

SAS

编程演义

Romance of SAS Programming

人生若只如初见
清歌苦调两不厌
苔点红香纳线青
行舟来去泛纵横
亦应帷幄运鸿筹
间有山川亦奇秀
刷扫翠袖拂云裳
菱花存曼随双桨
袖中藏巧混天成
一缕檀烟万佛名

SAS 中国研发中心总经理 刘政

SAS 大数据与可视化分析技术总监 巫银良

著名 SAS 博主 胡江堂

SAS 中文论坛创始人、前海征信副总经理 施奕明

《SAS 编程与数据挖掘商业案例》作者 姚志勇

SASiitk 创始人 邵小龙

Fernwood Consulting Group Inc. Founder Peter Eberhardt

联袂诚意推荐

清华大学出版社



SAS

编程演义

Romance of SAS Programming

谷鸿秋◎编著

清华大学出版社
北京

内 容 简 介

本书以Base SAS为基础,重点讲解SAS编程技术,内容涵盖SAS 基础知识、数据导入导出、变量与观测操作、数据集操作与管理、函数与例程、SAS格式以及宏编程等。另外,对于目前市面上SAS书籍涉及很少,但是应用频繁的统计表格和统计图形,本书从设计原则、选择思路、绘制方法以及实例步骤等方面做了系统、详细的介绍。

本书打破语法关键字的字典式编排方式,精心凝练10个既相对独立又互相联系的专题,就地取材,采用SAS自带的小数据集,循序渐进,层层递进地进行讲解讨论。无论是SAS初学者,还是江湖老手,都可从中获益。SAS初学者可以较为全面系统地了解SAS编程技术及其应用场景,江湖老手亦可从高阶技能中获得启发感悟。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

SAS编程演义 / 谷鸿秋编著. — 北京:清华大学出版社,2017(2017.7重印)

ISBN 978-7-302-47057-1

I. ①S… II. ①谷… III. ①统计分析—统计程序—研究 IV. ①C819

中国版本图书馆 CIP 数据核字(2017)第 108485 号

责任编辑:刘 洋

封面设计:李召霞

版式设计:方加青

责任校对:宋玉莲

责任印制:杨 艳

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社总机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印装者:三河市春园印刷有限公司

经 销:全国新华书店

开 本:187mm×235mm 印 张:21 字 数:428千字

版 次:2017年6月第1版 印 次:2017年7月第2次印刷

印 数:4001~7000

定 价:79.00元

产品编号:072726-01

谨以此书献给

每次我外出时

比我妈妈还唠叨

煮的鸡蛋我一个路上来回都吃不完

总是叮嘱“好好谈朋友、好好锻炼身体、好好工作”的奶奶

前 言

蠢蠢欲动一年，奋指敲键三月，夜深人静百天，所幸的是这本书稿没有胎死腹中，终于写完了。动笔之前，我曾异常兴奋，我自以为满腹经纶无处释放的日子从此结束。完稿以后，我却沉静了，在接连填了一个又一个自己挖的坑以后，猛然抬头，发现后面其实还有更大的坑要去填，于是乎内心不禁更加焦虑。不过我很感激这份焦虑，虽然它不足以保证我所写出来的文字和代码是字字珠玑，篇篇精华，但是因为它，我可以挺起胸膛，拍着胸脯说：10 章专题 10 多万字，近 180 张图片、30 多张表格和 200 多段代码，20 多张语法卡片、30 个原创实用宏程序，这些都是热血铸就的良心作品，最起码它对得起我当初出发时的那份心意。

缘 起

我还记得初学 SAS 编程时，因为看不懂 SAS Help 而懊恼，因为不理解 @ 与 @@ 的区别而苦恼，因为分不清宏变量的 %STR、%NRSTR、%QUOTE、%BQUOTE、%NRQUOTE 以及 %NRBQUOTE 等诸多 quoting 函数而哀伤。然而，光阴似箭，似水流年，这才不过几年光景，那个曾经面对这些“简单问题”而烧心的少年，在面对后来同样烧心的学弟学妹时竟然一脸诧异：“啊？这个应该很容易理解的吧！”你看，时间是多么的狡诈，它就这样轻易地抹平了我们学习过程中的苦与痛，当我们走得越远，当初的苦与痛就忘记得越多。庆幸的是，我不是什么大神，走得也不远，那些苦与痛还没有忘得一干二净，那就趁现在，赶紧记录下来，分享出来吧。

问 题

此前知乎里有一个提问：SAS 入门书籍有哪些值得推荐？在回答里我把 SAS 学习分

成了三类（点到即止，套 PROC 型；深入应用，编程统计型；走火入魔，开发工具型）并推荐了相应的书籍。在整理市面上 SAS 相关的书籍时，我总结了三个缺陷：①专门介绍数据整理与图表呈现的书太少、太零碎，即便有，也鲜有高质量者；②几乎都采用语法关键词按字典式的编排方式论述，缺乏从实际问题凝练的良好专题；③编程技术与使用场景割裂，讲技术者纯讲技术，缺少对应的应用场景带入感。

特 色

本书试图在数据整理与图表呈现的内容上、编排方式上以及论述形式上有所突破和改进。

在内容上，顾名思义，专门讨论数据整理和统计图表的制作，不贪大求全、忌蜻蜓点水。精心提炼的 10 个专题总计 10 万字，涉及 SAS 的八卦见闻、SAS 的基础知识、数据的导入导出方式、变量与观测的各种操作、数据集的各种操作与管理、函数与例程、输入输出格式、统计表格的制作、统计图形的绘制原则、选择方法以及各系列统计图形的绘制实例，此外对 SAS 宏变量、宏程序以及开发宏程序的原则、步骤、技巧等内容均有较为详细的论述。

在编排上，推陈出新，打破按语法关键字的字典式编排方式，精心挑选的 10 个专题构成 10 个既相对独立又互相联系的章节。小节与小节之间、例子与例子之间，尽量由问题层层引入，逐步推进，减少割裂与唐突感，增加使用场景的带入感。

此外，很多 SAS 用户虽然都了解、接受甚至已经受益于 SAS 在数据整理和统计分析方面毋庸置疑的优势，但是在统计结果的呈现上，尤其是统计表格，特别是统计图形方面都或多或少存在不甚了解抑或是误解的情况。因此，本书在统计表格的制作，尤其是统计绘图方面花了大量的笔墨做串讲——是的，用一个又一个层层递进的疑问来串讲，避免单纯的介绍绘图语法和 SAS 技术，这在其他书中是很少见的。

最后，为了便于读者理解 SAS 运行机制与原理，本书在论述时都尽量采用小数据、小实例以便清晰简洁地说明问题，避免因行业背景的不同陷入具体实例的大坑。同时，为了方便读者练手测试，几乎所有数据均就地取材，采用 SASHelp 库中自带的数据集。

心 得

SAS Help 文档是学习 SAS 不可多得的手边精品材料，如果还没有深刻体会到这一点，那么赶紧去读读 R 包的 Help 文档。很多 SAS 书籍取材于 SAS Help 文档却闭口不提，这是一个巨大的失误。因此，本书会专门引导，鼓励读者去多读 SAS Help 文档、多查 SAS Help 文档。

学习知识的理想状况是单调线性、循序渐进的推进，然而现实情况却是：知识本身是错综复杂的网状结构。因此，我们经常需要迂回包抄、循环往复地学习。在介绍知识点时，本书努力做到直线推进、循序渐进，但由于作者精力、能力有限，加之知识网状结构的客观的、存在的现实，希望读者能有一个迂回包抄、循环往复的学习心态。

当然，我也有一个迂回包抄、循环往复、精进迭代的心态。本书还有很多的话题，比如 SAS 的综合矩阵语言（Integrated Matrix Language, IML）、输出传递系统（Output delivery System, ODS）、正则表达式等没能在此版付诸实践；已经付诸实践的，也会因笔者的见识、学识以及精力受限，而有所欠缺。因此，诚恳地欢迎诸位读者给出您的建设性建议以及批评性意见，送达地址 [guhongqiu\(at\)yeah\(dot\)net](mailto:guhongqiu(at)yeah(dot)net)。有您的反馈，下一版（如果有的话），肯定会更好。

致 谢

如果您读到这里了，请不要嫌我啰唆，因为一路走来，需要感谢的人特多，而且感谢应该是一个严肃的话题，因此，下面是一本正经的致谢。

感谢北京中医药大学曾光教授、刘仁权教授带我叩开流行病与卫生统计领域的大门；感谢中国疾病预防控制中心吴尊友教授教我公共卫生的大义；感谢北京协和医学院李卫教授携我走进临床研究的大门；感谢国家神经系统疾病临床医学研究中心王拥军、王伊龙教授给我机会在实践中提升临床研究思维与技能。

感谢 *The Little SAS Book* 的作者 Lora D. Delwiche 女士，著名 SAS 绘图博客 Graphically Speaking 的博主、众多 SAS 绘图专著的作者 Sanjay Matange 先生，以及 *The DS2 Procedure: SAS Programming Methods at Work* 的作者 Peter Eberhardt 先生在本书写作过程中给予的支持和帮助。

感谢 SAS 中国研发中心总经理刘政先生；感谢 SAS 中国研发中心分析产品开发部总监高燕女士、SAS 中国研发中心商业智能和可视化分析产品部技术总监巫银良先生、SAS 中国区培训经理赵丹先生、SAS 大中华区市场总监蒋顺利先生在我准备书稿过程中给予的支持；感谢 SAS 中文论坛创始人、前海征信副总经理施亦明先生，SAS 中文资讯网的创始人 sxlion 以及人大经济论坛里的一大波 ID（jingju11、pobel、hopewell、davi12000、kuhasu、ahuige、soporaeternus、YueweiLiu、oloolo、bobguy、Imasasor、playmore、crackman、dxystata）在 SAS 的江湖里传道解惑。

感谢本书的编辑，清华大学出版社的刘洋先生。没有他的信任，这本书可能会散落于江湖；没有他的信任，写作可能会被无数次的催稿打断。还好，他对我和这本书稿一直保

持足够的耐心。再次感谢清华大学出版社编辑部，精心挑选每章首页的山水画，配合标题，意境深远。

来北京十多年，感谢中国气象科学研究院谷湘潜研究员、首都医科大学附属北京地坛医院江宇泳教授给予的各方面关照；感谢中南大学谷潜平教授的建议；感谢国家神经系统疾病临床医学研究中心的王彩云主任早上的烤红薯——无上美味、香甜至极。

最后，感谢因为 SAS、因为此书，和我有了交集的你。

谷鸿秋

2017年5月24日

目 录

图目录	17
表目录	23
程序目录	25
语法卡片目录	33

第 1 章	人生若只如初见：初识 SAS	1
1.1	往事并不如烟	1
1.1.1	逗你玩的发音	1
1.1.2	有点趣的历史	2
1.1.3	逝不去的江湖	5
1.2	选择一厢情愿	5
1.3	软件架构	6
1.4	安装与许可	7
1.5	运行模式	12
1.6	编程界面	12
1.6.1	DMS 界面	12
1.6.2	EG 界面	14
1.6.3	SAS Studio 界面	14
1.7	版本	16
1.7.1	购买版与大学版	17
1.7.2	免费云端版	17

1.7.3	各操作系统平台版	20
1.7.4	各语言版	20
1.8	本章小结	21

第 2 章 | 清歌苦调两不厌：夯实基础

2.1	Foundation SAS	22
2.1.1	Foundation SAS 的构成	22
2.1.2	Base SAS	23
2.2	SAS 数据分析流程	23
2.3	逻辑库与数据集	24
2.3.1	逻辑库	24
2.3.2	数据集	27
2.3.3	变量	31
2.4	SAS 编程语言	32
2.4.1	SAS 程序结构	32
2.4.2	SAS 语法规则	33
2.4.3	SAS 语言元素	35
2.4.4	三种逻辑结构	40
2.4.5	数组结构	42
2.4.6	函数与 CALL 例程	46
2.4.7	结构化查询语言 SQL	46
2.4.8	SAS 宏 MACRO	48
2.5	理解 SAS 运行机制	49
2.5.1	PDV 与 DATA 步自循环	49
2.5.2	@ 与 @@ 的困惑	55
2.6	用好 SAS Help 的秘诀	57
2.6.1	SAS Help 知多少	57
2.6.2	看懂 SAS Help 的基本套路	60
2.6.3	检索 SAS Help 的小技巧	61
2.6.4	熟悉 SAS Help 下的数据集	62
2.7	本章小结	63

3.1	读取对象与读取方式.....	64
3.2	数据读取策略.....	65
3.3	读取 DBMS 数据文件.....	66
3.3.1	SAS/ACCESS 与 DBMS.....	66
3.3.2	LIBNAME 语句访问 DBMS 数据文件语法.....	67
3.4	读取 PC 数据文件.....	67
3.4.1	小试牛刀.....	68
3.4.2	PROC IMPORT 语法.....	68
3.4.3	LIBNAME 访问 PC 文件语法.....	70
3.4.4	实例扩展.....	70
3.5	读取 Flat 数据文件.....	72
3.5.1	读入 CSV 文件.....	72
3.5.2	读入 TXT 特殊字符分隔的文件.....	72
3.6	读取流式数据.....	74
3.6.1	流式数据初探.....	74
3.6.2	INPUT 语句一般语法.....	74
3.6.3	列表读入式.....	75
3.6.4	列读入式.....	77
3.6.5	格式读入式.....	77
3.6.6	命名读入式.....	78
3.6.7	DATALINES 数据综合案例.....	79
3.6.8	关于列表、指针及格式等.....	79
3.7	顺带说说数据导出.....	80
3.8	数据导入导出的宏.....	81
3.9	本章小结.....	81

4.1	变量与观测操作概览.....	82
4.2	创建数据集.....	83

4.2.1	DATA+SET 语句	84
4.2.2	SQL CREATE 语句	85
4.2.3	过程步的语句与选项	85
4.3	观测与变量的筛选	89
4.3.1	IF 与 WHERE 的辨析	90
4.3.2	KEEP、DROP 与 RENAME	91
4.3.3	PROC SQL 筛选变量与观测	93
4.4	横向的其他操作	93
4.4.1	新变量生成与赋值	94
4.4.2	变量类型转换	98
4.4.3	变量重新分组切割	100
4.4.4	缺失变量查找	102
4.5	纵向的其他操作	103
4.5.1	DATA 步实现累加	103
4.5.2	PROC 步实现累加	106
4.5.3	LAG 与 DIF 的隔行取物	107
4.6	本章小结	108

第 5 章 亦应帷幄运鸿筹：数据库集 109

5.1	修改数据集：MODIFY 语句	109
5.1.1	MODIFY 语句实例	109
5.1.2	MODIFY 语句注意事项	111
5.1.3	MODIFY 语句语法概览	112
5.2	更新数据集：UPDATE 语句	113
5.2.1	UPDATE 语句实例	113
5.2.2	UPDATE 语句注意事项	113
5.2.3	UPDATE 语句语法概览	114
5.3	行列互换（转置）：PROC TRANSPOSE	114
5.3.1	行转列（宽表转长表）	114
5.3.2	列转行（长表转宽表）	116
5.3.3	PROC TRANSPOSE 注意事项	116

5.3.4	PROC TRANSPOSE 语法概览	117
5.4	横向拼接（并接）：多 SET/MERGE/PROC SQL	117
5.4.1	一对一读入	118
5.4.2	一对一并接	119
5.4.3	匹配并接	119
5.5	纵向拼接（串接）：SET /PROC APPEND / PROC SQL	123
5.5.1	SET 语句	124
5.5.2	PROC APPEND	124
5.5.3	PROC SQL	125
5.6	PROC SQL 表连接	126
5.6.1	左连接（LEFT JOIN）	127
5.6.2	右连接（RIGHT JOIN）	128
5.6.3	内连接（INNER JOIN）	129
5.6.4	全连接（FULL JOIN）	129
5.6.5	表连接操作语法总结	130
5.7	PROC SQL 集运算	131
5.7.1	差（EXCEPT）	131
5.7.2	并（UNION）	132
5.7.3	交（INTERSECT）	133
5.7.4	外并（OUTER UNION）	134
5.8	数据集管理	134
5.8.1	初入门庭：制作变量字典	135
5.8.2	大雅之堂：数据集信息	136
5.8.3	大雅之堂：数据集操作	137
5.8.4	大雅之堂：变量信息	138
5.8.5	登峰造极：精通 SAS 字典	139
5.9	本章小结	143
第 6 章	间有山川亦奇秀：函数例程	145
6.1	函数和例程是什么？	145
6.2	为什么要用函数和例程？	146

6.2.1	一个例子感受函数的便利	146
6.2.2	一个例子感受例程的便利	147
6.3	怎么用函数和例程?	148
6.3.1	函数语法	148
6.3.2	例程语法	149
6.4	在哪里使用函数和例程?	149
6.5	有哪些函数例程可用?	150
6.6	高频函数和例程有哪些?	152
6.7	不够用怎么办?	153
6.8	本章小结	154

第 7 章 | 翩跹翠袖拂云裳：巧用格式

7.1	何为格式?	155
7.2	为何要用格式?	156
7.3	格式名的样子	157
7.4	常用系统格式	158
7.4.1	常用输出格式	158
7.4.2	常用输入格式	159
7.5	PROC FORMAT 自定义格式	159
7.5.1	自定义格式	159
7.5.2	用数据集定义格式	161
7.5.3	PROC FORMAT 语法	161
7.6	格式的妙用	162
7.6.1	格式使用位置	162
7.6.2	变量重组	163
7.6.3	统计分析过程	164
7.6.4	统计缺失观测	165
7.6.5	条件显示	166
7.7	本章小结	167

菱花苕蔓随双桨：百变绘图	168
8.1 绘图软件知多少.....	168
8.2 丑陋死板太难搞.....	169
8.3 绘图系统瞄一瞄.....	170
8.3.1 SAS/GRAPH.....	171
8.3.2 统计过程自动绘图.....	172
8.3.3 ODS Graphics System.....	173
8.4 基本原则不能少.....	179
8.4.1 Less is more.....	179
8.4.2 图形元素设计.....	180
8.5 统计图形的选择策略.....	180
8.5.1 统计图形选择总则.....	181
8.5.2 案例说明.....	183
8.6 条图系列 (Bar Chart).....	184
8.6.1 单式条图.....	184
8.6.2 频数图.....	185
8.6.3 带误差限的单式条图.....	186
8.6.4 单式百分比条图.....	187
8.6.5 簇拥式复式条图.....	188
8.6.6 簇拥式复式误差限条图.....	188
8.6.7 堆叠式复式条图.....	189
8.6.8 马赛克图.....	191
8.6.9 镜面式复式条图.....	191
8.6.10 面板条图.....	193
8.7 直方图系列 (Histogram).....	195
8.7.1 简单直方图.....	195
8.7.2 重叠直方图.....	196
8.7.3 镜面直方图.....	197
8.7.4 面板直方图.....	199
8.8 箱线图系列 (Box Plot).....	200
8.8.1 简单箱线图.....	200

8.8.2	分组箱线图	201
8.8.3	面板箱线图	202
8.9	散点图系列 (Scatter Plot)	202
8.9.1	X-Y 散点图	203
8.9.2	X-Y 散点回归图	203
8.9.3	分组散点图	204
8.9.4	面板散点图	205
8.9.5	泡泡图	206
8.9.6	矩阵散点图	207
8.10	折线图系列 (Line Plot)	208
8.10.1	简单折线图	208
8.10.2	误差限折线图	209
8.10.3	分组误差限折线图	211
8.10.4	面板误差限折线图	211
8.11	面积图系列 (Area Plot)	212
8.11.1	面积图	212
8.11.2	带状图	213
8.12	拟合图系列 (Fit Plot)	214
8.12.1	密度曲线	214
8.12.2	回归线	215
8.12.3	椭圆曲线	216
8.12.4	ROC 曲线	217
8.12.5	Kaplan-Meier 曲线	217
8.12.6	LOESS 曲线	219
8.12.7	Spline 曲线	221
8.13	森林图系列 (Forest Plot)	222
8.13.1	简单森林图	222
8.13.2	亚组分析森林图	223
8.14	地图系列 (Map)	226
8.14.1	纯地图	226
8.14.2	统计地图	227
8.14.3	热力地图	228

8.15	PROC SGPLOT 总结	229
8.16	八九打扮十分妖	231
8.16.1	更换样式文件	231
8.16.2	SG 过程语句选项	232
8.16.3	GTL 语句选项	233
8.16.4	创建样式文件	235
8.17	图片属性控制	238
8.18	本章小结	240

第9章 | 拙中藏巧混天成：统计表格 241

9.1	统计表格是什么？	241
9.2	统计表格如何设计？	244
9.3	统计表格有哪些用法？	245
9.3.1	基线信息表格	245
9.3.2	危险因素表格	246
9.3.3	结局效应表格	246
9.3.4	亚组分析表格	247
9.4	SAS 制表方式有哪些？	248
9.4.1	统计汇总过程	249
9.4.2	专用制表过程 PROC TABULATE	250
9.4.3	专用报告过程 PROC REPORT	251
9.5	如何完美实现统计表格？	252
9.6	手把手来说明	253
9.6.1	完整实例	253
9.6.2	重点解读	255
9.6.3	技能升华	256
9.7	本章小结	257

第10章 | 一缕檀烟万佛名：宏中奥秘 259

10.1	学习宏的几个理由	259
------	----------	-----