实用多元统计分析

王学仁　王松桂　编译

上海科学技术出版社

实用多元统计分析

王学仁　王松桂　编译

上海科学技术出版社

实用多元统计分析

王学仁　王松桂　编译

上海科学技术出版社出版

（上海瑞金二路 450 号）

新华书店上海发行所发行　上海东方印刷厂印刷

开本 850×1156 1/32　印张 14　字数 365,000

1990 年 9 月第 1 版　1990 年 9 月第 1 次印刷

印数：1—3,400

ISBN 7-5323—0373-X/O·27

定价：11.95 元

内 容 提 要

本书系统地叙述了多元统计分析的基本理论和方法，其重点在于实用方面。本书特点是概念清晰、方法明了、取材较新。主要内容包括：多元随机样本；多元正态分布；均值向量的推断；以及多元分析中的各重要应用分支，如回归分析、主成份分析、因子分析、判别分析和聚类分析等。各章末尾附有练习题和参考文献。练习题中有一部分是实际例子，需要借助电子计算机来完成。

本书的读者对象是应用数理统计工作者，一般科技人员，理、工、农、医、经济、管理等有关专业的师生。本书也可作为应用多元分析课程的教材或教学参考书。

编译者的话

随着电子计算机技术的普及，多元分析在国民经济和科学研究的很多部门都得到了日益广泛的应用．特别是近年来许多理、工、农、医、经济、管理等院校相继开设了多元分析课程或有关的统计课程，因此，很需要一本所涉及的数学知识不很深，且概念清晰、方法明了、取材较新的多元分析方法方面的专著．为适应这一需要，我们参考了 R. A. Johnson 和 D. W. Wichern 所著的《实用多元统计分析》一书及其他一些资料编译了本书．

本书的着眼点在于实用，在理论和应用两个方面做到了较为恰当的平衡．作者十分注意阐明概念的统计背景，在一般工科院校一、二年级数学水平的基础上，对多元分析方法作了清楚的阐述，在不涉及过深数学知识的地方，对一些理论也给予了必要的证明．全书收集了大量实际例子，以展示各种方法应用的广度和深度．为使本书能以更精炼的形式奉献给读者，我们除重新编写了部分内容之外，还对一些纯数字例子、练习题和附录做了筛选．鉴于矩阵理论已属一般高校（甚至各种业余大学）基础课程的内容，所以，编译时删去了原书第二章有关矩阵的一些初等内容．为方便读者，我们还特别增加了一些中文参考文献．

本书的读者对象主要是应用数理统计工作者、广大科技人员以及理、工、农、医、经济和管理等有关专业师生，当然也可作为高等学校有关课程的教材或参考书．对于主要从事理论研究的同志，本书对加深统计思想、拓广统计背景也有所裨益．

陈希孺教授对本书的编译工作给予了热情的关心和支持，项可风同志仔细校阅了全文，提出了许多宝贵意见，这对提高本书的编译质量有很大帮助，编译者谨致衷心的感谢．

本书第二、三、六和十章由王学仁同志编译，第一、四、五、七、八、九章由王松桂同志编译．

在编译过程中，我们对原书一些错误作了订正，但限于水平，书中仍会有错误及不妥之处，恳请广大读者不吝赐教.

编译者　1984. 12. 15

目　录

第一章 绪 论

§1.1 引 言

在工业、农业、经济、生物和医学等领域的实际问题中，常常需要处理多个变量的观测数据. 如果用一元统计方法，则势必要把多个变量分开分析，一次处理一个变量. 由于这种方法忽视了诸变量之间可能存在着的相关性，因此，一般说来，丢失信息太多. 另一种方法就是本书要讨论的多元分析方法，它同时对多个变量的观测数据进行分析. 这样的分析对诸变量之间的关系、相依性和相对重要性等都能提供有用信息.

由于多元分析研究的是多个变量的统计总体，这就使得它成为一个较难处理的课题. 一方面是数据浩繁，人们往往会淹没在数据堆中；另一方面，导出推断方法所需要的数学工具远比单变量情形时要多. 本书试图给出以代数概念为基础的统计解释，而避免使用较艰深的数学推导. 我们的目标是，尽量应用解释性例子和较少的数学，以较清晰的方式介绍一些很有用的多元统计方法，当然，在阅读本书的时候，数学上的理解和定量的思维仍然是需要的.

在某种意义上讲，多元分析是一些方法的"混合体". 我们难于对其所有方法做一种分类，使它能够恰如其分地反映各种方法的适用性. 有一种分类法，它是依其所研究的关系性质将各种方法进行分类. 另一种则是根据所研究的总体和变量个数来分类. 虽然本书的章次是按照均值的推断(四、五、六章)、协方差阵结构的推断(七、八章)和分类法(九、十章)安排的，然而，这并不意味着我们试图把每一种方法固定在特定的某一类中，相反，研究中所采用的分析类型和具体方法主要取决于研究目的. 下面我们列

举一些实际问题，用以解释选择统计方法和研究目的之间的关系．这些问题以及课文中的大量例子能够使读者对多元统计方法在各个领域中的广泛应用有一定的了解．

多元分析适用于下列目的的研究：

1．数据或结构性化简．尽可能简单地表示所研究的现象，但不致损失很多有价值的信息，并希望这种表示能够很容易地进行解释．

2．分类和组合．基于所测量到的一些特征，对相似的对象或变量分组，同时给出好的分组法则．

3．研究变量之间的依赖关系．变量之间的关系往往是我们所感兴趣的．究竟所有变量都是相互独立的，还是一个或几个变量依赖于另外一些变量？如果是后者，那么这种依赖关系又是怎样的？

4．预测．变量间关系的建立是为了从一些变量的观测值预测另外一个或一些变量的值．

5．假设的提出及检验．检验由多元总体参数表示的某种统计假设．据此，能够证实某些假设条件的合理性或支持原有的某种信念．

最后，我们援引 F. H. C. Marriott[1] 中第89页上的一段话来结束本段讨论．虽然他的这些话是针对聚类分析而言的，但是我们认为，它也适用于更广的一类方法．每当人们从事数据分析时，记起这些话是大有裨益的．它可以提醒人们对问题保持适度的看法，而不被一些理论的严谨和优美所迷惑．

"如果所得到的结论与一种合乎情理的看法不一致，劝君莫轻易接受一个简单的逻辑性解释，也不要画图表示它，因为这些结论可能都是错误的．绝对不存在数值方法的一种魔术，实际上有多种途径会导致这些方法的失效．数值方法的价值仅仅在于帮助人们去解释数据，它绝不是香肠灌装机，能够把一堆数据自动地变换到科学事实的锦囊之中．"

§1.2 多元分析的应用

统计方法是科学研究的一种重要工具，其应用颇为广泛．特别地，多元分析方法常常被应用于自然科学、社会科学和医学等领域的问题中．现在我们列举一些这样的实际问题．实践表明，多元分析方法对处理这些问题有一定的价值．另一方面，这些问题本身也显示了多元分析应用的广泛性．

医学

1．在一项癌症患者对放射疗法反应的研究中，对一批患者测量六项反应指标．因为难于同时解释所有变量的观测值，于是人们需要寻找刻画患者反应的一个简单综合性指标．多元分析能够帮助研究者达到这一目的，且不损失数据中蕴含的很多信息．（这里的目的是数据化简．）

2．人对视觉刺激（如闪光）的反应能够从头皮上应用电子计算机记录下来．在视觉脑电计算机分析（VEDA）中，这些反应被称为治疗者形象．在一项关于视觉系统多重硬化效应研究中，多元分析被用来考察视觉脑电计算机分析对于诊断视觉疾病是否为一种既实用又可靠的手段．（这里的目的是判别或称分类，即建立数值判别法则，把由视觉疾病引起多重硬化病的人与没有这种病的人区分开来．）

商业和经济

3．测量六个金融变量并建立一个多元模型，以帮助保险事务管理者判定可能破产的财产—债务承保者．应用这个模型，保险公司可分为有偿付能力的和"预后不良"的两类，而后采取相应的补救措施，以防止后者的破产．（此处的目的是，建立一个判别法则，以区分有偿付能力的和"预后不良"的两类保险公司．）

教育学

4. 多元分析常常用于体育运动项目的研究. 例如,对田径运动成绩的分析有助于确定各种运动的基本功. Linden[12] 于1977年对八届奥林匹克运动会十项全能成绩用多元分析方法确定了四个基本体力因子:短跑速度、臂力、长跑耐力、腿力. (在这里,目的是确定观测变量(田径运动成绩)对少数潜在变量(体力因子)的依赖关系.)

生物学

5. 在植物育种研究中,当植物收获时,我们要选择种子,以保证在一些特征上,下一代优于上一代. 这时往往要测量、评价很多特征. 育种者的目标是在最短的时间内,使遗传基因达到最大. 在一项蚕豆育种研究中[10],应用多元分析把几个与产量和蛋白质含量有关的变量变换为一个"选择指标",根据这个指标的得分来选择种子. (此例的目的也是数据化简,即构造一个综合指标代替原来的变量,并建立判别法则.)

6. 有两种被认为是难以区分的繁缕. 对每株繁缕测量四个形状变量. 应用多元分析可以构造这四个变量的一个函数,用它来区分两种繁缕[8],即对新拿来的一株繁缕,仅知它是两种繁缕中之一种,用这个函数就可判定它是那一种. (显然,这里的目的是判别.)

环境保护

7. 许多学者研究了洛杉矶地区大气中污染物质的浓度. 在某项研究中(见练习1.5),在较长的一段时间内,每天测量与污染有关的七个变量. 我们的兴趣首先是,空气污染的程度在一周内是否固定不变,或周末与平日是否有显著差异;其次,就是要知道这庞杂的测量数据能否用一种易解释的方式加以归纳化简. (此例是假设检验问题和数据化简.)

气象学

8. 研究人员已着手进行一项研究工作(参见[9]), 即把树木年轮与各种气候参数之间的关系定量化. 研究者感兴趣的是, 首先确定年轮所包含的气候信息的类型, 然后找出公元1700年以来的气候异常现象. 这里, 应用多元分析方法可以把多得惊人的数据化简到人们能够处理的范围之内. 在这个过程中, 重新定义了少数几个新变量, 并对它们做较为简单的统计分析和解释. (此例的目的是数据化简.)

地质学

9. 在一项关于沉积物体积等级分布的研究中(见[7]、[13]), 多元分析被用于构造10个变量的两个线性函数, 以便地质学家区分五种沉积环境, 所获得的结果使人们大大减少了为区别不同沉积类型所必须做的一些试验室工作. (此例的目的仍然是数据化简和判别)

心理学

10. 在一项冒险行为的研究中[15], 随机地分配每个学生去接受三种不同训练或"处理"中的一种. 而后, 用同一测验的两种形式去检查他们, 对错误反应给予高、低两种处罚分数. 这样的测验得分与带有不同风险的训练方案有关. 问题是, 在用测验得分度量的可觉察的风险方面, 不同的训练是否确有差别. (此例是假设检验问题.)

在浏览了多元分析在各个领域的应用之后, 读者能够看到, 虽然具体问题的内容各有不同, 但从数据分析角度看却有许多是相同的或相似的. 因此, 和大多数统计方法一样, 多元分析的应用具有相当的广泛性.

§1.3 描述统计量

庞杂的大型数据使得人们难于直观地从中提取有用信息. 然而,数据中蕴含的许多有用信息能够用一些数字来描述,通常称这些数为*描述统计量*. 例如,算术平均值或样本均值就是给出位置度量的描述统计量,它刻画了一组数据的"中心值". 又如,每个数到它们均值的距离平方的平均值给出了该组数散布程度(或变差)的度量.

度量位置、变差和线性相关程度的描述统计量是十分重要的. 下面给出它们的定义.

设 $x_{11}, x_{12}, \cdots, x_{1n}$ 为第一个变量的 n 个观测值. 它们的算术平均值用 $\bar{x}_1$ 表示,其定义为 $\bar{x}_1=\sum_{j=1}^{n} x_{1j}/n$. 如果这 n 个观测值是所有可能的观测值的一部分,则也称 $\bar{x}_1$ 为第一个变量的样本均值. 我们之所以采用这个术语,是因为本书的大部分内容旨在建立分析这种测量样本的方法.

对每个变量的 n 个测量值,都可以计算样本均值. 于是,如果我们有 p 个变量,就得到 p 个样本均值

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad i=1, 2, \cdots, p. \tag{1.1}$$

样本方差给出了数据散布程度的度量. 对第一个变量而言,它定义为 $s_1^2=\frac{1}{n}\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)^2$,这里 $\bar{x}_1$ 为其样本均值. 一般地,对 p 个变量,我们有

$$s_i^2=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2, \quad i=1, 2, \cdots, p. \tag{1.2}$$

这里有两点值得注意. 首先,在(1.2)的定义中,许多作者不用 n 而用 $(n-1)$ 去除和式. 后面我们将会看到用 n 的理论根据,并且当 n 较小时,(1.2)是特别适当的. 采用适当的记号,样本方差的这两种定义总是可以区分开的.

其次，虽然通常用 s^2 表示样本方差，但是，考虑到它在样本协方差阵中将排在对角线上，于是，为方便起见，我们采用双下标 s_{ii} 来代替 s_i^2，即

$$s_i^2=s_{ii}=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2,\quad i=1,2,\cdots,p. \tag{1.3}$$

样本方差的平方根 $\sqrt{s_{ii}}$ 称为样本标准离差，它与观测值具有相同的度量单位.

考虑前两个变量的 n 对观测值

$$\begin{pmatrix}x_{11}\\x_{21}\end{pmatrix},\ \begin{pmatrix}x_{12}\\x_{22}\end{pmatrix},\cdots,\ \begin{pmatrix}x_{1n}\\x_{2n}\end{pmatrix};$$

即 x_{1j} 和 x_{2j} 是在第 j 个试验单元上的观测值 $(j=1,2,\cdots,n)$. 这两个变量的线性相关性的度量是样本协方差

$$s_{12}=\frac{1}{n}\sum_{j=1}^{n}(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2),$$

它是观测值对各自均值的偏差的乘积的平均. 如果两个变量或同时观测到较大的值，或同时观测到较小的值，则 s_{12} 为正. 相反，如果一个变量观测到大的值，而伴随着另一个观测到小的值，则 s_{12} 为负. 如果两个变量观测值之间没有特殊的关系，则 s_{12} 近似为 0.

一般地，样本协方差

$$s_{ik}=\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k),\quad i,k=1,2,\cdots,p \tag{1.4}$$

刻画了第 i 和第 k 两个变量的线性相关性. 注意，当 $i=k$ 时，样本协方差化为样本方差，而且，对所有 i, k, $s_{ik}=s_{ki}$.

我们要讨论的最后一个描述统计量是所谓样本相关系数（或称 Pearson 乘积矩相关系数[2]），两个变量之间的线性相关性的这种度量不依赖测量单位. 第 i 和第 k 两个变量的样本相关系数定义为

$$r_{ik}=\frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}=\frac{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{kj}-\bar{x}_k)^2}}, \tag{1.5}$$

其中 $i=1,2,\cdots,p$, $k=1,2,\cdots,p$. 注意，对所有 i, k, $r_{ik}=r_{ki}$.