

SAS

统计分析应用

董大钧 主编

SAS

统计分析教材



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

<http://www.phei.com.cn>

附光盘



SAS 统计分析应用

董大钧 主编

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书基于 SAS 9.1.3 中文版编写,介绍了 SAS Learning Edition 基于窗口点击式环境以及 Base SAS、SAS\STAT 的用法。全书以统计分析方法为主线,通过大量实例,详细介绍了 SAS 程序设计方法及各种统计过程适用条件和使用方法,并对统计过程实例的输出结果做了详尽的解释。本书例题和数据均保存在所附光盘中,以方便读者学习。

本书既可作为高等院校本科生、研究生信息统计分析相关课程的教材,也是从事数据分析和数据管理的人员的参考用书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有,侵权必究。

图书在版编目(CIP)数据

SAS 统计分析应用/董大钧主编. —北京:电子工业出版社,2008.4
ISBN 978-7-121-06268-1

I. S… II. 董… III. 统计分析—应用软件,SAS IV. C812

中国版本图书馆 CIP 数据核字(2008)第 040107 号

责任编辑:杨丽娟 特约编辑:明足群

印 刷:北京市天竺颖华印刷厂

装 订:三河市金马印装有限公司

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本:787×1092 1/16 印张:31.5 字数:806 千字

印 次:2008 年 4 月第 1 次印刷

印 数:4000 册 定价:42.00 元(附光盘 1 张)

凡所购买电子工业出版社的图书,如有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至 zlt@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010)88258888。

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可,复制、销售或通过信息网络传播本作品的行为;歪曲、篡改、剽窃本作品的行为,均违反《中华人民共和国著作权法》,其行为人应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,保护权利人的合法权益,我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为,本社将奖励举报有功人员,并保证举报人的信息不被泄露。

举报电话:(010)88254396;(010)88258888

传 真:(010)88254397

E-mail: dbqq@phei.com.cn

通信地址:北京市万寿路173信箱

电子工业出版社总编办公室

邮 编:100036

编 委 会

主 编：董大钧

副主编：杨文万 郑凤仁 董 丽

编 委：刘树民 韩 松 刘寅生 谢进军

刘明信 杨 玥 田 丹 刘申菊

前 言

在信息时代的今天,人们在工作实践中会获取大量的信息。如何存储、整理和分析处理它们是一项极其重要的工作。由于对数据的分析大都是基于基本的统计原理进行的,国内外学者多年来编制了许多统计软件包,如 BMDP、SAS、SPSS 等。其中 SAS 软件包则是诸多统计软件包中的佼佼者。

SAS 系统全称为 Statistics Analysis System,是用于决策支持的大型集成信息系统,经过多年的发展,SAS 已被全世界 120 多个国家和地区的近三万家机构采用,直接用户则超过三百万人,遍及金融、医药卫生、生产、运输、通信、政府和教育科研等领域。在数据处理和统计分析领域,SAS 系统被誉为国际上的标准软件系统,是建立数据库的首选产品,堪称统计软件界的巨无霸。

SAS 系统具有灵活的功能扩展接口和强大的功能模块,在 Base SAS 的基础上,还可以增加不同的模块而实现不同的功能:SAS/STAT(统计分析模块)、SAS/GRAPH(绘图模块)、SAS/QC(质量控制模块)、SAS/ETS(经济计量学和时间序列分析模块)、SAS/OR(运筹学模块)、SAS/IML(交互式矩阵程序设计语言模块)、SAS/FSP(快速数据处理的交互式菜单系统模块)、SAS/AF(交互式全屏幕软件应用系统模块),等等。SAS 有一智能型绘图系统,能绘各种统计图。SAS 提供了许多统计过程,每个过程均含有极丰富的任选项。用户还可以通过对数据集的一连串加工,实现更为复杂的统计分析。此外,SAS 还提供了各类概率分析函数、分位数函数、样本统计函数和随机数生成函数,使用户能方便地实现特殊统计要求。

自 1993 年由董大钧教授主编的全国第一本高校 SAS 统计软件应用教材出版以来,SAS 软件被迅速推广,它大大地提高了用户的计算机应用水平和信息处理的工作效率和工作质量。

本书为高等院校计算机数据处理、统计分析课程的教科书,也可以作为 SAS 软件应用的指导书。本书介绍了 SAS 9.1.3 中文版的 SAS Learning Edition 窗口点击式环境用法及 Base SAS 和 SAS/STAT 的主要内容;通过大量实例详细阐述如何使用 SAS 语言编写程序及 SAS 统计软件包中许多基本统计过程的使用方法和结果的分析。

本书第 1 章和第 2 章由董丽编写,第 3 章由郑凤仁编写,第 12 章至第 15 章由杨文万编写,其余各章由董大钧编写。

由于我们知识和经验有限,加之时间仓促,本书难免存有各种问题,望读者提出宝贵的意见。

编 者

目 录

绪论	1
第 1 章 SAS Learning Edition	4
1.1 概述	4
1.1.1 关于 SAS Learning Edition	4
1.1.2 示例数据文件	5
1.1.3 启动 SAS Learning Edition	5
1.1.4 SAS Enterprise Guide 可访问的内容	9
1.2 创建项目并添加数据	10
1.2.1 打开项目设计器	10
1.2.2 向项目中添加数据	10
1.2.3 导入数据	16
1.3 运行 SAS 任务,创建数据报表	20
1.3.1 SAS 任务概述	20
1.3.2 创建报表	20
1.3.3 修改报表	23
1.3.4 更改任务属性,生成 PDF 输出	27
1.4 “查询生成器”任务	27
1.4.1 概述	27
1.4.2 使用“查询生成器”	28
1.4.3 按条件进行查询	36
1.4.4 从查询生成汇总表	38
1.5 图形	43
1.5.1 条形图	43
1.5.2 饼图	45
1.6 “线性模型”分析	47
1.6.1 概述	47
1.6.2 运行“线性模型”任务并将列分配至角色	47
1.6.3 指定模型	48
1.6.4 运行任务和查看结果	48
1.7 将多个报表组合成单个文档	50
1.7.1 创建 HTML 文档和添加结果	50
1.7.2 更改报表顺序和样式,预览报表	50
1.7.3 导出报表	52

1.8 使用过程流	52
习题	54
第2章 SAS 系统概述	55
2.1 SAS 窗口系统使用方法	55
2.1.1 SAS 系统的启动和退出	55
2.1.2 SAS 窗口环境	56
2.1.3 程序编辑器(Editor)窗口	59
2.1.4 日志(LOG)窗口	60
2.1.5 输出(OUTPUT)窗口	61
2.1.6 结果(Results)窗口	61
2.1.7 改变窗口显示	62
2.1.8 打印和保存 SAS 活动窗口中的内容	63
2.1.9 KEYS 窗口	63
2.2 SAS 逻辑库	65
2.2.1 SAS 逻辑库概述	65
2.2.2 创建逻辑库	65
2.2.3 用命令指定逻辑库引用名	67
2.2.4 删除逻辑库	67
2.2.5 创建逻辑库成员	67
2.2.6 数据集名	68
2.3 SAS 程序的输入与运行	68
2.3.1 SAS 程序结构	68
2.3.2 SAS 程序的输入修改和运行	70
2.4 SAS 编程基础	76
2.4.1 观测	76
2.4.2 变量	76
2.4.3 常量	78
2.4.4 函数	80
2.4.5 操作符	86
2.4.6 SAS 表达式	87
习题	88
第3章 建立数据集	90
3.1 建立数据表	90
3.1.1 用 Viewtable 编辑数据集	90
3.1.2 使用分析数据表	92
3.2 数据步基本语句	93
3.2.1 DATA 语句	93

3.2.2	INPUT 语句	96
3.2.3	利用指针控制读数据	102
3.2.4	DATALINES 和 DATALINES4 语句	106
3.3	数据步流程	107
3.4	提供信息的几个语句	108
3.4.1	LENGTH 语句	108
3.4.2	LABEL 语句	109
3.4.3	LOSTCARD 语句	110
3.4.4	INFORMAT \ FORMAT 语句	111
3.4.5	MISSING 语句	112
3.4.6	注释语句	113
3.4.7	OPTIONS 语句	113
3.5	建立数据集的另外几种途径	114
3.5.1	从外部文本数据文件中读入数据	114
3.5.2	从已建立的数据集中读入数据	115
3.5.3	从数据库文件中导入数据	116
	习题	120
第 4 章	整理数据集	122
4.1	数据集中的数据格式	122
4.2	建立、删除与保留变量	124
4.2.1	赋值语句建立变量	124
4.2.2	求和语句	125
4.2.3	RETAIN 语句	126
4.2.4	DROP 语句删除变量	127
4.2.5	KEEP 语句保留变量	128
4.3	条件语句	128
4.3.1	条件 IF 语句	128
4.3.2	求子集 IF 语句	132
4.4	删除观测	133
4.5	GOTO 语句	134
4.6	RETURN 语句	136
4.7	LINK 语句	136
4.8	循环	138
4.8.1	利用 GOTO 实现循环	138
4.8.2	DO/END 循环	139
4.8.3	DO WHILE 循环	140
4.8.4	DO UNTIL 循环	141

4.9	数组	142
4.9.1	数组与下标变量	142
4.9.2	在 DO 循环中使用数组	144
4.9.3	多维数组	145
4.10	数据排序	147
4.11	分组控制	149
4.12	数据集连接	150
4.12.1	变量相同	150
4.12.2	变量不同	150
4.12.3	交叉连接	151
4.13	数据集合并	151
4.13.1	一对一合并	152
4.13.2	匹配合并	152
4.14	输出信息	155
4.14.1	PUT 语句格式	155
4.14.2	输出形式	156
4.15	FILE 语句	158
4.16	OUTPUT 语句	160
4.16.1	在一个数据步中创建一个以上的数据集	161
4.16.2	从一个输入行中创建几个观测	162
4.16.3	从几个记录合并信息	162
4.17	DATASETS 过程	164
4.17.1	PROC DATASETS 语句	164
4.17.2	APPEND 语句	164
4.17.3	MODIFY 语句	165
4.17.4	FORMAT 语句	165
4.17.5	INFORMAT 语句	165
4.17.6	LABEL 语句	166
4.17.7	RENAME 语句	166
	习题	166
第 5 章	SAS 过程步中常用语句	168
5.1	PROC 语句	168
5.2	BY 语句	168
5.3	CLASS 语句	169
5.4	VAR 语句	170
5.5	WEIGHT 语句	170
5.6	FREQ 语句	171

5.7	ID 语句	172
5.8	OUTPUT 语句	173
5.9	TITLE 语句	174
5.10	OPTIONS 语句	175
	习题	175
第 6 章	SAS 实用过程	176
6.1	PRINT 过程	176
6.1.1	语句说明	176
6.1.2	应用举例	176
6.2	FORMAT 过程	178
6.2.1	概述	178
6.2.2	语句说明	179
6.2.3	举例	180
6.3	TRANSPOSE 过程	181
6.3.1	概述	181
6.3.2	语句说明	182
6.4	查询数据	186
6.4.1	概述	186
6.4.2	PROC SQL 语句	187
6.4.3	SELECT 语句	187
6.4.4	CREATE VIEW 语句	188
6.4.5	DELETE 语句	189
6.4.6	DROP 语句	189
6.4.7	举例	189
	习题	191
第 7 章	描述性统计过程	192
7.1	统计概念简介	192
7.1.1	总体和样本	192
7.1.2	定位测量	192
7.1.3	离散程度测量	193
7.1.4	形状测量	193
7.1.5	正态分布	194
7.1.6	样本均值的分布	195
7.1.7	置信度和置信区间	195
7.1.8	统计关键字	195
7.1.9	单变量统计对数据的要求	196
7.2	MEANS 过程	197

7.2.1	语句说明	198
7.2.2	举例	199
7.3	FREQ 过程	202
7.3.1	简介	202
7.3.2	语句说明	205
7.3.3	列联表卡方检验	207
7.3.4	举例	207
7.4	TABULATE 过程	211
7.4.1	简介	211
7.4.2	用菜单制作汇总表	214
7.4.3	语句说明	217
7.4.4	举例	219
7.5	UNIVARIATE 过程	220
7.5.1	简介	220
7.5.2	语句说明	221
7.5.3	应用举例	222
7.6	PLOT 过程	225
7.6.1	简介	225
7.6.2	语句说明	226
7.6.3	举例	228
7.7	CHART 过程	230
7.7.1	简介	230
7.7.2	语句说明	231
7.7.3	选择项说明	233
7.7.4	举例	236
7.8	RANK 过程	238
7.8.1	简介	238
7.8.2	语句说明	238
7.8.3	应用	240
7.8.4	举例	241
	习题	244
第 8 章	假设检验	245
8.1	假设检验简介	245
8.1.1	两样本均数差别的统计意义检验	245
8.1.2	t 检验及类型	245
8.1.3	关于非正态分布资料均数差别的检验	246
8.1.4	双侧检验和单侧检验	246

8.2	t 检验	247
8.2.1	用 MEANS 过程进行 t 检验	247
8.2.2	用 TTEST 过程进行 t 检验	249
8.3	NPARIWAY 过程	251
8.3.1	语句说明	251
8.3.2	举例	252
	习题	255
第 9 章	方差分析	256
9.1	概述	256
9.2	均衡数据的方差分析(ANOVA 过程)	258
9.2.1	ANOVA 过程的主要语句	258
9.2.2	语句说明	258
9.2.3	输出结果	260
9.3	ANOVA 过程应用实例	261
9.3.1	完全随机设计资料的方差分析(单因素方差分析)	261
9.3.2	随机区组设计资料的方差分析(两因素方差分析)	264
9.3.3	拉丁方资料的方差分析	268
9.3.4	析因实验设计——三因素实验分析	270
9.3.5	正交实验设计的方差分析	271
9.3.6	$R \times C$ 析因设计资料的方差分析(有交互因素的方差分析)	272
9.4	GLM 过程	274
9.4.1	概述	274
9.4.2	过程使用的语句	274
9.4.3	四种类型的可估计函数	276
9.4.4	举例	277
9.5	协方差分析	280
9.5.1	协方差分析简介	280
9.5.2	举例	280
	习题	285
第 10 章	相关与回归	288
10.1	概述	288
10.1.1	相关分析	288
10.1.2	回归分析	289
10.2	CORR 过程	290
10.2.1	语句说明	290
10.2.2	举例	291
10.3	CANCORR 过程	295

10.3.1	CANCORR 过程介绍	296
10.3.2	应用举例	297
10.4	REG 过程	303
10.4.1	REG 过程使用的语句	303
10.4.2	举例	306
10.5	STEPWISE 过程	310
10.5.1	STEPWISE 过程简介	310
10.5.2	STEPWISE 过程使用的语句	311
10.5.3	举例	312
10.6	非线性回归 NLIN 过程	315
10.6.1	非线性回归概述	315
10.6.2	过程说明	315
10.6.3	可能出现的问题	317
10.6.4	显示输出内容	318
10.6.5	举例	318
10.7	LOGISTIC 过程	321
10.7.1	Logistic 回归简介	321
10.7.2	模型的选择	322
10.7.3	LOGISTIC 过程使用的语句	323
10.7.4	使用注意	325
10.7.5	LOGISTIC 过程的应用	327
	习题	332
第 11 章	判别分析	335
11.1	判别分析概述	335
11.2	DISCRIM 过程	336
11.2.1	概述	336
11.2.2	DISCRIM 过程使用的语句	337
11.2.3	DISCRIM 过程的输出	340
11.2.4	应用举例	341
11.3	STEPDISC 过程	351
11.3.1	概述	351
11.3.2	STEPDISC 过程使用的语句	352
11.3.3	STEPDISC 过程应用举例	353
	习题	358
第 12 章	聚类分析	361
12.1	概述	361
12.1.1	聚类方法	361

12.1.2	SAS 提供的聚类过程	362
12.1.3	聚类过程的输出及聚类类数的确定	362
12.1.4	散点图和树形图	363
12.2	CLUSTER 过程	364
12.2.1	简介	364
12.2.2	语句说明	364
12.2.3	举例	366
12.3	FASTCLUS 过程	371
12.3.1	简介	371
12.3.2	过程语句说明	371
12.3.3	过程输出数据集说明	373
12.3.4	举例	373
12.4	VARCLUS 过程	379
12.4.1	简介	379
12.4.2	VARCLUS 过程使用的语句	379
12.4.3	举例	381
12.4.4	输出结果的进一步说明	385
12.5	TREE 过程	388
12.5.1	简介	388
12.5.2	过程使用的语句	389
12.5.3	补充说明	390
12.5.4	举例	391
	习题	395
第 13 章	因子分析	399
13.1	因子分析的简单原理	399
13.2	FACTOR 过程	400
13.2.1	FACTOR 过程所用的语句	400
13.2.2	输出内容	404
13.2.3	举例	406
13.3	PRINCOMP 过程	415
13.3.1	语句格式说明	415
13.3.2	举例	416
	习题	419
第 14 章	生存分析	428
14.1	概述	428
14.1.1	什么是生存分析	428
14.1.2	生存分析的过程	429

14.1.3 生存分析中使用的术语	429
14.2 LIFEREG 过程	430
14.2.1 简介	430
14.2.2 LIFEREG 过程使用的语句	430
14.2.3 补充说明	432
14.2.4 生存回归应用举例	432
14.3 LIFETEST 过程	439
14.3.1 简介	439
14.3.2 LIFETEST 过程用的语句	440
14.3.3 举例	441
14.3.4 单变量 Wilcoxon 和 logrank 检验	444
14.3.5 多变量的 Wilcoxon 和 logrank 检验	450
14.4 PHREG 过程	453
14.4.1 PHREG 过程概述	453
14.4.2 PHREG 过程使用的语句	454
14.4.3 实例	457
习题	459
第 15 章 分类资料分析	462
15.1 概述	462
15.2 CATMOD 过程的基本原理	463
15.2.1 单总体	463
15.2.2 多总体	463
15.2.3 参数估计及假设检验	464
15.3 CATMOD 过程语句	464
15.3.1 CATMOD 过程语句	464
15.3.2 关于设计矩阵	468
15.4 举例	469
习题	483
参考文献	486

绪 论

1. 什么是 SAS

SAS (Statistical Analysis System) 是一个用于决策支持的大型集成信息系统, 用来管理、分析数据和编写报告的组合软件系统。该软件系统最早的功能限于统计分析, 至今, 统计分析功能也仍是它的重要组成部分和核心功能。经过多年的发展, SAS 已被全世界 120 多个国家和地区的三万多家机构所采用, 直接用户超过三百万人, 遍及金融、医药卫生、生产、运输、通信、政府和教育科研等领域。在数据处理和统计分析领域, SAS 系统被誉为国际上的标准统计软件系统, 堪称统计软件界的巨无霸。对数据处理要求最严格的美国食品和药物管理局 FDA, 在进行新药报批时, 实验数据只有使用 SAS 软件处理才接受申请。

SAS 是唯一一家将领先的数据仓库、分析和传统的商业智能 (BI) 应用完全集成在一起的供应商, 可从海量数据中创造智能。SAS 提供了世界级的分析、数据操纵和报表展现能力, 它具有可伸缩性、高度的互用性和易管理性。

SAS 是一种模块化软件, 其基本部分是 Base SAS。在此基础上, 可以加上 SAS/STAT (统计)、SAS/GRAPH (图形) 等其他模块而组成一个专用数据处理系统。

Base SAS 软件是 SAS 提供的各种数据管理和分析软件组件的核心基础, 它是访问、管理、分析和显示数据的世界技术领先的信息传输系统。Base SAS 软件包括第四代编程语言和称为过程的现成程序。这些过程可处理数据、信息存储和检索、统计分析和报表编写等。

SAS/STAT 可用于高级统计分析和应用, 包括方差分析、回归分析、属性数据分析、多元分析、生存分析、聚类分析、判断分析及非参数分析等。

SAS/ACCESS 软件用于从数据库管理系统 (DBMS), 以及 PC 文件格式读取和写入数据。

SAS/GRAPH 以各种彩色图表展示信息, 完成各种用于业务管理和统计分析的中、外文图形报表, 包括直方图、饼形图、曲线图、三维图等。

SAS/ASSIST 提供任务导向式、菜单驱动用户友好界面, 利用图形式的选项画面, 自动生成完整的 SAS 程序。

SAS/OR 可作线性规划、项目管理、资源分配和决策支持。

SAS/IML 交互式矩阵语言, 可利用其内置的功能进行科学计算和统计分析。

SAS/QC 提供用于质量管理的各种专门工具, 可作实验设计、质量管理和过程控制, 并提供完整的实验设计和质量管理的菜单系统。