研究生教学用书

教育部学位管理与研究生教育司推荐

医用多元统计分析方法

(第二版)

陈 峰 编著　陈启光 审阅

中国统计出版社
China Statistics Press

研究生教学用书

教育部学位管理与研究生教育司推荐

医用多元统计分析方法

(第二版)

陈 峰 编著　陈启光 审阅

中国统计出版社
China Statistics Press

(京)新登字041号

图书在版编目(CIP)数据

医用多元统计分析方法/陈峰编著. —2版
—北京:中国统计出版社,2006.11
ISBN 978-7-5037-3982-8

Ⅰ.医…
Ⅱ.陈…
Ⅲ.医学统计—多元分析—方法—研究生—教材
Ⅳ.R311

中国版本图书馆CIP数据核字(2006)第128551号

责任编辑/吕　军
封面设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街57号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲6号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/16
字　　数/450千字
印　　张/19.25
印　　数/1—3000册
版　　别/2007年1月第2版
版　　次/2007年1月第1次印刷
书　　号/ISBN 978-7-5037-3982-8/R·7
定　　价/35.00元

第二版前言

本书自 2000 年 12 月第一版问世以来，得到了读者的普遍关注，来信、来电不断，给了作者很多鼓励、建议和批评，有些学校将其列为多元统计分析的教材。去年岁末，本书被教育部推荐为研究生教学用书。甚为欣慰。借此重印之际，对原书作了修订。

此次修订对几个统计学名词作了统一：①将“multiple regression”译为多重回归，避免与“multivariate regression（多元回归）”相混淆，前者指一个因变量与多个自变量的回归，而后者是指多个因变量与多个自变量的回归；②将“odds”译为优势，“odds ratio”译为优势比，不再用“比数”和“比数比”的说法；③将“statistically significant”译为有统计学意义，不再用“显著性”一词，以免引起与实际资料中指标间的“显著差别”相混淆。

陈启光教授、于浩教授、荀鹏程博士、柏建岭博士仔细阅读了第二版书稿并提出了很多中肯的建议。中国统计出版社为本书的再版做了大量的组织协调工作。作者在此一并表示衷心感谢。

统计学是一门从数据中学习的科学，它植根于概率论和数学，并受计算科学的影响。面向应用是统计学自身发展的最直接的原动力，从应用中发现自身的不足，从而激发人们对新理论的探索和新方法的研究，似乎是统计学发展的永恒的轨迹。随着计算机技术和信息技术的迅猛发展，各门学科均得到了前所未有的发展契机，统计学更是如此。而伴随着“-omics”时代的到来，统计学在生物医学领域中的应用研究越来越广泛和深入，并不断与信息、计算科学交叉融合，在统计信息学、计算基因组学、遗传流行病学、进化和种群遗传学、计算神经学等方面有了突破性的进展；在统计理论研究方面上，Bayes 统计、计算密集型统计、高维数据统计等方面也得到了一定的发展，新的理论和方法不断涌现。

但是，统计学在其发展过程中已形成的理论和方法，不会因为新方法的涌现而被淘汰，尤其是较为经典者。例如，1713 年 J. Bernoulli 提出的二项

分布，1761 年 T. Bayes 提出的 Bayes 定理，1800 年左右 Legendre 提出的最小二乘法，1809 年 CF. Gauss 提出的正态分布，1900 年 K. Pearson 发展的 χ^2 检验，1921～1935 年期间 RA. Fisher 提出的实验设计的基本原则、方差分析的基本思想、极大似然函数等，这些经典的理论和方法几乎每天都有成千上万的科学家们在使用。对经典理论和方法的深刻理解和领悟，必将有助于掌握现代统计思想和方法；相反，如不了解甚至误解经典理论和思想，则很难真正掌握和理解现代统计的精髓。因此，无论是统计理论工作者还是应用工作者，都应掌握并正确理解经典统计的概念、理论和方法。这一点再怎么强调都不为过！本书现有内容是多元统计分析中最为基础的，且是医学研究生必须要掌握的。因此，这次修订仅限于修辞层面，而没有增添新的内容。

学习本身也是多元的，学习统计学更应该如此。很多读者反映统计学难学，有些人听了好几遍课仍不得要领。事实上，统计学是一门集理论性与实践性于一体的学科。要学好统计学，既要掌握其基本概念、基本原理和基本方法，又要不断练习、反复实践。模仿是学习的开端，因此，要找一本好书细细读，慢慢品，做做练习，谈谈体会。但不唯书。只看看书，听听课，无法超越书本，也难以领会统计学的精髓。实践出真知。统计学的理论和方法只有以“鲜活、多样”的实际问题为载体，方有用武之地；只有在实践中，所学的理论和方法才能得到最直接、最生动的验证；只有不断在实践中学习，才会加深对统计理论、概念、方法和分析策略等的领悟，方能超然书外，得心应手，运用自如。学而时习之，不亦乐乎？与读者共勉。

陈 峰

2006 年 12 月 于南京

序

医学和生物现象变化万端，其因果关系更是错综复杂。人们所看到的某一结果的发生往往是众多因素综合作用的结果，通常并非某一因素的单一作用所致。例如，在大体相似的自然环境和社会条件下有些人得了某种疾病，而其余的人却不患该病，其致病因素往往不是单一的和唯一的，寻找这些致病因素的独立作用和联合作用是医学研究者的重要任务。

对于这些多因素共同作用的医学现象，要想探讨和澄清其中的必然规律，常用的单因素分析法（即一元分析）显然是无能为力的。因此，从上个世纪后半叶起，借助电子计算机这一有力计算工具，国内外许多统计学先驱陆续把数理统计中的多元统计分析方法引入到医学研究中来，并在实践中获得了广泛的应用和取得了丰硕的成果。

实践中的成果和进一步的需要反过来又促进了医用多元统计分析方法不断向广度和深度发展，丰富和完善新的、功能优异的统计软件包的陆续涌现更使其如虎添翼。

进入 20 世纪末期，医学统计学界更是硕果累累，人才辈出，他们不仅有坚实的数学基础，更掌握电脑操作及其统计软件包使用方面的熟练技巧，再加上在医学应用方面的广泛实践经验，使得医用多元统计分析这一边缘性学科得到了迅猛的发展。陈峰教授就是其中的杰出代表之一。他在攻读硕士、博士期间，师从医学统计学界的著名学者，并继承了师辈们严谨的治学态度和开拓进取的好学精神，对本学科的发展前沿作了不懈的努力。近年来，随着其工作面的不断扩展，实践经验和理论修养也日益深化。本书就是在此背景下的一个宝贵产物。

该书较系统、全面地总结和反映了医用多元统计分析方法在目前阶段的发展现状及其应用成果和前景。作者在介绍统计方法时，从讲解实例入手，逐步深入，理论与实际并茂，突出应用背景，强调分析思路，并精心安排了大量习题，提供了参考答案和相应的 SAS 程序，有助于培养统计思维，指

导研究设计，提高应用技巧。不同背景的读者均可从中获得系统的多元统计分析知识和实例借鉴。全书布局严谨，思路缜密；文字流畅，可读性强。为本专业的教学、科研人员提供了一本有益的教材和参考专著，广大医务人员和医学院校师生亦可从中得到不少启示。

欣闻书稿即将付梓，特撰数言，权且为序。正值新世纪来临之际，相信该书的出版对于医学科学的发展将起到添砖加瓦的作用。

史秉璋

2001 年 2 月于上海第二医科大学

前　言

第一次接触医用多元统计分析是在1985年史秉璋教授的讲习班上，他向我们展示了多元统计分析理论的博大精深，以及在医学各领域中应用的精妙，他精湛的授课艺术和严谨治学的作风对我产生了极大的影响。从此，我迷上了多元统计分析。学习、讲授、应用，乐此不疲。

多元统计学起源于20世纪20年代，Wishart，Hotelling，Fisher，Roy等是该领域的先驱。由于多元统计分析的计算量较大，开始时多局限于理论问题的研究。20世纪50年代以后，随着计算机及统计分析软件的日益发展，多元统计方法越来越广泛地得到应用，并渗透到自然科学和社会科学的各个领域。与数学的其它理论一样，统计学的很多内容都是因实际需要而产生的，伴随着应用面的扩大和深入，多元统计分析的理论亦得到了突飞猛进的发展。实践证明，多元统计分析方法是一种有效的数据处理工具。

多元统计分析的书籍很多，有偏于理论方面的，也有偏于应用方面的。本书立足于应用，略涉及一些简单的理论问题，以使读者知其然，亦知其所以然。

多元统计分析方法的内容非常丰富，在选材时，笔者根据实际情况，选择了目前医学研究中最为常用的一些方法，组成10个专题，每个专题一章。第1章是多元分析的基础，介绍多元分析中的基本统计量和多元正态分布及其应用。第2章介绍多元T检验和多元方差分析，详细讨论了成组设计、配对设计、区组设计、析因设计资料的多元方差分析。第3章介绍多元线性回归，详细讨论了衡量回归方程优劣的标准和变量筛选的技巧。第4、5章介绍主成分和因子分析，讨论了主成分的应用、因子分析的策略等。第6章介绍几种基于logistic分布的回归，讨论了条件的和非条件的logistic回归，多类结果和有序结果的logistic回归，以及建模策略等。第7章介绍广义线性模型的一般理论，包括参数估计，残差分析、拟合优度等，讨论了logistic回归与probit回归的区别，Poisson回归与负二项回归的关系等。第8章介绍生存资料的分析，包括指数分布、Weibull分布的拟合，指数回归、Weibull回归和Cox回归模型的建立。第9章介绍了聚类分析，讨论了8种系统聚类及其联系，快速聚类，有序样品的聚类及条件系统聚类。第10章介绍了判

别分析，讨论了距离判别，Bayes 判别，Fisher 判别等。其中 2～5、9、10 章介绍的方法属经典的多元分析方法，第 6、7、8 三章中介绍的方法是现代多元分析方法。为便于读者理解、掌握和正确应用，我们重点介绍各种方法所能解决的问题、应用条件及其局限性，在每章的最后安排了“正确应用”一节，指出应用中常遇到的一些具体问题及解决办法。

附录 A 是专门为那些希望深入了解极大似然估计的读者而撰写的，这在练习编程时可能有帮助。书中的例题和练习，除个别为说明一些理论问题而杜撰的外，均为实际资料。这些资料大都短小精悍，除囿于篇幅，更重要的原因是便于说明问题。作者亦有意识地安排了个别大型资料的分析实例，目的是使读者从中领略到多元分析的一些策略和思路。为便于读者学习和练习，书中所有需计算的例题均给出了 SAS 程序，练习亦给出了参考答案。然而，要建立统计学思维，理解统计学方法，合理解释统计分析结果，不仅仅是读几本书，做几个练习，恐怕最重要的是在实践中反复应用，不断加深理解。

在应用多元分析时应注意：(1)必须思路清晰，知道自己要干什么；(2)在作多元分析前，必须先作描述性分析。只有在充分了解资料性质的基础上，才有可能正确选择方法，得出有价值的结论；(3)当所得结果不符逻辑，或有悖于专业知识时，既不要轻易接受，亦不要轻易放弃，必须弄清楚为什么。

限于作者水平，书中错谬之处在所难免，恳请同行专家和广大读者不吝赐教。

本书获江苏省跨世纪学术带头人专项基金和交通部学术专著出版基金联合资助。陈启光教授仔细审阅了原稿，字字斟酌，提出了许多中肯而富有指导性的意见，使许多错误在付印前得以更正。中国统计出版社的范仲实先生为本书的出版做了大量的组织工作。习题参考答案是苟鹏程、沈毅两位同志完成的。书中还大量引用了其它文献中的原始资料。作者在此一并表示衷心感谢。

最后，感谢我的导师陆守曾教授和杨树勤教授，正是在他们的指导和熏陶下，我才懂得统计学是 20 世纪最伟大的学科之一。

陈 峰

2000 年岁末 于南通

目　录

1 多元正态分布

正态分布是统计学的重要理论分布之一。事实上，很多统计方法都是建立在正态分布的假设之上的。比如：t 检验，方差分析，线性相关与回归等。尽管实际数据不会严格服从正态分布，但有三个原因使正态分布在实际中有着广泛的应用：其一，正态分布在许多情况下确实能作为真实总体的一个近似；其二，根据中心极限定理，不论总体分布如何，许多统计量的分布是近似正态的；第三，很多检验统计量的分布对正态分布条件是稳健的(robust)，即原始资料稍微偏离正态对检验结果的影响不大。

本章介绍多元统计分析中常用的几种描述统计量，以及多元正态分布及其在确定参考值范围中的应用。

1.1 多元分析常用统计量

先看一个例子。

例 1.1 调查某地 16 岁中学生 12 名，测得其身高，体重，胸围资料如表 1.1。

表 1.1 12 名中学生的身高、体重、胸围测量资料

编号	身高(cm) x_1	体重(kg) x_2	胸围(cm) x_3
1	171.0	58.5	81.0
2	175.0	65.0	87.0
3	159.0	38.0	71.0
4	155.3	45.0	74.0
5	152.0	35.0	63.0
6	158.3	44.5	75.0
7	154.8	44.5	74.0
8	164.0	51.0	72.0
9	165.2	55.0	79.0
10	164.5	46.0	71.0
11	159.1	48.0	72.5
12	164.2	46.5	73.0

资料来源：郭祖超主编，《医学统计学》，人民军医出版社，1999，189 页。

这里有 3 个指标(变量)。对一元分析来说，只需计算各指标的均数、标准差。而对多元分析来说，除了要计算各指标的均数、标准差外，还要计算各指标间的协方差或相关系数。

与一元分析一样，多元分析所用统计量也是从样本计算而得。主要是均数、方差、标准差、相关系数等。只是在多元分析中，为了便于清晰地表达多指标(变量)间的关系，常将它们用数据阵即矩

阵(matrix)来表示。构成矩阵的每个数据称为元素(element)。

1.1.1 均向量

将各指标的均数用矩阵向量的形式排列,得均向量(means vector)。本例:

$$\overline{\boldsymbol{X}} = \begin{bmatrix} 161.8667 \\ 48.0833 \\ 74.3750 \end{bmatrix}$$

有时为方便印刷,均向量不用列向量表示而用行向量表示:

$$\overline{\boldsymbol{X}}' = (161.8667 \quad 48.0833 \quad 74.3750)$$

$\overline{\boldsymbol{X}}'$称为均向量的转置。

1.1.2 方差-协方差矩阵

将各指标的方差、协方差用矩阵的形式排列,得方差-协方差矩阵(variance-covariance matrix),有时简称为协方差阵(covariance matrix),用字母 $\boldsymbol{V}$ 表示。其中:

$$v_{ii} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)^2}{n-1}, \quad v_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{n-1} \tag{1.1}$$

n 为样本含量,$1 \leqslant i, j \leqslant m$;$m$ 为变量数。本例 $n=12, m=3$,且有:

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix} = \begin{bmatrix} 45.7224 & 50.3621 & 32.2318 \\ 50.3621 & 69.6288 & 45.4659 \\ 32.2318 & 45.4659 & 35.3239 \end{bmatrix}$$

显然:$v_{ij}=v_{ji}$,即协方差阵是对称(symmetry)矩阵,所以常给出矩阵的左下半,另一半对称的部分就不再写出,称为下三角矩阵。例如,

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & & \\ v_{21} & v_{22} & \\ v_{31} & v_{32} & v_{33} \end{bmatrix} = \begin{bmatrix} 45.7224 & & \\ 50.3621 & 69.6288 & \\ 32.2318 & 45.4659 & 35.3239 \end{bmatrix}$$

1.1.3 离均差平方和与离均差积和矩阵

将各指标的离均差平方和与离均差积和用矩阵的形式排列,得离均差平方和与离均差积和矩阵(deviation sum of squares and cross-products matrix, DSSCP),简称离差阵。离差阵为对称矩阵。离差阵用字母 $\boldsymbol{SS}$ 表示(有时用 $\boldsymbol{L}$ 表示)。其中:

$$ss_{ii} = \sum_{k=1}^{n}(x_{ik} - \overline{x}_i)^2, \; ss_{ij} = \sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j) \tag{1.2}$$

$\boldsymbol{SS}$ 与 $\boldsymbol{V}$ 有如下关系:

$$\boldsymbol{SS} = (n-1)\boldsymbol{V}$$

请自行验证。本例:

$$\boldsymbol{SS} = \begin{bmatrix} ss_{11} & ss_{12} & ss_{13} \\ ss_{21} & ss_{22} & ss_{23} \\ ss_{31} & ss_{32} & ss_{33} \end{bmatrix} = \begin{bmatrix} 502.9464 & 553.9831 & 354.5498 \\ 553.9831 & 765.9168 & 550.1249 \\ 354.5498 & 500.1249 & 388.5629 \end{bmatrix}$$

或:

$$\boldsymbol{SS}=\begin{pmatrix} ss_{11} & & \\ ss_{21} & ss_{22} & \\ ss_{31} & ss_{32} & ss_{33} \end{pmatrix}=\begin{pmatrix} 502.9464 & & \\ 553.9831 & 765.9168 & \\ 354.5498 & 500.1249 & 388.5629 \end{pmatrix}$$

1.1.4 相关系数矩阵

将各指标间的相关系数用矩阵的形式排列，得相关系数矩阵(correlation coefficients matrix)，简称相关阵(correlation matrix)。变量自身的相关系数为1。相关阵也是对称矩阵，用字母 $\boldsymbol{R}$ 表示。

本例：

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}=\begin{pmatrix} 1 & 0.8926 & 0.8020 \\ 0.8926 & 1 & 0.9168 \\ 0.8020 & 0.9168 & 1 \end{pmatrix}$$

或：

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & & \\ r_{21} & r_{22} & \\ r_{31} & r_{32} & r_{33} \end{pmatrix}=\begin{pmatrix} 1 & & \\ 0.8926 & 1 & \\ 0.8020 & 0.9168 & 1 \end{pmatrix}$$

对每个变量作标准化变换，即减去其均数，除以其标准差，则标准化变换后变量的协方差矩阵就等于原变量的相关矩阵。

有时，为了节约版面，最大限度地表达资料的信息，可将相关矩阵和协方差矩阵用一个矩阵来表示，对角线上表示变量的方差，下三角矩阵表示变量间的相关，而上三角矩阵表示变量间的协方差。如：

$$\begin{pmatrix} 45.7224 & 50.3621 & 32.2318 \\ 0.8926 & 69.6288 & 45.4659 \\ 0.8020 & 0.9168 & 35.3239 \end{pmatrix}$$

用这种方法表示方差-协方差及相关系数，需要特别说明。

1.2 多元正态分布

1.2.1 定义

设变量 x 服从均数为 μ，方差为 σ^2 的正态分布，则其密度函数为：

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty<x<+\infty$$

这是一元正态分布的密度函数。为了与多元正态分布相比较，将一元正态分布表示为：

$$f(x)=\frac{1}{(2\pi)^{\frac{1}{2}}(\sigma^2)^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)}$$

多元正态分布是一元正态分布的直接推广。设随机向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)'$ 服从 m 维正态分布，则：

$$f(\boldsymbol{X})=\frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{X}-\mu)'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{X}-\mu)} \tag{1.3}$$

其中，$\boldsymbol{\Sigma}$ 是变量 $x_1,x_2,\cdots,x_m$ 的协方差阵。$|\boldsymbol{\Sigma}|$ 是 $\boldsymbol{\Sigma}$ 的行列式。