多元统计分析

杨维权　刘兰亭　林鸿洲　编

高等教育出版社

本书的出版得到了
加拿大驻华大使馆发展处
公民社会项目的资助
在此 表示衷心的感谢

中国和加拿大的社区发展

主编 陈启能 姜芃

民族出版社
The Ethnic Publishing House

内容提要

全书内容共分十章、三个附录以及两个补充，前四章主要讨论多元统计分析的理论基础，包括多元正态分布及其性质、参数估计、抽样分布、假设检验；后六章介绍较实用的多元统计分析方法，包括回归分析、线性模型、经济计量模型、判别分析与聚类分析、主成分分析与因子分析、典型相关分析。

本书可供数学、概率统计、计算数学等专业作为选修课教材使用，也可供工、农、医、经济等学科教学多元统计分析时选用.

多元统计分析

杨维权　刘兰亭　林鸿洲　编

*

高等教育出版社出版

新华书店北京发行所发行

二二〇七工厂印装

*

开本 850×1168　1/32　印张 13　字数 310 000

1989 年6月第1版　1989 年6月第1次印刷

印数 0 001— 2 600

ISBN7-04-001023-2/O·654

定价 3.45 元

说　　明

本书是为数学、概率统计、计算数学等专业开设多元统计分析选修课而编写的. 对于工科、农医、经济等学科开设数理统计选修课，本书也可适用. 全书由两大部分组成：第一部分是前四章，讲述了多元统计分析的理论基础，包括多元正态分布及其性质、参数估计、抽样分布、假设检验；第二部分是后六章，包括回归分析、线性模型、经济计量模型、判别分析与聚类分析、主成分分析与因子分析、典型相关分析，它们是多元统计分析中较实用的各种方法. 各章之间的相互关系如下图所示：

$$
\begin{tikzcd}
 & & & & & 6 \\
 & & & & 5 \arrow[ur] \arrow[dr] & \\
1 \arrow[r] & 2 \arrow[r] & 3 \arrow[r] & 4 \arrow[ur] \arrow[dr] & & 7 \\
 & & & & 8,9,10 &
\end{tikzcd}
$$

在全书的安排上，既注意到教学方便，各章都配有习题；还注意到理论与实践相结合，后六章都有实例讲解，这些实例涉及到地质、生化、农医、经济及教育心理学等方面. 对于概率统计基础知识不大熟悉的应用工作者，可直接阅读第 8、9、10 后三章，通过实例掌握这几种实用的统计方法.

为使适用面更广，本书没有采用测度论工具，只需要普通的微积分及概率统计基础知识. 本书编写了三个附录、两个补充，不占用教学时数，可供教学参考. 附录 I 给出了矩阵代数方面的内容，为教学提供方便. 考虑到教学时数的限制，将前四章中几个定理的详细证明作为附录 II，供教学参考. 附录 III 是介绍参数检验的并交原理，充实假设检验的内容. 补充 I 是简介随机成分向量的统计分析及其应用，这个专题的应用面较广. 补充 II 是简介椭

球等高分布族，这是多元统计分析近代发展的一个新课题，有利于扩大知识面。

在编写过程中，我们遵守下述基本原则：1. 保证理论的完整性和严谨性，对于多元统计分析中的基本概念和定理，作较全面且较严格的叙述；2. 为使本书的应用性较强，尽量做到阐述基本概念、定理和方法的客观实际意义，所述各种统计方法都用实例作示范；3. 全书力求简明扼要，重点突出，叙述清楚。

本书由杨维权（中山大学）主持编写。其中，第一、二、三、四、七各章及附录II、补充I、补充II由杨维权执笔，余锦华协助配前四章习题；第六、九、十各章及附录I、附录III由刘兰亭（山西大学）执笔；第五、八章由林鸿洲（山东海洋学院）执笔。杨维权负责全书修改定稿。

本书初稿写于1983年，在中山大学、山西大学、山东海洋学院、山西师范大学等院校几经试用与修改，并于1986年9月在太原市召开了审稿会，到会专家认真地审阅了全书，并就书的结构、选材和论述提出了许多重要的改进意见。在此基础上，我们对全书又作了进一步的修改。谨向他们表示衷心的感谢。

梁之舜教授及张尧庭、方开泰、林少宫、魏宗舒，常学将等教授对本书的编写一直关注与支持，王嵩骧、王学仁、邓集贤、邓炜材、周光亚等同行专家及高尚华、周纪乡、谢平民等同志对本书初稿提出了许多宝贵的意见，我们衷心感谢他们。

由于编者水平有限，本书难免还会有不少缺点及不当之处，恳请读者批评指正。

编　　者

1986年12月

目　录

第一章　多元正态分布

数理统计学的重要基础之一是一元正态分布，因为大多数实际问题中的随机变量服从正态分布，如长度(高度、厚度、宽度等)、重量、含量、容量等测量误差作为随机现象研究其统计规律性，都属于这种情形. 对于一元正态分布的研究早已较为完美，并建立了一整套的统计推断方法(如估计理论、假设检验原理及其在回归分析、试验设计等各种统计方法中的应用). 由于正态分布在理论研究及实际应用中起着重要的作用，因此对于非正态分布的随机变量，在什么样的条件下渐近地服从正态分布，也研究得较为充分. 多元统计分析是研究客观事物中多个变量(或多个因素)之间相互依赖的统计规律性，它的重要基础之一是多元正态分布.

多元正态分布是最常用的一种多元概率分布. 除此之外，还有多元对数正态分布、多项式分布、多元 β 分布、多元 χ^2 分布、多元指数分布等.

§1.1　多元分布的基本概念

多元分布的基本概念，可由二元概率分布的自然推广而得到，如联合分布、边沿分布、条件分布、独立性、特征函数、数字特征等. 对此，只作扼要的综述.

定义 1.1.1　随机向量、随机矩阵. 称由 p 个随机变量 x_1, $x_2, \cdots, x_p$ 构成的 p 维列向量

$$X=(x_1,x_2,\cdots,x_p)'=\begin{pmatrix}x_1\\x_2\\\vdots\\x_p\end{pmatrix}$$

为 p 维随机向量，它的概率分布称为 p 元分布. 称

$$X=(x_{ij})_{p\times n}$$

为随机矩阵，是指矩阵中每个元素 x_{ij} 都为一随机变量，常记

$$X=(x_{ij})_{p\times n}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&&\vdots\\x_{p1}&x_{p2}&\cdots&x_{pn}\end{pmatrix}_{p\times n}$$

$$=(X^{(1)},X^{(2)},\cdots,X^{(n)})_{p\times n}=\begin{pmatrix}X'_{(1)}\\X'_{(2)}\\\vdots\\X'_{(p)}\end{pmatrix}_{p\times n}\tag{1.1.1}$$

在此以 $X^{(1)},\cdots,X^{(n)}$ 表示 X 的列向量，$X_{(1)},\cdots,X_{(p)}$ 表示 X 的行向量的转置向量. X 是 $p\times n$ 阶随机矩阵，它可看成 n 个 p 维(列)随机向量所构成的，也可看成 p 个 n 维(行)随机向量构成的. X 的概率分布是指按列拉直的全体元素(随机变量) $x_{11},\cdots,x_{p1},x_{12},\cdots,x_{p2},\cdots,x_{1n},\cdots,x_{pn}$ 组成的 pn 元分布.（注：以后称向量都是指列向量.）

定义 1.1.2 数字特征. 依次称

$$E(X)=\begin{pmatrix}Ex_1\\Ex_2\\\vdots\\Ex_p\end{pmatrix},\qquad E(Y)=\begin{pmatrix}Ey_1\\Ey_2\\\vdots\\Ey_q\end{pmatrix}$$

为 $X=(x_1,x_2,\cdots,x_p)'$ 及 $Y=(y_1,y_2,\cdots,y_q)'$ 的均值向量，称

$$E(X)=(Ex_{ij})_{p\times n}$$

为随机矩阵 **X** 的均值矩阵. 称

$$\mathrm{Cov}(X,X)=E[X-E(X)][X-E(X)]'$$

$$=\begin{pmatrix}\mathrm{Cov}(x_1,x_1) & \cdots & \mathrm{Cov}(x_1,x_p)\\ \mathrm{Cov}(x_2,x_1) & \cdots & \mathrm{Cov}(x_2,x_p)\\ \vdots & & \vdots\\ \mathrm{Cov}(x_p,x_1) & \cdots & \mathrm{Cov}(x_p,x_p)\end{pmatrix}_{p\times p} \qquad (1.1.2)$$

为 p 维随机向量 X 的自协方差阵，简称为 X 的协方差阵. 称 $|\mathrm{Cov}(X,X)|$ 为 X 的广义方差，它是协方差阵的行列式之值. 称

$$\mathrm{Cov}(X,Y)=E[X-E(X)][Y-E(Y)]'$$

$$=\begin{pmatrix}\mathrm{Cov}(x_1,y_1) & \cdots & \mathrm{Cov}(x_1,y_q)\\ \mathrm{Cov}(x_2,y_1) & \cdots & \mathrm{Cov}(x_2,y_q)\\ \vdots & & \vdots\\ \mathrm{Cov}(x_p,y_1) & \cdots & \mathrm{Cov}(x_p,y_q)\end{pmatrix}_{p\times q} \qquad (1.1.3)$$

为 X 与 Y 的互协方差阵，简称为 X 与 Y 的协方差阵.

注意：有时记 $\mathrm{Cov}(X,X)=D(X)$，其中对角线上的元素 $D(x_i)$ 是随机变量 x_i 的方差，$i=1,\cdots,p$. 一般说来，$\mathrm{Cov}(X,Y)\neq\mathrm{Cov}(Y,X)$，因为

$$\mathrm{Cov}(X,Y)=E[X-E(X)][Y-E(Y)]'$$

$$=\{E[Y-E(Y)][X-E(X)]\}'$$

$$=\{\mathrm{Cov}(Y,X)\}'$$

对于 p 维随机向量 $X=(x_1,\cdots,x_p)'$，称

$$\rho_{ij}=\frac{\mathrm{Cov}(x_i,x_j)}{\sqrt{D(x_i)}\sqrt{D(x_j)}},\quad i,j=1,2,\cdots,p$$

为分量 x_i 与 x_j 之间的相关系数，称 $R=(\rho_{ij})_{p\times p}$ 为 X 的相关矩阵. 对于两个不同的随机向量 X 及 Y，它们之间的相关问题将在第十章典型相关分析中进一步讨论.

定义 1.1.3　特征函数. 随机向量X的特征函数定义为

$$\varphi(t)=E(e^{it'X}) \tag{1.1.4}$$

其中 $i=\sqrt{-1}$，t 是实数列向量，其维数与X的维数相同，即 $t=(t_1,\cdots,t_p)'$，则

$$\varphi(t)=E(e^{i\sum_{j=1}^{p}t_jx_j})$$

随机矩阵 $\boldsymbol{X}=(x_{\alpha\beta})_{p\times n}$ 的特征函数是用 pn 维随机向量 $(x_{11},\cdots,x_{p1},\cdots,x_{1n},\cdots,x_{pn})'$ 的特征函数来定义的，即为：

$$\varphi(T)=E(e^{i\sum_{\alpha=1}^{p}\sum_{\beta=1}^{n}t_{\alpha\beta}x_{\alpha\beta}})=E(e^{it_r(T'X)}) \tag{1.1.5}$$

其中T是与 $\boldsymbol{X}$ 的阶数相同的实数矩阵，t_r 是矩阵 $T'\boldsymbol{X}$ 的迹，即 $T'\boldsymbol{X}$ 的对角线上元素之和.

由上述定义及向量、矩阵有关运算的基本知识，可以证明下述等式成立. 设A、B、C为常数矩阵，$\boldsymbol{X}$、$\boldsymbol{Y}$ 为随机矩阵，X、Y、U、V 为随机向量，则有：

1° $E(A\boldsymbol{X}B+C)=AE(\boldsymbol{X})B+C$ (1.1.6)

2° $E(A\boldsymbol{X}+B\boldsymbol{Y})=AE(\boldsymbol{X})+BE(\boldsymbol{Y})$ (1.1.7)

3° $\mathrm{Cov}(X,Y)=E[XY']-[E(X)][E(Y)]'$ (1.1.8)

4° $\mathrm{Cov}(AX,BY)=A\mathrm{Cov}(X,Y)B'$ (1.1.9)

5° $\mathrm{Cov}(aX+bY,dU+eV)=ad\mathrm{Cov}(X,U)$
$+ae\mathrm{Cov}(X,V)+bd\mathrm{Cov}(Y,U)+be\mathrm{Cov}(Y,V)$

(1.1.10)

其中 a、b、d、e 皆为实常数.

6° $D(BX+K)=BD(X)B'$ (1.1.11)

其中K为实常数向量.

这里假定上述各式的运算总可进行(如协方差阵的存在及阶数、维数协调一致等). 详细证明留作练习.

p 维随机向量 $X=(x_1,\cdots,x_p)'$ 的协方差阵及相关矩阵都是非负定矩阵. 事实上

$$\mathrm{Cov}(X,X)=(\mathrm{Cov}(x_i,x_j))_{p\times p}$$

记

$$\mathrm{Cov}(x_i,x_j)=\sigma_{ij},\quad i,j=1,\cdots,p$$

则

$$\sigma_{ij}=\mathrm{Cov}(x_i,x_j)=\mathrm{Cov}(x_j,x_i)=\sigma_{ji}$$

可见协方差阵是对称的. 又对任意 p 维实向量 $C=(c_1,\cdots,\ c_p)'$, 则有

$$C'\mathrm{Cov}(X,X)C=\sum_{i=1}^{p}\sum_{j=1}^{p}c_ic_jE(x_i-Ex_i)(x_j-Ex_j)$$

$$=E\left[\sum_{i=1}^{p}c_i(x_i-Ex_i)\right]^2\geqslant 0 \tag{1.1.12}$$

对任意 c 都有 ≥0 ……

可见, 协方差阵是非负定阵(常用 $\mathrm{Cov}(X,X)\geqslant\square$ 表示非负定阵, $\square$表示元素为零的矩阵). 由此可知, 相关矩阵 R 也是非负定阵.

定义 1.1.4　条件分布、独立性. 将 p 维随机向量 $X=(x_1,\cdots,x_p)'$ 分成两个子向量: $1\leqslant q\leqslant p$,

$$X=\begin{pmatrix}X_{(1)}\\ \cdots\cdots\\ X_{(2)}\end{pmatrix},\quad X_{(1)}=\begin{pmatrix}x_1\\ \vdots\\ x_q\end{pmatrix},\quad X_{(2)}=\begin{pmatrix}x_{q+1}\\ \vdots\\ x_p\end{pmatrix} \tag{1.1.13}$$

在此 $X_{(1)}$、$X_{(2)}$ 分别表示 X 的 两个子 向量, 也为随 机向量. 记 $F(x_1,\cdots,x_p)$ 为 X 的联合分布函数, $F_1(x_1,\cdots,x_q)$, $F_2(x_{q+1},\cdots,x_p)$ 分别为 X 的边沿分布函数. 随机向量 $X_{(1)}$ 与 $X_{(2)}$ 相互独立是指

$$F(x_1,\cdots,x_p)=F_1(x_1,\cdots,x_q)F_2(x_{q+1},\cdots,x_p) \tag{1.1.14}$$

若记 $\varphi(t)$ 为 X 的特征函数, $\varphi_1(t_{(1)})$ 及 $\varphi_2(t_{(2)})$ 分别表示 $X_{(1)}$ 及 $X_{(2)}$ 的特征函数, 其中

$$t=\begin{pmatrix}t_{(1)}\\ \cdots\\ t_{(2)}\end{pmatrix},\quad t_{(1)}=\begin{pmatrix}t_1\\ \vdots\\ t_q\end{pmatrix},\quad t_{(2)}=\begin{pmatrix}t_{q+1}\\ \vdots\\ t_p\end{pmatrix}$$

则 $X_{(1)}$ 与 $X_{(2)}$ 相互独立等价于

$$\varphi(t)=\varphi_1(t_{(1)})\varphi_2(t_{(2)}) \tag{1.1.15}$$

条件分布通常是分别就离散型、连续型给出定义. 现在就连续型情形讨论条件分布问题，对于离散型情形可类似地给出有关结论. 设X具有分布密度函数 $f(x_1,\cdots,x_p)$,则 $X_{(1)}$ 及 $X_{(2)}$ 具有各自分布的密度函数,分别记作 $f_1(x_1,\cdots,x_q)$及 $f_2(x_{q+1},\cdots,x_p)$. 由(1.1.14)式知,$X_{(1)}$ 与 $X_{(2)}$ 相互独立又等价于

$$f(x_1,\cdots,x_p)=f_1(x_1,\cdots,x_q)f_2(x_{q+1},\cdots,x_p) \tag{1.1.16}$$

若(1.1.16)式不成立，则 $X_{(1)}$ 及 $X_{(2)}$ 中一方的概率分布必与另一方有关. 当给定 $X_{(2)}=x_{(2)}$ 的数值向量时，若 $f_2(x_{q+1},\cdots,x_p)\neq 0$,则 $X_{(1)}$ 具有条件分布密度函数 $f_1(x_1,\cdots,x_q|x_{q+1},\cdots,x_p)$，可以证明

$$f_1(x_1,\cdots,x_q|x_{q+1},\cdots,x_p)=\frac{f(x_1,\cdots,x_p)}{f_2(x_{q+1},\cdots,x_p)} \tag{1.1.17}$$

同理,在给定 $X_{(1)}=x_{(1)}$ 的数值向量时,若 $f_1(x_1,\cdots,x_q)\neq 0$,则$X_{(2)}$ 具有条件分布密度函数 $f_2(x_{q+1},\cdots,x_p|x_1,\cdots,x_q)$,可以证明

$$f_2(x_{q+1},\cdots,x_p|x_1,\cdots,x_q)=\frac{f(x_1,\cdots,x_p)}{f_1(x_1,\cdots,x_q)} \tag{1.1.18}$$

对照(1.1.16)、(1.1.17)、(1.1.18)三式可见,$X_{(1)}$ 与 $X_{(2)}$ 相互独立等价于

$$f_2(x_{q+1},\cdots,x_p|x_1,\cdots,x_q)=f_2(x_{q+1},\cdots,x_p)$$
$$f_1(x_1,\cdots,x_q|x_{q+1},\cdots,x_p)=f_1(x_1,\cdots,x_q)$$

§1.2 多元正态分布

多元正态分布是一元正态分布及二元正态分布的自然推广，当然内容更为丰富. 我们知道,若 y 服从$N(0,1)$，记 $x=\mu+\sigma y$,其中 μ 及 $\sigma>0$ 为实数,则 x 服从 $N(\mu,\sigma^2)$. 就是说,可用标准正