多元统计分析

史宁中 主编

概率统计系列 2

研究生教学丛书

王静龙 /著

科学出版社

www.sciencep.com

概率统计系列研究生教学丛书 2

多元统计分析

王静龙 著

科学出版社
北 京

内 容 简 介

本书系统讲解多元统计分析的基本理论与应用方法，同时包含了一些新近发展起来的理论丰富且有实用价值的内容. 本书内容包括多元正态分布及由其导出的分布、多元正态分布的参数估计与检验问题、线性模型、相关分析、判别分析以及聚类分析，结合案例分析讲解多元统计分析的理论与方法.

本书可作为统计专业研究生和高年级本科生的教材使用，同时也可供统计工作者、科技人员和高校相关专业的教师与学生阅读参考.

图书在版编目（CIP）数据

多元统计分析/王静龙著. —北京：科学出版社，2008

（概率统计系列研究生教学丛书；2）

ISBN 978-7-03-021555-0

Ⅰ. 多… Ⅱ. 王… Ⅲ. 多元分析：统计分析 Ⅳ. O212.4

中国版本图书馆 CIP 数据核字(2008) 第 044770 号

责任编辑：范庆奎 房 阳 / 责任校对：陈玉凤

责任印制：赵德静 / 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

丽源印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2008 年 6 月第 一 版 开本：B5（720×1000）

2008 年 6 月第一次印刷 印张：31 1/4

印数：1—4 000 字数：592 000

定价：68.00 元

（如有印装质量问题，我社负责调换〈新欣〉）

《概率统计系列研究生教学丛书》序

概率论与数理统计是一门研究随机现象规律性的数学学科. 它一方面有自己独特的概念和方法, 形成了结构宏大的理论; 另一方面, 它与其他数学分支又有紧密的联系, 它是近代数学的重要组成部分. 在培养高素质科学技术人才中具有其独特的、不可替代的重要作用. 它不单是一种知识、方法或工具, 更在于它可以有效地培养和训练学生的随机思维模式、培养学生的一种素养.

大体上说, 概率论是统计学的理论和方法的依据, 而统计学可视为概率论的一种应用. 统计方法的应用促进了科学技术的进步; 反过来科技的进步推动了统计学突飞猛进的发展. 统计学的一些新方法应运而生, 比如 EM 法、GEE 方法、MCMC 方法、经验似然、贝叶斯网络、大维数据分析等. 而计算机技术和信息技术的飞速发展为数据分析的复杂化和多样化提供了强有力的平台, 过去许多不敢想像的方法成为可能, 如 Data Mining, Bootstrap 和 Jack-knife 等方法. 英国统计学家哈斯利特说:"统计方法的应用是这样普遍, 在我们的生活和习惯中, 统计的影响是这样巨大, 以致统计的重要性无论怎样强调也不过分."

为了适应国内概率统计教学的现状以及社会对人才培养的需求, 并拓宽统计学应用的领域, 在科学出版社的大力支持下, 我们组织了一批专家编写了该系列适用于概率统计专业高年级本科生、研究生以及有关教师的教材 (教学参考书). 该丛书力求提高理论水平、突出前沿思想、侧重实际应用和学科渗透, 其中凝聚了该系列丛书作者的多年教学和科研经验.

我们衷心希望该系列丛书的出版能为我国高等院校教学改革作出贡献, 更希望能促进统计学在诸多领域的广泛应用.

史宁中

2008 年 5 月

于东北师范大学

前言

多元统计分析的理论很丰富, 应用非常广泛. 国内绝大多数高校, 包括华东师范大学统计系, 都把多元统计分析当作统计方向硕士研究生的学位必修课. 我们先后采用了张尧庭与方开泰两位教授所著的《多元统计分析引论》和 Johnson 与 Wichern 的*Applied Multivariate Statistical Analysis* 作为多元统计分析课的教学用书. 在二十多年教学的过程中, 我们参考了很多国内外有关多元统计分析的经典著作, 其中有 Anderson 的*An Introduction to Multivariate Statistical Analysis*, Giri 的*Multivariate Statistical Inference*, Muirhead 的*Aspects of Multivariate Statistical Theory*, 方开泰编著的《实用多元统计分析》, 王学仁和王松桂编译的《实用多元统计分析》. 此外, 特别需要提到的是, 同济大学的王福宝曾将 Srivastava 和 Khatri 的*An Introduction to Multivariate Statistics* 翻译成中文, 承蒙他给了我此书的翻译打印稿《多元统计学导引》, 它也是我们多元统计分析课教学的主要参考用书之一. 在教学过程中, 我们陆续添加了一些新近发展起来的、理论丰富且有实用价值的内容, 并在听取学生意见的基础上尝试用一些新的方法去处理教材. 本书就是结合我和我学生的研究工作, 将陆续编写的讲义经过多次修改整理加工后写成的.

本书系统地讲解了多元统计分析的基本理论与应用方法. 全书共分 9 章. 第 1 章通过实例阐述多元统计分析的特点, 激发学生学习多元统计分析的兴趣; 第 2 章介绍多元正态分布及其有关性质; 第 3 章介绍由多元正态分布导出的 Wishart 分布、Hotelling T^2 分布、Wilks 分布以及这些分布的性质; 第 4 章讨论多元正态分布参数的估计问题; 第 5 章讨论多元正态分布均值参数的检验问题; 第 6 章讨论多元正态分布协方差阵参数的检验问题; 第 7 章讨论多元线性模型、多元线性回归模型和重复测量模型的估计与检验问题; 第 8 章讨论相关分析, 包括典型相关分析、主成分分析、因子分析与协方差选择模型; 第 9 章简要介绍多元统计分析的一些应用, 有判别分析与聚类分析. 本书标有 “*” 号的章节和习题可以跳过去, 这并不影响全书阅读的连贯性.

本书力求讲清楚多元正态分布和由它导出的分布. 关于多元正态分布, 本书在讲相关系数的时候还讲了条件独立性, 即偏相关系数与精度矩阵的关系. 关于 Wishart 分布的密度函数的推导, 本书正文采用的是数学归纳法. 这种方法比较初等, 仅需要基本的概率论与数理统计的知识, 读者容易理解. 考虑到二元正态分布用得比较多, 并且它是多元正态分布的一个缩影, 本书在推导任意阶 Wishart 分布的密度函数之前, 先行推导二阶 Wishart 分布的密度函数, 这能使读者对 Wishart 分布有一个直观的理解. 此外, 在习题和附录中还分别用 Bartlett 分解、许氏公

式和不变测度推导了 Wishart 分布的密度函数. 多元统计检验问题的渐近 p 值的计算是一个很重要的问题. 本书在讲 Wilks 分布时, 还讲了它的分布函数的渐近展开, 为以后计算渐近 p 值作准备. 本书还简要地介绍了非中心的 Wishart 分布和 Hotelling T^2 分布.

随着经济的发展, 处理多个变量的观察数据的多元统计理论和方法也不断完善与发展. 除了常见的多元统计方法之外, 本书还介绍了多重比较、变点检验、序约束下有方向的检验问题、重复测量模型和协方差选择模型等多元统计方法. 本书力求讲清楚多元统计方法的实际背景和统计思想. 本书以成年人上衣服装号型的制定为例引入主成分分析, 这个例子在本书进行了多次讨论, 这不仅有助于激发读者学习多元统计分析的兴趣, 而且有助于读者理解多元统计分析的有关概念与方法. 通常人们是用关于智力的定义和测量的实例引入因子分析的, 除此之外, 本书还用新近发展起来的满意度指数模型引入因子分析. 根据解实际问题的需要, 本书对典型相关变量个数的检验、主成分特征根的统计推断、因子模型的估计与检验、费希尔判别函数个数的检验等问题进行了较为详细的讨论.

本书将与多元统计分析有关的一些内容放在附录中, 其中有多元特征函数、矩阵代数、二次型及其极值、变换的雅可比行列式和求导、指数分布族与条件独立性等. 这些内容对于理解正文的有关内容是很重要的, 而将它们放入附录的目的是为了在正文中突出多元统计分析的理论与方法. 用许氏公式和不变测度推导 Wishart 分布密度函数的证明比较抽象, 也放入附录.

我要特别感谢张尧庭和方开泰两位教授, 他们的书《多元统计分析引论》和他们关于多元统计分析的讲座影响了中国好几代从事多元统计分析的工作者. 我就是被他们的著作和讲座所吸引, 从而对多元统计分析产生了浓厚的兴趣. 如今, 张尧庭教授已过世, 仅以此书寄托我们对他的怀念与哀思. 方开泰教授给了我很多指导与帮助, 谨向他致以衷心的感谢.

我也要特别感谢林举干教授, 在我之前, 华东师范大学统计系硕士研究生的多元统计分析课是他任教的. 我的教学工作得到了他很多的帮助与启发. 我还要感谢徐进博士, 他仔细且评判性地阅读了本书底稿, 提出了很多修改意见. 我还要感谢华东师范大学统计系的历届研究生、访问学者和进修教师, 正因为有了他们的参与, 我对多元统计分析课的科研和教学越来越有体会, 否则难以想像本书能够成稿.

本书得到了华东师范大学研究生教材基金的资助, 在此对华东师范大学研究生院以及培养处束金龙教授表示衷心的感谢.

限于作者水平, 书中必有疏漏和需要改进之处, 恳请大家批评指正.

王静龙

2007 年 12 月

目录

第 1 章　引　　言

在生产、技术、社会、经济以及管理等领域中, 人们常常需要同时观察多个变量. 通常有两种不同的方法来处理多个变量的观察数据. 把多个变量分开来进行研究, 一次分析一个变量, 这是一种方法. 另一种方法就是本书介绍的多元统计分析方法, 它把多个变量合在一起进行研究.

前一种方法仅需使用分析单个变量的单元统计分析方法, 比较简单, 但它没有考虑变量之间的相互关系. 在变量之间具有相关关系时, 倘若把它们分开来进行研究, 就会丢失变量之间相关的信息, 其分析结果很可能不是有效的. 后一种方法将变量合在一起进行研究, 研究它们之间的相互关系, 正确地揭示这些变量内在的相关数量变化规律, 其分析结果通常是有效的. 看下面的例子:

英国著名统计学家 K. Pearson(1857~1936) 曾进行了一项研究[114], 研究家庭成员间的相似性. 作为这项研究的一部分, 他测量了 1078 个父亲及其成年儿子的身高. 经计算,

父亲平均身高为 68in(即 172.7cm), 标准差 SD 为 2.7in(即 6.86cm);

儿子平均身高为 69in(即 175.3cm), 标准差 SD 为 2.7in(即 6.86cm);

它们之间的相关系数为 0.5.

我们的讨论与 Pearson 的讨论有所不同. 我们欲解决的问题是, 希望得到一个区域 D, 使得有如 95%的家庭, 其家庭成员中父亲及其成年儿子的身高在这个区域 D 中, 即使得

$$P\left((X,Y)\in D\right)=95\%, \tag{1.1}$$

其中, X, Y 分别表示家庭成员中父亲及其成年儿子的身高. 下面的第一个做法就是将父亲和儿子的身高分开来进行分析, 然后将所得到的结果合在一起从而求得区域 D. 看看这样的分析方法有什么缺陷.

在正常的情况下, 人的生理测量值, 如身高、体重等都服从正态分布. 设父亲的身高 $X\sim N(\mu,\sigma^2)$. 由于此项研究测量了很多 (1078 个) 父亲的身高, 故不妨认为

$$\mu=172.7\text{cm},\quad \sigma=6.86\text{cm}.$$

在 $X\sim N(172.7, 6.86^2)$ 时, 若记标准正态分布 $N(0,1)$ 的 γ 分位点为 U_γ, 则有

$$P\left(X\in\left[172.7-6.86U_{1-\alpha/2}, 172.7+6.86U_{1-\alpha/2}\right]\right)=1-\alpha, \tag{1.2}$$

所以有 $1-\alpha$ 比例的家庭, 其家庭成员中父亲的身高在 $172.7-6.86U_{1-\alpha/2}$ 和 $172.7+6.86U_{1-\alpha/2}$ 之间. 同理, 在成年儿子的身高 $Y\sim N(175.3,6.86^2)$ 时, 有

$$P\left(Y\in\left[175.3-6.86U_{1-\alpha/2},175.3+6.86U_{1-\alpha/2}\right]\right)=1-\alpha, \tag{1.3}$$

所以有 $1-\alpha$ 比例的家庭, 其家庭成员中成年儿子的身高在 $175.3-6.86U_{1-\alpha/2}$ 和 $175.3+6.86U_{1-\alpha/2}$ 之间.

考虑到 $\sqrt{95\%}=97.47\%$, 故取 $\alpha=1-97.47\%=2.53\%$, 从而得

$$U_{1-\alpha/2}=U_{0.98735}=2.237,$$

则由 (1.2) 式和 (1.3) 式知

$$\begin{cases}P\left(X\in[159.29,186.11]\right)=97.47\%,\\ P\left(Y\in[161.89,188.71]\right)=97.47\%.\end{cases} \tag{1.4}$$

若取 D 为矩形 (图 1.1)

$$D=\{(x,y):159.29\leqslant x\leqslant 186.11,161.89\leqslant y\leqslant 188.71\},$$

则有

$$P\left((X,Y)\in D\right)=97.47\%\times 97.47\%=95\%, \tag{1.5}$$

所以 (1.1) 式成立. 事实上, 对本例来说, D 是正方形.

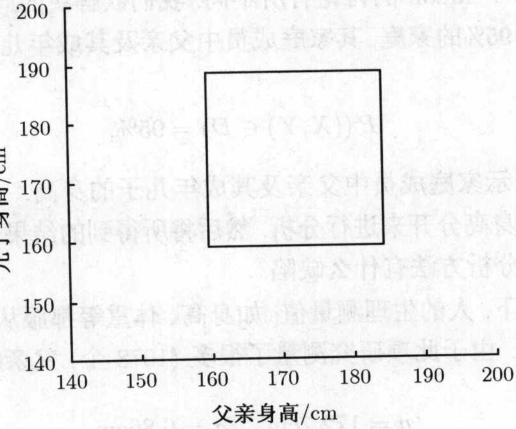

图 1.1

下面分析上述解法有什么不足之处. 可以知道, 家庭成员中父亲的身高与其成年儿子的身高正相关, 它们并不相互独立, 所以由 (1.4) 式成立并不能保证 (1.5) 式

成立. 这是上述解法的一个很明显的不足之处. 同时, 也正因为家庭成员中父亲的身高与其成年儿子的身高正相关, 所以如果父亲长得高, 其成年儿子往往也比较高, 而如果父亲比较矮, 其成年儿子往往长得也比较矮, 因此, 这个正方形的左上角 (父亲矮但成年儿子高) 和右下角 (父亲高但成年儿子矮) 不大可能发生, 没有必要保留. 这也就是说, 这个正方形应该被切成如图 1.2 所示的形状.

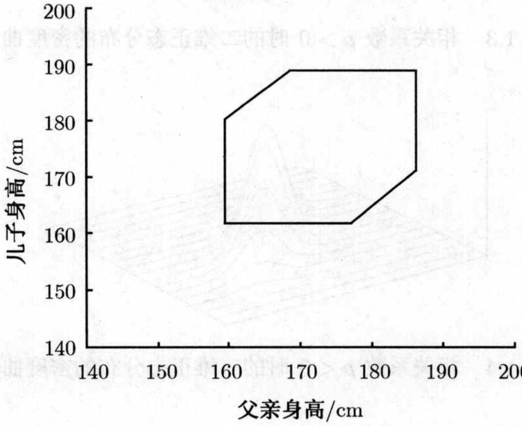

图 1.2

这启发我们, 使得 (1.1) 式成立的区域 D 取成长轴方向自左向右向上的椭圆比较恰当. 这是上述解法的又一个不足之处. 所以在变量之间存在相关关系时, 倘若分开来进行研究, 分析结果很可能不是有效的.

下面的第 2 个做法就是把父亲的身高 X 与其成年儿子的身高 Y 合在一起进行研究, 从而构造椭圆型区域 D. 二元正态分布 $(X,Y) \sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ 的密度函数为

$$
\begin{aligned}
p(x,y) = &\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\
&\cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}. \quad (1.6)
\end{aligned}
$$

如同一元正态分布, 二元正态分布的密度曲面也是单峰对称的 (图 1.3 ~ 图 1.5). 图 1.3 是相关系数 $\rho > 0$ 时的二维正态分布的密度曲面. 由此图可见, $\rho > 0$ 时二维正态分布的值集中在方向自左向右向上 (斜率大于 0) 的直线附近. 图 1.4 是相关系数 $\rho < 0$ 时的二维正态分布的密度曲面. 由此图可见, $\rho < 0$ 时二维正态分布的值集中在方向自左向右向下 (斜率小于 0) 的直线附近. 图 1.5 是相关系数 $\rho = 0$, 即 X 与 Y 相互独立时的二维正态分布的密度曲面.

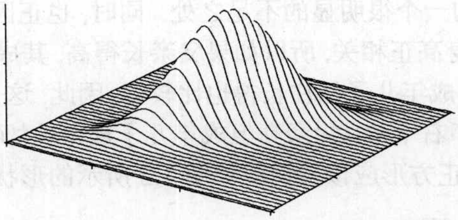

图 1.3　相关系数 $\rho > 0$ 时的二维正态分布的密度曲面

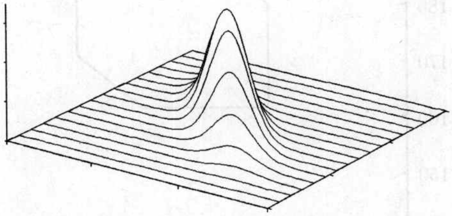

图 1.4　相关系数 $\rho < 0$ 时的二维正态分布的密度曲面

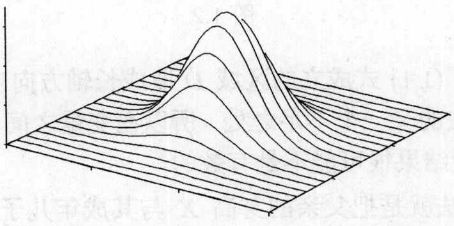

图 1.5　相关系数 $\rho = 0$ 时的二维正态分布的密度曲面

图 1.6 ~ 图 1.8 分别是与图 1.3 ~ 图 1.5 相对应的二维正态分布密度等高线. 密度等高线是椭圆曲线. $\rho > 0$ 时椭圆曲线长轴的方向自左向右向上 (图 1.6), $\rho < 0$ 时椭圆曲线长轴的方向自左向右向下 (图 1.7). $\rho = 0$ 时椭圆曲线的长轴和短轴分别平行于坐标 X 轴和 Y 轴或 Y 轴和 X 轴 (图 1.8). 由二维正态分布的密度函数表达式 (1.6) 知, 密度等高椭圆曲线的方程为

$$\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} = c. \tag{1.7}$$

取不同的 c 就得到不同的密度等高椭圆曲线. c 越大, 密度等高椭圆曲线就越大. 小的 c 的密度等高椭圆曲线在大的 c 的密度等高椭圆曲线的里面.