SAS

高级统计分析教程

（第2版）

◎ 胡良平 主编
◎ 高 辉 审校

SAS
统计分析系列

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

SAS

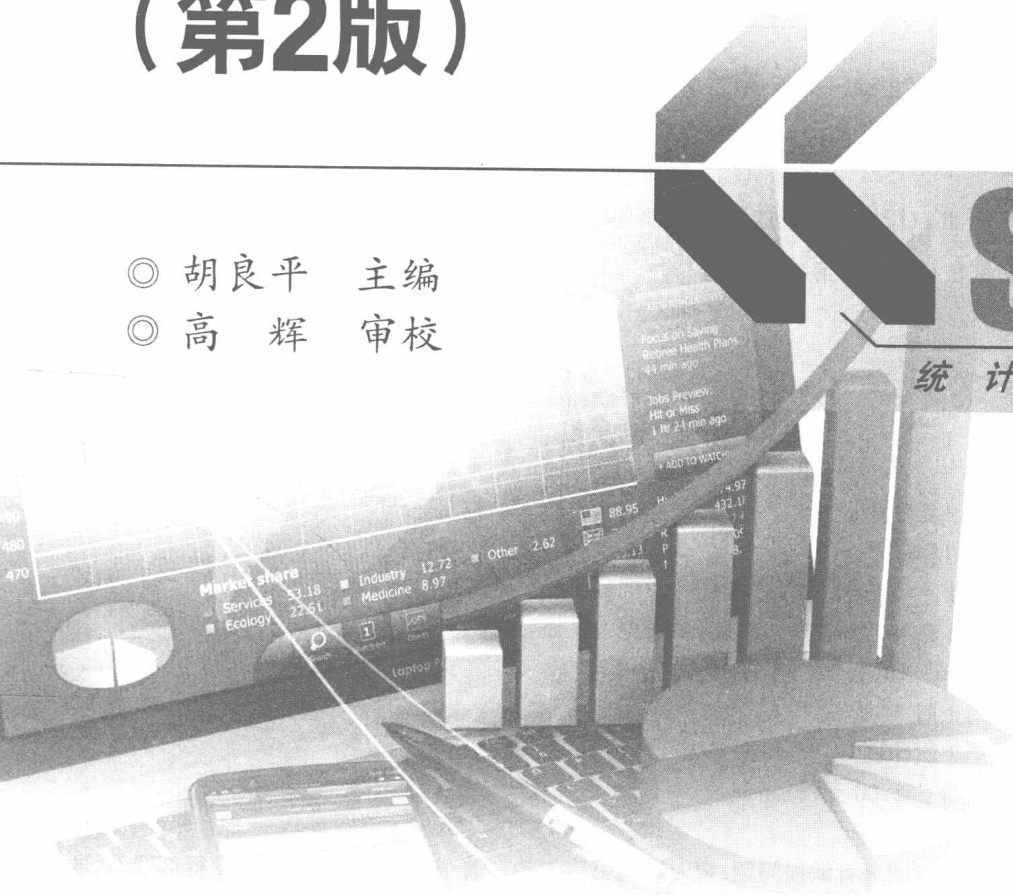
高级统计分析教程

（第2版）

◎ 胡良平　主编
◎ 高　辉　审校

SAS

统计分析系列



電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书基于SAS 9.3版本，内容丰富且新颖，适用面宽且可操作性强，涉及统计学基础和现代多元统计分析。这些内容高质量、高效率地解决了各种多元统计分析、数据挖掘、遗传资料统计分析和SAS实现及结果解释等人们迫切需要解决却又十分棘手的问题。

本书内容共6篇，第1篇包括第1～4章，回答了4个基础性问题，即“如何确保数据是值得分析的”、“如何选择统计图并用SAS绘制”、“如何给统计分析方法分类与合理选用统计分析方法”和“如何基于偏好数据确定多因素的最佳水平组合”；第2篇包括第5～12章，介绍了研究变量之间相互和依赖关系的8种多元统计分析方法；第3篇包括第13～16章，介绍了评价样品间亲疏、优劣或相对位置的4种多元统计分析方法；第4篇包括第17～19章，介绍了评价变量与样品之间关联性的3种多元统计分析方法；第5篇包括第20～24章，第6篇包括第25～26章，介绍了数据挖掘、生物信息学和遗传资料分析3大领域方面的知识和技术。另有配套的辅助资料，可在华信教育资源网 www.hxedu.com.cn 查询。

本书适合需要运用现代多元统计分析以及数据挖掘和遗传资料分析等相关领域知识解决实际问题的本科生、研究生、博士生、科研和管理工作者、临床医生和杂志编辑学习和使用。

图书在版编目(CIP)数据

SAS高级统计分析教程/胡良平主编. —2版. —北京：电子工业出版社，2016.1
(统计分析系列)
ISBN 978-7-121-27640-8

Ⅰ. ①S… Ⅱ. ①胡… Ⅲ. ①统计分析-应用软件-高等学校-教材 Ⅳ. ①C819

中国版本图书馆CIP数据核字(2015)第281679号

策划编辑：秦淑灵
责任编辑：苏颖杰
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
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：24.25　字数：621千字
版　　次：2016年1月第1版
印　　次：2016年1月第1次印刷
印　　数：3000册　　定价：55.00元

凡所购买电子工业出版社图书有缺损问题，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zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

服务热线：(010)88258888。

编 委 会

主　编　胡良平

审　校　高　辉

副主编　葛　毅　李长平　柳伟伟　胡纯严　郭　晋

编　委（以单位及姓氏笔画为序）

广东医学院	徐秀娟
山西医科大学	张岩波　罗艳虹
中日友好医院	周诗国
天津中医药大学	赵铁牛
天津医科大学	李长平
北京邮电大学	张晓航
军事医学科学院	李伍举　柳伟伟　胡良平　胡纯严
解放军卫生信息中心	葛　毅
解放军95969部队	高　辉
哈尔滨医科大学	李　霞　张瑞杰
济南军区疾控中心	李子建
首都医科大学	刘惠刚　罗艳侠　郭秀花　郭　晋
第三军医大学	伍亚舟　易　东
PPD医药公司	毛　玮

第2版前言

本书第1版于2010年6月由电子工业出版社正式出版后的2~3年间，经受住了市场的考验，得到了广大读者的褒奖、支持和鼓励。大约在2013年下半年，因市场上此书脱销，故出版社征求本人意见后，重印了数千册。最近我在主讲全国统计学培训班期间，听见部分学员说此书在全国各大网店缺货。对于可能出现脱销现象，在2014年年底，此书编辑早就预见到了，并敦促我抓紧时间修订。因为SAS软件内容更新很快，第1版主要基于SAS 9.1.3版本，而现在，SAS 9.2和SAS 9.3版本较之前增添了许多新功能。另外，本书第1版出版后，在使用过程中也发现了个别不妥和疏漏之处，有必要对其中的差错之处进行纠正；同时，有必要结合SAS软件的特点和实际使用统计学的需求，对全书内容和布局进行必要的修改、补充和调整。

由于增加的新内容比较多，修订后的篇幅大大超过了原书，本书编辑提出了一个大胆的方案——将第2版改为两册出版，书名分别为《SAS常用统计分析教程(第2版)》与《SAS高级统计分析教程(第2版)》，本书是后者。

本书的核心内容是现代多元统计分析(不包括多元方差和协方差分析、判别分析，这些内容已被收录入《SAS常用统计分析教程(第2版)》中)，其内容散布在第2~4篇共15章之中，分别冠以篇名“变量间相互与依赖关系分析”、“样品间亲疏优劣或相对位置分析”和“样品与变量或原因与结果之间的关联性分析”，以使条理更加清楚，便于读者学习和使用。

为了适应科学技术的发展对统计学提出的需求，书中保留了数据挖掘、生物信息学和遗传资料统计分析这三大领域方面的知识和技术，有利于开拓读者思路和激发其对统计学更广泛的兴趣。

在本书出版之际，我情不自禁地要提及，本室2013和2014级硕士研究生刘一松、胡完和郭春雪，他们对全书做了认真的校对。最后，请允许我感谢直接和间接为本书第1版和第2版付出过辛勤劳动的所有同志和朋友！

由于编者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。为便于与读者沟通和交流，特将我的电子邮箱地址和有关网址呈现在此：lphu812@sina.com；www.statwd.com；www.huasitai.com。

主编　胡良平
于北京军事医学科学院
2015年8月10日

第 1 版前言

众所周知，统计学内容非常丰富，学习和正确使用它难度很大；SAS 软件功能非常强大，实用面极宽，SAS 语言又十分繁杂，学习和全面掌握其用法并非易事；显然，实际工作者要想在较短的时间内，学会用 SAS 软件方便快捷且正确地解决各种实验设计、统计表达与描述、常见和多元统计分析、现代回归分析、数据挖掘和基因表达谱分析等问题，几乎是天方夜谭，然而，本书却使其成为现实！

笔者有何灵丹妙药呢？“面向问题”的思维模式和写作手法是解决复杂问题、并使其化繁为简、实用方便的“锦囊妙计”。本书各章针对拟解决的每个具体问题，首先给出“问题、数据和统计分析方法的选择”，接着，用编制好的 SAS 程序分析给定的资料，并给出程序修改指导、主要输出结果及其解释。为了降低书的出版费用，笔者千方百计对书稿进行精简，使以纸质印刷的篇幅约为原稿的三分之一，而原稿中三分之二的内容以附录的形式放在与本书配套的光盘上。本书正文内容共分 8 篇 47 章。第 1 篇 对定量结果进行差异性分析；第 2 篇 对定性结果进行差异性分析；第 3 篇 对定量结果进行预测性分析；第 4 篇 对定性结果进行预测性分析；第 5 篇 多变量间相互与依赖关系分析；第 6 篇 变量或样品间亲疏关系或近似程度分析；第 7 篇 数据挖掘技术与基因表达谱分析简介；第 8 篇 用编程法绘制统计图与实现实验设计。各章末尾均注明参编者详细名单。值得一提的是：本书中的所有计算基于 SAS 9.2 版本，少量 SAS 过程(特别是涉及 SAS/Genetics 模块)在 SAS 9.2 以前的环境下不能正常运行。

与本书配套光盘上的内容有：附录 1 与 SAS 语言有关的内容简介，包括第 48 章 SAS 语句简介、第 49 章 SAS 过程简介、第 50 章 SAS 函数简介、第 51 章 SAS 宏简介、第 52 章 SAS ODS 简介、第 53 章 SAS SQL 简介、第 54 章 SAS 数组简介和第 55 章 SAS/IML 简介；附录 2 四个非编程模块简介，包括第 56 章 SAS/ASSIST 模块用法简介、第 57 章 SAS/ANALYST 模块用法简介、第 58 章 SAS/INSIGHT 模块用法简介和第 59 章 SAS/ADX 模块用法简介；附录 3 数据挖掘技术与基因表达谱分析，包括第 60 章 数据挖掘的概念及常用统计分析技术、第 61 章 基因表达谱的概念及数据分析技术和第 62 章 生物信息学；附录 4 各章实例和数据；附录 5 各章 SAS 程序；附录 6 各章 SAS 输出结果；附录 7 各章计算原理和计算公式；附录 8 各章参考文献和附录 9 胡良平统计学专著与配套软件简介。

在本书即将出版之际，笔者真挚地感谢为本书作出过突出贡献的来自全国十多所高等院校的教授、副教授和青年学者，如山西医科大学张岩波、余红梅和郭东星教授，中山医科大学张晋昕教授，武汉大学毛宗福教授，哈尔滨医科大学李霞和张瑞杰教授，第三军医大学易东教授，首都医科大学郭秀花教授，北京大学医学部曹波副教授，北京邮电大学张晓航副教授等；还要感谢所有为本书付出过辛勤劳动的人们，他们的名字被写在本书的编委名单之中。特别是高辉、柳伟伟、周诗国、郭晋为本书的审阅工作付出了大量细致而又卓有成效

的劳动，而本室2009级四名硕士研究生（鲍晓蕾、贾元杰、关雪、梦冰）也为本书的校对做出了很多贡献。正是由于他们的积极参与、不懈努力和真心奉献，才使这部历时四年的专著能够问世！

由于笔者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。

主编　胡良平

于北京军事医学科学院

生物医学统计学咨询中心

2010年1月14日

目　录

第1篇　统计分析基础

第 2 篇 变量间相互与依赖关系分析

第 3 篇　样品间亲疏、优劣或相对位置分析

第 4 篇 样品与变量或原因与结果之间的关联性分析

第 5 篇 数据挖掘与分析

第6篇 遗传资料统计分析

附　录

第1篇　统计分析基础

第1章　应确保数据是值得分析的

一提起统计分析，很多人脑海中闪现出来的第一个问题是：应该选择什么统计分析方法分析面前的资料？这是对统计学一知半解的一个明显标志！统计分析应该被使用在“值得分析的资料”上，而不是被使用在任何资料上，更不是被使用在不知任何专业背景的所谓“大数据”上！

1.1　什么是数据和/或统计资料

1.1.1　数据不等于统计资料

【问题】　下面列出了300个数据，请问：利用这些数据可以做哪些统计分析？

0.6	0.8	0.9	1.1	1.2	1.4	1.5	1.6	1.8	1.9	2.1	2.2	2.4	2.5
2.6	2.8	2.9	3.1	3.2	3.4	3.5	3.6	3.8	3.9	10.1	10.2	10.4	10.1
4.2	4.2	4.3	4.4	4.5	4.5	4.6	4.7	4.8	4.8	4.9	5.0	5.1	5.2
5.2	5.3	5.4	5.5	5.5	5.6	5.7	5.8	5.8	5.9	6.0	6.1	6.2	6.2
6.3	6.4	6.5	6.5	6.6	6.7	6.8	6.8	6.9	7.0	7.1	7.2	7.2	7.3
7.4	7.5	7.5	7.6	7.7	7.8	8.1	8.1	8.2	8.3	8.3	8.4	8.5	8.5
8.6	8.7	8.7	8.8	8.9	8.9	9.0	9.1	9.1	9.2	9.3	9.3	9.4	9.5
9.5	9.6	9.7	9.7	9.8	9.9	9.9	10.0	10.1	10.1	10.2	10.3	10.3	10.4
10.5	10.5	10.6	10.7	10.7	10.8	10.9	10.9	11.0	11.1	11.1	11.2	11.3	11.3
11.4	11.5	11.5	11.6	11.7	11.7	11.8	11.9	12.1	12.2	12.2	12.3	12.4	12.5
12.5	12.6	12.7	12.8	12.8	12.9	13.0	13.1	13.2	13.2	13.3	13.4	13.5	13.5
13.6	13.7	13.8	13.8	13.9	14.0	14.1	14.2	14.2	14.3	14.4	14.5	14.5	14.6
14.7	14.8	14.8	14.9	15.0	15.1	15.2	15.2	15.3	15.4	15.5	15.5	15.6	15.7
15.8	15.8	16.1	16.2	16.3	16.3	16.4	16.5	16.6	16.7	16.8	16.8	16.9	17.0
17.1	17.2	17.3	17.3	17.4	17.5	17.6	17.7	17.8	17.8	17.9	18.0	18.1	18.2
18.3	18.3	18.4	18.5	18.6	18.7	18.8	18.8	18.9	19.0	19.1	19.2	19.3	19.3

19.4　19.5　19.6　19.7　19.8　20.2　20.3　20.5　20.7　20.8　21.0　21.2　21.3　21.5
21.7　21.8　22.0　22.2　22.3　22.5　22.7　22.8　23.0　23.2　23.3　23.5　23.7　24.2
24.4　24.7　24.9　25.1　25.3　25.6　25.8　26.0　26.2　26.4　26.7　26.9　27.1　27.3
28.3　28.6　28.9　29.1　29.4　29.7　30.0　30.3　30.6　30.9　31.1　31.4　32.4　32.8
33.2　37.6　34.0　34.4　34.8　35.2　35.6　27.8　27.7　37.7　38.2　38.8　39.3　 9.9
40.6　 9.9　41.9　42.5　45.3　46.5

【解答】 研究奥林匹克数学的人们会试图找出任何几个相邻数据之间的内在关系，最理想的结果是“推导”出一个“递推公式”，给定数据所在的序号 i，就能推算出 x_i 等于多少。而纯数理统计学家们会不假思索地说，可以绘制频数直方图或箱式图，形象化地呈现这些定量数据的频数分布规律；还可计算三种区间(即置信区间、容许区间和预测区间)，进而做出某些统计推断；若给定了标准值，还可对样本所抽自的总体均值与给定标准值之间差别是否具有统计学意义进行差异性分析，简称为单样本均值的假设检验。

其实，上面两种人所做的“统计工作”都有一个重要的前提条件，那就是此数据是值得分析的。然而，并非所有数据都是值得分析的！换一句话说，对数据进行分析之前，必须证明前提条件是成立的，否则，就是数字游戏！

事实上，值得分析的数据最起码要明确如下几条前提条件：收集这些数据的目的是什么？数据取自的总体是什么？这些数据分别在哪些条件下收集到的？数据的专业含义是什么？数据的名称是什么？数据的性质(定量或定性)是什么？若属于定量数据其单位是什么？测量数据的精确度有多高？由此可知，不明确这些前提条件的“一串数”可以称为“数据”；而明确这些前提条件的“一串数”才可以称为“统计资料”。

现在，全球有很多人在炒作“大数据”。大多数人或研究课题所呼喊的“大数据”，其实对应的正是前述提及的“数据”，而不是值得分析的“统计资料”！

1.1.2　统计资料的要素

1. 统计资料的基本要素

统计资料的基本要素有三条，即变量名称、专业含义和度量单位。例如，前面提及的那 300 个数据。若用变量 x 代表每个数据，且已知其专业含义为“尿汞值”，又知其单位是“μg/L”，此时就可以说，这 300 个数据提供的信息已基本清晰了，它们基本上可以被称为一组统计资料了。

2. 统计资料的全部要素

严格地说，仅仅知道一组统计资料的基本要素还是不能确定它是否值得分析。例如，前面提及的那 300 个数据，即使知道了它们是 x (尿汞值，μg/L)的具体取值，但不知道收集这些数据的目的是什么、这些数据所测自的受试对象是什么、这些受试对象是否来自同一个总体。如果研究目的是估计某地区正常成年人尿汞值的 80% 参考值范围，这个地区全部正常成年人构成了一个特定的总体，而这 300 个数据所测自的受试对象正是从这个总体中完全随机或按重要非试验因素分层随机抽取的(目的是为了提高样本对于总体的代表性)，那么便有理由认为，对这 300 个数据组成的统计资料进行各种统计表达与描述和各种统计分析，是值得的。

由此可知，统计资料的全部要素除了它应具备基本要素外，还应具备：有明确定义的与特