UMSS

大学数学科学丛书 —18

多元统计分析

张润楚 编著

科学出版社
www.sciencep.com

大学数学科学丛书　18

多 元 统 计 分 析

张润楚　编著

科 学 出 版 社

北　京

内 容 简 介

本书系统讲述统计中多元分布的基本理论和常用的多元数据分析方法. 多元分布理论包括 Wishart 分布、T^2分布、Λ分布、多元 Beta 分布、多元正态的参数估计和假设检验及一般多元分布的参数估计和假设检验理论. 多元数据分析方法包括多元线性回归模型、判别分析、主成分分析、因子分析、相应分析、聚类分析、典型相关分析和多维标度法. 既强调作为一个学科分支的理论系统性, 对一些基本定理给出了必要而简明的数学推导, 又注重数据分析方法的多样性, 对各方法从背景、数学工具的使用、计算步骤到应用技巧及各种方法之间的联系, 都有较详细的阐述, 包括近期的一些新发展. 书中给出一些有启发性的实例和习题. 书末附录给出一些代数补充知识.

本书可作为高等院校数学系、数理统计或统计系、计量经济系、生物统计系等有关学科专业的高年级本科生、研究生学位课程的教材, 也可作为数学、生物、医学、经济、金融、工程等领域的教师或科技工作者的参考书.

图书在版编目(CIP)数据

多元统计分析/张润楚编著. 一北京: 科学出版社, 2006. 9

(大学数学科学丛书; 18 / 李大潜主编)

ISBN 7-03-017779-7

I. 多… II. 张… III. 多元分析: 统计分析 IV. O212. 4

中国版本图书馆 CIP 数据核字(2006) 第 090063 号

责任编辑: 吕 虹 赵彦超 / 责任校对: 赵桂芬

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.sciencep.com

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2006 年 9 月第 一 版 开本: B5(720×1000)

2006 年 9 月第一次印刷 印张: 23

印数: 1—3 000 字数: 522 000

定价: 46.00 元

(如有印装质量问题,我社负责调换〈路通〉)

作 者 简 介

张润楚，南开大学数学科学学院教授，博士生导师. 1966 年毕业于南开大学数学系并留校任教至今. 长期担任概率信息统计教研室副主任、统计学系主任、学校数学学科评议组委员等职.

1989 至 2005 年间曾先后应邀赴加拿大滑铁卢大学、西蒙佛雷泽大学，美国加州大学 (Berkeley)、密歇根大学，香港浸会大学等多所国际知名大学长期或短期访问和合作研究. 现兼任教育部数学与统计学教学指导委员会委员、天津市统计学会副会长、中国现场统计研究会常务理事、中国统计学会理事等职. 先后主持国家自然科学基金项目六项、教育部博士点基金项目两项等多项研究课题. 1997 年获国务院科学研究突出贡献政府津贴，2001 年获"中国高校科学技术奖"二等奖，2002 年获"全国统计科学研究优秀成果课题奖"一等奖和"博士论文优秀奖"二等奖 (指导教师) 等科研教学奖励.

曾从事多指标随机过程研究，后长期从事试验设计和多元统计分析等领域的研究，特别在试验设计领域的研究取得大量成果. 已在《中国科学》、"Biometrica"、"Statistica Sinica" 等国内外重要刊物上发表论文 80 多篇. 研究成果被 "Experiments: Planning, Analysis and Parameter Design Optimization" (C. F. J. Wu & M. Hamada, John Wiley & Sons, Inc. 2000 年出版) 等 3 部当代试验设计领域国际权威著作和许多 SCI 论文所引用. 出版著作《多元数据分析方法》(南开大学出版社)、出版译著: "Birth and Death Processes and Markov Chains" (中译英，科学出版社和 Springer 出版社)、《经典位势论与概率位势论》(科学出版社) 和《试验设计与分析及参数优化》(中国统计出版社) 等. 至今已指导博士研究生 23 名、硕士研究生 40 余名.

《大学数学科学丛书》序

按照恩格斯的说法，数学是研究现实世界中数量关系和空间形式的科学．从恩格斯那时到现在，尽管数学的内涵已经大大拓展了，人们对现实世界中的数量关系和空间形式的认识和理解已今非昔比，数学科学已构成包括纯粹数学及应用数学内含的众多分支学科和许多新兴交叉学科的庞大的科学体系，但恩格斯的这一说法仍然是对数学的一个中肯而又相对来说易于为公众了解和接受的概括，科学地反映了数学这一学科的内涵．正由于忽略了物质的具体形态和属性、纯粹从数量关系和空间形式的角度来研究现实世界，数学表现出高度抽象性和应用广泛性的特点，具有特殊的公共基础地位，其重要性得到普遍的认同．

整个数学的发展史是和人类物质文明和精神文明的发展史交融在一起的．作为一种先进的文化，数学不仅在人类文明的进程中一直起着积极的推动作用，而且是人类文明的一个重要的支柱．数学教育对于启迪心智、增进素质、提高全人类文明程度的必要性和重要性已得到空前普遍的重视．数学教育本质是一种素质教育；学习数学，不仅要学到许多重要的数学概念、方法和结论，更要着重领会到数学的精神实质和思想方法．在大学学习高等数学的阶段，更应该自觉地去意识并努力体现这一点．

作为面向大学本科生和研究生以及有关教师的教材，教学参考书或课外读物的系列，本丛书将努力贯彻加强基础、面向前沿、突出思想、关注应用和方便阅读的原则，力求为各专业的大学本科生或研究生（包括硕士生及博士生）走近数学科学、理解数学科学以及应用数学科学提供必要的指引和有力的帮助，并欢迎其中相当一些能被广大学校选用为教材，相信并希望在各方面的支持及帮助下，本丛书将会愈出愈好．

李大潜

2003 年 12 月 27 日

前　言

多元统计分析, 是数理统计近几十年来发展起来的最重要的分支之一. 它在社会经济、工农业、生物医药、环保、气象等各个领域都有广泛的应用, 尤其在信息和计算机技术发展的今天, 更显示这一分支的重要作用. 可以说, 凡是有数据信息资料的地方, 都离不开数据处理, 即科学地进行数据分析. 要有效地提取信息和用数据科学地进行推断, 就会用到多元统计理论和多元数据分析方法. 因此, 在大学理科的数理统计专业、经济和管理学科的统计专业以及生物医学学科和工科学科的有关专业, 都开设了多元分析的课程. 本书是作者在总结近些年来这一学科发展和多年从事这一课程教学的基础上编写而成的一本教材, 其内容在南开大学统计学专业讲授多年, 也被国内其他院校采用. 本书吸取了国内外有关教材的优点, 既强调多元统计作为一个学科分支的一些基本理论, 对一些基本定理给出了必要而简明的数学证明, 同时又注重各种常用多元数据分析方法基本内容的介绍, 把各种方法的背景、数学工具的使用、计算步骤以及应用技巧, 包括各种方法的联系, 都作了较详细的阐述, 还给出了一些有启发性的应用例子和习题.

本书共十二章分两部分. 第一部分从第一章到第四章, 讲述多元统计的一些基本理论, 包括多元分布理论, 主要是与多元正态分布有关的几种常用分布：Wishart 分布、T^2 分布、多元 Beta 分布和 Λ 分布, 正态向量的参数估计和参数假设检验以及一些一般理论. 第二部分从第五章到第十二章, 讲述多元数据分析的一些常用方法, 包括多元线性回归模型、判别分析、主成分分析、因子分析、相应分析、聚类分析、典型相关分析和多维标度法. 这两部分内容既有区别而又相互联系, 后者的部分内容以前者为基础. 本书还特别地给出了一些代数补充知识以及对这些结果的简洁证明 (见附录 A), 这对读者阅读本书会很有帮助.

在本书的编写过程中, 曾鹏协助作者收集和编辑了许多习题, 并阅读了书稿和完成部分图形的计算机制作, 李鹏飞、李艳婷、孙剑平、林怡为本书的习题解答、制图等做了许多工作, 特别是孙剑平、谭鲜明最后阅读书稿和为本书的计算机排版付出许多努力. 此外, 本书的出版得到科学出版社和吕虹编审的大力支持和关心, 在此作者谨向他们表示诚挚的感谢.

本书可以作为大学统计学专业本科高年级学生的多元统计分析课程和有关专业研究生的相关课程的教科书, 也可以作为研究工作者的参考书.

受作者的学识水平和时间所限，书中会存在不足，恳请同行专家和广大读者批评指正.

张润楚

2006 年 4 月于南开大学

符 号 表①

$\equiv$	恒等
$\triangleq$	“定义为”或“记为”
$\Leftrightarrow$	充分必要条件
$\perp$	“垂直于”或“正交于”
$\stackrel{d}{=}$	表示依分布相等
$A \geqslant 0$	A 为对称非负定方阵, 简称非负定阵
$A > 0$	A 为对称正定方阵, 简称正定阵
$A \geqslant B$	$A \geqslant 0, B \geqslant 0$ 且 $A - B \geqslant 0$
$A_{(n\times p)}$ 或者 $A_{n\times p}$	$n \times p$ 阶矩阵 A
$A_{(n\times p)} = (\mathrm{a}_1, \cdots, \mathrm{a}_n)'$	矩阵 A 的行向量表示
$A_{(n\times p)} = (\mathrm{a}_{(1)}, \cdots, \mathrm{a}_{(p)})$	矩阵 A 的列向量表示
A^-	矩阵 A 的广义逆
A^+	矩阵 A 的加号逆, 即 Moore-Penrose 广义逆
$\|A\|$ 或者 $\det A$	矩阵 A 的行列式
$\|\|A\|\|$	矩阵 A 的范数
$\mathrm{rk}(A)$	矩阵 A 的秩
$\mathrm{tr}(A)$	矩阵 A 的迹
$\mathcal{M}(A)$	矩阵 A 的列向量张成的子空间
$\mathcal{M}(A)^\perp$	矩阵 A 的列向量张成的子空间的正交补空间
$\lambda_i(A)$	矩阵 A 第 i 个顺序特征值
$\mathrm{Vec}(A)$ 或者 A^{V}	矩阵 A 的拉直, 即将 A 的列向量依次相接排成的列向量
$A \otimes B$	矩阵 A 和 B 的 Kronecker 乘积
$\mathbf{1} = (1, \cdots, 1)'$	分量均为 1 的列向量
X	表示非随机矩阵
$\mathbf{X}$	表示随机矩阵
x	表示非随机向量

① 符号表中的符号是书中的一般记法和意义, 阅读时还要注意上下文.

$\mathbf{x}$	表示随机向量
x	表示标量
$\boldsymbol{x}$	表示随机变量或随机向量的分量 (并有下标时)
$\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\mu}, \cdots$	表示参数向量
$E(\boldsymbol{x}), E(\mathbf{x}), E(\mathbf{X})$	$\boldsymbol{x}, \mathbf{x}, \mathbf{X}$ 的数学期望
$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}), \mathrm{Var}(\boldsymbol{x})$	$(\boldsymbol{x}, \boldsymbol{y})$ 两分量的协方差, $\boldsymbol{x}$ 的方差
$\mathrm{Cov}(\mathbf{x})$ 或 $\mathrm{Var}(\mathbf{x})$	$\mathbf{x}$ 的协方差阵
$\mathrm{Cov}(\mathbf{x}, \mathbf{y})$	$\mathbf{x}, \mathbf{y}$ 的协方差阵
$\mathrm{Corr}(\boldsymbol{x}, \boldsymbol{y})$	$\boldsymbol{x}, \boldsymbol{y}$ 的相关系数
R^p	p 维向量空间
$\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$	$\mathbf{x}$ 为均值为 $\boldsymbol{\mu}$ 协方差为 Σ 的随机向量
$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$	$\mathbf{x}$ 服从均值为 $\boldsymbol{\mu}$、协方差为 Σ 的 p 维正态分布
$N_{n,p}(M, \Sigma, V)$	参数为 M, Σ, V 的 $n \times p$ 正态随机阵
$t(n), \quad t_\alpha(n)$	自由度为 n 的中心 t 分布及其上侧 α 分位点
$t(n, \delta)$	自由度为 n 和参数 δ 的非中心 t 分布
χ^2_p 或 $\chi^2(p)$	自由度为 p 的中心 χ^2 分布
$\chi^2_\alpha(n)$	自由度为 n 的中心 χ^2 分布上侧 α 分位点
$\chi^2_p(\lambda)$ 或 $\chi^2(p, \lambda)$	自由度为 p 和参数 λ 的非中心 χ^2 分布
$F(m, n), \quad F_\alpha(m, n)$	自由度m,n的中心F分布及其上侧α分位点
$F(m, n, \lambda)$	自由度m,n和参数λ的非中心F分布
$W_p(n, \Sigma)$,	参数为p, n, Σ的中心Wishart分布
$W_p(n, \Sigma, Q)$	参数为 p, n, Σ, Q 的非中心 Wishart 分布
$T^2(p, n), \ T^2_\alpha(p, n)$	参数 p, n 的中心 T^2 分布及其上侧 α 分位点
$T^2(p, n, \lambda)$	参数 p, n, λ 的非中心 T^2 分布
$\Lambda(p, n, m), \ \Lambda_\alpha(p, n, m)$	参数 p, n, m 的 Wilks Λ 分布及其下侧 α 分位点
$\mathrm{Plim}_{n \to \infty} \hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$ 依概率的极限为 $\boldsymbol{\theta}$
$\xrightarrow{\mathcal{L}}$	依分布收敛

$\xrightarrow{\text{a.s.}}$	几乎处处收敛
$\xrightarrow{P}$	依概率收敛
$J_{Y\to X}$ 或者 $J(Y\to X)$	由 Y 到 X 变换的 Jacobi 行列式
cf	特征函数
iid	独立同分布
pdf	概率分布密度函数
MLE	极大似然估计
LSE	最小二乘估计
MVUE	最小方差无偏估计
BLUE	最优线性无偏估计
MSE	均方误差
LRT	似然比检验或检验统计量
UIT	并交检验或检验统计量
UMP 检验	一致最大功效检验
PCR	主成分回归
PLSR	偏最小二乘回归

目　录

*　　*　　*

第一章　随机向量和多元正态分布

§1.1　随机向量及有关概念

1.1.1　随机向量定义

许多随机现象用一个变量来反映是不够的, 而需用多个变量来反映. 例如人的身材用身高体重两个变量来反映, 一个国家的经济状态则需要用多个经济指标来刻画.

定义 1.1.1　向量 $\mathbf{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$ 称为一个 p 维随机向量, 如果该向量在空间 R^p 中存在概率分布, 即对任何 $(y_1, \cdots, y_p) \in R^p$, 概率

$$F(y_1, \cdots, y_p) = P(\boldsymbol{x}_1 \leqslant y_1, \cdots, \boldsymbol{x}_p \leqslant y_p) \tag{1.1.1}$$

存在. 并称 p 元函数 (1.1.1) 为 $\mathbf{x}$ 的分布函数, 记 $\mathbf{x} \sim F$, 称 $\mathbf{x}$ 服从 F 分布, 也称 F 为 $\mathbf{x}$ 的联合分布函数.

如果一个随机向量 $\mathbf{x}$, 有 R^p 中的非负函数 $f(\mathrm{x})$, 使得其分布函数可表示为积分

$$F(\mathrm{x}) = \int_{-\infty}^{\mathrm{x}} f(\mathrm{u}) du, \qquad \mathrm{x} \in R^p, \tag{1.1.2}$$

则称 $\mathbf{x}$ 为连续型随机向量. 称 $f(\mathrm{x})$ 为 $\mathbf{x}$ 的概率分布密度函数 (简记 pdf).

例 1.1.1　设二维随机向量 $\mathbf{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)'$ 有密度

$$f(x_1, x_2) = \begin{cases} e^{-(x_1+x_2)}, & x_1 \geqslant 0, x_2 \geqslant 0, \\ 0, & \text{其他}, \end{cases}$$

则由 (1.1.2) 可算得它的分布函数为

$$F(x_1, x_2) = \begin{cases} 1 - e^{-x_1} - e^{-x_2} + e^{-(x_1+x_2)}, & x_1 \geqslant 0, x_2 \geqslant 0, \\ 0, & \text{其他}. \end{cases}$$

称一个随机向量 $\mathbf{x}$ 是离散的, 如果它的全部概率集中在一个有穷或可数个点的集合 $\mathcal{D} = \{\mathrm{x}_i, i = 1, 2, \cdots\}$ 上.

$$\begin{cases} f(\mathrm{x}_i) = P(\mathbf{x} = \mathrm{x}_i), & i = 1, 2, \cdots \\ f(\mathrm{x}) = 0, & \text{其他} \end{cases} \tag{1.1.3}$$

称为 $\mathbf{x}$ 的概率函数. 离散随机向量也可以用连续型的方法来处理, 只要把概率函数 (1.1.3) 视为 $\mathbf{x}$ 的 (离散) pdf, 而且 (1.1.2) 中的积分号用和号来代替, 即

$$P(\mathbf{x} \in D_{\mathrm{y}}) = \sum_{\mathrm{x}_i \in D_{\mathrm{y}} \cap \mathcal{D}} f(\mathrm{x}_i), \text{ 其中 } D_{\mathrm{y}} = \{\mathrm{x}: \mathrm{x} \leqslant \mathrm{y}\}. \tag{1.1.4}$$

$\mathbf{x}$ 的支撑集定义为

$$S = \{\mathrm{x} \in R^p: f(\mathrm{x}) > 0\}.$$

以后我们写一个 pdf, 只写它在其支撑集上的值, 其他点上的值自然为零.

1.1.2 边际分布和条件分布

考虑分块的 p 维随机向量 $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$, 其中 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 分别有 k 个和 $p-k$ 个分量. 称函数

$$P(\mathbf{x}_1 \leqslant \mathrm{x}_1) = F(x_1, \cdots, x_k, \infty, \cdots, \infty)$$

为 $\mathbf{x}_1$ 的边际分布函数 (marginal cdf), 记为 $F_1(\mathrm{x}_1)$, 其中 $\mathbf{x}_1 \leqslant \mathrm{x}_1$ 表示 $\mathbf{x}_1$ 的各分量小于或等于 x_1 的对应分量. 设 $\mathbf{x}$ 有联合 pdf $f(\mathrm{x})$, 则 $\mathbf{x}_1$ 的边际 pdf 由 $f(\mathrm{x})$ 对 x_2 的积分给出, 即

$$f_1(\mathrm{x}_1) = \int_{-\infty}^{\infty} f(\mathrm{x}_1, \mathrm{x}_2) d\mathrm{x}_2. \tag{1.1.5}$$

$\mathbf{x}_2$ 的边际 pdf $f_2(\mathrm{x}_2)$ 的定义类似. 给定 $\mathbf{x}_1 = \mathrm{x}_1^0$ 下, $\mathbf{x}_2$ 的条件 pdf 为

$$f(\mathrm{x}_2|\mathrm{x}_1 = \mathrm{x}_1^0) = \frac{f(\mathrm{x}_1^0, \mathrm{x}_2)}{f_1(\mathrm{x}_1^0)}, \quad f_1(\mathrm{x}_1^0) \neq 0.$$

类似地, 在给定 $\mathbf{x}_2 = \mathrm{x}_2^0$ 下, $\mathbf{x}_1$ 的条件 pdf 为 $f(\mathrm{x}_1|\mathrm{x}_2 = \mathrm{x}_2^0) = \frac{f(\mathrm{x}_1, \mathrm{x}_2^0)}{f_2(\mathrm{x}_2^0)}$, $f_2(\mathrm{x}_2^0) \neq 0$.

注意, 有相同边际分布的两个随机向量未必有相同联合分布. 例如, 两个联合 pdf 分别为

$$f(x_1, x_2) = 1, \quad 0 < x_1 < 1, 0 < x_2 < 1 \tag{1.1.6}$$

和

$$f(x_1, x_2) = 1 + \alpha(2x_1 - 1)(2x_2 - 1), \quad 0 < x_1 < 1, 0 < x_2 < 1, -1 \leqslant \alpha \leqslant 1, \alpha \neq 0 \tag{1.1.7}$$

的随机向量有相同的边际分布, 即 $(0,1)$ 上的均匀分布, 但联合分布却不同.

1.1.3 独立性

对于随机向量 $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$, 如果条件 pdf $f(\mathrm{x}_1|\mathrm{x}_2 = \mathrm{x}_2^0)$ 与 x_2^0 无关, 且等于 $f_1(\mathrm{x}_1)$, 则说 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 独立. 容易证明以下定理: