

DPS[©]数据处理系统

第三卷 专业统计及其他

第三版

DPS[©] Data Processing System

Volume III

Specialized Statistics and Miscellanea

Third edition

Tang Qiyi

唐启义 著

科学出版社

北京

内 容 简 介

本书从应用角度简要地阐述了现代统计学 500 多种实验数据统计分析和模型模拟方法,如包括混料实验设计在内的各种实验设计、方差分析、列联表分析及非参数检验;专业统计包括量表分析、数据包络分析、随机前沿面分析、顾客满意指数模型、试验诊断 ROC 曲线、生物测定、生存分析、动植物遗传育种、品种区域试验、空间分布型、地理统计、数值生态学方法等;各种多元分析技术;非线性回归模型参数估计、模型模拟技术;单目标和多目标线性规划、非线性规划等运筹学方法;以及状态方程、数值分析、时间序列分析、模糊数学、**随机森林**、BP 神经网络、径向基函数(RBF)、支持向量机(SVM)、层次分析、小波分析、灰色理论等方法。

本书是自然科学、社会科学各个专业科研、教学、管理和技术推广人员案头必备工具书。全书分第一卷(基础统计和实验设计),第二卷(现代统计与数据挖掘),以及第三卷(专业统计及其他)共 3 卷出版,并配以作者自主开发的具有自主知识产权的 DPS 统计软件系统试用版光盘。第一卷可作为大学本科和研究生、一般科研工作者掌握常用统计技术的教材和实验指导书;第二卷可供科研人员、博士、硕士研究从事科学试验时数据分析使用;第三卷适用于各领域科研人员解决本专业试验统计和数据分析问题。

图书在版编目(CIP)数据

DPS 数据处理系统:实验设计、统计分析及数据挖掘/唐启义著. —3 版,
—北京:科学出版社,2014
ISBN 978-7-03-026443-5

I. ①D… II. ①唐… III. ①计算机应用—统计分析(数学) IV. ①
0212.1-39

中国版本图书馆 CIP 数据核字(2010)第 010039 号

责任编辑:赵彦超 鄢德平/责任校对:李奕莹
责任印制:安春生/封面设计:王浩

科学出版社出版

北京东黄城根北街 16 号

邮政编码:100717

<http://www.sciencecp.com>

双青印刷厂印刷

科学出版社编务公司排版制作

科学出版社发行 各地新华书店经销

*

2007 年 1 月第 一 版 开本:787×1092 1/16

2014 年 1 月第 三 版 印张:73

2014 年 1 月第一次印刷 字数:650 000

印数:1—2 000

定价:99.80 元(含光盘)

如有印装质量问题,我社负责调换

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。早在文革期间，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汇编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，绝大多数团队均自然衰退。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、试验的设计和建模、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创造、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”的精神创立和发展起来的。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力最大的技术。从 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，影响领域也从开始的军事领域拓及到经济社会，以及其他各个领域。1974 年，Intel 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。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。进入新世纪以来，计算机发展更为迅猛，其与通讯技术等相互渗透和融合，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，使得统计分析方法并没有在实践中有效推广开来。而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实现自己的理论，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中。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 比较受到生物医药和社会科学领域人员的喜爱，而主流的应用统计人员则更倾向于 SAS, SPlus 则

更适合于进行统计研究的科研工作者。其中 SAS 更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, 由于这些软件比较昂贵, 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 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试验设计、混料试验设计、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, 要几年功夫?’”

澄观道: ‘这在般若堂的典籍中是有得记载的。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。’”

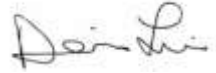
统计学之深奥, 多少人穷极一生都无法参透, 吾等凡夫俗子又如何能在有生之年得到真谛呢? 《DPS[®]数据处理系统——实验设计、统计分析及数据挖掘》让我们看到了一线曙光。该书内容涵盖之广, 仿佛一本统计百科全书。全书 8 篇 44 章, 包山包海, 几乎包括了所有目前统计应用的主要方向。全书由浅入深, 精辟易懂, 配合软件操作, 让一门

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

第三版前言

《DPS[®]数据处理系统——实验设计、统计分析及数据挖掘》(第3版)是作者基于自主开发的 DPS 统计软件,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,并将整个文稿分成3卷出版:即第一卷为基础统计;第二卷为现代统计;第三卷为专业统计及其他。

第一卷包括了第2版的第一篇和第二篇,和原来“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)。第6章系统介绍了相关和回归分析原理、应用,及应用中须注意的一些问题。另外在第1章里面增加了“工作表中图形输出”、“DPS 系统应用常见问题解答”两个章节,在第2章里面增加了“缺失值处理”,在第3章里面增加了“Poisson 分布抽样下多样本检验”。在第11章(多因素优化设计与分析)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2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2版第八篇中的第43章(神经网络和支持向量机)章节。

在第二卷中新增内容主要是第19章“结构方程模型”,介绍了在 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”章节。并以第2版43章(神经网络和支持向量机)为主体,增加了“随机森林”章节,将其命名为“数据挖掘”一章。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,作者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。

第三卷主体内容为上一版的第三篇(专业试验统计)、第六篇(常用数值分析)和第八篇(其他数据分析方法)。

在专业试验统计中,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。

在广大读者和 DPS 统计软件用户支持下、经过作者这三年多的努力, DPS 系统功能又得到了进一步的增强。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;研究生学生宛庭利、刘顺、罗巍,湖南农业大学实习生徐西林协助第三版文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

第二版前言

《DPS[®]数据处理系统——实验设计、统计分析及数据挖掘》(第二版)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“混料试验设计与分析”。在第一版的均匀试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、较目前已有的混料试验设计方法更有效的混料试验设计的新技术,该技术能适合各种情形的混料设计,特别是组分多、试验次数大、限制条件“苛刻”的情形。同时提供了混料试验数据分析和优化的数学模型,为工业、农业和科学试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“序贯试验分析”。在进行下一次试验以前,已经知道前面试验的结果,这种一次接着一次的抽样试验为序贯试验。本章提供了在序贯试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、多因素优化与分析、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、多因素分析、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,如试验设计和方差分析中的 Box-Behnken 设计、平衡不完全区组设计试验方差分析、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害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

第一版前言

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)，只要存在着带有随机性的数据，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 等以农业科学试验为对象，探讨试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，Engedow 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试验方法的基础上，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像的。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，试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试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 **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 **Window 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试验统计的完整性和系统性, 提供了比较全面的试验设计和试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、最优回归试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、试验评价的 **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试验的统计分析技术, 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试验中, 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)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素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。读者可以借助于 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、市场潜力预测、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、品种区域试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待清楚是不太可能的，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，竞争十分剧烈，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，及时解答用户在实际应用中的技术问题，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	I
序 二	III
第三版前言	VI
第二版前言	VIII
第一版前言	811
第六篇 常用数值分析	825
第 29 章 矩阵计算	827
29.1 矩阵转置	827
29.2 矩阵基本运算	828
29.2.1 矩阵加	828
29.2.2 矩阵减	829
29.2.3 矩阵相乘	829
29.3 矩阵自乘	830
29.4 矩阵样本方差	831
29.5 矩阵总体方差	832
29.6 解正规方程组	833
29.7 矩阵求逆	833
29.7.1 矩阵求逆及行列式值	833
29.7.2 矩阵广义逆	834
29.8 奇异值分解	835
29.9 实对称矩阵特征值和特征向量	836
29.9.1 方法简介	836
29.9.2 DPS 平台的操作示例	837
29.10 实矩阵特征值和特征向量	838
29.10.1 方法简介	838
29.10.2 DPS 平台的操作示例	841
29.11 应用矩阵运算组建多元线性回归模型	842
参 考 文 献	844
第 30 章 方程求解及多项式求根	846
30.1 求解线性方程组	846

30.1.1	方法简介.....	846
30.1.2	DPS 平台的操作示例.....	846
30.2	非线性方程组求解.....	847
30.3	实系数多项式求根.....	849
30.3.1	方法简介.....	849
30.3.2	方程求根的计算机处理.....	849
参 考 文 献	851
第 31 章	微积分数值计算.....	852
31.1	定 积 分.....	852
31.1.1	定积分基本原理和方法.....	852
31.1.2	DPS 平台的操作示例.....	853
31.2	多 重 积 分.....	853
31.2.1	方法简介.....	853
31.2.2	DPS 平台的操作示例.....	854
31.3	数 值 微 分.....	855
31.3.1	方法简介.....	855
31.3.2	DPS 平台的操作示例.....	856
31.4	微分方程(组)初值求解.....	856
31.4.1	微分方程模型的参数求解原理及步骤.....	857
31.4.2	DPS 平台操作示例.....	858
参 考 文 献	859
第七篇	专业试验统计.....	860
第 32 章	数据包络分析和随机前沿面分析.....	862
32.1	生产效率分析基本原理.....	863
32.2	数据包络分析 CCR 模型.....	866
32.2.1	CCR 模型.....	866
32.2.2	松弛变量.....	868
32.2.3	面向产出的 DEA 模型.....	869
32.2.4	超效率分析.....	870
32.2.5	CCR 模型应用示例.....	871
32.3	数据包络分析扩展模型.....	873
32.3.1	可变规模收益(VRS)模式下的 BCC 模型.....	873
32.3.2	可变规模收益(VRS)模式下的 IRS 模型和 DRS 模型.....	877
32.3.3	基于松弛测度的 SBM 模型.....	878
32.3.4	DEA 分析非任意控制变量模型.....	879
32.4	考虑价格因素时的 DEA 模型.....	880

32.4.1	成本最小化模型.....	880
32.4.2	收入最大化模型.....	881
32.4.3	利润最大化模型.....	881
32.4.4	考虑价格因素时 DEA 模型操作示例.....	882
32.5	面板数据的 MALMQUIST 指数.....	883
32.5.1	Malmquist 指数简介.....	883
32.5.2	Malmquist 指数计算示例.....	887
32.6	随机前沿面模型.....	888
32.6.1	随机前沿生产函数.....	889
32.6.2	极大似然估计.....	890
32.6.3	平均技术效率的估计.....	891
32.6.4	单位级技术效率的预测.....	891
32.6.5	假设检验.....	892
32.6.6	随机前沿面模型分析示例.....	893
32.7	面板数据随机前沿面模型.....	895
32.7.1	面板数据随机前沿面模型.....	895
32.7.2	截断正态分布.....	897
32.7.3	假设检验.....	898
32.7.4	成本函数.....	898
32.7.5	面板数据随机前沿分析中效率预测.....	898
32.7.6	随机前沿面模型分析示例.....	899
	参 考 文 献.....	901
第 33 章	量表分析和顾客满意指数模型.....	903
33.1	项目分析.....	903
33.1.1	测验项目的难度.....	903
33.1.2	测验项目的区分度.....	903
33.1.3	在 DPS 系统中进行项目分析(item analysis).....	904
33.2	量表可信度分析.....	906
33.2.1	信度指标.....	906
33.2.2	分半信度、Cronbach α 系数及荷伊特信度分析.....	908
33.2.3	库德-理查逊信度.....	910
33.3	顾客满意指数模型.....	911
33.3.1	顾客满意指数模型分析原理.....	911
33.3.2	DPS 处理应用.....	914
33.4	结 合 分 析.....	919
33.4.1	确定要研究的属性和属性水平.....	919
33.4.2	实验设计.....	920
33.4.3	数据分析.....	920

33.4.4	模型评价.....	920
33.4.5	应用示例.....	921
参 考 文 献.....		924
第 34 章	生物测定与药物动力学.....	926
34.1	质反应生测机率值分析.....	927
34.1.1	基本原理.....	927
34.1.2	数据整理与分析.....	928
34.1.3	结果分析解释.....	932
34.2	数量反应生测机率值分析.....	932
34.2.1	概述.....	932
34.2.2	数据整理与分析.....	932
34.2.3	分析结果解释.....	934
34.3	混剂互作毒力系数统计分析.....	934
34.3.1	增效性的判断(S. Sakai 公式).....	934
34.3.2	增效倍数.....	934
34.3.3	共毒系数(Co-toxicity Coefficient).....	935
34.3.4	Finney 调和平均数法.....	935
34.3.5	Chou-Talalay 结合指数 CI(combination index) 法.....	936
34.3.6	制剂混合毒力系数的统计检验.....	936
34.3.7	混合制剂毒力统计的实验设计.....	938
34.3.8	混合制剂毒力系数和 Finney 调和平均数的关系.....	938
34.3.9	混剂结合毒力统计的数据整理与分析.....	939
34.4	时间-剂量-死亡率模型分析.....	944
34.4.1	基本原理简介.....	944
34.4.2	数据整理与操作.....	946
34.5	药物代谢动力学(房室模型).....	948
34.5.1	基本原理简介.....	948
34.5.2	房室模型分析示例.....	950
34.6	药物代谢动力学(非房室参数估计).....	951
34.6.1	非房室统计矩方法简介.....	951
34.6.2	非房室统计矩参数的变异估计.....	953
34.6.3	DPS 应用实例.....	953
参 考 文 献.....		955
第 35 章	诊断试验评价.....	957
35.1	诊断试验常用指标.....	957
35.1.1	一致百分率.....	958
35.1.2	灵敏度.....	958

35.1.3	特异度.....	959
35.1.4	Youden 指数.....	959
35.1.5	阳性似然比.....	960
35.1.6	阴性似然比.....	960
35.1.7	预报值.....	960
35.1.8	阴性预报值.....	961
35.2	有序分类资料 ROC 曲线.....	962
35.2.1	实验资料 ROC 工作点计算及 ROC 曲线构建.....	962
35.2.2	ROC 曲线下面积的计算原理.....	963
35.2.3	单个有序分类样本 ROC 曲线估计.....	966
35.2.4	多组有序分类样本 ROC 曲线估计及其比较.....	967
35.3	定量数据 ROC 曲线分析.....	968
35.4	汇总多个样本的 SROC 曲线分析.....	970
	参 考 文 献.....	975
第 36 章	序贯试验分析.....	977
36.1	开放型序贯试验.....	978
36.1.1	开放型质反应单向序贯试验.....	978
36.1.2	开放型数量资料单向序贯试验.....	980
36.1.3	开放型质反应双向序贯试验.....	981
36.1.4	开放型量反应双向序贯试验.....	982
36.2	闭锁型序贯试验.....	984
36.2.1	闭锁型双向质反应序贯试验(翼形设计序贯试验).....	985
36.2.2	闭锁型双向数量反应序贯试验.....	986
36.3	成组序贯试验.....	987
36.3.1	成组序贯试验的意义.....	987
36.3.2	成组序贯试验各时段每组样本数 n 的估计.....	988
36.3.3	总体未知时两均数比较.....	989
36.3.4	两样本率比较的序贯试验.....	991
	参 考 文 献.....	993
第 37 章	生存分析.....	994
37.1	生存率估计.....	995
37.1.1	小样本资料生存率的估计.....	995
37.1.2	大样本分组资料生存分析.....	995
37.2	两样本生存率 LOG-RANK 检验.....	998
37.3	寿命表的编制与分析.....	1001
37.3.1	现时寿命表.....	1001
37.3.2	现时寿命表中平均寿命的估计.....	1004

37.3.3 去死因寿命表.....	1007
37.3.4 健康预期寿命表的编制.....	1009
37.4 比例风险模型——COX 回归.....	1011
37.5 指数模型.....	1015
37.6 WEIBULL 模型.....	1018
附 人口金字塔图的绘制.....	1020
参 考 文 献.....	1020
第 38 章 数学生态学方法.....	1021
38.1 种群空间分布型聚集度指标测定.....	1021
38.1.1 聚集度测定的指标.....	1021
38.1.2 试验数据资料整理.....	1023
38.2 种群空间分布型-频次分布检验.....	1024
38.2.1 方法简介.....	1024
38.2.2 试验数据资料分析整理.....	1027
38.3 负二项分布公共 κ 值估计.....	1030
38.3.1 估计方法.....	1030
38.3.2 试验数据资料整理及应用.....	1030
38.4 二元变量距离系数.....	1031
38.4.1 方法简介.....	1031
38.4.2 DPS 平台操作示例.....	1033
38.5 距离系数计算.....	1034
38.5.1 方法简介.....	1034
38.5.2 DPS 平台的操作示例.....	1036
38.6 极点排序.....	1037
38.6.1 方法简介.....	1037
38.6.2 DPS 平台的操作示例.....	1038
38.7 物种丰富度估计.....	1038
38.7.1 方法简介.....	1038
38.7.2 DPS 平台的处理方法.....	1039
38.8 对数序列参数估计.....	1040
38.8.1 方法简介.....	1040
38.8.2 DPS 处理平台的操作示例.....	1041
38.9 对数正态分布模型参数估计.....	1041
38.9.1 方法简介.....	1041
38.9.2 DPS 处理平台操作方法.....	1042
38.10 群落多样性指数.....	1042
38.10.1 方法简介.....	1042
38.10.2 DPS 平台的操作示例.....	1044

38.11 生态位宽度指数	1046
38.11.1 方法简介	1046
38.11.2 DPS 平台的操作示例	1047
38.12 生态位重叠指数	1048
38.12.1 方法简介	1048
38.12.2 DPS 平台的操作示例	1049
参 考 文 献	1050
第 39 章 地 理 统 计	1053
39.1 空间自相关分析	1053
39.1.1 Moran 的 I 指数	1053
39.1.2 Geary 的 C 指数	1054
39.1.3 DPS 中空间自相关分析过程	1055
39.2 空间联系统计分析	1057
39.3 局部空间相关分析	1060
39.3.1 局部 Moran	1060
39.3.2 局部 Geary 统计量	1061
39.3.3 局部空间统计量, 局部 Moran 指标和局部 Geary 指标的计算	1062
39.4 实验半变异函数	1062
39.5 协方差函数及相关系数	1065
39.6 变异函数理论模型的最优拟合	1066
39.6.1 变异函数理论模型简介	1066
39.6.2 变异函数理论模型参数拟合	1068
39.7 交叉验证	1069
39.8 克立格插值	1072
参 考 文 献	1073
第 40 章 品 种 比 较 试 验	1074
40.1 一年多点试验稳定性分析	1074
40.1.1 Eberhart 和 Russell 模型	1074
40.1.2 George C. C. Tai 模型	1076
40.1.3 Shukla 模型	1078
40.2 一年多点区域试验的统计分析	1082
40.2.1 方法简介	1082
40.2.2 DPS 平台的操作示例	1083
40.3 多年多点品种区域试验的统计分析	1085
40.3.1 方法简介	1085
40.3.2 DPS 平台的资料整理与分析步骤	1086
40.4 品种区域试验 AMMI 模型分析	1088

40.4.1	方法简介.....	1088
40.4.2	DPS 平台的数据格式及其编辑.....	1090
40.4.3	应用示例.....	1091
40.4.4	无区组品比试验的 AMMI 模型分析.....	1094
40.5	SHMM 模型.....	1095
40.6	增广随机区组设计试验.....	1097
40.6.1	方法简介.....	1097
40.6.2	增广随机区组试验统计分析.....	1098
40.6.3	增广随机区组试验统计分析例子.....	1099
	参 考 文 献.....	1102
第 41 章	遗 传 统 计.....	1103
41.1	世代平均数分析方法.....	1103
41.1.1	世代平均数分析法的基本原理.....	1103
41.1.2	DPS 处理操作.....	1105
41.2	遗 传 力.....	1107
41.2.1	遗传力估计的基本方法及 DPS 系统的实现.....	1108
41.2.2	阈值性状遗传力.....	1114
	附 选择强度计算函数 ios(x) 介绍.....	1117
41.3	重复力(率).....	1117
41.3.1	数量性状重复力.....	1117
41.3.2	阈值性状重复力.....	1119
41.4	遗 传 相 关.....	1121
41.4.1	亲子资料遗传相关估计.....	1121
41.4.2	同胞资料遗传相关估计.....	1123
41.4.3	植物数量性状遗传相关分析.....	1125
41.5	选 择 指 数.....	1127
41.5.1	综合选择指数.....	1127
41.5.2	约束选择指数.....	1129
41.6	最佳线性无偏预测(BLUP).....	1131
41.7	NCI 设计(两因素巢式设计).....	1133
41.8	NCII 设计(不完全双列杂交设计).....	1135
41.9	NCIII(回交系统)设计.....	1138
41.10	完全双列杂交 GRIFFING 配合力分析.....	1140
41.11	GRIFFING 分析: 亲本+正反交 F_1 组合.....	1144
41.12	GRIFFING 分析: 亲本+正交 F_1 组合.....	1148
41.13	GRIFFING 分析: 无亲本, 仅正反交 F_1 组合.....	1151
41.14	GRIFFING 分析: 无亲本, 仅一组 F_1 试验.....	1154
41.15	双列杂交设计 HAYMAN 分析法.....	1157

41.15.1 双列杂交 Hayman 分析法数学模型.....	1157
41.15.2 Hayman 分析方法的原理.....	1158
41.15.3 亲本及正反交 F_1 组合 Hayman 分析应用举例.....	1161
参 考 文 献.....	1164
第八篇 其他数据分析方法	1165
第 42 章 模糊数学方法.....	1167
42.1 模糊聚类分析.....	1167
42.1.1 模糊集的概念.....	1167
42.1.2 模糊分类关系.....	1168
42.1.3 模糊聚类.....	1170
42.1.4 DPS 平台操作示例.....	1172
42.2 模糊模式识别.....	1174
42.2.1 方法简介.....	1174
42.2.2 DPS 平台的操作示例.....	1176
42.3 模糊相似优先比方法.....	1178
42.3.1 方法简介.....	1178
42.3.2 DPS 平台的操作示例.....	1178
42.4 模糊综合评判.....	1180
42.4.1 方法简介.....	1180
42.4.2 DPS 平台的操作示例.....	1181
42.5 模糊关系方程求解.....	1184
42.5.1 方法简介.....	1184
42.5.2 DPS 平台的操作示例.....	1185
42.6 综合评判逆问题.....	1186
42.6.1 方法简介.....	1186
42.6.2 DPS 平台操作示例.....	1186
参 考 文 献.....	1187
第 43 章 灰色系统分析.....	1189
43.1 关联度分析.....	1189
43.1.1 原理与方法简介.....	1190
43.1.2 DPS 平台的操作步骤.....	1193
43.1.3 应用示例.....	1194
43.2 灰色动态(GM)建模基本原理.....	1197
43.2.1 灰色动态建模原理.....	1197
43.2.2 常见的线性微分拟合模型.....	1198
43.3 灰色数列 GM(1, 1)模型.....	1201

43.3.1 建立GM(1, 1)模型的基本步骤.....	1201
43.3.2 DPS 平台的操作示例.....	1203
43.4 灰色数列GM(2, 1)模型.....	1205
43.4.1 建立GM(2, 1)模型的基本步骤.....	1205
43.4.2 DPS 平台的操作示例.....	1208
43.5 灰色数列GM(1, N)模型.....	1208
43.5.1 建立GM(1, N)模型的基本步骤.....	1208
43.5.2 DPS 平台的操作示例.....	1210
43.6 灾变预测.....	1212
参 考 文 献.....	1214
第 44 章 多试验、多指标综合评价.....	1215
44.1 离散型变量 META 分析.....	1215
44.2 连续型变量资料分析.....	1218
44.3 含亚类资料 META 分析.....	1221
44.4 TOPSIS 法.....	1226
44.5 综合指数法.....	1228
44.6 投影寻踪分类.....	1231
44.7 层次分析法.....	1233
44.7.1 层次分析法的基本原理与步骤.....	1234
44.7.2 层次分析法的应用.....	1238
参 考 文 献.....	1246
配套光盘使用说明.....	1247

第六篇 常用数值分析

数值计算是计算机处理科学和工程技术研究的重要内容。作为实用的实验统计的计算机数据处理平台，我们把重点放在一些统计分布理论上，处理一些带有随机误差的试验数据，根据各个专业的规律建立数学模型，以期有助于解释自然界的某些现象的运动过程。而一般用于工程技术的、精确的数学计算问题的数值分析技术目前在生物统计分析方面应用较少。因此我们仅介绍一些最基本的数值分析技术，以便读者参考、应用。

考虑到本书的主要对象是各个领域的科研工作者，故本篇只对一些数值分析方法，在简述其原理的基础上，给出一些例子，使读者能在 DPS 计算机处理平台上灵活地应用。

DPS 平台主要包括以下几类数值分析工具：

- 矩阵计算：计算数据矩阵各列(变量)均值、方差、标准差及变异系数及矩阵的奇异值分解；计算数据矩阵各个变量间的相关系数、实对称矩阵以及一般实矩阵特征值、特征向量计算，矩阵求逆等。
- 方程求解：包括求解线性方程组、非线性方程组和实系数多项式求根。
- 常用微积分数值分析技术：包括定积分、多重积分、求导数以及求解微分方程组初值问题，这是本篇的重点。如在生物学中有很多生物现象的变化过程是动态的，如著名的逻辑斯蒂模型

$$\frac{dN_t}{dt} = N_t r \left(1 - \frac{N_t}{K} \right)$$

和 Lotka-Volterra 竞争模型

$$\begin{cases} \frac{dx}{dt} = ax - bx^2 - cxy \\ \frac{dy}{dt} = ey - fxy - gy^2 \end{cases}$$

等都是微分方程。

在 DPS 平台上，微积分数值分析技术是模拟分析生态过程、揭示生物运动规律的有用工具之一。


第 29 章 矩阵计算

线性代数是多变量统计分析和计算方法的重要基础。在统计学教学、科研及应用工作中经常碰到矩阵运算。这类运算尽管原理相当简单，但实施计算操作却相当费时费力。因此，DPS 系统包括了常用的矩阵运算功能，供读者在处理平台上灵活运用。

DPS 数据处理系统提供了许多种常用的矩阵计算功能，所有功能以菜单和工具按钮两种形式给出，这里仅介绍一些主要计算功能。

进行数据矩阵计算前需要先定义数据矩阵块。在作矩阵的加、减、乘运算时需要两个数据矩阵，数据块的定义最通用的方法是将鼠标位置移到数据块的块首，按下鼠标左按钮并拖动鼠标到数据块的块尾位置，然后松开按钮。如需定义第二个数据块，可在按下 Ctrl 键时再按下并拖动鼠标，即可定义第二个数据块。在此将有关计算功能的定义及操作使用方法予以介绍，并以多元线性回归参数求解为例，对矩阵运算的应用技巧进行说明。

29.1 矩阵转置

如欲对数据矩阵进行转置，先将数据编辑、定义成数据块，再用鼠标点击工具按钮。如对图 29-1 所示的数据矩阵进行转置，可将数据按图 29-1 方式定义数据块。

	A	B	C	D	E	F	G	H
1		虚拟变量	x1	x2	x3	x4	Y	
2		1	7	26	6	60	78.5	
3		1	1	29	15	52	74.3	
4		1	11	56	8	20	104.3	
5		1	11	31	8	47	87.6	
6		1	7	52	6	33	95.9	
7		1	11	55	9	22	109.2	
8		1	3	71	17	6	102.7	
9		1	1	31	22	44	72.5	
10		1	2	54	18	22	93.1	
11		1	21	47	4	26	115.9	
12		1	1	40	23	34	83.8	
13		1	11	66	9	12	113.3	
14		1	10	68	8	12	109.4	
15								

图 29-1 矩阵转置的数据块编辑、定义方式

然后，点击按钮, 即可得到结果如下：

虚拟变量	1	1	1	1	1	1	1	1	1	1	1	1	1
x1	7	1	11	11	7	11	3	1	2	21	1	11	10
x2	26	29	56	31	52	55	71	31	54	47	40	66	68
x3	6	15	8	8	6	9	17	22	18	4	23	9	8
x4	60	52	20	47	33	22	6	44	22	26	34	12	12
Y	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

29.2 矩阵基本运算

29.2.1 矩阵加

设有两个 n 行 m 列的矩阵 $A=(a_{ij})$ 和 $B=(b_{ij})$, 那么 A 和 B 的和为

$$\begin{aligned}
 A+B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\
 &= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2m}+b_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \cdots & a_{nm}+b_{nm} \end{pmatrix}
 \end{aligned}$$

需要注意, 只有当两个矩阵的行数和列数都相同时, 这两个矩阵才能进行相加运算。在 DPS 平台上, 定义两个数据矩阵块, 并按按钮 \oplus , 即可完成矩阵的加法运算。例如, 我们希望对图 29-2 中的两个矩阵实施相加运算, 只需按图 29-2 方式编辑定义数据块。

	A	B	C	D	E	F	G	H
1		矩阵 A				矩阵 B		
2		3	6	5		3	6	5
3		4	2	1		4	2	1
4		1	1	7		1	1	7
5								

图 29-2 实施矩阵加法运算数据编辑定义图


然后在 DPS 数据处理系统中用鼠标点击按钮 \oplus , 即可得到计算结果:

5	7	12
7	8	3
2	3	7

29.2.2 矩阵减


设有两个 n 行 m 列的矩阵 $A=(a_{ij})$ 和 $B=(b_{ij})$, 那么 A 和 B 之差为

$$\begin{aligned}
 A - B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\
 &= \begin{pmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1m} - b_{1m} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2m} - b_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} - b_{n1} & a_{n2} - b_{n2} & \cdots & a_{nm} - b_{nm} \end{pmatrix}
 \end{aligned}$$

需要注意, 只有当两个矩阵的行数和列数都相同时, 这两个矩阵才能进行相减运算。在 DPS 平台上, 定义两个数据矩阵块, 第一个矩阵是被减数, 第二个是减数, 并按按钮 , 即可完成矩阵的减法运算。例如, 我们希望对图 29-3 中的两个矩阵实施相减运算, 只需按图 29-3 方式编辑定义数据块。

	A	B	C	D	E	F	G	H
1		矩阵 A				矩阵 B		
2		3	6	5		3	6	5
3		4	2	1		4	2	1
4		1	1	7		1	1	7
5								

图 29-3 实施矩阵减法运算数据编辑定义图

然后在 DPS 数据处理系统中用鼠标点击按钮 , 即可得到计算结果

计算结果		
1	5	-2
1	-4	-1
0	-1	7


29.2.3 矩阵相乘

若实施矩阵 $A=(a_{ij})_{m \times p}$ 和 $B=(b_{ij})_{p \times n}$ 相乘, 必须满足前一个矩阵(矩阵 A)的列数 p 和后一个矩阵(矩阵 B)的行数相等, 所得到的结果矩阵为 $C=(c_{ij})_{m \times n}$:

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}$$


$$= a_{i1} b_{1j} + a_{i2} b_{2j} + \dots + a_{ip} b_{pj}$$

可见，乘积矩阵 C 的元素 c_{ij} 等于矩阵 A 中的第 i 行与矩阵 B 中的第 j 列对应元素乘积的代数和。

矩阵乘法在生态学中的应用是广泛的，如在种群生态学中应用甚广的 Leslie(1945, 1948)种群矩阵模型，其实质是简单的矩阵乘法运算过程。在 DPS 平台上，定义两个数据块，第一个矩阵是被乘数，第二个是乘数，并按按钮 ，即可完成矩阵的乘法运算。例如，对图 29-4 中的两个矩阵实施乘法运算，只需按图 29-4 方式编辑定义数据块。

	A	B	C	D	E	F	G
1							矩阵 B
2	矩阵 A						2
3		1	2	3	4		1
4		2	1	0	3		2
5		1	1	1	1		1
6							

图 29-4 实施矩阵乘法运算数据编辑定义图

然后在 DPS 数据处理系统中用鼠标点击按钮 ，系统将输出计算结果 $\begin{pmatrix} 14 \\ 8 \\ 6 \end{pmatrix}$ 。


29.3 矩阵自乘

对任意实矩阵 $A = (a_{ij})_{n \times m}$ ，将它转置后得 $A^T = (a_{ij})_{m \times n}$ ，在与它原来的矩阵相乘，所得到的结果矩阵为

$$C = (c_{ij})_{m \times m}$$

$$= \sum_{k=1}^n a_{ik} a_{kj}$$


$$= \sum_{k=1}^n a_{i1} a_{1j} + a_{i2} a_{2j} + \dots + a_{in} a_{nj}$$

即 C 为一方阵。矩阵自乘在多元统计分析中的应用很多，如多元线性回归模型参数最小二乘估计过程中建立联立方程组即是对变量矩阵实施自乘($X'X$)的过程。在 DPS 平台上，工具按钮  可完成矩阵自乘运算。例如，在建立一个有 4 个自变量，1 个因变量的回归

模型时，其联立方程组可通过对图 29-5 中的数据矩阵自乘运算实现。计算时只需按图 29-5 方式编辑数据(第 1 列是虚拟变量，全为 1)。

	A	B	C	D	E	F	G	H
1		虚拟变量	x1	x2	x3	x4	Y	
2		1	7	26	6	60	78.5	
3		1	1	29	15	52	74.3	
4		1	11	56	8	20	104.3	
5		1	11	31	8	47	87.6	
6		1	7	52	6	33	95.9	
7		1	11	55	9	22	109.2	
8		1	3	71	17	6	102.7	
9		1	1	31	22	44	72.5	
10		1	2	54	18	22	93.1	
11		1	21	47	4	26	115.9	
12		1	1	40	23	34	83.8	
13		1	11	66	9	12	113.3	
14		1	10	68	8	12	109.4	
15								

图 29-5 为建立多元线性回归模型联立方程组而编辑定义的数据矩阵

然后用鼠标点击数据矩阵自乘按钮 ，系统将输出计算结果如下：

13.0000	97.0000	626.0000	153.0000	390.0000	1240.5000
97.0000	1139.0000	4922.0000	769.0000	2620.0000	10032.0000
626.0000	4922.0000	33050.0000	7201.0000	15739.0000	62027.8008
153.0000	769.0000	7201.0000	2293.0000	4628.0000	13981.5000
390.0000	2620.0000	15739.0000	4628.0000	15062.0000	34733.3008
1240.5000	10032.0000	62027.8008	13981.5000	34733.3008	121088.0938

29.4 矩阵样本方差


对资料矩阵

$$\begin{array}{cccc}
 x_{11} & x_{12} & \cdots & x_{1p} \\
 x_{21} & x_{22} & \cdots & x_{2p} \\
 \vdots & \vdots & & \vdots \\
 x_{N1} & x_{N2} & \cdots & x_{Np}
 \end{array}$$

根据公式

$$S_{ij} = \frac{1}{N-1} \sum_{a=1}^N (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)$$

计算样本的协方差阵。

在 DPS 平台上，工具按钮  可完成该功能。例如，有 4 个变量、13 个样本的数据矩阵，可编辑成图 29-6 中数据块格式。

	A	B	C	D	E	F
1		x1	x2	x3	x4	
2		7	26	6	60	
3		1	29	15	52	
4		11	56	8	20	
5		11	31	8	47	
6		7	52	6	33	
7		11	55	9	22	
8		3	71	17	6	
9		1	31	22	44	
10		2	54	18	22	
11		21	47	4	26	
12		1	40	23	34	
13		11	66	9	12	
14		10	68	8	12	
15						

图 29-6 计算样本协方差阵数据编辑方式



计算时点击工具按钮  就可完成该功能计算，给出如下结果：

34.60256	20.92308	-31.05128	-24.16667
20.92308	242.1410	-13.87820	-253.4167
-31.05128	-13.87820	41.02564	3.166667
-24.16667	-253.4167	3.166667	280.1667

29.5 矩阵总体方差

对资料矩阵

$$\begin{array}{cccc}
 x_{11} & x_{12} & \cdots & x_{1p} \\
 x_{21} & x_{22} & \cdots & x_{2p} \\
 \vdots & \vdots & & \vdots \\
 x_{N1} & x_{N2} & \cdots & x_{Np}
 \end{array}$$

根据公式 $S_{ij} = \frac{1}{N} \sum_{a=1}^N (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)$ 计算总体协方差阵。矩阵总体协方差阵的计算，在 DPS 平台上，工具按钮  可完成该功能。若对图 29-6 中数据块进行计算。计算时点击工具按钮  可得到计算结果如下：


31.94083	19.31361	-28.66272	-22.30769
19.31361	223.5148	-12.81065	-233.9231
-28.66272	-12.81065	37.86982	2.923077
-22.3076	-233.9231	2.923077	258.6154

29.6 解正规方程组

本功能采用 Cholesky 分解法(即平方根法)矩阵为对称正定的线性代数方程组。计算时将系数矩阵放在左边，常数放在右边。如将图 29-5 数据进行矩阵自乘得到的结果，定义成数据矩阵图 29-7。

	A	B	C	D	E	F	G	H
1								
2		13.0000	97.00	626.00	153.00	390.00	1240.5000	
3		97.0000	1139.00	4922.00	769.00	2620.00	10032.0000	
4		626.0000	4922.00	33050.00	7201.00	15739.00	62027.8008	
5		153.0000	769.00	7201.00	2293.00	4628.00	13981.5000	
6		390.0000	2620.00	15739.00	4628.00	15062.00	34733.3008	
7								

图 29-7 解线性方程组数据编辑格式

再点击解线性方程组工具按钮 ，即可得到计算结果：

方程系数	
x1=	64.97440
x2=	1.524686
x3=	0.483688
x4=	0.074888
x5=	-0.170013

29.7 矩阵求逆

29.7.1 矩阵求逆及行列式值

对 n 阶方阵，采用高斯-约当选主元消去法对数据矩阵求逆并计算其行列式值。但应

注意，实施求逆运算的数据矩阵必须是方阵，且非奇异矩阵。

在 DPS 平台上，定义待求逆的数据矩阵块，然后用鼠标点击按钮 x^{-1} ，即可完成矩阵求逆运算。例如，对图 29-8 中 4 阶方阵实施求逆运算，只需按图 29-8 方式编辑定义数据块。

	A	B	C	D	E	F
1						
2		1	2	3	5	
3		3	2	1	3	
4		3	4	3	2	
5		9	7	4	9	
6						

图 29-8 实施矩阵求逆运算数据编辑定义图

然后在 DPS 数据处理系统中用鼠标点击按钮 x^{-1} ，系统将输出结果：

逆矩阵系数			
0.2500	13.2500	2.0000	-5.0000
-0.7500	-20.7500	-3.0000	8.0000
0.7500	17.7500	3.0000	-7.0000
0.0000	-5.0000	-1.0000	2.0000
矩阵行列值=4.0000000			

29.7.2 矩阵广义逆

在前面矩阵求逆运算中，一个矩阵的 A^{-1} 与矩阵 A 相乘的积是单位矩阵，即 $A^{-1}A = AA^{-1} = I$ 。如果 $\rho(A) = \gamma = n$ ，矩阵 A 是满秩(full rank)矩阵，存在 A^{-1} ；如果 $\rho(A) = \gamma < n$ ，则不存在 A^{-1} 。

但对于任何矩阵，无论是否满秩，都可以找到其广义逆。如果 $AA^{-}A = A$ ，那么 A^{-} 是矩阵 A 的一个广义逆矩阵。矩阵 A 的广义逆 A^{-} 数目是无穷的。即对于任意矩阵 Z ， $A^{-} + (I - A^{-}A)Z$ 也是矩阵 A 的一个广义逆。

对于任何矩阵 A 具有 $\rho(A) = \gamma \leq n$ ，由奇异值分解可以得到

$$A = P \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} Q^T$$

$n \times m$ $n \times n$ $m \times m$

其中， D 是对角线矩阵， P 和 Q 是正交矩阵， $P^T P = I$ ， $Q^T Q = I$ 。

矩阵 A 的一个广义逆可以由式

$$A^{-} = Q \begin{bmatrix} D^{-1} & E \\ B & C \end{bmatrix} P^T$$

得到，其中 B ， C ， E 可以是任何值的矩阵。这里 A^{-} 为矩阵 A 的减号逆。

对于任何矩阵，都可以找到唯一的一个广义逆，称为 Moore-Penrose 广义逆，一般

用 A^+ 表示, 该广义逆满足以下全部条件:

$$AA^+A=A, \quad A^+AA^+=A^+, \quad (A^+A)^T=(A^+A), \quad (AA^+)^T=(AA^+)$$

矩阵广义逆 A^+ 可以由式

$$A^+ = Q \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix} P^T$$


求出。

在 DPS 平台上, 定义待求广义逆的数据矩阵块, 然后用鼠标点击按钮 A^- 或 A^+ , 即可求得矩阵的广义逆 A^- 或 A^+ 。例如, 对图 29-6 中矩阵实施求广义逆运算, 只需按图 29-6 方式编辑定义数据块。然后用鼠标点击按钮 A^+ , 即可得到矩阵的 Moore-Penrose 逆 A^+ 如下:

-0.0193	-0.0176	0.0050	0.0158	-0.0315	0.0085	-0.0159	0.0130	-0.0086	0.0554	0.0132	0.0023	-0.0101
0.0040	0.0017	0.0018	-0.0036	0.0099	0.0006	0.0059	-0.0066	0.0020	-0.0094	-0.0059	0.0033	0.0066
-0.0261	-0.0065	-0.0026	0.0031	-0.0329	0.0012	0.0008	0.0283	0.0072	0.0233	0.0309	-0.0028	-0.0133
0.0111	0.0067	-0.0007	0.0032	0.0074	-0.0011	-0.0032	-0.0011	-0.0013	-0.0052	-0.0034	-0.0022	-0.0002


29.8 奇异值分解

本系统应用 Householder 变换、结合变形 QR 算法, 对一般实矩阵进行奇异值的分解。

在 DPS 平台上, 对待求奇异值的数据矩阵进行编辑、定义数据块, 然后用鼠标点击按钮 , 即可完成矩阵奇异值分解。

	A	B	C	D	E	F
1						
2		1	2	3	4	
3		6	7	8	9	
4		1	2	13	0	
5		16	17	8	9	
6		2	4	3	4	
7						

图 29-9 矩阵奇异值分解数据编辑、定义格式

例如, 对图 29-9 中 5×4 阶矩阵实施奇异值分解运算, 编辑定义数据块, 然后点击按钮 , 系统将输出结果。

计算结果			
矩阵 U			
-0.148914	-0.104745	-0.127199	0.448480
-0.462152	-0.157452	0.470364	0.640293

	-0.241480	-0.903891	-0.046684	-0.349218
	-0.815321	0.382755	-0.012546	-0.417256
	-0.202927	-0.026770	-0.871918	0.304687
矩阵奇异值 W				
	31.61621	11.70207	1.042008	5.602748
变换矩阵 V				
	-0.525499	-0.591090	-0.455922	-0.408165
	0.351835	0.280320	-0.883836	0.128325
	0.675363	-0.725683	0.055951	0.118902
	-0.379404	-0.213115	-0.088534	0.895989

29.9 实对称矩阵特征值和特征向量

29.9.1 方法简介

设 A 是实对称矩阵，对 A 实施一系列雅可比(Jacobi)变换，即对 A 施行一相似矩阵

$$A_k = R_k^{-1} A_{k-1} R_k$$

的序列变换。其中， R_k 是平面旋转矩阵，它可由下面的旋转(正交)变换逐步求得。对于某一指定的 r_{pq} ，可取正交阵

$$R_k = \begin{bmatrix} 1 & & & & & & & \\ & \text{O} & & & & & & \\ & & \cos \theta & & \sin \theta & & & \\ & & & 1 & & & & \\ & & & & \text{O} & & & \\ & & & & & 1 & & \\ & & -\sin \theta & & \cos \theta & & & \\ & & & & & & \text{O} & \\ & & & & & & & 1 \end{bmatrix} \begin{matrix} \text{第 } p \text{ 行} \\ \text{第 } q \text{ 行} \end{matrix}$$

第 p 列 第 q 列

其中， θ 为待定角，它的元素为

$$r_{pp} = r_{qq} = \cos \theta, \quad r_{pq} = \sin \theta, \quad r_{qp} = -\sin \theta, \quad r_{ii} = 1, \quad i \neq p, q$$

其余元素为零。显然这是正交变换，即 $R^{-1}=R^T$ ，所有 $A_k(k=1, 2, \dots)$ 都是对称矩阵， A_k 与 A_{k-1} 只是在第 p, q 行及列上不一样，它们之间的关系如下：

$$\begin{aligned} a_{ip}^{(k)} &= a_{pi}^{(k)} = a_{ip}^{(k-1)} \cos \theta - a_{iq}^{(k-1)} \sin \theta \\ a_{iq}^{(k)} &= a_{qi}^{(k)} = a_{ip}^{(k-1)} \sin \theta - a_{iq}^{(k-1)} \cos \theta \end{aligned} \quad (i \neq p, q)$$

$$\begin{aligned} a_{pp}^{(k)} &= a_{pp}^{(k-1)} \cos^2 \theta + a_{qq}^{(k-1)} \sin^2 \theta - 2a_{pq}^{(k-1)} \sin \theta \cos \theta \\ a_{qq}^{(k)} &= a_{pp}^{(k-1)} \sin^2 \theta + a_{qq}^{(k-1)} \cos^2 \theta + 2a_{pq}^{(k-1)} \sin \theta \cos \theta \\ a_{pq}^{(k)} &= a_{qp}^{(k-1)} = (a_{pp}^{(k-1)} - a_{qq}^{(k-1)}) \sin \theta \cos \theta + a_{pq}^{(k-1)} (\cos^2 \theta - \sin^2 \theta) \end{aligned}$$

若旋转角 θ 满足

$$\begin{aligned} \tan 2\theta &= -V_2 / U, \quad V_2 = a_{pq}^{(k-1)} \\ U &= \frac{1}{2}(a_{pp}^{(k-1)} - a_{qq}^{(k-1)}) \end{aligned}$$

则有 $a_{pq}^{(k)} - a_{qp}^{(k)} = 0$

若 $-\frac{\pi}{4} < \theta \leq \frac{\pi}{4}$, 那么在实际计算中, 并不需要计算三角函数, 只需利用三角函数之

间的纯粹代数关系, 即 $G = -\sin(U) \frac{V_2}{\sqrt{V_2^2 + U^2}} \quad (U \neq 0), \quad G=1 \quad (U=0),$

$\sin \theta = \frac{G}{\sqrt{2(1 + \sqrt{1 - G^2})}}, \quad \cos \theta = \sqrt{1 - \sin^2 \theta}$ 。每作一次正交变换, 主对角元素平方和增

加 $2(a_{pq}^{(k-1)})^2$, 同时对角线外元素的平方和相应减少此量。不断进行正交变换, 主对角线外的元素趋于零, 矩阵 A 趋于一对角矩阵, 对角线上的元素就是矩阵的各个特征值, 而 $S_k = R_1 R_2 \cdots R_k$ 的每一列就是相应的特征向量。


由此, 用雅可比旋转变换方法可求出实对称矩阵的全部特征值及其对应的特征向量, 求出的特征值按降序排列, 并重排相应的特征向量。

29.9.2 DPS 平台的操作示例

在 DPS 的电子表格中按待运算矩阵行列顺序输入各元素数据并定义成数据块(矩阵)如图 29-10:

	A	B	C	D	E	F	G
1							
2		10	1	2	3	4	
3		1	9	-1	2	-3	
4		2	-1	7	3	-5	
5		3	2	3	12	-1	
6		4	-3	-5	-1	15	
7							

图 29-10 实对称矩阵特征值及特征向量计算数据编辑、定义

然后用鼠标点击工具按钮, 或进入主菜单, 在菜单方式下选择执行“实对称矩阵求逆”功能, 系统立即给出分析结果。本例经计算后输出结果为:

No	特征值: 百分率		累计百分率	
1	19.17541890	36.1800366	36.1800	
2	15.8089209	29.8281534	66.0082	
3	9.3655548	17.6708586	83.6790	
4	6.9948378	13.1978075	96.8769	
5	1.6552662	3.12314380	100.0000	
特征向量				
0.1745051	0.6237025	-0.0521511	0.6540829	-0.3872969
-0.2473025	0.1591011	0.8599639	0.1996812	0.3662210
-0.3616418	0.2272975	-0.5055751	0.2565105	0.7043773
-0.2644109	0.6926845	-0.0002012	-0.6604028	-0.1189262

29.10 实矩阵特征值和特征向量

29.10.1 方法简介

DPS 采用收敛较快并带原点位移的两步 QR 算法, 求解 n 阶实矩阵 A 的全部特征值和特征向量。计算方法如下:

(1) 对矩阵 A 进行平衡处理。置换矩阵的行和列, 把部分特征值分离到对角线上, 这样分离出来的特征值, 不需要进一步计算便可使用, 从而降低了求特征值的阶数。对矩阵的其余部分进行平衡处理, 由一系列的对角相似变换组成。这些变换使被平衡部分所对应的行和列的元素数值之和接近相等, 因而减少矩阵的范数, 提高计算精度。变换矩阵的有关信息保留下来, 在求特征向量进行反变换时使用。

(2) 用初等相似变换把矩阵化为海森伯格(Hessenberg)型矩阵(以下简写上 H 阵)。计算一般矩阵的特征值(或同时计算其特征向量)之前, 先把矩阵化为上 H 阵(即当 $i > j+1$ 时, 其元素 $h_{ij}=0$), 可节省较多的计算量。

用初等相似变换把一个矩阵化为上 H 阵, 需要经过 $n-2$ 步, 它们依次把矩阵 A 的第 1, 2, \dots , $n-2$ 列化为上 H 阵。设 A_r 为 A ($A_1=A$) 经过 $r-1$ 步变换所得的矩阵, 其第 r 步变换的计算方法如下:

- ① 选 $|a_{ir}^{(r)}|$ ($i=r+1, \dots, n$) 中最大值(若不止一个, 取先遇到的那一个), 以 $|a_{(r+1),r}^{(r)}|$ 表示。
- ② 交换第 $(r+1)$ ' 行与第 $r+1$ 行, 交换第 $(r+1)$ ' 列与第 $r+1$ 列。
- ③ 对于从 $r+2$ 到 n 的每一个 i 值, 计算: $n_{i,r+1} = a_{ir}^{(r)} / a_{r+1,r}^{(r)}$, (第 i 行) $-n_{i,r+1} \times$ (第 $r+1$ 行), $n_{i,r+1} \times$ (第 i 列) 加到 (第 $r+1$ 列)。这些运算等价于

$$A_{r+1} = N_{r+1}^{-1} I_{r+1, (r+1)} A_r I_{r+1, (r+1)} N_{r+1}$$

此处矩阵 $I_{r+1, (r+1)}$ 是初等置换矩阵(即交换行或列的变换矩阵), N_{r+1} 是一初等矩阵, 即

$$(N_{r+1})_{i, r+1} = n_{i, r+1} \quad (i=r+2, \dots, n)$$

其余元素与单位矩阵相同, 即 $(N_{r+1})_{ij} = \delta_{ij}$ 。

(3) 海森伯格矩阵的 QR 算法。

①基本思路。Francis 提出的 QR 算法的基本思想是对矩阵进行分解, 获得一个类似于原矩阵的序列 $\{A_k\}$, 使其收敛到一个易于求得特征值的形式。LR 算法比 QR 算法收敛速度快, 但不稳定。从目前来说, QR 算法是求一般实矩阵全部特征值和特征向量的最有效方法。

由于基本 QR 算法的收敛往往过慢, 因此常用带原点位移的 QR 算法。设 A 为 $n \times n$ 阶的实矩阵, 令 $A_1 = A$, 并取一适当的数 k_1 , 对矩阵 $A_1 - k_1 I$ 作 QR 分解 $A_1 - k_1 I = Q_1 R_1$, 然后令 $A_2 = Q_1 R_1 + k_1 I$, 并选取适当数 k_2 , 对矩阵 $A_2 - k_2 I$ 作 QR 分解 $A_2 - k_2 I = Q_2 R_2$, 其一般关系式为

$$A_s - k_s I = Q_s R_s$$

$$A_{s+1} = Q_s R_s + k_s I$$

$$A_{s+1} - k_{s+1} I = Q_{s+1} R_{s+1}$$

推出

$$R_s = Q_s^T (A_s - k_s I)$$

$$\begin{aligned} A_{s+1} &= Q_s^T (A_s - k_s I) Q_s + k_s I \\ &= Q_s^T A_s Q_s \end{aligned}$$

这里 Q_s 是正交阵, R_s 是上三角阵。由此可见 A_{s+1} 与 A_s 相似, 具有相同的特征值。适当选取位移量 k_s , 可使 A_s 最末一行的非对角元素极其迅速地趋于零, 而次对角元素也逐渐趋于零。若原始矩阵 A_1 是上 H 阵, 则容易证明, 所有 A_s 也是上 H 阵。对上 H 阵进行 QR 分解, 计算量比较小。

因为 A_1 是实矩阵, 它可能有复特征值, 如果上面变换带有复位移量 k_s (一般 k_s 取 A_s 的一个粗糙的近似特征值), 则造成复数运算。所以我们采用双步 QR 算法, 它避免了复数运算, 在计算量上也是比较经济的。

② 双步 QR 算法。实的上 H 阵 A_s 经两步带位移的 QR 迭代得到的上 H 阵 A_{s+2} :

$$\begin{aligned} A_{s+2} &= Q_{s+1}^T Q_s^T A_s Q_s Q_{s+1} \\ &= (Q_s Q_{s+1})^T A_s (Q_s Q_{s+1}) \end{aligned}$$

我们的目的是要找出 $Q_s Q_{s+1}$ 的计算方法。由

$$Q_s(A_{s+1}-k_{s+1}I) = (A_s-k_{s+1}I)Q_s$$

得到

$$Q_s Q_{s+1} R_{s+1} R_s = (A_s - k_{s+1} I) Q_s R_s$$

即

$$(Q_s Q_{s+1})(R_{s+1} R_s) = (A_s - k_{s+1} I)(A_s - k_s I) \equiv M$$

因此, $Q_s Q_{s+1}$ 由它的第一列唯一确定, $Q_s Q_{s+1}$ 是将矩阵 M 约化为上三角阵 $R_{s+1} R_s$ 的酉矩阵, 它的第一列由 M 阵的第一列确定, M 的第一列是(下面省略了 A_s 的元素的 s 上指标)

$$\begin{aligned} m_{11} &= a_{11}^2 - (k_s + k_{s+1})a_{11} + a_{12}a_{21} + k_s k_{s+1} \\ m_{21} &= a_{21}[a_{11} + a_{22} - (k_s + k_{s+1})] \\ m_{31} &= a_{32}a_{21} \\ m_{41} &= L = m_{n1} = 0 \end{aligned}$$

设 $P_1 = I - 2u_1 u_1^T$ 是将向量 $(m_{11}, m_{21}, m_{31}, 0, \dots, 0)^T$ 变换为 $a_1 e_1$ 的 Householder 变换, 易知 u_1 除开始三个分量外, 其余分量全为零。由于 A_s 是上 H 阵, 所以 $P_1^T A_s P_1$ (例如, 对 $n=6$) 有如下形式:

$$P_1^T A P_1 = \begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \otimes & \times & \times & \times & \times & \times \\ \otimes & \otimes & \times & \times & \times & \times \end{bmatrix}$$

然后再对 $P_1^T A P_1$ 应用一系列 Householder 变换矩阵 P_2, P_3, \dots, P_{n-1} (它们的第一行和第一列元素除对角元素为 1 外, 其余都为零), 使

$$H \equiv P_{n-1}^T L P_2^T (P_1^T A P_1) P_2 L P_{n-1}$$

是上 H 阵, 由于 P_2, P_3, \dots, P_{n-1} 的第一列与 P_1 的第一列相同, 得到

$$H = A_{s+2}$$

由此完成一次典型的双步 QR 迭代。

③ 原点位移量 k_s 和 k_{s+1} 的选择。点位移量 k_s 和 k_{s+1} 由 A_s 的右下角二阶矩阵的特征值所确定, 即有

$$\begin{aligned} k_s + k_{s+1} &= a_{n-1, n-1}^{(s)} + a_{n, n}^{(s)} \\ k_s k_{s+1} &= a_{n-1, n-1}^{(s)} a_{n, n}^{(s)} - a_{n-1, n}^{(s)} a_{n, n-1}^{(s)} \end{aligned}$$

通常, 使用上述方法确定的原点位移, 迭代过程收敛是很快的。当 $a_{n, n-1}^{(s)}$ 或 $a_{n-1, n-2}^{(s)}$ 小

到可以忽略时,即求得一个实特征值或一对共轭复特征值,相应的矩阵降低一阶或二阶,继续上述迭代,直到求出全部特征值。若迭代多次仍求不出特征值,则原点位移量改用

$$k_1 + k_2 = 0.75 \left(|a_{n,n-1}| + |a_{n-1,n-2}| \right)$$

$$k_1 k_2 = \left(|a_{n,n-1}| + |a_{n-1,n-2}| \right)^2$$

再继续进行迭代。

④ 特征向量的计算。QR 算法最后得出的矩阵 $T=(t_{ij})$ 一般不是严格的三角形矩阵,其对角线上可能出现二阶小方阵,每一小方阵对应一对实特征值或一对共轭复特征值。计算对应于实特征值 λ_i 的特征向量 X_i 的算法是

$$x_{ji} = 0 \quad (j = i+1, \dots, n)$$

$$x_{ji} = 1$$

$$x_{ji} = - \left(\sum_{k=j+1}^i t_{jk} x_{ki} \right) / (\lambda_j - \lambda_i) \quad (j = i-1, \dots, 1)$$

(a) 如果在第 j 行和第 $j-1$ 行上有一个二阶小方阵,则 x_{ji} 和 $x_{j-1,i}$ 改由下列二阶方程组解出:

$$\begin{cases} (t_{j-1,j-1} - \lambda_i)x_{j-1,i} + t_{j-1,j}x_{ji} = - \sum_{k=j+1}^i t_{j-1,k}x_{ki} \\ t_{j,j-1}x_{j-1,i} + (t_{jj} - \lambda_i)x_{ji} = - \sum_{k=j+1}^i t_{jk}x_{ki} \end{cases} \quad (29.1)$$

(b) 对于一对共轭复特征值,仅需解出虚部为正的实特征值所对应的特征向量,令 $x_{ii} = 1+0i$, $x_{i-1,i}$ 满足下列方程:

$$\begin{cases} (t_{i-1,i-1} - \lambda_i)x_{i-1,i} + t_{i-1,i} = 0 \\ t_{i,i-1}x_{i-1,i} + (t_{ii} - \lambda_i) = 0 \end{cases} \quad (29.2)$$

$x_{i-1,i}$ 由上式中系数较大的一个方程决定,特征向量的其他分量可由方程组(29.1)和(29.2)给出。求出 T 的对应于 λ_i 的特征向量 x_i 后,根据

$$S^{-1}AS = H, \quad P^{-1}HP = T$$


便可求出 A 对应于 λ_i 的特征向量 SPx_i 。

29.10.2 DPS 平台的操作示例

在 DPS 的电子表格中按行列顺序输入所求矩阵元素数据,编辑定义成数据块(矩阵)如图 29-11。

	A	B	C	D	E	F
1						
2		3	1	2	5	
3		2	1	3	7	
4		3	1	2	4	
5		4	1	3	2	
6						

图 29-11 一般实矩阵特征值及特征向量计算数据编辑、定义

然后单击工具按钮 , 或进入主菜单, 选择执行“一般实矩阵求逆计算”, 系统立即在当前光标下给出分析结果。本例经计算后输出结果为:

特征值	X1	X2	X3	X4
实部:	10.59198	0.19135	-2.36630	-0.41703
	0	0	0	0
特征向量				
	-0.71229	-0.36607	-0.53163	0.02885
	-0.83019	-1.01640	1.80181	-2.57435
	-0.65042	-0.09763	0.19167	0.62077
	-0.65533	0.63522	-0.13840	0.24684

29.11 应用矩阵运算组建多元线性回归模型

一般多元线性回归模型 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_m x_{im}$ 。式中, 令

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

推导可得回归系数 β 的最小二乘估计值

$$\hat{\beta} = (X'X)^{-1}X'Y$$

即方程参数过程可归结为一次矩阵自乘、一次矩阵求逆和两次矩阵相乘。例如, 若有 7 个样本, 每个样本含 3 个自变量, 1 个因变量如下:

$$X = \begin{bmatrix} 1 & 1.0 & 22 & 2.0 \\ 1 & 1.4 & 13 & 1.0 \\ 1 & 1.8 & 28 & 3.0 \\ 1 & 2.2 & 16 & 3.5 \\ 1 & 2.6 & 25 & 0.5 \\ 1 & 3 & 10 & 2.5 \\ 1 & 3.4 & 19 & 1.5 \end{bmatrix}, \quad Y = \begin{bmatrix} 0.614 \\ 0.350 \\ 0.753 \\ 0.819 \\ 0.097 \\ 0.711 \\ 0.418 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

试运用矩阵运算方法建立多元线性回归方程。

第 1 步. 计算 $X'X$, 如图 29-12 所示, 将原始数据矩阵增加一列虚拟变量 1, 然后定义数据块。

	A	B	C	D	E	F
1		虚拟变量	x1	x2	x3	y
2		1	1.0	22	2.0	0.6146
3		1	1.4	13	1.0	0.3506
4		1	1.8	28	3.0	0.7537
5		1	2.2	16	3.5	0.8195
6		1	2.6	25	0.5	0.0970
7		1	3.0	10	2.5	0.7114
8		1	3.4	19	1.5	0.4186
9						

图 29-12 计算矩阵自乘的数据块定义

再用鼠标点击工具按钮 , 得到计算结果如下:

	A	B	C	D	E	F
1						
2		7.000	15.40	133.0	14.00	
3		15.40	38.36	285.4	30.40	
4		133.0	285.4	2779	263.0	
5		14.00	30.40	263.0	35.00	
6						

图 29-13 $X'X$ 矩阵

第 2 步. 将上步得到的结果定义成数据块(图 29-13), 再按 x^{-1} 按钮进行矩阵求逆运算, 得到逆矩阵系数如下:

4.267744	-0.683453	-0.099297	-0.367325
-0.683453	0.235826	0.006934	0.016447
-0.099297	0.006934	0.004192	0.002193
-0.367325	0.016447	0.002193	0.144737
行列式值=52093.44			

第 3 步. 将前一步计算所得的逆矩阵系数和原始数据的转置矩阵相乘, 即按图 29-14 方式定义数据矩阵块。

	A	B	C	D	E	F	G
1							
2		4.267744	-0.683453	-0.099297	-0.367325		
3		-0.683453	0.235826	0.006934	0.016447		
4		-0.099297	0.006934	0.004192	0.002193		
5		-0.367325	0.016447	0.002193	0.144737		
6							
7	1	1	1	1	1	1	1
8	1	1.4	1.8	2.2	2.6	3	3.4
9	22	13	28	16	25	10	19
10	2	1	3	3.5	0.5	2.5	1.5
11							

图 29-14 逆矩阵系数与转置矩阵相乘的数据块定义

在 DPS 平台上用鼠标点击  按钮, 即可得到计算结果。

0.665109	1.652726	-0.844759	-0.110239	-0.175319	0.306105	-0.493624
-0.262190	-0.246711	-0.015480	0.003870	0.111262	0.134481	0.274768
0.004257	-0.032895	0.037152	-0.009288	0.024639	-0.031089	0.007224
-0.013158	-0.171053	0.157895	0.210526	-0.197368	0.065789	-0.052632

最后, 将上面的结果和与因变量(y)定义成数据块, 再次相乘, 便可得到多元线性回

归方程参数 β 的估计值, $\hat{\beta} = \begin{pmatrix} 0.255313 \\ -0.034654 \\ -0.005229 \\ 0.229100 \end{pmatrix}$ 。

参 考 文 献

陈明遼, 凌永祥. 1992. 实用计算机数值方法[M]. 西安: 西安交通大学出版社.
 上海机械学院, 安徽省计算中心. 1984. FORTRAN 应用程序库[M]. 上海: 上海科学技术文献出版社.

上海计算技术研究所. 1982. 电子计算机算法手册[M]. 上海: 上海教育出版社.

郑维敏等. 1986. 系统工程 FORTRAN 程序集[M]. 北京: 清华大学出版社.

第 30 章 方程求解及多项式求根

方程求解包括求解线性方程组、非线性方程组和实系数多项式求根等最常用的数值算法和分析技术。

30.1 求解线性方程组

30.1.1 方法简介

DPS 平台采用全主元高斯-约当消去法求解 $A[XY]=[BI]$ ，其中 A 为 $n \times n$ 非奇异矩阵， B 为 $n \times m$ 矩阵，均已知； $X_{n \times m}$ ， $Y_{n \times n}$ 未知。由于消去过程是在全矩阵中选主元(绝对值最大的元素)来进行的，故将舍去误差使其对结果的影响减少到最小。具体步骤是：

(1) 施行初等变换把 A 变为单位矩阵，则

$$X=A^{-1}B, \quad Y=A^{-1}$$

(2) 记 $A^{(0)}=(a_{ij}^{(0)})=A=(a_{ij})$ ， $B^{(0)}=B=(b_{ij}^{(0)})$ 。

第 k 步的矩阵为 $A^{(k)}=(a_{ij}^{(k)})_{n \times n}$ ， $B^{(k)}=(b_{ij}^{(k)})_{n \times m}$ ($k=1, 2, \dots, n$)。第 k 步计算：

① 选主元，设为 $a_{i_0 j_0}^{(k-1)}$ 。

② 若 $i_0=j_0$ ，则转③；否则交换矩阵 $[A^{(k-1)}B^{(k-1)}]$ 中的第 i_0 行与第 j_0 行，则 $a_{i_0 j_0}^{(k-1)}$ 移至矩阵 $A^{(k-1)}$ 的对角线上，得到的矩阵仍记为 $[A^{(k-1)}B^{(k-1)}]=[a_{ij}^{(k-1)}(b_{ij}^{(k-1)})]$ ，主元为 $a_{i_0 j_0}^{(k-1)}$ 。

③ 按下面计算公式消元：

$$\begin{aligned} p^{(k)} &= 1 / a_{i_0 i_0}^{(k-1)} \\ a_{i_0 j}^{(k)} &= a_{i_0 j}^{(k-1)} \cdot p^{(k)} \\ b_{i_0 l}^{(k)} &= b_{i_0 l}^{(k-1)} \cdot p^{(k)} \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - a_{i_0 j}^{(k)} \cdot a_{i_0 i}^{(k-1)} \\ b_{il}^{(k)} &= b_{il}^{(k-1)} - a_{i_0 l}^{(k)} \cdot a_{i_0 i}^{(k-1)} \end{aligned}$$

式中 $j=1, 2, \dots, n$ ； $j=1, 2, \dots, n$ ； $l=1, 2, \dots, m$ ； $i=1, 2, \dots, n$ ； $i \neq i_0$ 。

30.1.2 DPS 平台的操作示例

解线性方程组的数据编辑只需将方程组的有关系数值按要求编辑即可。如有线性方

程组

$$10x_1+x_2+2x_3+3x_4=4$$

$$x_1+9x_2-x_3+2x_4=-3$$

$$2x_1-x_2+7x_3+3x_4=-5$$

$$3x_1+2x_2+3x_3+12x_4=-1$$

在求解方程参数时，只需在电子表格中编辑有关的系数数据，并将系数定义成数据块如下：

	A	B	C	D	E	F	G
1							
2		10	1	2	3	4	
3		1	9	-1	2	-3	
4		2	-1	7	3	-5	
5		3	2	3	12	-1	
6							

然后进入菜单操作，在“数值分析”栏下选择“解线性方程组”功能项，按回车键后立即得到如下分析结果：

$$x_1=0.6257, \quad x_2=-0.5386, \quad x_3=-1.0144, \quad x_4=0.1036$$

目标函数值为 5.58e-13。

30.2 非线性方程组求解

非线性方程(组)求解实质上是函数的极值和最优化过程，最优化的方法又分为用导数的方法和不用导数的方法两种。一般说来，用导数的方法比不用导数(仅用函数值)的方法更有效，但它不一定能补偿求导数的附加运算。为稳妥起见，本系统提供给读者的是加速单纯形方法。该方法计算原理及其基本步骤，可参阅第 27.1 节中有关加速单纯形方法的内容介绍，在此不再赘述。目前，无论是从运行的可靠性以及运行时间考虑加速单纯形方法都是较优的，基本上能满足用户的要求。

在 DPS 计算机数据处理平台上求解非线性方程组时，应先将非线性方程组的各个方程按自然的数学表达方式编辑(注意，方程中不可出现等号，又应将等号转化为减号，公式中可使用本系统提供的标准函数)，一行写一个方程，然后将它们定义为公式块，再执行求解非线性方程的操作，这时系统会立即给出各个方程系数的初值。如有一非线性方程组

$$x_1^2 + x_2^2 + x_3^2 - 1 = 0$$

$$2x_1^2 + x_2^2 - 4x_3 = 0$$

$$3x_1^2 - 4x_2 + x_3^2 = 0$$

在求解非线性方程组的解时，在屏幕下部的文本编辑器中按如下方式编辑、定义公式块：

```
x1x1+x2x2+x3x3-1
2x1x1+x2x2-4x3
3x1x1-4x2+x3x3
```

然后进入菜单，选择“非线性方程求解”项，回车后系统会提示用户输入各个参数的初值。一般在第一次计算时不用去管它，只有在计算过程不收敛时，再重新分析，并在分析时输入初值。

计算后输出的结果有包括非线性方程组的总体目标函数(越小越好)和各个系数的数值解。本例运行结果为：各个待求参数估计值分别是 $x_1=0.78519686$ ， $x_2=0.49661123$ ， $x_3=0.36992273$ 。

下面再介绍一个在非线性方程组中含有数学函数的例子，其原方程组为

$$\frac{1}{2} \sin(x_1 x_2) - \frac{1}{4\pi} x_2 - \frac{1}{2} x_1 = 0$$

$$\left(1 - \frac{1}{4\pi}\right) (e^{2x_1} - e) + \frac{1}{\pi} x_2 e - 2x_1 e = 0$$

在 DPS 平台的文本编辑器中，求解方程参数估计值的用户界面如下：

```
0.5*sin(x1*x2)-1/4/3.14159265*x2-0.5x1
(1-1/4/3.14159265)*(exp(2*x1)-exp(1))+1/3.14159265*x2*exp(1)-2*x1*exp(1)
```

注意，自然数 e (自然对数之底)在公式中可定义成 $\exp(1)$ 。定义公式后进入菜单，选择“非线性方程求解”项，回车后系统会提示用户输入各个参数的初值。同样第一次计算时可以忽略，如发现计算过程不收敛，可按 ESC 退出迭代，进行重新分析时再输入初值。本例要求输入初值。在输入初值 $x_1=0.3$ ， $x_2=3$ 之后，系统很快输出结果：目标函数值 = 0.0000000000， $x_1=0.49999987$ ， $x_2=3.14159303$ 。

本例 x_1 的精确值为 0.5， x_2 的精确值是 π 。从分析结果可以看出，拟合误差均在 10^{-6} 以内，其精度是较高的。

值得注意的是，对有的非线性方程组，如果初值取得不合理，可能得到的是局部最优值，而不是全局最优解。

30.3 实系数多项式求根

30.3.1 方法简介

在本系统中, 采用牛顿下山法求实系数高次方程

$$f(x)=a_0x^{n-1}+a_1x^{n-2}+a_2x^{n-3}+\cdots+a_{n-2}x+a_{n-1}=0$$

的全部根。式中 $a_0 \neq 0$, $n < 100$ 。其方法的原理及步骤为:

(1) 将牛顿法和撒网法连用, 寻找下山路线, 求方程全部根, 若令 $f(x)=u(x)+iv(x)$, 则 $s(x)=[f(x)]^2=[u(x)]^2+[v(x)]^2$ 。如果 $x_0=w_0+iz_0$ 为方程的根, 那么 $|f(x_0)|^2=s(w_0, z_0)=0$ 。

牛顿法是将公式 $f(x)$ 按泰勒公式展开, 并取其第 1 项, 得

$$x_{i+1}=x_i - \frac{f(x_i)}{f'(x_i)}$$

根据 $|f(x)|$ 的下降情况来判断迭代是否成功, 即如有 $s(x_{i+1}) < s(x_i)$, 则认为迭代是成功的。实际迭代时, 再引入一个因子 t , 即

$$x_{i+1}=x_i - t \frac{f(x_i)}{f'(x_i)}$$

只要选取适当的 t , 可以保证有 $|f(x_{i+1})|^2 < |f(x_i)|^2$ 。

(2) 牛顿法公式在鞍点或接近重根点附近常常失败。本程序按下山原则, 采用撒网格点方法, 也就是选择适当的 D 与 C , 用以下公式:

$$\begin{aligned} w_{i+1} &= w_i + D \cdot \cos(C) \\ z_{i+1} &= z_i + D \cdot \sin(C) \end{aligned}$$

使 $s(w_{i+1}, z_{i+1}) < s(w_i, z_i)$ 而冲过鞍点, 或得到 $s^* \leq \varepsilon$ (ε 是给定精度), 即求得一个根。每求得一个根 x^* 之后, 在 $f(z)$ 中劈去一个因子 $(x-x^*)$, 再求另一个根。

(3) 当某个 $x^*=w^*+z^*$ 使 $s(x^*) \leq \varepsilon$, 则认为 x^* 是所求的全部根。

在实际计算时, 每求一个根都要作变换 $x = n \cdot \sqrt[n]{|a_{n-1}|} \cdot x$, 使得当 $a_n = 1$ 时 $|a_{n-1}| = 1$, 保证寻根在单位圆内进行。

30.3.2 方程求根的计算机处理

先将实系数多项式方程按自然的数学表达方式(依高次至低次幂排列顺序)进行编辑(注意方程中不可出现等号, 应将等号转化为减号), 然后将其定义成公式块, 再执行“求解实系数多项式根”的操作, 这时系统立即给出方程的各个根(包括实部和虚部)。为检验本功能模块的性能, 现列举不同形态的高次方程如下:

例 1 求方程 $x^3-3x-1=0$ 的全部根。在本计算机数据处理平台上, 只需在屏幕下部的文本编辑器中按如下方式编辑定义公式块即可求解:

$$x^3-3x-1$$

经计算机处理，输出结果为 $x_1=1.8793852$ ， $x_2=-1.5320889$ ， $x_3=-0.3472964$ 。

例 2 求方程 $x^4-4.87x^3+8.8937x^2-7.218497x+2.19702846=0$ 的全部根。

本例是一个具有稠密根的多项式，同时多项式的病态性也很严重。现列出以考验本计算功能块的性能，在平台上，只需按如下方式编辑定义公式块即可求解：

$$x^4-4.87x^3+8.8937x^2-7.218497x+2.19702846$$

经计算机处理，输出结果为： $x_1=1.23000171$ ， $x_2=1.21999328$ ， $x_3=1.21018415$ ， $x_4=1.20982086$ 。分别与该多项式各根的真值 1.23，1.22，1.21，1.21 相比，可以看出求解的精度是较高的。

例 3 求方程 $x^{10}-1=0$ 的全部根。

本例是一个含有复根的实系数多项式。求解本例，只需在计算机数据处理平台上按如下方式编辑定义公式块即可求解：

$$x^{10}-1$$

经计算机处理，输出结果如下：

$$x_1=0.309016994375+0.951056516295i$$

$$x_2=0.309016994375-0.951056516295i$$

$$x_3=-0.809016994375+0.587785252292i$$

$$x_4=-0.809016994375-0.587785252292i$$

$$x_5=0.809016994375+0.587785252292i$$

$$x_6=0.809016994375-0.587785252292i$$

$$x_7=-1.000000000000$$

$$x_8=1.000000000000$$

$$x_9=-0.309016994375+0.951056516295i$$

$$x_{10}=-0.309016994375-0.951056516295i$$

与该多项式各个根的精确值相比，误差在 10^{-8} 以内。

例 4 求方程

$$x^{16}+8x^{15}+48x^{14}+196x^{13}+664x^{12}+1800x^{11}+4198x^{10}+8208x^9+13992x^8+20228x^7+25480x^6+26904x^5+24385x^4+17688x^3+10584x^2+4320x+1296=0$$

的全部根。本例是一个病态性非常严重的例子，用它考验 DPS 处理平台的性能。在平台上，此例的求解仍只需按如下方式编辑定义公式块即可：

$$x^{16}+8x^{15}+48x^{14}+196x^{13}+664x^{12}+1800x^{11}+4198x^{10}+8208x^9+13992x^8+20228x^7+25480x^6+26904x^5+24385x^4+17688x^3+10584x^2+4320x+1296$$

经计算机处理，输出结果如下：

$x_1=-0.499809845220$	$1.658276931600i$
$x_2=-0.499809845220$	$-1.658276931600i$
$x_3=-0.500190127590$	$1.658348043887i$
$x_4=-0.500190127590$	$-1.658348043887i$
$x_5=-0.499964457405$	$1.658502443986i$
$x_6=-0.499964457405$	$-1.658502443986i$
$x_7=-0.500035569785$	$1.658122161238i$
$x_8=-0.500035569785$	$-1.658122161238i$
$x_9=-0.499841545249$	$1.322879853020i$
$x_{10}=-0.499841545249$	$-1.322879853020i$
$x_{11}=-0.500004274374$	$1.323034216836i$
$x_{12}=-0.500004274374$	$-1.323034216836i$
$x_{13}=-0.500158484181$	$1.322871274748i$
$x_{14}=-0.500158484181$	$-1.322871274748i$
$x_{15}=-0.499995696195$	$1.322717277526i$
$x_{16}=-0.499995696195$	$-1.322717277526i$

即本例多项式各个根的真值是 4 个 $-0.5 \pm i\sqrt{1.75}$ 和 4 个 $-0.5 \pm i\sqrt{2.75}$ 。由此可见，DPS 平台所提供的求解多项式工具的效果是相当令人满意的。

参 考 文 献

- 陈明逵, 凌永祥. 1992. 实用计算机数值方法[M]. 西安: 西安交通大学出版社.
 上海机械学院, 安徽省计算中心. 1984. FORTRAN 应用程序库[M]. 上海: 上海科学技术文献出版社.
 上海计算技术研究所. 1982. 电子计算机算法手册[M]. 上海: 上海教育出版社.
 郑维敏等. 1986. 系统工程 FORTRAN 程序集[M]. 北京: 清华大学出版社.
 William H P et al. 2004. C 数值算法(第 2 版)[M]. 傅祖芸等译. 北京: 电子工业出版社.

第 31 章 微积分数值计算

DPS 系统的微积分运算程序提供常用的微积分数值运算和分析技术, 包括定积分、多重积分、求导数以及求解微分方程组初值等计算方法。

在 DPS 数据处理系统上进行数值积分运算时, 只需求出被积函数在特定点的数值序列, 然后根据这些值计算积分的近似值。如果被积函数在形式上呈病态, 系统给出的结果就有可能不正确。这是应用中值得注意的。

本章所有的操作都在 DPS 数据处理系统下部文本编辑器中进行。

31.1 定 积 分

31.1.1 定积分基本原理和方法

DPS 系统采用龙贝格(Romberg)积分法计算定积分

$$I = \int_a^b f(x)dx$$

的近似值, 使其误差小于给定的允许误差。该方法先用梯形公式计算

$$T_1 = \frac{b-a}{2}[f(a) + f(b)]$$

然后采用将求积区间 $[a, b]$ 逐次折半的办法, 令区间长度

$$h = \frac{b-a}{2^i}, \quad i = 0, 1, 2, \dots$$

计算

$$T_{2n} = \frac{1}{2}T_n + \frac{h}{2} \sum_{i=1}^n f\left(a + h\left(k - \frac{1}{2}\right)\right), \quad n=2^i$$

由此得到辛普森(Simpson)求积公式

$$S_n = T_{2n} + (T_{2n} - T_n)/3$$

及柯特斯求积公式

$$C_n = S_{2n} + (S_{2n} - S_n)/15$$

再得到龙贝格求积公式

$$R_n = C_{2n} + (C_{2n} - C_n) / 63$$

直至相邻两次积分近似值 R_{2n} 和 R_n 满足：当 $|R_{2n}| \leq 1$ 时， $|R_{2n} - R_n| < \varepsilon$ ；当 $|R_{2n}| > 1$ 时， $|(R_{2n} - R_n) / R_{2n}| < \varepsilon$ ，其中 ε 为允许误差。

31.1.2 DPS 平台的操作示例

在 DPS 窗口下部的文本编辑器中，先将被积函数按自然的数学表达方式编辑，在方程后面，将积分的下限和上限放在方括号 “[]” 内，将其定义为公式块，然后执行“定积分”操作。这时系统会给出积分值和积分误差，如计算在 0~1 区间的余弦函数即

$\int_0^1 \cos(x) dx$ 的定积分，可按以下方式编辑、定义被积分的公式及其上下限

cos(x)
[0, 1]

然后进入菜单操作，选择执行“定积分”功能，得到的积分值为 0.84147098。

又如计算在 0~2 区间的正态分布函数即 $\int_0^2 \frac{2}{\sqrt{\pi}} e^{-x^2} dx$ 的定积分，其公式及积分区间的编辑定义如下：

2/sqrt(3.14159265)*exp(-x*x)
[0, 2]

在菜单下选择执行“定积分”功能后得到的积分值为 0.99532227。

31.2 多重积分

31.2.1 方法简介

DPS 系统采用高斯法计算多重积分

$$\int_{FL_1}^{FU_1} f_1(x_1) dx_1 \int_{FL_2(x_1)}^{FU_2(x_1)} f_2(x_1, x_2) dx_2 \cdots \int_{FL_n(x_1, \dots, x_{n-1})}^{FU_n(x_1, \dots, x_{n-1})} f_n(x_1, \dots, x_n) dx_n$$

的近似值，其中 n 为积分重数，为正整数。根据高斯求积公式，在闭区间 $[-1, 1]$ 上取 mp 个插值点，计算 $\int_{-1}^1 f(t) dt$ ，则有

$$\int_{-1}^1 f(t) dt \approx \sum_{k=1}^{mp} A_k f(t_k)$$

式中插值点 t_1, t_2, \dots, t_{mp} 为勒让德(Legendre)多项式 $x_{mp}(t)$ 的 mp 个根，称为高斯型点，而

$$A_k = \int_{-1}^1 L_k(x) dx$$

称为高斯系数，其中

$$L_k(x) = \prod_{j=1}^{mp} [(x-x_j)/(x-x_k)] \left[\prod_{j=1}^{mp} (x-x_j) \right]_{x=x_k}$$

当积分区间为 $[\alpha, \beta]$ 时，应进行

$$x_k = \frac{\beta - \alpha}{2} t_k + \frac{\beta + \alpha}{2} \quad (31.1)$$

变换，以求高斯型点 x_1, x_2, \dots, x_{mp} 。

在计算单积分 $\int_a^b f(x) dx$ 时，先将积分区间 $[a, b]$ 划为长度相等的 ks_1 个子区间，每个子区间上用 mp 个高斯公式来求积分的近似值，然后再求整个区间上的积分近似值。

计算 n 重积分时，分别将 $1, 2, \dots, n$ 层积分区间分为相等的 ks_1, ks_2, \dots, ks_n 个子区间。按式(31.1)先计算出各个积分区间上的第一个子区间中第一个高斯型点 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$ ，然后暂时固定 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{n-1}$ ，按高斯公式计算最内层的积分，再从内到外计算各层积分值，直到最后得到所要求的积分近似值。

31.2.2 DPS 平台的操作示例

在 DPS 平台上求多重积分，先在 DPS 窗口下部的文本编辑器中将被积函数按系统规定的格式对数学表达式进行编辑，即被积函数放在公式块的第一行，从第二行开始，一行一行地将积分函数各个变量的下限和上限写入(上下限两个表达式之间用逗号隔开)，然后将它们整个定义为公式块，在主菜单下选择执行“多重积分”功能项。计算积分时系统要求用户输入高斯-勒让德求积公式的结点数和各层积分区间的子区间数，用户可根据积分区域的形状、被积函数形态和精度要求选择这两个值，一般值越大精度越高。输入参数后，系统将给出积分值。

(1) 一般定积分。

此项操作功能既可处理多重积分，也可处理一般的定积分。例如，对振荡函数 $\int_0^1 x^2(1-x)\sin(64\pi x) dx$ 进行积分运算，我们可令 $x_1=x$ ，并编辑定义积分公式块如下：

$$\begin{array}{c} x1x1(1-x1)*\sin(64*3.14159265*x1) \\ 0, 1 \end{array}$$

然后进入菜单操作，选择并执行“多重积分”功能后系统将提示用户输入高斯-勒让德求积公式的结点数和各层积分区间的子区间数，本例中分别输入 15 和 8，然后可得到积分值为 -0.000000738173 (准确值约等于 0.0000007381790413)。

(2) 二重积分。

被积函数为 $\int_0^2 \int_{x^2}^{2x} xy^2 dy dx$ ，其积分真值等于 6.4。在本系统中，令 $x_1 = x$ ， $x_2 = y$ ，然后按系统规定格式定义公式块如下：

$$\begin{array}{c} x1*x2*x2 \\ 0, 2 \\ x1*x1, 2*x1 \end{array}$$

进入菜单操作，选择执行“多重积分”功能并按系统提示输入高斯-勒让德求积公式的结点数和各层积分区间的子区间数，本例中分别输入 8 和 8，即可获得该积分值为 6.4000000000。

(3) 三重积分。

如对被积函数

$$\int_{-1}^1 dx \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy \int_{-\sqrt{1-x^2-y^2}}^{\sqrt{1-x^2-y^2}} \frac{dz}{x^2 + y^2 + \left(z - \frac{1}{2}\right)^2}$$

进行积分，先令 $x_1 = x$ ， $x_2 = y$ ， $x_3 = z$ ，按 DPS 系统所规定格式定义公式块，即写成如下形式：

$$\begin{array}{c} 1/(x1x1+x2x2+sqr(x3-0.5)) \\ -1, 1 \\ -sqr(1-x1x1), sqrt(1-x1x1) \\ -sqr(1-x1x1-x2x2), sqrt(1-x1x1-x2x2) \end{array}$$

然后进入菜单，选择执行“多重积分”并按系统提示输入高斯-勒让德求积公式的结点数 8 和各层积分区间的子区间数 8，即可得到积分值为 11.45412(该积分的精确值为 11.46027)。

用户应当注意的是，随着被积函数的重数越高，积分计算所需时间越长。另外，高斯-勒让德求积公式的结点数最大可取 15，各层积分区间的子区间数最大可取 16。

31.3 数值微分

31.3.1 方法简介

按微积分的观点，函数的微分比积分问题容易，且大多数函数是可微的，求导数也比较容易。但在数值计算中，求微分时常受到病态问题的困扰。因此，DPS 系统采用 Borland Pascal ver. 7.0 中的双精度实数处理数值微分计算，这一般是可行的。

对于任意给定的函数 $f(x)$ ，计算在点 $x=a$ 处的微分值时，我们会自然想到使用公式

$$f'(a) \approx \frac{f(a+h) - f(a)}{h}$$

计算函数微分的近似值，式中 h 为足够小的值。

DPS 系统仍采用以上方法，但为了减少误差，实际采用下式计算微分值

$$f'(a) \approx \frac{[f(a+h) - f(a)] + [f(a-h) - f(a)]}{2h}$$

式中， h 的取值为 a 的 $\frac{1}{10000000}$ ，经对各种函数的数值试验，一般可满足实用的要求。

31.3.2 DPS 平台的操作示例

在 DPS 平台上进行数值微分，先在文本编辑器中将函数按系统规定的格式编辑，如公式中有多个变量，则分别以 x_1, x_2, \dots 表示，在方程后面，将待求导数的变量值放在第二行的方括号“[]”内，然后将它们整个定义为公式块，再执行“求数值微分”的操作，系统将给出各个变量的微分值。

例如，求函数 $f(x) = \sin x$ 当 $x = \frac{\pi}{4}$ 时的微分值。编辑定义公式时，需将 $x = \frac{\pi}{4}$ 化成

$3.14159265/4=0.78539816$ ，然后按如下方式编辑定义计算公式：

$$\begin{array}{c} \sin(x1) \\ [0.78539816] \end{array}$$

进入菜单操作，在菜单方式下选择“数值微分”，执行计算后系统将立即输出计算结果，即导数值等于 0.70710678，在本例中， $\sin x$ 的导数是 $\cos x$ ，而 $\cos(0.78539816) = 0.70710678$ ，可见微分计算的精度达 8 位有效数值。

当函数中有多个变量时，用 x_1, x_2, \dots 分别表示各个变量。如求函数

$$f(x) = \frac{1}{x^2 + y^2 + (z - 0.5)^2}$$

在点 $x=1, y=1, z=0.5$ 处各个变量的偏导数值。可令 $x_1=x, x_2=y, x_3=z$ ，并定义公式块如下：

$$\begin{array}{c} 1/(x1x1+x2x2+sqr(x3-0.5)) \\ [1, 1, 0.5] \end{array}$$

然后进入菜单操作，选择“数值微分”，执行计算功能后系统将立即输出各变量的导数值 $x_1 = -0.5, x_2 = -0.5, x_3 = 0.0$ 。

31.4 微分方程(组)初值求解

微分方程(组)是描述生物(种群)发生发展过程的一个重要工具。在数学生态学中无论是单种种群还是复合种群，其数学模型的定义绝大多数来源于微分方程。如单种种群的 Logistic 方程、Malthus 方程、描述捕食者-猎物关系的 Lotka-Volterra 方程等。因此在生态学模拟模型的研究中，为模拟种群的发生发展动态，求解微分方程(组)初值是必不可

少的模拟手段。

31.4.1 微分方程模型的参数求解原理及步骤

DPS 系统采用特雷纳(Treanor)方法求微分方程的数值解。基本方法如下所述：设有 n 个微分方程的方程组为

$$y_i = f_i(t, y_i) \quad (i=1, 2, \dots, n)$$

$$y(t_0) = y_0$$

设已知时间 t_j 点处的函数值 $y_{i,j} (i=1, 2, \dots, n)$ ，则计算 $t_{j+1} = t_j + h$ (h 为 t 的增量) 点的未知函数 $y_{i,j+1}$ 的计算公式为

$$y_{i,j+1} = y_{i,j} + \Delta y, \quad i=1, 2, \dots, n$$

式中，增量函数 Δy 的定义是，当 $P \leq 0$ 时，

$$\Delta y = h[f_1 + 2(f_2 + f_3) + f_4]/6$$

否则

$$\begin{aligned} \Delta y = h\{ & f_1 F_1 + [-3(f_1 + y_1 P) + 2(f_2 + y_2 P) + 2(f_3 + y_3 P) - (f_4 + y_4 P)] F_2 \\ & + 4[(f_1 + y_1 P) - (f_2 + y_2 P) - (f_3 + y_3 P) + (f_4 + y_4 P)] F_3 \} \end{aligned}$$

并且

$$P = (f_3 - f_2) / (y_3 - y_2)$$

$$F_0 = e^{-Ph}, \quad F_1 = (F_0 - 1) / (-Ph)$$

$$F_2 = (F_1 - 1) / (-Ph), \quad F_3 = (F_2 - 0.5) / (-Ph)$$

$$t_1 = t_0, \quad y_1 = y(t_1), \quad f_1 = f_i(t_1, y_1)$$

$$t_2 = t_1 + h/2, \quad y_2 = y_1 + (h/2)f_1, \quad f_2 = f_i(t_2, y_2)$$

$$t_3 = t_1 + h/2, \quad y_3 = y_1 + (h/2)f_2, \quad f_3 = f_i(t_3, y_3)$$

$$t_4 = t_1 + h, \quad y_4 = y_1 + hf_1, \quad f_4 = f_i(t_4, y_4)$$

参数 q 又由公式

$$q = f_3, \quad P \leq 0$$

$$q = 2(f_3 - f_1)F_2 + (f_1 - f_2)F_1 + f_2, \quad P > 0$$

所确定。

31.4.2 DPS 平台操作示例

在 DPS 平台上求解微分方程模型的参数,先在文本编辑器中将微分方程组按自然的数学表达方式编辑,每个公式放一行,然后在公式下面的第一行放微分方程中各个变量的初始值,在公式下面的第二行放积分区间、步长及每间隔多少步输出一次积分结果,然后将它们定义为公式块,再执行“求解微分方程初值”的操作,系统将给出各步各个变量的积分值。以 Lotka-Volterra(1926, 1935)模型为例介绍如下:

该模型微分方程形式为

$$\begin{cases} \frac{dN}{dt} = (r_1 - \alpha P)N \\ \frac{dP}{dt} = (-r_2 + \beta N)P \end{cases}$$

该模型在一定的假设(条件)下描述了捕食者和猎物之间的互动关系,其结果无论是捕食者种群还是猎物种群都作周期性振荡(图 31-1(a)),即捕食者-猎物种群作用系统的行为动态是随时间序列的推移,猎物密度逐渐增加,捕食者的密度也追随它而增加,只是在时间上落后一步。由于捕食者密度增加,必然降低猎物的密度,而猎物密度的降低,又造成捕食者密度因食物减少而种群数量降低,这样又形成了猎物增加的条件。如此往复循环,形成了一定周期(图 31-1(b))。

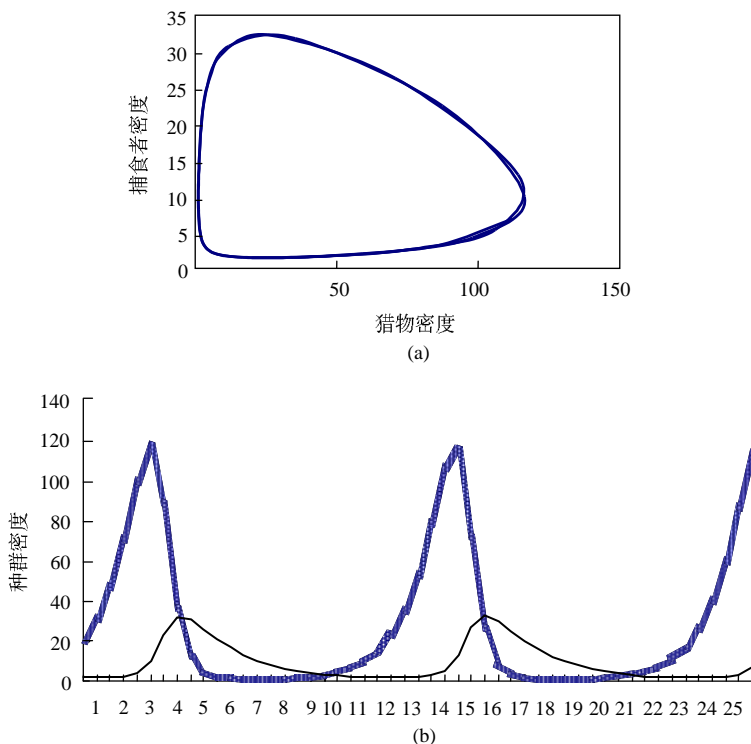


图 31-1 模拟曲线图

上述方程中, 若 $r_1=1.0$, $r_2=0.5$, $\alpha=0.1$, $\beta=0.02$, 则有 Lotka-Volterra 模型

$$\begin{cases} \frac{dN}{dt} = (1 - 0.1P)N \\ \frac{dP}{dt} = (-0.5 + 0.02N)P \end{cases} \quad (31.2)$$

模拟该模型所描述系统的行为, 就是在一定初始值的条件下求上述微分方程组的初值。在 DPS 平台上求式(31.2)微分方程组的初值, 首先令 $x_1=N$, $x_2=P$, 取种群初值 $N_0=20$, $P_0=1.5$, 按系统规定格式编辑定义公式及其参数块如下:

```
(1-0.1*x2)x1
(-0.5+0.02x1)x2
[20, 1.5]
[0, 25, 0.01, 50]
```

以上公式块的第一、二行是微分方程公式, 第三行是 N 和 P 的初值 $N_0=20$, $P_0=1.5$, 公式块的第四行中的 4 个数值分别是积分区间, 即微分方程组(31.2)中时间 t 的起始值(0)和终止值(25), 积分步长(0.01)和要求多少步输出积分结果一次(每 50 步输出积分结果一次)。然后进入主菜单, 选择“微分方程组数值计算”功能项, 执行后系统即输出计算结果。将计算结果作成 x - y 折线图便可得到图 31-1(a), 即 Lotka-Volterra 模型描述的捕食者和猎物种群动态相位图。将两个种群模拟值作折线图便可观察到两个种群数量消长过程中往复循环、形成周期的现象(图 31-1(b))。

参 考 文 献

- 陈明达, 凌永祥. 1992. 实用计算机数值方法[M]. 西安: 西安交通大学出版社.
 丁岩钦. 1994. 昆虫数学生态学[M]. 北京: 科学出版社.
 李淑霞. 1991. 生物学中常用数学方法[M]. 青岛: 青岛海洋大学出版社.
 上海机械学院, 安徽省计算中心. 1984. FORTRAN 应用程序库[M]. 上海: 上海科学技术文献出版社.
 上海计算技术研究所. 1982. 电子计算机算法手册[M]. 上海: 上海教育出版社.
 邬祥光. 1985. 昆虫生态学的常用数学分析方法[M]. 北京: 农业出版社.
 郑维敏等. 1986. 系统工程 FORTRAN 程序集[M]. 北京: 清华大学出版社.

第七篇 专业试验统计

与第二篇介绍的统计分析方法相比,本篇介绍一系列应用于各个学科、专业领域所特有的统计分析技术。如对于从事遗传、育种的工作者,除应用一般的方差分析方法之外,还需要考虑更多因素、更为复杂的分析技术,如品种区域试验的统计分析,分析品种效应、品种 \times 地点的互作效应、品种 \times 年份的互作效应。近年来,一种新的作物品种区域化试验统计分析技术——加性主效应乘积交互作用(AMMI)模型,已开始被用于多点区域试验资料的分析。该模型与方差分析模型、线性回归模型相比,应用范围更广、更为有效。这是因为该模型将方差分析和主成分分析有机地结合在一起。第 21 章将详细介绍这一新的分析技术。

数据包络分析(data envelopment analysis, DEA)可以用于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。DEA 作为一种有效性评价方法,20 多年来一直为广大理论和应用工作者所瞩目,不仅理论上得到了充实和完善,而且已广泛应用于各个领域。

生态学工作者除应用方差分析技术之外,重点是考察动植物种群数量的抽样精度,研究动植物种群数量的空间分布规律以及生物群落的消长演替过程。因此,本篇中对空间格局分析、抽样设计等专题的讨论将是十分有益的。从事群落生态学研究的工作者经常面对一大堆群落统计分析的问题,本书第 19 章的“数学生态学方法”,将在资料处理上提供极大的方便。

对从事“生物测定”研究的科技人员来说,作者所提供的统计分析技术将使读者喜出望外。第 14 章不仅介绍如何非常方便地使用经典的 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。

在目前的 DPS 数据处理系统中,对原来的“遗传育种”和“生存分析”两章的内容进行了扩充,使之方法更为完善,提供的功能更为系统。同时,在专业统计部分增加了更多的数据处理技术,如心理学统计方面的量表分析、数量经济学中的顾客满意指数模型、实验诊断评价工作中的 ROC 曲线分析以及地理、地质统计分析方面的数据处理技术。这将对各个专业领域的技术人员处理科研数据有所帮助。

第 32 章 数据包络分析和随机前沿面分析

生产效率评价与分析的主要目的是根据实际观测数据,使用有关的效率测度方法,并借助有效的计算技术进行评估、决策。生产效率通常用一些前沿面的形式来描述,在过去 50 多年里,许多不同方法被应用在前沿面的估计上,其中两个主要的方法是数据包络分析(data envelopment analysis, DEA)和随机前沿面分析(stochastic frontier analysis, SFA)。

数据包络分析是美国著名运筹学家得克萨斯大学教授查恩斯(Charnes)、库伯(Cooper)和罗兹(Rhodes)于 1978 年在权威的《欧洲运筹学杂志》上发表的一篇重要论文: *Measuring the efficiency of decision making units*(决策单元的有效性度量)中提出的评价生产效率的重要的非参数方法。该方法的原理主要是通过保持决策单元(decision making units, DMU)的输入或者输出不变,借助于数学规划方法确定相对有效的生产前沿面,将各个决策单元投影到 DEA 的生产前沿面上,并通过比较决策单元偏离 DEA 前沿面的程度来评价它们的相对有效性。

DEA 方法是以相对效率概念为基础,以凸分析和线性规划为工具的一种评价方法,应用数学规划模型计算比较决策单元之间的相对效率,对评价对象做出评价,能充分考虑对于决策单元本身最优的投入产出方案,因而能够更理想地反映评价对象自身的信息和特点;同时,对于评价复杂系统的多投入多产出分析也有独到之处,适用于多输出-多输入的有效性综合评价问题,在处理多输出-多输入的有效性评价方面具有绝对优势。

例如,评价一个国家不同城市的经济实力时,输入的指标可以是:固定资产投资额、劳动力数、城市占地面积、科技人员比例、工业基础等指标;输出的指标可以是:国民生产总值、国民收入、人均收入、居民消费、居民人均住房面积、上交国税等指标。这些问题可以表达为: n 个单位(企业)为被评价对象,称为决策单元;输入指标是这些单位“资源”的投入,如抽取 n 家企业(单位)的 m 项输入指标样本,得到资源的投入表如下:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

各个单位(企业)的产出(输出)指标是该企业单位投入了资源后的“成效”或“回报”,同理,由 n 家企业 s 项输出指标的样本,可得到资源的产出表如下:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{s1} & y_{s2} & \cdots & y_{sn} \end{bmatrix}$$

注意, 这里资源的投入和资源的产出是广义的。输入指标和输出指标组成一个指标体系。有了投入表和产出表, 就可以采用 DEA 对企业效率进行有效性评价。

很明显, 一般的综合评价问题是将这里的输入输出指标都作为输入指标, 即评价指标(有时将输入指标作为低优指标进行预处理), 但它体现不出相对有效性。DEA 正是提供了多输入-多输出指标的有效性综合评价方法。

在 DEA 分析中, 决策单元的最优效率指标与投入指标值及产出指标值的量纲选取无关, 应用 DEA 方法建立模型前也无需对数据进行无量纲化处理。

生产效率评价的参数方法主要有随机前沿面分析(SFA)技术。随机前沿生产函数分别由 Aigner, Lovell, Schmidt, Meeusen 和 van den Broeck 于 1977 年独立提出。该生产函数对包含由两个成分组成的误差进行描述, 可应用于面板数据, 估计随时间变化和不变的效率、成本和产量函数等。

实际工作中, 一般多应用数据包络分析方法, 以计算生产效率为目的, 并考虑到成本效率和分配效率, 以及应用面板数据来计算总的因素生产率(TFP)变化、技术变化、技术效率变化和规模效率变化的指数, 即由 Färe et al(1994)提出的 Malmquist 指数。

32.1 生产效率分析基本原理

现代效率的测度由 Farrell(1957)首创, 他在 Debreu 和 Koopmans 方法的基础上定义了一个单位化(取值在 0~1 之间)的能考虑到多种投入的效率测度方法, 该单位效率由两部分组成, 即技术效率和分配效率, 前者反映了一个单位在给定一系列的投入后获得最大产出量的能力。后者反映了一个单位以最佳的比例投入后所获得可观价值的的能力。这两个方法都与提供一个总的经济效率测度方法有关。

根据 Farrell 的基本观点, 采用投入/产出加以说明, 并以减少投入为中心, 这在计量经济学中通常称为面向投入的方法。Farrell 使用了一个简单的例子, 即各个单位以两项投入(x_1 和 x_2)得到单一产出(y), 其前提是具有固定规模收益, 即具不变的规模报酬(图 32-1)。

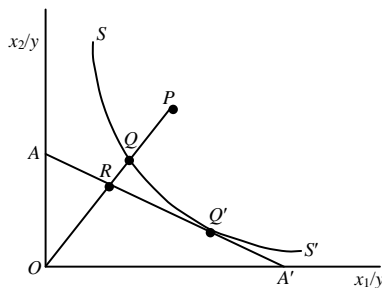


图 32-1 技术效率和分配效率示意图

图 32-1 中 SS' 表示完全有效的公司(或单位)的产出等高线, 代表生产前沿面。点 P 表示一个公司用相应数量投入来生产一单位产出, 技术非有效可以用距离 QP 来表示, 表示在不减少产出的前提下所有的投入可以成比例缩减的数量, 通常用百分比率 QP/OP

表示。故一个公司的技术效率通常用下述比率表示

$$TE=OQ/OP=1-QP/OP$$

其取值范围在 0 和 1 之间。如果值等于 1，则表示这个公司是完全技术有效率的，图 32-1 中点 Q 就是技术有效的点。技术有效的所有点就构成了生产前沿面，如图 32-1 中曲线 SS'上的点。

如果已经知道了投入的价格信息，如图 32-1 中直线 AA' 所示，则可计算分配效率 (allocative efficiency, AE)和经济效率(economic efficiency, EE)。在点 P 的分配效率(AE)定义为

$$AE = OR/OQ$$

而经济效率定义为

$$EE = OR/OP$$

因为有

$$(OQ/OP) \times (OR/OQ) = OR/OP$$

所以

$$EE=TE \times AE$$

图 32-1 中定义的这些效率方法是假定完全技术有效单位的生产函数已知。但实际情况下并非如此，因此，完全有效的等值曲线必须从样本数据中估算。Farrell 建议使用(a)建立一个非参数的分段线性凸面等值曲线，使不可观测的点位于它的左侧或者下方(图 32-2)；或者(b)建立一个参数函数，如 Cobb-Douglas 形式的模型拟合数据，同样使得不可观测的点应该位于它的左侧或者下方。

以上是面向投入的技术效率分析方法。这就提出了一个疑问：在没有改变产出量的前提下有多少投入量能够适当地减少？或者在投入量不变的情况下如何使产出量以适当比例增加？这是一个与上面讨论的面向投入的方法相对应的面向产出的方法。

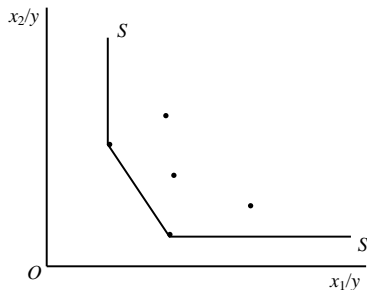


图 32-2 分段线性凸面等值曲线

面向产出和面向投入方法之间的不同可以用一个简单的关于单一投入和单一产出的

例子来阐述(图 32-3)。

在图 32-3(a)中描述了用 $f(x)$ 表示规模技术收益的减少, 点 P 表示一个非效率单位。Farrell 的技术效率, 如果是面向投入的方法, 那么其比值为 AB/AP , 但如果采用面向产出的方法, 则等于 CP/CD 。当固定规模收益存在时, 面向产出和面向投入方法得到的技术效率是一样的。但当规模收益出现增加或降低时它们将会不相等(Färe, Lovell, 1978)。固定规模收益情况可用图 32-3(b)中来描述, 对于我们谨慎选择的某一非效率点 P , 可以看出 $AB/AP=CP/CD$ 。

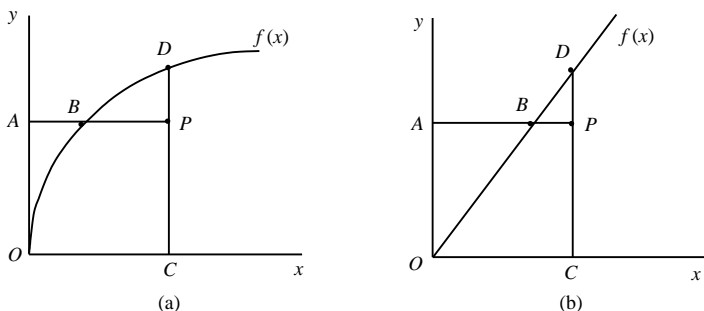


图 32-3 面向投入、产出的技术效率方法和规模收益

要深入理解面向产出的方法, 可以考虑包括两项产出(y_1 和 y_2)和一个投入(x_1)的情况, 并且假设有固定的规模收益。这时可用在二维空间中一个单位产出的可能性曲线来表示(如图 32-4)。

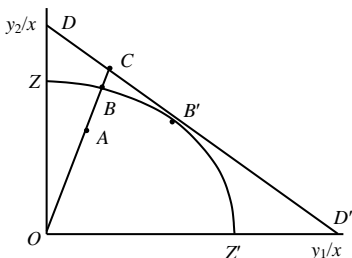


图 32-4 产出定向的技术和分配效率

图 32-4 中, ZZ' 是单位产出曲线, 点 A 则对应一个非有效单位。需要注意的是, 在这种情况下非有效单位点(A)位于曲线之下, 因为 ZZ' 代表产出可能性以上的区域。

Farrell 面向产出的效率方法如图 32-4 中的定义: AB 的距离代表技术非有效, 即在无需额外投入的条件下产出所增加的量。因此, 面向产出的技术效率即比值 $TE_o = OA/OB$ 。如果已知价格信息, 那么就可绘出收益等高线 DD' , 并且可定义分配效率为 $AE_o = OB/OC$, 且经济效益可以定义为

$$EE_o = (OA/OC) = (OA/OB) - (OB/OC) = TE_o \times AE_o$$

这些效率的取值都在 0~1 之间。各种效率, 从原点到可观测点之间的线都是规格化的, 也就是说, 即使改变测度单位(例如在测量劳动者数量时用年人数来取代小时人数)也不

对效率测量值有影响。

32.2 数据包络分析 CCR 模型

前沿面估算的分段线性凸面体法是由 Farrell(1957)提出的,但在提出后的近 20 年里,认可 Farrell 的方法的人并不多。如 Boles 和 Afriat(1972)提出的数学规划方法虽能完成估算,但是这种方法并没有得到广泛的关注。直到 Charnes, Cooper 和 Rhodes(1978)提出了数据包络分析(DEA)技术,即 CCR 模型之后,情形才得到了根本的改变。

Charnes, Cooper 和 Rhodes 提出的应用于前沿面估算的非参数数学规划的数据包络分析方法是 DEA 模型类中最基本、最重要的技术。随后有大量的论文,考虑到了其他一系列假设,对 DEA 方法进行扩展。如 Banker, Charnes 和 Cooper (1984)提出的可变规模收益(VRS)模式下数据包络分析技术,即 BCC 模型。但 DEA 中的面向投入的 CCR 模型的提出最早,且应用最广泛。

32.2.1 CCR 模型

若有 N 个公司或单位(决策单元),每一公司单位或决策单位都有 M 项投入和 S 项产出。分别用向量 x_{ij} 和 y_{rj} 表示决策单元的投入和产出,记

x_{ij} : 为第 j 决策单元 i 项输入指标的投入量, $x_{ij} > 0, i=1,2,L, M; j=1,2,L, N$;

v_i : 为第 i 项输入指标的权系数, $i=1,2,L, M$;

y_{rj} : 为第 j 决策单元 r 项输出指标的产出量, $y_{rj} > 0, r=1,2,L, S; j=1,2,L, N$;

u_r : 为第 r 项输出指标的权系数, $r=1,2,L, S$;

x_{ij}, y_{rj} 为样本已知数据, v_i, u_r 为待求的未知参数,那么可以得到 DEA 投入和产出信息表(表 32-1)。

表 32-1 DEA 投入产出信息表

决策单元→		1	2	...	j	...	n		
v_1	1 →	x_{11}	x_{12}	...	x_{1j}		x_{1n}	...	
v_2	2 →	x_{21}	x_{22}	...	x_{2j}		x_{2n}	...	
⋮	⋮	⋮	⋮	⋮	⋮		⋮		
v_m	m →	x_{m1}	x_{m2}	...	x_{mj}		x_{mn}	...	
		↓							
		y_{11}	y_{12}	...	y_{1j}		y_{1n}	→ 1 u_1	
		y_{21}	y_{22}	...	y_{2j}		y_{2n}	→ 2 u_2	
		⋮	⋮	⋮	⋮		⋮	⋮	
		y_{s1}	y_{s2}	...	y_{sj}		y_{sn}	→ s u_s	

CCR 模型建立在各决策单元相互比较的基础上, 具有相对有效性, 且各决策单元的效率评价指数依赖于它的输出综合与输入综合之比 Z_j , 即

$$Z_j = \frac{\sum_{r=1}^S u_r y_{rj}}{\sum_{i=1}^M v_i x_{ij}}, \quad j=1, 2, L, M$$

式中分子是第 j 个决策单元输出的总和, 分母是输入的总和, 效率评价指数 Z_j 即是相对有效性评价, 等价于我们常说的“综合评价”值。

对第 $j_0 (1 \leq j_0 \leq M)$ 个决策单元进行有效性评价。评价模型是: 以第 j_0 个决策单元的有效评价值为目标函数并且求最大值(使有效评价值最优), 以所有决策单元的有效评价值(包括第 j_0 个决策单元)小于等于 1 为约束。为了书写简便, 不妨将下标 j_0 记为 0, 将 y_{rj_0} 记为 y_{r0} , x_{ij_0} 记为 x_{i0} 。得到第 j_0 个决策单元的相对有效评价模型

$$\begin{aligned} \max Z_0 &= \frac{\sum_{r=1}^S u_r y_{r0}}{\sum_{i=1}^M v_i x_{i0}} \\ \text{s.t.} \quad &\frac{\sum_{r=1}^S u_r y_{rj}}{\sum_{i=1}^M v_i x_{ij}} \leq 1, \quad j=1, 2, L, M \\ &v_i, u_r \geq 0, \quad i=1, 2, L, M; \quad r=1, 2, L, S \end{aligned}$$

如令

$$\begin{aligned} x_j &= (x_{1j}, x_{2j}, L, x_{Mj})^T, \quad j=1, 2, L, N \\ y_j &= (y_{1j}, y_{2j}, L, y_{Sj})^T, \quad j=1, 2, L, N \\ v' &= (v_1, v_2, L, v_M)^T \\ u' &= (u_1, u_2, L, u_S)^T \end{aligned}$$

则用矩阵形式更为简洁:

$$\begin{aligned} \max_{u, v} \quad &(u'y_i/v'x_i) \\ \text{s.t.} \quad &u'y_j/v'x_j \leq 1, \quad j=1, 2, L, M \\ &u, v \geq 0 \end{aligned}$$

找出 u 和 v 的值, 使得第 i 个决策单元的效率测度值达到最大, 其受限于所有的效率测度值都小于或等于 1。求这个特殊的比值公式的一个问题是在于它有无穷多个解。为了避免这个问题, 可以限定 $v'x_i = 1$, 这时有

$$\begin{aligned}
 & \max_{\mu, \nu} \quad (\mu' y_i) \\
 & \text{s.t.} \quad \nu' x_i = 1 \\
 & \quad \quad \mu' y_j - \nu' x_j \leq 0, \quad j = 1, 2, L, N \\
 & \quad \quad \mu, \nu \geq 0
 \end{aligned}$$

其中从 u 和 v 到 μ 和 ν 的符号变化反映了它的转变。这种形式通常称为线性规划问题的乘数形式。利用线性规划中对偶性质，可得到这个问题的一个相同的包络形式：

$$\begin{aligned}
 & \min_{\theta, \lambda} \quad \theta \\
 & \text{s.t.} \quad -y_i + Y\lambda \geq 0 \\
 & \quad \quad \theta x_i - X\lambda \geq 0 \\
 & \quad \quad \lambda \geq 0
 \end{aligned} \tag{32.1}$$

其中 θ 是一个标量而 λ 是一个 $N \times 1$ 常数向量。这种包络形式要比乘数形式的少许多约束条件 ($M \times S < N \times 1$)，所以它通常是首选的解题形式。获得的 θ 值就是第 i 个决策单元的效率值。根据 Farrell (1957) 的定义，它满足 $\theta \leq 1$ ，当取值为 1 时表示该点在前沿面上，也就是说该决策单元是技术有效的。

32.2.2 松弛变量

DEA 中的非参数前沿面分段线性形式会导致在效率测度中出现一些问题，如分段线性前沿面部分地平行于轴线(图 32-2)，而在大部分参数函数中并非如此(见 32-1)。这种现象可参见图 32-5，其中决策单元 C 和 D 是有效的，并且由它们定义了前沿面，但决策单元 A 和 B 是非有效的。

Farrell(1957)的技术效率测度分别赋予决策单元 A 和 B 为 OA'/OA 和 OB'/OB 。然而，问题是点 A' 是否为一个有效率点，因为可以减少已投入量 x_2 (通过 CA')而仍然得到相同的产出，在文献中称为投入的松弛变量。当考虑包括更多投入和多种产出的情况时，图表就不再是这样简单，但松弛变量也可能出现。这样，在某一项 DEA 分析中，为提供精准的技术效率信息，同时列出的 Farrell 的技术效率测度(θ)和其他非零的投入或产出松弛变量就会有冲突的地方，这是因为对第 i 个决策单元而言，仅当 $Y\lambda - y_i = 0$ 时，其产出的松弛变量为 0；以及仅当 $\theta x_i - X\lambda = 0$ (给 θ 和 λ 赋适当的值)时，投入松弛变量为 0。

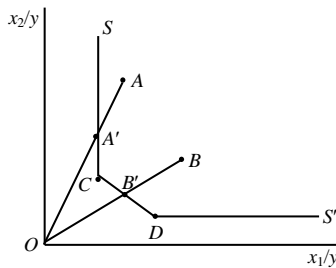


图 32-5 效率测度和投入的松弛变量

在图 32-5 中，与点 A' 相关的投入松弛变量是投入 x_2 的 CA' 。当有比这个简单例子更

多的投入和产出时, 视作与“最近”的有效前沿面点(例如 C)是相同的, 这样, 接下来的松弛变量计算就不再是一个繁琐的工作了。一些作者(Ali, Seiford, 1993)提出第二级的线性规划用于解决该问题, 通过增大所需的总松弛变量来移动一个有效前沿面点, 把一个非效率前沿面点(例如图 32-5 中的 A)移至一个有效率前沿面点(C 点)上。这个第二级线性规划可以通过以下定义实现:

$$\begin{aligned} \max_{\lambda, s^-, s^+} \quad & w = es^- + es^+ \\ \text{s.t.} \quad & s^- = \theta x_i - X\lambda \\ & s^+ = Y\lambda - y_i \\ & \lambda \geq 0, s^- \geq 0, s^+ \geq 0 \end{aligned} \quad (32.2)$$

其中 s^+ 是产出的松弛变量, s^- 是投入的松弛变量。

松弛变量也许能被视作所选择前沿面构造方法(DEA)和供有限样本量使用的人为制造的变量。如果有一个无限样本量和/或使用一个交替前沿面构造方法, 其中包含了一个平滑函数表面, 松弛变量问题就不存在了。另外, Ferrier 和 Lovell 提出的松弛变量本质上可看作分配非有效这一说法是相当合理的。这样使我们相信将技术效率分析的重心放在第一级 DEA 线性规划提供的径向效率大小上是合理的。

32.2.3 面向产出的 DEA 模型

上述面向输入的技术有效性测度说明了这样一个问题: 在不改变产出量的情况下可以按比例地减少多少投入量? 也可以换一个方式提出这个问题: 在不改变投入量的情况下产出量可以按比例地增加多少? 第二种提问方式给出了面向输出测算与上述所讨论的面向输入测算的不同之处。我们可以用一个简单的单输入(x)单输出(y)例子来解释面向输入与面向输出测度的区别。如图 32-3 所示, 一个规模收益递减的技术条件以 $f(x)$ 表示, 图中位于 P 点的企业是非有效的。Farell 面向输入的技术效率 TE 等于 AB/AP , 而面向输出的 TE 等于 CP/CD 。只有在规模收益不变的条件两种方法对技术有效性的测算是相等的。图 32-3(b)描绘了规模收益不变(CRS)的情况, 从图中可看出对处于 P 点的非有效企业有 $AB/AP = CP/CD$ 。

我们可以就涉及两种产出(y_1 和 y_2)和一种投入(x)的生产情况来阐述面向输出的测算。如果把投入固定在某一特定水平, 可以用二维生产可能曲线代表技术。图 32-4 描述了该例的情况, 图中 ZZ' 代表生产可能曲线, 点 A 代表非有效企业。因为 ZZ' 表示生产可能集的上边界, 点 A 位于曲线的下方, 所以处于 A 点的企业是非有效的。

Farell 面向输出效率测算的解释如图 32-4, 间距 AB 代表非技术有效, 即在没有额外投入的情况下产出可能增加的量。因而, 面向输出的技术效率可用下述比率测算:

$$TE_o = OA/OB$$

如果知道有关的价格信息, 则可以画出等收入曲线 DD' 并且定义分配效率如下:

$$AE_o = OB/OC$$

与面向输入情况中分配无效成本递减的解释相似，面向输出测算具有收入递增的特性。进而，定义总经济效率为这两种测度的乘积

$$EE_o = OA/OC = (OA/OB) \times (OB/OC) = TE_o \times AE_o$$

注意，所有 3 个测算指标的取值范围均为 0~1，同时也可发现面向输出的技术效率恰好等于输出间距函数，并且输出分配有效性概念与收入最大化一致的。如上定义的效率测算，首先，技术效率是建立在投入(或产出)的相对比例不变的基础上。单位无关性是通过影射进行效率测算方法的一个优点，即改变测算指标的单位不会改变效率测算的数值(如劳动力既可以人·小时为单位也可以人·年为单位)。其次，我们从成本最小化和收入最大化的角度对分配效率进行了探讨，但没有从利润最大化的角度进行研究(此处要同时假设成本最小化和收入最大化)。利润最大化问题将在后面的 SBM 模型实现。

与面向投入的分析模型比较，面向产出的分析模型在 CCR 中这两种测度方法提供了同样的测度值，但在后面介绍的可变规模收益(VRS)模式下，不同方法的测度值就不相等了(图 32-3)。

DEA 并非统计方法，故选择何种定向并不十分重要。一般来说，分析者倾向于选择面向投入的模式。这是因为投入量是主要的决定变量，尽管这种观点在所有的产业中并非无懈可击。在一些产业中，决策单元被给予了一定量的资源并要求它们尽其所能的获得产出，在这种情况下，使用产出定向可能更为恰当。本质上来说，一个 DUM 所选择的定向应该依管理者们能够掌控的量是否是投入或产出来定。但有时，对于定向的选择是出于使所获得的量受到较小的影响。

32.2.4 超效率分析

在实践中，有时会出现多个决策单元的技术效率值 θ 都等于 1 的情况，从而不能按效率值高低对各个决策单元进行排序，对于这种情况，DEA 提供了一种超效率模型，按此模型计算出来的效率值 θ 可大于 1，基本上能实现对所有决策单元的排序。 θ 值大于 1 的经济意义在实践中十分有用，它可以测算出各项投入指标在同时按多大比例增加情况下决策单元仍能保持 DEA 有效。

超效率含义可从图 32-6 看出，这里有 5 个决策单元(A, B, C, D 和 E)，每个单元有两项投入、一个产出，应用 5 个决策单元数据得到的标准的 DEA 模型，B, C 和 D 这 3 个决策单元形成了前沿面，因而这 3 个决策单元的效率得分值都是 1。不过，我们应用超效率的 DEA 模型求解，这 3 个决策单元的超效率得分值有可能大于 1，如对于决策单元 C 即是这样。对 C 来说，测定其超效率值，它不再是处于前沿面上。前沿面仅包含 B 和 D 两个决策单元，而 C 的投影点在 C' 处。决策单元 C 的超效率得分值等于 OC'/OC ，大致等于 1.2。这表明该单位的各项投入指标同步增加 20%，决策单元仍能保持 DEA 有效。

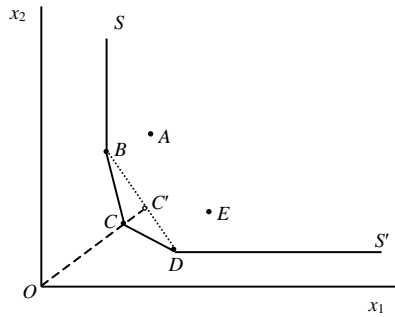


图 32-6 超效率图解

在超效率模型分析中，原来没有处于前沿面上的决策单元的效率得分值(如 A 和 E)不会发生变化，因为它们不是构成前沿面的单位。

32.2.5 CCR 模型应用示例

CCR 模型是最基本的 DEA 模型。我们将使用一个简单的例子来说明 CCR 的面向投入的 DEA 问题，这个例子仅涉及由两项投入获得一项单一产出的决策单元的 5 份报表。分析时在 DPS 中数据编辑格式如图 32-7。注意，数据块的左边是 2 列投入指标，右边是 1 列产出指标。



图 32-7 数据包络分析 CCR 模型分析数据格式及用户界面

按如图 32-7 左边所示编辑、选择数据，执行专业统计→数据包络分析及计量经济统计→CCR 模型，系统弹出如图 32-7 右边所示用户对话框。输入相应参数后即可得到计算结果如下：

输入 DEA_CCR 模型					
DMU _i	x1 权重	x2 权重	y1 权重		
DMU1	0.5000	0.0000	0.5000		
DMU2	0.1000	0.2000	0.5000		
DMU3	0.0556	0.1111	0.2778		
DMU4	0.1429	0.2857	0.7143		
DMU5	0.1000	0.2000	0.5000		
S*					
DMU1	0.0000	0.5000	0.0000	0.0000	0.0000
DMU2	0.0000	1.0000	0.0000	0.0000	0.0000
DMU3	0.0000	1.0000	0.0000	0.0000	0.5000
DMU4	0.0000	0.2143	0.0000	0.0000	0.2857
DMU5	0.0000	0.0000	0.0000	0.0000	1.0000

非有效性决策单元 i 的优化					
DMU _i	x ₁	x ₂	y ₁		
DMU1	1.0000	2.0000	1.0000		
DMU3	5.0000	5.0000	3.0000		
DMU4	2.1429	1.4286	1.0000		
S*-	DMU1	DMU2	DMU3	DMU4	DMU5
x ₁	0.0000	0.0000	0.0000	0.0000	0.0000
x ₂	0.5000	0.0000	0.0000	0.0000	0.0000
S*+	DMU1	DMU2	DMU3	DMU4	DMU5
y ₁	0.0000	0.0000	0.0000	0.0000	0.0000
DMU _i	相对效率	DEA 有效性			
DMU1	0.5000	非弱 DEA 有效 可减少一些投入, 并保持产出不变。			
DMU2	1.0000	DEA 有效			
DMU3	0.8333	非弱 DEA 有效			
DMU4	0.7143	非弱 DEA 有效			
DMU5	1.0000	DEA 有效			

结果解释以第 3 个决策单元为例, 其技术效率 $TE_3=0.833$, 也就是说, 在不减少产出的条件下, 决策单元 3 可以减少所有投入消耗的 16.7%。其含义是产出在图 32-8 的 3' 处。投影点 3' 位于点 2 和 5 连线之间。这样决策单元 2 和决策单元 5 就可看成和决策单元 3 的前沿面。决策单元 2 和决策单元 5 定义了与决策单元 3 相关的前沿面位置, 从而定义决策单元 3 的有效率的产出。

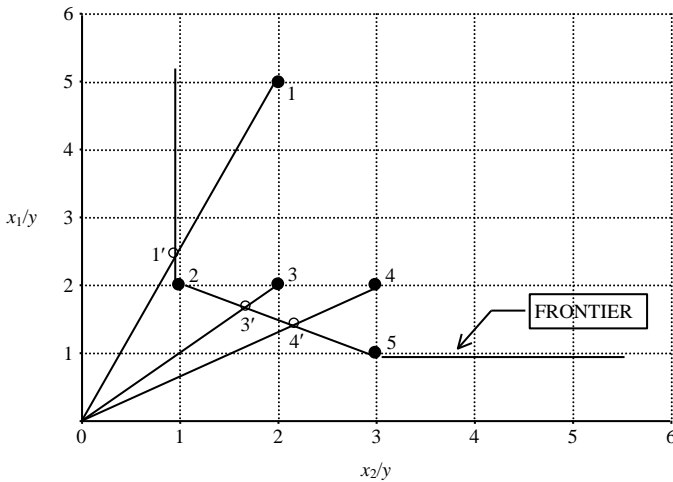


图 32-8 面向投入 CCR 模型分析结果图示

可以用同样的方法讨论其他两个非有效决策单元。决策单元 4 的 $TE_4=0.714$, 且与决策单元 3 有相同的前沿面。决策单元 1 的 $TE_1=0.5$, 它与决策单元 2 有相同的前沿面。同时也注意到决策单元 1 的投影点 1' 位于与 x_2 轴平行的前沿面上。这样, 它不能代表一个有效率的点 (根据 Koopman 的定义), 因为我们可以减少 0.5 个单位的 x_2 的投入且获得相同的产出, 这样决策单元 1 称为非径向有效。决策单元 2 和 5 的 TE_2 值为 1.0, 它们的前沿面就是其本身, 它们正是定义前沿面有效率的点。

总的说来,在输出结果中,这里从应用的角度给出由 CCR 模型来判断弱 DEA 有效和 DEA 有效的定义。

(1) 若 DEA 模型最优值 $\theta^*=1$, 则第 i 个决策单元为弱 DEA 有效, 反之亦然;

(2) 若 DEA 模型最优值 $\theta^*=1$, 并且满足所有的 $s^{-*}=0, s^{+*}=0$ (每个松弛变量都为零), 则第 j 个决策单元为 DEA 有效, 反之亦然。

如果模型最优值 $\theta^*<1$, 这说明第 i 个决策单元不是弱 DEA 有效, 当然更不是 DEA 有效的。也就是说, 它是非 DEA 有效的, 并且 θ^* 越小, 其有效性越差。但若用模型判定某个决策单元为 DEA 有效, 就需要检查所有的解 $\lambda^*, s^{+*}, s^{-*}, \theta^*$ 都满足 $\theta^*=1, s^{-*}=0, s^{+*}=0$ 。如果只有 $\theta^*=1$, 但并非所有的 $s^{-*}=0, s^{+*}=0$, 只能说明第 i 个决策单元是弱 DEA 有效, 不能说是 DEA 有效。但对于模型要判断所有的 $s^{-*}=0, s^{+*}=0$ 并不是一件容易事, 因此, 在实际中经常直接使用 CCR 模型, 而是一个稍加变化了的模型, 这个模型就是所谓具有非阿基米德无穷小 ε 的 CCR 模型。

DEA 分析中 CCR 模型结果有效性的经济含义。如果把相同类型的决策单元看成是某种“生产”活动, 则 DEA 有效性具有一定的经济含义: 在固定规模收益的 CCR 模型, 结果为 DEA 有效的决策单元, 从生产函数角度来讲, 既是技术有效的, 也是规模有效的。

技术有效是指生产处于最好状态下, 对投入 x 后获得最大产出 y , 如生产函数就是处于技术有效状态。

规模有效是指从规模收益递增到规模收益递减的点 (x_0, y_0) 。规模有效的几何意义是在 $(0, x_0)$ 区间内生产函数上凸, 这时边际收益是投资 x 的增函数, 即规模收益递增。而在 (x_0, ∞) 区间内生产函数下凹, 这时边际收益是投资 x 的减函数, 即规模收益递减。只有在点 x_0 处是规模收益的最佳点, 即规模有效。

DEA 有效性定理可分为以下三层含义:

(1) 如果 $\theta < 1$, 在保持产出 y_i 不变的前提下, 可以将投入 x_i 的各个分量均按同一比例减少, 则表明可以用比决策单元 i 更少的投入而使产出不变, 这正说明当前的决策单元 i 不是有效的生产活动, 其既非技术有效也非规模有效。

(2) 如果 $\theta = 1$, 要保持产出 y_i 不变, 投入 x_i 的各个分量不可以全部继续减少, 但可以做到减少部分投入, 而保持产出 y_i 不变, 则表明当前的决策单元 i 是弱 DEA 有效, 但不是 DEA 有效。从生产理论来讲, 它是技术有效而非规模有效。

(3) 如果 $\theta > 1$, 要保持产出 y_i 不变, 投入 x_i 的各个分量不仅不能整体按比例地减少, 而且连部分投入也不能再减少, 则表明当前的决策单元 i 是 DEA 有效的, 从生产理论来讲, 它既是技术有效的也是规模有效的。

32.3 数据包络分析扩展模型

32.3.1 可变规模收益(VRS)模式下的 BCC 模型

固定规模收益(CRS)模式设想仅当所有的决策单元是以最佳的规模运行时才合适。但由于不平等竞争、财政约束等都可能导致某个决策单元不能以最佳规模运行。Banker, Charnes, Cooper(1984)提出了一个对固定规模收益模式的 DEA 分析的扩展, 其中就考虑

到了可变规模收益(VRS)情况, 即当不是所有的决策单元都以最佳的规模运行时, 就使得对技术效益(TE)的测度受到规模效率(scale efficiency, SE)的影响。VRS 模式允许技术效益的计算不受规模效益的影响。

可变规模收益(VRS)模式下的 DEA 分析, 实际上只要对固定规模收益模式下的线性规划问题(CCR 模型)进行简单的改进就可以了, 也就是通过对式(32.2)增加凸面条件 $I1'\lambda=1$ 即可得到 BCC 模型:

$$\begin{aligned} \min_{\theta, \lambda} \quad & \theta \\ \text{s.t.} \quad & -y_i + Y\lambda \geq 0 \\ & \theta x_i - X\lambda \geq 0 \\ & I1'\lambda = 1 \\ & \lambda \geq 0 \end{aligned} \quad (32.3)$$

式中 $I1$ 是 $I \times 1$ 的 1 的向量。此方法形成了一个由相交面组成的凸面外壳, 这样, 它所包络的数据就比固定规模收益模式下的圆锥形外壳更紧密, 使得提供的技术效率量大于或等于使用固定规模收益模式所获得的技术效益。

规模效率(SE)可以通过对同样的数据进行一个 CCR 模型和一个 BCC 模型的 DEA 分析来完成。如果对于一个特殊的决策单元, CCR 和 BCC 的两个技术效益之间有差异, 则表明此决策单元规模无效的, 并且规模无效效率可以由 BCC 模型的技术效益和 CCR 模型的技术效益之间的差异计算出来(图 32-9)。

图 32-9 举出了一个单项投入和单项产出的例子, 并且绘出了在 CRS 模式下的 CCR 模型和在 VRS 模式下的 BCC 模型的 DEA 分析前沿面。在 CRS 模式下, 点 P 是面向投入的技术非有效, 其值是 P 到 P_C 间的距离, 在 VRS 模式下方的技术非有效测度是 P 到 P_V 间的距离。 P_C 和 P_V 两者之间的区别归因于规模非有效。

结合图 32-9, 可将所有这些效率用比值形式来表示:

$$\begin{aligned} TE_{CRS} &= AP_C/AP \\ TE_{VRS} &= AP_V/AP \\ SE &= AP_C/AP_V \end{aligned}$$

且所有这些指标的取值都在 0~1 之间, 同时注意

$$TE_{CRS} = TE_{VRS} \times SE$$

这是因为

$$AP_C/AP = (AP_V/AP) \times (AP_C/AP_V)$$

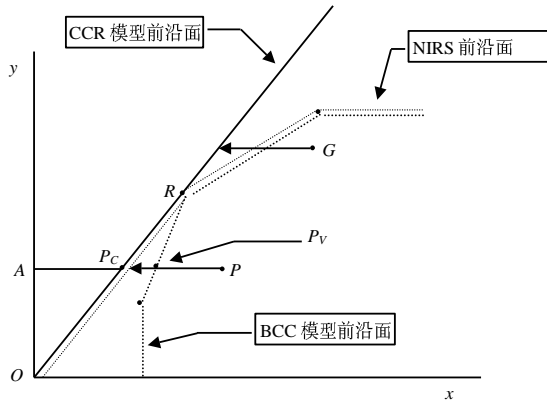


图 32-9 用 DEA 计算规模(经济)效益

这样，CRS 模式下的技术效率测度被分解成了“纯的”技术效率和规模效率。这种规模效率测度方法的一个缺点是所求的值并不能表明决策单元是在增长的还是在减少的规模收益情况中取得的。这需要由另外一个非增长规模收益(NIRS)的 DEA 分析方法来确定，这将在下面进行介绍。

前面提到，在 CCR 模型中，面向产出的模式与它们的面向投入的模式的结果相同，但在 BCC 模型中，其结果不一定相同。考虑以下面向产出的 BCC 模型：

$$\begin{aligned}
 \max_{\theta, \lambda} \quad & \theta \\
 \text{s.t.} \quad & -\theta y_i + Y \lambda \geq 0 \\
 & x_i - X \lambda \geq 0, \\
 & I' \lambda = 1 \\
 & \lambda \geq 0
 \end{aligned} \tag{32.4}$$

其中 $1 \leq \theta < \infty$ ，且 $\theta - 1$ 是由决策单元实现的，在投入中以适当的比例增长，而投入量恒定不变。注意 $1/\theta$ 定义了一个在 0~1 之间变化的值，即技术效益(TE)的得分值。一个有两项投入的面向产出的 DEA 可以用一个分段线性生产可能性曲线表示，如图 32-10 所示。

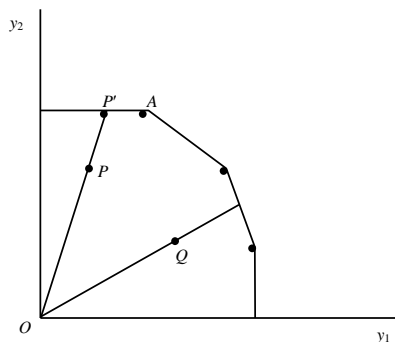


图 32-10 面向产出的 DEA 分析图示

注意可观测点在曲线的下方，当一个生产点通过产出的一个径向扩展被投影到与轴

垂直的曲线部分时就能计算出产出的松弛变量。例如，点 P 被投影在位于前沿面的点 P' ，但并不在效率前沿面上，因为在没有更多投入的条件下， y_1 的生产可以通过 AP' 的量得到增长。这也就是 AP' 在产出量 y_1 中出现松弛变量的原因。

有一点需说明的是面向产出和投入的模式将能够准确地估计出相同的前沿面，然后通过定义来确定一系列相同的决策单元是有效率的。但非有效的决策单元的有关的效率测定值在这两种方法上有区别。

BCC 模型操作示例 我们将使用一个简单的例子来说明 BCC 面向投入的 DEA 分析，这个例子仅涉及到一项投入、一项产出的 5 个决策单元。分析时在 DPS 中数据编辑格式如图 32-11。注意，数据块的左边是投入指标，右边是产出指标。

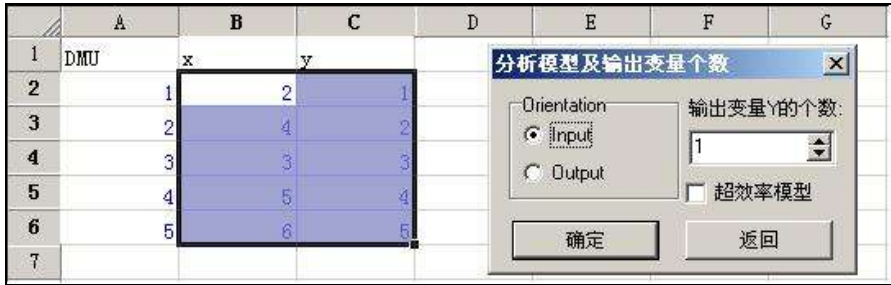


图 32-11 数据包络分析 BCC 模型分析数据格式及用户界面

按如图 32-11 左边所示编辑、选择数据，执行专业统计→数据包络分析及计量经济统计→BCC 模型，系统弹出如图 32-11 右边所示用户对话框。输入相应参数后即可得到计算 VRS 模式下的结果，这里仅列出 VRS 模式下的(相对)技术效率如下：

DMU _i	相对效率	名次	DEA 有效性
DMU1	1.0000	1	DEA 有效
DMU2	0.6250	5	非弱 DEA 有效
DMU3	1.0000	1	DEA 有效
DMU4	0.9000	4	非弱 DEA 有效
DMU5	1.0000	1	DEA 有效

为计算规模效率，还需估计在 CRS 模式下的技术效益，这时只要按图 32-11 所示，选中数据并执行 CCR 模型，即可得到各个决策单元的技术效益为 0.500，0.500，1.000，0.800 和 0.833。根据公式

$$TE_{CRS} = TE_{VRS} \times SE$$

可得

$$SE = TE_{CRS} / TE_{VRS}$$

这样可得到各个决策单元的规模效益(SE)分别为 0.500，0.800，1.000，0.889 和 0.833。其结果可由图 32-12 表示出来。

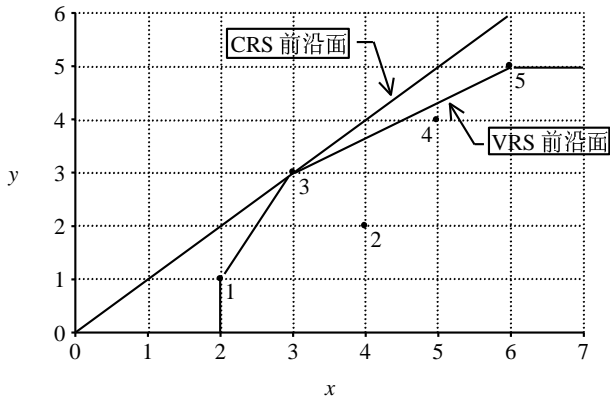


图 32-12 VRS 模式下(BCC 模型)面向投入的 DEA 分析

从图 32-12 可见, 当假设 CRS 时, 决策单元 3 是唯一的有效率单位(处于 DEA 前沿面上), 但是当假设 VRS 时, 决策单元 1, 3 和 5 都是有效率的(技术效率值都等于 1, 即它们都处在 DEA 前沿面上)。

各种效率测定计算。以决策单元 2 为例, 决策单元 2 在 CRS 和 VRS 技术下都是非有效的。CRS 模式下技术效率 TE 等于 $2/4 = 0.5$; VRS 模式下的 TE 为 $2.5/4 = 0.625$; 规模效率等于 CRS 模式下的 TE 和 VRS 模式下的 TE 的比值, 即 $0.5/0.625 = 0.8$ 。

32.3.2 可变规模收益(VRS)模式下的 IRS 模型和 DRS 模型

规模效益测度的缺点之一是其值不能说明某企业(单位)在某一程度(范围)内规模效益是增加还是减少。为解决非增加返回的规模(non-increasing returns to scale, NIRS)效益这一问题, 我们可以通过一个 DEA 分析模型进行处理。这个 DEA 模型(decreasing returns to scale 模型, DRS 模型)可通过替换 BCC 模型中的约束条件 $I'\lambda = 1$ 为 $I'\lambda \leq 1$ 得到:

$$\begin{aligned}
 \min_{\theta, \lambda} \quad & \theta \\
 \text{s.t.} \quad & -y_i + Y\lambda \geq 0 \\
 & \theta x_i - X\lambda \geq 0, \\
 & I'\lambda \leq 1 \\
 & \lambda \geq 0
 \end{aligned} \tag{32.5}$$

图 32-9 中绘出了 NIRS 模式下的 DEA 前沿面。对于一个指定的决策单元, 规模非有效(比如规模效益的增长或减少)的本质是由 NIRS 模式下的技术效益是否与 VRS 模式下的技术效益相等决定的。如果它们不相等(如图 32-9 中 P 点的情况), 那么对于该决策单元单位就存在增长的规模收益; 如果它们相等(如图 32-9 中 Q 点的情况), 那么就存在规模收益减少的情形。该模型的分析过程和前面介绍的 CCR 模型及 BCC 模型的操作完全相同。

与 DRS 模型 DEA 分析技术相对应的还有 IRS 模型, 该模型只需将 DRS 模型中的约束条件 $I'\lambda \leq 1$ 改为 $I'\lambda \geq 1$ 即可。这时, 如果得到的效率值小于 1, 那么就存在着规模收益增加的现象。

32.3.3 基于松弛测度的 SBM 模型

基于松弛变量测度(slacks-based measure, SBM)的 DEA 效率分析方法可理解为利润(profit)最大化的一种分析技术。因为在数据包络分析中, 松弛变量直接反映了我们关心的决策单元的投入的过度或产出的不足的程度, 且这种投入的过度或产出的不足的量只受指定的评价决策单元的影响, 而与整个数据集中其他决策单位无关。因此, 新的 SBM 模型的效率测度虽与前面介绍的基本的 CCR 模型有密切的联系, 但和 CCR 模型相比较, SBM 模型的优化更偏重于利润最大化的考虑。SBM 模型以优化其松弛变量为目标函数, 其分式规划的形式为

$$\begin{aligned} \min \quad & \rho = \frac{1 - (1/m) \sum_{i=1}^m s_i^- / x_{i0}}{1 - (1/s) \sum_{r=1}^s s_r^+ / y_{r0}} \\ \text{s.t.} \quad & x_0 = X\lambda + s^- \\ & y_0 = Y\lambda - s^+ \\ & \lambda \geq 0, s^- \geq 0, s^+ \geq 0 \end{aligned}$$

运用 Charnes-Cooper 转换方法, 其分式规划可以转换为线性规划问题, 再通过一些变换, Tone 得到了如下线性规划形式:

$$\begin{aligned} \min \quad & \tau = t - (1/m) \sum_{i=1}^m s_i^- / x_{i0} \\ \text{s.t.} \quad & 1 = t + (1/s) \sum_{r=1}^s s_r^+ / y_{r0} \\ & tx_0 = X\lambda + s^- \\ & ty_0 = Y\lambda - s^+ \\ & \lambda \geq 0, s^- \geq 0, s^+ \geq 0, t > 0 \end{aligned} \quad (32.6)$$

对该线性规划进行求解, 即可得到 SBM 模型效率参数估计值。与基本模型 CCR 相比较, SBM 模型的基本特性主要考虑投入/产出的利润最大化, 而 CCR 模型考虑的是投入和产出之比的比率最大化。且在效率评价中, SBM 模型的效率得分值小于等于 CCR 模型的效率得分值, 这在有些时候更便于各个决策单元的相互评价、比较。

SBM 模型操作示例 这里使用一个简单例子来说明 SBM 面向投入 DEA 分析, 这个例子是有 3 项投入、3 项产出的 6 个决策单位。分析时在 DPS 中数据编辑格式如图 32-13。注意, 数据块的左边是投入指标, 右边是产出指标。

	A	B	C	D	E	F	G	H	I	J	K	L
27	DMU	x1	x2	x3	y1	y2	y3					
28	1	1	1	1	2	2	2					
29	2	1	2	1	2	2	2					
30	3	1	1	2	1	2	2					
31	4	1	1	1	1	2	1					
32	5	1	1	1	2	1	0.5					
33	6	2	2	2	0.5	0.5	1					
34												

图 32-13 数据包络分析 SBM 模型分析数据格式及用户界面

按如图 32-13 左边所示编辑、选择数据，执行专业统计→数据包络分析及计量经济统计→SBM 模型，系统弹出如图 32-13 右边所示用户对话框。输入相应参数后即可得到 SBM 模型分析结果，这里仅列出 SBM 模型技术效率，并同时列出 CCR 模型的技术效率结果，以资比较。

DMU _i	相对效率	CCR 模型
DMU1	1.0000	1
DMU2	0.8333	1
DMU3	0.6250	1
DMU4	0.6000	1
DMU5	0.4286	1
DMU6	0.1500	0.25

从 SBM 模型和 CCR 模型的输出结果可以看出，根据 SBM 模型得到的相对效率值可以对各个决策单元进行排序，而 CCR 模型则不能。因此，在效率评价中，SBM 模型比 CCR 模型有更强的分辨能力。

32.3.4 DEA 分析非任意控制变量模型

这里介绍另一种 DEA 分析模型。在该模型中，有的投入项目可由管理者控制，而另一些投入项目则不能控制。对收益规模可变的 BCC 模型，假定其产出是固定的而投入是可变动的。我们要求对于一定水平的产出，决策者怎样尽量减少可变动的投入，即管理、决策者需要自主决定控制哪些投入并进行 DEA 分析。在这种情况下，我们将投入项目分为可任意处置(X^D)的和不能任意处置(X^{ND})的两类，并据此对 BCC 模型进行改进，得到

$$\begin{aligned}
 & \min_{\theta, \lambda} \quad \theta \\
 & \text{s.t.} \quad -y_i + Y\lambda \geq 0 \\
 & \quad \quad \theta x_i^D - X^D \lambda \geq 0 \\
 & \quad \quad x_i^{ND} - X^{ND} \lambda \geq 0 \\
 & \quad \quad I1' \lambda = 1 \\
 & \quad \quad \lambda \geq 0
 \end{aligned} \tag{32.7}$$

因此，在该 DEA 分析模型中，参数 θ 只与可以处置的投入项目有关。

非随意处置变量也可以理解为“外部”的环境变量，因此，这也是处理在有非可控的环境变量影响下的数据包络分析技术。

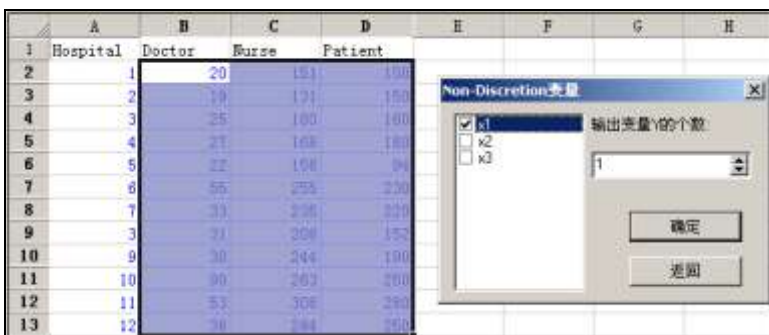


图 32-14 含非处置投入项目的 DEA 模型分析数据格式及用户界面

如图 32-14 左边所示选择数据，执行专业统计→数据包络分析及计量经济统计→非随意处置变量模型，系统弹出如图 32-14 右边所示用户对对话框。在对话框中，左边的 x_1, x_2, \dots 对应选中的数据列数。如果该列数据是非随意处置的投入项目，就在里面打上“√”。这里的第一项是不能随意处置的投入项目，故在“ x_1 ”前打上“√”，再输入产出项目个数，点击“确定”即可得到分析结果。

DMU_j	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
相对效率	0.578	1.000	0.776	0.936	0.519	0.788	0.818	0.644	0.680	0.830	0.799	0.769

32.4 考虑价格因素时的 DEA 模型

如果参与决策的部门提供了有关价格方面的信息，在考虑决策、行动时，更为合理的是需要利用其价格信息，以使决策结果达到成本最小化或收益最大化或利润最大化。那么对这些决策单元进行评价时，既要考虑技术效率也要考虑分配效率。

32.4.1 成本最小化模型

可变规模收益情形下成本最小化问题可以通过求解两次 DEA 模型来实现。首先是应用固定规模收益下的 CCR 模型进行分析，得到参与决策各个决策单元技术效益(TE)，然后再通过建立求解成本效益最小化的 DEA 模型得到成本效率。成本效率最小化模型的定义是

$$\begin{aligned}
 \min_{\lambda, x_i^*} \quad & w'_i x_i^* \\
 \text{s.t.} \quad & -y_i + Y\lambda \geq 0 \\
 & x_i^* - X\lambda \geq 0, \\
 & I'\lambda = 1 \\
 & \lambda \geq 0
 \end{aligned} \tag{32.8}$$

式中 w_i 是决策单元投入价格, x_i^* 是第 i 个决策单元的投入的最小化成本, 假设投入价格为 w_i 产出水平为 y_i , 第 i 个决策单元的总的成本效率(CE)计算如下

$$CE = w_i' x_i^* / w_i' x_i$$

也就是最低成本与已知成本的比值。然后使用下式计算分配效率, 即

$$AE = CE/TE$$

AE, CE 和 TE 的取值在 0~1 之间, 当取值为 1 时, 表明是充分有效的。此外, 这个方法已隐含地将松弛变量引入分配效率测度之中。

32.4.2 收入最大化模型

如果收入最大化是管理、决策者们的目的, 则在混杂有产出的分配非有效可以用类似方式进行解释。对于可变规模收益情形下的产出最大化问题, 首先是通过固定规模收益下的面向产出的 BCC 模型分析得到参与决策单位的技术效益(TE), 然后再求解产出效益最大化的 DEA 模型:

$$\begin{aligned} \min_{\lambda, y_i^*} \quad & p_i y_i^* \\ \text{s.t.} \quad & -y_i^* + Y\lambda \geq 0 \\ & x_i - X\lambda \geq 0, \\ & I'\lambda = 1 \\ & \lambda \geq 0 \end{aligned} \tag{32.9}$$

式中 p_i 是某个决策单元产出的价格, y_i^* 是第 i 个决策单元的产出的最大化收益, 假设产出价格为 p_i 投入水平为 x_i , 第 i 个决策单元的总的收入效率(RE)计算如下

$$RE = p_i' y_i^* / p_i' y_i$$

也就是实测收入与模型最大收益的比值, 然后使用下式计算分配效率, 即

$$AE = RE/TE$$

AE, RE 和 TE 的取值在 0~1 之间, 当取值为 1 时, 表明是充分有效的。

32.4.3 利润最大化模型

如果决策者已经同时掌握了投入和产出的价格, 对于制定的决策单元, 可以用 DEA 分析方法计算其利润最大化。利润最大化的 DEA 模型为

$$\begin{aligned}
 & \min_{\lambda, x_i^*, y_i^*} \quad P_i' y_i^* - w_i' x_i^* \\
 & \text{s.t.} \quad -y_i^* + Y\lambda \geq 0 \\
 & \quad \quad x_i^* - X\lambda \geq 0 \\
 & \quad \quad I1'\lambda = 1 \\
 & \quad \quad \lambda \geq 0
 \end{aligned} \tag{32.10}$$

模型中各个参数含义和前面的相同。在利润最大化的 DEA 分析中，计算得到的目标函数值不一定在 0~1 之间，如果该值小于 0，则利润为负；如果该值等于 0，表明利润不定。

32.4.4 考虑价格因素时 DEA 模型操作示例

考虑价格因素时，投入、产出数据编辑格式仍然是左边为投入指标，右边为产出指标。各个决策单位的价格因素值可放在 DPS 电子表格任意位置(一般放在投入产出的数据下面或右边)，分析时，在选中投入、产出指标值之后，按下 Ctrl 键的同时，用鼠标选中价格信息数据，进行成本最小化分析时，只选中成本价格数据；进行产出最大化分析时，只选中产出的价格数据。如进行利润最大化分析，需选中投入、产出的价格数据。如果各个决策单位的价格相等，那么只要输入一行价格数据，并选中即可。

这里的例子是 2 项投入 2 项产出的 12 个决策单位的成本最小化分析。分析时按图 32-15 格式编辑数据。注意其价格数据的编辑、选择方式。

	A	B	C	D	E	F	G	H	I
1	DMU _i	x1	x2	y1	y2		x1 的价格	x2 的价格	
2	1	20	151	100	90		500	100	
3	2	19	131	150	50		350	80	
4	3	25	160	160	55		450	90	
5	4	27	168	180	72		600	120	
6	5	22	158	94	66		300	70	
7	6	55	255	230	90		450	80	
8	7	33	235	220	88		500	100	
9	8	31	206	152	80		450	85	
10	9	30	244	190	100		380	76	
11	10	50	268	250	100		410	75	
12	11	53	306	260	147		440	80	
13	12	38	284	250	120		400	70	
14									

图 32-15 成本最小化 DEA 模型分析数据格式

如图 32-15 所示选择数据，执行专业统计→数据包络分析及计量经济统计→成本最小化模型，系统弹出用户对话框后键入产出的指标数量，再点击“确定”即可得到分析结果。

DMU _i	技术效率	分配效率	成本效率
------------------	------	------	------

DMU1	1.0000	1.0000	1.0000
DMU2	1.0000	1.0000	1.0000
DMU3	0.8827	0.9366	0.8268
DMU4	1.0000	0.9146	0.9146
DMU5	0.7635	0.9799	0.7481
DMU6	0.8348	0.6982	0.5828
DMU7	0.9020	0.9980	0.9002
DMU8	0.7963	0.9422	0.7503
DMU9	0.9604	0.9838	0.9448
DMU10	0.8706	0.7983	0.6951
DMU11	0.9551	0.8242	0.7872
DMU12	0.9582	0.9946	0.9531

32.5 面板数据的 Malmquist 指数

32.5.1 Malmquist 指数简介

面板数据(panel data)的数据包络分析, 可用 Malmquist 的 TFP 指标来测度生产力的改变, 并且将生产力的改变分解成技术改变和技术效率改变。Malmquist 指数通常用于测量企业的全要素生产率(TFP)以及技术变化和技术效率变化(包括纯效率和规模效率变化)。

Malmquist 生产率指数首先由 Caves, Christensen, Diewert(1982)引入, 由 Färe 等进一步发展而来。该指数运用 Shephard 提出的距离函数(distance functions)定义, 用来描述不需要说明具体行为标准(例如成本最小化和利润最大化)的多个输入变量和多个输出变量生产技术。运用面向产出(output-oriented)方法或面向投入(input-oriented)方法能够定义距离函数。给定投入变量矩阵, 一个产出的距离函数定义为产出指标变量矩阵的最优比例项。同样, 给定产出变量矩阵, 产出变量距离函数可以看作是投入指标变量矩阵的最小比例项。

一般来说我们运用面向产出的方法来测度各决策单元的效率。根据 Shephard 和 Färe 的方法, 产出指标变量的距离函数定义如下:

$$D_0(x, y) = \inf \{ \delta : (x, y/\delta) \in P(x) \}$$

其中 x 和 y 分别表示输入变量和输入变量矩阵, δ 表示 Farrell 的面向输出的效率指标, $P(x)$ 定义为可能生产集合。如果 y 是 $P(x)$ 的组成部分, 则函数的值将小于或等于 1。如果 y 位于可能生产集合的外部前沿面上, 那么函数值将等于 1; 反之, 如果 y 位于 $P(x)$ 外部, 那么函数的值将大于 1。

Caves, Christensen, Diewert(1982)第一次定义了一个投入、一个产出, 基于产出指标变量的 Malmquist 生产率指数:

$$M_0^t = \frac{D_0^t(x_{t+1}, y_{t+1})}{D_0^t(x_t, y_t)} \quad (32.11)$$

$$M_0^{t+1} = \frac{D_0^{t+1}(x_{t+1}, y_{t+1})}{D_0^{t+1}(x_t, y_t)} \quad (32.12)$$

其中 $D_0^t(x_t, y_t)$ 和 $D_0^{t+1}(x_{t+1}, y_{t+1})$ 是分别根据生产点在相同时间段(即 t 和 $t+1$)同前沿面技术相比较得到的输出距离函数, $D_0^t(x_{t+1}, y_{t+1})$ 和 $D_0^{t+1}(x_t, y_t)$ 分别是根据生产点在混合期间同前沿面技术相比较得到的输出距离函数。

为了避免随意选择一种参照技术, Färe 等运用两个前面所定义的 Malmquist 生产率指数的几何平均值来计算定向输出 Malmquist 指数:

$$M_0(x_t, y_t, x_{t+1}, y_{t+1}) = \frac{D_0^{t+1}(x_{t+1}, y_{t+1})}{D_0^t(x_t, y_t)} \sqrt{\frac{D_0^{t+1}(x_{t+1}, y_{t+1})}{D_0^{t+1}(x_t, y_t)} \times \frac{D_0^t(x_{t+1}, y_{t+1})}{D_0^t(x_t, y_t)}} \quad (32.13)$$

式中根号前的第一项测算技术效率的面向产出指标在区间 t 和 $t+1$ 的(逐渐趋近于生产前沿面)变化, 而根号里面的项作为技术变化(创新)指标。它是前沿面在区间 t 和 $t+1$ 变化的几何平均值。

根据 Färe 等(1994)的研究, 上方方程中的 Malmquist 指数中技术效率变化部分能够进一步分解为纯效率变化和规模效率变化:

$$M_0(x_t, y_t, x_{t+1}, y_{t+1}) = \frac{S_0^t(x_t, y_t)}{S_0^t(x_{t+1}, y_{t+1})} \times \frac{D_0^t(x_{t+1}, y_{t+1}/VRS)}{D_0^t(x_t, y_t/VRS)} \times \sqrt{\frac{D_0^t(x_{t+1}, y_{t+1})}{D_0^{t+1}(x_{t+1}, y_{t+1})} \times \frac{D_0^t(x_t, y_t)}{D_0^{t+1}(x_t, y_t)}} \quad (32.14)$$

其中第一项表示规模效率变化, 第二项表示纯技术效率变化, 最后一项表示技术变化。生产率的变化, 即 M_0 , 分解为纯技术变化, 规模效率变化以及技术变化可能等于 1、大于 1 和小于 1, 分别表示没有变化、有改进和倒退了。图 32-16 显示了距离函数和 Malmquist 指数是如何构建的。

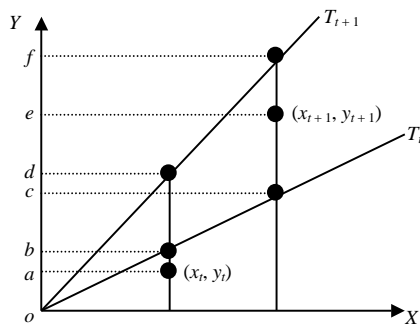


图 32-16 Malmquist 指数和距离函数的关系

从图 32-16 可知

$$\begin{aligned}
 D_0^t(x_t, y_t) &= oa / ob \\
 D_0^{t+1}(x_{t+1}, y_{t+1}) &= oe / of \\
 D_0^t(x_{t+1}, y_{t+1}) &= oe / oc \\
 D_0^{t+1}(x_t, y_t) &= oa / od
 \end{aligned}$$

故可分别得到技术效率变化:

$$\frac{D_0^{t+1}(x_{t+1}, y_{t+1})}{D_0^t(x_t, y_t)} = \frac{oe / of}{oa / ob}$$

和技术变化:

$$\sqrt{\frac{D_0^t(x_{t+1}, y_{t+1})}{D_0^{t+1}(x_{t+1}, y_{t+1})} \times \frac{D_0^t(x_t, y_t)}{D_0^{t+1}(x_t, y_t)}} = \sqrt{\frac{oe / oc}{oe / of} \times \frac{oa / ob}{oa / od}}$$

以及 Malmquist 指数(M_0):

$$\frac{D_0^{t+1}(x_{t+1}, y_{t+1})}{D_0^t(x_t, y_t)} \sqrt{\frac{D_0^t(x_{t+1}, y_{t+1})}{D_0^{t+1}(x_{t+1}, y_{t+1})} \times \frac{D_0^t(x_t, y_t)}{D_0^{t+1}(x_t, y_t)}} = \frac{oe / of}{oa / ab} \sqrt{\frac{oe / oc}{oe / of} \times \frac{oa / ob}{oa / od}}$$

参数和非参数方法都能够用于计算构成 Malmquist 指数基础的距离函数在 DPS 中, 我们采用 Färe 等(1994)推荐的非参数 DEA 线性规划来计算各个时间段 t 和 $t+1$ 间的距离函数来检验 TFP 的变化。这里仅简述用于计算 cross-sectional 效率指标和 Malmquist 指标的 DEA 方法。

假定有 K 个待决策的单元, 每个决策单元在每个时段 t 里有 M 项投入指标, 即 $x_{m,kt}$, $m=1,2,\dots,M$, S 项产出指标, 即 $y_{s,kt}$, $s=1,L,S$ 。 $F_0^t(x_{kt}, y_{kt})$ 表示面向产出的 Farrell 效率指标, 该效率指标给出了第 k 个决策单元在时期 t 内 y_t 的最大可能范围。

$$\begin{aligned}
 F_0^t(x_{kt}, y_{kt}) & \quad \max \delta \\
 \text{s.t.} \quad \delta y_{s,kt} & \leq \sum_{k=1}^K \lambda_{kt} y_{s,kt}, s=1,L,S \\
 \sum_{k=1}^K \lambda_{kt} x_{m,kt} & \leq x_{m,kt}, m=1,L,M \\
 \lambda_{kt} & \geq 0, k=1,L,K
 \end{aligned} \tag{32.15}$$

由于 Farrell 效率指标是 Shephard 距离函数的反函数, 所以 $F_0^t(x_{kt}, y_{kt})$ 实际上等于 $[D_0^t(x_{kt}, y_{kt})]^{-1}$ 。根据距离函数和 Farrell 技术效率指标的关系可建立下面四个线性规划问题, 并且得到它们的解的倒数, 那么把它们代入方程(32.13)中就可以得到 Malmquist 效率指数。这四个线性规划问题如下所示:

$$\begin{aligned}
 F_0^t(x_{kt}, y_{kt}) &= \left[D_0^t(x_{kt}, y_{kt}) \right]^{-1} && \max \delta \\
 \text{s.t. } \delta y_{s,kt} &\leq \sum_{k=1}^K \lambda_{kt} y_{s,kt}, s=1, \dots, S \\
 \sum_{k=1}^K \lambda_{kt} x_{m,kt} &\leq x_{m,kt}, m=1, \dots, M \\
 \lambda_{kt} &\geq 0, k=1, \dots, K
 \end{aligned} \tag{32.16}$$

$$\begin{aligned}
 F_0^{t+1}(x_{kt+1}, y_{kt+1}) &= \left[D_0^{t+1}(x_{kt+1}, y_{kt+1}) \right]^{-1} && \max \delta \\
 \text{s.t. } \delta y_{s,kt+1} &\leq \sum_{k=1}^K \lambda_{kt+1} y_{s,kt+1}, s=1, \dots, S \\
 \sum_{k=1}^K \lambda_{kt+1} x_{m,kt+1} &\leq x_{m,kt+1}, m=1, \dots, M \\
 \lambda_{kt+1} &\geq 0, k=1, \dots, K
 \end{aligned} \tag{32.17}$$

$$\begin{aligned}
 F_0^t(x_{kt+1}, y_{kt+1}) &= \left[D_0^t(x_{kt+1}, y_{kt+1}) \right]^{-1} && \max \delta \\
 \text{s.t. } \delta y_{s,kt+1} &\leq \sum_{k=1}^K \lambda_{kt+1} y_{s,kt}, s=1, \dots, S \\
 \sum_{k=1}^K \lambda_{kt} x_{m,kt} &\leq x_{m,kt+1}, m=1, \dots, M \\
 \lambda_{kt} &\geq 0, k=1, \dots, K
 \end{aligned} \tag{32.18}$$

$$\begin{aligned}
 F_0^{t+1}(x_{kt}, y_{kt}) &= \left[D_0^{t+1}(x_{kt}, y_{kt}) \right]^{-1} && \max \delta \\
 \text{s.t. } \delta y_{s,kt} &\leq \sum_{k=1}^K \lambda_{kt+1} y_{s,kt+1}, s=1, \dots, S \\
 \sum_{k=1}^K \lambda_{kt+1} x_{m,kt+1} &\leq x_{m,kt}, m=1, \dots, M \\
 \lambda_{kt+1} &\geq 0, k=1, \dots, K
 \end{aligned} \tag{32.19}$$

注意到，前面两个线性规划是计算同一时间段内的 Farrell 效率指标，后两个线性规划是计算前后两个时间段交叉时的 Farrell 效率指标。为了进一步检验规模效率的变化，我们需要计算在规模技术收益变化时的逆产出距离函数，这需要在式(32.16)和式(32.17)中增加一个限制条件

$$\sum_{k=1}^K \lambda_{kt} = 1$$

在每一个时间段，规模效率变化用规定规模收益条件下的距离函数与可变规模收益条件下的距离函数的比率来表示；同样，纯技术效率的变化也可以通过计算可变规模收益下

的本期(own-period)距离函数比率得到。

通过上面的计算,我们能够将来 Malmquist 指数分解成技术变化、纯效率变化和规模效率变化三个部分,即式(32.14)。

32.5.2 Malmquist 指数计算示例

Malmquist 指数计算的数据编辑格式:第 1 列是按 1,2,⋯,N 自然数顺序的决策单元的序号,第 2 列是按 1,2,⋯,T 自然数顺序的时期序号,从第 3 列开始放入各项指标:左边为投入指标,右边为产出指标。这里的例子是 4 个决策单元、5 个时期、1 项投入 1 项产出的 Malmquist 指数数据(图 32-17)。

	A	B	C	D	E	F
1		Firm	Year	Input	Output	
2		1	1	719	400	
3		2	1	358	92	
4		3	1	264	74	
5		4	1	190	44	
6		1	2	800	539	
7		2	2	401	139	
8		3	2	275	100	
9		4	2	203	49	
10		1	3	850	339	
11		2	3	418	120	
12		3	3	280	65	
13		4	3	231	66	
14		1	4	894	125	
15		2	4	427	34	
16		3	4	291	54	
17		4	4	255	56	
18		1	5	903	103	
19		2	5	390	10	
20		3	5	269	33	
21		4	5	262	57	
22						

图 32-17 Malmquist 指数计算数据格式

如图 32-17 所示选择数据,执行专业统计→数据包络分析及计量经济统计→Malmquist 指数,系统弹出用户对话框后键入产出的指标数量,再点击“确定”即可得到分析结果。

Malmquist 指数分析计算					
距离汇总(略):					
Malmquist 指数汇总:					
年份(时期)=2					
决策单元 i	技术效率	技术变化	纯技术效率	规模效率	生产率

DMU1	1.0000	1.2111	1.0000	1.0000	1.2111
DMU2	1.1138	1.2111	1.1219	0.9928	1.3489
DMU3	1.0712	1.2111	1.1726	0.9135	1.2973
DMU4	0.8607	1.2111	1.0000	0.8607	1.0423
几何平均	1.0066	1.2111	1.0710	0.9399	1.2191
年份(时期)=3					
DMU _i	技术效率	技术变化	纯技术效率	规模效率	生产率
DMU1	1.0000	0.5919	1.0000	1.0000	0.5919
DMU2	1.3991	0.5919	1.2299	1.1376	0.8282
DMU3	1.0785	0.5919	0.8020	1.3447	0.6384
DMU4	1.9996	0.5919	1.0000	1.9996	1.1837
几何平均	1.3180	0.5919	0.9966	1.3225	0.7802
.....					
年份(时期)=5					
DMU _i	技术效率	技术变化	纯技术效率	规模效率	生产率
DMU1	0.8235	0.9907	1.0000	0.8235	0.8158
DMU2	0.3251	0.9907	0.3314	0.9809	0.3220
DMU3	0.6673	0.9907	0.6365	1.0485	0.6611
DMU4	1.0000	0.9907	1.0000	1.0000	0.9907
几何平均	0.6501	0.9907	0.6777	0.9593	0.6440
Malmquist 指数汇总结果					
DMU _i	技术效率	技术变化	纯技术效率	规模效率	生产率
DMU1	0.8509	0.7908	1.0000	0.8509	0.6729
DMU2	0.7107	0.7908	0.7127	0.9973	0.5620
DMU3	1.0285	0.7908	0.9235	1.1137	0.8134
DMU4	1.2450	0.7908	1.0000	1.2450	0.9845
几何平均	0.9381	0.7908	0.9007	1.0415	0.7418

分析结果前面部分是各决策单元技术效率估计值,后面部分是各个时期的技术效率、技术变化、纯效率变化、规模效率变化和全要素生产率变化(基于 Malmquist 指数)的统计量。结果显示全要素生产率(TFP)、技术效率以决策单元 4 最高,决策单元 3 次之,决策单元 2 最低。

32.6 随机前沿面模型

前面用于生产效率估计采用的非参数数据包络分析方法,这里介绍的是参数分析技术,即随机前沿面分析。随机前沿建模及有效性度量, Farrell 于 1957 年就提出了度量一个单元效率的方法。该度量由两个部分组成:技术效率(technical efficiency)反映单元从给定的一组输入获取最大产出的能力;而配置效率(allocative efficiency)反映的是当给定各输入的价格后,单位最优配置诸输入的能力。将这两个量组合在一起就构成了测量总经济效率(total economic efficiency)的一种量化指标。这里的效率测度方法使用的前提是有效率单元的生产函数已知。但实际上生产函数总是未知的。因此 Aigner 等提出了随机前

沿面模型的参数方法以解决这一问题。

Aigner 和 Chu 针对 N 个决策单元的样本数据考察了柯布-道格拉斯(Cobb-Douglas)型前沿面生产函数的参数估计问题, 模型的定义是

$$\ln(y_i) = x_i\beta - u_i, \quad i = 1, 2, \dots, N \quad (32.20)$$

其中 $\ln(y_i)$ 是第 i 个单元输出的对数; x_i 是 $k+1$ 维的行向量, 其第一个元素是 1, 其余元素是第 i 个单元 k 个输入的对数; $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)'$ 是 $k+1$ 列向量, 它是未知的待估参数; u_i 是非负随机变量, 与工业企业生产中的技术无效性有关。

在已知输入向量 x_i 后, 第 i 个单元的技术效率可定义为输出的观测值与前沿面生产函数值之比:

$$TE_i = \frac{y_i}{\exp(x_i\beta)} = \frac{\exp(x_i\beta - u_i)}{\exp(x_i\beta)} = \exp(-u_i) \quad (32.21)$$

此值度量的是面向输出的 Farrell 技术效率, 并在 0 和 1 之间取值。它反映的是第 i 个单元与具有相同输入向量的充分有效单元的产出比。方程(32.21)定义的技术效率可由观测到的输出值 y_i 除以估计出的前沿产出 $\exp(x_i\beta)$ 得到。

但模型(32.21)没有把测量误差和其他噪声对前沿分析的影响考虑进去。所有偏离前沿的因素都归因于技术无效。处理噪声问题的方法就是著名的随机前沿面分析法。

32.6.1 随机前沿生产函数

Aigner, Lovell 和 Schmidt 以及 Meeusen 和 van den Broeck 于 1977 年分别提出了随机前沿生产函数的概念。在随机前沿生产函数模型中, 除了保留(32.20)方程中的非负随机变量 u_i 外, 还增加了随机误差项 v_i , 于是模型变为

$$\ln(y_i) = x_i\beta + v_i - u_i, \quad i = 1, 2, \dots, N \quad (32.22)$$

随机误差 v_i 不仅包含了测量误差以及对输出产生影响的其他因素, 诸如天气、运气等, 还包含了在生产函数中未被指明的输入变量的效应。Aigner, Lovell 和 Schmidt(1977) 在模型中假定 v_i 独立同分布于均值为 0、方差为 σ_v^2 的正态分布; u_i 独立同分布于指数分布或半正态分布, 且与 v_i 相互独立。方程(32.22)定义的模型为随机前沿生产函数, 这是因为输出值的上界是随机变量 $\exp(x_i\beta + v_i)$ 。随机误差 v_i 可能是正的, 也可能是负的, 所以随机前沿的产出会围绕随机模型的决定部分 $\exp(x_i\beta)$ 上下波动。

随机前沿模型的基本特性可由二维图形(如图 32-18)来描述。横轴代表输入, 纵轴代表输出。前沿模型中的确定部分 $\exp(x_i\beta)$ 是按规模报酬递减画出的。图形上有两个单元的输入与输出。第 i 个单元在输入水平 x_i 上, 产出为 y_i 。观测到的输入输出值用在 x_i 上方的 \times 来表示。因为随机误差 v_i 是正的, 随机前沿的输出值

$$y_i^* = \exp(x_i\beta + v_i)$$

用生产函数上方的产出值来表示。同样, 第 j 个单元在输入水平 x_j 上, 产出为 y_j 。然而,

因为随机误差 v_j 是负的，前沿产出值

$$y_j^* = \exp(x_j\beta + v_j)$$

将位于生产函数的下方。当然，由于随机误差 v_i 和 v_j 是不可观测的，因此，随机前沿的输出 y_i^* 和 y_j^* 也是不可观测的。但是，我们可以看到，随机前沿模型的确切部分是在这两个随机前沿的输出值之间。当随机误差大于相应的无效效应时，即如果 $v_i > u_i$ 则 $y_i > \exp(x_i\beta)$ ，其输出的观测值就会大于前沿的确切部分。

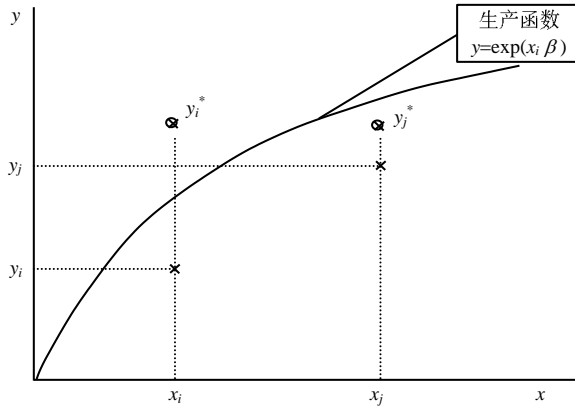


图 32-18 随机前沿生产函数

在随机前沿模型中，可以利用传统的最大似然法去估计标准误差，还可以进行假设检验。而这些在前面的确定模型中却是不可能的，因为确定模型违反最大似然法的某些正则条件。然而，随机前沿模型并不是没有问题的，其主要的问题是对 u_i 分布类型的选择没有一个通用的判断准则。一般分布形式如截断正态分布和双参数 Gamma 分布在某种程度上解决了这个问题，但最后的效益指标仍可能对分布的假设很敏感。截断正态分布将在下节讨论。

32.6.2 极大似然估计

随机前沿生产函数(32.22)中的参数可采用修正最小二乘估计或极大似然估计。Coelli 通过 Monte Carlo 实验得到了半正态模型下的有限样本性质，即当技术无效效应对总方差项的贡献较大时，极大似然估计显著地优于修正最小二乘估计。因此，一般可用极大似然估计来获取随机前沿模型中的参数。

实际应用中常假设技术无效效应是半正态的，因此，这里仅就这种情形进行讨论。假设 u_i 是独立同分布于截断(于原点处) $N(0, \sigma^2)$ 的随机变量， v_i 独立同分布于 $N(0, \sigma_v^2)$ ，且与 u_i 相互独立。Aigner 等得到了模型(32.22)的似然函数的对数形式，他们把似然函数按两方差参数的形式表示，即

$$\sigma_s^2 \equiv \sigma^2 + \sigma_v^2 \text{ 且 } \lambda = \sigma/\sigma_v$$

Battese 和 Corra 建议用参数 $\gamma = \sigma^2 / \sigma_s^2$, 因为 γ 是 0 和 1 之间的一个量, 而 λ 可能是任何一个非负数。为了用迭代最大化算法得到极大似然估计, 搜寻 γ 的参数空间以寻找一个合适的初值优于搜寻 λ 的空间。在这种参数化原则下, Battese 和 Corra(1977)证明似然函数的对数等于

$$\ln(L) = \frac{N \ln\left(\frac{\pi}{2}\right)}{2} - \frac{N \log(\sigma_s^2)}{2} + \sum_{i=1}^N \ln[1 - \Phi(z_i)] - \frac{\sum_{i=1}^N (\ln y_i - x_i \beta)^2}{2\sigma_s^2} \quad (32.23)$$

其中

$$z_i = \frac{\ln(y_i - x_i \beta)}{\sigma_s} \sqrt{\frac{\gamma}{1-\gamma}}$$

$\Phi(\cdot)$ 是标准正态随机变量的分布函数。

通过求对数似然函数(32.23)的最大值, 可以得到参数 β , σ_s^2 以及 γ 的极大似然估计。极大似然估计量是一致的、渐近有效的。

参数估计结果, 如果 γ 值为 0 意味着偏离前沿完全是由噪声引起的, 而其值为 1 意味着全部偏离来源于技术无效。然而需要强调的是 γ 不等于技术无效效应的方差与总残差方差的比。这是因为 u_i 的方差等于 $[(\pi-2)/\pi] \sigma^2$ 而不是 σ^2 。无效效应对总方差项(γ^*)的相对贡献为

$$\gamma^* = \gamma / [\gamma + (1-\gamma)\pi / (\pi-2)]$$

32.6.3 平均技术效率的估计

自从随机前沿生产函数提出以来, 在不同的实际应用中, 利用工业企业的截面数据不仅能估计模型中的参数, 而且同时还能估计平均技术效率。因此, 原来认为预测各个单元的技术效率是不可能的说法是不正确的。

若给出技术无效效应的分布假设, 技术效率 $TE_i = \exp(-u_i)$ 的数学期望是可以计算的。可以证明, 如果 u_i 是独立同分布于半正态分布的随机变量, 那么

$$E[\exp(-u_i)] = 2[1 - \Phi(\sigma_s \sqrt{\gamma})] \exp(-\gamma \sigma_s^2 / 2) \quad (32.24)$$

将方程(32.24)中的相关参数用相应的极大似然估计值代替, 就可以得到平均技术效率的极大似然估计。因为可预测每一个样本单元的技术效率, 那么平均技术效率的另一个估计量是各样本单元技术效率预测值的算术平均。但当样本单元的规模显著不同或不是利用简单随机抽样的方法选择样本单元时, 这个算术平均可能不是最好的估计。

32.6.4 单位级技术效率的预测

由定义可知第 i 个单元的技术效率是 $TE_i = \exp(-u_i)$ 。这里包含不可观测的技术无效效应 u_i 。即使是在随机前沿模型中参数向量 β 的真值已知, 也仅仅能观测到差值 $e_i \equiv v_i - u_i$ 。最好的 u_i 预测值是在给定值 $v_i - u_i$ 下 u_i 的条件期望。Jondrow 等(1982)首次

将这个结果用在随机前沿模型中，并得到如下结论

$$E[u_i|e_i] = -\gamma e_i + \sigma_A \left\{ \frac{\phi(\gamma e_i / \sigma_A)}{1 - \Phi(\gamma e_i / \sigma_A)} \right\} \quad (32.25)$$

其中

$$\sigma_A = \sqrt{\gamma(1-\gamma)\sigma_s^2}, \quad e_i = \ln(y_i) + x_i\beta$$

$\phi(\cdot)$ 是标准正态随机变量的密度函数。

Jondrow 等建议用 $1 - E[u_i|e_i]$ 预测第 i 个单元的技术效率。Battese 和 Coelli 指出，可用下式可得到 $\exp(-u_i)$ 最好的估计：

$$E[\exp(-u_i)|e_i] = \frac{1 - \Phi(\sigma_A + \gamma e_i / \sigma_A)}{1 - \Phi(\gamma e_i / \sigma_A)} \exp(\gamma e_i + \sigma_A^2 / 2) \quad (32.26)$$

这个预测与先由方程(32.25)估计 u_i ，然后再代入 $TE_i = \exp(-u_i)$ 的结果有差异。这是因为随机变量非线性函数的数学期望不等于该随机变量期望的函数，即若 $g(\cdot)$ 是非线性函数，则 $E[g(x)] \neq g[E(x)]$ 。在实际计算中，可把(32.26)式中的未知参数用它们的极大似然估计代替以得到技术效率预测值。

32.6.5 假设检验

对随机前沿模型(32.22)，可以对零假设“模型中不存在技术无效效应”，即 $H_0: \sigma^2 = 0$ 与备选假设 $H_1: \sigma^2 > 0$ 进行检验。这个假设检验可以用若干个检验统计量进行，常用的是 Wald 统计量，它等于 σ^2 的极大似然估计除以其标准差估计。但在许多情况下，还要对随机前沿模型中的参数，采用下面等价的两个假设检验： $H_0: \lambda = 0$ 与备选假设 $H_1: \lambda > 0$ 或 $H_0: \gamma = 0$ 与备选假设 $H_1: \gamma > 0$ 。这里仅考虑关于 γ 的假设检验。

为进行 Wald 检验，需计算 γ 的估计与其标准差估计的比值。如果 $H_0: \gamma = 0$ 是真的，这个统计量近似地服从标准正态分布。然而，由于 γ 不能取负值，因此，只能进行单边检验。Coelli 通过进行 Monte Carlo 研究，认为 Wald 检验的水平(即犯第一类错误的概率)性质很差。例如，对一个含有 100 个观测的样本，对水平是 $\alpha = 0.05$ 的 Wald 检验却拒绝零假设 20%(而不是 5%)。Coelli(1995)认为，当涉及极大似然估计时，为了具有正确的检验水平，应该进行单边广义似然比检验。

单边广义似然比检验。广义似然比检验需要在零假设和备选假设下同时对模型进行估计。在零假设 $H_0: \gamma = 0$ 下，因为没有技术无效效应 u_i ，所以模型等同于传统的平均响应函数，其检验统计量是

$$LR = -2 \ln \frac{L(H_0)}{L(H_1)} = -2 \{ \ln[L(H_0)] - \ln[L(H_1)] \} \quad (32.27)$$

其中 $L(H_0)$ 和 $L(H_1)$ 分别是零假设 H_0 和备选假设 H_1 下的似然函数值。

如果 H_0 是真的，通常假设这个检验统计量近似地服从卡方分布，自由度为约束的个数(在这个例中，自由度为 1)。因为 $\gamma = 0$ 位于 γ 参数空间的边界上，检验 $H_0: \gamma = 0$ 就产

生了困难。在这种情况下，如果 $H_0: \gamma=0$ 是真的，广义似然比统计量 LR 近似地服从混合的卡方分布，即

$$LR = \frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2$$

计算单边广义似然比检验的临界值相对简单。检验水平为 α 的临界值等于 $\chi_1^2(2\alpha)$ ，即服从 χ_1^2 的随机变量超过这个值的概率等于 2α 。这样，水平为 α 的单边广义似然比检验是：如果 LR 超过 $\chi_1^2(2\alpha)$ ，则拒绝 $H_0: \gamma=0$ ，而接受 $H_1: \gamma>0$ 。从而，检验水平为 $\alpha=0.05$ 的临界值是 2.71，而不是 3.84。

32.6.6 随机前沿面模型分析示例

这里利用一个具体的数据集，阐述随机前沿生产函数的估计过程。随机前沿分析不能像 DEA 那样仅用含有 3~5 个数据就可分析，因为这种估计方法需要相当多的观测值，才能成功地估计出模型参数及方差。因此，在此例中考察 Whiteman 和 Pearson 所研究的数据。这些数据来自 1990 年 21 个国家的电讯供应商。Whiteman 和 Pearson(1993)利用 DEA 方法评价澳大利亚电讯相对于全球的水平。分析包括两个输入，即资本(x_1 ：由主干线的公里数度量对数值)、劳动力(x_2 ：雇员人数对数值)和一个输出(y ：收入对数值)。利用这些数据估计 Cobb-Douglas 随机前沿生产函数。

应用 DPS 进行随机前沿模型分析，其数据编辑格式：左边为投入指标，右边为产出指标。这里的例子是 21 个决策单元、2 项投入 1 项产出，数据编辑如图 32-19 所示。

	A	B	C	D	E
1	序号	x1	x2	y	
2	1	-0.2527	-0.1625	-0.3011	
3	2	-1.1323	-1.7148	--1.4271	
4	3	-0.9188	-1.3471	-1.0217	
5	4	0.425	0.0488	0.2311	
6	5	-1.2341	-1.7148	-0.9416	
7	6	-1.3205	-1.6094	-1.2379	
8	7	1.0327	0.4447	0.7227	
9	8	1.098	0.7514	0.5481	
10	9	-4.3741	-3.912	-3.912	
11	10	-2.3197	-2.0402	-2.2073	
12	11	0.8042	0.1655	0.392	
13	12	1.6721	1.0188	1.0043	
14	13	-0.3653	-1.1394	-0.2614	
15	14	-1.9153	-1.772	-1.8326	
16	15	-1.5455	-1.8971	-1.3093	
17	16	-1.4359	-1.4697	-1.6607	
18	17	0.2313	-0.2877	-0.5276	
19	18	-0.5363	-0.8675	-0.3425	
20	19	-0.9306	-1.5141	-0.5798	
21	20	-0.3721	-1.0217	-1.8971	
22	21	0.9323	0.8198	0.9282	
23					

图 32-19 随机前沿面模型分析数据格式

如图 32-19 所示选择数据，执行专业统计→数据包络分析及计量经济统计→随机前沿面模型，系统弹出用户对话框后，模型类别选“误差组分前沿面”，函数类型选“生产函数”，在“因变量是否进行了对数转换”前面打勾，再点击“确定”即可得到分析结果。

计算结果	当前日期 2007-12-1 9:27:52			
误差组分前沿面, 生产函数模型				
因变量已取对数				
OLS 估计值:				
项目	系数	标准误	T 值	P 值
β_0	-0.1637	0.1717	-0.9535	0.3530
β_1	0.6327	0.2975	2.1266	0.0475
β_2	0.2245	0.3444	0.6517	0.5228
σ -SQUARED	0.1597			
对数似然函数	-8.9186			
网格搜索估计值:				
β_0	0.3194			
β_1	0.6327			
β_2	0.2245			
σ^2	0.3703			
γ	0.9900			
Iter=0	对数似然函数=-6.111893			
0.3194	0.6327	0.2245	0.3703	0.9900
梯度步长				
Iter=5	对数似然函数=-4.780161			
0.2478	0.6269	0.1595	0.3554	1.0000
Iter=10	对数似然函数=-4.369869			
0.2145	0.6623	0.1175	0.3413	1.0000
Iter=14	对数似然函数=-3.261913			
0.1169	1.0141	-0.163	0.3213	1.0000
		3		
极大似然估计结果:				
项目	系数	标准误	T 值	p 值
β_0	0.1169	0.0689	1.6963	0.1070
β_1	1.0141	0.2479	4.0904	0.0007
β_2	-0.1633	0.1979	-0.8252	0.4200
σ^2	0.3213	0.0876	3.6696	
γ	1.0000	0.0000		
$\sigma u^2=$	0.3213			
$\sigma v^2=$	0.0000			
对数似然比函数(LLF)=-3.26191425				
单侧似然比(LR)检验偏差= 11.31334				
受限制的条件数=1 (注意: 该统计量是混合卡方分布)				
迭代次数=14				
协方差阵(略):				
技术效益估计:				
单位	效率估计值			
1	0.8284			
2	0.5087			
...				
21	0.9997			

平均效率= 0.70485

上面的输出结果，第一项是模型参数的最小二乘估计，它是在假设技术无效效应不存在(即方程 32.22 不包含 u_i 项)的前提下得到的。注意 β_1 和 β_2 (资本和劳动的弹性系数)的最小二乘估计是无偏的，但截距 β_0 以及方差参数 σ_s^2 是有偏的。尽管我们称其方差参数为“sigma-squared”，其实它就是 σ_v^2 与 σ_s^2 的和 σ_s^2 的估计。

这些最小二乘估计用作迭代过程的初值，进一步分析可得到随机前沿面模型的极大似然估计。分析结果还包括极大似然估计的中间结果，即采用格点搜寻法选择 0 和 1 之间的 γ 值，以及使对数似然函数的函数值达到最大，其格点搜索值，并用这些估计作为初值，进行迭代最大化过程。每 5 步迭代给出一次中间结果，极大似然估计列在其后。

考察 γ 的估计是 1.000 和估计标准差是 0.000 的情形。这些结果表明，残差波动的最主要原因是由技术无效效应 u_i 造成的，而随机误差 v_i 近似为 0。我们还观察到， $\gamma=0$ 的单边广义似然比检验的检验统计量是 11.3133， $p<0.0001$ 。因此，传统的平均响应函数不足以表达数据。然而，随机前沿模型与不包含随机误差的确定性前沿看上去没有显著的差异。

再下面是预测的技术效率。我们看到澳大利亚(1 号)的技术效率得分是 0.8284，超过行业平均值 0.7049，但却小于英国(21 号，0.9997)等国家的技术效率得分。

32.7 面板数据随机前沿面模型

32.7.1 面板数据随机前沿面模型

上面讨论的是用 N 个单位的截面数据来估计随机前沿面中的参数。若还能观测到这些单元连续多个时期(如几年)的数据，则称这样的数据为面板数据(panel data)。在估计随机前沿模型的参数时，面板数据优于截面数据。一般来说，用面板数据进行参数有更大的自由度。更重要的是，当技术改变由一个适当的参数模型定义且随机前沿模型中的技术无效效应是随机的，并有指定的分布时，面板数据允许同时考察与时间有关的技术改变和技术效率变化。

Pitt 和 Lee 确定了一个面板数据版的半正态模型：

$$\ln(y_{it}) = x_{it}\beta + v_{it} - u_{it}, \quad i=1,2,L,N; t=1,2,L,T \quad (32.28)$$

其中 y_{it} 表示第 i 个单位于 t 时刻的产出； x_{it} 表示一个 $(1 \times K)$ 维的输入向量； β 是一个 $(K \times 1)$ 维尚待估计的标量参数向量； v_{it} 是随机误差，相互独立且服从 $N(0, \sigma_v^2)$ ； u_{it} 是模型中的技术无效效应，与 v_{it} 相互独立。

在面板数据模型中，还可以假设 u_{it} 具有几种不同的分布形式，其基本形式是假定它们是独立同分布的随机变量，即相对于其他单位在特定时间段的观测，这个单元没有特殊的优势得到额外的观测。其次可假设技术无效效应不随时间改变，即

$$u_{it} = u_i, \quad i=1, 2, \dots, N; t=1, 2, \dots, T \quad (32.29)$$

Battese 等(1988)改进了该模型, 使 u_i 服从广义截断正态分布。Battese 等(1989)进一步扩展了模型, 且允许使用不平衡面块数据。Kumbhakar 提供了一个面板数据的随机前沿模型。在模型中, 技术无效效应随时间系统地发生变化, 即

$$u_{it} = [1 + \exp(bt + ct^2)]^{-1} u_i \quad (32.30)$$

这里假设 u_i 服从半正态分布, b 和 c 是待估的未知参数。Battese 和 Coelli(1992)建议在该模型中假设 u_{it} 是时间的指数函数, 仅含有一个未知参数, 这时面板数据的随机前沿模型方程(32.28)关于技术无效效应的时变模型可定义为

$$u_{it} = \{\exp[-\eta(t - T)]\} u_i, \quad i=1,2, \dots, N; \quad t=1,2, \dots, T \quad (32.31)$$

如前类似定义, u_i 是独立同分布的广义截断正态随机变量, η 是未知的待估标量参数。

时变模型方程(32.31)基本性质由图 32-20 给出。

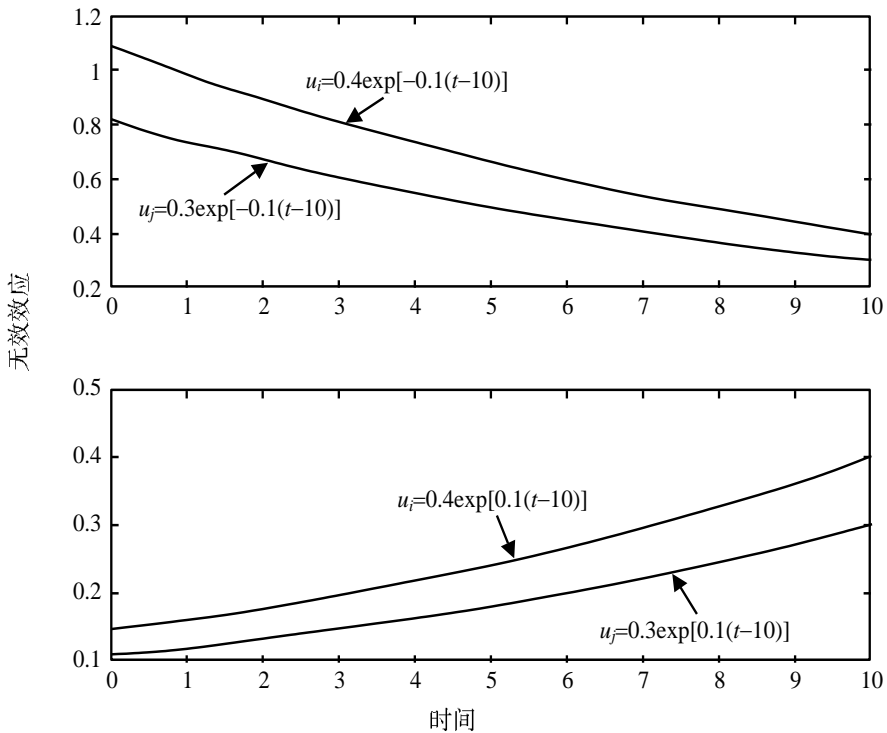


图 32-20 Battese 和 Coelli(1992)模型中的时变技术无效效应

图中描述了技术无效效应分别是递增和递减的两种情形。每一情形都有两条线, 刻画的是两个单元在不同时刻的技术无效效应值。面板最后一期($T=10$)的技术无效效应分别是 0.4 和 0.3。两种情形下的参数 η 值均指定为 0.1。

技术无效效应下降时, 两条指数曲线单调下降到面板最后一期的值, 且两指数值之差也是递减的。在技术无效效应递增时, 两条指数曲线单调上升到各自面板最后一期的值, 且两线之差也是递增的。

时变模型(32.31)的一个特例是时间不变模型, 这时 $\eta=0$ 。因此, 在给定时变模型时, 检验零假设“技术无效效应不随时间改变”, 即 $H_0: \eta=0$ 是研究要点之一。另一个有趣的复合零假设是 $H_0: \eta=\mu=0$ 。其含义是: 给定由方程(32.28)和(32.31)定义的广义前沿模型, $\eta=\mu=0$ 意味着随机前沿中的技术无效效应不随时间改变, 且服从半正态分布。

时变技术无效效应(32.31)的一个优点是: 只要在前沿函数(32.28)的向量 x_{it} 中适当地指定技术改变, 就可以区别技术无效随时间的变化和技术变化。而只有当技术无效效应是随机的且有确定的分布时, 才有可能区分技术无效随时间的变化和技术变化。

32.7.2 截断正态分布

随机前沿方法的一个弊端是对技术无效效应 u_i 分布的选择没有先验的判断准则。半正态分布和指数分布是任意选择的。由于这两个分布都在 0 点处出现众数, 这就意味着技术无效效应以很高的概率在 0 点附近取值, 这就意味着有相当高的技术效率。事实上, 可能会有几个非常有效的单元, 但大多数单元相当无效。

研究人员试图通过对技术无效效应 u_i 指定更一般的分布形式来解决这一问题。截断正态分布因具有更宽泛的分布形状(包括非 0 众数), 且计算相对较为容易, 因此目前应用较多。截断正态分布是半正态分布的推广。把均值为 μ , 方差为 σ^2 的正态分布在 0 处截去负值部分就得到截断正态分布。如果 μ 事先指定为 0, 则分布就是半正态的。根据 μ 的大小和符号, 截断正态分布将呈现不同的形状。

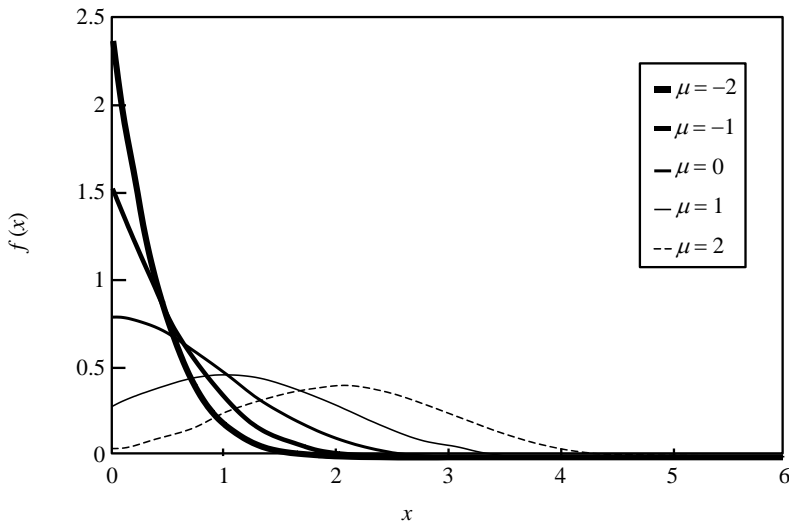


图 32-21 不同 μ 值时的截断正态分布

图 32-21 给出了当 $\sigma^2=1$, 而 $\mu=-2,-1,0,1$ 和 2 时的截断正态分布图形。截断正态随机前沿的估计不仅包括模型参数的估计, 而且还要估计参数 μ 。Stevenson 首先给出了用于求模型参数极大似然估计的对数似然函数。接着, Battese 和 Coelli(1988)给出了合理的预测单元技术效率的表达式。

32.7.3 假设检验

随机生产前沿函数的应用还需统计检验，其零假设是“给定广义截断正态分布，简单的半正态模型足以表达数据”，而这相当于利用 Wald 或广义似然比检验对零假设 $H_0: \mu=0$ 进行检验即可。

在上一节，单边广义似然比检验用于检验零假设“在半正态模型中，技术无效效应不存在”。这个检验可推广到截断正态模型中。如果零假设(模型中没有技术无效效应)是真的，那么，广义似然比统计量渐近服从混合卡方分布。混合卡方分布显著水平为 5% 的临界值是 5.138。这一值小于 χ_2^2 分布的上 5% 分位数 5.99。临界值 5.138 取自于 Kodde 和 Palm 的表 1。

32.7.4 成本函数

值得指出的是，前面介绍的所有模型都是以产出函数的形式表达的， u_i 代表技术无效性影响。此影响使得各个单位在随机产量前沿函数下操作。如果我们希望确定一个随机前沿成本函数，只需简单的把误差项从 $(v_i - u_i)$ 变化为 $(v_i + u_i)$ 。例如，这个置换可以将公式(32.28)定义的产出函数转变为成本函数：

$$\ln(y_i) = x_{it}\beta + (v_{it} + u_{it}), \quad i = 1, 2, \dots, N; t = 1, 2, \dots, T \quad (32.32)$$

在此成本函数中， u_i 定义了各个单位离成本前沿的距离，且如确定了位置效应，则 u_i 与技术无效性是一致的。如不做此假定，成本函数中的 u_i 就得不到清楚的解释，因为涉及到了技术无效性和位置无效性两个因素。

DPS 对随机前沿产量和成本函数相关参数进行最大似然估计，并可适用于面板数据中的随时间变化和不变的效率情形、面向成本或产出函数、半正态和截尾正态分布。

32.7.5 面板数据随机前沿分析中效率预测

随机前沿分析得到参数之后，可进一步预报每个单位的技术效率。相对于产出前沿的技术效率和相对于成本前沿的成本效率，其估计量可同时定义如下：

$$EFF_i^* = E(y_i^* | u_i, x_i) / E(y_i^* | u_i = 0, x_i)$$

这里 y_i^* 是第 i 个单位的产出(或成本)，当使用原始单位时， y_i^* 相当于 y_i ；当是对数形式时，就相当于 $\exp(y_i)$ 。在产出前沿例子中， EFF_i 的值在 0 和 1 之间，在成本前沿中，其值为 1 到无穷大。这些效率估量可以下列形式出现(如表 32-2)：

表 32-2 效率计算的表达形式

成本或产出	变量是否取了对数	效率(EFF _i)
产出	是	$\exp(-u_i)$
成本	是	$\exp(u_i)$
产出	否	$(x_i\beta - u_i)/(x_i\beta)$
成本	否	$(x_i\beta + u_i)/(x_i\beta)$

EFF_i 的上述四种表达形式全都依赖于无法观察的 u_i 的预报值。

32.7.6 随机前沿面模型分析示例

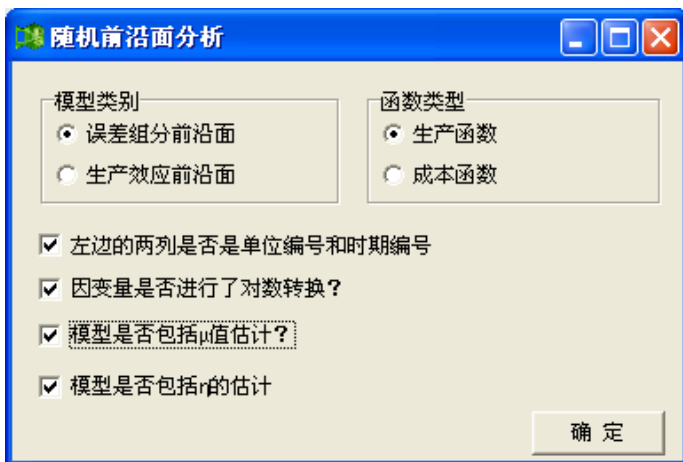
面板数据随机前沿面模型分析数据编辑格式：第 1 列是按 1,2,...,N 自然数顺序的各个企业(单位)序号，第 2 列是按 1,2,...,T 自然数顺序的时期序号。从第 3 列开始放入各项指标：左边为投入指标，右边为产出指标。这里的例子是 10 个单位、3 个时期、2 项投入 1 项产出的面板数据随机前沿面模型分析数据(图 32-22)。

	A	B	C	D	E	F	G
		DMU	t	Ln(K)	Ln(L)	Ln(GDP)	
2		1	1	8.310661	8.223895	7.531016	
3		2	1	7.662468	7.872836	7.06732	
4		3	1	7.018402	8.000685	7.06732	
5		4	1	6.806829	7.3066531	6.448889	
6		5	1	6.109248	7.533694	6.25575	
7		6	1	7.955425	8.390496	7.487734	
8		7	1	7.56008	8.386401	6.993015	
9		8	1	7.214504	7.850493	6.869014	
10		9	1	7.198931	8.104401	6.747587	
11		10	1	8.386401	8.130648	7.652546	
12		11	1	6.639876	7.70391	6.261492	
13		1	2	8.979794	8.198364	8.348064	
14		2	2	8.486734	7.882696	7.924072	
15		3	2	7.667158	8.105006	7.401231	
16		4	2	7.903966	7.391415	7.398786	
17		5	2	7.037028	7.586296	6.990257	
18		6	2	8.687106	8.446127	8.265136	
19		7	2	8.291296	8.517193	7.71913	
20		8	2	8.120291	7.869402	7.625107	
21		9	2	7.818832	8.160232	7.38709	
22		10	2	9.189117	8.226038	8.466321	
23		11	2	7.506042	7.812378	6.972606	
24		1	3	9.513994	8.191463	8.879333	
25		2	3	9.200795	7.99362	8.475954	
26		3	3	8.1898	8.136226	7.811973	
27		4	3	8.516593	7.471363	7.880426	
28		5	3	7.835975	7.586804	7.448916	
29		6	3	9.355133	8.48694	8.795734	
30		7	3	8.829812	8.619027	8.1665	
31		8	3	8.70649	7.838738	8.057694	
32		9	3	8.3676	8.165079	7.81682	
33		10	3	9.681219	8.323608	8.99293	
34		11	3	8.041091	7.863651	7.393263	
35							

图 32-22 面板数据随机前沿面模型分析数据格式

应用时应该注意，如果要求得函数形式不是线性函数形式，那么数据要做合适的变换。随机前沿分析中最常用的形式是 Cobb-Douglas 和对数转换函数。这些转换在 DPS 的电子表格里很容易实现。

数据编辑并选中之后，执行专业统计→数据包络分析及计量经济统计→面板数据的随机前沿面模型，系统弹出用户对话框如下：



在用户对话框中，用户可以根据实际情况选择相关参数。模型类别选“误差组分前沿面”，函数类型选“生产函数”，在“因变量是否进行了对数转换”前面打勾。这里，如不包括 μ 则表示假定服从半正态分布、不包括 η 表示只有唯一一个代表性数据，因此不考虑时间变化的有效性。

选定参数、执行分析计算，系统输出分析结果如下：

误差组分前沿面，生产函数模型

因变量已取对数

OLS 估计值:

项目	系数	标准误	T 值	P 值
β_0	0.9317	0.5726	1.6273	0.1141
β_1	0.8384	0.0331	25.3436	0.0000
β_2	-0.0164	0.0841	-0.1949	0.8468

σ -SQUARED 0.0183

对数似然函数 20.7899

网格搜索估计值:

β_0	1.0052
β_1	0.8384
β_2	-0.0164
σ^2	0.0220
γ	0.3850
μ	0.0000
η	0.0000

Iter=0 对数似然函数=21.330826

1.0052	0.8384	-0.0164	0.0220	0.3850	0.0000	0.0000
梯度步长						
Iter=5	对数似然函数=23.276274					
1.0337	0.8483	-0.0241	0.0158	0.3533	0.0339	0.4459
Iter=10	对数似然函数=24.723658					
1.0572	0.8197	0.0001	0.0079	0.1002	0.0517	0.5970
Iter=15	对数似然函数=26.079181					
0.8538	0.7819	0.0668	0.0071	0.1239	0.0595	0.6164
极大似然估计结果:						
项目	系数	标准误	T 值	p 值		
β_0	0.8538	0.6611	1.2914	0.2064		
β_1	0.7818	0.0341	22.9356	0.0000		
β_2	0.0668	0.0979	0.6825	0.5001		
σ^2	0.0071	0.0026	2.7791			
γ	0.1239	0.1063	1.1653			
$\sigma_u^2=$	0.0009					
$\sigma_v^2=$	0.0063					
μ	0.0595	0.0306	1.9426			
η	0.6164	0.2029	3.0385			
对数似然比函数(LLF)=26.07918114						
单侧似然比(LR)检验偏差=10.57858						
受限制的条件数=3 (注意: 该统计量是混合卡方分布)						
迭代次数=15						
横截面数=11						
时期长度=3						
协方差阵:						
0.4371	0.0044	-0.0600	0.0001	-0.0368	-0.0020	0.0040
0.0044	0.0012	-0.0018	0.0001	-0.0001	-0.0002	-0.0026
-0.0600	-0.0018	0.0096	-0.0001	0.0047	0.0006	0.0022
0.0001	0.0001	-0.0001	0.0000	0.0001	-0.0000	-0.0003
-0.0368	-0.0001	0.0047	0.0001	0.0113	-0.0003	-0.0050
-0.0020	-0.0002	0.0006	-0.0000	-0.0003	0.0009	-0.0035
0.0040	-0.0026	0.0022	-0.0003	-0.0050	-0.0035	0.0412
技术效益估计:						
效率估计(年份 1):						
单位	时期 1	时期 2	时期 3	地区		
1	0.7696	0.8678	0.9262	江苏		
2	0.7754	0.8713	0.9282	浙江		
...		
11	0.7220	0.8384	0.9091	广西		
平均效率	0.80538	0.88867	0.93795			

参 考 文 献

- Ali A I and Seiford L M. 1993. The mathematical programming approach to efficiency analysis//Fried HO, CAK Lovell and Schmidt SS (Eds.). The Measurement of Productive Efficiency: Techniques and Applications[M]. New York: Oxford University Press, 120-159.
- Andersen P and Petersen N. 1993. A procedure for ranking efficient units in data envelopment analysis[J]. Management Science, 39: 1261-1264.
- Balk B M. 2001. Scale efficiency and productivity change[J]. Journal of Productivity Analysis, 15: 159-183.

- Banker R D, Charnes A and Cooper W W. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis[J]. *Management Science*, 30: 1078-1092.
- Banker R D and Thrall R M. 1992. Estimation of returns to scale using data envelopment analysis[J]. *European Journal of Operational Research*, 62: 74-84.
- Battese G E. 1992. Frontier production functions and technical efficiency: a survey of empirical applications in agricultural economics[J]. *Agricultural Economics*, 7: 185-208.
- Battese G E and Coelli T J. 1988. Prediction of firm-level technical efficiencies with a generalised frontier production function and panel data[J]. *Journal of Econometrics*, 38: 387-399.
- Battese G E and Coelli T J. 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in india[J]. *Journal of Productivity Analysis*, 3: 153-169.
- Battese G E and Coelli T J. 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data[J]. *Empirical Economics*, 20: 325-332.
- Coelli T J, Rao D S P and Battese G E. 2005. *An Introduction to Efficiency and Productivity Analysis*[M]. New York: Springer.
- Cooper W W, Seiford L M and Tone K. 2007. *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, and DEA-Solver Software*[M]. Boston: Kluwer Academic Publishers.
- Cooper W W, Seiford L M and Zhu J. 2004. *Handbook on Data Envelopment Analysis*[M]. Boston: Kluwer Academic Publishers.
- Färe R and Lovell C A K. 1978. Measuring the technical efficiency of production[J]. *Journal of Economic Theory*, 19 : 150-162.
- Färe R, Grosskopf S, Norris M and Zhang Z. 1994. Productivity growth, technical progress, and efficiency changes in industrialised countries[J]. *American Economic Review*, 84: 66-83.
- Farrell M J. 1957. The measurement of productive efficiency[J]. *Journal of the Royal Statistical Society*, 120: 253-281.
- Tone K. 2001. A slacks-based measure of efficiency in data envelopment analysis[J]. *European Journal of Operational Research*, 130: 498-509.
- Tone K. 2002. A slacks-based measure of super-efficiency in data envelopment analysis[J]. *European Journal Operational Research*, 143: 32-41.

第 33 章 量表分析和顾客满意指数模型

33.1 项目分析

测验项目分析的主要内容就是测验题目的难度及区分度，也包括对多重选择题的项目分析等。项目分析的主要作用就是提供给被试者最合适的测验题目，使测量的结果能够达到测量的目的，同时又能够符合被试的需要。在测验编制中，项目分析是一个主要的任务。

33.1.1 测验题目的难度

测量中题目或项目的难度是指被试者完成题目或项目任务时所遇到困难的程度，可分为绝对难度和相对难度两种。绝对难度又称认知难度，是指项目本身固有的难度；相对难度又称统计难度，是学生实测后经计算得到的难度指标。定量刻画被试者作答一个题目所遇到的困难程度的量数，就叫题目的难度系数，也常称为难度值，用符号 P 表示。测验难度的计算，本质上是正确作答人数的比例，即 $P=R/N$ ，式中 P 代表项目难度， R 为答对人数， N 为被试人数。

33.1.2 测验题目的区分度

区分度是指项目对被试者特性的区分程度，是区别被试水平能力的量度，常记为 D 。区分度高的项目能有效地区分水平不同能力的被试者。因此，区分度是测验是否具有效度的指示器，并作为评价项目质量、筛选项目的主要指标与依据，其计算方法主要有相关法。

当测量分数为一个连续变量(百分制分数)，被试者为两分变量(及格与否)时，多采用二列相关，二列相关适用于两个变量都是正态连续变量，计算公式为

$$r_b = \frac{\bar{x}_p - \bar{x}_q}{S_t} \cdot \frac{p}{y}$$

式中， r_b 为二列相关系数， p 为通过该项目的人数百分比， q 为未通过该项目被试人数的百分比， \bar{x}_p 为通过该项目被试的平均效标分数， \bar{x}_q 为未通过该项目被试的平均效标分数， S_t 为全体被试者的效标分数的标准差， y 为 p 与 q 交界处正态曲线的高度，可通过正态分布表查得。

如果测验总分与项目分数均为两分变量，则可采用点二列相关计算项目的区分度。计算公式为

$$r_b = \frac{\bar{x}_p - \bar{x}_q}{S_t} \sqrt{pq}$$

判断项目区分度高低的经验性标准很多。一般认为二列相关系数大于 0.25 时，该项目便具有良好的区分度。项目的区分度具有相对性，它与难度关系，从理论上说，中等难度($p = 0.5$)的项目具有较好的区分度；难度接近 1 或 0 时，区分度会趋向为小，即过易或过难的项目区分度一般都不高，鉴别指数的最大值与项目难度的关系如下：

项目通过率 p	1	0.9	0.7	0.6	0.5	0.4	0.3	0.1	0.0
D 的最大值	0	0.2	0.6	0.8	1	0.8	0.6	0.2	0

33.1.3 在 DPS 系统中进行项目分析(item analysis)

如某项测验，有 10 个多选题，参加测验的有 30 人。经测试，30 名被测试者的结果如图 33-1 所示，试用项目分析方法进行分析。分析前将数据编辑如图 33-1 格式。

	A	B	C	D	E	F	G	H	I	J	K
1	No.	I1	I2	I3	I4	I5	I6	I7	I8	I9	
2	1	3	3	3	3	3	4	4	2	3	
3	2	3	3	4	4	1	1	1	1	4	
4	3	3	3	4	4	1	1	1	1	4	
5	4	3	3	1	1	3	4	4	1	4	
6	5	2	4	1	2	4	4	4	4	4	
7	6	3	3	1	1	3	4	4	1	4	
8	7	3	4	3	3	3	3	3	2	3	
9	8	4	3	4	4	4	4	4	3	4	
10	9	4	3	4	4	3	4	1	1	4	
11	10	3	3	3	2	3	4	3	2	3	
12	11	4	2	1	1	3	3	4	1	1	
13	12	3	3	3	3	3	3	3	2	4	
14	13	1	4	3	3	3	4	4	3	4	
15	14	2	3	1	1	4	3	3	1	4	
16	15	4	4	4	4	4	4	4	4	4	
17	16	2	3	1	1	4	3	3	1	4	
18	17	1	1	1	1	1	1	2	2	1	
19	18	3	4	3	2	4	2	1	2	4	
20	19	2	3	2	2	3	3	1	1	3	
21	20	3	4	3	4	4	0	1	1	4	
22	21	3	4	1	1	4	4	2	1	3	
23	22	4	1	4	4	1	4	4	1	4	
24	23	3	4	1	4	3	1	1	1	2	
25	24	3	4	1	4	4	1	1	1	2	
26	25	4	3	4	4	2	1	1	1	4	
27	26	4	4	4	4	4	4	4	4	4	
28	27	1	1	4	4	1	4	1	1	4	
29	28	4	3	4	3	1	3	1	1	4	
30	29	2	3	3	2	1	1	4	4	2	
31	30	4	4	1	1	4	4	4	1	1	
32											

图 33-1 项目分析数据编辑格式

然后执行“专业统计”下面“量表分析及顾客满意度模型”中的“量表项目分析”功能，这时系统出现对话框，确认是否输出详细结果。确认后系统给出分析结果：

含缺失调查项目样本：				
id	R	O	U	I

20	5	0	0	1								
样本数=	30	均值=	4.7667	方差=	2.5789	标准差=	1.6059					
Items=	10	Alpha	0.0766	SEM	1.5432							
P(正确%)	均值=0.4767	RPBI(点二列相关)	均值=0.0282	RBIS(二列相关)	均值=0.0358							
项目	反应值	Q1	Q2	Q3	Q4	Q5	和	均值	PROP	RPBI	RBIS	
Item1	[1]	1	0	1	1	0	3	3.3333	0.1000	-0.2113	-0.3613	
	[2]	1	1	2	0	1	5	4.0000	0.1667	-0.0945	-0.1409	
	[3+]	4	4	2	1	2	13	5.1538	0.4333	-0.0995	-0.1253	
	[4]	0	1	1	4	3	9	5.1111	0.3000	0.3227	0.4253	
	和	6	6	6	6	6	30	4.7667				
项目	反应值	Q1	Q2	Q3	Q4	Q5	和	均值	PROP	RPBI	RBIS	
Item2	[1]	1	0	0	2	0	3	3.3333	0.1000	-0.2039	-0.3486	
	[2]	0	1	0	0	0	1	5.0000	0.0333	0.0892	0.2159	
	[3+]	2	4	2	2	5	15	5.2667	0.5000	0.0000	0.0000	
	[4]	3	1	4	2	1	11	4.4545	0.3667	0.0937	0.1199	
	和	6	6	6	6	6	30	4.7667				
项目	反应值	Q1	Q2	Q3	Q4	Q5	和	均值	PROP	RPBI	RBIS	
Item3	[1+]	2	2	3	1	3	11	5.1818	0.3667	-0.1049	-0.1342	
	[2]	1	0	0	0	0	1	3.0000	0.0333	-0.1642	-0.3974	
	[3]	3	2	2	0	1	8	4.0000	0.2667	-0.1524	-0.2050	
	[4]	0	2	1	5	2	10	5.1000	0.3333	0.3126	0.4053	
	和	6	6	6	6	6	30	4.7667				
...												
项目	反应值	Q1	Q2	Q3	Q4	Q5	和	均值	PROP	RPBI	RBIS	
Item9	[1]	1	1	0	0	1	3	3.3333	0.1000	-0.1863	-0.3185	
	[2]	1	0	0	1	1	3	3.6667	0.1000	-0.1087	-0.1858	
	[3]	3	1	1	0	0	5	4.0000	0.1667	-0.0417	-0.0621	
	[4+]	1	4	5	5	4	19	5.3684	0.6333	0.2158	0.2763	
	和	6	6	6	6	6	30	4.7667				
项目	反应值	Q1	Q2	Q3	Q4	Q5	和	均值	PROP	RPBI	RBIS	
Item10	[1]	1	0	0	1	0	2	3.0000	0.0667	-0.2132	-0.4115	
	[2]	2	0	0	2	0	4	4.0000	0.1333	-0.0447	-0.0706	
	[3]	1	3	0	0	2	6	4.1667	0.2000	0.0000	0.0000	
	[4+]	2	3	6	3	4	18	5.3333	0.6000	0.1396	0.1770	
	和	6	6	6	6	6	30	4.7667				
分位数范围		Q1	Q2	Q3	Q4	Q5						
低, 高		1,4	4,5	5,5	5,5	5,9						

分析结果。首先是汇总后的 10 个多选题平均值, 然后是每个多选题的计算结果(正确答案项的编号后带“+”号, 并以粗体显示)。结果解释集中在答对的测试者, 并注意, 难度系数(PROP)实际上是答对人数除以总人数, 如本例的第 10 个项目, 答对的有 18 个人, 除以总数 30, 其难度系数就等于 0.6。注意难度系数越大, 该项目就越容易。

分辨度。这里有点二列相关(RPBI)和二列相关(RBIS)两个指标, 用户可根据前面的介绍酌情选用并解释。如本例的第 9 个项目, 其点二列相关系数等于 0.2158, 二列相关等于 0.2763, 其难度是合适的。分辨系数中的点二列相关取值是从 -1 到 +1, 当其值大于 0 时, 得分越高, 其分辨效果越好。相反, 如果点二列相关系数小于 0, 其负数数字越大, 则项目就越不好(不正确的频率增大), 这样的项目还容易误导。

33.2 量表可信度分析

信度(reliability)指的是测量结果的稳定性程度,如果用同一测量工具反复测量同一种特质对象,则多次测量结果间的一致性程度就叫信度。如果每次测量结果中误差部分都很小,那么测量结果必然是稳定的。但是由于测试对象的特殊性,出现测量误差的可能性更大,如施测环境、完成时限、主被试关系、被试的动机和情绪等因素都可能影响到测量的结果。在测试中,各方面偶然误差因素控制得好,多次测试所得分数的一致性就高,测量信度也就高,测验就比较可靠。

33.2.1 信度指标

信度是衡量一个测验量表质量高低的重要指标。目前一般应用同质性信度也叫内部一致性信度(internal consistency reliability),是指测验内部所有题目间的一致性程度,包括两层意思:一是所有题目测的是同一种心理特质;二是指所有题目得分之间都具有较高的正相关。同质性信度就是指一个测验所测或特质的相同程度。同质性信度估计参数主要有:

1. 分半信度(split-half reliability)

分半信度是指将一个测验分成对等两半后,所有被试在这两半上所得分数的一致性程度。这是内部一致性系数的一种粗略估计方法,分半信度反映的是测验内部两半题目间的一致性,属于同质性信度。如果把对等的两半测验看成是在最短时间内施测的两个平行测验,也可以将分半信度和等值性系数一样解释。

分半信度计算的是两个“半测验”上得分的相关系数,只是半个测验的信度,还必须用斯皮尔曼-布朗公式加以校正

$$r_{xx} = \frac{2r_{hh}}{1 + r_{hh}}$$

式中, r_{xx} 为整个测验的信度系数; r_{hh} 为两个“半测验”得分的相关系数。分半信度通常是在只能施测一次或没有复本的情况使用。测验分成相等的两半的方法很多,如按题号的奇偶分半、按题目的难度分半、按题目的内容分半等。因此,同一个测验通常会因为分半方式不同而产生多个分半信度值。在测验题目分半后一般都应考察分半的情况,看是否需要作适当调整,分半最终的目标是要将题目分成对等的两半。如果一个测验无法分成对等的两半,使用分半信度就存在困难。在实际中一般是采用奇偶分半并使两半测验的题目在难度上基本相等。

由于分半方法的多样,所得结果不稳定,故有提议:计算所有可能的分半信度,并用其平均值作为内部一致性的估计值。由此提出的做法最后导出了两个十分方便的信度系数计算公式。

2. 库德-理查逊信度(Kuder & Richardson reliability)

该方法适合于测验题目全部为二分(0, 1)记分题(dichotomously scoring)的测验的内部一致性信度分析。常用的有 KR_{20} 和 KR_{21} 公式

$$KR_{20} = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum p_i q_i}{ST^2} \right)$$

式中, r_{xx} 为测验的信度, k 为题目数, p_i 和 q_i 分别表示答对和答错第 i 题的被试人数比例, ST^2 为测验总分的方差。

$$KR_{21} = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum pq}{ST^2} \right)$$

式中, k 为题目数, pq 分别表示题目的平均通过率和失败率, ST^2 为测验总分的方差。此公式只有当所有题目的难度接近时才适用。

3. 克龙巴赫(Cronbach) α 系数

当测验题型较多, 并非都是二分记分题时, 估计测验信度可采用克龙巴赫 α 系数, 这个公式可以处理任何测验的内部一致性系数的计算问题。 KR_{20} , KR_{21} 只是这个公式计算的特例。但要注意的是, α 值大, 必然测量信度高; α 值小时, 却不能断定测量信度不高。其计算公式为

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum S_i^2}{ST^2} \right)$$

其中, k 为量表中题项的总数, S_i^2 为第 i 题得分的题内方差, ST^2 为全部题项总得分的方差。从公式中可以看出, α 系数评价的是量表中各题项得分的一致性, 属于内在一致性系数。这种方法适用于态度、意见式问卷(量表)的信度分析。

4. 荷伊特信度

1941 年荷伊特(Hoyt)提出用方差分量比来估计测验内部一致性的方法: 设有 n 名被试者参加一个有 k 个题目的测验, 测验分数总变异可以分解为被试间差异 $SS_{人}$, 项目间变异 $S_{测题}$ 和人与试题交互作用 $SS_{人 \times 测题}$ 。荷伊特认为可用 $MS_{人}$ 作为被试方差的估计值, 用 $MS_{人 \times 测题}$ 误差方差估计值, 则测验信度的估计值就为

$$r_{xx} = 1 - MS_{人 \times 测题} / MS_2$$

另外, 肯德尔和谐系数(第 8 章)也是信度的一个估计。肯德尔和谐系数是表示多列等级变量相关程度的。

信度是指测验所得到结果的一致性或稳定性, 而非测验或量表本身。信度值是指在某一特定类型下的一致性, 信度系数会因不同时间、不同受试者或不同评分者而出现不

同的结果；目前信度用的多是 Cronbach α 系数。信度系数越大，表明测量的可信程度越大。一般来说，一份量表或问卷，其信度系数最好在 0.80 以上，0.70 至 0.80 之间还算是可以接受的范围；分量表最好在 0.70 以上，0.60 至 0.70 之间可以接受。若分量表的内部一致性系数在 0.60 以下或者总量表的信度系数在 0.80 以下，应考虑重新修订量表或增删题项。

33.2.2 分半信度、Cronbach α 系数及荷伊特信度分析

如某项测试有 8 个题项，参加测验的有 30 人。经测试，30 名被测试者的结果如图 33-2 所示，试对该测试项目的信度指标进行分析。分析前将数据编辑如图 33-2 格式。

	A	B	C	D	E	F	G	H	I	J
1	No.	x1	x2	x3	x4	x5	x6	x7	x8	
2	1	4	3	4	4	4	4	4	4	
3	2	3	3	3	3	4	4	4	3	
4	3	4	3	4	4	4	4	4	4	
5	4	4	3	4	4	4	4	4	4	
6	5	4	4	4	2	3	3	4	4	
7	6	4	4	4	3	3	3	4	4	
8	7	4	3	3	4	4	4	4	4	
9	8	4	4	4	4	4	4	4	4	
10	9	3	3	3	4	4	4	4	4	
11	10	4	4	4	4	4	3	4	3	
12	11	2	2	2	2	2	2	2	2	
13	12	3	3	3	3	3	3	3	3	
14	13	3	3	2	3	3	3	3	3	
15	14	1	1	4	4	4	4	4	3	
16	15	1	1	4	4	3	3	3	2	
17	16	4	4	4	4	4	4	4	4	
18	17	4	4	3	3	3	4	4	4	
19	18	3	3	3	4	4	4	4	4	
20	19	3	3	4	3	4	3	3	3	
21	20	2	2	2	3	3	3	3	3	
22	21	4	4	4	1	4	1	4	4	
23	22	4	4	4	4	1	4	4	4	
24	23	4	4	4	4	4	4	4	4	
25	24	3	2	2	4	4	4	4	4	
26	25	4	4	4	4	3	3	4	4	
27	26	4	4	4	4	2	4	4	4	
28	27	4	4	4	4	3	3	3	4	
29	28	4	3	4	4	4	3	4	3	
30	29	3	3	3	4	4	3	3	3	
31	30	4	4	4	4	4	4	4	4	
32										

图 33-2 信度计算数据编辑格式

然后执行“专业统计”下面“量表分析及顾客满意度模型”中的“分半信度及 Cronbach α 系数”计算功能，这时系统输出详细结果如下：

计算结果	当前日期 2006-3-17 23:06:07		
项目	平均值	方差	标准差(略)

全部项目的 Cronbach $\alpha=0.7841$			标准化 $\alpha=0.8015$		
去掉因子与剩下全部因子			剩余因子		
去掉	相关系数 r		Cronbach α		
x1	0.6640		0.7269		
x2	0.5034		0.7593		
x3	0.4640		0.7644		
x4	0.3430		0.7843		
x5	0.1777		0.8098		
x6	0.4315		0.7694		
x7	0.7616		0.7322		
x8	0.7498		0.7252		
相关系数矩阵(略)					
依前后分 2 组:		Split-half r=0.56713	Spearman-Brown r=0.72378		
奇-偶间隔分 2 组:		Split-half r=0.67718	Spearman-Brown r=0.80753		
Hoyt 可靠性方差分析表					
变异来源	D.F.	SS	MS	F	PROB
个体	29	52.3500	1.8052	4.6318	0.0000
个体内	210	83.5000	0.3976		
测量	7	4.3833	0.6262	1.6067	0.1352
误差	203	79.1167	0.3897		
总的	239	135.8500	0.5684		
Hoyt 未调整可靠性检验=0.7797 SE(测度)=1.7835					
Hoyt 调整后可靠性检验=0.7841 SE(测度)=1.7658					
Hoyt 未调整项目可靠性=0.3068 SE(测度)=3.1641					
Hoyt 调整后项目可靠性=0.3122 SE(测度)=3.1516					
2 个项目时 Cronbach $\alpha=0.9450$ 均值=6.6000					方差=3.0069
Item	Mean	Variance	Pt.Bis.r		
1	3.4000	0.8000	0.9738		
2	3.2000	0.7862	0.9733		
3 个项目时 Cronbach $\alpha=0.9031$ 均值=10.1667					方差=4.9713
Item	Mean	Variance	Pt.Bis.r		
1	3.4000	0.8000	0.9683		
2	3.2000	0.7862	0.9419		
8	3.5667	0.3920	0.8440		
4 个项目时 Cronbach $\alpha=0.8808$ 均值=13.8667					方差=6.6713
Item	Mean	Variance	Pt.Bis.r		
1	3.4000	0.8000	0.9493		
2	3.2000	0.7862	0.9004		
8	3.5667	0.3920	0.8800		
7	3.7000	0.2862	0.7187		
5 个项目时 Cronbach $\alpha=0.8612$ 均值=17.3667					方差=8.9989
Item	Mean	Variance	Pt.Bis.r		
1	3.4000	0.8000	0.9202		

2	3.2000	0.7862	0.8790	
8	3.5667	0.3920	0.8403	
7	3.7000	0.2862	0.7370	
3	3.5000	0.5345	0.6525	
6个项目时	Cronbach $\alpha=0.8249$	均值=20.8000		方差=10.6483
Item	Mean	Variance	Pt.Bis.r	
1	3.4000	0.8000	0.8790	
2	3.2000	0.7862	0.8128	
8	3.5667	0.3920	0.8676	
7	3.7000	0.2862	0.7941	
3	3.5000	0.5345	0.6215	
6	3.4333	0.5299	0.4587	
7个项目时	Cronbach $\alpha=0.8098$	均值=24.3333		方差=12.8506
Item	Mean	Variance	Pt.Bis.r	
1	3.4000	0.8000	0.8174	
2	3.2000	0.7862	0.7269	
8	3.5667	0.3920	0.8348	
7	3.7000	0.2862	0.7911	
3	3.5000	0.5345	0.6184	
6	3.4333	0.5299	0.5770	
4	3.5333	0.6023	0.5040	
8个项目时	Cronbach $\alpha=0.7841$	均值=27.8000		方差=14.4414
Item	Mean	Variance	Pt.Bis.r	
1	3.4000	0.8000	0.7852	
2	3.2000	0.7862	0.6672	
8	3.5667	0.3920	0.8174	
7	3.7000	0.2862	0.8175	
3	3.5000	0.5345	0.6082	
6	3.4333	0.5299	0.5809	
4	3.5333	0.6023	0.5168	
5	3.4667	0.6023	0.3718	

用户注意，结果的后面部分是当取不同数量项目(item)得到的克隆巴赫 α 系数达到最大(一致性最好)的“最优”组合。在每个组合中，将每个项目按点二列相关系数(Pt.Bis.r)大小列出。如果是进行的初步调查分析，由此可以从初步拟定的项目中“筛选”出信度较好的项目进行组合，有助于进一步开展调查研究。这可以应用于下一节介绍的顾客满意度模型的研究工作中。

33.2.3 库德-理查逊信度

库德-理查逊信度是针对测验项目是 2 分法(0,1)类型时应用。例如，上面的测试试题中，选“4”为正确，记为“1”，选其他为错，记为“0”，这时 8 个题项，参加测验 30 人，测试得分结果在 DPS 电子工作表中如图 33-3 方式编辑，便可对该测试项目进行库

德-理查逊信度计算。

	A	B	C	D	E	F	G	H	I	J
1		1	0	1	1	1	1	1	1	
2		0	0	0	0	1	1	1	0	
3		1	0	1	1	1	1	1	1	
4		1	0	1	1	1	1	1	1	
5		1	1	1	0	0	0	1	1	
6		1	1	1	0	0	0	1	1	
7		1	0	0	1	1	1	1	1	
8		1	1	1	1	1	1	1	1	
9		0	0	0	1	1	1	1	1	
10		1	1	1	1	1	0	1	0	
11		0	0	0	0	0	0	0	0	
12		0	0	0	0	0	0	0	0	
13		0	0	0	0	0	0	0	0	
14		0	0	1	1	1	1	1	0	
15		0	0	1	1	0	0	0	0	
16		1	1	1	1	1	1	1	1	
17		1	1	0	0	0	1	1	1	
18		0	0	0	1	1	1	1	1	
19		0	0	1	0	1	0	0	0	
20		0	0	0	0	0	0	0	0	
21		1	1	1	0	1	0	1	1	
22		1	1	1	1	0	1	1	1	
23		1	1	1	1	1	1	1	1	
24		0	0	0	1	1	1	1	1	
25		1	1	1	1	0	0	1	1	
26		1	1	1	1	0	1	1	1	
27		1	1	1	1	0	0	0	1	
28		1	0	1	1	1	0	1	0	
29		0	0	0	1	1	0	0	0	
30		1	1	1	1	1	1	1	1	
31										

图 33-3 KR20, KR21 系数计算数据格式

分析前将数据编辑如图 33-3 格式。然后执行“专业统计”下面“量表分析及顾客满意度模型”中的“Kuder&Richardson 信度计算”功能，这时系统给出计算结果为： $KR_{20}=0.8081$ ， $KR_{21}=0.7979$ 。

33.3 顾客满意指数模型

顾客满意指数是通过一定的测量模型测量的，其具体结构可以随目的、行业不同而不同。模型中的满意指数是最终要得到的目标变量，产品形象、预期质量、感知质量和感知价值是顾客满意指数的原因变量，顾客忠诚则是顾客满意指数的结果变量。该模型对结构变量选择和结构变量之间关系的定义可以有助于建立宏观层次上的顾客满意指标体系，进行行业间顾客满意比较，以及分析顾客满意指数影响因素以及顾客满意指数同顾客忠诚之间的关系。

33.3.1 顾客满意指数模型分析原理

近年来，对顾客满意指数建模分析越来越多地借助于结构方程模型来进行。结构方

程模型(SEM)是应用统计领域近来发展迅速的一个分支,并被广泛应用于心理学、社会学等领域,尤其是顾客满意指数(CSI)分析模型(Claes Fornel)(Michael D Johnson et al., 1996)。关于 SEM 的数值分析方法,主要有 ML 算法、EM 算法(Liang et al., 2004)等,但实际应用中 PLS 方法仍是有效的方法(Michel et al., 2005)。例如,在中国顾客满意指数模型研究中,顾客满意指数基本模型是包含形象、预期质量、感知质量、感知价值、顾客满意度和顾客忠诚 6 个结构变量的结构方程模型,模型中的各个结构变量和相关关系见图 33-4。

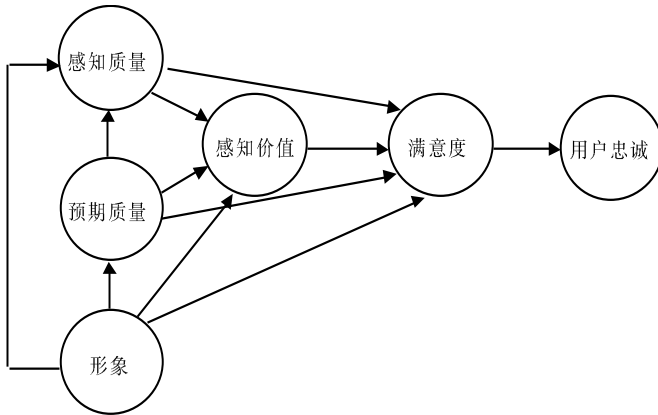


图 33-4 中国顾客满意指数基本模型结构

上述基本模型中的变量关系也可以用线性方程组形式来表达。其中结构变量之间关系可以表达为两个方程组,一个是结构变量之间的关系方程组,称为结构方程组;一个是结构变量与观测变量之间的关系方程组,称为观测方程组。在典型的顾客满意指数模型里,结构方程组包含 6 个结构变量(隐含变量) $\xi_1、\eta_1 \sim \eta_5$ 与 11 个关系(自变量作用的关系为 $\gamma_1 \sim \gamma_4$, 因变量之间的作用关系为 β_{ij}),其矩阵表达方式如下:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 0 & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 & 0 & 0 \\ \beta_{41} & \beta_{42} & \beta_{43} & 0 & 0 \\ 0 & 0 & 0 & \beta_{54} & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{pmatrix} + \begin{pmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{14} \\ 0 \end{pmatrix} \xi_1 + \begin{pmatrix} \varepsilon_{\eta_1} \\ \varepsilon_{\eta_2} \\ \varepsilon_{\eta_3} \\ \varepsilon_{\eta_4} \\ \varepsilon_{\eta_5} \end{pmatrix}$$

式中, γ 表示形象, η_1 表示预期质量, η_2 表示感知, η_3 表示感知价值, η_4 表示用户满意指数, η_5 表示用户忠诚。 $\varepsilon_{\eta_1} \sim \varepsilon_{\eta_5}$ 为误差项。

一般说来,结构变量不一定是 5 个,结构方程系数形式除了要求对角线是 0 外,下三角矩阵元素的定义也不相同。设因变量有 m 个,将 η_1, \dots, η_m 排成列向量,记为 η ; 自变量有 k 个,将 ξ_1, \dots, ξ_k 排成列向量,记为 ξ ; η 的系数矩阵为 m 阶方阵记为 B , ξ 的系数矩阵为 $m \times k$ 阶矩阵记为 Γ , 残差向量为 ε_η , 则结构方程组可以扩展为

$$\eta = B\eta + \Gamma\xi + \varepsilon_\eta \quad (33.1)$$

SEM 的结构变量是隐含的,不能直接观测,而是对应若干个观测变量。设一共有 M 个观测变量,对每一个观测变量有 N 个观测,在顾客满意指数分析中就是有 N 个顾客的测评,这样我们手里数据是一个 $N \times M$ 矩阵。

结构变量与观测变量之间的作用关系也可以用方程表示出来,按作用的因果路径有两种表示方式。设与自变量 ξ_t 对应的 $S(t)$ 个观测变量为 $x_{tj} (j=1, \dots, S(t))$, 与因变量 η_i 对应的 $L(i)$ 个观测变量为 $y_{ij} (j=1, \dots, L(i))$, 于是从观测变量到结构变量的观测方程组可以表达为

$$\xi_t = \sum_{j=1}^{S(t)} \psi_{tj} x_{tj} + \varepsilon_{\xi t}, \quad t=1, \dots, k \quad (33.2)$$

$$\eta_i = \sum_{j=1}^{L(i)} \omega_{ij} y_{ij} + \varepsilon_{\eta i}, \quad i=1, \dots, m \quad (33.3)$$

反之,从结构变量到观测变量的观测方程可以表达为

$$\begin{pmatrix} x_{t1} \\ \mathbf{M} \\ x_{tS(t)} \end{pmatrix} = \begin{pmatrix} \nu_{t1} \\ \mathbf{M} \\ \nu_{tS(t)} \end{pmatrix} \xi_t + \begin{pmatrix} \varepsilon_{xt1} \\ \mathbf{M} \\ \varepsilon_{xtS(t)} \end{pmatrix}, \quad t=1, \dots, k \quad (33.4)$$

$$\begin{pmatrix} y_{i1} \\ \mathbf{M} \\ y_{iL(i)} \end{pmatrix} = \begin{pmatrix} \lambda_{i1} \\ \mathbf{M} \\ \lambda_{iL(i)} \end{pmatrix} \eta_i + \begin{pmatrix} \varepsilon_{yi1} \\ \mathbf{M} \\ \varepsilon_{yiL(i)} \end{pmatrix}, \quad i=1, \dots, m \quad (33.5)$$

其中, ν_{tj} , λ_{ij} 为载荷项。

我们将式(33.1)~式(33.3)合在一起(或者式(33.1)、式(33.4)、式(33.5)合在一起)称为结构方程模型,有时也称为路径分析模型。

目前,对顾客满意指数的结构方程模型求解一般采用偏最小二乘(PLS)迭代算法。PLS 方法是先在式(33.2)、(33.3)给 ψ_{tj}, ω_{ij} 任意初值,于是可以根据已知的观测变量 x_{tj}, y_{ij} 的值,利用公式(33.2)、(33.3)求出结构变量 ξ_t, η_i 的值;结构方程的变量 ξ_t, η_i 有了数值,就可以对方程组(33.1)求最小二乘解,于是有了系数 γ_{tj}, β_{ij} 的估计值,进而有了变量 ξ_t, η_i 的估计值 $\hat{\xi}_t, \hat{\eta}_i$ 。结构变量 ξ_t, η_i 与观测变量 x_{tj}, y_{ij} 都有了数值,就可以对方程组(33.2), (33.3)求解,此时的 ψ_{tj}, ω_{ij} 就有了估计值,而不再是任给的初值。 ψ_{tj}, ω_{ij} 有了估计值,就可以回到迭代起点,开始新一轮迭代。迭代收敛控制可以选取迭代过程中所有对应元素的差值小于指定的误差精度。上述迭代过程可以表示为

$$(\psi_{tj}, \omega_{ij})^{(0)} \xrightarrow{(13.3)(13.4)} (\hat{\xi}_t^{(外生)}, \hat{\eta}_i^{(外生)})^{(0)} \xrightarrow{(13.2)} (\gamma_{tj}, \beta_{ij})^{(0)}$$

$$\xrightarrow{(13.2)} \left(\beta_{ij}, \gamma_{ij} \right) \xrightarrow{(13.3)(13.4)} (\psi_{ij}, \omega_{ij})^{(1)}$$

但是，这里给出的迭代初值是最小二乘意义下最优的。SEM 的最后解要满足两个方程组，其中结构方程组在开始时其变量与系数都是未知的，不可能计算出迭代初值。我们根据观测方程组计算出来的迭代初值在最小二乘意义下已经满足了这个方程组，当然就是最小二乘意义下最优的。当结构方程组根据这个初值计算出系数时，结构方程组也就满足了。但由于在第一次计算时没有体现结构变量之间的相互作用，所以需要继续迭代，直到两个方程组都被满足，特别是结构方程组内部平衡时为止。

得到模型参数估计值之后，我们就可以由结构变量(潜变量)估计值来计算顾客满意指数(CSI)，Fornel 提出的计算公式如下：

$$CSI = \frac{E(\xi) - \min(\xi)}{\max(\xi) - \min(\xi)} \times 100$$

式中， ξ 是顾客满意指数模型里的潜变量， $\max(\xi)$ ， $\min(\xi)$ 和 $E(\xi)$ 分别是潜变量的最大、最小和期望值，这里 $\max(\xi) = \sum_{i=1}^n \omega_i \max(X_i)$ ， $\min(\xi) = \sum_{i=1}^n \omega_i \min(X_i)$ ， X_i 是实测变量。

在估计出顾客满意指数模型参数之后，为得到各个系数的误差估计，DPS 系统采用了 Bootstrap 模拟抽样方法对模型的参数进行模拟抽样估计。Bootstrap 抽样的有关参数可由用户设定。

33.3.2 DPS 处理应用

由于顾客满意指数模型较为复杂，因此 DPS 系统将约定：用户用鼠标选中的部分是结构方程定义的有关参数，下面是原始数据。原始数据自变量放在左边，因变量放在右边。如选中的参数部分有 n 行 m 列，则其左上方的 $(n-1) \times (n-1)$ 方阵是标记因变量作用系数 β_{ij} ；右边 1 列是自变量作用系数 γ_{ij} ；上面用户界面里面的系数是根据如下结构方程确定的(注意， β_{ij} 矩阵对角线上的值必须为 -1)。

参数部分的第 n 行是和下面的数据相对应的每个结构变量所包含的观测变量个数，各个数字之间用空格隔开，其总和等于下面原始数据总列数。在本例中，该行参数“2 4 4 2”，其意思是表示原始观测数据左边两列是自变量 x_{11} ， x_{12} ；然后依次是 y_{11} ， y_{12} ， y_{13} ， y_{14} ； y_{21} ， y_{22} ， y_{23} ， y_{24} ； y_{31} ， y_{32} ； y_{41} ， y_{42} ， y_{43} ， y_{44} ； y_{51} ， y_{52} 。

在选中的数据块后面，接下来的第一行不是调查数据，而是每个观测变量(理论设计的)最大值，如这里的每个观测指标是分 9 个级别(最高级是 9)调查的，所以该行都是 9。

再下面就是各个观察(调查)指标的原始数据方阵，存放顺序按系统约定：自变量放在左边，因变量放在右边，其中间不可有缺失数据。

本例系统执行该顾客满意指数模型分析功能时，用户提供的参数、数据的电子工作表界面如图 33-5。

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

线波动情况按“停止模拟”按钮停止迭代。迭代结束，用户界面上部还有几个页面，存放着有关结果的图形(图 33-7~图 33-9)。

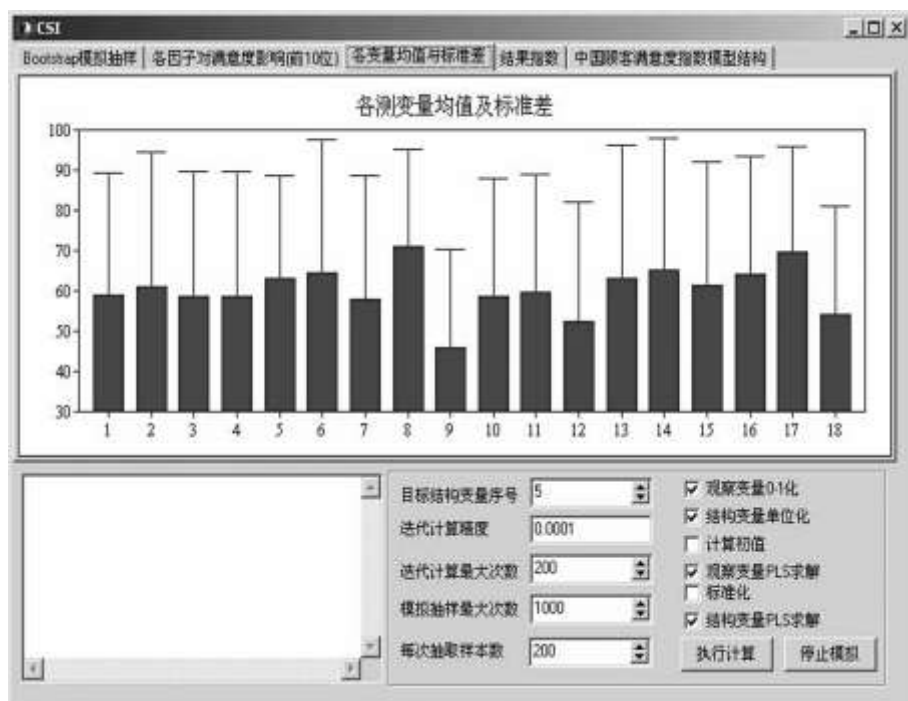


图 33-6 顾客满意指数分析用户交互对话界面

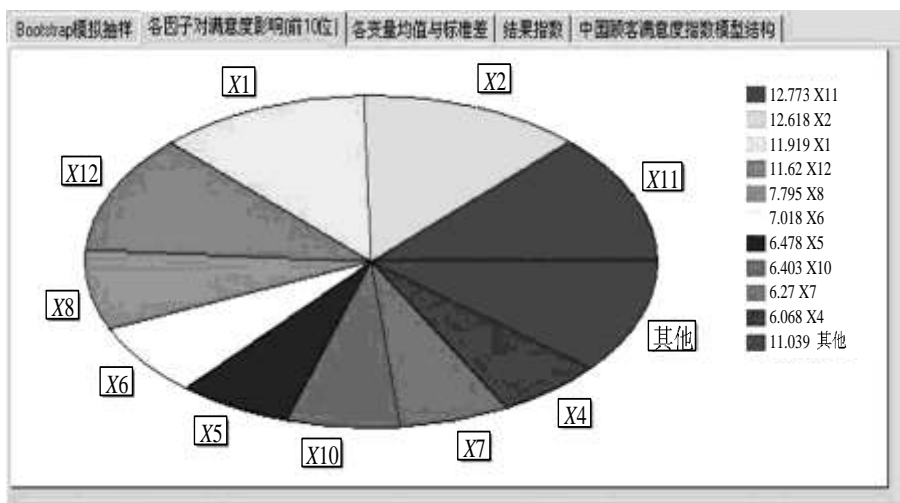


图 33-7 各个观察变量对目标结构变量影响力(按大小取前 10 位)

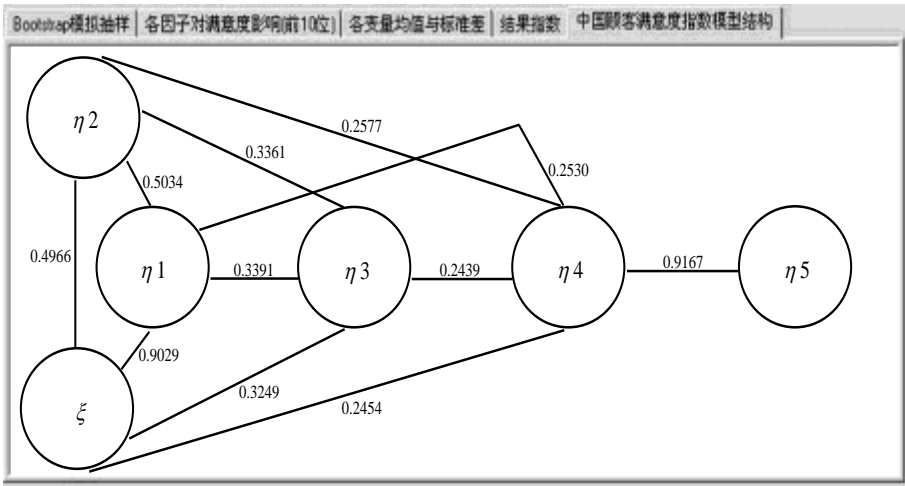


图 33-8 各观测变量的模型误差方差

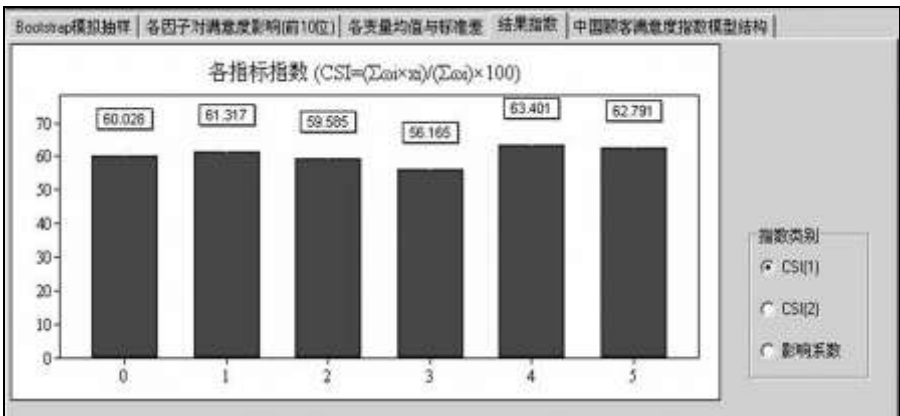


图 33-9 各结构变量所对应的指数

这些图形可供用户参考、使用。当用户点击界面右上角的窗口关闭按钮，系统将返回编辑状态，同时输出数值结果如下：

外部变量 0-1 化。

内部变量重新单位化。

结构变量 5 为目标变量(如顾客满意度)

结构变量矩阵(外生值)已经单位化。

偏最小二乘迭代, 控制精度=0.0001 次数控制≤ 200

结构变量矩阵(外生值)已经单位化。

第 1 次迭代误差=0.772339

第 2 次迭代误差=0.054690

第 3 次迭代误差=0.006445

第 4 次迭代误差=0.000757

第 5 次迭代误差=0.000089

迭代次数 = 5

变量	观察变量	权重	模型误差	观察变量对目标结构变量				
序号	平均值	ω_i	方差	影响系数	影响力(%)			
1	0.5900	0.4858	0.3397	0.1192	11.9187			
2	0.6100	0.5142	0.3687	0.1262	12.6180			
...			
18	0.5400	0.4329	0.2796	0.0000	0.0000			
结构变量	S1	S2	S3	S4	S5	S6		
1	0.1717	0.1427	0.2877	0.1460	0.2597	0.2579		
2	0.2174	0.2214	0.2390	0.1860	0.2329	0.1603		
...		
25	0.3122	0.2694	0.1837	0.2120	0.1714	0.1452		
平均	0.1874	0.1933	0.1964	0.1887	0.1954	0.1887		
结构变量(隐含变量)间内生系数矩阵 $\beta(ij)$, $\gamma(ij)$								
β_{ij}	j=2	j=3	j=4	j=5	j=6	γ_{ij}		
i=1	0	0	0	0	0	0.9029		
i=2	0.5034	0	0	0	0	0.4966		
i=3	0.3391	0.3361	0	0	0	0.3249		
i=4	0.2530	0.2577	0.2439	0	0	0.2454		
i=5	0	0	0	0.9167	0	0		
指数 $CSI(1)=(\sum\omega_i \times x_i)/(\sum\omega_i) \times 100$								
60.0285	61.3170	59.5848	56.1653	63.4009	62.7907			
指数 $CSI(2)=(E[\xi]-\text{Min}[\xi])/(\text{Max}[\xi]-\text{Min}[\xi]) \times 100$								
54.3183	54.1120	51.7257	49.9032	58.1725	57.4751			
影响系数								
0.2454	0.2530	0.2577	0.2439					
下面是 Bootstrap 抽样结果, 抽样次数 = 850								
变量	观察变量平均值		模型误差方差		观察变量对目标结构变量影响系数及其影响力			
序号	平均值	标准误	平均值	标准误	平均值	标准误	平均值	标准误
1	0.5893	0.0216	0.3972	0.0248	0.1191	0.0033	11.9093	0.3310
2	0.6100	0.0231	0.4346	0.0252	0.1263	0.0033	12.6285	0.3266
...
18	0.5393	0.0184	0.3278	0.0203	0.0000	0.0000	0.0000	0.0000
结构变量(隐含变量)间内生系数矩阵 $\beta(ij)$, $\gamma(ij)$ Bootstrap 抽样平均值								
β_{ij}	j=2	j=3	j=4	j=5	j=6	γ_{ij}		
i=1	0	0	0	0	0	0.9029		
i=2	0.5035	0	0	0	0	0.4965		
i=3	0.3391	0.3361	0	0	0	0.3248		
i=4	0.2530	0.2577	0.2439	0	0	0.2454		
i=5	0	0	0	0.9168	0	0		
与之对应结构变量(隐含变量)间内生系数矩阵 $\beta(ij)$, $\gamma(ij)$ Bootstrap 抽样标准误								
β_{ij}	j=2	j=3	j=4	j=5	j=6	γ_{ij}		

i=1	0	0	0	0	0	0.0094
i=2	0.0016	0	0	0	0	0.0016
i=3	0.0017	0.0012	0	0	0	0.0020
i=4	0.0012	0.0008	0.0016	0	0	0.0015
i=5	0	0	0	0.0081	0	0
指数 $CSI(1)=(\sum \omega_i \times x_i)/(\sum \omega_i) \times 100$						
序号	平均值	标准误				
S1	60.0346	1.6227				
S2	61.3413	1.1556				
S3	59.6529	0.7923				
S4	56.1864	1.4292				
S5	63.4685	0.9732				
S6	62.7743	1.5549				
指数 $CSI(2)=(E[\xi]-\text{Min}[\xi])/(\text{Max}[\xi]-\text{Min}[\xi]) \times 100$						
序号	平均值	标准误				
S1	54.3253	1.8546				
S2	54.1396	1.3847				
S3	51.8064	0.9552				
S4	49.9274	1.6333				
S5	58.2498	1.1122				
S6	57.4563	1.7771				
影响系数						
序号	平均值	标准误				
S1	0.2454	0.0015				
S2	0.2530	0.0012				
S3	0.2577	0.0008				
S4	0.2439	0.0016				

33.4 结合分析

结合分析(conjoint analysis)是评价消费者在多属性产品(或服务)中做出决策的一种重要手段,是在已知应答者对某一产品整体评估结果情况下,经过分解去估计其偏好结构的一种多元统计分析方法。它用于确定哪些产品的属性(attributes)与水平(levels)对于消费者来说是最重要的。通过测量消费者对产品属性的偏好,其数值化参数用分值得(part-worths)、权重(importance weights)等进行定量描述。其分析步骤如下:

33.4.1 确定要研究的属性和属性水平

属性应是影响被调查者偏好的突出属性,既不能太多,也不能太少。属性水平的确定,其范围应稍大于医疗卫生服务领域的流行范围,但又不能太大,以致影响评价的可信度。

33.4.2 实验设计

当属性和属性水平数较多时,如果要评价全部可能组合,即采用全因子设计(full-factorial design)势必造成应答任务繁重,从而影响到应答率和资料完整性、准确性。当属性数目在 3 个以上时,可以在一定条件下挑选部分组合做调查。挑选部分组合可采用正交设计方法,但当遇到诸如某些属性组合是不可能的、要设计一个不是标准数目的产品组合,正交设计则不可取,此时可采用均匀设计方法,其功效(efficiency)通常也较高。

33.4.3 数据分析

依据设计,将服务的属性和属性水平进行组合,每一种组合就代表了一个待评价的产品轮廓,并采用调查表格或评价卡片的形式由被调查者对产生的这一系列产品轮廓进行评价,完成资料的收集。分析如下:

(1) 水平效用:要获得各种组合在被调查者心目中的效用,需先计算组成该产品组合的每个属性水平的效用值,其效用值越大,说明被调查者对该属性该水平越偏好。计算方法通常采用分值函数模型(path-worth function model),因分值函数模型对函数关系没有特殊的规定,受到的限制少,属性的每个水平都单独作为一个哑变量进入模型,估计各属性水平的效用分值。

(2) 轮廓效用:最常用的是加法模型(addictive model),即把每个属性的成分效用值加起来得到某种属性组合的总轮廓效用。通过计算轮廓效用来比较消费者对不同产品组合的偏好。

常用的估计产品/服务组合总效用的模型为

$$U(x) = \sum_{i=1}^m \sum_{j=1}^k a_j x_{ij}$$

$U(x)$ 为某产品的总分值效用, m 和 k 分别代表属性数和水平数, a_j 为属性 i 第 j 个水平的分值效用。

(3) 属性的相对重要性:属性的水平效用的差值越大,表明该属性在整体轮廓中的重要性越大,则对被调查者决策的影响也越大,各属性的相对重要性之和为 1,其计算公式为

$$w_i = \frac{\max(v_i) - \min(v_i)}{\sum_{i=1}^m (\max(v_i) - \min(v_i))} \times 100\%$$

其中, w_i 是第 i 个属性的相对重要性, $i=1, 2, 3, \dots, m$ (m 为研究属性的个数); $\max(v_i)$ 是第 i 个属性的最大水平效用值; $\min(v_i)$ 是第 i 个属性的最小水平效用值。

33.4.4 模型评价

模型评价常用的方法有:① 评价模型的拟合优度(goodness of fit),即用决定系数 R^2

值说明模型对数据的拟合程度。一般而言, R^2 值不宜小于 0.50 才可信; ② 内部有效性, 将大部分实验卡片用于估计模型, 少部分卡片用于检查。检查卡片代入模型得到的预测值与实际观测值之间的 pearson 相关系数或 Kendall Tau 等级相关系数判断模型的内部有效性。

33.4.5 应用示例

这里用一个 MP3 产品的例子来说明结合分析在产品测试上的应用, 并介绍在 DPS 系统中结合分析的过程。

假定某公司打算在市场上推出一款新型 MP3 产品, 首先要了解消费者对 MP3 产品的喜好, 消费者更重视 MP3 产品的哪些属性或特征, 他们对这些属性或特征又有什么特别的偏好, 为了设计出受消费者欢迎的 MP3, 该公司需要开展一次市场调研, 对各种配置的 MP3 产品进行测试。

1. 确定产品或服务的属性与属性水平

通过查阅有关广告、收集二手资料和走访 MP3 的零售商确定: 容量、音质、价格、外形、品牌、功能、产地、电池使用时间、线控、屏显等产品特征是 MP3 的潜在的重要属性, 随机抽取若干消费者对这 10 个属性的重要度进行前期预调查(采用目前市场调查中常用的对单个属性的重要度分别打分, 按分数的平均值排序), 从而确定对容量、音质、外形和价格这 4 个属性进行结合分析, 这 4 个属性及其属性水平如表 33-1 所示:

表 33-1 MP3 产品属性及各属性水平

属性	属性水平		
	64M	128M	256M
音质	好	一般	差
外形	时尚	一般	传统
价格	125 元	175 元	250 元

2. 产品模拟

利用上述属性与属性水平可以组合起 81 种模拟产品(3×3×3×3)。如果受访者对所有 81 种模拟产品进行一一评价, 那将是十分麻烦的。因此需采用正交设计来减少模拟产品数量。查正交表, 所需要测试的模拟产品可以减少到 9 种。图 33-10 里面表格的 B~E 列是使用正交设计得出的一个正交设计方案。

将上面正交设计产生的 9 种模拟产品, 每个模拟产品设计 1 张卡片, 进行调查。通过调查可得到消费者对这 9 种模拟产品评价数据(评价分 1~9 级: 最愿意购买的评价等级为 9, 最不愿意购买的等级为 1)。如某消费者甲的评价数据如下:

模拟产品的编号	A	B	C	D	E	F	G	H	I
购买的可能性	5	1	3	4	3	9	1	4	8

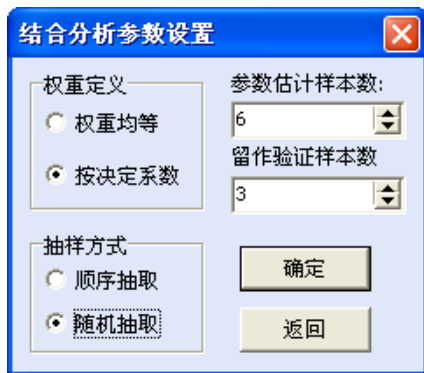
3. 调查数据 DPS 分析示例

如果我们调查了 9 个消费者，得到评价数据之后，录入到 DPS 电子表格，相应的试验设计表的右边。即在电子表格的左边是模拟产品的属性的各个水平，左边放各个调查对象(消费者)的评价值。分析时先用鼠标选中模拟产品属性水平及其评价数据(图 33-10)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	MP3					消费者									
1	产品	价格	容量	音质	外形	1	2	3	4	5	6	7	8	9	
2	A	250	128M	好	守旧	5	4	5	5	2	5	5	4	5	
3	B	250	256M	差	一般	1	1	2	2	3	1	2	1	2	
4	C	175	64M	好	一般	3	4	3	3	4	3	3	3	3	
5	D	175	256M	一般	守旧	4	3	4	4	3	4	4	3	4	
6	E	175	128M	差	时尚	3	2	3	3	2	3	3	2	3	
7	F	125	256M	好	时尚	9	9	9	9	9	9	9	9	9	
8	G	125	64M	差	守旧	1	1	1	1	1	1	1	1	1	
9	H	250	64M	一般	时尚	4	5	3	4	5	3	4	5	3	
10	I	125	128M	一般	一般	8	7	8	8	7	8	8	7	8	
11															

图 33-10 结合分析试验设计及数据格式

分析时执行专业统计→量表分析和顾客满意指数模型→结合分析，系统弹出如下所示的用户对话框。



在用户对话框中，可输入用于估计模型参数的样本数(该例中为 6 个样本)，及少部分用于检验预测值与实际观测值之间相关性样本数(该例中为 3 个样本)。另外，在对话框中，可由用户确定各样本综合加权方式，以及参数估计(或留作检验)的样本是否随机地分配。然后点击“确定”按钮，即可得到计算结果如下：

结合分析结果							
Factor	Part-Worth	S9	S8	S7	S6	S4	S5
价格:250	-0.7222	-0.8889	-0.5556	-0.6667	-1.1111	-0.6667	-0.4444
价格:175	-1.0000	-0.8889	-1.2222	-1.0000	-0.7778	-1.0000	-1.1111
价格:125	1.7222	1.7778	1.7778	1.6667	1.8889	1.6667	1.5556

容量:128M	0.8333	1.1111	0.4444	1.0000	1.2222	1.0000	0.2222
容量:256M	0.6111	0.7778	0.4444	0.6667	0.5556	0.6667	0.5556
容量:64M	-1.4444	-1.8889	-0.8889	-1.6667	-1.7778	-1.6667	-0.7778
音质:好	1.4444	1.4444	1.4444	1.3333	1.5556	1.3333	1.5556
音质:差	-2.3889	-2.2222	-2.5556	-2.3333	-2.4444	-2.3333	-2.4444
音质:一般	0.9444	0.7778	1.1111	1.0000	0.8889	1.0000	0.8889
外形:守旧	-1.0556	-0.8889	-1.2222	-1.0000	-0.7778	-1.0000	-1.4444
外形:一般	0.0000	0.1111	-0.2222	0.0000	-0.1111	0.0000	0.2222
外形:时尚	1.0556	0.7778	1.4444	1.0000	0.8889	1.0000	1.2222
Factor	Part-Worth Range						
价格	2.7222	2.6667	3.0000	2.6667	3.0000	2.6667	2.6667
容量	2.2778	3.0000	1.3333	2.6667	3.0000	2.6667	1.3333
音质	3.8333	3.6667	4.0000	3.6667	4.0000	3.6667	4.0000
外形	2.1111	1.6667	2.6667	2.0000	1.6667	2.0000	2.6667
Factor	Factor Importance						
价格	24.8731	24.2424	27.2727	24.2424	25.7143	24.2424	25.0000
容量	20.8122	27.2727	12.1212	24.2424	25.7143	24.2424	12.5000
音质	35.0254	33.3333	36.3636	33.3333	34.2857	33.3333	37.5000
外形	19.2893	15.1515	24.2424	18.1818	14.2857	18.1818	25.0000
决定系数	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
样本号	估计值	S3	S2	S1			
1	4.6667	5.0000	4.0000	5.0000			
2	1.6667	2.0000	1.0000	1.0000			
3	3.1667	3.0000	4.0000	3.0000			
...			
9	7.6667	8.0000	7.0000	8.0000			
相关系数		0.9864	0.9679	0.9937			
Kandell Tau 系数		0.9444	0.8733	0.9444			
Tau 系数 Z 值		3.5447	3.2778	3.5447			
Tau 系数 p 值		0.0002	0.0005	0.0002			

从分析结果可以看出,对该消费者而言,MP3 的音质是消费者最关心的,相对重要程度为 35.03%,其次是 MP3 价格,相对重要程度为 24.87%,容量和外形似乎并不十分重视,其重要程度靠后。

属性水平的效用栏表示该属性水平对于该消费者而言的效用。效用越高,则表示该特征水平的越受欢迎。注意,这里对于不同属性的属性水平的效用不能进行直接比较,比如不能说音质好的效用比外形时尚的效用好。但在其他属性水平相同的条件下,可以对同一属性下的不同属性水平进行比较,比如在价格、音质和容量属性水平不变的情况下,该消费者认为外形时尚的效用(1.0556)比外形一般(0.0000)的效用要高。可以看出最受该消费者欢迎的 MP3 产品属性特征是:价格 125 元(1.722);容量 128M(0.833);音质

好(1.444); 外形时尚(1.056)。

另外,从 3 个检验样本的相关系数可以看出,检验样本和模型具有极显著的相关性,这说明模型表现良好。

参 考 文 献

- 方积乾. 2001. 生存质量测定方法及应用[M]. 北京: 北京医科大学出版社.
- 国家质检总局质量管理司. 清华大学中国企业研究中心. 2003. 中国顾客满意指数指南[M]. 北京: 中国标准出版社.
- 柯惠新, 黄京华, 沈浩. 1992. 调查研究中的统计分析法[M]. 北京: 北京广播学院出版社.
- 柯惠新, 保罗·弗悉诺. 1994. 市场营销研究中的结合分析法[J]. 数理统计与管理, 13(6): 56-65.
- 刘宇. 2001. 顾客满意度测评方法研究[J]. 数量经济与技术研究, 2: 87-90.
- 刘宇. 2003. 顾客满意度测评[M]. 北京: 社会科学文献出版社.
- 沈浩, 柯惠新. 1998. 结合分析的原理与应用[J]. 数理统计与管理, 17(4): 39-45.
- 唐启义, 唐洁. 2005. 偏最小二乘回归分析在均匀设计试验建模分析中的应用[J]. 数理统计与管理, 25(5): 45-49.
- 童恒庆, 周瑾, 林卉. 2005. 顾客满意指数模型的最佳迭代初值问题[J]. 统计与决策, 189(5): 10-11.
- 童恒庆. 2005. 理论计量经济学[M]. 北京: 科学出版社.
- 王惠文. 1999. 偏最小二乘回归方法及其应用[M]. 北京: 国防工业出版社.
- Claes Fornel, Michael D. Johnson et al. 1996. The american customer satisfaction index: nature, purpose and findings[J]. Journal of Marketing, 60: 7-18.
- Green P E and Rao V R. 1971. Conjoint measurement for quantifying judgmental data[J]. Journal of Marketing Research, 8: 355-363.
- Green P E and Srinivasan V. 1990. Conjoint analysis in marketing: new developments with implications for research and practice[J]. Journal of Marketing, 54: 3-19.
- Manoukiam E B. 1986. Modern Concepts and Theorems of Mathematical Statistics[M]. New York: John Wiley and Sons. Inc.

第 34 章 生物测定与药物动力学

生物测定(bioassay)是生物、医学研究中最普遍的一类实验,通常包括供试因子(如化学或生物性杀虫剂、杀菌剂、除草剂、植物生长调节剂、新医药、保健品或可能具有某种活性的物质或产品)和受试靶体(如一种昆虫、植物或小白鼠之类的实验动物)。这类实验有定性与定量水平之分。定性测定较为简单,如仅测定某种微生物是否对某种动物有致病性,其结果仅仅回答是或非两种可能;定量测定则较为复杂,通常包括供试因子的一系列剂量,每剂量要处理合理数量(越多实验误差越小)的受试靶体并且整个实验要持续观察一段时间。单就受试靶体的个体而言,只有对供试因子反应和不反应两种结果,但就各剂量水平下受试靶体群而言,结果是发生反应或不反应的比例。例如测定某剂量杀虫剂处理某害虫后死亡与存活比例、病菌孢子喷药后的萌发与不萌发比例等。概言之,定量生物测定包含的主体因素是剂量、受试靶体和反应(或不反应)比例。

关于定量生物测定数据的整理分析,英国学者 Finney 的 *Probit Analysis* 一书曾被誉为该学科的“圣经”,20 世纪 40 年代末首版后曾三次修改再版并先后印刷 10 余次(最后一次印刷是 1981 年),可见其影响巨大。

但是,近年来国内外学者对 Finney 创立的概率分析方法提出了挑战,认为概率分析的明显缺点是不能充分利用实验信息,尤其是时间与剂量效应的评价相互脱节,无法揭示时间与剂量间的互作效应。因此,时间-剂量-死亡率模型(time-dose-mortality model, TDM)分析方法正被用来替代传统的概率分析方法并逐渐为学术界所接受(冯明光等, 1996; Nowierski 等, 1996; Robertson et al., 1992),但概率分析方法已为人们所习惯且仍有一定的实用性。

评价两种或两种以上的化学物同时作用于有机体所产生的综合毒性作用时,一般称之为化学物的联合毒性作用(joint toxic effect 或 combination toxic effect)时,目前。根据多种化学物同时作用于机体时所产生的毒性效应,将化学物的联合毒性作用分为相加作用、协同作用、增强作用、拮抗作用以及独立作用等。DPS 系统提出了联合毒性作用的统计学检验技术,解决了以往采用各种指数进行评价的问题。

药物动力学是研究药物在体内吸收、分布、代谢和排泄过程动态变化规律的一门学科;研究的主要目标就是揭示药物在体内的动态变化规律性。药物在体内经历吸收(absorption)、分布(distribution)、代谢(metabolism)和排泄(excretion)过程。随着细胞生物学和分子生物学发展,药物在体内代谢机理研究已有了长足进展。药物进入体内后,经吸收进入血液,并随血流透过生物膜进入靶组织与受体结合,从而产生药理作用,作用结束后从体内消除的过程,可在实验基础上,建立数学模型进行模拟,以对药物在体内过程进行预测。因此,为了解药物代谢过程,药物代谢动力学已成为临床医学的重要组成部分。

34.1 质反应生测机率值分析

34.1.1 基本原理

生物测定中,我们可通过一系列不同剂量对受试生物组群的反应的处理,得到剂量-反应曲线。一般而言,剂量反应曲线是偏右的,而不是正态分布曲线,即在较低剂量下提高单位剂量所增加的反应率比在较高剂量下提高单位剂量所增加的反应率要大,故剂量-反应曲线为偏右型。如果将剂量(浓度)进行对数转换,则转换以后的剂量-反应曲线呈正态分布曲线。因此,在杀虫剂毒力测定中,一般所设的农药浓度呈几何级数增加。

由于反应比例(如虫死亡率)与对数浓度的关系呈累积次数曲线形式,通常对该反应比例还需要一种适当的统计转换方法,使其与对数浓度之间的关系成为一条直线。Gaddum(1933)提出用正态等值偏差(normal equivalent deviation, NED)代换反应比例(0~1),可使代换后正态等值偏差与浓度对数的关系呈一条直线。具体代换方法为

$$\begin{aligned} \frac{1}{2}(1 + \alpha) &= p \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{NED}} e^{-\frac{1}{2}u^2} du \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x}{\delta}} e^{-\frac{x^2}{2\delta^2}} d\left(\frac{x}{\delta}\right) \end{aligned}$$

式中, $x = x - \mu$, 而 x/δ 即为正态等值偏差。由于 x/δ 可为正数亦可为负数,故计算不便,于是又有人提出了概率值(probit)代换法。设概率单位为 y , 概率值代换公式为 $y = 5 - x/\delta$ 或 $y = -5 + x/\delta$ 。通常, x/δ 的负值总是大于 -5 , 故概率值不会出现负数。代换后的概率值

$$\begin{aligned} P &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y-5} e^{-\frac{x^2}{2\delta^2}} d\left(\frac{x}{\delta}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y-5} e^{-\frac{\mu^2}{2}} d\mu \end{aligned}$$

概率分析的经验模型及其参数拟合方法由 Finney(1947)提出,且长期以来是杀虫剂毒力测定的通用方法。在此,我们应用最大似然(maximum likelihood)方法对供试因子(剂量)和受试生物组群反应比例之间的关系建立模型并进行拟合。

如上所述,概率分析中,对于给定剂量范围,概率模型是关于受试生物组群对供试因子反应的期望比例 $E(y)$ 对剂量 x 的函数,即

$$E(y) = \text{norm}(A - 5 + B \lg(x))$$

式中, $\text{norm}()$ 是标准的正态分布函数。当观察到的机率值 $\text{norm}^{-1}(y_i) + 5$ 和对数剂量 $\lg(x_i)$ 呈线性相关时可使用上述概率模型(y_i 是剂量为 x_i 的反应比例)。

概率或机率分析假定受试生物组群对给定剂量的反应呈二项分布(即有无反应),通

过最大似然的逼近方法估计参数 A 和 B 。在实际工作中，可能是一种化合物的几个剂量水平的处理，抑或是几种化合物且每种化合物又包含几个剂量水平的处理。对于后者，可对每种化合物的测定结果分别拟合机值模型，计算共毒系数（见 34.3.3 章节）。再对各机值模型的平行性进行假设统计检验。

34.1.2 数据整理与分析

在 DPS 计算机数据处理平台上，概率分析不再是令人烦恼的事情。只需将各供试药剂浓度(对照组的浓度为 0)、各浓度下处理的受试生物个体数(如试虫的头数)和相应的死亡个体数，按系统规定的格式进行编辑排列，并定义为数据块，就可进行概率分析运算。

例 1 一般情况下，对只包含一种化合物几个剂量水平的生测数据进行机值分析，其数据编辑、数据块定义格式见图 34-1。

	A	B	C	D	E
1		浓度	总虫数	死虫数	
2		10.2	50	44	
3		7.7	49	42	
4		5.1	46	24	
5		3.8	48	16	
6		2.6	50	6	
7		0.0	49	0	
8					

图 34-1 单一药剂概率分析的数据编辑和格式定义

按图 34-1 方式将数据定义成数据块之后，选择功能菜单中“生物测定”栏下的“质反应生测机率值分析”，按回车键后，系统给出用户界面如图 34-2。

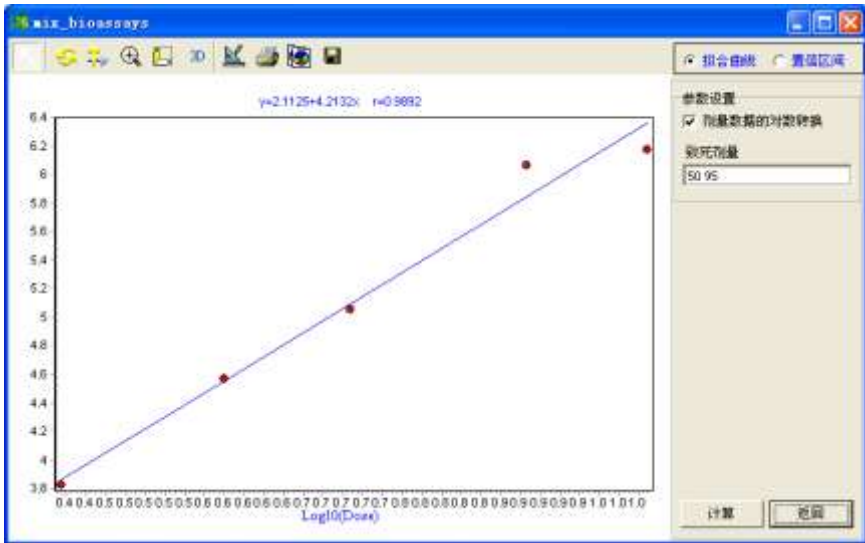



图 34-2 质反应机率值分析用户界面

在用户界面里面，用户可以确认是否对剂量浓度数据进行对数转换。并允许用户输

入待求的多个致死剂量效应指标,如 10, 50, 90 等,各值之间用空格隔开。并在用户界面显示拟合图,以及反应方程数值结果。

用户界面中显示的当前拟合结果图,可点击按钮将其复制到文字报告中。最后按“返回”按钮,将计算输出到工作表。

在分析中,计算机程序自动用各组输入数据计算反应比例(本例中为死亡比例)并转换为机率值。如果有对照,即有处理浓度为 0 的组合,系统可按(处理-对照)/(1-对照)的公式对所有浓度处理的结果进行校正,再转换为机率值。

本例数据,经 DPS 处理计算,输出结果经整理,得毒力回归方程 $y = 2.1125 + 4.2132x$ 。

拟合程度的卡方值为 $\chi^2 = 1.729$, $p = 0.6305$, 现 $p > 0.05$, 表明所求的毒力回归曲线是合适的,即否定所求模型异质性(heterogeneity)的假设。在计算结果中同时给出 LD_{50} 的对数浓度 (0.6853 ± 0.0005) 及正常浓度 (4.8455 ± 0.2462), 并给出了 95% 置信限 ($4.3517 \sim 5.3508$)。

例 2 根据甘蔗工业科学研究所用 0.65% 农药液剂、烟草粉肥皂液和烟脉肥皂液三种药液对棉蚜(*aphis gossypii*)进行毒力测定的数据(王鉴明, 1985), 进行机值分析并检验三种毒力回归直线的平行性。这一分析如果用手工计算,不但计算过程冗长,而且极易出错(事实上,原作者计算结果有误),同时会引起舍入误差。如今在计算机处理平台上用机值分析方法比较三种药剂的相对毒力,处理起来就方便多了。其过程如下:

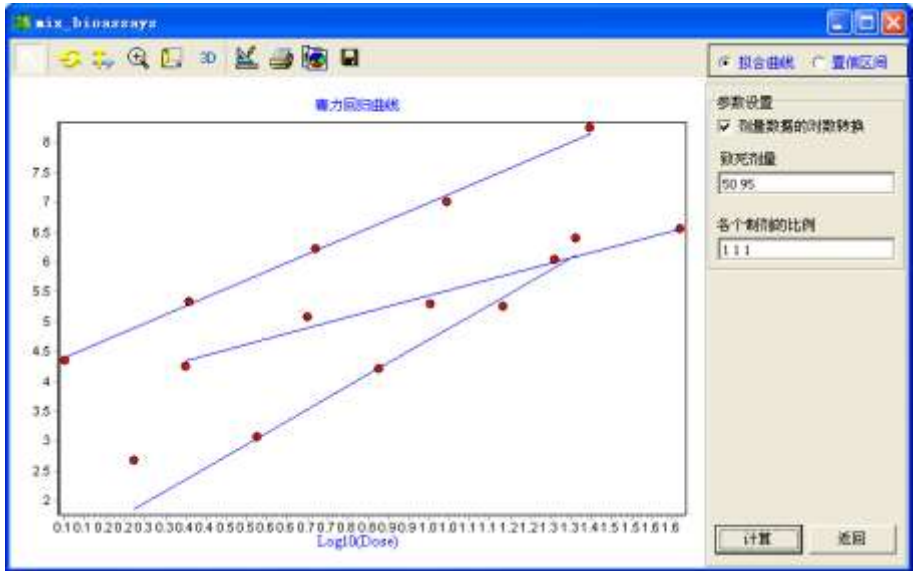
对多药剂的毒力测定,其中每种药剂有几个剂量水平处理的试验结果进行机值分析时,其数据编辑格式为:第 1 列为农药药剂(依次从 1, 2, ...)编号、第 2 列为农药处理浓度,第 3、4 列分别为供试虫数和死亡虫数。本例的数据定义格式如图 34-3 所示。

	A	B	C	D	E	F
1		药剂序号	浓度	试虫数	死亡数	
2		1	41.666	1097	1031	
3		1	20.408	1026	872	
4		1	10.101	979	605	
5		1	5.025	1010	532	
6		1	2.506	992	227	
7		2	23.018	1445	1325	
8		2	15.228	1459	875	
9		2	7.556	1002	212	
10		2	3.764	960	26	
11		2	1.879	923	9	
12		3	25	1706	1705	
13		3	11.1111	1185	1158	
14		3	5.2631	1192	1057	
15		3	2.5641	1348	849	
16		3	1.2658	1026	266	
17						


图 34-3 多药剂质反应机率分析的数据格式

定义数据块后,选择功能菜单中“生物测定”栏下的“质反应生测机率值分析”,按

回车键后，系统显示类似图 34-2 所示的用户界面如下：



在这里用户可以确认是否对剂量浓度数据进行对数转换，并输入待求的致死剂量效应指标(致死剂量可同时输入 10，50，90 等，各值间用空格隔开)以及各个制剂含量的比例。然后点击“确认”后可立即得到分析结果。

用户界面中显示的当前拟合结果图，可点击按钮将其复制到文字报告中。最后按“返回”按钮，将计算输出到工作表。

处理	截距 A	斜率 B	SE(B)	卡方值	自由度	P
样品 1	3.625090	1.808103	0.050940	36.62600	3	0.000000
注意：异质因子较大，H=12.2085708620336，故计算置信限时 t 值取 3.18243059406076						
LD50	对数浓度=	0.760416	95% 置信区	0.578126	~	0.889079
	浓度=	5.759914	95% 置信区	3.785525	~	7.746033
LD90	对数浓度=	1.469198	95% 置信区	1.313009	~	1.726023
	浓度=	29.45767	95% 置信区	20.55931	~	53.21367
处理	截距 A	斜率 B	SE(B)	卡方值	自由度	P
样品 2	0.794713	3.909065	0.093747	140.3370	3	0.000000
注意：异质因子较大，H=46.779085100109，故计算置信限时 t 值取 3.18243059406076						
LD50	对数浓度=	1.075778	95% 置信区	0.945432	~	1.371864
	浓度=	11.90634	95% 置信区	8.819261	~	23.54310
LD90	对数浓度=	1.403619	95% 置信区	1.190374	~	2.028178
	浓度=	25.32906	95% 置信区	15.50150	~	106.7033
处理	截距 A	斜率 B	SE(B)	卡方值	自由度	p
样品 3	4.104951	2.893126	0.078639	5.819000	3	0.120757
LD50	对数浓度=	0.309371	95% 置信区	0.280791	~	0.335526
	浓度=	2.038782	95% 置信区	1.908936	~	2.165338

LD90	对数浓度=	0.752335	95%置信区	0.737149	~	0.767618
	浓度=	5.653729	95%置信区	5.459455	~	5.856228
各个样品结合起来分析						
处理	截距 A	斜率 B	SE(B)	卡方值	自由度	p
样品 1	2.909186	2.601692	0.037971	41.84900	3	0.000000
注意: 异质因子较大, $H=13.9495170337498$, 故计算置信限时 t 值取 3.18243059406076						
LD50	对数浓度=	0.803636	95%置信区	0.662322	~	0.908161
	浓度=	6.362625	95%置信区	4.595382	~	8.093964
LD90	对数浓度=	1.296220	95%置信区	1.183487	~	1.461358
	浓度=	19.77972	95%置信区	15.25761	~	28.93061
处理	截距 A	斜率 B	SE(B)	卡方值	自由度	p
样品 2	2.214366	2.601692	0.037971	84.36300	3	0.000000
注意: 异质因子较大, $H=28.1209912176015$, 故计算置信限时 t 值取 3.18243059406076						
LD50	对数浓度=	1.070701	95%置信区	0.940002	~	1.291871
	浓度=	11.76795	95%置信区	8.709684	~	19.58262
LD90	对数浓度=	1.563285	95%置信区	1.329182	~	2.094001
	浓度=	36.58347	95%置信区	21.33937	~	124.1654
处理	截距 A	斜率 B	SE(B)	卡方值	自由度	p
样品 3	4.235837	2.601692	0.037971	5.450000	3	0.141663
LD50	对数浓度=	0.293718	95%置信区	0.263657	~	0.321253
	浓度=	1.966608	95%置信区	1.835089	~	2.095333
LD90	对数浓度=	0.786302	95%置信区	0.769784	~	0.803111
	浓度=	6.113666	95%置信区	5.885509	~	6.354940
方差分析结果						
变异来源	自由度	平方和	均方	F 值	显著水平	
平行回归	2	427.2733	213.6367	10.51925	0.00441	
平行偏差	9	182.7820	20.30911			
总和	11	610.0553				
致死剂量(LD)比率测定 Lethal Dose Ratios:						
比较组别	LD	比率	95%置信区间			
1:2	50	0.483769	0.4552~0.5141*			
1:3	50	2.825175	2.6407~3.0226*			
2:3	50	5.839927	5.5717~6.1211*			
1:2	90	1.161765	1.0546~1.2798*			
1:3	90	5.206449	4.6993~5.7683*			
2:3	90	4.481500	4.1911~4.7921*			
说明: 若两浓度致死剂量(LD)的比率 95%置信区间包含 1, 则 LD 之间的差异不显著						

从计算结果可以看出: 三种农药各自的概率分析模型的拟合度以第一、二个较差, 第三个较好, 卡方检验 χ^2 值分别为 36.63, 140.54 和 5.82。

对这三种药剂继续进行回归直线平行性假设检验, 其结果可从方差分析结果看出,

害虫对三种农药反应的差异极显著($F=10.5023$, 大于 $F_{\alpha=0.01}=8.05$), 所以这三条回归线的斜率不同, 即不存在共同斜率。

34.1.3 结果分析解释

(1) 拟合程度卡方值及其显著水平。拟合程度卡方值及其显著水平表明了所求的毒力回归曲线是合适的, 即否定所求模型异质性(heterogeneity)的假设。如显著水平大于 0.05, 则表明所建立的模型合适; 反之, 则认为所建立的模型不合适。

(2) 对多种药剂进行机值分析并检验毒力回归直线平行性。可以其方差分析结果推断: 当方差分析表中的 F 值的显著水平 p 值大于等于 0.05 时, 表明毒力回归直线互相平行。

(3) 上述方差分析只能揭示毒力回归直线的总的趋势, 而两种药剂其致死剂量之间的差异, DPS 数据处理系统则采用致死剂量比率测定方法来检验: 当致死剂量比率的 95% 置信区间包含 1, 则两种药剂的某致死剂量之间的差异不显著。

(4) 注意, 有时因异质因子较大, 计算的置信限会溢出, 以致不能给出其置信区间。

34.2 数量反应生测机率值分析

34.2.1 概述

除草剂、杀菌剂等非计数资料只有死亡(或抑制)比例的百分率, 而没有计数数据。由于反应比例与对数浓度的关系与上述的计数数据一样呈累积次数曲线形式, 因此需要一种适当的统计转换方法进行生物测定统计分析。

数量型概率分析的经验模型及其参数拟合方法可用非线性参数估计方法计算。非线性方程为

$$y = \text{norm}(A - 5 + B \lg(x))$$

式中, $\text{norm}()$ 是标准的正态分布函数。当观察到的概率值 $\text{norm}^{-1}(y_i) + 5$ 和对数剂量 $\lg(x_i)$ 呈线性相关时可使用上述概率模型(y_i 是剂量为 x_i 的反应比例)。


34.2.2 数据整理与分析

在 DPS 计算机数据处理平台上, 对一种药剂几个剂量水平的数量型生测数据进行机值分析时, 其数据编辑格式是每一个浓度在一行中按浓度和死亡(抑制)率顺序填入电子表格。对多种药剂, 且每种药剂有几个剂量水平处理的试验结果进行机值分析时, 数据编辑格式为: 第 1 列为药剂(依次从 1, 2, ...)编号, 第 2 列为药剂浓度, 第 3 列为死亡(抑制)率(详见图 34-4)。

定义数据块后, 选择功能菜单“生物测定”中的“数量反应生测机率值分析”, 回车后系统出现和前面质反应生测分析一样的用户界面。在用户界面里面, 用户可以确认是否对剂量浓度数据进行对数转换。并允许用户输入待求的多个致死剂量效应指标, 如 10, 50, 90 等, 各值之间用空格隔开。并在用户界面显示拟合图, 以及反应方程数值结果。

	A	B	C	D	E
1		药剂序号	浓度	死亡率	
2		1	41.666	93.983592	
3		1	20.408	84.990253	
4		1	10.101	61.797753	
5		1	5.025	52.673267	
6		1	2.506	22.883065	
7		2	23.018	91.695502	
8		2	15.228	59.972584	
9		2	7.556	21.157685	
10		2	3.764	2.708333	
11		2	1.879	0.975081	
12		3	25	99.941383	
13		3	11.1111	97.721519	
14		3	5.2631	88.674497	
15		3	2.5641	62.982196	
16		3	1.2658	25.925926	
17					

图 34-4 几种药品不同剂量对某生物死亡率生物测定

用户界面中显示的当前拟合结果图，可点击按钮将其复制到文字报告中。最后按“返回”按钮，将计算输出到工作表。

样品 1	回归截距 A	标准误	回归系数 B	标准误	F 检验值	显著水平 p
	3.658180	0.458157	1.773075	0.486315	110.1803	0.001846
LD50	对数浓度=	1.341820	95%置信区	0.547978	~	1.103047
	浓度=	5.711834	95%置信区	3.531651	~	12.67789
LD90	对数浓度=	2.623372	95%置信区	1.126941	~	2.100952
	浓度=	30.16895	95%置信区	13.39494	~	126.1688
样品 2	回归截距 A	标准误	回归系数 B	标准误	F 检验值	显著水平 p
	0.527118	0.727284	4.097493	0.681823	350.9870	0.000332
LD50	对数浓度=	4.472882	95%置信区	0.976071	~	1.233027
	浓度=	12.34850	95%置信区	9.463910	~	17.10120
LD90	对数浓度=	5.754434	95%置信区	1.260190	~	1.581848
	浓度=	25.37342	95%置信区	18.20497	~	38.18110
样品 3	回归截距 A	标准误	回归系数 B	标准误	F 检验值	显著水平 p
	4.064015	0.190394	3.032891	0.304517	3717.930	0.000010
LD50	对数浓度=	0.935985	95%置信区	0.290798	~	0.327010
	浓度=	2.035221	95%置信区	1.953433	~	2.123293
LD90	对数浓度=	2.217537	95%置信区	0.704562	~	0.759150
	浓度=	5.384714	95%置信区	5.064798	~	5.743151
致死剂量(LD)比率测定 Lethal Dose Ratios:						
比较组	LD	比率	95%置信区间			
别						
1 : 2		50	0.462600	0.3563~0.6005*		

1 : 3	50	2.806500	2.1925~3.5924*
2 : 3	50	6.067400	5.5623~6.6184*
1 : 2	90	1.187600	0.8328~1.6937
1 : 3	90	5.598000	3.9798~7.8743*
2 : 3	90	4.713600	4.2622~5.2127*

说明：若两浓度致死剂量(LD)的比率 95% 置信区间包含 1，则 LD 之间的差异不显著

34.2.3 分析结果解释

(1) 毒力回归方程方差分析 F 值及其显著水平。数量型数据毒力回归方程诊断采用方差分析方法进行，这是数量型资料和计数资料在生物测定分析进行模型诊断方面的差异。当 F 值的显著水平小于等于 0.05 时，表明所求的毒力回归曲线是合适的；反之则表明所建立的模型不合适。

(2) 对多种药剂进行机值分析，DPS 系统采用致死剂量比率测定方法来检验两种药剂在致死剂量方面的差异：当致死剂量比率的 95% 置信区间包含 1，则两种药剂的某致死剂量之间的差异不显著。

34.3 混剂互作毒力系数统计分析

在混合制剂筛选过程中，一个很重要的工作就是制剂混配后的联合毒力的生物测定，而混剂的增效性可以通过联合毒力生物测定的结果求得。如何评价制剂混配后对生物有机体作用大小，目前尚无公认的一套完整的统计方法。这里先介绍目前常用的几种方法。

34.3.1 增效性的判断(S. Sakai 公式)

据 Bliss 的独立联合作用概念，两种药剂或以上混用后的理论死亡率(P_m)公式如下： $P_m = 1 - (1 - P_1)(1 - P_2) \cdots (1 - P_m)$ 。式中 P_1, P_2, \cdots, P_m 分别是单药对生物的死亡率。当实际混配后制剂对生物致死率大于 P_m 时即为增效，小于 P_m 时为拮抗，接近 P_m 为相加作用。例如用 0.5% 敌敌畏杀灭家蝇，其杀灭率为 55% (0.55)，用氯菊酯灭蝇率为 0.75% (0.75)，敌敌畏加氯菊酯混剂的杀灭率为 95% (0.95)。则 $P_m = 1 - (1 - 0.55)(1 - 0.75) = 0.89$ 。这里混剂杀灭率 0.95 大于理论值 0.89，表明增效作用。此方法简单，适用于大量配方筛选，但其结果仅是定性分析。

34.3.2 增效倍数

如果令 LD_{50}^A 和 LD_{50}^B 分别为药剂 A 和药剂 B 单独用药时的半数(50%) 效应的剂量，

LD_{50}^{A+B} 为两药剂联用时产生半数效应的剂量， \hat{LD}_{50}^{A+B} 为混剂的理论致死剂量， P_A 和 P_B 为两药联用配比。则增效倍数计算公式如下：

$$\text{增效倍数} = \hat{D}_{50}^{A+B} / D_{50}^{A+B}, \quad \hat{LD}_{50}^{A+B} = \frac{D_{50}^A \times P_A + D_{50}^B \times P_B}{P_A + P_B}$$

例：AB 两种杀虫剂按 2: 1 比例混配；经测定 A 剂对家蝇 LD_{50} 为 $0.12\mu\text{g}/\text{♀}$ ，B 剂 LD_{50} 为 $0.41\mu\text{g}/\text{♀}$ ，混剂 LD_{50} 为 $0.029\mu\text{g}/\text{♀}$ 。先计算混剂理论 LD_{50}

$$\hat{D}_{50}^{A+B} = (0.12 \times 2 + 0.41 \times 1) / (3 + 1) = 0.217$$

$$\text{混剂的增效倍数} = 0.217 / 0.029 = 7.48 \text{ 倍}$$

两种或两种以上制剂混配后对某一生物增效系数，它在杀虫剂混配研究中应用较为普遍，而且经常用于增效剂的筛选和最佳配比的确定。

34.3.3 共毒系数(Co-toxicity Coefficient)

共毒系数 (CTC) 可按如下公式计算

$$\text{CTC} = \frac{1 / LC_{50}^{A+B}}{P_A / LD_{50}^A + P_B / LD_{50}^B} \times 100$$

式中各个变量与增效倍数公式计算的定义相同。

若共毒系数明显大于 100(120 以上)为增效作用；显著小于 100(80 以下)为拮抗作用；接近 100 时为相加作用。但一般公认共毒系数达到 200 才算明显增效，方有实用意义。例如已知辛硫磷对家蝇 LC_{50} 为 0.052%，二溴磷对家蝇的 LC_{50} 为 0.205%，辛硫磷与二溴磷以 1: 1 混合后其 LC_{50} 为 0.034%。则共毒系数 CTC 为

$$\begin{aligned} \text{CTC} &= \frac{1 / 0.034}{0.5 / 0.052 + 0.5 / 0.205} \times 100 \\ &= 243.99 \end{aligned}$$

得到的混剂共毒系数为 243.99，明显高于 100，既表示辛硫磷与二溴磷按 1: 1 混配后具有增效作用。

34.3.4 Finney 调和平均数法

Finney 提出的调和平均数公式为：

$$q = D_{50}^{A+B} \left[P_A / D_{50}^A + P_B / D_{50}^B \right]$$

式中各个变量与增效倍数公式计算的定义相同。

当 $q=1$ ，药物联用性质为相加性； $q<1$ 为协同性； $q>1$ 为拮抗性。实际上 q 的倒数乘上 100 即为共毒系数 (Co-toxicity Coefficient)。上面硫磷与二溴磷混合例子，其 Finney 调和平均数

$$q = 0.034(0.5/0.052 + 0.5/0.205) \\ = 0.034(9.6154 + 2.4390) = 0.4099$$

34.3.5 Chou- Talalay 结合指数 CI(combination index) 法

合并指数法是从通用的质量作用定律概念衍化出来, 由Chou和Talalay 根据中效方程式 $f_a/f_u = (D/D_{50})^m$, 推导而得到, 基本公式如下:

$$CI = \frac{(D)_1}{(D_x)_1} + \frac{(D)_2}{(D_x)_2} + \alpha \frac{(D)_1(D)_2}{(D_x)_1(D_x)_2}$$

f_a , f_u 为反应系统中反应和未反应部分, $f_a+f_u=1$ 。D为药物剂量, $D_x = [f_a / (1 - f_a)]^{\frac{1}{m}}$, D_{50} 代表产生中度效应的剂量(与 IC_{50} 、 EC_{50} 或 LD_{50} 相似), m 代表一种Hill型(多级动力学)的系数, 表示量-效应s型曲线特点。 $m=1$ 时, 剂量与反应关系为双曲线, $m>1$ 或 $m<1$ 代表正与反s型曲线。 α 表示联合作用参数, 当药物作用是相互排斥性(即相似的联合作用)时, $\alpha=0$, 该方法等同于等效线法(一般 α 取0)。

结合指数(CI)毒理意义, 当它们等于1时, 表示相加; 大于1时, 表示拮抗; 小于1时, 表示协同。这两个指标目前应用较为广泛。当 $\alpha=0$ 时, 结合指数(CI)实际上是根据Logistic模型计算出来的Finney调和平均数。

34.3.6 制剂混合毒力系数的统计检验

上面介绍的各种方法, 对于混配后是否有增效作用, 只是得到一个数值指标, 即增效(或毒力)系数, 并没有对得到的指标(系数)进行误差估计、以及统计检验。这使得当系数偏离1(或100)的多大程度时具有协同(或拮抗)作用的判断, 没有和试验误差联系起来。以至于在进行联合作用判断时, 凭主观臆断, 缺乏科学的统计判断。

生物测定是生物学实验试验在不同程度上存在试验误差, 如常用的共毒系数, 不同的生物测定实验均规定大于120时可认为有增效作用。这显然与实际情况有出入, 因不同实验、试验误差大小不同。生物测定的共毒系数是否显著地大于100, 应考虑试验误差的影响, 即须做统计学检验。在一定的概率水平下(一般为5%)推断是否大于100。

我们在研究混剂结合毒力系数时发现, Bliss早在1939年就提出的一个混合制剂的剂量-死亡曲线概念性公式

$$y = a' + b \lg(D_A + \tau D_B) \quad (34.1)$$

式中 D_A 和 D_B 是混剂中药剂 A 和 B 的剂量, τ 是两制剂成分的毒性比值。据此定义, Bliss 对系数 τ 赋予了一系列解释和推导。

但是将 τ 直接定义为毒性的比值是欠精确的。在固定比例混合的混剂试验中, τ 的精确含义仅是制剂 D_B 相对于 D_A 的当量系数。如果采用机值(Probit)模型评

价混剂结合毒性，混剂的毒力回归模型可表示为：

$$y = \Phi(\alpha + \beta \lg(D_A + \tau_B D_B)) \quad (34.2)$$

当自然对照死亡率为 C 时，有

$$y = C + (1 - C) \cdot \Phi(\alpha + \beta \lg(D_A + \tau_B D_B)) \quad (34.3)$$

式中 y 是死亡率， C 是对照(自然)死亡率， α 是毒力回归截距， β 是毒力回归系数； D_A 是基准药剂， D_B 是添加药剂， τ_B 是添加药剂 D_B 相对于基准药剂 D_A 的当量系数。

作者认为，添加药剂 D_B 的相对基准药剂 D_A 的毒力系数 ρ_B ，等于当量系数 τ_B 除以两种药剂单剂时的致死剂量 (LD_0) 的比值

$$\rho_B = \frac{\tau_B}{LD_{B0}/LD_{A0}} = \tau_B \frac{LD_{A0}}{LD_{B0}} \quad (34.4)$$

显然，仅当两药剂单剂的致死剂量 (LD_{A0} , LD_{B0}) 的比值相等时，当量系数 τ_B 即为相对毒力系数 ρ_B ，即 $\rho_B = \tau_B$ 。

同理，如 D_B 是基准药剂， D_A 是添加药剂， D_A 的当量系数为 τ_A 。同样可得到 D_A 相对毒力系数 ρ_A 。

$$\rho_A = \frac{\tau_A}{LD_{A0}/LD_{B0}} = \tau_A \frac{LD_{B0}}{LD_{A0}}$$

得到两个单剂相对另一个制剂(标准剂)的相对毒力系数之后，两制剂混用之后的结合毒力系数(Combination Potency, CP)可以定义为它们的乘积

$$\begin{aligned} CP &= \rho_A \cdot \rho_B \\ &= \frac{\tau_A}{LD_{A0}/LD_{B0}} \cdot \frac{\tau_B}{LD_{B0}/LD_{A0}} \\ &= \tau_A \cdot \tau_B \end{aligned} \quad (34.5)$$

结合毒力系数 CP 的生物学意义：如果结合毒力系数 CP 等于 1，表明两种药剂混合后的毒力效应是相加作用。如果结合毒力系数 CP 大于 1，那么为协同作用；反之为拮抗作用。

结合毒力系数 CP 是否等于 1，需要进行统计学检验。进行结合毒力系数 CP 的统计检验，须先对混剂的毒力回归方程(式 34.1 或式 34.2)采用极大似然方法求解。利用求解过程中的 Fisher 得分信息阵，估计当量系数 τ_A 和 τ_B 参数估计值及其标准误 SE_{τ_A} 和 SE_{τ_B} 。在 τ_A 和 τ_B 是独立的随机变量的假设下，根据混剂结合毒力系数 CP 和 τ_A 和 τ_B 的关系式， $CP = \tau_A \cdot \tau_B$ ，可得到混剂结合毒力系数 CP 的标准误 SE_{CP} 为

$$SE_{CP}^2 = \tau_B^2 SE_{\tau_A}^2 + \tau_A^2 SE_{\tau_B}^2 + SE_{\tau_A}^2 \cdot SE_{\tau_B}^2 \quad (34.6)$$

在混剂结合毒力系数 CP 及其标准误 SE_{CP} 已知时, 可对结合毒力系数 CP 进行是否等于 1 的 Z 得分统计检验:

$$Z = |CP - 1| / SE_{CP} \quad (34.7)$$

根据 Z 得分值可计算得到显著性概率 p 值。因此, 混剂结合毒力系数 CP 的统计检验: 如果 p>0.05 时, 药剂的混合是相加作用。如果 p<0.05, 且 CP<1, 那么两药剂混合之后药效下降, 亦即药剂之间存在拮抗作用; 如果 CP>1, 两药剂混合有协同 (促进) 作用。

34.3.7 混合制剂毒力统计的实验设计

为有效地估计当量系数, 进行混剂结合毒力系数的统计检验, 须采用Chou等(1996)提出的恒定比例混合设计 (Constant Ratio Combination Design) 制剂混合方式的生测实验。该设计以两种制剂各自的LD₅₀剂量, 按比例进行实验。比例的设置通常以两药剂LD₅₀的4倍 (或 8倍) 混合, 然后进行一系列的二分之一浓度的稀释。如稀释到LD₅₀的2倍、1倍、0.5倍、0.25倍或0.125倍。以获得良好的剂量范围和剂量大小。恒定比例设计的优点是当量系数的估计, τ 近似地等于相对毒理系数ρ。下面给出恒定比率 (LD₅₀比率) 设计是两种药物组合方式。这种设计可以扩大到更大剂量范围。

表34-1 两药剂实验剂量范围及剂量大小实验设计

		药 剂 A					
		0	0.25×LD ₅₀	0.5×LD ₅₀	1.0×LD ₅₀	2.0×LD ₅₀	4.0×LD ₅₀
药 剂 B	0	T _(对照)	T _{DA(1)}	T _{DA(2)}	T _{DA(3)}	T _{DA(4)}	T _{DA(5)}
	0.25×LD ₅₀	T _{DB(1)}	T _{DA,DB(1)}				
	0.50×LD ₅₀	T _{DB(2)}		T _{DA,DB(2)}			
	1.00×LD ₅₀	T _{DB(3)}			T _{DA,DB(3)}		
	2.00×LD ₅₀	T _{DB(4)}				T _{DA,DB(4)}	
	4.00×LD ₅₀	T _{DB(5)}					T _{DA,DB(5)}

在该设计模式下, 取得实验结果之后, 使用表中的第一列里面和对角线上的数据拟合毒力回归方程 (式 34.1 或式 34.2) 可得到药剂 A 的当量系数 τ_A; 使用表中第一行里面和对角线上的数据拟合毒力回归方程 (式 34.1 或式 34.2) 可得到药剂 B 的当量系数 τ_B。以及它们估计的标准误 (SE_{τA} 和 SE_{τB})。

34.3.8 混合制剂毒力系数和 Finney 调和平均数的关系

作者从已出版的文献中收集了 30 组两制剂混合毒力系数生物测定原始数据, 同时

计算实验数据的混剂结合毒力系数 CP 和致死剂量为 50% 时的 Finney 调和平均数 q 。并应用指数曲线方程计算了两者之间的关系。得到关系曲线方程为：

$$q = 0.9947CP^{-0.2987}, \quad r^2 = 0.96$$

根据 30 组实验数据拟合得到的曲线方程，当混剂结合毒力系数 CP 为 1 时，Finney 调和平均数为 0.9947，非常接近于 1。这表明我们提出的混剂结合毒力系数 CP 和 Finney 调和平均数在描述混剂增效效应时是一致的。图 34-5 显示了两种评价方法的关系。

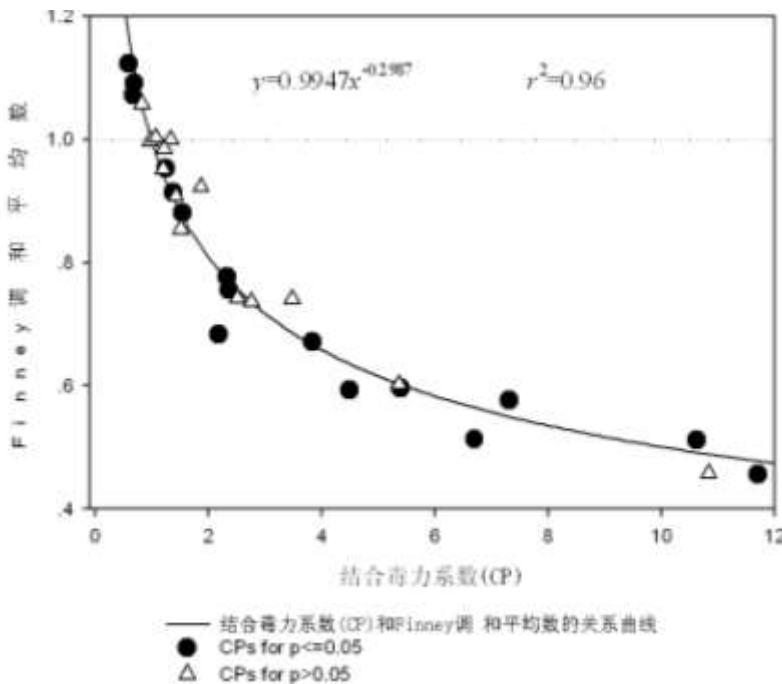


图 34-5 混剂结合毒力系数(CP)和 Finney 调和平均数(q)关系曲线

但是，这里的结合毒力系数 CP 可进行统计检验。以置信概率 $\alpha=0.05$ 水平进行统计推断，图中点 \triangle 表示概率 p 值大于 α ，亦即结合毒力系数 CP 和 1 相比，没有显著差异；点 \bullet 表示概率 p 值小于 α ，亦即结合毒力系数 CP 和 1 相比有显著的差异。30 个实验中，毒力系数 CP 数值大于 1 的有 25 次，其中 13 次有显著差异，而另外 12 次没有显著差异。尤其是其中有 5 次实验，尽管其结合毒力系数大于 5，但统计检验仍有概率 p 值大于 0.05 的样本，即认为没有显著地大于 1 的情形。

34.3.9 混剂结合毒力统计的数据整理与分析

例 1 质反应混剂结合毒力统计分析 在 DPS 计算机数据处理平台上，对两种药剂几个剂量水平的混合毒力统计检验，其数据编辑格式是每一行中按制剂 A 浓度、制剂 B 浓度、试虫数和死亡(抑制)数顺序填入电子表格即可。图 34-6 是对硫磷和敌百虫两农药混合毒力生测的数据编辑格式。

	A	B	C	D	E	F
1		对硫磷	敌百虫	N	R	
2		3.2410	0	50	5	
3		6.4820	0	50	17	
4		12.9660	0	50	39	
5		18.7600	0	50	47	
6		0	13.0930	50	6	
7		0	26.1870	50	12	
8		0	39.2790	50	16	
9		0	52.3730	50	24	
10		0	65.4660	50	38	
11		3.4488	9.6222	50	9	
12		5.1734	14.4336	50	23	
13		6.8420	19.0890	50	27	
14		8.6224	24.0566	50	36	
15		10.3472	28.8688	50	41	
16						

图 34-6 质量反应类型的两制剂混合毒力统计数据格式

选中数据块后，选择功能菜单中“生物测定”栏下的“质反应混配实验生测分析”，按回车键后，系统显示图 34-7 所示的用户界面如下：

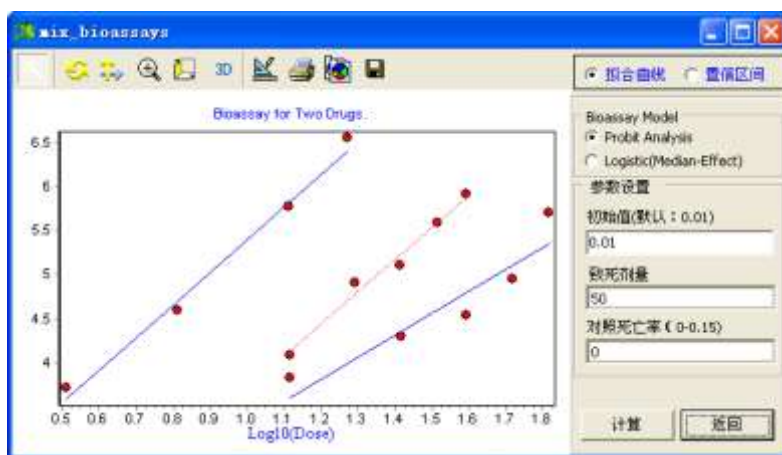



图 34-7 结合毒力系数统计用户界面

在这里用户可选择生测分析的模型，Probit 或 Logistic。对照死亡率不为 0 时，可输入对照死亡率的值，以供系统计算校正死亡率。并输入待求的致死剂量效应指标(致死剂量可同时输入 10, 50, 90 等，各值间用空格隔开)。然后点击“计算”可立即得到分析结果。

如果计算过程中出现数值计算溢出，可能是由于初始值不好所致，这时可在初始值

下面的编辑框中更改初始值之后再计算。

用户界面中显示的当前拟合结果图，可点击按钮将其复制到文字报告中。最后按“返回”按钮，将计算输出到工作表。

Mixture	方程系数	标准误	Z-score	p 值	
Alpha	-4.9527	0.7459	6.6402	0.0000	
Beta	3.6576	0.5345	6.8429	0.0000	
相关系数	0.9916				
皮尔逊(Pearson)卡方值=1.0312				df=3	p=0.7937
LD50	22.5989				
95% CI	20.3742	~	24.8140		
对硫磷	方程系数	标准误	Z-score	p 值	
Alpha	-3.3043	0.3997	8.2662	0.0000	
Beta	3.6970	0.4187	8.8299	0	
相关系数	0.9947				
皮尔逊(Pearson)卡方值=0.9937				df=2	p=0.6084
LD50	7.8303				
95% CI	7.0302	~	8.6798		
敌百虫	方程系数	标准误	Z-score	p 值	
Alpha	-4.1952	0.6197	6.7692	0.0000	
Beta	2.5000	0.3897	6.4158	0.0000	
相关系数	0.9311				
皮尔逊(Pearson)卡方值=7.1992				df=3	p=0.0658
LD50	47.6530				
95% CI	45.5103	~	50.1142		
对硫磷+敌百虫	方程系数	标准误	Z-score	p 值	
Alpha	-3.2904	0.3215	10.2345	0	
Beta	3.6819	0.3294	11.1763	0	
Equivalent	0.1119	0.0405	2.7662	0.0028	
相关系数	0.9936				
皮尔逊(Pearson)卡方值=2.0230				df=6	p=0.9176
LD50	7.8281				
敌百虫+对硫磷	方程系数	标准误	Z-score	p 值	
Alpha	-4.8742	0.5100	9.5578	0	
Beta	2.9288	0.3181	9.2076	0	
Equivalent	5.0919	0.7880	6.4620	0.0000	
相关系数	0.9470				
皮尔逊(Pearson)卡方值=12.1473				df=7	p=0.0958
LD50	46.1541				

Potency	Coefficient	Se	95%confidence interval		Z-score	p-value
对硫磷+敌百虫	0.6810	0.2462	0.1985	1.1635	1.2960	0.1950
敌百虫+对硫磷	0.8367	0.1295	0.5829	1.0905	1.2611	0.2073
Mixture	0.5698	0.2263	0.1262	1.0133	1.9011	0.0573
Adj. by Ratio	0.5422	0.2153	0.1201	0.9642	2.1261	0.0335
LD50的共毒系数=		90.0409				
LD50的 Finney q=		1.1106				
Chou's CI(D50)=		1.1382				

统计分析结果中，系统分别给出了单剂、混剂的生物测定结果。对于结合毒力测定，给出了 LD50 的共毒系数=90.0409，Finney 调和平均数 $q=1.1106$ ，及 Chou's 结合指数 $CI=1.1382$ 。但我们最需要关注的是结合毒力系数（CP）。本例中， $CP=0.5698$ ，标准误 $S_e=0.2263$ ，Z 得分等于 1.9011。 $p=0.0573$ ，95%置信区间为 0.1262~1.0133。统计检验 $p>0.05$ ，即结合毒力是相加作用；即该两种农药混合后有的毒力效应是相加的。


例 2 数量反应混剂结合毒力统计分析 在 DPS 计算机数据处理平台上，对具有数量反应的两种药剂几个剂量水平的混合毒力统计检验，其数据编辑格式是每行中按制剂 A 浓度、制剂 B 浓度、反应(抑制)率顺序填入电子表格即可。图 34-8 是对硫磷和敌百虫两农药混合毒力生测的数据编辑格式。

	A	B	C	E	F
1		AZT	ABT-358	反应率	
2		0.0125	0	0.9984	
3		0.0063	0	0.9930	
4		0.0031	0	0.8787	
5		0.0016	0	0.6213	
6		0.0008	0	0.2498	
7		0.0004	0	0.0839	
8		0	0.0500	0.9974	
9		0	0.0250	0.9921	
10		0	0.0125	0.7589	
11		0	0.0063	0.0125	
12		0	0.0031	0.0067	
13		0.0063	0.0250	0.9995	
14		0.0031	0.0125	0.9960	
15		0.0016	0.0062	0.9133	
16		0.0008	0.0031	0.7871	
17		0.0004	0.0016	0.3569	
18		0.0002	0.0008	0.2317	
19					

图 34-8 数量反应类型的两制剂混合毒力统计数据格式

选中数据块后,选择功能菜单中“生物测定”栏下的“数量反应混配实验生测分析”,按回车键后,系统显示和图 34-7 相同的用户界面。在这里用户可选择生测分析的模型,Probit 或 Logistic。对照死亡率不为 0 时,可输入对照死亡率的值,以供系统计算校正死亡率。并输入待求的致死剂量效应指标(致死剂量可同时输入 10, 50, 90 等,各值间用空格隔开)。然后点击“计算”可立即得到分析结果。

如果计算过程中出现数值计算溢出,可能是由于初始值不好所致,这时可在初始值下面的编辑框中更改初始值之后再计算。

用户界面中显示的当前拟合结果图,可点击按钮将其复制到文字报告中。最后按“返回”按钮,将计算输出到工作表。

Mixture	方程系数	标准误	Z-score	p 值
Alpha	12.5679	0.4153	30.2603	0
Beta	2.8371	0.1795	15.8040	0
相关系数	0.9921			
F-Ratio=249.7659			df(1,4)	p=0.0001
LD50	0.0022			
95% CI	0.0018	~	0.0026	
AZT	方程系数	标准误	Z-score	p 值
Alpha	13.8387	0.3895	35.5333	0
Beta	3.0250	0.1440	21.0122	0
相关系数	0.9955			
F-Ratio=441.5112			df(1,4)	p=0.0000
LD50	0.0012			
95% CI	0.0011	~	0.0014	
ABT-358	方程系数	标准误	Z-score	p 值
Alpha	14.8462	1.6335	9.0887	0
Beta	5.0484	0.8377	6.0265	0.0000
相关系数	0.9611			
F-Ratio=36.3187			df(1,3)	p=0.0092
LD50	0.0112			
95% CI	0.0081	~	0.0154	
AZT+ABT-358	方程系数	标准误	Z-score	p 值
Alpha	13.5905	0.3107	43.7421	0
Beta	2.9315	0.1128	25.9779	0
Equivalent	0.4118	0.0606	6.7918	0.0000
相关系数	0.9935			
F-Ratio=684.6231			df(2,9)	p=0.0000
LD50	0.0012			
ABT-358+AZT	方程系数	标准误	Z-score	p 值
Alpha	12.1762	1.0107	12.0468	0
Beta	3.6453	0.4982	7.3176	0.0000
Equivalent	16.2621	6.2890	2.5858	0.0049
相关系数	0.9368			
F-Ratio=57.3640			df(2,8)	p=0.0000

LD50	0.0108					
Potency	Coefficient	Se	95% confidence interval		Z-score	p-value
AZT+ABT-358	3.8574	0.5679	2.7442	4.9706	5.0311	0.0000
ABT-358+AZT	1.7362	0.6715	0.4202	3.0523	1.0965	0.2729
Mixture	6.6974	2.7975	1.2142	12.1806	2.0366	0.0417
Adj. by Ratio	8.2212	3.4341	1.4905	14.9519	2.1028	0.0355
LD50 的共毒系数=	195.0267					
LD50 的 Finney q=	0.5128					
Chou's CI(D50)=	0.5279					

统计分析结果中, 系统分别给出了单剂、混剂的生物测定结果。对于结合毒力测定, 给出了 LD50 的共毒系数=195.04, Finney 调和平均数 $q=0.5128$, 及 Chou's 结合指数 $CI=0.5279$ 。但我们最需要关注的是结合毒力系数 (CP)。本例中, $CP=6.6974$, 标准误 $S_e=2.7975$, Z 得分等于 2.0366。 $p=0.0417$, 95% 置信区间为 1.2142~12.1806。统计检验 $p<0.05$, 即结合毒力是相加作用; 即该两种农药混合后有的毒力效应为协同作用。

34.4 时间-剂量-死亡率模型分析

34.4.1 基本原理简介

对于包含供试毒品剂量和受试生物在试验期间死亡(反应)率的生测数据分析, 一般用上节中概率分析(probit analysis)的方法, 即将累计死亡率进行概率值转换后分别对剂量或时间作线性回归分析, 通过回归参数的估计而建立直线回归模型, 从而估计剂量效应(如 LD50)或时间效应(如 LT50)。这样的分析方法导致时间与剂量的效应相互分开, 无法使所建模型充分体现实验数据的完整性。因此, 有必要将时间和剂量的效应统一到同一个模型中来, 下面将系统介绍时间-剂量-死亡率(time-dose-mortality, TDM)模型, 按其数学结构又称互补重对数模型 (complementary log-log model), 简称 CLL 模型。

(1) TDM 模型。受试生物在时间 t_j ($j=1, 2, \dots, J$) 被供试因子(如化学或生物制品)的剂量 d_i ($i=1, 2, \dots, I$) 致死的概率可表达如下:

$$p_{ij} = 1 - \exp[-\exp(\tau_j + \beta \lg(d_i))] \quad (34.8)$$

式中, β 被认为是描述“假设存在的每个关键生物受体所接受的活(毒)性分子的平均数目”(Robertson, Preisler, 1992), 实际上可理解为描述剂量效应的斜率, τ_j 是至时间 t_j 的时间效应参数, $\tau_j + \beta \lg(d_i)$ 部分称为线性预测因子(Linear predictor)。

式(34.8)中的 p_{ij} 是剂量 d_i 在第 j 个时间单位内产生的累计死亡概率(Cumulative mortality probability)。由于累计死亡概率在时间上是连续的, 若直接拟合则不符合模型模拟变量的独立性假设, 因此转而考虑剂量 d_i 使受试个体在时间区间 $[t_{j-1}, t_j]$ 内可能遭受的死亡率即条件死亡概率(Conditional mortality probability), 可表示为

$$q_{ij} = 1 - \exp[-\exp(\gamma_j + \beta \lg(d_i))] \quad (34.9)$$

式中, β 的含义与式(34.8)相同, γ_j 为描述时间区间 $[t_{j-1}, t_j]$ 内时间效应的待估参数, 与 τ_j 在时间的含义上有所区别。由于条件死亡概率是独立的随机变量, 所以式(34.9)才更适于进行模型的拟合。获得参数 γ_j 的估计后通过下式估计参数 τ_j :

$$\tau_j = \ln \left(\sum_{k=1}^j \exp(\gamma_k) \right) \quad (34.10)$$

γ_j 和 β 的最大似然估计。通过二项分布最大似然函数

$$\prod_{j=1}^J \prod_{i=1}^I q_{ij}^{r_{ij}} (1 - q_{ij})^{n_{ij} - r_{ij}} \quad (34.11)$$

的逼近, 可获得 γ_j 和 β 的估计。式中, n_{ij} 是在时间 t_1 以剂量 d_i 处理 n_i 头个体后在 t_{j-1} 时刻仍然存活的个体数, r_{ij} 是在时间区间 $[t_{j-1}, t_j]$ 内死亡的个体数。但计算时, 只需输入各剂量下每天的存活数和死亡数, 即 $[d_i, n_{ij}, r_{ij}]$ 。

(2) 参数的变异。在模型拟合中伴随参数 $\hat{\gamma}_j$ 与 $\hat{\beta}$ 估计同时而获得了它们的方差和协方差。由于参数 $\hat{\tau}_j$ 是由式(34.10)计算的, 它的方差及其与 $\hat{\beta}$ 间的协方差需要用 Robertson 和 Preisler (1992) 给出的方程计算, 即

$$\text{var}(\hat{\tau}_j) = \frac{1}{e^{2\hat{\tau}_j}} \sum_{m=1}^j \sum_{n=1}^j e^{\hat{\gamma}_m + \hat{\gamma}_n} \text{cov}(\hat{\gamma}_m, \hat{\gamma}_n) \quad (34.12)$$

$$\text{cov}(\hat{\tau}_j, \hat{\beta}) = \frac{1}{e^{\hat{\tau}_j}} \sum_{n=1}^j e^{\hat{\gamma}_n} \text{cov}(\hat{\beta}, \hat{\gamma}_n) \quad (34.13)$$

(3) 有效致死剂量的估计。在理论上, 在任意时刻 t_j , 导致试虫死亡率 p_{ij} 的有效剂量 $\hat{\theta}_j$ 为

$$\hat{\theta}_j = (\ln(-\ln(1 - p_{ij})) - \hat{\tau}_j) / \hat{\beta} \quad (34.14)$$

并有如下方差:

$$\text{var}(\hat{\theta}_j) = \left((\hat{\theta}_j^2 \text{var}(\hat{\beta}) + \text{var}(\hat{\tau}_j) + 2\hat{\theta}_j \text{cov}(\hat{\tau}_j, \hat{\beta})) / \hat{\beta}^2 \right) \quad (34.15)$$

若 $p_{ij} = 0.5$ 即试虫死亡 50%, 式(34.14)便给出致死中浓度 LD_{50} 的解, 即

$$\hat{\theta}_j = (-0.3665 - \hat{\tau}_j) / \hat{\beta} \quad (34.16)$$

LD_{50} 是考察被测物活性的一个重要指标。显然, 有效致死剂量是一个时间相关函数, 由模型参数 $\hat{\tau}_j$ 和 $\hat{\beta}$ 确定。其他常见指标还有 LD_{90} 等, 变换式(34.14)中的 p_{ij} 即可解出。

(4) 剂量的时间效应估计。剂量 d_i 导致试虫死亡的速度常用使试虫达到死亡率 p_{ij} 所需的时间表示, 如 LT_{50} 和 LT_{90} 等, 它们是 τ_j 的函数。若令

$$\hat{\tau}_j = \tau \ln(d_i)$$

则对于任意剂量 d_i ，以下关系式成立

$$\lg(\text{LT}_{p_{ij}}) = \frac{\ln(-\ln(1-p_{ij})) - \hat{\beta} \lg(d_i)}{\hat{\tau}} \quad (34.17)$$

当 $p_{ij}=0.5$ 和 0.9 时, $\ln(-\ln(1-p_{ij}))$ 分别为 -0.3665 和 -0.8340 , $\text{LT}_{p_{ij}}$ 则分别为 LT_{50} 和 LT_{90} 。以 LT_{50} 的解法为例, 由于 $\hat{\tau}_j$ 是对应于相应的时间变量而拟合求出的, 对于 d_i , 总能找到一个时间区间 $[t_j, t_{j+1}]$ 满足 $p_{ij} < 0.5$ 和 $p_{ij+1} \geq 0.5$, 通过线性插值, 解出相应于 d_i 的 $\hat{\theta}_i$ 即 LT_{50} 为

$$\hat{\theta}_i = t_j + (t_{j+1} - t_j)(p_{ij+1} - 0.5)/(p_{ij+1} - p_{ij}) \quad (34.18)$$

同理可求出 LT_{90} 及其他代表时间效应的指标。

34.4.2 数据整理与操作

TDM 模型处理数据, 其数据编辑和存放格式为: 数据块第一列为供试药剂的浓度, 以后各列为用药后不同时间的试虫存活数。如用乙酰甲胺磷的 5 个浓度处理西红柿上害虫 *O. herbivorous*(Robertson et al., 1992)。处理后试虫死亡过程缓慢, 故适合用 TDM 模型分析数据。分析前, 先将数据编辑定义成图 34-9 的格式。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		剂量	时 间											
2			0	4	8	12	24	48	72	96	120	144	168	
3		1	60	57	46	41	36	36	34	33	33	29	24	
4		2	60	49	35	30	23	20	17	16	16	10	8	
5		3	60	48	31	24	20	17	11	9	9	6	5	
6		5	60	47	35	20	13	8	6	6	3	2	0	
7		16	60	47	24	21	5	0	0	0	0	0	0	
8														

图 34-9 生物测定数据拟合 TDM 模型的数据编辑与格式定义

定义数据块后在功能菜单“生物测定”栏下选择“时间-剂量-死亡率模型”, 执行该功能选项后, 输出结果, 包括: ①模型系数的协方差矩阵; ②所求系数即各参数的估计值、参数标准误、 t 测验值、显著水平; ③模型拟合值、残差、标准残差。本例分析结果如下:

至各个处理时间累积死亡率(略):				
各个处理时间内的相对死亡率(略)				
系数 c_i 的协方差阵(略)				
	系数	标准误	t 值	p
c_1	1.2359	0.1618	7.6373	0.0000
c_2	-2.4390	0.1808	13.4924	0.0000

c3	-1.7178	0.1583	10.8502	0.0000			
c4	-2.1597	0.1985	10.8822	0.0000			
c5	-1.6708	0.1912	8.7365	0.0000			
c6	-2.1337	0.2766	7.7130	0.0000			
c7	-2.0725	0.2855	7.2581	0.0000			
c8	-3.1098	0.5137	6.0533	0.0000			
c9	-3.3210	0.6685	4.9675	0.0000			
c10	-1.5918	0.2726	5.8396	0.0000			
c11	-1.6383	0.3243	5.0513	0.0000			
No	死亡率	拟合值	残差	残差	死亡数	理论值	卡方
X1, 1	0.0500	0.0835	0.0336	0.2229	3.0000	5.0100	0.8822
X1, 2	0.1833	0.1189	-0.0644	-0.4087	11.0000	7.1300	2.3783
...
X10, 4	1.0000	0.3693	-0.6307	-4.6459	2.0000	0.7400	3.4152
X10, 5	数据不合理, 未进行处理.						
对数似然函数值[-2*Lg(Likelihood)]=1187.79270, 偏差 Scaled Deviance= 1187.79270, AIC=1209.79270							
Hosmer & Lemeshow 拟合度测试							
组别	合计	实际死亡	理论死亡	实际存活	理论存活	卡方值	
1	126	7	6.45	119	119.55	0.0501	
2	112	8	10.63	104	101.37	0.7178	
...
10	169	69	74.67	100	94.33	0.7711	
卡方统计量=9.6586 df= 8 p = 0.28981							
系数 τ var(τ) cov(τ , β)							
τ 1	1.2359	0.0305	0.0305				
τ 2	-2.4390	0.0380	-0.0216				
τ 3	-1.3216	0.0237	-0.0208				
τ 4	-0.9621	0.0211	-0.0202				
τ 5	-0.5618	0.0195	-0.0196				
τ 6	-0.3731	0.0188	-0.0189				
τ 7	-0.2053	0.0174	-0.0176				
τ 8	-0.1519	0.0169	-0.0171				
τ 9	-0.1108	0.0170	-0.0169				
τ 10	0.0942	0.0155	-0.0152				
τ 11	0.2570	0.0147	-0.0138				
各时段致死对数剂量估计							
时段	LD50	SE	LD90	SE	各对数剂量致死中时间	对数剂量	LT50
1.0000	1.6768	0.1831	2.6482	0.2997	0.0000		5.0386
2.0000	0.7727	0.0796	1.7441	0.1692	0.3010		3.5429
3.0000	0.4819	0.0756	1.4533	0.1325	0.4771		3.0142
4.0000	0.1580	0.0960	1.1294	0.0959	0.6990		2.2448
5.0000	0.0054	0.1104	0.9767	0.0849	1.2041		1.4817
6.0000	-0.1305	0.1213	0.8409	0.0785	对数剂量		LT90
7.0000	-0.1736	0.1248	0.7978	0.0771	0.4771		9.9291
8.0000	-0.2069	0.1287	0.7644	0.0766	0.6990		8.4278
9.0000	-0.3727	0.1426	0.5986	0.0736	1.2041		3.8081
10.0000	-0.5045	0.1544	0.4669	0.0746			

上述结果中, 表列的累计死亡率和相对(即条件)死亡率系根据所输入数据按其定义计算的。系数 c_i 协方差阵是对式(34.9)进行最大似然逼近拟合式(34.11)的同时所估计的各系数变异及相互关联程度, 用于计算各参数估计值的方差、标准误并进行 t 检验。方程系数 c_i 是式(34.9)中的各参数估计值, 其中 c_1 为 β 的估计值 $\hat{\beta}$, $c_2 \sim c_9$ 是该式中 γ_j 的估计值 $\hat{\gamma}_j$, c 后的数码对应于 γ 的下标, 即表示试虫接受处理后的相应时间区间。

在方程系数表后给出的是由所估参数代入式(34.9)后计算出的模拟值与观察值之间的比较, 以便进行常规的 Pearson 卡方检验及表明模型拟合度的统计量。注意表列中有些行显示“数据不合理, 未进行处理”的字样, 这主要因为在最高剂量下试虫在整个实验观察完成前的数个时间区间已经全部死亡, 即累计死亡率在前一个时间区间已达 100%, 此后各时间区间的条件死亡率在数学上无法计算(0/0)。每当这种情况出现时, 本数据处理系统自动将该数据视为无效, 但截至前一时间区间的拟合仍然有效并且不影响其余各剂量数据的继续拟合。这是本系统的灵活性所在, 而不是模型拟合本身出了问题。

上述实例的 Pearson 卡方检验未能通过, 因为 $\chi^2=76.10$, 在 $df=34$ 时, $\chi_{0.05}^2=49.80$, 现 $\chi^2 > 49.8$, 表明模型异质性的假设成立。一般认为, 这样复杂的模型要通过 Pearson 卡方检验显得过于苛刻。于是, Nowierski 等(1996)提出用 Hosmer 和 Lemeshow(1989)改进的方法进行拟合度测验。本系统采用了这种方法。Hosmer-Lemeshow 检验表明该实例的数据拟合是成功的, $\chi^2=9.66$, 在 $df=8$ 时, $\chi_{0.05}^2=15.51$, 现 $\chi^2 < 15.51$, 故否决所拟合模型异质性的假设。

结果中最后表列出根据式(34.9)参数估计值由式(34.10)、式(34.12)、式(34.13)所计算出的式(34.8)的各个参数及其方差和协方差。其中, τ_1 代表 $\hat{\beta}$, $\tau_2 \sim \tau_9$ 表示该式中 τ_j 的估计值 $\hat{\tau}_j$, τ 后的数码表示试虫接受处理后的时间长度(如处理后天数)。

如果 Pearson 卡方检验能通过, 也就是说其显著水平大于等于 0.05, 则模型的拟合效果很好。而实际上, 一般 Pearson 卡方检验不能通过(显著水平小于 0.05), 这表明模型异质性的假设成立。但 Nowierski 等(1996)提出用 Hosmer 和 Lemeshow 改进的方法进行拟合度测验, 如 Hosmer-Lemeshow 检验的卡方值的显著水平大于等于 0.05, 则可认为数据拟合是成功的。

系统还根据式(34.8)计算各个参数及其方差和协方差, 根据式(34.14)~(34.18)计算剂量和时间效应指标, 输出剂量指标(LD₅₀, LD₉₀)和时间效应指标(LT₅₀, LT₉₀)。

34.5 药物代谢动力学(房室模型)

34.5.1 基本原理简介

为了定量地描述药物体内过程的动态变化规律性, 常常要借助数学的原理和方法来系统地阐明体内药量随时间而变化的规律性。房室模型理论从速度论的角度出发, 建立数学模型来模拟药物在机体内的运行过程。它将整个机体视为一个系统, 并将该系统按动力学特性划分为若干个房室(compartment), 把机体看成是由若干个房室组成的一个

完整系统，即如下图所示的房室模型（compartment model）。

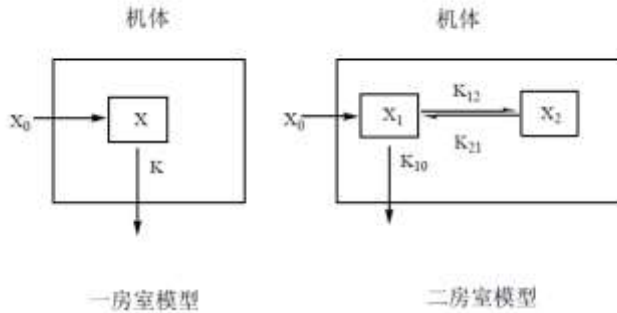


图 34-10 一房室和二房室模型示意图

房室模型中的房室划分主要是依据药物在体内各组织或器官的转运速率而定，只要药物在其间的转运速率相同或相似，就可归纳成为一个房室，但这里的房室只是数学建模过程中的一个抽象概念，并不代表解剖学上的任何一个组织或器官，因此房室模型的划分具有抽象性和主观性。但房室的概念又是与体内各组织器官的生理解剖学特性（如血流量、膜通透性等）有一定联系。同一房室中各组织部位的药物浓度并不一定相同，但药物在其间的转运速率相同或相似。示意图如图 34-10。

在应用房室模型研究药物的动力学特征时，最常采用的方法是把机体表述为由一些房室组成的系统，并假定药物在各房室间的转运速率以及药物从房室中消除的速度均符合一级反应动力学。一级反应动力学数学式表达为 $dx/dt = -kx$ ，式中 x 为反应物的量， dx/dt 表示反应速度， k 为速度常数，负号表示反应朝反应物量减少的方向进行。在房室模型的理论中假设药物在各房室间的转运速率以及药物从房室中消除的速率均符合一级反应动力学，因此其动力学过程属于线性动力学，故房室模型又称线性房室模型。因此该模型只适合于描述属于线性动力学药物的体内过程。

一房室模型基本特征是静脉注射给药后血药浓度对数-时间为近似直线(图 34-11A)；而二房室模型静脉注射给药后的血药浓度对数-时间呈双指数曲线；这有助于判别是否是一室或二室模型。

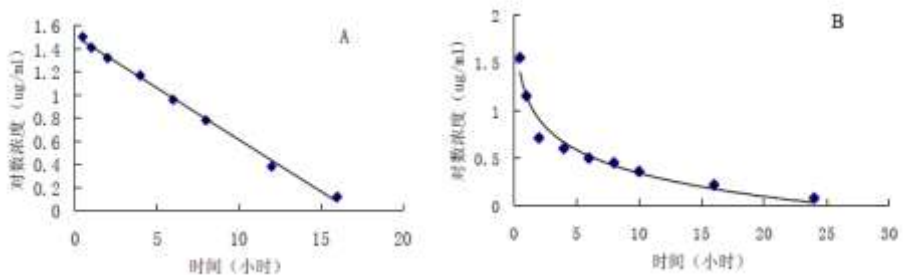


图 34-11 时间-血药浓度(对数)曲线图

在运用房室模型估计药物动力学参数时，用户所选择的模型将直接影响到药物动力学参数(如半衰期等)的估算值，因此模型的选择显得尤为重要。在进行药物动力学分析

时应首先确定所研究的药物属于模型种类。这一般可先用半对数图进行初步判断，但尚需计算机拟合后加以进一步地确定。在用计算机进行药物动力学分析时常用的判别标准有拟合度 $r^2 = 1 - \frac{\sum_{i=1}^n (C_i - \hat{C}_i)^2}{\sum_{i=1}^n (C_i - \bar{C})^2}$ ， F 统计量及其 p 值，以及 AIC (Akaike 信息准则)。应用时一般可用 p 值识别模型($p < 0.05$)。当血药高低浓度悬殊大时可考虑不同权重系数影响，采用加权法估计参数。

34.5.2 房室模型分析示例

DPS 可处理不同房室数的静脉推注、静脉滴注及非静脉用药 (包括有或无滞后时间) 的各种房室模型的时间血浓数据，计算药动学参数。计算主要结果包括：Aikake's 信息数据(AIC)、拟合优度值、 F 检验、 $C-T$ 图等。药物动力学房室模型建模分析，其数据编辑和存放格式为：数据块第一列为时间，第二列上各个时刻的药物浓度。如某口服药，剂量 400mg，测得各个时刻的血药浓度，编辑成图 34-12 格式所示。

	A	B	C	D
1		药剂序号	死亡率	
2		0.10	4.70	
3		0.30	13.20	
4		0.50	20.80	
5		1	36.30	
6		2.50	61.40	
7		5	68.10	
8		7.50	61.10	
9		10	52.10	
10		20	27.50	
11		25	21.10	
12		30	16.90	
13		40	11.40	
14		50	8.20	
15		60	5.90	
17				

图 34-12 房室建模数据格式

选中数据块后在功能菜单“生物测定与药物动力学”栏下选择“药物代谢动力学模型分析”，执行该功能选项后，系统弹出用户参数输入、调整界面如图 34-13。

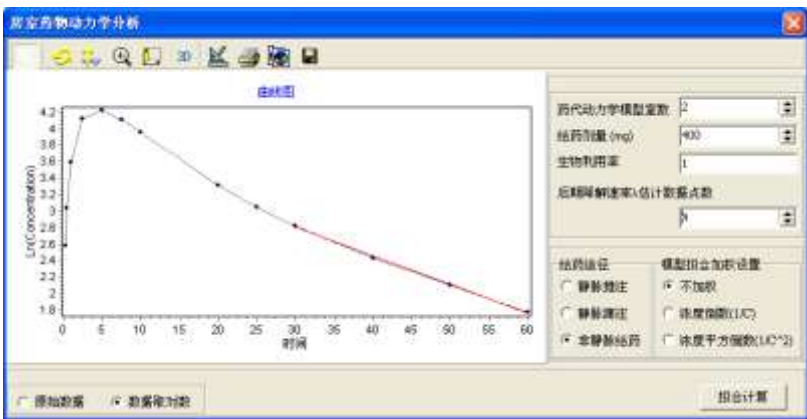


图 34-13 非房室分析用户界面

在图 34-13 用户界面中,左侧的曲线图可辅助用户识别血液浓度变化特性,以确定房室数量、后期降解速率计算的数据点数。在右边的对话框里面,用户可根据曲线特征、实验类型调整相关参数,如用户根据半对数图选择终末消除相。然后点击“拟合计算”执行拟合计算。本例分析结果如下:

药物动力学加权最小二乘模型分析					
"模型:非静脉给药-2室模型"					
参数	估计值	参数	估计值		
吸收常数(K _a)	0.4775	总清除率	0.2403		
中央室容积	3.9844	末期分布容积(V _z)	7.9889		
消解常数(K _e)	0.0603	稳态分布容积(V _{ss})	5.9974		
K ₁₂	0.0306	平均驻留时间(MRT)	24.9578		
K ₂₁	0.0605	半衰期(0) =	1.4515		
样点数 M=	14	半衰期(1) =	5.7158		
残差误差 s =	0.0473	半衰期(2) =	23.0440		
决定系数	1.0000	后期降解速率 Lambda=	0.0349		
方差比值 F=	737537.24	AUC(0-N)=	1482.3500		
	Prob(>F) =	AUC(0-∞)=	1651.5711		
	0				
Akaike 信息准则(AIC)	-44.6954				
Ti	观察值	拟合值	残差	标准偏差	标准残差
0.1000	4.7000	4.6595	0.0405	0.0473	0.8578
0.3000	13.2000	13.2129	-0.0129	0.0473	-0.2734
...
60	5.9000	5.9380	-0.0380	0.0473	-0.8041

上述 DPS 输出的房室模型分析结果分为 4 个部分:

1. 房室模型类型及参数估计值,本例中模型药物动力学模型上非静脉给药-2室模型;拟合方法上加权最小二乘。计算得到的模型参数有:吸收常数(K_a) = 0.4775;中央室容积=3.9844;消解常数(K_e)=0.0603; K₁₂ = 0.0306; K₂₁=0.0605。

2. 模型拟合结果的检验,这里拟合残差误差 s=0.0473;决定系数=1.0000;方差比值 F=737537.2386; Prob(>F)=0.0001;远小于 0.05;故该房室模型可以应用。

3. 根据模型参数计算的药物动力学指标:总清除率 = 0.2403;末期分布容积(V_z) = 7.9889;稳态分布容积(V_{ss}) = 5.9974;平均驻留时间(MRT) = 24.9578;半衰期(0) = 1.4515;半衰期(1) = 5.7158;半衰期(2) = 23.0440;后期降解速率 Lambda = 0.0349; AUC(0-N)=1482.3500 即 AUC(0-∞)=1651.5711。

4. 各采样时间点数据的拟合情况,即观察值、拟合值、残差、标准偏差及标准残差。

34.6 药物代谢动力学(非房室参数估计)

34.6.1 非房室统计矩方法简介

实际工作中由于情况复杂,且模型拟合具不确定性,有时实际数据和经典房室模型吻合很不理想,于是上世纪 70 年代前后有人提出用非房室模型来处理药动学数据。非房

室模型仅假设药物末端以单指数形式消除,不受经典房室模型限制,故可用于任何房室。目前体内数据解析中非房室模型已渐成主流处理方法而被推荐采用。但要注意统计矩方法和房室模型各有优缺点,并不互相排斥。

统计矩方法以概率论和数理统计学中的统计矩(Statistical Moment)方法为理论基础,对数据进行解析,包括零阶矩、一阶矩和二阶矩,体现平均值、标准差等概念,反映了随机变量的数字特征。在药动学中,零阶矩为 AUC,和给药剂量成正比,是一个反映量的函数;一阶矩为 MRT,反映药物分子在体内的平均停留时间,是一反映速度的函数;二阶矩为 VRT,反映药物分子在体内的平均停留时间的差异大小。各阶统计矩定义以及计算公式如下:

曲线下面积 AUC (Area Under Curve) 给药以后,血药浓度的经时过程可以看成随机分布曲线,不管何种给药途径或何种房室模型,其零阶矩 AUC 定义如下(和普通的零阶矩的定义不同,它未经标准化):

$$\begin{aligned} AUC &= \int_0^{\infty} C(t) dt \\ &= \int_0^{t^*} C(t) dt + \int_{t^*}^{\infty} C(t) dt \end{aligned}$$

这是因为在通常单剂量给药的药动学研究中,血药浓度只能观察至某一个时间 t^* ,于是计算 $0-\infty$ 时间内的血药浓度-时间曲线下面积 AUC 时划分为两个阶段,即从 $0-t^*$ 和 $t^*-\infty$ 时间内曲线下面积之和(图 34-14)。

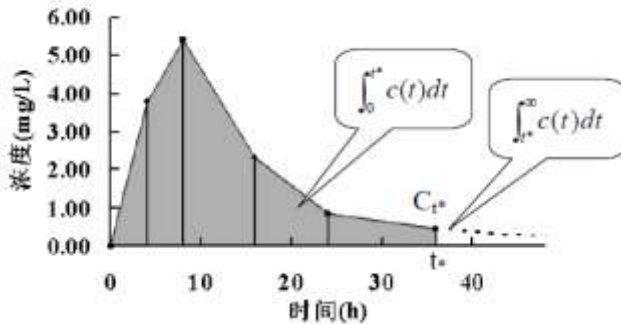


图 34-14 时间-血药浓度曲线及 AUC 定义示意图

AUC 的计算方式有线性梯形法和样条插值等,后者由于计算繁杂,而且对于提高数据的准确性也未见明显益处,目前应用较少。因此国际上通用线性梯形法。

平均驻留时间 MRT (Mean Residence Time), 称为一阶原点矩,其计算方法如下:

$$MRT = \int_0^{\infty} tc(t) dt / AUC = AUMC / AUC$$

$$AUMC = \int_0^{t^*} tc(t) dt + \int_{t^*}^{\infty} tc(t) dt$$

MRT 代表药物分子在体内的平均驻留时间,可理解为一次给药含有无数个药物分子,例如对于分子量为 300g/mol,即使 1mg 的药物也含有 2×10^{18} 个药物分子($=0.001/300 \times 6.023 \times 10^{23}$)。这些药物在体内停留的时间并不一致,有些被迅速排泄,而有一些可能停留较长的时间,极少数甚至可能停留终生。上述平均驻留时间 MRT 中“平均”就是这些药物分子停留时间的平均值。对于线性药物动力学过程,静注后 MRT 就表示消除给药量的 63.2% 所需要的时间,但如果存在吸收项, MRT 会大于消除给药量的 63.2% 所需要的时间。

清除率 指单位时间内,从体内消除的药物的表观分布容积数,其单位为 L/h 或 L/h/kg,表示从血中清除药物的速率或效率,它是反映药物从体内消除的另一个重要的参数。总清除率 Cl 等于总消除速率常数 dx/dt 对全血或血浆药物浓度 C 的比值, $Cl = (dx/dt) / C$ 。

稳态表观分布容积 在药物动力学参数中,稳态表观分布容积 V_{ss} 亦是重要参数之一。根据统计矩原理, V_{ss} 可在药物单剂量静注后通过清除率与平均驻留时间的简单乘积进行计算,即 $V_{ss} = Cl \cdot MRT$ 。

此外,药物动力学参数还有两个直观指标,即药峰时间(t_{max})和药峰浓度(C_{max})。药物经血管外给药吸收后出现的血药浓度最大值称为药峰浓度,达到药峰浓度所需的时间为药峰时间,两者是反映药物在体内吸收速率的两个重要指标,常被用于制剂吸收速率的质量评价。

34.6.2 非房室统计矩参数的变异估计

非房室模型分析,当处理有重复时,理论上我们可以估计出各个参数估计值的变异程度,即标准误。但实际上如何估计各个参数的标准误,其统计方法在这些年来一直处于探索之中。

较早的研究是对曲线下面积(AUC)的标准误的估计,如 Bailer (1988)和 Yuan (1993)提出的标准误估计方法。这些年来一直在应用。

近几年, Wolfsegger 等(2009)和 Jaki 等(2012)在总结前人工作基础上,提出了基于弱正态分布假设为前提的非房室模型参数的标准误的估计方法。这时可以得到各个参数的标准误的估计。

鉴于药物浓度曲线数据,一般来说可能与符合正态分布有一定距离,在 DPS 系统中,作者在研究非房室模型参数估计及标准误计算过程中,发现应用 Jackknife 抽样方法来估计参数及其标准误可能是一种较好的方法。因为用该方法和样本的分布关系不是很密切,且一般来说得到未知参数的估计量偏性较小或具有无偏性,可能得到的参数及标准误的精确较高,估计结果更具有应用价值。

34.6.3 DPS 应用实例

DPS 提供的非房室分析(Non-compartmental Analysis)可根据数据计算 AUC_{0-n} 、 $AUC_{0-\infty}$ 、 $AUMC_{0-n}$ 、 $AUMC_{0-\infty}$ 、MRT, CL 和 V_{ss} 等参数;用户界面给出了半对数图,用户根据它可选择终末消除相。试验有重复时,系统采用 Wolfsegger(2009)和 Jaki 等(2012)提出

的方法，或作者提出的非参数方法计算各个参数估计的近似标准误和 95% 置信区间。

如果试验没有重复，数据格式和房室模型参数估计相同(图 34-12)。如实验有重复，数据格式为：第一列为血药测定的时间、第二列以后各列为相应时刻的血药浓度测定值。如图 34-15 所示的上 3 个重复的血药浓度资料，作为例子进行非房室模型参数估计。

	A	B	C	D	E	F
1		Time	Conc.1	Conc. 2	Conc. 3	
2		0	0	0	0	
3		0.08	2.01	2.85	2.43	
4		3	0.85	1.00	0.91	
5		6	0.46	0.35	0.63	
6		9	0.39	0.32	0.45	
7		16	0.11	0.18	0.19	
8		24	0.08	0.09	0.06	
17						

图 34-15 非房室分析数据格式

图 34-15 是某静脉注射药，剂量 200mg，重复 3 次，测得的各个时刻的血药浓度。在 DPS 系统下，选中数据块后在功能菜单“生物测定与药物动力学”栏下选择“药物代谢动力学参数非房室模型分析”，执行该功能选项后，系统弹出用户参数输入、调整界面如图 34-16。

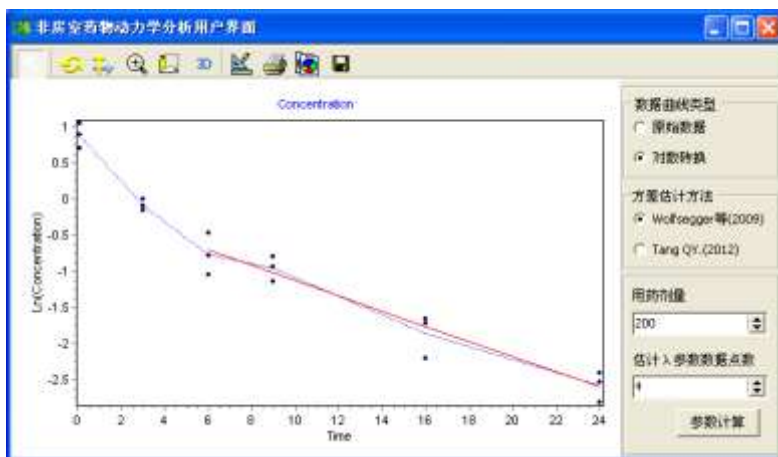


图 34-16 非房室分析用户界面

在图 34-16 用户界面中，左侧的曲线图可辅助用户识别血液浓度变化特性，后期降速率计算的数据点数。在右边的对话框里面，用户可根据曲线特征、用户根据半对数图选择终末消除相。然后点击“参数计算”执行计算。本例分析结果如下：

Dose=200			
Num of data points for lambda estimation=4			
Lambda=0.105041		SE=0.004634	
Ti	Conc.	AUC	AUMC
0	0	0	0

0.0833	2.4300	3.6450	0.3038		
3	0.9200	6.3667	8.4688		
6	0.4800	7.8067	17.1088		
9	0.3867	9.7400	34.5088		
16	0.1600	10.9400	53.7088		
24	0.0767	11.2467	61.0688		
"	0-∞"	11.9765	85.5343		
Parameter	Estimate	SE	Asymptotic 95% CI		
AUC	11.9765	0.5338	10.9303	to	13.0228
AUMC	85.5343	6.7574	72.2898	to	98.7788
MRT	7.1418	0.4941	6.1733	to	8.1103
CL	16.6993	0.7443	15.2405	to	18.1581
Vss	119.2634	10.1312	99.4063	to	139.1205

上述 DPS 输出的房室模型分析结果分为 2 个部分:

1. 各个时期的 AUC 和 AUMC 值: 如曲线下面积 $AUC_{(0-N)} = 11.2467$, $AUC_{(0-\infty)} = 11.9765$; 平均驻留时间 $MRT_{(0-N)} = 61.0688$, $MRT_{(0-\infty)} = 85.5343$ 。
2. 各个药动力学指标的参数估计值、标准误, 及其 95% 的置信区间。如这里的药动力学参数 $AUC_{(0-\infty)}$ 的估计值为 11.9765, 标准误 $S_e = 0.5338$; 95% 置信区间从 10.9303 到 13.0228。

应用作者提出的非参数估计方法, 得到的参数 95% 置信区间, 比 Wolfsegger 等(2009)提出的近似法大, 但比 Bootstrap-*t* 法的置信区间小。

Parameter	Estimate	SE	95% confidence interval		
Lambda	0.1050	0.0116	0.0823	to	0.1278
AUC(0-N)	11.2467	0.6333	10.0054	to	12.4880
AUC(0-∞)	11.9775	0.6361	10.7306	to	13.2243
AUMC(0-N)	61.0687	4.6282	51.9974	to	70.1399
AUMC(0-∞)	85.5782	8.0360	69.8277	to	101.3287
MRT(0-N)	5.4300	0.3136	4.8154	to	6.0446
MRT(0-∞)	7.1451	0.5898	5.9892	to	8.3010
CL	16.7004	0.8864	14.9631	to	18.4376
VSS	119.3281	12.2403	95.3372	to	143.3191

参 考 文 献

- M. 吉伯尔迪, D. 佩里尔著, 朱家璧译. 1987. 药动力学(第 2 版)[M]. 北京: 科学出版社.
- 崔安义; 许亚. 1986. 杀虫剂混用及其毒效的计算方法[J]. 医学动物防制, (03) : 37-39
- 冯明光, 唐启义, 胡国成, 黄世文. 1996. 球孢白僵菌对七种蚜虫的感染反应: 时间-剂量-死亡率模型分析. 应用基础与工程科学学报[J], 3: 22-33.
- 顾中言, 许小龙, 徐德进等. 2009. 杀虫剂混合使用的增效作用评判分析. 农药学报[J]. 11(3) : 304-311.
- 欧晓明, 林雪梅, 盛书祥, 余淑英. 1997. 评价杀虫剂混用联合作用的六种方法比较. 农药科学与管理, 20-23
- 谭福杰. 1986. 杀虫剂混用的生物测定方法. 昆虫知识, 23(9): 279-280
- Agresti, A. 1990. Categorical Data Analysis[M]. New York: John Wiley & Sons.
- Berenbaum MC. What is synergy. Pharmacol Rev[J], 1989, 41(2): 93-141
- Bliss C.I. 1939. The toxicity of poisons applied jointly. Annals of Applied Biology[J], 26 : 585-615.

Collett, D. 1991. *Modelling Binary Data*, London: Chapman and Hall.

- EL-Lakwah F.A., Mohamed R.A., EL-AZIZ A.E. 2001. Toxicity and action of cumin seeds extract with certain controlled atmospheres against stored-product insects. Donahaye, E.J., Navarro, S. and Leesch J.G. [Eds.] (2001) Proc. Int. Conf. Controlled Atmosphere and Fumigation in Stored Products, Fresno, CA. 29 Oct. - 3 Nov. 2000, Executive Printing Services, Clovis, CA, U.S.A.[M].
- Finney D.J. 1971. *Probit Analysis*[M]. 3rd edition. Britain: Cambridge University Press.
- Finney D.J. 1942. The analysis of toxicity tests on mixtures of poisons. *Annals of Applied Biology*[J], 29,82-94
- Giltinan D.M., Capizzi T.P., Malani H. 1988. Diagnostic tests for similar action of two compounds, *Applied statistics*[J],37, 39-50
- Greco WR, Bravo G, Parsons JC. 1995. The search for synergy: a critical review from a response surface perspective. *Pharmacol Rev*[J], 47:331-385
- Hewlett P.S. 1969. Measurement of the potencies of drug mixtures. *Biometrics*[J], 25, 477-487
- Hewlett P.S., Plackett R.L. 1952. Similar Joint Action of Insecticides, *Nature*[J], 169 : 198-199
- Hewlett P.S., Plackett R.L. 1957. Quantal Responses to Mixtures of drugs. *Nature*[J], 180:712-713.
- Hewlett P.S., Plackett R.L. 1959. A Unified Theory for Quantal Responses to Mixtures of Drugs Non-Interactive Action. *Biometrics*[J],15:591-610
- Jaki T. and Wolfsegger M. J. 2009. A theoretical framework for estimation of AUCs in complete and incomplete sampling designs. *Statistics in Biopharmaceutical Research*[J], 1(2):176-184.
- Jaki T. and Wolfsegger M. J. 2012. Non-compartmental estimation of pharmacokinetic parameters for flexible sampling designs. *Statistics in Medicine*[J], 31(11-12):1059-1073.
- Koouman S.A.L.M. 1981. Parametric analyses of mortality rates in bioassays. *Water Research*[J], 15: 107-119
- Lee J.J. and M. Kong. 2009. Confidence Intervals of Interaction Index for Assessing Multiple Drug Interaction. *Stat Biopharm Res*[J], 1(1): 4-17. doi:10.1198/sbr.2009.0001
- Nowierski R M, Zeng Z, Jaronski S, Delgado F & Swearingen W. 1996. Analysis and modeling of time-dose-mortality of *Melanoplus sanguinipes*, *Locusta migratoria migratorioides*, and *Schistocerca gregaria* (Orthoptera: Arididae) from Beauveria, Metarhizium, and Paecilomyces isolates from Madagascar[J]. *J. Invertebr. Pathol*[J], 67: 236-252.
- Plackett R.L., Hewlett P.S. 1952. Quantal Responses to mixtures of poisons. *Journal of the Royal Society. Journal of the Royal Statistical Society, series B*[J],14:141-161
- Plackett R.L., Hewlett P.S. 1963. A Unified Theory for Quantal Responses to Mixtures of Drugs: The Fitting to Data of Certain Models for Two Non-Interactive Drugs with Complete Positive Correlation of Tolerances. *Biometrics*[J], 19, 517-531
- Preisler H K, Robertson J.L. 1989. Analysis of time-dose-mortality data. *J. Econ. Entomol*[J], 83: 1534-1542.
- Sampford M.R.(1952). Studies in the principles of Phytotoxicity II. Experimental designs and techniques of statistical analysis for the assessment of toxicity. *Journal of Experimental Botany*[J]. 3, 28-46
- Shitara Y, Sato H, Sugiyama Y. 2005. Evaluation of drug-drug interaction in the hepatobiliary and renal transport of drugs. *Annu Rev Pharmacol Toxicol*[J]. 45:689-722
- Sun Y.P., Johnson E.R. 1960. Analysis of Joint Action of Insecticides Against House Flies. *Journal of Economic Entomology*[J], 53: 887-892
- Tallarida R.J., Kimmel H.L., Holtzman S.G. 1997. Theory and statistics of detecting synergism between two active drugs: cocaine and buprenorphine. *Psychopharmacology*[J], 133 :378-382
- Ting-chao Chou and Talalay P. 1984. Quantitative Analysis of Dose-Effect Relationships: The combined effects of multiple drugs or enzyme inhibitors. *Advances in enzyme regulation*[J], 22:27-55
- Ting-chao Chou. 2006.Theoretical Basis, Experimental Design, and Computerized Simulation of Synergism and Antagonism in Drug Combination Studies. *Pharmacol. Rev*[J],58:621-681
- William R.G. 1991. Noncompartmental Versus Compartmental Modelling in Clinical Pharmacokinetics *Clin. Pharmacokinetic*[J]. 20 (4) :253-262
- Wolfsegger MJ, Jaki T. 2009. Non-compartmental estimation of pharmacokinetic parameters in serial sampling designs. *Journal of Pharmacokinetics and Pharmacodynamics*[J]. 36(5):479-494.
- Yuan J.1993. Estimation of variance for AUC in animal studies. *J. Pharm. Sci*[J], 82:761-763

第 35 章 诊断试验评价

随着先进技术的迅猛发展,各种诊断试验(包括诊断设备、试剂、方法等)层出不穷,但并不是每一种诊断试验肯定比常规或旧试验好,经过一段时间的临床应用,不少试验被淘汰。如果医学工作者,特别是临床医生,能够掌握一些评价与解释诊断试验的有关知识,及时取舍或灵活运用新的诊断试验,不但可提高自身的业务素质,而且还能减少患者的不必要经济开支,更好地为患者服务。

35.1 诊断试验常用指标

对于诊断试验(diagnostic test)的评价,首先应知道受试者(人、动物或影像等)的真实类别,即哪些属于对照组(或无病组、正常组、噪声组等),哪些属于病例组(或有病组、异常组、信号组等)。划分病例与对照这两个组的标准就是金标准(gold standard)。研究中常见的金标准有:活组织检查、尸体解剖、手术探查和跟踪随访结果等。尽管金标准不需要十全十美,但是它们应比评价的诊断试验更加可靠,且与评价的诊断试验无关。对于按金标准确定的二项分类总体,如病例与对照(分别记为 D_+ 与 D_-),采用诊断试验检测的结果可分别写成阳性与阴性(记为 T_+ 与 T_-),资料可列成表 35-1 的四格表形式。表中有四个可能结果,其中两个是正确的,即病例被诊断为阳性(真阳性, TP)和对照被诊断为阴性(真阴性, TN);两个是错误的,即病例被诊断为阴性(假阴性, FN)和对照被诊断为阳性(假阳性, FP)。

表 35-1 诊断资料 2×2 四格表

诊断结果(T)	金标准(D)		合计
	病例(D_+)	对照(D_-)	
阳性(T_+)	TP(真阳性)	FP(假阳性)	TP+FP
阴性(T_-)	FN(假阴性)	TN(真阴性)	FN+TN
合计	TP+FN	FP+TN	N

为了确定诊断试验检测结果阳性时患病的概率有多大,可利用下列贝叶斯公式

$$P(D_+|T_+) = P(T_+|D_+)P(D_+)/[P(T_+|D_+)P(D_+)+P(T_+|D_-)P(D_-)]$$

式中 $P(D_+)$ 称为先验概率(prior probability),它不依赖于其他变量的变化而变化,在人群中称为患病率,此概率一般可从各种有关报道或参考书中获得,也可通过流行病学调查得到; $P(D_-)=1-P(D_+)$; $P(T_+|D_+)$ 是病例组中诊断试验检测结果为阳性的概率,也就是后面提到的灵敏度或真阳性率; $P(T_+|D_-)$ 是对照组中诊断试验检测结果为阳性的概率,

也就是后面提到的(1-特异度)或假阳性率; $P(D_+|T_+)$ 为后验概率(posterior probability),也就是后面提到的阳性预测值,一般情况下该指标难以直接得到,需要采用贝叶斯公式计算。例如,采用心电图(electrocardiogram, ECG)作为诊断工具,对具有急性持久胸痛的 700 名患者进行诊断,经证实有 520 例出现心肌梗塞,其余 180 例没有出现(表 35-2),基于贝叶斯公式,可计算得到 ECG 诊断试验的几个常用评价指标。

表 35-2 ECG 诊断试验的结果

ECG 诊断结果	心肌梗塞		合 计
	出现	不出现	
阳性	416	9	425
阴性	104	171	275
合计	520	180	700

评价诊断试验的常用指标有一致百分率、灵敏度、特异度、Youden 指数、阳性似然比、阴性似然比、阳性预测值和阴性预测值。

35.1.1 一致百分率

一致百分率是病例正确诊断为阳性与对照正确诊断为阴性的例数之和占总例数的百分率。计算公式为

$$(TP+TN)/N \times 100\%$$

其标准误为

$$SE_{\text{一致百分率}} = \sqrt{(TP+TN)(FP+FN)/N^3}$$

一致百分率很大程度上依赖于患病率,如某病的患病率为 5%,即使不采用诊断试验,且将所有研究个体划归为阴性,也可得到一致百分率为 95%;其次,它没有利用假阴性和假阳性的信息,相同的一致百分率可能有十分不同的假阴性和假阳性;第三,它还受诊断界点(cut-off point)(图 35-1)的限制。因此,诊断试验评价只用该指标粗略地表达诊断试验的一致性,更常用的诊断试验评价指标是灵敏度、特异度等。

35.1.2 灵敏度

实际患病且被诊断为阳性的概率就是灵敏度(sensitivity, Sen),也称为真阳性率(true positive rate, TPR),即

$$Sen = P(T_+|D_+) = TP / (TP + FN) = TPR$$

其标准误

$$SE_{Sen} = \sqrt{TP \times FN / (TP + FN)^3} = \sqrt{Sen(1 - Sen) / (TP + FN)}$$

灵敏度指标只与病例组有关，反映了诊断试验检出病例的能力。

35.1.3 特异度

实际未患病且被诊断为阴性的概率就是特异度(specificity, Spe), 即

$$\text{Spe} = P(T_- | D_-) = \text{TN} / (\text{FP} + \text{FN})$$

特异度的标准误

$$\begin{aligned} SE_{\text{spe}} &= \sqrt{\text{FP} \times \text{TN} / (\text{FP} + \text{TN})^3} \\ &= \sqrt{\text{Spe}(1 - \text{Spe}) / (\text{FP} + \text{TN})} \end{aligned}$$

特异度指标只与对照组有关，反映了诊断试验排除非病例的能力。

由灵敏度指标可以得到漏诊率 $\beta = 1 - \text{Sen} = \text{FN} / (\text{TP} + \text{FN})$, 漏诊率也就是假阴性率; 由特异度指标可导出误诊率 $\alpha = 1 - \text{Spe} = \text{FP} / (\text{FP} + \text{TN})$, 误诊率也就是假阳性率(false positive rate, FPR)。在本例漏诊率 $\beta = 1 - \text{Sen} = 1 - 0.8 = 0.2$; 误诊率 $\alpha = 1 - \text{Spe} = 1 - 0.95 = 0.05$, 即假阳性率 $\text{FPR} = 0.05$ 。灵敏度、特异度、漏诊率、误诊率之间的关系可用图 35-1 表示。此图中间的垂线与横轴的交点称为诊断界点, 它是定义诊断试验为阳性与阴性的临界点。

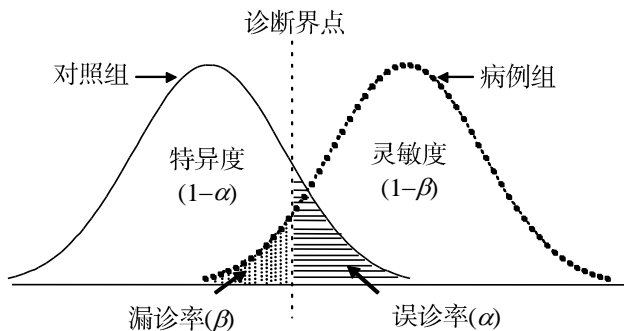


图 35-1 灵敏度、特异度、漏诊率及误诊率图示

灵敏度与特异度具有不受患病率影响的优点, 其取值范围均在(0, 1)之间, 其值越接近于 1, 说明其诊断试验的价值越好。

当比较两个诊断试验时, 单独使用灵敏度或特异度, 可能出现一个诊断试验的灵敏度高、特异度低, 而另一个诊断试验的灵敏度低、特异度高, 无法判断哪一个诊断试验更好。由此, 有人提出了将灵敏度和特异度结合的诊断试验评价指标, 如 Youden 指数、阳性似然比、阴性似然比等。

35.1.4 Youden 指数

真阳性率与假阳性率之差就是 Youden 指数(J):

$$J = \text{Sen} + \text{Spe} - 1 = \text{TPR} - \text{FPR}$$

其标准误为

$$SE_J = \sqrt{TP \times FN / (TP + FN)^3 + FP \times TN / (FP + TN)^3}$$

Youden 指数的取值范围在(-1, +1)之间, 其值越接近于+1, 诊断准确性越好。

35.1.5 阳性似然比

真阳性率与假阳性率之比, 即灵敏度与误诊率之比就是阳性似然比(positive likelihood ratio, LR_+)

$$\begin{aligned} LR_+ &= TPR / FPR \\ &= Sen / (1 - Spe) \end{aligned}$$

LR_+ 的取值范围为(0, $+\infty$), 其值越大, 检测方法证实疾病的能力越强。 LR_+ 的标准误涉及对数变换, 这里不予给出。以下几个指标的标准误计算也较复杂, 也不予给出。

35.1.6 阴性似然比

假阴性率与真阴性率之比, 即漏诊率与特异度之比为阴性似然比(negative likelihood ratio, LR_-)

$$\begin{aligned} LR_- &= (1 - TPR) / (1 - FPR) \\ &= (1 - Sen) / Spe \end{aligned}$$

LR_- 的取值范围为(0, $+\infty$), 其值越小, 检测方法排除疾病的能力越好。

35.1.7 预报值

在通常的情况下, 当要对某疾病作出诊断时, 并不知道金标准的结果, 只知道诊断试验是阳性或阴性。而临床医生更想知道的是: 当诊断试验阳性时, 受试者真正有病的概率有多大; 阴性时又有多大把握排除此病。这就需要引入阳性预报值(positive predictive value, PV_+)与阴性预报值的概念。

试验结果阳性时, 受试者实际为病例的概率就是阳性预报值, 这实际上就是本章开始时提到的贝叶斯公式的后验概率

$$\begin{aligned} PV_+ &= P(D_+ | T_+) \\ &= \frac{T(T_+ | D_+)P(D_+)}{P(T_+ | D_+)P(D_+) + P(T_+ | D_-)P(D_-)} \\ &= \frac{SenP_0}{SenP_0 + (1 - Spe)(1 - P_0)} \\ &= 1 / (1 + (1 - Spe)(1 - P_0) / SenP_0) \end{aligned}$$

式中 $P_0 = P(D_+)$ 表示先验概率, 在医院中为怀疑患有某病的概率, 而在总体人群中就是患病率; $P(D_-) = 1 - P(D_+) = 1 - P_0$; Sen 与 Spe 分别表示灵敏度与特异度。由上式可以看出,

当灵敏度与特异度为常数时，增加患病率，将降低 $(1-\text{Sen})(1-P_0)$ ，增加 $\text{Sen} \times P_0$ 的值，从而整个分母的值减少，阳性预报值增加。

35.1.8 阴性预报值

诊断试验结果阴性时，受试者实际为非病例的概率就是阴性预报值(negative predictive value, PV_-)。如果诊断试验金标准为 D_- 与 D_+ ，根据贝叶斯定理有

$$\begin{aligned} PV_- &= P(D_- | T_-) \\ &= \frac{T(T_- | D_-)P(D_-)}{P(T_- | D_-)P(D_-) + P(T_- | D_+)P(D_+)} \\ &= \frac{\text{Spe}(1 - P_0)}{\text{Spe}(1 - P_0) + (1 - \text{Sen})P_0} \\ &= 1 / \left(1 + \frac{(1 - \text{Sen})P_0}{\text{Spe}(1 - P_0)} \right) \end{aligned}$$

式中符号的意义与前面相同。当灵敏度与特异度为常数时，增加患病率将降低阴性预报值。

PV_+ 和 PV_- 的取值范围在(0, 1)之间；对于相同的患病率，其值越接近 1，检测方法的诊断价值越高。

在 DPS 中，各项诊断指标的计算，只需要将原始数据编辑成图 35-2 左上角阴影部分所示格式，然后在菜单方式下执行“专业统计”下面“诊断试验评价”里面的“诊断试验常用指标”功能，即可得到图 35-2 所示结果：一致百分率等于 0.8386，灵敏度等于 0.80，特异度等于 0.95，Youden 指数等于 0.75，阳性似然比等于 16，阴性似然比等于 0.2106。系统还显示了总体人群患病率在 0~0.25 时阳性预报值和阴性预报值的变化趋势(图 35-2 右边)。

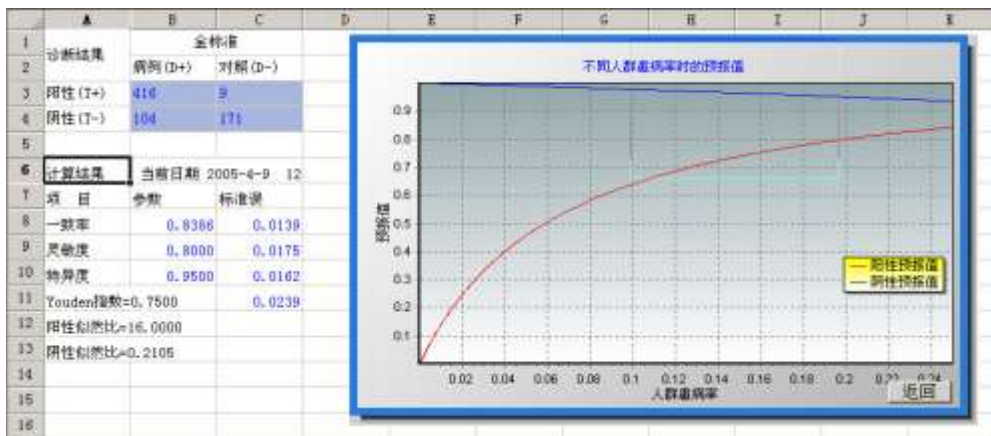


图 35-2 诊断指标计算数据编辑格式及其计算结果

35.2 有序分类资料 ROC 曲线

尽管前面所列的 Youden 指数、阳(阴)性似然比、阳(阴)性预报值等指标综合利用了真阳性率(TPR)与假阳性率(FPR)的信息,但这些指标都与诊断界点(或阈值)的选取有关。例如,同一项检测方法,采用不同的诊断界点就有不同的 TPR 与 FPR。为了更全面地评价检测方法的诊断价值,必须考虑各种可能的诊断界点。

接收者工作特征(receiver operating characteristic)曲线,简称 ROC 曲线,其分析始于 20 世纪 50 年代的统计决策理论,后来应用于雷达信号接收能力的评价。自从 20 世纪 80 年代起,该方法广泛应用于医学诊断试验性能的评价。通过改变诊断界点,可获得多对 TPR 与 FPR 值。以 FPR 为横坐标,TPR 为纵坐标,绘制 ROC 曲线,计算与比较 ROC 曲线下面积,以此反映诊断试验的价值。

35.2.1 实验资料 ROC 工作点计算及 ROC 曲线构建

ROC 分析资料可大致分为连续型资料与有序分类资料两种形式。连续型资料常见于某些定量检验。有序分类资料多见于医学影像诊断或心理学评价。这里先介绍有序分类资料的分析过程。

例如,我们有 109 份 CT 影像数据,其中有 51 份采用金标准确诊为异常,58 份确诊为正常。某放射科医生对这些 CT 影像的异常程度按 1, 2, 3, 4, 5 的顺序进行分类,结果见表 35-3。试计算所有可能的 TPR 和 FPR 值。

表 35-3 109 份 CT 影像分类资料

金标准	诊断分类					合计
	1	2	3	4	5	
异常	3	2	2	11	33	51
正常	33	6	6	11	2	58

对于这种 5 级分类资料,按级别从大到小排列,以前 4 个分类(即不考虑最小的分类)作为诊断界点,大于等于诊断界点者为阳性,小于该值者为阴性。这样,可整理出以下 4 个四格表:

诊断界点=5			诊断界点=4			诊断界点=3			诊断界点=2		
诊断结果	金标准		诊断结果	金标准		诊断结果	金标准		诊断结果	金标准	
	异常	正常		异常	正常		异常	正常		异常	正常
+	33	2	+	44	13	+	46	19	+	48	25
-	18	56	-	7	45	-	5	39	-	3	33

每个四格表对应的 ROC 曲线的工作点见表 35-4。

表 35-4 表 35-3 资料不同诊断界点的 FPR 与 TPR 值

	诊断界点(诊断分类)			
	5	4	3	2
FPR	0.0345	0.2241	0.3296	0.4310
TPR	0.6471	0.8627	0.9020	0.9412

注意，以上都假设诊断试验结果较大者被判为阳性，如果检测结果较小者被判为阳性，则应按检测结果从小到大排列，小于等于诊断界值者为阳性，大于诊断界值者为阴性。

如果以假阳性率 FPR，即 1-特异度为横轴，真阳性率 TPR(即灵敏度)为纵轴，横轴与纵轴长度相等，形成正方形，在图中将 ROC 曲线工作点标出，用直线连接各相邻两点构建未光滑的 ROC 曲线。ROC 曲线一定通过(0, 0)和(1, 1)两点，这两点分别对应于灵敏度为 0 而特异度为 1，和灵敏度为 1 而特异度为 0。表 35-3 资料对应的未光滑 ROC 曲线见图 35-3。

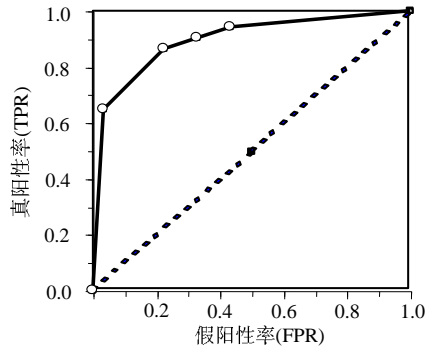


图 35-3 未光滑的 ROC 曲线

理论上，当诊断试验完全无价值时，有 $TPR=FPR$ ，是一条从原点到右上角的对角线，这条线称为机会线(chance line)；ROC 曲线一般位于机会线的上方，离机会线越远，说明诊断准确度越高；最好的诊断试验在图中表现为 ROC 曲线从原点垂直上升至左上角，然后水平到达右上角。

35.2.2 ROC 曲线下面积的计算原理

图 35-3 中的 ROC 曲线下面积(记为 A_z)可反映诊断试验的准确性大小。这一指标取值范围在 0.5~1，完全无价值的诊断 $A_z=0.5$ ，完全理想的诊断 $A_z=1$ 。一般认为 A_z 为 0.5~0.7 时，表示诊断准确性较低；为 0.7~0.9 时，表示诊断准确性为中等；为 0.9 以上时表示诊断准确性较高(Swets, 1988)。

A_z 的计算方法主要有梯形法(Wilcoxon 法)、Delong, Delong 和 Clarke-Pearson 非参数法等非参数方法，以及双正态模型参数法。非参数法计算较为简单，但结果精度较低；参数法计算比较复杂，需要采用计算机程序来完成，但精度较高。在 DPS 里面，我们给出了非参数梯形法和双正态分布模型参数方法的计算结果。

梯形法(Wilcoxon 法), 这类非参数方法计算相对简单, 容易理解。该方法计算过程如下:

假设对照组有 n_c 个观察值, 记为 $x_j(j=1, 2, \dots, n_c)$; 病例组有 n_a 个观察值, 记为 $X_i(i=1, 2, \dots, n_a)$ 。如果观察值大判归为病例, 根据 Wilcoxon Mann-Whitney 统计量, ROC 曲线下面积(A_z)就是病例组每个观察值大于对照组每个观察值的概率, 用公式表示为

$$A_z = \frac{1}{n_a n_c} \sum_{j=1}^{n_c} \sum_{i=1}^{n_a} \psi(X_i, X_j)$$

$$\psi(X_i, X_j) = \begin{cases} 1, & X_i > X_j \\ 0.5, & X_i = X_j \\ 0, & X_i < X_j \end{cases}$$

该公式为病例组的 n_a 个 X_i 与对照组的 n_c 个 x_j 比较, 如果前者大于后者, 则比较结果为 1, 相等时为 0.5, 否则为 0, 将 $n_a \times n_c$ 个比较结果相加取平均即得 A_z 。如果观察值小为异常, 则改变公式中的大于与小于符号即可。

A_z 的标准误的计算公式为

$$SE(A_z) = \sqrt{\frac{A_z(1-A_z) + (n_a - 1)(Q_1 - A_z^2) + (n_c - 1)(Q_2 - A_z^2)}{n_a n_c}}$$

其中 $SE(A_z)$ 为 A_z 的标准误, Q_1 是两个随机选择的病例组观察值比一个随机选择的对照组观察值都将有更大可能划归为病例的概率。 Q_2 是一个随机选择的病例组观察值比两个随机选择的对照组观察值将有更大可能划归为病例的概率。

得出的 ROC 曲线下面积是否与完全随机情况下获得的 $A_z=0.5$ 有统计学差异, 可近似采用标准正态离差 $Z=(A_z-0.5)/SE(A_z)$ 作检验。根据 $A_z \pm u_{\alpha} SE(A_z)$ 可计算 A_z 的 $100(1-\alpha)\%$ 可信区间。

双正态模型参数方法是假定 ROC 曲线为一函数。其做法是假定“有病”与“无病”患者有序试验结果 T_1 和 T_0 分别来自潜隐连续尺度随机变量 T_1^* 和 T_0^* , 且假定潜隐变量 T_1^* 和 T_0^* 的分布部分重叠(图 35-4)。

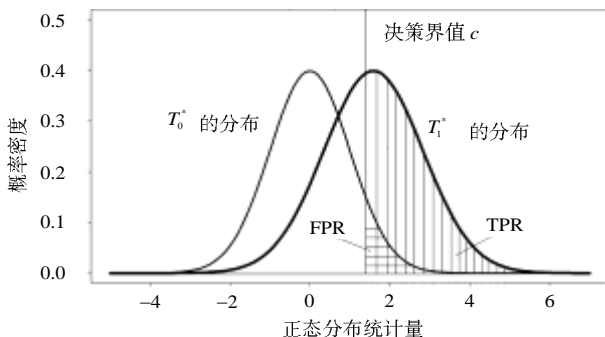


图 35-4 潜隐连续随机变量 T_1^* 和 T_0^* 分布曲线

为了更加明确, 假定 T_1^* 是“有病”者对诊断试验的基本潜隐反应, 其分布为正态分布 F_1 ; T_0^* 是“无病”者对诊断试验的基本潜隐反应, 其分布为正态分布 F_0 。当决策界值(或与阈值)为 c 时, c 与假阳性率(FPR)、灵敏度的关系是

$$\text{FPR}(c) = P(T_0^* > c | D=0) = 1 - F_0(c)$$

$$\text{TPR}(c) = P(T_1^* > c | D=1) = 1 - F_1(c)$$

这里有序尺度试验结果 T_i 为 K 个有序值中的一个值, $i=0, 1$, 假如有 $K-1$ 个未知决策界值 $\theta_0, \theta_1, \theta_2, \dots, \theta_{K-1}$, 则有: 如果 $T_i \leq \theta_0$, 那么 $T_i=1$; 如果 $\theta_{j-1} < T_i \leq \theta_j$, 那么 $T_i=j$, $j=1, 2, \dots, K-1$; 如果 $T_i > \theta_{K-1}$, 那么 $T_i=K$ 。

因这里假定 F_1 和 F_0 均呈正态分布, 故称量 T_1^* 和 T_0^* 服从双正态分布。这样, 通过适当转换, 有

$$T_1^* \sim N(\mu_1, \sigma_1^2), \quad T_0^* \sim N(\mu_0, \sigma_0^2)$$

式中, μ_1 和 μ_0 分别为“有病”和“无病”者潜隐试验反应的正态分布均数, σ_1^2 和 σ_0^2 为方差。基于这种概率分布, 试验中观察到某一有序尺度 j 的试验结果概率 P_{0j} 和 P_{1j} 分别为 $P_{0j} = P(T_0=j | D=0) = \Phi(c_j) - \Phi(c_{j-1})$ 和 $P_{1j} = P(T_1=j | D=1) = \Phi(bc_j - a) - \Phi(bc_{j-1} - a)$ 。其中 $c_j = (\theta_j - \mu_0) / \sigma_0$, $b = \sigma_0 / \sigma_1$, $a = (\mu_1 - \mu_0) / \sigma_1$; $c_0 = -\infty$, $c_K = +\infty$; Φ 为累积正态分布函数。其似然函数为

$$\begin{aligned} L &= \prod_{j=1}^K p_{0j}^{r_j} \prod_{jk=1}^K p_{1j}^{s_j} \\ &= \prod_{j=1}^K [\Phi(c_j) - \Phi(c_{j-1})]^{r_j} \prod_{j=1}^K [\Phi(bc_j - a) - \Phi(bc_{j-1} - a)]^{s_j} \end{aligned}$$

式中 s_j 和 r_j 分别是第 j 类“有病”和“无病”的个体数, 从这里介绍可以看出, 拟合 ROC 双正态分布模型中的参数 a , b , c_j , 可采用极大似然(maximum likelihood)法进行。

双正态分布模型中的参数 a , b 求解出来之后, ROC 曲线下的面积

$$A = \Phi\left(a / \sqrt{1+b^2}\right)$$

方差

$$\text{var}(A) = f^2 \text{var}(\hat{a}) + g^2 \text{var}(\hat{b}) + 2fg \text{cov}(\hat{a}, \hat{b})$$

式中

$$f = \frac{e^{-a^2/2(1+b^2)}}{\sqrt{2\pi(1+b^2)}}, \quad g = \frac{abe^{-a^2/2(1+b^2)}}{\sqrt{2\pi(1+b^2)^3}}$$

如果需要比较的两个诊断试验曲线下面积分别为 A_{z_1} 和 A_{z_2} , 对应的标准误分别为

SE₁ 和 SE₂，且面积之间的相关系数很小时，可利用公式

$$Z = \frac{|A_{z1} - A_{z2}|}{\sqrt{SE_1^2 + SE_2^2}}$$

来检验两 ROC 曲线下面积是否相等。

为检验拟合结果的合理性，Dorfman 和 Alf. Jr. 提出了拟合优度检验公式

$$\chi^2 = \sum_{i=0}^1 \sum_{k=1}^K \frac{[\hat{P}(T_i = k | D = i) - P(T_i = k | D = i)]^2}{\hat{P}(T_i = k | D = i)}$$

其中 $P(T=k|D=1)=s_k/n_1$, $P(T=k|D=0)=r_k/n_0$ 为试验结果观察值的概率, $\hat{P}(T_i = k | D = i)$ 是根据双正态模型计算得到的理论值。该统计量近似服从自由度为 $K-3$ 的卡方分布。

35.2.3 单个有序分类样本 ROC 曲线估计

单个有序分类样本 ROC 曲线估计，以乳腺 X 线照相数字工作站观察恶性病灶的准确度评价研究结果为例进行分析。该例中阅片者评阅了 58 例患者影像，按正常、良性、可能良性、恶性可疑和恶性 5 个级别。随后确认 13 例有病，45 例无病(图 35-5 上部阴影部分)。

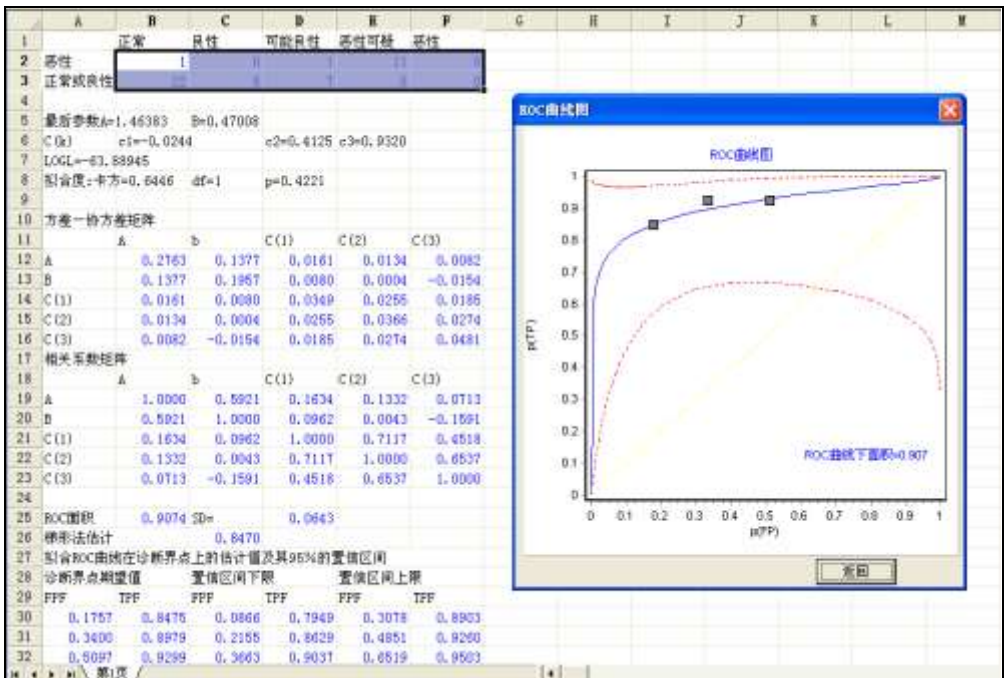


图 35-5 X 线照相数据 ROC 曲线分析结果

分析时，执行“专业统计”下面“诊断试验评价”中的“有序分类资料 ROC 曲线”功能，系统将显示图 35-5 右边的 ROC 曲线(该图在完成分析后以文件名 roc1.bmp 自动

保存在当前文件夹里面), 其中实线为理论值, 虚线为 95% 置信限, 点为观测的 FPR 和 TPR 值。

图 35-5 的左上部为重新编辑过的主要结果, 即 $\hat{a}=1.46383$, $\hat{b}=0.4701$; 3 个决策界点的估计值 $\hat{c}_1=0.0244$, $\hat{c}_2=0.4125$, $\hat{c}_3=0.9320$; 双正态模型的极大对数似然函数等于 -63.88945 。为检验拟合结果的合理性计算得到的卡方值为 0.6446, p 值为 0.4221, 无统计学意义, 即可认为观察值和理论值的拟合程度良好。

分析结果还给出了 ROC 曲线下面积估计值, 为 0.9074, 及其标准误 0.0643。并给出了非参数方法估计 ROC 曲线下面积估计值 0.8470。

35.2.4 多组有序分类样本 ROC 曲线估计及其比较

多组有序分类样本 ROC 曲线估计, 其方法和前面的一样。只是增加了比较的两个诊断试验曲线下面积的功能, 但这要求两组样本面积之间的相关系数很小。例如对两组资料进行 ROC 分析, 其数据编辑、选择格式如图 35-6 左上部阴影部分所示。分析时, 仍执行“专业统计”下面“诊断试验评价”中的“有序分类资料 ROC 曲线”功能, 系统将显示图 35-6 右边的 ROC 曲线(该图在完成分析后以文件名 roc2.bmp 自动保存在当前文件夹里面), 其中实线为各组样本的理论值, 点为观测的 FPR 和 TPR 值。

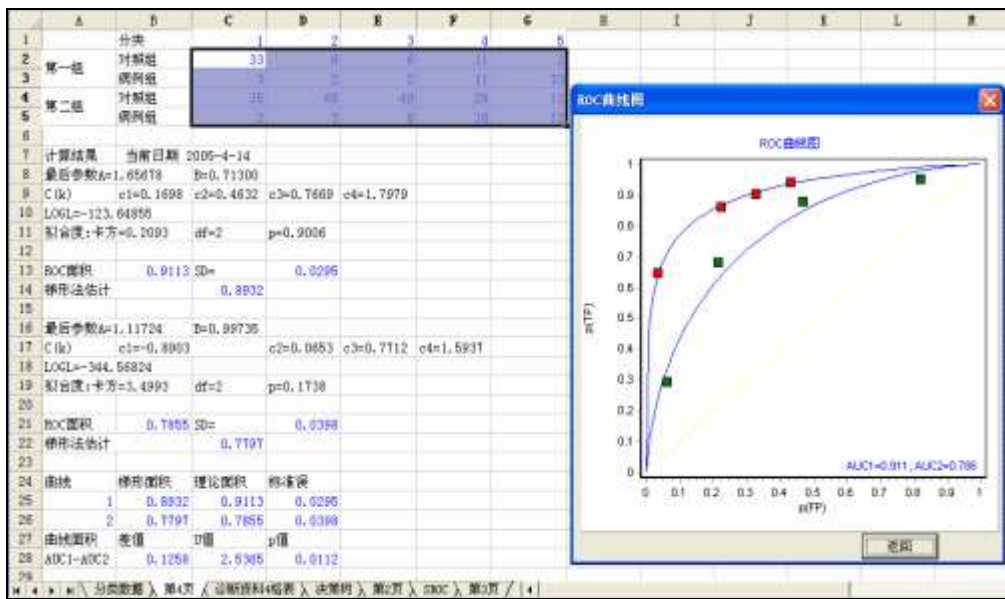


图 35-6 两组数据 ROC 曲线分析及其结果

图 35-6 的左上部为重新编辑过的主要结果, 即各组的参数 \hat{a} , \hat{b} , 决策界点的估计值 \hat{c}_j , 双正态模型的极大对数似然函数, 及为检验拟合结果的合理性计算得到的卡方值等。分析结果还给出了各组样本 ROC 曲线下面积估计值 \pm 标准误, 这里分别是 0.9113 ± 0.0296 和 0.7855 ± 0.0398 。

系统最后给出了两组资料 ROC 曲线下面积差异统计检验结果, 差值为 0.1258, 统计量 U 等于 2.5365, p 值为 0.0112。结论是差异显著。

35.3 定量数据 ROC 曲线分析

例如, 假设某诊断试验的病例组和对照组分别有 5 个和 4 个受试者, 其检测项目为定量指标, 检测结果, 病例组为 16.5, 13.5, 12.8, 11.2 和 5.0; 对照组为 8.5, 6.4, 4.6, 1.7。这时所有可能的 TPR 和 FPR 值的计算方法是将这 9 个数据从大到小排列, 以前 8 个数(不考虑最小值 1.7)分别作为诊断界点, 大于等于诊断界点者判为阳性, 小于该值者判为阴性。这样, 可整理成 8 个四格表。

诊断界点=16.5			诊断界点=13.5			诊断界点=12.8			诊断界点=11.2		
诊断 结果	金标准		诊断 结果	金标准		诊断 结果	金标准		诊断 结果	金标准	
	病例	对照		病例	对照		病例	对照		病例	对照
+	1	0	+	2	0	+	3	0	+	4	0
-	4	4	-	3	4	-	2	4	-	1	4

诊断界点=8.5			诊断界点=6.4			诊断界点=5.0			诊断界点=4.6		
诊断 结果	金标准		诊断 结果	金标准		诊断 结果	金标准		诊断 结果	金标准	
	病例	对照		病例	对照		病例	对照		病例	对照
+	4	1	+	4	2	+	5	2	+	5	3
-	1	3	-	1	2	-	0	2	-	0	1

每个四格表可计算一对(FPR, TPR), 称为 ROC 曲线工作点(表 35-5)。如果有多个检测结果相同(结), 则只保留一个值作为诊断界点。

表 35-5 不同诊断界点的 FPR 和 TPR 值

	诊断界值							
	16.5	13.5	12.8	11.2	8.5	6.4	5.0	4.6
FPR	0	0	0	0	1/4	2/4	2/4	3/4
TPR	1/5	2/5	3/5	4/5	4/5	4/5	5/5	5/5

和前面的有序分类资料一样, 如果以假阳性率 FPR(即(1-特异度))为横轴, 真阳性率 TPR(即灵敏度)为纵轴, 横轴与纵轴长度相等, 形成正方形, 在图中将 ROC 曲线工作点标出, 用直线连接各相邻两点构建未光滑的 ROC 曲线。ROC 曲线一定通过(0, 0)和(1, 1)两点, 这两点分别对应于灵敏度为 0 而特异度为 1, 和灵敏度为 1 而特异度为 0。得到的未光滑 ROC 曲线形式也如图 35-3。理论上, 当诊断试验完全无价值时, 有 TPR=FPR, 是一条从原点到右上角的对角线, 这条线称为机会线(chance line); ROC 曲线一般位于机会线的上方, 离机会线越远, 说明诊断准确度越高; 最好的诊断试验在图中表现为 ROC 曲线从原点垂直上升至左上角, 然后水平到达右上角。

和有序分类资料 ROC 曲线下面积估计一样, 定量资料 ROC 曲线下面积 A_z 的计算主

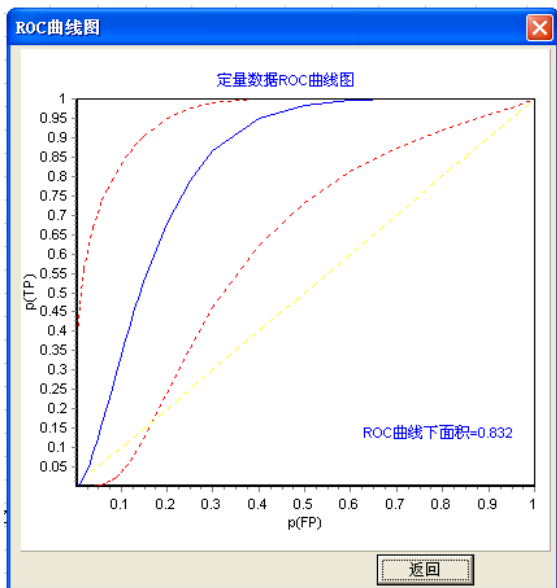
要方法也是梯形法(Wilcoxon 法)、DeLong, DeLong 和 Clarke-Pearson 非参数法等非参数方法, 以及双正态模型参数法。其计算原理也相同。在 DPS 里面, 我们给出了梯形法和双正态分布参数法的计算结果。

定量数据 ROC 曲线估计, 假设这里的真实阳性组和真实阴性组各检测了 20 例某生化指标。在 DPS 电子表格里, 将真实阳性组和真实阴性组 2 组结果按 2 行或 2 列方式输入, 然后用鼠标选中, 进行分析。这里真实阳性组和真实阴性组 2 组数据按 2 行输入(图 35-7)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																						
2		0.38	3.23	1.36	1.02	2.06	2.17	0.44	2.73	2.16	0.52	0.46	2.49	0.64	2.04	1.01	1.64	0.90	2.89	1.95	2.74	
3		0.86	0.88	1.07	1.17	0.69	0.40	0.49	1.11	0.61	0.40	0.66	0.41	0.17	0.64	0.40	0.85	0.80	0.68	0.45	0.15	
4																						

图 35-7 定量数据 ROC 曲线分析

分析时, 执行“专业统计”下面“诊断试验评价”中的“定量资料 ROC 曲线”功能, 系统将显示如下面图形所示的 ROC 曲线(该图在完成分析后以文件名 qroc.bmp 自动保存在当前文件夹里面), 其中实线为理论值, 虚线为 95%置信限。



返回编辑后, 得到系统给出的计算结果。该例主要结果是, 双正态模型参数 $\hat{a}=2.1609$, $\hat{b}=2.0151$; ROC 曲线下面积估计值为 0.8316, 及其标准误 0.0670。并给出了非参数方法估计 ROC 曲线下面积估计值 0.8088, 及其标准误 0.0697。

35.4 汇总多个样本的 SROC 曲线分析

为了评价一个诊断试验的准确度，首先需要知道受试者的真实分类情况，即哪些属于对照组，哪些属于病例组，常采用金标准划分受试者属于病例组还是对照组。医学研究中常见的金标准有：活组织检查、尸体解剖、手术探查和跟踪随访结果等。尽管金标准不需要十全十美，但是它们应比评价的诊断试验更可靠，且与评价的诊断试验无关。对于按金标准确定的二项分类总体，如病例与对照，采用诊断试验检测的结果分别写成阳性与阴性，资料可列成四格表形式。表中有四个可能结果，其中两个是正确的，即病例被诊断为阳性(真阳性, TP)和对照被诊断为阴性(真阴性, TN)；两个是错误的，即对照被诊断为阳性(假阳性, FP)和病例被诊断为阴性(假阴性, FN)。

根据 35.1 节介绍的公式，可计算真阳性率(TPR)、假阳性率(FPR)，对 TPR 和 FPR 采用 logit 变换后，

$$\text{logit}(\text{TPR})=\ln[\text{TPR}/(1-\text{TPR})]$$

$$\text{logit}(\text{FPR})=\ln[\text{FPR}/(1-\text{FPR})]$$

并令 $D=\text{logit}(\text{TPR})-\text{logit}(\text{FPR})$ ， $S=\text{logit}(\text{TPR})+\text{logit}(\text{FPR})$ ，以 D 为因变量， S 为自变量，建立 SROC 线性回归模型

$$\hat{D}=A+B\times S$$

式中， D 为对数优势比，表示病例组中阳性结果的优势(真阳性率与假阴性率之比)与对照组中阳性结果的优势(假阳性率与真阴性率之比)的比值。 D 值的大小反映诊断试验的“判别”能力。 S 与诊断界点的选择有关。当灵敏度和特异度相等，即真阳性率与真阴性率相等时， $S=0$ ；在此基础上增加灵敏度或降低特异度有 $S>0$ ，为正值；降低灵敏度或增加特异度则 $S<0$ ，为负值； S 可视为诊断的阈值。 A 为线性回归模型的截距项，是 $S=0$ 即灵敏度等于特异度时的对数优势比。 B 为回归系数，反映对数优势比(D)依赖于阈值(S)的程度，如果回归系数为 0 或与 0 的差异无统计学意义，则可采用截距项 A 作为多个原研究的公共对数优势比。

SROC 线性回归模型参数，可采用一般最小二乘法、加权最小二乘法、稳健法求解。

(1) 一般最小二乘(OLS)法。一般统计学教材均有这种方法，其基本思想是使实测值与估计值之差(残差)的平方和最小，求解参数 A 与 B 。该方法的缺点是没有充分考虑原研究样本含量的大小。

(2) 加权最小二乘(WLS)法。为了利用原研究样本含量的信息，可采用对数优势比 D 的方差的倒数为权重，采用加权最小二乘法进行分析。加权最小二乘法的基本思想是通过加权的残差平方和最小来求解参数。如果用 TP, FP, FN, TN 分别表示真阳性数、假阳性数、假阴性数、真阴性数，以方差倒数为权重 W ，则

$$W = \frac{1}{\text{var}(D)}$$

$$= \frac{1}{1/TP + 1/FP + 1/FN + 1/TN}$$

如某一原始研究的任一个数据为 0，为了避免分母为 0 的情况，可将所涉及研究的每个观察值加 0.5。

加权法估计也不够精确，某些情况下也可导致偏倚，由此 Moses, Shapiro 和 Littenberg 于 1993 年提出了 SROC 分析的稳健法。

(3) 稳健法。其参数拟合如图 35-8 所示

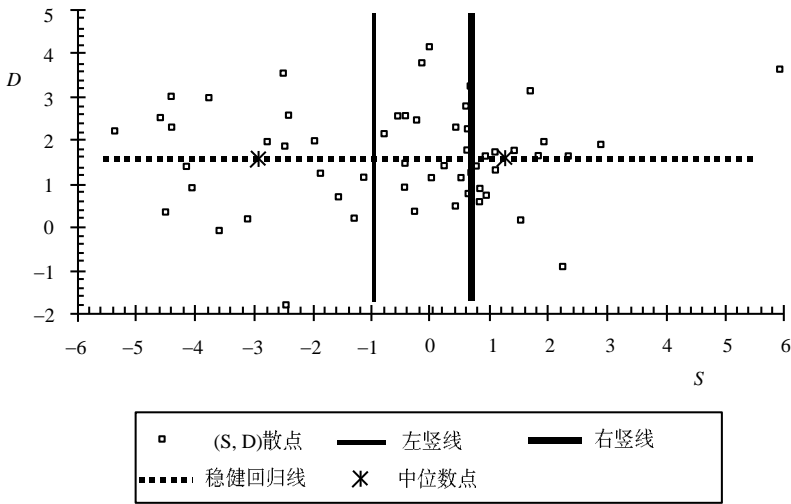


图 35-8 稳健法求解 SROC 曲线参数的示意图

图中，以 S 为横轴， D 为纵轴，将各研究的 (S, D) 散点描在坐标系中。用两条竖线将 (S, D) 散点分成三组，并使两条竖线之间及左右侧的散点各占 $1/3$ 左右。总例数除以 3 后，其得数四舍五入取整可得左右侧的散点个数。比如，有 10 个散点时两条竖线之间占 4 个点，左右侧各占 3 个点。分别找到左右侧散点的 S 与 D 的中位数，并在图中作上标记，将标记的两点连成一条线，该线的斜率就是回归系数 B 。观察这条线的上下散点是否相等，如果不相等，应将该直线上下平移，使上下散点数基本相等。如果用 (S_1, D_1) 和 (S_2, D_2) 分别表示直线上距离相隔较远的两个点(如左右侧 S 和 D 的中位数所在的点)，采用下列公式可计算回归参数 A 与 B ：

$$A = \frac{D_1 S_2 - D_2 S_1}{S_2 - S_1}, \quad B = \frac{D_2 - D_1}{S_2 - S_1}$$

在采用上述几种方法求得回归参数 A, B 后，根据下式可建立 SROC 曲线的回归方

程为

$$\text{TPR} = \frac{1}{1 + e^{-A/(1-B)} \left(\frac{1 - \text{FPR}}{\text{FPR}} \right)^{(1+B)/(1-B)}}$$

对于一般 ROC 分析, 采用 ROC 曲线下面积表示诊断试验的准确性; 对于 SROC 分析, 比较简单的做法是采用 TPR^* 表示诊断试验的准确性。为了不与一般的 TPR 混淆, SROC 曲线的诊断试验准确性指标为

$$\text{TPR}^* = (1 + e^{-A/2})^{-1}$$

它对应的标准误可通过

$$\text{SE}(\text{TPR}^*) = \frac{\text{SE}(\hat{A})}{8[\cosh(A/4)]^2}$$

计算, 式中 $\text{SE}(\hat{A})$ 为直线回归方程的截距项 A 对应的标准误, \cosh 为双曲余弦函数。如果需要比较两组独立诊断试验的准确性, 当研究数不太少(如 >10)时, 可采用 Z 统计量

$$Z = \frac{\text{TPR}_1^* - \text{TPR}_2^*}{\sqrt{\text{SE}^2(\text{TPR}_1^*) + \text{SE}^2(\text{TPR}_2^*)}}$$

Z 为正态离差值, TPR_1^* , TPR_2^* 为比较的 SROC 曲线诊断准确性指标, $\text{SE}(\text{TPR}_1^*)$ 与 $\text{SE}(\text{TPR}_2^*)$ 分别为 TPR_1^* 与 TPR_2^* 对应的标准误。

更精确的做法是采用数值积分的方法计算 ROC 曲线下的面积 AUC:

$$\text{AUC} = \int_0^1 \frac{1}{1 + e^{-A/(1-B)} \left(\frac{1-x}{x} \right)^{(1+B)/(1-B)}} dx$$

ROC 曲线下的面积 AUC 的标准误, 可根据参数 \hat{a} 和 \hat{b} 的样本方差, 采用 delta 方法进行估计

$$\text{var}(\widehat{\text{AUC}}) = \left(\frac{\partial \text{AUC}}{\partial a} \right)^2 \text{var}(\hat{a}) + \left(\frac{\partial \text{AUC}}{\partial b} \right)^2 \text{var}(\hat{b}) + 2 \left(\frac{\partial \text{AUC}}{\partial a} \right) \left(\frac{\partial \text{AUC}}{\partial b} \right) \text{cov}(\hat{a}, \hat{b})$$

实例分析。Pap 试验是一种收集、制备和检验子宫颈脱落细胞的试验, 因该试验快速、价廉、无创伤, 所以医生常用它作为早期宫颈癌筛选, 目前也有人用它作为女性患者随访前的检查。因该试验受医生宫颈癌知识、宫颈癌发病率和检验时细胞数量多少限制, 报道的该诊断试验的准确性有较大差异, 灵敏度和特异度变化范围分别为 11%~99% 和 14%~97%。下面借用 Fahey, Irwig 和 Macaskill 报道的 59 份研究资料说明 SROC 分

析方法。

在 Fahey, Irwig 和 Macaskill 的资料中没有给出真阳性数、假阳性数、假阴性数和真阴性数的值,但列出了病例数、非病例数即灵敏度、(1-特异度)等值,因为方法的需要,我们根据他们所列数据计算了 TP, FP, FN, TN。文献第 13、24 号和 42 号研究中有 0,为了避免计算方差倒数时出现分母为 0 的情况, DPS 会对这 3 个研究的各项加上 0.5 作矫正。

应用 DPS 进行 SROC 曲线分析,数据格式是每个案例一行,每行按真阳性数 TP、假阳性数 FP、假阴性数 FN、真阴性数 TN 顺序输入。然后用鼠标选中(图 35-9)。

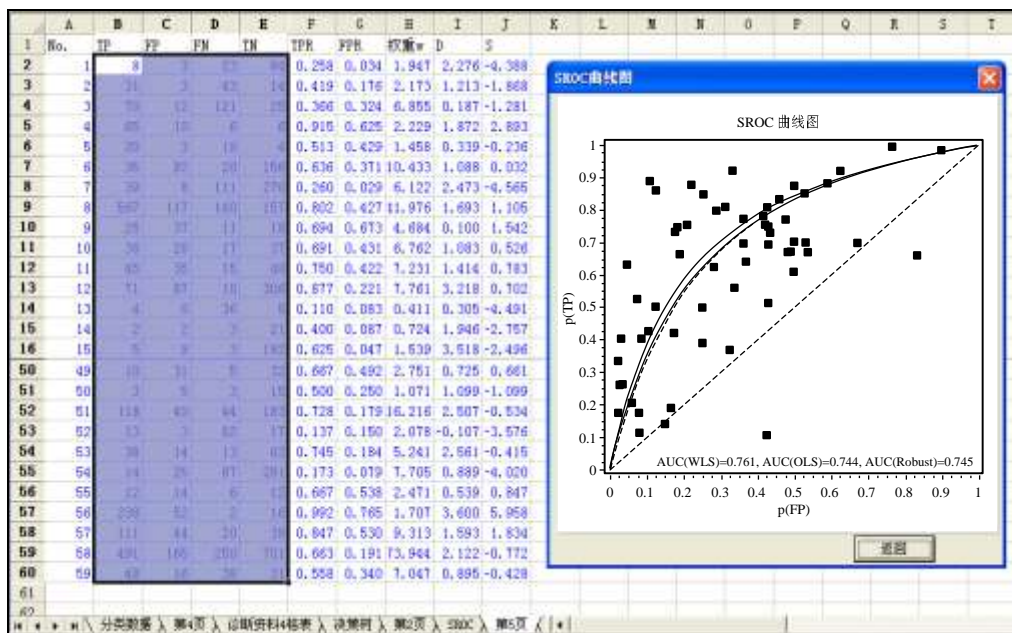


图 35-9 SROC 曲线分析数据编辑与分析示意图

在 DPS 菜单方式下执行“专业统计”下面“诊断试验评价”中的“综合接收者工作特征 SROC 曲线”功能,系统将显示图 35-9 右边的图形分析结果(该图在完成分析后以文件名 sroc.bmp 自动保存在当前文件夹里面),其中实线为理论值,点为观察值。数值分析的主要结果如下:

加权回归估计:						
项目	系数	标准误	t 值	概率	95% 置信区间	
A	1.7198	0.0998	17.2274	0.0001	1.5200	1.9196
B	-0.0151	0.0700	0.2154	0.8303	-0.1553	0.1251
TPR*=0.7026 Se(TPR*)=0.0104			优势比=5.5833		剩余标准差=2.4297	
SROC 曲线下的面积=0.7611			标准误= 0.0124			
非加权回归估计:						
项目	系数	标准误	t 值	概率	95% 置信区间	

A	1.5900	0.1511	10.5225	0.0001	1.2875	1.8924
B	0.0201	0.0630	0.3189	0.7509	-0.1060	0.1461
TPR*=0.6889	Se(TPR*)=0.0162		优势比=4.9037		剩余标准差=1.1444	
SROC 曲线下的面积=0.7445			标准误=0.0196			
两种方法面积差值=0.0166		U 统计量=0.71284		p=0.47596		
稳健估计:						
项目	系数					
A	1.5914					
B	0.0011					
TPR*=0.6890	优势比=4.9104					
SROC 曲线下的面积=0.74470						

以 D 为因变量, S 为自变量, 并利用所计算的权重拟合加权最小二乘直线回归模型, 得诊断试验阈值(S)引起的对数优势比(D)变异的剩余标准差为 2.430。截距 $A=1.720$, 其标准误 $SE(A)=0.100$, t 检验得 $t=17.227$, $p \leq 0.001$, 用 $A \pm_{0.05, 58} SE(A)$ 计算得 A 的 95% 置信区间为 1.520~1.920; 结果表明: 在 0.05 检验水准, A 与 0 间的差异有统计学意义。回归系数 $B=-0.015$, 其标准误 $SE(B)=0.070$, t 检验得 $t=0.215$, $p \leq 0.830$; 结果表明: 在 0.05 检验水准下, 参数 B 与 0 间的差异无统计学意义。从试验分析结果的优势比可以看出

$$\exp(A)=\exp(1.720)=5.585$$

可以认为异常组阳性优势明显高于正常组, 根据数值积分计算得到 ROC 曲线下面积为 0.7611, 其标准误 $SE(AUC)=0.0124$ 。

一般(非加权)最小二乘估计结果得诊断试验阈值(S)引起的对数优势比(D)变异的剩余标准差为 1.144。拟合模型得截距 $A=1.590$, 其标准误 $SE(A)=0.151$, t 检验得 $t=10.522$, $p \leq 0.001$, 用 $A \pm_{0.05, 58} SE(A)$ 计算得 A 的 95% 置信区间为 1.288~1.892, 结果表明: 在 0.05 检验水准下, A 与 0 间的差异有统计学意义。回归系数 $B=0.020$, 其标准误 $SE(B)=0.063$, t 检验得 $t=0.319$, $p \leq 0.751$; 结果表明: 在 0.05 检验水准下, 参数 B 与 0 间的差异无统计学意义。从试验分析结果的优势比可以看出

$$\exp(A) = \exp(1.590) = 4.904$$

可以认为异常组阳性优势明显高于正常组, 根据数值积分计算得到 ROC 曲线下面积为 0.7455, 其标准误 $SE(AUC)=0.0196$ 。

为了获得稳健法的参数值, 以 S 为横轴, D 为纵轴, 将 59 个研究的 S, D 点描在坐标中。用两条竖线将 (S, D) 散点分成三组, 其中左右两侧各占 20 个点(59, 3)四舍五入取整得 20, 对左右侧散点的 S 与 D 分别取中位数有: 左侧(S_1, D_1)= $(-2.916, 1.588)$, 右侧(S_2, D_2)= $(1.265, 1.593)$ 。计算得到回归参数 A 与 B 分别为 $A=1.5914, B=0.0011$, 由此可建立直线回归方程。

为了使直线的上下散点数相等, 需要固定斜率 0.0011 不变, 改变截距项, 使直线上下平移。这实际上相当于将各散点代入方程, 改变截距项 A 值, 所得估计值 \hat{D} 与实际值 D 之差(即残差)的正负号相等。经多次改变截距项 A 值, 检查直线上下的观察点, 使这

些点分布在直线上下基本相等, 此时得 $A=1.5914$, 优势比为

$$\exp(A)=\exp(1.5914)=4.9106$$

根据数值积分计算得到 ROC 曲线下面积为 0.7447。根据以上三种方法得到的参数 A , B , 可分别建立加权法、非加权法和稳健法的 SROC 曲线回归方程

$$\begin{aligned} \text{TPR}_{\text{加权}} &= \left[1 + e^{-1.694 \left(\frac{1-\text{FPR}}{\text{FPR}} \right)^{0.970}} \right]^{-1} \\ \text{TPR}_{\text{非加权}} &= \left[1 + e^{-1.623 \left(\frac{1-\text{FPR}}{\text{FPR}} \right)^{1.041}} \right]^{-1} \\ \text{TPR}_{\text{稳健}} &= \left[1 + e^{-1.593 \left(\frac{1-\text{FPR}}{\text{FPR}} \right)^{1.002}} \right]^{-1} \end{aligned}$$

从 SROC 曲线可以看出, 该资料加权法、非加权法、稳健法 SROC 曲线和 59 个研究工作点(1-特异度, 灵敏度), 加权法获得的曲线下面积较大, 非加权法与稳健法获得的曲线下面积相似, 几乎重叠, 这与数值结果中的 AUC 大小一致。

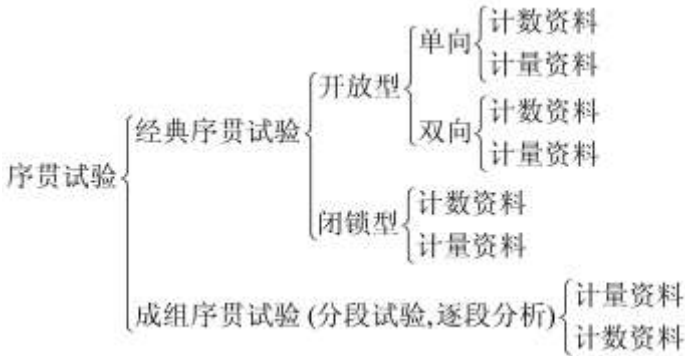
如果忽略加权法与非加权法的相关性, 并且假定两法所得 AUC 近似服从正态分布, 可粗略采用检验它们所得 AUC 是否有差异, 经检验得 $U=0.7128$, 双侧 $p=0.4760$, 结果表明加权法与非加权法计算所得 AUC 的差异无统计学意义。

参 考 文 献

- 方积乾, 陆盈. 2002. 现代医学统计学[M]. 北京: 人民卫生出版社.
- 余松林. 2002. 医学统计学[M]. 北京: 人民卫生出版社.
- Metz C E. 1989. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol*[J], 24: 234-245.
- Metz C E, Herman B A, Shen J H. 1998. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*[J], 17: 1033-1053.
- Rosner B. 2004. 生物统计学基础[M]. 孙尚拱译. 北京: 科学出版社.
- Zhou X H, Obuchowski N A, McClish D K. 2002. *Statistical methods in diagnostic medicine*[M]. John Wiley & Sons. Inc..

第 36 章 序贯试验分析

这里仅介绍经典序贯试验。经典序贯试验，根据观察指标的性质、是否预先限制最大样本例数、实验样本逐一分析还是阶段性逐段分析，一般可分为如下几类：



经典序贯试验是基于 Wald 的序贯概率比理论，由 H_1 和 H_0 的备选概率函数 $f(x_i, \theta_1)$ 与无效概率函数 $f(x_i, \theta_0)$ 之比构成对数似然比率

$$\lambda = \ln \prod \frac{f(x_i, \theta_0)}{f(x_i, \theta_1)}$$

来确定拒绝或接受 H_0 的界限线。对数似然比与我们规定的假阳性概率 α 和假阴性概率 β 有关：接受 H_0 的界限取决于对数似然比小于或等于 II 型错误 β 与置信度 $1-\alpha$ 概率比的对数，拒绝 H_0 的界限取决于对数似然比大于或等于检验效能 $1-\beta$ 与 I 型错误 α 概率比的对数，即当 $\ln \frac{\beta}{1-\alpha} < \lambda < \ln \frac{1-\beta}{\alpha}$ 时，不能作出结论，继续添加新的试验样本；而当 $\lambda \leq \ln \frac{\beta}{1-\alpha}$ 时，接受 H_0 ，停止试验；或者当 $\lambda \geq \ln \frac{1-\beta}{\alpha}$ 时，拒绝 H_0 ，接受 H_1 ，停止试验。

开放型还是闭锁型根据研究性质与预试情况而定。预试验提示差异明显或差异很小者，宜选用开放型设计，一般说来，开放型所需样本数相对较少，比闭锁型序贯试验能较早地得到结论。但在试验结果方向性非常模糊时，一定条件下，为避免迟迟不能得出结论，则宜选用闭锁型设计。这样能保证不超过一定试验样本例数得到明确的结论。

单向还是双向试验：仅回答有效或无效，或者 $A \geq B$ 、 $A \leq B$ 中的一种时，用单向试验；如 $A \geq B$ 和 $A \leq B$ 两种可能都存在，比较两者谁优谁劣时，常用双向试验。

试验结果属于质反应的为计数资料，属于量反应的结果为计量资料。但这还与我们的思路有关。例如疗效的判定，如以存活与死亡为指标(计数)，则为质反应；若以存活时间为指标(计量)，则为量反应。

如果从试验的不同方式分类, 序贯试验又可分为自身前后对照、异体配对比较、成组比较。经典序贯试验常采用自身对照或异体配对的试验。自身前后对照用于初步筛选某种单一处理有无作用, 或药物作用时间短暂、无后遗症影响, 用药后病情回到原水平的药物评价。异体配对比较的序贯试验用于比较两种处理间试验指标的差别。

序贯试验所用样本数明显少于其他设计, 但必须在严格保证样本的代表性与两组间可比性前提下进行。经典序贯设计常用于: ① 试验药与对照药的疗效相差较大, 但病例稀少且临床观察时间较短; ② 怀疑试验药有较高的不良反应发生率, 采用成组序贯设计可以较早终止试验。

经典序贯试验样本例数的估计方法。序贯试验界限线一半可用一元一次方程表达, 开放型质反应单向上界线 U 的方程为 $Y=a+bn$, 若试验的有效率为 π_1 , 触及上界线 U 时必然 $Y=\pi_1 n$ 。由 $a+bn=\pi_1 n$ 可求出样本例数 $n=a/(\pi_1-b)$ 。其他序贯设计估计 n 的基本方法也类似。

36.1 开放型序贯试验

36.1.1 开放型质反应单向序贯试验

第一步, 根据专业要求规定试验接受和拒绝水平: 例如规定接受水平为 $\pi_1 = 80\%$, 拒绝水平 $\pi_0 = 30\%$ 。当处理效果 $\geq 80\%$ 时, 拒绝 H_0 , 接受 H_1 ; 当有效率 $\leq 30\%$ 时, 结论为没有效果。

第二步, 确定假阳性率 α 与假阴性率 β 水准: 一般选用 α 与 β 均为 0.05, 必要时 α 与 β 可取不同水平。预计可信度与把握度很高时, α 与 β 可均用 0.01; 反之亦然。

第三步, 根据公式计算序贯试验界限线。用 y 代表反应指标。对给定的 π_1 , π_2 , α , β 值计算开放型单向质反应试验有效和无效的两界限线直线方程:

上界(有效) U 线: $y = a+bn$;

下界(无效) L 线: $y = -a+bn$;

上式中 a , b 称为边界系数:

$$a = \ln[(1-\beta)/\alpha] / \ln r, \quad b = \ln[(1-\pi_0)/(1-\pi_1)] / \ln r, \quad r = \pi_1(1-\pi_0) / [\pi_0(1-\pi_1)]$$

例如, 为了解冠心苏合丸治疗冠心病的作用, 用冠状窦流量为观察指标, 选用犬心肌梗死模型自身比较, 进行序贯试验。

(1) 规定观察指标的界限: 本例目的是检验冠心苏合丸是否能有效改善心肌梗死犬的冠状窦流量。根据预试验结果, 规定每个受试对象用药后冠状静脉窦流量增加值 $\geq 1.8\text{mL}/\text{min}$ 为有效; 流量增加值 $< 1.8\text{mL}/\text{min}$ 为无效。属于质反应试验。

(2) 确定试验类型: 根据临床经验与动物试验初步预试, 提示冠心苏合丸对改善冠状窦循环的作用是较为肯定的, 故采用开放型单向序贯试验。

(3) 规定试验的接受和拒绝水平: 规定接受水平为 $\pi_1 = 80\%$, 拒绝水平 $\pi_0 = 30\%$ 。即当冠心苏合丸改善冠状窦循环的有效率 $\geq 80\%$ 时, 结论为冠心苏合丸可改善急性心肌梗

死犬的冠状窦流量；当有效率 $\leq 30\%$ 时，结论为冠心苏合丸不能改善急性心肌梗死犬的冠状窦流量。

(4) 确定假阳性率 α 与假阴性率 β 水准：本例中取 $\alpha = \beta = 0.05$ 。

(5) 计算本例界限线： $r = 0.8(1-0.3)/[0.3(1-0.8)] = 9.33$ ， $a = \ln[0.95/0.05]/\ln 9.33 = 1.32$ ， $b = \ln[(1-0.3)/(1-0.8)]/\ln 9.33 = 0.56$ 。将边界系数代入界限线方程，得上界 U 线 $y = 1.32 + 0.56n$ ，下界 L 线为 $y = -1.32 + 0.56n$

在 DPS 中，我们只需将实验结果用 1(有效)或 0(无效)表示，并按图 36-1 中阴影部分，编辑、选中数据→专业统计→序贯检验及成组序贯分析→开放型→单向→计数资料，出现如图 36-1 右边菜单，根据专业要求填写有关参数，点击“确定”，即可得到如图 36-2 的序贯检验图。

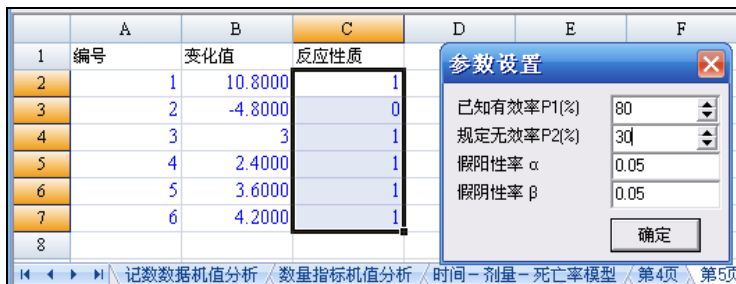


图 36-1 单向计数资料开放性序贯检验数据格式及用户界面图

根据图 36-2 中的红色实验线触及不同界线的情况可作出推论：不同序贯试验设计，其界线的意义有不同。在开放型单向试验中，凡触及上界 U 线，表示接受试验措施；若触及下界 L 线，表示拒绝试验措施，若是 $\alpha = \beta = 0.05$ ，则 $p < 0.05$ 。从图 36-2 可知，本例用冠心苏合丸后，急性心肌梗死组 6 条狗实验反应曲线已触及上界线 U，可认为冠心苏合丸可增加急性心肌梗死狗的冠状窦流量 ($p < 0.05$)。

单向计数资料开放性序贯检验图： $\alpha = 0.05$ ， $\beta = 0.05$ ，上界 $p_1 = 0.8$ ，下界 $p_2 = 0.3$

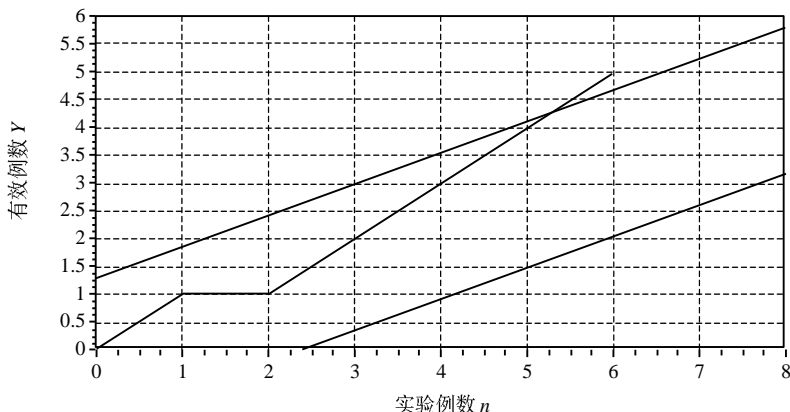


图 36-2 单向计数资料开放性序贯检验图

退出图表状态，系统会给出直线方程的数字结果如下：

单向计数资料开放性序贯检验图: $\alpha=0.05, \beta=0.05$, 上界 $p_1=0.8$, 下界 $p_2=0.3$
 直线方程:
 上界 U: $Y=1.3183+0.5609N$
 下界 L: $Y=-1.3183+0.5609N$

36.1.2 开放型数量资料单向序贯试验

开放型单向计量资料序贯检验和质反应单向序贯试验基本相同, 不同之处是需通过预试验计算或采用近似估计效应量的总体标准差 σ ; 效应量 y 有效时的指标 θ_1 及反应量无效时的指标 θ_2 。检验界线的参数分别是

$$a_1 = \frac{\sigma^2}{\theta_1 - \theta_2} \ln \frac{1 - \beta}{\alpha}, \quad a_2 = \frac{\sigma^2}{\theta_1 - \theta_2} \ln \frac{\beta}{1 - \alpha}, \quad b = \frac{\theta_1 + \theta_2}{2}$$

例如用肾型高血压狗试验胍乙啶降压效果, 规定服药后血压下降 $\geq 2.7\text{kPa}$ 以上为有效, 下降 ≥ 0.7 为无效, 估计总体标准差 $\sigma=2.8$, 拟定 $\alpha=0.1, \beta=0.05$, 这时可在 DPS 电子表格中, 逐一输入试验结果。并按图 36-3 中阴影部分, 编辑、选中数据 \rightarrow 专业统计 \rightarrow 序贯检验及成组序贯分析 \rightarrow 开放型 \rightarrow 单向 \rightarrow 计量资料, 出现如图 36-3 右边菜单, 根据专业要求填写有关参数, 点击“确定”, 即可得到如图 36-4 的序贯检验图。



图 36-3 单向定量开放性序贯检验数据格式及用户界面

单向计量资料开放性序贯检验图: $\alpha=0.1, \beta=0.05$, 上界 $\theta_1=2.7$, 下界 $\theta_2=0.7$, 差数方差 = 2.8

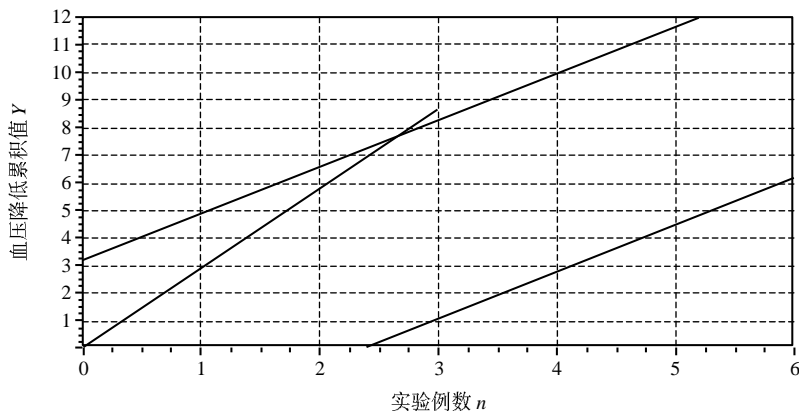


图 36-4 单向定量开放性序贯检验图

根据图 36-4 中的红色实验线触及不同界线的情况可作出推论：不同序贯试验设计，其界线的意义有不相同。在开放型单向试验中，凡触及上界 U 线，表示接受试验措施；若触及下界 L 线，表示拒绝试验措施，该例结论是胍乙啶对肾型高血压狗降压有效。

退出图表状态，系统会给出直线方程的数字结果如下：

单向计量资料开放性序贯检验图： $\alpha=0.1, \beta=0.05, \theta_1=2.7, \theta_2=0.7$
差数方差=2.8
直线方程：
上界 U: $Y=3.1518+1.7000N$
下界 L: $Y=-4.0465+1.7000N$

36.1.3 开放型质反应双向序贯试验

开放型质反应双向序贯试验方法基本同单向序贯试验，不同之处主要是试验标准和计数方式。一般记 $\theta = \text{SF 数} / (\text{SF 数} + \text{FS 数})$ ，由试验者的专业知识与预试情况决定 θ 达到多大的标准值时，认为试验措施优于对照；若 SF 数 > FS 数，则 $\theta > 0.5$ ；反之，则 $\theta < 0.5$ 。若 SF 数 = FS 数，则 $\theta = 0.5$ 。

I 型和 II 型错误的概率，在开放型双向序贯试验的假阳性率有 2 个，故 I 型错误概率是 2α 。

双向序贯试验，纵轴为 $Y = \text{SF 数} - \text{FS 数}$ 。根据试验标准规定，不仅有上界 U 线(试验措施优于对照)与下界 L 线(对照优于试验措施)，而且有两条中界线 M 和 M'(表示试验措施与对照无差别)。计算公式为

$$\begin{aligned} \text{U: } y &= a_1 + bn, & \text{L: } y &= -a_1 - bn \\ \text{M: } y &= -a_2 + bn, & \text{M': } y &= a_2 - bn \end{aligned}$$

式中

$$a_1 = \frac{2 \ln \frac{1-\beta}{\alpha}}{\ln \frac{\theta_1}{1-\theta_1}}, \quad a_2 = \frac{2 \ln \frac{1-\alpha}{\beta}}{\ln \frac{\theta_1}{1-\theta_1}}, \quad b = -\ln \frac{4\theta_1(1-\theta_1)}{\ln \frac{\theta_1}{1-\theta_1}}$$

例如，比较新旧 2 个镇痛复方对风湿性关节炎患者镇痛效果的序贯试验。本序贯试验采用一部分病例先服新复方试验措施，后服原复方对照；另一部分病例先服原复方，后服新复方。每药连续服用 2 天，随机决定各个患者先服何方。两药间隔期为 3 天。由受试者判断两复方中何者镇痛作用较好，记录为：病人判断新复方较原复方好时，记为 SF；病人判断原复方较好时，记为 FS；当病人不能区分两者优劣时不纳入记录。试验标准取 θ 为 0.8，规定计算得 $\theta \geq 0.8$ 时，结论为新复方较原复方好。规定 I 型和 II 型错误的概率 $2\alpha = \beta = 0.05$ 。

分析时在 DPS 电子表格中，按对子序号记录依次逐一输入试验结果。并按图 36-5 中阴影部分，编辑、选中数据 → 专业统计 → 序贯检验及成组序贯分析 → 开放型 → 双向 →

计数资料，出现如图 36-5 下部菜单，根据专业要求填写有关参数，点击“确定”，即可得到如图 36-6 的序贯检验图。

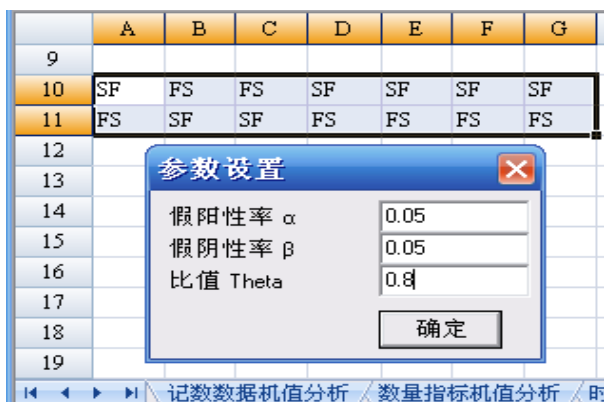


图 36-5 双向质反应序贯检验数据格式及用户界面图

两处理配对计数资料开放性序贯检验图： $2\alpha = 0.05$, $\beta = 0.05$, 比值 $\theta = 0.8$

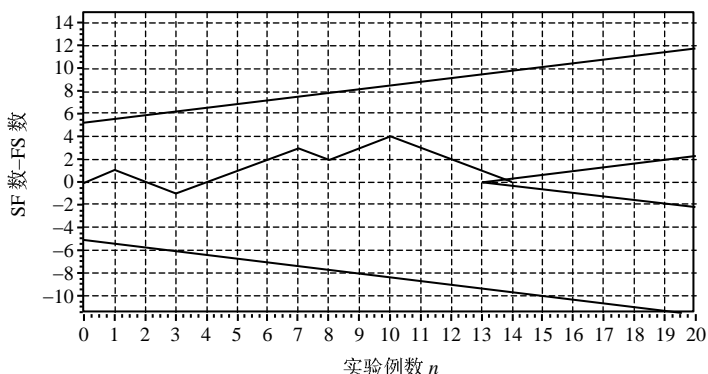


图 36-6 双向质反应开放性序贯检验图

根据图 36-6 中的红色实验线触及不同界线的情况可作出推论：如触及上界 U 线，表示处理优于对照；若触及下界 L 线表示处理劣于对照；若触及 M 界线，表示处理和对照间没有差异。该例结论是两药的效果没有差异。

退出图表状态，系统会给出直线方程的数字结果，供分析参考。

两处理配对计数资料开放性序贯检验图： $2\alpha = 0.05$, $\beta = 0.05$, 比值 $\theta = 0.8$
 直线方程：
 上界 U: $Y = 5.2479 + 0.3219N$
 下界 L: $Y = -5.2479 + 0.3219N$
 中界 M: $Y = 4.1699 + 0.3219N$
 中界 M': $Y = -4.1699 + 0.3219N$

36.1.4 开放型量反应双向序贯试验

配对比较的序贯检验用来比较两种处理反应量的差别。若 A 处理反应量是 X_1 , B 处理反应量是 X_2 , 其差值 $Y = X_1 - X_2$, 若 $Y > 0$, 说明处理 A 的反应量大于 B; 若 $Y < 0$, 说明

处理 A 的反应量小于 B；若 $Y=0$ ，说明处理 A 的反应量和 B 相等。如果已知 Y 的标准差为 σ ，令 $\theta=Y/\sigma$ 可得以标准差为单位的两处理反应量的差值，该差值也作为序贯分析的标准之一。

I 型和 II 型错误的概率，在开放型双向序贯试验的假阳性率有 2 个，故 I 型错误概率是 2α 。双向序贯试验，纵轴为 $Y=X_1-X_2$ 。根据试验标准规定，不仅有上界 U 线(试验措施优于对照)与下界 L 线(对照优于试验措施)，而且有两条中界线 M 和 M'(表示试验措施与对照无差别)。计算公式为

$$\begin{aligned} U: y &= a_1 + bn, & L: y &= -a_1 - bn \\ M: y &= -a_2 + bn, & M': y &= a_2 - bn \end{aligned}$$

式中

$$a_1 = \frac{\sigma}{\theta} \ln \frac{1-\beta}{\alpha/2}, \quad a_2 = -\frac{\sigma}{\theta} \ln \frac{\beta}{1-\alpha/2}, \quad b = \frac{\sigma\theta}{2}$$

例如天麻钩藤饮与六味地黄汤对 II 期高血压病降压影响比较的序贯试验，选用动脉压作为试验指标。首先各患者在服药前测 3 次清晨动脉压，每个患者先服两药中之一 5 天；而后停用 4 天，再服另一药。各患者服用药的次序依随机决定，服药后第 4 天与第 5 天清晨复查动脉压，根据服药前后动脉压变化判断疗效。

拟定指标界限：用 A 与 B 分别代表天麻钩藤饮与六味地黄汤各自的血压下降值，两药血压下降值之差 $Y=A-B$ 。试验者要求 $|Y| \geq 2\text{mPa}$ 时，才认为一药优于另一药。所以，本例可用开放型双向量反应序贯试验。

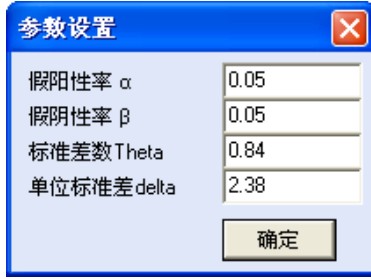
估计总体标准差 σ ：已知两药降压值之差 Y 的标准差 $\sigma \approx 2.38\text{mPa}$ ，试验标准采用 $\theta = Y/\sigma \approx 2.0 \div 2.38 = 0.84$ ，即以标准差为单位的天麻钩藤饮血压下降值与六味地黄汤血压下降值之差 $\theta \geq 0.84$ 时，结论为天麻钩藤饮的降压作用优于六味地黄汤；若 $\theta \leq -0.84$ 时，结论为六味地黄汤的降压作用优于天麻钩藤饮；若 $\theta = 0$ 时，结论为天麻钩藤饮的降压作用与六味地黄汤无差别。

在确定 2α 与 β 水平， $2\alpha = \beta = 0.05$ ，可在 DPS 系统下进行序贯检验。分析时在 DPS 电子表格中，按处理对子序号记录依次逐一输入试验结果。并按图 36-7 编辑、选中数据。

	A	B	C	D	E	F	G	H	I	J	K	L
1	受试者序号	1	2	3	4	5	6	7	8	9	10	
2	天麻钩藤饮	3.93	2.32	5.33	5.40	3.73	2.93	1.23	4.90	2.95	5.33	
3	六味地黄汤	-0.01	3.47	1.30	0.16	2.45	1.23	3	2.67	0.67	-0.82	
4												

图 36-7 双向数量反应序贯检验数据格式

然后在菜单方式下，选择专业统计→序贯检验及成组序贯分析→开放型→双向→计量资料，这是弹出如图所示参数设置用户界面。



根据专业要求填写有关参数，这里为两处理配对计量资料开放性序贯检验，设定 $2\alpha=0.05$ ， $\beta=0.05$ ，差值 = 0.84。点击“确定”，即可得到如图 36-8 的序贯检验图。

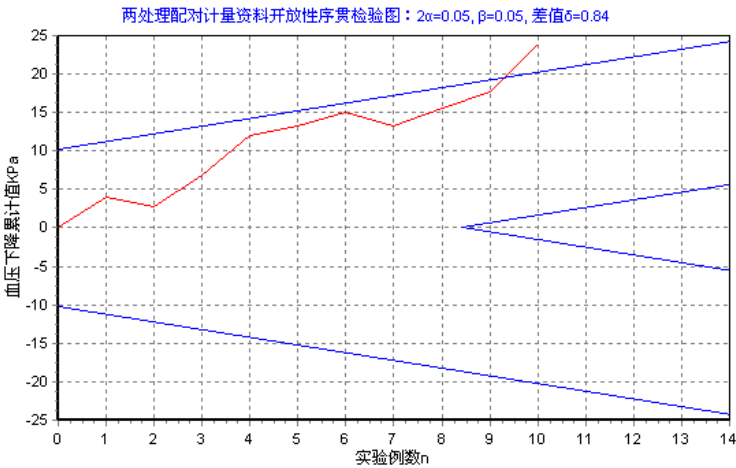


图 36-8 双向数量反应开放性序贯检验图

本例序贯试验结果见图 36-8，到第 10 个处理对时，实验线触及上界 U 线，因此结论为天麻钩藤饮对 II 期高血压病降压疗效优于六味地黄汤。

退出图表状态，系统会给出直线方程的数字结果，供分析参考。

两处理配对计量资料开放性序贯检验图： $2\alpha=0.05$ ， $\beta=0.05$ ，差值 $\delta=0.84$
 直线方程：
 上界 U: $Y=10.3065+0.9996N$
 下界 L: $Y=-10.3065+0.9996N$
 中界 M: $Y=8.4162+0.9996N$
 中界 M': $Y=-8.4162+0.9996N$

36.2 闭锁型序贯试验

闭锁型序贯试验设计通常是双向的，适用于试验结果方向性捉摸不定的双向试验，常采用异体配对比较。

闭锁型序贯试验实验图的边界坐标围成一个封闭的界限，保证了受试对象超过一定数量之前，试验曲线必定要触及上、中、下界限之一，从而克服了开放型序贯试验迟迟

得不到结论的缺点。

闭锁型序贯检验图的边界坐标，不能用一直线方程表达。前人已将设计的界点制成坐标表，在指定的 $2\alpha=\beta$ 水平及比值 θ 时，可根据表中的界点坐标绘制序贯试验界线。

36.2.1 闭锁型双向质反应序贯试验(翼形设计序贯试验)

质反应异体配对比较的闭锁型序贯试验设计的方法很多，这里介绍常用的翼形设计序贯试验。翼形设计(wing design)又称 Armitage 设计。因其设计的上下边界线 U, L(为折线)和中界线 M, M'(为直线)围成的序贯试验图形似机翼而得名。

例如研究丁香柿蒂汤和六君子汤治慢性浅表性胃炎疗效比较的序贯试验。当丁香柿蒂汤优于六君子汤时，记为 SF；反之记为 FS；凡综合评分相等者舍去。用 θ 代表丁香柿蒂汤较六君子汤好的比值，在较少的配对病人中比较两药的优劣时用 $\theta=0.95$ ，取 $2\alpha=\beta=0.05$ 。

进行实验，本例结果按病例对序号记录依次逐一输入。并按图 36-9 中阴影部分，编辑、选中数据。

	A	B	C	D	E	F	G	H
1								
2		SF	SF	FS	SF	SF	SF	
3		SF	SF	SF	SF	SF		
4								

图 36-9 双向质量反应闭锁型序贯检验数据格式及用户界面

然后在菜单方式下，选择专业统计→序贯检验及成组序贯分析→闭锁型→配对计数资料 Armitage 设计，出现如下用户界面。



根据专业要求填写有关参数，点击“确定”，即可得到如图 36-10 的序贯检验图。

DPS 根据试验结果绘制的实验线，至 11 对时，实验线触及界线，停止实验，如图 36-10。

根据反应线触及不同界线的情况作出推断结论：闭锁式序贯试验若反应线触及穿过上边界 U 线，则结论为 SF；若反应线触及穿过下边界 L 线，则结论为 FS，若越过中界 M' 或 M 线，则两种药物的效果无区别。本例实验线触及上界 U 线，结论为丁香柿蒂汤对浅表性胃炎的疗效优于六君子汤。

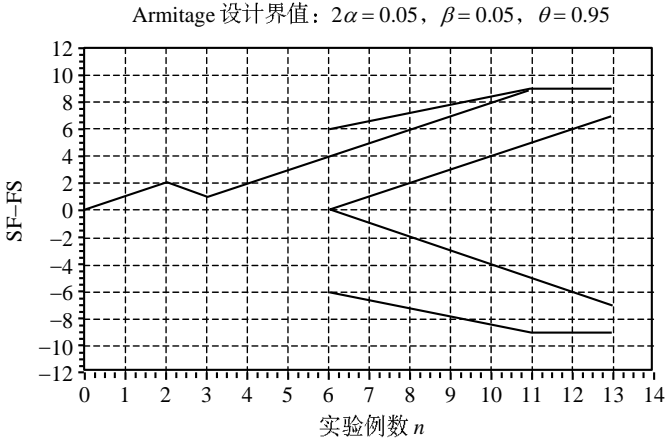


图 36-10 双向质量反应闭锁型序贯检验图

36.2.2 闭锁型双向数量反应序贯试验

闭锁型数量双向反应序贯试验，在 DPS 中提供了楔形闭锁型序贯 t 检验方法。楔形设计又称 Schreiderman-Armitage 设计，它由中界 M 向内推进呈楔形而得名。

山莨菪碱抑制组织蛋白酶的序贯试验，将家兔依品种、窝别、性别与体重配对，用内毒素从静脉注入家兔造成内毒素血症之后，将配对的家兔分别以容量与速度相同的静脉滴注山莨菪碱液和生理盐水(配对对照)。在注入内毒素前后 6 小时分别取血检查血清组织蛋白酶(U/L)的含量，以其增加值作指标。

规定指标界限：配对接受山莨菪碱家兔较接受盐水家兔的血清组织蛋白酶活性增加值 $d \geq 0.60U/L$ 时，认为山莨菪碱有效。

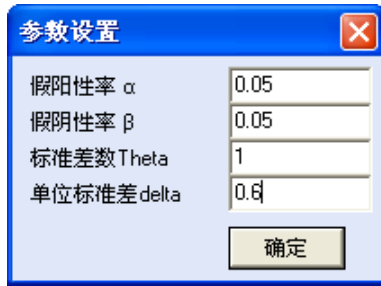
估计 y 的总体标准差 $\sigma_y \approx \sqrt{2} \times 0.424 = 0.60U/L$ 。确定试验标准： $\delta = d/\sigma_y = 0.60/0.60 = 1.0$ 。以标准差为单位的配对接受山莨菪碱家兔较接受盐水家兔的血清组织蛋白酶活性增加值之相对差 ≥ 1.0 时，结论为山莨菪碱降血清组织蛋白酶的作用优于省里盐水；相对差 ≤ -1.0 时，结论为生理盐水作用优于山莨菪碱降血清组织蛋白酶。相对差 $= 0$ 时，两者无差别。

本例取 $2\alpha = \beta = 0.05$ 。进行实验，本例结果按病例对序号记录依次逐一输入。并按图 36-11 中阴影部分。

	A	B	C	D	E	F	G	H	I	J	K
1											
2	处理组	3.4	3.8	3.3	2.8	2	2.5	3	2.2	2.5	
3	对照组	2.2	2.8	3	2.6	1.8	2.5	2.5	1.5	1.5	
4											

图 36-11 双向数量反应序贯检验数据格式

编辑、选中数据后，在菜单方式下选择专业统计→序贯检验及成组序贯分析→闭锁型→配对计量闭锁型楔形 Armitage 设计，执行后出现如下用户界面



根据专业要求填写有关参数，点击“确定”，即可得到如图 36-12 的序贯检验图。

两处理配对计量资料开放性序贯检验图： $2\alpha=0.05$ ， $\beta=0.05$ ，差值 $\delta=1$

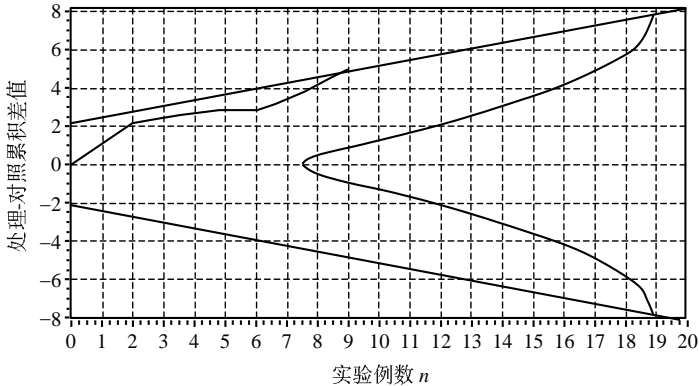


图 36-12 双向数量反应开放性序贯检验图

DPS 根据试验结果绘制的实验线，至 9 对时，实验线触及界线，停止实验，如图 36-12。

根据反应线触及不同界线的情况作出推断结论的原则同前。本实验反应线触及上界 U 线，故可认为山莨菪碱具有降低内毒素血症血清组织蛋白酶的作用。

退出图表状态，系统会给出直线方程的数字结果，供分析参考。

两处理配对计量资料开放性序贯检验图： $2\alpha=0.05$ ， $\beta=0.05$ ，差值 $\delta=1$
 直线方程：
 上界 U: $Y=2.1826+0.3000N$
 下界 L: $Y=2.1826+0.3000N$

36.3 成组序贯试验

36.3.1 成组序贯试验的意义

前面介绍的经典序贯方法是要求将受试者配对后随机分配到两个处理组(或同一受试者先后接受两种处理)，每得到一对试验结果就进行一次统计分析，直到以一定的显著性水平得出拒绝或接受零假设的结论，即可结束试验。这样的优点是当两种处理间确实

存在差异时常可较早地得出结论，从而可减少样本量，缩短试验周期。但该方法要求受试者逐对进入试验，并且只有当得到该试验结果并统计分析后才可确定结束试验还是继续下一个试验。当试验周期较长时就不太适用了。此外，实验者通常是希望不是一对一的逐一分析，而是每隔一段时间把以前所积累起来的资料进行一次统计分析，在这种情况下需采用成组序贯试验方法(group sequential method)。该方法可用于得到试验结果的时间较长及整个试验过程中分几个时间段来重复分析试验结果的情况，它不要求受试者必须配对，因此成组序贯试验方法既保留了传统序贯方法的优点又避免其局限性，且正好与期中分析相配合，可望在临床试验中有较大的发展前途。

成组序贯试验(group sequential trails)首先由 Pocock 在 1977 年提出，其基本方法是将试验的全过程划分为 $N(N \leq 10)$ 个连贯的时段，整个试验最多进行 N 个时段，每个时段内都有 $2n$ 个受试者，随机分配到两组(每组 n 个， $n > 1$)，分别接受 A, B 两种处理。当第 i 个($i=1, 2, \dots, N$)时段内试验完毕时，即将第 1 个到第 i 个时段累积的 $2n_i$ 个观察值进行一次统计分析，若不能拒绝 H_0 ，而且 $i < N$ ，则继续下一阶段试验；如拒绝 H_0 ，即可结束试验；如到 N 个时段结束，仍不能拒绝 H_0 ，则可接受 H_0 。这样，既可避免盲目加大样本而造成浪费，又不致于因样本过小而得不到应有的结论。

因成组序贯试验的统计分析是按时段进行假设检验，通常需要多次假设检验，这会引来 I 类错误的概率增大。为控制总的 I 类误差不超过预先设定的水准 α ，需对每次检验的水准进行调整，记为 α' ，称为名义检验水准。当 α 和 N 确定后，名义水准 α' 也确定了。表 36-1 列出了成组序贯试验的 α' 及相应的临界值 Z' 。只有检验统计量 $Z \geq Z'$ ，才能表明 $P \leq 0.05$ 。

表 36-1 $\alpha=0.05$ 时，已知方差的正态分别变量成组序贯试验设计参数

时段数 N	名义水准 α'	校正界值 Z'	Δ 值			
			$\beta=0.25$	$\beta=0.10$	$\beta=0.05$	$\beta=0.01$
1	0.050	1.960	2.634	3.242	3.605	4.286
2	0.0294	2.178	1.967	2.404	2.664	3.152
3	0.0221	2.289	1.647	2.007	2.221	2.622
4	0.0182	2.361	1.449	1.763	1.949	2.297
5	0.0158	2.413	1.311	1.592	1.759	2.071

表 36-1 中我们只列出 5 个时段($N=5$)的名义水准等参数，因为实际工作中最好 $N \leq 5$ 。 $N > 5$ 之后平均样本量减少并不多，难以体现序贯设计节约样本的优点。虽然表 36-1 是基于方差已知的正态变量情况下导出，但也可用于方差未知的正态变量及二项变量等情况，其误差很小。

36.3.2 成组序贯试验各时段每组样本数 n 的估计

成组序贯试验，每个时段该处理多少样本合适，可有表 36-1 中的 Δ 来估计。两样本均数比较时，如果两处理组 A 和 B 的处理结果 x_A 和 x_B 均服从正态分布，且方差齐性时，反应变量为两处理组总体均数之差 $\mu_A - \mu_B$ (两总体均数 μ_A 和 μ_B 未知)。 H_0 为两处理组效应相同，即 $\mu_A - \mu_B = \mu = 0$ ；双侧检验时， H_1 为两处理组效应不同，即 $\mu_A - \mu_B$ 的总体均数 $\mu = \delta$ 或 $\mu = -\delta$ ， $\delta \neq 0$ ，为试验者所期待的反应变量值和 0 之间的差异。检验效能为 $1 - \beta$ 时，估

算每个时段各组的 n 所设的 $\Delta = \sqrt{n} \times \delta / \sigma$, 两样本率比较时, 将 μ 换成 π , π_A 和 π_B 分别表示两处理组 A 和 B 的总体率, 反应变量为两总体率之差 $(\mu_A - \mu_B)$, $\bar{\pi} = (\pi_A + \pi_B) / 2$, 总体方差 $\sigma^2 = \bar{\pi}(1 - \bar{\pi})$; 检验效能为 $1 - \beta$ 时, 估算每个时段各组的 n 所设的 $\Delta = \sqrt{n} \times \bar{\pi}(1 - \bar{\pi}) / (\mu_A - \mu_B)$ 。从表 36-1 可查得 $\alpha = 0.05$, $\beta = 0.10$ 和 0.05 时的 Δ 值。DPS 中提供了样本数估计的功能, 应用时只要在专业统计 \rightarrow 序贯检验及成组序贯分析 \rightarrow 量反应成组序贯设计, 出现如图 36-13 右边所示用户界面, 输入相关参数后, 即可得到 n 的估计值。



图 36-13 成组序贯试验设计界面(左边是均值检测, 右边是两率比较)

得到 n 估计值之后, 系统将生成各个待样本的编码、分组方案。处理样本编码方式有整体, 即对 $2nN = 2 \times 14 \times 5 = 140$ 个样本, 从 1 到 140 编码, 然后分组。如果选中分段编码, 则对待处理样本, 分时段, 按每时段处理样本数 $(2n)$ 进行编码, 这是即按 1~28 进行编码。然后随机分组。

36.3.3 总体未知时两均数比较

当两处理组 A 和 B 的处理结果 x_A 和 x_B 均服从正态分布, 且方差齐性时, 总体方差未知时, 可进行两样本均数比较的成组序贯试验分析。

成组序贯分段试验逐段分析 每完成一个时段试验, 都分别对两组计算从第 1 时段累计至该时段(记为第 i 时段)的均数 \bar{x}_{Ai} , \bar{x}_{Bi} 和标准差 s_{Ai} , s_{Bi} :

$$\bar{x}_{Ai} = \frac{1}{i \times n} \sum_{k=1}^{i \times n} x_{Ak}, \quad \bar{x}_{Bi} = \frac{1}{i \times n} \sum_{k=1}^{i \times n} x_{Bk},$$

$$s_{Ai} = \sqrt{\sum_{k=1}^{i \times n} (x_{Ak} - \bar{x}_{Ai})^2 / (i \times n - 1)}, \quad s_{Bi} = \sqrt{\sum_{k=1}^{i \times n} (x_{Bk} - \bar{x}_{Bi})^2 / (i \times n - 1)}$$

进行 t 检验, H_0 为 $\mu_{Ai} = \mu_{Bi}$, H_1 为 $\mu_{Ai} \neq \mu_{Bi}$ 。 $\alpha = 0.05$, 第 i 时段的 t 检验统计量为

$$t_i = (\bar{x}_{Ai} - \bar{x}_{Bi}) / \sqrt{(s_{Ai}^2 + s_{Bi}^2) / (i \times n)}, \quad df_i = 2(i \times n - 1)$$

将 t_i 与名义水准 α' 和自由度 $df_i = 2 \times i \times n - 2$ 时的 t 值 $ta'(2 \times i \times n - 2)$ 比较, 若 $t_i < ta'(2 \times i \times n - 2)$, 如 $i < N$, 需继续进行第 $i+1$ 时段的试验; 若 $t_i \geq ta'(2 \times i \times n - 2)$, 则结束试验, 结论为两种处理效应不同。如实验进行到 $i=N$ 时仍是 $t_i < ta'(2 \times i \times n - 2)$, 则结束试验, 结论为两种处理效应相同。

例如比较新药(A)与传统药(B)治疗再生障碍性贫血的疗效, 以血红蛋白(g/L)增加作为观察指标。要求新药疗效平均比传统药高出值 $\delta = 10\text{g/L}$ 。已知传统药标准差的估计值 $s = 15\text{g/L}$, 规划进行 5 期试验, 则每个时期每组处理需要样本数 n 用的估计如图 36-13 左边所示, 即每个时段每组需分配 14 个病例进行试验, 最大样本数为 $2nN = 2 \times 14 \times 5 = 140$ 。采取整体编码方式, 给出成组序贯试验各个待处理样本编号、分组如下:

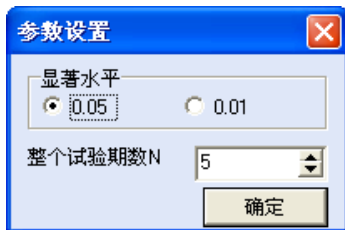
5 个时期, 每组 14 个样本, 每个时期 28 个样本, 共 140 个样本。														
时期 1														
对照	102	11	44	90	74	6	42	39	114	116	81	100	101	30
处理	23	89	13	40	115	96	38	117	68	9	103	92	59	85
时期 2														
对照	120	2	125	5	91	83	3	22	118	111	73	18	126	26
处理	32	67	58	56	28	130	95	122	24	50	14	80	86	132
时期 3														
对照	64	43	61	121	137	7	93	138	139	29	108	63	41	87
处理	49	35	135	8	94	52	77	106	16	10	131	110	123	25
时期 4														
对照	47	127	33	69	4	107	128	119	21	79	72	136	99	19
处理	134	66	71	48	98	97	46	51	82	109	60	54	57	133
时期 5														
对照	129	31	75	104	84	20	65	53	36	62	45	124	37	76
处理	105	140	70	113	34	55	27	1	17	88	15	78	112	12

经过 2 期实验, 得到实验结果应用序贯检验进行期中分析。本例结果按时期, 每个时期按处理、对照组的顺序, 逐行输入(图 36-14)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1																	
2	第 1 时段	处理药组 A	85.2	81.1	88.5	94.8	47.8	93.5	84.6	75.3	49.3	76.4	90.1	82.1	94.3	85.2	
3		对照药组 B	76	87.1	94.1	68.6	59.9	80.7	78.5	42.9	43.5	72.3	75	69.9	56.3	77.9	
4	第 2 时段	处理药组 A	76	40.1	94.1	95.2	84	80.7	78.5	70.5	79.4	76	69	87.9	80.5	91.3	
5		对照药组 B	73.9	76.4	86.5	51.7	41	85.3	86.5	51.3	54.7	71.3	65.9	85.1	43.8	89.5	
6																	

图 36-14 成组序贯 t 检验数据格式

按图 36-14 中阴影编辑、选中数据后, 进入菜单, 选择专业统计→序贯检验及成组序贯分析→成组序贯设计分析→总体方差未知时的 t 检验, 执行后出现如下用户参数设置界面



根据要求填写有关参数，点击“确定”，即可得到成组序贯检验结果如下：

校正 $\alpha=0.0158$	计划总期数=5	每组例数=14					
批次	A 组均值	B 组均值	合并标准差	Ti	df	T(0.0158)	备注
1	80.5857	70.1929	5.6388	1.8431	26	2.5823	需继续实验
2	79.6929	69.4857	4.0083	2.5465	54	2.4921	可终止实验

从分析结果可以看出，第 1 时段，计算得 $\bar{x}_{A1}=80.5857$ ， $\bar{x}_{B1}=70.1929$ ，计算得 t_1 值 = 1.843，自由度 $df_1=2(1 \times 14 - 1)=26$ ，相应自由度下，名义显著水平为 0.0158 时的 t 界值等于 2.5823， $t_1 < t_{(\alpha=0.0158)}$ 界值， $P_1 > 0.05$ ，不拒绝 H_0 ，继续进行第 2 时段的试验。

第 2 时段， $i=2$ ， $n=14$ ，试验结果见图 36-14 中阴影所示的第 3~4 行，计算得到检验统计量 $t_2=2.5465$ ，在自由度 $df_2=2(2 \times 14 - 1)=54$ 下，名义显著水平为 0.0158 时的 t 界值等于 2.4921， $t_2 > t_{(\alpha=0.0158)}$ 界值， $P_2 < 0.05$ ，拒绝 H_0 ，结束试验，可认为新复方比传统复方升高血红蛋白平均比出 $\delta=10\text{g/L}$ 。

如果总体方差已知，这时两均数比较序贯试验分析，计算过程更简单一些。当两处理组 A 和 B 的处理结果之差 θ 服从正态分布，其方差 σ^2 已知，对于双侧检验， $H_0: \mu=0$ ，对应假设 $H_1: \mu \neq 0$ 。当实验进行到第 i 个时段时，累积样本均数之差为 \bar{d}_i ，这时可用 Z 检验进行统计分析

$$Z_i = \sqrt{i \cdot n \cdot \bar{d}_i} / \sqrt{2\sigma^2}$$

若 Z_i 大于表 36-1 中的界值 Z' ，则可拒绝零假设，结束实验，结论为两种处理效应不同；否则再进行下一个时期的实验。若到第 N 个时段结束仍不能拒绝零假设，结论为两种处理效应相同。

总体方差已知时两均数比较序贯试验分析，数据格式和总体方差未知时两均数比较序贯试验分析相同。功能在菜单专业统计→序贯检验及成组序贯分析→成组序贯设计分析→总体方差已知时 t 检验里面，分析时需要给出总体方差值，结果分析也类似总体方差未知时两均数比较，故在此从略。

36.3.4 两样本率比较的序贯试验

当实验结果是质反应的，进行一个时段试验，就可分别各组计算累计至该时段的样本率，进行统计分析。成组序贯试验各时段两率比较的正态近似检验统计量为

$$Z_i = (P_A - P_B) \sqrt{i \times n} / \sqrt{2\bar{P}_i(1 - \bar{P}_i)}$$

式中 P_{A_i} , P_{B_i} 分别为 A 组、B 组从第 1 时段到第 i 时段试验结果累计计算的样本率;
 $\bar{P}_i = (P_{A_i} + P_{B_i})/2$, 为第 i 时段的两组平均率。

例如比较新药 A 与常规药 B 治疗某病证的疗效。已知 B 药的有效率为 60%，经预试可期望 A 药的有效率达到 80%。用成组序贯试验作检验、分析。

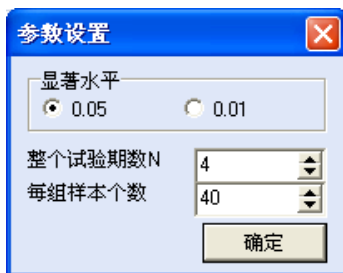
首先确定时段，这里取 $N=4$ ，包括中间 3 次分析与最后 1 次总的统计分析。确定 α 与 β 均等于 0.05。则每个时期每组处理需要样本数 n 用的估计如图 36-13 右边所示，即每个时段每组需分配 40 个病例进行试验。 $2nN=2 \times 40 \times 4=320$ ，可知本例成组序贯试验的样本数不超过 320 例，并类似量反应成组序贯设计，给出各个处理样本分组、编号(略)。

经过 2 期实验，得到实验结果应用序贯检验进行期中分析。本例结果按时期，每个时期按处理、对照组的顺序，逐行输入(图 36-15)。

	A	B	C	D	E	F
1		处理组 A		对照组 B		
2		有效	无效	有效	无效	
3	第一时段	32	8	24	16	
4	第二时段	33	7	25	15	
5						

图 36-15 成组两样本率比较序贯检验数据格式及用户界面

按图 36-15 中阴影部分编辑、选中数据，在菜单方式下选择专业统计→序贯检验及成组序贯分析→成组序贯设计分析→两样本率比较。执行后出现如下用户界面。



根据要求填写有关参数，点击“确定”，即可得到成组序贯检验结果如下：

校正 $\alpha=0.0182$		计划总期数=4		每组例数=40			
大样本近似正态 Z 检验							
批次	rA	率 A(%)	rB	率 B(%)	Z(0.0182) 备注		
1	32	80	24	60	1.9518 2.3615 需继续实验		
2	65	81.2500	49	61.2500	2.7948 2.3615 可终止实验		
卡方检验							
批次	rA	率 A(%)	rB	率 B(%)	卡方	卡方临界值	备注
1	32	80	24	60	3.8095	5.5768	需继续实验

2	65	81.2500	49	61.2500	7.8108	5.5768	可终止实验
---	----	---------	----	---------	--------	--------	-------

本例第 1 时段, 试验结果序贯检验结果 $P_{A1}=0.80$, $P_{B1}=0.60$, $\bar{P}_1=0.70$, $Z_1=1.9518$ 。因 $Z_1 <$ 校正界值 2.361, 故不拒绝 H_0 , 继续进行第 2 个时段试验;

第 2 个时段, 根据试验结果计算得 $P_{A2}=0.8125$, $P_{B2}=0.6125$, $\bar{P}_2=0.712$, $Z_2=2.7948$ 。因 $Z_2 >$ 校正界值 2.361, 拒绝 H_0 , 可认为 A 药疗效优于 B 药。

有的文献还介绍了应用卡方检验方法进行两样本率比较的序贯检验, DPS 也给出了其分析结果, 供用户选用。

参 考 文 献

- 金丕焕. 1993. 医用统计方法[M]. 上海: 上海医科大学出版社.
- 刘明芝, 周仁郁. 2006. 中医药统计学与软件应用[M]. 北京: 中国中医药出版社.
- Elfring G L, Schultz J R et al. 1973. Group sequential designs for clinical trials[J]. Biometrics, 29(3):471-477.
- Mazumdar M and Bang H. 2008. Sequential and Group Sequential Designs in Clinical Trials: Guidelines for Practitioners//Rao C R, Miller J P and Rao D C (Eds.). Handbook of Statistics Vol. 28. [M]. Elsevier Sciences, 491-512.
- Pocock S J. 1977. Group sequential methods in the design and analysis of clinical trials[J]. Biometrika, 64:191-199.
- Schoenfeld D A. 2001. A simple algorithm for designing group sequential clinical trials[J]. Biometrics, 57(3):972-974.
- Schneiderman M A and Armitage P. 1962. A family of closed sequential procedures[J]. Biometrika, 49:41-56.

第 37 章 生存分析

生存分析(survival analysis)是研究特定人群生存过程的统计分析方法。在生存分析中,常用生存率(或称生存函数)、风险率(或称危险率、风险函数)、生存时间 T 的分布函数或密度函数描述生存过程。

T 的分布函数 $F(t)$ 为生存时间 T 不超过时点 t 的概率,即 $F(t)=Pr(T\leq t)$ 。生存时间的密度函数 $f(t)$ 是 $F(t)$ 的导数,即 $f(t)=F'(t)$, $f(t)$ 描述所有个体在时点单位时间内死亡的危险性。生存率 $S(t)$ 是指生存时间 T 超过时点 t 的概率,即 $S(t)=Pr(T>t)=1-F(t)$ 。例如 5 年生存率是指生存时间 T 超过 5 年的概率,记为 $S(5)$ 。风险率 $h(t)$ 是密度函数与生存率的比,即 $h(t)=f(t)/s(t)$, 风险率 $h(t)$ 描述已活过时点 t 的个体在时点 t 后单位时间内死亡的危险性。

实际工作中,我们并不能得到所有患者的生存时间资料,即不可能得到 T 的总体分布,因此只能利用样本资料近似地描述总体 T 的分布,对 T 的分布作出有关的统计推断。若有大样本的完全数据资料,则可以借助频数分布表近似描述 T 的分布(表 37-1)。

表 37-1 100 名某病患者生存时间数据资料

I	确诊后(年) $t_i\sim$	死于本病例数 f_i	频数/年 f_i /年	频率/年 $f(t_i)$	累计频率/年(t_i+1)
1	0~	5	5	0.05	0.05
2	1~	10	10	0.10	0.15
3	2~	20	20	0.20	0.35
4	3~	30	30	0.30	0.65
5	4~	10	10	0.10	0.75
6	5~	5	5	0.05	0.80
7	6~	4	4	0.04	0.84
8	7~	3	3	0.03	0.87
9	8~	2	2	0.02	0.89
10	9~	2	2	0.02	0.91
11	10~	1	1	0.01	0.92
12	11~12	8	0.8	0.008	1.00

表 37-1 是一组完全资料,据此资料,可估计生存时间不超过时点 t_i 的概率为 $F(t_i)$, 在 j 时点 t_i 后单位时间内死亡的危险性为 $f(t_i)$ 。例如,估计该病患者生存时间不超过 2 年的概率 $F(2)=0.15$, 估计该病患者在时点 2(年)死亡的危险性 $f(2)=0.20$, 估计生存时间超过 2 年的概率,即估计 2 年生存率 $S(2)=1-0.15=0.85$, 在时点 2(年)的风险率,即估计活过 2 年者在接下去的时刻死亡的危险率 $h(2)=0.20/0.85=0.24$ 。

一般来讲,生存资料中包含截尾数据,并且我们的目的不仅仅是为了描述某特定人群的生存时间分布,还要研究影响生存时间的因素,故相应的统计方法较复杂。生存分析方法一般可分为参数、非参数和半参数三大类。如果生存时间的分布符合某一特定分布类型,如正态分布、Weibull 分布、指数分布等,则可利用特定的分布函数来分析,这称之为参数法;若不知道生存时间的分布类型,而对分布或其某特征作推断,就只能用半参数法或非参数法。如用寿命表估计期望寿命;用 Kaplan-Meier 法求生存率,作出生存曲线;用 Logrank 检验(或 Gehan 检验)等作不同组别生存过程差异的显著性检验,均属非参数法。非参数法无法分析多个危险因素对生存时间的影响。本系统提供多因素生存分析方法,即半参数法比例风险模型(cox model)、参数法的指数分布模型和 Weibull 分布模型。

生存分析资料,要求每个观察个体除了有考察影响因素外,还需有表明生存时间(或观察时间)长短和表明结局(“失效”与否)。

37.1 生存率估计

37.1.1 小样本资料生存率的估计

对小样本资料,生存率的估计主要是用乘积极限法(product-limit estimate),又称 Kaplan-Meier 法,它是一种非参数方法。

生存率曲线:小样本资料生存率曲线也称 Kaplan-Meier 曲线,它是以前时间 t 为横轴,生存率 $p(X>t)$ 为纵轴,水平横线的长短代表一个 t 时点到下一个 t 时点的距离,从而表示时间与生存率关系的曲线。由图可直观地比较各样本的自下而上率曲线,也可对某一病例任意时刻的生存率作出估计,反之,也可由任意生存率估计生存时间。一般将生存率 0.5 所对应的生存时间称为中位生存时间,它可用内插法求得。

总体生存率可信区间的估计,可采用正态近似原理估计某时点总体生存率的可信区间,公式为 $P(X>t) \pm u_{\alpha} S_p(X>t)$ 。注意,对于生存率曲线末端尾部的生存率,不适合用该方法计算总体生存率的可信区间,因为例数较少时误差较大,会出现一些不合理的现象。

对未分组资料生存分析进行生存率的估计,在 DPS7.05 以后的版本中,其数据编辑如图 37-1 阴影部分。然后在 DPS 系统支持下,执行主菜单“专业统计”的“生存分析”下的“Kaplan-Meier 法生存率估计”功能,可得到 Kaplan-Meier 生存曲线(图 37-1 右上部),并输出数值计算结果(图 37-1 下部)。

DPS 系统根据数据直接生成的 Kaplan-Meier 曲线,其横轴标注可放在数据块的第一列的上面。但要注意,如果该标注是中文,那么 DPS 在纵轴上自动给出的标注为“生存率”;如果横轴上的标注是纯英文,那么 DPS 在纵轴上的标注相应为英文“proportion of surviving”。

37.1.2 大样本分组资料生存分析

当随访资料的例数较多(如 $n>50$)时,可先将原始资料进行分组,然后再进行生存率的估计。另外,有很多随访研究设计的随访时间是一年或一个月一次,随访结果也只有



图 37-1 小样本资料生存分析图示

该年或该月期间的若干观察人数、发生失败事件人数和截尾人数，没有各个病例的准确观察时间，因此所获得的资料只能视为分组资料。

分组资料中，一般记 l_0 或 $l(t_0)$ 为开始时(或 t_0 时)观察的存活的人数， l_i 或 $l(t_i)$ 为 t_i 时刻仍存活的人数， $p_{ij}=l_j/l_i$ 表示 t_i 时存活者到 t_j 时仍存活的概率，也称条件生存概率， $p_i=l_{i+1}/l_i$ ，也记为 $n_i p_i$ ， $n_i=t_{i+1}-t_i$ ， $q_i=1-p_i$ ，为 t_i 时存活者到 t_{i+1} 时死亡的概率， $d_i=l_i-l_{i+1}$ 为死亡人数， $s(t)=l(t)/l_0 = Pr(T \geq t)$ 为开始的 l_0 个人到 t 时仍存活的概率，称为生存函数， $s_{(\infty)}=0$ ， $f(t) = \frac{-ds}{dt}$ 为生存的密度函数， $h(t) = \lim_{n \rightarrow 0} \frac{nq_t}{n}$ ， nq_t 为 $[t, t+n)$ 上的死亡概率， $h(t)$

为危险(率)函数， $H(t) = \int_{t_0}^t h(u)du$ 是累加危险率函数， $m_i = d_i / \int_{t_i}^{t_{i+1}} l(t)dt$ 是 $[t_i, t_{i+1})$ 上的年龄别死亡率。一般，抽样总是在有限个时刻点上抽取的，设抽样点分别为

$$0 = t_0 < t_1 < t_2 < \dots < t_i < t_{i+1} < \dots < t_w < t_{w+1}$$

t_{w+1} 是抽样截止时间。在实际抽样时，常会有“漏失”或“失访”了的资料，如记 w_i 为在 (t_i, t_{i+1}) 内“漏失”及“失访”的人数，一般用

$$\hat{q}_i = d_i / (l_i - w_i / 2)$$

估计死亡概率 q_i ，再用 $p_i = 1 - q_i$ 计算生存概率。经典的计算 $s(t) = l(t)/l_0 = Pr(T \geq t)$ 的方法称为 Kaplan-Meier(简记 K-M)公式

$$\hat{S}(t_i) = l_i/l_0 = \frac{l_1}{l_0} \cdot \frac{l_2}{l_1} \cdot \frac{l_3}{l_2} \cdots \frac{l_i}{l_{i-1}} = p_0 p_1 p_2 \cdots p_{i-1}, \quad i=1, 2, \dots, w$$

这是一个不连续的阶梯函数。它的方差渐近公式为

$$Se(\hat{S}(t_i)) = \hat{S}(t_i) \left\{ \sum_{j=0}^{i-1} \frac{q_j}{l_j p_j} \right\}^{1/2}, \quad i=1, 2, \dots, w$$

孙尚拱对以上公式提出了改进，他认为在每一个小区间内，可用指数分布代替，对任一时间 t ，先找到 t 所在的子区间，记 $t \in [t_i, t_{i+1})$ ，则令

$$\begin{aligned} \hat{S}(t) &= p_0 p_1 \cdots p_{i-1}, & \text{如果 } t = t_i \\ \hat{S}(t) &= p_0 p_1 \cdots p_{i-1} \times \exp\{-\bar{h}_i(t - t_i)\}, & \text{如果 } t_i \leq t < t_{i+1} \end{aligned}$$

其中 $\bar{h}_i = \frac{-1}{t_{i+1} - t_i} \ln(p_i)$ 为 $[t_i, t_{i+1})$ 上的平均危险率，这是一个连续的函数。它的方差渐近公式为

$$Se(\hat{S}(t)) = \hat{S}(t_i) \exp\{-\bar{h}_i(t - t_i)\} \left\{ \sum_{j=0}^{i-1} \frac{q_j}{l_j p_j} + \frac{(t - t_i)^2}{n_i} \times \frac{q_i}{l_i p_i} \right\}^{1/2}, \quad i=1, 2, \dots, w$$

DPS7.05 以后的版本吸收了孙尚拱的研究成果，给出了生存分析更多的信息。DPS 中生存率估计的数据编辑格式和以前的版本不相同。例如 Carbone 等(1967)对 112 名骨髓癌患者给出了他们的生存数据(单位：月)，其数据如图 37-2 阴影部分所示。

时间 t	区间长 n	条件死 亡率 q	条件生 存率 p	se_p	hazard	se_haza	生存率 St	se_sur	iei	e	se_ei	A
0.0000	5.5000	0.1614	0.8386	0.0348	0.0320	0.0076	1.0000	0.0000	5.0430	20.8537	1.5562	21.7732
5.5000	5.0000	0.1730	0.8270	0.0393	0.0380	0.0095	0.8386	0.0348	4.5539	18.8545	1.6203	19.9576
10.5000	5.0000	0.2416	0.7584	0.0496	0.0553	0.0131	0.6935	0.0438	4.3682	17.2916	1.7141	19.7879
15.5000	5.0000	0.1818	0.8182	0.0520	0.0401	0.0127	0.5260	0.0478	4.5303	17.0406	1.8530	17.9664
20.5000	5.0000	0.2444	0.7556	0.0641	0.0561	0.0170	0.4303	0.0477	4.3604	15.2904	1.9558	17.2171
25.5000	5.0000	0.2388	0.7612	0.0737	0.0546	0.0194	0.3251	0.0454	4.3759	14.4662	2.1377	15.9997
30.5000	10.0000	0.5106	0.4894	0.1031	0.0715	0.0211	0.2475	0.0420	7.1453	13.2560	2.3429	18.9254
40.5000	10.0000	0.3750	0.6250	0.1712	0.0470	0.0274	0.1211	0.0328	7.9787	12.4871	2.6492	13.0982
50.5000	10.0000	0.5000	0.5000	0.3536	0.0693	0.0707	0.0757	0.0291	7.2135	7.2135	2.2581	6.3867
60.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0378	0.0000	0.0000	0.0000	0.0000	0.0000

分位点估计结果略。

对大样本资料进行生存率估计，可在 DPS 系统支持下，执行主菜单“专业统计”的“生存分析”下的“大样本寿命表法生存率估计”功能，给出数值计算结果，以及生存函数的分位点、标准误，输出到电子表格里面。最后给出生存率与时间关系及生存函数与分位点关系曲线(图 37-3)。图的文件名分别是“生存曲线.bmp”和“生存函数与分位点关系曲线.bmp”。

1	时间	期初病例	死亡数	截尾数
2	0	112	18	1
3	5.5	93	16	1
4	10.5	76	18	3
5	15.5	55	10	0
6	20.5	45	11	0
7	25.5	34	8	1
8	30.5	25	12	3
9	40.5	10	3	4
10	50.5	3	1	2
11	60.5	0	0	0
12				

图 37-2 大样本分组资料生存分析图示

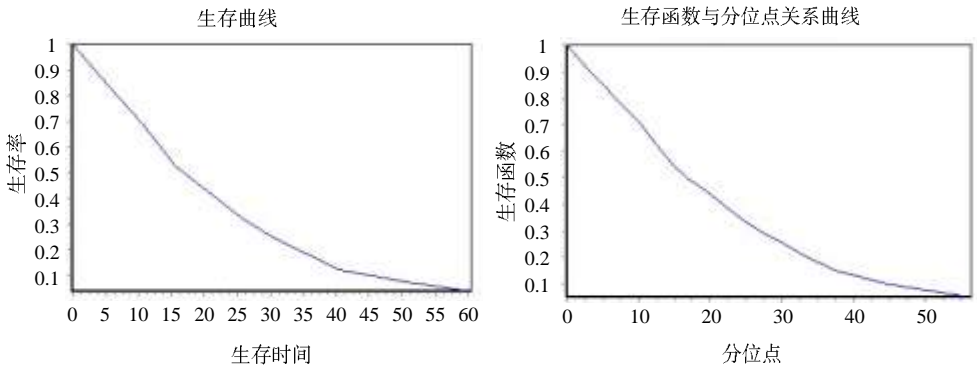


图 37-3 生存率与时间关系(左)及生存函数与分为数关系(右)曲线图

分析结果中， t 是指时间区间的起点， n 是区间长度，即 $n_i = t_{i+1} - t_i$ ； q, p 是区间内的死亡概率与生存概率； ie_i 为 $n_i e_i$ ，即存活者在该区间内的平均生存时间， $t=0$ 时 $n_i e_i = 5.04303$ ，说明手术后的 5 个半月之内，病人平均可存活 5.0420 月； e 是存活者在该区间的起点到末端，即在 $(t_i, 60.5)$ 内的平均生存时间， $t=0$ 时 $e = 20.8537$ ，说明手术后的 60.5 月内，病人可平均存活 20.8537 月； se_{e_i} 是 e 的标准误； a 是生存概率 p 对于生存时间的“灵敏度”指标。上面结果中的第一个观察值的 $a = 21.7732$ 为最大，说明起始区间上的生存概率对平均生存时间的影响比另外的区间上的影响都要大。 a 值(如同下面看到的)对于估计平均生存时的各种变化特别有用。从 hazard(危险率)值看，0.071654 值最大，它对应于区间(30.5, 40.5)，即骨癌病人在术后生存时间 30 至 40 个月期间的死亡危险性最大，这在 q 值(死亡概率)的大小中也可以看出。

37.2 两样本生存率 Log-rank 检验

DPS 系统采用对数秩检验(log-rank test)方法进行两样本生存率曲线比较，其基本思想是如果无效假设成立，即两总体生存曲线无差别时，根据不同日期两种处理的期初人数和残废人数计算各种处理在各个时期的理论残废数。如无效假设成立，则实际死亡数

与理论死亡数不会相差太大，否则，应认为无效假设不可能成立，两条生存率曲线差异有统计学意义。对数秩检验计算步骤是：

(1) 按时间排序将两组未截尾的完全生存时间从小到大混合排序。

(2) 根据不同生存时间的死亡数、期初观察数，计算每个时间区间上的 2×2 列联表。这时第 i 个时间区间的 2×2 表如表 37-2。

表 37-2 第 i 个时间区间的 2×2 表形式

	死亡	生存	合 计
组 I	a_i	b_i	n_{i_1}
组 II	c_i	d_i	n_{i_2}
合 计	D_j	S_j	n_i

这里， n_{i_1} 和 n_{i_2} 分别是第 i 个时间区间在开始时第 1 组和第 2 组人数，他们开始时没有发生事件也没有失访； a_i 和 c_i 分别是第 i 个时间区间第 1 组和第 2 组的死亡人数； b_i 和 d_i 分别是第 i 个时间区间第 1 组和第 2 组的生存人数。

对上面的 2×2 表做 Mantel-Haenzsel 检验，即计算检验统计量

$$\chi_{\text{LR}}^2 = \frac{(|O - E| - 0.5)^2}{\text{var}_{\text{LR}}}$$

此处 H_0 ：两组的死亡率没有差异。式中

$$O = \sum_{i=1}^k a_i, \quad E = \sum_{i=1}^k E_i = \sum_{i=1}^k \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

$$\text{var}_{\text{LR}} = \sum_{i=1}^k v_i = \sum_{i=1}^k \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2(n_i - 1)}$$

在 H_0 下， χ_{LR}^2 是具有自由度为 1 的卡方分布。根据 χ_{LR}^2 值可计算确切的概率值，并进行统计推断。

目前，大多数统计软件，如 SAS，SPSS 进行 Log-Rank 检验，计算卡方值都是使用的未校正的公式 $\chi_{\text{LR}}^2 = (O - E)^2 / \text{var}_{\text{LR}}$ ，故 DPS 也是采用该式计算卡方值。

如果随机变量以四格表的例数 N_i 为权重，便可获得 Breslow 统计量

$$\chi_B^2 = \frac{(\sum N_i O - \sum N_i E)^2}{\sum N_i^2 \text{var}_{\text{LR}}^2}$$

在 H_0 下， χ_B^2 是具有自由度为 1 的卡方分布。

如有 22 例 III 期非小细胞肺癌患者在不同日期经随机化分配到放疗组和放化疗联合

组，从缓解出院日开始随访，随访结果如下：

甲组：放疗组 1, 2, 3, 5, 6, 9⁺, 11, 13, 16, 26, 37⁺；

乙组：放化疗联合组 10, 11⁺, 14, 18, 22, 22, 26, 32, 38, 40⁺, 42⁺。

试比较放疗和放化疗联合两种治疗方案的生存率曲线有无差别。

在分析之前，建立检验假设。 H_0 ：单放疗与放化疗联合两治疗方案的生存率曲线分布相同； H_1 ：单放疗与放化疗联合两治疗方案的生存率曲线分布不同。

分析时，将上述数据整理成如图 37-4 格式(阴影部分)，然后执行 Log-Rank 检验。系统先输出两样本的生存曲线(图 37-4 右上部)，然后输出分析得到的数值结果如下(图 37-4 下部)。

从输出的结果来看，其卡方值的显著水平 p 均小于 0.05，按 $\alpha = 0.05$ 水准拒绝 H_0 ，接受 H_1 ，故可以认为单独放疗与放化疗联合治疗 III 期非小细胞肺癌的生存率曲线分布不同。

又因为放疗组的观察值 $O=9$ ，大于期望值 $E=5.128$ ，这说明放疗组的死亡率更高，即放化疗联合治疗的效果较好。

另外，DPS 系统根据数据直接生成的两组资料的 Kaplan-Meier 生存曲线，其横轴标注可放在数据块的第一列的上面。但要注意，如果该标注是中文，那么 DPS 在纵轴上自动给出的标注为“生存率”；如果横轴上的标注是纯英文，那么 DPS 在纵轴上的标注相应为英文“proportion of surviving”。两条曲线的标注(名称)，如图 37-4 所示，可分别放在数据资料的第 2 列、第 4 列上面。

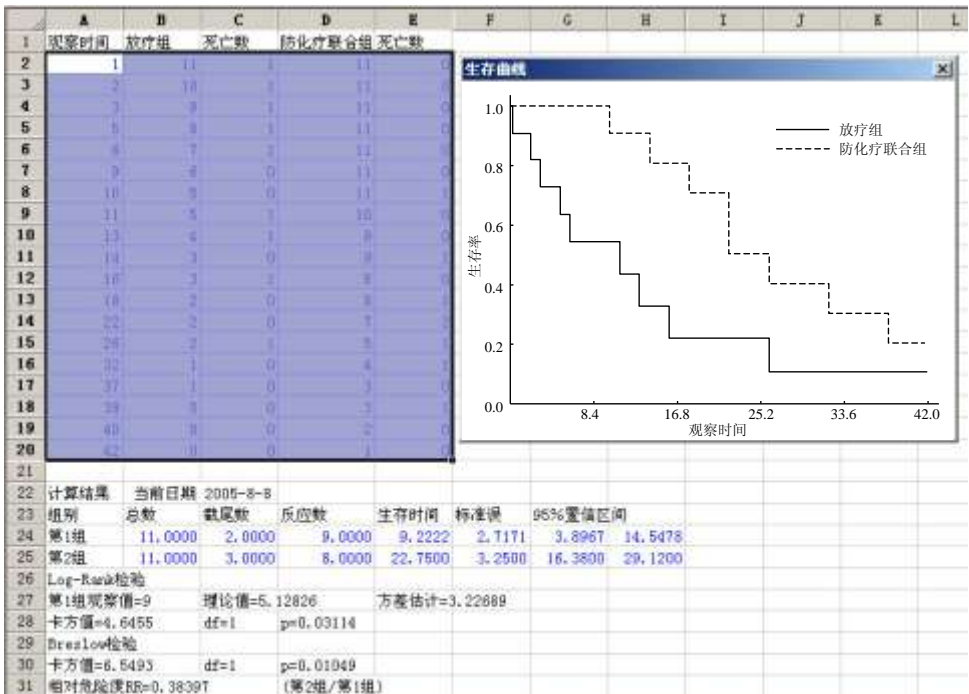


图 37-4 两组资料 Log-Rank 检验数据编辑及结果输出格式

37.3 寿命表的编制与分析

寿命表(life table)是根据某地特定人群的年龄组死亡率编制而成的。一般可分为队列寿命表(cohort life table)和现时寿命表(current life table)。由于人的生命周期很长,用队列寿命表的方法去研究人群的生命过程,不仅随访人数要很多,而且随访时间要数十年。因此,在编制寿命表时,一般不使用队列寿命表,而是使用现时寿命表。

现时寿命表是假定同时出生的一代人(一般为10万)按照某年(或某一时期内)特定人群的年龄组死亡率先后死去,直至死完为止,用寿命表方法计算出这一代人在不同年龄组的“死亡概率”、“死亡人数”、刚满某年龄时的“尚存人数”及其“预期寿命”等指标。这些指标不受人群年龄构成的影响,不同的地区可以相互比较。

37.3.1 现时寿命表

根据年龄分组方法的不同,现时寿命表可分为完全寿命表(complete life table)及简略寿命表(abridged life table)两种。完全寿命表年龄分组的组距是1岁,而简略寿命表习惯上将年龄组距定为5岁。因为婴儿死亡率对寿命表的影响相当大,所以简略寿命表也将1个5岁年龄组拆分为组距为1岁的“0~”岁组和组距为4岁的“1~”岁组,从5岁开始年龄组距才为5岁(图37-6)。

编制全国或地区、城市或农村的简略寿命表,因不同性别年龄组死亡率不同,故应分性别编制,现时简略寿命表的编制步骤如下:

(1) 基本数据准备。

年龄分组:记为“X~”,其中X为“刚满年龄(exact age)”,如表中的“5~”表示刚满5岁(即刚过第5个生日)至不到10岁的儿童。以及年平均人口数、实际死亡人数。

(2) 计算年龄组死亡概率。

先计算死亡率,死亡率 = $\frac{\text{同年内该年龄组死亡人数}}{\text{某年某年龄组平均人口数}} \times 1000\%$,记为 ${}_i m_x$ 。然后计算

年龄组死亡率(age specific probability of dying)记为 ${}_i q_x$, ${}_i q_x$ 是指X岁尚存活者在今后i年内死亡的概率,这里i为年龄组组距。该指标与年龄组死亡率 ${}_i m_x$ 的意义完全不同,年龄组死亡率只是表示某年龄组人口平均死亡水平。但这两个指标间存在下列函数关系

$${}_i q_x = \frac{2 \times i \times {}_i m_x}{2 + i \times {}_i m_x}$$

其中的“0~”岁年龄级组死亡率,通常由公式

$$\frac{\text{同年内未满1周岁的婴儿死亡率}}{\text{某年活产总数}} \times 1000\%$$

计算的婴儿死亡率替代,本例婴儿死亡率为0.010168=10.168%。但最后一个年龄组死亡概率为1,本例 ${}_5 q_{85}=1.000000$ 。

(3) 计算尚存人数与死亡人数。

尚存人数(number of survivors)记为 l_x , 表示同时出生的一代人活到满 X 岁时的存活人数, 一般假定“0~”岁组, 即刚出生的一代为 100000。

死亡人数(number of dying)记为 ${}_i d_x$, 与年龄组死亡概率 ${}_i q_x$ 和尚存人数 l_x 有如下关系 ${}_i d_x = {}_i q_x \times l_x$ 。

(4) 计算生存人年数。

生存人年数(number of survival person-years)记为 ${}_i L_x$, 是指 X 岁尚存活者在今后 i 年内的生存人年数, 计算公式为 ${}_i L_x = \frac{i}{2}(l_x + l_{x+i})$ 。

但“0~”岁组的生存人年数应采用 ${}_1 L_0 = {}_4 l_1 + a_0 \times {}_1 d_0$ 计算, 式中的 a_0 为经验系数, 中国的 $a_0 = 0.15$, 表示“0~”岁组死亡者的平均存活年数。这是由我国 1981 年与 1982 年部分地区的婴儿死亡资料计算出来的, 男性为 0.1450, 女性为 0.1525, 取 2 位小数, 则男女统一用 0.15, 本例有

$${}_1 L_0 = {}_4 l_1 + 0.15 \times {}_1 d_0 = 98983 + 0.15 \times 1017 = 99136$$

最后一个年龄组的生存人年数应为该组的尚存人数与该组的死亡率的比值, 本例有 ${}_5 L_{85} = l_{85} / m_{85} = 7264 / 0.259042 = 28044$ 。

(5) 计算生存总人年数。

生存总人年数(total number of survival person-years)记为 T_x , 是指 x 岁尚存活者今后存活的总人年数, 它是 x 岁及以上各年龄组生存人年数的总和, 即

$$T_x = \sum {}_i L_x = T_{x+1} + {}_i L_x$$

计算时, 由该式从下向上累积求得。

(6) 计算预期寿命。

预期寿命(life expectancy)记为 e_x , 是指 x 岁尚存活者预期平均尚能存活的岁数, 即 $e_x = T_x / l_x$ 。

注意, 如果婴儿死亡率特别高, 有可能会出 $e_0 < e_1$ 的情况。

寿命表中的各项指标(如死亡概率、死亡人数、尚存人数、期望寿命先等)都是以年龄组死亡率计算出来的, 不受年龄构成的影响, 同样说明了人群的死亡水平, 具有较好的可比性。特别是预期寿命 e_x , 既能综合地反映各个年龄组的死亡水平, 又能预期寿命长短的形式说明人群的健康水平, 因此, 它是评价不同国家或不同地区社会卫生状况的主要指标之一。尤其是出生时预期寿命(即“0~”岁组预期寿命 e_0)更为常用。

在 DPS 系统中, 计算简略寿命表只要将数据按图 37-5 方式编辑、定义即可。图 37-5 是 1991 年中国人口普查资料。只要在 DPS 电子表格中第 1 列放年龄区间的起点, 末行中的年龄为 90 岁, 它对应于 90 岁及以后的区间; 第 2 列为 1991 年普查时该区间中人口数; 第 3 列为对应区间中的死亡数。

	A	B	C	D	E
1		年龄区间的起点	人口数	死亡数	
2		0	23220850	650203	
3		1	93217570	218220	
4		5	99336740	73615	
5		10	97226690	55939	
6		15	120158400	120289	
7		20	125761200	167039	
8		25	104267500	130619	
9		30	83875700	133916	
10		35	86351810	166384	
11		40	63707660	174673	
12		45	49087940	212186	
13		50	45619560	316368	
14		55	41709340	461670	
15		60	33976260	628003	
16		65	26332520	778748	
17		70	18050580	886707	
18		75	10933920	819081	
19		80	5352690	633754	
20		85	1907544	319635	
21		90	416134	101354	
22					

图 37-5 现时寿命表数据编辑格式

然后执行“生存分析”下的“现时寿命表”分析，即可得到结果如表 37-3。

表 37-3 简略寿命表计算结果

年龄组 $X(\text{岁})(1)$	年均人口 $P_x(2)$	实际死亡 $D_x(3)$	死亡率 $m(4)$	死亡概率 $q_x(5)$	尚存人数 $l_x(6)$	死亡人数 $d_x(7)$	生存人年 $L_x(8)$	总人年 $T_x(9)$	预期寿命 $e_x(10)$
0~	23220850	650203	0.0280	0.0280	100000	2800	97620	7024149	70.2415
1~	93217570	218220	0.0023	0.0093	97200	906	386988	6926529	71.2607
5~	99336740	73615	0.0007	0.0037	96294	356	480580	6539542	67.9123
10~	97226690	55939	0.0006	0.0029	95938	276	479000	6058962	63.1551
15~	120158400	120289	0.0010	0.0050	95662	478	477117	5579962	58.3298
20~	125761200	167039	0.0013	0.0066	95185	630	474348	5102845	53.6100
25~	104267500	130619	0.0013	0.0062	94555	590	471297	4628497	48.9505
30~	83875700	133916	0.0016	0.0080	93964	747	467953	4157200	44.2424
35~	86351810	166384	0.0019	0.0096	93217	894	463851	3689247	39.5770
40~	63707660	174673	0.0027	0.0136	92323	1257	458474	3225396	34.9359
45~	49087940	212186	0.0043	0.0214	91066	1947	450463	2766922	30.3836
50~	45619560	316368	0.0069	0.0341	89119	3038	438002	2316459	25.9928

55~	41709340	461670	0.0111	0.0539	86082	4636	418818	1878457	21.8218
60~	33976260	628003	0.0185	0.0883	81446	7195	389242	1459639	17.9216
65~	26332520	778748	0.0296	0.1377	74251	10224	345697	1070396	14.4159
70~	18050580	886707	0.0491	0.2188	64028	14006	285123	724699	11.3185
75~	10933920	819081	0.0749	0.3155	50021	15781	210656	439576	8.7878
80~	5352690	633754	0.1184	0.4568	34241	15641	132102	228921	6.6856
85~	1907544	319635	0.1676	0.5905	18600	10983	65543	96818	5.2053
90~	416134	101354	0.2436	1.0000	7617	7617	31275	31275	4.1057

37.3.2 现时寿命表中平均寿命的估计

现时寿命表中, 已知 t_i 表示第 i 年龄段(t_i, t_{i+1})左端点的年龄, t_{i+1} 是右端点的年龄。在每一区间上已知死亡人数 d_i , 它的实测值常记为 D_i ; 而记 P_i 表示年龄段(t_i, t_{i+1})上的普查资料中计算得的人口数, 则普查资料中首先估计年龄段(t_i, t_{i+1})上的死亡率 M_i , 它是用下式估计

$$M_i \approx \frac{D_i}{P_i}$$

其次再估计(t_i, t_{i+1})上的生存概率及死亡概率

$$p_i = \exp\{-n_i \cdot M_i\}, \quad q_i = 1 - p_i, \quad n_i = t_{i+1} - t_i$$

由此计算平均寿命如下:

$$\begin{cases} e_i = p_i e_{i+1} + q_i / M_i, \\ e_w = 1 / M_w, \end{cases} \quad i = 1, 2, \dots, w$$

其中 e_w 是最后一个年龄段 $[t_w, +\infty)$ 上 t_w 岁后仍然可存活的平均岁数。一般情况下, 取 $t_w = 80, 85$ 或 90 岁中之一, 这可视 $[t_w, +\infty)$ 上的人口数多少而定。其标准误(方差 $\text{var}()$ 的开方)为

$$\begin{aligned} \text{var}(e_i) &= p_i^2 \text{var}(e_{i+1}) + A_i^2 \text{var}(M_i), \quad i < w \\ \text{var}(e_w) &= \text{var}(M_w) / M_w^4 \end{aligned}$$

这里

$$\begin{aligned} \text{var}(M_i) &\approx (1 - q_i) M_i / P_i, \quad i = 1, 2, \dots, w \\ A_i &= \begin{cases} -1 / M_w^2, & i = w \\ -n_i p_i \left\{ e_{i+1} - \frac{1}{M_i} \right\} - \frac{q_i}{M_i^2}, & i < w \end{cases} \end{aligned}$$

上式是考察平均寿命的公式。如果我们的着重点不是考察“寿命”而是考察“生存时间”，这时可用孙尚拱提出的一个用于计算任何一个年龄段上的平均生存时间，比如要估计指定的 $[t_a, t_b]$ 年龄区间上平均生存时间的公式

$$\begin{cases} e_i = p_i e_{i+1} + q_i / M_i, & a \leq i < b \\ e_{b-1} = q_{b-1} / M_{b-1}, & i = b-1, \text{ 当 } p_{b-1} > 0 \text{ 时} \end{cases}$$

$$\begin{cases} \text{var}(e_i) = p_i^2 \text{var}(e_{i+1}) + A_i^2 \text{var}(M_i), & a \leq i \leq b-1 \\ \text{var}(e_{b-1}) = A_{b-1}^2 \text{var}(M_{b-1}), \end{cases}$$

例如，对 1991 年中国普查资料的分析。只要在 DPS 电子表格中第 1 列放年龄区间的起点，末行中的年龄为 90 岁，它对应于 90 岁及以后的区间；第 2 列为 1991 年普查时该区间中人口数；第 3 列为对应区间中的死亡数(图 37-5)。然后执行“生存分析”下的“现时寿命表分析”功能，这时系统提示输入待估计平均生存时间的年龄区间。



如我们希望限定[60 岁, 80 岁]上估计 t_i 时刻的存活者到 80 岁时的平均存活时间，则输入“60 80”，确认后得到结果如下：

T	N	Q	P	Hazard	IEI	E	Se	A
0.0000	1.0000	0.0276	0.9724	0.0280	0.9861	70.2349	0.0051	-69.7395
1.0000	4.0000	0.0093	0.9907	0.0023	3.9813	71.2152	0.0046	-276.8855
5.0000	5.0000	0.0037	0.9963	0.0007	4.9907	67.8663	0.0044	-326.8472
10.0000	5.0000	0.0029	0.9971	0.0006	4.9928	63.1090	0.0043	-303.0570
15.0000	5.0000	0.0050	0.9950	0.0010	4.9875	58.2836	0.0043	-278.9389
20.0000	5.0000	0.0066	0.9934	0.0013	4.9834	53.5635	0.0042	-255.3454
25.0000	5.0000	0.0062	0.9938	0.0013	4.9844	48.9038	0.0042	-232.0451
30.0000	5.0000	0.0080	0.9920	0.0016	4.9801	44.1954	0.0041	-208.5102
35.0000	5.0000	0.0096	0.9904	0.0019	4.9760	39.5296	0.0041	-185.1881
40.0000	5.0000	0.0136	0.9864	0.0027	4.9659	34.8881	0.0040	-161.9975
45.0000	5.0000	0.0214	0.9786	0.0043	4.9464	30.3353	0.0039	-139.2659
50.0000	5.0000	0.0341	0.9659	0.0069	4.9143	25.9436	0.0038	-117.3613
55.0000	5.0000	0.0538	0.9462	0.0111	4.8642	21.7713	0.0036	-96.5838
60.0000	5.0000	0.0883	0.9117	0.0185	4.7759	17.8692	0.0035	-77.2223
65.0000	5.0000	0.1375	0.8625	0.0296	4.6479	14.3610	0.0033	-59.8991
70.0000	5.0000	0.2178	0.7822	0.0491	4.4333	11.2610	0.0032	-44.7686
75.0000	5.0000	0.3124	0.6876	0.0749	4.1703	8.7286	0.0031	-32.5679
80.0000	5.0000	0.4468	0.5532	0.1184	3.7735	6.6293	0.0031	-22.7876
85.0000	5.0000	0.5673	0.4327	0.1676	3.3859	5.1622	0.0032	-16.1781
90.0000	0.0000	1.0000	0.0000	0.2436	4.1057	4.1057	0.0000	0.0000
限定年龄在[60, 80]区间的平均存活时间								
年龄	平均寿命	S.E.						
60.0000	15.0652	0.0035						
65.0000	11.2855	0.0033						
70.0000	7.6954	0.0032						

75.0000	4.1703	0.0031
---------	--------	--------

输出结果可以看出, $t=0$ 时, $e=70.2349$, 说明 1991 年出生的婴儿在 1991 年的死亡模式下, 平均可存活 70.2 岁; 而 1991 年时 90 岁的老人, 仍可存活(平均)4.1 岁。末列 A 的一切数全为负, 与队列寿命中 A 为正不同的原因是: 此处的自变量为 m_i (死亡率), 它与队列寿命中的 p_i (生存概率)意义反向。 A 的最大用处是可以估计死亡率 m_i 发生变化时, 如何估计生存时间的变化, 其公式为

$$\Delta e_i(\Delta m_i, \Delta m_{i+1}, L, \Delta m_w) \approx \sum_{j=i}^w p_{ij} A_j \Delta m_j$$

上式也可以写成

$$\Delta e_i(\Delta m_i, \Delta m_{i+1}, L, \Delta m_w) \approx A_i \Delta m_i + p_i \Delta e_{i+1}(\Delta m_{i+1}, L, \Delta m_w)$$

例 1 如只考虑出生时死亡率 m_0 有变化, 变化值为 Δm_0 , 而另外的死亡率都不变化, 则有公式

$$\Delta e_0(\Delta m_0) \approx A_0 \Delta m_0$$

注意, Δm_0 对 e_1 及以下的 e_i 没有影响。

例 2 如只考虑最后一个年龄段 $[t_w, +\infty)$ 上的死亡率 m_w 对前面任一个年龄上的生存时间的影响, 则有

$$\Delta e_i(\Delta m_w) \approx \frac{-p_{iw}}{m_w^2} \Delta m_w, \quad i > 0$$

其中 $p_{iw} = p_0 p_1 \cdots p_{w-1}$ 。

如果我们考察 1991 年普查资料中出生婴儿死亡率 m_0 变化 10%(即 $\Delta m_0 = 0.1m_0$), 其他不变, 则有

$$\Delta e_0(\Delta m_0) \approx A_0 \Delta m_0 = -69.739 \times 0.1 \times 0.028 = -0.195 (\text{岁})$$

即出生婴儿死亡率减少 10% 时, 可望提高出生婴儿平均寿命 0.195 年。此公式对于漏报出生死亡人数而估计对平均寿命的影响很有用。同样, 当 90 岁及以上年龄段中死亡率 m_w 减少 10% 时, 可得

$$\begin{aligned} \Delta e_0(\Delta m_w) &\approx -p_w \Delta m_w / m_w^2 = -p_0 p_1 \cdots p_{w-1} \times 0.1 / m_w \\ &= -0.97239 \times 0.99068 \times 0.9963 \times \cdots \times 0.43265 \times 0.1 / 0.24346 = -0.0339 (\text{岁}) \end{aligned}$$

另外, 当所有年龄段中, 死亡率都改变 10% 时 ($\Delta m_i = 0.1m_i$), 则可得

$$\Delta e_0(\Delta m_0, L, \Delta m_w) \approx -1.40 (\text{岁})$$

可见, 只有全面减低死亡率时, 才能对平均寿命产生大的影响。

最后的输出结果是限定[60 岁, 80 岁]上估计 t_i 时刻, 即存活者到 80 岁时的平均存活时间。

37.3.3 去死因寿命表

研究某种死因对居民死亡的影响, 可编制去死因寿命表(cause eliminated life table)。其基本思想是, 假使消除了某种死因, 则原死于该原因的人不死于该原因, 寿命就会有所延长。显然, 如果消除了对生命威胁大的死因, 寿命就会延长更多。去死因寿命表的优点是: ① 以某死因耗损预期寿命和尚存人数合理地说明了该死因对人群生命的影响程度; ② 去死因寿命表的指标既能综合说明某死因对全人口的作用, 又能分别说明某死因对各年龄组人口的作用; ③ 去死因寿命表的指标同样不受人口年龄构成的影响, 便于相互比较。

去死因寿命表中各项指标的意义与全死因寿命表相同, 为了区分, 在有关符号的右上角加上标记“'”表示某死因或去某死因。

编制去死因寿命表的关键是求去某死因后各年龄组生存概率(${}_i p'_x$), 有了 ${}_i p'_x$, 就可仿照前面介绍的编制全死因寿命表的方法来编制去某死因寿命表。

(1) 整理基本数据。数据按年龄组、年均人口、实际死亡及肿瘤死亡数格式输入(图 37-6)。

(2) 计算去肿瘤后的死亡人数数据占全死因人数的比例, 公式为

$${}_i r'_x = \frac{{}_i D_x - {}_i D'_x}{{}_i D_x}$$

(3) 计算全死因寿命表死亡概率 ${}_i q_x$ 及生存概率 ${}_i p_x = 1 - {}_i q_x$ 。

(4) 计算去肿瘤后的生存概率 ${}_i p'_x$, 这是编制去死因寿命表的关键指标, 计算公式为 ${}_i p'_x = ({}_i p_x)^{{}_i r'_x}$ 。

(5) 计算去肿瘤死亡后的尚存人数 l'_x 与死亡人数 ${}_i d'_x$ 。令 $l'_0 = 100000$, 利用 ${}_i p'_x$, l'_x , 按公式: ${}_i d'_x = (1 - {}_i p'_x) \times l'_x$ 。

计算“0~”岁组去肿瘤死亡人数为

$${}_1 d'_0 = (1 - {}_1 p'_0) \times l'_0 = (1 - 0.989877) \times 100000 = 1012$$

利用“0~”岁组的去肿瘤死亡后尚存人数与死亡人数之差, 计算“1~”岁组的去肿瘤死亡后尚存人数

$$l'_1 = l'_0 - {}_1 d'_0 = 100000 - 1012 = 98988$$

(6) 计算去肿瘤死亡后的生存人年数 ${}_i L'_x$ 。

(7) 计算去肿瘤死亡后的生存总人数。从最后一组开始, 对去肿瘤死亡后生存人年数 ${}_i L'_x$ 进行累加。

(8) 计算去肿瘤死亡后的预期寿命 e'_x 。由 ${}_i L'_x$ 与 l'_x 之比获得。如去肿瘤死亡后的出生时预期寿命

$$e'_0 = 7227641 / 100000 = 72.28$$

例如，已知某地 2000 年男性的各年龄组年均人口数、全死因死亡数及肿瘤死亡数，请用这些数据求各年龄组尚存者去肿瘤死亡后的预期寿命。在 DPS 系统中，计算去死因寿命表只要将数据按图 37-6 方式编辑、定义即可。然后执行“生存分析”下的“去死因寿命表”分析，即可得到结果如表 37-4。

编制去死因寿命表的关键是求去某死因后各年龄组生存概率($i p'_x$)，有了 $i p'_x$ ，就可仿照前面的介绍的编制全死因寿命表的方法来编制去某死因寿命表。如某地 2000 年男性因肿瘤死亡而减少的出生时预期寿命为

$$e'_0 - e_0 = 72.28 - 69.98 = 2.30(\text{岁})$$

	A	B	C	D	E	F
1		年龄组(岁) X	年平均人口数 P_x	全死因死亡数 iD_x	肿瘤死亡数 iD'_x	
2		0	43861	446	2	
3		1	98053	124	4	
4		5	100407	75	8	
5		10	16626	116	11	
6		15	199860	162	13	
7		20	249279	250	21	
8		25	194560	210	36	
9		30	141843	165	41	
10		35	99965	154	44	
11		40	95662	231	80	
12		45	95652	376	142	
13		50	85074	563	210	
14		55	69403	852	315	
15		60	51560	1129	360	
16		65	39865	1652	381	
17		70	28956	1789	248	
18		75	14000	1564	127	
19		80	4465	986	60	
20		85	1023	265	4	
21						

图 37-6 去死因寿命表数据编辑格式

表 37-4 去死因寿命表计算结果

年龄组	年均人口	实际死亡	因某疾病	死因比	死亡概率	去掉特定死亡因子后					预期寿命	
						生存概率	生存概率	尚存人数	死亡人数	生存人年		总人年
0~	43861	446	2	0.9955	0.0102	0.9898	0.9899	100000	1012	99140	7227641	72.2764
1~	98053	124	4	0.9677	0.0050	0.9950	0.9951	98988	483	394984	7128501	72.0140
5~	100407	75	8	0.8933	0.0037	0.9963	0.9967	98504	328	491701	6733517	68.3576
10~	166626	116	11	0.9052	0.0035	0.9965	0.9969	98176	309	490109	6241816	63.5777

15~	199860	162	13	0.9198	0.0040	0.9960	0.9963	97867	364	488426	5751707	58.7704
20~	249279	250	21	0.9160	0.0050	0.9950	0.9954	97503	447	486399	5263281	53.9806
25~	194560	210	36	0.8286	0.0054	0.9946	0.9955	97056	433	484199	4776882	49.2176
30~	141843	165	41	0.7515	0.0058	0.9942	0.9956	96623	421	482063	4292683	44.4270
35~	99965	154	44	0.7143	0.0077	0.9923	0.9945	96202	528	479690	3810619	39.6106
40~	95662	231	80	0.6537	0.0120	0.9880	0.9921	95674	752	476490	3330929	34.8154
45~	95652	376	142	0.6223	0.0195	0.9805	0.9878	94922	1154	471725	2854439	30.0714
50~	85074	563	210	0.6270	0.0326	0.9674	0.9795	93768	1926	464026	2382715	25.4108
55~	69403	852	315	0.6303	0.0596	0.9404	0.9620	91842	3486	450496	1918689	20.8911
60~	51560	1129	360	0.6811	0.1038	0.8962	0.9281	88356	6355	425892	1468193	16.6168
65~	39865	1652	381	0.7694	0.1877	0.8123	0.8522	82001	12123	379695	1042301	12.7109
70~	28956	1789	248	0.8614	0.2676	0.7324	0.7647	69877	16440	308285	662606	9.4824
75~	14000	1564	127	0.9188	0.4366	0.5634	0.5902	53437	21896	212444	354321	6.6307
80~	4465	986	60	0.9391	0.7114	0.2886	0.3113	31541	21723	103396	141877	4.4982
85~	1023	265	4	0.9849	1.0000	0.0000	0.0000	9818	9818	38481	38481	3.9195

也就是说, 由于肿瘤损失的出生时预期寿命为 2.30 岁, 如果肿瘤的死亡得到了有效的控制, 出生时预期寿命可增加

$$(e' - e_0)/e_0 = (72.28 - 69.98)/69.98 = 3.29\%$$

37.3.4 健康预期寿命表的编制

1964 年, Sanders 首次提出了健康预期寿命表的概念, 1971 年, Sullivan 提出了编制健康预期寿命表的方法, 该法称为 Sullivan 方法, 由于其简便易行, 目前应用广泛。

当年龄(岁) X , 死亡率 ${}_i m_x$, ADL 依赖率 ${}_i R_x$ 为已知时, 编制 Sullivan 健康寿命表, 计算死亡概率、尚存人数、死亡人数、生存人年、总人年、预期寿命这几项和前面的简略寿命表相应项相同, 这里只是年龄组距为 $i=1$, 且开始年龄不是“0~”岁组。

ADL 无依赖生存人年数的计算公式为 ${}_i h_x = {}_i L_x (1 - {}_i R_x)$, 式中 ${}_i L_x$ 为生存人年数, ${}_i R_x$ 为 ADL 依赖率。如“≥90”组的无依赖生存人年数为

$${}_1 h_{90} = {}_1 L_{90} (1 - {}_1 R_{90}) = 43421$$

从倒数第二行起, 无依赖累积生存人年数的计算公式为

$$H_x = \sum {}_i h_x = H_{x+1} + {}_i h_x$$

最后一组“≥90”的无依赖累积生存人年数为 $H_{90} = {}_1 h_{90} = 43421$ 。

健康预期寿命采用累计人年的 H_x 与尚存人数 l_x 的比值计算, 即 $E_x = H_x / l_x$ 。

利用预期寿命 e_x 与健康预期寿命 E_x 的差值可获得 x 岁的依赖预期寿命(dependent life expectancy, DLE), 即 $DLE_x = e_x - E_x$ 。

例如, 已知某地 2000 年男性 75 岁以上老年人的各年龄死亡率, 并于 2000 年 5 月采

用随机抽样方法获得了各年龄的 ADL 依赖率。在 DPS 系统中, 计算各年龄的 ADL 依赖预期寿命, 只要将数据按图 37-7 方式编辑、定义即可。

	A	B	C	D	E
1		年龄 (岁) X(1)	死亡率 i X(2)	ADL 依赖率 i x(3)	
2		75	0.036	0.104	
3		76	0.040	0.104	
4		77	0.044	0.104	
5		78	0.049	0.104	
6		79	0.056	0.104	
7		80	0.062	0.203	
8		81	0.071	0.203	
9		82	0.080	0.203	
10		83	0.089	0.203	
11		84	0.098	0.203	
12		85	0.112	0.543	
13		86	0.126	0.543	
14		87	0.134	0.543	
15		88	0.148	0.543	
16		89	0.169	0.543	
17		90	0.233	0.623	
18					

图 37-7 健康预期寿命表数据编辑格式

然后执行“生存分析”下的“预期寿命表”分析, 即可得到结果如表 37-5。

表 37-5 Sullivan 健康预期寿命表

年龄	死亡率	死亡 概率	尚存 人数	死亡 人数	生存 人年	总人年	预期 寿命	ADL	无 依 赖 生 存			
									生存 人年	累计 人年	预期 寿命	依赖 寿命
75	0.0360	0.0354	100000	3536	98232	1101397	11.0140	0.1040	88016	806604	8.0660	2.9479
76	0.0400	0.0392	96464	3783	94572	1003165	10.3994	0.1040	84737	718588	7.4493	2.9501
77	0.0440	0.0431	92681	3990	90686	908593	9.8035	0.1040	81254	633852	6.8391	2.9644
78	0.0490	0.0478	88691	4242	86570	817907	9.2220	0.1040	77566	552597	6.2306	2.9914
79	0.0560	0.0545	84449	4600	82149	731338	8.6601	0.1040	73605	475031	5.6251	3.0351
80	0.0620	0.0601	79848	4802	77447	649189	8.1303	0.2030	61726	401426	5.0274	3.1029
81	0.0710	0.0686	75047	5146	72474	571742	7.6185	0.2030	57762	339700	4.5265	3.0920
82	0.0800	0.0769	69901	5377	67212	499268	7.1425	0.2030	53568	281939	4.0334	3.1091
83	0.0890	0.0852	64524	5498	61775	432056	6.6960	0.2030	49235	228370	3.5393	3.1567
84	0.0980	0.0934	59026	5514	56269	370281	6.2732	0.2030	44846	179136	3.0349	3.2383

85	0.1120	0.1061	53512	5675	50674	314012	5.8681	0.5430	23158	134289	2.5095	3.3586
86	0.1260	0.1185	47836	5670	45001	263338	5.5050	0.5430	20566	111131	2.3232	3.1818
87	0.1340	0.1256	42166	5295	39518	218337	5.1780	0.5430	18060	90566	2.1478	3.0302
88	0.1480	0.1378	36871	5081	34330	178818	4.8499	0.5430	15689	72506	1.9665	2.8834
89	0.1690	0.1558	31790	4954	29313	144488	4.5451	0.5430	13396	56817	1.7873	2.7578
90	0.2330	1.0000	26836	26836	115175	115175	4.2918	0.6230	43421	43421	1.6180	2.6738

37.4 比例风险模型——COX 回归

生存分析的目的不仅是描述某特定人群的生存时间分布，还要研究影响生存率、生存时间的因素。前面介绍的生存率的乘积限法、寿命表法估计及生存率比较的时序检验属非参数法，可用于服从任何分布的生存时间资料，但难以用于生存率的多因素分析。若生存时间的分布符合某个特定的分布类型，如符合对数正态分布、指数分布、Weibull 分布等，则可利用特定分布函数来分析，称为参数法。若不知生存时间的分布类型，或生存时间不服从参数法所规定的分布类型，但又要进行生存率的多因素分析，例如，研究接受某种治疗的某种肿瘤病人的治疗效果(生存时间)与病人的性别、年龄、血型、生活习惯(饮酒、吸烟)、生活环境等因素的关系，而不知相应的生存时间服从什么分布。此时可考虑用 Cox 模型，模型结构为

$$h(t, X) = h_0(t) \exp(\beta' X)$$

其中， $X = \begin{pmatrix} x_1 \\ \mathbf{M} \\ x_p \end{pmatrix}$ 是协变量向量； $\beta = \begin{pmatrix} \beta_1 \\ \mathbf{M} \\ \beta_p \end{pmatrix}$ 是系数向量； $h(t, X)$ 表示时刻 t ， X 下的风险

率； $h_0(t)$ 表示时刻 $t, X=0$ 的风险率，又称为基准风险率。生存率

$$S(t, X) = \exp \left[- \int_0^t h(t, X) dt \right] = \exp \left[- \exp(\beta' X) \int_0^t h_0(t) dt \right]$$

$\exp(\beta' X)$ 描述了协变量对风险率、生存率的影响。

因为 $\chi^2 = 2H_0 \left[\ln L(\hat{\beta}_{(1)}, \hat{\beta}_{(2)}) - \ln L(\beta_{(1)0}, \hat{\beta}_{(2)0}) \right]$ ， $\frac{h(t, X)}{h_0(t)} = \exp(\beta' X)$ 与时刻 t 无关，

也就是说，只要 X 固定，无论在哪个时点 t ， X 的风险率与基准风险率之比保持不变，而且对任一时点 t ，不同的 X_1 与 X_2 的相对风险率之比 $\exp(\beta'(X_1 - X_2))$ 也与时点 t 的大小无关，故此模型也被称之为比例风险模型。

上述模型表明无论哪个协变量(或协变量组)对风险率的影响都不随时间变化，这是应用 Cox 模型时要注意的地方。

模型中参数 β_i 的实际意义：设 $X_1 = (x_1, x_2, \dots, x_p)$ ， $X_2 = (x_1+1, x_2, \dots, x_p)$ ，即 X_1 与 X_2 的区别仅在于后者的第 1 分量比前者多一个单位，则 X_1 与 X_2 的风险率之比，即相

对风险度(相对危险度)为

$$\frac{h(t, X_2)}{h(t, X_1)} = \exp(\beta_1)$$

这表明, $\exp(\beta_1)$ 是其他协变量不变, 仅 x_1 增加一个单位时两个个体的相对危险度。同样, $\exp(\beta_i)$ 是其他协变量不变, 仅 x_i 增加一个单位时两个个体的相对危险度。例如, 假设某病受吸烟(x_1)和饮酒(x_2)的影响。相应的二因子 Cox 模型为

$$h(t, X) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2)$$

据样本资料得到 β_1, β_2 的估计: $\beta_1 = 0.8755, \beta_2 = 0.5108$, 代入模型表达式得

$$\begin{aligned} h(t, X) &= h_0(t) \exp(0.8755x_1 + 0.5108x_2) \\ RH(t) &= \exp(0.8755x_1 + 0.5108x_2) \end{aligned}$$

其中基准风险率 $h_0(t)$ 表示不吸烟也不饮酒者($x_1=0, x_2=0$)在时点 t 的发病风险率。由此可估计得

(1) 吸烟但不饮酒者($x_1=1, x_2=0$)相对于不吸烟也不饮酒者发病的风险比为

$$\exp(\beta_1(1-0) + \beta_2(1-0)) = \exp(\beta_1) = \exp(0.8755) = 2.40$$

(2) 不吸烟但饮酒者($x_1=0, x_2=1$)相对于不吸烟也不饮酒者发病的风险比为

$$\exp(\beta_1(0-0) + \beta_2(1-0)) = \exp(\beta_2) = \exp(0.5108) = 2.40$$

(3) 吸烟又饮酒者($x_1=1, x_2=1$)相对于不吸烟也不饮酒者发病的风险比为

$$\exp(\beta_1(1-0) + \beta_2(1-0)) = \exp(\beta_1 + \beta_2) = \exp(0.8755 + 0.5108) = 4.00$$

一般地, 模型 $h(t, X) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_p x_p)$ 中参数 β_i 的指数 $\exp(\beta_i)$ 表示 $x_1, \dots, x_{i-1}, x_{i+1}, x_p$ 不变时, x_i 变动一个单位引起的风险率变动的倍数。

模型的参数估计. 可用最大似然法求模型中参数 β 的估计. 设共有 d 个“失败”(死亡)个体, 将 d 个“失败”时间从小到大排列

$$t_{(1)} < t_{(2)} < \dots < t_{(d)}$$

记相应的协变量为 $X_{(1)}, X_{(2)}, \dots, X_{(d)}$; 记 R_j 为观察时间不小于 $t_{(j)}$ 的个体的集合。则似然函数为

$$L(\beta) = \prod_{j=1}^d \frac{\exp(\beta' X_{(j)})}{\sum_{k \in R_j} \exp(\beta' X_k)}$$

似然方程组为

$$\frac{\partial \ln L(\beta)}{\partial \beta_i}, \quad i=1, 2, \dots, p$$

解此方程组, 所得之解就是 β 的最大似然估计。

参数检验. 设 $\beta = (\beta_1, \dots, \beta_k)$, 记 $\beta_{(1)} = (\beta_1, \dots, \beta_s)$, $\beta_{(2)} = (\beta_{s+1}, \dots, \beta_k)$ 。关于参数 $\beta_{(1)}$ 的检验可用似然比检验。检验 $H_0: \beta_{(1)} = \beta_{(1)0}$ 的似然比 χ^2 统计量为

$$\chi^2 = 2 \left[\ln L(\hat{\beta}_{(1)}, \hat{\beta}_{(2)}) - \ln L(\beta_{(1)0}, \hat{\beta}_{(2)0}) \right]$$

其中, $\hat{\beta}_{(2)0}$ 是 H_0 成立时 $\beta_{(2)}$ 的最大似然估计。 χ^2 的自由度=向量 $\beta_{(1)}$ 的维数 s 。若按样本算得的 χ^2 值大于等于自由度为 s 的 χ^2 分布的第 $1-\alpha$ 百分位数, 则 $P \leq \alpha$, 拒绝 H_0 , 否则无理由拒绝 H_0 。

例如, 为研究某种药物是否会改进急性白血病病人的预后, 延长其缓解时间, 我们将确诊病人随机给予不同的治疗。一组为用药组(传统治疗加某药), 另一组为对照组(传统治疗)。治疗前检测病人白细胞计数(wbc), 经一定时间随访, 白血病病人的缓解时间见图 37-8, 试作统计分析。

	A	B	C	D	E	F
1		缓解时间	是否截尾 (0-截尾)	用药组 1, 对照组 0	Log10(wbc)	白细胞计数 (wbc)
2		1	1	0	2.799	630
3		1	1	0	5	100000
4		2	1	0	4.910	81300
5		2	1	0	4.480	30200
6		3	1	0	4.010	10240
7		4	1	0	4.360	22910
8		4	1	0	2.423	265
9		5	1	0	3.490	3090
10		5	1	0	3.970	9330
11		6	0	1	3.204	1600
12		6	1	1	2.312	205
13		6	1	1	4.061	11500
14		6	1	1	3.279	1900
15		7	1	1	4.431	27000
16		8	1	0	3.520	3310
17		8	1	0	3.050	1122
18		8	1	0	2.322	210

19	8	1	0	3.260	1820
20	9	0	1	2.799	630
21	10	0	1	2.699	500
22	10	1	1	2.959	910
23	11	0	1	2.602	400
24	11	1	0	3.490	3090
25	11	1	0	2.121	132
26	12	1	0	1.491	31
27	12	1	0	3.061	1150
28	13	1	1	2.881	760
29	15	1	0	2.301	200
30	16	1	1	3.601	3990
31	17	0	1	2.161	145
32	17	1	0	2.949	890
33	19	0	1	2.057	114
34	20	0	1	2.013	103
35	22	1	1	2.322	210
36	22	1	0	2.732	540
37	23	1	1	2.568	370
38	23	1	0	1.968	93
39	25	0	1	1.778	60
40	32	0	1	2.204	160
41	32	0	1	2.531	340
42	34	0	1	1.477	30
43	35	0	1	1.447	28
44					

图 37-8 比例风险-Cox 模型分析编辑格式

不同治疗组急性白血病病人的缓解时间和 wbc 计数

比例风险-Cox 模型分析数据编辑格式是：各个样本按时间的先后顺序排列。数据块的第一列为时间，需从小到大，按顺序排列；第二列是相应时间的病例是否截尾的标志(0 表示截尾，1 表示失效，即死亡)，第 3 列以后是各个样本的协变量观察值。这里，将有关数据编辑、定义成图 37-8 的格式。

然后在 DPS 数据处理系统中执行“专业统计”下的生存分析，比例风险-Cox 模型分析，即可得到计算结果如下：

计算结果	当前日期 2005-8-7 18: 19: 13
Cox 比例风险回归模型分析	
描述性统计量	
变量	平均值±标准差
x1	0.5000±0.5000
x2	2.9309±0.9064
模型整体检验	

Log Likelihood = -72.2966						
检验项目	卡方值=	df	p			
-2 Ln L	144.5932	40	0.0001			
模型卡方	43.3769	2	0.0001			
变量	系数	标准误	Z 值	p 值	95% 置信区间	
x1	-1.2954	0.4221	3.0688	0.0021	-2.1227	-0.4680
x2	1.6032	0.3293	4.8687	0.0001	0.9578	2.2486
风险比及其置信区间						
变量	RR	95% 置信区间				
x1	0.2738	0.1197	0.6262			
x2	4.9689	2.6059	9.4745			
本底生存函数(协变量为均值时)						
1.0000	0.9904					
2.0000	0.9776					
3.0000	0.9653					
.....					
23.0000	0.0275					

对数似然度 $L = -72.2966$ ，模型卡方 = 43.3769，显著水平 $p = 0.0001$ 。

输出结果中，似然比卡方值描述对 $H_0: \beta_1 = \dots = \beta_p = 0$ 检验的结果，本例 $p = 2$ ，据此结果卡方 = 43.3769，显著水平 $p = 0.0001$ ，故可拒绝 H_0 。

方程系数描述协变量的作用在总变异中的比例，列出协变量的系数的估计，本例变量 x_1 的系数 β_1 估计为 -1.2954， x_2 的系数 β_2 估计为 1.6032。

各个系数的卡方值描述协变量系数的检验结果，其后的值为显著水平。

由以上结果，可得

$$h(t, X) = h_0(t) \exp(-1.2954x_1 + 1.6032x_2)$$

治疗前病人的白细胞计数与治疗方法对病人的预后均有影响。平衡治疗前病人的白细胞计数后，加药治疗相对于传统疗法的风险比为

$$\exp(\beta_1) = \exp(-1.2954) = 0.274$$

即加药治疗相对于传统疗法发生“结束缓解”的风险小，前者是后者的 0.274 倍，或者说前者是后者的 27.4%，这说明前者不容易发生“结束缓解”，即此药物能延长缓解时间，改进急性白血病人的预后。

系统最后给出了当各个协变量取均值时的生存函数理论值，即本底生存率。

37.5 指数模型

若生存时间服从指数分布，则带协变量的生存分析可用多因子指数分布模型表达：

$$f(t, X) = \lambda(X) \exp(-\lambda(X)t)$$

其中， X 是暴露因素向量(协变量向量)， $X = (1, x_1, \dots, x_p)$ ， $\lambda(X) = \exp(X\beta) = \exp(\beta_0 + \beta_1 + \dots + \beta_p)$ ，

$\beta=(\beta_0, \beta_1, \dots, \beta_p)$ 是参数向量。可以证明其分布函数为

$$F(t, X)=1-\exp(-\lambda(X)t)$$

生存函数为

$$S(t, X)=\exp(-\lambda(X)t)$$

相应的风险率函数为

$$h(t, X)=\exp(X\beta')=\exp(\beta_0+\beta_1+\dots+\beta_p)$$

由风险率函数表达式可发现，风险率的大小与生存时间 t 无关，即风险率不随生存时间变化，这是指数分布的特点，可利用此特点识别某生存时间分布是否服从指数分布。

若某生存时间服从指数分布，则其平均生存时间 $E(T)=1/(\lambda(X))$ 。

模型中参数的意义：由模型可知，不同的暴露者有不同的生存时间分布及不同的风险率函数。基准状态，即各协变量均为 0 时的风险率，这表明参数 β_0 的指数 $\exp(\beta_0)$ 描述了基准状态的风险率。

参数 β_i 的意义与 Cox 模型中一样， $\exp(\beta_i)$ 描述当 $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p$ 不变时 x_i 变动一个单位引起的风险率变动的倍数。

例如，考察 40 例严重肺癌患者的生存时间及其影响因素，其中

x_1 : 患者在诊断时的身体状况，用 0~100 的整数表示(取值越大，状况越好)；

x_2 : 患者的年龄(岁)；

x_3 : 患者从诊断到进入治疗所经历的时间(月)；

x_4, x_5, x_6 : 描述癌细胞的种类：

$x_4=1, x_5=0, x_6=0$: 癌细胞是扁平的；

$x_4=0, x_5=1, x_6=0$: 癌细胞是小的；

$x_4=0, x_5=0, x_6=1$: 癌细胞是腺样的；

$x_4=0, x_5=1, x_6=0$: 癌细胞是大的；

x_7 : 治疗方法， $x_7=1$ 表示标准方法， $x_7=0$ 表示用新试验的方法。

在 DPS 数据处理平台上进行指数模型分析数据编辑格式：各个样本按时间的先后顺序排列。数据块的第一列为时间，以从小到大顺序排列；第二列是相应时间的病例是否截尾的标志(0 表示截尾，1 表示失效，即死亡)，第三列以后是各个样本的协变量观察值。这里，将有关数据编辑定义成图 37-9 格式。

然后在 DPS 数据处理系统中执行“专业统计”下的生存分析→指数模型分析，即可得到计算结果如下：

计算结果		当前日期 02-1-31 15: 36: 37					
	方程系数	标准误	OR	Z	显著水平	95%置信区间	
β_1	-0.052106	0.010818	0.949228	-4.81672	0.000002	-0.073311	-0.030901
β_2	-0.020501	0.019345	0.979707	-1.059787	0.289241	-0.058421	0.017418
β_3	-0.000994	0.011801	0.999007	-0.084218	0.932883	-0.024126	0.022138
β_4	-0.103971	0.443495	0.901251	-0.234436	0.814646	-0.973300	0.765357
β_5	0.320940	0.481631	1.378423	0.666360	0.505181	-0.623142	1.265022

β_6	1.166124	0.564162	3.20953	2.067003	0.038734	0.060268	2.271981
β_7	0.503663	0.372975	1.654772	1.350393	0.176890	-0.227434	1.234760
β_0	-0.944074	1.372909		-0.687645	0.491677	-3.63522	1.747069
对数似然度 L=	-57.2714						
检验项目	卡方值	df	p				
-2Ln L	114.5428	31	0.000000				
模型卡方	38.9909	7	0.000002				
Lambda=5.0537							

	A	B	C	D	E	F	G	H	I	J	K
1		生存	截尾	协变量							
2		时间	(0 截尾)	x1	x2	x3	x4	x5	x6	x7	
3		1	1	20	65	21	1	0	0	0	
4		2	1	40	44	36	0	1	0	0	
5		8	1	40	63	58	1	0	0	1	
6		8	1	20	61	19	0	0	1	1	
7		10	1	40	67	23	0	1	0	1	
8		11	1	70	48	11	1	0	0	1	
9		12	1	50	63	4	0	0	1	1	
10		12	1	40	68	12	0	0	0	1	
11		15	1	50	40	13	1	0	0	0	
12		16	1	30	53	4	0	1	0	1	
13		18	1	40	69	5	0	0	1	0	
14		19	1	30	39	4	0	0	0	0	
15		20	1	30	54	9	0	1	0	0	
16		21	1	40	55	2	0	1	0	1	
17		25	0	70	48	9	1	0	0	1	
18		43	1	60	49	11	0	0	0	0	
19		44	1	60	70	13	1	0	0	0	
20		51	1	30	59	87	0	1	0	0	
21		54	1	80	63	4	0	1	0	1	
22		56	1	80	43	12	0	1	0	1	
23		84	1	80	62	4	0	0	1	0	
24		90	1	60	50	22	0	0	1	0	
25		92	1	40	69	10	1	0	0	1	
26		100	1	60	37	13	0	0	0	1	
27		103	0	70	36	22	0	1	0	0	
28		118	1	70	65	11	1	0	0	1	
29		126	1	60	63	9	1	0	0	1	
30		153	1	60	63	14	0	1	0	1	
31		164	1	40	68	15	0	0	0	0	
32		177	1	50	66	16	0	0	0	1	
33		200	1	80	41	12	0	0	0	1	
34		201	1	80	52	28	1	0	0	0	
35		231	0	50	52	8	1	0	0	0	
36		231	1	70	64	18	0	0	0	0	
37		250	1	70	53	8	0	0	0	1	
38		287	1	60	66	25	0	1	0	1	
39		340	1	80	64	10	0	0	0	0	
40		411	1	70	64	5	1	0	0	1	
41		991	1	70	50	7	1	0	0	0	
42		999	1	90	54	12	1	0	0	0	
43											

图 37-9 指数模型分析数据编辑定义格式

输出结果中, 模型卡方描述对 $H_0: \beta_1 = \dots = \beta_p = 0$ 检验的结果, 本例 $p=7$, 据此结果模型卡方=38.9909, 即相应 p 值等于 0.0000, 故可拒绝 H_0 。

方程系数列出协变量的系数的估计, 本例变量 x_1 的系数 β_1 估计为 -0.0521, x_2 的系数 β_2 估计为 -0.0205, \dots , β_7 估计为 0.5037, β_0 估计为 -0.9441。

各协变量的检验结果, 其后的值为显著水平及 95% 置信区间。这里, 有统计意义的两个协变量是 x_1 (患者在诊断时的身体状况), x_6 (癌细胞是否是腺样的)。相应的参数估计 $\beta_1 = -0.052$, $\beta_6 = 1.166$ 。这表明在其他因素保持不变时, x_1 增加 1 个单位, 即病人在诊断时身体状况好 1 分, 病人死亡的风险率将缩为原来的 $\exp(-0.0521) = 0.949$ 倍; 在其他因素不变时, 癌细胞是腺样的患者的风险率是大癌细胞患者的风险率的 $\exp(1.166) = 3.209$ 倍。

据上述输出结果还可以估计在某暴露 x 下患者的平均生存时间。例如, 某患者在诊断时身体状况 $x_1 = 50$, 年龄 $x_2 = 60$, 从诊断到进入治疗所经时间 $x_3 = 5$ (月), 癌细胞是大的 (即 $x_4 = x_5 = x_6 = 0$), 用新试验方法治疗 ($x_7 = 0$), 则其相应的 $X_1 = (50, 60, 5, 0, 0, 0, 0)$, 估计这种暴露患者的平均生存时间为

$$\frac{1}{\exp(-0.944 + (-0.052 \times 50) + (-0.021 \times 60) + (-0.001 \times 5) + 0.104 \times 0 + 0.321 \times 0 + 1.166 \times 0 + 0.504 \times 0)}$$

$$= 1/\exp(-4.809) = 1/0.008 = 122.609(\text{月})$$

若暴露 $X_2 = (50, 60, 5, 0, 0, 1, 0)$, 估计相应的平均生存时间为

$$\frac{1}{\exp(-0.944 + (-0.052 \times 50) + (-0.021 \times 60) + (-0.001 \times 5) + 0.104 \times 0 + 0.321 \times 0 + 1.166 \times 1 + 0.504 \times 0)}$$

$$= 1/\exp(-3.643) = 1/0.026 = 38.206(\text{月})$$

暴露 X_1 与暴露 X_2 的区别仅在于前者的癌细胞是大的, 后者的癌细胞是腺样的。腺样癌细胞患者的死亡风险率是大癌细胞患者的 3.209 倍, 平均生存时间是前者的 $(38.206/122.609) = 31.16\%$ (即是前者的 3.209 分之一)。

输出结果的最后 1 个参数给出了不考虑协变量作用时生存数据指数分布参数 λ 的极大似然估计。

37.6 Weibull 模型

Weibull 分布模型结构为 $F(t, X) = (\lambda(X)(1/\delta)t)^{(1/\delta)-1} \exp(-\lambda(X)t^{(1/\delta)})$, 其中 $X = (x_1, \dots, x_p)$ 是协变量向量, $\lambda(X) = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$, $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ 是与协变量有关的参数向量, δ 是形状参数。可以证明

分布函数为 $F(t, X) = 1 - \exp[-\lambda(X)t^{(1/\delta)}]$;

生存函数为 $S(t, X) = \exp[-\lambda(X)t^{(1/\delta)}]$;

风险率函数为 $h(t, X) = (-\lambda(X))(1/\delta)t^{(1/\delta)-1}$ 。

模型中参数的意义：模型中共有 $p+2$ 个参数，其中 δ 是形状参数，描述分布的形状，且决定风险率随时间变化的速度， $\delta < 1$ 时，风险率随时间递增， $\delta > 1$ 时，风险率随时间递减。模型中参数 β_i 的意义与指数分布模型相同。

应用前面指数模型分析 40 例严重肺癌患者的生存时间及其影响因素资料，数据整理和数据块定义方式均和拟合指数模型相同(图 37-9)。在 DPS 数据处理系统下拟合 Weibull 分布模型，得到结果如下：

	方程系数	标准误	OR	Z 值	显著水平	95%置信区间	
β_1	-0.059156	0.011003	0.94	-5.37649	0.000000	-0.080772	-0.037539
β_2	-0.024556	0.019550	0.97	-1.256064	0.209092	-0.062965	0.013853
β_3	-0.001718	0.012055	0.99	-0.142484	0.886698	-0.025401	0.021966
β_4	-0.149725	0.446820	0.86	-0.335091	0.737556	-1.027589	0.728139
β_5	0.370254	0.483209	1.45	0.766239	0.443534	-0.579105	1.319613
β_6	1.341298	0.564521	3.82	2.375992	0.017502	0.232186	2.450410
β_7	0.574002	0.377545	1.77	1.520353	0.128422	-0.167759	1.315762
β_0	-1.042551	1.386604		-0.751874	0.452127	-3.76681	1.681702
δ	0.877916	0.070272					
对数似然度 L=		-56.8154					
检验项目		卡方值	df	p			
-2Ln L		113.6307	30	0.000000			
模型卡方		31.58281	7	0.000049			
Lambda=		3.4964					
Sigma=		1.3897					

输出结果中，模型卡方描述对 $H_0: \beta_1 = \Lambda = \beta_p = 0$ 检验的结果，本例 $p=7$ ，据此结果模型卡方=31.5828，即相应 p 值等于 0.0000，故可拒绝 H_0 。方程系数描述协变量的作用在总变异中的比例，列出协变量的系数的估计，本例变量 x_1 的系数 β_1 估计为 -0.0592， x_2 的系数 β_2 估计为 -0.0246， \dots ， β_7 估计为 0.5740， β_0 估计为 -1.0426。

各系数 t 描述协变量系数的检验结果，其后的值为显著水平，95%置信区间。本例仅两个 P 值均小于 0.05，故仅两个协变量(x_1, x_6)在 $\alpha = 0.05$ 水平有统计意义。

有统计意义的两个协变量是 x_1 (患者在诊断时的身体状况)， x_6 (癌细胞是否是腺样的)。相应的参数估计 $\beta_1 = -0.0592$ ， $\beta_6 = 1.3413$ 。这表明在其他因素保持不变时， x_1 增加 1 个单位，即病人在诊断时身体状况好 1 分，病人死亡的风险率将缩为原来的 $\exp(-0.0592) = 0.943$ 倍。在其他因素不变时，癌细胞是腺样的患者的风险率是大癌细胞患者风险率的 $\exp(1.3413) = 3.824$ 倍。

输出结果中，参数 $\delta = 0.8779$ ，根据估计形状参数 $\delta < 1$ ，说明随时间的增加，死亡的危险也增加。

输出结果的最后 2 个参数给出了不考虑协变量作用时生存数据 Weibull 分布参

数 λ 和 δ 的极大似然估计。

附 人口金字塔图的绘制

绘制人口金字塔(population pyramid), 其数据编辑格式如图 37-10 阴影部分所示。该功能菜单也放在生存分析里面。用 DPS 绘制人口金字塔图, 其横轴标注可放在数据块第一列的上面。这时标注是中文, 那么图中给出的标注为“男”、“女”; 如果横轴标注是纯英文, 那么图中给出的标注为英文“Males”和“Females”。图 37-10 右边是输出示例。

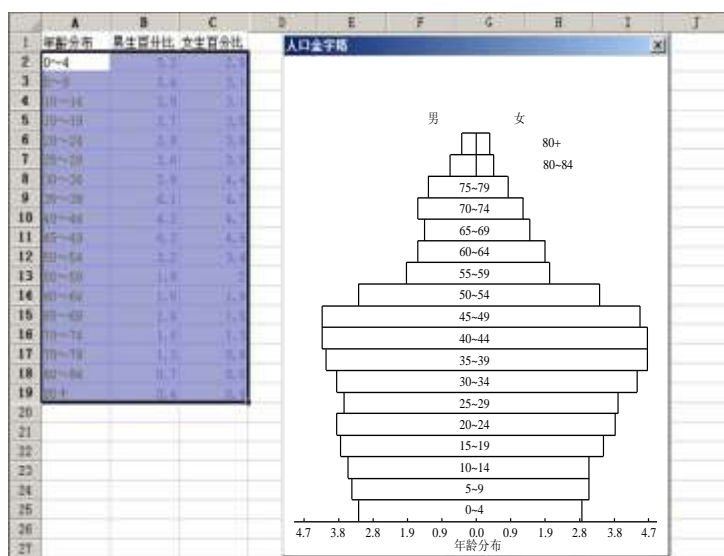


图 37-10 人口金字塔数据编辑格式及图形输出示例

参 考 文 献

- 曹素华等. 1998. 实用医学多因素统计方法[M]. 上海: 上海医科大学出版社.
- 金丕焕等. 1993. 医用统计方法[M]. 上海: 上海医科大学出版社.
- 茆诗松. 2003. 统计手册[M]. 北京: 科学出版社.
- 孙尚拱. 1990. 实用多变量统计方法与计算程序[M]. 北京: 北京医科大学、中国协和医科大学联合出版社.
- 孙尚拱. 2000. 医学多变量统计与统计软件[M]. 北京: 北京医科大学出版社.
- 孙振球. 2002. 医学统计学[M]. 北京: 人民卫生出版社.
- 周仁郁. 2004. 中医药统计学[M]. 北京: 中国中医药出版社.
- Elisa T Lee. 1997. 生存数据分析的统计方法[M]. 陈家鼎, 戴中维译. 北京: 中国统计出版社.
- Lawless J F. 1997. 寿命数据中的统计模型与方法[M]. 茆诗松译. 北京: 中国统计出版社.
- Rosner B. 2004. 生物统计学基础[M]. 孙尚拱译. 北京: 科学出版社.
- Sun Shang-Gong. 2000. Estimation of life expectancy I. Cohort life table. *Mathematical Population Studies*[J], 8(4): 1-20.

第 38 章 数学生态学方法

生物种群空间分布型常因种类和发育阶段的差异而不同，亦随种群密度的大小而有所变化，同时还受地形、土壤和气候等环境因素的影响。了解不同生物种群空间分布型的差异，不仅可以认识它的生活习性和对环境的适应性，还可以根据不同分布型进行调查取样及有关生物学试验的设计。研究生物种群空间分布型的聚集度指标始见于 20 世纪 50 年代后期，近 20 多年来发展迅速。它既可用于判断种群的空间分布型，又可为了解种群群体行为、种群扩散型的时间序列变化等提供有用的信息。

群落生态学家在解释和测度动植物群落特征特性时面对的是一类特殊的统计问题。有些群落研究，如能量分析，只需要使用一般的统计学方法、原理，就可以对群落内物种的丰富度进行估计。但是，有些群落研究，需要使用一些新的参数，在群落水平上来度量各个群落之间的相似性(similarity)。DPS 系统提供了常用的群落相似性参数估计，包括二项分布系数和距离系数。前者可根据定性或定量调查资料，通过计算各个生物群落间的相似系数和不相似(距离)系数来描述各个样本间的相似性测度；后者则根据定量资料，对群落的相似性进行测度。

植物生态学家为分析群落格局与环境关系，使用了大量的多变量统计分析技术。如聚类分析、主成分分析等。这些分析技术将在后面章节中介绍。本章仅介绍极点排序分析技术和对排序效果进行检验的方法。

群落物种多样性(diversity)是群落最直观的特征。目前描述群落特征、比较群落之间的差异时，一般都是使用多样性指数，因此多样性指数在当今群落生态学中的应用十分广泛。多年来，为检验群落的异质性所设计的多样性指数沿着两个不同途径发展：一是应用统计抽样理论方法观察分析群落结构如何发生发展变化的，如对数序列模型(Fisher et al., 1943)和对数正态分布模型(Preston, 1948)；二是应用信息理论来评价群落结构的多样性，如 Simpson 多样性指数、Shannon 多样性指数。这两种方法的可应用性一直处在争论之中，但都由于理论依据不够充分而谁也说服不了谁。因此两大类方法目前均在应用，我们在此也一并进行介绍。

自从 MacArthur(1967)开拓性的工作以来，生态位理论目前已是分析群落结构最有力的方法之一。群落结构的分析以及群落内竞争物种之间动态的互作的描述都离不开物种生态位参数的度量。在 DPS 数据处理平台上，我们提供了生态位宽度(niche breadth)和生态位重叠(niche overlap)指标的估计技术。

38.1 种群空间分布型聚集度指标测定

38.1.1 聚集度测定的指标

用于病虫空间分布型的聚集度测定的指标有多种，现归纳介绍如下：

1. 平均拥挤度 m^* (Lloyd, 1967)

通俗地讲, 平均拥挤度表示生物个体在一个样方中的平均邻居数, 它反映了样方内生物个体的拥挤程度,

$$m^* = \frac{\sum_{j=1}^{\theta} x_j(x_{j-1})}{\sum_{j=1}^{\theta} x_j} = \bar{x} + S^2/\bar{x} - 1$$

式中 x_j 为第 j 个样方的个体数, θ 为样方总数, \bar{x} 为平均密度(有时也记为 m), S^2 为样本方差。

2. I 指标

$$I = S^2/\bar{x} - 1$$

当 $I < 0$ 时为均匀分布, 当 $I = 0$ 时为随机分布, 当 $I > 0$ 时为聚集分布(Moore, 1954)。

3. m^*/m 指标

在 Moore(1954)的 m^* 指标的基础上, Lloyd(1967)提出了 m^*/m 指标, 即平均拥挤度与其平均值之比, 即

$$m^*/m = m^*/\bar{x}$$

当 $m^*/m < 1$ 时为均匀分布; 当 $m^*/m = 1$ 时为随机分布; 当 $m^*/m > 1$ 时为聚集分布。

4. C_A 指标

$$C_A = (S^2/\bar{x} - 1)/\bar{x}$$

Kuno(1968)最早提出并认为, 当 $C_A < 0$ 时为均匀分布, 当 $C_A = 0$ 时为随机分布, 当 $C_A > 0$ 时为聚集分布。

5. 扩散系数 C

该指标 $C = S^2/\bar{x}$ 用于检验种群是否偏离随机型。当 $C < 1$ 时为均匀分布, 当 $C = 1$ 时为随机分布, $C > 1$ 时为聚集分布。

6. 负二项分布中的 K 指标

在负二项分布中, $K = \bar{x}^2/(S^2 - \bar{x})$ 。当 $K < 0$ 时为均匀分布, 当 $K \rightarrow +\infty$ 时为随机分布, 当 $K > 0$ 时为聚集分布。

7. m^*-m 回归分析法

Iwao(1968, 1971, 1976)建立了 m^*-m 回归式 $m^* = \alpha + \beta\bar{x}$, 用于研究 m^* 与平均值之间的相关关系。式中 α 为分布的基本成分按大小分布的平均拥挤度: 当 $\alpha = 0$ 时, 分布的基本成分为单个个体; 当 $\alpha > 0$ 时, 个体间相互吸引, 分布的基本成分为个体群; 当 $\alpha < 0$

时，个体之间相互排斥。式中 β 为基本成分的空间分布图式：当 $\beta < 1$ 时，为均匀分布；当 $\beta = 1$ 时，为随机分布；当 $\beta > 1$ 时，为聚集分布。

8. 幂法则

Taylor(1961)在大量生物种群资料的统计分析中，发现样本平均数与方差的对数值之间存在着以下很有意义的回归关系 $\lg S^2 = \lg a + b \lg \bar{x}$ ，即乘幂函数关系 $S^2 = a \bar{x}^b$ 。在幂法则中，当 $b \rightarrow 0$ 时为均匀分布， $b = 1$ 时为随机分布， $b > 1$ 时为聚集分布。

38.1.2 试验数据资料整理

根据田间生物种群的实际观测资料计算各个田块的平均值和方差。在数据编辑器中，每块田的数据放一行，每一行中放平均值和方差两个数据，然后进入菜单操作，选择相应功能后按回车键即可。应用示例如图 38-1。

	A	B	C	D	E	F
1	田号	调查丛数	幼虫数	虫/m2	方差	
2	1	2000	294	0.147	0.519	
3	2	2000	501	0.251	1.115	
4	3	2000	818	0.409	1.708	
5	4	2000	1041	0.521	2.343	
6	5	2000	1631	0.807	3.291	
7	6	2000	1688	0.834	3.682	
8	7	2000	2698	1.346	7.859	
9	8	2000	2827	1.414	7.094	
10	9	2000	2939	1.47	5.653	
11	10	2000	3461	1.731	8.494	
12	11	2000	3891	1.946	10.693	
13	12	2000	6264	3.132	14.218	
14						

图 38-1 聚集度指标分析数据编辑、定义格式

定义数据块后，选择功能菜单“空间分布型”中的“聚集度指标”分析，得到计算结果如下：

各项聚集度指标						
No.	拥挤度 M*	I 指标	M*/M 指标	Ca 指标	扩散系数 C	K 指标
1	2.6776	2.5306	18.2150	17.2150	3.5306	0.0581
2	3.6932	3.4422	14.7141	13.7141	4.4422	0.0729
3	3.5850	3.1760	8.7654	7.7654	4.1760	0.1288
4	4.0181	3.4971	7.7123	6.7123	4.4971	0.1490
5	3.8851	3.0781	4.8142	3.8142	4.0781	0.2622
6	4.2489	3.4149	5.0946	4.0946	4.4149	0.2442
7	6.1848	4.8388	4.5949	3.5949	5.8388	0.2782
8	5.4310	4.0170	3.8409	2.8409	5.0170	0.3520

9	4.3156	2.8456	2.9358	1.9358	3.8456	0.5166
10	5.6380	3.9070	3.2571	2.2571	4.9070	0.4431
11	6.4409	4.4949	3.3098	2.3098	5.4949	0.4329
12	6.6716	3.5396	2.1301	1.1301	4.5396	0.8848
M*-M 回归分析法(IWAO)						
M*=3.17225+1.33657M R=0.8870						
TAYLOR 幂法则						
lg(v)=0.66152+1.09202*lg(M) R=0.9931						

分析结果有 m^*-m 回归式(IWAO)为 $m^*=3.17225+1.33657m$, $R=0.8870$; 应用 Taylor 幂法则, 得 $lg(v)=0.66152+1.09202*lg(m)$, $R=0.9931$ 。从各个样地分析结果可以看出, 上述两式的回归系数均大于 1, 且各个田块的 m^*/m 指标大于 1, C_A 指标 > 0 , 扩散系数 $C > 1$, 负二项分布 K 值 > 0 时, 因此, 可以推断稻负泥虫空间分布均表现为聚集分布。

38.2 种群空间分布型-频次分布检验

38.2.1 方法简介

种群空间分布型-频次分布检验是一种常用的病虫害空间分布型检测方法。使用该方法时, 首先根据各个分布型的理论概率分布公式计算出观察样本的理论频次, 再用卡方统计量检验各种分布型理论假设总体 X 的分布函数 $F(x)$ 。若 x_1, x_2, \dots, x_n 为其样本观察值, 为了检验 $F(x)$ 是否与预先给定的分布函数 $F_0(x)$ 相同, 可以检验假设 $H_0: F(x)=F_0(x)$, $H_1: F(x) \neq F_0(x)$ 。下面给出卡方检验基本原理与步骤:

步骤 1. 根据样本的频次分布情况分成 s 个区间即 $(-\infty, a_1), [a_1, a_2], \dots, [a_{s-1}, \infty]$, 用 V_i 表示样本落在这些区间的频数, 一般希望 $V_i \geq 5 (i=1, 2, \dots, s)$, 若满足不了这个条件, 可将相邻的区间适当合并(有时可放松至 $V_i > 2$)。

步骤 2. 若分布函数 $F_0(x)$ 中有 m 个未知参数 $(0 \leq m < s)$, 则用样本估计它们, 再用估计值代入分布函数之中。

步骤 3. 在 H_0 下计算理论概率

$$p_i = P(a_{i-1} < X \leq a_i) = F_0(a_i) - F_0(a_{i-1}), \quad i = 1, 2, \dots, s$$

其中 $a_0 = -\infty, a_s = \infty$, 并计算理论频数 np_i 。

步骤 4. 计算卡方统计量

$$\chi^2 = \sum_{i=1}^s \frac{(v_i - np_i)^2}{np_i}$$

当 n 充分大时 ($n \geq 50$), 则不论总体为何种分布, 当原假设成立时, 统计量 χ^2 总是近似地服从自由度为 $s-m-1$ 的 χ^2 分布。

步骤 5. 对给定 a , 在 χ^2 分布表中查得 $\chi^2_{s-m-1}(a)$, 若 $\chi^2 > \chi^2_{s-m-1}(a)$, 否定原假设, 相反则接受原假设。理论频次和实测频次间的适合度, 凡是适合者(即差异不显著者)可判

断为实测样本属于该种分布类型。在 DPS 平台上有下列几种空间分布型可供检验。

1. 二项分布

随机均匀分布：种群中个体间相互独立，因而在空间分布上是均匀的。其概率公式为

$$P_r = \binom{n}{r} p^r q^{n-r}, \quad r=0, 1, 2, \dots, n$$

式中 p 为样本中个体出现的概率， q 为不出现的概率。均值 $\bar{x}=np$ ，方差 $S^2=npq$ ，其理论频次公式为 $NP_r = N \binom{n}{r} p^r q^{n-r}$ 。

2. Poisson 分布

Poisson 分布是由法国数学家 Poisson 在研究二项分布尾部概率的近似计算时提出的，常用于研究单位时间(空间)内某现象发生次数的分布。例如，研究细菌、血细胞、粉尘等在单位面积或容积内计数结果的分布；单位空间内某种野生动物或昆虫数的分布；在特定人群中某种非传染性罕见病的患病数或死亡数的分布。该分布用于描述种群的随机分布，它不同于均匀分布。该分布的特征是种群中的任何个体占据空间任何一点的概率是相等的，并且一个个体的存在决不影响其他个体的分布。由于 Poisson 分布的均值和方差相等，故概率公式仅一个参数 m (总体平均数)，有概率 $P_r = e^{-m} \frac{m^r}{r!}$ ，理论频次

$$NP_r = Ne^{-m} \frac{m^r}{r!}。$$

3. 负二项分布(嵌纹分布)

服从负二项分布的生物种群在空间结构上呈聚集分布，特点是呈疏密相间的极不均匀状或嵌纹状。负二项分布的概率公式为

$$\begin{aligned} P_r &= (-1)^r \binom{-k}{r} p^r q^{-k-r} \\ &= \frac{(k+r-1)!}{r!(k-1)!} p^r q^{-k-r} \end{aligned}$$

其理论频次

$$NP_r = N \frac{(k+r-1)!}{r!(k-1)!} p^r q^{-k-r}$$

式中 p ， q ， k 均为参数，可用矩法、零频率法和最大似然法给予估计。一般用最大似然法估计出来的参数比较精确。

4. 核心分布(Neyman 分布)

Neyman 分布是 Poisson 分布的特例, 即由 Poisson 分布的群所组成, 也称 Poisson-Poisson 分布。本分布的个体群之间是随机的, 个体群大小约相等, 个体群周围呈放射状蔓延, 其方差与均数之比通常在 1.5~3。

Neyman 分布的最大特点是 n 为参数, 即由于 n 的不同, 其理论值可能有很大的不同。Neyman 在 1939 年第一次提出本种分布时, 仅提出 $n=0$, $n=1$ 和 $n=2$ 的分布公式, 即 Neyman 分布的 A 型、B 型和 C 型。Beall 等于 1953 年将本分布进行推广而提出 Neyman 分布理论公式的通式, 即 $n=0, 1, 2, \dots, +\infty$, 这样就更增加了本分布型理论公式的适用范围。但一般是应用 Neyman A 型分布, 因此, 有些本为 Neyman 分布的抽样资料, 也不能适合, 这限制了本种分布的应用。

为了使本分布更广泛地在科研上应用, DPS 提供了不同 n 值的计算方法, 以适应各种呈核心分布的生物种群空间分布疏密不一, 形成一个个核心, 核心周围呈放射状扩散的空间分布情形。其理论分布为

$$NP_0 = Ne^{-m_1} e^{m_1 F(0)}$$

$$NP_{r+1} = N \frac{m_1}{r+1} \sum_{k=0}^r F_k p_{r-k}$$

式中, $F_k = f_{(0)}^{(k+1)} / k!$, $f(0) = n!(-m_2)^{-n} \left\{ e^{-m_2} - \sum_{i=0}^{n-1} \frac{(-m_2)^i}{S!} \right\}$, 且由于式中的 n 不同, 且

衍生为不同类型的 Neyman 分布。

5. Poisson 二项分布

该分布型是随着 n 的不同而得到各种 n 型的 Poisson 二项分布。当 $n \rightarrow \infty$ 时, 它趋于 Neyman 分布; 在 $0 < n < 1$ 范围内, 当 $n \rightarrow 0$ 时, 则趋于负二项分布。因此, 这种分布正处于 Neyman 分布和负二项分布之间。该分布的理论公式为

$$F_0 = e^{-a(1-q^n)}, \quad \text{当 } r=0 \text{ 时}$$

$$F_r = \bar{x} \sum_{i=0}^{r-1} \binom{r-1}{i} (n-1)^{[r-i-1]} p^{(r-i-1)} q^{(n-r+i)} F_i, \quad \text{当 } r>0 \text{ 时}$$

式中, $(n-1)^{[0]}=1$, $(n-1)^{[1]}=n-1$, $(n-1)^{[2]}=(n-1)(n-2)$, $(n-1)^{[3]}=(n-1)(n-2)(n-3)$, \dots , $(n-1)^{[r]}=(n-1)(n-2)\dots(n-r)$, 故含 r 个个体的样本出现的概率为 $P_r = F_r / r!$ 。其中, μ_1^n 可用 $\bar{x} = (\sum x_i) / n$ 估计, 参数 a, p, q 可用 $\hat{a}, \hat{p}, \hat{q}$ 来估计, 而 $\hat{a} = (n-1)\bar{x}^2 / n(S^2 - \bar{x})$, $\hat{p} = (S^2 - \bar{x}) / (n-1)\bar{x}$, $\hat{q} = (1 - \hat{p})$ 。

一般 n 的选择必须使 $\hat{p} < 1$, n 通常位于 2~4。

6. 复合 Poisson 分布

这种分布型最初由 Thomas 发现。假设在空间有许多随机分布小区，在各小区上还有随机的点分布，调查空间小区上分割开的点数，可得到下面两个 Poisson 分布的组合： $(e^{-m}m^n)/n!$ 和 $e^{-\lambda}\lambda^y/y!$ 。首先考虑总体中小区的分布，再考虑小区内个体的分布，由此可得到每样方出现 k 个个体的理论概率 p_k 为

$$p_k = \sum_{r=1}^k \frac{m^r e^{-m}}{r^2} \cdot \frac{(r\lambda)^{k-r} e^{-r\lambda}}{(k-r)!}$$

式中 r 个集群中有 a_1, a_2, \dots, a_r 个个体，并且 $a_1+a_2+\dots+a_r=k$ 。

该分布型有 m 和 λ 两个参数，设 0 个个体的样方频率为 n_0 ，1 个个体的样方频率为 n_1 ，即有 $e^{-m} = n_0/n$ ， $me^{-m} = n_1/n_0$ 。各自取自然对数，得 $m = -\ln(n_0/n_1)$ ， $\lambda = \ln m - \ln(n_1/n_0)$ 。

当 $\lambda = 0$ 时，该分布就是参数为 m 的 Poisson 分布。

卡方检验时，复合 Poisson 分布的自由度为 $n-3$ 。

38.2.2 试验数据资料分析整理

在 DPS 平台上，用户既可对经整理后的频次分布数据进行分析，又可对田间原始调查数据(如各样方的害虫种群数)进行分析。在编辑数据时，如果是经整理后的频次分布资料，编辑的格式是每行放两个数据，即每样方数量(x)和相应的频次数(f)。按样方数量从小到大的顺序，即从 0, 1, 2, ... 的顺序依次输入(图 38-2)。

	A	B	C	D
1		每样方虫量(x)	频次(f)	
2		0	493	
3		1	354	
4		2	236	
5		3	146	
6		4	88	
7		5	57	
8		6	33	
9		7	20	
10		8	11	
11		9	7	
12				

图 38-2 频次数据编辑整理格式

在菜单方式下，执行“专业统计”→“空间分布型”→“频次数据空间分布型”功能项后，系统给出空间分布型选项的用户界面(图 38-3)。在该用户界面支持下，用户可根据拟合图形以及卡方值及其概率水平(p 值越大越好)对几个待选的空间分布型进行比较，确定选择哪种分布型来拟合试验数据。

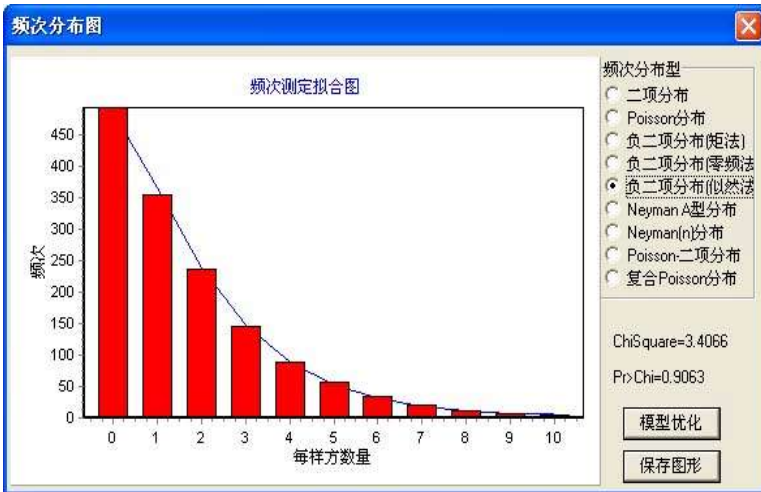


图 38-3 拟合空间分布型的用户对话框

在本例中，显然，采用似然法估计的负二项分布的拟合效果最好，其计算结果如下：

样方均值=1.6770		样本方差=3.5973	
负二项分布频次检验：			
最大或然法估计参数 K:			
$K_0=$	1.4646		
$K=$	1.3557		
每样方数	观察频数	理论频数	卡方值
0	493.0000	486.4332	0.0886
1	354.0000	364.6631	0.3118
2	236.0000	237.5130	0.0096
3	146.0000	146.9113	0.0057
4	88.0000	88.4625	0.0024
5	57.0000	52.3976	0.4043
6	33.0000	30.6923	0.1735
7	20.0000	17.8345	0.2629
8	11.0000	10.3006	0.0475
9	7.0000	5.9211	0.1966
10	4.0000	7.8708	1.9037
理论频次总和=1449.0000		卡方值=3.4066	$p=0.9063$

如果是田间原始调查数据(如各样方的害虫种群数)，分析前数据按图 38-4 格式，依次输入各样方的观察数据(x)。

然后在菜单方式下，执行“专业统计”→“空间分布型”→“原始数据空间分布型”功能项后，系统给出类似图 38-4 空间分布型选项的用户界面。在该用户界面支持下，用户可根据拟合图形以及卡方值及其概率水平(p 值越大越好)对几个待选的空间分布型进行比较，确定选择哪种分布型来拟合试验数据。分析结束系统输出计算结果如下：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1																				
2		5	4	1	1	0	0	0	2	0	0	0	0	0	0	0	1	0	1	
3		0	0	0	4	1	0	0	0	0	0	1	0	0	0	1	0	2	0	
4		0	1	0	0	0	0	1	0	1	0	0	0	0	2	0	0	0	0	
5		0	1	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	
6		0	1	0	0	1	1	2	1	0	0	0	0	1	0	1	0	0	1	
7		0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	2	1	
8		0	0	0	0	1	1	0	0	0	2	0	0	0	0	0	2	2	0	
9		0	2	0	0	0	0	0	1	1	0	4	3	1	0	1	1	2	0	
10		0	0	2	0	0	0	0	0	1	0	0	1	2	0	1	0	1	2	
11		0	1	0	0	0	0	0	0	1	3	3	2	0	0	0	0	0	0	
12		1	1	1	2	0	4	0	0	0	0	1	1	0	0	0	0	0	0	
13		0	0	0	0	1	0	0	1	1	3	0	2	0	1	0	0	0	4	
14		0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	
15		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16																				

图 38-4 田间实测数据编辑整理格式

计算结果	当前日期 2006-1-10 12: 47: 26		
每样方数	实测频数	Fx	Fx*x
0	179	0	0
1	44	44	44
2	19	38	76
3	4	12	36
4	5	20	80
5	1	5	25
样方均值=0.4722	样本方差=0.8160		
负二项分布频次检验:			
最大或然法估计参数 K:			
K ₀ =	0.6487		
K=	0.5769		
每样方数	观察频数	理论频数	卡方值
0	179.0000	178.4703	0.0016
1	44.0000	46.3426	0.1184
2	19.0000	16.4467	0.3964
3	4.0000	6.3589	0.8750
4	5.0000	2.5595	2.3271
>=5	1.0000	1.8220	0.3709
理论频次总和=252.0000	卡方值=4.0894		p=0.1294

38.3 负二项分布公共 k 值估计

38.3.1 估计方法

负二项分布是生物种群中最常见的一种空间分布型。当研究对象符合负二项分布时，它的抽样设计必须首先获得一个适合各个田块的公共 k 值，然后再计算其他参数。在此我们采用 Bliss & Owen (1958) 提出的矩法估计公共 k 。

对服从负二项分布的单组资料， k 的矩法估计公式

$$k_1 = \bar{x}^2 / (S^2 - \bar{x})$$

式中， \bar{x} 为样本的平均值， S^2 为样本的方差。

对公共 k 值的计算，首先逐个计算 n 个种群的 $x = \bar{x}^2 - S^2/n$ 与 $y = S^2 - \bar{x}$ 并求其总和，于是得出 k 的最初估计值

$$k = \frac{\sum_{j=1}^n x}{\sum_{j=1}^n y}$$

此式将方差与平均值间的二次函数关系，通过 x' 和 y' 的坐标原点(0, 0)而直线化。若观察值成为直线，则用直线化后的值。若仍然分散，则重复校正。校正的方法是先求 A ，即

$$A = \frac{0.5(n-1)k^4}{\frac{k(k+1) - (2k-1)}{(n-3)/n^2}}$$

若各组资料的样本数量相等， A 只需计算 1 次，然后用 A 通过求 wx'^2 及 $wx'y'$ ，最后求出多个种群的公共 k 值(k_c):

$$k_c = \frac{\sum(wx'^2)}{\sum(wxy')}$$

k_c 的 χ^2 检验(自由度为处理组数减 1)

$$\chi^2 = \sum(wy'^2) - B_0^2$$

式中， $B_0^2 = \frac{\sum^2(wxy')}{\sum(wx'^2)}$ ， $wy' = \frac{A}{(\bar{x} + k)^2} \cdot y/x$ 。

38.3.2 试验数据资料整理及应用

根据田间实际观测资料计算各田块的理论取样数、平均值和方差。在编辑数据时，每块田的资料放一行，每一行中输入该田块的实际取样数、平均值和方差三个数据。示范例子数据如图 38-5。

	A	B	C	D	E
1		调查株数	平均值	方差	
2		2100	0.0771	0.0846	
3		2100	0.1924	0.2611	
4		2050	0.302	0.3825	
5		2100	0.1762	0.2072	
6		2100	0.189	0.2344	
7		2100	0.1862	0.2173	
8		2100	0.1109	0.1263	
9		2100	0.1076	0.1161	
10		2100	0.0367	0.0439	
11		2000	0.0655	0.0752	
12		2000	0.019	0.0227	
13					

图 38-5 负二项分布公共 k 值计算数据编辑格式

在菜单方式下，执行“专业统计”→“空间分布型”→“负二项分布公共 k 值”功能项后，系统输出计算结果如下：

公共 k 参数的初始估计值: $K' = 0.8572$			
A(i)	wXX	wXY	wYY
356.0480	2.408000	3.058900	3.885700
356.0480	11.92290	22.20180	41.34250
347.5684	23.54080	20.82060	18.41470
356.0480	10.31740	10.33480	10.35230
356.0480	11.58280	14.76750	18.82770
356.0480	11.30420	10.17040	9.150300
356.0480	4.649100	5.850000	7.361100
356.0480	4.407100	3.251000	2.398200
356.0480	0.590800	3.208000	17.41900
339.0889	1.693600	3.863100	8.811400
339.0889	0.154400	1.634100	17.29190
和=	82.57110	99.16010	155.2548

卡方测验结果是判断模型是否合适的依据，如 $\chi^2 > \chi_{0.05}^2$ 表明负二项分布公共 k 值的公共性不显著。本例中 $k = 115.255$ ， $\chi^2 = 155.2548$ ， $\chi^2 > \chi_{0.05}^2$ ，因此推断负二项分布公共 k 值的公共性不显著。

38.4 二元变量距离系数

38.4.1 方法简介

群落相似性表示两个群落共有的基本特征。相似性指标有两类：一类是相似性指标，其大小直接反映群落间的相似程度；另一类是相异性指标，反映群落间的差异程度。从数学上来讲，相似性和相异性是互补概念。若有 2 个生物群落 A 和 B，则最简单的相似

性测度是处理存在或不存在的一类资料。计算二项系数的基本数据可用以下 2×2 列联表:

		群落 A	
		具有的种数	没有的种数
群落 B	具有的种数	a	b
	没有的种数	c	d

表中, a 是群落 A 和群落 B 都有的物种数量, b 是群落 B 有但群落 A 无的物种数量, c 是群落 A 中有但群落 B 中无的物种数量, d 是群落 A 和群落 B 中都没有的物种数量, 且 $a+b+c+d=n$ 。关于 d 是否具有生物学意义争论颇多, 因此有人将 d 忽略不计。

现有文献中有 20 多种二项分布系数, DPS 系统仅提供以下常用的 10 种相似性系数:

(1) Jaccard (1901) 相似系数(S_j)

$$S_j = \frac{a}{a+b+c}$$

式中 a , b , c 与上述列联表中的定义相同。

(2) Czekanowski (1913) 相似系数(C_s)

$$C_s = \frac{2a}{2a+b+c}$$

该相似性测度类似于 Jaccard 相似系数(S_j), 它首先由 Czekanowski 于 1913 年提出, 并由 Sorensen 在 1948 年重新发现。这个系数偏重于考虑两个群落共有的物种数量, 何种情况下使用它则取决于数据的特性。如果一个群落里有很多物种, 而另一个群落里的物种少, 用 Czekanowski 相似系数可能较为有用。

(3) Sokal(1958) 相似系数(C_1)

$$C_1 = \frac{a+d}{n}$$

这是一个简单的相似系数, 它不仅利用两群落全都存在的物种信息, 同时还利用两群落都没有的物种信息。

(4) Baroni-Urbani 和 Buser 相似系数(S_b)

$$S_b = \frac{\sqrt{ad} + a}{a+b+c + \sqrt{ad}}$$

该相似系数较为复杂, 同样也利用了两群落都有和都无的物种信息。

(5) Ochilai (1957) 相似系数(C_o)

$$C_o = \frac{a}{\sqrt{a+b} \cdot \sqrt{a+c}}$$

(6) Dagnelie (1962)相似系数(C_d)

$$C_d = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

以上两种相似系数的计算也较复杂,但都是利用上述列联表中的信息。

(7) Rogers 和 Tanimoto (1960)系数

$$S_{RT} = \frac{a+d}{n+b+c}$$

(8) Kulczynski(1927)系数

$$S_K = \frac{a}{2} \left(\frac{1}{a+b} + \frac{1}{a+c} \right)$$

(9) Sokal 和 Sneath(1963)系数

$$S_s = \frac{ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(10) Watson 等(1966)系数

$$S_w = 1 - \frac{b+c}{2a+b+c}$$

38.4.2 DPS 平台操作示例

DPS 平台处理原始的群落资料,不必将资料整理成前面的列联表形式,因为相似系数计算极其容易,利用系统中设置的运算程序即可。

分析前先输入数据,一行为一个(群落)样本,一列为一个物种的数量,输完一个样本后再输下一个样本,然后将所有输入的数据定义成数据矩阵块(图 38-6)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1		物种数量																									
2	Erie	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
3	Ontario	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0
4																											

图 38-6 二元变量距离系数数据定义格式

然后,在菜单下选“专业统计”→“群落统计分析”→“二元数据相似系数”项操作,运算完成后即输出结果。

系统输出的相异系数矩阵,可用 38.3 节介绍的极点排序方法进行进一步分析,或用第 26 章介绍的系统聚类方法直接进行聚类分析,作出系统聚类谱系图。

注意,系统最多可处理 80 个群落、500 个物种的样本数据。

例如, Watson(1974)在 Erie 湖和 Ontario 湖调查甲壳纲动物分布情况,得到了 25 种

甲壳动物在湖中生存的资料, 其中 18 种在两个湖中均有分布, 一个种只在 Erie 湖中发现, 另一个种只在 Ontario 湖中发现, 5 个种在两个湖中均未发现。现计算两湖中甲壳动物群落的相似系数及相异系数, 数据的输入、编辑及数据块的定义如图 38-6 所示。

在菜单下, 选择“专业统计”→“群落参数统计”→“二元数据相似系数”功能项, 回车执行后系统立即输出如下结果:

Jaccard(1901)相异系数及相似系数: 相异系数 0.1000, 相似系数 0.9000;
 Czekanowski(1913)相异系数及相似系数(Cs): 相异系数 0.0526, 相似系数 0.9474;
 Sokal(1958)相异系数及相似系数(Cj): 相异系数 0.0800, 相似系数 0.9200;
 Baroni-Urbani 和 Buser 相异系数及相似系数(Sb): 相异系数 0.0678, 相似系数 0.9322;
 Ochilai(1957)相异系数及相似系数(Co): 相异系数 0.0526, 相似系数 0.9474;
 Dagnelie(1962)相异系数及相似系数(Cd): 相异系数 0.2193, 相似系数 0.7807;
 Rogers 和 Tanimoto, (1960)相异系数及相似系数: 相异系数 0.148150, 相似系数 0.851850;
 Kulczynski, (1927)相异系数及相似系数: 相异系数 0.052630, 相似系数 0.947370;
 Sokal 和 Sneath, (1963)相异系数及相似系数: 相异系数 0.842110, 相似系数 0.157890;
 Watson 等(1966)相异系数及相似系数: 相异系数 0.052630, 相似系数 0.947370。

38.5 距离系数计算

38.5.1 方法简介

各个群落间的距离系数(即不相似性系数)也可表达各个样本间的相似性测度。本系统在计算距离系数的同时, 还根据距离系数, 采用极点排序分析方法进行排序, 并进行排序效果的检验。在进行分析时, 可用系统提供的以下 6 种距离系数:

(1) 欧氏距离系数

$$\Delta_{jk} = \sqrt{\sum_{i=1}^n (x_{ij}^2 - x_{ik}^2)}$$

式中 Δ_{jk} 为样本 j 和样本 k 之间的欧氏距离, x_{ij} 为物种 i 在样本 j 的个体数量或生物量, x_{ik} 为物种 i 在样本 k 的个体数量或生物量, n 为物种总数。

从公式可以看出, 欧氏距离随着物种数的增加而增加, 因此一个补偿性的方法是计算其平均距离, 即

$$d_{jk} = \sqrt{\Delta_{jk}^2 / n}$$

式中 d_{jk} 为平均欧氏距离, Δ_{jk} 和 n 与上式中含义相同。

距离系数的测度很多, 一个最简单的距离系数是马氏距离(Manhattan distance)。马氏距离的计算公式是

$$d_M(j, k) = \sum_{i=1}^n |x_{ij} - x_{ik}|$$

这里, $d_M(j, k)$ 为样本 j 和样本 k 之间的马氏距离, x_{ij} 为物种 i 在样本 j 的个体数量或生物量, x_{ik} 为物种 i 在样本 k 的个体数量或生物量, n 为物种总数。

在马氏距离基础上, 有人提出了目前已被广泛应用的下面两个距离系数。

(2) Bray-Curtis Measure 距离系数

$$B = \frac{\sum_{i=1}^n |x_{ij} - x_{ik}|}{\sum_{i=1}^n (x_{ij} + x_{ik})}$$

B 为 Bray-Curtis 相异性测度, 是 Bray 和 Curtis 于 1957 年提出的标准化马氏距离系数。式中 n 为样本中物种数量, x_{ij} , x_{ik} 为样本 j 和样本 k 中第 i 种物种的个体数。

Bray-Curtis 相异性测度忽略了两个群落中物种都不存在的情形, 且丰富种在测度中占优势, 而稀有种在测度中起不了什么作用。

(3) Canberra 测度系数

$$C = \frac{1}{n} \left[\sum_{i=1}^n \left(\frac{|x_{ij} - x_{ik}|}{x_{ij} + x_{ik}} \right) \right]$$

这里, C 为样本 j 和样本 k 之间的 Canberra 测度系数(相异性系数), n 为样本中物种数量, x_{ij} , x_{ik} 为样本 j 和样本 k 中第 i 种物种的个体数。

(4) 百分率系数(percentage similarity)

$$P_{jk} = \sum_{i=1}^n \text{选取最小的 } P_{ij} \text{ 或 } P_{ik}$$

这是由 Renkonen(1938)提出的, 这里 P_{jk} 为样本 j 和样本 k 之间的百分率系数, P_{ij} 为物种 i 在样本 j 的百分率, P_{ik} 为物种 i 在样本 k 的百分率, n 为物种总数。

为统一起见, 取 $1 - P_{jk}$ 作为距离系数。

(5) Morisita 相似性指数

$$C_\lambda = \frac{2 \sum_{i=1}^n x_{ij} x_{ik}}{(\lambda_1 + \lambda_2) \sum_{i=1}^n x_{ij} \sum_{i=1}^n x_{ik}}$$

这里, C_λ 为样本 j 和样本 k 之间的 Morisita 相似性指数, n 为样本中物种数量, x_{ij} , x_{ik} 为样本 j 和样本 k 中第 i 种物种的个体数, 式中

$$\lambda_1 = \frac{\sum_{i=1}^n [x_{ij}(x_{ij} - 1)]}{\sum_{i=1}^n x_{ij} \left(\sum_{i=1}^n x_{ij} - 1 \right)}, \quad \lambda_2 = \frac{\sum_{i=1}^n [x_{ik}(x_{ik} - 1)]}{\sum_{i=1}^n x_{ik} \left(\sum_{i=1}^n x_{ik} - 1 \right)}$$

为统一起见，取 $1-C_\lambda$ 作为距离系数。

(6) Horn 指数

$$R_0 = \frac{\sum_{i=1}^n [(x_{ij} + x_{jk}) \ln(x_{ij} + x_{jk})] - \sum_{i=1}^n (x_{ij} \ln x_{ij}) - \sum_{i=1}^n (x_{ik} \ln x_{ik})}{[(N_j + N_k) \ln(N_j + N_k)] - N_j \ln N_j - N_k \ln N_k}$$

式中， R_0 为样本 j 和样本 k 之间的 Horn 相似性指数， n 为样本中物种数量， x_{ij} ， x_{ik} 为样本 j 和样本 k 中第 i 种物种的个体数， $N_j = \sum_{i=1}^n x_{ij}$ ， $N_k = \sum_{i=1}^n x_{ik}$ 。为统一起见，取 $1-R_0$ 作为距离系数。

38.5.2 DPS 平台的操作示例

分析数据前先编辑数据，数据一行为一个(群落)样本，一列为一个物种的数量，输完一个样本后再输下一个样本。编辑完成后，定义数据矩阵块。在菜单下执行本项计算。

这时，如果只有两个样本(群落)，系统将会把所有的相异性测度计算出来，但不进行极点排序(实际上不必进行)操作。如果样本数量在 3 个以上，分析过程中系统会提示输入距离系数计算方法，并执行极点排序操作，给出相应的结果。

系统输出的距离(相异)系数矩阵，可用第 26 章介绍的系统聚类方法直接进行系统聚类分析，作出系统聚类谱系图。

例如，赵志模等(1985)采用主成分分析技术分析了柑橘园昆虫群落随树龄的演替规律。在此我们根据其原始数据采用 Bray & Curtis (1957)提出的公式计算距离系数，并用极点排序和系统聚类(离差平方和)方法对其排序、聚类分析。计算前先编辑数据，数据编辑格式是一行放一个橘园昆虫生物群落各个种类的数据，再定义数据块(如图 38-7)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		物种种类															
2		30	0	0	0	850	3	5	0	0	0	17	10	1	0	0	
3		40	0	0	0	952	3	87	0	4	0	15	0	1	2	0	
4		11	0	1	0	1064	5	55	2	0	0	15	13	1	3	0	
5	橘 园 类 型	9	0	8	0	2212	64	77	2	0	0	4	38	0	1	0	
6		67	21	5	16	694	110	327	4	33	0	4	69	2	15	0	
7		173	24	6	5	2132	86	124	30	22	0	2	88	2	0	1	
8		68	0	0	44	252	22	184	0	0	1	6	43	1	0	0	
9		41	5	1	3	76	62	82	6	1	4	2	91	1	1	1	
10		580	24	1	33	271	16	66	6	7	6	0	17	0	3	0	
11		2150	0	4	8	207	25	128	8	0	6	1	43	0	2	3	
12																	

图 38-7 聚类系数计算数据定义格式

然后在菜单方式下选择距离系数计算，并选择按 Bray 和 Curtis(1957)提出的公式执行距离系数计算分析后，系统立即给出所得到的系数矩阵：

Bray-Curtis Measure 系数矩阵
距离系数矩阵

0.0000	0.1050	0.1419	0.4710	0.3456	0.5010	0.6005	0.8036	0.6721	0.8538
0.1050	0.0000	0.0862	0.4055	0.3242	0.4267	0.5490	0.7218	0.6382	0.8157
0.1419	0.0862	0.0000	0.3562	0.3780	0.4028	0.6170	0.7841	0.6718	0.8418
0.4710	0.4055	0.3562	0.0000	0.5272	0.0881	0.7352	0.8080	0.7776	0.8544
0.3456	0.3242	0.3780	0.5272	0.0000	0.4579	0.4074	0.6009	0.5920	0.7525
0.5010	0.4267	0.4028	0.0881	0.4579	0.0000	0.6882	0.7598	0.6854	0.7761
0.6005	0.5490	0.6170	0.7352	0.4074	0.6882	0.0000	0.4569	0.4512	0.7018
0.8036	0.7218	0.7841	0.8080	0.6009	0.7598	0.4569	0.0000	0.6631	0.8082
0.6721	0.6382	0.6718	0.7776	0.5920	0.6854	0.4512	0.6631	0.0000	0.4971
0.8538	0.8157	0.8418	0.8544	0.7525	0.7761	0.7018	0.8082	0.4971	0.0000

38.6 极点排序

38.6.1 方法简介

极点排序是 20 世纪 50 年代由美国 Wisconsin 学派创立的方法, 以其代表作者姓氏命名, 称为 Bray-Curtis 方法。在进行极点排序时, 首先需确定群落间的相似性测度, 并将其转换成相异系数, 然后根据相异系数进行排序。排序的基本步骤如下:

(1) 排序轴 x 的选择。选择相异系数最大的两个群落作为第一排序轴 x 的两个端点。如群落中相异系数最大的是群落 A 和群落 B, 两者之间的相异系数为 L_x , 可令群落 A 为 x 轴的 0 端(点 a), 它在该轴上排序坐标为 0, 群落 B 为 x 轴的另一端点(点 b), 它在 x 轴的坐标是 L_x , 其余群落位置可根据它们与两 endpoint 群落的相异系数来定, 按公式

$$X_i = \frac{L_x^2 + D_{ia}^2 - D_{ib}^2}{2L_x} \quad (i \neq a, b)$$

式中 D_{ia} 和 D_{ib} 分别代表所求群落 i 与群落 A 和群落 B 之间的相异系数。

(2) 排序轴 y 的选择。 y 轴应尽量与 x 轴垂直, 因此选择偏离 x 轴最大者为 y 轴的第一个端点 a , y 轴的另一个端点 b (第二个端点) 则选取与第一个端点相异系数最大者。若两者之间的相异系数为 L_y , 可令群落 A 为 y 轴的 0 端, 它在该轴上排序坐标为 0, 群落 B 为 y 轴的另一端点, 它在 y 轴的坐标是 L_y , 其余群落位置可根据它们与两 endpoint 群落的相异系数来定, 按公式

$$X_j = \frac{L_y^2 + D_{ja}^2 - D_{jb}^2}{2L_y} \quad (j \neq a, b)$$

式中 D_{ja} 和 D_{jb} 分别代表所求群落 j 与群落 A 和群落 B 之间的相异系数。

由此, 可将群落在 x 和 y 轴上的位置全部求出。

(3) 排序效果的检验。由于极点排序往往只求出两维坐标, 它是否能很好地反映原群落间的相异性, 则需要进行统计检验。检验的方法是将按排序坐标求出来的群落之间

的欧氏距离，与原来群落之间相应的欧氏距离进行相关分析，得到两个欧氏距离系数矩阵之间的相关系数，最后根据相关系数的大小进行检验。

38.6.2 DPS 平台的操作示例

在 DPS 平台上，根据二元数据相异系数矩阵，或距离系数矩阵进行极点排序分析，及将数据矩阵定义成数据块，如将 38.5 节例子中的 Bray-Curtis Measure 距离系数定义成数据块(如图 38-8):

	A	B	C	D	E	F	G	H	I	J	K	L
1		A	B	C	D	E	F	G	H	I	J	
2	1	0.0000	0.1050	0.1419	0.4710	0.3456	0.5010	0.6005	0.8036	0.6721	0.8538	
3	2	0.1050	0.0000	0.0862	0.4055	0.3242	0.4267	0.5490	0.7218	0.6382	0.8157	
4	3	0.1419	0.0862	0.0000	0.3562	0.3780	0.4028	0.6170	0.7841	0.6718	0.8418	
5	4	0.4710	0.4055	0.3562	0.0000	0.5272	0.0881	0.7352	0.8080	0.7776	0.8544	
6	5	0.3456	0.3242	0.3780	0.5272	0.0000	0.4579	0.4074	0.6009	0.5920	0.7525	
7	6	0.5010	0.4267	0.4028	0.0881	0.4579	0.0000	0.6882	0.7598	0.6854	0.7761	
8	7	0.6005	0.5490	0.6170	0.7352	0.4074	0.6882	0.0000	0.4569	0.4512	0.7018	
9	8	0.8036	0.7218	0.7841	0.8080	0.6009	0.7598	0.4569	0.0000	0.6631	0.8082	
10	9	0.6721	0.6382	0.6718	0.7776	0.5920	0.6854	0.4512	0.6631	0.0000	0.4971	
11	10	0.8538	0.8157	0.8418	0.8544	0.7525	0.7761	0.7018	0.8082	0.4971	0.0000	
12												

图 38-8 极点排序数据定义格式

进入菜单操作，执行“专业统计”下的“群落参数统计”中的“极点排序”分析，得到如下极点排序结果：

X	Y	偏离值
0.130422	0.378510	0.452583
0.134049	0.342713	0.382702
0.086757	0.372299	0.345473
0.000000	0.382060	0.000000
0.258475	0.307130	0.459489
0.079254	0.412549	0.038476
0.455287	0.261139	0.577263
0.427011	0.000000	0.685949
0.636442	0.539907	0.446769
0.854400	0.854400	0.000000
<i>RR=0.8819</i>		

38.7 物种丰富度估计

38.7.1 方法简介

有些群落的组成非常简单明了，以至于所有存在的物种全都可以调查计数，那么这个值就是最简单的物种丰富度(richness)的指标。但并非所有群落内物种都能查清楚，如昆虫群落、热带植物、土壤无脊椎动物群落。当只有抽样资料时，如何估计整个群落中

物种的丰富度? DPS 系统对群落内样方调查资料, 采用一种非参数估计方法——刀切法(jackknife), 根据稀有种在各样方出现的频率来估计群落内物种的丰富度(Heltshe, Forrester, 1983)。

在应用刀切法估计群落丰富度前, 需将一系列群落样方资料整理成表 38-1 形式。即对每个样方只记载有(用 1 表示)或没有(用 0 表示), 这在 DPS 平台上是自动进行的。现要计算抽样中稀有种(unique species)数量, 在这里稀有种是指样方中出现或不出现, 而不是指数量上的多寡, 具体来说, 指只在一个样方中仅出现一次。按 Heltshe 和 Forrester (1983b)提出的刀切法, 物种数量可根据下式进行估计

$$\hat{S} = S + \left(\frac{n-1}{n} \right) \cdot k$$

式中, \hat{S} 为物种丰富度的刀切法估计值, S 为在 n 个样方中已经观察到的物种总数, n 为抽样的总样方数, k 为稀有种的数量。

表 38-1 采用刀切法估计群落丰富度样方抽样资料整理格式*

物种	样方							和
	A	B	C	D	E	F		
1	1	0	0	1	1	0	3	
2	0	1	0	0	0	0	1	
3	1	1	1	1	1	0	5	
4	0	1	0	0	1	0	2	
5	1	1	1	1	1	1	6	
6	0	0	0	0	1	0	1	
7	0	0	1	1	1	1	4	
8	1	1	0	0	1	1	4	

* 表中 1 表示存在, 0 表示不存在。

采用刀切法估计物种丰富度的估计方差

$$VS = \left(\frac{n-1}{n} \right) \left[\sum_{j=1}^s (j^2 f_j) - \frac{k^2}{n} \right]$$

式中 f_j 为含 j 个稀有种样方的数量($j=1, 2, \dots, S$), k 为稀有种数量, n 为抽样总样方数。

根据方差 VS 可以得到物种丰富度刀切法估计的置信限 $\hat{S} \pm t_{\alpha} \sqrt{VS}$, 式中 t_{α} 为自由度为 $n-1$, 置信水平为 α 的 Student t 值。

38.7.2 DPS 平台的处理方法

计算前先编辑数据, 数据编辑格式是每一行放一个样方所查得的生物物种数量, 然后定义数据块(图 38-9)。如在 10 个样方中, 共查到 14 个生物物种(Heltshe, Forrester,

1983)资料, 试根据这批数据采用刀切法估计群落的丰富度。首先, 将调查资料按如下方式编辑数据并定义数据块, 再在菜单方式下选择“专业统计”→“群落参数估计”→“物种丰富度估计”菜单。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	样方	Sb	Ns	Pl	Sr	Eh	Hf	Cc	Sv	Hg	Bc	Mb	Aa	Nt	Ts	
2	1	0	2	0	1	0	1	1	2	0	0	0	0	0	8	
3	2	13	2	1	0	0	1	0	0	1	0	0	0	0	36	
4	3	21	4	0	1	1	2	0	0	0	1	3	5	0	14	
5	4	14	4	0	2	2	1	0	0	0	0	0	1	0	19	
6	5	5	1	0	0	0	0	0	0	0	0	0	0	0	3	
7	6	22	1	0	6	0	1	0	0	0	0	0	2	0	22	
8	7	13	1	0	0	1	0	0	0	0	0	0	0	0	6	
9	8	4	0	1	0	0	0	0	0	0	0	0	0	1	8	
10																

图 38-9 物种丰富度估计数据定义格式

执行分析计算后, 给出结果为: 实际物种数 14, 理论物种数 18.5000, 理论值的标准误 2.0125, 理论值 95%置信限 13.9475~23.0525。

注意, 应用本方法估计群落的丰富度, 其群落丰富度的估计值通常是偏大(Heltshe, Forrester, 1983), 而估计误差(S)通常偏小。从刀切法的估计公式可以看出, 最大的估计值可能达到观察值(实际物种数量)的两倍, 故该方法不宜用于稀有种数量很多的情形, 同样也不宜用于抽样样本太小的情形。

38.8 对数序列参数估计

38.8.1 方法简介

Fisher(1943)提出对数序列模型时认为, 当群落中优势种类很少但优势种在个体数量或相对重要性上表现很突出时, 可假设不同种包含 r 个个体的概率 P_r 是负二项分布。在考虑略去 0 项的条件下, 则一个种包含 r 个个体的概率 P_r' 是截尾负二项分布项。于是, 参数 k 就能测量出物种间个体数量的变异性。当不同种的密度彼此只有轻微差别时, 可望 k 值很大; 当不同种间密度差异很大, 令负二项分布参数 $k \rightarrow 0$, 这时可得到一个两参数的种——多度公式

$$n_r = \alpha \frac{X^r}{r}$$

即对数序列模型。它表示的数列为 $\alpha x, \alpha^2 x / 2, \alpha^3 x / 3, \dots$ 其中 αx 表示在整个抽样中由一个个体所代表的物种数量, $\alpha^2 x / 2$ 表示由两个个体所代表的物种数量, 各项的总和等于物种种群数量的总和。对数序列数据是由物种数目和物种的个体数量决定, 关系式为

$$S = \alpha \ln \left(1 - \frac{N}{\alpha} \right)$$

式中 α 为多样性指数, α 大时稀有种比富集种多, 若 α 小则富集种相对较多。

38.8.2 DPS 处理平台的操作示例

计算前先编辑数据, 数据编辑格式是将群落的数据全都放在一行里, 并定义数据块, 在菜单方式下执行计算。如有一个群落, 共有 36 个物种(丁岩钦, 1994), 其数据编辑定义格式如下:

1 1 2 2 3 3 3 5 5 6 6 6 7 7 9 9 10 10 11 12 13 14 19 20 23 24 25 28 29 35 50 55 60 95 110 200

执行计算后, 系统立即输出计算结果如下:

序号	总个数	种类数	x	α	方差
0	918	36	0.9919287	7.469746	1.549920
序号	观察值	拟合值	机率值	系数	
1	200.0000	196.02	0.213526	7.409456	
2	110.0000	132.45	0.144285	3.674826	
...	
36	1.0000	0.68	0.000743	0.154989	
卡方检验					
	组距	观察值	理论值	卡方值	
	0.5~1.5	2	6	2.6667	
	1.5~3.5	5	7	0.5714	
	3.5~7.5	7	5	0.8000	
	7.5~15.5	8	5	1.8000	
	15.5~31.5	7	5	0.8000	
	31.5~63.5	4	4	0.0000	
	63.5~127.5	2	2	0.0000	
	>= 127.5	1	2	0.5000	

计算得卡方值为 7.138, 小于 $\chi_{0.05}^2 = 12.59$, 故该群落物种数量符合对数正态序列模型。

38.9 对数正态分布模型参数估计

38.9.1 方法简介

对数序列模型反映了具有很小的丰富度的物种数量最大, 单个样本所表示的物种数量总是最大的, 但实际情况并不总是这样。因此, Preston (1948)提出的对数正态分布模型可用于描述群落中优势种和稀有种都很少, 而中等程度的种很多的生物群落。对数正态分布模型为

$$S_{(R)} = S_0 e^{(-a^2 R^2)}$$

式中 $S_{(R)}$ 是第 R 倍频程中的物种数量; S_0 是对数正态分布的众数倍频程物种数; a 是与分布有关的参数, 若分布的标准差为 δ , 则 $\hat{a} = \frac{1}{\delta\sqrt{2}}$ 。

为了使田间资料适合对数分布, 可用线性回归

$$\ln S_{(R)} = \ln S_0 - a^2 R^2$$

模型求解参数, 参数求解时, 需采用以倍频程(octaves)分组 R 值的办法, 列出每个倍频程代表的种数。它是让每一倍频程的中点值恰好等于前一组的中点值的两倍, 即 $R=2^x$ ($x=0, 1, 2, \dots$), 亦即 $R=1, 2, 4, 8, \dots$ 这样, 各组的中点值分别为 1.5, 3, 6, \dots 如果一个种的个体数恰好落在边界上, 即它有 2^x 个个体, 可将一半分配在倍频程($2^{x-1}, 2^x$)内, 另一半分配在倍频程($2^x, 2^{x+1}$)内。这种办法实际上是对各个种的个体数取以 2 为底的对数。

拟合对数正态模型参数, 本系统采用由 Cohen(1959, 1961)提出的方法。

38.9.2 DPS 处理平台操作方法

计算前先编辑数据, 数据编辑格式是将群落的数据全都放在一行里, 并定义数据块, 再在菜单方式下执行计算。如对上一例的数据进行分析, 其数据编辑定义格式如下:

```
1 1 2 2 3 3 3 5 5 6 6 6 6 7 7 9 9 10 10 11 12 13 14 19 20 23 24 25 28 29 35 50 55 60 95 110 200
```

在执行计算后, 系统立即输出结果如下:

种数	个数	中位数	分布参数 a	方差 δ	种数估计值
36	918	11	1.068224	0.307053	36.24
种类分布观察值和期望值					
	分组	区间	期望值频次	观察值频次	
	0	0—0.5	0.24	—	
	1	0.5—1.5	1.702	2.00	
	2	1.5—3.5	4.291	5.00	
	3	3.5—7.5	6.944	7.00	
	4	7.5—15.5	8.101	8.00	
	5	15.5—31.5	7.030	7.00	
	6	31.5—63.5	4.579	4.00	
	7	63.5—127.5	2.243	2.00	
	8	127.5—255.5	0.826	1.00	
	9	255.5—511.5	0.228	0.00	
	10	511.5—1023.5	0.047	0.00	
	11	1023.5—2047.5	0.007	0.00	

38.10 群落多样性指数

38.10.1 方法简介

1. Simpson 多样性指数(D)

前面介绍的对数序列模型和对数正态分布模型, 或多或少由于其参数估计复杂以及

统计方法缺乏理论上的完整、充分的特性, 所以, 群落生态学家转而趋向使用一些非参数的统计方法来描述群落的生物多样性(即异质性的度量), 因为这些方法对群落中各个物种丰富度的分布曲线没有什么假设要求。第一个非参数多样性测度由 Simpson (1949) 提出, 人们称之为 Simpson 指数(D)的测度。Simpson 多样性指数的定义为

$$D = 1 - \sum_{i=1}^s p_i^2$$

式中, p_i 为第 i 种在群落中所占比例。严格地说, 这个公式只能用于估计无限总体的多样性指数。Pielou(1969)指出, 对于有限总体, 其多样性指数估计公式为

$$D = 1 - \sum_{i=1}^s \left[\frac{n_i(n_i - 1)}{N(N - 1)} \right]$$

式中, n_i 为抽样中第 i 个物种的个体数量, N 为抽样中所有物种的个体总和, s 为物种总数。本系统是按 Pielou(1969)提出的公式计算 Simpson 多样性指数的。

2. Shannon-Wiener 多样性指数(H')

物种 Shannon-Wiener 多样性指数是根据信息论建立起来的最通用的度量。该指数表征在信息通信中的某一瞬间, 一定符号出现的不定度以及它传递的信息总和。在群落的多样性测度上, 借用了信息论中这种不确定性的度量方法。Pielou (1966)认为多样性应该是“等于从一个群落中随机抽取一个个体的种的不定量”。显然, 群落的多样性指数越高, 其不定性就越不稳定。Shannon-Wiener 多样性指数(H')计算公式为

$$H' = - \sum_{i=1}^s p_i \text{lb}(p_i)$$

式中, H' 为样本中的信息容量(bit/个体), 即种的多样性指数, s 为物种总数, p_i 为第 i 种物种个体数占群落总个体数的比例。

严格地讲, Shannon-Wiener 信息指数只能用于随机取样、群落很大且群落中所有的物种已知情况, 但是大多数群落调查并不是这种情况。因此, Pielou (1966)推荐使用更为精确的 Brillouin 指数。

3. Brillouin 指数(H)

按照 Pielou(1966)的观点, 许多群落抽样样本应该作为一个集合, 而不应作为一个来自很大的生物群落的样本来对待。在任何情形下, 群落数据应假定是来自一个有限的集合(collection), 且是无放回抽样。因此合适的信息理论测度应是 Brillouin 指数(H)

$$H = \frac{1}{N} \text{lb} \left(\frac{N!}{n_1! n_2! n_3! \dots} \right)$$

式中, n_1 为抽样中第 1 个物种的个体数量, n_2 为抽样中第 2 个物种的个体数量, n_3 为抽

样中第 3 个物种的个体数量, 如此类推, N 为抽样中所有物种的个体总和。

4. 均匀度

均匀度计算的通式为

$$\text{均匀度} = \frac{D}{D_{\max}} \quad \text{或} \quad \text{均匀度} = \frac{D - D_{\min}}{D_{\max} - D_{\min}}$$

式中, D 为多样性指数, D_{\max} 为给定 S 个物种 N 总个体数时多样性指数最大值, D_{\min} 为给定 S 个物种 N 总个体数时多样性指数最小值。

在本系统中, 我们为用户提供了目前文献中最常见到的一种根据 Shannon-Wiener 信息指数计算均匀度的方法, 其公式是

$$J' = \frac{H'}{H'_{\max}} = \frac{H'}{\text{lb}S}$$

式中, J' 为均匀度, H' 为 Shannon-Wiener 信息指数, H'_{\max} 为 H' 的最大值, 等于 $\text{lb}S$, S 为物种总数。

5. McIntosh 指数(D_{Mc})

McIntosh(1967)认为, 一个群落(集合)可看成是 S 维空间的一个点, 每一维坐标由一个物种的丰富度(即个体数)表示。这样, 由点到 S 维坐标系原点的距离, 可由欧氏距离表示

$$U = \sqrt{\sum_{i=1}^s (N_i - 0)^2} = \sqrt{\sum_{i=1}^s N_i^2}$$

显然, 对已知的 $N \left(N = \sum_{i=1}^s N_i \right)$, 物种越多, U 越小。因此 U 是集合一致性的一个度量。当集合只含一个物种时, 达到它的最大值; 当每个个体都属于不同种时, 达到最小值。因为多样性是一致性的互补, 所以该多样性的度量公式为

$$D_{\text{Mc}} = \left(N - \sqrt{\sum_{i=1}^s n_i} \right) / \left(N - \sqrt{N} \right)$$

式中, n_i 为抽样中第 i 个物种个体数量, N 为抽样中所有物种的个体总和, s 为物种数。

物种多样性指数的 95% 置信区间, 可采用 Bootstrap 随机模拟抽样方法, 进行模拟抽样。然后根据模拟抽样样本, 采用基于正态分布的方法、或用分位数方法进行 95% 置信区间的估计。

38.10.2 DPS 平台的操作示例

计算前先编辑数据, 数据编辑格式是一行放一个生物群落中各个种类的数据。因此

每次可计算多个群落的多样性指数。然后定义数据块，在菜单方式下选择“专业统计”→“群落参数估计”→“群落多样性指数”。执行多样性指数计算分析后，系统立即给出各个多样性指数测度。

例如，作者于 1995 年 6 月 26 日在浙江富阳中国水稻研究所试验区研究稻田生境与非稻田生境节肢动物群落的关系，按试验设计进行取样，得到 7 个稻田节肢动物群落样本资料，现以其中害虫亚群落中 14 个种数据为例计算不同生物群落的多样性指数(图 38-10)。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	群落类型	物 种														
2		s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	
3	非稻田生境	0	0	0	52	79	2	0	11	0	0	89	9	301	6	
4	距非稻田生境 1m	0	7	0	2	0	31	0	140	0	5	2	2	2	2	
5	距非稻田生境 5.3m	0	22	0	1	0	10	1	168	0	5	4	0	1	3	
6	距非稻田生境 9.3m	0	23	5	6	0	26	0	151	0	1	0	0	0	2	
7	距非稻田生境 13.3m	0	28	0	4	0	32	0	229	0	9	0	1	0	2	
8	距非稻田生境 17.3m	0	24	0	5	0	63	0	139	0	9	1	0	1	0	
9	距非稻田生境 21.3m	0	23	7	2	0	55	0	95	0	9	1	2	0	0	
10																

图 38-10 群落多样性指数估计数据定义格式

计算前按如下方式编辑数据并定义数据块，再在菜单方式下选择“专业统计”→“群落参数估计”→“群落多样性指数”功能项，回车即可执行运算。计算时出现如下用户界面：



这里用户可通过各种选项的确定，得到物种多样性指数及其 95%置信区间的估计结果。用户界面中一般可选择模拟 5000 次、采用正态分布法计算多样性指数 95%置信区间。本例输出结果如下：

序号	物种	个体数	丰富度	SIMPSON	95%置信区间		SHANNON	95%置信区间	
S1	8	549	1.1097	0.6438	0.4537	0.8339	1.3422	0.8653	1.8191
S2	9	193	1.5201	0.4478	0.1870	0.7086	0.9783	0.3080	1.6486
S3	9	215	1.4896	0.3774	0.0179	0.7369	0.8649	0.0699	1.6598
S4	7	214	1.1182	0.4766	0.1721	0.7811	0.9986	0.3550	1.6421

S5	7	305	1.0489	0.4171	0.1810	0.6532	0.8835	0.4094	1.3575
S6	7	242	1.0931	0.5931	0.3670	0.8192	1.1459	0.6642	1.6277
S7	8	194	1.3288	0.6655	0.3921	0.9389	1.3436	0.6910	1.9962
均匀度	95%置信区间		BRILLOUIN	95%置信区间		McIntosh	95%置信区间		
0.6455	0.4692	0.8218	1.8940	1.2191	2.5688	0.4201	0.2584	0.5819	
0.4453	0.1526	0.7379	1.3168	0.3827	2.2509	0.2752	0.0496	0.5007	
0.3936	0.0479	0.7393	1.1652	0.0432	2.2873	0.2252	-0.0689	0.5192	
0.5132	0.2044	0.8219	1.3682	0.4669	2.2696	0.2952	0.0479	0.5424	
0.4540	0.2139	0.6941	1.2222	0.5526	1.8919	0.2499	0.0534	0.4464	
0.5889	0.3830	0.7948	1.5877	0.9134	2.2620	0.3849	0.1978	0.5721	
0.6461	0.4031	0.8891	1.8447	0.9372	2.7522	0.4511	0.2104	0.6918	
模拟次数=5000									
"95%CI 计算方法:正态分布法"									

38.11 生态位宽度指数

38.11.1 方法简介

生态位宽度(niche breadth)指数是指生物利用资源多样性的一个测度指标,若现有的资源谱中,仅能利用一小部分资源,称狭生态位,能利用很大部分资源的,则称为广生态位。本系统提供下列生态位宽度指标。

1. Levin 测度(B)

$$B = \frac{1}{\sum p_j^2}$$

式中, P_j 为利用资源 j 的个体的比例。Levin 测度指标比较简单, Hurlbert 建议下述测度作为标准生态位宽度

$$B_A = \frac{B-1}{n-1}$$

式中, n 为可能的资源状态总数。

2. Shannon-Wiener 信息指数(H')及均匀度测度(J')

$$H' = -\sum_{j=1}^n p_j \ln(p_j), \quad J' = H' / \ln(n)$$

式中, P_j 为利用资源 j 的个体的比例, n 为可能的资源状态总数。

3. Hurlbert 生态位宽度指数

$$B = \frac{1}{\sum_{j=1}^n \frac{p_j^2}{a_j}}$$

式中, p_j 为利用资源 j 的个体的比例, a_j 为资源 j 可利用的项目数 ($\sum a_j = 1.0$), 为标准化的 Hurlbert 生态位宽度指数, 可使用下述方程

$$B_A = \frac{B - a_{\min}}{1 - a_{\min}}$$

a_{\min} 是资源中最小者。

4. Smith 生态位宽度测度(FT)

$$FT = \sum_{j=1}^n \sqrt{p_j \alpha_j}$$

式中, P_j 为利用资源 j 的个体的比例, α_j 为整个资源中资源 j 所占比例, n 为可能的资源状态总数。对于大样本, 95% 的置信限可根据公式 $\sin\left(x - 1.96 \frac{1}{\sqrt{4y}}\right)$ 和 $\sin\left(x + 1.96 \frac{1}{\sqrt{4y}}\right)$ 进行估计。其中 $x = \arcsin(FT)$, $y = \sum_{j=1}^n N_j$ (即个体总数)。

38.11.2 DPS 平台的操作示例

计算前先编辑数据, 数据编辑格式是一行放一个生物物种利用的各个食物资源数据, 然后定义数据块。如有 3 种食虫鸟利用昆虫资源, 按虫体大小, 可将资源分成 4 个等级。现将不同猎物资源利用多少的定值资料编辑和定义, 如图 38-11。

猎物资源按个体大小分等级表

	A	B	C	D	E	F
1		0~10 mm	11~20 mm	21~30 mm	31~40 mm	
2	Empidonax minutus	11	2	0	0	
3	Contopis virens	5	2	1	0	
4	Mylarchus crinitus	3	1	1	3	
5						

图 38-11 群落生态位宽度估计数据定义格式

再在菜单方式下选择“专业统计”→“群落参数估计”→“生态位宽度”。执行计算分析后, 系统输出各个指数测度如下:

a. 各项指标

b. 各个物种常用资源情况

物种	Levins	标准值	均匀度	HURL-BERT'S	标准值	Smith	FT 95%置信限	频数	截断比例	常用资源种类
1	1.3520	0.352	0.6194	0.6760	0.352	0.927	0.7936~0.9939	1	0.20	S1=84.62%
2	2.1333	0.566	0.8194	0.7111	0.566	0.949	0.7860~0.9996	2	0.20	S1=62.50%, S2=25.00%
3	3.2000	0.733	0.9056	0.8000	0.733	0.965	0.8206~0.9964	2	0.18	S1=37.50%, S4=37.50%

38.12 生态位重叠指数

38.12.1 方法简介

了解群落组织结构的一个方面是测度生物群落不同生物物种在资源利用方面的重叠程度。它表示种间竞争中，某一物种与竞争种共存的极限相似性。为了计算重叠程度，最常用的资源测度是食物和空间(微生境)。有几个生态位重叠指标已被提出，但是哪个更好尚有相当大的争议。本系统提供了下述几个生态位重叠指标，供用户参考应用。

1. MacArthur-Levins 重叠度指数

为度量生态位重叠，MacArthur 和 Levins(1967)提出了第一个公式

$$M_{jk} = \sum_{i=1}^n p_{ij} p_{ik} / \sum_{i=1}^n p_{ij}^2$$

式中 M_{jk} 表示种类 k 对种类 j 的生态位重叠； p_{ij} , p_{ik} 等于由种类 k 或种类 j 所利用的整个资源中第 i 种资源所占比例； n = 资源状态总数。

由于 MacArthur-Levins 生态位重叠指数不对称，于是 Pianka (1973)对它进行了修正，随即有了 Pianka 测度

$$O_{jk} = \frac{\sum_{i=1}^n p_{ij} p_{ik}}{\sqrt{\sum_{i=1}^n p_{ij}^2 \sum_{i=1}^n p_{ik}^2}}$$

2. 百分率重叠测度 P

Renknen(1938)提出了一个非常简单，但是又非常引人注意的生态位重叠测度。该测度利用百分率来进行计算，其公式为

$$P_{jk} = \left[\sum_{i=1}^n (p_{ij}, p_{ik} \text{ 中较小值}) \right] \times 100\%$$

式中， P_{jk} = 种类 k 和种类 j 之间重叠比例； p_{ij} , p_{ik} = 由种类 k 或种类 j 所利用的整个资源中第 i 种资源所占比例； n = 资源状态总数。

3. Morisita 相似性测度

Morisita 相似性系数(见 38.2 节)也可用来作为生态位测度的指标，计算公式为

$$C = \frac{2 \sum_{i=1}^n P_{ij} P_{ik}}{\sum_{i=1}^n P_{ij} [(n_{ij} - 1)/(N_j - 1)] + \sum_{i=1}^n P_{ik} [(n_{ik} - 1)/(N_k - 1)]}$$

式中, C = Morisita 生态位重叠; P_{ij} = 由种类 j 所利用的整个资源中第 i 种资源所占比例; P_{ik} = 由种类 k 所利用的整个资源中第 i 种资源所占比例; n_{ij} = 使用第 i 类资源的种类 j 的个体数; n_{ik} = 使用第 i 类资源的种类 k 的个体数; N_j, N_k 为样本中种类 j 和种类 k 的个体总数。

4. Morista-Horn 指数

1966 年, Horn 提出一个简化的用于测定生态位重叠的 Morista 指数, 其计算公式是

$$C_H = \frac{2 \sum_{i=1}^n P_{ij} P_{ik}}{\sum_{i=1}^n P_{ij}^2 + \sum_{i=1}^n P_{ik}^2}$$

式中, C_H 等于简化 Morisita 种类 k 对种类 j 的生态位重叠指数; P_{ij}, P_{ik} 等于由种类 k 或种类 j 所利用的整个资源中第 i 种资源所占比例; n = 资源状态总数。

5. Horn 重叠指数

它是 Horn(1966)提出的信息论指数, 公式为

$$R_o = \frac{\sum_{i=1}^n (P_{ij} + P_{ik}) \ln(P_{ij} + P_{ik}) - \sum_{i=1}^n P_{ij} \ln(P_{ij}) - \sum_{i=1}^n P_{ik} \ln(P_{ik})}{2 \ln(2)}$$

式中, R_o = Horn 种类 k 对种类 j 的生态位重叠指数; P_{ij} = 由种类 j 所利用的整个资源中第 i 种资源所占比例; P_{ik} = 由种类 k 所利用的整个资源中第 i 种资源所占比例; n = 资源状态总数。

6. Hurlbert 指数(Hurlbert, 1978)

$$L = \sum_{i=1}^n \left(\frac{P_{ij} P_{ik}}{a_i} \right)$$

式中, L = 简化 Morisita 种类 k 对种类 j 的生态位重叠指数; P_{ij}, P_{ik} = 由种类 k 或种类 j 所利用的整个资源中第 i 种资源所占比例; a_i = 资源状态 i 的比例数; n = 资源状态总数。

38.12.2 DPS 平台的操作示例

分析前先编辑数据, 数据一行为一个物种, 一列为一个资源种类, 输完一个物种样

本后再输下一个样本。编辑完成后，定义数据矩阵块。然后进入主菜单，执行本项计算后系统会给出各个指标的计算结果。

注意，系统最多可处理 100 个群落，共 5000 个样本数据。

例如，调查了 2 个生物物种，获得每种利用 7 个资源种类的数据，可按如下方式编辑数据并定义数据块(图 38-12)：

	A	B	C	D	E	F	G	H	I
1									
2		7	1	286	71	0	0	0	
3		0	0	38	24	30	140	5	
4									

图 38-12 群落生态位重叠估计数据定义格式

再在菜单方式下选择“专业统计”→“群落参数估计”→“生态位重叠”。执行计算分析后，系统输出各个指数测度如下：

Pianka 测度(源于 MACARTHUR-LEVINS 重叠指数)	$O_{12} = 0.2841$
PERCENTAGE OVERLAP MEASURE(百分率重叠测度)	$P_{12} = 26.1603$
MORISITA 相似性测度	$C_{12} = 0.2768$
MORISITA-HORN INDEX	$C_{H12} = 0.2759$
HORN'S INDEX OF NICHE OVERLAP	$R_{O12} = 0.4474$
HURLBERT 指数	$L_{12} = 1.0173$

参 考 文 献

丁岩钦. 1994. 昆虫数学生态学[M]. 北京: 科学出版社.

管致和等. 1983. 玉米螟分布型及其应用研究 I: 一代螟卵块在春玉米上的田间分布型[J]. 北京农业大学学报, 7: 1-12.

李典谟. 1987. 生态的多样性度量[J]. 生态学杂志, 6: 49-52.

邬祥光. 1985. 昆虫生态学的常用数学分析方法[M]. 北京: 中国农业出版社.

徐克学. 1999. 生物数学[M]. 北京: 科学出版社.

徐汝梅. 1987. 昆虫种群生态学[M]. 北京: 北京师范大学出版社.

阳含熙, 卢泽愚. 1983. 植物生态学的数量分类方法[M]. 北京: 科学出版社.

尹汝湛. 1954. 稻田三化螟集团分布的初步调查及在实践上的意义[J]. 昆虫学报, 4: 337-364.

赵士熙, 吴忠孚. 1989. 农作物病虫害数理统计 BASIC 程序库[M]. 福州: 福建科学技术出版社.

赵志模, 郭依泉. 1990. 群落生态学原理和方法[M]. 重庆: 科学技术文献出版社.

赵志模等. 1985. 桔园昆虫群落演替的初步研究[J]. 西南农学院学报, 3: 135-143.

Bliss C I, Owen A R C. 1958. Negative binomial distribution with a common k[J]. Biometrika, 45: 59-63.

Colwell R K, Futuyma D J. 1971. On the measurement of niche breadth and overlap[J]. Ecology, 52: 567-576.

Feinsinger P, Spear E E, Poole R W. 1981. A simple measure of niche breadth[J]. Ecology, 62: 26-32.

Hanski I. 1978. Some comments on the measurement of niche metrics[J]. Ecology, 59: 168-174.

Heltsh J F, Forrester N E. 1983. Estimating diversity using quadrat sampling[J]. Biometrics, 39: 1073-1076.

Heltsh J F, Forrester N E. 1983. Estimating species richness using the jackknife procedure[J]. Biometrics, 39: 1-11.

Heltsh J F, Forrester N E. 1985. Statistical evaluation of the jackknife estimate of diversity when using quadrat samples[J]. Ecology, 66: 107-111.

Horn H S. 1966. Measurement of "overlap" in comparative ecological studies[J]. Am. Nature, 100: 419-424.

Huhta V. 1979. Evaluation of different similarity indices as measures of succession in arthropod communities of the forest floor after

- clear-cutting[J]. *Oecologia*, 41: 11-23.
- Hurlbet S H. 1978. The measurement of niche overlap and some relatives[J]. *Ecology*, 59: 67-77.
- Iwao S. 1971. An approach to the analysis of aggregation pattern in biological populations[J]. *Statist. Ecol*, 1: 461-513.
- Iwao S. 1976. Relation of frequency index to population density and distribution pattern[J]. *Physiol. Ecol. Jap*, 17: 457-464.
- Kempton R A, Taylor L R. 1974. Log-series and log-normal parameters as diversity discriminates for the Lepidoptera[J]. *J. Anim. Ecol*, 43: 381-399.
- Krebs C J. 1989. *Ecological Methodology*[M]. New york: Harper & Row.
- MacArthur R H. 1965. Patterns of species diversity[J]. *Biol. Rev*, 40: 510-533.
- Moristia M. 1959. Measuring of interspecific association and similarity between communities[J]. *Mem. Fac. Sci. Kyushu Univ. Ser. E, Biol*, 3: 65-80.
- Peet R K. 1974. The measurement of species diversity[J]. *Annu. Rev. Ecol. Systematr*, 5: 285-307.
- Simpson E H. 1949. Measurement of diversity[J]. *Nature*, 163: 688.
- Smith E P. 1982. Niche breadth, resource availability, and inference[J]. *Ecology*, 63: 1675-1681.
- Zahl S. 1977. Jackknifing an index of diversity[J]. *Ecology*, 58: 907-913.

第 39 章 地理统计

39.1 空间自相关分析

空间自相关分析(spatial autocorrelation analysis)是检验某一空间变量的观测值是否显著地与其相邻空间点上的观测值相关联(Cliff, Ord, 1981)。如果相邻两点上的值均高或均低,则称其为空间正相关,否则称为空间负相关。空间自相关分析在地理统计学科中应用较多,现已有多种指数可以使用,但最主要的有两种指数,即 Moran 的 I 指数和 Geary 的 C 指数。

39.1.1 Moran 的 I 指数

与统计学上的相关系数相近,其值变化于 0 和 ± 1 之间。当 $I = 0$ 时代表空间无关,当 $I > 0$ 为正相关,而 $I < 0$ 时为负相关。其计算公式为

$$I = \frac{N \sum_{i=1}^N \sum_{j=1}^N W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^N \sum_{j=1}^N W_{ij} \right) \sum_{i=1}^N (x_i - \bar{x})^2} \quad (i \neq j)$$

式中 x_i 和 x_j 分别代表空间要素 x 在空间单元 i 和 j 中的观测值, \bar{x} 为 x 的平均值, W_{ij} 为相邻权重, N 为空间单元总数。根据上式得到空间自相关 Moran 的 I 系数之后,然后可以对它进行显著性检验。Moran 的 I 系数的统计学上的显著性可以通过一套公式进行估计,这些公式根据正态分布假设和随机分布假设推导出来。在正态分布假设条件下, Moran 的 I 系数的方差 $\text{var}_N(I)$ 为

$$\text{var}_N(I) = \left(\frac{1}{S_0^2(n^2 - 1)} (n^2 S_1 - n S_2 + 3 S_0^2) \right) - E_N(I)^2$$

然而,在随机分布的假设条件下, Moran 的 I 系数的方差 $\text{var}_R(I)$ 为

$$\text{var}_R(I) = \frac{\left\{ \left[n(n^2 - 3n + 3) S_1 - n S_2 + 3 S_0^2 \right] - \left[k \left((n^2 - n) S_1 - 2n S_2 + 6 S_0^2 \right) \right] \right\}}{(n-1)(n-2)(n-3) S_0^2} - E_R(I)^2$$

式中 n 为样本数量, $E_N(I) = \frac{-1}{n-1}$; $E_R(I) = E_N(I)$; $S_0 = \sum_{i=1}^n \sum_{j=1}^n W_{ij}$, 它是空间权重矩阵

之和; $S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (W_{ij} + W_{ji})^2$; $S_2 = \sum_{i=1}^n (W_{i.} + W_{.i})^2$, 即权重矩阵的第 i 列和第 i 行相加后

的平方和; $k = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 / n}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 / n \right)^2}$ 。Moran 的 I 自相关系数的标准偏差为

$$SD_{N \text{ 或 } R} = \sqrt{\text{var}_{N \text{ 或 } R}(I)}$$

Moran 的 I 系数的 z -score 等于

$$z = \frac{I - E_{N \text{ 或 } R}(I)}{\sqrt{\text{var}_{N \text{ 或 } R}(I)}}$$

一般情况, 我们可以应用随机分布的假设条件。但是, 正态分布条件假设下 Moran 的 I 自相关系数的标准偏差和 z 得分值计算简单得多。

39.1.2 Geary 的 C 指数

Geary 的 C 指数取值一般在 0~2, 大于 1 表示负相关, 等于 1 表示不相关, 小于 1 表示正相关, 其计算公式为

$$C = \frac{(n-1) \sum_{i=1}^N \sum_{j=1}^N W_{ij} (x_i - x_j)^2}{2 \left(\sum_{i=1}^N \sum_{j=1}^N W_{ij} \right) \sum_{i=1}^N \sum_{j=1}^N (x_i - \bar{x})^2} \quad (i \neq j)$$

式中 x_i 和 x_j 分别代表空间要素 x 在空间单元 i 和 j 中的观测值, W_{ij} 为相邻权重, n 为空间单元总数。根据公式得到空间自相关 Geary 的 C 系数之后, 然后可以对它进行显著性检验。Geary 的 C 系数的统计学上的显著性可以通过一套公式进行估计, 这些公式根据正态分布假设和随机分布假设推导出来。和 Moran 的 I 系数类似, 在正态分布和随机分布的假设条件下, Geary 的 C 系数的方差 $\text{var}_N(C)$ 和 $\text{var}_R(C)$ 分布为

$$\text{var}_N(C) = \left(\frac{1}{2(n+1)S_0^2} (2S_1 + S_2)(n-1-4S_0^2) \right)$$

令

$$VC_1 = (n-1)S_1(n^2 - 3n + 3 - (n-1)k)$$

$$VC_2 = \frac{1}{4} \left((n-1)S_2 \left(n^2 + 3n - 6 - (n^2 - n + 2)k \right) \right)$$

$$VC_3 = S_0(n^2 - 3 - (n-1)^2 k)$$

则

$$\text{var}_R(C) = \frac{VC_1 + VC_2 + VC_3}{n(n-1)(n-1)S_0^2}$$

式中所有的变量的定义和 Moran 的 I 的显著性测定中的公式相同。Geary 的 C 系数的显著性也可以通过和 Moran 的 I 系数显著性类似方法进行。实际上, 在 DPS 系统中, 同时给出了 Moran 的 I 系数和 Geary 的 C 系数后, 还给出了这两个系数分布在正态分布假定和随机分布假定条件下的显著性检验结果。

Griffith(1987)通过模拟试验注意到 Moran 的 I 系数和 Geary 的 C 系数呈负相关, 基本上是线性的趋势(图 39-1)。

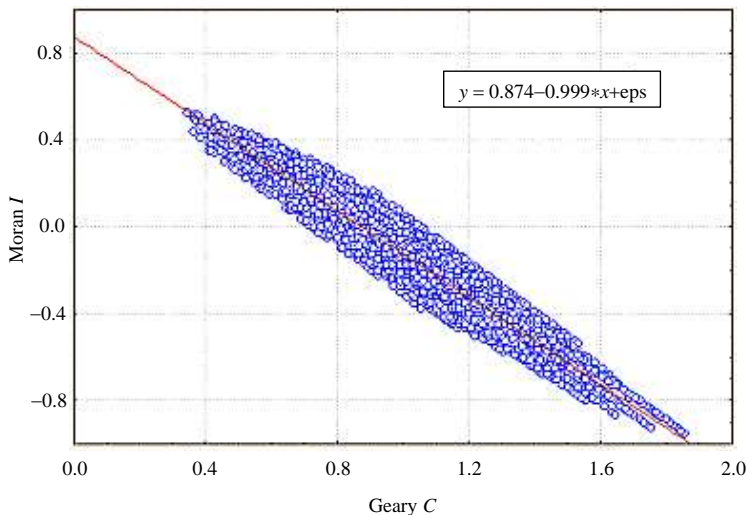


图 39-1 Moran 的 I 系数和 Geary 的 C 系数的关系

39.1.3 DPS 中空间自相关分析过程

DPS 系统中进行空间自相关分析的操作是在电子表格中, 一行一个样本, 每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应的观察值(Z)。然后选定为数据块(图 39-2)。

	A	B	C	D	E
1	1	经度	50-纬度	观察指标	
2	2	119.1	22	1.12	
3	3	119	22.4	0	
4	4	119.4	21.6	12.48	
5	5	119.6	22.1	0	
6	⋮	⋮	⋮	⋮	
7	67	121.3	20.9	30.01	
8	68	121.8	20.6	54.66	
9	69	121.4	21.4	83.42	
10					

图 39-2 实验半变异函数计算数据编辑、定义格式

计算结束后, 系统将会给出如表 39-1 和表 39-2。

表 39-1 Moran 的 I 系数及其显著性测定结果

序号	滞后距离	Moran I 系数	正态分布假定		P	随机分布假定		P
			方差	z -score		方差	z -score	
1	0.35403	0.45525	0.01709	3.59472	0.00032	0.01695	3.60989	0.00031
2	0.70796	0.16179	0.00400	2.79206	0.00524	0.00396	2.80370	0.00505
3	1.06189	0.00104	0.00293	0.29077	0.77123	0.00291	0.29196	0.77032
4	1.41582	-0.17531	0.00219	3.43145	0.00060	0.00217	3.44526	0.00057
5	1.76975	-0.10077	0.00223	1.82315	0.06828	0.00221	1.83075	0.06714
6	2.12368	-0.03931	0.00245	0.49733	0.61896	0.00243	0.49942	0.61749
7	2.47761	-0.07983	0.00305	1.17920	0.23832	0.00303	1.18405	0.23639
8	2.83154	0.08213	0.00399	1.53262	0.12537	0.00396	1.53878	0.12386
9	3.18547	0.04954	0.00715	0.75982	0.44736	0.00709	0.76276	0.44560
10	3.53940	0.31713	0.01513	2.69733	0.00699	0.01502	2.70785	0.00677
11	3.89333	0.23330	0.03188	1.38894	0.16485	0.03164	1.39427	0.16324
12	4.24726	-0.81525	0.11774	2.33306	0.01965	0.11679	2.34251	0.01915
13	4.60119	-0.33983	0.15020	0.83889	0.40153	0.14906	0.84211	0.39973

表 39-2 Geary 的 C 系数及其显著性测定结果

序号	滞后距离	Geary C 系数	正态分布假定		P	随机分布假定		P
			方差	z -score		方差	z -score	
1	0.35403	0.42344	0.02843	3.41973	0.00063	0.03168	3.23947	0.00120
2	0.70796	0.74607	0.00816	2.81141	0.00493	0.00935	2.62579	0.00864
3	1.06189	0.88982	0.00766	1.25870	0.20814	0.00902	1.16012	0.24600
4	1.41582	1.06682	0.00690	0.80419	0.42128	0.00826	0.73539	0.46210
5	1.76975	1.06343	0.00461	0.93459	0.35000	0.00529	0.87224	0.38308
6	2.12368	0.96227	0.00440	0.56862	0.56961	0.00496	0.53554	0.59228
7	2.47761	1.07444	0.00760	0.85393	0.39314	0.00891	0.78886	0.43019
8	2.83154	1.07640	0.01269	0.67831	0.49757	0.01518	0.62008	0.53521
9	3.18547	1.15113	0.03075	0.86189	0.38875	0.03752	0.78025	0.43524
10	3.53940	0.96116	0.06268	0.15513	0.87672	0.07633	0.14058	0.88820
11	3.89333	1.22307	0.14862	0.57865	0.56283	0.18211	0.52273	0.60116
12	4.24726	2.66825	0.35089	2.81627	0.00486	0.41779	2.58097	0.00985
13	4.60119	3.07361	0.70000	2.47843	0.01320	0.85776	2.23895	0.02516

分析时, 可将自相关系数与滞后距离尺度结合起来, 以分析不同尺度下的空间相关关系的变化, 这样的结果可以用尺度-自相关系数图表示, 可以直观地看出空间相关性随尺度的变化。

39.2 空间联系统计分析

Getis 和 Ord(1992)提出了一个空间统计量指标 $G_i(d)$, 它是在距离为 d 的范围内, 度量每一个 i 与每个 j 的相关程度空间统计量。 $G_i(d)$ 可定义如下:

$$G_i(d) = \frac{\sum_{j=1}^n w_{ij}(d)x_j}{\sum_{j=1}^n x_j}, \quad j \neq i$$

$$E(G_i(d)) = \frac{\sum_{j, j \neq i}^n w_{ij}E(x_j)}{\sum_{j, j \neq i}^n x_j}$$

$$= \left(\sum_{j, j \neq i}^n w_{ij} \right) \left(\frac{x_1}{n-1} + L + \frac{x_{i-1}}{n-1} + \frac{x_{i+1}}{n-1} + L + \frac{x_n}{n-1} \right) / \sum_{j, j \neq i}^n x_j$$

$$= \frac{W_i}{n-1}$$

$$\text{var}(G_i(d)) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)} \left(\frac{Y_{i2}}{Y_{i1}^2} \right)$$

其中 n 是观察的值个数, x_i 是在地点 i 的观察值, $\{w_{ij}\}$ 是一个对称的二值空间权重矩阵, 如果地点 j 与地点 i 的距离在 d 的范围之内(或者与 i 是邻接的), w_{ij} 的值为 1, 否则为 0,

$$W_i = \sum_{j, j \neq i}^n w_{ij}, \quad Y_{i1} = \sum_{j \neq i}^n x_j^2 \quad \text{且} \quad Y_{i2} = \sum_{j \neq i}^n x_j^2 / (n-1) - Y_{i1}^2$$

不像 Moran 的 I 系数那样的全局指标, G 统计量在很多情况下也可以用作局部空间关系的度量。若将 i 本身也列入计算范围之内, 此时公式则可改写为

$$G'_i(d) = \frac{\sum_{j=1}^n w_{ij}(d)x_j}{\sum_{j=1}^n x_j}$$

为了更好地解释, Ord 和 Getis(1994)定义了 $G_i(d)$ 如下的标准化形式。在这种情况下 ($i \neq j$)

$$Z(G_i) = \frac{G_i - E(G_i)}{\sqrt{\text{var}(G_i)}} = \frac{\sum_{j \neq i}^n w_{ij}(d)(x_j - \bar{x}_i)}{S_i \sqrt{w_i(n-1-w_i)/(n-2)}}$$

这里, x_i 是在地点 i 的观察值

$$\bar{x}_i = \frac{1}{n-1} \sum_{j, j \neq i}^n x_j$$

如前面定义的一样, $\{w_{ij}\}$ 是对称二值空间加权矩阵,

$$W_i = \sum_{j, j \neq i}^n w_{ij}(d), \quad S_i^2 = \frac{1}{n-1} W_i = \sum_{j, j \neq i}^n (x_j - x_i)^2$$

结果表明(Ord & Getis, 1994), 在原假设为 x_i 没有空间关系时, G_i 的排列分布接近于正态分布。当 $i=j$ 时, 该情况也是允许的, $Z(G'_i)$ 可以定义如下:

$$Z(G'_i) = \frac{G'_i - E(G'_i)}{\sqrt{\text{var}(G'_i)}} = \frac{\sum_{j=1}^n w_{ij}(d)(x_j - \bar{x})}{S \sqrt{w'_i(n-1-w'_i)/(n-1)}}$$

其中 w'_i 的定义与 W_i 相似, 除了允许 $i=j$ 之外,

$$w_i = \sum_{j=1}^n w_{ij}(d), \quad \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j, \quad S^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$$

当权重不是二值时, $Z(G_i)$ 可定义如图 39-3,

$$Z(G_i) = \frac{\sum_{j \neq i}^n w_{ij}(d)(x_j - w'_{i\bar{x}_i})}{S_i \sqrt{((n-1)S_{li} - w_i^2)/(n-2)}}$$

及

$$Z(G_i^*) = \frac{\sum_{j \neq i}^n w_{ij}(d)(x_j - w'_{i\bar{x}_i})}{S_i \sqrt{(nS'_{li} - w_i^2)/(n-1)}}$$

其中 $S_{li} = \sum_{j \neq i}^n w_{ij}^2$, $S'_{li} = \sum_{j=1}^n w_{ij}^2$, \bar{x}_i , \bar{x} , w_i , w'_i , S_i , S 和前面的定义一样。

上面定义的标准 G 统计量 $Z(G_i)$ 既可用在正效应上也可用在负效应上。可以在原假设 $H_0: G_i = 0$ 情况下进行 t 检验。显著的正的 $Z(G_i)$ 说明地点 i 被相对比较大的值包围, 因而显著的负的 $Z(G_i)$ 说明地点 i 被相对比较小的值包围。 $Z(G_i)$ 和 $Z(G_i^*)$ 的唯一区别在于 $Z(G_i^*)$ 不排除指数中观察值 i 。 G 统计量可以用来鉴别高值簇还是低值簇空间聚集模式。但是 G 统计量不能检测空间模式的相似性(正相关)或相异性(负相关)。

空间统计量指标 $G_i(d)$ 的计算 DPS 系统中进行空间统计量指标 $G_i(d)$ 的计算的操作是在电子表格中, 一行一个样本, 每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应的观察值(Z), 并用鼠标选中待分析的数据(图 39-3 中阴影部分)。

	A	B	C	D	E
1	Neighborhood	x	y	Price	
2	Bay Park	6	15	152	

3	Beaches	4	18	249
4	Bonita	17	8	297
5	Cardiff	2	36	260
6	Chula Vista	15	6	165
7	Chula Vista	17	4	184
8	Clairemont	6	20	192
9	College	14	16	138
10	Coronado	7	10	374
11	Del Cerro	14	18	187
12	Del Mar	5	32	309
13	Downtown	8	12	175
14	East San Diego	13	14	111
15	East San Diego	12	12	88
16	East San Diego	15	11	126
17	El Cajon	24	19	151
18	El Cajon	22	17	150
19	Encinitas	1	39	264
20	Hillcrest	8	14	225
21	Imperial Beach	11	1	164
22	Jamul	24	12	291
23	Kearny Mesa	8	19	138
24	Kensington	11	16	192
25	La Jolla	3	22	398
26	La Mesa	18	16	169
27	Lake Hodges	10	34	265
28	Lakeside	23	24	147
29	Lemon Grove	17	13	137
30	Logan Heights	11	10	84
31	Mira Mesa	8	28	162
32	Mission Valley	9	16	89
33	Mission Village	10	19	131
34	National City	13	8	99
35	North Park	11	14	152
36	Ocean Beach	3	14	229
37	Otay Mesa	15	2	126
38	Paradise Hills	16	9	117
39	Point Loma	3	12	338
40	Poway	17	32	236
41	R. Bernardo	15	35	191
42	R. Penasquitos	12	32	194
43	San Carlos	17	19	182
44	Santee	20	22	124
45	Scripps Ranch	13	29	270
46	Solana Beach	3	34	261
47	Spring Valley	20	13	150
48	Tierrasanta	13	20	221
49	University City	6	23	201
50				

图 39-3 Getis-Ord 分析数据编辑格式

然后执行专业统计下面的“地理统计”→“Getis-Ord 空间相关分析”功能，计算结束后，系统给出的结果有各个地点、以不同距离为半径时的 $G_i(d)$ 、 $G_i'(d)$ 值，以及“标准化”后的 $Z(G_i(d))$ 和 $Z(G_i'(d))$ 估计值。同时给出了进行处理计算时所依据的邻接结点数。如各个点在不同距离半径 3.29667, 6.59335, ..., 19.78004 时的 $Z(G_i'(d))$ 估计值的

结果如表 39-3 所示(其他输出结果的格式相同, 结果从略)。

表 39-3 $Z(G'_i(d))$ 估计值

$Z(G'_i(d))$	3.29667	6.59335	9.89002	13.18669	16.48337	19.78004
1	-0.53579	0.56568	-0.20220	-1.06859	-2.07041	-1.26307
2	0.53819	1.02149	1.17433	-0.36317	-0.01549	-0.50821
3	0.00451	-2.53191	-2.79859	-3.03635	-2.84696	-3.18402
⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	1.70543	0.66754	1.07116	1.63277	0.88088	0.04392

39.3 局部空间相关分析

根据 Anselin(1994)建议, 局部 Moran 和局部 Geary 统计量是两个供选择的比上面定义的 $G_i(d)$ 统计量有更多优点的局部相关指标。采用特定的排列组合方法, 可以对原假设为所有值在空间是否是随机分布的情况进行检验。

39.3.1 局部 Moran

对于每一个观察值, 局部 Moran 统计量 I 可以定义如下:

$$I_i(d) = Z_i \sum_{j \neq i}^n w_{ij} Z_j$$

其中观察值 Z_i 和 Z_j 都是标准形式(平均值为 0 方差为 1), 空间权重 w_{ij} 行标准形式。所以, I_i 是 Z_i 和周围点观察值平均数的积。

可以看出局部 Moran 统计量的平均值正好与相应的全局 Moran 的 I 统计量(Anselin, 1994)是一样的

$$I_i(d) = \frac{\sum_i^n \sum_{j \neq i}^n w_{ij} Z_i Z_j}{S^2 \sum_i^n \sum_{j \neq i}^n w_{ij}} = \frac{1}{n} \sum_i^n \left(Z_i \sum_{j \neq i}^n w_{ij} Z_j \right) = \frac{1}{n} \sum_i^n I_i(d)$$

$$S^2 = \sum_i^n Z_i^2 = 1, \quad \text{且} \quad \sum_i^n \sum_{j \neq i}^n w_{ij} = n$$

在“特定”的随机和排列方法中可能会得到 I_i 假显著水平(Anselin, 1994)。 Z_i 在地点 i 的观察值是固定的, 剩余的值在所有地点是随机排列的。在实际的计算中, 每一个重复取样的数据集可以从总体中随机和无重复的选择。 p 值的显著水平可以通过数据排列在数据集中占的比例获得, 仿效的 I_i 比实际的 I_i 值大(或小)或接近。 Z_i 在每一个被模仿的 I_i 中是固定的, p 值显著水平的计算可以简化成通过计算数据排列在数据集占的比例, 这使仿效的平均值比实际周围观察点 I 的平均值大(或小)或接近相等。试验的 p 值提供了

原假设地点 I 周围的观察值的平均值不是在极端(也就是说所有的值是随机分布在整个空间中)。

局部 Moran 的解释与 G 统计量相似。小 p 值(如 $p < 0.05$)说明地点 I 与周围地点比较大的值关联。大 p 值(如 $p > 0.95$)说明地点 I 与周围比较小的值关联。与地点 I 观察值和其他的统计量一起, 局部 Moran 可以用来鉴别可能存在的空间关系的类型。

39.3.2 局部 Geary 统计量

每个观察点 i 的局部 Geary 统计量可以定义如下(Anselin, 1994):

$$C_i(d) = \sum_{j \neq i}^n w_{ij} (Z_i - Z_j)^2$$

局部 Geary 统计量与全局 Geary 统计量关联情况如下:

$$C(d) = \frac{(n-1) \sum_i \sum_j^n w_{ij} (Z_i - Z_j)^2}{2nS^2 \sum_i \sum_j^n w_{ij}} = \frac{(n-1) \sum_i \sum_j^n w_{ij} (Z_i - Z_j)^2}{2n^2} = \frac{(n-1)}{2n^2} \sum_i C_i(d)$$

其中 S^2 和前面一样, Z_i 和 Z_j 是标准化的值, w_{ij} 是行标准化的空间权重矩阵的元素。

如果每一个观察值对全局统计量的贡献一样, 它的值应是 $I(d) = \frac{1}{n} \sum_{i=1}^n I_i$, 或者是

$$\bar{C} = \frac{2n}{n-1} \sum_{i=1}^n C_i, \text{ 每个 } C_i \text{ 统计量的显著性可以同样地用上面列出 } I_i \text{ 排列的方法来获得。}$$

与局部 Moran 不同, 局部 Geary 统计量是在地点 i 和它周围地点差异的加权平方和的度量。 p 值假的显著性的计算与局部 Moran 统计量类似。大 p 值(如 $p > 0.95$)说明在极端的 C_i 小, 这说明观察值和周围值有正的空间相关关系(例如, 相似性++或--), 小 p 值(如 $p < 0.05$)说明在极端的 C_i 大, 这说明观察值和周围值有负的空间相关关系(如相异性+ -或- +)。

通过总结局部 Moran 统计量和局部 Geary 统计量, 地点 i 和附近地点的空间关系鉴定标准如下:

I_i 的 p 值小, C_i 的 p 值大——正的空间关系(++);

I_i 的 p 值大, C_i 的 p 值大——正的空间关系(--);

I_i 的 p 值大, C_i 的 p 值小——负的空间关系(+ -);

I_i 的 p 值小, C_i 的 p 值小——负的空间关系(- +)。

当有外来作用时, C_i 的值会被夸大, 计算出的虚假的显著水平可能是有偏的。对局部 Geary 统计量可以选择的指示量是

$$C_i^2(d) = \sum_{j \neq i}^n w_{ij} |Z_i - Z_j|$$

同样的排列方法可以被用来作显著性检验。 C_i^2 的可以用与 C_i 同样的方式解释但是前者对外来作用有更高的抵干扰能力。

39.3.3 局部空间统计量，局部 Moran 指标和局部 Geary 指标的计算

DPS 系统中进行局部 Moran 指标和局部 Geary 指标的计算的操作是在电子表格中，一行一个样本，每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应的观察值(Z)。然后选定为数据块(图 39-3)。计算结束后，系统将会给出如下结果(仅列出局部 Moran 指标的计算结果)：

zLocal I	3.297	6.593	9.890	13.187	16.483	19.780	23.077	26.373	29.670	32.967	36.263	39.56
	0.20011	-0.34636	0.15479	0.67981	1.24660	0.84482	0.69054	0.46647	0.46647	0.46647	0.46647	0.46647
	0.01315	0.69150	0.85138	-0.30832	0.02475	-0.21074	0.28059	0.28059	0.28059	0.28059	0.28059	0.28059
	-1.43018	-4.43480	-4.58164	-4.70088	-4.31091	-4.71324	-3.40224	-3.34808	-1.88624	-1.37183	-0.67883	-0.67883
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0.25764	0.13203	0.21977	0.33880	0.31667	0.39242	0.66571	0.66571	0.66571	0.66571	0.66571	0.66571

39.4 实验半变异函数

当一个变量呈空间分布时，就称之为区域化变量，这种变量反映了空间某种属性的分布特征。矿产、地质、海洋、土壤、气象、水文、生态、温度、浓度等领域都具有某种空间属性。区域化变量具有双重性，在观测前区域化变量 $Z(x)$ 是一个随机场，观测后是一个确定的空间点函数值。

区域化变量具有两个重要的特征：一是区域化变量 $Z(x)$ 是一个随机函数，它具有局部的、随机的、异常的特征；二是区域化变量具有一般的或平均的结构性质，即变量在点 x 与偏离空间距离为 h 的点 $x+h$ 处的随机量 $Z(x)$ 与 $Z(x+h)$ 具有某种程度的自相关，而且这种自相关性依赖于两点间距离 h 与变量特征。在某种意义上，这就是区域化变量的结构性特征。

变异函数又称变差函数、变异矩，是地理统计分析所特有的基本工具。在一维条件下变异函数定义为：当空间点 x 在一维 x 轴上变化时，区域化变量 $Z(x)$ 在点 x 和 $x+h$ 处的值 $Z(x)$ 与 $Z(x+h)$ 差的方差的一半为区域化变量 $Z(x)$ 在 x 轴方向上的变异函数，记为 $\gamma(h)$ ，即

$$\begin{aligned} \gamma(x, h) &= \frac{1}{2} \text{var}[Z(x) - Z(x+h)] \\ &= \frac{1}{2} E[Z(x) - Z(x+h)]^2 - \frac{1}{2} \{E[Z(x)] - E[Z(x+h)]\}^2 \end{aligned}$$

在二阶平稳假设条件下，对任意的 h ， $E[Z(x+h)] = E[z(x)]$ 。因此上式可以改写为

$$\gamma(x, h) = \frac{1}{2} E[Z(x) - Z(x+h)]^2$$

从上式可知, 变异函数依赖于两个自变量 x 和 h , 当变异函数 $\gamma(x, h)$ 仅仅依赖于距离 h 而与位置 x 无关时, 可改写成 $\gamma(h)$, 即

$$\gamma(h) = \frac{1}{2} E[Z(x) - Z(x+h)]^2$$

设 $Z(x)$ 是属性 Z 在空间位 x 处的值, $Z(x)$ 为一区域化随机变量, 并满足二阶平稳假设, h 为两样本点空间分隔距离, $Z(x_i)$ 和 $Z(x_i+h)$ 分别是区域化变量在空间位置 x_i 和 x_i+h 处的实测值 ($i=1, 2, \dots, N(h)$), 那么根据上式的定义, 变异函数 $\gamma(h)$ 的离散公式为

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i+h)]^2$$

变异函数揭示了在整个尺度上的空间变异格局, 而且变异函数只有在最大间隔距离 $1/2$ 处才有意义。DPS 数据处理系统提供的求二维实验变差函数值的功能。在 DPS 中, 可以计算 0° 、 45° 、 90° 和 135° 角方向上的实验变差函数。实验变差函数的滞后距 h , 在程序运行时, 首先由程序根据样本点分布情况, 分别对 X 坐标和 Y 坐标进行排序, 自动算出。由屏幕提示、用户可以修改。

在实际工作中, 因观测点并不完全呈规则的矩形网格分布, 所以在求某个方向上的实验变差函数时, 由于观测点不一定完全位于这个方向的同一条直线上, 本程序采用角度允许误差限和距离允许误差限的方式进行调整。角度允许误差限在数据处理程序中给定。若角度允许误差限为 10° 。要计算 90° 方向上的实验变差函数, 则从某一点出发, 对位于 $80^\circ \sim 100^\circ$ 扇形区域内的任一点都可以看成是在 90° 方向上的点。距离允许误差限取基本滞后距的一半, 若基本滞后距为 400m , 要求两点间的距离为 1000m 的实验变差函数值, 则接点相距 $800 \sim 1200\text{m}$ 范围内的任一点都可以近似看成是这两点间的距离为 1000m 。

DPS 系统中计算半变异函数的操作是在电子表格中, 一行一个样本, 每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应观察值。然后选定为数据块(图 39-4)。

	A	B	C	D	E
1	1	经度	50-纬度	观察指标	
2	2	119.1	22	1.12	
3	3	119	22.4	0	
4	4	119.4	21.6	12.48	
5	5	119.6	22.1	0	
6	⋮	⋮	⋮	⋮	
7	67	121.3	20.9	30.01	
8	68	121.8	20.6	54.66	
9	69	121.4	21.4	83.42	
10					

图 39-4 实验半变异函数计算数据编辑、定义格式

计算时，系统将会给出如图 39-5 所示的用户操作界面。在这里，用户可以自行调整有关参数，如基本滞后距离，距离允许误差和角度允许误差。调整好之后，双击“画图”按钮，就可以得到当前设置下的半变异函数(图 39-5)。

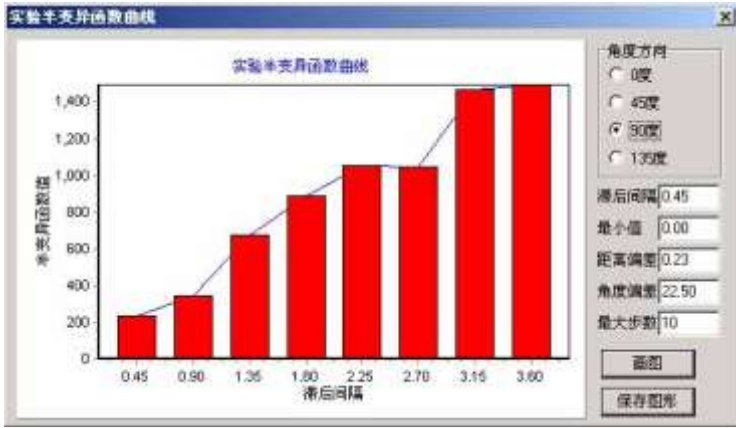


图 39-5 半变异函数计算的用户工作界面

最后确认后，点击图 39-5 右上角的关闭按钮，即返回操作的电子表格界面，并输出结果如下：

数据个数=69	平均值= 32.6932	方差= 779.1324	
基本滞后=0.45			
距离偏差限=0.23			
角度偏差限=22.50			
半变异函数计算结果			
Alpha= 0			
No.	滞后距离	Variogram	数据对数
1	0.20000	275.01050	3
2	0.48918	299.85552	58
3	0.92052	473.26980	99
4	1.36348	585.90474	105
5	1.79573	601.68676	108
6	2.25449	612.95724	76
7	2.68121	519.50553	56
8	3.10846	873.19956	22
Alpha= 45			
No.	滞后距离	Variogram	数据对数
1	0.19621	56.35225	12
2	0.49794	242.56537	48
...
7	2.61304	355.80146	24
Alpha= 90			
No.	滞后距离	Variogram	数据对数
1	0.15000	236.22365	2
2	0.48344	357.32835	63

...
9	3.52657	2738.54272	14
Alpha=	135		
No.	滞后距离	Variogram	数据对数
1	0.20013	108.98754	7
2	0.51359	251.43343	60
...
9	3.61276	1111.98072	20

39.5 协方差函数及相关系数

协方差又称半方差,是用来描述区域化随机变量之间的差异的参数。在概率理论中,随机向量 X_1 与 X_2 的协方差被定义为

$$\begin{aligned} C(x_1, x_2) &= E\{[Z(x_1) - m(x_1)] \cdot [Z(x_2) - m(x_2)]\} \\ &= E\{Z(x_1) \cdot Z(x_2)\} - E\{Z(x_1)\}m(x_2) - E\{Z(x_2)\}m(x_1) + m(x_1)m(x_2) \\ &= E\{Z(x_1) \cdot Z(x_2)\} - m(x_1)m(x_2) \end{aligned}$$

区域化变量 $Z(x)=Z(x_u, x_v, x_w)$ 在空间点 x 和 $x+h$ 处的两个随机变量 $Z(x)$ 和 $Z(x+h)$ 的二阶混合中心矩定义为 $Z(x)$ 的自协方差函数,即

$$\text{cov}\{Z(x), Z(x+h)\} = E[Z(x)Z(x+h)] - E[Z(x)]E[Z(x+h)]$$

区域化变量 $Z(x)$ 的自协方差函数也简称为协方差函数。一般来说,它是一个依赖于空间点 x 和向量 h 的函数。

设 $Z(x)$ 为区域化随机变量,并满足二阶平稳假设,即随机函数 $Z(x)$ 的空间分布规律不因位移而改变, h 为两样本点空间分隔距离或距离滞后, $Z(x_i)$ 为 $Z(x)$ 在空间位置 x_i 处的实测值, $Z(x_i+h)$ 是 $Z(x)$ 在 x_i 处距离偏离 h 的实测值 ($i=1, 2, 3, \dots, N(h)$), 根据协方差函数的定义公式,可得到协方差函数的计算公式为

$$C^\#(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - \bar{Z}(x_i)][Z(x_i+h) - \bar{Z}(x_i+h)]$$

在上面的公式中, $N(h)$ 是分隔距离为 h 时的样本点对的总数, $\bar{Z}(x_i)$ 和 $\bar{Z}(x_i+h)$ 分别为 $Z(x_i)$ 和 $Z(x_i+h)$ 的样本平均数,即

$$\bar{Z}(x_i) = \frac{1}{N} \sum_{i=1}^N Z(x_i), \quad \bar{Z}(x_i+h) = \frac{1}{N} \sum_{i=1}^N Z(x_i+h)$$

在公式中 N 为样本单元数。一般情况下 $\bar{Z}(x_i) \neq \bar{Z}(x_i+h)$ (特殊情况下可以认为近似相等)。若 $\bar{Z}(x_i) = \bar{Z}(x_i+h) = m$ (常数), 协方差函数可改写为如下:

$$C^\#(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} [Z(x_i)Z(x_i+h)] - m^2$$

式中, m 为样本平均数, 可由一般算术平均数公式求得, 即 $m = \frac{1}{N} \sum_{i=1}^N Z(x_i)$ 。由于协方差与相关系数有密切的关系, 因此通过协方差函数可以导出相关函数的计算公式为

$$\rho^{\#}(h) = \frac{C^{\#}(h)}{C^{\#}(0)}$$

式中 $C^{\#}(0)$ 为先验方差, 可以根据公式

$$C^{\#}(0) = \frac{1}{N(h)} \sum_{i=1}^N [Z(x_i)]^2 - m^2$$

计算, 式中 m 和上面协方差计算中的相同, 为样本平均数, N 为样本单元数。

协方差函数的计算和前面的半变异函数计算的数据格式相同。实际上, 在计算协方差函数的同时, 也计算出了半变异函数的计算结果。

39.6 变异函数理论模型的最优拟合

39.6.1 变异函数理论模型简介

前面计算的变异函数是实验变异函数(experimental variogram), 该变异函数可用变异曲线来表示。它是一定滞后距 a 的变异函数值 $\gamma(h)$ 与该 h 的对应图, 一个理想化的变异曲线应如图 39-6 所示。

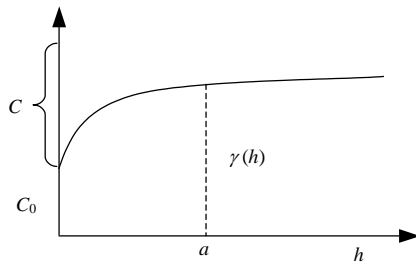


图 39-6 变异曲线示意图

图 39-6 中的曲线包括如下几部分: C_0 称为块金效应(nugget effect), 表示 h 很小时两点间观察指标的变化; a 称为变程(range), 当 $h \leq a$ 时, 任意两点间的观测值有相关性, 这个相关性随 h 的变大而减小, 当 $h > a$ 时就不再具相关性, a 的大小反映了研究对象中某一区域化变量(如品位)的变化程度。从另一个意义看, a 反映了影响范围, 估计 C 称为总基台值, 它反映某区域化变量在研究范围内变异的强度, 它是最大滞后距的可迁性变异函数的极限值。当 h 趋于无穷大时

$$\gamma(\infty) = C(0) = \text{var}[Z(x)] = C$$

即当 h 趋于无穷大时, 变异函数值近于先验方差 $C(0)$, 当无块金效应(常数) C 时, $\gamma(h) = C$, 当有块金效应时, $\gamma(h) = C + C_0$; 而 C 称为基台值(sill), 它是先验方差与块金效应(常数)之差: $C = \gamma(h) - C_0$ 。

尽管变异函数有助于解决一个矿床某区域化变量(例如品位)的变化特征及结构性状, 但它纯粹是一个数据的概括技术, 当定量地描述全矿床的特征时, 有关全矿床的变异结构还必须借助于推断, 这个过程类似于用样品值构制直方图, 再从直方图推断全矿床的理论分布。如果已绘制了试验半变异曲线, 那么为了要得到最后的结论, 就必须给试验半变异曲线配以相应的理论模型, 这些理论模型将直接参与克里格(Kriging)计算或其他地理统计学研究。

如同经典统计学那样, 理论变异函数也有几个简单的模型。DPS 提供了如下 6 个常用的理论模型, 其基本公式简介如下:

(1) 球状模型: 一般公式为

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C \left(\frac{3h}{2a} - \frac{1}{2} \frac{h^3}{a^3} \right), & 0 < h \leq a \\ C_0 + C, & h > a \end{cases}$$

该模型在零点处($h=0$), 切线的斜率为 $3C/2a$, 切线到达 C 值的距离为 $2a/3$ 。

(2) 高斯模型: 其公式为

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C(1 - e^{-h/a}), & h > 0 \end{cases}$$

(3) 指数模型:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C_0 + C(1 - e^{-h^2/a^2}), & h > 0 \end{cases}$$

(4) 球状套合模型:

$$\gamma(h) = \begin{cases} 0, & 0 \\ C_0 + \frac{3}{2} \left(\frac{C_1}{a_1} + \frac{C_2}{a_2} \right) h - \frac{1}{2} \left(\frac{C_1}{a_1^3} + \frac{C_2}{a_2^3} \right) h^3, & 0 < h \leq a_1 \\ C_0 + C_1 + C_2 \left(\frac{3h}{2a_2} + \frac{1}{2} \left(\frac{h}{a_2} \right)^3 \right), & a_1 < h \leq a_2 \\ C_0 + C_1 + C_2, & h > a_2 \end{cases}$$

(5) 球状+指数套合模型:

$$\gamma(h) = \begin{cases} 0, & 0 \\ C_0 + C_1 \left(\frac{3}{2} \frac{h}{a_1} - \frac{1}{2} \frac{h^3}{a_1^3} \right) + C_2 (1 - e^{-\frac{h}{a_1}}), & 0 < h \leq a_1 \\ C_0 + C_1 + C_2 (1 - e^{-\frac{h}{a_1}}), & a_1 < h \leq a_2 \\ C_0 + C_1 + C, & h > a_2 \end{cases}$$

(6) 线性有基台模型:

$$\gamma(h) = \begin{cases} C_0, & h = 0 \\ C_0 + Ah, & 0 < h \leq a \\ C_0 + C, & h > a \end{cases}$$

39.6.2 变异函数理论模型参数拟合

半变异函数理论模型参数估计, 实质上是一元非线性回归分析。DPS 系统采用加权的非线性最小二乘法(麦夸特法)进行估计, 其数据格式入图 39-7。

	A	B	C	D	E
1		Alpha= 0			
2		滞后距离	Variogram	数据对数	
3		0.28734	259.69786	16	
4		0.46124	387.40028	19	
5		0.66846	328.77115	38	
6		0.94516	436.37728	46	
7		1.14652	631.01630	39	
8		1.38201	665.53265	52	
9		1.60368	776.26233	47	
10		1.85090	503.27127	40	
11		2.06597	610.13070	38	
12		2.29881	600.24048	42	
13		2.53747	534.31003	26	
14		2.76581	528.17883	21	
15					

图 39-7 理论变异函数拟合数据编辑、定义及操作的用户界面

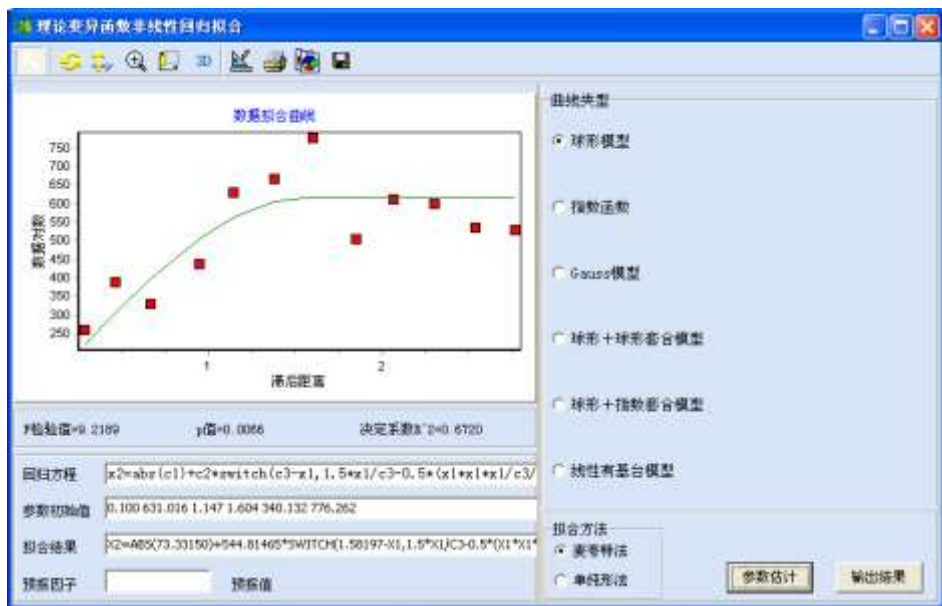
理论半变异模型的统计检验, 可类似线性回归分析, 将总平方和分解成残差平方和和回归平方和两部分, 即

$$\sum_{i=1}^n (\gamma_i - \bar{\gamma})^2 = \sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2 + \sum_{i=1}^n (\hat{\gamma}_i - \bar{\gamma})^2$$

这样可以构造一个 F 统计量

$$F = \frac{\sum_{i=1}^n (\hat{\gamma}_i - \bar{\gamma}_i)^2 / (k-1)}{\sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2 / (n-k)}$$

进行统计检验，并可计算出拟合结果的决定系数 R^2 。这些指标都在计算过程中及时反馈到一元非线性回归用户工作界面如下：



该用户对话界面绘出了理论变差函数拟合曲线图，用户可以根据理论变差函数曲线的拟合情况及有关统计检验结果，选择适当的半变函数模型，直到得出理想的理论变差函数。

39.7 交叉验证

由于实验变异函数并不是连续的图形，且无法满足条件半正定的性质，因此在实际应用时必须套配成连续性的理论变异函数模式后方可应用。理论变异函数模式的套配即是利用数值方法来选取固定模式下的最佳参数，因此，为了验证所选取的变异函数模型的优劣及利用克立格估计法推估时所做的假设是否合理，可以利用交叉验证法(cross validation)来进行鉴定。

半变异函数模型的验证往往因为观察资料的数量太少，故常常采用交叉验证法来进行鉴定，其步骤如下：

- (1) 利用所有 n 个观察资料可求得一代表性的半变异函数模型；
- (2) 自观察资料点中任意取出一观察值 Z_1 ，以其余 $n-1$ 个观察资料来进行克立格推估，则可求得该观察点 Z_1 的估计值；

(3) 计算该观察点的估计误差及克立格变异方差;

(4) 将该点 Z_1 放回, 并取出另一个观察点, 记为 Z_2 ; 重复上述步骤, 直到所有观察点均完成克立格估计及其估计误差与克立格方差的计算;

(5) 利用上述计算结果分别鉴定其无偏性与一致性。

无偏性与一致性的判断准则如下:

鉴定研究中所配套的理论半变异函数模型的优劣及利用克立格估计法推估时所做的假定是否合理, 可利用推估结果的无偏性及一致性来鉴定。

(1) 无偏性(unbiasness)。当克立格平均误差, 即估计值与观察值的期望值越接近 0 时, 表示其无偏估计的效果越佳。无偏性(KAE)的计算公式为

$$KAE = \frac{1}{n} \sum_{i=1}^n (Z_i^* - Z_i) \approx 0$$

(2) 一致性(coherence)。若克立格估计误差的平方与克立格方差的比值的期望值越趋近于 1 时, 则表示模型愈佳。其一致性(KRMSE)的计算公式为

$$KRMSE = \frac{1}{n} \sum_{i=1}^n \left[\frac{(Z_i^* - Z_i)^2}{\sigma_k^2} \right] \approx 1$$

该程序根据上述原理, 用普通克立格法、对观测点上的数据进行估值、求出每个观测点上的估计值、估计方差和偏差(估计值与观测值之差), 并绘出偏差直方图及当前的 KAE 值和 KRMSE 值。并可调整有系统自动优化有关模型参数, 使之更为接近。

	A	B	C	D	E	F	G	H	I
1	德清县	120	19.50	59.05					
2	长兴县	119.90	19	50.14					
3	余姚市	121.10	20	54.10					
4	慈溪市	121.20	19.90	85.28					
5	上虞市	120.80	20	62.41					
6	鄞县	121.50	20.20	53.60					
7	绍兴县	120.50	20	48.17					
8	定海区	122.10	20	45.89					
9	普陀区	122.20	20.10	32.85					
10	岱山县	122.10	19.80	37.78					
11	玉环县	121.20	21.90	48.01					
12	嵊泗县	122.40	19.30	18.64					
13	洞头县	121.10	22.20	0					
14	瑞安市	120.60	22.30	34.82					
15	温岭市	121.30	21.70	57.93					
16	平阳县	120.50	22.40	27.60					
17	苍南县	120.30	22.50	28.51					
18	乐清市	120.90	21.90	30.01					
19	瓯海区	120.60	22	37.51					
20	临海市	121.10	21.20	29.22	X	a	b	C0	
21	黄岩市	121.20	21.40	44.80	Y	1.4	30.16	553.9	
22	三门县	121.30	20.90	30.01		4.82	194.58	1551.1	
23	象山县	121.80	20.60	54.66					
24									

图 39-8 交叉验证分析的数据编辑、定义用户界面

在搜索信息点时,采用等分象限的技术,即将每个象限分成若干份,在每一份中最多只取 n_0 个点作为信息点参与估值。程序运行时、用户可以根据屏幕提示,确定估值的搜索方位数(4, 8, 12 或 16),输入象限等分(角)数(1, 2, 3 或 4),即将每个象限一等分、表示 4 方位搜索;如将每个象限二等分,表示 8 方位搜索……程序设计成最多可以将每个象限 4 等分,即 16 方位搜索。

DPS 系统中进行交叉验证分析,其操作是在电子表格中,一行一个样本,每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应的观察值,然后选定为数据块(数据块的左边一列可以放各个样本点的名称,但不要鼠标选进来,系统会自动将名称放到交叉验证的结果里面去的)。如果将半变异函数模型参数放在电子表格里的话,第一行放 X 轴的值,第二行放 Y 轴的值。每行分别放入参数 a , C 和 C_0 的值,然后再按下 Ctrl 键的状态下用鼠标选定(图 39-8)。

计算时,系统将会给出如图所示的用户操作界面(图 39-9)。

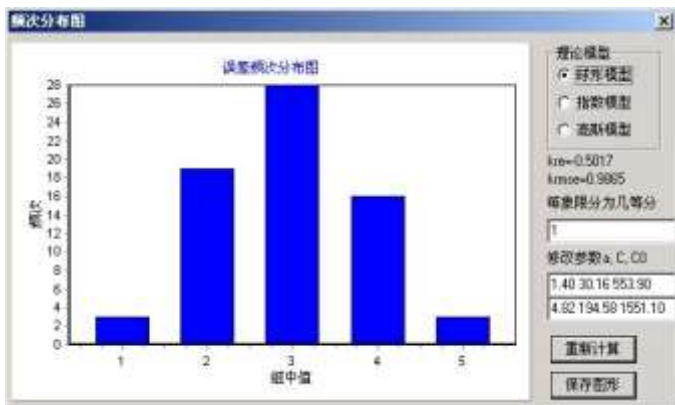


图 39-9 交叉验证分析用户参数对话框

在这里,用户可以自行调整有关参数,调整好之后,双击“理论模型优化”按钮,就可以得到当前设置下的新的理论半变异函数参数、交叉验证残差直方图及无偏性指标 KRE 和一致性指标 KRMSE 的值。

完成调整、模型优化之后,点击用户操作界面右上方的关闭按钮,即输出结果如下:

X 方向:	a=1.40	基台 $c+c_0=584.0600$	块金=553.9000			
Y 方向:	a=4.82	基台 $c+c_0=1745.6800$	块金=1551.1000			
套 合:	a=1.40	基台 $c+c_0=1164.8700$	块金=1052.5000			
每个象限分 1 等分角						
样本个数=69						
偏差均值 KRE=-0.5017						
偏差标准差=244.6001						
一致性系数 KRMSE=0.9865						
克立格估计结果						
序号	样本号	Zi 值	估计值	估计方差	偏差	参与估计值的样本

1	龙泉市	1.1200	1.1568	1431.2523	-0.0368	云和县	庆元县	遂昌县	
2	庆元县	0.0000	0.5852	1635.5725	-0.5852	景宁县	龙泉市		
3	松阳县	12.4800	16.2657	1340.9978	-3.7857	丽水市	云和县	遂昌县	兰溪市
...
69	椒江市	83.4200	44.2874	1429.5283	39.1326	温岭市	黄岩市	三门县	

39.8 克里格插值

克里格法，也称空间局部估计或空间局部插值，是建立在变异函数理论及结构分析基础之上，在有限区域内对区域化变量的取值进行无偏最优估计的一种方法。克里格法最大限度地利用了空间取样所提供的各种信息。在估计未知样点的数值时，它不仅考虑了落在该样点的数据，而且还考虑了邻近样点的数据；不仅考虑了待估计样点与邻近已知点的空间位置，而且考虑了各个邻近样点彼此之间的相互关系。

克里格法是将任一个点的估计值通过该点影响范围内的 n 个有效样本值 $Z(x_i)$ 的线性组合得到，即

$$Z_v^* = \sum_{i=1}^n \lambda_i Z(x_i)$$

式中， λ_i 是与样点观察 $Z(x_i)$ 有关的加权系数，用来表示各个样点值 $Z(x_i)$ 对估计值 Z_v^* 的贡献。对于任一给定的区域和数据信息 $Z(x_i)$ ， $i=1, 2, \dots, n$ ，存在一组加权系数 λ_i 。如果使得估计方差为最小，其区域内的真值就能在最小的可能置信区间内产生。这种给出最佳线性无偏估计权系数的即为克里格法。

在 DPS 系统中进行克里格插值，其操作是在电子表格中，一行一个样本，每个样本包含经度(X 轴方向)、纬度(Y 轴方向)和相应的观察值。然后选定为数据块(图 39-2)。计算时，系统将会给出如图 39-10，图 39-11 所示的用户操作界面。

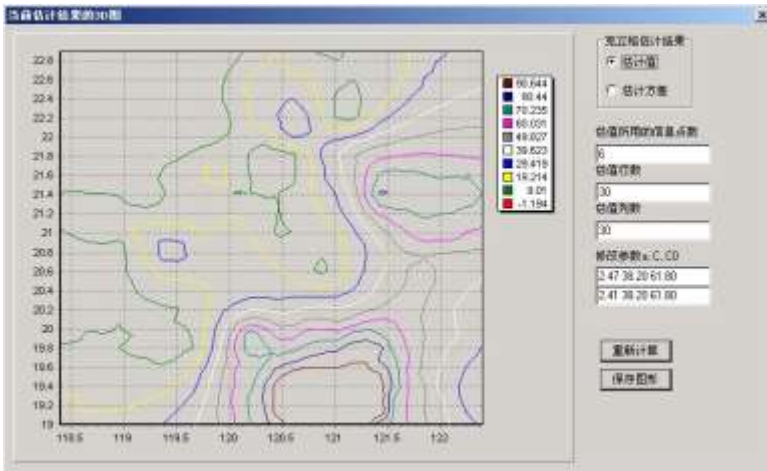


图 39-10 克里格插值用户操作界面(a.估计值)

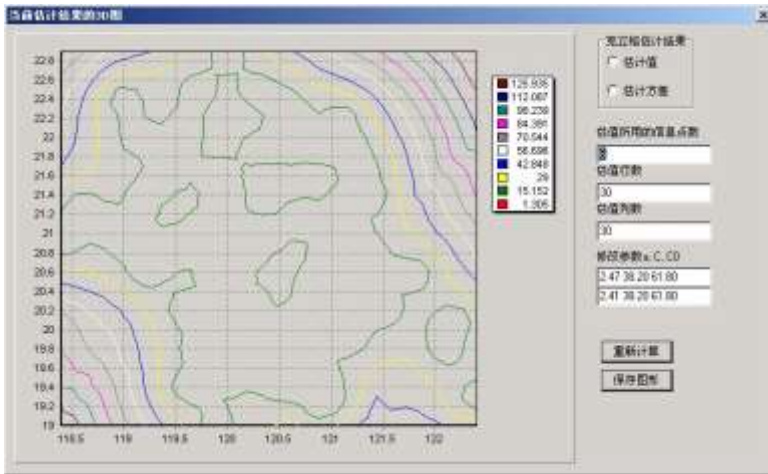


图 39-11 克里格插值用户操作界面(b.估计方差)

在这里，用户可以自行调整有关参数，如估计所用的信息点数，插值后生成的数据矩阵的行、列数，并以图形方式显示出来。

参 考 文 献

- 陈述彭, 鲁学军, 周成虎. 2000. 地理信息系统导论[M]. 北京: 科学出版社.
- 黄杏元, 马劲松, 汤勤. 2001. 地理信息系统概论(修订版)[M]. 北京: 高等教育出版社.
- 孙洪泉等. 1997. 实用地质统计学程序集[M]. 北京: 地质出版社.
- 王政权. 1999. 地统计学及在生态学中的应用[M]. 北京: 科学出版社.
- 徐建华. 2002. 现代地理学中的数学方法(第2版)[M]. 北京: 高等教育出版社.
- 於崇文. 1980. 数学地质的方法与应用[M]. 北京: 冶金工业出版社.
- A Stewart Fortheringham, Chris Brunsdon, Martin Charlton. 2000. Quantitative geography, perspective on spatial data analysis, SAGE Publications Ltd, 6 Bonhill street London EC2A 4PU.
- Arc/info. 中国技术咨询与培训中心[M]. 1995. Arc/info 地理信息系统应用教程.
- H Wackernagel. 1998. Multivariate Geostatistics. revised edition. Springer-Verlag.
- M L. Stein. 1999. Interpolation of Spatial Data[M]. Springer-Verlag.
- N Cressie. 1993. Statistics for Spatial Data, revised edition. Wiley.
- P J Diggle. 1983. Statistical Analysis of Spatial Point Patterns[M]. Academic Press.
- Paul A Longley, Michael F Goodchild, David J Maguire, David W Rhind. 2001. Geographic Information System and Science. John Wiley & Sons, Ltd. Baffins Lane, Chichester, West Sussex PO19 1UD, England.

第 40 章 品种比较试验

新引入或新育成的品种或品系，在大面积推广之前，都需要经过区域试验来鉴定其产量水平和适应性，为其推广和合理利用提供依据。

区域试验中需要研究的主要效应有：① 品种效应，即品种的产量或品质效应，因区域试验中供试品种是一定的，故品种效应是固定型效应；② 地区效应，即地区之间土壤类型、耕作制度、生产水平和管理方法等可以预见的环境差异对品种的影响效应，一般亦属固定型效应；③ 年际效应，即不同年份的温度、雨量、偶然性灾害等难以预知的随机性环境差异对品种的影响效应，一般属随机型效应；④ 品种与地区的互作效应，即研究品种对于可预见的环境差异是否具有特殊的适应性，若品种与地区互作效应显著，说明品种对地区有特殊适应性，反之则说明适应性广泛，故该互作效应一般也属固定型效应；⑤ 品种与年份的互作效应，反映品种对难以预测的环境差异是否有特殊适应性，一般属随机型效应；⑥ 品种×地点×年份的互作效应，反映品种与地点的互作效应是否随难以预知的环境变化而改变，因此该效应属于随机型效应。

评定一个品种的应用价值，主要考虑以下三个效应值：① 品种效应；② 品种×地区点的互作效应及方差；③ 品种×年份的互作效应及方差。品种效应显著而互作效应小的品种是具有广泛适应性的丰产型品种，适于大面积推广，而互作效应显著的品种具有特殊适应性(如对环境条件有特殊要求)，只能在特定地区推广才能发挥增产作用。

40.1 一年多点试验稳定性分析

40.1.1 Eberhart 和 Russell 模型

Eberhart 和 Russell(1966)认为品种相对于环境的稳定性分析，简单采用回归系数还不够，建议采用回归系数和校正的离回归方差 σ_{di}^2 作为测定稳定性的两个参数，它们可由以下的线性模型确定：

$$y_{ij} = \mu_i + \beta_i I_j + \delta_{ij}$$

其中 y_{ij} 表示第 i 个品种在第 j 个环境中的平均产量； μ_i 表示第 i 个品种在所有环境中的平均产量； β_i 表示第 i 个品种对各种变化环境反应的回归系数； δ_{ij} 表示第 i 个品种在第 j 个环境的回归离差； I_j 表示第 j 个环境的环境指数，等于 j 环境中所有品种的平均产量与总平均产量之差，即

$$I_j = \sum_i y_{ij} / v - \sum_i \sum_l y_{il} / vl, \quad \text{且} \quad \sum_i I_j = 0$$

第一个稳定性参数 β_i 的估计值 b_i 可用通常的回归方法求得

$$b_i = \frac{\sum_j y_{ij} I_j - \sum_j y_{ij} \sum_j I_j / l}{\sum_i I_j^2 - \left(\sum_j I_j \right)^2 / l}$$

因为 $\sum_i I_j = 0$ ，所以

$$b_i = \sum_j y_{ij} I_j / \sum_i I_j^2$$

为了估计另一个稳定性参数——校正的离回归方差 σ_{di}^2 ，首先根据表 40-1 作方差分析。

表 40-1 Eberhart 和 Russell 模型方差分析

变异来源	自由度	平方和	均方
总变异	$ev-1$	$\sum_i \sum_j y_{ij}^2 - C$	
品种间	$v-1$	$\frac{1}{l} \sum_i y_i^2 - C$	MS_1
环境间+(品种×环境)	$v(e-1)$	$\sum_i \sum_j y_{ij}^2 - \sum_i y_i^2 / l$	
环境(线性)	1	$\frac{1}{v} \left(\sum_i y_j I_j \right)^2 / \sum_j I_j^2$	
品种×环境(线性)	$v-1$	$\sum_i \left[\left(\sum_j y_{ij} I_j \right)^2 / \sum_j I_j^2 \right] - \text{环境线性 SS}$	MS_2
合并离差	$v(e-2)$	$\sum_i \sum_j \delta_{vj}^2$	MS_3
品种 1	$e-2$	$\sum_j \delta_{1j}^2$	
品种 2	$e-2$	$\sum_j \delta_{2j}^2$	
⋮	⋮	⋮	
品种 v	$e-2$	$\sum_j \delta_{vj}^2$	
试验误差	$e(v-1)(r-1)$		

然后按下式求出第二个稳定性参数 σ_{di}^2 的估计值 S_{di}^2 ：

$$S_{di}^2 = \left[\sum_j \hat{\delta}_{di}^2 / (l-2) \right] - S_e^2 / r$$

式中 S_e^2 / r 为 y_{ij} 的误差均方, $\sum_j \hat{\delta}_{di}^2 / (l-2)$ 为第 i 个品种的离回归均方。由于 S_e^2 / r 为一个常数, 比较品种稳定性实际上只比较前一部分的大小。式中第一项可能小于第二项, 即 S_{di}^2 可能为负值, 因此 S_{di}^2 仅作为一个指标值。实际上, S_{di}^2 是减去 y_{ij} 的误差均方后的离回归方差校正值, 其中离差 δ_{ij} 由下式估算

$$\sum_j \hat{\delta}_{ij}^2 = \left(\sum_j y_{ij}^2 - y_i^2 / l \right) - \left(\sum_j y_{ij} I_i \right)^2 / \sum_j I_j^2$$

方差分析模式如表 40-1 所示。

对于品种平均数之间的差异, 可用 $F = MS_1 / MS_3$ 测验无效假设 $H_0: \mu_1 = \mu_2 = \dots = \mu_t$ 是否成立。对于各个回归系数间的差异, 则用 $F = MS_2 / MS_3$ 来近似测验无效假设 $H_0: \beta_1 = \beta_2 = \dots = \beta_v$ 。若测验显著, 则表明参试品种间的稳定性有显著差异, 需继续对回归系数进行 t 测验和 F 测验。而当测验不显著时, 则一般无需进行下述的 t 测验和 F 测验。

对于各回归系数平均值(其值为 1)的差异采用 t 测验, $t = (b_i - 1) / SE(b)$, 这里 $SE(b) = \sqrt{MS_3 \sum_i I_j^2}$ 。

每个品种校正的离回归方差的显著性测验公式为

$$F = \left[\sum_j \hat{\delta}_{ij}^2 / (l-2) \right] / (S_e^2 / r)$$

关于两个稳定性参数的应用, Eberhart 等认为回归系数 b_i 可预测品种在不同环境下的表现。回归系数小于 1 的品种或杂交系, 其产量水平较稳定, 在不利条件下种植也有一定的产量; 回归系数大于 1 的品种在有利条件下有好的表现, 但不利条件下表现较差。校正的离回归方差 S_{di}^2 则衡量利用回归方程进行预测的有效性。 S_{di}^2 等于零, 说明可利用回归方程进行有效的预测, S_{di}^2 与零的差异显著, 则线性预测效果差。在生产过剩的地区, 应着重探求稳定性好的品种, 而不以高产为其主要目标。在环境条件好的地区, 人们常常要求特别适应于该地区的高产品种。但在一般条件下, 育种工作者往往希望一个品种在各种环境中相对稳定, 其产量不低于平均产量, 这样的品种应该有较高的产量和相当程度的稳定性。由于稳定性好的品种($b < 1$)其产量往往低于平均产量, 因此, 一般将具有 $b=1$ 和 $S_{di}^2=0$ 的品种定义为稳定性品种。

40.1.2 George C. C. Tai 模型

George C C Tai(1971)提出的方法是根据以下模型来进行分析的:

$$y_{ijk} = \mu + g_i + l_j + (gl)_{ij} + b_{jk} + e_{ijk} \quad (40.1)$$

其中 y_{ijk} 表示第 i 个品种在第 j 个环境、第 k 次重复的平均产量； μ 表示所有品种在所有环境、所有重复中的总平均产量； g_i 表示第 i 个品种的遗传效应 ($i=1, 2, \dots, v$)； l_j 表示第 j 个环境的环境效应 ($j=1, 2, \dots, n$)； $(gl)_{ij}$ 表示第 i 个品种与第 j 个环境的交互作用效应； b_{jk} 表示第 j 个环境内第 k 次重复的效应 ($k=1, 2, \dots, r$)； e_{ijk} 表示第 i 个品种在第 j 个环境、第 k 次重复的随机误差。

在方差分析中，假定品种效应为固定模型、重复和环境效应为随机模型，那么有

$$\sum_{i=1}^v g_i = \sum_{i=1}^v (gl)_{ij} = 0, \quad \sum_{j=1}^n l_j \neq 0, \quad \sum_{j=1}^n (gl)_{ij} \neq 0, \quad \sum_{j=1}^n \sum_{k=1}^r b_{jk} \neq 0$$

在稳定性分析中，进而把品种与环境的交互作用项分割成两部分：一部分是环境效应的线性部分，另一部分是线性回归离差。以 α_i 表示第 i 个品种的线性回归系数， δ_{ij} 表示第 i 个品种在第 j 个环境的线性回归离差，于是有

$$(gl)_{ij} = \alpha_i l_j + \delta_{ij} \quad (40.2)$$

把(40.2)式代入(40.1)式，则可改写为

$$y_{ijk} = \mu + g_i + l_j + \alpha_i l_j + \delta_{ij} + b_{jk} + e_{ijk} \quad (40.3)$$

品种效应固定，重复和环境效应为随机的混合模型方差分析，其品种效应间的差异显著性可以用 $F = MS_v / MS_{vl}$ 进行测验。而遗传型与环境交互作用的显著性则可用 $F = MS_{vl} / MS_e$ 进行测验。并按修改后的(40.3)模式，品种与环境交互作用平方和能够进一步分解为两个组成部分，即直线回归平方和和直线回归离差平方和。但是这种分解有时并非易事，而在实际计算中往往采用的是：只要品种与环境交互作用均方的 F 测验显著时，可直接从试验中估计每一个品种的稳定性参数 α_i 和 λ_i 。第 i 个品种的参数 α_i 和 λ_i 的估计值

$$\hat{\alpha}_i = \frac{S_{l(gl)_i}}{(MS_l - MS_b) / vr}, \quad \hat{\lambda}_i = \frac{S_{(gl)_i}^2 - \hat{\alpha}_i S_{l(gl)_i}}{(v-1)MS_e / vr}$$

这里 MS_l 、 MS_b 和 MS_e 分别表示环境效应均方、环境内重复间均方和误差均方， α_i 用以量测第 i 个品种对环境效应的线性响应，而 λ_i 是表示以误差均方为单位的线性回归离差。 S_l^2 是环境效应样本方差， $S_{(gl)_i}^2$ 是第 i 个品种与 n 个环境互作效应的样本方差， $S_{l(gl)_i}$ 是环境与互作效应间样本协方差。

Tai(1971)提出的品种稳定性分析方法，也是根据回归分析的方法提出对第 i 个品种的两个稳定性参数：

- (1) 对环境效应的直线响应，采用参数 α_i 测定；
- (2) 对直线响应的离差，采用另一个参数 λ_i 测定。

Tai 把具有 $\hat{\alpha} = 0, \hat{\lambda} = 1$ 的品种视为具有平均稳定品种, 而把具有 $\hat{\alpha} = -1, \hat{\lambda} = 1$ 的品种称为完全稳定品种。

通过考虑第 i 个品种成对效应 l_i 和 $(gl)_{ij}$ 间的相关性, 可以得到 $\alpha_i = 0$ 时概率水平 P 的预测区间。记 r 为 l_i 与 $(gl)_{ij} (j = 1, 2, \dots, n)$ 的相关系数, 则

$$\left[(n-2)r^2 / (1-r^2) \right]^{1/2}$$

服从自由度为 $df=n-2$ 的 t 分布。把 t 值表中自由度为 $df=n-2$, 显著水平为 $\alpha=1-p$ 的值记作 t_α , 可得到对于 $\alpha_i = 0$ 时的两个置信上、下限, 如下式所示:

$$\pm t_\alpha \left\{ \frac{\lambda(v-1)MS_e MS_l}{(MS_l - MS_b) \left[(n-2)MS_l - (t_\alpha^2 + n-2)MS_b \right]} \right\}^{1/2}$$

上述置信限是一个 λ 的函数。

对于 $\lambda_0 = 1$, 在概率水平 P 时的置信区间为

$$\frac{1}{F_{\alpha(n_2, n_1)}} \leq \lambda_0 \leq F_{\alpha(n_1, n_2)}$$

式中 $\alpha = (1-p)/2, n_1 = n-2, n_2 = n(v-1)(r-1)$ 。

对于 $\lambda_0 > 1$ 时的置信上限为 $F'_{\alpha(n_2, n_1)} = \lambda_0 F_{\alpha\left(\frac{n_1 \lambda_0^2}{2\lambda_0 - 1}, n_2\right)}$, $\alpha_i = 0$ 时的预测区间以及 $\lambda_0 \geq 1$

时的置信区间上限, 有助于评价区域试验一组品种的相对稳定性。

图 40-2 表示 α 与 λ 的分布图, 它用以解释这种方法对于品种稳定性分析的应用。图中两条正交轴分别以 α, λ 值表示, 如果 $\alpha_i = 0$ 为真, 那么将有 95% 的 $\hat{\alpha}$ 落在双曲线内。第一、二条垂直线是 $\lambda_0=1$ 时 90% 置信区间的上、下限, 其余垂直线是 $\lambda_0 > 1$ 时的置信上限。这些直线把 α, λ 区分成不同的稳定性区域, 对应于期望值为 $(\alpha_i, \lambda_i) = (0, 1)$ 和 $(<0, 1)$ 的区域分别以图中的 A, B 表示。某一品种的 α_i, λ_i 落在 A 区域表示具有平均稳定性, 落在 B 区域的品种其稳定性显著地高于平均稳定性。

40.1.3 Shukla 模型

Shukla(1972)提出的模型为

$$y_{ijk} = \mu + d_i + \varepsilon_j + g_{ij} + e_{ijk}$$

其中 y_{ijk} 表示第 i 个品种在第 j 个环境、第 k 次重复的平均产量; μ 表示所有 y_{ijk} 的总平均产量; d_i 表示第 i 个品种的遗传效应($i=1, 2, \dots, v$); ε_j 表示第 j 个环境的环境效应($j=1, 2, \dots, l$); g_{ij} 表示第 i 个品种与第 j 个环境的交互作用效应; e_{ijk} 表示第 i 个品种在第 j 个环境、第 k 次重复的试验误差($k=1, 2, \dots, r$); 记 \bar{y}_{ij} 为 r 次重复 y_{ijk} 的均值, \bar{e}_{ij} 是 r

次重复 e_{ijk} 的均值, 则(40.1)式又可表示为 $\bar{y}_{ij} = \mu + d_i + \varepsilon_j + g_{ij} + \bar{e}_{ij}$ 。

若假定 \bar{e}_{ij} 的期望值 $E(\bar{e}_{ij}) = 0$, \bar{e}_{ij} 的期望方差 $V(\bar{e}_{ij}) = \sigma_0^2$, 这里 σ_0^2 是 r 次重复平均的环境内误差方差, 且可通过下式进行估算:

$$\sigma_0^2 = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij})^2 / vl(r-1)$$

令 $g_{ij} + \bar{e}_{ij} = r_{ij}$, 并假设 $E(r_{ij}) = 0$, $V(r_{ij}) = \sigma_i^2$, $V(g_{ij}) = \sigma_i'^2$, 于是有 $\sigma_i^2 = \sigma_i'^2 + \sigma_0^2$, 其中 $\sigma_i'^2$ 的估计式为

$$\hat{\sigma}_i'^2 = \frac{1}{(l-1)(v-1)(v-2)} \left[v(v-1) \sum_j (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - \sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \right]$$

并有 $\hat{\sigma}_i'^2 / \hat{\sigma}_0^2$ 渐近于 F 分布, 其自由度为 $[l-1, lv(r-1)]$, 所以有 $F = \hat{\sigma}_i'^2 / \hat{\sigma}_0^2$ 。

当实际计算中出现 $\hat{\sigma}_i'^2$ 有负值情况时, 一般取零值处理, 这样, 当 F 值越小, 则表明第 i 个品种的基因越稳定。一般情况下, F 值不显著, 则认为该品种稳定; F 值达到显著水平, 则该品种不稳定。

例如, 对 5 个品种 5 个地点的品种区域试验结果进行稳定性分析, 得到结果如图 40-1。

分析时按图 40-1 中的阴影部分编辑并用鼠标选中待分析的数据, 然后执行专业统计下面“品种比较试验”里面的“品种稳定性分析”, 这时系统会要求输入地点数、品种数(这里分别是 5、5)。输入后系统立即给出 George C C Tai 模型的品种稳定性分析图(图 40-2)。它可以解释该方法对于品种稳定性分析的应用, 从图中各个品种分布位置可以看出, 供试品种 1 和品种 4 具有平均稳定性; 而品种 2 的稳定性最差。

	A	B	C	D	E
1	地点	品种	区组 I	区组 II	
2	1	1	750	772	
3	1	2	880	858	
4	1	3	770	544	
5	1	4	685	707	
6	1	5	673	533	
7	2	1	950	974	
8	2	2	862	892	
9	2	3	896	854	
10	2	4	950	914	
11	2	5	933	907	
12	3	1	924	900	
13	3	2	1023	1017	
14	3	3	934	914	
15	3	4	997	969	
16	3	5	760	816	
17	4	1	930	980	
18	4	2	890	938	
19	4	3	890	920	
20	4	4	975	949	
21	4	5	892	870	
22	5	1	759	799	
23	5	2	739	699	
24	5	3	625	683	
25	5	4	800	858	
26	5	5	660	682	
27					

图 40-1 品种稳定性分析数据编辑格式

结束图形分析, 返回到编辑状态, 则可得到几种方法的稳定性分

析结果。分析结果的第一部分是各个试验点误差均方的同质性检验。一般是要求卡方值较小, 概率 p 值大于 0.05。本例中的 $p=0.6845$ 表明, 各个试验点的误差均方是同质的, 可以进行进一步分析。

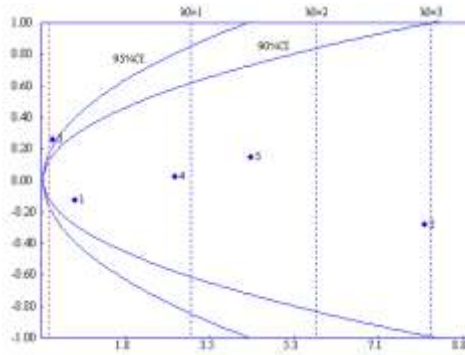


图 40-2 Tai 模型品种稳定性分析图

随后，系统给出了各个试点、品种的均值，并给出了各个试点参试品种的方差分析结果和几个试点的联合方差分析结果，从这里可以初步分析各个试点各个品种的表现情况，以及它们的综合表现情况。

下面是 5 个品种稳定性分析方法分析结果：

误差均方同质性测验						
地点	误差均方	自由度 V	V*ln(S ²)			
1	6057.6000	4	34.8363			
2	589	4	25.5137			
3	604.4000	4	25.6169			
4	698	4	26.1929			
5	825.4000	4	26.8635			
df=4	样本卡方值=9.4362		p=0.0511			
平均产量=841.9200			试验误差=877.4400			
品种\环境	1	2	3	4	5	平均
1	761	962	912	955	779	873.8000
2	869	877	1020	914	719	879.8000
3	657	875	924	905	654	803
4	696	932	983	962	829	880.4000
5	603	920	788	881	671	772.6000
平均	717.2000	913.2000	925.4000	923.4000	730.4000	841.9200
第 1~5 个地点(环境)各品种差异方差分析表略						
资料的联合方差分析						
变异来源	df	SS	MS	F 值	p 值	
环境区组内	5	14676.4000	2935.2800	1.6726	0.1872	
环境	4	466805.2800	116701.3200	17.2931	0.0000	
品种	4	102519.6800	25629.9200	3.7979	0.0235	
品种×环境	16	107974.7200	6748.4200	3.8455	0.0027	
试验误差	20	35097.6000	1754.8800			

总的	49	727073.6800							
不考虑区组效应的联合方差分析									
变异来源	df	SS	MS	F 值	p				
环境	4	466805.2800	116701.3200	17.2931	0.0000				
品种	4	102519.6800	25629.9200	3.7979	0.0235				
品种×环境	16	107974.7200	6748.4200	3.3895	0.0031				
试验误差	25	49774	1990.9600						
总的	49	727073.6800							
估计品种稳定性参数的方差分析(一)									
变异来源	df	SS	MS	F 值	p				
总的	24	338649.8400							
品种	4	51259.8400	12814.9600	4.2047	0.0176				
环境+(品种×环境)	20	287390	14369.5000						
环境(线性)	1	233402.6400	233402.6400						
品种×环境(线性)	4	8270.9058	2067.7264	0.6784	0.6174				
合并离差	15	45716.4542	3047.7636						
估计品种稳定性参数的方差分析(二)									
不考虑环境区组的									
变异来源	df	SS	MS	F 值	p	Sdi ²	F 值	p	
1	3	2000.3960	666.7987	0.7599	0.5297	-210.6413	0.6698	0.5785	
2	3	22664.6912	7554.8971	8.6102	0.0007	6677.4571	7.5892	0.0009	
3	3	759.8172	253.2724	0.2886	0.8330	-624.1676	0.2544	0.8574	
4	3	7910.0055	2636.6685	3.0050	0.0546	1759.2285	2.6486	0.0709	
5	3	12381.5444	4127.1815	4.7037	0.0121	3249.7415	4.1459	0.0163	
试验误差	20	17548.8000	877.4400						
不考虑环境区组的									
试验误差	25	24887	995.4800						
检验各个回归系数与 1 的差异显著性									
No.	品种均值	回归系数	(b-1)/seb	p					
1	873.8000	0.8726	0.4987	0.6252					
2	879.8000	0.7192	1.0991	0.2890					
3	803	1.2504	0.9800	0.3426					
4	880.4000	1.0202	0.0790	0.9381					
5	772.6000	1.1377	0.5388	0.5979					
Seb=0.255519									
品种稳定性 Tai 模式分析									
No.	品种均值	α	$p(\alpha = 0)$	λ	$p(\lambda = 1)$				
1	873.8000	-0.1307	0.3646	0.7055	0.5599				

2	879.8000	-0.2881	0.5354	8.0382	0.0010
3	803	0.2569	0.0425	0.2437	0.8648
4	880.4000	0.0207	0.9377	2.8170	0.0653
5	772.6000	0.1412	0.6749	4.4016	0.0156
Shukla 模型分析					
No.	品种均值	方差	F 值	p 值	
1	873.8000	24.5500	0.0247	0.9987	
2	879.8000	9853.0500	9.8978	0.0001	
3	803	411.4833	0.4134	0.7973	
4	880.4000	2179.0167	2.1889	0.0994	
5	772.6000	4402.9500	4.4229	0.0077	

Eberhart 和 Russell 模型分析, 先看“环境×品种平均产量方差分析”中方差分析表中的交互项是否显著, 若显著, 说明存在品种×环境交互, 这时才有必要进一步分析以后的内容; 若不显著, 说明交互不重要, 则无需再作分析。该例中, $F=3.8455$, $p=0.0027$, $p<0.05$, 说明交互项显著, 可继续分析。

由于环境×品种平均产量方差分析中交互显著, 可进一步对“估计品种稳定性参数的方差分析(一)”进行分析。如表中品种×环境(线性)项显著, 说明品种间回归系数的差异显著, 反之, 则说明品种间回归系数差异不显著。这里的 F 统计量为 0.6784, $p=0.6174$; $p>0.05$ 说明品种间回归系数差异不显著。一般来说, 只有在品种间回归系数的差异显著, 并且回归模型适合的情况下, 下面的回归系数才有可比性。

下面的“估计品种稳定性参数的方差分析(二)”表中的回归离差是各品种的回归剩余均方与误差均方之差, 若 P 值小于 0.05, 则回归关系显著; 再下面是检验各回归系数与 1 的差异显著性, t 测验概率值 p 若小于 0.05, 则与 1 差异显著。

回归模型中回归系数 b 值越大, 说明该品种对环境反应越敏感, 在不同环境条件下, 产量差异大, 因而稳定性低。如果 b 值小, 说明该品种对环境反应迟钝, 在不同环境条件下, 产量差异小, 稳定性好, 即如果回归系数 $b=1$, 品种具有平均稳定性; $b<1$ 表示品种超平均稳定; 而 b 值大于 1 表示品种不稳定。

若回归模型不适合时, 由于品种稳定性同时与回归系数和回归离差有关, 同时权衡两个参数, 不便于品种间稳定性的直接比较。此时, 可结合利用 Shukla 模型进行分析。

40.2 一年多区域试验的统计分析

40.2.1 方法简介

设有 v 个品种, 在 u 个地点做比较试验, 每个地点皆设 r 个完全区组(重复), 按随机区组设计进行试验, 则第 i 个品种($i=1, 2, \dots, v$) 在第 j 个地点($j=1, 2, \dots, u$), 第 k 区组($k=1, 2, \dots, r$) 的观测值为 x_{ijk} , 它的线性模型为

$$x_{ijk} = \mu + \tau_i + \nu_j + (\tau\nu)_{ij} + \rho_{jk} + \varepsilon_{ijk}$$

式中 μ 为群体的平均值, τ_i 为品种 i 的效应值, v_j 为地点 j 的效应, $(\tau v)_{ij}$ 为品种 \times 地点互作效应, ρ_{jk} 为地点内的区组效应, ε_{ijk} 为随机误差。由此可以得到一年多点区域试验的方差分析表(表 40-2)。

表 40-2 一年多点试验的方差分析

变异来源	自由度 df	平均和 SS	均方 MS	ems(固定模型)
地点内区组	$u(r-1)$	SS_r	MS_r	
地 点	$u-1$	SS_u	MS_u	$\sigma_e^2 + rv \chi_u^2$
品 种	$v-1$	SS_v	MS_v	$\sigma_e^2 + ru \chi_v^2$
品种 \times 地点	$(u-1)(v-1)$	SS_{uv}	MS_{uv}	$\sigma_e^2 + r \chi_{uv}^2$
试验误差	$u(r-1)(v-1)$	SS_e	MS_e	σ_e^2
总 计	$ruv-1$	SS_t		

若方差分析中各效应的差异显著, 就可进一步进行多重比较, 以检查不同品种间或不同地区间的各效应是否存在差异。多重比较一般采用 Duncan 新复极差方法, 为此要计算出各效应差数的标准误。

方差分析可以反映品种对不同地点的适应性, 但不能直观反映品种的稳定性。我们可以定义另外两个参数, 即某一品种与地点的互作方差及其相对变异系数。对于第 i 个品种(或品系), 可以算出在 u 个地区的品种 \times 地点互作效应值, 若计算出这些离差平方和, 就能评价某品种在 u 个地点的稳定性。

各品种与地区的互作方差的大小反映该品种在这些地区的稳定性。互作方差小, 说明品种具有广泛适应性, 属稳定类型, 但仅比较不同品种的互作方差还不足以对各品种稳定性的好坏作出评价, 因为若两品种的互作方差相同, 而产量水平不同, 那么产量较大的品种的相对变异较小, 而产量较小的品种的相对变异较大, 因此, 要引入相对变异系数来评价品种的稳定性。

40.2.2 DPS 平台的操作示例

分析前先建立数据文件, 按地点、品种、区组顺序(图 40-1)输入数据。将所输入的整个数据阵定义成数据块, 在菜单下, 选择执行“专业统计 \rightarrow 品种比较试验一年多点品比试验”功能项, 再按系统提示输入试验点数(u)、试验品种数(v)。输入这些参数后系统很快完成运算, 并在当前光标下输出计算结果。输出结果包括双向列表、方差分析表、F 统计量、多重比较和稳定性分析的结果以及最后的综合评价。

DPS 系统最多可容纳 100 个处理、5000 个样本的数据。可以允许含有缺失数据, 有缺失数据时, 就让该单元格空着。这里仍以前面品种稳定性分析数据为例, 在 5 个地点进行 5 个品种的区域品种比较试验, 每个地点 2 个区组, 获得一组试验资料。现对该试验资料进行一年多点试验的方差分析, 分析前先按图 40-1 方式输入、编辑数据。

定义数据块后，进入菜单，选择“专业统计→品种比较试验”下面的“一年多点点比试验”，系统给出用户输入地点、品种数及选择统计分析方法的界面如图 40-3。

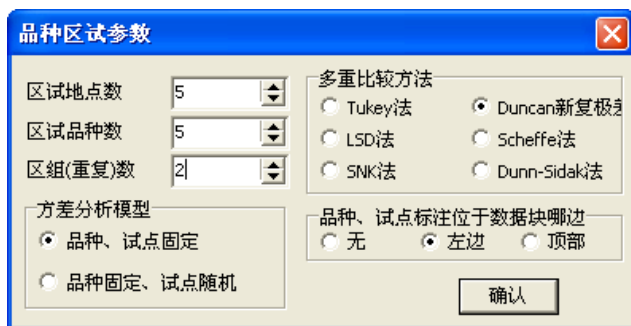


图 40-3 一年多点点品种区试数据分析用户界面

这里我们输入各个参数及统计分析方法选项如上图，点击“确认”按钮、执行计算后得到主要结果如下：

误差均方同质性测验						
df=4	卡方 0.05=9.488		样本卡方值=9.436			
表 1 品种试验方差分析表						
(品种、地点固定，随机区组模型)						
变异来源	df	SS	MS	F	Prob.	
地点内区组	5	14676.40	2935.28	1.6726	0.1872	
地点	4	466805.28	116701.32	66.5010	0.0000	
品种	4	102519.68	25629.92	14.6049	0.0000	
品种×地点	16	107974.72	6748.42	3.8455	0.0027	
试验误差	20	35097.60	1754.88			
总的	49	727073.68				
品种×地点平均产量两向列表						
	1	2	3	4	5	E.
1	761	962	912	955	779	873.8
2	869	877	1020	914	719	879.8
3	657	875	924	905	654	803
4	696	932	983	962	829	880.4000
5	603	920	788	881	671	772.6000
V.	717.2000	913.2000	925.4000	923.4000	730.4000	841.9200
品种×地点互作效应两向列表						
	1	2	3	4	5	E.
1	11.9200	16.9200	-45.2800	-0.2800	16.7200	31.8800
2	113.9200	-74.0800	56.7200	-47.2800	-49.2800	37.8800
3	-21.2800	0.7200	37.5200	20.5200	-37.4800	-38.9200
4	-59.6800	-19.6800	19.1200	0.1200	60.1200	38.4800
5	-44.8800	76.1200	-68.0800	26.9200	9.9200	-69.3200
V.	-124.7200	71.2800	83.4800	81.4800	-111.5200	
各个品种差异显著性 Duncan 多重比较						

处理	均值	5%显著水平	1%极显著水平				
4	880.4	a	A				
2	879.8	a	A				
1	873.8	a	A				
3	803	b	B				
5	772.6	B	B				
各个地点间差异显著性 Duncan 多重比较							
处理	均值	5%显著水平	1%极显著水平				
3	925.4	A	A				
4	923.4	A	A				
2	913.2	A	A				
5	730.4	B	B				
1	717.2	B	B				
品种丰产性及其稳定性的回归分析							
Y1= 139.1570 +0.87258x							
Y2= 274.3329 +0.71915x							
Y3=-249.7452 +1.25041x							
Y4= 21.4933 +1.02018x							
Y5=-185.2380 +1.13768x							
品种丰产性及其稳定性分析							
品种	丰产性参数		稳定性参数			适应地区	综合评价 (供参考)
	产量	效应	方差	变异度	回归系数		
4	880.4000	38.4800	1982.2520	5.0571	1.0202	E1~E5,	很好
2	879.8000	37.8800	6586.6720	9.2246	0.7192	E1~E5,	很好
1	873.8000	31.8800	689.5720	3.0052	0.8726	E1~E5,	很好
3	803	-38.9200	921.7320	3.7808	1.2504	E1~E5,	一般
5	772.6000	-69.3200	3316.6120	7.4541	1.1377	E1~E5,	较差

40.3 多年多点品种区域试验的统计分析

40.3.1 方法简介

品种区域试验常常需要在多个地点连续进行数年，以便为品种的评定和应用提供更广泛的信息，尤其是品种对不同年份随机变化的气候条件适应性的信息。

设 v 个品种在 u 个地点连续进行 w 年试验，每年每地设 r 个区组，按完全随机区组设计进行试验。第 i 个品种在第 j 地点、第 k 年份、第 l 区组的观测值 x_{ijkl} 的线性模型为

$$x_{ijkl} = \mu + \tau_i + \nu_j + (\tau\nu)_{ij} + \omega_k + (\tau\omega)_{ik} + (\nu\omega)_{jk} + (\tau\nu\omega)_{ijk} + \rho_{jkl} + \varepsilon_{ijkl}$$

($i=1, 2, \dots, v; j=1, 2, \dots, u; k=1, 2, \dots, w; l=1, 2, \dots, r$)

式中， μ 为群体的平均值， τ_i 为品种 i 的效应值， ν_j 为地点 j 的效应， $(\tau\nu)_{ij}$ 为品种×地点交互效应， $(\tau\omega)_{ik}$ 为品种×年份的交互效应(随机效应)， $(\nu\omega)_{jk}$ 为地点×年份的交互效应(随机效应)， $(\tau\nu\omega)_{ijk}$ 为品种×地点×年份的交互效应(随机效应)， ρ_{jkl} 为品种×地点×年份的交互效应(随机效应)， ε_{ijkl} 为随机误差。根据该线性模型，可以对多年多地品种区域试验资料的变异来源给予分解，从而列出方差分析表(表 40-3)。

表 40-3 多年多地区域试验方差分析表

变异来源	自由度 df	SS	MS	EMS(vu 固定, w 随机)
点内年内区组	$Uw(r-1)$	SS_r	MS_r	
年 份	$w-1$	SS_w	MS_w	$\sigma^2\varepsilon+ruv\sigma^2w$
地 点	$u-1$	SS_u	MS_u	$\sigma^2\varepsilon+rv\sigma^2uw+rvw\chi_u^2$
地点×年份	$(u-1)(w-1)$	SS_{uw}	MS_{uw}	$\sigma^2\varepsilon+ru\sigma^2vw$
品 种	$v-1$	SS_v	MS_v	$\sigma^2\varepsilon+ru\sigma^2vw+ruv\chi_v^2$
品种×年份	$(v-1)(w-1)$	SS_{vw}	MS_{vw}	
品种×地点	$(v-1)(u-1)$	SS_{vu}	MS_{vu}	
品种×年份×地点	$(v-1)(u-1)(w-1)$	SS_{vuw}	MS_{vuw}	$\sigma^2\varepsilon+ru\sigma^2vuw+rvw\chi_{uv}^2$
试验误差	$uw(r-1)(v-1)$	SS_ε	MS_ε	
总 和	$vuwr-1$	SS_T		

若方差分析显示各效应的差异显著, 可进一步进行多重比较, 以检查不同品种间或不同地点间的效应是否存在差异。多重比较一般采用 Duncan 新复极差方法, 为此要计算出各效应差数的标准误。方差分析可以反映品种在不同地区的适应性, 但不能直观反映品种的稳定性, 因此, 可另外定义两个参数, 即一个品种与地区的互作方差及其相对变异系数。

对于第 i 个品种(或品系), 可以算出在 u 个地区的品种×地点互作效应值, 若计算出这些离差平方和, 就能评价该品种在 u 个地区的稳定性。品种×地点互作方差的大小反映了该品种在各地区的稳定性。互作方差小说明品种具有广泛适应性, 属稳定类型; 反之亦然。但是, 仅比较不同品种的互作方差还不足以对各品种稳定性的好坏作出准确评价, 还要考虑品种的相对变异性。两品种的互作方差可能完全相同, 但产量水平却不同。显然, 产量较大品种的相对变异性较小, 稳定性较强; 而产量较小品种的相对变异性可能较大, 稳定性可能较差。

40.3.2 DPS 平台的资料整理与分析步骤

对于多年多点的品比试验资料, DPS 系统要求按地点、品种、区组顺序输入试验数据并按规定格式编辑和定义数据块。在菜单下选择执行多年多点品比试验的方差分析功能项, 并按系统提示输入试验点数(u)、试验品种数(v), 按回车执行运算, 系统将在当前光标下输出计算结果。输出结果包括双向列表、方差分析表、 F 统计量、多重比较和稳定性分析结果以及最后的综合评价。

本系统最多可容纳 100 个处理、5000 个样本的数据。可以允许含有缺失数据, 有缺失数据时, 就让该单元格空着。

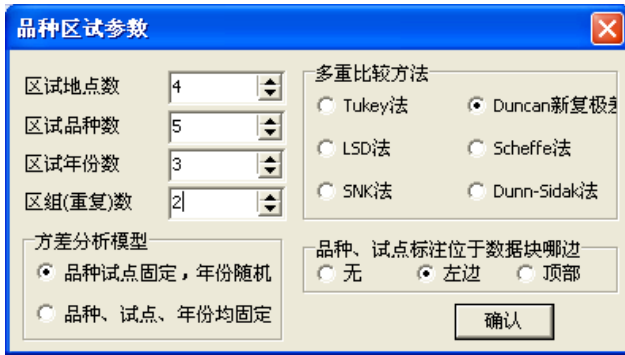
如有 5 个玉米品种在 4 个地点连续进行了 3 年试验, 每次试验设 2 个区组。试验结果数据按图 40-4 格式输入并编辑定义成数据块, 注意品种、地点、年份、重复的排列

格式。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	品种	1		2		3		4		5		6		
2		地点	I	II	I	II	I	II	I	II	I	II	I	II
3	1		6	7	5	6	6	4	9	8	15	13	10	11
4	2	7	8	7	6	8	6	9	9	14	11	10	11	
5	3	10	12	13	11	9	10	9	10	13	15	9	10	
6	4	5	6	6	6	7	5	8	9	10	11	10	10	
7														

图 40-4 多年多点品种区试数据编辑格式

然后，进入菜单操作，选择“专业统计→品种比较试验”中的多年多点品比试验的方差分析项，



按系统提示输入各参数后按回车执行运算，分析结束时输出如下结果：

①方差分析表					
方差分析表 1					
变异来源	SS	df	MS	F	prob.
点内年内区组间	12.3000	12	1.0250		
年份间	17.0667	2	8.5333	5.52021	0.00695
地点间	22.8917	3	7.6306	5.77101	0.03348
品种间	1495.6167	4	373.9042	49.49644	0.00001
地点×年份	7.9333	6	1.3222	0.85535	0.53429
品种×年份	60.4333	8	7.5542	4.88679	0.00019
地点×品种	241.9833	12	20.1653	12.54882	0.00000
地点×品种×年份	38.5667	24	1.6069	1.03953	0.44096
误差	74.2000	48	1.5458		
总计	1970.9917	119			

注：由于方差分析表①中的地点×年份和品种×地点×年份两项不显著，因此将其归入误差平方和项，相应的自由度也合并，从而得到方差分析表 2 (该分析过程由系统自动完成)。

方差分析表 2					
变异来源	SS	df	MS	F	prob.
年份间	17.0667	2	8.5333	5.52021	0.00695
地点间	22.8917	3	7.6306	5.77101	0.03348
品种间	1495.6167	4	373.9042	49.49644	0.00001
地点×年份	7.9333	6	1.3222	0.85535	0.53429
品种×年份	60.4333	8	7.5542	4.88679	0.00019
地点×品种	241.9833	12	20.1653	12.54882	0.00000
地点×品种×年份	38.5667	24	1.6069	1.03953	0.44096
误差	74.2000	48	1.5458		
总计	1970.9917	119			

点内年内区组间	12.3000	12	1.0250		
年份间	17.0667	2	8.5333	5.51450	0.00575
地点间	22.8917	3	7.6306	4.93109	0.00345
品种间	1495.6167	4	373.9042	49.49644	0.00001
品种 × 年份	60.4333	8	7.5542	4.88173	0.00007
地点 × 品种	241.9833	12	20.1653	13.03141	0.00000
误差	120.7001	78	1.5474		
总计	1970.9917	119			

②每个变量的方差估计值:

年份=0.17316 品种×年份=0.75084

地点×品种=3.09306 品种=15.26458 地点=0.21028

③各个品种、地点间差异显著性 DunCan's 多重比较:

		品 种 间		
处理	平均	5%显著	1%极显著	
4	15.0417	<i>a</i>	<i>a</i>	
2	10.5833	<i>b</i>	<i>b</i>	
1	7.3333	<i>c</i>	<i>bc</i>	
5	7.1250	<i>c</i>	<i>bc</i>	
3	4.8750	<i>c</i>	<i>c</i>	
		年份间		
处理	平均	5%显著	1%极显著	
2	0.2381	<i>a</i>	<i>a</i>	
1	0.2181	<i>a</i>	<i>a</i>	
3	0.2181	<i>a</i>	<i>a</i>	

④品种丰产性及其稳定性分析:

品 种	丰产性参数		稳定性参数		适应地区及评价 (供参考)
	产量	效应	方差	变异度	
v4	15.042	6.050	4.160	13.560	e1-e4, 很好
v2	10.583	1.592	0.077	2.615	e1-e4, 较好
v1	7.333	-1.658	4.565	29.137	e3, 较差
v5	7.125	-1.867	0.462	9.542	e1-e4, 一般
v3	4.875	-4.117	4.179	41.933	e1, 不好

40.4 品种区域试验 AMMI 模型分析

40.4.1 方法简介

作物品种区域化试验旨在鉴定品种的丰产性、稳定性和适应性。参加区试的品种在不同地点的产量往往表现得 不一致。这表明品种的基因型和环境互作($g \times e$)效应的存在。以往对这种互作效应大多采用线性模型进行分析,这在前面已经作了介绍。但线性模型一般仅能解释很少一部分交互作用的变化。近年来一种更为有效的加性主效应乘积交互作用(additive main effects and multiplicative interaction, AMMI)模型已开始被用于多年多点的区域试验资料的分析(Crossa et al., 1990; Zobel et al., 1988)。该模型与方差分析模型、

线性回归模型相比,应用范围更广且更有效。AMMI 模型的分析结果可以用直观简洁的图形表达和解释。AMMI 分析不仅可以帮助我们进行更可靠的稳定性分析,而且可以鉴别一些具有特殊品种基因型和环境($g \times e$)互作效应的基因型,为针对某一特殊环境的特殊适应性的品种的育种提供有价值的信息。该模型的主要特点是将方差分析和主成分分析有机地结合在一起,具有如下形式:

$$y_{ge} = \mu + \alpha_g + \beta_e + \sum_{i=1}^N \lambda_n \gamma_{gn} \delta_{gn} + \theta_{ge}$$

式中, y_{ge} 是在环境 e 中基因型 g 的产量, μ 代表总体平均值, α_g 是基因型平均偏差(各个基因型平均值减去总的平均值), β_e 是环境的平均偏差(各个环境的平均值减去总的平均值), λ_n 是第 n 个主成分分析的特征值, δ_{gn} 是第 n 个主成分的基因型主成分得分, γ_{gn} 是第 n 个主成分的环境主成分得分, n 是在模型主成分分析中主成分因子轴的总个数, θ_{ge} 为残差。如果试验设有重复,则误差项为 ε_{ger} , 等于 y_{ge} 平均值与 r 个重复的单个观察值之间的偏差,并具有可加性。

在 DPS 系统中,为提供充分有用的信息给用户,品种比较试验的统计分析从方差分析模型、回归分析模型入手,步步深入。在此依次介绍有关模型。

(1) 方差分析(可加性)模型。方差分析提供一个可加性两向数据处理模型,它将产量观察值 y_{ger} 分解为三个部分:① 总体平均值 μ , 基因型偏差 α_g , 环境偏差 β_e ; ② 非加性残差部分 θ_{ge} ; ③ 误差项 ε_{ger} , 即

$$y_{ger} = \mu + \alpha_g + \beta_e + \theta_{ge} + \varepsilon_{ger}$$

式中, μ 由样本均值 \bar{y} 估计, 偏离总体平均值的基因型离差 α_g 由 $\bar{y}_g - \bar{y}$ 估计, 而环境离差 β_e 由 $\bar{y}_e - \bar{y}$ 估计, 交互作用部分 θ_{ge} 由 $\bar{y}_{ge} - \bar{y}_g - \bar{y}_e + \bar{y}$ 估计。由此列出方差分析表 40-4。

(2) 线性回归模型。过去,分离和分析交互作用的最初的统计方法是简单的线性回归方法,作物育种学家注意到产量中很大一部分变异可以用单个基因型产量对环境平均的线性回归来解释。回归模型将交互作用效应分离成回归和残差两个部分,回归效应等于基因型斜率 ξ_g 乘以环境离差 β_e , 所以有了 Finlay-Wilkinson (1963) 回归模型

$$y_{ger} = \mu + \alpha_g + \beta_e + \xi_g \beta_e + \rho_{ge} + \varepsilon_{ger}$$

对该模型一般先用最小二乘方法拟合出加性效应参数 α_g 和 β_e , 然后估计基因型斜率 ξ_g 。

当考察焦点集中于环境而非基因型时,环境回归模型将交互作用效应分离成环境斜率 ϕ_e 乘以基因型离差 α_g , 所以有相应的回归模型

$$y_{ger} = \mu + \alpha_g + \beta_e + \phi_e \alpha_g + \rho_{ge} + \varepsilon_{ger}$$

表 40-4 加性模型方差分析

变异来源	自由度 df	平方和 SS
------	----------	--------

总的	$ger-1$	$\sum_{i=1}^g \sum_{j=1}^e \sum_{k=1}^r (y_{ijk} - \bar{y})^2$
处理	$ge-1$	$r \sum_{i=1}^g \sum_{j=1}^e (\bar{y}_{ij} - \bar{y})^2$
基因	$g-1$	$er \sum_{i=1}^g (\bar{y}_i - \bar{y})^2$
环境	$e-1$	$gr \sum_{j=1}^e (\bar{y}_j - \bar{y})^2$
交互作用	$(g-1)(e-1)$	$r \sum_{i=1}^g \sum_{j=1}^e (\bar{y}_{ge} - \bar{y}_g - \bar{y}_e + \bar{y})^2$
误差	$ge(r-1)$	$\sum_{i=1}^g \sum_{j=1}^e \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij})^2$

除基因型和环境回归模型以外, 还有第三种回归模型。这一模型包含了基因型和环境两者的离差。这种联合回归将交互作用效应解释为联合回归系数 κ 乘以基因型离差 α_g 和环境型离差 β_e , 即

$$y_{ger} = \mu + \alpha_g + \beta_e + \kappa \alpha_g \beta_e + \rho_{ge} + \varepsilon_{ger}$$

虽然通常的基因型回归模型和与之对应的环境型回归模型都可用于解释交互作用, 但是为了更有效地分析交互作用效应, 有必要结合联合回归模型。因此, 线性回归最完整的回归模型是

$$y_{ger} = \mu + \alpha_g + \beta_e + \xi_g \beta_e + \phi_e \alpha_g + \kappa \alpha_g \beta_e + \rho_{ge} + \varepsilon_{ger}$$

在品种区域试验的资料分析中, 尽管使用了上述比较复杂的线性回归模型, 但结果往往仍不尽人意。如下面给出的实例中, 基因型、环境和联合回归模型三者加起来才解释了交互作用效应平方和的 12.9%, 剩余残差仍很大。

为了充分解释交互效应变异, 近年来有人结合方差分析和主成分分析(PCA)技术, 提出了如本节开始时所描述的加性主效应乘积交互作用(AMMI)模型, 并在一般情形下能很好地解释基因×环境的互作效应。再结合双标图(biplot)(bardu, 1978), 就可以增进人们对有关效应的理解。

40.4.2 DPS 平台的数据格式及其编辑

分析前先在编辑状态下输入数据, 数据存放格式要求以行代表环境(地点), 以列代表基因型(品种), 按试验地点的顺序, 输完一个地点的数据后再输下一个。若一个品比试验有 e 个地点、 g 个品种、 r 个区组(重复), 其资料整理格式和一年多品种区域实验分析相同。将输入数据定义成数据块, 在菜单方式下执行 AMMI 功能项, 再按提示输入试验点数(E)、试验品种数(G)。运算结束时, 系统在当前光标下输出计算结果。

40.4.3 应用示例

这里仍以图 40-1 所示数据进行 AMMI 模型分析。在菜单下选择“专业统计→品种比较试验”中的 AMMI 模型分析(有区组)项,执行后输入环境和品种个数之后,获得如下输出结果有:① 平均产量两向列表(注:列表示品种,行表示地点);② 离差两向列表(注:列表示品种,行表示地点);③ 各种方差分析表,如品种试验方差分析表、品比试验结果线性回归分析模型方差分析表、品比试验结果 AMMI 分析;④ 品种和环境各个变量的平均值、离差、主成分值及特征向量;⑤ AMMI 模型拟合两向列表(注:列表示品种,行表示地点环境)。本例输出结果如下:

平均产量两向列表(略)					
离差两向列表(略)					
表 2 品种试验方差分析表					
变异来源	DF	SS	MS	df	Prob.
总的	49	727073.6800	14838.2384		
处理	24	677299.6800	28220.8200	14.1745	0
基因	4	102519.6800	25629.9200	12.8731	0
环境	4	466805.2800	116701.3200	58.6156	0
交互作用	16	107974.7200	6748.4200	3.3895	0.0031
误差	25	49774	1990.9600		
线性回归模型分析结果					
变量	均值	离差	斜率	r ² (%)	
1	873.8000	31.8800	0.8726	27.4800	
2	879.8000	37.8800	0.7191	13.9800	
3	803	-38.9200	1.2504	79.3900	
4	880.4000	38.4800	1.0202	0.2400	
5	772.6000	-69.3200	1.1377	6.6700	
1	717.2000	-124.7200	1.6182	20.4600	
2	913.2000	71.2800	0.1876	56.5900	
3	925.4000	83.4800	1.4584	18.4500	
4	923.4000	81.4800	0.5650	57.3800	
5	730.4000	-111.5200	1.1708	3.8200	
联合回归系数=-0.00323981					
表 品比试验结果线性回归分析					
变异来源	df	SS	MS	F	Prob.
总的	49	727073.68	14838.2384		
处理	24	677299.68	28220.82	14.1745	0
基因	4	102519.68	25629.92	12.8731	0
环境	4	466805.28	116701.32	58.6156	0
交互作用	16	107974.72	6748.4200	3.3895	0.0031
联合回归	1	10046.4671	10046.4671	5.0460	0.0338
基因回归	3	6495.3445	2165.1148	1.0875	0.3726
环境回归	3	20110.4153	6703.4718	3.3670	0.0344
残差	9	71322.4931	7924.7215	3.9804	0.0030
误差	25	49774	1990.9600		

表 品比试验结果 AMMI 分析					
变异来源	df	SS	MS	F	Prob.
总的	49	727073.6800	14838.2384		
处理	24	677299.6800	28220.8200	14.1745	0
基因	4	102519.6800	25629.9200	12.8731	0
环境	4	466805.2800	116701.3200	58.6156	0
交互作用	16	107974.7200	6748.4200	3.3895	0.0031
PCA1	7	78761.3691	11251.6241	5.6514	0.0005
PCA2	5	19308.9133	3861.7827	1.9397	0.1233
PCA3	3	9786.6811	3262.2270	1.6385	0.2057
残差	1	117.7566	117.7566		
误差	25	49774	1990.9600		
变量	产量平均	离差	PCA1	PCA2	PCA3
1	873.8000	31.8800	-1.8200	-3.4874	3.5421
2	879.8000	37.8800	11.4364	-1.7721	1.0544
3	803	-38.9200	0.6089	2.0037	-6.7772
4	880.4000	38.4800	-2.6946	7.6959	3.0886
5	772.6000	-69.3200	-7.5307	-4.4400	-0.9079
1	717.2000	-124.7200	8.9040	-5.5579	2.3298
2	913.2000	71.2800	-7.0436	-4.2309	-2.1865
3	925.4000	83.4800	6.1231	5.9232	-3.3451
4	923.4000	81.4800	-3.6824	0.0740	-3.0590
5	730.4000	-111.5200	-4.3011	3.7916	6.2607
Results for Mega-environments					
Gentype	Wins(%)	Exp. Yield	Environments		
1	13.9610	892.0554	2		
2	52.3528	951.5429	1	3	
4	33.6863	916.7343	4	5	

系统然后给出 AMMI 模型分析图形界面，其中包括双标图分析(图 40-5)、品种适应性分析(图 40-6)和各品种 AMMI 排序分析(图 40-7、40-8)。

AMMI 双标图的纵、横轴都是平均产量、离差、IPCA1、IPCA2 和 IPCA3，用户可以指定任意两个变量，作出双标图。双标图有助于揭示品种和环境之间的相互关系。双标图可能是解释 AMMI 模型最有效的工具。AMMI 双标图是以品种和地点的主效应值或平均产量为 x 坐标， y 坐标为品种和地点的 IPCA1 值所作的图形。这一双标图同时表达了品种的丰产性和 $g \times e$ 交互作用的信息。解释 AMMI1 双标图的原则：

(1) 在水平方向，品种和地点的分布范围表明相应的主效应；

(2) 在垂直方向，如果以 IPCA1=0 作一条水平线，则品种与同在此水平线一侧的地点有正的互作，与另一侧的地点的互作则为负；

(3) 靠近水平线的品种为较为稳定的品种，所以，高产稳产的品种对应最右边且靠近水平线的品种图标。

如图 40-5 是以平均产量为 x 轴，以 IPCA1 为 y 轴的双标图。根据上述 3 个要点，不难分析出它们之间的对应关系。

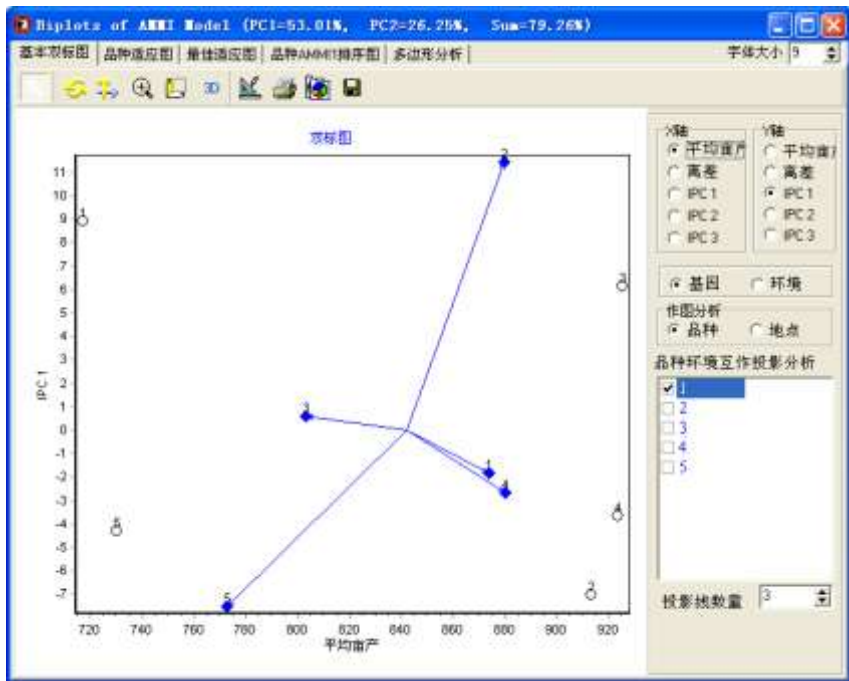


图 40-5 AMMI 模型用户图表分析及双标图示例

AMMI 品种适应图。如果忽略地点主效应，用品种主效应和 IPCA1 值估计产量比较品种的产量

$$y_i = u + \alpha_i + u_i x$$

其中， x 是以所有地点的 IPCA1 的可能值作为值域变量。这里不需要考虑地点效应是因为它只是将产量直线沿着产量坐标一块移上或移下，而我们只对每一地点所处于最高的直线感兴趣。这样，每一品种对应一条直线，在同一图上画出所有品种对应的直线，并在 x 轴内侧上标出的是参试地点代码，则处于最上端的折线对应的品种即为有最佳适应性的品种。本例中，从产量水平角度来看，在参试地点环境范围内，V1 和 V2 两个品种表现较佳（图 40-6）。

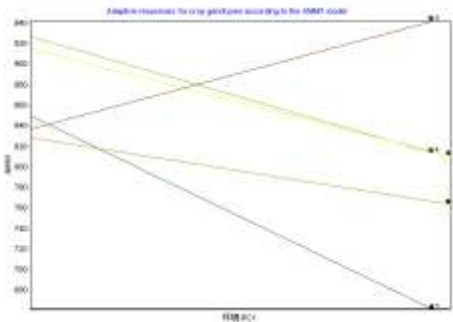


图 40-6 品种适应性分析图

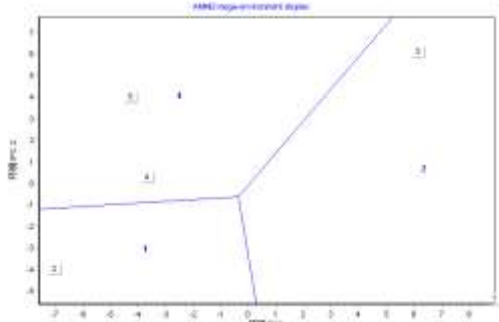


图 40-7 AMMI 模型分析的品种最佳适应图

图 40-7 是对同一区域的地点给出的具有最高理论产量 AMMI2 预测值的品种的最佳适应图。如这里，在地点 4 和地点 5，品种 4 具有最高的 AMMI2 理论产量。

图 40-8 所示的 AMMI 分析排序图，是以品种的平均产量为 x 轴，以 IPCA1 为 y 轴的 x - y 曲线图。在图中当指定某一品种后，作出通过该点的两条曲线，这时通过它们的空间位置可以分析其品种特性，即指定品种特性优于那些落在它左边的品种，如该图表示品种 V1 好于 V3 和 V5。

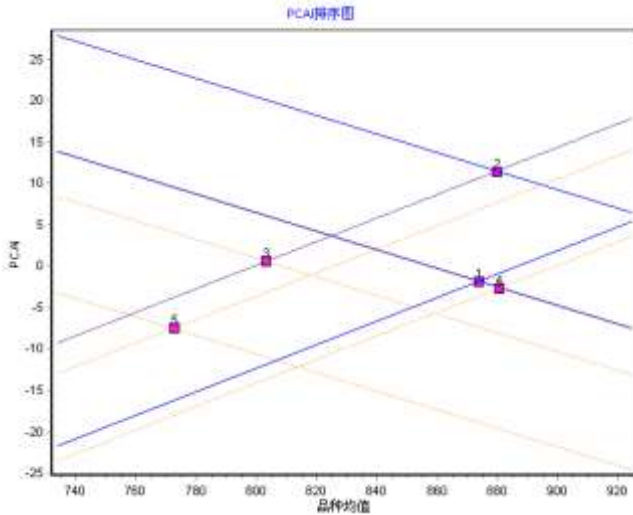


图 40-8 品种 AMMI 分析排序图

40.4.4 无区组品比试验的 AMMI 模型分析

当参与区试的品种没有设置区组时，无区组品比试验的 AMMI 模型分析，其数据编辑格式是以行代表环境(地点)，以列代表基因型(品种)，按试验地点的顺序，输完一个地点的数据后再输下一个地点的数据，在一个地点中依次输入各品种的数据。编辑数据并定义数据块后，进入菜单操作，选择相应功能项，按系统提示输入试验点数(E)、试验品种数(G)，然后运行并输出分析结果。

例如，对 5 个品种在 12 个地点(无重复)品比资料(Cauch, 1992)进行 AMMI 模型分析，可按如下方式输入数据并编辑定义数据块(图 40-9)。然后在菜单下选择“品种比较试验 AMMI 模型分析(无区组)”功能项，系统运算后给出计算结果，包括各方差分析表、品种和地点的两向列表、各个阶次 AMMI 模型预测估计值等。

	A	B	C	D	E	F	G	H	I
1	地	品	第一年		第二年		第三年		
2	点	种	R1	R2	R1	R2	R1	R2	
3	1	1	6	7	5	6	6	4	
4		2	9	8	15	13	10	11	
5		3	7	8	9	9	7	7	
6		4	14	14	16	12	13	12	

7		5	8	6	7	8	6	8	
8		1	7	8	7	6	8	6	
9		2	9	9	14	11	10	11	
10	2	3	3	4	3	5	3	3	
11		4	20	18	16	18	19	18	
12		5	9	5	8	8	7	6	
13		1	10	12	13	11	9	10	
14		2	9	10	13	15	9	10	
15	3	3	4	2	6	4	4	5	
16		4	13	14	15	11	17	14	
17		5	8	8	6	7	8	4	
18		1	5	6	6	6	7	5	
19		2	8	9	10	11	10	10	
20	4	3	5	4	3	4	4	4	
21		4	15	15	14	13	16	14	
22		5	7	6	9	8	8	6	
23									

图 40-9 无区组 AMMI 模型分析数据编辑格式

40.5 SHMM 模型

SHMM 模型是一种广义的主成分分析模型。主成分分析之前，将所有的数据均减去数据集的总的平均值(μ)，再获得主成分的特征值和特征向量。如果选择一个一般的常数 ψ (与总的平均值(μ)不同)，以使主成分分析提取更多的处理平方和，这样就得到 SHMM 模型

$$y_{ger} = \varphi + \sum_n \lambda_n \xi_{gn} \eta_{en} + \rho_{ge} + \varepsilon_{ger}$$

注意，正规的主成分分析模型和 SHMM 模型有相同的模型结构，只是在进行主成分分析之前，先求出 ψ 。由于 ψ 本身依赖于主成分分析结果，因此只能采用迭代算法来求出 ψ 。DPS 中采用的即是这种算法。

在 DPS 数据处理平台上进行 SHMM 模型分析，数据排列与无重复的 AMMI 模型分析相同：列为品种，行为地点，按试验地点的顺序，输完一个地点的数据后再输下一个地点的数据，在一个地点中依次输入各品种的数据。若有 E 个地点， G 个品种，其资料输入顺序为(如有缺区，在相应的位置输入-1)

	A	B	C	D	E	F	G
1		品种 1	2	3	4	5	
2	地点 1	352	249	266	231	182	
3	2	298	295	226	254	199	
4	3	331	251	219	196	216	
5	4	268	253	243	199	234	
6	5	271	270	215	205	184	

7	6	232	218	205	217	221
8	7	201	165	154	157	171
9	8	170	214	206	176	159
10	9	169	206	153	144	195
11	10	125	180	159	170	121
12	12	121	133	146	180	138
13	13	102	158	124	139	164
14						

输入数据并定义数据块后，进入菜单操作，选择相应功能项，按系统提示输入试验点数(E)、试验品种数(G)，然后运行并输出分析结果。例如，对 5 个品种在 12 个地点(无重复)品比资料(Cauch, 1992)进行 SHMM 模型分析，可按上图方式输入数据并编辑定义数据块(图 40-9)。

在菜单下选择“专业统计→品种比较试验→SHMM 模型分析”功能项，系统运算后给出计算结果，包括参数 ψ 的估计值，选定的前 n 个特征值，SHMM 模型的方差分析表，各个品种和地点的主成分及特征向量，以及前 n 阶次 SHMM 模型拟合值等。本例计算结果如下：

参数 B=	154.4558				
特征值 v1=	517.7407				
特征值 v2=	100.4812				
SHMM 分析结果					
变异来源	DF	SS	MS	F	p
处 理	59	167438.00			
SHMM1	16	143599.20	8974.950	18.93954	0.000000
SHMM2	14	10096.47	721.1761	1.521874	0.164910
误 差	29	13742.34	473.8737		
变 量	PCA1	PCA2	特征向量 1	特征向量 2	
G_1	15.59593	-6.642890	0.685418	-0.662696	
G_2	11.35580	6.995609	0.499070	0.697884	
G_3	8.427692	0.110168	0.370384	0.010990	
G_4	6.717521	2.345329	0.295225	0.233971	
G_5	5.422390	1.379114	0.238306	0.137581	
E_1	11.12161	-4.190572	0.488778	-0.418053	
E_2	10.32925	3.308294	0.453955	0.330036	
E_3	9.669813	-3.064816	0.424974	-0.305747	
...	
E_12	-2.098726	3.451493	-0.092236	0.344322	
模型拟合双向列表[列表示品种(基因)，行表示地点(环境)]					
SHMM1					
327.9078	280.7506	248.1854	229.1655	214.7616	
315.5502	271.7527	241.5076	223.8428	210.4651	
305.2656	264.2643	235.9500	219.4130	206.8893	
285.9925	250.2310	225.5353	211.1116	200.1885	
279.1559	245.2532	221.8410	208.1670	197.8115	
248.9792	223.2807	205.5342	195.1692	187.3197	

183.0304	175.2617	169.8969	166.7635	164.3906
200.3140	187.8463	179.2365	174.2080	170.3998
183.0575	175.2814	169.9115	166.7752	164.4001
148.1899	149.8934	151.0698	151.7569	152.2773
131.7333	137.9110	142.1771	144.6687	146.5557
121.7242	130.6231	136.7684	140.3576	143.0757
SHMM2(略)				

即本例的参数 ψ 的估计值=153.9180(总的平均值为 200.0000), 特征值 $v_1=579.7533$, $v_2=109.6370$ 。

SHMM 模型结果方差分析表中, 来源于 SHMM1 均方为 10062.98, F 值为 22.53932, 显著水平 $p=0.000000$; 来源于 SHMM2 的均方为 801.3517, F 值为 1.794887, 显著水平 $p=0.080882$ 。

另外给出了品种和环境各个变量的平均值、离差、主成分值及特征向量, 以及 SHMM 模型拟合两向列表(注: 列表示品种, 行表示地点环境)。

40.6 增广随机区组设计试验

增广设计(augmented design)主要是为了克服当有多个对照时对照小区占用试验处理较多而提出的。增广设计中, 若干个对照或需要比较的处理(对照处理)仍采用标准设计, 而“增广”的处理则安排选系。增广设计的主要特点是试验可以同时安排有重复的对照(品种)和无重复的增广处理(选系)。对照按一般设计在各个区组中进行重复试验, 增广处理一般较多, 但仅在试验的某一区组里面出现一次。

增广试验结果可以进行 3 种不同类型比较, 即增广处理间、对照处理间和增广处理与对照处理间的比较。

增广设计适用于大量的无重复的处理筛选试验, 因为增广处理数目任意, 所以处理较灵活。增广设计有增广随机区组设计、增广拉丁方设计等, DPS 目前提供的使用较多的增广随机区组设计。

40.6.1 方法简介

增广随机区组设计, 设试验有 b 个区组, v^* 个处理($v^*=v+v_1$)。这里 v 为对照处理, 即 v 个品种, 另外有 v_1 个增广处理, 即 v_1 个选系。增广随机区组设计是使 v 个对照品种在每一区组中均有一个小区, 而选系 v_1 仅在某一区组中各有(且仅有)一个小区。

增广随机区组设计需要对对照品种和选系进行随机化, 试验方案的设计较为复杂。DPS 中提供了增广随机区组设计生成方法。用户在第一行依次放入各个对照品种名称, 第二行依次放入各个选系的名称, 再执行增广随机区组设计即可。例如有 4 个区组($b=4$), 4 个对照品种(Va, Vb, Vc, Vd), 22 个选系(编号为 1~22)。进行增广随机区组设计, 首先将各个品种、选系名称编辑如图 40-10 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

1	对照	Va	Vb	Vc	Vd																				
2	选系	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22		
3																									

图 40-10 增广随机区组设计数据编辑格式

然后执行“试验设计”下“随机区组设计”中的“增广随机区组设计”，并在系统提示学输入区组数“4”，即可得到如下试验设计方案。根据试验设计方案，可实施试验。

编号	区组 1	区组 2	区组 3	区组 4
1	Vb	18	16	Vb
2	9	Vb	8	19
3	4	13	Vc	6
4	Vc	Vd	Vd	1
5	3	12	7	Vd
6	Vd	5	11	Vc
7	2	14	Vb	Va
8	Va	20	15	21
9	10	Vc	Va	17
10	22	Va		

40.6.2 增广随机区组试验统计分析

增广随机区组设计试验结果，因其中的增广处理效应和区组效应相互混杂，分析时一般采用不完全区组的区组内方差分析，对各个处理效应进行调整，以消除区组效应的混杂。

若增广随机区组试验有 b 个区组， v 个对照品种， v_1 个增广处理(选系)。设第 i 个对照品种在第 j 个区组的观察值为 y_{ij} ，而在第 j 个区组的第 k 个选系观察值为 z_{kj} ，并设 T_i 为第 i 个对照处理的总和， $T_y = \sum_{i=1}^v T_i$ ，为各个处理的总和； B_{yj} 为对照品种在第 j 个区组总和， B_j 为第 j 个区组的对照品种和选系的总和， G 为全部试验数据的总和。这时方差分析的平方和分解计算为

$$\begin{aligned}
C &= G^2 / (bv + v_1), & SS_{B_y} &= \frac{1}{v} \sum_{j=1}^b B_{yj}^2 - \frac{1}{bv} T_y^2 \\
SS_T &= \sum_{i=1}^v \sum_{j=1}^b y_{ij}^2 + \sum_{k=1}^{v_1} z_{kj}^2 - C, & SS_{I(adj)} &= SS_I - SS_{B_y} - SS_b \\
SS_b &= \sum_{j=1}^b \frac{1}{N_j} B_j^2, & SS_v &= \frac{1}{b} \sum_{i=1}^v T_i^2 - \frac{1}{bv} T_y^2 \\
SS_I &= \frac{1}{b} \sum_{i=1}^v T_i^2 + \sum_{k=1}^{v_1} z_{kj}^2 - C, & SS_s &= SS_{I(adj)} - SS_v \\
& & SS_e &= SS_T - SS_b - SS_{I(adj)}
\end{aligned}$$

根据各项平方和，列出方差分析表如表 40-5。

表 40-5 增广随机区组设计区组内方差分析表

变异来源	自由度	平方和	均方	F 值
------	-----	-----	----	-----

区 组	$b-1$	SS_b		
处理调整	$v+v_1-1$	$SS_{t(adj)}$	$MS_{t(adj)}$	$MS_{t(adj)}/MS_e$
其中: 对照品种	$v-1$	SS_v	MS_v	MS_v/MS_e
选 系	v_1	SS_s	MS_s	MS_s/MS_e
误 差	$(b-1)(v-1)$	SS_e	MS_e	
总 的	$bv+v_1-1$	SS_T		

表 40-5 中的“处理调整”项及其所属的“对照品种”、选系项的变异表示相应处理的选择潜力；若它们的 F 检验显著($p < 0.05$)，要进一步做多重比较。因为选系效应与区组效应混杂，故需对选系观察值 z_{kj} 进行矫正，再做多重比较。矫正的方法是：先求出第 j 个区组的效应 r_j ， $r_j = B_{.yj} / v - T_{.y} / (bv)$ ，注意，这里的 r_j 满足条件 $\sum_{j=1}^b r_j = 0$ ；然后计算

z_{kj} 进行矫正值 z'_{kj} ， $z'_{kj} = z_{kj} - r_j$ 。

多重比较时，各种差数的方差分别为：

两个对照品种平均值比较时， $D(\bar{y}_i - \bar{y}_j) = 2MS_e / b$ ；

同一区组两个选系平均值比较时， $D(z'_{ij} - z'_{kj}) = 2MS_e$ ；

不同区组两个选系平均值比较时， $D(z'_{ij} - z'_{kl}) = 2MS_e(1 + 1/v)$ ；

对照品种和选系之间平均值比较时， $D(\bar{y}_i - z'_{ij}) = 2MS_e(1 + 1/b + 1/v - 1/(bv))$ 。

40.6.3 增广随机区组试验统计分析例子

这里以一个简单例子说明增广随机区组设计的统计分析过程。若有一个对照品种 3 个，选系 8 个，3 个区组试验，先将原始数据整理成如图 40-11 左上部所示。

	A	B	C	D	E	F	G	H	I
1	对照品种	I	II	III		处理	区组	观察值	
2	a	6.8	6.8	7.7		a	I	6.8	
3	b	7.3	8.2	7.3		a	II	6.8	
4	c	5.4	6.8	6.4		a	III	7.7	
5		6.4(L1)	9.5(L6)	6.0(L8)		b	I	7.3	
6		7.7(L7)	4.1(L3)	7.3(L4)		b	II	8.2	
7		8.6(L2)	5.8(L5)			b	III	7.3	
8						c	I	5.4	
9						c	II	6.8	
10						c	III	6.4	
11						L1	I	6.4	
12						L7	I	7.7	
13						L2	I	8.6	
14						L6	II	9.5	
15						L3	II	4.1	
16						L5	II	5.8	

17						L8	III	6	
18						L4	III	7.3	
19									

图 40-11 增广随机区组数据及 DPS 处理格式

在 DPS 系统中, 增广随机区组设计试验的数据分析, 其数据格式须编辑成第 5 章中一般线性模型方差分析格式, 即图 40-11 中的阴影部分所示。这里共有 3 列, 第 1 列是处理(即对照品种、选系), 第 2 列是区组, 第 3 列是处理结果观察值。例如, 图 40-11 中左上部原始整理表中的第 2 个区组、品种 b 的观察值 8.2, 放在左边的第 5 行。

数据编辑完成之后, 用鼠标选中待分析数据, 在菜单方式下执行“专业统计→品种比较试验”下面的增广随机区组试验统计分析项, 系统会立即给出所有的统计分析结果。

输出结果的第一部分是方差分析表, 从方差分析结果可以看出, 处理(调整)间 F 值等于 8.8043, $p=0.0252(p<0.05)$, 差异显著, 其中(对照)品种间 F 值等于 5.4007, $p=0.0730(p>0.05)$, 说明对照品种间差异不显著; 选系间 F 值等于 9.6552, $p=0.0219(p<0.05)$, 表明选系间差异显著。

变异来源	平方和	自由度	均方	F 值	p 值	备注
A	24.5053	10	2.4505	8.8043	0.0252	处理
B	1.0067	2	0.5033	1.8084	0.2758	区组
品种	3.0064	2	1.5032	5.4007	0.0730	品种
选系	21.4990	8	2.6874	9.6552	0.0219	选系
ERR	1.1133	4	0.2783			
总变异	25.7024	16				

各个处理间多重比较, 系统首先给出了对照品种间的比较结果。作为例子, 尽管对照品种间 F 检验不显著, 仍加以解释: 在 5% 显著水平下, 品种 b 和品种 c 的均数有差异, 其他比较组别之间没有差异。

类别	理论均值	样本均值	标准差	样本数
A	7.1000	7.1000	0.5196	3
B	7.6000	7.6000	0.5196	3
C	6.2000	6.2000	0.7211	3
LSD05=1.1960		LSD01=1.9833		
处理	均值	5% 显著水平	1% 极显著水平	
2	7.6000	a	A	
1	7.1000	ab	A	
3	6.2000	b	A	

各区组内选系均数之间的比较, DPS 输出了每个区组内选系的多重比较结果, 并给出了区组内选系两两之间差异 LSD 测定的精确显著水平。本例中, 区组 I 中, 选系 L1 和 L7 间均数差异显著, 其他选系间差异不显著; 区组 II 中, 选系 L6 和 L5、L3 间均数差异显著, 但 L3 和 L5 间差异不显著; 区组 III 中, 选系 L4 和 L8 间均数差异不显著。

区组 1 内选系因素的多重比较			
类别	理论均值	样本均值	
L1	6.8667	6.4000	
L7	8.1667	7.7000	
L2	9.0667	8.6000	
LSD05=2.0715		LSD01=3.4351	
处理	均值	5%显著水平	1%极显著水平
3	9.0667	a	A
2	8.1667	ab	A
1	6.8667	b	A
区组 2 内选系因素的多重比较			
类别	理论均值	样本均值	
L6	9.2000	9.5000	
L3	3.8000	4.1000	
L5	5.5000	5.8000	
LSD05=2.0715		LSD01=3.4351	
处理	均值	5%显著水平	1%极显著水平
1	9.2000	a	A
3	5.5000	b	B
2	3.8000	b	B
区组 3 内选系因素的多重比较			
类别	理论均值	样本均值	
L8	5.8333	6	
L4	7.1333	7.3000	
LSD05=2.0715		LSD01=3.4351	
处理	均值	5%显著水平	1%极显著水平
2	7.1333	a	A
1	5.8333	a	A

区组间各个选系均数间多重比较，以矩阵形式给出。比较表中的数值为差值，差值后面跟有一个“*”表示在 5%显著水平上有差异，有两个“**”表示在 1%显著水平上有差异，没有星号表示差异不显著。

不同区组两选系比较，LSD05=2.3920，LSD01=3.9665						
不同区组间各个选系均值差异比较						
选系	L1	L7	L2	L6	L3	L5
L6	2.3333	1.0333	0.1333			
L3	-3.0667*	-4.3667**	-5.2667**			
L5	-1.3667	-2.6667*	-3.5667*			
L8	-1.0333	-2.3333	-3.2333*	-3.3667*	2.0333	0.3333
L4	0.2667	-1.0333	-1.9333	-2.0667	3.3333*	1.6333

系统最后以矩阵形式输出了对照品种和选系间均数差异比较。比较表中给出了对照品种均数、选系的矫正均数，以及它们之间数值的差值。差值大于零表示选系均数高于对照品种的均数，反之亦然。差值后面跟有一个“*”表示在 5%显著水平上有差异，有两个“**”表示在 1%显著水平上有差异，没有星号表示差异不显著。

品种和选系之间差异比较				
	LSD05=2.5836		LSD01=4.2843	
	品种	a	b	c
选系	校正均值	7.1000	7.6000	6.2000
L1	6.8667	-0.2333	-0.7333	0.6667
L7	8.1667	1.0667	0.5667	1.9667
L2	9.0667	1.9667	1.4667	2.8667*
L6	9.2000	2.1000	1.6000	3.0000*
L3	3.8000	-3.3000*	-3.8000*	-2.4000
L5	5.5000	-1.6000	-2.1000	-0.7000
L8	5.8333	-1.2667	-1.7667	-0.3667
L4	7.1333	0.0333	-0.4667	0.9333

参 考 文 献

- 胡秉民, 张全德. 1985. 农业试验统计分析方法[M]. 杭州: 浙江科学技术出版社.
- 黄金龙等. 1991. 电子计算机在遗传育种中的应用[M]. 北京: 中国农业出版社.
- 莫惠栋. 1992. 农业试验统计(第二版)[M]. 上海: 上海科技出版社.
- 王磊, 杨仕华, 沈希宏, 谢美贤. 1998. 主效可加互作可乘图形[J]. 南京农业大学学报, 21(2): 18-23.
- 王磊, 杨仕华, 沈希宏, 谢美贤. 1997. AMMI 模型及其在作物区试数据分析中的应用[J]. 应用基础与工程科学学报, 5(1): 39-46.
- 杨德. 2000. 试验设计与分析[M]. 北京: 中国农业出版社.
- 杨德, 戴君锡. 1991. 作物的典范性状及其遗传分析[C]// 刘垂珩. 作物数量性状的多元统计分析. 北京: 中国农业出版社.
- Federer W T. 1956. Augmented (or hoonuiaku) designs[J]. Hawaii. Plant. Rec, 2:191-208.
- Federer W T. 1961. Augmented designs with one-way elimination of heterogeneity[J]. Biometrics, 17:447-473.
- Gauch H G. 1988. Model selection and validation for yield trials with interaction[J]. Biometrics, 44: 705-715.
- Gauch H G. 2001. 产量区域试验统计分析——因子设计的 AMMI 分析[M]. 王磊等译. 北京: 中国农业出版社.
- Shukla G K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability[J]. Heredity 29:237-245.

第 41 章 遗传统计

数量遗传统计分析涵盖的内容很多，DPS 数据处理系统当前版本提供了目前应用较多的遗传参数估计以及遗传交配设计试验的统计分析技术。

世代平均数分析方法是检验遗传模型是否属于加性-显性模型的一种方法，因为育种学家在制定育种方案时，需要了解性状的遗传规律，即了解这些性状的遗传模型是属于加性-显性模型还是存在上位性效应，这不仅对制定杂交方案，而且对后代的选择都具有一定的指导意义。

重复力(率)、遗传力和遗传相关是非常重要的遗传参数。这些参数对于研究群体的数量性状的性质是必不可少的，在育种中也是估计其他育种参数的前提。

数量遗传统计分析涵盖的内容很多，DPS 数据处理系统当前版本提供了目前应用较多的遗传参数估计以及遗传交配设计试验的统计分析技术。

遗传交配设计试验，包括因为选择配合力高的自交系或品种进行杂交，有可能选育出杂种优势强的品种或新品系。

Comstock 和 Robinson(1952)在杂种优势利用的论文中，提出了三种试验设计，又称为 North Carolina 设计，分别简称为 NC I、NC II、NC III 设计，这里首先对这三个试验设计进行介绍。

在配合力育种分析中，介绍了完全双列杂交试验的 Griffing(1956)分析方法及 Hayman 分析法。

41.1 世代平均数分析方法

为了对数量性状进行分析，首先要确定其遗传模型。最简单的是加性-显性模型。因此要根据资料检验加性-显性模型是否合适。世代平均数分析就是检验加性-显性模型的统计方法。

41.1.1 世代平均数分析法的基本原理

世代平均数分析法的基本原理是认为任一数量遗传性状都可看成是加性效应和显性效应的一个线性函数。对加性-显性模型，可导出两亲本(P_1 , P_2)、自交一代(F_1)、自交二代(F_2)，以及 2 个回交一代(F_1 分别与大值亲本 P_1 和小值亲本 P_2 回交的子代 B_1 和 B_2)，这 6 个世代的样本均数与 Fisher 坐标值 d 、 h 、中亲值 m 之间关系，对多对独立基因，有 $\bar{P}_1 = M + [d]$ ， $\bar{P}_2 = M - [d]$ ， $\bar{F}_1 = M + [h]$ ， $\bar{F}_2 = M + [h]/2$ ， $\bar{B}_1 = M + [d]/2 + [h]/2$ ， $\bar{B}_2 = M + [d]/2 - [h]/2$ 。

加性-显性模型检验最简单的方法是 ABC 尺度检验。根据加性-显性遗传规律，各个世代平均数的关系为

$$\begin{aligned}\bar{B}_1 &= \frac{1}{2}\bar{P}_1 + \frac{1}{2}\bar{F}_1 \\ \bar{B}_2 &= \frac{1}{2}\bar{P}_2 + \frac{1}{2}\bar{F}_1 \\ \bar{F}_2 &= \frac{1}{4}\bar{P}_1 + \frac{1}{2}\bar{F}_1 + \frac{1}{4}\bar{P}_2\end{aligned}$$

这时为检验是否符合加性-显性模型的假设，引入 3 个参数

$$\begin{aligned}A &= \bar{P}_1 + \bar{F}_1 + 2\bar{B}_1 \\ B &= \bar{P}_2 + \bar{F}_2 - 2\bar{B}_2 \\ C &= \bar{P}_1 + \bar{P}_2 + 2\bar{F}_1 - 4\bar{F}_2\end{aligned}$$

如果试验所得数据的数量性状符合加性-显性遗传模型，那么上式应该成立，即参数 $A=0$ ， $B=0$ ， $C=0$ 成立。因此我们可以通过统计检验，推断 A ， B ， C 与期望值 0 是否有显著差异。

如尺度检验结果是 A ， B ， C 与期望值 0 有的有差异，出现矛盾的结果时，Cavalli 提出了同时利用 6 个世代资料来检验加性-显性模型的联合尺度检验方法。该方法将 6 个世代的平均数看成是 3 个系数遗传参数的因变量，求出回归系数，并以各个世代的变异大小加权，其遗传模型矩阵 G 、遗传参数矩阵 M 、世代均数矩阵 Y 及加权矩阵 V_Y^{-1} 分别为

$$G = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0.5 \\ 1 & 0.5 & 0.5 \\ 1 & -0.5 & 0.5 \end{bmatrix}, \quad M = \begin{bmatrix} m \\ [d] \\ [h] \end{bmatrix}, \quad Y = \begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \\ \bar{F}_1 \\ \bar{F}_2 \\ \bar{B}_1 \\ \bar{B}_2 \end{bmatrix}, \quad V_Y^{-1} = \begin{bmatrix} V_{P_1}^{-1} & & & & & \\ & V_{P_2}^{-1} & & & & \\ & & V_{F_1}^{-1} & & & \\ & & & V_{F_2}^{-1} & & \\ & & & & V_{B_1}^{-1} & \\ & & & & & V_{B_2}^{-1} \end{bmatrix}$$

其多元回归方程为

$$G' V_Y^{-1} G M = G' V_Y^{-1} Y$$

解方程得

$$\hat{M} = (G' V_Y^{-1} G)^{-1} (G' V_Y^{-1} Y)$$

遗传模型的拟合度可用卡方检验来检验。

当上面这两种检验都不能接受加性-显性模型时，Mather 和 Jinks 提出了所谓的 6 参数模型方法，其遗传模型系数矩阵为

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0.5 & 0 & 0 & 0.25 \\ 1 & 0.5 & 0.5 & 0.25 & 0.25 & 0.25 \\ 1 & -0.5 & 0.5 & 0.25 & -0.25 & 0.25 \end{bmatrix}, \quad M = \begin{bmatrix} [m] \\ [d] \\ [h] \\ [dd] \\ [dh] \\ [hh] \end{bmatrix}$$

6 参数模型的参数既可以通过解方程得到，也可以从 6 个世代的均值计算得到。应用 6 参数模型分析，因是饱和模型，故不能进行卡方检验。

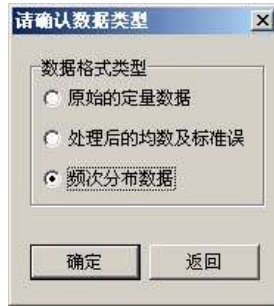
41.1.2 DPS 处理操作

在 DPS 上进行世代平均数的分析，其数据可以是定量的原始数据，也可以是频次分布数据，或是通过计算得到的平均值和标准误资料。各种类型数据都需按世代顺序整理、排列，并定义成数据块(如图 41-1 中的 A, B 和 C 所示)。

	A	B	C	D	E	F	G	H	I	J
1	世代	观察值					世代	均值	标准误	
2	P1	3.696	3.576	3.628	3.772		P1	10.36	0.581	
3	P2	4.516	4.588	4.62	3.688		P2	0.45	0.017	
4	F1	3.296	3.27	3.302	3.278		F1	2.33	0.13	
5	F2	3.53	3.238	3.269	3.336		F2	2.12	0.105	
6	B1	3.992	4.062	4.056	4.036		B1	4.82	0.253	
7	B2	4.078	4.374	4.026	3.912		B2	0.97	0.045	
8	A.原始数据编辑格式					B.计算处理后的数据编辑格式				
9										
10	组中值	P1	P2	F1	F2	B1	B2			
11	-11				3					
12	-9									
13	-7		9		4					
14	-5		43	1	14		4			
15	-3		34		26	1	5			
16	-1		28	2	74	6	20			
17	1		21	6	142	5	14			
18	3	7		6	128	31	13			
19	5	30			54	22				
20	7	46			22	21				
21	9	46			4	14				
22	11	4				1				
23	13	5								
24	15					1				
25	C.频次分布数据编辑格式									

图 41-1 世代平均数分析的数据编辑、定义格式

如果我们对第 3 组频次分布数据进行分析，在进入菜单操作，选择“专业统计”中的“数量遗传参数估计”下面的“世代平均数分析”功能后，系统会给出一个如下图所示的用户界面，让您确认一下当前分析的数据类型。



在该用户界面里，选择“频次分布数据”项，点击“确认”按钮后系统将输出如下结果：

计算结果	当前日期 2005-11-23 22: 24: 27			
世代	样本数	均值	标准误	S ² /n
P1	138	7.3623	0.1843	0.0340
P2	135	-2.8667	0.2052	0.0421
F1	15	1.1333	0.5679	0.3225
F2	471	1.4926	0.1380	0.0190
B1	102	4.8824	0.2999	0.0899
B2	56	-0.0357	0.3100	0.0961
ABC 尺度检验				
	均值	S.E.	t 值	p 值
A	1.2691	0.8463	1.4995	0.1690
B	1.6619	0.8654	1.9203	0.0755
C	-0.7920	1.2926	0.6128	0.4862
遗传参数	[d]=5.1145		[h]=-1.1145	
联合尺度检验				
模型拟合卡方值=0.0637			df=3	p=0.9958
M=2.7211 ±2.8921				
[d]=5.0604 ±2.8653				
[h]=-1.4748 ±3.6722				
Mather & Jinks(1977)模型				
M=-1.4752 ±3.6590				
[d]=5.1145 ±3.6471				
[h]=9.2625 ±65.2689				
[dd]=3.7230 ±30.4413				
[dh]=-0.3929 ±11.8030				

$[hh]=-6.6540 \pm 35.3637$

系统首先给出 ABC 尺度检验结果。ABC 尺度检验里面 A, B, C 均等于 0 是符合加性-显性模型的必要条件, 因此, 如果 A, B, C 中的任意一个明显地不等于 0 时就可以推断该模型是不合适的。在本例中 A, B 和 C 各个参数的 p 值均大于 0.05, 因此, 可认为该资料符合加性-显性遗传模型。

如果 ABC 尺度检验中的某个参数的 p 值小于 0.05, 则需继续看下面的联合尺度检验结果, 或 Mather & Jinks(1977)模型拟合结果。

41.2 遗 传 力

生物的各种性状同时受到基因和环境条件的影响, 当假定基因与环境的互作不存在时, 一般性状均有: 表现型值 $P =$ 基因型值 $G -$ 环境效应 E 。在现有的科学技术水平条件下, 人们能直接观测到的仅是生物个体的表现型。育种家们通常就是用这些个体的表现型值来代表个体的基因型值, 这就需要引入一个新参数来度量这种代替的可能性。所谓遗传学中的广义遗传力即是指通过表现型值预测基因型值的可靠性大小的比值, 更一般的定义是遗传方差 V_G 占表现型总方差 V_p 的百分比, 即

$$h_B^2 = V_G / V_p \times 100$$

由育种实践知, 遗传变异中只有加性效应引起的那部分变异才可能在上下代之间稳定传递, 即只有这部分变异才可能通过选择来加以固定。因此, 育种上更常用的是另一种遗传力上狭义遗传力。它的定义是观察群体中由加性效应引起的那部分方差(V_D)占表现型总方差的比例, 其公式为

$$h_N^2 = V_D / V_p \times 100$$

遗传力的第三个概念是指某性状的子代观察值平均数与双亲观察值平均数(中亲值)之间回归直线的斜率。这个概念刻画了一个群体中某性状从亲代稳定传递给子代的遗传传递能力, 它在其他某些遗传参数的估算中有重要作用。它的意义和计算公式见图 41-2。

设有一个组合两个世代的资料, 用坐标表示亲子两代性状之间的关系, 横坐标 x 表示双亲的中亲值, 纵坐标 y 表示子代平均数。现考虑一个平衡群体, 即如果不进行选择时两个世代的群体平均数应是相同的, 有 $\bar{x} = \bar{y} = \mu$ 。此时, 如果将各子代家系平均值 y 对亲代的中亲值 x 做回归, 回归直线应通过 (μ, μ) 点。因此有 $y - \mu = k(x - \mu)$ 。该线性方程中的斜率 k 表明了亲代与子代的相像程度, 如 $k=0$, 无论中亲值如何变化, 子代平均数均为一个常数, 说明亲子代无相像性; 当 $k=1$ 时, $y=x$, 即亲子完全相像。因此 k 作为衡量亲子代相像的指标, 通常被称作生统遗传力, 值得一提的是, 第三个概念下的遗传力的唯一基本假定是子代平均数与中亲值之间存在线性回归关系, 而按中心极限定理, 这种关系在大多数情况下是成立的。

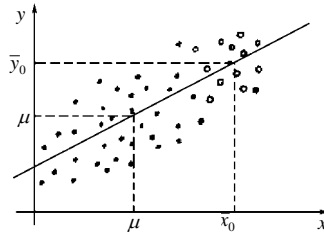


图 41-2 亲-子代关系图

遗传力是数量性状遗传分析的一个基本遗传参数。遗传力可以用于确定选择方法和预测选择进展。

41.2.1 遗传力估计的基本方法及 DPS 系统的实现

从用于遗传力估计的两类个体间的遗传关系来看，有亲子资料、同胞资料及同卵双生资料等；从计算的统计方法来看，有回归和相关分析法、方差分析法、最小二乘法、最大似然法等。在育种实践中，估计遗传力用得最多的资料类型是亲子资料和同胞资料。

1. 亲子资料估计遗传力

亲子资料估计遗传力，这里提供的是子女表型值对一个亲本的回归或相关估计。在随机交配情况下，亲子间的亲缘系数近似等于 0.5，假设 P_p 表示单亲(任何一个亲本)表型值， P_o 表示子女表型值，由于

$$\text{cov}(P_o, P_p) = \text{cov}[A_p + R_p, (A_p + A'_p)/2 + R_o] = \text{cov}(A_p, A_p/2) = V(A)/2$$

这里， A'_p 为另一亲本的育种值。如各 R 之间及 R 和育种值之间不相关，这时子代对一个亲本的回归估计为

$$b_{P_o P_p} = \text{cov}(P_p, P_o) / V(P_p) = 0.5V(A) / V(P_p) = h^2/2$$

这样可以得到利用简单回归或相关方法估计的遗传力。

在 DPS 系统中，应用亲子回归，估计遗传力，其数据编辑整理格式如图 41-3 中的阴影部分。

	A	B	C	D	E	F	G	H	I	J	K	L
1	观 察 值											
2	亲代	1.2	1.3	1.3	1.3	1.2	1.2	1.3	1.2	1.3	1.3	
3	子代	1.2	1.2	1.3	1.2	1.3	1.2	1.3	1.2	1.3	1.1	
4												

图 41-3 根据亲代和子代估计遗传力数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“根据亲代和子代估计遗传力”功能后，得到该例的遗传力估计值 $h^2=0.08333$ 。

2. 公畜内女母回归或相关估计遗传力

应用公畜内女母回归和相关法估计遗传力, 应剔除公畜效应, 计算公畜内母、女各自的方差及协方差, 然后计算公畜内女母回归系数或相关系数, 具体计算方法如下:

设有 S 头公畜, 各配 d_i 头母畜, 各有一个女儿, D_{ij} 和 O_{ij} 分别表示第 i 头公畜所配的第 j 头母畜表型值和其女儿表型值, $i=1, 2, \dots, d_i$; 记 $N = \sum d_i$ 。表 41-1 是其平方和、乘积和分析表。

表 41-1 公畜内女母平方和及乘积和分析表

变异来源	df	母亲平方和(SS _D)	女儿平方和(SS _O)	母女乘积和(SP)
公畜间	$S-1$	$SS_{T(D)}-SS_{W(D)}$	$SS_{T(O)}-SS_{W(O)}$	SP_T-SP_W
公畜内	$N-S$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (D_{ij} - \bar{D}_{i.})^2$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (O_{ij} - \bar{O}_{i.})^2$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (D_{ij} - \bar{D}_{i.})(O_{ij} - \bar{O}_{i.})$
总 和	$N-1$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (D_{ij} - \bar{D}_{..})^2$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (O_{ij} - \bar{O}_{..})^2$	$\sum_{i=1}^s \sum_{j=1}^{d_i} (D_{ij} - \bar{D}_{..})(O_{ij} - \bar{O}_{..})$

实际上, 只需要计算出与公畜内有关的平方和以及乘积和即可, 由于

$$\sigma_{W(D)}^2 = MS_{W(D)} = \frac{SS_{W(D)}}{df_W}, \quad \sigma_{W(O)}^2 = MS_{W(O)} = \frac{SS_{W(O)}}{df_W}, \quad \text{cov}_{W(D)} = MP_{W(D,O)} = \frac{SP_W}{df_W}$$

所以可得到

$$h^2 = 2b_{W(OD)} = \frac{2\text{cov}_{W(D)}}{\sigma_{W(D)}^2} = 2 \frac{SP_W}{SS_{W(D)}}$$

即

$$h^2 = 2r_{W(OD)} = \frac{2\text{cov}_{W(D)}}{\sigma_{W(O)}^2} = 2 \frac{SP_W}{\sqrt{SS_{W(D)}SS_{W(O)}}$$

通过上述方法估计出遗传力后, 一般均需要进行显著性检验, 以便说明两个问题: ①用来估计 h^2 的样本是否来自一个参数为 0 的总体, 即估计值是否真实存在, 与 0 有无实质性差异, 这检验了参数的真实性; ②总体参数 h^2 是否处在一个以估计值为中心的一定范围内, 即估计值是否能代表参数, 这检验了参数估计的准确性。但是, 显著性检验只能对参数本身检验, 而并不能检验设计和统计方法是否正确。当然, 不同的统计方法其显著性检验方法也不一样, 一般可以采用统计量与统计量的标准误比值进行 t 检验。如果参数是由于方差分析估计的, 也可以用 F 检验替代, 但是这两种检验方法的结果并不总是一致的。

利用亲子资料估计遗传力时的显著性检验, 均可以转化为相应的回归系数(b)或相关系数(r)的显著性检验。设有常数 c 使得 $h^2=cb$ 或 $h^2=cr$, 则

$$t_b = \frac{h^2}{\sigma_{h^2}} = \frac{cb}{c\sigma_b} = \frac{b}{\sigma_b}, \quad t_r = \frac{h^2}{\sigma_{h^2}} = \frac{cr}{c\sigma_r} = \frac{r}{\sigma_r}$$

而

$$\sigma_b = \sqrt{\frac{1}{N-2} \left(\frac{SS_O - b^2 SS_P}{SS_P} \right)} = \sqrt{\frac{1}{N-2} \left(\frac{\sigma_O^2}{\sigma_P^2} - b^2 \right)}, \quad \sigma_r = \sqrt{\frac{1-r^2}{N-2}}$$

这里 N 为亲子总对子数。但是，当用公畜内女母相关或回归估计时，误差自由度(离回归自由度)不是 $N-2$ ，而是 $N-S-1$ ，需要它代替上两式中的 $N-2$ 。

在 DPS 系统中，应用公畜内女母回归和相关法估计遗传力，其数据编辑整理格式如图 41-4 中的阴影部分。

	A	B	C	D	E	F	G	H	I	J	K	
1			母 女 对									
2	公畜 A	母亲 p	20	14.7	16.8	17.9	11.2	13.4	18	15		
3		女儿 O	16.4	17	18.5	19.5	16.3	12.5	14	11		
4	公畜 B	母亲 p	19.3	10.9	17.6	16.3						
5		女儿 O	16.4	14.6	17.3	15.5						
6												

图 41-4 公畜内女母回归或相关估计遗传力数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“公畜内女母相关、回归估计遗传力”功能后，应用相关法估计得到的遗传力估计值 $h^2=0.7046$ ，标准误为 0.4488， t 检验统计量为 1.5702，概率 p 值为 0.1197；应用回归法得到的遗传力估计值 $h^2=0.5752$ ，标准误为 0.5093， t 检验统计量为 1.1294，概率 p 值为 0.2615。

3. 同胞资料估计遗传力

如前所述，估计遗传力需要两类遗传关系明确的亲属个体资料，而同胞之间的亲缘关系是可以确定的，因此，计算出同胞组内相关系数就可估计出遗传力。在动物育种中，同胞资料是比较易于得到而且数量较大，因而广泛用于遗传参数估计。一般有半同胞组内相关法和全同胞组内相关法两种。然而，这两种方法的应用都是有一定条件的，下面分别讨论这两种估计方法。

(1) 半同胞组内相关法。

在近似条件下半同胞间的亲缘系数为 $r_{(HS)}=0.25$ ，则 $cov(A_i, A_j)=r_{(HS)}\sigma_A^2=0.25\sigma_A^2$ 。因半同胞个体很多， r_{HS} 实际上是一个组内相关系数，一般通过单因方差分析估计方差组分来计算，它应满足下列条件：① 所有变异来源除公畜个体效应外，其余的应全部为随机误差效应；② 公畜内个体间的遗传关系必须完全一致，即不应有全同胞个体存在，否则会使组内变异 σ_W^2 减小，遗传力估计偏高；③ 公畜间应无亲缘相关，否则有亲缘相关的公畜组的半同胞间一致性增加，导致公畜间变异 σ_S^2 下降，遗传力估计偏低。

从实际应用来说，主要要求公畜间没有亲缘相关，所配母畜是从群体中随机选留和随机选配的，每头母畜只选取一个后代，这些后代都在同一环境条件下培育。可见，这些条件是相当苛刻的，实际育种工作很难满足这些条件。但是，如果能基本满足的话，即可按下面的方法估计遗传力。

设共有 S 头公畜，每一头公畜有 n_i 个半同胞子女， $i=1, 2, \dots, S$ 。根据各子女观察值，计算得到其均方值如表 41-2。这里

$$k_0 = \frac{1}{df_S} \left(\sum_{i=1}^S n_i - \frac{\sum_{i=1}^S n_i^2}{\sum_{i=1}^S n_i} \right)$$

因此可得到

$$h^2 = 4r_{HS} = \frac{4\sigma_S^2}{\sigma_S^2 + \sigma_e^2} = \frac{4(MS_S - MS_e)}{MS_S + (k_0 - 1)MS_e}$$

估计出遗传力后均需进行显著性检验。用半同胞资料估计，由于其估计方法与组内相关系数完全一样，因此遗传力的显著性检验可转化为组内相关系数的显著性检验，即

$$t = \frac{h^2}{\sigma_{h^2}}, \quad \sigma_{h^2} = \sigma_{4r_{HS}} = 4\sqrt{\frac{2(1-r_{HS})^2 [1 + (k_0 - 1)r_{HS}]^2}{(S-1)k_0(k_0-1)}}$$

表 41-2 半同胞资料方差分析

变异来源	自由度	平方和	均方	期望均方
公畜间	$S-1$	$\frac{\sum_{i=1}^S \left(\sum_{j=1}^{n_i} x_{ij} \right)^2}{n_i} - \left(\frac{\sum_{i=1}^S \sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^S n_i} \right)^2$	SS_S/df_S	$\sigma_e^2 + k_0\sigma_S^2$
公畜内	$\sum n_i - S$	$\sum_{i=1}^S \sum_{j=1}^{n_i} x_{ij}^2 - \frac{\sum_{i=1}^S \left(\sum_{j=1}^{n_i} x_{ij} \right)^2}{n_i}$	SS_e/df_e	σ_e^2
总和	$\sum n_i - 1$	$SS_S + SS_e$		

在 DPS 系统中，根据半同胞资料估计遗传力，其数据编辑整理格式如图 41-5 中的阴影部分。

	A	B	C	D	E	F	G	H	I
1	父本	半同胞子女观察值							
2	1	3.8	3.5	3.6	3.4	3.5			
3	2	3.3	3.4	3.7	3.2	3.5	3.4		
4	3	3.6	3.7	3.5	3.6	3.4	3.8	3.5	
5	4	3.9	3.4	3.6	3.8	3.5			
6									

图 41-5 半同胞资料遗传力估计的数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“半同胞资料估计遗传力”功

能后, 得到遗传力估计值 $h^2=0.5460$, 方差 $vs=0.8934$, t 值为 0.6111, 概率 p 值为 0.7997。

(2) 全同胞组内相关法。

在近似条件假设下, 全同胞个体间的亲缘系数为 $r_{(FS)}=0.5$, 则有

$$\text{cov}_{FS}(A_i, A_j) = r_{(FS)}\sigma_A^2 = 0.5\sigma_A^2$$

因全同胞个体也很多, r_{FS} 实际上也是一个组内相关系数, 一般通过二因素系统分组方差分析估计出各方差组分来计算。这样的分析应满足下列条件: ① 所有变异来源除公畜个体效应和母畜个体效应外, 其余效应全部为随机误差效应; ② 公畜间的母畜不交叉; ③ 母体效应不存在或很小, 可以忽略不计, 否则, 全同胞间会因母体效应导致组内变异 σ_W^2 变小, 遗传力估计偏高; ④ 不同公畜, 母畜间均无亲缘相关, 否则存在亲缘相关的两组间的全同胞个体一致性增加, 导致组间变异 σ_S^2 或 σ_D^2 下降, 遗传力估计偏低。

此外, 由遗传学知识可知, 全同胞亲缘协方差中实际上包含有较大的显性效应组分, 在实际估计遗传力时应引起注意。由此可见, 在实际应用中, 上述条件同样是相当苛刻的。只要基本能满足这些条件, 就可以按下列方法估计遗传力。

设有 S 头公畜, 各配 d_i 头母畜, $i=1, 2, \dots, S$; 每头母畜有 n_{ij} 头子女, $j=1, 2, \dots, d_i$ 。记为

$$d_i = \sum_{j=1}^{d_i} d_i, \quad n_{i.} = \sum_{j=1}^{d_i} n_{ij}, \quad n_{..} = \sum_{i=1}^S n_{i.}$$

各子女观察值的方差分析结果如表 41-3。

表 41-3 全同胞资料遗传力估计

变异来源	自由度	平方和	均方	期望均方
公畜间	$S-1$	$\sum_{i=1}^S \frac{x_{i..}^2}{n_{i.}} - \frac{x_{..}^2}{n_{..}}$	SS_S/df_S	$\sigma_e^2 + k_2\sigma_D^2 + k_3\sigma_S^2$
公畜内母畜间	$d-1$	$\sum_{i=1}^S \sum_{j=1}^{d_i} \frac{x_{ij.}^2}{n_{ij}} - \sum_{i=1}^S \frac{x_{i..}^2}{n_{i.}}$	SS_d/df_d	$\sigma_e^2 + k_1\sigma_D^2$
公母畜内子女间	$n_{..}-d$	$\sum_{i=1}^S \sum_{j=1}^{d_i} \sum_{k=1}^{n_{ij}} x_{ijk}^2 - \sum_{i=1}^S \sum_{j=1}^{d_i} \frac{x_{ij.}^2}{n_{ij}}$	SS_e/df_e	σ_e^2
总和	$n_{..}-1$	$SS_S+SS_D+SS_e$		

这里

$$k_1 = \frac{1}{df_D} = \left(n_{..} - \sum_{i=1}^S \sum_{j=1}^{d_i} \frac{n_{ij}^2}{n_{i.}} \right), \quad k_2 = \frac{1}{df_S} = \left(\sum_{i=1}^S \sum_{j=1}^{d_i} \frac{n_{ij}^2}{n_{i.}} - \sum_{i=1}^S \sum_{j=1}^{d_i} \frac{n_{ij}}{n_{..}} \right), \quad k_3 = \frac{1}{df_S} = \left(n_{..} - \frac{1}{n_{..}} \sum_{i=1}^S n_{i.}^2 \right)$$

由全同胞资料作二因系统分组方差分析可以得到三个遗传力估计值, 即

$$h_S^2 = \frac{4\sigma_S^2}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2}, \quad h_D^2 = \frac{4\sigma_D^2}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2}, \quad h_{SD}^2 = \frac{2(\sigma_S^2 + \sigma_D^2)}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2}$$

从理论上讲，这三个遗传力估计值应该相等。然而，在实际估计时，偏差总是存在的，在满足全同胞资料估计遗传力的条件下，一般以 h_{SD}^2 为准，因为 $h_{SD}^2 = (h_S^2 + h_D^2)/2$ ，它是公畜组分估计的平均值。但是，当这三个估计值相差太大时，应设法进一步收集资料，重新估计，或者应考虑到有显著的显性效应或母体效应存在。

全同胞资料估计遗传力的显著性检验，Kempthorne(1957)提出采用大样本理论，利用均方来估计遗传力标准误。Dickerson(1959)证明，当分子均方的自由度中等大时，该方法可以提供满意的近似估计值。因此，三个遗传力估计值的标准误可如下计算

$$\textcircled{1} \text{ 公畜组分估计 } h_S^2 = \frac{4\sigma_S^2}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2} = \frac{4\sigma_S^2}{\sigma_P^2};$$

$$\textcircled{2} \text{ 母畜组分估计 } h_D^2 = \frac{4\sigma_D^2}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2} = \frac{4\sigma_D^2}{\sigma_P^2};$$

$$\textcircled{3} \text{ 公母畜组合并估计 } h_{SD}^2 = \frac{2(\sigma_S^2 + \sigma_D^2)}{\sigma_S^2 + \sigma_D^2 + \sigma_e^2} = \frac{2(\sigma_S^2 + \sigma_D^2)}{\sigma_P^2}.$$

由上可见，估计遗传力的显著性 t 检验是比较麻烦的，因此，在要求不太严格时，可用方差分析的 F 检验近似替代，或用组内相关系数的显著性检验替代。但后两种检验方法的误差，特别是 F 检验的误差是很大的。若用组内相关系数检验替代，则有

$$t = \frac{h^2}{\sigma_{h^2}}, \quad \sigma_{h^2} = \sigma_{2r_{FS}} = 2\sqrt{\frac{2(1-r_{FS})^2 [1+(k_1-1)r_{FS}]^2}{k_1(k_1-1)(d-1)}}$$

即用 k_1 替代 k_0 ， $d-1$ 替代 $n-1$ 。

在 DPS 系统中，根据全同胞资料估计遗传力，其数据编辑整理格式如图 41-6 中的阴影部分。

	A	B	C	D	E	F	G	H	
1	父本	母本	全同胞子女观察值						
2	1	1	403	392	382	395			
3	1	2	404	395					
4	1	3	397	406	425	418	394		
5	2	1	395	412	407				
6	2	2	384	395					
7	2	3	410	382	394	390			
8	2	4	405	396	387				
9	2	5	382	395	406	392	403		
10	3	1	405	415	396	418	407		
11	3	2	423	397	409	411			
12	3	3	405	418	392				
13	3	4	385	402	394				
14									

图 41-6 半同胞资料遗传力估计的数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“全同胞资料估计遗传力”功能后，得到父本、父本内母本及全同胞的遗传力估计值分别为 0.3555，0.4161 和 0.3858，并给出了它们的标准误和统计检验结果。

方差分析表					
变异来源	平方和	自由度	均方	F 值	p 值
父本间	642.4938	2	321.2469	2.1137	0.1768
父本内母本间	1367.8706	9	151.9856	1.4478	0.2112
父母本内	3254.3333	31	104.9785		
总变异	5264.6977	42			
各有效平均子女数					
母本平均子女数=3.4744					
父本内每母本平均子女数=3.9232					
每父本平均子女数=14.1163					
方差组分					
$\sigma e^2=104.9785$					
$\sigma D^2=13.5294$					
$\sigma S^2=11.5604$					
$\sigma T^2=130.0683$					
遗传力组份	遗传力	方差	t 值	p 值	
父本 h_s^2	0.3555	0.7170	0.4958	0.6235	
父本内母本 h_d^2	0.4161	0.6767	0.6149	0.5431	
全同胞 h_{sd}^2	0.3858	0.5408	0.7133	0.4810	

41.2.2 阈性状遗传力

如前所述，无论是单阈或是多阈性状，均假定它们是由于多基因系统控制，具有一个类似于数量性状的潜在连续分布，只不过需要超过一定阈值，表现型才发生改变，呈现出非连续性分布。当然，表现型同样地受到遗传控制和环境效应的共同作用。因此，与普通数量性状一样，需要确定这两种效应作用的大小，即估计出阈性状的遗传力，仍可用方差分析来估计。

对给定资料作半同胞方差分析，设 x_{ij} 为第 i 公畜的第 j 个后代取值，则第 i 公畜的全部后代里，有 a_i 个 x_{ij} 取值 1， $n_i - a_i$ 个 x_{ij} 取值 0，因此 $\sum_{j=1}^{n_i} x_{ij} = a_i$ 。针对这种由 0~1 取值构成的半同胞资料，可作如下的方差分析，这时的公畜组间平方和 (SS_S) 及组内平方和 (SS_W) 为

$$SS_S = \sum_{i=1}^s \left\{ \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 / n_i \right\} - \left(\sum_{i=1}^s \sum_{j=1}^{n_i} x_{ij} \right)^2 / \sum_{i=1}^s n_i, \quad SS_W = \sum_{i=1}^s \sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^s \left\{ \left(\sum_{j=1}^{n_i} x_{ij} \right)^2 / n_i \right\}$$

自由度为 $df_S = s-1$ 和 $df_W = \sum_{i=1}^s n_i - s$ ，由此可估计出组内方差和组间方差组分为

$$\hat{\sigma}_W^2 = MS_W = \frac{SS_W}{df_W}, \quad \hat{\sigma}_S^2 = \frac{MS_S - MS_W}{k_0} = \frac{1}{k_0} \left(\frac{SS_S}{df_S} - \frac{SS_W}{df_W} \right)$$

式中， k_0 为有效平均子女数，如果直接由组内相关法估计遗传力，则

$$h^2 = 4r_1 = \frac{4\hat{\sigma}_S^2}{\hat{\sigma}_S^2 + \hat{\sigma}_W^2} = \frac{4(df_W SS_S - df_S SS_W)}{df_W SS_S + (k_0 - 1)df_S SS_W}$$

该遗传力估计值的显著性检验可用前述同胞资料的检验方法，若利用上述选择反应估计，则因 $\hat{\sigma}_A^2 = 4\hat{\sigma}_S^2$ ，即公畜只决定后代育种值的一半，整理后得到

$$h^2 = \frac{4}{df_S k_0} \left[\frac{\sum_{i=1}^s a_i^2 - \left(\sum_{i=1}^s a_i \right)^2 / \sum_{i=1}^s n_i}{\bar{P}(1-\bar{P})} - \frac{df_S \left(\sum_{i=1}^s a_i - \sum_{i=1}^s \frac{a_i^2}{n_i} \right)}{df_W \bar{P}(1-\bar{P})} \right]$$

此外，上面遗传力显著性检验均可由组内相关的方法检验。相当于 $2 \times s$ 表中的异质性 χ^2 ，即

$$\chi^2 = \left[\sum_{i=1}^s \frac{a_i^2}{n_i} - \left(\sum_{i=1}^s a_i \right)^2 / \sum_{i=1}^s n_i \right] / [\bar{P}(1-\bar{P})]$$

这里 χ^2 的自由度为 $s-1$ 。

此外，注意到依据上述半同胞方差分析估计得到的遗传力是以潜在的连续分布为基础的，它与表型分布遗传力 h_p^2 是直接相关的，Robertson 和 Lerner(1949) 得到两者间的关系如下 $h_p^2 = r_{XP}^2 h_X^2 = Z^2 / [\bar{P}(1-\bar{P})] \cdot h_X^2$ 。式中 h_X^2 为潜在连续分布的遗传力， r_{XP} 为 h_X^2 和 h_p^2 间的相关系数，即 $r_{XP} = Z / \sqrt{\bar{P}(1-\bar{P})}$ ， Z 是当潜在分布上右端面积为 \bar{P} 时的截点处纵高，可由 \bar{P} 值查标准正态分布表得到。由此关系可知，观察到的表型分布的遗传力总是低于其潜在连续分布的遗传力，其比例为 r_{XP} 。此外，对于一个给定潜在连续分布的遗传力，其表型二项分布的遗传力是对称的，最大值为发生率等于 50% 时的遗传力。

例如，调查某鸡群马立克病发生情况，得到 10 只种公鸡的 816 只后代中发病鸡数，试计算其发病力的遗传力。计算前将数据按图 41-7 阴影部分方式整理编辑。

	A	B	C	D	E	F	G
90	种公鸡	后代总数	发病数				
91	1	98	35				
92	2	75	11				
93	3	117	38				
94	4	46	10				
95	5	66	10				
96	6	85	21				
97	7	93	24				
98	8	105	15				
99	9	73	31				
100	10	58	9				
101	阈值性状遗传力分析						
102							
103	计算结果	当前日期 2005-10-30					
104	k0=	81.0014					
105	参数估计法	遗传力	标准误	t 值	p 值		
106	直接估计法	0.1655	0.0968	1.7101	0.0872		
107	选择反应法	0.1645	0.0964	1.7073	0.0878		

图 41-7 阈值性状遗传力估计数据格式及计算结果

执行“专业统计”中的“数量遗传参数估计”下面的“阈值性状遗传力估计”功能后，得到直接估计方法估计的遗传力为 0.1655，间接估计方法估计的遗传力为 0.1645，两者相差不大。经统计检验，其遗传力估计不显著。

但是，假设潜在分布的变量 x 一般是不能度量到的。因而 Falconer(1965)提出用类似估计实现遗传力的方法来估计阈性状遗传力，以及重复率和遗传相关。他用“易感性”来代替“敏感性”表示各种内外环境对阈性状的影响。如果假定阈值在各种情况下保持恒定，而且表型方差在全群和感染个体亲属群(如子代群)中相同，全群的均值(m_p)将低于阈值 x_R 各单位。因此，对感染个体进行选择得到的选择反应(R)可以用其亲属群体均值的改变来度量，即 $R=m_R-m_P=x_P-x_R$ 个单位。感染个体群均值超过群体均值部分为选择差(S)，即 i_p ，这里的单位以表型标准差表示， i 就是选择强度。所以选择反应对选择差的回归系数 b 为： $b=R/S=(x_P-x_R)/i_p$ ，但当感染个体亲属的方差与全群的方差相同的假定不能成立时，由该式估计出的回归系数大约会偏低 5%~10%，为此，Reich 等(1972)提出了如下的修正公式：

$$b = \frac{x_P - x_R \sqrt{1 - (x_P^2 - x_R^2) \left(1 - \frac{x_P}{i_P}\right)}}{i_P + x_R^2 (i_P - x_P)}$$

式中，平方根取正或负要满足于使 b 值在 0~1 取值。由此可得到遗传力估计值为 $h^2=(x_P-x_R)/(r_A i_P)$ 。由该式估计得到的遗传力显著性 t 检验可采用 Falconer 提出的近似抽样

方差得到，即 $V(\hat{h}^2) = \frac{1}{r_A^2} \left\{ \left[\frac{1}{i_P} - b(i_P - x_P) \right]^2 \frac{(1-P_P)}{i_P^2 a_P} + \frac{(1-P_R)}{\sqrt{2} i_R^2 a_P} \right\}$ ，式中 a 表示感染个体数

目，下标 P 和 R 分别表示与全群和亲属有关部门的各个取值。

例如，Mikami 等(1979)对公猪隐睾症的研究结果：在全群中隐睾发生率是 3.9%，而

在隐睾公猪的全同胞间的发生率是 11.6%，根据正态分布可计算其截点值和选择强度。在 DPS 系统中分析时，将数据按图 41-8 阴影部分整理编辑。

	A	B	C	D	E	F
111		总个体数	反应数			
112	全群	1129	44			
113	全同胞	215	25			
114		2x2列联表遗传力分析				
115						
116	计算结果	当前日期	2005-10-30			
117	群体类型	总个体	反应数	反应率	截点值	选择强度
118	全群	1129	44	0.0390	1.7627	2.1648
119	全同胞	215	25	0.1163	1.1938	1.6824
120						
121	估计方法	遗传力	标准误	t值	p-值	
122	Falconer	0.5256	0.1941	2.7075	0.0068	
123	Reich	0.5646	0.1939	2.9121	0.0036	

图 41-8 2×2 列联表遗传力估计数据格式及计算结果

执行“专业统计”中的“数量遗传参数估计”下面的“2×2 列联表遗传力估计”功能后，得到 Falconer 法估计的遗传力为 0.5256，Reich 法估计的遗传力为 0.5646，两者相差不大。经统计检验，其遗传力极显著。

附 选择强度计算函数 ios(x)介绍

选择强度计算，在 DPS 系统里以专用函数 ios(x)形式提供给用户。例如计算发生率 25.8%时的选择强度，只要在 DPS 的电子表格里输入 ios(25.8)，即可得到其选择强度值为 1.252218。

41.3 重复力(率)

41.3.1 数量性状重复力

从统计学角度来说，重复力是同一个体同一性状的不同次生产记录的组内相关系数。重复力的计算公式是

$$r_e = V_B / [V_B + V_W]$$

式中 r_e 为重复力； V_B 是个体间方差； V_W 是个体内度量间方差。在实际计算中，可使用下面公式计算

$$r_e = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$$

式中 MS_B 为个体间均方； MS_W 为个体内度量间均方； k 为各个体的度量次数，如果各个体度量次数不相等，可求加权平均数 k_0 ：

$$k_0 = \frac{\left(\frac{\sum k - \sum k^2}{\sum k} \right)}{df_B}$$

从遗传学角度来说，重复力是表型变量中遗传变量和环境变量所占的百分数。所以，一般环境变量和遗传变量是造成个体间变异的原因，即 $V_B=V_G+V_{Eg}$ ，而特殊环境变量是造成个体内度量间变异的原因，即 $V_W=V_{ES}$ 。公式可写成

$$r_e = \frac{V_B}{(V_B + V_W)} = \frac{(V_G + V_{Eg})}{(V_G + V_{Eg} + V_{ES})} = \frac{(V_G + V_{Eg})}{V_P}$$

测定重复力的方法，如有 n 个家系，每个家系有 k 个成员：

第 1 家系成员 X_{11} X_{12} X_{13} \cdots X_{1k} ；

第 2 家系成员 X_{21} X_{22} X_{23} \cdots X_{2k} ；

Λ Λ Λ Λ Λ

第 n 家系成员 X_{n1} X_{n2} X_{n3} \cdots X_{nk} 。

像方差分析那样，列出方差分析表(表 41-4)。

表 41-4 重复力测定方差分析表

变 异	自由度	均方	期望均方
家系间(个体间)	$N-1$	MS_B	$MS_B = \sigma_{W^2} + k_0 \sigma_{B^2}$
家系内(个体内)	$n(k-1)$	MS_W	$MS_W = \sigma_{W^2}$
总计	$nk-1$		

若各家系内成员相等，则 $k_0=k$ ，不等时利用公式 $k_0 = (\sum k - \sum k^2 / \sum k) / df_B$ 计算。重复力按如下公式计算：

$$r_e = \frac{(MS_B - MS_W)}{[MS_B + (k - 1)MS_W]}$$

重复力参数的作用：

① 可以确定性状应该度量的次数。重复力反映的是性状多次度量值之间的相关程度。重复力高的性状可以根据少数几次度量结果进行选种。重复力低的性状，需要根据多次度量的结果作出判断，较为可靠。不同重复力 r_e 和性状需要度量次数 n 关系大体为 $n \approx (1-r_e) * 10$ 。

② 估计个体最大可能生产力。若某性状重复力为 r_e ， k 次度量均值为 \bar{P}_k ，则最大可能生产力 $MPPA = \bar{P} + b(\bar{P}_k - \bar{P})$ 。式中 \bar{P} 为全群平均数

$$b = \frac{kr_e}{[1 + (k-1)r_e]}$$

③ 综合评定家畜的育种值。如家畜一生共有 k 次记录，则该家畜的育种值的准确度

$$r_A \bar{P}_k = \sqrt{k/[1 + (k-1)r_e]} \cdot h$$

例如，统计 13 头小梅山母猪各胎产仔数，估计母猪窝产仔数的重复力。分析前数据按图 41-9 阴影部分整理编辑。

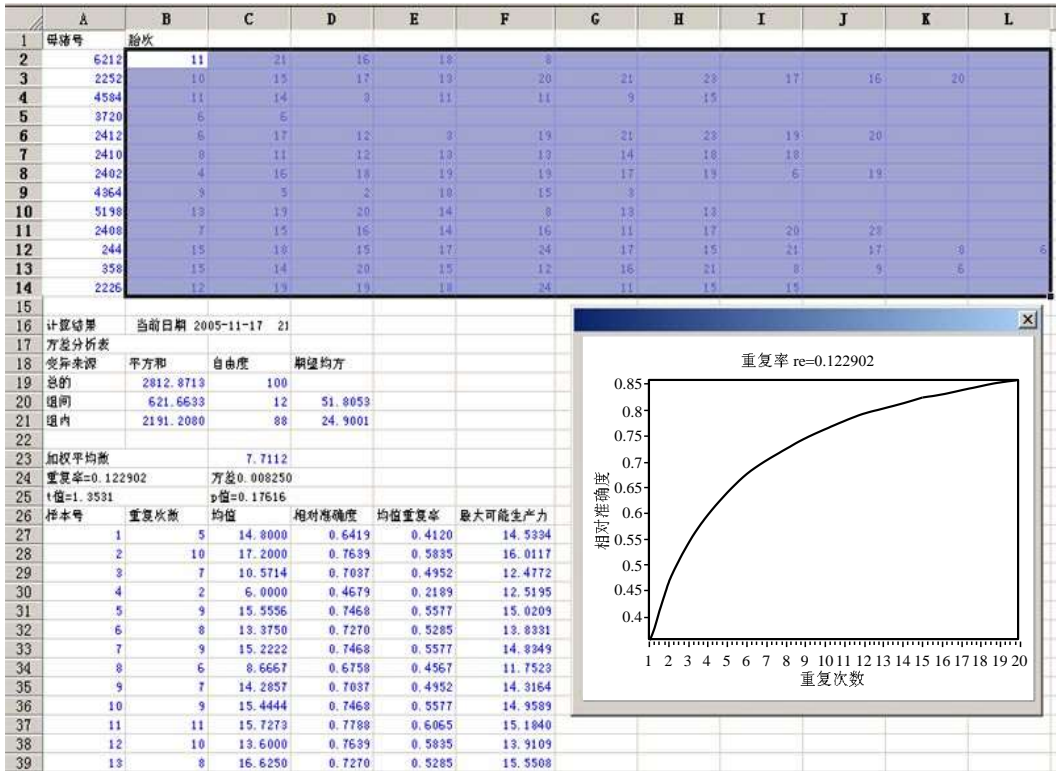


图 41-9 重复力估计数据格式及计算结果

执行“专业统计”中的“数量遗传参数估计”下面的“重复力估计”功能后，系统先给出方差分析表，然后给出加权平均数 $k_0 = 7.711$ ，以及重复力估计值 0.1229，并对重复力进行统计检验。检验结果是重复力不显著。系统最后给出了每个样本的重复次数、均值、相对准确度、均值重复力和最大可能生产力，并在当前的重复力情形下，将重复次数和相对准确度关系以图形方式表示。

41.3.2 阈值性状重复力

1. 回归法

Lush(1956)对母羊产羔数的重复进行了估计，他提出这一重复率(r)可以用将来平均产羔数对头胎或其他胎次产羔数的回归来直接估计。设 y = 将来平均产羔数， x_0 = 头胎

或其他胎次的产羔数, $k =$ 计算 y 的胎次数。因此 y 对 x_0 的回归系数为 $b_{yx_0} = \text{cov}(y, x_0) / V(x_0)$ 。由于第 i 母羊的第 j 胎次产羔数模型为 $X_{ij} = \mu + d_i + e_{ij}$ 。 μ 为全群母羊产羔数平均数, d_i 为 i 母羊的产羔效应, e_{ij} 为随机偏差。由重复率定义知道 $r = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_e^2}$ 。实际的重复率估计方法计算:

法计算:

- ① 设有 n 头母羊, k 年中每年产羔数为 l_1, l_2, \dots, l_k ;
- ② 按头胎的产羔数(即 l_1)为 0, 1, 2 分类统计母羊数为 n_0, n_1, n_2 , 而 $\sum_{i=0}^2 n_i = n$;
- ③ 计算其他 $k-1$ 胎次的各组母羊平均产羔数, 分别记为 y_0, y_1, y_2 ;
- ④ 计算 $\hat{r}_{11} = y_1 - y_0, \hat{r}_{12} = y_2 - y_1, r$ 前一个下标(1)表示按产羔数分类;
- ⑤ 计算加权均数 \hat{r}_1 , 为 $\hat{r}_1 = \frac{\sum_{i=1}^2 w_{1i} \hat{r}_{1i}}{\sum_{i=1}^2 w_{1i}}$, 其权重为 $w_{11} = (k-1)n_0n_1 / (n_0+n_1), w_{12} = (k-1)n_1n_2 / (n_1+n_2)$;

⑥ \hat{r}_1 的抽样方差则为 $\sigma_{\hat{r}_1}^2 = \frac{\sum_{i=1}^m w_{1i} (\hat{r}_{1i} - \hat{r}_1)^2}{(m-1) \sum_{i=1}^m w_{1i}}$, 这里 $m=2$ 。

例如, 收集某羊场资料, 选取其中 10 头母羊的 9 胎产羔记录, 利用头胎和其余 8 胎的记录估计重复力。分析前数据按图 41-10 阴影部分整理编辑。

	A	B	C	D	E	F	G	H	I	J	K
1	样本号	胎次									
2	1	1	2	2	1	0	1	1	1	2	
3	2	1	1	2	1	1	2	2	0	2	
4	3	2	1	2	2	1	1	2	2	1	
5	4	0	1	1	1	2	1	1	2	1	
6	5	1	2	1	2	0	2	1	1	1	
7	6	2	2	1	2	2	2	0	1	2	
8	7	1	1	1	0	2	1	1	2	2	
9	8	0	1	1	2	1	1	0	2	1	
10	9	1	1	2	2	2	1	2	2	1	
11	10	2	1	1	2	1	2	2	2	2	
12											

图 41-10 阈值性状重复力估计数据格式及计算结果

执行“专业统计”中的“数量遗传参数估计”下面的“阈值性状重复力”功能后, 系统给出阈值性状重复力估计值 0.1791, 标准差为 0.0114, 统计检验 t 值为 12.3917, $p < 0.0001$ 。重复力估计值极显著。

2. 列联表法

当观察值只分为两组时, Plackett(1965)提供了一个相对简单的方法。首先将同一些家畜个体的两年记录列成 2×2 列联表(表 41-5), n_{ij} 为每格的观察次数。

表 41-5 重复率估计的 2×2 列联表

第 2 年记录	第 1 年记录	
	0 组	1 组
0 组	n_{00}	n_{01}
1 组	n_{10}	n_{11}

表 41-5 的关联系数 $\psi = n_{00}n_{11}/n_{10}n_{01}$, 则重复率 r 估计值为

$$\hat{r} = -\cos\left(\pi\sqrt{\psi} / (1 + \sqrt{\psi})\right)$$

重复率 r 的标准误为

$$\hat{\sigma}_r = \frac{\pi\sqrt{\psi} \sin\left(\frac{\pi\sqrt{\psi}}{1 + \sqrt{\psi}}\right)}{2(1 + \sqrt{\psi})^2} \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{11}}}$$

例如, 某一奶牛场两年统计母牛产犊情况(顺产和难产), 试根据资料估计奶牛难产的重复力。分析前数据按图 41-11 阴影部分整理编辑。

	A	B	C	D
1	第 2 年	第 1 年		
2		难产 ²	顺产 ²	
3	难产 ²	3	5	
4	顺产 ²	10	28	
5				

图 41-11 2×2 列联表重复力分析数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“ 2×2 列联表重复力分析”功能后, 系统给出阈值性状重复力估计值 0.2012, 标准误为 0.3094, 统计检验 t 值为 0.6503, $p=0.5155$, 重复力估计值没有统计学意义。

41.4 遗传相关

41.4.1 亲子资料遗传相关估计

设 P, O 表示亲子表型值, x, y 表示两性状, A 表示育种值, r_A 为亲子间的亲缘系数, 在随机交配群体中等于 0.5, $r_{(xy)}$ 表示 x, y 性状间的遗传相关。假定亲子间的同一性

状无环境相关，这样，亲子两性状间的关系，根据通径分析原理有

$$\begin{cases} r_{P_x O_y} = h_x r_{(xy)} r_A h_y, & \begin{cases} r_{P_x O_x} = r_A h_x^2 \\ r_{P_y O_y} = r_A h_y^2 \end{cases} \\ r_{P_y O_x} = h_y r_{(xy)} r_A h_x, \end{cases}$$

所以

$$\begin{cases} r_{(xy)} = \frac{r_{P_x O_y}}{h_x r_A h_y} \\ r_{(xy)} = \frac{r_{P_y O_x}}{h_y r_A h_x} \end{cases}$$

从理论上讲，用这两个通径估计的遗传相关应该相等。但实际上，由于亲子间抽样误差等原因，它们并不一定相等。另外，从公式形式看，各式本身不是对称的，而带有回归性质。为了把两式统一起来，一般采用几何平均的方法，即

$$r_{(xy)} = \sqrt{\frac{r_{P_x O_y} r_{P_y O_x}}{h_x r_A h_y h_y r_A h_x}} = \sqrt{\frac{r_{P_x O_y} r_{P_y O_x}}{r_{P_x O_x} r_{P_y O_y}}} = \sqrt{\frac{\text{cov}(P_x, O_y) \text{cov}(P_y, O_x)}{\text{cov}(P_x, O_x) \text{cov}(P_y, O_y)}} = \sqrt{\frac{\text{SP}_{P_x O_y} \text{SP}_{P_y O_x}}{\text{SP}_{P_x O_x} \text{SP}_{P_y O_y}}}$$

因此，遗传相关估计可通过上述 4 个协方差或乘积和的计算得到。其中， $\text{cov}(P_x, O_y)$ ， $\text{cov}(P_y, O_x)$ 是亲子两性状的交叉协方差， $\text{cov}(P_x, O_x)$ ， $\text{cov}(P_y, O_y)$ 是两性状各自亲子间协方差。上式是对称的，符合相关的性质。但是由于涉及的统计量增加，因而也加大了估计误差。此外，式中分子部分的交叉协方差有时可能一正一负，结果出现虚数，为避免这种情况，分子部分可用算术均数替代，即

$$r_{(xy)} = \frac{\text{cov}(P_x, O_y) + \text{cov}(P_y, O_x)}{2\sqrt{\text{cov}(P_x, O_x) \text{cov}(P_y, O_y)}}$$

估计出遗传相关后需要作显著性检验。Reeve(1955)推导了 $r_{(xy)}$ 的抽样方差，其大样本近似抽样方差为

$$V(\hat{r}_{(xy)}) = \frac{1}{n} \left\{ \frac{1}{2}(1 - r_{(xy)}^2) + \frac{2(1 - r_{(xy)}^2)(1 - r_{xy}^2)}{h_x^2 h_y^2} + 4 \left[\frac{r_{(xy)}}{2} \left(\frac{1}{h_x^2} + \frac{1}{h_y^2} \right) - \frac{r_{xy}}{h_x h_y} \right]^2 \right\}$$

这里， n 为亲子对数， r_{xy} 为 x ， y 性状的表型相关。

例如，收集了某奶牛场母女头胎 300 天和 60 天产奶量资料，利用亲子测定法估计两个性状的遗传相关。分析前数据按图 41-12 阴影部分整理编辑。

	A	B	C	D	E	F
3	母60天产奶量	800	800	1000	1000	900
4	女60天产奶量	4000	4300	5000	4800	4700
5	母300天产奶量	900	800	1100	1000	1000
6	女300天产奶量	4000	4300	5300	5000	4800
7	亲子关系估计遗传相关					
8						
9	计算结果	当前日期 2005-11-18 1				
10	遗传相关=0.94781	标准误=0.05667				
11	t值=16.72409	p值=0.04392				
12	亲本表型相关=	0.9288				
13	子女表型相关=	0.8582				
14	合并表型相关=	0.8935				

图 41-12 亲子法遗传相关分析数据格式及计算结果

执行“专业统计”中的“数量遗传参数估计”下面的“亲子关系估测性状间的遗传相关”功能后，系统给出结果(图 41-12 的下部)。其遗传相关估计值 0.9478，标准误为 0.05667，统计检验 t 值为 16.7241， $p=0.0439$ 。遗传相关有显著的统计学意义。输出结果还给出了亲本、子女及合并后的表型相关系数。

41.4.2 同胞资料遗传相关估计

利用同胞资料估计遗传相关，一般采用方差、协方差分析方法。对于大家畜而言，大量的全同胞资料较难得到，为扩大样本含量一般采用半同胞资料估计遗传相关。由于半同胞个体很多，与遗传力估计时一样，需要计算公畜组内两性状方差组分和协方差组分来估计性状的遗传方差及它们间的遗传协方差(表 41-6)。

表 41-6 半同胞资料遗传相关估计方差协方差分析

变异来源	自由度	x 平方和	x 期望均方	y 平方和
公畜间	$S-1$	$\sum_{i=1}^n \frac{x_i^2}{n_i} - \frac{x^2}{n}$	$\sigma_{ex}^2 + k_0 \sigma_{S_x}^2$	$\sum_{i=1}^n \frac{y_i^2}{n_i} - \frac{y^2}{n}$
公畜内	$n-S$	$\sum_{i=1}^n \sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^n \frac{x_i^2}{n_i}$	σ_{ex}^2	$\sum_{i=1}^n \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^n \frac{y_i^2}{n_i}$
总 和	$n-1$	$SS_{S_x} + SS_{ex}$		$SS_{S_y} + SS_{ey}$
y 期望均方		xy 乘积和		xy 期望均积
$\sigma_{ey}^2 + k_0 \sigma_{S_y}^2$		$\sum_{i=1}^n \frac{x_i \cdot y_i}{n_i} - \frac{x \cdot y}{n}$		$cov_e(x, y) + k_0 cov_{S_y}(x, y)$
σ_{ey}^2		$\sum_{i=1}^n \sum_{j=1}^{n_i} x_{ij} y_{ij} - \sum_{i=1}^n \frac{x_i \cdot y_i}{n_i}$		$cov_e(x, y)$
		$SP_{S_{xy}} + SP_{exy}$		

由方差、协方差分析可得到

$$\sigma_{Sx}^2 = \frac{MS_{Sx} - MS_{ex}}{k_0} = \frac{df_e SS_{Sx} - df_S SS_{ex}}{k_0 df_S df_e}$$

$$\sigma_{Sy}^2 = \frac{MS_{Sy} - MS_{ey}}{k_0} = \frac{df_e SS_{Sy} - df_S SS_{ey}}{k_0 df_S df_e}$$

$$\text{cov}_S(x, y) = \frac{MP_{Sxy} - MP_{exy}}{k_0} = \frac{df_e SP_{Sxy} - df_S SP_{exy}}{k_0 df_S df_e}$$

加权均数 $k_0 = \frac{1}{df_S} \left(\sum_{i=1}^S n_i - \sum_{i=1}^S \frac{n_i^2}{n_i} \right) / \sum_{i=1}^S n_i$ 。

在随机交配群体中，如忽略显性效应和各高次项，公畜只决定子女育种值的一半，即 $\sigma_{Sx}^2 = \sigma_{Ax}^2 / 4$ ， $\sigma_{Sy}^2 = \sigma_{Ay}^2 / 4$ ， $\text{cov}_S(x, y) = \text{cov}(A_x, A_y) / 4$ ，

$$r_{(xy)} = \frac{\text{cov}(A_x, A_y)}{\sigma_{Ax} \sigma_{Ay}} = \frac{\text{cov}_S(x, y)}{\sigma_{Sx} \sigma_{Sy}} = \frac{MP_{Sxy} - MP_{exy}}{\sqrt{(MS_{Sx} - MS_{ex})(MS_{Sy} - MS_{ey})}}$$

计算出遗传相关后，需进行显著性检验，一般采用 t 检验，Robertson(1959)得到 $r_{(xy)}$ 抽样方差近似计算公式为

$$V(r_{(xy)}) \approx (1 - r_{(xy)}^2) \sqrt{V(h_x^2) V(h_y^2)} / (2h_x^2 h_y^2)$$

Tallis(1959)得到半同胞估计时 $r_{(xy)}$ 的大样本方差为

$$V(r_{(xy)}) = \frac{1}{df_S k_0^2 t_x} \left\{ (1 + r_{(xy)}^2)(1 + r_S^2) [1 + (k_0 - 1)t_x] [1 + (k_0 - 1)t_y] \right.$$

$$- 2r_{(xy)} r_S \sqrt{t_x t_y [1 + (k_0 - 1)t_x] [1 + (k_0 - 1)t_y]} \times \left(\frac{1}{t_x} + \frac{1}{t_y} + 2k_0 - 2 \right) + \frac{r_{(xy)}^2 (t_x - t_y)^2}{2t_x t_y} \left. \right\}$$

$$+ \frac{1}{df_e k_0^2 t_x t_y} \left\{ (1 + r_{(xy)}^2)(1 + r_e^2)(1 - t_x)(1 - t_y) \right.$$

$$- 2r_{(xy)} r_e \sqrt{t_x t_y (1 - t_x)(1 - t_y)} \times \left(\frac{1}{t_x} + \frac{1}{t_y} - 2 \right) + \frac{r_{(xy)}^2 (t_x - t_y)^2}{2t_x t_y} \left. \right\}$$

$$t_x = \frac{h_x^2}{4}, \quad t_y = \frac{h_y^2}{4}, \quad r_S = \frac{MP_{Sxy}}{\sqrt{MS_{Sx} MS_{Sy}}}, \quad r_e = \frac{MP_{exy}}{\sqrt{MS_{ex} MS_{ey}}}$$

例如，某奶牛场的部分产奶和体重资料，估计奶牛的头胎产奶量(x)和 12 月龄体重(y)的遗传相关。分析前数据按图 41-13 阴影部分整理编辑。

执行“专业统计”中的“数量遗传参数估计”下面的“半同胞关系估测遗传相关”功能后，系统给出结果(图 41-13 的下部)。其遗传相关估计值 0.7467，Robertson 法估计的标准误为 0.5373，统计检验 t 值为 1.3897， $p=0.1849$ ；应用 Tallis 大样本方法估计的标

准误差为 0.6973，统计检验 t 值为 1.0710， $p=0.3011$ 。两种方法检验，其遗传相关均没有显著的统计学意义。

	A	B	C	D	E	F	G	H	I	
23	父本1	性状x	4.9	4.7	5.3	5.3	4.4	5.6	5.4	
24		性状y	3.6	3.7	4	3.2	3.5	4.1	3.2	
25		性状x	3	3.8	4.2	4.5	3.7	3.8		
26	父本2	性状y	3.4	3.3	3.4	3.5	3.7	3.2		
27		性状x	4.7	5	4.2	3.9	5.6			
28	父本3	性状y	3.8	3.3	3.2	3.5	3.8			
29	半同胞资料估计遗传相关									
30										
31	计算结果	当前日期	2005-11-18 1							
32	性状x和y的协方差分析									
33	变异来源	df	SSx	MSx	SSy	MSy	SSxy	MSxy		
34	父本间	2	2.0053	1.0027	0.4364	0.2182	0.6160	0.3080		
35	父本内	15	7.4441	0.4963	1.4236	0.0949	1.8207	0.1214		
36	总 计	17	9.4494		1.8600		2.4367			
37										
38	遗传参数	系数值	标准误	t值	p-值					
39	h ² (x)	0.5880	1.0859	0.5397	0.5974					
40	h ² (y)	0.7175	1.1425	0.6280	0.5304					
41	遗传相关	0.7487	0.5373	1.3897	0.1849 (Robertson近似方差)					
42		0.7487	0.6973	1.0710	0.3011 (Tallix大样本方差)					
43	表型相关	0.5812								

图 41-13 半同胞关系估测遗传相关分析数据格式及计算结果

41.4.3 植物数量性状遗传相关分析

前面介绍的遗传相关，只考虑基因型、表现型两个因素，因此这类相关分析只适用于动物育种。在植物育种工作中，不仅要考虑基因遗传、表现型的影响，而且要考虑环境因素的作用，因此对某两个性状，不仅要分解其表现型和遗传型方差，而且要分解其环境协方差。

表型相关系数可以直接从观测值计算得到，遗传相关系数和环境相关系数即估计遗传协方差和环境协方差，DPS 里面是采用单因素遗传设计方差-协方差方法。该方法方差-协方差分析模式如表 41-7。

表 41-7 单因素遗传设计方差-协方差分析

变异来源	自由度	方差	期望方差	协方差	期望协方差
重复	$R-1$				
品种间	$N-1$	$V_1(x) V_1(y)$	$\sigma^2 + r \sigma_g^2$	W_1	$\text{cov}_e + \text{cov}_g$
机误	$(r-1)(n-1)$	$V_2(x) V_2(y)$	σ^2	W_2	cov_e

从表 41-7 可见，环境协方差由误差协方差估计(W_2)；遗传协方差的估计值为 $(W_1 - W_2)/r$ ；表型协方差为 $[W_1 + (r-1)W_2]/r$ 。从而可求出 3 种相关系数分别为

$$\hat{r}_p = \frac{W_1 + (r-1)W_2}{\sqrt{[V_1(x) + (r-1)V_2(x)][V_1(y) + (r-1)V_2(y)]}}$$

$$\hat{r}_g = \frac{W_1 - W_2}{\sqrt{[V_1(x) - V_2(x)][V_1(y) - V_2(y)]}}$$

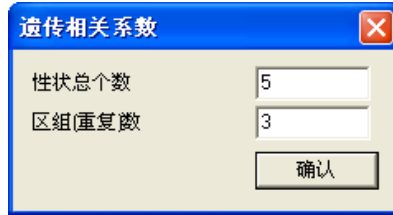
$$\hat{r}_e = \frac{W_2}{\sqrt{V_2(x)V_2(y)}}$$

在 DPS 系统中, 计算植物数量遗传相关系数时数据编辑、定义格式如图 41-14 所示。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	品 种	性状 x1			性状 x2			性状 x3			性状 x4			性状 y			
2		重复 I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	
3	A	10.7	8.2	10.9	41	41.6	41.1	31.9	33.5	32.4	0.4	0.5	0.4	14	11.5	14.5	
4	B	8.9	10.3	9.3	40.9	49.3	51.9	29.5	32	31.3	0.4	0.4	0.4	10.8	16.3	15.1	
5	C	8.7	13.6	10.1	36	42.1	30.6	35.7	32.8	35.1	0.3	0.4	0.3	11.2	18.8	14.4	
6	D	9.1	8.5	8.8	50.3	54.8	53.2	36.3	34.5	36.5	0.3	0.4	0.4	16.5	16.1	17.1	
7	E	11.9	12.7	8.5	48.5	47.7	50.1	36.4	34.8	37.3	0.4	0.4	0.4	21	21.1	15.9	
8	F	10.6	11.5	9.9	43	42.6	48.5	37	35.2	36.3	0.4	0.5	0.4	16.8	17.3	17.4	
9	G	11.8	12.5	12.1	48.2	44	40.8	32.2	33.1	31.4	0.5	0.4	0.4	18.3	18.2	15.5	
10	H	11	10.4	9	50	48.9	49.6	40.5	40.5	39	0.4	0.4	0.4	22.3	20.6	17.4	
11	I	14.7	10.6	12.8	44.7	46.5	44.8	29.8	31	31.9	0.3	0.3	0.3	19.3	15.3	18.3	
12																	

图 41-14 植物数量遗传相关系数分析数据格式

执行“专业统计”中的“数量遗传参数估计”下面的“植物数量遗传相关分析”功能后, 进入用户输入性状个数及其区组(重复)个数的对话框:



输入参数后点击“确认”按钮, 系统给出了包括表型相关系数、遗传相关系数及其标准误的刀切法估计值、环境相关系数和遗传相关系数的通径分析等计算结果如下:

表型相关系数					
性状	x1	x2	x3	x4	x5
x1	1	-0.4374	-0.0851	0.2636	0.3101
x2	-0.4374	1	0.3704	0.3360	0.5378
x3	-0.0851	0.3704	1	-0.0283	0.7843
x4	0.2636	0.3360	-0.0283	1	0.3100
x5	0.3101	0.5378	0.7843	0.3100	1
遗传相关系数					
性状	x1	x2	x3	x4	x5
x1	1	-0.5944	-0.3863	-0.4697	-0.0701
x2	-0.5944	1	0.3762	0.2343	0.5892
x3	-0.3863	0.3762	1	0.3229	0.8283
x4	-0.4697	0.2343	0.3229	1	0.1029
x5	-0.0701	0.5892	0.8283	0.1029	1

遗传相关标准误(Jacknife 估计)					
性状	x1	x2	x3	x4	x5
x1	0	0.1686	0.1903	0.4959	0.1739
x2	0.1686	0	0.1036	0.2255	0.1207
x3	0.1903	0.1036	0	0.1629	0.0963
x4	0.4959	0.2255	0.1629	0	0.2231
x5	0.1739	0.1207	0.0963	0.2231	0
环境相关系数					
性状	x1	x2	x3	x4	x5
x1	1	0.1229	-0.4424	-0.2309	0.8489
x2	0.1229	1	0.0864	0.3031	0.4501
x3	-0.4424	0.0864	1	0.1785	-0.0709
x4	-0.2309	0.3031	0.1785	1	0.0072
x5	0.8489	0.4501	-0.0709	0.0072	1
相关遗传力(对角线为遗传力)					
性状	x1	x2	x3	x4	x5
x1	0.3855	-0.4763	-0.0045	0.3675	-0.0610
x2	-0.4763	0.8369	0.3623	0.2658	0.4364
x3	-0.0045	0.3623	0.9460	-0.0521	0.7935
x4	0.3675	0.2658	-0.0521	0.6705	0.3077
x5	-0.0610	0.4364	0.7935	0.3077	0.6891
遗传系数通径分析					
变量	直接系数	通过 x1	通过 x2	通过 x3	通过 x4
x1	0.6306		-0.3901	-0.3216	0.0111
x2	0.6563	-0.3748		0.3132	-0.0055
x3	0.8326	-0.2436	0.2469		-0.0076
x4	-0.0236	-0.2962	0.1538	0.2688	

41.5 选择指数

41.5.1 综合选择指数

育种选择往往是多性状的。这就需要尽量准确地估计出个体的这些多性状的一个综合育种值，并依据它进行选种，以获得最佳效果。由于不同性状在育种上和经济上的重要性差异，因而，实际的育种目标对各性状的选育提高要求不一致的，即要对性状进行加权。

设需要选择提高的性状共有 n 个，每个性状育种值为 a_1, a_2, \dots, a_n ，这些性状也称为目标性状，相应的经济加权值为 w_1, w_2, \dots, w_n ，故综合育种值可定义如下：

$$H = \sum w_i a_i = w'a$$

其中， $w'=[w_1, w_2, \dots, w_n]$ ， $a=[a_1, a_2, \dots, a_n]$ 。

为估计个体的综合育种值 H ，经典的选择指数是依据个体的 n 个表现型性状，建立一个信息性状的判别函数

$$H = \sum b_i x_i = b'x'$$

这里

$$b'=[b_1, b_2, \dots, b_n], \quad x'=[x_1, x_2, \dots, x_n]$$

b_i 是性状 x_i 的加权系数。显然，我们的目的是要获得一个指数 I ，用它估计 H 可获得最大的综合育种值进展 ΔH ：

$$\Delta H = R_H = i r_{HI} \sigma_H$$

而

$$r_{HI} = \text{cov}(H, I) / (\sigma_I \sigma_H)$$

对于由不同的 b 确定的不同选择指数而言， i 和 σ_H 都是相等的。故 ΔH 最大化和 r_{HI} 的最大化是一致的。 r_{HI} 取极大时的 b 值，由于

$$\text{cov}(I, H) = \text{cov}[b'x, (w'a)'] = b' \text{cov}(x, a')w$$

而表现型向量可分解为 $x = a + e$ 。这里， e 为环境偏差。假定各性状育种值与环境偏差不存在相关，并令

$$A = \text{cov}(a, a') = V(a)$$

它表示性状育种值之间的方差、协方差矩阵，由此可推导得到

$$\text{cov}(I, H) = b'Aw, \quad V(H) = w'Aw, \quad V(I) = b'Pb,$$

式中， $P = V(x)$ ，它表示性状表型值之间的方差、协方差矩阵，所以

$$r_{HI} = \frac{\text{cov}(I, H)}{\sigma_H \sigma_I} = \frac{b'Aw}{\sqrt{w'Aw} \sqrt{b'Pb}}$$

求 b 使得 r_{HI} 极大，根据极值原理可以得到 $b = P^{-1}Aw$ 。得到 b 值后，根据每一性状表型值来计算个体指数 I ，并以该值的大小确定个体是否选留。除 I 外，衡量选择效果的指标还有综合育种值估计准确度 r_{HI} 、综合育种值选择进展 ΔH 、各性状育种值选择进展 Δa ，以及指数遗传力 $h^2 I$ 等。

例如，由某猪场资料分析得到 3 个性状：饲料利用率(肥育期间总消耗量 kg/肥育期总增重 kg) x_1 、平均日增重 x_2 (g)和背膘厚度 x_3 (cm)的表型值、性状的边际效益、遗传力、表型方差系数，以及表型相关、遗传相关系数。在 DPS 系统中，分析数据按图 41-15 阴影部分整理编辑(注意阴影右边 3 项的对角线右上部分是表型相关，对角线左下部分是遗传相关系数)。

执行“专业统计”中的“数量遗传参数估计”下面的“综合选择指数”功能后，系统会提示输入要给定的参数($x_1=2.85$, $x_2=670$, $x_3=2.8$)，最后给出分析结果(图 41-15 下部)。

从分析结果可以看出，这 3 个性状的综合选择指数为

$$I = -5.8824x_1 + 0.0326x_2 - 17.5925x_3$$

系统还给出了综合选择指数的估计准确度 $r_{HI}=0.6352$ 及其他结果。最后给出了给定参数($x_1=2.85$, $x_2=670$, $x_3=2.8$)时的选择指数 $I=6.68$ 。

	A	B	C	D	E	F	G	H
1	性状	均值	权重w	遗传力H ²	方差	x1	x2	x3
2	饲料利用率x1	3	-40	0.3	0.05		-0.65	0.45
3	平均日增重x2	600	0.05	0.35	5000	-0.75		-0.35
4	背 膘 厚x3	3	-30	0.5	0.05	0.4	-0.4	
5	综合选择指数计算数据编辑界面							
6								
7	计算结果	23:00:04						
8	综合选择指数							
9	I=-5.8824x1+0.0326x2-17.5925x3							
10	综合选择指数估计的准确度=0.6352							
11	综合育种值选择进展ΔH=6.1184							
12	各个性状育种值选择进展							
13	Δ a1=-0.0571		Δ a2=20.6127		Δ a3=-0.0934			
14	指数遗传力=0.4332							
15	性状x1=2.85		性状x2=670.00		性状x3=2.80		预测值=6.6796	

图 41-15 综合选择指数计算数据格式及计算结果

41.5.2 约束选择指数

综合选择指数是以所考虑性状的综合育种值进展最大为目的，其各个性状的进展不受人为控制。但实际上，人们需要对不同性状的改进作适当的控制——在一些性状改进的同时保持另一些性状不变，即所谓的约束选择指数。约束选择的目的是在给定的约束条件下使得确定的选择指数尽可能准确地估计综合育种值，即使得估计方差 $Q=V(I-H)$ 最小，这时

$$Q=V(I-H)=V(I)+V(H)-2cov(I, H)=b'Pb+w'Aw-2b'Aw$$

为了对某些性状的遗传进展施加一定的约束，可引入如下约束矩阵 R 。 R 中每一列向量对应于一个约束性状。在该列向量中，对应于约束性状的元素取 1，其余元素取 0，从而使 $R'a$ 只含右约束性状的育种值向量。由公式

$$\Delta a' = i \cdot b'A / \sqrt{b'Aw}$$

得到

$$R'\Delta a = iR'Ab/\sigma_I = \alpha k$$

这里的 $R'\Delta a$ 显然是所需要的约束性状的与其遗传进展，若 $\alpha = i/\sigma_I$ ，可得到使约束性状按比例向量 $k'=[k_1, k_2, \dots, k_n]$ 变化的限制条件，即

$$R'Ab - k = 0$$

这样约束选择指数模型是求 b ，并使得

$$\begin{cases} \min & Q = b'Pb + w'Aw - 2b'Aw \\ \text{s.t.} & R'Ab - k = 0 \end{cases}$$

解此模型，可得到

$$\begin{bmatrix} b \\ \lambda \end{bmatrix} = \begin{bmatrix} P & AR \\ R'A & 0 \end{bmatrix}^{-1} \begin{bmatrix} Aw \\ k \end{bmatrix}$$

应用分块求逆方法解出其逆矩阵可得到

$$b = [E - P^{-1}AR(R'AP^{-1}AR)^{-1}R'A]P^{-1}Aw + P^{-1}AR(R'AP^{-1}AR)^{-1}k$$

式中 E 为单位矩阵。显然当 $k=0$ 时，即保持所有的约束性状不变，则有

$$b = [E - P^{-1}AR(R'AP^{-1}AR)^{-1}R'A]P^{-1}Aw$$

得到 b 值后，根据每一性状表型值来计算个体指数 I ，并以该值的大小确定个体是否选留。除 I 外，衡量选择效果的指标还有综合育种值估计准确度 r_{HI} 、综合育种值选择进展 ΔH 、各性状育种值选择进展 Δa ，以及指数遗传力 h_I^2 等。

这里仍以前面例子来说明约束选择指数的制定方法。假设经过选择已使平均背膘厚度下降到 1.5cm， $\bar{x}_3=1.5$ 。为保持肉质不发生大的改变，在以后的选择中希望保持背膘厚度不再下降，试计算其约束选择指数并计算当 $x_1=2.85$ ， $x_2=670$ ， $x_3=1.3$ 时的约束选择指数值。

在 DPS 系统中，分析数据按图 41-16 阴影部分整理编辑。要注意的是，数据块的第二列放“是否约束”的选择，这里为 1 时表示对该因子进行约束。另外，阴影右边 3 项的对角线右上部分是表型相关，对角线左下部分是遗传相关系数。

执行“专业统计”中的“数量遗传参数估计”下面的“约束选择指数”功能后，系统会提示输入要给定的参数($x_1=2.85$ ， $x_2=670$ ， $x_3=1.2$)，最后给出分析结果(图 41-16 下部)。从分析结果可以看出，这 3 个性状的约束选择指数 I 为

$$I = -11.47x_1 + 0.02x_2 + 6.08x_3$$

	A	B	C	D	E	F	G	H	I
8	性状	均值	是否约束	权重w	遗传力H2	方差	x1	x2	x3
9	饲料利用率x1	3	0	-40	0.3	0.05		-0.65	0.45
10	平均日增重x2	600	0	0.05	0.35	5000	-0.75		-0.35
11	背 膘 厚x3	1.5	1	-30	0.5	0.05	0.4	-0.4	
12	约束选择指数计算数据编辑界面								
13									
14	计算结果	当前日期 2005-11-22							
15	约束选择指数								
16	I=-11.4667x1+0.0239x2+6.0784x3								
17	约束选择指数估计的准确度=0.3619								
18	约束育种值选择进展ΔH=3.4860								
19	各个性状育种值选择进展								
20	Δ a1=-0.0621	Δ a2=20.0067		Δ a3=0.0000					
21	指数遗传力=0.3414								
22	性状x1=2.85	性状x2=670.00		性状x3=1.30		预测值=2.1748			

图 41-16 约束选择指数计算数据格式及计算结果

系统还给出了约束选择指数的估计准确度 $r_{HI}=0.3619$ 及其他结果。最后给出了给定参数($x_1=2.85$ ， $x_2=670$ ， $x_3=1.3$)时的选择指数 $I=2.17$ 。

比较约束与非约束选择指数，其变化是相当大的。特别是对于背膘厚度而言，其偏回归系数由无约束时的-17.59 变为约束时的 6.08。而且约束后的选择效果大大降低，其

牧场 I	36	176999	0	0	5	18637	11	47637	2	7373
牧场 II	32	145031	2	7474	3	11189	5	21192	3	9878
牧场 III	32	162463	2	8002	1	4180	27	123052	8	41044
牧场 IV	1	4678	21	91216	4	9369	3	9688	4	11813
牧场 V	1	4938	16	68115	2	8050	1	4338	0	0

在 DPS 系统中进行 BLUP 育种值预测，其数据编辑格式如图 41-17 阴影部分所示。数据的第一行的数字是牧场的个数；下面的下三角阵表示各个公畜间亲缘关系矩阵 A 的下三角部分。

在 DPS 系统中执行“专业统计”中的“数量遗传参数估计”下面的“两因子(无互作)BLUP 估计”功能后，系统会提示输入遗传力参数(本例中输入 0.3)，最后给出分析结果(图 41-17 下部)。

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	36	176999	0	0	5	18637	11	47637	2	7373		
3	32	145031	2	7474	3	11189	5	21192	3	9878		
4	32	162463	2	8002	1	4180	27	123052	8	41044		
5	1	4678	21	91216	4	9369	3	9688	4	11813		
6	1	4938	16	68115	2	8050	1	4338	0	0		
7												
8												
9												
10												
11												
12												
13	计算结果	当前日期 2006-11-21										
14	求解一般线性方程											
15	222.0000	84.0000	45.0000	70.0000	35.0000	26.0000	102.0000	41.0000	16.0000	47.0000	17.0000	39239.0000
16	84.0000	84.0000	0.0000	0.0000	0.0000	6.0000	34.0000	0.0000	6.0000	11.0000	2.0000	25944.0000
17	45.0000	6.0000	45.0000	0.0000	0.0000	6.0000	31.0000	2.0000	3.0000	6.0000	3.0000	19474.0000
18	70.0000	6.0000	3.0000	70.0000	0.0000	6.0000	32.0000	2.0000	1.0000	27.0000	8.0000	11874.0000
19	35.0000	6.0000	3.0000	0.0000	35.0000	6.0000	1.0000	21.0000	4.0000	2.0000	4.0000	12474.0000
20	20.0000	6.0000	3.0000	0.0000	0.0000	26.0000	1.0000	16.0000	2.0000	3.0000	6.0000	9544.0000
21	102.0000	16.0000	32.0000	32.0000	1.0000	1.0000	114.3333	0.0000	0.0000	6.0000	6.0000	49410.0000
22	41.0000	6.0000	2.0000	2.0000	21.0000	16.0000	0.0000	63.3333	6.0000	6.0000	6.0000	17497.0000
23	16.0000	6.0000	3.0000	1.0000	4.0000	2.0000	0.0000	21.3333	6.0000	6.0000	6.0000	51426.0000
24	47.0000	11.0000	5.0000	27.0000	3.0000	1.0000	0.0000	0.0000	0.0000	19.3333	6.0000	30897.0000
25	11.0000	2.0000	3.0000	8.0000	4.0000	6.0000	0.0000	0.0000	0.0000	6.0000	29.3333	78109.0000
26												
27	地点1年, 第11环效应			223.3608								
28	地点1年, 第12环效应			-136.1489								
29	地点1年, 第13环效应			692.9123								
30	地点1年, 第14环效应			-668.9726								
31	地点1年, 第15环效应			-113.1936								
32												
33	公畜1效应	381.3690	育种值	782.7990								
34	公畜2效应	238.8696	育种值	677.6192								
35	公畜3效应	-481.6009	育种值	-903.5019								
36	公畜4效应	-181.6875	育种值	-303.1945								
37	公畜5效应	-127.8104	育种值	-269.0209								

图 41-17 两因子 BLUP 育种值预测数据编辑格式及分析结果

分析结果中首先给出了正规方程。然后给出了 5 个牧场的环境效应值，分别是

从育种值 BLUP 预测结果可以看出，5 头种公牛以 351 号为最佳，育种值为 782.798、75-204 号次之，它们均有提高牛群产奶量的改进作用。而 1312 号、593 号及 76-309 号公牛的女儿产奶量预测值均在平均数以下，相对而言是比较差的。

41.7 NC I 设计(两因素巢式设计)

NC I 设计属于巢式设计(nested design), 其主要特点是以两个纯合亲本杂交所产生的 F_2 或者以随机交配植物的高世代作为基础群体, 从中随机挑选出一批个体作为父本, 再随机挑选若干个体作为每一父本相应的母本, 因此有多少个父本就有多少组相应的母本, 但每一母本只限交一次, 所以如有 m 个父本, 每一父本同 f 个母本杂交则计有 mf 个母本, 得 mf 个家系, 如每个家系内有 r 次重复, 每个重复内考察 n 个个体, NCI 设计的统计模型为

$$x_{ijkl} = u + b_i + r_{ip} + m_{ij} + f_{ijk} + e_{ijkl}$$

式中, u 表示总的平均数; b_i 表示第 i 个区组效应; r_{ip} 表示区组中第 i 个区组第 p 个重复的效应; m_{ij} 表示第 i 个区组第 j 个父本的效应; f_{ijk} 表示第 i 个区组内第 j 个父本内的第 k 个母本的效应; e_{ijkl} 表示每个观察值的误差。

如果是完全随机设计, 不考虑区组效应, 则其方差分析表简单一些(表 41-9)。

表 41-9 NC I 完全随机设计方差分析

变异来源	自由度	均方	期望均方
处理	$mf-1$		
父本间	$m-1$	M_1	$\sigma_w^2 + r\sigma_{D/S}^2 + rf\sigma_s^2$
母本间/父本内	$m(f-1)$	M_2	$\sigma_w^2 + r\sigma_{D/S}^2$
试验误差	$(r-1)(mf-1)$	M_3	σ_w^2

其父本间的方差 $\sigma_m^2 = (M_1 - M_2) / (rf)$, 父本内母本间的方差 $\sigma_{f/m}^2 = (M_2 - M_3) / r$ 。根据 σ_m^2 和 $\sigma_{f/m}^2$ 可估计出加性方差 V_d 和显性方差 V_h :

$$V_d = 4\sigma_m^2, \quad V_h = 4(\sigma_{f/m}^2 - \sigma_m^2)$$

DPS 系统处理 NCI 设计, 其数据格式如图 41-18。图 41-18 是 2 个区组($r=2$), 3 个父本($m=3$), 2 个母本($f=2$), 在每个家系中取 3 个后代($n=3$)的 NCI 设计试验结果。分析前按图 41-18 中阴影部分编辑、定义数据。

	A	B	C	D	E	F	G	H	I
1			I			II			
2	父本	母本	1	2	3	1	2	3	
3	m1	f1	2.6	2.5	2.3	2.5	2.1	2	
4	m1	f2	3.4	3.8	4	2.1	2.7	3	
5	m2	f3	2.7	3	3.4	2.5	2	2.4	

6	m2	f4	3	3.5	3.7	2.5	2.9	3
7	m3	f5	3.5	3.7	4	3.5	3.7	4.1
8	m3	f6	4	4.6	4.8	3.4	4	4.2
9								

图 41-18 NCI 设计数据编辑定义格式

在 DPS 系统支持下，执行“专业统计”中的“遗传交配设计统计”下面的“NCI 设计(巢式设计)”功能后，系统给出要求用户输入父本数、母本数及重复区组数的用户界面。用户输入正确的参数后，系统立即给出分析结果如下：

NCI 方差分析表(随机区组设计)					
变 异 来 源	平方和	自由度	均 方	F 值	p 值
区 组 间	2.7225	1	2.7225		
雄 性 品 系	10.5339	2	5.2669	4.9884	0.1112
雄性内雌性品系	3.1675	3	1.0558	4.4024	0.0721
重复 × 组合	1.1992	5	0.2398		
误 差	2.5867	24	0.1078		
总 变 异	20.2097	35			
σ_m 方差=	0.3509				
σ_f/m 方差=	0.1360				
当亲本的近交系数 F=0 时					
加性方差=	1.4037				
显性方差=	-0.8597				
组内与组间方差之和	0.7268				
半同胞系数=0.4829 Hn 方差=1.9315					
全同胞系数=0.6700					
狭义遗传力估计值=1.9315±1.3718					
NCI 方差分析表(完全随机设计)					
变 异 来 源	平方和	自由度	均 方	F 值	p 值
雄 性 品 系	10.5339	2	5.2669	4.9884	0.1112
雄性内雌性品系	3.1675	3	1.0558	4.8668	0.0071
误 差	6.5083	30	0.2169		
总 变 异	20.2097	35			
σ_m 方差=	0.3509				
σ_f/m 方差=	0.1398				
当亲本的近交系数 F=0 时					
加性 方差=	1.4037				
显性 方差=	-0.8444				
组内与组间方差之和	0.7077				
半同胞系数=0.4959 Hn 方差=1.9835					

全同胞系数=0.6934
狭义遗传力估计值=1.9835±1.3490

分析结果中, DPS 分别执行了完全随机设计和随机区组设计两种试验设计情形, 给出了各自的方差分析表、父本间方差(σ_m)、父本内母本间方差(σ_{fm}), 以及当亲本的近交系数 $F=0$ 时的加性方差、显性方差、组内与组间方差之和, 以及半同胞系数及其方差、全同胞系数和狭义遗传力估计值。

41.8 NCII 设计(不完全双列杂交设计)

NCII 设计包括 n_1 个父本、 n_2 个母本, 将这 n_1+n_2 个亲本进行 $n_1 \times n_2$ 次杂交, 共得 $n_1 n_2$ 个家系, 组成试验设计方案, 因此属于不完全双列杂交设计。该设计经常在作物育种中应用。在试验中, 完全双列杂交在方法和理论上是较完整的, 可由它获得较多、较完整的遗传信息。但是, 不难发现, 随着亲本的增加, 其杂交组合随之急剧增加, 这使得一方面由于组合过多, 难于控制试验误差, 从而影响试验的准确性和精确性; 另一方面, 组合过多, 给实际杂交工作和后代的处理工作增加了许多困难。事实上, 依据育种学理论, 可以预见一些组合可能无多大潜在价值而没有必要去处理它们, 同时, 亲本数过少而造成抽样误差较大, 可能使得配合力及其他遗传参数的估计失去准确性。为了协调二者之间的关系, 一些遗传、育种工作者设计了改良双列杂交, 如不完全双列杂交和部分双列杂交设计等。在这样的设计中, 仅仅包含了所有可能杂交组合的一部分而不是全部。不完全双列杂交设计比部分双列杂交更适合于某种要求, 所以, 本系统提供了不完全双列杂交分析功能。

不完全双列杂交是指把供试纯育亲本按试验要求分为两组, 只进行组间的各种可能杂交。它可以大大减少组合数, 例如, 当亲本数为 8 时, 即使按完全双列杂交设计试验方法 4, 供试组合仍有 28 个, 如按不完全双列杂交设计, 将亲本分为两级的可能分法有 1×7 、 2×6 、 3×5 、 4×4 , 最高组合为 16, 比前者的组合数要少近一半。可见, 在相同试验规模上, 采用不完全双列杂交设计比完全双列杂交设计多处理较多的亲本, 并能降低抽样误差和扩大试验范围, 以更准确估计群体的遗传参数。

按随机区组方式进行 NCII 设计, 区组(重复)为 b , 每个小区只取一个观察值, 计算步骤为:

- (1) 列出方差分析表, 测定组合的差异显著性(表 41-10)。

表 41-10 各个组合理论方差

变异来源	自由度	方差	理论方差组成成分
区组	$b-1$	V_b	
组合	$n_1 n_2 - 1$	V_c	
P_1 (或 g_i)	$n_1 - 1$	V_{P_1} (或 V_{g_i})	$\sigma_c^2 + b\sigma_{12}^2 + bn_2\sigma_1^2$
P_2 (或 g_j)	$n_2 - 1$	V_{P_2} (或 V_{g_j})	$\sigma_c^2 + b\sigma_{12}^2 + bn_1\sigma_2^2$

$P_1 \times P_2$ (或 s_{ij})	$(n_1-1)(n_2-1)$	$V_{P_1 \times P_2}$ (或 V_{sij})	$\sigma_e^2 + b\sigma_{12}^2$
机误(即区组×组合互作)	$(b-1)(n_1n_2-1)$	$V_{bc} = V_e$	σ_e^2
总数	bn_1n_2-1		

组合间的方差是由两套亲本品种的一般配合力和相互间的特殊配合力组成，所以需将组合间的方差区分为亲本间一般配合力和特殊配合力方差。在计算时，列出各个组合的二向表(表 41-11)：

表 41-11 亲本组合均数二向表

	P_{11}	P_{12}	...	P_{1n_1}	P_1 (或 $X_{1.}$)
P_{21}	x_{11}	x_{12}	...	x_{1n_1}	$X_{1.}$
P_{22}	x_{21}	x_{22}	...	x_{2n_1}	$X_{2.}$
\mathbf{N}	\mathbf{N}	\mathbf{N}		\mathbf{N}	\mathbf{N}
P_{2n_2}	x_{n_21}	x_{n_22}	...	$x_{n_2n_1}$	$X_{n_1.}$
P_2 (或 $X_{.j}$)	$X_{.1}$	$X_{.2}$...	$X_{.n_2}$	$X_{..}$

根据二向表，进行方差分解，得到方差分析表 41-10。

(2) 根据表 41-10 理论方差的组成模式，估算配合力方差。各个配合力方差估计公式

$$\sigma_{gi}^2(\sigma_{P1}^2) = \frac{V_{gi} - V_{sij}}{bn_2}, \quad \sigma_{gj}^2(\sigma_{P2}^2) = \frac{V_{gj} - V_{sij}}{bn_1}, \quad \sigma_{sij}^2(\sigma_{P1 \times P2}^2) = \frac{V_{sij} - V_e}{b}$$

一般配合力方差(%)

$$V_g = \frac{\sigma_{gi}^2 + \sigma_{gj}^2}{\sigma_{gi}^2 + \sigma_{gj}^2 + \sigma_{sij}^2} \times 100$$

特殊配合力方差

$$V_s = \frac{\sigma_{sij}^2}{\sigma_{gi}^2 + \sigma_{gj}^2 + \sigma_{sij}^2} \times 100$$

(3) 根据各个组合的平均值，计算各个亲本一般配合力和各组合特殊配合力相对效应，即以全试验的总平均数为基础，计算各个组合相对平均值，即从求得的百分数中减去 100。然后，根据各个组合的相对平均值，计算一般配合力的相对效应和特殊配合力的相对效应。

在 DPS 系统平台，数据整理格式为：设 5 个父本、4 个母本，设 3 个区组，每个小区 1 个观察值，其数据在电子表格中的输入整理格式如图 41-19。

	A	B	C	D	E	F
--	---	---	---	---	---	---

1	父本 A	母本 B	区组 I	区组 II	区组 III
2	A1	B1	31	29	24
3	A1	B2	27	25	26
4	A1	B3	18	18	10
5	A1	B4	14	16	14
6	A2	B1	19	21	21
7	A2	B2	21	17	27
8	A2	B3	14	13	12
9	A2	B4	16	9	14
10	A3	B1	20	14	16
11	A3	B2	16	18	12
12	A3	B3	16	11	14
13	A3	B4	12	9	10
14	A4	B1	26	35	36
15	A4	B2	30	31	34
16	A4	B3	20	18	17
17	A4	B4	26	24	21
18	A5	B1	12	8	11
19	A5	B2	10	10	15
20	A5	B3	7	13	10
21	A5	B4	11	11	8
22					

图 41-19 NCII(不完全双列杂交)数据编辑格式

将数据按图 41-19 格式输入，并将待分析的数据定义成数据块，再进入菜单，执行“专业统计”中的“遗传交配设计统计”下面的“NCII 设计”功能后，系统给出要求用户输入父本数、母本数及重复区组数的用户界面。用户输入正确的参数后，系统立即给出分析结果如下：

各个组合均值列表					
均值	p2, 1	p2, 2	p2, 3	p2, 4	p1.
p1, 1	28.00000	26.00000	15.33333	14.66667	21.00000
p1, 2	20.33333	21.66667	13.00000	13.00000	17.00000
p1, 3	16.66667	15.33333	13.66667	10.33333	14.00000
p1, 4	32.33333	31.66667	18.33333	23.66667	26.50000
p1, 5	10.33333	11.66667	10.00000	10.00000	10.50000
p2.	21.53333	21.26667	14.06667	14.33333	17.80000
方差分析表					
变异来源	自由度	平方和	均方	F 值	显著水平
区组	2	7.600000	3.800000	0.427556	0.65519
组合	19	2964.267	156.0140	17.55389	0.000000
p1	4	1851.600	462.9000	16.63114	0.000077
p2	3	778.6667	259.5556	9.325349	0.001849
P1xP2	12	334.0000	27.83333	3.131662	0.003583
误差	38	337.7333	8.887719		
合计	59	3309.600			

从方差分析表结果可以看出, 各个组合间的 F 值=17.5539, 差异极显著($p=0.0000$); 一般配合力(P_1 , 亦即 g_i)间的 F 值=16.63114, 差异极显著($p=0.0001$); 特殊配合力($P_1 \times P_2$, 亦即 s_{ij})间的 F 值=3.131662, 差异也是极显著($p=0.0036$);

配合力方差估计

P_1 一般配合力方差基因型方差=36.2556

P_2 一般配合力方差基因型方差=15.4481

$P_1 \times P_2$ 特殊配合力基因型方差=6.3152

环境方差=8.8877

一般配合力方差=89.12%, 特殊配合力方差=10.88%

广义遗传力=86.72% , 狭义遗传力=77.28%

各个亲本一般配合力和各个组合特殊配合力相对效应

各个组合的相对平均值

	p2, 1	p2, 2	p2, 3	p2, 4	p1.
p1, 1	57.30337	46.06742	-13.85768	-17.60300	17.97753
p1, 2	14.23221	21.72285	-26.96629	-26.96629	-4.49438
p1, 3	-6.36704	-13.85760	-23.22097	-41.94757	-21.34831
p1, 4	81.64794	77.90262	2.99626	32.95880	48.87640
p1, 5	-41.94750	-34.45690	-43.82022	43.82022	-41.01124
p2.	20.97378	19.47566	-20.97378	-19.47566	

从各个组合的相对平均值可以看出, P_1 亲本中的 p_{11} 、 p_{14} 的一般配合力的相对效应分别是 17.98 和 48.88, P_2 亲本中的 p_{21} 、 p_{22} 的一般配合力的相对效应分别是 20.90 和 19.55, 这些亲本的一般配合力较优。特别是 p_{14} 的一般配合力的相对效应最大。

各个组合一般配合力和特殊配合力的相对平均值

	p2, 1	p2, 2	p2, 3	p2, 4	p1.
p1, 1	18.35206	8.61423	-10.86142	-16.10487	17.97753
p1, 2	-2.24719	6.74157	-1.49813	-2.99626	-4.49438
p1, 3	-5.99251	-11.98502	19.10112	-1.12360	-21.3483
p1, 4	11.79775	9.55056	-24.90637	3.55805	48.87640
p1, 5	-21.91011	-12.92135	18.16479	16.66667	-41.0112
p2.	20.97378	19.47566	-20.97378	-19.4756	

从各个组合一般配合力和特殊配合力的相对平均值可以看出, P_2 亲本中的 p_{23} 、 p_{24} (即亲本 2 中的第 3 个和第 4 个亲本)、 P_1 亲本中的 p_{15} (即亲本 1 中的第 5 个亲本) 的一般配合力不大, 但它们的特殊配合力的表现还可以。

41.9 NCIII(回交系统)设计

NCIII 设计又称回交系统类型的遗传设计, 从两个纯系杂交后代的 F_2 中随机抽取部分植株作为亲本, 与原来的两个纯系亲本回交, 可以形成两个回交系统。如果从 F_2 抽取 n 个单株, 回交于 P_1 和 P_2 , 一共有 $2n$ 个组合, 如每个组合重复 r 次, 一共有 $2(n+1)r$ 个试验小区, NCIII 设计的统计模型为

$$x_{ijkl} = u + b_i + r_{ip} + m_{ij} + f_{ii} + (m_f)i_{kl} + e_{ijkl}$$

式中, u 表示总的平均数; b_i 表示第 i 个区组效应; r_{ip} 表示第 i 个区组中第 p 个重复的效应; m_{ij} 表示第 i 个区组第 j 个父本的效应; f_{ii} 表示第 i 个区组内自交系效应; e_{ijkl} 表示每个观察值的误差。其方差分析如表 41-12。

表 41-12 NCIII 设计的方差分析

变异原因	Df	MS	EMS
重复	$R-1$		
自交系间(亲本)	1	M_1	
雄性 F_2 家系	$n-1$	M_2	$\sigma^2 + 2r\sigma_n^2$
F_2 家系×自交系(亲本)	$n-1$	M_3	$\sigma^2 + r\sigma_{nl}^2$
误差	$(r-1)(2n-1)$	M_4	σ^2

表 41-12 中, σ_n^2 是 F_2 家系(亲本)加性遗传方差, σ_{nl}^2 为 F_2 家系与自交系互作的遗传方差(分量), 即显性遗传方差, 估计方法如下:

$$\sigma_n^2 = \frac{(M_2 - M_4)}{2r}, \quad \sigma_{nl}^2 = \frac{(M_3 - M_4)}{r}$$

方差分量估计的相应遗传成分为 $\sigma_n^2 = \frac{1}{8}D$, $\sigma_{nl}^2 = \frac{1}{4}H$ 。

DPS 系统处理 NCIII 设计, 其数据格式如图 41-20。图 41-20 是 2 个区组($r=2$), 4 个父本($m=4$), 2 个自交系($f=2$)的 NCIII 设计试验结果。分析前按图 41-20 中阴影部分编辑、定义数据。

	A	B	C	D	E	F	G
1			I		II		
2	父本	自交系	1	2	1	2	
3	m1	f1	4.2	5	3.8	4.6	
4	m1	f1	3.8	3	6	2.5	
5	m2	f2	5.7	3.7	5	1.8	
6	m2	f2	3.7	4	3.5	5	
7	m3	f3	5.3	3.3	4.3	4.2	
8	m3	f3	3.3	5	2.8	3.4	
9	m4	f4	2.2	4	1	3.6	
10	m4	f4	5.8	2.5	1.8	3.2	
11							

图 41-20 NC III 设计试验数据编辑格式

将数据按图 41-20 格式输入, 并将待分析的数据定义成数据块, 再进入菜单, 执行

“专业统计”中的“遗传交配设计统计”下面的“NCIII 设计”功能后，系统给出要求用户输入父本数、自交系数及重复区组数的用户界面。用户输入正确的参数后，系统立即给出分析结果如下：

计算结果	当前日期 2005-11-26 19: 20: 19				
	NCIII 方差分析表(随机区组设计)				
变异来源	平方和	自由度	均方	F 值	p 值
区组间	2.0000	1	2.0000		
自交系	0.1800	1	0.1800	0.2423	0.6376
父本	6.4113	3	2.1371	2.8768	0.1128
父本×自交系	2.1075	3	0.7025	0.9457	0.4685
重复 × 组合	5.2000	7	0.7429		
误差	30.4300	16	1.9019		
总变异	46.3288	31			
σ _m 方差=	0.3486				
σ _{mf} 方差=	-0.0202				
V(A)方差=	1.3942				
V(d)方差=	-0.0404				
V(p)方差=	2.1371				
遗传力 h ² 估计值=		0.6524			
	NCIII 方差分析表(完全随机设计)				
变异来源	平方和	自由度	均方	F 值	p 值
自交系	0.1800	1	0.1800	0.1148	0.7377
父本	6.4113	3	2.1371	1.3630	0.2779
父本×自交系	2.1075	3	0.7025	0.4480	0.7210
误差	37.6300	24	1.5679		
总变异	46.3288	31			
σ _m 方差=	0.1423				
σ _{mf} 方差=	-0.4327				
V(A)方差=	0.5692				
V(d)方差=	-0.8654				
V(p)方差=	2.1371				
遗传力 h ² 估计值=		0.2663			

从分析结果可以看出，如试验采用随机区组方式设计实施，实验结果的 F_2 家系(亲本)加性遗传方差为 1.3942， F_2 家系与自交系互作的遗传方差(分量)，即显性遗传方差为 0，遗传力 h^2 估计值为 0.6524。

41.10 完全双列杂交 Griffing 配合力分析

Griffing(1956)把进行配合力分析的双列杂交设计归纳为四种交配设计方法。假定有 P 个供试的纯育亲本，它们彼此之间进行可能的杂交，最多可得 P^2 个杂交与自交组合。

依据其遗传特性, P^2 个组合又大致可分三组: ① P 个亲本, 即自交组合; ② 一组所有可能的杂交(正交)组合, 共有 $P(P-1)/2$ 个, 该组合简称为正交组合; ③ 一组与上述杂交组合相对应的反交, 也共有 $P(P-1)/2$ 个, 该组合简称反交组合。所以, 共有 $P+(P(P+1))/2+(P(P-1))/2=P^2$ 个组合。它们各个小区的平均数组成了 $P \times P$ 矩阵, 这个矩阵称为双列杂交表(表 41-13)。表中, x_{ii} = 第 i 个纯育亲本自交后代均数, x_{ij} = 第 i 个亲本与第 j 个亲本杂交 F_1 代的均数。

表 41-13 双列杂交表

雄雌	1	2	...	P	Σ
1	x_{11}	x_{12}	...	x_{1P}	$x_{1.}$
2	x_{21}	x_{22}	...	x_{2P}	$x_{2.}$
\vdots	\vdots	\vdots		\vdots	\vdots
P	x_{P1}	x_{P2}	...	x_{PP}	$x_{.P}$
Σ	$x_{.1}$	$x_{.2}$...	$x_{.P}$	$x_{..}$

依设计方法所包括的组合类型, 可以组成四种试验方法:

方法 1. 包含亲本自交及正反交组合, 共 P^2 个试验材料。

方法 2. 包含亲本自交及正交组合, 共 $P(P+1)/2$ 个试验材料。

方法 3. 不包含亲本, 仅包含正反交组合, 共 $P(P-1)$ 个试验材料。

方法 4. 仅包含正交组合, 共 $P(P-1)/2$ 个试验材料。

假定对上述试验采用随机区组设计, a 个供试材料, b 个区组, 每组合在一个区组中一个小区, 每小区观察 c 个个体, 其数据的数学模型为

$$x_{ijkl} = \mu + v_{ij} + b_k + (bv)_{ijk} + e_{ijkl}$$

式中, μ 为总体平均数; v_{ij} 为第 ij 个品种(遗传型)效应; b_k 为区组效应; $(bv)_{ijk}$ 为第 ij 遗传型在第 k 个区组效应; e_{ijkl} 为小区中第 l 个观察值的环境效应, 服从 $N(0, \sigma_e^2)$ 分布。

根据配合力的概念, 对于品种效应, 可进一步分解为 $v_{ij} = g_i + g_j + s_{ij}$ (不包括反交效应) 或 $v_{ij} = g_i + g_j + s_{ij} + r_{ij}$ (包括反交效应) 两种形式。

这里, g_i = 第 i 个亲本的一般配合力效应, g_j = 第 j 个亲本的一般配合力效应, g_{ij} = 第 i 个亲本与第 j 个亲本之间杂交的特殊配合力效应, r_{ij} = 第 i 个亲本与第 j 个亲本之间的反交效应。

上述 4 种试验方法, 因其抽样假定不同, 可分为固定模型与随机模型两种, 从而形成 8 种不同的分析方法。

一种假定我们研究目的是比较供试亲本的配合力或选择最优组合, 也就是估算这些亲本的效应值以及了解它们之间差异的状况。这里, 我们所得的结论只论及试验材料本身。因此, 各亲本、各组合在群体中的效应都是相对稳定的, 即 g_i, s_{ij} 都是一个固定常

数。这样的抽样假定称为固定模型，亦称为模型 I，且

$$\sum_i g_i = 0, \quad \sum_j g_j = 0, \quad \sum_i s_{ij} = \sum_j s_{ij} = 0, \quad \sum_k b_k = 0, \quad \sum_{ij} (bv)_{ijk} = \sum_k (bv)_{ijk} = 0$$

另一种假定是将试验材料看成从某个群体中抽出来的一个随机样本。这时，我们的试验目的并不在于探索材料本身的表现如何，而是从育种的角度，通过这些样本对总体的某些遗传状况作出估计，特别是估计复杂的群体方差的组成部分。如加性方差分量、非加性方差分量等，这时 v_{ij} 和 b_k 等效应值表现为随机变量。在此种假定下对数据所配合的模型称为随机模型，亦称为模型 II。在模型 II 中，各效应方差满足

$$V_{(gi)} = \sigma_g^2, \quad V_{(sij)} = \sigma_s^2, \quad V_{(bk)} = \sigma_b^2, \quad V_{(vij)} = \sigma_v^2, \quad V_{[(bv)ijk]} = \sigma_v^2$$

但无论固定模型或随机模型，都假定 e_{ijkl} 为正态随机变量，数学期望为 0，方差为 σ_e^2 ，即 $V_{(e_{ijkl})} = \sigma_e^2$ 。

对双列杂交设计进行配合力分析之前，首先要进行基因型间的方差分析，检验遗传型(组合)间的差异显著性，如果基因型间差异显著，然后才能进行配合力分析。现扼要介绍分析步骤及有关公式。

(1) 基因型间差异显著性检验。如前所述，在双列杂交试验中某田间小区内观察值的数学结构为

$$x_{ijkl} = \mu + v_{ij} + b_k + (bv)_{ijk} + e_{ijkl}$$

根据单因素随机区组方差分析原理，各类平方和计算方法如下，总平方和

$$CT = \left(\sum x_{ijkl} \right)^2 / (abc)$$

式中， a =供试组合数(因所用方法不同而异，后面将分别说明)， b =重复数， c =小区内观察值个数， $SS_T = \mu \sum x_{ijkl}^2 - CT$ ，再列区组与组合二向表，先求二向表的总平方和，然后再计算其余各种平方和

$$SS'_T = \frac{1}{c} \sum x_{ijk.}^2 - CT \quad (\text{两向表的总平方和})$$

$$SS_b = \frac{1}{ac} \sum x_{.k.}^2 - CT \quad (\text{区组平方和})$$

$$SS_v = \frac{1}{bc} \sum x_{ij.}^2 - CT \quad (\text{组合平方和})$$

$$SS_{bv} = SS'_T - SS_b - SS_v \quad (\text{区组} \times \text{组合平方和})$$

$$SS_e = SS_T - SS'_T \quad (\text{机误平方和})$$

并可列出方差分析表。注意，在固定模型中用 ϕ 代替随机模型中的 σ^2 ， ϕ 的计算公

式如下:

$$\begin{aligned}\phi(v) &= \frac{1}{a-1} \sum_i v_i^2 \\ \phi(b) &= \frac{1}{b-1} \sum_k b_k^2 \\ \phi_1(bv) &= \frac{1}{(a-1)(b-1)} \sum_{i<j} \sum_k (bv)_{ijk}^2 \\ \phi_2(bv) &= \frac{1}{(a-1)} \sum_{i<j} \sum_k (bv)_{ijk}^2 \\ \phi_3(bv) &= \frac{1}{(b-1)} \sum_k (bv)_{ijk}^2\end{aligned}$$

在方差分析中, 只需对 $\phi(v)=0$ (模型 I), 或对 σ_v^2 (模型 II) 进行检验。由于两者期望均方的构成不同, 其 F 检验亦不同, 在模型 I 中, $F_{((a-1), m)} = MS_v / MS_e$, 而在模型 II 中, $F_{((a-1), (a-1)(b-1))} = MS_v / MS_{bv}$ 。这里, $m=ab(c-1)$, MS 为均方。

(2) 配合力分析。若前面方差分析发现遗传型间差异显著, 就可进行配合力分析。配合力分析是以组合(遗传型)平均数为单位进行的。以下按不同方法分别叙述。

在双列杂交分析中, 使用模型 II 的目的主要在于估计群体方差的分量(即组成部分), 群体的加性遗传方差与非加性遗传方差能够从配合力方差组成部分估计。假定群体是自交系或纯系, 即 $F=1$, 那么, 它们的关系如下:

$$\begin{aligned}\hat{\sigma}_d^2 (\text{加性遗传方差}) &= 2\hat{\sigma}_g^2 \\ \hat{\sigma}_h^2 (\text{非加性遗传方差}) &= \hat{\sigma}_s^2 \\ \hat{\sigma}_G^2 &= 2\hat{\sigma}_g^2 + \hat{\sigma}_s^2 \\ \hat{\sigma}_p^2 &= \hat{\sigma}_G^2 + \hat{\sigma}_e^2 = 2\hat{\sigma}_g^2 + \hat{\sigma}_s^2 + \hat{\sigma}_e^2\end{aligned}$$

这些估计值在假定上位效应等于零时才是无偏的, 因此, 群体的遗传型方差与表型方差从下式估计。遗传力值的估计如下:

$$\text{广义遗传力, } \hat{h}_B^2(\%) = (2\hat{\sigma}_g^2 + \hat{\sigma}_s^2) / \hat{\sigma}_p^2 \times 100;$$

$$\text{狭义遗传力, } \hat{h}_N^2(\%) = 2\hat{\sigma}_g^2 / \hat{\sigma}_p^2 \times 100。$$

Pooni 和 Jinks(1984)从理论上证实了 Griffing 双列杂交分析法中的 σ_g^2 和 σ_s^2 与 D 和 H 的关系为: σ_g^2 在方法 1 中遗传分量为 $D/4$, 在方法 2, 3, 4 中为 $D/4$; σ_s^2 在方法 1 中遗传分量为 $H/2$, 在方法 2, 3, 4 中为 $H/4$ 。

对包含亲本自交及正反交组合的试验, 上述关系式在任何情况下均成立。对其他情况, 包含: ① 亲本自交及正交组合的方法; ② 不包含亲本而仅包含正反交组合的方法; ③ 以及仅包含正交组合的方法; ④ 指仅在设计中亲本为大样本时, 上述关系式才成立。Pooni 和 Jinks 用黄花烟草为供试材料, 进行双列杂交设计分析, 证实了上述结论。这在进行试验设计时是应当予以考虑的。

41.11 Griffing 分析：亲本+正反交 F_1 组合

该方法包括亲本自交及正反交共 P^2 个遗传型(组合)，即 $a=P^2$ ，各组合均数将构成一个 $P \times P$ 表，如表 41-13。在表 41-13 中，各边际总数的计算式为

$$X_{i\cdot} = \sum_j x_{ij}, \quad X_{\cdot j} = \sum_i x_{ij}, \quad X_{\cdot\cdot} = \sum_i \sum_j x_{ij}$$

表中，各项平方和计算公式如下：

一般配合力

$$SS_g = \frac{1}{2p} \sum_i (X_{i\cdot} + X_{\cdot i})^2 - \frac{2}{p^2} X_{\cdot\cdot}^2$$

特殊配合力

$$SS_s = \frac{1}{2} \sum_i \sum_j X_{ij} (X_{ij} + X_{ji})^2 - \frac{1}{2p} \sum_i (X_{i\cdot} + X_{\cdot i})^2 + \frac{1}{p^2} X_{\cdot\cdot}^2$$

反交效应

$$SS_r = \sum_{i < j} (X_{ij} - X_{ji})^2$$

根据各项平方和的分解，可列出配合力分析表如表 41-14。

表 41-14 配合力分析表

变异来源	自由度	平方和	均方
一般配合力 Gca	$(p-1)$	SS_g	MS_g
特殊配合力 Sca	$p(p-1)/2$	SS_s	MS_s
反交效应	$p(p-1)/2$	SS_r	MS_r
试验误差	m	SS_e	MS_e

表中试验误差均方 MS'_e 等于 $M_e/(b_c)$ ， MS'_e 的期望为 σ^2 ，试验误差的自由度 $m=ab(c-1)$ 。在进一步进行参数估计和假设测验时参数估计和假设测验依模型而定，因此对模型 I，组合均数的数学模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

式中， $i, j=1, 2, \dots, p$ ； $k=1, 2, \dots, b$ ； $l=1, 2, \dots, p$ ； μ 为总体平均数； v_{ij} 为第 ij 个品种(遗传型)效应； b_k 为区组效应； $(bv)_{ijk}$ 为第 ij 遗传型在第 k 个区组效应； e_{ijkl} 为小区中第 l 个观察值的环境效应，它服从 $N(0, \sigma_e^2)$ 分布。显然，品种效应为 $v_{ij}=g_i+g_j+s_{ij}+r_{ij}$ ，

限制条件为: $\sum g_i = 0$, 对每一个 j , $\sum S_{ij} = 0$ 。各种效应的假设测验以下面各式进行:

测验 G_{ca} 效应时, $F_{[(p-1), m]} = MS_g / MS'_e$;

测验 S_{ca} 效应时, $F_{[p(p-1)/2, m]} = MS_s / MS'_e$;

测验反交效应时, $F_{[p(p-1)/2, m]} = MS_r / MS'_e$ 。

对模型 II, 效应值的估计按下式进行:

群体均数 $\hat{\mu} = X_{..} / p^2$

一般配合力, $\hat{g}_i = \frac{1}{2p}(X_{i.} + X_{.i}) - \frac{1}{p^2} X_{..}$

特殊配合力, $\hat{s}_{ij} = \frac{1}{2}(X_{ij} + X_{ji}) - \frac{1}{2p}(X_{i.} + X_{.i} + X_{.j} + X_{.j}) + \frac{1}{p^2} X_{..}$

反交效应, $\hat{r}_{ij} = \frac{1}{2}(X_{ij} - X_{ji})$

其每一组合均数的统计模型如下:

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{b} \sum_k b_k + \frac{1}{b} \sum_k (bv)_{ijk} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

本模型除 μ 外, 其他一切效应皆属随机变量。由于 MS'_e 的期望等于 σ^2 , 而 σ^2 由式 $MS_{bc}/(bc)$ 估计, 这里 MS_{bc} 是品种与区组的互作均方。

为测验有关方差的假设, 可以采用 F 检验: $F = MS_g / MS^*$ 。其中, $MS^* = (1-A) MS'_e + AMS_s$, $A = p(p-1)/(p^2-p+1)$, 自由度 f 可由式

$$f = \frac{mp^3(p-1)[p+2(p-1)k]^2}{p^5(p-1)(1-A)^2 + 2mA^2(p^2+2ck)}$$

估算。式中

$$k = \frac{p^2}{2c} \left(\frac{MS_s - MS'_e}{MS'_e} \right), \quad c = p^2 \frac{p+1}{F_1}$$

F 测验时,

(1) 测验一般配合力, $F_{[(p-1), f]} = MS_g / m$;

(2) 测验特殊配合力, $F_{[p(p-1)/2, m]} = MS_g / MS'_e$;

(3) 测验反交效应时, $F_{[p(p-1)/2, m]} = MS_r / MS'_e$,

这里 $m = ab(c-1)$ 。

例如, 有 5 个亲本, 亲本及正反交 F_1 组合共有 5×5 个材料。试验设计为完全随机设计, 3 次重复, 每个小区内有 2 株 (2 个观察值), 其数据在电子表格中的输入整理格式如图 41-21。

	A	B	C	D	E	F	G	H
1	组	区组 I		区组 II		区组 III		
2	合	单株 1	单株 2	单株 1	单株 2	单株 1	单株 2	
3	1X1	8	10	11	13	11	7	
4	1X2	15	11	10	8	10	12	
5	1X3	14	10	16	18	14	18	
6	1X4	34	36	31	35	30	26	
7	1X5	14	16	18	16	15	17	
8	2X1	14	12	9	13	16	14	
9	2X2	12	16	18	14	16	20	
10	2X3	27	21	17	21	21	19	
11	2X4	7	8	9	11	9	10	
12	2X5	10	14	9	11	12	16	
13	3X1	21	17	24	26	24	25	
14	3X2	26	24	25	13	10	18	
15	3X3	13	10	18	21	22	24	
16	3X4	21	22	25	28	27	31	
17	3X5	13	10	18	26	24	25	
18	4X1	21	22	25	13	10	18	
19	4X2	26	24	25	21	22	25	
20	4X3	13	10	26	24	25	29	
21	4X4	21	22	13	10	18	16	
22	4X5	26	21	25	26	24	22	
23	5X1	23	24	25	13	10	18	
24	5X2	13	10	18	21	22	25	
25	5X3	21	22	25	26	24	25	
26	5X4	13	10	18	13	10	18	
27	5X5	21	22	25	21	22	25	
28								

图 41-21 亲本及正反交 F_1 组合配合力分析数据整理格式

输入完成后，并将待分析的数据定义成数据块，再进入菜单，选择“专业统计”→“遗传交配设计统计分析”→“亲本及正反交 F_1 组合 Griffing 配合力分析”功能，执行时系统会提示输入每个小区的单株数，输入后即可得到计算结果：

原始数据方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
区 组	2	79.41333	39.70667	4.474800		
处 理	24	4289.240	178.7183	20.14110	0.000100	固定模型
				6.123500	0.000100	随机模型
区组 X 处理	48	1400.920	29.18583	3.289162		
误 差	75	665.5000	8.873333			
总 的	149	6435.073				
配合力分析： 均值双向列表						
亲本 1	2	3	4	5	y(i.)	
p=1	10.000000	11.00000	15.00000	32.00000	16.00000	84.00000

p=2	13.00000	16.00000	21.00000	9.000000	12.00000	71.00000
p=3	22.83333	19.33333	18.00000	25.66667	19.33333	105.1667
p=4	18.16667	23.83333	21.16667	16.66667	24.00000	103.8333
p=5	18.83333	18.16667	23.83333	13.66667	22.66667	97.16667
y(.j)=	82.83333	88.33333	99.00000	97.00000	94.00000	
yi+yi.i=	166.8333	159.3333	204.1667	200.8333	191.1667	
固定模型分析结果						
配合力方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
一般配合力	4	164.3467	41.08667	27.78212	0.000100	
特殊配合力	10	214.0961	21.40961	14.47682	0.000100	
反交	10	336.4306	33.64306	22.74887	0.000000	
误差	75	110.9167	1.478889			
一般配合力效应及其比较						
平均数标准误 $S_x=0.3846$ ，平均数间差数标准误 $S_d=0.5439$						
最小显著差 $LSD(5\%)=1.0847$ $LSD(1\%)=1.4401$ ，LSD 最小显著性差异测验结果						
处理	平均	5%显著水平		1%极显著水平		
3	1.970000	a		A		
4	1.636700	ab		A		
5	0.670000	b		A		
1	-1.76330	c		B		
2	-2.51330	c		B		
自交组合对非自交组合间比较的临界值 $LSD_{0.05}=2.7622$ $LSD_{0.01}=3.6647$						
自交组间比较的临界值 $LSD_{0.05}=2.6538$ $LSD_{0.01}=3.5209$						
无共同亲本间比较的临界值 $LSD_{0.05}=2.2983$ $LSD_{0.01}=3.0492$						
非自交组间(有 1 共同亲本)比较的临界值 $LSD_{0.05}=2.1668$ $LSD_{0.01}=2.8748$						
非自交组间(无共同亲本)比较的临界值 $LSD_{0.05}=1.8765$ $LSD_{0.01}=2.4896$						
特殊配合力						
亲本	2	3	4	5		
1	-2.170000	0.263300	6.763300	0.063300		
2		2.263300	-1.153300	-1.520000		
3			1.363300	0.496700		
4				-1.920000		
反交效应						
亲本	2	3	4	5		
1	-1.000000	-3.916700	6.916700	-1.416700		
2		0.833300	-7.416700	-3.083300		
3			2.250000	-2.250000		
4				5.166700		
每种效应的方差估计						
均值 18.4467，一般配合力方差 0.1183，特殊配合力方差 0.9465，反交效应方差 0.7394						
随机模型分析结果						
各个方差成分的估计与测验						
成分	方差		F 值	F_1 显著水平		
一般配合力	2.062600		2.008100	0.106700		
特殊配合力	11.86350		14.47680	0.000100		
反交	402.0521		22.74890	0.000100		
误差	0.058300					
遗传参数估计						

加性方差=4.1252, 显性方差=11.8635, 遗传方差=15.9888, 环境方差=8.8733, 表型方差=24.8621
 广义遗传力=64.31%, 狭义遗传力=16.59%

41.12 Griffing 分析: 亲本+正交 F_1 组合

该方法包括亲本自交及正交组合, 因此, 共有 $p(p+1)/2$ 个遗传型(组合)。各组合均数列成如表 41-15。

表 41-15 亲本+正交 F_1 组合试验各组合均数二向表

亲本	1	2	...	p	Σ
1	x_{11}	x_{12}	...	x_{1p}	$X_{1.}$
2		x_{22}	...	x_{2p}	$X_{2.}$
⋮			...	⋮	⋮
P				x_{pp}	$x_{p.}$

在进行配合力分析之前, 先进行方差分析。在此基础上, 将小区各个区组合并平均, 构成表 41-15 的两向平均值表, 然后进行配合力分析。

对模型 I, 中进行配合力分析, 假定每个均数的统计模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

式中, $i, j=1, 2, \dots, p$; $k=1, 2, \dots, b$; $l=1, 2, \dots, p$; μ 为总体平均数; v_{ij} 为第 ij 个品种(遗传型)效应; b_k 为区组效应; $(bv)_{ijk}$ 为第 ij 个遗传型在第 k 个区组内的效应; e_{ijkl} 为小区中第 l 个观察值的环境效应, 它服从 $N(0, \sigma_e^2)$ 分布。显然, 其品种效应为

$$v_{ij} = g_i + g_j + s_{ij} + r_{ij}$$

并有 $s_{ij} = s_{ji}$ 。限制条件为 $\sum g_i = 0$, 对每一个 i , $\sum s_{ij} + s_{ji} = 0$ 。同时 $MS'_e = MS_{e/(bc)}$ 。模型中各种效应的假设测验以下面各式进行:

检测一般配合力 G_{ca} 效应, $F_{[(p-1), m]} = MS_g / MS'_e$;

检测特殊配合力 S_{ca} 效应时, $F_{[p(p-1)/2, m]} = MS_s / MS'_e$ 。

各个效应值的估计方法为:

群体均值: $\hat{\mu} = \frac{2}{p(p+1)} X_{..}$;

一般配合力: $\hat{g}_i = \frac{1}{p+2} (X_{i.} + X_{.i} - \frac{2}{p} X_{...})$;

特殊配合力: $\hat{s}_{ij} = X_{ij} - \frac{1}{p+2} (X_{i.} + X_{.i} + X_{j.} + X_{.j}) + \frac{2}{(p+1)(p+2)} X_{..}$ 。

对模型 II,

$$\text{群体均值: } \hat{\mu} = \frac{1}{p^2} X_{..};$$

$$\text{一般配合力: } \hat{g}_i = \frac{1}{2p} (X_{i.} + X_{.i}) - \frac{1}{p^2} X_{..};$$

$$\text{特殊配合力: } \hat{s}_{ij} = \frac{1}{2} (X_{ij} + X_{ji}) - \frac{1}{2p} (X_{i.} + X_{.i} + X_{.j} + X_{.j}) + \frac{1}{p^2} X_{..}。$$

其每一组合均数的统计模型如下:

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{b} \sum_k b_k + \frac{1}{b} \sum_k (bv)_{ijk} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

本模型除 μ 外, 其他一切效应皆属随机变量。由于 MS'_e 的期望等于 σ^2 , 而 σ^2 由式 $MS_{bc}/(bc)$ 估计, 这里 MS_{bc} 是品种与区组的交互均方。

为测验有关方差的假设, 可以采用 F 检验:

(1) 测验一般配合力, F 测验为 $F_{[(p-1), p(p-1)/2]} = MS_g/m, m=ab(c-1)$;

(2) 测验特殊配合力, F 测验为 $F_{[p(p-1)/2, m]} = MS_g/MS'_e$ 。

例如, 有 5 个亲本, 亲本及正反交 F_1 组合共有 $5 \times (5+1)/2$ 个材料。试验设计为完全随机设计, 3 次重复, 每个小区内有 2 株(2 个观察值), 其数据在电子表格中的输入整理格式如图 41-22。

	A	B	C	D	E	F	G	H
1	组	区组 I		区组 II		区组 III		
2	合	单株 1	单株 2	单株 1	单株 2	单株 1	单株 2	
3	1X1	8	10	11	13	11	7	
4	1X2	15	11	10	8	10	12	
5	1X3	14	10	16	18	14	18	
6	1X4	34	36	31	35	30	26	
7	1X5	14	16	18	16	15	17	
8	2X2	14	12	9	13	16	14	
9	2X3	12	16	18	14	16	20	
10	2X4	27	21	17	21	21	19	
11	2X5	7	8	9	11	9	10	
12	3X3	10	14	9	11	12	16	
13	3X4	21	17	24	18	26	20	
14	3X5	8	10	11	14	12	11	
15	4X4	26	24	25	29	27	31	
16	4X5	13	10	18	16	14	13	
17	5X5	21	22	25	28	20	22	
18								

图 41-22 亲本及一组正交组合配合力分析数据整理格式

将数据按图 41-22 格式输入, 并将待分析的数据定义成数据块, 再进入菜单, 选择“专业统计” → “遗传交配设计统计分析”下的“亲本及一组 F_1 组合 Griffing 配合力分

析”功能，执行时系统会提示输入每个小区的单株数，输入后即可得到计算结果：

原始数据方差分析表							
变异来源	自由度	平方和	均方	F 值	显著水平		
区组	2	22.86667	11.43333	2.198700			
处理	14	3797.600	271.2571	52.16480	0.000100	固定模型	
				26.63740	0.000100	随机模型	
区组 X 处理	28	285.1333	10.18333	1.958333			
误差	45	234.0000	5.200000				
总的	89	4339.600					
配合力分析均值双向列表							
亲本	1	2	3	4	5	yi	yi+yii
p=1	10.0000	11.000	15.000	32.000	16.000	84.000	94.000
p=2		13.000	16.000	21.000	9.0000	70.000	83.000
p=3			12.000	21.000	11.000	75.000	87.000
p=4				27.000	14.000	115.00	142.00
p=5					23.000	73.000	96.000
固定模型分析结果							
配合力方差分析表							
变异来源	自由度	平方和	均方	F 值	显著水平		
一般配合力	4	324.7429	81.18571	93.67582	0.000000		
特殊配合力	10	308.1905	30.81905	35.56044	0.000000		
误差	45	39.00000	0.866667				
一般配合力效应及其比较							
平均数标准误 $S_x=0.3519$							
平均数间差数标准误 $S_d=0.4976$							
最小显著差 $LSD(5\%)=1.0057$, $LSD(1\%)=1.3458$							
LSD 最小显著性差异测验结果							
处理	平均	5%显著水平		1%极显著水平			
4	5.942900	a		A			
5	-0.628600	b		B			
1	-0.914300	bc		B			
3	-1.914300	cd		BC			
2	-2.485700	d		C			
自交组间比较的临界值 $LSD_{0.05}=1.7359$ $LSD_{0.01}=2.3181$							
具共同亲本间比较的临界值 $LSD_{0.05}=2.4550$ $LSD_{0.01}=3.2783$							
非自交组间(无共同亲本)比较的临界值 $LSD_{0.05}=2.2411$ $LSD_{0.01}=2.9927$							
特殊配合力							
亲本	2	3	4	5			
1	-2.333300	1.095200	10.238100	0.80950			
2		3.666700	0.809500	-4.61900			
3			-9.761900	12.80950			
4				-22.04760			
每种效应的方差估计 均值 16.7333 ($x_{..}$)=0.0578, 一般配合力方差 0.0990							
ii 特殊配合力方差 0.4127							
ij 特殊配合力方差 0.4540							
随机模型分析结果							
各个方差成分的估计与测验							
成分	方差	F 值	显著水平				

一般配合力	81.18571	2.634271	0.097592
特殊配合力	29.95238	35.56044	0.000000
误差	0.033383		
遗传参数估计			
加性方差=14.3905, 显性方差=29.9524, 遗传方差=44.3429			
环境方差=5.2000, 表型方差=42.3476, 遗传决定度=70.73%, 狭义遗传力=33.98%			

41.13 Griffing 分析: 无亲本, 仅正反交 F_1 组合

本试验仅包括亲本间正反交组合, 不包括自交亲本, 故有 $p(p-1)$ 个组合, 各组合均数列成如表 41-16 所示。

表 41-16 仅正交 F_1 组合试验各组合均数二向表

雌亲 \ 雄亲	雄亲					Σ
	1	2	...	p		
1	x_{11}	x_{12}	...	x_{1p}		$X_{1.}$
2	x_{21}	x_{22}	...	x_{2p}		$X_{2.}$
⋮	⋮	⋮		⋮		⋮
p	x_{p1}	x_{p2}	...	x_{pp}		$x_{p.}$
Σ	$X_{.1}$	$X_{.2}$...	$x_{.p}$		

在进行配合力分析之前, 先进行方差分析。在此基础上, 将小区各个区组合并平均, 构成表 41-16 的两向平均值表, 然后进行配合力分析。

在模型 I 中, 进行配合力分析, 假定每个均数的统计模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

式中, $i, j=1, 2, \dots, p; k=1, 2, \dots, b; l=1, 2, \dots, p; \mu$ 为总体平均数; v_{ij} 为第 ij 个品种(遗传型)效应; b_k 为区组效应; $(bv)_{ijk}$ 为第 ij 个遗传型在第 k 个区组内的效应; e_{ijkl} 为小区中第 l 个观察值的环境效应, 它服从 $N(0, \sigma_e^2)$ 分布。

显然, 品种效应为 $v_{ij}=g_i+g_j+s_{ij}+r_{ij}$, 并有 $s_{ij}=s_{ji}, r_{ij}=-r_{ji}$, 限制条件: $\sum g_i = 0$, 对每一个 $j, \sum_{i \neq j} s_{ij} = 0$, 同时 $MS'_e = MS_e / (bc)$ 。各种效应的假设测验以下面各式进行:

检测一般配合力 Gca 效应, $F_{[(p-1), m]} = MS_g / MS'_e$;

检测特殊配合力 Sca 效应时, $F_{[p(p-3)/2, m]} = MS_s / MS'_e$;

测验反交效应时, $F_{[p(p-1)/2, m]} = MS_r / MS'_e$;

群体均值 $\hat{\mu} = \frac{1}{p(p-1)} X_{..}$;

$$\text{一般配合力 } \hat{g}_i = \frac{1}{2p(p-2)}(p(X_{i\cdot} + X_{\cdot i}) - 2X_{\cdot\cdot});$$

$$\text{特殊配合力 } \hat{s}_{ij} = \frac{1}{2}(x_{ij} + x_{ji}) - \frac{1}{2(p-2)}(X_{i\cdot} + X_{\cdot i} + X_{j\cdot} + X_{\cdot j}) - \frac{1}{(p-1)(p-2)}X_{\cdot\cdot};$$

$$\text{反交效应 } \hat{r}_{ij} = \frac{1}{2}(x_{ij} - x_{ji}).$$

各个效应值的估计方法为：对模型 II，j 假定每一组合均数的统计模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{b} \sum_k b_k + \frac{1}{b} \sum_k (bv)_{ijk} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

本模型除 μ 外，其他一切效应皆属随机变量。为测验有关方差的假设，可以采用 F 检验：

- (1) 测验一般配合力， F 测验为 $F_{[(p-1), p(p-3)/2]} = MS_g / MS_s$;
- (2) 测验特殊配合力， F 测验为 $F_{[p(p-3)/2, m]} = MS_s / MS'_e$;
- (3) 测验反交配合力， F 测验为 $F_{[p(p-1)/2, m]} = MS_r / MS'_e$ 。

例如，有 5 个亲本，其正反交 F_1 组合共有 $5 \times (5-1)$ 个材料。试验设计为完全随机设计，3 次重复，每个小区内有 2 株(2 个观察值)，其数据在电子表格中的输入整理格式如图 41-23。

	A	B	C	D	E	F	G	H
1	组	区组 I		区组 II		区组 III		
2	合	单株 1	单株 2	单株 1	单株 2	单株 1	单株 2	
3	1X2	15	11	10	8	10	12	
4	1X3	14	10	16	18	14	18	
5	1X4	34	36	31	35	30	26	
6	1X5	14	16	18	16	15	17	
7	2X1	14	12	9	13	16	14	
8	2X3	27	21	17	21	21	19	
9	2X4	7	8	9	11	9	10	
10	2X5	10	14	9	11	12	16	
11	3X1	21	17	24	26	24	25	
12	3X2	26	24	25	13	10	18	
13	3X4	21	22	25	28	27	31	
14	3X5	13	10	18	26	24	25	
15	4X1	21	22	25	13	10	18	
16	4X2	26	24	25	21	22	25	
17	4X3	13	10	26	24	25	29	
18	4X5	26	21	25	26	24	22	
19	5X1	23	24	25	13	10	18	
20	5X2	13	10	18	21	22	25	
21	5X3	21	22	25	26	24	25	
22	5X4	13	10	18	13	10	18	
23								

图 41-23 仅有正反交组合试验数据编辑格式

将数据按图 41-23 格式输入，并将待分析的数据定义成数据块，再进入菜单，选择“专业统计”→“遗传交配设计统计分析”下的“无亲本，仅正反交 F_1 组合 Griffing 配合力分析”功能，执行时系统会提示您输入每个小区的单株数，输入后即可得到计算

结果:

原始数据方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
区组	2	60.51667	30.25833	3.033400		
处理	19	3674.425	193.3908	19.38750	0.000100	固定模型
				6.400600	0.000100	随机模型
区组 X 处理	38	1148.150	30.21447	3.029020		
误差	60	598.5000	9.975000			
总的	119	5481.592				
配合力分析均值两向列表						
亲本	1	2	3	4	5	y(i)
p=1		11.00000	15.00000	32.00000	16.00000	74.00000
p=2	13.00000		21.00000	9.000000	12.00000	55.00000
p=3	22.83333	19.33333		25.66667	19.33333	87.16667
p=4	18.16667	23.83333	21.16667		24.00000	87.16667
p=5	18.83333	18.16667	23.83333	13.66667		74.50000
y(.j)	72.83333	72.33333	81.00000	80.33333	71.33333	
固定模型分析结果						
配合力方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
一般配合力	4	195.1704	48.79259	29.34893	0.000100	
特殊配合力	5	80.80324	16.16065	9.720691	0.000100	
反交	10	336.4306	33.64306	20.23642	0.000000	
误差	60	99.75000	1.662500			
一般配合力效应及其比较						
平均数标准误 $S_x=0.5264$, 平均数间差数标准误 $S_d=0.7444$						
最小显著差 $LSD(5\%)=1.4919$, $LSD(1\%)=1.9863$						
LSD 最小显著性差异测验结果						
处理	平均	5% 显著水 平	1% 极显著水 平			
3	2.838900	a	A			
4	2.727800	a	A			
1	-0.716700	b	B			
5	-0.883300	b	B			
2	-3.966700	c	C			
具共同母本的特殊配合力差异比较的临界值 $LSD_{0.05}=2.1059$ $LSD_{0.01}=2.8007$						
不同父母本各个组间比较的临界值 $LSD_{0.05}=1.4891$ $LSD_{0.01}=1.9804$						
特殊配合力						
亲本	2	3	4	5		
1	-2.208300	-2.097200	4.180600	0.125000		
2		2.402800	-1.236100	1.041700		

3			-1.041700	0.736100	
4				-1.902800	
反交效应					
亲本	2	3	4	5	
1	-1.000000	-3.916700	6.916700	-1.416700	
2		0.833300	-7.416700	-3.083300	
3			2.250000	-2.250000	
4				5.166700	
每种效应的方差 21.8 Griffing 分析: 无亲本, 仅一组 F_1 试验					
一般配合力方差 0.2217, 特殊配合力方差 0.4156, 反交效应方差 0.8313					
随机模型分析结果					
配合力方差分析表					
变异来源	自由度	平方和	均方	F 值	显著水平
一般配合力	4	195.1704	48.79259	3.019223	0.024600
特殊配合力	5	80.80324	16.16065	9.720691	0.000100
反交	10	336.4306	33.64306	20.23642	0.000000
误差	60	99.75000	1.662500		
各个方差成分的估计与测验					
成分	方差	F 值	显著水平		
一般配合力	5.438700	3.019200	0.128500		
特殊配合力	9.720700	9.720700	0.000100		
反交	20.23640	0.000100			
误差	0.092100				
遗传参数估计					
加性方差=10.8773, 显性方差=7.2491, 遗传方差=18.1264, 表型方差=28.1014, 广义遗传力=64.50%, 狭义遗传力=38.71%					

41.14 Griffing 分析: 无亲本, 仅一组 F_1 试验

本试验仅包括各亲本的正交组合, 不包括自交亲本, 故有 $p(p-1)/2$ 个组合, 各组合均数列成表 41-17。

表 41-17 仅正交 F_1 组合试验各组合均数二向表

亲本	1	2	...	p	Σ
1		x_{12}	...	x_{1p}	$X_{1.}$
2			...	x_{2p}	$X_{2.}$
\vdots				\vdots	\vdots
$p-1$				$x_{p-1, p}$	$x_{p-1.}$
p					$x_{p.}$

在进行配合力分析之前, 先进行方差分析。在此基础上, 将小区各个区组合并平均, 构成表 41-17 的两向平均值表, 然后进行配合力分析。

在模型 I 中进行配合力分析, 假定每个均数的统计模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{bc} \sum_k \sum_{F_1} e_{ijkl}$$

式中, $i, j=1, 2, \dots, p$; $k=1, 2, \dots, b$; $l=1, 2, \dots, p$; μ 为总体平均数; v_{ij} 为第 ij 个品种(遗传型)效应; b_k 为区组效应; $(bv)_{ijk}$ 为第 ij 个遗传型在第 k 个区组效应; e_{ijkl} 为小区中第 l 个观察值的环境效应, 它服从 $N(0, \sigma_e^2)$ 分布。显然, 其品种效应为

$$v_{ij} = g_i + g_j + s_{ij} + r_{ij}$$

并有 $s_{ij} = s_{ji}$, $r_{ij} = -r_{ji}$, 限制条件是 $\sum g_i = 0$, 对每一个 j , $\sum_{i \neq j} s_{ij} = 0$, 同时 $MS'_e = MS_e / (bc)$ 。

模型中各种效应的假设测验可根据下面各式进行:

检测一般配合力 G_{ca} 效应, $F_{[(p-1), m]} = MS_g / MS'_e$;

检测特殊配合力 S_{ca} 效应时, $F_{[p(p-3)/2, m]} = MS_s / MS'_e$;

群体均值: $\hat{\mu} = \frac{2}{p(p-1)} X_{..}$;

一般配合力: $\hat{g}_i = \frac{1}{2p(p-2)} (pX_{i.} - 2X_{..})$;

特殊配合力: $\hat{s}_{ij} = \frac{1}{(p-2)} (X_{i.} + X_{.j}) - \frac{2}{(p-1)(p-2)} X_{..}$ 。

各个效应值, 对模型 II, 假定每一组合均数的统计模型为

$$X_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + \frac{1}{b} \sum_k b_k + \frac{1}{b} \sum_k (bv)_{ijk} + \frac{1}{bc} \sum_k \sum_l e_{ijkl}$$

本模型除 μ 外, 其他一切效应皆属随机变量。为测验有关方差的假设, 可以采用 F 检验:

(1) 测验一般配合力, $F_{[(p-1), p(p-3)/2]} = MS_g / MS_s$;

(2) 测验特殊配合力, $F_{[p(p-3)/2, m]} = MS_s / MS'_e$ 。

例如, 有 5 个亲本, 其正交 F_1 组合共有 $5 \times (5-1)$ 个材料。试验设计为完全随机设计, 3 次重复, 每个小区内有 2 株(2 个观察值), 其数据在电子表格中的输入整理格式如图 41-24。

	A	B	C	D	E	F	G	H
1	组	区组 I		区组 II		区组 III		
2	合	单株 1	单株 2	单株 1	单株 2	单株 1	单株 2	
3	1X2	14	12	9	13	16	14	
4	1X3	12	16	18	14	16	20	
5	1X4	27	21	17	21	21	19	
6	1X5	7	8	9	11	9	10	
7	2X3	10	14	9	11	12	16	
8	2X4	21	17	24	18	26	20	
9	2X5	8	10	11	14	12	11	
10	3X4	26	24	25	29	27	31	
11	3X5	13	10	18	16	14	13	
12	4X5	21	22	25	28	20	22	
13								

图 41-24 仅有正交组合试验数据编辑格式

将数据按上述表格的格式输入，并将待分析的数据定义成数据块，再进入菜单，选择“专业统计”→“遗传交配设计统计分析”下的“无亲本，仅一组 F_1 组合 Griffing 配合力分析”功能，执行时系统会提示输入每个小区的单株数，输入后即可得到计算结果：

原始数据方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
区组	2	35.10000	17.55000	3.133900		
处理	9	1908.600	212.0667	37.86900	0.000100	固定模型
				23.72410	0.000100	随机模型
区组 X 处理	18	160.9000	8.938889	1.596230		
误差	30	168.0000	5.600000			
总的	59	2272.600				
配合力分析						
均值双向列表						
亲本	1	2	3	4	5	
p=1		13.00000	16.00000	21.00000	9.000000	
p=2			12.00000	21.00000	11.00000	
p=3				27.00000	14.00000	
p=4					23.00000	
固定模型分析结果						
配合力方差分析表						
变异来源	自由度	平方和	均方	F 值	显著水平	
一般配合力	4	297.6000	74.40000	79.71429	0.000100	
特殊配合力	5	20.50000	4.100000	4.392857	0.004100	
误差	30	28.00000	0.933333			
一般配合力效应及其比较						
平均数标准误 $S_x=0.5578$						
平均数间差数标准误 $S_d=0.7888$						
最小显著差异		LSD(5%)=1.6246		LSD(1%)=2.1988		

LSD 最小显著性差异测验结果				
处理	平均	5%显著水平		1%极显著水平
4	8.400000	a		A
3	0.733300	b		B
1	-2.600000	c		C
2	-3.266700	c		C
5	-3.266700	c		C
具共同亲本间比较的临界值, $LSD_{0.05}=2.2783$ $LSD_{0.01}=3.0677$				
非自交组合间(无共同亲本)比较的临界值 $LSD_{0.05}=1.6110$ $LSD_{0.01}=2.1692$				
特殊配合力				
亲本	2	3	4	5
1	2.166667	1.166667	-1.500000	-1.833333
2		-2.166667	-0.833333	0.833333
3			1.166667	-0.166667
4				1.166667
每种效应的方差估计				
均值 16.7000				
(x..)=0.0933				
一般配合力方差 0.2489, 特殊配合力方差 0.4667				
随机模型分析结果(各个方差成分的估计与测验)				
成分	方差	F 值	显著水平	
一般配合力	74.40000	18.14630	0.003500	
特殊配合力	3.166700	4.392900	0.004000	
误差	0.058100			
遗传参数估计				
加性方差=46.8667, 显性方差=3.1667, 遗传方差=50.0333, 表型方差=55.6333				
遗传决定度=89.93%, 狭义遗传力=84.24%				

41.15 双列杂交设计 Hayman 分析法

双列杂交的 Hayman(1954, 1960)分析法, 目的是要弄清基因的效应, 如正反交效应、加性效应、显性效应等, 主要从遗传角度考虑问题。Hayman 分析法研究的主要是自交群体, 前面介绍的 Griffing 法研究的主要是随机交配群体。应用何种方法要依据研究目的和试验设计而定, 如玉米自交系, 配合力分析显得格外重要, 但对于估计自交群体的遗传参数, Hayman 的分析法则更可行。

41.15.1 双列杂交 Hayman 分析法数学模型

Hayman 分析方法是采用随机模型进行分析。考虑母体效应(或细胞质效应)的线性模型为

$$y_{ijkl} = \mu + c_i + g_{ij} + b_k + (bg)_{ijk} + \varepsilon_{ijk}$$

式中, Y_{ijkl} 为第 i 个母本和第 j 个父本杂交组合在第 k 个区组中的第 l 个观察值, μ 为群体平均数, g_{ij} 为该组合的平均数, b_k 为区组效应, $(bg)_{ijk}$ 为组合效应与区组效应的交互,

c_i 为第 i 个母本的细胞质效应。试验设计为完全随机设计。这是以个体为单位的模型，若以小区为单位，则模型为

$$y_{ijk} = \mu + c_i + g_{ij} + b_k + \varepsilon_{ijk}$$

式中， Y_{ijk} 为第 i 个母本和第 j 个父本杂交组合在第 k 个区组中的观察值， μ 为群体平均数， g_{ij} 为该组合的平均数， b_k 为区组效应， c_i 为第 i 个母本的细胞质效应。此模型采用随机区组设计。在应用上，双列杂交也可以采用随机区组设计。

Hayman 方法一般采用第一种完全双列杂交设计。模型中，剔出细胞质效应后，使 $g_{ij}=g_{ji}$ ，即去除细胞质效应后细胞核效应不因为正反交而改变。去除的方法是：将正反交组合平均作为两个亲本杂交组合的表现型值。组合的遗传效应 g_{ij} 可以进一步分解 $g_{ij}=D_i+S_j+(DS)_{ij}$ 。其中， D_i 为第 i 个母本的加性效应， S_j 为第 j 个父本的加性效应， $(DS)_{ij}$ 为第 i 个亲本与第 j 个亲本间显性效应。

双列设计分析时，(a)可以初步整理数据，做成两张表：一是正反交之差，二是正反交之和： $X_{ij}-X_{ji}$ 和 $X_{ij}+X_{ji}$ ，使用差值表可以分析母本(细胞质)效应和正反交效应，使用和价值以分析基因的加性效应、显性效应、显性方向等；(b)对变异和自由度进行分解，按照项目可以分解为以下几部分：(c)测验平均的母性效应(每个家系)；(d)测定正反中的效应中不属于(c)的其余效应；测定基因的加性效应的显著性，即 D_R ；(e)测定显性效应。遗传分析中，可以正反交平均数做分析，这样可以除去细胞质效应，得到表 41-18。

表 41-18 完全双列杂交自由度分解

	数据的个数	自由度	测定的效应	
正反交差异	$p(p-1)/2$	$p-1$	(c)测定母性效应	
		$(p-1)(p-2)/2$	(d)测定正反交效应-母性效应	
	$p(p+1)/2$	$p-1$	(a)加性效应	
		$P(p-1)/2$	(b)显性效应	
正反交和	b_1	1	自交对杂交或者 F_1 对 MP 的方向显性	
	b	b_2	$p-1$	测定双列中 F_1 显性离差 $X_{ij}-MP$
	b_3	$p(p-3)/2$	测定显性离差部分，相当于 Griffing 方法的特殊配合力	

双列杂交组合平均数，如果按上述模型像 NCII 的形式分析，所得到的信息与 NCII 设计基本相同。据此模型，依方差分析原理可以求出加性效应的方差和显性效应的方差估计值。计算得到的加性遗传方差为 $\sigma_A^2 = D_R / 4$ ，为父、母本半同胞方差之和。如果将双列杂交视为(A)(父本)、(B)(母本)两因素的试验，那么这里加性方差为 $\sigma_D^2 + \sigma_S^2$ 。显性方差为父本与母本间互作的遗传方差： $\sigma_{DS}^2 = H_R / 16$ 。

41.15.2 Hayman 分析方法的原理

Hayman 分析方法要求亲本是纯合的，所以下面以纯合亲本为例说明世代的遗传方

差和协方差(Jinke, Hayman, 1953)。以下以自交体系为例说明原理, 它对于随机交配得到的纯合群体的分析也是适用的, 只是方差的涵义改变了。自交多代得到纯系群体, AA 和 aa 频率均为 0.5。现将一对等位基因的情况列于表 41-19, 进一步计算方差和协方差, 并依据方差进行推理。

表 41-19 一对等位基因双列杂交的方差-协方差

亲本	AA d	aa $-d$	均数
AA d	AA D	Aa h	$(d+h)/2$
aa $-d$	Aa H	Aa $-d$	$(-d+h)/2$
均数 r	$(d+h)/2$	$(h-d)/2$	$h/2$
V_r	$(d-h)^2/2$	$(d+h)^2/2$	$(d^2+h^2)/4$
W_r	$d(d-h)/2$	$d(d+h)/2$	$d^2/4$

aa 行的方差 $V_{r(aa)}=(d+h)^2/4$, AA 行的方差 $V_{r(AA)}=(h-d)^2/4$, 行方差的平均为 $\bar{V}_r = (d^2 + h^2)/4$ 。

半同胞方差, 乃对应于亲本的家系平均数的方差, 或者说是行平均数的方差 $\bar{V}_r = d^2/4$ 。

因此, 如果 $h=0$, 则 $V_{r(AA)}=V_{r(aa)}$; 若 $h \neq 0$, 则 $V_{r(AA)} \neq V_{r(aa)}$, 因此可以对 V_{ri} (行方差间) 进行同质性检测, 即对各行方差的同质性进行检验, 如果显著, 则说明显性效应存在, 如果不显著则说明显性效应不存在。

计算 AA 行的家系与其对应的非共同亲本之间的协方差。共同亲本为 AA, 非共同亲本为各列的亲本, 那么协方差 $W_{r(AA)}=d(d-h)/2$, aa 行的家系与其对应的非共同亲本之间的协方差 $W_{r(aa)}=d(h+d)/2$ 。AA 与 aa 行的方差平均为 $\bar{W}_r = d^2/2$ 。

因此, 如果 $h=0$, 则 $W_{r(AA)}=W_{r(aa)}$; 若 $h \neq 0$, 则这种关系不复存在。因此, 利用协方差也可以推断显性效应是否存在。行方差(家系间的)与行协方差(家系与非共同亲本间的)的关系还可以作进一步推导, 行间方差的差值和协方差的差值分别为 $\Delta V_r = dh$, $\Delta W_r = dh$, 即 $\Delta V_r = \Delta W_r$ 。

综合分析, 可得三点原理: ① 若 k 对控制性状的基因独立时, 则属加性-显性模型, W_r 在 V_r 上回归为 $\beta=1$ 的直线, 此标准可以检验基因的效应是否属于加性-显性模型; ② 若 $H_0: \beta=1$ 被接受, 则具有较多显性基因的公共亲本位于回归线的左下方, 而较少显性基因的公共亲本位于回归线的右上方; ③ 回归截距 a 对显性的度量, 因 $a=(d^2-h^2)/4$, 如 k 对基因则 $a=(D-H)/4$, 若 $|h|<d$, 则 $a>0$ 为部分显性; $|h|=d$, 则 $a=0$ 为完全显性; $|h|>d$, 则 $a<0$ 为超显性。这些原理也适用于组合数为 $p(p+1)/2$ 的双列杂交设计。但是, 对于组合数为 $p(p-1)/2$ 的双列杂交设计, 情况则有所不同, 该设计不包括亲本, 可以序列均数 (\bar{r}) 代替亲本平均数, 可以推导得到

$$\beta = \Delta W_r / \Delta V_r = 0.5, \quad a = \bar{W}_r - \beta \bar{V}_r = (d^2 - h^2) / 8$$

当基因频率不为 0.5 时, 有 $\beta=0.5-(u-v)h/(2d)$ 。因此, 对于该设计应用 Hayman 方法分析可能出现问题。

进一步分析 W_r 和 V_r 的关系, 以用于检验非加性效应是否存在。计算行方差与行协方差二者之和以及二者之差, 则有

$$W_{r(AA)} + V_{r(AA)} = (3d/4 - h/4)(d - h)$$

$$W_{r(aa)} + V_{r(aa)} = (3d/4 + h/4)(d + h)$$

显然, 若各个公共亲本的 $W_{ri}+V_{ri}$ 彼此间差异显著, 则可表明显性效应存在, 即 $h \neq 0$; 若无显著差异, 则显性效应不存在。

同样, 由 $W_{ri}-V_{ri}$ 得

$$W_{r(AA)} - V_{r(AA)} = d(d-h)/2 + (d-h)^2/4 = (d^2 - h^2)/4$$

$$W_{r(aa)} - V_{r(aa)} = d(d+h)/2 + (d+h)^2/4 = (d^2 + h^2)/4$$

可见, 各公共亲本的 $W_{ri}-V_{ri}$ 间有显著的差异, 则表明有上位性存在; 如无显著差异, 则表明不存在上位性效应。

按照随机区组设计, 有 r 次重复, 那么双列杂交试验的每个区组可得到一套 $(W_{ri}+V_{ri})$ 和 $(W_{ri}-V_{ri})$, 共 p 个; r 个区组产生 rp 个, 形成 $r \times p$ 的二向表, 进行方差分析, 通过 F 检验评定公共亲本 $(W_{ri}+V_{ri})$ 将和 $(W_{ri}-V_{ri})$ 间的显著性, 如果不显著, 说明不存在上位性, 则可按下列方程求出 D 和 H :

$$\bar{V}_r = \frac{1}{4}D + \frac{1}{4}H + E', \quad \bar{W}_r = \frac{1}{2}D + E$$

其中, $E=E_w/rp$, E_w 为抽样误差。 E' 也为抽样误差, 但它与双列杂交组合数有关。

如组合数为 p^2 (完全双列杂交), 则 V_{ri} 的抽样误差 E' 计算方式为: 计算 V_{ri} 的样本中包括 $(p-1)$ 个组合观察值 X_{ij} 、一个亲本观察值 X_{ii} ; $X_{ij}(i \neq j)$ 为正、反交组合、 r 次重复平均数, 所以 X_{ij} 的抽样误差为 $(p-1)E_w/2pr$; 而亲本平均数 X_{ii} 是 r 个区组的平均数, 其抽样误差为 E_w/pr , 故 V_{ri} 的抽样误差 E' 为 $(p-1)E_w/2p+E_w/p$, $r=(p-1)E_w/2pr$ 。

如组合数为 $p(p-1)/2$ (设计 2), 则 X_{ij} 的抽样误差为 $(p-1)E_w/pr$, 亲本平均数 X_{ii} 的抽样误差为 E_w/rp , V_{ri} 的抽样误差为 $E'=(1/rp+(p-1)/pr)E_w = E_w/r$ 。其他类型双列设计的行(或列)方差的抽样误差可以类比计算。

具体分析计算时, 可以首先求正反交组合的平均数; 然后计算 W_{ri} 和 V_{ri} ; 最后, 对 W_{ri} 依 V_{ri} 做分析, 可以 t 测验检验 W_{ri} 依 V_{ri} 的回归系数是否等于 1.0, 若不显著, 说明不存在上位性, 可以继续分析加性、显性方差; 对 W_{ri} 依 V_{ri} 的分析, 也可采用方差分析的方法, 分析 $W_{ri}+V_{ri}$ 间和 $W_{ri}-V_{ri}$ 间的显著性。若 $W_{ri}-V_{ri}$ 间差异不显著说明加性-显性模型是适合的, 若 $W_{ri}+V_{ri}$ 间差异显著, 说明显性作用存在。这里, 回归分析和 $W_{ri}-V_{ri}$ 间方差分析结果可以相互印证, 结论一致时说明结论很可靠。需指出的是, 回归分析是使用各个区组以及正、反交平均数进行的分析, 而方差分析是使用各个区组 W_{ri} 依 V_{ri} 的数据进行分析。

41.15.3 亲本及正反交 F_1 组合 Hayman 分析应用举例

从一个小麦自交群体 F_2 代(未实施选择)随机抽取 5 个纯系品系, 配制成 5×5 小麦双列杂交试验, 4 次重复, 随机区组设计, 性状为籽粒产量。现对此试验进行分析。分析时, 在 DPS 电子表格中编辑、选择数据格式如图 41-25 阴影部分所示。

	A	B	C	D	E	F
1		I	II	III	IV	
2	1X1	4.2	8.6	7.1	8.7	
3	1X2	-3.6	3.1	-1.1	4.4	
4	1X3	2.2	5.5	1.5	4.6	
5	1X4	8.4	5.2	7.4	8.1	
6	1X5	2.5	7.2	6.1	7.4	
7	2X1	0.1	1.2	6.8	2.3	
8	2X2	-6.9	-5.9	-7.6	-1.3	
9	2X3	-4.5	1.2	0.1	1.2	
10	2X4	3.5	4.8	9.3	9.5	
11	2X5	4.7	5.5	-0.9	4.7	
12	3X1	1.6	2.3	1.5	3.9	
13	3X2	-3.2	-0.5	1	3.5	
14	3X3	-2.3	0.5	-1.6	3.5	
15	3X4	5.8	8.2	7.4	7.7	
16	3X5	5.8	6.9	5.6	10.6	
17	4X1	6.6	6.4	7.7	8.8	
18	4X2	3	6.4	5.1	7.8	
19	4X3	6.6	6.3	11.2	9.5	
20	4X4	5	7.1	-0.7	7.2	
21	4X5	7.3	9.6	9.5	10.5	
22	5X1	4.4	4.6	8.3	7.6	
23	5X2	8.5	8.1	3.3	9.7	
24	5X3	8.9	7	3	7.7	
25	5X4	5.9	8.9	8.3	5.6	
26	5X5	1.4	6.3	2	7.2	
27						

图 41-25 Hayman 分析数据编辑格式

在 DPS 系统中执行“专业统计”下面的“遗传交配设计统计分析”中的“亲本及正反交 F_1 组合 Hayman 分析”。分析结果如下:

1	原始数据方差分析表				
2	变异来源	自由度	平方和	均方	F 值 p 值
3	区组	3	155.8403	51.9468	12.6873
4	处理	24	1174.1204	48.9217	11.9484 0.0001
5	误差	72	294.7972	4.0944	
6	总的	99	1624.7579		
7					

8	配合力分析							
9	均值两向列表							
10	亲本	p1	p2	p3	p4	p5	y(i)	
11	p=1	7.1500	0.7000	3.4500	7.2750	5.8000	24.3750	
12	p=2	2.6000	-5.4250	-0.5000	6.7750	3.5000	6.9500	
13	p=3	2.3250	0.2000	0.0250	7.2750	7.2250	17.0500	
14	p=4	7.3750	5.5750	8.4000	4.6500	9.2250	35.2250	
15	p=5	6.2250	7.4000	6.6500	7.1750	4.2250	31.6750	
16	y(j)=	25.6750	8.4500	18.0250	33.1500	29.9750		
17	y _i +y _j =	50.0500	15.4000	35.0750	68.3750	61.6500		
18								
19	配合力方差分析表(固定模型)							
20	变异来源	自由度	平方和	均方	F 值	p 值		
21	一般	4	181.7621	45.4405	44.3928	0.0001		
22	特殊	10	97.7655	9.7766	9.5511	0.0001		
23	正反交	10	14.0025	1.4003	1.3680	0.2127		
24	误差	72	73.6993	1.0236				
25								
26	正反交之间无显著差异, 以其均值进行分析.							
27		p1	p2	p3	p4	p5	V _r W _r	
28	p1	7.1500	1.6500	2.8875	7.3250	6.0125	6.6846 12.1323	
29	p2	1.6500	-5.4250	-0.1500	6.1750	5.4500	22.0377 19.1505	
30	p3	2.8875	-0.1500	0.0250	7.8375	6.9375	14.1008 12.4870	
31	p4	7.3250	6.1750	7.8375	4.6500	8.2000	2.0795 0.6723	
32	p5	6.0125	5.4500	6.9375	8.2000	4.2250	2.2590 1.0185	
33	平均						9.4323	9.0921
34								
35	回归方程 W _r =0.9330+0.8650V _r					r=0.92300	p=0.0254	
36	检验回归截距 a 与 0 的差异: Sa=2.5307					t=0.36868	p=0.7369	
37	检验回归系数 b 与 0 的差异: Sb=0.2082					t=4.15445	p=0.0254	
38	检验回归系数 b 与 1 的差异: Sb=0.2082					t=0.64828	p=0.5630	
39								
40	D=17.7748, H=17.4978, 平均显性度=0.9922							
41								
42	各个区组的 W _r +V _r 和 W _r -V _r 值							
43	亲本	W _r +V _r				W _r -V _r		
44	p1	27.2595	19.3678	17.9985	15.6200	4.4360	7.6788	1.8300 3.3125
45	p2	48.3900	51.7875	47.7710	28.4500	-10.7165	1.3100	-10.1975 -3.0925
46	p3	44.3520	27.1745	20.9950	17.8605	-5.0915	4.8610	-14.3110 -1.3360
47	p4	7.9633	5.3713	18.6188	-0.8975	2.5673	1.1038	-14.8912 -1.6210

48	p5	1.4665	2.3483	20.0635	0.6405	-11.5200	-1.0502	-1.7305	-0.7060
49	W _r +V _r 方差分析表								
50	变 异 来 源	平方和	自由度	均 方	F 值	p 值			
51	公共亲本间	3896.8368	4	974.2092	11.3240	0.0002			
52	公共亲本内	1290.4093	15	86.0273					
53	总 变 异	5187.2460	19						
54	W _r -V _r 方差分析表								
55	变 异 来 源	平方和	自由度	均 方	F 值	p 值			
56	公共亲本间	242.9261	4	60.7315	1.5570	0.2365			
57	公共亲本内	584.9249	15	38.9950					
58	总 变 异	827.8511	19						

分析结果中, 首先, 按两向分组进行方差分析(分析结果的 1~6 行)。可见, 基因型间差异显著, 表明可进一步分析。

其次, 去除母体效应, 将各组合子代包括正反交和重复的数据列成半双列杂交, 及正反交平均(分析结果的 9~17 行), 然后进行配合力检验(19~24 行)。计算 W_r 与 V_r 间的回归方程并作显著性测验(35~38 行), 结果表明 α 的估计值为 0.9330, β 的估计值为 0.8650, $r=0.9230^{**}$ 。假设测验, 首先可以检验回归系数是否显著, 则 $H_0: \beta=0$, 而 $t=b/Sb=4.15445$, $p=0.0254$, 说明回归关系是显著存在的。

然后, 检验回归系数是否为 1, 则 $H_0: \beta=1$, 而 $t=(b-1)/Sb=0.64828$, $p=0.563$, 说明回归系数与 1 差异不显著, 加性-显性模型适合。

因此, 上述结果表明亲本间性状遗传属于加性-显性遗传, α 与 0 无显著差异, 但 α 是正值, 说明显性作用可能在部分显性到完全显性之间。从图 41-26 中还可知道亲本 3 具较多增效基因而亲本 2 具较多减效基因。

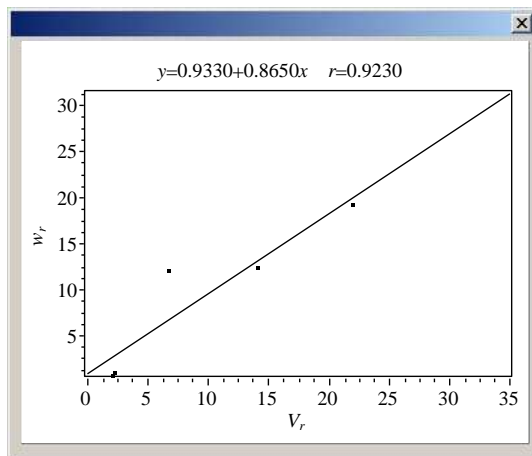


图 41-26 V_r 和 W_r 回归分析

D , H 和 $\sqrt{H/D}$ 的计算。因为分析结果表明已知上位性作用不存在, 故可通过对 V_r 和 W_r 的分析, 求出 D , H 和平均显性度(结果的 40 行), 且平均显性度接近于 1。

系统然后计算双列杂交试验各个区组的 $W_{ri}+V_{ri}$ 与 $W_{ri}-V_{ri}$ 结果(42~48 行), 并对 $W_{ri}+V_{ri}$ 与 $W_{ri}-V_{ri}$ 进行随机区组试验方差分析, 其结果见 48~58 行。结果表明: 各个共同亲本 $W_{ri}+V_{ri}$ 间差异显著, 说明显性效应是存在的, 而 $W_{ri}-V_{ri}$ 不显著, 表明不存在上位性效应, 这与回归分析的结果是一致的。

参 考 文 献

- 戴君惕, 王身立. 1989. 遗传分析方法[M]. 长沙: 湖南科学技术出版社.
- 范濂. 1983. 农业试验统计方法[M]. 郑州: 河南科学技术出版社.
- 胡秉民, 张全德. 1985. 农业试验统计分析方法[M]. 杭州: 浙江科学技术出版社.
- 黄金龙等. 1991. 电子计算机在遗传育种中的应用[M]. 北京: 中国农业出版社.
- 马育华. 1982. 试验统计[M]. 北京: 中国农业出版社.
- 毛盛远, 黄远樟. 1987. 群体遗传学及其程序设计[M]. 北京: 北京师范大学出版社.
- 莫惠栋. 1984. 农业试验统计[M]. 上海: 上海科学技术出版社.
- 裴新湖. 1987. 数量遗传与育种[M]. 上海: 上海科学技术出版社.
- 盛志康, 陈瑶生. 1999. 数量遗传学[M]. 北京: 科学出版社.
- 吴仲贤. 1979. 统计遗传学[M]. 北京: 科学出版社.
- 俞渭江, 洪振中. 1984. 种公牛选择方法研究[J]. 中国农业科学, (5): 77-83.
- 朱军. 2000. 线性模型分析原理[M]. 北京: 科学出版社.

第八篇 其他数据分析方法

从实用的观点出发,本篇向读者提供一些新近发展起来的建模技术和方法。这些方法包括模糊数学方法、神经网络分析和灰色系统分析,供读者在研究工作中参考、选用。

模糊数学方法是 20 世纪 60 年代由 Zadeh 首先提出来的。目前,模糊集合论在各个领域的应用已经十分广泛。实践证明,模糊数学在农业中用于病虫测报、种植区划、品种选育等方面,在图像识别、天气预报、地质地震、交通运输、医疗诊断、信息控制、人工智能等其他领域的应用已初见成效。目前看来,该学科具有强大的生命力和渗透力,发展前景广阔。因此,DPS 平台模糊数学部分将向读者提供一些目前已被广泛应用于解决科研生产实际问题的功能模块,包括模糊聚类、模糊识别、综合评判及综合评判逆问题的建模分析技术,作为预报和控制的有力工具。

灰色系统理论是邓聚龙先生于 20 世纪 80 年代前期提出的用于控制和预测的新技术。灰色系统理论及其方法目前已广泛应用于我国农业、社会和经济领域,并已取得显著成就。灰色关联分析同经典的因素分析方法相比较,将研究对象作为一个发展变化的系统,可对事物发展态势进行量化比较分析。目前它在农业中应用较广。灰色预测建模常用 GM(1, 1) 单序列一阶线性动态模型,有时也可建立 GM(2, 1) 模型。后者有 2 个特征根,其动态过程能反映不同的情况,即可能是单调的、非单调的或摆动的(振荡的)情况。灰色系统分析,作为状态分析模型,作者将 GM(1, h) 模型介绍给读者,它可以反映 $h-1$ 个变量对于因变量一阶导数的影响。

最后,作者介绍了目前一些常用的对多试验、多指标分析进行综合的方法,如 Meta 分析、Topsis 综合评价以及层次分析方法(AHP)等。

上述新的分析技术均可在 DPS 平台上实现。

第 42 章 模糊数学方法

在自然科学或社会科学研究中,存在着许多定义不很严格或者说具有模糊性的概念。这里所谓的模糊性,主要是指客观事物的差异在中间过渡中的不分明性,如某一生态条件对某种害虫、某种作物的存活或适应性可以评价为“有利、比较有利、不那么有利、不利”;灾害性霜冻气候对农业产量的影响程度为“较重、严重、很严重”,等等。这些通常是本来就属于模糊的概念,为处理分析这些“模糊”概念的数据,便产生了模糊集合论。

根据集合论的要求,一个对象对应于一个集合,要么属于,要么不属于,二者必居其一,且仅居其一。这样的集合论本身并无法处理具体的模糊概念。为处理这些模糊概念而进行的种种努力,催生了模糊数学。模糊数学的理论基础是模糊集。模糊集的理论是 1965 年美国自动控制专家 Zadeh 教授首先提出来的,近三十年来发展很快。

模糊集合论的提出虽然较晚,但目前在各个领域的应用十分广泛。实践证明,模糊数学在农业中主要用于病虫测报、种植区划、品种选育等方面,在图像识别、天气预报、地质地震、交通运输、医疗诊断、信息控制、人工智能等诸多领域的应用也已初见成效。从该学科发展趋势来看,它具有极其强大的生命力和渗透力。

在侧重于应用的模糊数学分析中,经常应用到聚类分析、模式识别和综合评判等方法。在 DPS 系统中,我们将模糊数学的分析方法与一般常规统计方法区别开来,列专章介绍其分析原理及系统设计的有关功能模块程序的操作要领,供用户参考和使用。

42.1 模糊聚类分析

42.1.1 模糊集的概念

对于一个普通的集合 A , 空间中任一元素 x , 要么 $x \in A$, 要么 $x \notin A$, 二者必居其一。这一特征可用一个函数 $A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$ 表示, 这里 $A(x)$ 即为集合 A 的特征函数。

将特征函数推广到模糊集, 在普通集合中只取 0, 1 两值推广到模糊集中为 $[0, 1]$ 区间。

定义 1 设 X 为全域, 若 A 为 X 上取值 $[0, 1]$ 的一个函数, 则称 A 为模糊集。

如给 5 个同学的性格稳重程度打分, 按百分制给分, 再除以 100, 这样给定了一个从域 $X = \{x_1, x_2, x_3, x_4, x_5\}$ 到 $[0, 1]$ 闭区间的映射。

x_1 : 85 分, 即 $A(x_1) = 0.85$;

x_2 : 75 分, 即 $A(x_2) = 0.75$;

x_3 : 98 分, 即 $A(x_3) = 0.98$;

x_4 : 30 分, 即 $A(x_4) = 0.30$;

x_5 : 60 分, 即 $A(x_5) = 0.60$ 。

这样确定出一个模糊子集 $A = (0.85, 0.75, 0.98, 0.30, 0.60)$ 。

定义 2 若 A 为 X 上的任一模糊集, 对任意 $0 \leq \lambda \leq 1$, 记 $A\lambda = \{x \mid x \in X, A(x) \geq \lambda\}$, 称 $A\lambda$ 为 A 的 λ 截集。

$A\lambda$ 是普通集合而不是模糊集。由于模糊集的边界是模糊的, 如果要把模糊概念转化为数学语言, 需要选取不同的置信水平 $\lambda (0 \leq \lambda \leq 1)$ 来确定其隶属关系。 λ 截集就是将模糊集转化为普通集的方法。模糊集 A 是一个具有游移边界的集合, 它随 λ 值的变小而增大, 即当 $\lambda_1 < \lambda_2$ 时, 有 $A\lambda_1 \supset A\lambda_2$ 。

定义 3 模糊集运算定义。若 A, B 为 X 上两个模糊集, 它们的和集、交集和 A 的余集都是模糊集, 其隶属函数分别定义为

$$(A \vee B)(x) = \max(A(x), B(x))$$

$$(A \wedge B)(x) = \min(A(x), B(x))$$

$$A^C(x) = 1 - A(x)$$

关于模糊集的和、交等运算, 可以推广到任意多个模糊集合中去。

定义 4 若一个矩阵元素取值为 $[0, 1]$ 区间内, 则称该矩阵为模糊矩阵。同普通矩阵一样, 有模糊单位阵, 记为 I ; 模糊零矩阵, 记为 0 ; 元素皆为 1 的矩阵用 J 表示。

定义 5 若 A 和 B 是 $n \times m$ 和 $m \times l$ 的模糊矩阵, 则它们的乘积 $C = AB$ 为 $n \times l$ 阵, 其元素为

$$C_{ij} = \bigvee_{k=1}^m (a_{ik} \wedge b_{kj}), \quad i=1, 2, \dots, n; j=1, 2, \dots, l$$

符号“ \vee ”和“ \wedge ”含意的定义为 $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$ 。模糊矩阵乘法性质包括:

$$(1) (AB)C = A(BC);$$

$$(2) AI = IA = A;$$

$$(3) A0 = 0A = 0;$$

$$(4) AJ = JA;$$

(5) 若 A, B 为模糊矩阵且 $a_{ij} \leq b_{ij}$ (一切 i, j), 则 $A \leq B$, 又若 $A \leq B$, 则 $AC \leq BC$, $CA \leq CB$ 。

42.1.2 模糊分类关系

模糊聚类是在模糊分类关系基础上进行的。由集合概念可给出如下定义:

定义 6 n 个样品的全体所组成的集合 X 作为全域, 令 $X \times Y = \{(X, Y) \mid x \in X, y \in Y\}$, 则称 $X \times Y$ 为 X 的全域乘积空间。

定义 7 设 R 为 $X \times Y$ 上的一个集合, 并且满足:

(1) 反身性: $(x_i, y_i) \in R$, 即集合中每个元素和它自己同属一类; (2) 对称性: 若 $(x, y) \in R$, 则 $(y, x) \in R$, 即集合中 (x, y) 元素同属于类 R 时, 则 (y, x) 也同属于 R ; (3) 传递性: $(x, y) \in R, (y, z) \in R$, 则有 $(x, z) \in R$ 。

上述 3 条性质称为等价关系, 满足这 3 条性质的集合 R 为一分类关系。

聚类分析的基本思想是用相似性尺度来衡量事物之间的亲疏程度，并以此来实现分类，模糊聚类分析的实质就是根据研究对象本身的属性来构造模糊矩阵，在此基础上根据一定的隶属度来确定其分类关系。

定理 1 R 为 X 上的一个分类关系的充分必要条件是：① $R_{ij} = 1$ ；② $R_{ij} = R_{ji}$ ；③ $R \times R \leq R$ 。

这三个结论等价于分类关系的反身性、对称性和传递性。

定理 2 R 为 X 上的一个模糊分类关系的充分必要条件是：对每一个 λ ($0 \leq \lambda \leq 1$) 都使 $R\lambda$ 为普通的分类关系。

这说明当模糊关系 R 确定之后，对于给定 $\lambda \in [0, 1]$ ，便可相应地得到普通分类关系 R ，即可以决定一个 λ 水平的分类。例如给定下面一个模糊分类关系：

$$R = \begin{bmatrix} 1 & 0.48 & 0.62 & 0.41 & 0.47 \\ 0.48 & 1 & 0.48 & 0.41 & 0.47 \\ 0.62 & 0.48 & 1 & 0.41 & 0.47 \\ 0.41 & 0.41 & 0.41 & 1 & 0.41 \\ 0.47 & 0.47 & 0.47 & 0.41 & 1 \end{bmatrix}$$

根据不同水平的 λ 进行分类：

(1) 当 $0.62 < \lambda \leq 1$ 时，

$$R\lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

此时共分为 5 类，即 $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$ 。

(2) 当 $0.48 < \lambda \leq 0.62$ 时，

$$R\lambda = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

此时分为 4 类，即 $\{x_1, x_3\}$, $\{x_2\}$, $\{x_4\}$, $\{x_5\}$ 。

(3) 当 $0.47 < \lambda \leq 0.62$ 时，

$$R\lambda = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

此时分为 3 类，即 $\{x_1, x_2, x_3\}$, $\{x_4\}$, $\{x_5\}$ 。

(4) 当 $0.41 < \lambda \leq 0.47$ 时,

$$R\lambda = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

此时分为 2 类, 即 $\{x_1, x_2, x_3, x_5\}, \{x_4\}$ 。

(5) 当 $0 \leq \lambda \leq 0.41$ 时, 此时 $R\lambda$ 元素全为 1, 即 5 元素合为 1 类: $\{x_1, x_2, x_3, x_4, x_5\}$ 。

42.1.3 模糊聚类

利用模糊集理论进行聚类分析的具体步骤如下:

(1) 若定义相似系数矩阵用的是定量资料, 在定义相似系数矩阵之前, 可先对原始数据进行转换, 转换方法同系统聚类分析, 可参考 26 章系统聚类分析部分。

(2) 计算模糊相似矩阵。设 U 是需要被分类对象的全体, 建立 U 上的相似系数 R , $R(i, j)$ 表示 i 与 j 之间的相似程度, 当 U 为有限集时, R 是一个矩阵, 称为相似系数矩阵。定义相似系数矩阵的工作, 原则上可以按系统聚类分析中的相似系数确定方法, 但也可以用主观评定或集体打分的办法。DPS 平台对数据集

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}_{n \times m}$$

提供了以下 8 种建立相似矩阵的方法:

$$\textcircled{1} \text{ 相关系数法: } r_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}} \quad (i, j \leq n)$$

式中

$$\bar{x}_i = \frac{1}{m} \sum_{k=1}^m x_{ik}, \quad \bar{x}_j = \frac{1}{m} \sum_{k=1}^m x_{jk}$$

$$\textcircled{2} \text{ 最大最小法: } r_{ij} = \frac{\sum_{k=1}^m \min(x_{ik}, x_{jk})}{\sum_{k=1}^m \max(x_{ik}, x_{jk})} \quad (i, j \leq n)$$

$$\textcircled{3} \text{ 算术平均最小法: } r_{ij} = \frac{\sum_{k=1}^m \min(x_{ik}, x_{jk})}{\frac{1}{2} \sum_{k=1}^m (x_{ik} + x_{jk})} \quad (i, j \leq n)$$

$$\textcircled{4} \text{ 几何平均最小法: } r_{ij} = \frac{\sum_{k=1}^m \min(x_{ik}, x_{jk})}{\sum_{k=1}^m \sqrt{x_{ik} x_{jk}}} \quad (i, j \leq n)$$

$$\textcircled{5} \text{ 绝对指数法: } r_{ij} = e^{-\sum_{k=1}^m |x_{ik} + x_{jk}|} \quad (i, j \leq 0)$$

$$\textcircled{6} \text{ 绝对值减数法: } r_{ij} = \begin{cases} 1, & i = j \\ 1 - \frac{\sum_{k=1}^m |x_{ik} - x_{jk}|}{c}, & i \neq j \end{cases}$$

式中, c 等于 $\sum_{k=1}^m |x_{ik} - x_{jk}|$ 中的最大值。

$$\textcircled{7} \text{ 夹角余弦法: } r_{ij} = \frac{\sum_{k=1}^m x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^m x_{ik}^2 \sum_{k=1}^m x_{jk}^2}} \quad (i, j \leq n)$$

$$\textcircled{8} \text{ 欧氏距离: } r_{ij} = 1 - \frac{\sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}}{\max D} \quad (i, j \leq n)$$

式中, $\max D$ 等于 $\sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$ 中的最大值。

(3) 聚类分析。用上述方法建立起来的相似关系 R , 一般只满足反射性和对称性, 不满足传递性, 因而还不是模糊等价关系。为此, 需要将 R 改造成 R^* 后得到聚类图, 在适当的阈值上进行截取, 便可得到所需要的分类。将 R 改造成 R^* , 可用求传递闭包的方法。 R 自乘的思想是按最短距离法原则, 寻求两个向量 x_i 与 x_j 的亲密度。

假设 $R^2 = (r_{ij})$, 即 $r_{ij} = \bigvee_{k=1}^n (r_{ik} \wedge r_{kj})$, 说明 x_i 与 x_j 是通过第三者 K 作为媒介而发生关系, $r_{ik} \wedge r_{kj}$ 表示 x_i 与 x_j 的关系密切程度是以 $\min(r_{ik}, r_{kj})$ 为准则, 因 k 是任意的, 故从一切 $r_{ik} \wedge r_{kj}$ 中寻求一个使 x_i 和 x_j 关系最密切的通道。 R^m 随 m 的增加, 允许连接 x_i 与 x_j 的链的边就越多。由于从 x_i 到 x_j 的一切链中, 一定存在一个使最大边长达到极小的链, 这个边长就是相当于 r_{ij}^α 。

在实际处理过程中, R 的收敛速度是比较快的。为进一步加快收敛速度, 通常采取如下处理方法:

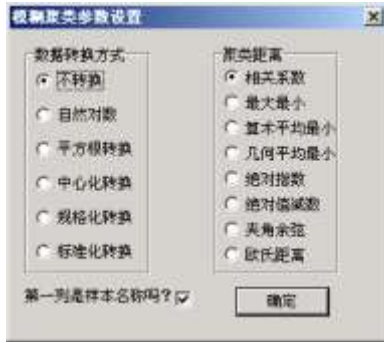
$$R \rightarrow R^2 \rightarrow R^4 \rightarrow R^8 \rightarrow \dots \rightarrow R^{2^k}$$

即先将 R 自乘改造为 R^2 , 再自乘得 R^4 , 如此继续下去, 直到某一步出现 $R^{2^k} = R^k = R^*$ 。此时 R^* 满足了传递性, 于是模糊相似矩阵 (R) 就被改造成了一个模糊等价关系矩阵 (R^*)。

(4) 模糊聚类。对满足传递性的模糊关系 R^* 进行聚类处理, 给定不同置信水平的 λ , 求 R_λ^* 阵, 找出 R^* 的 λ 显示, 得到普通的分类关系。当 $\lambda = 1$ 时, 每个样品自成一类, 随 λ 值的降低, 由细到粗逐渐归并, 最后得到模糊聚类的谱系图。

42.1.4 DPS 平台操作示例

首先在编辑状态下输入编辑数据，格式是每一行为一个样本，每一列为一个变量，然后将待分析的数据定义成数据矩阵块，在菜单方式下选择“模糊数学”→“模糊聚类”功能项，回车执行时，系统将提示用户选择数据转换方法及聚类距离计算方法的界面如下：



作出数据转换方式的选择后，系统又将提示选择建立模糊相似关系的计算方法，共有上面所述的 8 种方法可供选择。

分析输出的结果包括各个样本的联结序号、联结水平、聚类谱系图索引及在屏幕上显示聚类谱系图(拷屏可得到谱系图硬拷贝，或按 S 将图形文件以“.bmp”格式存放在盘上，然后可在 Windows 有关应用软件中调出)。

例如，根据某县气候地貌呈明显变化，即气候特征由中亚热带向北亚热带过渡、地貌由洞庭湖平原向湘西山地过渡的特点，研究稻瘟病在该县的自然生态区划中对当地生产所具有的重要意义。作者从该县 60 多个乡镇中抽取了 12 个代表性地点的地理气象资料作为样本，选取对稻瘟病流行影响比较大的 6 个自然生态因子，对稻瘟病自然生态系进行模糊聚类分析。选取样本的有关数据按图 42-1 的方式编辑和定义。

	A	B	C	D	E	F	G	H	I
1		区域	年均温	年降水量	海拔高度	山区%	丘陵%	平区%	
2		西北山区	14.50	1450	380	0.96	0.04	0	
3		西部丘陵	15.40	1500	245	0.35	0.54	0.11	
4		中部丘陵	16.10	1600	134	0	0.90	0.10	
5		西南山区	14.10	2000	510	1	0	0	
6		东部平原	16.50	1460	65	0	0.07	0.93	
7		东北部丘陵	16	1425	130	0	0.85	0.15	
8		东部丘平区	16.20	1440	110	0	0.60	0.40	
9		南部丘平区	16.20	1600	100	0	0.80	0.20	
10		西部山丘区	14.80	1500	280	0.60	0.35	0.05	
11		西部山区	14.20	1450	330	0.80	0.18	0.02	
12		南部山丘区	14.60	1600	330	0.70	0.25	0.05	
13		西南山丘区	15.20	1750	290	0.40	0.35	0.25	
14									

图 42-1 某县气候、地貌生态因子资料模糊聚类分析的数据块定义图

执行选项功能运算时,按提示选择标准化转换和采用欧氏距离计算模糊相似矩阵 R ,并对模糊相似矩阵用求传递闭包的方法加以改造,得到模糊分类关系结果如下:

数据标准化											
-0.9438	-0.6845	1.0152	1.3955	-1.1494	-0.7204						
0.0963	-0.3858	0.0221	-0.1269	0.4003	-0.2996						
0.9052	0.2116	-0.7945	-1.0004	1.5161	-0.3379						
-1.4060	2.6010	1.9716	1.4953	-1.2733	-0.7204						
1.3675	-0.6247	-1.3022	-1.0004	-1.0564	2.8370						
0.7897	-0.8338	-0.8240	-1.0004	1.3611	-0.1466						
1.0208	-0.7442	-0.9711	-1.0004	0.5863	0.8097						
1.0208	0.2116	-1.0447	-1.0004	1.2062	0.0446						
-0.5971	-0.3858	0.2796	0.4971	-0.1885	-0.5292						
-1.2905	-0.6845	0.6474	0.9962	-0.7154	-0.6439						
-0.8282	0.2116	0.6474	0.7466	-0.4985	-0.5292						
-0.1348	1.1076	0.3531	-0.0021	-0.1885	0.2359						
欧氏距离											
1.00											
0.85	1.00										
0.57	0.92	1.00									
0.75	0.57	0.31	1.00								
0.39	0.66	0.63	0.00	1.00							
0.61	0.93	0.98	0.21	0.68	1.00						
0.60	0.92	0.93	0.21	0.85	0.97	1.00					
0.56	0.92	0.99	0.30	0.71	0.97	0.96	1.00				
0.95	0.97	0.81	0.69	0.57	0.83	0.81	0.81	1.00			
0.99	0.89	0.65	0.73	0.43	0.68	0.67	0.64	0.97	1.00		
0.96	0.93	0.74	0.81	0.50	0.74	0.73	0.74	0.99	0.98	1.00	
0.83	0.94	0.84	0.77	0.65	0.80	0.82	0.85	0.93	0.86	0.95	1.00
模糊等价关系矩阵											
1.00											
0.97	1.00										
0.93	0.93	1.00									
0.81	0.81	0.81	1.00								
0.85	0.85	0.85	0.81	1.00							
0.93	0.93	0.98	0.81	0.85	1.00						
0.93	0.93	0.97	0.81	0.85	0.97	1.00					
0.93	0.93	0.99	0.81	0.85	0.98	0.97	1.00				
0.98	0.97	0.93	0.81	0.85	0.93	0.93	0.93	1.00			
0.99	0.97	0.93	0.81	0.85	0.93	0.93	0.93	0.98	1.00		
0.98	0.97	0.93	0.81	0.85	0.93	0.93	0.93	0.99	0.98	1.00	
0.95	0.95	0.93	0.81	0.85	0.93	0.93	0.93	0.95	0.95	0.95	1.00
不同水平下的联结情况											
	I = 3		J = 8		Mx = 0.993333						
	I = 1		J = 10		Mx = 0.987243						
	I = 9		J = 11		Mx = 0.985257						
	I = 1		J = 9		Mx = 0.976143						
	I = 3		J = 6		Mx = 0.975546						
	I = 1		J = 2		Mx = 0.972023						
	I = 3		J = 7		Mx = 0.966531						
	I = 1		J = 12		Mx = 0.945291						
	I = 1		J = 3		Mx = 0.934930						

I = 1	J = 5	Mx = 0.852289
I = 1	J = 4	Mx = 0.811624
谱系图联结序号： 4 5 7 6 3 8 12 2 9 11 1 10		

在模糊分类关系基础上，给定不同置信水平 λ ，求 λ 截阵 $R\lambda$ ，找出 R 的 λ 显示，逐渐归并，得到动态性的聚类分析结果，可绘成聚类谱系图(图 42-2)。

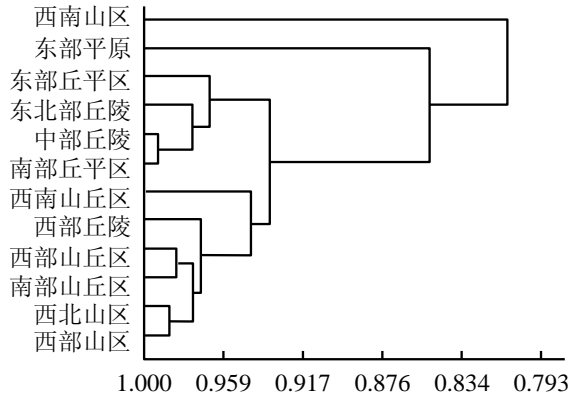


图 42-2 模糊聚类分析的动态谱系图

根据上述聚类谱系图，取 $\lambda = 0.935$ ，可将 12 个样本分成以下 4 类：

第 1 类即类{4}，为稻瘟病山区生态系。气候特点是雨水多、湿度大、雾多日照少；地理特点是平均海拔高、多高山、山阴冷浸田多。此类地区的自然生态因子有利于稻瘟病的流行，为常年重灾区。

第 2 类即类{1, 2, 9, 10, 11, 12}，为稻瘟病山丘区生态系。该区属山区到丘陵区的过渡地带，地理气候特点类似山区，但程度上有差异，为稻瘟病发生流行程度较重区。

第 3 类即类{3, 6, 7, 8}，为稻瘟病丘平区生态系。该区属丘陵到平坝的过渡地带，地理气候特点是雨水较少、湿度较低、日照较多，一般年份稻瘟病流行程度不重，但多雨年份亦可流行成灾，且局部地区发病较重，故属于局部流行区。

第 4 类即类{5}，为稻瘟病滨湖平原生态系。该区属洞庭湖平原区，地理气候特点是降雨少、湿度小、温度高，常出现高温干燥天气，对稻瘟病发病流行不利，但多雨年份也有局部丘块稻瘟病发生危害的可能，属偶发区。

根据模糊聚类分析结果，可分区制定稻瘟病的综合治理对策，为农业生产提供依据。

42.2 模糊模式识别

42.2.1 方法简介

“模式”一词来源于英文 pattern，原意是典范、式样、样品，在不同场合有其不同的含义。在此讲的模式是指具有一定结构的信息集合。

模式识别就是识别给定的事物以及与它相同或类似的事物，也可以理解为模式的分类，即把样品分成若干类，判断给定事物属于哪一类，这与前面介绍的判别分析很相似。

模式识别的方法大致可以分为 2 种, 即根据最大隶属原则进行识别的直接法和根据择近原则进行归类的间接法, 分别简介如下:

(1) 若已知 n 个类型在被识别的全体对象 U 上的隶属函数, 则可按隶属原则进行归类。此处介绍的是针对正态型模糊集的情形。对于正态型模糊变量 x , 其隶属度为 $A(x) = e^{-[(x-a)/b]^2}$, 其中 a 为均值, $b^2 = 2\sigma^2$, σ^2 为相应的方差。按泰勒级数展开, 取近似值得

$$A(x) = \begin{cases} 1 - \left(\frac{x-a}{b}\right)^2, & x-a < b \\ 0, & x-a > b \end{cases}$$

若有 n 种类型 m 个指标的情形, 则第 i 种类型在第 j 种指标上的隶属函数是

$$A_{ij}(x) = \begin{cases} 0, & x \leq a_{ij}^{(1)} - b_{ij} \\ 1 - \left(\frac{x - a_{ij}^{(1)}}{b_{ij}}\right)^2, & a_{ij}^{(1)} - b_{ij} < x < a_{ij}^{(1)} \\ 1, & a_{ij}^{(1)} \leq x \leq a_{ij}^{(2)} \\ 1 - \left(\frac{x - a_{ij}^{(2)}}{b_{ij}}\right)^2, & a_{ij}^{(2)} < x < a_{ij}^{(2)} + b_{ij} \\ 0, & a_{ij}^{(2)} + b_{ij} < x \end{cases}$$

其中, $a_{ij}^{(1)}$ 和 $a_{ij}^{(2)}$ 分别是第 i 类元素第 j 种指标的最小值和最大值, $b_{ij}^2 = 2\sigma_{ij}^2$, 而 σ_{ij}^2 是第 i 类元素第 j 种指标的方差。

(2) 若有 n 种类型 (A_1, A_2, \dots, A_n) , 每类都有 m 个指标, 且均为正态型模糊变量, 相应的参数分别为 $a_{ij}^{(1)}, a_{ij}^{(2)}, b_{ij}$ ($i=1, 2, \dots, n; j=1, 2, \dots, m$)。其中, $a_{ij}^{(1)} = \min(x_{ij})$, $a_{ij}^{(2)} = \max(x_{ij})$, $b_{ij}^2 = 2\sigma_{ij}^2$, 而 σ_{ij}^2 是 x_{ij} 的方差。待判别对象 B 的 m 个指标分别具有参数 a_j, b_j ($j=1, 2, \dots, m$), 且为正态型模糊变量, 则 B 与各个类型的贴近度为

$$(A_{ij}, B) = \begin{cases} 0, & a_j \leq a_{ij}^{(1)} - b_{ij} \\ 1 - \frac{1}{2} \left(\frac{a_j - a_{ij}^{(1)}}{b_j + b_{ij}}\right)^2, & a_{ij}^{(1)} - b_{ij} < a_j < a_{ij}^{(1)} \\ 1, & a_{ij}^{(1)} \leq a_j < a_{ij}^{(2)} \\ 1 - \frac{1}{2} \left(\frac{a_j - a_{ij}^{(2)}}{b_j + b_{ij}}\right)^2, & a_{ij}^{(2)} \leq a_j < a_{ij}^{(2)} + (b_j + b_{ij}) \\ 0, & a_{ij}^{(2)} + (b_j + b_{ij}) \leq a_j \end{cases}$$

记 $S_i = \min_{1 \leq j \leq m} (A_{ij}, B)$, 又有 $S_{i0} = \min_{1 \leq j \leq n} (S_i)$, 按贴近原则可认为 B 与 A_{i0} 最贴近。

根据如上介绍, DPS 系统中设计了两个功能模块: 一是根据在集合上的隶属函数,

按隶属原则识别对象，判定样本的类别归属；二是根据模糊集两两之间的贴进度，按接近原则，确定出最接近的两个模糊集。

42.2.2 DPS 平台的操作示例

系统规定数据输入的格式是一行一个样本，一列一个变量。最右边的一列为样本的已知类别(如 1, 2, ...)(注意每一类中至少要有 3 个样本)。对于待判别的样本，其分类类别用 0 表示。所有待分析数据(连同类别一起)需定义成数据块，然后进入菜单操作，选择“模糊数学”→“模糊识别”功能项，回车执行后即可输出分析结果。输出结果包括各类参数(变量名、最小值、最大值、标准差和参数 *B*)和各待判样本归类结果(样本序号、对各类贴进度的最大值、最贴进的类号)。

	A	B	C	D	E	F	G	H	I	J
1	序	年	幼虫发生量		发生期		增殖系数		类别	
2	号	份	第二代	第三代	第二代	第三代	二至三代	三至四代		
3	1	1962	344	3333	29	9	1	1962	344	
4	2	1963	121	1497	27	19	2	1963	121	
5	3	1964	187	1813	32	18	3	1964	187	
6	4	1965	500	4000	34	14	4	1965	500	
7	5	1966	441	3750	36	14	5	1966	441	
8	6	1967	404	4600	33	16	6	1967	404	
9	7	1968	328	986	35	18	7	1968	328	
10	8	1969	806	1790	32	15	8	1969	806	
11	9	1970	730	1970	36	20	9	1970	730	
12	10	1971	263	333	29	15	10	1971	263	
13	11	1972	486	600	32	19	11	1972	486	
14	12	1973	248	585	33	20	12	1973	248	
15	13	1974	2100	2700	22	14	13	1974	2100	
16	14	1975	333	287	38	19	14	1975	333	
17	15	1976	90	77	40	24	15	1976	90	
18	16	1977	19	25	40	27	16	1977	19	
19	17	1978	230	2525	39	20	17	1978	230	
20	18	1979	1392	1041	33	18	18	1979	1392	
21	19	1980	308	41	31	28	19	1980	308	
22	20	1981	415	916	36	18	20	1981	415	
23	21	1982	34	401	38	29	21	1982	34	
24	22	1983	267	803	37	26	22	1983	267	
25	23	1984	1043	3500	39	26	23	1984	1043	
26	24	1985	2243	7452	31	20	24	1985	2243	
27	25	1986	236	599	35	26	25	1986	236	
28	26	1987	558	1061	33	24	26	1987	558	
29	27	1988	162	2817	34	21	27	1988	162	
30	28	1989	760	877	39	33	28	1989	760	
31	29	1990	458	199	35	27	29	1990	458	
32										

图 42-3 某县三化螟模糊识别分析的数据编辑定义图

例如, 在“有序样本最优分割”一节中, 将历年三化螟发生动态用最优分割分成 3 类, 即将三化螟种群消长过程划分为猖獗—缓和—猖獗三个阶段, 这样划分结果与该县历年水稻种植制度(一季中稻为主→纯双季稻→单双季混栽)变化是相吻合的。为识别 1988 年之后三化螟发生动态, 也可用模糊识别方法进行分析。现将待识别数据和原来的历史资料按图 42-3 方式整理编辑和定义。完成数据编辑定义之后, 执行选项功能“模糊识别”, 便可得到如下结果:

第 1 类				
变量名	最小值	最大值	标准差	参数 B
X(1)	121.000000	500.000000	148.979746	210.689178
X(2)	1497.000000	4600.000000	1243.947346	1759.207208
X(3)	27.000000	36.000000	3.311596	4.683304
X(4)	9.000000	19.000000	3.577709	5.059644
X(5)	8.000000	12.370000	1.671208	2.363445
X(6)	1.060000	1.910000	0.341155	0.482466
第 2 类				
变量名	最小值	最大值	标准差	参数 B
X(1)	19.000000	2100.000000	587.039979	830.199901
X(2)	25.000000	2700.000000	948.248286	1341.025587
X(3)	22.000000	40.000000	5.047213	7.137837
X(4)	14.000000	28.000000	4.323460	6.114296
X(5)	0.130000	10.960000	2.754776	3.895841
X(6)	0.550000	4.170000	1.092898	1.545592
第 3 类				
变量名	最小值	最大值	准差	参数 B
X(1)	34.000000	2243.000000	725.470408	1025.970090
X(2)	401.000000	7452.000000	2400.924093	3395.419414
X(3)	31.000000	39.000000	2.669270	3.774917
X(4)	18.000000	29.000000	3.732100	5.277987
X(5)	1.900000	11.790000	3.250108	4.596346
X(6)	0.000000	1.090000	0.462817	0.654523
各个待判样本的归类结果				
样本序号	对各类贴近度的最大值		最贴近的类号	
1	0.42564		3	
2	0.89772		3	

从分析结果可以看出, 1989 年和 1990 年三化螟发生动态仍和前几年相似, 表明农业生态系统是相对稳定的。

42.3 模糊相似优先比方法

42.3.1 方法简介

相似优先比是模糊性度量的一种形式,它是成对的样本与一个固定的样本作比较,确定哪一个与固定样本更相似,从而选择与固定样本相似程度较大者。

假定样本 x_i 和 x_j 与固定样本 x_k 比较,其相似优先比 R_{ij} 必须满足如下要求:

- (1) 若 R_{ij} 在 $[0.5, 1.0]$ 之间,则表示 x_i 比 x_j 优先;
- (2) 若 R_{ij} 在 $[0.0, 0.5]$ 之间,则表示 x_j 比 x_i 优先;
- (3) 在极值情形下有 3 种可能:如 $R_{ij} = 1$,则表示 x_i 比 x_j 显然优先;如 $R_{ij} = 0$,则表示 x_j 比 x_i 显然优先;如 $R_{ij} = 0.5$,则 x_i 和 x_j 不分伯仲,优先无法确定。

在模糊优先比分析中,一般采用海明距离作为相似优先比中 R_{ij} 的测度。如对样本 x_i 和样本 x_j 与固定样本 x_k 之间进行比较,海明距离可定义为

$$r_{ij} = \frac{d_{ki}}{d_{ki} + d_{kj}}$$

$$R_{ji} = 1 - R_{ij}$$

式中, $d_{ki} = |x_k - x_i|$, $d_{kj} = |x_k - x_j|$, 接下来,对给定的一样本集合 $X = \{x_1, x_2, \dots, x_n\}$ 和固定样本 x_k ,令任意 $x_i, x_j \in X$ 和 x_k 作比较,即计算两两样本间的相似优先比,从而得到模糊相关矩阵

$$R = (r_{ij}) \begin{cases} r_{ij} \in [0, 1] \\ i, j = 1, 2, \dots, n \end{cases}$$

建立模糊相似矩阵之后,由 λ 水平集选出相似样本,即在相似矩阵中,从大到小地选定 λ 值,以在 λ 值下降过程中首先到达的除主对角线元素外全行都为 1 的那一行的样本最相似,然后删除矩阵相应的行和列,并降低 λ 水平值,继续寻找。依此类推,直至截距处理完毕。

一般情形下,若每个样本有 m 个因素,则对每一因素都有一个模糊相似矩阵,所以,每一样本的每一因素都将产生一个反映相似程度的序号值,最后将每一样本各个因素的序号值相加,其结果便是该样本与固定样本相似程度的综合反映。

样本的序号值越小,该样本与固定样本就越相似,但严格地说,各个因素对样本的影响程度是不一样的,因此有必要给各个因素赋予一定的权重,这样得到的结果将更符合实际情况。所以,当用户在对有关因素影响的轻重程度有比较大的把握,或在分析中需突出某个因素时,可对各个因素进行加权处理以达到更好的分析效果。

42.3.2 DPS 平台的操作示例

数据的输入编辑格式是每一行为一个样本,每一列为一个变量,最右边的一列为已

知样本的代码(用 1 表示)和待识别样本的代码(用 0 表示)。数据块的第一行可以放入各个因子的权重系数(如图 42-4)。

在菜单下选择“模糊数学”→“相似优先比分析”，执行该项功能后系统将输出分析结果。结果包括待识别样本与各样本间的海明距离以及待识别样本与其他样本各个因素的模糊优先比矩阵 R ，最后给出待判样品对各已知样品各变量相似程度和待判样品对各已知样品的优先比值，并按顺序排列。

例如，高素华(1981)对日本柑橘主要产地之一福岡和我国合肥、武汉、长沙、桂林、温州和成都等 7 地柑橘生长的农业气候相似程度进行了分析，选用各地年均温、年降水量、年日照时数、年极端最低气温和 1 月均温作为相似因子。现运用模糊相似优先比方法在 DPS 平台上进行分析。其数据输入、编辑整理和数据块的定义如图 42-4 所示，

	A	B	C	D	E	F	G	H
1	地点	年均温	年降水量	年日照时数	年最低气温	1 月均温	识别标识	
2	合肥	15.7	970	2309	-20.6	1.9	1	
3	武汉	16.3	1260	2085	-17.3	2.8	1	
4	上海	15.7	1129	2039	-9.4	3.3	1	
5	长沙	17.2	1422	1726	-9.5	4.6	1	
6	桂林	18.8	1874	1709	-4.9	8.0	1	
7	温州	17.9	1698	1846	-4.5	7.5	1	
8	成都	16.3	976	1239	-4.6	5.6	1	
9	日本福岡	16.2	1492	2000	-8.2	6.2	0	
10								

图 42-4 模糊优先比分析的数据编辑定义图

在执行运算时，系统会提示用户选择是否输出相似程度矩阵(本例下面结果为不输出)，分析结果如下：

各因子的权重系数分别为:					
年均温	0.2000				
年降水量	0.2000				
年日照时数	0.2000				
年最低气温	0.2000				
1 月均温	0.2000				
待报样本 1 与各个样本间绝对值:					
合肥	0.5000	522.0000	309.0000	12.4000	4.3000
武汉	0.1000	232.0000	85.0000	9.1000	3.4000
上海	0.5000	363.0000	39.0000	1.2000	2.9000
长沙	1.0000	70.0000	274.0000	1.3000	1.6000
桂林	2.6000	382.0000	291.0000	3.3000	1.8000
温州	1.7000	206.0000	154.0000	3.7000	1.3000
成都	0.1000	516.0000	761.0000	3.6000	0.6000

待判样品 x_1 对各个已知样品各变量相似程度:

样本	年均温	年降水量	年日照时数	年最低气温	1月均温	相似程度
合肥	2	7	6	7	7	5.8000
武汉	1	3	2	6	6	3.6000
上海	2	4	1	1	5	2.6000
长沙	3	1	4	2	3	2.6000
桂林	5	5	5	3	4	4.4000
温州	4	2	3	5	2	3.2000
成都	1	6	7	4	1	3.8000

预测样本与该样本最接近.

从分析结果来看, 上海、长沙与福冈柑橘生产的农业气候条件最相似, 而合肥的相似程度最小。

42.4 模糊综合评判

42.4.1 方法简介

综合评判就是对受到多个因素制约的事物或对象作出一个总的评价, 这是在日常生活和科研工作中经常遇到的问题, 如产品质量评定、科技成果鉴定、某种作物种植适应性的评价等, 都属于综合评判问题。由于从多方面对事物进行评价难免带有模糊性和主观性, 采用模糊数学的方法进行综合评判将使结果尽量客观, 从而取得更好的实际效果。

模糊综合评判的数学模型可分为一级模型和多级模型, 在此仅介绍一级模型。采用一级模型进行综合评判, 一般可归纳为以下几个步骤:

(1) 建立评判对象因素集 $U = \{u_1, u_2, \dots, u_n\}$ 。因素就是对象的各种属性或性能, 在不同场合, 也称为参数指标或质量指标, 它们能综合地反映出对象的质量, 因而可由这些因素来评价对象。

(2) 建立评判集 $V = \{v_1, v_2, \dots, v_m\}$ 。如工业产品的评价, 评判集是等级的集合; 农作物种植区域适应性的评价, 评判集是适应程度的集合。

(3) 建立单因素评判, 即建立一个从 U 到 $F(V)$ 的模糊映射

$$f: U \rightarrow F(V), \quad \forall u_i \in U$$

$$u_i \mapsto f(u_i) = \frac{r_{i1}}{v_1} + \frac{r_{i2}}{v_2} + \Lambda + \frac{r_{im}}{v_m}$$

$$0 \leq r_{ij} \leq 1, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m,$$

由 f 可以诱导出模糊关系, 得到模糊矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}$$

称 R 为单因素评判矩阵, 于是 (U, V, R) 构成了一个综合评判模型。

(4) 综合评判。由于对 U 中各个因素有不同的侧重, 需要对每个因素赋予不同的权重, 它可表示为 U 上的一个模糊子集 $A = (a_1, a_2, \dots, a_n)$, 且规定 $\sum_{i=1}^n a_i = 1$ 。

在 R 与 A 求出之后, 则综合评判模型为 $B = A \circ R$ 。记 $B = (b_1, b_2, \dots, b_m)$, 它是 V 上的一个模糊子集, 其中

$$b_j = \bigvee_{i=1}^n (a_i \wedge r_{ij}) \quad (j = 1, 2, \dots, m)$$

如果评判结果 $\sum_{j=1}^m b_j \neq 1$, 就对其结果进行归一化处理。

从上述模糊综合评判的 4 个步骤可以看出, 建立单因素评判矩阵 R 和确定权重分配 A 是两项关键性的工作, 但同时又没有统一的格式可以遵循, 一般可采用统计实验或专家评分的方法求出。

但当被评价的对象中, 如果因素较多, 这时会在“取小”、“取大”的运算过程中, 筛选掉过多的信息, 甚至使评判结果完全失去作用。这是因为当因素较多时, 因子权重的和为 1 的条件使得每个因素的权重变得很小, 这样在“取小”运算时就被取上了。这时, 一般的做法是应用多级模型来解决这个问题。

经作者对许多事例进行分析发现, 可采取适当的方法对因素权重进行调整, 这里, 作者对因素个数 n 大于评判集合的因子数 m 时, 将各个因素的权重 a_i 按下式进行调整:

$$a'_i = a_i \cdot n/m$$

这时综合评判模型 $B = A \circ R$ 的运算公式为

$$b_j = \bigvee_{i=1}^n (a'_i \wedge r_{ij}) \quad (j = 1, 2, \dots, m)$$

在 DPS 平台上, 只是根据给出的评判矩阵 R 和确定的权重分配 A 进行综合评判处理。但是, 用户可以在平台上利用公式计算, 根据定义公式, 计算其隶属函数。

42.4.2 DPS 平台的操作示例

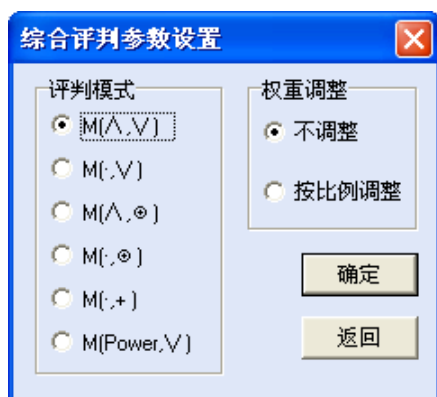
根据给定的评估数据(评判矩阵 R)和最终的评价结果(权重分配 A), 确定出该类评估近似的权分配系数。

例1 对某教师讲课质量进行综合评定。已知因素集合 U 为: $U = \{\text{教材熟练, 逻辑性强, 启发性强, 语言生动, 板书整齐}\}$, 评语集合 $V = \{\text{很好, 较好, 一般, 不好}\}$ 。首先, 建立评估集合(评判矩阵), 即评价矩阵数据编辑和定义。在 DPS 平台上, 以行代表因素集, 列代表评语集。每一行为因素集 U 中某一单因素的评价结果, 最后一列存放最终评价价值(建立权重分配), 本例的权重分配为 $A = (0.3, 0.2, 0.2, 0.2, 0.1)$ 并放在数据块的最后一列。最后, 评价矩阵数据块可按图 42-5 格式输入在分析前选中。

	A	B	C	D	E	F	G
1	因素\评语	很好	较好	一般	差	因子权重	
2	教材熟练	0.45	0.25	0.2	0.1	0.3	
3	逻辑性强	0.5	0.4	0.1	0	0.2	
4	启发性强	0.3	0.4	0.2	0.1	0.2	
5	语言生动	0.4	0.4	0.1	0.1	0.2	
6	板书整齐	0.3	0.5	0.1	0.1	0.1	
7							

图 42-5 模糊综合评判的数据编辑定义图

完成评估数据(评判矩阵 R)和最终的评价结果(权重分配 A)数据编辑及选中之后,在菜单下选择“模糊数学”→“模糊综合评判分析”功能项。系统显示如下综合评判参数设置界面。



用户界面的左边是综合评判的不同模糊运算组合模式。右边是确定权重是否调整?当被评价对象中因素较多时,建议进行调整。本例因素不是很多,可不调整。设置好参数后,点击“确定”按钮后即得分析结果。

评判结果表明,对该教师的课堂教学认为“很好”的占 40.54%，“较好”的占 27.03%，“一般”的占 21.62%，“不好”的占 10.81%。根据最大隶属原则,结论是“很好”。

例 2 假定有一种工业品,它的质量由 9 个指标 X_1, X_2, \dots, X_9 确定,产品的级别分为 1 级、2 级、等外和废品。现请有关专家、检验人员、用户等组成一个单因素评判小组,得到单因素评判矩阵及各个指标重要性的权重如图 42-6 上部。

完成评估数据(评判矩阵 R)和最终的评价结果(权重分配 A)数据编辑及选中之后,在菜单下选择“模糊数学”→“模糊综合评判分析”功能项。系统显示综合评判参数设置界面。这时,如果按上例模式,权重不调整,因为因素指标较多,每个因素的权重较小,经先“取小”后“取大”的运算之后得到的结果是

$$b = a \circ R = (0.2500, 0.2500, 0.2500, 0.2500)$$

这显然不是应有的结论。

	A	B	C	D	E	F	G
1	因素	1级	2级	等外	废品	因素权重	
2	x1	0.36	0.24	0.13	0.27	0.10	
3	x2	0.20	0.32	0.25	0.23	0.12	
4	x3	0.40	0.22	0.26	0.12	0.07	
5	x4	0.30	0.28	0.24	0.18	0.07	
6	x5	0.26	0.36	0.12	0.20	0.16	
7	x6	0.22	0.42	0.16	0.10	0.10	
8	x7	0.38	0.24	0.08	0.20	0.10	
9	x8	0.34	0.25	0.30	0.11	0.10	
10	x9	0.24	0.28	0.30	0.18	0.18	
11							

图 42-6 产品质量评价数据编辑与定义

这时，我们可以在 DPS 处理平台上，对权重系数进行调整，即将图 42-6 权重调整选项选定为“按比例调整”，应用调整后的权重进行运算。



通过计算，可得到结果如下：

$$b = (0.2645, 0.2975, 0.2479, 0.1901)$$

根据最大隶属原则，该产品属于 2 等品。这与葛苏林将 9 个指标分为 3 类，采用 2 级综合评判结果

$$b = (0.26087, 0.304348, 0.26087, 0.173913)$$

是一致的。

42.5 模糊关系方程求解

42.5.1 方法简介

模糊关系方程是模糊数学的一个重要组成部分。如下所示，它其实是模糊综合评判的逆问题。



这类问题具有普遍实际意义。如一些老专家、老中医或具有丰富经验的实际工作人员，在他们的头脑里，经验和技能常归结于对诸因素有一种优越的权数分配方案，这些难以言传的经验和技能，可望利用模糊数学原理，借助计算机技术进行模拟并保存下来。

在 DPS 系统中，求解模糊关系方程采用徐罗曹李方法(汪培庄，1983)。其求解过程首先考虑 $X \circ R=B$ 类型的模糊关系方程

$$(x_1, x_2, \dots, x_n) \circ \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nm} \end{bmatrix} = (b_1, b_2, \dots, b_m)$$

按照模糊合成运算的最大-最小法则，上式可化为下面一组线性等式，称为模糊线性方程：

$$\begin{cases} (r_{11} \wedge x_1) \vee (r_{21} \wedge x_2) \vee \dots \vee (r_{n1} \wedge x_n) = b_1 \\ (r_{12} \wedge x_1) \vee (r_{22} \wedge x_2) \vee \dots \vee (r_{n2} \wedge x_n) = b_2 \\ \dots \dots \dots \\ (r_{1m} \wedge x_1) \vee (r_{2m} \wedge x_2) \vee \dots \vee (r_{nm} \wedge x_n) = b_m \end{cases}$$

因此，求解模糊关系方程实际上就是对上述方程组的求解，其步骤简述如下：

(1) 标准化排列。将向量 B 写于 A 的上方，按 $\bar{b}_1 \geq \bar{b}_2 \geq \dots \geq \bar{b}_m$ 的次序更换向量 B 的分量，同时矩阵 R 也作相应的更换，由此得到矩阵的标准化排列。

(2) 上铤。按上述规则更换后的矩阵，对每一个 $j(j=1, 2, \dots, m)$ ，用 \bar{b}_j 的上铤第 j 列，即若 $\bar{r}_{ij} > \bar{b}_j$ ，将 \bar{r}_{ij} 变成 \bar{b}_j ；若 $\bar{r}_{ij} \leq \bar{b}_j$ ，将 \bar{r}_{ij} 变成空元 ϕ 。从而得到新的 r_{ij} ，用 r'_{ij} 表示

$$r'_{ij} = \begin{cases} \bar{b}_j, & \bar{r}_{ij} > \bar{b}_j \\ \phi, & \bar{r}_{ij} \leq \bar{b}_j \end{cases}$$

(3) 求下确界。对上铤后所得矩阵的各行求下确界，得向量

$$UR = (ur_1, ur_2, \dots, ur_n)$$

并定义空元的下确界为 1。

(4) 平铣。将上铣出来的矩阵称为平铣矩阵，即对每一个元素 j 分别用 \bar{b}_j 平铣 \bar{R} 的第 j 列：若 $\bar{r}_{ij} \geq \bar{b}_j$ ，将 \bar{r}_{ij} 变成 \bar{b}_j ；若 $\bar{r}_{ij} < \bar{b}_j$ ，将 \bar{r}_{ij} 变成空元 ϕ 。这样得到新的 r_{ij} ，以 r_{ij} 表示

$$r_{ij} = \begin{cases} \bar{b}_j, & \bar{r}_{ij} \geq \bar{b}_j \\ \phi, & \bar{r}_{ij} < \bar{b}_j \end{cases}$$

(5) 划元。对平铣后的矩阵，逐行划去该行大于上界即第 i 行大于 uR_i 的元素。

(6) 判别。若原方程有解，则上一步所得矩阵的每一列都有未被划去的非空白元素。

(7) 求解。从经过划元后所得的矩阵中的每一列，选定一个非空白且未被划元的元素，对这些当选元素逐行取上确界，并规定空集 \emptyset 的上确界为 0。这样得到的一组解为“拟极小解”。当被选元素不同时，可得到多个“拟极小解”。

最后得到的解之所以称为“拟极小解”，因为它们不一定是极小解，它们相互之间还可能存在重合或优劣的模糊关系，需进一步进行筛选。实际上，当方程经判别有解时，第 3 步所求的上界就是方程的最大解。

应用 DPS 平台求解模糊关系方程时，当模糊方程有解时，系统将给出最大解和最小解，如模糊方程无解时则提示没有解。

42.5.2 DPS 平台的操作示例

在平台上求解模糊关系方程时，其数据编辑格式是：行为因素集，列为评语集，每一行为因素集 U 中某一单因素的评价结果，最后一行存放整个因素集 U 的综合评价结果，然后将数据定义成数据块。如对某单位的管理工作进行“民意测验”，得综合评价矩阵和最终评价结果，并按系统规定格式编辑整理数据和定义数据块，如图 42-7。

	A	B	C	D	E	F
1		很好	较好	一般	差	
2	政治思想工作(X1)	0.04	0.30	0.40	0.26	评价 矩阵
3	精神文明建设(X2)	0.24	0.28	0.28	0.20	
4	行政及其管理(X3)	0.10	0.34	0.40	0.16	
5	职工组织纪律(X4)	0.02	0.26	0.40	0.32	
6	人才培训工作(X5)	0.13	0.28	0.42	0.17	
7	学术科研工作(X6)	0.14	0.27	0.30	0.29	
8	后勤管理工作(X7)	0.03	0.14	0.42	0.41	
9	整个因素集 U 综合评价	0.15	0.29	0.34	0.22	
10						

图 42-7 求解模糊关系方程的数据编辑、定义示意图

然后，在菜单下选择“模糊数学”→“模糊关系方程”功能项，回车执行后即得到归一化之后的分析结果，详见于后。

输出结果			
最小解	X_i	最大解	归一处理后结果
	0.2200 <-- $x(1)$ -->	0.2200	0.1325
	0.1500 <-- $x(2)$ -->	0.1500	0.0904
	0.2900 <-- $x(3)$ -->	0.2900	0.1747
	0.0000 <-- $x(4)$ -->	0.2200	0.1325
	0.3400 <-- $x(5)$ -->	0.3400	0.2048
	0.0000 <-- $x(6)$ -->	0.2200	0.1325
	0.0000 <-- $x(7)$ -->	0.2200	0.1325

42.6 综合评判逆问题

42.6.1 方法简介

前面的模糊综合评判和模糊关系求解是综合评判的正反两个方面。由于权重分配 A 的确定并无通用公式，所以它的正确与否往往只能取决于专家的判断或经验，而这些又是很难用数学公式表达出来的。与之相反，权重分配 B 可以通过实践的检验建立起来。由 R 和 B 反过来求 A ，有利于总结专家的经验，使它得到量化。同样道理，由 A 和 R 而求 R 可以帮助我们检验建立起来的数学模型是否合适。从这个角度来看，综合评判的逆问题比其正问题更有意义。

上面介绍的模糊关系方程求解，即求解综合评判逆问题时，其方程可能有解，也可能无解。有解时，解也可能有多个，这需根据实际情况作出恰当选择。若无解，又应当怎么办呢？DPS 系统提供根据贴近原则解模糊关系方程的方法。

首先选择一些备择解的集合，设它们为 $J=\{A_1, A_2, \dots, A_n\}$ ，然后用贴近度原则在备择解集 $\{A_1, A_2, \dots, A_n\}$ 中选取一个 A_i 作为模糊关系方程 $X \circ R = B$ 中的 X 。将 A_i 与

$$\text{进行合成: } \begin{cases} A_1 \circ R = B_1 \\ A_2 \circ R = B_2 \\ \vdots \\ A_n \circ R = B_n \end{cases}$$

求出 A 之后，再用贴近度公式 $(A_i, B) = \frac{1}{2} [A_i, B + (1 - A_i \odot B)]$ 分别求出 B 与 A_1, A_2, \dots, A_n 的贴近度，选最大者作为近似的 X^* ，将它作为方程的近似解。

42.6.2 DPS 平台操作示例

根据给定的评估数据(评判矩阵 R 、最终的评价结果 B 和备择权重方案 A ，给出与预定若干个权重数分配方案的贴近程度，从而确定出该类评估近似的权分配系数。这是一种近似求解模糊关系方程的方法，也可用来评价某一类量对于某些标准定量的贴近程度。

(1) 评估集合。评价矩阵的数据编辑和定义格式是行为因素集，列为评语集。每一行为因素集 U 中某一单因素的评价结果，最后一行存放整个因素集 U 的综合评价结果，再将数据定义成数据块。例如，对某品牌的自行车进行评价，已知因素集合 $U=\{\text{外型, 质量, 价格}\}$ ，评语集合 $V=\{\text{很好, 较好, 一般, 较差}\}$ ，其评价矩阵数据可

按图 42-8 格式定义。

	A	B	C	D	E	F
1	因素\评语	很好	较好	一般	差	
2	外型	0.2	0.7	0.1	0.0	
3	质量	0.0	0.4	0.5	0.1	
4	价格	0.2	0.3	0.4	0.1	
5	最终评价值	0.0	0.8	0.2	0.0	
6						

图 42-8 贴适度解模糊关系方程评语集及评价结果数据编辑定义图

(2) 可能的权分配方案集合(A)的矩阵数据编辑和定义。每一行为一个权的分配方案，每一列为某一个因素的权重。数据编辑后将数据定义成公式块。如上例根据对顾客的心理分析，提出了 4 种可能的权分配方案，其数据按图 42-9 方式编辑和定义。注意，此处的数据块定义时，需在按下“Ctrl”的同时拖动鼠标。

	A	B	C	D	E
1		外型	质量	价格	
2	第一种权分配方案:	0.2	0.50	0.30	
3	第二种权分配方案:	0.5	0.30	0.20	
4	第三种权分配方案:	0.2	0.30	0.50	
5	第四种权分配方案:	0.7	0.25	0.05	
6					

图 42-9 贴适度解模糊关系方程权重备择方案数据编辑定义图

完成评估数据(评判矩阵 R)、最终的评价结果 B 以及各择权重分配 A 的数据的编辑、定义之后，再进入菜单操作，在菜单方式下选择“模糊数学”→“综合评判逆问题”分析功能项，回车后便可得到归一化之后的分析结果。

贴适度分析：第 1 组的贴适度为 0.6500，第 2 组的贴适度为 0.7000，第 3 组的贴适度为 0.6000，第 4 组的贴适度为 0.8000。

从贴适度分析结果可以看出，在权重备择集中应选 A_4 作为近似的 X^* ，即

$$X^* = (0.7, 0.25, 0.05)$$

采用此方法，如果备择权分配方案太少，计算结果会不太理想。因此在实际应用时，为取得较满意的结果，应当尽可能多地建立一些权重分配方案。

参 考 文 献

- 高素恬. 1981. 模糊相似优先比在农业气候相似分析中的应用[J]. 气象, 7.
 葛苏林. 1985. 模糊子集-模糊关系-模糊映射[M]. 北京: 北京师范大学出版社.
 贺仲雄. 1983. 模糊数学及其应用[M]. 天津: 天津科技出版社.
 汪培庄. 1983. 模糊集合论及其应用[M]. 上海: 上海科学技术出版社.

吴万铎, 吴万钊. 1988. 模糊数学与计算机应用[M]. 北京: 电子工业出版社.

肖位枢. 1992. 模糊数学基础及其应用[M]. 北京: 航空工业出版社.

杨崇瑞, 耿济国. 1991. 农业病虫模糊分析预报[M]. 合肥: 安徽教育出版社.

第 43 章 灰色系统分析

现代科学技术在高度分化的基础上高度综合的大趋势，导致了具有方法论意义的横断学科群的出现。横断学科揭示了事物之间更为深刻、更具本质性的内在联系，大大促进了科学技术的整体化进程。许多科学领域中长期难以解决的复杂问题随着新兴横断学科的出现迎刃而解，人们对自然界和客观事物演化规律的认识也由于横断学科的出现而逐步深化。我国学者邓聚龙教授于 20 世纪 80 年代前期提出的用于控制和预测的新理论、新技术的灰色系统理论和方法，目前已广泛地应用于农业和社会经济等领域，并取得了显著成就。与研究“随机不确定性”的概率统计和研究“认知不确定性”的模糊数学不同，灰色系统理论的研究对象是“部分信息已知，部分信息未知”的“小样本”、“贫信息”不确定性系统。它通过对“部分”已知信息的生成、开发去了解、认识现实世界，实现对系统运行行为和演化规律的正确把握和描述。灰色系统模型对试验观测数据及其分布没有什么特殊的要求和限制，作为一种十分简便、易学易用的新理论，灰色系统理论具有十分广泛的应用领域，并深受各领域研究人员和实际工作者的喜爱。本章向用户提供一些常用灰色系统的分析处理技术及其功能模块，供用户参考使用。

43.1 关联度分析

客观世界中存在着大大小小的各类系统，都是由许多因素组成的。这些系统及系统因素之间，相互关系非常复杂。特别是表面现象变化的随机性容易混淆人们的直觉，掩盖事物的本质，使人们在认识、分析、预测和决策时得不到充分全面的信息，不容易形成明确的概念。因此，不仅不同系统之间的关系是灰的，同系统中不同因素之间的关系也是灰的。人们一时会分不清哪些因素关系密切，哪些因素关系不密切，也就是说难以找到主要矛盾，抓住主要特征与主要关系。为此，灰色系统理论提出了关联度分析的概念，其目的就是通过一定的方法理清系统中各因素间的主要关系，找出影响最大的因素，把握矛盾的主要方面。如各类产业中哪个项目的收入影响产值最明显，这种影响程度表明有关生产和销售系统之间或系统内部各因素之间的关联性。

对两个系统或两个因素之间关联性大小的量度称为关联度。它描述系统发展过程中因素间相对变化的情况，也就是变化大小、方向及速度等指标的相对性。如果两者在系统发展过程中相对变化基本一致，则认为两者关联度大；反之，两者关联度就小。可见，灰色关联度分析是对于一个系统发展变化态势的定量描述和比较。只有弄清楚系统或因素间的这种关联关系，才能对系统有比较透彻的认识，分清哪些是主导因素，哪些是潜在因素，哪些是优势而哪些又是劣势。所以，对于一个灰色系统进行分析研究时，首先要解决如何从随机的时间序列中找到关联性，计算关联度，以便为因素判别、优势分析和预测精度检验等提供依据，为系统决策打好基础。因此说，灰色因素间的关联度分析，实质上是灰色系统分析、预测、决策的基础。

灰色系统理论的关联度分析与数理统计学的相关分析是不同的，两者的区别在于：第一，它们的理论基础不同。关联度分析基于灰色系统的灰色过程，而相关分析则基于概率论的随机过程；第二，分析方法不同。关联分析是进行因素间时间序列的比较，而相关分析是因素间数组的比较；第三，数据量要求不同。关联分析不要求数据太多，而相关分析则需有足够的数量；第四，研究重点不同。关联度分析主要研究动态过程，而相关分析则以静态研究为主。因此，关联度分析适应性更广，在用于社会经济系统中的应用更有其独到之处。

目前，关联度分析应用十分广泛，几乎渗透到社会和自然科学各个领域，如农业、教育、卫生、政法、环保、军事、地理、地质、石油、水文、气象、生物，等等。尤其在社会经济领域，如预测宏观经济的发展态势、国民经济各部门投资效益、区域经济优势分析、技术经济的方案评价、产业结构的调整方向以及微观经济的因素分析等方面，都取得了较好的应用效果。

43.1.1 原理与方法简介

关联度分析一般包括下列计算和步骤：①原始数据变换；②计算关联系数；③求关联度；④排关联序；⑤列关联矩阵。在应用中是否进行所有步骤，可视具体情况而定。

设有 m 个时间序列

$$\begin{array}{cccccc}
 t & x_1^{(t)} & x_2^{(t)} & L & x_n^{(t)} \\
 1 & x_1^{(1)} & x_2^{(1)} & L & x_n^{(1)} \\
 2 & x_1^{(2)} & x_2^{(2)} & L & x_n^{(2)} \\
 M & M & M & & M \\
 m & x_1^{(m)} & x_2^{(m)} & L & x_n^{(m)}
 \end{array}$$

即

$$\{X_1^{(0)}(t)\}, \{X_2^{(0)}(t)\}, L, \{X_m^{(0)}(t)\}, \quad t=1, 2, \dots, N$$

式中， N 为各序列的长度即数据个数，这 m 个序列代表 m 个因素(变量)。另设定时间序列

$$\{X_0^{(0)}(t)\}, \quad t=1, 2, \dots, N$$

该时间序列称为母序列，而上述 m 个时间序列称为子序列。关联度是两个序列关联性大小的度量。根据这一观点，可给关联度一个量化模型，其计算方法与步骤具体叙述如下：

(1) 原始数据变换。

由于系统中各因素的量纲(或单位)不一定相同，如劳动力为人，产值为万元，产量为吨等，且有时数值的数量级相差悬殊，如人均收入为几百元，粮食每公顷产量为几千公斤，费用为几十万元，有些产业产值达百亿元，有些产业才几万元，等等，这样的数据很难直接进行比较，且它们的几何曲线比例也不同。因此，对原始数据需要消除量纲(或单位)，转换为可比较的数据序列。目前，原始数据的变换有以下几种常用方法：

① 均值化变换。先分别求出各个序列的平均值，再用平均值去除对应序列中的各个原始数据，得到新的数据列，即为均值化序列。其特点是量纲为一，其值大于 0，并且大部分近于 1，数列曲线互相相交。

② 初值化变换。分别用同一序列的第一个数据去除后面的各个原始数据，得到新的倍数数列，即为初值化数列。其特点是量纲为一，各值均大于 0，且数列有共同的起点。

③ 标准化变换。先分别求出各个序列的平均值和标准差，然后将各个原始数据减去平均值后再除以标准差，这样得到的新数据序列即为标准化序列。其特点是量纲为一，其均值为 0，方差为 1。

一般情况下，对于较稳定的社会经济系统数列作动态序列的关联度分析时，多采用初值化变换，因为这样的数列多数是增长的趋势。若对原始数列只作数值间的关联比较，可用均值化变换，譬如进行产业结构变化的关联分析，自然因素周期性变化的关联分析等。

(2) 计算关联系数。

经数据变换的母数列记为 $\{X_0(t)\}$ ，子数列记为 $\{X_i(t)\}$ ，则在时刻 $t = k$ 时母序列 $\{X_0(k)\}$ 与子序列 $\{X_i(k)\}$ 的关联系数 $L_{0i}(k)$ 可由下式计算：

$$L_{0i}(k) = \frac{\Delta_{\min} + \rho \Delta_{\max}}{\Delta_{0i}(k) + \rho \Delta_{\max}}$$

式中， $\Delta_{0i}(k)$ 表示 k 时刻两比较序列的绝对差，即 $\Delta_{0i}(k) = |x_0(k) - x_i(k)| (1 \leq i \leq m)$ ； Δ_{\max} 和 Δ_{\min} 分别表示所有比较序列各个时刻绝对差中的最大值与最小值。因为比较序列相交，故一般取 $\Delta_{\min} = 0$ ； ρ 称为分辨系数，其意义是削弱最大绝对差数值太大引起的失真，提高关联系数之间的差异显著性， $\rho \in (0, 1)$ ，一般情况下可取 0.1~0.5。

关联系数反映两个被比较序列在某一时刻的紧密(靠近)程度。如在 Δ_{\min} 时刻， $L = 1$ ；而在 Δ_{\max} 时刻则关联系数为最小值。因此，关联系数的范围为 $0 < L \leq 1$ 。

(3) 求关联度。

由以上所述可知，关联度分析实质上是对时间序列数据进行几何关系比较，若两序列在各个时刻点都重合在一起，即关联系数均等于 1，则两序列的关联度也必等于 1。另一方面，两比较序列在任何时刻也不可垂直，所以关联系数均大于 0，故关联度也都大于 0。因此，两序列的关联度便以两比较序列各个时刻的关联系数之平均值计算，即

$$r_{0i} = \frac{1}{N} \sum_{k=1}^N L_{0i}(k)$$

式中， r_{0i} 为子序列 i 与母序列 0 的关联度， N 为比较序列的长度(即数据个数)。

用几何坐标表示，即在横坐标为时间 t 、纵坐标为关联系数 L 的坐标图中(图 43-1)绘出关联系数曲线(虚线)。该折线与横坐标间围成的面积，称为关联面积，记作 S_{0i} ，而母序列自身的关联系数处为 1。所以，取纵坐标 $L=1$ ，作水平线与横坐标间围成的面积为重合面积，记为 S_{00} ，则关联度的几何意义为两面积之比，即 $r_{0i} = \frac{S_{0i}}{S_{00}}$ 。因关联系数曲

线为等时距，且 $S_{00} = 1$ ，故 $r_{0i} = S_{0i} = \frac{1}{N} \sum_{k=1}^N L_{0i}(k)$ 。

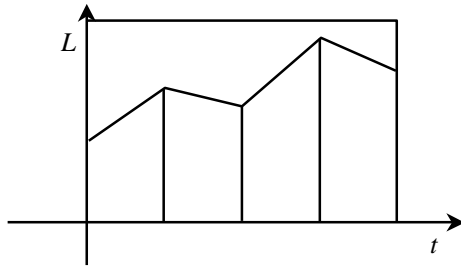


图 43-1 关联度几何图示

不难看出，关联度与下列因素有关：

- ① 母序列 X_0 不同，则关联度不同；
- ② 子序列 X_i 不同，则关联度不同；
- ③ 参考点 0 (或数据变换) 不同，关联度不同；
- ④ 数据序列长度 N 不同，关联度不同；
- ⑤ 分辨系数 ρ 不同，关联度不同。

一般来说，关联度也满足等价“关系”三公理，即① 自反性： $r_{00} = 1$ ；② 对称性： $r_{0i} = r_{i0}$ ；③ 传递性： $r_{0a} > r_{0b}$ ， $r_{0b} > r_{0c}$ ，则 $r_{0a} > r_{0c}$ 。

(4) 排关联序。

将 m 个子序列对同一母序列的关联度按大小顺序排列起来，便组成关联序，记为 $\{X\}$ 。它直接反映各个子序列对于母序列的“优劣”关系。若 $r_{0a} > r_{0b}$ ，则称 $\{X_a\}$ 对于相同母序列 $\{X_0\}$ 有优于 $\{X_b\}$ 的特点，记为 $\{X_a | X_0\} f \{X_b | X_0\}$ ；若 $r_{0a} < r_{0b}$ ，则称 $\{X_a\}$ 对于母序列 $\{X_0\}$ 劣于 $\{X_b\}$ ，记为 $\{X_a | X_0\} p \{X_b | X_0\}$ ；若 $r_{0a} = r_{0b}$ ，则称 $\{X_a\}$ 对于母序列 $\{X_0\}$ 等价于(或等于) $\{X_b\}$ ，记为 $\{X_a | X_0\} \sim \{X_b | X_0\}$ ；若有 $r_{0a} \geq r_{0b}$ ，称 $\{X_a\}$ 对于母序列 $\{X_0\}$ 优于或等于 $\{X_b\}$ ，记为 $\{X_a | X_0\} f \tilde{p} \{X_b | X_0\}$ ；若有 $r_{0a} \leq r_{0b}$ ，则称 $\{X_a\}$ 对于母序列 $\{X_0\}$ 劣于或等于 $\{X_b\}$ ，记为 $\{X_a | X_0\} \tilde{p} \{X_b | X_0\}$ 。

根据上述几种关系，可定义两种有代表性的关联序，即“有序”与“偏序”。若关联序 $\{X\}$ 为有序，那么所有元素之间必存在以下几种关系之一：“优于”(f)，“劣于”(p)，或“等价于”(~)。若关联序 $\{X\}$ 为偏序，则不是所有元素都可比较的。

一般而言，各因素只要能构成关系，算出关联度，则总是“有序”的。只有在无“参考点”或无参考母序列的情况下，才可能出现“偏序”现象。

(5) 列出关联矩阵。

若有 n 个母序列 $\{Y_1\}, \{Y_2\}, \dots, \{Y_n\} (n \neq 2)$ 及其 m 个子序列 $\{X_1\}, \{X_2\}, \dots, \{X_m\} (m \neq 1)$ ，则各子序列对母序列 $\{Y_1\}$ 有关联度 $[r_{11}, r_{12}, \dots, r_{1m}]$ ，各子序列对于母序列 $\{Y_2\}$ 有关联度 $[r_{21}, r_{22}, \dots, r_{2m}]$ ，类似地，各子序列对于母序列 $\{Y_n\}$ 有关联度 $[r_{n1}, r_{n2}, \dots, r_{nm}]$ 。

将 $r_{ij} (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$ 作适当排列，可得到关联度矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & L & r_{1m} \\ r_{21} & r_{22} & L & r_{2m} \\ M & M & & M \\ r_{n1} & r_{n2} & L & r_{nm} \end{bmatrix} \quad \text{或} \quad R = \begin{bmatrix} r_{11} & r_{12} & L & r_{1n} \\ r_{21} & r_{22} & L & r_{2n} \\ M & M & & M \\ r_{m1} & r_{m2} & L & r_{mn} \end{bmatrix}$$

关联度矩阵(简称关联矩阵),不仅可以作为优势分析的基础,而且可作为决策的依据。若关联矩阵 R 中第 i 列满足

$$\begin{bmatrix} r_{1i} \\ r_{2i} \\ M \\ r_{mi} \end{bmatrix} > \begin{bmatrix} r_{1j} \\ r_{2j} \\ M \\ r_{mj} \end{bmatrix} \quad (\forall i, j \in 1, 2, L, n, i \neq j)$$

则称母序列 $\{Y_i\}$ 相对于其他母序列为最优,或者说从 Y_i 对于子序列 $X_j (j=1, 2, \dots, m)$ 的关联度来看,序列 $\{Y_i\}$ 是系统最优序列,并记为

$$\{Y_i\} f \{Y_j\}, \quad j=1, 2, \dots, n; j \neq i$$

若有

$$\frac{1}{n} \sum_{k=1}^n r_{ki} > \frac{1}{n} \sum_{k=1}^n r_{kj}, \quad j=1, 2, \dots, n; j \neq i$$

则称母序列 $\{Y_i\}$ 相对于其余母序列,或相对于子序列 $\{X_i\} (i=1, 2, \dots, m)$ 的关联度是准最优的,并记为

$$\{Y_i\} f_{\%} \{Y_j\}, \quad j \in \{1, 2, \dots, n\}, j \neq i$$

若关联矩阵 R 为下三角矩阵,即

$$\begin{array}{cccc} r_{11} & & & \\ r_{21} & r_{22} & & \\ r_{31} & r_{32} & r_{33} & \\ M & M & M & O \\ r_{n1} & r_{n2} & r_{n3} & r_{nm} \end{array}$$

则称 $[Y_1]$ 相对于 $[Y_i] (i \in \{2, 3, \dots, n\})$ 是最优势的。

43.1.2 DPS 平台的操作步骤

在 DPS 平台上,数据编辑方式一般是前面 p 列为子序列,后面 q 列为母序列,每一行为一个样本(或某一时刻)的数据。若无指定的母序列数据,对有 q 个因子序列、 n 个时刻的资料输入顺序为

$$\begin{array}{ccccccc} & & & & \text{因子序列} & & \\ \text{时} & x_{11} & x_{12} & x_{13} & \dots & x_{1q} & \end{array}$$

刻	x_{21}	x_{22}	x_{23}	...	x_{2q}
	\mathbf{M}	\mathbf{M}	\mathbf{M}		\mathbf{M}
	x_{n1}	x_{n2}	x_{n3}	...	x_{nq}

这时系统会自动依次将各组作为母序列，其余为子序列，试计算关联矩阵并对因子进行排序。若有 p 个子序列、 q 个母序列、 n 个时刻的时间序列资料，其数据输入、编辑、定义数据块的顺序为子序列在左，母序列在右，方式为

		子序列				母序列				
	x_{11}	x_{12}	x_{13}	...	x_{1p}	y_{11}	y_{12}	y_{13}	...	y_{1q}
时	x_{21}	x_{22}	x_{23}	...	x_{2p}	y_{21}	y_{22}	y_{23}	...	y_{2q}
刻	\mathbf{M}	\mathbf{M}	\mathbf{M}		\mathbf{M}	\mathbf{M}	\mathbf{M}	\mathbf{M}		\mathbf{M}
	x_{n1}	x_{n2}	x_{n3}	...	x_{np}	y_{n1}	y_{n2}	y_{n3}	...	y_{nq}

数据输入并编辑好之后定义数据块，然后进入主菜单，选择“其他方法”→“灰色系统”→“关联分析”功能项。按回车执行时，系统还需用户根据分析要求作出一些选择(对每项选择系统都有默认值，即每次按回车也可得到分析结果)：

- (1) 在系统提示下输入母序列个数后回车。如果不指定母序列数目，按回车键即可。
- (2) 在系统提示下选择数据转换方式。一般为均值化。
- (3) 选择数据转换方式之后，系统提示用户输入分辨系数，这时如直接回车则取分辨系数的系统默认值(0.1)。
- (4) 最后，系统询问是否将分析参数 Δ_{\min} 取值为 0。如直接回车或键入“Y”后再回车则将参数 Δ_{\min} 取值为 0；如键入“N”后再回车，则参数 Δ_{\min} 的取值由各个数据序列各个时刻绝对差值的比较来确定。

最后系统输出分析结果，包括数据转换结果、母序列与其他因子序列的绝对差值、最大差值 Δ_{\max} 、关联系数及其关联序等。

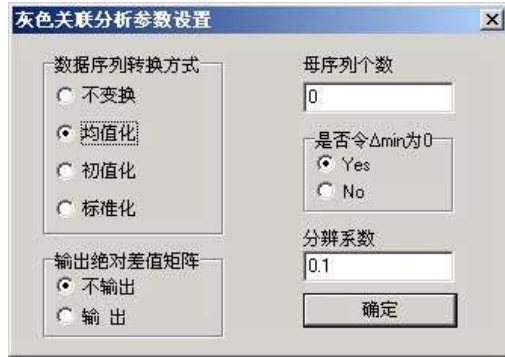
43.1.3 应用示例

(1) 无母序列情形。设有下述 4 组时间序列，若依次将各组作为母序列，其余为子序列，试计算关联矩阵并对因子进行排序。在 DPS 平台的编辑状态下，按图 43-2 方式编辑定义待分析数据块。

	A	B	C	D	E	F
1		无母序列数据关联分析数据编辑格式				
2		X1	X2	X3	X4	
3		39.1	45.8	3.4	6.7	
4		41.6	43.4	3.3	6.8	
5		43.9	42.3	3.5	5.4	
6		44.9	41.9	3.5	4.7	
7						

图 43-2 关联分析的数据编辑、定义示意图

然后在菜单下执行灰色关联分析选项功能。



在分析过程中对数据进行均值化变换，分辨系数取 0.1，参数 Δ_{\min} 取值为 0，不输出绝对差值矩阵，输出结果如下：

均值化变换结果(略)				
最大差值 $\Delta_{\max} = 0.26298$				
x1 和其他因子的关联序：				
No.	因子	关联系数		
X3	x3	0.4815		
X2	x2	0.3159		
X4	x4	0.1283		
最大差值 $\Delta_{\max} = 0.16994$				
x2 和其他因子的关联序：				
No.	因子	关联系数		
X3	x3	0.2564		
X1	x1	0.2384		
X4	x4	0.1470		
最大差值 $\Delta_{\max} = 0.22529$				
x3 和其他因子的关联序：				
No.	因子	关联系数		
X1	x1	0.4465		
X2	x2	0.3132		
X4	x4	0.1270		
最大差值 $\Delta_{\max} = 0.26298$				
x4 和其他因子的关联序：				
No.	因子	关联系数		
X2	x2	0.2086		
X3	x3	0.1450		
X1	x1	0.1283		
关联矩阵	x1	x2	x3	x4

x1	1.0000	0.3159	0.4815	0.1283
x2	0.2384	1.0000	0.2564	0.1470
x3	0.4465	0.3132	1.0000	0.1270
x4	0.1283	0.2086	0.1450	1.0000

(2) 有母序列情形。当指定某一组或多组数据为母序列时，这时母序列数据必须放在数据矩阵右边。例如，王淑荣(1995)应用灰色关联分析大豆主要数量性状的选择，通过田间调查株高、主茎节数、单株荚数、单株粒数及室内考种测得百粒重等诸项指标。现以这些因素作为子因素，以大豆产量作为母因素进行分析。先将试验观察数据在 DPS 平台上按规定格式编辑和定义，如图 43-3 所示。

	A	B	C	D	E	F	G	H
1	有母序列数据关联分析数据编辑格式							
2	品系	株高	主茎节数	单株荚数	单株粒数	百粒重	产量	
3	v1	78.8	14.3	28.9	74.8	17.7	1974	
4	v2	60.5	13.4	33.9	89.9	16.8	1688.4	
5	v3	74.3	13.7	26.5	70.9	17.7	1786.5	
6	v4	69.3	14.3	34.8	83.6	19.9	1796.6	
7	v5	77	12.6	27.9	75.4	20	1796.3	
8	v6	76.5	12.4	21.5	55.3	18.1	1826.1	
9	v7	67.5	11	24.8	70	20.3	1715.5	
10	v8	78.7	12.2	30.6	81.4	18.3	1991.5	
11	v9	76.1	13.3	38	94.7	18.4	1847.4	
12	v10	75.5	13.1	23.1	62.8	17	1854.6	
13								

图 43-3 灰色关联分析的数据编辑与定义(子序列在左，母序列在右)

按图 43-3 方式编辑定义数据之后，在菜单下选择“其他方法”→“灰色系统”→“关联分析”功能项。执行灰色关联分析时，按系统提示输入母序列个数 1 之后再回车，数据转换方式选择标准化，分辨系数取 0.1，参数 Δ_{\min} 的取值从各个数据序列各个时刻的绝对差值的比较来确定。分析结束时系统输出如下结果：

均值化变换结果：					
株高	主茎节数	单株荚数	单株粒数	百粒重	产量
1.0733	1.0975	0.9966	0.9858	0.9609	1.0801
0.8240	1.0284	1.1690	1.1848	0.9121	0.9238
1.0120	1.0514	0.9138	0.9344	0.9609	0.9775
0.9439	1.0975	1.2000	1.1017	1.0803	0.9830
1.0488	0.9670	0.9621	0.9937	1.0858	0.9828
1.0420	0.9517	0.7414	0.7288	0.9826	0.9991
0.9194	0.8442	0.8552	0.9225	1.1021	0.9386
1.0719	0.9363	1.0552	1.0727	0.9935	1.0896
1.0365	1.0207	1.3103	1.2480	0.9989	1.0108
1.0283	1.0054	0.7966	0.8276	0.9229	1.0147

产量 与其他因子的绝对差值:											
X1	株高	0.0068	0.0998	0.0345	0.0391	0.0659	0.0428	0.0192	0.0177	0.0257	0.0136
X2	主茎节数	0.0174	0.1046	0.0740	0.1145	0.0158	0.0475	0.0944	0.1533	0.0099	0.0094
X3	单株荚数	0.0835	0.2452	0.0637	0.2170	0.0208	0.2578	0.0834	0.0345	0.2996	0.2182
X4	单株粒数	0.0943	0.2610	0.0431	0.1188	0.0108	0.2703	0.0161	0.0169	0.2372	0.1871
X5	百粒重	0.1191	0.0117	0.0166	0.0974	0.1030	0.0165	0.1634	0.0961	0.0119	0.0918
最大差值 $\Delta_{\max}=0.29956$											
产量 和其他因子的关联序:											
No.	因子	关联系数									
X1	株高	0.5132									
X2	主茎节数	0.4309									
X5	百粒重	0.4024									
X4	单株粒数	0.3330									
X3	单株荚数	0.2450									
关联矩阵											
	株高	主茎节数	单株荚数	单株粒数	百粒重						
产量	0.5132	0.4309	0.2450	0.3330	0.4024						

从分析结果来看,大豆主要数量性状与产量之间的关联度第一位是株高,其次是主茎节数,再其次分别是百粒重、单株粒数和单株荚数。

43.2 灰色动态(GM)建模基本原理

预测实际上是建立模型的一种过程,即根据过去和现在已有的信息建立一个从过去引申到未来的模型,灰色系统的 GM 模型就具有这种作用。但作为预测模型,主要指标必须具有足够的精度。为此,应首先弄清楚灰色动态 GM 模型的建模原理与方法。

43.2.1 灰色动态建模原理

灰色预测建模是以灰色模块概念为基础的。灰色系统理论认为一切随机量都是在一定范围内、一定时段上变化的灰色量及灰色过程。对于灰色量的处理,不是去寻求它的统计规律和概率分布,而是从无规律的原始数据中找出规律,即对数据通过一定方式处理后,使其成为较有规律的时间序列数据,再建立模型。因为无论客观系统怎样复杂,它总是有关联、有整体功能、因而有序的。因此,作为表现系统行为特征的数据,总是蕴含着某种规律。经过一定方式处理而生成的序列数据,称为“模块”。它的几何意义是指生成序列数据在时间和数据二维平面上所给的连续曲线与其底部(即横坐标)的总称。由已知数据列构成的模块,称为白色模块,而由白色模块外推到未来的模块,即由预测值构成的模块,称为灰色模块。

一般情况下,对于给定的原始数据列

$$X_{(0)} = \{x_{(0)}^{(1)}, x_{(0)}^{(2)}, x_{(0)}^{(3)}, \dots, x_{(0)}^{(N)}\}$$

不能直接用于建模,因为这些数据多为随机的、无规律的。若将原始数据列经过一次累加生成,可获得新数据列

$$X_{(1)} = \{x_{(1)}^{(1)}, x_{(1)}^{(2)}, x_{(1)}^{(3)}, \dots, x_{(1)}^{(N)}\}$$

其中

$$x_{(1)}^{(i)} = \sum_{k=1}^i x_{(0)}^{(k)}$$

新生成的数据列为一单调增长的曲线，显然它增强了原始数列的规律性，而随机性被弱化了。对于非负的数据列，累加的次数越多，随机性弱化就越明显且规律性越强，因而比较容易用指数函数去逼近。可见，这样的数据处理能够达到两个目的：

- (1) 弱化了原始数据列的随机性而找到了其变化的规律性。
- (2) 为建立动态模型提供了中间信息。

对于生物学、社会经济学等系统，应用微分方程便于描述其内部物理或化学过程动态特征。灰色理论之所以能够用来建立微分方程模型，是基于下述概念、观点、方法和途径。

- 灰色理论将随机量当作是在一定范围内变化的灰色量，将随机过程当作是在一定幅区和一定时区变化的灰色过程。
- 灰色理论将无规律的原始数据生成后，会先使其成为较有规律的生成数列再建模。所以，灰色 GM 建模实际上是生成数据模型，而一般建模所用的是原始数据模型。
- 通过 GM 模型得到的数据，必须经过逆生成还原后才能有用。
- 灰色理论是针对符合光滑离散函数条件的一类数列建模，一般原始数据作累加生成后可以得到这样的光滑离散函数。
- 灰色理论认为微分方程是背景与各阶导数(灰导数)的某种组合(线性或非线性的)。
- 灰色理论通过灰数的不同生成，数据的不同取舍，不同级的残差模型的补充，来调整、修正、提高模型的精度。
- 对于高阶系统，可由一阶的 GM 模型群建立状态方程解决。

43.2.2 常见的线性微分拟合模型

GM (n, h)模型为 n 阶 h 个变量的微分方程，不同 n 与 h 的 GM 模型有不同的意义和用途，要求有不同的数据处理。作为预测模型，常用 GM ($n, 1$)模型，即只有一个变量的 GM 模型，对数据列要求是“综合效果”的时间序列。由于 n 越大，计算越复杂，但精度未必就高，因此一般取 n 在 3 阶以下。最常用的 $n=1$ 阶模型，计算简单，适用性广，记为 GM (1, 1)，称为单序列一阶线性动态模型。

GM (1, 1)模型的微分方程为

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u$$

系数向量

$$\hat{a} = [a, \mu]^T$$

因 $n = 1, h = 1$ ，故矩阵 $A = 0$ ，累加矩阵 B 为

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -\frac{1}{2}(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ \quad \quad \quad M & M \\ -\frac{1}{2}(x^{(1)}(N-1) + x^{(1)}(N)) & 1 \end{bmatrix}$$

常数项向量

$$Y_N = (x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(N))^T$$

用最小二乘法求解 \hat{a} :

$$\hat{a} = (B^T B)^{-1} B^T Y_N$$

并代入微分方程的解, 得到时间函数

$$\hat{x}^{(1)}(t+1) = \left(x^{(1)}(0) - \frac{u}{a} \right) e^{-at} + \frac{u}{a}$$

令 $x^{(1)}(0) = x^{(0)}(1)$, 则

$$\hat{x}^{(1)}(t+1) = \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-at} + \frac{u}{a}$$

求导还原后可得到

$$\hat{x}^{(0)}(t+1) = -a \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-at}$$

上述两个方程即为 GM(1, 1)模型灰色预测的基本计算公式。

GM(1, 1)模型只有一个指数分量, 故变化是单调的。GM(2, 1)为二阶模型, 有两个特征根, 其动态过程能反映单调的、非单调的或摆动(振荡的)等不同情况, GM(2, 1)模型的微分方程为

$$\frac{d^2 x^{(1)}}{dt^2} + a_1 \frac{dx^{(1)}}{dt} + a_2 x^{(1)} = u$$

其系数向量 $\hat{a} = (a_1, a_2, u)^T$ 可用最小二乘法求解:

$$\hat{a} = \begin{bmatrix} a_1 \\ a_2 \\ u \end{bmatrix} = \left[X(A, B)^T X(A, B) \right]^{-1} X(A, B)^T Y_N$$

式中

$$A = \begin{bmatrix} -\alpha^{(1)}(x^{(1)}, 2) \\ -\alpha^{(1)}(x^{(1)}, 3) \\ \mathbf{M} \\ -\alpha^{(1)}(x^{(1)}, N) \end{bmatrix}$$

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(2) + x^{(1)}(1)) & 1 \\ -\frac{1}{2}(x^{(1)}(3) + x^{(1)}(2)) & 1 \\ \mathbf{M} & \mathbf{M} \\ -\frac{1}{2}(x^{(1)}(N) + x^{(1)}(N-1)) & 1 \end{bmatrix}$$

$$Y_N = [\alpha^{(2)}(x^{(1)}, 2), \alpha^{(2)}(x^{(1)}, 3), \dots, \alpha^{(2)}(x^{(1)}, N)]^T$$

其时间响应函数为

$$x^{(1)}(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} + \frac{u}{a^2}$$

式中, λ_1, λ_2 为两个特征根, 按以下不同情况可分析系统的主要动态特征:

- (1) 若 $\lambda_1 = \lambda_2$, 则动态过程是单调的;
- (2) 若 $\lambda_1 \neq \lambda_2$, 且为实数, 则动态过程可能是非单调的;
- (3) 若 λ_1, λ_2 为共轭复根, 则动态过程是周期摆动的。

上面介绍的 GM(1, 1) 和 GM(2, 1) 多用于预测, 而作为状态分析模型, 常用 GM(1, h) 模型, 它可以反映 $h-1$ 个变量对于因变量一阶导数的影响。由于 $h > 1$, 故称为 h 个序列的一阶线性动态模型。其建模步骤如下:

设有 h 个变量 X_1, X_2, \dots, X_h 组成原始数列

$$x_i^{(0)} = \{x_i^{(0)}(1), x_i^{(0)}(2), \dots, x_i^{(0)}(N)\}, \quad i=1, 2, \dots, h$$

对 $X_i^{(0)}$ 分别作一次累加生成, 得到新的数列

$$X_i^{(1)} = \{x_i^{(1)}(1), x_i^{(1)}(2), \dots, x_i^{(1)}(N)\}, \quad i=1, 2, \dots, h$$

建立微分方程

$$\frac{dx_1^{(1)}}{dt} + ax_1^{(1)} = b_1 x_2^{(1)} + b_2 x_3^{(1)} + L + b_{h-1} x_h^{(1)}$$

其系数向量

$$\hat{a} = (b_1, b_2, \dots, b_{h1})^T$$

用最小二乘法求解, 即

$$\hat{a} = (B^T B)^{-1} B^T Y_N$$

式中, B 为累加矩阵, Y_N 为常数项向量, 分别为

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(1) + x^{(1)}(2)) & x_2^{(1)}(2) & L & x_h^{(1)}(2) \\ -\frac{1}{2}(x^{(1)}(2) + x^{(1)}(3)) & x_2^{(1)}(3) & L & x_h^{(1)}(3) \\ & M & & M \\ -\frac{1}{2}(x^{(1)}(N-1) + x^{(1)}(N)) & x_2^{(1)}(N) & L & x_h^{(1)}(N) \end{bmatrix}$$

$$Y_N = [x_1^{(0)}(2), x_1^{(0)}(3), \dots, x_1^{(0)}(N)]^T$$

求得微分方程的解

$$x_1^{(1)}(t+1) = (x_1^{(0)}(1) - \sum_{i=2}^h \frac{b_{i-1}}{a} x_i^{(1)}(t+1))e^{-at} + \sum_{i=2}^h \frac{b_{i-1}}{a} x_i^{(1)}(t+1)$$

43.3 灰色数列 GM(1, 1)模型

灰色数列预测主要是指利用 GM(1, 1)模型, 对时间序列数据进行数量大小的预测。例如, 人口预测、劳力预测、产品产量预测、各业产值预测及病虫发生趋势预测等, 就是利用历年统计资料, 对其未来的发展进行预测。这类预测不仅应用最广, 而且方法步骤也有普遍意义。因此, 下面作比较详细的介绍。

43.3.1 建立 GM(1, 1)模型的基本步骤

第 1 步. 对数据序列 $X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(N)\}$ 作一次累加生成, 得到 $X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(N)\}$, 其中 $x^{(1)}(t) = \sum_{k=1}^t x^{(0)}(k)$ 。

第 2 步. 构造累加矩阵 B 与常数项向量 Y_N , 即

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -\frac{1}{2}(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ & M & M \\ -\frac{1}{2}(x^{(1)}(N-1) + x^{(1)}(N)) & 1 \end{bmatrix}$$

$$Y_N = [x_1^{(0)}(2), x_1^{(0)}(3), \dots, x_1^{(0)}(N)]^T$$

第 3 步. 用最小二乘法解灰参数 \hat{a} :

$$\hat{a} = \begin{bmatrix} a \\ u \end{bmatrix} = (B^T B)^{-1} B^T Y_N$$

第 4 步. 将灰参数代入时间函数

$$\hat{x}^{(1)}(t+1) = \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-at} + \frac{u}{a}$$

第 5 步. 对 $\hat{X}^{(1)}$ 求导还原得到

$$\hat{x}^{(0)}(t+1) = -a \left(x^{(0)}(1) - \frac{u}{a} \right) e^{-at} \quad \text{或} \quad \hat{x}^{(0)}(t+1) = \hat{x}^{(1)}(t+1) - \hat{x}^{(1)}(t)$$

第 6 步. 计算 $x^{(0)}(t)$ 与 $\hat{x}^{(0)}(t)$ 之差 $\varepsilon^{(0)}(t)$ 及相对误差 $e(t)$:

$$\begin{aligned} \varepsilon^{(1)}(t) &= x^{(0)}(t) - \hat{x}^{(0)}(t) \\ e(t) &= \varepsilon^{(0)}(t)/x^{(0)}(t) \end{aligned}$$

第 7 步. 模型诊断及应用模型进行预报。

为了分析模型的可靠性, 必须对模型进行诊断。目前较通用的诊断方法是对模型进行后验差检验, 即先计算观察数据离差 s_1 :

$$s_1^2 = \sum_{t=1}^m (x^{(0)}(t) - \bar{x}^{(0)}(t))^2$$

及残差的离差 s_2 :

$$s_2^2 = \frac{1}{m-1} \sum_{t=1}^{m-1} (q^{(0)}(t) - \bar{q}^{(0)}(t))^2$$

再计算后验比

$$c = \frac{s_1}{s_2}$$

及小误差概率

$$p = \left\{ \left| q^{(0)}(t) - \bar{q}^{(0)} \right| < 0.6745s_1 \right\}$$

根据后验比 c 和小误差概率 p 对模型进行诊断, 当 $p > 0.95$ 和 $c < 0.35$ 时, 模型可靠, 这时可根据模型对系统行为进行预测。

上述 7 步为整个建模、预测的分析过程。当所建立的模型残差较大、精度不够理想时, 为提高精度, 一般可以对其残差进行残差 GM(1, 1) 模型建模分析, 以修正预报模型。

43.3.2 DPS 平台的操作示例

在 DPS 平台上, 先将系列原始数据资料按时间顺序从左到右、从上到下地在编辑器中输入, 每行的数据个数不一定相同。编辑好数据后, 将待分析的数据定义为数据矩阵块, 如图 43-4 上部所示。

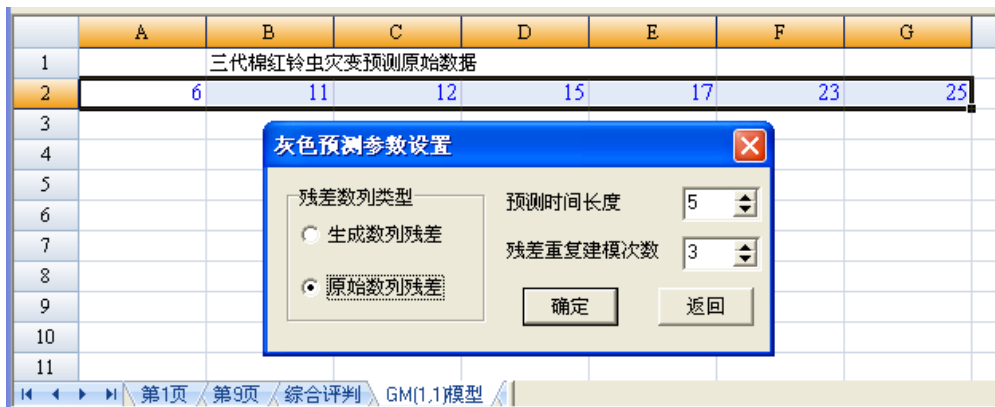


图 43-4 GM(1, 1)建模分析数据编辑、定义示意图

然后, 在菜单下选择 GM(1, 1)建模分析功能项, 执行运算时, 系统会出现用户参数设置界面(图 43-6 下部)。这里用户可以输入待预测时刻(样本)的个数, 再选择待识别的残差类型, 以及对残差重复建模的次数。点击“确定”按钮执行分析。

最后系统输出模型参数、观察值、拟合值、误差并对模型作出评价, 同时还给出未来的预测值。

例如, 汪笃栋等(1994)对彭泽县 1960~1986 年 3 代棉红铃虫百株卵量, 运用灰色系统理论和方法, 建立了 GM(1, 1)模型, 进行长期灾变预测。根据该县数据, 在 DPS 平台上作同样分析。首先将原始数据按年份顺序整理成表 43-1。

表 43-1 彭泽县 1960~1986 年三代棉红铃虫百株卵量(粒)

序号	年份	百株卵量	序号	年份	百株卵量	序号	年份	百株卵量
1	1960	584	10	1969	775	19	1978	2888
2	1961	1343	11	1970	4435	20	1979	2652
3	1962	1530	12	1971	3117	21	1980	1954
4	1963	410	13	1972	428	24	1983	2920
5	1964	464	14	1973	270	23	1982	5870
6	1965	12385	15	1974	4284	24	1983	2920
7	1966	1799	16	1975	1527	25	1984	6148
8	1967	2242	17	1976	7295	26	1985	2378
9	1968	2533	18	1977	453	27	1986	1384

然后, 根据原作者研究, 将全代卵量大于 3000 粒以上的年份作为大发生年。据此, 以全代卵量 3000 粒为阈值, 对离散函数

$$Y^{(0)} = \{y^{(0)}(1), y^{(0)}(2), y^{(0)}(3), \dots, y^{(0)}(27)\}$$

$$= \{584, 1343, 1530, 410, 464, 12385, 1799, 2242, 2533, 775, 4435, 3117, 428, 270, 4284, 1527, 7295, 453, 2888, 2652, 1954, 2920, 5870, 2920, 6148, 2378, 1384\}$$

作灾变映射 ξ : $\{Y^{(0)}\} = \{Y_{\xi}^{(0)}\}$, 按 $\xi \geq 3000$ 得到灾变集 $\{Y_{\xi}^{(0)}(1), Y_{\xi}^{(0)}(2), \dots, Y_{\xi}^{(0)}(27)\} = \{12385, 4435, 3117, 4284, 7295, 5870, 6148\}$ 。其中 $\forall Y_{\xi}^{(0)}(i) \geq \xi$ 且 $Y_{\xi}^{(0)}(i) \in Y^{(0)}, i \in \{1, 2, 3, 4, 5, 6, 7\}$, 经对 $Y^{(0)}$ 编序后再作映射 X , 得灾变日期集

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(7)\}$$

$$= \{6, 11, 12, 15, 17, 23, 25\}$$

这即为我们建立灰色模型 GM(1, 1) 时所需的原始数据(图 43-4)。

将数据编辑定义成数据块(图 43-4)之后, 再执行 GM(1, 1) 建模分析功能。按系统提示输入待预测的样本数(3), 然后选择待识别的残差类型为累加序列与其理论值的误差, 作两次残差序列的建模分析时得到如下结果:

模型参数: $a = -0.178152$ $b = 8.481744$				
$x(t+1) = 53.609524e0.178152t + -47.609524$				
原始数据及其拟合	观察值	拟合值	误差	误差%
$X(2) =$	11.0000	10.4542	0.54575	4.96
$X(3) =$	12.0000	12.4929	-0.49290	-4.11
$X(4) =$	15.0000	14.9291	0.07089	0.47
$X(5) =$	17.0000	17.8404	-0.84039	-4.94
$X(6) =$	23.0000	21.3194	1.68060	7.31
$X(7) =$	25.0000	25.4768	-0.47683	-1.91
对当前模型的评价				
$C = 0.1353$ 很好				
$p = 1.0000$ 很好				
未来 3 个时刻预测值				
$X(t+1) = 30.44500$				
$X(t+2) = 36.38200$				
$X(t+3) = 43.47675$				
*** 继续对残差序列进行建模分析 ***				
第 1 次残差序列分析结果				
模型参数				
$a = -0.214610$ $b = 0.219015$				
$x(t+1) = 2.282920e0.214610t + -1.020524$				
原始数据及其拟合	观察值	拟合值	误差	误差%
$X(2) =$	11.0000	10.4542	0.54575	4.96
$X(3) =$	12.0000	11.7770	0.22301	1.86

$X(4) =$	15.0000	14.8369	0.16309	1.09
$X(5) =$	17.0000	17.8394	-0.83943	-4.94
$X(6) =$	23.0000	22.3598	0.64023	2.78
$X(7) =$	25.0000	25.0856	-0.08564	-0.34
对当前模型的评价				
$C = 0.0783$ 很好				
$p = 1.0000$ 很好				
未来 3 个时刻预测值				
$X(t+1) = 31.32643$				
$X(t+2) = 37.64597$				
$X(t+3) = 45.21485$				
*** 继续对残差序列进行建模分析 ***				
第 2 次残差序列分析结果				
模型参数				
$a = 0.057166 \quad b = 0.606903$				
$x(t+1) = -9.940109e - 0.057166t + 10.616449$				
原始数据及其拟合	观察值	拟合值	误差	误差%
$X(2) =$	11.0000	10.4542	0.54575	4.96
$X(3) =$	12.0000	11.7770	0.22301	1.86
$X(4) =$	15.0000	14.7129	0.28713	1.91
$X(5) =$	17.0000	17.5216	-0.52161	-3.07
$X(6) =$	23.0000	22.8524	0.14760	0.64
$X(7) =$	25.0000	24.9107	0.08933	0.36
对当前模型的评价				
$C=0.0523$ 很好				
$p=1.0000$ 很好				
未来 3 个时刻预测值				
$X(t+1) = 31.31251$				
$X(t+2) = 37.60764$				
$X(t+3) = 45.15345$				

根据计算结果，第一个原始数据序列模型与原作者所建立的模型完全相同，但是在对残差数据序列进行建模时，由于采用的残差类型不同，所得到的结果也不同。从分析可以看出，采用该系统，经过一次残差拟合，就得到了较原作者经过两次残差修正后的模型还要高的精度。

从对历史资料回测验证和 1987~1991 年预测结果验证来看，实况和预测值的吻合程度很高。

43.4 灰色数列 GM(2, 1)模型

43.4.1 建立 GM(2, 1)模型的基本步骤

GM(1, 1)模型一般仅用于预测，而灰色数列 GM(2, 1)模型为单序列二阶线性动态模型，它不仅可以进行预测，还可以进行动态分析。建立 GM(2, 1)模型的步骤如下：

第 1 步. 对数据序列

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(N)\}$$

作一次累加生成，得到

$$X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(N)\}$$

其中, $x^{(1)}(t) = \sum_{k=1}^t x^{(0)}(k)$, 然后作累加生成数据序列的一次累差

$$\alpha^{(1)}(x^{(1)}, 1) = x^{(1)}(1) - x^{(1)}(0) = x^{(0)}(1)$$

$$\alpha^{(1)}(x^{(1)}, 2) = x^{(1)}(2) - x^{(1)}(1) = x^{(0)}(2)$$

L L L L

$$\alpha^{(1)}(x^{(1)}, N) = x^{(1)}(N) - x^{(1)}(N-1) = x^{(0)}(N)$$

再作累加数据序列的二次累差

$$\alpha^{(2)}(x^{(1)}, 2) = \alpha^{(1)}(x^{(1)}, 2) - \alpha^{(1)}(x^{(1)}, 1) = x^{(0)}(2) - x^{(0)}(1)$$

$$\alpha^{(2)}(x^{(1)}, 3) = x^{(0)}(3) - x^{(0)}(2)$$

L L L L

$$\alpha^{(2)}(x^{(1)}, N) = x^{(0)}(N) - x^{(0)}(N-1)$$

第 2 步. 构造矩阵 $X(A, B)$

$$X(A, B) = \begin{bmatrix} -\alpha^{(1)}(x^{(1)}, 2) - \frac{1}{2}(x^{(1)}(2) + x^{(1)}(1)) & & & 1 \\ -\alpha^{(1)}(x^{(1)}, 3) - \frac{1}{2}(x^{(1)}(3) + x^{(1)}(2)) & & & 1 \\ & M & M & M \\ -\alpha^{(1)}(x^{(1)}, N) - \frac{1}{2}(x^{(1)}(N) + x^{(1)}(N-1)) & & & 1 \end{bmatrix}$$

和

$$Y_N = [\alpha^{(2)}(x^{(1)}, 2), \alpha^{(2)}(x^{(1)}, 3), \dots, \alpha^{(2)}(x^{(1)}, N)]^T$$

第 3 步. 求

$$\frac{d^2 x^{(1)}}{dt^2} + a_1 \frac{dx^{(1)}}{dt} + a_2 x^{(1)} = u$$

的系数向量

$$\hat{a} = \begin{bmatrix} a_1 \\ a_2 \\ u \end{bmatrix} = \left[X(A, B)^T X(A, B) \right]^{-1} X(A, B)^T Y_N$$

第 4 步. 解系数特征方程并得到系统响应方程

$$\gamma^2 + a_1 \gamma + a_2 = 0$$

得

$$\gamma_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}$$

令 $W = a_1^2 - 4a_2$, 这时有 2 种情形: 若 $W > 0$, 则微分方程的解为

$$x^{(1)}(t+1) = c_1 e^{\lambda t} + c_2 e^{\lambda_2 t} + c^*$$

式中, c_1, c_2 为系数。当 $t=0$ 时, 系数的计算根据

$$x^{(1)}(1) = c_1 + c_2 + c^* = X^{(0)}(1) \quad (c^* \text{ 为待定常数})$$

其导数

$$\frac{dx^{(1)}}{dt} = c_1 \gamma_1 + c_2 \gamma_2 \approx x^{(1)}(2) - x^{(1)}(1) = x^{(0)}(2)$$

故由联立方程

$$\begin{cases} c_1 + c_2 + c^* = x^{(0)}(1) \\ c_1 \gamma_1 + c_2 \gamma_2 = x^{(0)}(2) \end{cases}$$

求解, 得

$$c_1 = x^{(0)}(1) - c^*$$

又由

$$(x^{(0)}(2) - c^*) \gamma_1 + c_2 \gamma_2 = x^{(0)}(2)$$

得

$$c_2 = \frac{x^{(0)}(2) - (x^{(0)}(1) - c^*) \gamma_1}{\gamma_2 - \gamma_1}$$

应用待定系数法求得的 c^* 的特解为

$$c^* = u/a_2$$

若 $W < 0$, 有微分方程的解为

$$x^{(1)}(t+1) = e^{\alpha t} (A_1 \cos \beta t + A_2 \sin \beta t) + c^*$$

式中, $\alpha = \frac{-a_1}{2}, \beta = \frac{\sqrt{4a_2 - a_1^2}}{2}$; A_1, A_2 为系数, 经推导可得

$$A_1 = x^{(0)}(1) - c^*$$

$$A_2 = \frac{x^{(0)}(2) - \alpha A_1}{\beta} = \frac{x^{(0)}(2) - \alpha(x^{(0)}(1) - c^*)}{\beta}$$

第 5 步. 将灰参数代入时间函数, 求生成数据序列计算值 $\hat{x}^{(1)}(t)$, 再对 $\hat{x}^{(1)}$ 求导还原得到 $\hat{x}^{(0)}(t)$ 。

第 6 步. 计算 $x^{(0)}(t)$ 与 $\hat{x}^{(0)}(t)$ 之差 $\varepsilon^{(0)}(t)$ 及相对误差 $e(t)$:

$$\varepsilon^{(1)}(t) = x^{(0)}(t) - \hat{x}^{(0)}(t)$$

$$e(t) = \varepsilon^{(0)}(t) / x^{(0)}(t)$$

第 7 步. 利用上述模型进行预测及系统分析。

43.4.2 DPS 平台的操作示例

原始数据资料的编辑定义方式与前述 GM(1, 1)模型的数据编辑定义方式完全相同, 即按时间顺序从左到右、从上到下地在编辑器中输入数据(数据放在一行或一列都可以), 每行的数据个数不一定相同。将待分析的数据定义成数据块, 如图 43-5 所示。

	A	B	C	D	E	F	G
2		2.874	3.278	3.337	2.874	3.278	
3							

图 43-5 GM(2, 1)建模分析数据编辑定义示意图

邓聚龙(1985)应用 GM(2, 1)模型分析某地某作物历年种子用量情况, 我们引用其数据在 DPS 平台上进行分析。将原始数据按图 43-5 方式编辑定义, 然后在菜单下选择执行 GM(2, 1)建模分析功能项, 可立即得到同样的计算结果。输出结果有模型参数、观察值、拟合值、误差等, 并对模型作出评价, 最后给出预测结果。

模型参数				
$a1 = 25.3658 a2 = -0.9621 U = 75.41002$				
$r1 = 0.03787$		$r2 = -25.40371$		
$C1 = 81.35986$		$C2 = -0.00435$		
$x1(t + 1) = 81.35986e^{0.03787t} + -0.00435e^{-25.40371t} + -78.38051$				
$x0(2) = 3.24574$...	3.27800	$e = 0.03226$	$q = 0.9841$
$x0(3) = 3.26161$...	3.33700	$e = 0.07539$	$q = 2.2593\%$
$x0(4) = 3.38750$...	3.39000	$e = 0.00250$	$q = 0.0737\%$
$x0(5) = 3.51826$...	3.67900	$e = 0.16074$	$q = 4.3692\%$
$C = 0.2381$ 很好				
$p = 1.0000$ 很好				
预测: $G0 = 16.2871067$				
$\hat{X}1(6) = 19.94116$		$\hat{x}0(6) = 3.65406$		
$\hat{X}1(7) = 23.73626$		$\hat{x}0(7) = 3.79510$		
$\hat{X}1(8) = 27.67785$		$\hat{x}0(8) = 3.94158$		

43.5 灰色数列 GM(1, N)模型

43.5.1 建立 GM(1, N)模型的基本步骤

GM(1, 1)模型和 GM(2, 1)模型均为单序列线性动态模型, 而 GM(1, N)模型是描述多元(多变量)一阶线性动态模型, 它主要用于系统的动态分析。

对于具有 N 个数列、数列长度为 M 的原始数据, 可用如下数据矩阵描述:

$$X_N^{(0)} = \begin{bmatrix} x_1^{(0)}(1) & x_2^{(0)}(1) & L & x_N^{(0)}(1) \\ x_1^{(0)}(2) & x_2^{(0)}(2) & L & x_N^{(0)}(2) \\ M & M & & M \\ x_1^{(0)}(M) & x_2^{(0)}(M) & L & x_N^{(0)}(M) \end{bmatrix}$$

第 1 步. 计算一次累加生成的数据矩阵

$$X_N^{(1)} = \begin{bmatrix} \sum_{i=1}^2 x_1^{(0)}(i) & \sum_{i=1}^2 x_2^{(0)}(i) & L & \sum_{i=1}^2 x_N^{(0)}(i) \\ \sum_{i=1}^3 x_1^{(0)}(i) & \sum_{i=1}^3 x_2^{(0)}(i) & L & \sum_{i=1}^3 x_N^{(0)}(i) \\ M & M & & M \\ \sum_{i=1}^M x_1^{(0)}(i) & \sum_{i=1}^M x_2^{(0)}(i) & L & \sum_{i=1}^M x_N^{(0)}(i) \end{bmatrix}$$

再计算 $x_1^{(1)}$ 的一次累差数据序列

$$\{\alpha^{(1)}(x_1^{(1)}, i)\}, \quad i = 2, 3, L, M$$

$$\{\alpha^{(1)}(x_1^{(1)}, i)\} = x_1^{(1)}(i) - x_1^{(1)}(i-1) = \sum_{k=1}^i x_1^{(0)}(k) - \sum_{k=1}^{i-1} x_1^{(0)}(k) = x_1^{(0)}(i)$$

第 2 步. 构造矩阵 $X(A, B) = B$, 即

$$\begin{bmatrix} -\frac{1}{2}(x_1^{(1)}(2) + x_1^{(1)}(1)) & x_2^{(1)}(2) & L & x_N^{(1)}(2) \\ -\frac{1}{2}(x_1^{(1)}(3) + x_1^{(1)}(2)) & x_2^{(1)}(3) & L & x_N^{(1)}(3) \\ M & M & & M \\ -\frac{1}{2}(x_1^{(1)}(M) + x_1^{(1)}(M-1)) & x_2^{(1)}(M) & L & x_N^{(1)}(M) \end{bmatrix}$$

$$Y_M = [x_1^{(0)}(2), x_1^{(0)}(3), \dots, x_1^{(0)}(M)]^T$$

第 3 步. 用最小二乘法解灰参数 \hat{a} :

$$\hat{a} = \begin{bmatrix} a \\ b_1 \\ M \\ b_{n-1} \end{bmatrix} = (B^T B)^{-1} B^T Y_M$$

第 4 步. 将灰参数代入时间函数

$$\hat{x}^{(1)}(t+1) = \left(x^{(0)}(1) - \sum_{i=2}^N \frac{b_i - 1}{a} x_i^{(1)}(t+1) \right) e^{-at} + \sum_{i=2}^N \frac{b_i - 1}{a} x_i^{(1)}(t+1)$$

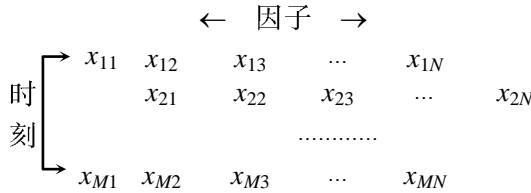
第 5 步. 将灰参数代入时间函数, 求生成数据序列计算值 $\hat{x}^{(1)}(t)$ 后, 再对 $\hat{x}^{(1)}$ 求导还原得到 $\hat{x}^{(0)}(t)$, 并计算 $x^{(0)}(t)$ 与 $\hat{x}^{(0)}(t)$ 之差 $\varepsilon^{(0)}(t)$ 以及相对误差 $e(t)$ 。

第 6 步. 在建立模型同时, 系统还将 GM(1, N)模型参数作拉普拉斯变换, 给出在零初始条件下, 第 i 个影响因素对其作用对象的动态环节传递函数 $w_i(s)$:

$$w_i(s) = \frac{x_1^{(1)}(s)}{x_i^{(1)}(s)} = \frac{b_i/a}{1+1/as}, \quad i=2, 3, \dots, N$$

43.5.2 DPS 平台的操作示例

在 DPS 平台上, 数据编辑方式是第一列为待分析的对象, 后面各列存放有关影响因素, 每一行为一个样本(或某一时刻)的数据。对有 N 个因子序列、 M 个时刻的资料, 输入顺序为



数据编辑之后将所有待分析数据定义成数据块, 在菜单下选择“其他方法” → “灰色系统” → “GM(1, N)模型分析”功能项, 按回车执行后系统给出分析结果。

如对山西省太谷县农业产值构成动态进行分析。现抽取 1978~1984 年农业产值资料, 并按图 43-6 方式进行编辑、定义, 然后在菜单方式下选择执行 GM(1, N)建模分析功能项, 最后系统将给出矩阵 B 的各个参数、GM(1, N)模型参数、观察值、拟合值、误差以及各个因素对作用对象的动态环节传递函数, 等等。

	A	B	C	D	E
1		山西太谷 1978~1984 年农业产值			
2	年份	农业总产值	农业产值	农村工业产值	
3	1978	5344	4500	734	
4	1979	6724	5279	1345	
5	1980	5968	4442	1526	
6	1981	6373	4697	1676	
7	1982	7309	5485	1824	
8	1983	8509	6019	2490	
9	1984	11542	7875	3668	
10					

图 43-6 GM(1, N)建模分析数据编辑、定义示意图

Matrix $A_{ij} =$					
- 8706.000	- 15052.000	- 21222.500	- 28063.500	- 35972.500	- 45998.000
9779.000	14221.000	18918.000	24403.000	30422.000	38297.000
2079.000	3605.000	5281.000	7105.000	9595.000	13263.000
=====系数向量=====					
$\hat{\alpha} = 2.08515$		$\hat{b}_1 = 2.13172$		$\hat{b}_2 = 1.94612$	
系统动态环节及其传递函数					
X1	1.0223				

X2	$1 + 0.4796s$				

X1	0.9333				

X3	$1 + 0.4796s$				
$\hat{x}_1(t + 1) = (5344.00000 - 1.0223x_2 - 0.9333x_3)\exp(-2.08515t) + 1.0223x_2 + 0.9333x_3$ ----- $\hat{x}_1(i)$ -----					
$\hat{x}_1(1) = 5344.00000 \dots 5344.00000$					
$\hat{x}_1(2) = 11118.23793 \dots 12068.00000$					
$\hat{x}_1(3) = 17709.20981 \dots 18036.00000$					
$\hat{x}_1(4) = 24233.02974 \dots 24409.00000$					
$\hat{x}_1(5) = 31572.98063 \dots 31718.00000$					
$\hat{x}_1(6) = 40055.60594 \dots 40227.00000$					
$\hat{x}_1(7) = 51530.76365 \dots 51769.00000$					
... \hat{x}_0	e	%	
$\hat{x}_0(2) = 5774.2379$	6724.0000	$e = 949.7621$	$q = 14.12\%$		
$\hat{x}_0(3) = 6590.9719$	5968.0000	$e = -622.9719$	$q = -10.44\%$		
$\hat{x}_0(4) = 6523.8199$	6373.0000	$e = -150.8199$	$q = -2.37\%$		
$\hat{x}_0(5) = 7339.9509$	7309.0000	$e = -30.9509$	$q = -0.42\%$		
$\hat{x}_0(6) = 8482.6253$	8509.0000	$e = 26.3747$	$q = 0.31\%$		
$\hat{x}_0(7) = 11475.1577$	11542.0000	$e = 66.8423$	$q = 0.58\%$		

从分析结果可以得到两个因素的系统影响的动态环节传递函数

$$\frac{x_1^{(1)}}{x_2^{(1)}} = \frac{1.0223}{1 + 0.4796s}, \quad \frac{x_1^{(1)}}{x_3^{(1)}} = \frac{0.9333}{1 + 0.4796s}$$

由这两个函数看出，系统在 1.5~2 年内呈摆动状态，但 1.5~2 年后会形成稳态经济状态而逐步发展。

43.6 灾变预测

设有原始数列

$$X^{(0)}(i) = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$$

若规定一个阈值 ζ , 把 $X^{(0)}(i)$ 中那些大于或小于 ζ 的点认为是异常值的点, 然后将所有符合这一条件的数据挑出来, 另外组成一个数列称为上限或下限灾变数列。

例如, 某地 4 年雨量资料 $X^{(0)}(t) = \{800(1), 300(2), 700(3), 600(4)\}$, 若规定年雨量 $\zeta \geq 600\text{mm}$ 为涝年, 并规定 $x^{(0)}(t)$ 中凡是大于 600mm 的数据是异常值, 则有

$$\begin{aligned} X_{\zeta}^{(0)}(t') &= \{x_{\zeta}^{(0)}(1'), x_{\zeta}^{(0)}(2'), x_{\zeta}^{(0)}(3')\} \\ &= \{800, 700, 600\} \\ &= \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3)\} = X^{(0)}(t) \end{aligned}$$

上述数据表明, $X_{\zeta}^{(0)}(t')$ 的数据与 $X^{(0)}(t)$ 中的数据具有下述对应关系: $X^{(0)}(t)$ 中的第 1' 个异常值 800mm 出现在第 1 点; 第 2' 个异常值 700mm 出现在第 3 点; 第 3' 个异常值 600mm 出现在第 4 个点。那么, 未来第 4' 个异常值将出现在哪个点(或称年份)上呢? 所以灾变预测不是预测数据本身的变化, 而是预测未来异常值出现的时刻。对于本例, 不是去预测未来雨量的多少, 而是预测未来的涝年将出现在哪一年。

若把垂直坐标的灾变点全体 $\{1, 3, 4\}$ 记为 q , 把水平坐标的异常值的全体 $\{1', 2', 3'\}$ 记为 i' , 则根据水平坐标 i 与垂直坐标 q 对应的点(图 43-7)。

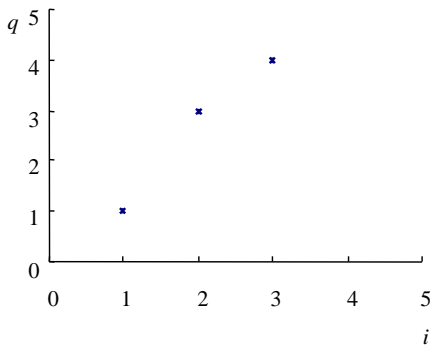


图 43-7 时序与灾变点关系图

这样可得到异常值到灾变点的映射

$$P: \{i'\} \rightarrow \{q\}$$

或记为

$$p(i) = q$$

则 $p^{(0)}(i') = \{p^{(0)}(1'), p^{(0)}(2'), \dots, p^{(0)}(n')\}$ 称为灾变日期集。灾变预测就是按灾变日期集建立 GM(1, 1) 预测模型进行的。

例如, 对某地旱年预测, 取 1957~1980 年该县年降水量(mm)的原始数据为

$$X^{(0)}(t) = \{x^{(0)}(1957), x^{(0)}(1958), \dots, x^{(0)}(1980)\}$$

$$= \{386.6, 514.6, 434.1, 484.1, 647.0, 399.7, 498.7, 701.6, 254.5, 463.0, 745.0, 398.3, 554.5, 471.1, 384.5, 242.5, 671.7, 374.7, 458.9, 511.3, 530.8, 586.0, 387.1, 454.4\}$$

规定年降水量 $\zeta \leq 390(\text{mm})$ 为异常值，则有

$$x_{\zeta}^{(0)}(t') = \{x_{\zeta}^{(0)}(1'), x_{\zeta}^{(0)}(2'), \dots, x_{\zeta}^{(0)}(6')\}$$

$$= \{386.6, 254.5, 384.5, 242.5, 374.7, 387.1\}$$

$$= \{x^{(0)}(1), x^{(0)}(9), x^{(0)}(15), x^{(0)}(16), x^{(0)}(18), x^{(0)}(23)\}$$

作异常值 i' 到出现灾变点 q 的映射 $P^{(0)}: \{i'\} \rightarrow \{q\}$ ，得灾变日期集 $p^{(0)}(i')$ 为

$$p^{(0)}(i') = \{p^{(0)}(1'), p^{(0)}(2'), p^{(0)}(3'), p^{(0)}(4'), p^{(0)}(5'), p^{(0)}(6')\}$$

$$= \{1, 9, 15, 41, 59, 82\}$$

据此对 $p^{(0)}(i')$ 建立 GM(1, 1) 模型，进行灾变预测。

灾变预测，先按图 43-8 方式进行编辑，然后在菜单方式下选择执行“灾变预测”建模分析功能项，这时系统提示输入阈值，并在阈值前加上“>”或“<”号，最后系统给出一次累加生成向量，GM(1, 1)模型参数、观察值、拟合值、模型精度检验以及预测值等。本例输出结果如下：

	A	B	C	D	E	F	G	H
1								
2		386.6	514.6	434.1	484.1	647	399.7	
3		498.7	701.6	254.5	463	745	398.3	
4		554.5	471.1	384.5	242.5	671.7	374.7	
5		458.9	511.3	530.8	586	387.1	454.4	
6		某地 1957~1980 年降雨量						

图 43-8 灾变预测数据编辑格式

灰色灾变预测			
灾变点	累加值		
1.0000	1.0000		
9.0000	10.0000		
15.0000	25.0000		
16.0000	41.0000		
18.0000	59.0000		
23.0000	82.0000		
系数向量			
A = -0.188422	U = 9.548719		
预测方程			
$X(K+1) = 51.677419 * \text{Exp}(0.18842157 * K) - 50.677419$			
$\wedge X1(K)$	$\wedge X0(K)$	理论值	相对误差%
1.0000	1.0000	1.0000	0.0000

11.7149	10.7149	9.0000	19.0547
24.6515	12.9366	15.0000	- 13.7562
40.2704	15.6189	16.0000	- 2.3820
59.1277	18.8573	18.0000	4.7630
81.8950	22.7673	23.0000	- 1.0119

阈值 ≤ 390.00 的灾变点预测

$\hat{X}_0(7)=27.4879$ 第 1 次灾变将发生在序号为 27.5 的时刻左右
即出现以后的第 3.5 个单位时间左右

$\hat{X}_0(8)=33.1873$ 第 2 次灾变将发生在序号为 33.2 的时刻左右
即出现以后的第 9.2 个单位时间左右

$\hat{X}_0(9)=40.0684$ 第 3 次灾变将发生在序号为 40.1 的时刻左右
即出现以后的第 16.1 个单位时间左右

参 考 文 献

邓聚龙. 1990. 灰色系统理论教程[M]. 武汉: 华中科技大学出版社.

刘思峰等. 1999. 灰色系统理论及其应用[M]. 北京: 科学出版社.

王学萌等. 2001. 灰色系统分析及其实用计算程序[M]. 武汉: 华中科技大学出版社.

袁嘉祖. 1991. 灰色系统理论及其应用[M]. 北京: 科学出版社.

第 44 章 多试验、多指标综合评价

44.1 离散型变量 Meta 分析

对离散型变量的资料的 Meta 分析，主要讨论四格表资料的 Meta 分析方法。四格表形式资料包括了最常见的临床试验研究、病例对照研究和队列研究。假定数据的基本格式如表 44-1 所示。

表 44-1 四格表资料的基本格式

	暴露(事件发生)	未暴露(事件未发生)	合计
病例组(处理组)	a_i	b_i	n_{1i}
对照组	c_i	d_i	n_{2i}
合计	m_{1i}	m_{2i}	T_i

在 Meta 分析中，固定效应模型在实际中应用最为广泛，适用于固定效应模型的 Meta 多种。

(1) Mantel-Haenszel 法。利用分层分析的原理，将纳入 Meta 分析的每一个研究视为一层，计算合并的优势比的方差估计采用 Robins 等提出的方法。

第 i 个研究的优势比

$$OR_i = \frac{a_i d_i}{b_i c_i}$$

每项研究的权重

$$\omega_i = \frac{b_i c_i}{T_i}$$

则合并 OR 值为

$$OR_{MH} = \frac{\sum(w_i OR_i)}{\sum w_i} = \frac{\sum(a_i d_i / T_i)}{\sum(b_i c_i / T_i)}$$

合并 OR 对数的方差为

$$\text{var}(\ln(OR_{MH})) = \frac{\sum F}{2 \sum R^2} + \frac{\sum G}{2 \sum R \sum S} + \frac{\sum H}{2 \sum S^2}$$

式中

$$F = \frac{a_i d_i (a_i + d_i)}{T_i^2}, \quad G = \frac{a_i d_i (b_i + c_i) + b_i c_i (a_i + d_i)}{T_i^2}$$

$$H = \frac{b_i c_i (b_i + c_i)}{T_i^2}, \quad R = \frac{a_i d_i}{T_i}, \quad S = \frac{b_i c_i}{T_i}$$

于是, 合并 OR 值的 95% CI 为

$$\exp\left(\ln(\text{OR}_{MH}) \pm 1.96 \sqrt{\text{var}(\ln(\text{OR}_{MH}))}\right)$$

用 Mantel-Haenszel 法进行 Meta 分析, 异质性检验的 Q 统计量可表达为

$$Q = \sum \omega'_i (\ln(\text{OR}_{MH}) - \ln(\text{OR}_i))^2 = \sum w'_i [\ln(\text{OR}_i)]^2 - \frac{[\sum w'_i \ln(\text{OR}_i)]^2}{\sum w'_i}$$

式中, $w'_i = [\text{var}(\ln(\text{OR}_i))]^{-1} = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}\right)^{-1}$ 。

零假设成立时, K 组资料的 Q 统计量服从自由度为 $K-1$ 的 χ^2 分布。

Mantel-Haenszel 法在实际中应用很多, 有较好的统计学特性, 其结果不受数据本身的影响。只要每个纳入的研究都能提供完整的四格表资料, 用 Mantel-Haenszel 法得到的效应合并值是比较精确的。

Mantel-Haenszel 法除了适用于结局变量为优势比的资料外, 还适用于结局变量是相对危险度和率差的资料, 具体的方法这里不再详述。

(2) Peto 法。也称改进的 Mantel-Haenszel 法, 适用于结局变量是优势比的资料。数据的基本格式同 Mantel-Haenszel 法。假设 O_i 为第 i 个研究病例组的实际暴露数, E_i 为第 i 个研究病例组暴露数的期望。

第 i 个研究的期望为

$$E_i = \frac{n_i m_{1i}}{T_i}$$

第 i 个研究的优势比和方差为

$$\text{OR}_i = \exp\left(\frac{O_i - E_i}{V_i}\right), \quad V_i = \frac{n_i m_{1i} n_{2i} m_{2i}}{T_i^2 (T_i - 1)}$$

则合并 OR 值为

$$\text{OR}_P = \exp\left(\frac{\sum (O_i - E_i)}{V_i}\right)$$

合并 OR 值的 95% CI 为

$$\exp\left(\ln \text{OR}_P \pm \frac{1.96}{\sqrt{\sum V_i}}\right) = \exp\left(\frac{\sum (O_i - E_i) \pm 1.96 \sqrt{\sum V_i}}{\sum V_i}\right)$$

用 Peto 法进行 Meta 分析时, 异质性检验的统计量 Q 可以表达为

$$Q = \sum \frac{(O_i - E_i)^2}{V_i} - \frac{(\sum(O_i - E_i))^2}{\sum V_i}$$

同样，零假设成立的时 Q 统计量服从自由度为 $K - 1$ 的 χ^2 分布。

当 Meta 分析异质性检验结果为研究间差异显著时，需采用随机效应模型。该模型由 DerSimonian-Laird 于 1986 年提出。D-L 方法的关键是对每个研究的权重进行修正，即

$$w_i^* = \frac{1}{D + 1/w_i}$$

式中， D 为校正因子

$$D = \frac{(Q - (k - 1)) \sum w_i}{(\sum w_i)^2 - \sum w_i^2}$$

根据校正后的权重计算合并后的 OR 值及其 95% 置信区间。

实际应用中，Peto 法的计算较为简单。在 Cochrane 协作网提供的 Meta 分析软件 Metaview 中，固定效应模型采用的就是 Peto 法。但必须注意，如果各个纳入研究的 OR 值都远偏离 1 (一般认为，OR 均大于 5 或小于 0.2)，或者四格表数据很不平衡时，用 Peto 法得到的效应合并值是有偏的。在这种情况下，应直接应用 Mantel-Haenszel 法进行效应值合并。

例如，图 44-1 阴影部分给出了 20 世纪 70 年代到 80 年代完成的有关阿司匹林预防心肌梗死后死亡的 7 个临床试验的研究结果，现希望通过 Meta 分析，综合这些资料。在 DPS 数据处理系统中，只要将数据按图 44-1 方式编辑、定义数据块，就可进行分析。

	A	B	C	D	E	F
1	试验研究批次	阿司匹林		安慰剂		
2		死亡数 a_i	存活数 b_i	死亡数 c_i	存活数 d_i	
3	1	49	566	1	49	
4	2	44	714	2	44	
5	3	102	730	3	102	
6	4	32	285	4	32	
7	5	85	725	5	85	
8	6	246	2021	6	246	
9	7	1570	7017	7	1570	
10						

图 44-1 阿司匹林预防心肌梗死后死亡的 7 个临床试验研究结果

然后在 DPS 系统中调用“数据分析”→“Meta 分析”→“四格表”分析功能项，即可得到结果如下：

组号	OR	OR 值 95%置信区间	M-H 卡方	p-值	V_i
1	0.719714	1.058131 0.489532	2.797744	0.094397	26.30475
2	0.680760	1.011356 0.458230	3.62555	0.056899	25.10748
3	0.802870	1.062608 0.606621	2.357315	0.124696	49.2971
4	0.800739	1.318184 0.486413	0.763471	0.382245	15.56546
5	0.798143	1.152406 0.552785	1.447413	0.228943	27.05887
6	1.132736	0.934726 1.372693	1.616488	0.203582	104.3238
7	0.894969	0.965685 0.829432	8.17918	0.004237	665.092
合并	0.896866	0.956965 0.840541	10.81933	0.001005	
合并卡方检验:					
方法	卡方值	df	p 值		
Peto 法	9.96782	6	0.126014		
M-H 法	10.82142	6	0.094055		
Mantel & Fleiss	样本含量检验参数=		2227.375		
随机效应模型					
异质性校正因子 $D = 0.009690$					
合并 OR	0.876192	95%置信限: 0.7523~1.0001			

从计算结果可以看出，7 个预防心肌梗死后死亡的研究资料合并的异质性检验，Peto 法的卡方值等于 9.9678， $df = 6$ ， $p = 0.1260$ 。因 $p > 0.05$ ，故可认为 7 个研究来自同一总体，即效应的总体水平相同。直接应用 Mantel-Haenszel 法进行效应值合并检验的结论相同。

合并后的 OR 值等于 0.8969，其 95% 置信限为 0.840541~0.956965。

对合并后的 OR 值进行检验，其卡方值等于 10.81933， $df = 1$ ， $p < 0.01$ ，即阿司匹林能减少病人心肌梗死后死亡的危险。

若应用随机效应模型，得到的合并后 OR 值等于 0.8762，其 95% 置信限为 0.7523~1.0001。结论与固定效应模型给出的结果相同，但随机效应模型的结果趋于保守。

44.2 连续型变量资料分析

对连续型变量资料的 Meta 分析，主要介绍以标准化的均数之差作为效应尺度的情形。如前所述，用 n_{1i} ， n_{2i} 表示纳入 Meta 分析的第 i 个研究处理组(暴露组)和对照组的样本大小， \bar{x}_{1i} ， \bar{x}_{2i} 表示第 i 个研究处理组(暴露组)和对照组的均数， s_{1i}^2 ， s_{2i}^2 为第 i 个研究处理(暴露组)和对照组的方差，两组的合并方差为 s_{pi}^2 ，令 $y_i = \bar{x}_{1i} - \bar{x}_{2i}$ ， y_i 的标准误 s_i 为

$$s_i^2 = s_{pi}^2 \left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}} \right)$$

而

$$s_{pi}^2 = \frac{(n_{1i} - 1)s_{1i}^2 + (n_{2i} - 1)s_{2i}^2}{n_{1i} + n_{2i} - 2}$$

(1) 固定效应模型。当纳入的研究结局变量的量纲相同时，可直接用两组的均数之差 y_i 作为效应，则效应合并值为

$$y_{\text{合并}} = \frac{\sum w_i y_i}{\sum w_i}, \quad w_i = \frac{1}{s_i^2}$$

最后计算异质性检验的统计量 Q :

$$Q = \sum w_i (y_{\text{合并}} - y_i)^2 = \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2}{\sum w_i}$$

效应合并值的 95% CI 为

$$y_{\text{合并}} \pm 1.96 \sqrt{\sum w_i}$$

(2) 随机效应模型。如果异质性检验拒绝 H_0 ，宜采用下面的随机效应模型。效应尺度的计算同前，即

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{s_{p_i}}$$

d_i 的加权均数和加权方差的估计值为

$$\bar{d} = \frac{\sum w_i d_i}{\sum w_i} J_x$$

这里， J_x 是校正系数

$$J_x = 1 - \frac{3}{4(n_{1i} + n_{2i} - 2) - 1}$$

$$s_d^2 = \frac{\sum w_i (d_i - \bar{d})^2}{\sum w_i} = \frac{\sum w_i d_i^2}{\sum w_i} - \bar{d}^2$$

权重 $w_i = N_i$ 为各个研究的样本含量。

由随机效应模型 $d_i = \delta_i + e_i$ ， δ_i 和 e_i 的均数和方差分别为

$$\bar{\delta} = \bar{d}, \quad \bar{e} = 0, \quad s_e^2 = \frac{4k}{\sum w_i} \left(1 + \frac{\bar{d}^2}{8} \right)$$

① 若 $s_d^2 > s_e^2$ ，则 $s_\delta^2 = s_d^2 - s_e^2$ ，此时效应尺度的合并值的 95% CI 为 $\bar{d} \pm 1.96 s_\delta$ 。

② 若 $s_d^2 \leq s_e^2$ ，则 $s_\delta^2 = 0$ ，此时随机效应模型退化为固定效应模型 $d_i = \delta + e_i$ ， \bar{d} 的标准误差为 $s_{\bar{d}} = s_e / \sqrt{k}$ ，则效应尺度的合并值的 95% CI 为 $\bar{d} \pm 1.96 s_{\bar{d}}$ 。

在连续型变量的随机效应模型中，异质性检验可以采用统计量

$$x^2 = \frac{k s_d^2}{s_e^2}$$

H_0 成立时，上述统计量服从自由度为 $df = k - 1$ 的 χ^2 分布。

例如，有人进行了 11 项高氟区和适氟区女童 II 掌骨皮质厚度的研究，其结果如图 44-2，试进行 Meta 分析。

	A	B	C	D	E	F	G	H
1	研究	样本数量		平均值		标准差		
2	批次	高氟区	适氟区	高氟区	适氟区	高氟区	适氟区	
3	1	26	42	2.26	2.33	0.32	0.33	
4	2	55	40	2.39	2.49	0.31	0.32	
5	3	46	50	2.5	2.67	0.3	0.35	
6	4	45	50	2.64	2.9	0.26	0.45	
7	5	45	45	2.81	2.93	0.35	0.36	
8	6	52	55	2.95	3.27	0.46	0.37	
9	7	46	42	3.15	3.48	0.39	0.48	
10	8	45	51	3.47	3.73	0.46	0.54	
11	9	45	45	3.63	3.81	0.38	0.4	
12	10	42	45	3.81	4.16	0.41	0.42	
13	11	44	25	3.99	4.18	0.56	0.41	
14								

图 44-2 高氟区和适氟区女童 II 掌骨皮质厚度的 11 个研究结果

在 DPS 数据处理系统中，只要将数据按图 44-2 方式，编辑、定义数据块，就可进行分析。

然后在 DPS 系统中调用“数据分析”→“Meta 分析”→“单组资料分析”功能项，即可得到如下结果：

计算结果	当前日期	02-9-8 20: 20: 50				
试验号	样本	<i>d</i>	<i>W</i>	<i>wd</i>	<i>wdd</i>	
1	68	-0.212113	15.97396	-3.38829	0.718700	
2	95	-0.315663	22.88002	-7.22237	2.279831	
3	96	-0.515689	23.18883	-11.95822	6.16671	
4	95	-0.692542	22.34811	-15.47700	10.71847	
5	90	-0.335106	22.18854	-7.43551	2.491682	
6	107	-0.763443	24.91519	-19.02133	14.52171	
7	88	-0.751580	20.50939	-15.41445	11.58520	
8	96	-0.511608	23.15173	-11.84461	6.05980	
9	90	-0.457443	21.92647	-10.03012	4.58821	
10	87	-0.835494	19.98260	-16.69535	13.94887	
11	69	-0.367389	15.69727	-5.76701	2.118737	
合计	981		232.7621	-124.2543	75.1979	
固定效应模型						
合并效应 <i>d</i> =		-0.5338				
95%置信限: -0.6623 ~ -0.4054						
异质性检验 <i>Q</i> =		8.86790				
df = 10	<i>p</i> = 0.5447					

如果研究间同质性不好, 宜采用下面的随机效应模型

合并效应 $d = -0.5362$

95% 置信限: $-0.7113 \sim -0.3611$

从计算结果可以看出, 11 个试验研究资料合并的异质性检验 $Q = 8.8679$, $df = 10$, $p = 0.5447$ 。因 $p > 0.05$, 故可认为 11 个研究来自同一总体, 即效应的总体水平相同。合并效应 $d = -0.5338$, 其 95% 置信限为 $-0.4054 \sim -0.6623$ 。

Meta 分析结果表明, 高氟区女童骨皮质平均厚度比适氟区低 0.533, 95% 置信限为 $-0.4054 \sim -0.6623$, 均小于 0。

若应用随机效应模型, 得到的 Meta 分析结果也表明, 高氟区女童骨皮质平均厚度比适氟区低 0.536, 95% 置信限为 $-0.3611 \sim -0.7113$, 均小于 0。结论与固定效应模型给出的结果相同, 但随机效应模型的结果趋于保守。

44.3 含亚类资料 Meta 分析

前面介绍的 Meta 分析, 是将多个试验研究结果合并在一起。但有时, 当我们进行 Meta 分析的目的是对几个大类分别综合后再统一合并, 这在生态学研究是常见的现象, 如将要进行比较的对象分成 3 类, 须先按子类合并后再继续合并(综合)。当研究资料收集得到之后, 从这些试验中获取的基本信息列成表格。对于每一个研究, 首先必须计算无偏效应值(d_{ij})

$$d_{ij} = \frac{\bar{X}_{ij}^1 - \bar{X}_{ij}^2}{S_{ij}} J_x, \quad J_x = 1 - \frac{3}{4(N_{ij}^1 + N_{ij}^2 - 2) - 1}$$

式中, d 是指第 i 类中的第 j 个研究计算值, $\bar{X}_{ij}^1 =$ 对照平均值, $\bar{X}_{ij}^2 =$ 处理平均值, $S_{ij} =$ 对照和处理的组内标准差

$$S_{ij} = \sqrt{\frac{(N_{ij}^1 - 1)(s_{ij}^1)^2 + (N_{ij}^2 - 1)(s_{ij}^2)^2}{N_{ij}^1 + N_{ij}^2 - 2}}$$

式中, $N_{ij}^2 =$ 对照总个体数, $N_{ij}^1 =$ 处理总个体数, $S_{ij}^2 =$ 对照标准差, $S_{ij}^1 =$ 处理标准差。注意, 可以用研究的条件、试验和比较相互替换来表示一个简单处理与其对照之间的比较(Hedges, Olkin, 1985)。抽样越大 J_x 越趋近于 1。最后, 第 i 组的第 j 个研究结果的方差可以近似的用 v_{ij} 来表示

$$v_{ij} = \frac{N_{ij}^1 + N_{ij}^2}{N_{ij}^1 N_{ij}^2} + \frac{d_{ij}^2}{2(N_{ij}^1 + N_{ij}^2)}$$

(Hedges, Olkin, 1985), 这一项可以直接用来比较结果的大小, 估计结果的重要性(如结

果是否显著大于 0)和将整个研究结果结合起来。

结合固定模型中的结果,在第 i 组的结果的累加值 d_{ij} 是进行组内结果估计的重要平均值。被用来结合结果的砝码 w_{ij} 与抽样误差是相互的, $w_{ij} = 1/v_{ij}$, 真值 δ_i 的重要估计是

$$d_{i+} = \sum_{j=1}^{k_i} w_{ij} d_{ij} / \sum_{j=1}^{k_i} w_{ij}$$

式中, k_i 等于第 i 组内比较结果的个数(Hedges, Olkin, 1985)。

假设较大抽样能产生比较精确的结果(Hedges, Olkin, 1985)。那么在较难计算结果的较大研究中, 累积值是一个很好的平均值。 d_{i+} 的方差 $S(d_{i+})^2$ 等于

$$s_{(d_{i+})}^2 = 1 / \sum_{j=1}^{k_i} w_{ij}$$

d_{i+} 的 95% 置信区间的置信上限和置信下限分别用 d^U 和 d^L 来表示

$$d^L = d_{i+} - [Z_{\alpha/2} s_{(d_{i+})}]$$

$$d^U = d_{i+} + [Z_{\alpha/2} s_{(d_{i+})}]$$

z 是标准正态分布的两尾临界值。总体平均值 d_{++} 为

$$d_{++} = \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij} d_{ij} / \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij}$$

m 是总组数, 总体平均值方差

$$s_{(d_{++})}^2 = 1 / \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij}$$

比较各组间的差异, 可用统计值 Q 来测验(Q 近似等于 x^2 除以一个比研究数小 1 的自由度)。 Q 值越大, 进行比较的结果之间的差异越大。类似 ANOVA 分解方差, 可将总差异 Q_T 分解成组内差异 Q_W 和组间差异 Q_B , 即

$$Q_B + Q_W = Q_T$$

组内差异

$$Q_W = \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij} (d_{ij} - d_{++})^2$$

Q_W 是用来测量组内变异和程度, 而组间差异 Q_B 是用一个平均来测量组间进行测量的

$$Q_B = \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij} (d_{i+} - d_{++})^2。$$

Q_B 等于 x^2 除以总组数减去 1 的自由度(Hedges, Olkin, 1985)。统计值 Q_T 可按公式

$$Q_T = \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij} d_{ij}^2 - \left(\sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij} d_{ij} \right)^2 / \sum_{i=1}^m \sum_{j=1}^{k_i} w_{ij}$$

计算。

总组内差异值 Q_W 可以用各组的组内差异值 $Q_{W1}, Q_{W2}, \dots, Q_{Wi}$ 的总和来计算。每组的组内差异值 Q_{Wi} 可以用下面的公式来进行计算:

$$Q_{Wi} = \sum_{j=1}^{k_i} w_{ij} d_{ij}^2 - \left(\sum_{j=1}^{k_i} w_{ij} d_{ij} \right)^2 / \sum_{j=1}^{k_i} w_{ij}$$

自由度为 $k-1$, $Q_B=Q_T-Q_W$ 。这些统计量的统计意义用一个标准 χ^2 表来进行估价。我们对前两类结果的意义解释是类似的, 它们结果大小的不同可能是由于随机抽样误差的存在(但我们可能没有办法找出它们之间的真正差异, 我们都看到一些用较小数量研究而得到显著的 Q 统计值的例子)。在对第 3 类的研究中明显显露出方差大于抽样误差。

d_{ij} 的固定效应方差 v_{ij} 实际上是一个条件方差, 因为它是在假设同一组中的所有研究都有一个共同的真值 σ 的条件下进行的。 d_{ij} 的无条件方差 v_{ij}^* (也被看作混合模型的方差), 它是假设在研究中因受兴趣的影响而存在随机误差, 因此就没有一个共同的真值。为了得到 v_{ij}^* , 必须增加常数项 C_i 作为处理组内方差 $\hat{\sigma}_{\text{pooled}}^2$ 的一部分

$$C_i = \sum_{j=1}^{k_i} w_{ij} d_{ij}^2 - \sum_{j=1}^{k_i} w_{ij}^2 / \sum_{j=1}^{k_i} w_{ij}$$

i 是第 i 组, k_i 是第 i 组的试验数, 然后我们计算方差 σ_{pooled}^2 的估计值

$$\sigma_{\text{pooled}}^2 = \left[Q_W - \sum_{i=1}^m (k_i - 1) \right] / \sum_{i=1}^m C_i$$

m 是组数, Q_W 是固定模型分析所得到的组内差异。

Meta 分析与随机误差估计是相互的(正如固定模型中的砝码与固定模型效应的方差估计的关系)。

在混合模型中每一类的累加结果 d_{++}^* 和它们的方差 $s_{(d_{++}^*)}^2$ 的计算与固定模型中的一样, 总体累加结果 d_{++}^* 和它的方差 $s_{(d_{++}^*)}^2$, s 可以用下列公式计算:

$$d_{++}^* = \sum_{i=1}^m w_{ij}^* d_{i+}^* / \sum_{i=1}^m w_{i+}^*, \quad s_{(d_{++}^*)}^2 = 1 / \sum_{i=1}^m w_{i+}^*$$

最后, 组间差异可用下面的公式来求得

$$Q_B^* = \sum_{i=1}^m w_{i+}^* d_{i+}^{*2} - \left(\sum_{i=1}^m w_{i+}^* d_{i+}^* \right)^2 / \sum_{i=1}^m w_{i+}^*$$

在混合模型中，我们无法计算 Q_w 值，因为我们没有假设同一组内的所有研究具有一个共同的真值。组内结果差异的测量，用固定模型中的 Q_w 来代替，然而，在固定模型中的解释是组内方差部分 σ^2 大于零，这个检验很少有用是因为它总是忽略 $\sigma^2 = 0$ 的假设，所以是不可应用的。

例如，关于竞争效应的研究，原作者将整个试验要进行比较的对象分成 3 类(陆生生物、淡水生物和海洋生物)。从这些试验中获取的基本信息列于图 44-3。

1	A	B	C	D	E	F	G	H	I
		类别	N^C	\bar{X}^C	S^C	N^E	\bar{X}^E	S^E	
2		1	7	79.71	40.65	7	78.14	40.65	
3		1	7	26	9.17	7	18.86	9.17	
4		1	6	2.1	0.49	6	1.8	0.49	
5		1	5	2.8	0.447	5	2.2	0.224	
6		1	7	3	0.529	7	2.1	0.265	
7		1	6	4.2	1.225	6	2.3	0.49	
8		1	3	285.7	153.806	3	85.3	115.008	
9		1	3	3	2.425	3	0	0	
10		1	3	2	2.078	3	0	0	
11		1	3	1.67	1.732	3	0	0	
12		1	5	17	5.367	5	17	7.603	
13		1	5	37	9.391	5	47	10.286	
14		1	4	272	183.532	4	87	37.712	
15		1	20	0.294	0.215	18	0.113	0.255	
16		1	20	0.412	0.218	20	0.163	0.588	
17		1	20	0.632	0.359	18	0.14	0.38	
18		1	20	0.259	0.238	20	0.184	0.326	
19		1	20	0.354	0.182	20	0.075	0.487	
20		1	20	0.541	0.299	20	0.147	0.34	
21		2	4	201.03	27.52	4	281.11	158.038	
22		2	4	155.32	41.252	4	187.31	80.163	
23		3	7	16	3.37	7	11.8	3.08	
24		3	10	9.5	7.23	3	0.4	1.47	
25		3	20	14.1	7.603	20	0	7.603	
26		3	20	7.1	3.13	20	0	3.13	
27		3	20	1.4	1.789	20	0	1.789	
28		3	10	94	9.171	10	82.2	29.093	
29		3	10	10.5	11.068	10	8.3	15.546	
30		3	10	20	42.691	10	0	42.691	
31		3	2	18.5	4.257	2	3.63	3.352	
32		3	2	0.25	0.354	2	0	0	
33		3	2	2.25	0.707	2	3.63	3.352	
34		3	4	34.8	58.2	4	0	0	
35		3	4	25.3	35.8	4	0	0	
36		3	4	23.6	47	4	5.4	10.88	
37		3	4	10.5	24.2	4	1.8	5.2	
38		3	4	10.3	17.4	4	0	0	
39		3	4	8.7	17	4	0	0	
40		3	4	5.7	14	4	0	0	
41		3	4	5.4	8	4	10.8	15.8	
42		3	4	38.25	22.02	4	21.25	9.54	
43		3	4	20.25	9	4	40.25	8.78	
44		3	5	11.9533	6.24	5	15.8445	10.787	
45									

图 44-3 竞争效应的研究结果 Meta 分析数据编辑

图 44-3 中第 1 列为资料类别, N^C , N^E 为对照和处理的个体数, \bar{X}^C , \bar{X}^E 为对照和处理的均值, S^C , S^E 为对照和处理的标准差。在 DPS 数据处理系统中, 只要将数据按图 44-3 方式编辑、定义数据块, 就可进行分析。

然后在 DPS 系统中调用“数据分析”→“Meta 分析”→“单组资料分析”功能项, 即可得到如下结果:

计算结果				
当前日期 02-9-3 22: 44: 59				
试验号	样本	d	w	wd
1, 1	14	0.036157	3.49943	0.126529
1, 2	14	0.728926	3.28202	2.392350
1, 3	12	0.565149	2.884826	1.630357
1, 4	10	1.532871	1.932424	2.962157
1, 5	14	2.013897	2.322537	4.67735
1, 6	12	1.879929	2.080781	3.91172
1, 7	6	1.180562	1.277448	1.508107
1, 8	6	1.399634	1.204943	1.686480
1, 9	6	1.088904	1.306377	1.422518
1, 10	6	1.090871	1.305767	1.424424
1, 11	10	0.000000	2.500000	0.000000
1, 12	10	-0.917104	2.262167	-2.074643
1, 13	8	1.214219	1.688774	2.050542
1, 14	38	1.697460	6.97018	11.83160
1, 15	40	1.270935	8.32009	10.57430
1, 16	38	1.305129	7.81444	10.19885
1, 17	40	1.521305	7.75617	11.79950
1, 18	40	1.143773	8.59456	9.83022
1, 19	40	1.206196	8.46121	10.20588
2, 1	8	3.69609	0.738653	2.730125
2, 2	8	4.67364	0.536141	2.505729
3, 1	14	1.217968	2.952513	3.59607
3, 2	13	1.288499	2.011310	2.591571
3, 3	40	1.817686	7.07715	12.86404
3, 4	40	2.223304	6.18091	13.74204
3, 5	40	0.767013	9.31499	7.14471
3, 6	20	0.523947	4.83412	2.532820
3, 7	20	0.156146	4.98481	0.778358
3, 8	20	0.448688	4.87726	2.188368
3, 9	4	2.217814	0.619258	1.373398
3, 10	4	0.570708	0.960879	0.548381
3, 11	4	-0.325537	0.986926	-0.321281
3, 12	8	0.735315	1.873385	1.377528
3, 13	8	0.869070	1.827468	1.588197
3, 14	8	0.463933	1.947601	0.903557
3, 15	8	0.432235	1.954359	0.844742
3, 16	8	0.727956	1.875750	1.365464
3, 17	8	0.629343	1.905653	1.199310
3, 18	8	0.500684	1.939233	0.970944
3, 19	8	-0.374969	1.965457	-0.736985
3, 20	8	0.871155	1.826711	1.591349
3, 21	8	-1.956127	1.352901	-2.646448
3, 22	10	-0.398853	2.451256	-0.977692
合计	689		142.4588	143.9125

类 1	364		75.4641	86.1583		
类 2	16		1.274794	5.23585		
类 3	309		65.7199	52.5184		
固定效应模型:						
类	合并效应	95%置信限:	<i>Q</i>	df	<i>p</i>	
类 1	1.141711	0.9161 ~ 1.3673	25.59046	18	0.109497	
类 2	4.10722	2.3713 ~ 5.8432	0.296865	1	0.585854	
类 3	0.799125	0.5574 ~ 1.0409	43.7017	21	0.002555	
总的	1.010204	0.8460 ~ 1.1744	86.0494	42	0.000072	
类间			16.46035	2	0.000266	
混合效应模型:						
类 1	1.083759	0.7571 ~ 1.4104				
类 2	4.11639	2.2625 ~ 5.9703				
类 3	0.704125	0.3767 ~ 1.0315				
类间			14.01390	2	0.000906	

从分析结果可以看出，就整个 3 类初级生产者看来，在这 3 组之间有很明显的不同 ($Q_B = 16.4604$ ，自由度为 2， $p < 0.001$)。因此，可以从这些研究中得出这样的结论：当就生长和繁殖来说，陆地植物、淡水植物和海洋植物对竞争能力有统计性的明显不同。尽管陆地植物和海洋植物的置信限相同，但淡水植物的置信区间与它们都有不相同。这就表明淡水植物比彼此相等的陆地植物和海洋植物竞争能力都要强。

44.4 Topsis 法

Topsis(technique for order preference by similarity to ideal solution)法是系统工程有限方案多目标决策分析的一种常用方法，可用于效益评价、决策、管理多个领域。

Topsis 法的基本思想是：基于归一化后的原始数据矩阵，找出有限方案中的最优方案和最劣方案(分别用最优向量和最劣向量表示)，然后分别计算诸评价对象与最优方案和最劣方案的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。

具体方法为：设有 n 个评价对象、 m 个评价指标，原始数据可写为矩阵 $X = (X_{ij})_{n \times m}$ 。对高优(越大越好)、低优(越小越好)指标分别进行归一化变换，即

$$Z_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^n X_{ij}^2}}$$

或

$$Z_{ij} = \frac{1/X_{ij}}{\sqrt{\sum_{i=1}^n (1/X_{ij})^2}}$$

归一化得到矩阵 $Z = (Z_{ij})_{n \times m}$ ，其各列最大、最小值构成的最优、最劣向量分别记为

$$Z^+ = (Z_{\max 1}, Z_{\max 2}, \dots, Z_{\max n}), \quad Z^- = (Z_{\min 1}, Z_{\min 2}, \dots, Z_{\min m})$$

第 i 个评价对象与最优、最劣方案的距离分别为

$$D_i^+ = \sqrt{\sum_{j=1}^m (Z_{\max j} - Z_{ij})^2} \quad \text{和} \quad D_i^- = \sqrt{\sum_{j=1}^m (Z_{\min j} - Z_{ij})^2}$$

第 i 个评价对象与最优方案的接近程度 C_i (值越大综合效益越好) 为 $C_i = D_i^- / (D_i^+ + D_i^-)$ 。

某儿童医院 1994~1998 年 7 项指标的实际值如图 44-4 所示, 用 Topsis 法比较该医院这 5 年的医疗质量, 分析前将数据编辑如图 44-4 格式。

	A	B	C	D	E	F	G	H	I
1	年份	出院人数	病床使用率	平均住院日	病死率	抢救成功率	治愈好转率	院内感染率	
2	1994	21584	76.6	7.3	1.01	78.3	97.5	2	
3	1995	24372	86.3	7.4	0.8	91.1	918	2	
4	1996	22041	81.8	7.3	0.62	91.1	97.3	3.2	
5	1997	21115	84.5	6.9	0.6	90.2	97.7	2.9	
6	1998	24633	90.3	6.9	0.25	95.5	97.9	3.6	
7									

图 44-4 某儿童医院 1994~1998 年 7 项指标的实际值

然后执行“其他方法”下面的“Topsis 法”功能, 这时系统出现供用户对各个指标特性进行设置的对话框(图 44-5)。这里可根据每个指标的特性, 设置该指标是否是“低优指标”以及该指标所占权重。本例中, 平均住院日、病死率、院内感染率, 即 x_3 , x_4 和 x_7 为低优指标, 其余为高优指标, 各个指标的权重均为 1, 这时用户综合评价参数设置如图 44-5 所示。



图 44-5 Topsis 分析指标特性分析对话框

完成设置, 点击“确认”按钮后, 系统给出分析结果:

变换矩阵							
No.	x1	x2	x3	x4	x5	x6	x7
N1	0.4234	0.4081	0.4380	0.2024	0.3916	0.4464	0.5612
N2	0.4781	0.4592	0.4321	0.2556	0.4556	0.4487	0.5612
N3	0.4324	0.4353	0.4380	0.3298	0.4556	0.4455	0.3508
N4	0.4142	0.4496	0.4634	0.3408	0.4511	0.4473	0.3871
N5	0.4833	0.4805	0.4634	0.8178	0.4776	0.4482	0.3118
最优向量	0.4833	0.4805	0.4634	0.8178	0.4776	0.4487	0.5612
最劣向量	0.4142	0.4081	0.4321	0.2024	0.3916	0.4455	0.3118
各个样本排序指标值							
样本	D+	D-	指标 CI	名次			
N1	0.6289	0.2497	0.2842	3			
N2	0.5640	0.2754	0.3281	2			
N3	0.5369	0.1514	0.2200	5			
N4	0.5141	0.1762	0.2552	4			
N5	0.2494	0.6302	0.7164	1			

分析结果包括各变量归一化变换后的矩阵，各列最大、最小值构成的最优、最劣向量，分别为

$$Z_+ = (0.4833 \quad 0.4805 \quad 0.4634 \quad 0.8178 \quad 0.4776 \quad 0.4487 \quad 0.5612)$$

$$Z_- = (0.4142 \quad 0.4081 \quad 0.4321 \quad 0.2024 \quad 0.3916 \quad 0.4455 \quad 0.3118)$$

计算与最优、最劣向量的距离及与最优向量的接近程度，如 1994 年

$$D_1^+ = \sqrt{(0.4833 - 0.4234)^2 + L + (0.5612 - 0.5612)^2} = 0.6289$$

$$D_1^- = \sqrt{(0.4142 - 0.4234)^2 + L + (0.3118 - 0.5612)^2} = 0.2497$$

$$C_1 = 0.2497 / (0.6289 + 0.2497) = 0.2842$$

依 C_i 值的大小，对各个样本排序结果。可以看出，1998 年的综合效益最好，其次为 1995 年，随后为 1994 年、1997 年，1996 年最差。

44.5 综合指数法

用来测定一个或一组变量对某个特定变量值大小的相对数，称为指数。如环境质量指数，是以地形、地貌、水文、气象、生态、污染源、污染物等环境监测数据计算出的无量纲相对数，用以反映环境质量。反映某一事物或现象动态变化的指数，称为个体指数；综合反映多种事物或现象动态平均变化程度的指数，称为总指数。综合指数是编制总指数的基本计算形式，能定量地反映几个指标的综合平均变动程度。利用综合指数的

计算形式, 定量地对某现象进行综合评价的方法称为综合指数法(synthetic index)。综合指数法可用于社会、环境评价, 工作效率评价等。

高优指标的个体指数 p , 用实测值 X 与标准值 M 的商计算, 即 $p=X/M$;

低优指标的个体指数 p , 可用标准值 M 与实测值 X 的商计算, 即 $p=M/X$;

综合指数 I 较为复杂, 没有统一的表达形式, 可根据实际问题确定计算模式, 可表示为各个指标的相加或相乘, 如取相加, 则有

$$I = \frac{1}{n} \sum_{i=1}^m y_i \quad \text{或} \quad I = \sum_{i=1}^k \prod_{j=1}^l y_{ij}$$

式中, y 为个体指数, m 为指标数, n 为分组数, k 为指标类别数, l 为各类内的指标数。

模型建立后, 要用已知评价结果的历史资料计算总体指数, 对比符合程度。

例如, 某医院根据 11 项指标的标准值、实际值(图 44-6), 试用平均指数法对该医院 1994~1998 年的工作质量进行综合评价。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	指标类型	医疗质量			工作质量			动态指标		病床使用			
2	指标名称	治疗率	病死率	成功率	感染率	愈合率	死亡率	门诊量	出院人数	使用率	周转率	住院日	
3	标准值	95	2	84	10	95	0.5000	741	16452	85	17	20	
4	1994年	94.70	2.30	88.10	5.20	97.60	0.23	606.0	15236	89.8	30.6	10.6	
5	1995年	95.30	1.70	89.50	4.50	98.80	0.19	727.0	15201	86.6	30.1	10.3	
6	1996年	95.00	1.60	87.80	4.10	98.00	0.15	776.0	16940	90.1	33.5	9.7	
7	1997年	95.40	1.30	90.80	3.60	97.10	0.18	778.0	17101	82.8	32.3	9.1	
8	1998年	95.10	1.30	88.40	3.60	98.30	0.17	814.0	17783	82.1	33.2	8.1	
9													

图 44-6 某医院 1994~1998 年 11 项指标的标准值、实际值

执行综合评价, 数据编辑方式是, 数据块的第 1 行是各个指标的标准值, 以后各行为各个样本实际值。按图 44-7 方式编辑并选中数据后, 执行“其他方法”下面的“综合评价法”功能, 这时系统出现供用户对各个指标特性进行设置的对话框(图 44-7)。这里可根据每个指标的特性, 设置该指标是否是“低优指标”、该指标的标准值、所属组别(亚类)。本例中 11 项指标中, 病死率、院内感染率、新生儿死亡率、平均住院日为低优指标, 按 M/X 计算 p ; 其余为高优指标, 按 X/M 计算 p 值。

这里评价指标共 4 个亚类, 则需在各个指标后面填入所属亚类序号(如果没有分亚类, 则组别值均为 1)。

图 44-7 左下角的两个选择按钮, 第 1 个按钮表示各个亚类间(组间)的指标是采用相加还是相乘方式进行综合; 第 2 个按钮表示各个亚类内(组内)指标是采用相加还是相乘方式进行综合。这里我们对各个指标采取组间亚类指标相加, 组内各个指标相乘的综合方法。



图 44-7 综合评价用户对话框

完成设置后点击“确认”按钮，这时系统给出分析结果如下：

各个样本指数											
No.	治疗率	病死率	成功率	感染率	愈合率	死亡率	门诊量	出院人数	使用率	周转率	住院日
	优劣	高优	低优	高优	低优	高优	低优	高优	高优	高优	低优
标准	95.00	2.00	84.00	10.00	95.00	0.50	741.00	16452.00	85.00	17.00	20.00
N1	0.9968	0.8696	1.0488	1.9231	1.0274	2.1739	0.8178	0.9261	1.0565	1.8000	1.8868
N2	1.0032	1.1765	1.0655	2.2222	1.0400	2.6316	0.9811	0.9240	1.0188	1.7706	1.9417
N3	1.0000	1.2500	1.0452	2.4390	1.0316	3.3333	1.0472	1.0297	1.0600	1.9706	2.0619
N4	1.0042	1.5385	1.0810	2.7778	1.0221	2.7778	1.0499	1.0394	0.9741	1.9000	2.1978
N5	1.0011	1.5385	1.0524	2.7778	1.0347	2.9412	1.0985	1.0809	0.9659	1.9529	2.4691
各个样本综合指标											
No.	综合指标	亚类 I1	亚类 I2	亚类 I3	亚类 I4						
N1	9.5495	0.9091	4.2950	0.7574	3.5880						
N2	11.7486	1.2575	6.0819	0.9065	3.5028						
N3	15.0785	1.3065	8.3868	1.0783	4.3069						
N4	14.7157	1.6700	7.8866	1.0914	4.0677						
N5	15.9194	1.6208	8.4537	1.1874	4.6576						

分析结果中综合指数的计算，如 1994 年的计算方法为

$$I_1 = 0.9968 \times 0.8696 \times 1.0488 \times 1.9231 \times 1.0274 \times 2.1739$$

$$0.9811 \times 0.9240 \times 1.0188 \times 1.7706 \times 1.9417 = 9.5495$$

类似计算 1995~1998 年的综合指数得到

$$I_2=11.7486, \quad I_3=15.0785, \quad I_4=14.7157, \quad I_5=15.9194$$

各年的综合指数呈上升趋势, 可以认为自 1994 年以来, 该医院的工作状况逐步进入佳境。

44.6 投影寻踪分类

投影寻踪分类(projection pursuit clustering, PPC)法是 Friedman 和 Turkey 于 1974 年提出的一种既可作探索性分析, 又可作确定性分析的聚类 and 分类分析方法。所谓投影实质上就是从不同的角度去观察数据, 寻找能够最大程度地反映数据特征和最能充分挖掘数据信息的最优投影方向。PPC 法是一种可用于高维数据分析的、有效的降维技术, 适用于高维、非线性、非正态问题的分析和处理, 评价结果与实际相符率高, 已经广泛应用于水质评价、大气环境质量综合评价、灾情评估、工业经济、企业竞争力等方面。

PPC 法的特点是在未知权重系数的情况下, 通过把高维数据投影到低维(1~3 维)子空间上, 对于投影到的构形, 采用投影指标函数来衡量投影暴露某种结构的可能性大小, 寻找出使投影指标函数达到最优(即能反映高维数据结构或特征)的投影值, 然后根据该投影值来分析高维数据的结构特征, 或根据该投影值与研究系统的输出值之间的散点图构造数学模型以预测系统的输出。它避免了专家打分的人为干扰因素, 省去了利用专家打分评定的步骤, 更为准确和便捷, 因而在定量评价指标数据的处理上更具有优势。

PPC 法的计算步骤如下:

(1) 样本评价指标集的归一化处理, 计算公式为

$$x(i, j) = \frac{x^*(i, j) - x_{\min}(j)}{x_{\max}(j) - x_{\min}(j)}$$

式中, $x^*(i, j)$, $i = 1, \dots, n$; $j = 1, \dots, p$ 为第 i 个样本的第 j 个评价指标值, n , p 分别为样本的数目和评价指标的数目; $x_{\max}(j)$ 与 $x_{\min}(j)$ 分别为样本集中第 j 个评价指标的最小值和最大值。

(2) 构造投影目标函数 $Q(a)$ 。实质是把 p 维数据 $x(i, j)$, $j = 1, \dots, p$ 综合成为综合成 $a = (a(1), a(2), \dots, a(p))$ 为投影方向的一维投影值 $z(i)$ 。计算公式为

$$z(i) = \sum_{j=1}^n a(j)x(i, j), \quad i = 1, \dots, p$$

式中 a 为单位长度向量。

在综合投影值时, 要求投影值 $z(i)$ 的散布特征应为: 局部投影点尽可能密集, 最好凝聚成若干个点团, 而在整体上投影点团之间尽可能散开。基于此, 投影指标函数可构造为

$$Q(a) = S_z D_z$$

式中, S_z 为投影值 $z(i)$ 的标准差, D_z 为投影值 $z(i)$ 的局部密度。计算公式分别为

$$S_z = \sqrt{\sum_{i=1}^n z(i) - \bar{z} / (n-1)}$$

$$D_z = \sum_{i=1}^n \sum_{j=1}^n (R - r_{ij}) u(R - r_{ij})$$

$$r_{ij} = |z(i) - z(j)|$$

式中， a 为单位长度向量； \bar{z} 为系列 $z(i), i = 1, \dots, n$ 的均值； R 为求局部密度的窗口半径。 u_t 为单位阶跃函数，当 $t = (R - r_{ij}) \geq 0$ 时其函数值为1，当 $t < 0$ 时其函数值为0。在此需指出的是，局部密度的窗口半径(R)取值既要使包含在窗口内的投影点的平均个数不能太少，避免滑动平均偏差太大，又不能使它随着指标数(n)的增大而增加太高，为此，密度阈值(α)可以取 $0.1S_z, 0.01S_z, 0.001S_z$ 等；

(3)通过求解投影指标函数最大化来估计最佳投影方向，以最大可能揭示高维数据某类特征结构。计算公式为

$$\max Q'(a) = S_z D_z$$

$$\text{s.t. } \sum_{j=1}^p a^2(j) = 1, \quad a(j) \geq 0$$

式中，s.t. 代表约束条件，这是一个以 $a(j), j = 1, \dots, p$ 为优化变量的复杂非线性优化问题，一般常用模拟生物优胜劣汰规则与群体内部染色体信息交换机制的遗传算法来求解，但在DPS里面采用的是复合单纯形法，在有些情形下，优化效果可能更好一些。

(4) 把由公式 $\max Q'(a) = S_z D_z$ 与 s.t. $\sum_{j=1}^p a^2(j) = 1, a(j) \geq 0$ 求得的最佳投影方向 a^* 代入公式 $z(i) = \sum_{i=1}^n a(j)x(i, j)$ 后即得各评价样本的投影值 $z(i)$ ，按由大到小排序，据此可对评价指标样本集进行统一评价。

这里仍引用图 44-4 中某儿童医院 1994~1998 年 7 项指标的实际值，用投影寻踪分类法比较该医院这 5 年的医疗质量，分析前亦将数据编辑如图 44-4 的格式。

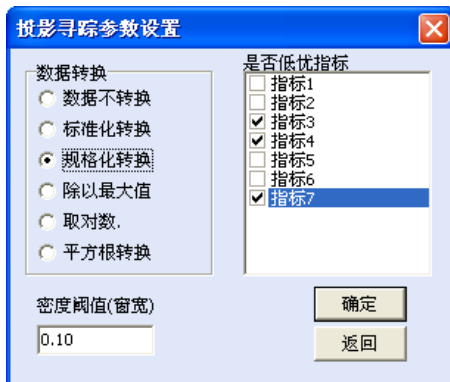


图 44-8 投影寻踪分类综合评价用户对话框

然后执行“其他方法”下面的“投影寻踪分类”功能，这时系统出现供用户对各个指标特性进行设置的对话框(图 44-8)。这里可根据每个指标的特性，设置该指标是否是“低优指标”以及该指标所占权重。本例中，平均住院日、病死率、院内感染率，即 x_3 ， x_4 和 x_7 为低优指标，其余为高优指标。数据转换方式为规格化转换，密度阈值取缺省值(0.1)。点击确定按钮后得到计算结果如下。

数据规格化转换.								
输入局部密度控制参数=0.100								
投影值标准差=0.7303			局部密度=0.5112					
目标函数 Q(a)=0.3734								
变量	投影方向							
X1	0.6654							
X2	0.5113							
X3	0.0000							
X4	0.3342							
X5	0.4019							
X6	0.1502							
X7	0.0003							
样本	X1	X2	X3(低优)	X4(低优)	X5	X6	X7(低优)	投影值
S1	0.1333	0.0000	0.2000	0.0000	0.0000	0.2857	1	0.1319
S2	0.9258	0.7059	0.0000	0.2763	0.7442	1	1	1.5189
S3	0.2632	0.3750	0.2000	0.5132	0.7442	0.0000	0.2500	0.8375
S4	0.0000	0.5735	1.0000	0.5395	0.6919	0.5714	0.4375	0.8375
S5	1	1	1.0000	1	1	0.8571	0.0000	2.0416

从结果的投影方向大小可以看出，第一个指标(x_1)对综合指标的影响最大， x_5 ， x_2 和 x_4 次之，其他因素影响较小。

从结果的投影值的大小，对各个样本排序分析可以看出，1998 年的综合效益最好，其次为 1995 年，随后为 1997 年，1996 年，1994 年最差。

44.7 层次分析法

层次分析法(analytic hierarchy process, AHP)是对一些较为复杂、较为模糊的问题作出决策的简易方法，它特别适用于那些难于完全定量分析的问题。它是美国运筹学家 Saaty 教授于 20 世纪 70 年代初期提出的一种简便、灵活而又实用的多准则决策方法。

44.7.1 层次分析法的基本原理与步骤

人们在进行社会、经济以及科学管理领域问题的系统分析中，面临的常常是一个由相互关联、相互制约的众多因素构成的复杂而往往缺少定量数据的系统。层次分析法为这类问题的决策和排序提供了一种新的、简洁而实用的建模方法。运用层次分析法建模，大体上可按以下几个步骤进行：

1. 递阶层次结构的建立与特点

应用 AHP 分析决策问题时，首先要把问题条理化、层次化，构造出一个有层次的结构模型。在这个模型下，复杂问题被分解为元素的组成部分。这些元素又按其属性及关系形成若干层次。上一层次的因素作为准则对下一层次有关因素起支配作用。这些层次可以分为 3 类：

(1) 最高层：这一层次中只有一个元素，一般它是分析问题的预定目标或理想结果，因此也称为目标层。

(2) 中间层：这一层次中包含了为实现目标所涉及的中环节，它可以由若干个层次组成，包括所需考虑的准则、子准则，因此也称为准则层。

(3) 最底层：这一层次包括了为实现目标可供选择的各种措施、决策方案等，因此也称为措施层或方案层。

递阶层次结构中的层次数与问题的复杂程度及需要分析的详尽程度有关，一般地层次数不受限制。每一层次中各元素所支配的元素一般不要超过 9 个。这是因为支配的元素过多会给两两比较判断带来困难。

2. 构造判断矩阵

层次结构反映了因素之间的关系，但准则层中的各准则在目标衡量中所占的比重不一定相同，在决策者的心目中，它们各占有一定的比例。

在确定影响某因素的诸因子在该因素中所占的比重时，遇到的主要困难是这些比重常常不易量化。此外，当影响某因素的因子较多时，直接考虑各因子对该因素有多大程度的影响时，常常会因考虑不周全、顾此失彼而使决策者提出与他实际认为的重要性程度不相一致的数据，甚至有可能提出一组隐含矛盾的数据。为看清这一点，可作如下假设：将一块重为 1kg 的石块砸成 n 小块，你可以精确称出它们的重量，设为 w_1, w_2, \dots, w_n ，现在，请人估计这 n 小块的重量占总重量的比例(不能让他知道各小石块的重量)，此人不仅很难给出精确的比值，而且完全可能因顾此失彼而提供彼此矛盾的数据。

设现在要比较 n 个因子 $X=\{x_1, x_2, \dots, x_n\}$ 对某因素 Z 的影响大小，怎样比较才能提供可信的数据呢？Saaty 等建议可以采取对因子进行两两比较建立成对比较矩阵的办法。即每次取两个因子 x_i 和 x_j ，以 a_{ij} 表示 x_i 和 x_j 对 Z 的影响大小之比，全部比较结果用矩阵 $A=(a_{ij})_{n \times n}$ 表示，称 A 为 Z - X 之间的成对比较判断矩阵(简称判断矩阵)。容易看出，若 x_i 与 x_j 对 Z 的影响之比为 a_{ij} ，则 x_j 与 x_i 对 Z 的影响之比应为 $a_{ji}=1/a_{ij}$ 。

定义 1 若矩阵 $A = (a_{ij})_{n \times n}$ 满足① $a_{ij} > 0$, ② $a_{ji} = 1/a_{ij}$ ($i, j=1, 2, \dots, n$), 则称之为正互反矩阵(易见 $a_{ii} = 1, i=1, 2, \dots, n$)。

关于如何确定 a_{ij} 的值, Saaty 等建议引用数字 1~9 及其倒数作为标度。表 44-2 列出了 1~9 标度的含义:

表 44-2 比较矩阵中标度的含义

标度	含 义
1	表示两个因素相比, 具有相同重要性
3	表示两个因素相比, 前者比后者稍重要
5	表示两个因素相比, 前者比后者明显重要
7	表示两个因素相比, 前者比后者强烈重要
9	表示两个因素相比, 前者比后者极端重要
2, 4, 6, 8	表示上述相邻判断的中间值
倒数	若因素 i 与因素 j 的重要性之比为 a_{ij} , 那么因素 j 与 i 重要性之比为 $a_{ji} = 1/a_{ij}$

从心理学观点来看, 分级太多会超越人们的判断能力, 既增加了作判断的难度, 又容易因此而提供虚假数据。Saaty 等还用实验方法比较了在各种不同标度下人们判断结果的正确性, 实验结果也表明, 采用 1~9 标度最为合适。

最后, 应该指出, 一般地作 $n(n-1)/2$ 次两两判断是必要的。有人认为把所有元素都和某个元素比较, 即只作 $n-1$ 个比较就可以了。这种做法的弊病在于, 任何一个判断的失误均可导致不合理的排序, 而个别判断的失误对于难以定量的系统往往是难以避免的。进行 $n(n-1)/2$ 次比较可以提供更多的信息, 通过各种不同角度的反复比较, 从而导出一个合理的排序。

3. 层次单排序及一致性检验

判断矩阵 A 对应于最大特征值 λ_{\max} 的特征向量 W , 经归一化后即为一层次相应因素对于上一层次某因素相对重要性的排序权值, 该过程称为层次单排序。

上述构造成对比较判断矩阵的办法虽能减少其他因素的干扰, 较客观地反映出一对因子影响力的差别。但综合全部比较结果时, 其中难免包含一定程度的非一致性。如果比较结果是前后完全一致的, 则矩阵 A 的元素还应当满足

$$a_{ij}a_{jk} = a_{ik}, \quad \forall i, j, k=1, 2, \dots, n$$

定义 2 满足上述关系式的正互反矩阵称为一致矩阵。

需检验构造出来的(正互反)判断矩阵 A 是否严重地非一致, 以便确定是否接受 A 。

定理 1 正互反矩阵 A 的最大特征根 λ_{\max} 必为正实数, 其对应特征向量的所有分量均为正实数。 A 的其余特征值的模均严格小于 λ_{\max} 。

定理 2 若 A 为一一致矩阵, 则

(1) A 必为正互反矩阵。

(2) A 的转置矩阵 A^T 也是一致矩阵。

(3) A 的任意两行成比例, 比例因子大于零, 从而 $\text{rank}(A) = 1$ (同样, A 的任意两列也成比例)。

(4) A 的最大特征值 $\lambda_{\max} = n$, 其中 n 为矩阵 A 的阶。 A 的其余特征根均为零。

(5) 若 A 的最大特征值 λ_{\max} 对应的特征向量为 $W = (w_1, w_2, \dots, w_n)^T$, 则 $a_{ij} = w_i / w_j$, $\forall i, j = 1, 2, \dots, n$, 即

$$A = \begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \dots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \dots & \frac{w_2}{w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \dots & \frac{w_n}{w_n} \end{bmatrix}$$

定理 3 n 阶正互反矩阵 A 为一致矩阵当且仅当其最大特征根 $\lambda_{\max} = n$, 且当正互反矩阵 A 非一致时, 必有 $\lambda_{\max} > n$ 。

根据定理 3, 我们可以由 λ_{\max} 是否等于 n 来检验判断矩阵 A 是否为一致矩阵。由于特征根连续地依赖于 a_{ij} , 故 λ_{\max} 比 n 大得越多, A 的非一致性程度也就越严重, λ_{\max} 对应的标准化特征向量也就越不能真实地反映出 $X = \{x_1, x_2, \dots, x_n\}$ 在对因素 Z 的影响中所占的比重。因此, 对决策者提供的判断矩阵有必要作一次一致性检验, 以决定是否接受它。

对判断矩阵的一致性检验的步骤如下:

(1) 计算一致性指标 CI:

$$CI = (\lambda_{\max} - n) / (n - 1)$$

(2) 查找相应的平均随机一致性指标 RI。对 $n = 1, 2, \dots, 15$, Saaty 给出了 RI 的值。在 DPS 系统中, 作者首先用随机地从 1~9 及其倒数中抽取数字方法构造正互反矩阵, 求最大特征根, 重复 20000 次, 计算平均值 λ'_{\max} , 并定义

$$RI = \frac{\lambda'_{\max} - n}{n - 1}$$

然后根据不同的 $n(n = 3, 4, \dots, 100)$ 所求得的 RI 估计值建立经验方程, 得

$$RI = \frac{-5.739303 + 4.303993n - 1.281630n^2 + 0.241633n^3}{1.771704n - 0.540019n^2 + 0.137736n^3}$$

(3) 计算一致性比例 CR:

$$CR = CI / RI$$

当 $CR < 0.10$ 时, 认为判断矩阵的一致性是可以接受的, 否则应对判断矩阵作适当修正。

4. 残缺判断矩阵的权重系数估计

当判断矩阵中有的比较单元缺失时, 和完全信息下排序方法类似, 可采用特征根方法获得不完全信息下的排序权值。特征根方法是由拟互反矩阵似一致性条件得出, 对任何非零元素均有 $a_j = w/w_j$, w 和 w_j 是排序权重系数。把这一结论扩展到“零”元素即残缺元素上, 即若 $a_j = 0$, 那么就用 w/w_j 替代它。设 $A = (a_{ij})$ 是一拟互反矩阵, 其排序向量为 $w = (w_1, w_2, \dots, w_n)^T$, 对 A 构造一个辅助矩阵 C :

$$C_{ij} = \begin{cases} a_j, & \text{当 } a_j \neq 0 \\ w/w_j, & \text{当 } a_j = 0 \end{cases}$$

例如, 矩阵 $A = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 & 2 \\ & 1/2 & 1 \end{bmatrix}$ 的辅助矩阵为 $C = \begin{bmatrix} 1 & 2 & w_1/w_3 \\ 1/2 & 1 & 2 \\ w_3/w_1 & 1/2 & 1 \end{bmatrix}$, 解特征根问

题 $Cw = \lambda_{\max}w$, 展开后得

$$(2w_1 + 2w_2, w_1/2 + w_2 + 2w_3, w_2/2 + 2w_3) = \lambda_{\max}(w_1, w_2, w_3)$$

实际上, 这里的特征根问题与矩阵

$$\bar{A} = \begin{bmatrix} 2 & 2 \\ 1/2 & 1 & 2 \\ & 1/2 & 2 \end{bmatrix}$$

的特征根问题 $\bar{A}w = \lambda_{\max}w$ 等价。故可把 \bar{A} 称为 A 的等价矩阵。 \bar{A} 中的“0”表示数值 0。若以 θ 表示矩阵中缺失元素, 这时

$$\bar{a}_{ij} = \begin{cases} a_{ij}, & i \neq j, a_{ij} \neq \theta \\ 0, & i \neq j, a_{ij} = \theta \\ m_i + 1, & i = j \end{cases}$$

式中, m_i 是矩阵 A 中第 i 行中 θ 的个数。得到 A 的等价矩阵 \bar{A} 后, 采用特征根方法即可得到各个因素的权重系数。

5. 层次总排序及一致性检验

上面得到的是一组元素对其上一层中某元素的权重向量。我们最终要得到各元素, 特别是最低层中各方案对于目标的排序权重, 从而进行方案选择。总排序权重要自上而下地将单准则下的权重进行合成。

设上一层次 (A 层) 包含 A_1, L, A_m 共 m 个因素, 它们的层次总排序权重分别为 a_1, L, a_m 。又设其后的下一层次 (B 层) 包含 n 个因素 B_1, L, B_n , 它们关于 A_j 的层次单排序权重分别为 b_{1j}, L, b_{nj} (当 B_i 与 A_j 无关联时, $b_{ij} = 0$)。现求 B 层中各因素关于总目标的权重, 即求 B 层各因素的层次总排序权重 b_1, L, b_n , 计算按表 44-3 所示方式进行, 即

$$b_i = \sum_{j=1}^m b_{ij} a_j, \quad i=1, L, n。$$

表 44-3 层次间权重计算关系表

层次 A \ 层次 B	A_1	A_2	...	A_m	B 层次总排序数值
	a_1	a_2	...	a_m	
B_1	b_{11}	b_{12}	...	b_{1m}	$\sum_{j=1}^m a_j b_{1j}$
B_2	b_{21}	b_{22}	...	b_{2m}	$\sum_{j=1}^m a_j b_{2j}$
\vdots	\vdots	\vdots		\vdots	\vdots
B_n	b_{n1}	b_{n2}	...	b_{nm}	$\sum_{j=1}^m a_j b_{nj}$

对层次总排序也需作一致性检验, 检验仍像层次总排序那样由高层到低层逐层进行。这是因为虽然各层次均已经过层次单排序的一致性检验, 各成对比较判断矩阵都已具有较为满意的一致性。但当综合考察时, 各层次的非一致性仍有可能积累起来, 引起最终分析结果较严重的非一致性。

设 B 层中与 A_j 相关因素的成对比较判断矩阵在单排序中经一致性检验, 求得单排序一致性指标为 $CI(j)$, ($j=1, L, m$), 相应的平均随机一致性指标为 $RI(j)$, $CI(j)$, $RI(j)$ 已在层次单排序时求得, 则 B 层总排序随机一致性比例为

$$CR = \frac{\sum_{j=1}^m CI(j)a_j}{\sum_{j=1}^m RI(j)a_j}$$

当 $CR < 0.10$ 时, 认为层次总排序结果具有较满意一致性并接受该分析结果。

44.7.2 层次分析法的应用

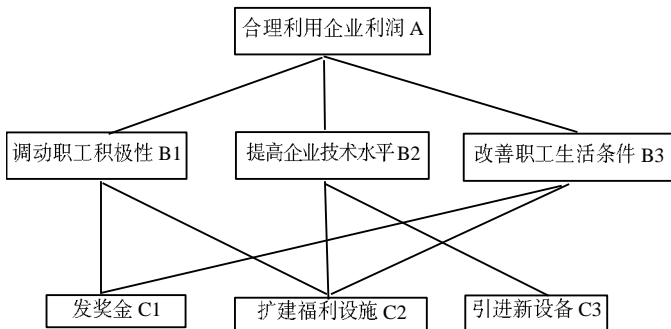
在应用层次分析法研究问题时, 遇到的主要困难有两个: (i) 如何根据实际情况抽象出较为贴切的层次结构; (ii) 如何将某些定性的量作比较接近实际定量化处理。层次分析法对人们的思维过程进行了加工整理, 提出了一套系统分析问题的方法, 为科学管理和决策提供了较有说服力的依据。但层次分析法也有其局限性, 主要表现在: (i) 它在很大程度上依赖于人们的经验, 主观因素的影响很大, 它至多只能排除思维过程中的严重非一致性, 却无法排除决策者个人可能存在的严重片面性; (ii) 比较、判断过程较为粗糙, 不能用于精度要求较高的决策问题。AHP 至多只能算是一种半定量(或定性定量结合)的方法。

在应用层次分析法时, 建立层次结构模型是十分关键的一步。现在分析一个实例, 以便说明如何从实际问题中抽象出相应的层次结构。

例 1 某工厂有一笔企业留成利润, 要由厂领导和职代会决定如何利用, 可供选择的方案有: 发奖金、扩建福利设施、引用新设备, 为进一步促进企业发展, 如何合理利

用这笔利润？

首先，对于这个问题采用层次分析法进行分析，所有措施的目的都是为了更好地调动职工生产积极性、提高企业技术水平和改善职工生活，当然最终目的是为了促进企业的发展，因此，建立的递阶层次结构如图所示。



其次，构造判断矩阵。在 DPS 系统支持下，为便于进行多个层次数据的分析，判断矩阵在 DPS 里编辑规则如图 44-9 和图 44-10。

	A	B	C	D	E	F
1						
2		A	B1	B2	B3	
3			1	0.2	0.3333	
4				1	3	
5					1	
6						

图 44-9 层次分析数据编辑格式(总目标层次)

	A	B	C	D	E
1					
2		B1	C1	C2	
3			1	3	
4				1	
5					
6		B2	C1	C2	
7			1	0.2	
8				1	
9					
10		B3	C1	C2	
11			1	2	
12				1	
13					

图 44-10 次级层次数据编辑格式

即数据在电子表格中必须规范存放，且一个层次分析项目需要建立一个数据文件(这一点不像其他功能)。数据文件中各层因素以 A, B, ... 英文字母顺序取名。每个层次，依顺序建立一个表单。总目标(即 AB)层次放在工作表的第一个表单里，次级层次(BC 层次、CD 层次, ...)放在工作表的第 2 个, 第 3 个表单里, ...。工作表的第 1 列放入上 1

层次的因子名称。按此规则，总目标层次，即第 1 个表单第 1 行第 1 列则是“A”；第 2 个表单，即 BC 层次第 1 行第 1 列则是“B1”；第 3 个表单，即 CD 层次第 1 行第 1 列则是“C1”；等等。要注意的是，图 44-9 和图 44-10 中，各个层次的因子的代码，如“B1”、“D5”等需在“英文数字”状态输入，不能是全角字符。

图 44-9 和图 44-10 是上述挑选合适的工作过程例子所取得数据进行层次分析的编辑格式。

图 44-9 是企业的 3 个方面(B1 是调动职工生产积极性，B2 是提高企业技术水平，B3 是改善职工生活)对总目标(A: 合理利用企业利润)的判断矩阵。

图 44-10 是第 2 个层次的数据编辑格式。图中为企业的 3 项措施(C1, C2 和 C3)，对企业的 3 个方面进行判断的判断矩阵。

数据编辑之后，不需要用鼠标选择数据，而是直接从菜单中选择“层次分析法”执行层次分析功能。分析时，系统出现要求用户输入层次数的界面，本例中的层次数为 2。输入 2 之后，点击“OK”按钮，系统会给出分析结果如下：

判断矩阵 A-Bi					
A	B(1)	B(2)	B(3)	W	位次
B(1)	1.0000	0.2000	0.3333	0.1047	3
B(2)	5.0000	1.0000	3.0000	0.6370	1
B(3)	3.0000	0.3333	1.0000	0.2583	2
$\lambda_{\max} = 3.0385$		CI = 0.0193	RI = 0.5157	CR = 0.0373	
判断矩阵 B1-Ci					
B1	C(1)	C(2)	W	位次	
C(1)	1.0000	3.0000	0.7500	1	
C(2)	0.3333	1.0000	0.2500	2	
$\lambda_{\max} = 2.0000$		CI = 0.0000	RI = 0.0000	CR = 0.0000	
判断矩阵 B2-Ci					
B2	C(2)	C(3)	W	位次	
C(2)	1.0000	0.2000	0.1667	2	
C(3)	5.0000	1.0000	0.8333	1	
$\lambda_{\max} = 2.0000$		CI = 0.0000	RI = 0.0000	CR = 0.0000	
判断矩阵 B3-Ci					
B3	C(1)	C(2)	W	位次	
C(1)	1.0000	2.0000	0.6667	1	
C(2)	0.5000	1.0000	0.3333	2	
$\lambda_{\max} = 2.0000$		CI = 0.0000	RI = 0.0000	CR = 0.0000	
B-C 层次总排序					

图 44-11 电力修复问题层次分析数据编辑格式

本例数据组织、编辑，要注意的是影响人数(B2)的输入格式：它是各个单位的人数，即以人数直接作为权重指标，以一行的形式输入，而不再是上三角矩阵。

数据输入后，执行“层次分析法”功能，系统给出分析结果如下：

A-Bi 的判断矩阵								
3A	B(1)	B(2)	W	位次				
B(1)	1.0000	5.0000	0.8333	1				
B(2)	0.2000	1.0000	0.1667	2				
$\lambda_{max}=2.0000$		CI=0.0000	RI=0.0000	CR=0.0000				
B(1)-Ci 的判断矩阵								
B(1)	C(1)	C(2)	C(3)	C(4)	C(5)	C(6)	W	位次
C(1)	1.0000	0.1667	0.1667	0.2500	0.3333	0.1111	0.0245	6
C(2)	6.0000	1.0000	3.0000	5.0000	3.0000	0.1429	0.1817	2
C(3)	6.0000	0.3333	1.0000	6.0000	4.0000	0.1667	0.1401	3
C(4)	4.0000	0.2000	0.1667	1.0000	3.0000	0.1111	0.0590	4
C(5)	3.0000	0.3333	0.2500	0.3333	1.0000	0.1111	0.0425	5
C(6)	9.0001	7.0000	6.0000	9.0000	9.0000	1.0000	0.5522	1
$\lambda_{max}=6.8786$		CI=0.1757	RI=1.2482	CR=0.1408				
因子层 B(2)-Ci 的权重								
因子	原始值	归一化						
C(1)	1615.0000	0.0184						
C(2)	5900.0000	0.0671						
C(3)	1035.0000	0.0118						
C(4)	75820.0000	0.8629						
C(5)	3100.0000	0.0353						
C(6)	395.0000	0.0045						
B-C 层次总排序								
C\B	B(1)	B(2)	CW	位次				
Bi 权重	0.8333	0.1667						
C(1)	0.0245	0.0184	0.0235	6				
C(2)	0.1817	0.0671	0.1626	3				
C(3)	0.1401	0.0118	0.1187	4				
C(4)	0.0590	0.8629	0.1929	2				
C(5)	0.0425	0.0353	0.0413	5				
C(6)	0.5522	0.0045	0.4609	1				
层次总排序一致性		CI=0.0000	RI=1.2069	CR=0.0000				

这里请注意，在输出结果中，各单位人数权重，采取归一化方法进行了处理。该问题结论是可按照此权重系数大小顺序来安排与指派维修工人。

例 3 医院工作质量评价。首先，需对医院各项工作进行连续性分解，建立医院工作质量评价的目标树如图 44-12。

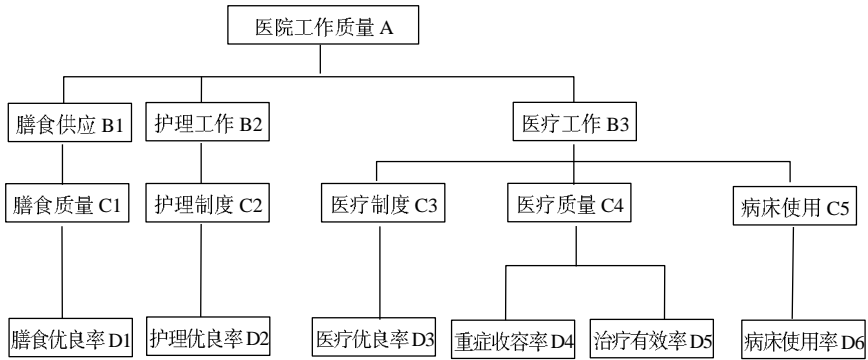


图 44-12 医院工作质量评价目标树

建立的目标树，其第 1 层是总的目标，即医院工作质量；第 2 层次为膳食供应、护理工作和医疗工作共 3 个子目标；医疗工作下的第 3 层次为医疗制度、医疗质量、病床使用子目标；在医疗质量下还有第四层次(重症收容率、疗效 2 个子目标)，最下面层次为医院工作质量评价的膳食优良率、护理优良率……病床使用率等 6 项指标。

从目标树可以看出，第 1 层子目标有 3 个评价目标须成对比较；第 2 层子目标中，医疗工作可分解为 3 个子目标，即医疗制度、医疗质量和病床使用；在第 3 层子目标中，医疗质量又可分解为 2 个子目标，即重症收容率和治疗有效率。因此该系统的层次分析需建立 3 个判别矩阵。3 个判别矩阵评判前面的目标树进行，方法如表 44-4。

表 44-4 各层子目标成对比较优选矩阵

第一层 (工作质量)	膳食 供应	护理 工作	医疗 工作	第二层 (医疗工作)	医疗 制度	医疗 质量	病床 使用	第三层 (医疗质量)	重症 收容率	治疗 有效率
膳食供应	1	1/3	1/5	医疗制度	1	1/3	1/2	重症收容率	1	1/2
护理工作	3	1	1/3	医疗质量	3	1	2	治疗有效率	2	1
医疗工作	5	3	1	病床使用	2	1/2	1			

在 DPS 系统中对该例进行层次分析，其数据组织、编辑可根据图 44-12 所示的目标树进行。从目标树我们可以看出，第 2 层的 B1→C1，B2→C2；以及第 3 层的 C1→D1，C2→D2，C3→D3 和 C5→D6 实际上并没有被分解，因此这几个节点处的权重系数都可认为是 1。按此方式组织、编辑，这时 DPS 电子工作表中从左到右各层数据格式如图 44-13 所示。

图 44-13 医院工作质量评价层次分析数据组织、编辑格式

用户可能注意到，前面两例我们作层次分析时，只按规则组织、编辑数据即可进行分析，没有使用鼠标将任何数据选中。但在本例中，把系统中各个层次的各个因子名称，以一层为一行、每行中依各个因子编码顺序的方式输入到电子表格中，并用鼠标将其选中，其效果会是怎样？当对该例执行层次分析功能之后，得到的结果如下：

工作质量-Bi 的判断矩阵					
工作质量	膳食供应	护理工作	医疗工作	W	位次
膳食供应	1.0000	0.3333	0.2000	0.1047	3
护理工作	3.0000	1.0000	0.3333	0.2583	2
医疗工作	5.0000	3.0000	1.0000	0.6370	1
$\lambda_{max} = 3.0385$		CI = 0.0193	RI = 0.5180	CR = 0.0372	
膳食供应-Ci 层的权重					
因子	原始值	归一化			
膳食质量	1.0000	1.0000			
护理工作-Ci 层的权重					
因子	原始值	归一化			
护理制度	1.0000	1.0000			
医疗工作-Ci 的判断矩阵					
医疗工作	医疗制度	医疗质量	病床使用	W	位次
医疗制度	1.0000	0.3333	0.5000	0.1634	3
医疗质量	3.0000	1.0000	2.0000	0.5396	1
病床使用	2.0000	0.5000	1.0000	0.2970	2
$\lambda_{max} = 3.0092$		CI = 0.0046	RI = 0.5180	CR = 0.0089	
B-C 层次总排序					
C\B	膳食供应	护理工作	医疗工作	CW	位次
Bi 权重	0.1047	0.2583	0.6370		
膳食质量	1.0000	0.0000	0.0000	0.1047	4
护理制度	0.0000	1.0000	0.0000	0.2583	2
医疗制度	0.0000	0.0000	0.1634	0.1041	5
医疗质量	0.0000	0.0000	0.5396	0.3437	1

病床使用	0.0000	0.0000	0.2970	0.1892	3		
层次总排序一致性 CI = 0.0029			RI = 0.6930	CR = 0.0461			
膳食质量-Di 层的权重							
因子	原始值	归一化					
膳食质量	1.0000	1.0000					
护理制度-Di 层的权重							
因子	原始值	归一化					
护理制度	1.0000	1.0000					
层医疗制度-Di 层的权重							
因子	原始值	归一化					
医疗制度	1.0000	1.0000					
医疗质量-Di 的判断矩阵							
医疗质量	重症收容	疗效	W	位次			
重症收容	1.0000	0.5000	0.3333	2			
疗效	2.0000	1.0000	0.6667	1			
$\lambda_{\max} = 2.0000$	CI = 0.0000	RI = 0.0000	CR = 0.0000				
病床使用-Di 层的权重							
因子	原始值	归一化					
病床使用	1.0000	1.0000					
C-D 层次总排序							
D\C	膳食质量	护理制度	医疗制度	医疗质量	病床使用	CW	位次
Ci 权重	0.1047	0.2583	0.1041	0.3437	0.1892		
膳食质量	1.0000	0.0000	0.0000	0.0000	0.0000	0.1047	5
护理制度	0.0000	1.0000	0.0000	0.0000	0.0000	0.2583	1
医疗制度	0.0000	0.0000	1.0000	0.0000	0.0000	0.1041	6
重症收容	0.0000	0.0000	0.0000	0.3333	0.0000	0.1146	4
疗效	0.0000	0.0000	0.0000	0.6667	0.0000	0.2292	2
病床使用	0.0000	0.0000	0.0000	0.0000	1.0000	0.1892	3
层次总排序一致性 CI = 0.0000			RI = 0.6563	CR = 0.0461			

输出结果中所有的名称不像前面那两个例子是以代码表示各个因子，而都变成了中文。这样，分析结果更加一目了然。但要注意，中文名称不要超过 4 个汉字(9 个英文字符)，否则输出结果将不整齐。

当判断矩阵中有的元素缺失(没有判断评价)时，就让它空着(输入 0 也可以)，DPS 会应用特征根方法处理含有缺失单元的判断矩阵，估计出权重系数。例如，对有缺失单元的判断矩阵 $A = \begin{bmatrix} 1 & 2 & 0 \\ & 1 & 2 \\ & & 1 \end{bmatrix}$ ，DPS 会采用特征根方法对它进行权重系数估计，得到权重

系数分别为 0.5714，0.2857 和 0.1429。

DPS 系统进行层次分析，其层次及因子的数量多少均没有限制。样本文件 dps_ahp.cll(下载网址 http://www.chinadps.net/dps_ahp.cll)是一个含有 5 大类因子，4 个层次，共 54 个因子项目的复杂系统。根据该例子可以进一步掌握较复杂系统的层次分析技术。

参 考 文 献

- 陈义华. 1997. 数学建模的层次分析法[J]. 甘肃工业大学学报, (3): 92-97.
- 方积乾. 2003. 卫生统计学[M]. 北京: 人民卫生出版社.
- 胡运权, 郭耀煌. 1998. 运筹学教程[M]. 清华大学出版社.
- 姜启源. 1993. 数学模型(第 2 版) [M]. 北京: 高等教育出版社.
- 刘关键, 吴秦相, 康德英. 2003. Meta-分析中的统计学过程[J]. 中国临床康复, 7(4), 538-539.
- 孙振球. 2002. 医学统计学[M]. 北京: 人民卫生出版社.
- 王莲芬, 许树柏. 1990. 层次分析法引论[M]. 北京: 中国人民大学出版社.
- 王毓基. 1985. 区域规划系统工程[M]. 长沙: 湖南大学出版社.
- 许树柏. 1988. 层次分析法原理[M]. 天津: 天津大学出版社.
- 周仁郁. 2004. 中医药统计学[M]. 北京: 中国中医药出版社.
- Fleiss J L, Gross A J. 1991. Meta analysis in epidemiology[J]. J Clin Epidemiology. 44(2): 127.
- Friedman J H and Tukey J W. 1974. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Transactions on Computers, C-23 (9): 881-890
- Saaty T L. 1988. 层次分析法—在资源分配、管理和冲突分析中的应用[M]. 北京: 煤炭工业出版社.
- Saaty T L. 1980. The Analysis Hierarchy Process[M]. New York: McGraw-Hill Inc.

配套光盘使用说明

本书配套光盘包含内容有：

1. DPS 统计软件试用、演示版。点击 **Setup.exe**，执行安装程序后便可试用。试用版可以对经过处理后的例子数据进行统计分析。

2. 本书各章例子数据文件，文件名是每一章的章名，例子中的数据都已经过处理，数据打开后就可使用，文件存放在“例子数据”文件夹。

3. 抽样技术的电子工作表，打开后就可使用，文件存放在“抽样技术”文件夹。

4. Meta 分析的电子工作表，打开后就可使用，文件存放在“Meta 分析”文件夹。

5. 常用统计分布电子工作表，打开后就可使用，文件存放在“统计分布与检验”文件夹。

6. 单样本、两样本统计检验及样本含量、功效计算的电子工作表，打开后就可使用，文件存放在“统计分布与检验”文件夹。

7. 本书初稿的 PDF 格式文档(文件名是对应 1~3 卷，分别是 **DPS_Volume_I(3rd).pdf**, **DPS_Volume_II(3rd).pdf** 和 **DPS_Volume_III(3rd).pdf**)。

试用版 DPS 及相关文档，可从网站 <http://www.dpsw.cn> 和 <http://www.statforum.com> 直接下载，DPS 统计软件购买与更新事宜可以浏览其官方网站并和开发商联系。DPS 软件应用范围很广，使用中如需技术咨询，请与软件开发、技术支持人员联系，通讯、联系方式如下：

信件：浙江大学紫金港校区农业与生物技术学院 C 座 1049，邮编 310058，唐启义 收
电话：13958168331

E-mail: qytang@zju.edu.cn

网站：<http://www.dpsw.cn>