

循序渐进掌握数据分析要义和精髓
从数据中获取洞见与智慧

SAS

技术内幕

从程序员到数据科学家

巫银良◎著

INSIDER OF
SAS TECHNOLOGY
FROM PROGRAMMER
TO DATA SCIENTIST

张铭

北京大学信息科学技术学院
博士生导师、教授

刘云中

国务院发展研究中心
研究员

唐朝晖

艾德思奇 (adSage)
董事长

刘政

SAS中国研发中心
总经理

李嘉

慧科集团技术产品中心
技术副总裁

谷鸿秋

《SAS 编程演义》
作者

付春光

北京冲和资产管理有限公司
量化投资总监

杨祉雄

致远互联
高级副总裁

林建伟

华为供应链管理
数据科学家

陈云凯

SAS 中国销售总监兼合作伙伴
与渠道负责人

联袂诚意推荐

清华大学出版社



SAS

技术内幕

从程序员到数据科学家

巫银良◎著

INSIDER OF
SAS TECHNOLOGY

FROM PROGRAMMER
TO DATA SCIENTIST

清华大学出版社

北京

内 容 简 介

本书共 27 章,分为上下两卷:上卷介绍 SAS 编程基础与使用方法,是广大程序员快速掌握 SAS 编程技术的简明开发教程;下卷阐述数据分析的关键基础知识并提供相应 SAS 代码实现,目的是激发读者兴趣,助其跨越传统编程与数据分析的鸿沟,从程序员华丽转身为数据科学家。书中演示代码力图简洁清晰地解释相关概念,追求大道至简。本书兼顾编程技术与数据分析,期待程序员、信息处理与统计分析人员以及对数据分析科学感兴趣的读者都能从本书中获益良多,循序渐进地掌握数据分析的要义和精髓,从数据中获取洞见与智慧。

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技术内幕:从程序员到数据科学家/巫银良著. —北京:清华大学出版社,2018
ISBN 978-7-302-50278-4

I. ①S… II. ①巫… III. ①统计分析—应用软件—程序设计 IV. ①C819

中国版本图书馆 CIP 数据核字(2018)第 111953 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:清华大学印刷厂

经 销:全国新华书店

开 本:185mm×260mm 印 张:38.75 字 数:891 千字

版 次:2018 年 12 月第 1 版 印 次:2018 年 12 月第 1 次印刷

定 价:168.00 元

产品编号:077684-01

推荐语

《SAS 技术内幕：从程序员到数据科学家》是值得一读的数据分析技术入门佳作。这本书是数据分析领域资深专家为融合程序员和统计分析人员不同视角而编写的，给读者在计算机程序设计世界和数据分析世界之间架起一座桥梁……罗马不是一天建成的！但阅读本书将使普通的程序员也能快速入门 SAS 语言并以最简洁的方式掌握 SAS 编程核心，实现向优秀数据科学家的华丽转身！

北京大学信息科学技术学院博士生导师、教授 张铭

本书把复杂的数据分析科学以最简单的应用和实例介绍给读者，是一本难得的告诉你用 SAS 将数据变为知识，将知识变为决策的好书！

国务院发展研究中心研究员 刘云中

SAS 是数据分析、数据挖掘和人工智能的领军平台。本书集结了作者多年从事数据分析和商业智能研究与实践的经验，用 SAS 平台深入浅出地给程序员介绍了数据挖掘的基础与精髓，帮助他们成为数据分析的真正专家，非常推荐！

艾德思奇（adSage）董事长 唐朝晖

现在很多企业都开始进入数据分析领域，由于没有现成的、足够多的数据分析人员，企业只能让现有的软件开发人员承担这些责任。然而，软件开发人员与数据分析人员的思维是有差异的，也就是确定性的结果和概率性的结果的差异。如何帮助这些软件开发人员成为合格的双重人才，需要一定的知识补充和理论指导。巫银良先生根据自己丰富的从业实践与行业观察总结出来的这本书正好填补了这一空白，值得推荐！

SAS 中国研发中心总经理 刘政

SAS 作为一门古老而高效的统计编程语言，在商业数据分析行业已经流行了几十年，历久弥新！它给人的感觉是神秘而复杂，深奥且与众不同，往往让人难以理解！但本书作者在多年编程经验和分析实践的基础上，用简单直白的语言将数据分析这件事说明白了。本书从程序员的角度出发，以简洁实例带领读者从程序员世界深入数据分析的世界，是一本写给程序员的数据分析技术入门佳作！

慧科集团技术产品中心技术副总裁 李嘉

很多 SAS 统计师其实是野蛮生长起来的，比如我，面对代码缺乏程序员的严谨思维；很多 SAS 程序员其实是半道出家的，比如我，面对数据缺乏统计师的分析思维。巫银良先生此书恰好弥合了这两方面的话题。本书不仅对常规 SAS 编程语言进行了细致的介绍，而且对各种数据结构用 SAS 编程语言进行了对接；不仅有常规统计方法的讲解，而且有统计方法背景知识的 SAS 论证，末章更是趣味大增、思路大开，花式展示用 SAS 高精度求解 π 值。本书内容丰富，构思严谨，必将是程序员、统计分析师以及未来的数据科学家们手中的一份上好读物。

《SAS 编程演义》作者 谷鸿秋

本书由基础到专业、循序渐进地介绍了数据分析的各种技术和应用，是从事量化投资专业人士必读的入门书籍。

北京冲和资产管理有限公司量化投资总监 付春光

企业数字化时代已到来！当每个企业每个人每个物体都在被不断数据化时，对数据化背后的洞察获取相比高速增长的数据积累显得迟钝。这本书将弥补这一不足，让更多程序员、统计分析师、软件产品设计者向数据科学专业人士方向发展，以科学的方法去发现数据的神秘和数据之美。

致远互联高级副总裁 杨祉雄

当我看到标题的一刻，心想这可能又是一本专门针对程序员讲解编程的书。但当我打开书，慢慢品读的时候，有的是惊喜；在书中，可以看到与数学概念相关的历史、公式，计算机技术以及 SAS 编程的有机融合。我一口气读下来，没有感到晦涩和枯燥，这是一本值得推荐的好书。

华为供应链管理部数据科学家 林建伟

SAS 的合作伙伴，期待一些系统化地解释 SAS 技术和应用案例的书籍，以帮助合作伙伴提升 SAS 应用技能，从而更好地服务客户，实现商业价值。本书有层次、有步骤地帮助 SAS 顾问和技术人员，兼顾程序员视角和数据科学家视角，逐步深入地探索 SAS 技术的应用，以充分发挥 SAS 技术在商业分析具体业务场景中的优势，这是一本非常值得推荐的好书！

SAS 中国销售总监兼合作伙伴与渠道负责人 陈云凯

序 言

《SAS 技术内幕：从程序员到数据科学家》是值得一读的数据分析技术入门佳作。这本书是数据分析领域资深专家为融合程序员和统计分析人员不同视角而编写，给读者在计算机程序设计世界和数据分析世界之间架起一座桥梁。

我在北京大学计算机系从事计算机教学和科研几十年，所主讲的“数据结构与算法”被评选为国家级精品课程。图灵奖获得者尼古拉斯·沃斯 (Niklaus Wirth) 提出“程序 = 算法 + 数据结构”，而计算机语言则是构建程序世界的主要工具。在计算机编程语言世界里，除了汇编、C/C++ 和 Java/C# 等主流通用语言外，还有很多面向特定领域的高级语言，如 SAS、R 和 MATLAB 等。我一直鼓励学生们开阔视野，多接触一些工业界的专业工具和编程语言，不要局限在 C/C++ 等通用语言的框架里。在这些专用语言的背后，往往蕴含特定的领域思维和设计哲学，基于专业人士在处理领域问题时积累的丰富经验进行了极其灵活的设计，专用语言的这些特性往往是通用编程语言所不考虑也不具备的。

近些年来，数据分析和商业智能发展迅猛，大数据和人工智能在学术界和产业界都生机勃勃。学术界科研人员注重的是引领世界科技发展的超前研究，在理论和创新方面有独到之处。工业界是前沿技术的成熟应用，学术界在培养应用型人才方面需要考虑工业界的真实需求。因此，北京大学很早就与全球数据分析行业的领导者 SAS 合作，开设了面向研究生和高年级本科生的统计分析选修课程“统计分析与商务智能”，取得了非常好的教学效果。SAS 是全球数据分析领域的领导者，它们以创新的软件和服务，在数据分析、商业智能、数据管理等领域耕耘四十余年，一直秉持的理念就是提供“慧识力量”（The Power to Know[®]），使用户能够对海量数据深入了解并获得洞见和价值（Insight & Value），为企业的运营发展提供决策利器。

如果按照数据的生产和消费进行划分，大部分应用软件和系统，包括互联网社交媒体和电子商务都在大量制造数据，这一进程经过信息化时代和互联网时代已经得到充分体现。如何消费这些数据却催生了其他软件系统的发展，它们包括各种分析系统如商业智能（BI）、决策系统、专家系统和人工智能系统（AI）等。近些年来数据分析在公众媒体眼中变得非常热门，原因就是分析利用数据并从中获取价值在“信息化时代”和“互联网时代”之后的“大数据时代”变得尤为迫切。在过去的几年中，处理数据、对数据进行可视化分析、理解数据，跟数据进行沟通、深入探索，并从中获得价值，这些过程已经变得极为重要。在软件领域，数据分析将是 21 世纪头十年经过商业智能产业大并购之后的下一个蓝海！不过，蓝海中不再是程序员划着小船在徜徉，而是数据科学家们开着利舰遨游在大数据的海洋之上。

现实已经告诉我们，计算机程序员需要在程序员思维之外，尽快拥有数据分析思维，

华丽转身成为数据科学家！文本分析、语音识别、神经网络、人工智能、自动驾驶等各种最新最热门的数据分析领域，说到底需要的是强大的综合能力，包括良好的计算机编程技能（Programming）、扎实的数学和统计知识（Mathematics & Statistics）、专业的业务领域知识（Business）以及与数据来源和消费方良好的沟通技能（Communication）等。除具备良好的编程能力和丰富的数据库知识之外，如果程序员掌握像 SAS 这样一种严谨且专门面向数据分析的语言，拥有扎扎实实修炼数学和统计分析的能力，那就能让自己站在数据分析的最前沿。

罗马不是一天建成的！但阅读本书将使普通的程序员也能快速入门 SAS 语言并以最简洁的方式掌握 SAS 编程核心，实现向优秀数据科学家的华丽转身！

北京大学信息科学技术学院

博士生导师、教授

张 铭

2018 年 8 月

前言

本书是写给程序员的数据分析技术入门书籍，成书于 2017 年作者在北京大学教授面向研究生和高年级本科生的“统计分析与商务智能”选修课期间，它试图在程序世界和数据分析世界之间架起一座坚实的桥梁。

本书主要包括上下两卷内容：程序员视角下的 SAS 编程技术和数据结构，数据科学家视角下的数据分析理论和 SAS 实践。

上卷主要包括 SAS 语言入门、数据集与 DATA 步、变量与表达式、流程控制、函数封装、SAS 宏、DS2、代码组织、文件读写、按位运算以及扩展 SAS 功能。另外，还从程序员的视角阐述了各种数据结构在 SAS 中的编程实现和应用，包括 SAS 数组、队列与堆栈、链表、二叉树、矩阵运算和图等。

下卷包括统计学基础、大数定律与中心极限定理、统计分布、方差分析、数据标准化、主成分分析与因子分析、相关分析与回归分析、聚类分析、神经网络，最后以 π 值高精度求解和探索分析结束。

本书内容涵盖理论和实践，章节组织采用从简单到复杂的方式。本书囊括数据分析技术方面较为核心的基础内容，试图与读者一起触及数据世界分析与智能的核心。各章为读者提供简洁可运行的 SAS 示例代码、算法实现以及快速指南，为广大受过计算机科学教育的程序员向数据科学家华丽转身提供了必要的快速入门指导。本书附录还提供二项分布、泊松分布和标准正态分布累积概率表以及 t 分布、 χ^2 分布和 F 分布临界值表的制作与查找方法。本书适用于各大大专院校统计分析专业和信息处理专业的学生，有志于从事数据分析的广大程序员、统计分析从业人员以及所有想成为数据科学家的专业人士。

本书与传统的 SAS 数据分析书籍不同之处在于，它从程序员的视角出发，循序渐进探讨数据分析的各个方面，避免“知其然而不知其所以然”。因此，阅读本书可使你从完全不了解 SAS 到对 SAS “有所了解”，从而掌握数据分析的要义和精髓，实现从程序员到数据科学家的华丽转身！

最后，诚挚感谢北京大学信息科学技术学院博士生导师、中国 ACM 教育专委会主席张铭教授为本书作序。感谢 SAS 中国研发中心总经理刘政博士给予的大力帮助和指导，感谢 SAS 大中华区前市场总监蒋顺利先生、高级市场经理曾秋媚女士在写作过程中给予的帮助和支持，感谢在写作过程中给予特别帮助的 SAS 中国研发中心的同仁们。感谢清华大学出版社编辑部的刘洋先生和全体同仁的辛勤工作，是他们的努力使本书得以与读者见面！感谢恩师陈永金先生以及所有在我生命中给予帮助的人们！特别感谢与我一路同行的知己和家人，是他们赋予了我生命的色彩与力量！

目 录

上 卷

第1章	SAS语言入门	002
1.1	语言概述	005
1.2	编程环境	010
1.3	SAS Studio编程	016
第2章	数据集与DATA步	032
2.1	SAS 逻辑库	032
2.2	SAS 数据集	036
2.3	DATA 步	041
2.3.1	内嵌数据行或外部数据文件	043
2.3.2	通过已有 SAS 数据集生成	048
2.3.3	通过 PROC IMPORT 或 PROC SQL 生成	051
2.4	DATA 步的运行机制	054
2.4.1	编译阶段	055
2.4.2	运行阶段	056
2.5	DATA 步语句快速索引	057
第3章	变量与表达式	062
3.1	常量与变量	062
3.1.1	变量长度与缺失值	063
3.1.2	数值常量	065
3.1.3	日期/时间/日期时间常量	067
3.1.4	字符常量	067
3.2	表达式	068
3.2.1	运算符	068
3.2.2	运算符优先顺序	072
3.2.3	WHERE 语句特定运算符	072
3.2.4	赋值语句	073
3.2.5	累加赋值语句	074
3.2.6	RETAIN 语句	075

3.3	SAS数组	077
3.3.1	数组名称	079
3.3.2	数组元素变量列表	079
3.3.3	数组长度	080
3.3.4	隐式下标变量	080
3.3.5	多维数组	081
3.3.6	临时数组	082
3.3.7	数组排序	083
3.3.8	注意事项	084
第4章	流程控制	087
4.1	DO-END语句块	087
4.2	分支控制	088
4.2.1	IF-THEN 分支控制	088
4.2.2	ELSE-IF 多分支控制	090
4.2.3	SELECT-WHEN 多分支控制	090
4.3	循环控制	091
4.3.1	指定次数的循环：DO-TO-BY	091
4.3.2	指定条件的循环：DO-WHILE 与 DO-UNTIL	092
4.3.3	指定集合的循环：DO-OVER	093
4.4	特殊的流程控制语句	094
4.4.1	跳出循环语句：LEAVE	094
4.4.2	继续循环语句：CONTINUE	094
4.4.3	返回语句：RETURN	095
4.4.4	中止执行语句：STOP 与 ABORT	095
4.4.5	跳转语句：GOTO 与 LINK	096
第5章	函数封装	099
5.1	LINK-RETURN 技术	101
5.2	SAS宏函数封装	103
5.3	FCMP函数	105
5.4	系统函数速查	110
第6章	SAS宏	113
6.1	宏变量	114
6.1.1	命名	114
6.1.2	作用域	114
6.1.3	系统宏	115
6.1.4	宏代码调试	117

6.1.5 宏表达式	117
6.2 宏函数	119
参数定义	120
6.3 逻辑控制	121
6.3.1 宏语句块	121
6.3.2 条件分支	121
6.3.3 循环控制	121
6.4 系统宏函数	122
第7章 DS2	124
7.1 程序结构	126
7.1.1 变量声明与类型	126
7.1.2 程序实体作用域	129
7.1.3 变量数组与标准数组	130
7.1.4 系统方法与用户自定义方法	131
7.2 数据程序	136
7.3 包程序	137
7.4 线程程序	143
第8章 代码组织	149
8.1 静态文件包含	149
8.2 程序中动态扩展代码	151
8.3 动态执行外部命令	153
第9章 文件读写	157
9.1 二进制文件读写	157
9.2 文本文件读写	159
9.3 顺序读取多个文件	162
9.4 并行读取多个文件	165
9.5 共享缓冲区读写	166
第10章 按位运算	168
10.1 按位运算	168
10.2 实现方法	169
10.3 按位运算应用	173
第11章 扩展SAS功能	177
11.1 通过 Module调用外部 DLL函数	177
11.2 用 C 语言开发用户函数库	182
11.2.1 准备64位 C 编译环境	182

11.2.2	开发用户自定义动态库	183
11.3	PROTO 编写 C 代码或注册外部 DLL	184
第12章	数据结构——数组	187
12.1	数组	187
12.1.1	DATA步数组	187
12.1.2	FCMP 数组	189
12.1.3	DS2 数组	192
12.1.4	SAS宏数组	193
12.2	数组应用：高精度数值计算	194
第13章	数据结构——队列与堆栈	196
13.1	队列	196
13.1.1	函数版实现与示例	196
13.1.2	宏版实现与示例	199
13.2	堆栈	202
	函数版实现与示例	202
第14章	数据结构——链表	206
14.1	基础知识	206
14.2	如何在 SAS 代码中内嵌 C 语言代码	207
14.3	单向链表和双向链表	209
14.4	链表应用：约瑟夫斯问题	216
第15章	数据结构——二叉树	221
15.1	PROTO 实现与封装	221
15.2	FCMP二叉树实现	227
15.3	二叉树应用：算术表达式求值	231
第16章	数据结构——矩阵运算	235
16.1	FCMP 矩阵运算	236
16.2	DS2 矩阵运算	243
16.3	矩阵应用：线性方程组求解	246
16.4	矩阵应用：非线性方程组求解	248
第17章	数据结构——图	255
17.1	深度优先和广度优先遍历	256
17.2	最短路径问题	260
17.2.1	Dijkstra算法	261
17.2.2	Bellman-Ford 算法	263
17.2.3	Floyd-Warshall 算法	265

下 卷

第18章 统计学基础	270
18.1 数据特征度量	270
18.1.1 集中趋势度量	272
18.1.2 离散趋势度量	274
18.1.3 分布特征度量	277
18.1.4 置信区间、置信水平与 p -值	279
18.2 统计学上的变量类型	280
18.3 基本数据处理	283
18.3.1 排序与排名	284
18.3.2 数据转置	285
18.3.3 堆叠与拆分	286
18.3.4 过滤数据	287
18.3.5 随机抽样	289
18.3.6 基本统计量	290
18.4 基本图形图表	292
18.5 SAS 产品与过程步概览	303
18.5.1 SAS核心产品功能简介	305
18.5.2 Base SAS 过程步速查	309
18.5.3 SAS/STAT过程步速查	318
第19章 大数定律与中心极限定理	327
19.1 大数定律	327
19.1.1 弱大数定律	327
19.1.2 三种大数定律	329
19.1.3 图形化证明	330
19.1.4 强大数定律	333
19.2 中心极限定理	334
19.2.1 大数定律与中心极限定理关系	335
19.2.2 图形化证明	336
19.2.3 实际用途	340
第20章 统计分布	342
20.1 均匀分布	342
20.2 离散型统计分布	345
20.2.1 伯努利分布	345

20.2.2	二项分布	347
20.2.3	几何分布	353
20.2.4	负二项分布	357
20.2.5	超几何分布	360
20.2.6	泊松分布	362
20.3	连续型统计分布	365
20.3.1	正态分布	365
20.3.2	对数正态分布	373
20.3.3	指数分布	376
20.3.4	卡方分布	379
20.3.5	学生t-分布	381
20.3.6	F分布	387
20.3.7	柯西分布	390
20.3.8	贝塔分布	392
20.3.9	伽马分布	395
20.3.10	爱尔朗分布	397
20.3.11	韦布尔分布	399
20.3.12	三角分布	400
20.3.13	Table 分布	401
附录:	各统计分布之间的关系	403
第21章	方差分析	404
21.1	假设检验	404
21.2	方差分析	406
21.2.1	学生t-检验	406
21.2.2	单因子方差分析	408
21.2.3	双因子方差分析	418
第22章	数据标准化	421
22.1	常用标准化方法	421
22.2	SAS数据标准化	424
22.3	自定义数据标准化	429
第23章	主成分分析与因子分析	433
23.1	主成分分析	434
23.1.1	主成分分析原理	435
23.1.2	主成分分析的具体步骤	436
23.2	因子分析	443

23.2.1	因子分析原理	443
23.2.2	巴特利球度检验和 KMO 检验	443
23.2.3	因子分析的具体步骤	445
第24章	相关分析与回归分析	450
24.1	变量关系	450
24.2	相关分析	451
24.2.1	线性相关性度量	451
24.2.2	非参数关联度量	452
24.2.3	定量数据的相关分析	455
24.2.4	类别数据的相关分析	457
24.3	回归分析	460
第25章	聚类分析	467
25.1	聚类度量	469
25.1.1	距离系数	469
25.1.2	相似性/相关系数	471
25.1.3	SAS实践	473
25.2	聚类形成方法	475
25.2.1	一次形成分类系统	475
25.2.2	K-均值聚类	477
25.2.3	逐步形成分类系统	485
25.2.4	R 型聚类分析	491
25.3	自己实现聚类算法	494
25.3.1	K-均值方法	494
25.3.2	逐步形成分类系统	501
附录:	聚类度量的自定义实现	509
第26章	神经网络	512
26.1	神经元模型	513
26.2	神经网络	517
26.2.1	训练神经网络	519
26.2.2	反向传播算法	519
26.3	SAS 代码实现与范例	524
第27章	π 高精度求解与探索分析	536
27.1	π 值计算	537
27.1.1	蒙特卡罗方法	543
27.1.2	蒲丰投针方法	544

27.1.3	微积分方法	545
27.1.4	幂级数方法	546
27.1.5	幂级数高精度方法	548
27.1.6	梅钦类公式高精度方法	550
27.1.7	迭代方法——贝拉公式	554
27.2	π 值分析	557
27.2.1	数字分布规律	558
27.2.2	可视化探索	561
附录		564
参考文献		598

上 卷