

医用多元统计分析方法

陈 峰 编著 陈启光 审阅



中国统计出版社
China Statistics Press

医用多元统计分析方法

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/陈峰著.
—北京:中国统计出版社,2000.12
ISBN 7-5037-3417-5/R.1

I. 医…
II. 陈…
III. 医学统计-多元分析:统计分析
IV. R195.1

中国版本图书馆 CIP 数据核字(2000)第 79930 号

责任编辑/吕 军
封面设计/张建民
责任校对/刘开颜
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 75 号 邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电 话/(010)63459084,63266600-22500(发行部)
印 刷/北京兴华印刷厂印刷
经 销/新华书店
开 本/787×1092 mm 1/16
字 数/450 千字
印 张/19.25
印 数/1-3000 册
版 别/2000 年 12 月第 1 版
版 次/2001 年 4 月第 1 次印刷
定 价/35.00 元

中国统计版图书,版权所有,侵权必究。

中国统计版图书,如有印装错误,本社发行部负责调换。

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

2 医用多元统计分析方法·序

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的应用的精妙,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 回归等,以及建模策略。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。习题参考答案是荀鹏程、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

序 前言

1 多元正态分布	(1)
1.1 多元分析常用统计量	(1)
1.1.1 均向量	(2)
1.1.2 方差、协方差矩阵	(2)
1.1.3 离均差平方和与离均差积和矩阵	(2)
1.1.4 相关系数矩阵	(3)
1.2 多元正态分布	(3)
1.2.1 定义	(3)
1.2.2 性质	(4)
1.3 二元正态相关变量的参考值范围	(6)
2 均向量的统计推断	(9)
2.1 多元 T 检验	(9)
2.1.1 多元配对设计的均向量检验	(9)
2.1.2 多元成组设计两样本的均向量检验	(11)
2.2 多元方差分析	(12)
2.2.1 多元成组设计资料的分析	(12)
2.2.2 多元区组设计资料的分析	(15)
2.2.3 多元析因设计资料的分析	(16)
2.3 协方差阵的检验	(19)
2.3.1 $V=V_0$ 的检验	(19)
2.3.2 $V=\sigma^2V_0$ 的检验	(20)
2.3.3 $V_1=V_2=\dots=V_g$ 的检验	(20)
2.4 多元方差分析的正确应用	(21)
3 多元线性回归	(22)
3.1 多元线性回归模型简介	(22)

3.2	回归系数的估计	(23)
3.2.1	矩阵算法	(24)
3.2.2	消去变换法	(25)
3.3	方程的假设检验	(27)
3.3.1	y 方面变异的分解	(27)
3.3.2	回归方程的方差分析	(28)
3.4	决定系数与剩余标准差	(28)
3.5	偏回归系数的假设检验与区间估计	(29)
3.6	标准偏回归系数与自变量的贡献	(30)
3.6.1	标准偏回归系数	(30)
3.6.2	自变量作用的分解	(30)
3.6.3	复相关系数的分解	(31)
3.7	因变量的区间估计	(31)
3.7.1	y 的可信区间估计	(32)
3.7.2	y 的容许区间估计	(32)
3.8	指标的量化	(32)
3.9	衡量回归方程的标准	(33)
3.9.1	复相关系数 R	(33)
3.9.2	校正复相关系数 R_{adj}	(34)
3.9.3	剩余标准差 $S_g \cdot x_1 x_2 \cdots x_p$	(34)
3.9.4	赤池信息准则	(34)
3.9.5	C_p 统计量 (C_p statistic)	(35)
3.10	逐步回归	(36)
3.11	回归系数反常的原因	(44)
3.12	岭回归	(46)
3.13	回归分析的正确应用	(48)
4	主成分分析	(50)
4.1	主成分的定义	(50)
4.2	主成分的计算	(52)
4.3	主成分的性质	(54)
4.4	主成分的应用	(56)
4.4.1	主成分评价	(56)
4.4.2	主成分回归	(60)
4.5	有关的统计推断	(61)
4.5.1	特征根的可信区间估计	(62)
4.5.2	等相关性检验	(62)
4.5.3	主成分相等的检验	(63)

4.6	主成分分析的正确应用	(63)
5	因子分析	(65)
5.1	因子模型	(65)
5.2	因子模型的估计	(68)
5.2.1	主成分法	(68)
5.2.2	极大似然法	(69)
5.2.3	主因子法	(70)
5.2.4	迭代主因子法	(72)
5.2.5	残差矩阵	(72)
5.3	因子旋转	(74)
5.3.1	方差最大正交旋转	(74)
5.3.2	斜交旋转	(75)
5.4	因子得分	(77)
5.5	因子分析的策略	(78)
5.6	因子分析的正确应用	(81)
6	logistic 族回归	(83)
6.1	多元 logistic 回归模型	(83)
6.1.1	多元 logistic 回归模型	(83)
6.1.2	系数的解释	(84)
6.1.3	变量的假设检验	(87)
6.1.4	建模策略	(89)
6.1.5	四格表资料的 logistic 回归	(95)
6.2	配比设计的条件 logistic 回归	(96)
6.2.1	条件 logistic 回归模型	(96)
6.2.2	配对四格表资料的条件 logistic 回归	(99)
6.3	多类结果变量的 logistic 回归	(101)
6.3.1	多类结果变量的 logistic 回归模型	(101)
6.3.2	系数的解释与检验	(102)
6.3.3	建模策略	(104)
6.4	有序结果的累积比数 logistic 回归	(105)
6.4.1	累积比数 logistic 回归模型	(105)
6.4.2	累积比数模型的应用条件	(107)
6.5	有序结果的相邻比数 logistic 回归模型	(109)
6.5.1	相邻比数 logistic 回归模型	(110)
6.6	logistic 族回归模型的正确应用	(111)
7	广义线性模型	(114)
7.1	线性模型与广义线性模型	(114)

7.1.1	线性模型	(114)
7.1.2	广义线性模型	(115)
7.1.3	指数分布族	(116)
7.1.4	联接函数	(117)
7.2	广义线性模型的建立	(118)
7.2.1	GLM 的参数估计	(118)
7.2.2	GLM 的假设检验	(120)
7.2.3	拟合优度	(121)
7.2.4	残差分析	(122)
7.3	logistic 回归与 Probit 回归	(123)
7.4	Poisson 回归	(125)
7.5	负二项回归	(128)
7.6	广义线性模型的正确应用	(130)
8	生存分析	(132)
8.1	随访研究的特点	(132)
8.1.1	截尾数据	(133)
8.1.2	几个基本概念	(134)
8.1.3	随访资料的特点	(134)
8.2	生存分析的理论体系与常用指标	(136)
8.3	指数模型	(139)
8.3.1	指数分布模型	(139)
8.3.2	指数分布模型的参数估计	(140)
8.3.3	两个指数分布模型的比较	(140)
8.3.4	指数回归模型	(142)
8.4	Weibull 模型	(146)
8.4.1	Weibull 分布模型	(146)
8.4.2	Weibull 分布模型的参数估计	(147)
8.4.3	Weibull 回归模型	(148)
8.4.4	Weibull 回归与指数回归的比较	(149)
8.5	Cox 比例风险模型	(155)
8.6	生存分析的正确应用	(157)
9	聚类分析	(160)
9.1	聚类的目的与方法	(160)
9.2	距离和相似系数	(161)
9.2.1	距离	(161)
9.2.2	相似系数	(163)
9.2.3	列联系数	(164)
9.3	系统聚类法	(165)

9.3.1	最短距离法	(166)
9.3.2	最长距离法	(168)
9.3.3	中间距离法	(170)
9.3.4	可变距离法	(170)
9.3.5	重心法	(170)
9.3.6	类平均法	(171)
9.3.7	可变类平均法	(171)
9.3.8	Ward 最小方差法	(171)
9.3.9	八种系统聚类方法的统一	(172)
9.4	动态聚类	(176)
9.5	有序样品的聚类	(180)
9.6	条件系统聚类	(188)
9.7	聚类分析的正确应用	(191)
10	判别分析	(193)
10.1	距离判别	(193)
10.1.1	两个总体的距离判别	(193)
10.1.2	多个总体的距离判别	(200)
10.2	Bayes 判别	(203)
10.3	Fisher 判别	(205)
10.4	逐步判别	(208)
10.4.1	基本思想	(208)
10.4.2	计算步骤	(208)
10.5	Bayes 公式法和极大似然法	(216)
10.5.1	Bayes 公式法	(217)
10.5.2	极大似然法及其简便算法	(220)
10.6	判别分析的正确应用	(221)
附录 A	极大似然方法	(224)
附录 B	习题	(232)
附录 C	习题参考答案	(251)
附录 D	部分例题 SAS 程序	(274)
参考文献		(293)

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),即原始资料对正态的偏离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:

$$\bar{X} = \begin{pmatrix} 161.8667 \\ 48.0833 \\ 74.3750 \end{pmatrix}$$

有时为方便印刷,均向量不用列向量表示而用行向量表示:

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

\bar{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 V 表示。其中:

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

本例:

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

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,

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

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(sum of squares and cross-products matrix, SSCP),简称离差阵。离差阵为对称矩阵。离差阵用字母 SS 表示(有时用 L 表示)。其中:

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

SS 与 V 有如下关系:

$$SS = (n-1)V$$

请自行验证。离差阵也是对称矩阵。本例:

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

或:

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

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 R 表示。

本例:

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

或:

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

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:

$$\begin{pmatrix} 502.9464 & 553.9831 & 354.5498 \\ 0.8926 & 765.9168 & 500.1249 \\ 0.8020 & 0.9168 & 388.5629 \end{pmatrix}$$

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:

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

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 $X = (x_1, x_2, \dots, x_m)'$ 服从 m 维正态分布, 则:

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

其中, Σ 是变量 x_1, x_2, \dots, x_m 的协方差阵。 $|\Sigma|$ 是 $|\Sigma|$ 的行列式。

4 医用多元统计分析方法

例 1.2 (二元正态分布) 设变量 x_1, x_2 的均数分别为 μ_1, μ_2 , 方差分别为 σ_1^2, σ_2^2 (为方便起见, σ_1^2, σ_2^2 分别表示为 σ_{11} 和 σ_{22}), x_1 与 x_2 的协方差为 σ_{12} ($\sigma_{12} = \sigma_{21}$), x_1 与 x_2 的相关系数为 ρ 。用矩阵表示, x_1, x_2 的协方差阵为:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

利用矩阵知识得协方差阵的逆矩阵为:

$$\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{11} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{22} \end{pmatrix}$$

行列式为:

$$|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho^2)$$

二元正态分布的密度函数为:

$$f(x_1, x_2) = \frac{1}{2\pi \sqrt{\sigma_{11}\sigma_{22}(1 - \rho^2)}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\} \quad (1.4)$$

二元正态分布中有 5 个参数, 即 x_1 的均数 μ_1 , 及方差 σ_1^2 ; x_2 的均数 μ_2 , 及方差 σ_2^2 ; 以及 x_1 与 x_2 的相关系数 ρ 。

如果随机向量 $\mathbf{X} = (x_1, x_2)$ 服从二元正态分布, 则记为 $\mathbf{X} \sim MN(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2; \rho)$ 。图 1.1~1.3 示意了均向量为 0, 不同方差及相关时二元正态分布的三维分布曲面。其中 z 表示 $f(x_1, x_2)$ 。

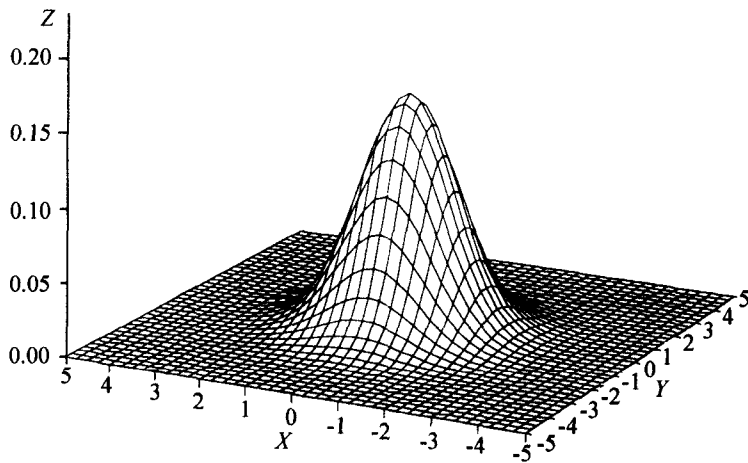
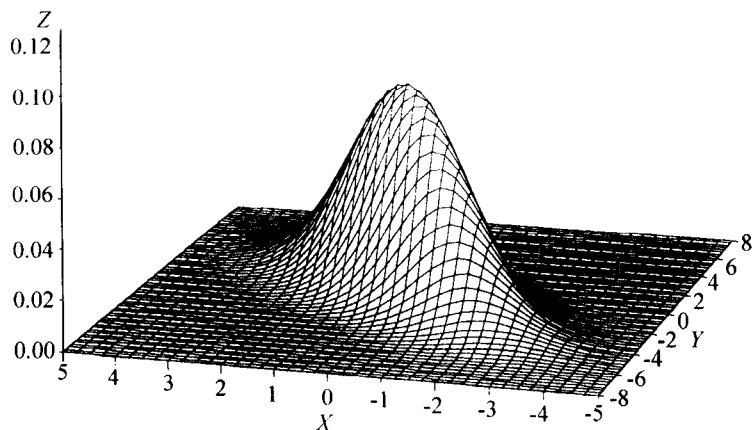
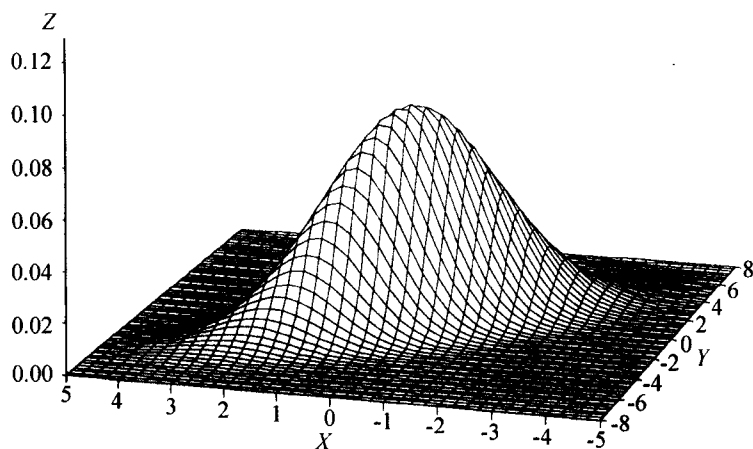


图 1.1 二元正态分布曲面 ($\sigma_{11} = 1, \sigma_{22} = 1, \rho_{12} = 0$)

1.2.2 性质

m 元正态分布具有如下几个性质:

(1) 每一个变量均服从正态分布。多元正态分布中, 每个变量的分布均服从正态分布, 见图 1.1~1.3。但反之不真。

图 1.2 二元正态分布曲面($\sigma_{11}=2, \sigma_{22}=4, \rho_{12}=0.75$)图 1.3 二元正态分布曲面($\sigma_{11}=2, \sigma_{22}=4, \rho_{12}=-0.75$)

(2) 变量的线性组合服从正态分布。显然,这一性质是性质(1)的推广。

(3) m 元正态分布中的任意 k ($0 < k < m$) 个变量服从 k 元正态分布。性质(1)是其特殊情况。

(4) m 元正态分布的条件分布仍服从正态分布。即在某些变量取值固定时,另外一些变量的分布服从(多元)正态分布。图 1.4 表示了二元正态分布 $MN(0,1;0,0.5;-0.75)$ 中 $x=1$ 时 y 的分布,和 $y=-6$ 时, x 的分布。

(5) 协方差为 0 的变量间相互独立。对多元正态分布而言,不相关(zero-correlation)与独立(independent)是等价的。

在讨论多元正态分布的理论和方法时,经常用到这些性质。

多元分析中的许多统计方法大都假设资料来自正态分布总体。但要判断已有的一批数据是否来自正态分布总体,并不是一件容易的事。可是,反过来要判断数据不是来自正态分布总体,倒有一些简单方法。比如,性质(1)表明,正态分布中的每个变量均服从正态分布,据此,我们对每个变量检验其是否服从正态分布。如果有一个变量不服从正态分布,则可判定资料也可能不服从多元正态分布。

有关多元正态分布的深入讨论见张尧庭,方开泰(1982)的著作。