

SAS 系统
与基础统计分析

北
京
大
学

965
36
96

SAS 系统 与基础统计分析

高惠璇 李东风 耿直
栾世武 张平 编译

北京大学出版社

SAS系统与基础统计分析

高惠璇 李东风 耿直
栾世武 张平 编译

北京大学出版社
北 京

内 容 简 介

本书是统计应用 SAS 丛书中的一本,它介绍了如何利用 SAS 系统进行基础统计分析。

本书具体介绍了几种基本的统计方法,并结合许多实际例子,给出了利用 SAS 系统进行这些分析所使用的 SAS 程序,最后对输出结果给予详细地解释。

本书的重点在数据分析上。介绍了如何把数据读入计算机;一些基本统计概念和本书所涉及统计方法的基础知识;如何对数据进行组间比较;如何对数据拟合回归线及检验回归模型拟合效果的基本方法;最后介绍了分析表格数据的基本方法。

本书适用于 SAS 系统的初学者、各领域中对统计分析方法了解不多的实际工作者以及非统计专业的大学生、研究生和教师。

图书在版编目(CIP)数据

SAS 系统与基础统计分析 / 高惠璇等编译. —北京:北京大学出版社,1995. 11
ISBN 7-301-02938-1

I. SAS… II. 高… III. 统计分析-数据处理-计算机应用 IV. C813

书 名: SAS 系统与基础统计分析

著作责任者: 高惠璇 李东风 耿直 栾世武 张平

责任编辑: 杨锡林

标准书号: ISBN 7-301-02938-1/TP·273

出版者: 北京大学出版社

地 址: 北京市海淀区中关村北京大学校内 100871

电 话: 出版部 2502016 发行部 2559712 编辑部 2502032

排 印 者: 北京经纬印刷厂印刷

发 行 者: 北京大学出版社

经 销 者: 新华书店

787×1092 毫米 16 开本 14.875 印张 372 千字

1995 年 11 月第一版 1995 年 11 月第一次印刷

定 价: 25.00 元

前 言

在当今的信息时代,人们在实践中收集到大量的信息。如何存储、整理和分析处理这些资料是一件很重要又是很复杂的工作。领先世界的统计分析系统——SAS 系统为您提供了灵活方便、功能齐全的整理数据、分析数据及报告结果的软件系统。

SAS 系统自美国 SAS 软件研究所(SAS Institute Inc.)于 1976 年商品化以来,就以其超凡的功能和可靠的技术支持著称于世。经过多年来的完善和发展,SAS 系统在国际上已被誉为统计分析的标准软件,在各个领域中得到了广泛的应用。

SAS 系统是集数据管理、数据分析和信息处理于一体的应用软件系统。它的最大特点是能以方便、直观的方式,有效地将各种数据转化为概括性的、有用的信息,为您提供准确、形象的分析结果,为您计划管理、预测和决策提供充分的信息。

SAS 系统是一种集成软件,用户可以将各种模块适当组合以满足各自不同的要求,使用户可以最大限度地利用计算机资源,以迎接日益增长的信息挑战。不管您用的是哪种型号的计算机,您都可以方便地安装运行 SAS 系统。

北京大学概率统计系自 1987 年在评估国际上几种统计与经济分析软件的基础上,对 SAS 系统进行了研究。并于 1990 年以来连续几年举办 SAS 系统学习班,参加的人数逐期增加。为了加快 SAS 系统在国内的推广使用,提高计算机应用水平和信息处理方面的工作效益和质量,在 SAS 软件研究所北京办事处的大力支持下,我们结合多年来使用 SAS 系统的体会,编译出版了《SAS 系统与基础统计分析》这本书,把它奉献给广大的用户。

本书是统计应用 SAS 系列丛书中的一本,主要讲述如何用 SAS 系统进行基础统计分析。内容浅显易懂,统计概念清楚严谨。本书适用于各个领域希望用 SAS 系统进行统计分析的人们,对于 SAS 新用户及非统计专业的实际工作者更是一本不可多得的好书。

本书由高惠璇负责组织编译、审校和排版工作。其中的第一至第四章由高惠璇、张平编译;第五至第六章由栾世武编译;第七至第八章由李东风编译;第九至第十章、附录 B 和附录 C 由高惠璇编译;第十一章由耿直编译;附录 A 由栾世武、高惠璇编译。参加审校工作的有高惠璇、李东风、栾世武和宫明。

本书在编译过程中得到了 SAS 软件研究所北京办事处的大力支持;得到了汪仁官和陈家鼎教授的热情指导。在此我们表示衷心的感谢。

编译者

1995 年 7 月

目 录

第一章 简介	1
§ 1.1 SAS 系统简介	1
1.1.1 什么是 SAS 系统	1
1.1.2 SAS 系统的特点	1
§ 1.2 本书简介	4
1.2.1 本书面向的对象	4
1.2.2 本书内容的安排	4
1.2.3 怎样使用本书	6
1.2.4 需要 SAS 系统的哪些部分	6
第二章 SAS 系统初步	7
§ 2.1 对所使用的计算机应了解的知识	7
2.1.1 你使用的计算机的类型	7
2.1.2 如何与计算机进行通信	7
2.1.3 怎样使用文本编辑器	8
2.1.4 怎样得到输出结果,在哪里打印	8
§ 2.2 对 SAS 软件应了解的知识	8
2.2.1 SAS 软件的顾问是谁	8
2.2.2 如何进入 SAS 系统	8
2.2.3 怎样使用 SAS 软件	9
§ 2.3 SAS 系统初步	9
2.3.1 SAS 系统结构	10
2.3.2 SAS 系统中句法和间隔的约定	10
2.3.3 SAS 软件产生的输出	11
2.3.4 SAS 软件的版本	11
2.3.5 SAS 程序的简例	12
2.3.6 介绍几个 SAS 语句	14
§ 2.4 总结	16
2.4.1 关键想法	16
2.4.2 句法	17
2.4.3 例子	17
第三章 创建 SAS 数据集	18
§ 3.1 什么是数据集	18
§ 3.2 掌握 SAS DATA 步	18
3.2.1 DATA 步概述	19
3.2.2 给变量命名	20
3.2.3 DATA 语句	20

3.2.4	INPUT 语句	20
3.2.5	CARDS 语句	21
3.2.6	数据行	22
3.2.7	空语句	22
§ 3.3	创建超速行驶罚款的数据集	22
§ 3.4	打印数据集	22
§ 3.5	数据集排序	24
§ 3.6	总结	26
3.6.1	关键想法	26
3.6.2	句法	26
3.6.3	例子	26
§ 3.7	进一步的专题	27
3.7.1	标记变量	27
3.7.2	变量值格式化	28
3.7.3	标记变量和格式化值的联合使用	31
§ 3.8	进一步的专题总结	33
3.8.1	关键想法	33
3.8.2	句法	33
3.8.3	例子	34
第四章	数据的概括	36
§ 4.1	量测水平	36
4.1.1	变量类型	36
4.1.2	响应尺度的类型	37
4.1.3	变量类型与响应尺度类型的对照	37
4.1.4	变量类型与概括方法的对照	38
§ 4.2	检查数据错误	38
§ 4.3	描述统计量	38
4.3.1	由 PROC UNIVARIATE 得到的描述统计量	39
4.3.2	由 PROC MEANS 得到的描述统计量	42
§ 4.4	频数表	43
4.4.1	由 PROC UNIVARIATE 得到的频数表	43
4.4.2	由 PROC FREQ 得到的频数表	44
4.4.3	用频数表检查错误	45
§ 4.5	条形图	45
4.5.1	用 PROC CHART 产生垂直条形图	46
4.5.2	用 PROC CHART 产生水平条形图	51
§ 4.6	探索性数据分析(EDA):简单图表	53
4.6.1	茎叶图	53
4.6.2	盒形图	55
4.6.3	EDA 图表小结	55
§ 4.7	总结	55

4.7.1 关键想法	55
4.7.2 句法	56
4.7.3 例子	57
第五章 基本统计概念	59
§ 5.1 总体和样本	59
5.1.1 定义	59
5.1.2 随机样本	60
5.1.3 参数和统计量	60
§ 5.2 正态分布	61
5.2.1 定义和性质	61
5.2.2 经验规则	63
5.2.3 正态性检验	64
§ 5.3 建立假设检验	68
5.3.1 统计显著性和实际显著性	71
5.3.2 参数和非参数统计方法	72
§ 5.4 总结	73
5.4.1 关键想法	73
5.4.2 句法	73
5.4.3 例子	73
第六章 均值的估计	75
§ 6.1 用单个数值估计总体均值	75
§ 6.2 样本容量的影响	75
6.2.1 减少样本容量	77
6.2.2 再次减少样本容量	77
§ 6.3 总体变异性的影响	81
6.3.1 标准差 σ 较小时均值的估计	81
6.3.2 标准差 σ 更小时均值的估计	82
6.3.3 估计误差的界限	87
§ 6.4 样本均值的分布	88
6.4.1 中心极限定理	88
6.4.2 经验规则和中心极限定理	90
§ 6.5 均值的置信区间	91
6.5.1 使用 SAS 软件计算置信区间(CIs)	92
§ 6.6 总结	94
6.6.1 关键想法	94
6.6.2 句法	94
6.6.3 例子	95
第七章 两组比较	96
§ 7.1 区分独立组与成对组	96
7.1.1 独立组	96
7.1.2 成对组	96

§ 7.2 从两个独立组概括数据	97
7.2.1 使用 PROC SORT 和 PROC MEANS	97
7.2.2 使用 BY 语句	99
7.2.3 用 PROC CHART 进行组间比较	99
7.2.4 概括独立组的其他方法	101
§ 7.3 概括成对观测数据	101
7.3.1 计算成对观测的差值	101
7.3.2 描述成对观测的差值	102
7.3.3 概括成对观测的其他方法	104
§ 7.4 建立两组比较的假设检验	104
7.4.1 如何选用统计检验	105
7.4.2 结果的解释	105
§ 7.5 进行两样本 t 检验	105
7.5.1 假定和假设检验	105
7.5.2 使用 PROC TTEST	106
7.5.3 PROC TTEST 小结	108
§ 7.6 进行 Wilcoxon 秩和检验	108
7.6.1 使用 PROC NPARIWAY	108
7.6.2 PROC NPARIWAY 小结	110
§ 7.7 进行成对差值的 t 检验	110
7.7.1 假定和假设检验	111
7.7.2 使用 PROC UNIVARIATE	111
7.7.3 成对差值 t 检验小结	112
§ 7.8 进行 Wilcoxon 符号秩检验	113
7.8.1 Wilcoxon 符号秩检验小结	114
§ 7.9 总结	114
7.9.1 关键想法	114
7.9.2 句法	114
7.9.3 例子	115
第八章 多组比较	117
§ 8.1 概括多组数据	117
8.1.1 使用 SORT 过程和 MEANS 过程	118
8.1.2 使用 SORT 过程和 UNIVARIATE 过程	120
8.1.3 使用 PROC CHART 比较几个组	120
§ 8.2 建立多组比较的假设检验	121
8.2.1 ANOVA 和 Kruskal-Wallis 检验	122
8.2.2 结果的解释	122
§ 8.3 进行单向方差分析	123
8.3.1 假定	123
8.3.2 使用 PROC ANOVA	123
8.3.3 PROC ANOVA 小结	125
§ 8.4 进行 Kruskal-Wallis 检验	126

8.4.1	假定	126
8.4.2	使用 PROC NPAR1WAY	126
8.4.3	PROC NPAR1WAY 小结	127
§ 8.5	多重比较过程	128
8.5.1	定义一些统计术语	128
8.5.2	用重复 t 检验控制 CER	129
8.5.3	用 Bonferroni t 检验控制 MEER	130
8.5.4	用 REGWF 检验控制 MEER	132
8.5.5	比较 VENEER 数据的结果	133
8.5.6	多重比较的更多知识	134
§ 8.6	总结	136
8.6.1	关键想法	136
8.6.2	句法	136
8.6.3	例子	137
第九章	相关与回归	139
§ 9.1	创建数据集	139
§ 9.2	绘制两个变量的散点图	140
§ 9.3	计算相关系数	141
9.3.1	使用 PROC CORR	142
9.3.2	关于相关的注意事项	143
9.3.3	相关不能回答的问题	143
§ 9.4	一元线性回归	143
9.4.1	回归分析解答的一些问题	144
9.4.2	最小二乘回归	144
9.4.3	使用 PROC REG 拟合直线	146
9.4.4	参数估计的置信界限	149
9.4.5	预测值和预测界限	149
9.4.6	估计均值和均值的置信界限	151
9.4.7	绘制回归线和置信界限的图形	152
9.4.8	一元线性回归的小结	154
§ 9.5	两个或多个自变量的回归	155
9.5.1	使用 PROC REG 进行多元回归	155
9.5.2	“显著性”的解释	156
9.5.3	一元线性回归与多元回归之间的差别	157
9.5.4	多元回归的小结	158
§ 9.6	曲线拟合	158
9.6.1	使用 PROC REG 拟合曲线	158
9.6.2	系数的显著性检验	161
9.6.3	预测值、预测界限和置信界限	162
9.6.4	绘制回归方程的图形	162
§ 9.7	PROC REG 在不同版本使用中的说明	163
9.7.1	不同版本差别的小结	164

9.7.2 使用 6.03 版本的 PROC REG	164
9.7.3 使用 6.03 以前版本的 PROC REG	167
§ 9.8 总结	169
9.8.1 关键想法	169
9.8.2 句法	169
9.8.3 例子	171
第十章 基本回归诊断	175
§ 10.1 诊断方法	175
§ 10.2 残差的特征	175
10.2.1 绘制残差对预测值的散点图	175
10.2.2 残差和异常点	176
10.2.3 序列趋势	177
§ 10.3 使用残差对预测值的散点图	177
§ 10.4 检验模型中是否需要增加一些项	179
10.4.1 绘制观测值和预测值的散点图	182
10.4.2 使用输出的残差值	182
§ 10.5 使用残差对自变量的散点图	182
10.5.1 指出需要二次项的例子	182
10.5.2 指出需要其他项的例子	185
§ 10.6 检查数据中的异常点	189
10.6.1 数据中含有异常点的例子	189
10.6.2 数据中没有异常点的例子	191
§ 10.7 时间序列的残差图	191
10.7.1 检查 KILOWATT 数据的序列效应	191
10.7.2 检查 ENGINE 数据的序列效应	193
§ 10.8 对 KILOWATT 数据和 ENGINE 数据进行回归诊断的小结	195
10.8.1 对 KILOWATT 数据进行回归诊断	195
10.8.2 对 ENGINE 数据进行回归诊断	195
§ 10.9 总结	196
10.9.1 关键想法	196
10.9.2 句法	196
10.9.3 例子	197
第十一章 交叉表的生成和分析	200
§ 11.1 列联表的生成	200
11.1.1 由原始数据生成列联表	201
11.1.2 由现有的表生成列联表	202
11.1.3 生成多个变量的交叉表	204
11.1.4 每页仅打印一张列联表	204
11.1.5 取消某些单元信息的打印	205
§ 11.2 检验分类变量间的独立性	206
11.2.1 执行卡方检验	206

11.2.2 执行 Fisher 精确检验	208
11.2.3 独立性检验的小结	208
§ 11.3 有序变量间的关联性量度	208
11.3.1 使用 PROC FREQ 计算 Kendall 的 Tau-b	209
§ 11.4 总结	211
11.4.1 关键想法	211
11.4.2 句法	211
11.4.3 例子	212
附录 A 错误诊断及其他	214
FA.1 程序中问题的快速检查方法	214
FA.2 无法使 SAS 软件运行	214
FA.3 出现有关句法错误的信息	214
FA.4 出现有关无效数据的出错信息	214
FA.5 出现不完全的格式化值	215
FA.6 出现一些“错误”的概括结果和一条出错信息	215
FA.7 得到总概括结果而不是分组概括结果	216
FA.8 产生“错误”的条形图	216
FA.9 没有获得方差分析结果	216
FA.10 没有进行多重对比检验	216
FA.11 无法在回归模型中增加变量	217
FA.12 用同样的 MODEL 语句却得到不同的回归结果	217
FA.13 列联表中没有列出变量的所有水平	217
FA.14 如何存储或调入 SAS 外部文件	217
FA.15 如何打印 SAS 程序或输出结果	218
FA.16 附加信息	218
附录 B SAS 术语汇编	219
附录 C 句法总结	221
参考文献	224

第一章 简介

§ 1.1 SAS 系统简介

1.1.1 什么是 SAS 系统

SAS(Statistical Analysis System)是一个由美国 North Carolina 州立大学于 1966 年开发的用于分析管理数据和编写报告的统计分析软件系统。1976 年,SAS 软件研究所(SAS Institute Inc.)成立,开始进行 SAS 系统的维护、开发、销售和培训工作。当时,SAS 统计分析系统只能运行在大型机上,1985 年又推出 SAS 的 VAX 机和 PC 机的版本。至 1990 年,SAS/PC 软件又推出了 6.04 版本,版本更新之快表明 SAS 的用户极为广泛。SAS 系统自 1976 年商品化以来,就以其超凡的功能和可靠的技术支持著称于世。经过多年来的完善和发展,SAS 系统在国际上已被誉为统计分析的标准软件,在各个领域中得到了广泛的应用。

SAS 系统是一种组合软件系统,其基本部分是 Base SAS 软件,在基本 SAS 软件的基础上,用户可以根据需要加上 SAS 系统的其他模块。比如用于统计分析的 SAS/STAT 软件,用于经济预测或时间序列分析的 SAS/ETS 软件,用于质量控制管理的 SAS/QC 软件,用于运筹学的 SAS/OR 软件,用于绘图的 SAS/GRAPH 软件,用于矩阵运算的 SAS/IML 软件等等。SAS 系统的核心是系统提供的大量灵活运用过程。这些过程可以直接使用,或稍加修改,以便满足用户的特殊需要。用户只要学会 SAS 系统,就可以替代你学习其他程序设计语言乃至其他统计软件包。SAS 系统可成为用户唯一需要的软件系统。

1.1.2 SAS 系统的特点

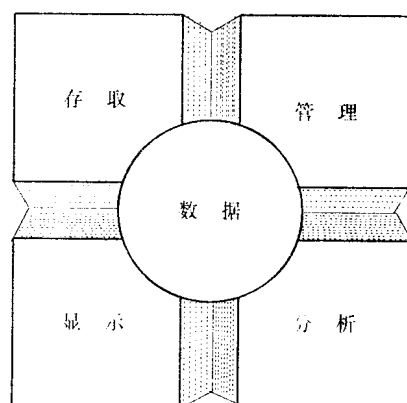
SAS 系统是一个功能非常齐全、应用极广、适用性很强、使用灵活、易于操作的综合软件系统。他的特点可归纳为以下几方面:

全面的管理数据

SAS 系统是集数据管理、信息处理和数据分析于一体的应用软件系统。SAS 系统提供了以数据为中心的四种基本处理功能。对于任何一批数据都必须对其进行四种基本的处理,才能得到有用又有意义的信息。这四种基本处理:存取、管理、分析和显示是以数据为中心的。这四种数据驱动处理对于开发业务的每种应用系统都是共同的。

□ **存取**——您需要快速且有效地获得数据。

SAS 系统能够存取用于应用系统的数据,无论他们存储在任何地方,以什么格式存在。



- **管理**——数据常常不会按您需要的格式提供给您,然而一旦使用 SAS 系统存取数据,在分析数据之前,您就可以利用他的数据管理功能对数据进行更新、重排、组合编辑或求数据子集,从而可以得心应手地分析数据。
- **分析**——通过数据分析揭示数据中蕴含的事实(信息),从而帮助您制定促进事业进一步发展的决策。SAS 系统所提供的分析方法包括从最简单的描述性统计到高级的或特殊的统计分析。
- **显示**——了解到数据中蕴含的这些信息仅仅是回答了问题的一部分,还需要把这些信息以简洁易懂的形式表示出来。数据的显示功能可在多种设备上实现,比如显示器、打印机或绘图仪等,其范围包括从简单的列表、表格到多维图形和精美的彩图。

灵活的用户界面

用户界面是指创建和运行 SAS 应用系统的一种方式。由于 SAS 系统提供了一系列灵活的界面,初学的用户和熟练的 SAS 程序员都可以全面地进行以数据为中心的四种基本处理。

SAS 系统提供的 SAS/ASSIST 软件是一种菜单驱动、面向任务的界面。SAS/ASSIST 菜单通过选择关键词来完成诸如:管理数据、打印报告或绘图等任务。由于 SAS/ASSIST 软件实际上用菜单选择的形式创建并存储了 SAS 程序,因此,SAS 系统的新用户不需要学习 SAS 语言的句法,就可以使用 SAS/ASSIST 界面对系统进行开发应用。较熟练的用户可使用 SAS/ASSIST 的程序作为特定应用系统的基础独立运行,或者在 SAS/ASSIST 环境下运行。

有经验的用户还可以使用 SAS 显示管理系统,这是一个方便的程序设计环境,也是一个交互式的窗口系统。通过他,您可以编写和修改程序,运行程序,检查产生的输出和信息。他使用方便的下拉菜单简化了通常的文件管理任务。

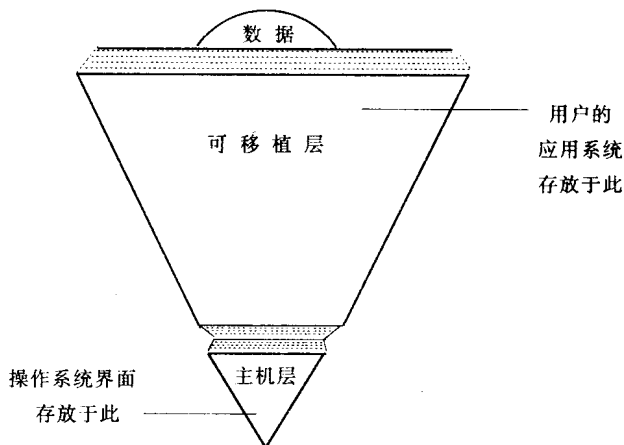
显示管理界面可以用户化。您可以利用菜单或在命令行上键入命令来显示窗口,变换他们的大小、形状和在屏幕上位置,编辑文本及搜索字词和短语。

SAS 系统提供了一个有效的文本编辑器,他可以方便地用于 SAS 系统中显示管理系统的固定窗口及一些其他的窗口环境中。可以用他来修改、重新整理或重新组合文本中的内容,或检查拼写和标记错误。

集成化的计算环境

由于 SAS 系统采用分层的多厂商结构(MultiVendor Architecture,简记为 MVA),故 SAS 应用系统是可移植的。SAS 系统的许多功能包含在可移植层,不依赖于硬件和操作系统;而主机层提供了计算机硬件和操作系统所需的界面。

应用系统存放在可移植部分。无论硬件和操作系统多么不同,不管您是使用什么型



号的计算机处理数据,SAS 应用系统的功能相同、显示相同且产生同样的结果。

主机层提供 SAS 系统和操作环境之间需要的所有接口。SAS 系统的主机层对每种操作环境是分别编写的,而且允许 SAS 系统使用各种操作系统提供的技术。这种层次结构意味着您所在的环境中使用 SAS 应用系统作业时能够使用一系列窗口、下拉菜单和图像功能。

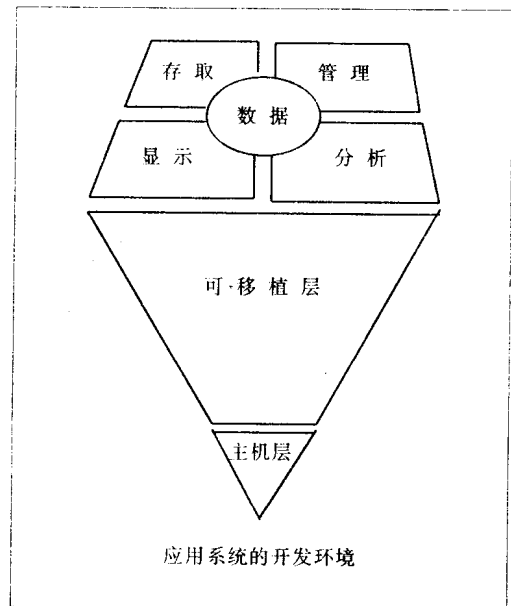
可移植的应用系统可以在任何计算环境中运行同样的应用系统。计算环境是由硬件及其运行的操作系统所决定的。您可以在某个环境下开发 SAS 应用系统,而在另一个环境下不必重新编写程序即可运行这个应用系统。

可以利用协作处理的优势。如果您所在的地区已经把几种不同型号的计算机联网,您就可以分享整个网络系统的数据和 SAS 应用系统。SAS 系统特有的 MVA 结构使应用系统保持可移植性,同时他又允许您的企业发挥每个特殊的硬件资源——从大型计算机到工作站以及个人计算机的所有优势。

完整的应用系统

SAS 系统支持应用系统的开发。SAS 系统具有强有力的程序设计语言和一些随时备用的被称为过程的程序。SAS 语言和他的过程,同 SAS 系统的其他特征结合在一起,使他成为一种几乎没有上限的可不断改进的应用系统。他包括通常的数据处理和众多应用领域中的非常特殊的分析。这种功能可以满足复杂的应用开发环境的需要。

SAS 系统为建立用户专用的,菜单驱动的应用系统提供了工具。应用系统开发者可以为用户组装完整的系统,从而用户只要输入一个指定的命令,就能够调用所要求的应用系统。为了开发应用系统,可以编写 SAS 程序并提供运行该系统的方式,而且经常增加一些用户交互运行应用系统的办法。应用开发环境能够——



☐ 控制一个交互程序的所有方面,包括制定菜单、处理错误、控制程序分支及用户化的信息和屏幕。

☐ 为批处理和作业设计前端菜单和数据输入屏幕,并使用交互会话过程提交给系统。

SAS 系统可以开发哪些类型的应用系统呢?由于 SAS 系统的组合模块设计使得他几乎可以在所有的领域中应用。因为 SAS 系统为用户提供了四种数据驱动处理的控制方式,可以满足您的任何一种需要。SAS 系统的所有模块能够按照用户的需要进行组合和再组合,以便解决所在机构的任何特定的任务。SAS 系统的这种组合模块结构保证能够按照需要的变化来增加功能。

SAS 系统为用户业务的每一方面提供处理功能。SAS 系统的部分功能包括:

☐ 决策支持

☐ 执行信息系统

☐ 财政金融分析和报告

☐ 市场分析和销售管理

☐ 计划管理

☐ 计算机性能评价

SAS 系统能为执行信息系统提供一种应用编码程序,这种编码程序是一种简便的交互式程序,他把目标程序连接在一起形成应用系统。利用这种应用编码程序,可以迅速地设计为决策制定提供必需信息的执行信息系统。

§ 1.2 本书简介

1.2.1 本书面向的对象

本书叙述怎样使用 SAS 系统完成基本的统计分析,同时对所使用的统计方法进行解释,包括什么情况下应使用哪种方法,假定了什么条件,怎样解释输出结果。

本书适用于在自己的专业领域内具有丰富的知识,但对于 SAS 系统或统计分析或两方面却了解很少的那些读者,包括管理人员、研究人员、企业中的分析化验员、以及学术界的大学生和教授。

作为一个面向任务的工具,本书介绍了几种基本的数据分析方法。他把分析方法与执行分析的 SAS 语句结合起来,然后对输出结果详细地给予解释。通过举例描述如何对这组数据进行分析以便得出结论。

本书还演示怎样使用 SAS 系统执行所涉及的数据分析方法。因为 SAS 软件提供的 SAS 语言是一种通用而且高效的语言,有时会发生这种情况:在 SAS 软件中某种分析可以用另一种方式完成。本书没有包含 SAS 软件进行某种特定分析的所有方式,而是给出执行那个分析任务的一种直接的方式。

因为本书是为向你提供一种解决数据分析问题的工具而编写,他的重点也放在这些问题上。因此,本书并没有列举 SAS 软件的全面功能,而是为读者提供解决问题所需的信息。

本书重点在数据分析上。还包括一些背景材料以使对于分析的讨论更有实际意义。虽然你只要使用本书提供的信息就可以进行数据分析,但对于所用方法的详细说明及更多的理论背景,你还可以参阅统计学教科书。除此之外,本书很少使用公式。书中所涉及统计方法的公式在许多教科书中以及 SAS 软件的其他文献中都是通用的。

本书侧重于使用 SAS 软件对数据进行统计分析,而不是研究方案的设计方法,几种实际的研究方案、设计方法方面的参考资料将在参考文献中给出。研究方案的设计与选择正确的分析方法同样重要。没有一种万能的统计分析方法能够解决设计时没有计划要解决的问题。如果读者对如何设计研究方案并不了解,建议你在开始工作之前请教统计学家或参阅参考文献中给出的资料。你还应该预先考虑数据的记录方式,有计划地用 SAS 系统易读的方式记录数据可以节省大量的时间。

本书中只介绍了几种基本的数据分析。比如除完全随机化设计以外的其他设计的方差分析、多元分析、时间序列方法、质量控制和运筹学书中就没有讨论。

1.2.2 本书内容的安排

本书分为以下十一章:

第一章 简介	第七章 两组比较
第二章 SAS 系统初步	第八章 多组比较
第三章 创建 SAS 数据集	第九章 相关与回归
第四章 数据的概括	第十章 基本回归诊断
第五章 基本统计概念	第十一章 交叉表的生成和分析
第六章 均值的估计	

第一至第四章说明怎样把数据读入计算机;第五、六章给出讨论第七至第十一章中分析方法时要用到的统计分析基础知识;本书其他部分说明怎样进行各种分析。章与章之间尽可能相互独立——你不必为了执行某一类分析而了解很多章的内容。

第一至第四章对于不熟悉 SAS 系统的读者来说十分重要,第一章是简单的介绍;第二章告诉读者使用 SAS 系统之前需要知道哪些知识,并对 SAS 系统做一介绍;第三章描述如何创建 SAS 数据集;第四章说明如何使用 SAS 软件来得到数据的描述性统计量、频数表、条形图和简单的散点图,并对所有的输出结果做详细的解释。

从第二章开始,在每一章最后的总结部分有三个小节:“关键想法”概括了该章的主要想法;每一章中讨论的所有 SAS 语句的一般形式在“句法”中做一小结,但有一个例外——创建 SAS 数据集的 SAS 语句只在第三章中说明;最后,在“例子”小节中列出产生该章所有输出的 SAS 语句。

第五、六章着重介绍统计概念和第七至第十一章要用到的统计分析基础知识,讨论的内容包括总体和样本、正态分布、经验规则、假设检验、样本容量和总体方差对均值估计的影响、中心极限定理和均值的置信区间。第五章叙述如何使用 SAS 软件进行正态性检验;第六章叙述怎样利用 SAS 软件计算均值的置信区间。

第七、八章说明怎样进行数据的组间比较。第七章描述两组间的比较;第八章介绍多于两组的比较。以下给出能够利用第七、八章的方法分析的各类问题的例子:

- ☐ 对于新雇员,在操作某种危险的和昂贵仪器前必须经过熟练程度检测,新雇员中有些能够熟练使用这种仪器,有些则不能。你希望发现这两种雇员(熟练和不熟练)中平均测试分数是否有差异。
- ☐ 测量骑脚踏车这种轻度运动前后的心率。一般认为同一个人运动之前比运动之后的心率慢,你想知道这种心率的变化是否比偶然发生的变化大。
- ☐ 作试验比较五种化肥对于盆栽天竺葵的作用。在六个星期末测量植物的高度,要比较施用五种不同化肥的植物的平均高度。
- ☐ 收集某流水线上从周一到周五的工作日中不合格品的数目,希望知道在一周的不同天中,产生的不合格品的数目是否不同。

第九、十章告诉我们对某个数据集如何用直线或曲线去拟合他。第九章给出了怎样进行相关分析和线性回归分析;第十章描述一些检查回归模型拟合效果的基本方法。以下是几个利用第九、十章中的方法能够分析的问题的实例:

- ☐ 为了确定孩子的体重与年龄间的关系,想拟合一条直线或曲线来定义这种关系。可以先绘制表示数据关系的散点图,然后用直线或曲线拟合,最后检验数据并找出离拟合线很远的异常点。
- ☐ 在某工艺制造过程中,改变了炉的温度,想确定由给定温度变化引起的产品硬度的变

化量,并对给定的温度预报未来产品的硬度。

第十一章叙述了分析表格数据的一些基本方法。以下列出利用这一章的方法可以解决的问题的例子:

- 收集男人和女人喜欢的香波颜色的数据,希望知道男人和女人喜欢的颜色的分布是否相同。
- 进行一次民意测验来征求某小镇居民对增加新税来建设运动场的意见如何。希望知道在所提的运动场场址附近居住的人是否与住得远的人意见相同。
- 公司正在试销一种新产品,做一项调查询问顾客是否购买了该产品,还设计了一系列问题以便了解人们对该产品包装的意见。希望从中发现使用过和没使用过该产品的人所持的意见是否不同的结论。

附录 A“错误诊断及其他”提供在 SAS 程序中出现的常见问题及处理方法;附加信息告诉你怎样从 SAS 软件研究所获得资料及培训目录,并描述了一些附加信息和服务的情况。这本书最后面是“SAS 术语汇编”、“句法总结”及“参考文献”。

1.2.3 怎样使用本书

本书中用到几种类型的印刷体,我们做以下约定:

书宋体——中文使用的基本字体。

英文斜体——用于表示 SAS 语句中用户应提供一些项目。

黑体——用于标题、输出的标题和表的标题。给出 SAS 语句的一般形式时,黑体用来表示必须按原样拼写的部分,除此之外,特别重要的短语和句子也用黑体给出。

本书并非统计教材的代用品,他只是在说明怎样使用 SAS 系统去完成各种统计分析的同时,描述了所使用的统计方法、SAS 语句和 SAS 输出。他并不能取代统计教科书。

如果读者以前从未使用过 SAS 软件,首先应该认真地阅读第一至第四章。在各章中,读者可以学到使用软件去创建和概括 SAS 数据集的基础知识。如果读者对统计也了解较少,则可参阅第五、六章。如果你对统计分析比较熟悉,则可跳过第五、六章而阅读本书的其他部分。“本书内容的安排”一节可能会有助于你选择解决问题所需的章节。

如果读者以前已使用过 SAS 软件,但对统计不熟悉,则可跳过第一至第四章。认真阅读第五、六章,这里对几个统计概念给出了解释。然后可根据你的需要选择有关的部分进行阅读。

如果读者以前既使用过 SAS 软件又熟悉统计方法,但不知怎样解决所遇到的特殊问题,则可跳过第一至第六章。直接阅读适合你需要的章节,了解怎样使用 SAS 软件解决自己的问题。

1.2.4 需要 SAS 系统的哪些部分

在所有机型中使用本书均需要 Base SAS 及 SAS/STAT 软件。