结论大多与之相同。如果出现矛盾的情况,必须仔细考察参数检验的条件是否符合。当总体分布为非正态分布时,也无法通过适当的变量变换达到正态分布,甚至于分布类型未知时;对于诸如“18 岁以下”或“大于 2 000 元”等无法精确测量的数据,以及数据是分类数据、样本量很小的情况,传统的参数检验方法作用将变得非常有限甚至无能为力,这时可以转而使用非参数统计检验。

非参数检验方法中最常用的是等级次序或符号秩,这样做方法简单,易于理解。但是由于没有利用实际数值,又会损失部分信息,因而检验的有效性就比较差。现将本章介绍的几种非参数方法简单总结如下。

(1) 关于两个独立样本的非参数检验,Mann-Whitney U 检验是功效最强、应用最广的非参数检验。其零假设和备择假设的基础是:如果两个样本有差异,它们的中心位置将不同。

(2) 关于两个配对样本的非参数检验。最常用的是 Wilcoxon 配对秩和检验,它是对 Sign 符号检验正负号的改进,其基本思想是:若检验假设成立,则两组的秩和不应相差太大。不仅考虑了样本配对数据差异的方向,同时又考虑到差数的顺序。

- (3) 关于多个独立样本的非参数检验。SPSS 提供了 Kruskal-Wallis 检验和 Median 中位数法等。
- (4) 关于多个配对样本的非参数检验。SPSS 提供了 Friedman 检验和 Kendall 协和系数以及 Cochran 检验方法。

思考与练习

1. 在熟悉假设检验的思想基础上,比较参数检验与非参数检验的适用条件,并根据某一种具体的检验方法举例。
2. 在关于放松(比如听音乐等)对成年女性入睡所需时间影响的研究中,抽取了 10 名女性组成样本。题表 1 给出了 10 个对象在有放松条件和无放松条件下入睡所需的时间(min)。就此数据可以得出什么结论?

题表 1

研究对象	无放松	有放松
1	15	10
2	12	10
3	22	12
4	8	11
5	10	9
6	7	5
7	8	10
8	10	7
9	14	11
10	9	6

3. 对于一个由冬季各月其中的某些天数组成的样本和一个由夏季各月其中的某些天数组成的样本,警察记录了如题表 2 所示的每日犯罪报告的数据。给定 0.05 的显著性水平,判断犯罪报告数量在冬季数月与夏季数月之间是否有显著的差异?

题表 2

冬季	夏季	冬季	夏季
18	28	20	29
20	18	12	23
15	24	16	38
16	32	19	28
21	18	20	18

4. 一名证券经纪人收集到了某年三大公司的股票每股所能获利的钱数,如题表 3 所示。

题表 3

计算机公司	1.94	2.76	8.95	3.23	3.04	0.69	1.52
药品公司	7.89	1.65	2.59	1.09	-1.70		
公共服务公司	2.26	4.66	2.22	1.77	-0.15		

试比较这 3 种不同类型的公司股票所挣的钱是否相同。

5. 在做一个智力游戏时,人们认为它与年龄以及是否是盲人有关,现以年龄为区组,研究该游戏与眼睛看见与否是否有关。首先第1组安排天生眼盲的儿童参加游戏,第2组安排眼睛正常但做游戏时把眼睛蒙上的儿童参加游戏,第3组安排眼睛正常而且不蒙住眼睛的儿童参加游戏,观察他们的得分,如题表4所示,试就此进行分析。

题表4

	年 龄											
	1	2	3	4	5	6	7	8	9	10	11	12
盲人	0	0	0	0	1	8	8	8	0	8	8	8
蒙眼	0	8	0	0	2	8	5	6	8	8	3	8
不蒙眼	8	1	8	8	0	8	8	8	8	8	8	8

第 16 章 无序分类变量的统计推断——卡方检验

通过前面的介绍可以知道,变量可被分为连续性变量和分类变量两大类,而后者又可被细分为有序、无序变量两种。对于各组所在总体定量变量的平均水平,可以使用 t 检验和方差分析方法进行比较,秩和检验则用于比较各组所在总体有序分类变量的分布情况是否相同。这里将要介绍的卡方检验主要用于无序分类变量的统计推断,是在应用的广泛程度上可以和 t 检验相媲美的另一种常用检验方法。

16.1 卡方检验概述

16.1.1 卡方检验的基本原理

1. 卡方检验的基本思想

卡方检验是以 χ^2 分布为基础的一种常用假设检验方法,它的无效假设 H_0 是:观察频数与期望频数没有差别。

该检验的基本思想是:首先假设 H_0 成立,基于此前提计算出 χ^2 值,它表示观察值与理论值之间的偏离程度。根据 χ^2 分布及自由度可以确定在 H_0 假设成立的情况下获得当前统计量及更极端情况的概率 P 。如果 P 值很小,说明观察值与理论值偏离程度太大,应当拒绝无效假设,表示比较资料之间有显著差异;否则就不能拒绝无效假设,尚不能认为样本所代表的实际情况和理论假设没有差别。

2. 卡方值的计算与意义

χ^2 值表示观察值与理论值之间的偏离程度。计算这种偏离程度的基本思路如下。

(1) 设 A 代表某个类别的观察频数, E 代表基于 H_0 计算出的期望频数, A 与 E 之差称为残差。

(2) 显然,残差可以表示某一个类别观察值和理论值的偏离程度,但如果将残差简单相加以表示各类别观察频数与期望频数的差别,则有一定的不足之处。因为残差有正有负,相加后会彼此抵消,总和仍然为 0,为此可以将残差平方后求和。

(3) 另一方面,残差大小是一个相对的概念,相对于期望频数为 10 时,期望频数为 20 的残差非常大,但相对于期望频数为 1 000 时 20 的残差就很小了。考虑到这一点,人们又将残差平方除以期望频数再求和,以估计观察频数与期望频数的差别。

进行上述操作之后,就得到了常用的 χ^2 统计量,由于它最初是由英国统计学家 Karl Pearson 在 1900 年首次提出的,因此也称之为 Pearson χ^2 ,其计算公式为

$$\chi^2 = \sum \frac{(A - E)^2}{E} = \sum_{i=1}^k \frac{(A_i - E_i)^2}{E_i} = \sum_{i=1}^k \frac{(A_i - np_i)^2}{np_i} \quad (i=1, 2, 3, \dots, k)$$

其中, A_i 为 i 水平的观察频数, E_i 为 i 水平的期望频数, n 为总频数, p_i 为 i 水平的期望频率。 i 水

平的期望频数 T_i 等于总频数 $n \times i$ 水平的期望概率 p_i , k 为单元格数。当 n 比较大时, χ^2 统计量近似服从 $k-1$ (计算 E_i 时用到的参数个数) 个自由度的卡方分布。



作为学术界的领袖, Pearson 先生当初发表在《哲学杂志》上的 χ^2 论文题目为: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling。

由卡方的计算公式可知, 当观察频数与期望频数完全一致时, χ^2 值为 0; 观察频数与期望频数越接近, 两者之间的差异越小, χ^2 值越小; 反之, 观察频数与期望频数差别越大, 两者之间的差异越大, χ^2 值越大。换言之, 大的 χ^2 值表明观察频数远离期望频数, 即表明远离假设。小的 χ^2 值表明观察频数接近期望频数, 接近假设。因此, χ^2 是观察频数与期望频数之间距离的一种度量指标, 也是假设成立与否的度量指标。如果 χ^2 值“小”, 研究者就倾向于不拒绝 H_0 ; 如果 χ^2 值大, 就倾向于拒绝 H_0 。至于 χ^2 在每个具体研究中究竟要大到什么程度才能拒绝 H_0 , 则要借助于卡方分布求出所对应的 P 值来确定。

3. 卡方检验的样本量要求

卡方分布本身是连续型分布, 但是在分类资料的统计分析中, 显然频数只能以整数形式出现, 因此计算出的统计量是非连续的。只有当样本量比较充足时, 才可以忽略两者间的差异, 否则将可能导致较大的偏差。具体而言, 一般认为对于卡方检验中的每一个单元格, 要求其最小期望频数均大于 1, 且至少有 4/5 的单元格期望频数大于 5, 此时使用卡方分布计算出的概率值才是准确的。如果数据不符合要求, 可以采用确切概率法进行概率的计算。

16.1.2 卡方检验的用途

卡方检验最常见的用途就是考察某无序分类变量各水平在两组或多组间的分布是否一致。实际上, 除了这个用途之外, 卡方检验还有更广泛的应用。具体而言, 其用途主要包括以下几个方面。

(1) 检验某个连续变量的分布是否与某种理论分布相一致。如是否符合正态分布、是否服从均匀分布、是否服从 Poisson 分布等。

(2) 检验某个分类变量各类的出现概率是否等于指定概率。如在 36 选 7 的彩票抽奖中, 每个数字出现的概率是否各为 $1/36$; 掷硬币时, 正反两面出现的概率是否均为 0.5。

(3) 检验某两个分类变量是否相互独立。如吸烟(二分类变量: 是、否) 是否与呼吸道疾病(二分类变量: 是、否) 有关; 产品原料种类(多分类变量) 是否与产品合格(二分类变量) 有关。

(4) 检验控制某种或某几种分类因素的作用以后, 另两个分类变量是否相互独立。如在上例中, 控制性别、年龄因素影响以后, 吸烟是否和呼吸道疾病有关; 控制产品加工工艺的影响后, 产品原料类别是否与产品合格有关。

(5) 检验某两种方法的结果是否一致。如采用两种诊断方法对同一批人进行诊断, 其诊断结果是否一致; 采用两种方法对客户进行价值类别预测, 预测结果是否一致。

本章主要介绍卡方检验的后 4 种应用, 有关分布检验的内容可参看相关章节。

16.1.3 SPSS 中的相应功能

由于卡方检验的用途很广, 因此在 SPSS 中经常用到, 但在很多地方都会以分布检验、方差齐

性检验等其他检验的名义出现(或者说这些检验方法的统计量是服从卡方分布的),直接以卡方检验的名称显示的主要是以下两处。

1. 非参数分布检验中的卡方检验

准确地说,这里提供的就是检验某个分类变量各类的出现概率是否等于指定概率的分布检验。

2. 交叉表过程

主要用于针对两个/多个分类变量的交叉表进行其关联程度的卡方检验,并可进一步计算出关联程度指标等,上面提到的卡方检验用途中的后3项都可以在该过程中实现,而人们一般所说的卡方检验也就是指该过程中的相应功能。

16.2 单样本案例:考察抽样数据的性别分布

在第12章中已经讲到,如果希望考察某个二分类变量的分布是否服从假设分布,可以考虑使用二项分布检验。但无论是两分类还是多分类,实际上都可以被归结为:从已知的样本数据出发,来判断总体各取值水平出现的概率是否与已知概率相符,即该样本是否的确来自已知的总体分布。这就是本节所说的单样本率与总体率的比较,也有人称它为拟合问题,在统计学上可以利用(单样本)卡方检验来回答此问题。



在实践工作中,有很多单样本率与总体率进行比较的例子。如骰子是否公平,检验各面出现的频率是否各等于 $1/6$;检验彩票中奖号码的分布是否均匀分布,以检验彩票开奖是否作弊;国家人口老龄化问题是否更严重了;某产品的市场占有率是否较以前更大;某病的发病率是否较前降低等。

16.2.1 用新对话框界面分析本案例

例 16.1 在第12章中用二项分布检验考察了2007年4月的性别分布是否均衡,这里使用卡方检验来完成相同的任务。

在有了第12章的分析基础之后,相应的对话框等均不再重复解释,操作如下。

- (1) 选择“数据”→“选择个案”菜单项。
- (2) “选择”框组:选中“如果条件满足”单选按钮。
- (3) 单击“如果”按钮,打开“如果”子对话框,设置 $\text{time} = 200704$ 。
- (4) 选择“分析”→“非参数检验”→“单样本”菜单项。
- (5) 在打开的对话框中选择“目标”选项卡,选中“自动比较观察数据和假想数据”单选框。
- (6) 在“字段”选项卡中设置使用定制字段分配,将“S2 性别”选入“检验字段”列表框中。
- (7) 在“设置”选项卡中选择“自定义检验”选项组中的第二项“卡方检验”,相应选项中的类别概率已经是所需的“所有类别概率相等”了,如图16.1所示因此不需要更改。
- (8) 运行。

最终的模型输出如图16.2所示,有兴趣可以和二项分布检验的结果相对比,会发现非常相似。

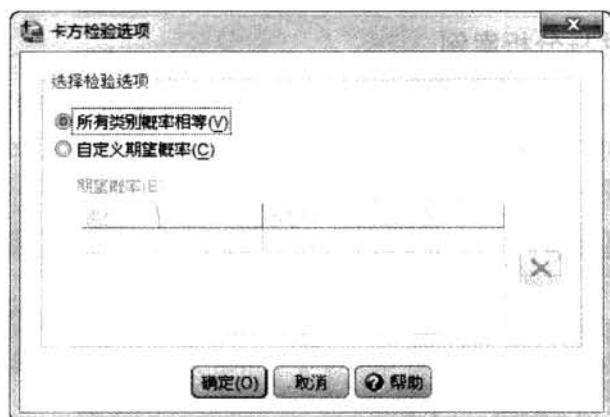


图 16.1 “卡方检验选项”子对话框

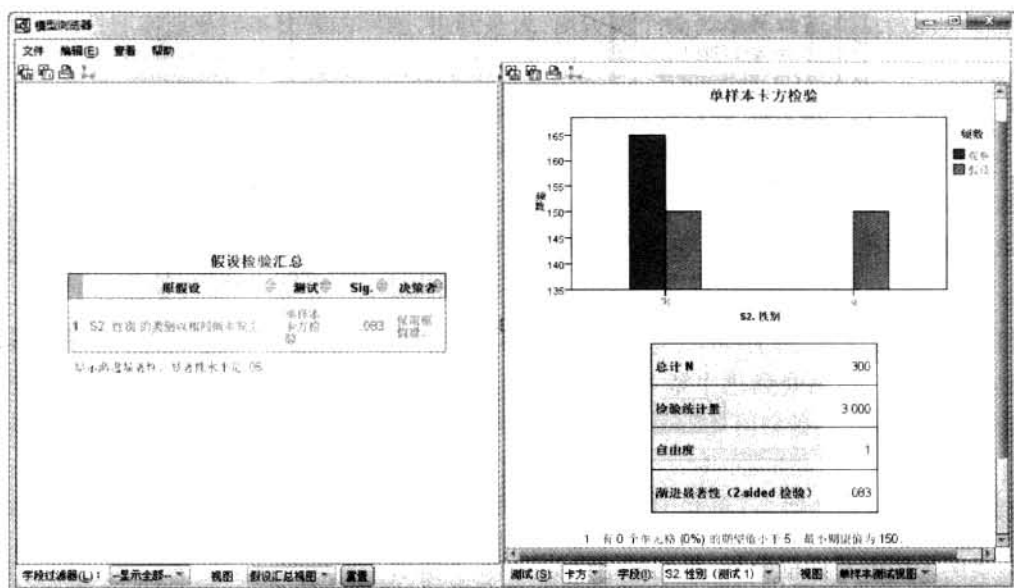


图 16.2 单样本卡方检验的分析结果

(1) 右侧辅助视图中给出了男性单元格的实际样本频数为 165 例, 而理论频数为 150, 因此该单元格的残差为 15, 相应的女性单元格的残差为 -15, 因此整个样本相应的标化残差平方和也就是卡方统计量, 即 $15^2/150 + (-15)^2/150 = 3$ 。

(2) 右侧辅助视图中已经给出了卡方统计量的数值 3.0, 最终基于样本和无效假设推导出的 P 值为 0.083, 因此不能拒绝无效假设, 尚不能认为 CCSS 抽样数据的性别分布有差异。

(3) 左侧是汇总视图的输出, 结论相同。



读者可以发现这里卡方检验的 P 值为 0.083, 和二项分布检验的 0.094 略有差异, 这是因为此处卡方检验给出的是近似 P 值, 而二项分布检验给出的是确切 P 值。由于本案例样本量充足, 因此两者差异不大。但如果要论正确性, 则显然二项分布检验的 P 值是更准确的。

16.2.2 使用旧对话框分析案例

选择“分析”→“非参数检验”→“旧对话框”→“卡方”菜单项,就会打开卡方分布的检验对话框,如图 16.3 所示,该中文界面非常简明,而且几乎所有内容都和第 12 章中介绍过的二项分布检验对话框相似,因此不再重复解释。



图 16.3 “卡方检验”对话框

本案例采用旧对话框分析,得到的结果如图 16.4 所示。

	观察数	期望数	残差
男	165	150.0	15.0
女	135	150.0	-15.0
总数	300		

图 16.4 S2. 性别

图 16.4 给出了样本中两个性别的观察频数、根据总体构成比计算出的期望频数,以及观察频数与期望频数之差——残差,和前面的表格相对应,应不难理解。

图 16.5 给出的是单样本卡方检验的结果,具体包括卡方统计量 (Chi-Square)、自由度 (df), 以及对应的卡方分布近似概率值 (Asymp. Sig)。可见 χ^2 统计量为 3.0,自由度为 1,对应的 P 值为 0.083,结果和前面完全相同。最下方的注解则就卡方检验所要求的 (单元格) 样本量进行了评估,可见本案例的样本量是满足需求的。

	S2. 性别
卡方	3.000 ^a
df	1
渐近显著性	.083


a. 0 个单元 (.0%) 具有小于 5 的期望频率。单元最小期望频率为 150.0。

图 16.5 检验统计量

16.3 两样本案例:不同收入级别家庭的轿车拥有率比较

前面介绍了样本率与已知总体率的检验方法,显然,其中所使用的卡方检验原理可以非常容易地推广到两样本或多样本比较的问题,也就是说,比较两个或多个样本所在总体的另一个分类变量的发生率/构成比是否相同,这在统计中都可以用卡方检验来分析。

例 16.2 在 CCSS 的分析报告中,所有受访家庭会按照家庭年收入被分为低收入家庭和高收入家庭两类,现希望考察不同收入级别的家庭其轿车拥有率是否相同。



需要注意的是,卡方检验仅仅告知使用者各类别的构成/分布是否相同,如果交叉表中存在有序分类变量,则使用卡方检验并不合适,而应当用第 15 章介绍的秩和检验方法加以分析。

该问题可以被简化为两组收入不同的受访家庭其轿车拥有率的比较,由前面的介绍可知,这应当采用交叉表过程来完成。虽然交叉表过程的界面在第 8 章中已经介绍过,但由于该过程中所涉及的卡方检验相关功能比较丰富,因此这里将进一步讲解一下和检验有关的对话框功能。

1. “统计量”子对话框界面说明

该对话框不仅包括常用的卡方检验,还包含了一大批用于度量行、列变量关联度的指标,如图 16.6(a)所示,这里主要介绍本章将会用到的一些方法。

(1) “卡方”复选框:进行卡方检验,对于 4 格表资料还会自动给出校正卡方检验和确切概率法的结果。

(2) “Kappa”复选框:计算 Kappa 值,即内部一致性系数。这是医学中非常常用的一致性指标,取值在 0~1 之间,除根据 P 值判断一致性有无统计学意义外,根据经验,Kappa ≥ 0.75 表明两者一致性较好;0.4 ≤ Kappa < 0.75 表明一致性一般;Kappa < 0.4 则表明两者一致性较差。

(3) “风险”复选框:计算 OR 值(比数比)和 RR 值(相对危险度),这些指标用于反映交叉表的行、列变量之间的关联强度。

(4) “McNemar”复选框:进行 McNemar 检验,即常用的配对卡方检验。

(5) “Cochran's and Mantel-Haenszel 统计量”复选框:为两个二分类变量进行分层卡方检验,即层间的独立性检验和同质性(齐性)检验,同时可进行分层因素的调整。该复选框下方的文本框用于设定相应 H₀假设的 OR 值,默认为 1。



图 16.6 交叉表过程的“统计量”子对话框和“单元显示”子对话框

2. “单元显示”子对话框界面说明

该子对话框用于定义交叉表单元格中需要显示的指标。

- (1) “计数”复选框组: 是否输出实际观察数 (Observed) 和理论频数 (Expected)。
- (2) “残差”复选框组: 选择残差的显示方式, 可以是实际数与理论数的差值 (Unstandardized)、标准化后的差值 (Standardized, 将差值转化为标准正态分布), 或者被标准误差的单元格残差 (Adj. Standardized)。
- (3) “百分比”复选框组: 是否输出行百分数 (Row)、列百分数 (Column) 以及合计百分数 (Total)。
- (4) “z-检验”框组: 对于组数超过两组的样本率的比较, 在整个交叉表的卡方检验有统计学意义之后, 后续的问题和方差分析非常类似, 也面临着组间两两比较的问题。在新版本的 SPSS 中, 率的两两比较已经可以通过选中“单元显示”子对话框中提供的“z-检验”复选框实现, 且可以进一步要求在两两比较中进行检验水准的 Bonferroni 调整。
- (5) “非整数权重”框组: 当所分析的数据为加权数据, 且权重变量可能有小数取值时, 将会导致单元格内的观测频数也出现小数, 该框组用于确定此时对小数权重的处理方式。

3. 操作说明与结果解释

由于已经有了第 8 章的基础, 这里的操作思路就非常清楚了, 具体如下。

- (1) 选择“分析”→“描述统计”→“交叉表”菜单项, 打开“交叉表”对话框。
- (2) 在“行”列表框中选中“家庭收入两级 Ts9”。
- (3) 在“列”列表框中选中“是否拥有家庭轿车 O1”。
- (4) 单击“单元显示”按钮, 打开“单元显示”子对话框, 选中“行百分比”复选框。
- (5) 单击“统计量”按钮, 打开“统计量”子对话框, 选中“卡方”复选框。

(6) 单击“确定”按钮。
相应的分析结果如图 16.7 所示。

			01. 是否拥有家用轿车		合计
			有	没有	
家庭收入 2 级	Below 48,000	计数	32	303	335
		家庭收入 2 级 中的 %	9.6%	90.4%	100.0%
	Over 48,000	计数	225	429	654
		家庭收入 2 级 中的 %	34.4%	65.6%	100.0%
合计		计数	257	732	989
		家庭收入 2 级 中的 %	26.0%	74.0%	100.0%

图 16.7 家庭收入 2 级 * 01. 是否拥有家用轿车 交叉制表

首先给出的是家庭收入分级和轿车拥有情况的交叉表,可见低收入家庭中只有 10% 拥有轿车,而中高收入家庭中有 34% 拥有轿车,样本数据的差异很明显,但该差异是否具有统计学意义尚需进行检验。

图 16.8 即为卡方检验结果表,其中给出了多种检验结果,在解释这些检验结果之前,先来说明一下最下方的脚注内容:在该 4 格表中,没有单元格(0%)的期望频数少于 5,其中期望频数最少的那个单元格的期望频数为 87.05。该脚注充分说明本样本的样本量(及其单元格分布)完全满足 Pearson 卡方的要求,因此可以放心地阅读最常用的 Pearson 卡方的检验结果。

	值	df	渐进 Sig. (双侧)	精确 Sig. (双侧)	精确 Sig. (单侧)
Pearson 卡方	71.134 ^a	1	.000		
连续校正 ^b	69.848	1	.000		
似然比	80.146	1	.000		
Fisher 的精确检验				.000	.000
线性和线性组合	71.062	1	.000		
有效案例中的 N	989				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 87.05。
b. 仅对 2 × 2 表计算。

图 16.8 卡方检验

下面是图 16.8 中详细输出的内容的解释。

(1) Pearson 卡方:最标准,也是最常用的卡方检验结果,当样本量充足时使用。

(2) 连续校正:由统计学家 Frank Yates 提出,故也称为 Yates 校正。本法只适用于 4 格表资料,在样本含量大于 40,所有单元格的期望频数均大于 1,且只有 1/5 以下的单元格的期望频数小于 5 大于 1 时,要对卡方统计量进行连续性校正。近年来蒙特卡罗随机模拟表明,Yates 校正似乎有一点矫枉过正,但在实践工作中依然经常用到。

(3) Fisher 的精确检验:对于 4 格表资料,即使不选中 Exact 子对话框中的 Exact 复选框,SPSS 也会给出 Fisher 精确概率法检验结果。如果安装了 SPSS Exact Test 模块,并在对话框中指定进行 Exact 检验时,对其他列联表也会给出 Fisher 精确概率检验结果。与 Pearson 卡方和似然比卡方相比,确切概率法的优点在于不需要近似,结果最准确,但计算时消耗的资源多。在样本含量小于 40,或有格子的期望频数小于 1 的 4 格表中,需要用 Fisher 精确概率法。对于其他列联

表,如果有单元格的期望频数小于1,或大于1小于5的期望频数较多,也可以采用该法。

(4) 似然比(Likelihood Ratio):与Pearson卡方相比,检验的是同样的 H_0 假设,即行变量与列变量之间相互独立,不同的是卡方的计算公式不一样,在处理多维表时有更大的优势。在大多数情况下,两者的结论是基本一致的。

(5) 线性卡方(Linear by Linear):检验的 H_0 假设是行变量与列变量之间无线性相关。在列联表分类变量中很少用,更多用于连续变量。

16.4 两分类变量间关联程度的度量

卡方检验可以从定性的角度告诉用户两个变量是否存在关联,当拒绝 H_0 时,在统计上有把握认为两个变量存在关联。但接下来的问题是,如果变量之间存在相关性,它们之间的关联强度有多大,有没有什么指标可以客观表示其大小?例如进行一个客户满意度的研究,研究者发现价格、质量、服务都与总体满意度相关,但哪项与总体满意度关系更密切一些呢?如果想要提高客户满意度,最需要做的是调整价格、提高服务水平,还是改进产品质量?这里就来深入探讨一下对分类变量关联程度的度量方式。

针对不同的变量类型,在SPSS中可以计算各种各样的相关指标,而且交叉表过程也对此提供了完整的支持,但此处只涉及测量两分类变量间关联强度的指标,更系统的相关程度指标体系介绍参见第17章。

16.4.1 相对危险度与优势比

在实际应用中,卡方值的大小可以粗略地反映两变量联系的强弱,但是这很难有更贴近实际的解释,只能从它的大小上获得一个关联强弱的印象。但是如果有一个指标能够告诉研究者:男性和女性相比,购买该产品的可能性是女性的3倍,这就非常容易理解。相对危险度(Relative Risk, RR)和优势比(Odds Ratio, OR,也翻译成比数比)就可以满足这一要求,它们与其他关联测量参数的最大不同之处在于,RR值和OR值关心的是行变量某一水平和列变量某一水平相对于基础水平的关联程度,即不同水平间的比较,而上述的关联测量参数关心的则是行变量各水平和列变量各水平的关联程度。

1. 相对危险度

RR值是一个概率的比值,是指实验组人群反应阳性概率与对照组人群反应阳性概率的比值。用公式表示为

$$RR = \frac{P_i}{P_c} = \frac{a/n_i}{c/n_c}$$

其中, P_i 为实验组人群反应阳性概率, P_c 为对照组人群反应阳性概率, n_i 为实验组总人数, a 为实验组反应阳性人数, n_c 为对照组总人数, c 为对照组反应阳性人数。RR值用于反映实验因素与反应阳性的关联程度。取值范围从0到无限大。数值为1时,表明实验因素与反应阳性无关联;小于1时,表明实验因素导致反应阳性的发生率降低;大于1时,表明实验因素导致反应阳性的发生率增加。

2. 优势比

显然,RR 的解释非常容易理解,但是 RR 的计算要求得到各组的反应概率,由于在回顾性研究中很难求得人群反应概率的估计值,因此也无法进行 RR 值的估计,此时研究者往往使用 OR 值代替 RR 值,来反映实验因素与对照因素的关联强度。OR 值是一个比值的比,是反应阳性人群中实验因素有无的比例与反应阴性人群中实验因素有无的比例之比。计算公式可以表示为

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

其中,a 为反应阳性组实验因素阳性人数,b 为反应阳性组实验因素阴性人数,c 为反应阴性组实验因素阳性人数,d 为反应阴性组实验因素阴性人数。显然,如果 OR 大于 1,则说明该试验因素更容易导致结果为阳性。或者说采用的试验因素和结果为阳性有关联。



由于优势比是两个比值的比值,因此它不太好解释,而解释相对危险度则要容易得多,因此在大多数情况下人们希望能够按照相对危险度的含义来解释优势比。当所关注的事件发生概率比较小时(<0.1),优势比可作为相对危险度的近似。

16.4.2 案例:计算家庭收入级别和轿车拥有情况的关联程度

16.3 节中已经对家庭收入级别和轿车拥有情况的 4 格表做了卡方检验,结果显示两者之间存在联系,中高收入家庭的轿车拥有比例更高。另外,还可以使用 RR、OR 等一系列指标来对其关联强度加以定量描述。利用 16.4.1 节所介绍的知识,可以手工计算出 OR、RR 等关联强度指标,还可以利用 SPSS 软件直接求得相应的数值,操作如下。

- (1) 选择“分析”→“描述统计”→“交叉表”菜单项,打开“交叉表”对话框。
- (2) 在“行”列表框中选中“家庭收入两级”。
- (3) 在“列”列表框中选中“是否拥有家庭轿车”。
- (4) 单击“统计量”按钮,打开“统计量”子对话框,选中“风险”复选框。
- (5) 单击“确定”按钮。



由于本案例不是前瞻性的研究设计,因此严格地说,这里得到的只是两种人群的轿车拥有比例,而不是购买概率,所得到的 RR 值在解释上和其初衷是不同的。但是现在 RR 的概念应用很广,虽然很多时候的应用和定义相比确实存在差异,但用 RR 表述影响因素的作用强度的确是很常见的。

分析结果如图 16.9、图 16.10 所示。

			01. 是否拥有家用轿车		合计
			有	没有	
家庭收入 2 级	Below 48,000	计数	32	303	335
		家庭收入 2 级 中的 %	9.6%	90.4%	100.0%
	Over 48,000	计数	225	429	654
		家庭收入 2 级 中的 %	34.4%	65.6%	100.0%
合计		计数	257	732	989
		家庭收入 2 级 中的 %	26.0%	74.0%	100.0%

图 16.9 家庭收入 2 级* 01. 是否拥有家用轿车交叉制表

	值	95% 置信区间	
		下限	上限
家庭收入 2 级 (Below 48,000/Over 48,000) 的几率比	.201	.135	.300
用于 cohort O1. 是否拥有家用轿车 = 有	.278	.196	.392
用于 cohort O1. 是否拥有家用轿车 = 没有	1.379	1.291	1.472
有效案例中的 N	989		

图 16.10 风险估计

由图 16.9、图 16.10 可以看出：

(1) 优势比 OR 是两个比数的比。某个事件的比数是它发生的概率除以不发生的概率。在本例中,低收入家庭拥有家庭轿车的比数是 $9.6\%/90.4\% = 0.106$,中高收入家庭拥有家庭轿车的比数是 $34.4\%/65.6\% = 0.524$,则 OR 值等于 $0.106/0.524 = 0.201$,该指标的 95% CI 同样不包括 1,说明该数值的确是不等于 1 的(有统计学差异)。

(2) 对于不同收入的家庭而言,其拥有家庭轿车的相对危险度是两组人群拥有轿车的概率之比,其估计值是 $9.6\%/34.4\% = 0.278$,即低收入家庭拥有轿车的概率是中高收入家庭的 0.278 倍,或者倒过来讲,中高收入家庭拥有家庭轿车的概率是低收入家庭的 $1/0.278 = 3.597$ 倍。且其 95% CI 不包括 1,具有统计学意义。

(3) 相应的,不同收入家庭不拥有家庭轿车的概率则是两个人群不拥有轿车的概率之比,其估计值为 $90.4\%/65.6\% = 1.379$,即低收入家庭不拥有轿车的概率是中高收入家庭的 1.379 倍(当然从这个案例的背景而言更应关心的是 0.278 这个数据),该数值的 95% CI 同样也不包括 1。



上述 3 个指标的假设检验实际上完全等价,此外 OR 的数值也等于有车与无车的相对危险度的比值($0.278/1.379 = 0.201$)。

16.5 一致性检验与配对卡方检验

16.5.1 Kappa 一致性检验

在 Pearson 卡方检验中对行变量和列变量的相关性已经做了检验,其中行变量和列变量是一个事物的两个不同属性,如果两个变量独立,可以期望第 i 行 j 列单元格中的频数为 $n \times p_i \times p_j$,其中 n 为总观察频数, p_i 为第 i 行的概率, p_j 为第 j 列的概率。

还有一种列联表,其行变量和列变量反映的是一个事物的同一属性的相同水平,只是对该属性各水平的区分方法不同,这相当于在研究设计中采用了配对设计。例如在一张表内显示某病的诊断结果,行变量为一种诊断方法,列变量为另一种诊断方法;或者在一张表内显示对某事物的评价等级,行变量和列变量分别显示不同裁判员的评价。如果希望检验这两种区分同一属性的方法给出的结果是否一致,则不应当使用 Pearson 卡方检验,因为 Pearson 卡方检验并不适用于这种配对设计的数据,它无法明确说明结果的一致程度如何。此时,可以采用 Kappa 一致性检验对两种方法结果的一致程度进行评价。



更准确地说, Pearson 卡方只能告诉用户两种测量结果之间是否存在关联,但不能判断其是否具有一致性。举一个简单的例子,如果对于诊断者甲分别诊断为轻度、中度、重度疾病的患者,诊断者乙一律会分别诊断为中度、重度、轻度,则两者的诊断结果显然不具有-致性,但如果使用卡方检验,则是有统计学意义的,因为它们两者的诊断结果的确存在关联!

例 16.3 某公司期望扩展业务,增开几家分店,但对开店地址不太确定。于是选了 20 个地址,请两位资深顾问分别对 20 个地址进行评价,将它们评为好、中、差 3 个等级,以便确定应对哪些地址进行更进一步调查,那么这两位资深顾问的评价结果是否一致? 数据参见 site. sav。

在 SPSS 中,依然用“交叉表”对话框,将两个顾问的评价结果分别作为行变量或列变量,并在“统计量”子对话框中指定进行 Kappa 统计分析。另外因本例样本量很小,故要求计算确切概率以保证结果的正确性。在 SPSS 中进行的具体操作如下。

- (1) 选择“数据”→“加权个案”菜单项。
- (2) 在打开的对话框中选中“加权个案”单选框。
- (3) 将“频数”选入“频率变量”列表框中。
- (4) 选择“分析”→“统计描述”→“交叉表”菜单项,打开“交叉表”对话框。
- (5) 在“行”列表框中选入“cons1”。
- (6) 在“列”列表框中选入“cons2”。
- (7) 单击“统计量”按钮,打开“统计量”子对话框,选中“Kappa”复选框。
- (8) 单击“确定”按钮。

结果如图 16.11、图 16.12 所示。

计数

		顾问二的评价			合计
		差	中	好	
顾问一的评价	差	6	0	0	6
	中	5	2	2	9
	好	1	0	4	5
合计		12	2	6	20

图 16.11 顾问一的评价*顾问二的评价交叉制表

	值	渐进标准误差 ^a	近似值 T ^b	近似值 Sig.
一致性度量 Kappa	.429	.131	3.333	.001
有效案例中的 N	20			

a. 不假定零假设。

b. 使用渐进标准误差假定零假设。

图 16.12 对称度量

注意这里 Kappa 检验的 H_0 假设是: $Kappa = 0$, 即两者完全无关。图 16.12 显示 Kappa 值为 0.429, P 值为 0.001, 拒绝 H_0 假设(两位顾问的评价结果不一致), 接受 H_1 假设, 认为两位顾问的评价结果是存在一致性的。但是根据经验, 一般认为当 $Kappa \geq 0.75$ 时两者的一致性较好; $0.4 \leq Kappa < 0.75$ 时一致性一般, $Kappa < 0.4$ 时两者一致性较差。此处的估计值为 0.429, 因此实

际上本例中数据的一致性并不是很强。特别是有一个地址两人竟给出了完全对立的评价。

一致性检验在医学研究中用得很多。如在一种简单易行的诊断方法是否可替代另一种结果可靠但操作繁杂的诊断方法的研究中,就会用到一致性检验。另外,在数据分析中,比较两种预测方法预测结果的一致性时也可能用到 Kappa 检验。


16.5.2 配对卡方检验

通过 Kappa 检验已经回答了两种测量间究竟有无关联的问题,但是通过对列联表的观察会发现,两位顾问的评价似乎不太一样,这种问题又如何来加以分析? McNemar 配对卡方检验就是经典的配对检验,专门用于解决此类问题。在统计量子对话框的左下角就是 McNemar 复选框,例如上例,选择后相应的结果输出如图 16.13 所示。


	值	df	渐进 Sig. (双侧)
McNemar-Bowker 检验	8.000	3	.046
有效案例中的 N	20		

图 16.13 配对卡方检验

此处的原假设为:两顾问的评价结果无差别,显然, P 值小于 0.05,因此拒绝了该假设,认为应当是有差别的,从样本数据看,应当是第一个顾问倾向于评价得更高。

 对于 4 格表数据,配对卡方会直接使用等价于二项分布检验的 McNemar 检验给出确切 P 值。但是对于较大的表格,则只能使用近似的 McNemar-Bowker 检验给出近似 P 值。

现在,Kappa 检验显示两者的评价存在一致性,而配对卡方检验则显示两者的结果是有差别的。实际上,这两个结论并不矛盾,参考前面对 Kappa 值的评价方式就可以理解。另外,这两者在信息的利用上也有差异。Kappa 检验会利用列联表中的全部信息,而 McNemar 检验只会利用非主对角线单元格上的信息,即它只关心两者不一致的评价情况,用于比较两个评价者间存在怎样的倾向。

 在应用中,对于一致性较好,即绝大多数数据都在主对角线上的列联表,McNemar 检验可能会失去实用价值。例如对 1 万个案例进行一致性评价,9 995 个都是完全一致的,在主对角线上,另有 5 个分布在左下侧的三角区,显然,此时一致性相当好。但如果使用 McNemar 检验,由于它并不考虑主对角线上的数据,只会利用上、下三角区的信息,此时反而会得出两种评价有差异的结论。而正确应用这些方法的关键点就在于弄清楚自己希望考察的究竟是一致性还是差异性。

16.6 分层卡方检验

在例 16.2 中,经卡方检验发现家庭收入级别的确会影响家庭轿车的拥有情况,随后又进一步计算出了两者的关联强度指标 OR 和 RR 。但在进一步考虑之后,研究者会发现还存在如下问题。

(1) 不同城市的轿车拥有情况是存在差异的,那么收入级别对其的影响在不同城市间是否存在差异?比如说在有的城市影响大些,而在其他城市的影响小些?

(2) 如果收入分级的影响在不同城市间没有差异,那么由于不同城市的人群的收入分布并不相同,直接将数据混合进行分析难免会影响结果的准确性。在考虑了此问题,或者说在控制了城市的混杂作用之后,校正后的 RR 或者 OR 应当是多少?

解决上述问题的统计方法有很多,而本节将要介绍的分层卡方检验就是最为基本和常用的一种。

分层卡方是把研究对象分解成不同层次,每层分别研究行变量与列变量的相关性。如按工资级别分成低、中、高层,分别研究低、中、高工资的人订购商品与邮件回应的关系;按受教育程度分成本科以下、本科、硕士、博士及以上,分别研究性别与职位类别的关系,借以排除这些分层因素(如工资级别、受教育水平)对行变量与列变量关联的干扰。分层因素在几个组之间的分布不均,既可能削弱了行变量与列变量间原本存在的关系,也可能使得原本不存在关系的两个变量的关系呈现统计学显著性。

例 16.4 在例 16.2 的基础上,进一步控制城市的影响,在控制城市影响的前提下得到更准确的家庭收入分级和轿车拥有情况的关联程度测量指标。

(1) 选择“分析”→“描述统计”→“交叉表”菜单项,打开“交叉表”对话框。

(2) 在“行”列表框中选中“家庭收入两级”。

(3) 在“列”列表框中选中“是否拥有家庭轿车”。

(4) 在“层”列表框中选中“城市 S0”。

(5) 单击“统计量”按钮,打开“统计量”子对话框,选中“风险”复选框、“Cochran's and Mantel-Haenszel 统计量”复选框。

(6) 单击“确定”按钮。

这里省略了分层交叉表的输出,直接给出了相应的分析结果。

图 16.14 为所选中的“风险”复选框所对应的输出,由于设定了分层变量,因此该输出会对每一层单独进行风险估计,并同时给出合计样本的风险估计。仅从 OR 值就可以看出,北京、上海、

		95% 置信区间		
S0. 城市		值	下限	上限
100 北京	家庭收入 2 级 (Below 48,000/Over 48,000) 的几率比	.156	.075	.326
	用于 cohort O1. 是否拥有家用轿车 = 有	.231	.121	.440
	用于 cohort O1. 是否拥有家用轿车 = 没有	1.477	1.308	1.666
	有效案例中的 N	319		
200 上海	家庭收入 2 级 (Below 48,000/Over 48,000) 的几率比	.089	.031	.251
	用于 cohort O1. 是否拥有家用轿车 = 有	.123	.046	.328
	用于 cohort O1. 是否拥有家用轿车 = 没有	1.384	1.261	1.519
	有效案例中的 N	337		
300 广州	家庭收入 2 级 (Below 48,000/Over 48,000) 的几率比	.333	.189	.586
	用于 cohort O1. 是否拥有家用轿车 = 有	.434	.275	.683
	用于 cohort O1. 是否拥有家用轿车 = 没有	1.302	1.151	1.474
	有效案例中的 N	333		
合计	家庭收入 2 级 (Below 48,000/Over 48,000) 的几率比	.201	.135	.300
	用于 cohort O1. 是否拥有家用轿车 = 有	.278	.196	.392
	用于 cohort O1. 是否拥有家用轿车 = 没有	1.379	1.291	1.472
	有效案例中的 N	989		

图 16.14 风险估计

广州三地的 OR 值虽然都不等于 1,但样本估计值并不相同,上海的 OR 值只有 0.089,而广州则高达 0.333。这种差异究竟代表的是抽样误差,还是真实存在的总体差异,仅靠普通的卡方检验/风险估计是无法回答的,这是分层卡方检验应当完成的任务。

分层卡方检验分析结果如图 16.15 所示,其中给出的是层间差异的检验结果,即不同层间收入级别与轿车拥有情况的联系强度是否相同,分别采用了两种检验方法,可见两者在本案例中结论相同,即在不同城市间,行 \times 列变量的联系强度并不相同,因此不应当考虑将不同城市的数据结合起来得到一个总的分析结果。

	卡方	df	渐进 Sig. (双侧)
Breslow-Day	6.165	2	.046
Tarone 的	6.161	2	.046

图 16.15 几率比的均一性检验



如果按照多变量统计模型的说法,上述结果实际上意味着城市这个影响因素和行 \times 列变量间存在交互作用,需要在模型中引入交互项。

图 16.16 给出的是分层卡方检验的结果,即考虑了(或者说去除了)分层因素的影响后,对行 \times 列变量关联强度的检验结果。图 16.16 中共给出 CMH 卡方检验和 MH 卡方检验两种结果,前者是后者的改进,可见 P 值均小于 0.05,即可以认为收入级别与轿车拥有情况有关联,但是由于前面的层间一致性检验结果为有统计学差异,因此这里的结论仅供参考。

	卡方	df	渐进 Sig. (双侧)
Cochran 的	72.397	1	.000
Mantel-Haenszel	70.879	1	.000

在条件的独立性假定下,仅当层数固定时 Cochran 的统计量才渐进分布为 1 df 卡方分布,而 Mantel-Haenszel 统计量始终渐进分布为 1 df 卡方分布。注意,当观测值和期望值差值之和为 0 时,将从 Mantel-Haenszel 统计量中删除连续校正。

图 16.16 条件的独立性检验

图 16.17 给出的是 OR_{MH} 值(调整了分层因素作用后的综合 OR 值)、 OR_{MH} 值的自然对数、可信区间及其相应的 P 值,可见统计检验结论和前面一致, $OR_{MH}=0.195$,即去除了不同分店的混杂效应后,和中高收入家庭相比,中低收入家庭拥有轿车的优势比为 0.195,或者说概率大约为前者的 1/5。当然,由于一致性检验有统计学差异,该结果也仅供参考。

估计			.195
ln(估计)			-1.636
ln(估计)的标准误差			.206
渐进 Sig. (双侧)			.000
渐进 .95% 置信区间	一般几率比	下限	.130
		上限	.292
	ln(一般几率比)	下限	-2.040
		上限	-1.232

Mantel-Haenszel 一般几率比估计在 1.000 假定的一般几率比下渐进地正态分布,因此是估计的自然对数。

图 16.17 Mantel-Haenszel 一般几率比估计

分层卡方检验是一种很好的控制其他因素的方法,从而能得到更准确的结果。如果数据量

足够大,还可以引入更多的分层因素加以控制。但是,和 SAS 中的 CMH 卡方不同,SPSS 提供的 CMH 卡方检验只能进行两分类变量的检验,而不能进行多分类变量的检验。这是因为分层卡方只对分层因素进行了简单的控制,当各层间效应的大小不同,或者说分层因素和要分析的变量间存在交互作用时,分层卡方检验就不再适用。而这种情况在多分类变量的分层分析中会经常遇到,此时应当使用对数线性模型或者 Logistic 模型来进行更为深入和准确的分析,关于这些方法可参见本丛书《SPSS 统计分析高级教程》的相关章节,这里不再详述。

16.7 本章小结

(1) 卡方检验是以卡方(χ^2)分布为基础的一种常用假设检验方法,常用于计数资料的显著性检验。其基本思想是:首先假设观察频数与期望频数没有差别。而统计量 χ^2 值表示观察值与理论值之间的偏离程度。当 n 比较大时, χ^2 统计量近似服从 χ^2 分布。在自由度固定时,每个 χ^2 值与一个概率值(P 值)相对应,此概率值即在 H_0 假设成立的前提下,出现这样一个样本或更大差别样本的概率。如果 P 值小于或等于用户所设的显著性水平,则应拒绝 H_0 ,接受 H_1 。

(2) 关联程度的测量:卡方检验从定性的角度指出是否存在相关性,而各种关联指标从定量的角度指出相关的程度如何。不同的指标适用于不同类型的变量。

① RR 值是一个概率的比值,是指实验组人群反应阳性概率与对照组人群反应阳性概率的比值,用于反映实验因素与反应阳性的关联程度。

② OR 值是比值的比,是反应阳性人群中实验因素有无的比例与反应阴性人群中实验因素有无的比例之比。在下列两个条件均满足时,可用于估计 RR 值:所关注的事件发生的概率比较小(<0.1),这个条件保证比数比能对相对危险度有一个好的近似;所涉及的研究是病例对照研究。

③ 在 SPSS 中,在交叉表过程的“统计量”子对话框中选中“风险”复选框会自动给出 OR 与 RR 值。

(3) Kappa 检验与配对卡方检验:Kappa 一致性检验用于对两种方法结果的一致程度进行评价;配对卡方检验则用于分析两种分类方法的分类结果是否有差异。

(4) 分层卡方检验:分层卡方是把研究对象分解成不同层次,按各层对象来进行行变量与列变量的独立性研究。可在去除分层因素混杂的影响下更准确地对行列变量的独立性进行研究。在 SPSS 中,在交叉表过程的“统计量”子对话框中选中“Cochran's and Mantel-Haenszel 统计量”复选框会自动给出分层卡方检验结果。

思考与练习

1. 在周六晚节目单修订前后,分别进行了收视率的调查。在节目被修改前,收视率记录为 ABC29%,CBS28%,NBC25%,独立电台 18%。节目被修改后,300 个家庭所组成的样本产生下列电视收视数据:ABC95 个家庭,CBS70 个家庭,NBC89 个家庭,独立电台 46 个家庭。取显著性水平 $\alpha=0.05$,检验电视收视率是否已经发生了变化。用软件 SPSS 进行分析,并解释各表的含义。

2. 在周六晚节目单修订前后,分别进行了收视率的调查。在节目被修改前,300 个家庭收视纪录为:ABC76 个家庭,CBS89 个家庭,NBC83 个家庭,独立电台 52 个家庭。节目被修改后,300 个家庭所组成的样本产生下列电视收视数据:ABC95 个家庭,CBS70 个家庭,NBC89 个家庭,独立电台 46 个家庭。取显著性水平 $\alpha=0.05$,检验电视收视率是否已经发生了变化。用软件 SPSS 进行分析,并解释各表的含义(将本题与第 1 题进行比较)。

3. 3 名推销员 3 个月内的销售数量报告如题表所示。取显著性水平 $\alpha=0.05$,检验推销员与产品类型的独立性。


题 表

推销员	产品		
	A	B	C
Michael	14	12	4
David	21	16	8
Alice	15	5	10

4. 一家生产性公司从 3 家供应商处购买某零件,但该零件经常出现次品。在记录的 435 件零件质量数据中,100 件来自 A 公司,其中 90 件质量等级为良好,3 件有小缺陷,7 件有大缺陷;195 件来自 B 公司,其中 170 件质量等级为良好,18 件有小缺陷,7 件有大缺陷;150 件来自 B 公司,其中 135 件质量等级为良好,6 件有小缺陷,9 件有大缺陷。取显著性水平 $\alpha=0.05$,检验供应商与零件质量的独立性。分析结果能给采购部门提供什么信息?

第17章 相关分析


唯物论者认为任何事物之间都是有联系的,这种联系间存在着强弱、直接或间接的差别。相关分析就是通过定量的指标来描述这种联系。在第16章中实际上已经讲到了相关分析的指标体系,根据变量的类型,可以选用各种各样的相关程度描述指标。本章将针对连续变量的情形就此问题进行进一步的深入探讨。

 提到相关分析,读者可能下意识地会认为研究的是两个变量间的关系。但实际上,广义的相关分析研究的可以是一个变量和多个变量之间的关系,也可以是研究两个变量群,甚至于多个变量群之间的关系。由于后两种情况涉及比较复杂的模型,因此不在本书介绍范围之内,读者可以在高级教程中学习相应的方法。

17.1 相关分析简介

17.1.1 相关分析的指标体系

尽管在提及相关分析时,考察的往往都是两个连续变量之间的相关关系,但实际上对于任何类型的变量,都可以使用相应的指标进行相关关系的考察。第16章中已经给出了一些相关指标。为了能使读者建立一个完整的相关分析体系,下面将首先介绍针对不同的变量类型可供使用的相关分析指标种类。

 测量相关程度的相关系数有很多,各种参数的计算方法、特点各异。有的基于卡方值,有的则主要考虑预测效果。有些是对称性的,有些是非对称性的(在将变量的位置互换时,对称性参数将不变,非对称性参数则会改变)。大部分关联强度参数的取值范围在0~1之间,0代表完全不相关,1代表完全相关,但是,对于反映定序变量或连续变量间关联程度的参数,其取值范围则在-1到1之间,绝对值代表相关程度,而符号则代表是正相关还是负相关。

1. 连续变量的相关指标

显然,这种情况是最多见的,此时一般使用积差相关系数,又称为 Pearson 相关系数来表示其相关性的 大小,其数值介于-1~1 之间,当两个变量间的相关性达到最大,散点呈一条直线时取值为-1 或 1,正负号表明了相关的方向;如两变量完全无关,则取值为 0。

积差相关系数应用非常广泛,但严格地讲只适用于两变量呈线性相关时,详见后面介绍。此外,作为参数方法,积差相关分析有一定的适用条件,当数据不能满足这些条件时,分析者可以考虑使用 Spearman 等级相关系数来解决这一问题。

2. 有序变量的相关指标

对于有序的等级资料的相关性,又往往称其为一致性,所谓一致性高,就是指行变量等级高

的列变量等级也高,行变量等级低的列变量等级也低。如果行变量等级高而列变量等级低,则称其为不一致。

在详细介绍所用指标之前先要搞清楚两个指标的含义:当按两个变量的取值列出交叉表后, P 代表两倍的一致对子数, Q 代表两倍不一致的对子数,所谓一致对子数就是指行变量等级高的列变量等级也高,反之亦然。按此可以计算下面的5个指标,它们实际上均是由最前面的Gamma统计量衍生出来的。

(1) Gamma统计量:描述有序分类数据联系强度的度量。介于 $-1 \sim 1$ 之间,当观察值集中于对角线处时,其取值为 -1 或 1 ,表示两者取值绝对一致或绝对不一致;如两变量完全无关,则取值为 0 。它的计算公式非常简单,即 $\gamma = (P - Q)/(P + Q)$ 。

(2) Kendall's Tau-b:要掌握该系数必须先了解 τ_a 系数,该系数以同序对 P 与异序对 Q 之差为分子:

$$\tau_a = \frac{P - Q}{n(n-1)/2}$$

理论上 τ_a 的取值范围是 ± 1 ,但是当相等级太多时,会使其的极大值与极小值不能达到 ± 1 ,为此在分母上按照相等级的对子数进行了校正,以保证取值范围能达到 ± 1 ,此即 τ_b 系数,因校正后公式比较复杂,这里不再给出。

(3) Kendall's Tau-c:在Kendall's Tau-b的基础上又进一步考虑了整张列联表的大小,并对其进行了校正。

(4) Somers's d(C|R): d 系数为Somer所创,因此称为Somer's d 。它是 τ_b 的不对称调整,只校正了自变量相等的对子。分别给出了 d_{yx} 和 d_{xy} 两个系数:

$$d_{yx} = \frac{P - Q}{P + Q + P_y}, d_{xy} = \frac{P - Q}{P + Q + P_x}$$

d_{yx} 表示 x 为自变量、 y 为因变量时的情况,其中 P_y 表示仅在 y 方向的分对。

3. 名义变量的相关指标

对于名义变量,实际上第16章中所学习的卡方检验中的 χ^2 值就可以用于测量两个变量的相关性,而这里介绍的更专业的指标实际上多数也就是从 χ^2 值衍生而来的。可以用以下几个指标来评价相关性。

(1) 列联系数(Contingency Coefficient):基于 χ^2 值得出公式为 $\sqrt{\chi^2/(\chi^2 + n)}$,其中 n 为总样本量。其值介于 $0 \sim 1$ 之间,越大表明两变量间的相关性越强。

(2) Phi和Cramer's V:这两者也是基于 χ^2 值的,Phi是基于卡方值和总观察频数计算而来的, $\phi = \sqrt{\chi^2/n}$ 。在4格表 χ^2 检验中介于 $0 \sim 1$ 之间,在其他列联表中其取值在理论上没有上限,值越大,关联程度越强。Cramer's V是Phi的一个调整,较用Phi进行关联程度的测量保守,经调整后使得取值在任何列联表中均不超过1。指标的绝对值越大,则相关性越强:

$$V = \sqrt{\phi^2 / \min[(r-1), (c-1)]}$$

分母中的 $\min[(r-1), (c-1)]$ 表示选择 $(r-1), (c-1)$ 中的较小者作为除数。经过这样的改进, V 的取值范围就为 $[0, 1]$ 了,因此 V 系数就克服了 ϕ 系数不能与其他相关系数间进行比较的缺点。

(3) λ 系数(Lambda):用于反映自变量对因变量的预测效果,即知道自变量取值时对因变量的预测有多少改进,或者说知道自变量的取值时期望预测误差个数减少的比例,Lambda 将误差定义为列(行)变量预测时的错误,其预测值是基于个体所在行(列)的众数。值为 1 时表明知道了自变量就可以完全确定因变量取值,为 0 时表明自变量对因变量完全无预测作用:

$$\lambda = \frac{\sum f_{im} - F_{ym}}{n - F_{ym}}$$

其中, f_{im} 为每一类 x 中 y 分布的众数次数, F_{ym} 为 y 次数分布的众数次数。 λ 相关来自消减误差比率,对计算结果自然也从消减误差比例的角度解释。即“根据 x 去估计 y 可以减少百分之 λ 的误差”。 λ 必定处于 0~1 之间。

另外需要注意的是,如果将表中两个变量的位置对换,计算出的 λ 值将会不同,也就是说,行变量为自变量、列变量为自变量时的结果是不一样的。当无法确定自变量与因变量时,可以取两个 λ 平均值作为 λ 相关量,SPSS 会同时给出这 3 种结果。

(4) 不确定系数(Uncertainty Coefficient):其值介于 0~1 之间,和 Lambda 类似,也用于反映当知道自变量后,因变量的不确定性下降了多少(比例),只是在误差的定义上稍有差异。以熵为不确定性大小的度量指标,共会输出行变量为自变量、列变量为自变量、对称不确定系数 3 个结果,后者为前两者的对称平均指标。



相信很多读者看到这里已经是一头雾水:这么多指标,究竟用哪个才合适?如同方差分析中的两两比较方法一样,统计方法如此多恰好说明了统计学在解决分类指标相关性方面所面临的困境。就现状而言,分析人员能做的事只能是根据具体的问题特征从上面这些各具特点的统计指标中挑选最为合适的一个加以使用。同时也要注意:不同的指标是不能简单地进行数值大小的对比的!

4. 其他特殊指标

除了以上较为系统的指标外,当希望测量一个名义变量和连续变量间的相关程度时,还可以使用一个叫做 Eta 的指标,它所对应的问题以前是用方差分析来解决的。实际上,Eta 的平方表示由组间差异所解释的因变量的方差的比例,即 SS 组间/SS 总。

第 16 章中还介绍过 Kappa、OR、RR 等统计指标,它们实际上也都是相关程度的测量指标,因第 16 章中已进行专门讲述,这里不再重复。

17.1.2 SPSS 中的相应功能

SPSS 的相关分析功能分散在几个过程中,但大致可归为以下两类。

1. “交叉表:统计量”子对话框

该子对话框在第 16 章中已经有所介绍,事实上,按照无序、有序、连续变量的分类,在对话框中提供了非常整齐的相关分析指标体系,如图 17.1 所示,在其中找到上面介绍的几乎全部指标,具体解释如下。

(1) “相关性”复选框:适用于两个连续性变量的分析,计算行、列变量的 Pearson 相关系数和 Spearman 等级相关系数。

(2) “按区间标定”框组:包含了一个变量为数值变量,而另一个变量为分类变量时度量两者



图 17.1 “交叉表:统计量”对话框中的相关指标体系

关联度的指标, Eta 的平方表示由组间差异所解释的因变量的方差的比例,即 $SS_{\text{组间}}/SS_{\text{总}}$ 。系统一共会给出两个 Eta 值,分别对应了行变量为因变量(数值变量)和列变量为因变量的情况。

(3) “有序”复选框组:包含了一组用于反映分类变量一致性的指标,这些指标只能在两个变量均属于有序分类时使用。它们均是由 Gamma 统计量衍生出来的。

(4) “名义”复选框组:包含了一组用于反映分类变量相关性的指标,这些指标在变量属于有序和无序分类时均可使用,但两变量均为有序分类变量时效率没有“有序”复选框组中的统计量高。

(5) “Kappa”:计算 Kappa 值,即内部一致性系数。

(6) “风险”:计算 OR 值(比值比)和 RR 值(相对危险度)。

2. “相关”子菜单

由于针对连续性变量的相关分析更为常用,因此 SPSS 还专门提供了“相关”子菜单中的 3 个过程用于满足相应的分析需求。

(1) 双变量(Bivariate)过程:此过程用于进行两个/多个变量间的参数/非参数相关分析,如果是多个变量,则给出两两相关的分析结果。这是相关分析中最为常用的一个过程,实际上人们对它的使用可能占到相关分析的 95% 以上。

(2) 偏相关(Partial)过程:如果需要进行相关分析的两个变量其取值均受到其他变量的影响,就可以利用偏相关分析对其他变量进行控制,输出控制其他变量影响后的相关系数,偏相关过程就是专门进行偏相关分析的。

(3) 距离(Distances)过程:调用此过程可对同一变量内部各观察单位间的数值或各个不同变量进行相似性或不相似性(距离)分析,前者可用于检测观测值的接近程度,后者则常用于考察各变量的内在联系和结构。该过程一般不单独使用,而是用于因子分析、聚类分析和多维尺度分析的预分析,以帮助用户了解复杂数据集的内在结构,为进一步分析做准备。

至于更复杂的相关分析问题,如两组变量间的相关分析等,在 SPSS 中还有线性回归模型、典型相关分析等更复杂的功能可供调用,但这已经超出了本书的讲授范围,对此感兴趣的读者可参

见《SPSS 统计分析高级教程》。

17.2 简单相关分析

17.2.1 方法原理

1. 一些基本概念

连续变量相关分析的一个显著特点是变量不分主次,被置于同等的地位。它的一些常用术语如下。

(1) 直线相关:这是最简单的一种情况,两变量呈线性共同增大,或者呈线性一增一减的情况。这里讨论的范围基本上限于直线相关。

(2) 曲线相关:两变量存在相关趋势,但并非线性,而是呈各种可能的曲线趋势。此时如果直接进行直线相关分析,有可能得出无相关性的结论。

(3) 正相关与负相关:如果 A 变量增加时 B 变量也增加,则称为正相关,如 A 变量增加时 B 变量减小,则称为负相关。

(4) 完全相关:两变量的相关程度达到了亲密无间的程度,当得知 A 变量取值时,就可以准确推算出 B 变量的取值。又分为完全正相关和完全负相关两种。



当数据为有序变量或者名义变量时,一般不再考虑直线、曲线相关的问题,但正、负相关和完全相关这些概念仍然适用。

2. 系数计算

当两个连续变量在散点图上的散点呈现直线趋势时,就可以认为两者存在直线相关趋势,也称为简单相关趋势。Pearson 相关系数,也称为积差相关系数就是人们定量地描述线性相关程度高低的一个常用指标。



一般认为,相关和回归的概念是在 1877—1888 年间由 Francis Galton 提出的,并在 1889 年出版的《自然遗传》一书中总结了自己的工作。但真正使这方面的理论系统化的是 Karl Pearson,正是后者的出色工作使得相关和回归理论大放光彩,并得到了广泛的应用。而为了纪念他的贡献,简单相关分析中所用的相关系数就被称为 Pearson 相关系数。

在介绍相关系数的计算方法前首先介绍方差。对于相关分析中的两个变量,其方差 SS_x 和 SS_y 分别反映了各自的变异程度,在相关与回归分析中方差又被记为 l_{xx} 和 l_{yy} 。以 x 的样本方差为例,其计算公式为

$$l_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$$

在相关分析中,协方差是一个非常重要的概念,用符号 l_{xy} 来表示,其计算公式和方差非常类似:

$$l_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)$$

可见,样本协方差是离均差乘积在样本中的平均,可以认为其近似反映了变量 x 与 y 之间的

联系强弱和方向。若离均差乘积平均后接近0,则表明变量 x 和 y 的部分取值同向,部分取值反方向,因而离均差乘积有正有负相互抵消,其和就接近于零。而如果 x, y 为同向变化,则离均差乘积大多为正,其和也为正,反之则离均差积和为负。

显然,协方差可以反映两变量相关性的 大小,但由于协方差的大小与 x, y 的量纲有关,不同问题中的协方差不可直接比较。因此考虑使用 x, y 的方差对其进行标准化,公式如下:

$$R^2 = l_{xy}^2 / (l_{xx} l_{yy})$$

由于是同时使用 x 和 y 的方差进行标准化的,所以分子为协方差的平方,该指标称为决定系数,取值范围为 $0 \sim 1$,可以很好地反映两变量间相关性的强弱:决定系数越大,表明两变量的相关程度越高;当两变量完全相关时,决定系数为1;当两变量不相关时,决定系数为0。

但是,决定系数仍然存在问题。由于协方差平方后均为正,从而决定系数不能反映相关 的方向。因此为了便于应用,在标准化协方差时不是将分子平方,而是将分母开根号用于 标准化,公式如下:

$$r = l_{xy} / \sqrt{l_{xx} l_{yy}}$$

上述指标就是相关系数,显然,它也是标准化之后的协方差,可以很好地反映相关程 度的强弱,而且数值介于 -1 和 $+1$ 之间,其正负就反映了相关 的方向,便于应用。

归纳起来相关系数具有如下特点:

- (1) 相关系数 r 是一个无单位的量值,且 $-1 < r < 1$ 。
- (2) $r > 0$ 为正相关, $r < 0$ 为负相关。
- (3) $|r|$ 越接近于1,说明相关性越好, $|r|$ 越接近于0,说明相关性越差。

3. 相关系数的检验方法

计算出样本的相关系数后必须对其进行检验,以确定其不是从一个数值为0的相关系 数的总体中抽出的(避免计算出的数值是由于抽样误差所导致的)。它的假设检验如下:

$H_0: \rho = 0$, 两变量间无直线相关关系。

$H_1: \rho \neq 0$, 两变量间有直线相关关系。

检验的方法主要是 t 检验,公式为: $t = \frac{r - 0}{s_r}, \nu = n - 2$ 。求出统计量后即可根据自由度得到 P 值,通过 P 值与临界值的比较就可以进行判断了。但是在 SPSS 的结果中只会给出相关系 数值和最终的 P 值,并不会给出统计量 t 的具体计算结果。

4. 积差相关系数的适用条件

任何一种统计方法都是有适用条件的,对统计方法运用得 好坏和正确不在于是否能写出公式或能否计算出结果,而在于针对数据特征懂得运用正确的统计方法。在相关分析中首先要考虑的问题就是两个变量是否可能存在相关关系,如果得到了肯定的结论,那么才有必要进行下一步的定量分析。

另外在进行相关分析前必须注意以下几个问题。

(1) 积差相关系数适用于线性相关的情形,对于曲线相关等更为复杂的情形,积差相关系 数的大小并不能代表其相关性的强弱。

(2) 样本中存在的极端值对积差相关系数的计算影响极大,因此要慎重考虑和处理,必要时 可以对其进行剔除,或者进行变量变换,以避免由一两个数值导致出现错误的结论。需要注意的

是,有的时候在分别观察每个变量时极端值并不明显,但是联合观察两个变量时就会凸显出来。

(3) 积差相关系数要求相应的变量呈双变量正态分布,注意双变量正态分布并非简单的要求 x 变量和 y 变量各自服从正态分布,而是要求服从一个联合的双变量正态分布,如图 17.2 所示。

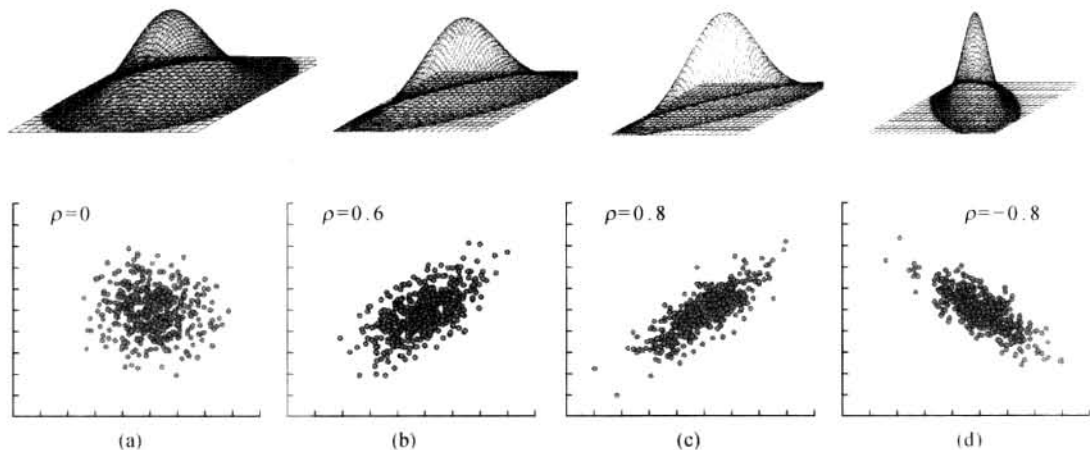


图 17.2 双变量正态分布及其样本散点图

在以上几条要求中,前两者要求最严,第三条比较宽松,违反时系数的计算结果也是比较稳健的。一般而言,分析者可以使用图形工具来对以上条件加以考察,散点图和直方图是最常用的工具。特别是散点图,它可以同时考察变量间是否存在线性相关、有无极端值、变量的分布是否接近正态,因此在相关分析中考察适用条件时更为常用。从散点图中可以发现如下重要信息。

(1) 两变量间是否存在相关趋势。

(2) 这种相关趋势呈现为线性趋势还是曲线趋势,是否可以直接使用线性相关的积差相关系数加以刻画。

(3) 在散点图上是否发现明显的异常值,或者说强影响点。

只有上述问题都得到解答,后续相关分析的结果才是可信的。

17.2.2 案例:考察信心指数值和年龄的相关性

例 17.1 利用相关分析考察总信心指数值和年龄的相关性。

对于本例,首先应该意识到的是题中的变量均为连续性变量,因此在相关指标体系中应当考虑使用描述两个连续性变量相关性的指标。在第 11 章中已经绘制出了总信心指数和年龄的散点图,并且从趋势线也已经确认两者之间应当是存在负相关关系的,这里就直接进行相关系数的计算。



此处因为已经在前面绘制过散点图,所以直接进入了后续的相关分析操作,否则必须首先进行散点图的考察。

1. 界面说明

双变量相关过程的对话框如图 17.3 所示,内容非常简单,这里简要介绍如下。

(1) “变量”列表框:用于选入需要进行相关分析的变量,至少需要选入两个。如果选入了多个,则分析结果会以相关矩阵的形式给出两两直线相关分析的结果。

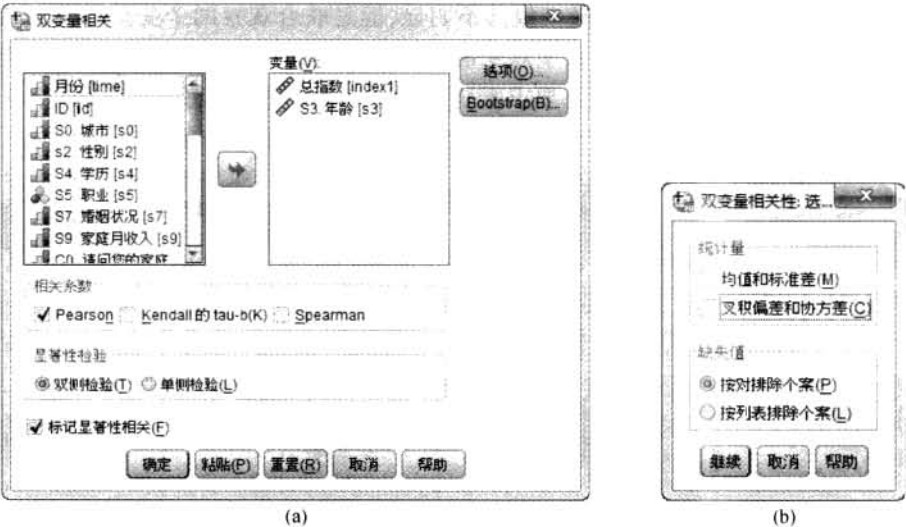


图 17.3 “双变量相关”对话框

(2) “相关系数”复选框组:用于选择需要计算的相关分析指标。Pearson 是默认选项,进行积差相关分析;Kendall 的 tau-b 要求计算 Kendall’s 等级相关系数;Spearman 则要求计算 Spearman 相关系数,即最常用的非参数相关分析(秩相关)。

(3) “显著性检验”单选框组:用于确定是进行相关系数的单侧(One-tailed)检验还是双侧(Two-tailed)检验,一般选中“双侧检验”单选框。

(4) “标记显著性相关”复选框:要求在结果中用星号标记有统计学意义的相关系数,一般选中。此时 $P < 0.05$ 的系数值旁会标记一个星号, $P < 0.01$ 则标记两个星号。

(5) “选项”按钮:单击后打开的子对话框用于选择需要计算的描述统计量和进行的统计分析,包括常用的均值和标准差,以及相关分析特有的交叉积和与协方差,下方的“缺失值”单选框组用于定义分析中对缺失值的处理方法,前面出现过许多次了。

2. 操作说明与结果解释

本例的操作步骤非常简单,具体如下。

- (1) 选择“分析”→“相关”→“双变量”菜单项,打开“双变量相关”对话框。
- (2) 将“总指数”、“年龄”选入“变量”列表框中。
- (3) 单击“确定”按钮。

最终相关分析的结果输出非常简单,如图 17.4 所示。

		总指数	S3. 年龄
总指数	Pearson 相关性	1	-.219**
	显著性(双侧)		.000
	N	1147	1147
S3. 年龄	Pearson 相关性	-.219**	1
	显著性(双侧)	.000	
	N	1147	1147

** . 在 .01 水平(双侧)上显著相关。

图 17.4 相关性

图 17.4 给出的就是积差相关系数的结果,也就是要求的 Pearson 相关系数。结果是以对角阵的形式给出的,由于这里只分析了两个变量,因此给出的是 2×2 的方阵。每个单元格共分为 3 行,分别是相关系数、 P 值和样本数。总信心指数和年龄的相关系数为 -0.219 ,对相关系数的检验的双侧 P 值小于 0.001 ,所以可以认为两变量间的负相关是有统计学意义的,随着年龄的增加,总指数值呈现减小的趋势。



如果在选项子对话框中选中了“叉积偏差和协方差”复选框,则会在输出表格中包括离均差平方和以及协方差值,大家可以按照前述公式计算出相关系数,结果就是这里得到的 -0.219 ,这可以帮助大家理解前面的计算过程。



在第 11 章的图形分析中曾经提到过在 30 岁左右信心指数出现过一个稳定平台,并未持续下降。但是在本章中如果计算线性相关系数,则模型应当假设信心指数是随年龄持续下跌的。如果希望考察不同年龄段信心指数的变化规律,则应当考虑更为复杂的分段分析或者非线性回归模型,而不是简单的线性相关系数。

17.2.3 秩相关系数

熟悉计算积差相关系数的整个过程后,有人可能会问:计算积差相关系数的要求那么高,要求 x 、 y 都要服从正态分布,如果数据达不到那么高的要求但是又要衡量两个变量之间的相关关系该如何解决呢?其实,SPSS 为用户提供了其他的方法,最常用的一个就是 Spearman 等级相关系数。

Spearman 相关系数又称为秩相关系数,是利用两变量的秩次大小进行线性相关分析的,对原始变量的分布不作要求,属于非参数统计方法。因此它的适用范围较 Pearson 相关系数要广得多。即使原始数据是等级资料,也可以计算 Spearman 相关系数。当然,对于服从 Pearson 相关系数的数据可以计算 Spearman 相关系数,但统计效能比 Pearson 相关系数要低一些(不容易检测出两者事实上存在的相关关系)。

Spearman 相关系数的计算公式可以完全套用 Pearson 相关系数的计算公式,将公式中的 x 和 y 用 x 和 y 相对应的秩次代替即可。当样本含量 n 小于等于 50 时,Spearman 相关系数的检验可以借助查界值表进行,大于 50 后检验公式与积差相关系数相同。

对于例 17.1,如果计算秩相关系数,则结果如图 17.5 所示。

		总指数	S3. 年龄
Spearman 的 rho	总指数	1.000	-.213**
	相关系数	.	.000
	Sig. (双侧)		
	N	1147	1147
S3. 年龄	总指数	-.213**	1.000
	相关系数	.000	.
	Sig. (双侧)		
	N	1147	1147

** . 在置信度(双侧)为 0.01 时,相关性是显著的。

图 17.5 相关系数

从结果中可以看到 Spearman 相关系数为 -0.213 , P 值小于 0.001 ,在 $\alpha = 0.05$ 的水平上是

拒绝无效假设的,结论和前面相同。



由于非参数方法对信息的利用效率要低于参数方法,因此对于同一个资料,在双变量正态分布成立的时候,在绝大多数情况下 Spearman 等级相关系数的绝对数值都是小于 Pearson 相关系数的。

17.2.4 Kendall 等级相关系数

在双变量相关的对话框中还提供了 Kendall's tau-b 等级相关系数的选项,本章开始已经介绍了相关分析的指标体系,显然这个 Kendall's tau-b 等级相关系数是用于反映分类变量相关性的指标,适用于两个变量均为有序分类的情况。对于上例,如果计算等级相关系数,则结果如图 17.6 所示。

		总指数		S3. 年龄
Kendall 的 tau_b	总指数	相关系数	1.000	-.152**
		Sig. (双侧)	.	.000
		N	1147	1147
S3. 年龄		相关系数	-.152**	1.000
		Sig. (双侧)	.000	.
		N	1147	1147

** . 在置信度(双侧)为 0.01 时,相关性是显著的。

图 17.6 相关系数

可见分析结论和前面相同。本例显然是定量数据,且并未违反积差相关系数的适用条件,因此使用积差相关系数来描述相关情况是合适的,这里仅演示计算。从表 17.3 中还可以发现,对于相同的数据,秩相关系数和等级相关系数的绝对值都小于积差相关系数,显然这是由于在秩变换或者数据按有序分类处理时会损失信息所导致的。

17.3 偏相关分析

17.3.1 方法原理

1. 偏相关所需要解决的问题

辩证法里有这样一对概念:现象和本质。之所以要通过现象看本质就是因为某些现象可能会干扰人们对于本质的认识。在相关分析中也存在这样的问题。就像世界上没有两片完全一模一样的树叶一样,也不存在完全独立于其他事物的个体和现象。在研究两个事物或现象之间的关系时,只有充分考虑到其他事物和现象对两者之间的影响,才可能发现两者真正的联系。

但是,前面介绍的相关分析是分析两个变量间的关系,在计算积差相关系数、Spearman 相关系数和 Kendall 相关系数的时候都没有考虑第三方的影响,这就有可能导致对事物的解释出现偏差。例如上面总信心指数和年龄的相关分析,在前面章节的分析中已经知道家庭收入 Qs9 和总信心指数也存在着负相关趋势,显然年龄可能和家庭收入存在一定的关联,那么前述 -0.219 的相关系数中究竟有多少反映的是年龄 - 家庭收入 - 信心值这样一种间接链条的影响?或者说

在控制了家庭收入的作用之后,年龄和信心值之间还有相关性吗?要解决这些问题就需要进行偏相关分析。


2. 偏相关分析的计算公式

偏相关分析是指在相关的基础上考虑两个因素以外的各种影响因素,或者说在扣除了其他因素的作用大小以后,重新来考察这两个因素间的关联程度。这种方法的目的就在于消除其他变量关联性的传递效应。

计算偏相关系数时可以首先分别计算3个因素之间的相关系数,然后通过这3个简单相关系数来计算偏相关系数,计算公式如下:

$$r_{12(3)} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{1 - r_{13}^2} \times \sqrt{1 - r_{23}^2}}$$

该公式就是在控制了第三个因素的影响后所计算的第一、第二个因素之间的偏相关系数。当考虑一个以上的控制因素时公式类推。

 事实上,如果从回归的角度来解释,偏相关系数就是首先以希望分析的变量为因变量,被控制的变量为自变量分别拟和两个回归方程,然后将所得的两组残差进行简单相关分析,有兴趣的读者可以自行尝试一下。

17.3.2 案例:控制家庭收入的影响之后考察年龄的作用

例 17.2 在控制家庭收入 Qs9 对总信心指数影响的前提下,考察总信心指数值和年龄的相关性。

可以首先对上述3个变量的相关性进行两两考察,结果如图 17.7 所示。

Pearson 相关性

	总指数	S3. 年龄	Qs9
总指数	1	-.219 **	.084 **
S3. 年龄	-.219 **	1	-.138 **
Qs9	.084 **	-.138 **	1

** . 在 .01 水平(双侧)上显著相关。

图 17.7 相关性

从图 17.7 中可见家庭收入和年龄呈负相关关系,同时和总指数呈正相关关系,且两者均有统计学意义。这样一来,年龄就完全可能通过上述相关关系和总指数建立数量上的关联,需要考虑利用偏相关分析来得到更加纯粹的分析结果。

1. 界面说明

偏相关分析所使用的对话框和相关分析极为相似,如图 17.8 所示,简介如下。

(1) 主对话框:大部分内容和“双变量相关分析”主对话框类似,只是新出现了一个“控制”列表框,用于选择在进行偏相关分析时需要控制的变量。如果不选入,则进行的是普通的相关分析(求出的是积差相关系数)。

(2) “选项”按钮:单击后打开的对话框除了给出均数、标准差等的描述外,还可以选中“零阶相关系数”复选框,要求给出包括协变量在内所有变量两两相关的系数阵。

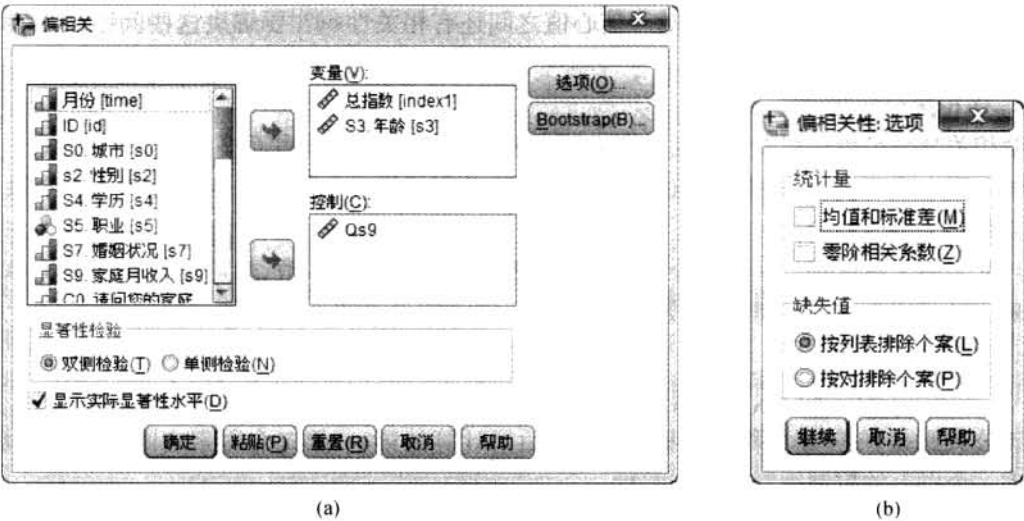


图 17.8 “偏相关”对话框

2. 操作说明与结果解释

本例的操作步骤非常简单,具体如下。

- (1) 选择“分析”→“相关”→“偏相关”菜单项,打开“偏相关”对话框。
- (2) 将“总指数”、“年龄”选入“变量”列表框中。
- (3) 将家庭收入“Qs9”选入“控制”列表框中。
- (4) 单击“确定”按钮。

相应的分析结果如图 17.9 所示。

控制变量		总指数		S3. 年龄
Qs9	总指数	相关性	1.000	-.216
		显著性(双侧)	.	.000
		df	0	989
	S3. 年龄	相关性	-.216	1.000
		显著性(双侧)	.000	.
		df	989	0

图 17.9 相关性

图 17.9 就是在控制了家庭收入 Qs9 的作用之后计算出的年龄和总指数间的偏相关系数矩阵,可见两者的偏相关系数为 -0.216,虽然绝对数值有所减小,但仍然具有统计学意义,因此在控制了家庭收入的作用之后,仍然可以确认年龄和总指数之间存在负相关关系。

17.4 Distance 过程

简单相关和偏相关有一个共同点,那就是对所分析的数据背景应当有一定程度的了解。但有时会遇到一种情况,在分析前对数据所代表的专业背景知识了解尚不充分,本身就属于探索性

的研究,这时往往需要先对各个指标或者案例的差异性/相似程度进行考察,对数据有一个初步的了解,然后再根据结果考虑如何进行深入分析。

距离过程可以用于计算记录(或变量)间的距离(或相似程度),根据变量的不同类型,可以有许多距离/相似程度测量指标供用户选择。但由于本模块只是一个预分析过程,因此距离分析并不会给出常用的 P 值,而只给出各变量/记录间的距离大小,以供用户自行判断相似性。

17.4.1 距离测量与相似性测量的指标体系

如前所述,距离过程可以计算距离测量指标或者相似性测量指标,这可以在主对话框中加以切换,会分别打开不同的子对话框,有多种指标可供设定,这里分述如下。

1. 不相似性(距离)测量指标

以案例间的距离测量为例,其基本原理就是将变量看成是构成空间的维度,然后案例就构成了这样一个多维空间中的散点,求出这些散点的空间距离,即为相应的距离测量值。根据不同的数据类型,距离测量指标也有所不同。

(1) 连续性变量:默认为欧式距离(欧几里得距离),具体有以下几种。

① Euclidean distance:欧几里得距离,以两变量差值平方和的平方根为距离,就是人们平常所理解的空间距离。

② Squared Euclidean distance:欧氏平方距离,以两变量差值平方和为距离,这种测量方法更重视较大的数值和距离。

③ Chebychev:切比雪夫距离,以两变量绝对差值的最大值为距离。

④ Block:以两变量绝对差值之和为距离。

⑤ Minkowski:闵可夫斯基距离,以两变量绝对差值 p 次幂之和的 p 次根为距离,用户可以在 Power 文本框中更改分量值之差的次方 P 的大小。当 $p=2$ 时即为欧氏距离。

⑥ Customized:自定义距离公式,用户需要在 Power 文本框中定义分量值之差的次方,在 Root 文本框中定义分量值之差次方的开方。以两变量绝对差值 p 次幂之和的 r 次根为距离。

(2) 频数表资料:默认为 χ^2 值测距,具体有以下几种。

① Chi-square measure: χ^2 值测距。

② Phi-square measure: ψ^2 值测距,即将 χ^2 测距值除以合计频数的平方根。

(3) 二分类变量:默认为欧氏距离,具体有以下几种。

① Euclidean distance:计算公式为 $\text{SQRT}(b+c)$,其中 b, c 分别为 4 格表中对角线上的元素,最小值为 0,最大无限。

② Squared Euclidean distance:即 $|b+c|$,最小为 0,最大无限。

③ Size difference:最小距离为 0,最大无限。

④ Pattern difference:从 0 至 1 的无级测距。

⑤ Variance:以方差为测距,最小为 0,最大无限。

⑥ Lance and Williams:Bray-Curtis 非等距系数,介于 0 至 1 之间。

2. 相似性测量指标

相似性测量指标实际上就是前面所讲的那些相关分析指标体系,只是更为详细一些,主要分为以下两类。

(1) 计量资料:可以采用 Pearson Correlation 即常用的积差相关系数,也可采用 Cosine,即以变量矢量的余弦值为距离,大小介于 -1 至 $+1$ 之间,数值越大表明相似性越高。

(2) 二分类变量:给出了一大堆测量指标,其实非常少用,这里完全没有罗列出来凑字数的必要。只需要使用默认的 Russell and Rao(以二分点乘积为配对系数)即可。

上面只是简单地解释了一下各种指标的含义,当使用不同的距离测量指标时,得到的结果可能完全不同,对于在不同分析问题中对各种距离/相似性测量指标的选择问题,感兴趣的读者可以参考《SPSS 统计分析高级教程》中聚类分析一章,这里不再详述。

17.4.2 案例:基因间距离的计算

例 17.3 某实验室制作了一张基因芯片,上面一共检测了上万个基因,现在从数据库中提取出 7 个基因的数据,由于对这 7 个基因的生物学功能现在一无所知,因此首先想对其进行距离测量,确定哪几个基因的“距离”比较接近,然后通过临床或实验室进一步验证。数据见 distance.sav。

1. 界面说明

比起前面两个相关分析过程,“距离”对话框界面要稍微复杂一些,如图 17.10 所示,但所针对的分析任务还是很清晰的。



图 17.10 “距离”对话框

(1) “变量”列表框:用于选入需要进行距离相关分析的变量,至少需要选入两个。

(2) “标注个案”列表框:选择一个变量,其取值会在输出结果中给相应记录加上标签,以方便阅读。该列表框只在分析记录间的距离时可用。

(3) “计算距离”单选框组:其中有两个选项,用于确定随后的分析是进行个案间的还是变量间的距离/相关分析。

(4) “度量标准”单选框组:用于选择分析时采用的距离类型,对于不相似性测距,数值越大表示距离越远;对于相似性测距,数值越大表示距离越近。

(5) “度量”按钮:单击后打开的子对话框分为不相似性和相似性两种,如图 17.11 所示,其中所提供的指标在 17.3 节已经介绍过了,这里不再重复。



图 17.11 距离过程中非相似性与相似性度量的子对话框

2. 操作说明与结果解释

对于本例这里不再详细讨论,只是给出分析过程的演示,相应的操作如下。

- (1) 选择“分析”→“相关”→“距离”菜单项,打开“距离”对话框。
- (2) 将 FPGS ~ IRF2 选入“变量”列表框中。
- (3) 在“计算距离”框组中选中“变量间”单选框。
- (4) 单击“确定”按钮。

图 17.12 即为两两变量间的距离计算结果,通过这张表可以看出代号为 CDK2AP1、TCEB1 和 IRF2 的 3 个基因比较接近,可以粗略地划分为一类,而 FPGS、ELF3 和 GFRA2 可以划为另外一类,而 NFE2 不确定,可能会作为单独的一类,这样就可以进一步考虑以后的研究了。

Euclidean 距离							
	FPGS	ELF3	CDK2AP1	GFRA2	TCEB1	NFE2	IRF2
FPGS	.000	.779	2.416	.749	1.006	.781	1.424
ELF3	.779	.000	1.749	.804	1.106	.933	1.578
CDK2AP1	2.416	1.749	.000	2.106	2.480	2.349	2.784
GFRA2	.749	.804	2.106	.000	1.312	.521	1.085
TCEB1	1.006	1.106	2.480	1.312	.000	1.400	1.864
NFE2	.781	.933	2.349	.521	1.400	.000	.962
IRF2	1.424	1.578	2.784	1.085	1.864	.962	.000

这是一个不相似性矩阵。

图 17.12 近似矩阵

17.5 本章小结

- (1) 虽然一般所说的相关分析均是指两个连续变量的相关性,但实际上任意测量尺度的两个变量都可以用相应的指标来描述其相关程度大小。
- (2) 相关系数 r 表示两变量间的直线相关程度, r 值的范围为从 -1 到 1 。 r 为正表示 x 与 y 之间为正相关, r 为负表示负相关。 r 近于零表示两变量间的关系不密切, r 的绝对值接近于 1 表示两变量间的关系较密切。但 r 有抽样误差,故算得相关系数之后,必须检验相应的总体相关系数 ρ 是否为 0 。
- (3) 研究中一般只涉及直线相关关系,但从理论上讲,可以进行变量间的曲线相关分析;如果希望扣除其他变量的影响,可以进行偏相关分析;如果变量不满足线性相关分析的适用条件,则可以进行 Spearman 秩相关分析。

思考与练习

某医师研究婴儿出生体重和双顶径的数量关系,收集了婴儿出生体重(X, g)和双顶径(Y, mm)数据,如题表所示,分析两者的数量关系。

题 表

X	273	299	226	315	294	260	383	273	234	329	302	357
Y	94	88	91	99	93	87	94	93	81	94	94	91

第 18 章 线性回归模型入门

18.1 线性回归模型简介

18.1.1 相关分析与回归分析的联系与区别

第 17 章介绍了相关分析,本章将要学习的回归分析也可以用来考察两个连续变量间的联系,但反映的是不同的侧面。以图 18.1 为例,这两幅散点图坐标尺度相同,都反映了 X 和 Y 两个变量的关联趋势,但它们有两个明显的差别。

(1) 图 18.1(a) 的散点明显要比图 18.1(b) 中稀疏一些,这表明左图中两变量在数量上的联系是弱于右图的。如果要用统计指标对这种差别进行表述,则应当进行相关分析,相关系数就可以反映散点的疏密,图 18.1(a) 的相关系数没有图 18.1(b) 的大。

(2) 如果在图中观察当 X 变动时 Y 的数量变化,则会发现在图 18.1(a) 中当 X 每增加一个单位时, Y 平均增加的较多,而在图 18.1(b) 中 X 增加一个单位时 Y 平均增加的较少,即图 18.1(a) 中 X 的变动对 Y 数值的影响要比图 18.1(b) 中大,这种差别在统计中可以使用回归分析来加以表述。

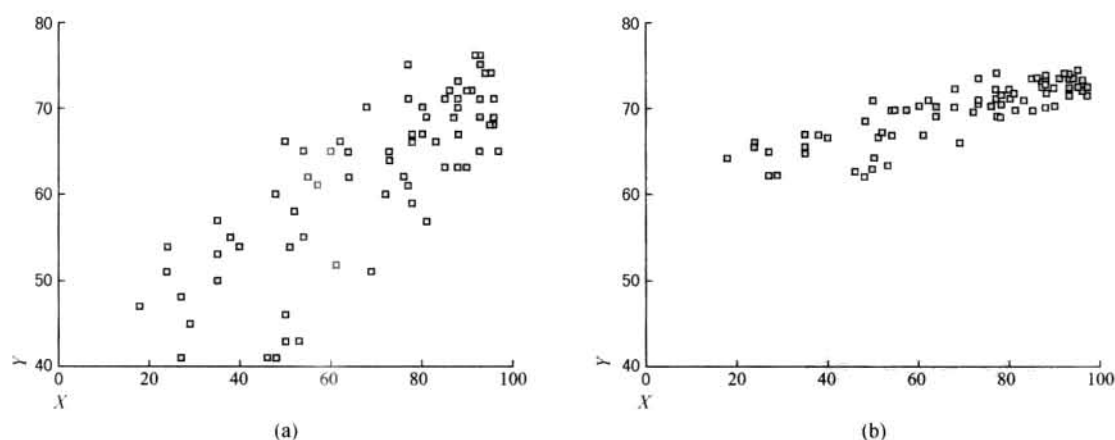


图 18.1 两变量间关系的示意图

从上面的比较可知,通过回归方程解释两变量之间的关系显得更为精确,例如可以计算出年龄每增加 1 岁时信心指数值平均下降的单位数量,这是相关分析无法做到的。除了描述两个变量间的关系以外,通过回归方程还可以进行预测和控制,预测就是在回归方程中控制了变量 x 的取值范围就可以相应的得到变量 y 的上下限,而控制则正好相反,也就是通过限制结果变量 y 的取值范围来得到 x 的上下限。这两点在实际应用中显得尤为重要。

18.1.2 简单回归分析的原理和要求

1. 模型的基本结构

如果将两个事物的取值分别定义为变量 x 和 y , 则可以用回归方程 $\hat{y} = a + bx$ 来描述两者的关系, 这里需要注意两点, 第一, 变量 x 称为自变量, 而 y 为因变量, 一般来讲应该有理由认为是由于 x 的变化而导致 y 发生变化的。第二, \hat{y} 不是一个确定的数值, 而是对应于某个确定 x 的群体的 y 值平均值的估计。该方程的含义可以从其等式右边的组成来理解, 即每个预测值都可以分解成如下两部分。

(1) 常量 (Constant): x 等于零时回归直线在 Y 轴上的截距 (Intercept), 即 x 取值为零时 y 的平均估计量。

(2) 回归部分: 它刻画因变量 y 的取值中, 由因变量 y 与自变量 x 的线性关系所决定的部分, 即可以由 x 直接估计的部分。 β 称为回归系数 (Coefficient of Regression), 又称为回归线的斜率 (Slope)。

估计值 \hat{y} 和每一个实测值之间的差称为残差。它刻画了因变量 y 除了自变量 x 以外的其他所有未进入该模型或未知但可能与 y 有关的随机和非随机因素共同引起的变异, 即不能由 x 直接估计的部分。通常假定 ε_i 服从正态分布 $N(0, \sigma^2)$ 。

既然模型中有无法消除的残差存在, 采用初中学过的那种两点确定一条直线的方法是无法求得方程中的具体参数值的。由于方程应当和大多数点尽量靠近, 从模型算得的预测值应当就是总体中相应个体 y 值的均数, 为此人们一般采用最小二乘法来拟合模型, 即保证各实测点至回归直线纵向距离的平方和最小。



究竟是应当把 Y 称为因变量还是应变变量? 其实这个问题不重要, 虽然在数学上可以给出很严格的区分定义, 但可以混用, 只要明白就可以了, 两者的英文其实是相同的, 即 Dependent Variable。

2. 回归系数的计算和检验

公式中 a 和 b 的数值分别通过下列公式算出:

$$b = l_{xy} / l_{xx}, a = \bar{y} - b\bar{x}, v = n - 2$$

回归系数 b 计算出来以后需要对其进行假设检验, 以确定求出了不为 0 的回归系数并不是由于抽样误差而导致的。对于回归系数的假设检验可以用 t 检验和方差分析, 公式分别如下:

(1) t 检验: 其检验统计量为 $t_b = (b - \beta) / S_b$, 其中 S_b 为回归系数的标准误, 其定义为 $S_b = S_{Y.X} \sqrt{1/l_{xx}}, v = n - 2$ 。

(2) 方差分析: 其原理和前面的单因素方差分析相同, $F = \frac{MS_{\text{回归}}}{MS_{\text{剩余}}} = \frac{SS_{\text{回归}}/v_{\text{回归}}}{SS_{\text{剩余}}/v_{\text{剩余}}}, v_{\text{回归}} = 1, v_{\text{剩余}} = n - 2$ 。

3. 总体回归线的可信区间

在应用回归分析的结果时, 经常会涉及区间估计的问题, 这里可以对回归线的总体进行可信

区间的估计,该区间估计范围在散点图上表现为一个二维空间的弧形区带,也称为回归线的置信带(Confidence Band)。以95%的区间为例,其含义是在满足线性回归的假设条件下,两条弧形曲线所形成的区域包含真实总体回归直线的置信度为95%。其标准误如下:

$$S_{\hat{Y}_X} = S_{Y.X} \sqrt{\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

相应的总体回归线 $100(1 - \alpha)\%$ 置信带为: $\hat{Y} \pm t_{\alpha(n-2)} S_{\hat{Y}}$ 。因为其方差是 x 的函数,所以其置信带在均数 (\bar{X}, \bar{Y}) 处的宽度最小,越远离该均数点,则其区间宽度越大。

4. 个体 Y 预测值的区间估计

该区间指的是当 X 为某定值时,个体 Y 值的参考值范围的波动范围,其分布的标准差 $S_{Y|X_p}$ 按下式估计:

$$S_{Y|X_p} = S_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum (X - \bar{X})^2}}$$

为了简化计算,当 X 与 \bar{X} 接近且 n 充分大时,可以用 $S_{Y.X}$ 代替 $S_{Y|X_p}$,其参考值区间为 $\hat{Y} \pm t_{\alpha(n-2)} S_{Y.X}$ 。该区间是由比总体回归线置信区带更远的两条弧形曲线构成的,以95%的区间为例,表示的是期望有95%的数据点所落入的范围。



在 SPSS 中可以使用散点图对简单回归分析进行非常直观的图形呈现,并通过编辑方式直接绘制出回归线、回归线的95%可信区间和个体值的95%参考值范围,并加绘出实测值和预测值的差距(残差),大大方便了实际使用。具体操作参见绘图一章。

5. 回归模型的适用条件与注意事项

即使进行简单回归分析,模型对数据也有一定的要求,基本的适用条件如下。

(1) 线性趋势:自变量与因变量的关系是线性的,如果不是,则不能采用线性回归来分析。这可以通过散点图来加以判断。

(2) 独立性:可表述为因变量 y 的取值相互独立,之间没有联系。反映到模型中,实际上就是要求残差间相互独立,不存在自相关性,否则应当采用自回归模型来分析。

(3) 正态性:就自变量的任何一个线性组合,因变量 y 均服从正态分布,反映到模型中,实际上就是要求 e_i 服从正态分布。

(4) 方差齐性:就自变量的任何一个线性组合,因变量 y 的方差均相同,实质就是要求残差的方差是齐性的。

如果只是建立方程,探讨自变量与因变量间的关系,而无须根据自变量的取值预测因变量的容许区间、可信区间等,则后两个条件可以适当放宽。

此外在进行回归分析时,还需要特别注意不能将回归模型的分析结果随意延伸到因果关系上去,也就是说,自变量和因变量有回归关系并不一定代表两者一定会有因果关联,但显然这是个很常见的误用和误解,虽然没有统计分析是万万不能的,但统计分析绝不是万能的!因果关系的推定需要统计学以外的专业知识,仅靠数据和模型还不能做到这一点。



遗憾的是,很多著名的学者都在这一点上犯过错,或许最知名的例子是太阳黑子活动周期和犯罪率的关系,称得上是非常漂亮的模型,毫无道理的结论。这个故事在 Google 上可以搜索到不细讲了。这里举另外一个有趣的例子:自冥王星被发现以来,最初人们以为它很大,随着观测手段的不断进步,观测工具的精度不断提高,冥王星的直径测量值就在不断地缩小:

1949 年:10 000 km

1950 年:6 000 km

1965 年:5 500 km

1977 年:2 700 km

.....

以至于后来有人用测量时间和观测直径做了回归分析,结论是在 1980 年,冥王星将会消失.....

18.2 案例:建立用年龄预测总信心指数值的回归方程

这里利用相关分析中的例题来进一步进行回归分析,计算它的回归方程。与相关分析类似,在进行回归分析前首先要考虑的问题就是两个变量之间是否可能存在某种趋势,通过前面的散点图分析已经得到了肯定的结论,因此可以直接进行回归分析。

1. 界面说明

选择“分析”→“回归”→“线性”菜单项,即可打开“线性回归”对话框,如图 18.2 所示。这里只介绍主界面上的内容,各子对话框的功能将在 18.3 节中加以介绍。

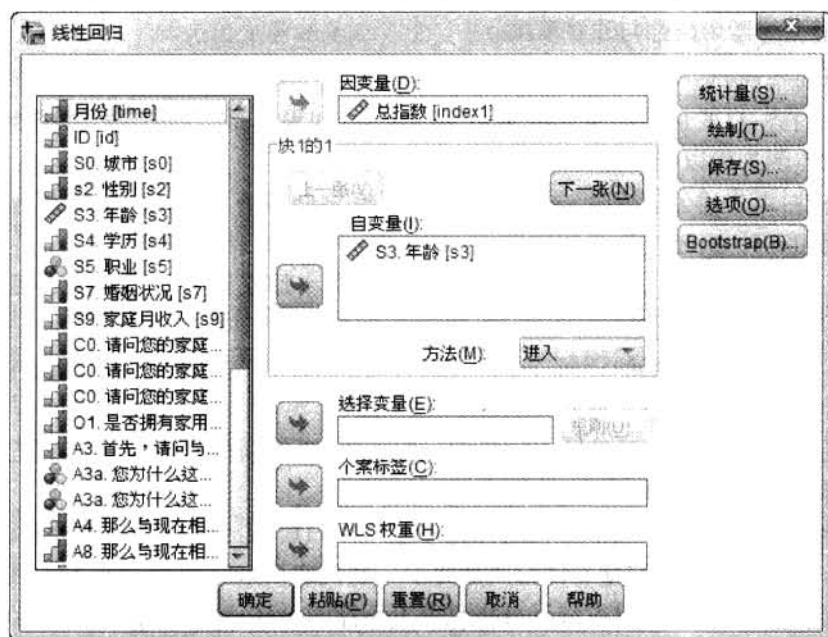


图 18.2 “线性回归”对话框

(1) “变量”列表框:用于选入回归分析模型中的因变量,只能选入一个。

(2) 块按钮组:由“上一张”和“下一张”这对按钮组成,用于将选入下面“自变量”列表框中的自变量分组。由于在多变量回归分析中自变量的选入方式有“进入”、“向后”、“逐步”等方法,如果对于不同的自变量选入的方法不同,则用该按钮组将自变量分组选入即可。

(3) “自变量”列表框:用于选入回归分析模型中的自变量,选入的方法可使用块按钮组中的按钮进行不同的定义。

(4) “方法”下拉列表框:用于选择对自变量的选入方法,包括“进入”、“后退”、“逐步”等方法,该选项对当前“自变量”列表框中的所有变量均有效。

(5) “选择变量”列表框:实际含义是进行案例筛选。选入一个筛选变量,并利用右侧的“规则”按钮建立一个选择条件,这样,只有满足该条件的记录才能进行回归分析。

(6) “个案标签”列表框:选择一个变量,它的取值将作为每条记录的标签。最典型的情况是使用记录 ID 号的变量。

(7) “WLS 权重”列表框:可选择权重变量以进行加权最小二乘法的回归分析。

2. 操作说明和结果解释

本例的操作是比较简单的,具体如下。

(1) 选择“分析”→“回归”→“线性”菜单项,打开“线性回归”对话框。

(2) 将“总指数”选入“因变量”列表框中,将“年龄”选入“自变量”列表框中。

(3) 单击“确定”按钮。

分析结果较为复杂,一共会出现 4 张表格,如图 18.3 ~ 图 18.6 所示,依次解释如下。

模型	输入的变量	移去的变量	方法
1	S3. 年龄 ^a	.	输入

a. 已输入所有请求的变量。

b. 因变量:总指数。

图 18.3 输入/移去的变量^b

首先是对模型中各个自变量纳入模型的情况进行汇总,由于本例只有一个自变量,所以结果显得比较单薄。从图 18.3 中可以看出,放入模型中的只有年龄一个变量,选择变量的方法为强行进入法,也就是将所有的自变量都放入模型中(尽管本例只有一个)。筛选自变量的方法有很多种,在不同的情况下可以选择不同的筛选方法,具体见下文。

图 18.4 所示的结果是对模型的简单汇总,其实就是对回归方程拟合情况的描述,通过这张表可以知道相关系数(绝对值)的取值(R),相关系数的平方即决定系数(R Square),调整后的决定系数(Adjusted R Square)和回归系数的标准误(Std. Error of the Estimate)。注意这里的相关系数绝对值大小和第 17 章相关分析中计算出的结果完全相同。决定系数的取值在 0 到 1 之间,

模型	R	R 方	调整 R 方	标准 估计的误差
1	.219 ^a	.048	.047	20.49596

a. 预测变量:(常量), S3. 年龄。

b. 因变量:总指数。

图 18.4 模型汇总^b

它的含义就是自变量所能解释的方差在总方差中所占的百分比,取值越大说明模型的效果越好。通俗一点来讲就是决定系数越大该因素所起的作用越大。



如果决定系数 R^2 为 0.8,则说明回归关系可以解释因变量 80% 的变异。换言之,如果能够成功地控制自变量的取值不变,则因变量的变异程度会降低 80%。调整后的决定系数主要用于对自变量数量不同的模型拟合效果进行对比,在简单回归模型中没有实际价值。

图 18.5 即为对模型进行方差分析的结果,对回归系数进行检验有两种方法,其中一种就是方差分析。在方差分析的结果中 F 值为 57.726, P 值小于 0.05,所以该模型是有统计意义的,由于只有一个自变量,也就等于说该自变量的回归系数是有统计意义的。在简单回归中方差分析的结果和 t 检验的结果完全等价,可以和图 18.6 所示的结果加以比较。

模型		平方和	df	均方	F	Sig.
1	回归	24249.673	1	24249.673	57.726	.000 ^a
	残差	480996.625	1145	420.084		
	总计	505246.298	1146			

a. 预测变量: (常量), S3. 年龄。

b. 因变量: 总指数。

图 18.5 Anova^b

模型		非标准化系数		标准化系数		t	Sig.
		B	标准 误差	试用版			
1	(常量)	108.898	1.816			59.982	.000
	S3. 年龄	-.358	.047	-.219		-7.598	.000

a. 因变量: 总指数

图 18.6 系数^a

图 18.6 为最后一张结果表格,个人认为也是最重要的一张,其中给出了回归方程中常数项、回归系数的估计值和检验结果,可见 $a = 108.898$, $b = -0.358$,由此就可以写出如下回归方程:

$$\hat{\text{总信心指数}} = 108.898 - 0.358 \times \text{年龄}$$

上述回归方程给出了如下信息。

(1) 年龄为 0 岁时,受访者的信心指数平均理论值为 108.9,显然这只是个理论值,因为 CCSS 项目的受访者必须年满 18 岁。

(2) 年龄每增加一个单位,信心指数值平均会下降 0.358 点。



注意图 18.6 中的“试用版”是一个严重的误译,其原文是 Beta,含义为标准化回归系数,实际上不需要翻译成中文。

在图 18.6 中还使用 t 检验对各参数进行了检验,其中对于常数项主要是检验其是否为 0,但这在回归问题中一般是没有实际意义的,因此不用加以关心。对回归系数的检验也拒绝了无效假设,认为上述影响的确是存在的,注意其统计量 t 值实际上就是前述方差分析 F 值的平方根,两个检验结果是完全等价的。



如果在一些特殊的建模分析中不希望模型中包含常数项,或者说想将常数项固定为0,则在“选项”子对话框中取消选中“在等式中包含常量”复选框即可。

3. 存储预测值和区间估计值

如果建立回归模型的任务不仅仅是寻找潜在的影响因素,而是希望对因变量进行预测,则往往需要在数据集中计算出预测值、个体参考值范围等。这项工作虽然可以手工进行,但显然SPSS 会提供更为便捷的功能,在“保存”子对话框中,可以将如下几类信息存储在数据集中,如图 18.7 所示。

(1) “预测值”复选框组:包含了各种可供存储的应变量预测值,包括未标准化预测值、标准化预测值(服从标准正态分布)、调节预测值(去掉当前记录时模型对该记录应变量的预测值),以及预测值的标准差(标准误)。

(2) “残差”复选框组:包含了可供存储的各种残差,可用于模型诊断。具体包括未标准化残差、标准化残差(服从标准正态分布)、学生化残差(服从 t 分布)、删除残差(调节预测值所对应的残差)、学生化已删除残差(上一个预测值进行 t 变换后的结果)。

(3) “距离”复选框组:给出一系列用于测量数据点离拟合模型距离的指标,包括“马哈拉诺夫距离”、“Cook 距离”、“杠杆值”等,这些指标主要用于强影响点的诊断。

(4) “影响统计量”复选框组:提供一些专门用于判断强影响点的统计量,如“DfBeta”、“DfFit”、“协方差比率”等。

(5) “预测区间”复选框组:要求给出均数的可信区间或个体参考值范围的上下界,默认为 95% 区间,用户可以自己设定概率值。

(6) “系数统计”框组:可以将回归系数等模型结果输出到一个新的数据文件中供后续分析使用。

(7) “将模型信息输出到 XML 文件”框组:实际含义是将所拟合的模型以 PMML 语言格式输出为 XML 文件,以便可以读取该数据交换格式的分析软件对该模型加以利用。该功能与数据挖掘应用有关,这里不再详述。

在本例中,如果希望存储预测值和个体参考值范围,则所需操作为:在“保存”子对话框中选中“未标准化”预测值、“单值”预测区间两个复选框。

这样在建模完成后,原数据集就会增加 PRE_1、LICI_1 和 UICI_1 这样 3 个新变量,分别代表每条案例的模型预测值、个体预测值 95% 参考值区间的下界和上界。



图 18.7 “保存”子对话框

18.3 多重线性回归模型入门

所谓多重线性回归模型,就是指包括一个或多个自变量的回归模型,由于自变量数可能超过一个,因此架构更为复杂,本节将对其进行概要介绍。

18.3.1 模型简介

以上面分析的年龄和信心指数的回归模型为例,如果进一步在模型中加入家庭收入,希望建立同时考虑年龄和家庭收入影响的线性回归方程,则所拟合的模型架构如下:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

这里 \hat{y} 称为 y 的估计值或预测值 (Predicted Value), 表示给定各自变量的值时, 因变量 y 的估计值; a 为截距 (Intercept), 在回归方程中又称为常数项 (Constant), 表示各自变量均为 0 时 y 的估计值; b_i 称为偏回归系数 (Partial Regression Coefficient), 简称为回归系数, 表示其他自变量不变, x_i 每改变一个单位时, 所预测的 y 的平均变化量。比如在该方程中 x_1 代表年龄, 并最终求得 $b_1 = -0.2$, 则表示当年龄上升 1 个单位 (即增加 1 岁) 时, 受访者的信心指数值平均下降 0.2 个单位。

如果从个体的角度来看待线性回归模型, 则上式可改写为如下形式:

$$y_i = \hat{y} + e_i = a + b_1x_{1i} + b_2x_{2i} + e_i$$

其中 e_i 为随机误差, 被假定为服从均数为 0 的正态分布。即对每一个个体而言, 在知道了所有自变量取值时, 能确定的只是因变量的平均取值, 个体的具体取值在其附近的一个范围内。而具体取值和平均取值间的差异 (即 e_i) 称为残差, 这一部分变异是当前模型不能确定的部分。



多重线性回归模型的适用条件和简单线性回归模型类似, 也是线性趋势、独立性、正态性、方差齐性 4 项, 但为了保证参数估计值的稳定, 还需要注意模型的样本量要求。虽然在这方面还没有精确的计算公式可供选择, 但根据人们的经验, 记录数应当在希望分析的自变量数的 20 倍以上为宜。比如希望在模型中纳入 5 个自变量, 则样本量应当在 100 以上, 少于此数则可能会出现检验效能不足的问题。此时得到的阳性结论并非不可信, 但在解释时要加倍小心, 需要时刻牢记得到的系数可能是不稳定的。

18.3.2 多重线性回归模型的标准分析步骤

多重线性回归分析被应用得非常广泛, 已经到了滥用的程度。作为一个严肃的统计学模型, 它有着自己严格的适用条件, 在拟合时也需要不断进行这些适用条件的判断。但是, 许多使用者往往忽视了这一点, 只注重把方程列出来, 这不仅浪费信息, 更有可能得出错误的结果。下面给出笔者认为比较合适的回归分析操作步骤, 仅供参考。

1. 关联趋势的图形考察

首先应当做出散点图, 观察变量间的趋势。如果是多个变量, 则还应当做出散点图矩阵、重叠散点图和三维散点图。具体做法参见绘图部分。

图 18.8 所示为 4 幅散点图,可见第 1 幅图中的两个变量间基本呈线性关系,可进行分析;第 2 幅图中实际上为曲线关系,应当进行曲线方程的拟合;第 3 幅图中的两个变量间虽然呈直线关系,但存在一个异常点,必须先对它进行考察后才能进行分析,并且在分析方法上可能要采用其他拟合方法;第 4 幅图中两变量间实际上存在的线性趋势非常微弱,但由于一个异常点的出现使得这种关系被虚假地大大增强。这种情况同样要先考察异常点,并考虑采用其他拟合方法来分析。第 4 种情况应当引起充分的注意,因为该异常点离线性趋势线不远,许多人都错误地把它当成是正常情况。

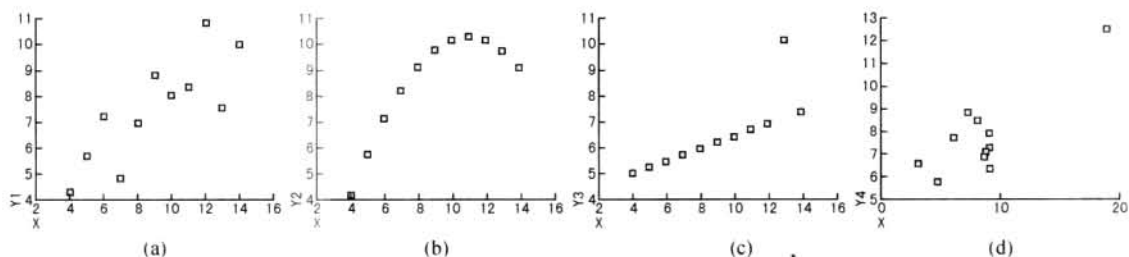


图 18.8 不同情况下的散点图

从上面这 4 幅图就可以看出,做出散点图是进行线性回归分析之前的必要步骤,不能随意省略。

2. 数据分布考察与预处理

用统计量或者图形考察数据的分布,进行必要的预处理,即分析变量的正态性、方差齐等问题,并确定是否可以直接进行线性回归分析。如果进行了变量变换,则应当重新绘制散点图,以确保线性趋势在变换后仍然存在。

3. 初步建模

对数据进行直线回归分析,包括变量的初筛、变量选择方法的确定等,这里不再重复。

4. 残差分析

建模完毕后就开始模型诊断的工作,残差分析是模型诊断过程的第一步,主要分析以两大方面。

(1) 残差间是否独立:一般采用 Durbin-Watson 残差序列相关性检验进行分析。

(2) 残差分布是否为正态:可以采用残差列表及一些相关指标来分析,但最重要和直观的方法为图示法。在图 18.9 所示的 3 幅残差图中,第 1 幅残差分布非常好,没有什么问题;第 2 幅图中的残差虽然围绕均线均匀分布,但波动范围随着拟合值增大而增大,提示方差不齐,模型假设不成立,应当进行变量变换,或采用加权最小二乘法分析;第 3 幅图最差,残差随着拟合值的不同有明显趋势,提示因变量与自变量间并非直线关系,应该按曲线趋势进行拟合。

5. 强影响点的诊断及多重共线性问题的判断

这两个步骤和残差分析往往混在一起,难以完全分出先后,由于操作较为复杂,具体的方法和操作可参见高级教程的相应内容。

只有以上 5 步全部通过,研究者才能认为得到的是一个统计学上无误的模型,下一步就是结合专业实际,将分析结果运用到现实中,来看看结果有无实用价值,以及是否存在应用中的其他问题。

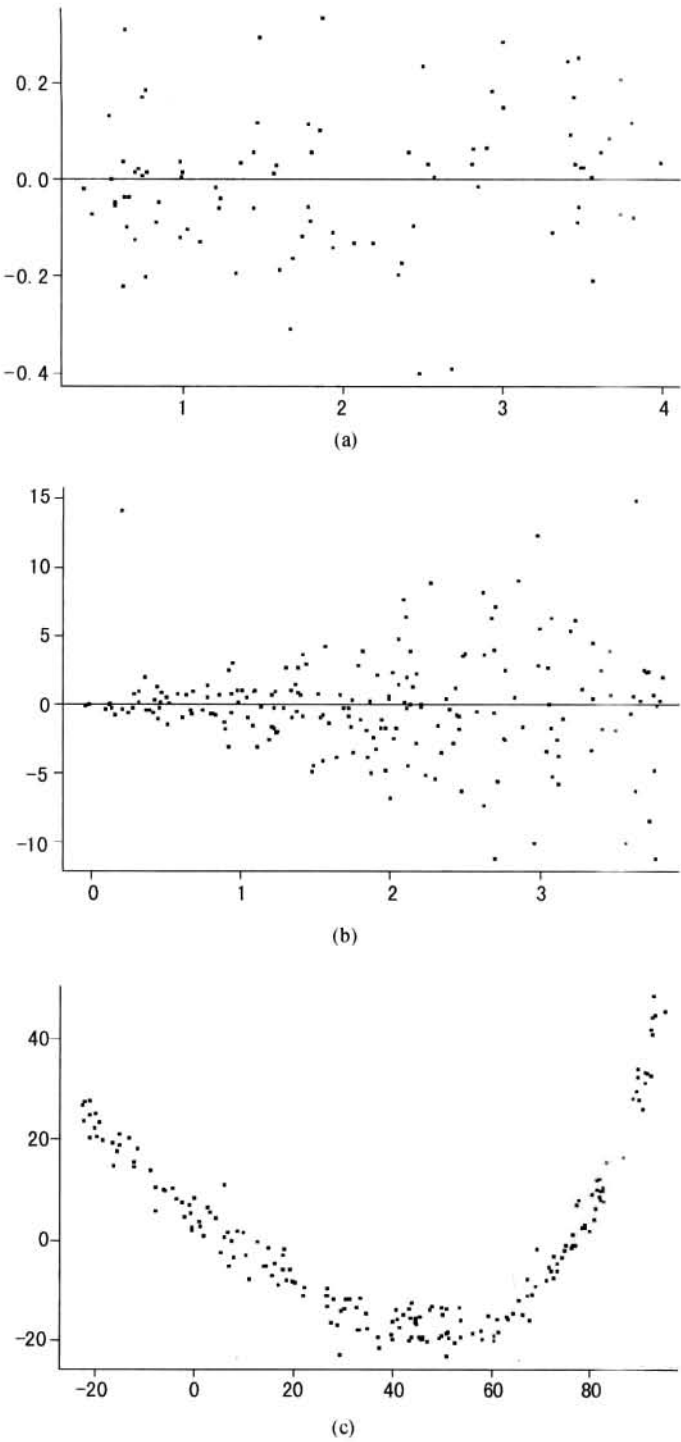


图 18.9 几种常见的残差分布情况

18.3.3 回归方程中的自变量筛选方法



这里郑重提醒一句:变量选择不是纯粹的数学问题,不能孤立于背景来考虑。许多时候,专业上的判断要优先于统计学检验的结果。

多重线性回归分析中一个重要的问题就是如何进行自变量的筛选,虽然最为稳妥的方法是分别建立简单线性回归方程,然后根据所得到的结果建立多重线性回归方程。但 SPSS 也提供了一些方法可以简化分析人员的工作,具体来说包括下面几种。

1. 进入法

所有选入“自变量”列表框中的变量均进入模型,不涉及变量筛选问题,为默认选项。

2. 向前法

(1) 首先分别对 p 个自变量 (x_1, x_2, \dots, x_p) 拟合与因变量的简单线性回归模型,共有 p 个。

(2) 考察其中有统计学意义的 k 个简单线性回归模型 $(k \leq p)$,将其中 P 值最小的模型所对应的自变量 (x_i) 首先引入模型。如果所有模型均无统计学意义,则运算过程终止,没有模型被拟合。

(3) 在已经引入模型的 x_i 的基础上,再分别拟合引入模型外的 $p-1$ 个自变量的线性回归模型。即自变量组合为 $x_i + x_1, \dots, x_i + x_{i-1}, x_i + x_{i+1}, \dots, x_i + x_p$ 的 $p-1$ 个线性回归模型。将这些模型中统计学检验 P 值最小且有统计学意义的那个自变量 (x_j) 引入模型。如果除 x_i 之外的 $p-1$ 个自变量中没有一个有统计学意义,则运算过程终止,SPSS 给出模型 $\hat{y} = a + b_i x_i$ 的参数估计。

(4) 如此反复进行,直至模型外的自变量均无统计学意义为止。

3. 向后法

(1) 与向前法相反,首先对因变量拟合包含全部 p 个自变量的线性回归模型。

(2) 考察其中无统计学意义的 k 个自变量 $(k \leq p)$,将其中 P 值最大的 (x_i) 剔除出模型。如果所有自变量 P 值均有统计学意义,则运算过程终止,SPSS 给出包含所有自变量的线性回归模型。

(3) 对因变量拟合包含剩下的 $p-1$ 个自变量的线性回归模型,同样剔除 P 值最大且无统计学意义的变量。

(4) 如此反复进行,直至模型中剩余的所有自变量均有统计学意义为止。

4. 逐步回归法

前进法只进不出,而后退法只出不进,在逻辑上似乎都有缺陷,而逐步法则将上面两种方法结合起来筛选自变量,显得更为完善。逐步法的前两步与前进法相同,从第 3 步开始的步骤如下。

(1) 考察第 1 步引入模型的自变量 (x_i) 是否仍有统计学意义。若没有统计学意义,则将其剔除出模型。

(2) 拟合包含第 2 步引入模型的自变量 (x_j) 与除 x_i 外的 $p-2$ 个自变量的模型,将其中 P 值最小且有统计学意义的引入模型。此时若没有自变量有统计学意义,则运算过程终止,SPSS 给出仅包含自变量 x_j 的模型参数估计结果。

(3) 如果第 1 步引入模型的自变量 (x_i) 有统计学意义,则进行第 4 步。在模型引入自变量

x_i, x_j 的基础上继续拟合包含其他 $p-2$ 个自变量的回归模型,考察剩余的 $p-2$ 个自变量是否有统计学意义。引入 P 值最小且有统计学意义的自变量。如果剩余的 $p-2$ 个自变量均无统计学意义,则运算过程终止,SPSS 输出包含 x_i, x_j 的回归模型参数估计结果。

(4) 如此反复进行,直至模型外的自变量均无统计学意义,而模型内的自变量均有统计学意义。

由此可见,与前进法、后退法相比,逐步回归是比较“负责任”的,每向模型引入一个新变量,均要考察原来在模型中的自变量是否还有统计学意义,是否可以将其剔除。

5. 删除法

规定为删除(Remove)的自变量将被强制剔除出模型,但 SPSS 会给出如果将其引入模型的参数估计及检验结果。该方法实际上是和将自变量分块的功能联合使用的。



对于不同的自变量纳入方法,在 SPSS 中可通过将其分为不同的“块”决定其进入模型的方式,同一区块中的自变量进入模型的方式需要相同,而不同块则可以完全不同。

18.3.4 SPSS 中与多重线性回归模型相关的功能

在 SPSS 的“回归”子菜单中提供了相当丰富和强大的回归建模功能,但就本章所涉及的多重线性回归模型而言,主要可能会用到以下两个过程。

1. 线性回归过程

前面几节使用的线性回归过程本身就是一个非常强大的回归模型拟合过程,除对模型进行拟合外,该过程的各子对话框还提供了非常强大的模型诊断、模型输出等功能,完全可以满足复杂多重线性回归模型的变量筛选和建模工作。

这里简单介绍一下前面未加以介绍的几个子对话框的功能,以便读者对该过程有一个较全面的了解。

(1) “统计量”子对话框(如图 18.10(a)所示):用于选择所需要的各种统计量,“回归系数”复选框组用于在结果中输出回归系数的估计值和检验结果,以及其可信区间、协方差矩阵等;“残差”复选框组则用于输出残差诊断的信息,包括 Durbin-Watson 残差序列相关性检验、超出规定的 n 倍标准误的残差列表;右侧的复选框用于输出模型诊断相关的指标,包括决定系数,自变量间的相关、部分相关和偏相关系数,一些共线性诊断统计量,如特征根(Eigenvalues)、方差膨胀因子(VIF)等。

(2) “图”子对话框(如图 18.10(b)所示):用于指定残差分析中所需的图形输出,可直接绘制残差的直方图和 PP 图;也可使用左侧列表中预设的模型变量绘制各种散点图;而右下角的“产生所有部分图”的含义是对于每一个自变量绘出它与因变量残差的散点图,主要也用于回归诊断。

(3) “选项”子对话框(如图 18.10(c)所示):设置回归分析的一些选项,“步进方法标准”框组用于设置进入和删除标准,可按 P 值或 F 值来设置,“在等式中包含常量”复选框用于决定是否在模型中包括常数项,默认选中;“缺失值”框组则用于选择对缺失值的处理方式,可以是不分析选入的任何有缺失值的变量的记录(Exclude Cases Listwise)而无论该缺失变量最终是否进入模型,不分析选入的变量有缺失值的记录(Exclude Cases Pairwise),或者将缺失值用该变量的均数代替(Replace with Mean)。



图 18.10 线性回归过程的“统计量”、“图”和“选项”子对话框

2. 自动线性建模过程

该过程是 SPSS 近几个版本向自动化、智能化操作平台转化的成果之一,利用该过程,用户可以采用几乎完全自动的方式进行自变量的预变换、筛选、模型优化、检验等工作。自变量也可以是连续、有序、无序等任何一种测量尺度,系统会自动选择相应的转换方式/算法来加以分析。

由于自动准备数据过程的自动化程度较高,中间又会涉及一些比较复杂的算法,因此笔者并不建议不了解回归模型细节的初学者使用,本书将不对其进行深入介绍,对该过程感兴趣的读者可参考《SPSS 统计分析高级教程》。

18.3.5 案例:建立自变量包括年龄、家庭收入的信心指数回归方程

例 18.1 建立候选自变量包括年龄、性别、家庭收入的总信心指数回归方程。

本例需要建立包括 3 个候选自变量的回归方程,由于事前不能确定这些自变量是否均具有统计学意义,因此出于简化模型的思路,可以考虑用向后法进行变量筛选。

1. 操作说明与结果解释

- (1) 选择“分析”→“回归”→“线性”菜单项,打开“线性回归”对话框。
- (2) 将“总指数”选入“因变量”列表框中,将“年龄”、“性别”、“Qs9”选入“自变量”列表框中。
- (3) 在“方法”下拉列表框选择“向后法”选项。
- (4) 单击“确定”按钮。

由于进行了变量筛选,最终输出的表格较多,依次解释如下。

图 18.11 给出了 SPSS 在回归过程中每个步骤所进行的操作,第 1 步是采用进入法选入了全部 3 个候选自变量,然后在第 2 步剔除了性别,原因是其检验概率大于 0.1 的剔除标准。

模型	输入的变量	移去的变量	方法
1	Qs9, S2, 性别, S3, 年龄		输入
2		S2, 性别	向后(准则: F-to-remove >= .100 的概率)。

- a. 已输入所有请求的变量。
- b. 因变量: 总指数。

图 18.11 输入/移去的变量^b

图 18.12 给出了两步分析中所拟合模型的决定系数,可见总决定系数几乎没有下降,而校正的决定系数在自变量较少的第 2 个模型中还略有上升,这说明被剔除的自变量的确不应当被选入模型中。

模型	R	R 方	调整 R 方	标准 估计的误差
1	.231 ^a	.053	.050	20.93061
2	.231 ^b	.053	.051	20.92005

- a. 预测变量: (常量), Qs9, S2. 性别, S3. 年龄。
- b. 预测变量: (常量), Qs9, S3. 年龄。

图 18.12 模型汇总

图 18.13 分别给出了检验所拟合的两个模型是否在整体上具有统计学意义的结果,显然两个模型都是具有一定预测价值的。

模型		平方和	df	均方	F	Sig.
1	回归	24360.728	3	8120.243	18.536	.000 ^a
	残差	432833.227	988	438.090		
	总计	457193.955	991			
2	回归	24359.741	2	12179.871	27.830	.000 ^b
	残差	432834.214	989	437.648		
	总计	457193.955	991			

- a. 预测变量: (常量), Qs9, S2. 性别, S3. 年龄。
- b. 预测变量: (常量), Qs9, S3. 年龄。
- c. 因变量: 总指数。

图 18.13 Anova^c

图 18.14 输出了两个模型中自变量的偏回归系数估计,其中显示为 0 的家庭收入回归系数可以进入表格编辑状态看到精确值。注意第一个模型中性别的检验 P 值高达 0.962,没有统计学意义;而从回归系数的估计值可见,当性别被剔除出模型之后,年龄、家庭收入的系数值基本未发生变化,这也间接支持了应当将性别剔除出模型。

模型		非标准化系数		标准系数		t	Sig.
		B	标准 误差	试用版			
1	(常量)	108.238	2.960			36.565	.000
	S2. 性别	.064	1.339	.001		.047	.962
	S3. 年龄	-.362	.052	-.217		-6.948	.000
	Qs9	.000	.000	.054		1.721	.086
2	(常量)	108.330	2.238			48.409	.000
	S3. 年龄	-.362	.052	-.217		-6.952	.000
	Qs9	.000	.000	.054		1.721	.085

- a. 因变量: 总指数

图 18.14 系数^a

图 18.15 则给出了所有被剔除出模型的变量的检验结果,包括如果将其选入模型之后的回归系数估计、偏相关系数、共线性统计量等。这里的偏相关系数是控制模型中所包含的自变量之后所计算出的模型残差与该自变量的偏相关系数,显然绝对数值越小,就越说明该自变量没有必

要被选入模型中。


					共线性统计量	
模型		Beta In	t	Sig.	偏相关	容差
2	S2. 性别	.001 ^a	.047	.962	.002	.996

- a. 模型中的预测变量: (常量), Qs9, S3. 年龄。
- b. 因变量: 总指数。

图 18.15 已排除的变量^b

本次分析最终得到的是包含两个自变量的回归方程,其表达式如下:

总信心指数 = 108.33 - 0.362 × 年龄 + 1.65 × 10⁻⁴ × 家庭月收入

 注意在最终得到的模型 2 中,家庭收入的 P 值是大于 0.05 的,之所以仍被保留在模型中,是因为此处向后法所用的剔除标准是 P ≥ 0.1,候选自变量相应的选入和剔除标准可以在“选项”子对话框中加以修改。

2. 残差的独立性检验

虽然上面已经得到了所需的回归方程,并进行了相应的假设检验,但分析工作不应当就此停止,因为数据是否满足回归模型的适用条件这一问题还未得到彻底回答,在上面的工作中至多只能算是完成了线性趋势的考察,而独立性、正态性和方差齐性方面均未涉及。下面就来完成这些工作,首先是残差的独立性检验,这可以通过选中“统计量”子对话框中的“Durbin-Watson 检验”复选框来进行,相应的输出结果如图 18.16 所示。

模型	R	R 方	调整 R 方	标准 估计的误差	Durbin-Watson
1	.231 ^a	.053	.051	20.92005	1.880

- a. 预测变量: (常量), Qs9, S3. 年龄。
- b. 因变量: 总指数。

图 18.16 模型汇总^b

可见在模型汇总表格的右侧增加了 Durbin-Watson 统计量的输出。该统计量的取值在 0 ~ 4 之间。具体应用可查相应统计用表,若大于界值上界,则说明残差间相互独立;若低于下界,则说明残差间存在自相关性。一般的,若自变量数少于 4 个,统计量大于 2,基本上可以肯定残差间相互独立,本例的计算结果为 1.88,显然独立性是没有问题的。

3. 残差分布的图形观察

下面利用图形来进行残差的独立性检验,首先在“图”子对话框中选中“直方图”和“正态概率图”复选框。

从输出的残差直方图和 P - P 图中可以看出,模型的残差基本上服从正态分布,没有严重偏离正态性假设,如图 18.17 所示。但是在左侧存在个别数值低于 -4 的案例,说明模型中有可能存在强影响点。

需要注意的是,自变量与因变量间的关系并非线性、残差方差不齐、观测值间不独立等情况均会导致残差的直方图、茎叶图、P - P 图等表现出非正态,因此建议在确认残差服从线性回归的其他几项条件后,再来研究残差分布是否为正态分布。

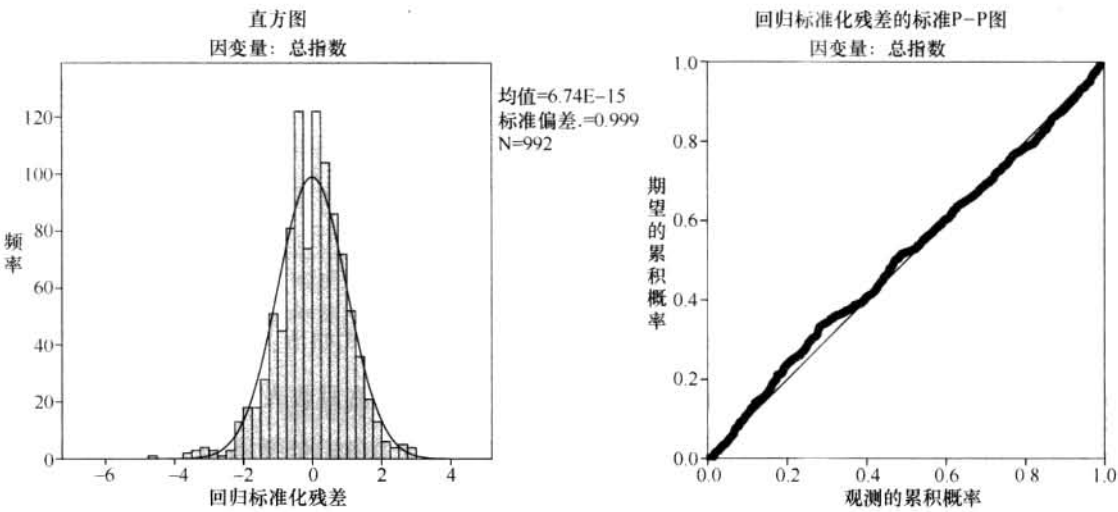


图 18.17 模型残差的直方图与 P-P 图

4. 方差齐性的图形观察

方差齐性同样也可以通过残差图来加以考察,操作方法如下:在“图”子对话框中将 ZPRED (标准化预测值) 选入“X2”列表框中,ZRESID(标准化残差)选入“Y”列表框中。

最终绘制出的散点图如图 18.18 所示,从中可以了解如下信息。

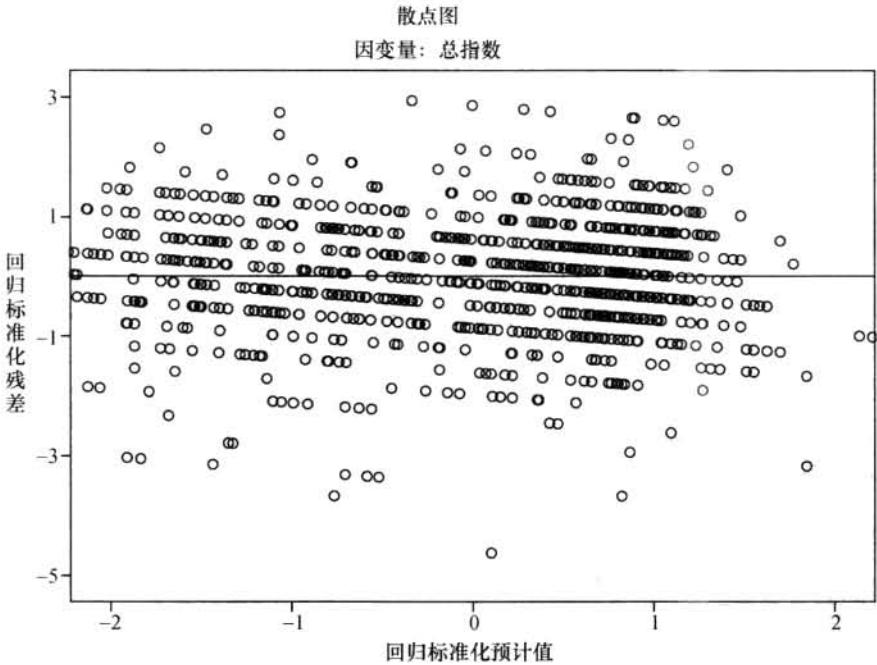


图 18.18 标准化预测值和标准化残差的散点图

(1) 随着预测值的上升,残差似乎存在轻微的减少趋势,这一般是暗示在现有变量中仍有信息需要选入,可能是变量的高次项,也有可能是交互项。

(2) 的确存在个别残差偏离较远的案例,一般而言残差绝对值大于3就需要加以注意,大于5则最好进行有针对性的分析评估,本例就属于此类情况。

由于篇幅和深度所限,对于在上述残差分析中发现的线索本书将不再继续进行深入分析,有兴趣的读者可自行完成相应工作。

18.4 本章小结

(1) 相关分析和回归分析具有密切的联系,如果要用统计指标对变量数量联系的密切程度进行表述,则应当进行相关分析;如果希望反映一个变量变化时对另一个变量数量的影响大小,则应当使用回归分析。相关系数 ρ 大小反映了两个变量之间的密切程度,而回归系数 β 反映了 X 与 Y 对应的平均数量变化关系,两者的正负号和假设检验是一致的,但两者没有定量对应关系。

(2) 多重线性回归模型可以使用向前法、向后法、逐步法等多种回归分析方法来协助进行自变量筛选,但这些方法在正式分析中应当处于辅助地位,自动筛选不能完全替代人工筛选。

(3) 回归模型有着自己严格的适用条件,在拟合时需要不断进行这些适用条件的判断。标准的回归模型建模步骤应当包括如下内容:做出散点图,观察变量间的趋势;考察数据的分布,进行必要的预处理;进行直线回归分析,建立基本模型;进行残差分析;进行强影响点的诊断及多重共线性问题的判断。只有以上5步全部通过,研究者才能认为得到的是一个统计学上无误的模型,下一步应当做的就是结合专业实际,将分析结果运用到现实中,来看看结果有无实用价值,以及是否存在应用中的其他问题。

思考与练习

按照回归分析的方式对上一章中偏相关分析的实例进行重新分析,在建立回归方程后,使用“保存”子对话框中右上角的功能存储残差(选中“未标准化”复选框),最后用这两组残差进行简单相关分析,并将结果和直接用偏相关分析得到的结果相比较。

第 19 章 统计实战案例集锦(三)

19.1 X 药物对原发性高血压治疗的临床试验研究

19.1.1 项目背景

1. X 药物的基本情况

X 药物是由 * 公司研发的长效二氢吡啶钙离子拮抗剂,国外临床研究表明该药可平稳、长效地降血压,而且对老年人高血压的治疗疗效稳定、安全性高,对伴有肾功能损害的高血压病患者具有肾脏保护作用。

2. 研究目的

以苯磺酸氨氯地平为对照药,通过多中心、随机、双盲、平行对照试验,验证 X 药物单独给药对原发性高血压病患者的疗效和安全性。

19.1.2 研究方法

1. 试验方案简述

(1) 使用药品:被试制剂为 X 药物,参比制剂为苯磺酸氨氯地平商品名 ***。

(2) 随机和分组:本试验在门诊原发性高血压患者中完成。经过筛选,159 例合格的原发性高血压患者在治疗期 0 周随访时被随机分组。研究者按照患者 0 周来访的先后顺序由小到大分配随机号码。试验组接受 X 药物,对照组接受氨氯地平。

(3) 诊室血压研究内容:患者进行病史回顾,接受全面体检。签署书面知情同意书。

① 试验开始前 2 周服用过任何降压药的患者($\text{SeDBP} \geq 90 \text{ mmHg}$ 且 $\text{SeSBP} < 180 \text{ mmHg}$),且完全符合观察期开始时的入选标准和排除标准,启动观察期,同时停药所有的抗高血压药物。观察期持续时间至少为中止服药后 2 周以上。观察期结束后进行血、尿、心电图及胸部平片检查,完全符合治疗期开始时的入选标准和排除标准者(SeDBP 为 $95 \sim 114 \text{ mmHg}$ 且 $\text{SeSBP} < 180 \text{ mmHg}$)启动治疗期。

② 试验开始前 2 周末服用过任何降压药的患者($\text{SeDBP} \geq 95 \text{ mmHg}$ 且 $\text{SeSBP} < 180 \text{ mmHg}$),且完全符合观察期开始时的入选标准和排除标准,进行血、尿、心电图及胸部平片检查,若完全符合治疗期开始时的入选标准和排除标准者(SeDBP 为 $95 \sim 114 \text{ mmHg}$ 且 $\text{SeSBP} < 180 \text{ mmHg}$)启动治疗期。

2. 试验对象

(1) 患者来源:4 个临床中心于 2003 年 9 月~2004 年 2 月选择中国人原发性高血压门诊患者 159 例(平均坐位舒张压(SeDBP)为 $95 \sim 114 \text{ mmHg}$ 且平均坐位收缩压(SeSBP) $< 180 \text{ mmHg}$)。经 2 周停药观察期后,随机双盲接受 X 药物($n=126$)或氨氯地平($n=125$),治疗 8 周。

(2) 入选标准、排除标准、退出实验标准:此处略。

3. 疗效评价和安全性评价

(1) 主要疗效评价指标:根据治疗期开始前和治疗结束时坐位舒张压的差值评价降压效果。

疗效按以下方式分类:显效、有效或无效(根据我国药审现行要求)。

- ① 显效:舒张压下降 ≥ 10 mmHg 并降到正常,或下降 ≥ 20 mmHg。
- ② 有效:舒张压下降虽未达到 10 mmHg 但降到正常,或下降 10 ~ 19 mmHg。
- ③ 无效:未达到上述标准。

如为收缩期性高血压,收缩压下降 ≥ 30 mmHg 时有效。

(2) 次要疗效评价指标:血压正常化率,以治疗期末的 SeDBP 达到正常范围(即低于 90 mmHg)的病例数为分子,以全部可供疗效评价的入选病例作为分母,统计血压正常化率。

(3) 安全性参数

- ① 不良事件。
- ② 体格检查发现。
- ③ 胸部 X 片(正位)。
- ④ 导联心电图检查。
- ⑤ 实验室检查:红细胞计数、血红蛋白、红细胞压积、白细胞计数、白细胞分类、血小板计数;总蛋白、白蛋白、ALT、AST、GGT、ALP、LDH、总胆红素、总胆固醇、HDL、TG、血糖、BUN、肌酐、尿酸、Na、K、Cl、CPK;尿蛋白、尿糖、尿胆原、尿沉渣。

4. 样本量计算

(1) 根据马来酸氨氯地平与苯磺酸氨氯地平治疗原发性高血压的临床试验结果(中国新药杂志 2000,9,12),每组 60 例的对照试验得到的有效率分别为 87% 和 85%。

(2) X 药物与尼卡地平对照的临床试验(Therapeutic res, 1990,11:1657)所得到的有效率为 84%。

(3) 本次试验采用苯磺酸氨氯地平为对照药,假设苯磺酸氨氯地平或 X 药物的有效率为 85%,而对照的另一方有效率为 70%,显著性水平为 0.05,把握度为 80% 时,可算出样本数为每组 100 例。假设脱落率为 20%,本试验至少需要选入有效病例 240 例。



为了保护信息,本案例只提供了原始的 251 例真实样本中的 159 例用于分析。

5. 统计学分析

(1) 疗效分析采用符合方案分析集(PPS)。安全性分析采用安全集(SS)。

(2) 计量资料用均数、标准差、中位数等表示,根据正态性检验结果采用 t 检验,或组间采用 Wilcoxon 秩和检验,组内前后比较采用配对 t 检验或符号秩和检验。计数资料用频数、构成比等表示,采用 χ^2 检验。

(3) 所有统计检验采用双侧检验,显著性水平定为 0.05。

19.1.3 数据准备

在实验过程中,出于数据安全的需要,将原始数据备份保存存在不同的文件中,具体在本案例中涉及如下几个数据文件。

(1) Xdrug_key:病例分组文件,包括病例的 ID 号(由流水号 + 中心号组成),以及具体的药品分组,在揭盲之前,两个药品组分别用代号 A、B 表示。

(2) Xdrug_main:存储了病例的所在中心号、入组流水号、受试者背景变量、实验室检查结果、不良事件和不良反应记录等。



实际上,本案例的原始数据存储方式要比现在描述的复杂得多,为了便于理解,同时也为了保护真实的试验信息,这里进行了大幅度的简化。

因此在进行正式的统计分析之前,分析师必须要将不同的数据库按照索引变量合并起来,并计算出分析时所需要的一些变量,如疗效结果等,具体的 SPSS 程序如下:

* 分别读入所需的数据文件,假设两个文件均放置在 E 盘根目录下。

GET

FILE = 'E:\XDRUG_KEY.SAV'.

DATASET NAME KEY.

GET

FILE = 'E:\XDRUG_MAIN.SAV'.

DATASET NAME MAIN WINDOW = FRONT.

* 计算 ID 变量。

STRING ID (A10).

COMPUTE ID = CONCAT(STRING(RANDONO,F2.0),"-",STRING(CENTERNO,F1.0)).

EXECUTE.

* 文件 KEY 中已排序,只需对 MAIN 进行排序操作。

SORT CASES BY ID(A).

MATCH FILES /FILE = *

/FILE = 'KEY'

/BY ID.

EXECUTE.

* 计算疗效指标。

COMP TREATRES = 0.

IF V1SBP > 180 & V2SBP - V1SBP >= 30 TREATRES = 1.

IF (V2DBP - V1DBP < 10 & V2DBP < 95) | RANGE(V2DBP - V1DBP, 10, 19) = 1
TREATRES = 1.

IF (V2DBP - V1DBP >= 10 & V2DBP < 95) | V2DBP - V1DBP >= 20 TREATRES = 2.

EXECUTE.

VARIABLE LEVEL TREATRES = SCALE.

COMP BPNORMAL = 0.

IF V2DBP < 95 & V2SBP < 180 BPNORMAL = 1.

EXECUTE.

19.1.4 基线情况比较

- 这里需要进行比较的指标分为以下几大类。
- (1) 性别、年龄、身高、体重、体重指数等基本背景变量。
 - (2) 高血压病史和病程、家族史、高血压患者服药情况、其他既往疾病史等疾病相关的背景变量。
 - (3) 基线血压、血细胞、实验室检查的相关变量。
- 分析以上指标的核心目的是确认两组患者的确符合随机入组的要求。因变量较多,这里只以少数变量的分析结果为例来说明具体的操作。

1. 性别的组间比较

- (1) 选择“分析”→“描述统计”→“交叉表”菜单项,打开“交叉表”对话框。
- (2) 在“行”列表框中选入分组变量“keys”。
- (3) 在“列”列表框中选入“sex”。
- (4) 单击“统计量”按钮,打开“统计量”子对话框中选中“卡方”复选框。
- (5) 单击“确定”按钮。

分析结果如图 19.1、图 19.2 所示,由于样本量满足卡方检验的要求,因此可以直接使用卡方检验的 P 值 0.474。性别分布在两组间无统计学差异。

计数		受试者性别		SEX	
		1	2		合计
keys	A	43	39		82
	B	36	41		77
合计		79	80		159

图 19.1 keys* 受试者性别 SEX 交叉制表

	值	df	渐进 Sig. (双侧)	精确 Sig. (双侧)	精确 Sig. (单侧)
Pearson 卡方	.514 ^a	1	.474		
连续校正 ^b	.311	1	.577		
似然比	.514	1	.473		
Fisher 的精确检验				.527	.289
有效案例中的 N	159				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 38.26。
b. 仅对 2 × 2 表计算。

图 19.2 卡方检验

2. 身高、体重、体重指数等的组间比较

按照分析方案规定,首先应当进行正态分布检验。

- (1) 选择“分析”→“非参数检验”→“单样本”菜单项。
- (2) 在打开的对话框中选择“目标”选项卡,选择“自动比较观察数据和假想数据”选项。
- (3) 选择“字段”选项卡,设置使用定制字段分配,将“身高”、“体重”、“体重指数”选入“检验变量”列表框中。
- (4) 选择“设置”选项卡,选择“自定义检验”选项组中的第二项“检验观察分布和假设分布”,相应选项中的假设分布选择“正态分布”。
- (5) 单击“运行”按钮。

结果如图 19.3 所示,可以清楚地看到这 3 个变量的 K-S 检验都没有拒绝正态性假设,因此可以直接进行标准的两样本 *t* 检验。

假设检验汇总				
	原假设	测试	Sig.	决策者
1	身高(cm) HEIGHT 的分布为正态分布, 平均值为165.97, 标准差为8.43。	单样本 1Kolmogorov-Smirnov 检验	.459	保留原假设。
2	体重(kg) WEIGH 的分布为正态分布, 平均值为72.93, 标准差为10.48。	单样本 1Kolmogorov-Smirnov 检验	.326	保留原假设。
3	体重指数 WINDEX 的分布为正态分布, 平均值为26.43, 标准差为2.96。	单样本 1Kolmogorov-Smirnov 检验	.779	保留原假设。

显示渐进显著性。显著性水平是 .05。

图 19.3 身高、体重等的正态分布检验结果



这里的分析思路和本书前面所介绍的不太一样,是严格按照正态分布检验的结果来进行后续方法的选择的,而本书更推荐的是灵活处理。这是因为临床试验在统计分析方面的相关规定是非常死板和严格的(主要是防止漏洞被制药公司利用),因此没有什么变化可讲,只能遵照指导原则进行。

- (1) 选择“分析”→“比较均值”→“独立样本 *t* 检验”菜单项,打开“独立样本 *t* 检验”对话框。
- (2) 在“检验变量”列表框中选入“身高”、“体重”、“体重指数”选项。
- (3) 在“分组变量”列表框中选入“keys”选项。
- (4) 单击“定义组”按钮,打开“定义组”对话框,在“组 1”、“组 2”文本框中分别输入“A”和“B”。
- (5) 单击“确定”按钮。

结果如图 19.4、图 19.5 所示,身高、体重、体重指数的两样本 *t* 检验均为方差齐,*P* 值分别为 0.306、0.593 和 0.770,均无统计学意义。

3. 基线血压的比较

基线血压比较的分析思路和上面的身高、体重等基本相同,这里省略对同类操作的介绍,直接给出结果。首先是正态性检验的结果,如图 19.6 所示。

		keys	N	均值	标准差	均值的标准误
身高(cm)	HEIGHT	A	82	166.63	8.719	.963
		B	77	165.26	8.108	.924
体重(kg)	WEIGHT	A	82	73.3598	10.45545	1.15461
		B	77	72.4675	10.55666	1.20304
体重指数	WINDEX	A	82	26.3659	2.71772	.30012
		B	77	26.5039	3.22284	.36728

图 19.4 组统计量

		方差方程的 Levene 检验				均值方程的 t 检验	
		F	Sig.	t	df	Sig. (双侧)	均值 差值
身高(cm)	假设方差相等	.395	.531	1.028	157	.306	1.374
	假设方差不相等			1.030	156.987	.305	1.374
体重(kg)	假设方差相等	.006	.940	.535	157	.593	.89222
	假设方差不相等			.535	156.169	.593	.89222
体重指数	假设方差相等	3.176	.077	-.293	157	.770	-.13804
	假设方差不相等			-.291	149.034	.771	-.13804

图 19.5 独立样本检验

假设检验汇总

	原假设	测试	Sig.	决策者
1	V1SBP 的分布为正态分布，平均值为153.80，标准差为12.86。	单样本 1Kolmogorov-Smirnov 检验	.120	保留原假设。
2	V2SBP 的分布为正态分布，平均值为135.17，标准差为13.33。	单样本 1Kolmogorov-Smirnov 检验	.136	保留原假设。

显示渐进显著性。显著性水平是 .05。

图 19.6 血压的正态分布检验结果

分析结果如图 19.7 所示,可见基线收缩压和舒张压的正态性检验 P 值均大于 0.05,不能拒绝其正态分布的假设。

		方差方程的 Levene 检验				均值方程的 t 检验	
		F	Sig.	t	df	Sig. (双侧)	均值 差值
V1SBP	假设方差相等	.146	.703	.758	157	.450	1.549
	假设方差不相等			.757	155.939	.450	1.549
V2SBP	假设方差相等	1.341	.249	1.385	157	.168	2.923
	假设方差不相等			1.389	156.991	.167	2.923

图 19.7 独立样本检验

结果显示基线收缩压、舒张压的分布在两组间均无统计学差异。

4. 基线比较结论

两组基线坐位血压无统计学差异。两组药前年龄、性别、身高、体重、体重指数及高血压病史和病程、家族史、高血压患者的服药情况以及其他既往疾病史等分布的差别均无统计学差异,说明两组患者符合随机入组的要求。

19.1.5 疗效比较

1. 总有效率

由于所有病例中无人被分为显效,只被分为无效和有效两类,因此分析可以只采用卡方检验来进行。

如图 19.8 所示,两种药物的治疗后总有效率分别为 80.5% 和 88.3%。

		TREATRES		合计	
		.00	1.00		
keys	A	计数	16	66	82
		keys 中的 %	19.5%	80.5%	100.0%
	B	计数	9	68	77
		keys 中的 %	11.7%	88.3%	100.0%
合计		计数	25	134	159
		keys 中的 %	15.7%	84.3%	100.0%

图 19.8 keys * TREATRES 交叉制表

如图 19.9 所示,总有效率组间比较 $P=0.176$,无统计学差异。

	值	df	渐进 Sig. (双侧)	精确 Sig. (双侧)	精确 Sig. (单侧)
Pearson 卡方	1.834 ^a	1	.176		
连续校正 ^b	1.292	1	.256		
似然比	1.859	1	.173		
Fisher 的精确检验				.197	.128
有效案例中的 N	159				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 12.11。
b. 仅对 2 × 2 表计算。

图 19.9 卡方检验

2. 坐位血压下降情况

这里需要首先将数据按组别进行拆分,以便分别比较两组治疗前后的血压下降情况。

- (1) 选择“数据”→“拆分文件”菜单项,打开“拆分文件”对话框。
- (2) 比较组,在“分组方式”列表框中选中“keys”选项。
- (3) 单击“确定”按钮。
- (4) 选择“分析”→“比较均值”→“配对样本 t 检验”菜单项,打开“配对样本 t 检验”对话框。

(5) 在“检验变量”列表框中成对选入 V1SBP - V1DBP、V2SBP - V2DBP。

(6) 单击“确定”按钮。

由图 19.10 可见,A、B 两药的收缩压/舒张压均较试验前明显下降, $P < 0.001$ 。A 药的收缩压下降 12.2mmHg,而 B 药下降幅度为 19.3mmHg。两组的舒张压下降幅度则分别为 7.3 和 7.7mmHg,相差不大。

		成对差分								
		均值	标准差	均值的 标准误	差分的 95% 置信区间					Sig. (双侧)
keys					下限	上限	t	df		
A	对 1 V1SBP – V2SBP	17.963	12.242	1.352	15.274	20.653	13.288	81		.000
	对 2 V1DBP – V2DBP	13.146	7.293	.805	11.544	14.749	16.323	81		.000
B	对 1 V1SBP – V2SBP	19.338	16.495	1.880	15.594	23.082	10.287	76		.000
	对 2 V1DBP – V2DBP	15.429	7.752	.883	13.669	17.188	17.465	76		.000

图 19.10 成对样本检验

随后将进一步比较两组的血压下降程度有无差异,需要首先将差值变量计算出来。

COMP SBPMIN = V2SBP - V1SBP.

COMP DBPMIN = V2DBP - V1DBP.

EXEC.

然后使用差值变量进行两组间的 t 检验(首先去除上面的文件拆分状态)。

结果如图 19.11 所示,两组治疗前后的收缩压/舒张压下降幅度组间比较无统计学差异, P 值均大于 0.05。

		方差方程的 Levene 检验		均值方程的 t 检验				均值差值	标准误差值
		F	Sig.	t	df	Sig. (双侧)			
sbpmin	假设方差相等	4.446	.037	.599	157	.550	1.37425	2.29432	
	假设方差不相等			.594	139.844	.554	1.37425	2.31545	
dbpmin	假设方差相等	.564	.454	1.913	157	.058	2.28223	1.19312	
	假设方差不相等			1.909	154.618	.058	2.28223	1.19542	

图 19.11 独立样本检验

3. 血压正常化率

两组的血压正常化率,组间卡方比较 $P > 0.05$,无统计学差异。

19.1.6 安全性评价

1. 实验室检查

实验室检查涉及血常规、尿常规、血生化三大类指标。这里只以红细胞、总蛋白为例给出分析结果,如图 19.12 所示。

假设检验汇总

	原假设	测试	Sig.	决策者
1	1 红细胞的分布为正态分布，平均值为4.48，标准差为0.51。	单样本1Kolmogorov-Smirnov 检验	.819	保留原假设。
2	2 红细胞的分布为正态分布，平均值为4.74，标准差为0.45。	单样本1Kolmogorov-Smirnov 检验	.776	保留原假设。
3	1 总蛋白的分布为正态分布，平均值为67.71，标准差为20.22。	单样本1Kolmogorov-Smirnov 检验	.000	拒绝原假设。
4	2 总蛋白的分布为正态分布，平均值为68.73，标准差为19.33。	单样本1Kolmogorov-Smirnov 检验	.000	拒绝原假设。

显示渐进显著性。显著性水平是 .05。

图 19.12 实验室指标的正态分布检验结果

K-S 检验结果显示,试验前后的红细胞数分布符合正态分布,但总蛋白数均不符合正态分布。因此前者将使用 *t* 检验进行组间比较,而总蛋白数则将使用秩和检验进行组间比较。分析结果如图 19.13 所示,试验前、后的红细胞数在两组间都是无统计学差异的。

		方差方程的 Levene 检验		均值方程的 t 检验			
		F	Sig.	t	df	Sig. (双侧)	均值差值
1 红细胞	假设方差相等	.177	.674	-1.235	156	.219	-.09913
	假设方差不相等			-1.236	155.888	.218	-.09913
2 红细胞	假设方差相等	.542	.463	.775	149	.439	.05698
	假设方差不相等			.773	145.912	.441	.05698

图 19.13 独立样本检验

- (1) 选择“分析”→“非参数检验”→“独立样本”菜单项。
- (2) 在打开的对话框中选择“字段”选项卡,将 1 总蛋白、2 总蛋白选入“检验变量”列表框中,将“keys”选项选入“组变量”列表框中。
- (3) 选择“设置”选项卡,选择“自定义检验”选项组中的第一项“M-H 检验”,该方法的结果等价于所需的秩和检验。
- (4) 单击“运行”按钮。

分析结果如图 19.14 所示,同样显示试验前、后的总蛋白红细胞数在两组间都是无统计学差异的。

最终对全部血常规、尿常规、血生化三大类指标的分析结果显示,绝大部分指标在试验前后均无统计学差异。极个别指标虽然存在试验后升高或下降的统计学意义,但变化幅度均无临床意义。此外,各指标的组间比较也均无统计学差异。

2. 不良事件和不良反应

如图 19.15 所示,A 药物组不良事件发生率为 35.4% (29/82 例),B 药物组为 49.4% (38/77 例),试验过程中无严重不良事件发生。

假设检验汇总

	原假设	测试	Sig.	决策者
1	1 总蛋白的分布在 keys 类别上相同。	独立样本 Mann-Whitney U 检验	.288	保留原假设。
2	2 总蛋白的分布在 keys 类别上相同。	独立样本 Mann-Whitney U 检验	.246	保留原假设。

显示渐进显著性。显著性水平是 .05。

图 19.14 总蛋白的组间秩和检验

		不良事件		合计
		0	1	
keys	A	计数	53	82
		keys 中的 %	64.6%	100.0%
	B	计数	39	77
		keys 中的 %	50.6%	100.0%
合计		计数	92	159
		keys 中的 %	57.9%	100.0%

图 19.15 keys * 不良事件交叉制表

对两组的不良事件率进行组间比较,如图 19.16 所示, $P>0.05$,无统计学意义。

	值	df	渐进 Sig. (双侧)	精确 Sig. (双侧)	精确 Sig. (单侧)
Pearson 卡方	3.185 ^a	1	.074		
连续校正 ^b	2.638	1	.104		
似然比	3.194	1	.074		
Fisher 的精确检验				.080	.052
有效案例中的 N	159				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 32.45。

b. 仅对 2×2 表计算。

图 19.16 卡方检验

对不良反应的分析本案例略。

19.1.7 分析结论与总结

1. 研究结论

(1) 每日口服 A 药物 4~8mg 一次能有效地降低中国人轻中度原发性高血压患者的坐位舒张压和收缩压,治疗 8 周的总有效率为 80.5% (66/82 例),其降压作用与 B 药物(88.3%,68/77 例)相比无明显差别。两组的血压下降幅度、血压正常化率等也无差别。

(2) 两种药物每日服药一次安全性和耐受性较好。A 药物组不良事件发生率为 35.4% (29/82 例),B 药物组为 49.4% (38/77 例),无统计学差异,在试验过程中无严重不良事件发生。

最终在盲底揭盲之后就可以知道,A、B两组所对应的药物究竟是试验的X药物,还是对照用的苯磺酸氨氯地平,并形成最终的X药物临床研究总报告了。

2. 项目总结

对于不熟悉临床试验项目的读者而言,本案例在许多方面都与众不同。

(1) 数据安全性和保密性在项目中非常关键,不仅真实的药物名称不会在分析中出现,而且所有原始数据被分成多个单独的数据文件进行管理、保存(本案例已经将其简化为只有两个)。

(2) 虽然是多中心研究,而且也涉及很多变量,但是在分析计划中完全以最简单的 t 检验、方差分析、秩和检验为主,在整个统计分析计划中根本就没有考虑过多因素模型,更不要说考虑随机效应的GEE、混合效应模型等了。

(3) 整个研究流程异常死板,以连续性变量为例,完全按照是否符合正态分布评判,如果检验拒绝正态分布假设,则必须采用秩和检验这种严格规定的流程来执行,不存在任何例外。

实际上,这正是临床试验这一特殊分析类型的特点所在,出于确保病人利益、杜绝方法体系上一切可能的漏洞的原因,整个临床试验统计分析体系都是尽量采用保守、稳妥的思路来构建的,只要是能够采用简单方法解决的问题,就必然不会考虑比较复杂的方法。在大家深入了解了这一行业的特点之后,就会真正明白上述做法的原因所在。



事实上,上述特点恰恰也就是笔者本人并不喜欢临床试验统计分析的重要原因,笔者更喜欢能够自由发挥的、有创造性和挑战性的分析项目,而做临床试验的统计分析虽然收入丰厚,但统计师在其中只是一个法规条例的执行者而已,会缺少很多统计原本应有的乐趣。

19.2 咖啡屋需求调查案例

19.2.1 项目背景

1. 研究目的

2003年,受毕业校友的委托,北京大学的几位在读研究生在校内进行了一次关于北京大学师生对咖啡屋及类似休闲场所的需求调查,以便对这些校友的创业决策(在北京大学校内开设一家咖啡屋)提供数据支持。

具体而言,本研究的需求如下。

- (1) 了解北京大学校内咖啡消费人群的基本背景状况。
- (2) 了解该消费人群的咖啡消费习惯,包括频次、额度、消费原因等。
- (3) 了解该消费人群可能存在,但目前尚未被满足的潜在需求。

2. 研究问卷

该调查共收集了302位受访者的回答,具体的问卷如下。

北京大学师生对咖啡屋及类似休闲场所的需求调查

第一部分:甄别问卷

F 您是否在过去的一年中去过咖啡店或类似的休闲场所?

1. 是 2. 否(跳至 Q9)

第二部分:主体问卷

Q1 以下休闲吧您光顾最频繁的是:

1. 星巴克 2. 仙踪林 3. 真锅咖啡 4. 雕刻时光 5. 绿叶谷 6. 师生缘
7. 勺园咖啡屋 8. 西门外酒吧 9. 闲情偶寄 10. 其他 _____

Q2 以下休闲吧您最喜欢的是:

1. 星巴克 2. 仙踪林 3. 真锅咖啡 4. 雕刻时光 5. 绿叶谷 6. 师生缘
7. 勺园咖啡屋 8. 西门外酒吧 9. 闲情偶寄 10. 其他 _____

Q3 您喜欢的原因是出于(多选,三项以内):

1. 那里有我最喜欢的饮料; 2. 我喜欢那里的情调和环境;
3. 那里的价格很公道; 4. 因为朋友喜欢,我就一起去了;
5. 因为离得近,方便; 6. 其他 _____

Q4 您去咖啡屋或休闲吧的主要目的是(多选,三项以内):

1. 喝喜欢的东西 2. 与朋友聊天 3. 自习或一个人看东西
4. 讨论案例或公事 5. 约会 6. 其他 _____

Q5 您去咖啡屋或休闲吧主要消费的是(多选):

1. 咖啡 2. 奶茶 3. 啤酒 4. 冰淇淋 5. 碳酸饮料 6. 果汁
7. 牛奶 8. 茶 9. 矿泉水 10. 爆米花 11. 面包小点 12. 薯条
13. 沙拉 14. 套餐 15. 其他 _____

Q6 您去咖啡屋或休闲吧平均每次的花费大约是(人均):

1. 20元以下 2. 20~39元 3. 40~59元 4. 60元及以上

Q7 您去咖啡屋或休闲吧平均每次停留的时间大约是:

1. 1小时以下 2. 1~2小时 3. 2~3小时 4. 3小时以上

Q8 一般来说,您得知学校附近开新店的消息通过的途径是(多选):

1. 路过看到 2. 朋友介绍 3. 校内海报 4. 网上广告 5. 校内 BBS 6. 其他

Q9 您觉得在校内开咖啡店的理想位置是:

1. 三角地 2. 学生宿舍楼区 3. 勺园周围 4. 理教·光华一带
5. 一教及图书馆一带 6. 三教、四教一带 7. 其他 _____

第三部分:个人信息

P1 性别:1. 男 2. 女

P2 年龄: _____

P3 您是:1. 本科 2. 研究生 3. MBA 学生 4. 博士生 5. 进修生

6. 教师 7. 留学生

P4 可支配的月收入:(人民币)

1. 500 元以下
2. 500 ~ 999 元
3. 1 000 ~ 2 999 元
4. 3 000 ~ 4 999 元
5. 5 000 元以上



为了便于讲解,问卷及原始数据均有所修改,以简化结果输出。

最终整理完毕的数据见文件 coffee.sav。

19.2.2 数据预分析

因本案例结构非常清楚,因此省略分析思路的讨论,直接开始数据预分析。这里需要了解一下受访者人群的人口背景特征,由于样本可以被分为过去是否去过咖啡店的两类人群,因此直接进行交叉表分析,结果如图 19.17 所示。



特别需要注意的是交叉表的总样本量,如果小于 302 例,则说明相应变量中存在缺失值,而这些含缺失值的案例是会被直接剔除出交叉表分析的。

			性别		合计
			男	女	
过去是否去过	是	计数	151	101	252
		过去是否去过中的 %	59.9%	40.1%	100.0%
	否	计数	39	11	50
		过去是否去过中的 %	78.0%	22.0%	100.0%
合计		计数	190	112	302
		过去是否去过中的 %	62.9%	37.1%	100.0%

图 19.17 过去是否去过 * 性别交叉制表

本研究中的受访者以男性为主,占 63%,由于性别比例不平衡,因此需要注意性别是否会对某些题目的答案有影响,以免数据解释错误。

此外,卡方检验的结果如图 19.18 所示,其中显示女性去过咖啡消费场所的比例要更高一些,且差异具有统计学意义,这或许提示以女性为新店客源的突破口比较可行。

	值	df	渐进 Sig. (双侧)	精确 Sig. (双侧)	精确 Sig. (单侧)
Pearson 卡方	5.845 ^a	1	.016		
连续校正 ^b	5.096	1	.024		
似然比	6.235	1	.013		
Fisher 的精确检验				.016	.010
线性和线性组合	5.825	1	.016		
有效案例中的 N	302				

a. 0 单元格(.0%) 的期望计数少于 5。最小期望计数为 18.54。

b. 仅对 2 × 2 表计算。

图 19.18 卡方检验



为了节约篇幅,本案例中使用的卡方检验、制表模块等的具体操作均不再列出,且交叉表所对应的卡方检验结果也不再列出,后续的交叉表除非特别指明,否则其卡方检验 P 值均小于 0.05。

如图 19.19 所示,无论过去是否去过咖啡店,两群体的平均水平和年龄分布基本相同。

		年龄						极大值
		均值	极小值	百分位 25	中值	百分位 75	百分位 95	
过去是否去过	是	25	18	22	24	28	33	45
	否	25	16	22	24	27	32	37

图 19.19 年龄的表格描述

如图 19.20 所示,受访者中本科、研究生一共占一半以上,此外 MBA/博士总共占 30%,进修生和留学生合计占 10%,因此对于分析结果还是首先考虑本科/硕士生的需求,MBA/博士由于经济状况、年龄等和本科/研究生相差会较大,因此作为次要研究人群考虑。

			身份						
			本科	研究生	MBA	博士	进修生	留学生	合计
过去是否去过	是	计数	64	81	38	34	7	28	252
		过去是否去过中的 %	25.4%	32.1%	15.1%	13.5%	2.8%	11.1%	100.0%
	否	计数	19	11	8	12	0	0	50
		过去是否去过中的 %	38.0%	22.0%	16.0%	24.0%	.0%	.0%	100.0%
合计		计数	83	92	46	46	7	28	302
		过去是否去过中的 %	27.5%	30.5%	15.2%	15.2%	2.3%	9.3%	100.0%

图 19.20 过去是否去过 * 身份交叉制表

如图 19.21 所示,受访者收入以 3 000 元以下为主,特别是 1 000 元以下的占了 2/3,非常符合学生的特征。收入的分布在去过/未去过咖啡吧的人群间无统计学差异(通过秩和检验得知)。

		可支配月收入						
		500 元以下	500 ~ 999 元	1 000 ~ 2 999 元	3 000 ~ 4 999 元	5 000 以上	合计	
过去是否去过	是	计数	63	100	55	9	21	248
		%	25.4%	40.3%	22.2%	3.6%	8.5%	100.0%
	否	计数	18	17	8	3	3	49
		%	36.7%	34.7%	16.3%	6.1%	6.1%	100.0%
合计	计数	81	117	63	12	24	297	
	%	27.3%	39.4%	21.2%	4.0%	8.1%	100.0%	

图 19.21 过去是否去过 * 可支配月收入交叉制表

综合上述对人口背景资料的分析,可以得到如下线索。

- (1) 整个研究接触到的核心人群应当就是本科/硕士在读学生,在抽样合理的情况下,这也应当是主要的咖啡消费人群。
- (2) 需要注意性别间可能存在的差异。

19.2.3 主体问卷分析

1. 受访者对现有酒吧的 U&A

下面开始就主体问卷中的题目进行分析,首先是对光顾频次和咖啡店偏好情况的交叉分析。如图 19.22 所示,从中可以看出。

计数

		最喜欢							
		星巴克	仙踪林	雕刻时光	师生缘	勺园咖啡屋	西门外酒吧	其他	合计
最频繁	星巴克	26	2	5	1	0	0	6	40
	仙踪林	2	17	5	0	0	0	1	25
	雕刻时光	3	1	19	1	0	0	0	24
	师生缘	13	9	10	27	3	2	16	80
	勺园咖啡屋	5	2	1	0	5	0	1	14
	西门外酒吧	1	2	6	0	0	17	4	30
	其他	3	2	4	0	1	0	23	33
合计		53	35	50	29	9	19	51	246

图 19.22 最频繁 * 最喜欢交叉制表

- (1) 在消费的频繁程度上,师生缘明显具有优势,处于第一集团,其次为星巴克,和其余竞争对手相比也明显具有优势。
- (2) 将消费的频繁程度和最受欢迎的程度交叉起来会发现,师生缘其实并不是最受欢迎的,星巴克、雕刻时光的受欢迎程度不相上下,均明显优于其他竞争对手。
- 上述结果带来两个需要进一步考察的问题。
- (1) 首先,师生缘消费频繁程度明显高于其受欢迎程度的表现,是由于价格还是风格等原因?
- (2) 其次,雕刻时光的受欢迎程度为什么无法转换为其实际消费行为?
- 下面进一步结合多选题 Q3 的答案来对上述问题进行解答。



图 19.23 是使用制表模块生成的,制表时需要注意将 Q1、Q2 等变量的测量尺度指定为正确的名义尺度,而不是直接使用数据中默认的度量尺度,否则将无法得到所需的表格,具体操作参见第 9 章。

由图 19.23 所示的结果可以看出:

- (1) 平均而言,受访者去咖啡吧最看重的就是情调和环境,平均提及率高达 70%。
- (2) 师生缘最大的优势就是距离近,得分较高的情调和环境实际上就已经低于均值,至于其他几项的提及率就更差。

		有喜欢的饮料	情调和环境	价格公道	因为朋友喜欢	距离近	其他原因
最喜欢	星巴克	22.6%	81.1%	9.4%	22.6%	3.8%	3.8%
	仙踪林	8.8%	82.4%	2.9%	20.6%	11.8%	8.8%
	雕刻时光	8.0%	82.0%	16.0%	24.0%	8.0%	2.0%
	师生缘	3.4%	51.7%	10.3%	17.2%	65.5%	.0%
	勺园咖啡屋	.0%	55.6%	11.1%	22.2%	77.8%	.0%
	西门外酒吧	15.8%	57.9%	21.1%	36.8%	36.8%	10.5%
	其他	14.3%	55.1%	30.6%	12.2%	40.8%	10.2%
	总计	12.3%	70.0%	15.2%	21.0%	25.9%	5.3%

图 19.23 最喜欢的原因

(3) 星巴克和雕刻时光在情调和环境这一项上都得分颇高,但在距离上明显不占优势,导致了其受欢迎程度无法充分转换为其实际消费行为。



多选题的交叉表也是可以进行假设检验的,但方法比较麻烦(标准分析模型应当是多水平模型),在大多数情况下,在探索性的调研项目中,通过数据描述来探知多选题中可能包括的趋势就足够了。

上述分析结果已经很清楚地提示距离远近是一个重要的影响因素,下面再利用 Q5 的结果来进一步剖析数据。

由图 19.24 所示的结果可以看出。

		喝喜欢的东西	与朋友聊天	自习或一个人看东西	讨论案例或公事	约会	其他目的
最频繁	星巴克	22.0%	80.5%	24.4%	26.8%	7.3%	7.3%
	仙踪林	12.0%	76.0%	16.0%	8.0%	56.0%	4.0%
	雕刻时光	4.2%	79.2%	25.0%	8.3%	16.7%	4.2%
	师生缘	4.8%	83.1%	8.4%	19.3%	24.1%	4.8%
	勺园咖啡屋	14.3%	100.0%	7.1%	28.6%	35.7%	.0%
	西门外酒吧	16.1%	87.1%	16.1%	6.5%	19.4%	9.7%
	其他	9.1%	90.9%	9.1%	9.1%	15.2%	6.1%
	总计	10.8%	84.1%	14.3%	15.9%	22.7%	5.6%

图 19.24 去咖啡店的主要目的

(1) 星巴克既适合于独坐,也适合于分享,但不适合于恋爱约会。

(2) 星巴克的另一个优势是饮料品种/口味更受欢迎,这方面仅有西门外酒吧的提及率与其相近。

(3) 相比之下,仙踪林是比较受恋人青睐的地方。

(4) 师生缘最合适的是好友同行,但实际上该项提及率还低于平均水平,总体而言看不出有明显优势。

上述结果进一步确认了前面的发现:距离足够近是消费频率的关键因素,所谓一俊遮百丑,师生缘虽然各项指标都不出色,但距离近就使得它成了为最常光顾的酒吧。

2. 受访者在酒吧消费的情况

下面再进一步考察受访者的具体消费情况,如图 19.25 所示。

		咖啡	奶茶	啤酒	冰激凌	碳酸饮料	果汁	茶	爆米花	薯条
最频繁	星巴克	85.4%	14.6%	14.6%	19.5%	14.6%	24.4%	26.8%	4.9%	2.4%
	仙踪林	52.0%	48.0%	16.0%	32.0%	12.0%	52.0%	20.0%	8.0%	20.0%
	雕刻时光	62.5%	25.0%	20.8%	16.7%	8.3%	33.3%	29.2%	25.0%	16.7%
	师生缘	38.1%	22.6%	26.2%	16.7%	14.3%	35.7%	21.4%	21.4%	25.0%
	勺园咖啡屋	53.8%	15.4%	15.4%	30.8%	7.7%	38.5%	30.8%	15.4%	23.1%
	西门外酒吧	32.3%	9.7%	61.3%	9.7%	25.8%	16.1%	16.1%	22.6%	22.6%
	其他	57.6%	21.2%	30.3%	27.3%	9.1%	27.3%	21.2%	33.3%	39.4%
	总计	52.2%	21.9%	27.1%	19.9%	13.9%	31.9%	22.7%	19.1%	21.5%

图 19.25 去咖啡店消费的主要饮料/食品种类

为了节省版面,图中已经删除了提及过少的饮料/食品种类,从图 19.25 中会发现一些很有趣的信息。

- (1) 星巴克已经牢牢占据了正宗咖啡的形象阵地,在咖啡的消费比例上星巴克非常高。
- (2) 仙踪林则是以奶茶、果汁、冰激凌的消费为主,看来这两样比较适合于和恋人同行时饮用,而且受访者实际上并未将该店定位为咖啡吧。换言之,北京大学的恋人们是不会去咖啡吧这类地方谈恋爱的。
- (3) 师生缘又一次走了中庸路线,没有发现他的消费人群更偏向于消费哪种饮料/食品。
- (4) 在西门外酒吧消费啤酒和碳酸饮料的比例很高,这应当是一个很合理的结果。

下面来考察人均花费的情况,如图 19.26 所示。

		人均花费			
		20 元以下	20 ~ 39 元	40 ~ 59 元	60 元以上
最频繁	星巴克	14.6%	39.0%	34.1%	12.2%
	仙踪林	4.0%	36.0%	32.0%	28.0%
	雕刻时光	8.3%	54.2%	29.2%	8.3%
	师生缘	19.0%	38.1%	28.6%	14.3%
	勺园咖啡屋	7.1%	64.3%	21.4%	7.1%
	西门外酒吧	6.5%	35.5%	35.5%	22.6%
	其他	12.1%	30.3%	33.3%	24.2%
	合计	12.7%	39.7%	31.0%	16.7%

图 19.26 人均花费情况

图 19.26 反映出咖啡吧的主要消费情况是人均 20 ~ 60 元。虽然从样本数据描述来看去雕刻时光、勺园的花费偏低,而去星巴克、仙踪林、西门外酒吧的则偏高,但秩和检验无统计学差异。

3. 酒吧/咖啡吧相关的信息来源

下面来进一步考察信息来源,如图 19.27 所示。

		路过看到	朋友介绍	校内海报	网上广告	校内 BBS	其他信息渠道
过去是否去过	是	36.5%	63.9%	38.6%	2.0%	14.5%	.8%
	否	37.5%	50.0%	45.8%	4.2%	14.6%	2.1%
最频繁	星巴克	50.0%	65.0%	40.0%	2.5%	7.5%	.0%
	仙踪林	32.0%	52.0%	36.0%	4.0%	24.0%	.0%
	雕刻时光	56.5%	47.8%	26.1%	.0%	13.0%	4.3%
	师生缘	28.9%	59.0%	41.0%	2.4%	13.3%	1.2%
	勺园咖啡屋	21.4%	78.6%	35.7%	.0%	14.3%	.0%
	西门外酒吧	41.9%	80.6%	38.7%	3.2%	22.6%	.0%
	其他	30.3%	72.7%	42.4%	.0%	12.1%	.0%

图 19.27 主要信息来源

从结果图 19.27 中可以看出：

- (1) 受访者对此类场所的了解主要还是通过路遇/朋友介绍/海报等传统方式,比较新的网上广告/BBS 所占比例并不高。
 - (2) 对于恋爱人群和酒吧人群而言,校内 BBS 是一个可能有价值的推广渠道。
 - (3) 无论过去是否去过酒吧,其信息来源渠道是非常接近的。
- 选址的结果如图 19.28 所示。


过去是否去过中的 %		选址								
		三角地	宿舍区	勺园	理教 / 光华	一教 / 图书馆	三教 四教	其他	未名湖	合计
过去是	是	9.2%	33.1%	16.3%	17.9%	11.6%	1.2%	5.6%	5.2%	100.0%
否去过	否	20.8%	43.8%	4.2%	10.4%	14.6%	2.1%	2.1%	2.1%	100.0%
合计		11.0%	34.8%	14.4%	16.7%	12.0%	1.3%	5.0%	4.7%	100.0%

图 19.28 过去是否去过 * 选址交叉制表

从图 19.28 中可见,有过咖啡吧消费经验的人非常倾向于新店开在宿舍区,而没有消费经验的人群会同时考虑三角地,但宿舍区仍然是其首选。

4. 加入背景资料进行结果验证

下面考虑将上面得到的线索和背景资料结合起来重新分析,以进一步验证该结果的真实性。这里为了节省篇幅,只给出比较重要的几个表格。

 背景资料和主体题目的交叉分析一般是融合在主体问卷分析中一并完成的,本案例为了节省篇幅,同时也为了讲解更为清晰,将其放在最后作为用于结果的核查/补充。

如图 19.29 所示,从样本数据看,似乎女性更喜欢星巴克,更少去师生缘,但卡方检验 P 值为 0.1,差异尚无统计学意义。

性别中的 %

		最频繁							合计
		星巴克	仙踪林	雕刻时光	师生缘	勺园咖啡屋	西门外酒吧	其他	
性别	男	11.9%	9.9%	9.3%	35.1%	4.0%	15.9%	13.9%	100.0%
	女	22.8%	9.9%	9.9%	30.7%	7.9%	6.9%	11.9%	100.0%
	合计	16.3%	9.9%	9.5%	33.3%	5.6%	12.3%	13.1%	100.0%

图 19.29 性别 * 最频繁交叉制表

如图 19.30 所示,其中的数据显示,相对而言,研究生/博士生更喜欢雕刻时光和师生缘,显然去前者的是恋爱人群,去后者的则是大众人群。而 MBA、留学生更倾向于去星巴克这类“正宗”的地方消费。

身份中的 %

		最喜欢							合计
		星巴克	仙踪林	雕刻时光	师生缘	勺园咖啡屋	西门外酒吧	其他	
身份	本科	21.9%	17.2%	26.6%	10.9%	1.6%	12.5%	9.4%	100.0%
	研究生	15.0%	13.8%	26.3%	16.3%	3.8%	3.8%	21.3%	100.0%
	MBA	36.8%	15.8%	7.9%	10.5%	2.6%	2.6%	23.7%	100.0%
	博士	6.3%	9.4%	25.0%	12.5%		3.1%	43.8%	100.0%
	进修生	33.3%	16.7%			16.7%	16.7%	16.7%	100.0%
	留学生	34.6%	11.5%	3.8%	3.8%	11.5%	19.2%	15.4%	100.0%
	合计	21.5%	14.2%	20.3%	11.8%	3.7%	7.7%	20.7%	100.0%

图 19.30 交叉表

如图 19.31 所示,样本数据显示,月收入 1 000 元以下人群的消费行为明显集中在师生缘,随着收入的增加,受访者对星巴克和西门外酒吧的兴趣似乎也在增加,但是上述趋势在假设检验中均无统计学意义。

可支配月收入中的%

		最频繁							合计
		星巴克	仙踪林	雕刻时光	师生缘	勺园咖啡屋	西门外酒吧	其他	
可支配月收入	500 元以下	7.9%	7.9%	12.7%	34.9%	6.3%	12.7%	17.5%	100.0%
	500 ~ 999 元	15.0%	8.0%	11.0%	38.0%	4.0%	11.0%	13.0%	100.0%
	1 000 ~ 2 999 元	20.0%	12.7%	3.6%	29.1%	7.3%	16.4%	10.9%	100.0%
	3 000 ~ 4 999 元	11.1%	11.1%	11.1%	22.2%	11.1%	22.2%	11.1%	100.0%
	5 000 以上	33.3%	14.3%	9.5%	23.8%	4.8%	4.8%	9.5%	100.0%
	合计	15.7%	9.7%	9.7%	33.5%	5.6%	12.5%	13.3%	100.0%

图 19.31 可支配月收入 * 最频繁交叉制表

相比之下,女性虽然去过咖啡吧的比例较高,但消费额度却比男性低,集中在20~39元范围内,如图19.32所示。

性别中的%

		人均花费				合计
		20元以下	20~39元	40~59元	60元以上	
性别	男	12.6%	33.1%	31.8%	22.5%	100.0%
	女	12.9%	49.5%	29.7%	7.9%	100.0%
合计		12.7%	39.7%	31.0%	16.7%	100.0%

图19.32 性别*人均花费交叉制表

19.2.4 项目总结与讨论

根据前几节的分析,可以得到如下研究结论。

(1) 校园咖啡吧的消费人群应当以本科/硕士学历、月收入1000元以下的人群为主。


(2) 校园咖啡吧大致有两种设计思路:便捷性为主(如师生缘),或者有突出的特色(如雕刻时光或者西门外酒吧),但是前者显然更贴合消费人群的需求。

(3) 主要消费人群的消费额度是人均30~60元的范围,相对而言,咖啡吧所提供的食品种类及特色并不重要,控制总价范围,或者说提供视频之外的消费选择可能更为重要。

(4) 咖啡吧选址应当尽量考虑宿舍区这种便捷性场所。

(5) 如果不是特殊定位,那么网络、BBS不是特别重要的宣传渠道。

(6) 在开业初期,女性群体可以作为首批推广的主要对象。

 读者可以看到,前面十几页的分析最终被汇总为很简单的几句话结论。实际上,对于一个真实的分析项目,在研究目标的指导下,分析员应当对可能涉及的分析维度尽可能加以考察,但是并不是所有的分析结果都需要在报告里呈现。能把100页报告的内容浓缩成两三页的分析人员,才是真正的大师,而篇幅只有一页的报告才是老总想看的東西。

在本研究中并未使用过于复杂的分析方法,主要是通过单选题交叉表、多选题交叉表,外加相应的卡方检验来对数据进行分析的。虽然类似于交叉表也可以用对应分析图来做到更好的呈现,但显然这些方法的应用只是锦上添花。只要能够解决实际问题,那么简单的方法就应当被优先考虑,比较复杂(往往也更加难懂)的方法只应当在需要的时候才加以应用。

19.3 牙膏新品购买倾向研究案例

19.3.1 研究背景

国内的牙膏市场在20世纪以来处于相对稳定的状态,高露洁、佳洁士等几大国际品牌占据了主要的市场份额,而蓝天六必治、两面针等品牌则因其某些方面的优势占据着某个细分市场。但是,即使是相对稳定的市场,用于老产品升级换代的新品的推出也是一件非常慎重的事情,错

误的产品投放决策完全可能导致原本稳固的市场份额出现流失。因此,在上市前对新品受消费者的欢迎程度,特别是具体的市场定位进行研究,就成为非常重要的一个决策依据。

在2003年,受客户的委托,我们对某牙膏新品的市场潜力进行了一次研究,研究目的非常明确,有以下几个。

- (1) 考察该牙膏新品的市场欢迎程度是否达到预期。
 - (2) 受访者对该牙膏新品的评价是否能超过现有市场品牌。
 - (3) 受访者对该新品的评价受到哪些因素的影响,是否存在比较合适进入的细分市场。
- 在具体的研究设计方面,和本案例有关的内容如下。

(1) 核心评判指标:本研究中的核心评价指标为对未来购买该新品的倾向性评分,按照1~10分设定,10分表示一定会购买,1分则表示一定不会购买。

(2) 考虑人口背景变量的影响:不同受访者所在的城市、性别、年龄、收入等人口背景资料显然应当被纳入分析范围,对其可能的影响进行分析。

(3) 考虑卫生习惯的影响:受访者的日常卫生习惯(如每日刷牙次数等)也应当是重要的潜在影响因素。

(4) 现在使用牙膏品牌的影响:不同品牌的牙膏实际上占据的是不同的细分市场,而新上市产品的细分市场定位是否正确,将会直接影响其上市是否成功,因此受访者目前最常使用的牙膏品牌也将是重要的潜在影响因素。

考虑到本项目的复杂性,这里只提取了年龄、目前最常用的牙膏品牌、购买倾向性这3个变量用于案例分析,分析目的如下。

- (1) 考察年龄、目前最常用的牙膏品牌这两个变量是否对购买倾向性有影响。
- (2) 如果有影响,则给出不同状况下的购买倾向性评分估计值。

本案例的数据见文件牙膏新品研究.sav。

19.3.2 分析思路

在有针对性的研究设计框架之下,本项目的数据分析实际上是比较简单的。

(1) 本研究中所关心的结局变量为购买倾向性评分,取值为1~10,由于范围较宽,因此可以直接按照连续性变量加以分析(为稳妥起见,最好列出频数表确认实际取值范围)。

(2) 由于该评分是从每一位受访者询问而来的,因此研究中的基本观察单位就是受访者。

(3) 在本案例中需要考虑的潜在影响因素有两个,最常用的牙膏品牌共分为6个水平,如果只分析该变量的作用,则分析目的就是考察这6组人群的平均购买倾向性评分有无差异,可以考虑使用单因素方差分析模型进行数据分析,即将牙膏品牌作为模型中的影响因素,考察它对评分有无影响。

(4) 另一个潜在影响因素为年龄,如果只分析该变量的作用,则可以考虑采用相关分析或者回归分析。

(5) 如果同时考虑上述两个影响因素的作用,则必须建立一个多变量分析模型,最基本的模型框架是一般线性模型,然后在该模型的架构下进行不同影响因素组合之下更准确的购买倾向均数(即边际期望均数)估计。

(6) 在实际分析中,没有必要直接去建立多变量模型,而应当先逐个进行变量的筛选和数据

理解,在了解到足够的信息之后再再来建立复杂模型。

下面将按此思路进行分析。

19.3.3 数据预分析

1. 因变量的统计描述

模型中的因变量为分值,属于连续性变量,对连续性变量可以考虑使用描述过程进行简单快速的统计描述,操作步骤如下。

- (1) 选择“分析”→“描述统计”→“描述”菜单项,打开“描述”对话框。
- (2) 在“变量”列表框中选入“上市后购买指数”。
- (3) 单击“确定”按钮。

图 19.33 反映出购买指数的数值分布介于 1~10 之间,因此分布范围比较理想,且标准差大小只有均数的 1/3 左右,因此大致可以看出不存在明显的极端值/偏态分布。当然,对此问题更好地观察方法是使用图形工具,在本例中条图或者直方图都是很好的选择,这里采用前者来加以考察。

	N	极小值	极大值	均值	标准差
上市后购买指数	484	1	10	6.37	2.088
有效的 N (列表状态)	484				

图 19.33 描述统计量



如果该变量的取值范围较小,例如实际取值范围只在 3~7 之间,则按照有序分类变量来加以分析可能更为合理。

- (1) 选择“图形”→“图表构建程序”,打开“图表构建程序”对话框。
- (2) 在图库中选择“条”图组,将简单条图图标拖入画布中。
- (3) 在变量“上市后购买指数”处右击,将测量尺度改为“序号”。
- (4) 将“上市后购买指数”拖入 X 轴框。
- (5) 单击“确定”按钮。

注意在上述操作中,必须先将上市后购买指数的测量尺度改为“序号”或者“名义”,否则按照默认的度量方式,系统仍然自动切换为绘制直方图。从图 19.34 中可以看出,虽然左侧 1、2 的取值频数略少,似乎略呈偏态,但不是非常强烈,而且也不存在明显的极端值(因取值范围所限),在实际数据中应当是比较好的分布情况了。



条图严格来说应当被用于分类变量,这里由于指数变量只有 10 种分数取值,使用条图反而比直方图更容易进行原始数据的观察,因此这里进行了工具的活用。

2. 分类自变量的统计描述

首先考察最常使用的品牌的分布情况,这可以采用频率过程来实现。

- (1) 选择“分析”→“描述统计”→“频率”菜单项,打开“频率”对话框。
- (2) 在“变量”列表框中选入最常使用的品牌。
- (3) 单击“确定”按钮。

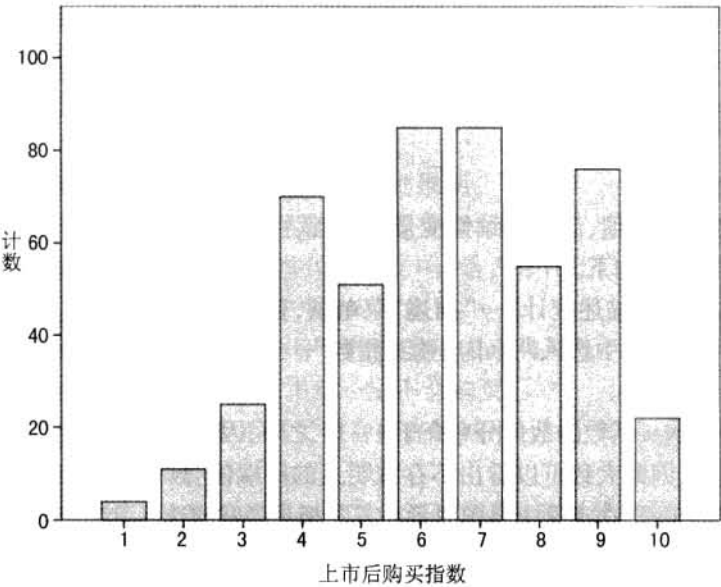


图 19.34 购买指数的条图

图 19.35 不进行过多解释,依次为频数、百分比、有效百分比、累积百分比的数值。由于在设计中对不同牙膏品牌根据其市场表现进行了配额,因此高*和佳*的样本量大约是其余几个品牌的两倍。这里的有效百分比指的是去除掉缺失样本后,各类别在有效样本中所占的比例,表格最后列出有 2 位受访者的该变量缺失,因此百分比和有效百分比的数值也有所差异。

		频数	百分比	有效百分比	累积百分比
有效	高*	108	22.3	22.4	22.4
	佳*	105	21.7	21.8	44.2
	黑*	51	10.5	10.6	54.8
	中*	56	11.6	11.6	66.4
	蓝*	47	9.7	9.8	76.1
	其他	115	23.8	23.9	100.0
	合计	482	99.6	100.0	
缺失	系统	2	.4		
合计		484	100.0		

图 19.35 最常使用的品牌

对于品牌,也可以使用条图进行图形描述,留给读者自行操作,此处略去。

3. 连续自变量的统计描述

下面考虑对年龄进行描述,可以使用探索过程做一个全面的统计描述。

- (1) 选择“分析”→“描述统计”→“探索”菜单项,打开“探索”对话框。
- (2) 在“因变量”列表框中选中“年龄”。
- (3) 单击“确定”按钮。

年龄的统计描述如图 19.36 所示,因内容较多,依次解释如下。

(1) 集中趋势指标:可见样本的年龄均值为 38.4 岁,而 5% 截尾均数为 38.0,中位数为 38,三者相差不明显,说明年龄变量基本上呈对称分布。

(2) 离散趋势指标:样本年龄在 19 ~ 72 岁之间分布,方差为 124.5,其平方根即标准差为 11.2,全距和四分位间距分别为 53 和 16。由于标准差只有均数的大约 1/3,因此同样可以粗略看出数据分布是比较好的。

(3) 参数估计:可见年龄总体均数的标准误为 0.5,相应的总体均数 95% 可信区间为 37.4 ~ 39.4。

(4) 分布特征指标:图 19.36 最下方还给出表示数据偏离正态分布程度的偏度系数和峰度系数,及其各自的标准误,这里不再详述。

			统计量	标准误
年龄	均值		38.42	.507
	均值的 95% 置信区间	下限	37.43	
		上限	39.42	
	5% 修整均值		37.99	
	中值		38.00	
	方差		124.526	
	标准差		11.159	
	极小值		19	
	极大值		72	
	范围		53	
	四分位距		16	
	偏度		.467	.111
	峰度		-.459	.222

图 19.36 描述

在统计描述表格之后,探索过程还会给出年龄的茎叶图和箱图,这里只给出箱图的结果,如图 19.37 所示。

(1) 每个箱形都由最中间的粗线、一个方框、外延出来的两条细线和最外端可能有的单独散点组成。

(2) 箱体中间的粗线表示当前变量的中位数,方框的两端分别表示上、下四分位数(Q1 和 Q3,即 25% 和 75% 百分位数)。显然,整个方框内包括了中间 50% 样本的数值分布范围。

(3) 方框外的上、下两条细线分别表示除去异常值外的最大、最小值。

(4) 在箱图中,凡是与四分位数值(图 19.37 中即为方框上下界)的距离超过 1.5 倍四分位间距的都会被定义为异常值,其中离方框上/下界的距离超过四分位数间距 1.5 倍的为离群值,在图中以“O”表示;超过 3 倍的则为极值,用“*”表示。

从箱图得到的年龄分布信息与茎叶图基本相似,即年龄的分布虽然略有偏态,但程度应当是非常轻微的。

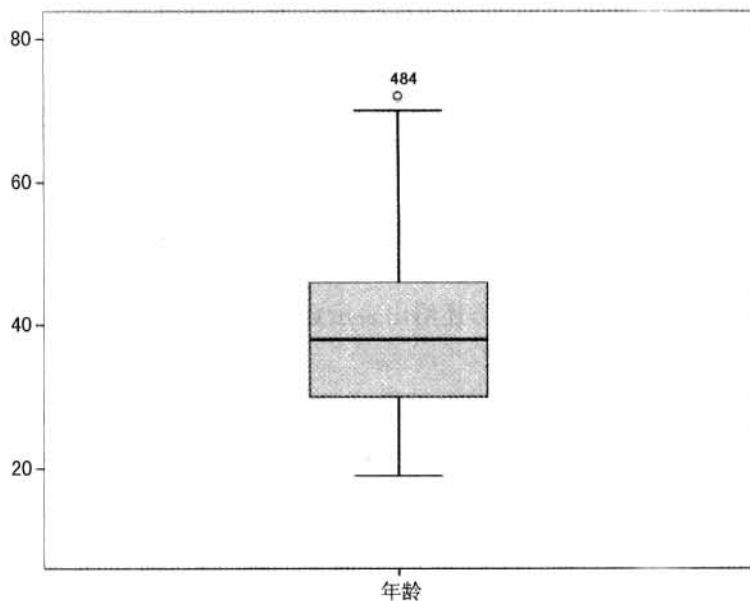


图 19.37 年龄的箱图

19.3.4 数据建模

1. 年龄对上市后指数影响程度的分析

由于年龄和上市后购买指数均为连续性变量,因此可以直接采用散点图对其关联性进行考察,当然,由于购买指数取值种类较少,可以考虑在散点图中进一步加绘回归线以使得趋势更为清晰。

- (1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。
- (2) 在图库中选择“散点图”图组,将简单散点图图标拖入画布中。
- (3) 将“年龄”拖入 X 轴框中,上市后购买指数拖入 Y 轴框中。
- (4) 单击“确定”按钮。
- (5) 双击生成的散点图进入编辑状态。
- (6) 选择“元素”→“总计拟合线”菜单项。
- (7) 在打开的对话框中选择“拟合线”选项卡,将置信区间更改为“均值”。
- (8) 单击“应用”按钮。

从图 19.38 中可以看到,如果在这两个变量进行回归分析,则相应模型的决定系数只有 0.002,如此低的决定系数意味着即使两者间的关联具有统计学意义,其联系也是非常微弱的。

下面通过相关分析来进一步给出假设检验的结果。

- (1) 选择“分析”→“相关”→“双变量”菜单项,打开“双变量相关分析”对话框。
- (2) 在“变量”列表框中选入“年龄”和“上市后购买指数”。
- (3) 单击“确定”按钮。

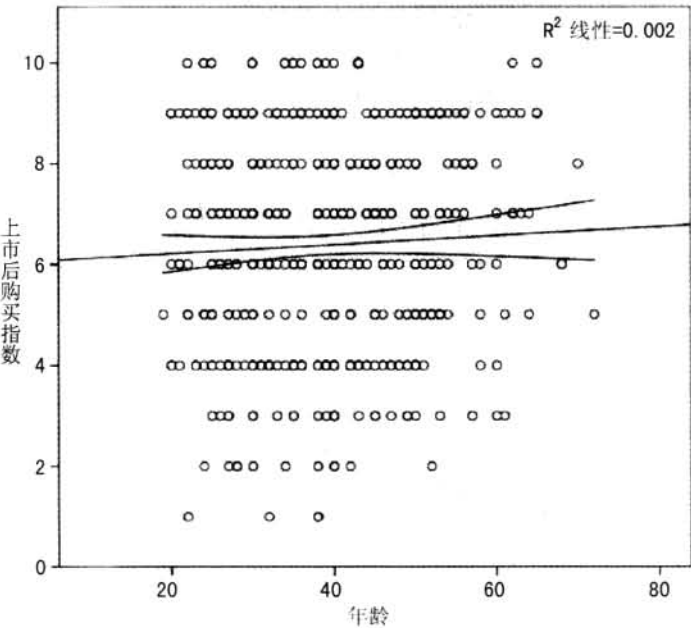


图 19.38 年龄与上市后购买指数的散点图

图 19.39 给出的就是积差相关系数的结果,是以对角阵的形式给出的,由于这里只分析了两个变量,因此给出的是 2×2 的方阵。每个单元格共分为 3 行,分别是相关系数、 P 值和样本数。可以看到年龄和上市后购买指数的相关系数为 0.047,对相关系数的检验的双侧 P 值为 0.3,远大于 0.05,因此这两个变量的线性关联趋势没有统计学意义。

年龄		上市后购买指数	
年龄	Pearson 相关性	1	.047
	显著性(双侧)		.300
	N	484	484
上市后购买指数	Pearson 相关性	.047	1
	显著性(双侧)	.300	
	N	484	484

图 19.39 相关性

2. 对品牌的作用进行总体检验

下面进一步考察品牌对上市后购买指数的作用大小,如前所述,该问题可被归纳为一般线性模型框架下的方差分析。由于在上面的分析中已经发现年龄可能不具有统计学意义,因此问题很可能会被简化成单因素方差分析。从标准的分析流程而言,这里首先应当对上市后购买指数进行品牌的分组描述,但这里将其合并,直接利用方差分析对话框中的描述功能来完成,操作步骤如下。

- (1) 选择“分析”→“比较均值”→“单因素 ANOVA”菜单项,打开“单因素 ANOVA”对话框。
- (2) 在“因变量”列表框中选出“上市后购买指数”。

- (3) 在“因子”列表框中选入最常使用的品牌。
- (4) 单击“选项”按钮,打开“选项”子对话框。
- (5) 在“统计量”框组中选“描述统计”、“方差同质性检验”复选框。
- (6) 单击“继续”按钮。
- (7) 单击“确定”按钮。

图 19.40 给出的就是各品牌组的上市后购买指数均值、标准差和样本量,从中可以很清晰地看出,目前使用高*或佳*品牌牙膏的受试者对测试新品的购买指数明显较低,但和其他品牌组相比是否具有统计学意义尚需检验后才能得知。

上市后购买指数								
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
高*	108	5.33	1.943	.187	4.96	5.70	1	9
佳*	105	5.67	1.736	.169	5.33	6.00	1	9
黑*	51	7.27	2.272	.318	6.64	7.91	3	10
中*	56	6.48	1.748	.234	6.01	6.95	2	10
蓝*	47	6.62	2.132	.311	5.99	7.24	2	10
其他	115	7.44	1.888	.176	7.09	7.79	3	10
总数	482	6.37	2.092	.095	6.19	6.56	1	10

图 19.40 描述

图 19.41 所示的是方差齐性检验的分析结果,此处的无效假设为:各组方差齐。可见 P 值为 0.057,大于 0.05,因此尚不能拒绝该无效假设,即可以认为方差是齐性的。

上市后购买指数			
Levene 统计量	df1	df2	显著性
2.163	5	476	.057

图 19.41 方差齐性检验

图 19.42 就是结果中最重要的方差分析表,可见该检验的 P 值远小于 0.05,因此组间的确是存在差异的,即最常使用品牌不同的人群,对该新产品的购买倾向是不同的。

上市后购买指数					
	平方和	df	均方	F	显著性
组间	345.819	5	69.164	18.717	.000
组内	1758.961	476	3.695		
总数	2104.780	481			

图 19.42 ANOVA

3. 品牌的组间两两比较

上面的结果表明品牌间是有差异的,但究竟哪些品牌间有差异还不确定。为了进一步详细地回答此问题,在进行方差分析后需要使用两两比较方法进行进一步的分析。这里采用 SNK 法

进行两两比较,操作步骤如下。

- (1) 打开“两两比较”子对话框。
- (2) 在“两两比较检验”框组中选择 S - N - K 选项。
- (3) 单击“继续”按钮。

图 19.43 是用 S - N - K 法进行两两比较的结果,该方法的输出比较特别,简单地说,首先就是将各组在表格的纵向上按照均数大小排序,然后在表格的横向上分成若干个亚组(Subset),不同亚组间的 P 值小于 0.05,而同一亚组内的各组均数则两两无差别,比较的 P 值均大于 0.05。从表 19.36 中可见,6 种品牌被分在了 3 个不同的亚组中,第一亚组由“高*”、“佳*”组成,评分最低,且它们两者间的比较无差异,在该列的最下方可见本亚组的检验 P 值为 0.302;第二亚组由“中*”、“蓝*”组成,评分居中;第三亚组由“黑*”、“其他”组成,评分最高。如果两个品牌被分在了完全不同的亚组中,则它们的均数有统计学差异,如“高*”和“黑*”,或者“高*”和“中*”均是如此。

Student - Newman - Keuls ^{a, b}				
最常使用的品牌	N	alpha = 0.05 的子集		
		1	2	3
高 *	108	5.33		
佳 *	105	5.67		
中 *	56		6.48	
蓝 *	47		6.62	
黑 *	51			7.27
其他	115			7.44
显著性		.307	.679	.604

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 69.589。

b. 组大小不相等。将使用组大小的调和均值。不保证 I 类错误级别。

图 19.43 上市后购买指数

19.3.5 项目总结与讨论

本案例分析了受访者目前使用品牌与年龄对新品上市后购买指数的影响,结果显示后者应当是没有作用的,而前者则有着明显的影响。根据以上分析结果可以得出如下分析结论。

- (1) 6 种现有牙膏品牌的使用人群大致可以分为 3 组,最常使用“黑*”、“其他”品牌的人对新产品表现出了较大兴趣,平均购买分值在 7 分以上。
 - (2) “高*”、“佳*”的使用者表现出了较高的忠诚度,上市后购买指数的平均分值仅在 5.5 分左右。
 - (3) “中*”、“蓝*”的适用者情况介于两者之间,评分居中。
- 综上,建议该新产品在上市后应当主攻“黑*”、“其他”品牌的定位人群,相对而言成功进入该细分市场的可能性较大,应当会有较好的收益。

19.4 证券业市场绩效与市场结构关系的实证分析

19.4.1 项目背景

在证券业中为了描述市场结构和绩效之间的关系,存在着各种各样的理论模型和假说,如市场结构假说(SCP假说)、有效结构假说(ES假说)、共谋假说(CH假说)等,这些假说存在着互相矛盾甚至于相反的一些前提假设,但是却都能适用于某些具体的市场状况。因此在开展进一步的研究工作之前,需要首先确定究竟所研究的证券市场适用于哪种假说。

2008年,某研究者针对国内证券市场进行了有关研究,经过文献检索,认为我国证券业市场结构与市场绩效之间的关系比较倾向于SCP假说与共谋假说(CH假说),并在借鉴Smirlock(1984)、Shepherd(1986)、Timme和Yang(1991)以及Berger(1995)等人所用模型的基础上,进一步构建了以下模型:

$$P = \beta_0 + \beta_1 CR + \beta_2 MS + \beta_3 EF + \beta_4 TA + \beta_5 ACR + \beta_6 GR + \beta_7 SR + \varepsilon$$

其中, P 是证券业的市场绩效, CR 是市场集中度, MS 是证券公司的市场份额, EF 代表证券公司的经营效率变量, TA 是证券公司总资产金额, ACR 为证券公司总资产金额与资本金的比率(反映证券公司利用财务杠杆来提高效益的能力,同时也反映证券公司的经营风险), GR 和 SR 是反映市场绩效的环境变量,其中 GR 是GDP增长率, SR 是股票市场增长率, β_i 为各待估系数, ε 是随机项。

在完成对上述模型的构建之后,研究者将进一步通过对模型中各系数的比较与计算确定其更加符合哪种细分的绩效模型,因为随后的工作相对偏离本书主题,此处不再赘述。

19.4.2 数据的采集

1. 模型中变量的进一步确定

19.4.1节的模型中列出的是较为理想的变量,但是在实际操作中,某些变量可能是难以获取的,或者是难以准确测量的,因此需要对模型中所需的变量根据实际情况进行进一步的调整,具体如下。

(1) 效率变量:在验证有效结构假说时,Smirlock(1984)等运用市场份额代表了效率变量,理由是效率高的企业会赢得更大的市场份额,两者之间高度相关。但是,Berger(1995)等人对此提出了质疑并认为,市场份额高的企业未必经营效率高,因此,市场份额不宜作为企业经营效率的代表。基于这一原因,最终研究者的模型中使用的是企业经营效率这一变量。但是在该指标的具体计算上,测定证券公司效率的方法涉及证券公司的投入、产出及劳动力价格、经营成本、业务收入等数据,而此类数据较难获取。相对比较容易获得的近似指标只能是证券公司的营业收入/资产总额。虽然严格说来,营业收入/资产总额过于笼统,但它与公司的综合经营效率间有较高的相关性。因此,在数据所限的情况下这这也是一个可行的选择。

(2) 市场绩效指标 P :对于证券公司而言,其效益主要体现在收益率上,因此在分析中选用净资产收益率(ROE)来衡量。

(3) 市场份额 MS :基于上述同样的理由,考虑采用证券公司总收入的总市场份额来代表。

(4) 其他指标:市场集中度 CR 采用 H 指数来表示, SR 则用上证综合指数每年收盘价的增

长率来衡量。

2. 实证检验的数据来源

由于国内证券市场自开始以来就波动频繁,有关公司变化很大,因此最终研究中所选取的研究对象是 2007 年总资产前 20 位的证券公司。选前 20 家公司的原因一是前 20 家证券公司能基本反映我国证券业的整体状况,二是考虑到数据的持续可得性,过于早期的数据不仅可能无法代表现在的市场状况,而且很难收集完整。

样本数据的采集区间为 6 年,即 2002 ~ 2007。因此理论上的样本总数为 120。数据主要来源于《中国证券期货统计年鉴 2007》,对于 GDP 等年度数据,则直接采用复制的方法赋值给每一个案例。最终整理完毕的数据见文件证券实证分析 . sav。

19.4.3 数据预分析

1. 原始变量描述

首先应当对所有需要纳入模型分析的变量进行基本的描述,采用描述过程,相应的结果如图 19.44 所示。

	N	极小值	极大值	均值	标准差
营业收入	105	- 1699. 05	3087111. 41	274433. 2303	462951. 44699
利润总额	103	- 70247. 80	1990417. 28	151051. 2413	319913. 76884
净利润	104	- 132463. 58	1354578. 59	103079. 4217	221767. 89590
总资产	107	146436. 64	18965388. 17	2226362. 7287	2849258. 45363
负债合计	104	45075. 41	13563045. 62	1818178. 6829	2308847. 27388
实收资本	83	51874. 52	873443. 89	255170. 6120	198925. 85845
所有者权益	85	- 614594. 09	1230994. 94	251783. 6613	214779. 14810
净资产	102	- 39329. 58	5402342. 55	399512. 9902	697876. 38109
净资产收益率	104	- . 85	. 67	. 1096	. 25949
H 指数	107	. 04	. 05	. 0427	. 00626
证券市场增长率	107	- . 18	1. 30	. 3963	. 61229
GDP 增长率	107	9. 10	11. 90	10. 4832	. 84580
市场份额	107	. 00	. 11	. 0273	. 02137
有效的 N (列表状态)	78				

图 19.44 描述统计量

在描述图 19.44 中可以发现,几乎所有的变量都存在一定数量的缺失值,这导致所采集到的 107 例样本只有 78 例具有完整的信息,数据缺失的原因是有一部分未能收集到,另一部分确实由于各种原因而无法计算得到。

一般而言,当总样本中因缺失而导致的样本损失超过 10% 时,就可能对分析结果造成明显的影响。本例已经达到了近 30%。但是除非能够继续收集到所缺失的数据,否则在统计上不应当采取什么主动措施来加以弥补,因为证券市场的波动非常剧烈,在没有弄清楚其具体规律之前,任何一种缺失值填补方法都可能会带来更大的误差。而且本研究的目的只是通过获取拟合方程的系数估计值来进行市场的模式推断,并不涉及数值预测,因此保持原数据状况进行模型估计应当是更合理的做法。

在数据表中从很多变量的标准差都接近甚至大于均值可以看出另一个问题,即这些变量都

存在着高度的离散性,很可能不服从正态分布。实际上,这恐怕是所有金融类模型都可能遇到的问题。这里同样不考虑进行事前处理,而是建模之后再进行考虑。

最后,从图 19.37 所示的数据可以得出一个关于该行业的结论:低市场集中度和低利润率并存是我国证券业的一大特点。

2. 数据变换

下面考虑对建模所需的变量进行计算,根据上面的讨论,相应的数值计算程序如下:

COMP 经营效率 = 营业收入/总资产.

COMP ACR = 总资产/所有者权益.

EXEC.

上述程序运行之后,数据集中就会增加经营效率和 ACR 两个新变量,注意这两个新变量中也存在缺失值,这是因为计算用原始变量也存在缺失值。

19.4.4 数据建模



比较熟悉统计模型的读者可能会想到,这应当是一个比较典型的重复测量数据结构:20 家公司在 6 个年度重复给出所测量的数据,构成了一个完美的重复测量结构,因此在分析的时候就应当考虑这种特殊的数据结构,采用重复测量的一般线性模型,最好能够采用重复测量的 GEE 模型或者混合效应模型来进行建模估计——这类想法完全可以理解,但模型是用来解决问题的,本案例分析目的如此简单,有必要建立这么复杂的分析模型吗?

下面考虑基于上述数据对前面构建的模型公式进行估计,采用回归过程来分析,注意此处不应当采用变量筛选,而是使用进入法将所有变量纳入模型,相应的结果如图 19.45 所示。

图 19.45 显示整个模型的决定系数为 39%,也就是说对样本的因变量变异有一定的解释能力,但是否具有统计学意义尚需进行检验。

模型	R	R 方	调整 R 方	标准 估计的误差
1	.624 ^a	.390	.333	.16951

a. 预测变量: (常量), 证券市场增长率, ACR, 市场份额, 经营效率, GDP 增长率, 总资产, H 指数。

图 19.45 模型汇总

如图 19.46 所示,模型总的检验结果提示,整个模型用于预测因变量是有价值的,具有统计学意义。

模型	平方和	df	均方	F	Sig.
回归	1.376	7	.197	6.844	.000 ^b
1 残差	2.155	75	.029		
总计	3.531	82			

a. 因变量: 净资产收益率。

b. 预测变量: (常量), 证券市场增长率, ACR, 市场份额, 经营效率, GDP 增长率, 总资产, H 指数。

图 19.46 Anova^a

图 19.47 给出了模型中每个系数的估计值和检验结果,可见大多数系数都没有统计学意义,那么是否需要在提出后建立简化模型呢?在本例中恰恰是不需要这样做的,至少在回答研究问题这一点上,简化模型并不重要。

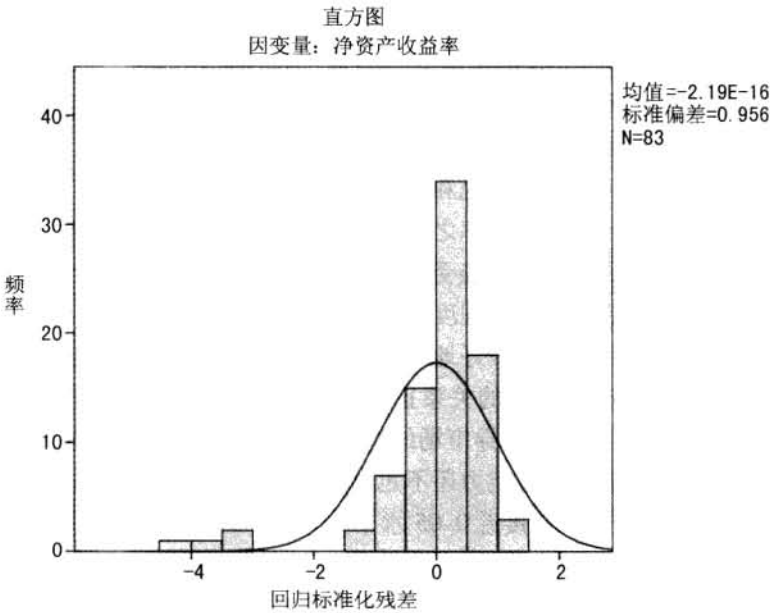
模型	非标准化系数		标准化系数		t	Sig.
	B	标准 误差	试用版			
1 (常量)	-.023		.553		-.041	.967
H 指数	-10.050	14.265	-.340		-.704	.483
市场份额	.362	3.939	.035		.092	.927
经营效率	1.090	.349	.314		3.120	.003
总资产	1.09E-008	.000	.060		.147	.883
ACR	-.002	.005	-.048		-.511	.611
GDP 增长率	.031	.086	.080		.362	.718
证券市场增长率	.232	.134	.687		1.737	.087

a. 因变量：净资产收益率。

图 19.47 系数^a

作为建模后的常规步骤,也可以考虑对模型的残差分布等进行考察,图形结果如图 19.48 所示,可见模型的残差分布并不理想,主要是在较小侧存在可疑的极端值。可以考虑删除相应的案例,重新拟合无极端值的模型,以比较回归系数的估计值是否会因极端值的出现而发生变化,读者可自行按此思路操作,此处不再详述。

除极端值以外,如果进行共线性分析,则会发现该模型可能会存在共线性问题,但是在本例中,共线性问题可能是无法解决的,因为从自变量的含义上,这些变量可能存在数量上的关联,因此基于本案例的分析目的,首先考虑的应当是模型分析结果能否协助找到合适的市场假说模型,只有当系数估计值不合理,无法达到分析目的时,才应当考虑是否需要去解决共线性问题。



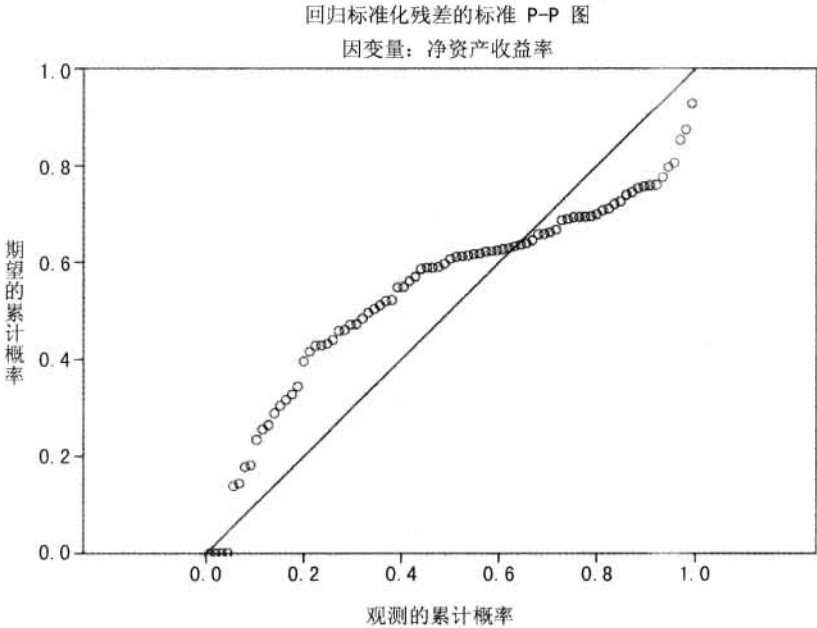


图 19.48 模型残差的直方图和 P-P 图

19.4.5 项目总结与讨论

1. 项目分析结论

由于上述回归模型总体检验具有统计学意义,因此模型可以用来解释我国证券业市场结构与市场绩效之间的关系。从具体回归方程的系数结果来看,可以得到以下结论。

(1) 我国证券业的市场绩效与市场集中度之间的正相关关系并未通过显著性检验($P = 0.483$),因此,根据有关理论,可以拒绝纯共谋假说。这也同时表明,较高的市场集中度在给我国证券公司带来高额垄断利润的同时并没有给投资者带来相应的投资回报。

(2) 我国证券业的市场绩效与市场份额之间没有表现出显著的正相关关系($P = 0.927$),因此,根据有关理论,拒绝纯有效结构假说和混合的共谋/有效结构假说。

(3) 市场绩效和基金公司经营效率之间的正相关关系有统计学意义,而与市场集中度之间的正相关关系不显著,因此,可以考虑接受修正的有效结构假说。

(4) 我国证券业并未体现出一定程度的规模经济($P = 0.883$)。

(5) 我国证券业总体上并不能较好地控制财务风险并提高资本创利能力($P = 0.611$)。

(6) 证券市场绩效与 GDP 之间的关系正相关但无统计学意义,这是因为,我国证券市场绩效的变化主要由证券市场主导,而证券市场的涨升与国民经济运行存在脱节。

(7) 我国证券业抗股市风险的能力还不足,市场绩效在很大程度上可能还依赖于股市行情(回归系数检验的 $P = 0.087$,非常接近 0.05 水平)。

2. 项目讨论

本分析项目实际上在统计分析上是一个比较简单的回归问题,笔者将它放在这里是为了演

示统计分析方法在不同的专业领域中可能遇到的具体应用方式。对于并不熟悉证券业的读者而言,可能在处理本案例的时候会考虑得非常复杂:缺失值应当如何处理?模型中的共线性、异常值这些问题应当如何处理?但是实际上,只需要抓住一点即可:本研究的目的是在几种可能的市场假说中找到最为正确的一种,并非要求建立一个高度精确的模型,因此在能够正确解答这一问题的前提下,并非所有的统计学问题都需要面面俱到的去考虑和解答,更何况在这里所假设要拟合的模型框架之中,本身就蕴含着可能出现共线性的变量设定。简言之,在将统计方法应用到具体实践的时候,需要根据情况灵活应用,客观取舍,而不是局限于教条的规定以至于寸步难行。

思考与练习

自行练习本章中涉及的案例数据操作。

附 录

附录 1 SPSS 函数一览表

本部分的内容是基于目前上市的最新版本 IBM SPSS Statistic20 编写的,在 SPSS 中共有上百个函数,可将其分为 10 多个类别。每个函数由两部分构成,一部分是函数名称,以大写字母表示;另一部分是参数,以小写字母表示,一个函数中可以有一个或几个参数,每个参数之间用逗号分隔,所有参数用括号括起来。

参数是使用函数时要替换和更改的部分,因此掌握函数,就必须掌握每个参数的意义。每个参数要求的表达式的形式是不一样的,有的要求是数值型(既可以是具体数字,也可以是数值型变量);有的要求是字符型(既可以是具体字符,也可以是字符型变量);有的要求是日期型(既可以是具体日期时间,也可以是日期型变量);还有的参数对其取值范围有具体要求。SPSS 的函数中涉及的参数大致可归纳为以下几类。

- (1) 数字或数值型变量作参数,如 num、radians、mod、high、low、test、pos、length、divisor、value、numexpr、numvar、variable。
- (2) 各种分布的参数,如 quant、prob、shapel、r、scale、loc、df、mean、std、sample、hits、total、threshold、size、min、max、zvalue、nc。
- (3) 字符或字符型变量作参数,如 high、low、test、char、needle、haystack、strexpr、value、variable。
- (4) 数值或时间日期型变量作参数,如 timevalue、day、month、year、quarter、weeknum、daynum、hours、min、sec、datevalue。
- (5) 变量作参数,如 variable。

下面分类介绍各类函数。

1. 数学函数

数学函数如附表 1.1 所示。

附表 1.1 数 学 函 数

函 数 形 式	返回值类型	函 数 说 明
ABS(num)	数值型	计算 num 的绝对值 例:ABS(-3)=3
ARSIN(num)	数值型	返回 num 的反正弦值,以弧度为单位,num 需介于 -1~1 之间
ARTAN(num)	数值型	返回 num 的反正切值,以弧度为单位
COS(radians)	数值型	返回 radians 的余弦值,以弧度为单位
EXP(num)	数值型	返回 e 的 num 次幂 例:EXP(2)= $e^2=7.389$

续表

函数形式	返回值类型	函数说明
LG10(num)	数值型	返回 num 以 10 为底的对数值,num 必须大于 0 例:LG10(100) = $\log_{10} 100 = 2$
LN(num)	数值型	返回 num 的自然对数值,num 必须大于 0 例:LN(7.389) = 2
LNGAMMA(num)	数值型	返回 num 的完全 Gamma 函数的自然对数值,num 必须大于 0 例:LNGAMMA(5) = 3.18
MOD(num, mod)	数值型	返回 num 除以 mod 以后的余数,mod 不能为 0 例:MOD(3,2) = 1
RND(num, [mult, fuzz])	数值型	只使用 num 参数,则返回最接近 num 的整数;mult 指定结果为该数值的整数倍;fuzz 用于设定四舍五入所需要考虑的小数位数阈值,默认为 6,该选项用在一些特殊情况下避免程序计算错误,一般不需要使用 例:RND(4.75) = 5,RND(4.75, 0.1) = 4.7
SIN(radian)	数值型	返回 num 的正弦值,参数必须为数值型 例:SIN(3.14) = 0
SQRT(num)	数值型	返回 num 的平方根,参数必须为数值型,又不为负数 例:SQRT(4) = 2
TRUNC(num, [mult, fuzz])	数值型	返回 num 向 0 方向截尾的值,用法类似于 RND 函数 例:TRUNC(4.7) = 4

2. 累积概率函数、逆分布函数、概率密度函数与显著性函数

基于统计分布的函数如附表 1.2 所示,共分为以下几大类。

(1) CDF 族函数:用于计算当概率函数值等于 quant 时指定分布函数的下侧(左侧)累积概率值,共有 25 种常用分布可供选择。

(2) 非中心 CDF 族函数:返回非中心化分布函数的下侧(左侧)累积概率值,和普通分布相比,其中非中心化分布需要多指定一个非中心参数。

(3) IDF 族函数:相当于 CDF 族函数的反函数,返回指定分布在下侧累积概率值为所给数值时的函数值。共有 18 个,基本上和 CDF 族函数一一对应,这里不再一一列出。

(4) PDF 族函数:即概率密度函数,用于计算当概率函数值等于 quant 时指定分布函数的点概率密度值,共有 25 种,和上面的 CDF 函数可一一对应,此处不再重复。

(5) 非中心 PDF 函数:共有 4 种,用于计算非中心化分布函数的点概率密度值,和上面的 NCDF 函数可一一对应,此处不再重复。

(6) SIG 族函数:用于计算相应分布的右侧累积概率值,目前只有 CHISQ 和 F 两个。

附表 1.2 累积概率函数等

函 数 形 式	返回值类型	函 数 说 明
CDF. BERNOULLI (quant , prob)	数值型	返回参数为“prob”的伯努利分布的“quant”分位点的累积概率值
CDF. BETA (quant , shapel , shape2)	数值型	返回参数为“shapel ,shape2”的贝塔分布的“quant”分位点的累积概率值
CDF. BINOM (quant , n , prob)	数值型	返回实验次数 (n) 和成功概率 (prob) 的二项分布的“quant”分位点的累积概率值
CDF. BVNOR(q1 , q2 , r)	数值型	返回相关系数为“r”的双变量标准正态分布的“q1 ,q2”分位点的累积概率值
CDF. CAUCHY (quant , loc , scale)	数值型	返回位置、比例参数分别为“loc”和“scale”的柯西分布的“quant”分位点的累积概率值
CDF. CHISQ(quant , df)	数值型	返回自由度为“df”的卡方分布的“quant”分位点的累积概率值
CDF. EXP(quant , shape)	数值型	返回参数为“shape”的指数分布的“quant”分位点的累积概率值
CDF. F(quant , df1 , df2)	数值型	返回自由度为“df1 \df2”的 F 分布的“quant”分位点的累积概率值
CDF. GAMMA (quant , shape , scale)	数值型	返回来自给定形状参数 (shape) 和比例参数 (scale) 的伽玛分布的“quant”分位点的累积概率值
CDF. GEOM(quant , prob)	数值型	返回概率参数为“prob”的几何分布的“quant”分位点的累积概率值
CDF. HALFNRM (quant , mean , std)	数值型	返回总体均数为“mean”和标准差为“std”的半正态分布的“quant”分位点的累积概率值
CDF. HYPER (quant , total , sample , hits)	数值型	返回总体为“total”和对应项大小为“hits”以及样本大小为“sample”的超几何分布的“quant”分位点的累积概率值
CDF. IGAUSS (quant , mean , scale)	数值型	返回来自给定均数和标准差的反高斯分布的“quant”分位点的累积概率值
CDF. LAPLACE (quant , mean , scale)	数值型	返回均值为“mean”和比例参数为“scale”的拉普林斯分布的“quant”分位点的累积概率值
CDF. LOGISTIC (quant , mean , scale)	数值型	返回均值为“mean”和比例参数为“scale”的 Logistic 分布的“quant”分位点的累积概率值
CDF. LNORMAL (quant , a , b)	数值型	返回参数为“a ,b”的对数正态分布的“quant”分位点的累积概率值

续表

函数形式	返回值类型	函数说明
CDF. NEGBIN (quant , thresh , prob)	数值型	返回次数参数为“thresh”和概率为“prob”时获取成功所需试验次数的“quant”分位点的累积概率
CDF. NORMAL (quant , mean , stddev)	数值型	返回均值为“mean”和标准差为“stddev”的正态分布的“quant”分位点的累积概率值
CDF. PARETO (quant , threshold , shape)	数值型	返回参数为“threshold”和形状参数 shape 的帕累托分布的“quant”分位点的累积概率值
CDF. POISSON (quant , mean)	数值型	返回均值为“mean”的泊松分布的“quant”分位点的累积概率值
CDF. SMOD (quant , size , df)	数值型	返回参数为“size、df”的 Studentized Maximum Modulus 的“quant”分位点的累积概率值
CDF. SRANGE (quant , size , df)	数值型	返回参数为“size、df”的 Studentized Range Statistic 的“quant”分位点的累积概率值
CDF. T(quant , df)	数值型	返回自由度为“df”的 T 分布的“quant”分位点的累积概率值
CDF. UNIFORM (quant , min , max)	数值型	返回最小值为“min”和最大值为“max”的均匀分布的“quant”分位点的累积概率值
CDF. WEIBULL (quant , a , b)	数值型	返回参数为“a、b”的威布尔分布的“quant”分位点的累积概率值
CDFNORM(zvalue)	数值型	返回标准正态分布的“zvalue”分位点的累积概率值
NCDF. BETA (quant , shape1 , shape2 , nc)	数值型	返回形状参数为“shape1、shape2”和非中心参数为“nc”的非中心贝塔分布的“quant”分位点的累积概率值
NCDF. CHISQ (quant , df , nc)	数值型	返回自由度为“df”和非中心参数为“nc”的非中心卡方分布的“quant”分位点的累积概率值
NCDF. F (quant , df1 , df2 , nc)	数值型	返回自由度为“df1、df2”和非中心参数为“nc”的非中心 F 分布的“quant”分位点的累积概率值
NCDF. T(quant , df , nc)	数值型	返回自由度为“df”和非中心参数为“nc”的非中心 T 分布的“quant”分位点的累积概率值
SIG. CHISQ(q , df)	数值型	返回自由度为“df”的卡方分布的“quant”分位点的右侧累积概率值
SIG. F(q , df1 , df2)	数值型	返回自由度为“df1、df2”的 F 分布的“quant”分位点的右侧累积概率值

3. 日期时间函数

主要有 6 类,如附表 1.3 所示,在计算时均以 1582 年 10 月 14 日午夜为基线时间,它们的功能如下。

(1) 当前日期与时间:用于返回函数执行时的系统时间。

(2) 日期运算:对时间变量进行四则运算。

(3) TIME:用于创建时间变量,返回数值型时间变量,参数“day”、“hour”、“min”、“sec”为数值,反映当前时间的秒数。

(4) DATE:用于创建日期,返回数值型时间变量,参数“day”、“month”、“quarter”、“year”、“weeknum”、“daynum”均为数值型,反映当前时间距基线日期的秒数。

(5) CTIME:用于提取时间段,返回数值型变量,参数“timevalue”为时间变量或表达式,即一般要使用 Time 或 Date 系列函数来设置变量参数,返回该日期和基线时间相差的累计日、时、分或秒数,计算时会上一层次的差异进行换算,如将小时换算为分。

(6) XDATE:用于提取日期,返回数值变量,参数“datevalue”必须为时间变量或表达式,即一般要使用 Time 或 Date 系列函数来设置变量参数,返回该日期和基线时间相差的日、时、分或秒数,注意只计算同级差异,如计算秒时不考虑分的差异。

附表 1.3 日期时间函数

函 数 形 式	返回值类型	函 数 说 明
\$ DATE	日期数值型	返回采用两位数计年格式的当前日期
\$ DATE11	日期数值型	返回采用四位数计年格式的当前日期
\$ JDATE	数值型	返回从 1582 年 10 月 14 日开始计算的当前日期,格式为 F6.0
\$ TIME	数值型	返回从 1582 年 10 月 14 日午夜到命令执行时间所经过的秒数,格式为 F20
DATEDIFF(time2, time1, " unit")	数值型	返回两个日期变量的差值,unit 则为引号括起来的有效时间单位值,如 years、months、minutes 等
DATESUM(time, value, " unit" , " method")	日期数值型	按照 unit 给定的时间单位,将 value 的值和 time 相加,备选的方法默认为 closest,即使用月中最接近的合法日期,可更改为 rollover,即将多余天数前移
DATE. DMY(day, month, year)	日期数值型	返回日期“year”年“month”月“day”日距基线日期的秒数 例:DATE. DMY(02,03,1982)意思是 1982 年 3 月 2 日
DATE. MDY(month, day, year)	日期数值型	返回日期:“year”年“month”月“day”日距基线日期的秒数 例:DATE. MDY(02,03,1982)意思是 1982 年 2 月 3 日
DATE. MOYR(month, year)	日期数值型	返回日期:“year”年“month”月 1 日距基线日期的秒数 例:DATE. MOYR(02,1982)意思是 1982 年 2 月
DATE. QYR(quarter, year)	日期数值型	返回日期:“year”年第“quarter”季第 1 天距基线日期的秒数 例:DATE. QYR(3,1982)意思是 1982 年第 3 季度

续表

函数形式	返回值类型	函数说明
DATE. WKYR(weeknum, year)	日期数值型	返回日期:“year”年第“weeknum”周第1天距基线日期的秒数 例:DATE. WKYR(21,1982)意思是1982年第21周
DATE. YRDAY(year, daynum)	日期数值型	返回日期:“year”年第“daynum”天距基线日期的秒数 例:DATE. YRDAY(1982,21)意思是1982年第21天
XDATE. DATE(datevalue)	数值型	返回“datevalue”距离1582年1月1日的秒数,与CTIME. SECONDS(timevalue)相似 例:XDATE. DATE(1952/02/03) = 11654150400.00
XDATE. HOUR(datevalue)	数值型	返回“datevalue”为本天第几时,为整数,介于0~23之间 例:XDATE. HOUR(02 - MAR - 2003 02:30:30) = 2
XDATE. JDAY(datevalue)	数值型	返回“datevalue”为本年度第几天,为整数,介于1~366之间 例:XDATE. JDAY(31 - DEC - 2004) = 366
XDATE. MDAY(datevalue)	数值型	返回“datevalue”为本月第几天,为整数,介于0~31之间 例:XDATE. MDAY(31 - DEC - 2003) = 31
XDATE. MINUTE(datevalue)	数值型	返回“datevalue”为本时第几分,为整数,介于0~59之间 例:XDATE. MINUTE(02 - MAR - 2003 02:30:29) = 30
XDATE. MONTH(datevalue)	数值型	返回“datevalue”为本年度第几月,为整数,介于1~12之间 例:XDATE. MONTH(02 - MAR - 2003 02:30:29) = 3
XDATE. QUARTER(datevalue)	数值型	返回“datevalue”为本年度第几季,为整数,介于1~4之间 例:XDATE. QUARTER(02 - MAR - 2003 02:30:29) = 1
XDATE. SECOND(datevalue)	数值型	返回“datevalue”为本分第几秒,为整数,介于1~60之间 例:XDATE. SECOND(02 - MAR - 2003 02:30:29) = 29
XDATE. TDAY(datevalue)	数值型	返回“datevalue”距离1582年1月1日的整数天数,与CTIME. SECONDS(timevalue)相似 例:XDATE. TDAY(02 - MAR - 2003 02:30:29) = 153541.00
XDATE. TIME(datevalue)	日期数值型	返回“datevalue”为当天的第几秒 例:XDATE. TIME(02 - MAR - 2003 02:30:29) = 26430.00
XDATE. WEEK(datevalue)	数值型	返回“datevalue”为当年的第几整周,取值为1~53 例:XDATE. WEEK(02 - MAR - 2003 02:30:29) = 9.00
XDATE. WKDAY(datevalue)	数值型	返回“datevalue”为星期几,取值为1~7的整数 例:XDATE. WKDAY(02 - MAR - 2003 02:30:29) = 1.00
XDATE. YEAR(datevalue)	数值型	返回“datevalue”4位数整数表示的年号 例:XDATE. YEAR(02 - MAR - 2003 02:30:29) = 2003.00

续表

函数形式	返回值类型	函数说明
YRMODA(year, month, day)	数值型	返回从 1582 年 10 月 15 日开始一直到参数 year、month 和 day 所代表的天数 例: XDATE. YEAR (03,3,2) = 153541.00
TIME. DAYS(days)	日期数值型	返回“days”天的秒数, “days”必须为数值 例: TIME. DAYS(1) = 86500(1 天为 86 400 秒)
TIME. HMS(hours, [min, sec])	日期数值型	返回“hour”小时“min”分“sec”秒所对应的秒数, 参数必须为数值, 且后两者可选。如果希望结果显示为时间, 则将其指定为时间格式 例: TIME. HMS (1,1,1) = 3661(1 小时 1 分 1 秒 = 3 661 秒)
CTIME. DAYS(timevalue)	数值型	返回“timevalue”距基线日期的累计天数, 包括小数, “timevalue”必须是数字或日期格式的表达式 例: CTIME. DAYS(1952/02/03) = 134886.00
CTIME. HOURS(timevalue)	数值型	返回“timevalue”距基线日期的累计小时数, 包括小数, “timevalue”必须是数字或日期格式的表达式 例: CTIME. HOURS(1952/02/03) = 3237264.00
CTIME. MINUTES(timevalue)	数值型	返回“timevalue”距基线日期的累计分钟数, 包括小数, “timevalue”必须是数字或日期格式的表达式 例: CTIME. MINUTES(1952/02/03) = 194235840.00
CTIME. SECONDS(timevalue)	数值型	返回“timevalue”距基线日期的累计秒数, 包括小数, “timevalue”必须是数字或日期格式的表达式 例: CTIME. SECONDS(1952/02/03) = 11654150400.00

4. 缺失值函数

缺失值函数如附表 1.4 所示。

附表 1.4 缺失值函数

函数形式	返回值类型	函数说明
\$ SYSMISS		返回系统缺失值
MISSING(variable)	逻辑型	判断变量“variable”是否为缺失值, 如是, 则返回“1”, 否则返回“0”
NMISS(variable(...))	数值型	返回“variable(...)”中含缺失值的变量个数
NALID(variable(...))	数值型	返回“variable(...)”中非缺失值的变量个数
SYSMIS(numvar)	逻辑型	判断数值型变量“numvar”是否为系统缺失值
VALUE(variable)	数值型/字符型	返回变量“vanable”的值, 即使是用户自定义的缺失值也返回, 并不再把它看做缺失值

5. 随机函数

主要为 RV 系列函数,用于返回随机数,RV 函数共有 25 个,同样可以和 CDF 系列一一对应,这里不再重复。

6. 检索函数

检索函数如附表 1.5 所示。

附表 1.5 检索函数

函数形式	返回值类型	函数说明
ANY (test, value, (value...))	逻辑型	如果“test”与各“value”中的任何一个匹配,则返回“1”,否则返回“0” 例:ANY('a','b','c','d') = 0 ANY('a','b','a','d') = 1
CHAR. INDEX (haystack, needle, [divisor])	数值型	返回 haystack 中第一次出现 needle 的位置,可选参数 divisor 用于指定可以将 needle 划分为单独字符串的字符数
CHAR. RINDEX (haystack, needle, [divisor])	数值型	返回 haystack 中最后一次出现 needle 的位置,可选参数 divisor 用于指定可以将 needle 划分为单独字符串的字符数
MAX (value, value [,...])	数值型/字符型	返回“value”中的最大值,需要两个或两个以上 value,可以为本函数指定有效变量的最小个数 例:MAX(2,3,7,9) = 9;MAX(a,b,c,d) = d
MIN (value, value [,...])	数值型/字符型	返回 value 中的最小值,需要两个以上 value,可以为本函数指定有效变量的最小个数 例:MAX(2,3,7,9) = 2;MAX(a,b,c,d) = a
RANGE (test, low, high, (low, high...))	逻辑型	如果“test”处于“low”与“high”确定的范围内,则返回“1”,否则返回“0”。如果“low”与“high”为字符,必须等长度 例:ANY(2,10,20) = 0 ANY('d','a','f') = 1
REPLACE (a1, a2, a3 [a4])	字符型	在 a1 中,将所有 a2 实例替换为 a3,a4 用于指定允许替换的次数,默认为全部替换

7. 统计函数

统计函数如附表 1.6 所示。

附表 1.6 统计函数

函数形式	返回值类型	函数说明
CFVAR (numexpr, numexpr [,...])	数值型	返回“numexpr”中有效值构成样本的变异系数,需要两个或两个以上 value,可以为本函数指定有效变量的最小个数
MAX (value, value [,...])	数值型/字符型	返回“value”中的最大值,需要两个或两个以上 value,可以为本函数指定有效变量的最小个数 例:MAX(2,3,7,9) = 9;MAX(a,b,c,d) = d

续表

函 数 形 式	返回值类型	函 数 说 明
MEDIAN (value, value [,...])	数值型	返回 value 中的中位数,需要两个以上 value,可以为本函数指定有效变量的最小个数
MEAN (numexpr, numexpr [,...])	数值型	返回“numexpr”这些数值型变量的算术平均值,需要两个或两个以上 numexpr,可以为本函数指定有效变量的最小个数 例:MEAN(3,4,8) = 5
MIN (value, value [,...])	数值型/字符型	返回 value 中的最小值,需要两个以上 value,可以为本函数指定有效变量的最小个数 例:MAX(2,3,7,9) = 2;MAX(a,b,c,d) = a
SD (numexpr, numexpr [,...])	数值型	返回“numexpr”这些数值型变量中含有有效值的变量的标准差,此函数需要两个或两个以上的数值型变量,可以为此函数指定有效变量的最小个数
SUM (numexpr, numexpr [,...])	数值型	返回“numexpr”这些数值型变量中含有有效值的变量的总和,此函数需要两个或两个以上的数值型变量,可以为此函数指定有效变量的最小个数
VARIANCE (numexpr, numexpr [,...])	数值型	返回“numexpr”这些数值型变量中含有有效值的变量的方差,此函数需要两个以上的数值型变量,可以为此函数指定有效变量的最小个数

8. 字符串函数

字符串函数如附表 1.7 所示。

附表 1.7 字符串函数

函 数 形 式	返回值类型	函 数 说 明
CHAR. INDEX (haystack, needle, [divisor])	数值型	返回 haystack 中第一次出现 needle 的位置,可选参数 divisor 用于指定可以将 needle 划分为单独字符串的字符数
CHAR. RINDEX (haystack, needle, [divisor])	数值型	返回 haystack 中最后一次出现 needle 的位置,可选参数 divisor 用于指定可以将 needle 划分为单独字符串的字符数
CHAR. LENGTH (strexpr)	数值型	返回“strexpr”的长度不包括尾部空格 例:CHAR. LENGTH('abcde') = 5
CHAR. LPAD (strexpr, length, [str2])	字符型	使用 str2 在“strexpr”左侧填充,直至其长度达到 length,如果省略 str2,则使用空格填充 例:LPAD('ab',5,'c') = 'cccab'
CHAR. RPAD (strexpr, length, [str2])	字符型	使用 str2 在“strexpr”右侧填充,直至其长度达到 length,如果省略 str2,则使用空格填充
CHAR. MBLEN (strexpr, pos)	数值型	返回在 strexpr 的字符位置 pos 处的字符中的字节数

续表

函数形式	返回值类型	函数说明
CHAR. SUBSTR(strexpr, pos, [length])	字符型	返回“strexpr”中从“pos”位置开始、长度为“length”的字符串,“pos”、“length”均为整数,省略length则返回至末尾
CONCAT (strexpr, strexpr [,...])	字符型	返回“strexpr”合并而成的字符串,需要两个或两个以上的表达式 例:CONCAT('a','b') = ab
LENGTH(strexpr)	数值型	返回“strexpr”的长度,在常用的代码页模式中,该函数返回的实际上是定义的字符串长度,包括尾部空格
LOWER(strexpr)	字符型	将“strexpr”中的大写字母转换为小写返回 例:LOWER('aBcD') = abcd
UPCAS(strexpr)	字符型	将“strexpr”中的小写字母变为大写返回 例:LOWER('aBcD') = ABCD
LTRIM(strexpr, [char])	字符型	删除“strexpr”左侧的“char”,“char”必须是一个单一字符,省略时则删除空格 例:LTRIM('aatt','a') = tt
MBLEN. BYTE (strexpr, pos)	数值型	返回在 strexpr 的字节位置 pos 处的字符中的字节数
NORMALIZE(strexpr)	字符串	返回 strexpr 标准化版本。在 Unicode 模式下,返回 Unicode NFC。在代码页模式下,无效应并返回未修改的 strexpr。结果长度可能与输入长度不同
NTRIM(varname)	字符串	返回 varname 值,不用删除拖尾空格。varname 的值必须是一个变量名,不可以是一个表达式
REPLACE (a1, a2, a3 [a4])	字符串	在 a1 中,将所有 a2 实例替换为 a3,a4 用于指定允许替换的次数,默认为全部替换
RTRIM(strexpr, char)	字符型	删除“strexpr”右侧的“char”,“char”必须是一个单一字符,省略则删除空格
SUBSTR(strexpr, pos)	字符型	返回“strexpr”中从“pos”位置开始至最后一个字符的字符串,“pos”是一个数字 例:SUBSTR('factory',2) = actory
STRUNC(strexpr, length)	字符串	返回截断至 length 长度(以字节为单位)的 strexpr,然后删除所有拖尾空格。截断将删除任何可能被截断的字符片段

9. 转换、特殊变量与其他函数

转换、特殊变量与其他函数如附表 1.8 所示。

附表 1.8 转换、特殊变量与其他函数

函 数 形 式	返回值类型	函 数 说 明
NUMBER (strexpr, format)	数值型	以“format”格式返回“strexpr”的数值 例:NUMBER('3.2', f8.1) = 3.2(为数值型)
STRING(num, format)	字符型	以“format”格式读取“num”数值,返回类型字符型 例:STRING(-1.5, F5.2) = -1.50(为字符型)
\$ CASENUM		返回当前个案的顺序号
LAG(variable, [n])	数值型/字符型	返回前面第 n 条记录的“variable”的取值,为前 n 条记录返回系统缺失值(对于数值型变量)或空格(对于字符型变量),第二个参数可选,默认为 1
VALUELABEL(varname)	字符型	返回变量值的值标签,如果没有值标签,则返回空字符串
APPLYMODEL(handle, "func", value)	数值型	使用句柄指定的模型将特定得分函数应用于输入个案数据,其中“function”是以下字符串文本值之一,以引号括起: predict, stddev, probability, confidence, nodeid, cumhazard, neighbor, distance。模型句柄是与外部 XML 文件相关联的名称,在 MODEL HANDLE 命令中定义。当功能是“概率”、“邻元素”或“距离”时,应用可选的第三参数。对于“概率”,将其指定为其计算概率的类别。对于“邻元素”和“距离”,将其指定为最近相邻模型的特定邻元素(作为整数)。如果无法计算值,则 APPLYMODEL 返回系统缺失值
STRAPPLYMODEL (han- dle, "func", value)	字符型	使用句柄指定的模型将特定得分函数应用于输入个案数据,其中“function”是以下字符串文本值之一,以引号括起: predict, stddev, probability, confidence, nodeid, cumhazard, neighbor, distance。模型句柄是与外部 XML 文件相关联的名称,在 MODEL HANDLE 命令中定义。当功能是“概率”、“邻元素”或“距离”时,应用可选的第三参数。对于“概率”,将其指定为其计算概率的类别。对于“邻元素”和“距离”,将其指定为最近相邻模型的特定邻元素(作为整数)。如果不能计算值,则 STRAPPLYMODEL 返回空字符串

附录2 各种情形下最常用统计检验方法索引^①

1. 单变量

连续	单样本 t 检验
有序多分类	单样本秩和检验
无序多分类	单样本卡方检验
二分类	二项分布确切概率法

2. 应变量:连续变量

单个自变量: 连续	相关分析, 回归分析
有序多分类	单因素方差分析, 解释结果时利用有序信息
无序多分类	单因素方差分析
二分类	两样本 t 检验
多个自变量: 连续变量为主	线性回归模型
分类变量为主	方差分析模型, 和回归模型实际上等价

3. 应变量:有序分类变量

单个自变量: 连续	有序分类的 Logistic 回归
有序多分类	秩相关分析、CMH 卡方
无序多分类	多样本秩和检验(H 检验)
二分类	两样本秩和检验(W 检验)
多个自变量: 连续变量为主	有序分类的判别分析, 有序分类的 Logistic 回归
分类变量为主	有序分类的 Logistic 回归

4. 应变量:无序分类变量

单个自变量: 连续	无序分类的 Logistic 回归
有序多分类	可将自/应变量交换后进行分析
无序多分类	卡方检验, 深入分析时可用对数线性模型
二分类	卡方检验
多个自变量: 连续变量为主	判别分析、无序分类的 Logistic 回归
分类变量为主	无序分类的 Logistic 回归

5. 应变量:二分类变量

单个自变量: 连续	两分类 Logistic 回归
有序多分类	可将自/应变量交换后进行分析
无序多分类	卡方检验, 两分类 Logistic 回归
二分类	四格表卡方检验, 确切概率法
多个自变量: 连续变量为主	判别分析、两分类 Logistic 回归
分类变量为主	两分类 Logistic 回归

① 这里给出的仅仅是各种情况下最常见的分析方法, 便于初学者选用, 并不意味着必须要使用相应的方法来分析。

6. 多元分析方法

考察的特征需要由多个应变量来表示,同时研究多个自变量对它们的影响:多元方差分析模型、多元回归模型。

希望将变量/记录分成若干个类别,但类别数不清楚,或各类别的特征不明:聚类分析。

已知分类情况,研究目的是希望建立判别方程,对之后新进入的案例进行所属类别的预测:判别分析。

需要探索多个连续变量间的内在联系或数据的内在结构:因子分析。

需要探索多个分类变量间的内在联系或数据的内在结构:对应分析。

考察多个概念间的相似程度,并寻找受访者用于评价相似性的标准:多维尺度分析。

生存时间和生存结局都是需要关心的因素,同时数据中存在大量的失访:生存分析。

得到的是时间序列数据,需要根据历史资料对之后的情形加以预测:时间序列模型。

附录3 统计术语英汉名词对照表

本对照表并不代表 SPSS 官方的意见,仅仅是为初学者提供的一份统计英文术语参考译名索引,考虑到国内统计界的使用习惯,许多术语在译法上和 SPSS 中文界面和输出方式都有所差异。

由于在不同行业内,统计术语的使用频率和翻译方式各不相同,因篇幅所限,本对照表中只提供了最为常用的术语。在可能有多种译法时,主要采用最为标准和常用的一种,并随后标出备选的其他译法,或者相应术语更为通俗的译法(但与字面含义无关)。如 censoring 对应的译法为删失、失访、终检。这表示以上 3 种译法均很常见,但以删失最为妥当。这种推荐次序仅仅是笔者个人的看法,供广大读者参考。

A			
accuracy	准确度	class mid - value	组中值
actual frequency	实际频数	cluster analysis	聚类分析
adjusted value	校正值	cluster sampling	整群抽样
alternative hypothesis	备择假设	coding	编码
analysis of covariance	协方差分析	coefficient of contingency	列联系数
analysis of variance, ANOVA	方差分析	coefficient of correlation	相关系数
arithmetic mean	算术平均数	coefficient of determination	决定系数
asymmetric distribution	非对称分布	coefficient of partial correlation	偏相关系数
autocorrelation	自相关	coefficient of product - moment correlation	积差相关系数
B		coefficient of rank correlation	等级相关系数
bar chart	条图	coefficient of regression	回归系数
bayes' theorem	Bayes 定理	coefficient of skewness	偏度系数
bias	偏倚, 偏性	coefficient of variation	变异系数
binominal distribution	二项分布	cohort study	队列研究
bivariate normal distribution	双变量正态分布	communality variance	公共方差
block	区组	comparability	可比性
box plot	箱图, 箱线图	complete association	完全相关
C		complete random design	完全随机设计
canonical correlation	典型相关	conditional likelihood	条件似然
case - control study	病例一对照研究	conditional probability	条件概率
categorical variable	分类变量	confidence interval, CI	可信(置信)区间
cell	单元	confidence limit, CL	可信(置信)限
censored data	截尾数据	confirmatory factor analysis	验证性因子分析
censoring	删失, 失访, 终检	confirmatory research	证实性研究
central limit theorem	中心极限定理	conjoint analysis	联合分析
central tendency	集中趋势	consistency test	一致性检验
chance error	随机误差	constraint	约束
		contingency table	列联表(R × C 表)

contribution rate	贡献率
control	对照
controlled experiments	对照实验
correction	校正
correction for continuity	连续性校正
correlation	相关
correlation analysis	相关分析
correlation coefficient	相关系数
correspondence analysis	对应分析
counts	计数/频数
covariance	协方差
Cox regression	Cox 回归
criteria for fitting	拟合准则
critical value	(临)界值
cross - over design	交叉设计
cross - section analysis	横断面分析
cross - section survey	横断面调查
crosstabulation table	交叉表
crosstabs	交叉表
cumulative frequency	累计频数
cumulative probability	累计概率
curve fit	曲线拟合
curvilinear regression	曲线回归
D	
data reduction	数据缩减
data transformation	数据变换
dataset	数据集
degree of freedom	自由度
degree of reliability	可靠度
density function	密度函数
dependent variable	应变变量
deviation	离差
discrete variable	离散型变量
discriminant analysis	判别分析
distribution	分布
distribution - free method	任意分布方法, 分布自由方法
dose response curve	剂量反应曲线
dummy variable	哑变量,虚拟变量
E	
eigenvalue	特征值

eigenvector	特征向量
equivariance	等方差
error	误差/错误
error of estimate	估计误差
estimated value	估计值
euclidean distance	欧氏距离
event	事件
expected values	期望值
experiment design	实验设计
exploratory data analysis	探索性数据分析
exponential curve	指数曲线
extrapolation	外推法
extremes	极端值,极值
F	
F distribution	F 分布
factor analysis	因子分析
factor score	因子得分
factorial	阶乘
factorial design	析因试验设计
false negative	假阴性
false positive	假阳性
finite population	有限总体
fitted value	拟合值
fitting a curve	曲线拟合
forecast	预测
fourfold table	四格表
frequency	频数
frequency distribution	频数分布
G	
general liner model, GLM	一般线性模型
generalized liner model	广义线性模型
geometric mean	几何均数
goodness of fit	拟合优度
H	
half - life	半衰期
harmonic mean	调和均数
hazard function	风险函数
hazard rate	风险率
heterogeneity	不同质
heterogeneity of variance	方差不齐

heteroscedasticity	方差不齐
hierarchical clustering method	分层聚类法
histogram	直方图
homogeneity	同质, 齐性
homogeneity of variance	方差齐性
homogeneity test	齐性检验
homoscedasticity	方差齐性
hypothesis test	假设检验
hypothesis testing	假设检验
I	
independence	独立性
independent variable	自变量
initial mean vectors	初始凝聚点
interaction	交互作用
intercept	截距
interpolation	内插法
inter - quartile range	四分位数间距
interval estimation	区间估计
inverse matrix	逆矩阵
iteration	迭代
K	
K means method	K - 均值聚类法
Kaplan - Merier curve	Kaplan - Merier 曲线
kendall 's rank correlation	Kendall 等级相关
Kolmogorov - Smirnov test	K - S 检验
Kruskal and Wallis test	K - W 检验, H 检验
Kurtosis	峰度
L	
lack of fit	拟合劣度, 失拟
Latin square design	拉丁方设计
least square method	最小二乘法
legend	图例
level	水平
level of significance	统计意义水平
life table	寿命表
likelihood function	似然函数
likelihood ratio test	似然比检验
line graph	线图
linear	线性
linear correlation	直线相关

linear equation	线性方程
linear programming	线性规划
linear regression	直线回归
linear trend	线性趋势
loading	载荷
log - rank test	时序检验
logarithmic scale	对数尺度
logistic regression	logistic 回归
logit transformation	logit 转换
loglinear model	对数线性模型
M	
main effect	主效应
matched data	配对资料
matching	配对
maximum likelihood method	最大似然法
maximum likelihood ratio test	最大似然比检验
mean	均数
mean square, MS	均方
measurement bias	测量性偏倚
median	中位数
median effective dose	半数效量
median lethal dose	半数致死量
median survival time	中位生存时间
median test	中位数检验
M - estimators	M 估计量
minimum lethal dose	最小致死量
missing value	缺失值
multidimensional scaling analysis, MDS	多维尺度分析
multinomial distribution	多项分布
multiple comparison	多重比较
multiple correlation	复相关, 多重相关
multiple covariance	多元协方差
multiple linear regression	多重线性回归
multiple response	多重应答, 多选题
multi - stage sampling	多阶段抽样
multivariate regression	多元回归
multivariate statistical analysis	多变量统计分析, 多元统计分析
N	
negative correlation	负相关

no statistical significance	无统计学意义
nominal variable	名义变量
nonlinear regression	非线性回归
nonparametric statistics	非参数统计
nonparametric test	非参数检验
normal distribution	正态分布
null hypothesis	无效假设
numerical variable	数值变量

O

observation unit	观察单位
observed value	观察值
odds ratio, OR	优势比, 比数比
one - sided test	单侧检验
one - way ANOVA	单因素方差分析
optimum allocation	最优分配
order statistics	顺序统计量
ordered categories	有序分类
orthogonal experimental design	正交试验设计
outlier	异常值
overall survey	普查

P

paired design	配对设计
paired(matched) t - test	配对 t 检验
parameter	参数
parametric statistics	参数统计
parametric test	参数检验
path analysis	路径分析
partial correlation	偏相关
partial likelihood	偏似然函数
partial regression coefficient	偏回归系数
percent bar graph	百分条图
percentage	百分比, 百分数
percentile	百分位数/位点
periodicity	周期性
pie graph	饼图, 圆图
placebo	安慰剂
point estimation	点估计
Poisson distribution	Poisson 分布
polynomial curve	多项式曲线
population	总体
population mean	总体均数

positive correlation	正相关
posterior distribution	后验分布
power of a test	检验效能
power of statistics	检验效能
precision	精度
principal component analysis	主成分分析
prior distribution	先验分布
product moment	乘积矩/协方差
product - limit method	乘积极限法
proportion	构成比
prospective study	前瞻性研究
P - value	P 值

Q

qualitative evaluation	定性评价
qualitative method	定性方法
quantile - quantile plot	QQ 图
quantitative analysis	定量分析
quantitative evaluation	定量评价
quartile	四分位数
questionnaire	问卷
quick cluster	快速聚类

R

random event	随机事件
random sampling	随机抽样
randomization	随机化
randomized allocation	随机分配
randomized block design	随机区组设计
randomized control trial	随机对照试验
randomized double blind control trial	随机双盲对照试验
range	极差(全距)
rank correlation	等级(秩)相关
rank sum test	秩和检验
ranked data	等级资料
rate	率
ratio	比
raw data	原始资料
regression analysis	回归分析
regression coefficient	回归系数
regression SS	回归平方和
relative number	相对数

relative risk, RR	相对危险度
reliability	可靠性, 信度
replacement level	更替水平
residual	残差
residual standard deviation	剩余标准差
residual sum of square	剩余平方和
ridge trace	岭迹
Ridit analysis	Ridit 分析
risk ratio	风险比, 危险比
rotation	旋转
R × C table	R × C 表
S	
sample	样本
sample size	样本量
sampling error	抽样误差
sampling fraction	抽样比
sampling study	抽样研究
sampling survey	抽样调查
scale	测量尺度
scatter diagram	散点图
score test	比分检验
screening	筛检
selection bias	选择性偏倚
semilogarithmic line graph	半对数线图
sequential design	序贯设计
sign test	符号检验
signed rank	符号秩
significance level	显著性水准
significance test	显著性检验
simple correlation	简单相关
simple regression	简单回归
skewness	偏度
slope	斜率
spearman rank correlation	等级相关
spherical distribution	球型分布
standard deviation, SD	标准差
standard error, SE	标准误差
standard normal distribution	标准正态分布
standardization	标准化
standardized partial regression coefficient	标准化偏回归系数

statistic	统计量
statistical control	统计控制
statistical graph	统计图
statistical inference	统计推断
statistical significance	统计学意义
statistical table	统计表
stem and leaf graph	茎叶图
stepwise method	逐步法
strata	层(复数)
stratification	分层
stratified cluster sampling	分层整群抽样
stratified sampling	分层抽样
structural equation modeling	结构方程模型
sum of squares	离差平方和
sum of squares of deviations from mean	离均差平方和
survey	调查
survival analysis	生存分析
survival curve	生存曲线
survival probability	生存概率
survival rate	生存率
survival time	生存时间
symmetry	对称
synthetic index	综合指数
synthetical evaluation	综合评价
systematic error	系统误差
systematic sampling	系统抽样
T	
t - distribution	t 分布
tendency of dispersion	离散趋势
test statistic	检验统计量
testing of hypotheses	假设检验
theoretical frequency	理论频数
time series analysis	时间序列分析
t - test	t 检验
two - sided test	双侧检验
two - stage least squares method	二阶段最小二乘法
two - stage sampling	二阶段抽样
two - step cluster	两步聚类法
two - tailed probability	双尾概率

two - tailed test	双侧检验	variance	方差
two - way ANOVA	两因素方差分析	variance component estimation	方差成分估计
two - way table	双向表	varimax orthogonal rotation	方差最大化正交旋转
type I error	I 类错误	W	
type II error	II 类错误	weight	权重
U		weighted linear regression method	加权直线回归
unbiased estimate	无偏估计	Z	
uniform distribution	均匀分布	weighting method	加权法
upper limit	上限	zero correlation	零相关
u test	u 检验	z - transformation	标准正态(Z)变换
V			
variable	变量		

附录4 IBM SPSS Statistics 19/20 介绍

1. 产品简介

IBM SPSS Statistics 统计分析软件是一款在调查统计行业、市场研究行业、医学统计、政府和企业的数据分析应用中久享盛名的统计分析工具,是世界上最早的统计分析软件,由美国斯坦福大学的3位研究生于1968年研制,并于1975年在芝加哥成立了SPSS公司总部,1984年SPSS首先推出了世界上第一个统计分析软件微型计算机版本SPSS/PC+,极大地扩充了它的应用范围,20世纪90年代,SPSS又推出了Windows版本,从SPSS 5.0开始,一直到现在的SPSS 20.0,它的功能一直在不断地增强、改进以满足不同客户的需求。世界上许多有影响的报纸杂志纷纷就SPSS的自动统计绘图、数据的深入分析、使用方便、功能齐全等方面给予了高度的评价与称赞。

在国际学术界有条不紊的规定,即在国际学术交流中,凡是用SPSS软件完成的计算和统计分析,可以不必说明算法,由此可见其影响之大和信誉之高。

迄今IBM SPSS Statistics软件已有40余年的成长历史。全球约有28万家产品用户,它们分布于通信、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,是世界上应用最广的统计分析软件。

2. 产品功能

IBM SPSS Statistics是一种权威的统计学工具,基本功能包括数据管理、统计分析、图表分析、输出管理等。由于其简单易用、界面友好,拥有强大的统计分析能力、数据管理功能、方便的图标展示功能,以及广阔的兼容性,满足了广大用户的需求,深受广大应用统计分析人员的喜爱。

SPSS统计分析软件是一款按照功能模块进行配置的产品,每个模块都可以独立安装和运行,或者是几个模块组合在一起。每个模块都拥有数据访问、数据管理和绘图功能。

(1) 更方便地访问和分析大型数据资料:IBM SPSS Statistics可以更快地访问、管理和分析任何类型的数据集,包括调查数据、公司数据、或者其他Web上下载的数据。用户可以让服务器去做繁重的计算工作,从而以尽可能快的速度进行分析。

(2) 轻松快捷地为分析做数据准备:IBM SPSS Statistics数据准备工作简单易行。

(3) 利用增强的报告能力更轻松地创建图表:IBM SPSS Statistics可实现多种图形功能,包括条形图、线型图、面积图、高低图、箱线图、散点图等,同时从IBM SPSS Statistics 17.0版本开始提供全新的图表创建界面,能够更轻松地创建常用的图表,同时预览将要生成的图表。利用图形生成语言(GPL),高级用户能够创建更多图表。

(4) 用户可设计自定义对话框构建程序,自定义对话框构建程序使普通用户能快速高效地学会常规分析操作,并为程序员(高级用户)高效部署他们自己的工作提供一条便捷的途径。

因此,在安装和使用方面,用户可以有更加灵活的选择。

3. 易用性

IBM SPSS Statistics是最早采用图形菜单驱动界面的统计软件,它的操作界面极为友好,输出结果清晰、美观,整个系统易学易用,只要对统计分析原理有基本的了解,就可以使用,是非专业统计人员的首选统计软件。

同时IBM SPSS Statistics使用Windows的窗口展示各种管理和分析数据的功能,使用对话框

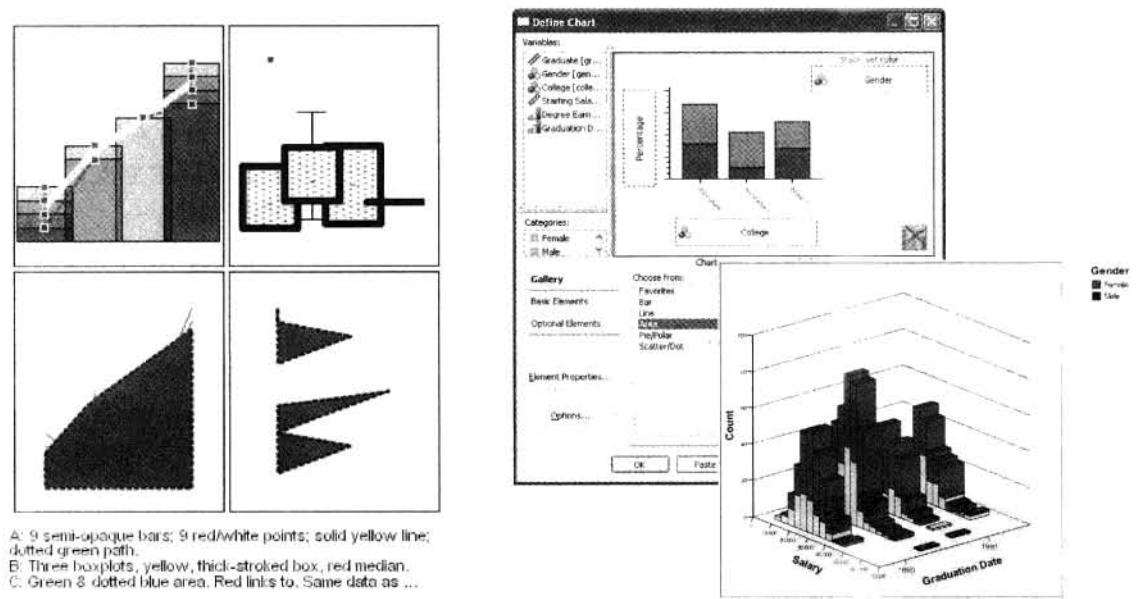
方式展示各种功能选项,它可以直接读取 Excel 和 DBF 数据文件,数据分析结果可以保存为多种文件格式,包含 Word、PPT、Excel、网页等格式,方便用户使用分析结果。另外,统计过程包括了常用的、较为成熟的统计过程,完全可以满足非统计专业人士的工作需要。

对于高级用户,IBM SPSS Statistics 提供了强大的程序语言,通过使用语法语言可使分析过程自动化、标准化。菜单操作过程能自动生成对应的操作程序,可以直接“粘贴”运行,供广大使用者学习、参考。

除此之外,IBM SPSS Statistics 还有很好的联机帮助系统,包括在线帮助、在线教程、学习向导、对话框帮助和语法手册。可以随时随地为不同层次的用户提供帮助。

4. 强大的统计分析功能

IBM SPSS Statistics 具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。自带 11 种类型 136 个函数,SPSS 提供了从简单的统计描述到复杂的多因素统计分析方法。IBM SPSS Statistics 统计分析过程包括描述性统计、均值比较、一般线性模型、方差分析、相关分析、回归分析、判别分析、聚类分析、因子分析、生存分析、时间序列分析、RFM 分析等几大类,每类中又包括几个统计过程,而且每个过程中又允许用户选择不同的方法及参数。IBM SPSS Statistics 也有专门的绘图系统,一种高度可视化的构造图表交互界面——图形构建程序,用户只需拖、拉各类图形要素,就可以根据用户自己的需求绘制各种各样的统计图形,如附图所示。



附图 统计图形

5. 兼容性与系统扩展能力

(1) 免费提供 SPSS Data Access Pack,其包含了当前流行数据库的 ODBC 和 OLE DB 驱动程序。

(2) 数据输入:Excel、Lotus、Oracle、SQL Server、Access、dBASE、文本以及各类统计软件的数据形式 SAS、Stata 文件。

(3) 数据输出: Word、HTML、XML、Excel、PowerPoint、PDF, 数据可回写到数据库中。

IBM SPSS Statistics 产品提供高度的可扩展性, 用户可以随着数据量和业务量的加大选择增加硬件配置来提高运算速度、升级服务器支持的 CPU 数目或者选择增加客户端来增加分析终端。

6. IBM SPSS Statistics 19 的新特性

(1) 广义线性混合模型: 在 Advanced Statistics 模块中增加了更多模型, 在因变量与自变量呈非线性关系的情况下, 对因变量的预测将更为精确。

(2) 自动线性模型: 在 IBM SPSS Statistics Base 及 IBM SPSS Statistics Server 中以更简单、更自动的方式建立线性模型。

(3) 新增评分功能: IBM SPSS Statistics 提供了用来构建预测模型(如回归、聚类、树和神经网络模型)的过程。构建模型后, 可以将模型规范保存在文件中, 该文件包含重建模型所需的所有信息。然后就可以使用该模型文件在其他数据集中生成预测得分。

(4) 支持更广的平台: IBM SPSS Statistics Server 19 支持 IBM System z, 可以为使用者提供强大的执行能力和优良的扩展性。

(5) 更智能更直观的结果输出: 更加直观地显示模型结果。

7. IBM SPSS Statistics 20 的新特性

(1) 地图: 统计地图功能现在以全新的面貌重新提供。图形画板模板选择器现在包含用于创建不同类型的地图直观表示的模板, 例如分区图(着色地图)、带有微型图表的地图和重叠地图等。IBM SPSS Statistics 附带了一些地图文件, 但用户可以使用地图转换实用程序来转换现有的外部地图文件以将其用于图形画板模板选择器。

(2) 更快呈现枢轴表: 枢轴表现在比以前版本中更快地呈现, 同时保持对透视和编辑操作的完全支持。如果使用过版本 19 的轻量表快速呈现功能, 会在版本 20 及更高版本中得到相当的枢轴表结果, 并且没有轻量表的限制。

(3) 在后台中无连接执行生产作业: 生产作业可以在远程服务器上的独立后台会话中运行。用户可以从本地计算机上提交作业, 断开与远程服务器的连接, 稍后再重新连接并检索结果。用户无需保持 SPSS Statistics 在本地计算机上运行, 甚至也不需要保持用户本地计算机处于打开。

(4) 广义线性混合模型的有序目标: 广义线性混合模型过程现在按照具有有序测量级别的目标类别的顺序来使用信息。有序目标采用有序多项式分布进行建模, 并且目标通过多个累积关联函数之一来与因子和协变量线性相关。

参考文献

- [1] SPSS Inc. IBM SPSS Statistics 20 简明指南[M]. Chicago, Illinois, 2011.
- [2] SPSS Inc. IBM SPSS Statistics 20 Core System 用户指南[M]. Chicago, Illinois, 2011.
- [3] SPSS Inc. IBM SPSS Statistics Base 20[M]. Chicago, Illinois, 2011.
- [4] SPSS Inc. IBM SPSS Statistics 20 Command Syntax Reference[M]. Chicago, Illinois, 2011.
- [5] SPSS Inc. IBM SPSS Custom Tables 20[M]. Chicago, Illinois, 2011.
- [6] SPSS Inc. IBM SPSS Bootstrapping 20[M]. Chicago, Illinois, 2011.
- [7] SPSS Inc. IBM SPSS Regression 20[M]. Chicago, Illinois, 2011.
- [8] SPSS Inc. IBM SPSS Data Preparation 20[M]. Chicago, Illinois, 2011.
- [9] SPSS Inc. Presenting Data with SPSS Tables™: Advanced[M]. Chicago, Illinois, 2003.
- [10] SPSS Inc. Statistical Analysis Using SPSS[M]. Chicago, Illinois, 2001.
- [11] Efron B., Tibshirani R J. An introduction to the bootstrap[M]. New York: Chapman & Hall, 1994.
- [12] David F G. Business Statistics: A Decision - making Approach[M]. 北京: 中国统计出版社, 2003.
- [13] Kleinbaum D G, Kupper L L, Muller K E. Applied Regression Analysis and Other Multivariable Methods[M]. California: Brooks/Cole, 1998.
- [14] Sit V. Analyzing ANOVA Designs, Biometrics Information Handbook[M]. Canada: Province of British Columbia 1995.
- [15] Sahai H, Ageel M I. The Analysis of Variance: Fixed, Random and Mixed Models[J]. Birkhasuser, 2000.
- [16] Steel R G D, Torrie J H. Principles and Procedures of Statistics: A Biometrical Approach[M]. 2nd Edition. McGraw - Hill, 1980.
- [17] 张文彤, 闫洁. SPSS 统计分析基础教程[M]. 北京: 高等教育出版社, 2004.
- [18] 张文彤. SPSS 统计分析高级教程[M]. 北京: 高等教育出版社, 2004.
- [19] 张文彤. SPSS 11 统计分析教程(基础篇)[M]. 北京: 北京希望电子出版社, 2002.
- [20] 张文彤. SPSS 11 统计分析教程(高级篇)[M]. 北京: 北京希望电子出版社, 2002.
- [21] Ronald M W. 商务统计导论[M]. 北京: 北京大学出版社, 2003.
- [22] Robert D M, Douglas A L. 商务经济统计方法(英文版)[M]. 9 版. 北京: 机械工业出版社, 1998.
- [23] 吴喜之. 统计学基本概念和方法[M]. 北京: 高等教育出版社, 2003.
- [24] 杨树勤. 中国医学百科全书·医学统计学分册[M]. 上海: 上海科学技术出版社, 1982.
- [25] 杨树勤. 卫生统计学[M]. 3 版. 北京: 人民卫生出版社, 1995.
- [26] 方积乾. 卫生统计学[M]. 5 版. 北京: 人民卫生出版社, 2003.
- [27] 曹素华, 赵耐青. 卫生统计学方法[M]. 上海: 复旦大学出版社, 2003.
- [28] 缪铨生. 概率与数理统计[M]. 2 版. 上海: 华东师范大学出版社, 1997.
- [29] 盛骤, 谢式千, 潘承毅. 概率论与数理统计[M]. 2 版. 北京: 高等教育出版社, 2000.
- [30] 茆诗松, 周纪芑. 概率论与数理统计[M]. 北京: 中国统计出版社, 1996.
- [31] Freedman D, 等. 统计学[M]. 魏宗舒, 等译. 2 版. 北京: 中国统计出版社, 1997.
- [32] 陆守曾. 医学统计学[M]. 北京: 中国统计出版社, 2002.
- [33] 茆诗松. 统计手册[M]. 北京: 科学出版社, 2003.

- [34] Anderson D R.,等. 商务与经济统计[M]. 张建华等译. 7版. 北京:机械工业出版社,2000.
- [35] 周润兰,喻胜华. 应用概率统计[M]. 北京:科学出版社,1999.
- [36] 何灿芝. 概率统计学习指导[M]. 长沙:湖南科学技术出版社,1984.
- [37] 吴喜之. 非参数统计[M]. 北京:中国统计出版社,1999.
- [38] 吴喜之,王兆军. 非参数统计方法[M]. 北京:高等教育出版社,1996.
- [39] 方开泰,金辉,陈庆云. 实用回归分析[M]. 北京:科学出版社,1988.
- [40] 陈希孺. 数理统计学简史[M]. 长沙:湖南教育出版社,2002.
- [41] 任仕泉. 非独立数据统计分析方法及其医学应用[D]. 四川:华西医科大学,1999.
- [42] 潘晓平,倪宗瓚,殷菲. 一种稳健的方差齐性检验方法[J]. 现代预防医学,2002,29(6):774-776.
- [43] 刘彤. 利用非参数方法对上海股市周末效应的研究[J]. 数理统计与管理,2003,1:28-32.
- [44] 拉里科夫(Rurikov, Y.). 仅仅依靠爱情?[J]. 文学杂志,1974,(29):13-28.
- [45] INTAGE 中国消费者信心调研[EB/OL]. <http://www.intage-china.com/product-service.php?id=102>.
- [46] 密歇根大学消费者信心指数主页[EB/OL]. <http://www.sca.isr.umich.edu/>.
- [47] CRISP-DM 方法论主页[EB/OL]. <http://www.crisp-dm.org/>.

郑重声明

高等教育出版社依法对本书享有专有出版权。任何未经许可的复制、销售行为均违反《中华人民共和国著作权法》，其为人将承担相应的民事责任和行政责任；构成犯罪的，将被依法追究刑事责任。为了维护市场秩序，保护读者的合法权益，避免读者误用盗版书造成不良后果，我社将配合行政执法部门和司法机关对违法犯罪的单位和个人进行严厉打击。社会各界人士如发现上述侵权行为，希望及时举报，本社将奖励举报有功人员。

反盗版举报电话 (010)58581897 58582371 58581879

反盗版举报传真 (010)82086060

反盗版举报邮箱 dd@hep.com.cn

通信地址 北京市西城区德外大街4号 高等教育出版社法务部

邮政编码 100120