

序 一

统计学有很长的历史,在工农业、科技、商业、社会科学、政府决策等领域已得到普遍的应用,已经或正在成为社会发展的必不可少的工具。统计学的普遍应用只是近 40 年的事,得益于计算机的飞速发展。毫不夸张地说,没有计算机及统计软件的支撑,统计学的发展绝大部分只能停滞在理论研究上。一个新的统计方法,如果没有统计软件的支撑,是很难普及的。统计软件成为统计学发展的重要方面。

中国学者对统计软件的关注由来已久。早在“文化大革命”期间,由于地质勘探的需要,中国许多地质大队和大的油田花了大量的人力编制应用多元统计分析的统计软件。可惜编制这些软件的力量没有能够集中、提高,绝大多数团队自然衰退。以中国科学院前计算中心(现属中国科学院数学与系统科学研究院)的张建中、杨自强研究员为首的团队,一直在发展中国人自己的大型系统统计软件。在 20 世纪 80 年代,他们推出了统计软件包 SASD,对中国应用统计的发展起了很大作用。随着中国的改革开放,许多国际通用的大型统计软件包如 SAS、SPSS、MATLAB、R 等逐渐进入中国的统计教育和统计软件市场,中国人自己发展的各种统计软件包的国内市场却日益缩小。

中国是一个大国,需要有中国人自己发展的视窗系统,同时也需要具有自主知识产权的大型统计软件系统。这样的想法不是“闭关自守”,也不是“狭隘的民族主义”。

统计是一门大的学科,包括了众多的方法,如数据的描述性统计、单变量的统计分析(如 t 检验、 F 检验等)、多变量的统计分析、定性数据的数据分析、生存分析、实验的设计和建模、非参数统计等。一个大的统计软件包大多是由一个公司、一个研究单位或一所大学的一个大的团队来研制、发展、推销及服务。使我万分惊讶的是,DPS 统计软件是由唐启义教授个人用了 20 多年的时间,以“愚公移山”的精神研制和发展起来的。该软件涉及面广,使用方便,很受广大用户欢迎。唐启义教授坚持不懈的精神,令人感动。相信在唐启义教授的呵护下,DPS 统计软件会与时俱进,更加人性化,不仅在中国处处开花结果,而且能深入到全世界广大华语地区。



2009 年 5 月于中国珠海市

序 二

——青目睹人少，问路白云头

有所谓“文字的学问”，有所谓“数字的学问”。从科举制度迄今，文字学问的重要性不言而喻。相形之下，数字学问的普及与认知，尚有一段漫长的路。常言道：“数字会说话”，此言不假。但是当数字说话时，您听得懂、看得明白吗？统计学，作为研究数字学问的一门主要学科，在中国历史源远流长。相传在伏羲时代，人们便开始“结绳记事”（详见《周易》），《管子》“问篇”亦提及：“不明于数，欲举大事，如舟之无楫而欲行于大海也。”商鞅在公元前 390 年更提出：“强国需知十三数”，为统计的重要性立下经典。近来在商场上盛行的古代兵书《孙子兵法》更直言：“多算胜，少算不胜。”这些都说明统计在古代中国受重视的程度，数千年后的今天读起来依然意味深长。

计算机当属 20 世纪对人类历史推动最大的技术。自 1946 年在美国宾夕法尼亚大学诞生第一台计算机以来，计算机在功能上以飞快的速度发展，影响领域也从最开始的军事领域拓展到经济社会，以及其他各个领域。1974 年，Intel 公司发布了其面向个人机的八位微处理器芯片 8080。随着超大规模集成电路和微处理器技术的进步，计算机技术更从过去主要集中在大型机和小型机高端领域逐渐进入普通人的生活。自 20 世纪 80 年代初第一台个人计算机面世以来，计算机的成本快速下降。之后，随着互联网技术、多媒体技术的飞速进步，计算机技术快速发展。1993 年互联网开始商业化运行，1995 年 Windows 95 发布，日新月异的技术促进了计算机的快速发展与普及。1997 年 IBM 深蓝计算机战胜了人类国际象棋世界冠军，标志着人工智能发展日趋成熟。21 世纪以来，计算机发展更为迅猛，其与通信技术等的相互渗透和融合，大大推动了人类社会的进步。

计算机硬件的发展为统计学的发展提供了难得的历史机遇。统计学发展初期的方法虽然很好，但是由于计算工作量大，统计分析方法并没有在实践中有效推广开来。电子计算机的发展促进了软件的开发与商业化，推广了统计学的应用。具体来说，计算机从两个方面促进了统计学的发展。一方面，统计计算技术使得统计学家可以在计算机上验证自己的理论，推动了统计理论的发展；另一方面，大量统计软件的普及使得统计分析广泛应用到社会的各个方面，这些实践活动又推动了统计理论的发展。总之，现代统计学已经离不开计算机的发展，计算机已成为统计学发展的基本工具。

根据 1977 年 Kohm 在美国统计学会上的介绍,当时美国有 55 个公众可用的统计软件。1981 年,Francis 的著作中介绍了当时世界各国已有的统计软件 117 个。随着统计方法的不断改进,统计软件也在不断改进。70 年代百家争鸣以来,研发出了一批目前占据主流地位的统计分析软件,比如 SAS、MINITAB、BMDP、SPlus、SPSS 等。由于使用目的的不同,这些统计软件有着不同的使用人群。一般而言,BMDP 和 SPSS 比较受生物医药和社会科学领域人员的喜爱,主流的应用统计人员则更倾向于 SAS,而 SPlus 则更适合于进行统计研究的科研工作者。SAS 公司是目前全世界最大的统计软件公司。其他享有较好声誉的统计分析软件还有 SYSTAT, DataDesk 和 JMP 等。这些软件都是商业软件,由于比较昂贵,在教育界未能充分普及。20 世纪 90 年代末期,出现了类似于 S 语言的 R 软件,该软件平台是免费的,大大增加了其普及性,而由于其灵活性及强大的编程实现能力,该软件获得迅速发展和推广。

在中国统计学快速发展的过程中,统计分析软件必不可少。虽然 20 世纪 80 年代以来,各科研单位所编制的统计分析软件和程序数不胜数,但是随着统计学的发展,绝大多数统计软件由于各种各样的原因而渐渐淡出市场,使得国外较规范的统计分析软件如 SAS、SPSS、BMDP、GLIM 等占据了国内统计分析软件的大部分市场。但是,国内也产生了一批优秀的统计分析软件,如 NOSA、SPLM、Qstat、PEMS、Markway、DPS 等,其中 DPS 分析软件便是最为出色的统计分析软件之一。它立足国内市场,从初期主要用于处理农林生物科研数据起步,经过不断改进,至今已发展成为一款综合性的大型统计软件,尤其是它还开发了具有全球领先地位的均匀实验设计、混料实验设计、动态聚类分析几个功能模块的独创算法。如今,DPS 应用已遍及全国各地,涉及自然科学和社会科学各个领域,如农林牧渔、医药卫生、气象、水文地质、工程计算、市场调研、社会经济、金融等。

改革 30 年,我国经济一天比一天繁荣,但我们的教育与科研并没有明显的同步提升,距离国际水平尚有一大段发展空间。许多方面尚需努力与世界接轨,更遑论领导世界潮流。不过从另一个角度看,我们也不该一味迷信西方理论,而应该从我国国情出发,建立一套适合我国学术与教育的统计软件。这正证实了《孙子兵法》“虚实篇”所言:“致人而不致于人”。换言之,如何自力更生,建立一套属于中国人特质的统计分析软件,而不将就于外国人的思维与文化,进而进军国际市场,影响全世界。作为国内最为优秀的统计分析软件之一,DPS 统计软件研发始于 1988 年,经过 20 多年的努力,发展到现在的规模,填补了国内具有自主知识产权的大型统计软件的空白。

统计大师赵民德先生在《探索真相》一书的导读中引用了《鹿鼎记》第二十二回中韦小宝为了学“一指禅”就询于少林寺首座澄观和尚的一段对话。此话给我印象极深,特别引用如下:

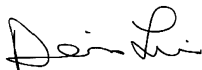
“韦小宝：‘你说那一指禅并不难学，可是从少林长拳练起，一路路拳法练将下来，练成这一指禅，要几年工夫？’

澄观道：‘这在般若堂的典籍中是有记载的。五代后晋年间，本寺有一位法慧禅师，生有宿慧，入寺不过三十六年，就练成了一指禅，进展神速，前无古人，后无来者。料想他前生一定是一位武学大宗师，许多功夫是前生带来的。其次是南宋建炎年间，有一位灵兴禅师，也不过花了三十九年时光。那都是天纵聪明、百年难遇的奇才，令人好生佩服。前辈典型，后人也只有神驰想象了。’”

统计学之深奥，多少人穷极一生都无法参透，吾等凡夫俗子又如何在有生之年得到真谛呢？《DPS[®]数据处理系统》让我们看到了一线曙光。该书内容涵盖之广，仿佛一本统计百科全书。全书8篇52章，包山包海，几乎包括了目前统计应用的主要方向。全书由浅入深，精辟易懂，配合软件操作，让一门高深的统计学可以在短期内得以理解并熟练地应用在生活周边的事务上。那统计学的“一指禅”在一年半载内借由该书得以学到六七成功力，不亦乐乎！

浙江大学唐启义教授是个认真的学者，同时也是热情而务实的工作者。从另一个角度看，唐教授是个理论扎实的统计学家，同时也是应用统计实务的执行者。更者，他是个计算机专家。通过他精心设计，妙笔生辉，该书充分反映了这些特质，也适时地满足了我国目前统计界的需求。

是以为序。

林共进 

2009年春写于美国宾夕法尼亚州

第 6 版前言

DPS 数据处理系统经过 30 多年的开发升级以及数以万计用户的应用反馈,目前已具备了较为完整、适用于各个专业试验设计及试验结果数据统计分析的能力。同时,DPS 数据处理系统已经可以在 Windows 系统及不同芯片架构的国产操作系统下运行。

DPS 数据处理系统反映了我国在具自主知识产权的统计分析软件开发方面实现核心技术的自主可控,解决了数据统计分析等关键领域长期依赖国外技术的问题,实现了各行各业数据统计分析的国产替代。

《DPS[®]数据处理系统》是基于我们开发的 DPS 数据处理系统计算机软件,介绍 DPS 数据处理系统软件试验设计及数据统计分析的工具书。目前版本是第 6 版,篇章结构仍沿用第 5 版。第 6 版主要新增内容有:

第一卷,完善了广义线性模型功能,将复杂的广义线性模型功能以方便操作的图形界面的形式提供给读者。另外,在第一卷中,作者增加了多指标 Kendall 趋势检验等功能。

第二卷,在 BP 神经网络建模中新增了目前常用的激励函数,以及对 BP 神经网络模型中各输入变量对输出变量的影响效应评估的改进。

第三卷,增加了大量的统计内容。这里介绍了作者在综合评价、多属性决策及机器学习中高维数据降维到一维方面的研究成果,即作者提出的最大熵-最小残差非线性数学模型。该数学模型解决了令人头疼的多指标综合评价、多属性决策过程中各个指标权重系数如何科学、客观、精确估计的世界性难题。在专业统计中,还增加了测量系统分析、统计质量控制、多因子分析的决策试验及评价实验室法(Dematel 法)和解释结构模型法(ISM)等内容,完善了 Meta 综合评价分析方法等。同时根据用户建议,对专业统计分析中一些统计功能结果输出进行了完善和充实。

本书各卷和前几版一样,用户界面中对于图中颜色的表述,均是 DPS 软件使用时在屏幕上的显示特征,和书中黑白印刷示意图有所差别,读者在阅读本书时需予以注意。

包括在国产操作系统下,各个版本的 DPS 数据处理系统软件试用版以及相关的帮助文档可从 DPS 数据处理系统网站(网址: www.dpsw.cn)下载。DPS 高级版下载安装之后在安装文件夹下面有一个“例子数据(高级版)”,里面有本书所附的所有例子数据,以供用户了解、练习 DPS 数据处理系统的功能。

DPS 数据处理系统既是我们 30 多年不懈努力的结晶,更是与广大 DPS 统计软件用户一如既往地反馈、支持密不可分。用户的需求是我们软件开发的动力!在此,我们对广大用户的使用、反馈致以衷心感谢。

本书再版得到了农业农村部“农作物病虫害监测、预警与防治服务”(代码: A120301)和杭州睿丰信息技术有限公司的资助。本书成稿过程中得到了作者所在单位和科学出版社的大力支持。这里作者感谢浙江大学生命科学学院冯明光教授和中国水稻研究所胡国文研究员多年来对作者工作的支持和帮助。对已故的浙江大学原副校长程家安教授、湖南省农业厅李绍石研究员、华南农业大学庞雄飞院士在作者从事科研的过程中给予的殷切关怀和鼓励致以衷心的感谢。同时对 30 多年来对作者在 DPS 软件开发、文稿撰写过程中给予帮助的各位朋友、同事、学生,以及我的家人,一一表示感谢。

最后,对 DPS 数据处理系统的广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心感谢。

唐启义

2025 年 11 月

第 5 版前言

这是一本介绍试验设计及数据统计分析的工具书,是基于我们开发的 DPS 数据处理系统计算机软件而成的。DPS 数据处理系统经过我们 30 多年的开发以及数以万计的用户的反馈、升级,已较好地满足各个专业试验设计及试验结果数据统计分析的需求。本书也反映了我们在具有自主知识产权的统计分析软件开发方面几十年来不懈努力的成果。

目前的第 5 版,篇章结构仍沿用第 4 版,主要新增内容有:

第一卷,改进了随机区组试验每个处理标准差、标准误的计算方法,即将区组效应去掉之后再计算;增加了两随机变量回归分析、Joinpoint 回归(折线回归)分析功能。

第二卷,这次再版主要新增了两个章节,即第 16 章“高维数据 Lasso 回归分析”和第 20 章“有序约束下的多峰识别模型”。前者既提供了高维数据在变量数量大(多重共线性)且大于样本数时的回归建模技术,又提供了自变量含有分类变量且有缺失数据情形下筛选重要因子、建立回归模型的途径。后者介绍了作者首次提出的有维度限制的多峰识别模型,它提供了解决一维、二维、三维及更高维有序样本的统计分析(如聚类分析)的技术。另外在时间序列分析章节中增加了向量自回归模型、格兰杰因果检验、协整检验和误差修正模型。并对 BP 神经网络算法进行了改进,使之可适合于含缺失数据的情形,并在 BP 神经网络模型中增加了估计各输入变量对输出变量的影响效应的功能。

第三卷,这次再版将面板数据一章调整到第二卷。功能上主要新增了水文气象频率分析,其中对 Wakeby 分布的参数估计提出了独创的新的算法,使拟合精度提高,进行百年一遇、千年一遇重现期估计时更精准。同时根据用户建议,对专业统计分析中一些统计功能结果输出进行了完善和充实。

本书各卷,和前几版一样,用户界面中对于图中颜色的表述,均是 DPS 软件使用时在屏幕上的显示特征。因此和书中黑白印刷示意图有所差别,读者在阅读本书时需予以注意。

DPS 数据处理系统软件试用版以及相关的帮助文档可从 DPS 数据处理系统网站(网址: www.dpsw.cn)中的“下载中心”下载。DPS 高级版下载安装之后在安装文件夹下面有一个“例子数据(高级版)”,里面有本书所附的所有例子数据,以供用户了解、练习 DPS 数据处理系统的功能。

DPS 数据处理系统既是我们 30 多年来不懈努力的结晶,更是广大 DPS 统计

软件用户一如既往地支持的成果。读者和用户的需求是我们软件开发的动力!

本书再版得到了国家重点研发计划项目(项目编号:2016YFD0201302)的资助。作者所在单位和科学出版社给予了大力支持。这里作者感谢浙江大学生命科学学院冯明光教授多年来对作者工作的支持和帮助。对浙江大学原副校长程家安教授、中国水稻研究所胡国文研究员及已故的湖南省农业厅李绍石研究员、华南农业大学庞雄飞院士对作者在 DPS 统计软件开发、升级,以及文稿写作过程中给予的殷切关怀和鼓励致以衷心的感谢。同时特别感谢郭若艳同志在这次再版过程中为校对文稿所付出的大量心血。最后对方开泰教授、王静龙教授、周仁郁教授、刘明芝教授、方萍教授、黄敬峰教授、张承恩高级工程师、董玉恒教授及夏结来博士等给予的帮助表示感谢。多年来浙江大学研究生高薇、孙传恒、唐洁、沈爱华、时嵩、彭奇、宛庭利、刘顺、罗巍等在协助文稿整理、图表编辑处理方面做了很多工作;浙江大学昆虫科学研究所、浙江大学环境毒理研究所的同事给作者工作以很大帮助,在此一并表示感谢。

最后,对 DPS 广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心感谢。

唐启义

2020 年春于浙江大学紫金港校区

第 4 版前言

光阴荏苒。DPS[®]数据处理系统软件及其配套专著自 1997 年首次出版以来,经过近 20 年不懈努力,DPS[®]数据处理系统软件得到了长足的发展。配套专著先后以《实用统计分析及其计算机处理平台》《实用统计分析及其 DPS 数据处理系统》《DPS[®]数据处理系统——实验设计、统计分析及数据挖掘》,以及三卷本的《DPS[®]数据处理系统》配套专著形式出版,其篇幅亦从 60 多万字增加到目前的近 200 万字。

目前的第 4 版,篇章结构仍沿用第 3 版,主要的新增内容有:

第一卷,在相关和回归分析(第 6 章)中增加了 Deming 回归、具重复测量试验回归;在分类数据模型分析(第 8 章)中增加了广义线性模型;非参数统计(第 9 章)增加了 Passing-Bablok 回归;混料试验设计与分析(第 12 章)增加了极端顶点设计功能。另外,在第 2 章、第 4 章和第 5 章改进了数据分析方法、用户界面,增加了各个处理直方图比较,以及模型残差作图分析。

第二卷中,作者对相关、回归分析章节内容按是否是多个因变量进行了重组,将典型相关分析、多重筛选逐步回归、偏最小二乘回归分析及线性联立方程合并成“多因变量统计分析”一章,并将原来结构方程模型与路径分析放到本章节。统计分析功能方面,作者改进了逐步回归分析和双重筛选逐步回归分析的用户界面,优化了双重筛选逐步回归相关算法,并在时间序列分析章节中新增了单位根检验等内容。

第三卷中的专业试验统计,增加了面板数据统计分析、多点试验统计分析、格子设计统计分析等方法。并根据用户建议,对一些统计功能结果输出进行了完善、充实,改进许多 DPS 使用时的用户界面。

本书各卷中,对于图中颜色的表述,均是 DPS 软件使用时在屏幕上的显示特征。因此和书中黑白印刷示意图有所差别,读者在阅读本书时需予以注意。

DPS 数据处理系统软件试用版以及相关的帮助文档,可从 DPS 数据处理系统网站(网址:<http://www.dpsw.cn>)中的“下载中心”下载。注意 DPS 数据处理系统有高级版(最新版本)和标准版(2007 年底以前)两个版本。DPS 系统高级版下载安装后,在安装文件夹下面有一个“例子数据(高级版)”文件夹,其中有本书所附的所有例子数据,供用户了解、练习 DPS 数据处理系统及其使用。

在广大读者和 DPS 统计软件用户支持下,经过这三年多的努力,DPS 系统功能又得到了进一步的增强。作者衷心希望广大读者一如既往地支持 DPS 这一国

产统计分析软件,读者和用户的需求就是我们软件开发的动力!读者和用户在实验设计、统计分析和数据挖掘方面如遇到困难,可以通过访问我们为 DPS 提供的专门网站 <http://www.dpsw.cn>,即可得到技术咨询和软件升级。

本书再版得到了国家重点研发计划项目(2016YFD0201300)的资助。作者所在单位和科学出版社给予的大力支持。这里,作者感谢浙江大学生命科学学院冯明光教授多年来对作者工作的支持和帮助。对浙江大学原副校长程家安教授、湖南省农业厅李绍石研究员、中国水稻研究所胡国文研究员及已故的中国科学院庞雄飞院士对作者在 DPS 统计软件开发、升级,以及本书写作过程中给予的殷切关怀和鼓励致以衷心的感谢。并对方开泰教授、王静龙教授、周仁郁教授、刘明芝教授、方萍教授、黄敬峰教授、张承恩高级工程师、董玉恒教授及夏结来博士等给予的帮助表示感谢。多年来浙江大学研究生高薇、孙传恒、唐洁、沈爱华、时嵩、彭奇、宛庭利、刘顺、罗巍等在协助文稿整理、图表编辑处理方面做了很多工作;浙江大学昆虫科学研究所、浙江大学环境毒理研究所的同事给作者的工作以很大帮助,在此一并表示感谢。

最后,对 DPS 广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心感谢。

唐启义

2016 年夏于浙大紫金港

第 3 版前言

《DPS[®]数据处理系统》(第 3 版)是作者基于自主开发的 DPS 统计软件而撰写的,以实用为目的的统计工具书。

随着 DPS 统计软件功能的不断完善,相应配套专著已变得越来越厚。许多 DPS 用户、读者反映书本太厚了一些,使用起来不甚方便。因此,这次出版对版面进行了较大调整,并将整个书稿分成三卷出版,第一卷为基础统计及实验设计;第二卷为现代统计及数据挖掘;第三卷为专业统计及其他。

第一卷包括了第二版的第一篇和第二篇及原来“抽样技术”一章,并将原“概率统计模型”一章中的加权列联表分析、多因子综合相关分析放在本卷第 8 章“分类数据模型分析”里面。

在第一卷中,新增了相关和回归分析(第 6 章),系统介绍了相关和回归分析原理、应用及应用中需注意的一些问题。另外,在第 1 章里面增加了“工作表中图形输出”和“DPS 系统应用常见问题解答”两节,在第 2 章里面增加了“缺失值处理”,在第 7 章里面增加了“Poisson 分布抽样下多样本检验”,在第 11 章“多因子优化设计与分析”里面,改写了实验设计工作界面,增加了区组设计的统计分析内容。

第二卷主要包括了第二版的第三篇(多元统计分析)、第四篇(数学模型模拟分析)和第七篇(时间序列分析)以及第八篇中的第 43 章“神经网络和支持向量机”。

在第二卷中新增的内容主要是第 19 章“结构方程模型”,介绍了如何在 DPS 系统下,借助良好的用户界面,实现结构方程的建模与分析。另外,在第 15 章里面增加了“线性联立方程”一节,并以第 2 版第 43 章“神经网络和支持向量机”为主体,增加了“随机森林”,将该章命名为“数据挖掘”。在第 3 版的第 21 章“回归方程模型”主要是原第 2 版的第 30 章“非线性回归模型”内容,但充实了许多内容,如联立方程模型功能得到加强。在时间序列分析领域,增加了时间序列数据的游程检验、谐波分析、奇异谱分析等功能。

第三卷主体内容为第 2 版的第三篇“专业实验统计”、第六篇“常用数值分析”和第八篇“其他数据分析方法”。

在专业实验统计中,这次我们主要增加了生物测定及药物代谢统计学分析、品种区试 AMMI 模型分析作图技术,并根据用户建议,对一些统计功能结果输出进行了完善、充实,修改了许多 DPS 使用时的用户界面。

在本书各卷中,对于图中颜色的表述,均是 DPS 软件使用时在屏幕上显示特征的描述。因此和书中黑白印刷的示意图有所差别,请读者在阅读本书时需予以

注意。

在广大读者和 DPS 统计软件用户的支持下,经过作者三年多的努力,DPS 系统功能又得到了进一步增强。希望广大读者一如既往地继续支持 DPS 这一国产统计分析软件,读者和用户的需求就是我们软件开发的动力!读者和用户在实验设计、统计分析和数据挖掘方面如遇到困难,可以通过访问我们为 DPS 提供的专门网站 <http://www.dpsw.cn> 得到技术咨询和软件升级。

本书这次再版得到了国家重点基础研究发展计划(“973”)课题(编号:2010CB126201)部分资助。作者所在单位和科学出版社给予了大力支持。这里,作者感谢浙江大学生命科学学院冯明光教授多年来对作者工作的支持和帮助。对浙江大学原副校长程家安教授、湖南省农业厅李绍石研究员、中国水稻研究所胡国文研究员及已故的中国科学院庞雄飞院士对作者在 DPS 统计软件开发、升级,以及本书写作过程中给予的殷切关怀和鼓励致以衷心的感谢。对方开泰教授、王静龙教授、周仁郁教授、刘明芝教授、方萍教授、黄敬峰教授、张承恩高级工程师、董玉恒教授及夏结来博士等给予的帮助表示感谢。浙江大学 2004 级软件工程专业唐睿和外语系沈宁宁协助完成了 DPS 数据处理系统软件英文版的翻译与转换工作;研究生宛庭利、刘顺、罗巍,湖南农业大学实习生徐西林、刘云协助第 3 版书稿里面的图表编辑处理;浙江大学昆虫科学研究所、浙江大学环境毒理研究所的同事给作者的工作以很大帮助,在此一并表示感谢。

最后,对 DPS 广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心感谢。

唐启义

2013 年春于浙大紫金港

第 2 版前言

《DPS[®]数据处理系统——实验设计、统计分析及数据挖掘》(第 2 版)是作者基于自主开发的 DPS 软件,以实用为目的的统计工具书。

根据读者、用户建议,这次再版除对原有的一些内容进行修改、补充外,还增加了作者在过去两年研发出来的若干新功能的介绍。全书新增了 3 章共 30 多节,并对一些章节的顺序进行了调整。新增的 3 章是:

第 11 章“混料实验设计与分析”。在第 1 版的均匀实验设计技术取得的重大进展的基础上,本章介绍了自主开发的、较目前已有的混料实验设计方法更有效的混料实验设计的新技术,该技术能适合各种情形的混料设计,特别是组分多、实验次数大、限制条件“苛刻”的情形。同时提供了混料实验数据分析和优化的数学模型,为工业、农业和科学实验中应用混料设计提供了新的技术。

第 12 章“数据包络分析与随机前沿面分析”。主要用来评价相同类型部门或单位(称为决策单元)中各成员之间的相对有效性,其中包括非参数的数据包络分析(data envelopment analysis, DEA)和属于参数方法的随机前沿面分析(stochastic frontier analysis, SFA),并给出了在实际工作中计算生产效率时应用较多的 Malmquist 指数。

第 16 章“序贯实验分析”。在进行下一次实验以前,已经知道前面实验的结果,这种一次接着一次的抽样实验为序贯实验。本章提供了在序贯实验中计算各接收时刻的接收概率和各时间区段的拒收概率的方法,并给出了工作特性曲线。

除这 3 章外,还在方差分析、非参数检验、多因子优化与分析、量表分析和顾客满意度指数模型、多变量统计检验、回归分析、多因子分析、非线性回归模型、时间序列及多指标综合评价等 10 多个章节,增加了新的统计分析功能,如实验设计和方差分析中的 Box-Behnken 设计、平衡不完全区组设计实验方差分析、增广随机区组设计及方差分析、多元方差分析的线性模型方法,回归分析中的稳健回归、积分回归和主导分析,多变量时间序列的独立分量分析,以及结合分析、展开法、非参数回归、支持向量机和投影寻踪分类等。

经过这两年多的努力,尽管 DPS 系统功能又进一步增强,但距离广大读者的要求仍差距甚大。希望广大读者一如既往地继续支持 DPS 这一国产统计分析软件,及时指出目前版本的缺点和不足,以便我们继续改进,加速开发新的数据分析技术。读者可通过访问网站 <http://www.chinadps.net> 得到技术咨询和软件升级。

本书再版得到了国家高技术研究发展计划(863 计划)项目(2006AA10Z217)和农业公益性行业科研专项(200803003)的资助。在本书出版过程中,作者所在单位和科学出版社给予了大力支持。这里,作者特别感谢浙江大学生命科学学院冯明光教授多年来对作者工作的支持和帮助。对浙江大学原副校长程家安教授、湖南省农业厅李绍石研究员、中国水稻研究所胡国文研究员及已故的中国科学院庞雄飞院士对作者在 DPS 统计软件开发、升级,以及本书写作过程中给予了殷切关怀和鼓励致以衷心的感谢。并对方开泰教授、王静龙教授、周仁郁教授、刘明芝教授、张承恩高级工程师、董玉恒教授及夏结来博士等给予的帮助表示感谢。浙江大学软件工程专业唐睿和外语系沈宁宁协助完成了 DPS 数据处理系统软件英文版的翻译与转换工作;研究生时嵩、彭奇、高薇、缪清玲协助完成了本书文稿的文字处理;浙江大学植物保护系、昆虫科学研究所农业部作物病虫分子生物学重点开放实验室以及浙江省农业遥感与信息技术重点研究实验室的同事给作者的工作以很大帮助,在此一并表示感谢。

最后,对 DPS 广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心感谢。

唐启义

2009 年春节于杭州华家池

第 1 版前言

呈现在读者面前的《DPS[®] 数据处理系统——实验设计、统计分析及数据挖掘》是作者于 2002 年出版的《实用统计分析及其 DPS 数据处理系统》的补充和发展。和原来版本一样,它既是一本以实用为目的的统计工具书,更是一份请柬。它邀请读者进入 DPS 数据处理系统漫游,寓复杂繁冗而又沉闷乏味的统计分析于轻松自如和无穷乐趣之中。

现代科学研究,不管是自然科学还是社会科学的各个领域(物理、化学、工程、医学、生物、经济、社会等),只要存在带有随机性的数据,就需要应用数理统计方法去有效地收集、整理和分析,以便得出正确的结论,为相关决策行动提供科学依据。

统计学的形成可以追溯到 20 世纪 20 年代。当时 Fisher 等以农业科学实验为对象,探讨实验资料的整理和统计分析方法。1922 年,Hayes 研究亲代与子代蛋白质含量的相关性,提出了相关分析的理论。翌年,Fisher 和 Maekenzie 研究肥料对马铃薯产量的影响,首次提出了方差分析和交互作用的概念。1925 年,Engel-dow 研究不同品种在不同地区和不同年份的产量表现时,进一步阐述了因素间的交互作用。Fisher 等正是在研究农业科学实验方法的基础上,建立起了统计学基本理论和方法。这些已被广泛地应用于科学研究、工农业生产和现代经济管理。

随着经典数理统计科学的发展,在 20 世纪四五十年代,多元统计分析在 Pearson 及 Fisher 等人开拓性工作基础之上又有了坚实的理论基础,如关于多元正态分布的理论著述、Anderson(1958)的《多元统计分析导论》等著作,标志着统计学进入到一个新的发展阶段。但是,多元统计分析的发展、完善和广泛应用,离不开计算机科学和计算数学的发展。由于多元统计分析的计算工作量十分惊人,没有计算机是不可想象的。正是由于计算机技术的飞速发展,一些独特的多元统计分析技术,如偏最小二乘法、主成分分析、典型相关(又称“典范相关”)、因子分析、对应分析、判别分析、聚类分析等,在农林、医药、气象、地质、生物、心理、工业、经济、社会等许多研究和生产领域获得了广泛应用。目前,多元统计分析已是统计学中最活跃的分支学科之一。

传统的自然科学和社会科学研究以定性描述为主。Fisher 等创立的一套经典统计方法在定量研究方面迈出了一大步。但囿于其固有缺陷,实验结果的统计推断不能对有关数量关系从机理上予以合理解释,因而其统计模型仍属经验范畴。因此,现代科学工作者们正力图使具体学科的理论及研究对象同数学有机地结合起来,以模型来描述、解释自然现象和社会、经济的运动规律,从而解决科研和生产

中的实际问题。

自 20 世纪六七十年代以来,一些新的统计理论和分析技术发展起来。如 20 世纪 60 年代 Zadeh 提出的模糊集合论在各个领域的应用已十分广泛。实践证明,模糊数学方法在农业、图像识别、天气预报、地质地震、交通运输、医疗诊断、信息控制、人工智能等领域的应用已初见成效,其发展前景广阔。作为预报和控制的有力工具,研究系统变化规律的时间序列分析技术目前正被广泛应用于地质、石油、气象、商业、工程控制等领域,在农业中已被用于作物病虫害的中长期预测预报,并获得了可喜的成功。灰色系统是我国学者邓聚龙先生于 20 世纪 80 年代初提出的用于控制和预测的新技术,目前亦在我国农业、社会和经济等领域广泛应用,并已取得显著成就。

统计学的发展源远流长,内容十分丰富。迄今为止,国内外统计学专著和教科书已有多种版本见诸于世。如果按照一般统计学著述的内容和体裁,我们没有必要再推出一本教科书似的理论专著,让广大科技工作者在复杂的数学公式堆中艰难跋涉。本书的编写宗旨强调实用性。作者在内容上刻意求新,形式上力求通俗易懂,在取材上尽可能全面,并注意反映国内外在实验设计和统计分析方面的最新研究进展。书中介绍的 400 多种统计分析方法,可以满足各个领域的统计分析需要,且都能在作者设计制作的 DPS 计算机数据处理系统中随意调用,自如操作。本着服务于读者、服务于用户的最高宗旨,作者对入选的 400 多种统计分析方法分门别类,每一类(种)统计分析技术按照方法或原理简介与操作示例两个基本部分编写,全书共分 8 篇 41 章。同时,作者按书中篇目类别设计 DPS 系统的套叠下拉式主菜单和子菜单,将带给读者和用户极大的便利。下面分篇给予简单介绍,以引导读者高效率地使用书中的各种分析技术和 DPS 数据处理系统。

第一篇主要介绍 DPS 系统的性能、特点和操作要领。DPS 数据处理系统是作者独立研究开发的大型通用多功能数据处理分析应用软件系统。该系统在 IBM PC 及其兼容机、中文 Windows 98/2000/XP 下运行。因此,在使用 DPS 系统之前,建议读者首先浏览本篇内容,熟悉该系统将使读者在数据处理和统计分析中事半功倍。

第二篇的编写充分考虑到经典实验统计的完整性和系统性,提供了比较全面的实验设计和实验结果统计分析的内容,如 t 测验、方差分析、列联表分析、非参数检验、最优回归实验设计与分析以及多元方差分析等,尤其对方差分析给予了较为详尽的介绍。因此,各个层次、各个领域的读者都可使用这些基本的统计学分析方法。

第三篇介绍了一些应用于某些较特殊专业领域的统计分析技术,如心理统计中的量表分析、顾客满意指数的结构方程模型分析、实验评价的 ROC 曲线分析、地理统计分析,还为育种学工作者提供了品种区域实验的统计分析技术,为生态学

工作者提供了动植物种群数量抽样、动植物种群数量空间分布研究、生物群落的消长演替过程分析、生存分析等方面的内容;对从事生物测定研究的科技人员来说,作者既提供了经典 Finney 概率分析方法,还重点引入近几年才逐渐发展起来的更为完善的分析工具,即时间—剂量—死亡率模型(TDM)分析技术。

第四篇以较大篇幅介绍多变量统计分析技术,并注意反映目前国内外的最新研究动态,将最新的技术提供给读者。近年来,多变量统计分析技术已被广泛应用于各个领域的科学实验中,因为人们越来越多地将研究对象作为一个有机结合的整体即多元素系统(亚系统)来对待,而不是分割开来作为单独事物处理。“从树木到森林”,抓住系统运动变化的本质,舍弃次要因素,通过简化系统的结构,探索系统运动变化的总趋势,为人类征服自然、改造自然服务。因此,偏最小二乘回归分析、主成分分析、因子分析、对应分析、典型相关分析、聚类与判别分析以及马尔可夫链分析等多因子分析技术备受自然科学和社会科学各领域科研人员的重视。在 DPS 中还提供了我们研发的含有定性变量的逐步回归分析技术。

第五篇和第六篇向读者介绍相当灵活的通用数学模型的模拟分析和数值计算工具。因前面介绍的分析方法和程式基本上都有完整的体系和要求,读者必须按部就班地照着去做。而利用第五、六篇的通用工具,读者将尽可能地发挥自己的专长和展现自己的学识水平,实现思维活动的飞跃。例如,模拟作物生长发育过程或经济活动的演替规律,建立结合自己专业和学科特点的数学模型,而不是套用别人的现成模型。读者可以借助 DPS 系统强大的数学建模功能,在计算机上实现自己的思维活动,获得所想即所见、所见即所得的效果。

第七篇介绍时间序列分析技术。时间序列分析是根据系统观测得到的时间序列数据,通过曲线拟合和参数估计来建立数学模型的理论和方法,常用于国民经济宏观控制、区域综合发展规划、企业经营管理、市场潜量预测、气象预报、水文预报、地震前兆预报、农作物病虫害灾害预报、环境污染控制、生态平衡、天文学和海洋学等方面。

第八篇为读者提供一些不属于经典统计学但近年来发展迅速且应用已经相当广泛的模型分析技术,即模糊数学分析、灰色系统分析和神经网络模型技术。只要读者有兴趣,DPS 系统可以帮助读者熟悉和掌握上述各项分析技术,并应用到研究实践中去。

综上所述,本书和 DPS 数据处理系统是密不可分的整体。作为通用的数据处理和统计分析的工作平台,作者在系统的设计方面注重专用性与通用性以及“傻瓜”性和灵活性的关系,以适应不同层次读者的需要。

专用性和通用性。DPS 的有些功能是专为某些领域或完成某一类工作而专门设计的,如“生物测定”是专为从事化学或生物制剂的生物活性测定的人员而设计的;种群空间分布型及群落参数统计分析、品种区域实验等章节都是服务于某个领

域的科学工作者。相对来说,它们的用途比较专一。但是,本书中绝大部分统计分析和建模技术都是通用的,例如丰富的数学函数、交互方式的矩阵运算功能以及通用数学模型的模拟分析技术。DPS 处理数学公式的功能非常独特,有利于读者发挥潜在的想象力和创造力。同时,数学计算功能,特别是矩阵交互式处理功能,在当前微机普及的情况下,将有助于高等院校数理统计课程的教学,使繁冗无味的矩阵运算得以在计算机上精确而快速地进行,并且在一定程度上提高学生的学习积极性和教学质量。

“傻瓜”性和灵活性。在 DPS 平台上进行经典统计分析的过程十分简单。它无需读者懂得数学推导,无需读者了解运算细节,只需读者将自己的实验数据输入电子表格,将它们定义成数据块(移动鼠标将待分析的数据“框”起来),然后调用分析功能(用鼠标点选菜单),按回车键就可在当前光标下面得到分析结果。虽然有些较复杂的统计分析需要读者输入参数,但系统会自动给以提示,如用户事先没有相关的经验,系统会给出缺省值,用户只管用鼠标点击“确定”按钮,就可以获得较优的分析结果;又如时间序列分析,系统以直观的图形方式让读者确定参数,只要读者稍有常识,无需牢记一大堆统计分析过程中的参数辨识准则,就可成为高水平的建模能手;再如在系统聚类分析中,聚类结果一次成“像”,给出的是非常直观的聚类谱系图等。系统的这些功能,好比一只傻瓜型照相机,使用时只需按快门。另一方面,DPS 系统在数据处理和模型模拟中又比较灵活,完全取决于读者的经验和知识水平。如数据行列的转换功能、通用方差分析的广义线性模型的应用、自定义数学模型的拟合、逐步回归及逐步判别分析中 F 检验临界值的随意调整等等,使读者大有用武之地,挥洒自如。

应当指出,本书不是一本基础理论教材,而是一本不需要高深数理统计知识即可看懂的统计方法论专著。事实上,要把本书中每种统计方法的理论背景在这样一本书中完全交代清楚是不太可能的,因为有些章节本身就是一本书的内容。作者的主要目的是要把读者从繁琐的统计分析中解放出来,既要用最适合、最先进的统计分析技术为自己的研究工作服务,又不被它们复杂的数学公式和运算所困扰。如果本书及其 DPS 数据处理系统能使读者在数据处理和统计分析方面有所获益或帮助,作者的基本目的就达到了。

DPS 统计软件相关图书自 1997 年中国农业出版社第一次出版以来,已经历了 10 个不平凡的年头。在这 10 年中,国内外统计软件品种繁多,竞争十分激烈,但由于有广大用户的大力支持,DPS 终于突破重围,在国内应用统计软件领域占有了一定地位。为此,我们对 DPS 广大用户以及关心和扶持 DPS 成长的各个领域的专家、学者表示衷心的感谢!

经过 4 年的努力,尽管 DPS 系统的功能又进一步增强,但离广大读者的要求还有较大差距。作者希望广大读者继续支持 DPS 数据处理系统这一国产统计分

析软件,使之不断完善。在使用过程中如发现缺点和不足请及时反馈给作者,以便再次修订时改进。作者将不辜负广大读者的期望,努力工作,不断充实新内容。为方便广大读者、用户使用 DPS 数据处理系统,及时解答用户在应用中的技术问题,加快更新 DPS 软件步伐,使之日臻完善,特设立了为 DPS 统计软件提供技术支持的网站 <http://www.chinadps.net> 和 <http://www.statforum.com>。读者可通过访问该网站及时得到 DPS 使用技术咨询和软件的升级服务。

本书的出版得到国家自然科学基金(30370914)和国家 863 计划项目(2002AA243041)的资助。在本书修订和出版过程中,作者所在单位和科学出版社给予了大力支持;浙江大学原副校长程家安教授、湖南省农业厅李绍石研究员、中国水稻研究所胡国文研究员及已故的中国科学院庞雄飞院士对作者在 DPS 统计软件开发、升级,以及本书写作过程中给予了殷切关怀和鼓励;香港浸会大学方开泰教授、华东师范大学王静龙教授、成都中医药大学周仁郁教授、湖南中医学院刘明芝教授以及第四军医大学夏结来博士等给予大力帮助;研究生孙传恒、唐洁和沈爱华协助本书文稿的文字处理;浙江大学昆虫科学研究所的各位同事给作者工作以很大帮助,作者在此一并表示感谢。

作 者

2006 年 3 月于杭州

目 录

序一

序二

第 6 版前言

第 5 版前言

第 4 版前言

第 3 版前言

第 2 版前言

第 1 版前言

第三篇 多元统计分析

第 14 章 多变量统计检验	649
14.1 多元计量资料的常用统计量	649
14.2 多元均值检验	651
14.3 多元方差分析简介	653
14.4 单向完全随机设计	656
14.5 单因素随机区组设计	658
14.6 轮廓分析	660
14.7 多元方差分析的线性模型方法	664
参考文献	671
第 15 章 多元回归分析	673
15.1 多元线性回归	673
15.2 逐步回归分析	683
15.3 二次多项式回归分析	694
15.4 含定性变量的逐步回归分析	698
15.5 积分(逐步)回归	703
15.6 趋势面分析	709
15.7 Tobit 回归	713
15.8 主成分回归	718
15.9 岭回归	721
15.10 特征根回归	726

15.11	稳健回归 M 估计	729
15.12	分位数回归	732
15.13	优势(主导)分析	737
15.14	分层逐步回归	740
	参考文献	750
第 16 章	高维数据 Lasso 回归分析	752
16.1	Lasso 回归分析简介	753
16.2	建立 Lasso 回归模型的弹性网络算法	754
16.3	高维数据线性回归分析	756
16.4	高维数据 Logistic 回归分析	764
16.5	高维数据 Poisson 回归分析	770
16.6	高维数据 Cox 比例风险回归分析	776
16.7	含分类变量的 Lasso 回归分析	781
16.8	含分类变量的 Lasso Logistic 回归分析	792
	参考文献	795
第 17 章	面板数据分析	796
17.1	面板数据分析基本模型	796
17.2	空间面板数据模型	805
	参考文献	811
第 18 章	多因变量统计分析	813
18.1	典型相关分析	813
18.2	双重筛选逐步回归	821
18.3	偏最小二乘回归	827
18.4	线性联立方程	832
18.5	结构方程模型与路径分析	841
	参考文献	853
第 19 章	聚类分析	855
19.1	系统聚类分析	855
19.2	0-1 型变量聚类分析	870
19.3	动态聚类分析	873
19.4	有序样本分类	879
19.5	非线性映射分析	883
19.6	两维图论聚类	886
	参考文献	889
第 20 章	有序约束下的多峰识别模型	891

20.1 问题的提出	891
20.2 一维有序约束下的多峰模型	894
20.3 多个受约束变量的综合	898
20.4 多峰识别模型在 DPS 系统中的实现	901
20.5 二维有序约束下的峰值识别模型	910
20.6 高维有序空间多峰识别模型	918
20.7 多维有序约束下数据分析与经典聚类算法的关系	924
参考文献	925
第 21 章 判别分析	927
21.1 两组判别	927
21.2 Fisher 线性判别	930
21.3 逐步判别分析	936
参考文献	944
第 22 章 多因子分析	945
22.1 主成分分析	945
22.2 因子分析	952
22.3 对应分析	967
22.4 展开法	972
参考文献	975
第 23 章 数据挖掘	977
23.1 BP 神经网络	977
23.2 径向基函数(RBF)网络模型	989
23.3 投影寻踪回归	993
23.4 支持向量机(SVM)	1000
23.5 随机森林	1011
参考文献	1018

第四篇 数学模型模拟分析

第 24 章 回归方程模型	1021
24.1 任意非线性回归模型参数估计实现	1021
24.2 非线性回归模型参数估计方法	1023
24.3 非线性回归模型参数全局优化估计	1026
24.4 非线性回归分析实例研究	1029
24.5 二值反应变量模型参数估计	1044
24.6 有约束条件模型参数估计	1051

24.7 联立方程模型	1055
参考文献	1063
第 25 章 数学模型模拟与优化	1066
25.1 模型模拟分析	1066
25.2 模型参数灵敏度分析	1071
25.3 模型优化	1073
参考文献	1077
第 26 章 数学规划	1078
26.1 线性规划	1078
26.2 多目标线性规划:评价函数法	1083
26.3 多目标线性规划:逐步宽容约束法	1087
26.4 多目标线性规划:分层评价法	1090
26.5 整数规划及混合整数规划	1095
26.6 指派问题匈牙利法	1101
26.7 运输问题	1104
26.8 非线性规划	1107
26.9 投入产出分析	1110
26.10 目标规划	1114
参考文献	1122
第 27 章 状态空间模型	1123
27.1 线性控制系统能控性	1124
27.2 线性控制系统能观性	1126
27.3 连续线性状态方程离散化	1128
27.4 离散状态方程求解	1130
参考文献	1133

第五篇 时间序列分析

第 28 章 时间序列趋势分析	1137
28.1 常用时间序列趋势分析	1137
28.2 时间序列数据游程检验	1145
28.3 最优气候均态模型	1147
28.4 均值生成函数预测模型	1149
28.5 马尔可夫链	1152
28.6 多元时空序列马尔可夫链分析	1155
参考文献	1160

第 29 章 时间序列周期分析	1161
29.1 谐波分析	1161
29.2 小波分析	1164
29.3 奇异谱分析	1167
29.4 时间序列周期方差分析外推法	1171
29.5 季节性水平模型	1174
29.6 季节性交乘趋势模型	1177
29.7 季节性叠加趋势模型	1181
参考文献	1184
第 30 章 平稳时间序列分析	1185
30.1 取样间隔与插值处理	1185
30.2 数据序列突变点的检测	1187
30.3 数据序列统计特性估计	1192
30.4 单位根检验	1198
30.5 差分自回归移动平均 (ARIMA) 模型	1205
30.6 季节-周期组合模型	1220
参考文献	1226
第 31 章 其他时间序列模型	1228
31.1 多变量时间序列 CAR 模型	1228
31.2 向量自回归模型	1234
31.3 格兰杰因果关系检验	1249
31.4 协整检验和误差修正模型	1255
31.5 协整和向量误差修正模型的最大似然估计	1262
31.6 门限自回归模型	1271
31.7 独立分量分析	1275
参考文献	1280

第三篇 多元统计分析

多元统计分析是运用数理统计方法来研究解决多指标问题的理论和方法。构成多元统计分析模型的数学方法并不新颖,如与多变量有关的基本概率分布——正态分布源自 20 世纪 30 年代。然而,当随机变量较多时,多变量分析的计算工作量极其繁冗,没有计算机根本无法完成。因此,直到有了计算机之后,多变量分析技术才进入实用阶段并迅速发展。近年来,随着计算机技术发展和科研生产的迫切需要,多元统计分析被广泛地应用于地质、气象、水文、医学、工业、农业和经济等许多领域,已经成为解决实际问题的有效方法。在采用多元统计分析进行数据处理、建立宏观或微观系统模型时,需要研究以下几个方面的问题:

- 简化系统结构,探讨系统内核。可采用主成分分析、因子分析、对应分析等方法,在众多因素中找出各个变量最佳的子集合,从子集合所包含的信息描述多变量的系统结果及各个因子对系统的影响。“从树木看森林”,抓住主要矛盾,把握主要矛盾的主要方面,舍弃次要因素,以简化系统的结构,认识系统的内核。

- 构造预测模型,进行预报控制。在自然和社会科学领域的科研与生产中,探索多变量系统运动的客观规律及其与外部环境的关系,进行预测预报,以实现系统的最优控制,是应用多元统计分析技术的主要目的。在多元统计分析中,用于预报控制的模型有两大类:一类是预测预报模型,通常采用多元线性回归或逐步回归分析、判别分析、双重筛选逐步回归分析等建模技术;另一类是描述性模型,通常采用聚类分析的建模技术。随着大数据时代的到来,面对海量数据,传统数据分析技术已不适应时代的要求,因此在本版中提供了 Lasso 回归分析技术,并对该功能进行了改进,使之可用于含分类变量的重要因子筛选及数据不完整(有缺失)的情形。

- 进行数值分类,构造分类模式。在多变量系统的分析中,往往需要将系统性质相似的事物或现象归为一类,以便找出它们之间的联系和内在规律。过去许多研究多是按单因素进行定性处理,以致处理结果反映不出系统的总的特征。进行数值分类,构造分类模式一般采用聚类分析和判别分析技术。为解决著者在科研中遇到的实际问题,本版书介绍了我们提出的一个新的统计分析模型,即在有序约束条件下的多峰识别模型(modeling peaks under order restrictions)。

如何选择适当的方法来解决实际问题,需要对问题进行综合考虑。对一个问题可以综合运用多种统计方法进行分析。例如,一个预报模型的建立,可先根据有关生物学、生态学原理,确定理论模型和实验设计。根据实验结果,收集实验资料,

对资料进行初步提炼;然后应用统计分析方法(如相关分析、逐步回归分析、Lasso 回归分析、偏最小二乘回归分析、主成分分析等)研究各个变量之间的相关性,选择最佳的变量子集合,在此基础上构造预报模型,最后对模型进行诊断和优化处理,并应用于生产实际。

近年来全球性的神经网络研究热潮的再度兴起,主要的原因在于发展新型计算机和人工智能新途径的迫切需要。基于神经科学研究成果基础上发展出来的人工神经网络模型,反映了人脑功能的若干基本特性,开拓了神经网络用于计算机的新途径。在这众多神经网络模型中,应用最广泛的是多层感知机神经网络。它的研究始于 20 世纪 50 年代,但一直进展不大,直到 1985 年,Rumelhart 等提出了误差反向传递学习算法(即 BP 算法),实现了 Minsky 的多层网络设想。在第五版中,我们对 BP 神经网络算法进行改进,以适应有数据缺失情况下建立非线性优化模型需要。

多元统计分析虽然较为复杂,但 DPS 平台将引导读者走向成功!

第 14 章 多变量统计检验

多变量计量资料(multivariate data)是指有多个反应变量(观测指标)同时受处理因素(分组变量)影响的测量资料。科学实验处理对象的观察指标通常有多个,如除草剂药效试验中,一种除草剂防治对象有几种杂草,要想同时评价除草剂对几种杂草的防治效果,就需要应用多元统计检验技术。同样,在作物栽培试验中,某项增产措施对作物的多个产量性状的影响也是多指标的。这类评价多指标的统计分析即为多元方差分析技术。

对于反应变量服从正态分布的资料,只考察一个反应变量时,常用 t 检验和一元方差分析。考察多个反应变量时,虽然也可以用 t 检验和一元方差分析,但将 p (≥ 2)个反应变量分开,进行 m 次单变量分析,这样做:①不能全面反映多个反应变量在多维空间上的信息;②当不同观测指标的检验结果不一致时,得不出一个概括性的结论;③不能处理多个观测指标间的相互关系。因此对多元资料宜用多元统计分析法。

14.1 多元计量资料的常用统计量

单个反应变量资料常用平均值±标准差来描述。多元计量资料的所有统计描述与统计推断都采用随机向量或多个随机向量放在一起组成的随机矩阵进行。常用描述多元计量资料集中性和离散性的统计量是均数向量、协方差矩阵和相关矩阵。

14.1.1 样本均数向量

为分析不同时期儿童的身体发育情况,随机检查 5 名 8 岁男孩的三项身体发育指标,资料见表 14-1。

表 14-1 5 名 8 岁男孩的 3 项测量值

儿童编号	身高(x_1/cm)	体重(x_2/kg)	胸围(x_3/cm)
1	119.8	22.6	60.5
2	121.7	21.5	55.5
3	129.5	24.5	51.0
4	133.9	27.2	65.8
5	126.5	25.0	63.9
$\bar{x}$	126.28	24.16	59.34
s^2	32.86	4.89	37.04

这时每个个体有 3 个观测指标 x_1, x_2 和 x_3 的资料, 第 j ($j=1, 2, 3$) 个个体的测量值可用一个 3 维列向量表示为 $X_j = (x_{1j}, x_{2j}, x_{3j})^T$ 。例如, 第 1 个男孩测量值可用一个 3 维列向量 $X_1 = (x_{11}, x_{21}, x_{31})^T = (119.8, 22.6, 60.5)^T$ 表示。5 名男孩身高均数 $\bar{x}_1 = 126.28$, 体重均数 $\bar{x}_2 = 24.16$, 胸围均数 $\bar{x}_3 = 59.34$, 将 3 个观测指标各自的样本均数合起来, 便构成了一个 3 维样本均数向量(mean vector):

$$\bar{X} = (\bar{x}_1, \bar{x}_2, \bar{x}_3)^T = (126.28, 24.16, 59.34)^T$$

一般地, 如果 n 个观察对象, 每个观察对象有 p 个观测指标 $x_1, x_2, \dots, x_p$ 时, 第 j ($j=1, 2, \dots, n$) 个观察对象的观察值为 $x_{1j}, x_{2j}, \dots, x_{pj}$, 可表示为一个 p 维列向量:

$$X_j = (x_{1j}, x_{2j}, \dots, x_{pj})^T$$

则 p 个观测指标各自的样本均数可用一个 p 维列向量表示为样本均数向量 $\bar{X}$:

$$\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T$$

14.1.2 样本协方差矩阵

一般情况下, 如果有 p 个观测指标 $x_1, x_2, \dots, x_p$, 则样本方差-协方差矩阵 (variance-covariance matrix, 简称协方差矩阵) 是一个 $p \times p$ 矩阵, 记为

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

式中对角线上是各变量的方差:

$$s_{ii} = \sum (x_{ij} - \bar{x}_i)^2 / (n-1), \quad i=1, 2, \dots, p$$

对角线两侧是变量与变量之间的协方差:

$$s_{ik} = s_{ki} = \sum (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) / (n-1) \\ i \neq k; i=1, 2, \dots, p; k=1, 2, \dots, p$$

14.1.3 样本相关矩阵

p 个观测指标 $x_1, x_2, \dots, x_p$ 的所有相关系数合起来, 写成如下所示矩阵, 便得到一个 $p \times p$ 的样本相关矩阵 R :

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

因 $s_{ik} = s_{ki}$, $r_{ik} = r_{ki}$, 所以 S 和 R 都是对称矩阵。

在 DPS 中, 以一行一个样本, 一列一个变量的方式在电子工作表中编辑数据, 然后用鼠标选中, 再点击“多元分析”→“相关分析”菜单下面的“两变量相关分析”或“多变量相关分析”菜单, 即可得到样本均值向量和样本相关系数矩阵, 如点击快捷按钮“S”即可得到样本协方差矩阵。

14.2 多元均值检验

14.2.1 一组样本和总体的检验

当我们从多元总体中抽取 1 组样本, 希望检验该样本是否和总体没有差异时, 可用多元分析中的 T^2 检验:

$$T^2 = n(n-1)(\bar{x} - \mu)^T A^{-1}(\bar{x} - \mu)$$

式中 n 为样本容量, A 为样本协方差阵。同时 T^2 检验可以转换为等价的 F 检验:

$$F = \frac{(n-p)}{(n-1)p} T^2$$

其自由度为 $(p, n-p)$, 从而我们可利用 F 统计量进行检验。

例如, 若一个实验有 n 个变量, m 个观察样本的检验, 分析数据按图 14-1 (第一行放入总体的各个指标, 以下逐个放入各个样本观察值) 所示方式编辑。本例是检验某品种成牛的体长、胸围是否来自总体 (145, 190) 的样本, 现从群体中随机抽取 20 头牛, 每头牛的测量数据按图 14-1 所示方式编辑。

	A	B	C	D
1		体长	胸围	
2	总体均值	145	190	
3	样本 1	151.5	186	
4	样本 2	156.2	186	
5	样本 3	146	193	
6	样本 4	138.1	193	
7	样本 5	146.2	172	
8	样本 6	149.8	188	
9	样本 7	155	187	
10	样本 8	144.5	175	
11	样本 9	147.2	175	
12				

图 14-1 一组样本和总体比较用户界面

然后执行“多元分析”→“多元统计检验”→“样本与总体均值差异检验”功能, 得到分析结果: Willks 统计量为 0.06297, F 统计量等于 11.3341, 其显著水平等于 0.00065, 小于 0.01 的极显著水平。因此推断样本的均值与总体不相等。

14.2.2 两组样本均值比较

如果我们从两总体中分别抽取样本容量为 n 和 m 的样本, 检验两样本是否相等。同样可以采用 T^2 检验方法进行检验,

$$T^2 = \frac{nm}{n+m} (\bar{x} - \bar{y}) S^{-1} (\bar{x} - \bar{y})$$

式中 $S = (A_x + A_y) / (n + m - 2)$ 。同样 T^2 统计量可以转换为 F 统计量,

$$F = \frac{n+m-p-1}{(n+m-2)p} T^2$$

其自由度为 $(p, n+m-p-1)$, 从而我们可利用 F 统计量进行检验。

例如, 若一个实验有 2 个变量, 每个处理有 4 个观察样本, 分析数据按图 14-2 所示方式编辑, 经计算得到如图 14-2 所示结果。

然后执行“多元分析”→“多元统计检验”→“两组样本均值检验”功能, 经计算得到计算结果, $T^2 = 116.7081$, F 统计量等于 48.63, 其显著水平等于 0.00053, 大于 0.01 的极显著水平。因此推断两组样本均值不相等。

	A	B	C	D	E
1		x1	x2		
2	1	131.5	9	第一组样本	
3	1	145	12		
4	1	141	30		
5	1	150	36		
6	2	40.5	54	第二组样本	
7	2	80	74.5		
8	2	50	64.5		
9	2	90	60.5		
10					

图 14-2 两样本均值比较界面

14.2.3 成对试验的统计检验

在两组样本的比较分析中, 有时两总体的样品必须成对出现才能作比较。这时统计检验的 T^2 统计量的构成为: 令 $\bar{d} = \bar{x} - \bar{y}$, $A_d = \sum_{i=1}^n (d_i - \bar{d})(d_i - \bar{d})^T$, 则

$$T^2 = n(n-1) \bar{d}^T A_d^{-1} \bar{d}$$

同样, 可以将 T^2 统计量转换为 F 统计量进行统计检验。

例如, 在进行配对检验之前, 需先在电子表格中计算出样品中各个变量两两配对时的差值 (d_i) , 然后将差值定义成数据块进行分析(图 14-3)。

	A	B	C	D	E	F	G
1	原始数据				用于统计分析数据		
2	喷矮壮素		对照		株高差异	产量差异	
3	株高	产量	株高	产量			
4	160	28	170	25	-10	3	
5	160	30	270	28	-110	2	
6	200	31	220	29	-20	2	
7	160	35	250	29	-90	6	
8	200	30	270	27	-70	3	
9	170	34	290	25	-120	9	
10	150	30	270	28	-120	2	
11	210	28	230	21	-20	7	
12							

图 14-3 成对试验的统计检验界面

然后执行“多元分析”→“多元统计检验”→“配对试验差异检验”功能,得到分析结果:差异均值: $x_1=-70;x_2=4.25$; F 统计量等于 14.075,其显著水平等于 0.0054,大于 0.01 的极显著水平。因此推断两组样本均值不相等。

14.3 多元方差分析简介

14.3.1 多元方差分析基本原理

前面介绍的是两样本的统计检验,但在实际工作中,我们常遇到多组样本的统计检验问题。在多元方差分析中,我们观察的如果不是一个指标(性状),而是 m 个指标,记为向量形式 $X=(x_1,x_2,\cdots,x_m)^T, m<n$,那么第 i 水平的第 j 次重复观察值可以表示为 $X_{ij}=(x_{ij1},x_{ij2},\cdots,x_{ijm})^T, i=1,2,\cdots,k;j=1,2,\cdots,n$ 。全实验共有 knm 个数据,并假定 X_{ij} 满足线性模型:

$$\begin{aligned} X_{ij} &= m + e_{ij} \\ e_{ij} &\sim N_m(0, S) \\ \sum_i a_i &= 0, \quad i=1,\cdots,k; j=1,\cdots,n \end{aligned}$$

统计假设为 $H_0: \mu_1=\mu_2=\cdots=\mu_k$ 或 $\alpha_1=\alpha_2=\cdots=\alpha_k=0$ 。为了构造 H_0 的检验量,采用类似于一元方差分析的做法,把反映全实验变异的总离差阵按变异来源分解成水平间离差阵 H 与水平内离差阵 E 之和,即

$$\begin{aligned} H &= n \sum_{i=1}^k (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T \\ E &= \sum_i \sum_j^n (X_u - \bar{X}_i)(X_u - \bar{X}_i)^T \end{aligned}$$

$$W = \sum_i \sum_j (X_u - \bar{X})(\bar{X}_{ij} - \bar{X})^T$$

并有 $W = H + E$, 当 H_0 成立时, W 服从于自由度为 $\text{df}_w = kn - 1$ 的 $W_m(kn - 1, \Sigma)$ 分布; E 是 n 个独立样本离差阵之和, 服从于自由度为 $\text{df}_E = kn - k$ 的 $W_m(kn - k, \Sigma)$ 分布; H 是 k 个独立均值向量的离差阵, 其自由度 $\text{df}_H = k - 1$ 。多组均值向量比较检验的 Wilks 统计量 Λ 为

$$\Lambda = \frac{E}{|H + E|}$$

该统计量 Λ 由 Wilks 首先提出, Wilks 统计量 Λ 分布较复杂, 在变量不多或相比较的资料组数不多时, 可以导出统计量的精确分布(表 14-2)。

表 14-2 Wilks 统计量 Λ 的精确分布

变量个数(m)	资料组数(g)	Λ 统计量分布
1	≥ 2	$\left(\frac{n-g}{g-1}\right) \left(\frac{1-\Lambda}{\Lambda}\right) - F_{g-1, n-g}$
2	≥ 2	$\left(\frac{n-g-1}{g-1}\right) \left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right) - F_{2(g-1), 2(n-g-1)}$
≥ 1	2	$\left(\frac{n-m-1}{m}\right) \left(\frac{1-\Lambda}{\Lambda}\right) - F_{m, n-m-1}$
≥ 1	3	$\left(\frac{n-m-2}{m}\right) \left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right) - F_{2m, 2(n-m-2)}$

当 m 和 g 超出上述范围时, Rao 提出了 Λ 统计量的近似分布

$$F = \frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \frac{v'_2}{v'_1} - F(v'_1, v'_2)$$

式中, $s = \sqrt{\frac{m^2 v_T^2 - 4}{m^2 + v_T^2 - 5}}$, $v'_1 = m v_T$, $v'_2 = s \left(v'_1 + v_E - \frac{m + v_T + 1}{2} \right) - \frac{m v_T - 2}{2}$ 。

这里 v_T 是处理的自由度, v_E 是误差自由度。

14.3.2 多元方差分析中的多重比较

1. Roy 氏最大 T^2 检验(Roy's $T_{\max}^2$)

在多元方差分析中, 如果 $\Lambda < \Lambda_{\alpha(m, kn-k, k-1)}$, 仅仅表现诸 μ_i 不全相等, 不能排除其中部分总体均值向量相等的情形。因此, 有必要检验两个水平的差异, 即检验 $H_{0ij}: \mu_1 = \mu_j (i < j, i, j = 1, 2, \dots, k)$ 。为此将 k 个样本均值向量 $X_1, X_2, \dots, X_k$ 两两配对, Roy 氏最大 T^2 检验是分别计算它们的 T^2 :

$$T_{\max(ij)}^2 = (kn - k) \frac{n_i n_j}{n_i + n_j} (\bar{X}_i - \bar{X}_j) E^{-1} (\bar{X}_i - \bar{X}_j)$$

服从于 $T^2_{(m, kn-k)}$, 或转换成等价的 F 统计量:

$$F = \frac{n-g-m+1}{(n-g)m} T^2_{\max}$$

如果 $F_{ij} > F_{\alpha(m, kn-k-m+1)}$, 在显著水平 α 下否定 H_{0ij} 。不难看出, T^2 法相当于一元统计检验中的最小显著差数法(LSD)。

2. 单一自由度独立比较

有些实验在设计时为达到研究目的就预先安排若干特定的比较。这时我们可用单一自由度独立比较方法进行分析, 又称为正交比较, 因为该方法具正交性(orthogonal comparison), 即在所有对比(比较)中, 两两比较的比较系数(C_i)乘积之和都为 0, 这种比较系数称为正交系数(orthogonal coefficient)。

正交分析比较中, 正确进行处理合并比较的关键是正确确定比较的内容和正确写出比较的正交系数。为此, 需满足下列 3 个条件:

(1) 比较的数目必须为 $k-1$, 以使每一比较占有而且仅占有 1 个自由度。

(2) 每一独立比较的正交系数之和必须为 0, 即 $\sum C_i = 0$, 以使每一比较都是均衡的。

(3) 任何两个独立比较的相应正交系数乘积之和必须为 0, 即 $\sum C_i C_j = 0$, 以保证 SS_t 恰好分解为 $k-1$ 个 SS_Q , 如果出现 $\sum C_i C_j \neq 0$, 则比较就不再是独立的, 其后果是各个比较的 SS_Q 之和必不等于 SS_t 。这一点在实验设计时需特别注意。如有各个处理数(n)相等的 3 个处理, 其总和数分别为 T_A, T_B, T_C , 则

$$\begin{array}{ccc} T_A & T_B & T_C \\ 2 & -1 & -1 \\ 0 & 1 & -1 \end{array} \quad \text{或} \quad \begin{array}{ccc} T_A & T_B & T_C \\ 1 & 1 & -2 \\ -1 & 1 & 0 \end{array}$$

等皆为两个独立的比较, 而

$$\begin{array}{ccc} T_A & T_B & T_C \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{array} \quad \text{或} \quad \begin{array}{ccc} T_A & T_B & T_C \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{array}$$

等则为不独立的, 因为前者的 $\sum C_i C_j = 0$, 后者的 $\sum C_i C_j \neq 0$ 。

在具体写出各个独立比较的正交系数时, 可按下列规则进行:

(1) 若被比较的两个组的处理数目相等, 则给一个组的各处理以系数+1, 另一个组的各处理都是系数-1。到底哪个组取+1, 哪个组取-1, 可以随便, 一般以“在前”的组(如施肥对不施肥的比较, 施肥在前)取“+”号。

(2) 若被比较的两个组的处理数目不相等,则第一组的系数为第二组的处理数,第二组的系数为第一组的处理数,如某个比较项目是前面的 2 个处理和后面的 3 个处理相比较,其 C_i 写作 $3, 3, -2, -2, -2$ 。

(3) 如果写出的 C_i 有公约数,则应将其约为最小的整数,如 4 个处理与 2 个处理的一个比较,按规则(2)其 C_i 为 $2, 2, 2, 2, -4, -4$,应简化成 $1, 1, 1, 1, -2, -2$ 。

(4) 如果某一处理已经和所有其余处理作过一次比较,则该处理不能再参加其余比较,否则一定破坏了 $\sum C_i C_j = 0$ 。

14.3.3 多元方差分析应用要点

(1) Wilks 统计量 Λ , F 统计量及其显著水平。如显著水平 p 小于等于 0.05,则表明各个处理之间的综合效应有显著的差异;如显著水平 p 大于 0.05,则表明各个处理之间的综合效应没有显著差异。一般来说,只要有一个变量有差异,则多变量综合效应就有差异。

除综合效应分析外,系统还对各个指标进行了单因素方差分析,看哪些因素的各个处理间的差异不显著,这时可以剔除某个 p 值最大的指标,再进一步作多元方差分析。

(2) 各个处理间两两差异的 Roy 氏最大 T^2 检验,其概率 p 值是根据各个处理间差异检验的 F 值计算出的,并标出了其显著水平(1 个星号表示 $\alpha=0.05$;2 个星号表示 $\alpha=0.01$)。

14.4 单向完全随机设计

例 1 为了研究某疾病,对一批人同时测了 4 个指标: x_1 (β 脂蛋白), x_2 (甘油三酯), x_3 (α 脂蛋白), x_4 (前 β 脂蛋白),并针对不同性别、不同年龄,将这批人分成 3 组测量:第一组是 20~35 岁女性,第二组是 20~25 岁男性,第三组是 35~50 岁男性。现欲检验 3 组之间有没有显著差异,在 DPS 数据处理平台上,需按图 14-4 所示方式编辑、定义数据块。

然后执行“多元分析”→“多元统计检验”→“单因素完全随机设计”功能,在分析时系统会提示用户输入处理“组数或各组的样本数”,这时可以输入 1 个数字(处理的组数)或输入多个数字(2 个以上,各个数字用空格隔开),然后系统输出计算结果。