多元统计分析方法

唐守正　编著

中 国 林 业 出 版 社

多元统计分析方法

唐守正　编著

中国林业出版社出版　(北京西城区刘海胡同 7 号)

新华书店北京发行所发行　　河北遵化印刷厂印刷

850×1168毫米32开本　　11印张　　254千字

1986年10月第1版　1986年10月遵化第1次印刷

印数　1—3,200　册

统一书号 16046·1266　　定价 2.55 元

序　言

多元统计分析或简称多元分析是本世纪初在数理统计的基础上逐渐发展起来的一门应用数学学科。从四十年代后期以来，由于数字电子计算机的出现和发展，使得多元统计分析方法在实际工作中的应用有了更为现实的意义，因而多元统计分析的应用范围日益广泛，同时，也促使多元统计分析在理论和方法上有了更为迅速的发展。目前在农业、林业及其它一些领域中多元分析方法已经成为分析实验数据的一种重要手段，在某些课题中甚至已经成为必要手段。由于这些原因，多元统计分析方法已越来越为广大农林工作者所重视。

我国于七十年代开始，逐渐推广多元分析方法的应用。但是系统地介绍多元分析理论或方法的书籍，直到七十年代后期才在国内出版，且为数不多。至于专门介绍多元分析方法在农林业中应用的书籍，则尚未有正式出版者。因此迫切需要编写和出版适用于农林院校学生、研究生及广大农林科学工作者的多元分析参考书。

本书是为农业林业的实际工作者编写的，重点在于介绍各种方法的基本思想和计算步骤。叙述力求简明，并结合了在农林科学研究和生产工作中的实例。为了加深对概念的理解，适当做了一些数学推导，但略去了复杂的证明。本书所使用的数学工具，一般未超出高等农林院校高等数学、线性代数和概率统计等课程的教学大纲范围。为便于对线性代数不太熟悉的读者学习，后附有线性代数的简明材料。

本书的写作曾得到黄中立教授的大力支持，符伍儒副教授详细阅读了初稿，提出了宝贵意见。徐冠华、张志鹏、郑世楷、王涛、王松龄等同志提供了部分实验数据。本书初稿曾在中国林业

科学研究院、北京林学院研究生班讲授过多次，根据反映删改了部分内容。对这些单位和个人的支持和鼓励在此一并致谢。

由于编者水平所限，书中有错误或不当之处，恳请读者批评指正。

编 著 者

1984年7月

目　录

第一章　多元总体和样本

§1.1　多元总体

1. 多元数据的表示方法

在农林业生产及科研中，往往同时观测n个对象的p个属性，然后再对这些数据进行整理分析，从而得出结论。多元统计分析是处理这类数据的一个有用工具。

为了讲述多元统计分析的基本原理和方法，本章先引入一些基本概念、公式和符号。

和一元统计学一样，把所研究的全部可能对象叫做一个总体，对象的每一个属性叫做一个变量。因此，如果同时研究一个总体的p个属性，就可以把这个总体看成一个p元变量。

从总体中随机抽取进行观测的对象叫做样本。一个样本单元的观测结果（p个数值）可以看做这个p元变量的一次取值，第i号样本单元的第α个属性的观测结果记为$x_{\alpha i}$。每个样本单元，例如第i号样本单元，可以得到p个观测值，用一个p维向量

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{bmatrix} \qquad (1.1.1)$$

来表示。因为p维向量是欧氏空间中的一个点，所以我们也把一个样本单元叫做一个(样本)点。由此，一个样本单元，一个p维向量$\mathbf{x}_i$，或p维空间中一个点，是同一个东西，分别为研究对象的实体，代数表示或几何表示。

对n个样本单元进行观测的全部结果，共有$p \times n$个数据。为

了方便，用一个矩阵

$$X=[x_1,x_2,\cdots,x_n]=\begin{bmatrix} x_{11} & x_{12} & \cdots x_{1n} \\ x_{21} & x_{22} & \cdots x_{2n} \\ \cdots\cdots & & \\ x_{p1} & x_{p2} & \cdots x_{pn} \end{bmatrix} \quad (1.1.2)$$

来表示，矩阵X是进行各种统计分析的基础资料，称为原始数据矩阵或（多元）样本数据。

2. 定量数据和定性数据

变量基本上可以分为二类：一类变量取值为实数，例如长度、百分含量等，这一类变量的观测值称为定量数据。另一类变量不是表示数量关系，而是表示样本是否具有某种性质。例如研究的对象是土壤，观测它的一个属性：是否灰化。这时可以这样记录：当样本具有这种属性时（例如有灰化）记为1，当样本没有这种属性时（例如无灰化）记为0，这里的1或0并不表示观测对象的数量关系，只是一种表示分类的方法。这类观测数据叫0、1型数据，或定性数据。

还有一种情况是变量的观测结果表示某种等级的编号，例如土壤中石砾含量分为无、少、中、多四级，自然可以用0，1,2，3分别表示这四个等级，这是介于定量数据与定性数据之间的一种数据。有时可以近似地把这类数据作定量数据处理，但更一般将它当成定性数据来处理。用定性数据来表示这类分级数字的方法如下：用一个向量来表示，它的每个分量分别对应一个等级，某个分量取值0表示不属于这个等级，取值1表示属于这个等级，每个分量都是0，1型数据。例如石砾含量少可以表示成（0，1，0，0）。今后将会经常用到这种表示定性数据的定性化方法。

定量数据也可以化为定性数据，只要将它取值的结果分成n个等级，然后用上述增加变量维数的办法即可化为0、1型数据。

3. 多元总体的概念

在上面已经提到，一个多元总体可以看成一个多元变量。这个多元变量在每个样本单元上取一个向量值。这种在不同样本单元上取不同向量值的多元变量，称为随机向量，或多元随机变量。多元随机变量的每一个分量，都是一个一元随机变量。简言之，多元总体就是一个多元随机变量。

综上所述，我们考查一个 p 元总体，就是考查这个总体中每个对象的 p 个属性。或者说考查一个 p 元随机变量。为此，从总体中随机抽取 n 个对象（样本）进行观测，得到 $p\times n$ 个观测数据 X。多元统计分析的主要任务，一是分析各观测数据之间的关系，二是推断总体的某些性质。

4. 随机向量的分布函数及密度

可以把随机变量分布函数的概念推广到随机向量。称 p 元函数：

$$F(x_1,\ x_2,\cdots,x_p)=P(\xi_1\leqslant x_1,\cdots,\ \xi_p\leqslant x_p) \qquad (1.1.3)$$

为 p 元随机变量 ξ 的分布函数。1.1.3 式也可以写成向量函数的形式：

$$F(x)=P(\xi\leqslant x) \qquad (1.1.4)$$

其中 $\xi\leqslant x$ 表示 ξ 的每个分量小于或等于 x 的相应分量。

同样，如果存在 p 元函数 $f(x_1,\ x_2,\cdots,x_p)$ 使

$$F(x_1,x_2,\cdots,x_p)=\int_{-\infty}^{x_p}\cdots\int_{-\infty}^{x_1}f(y_1,y_2,\cdots,y_p)dy_1dy_2\cdots dy_p \qquad (1.1.5)$$

或写成

$$F(x)=\int_{-\infty}^{x}f(y)dy$$

则称 $f(x_1\cdots x_p)$ 为随机向量 ξ 的分布密度函数，注意 1.1.5 是 p 重积分。

如果F(x) 是p次连续可微的，则：

$$f(\mathbf{x})=\frac{\partial^p}{\partial x_1\partial x_2\cdots\partial x_p}F(x_1,x_2,\cdots,x_p)$$

这些公式都与一元分布函数类似。以后我们总假定密度函数存在。

5. 总体平均向量（数学期望）

p元随机变量ξ的数学期望，即此总体的平均向量，定义为

$$E\xi=\begin{bmatrix}E\xi_1\\E\xi_2\\\vdots\\E\xi_n\end{bmatrix}\qquad(1.1.6)$$

其中 $E\xi_\alpha$ 为ξ的第α分量的数学期望，或

$$E\xi_\alpha=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}y_\alpha f(y_1,\cdots,y_p)dy_1\cdots dy_p\qquad(1.1.7)$$

6. 总体协方差矩阵

作为一元总体方差的推广，称下述p维矩阵Dξ为p维总体ξ的协方差矩阵：

$$D\xi=\Sigma=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\cdots\sigma_{1p}\\\sigma_{21}&\sigma_{22}&\cdots\sigma_{2p}\\\cdots\cdots&&\\\sigma_{p1}&\sigma_{p2}&\cdots\sigma_{pp}\end{bmatrix}\qquad(1.1.8)$$

其中对角线元素 $\sigma_{\alpha\alpha}$ 为ξ的第α分量 ξ_α 的方差：

$$\sigma_{\alpha\alpha}=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{\infty}(y_\alpha-E\xi_\alpha)^2f(y_1,\cdots,y_p)dy_1\cdots dy_p\quad(1.1.9)$$

非对角线元素 $\sigma_{\alpha\beta}(\alpha\neq\beta)$ 为ξ的第α与第β分量的协方差

$$\sigma_{\alpha\beta}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(y_\alpha-E\xi_\alpha)(y_\beta-E\xi_\beta)f(y_1,\cdots,y_p)dy_1\cdots dy_p$$

$$(1.1.10)$$

7. 平均向量与协方差矩阵的性质

在多元统计分析中，经常要对随机向量进行线性变换。所谓线性变换就是用一个新的随机向量 τ 代替原向量 ξ，使 τ 的每一分量是 ξ 的各分量的线性组合。用矩阵形式可将线性变换写成：

$$\tau = A\xi \qquad (1.1.11)$$

其中 A 一般是一个 $q\times p$ 矩阵，τ 是 q 之随机向量，矩阵 A 叫做线性变换的矩阵。

对于线性变换、平均向量与方差有下述性质：

i) $EA\xi = AE\xi \qquad (1.1.12)$

证明：令 $\tau = A\xi$，即 $\tau_\alpha = \sum_{k=1}^{p} a_{\alpha k}\xi_k$，$(\alpha = 1,2\cdots q)$

所以 $E\tau_\alpha = \sum_{k=1}^{p} a_{\alpha k}E\xi_k$，

$$E\tau = \begin{bmatrix} E\tau_1 \\ E\tau_2 \\ \vdots \\ E\tau_q \end{bmatrix} = \begin{pmatrix} \sum_{k=1}^{p} a_{1k}E\xi_k \\ \sum_{k=1}^{p} a_{2k}E\xi_k \\ \cdots\cdots \\ \sum_{k=1}^{p} a_{qk}E\xi_k \end{pmatrix} = AE\xi$$

ii) $DA\xi = A(D\xi)A' \qquad (1.1.13)$

证明：令 $\tau = A\xi$，于是 τ_α 和 τ_β 的协方差$\sigma_{\alpha\beta}^{(\tau)}$为

$$\sigma_{\alpha\beta}^{(\tau)} = \mathrm{cov}\left(\sum_{k=1}^{p} a_{\alpha k}\xi_k,\quad \sum_{j=1}^{p} a_{\beta j}\xi_j\right) = \sum_{k=1}^{p}\sum_{j=1}^{p} a_{\alpha k}a_{\beta j}\sigma_{kj}^{(\xi)}$$

于是

$$DA\xi = D(\tau) = [\sigma_{\alpha\beta}^{(\tau)}] = A[\sigma_{kj}^{(\xi)}]A' = A(D\xi)A'$$

其中$[\sigma_{\alpha\beta}^{(\tau)}]$或$[\sigma_{kj}^{(\xi)}]$表示由 $\sigma_{\alpha\beta}^{(\tau)}$ 或 $\sigma_{kj}^{(\xi)}$ 为元素的矩阵。以后为了书写简单，经常使用这种矩阵的记法。

iii）特别当 A 为 $1\times p$ 的矩阵，即行向量 a' 时，有

$$D(a'\xi)=a'(D\xi)a \qquad (1.1.14)$$

这是在一般概率论中经常出现的公式

$$D(a_1\xi_1+a_2\xi_2+\cdots+a_p\xi_p)=\sum_{j=1}^{p}\sum_{k=1}^{p}a_ja_k\sigma_{jk}$$

的矩阵形式。

8. 两个随机向量的协方差矩阵

若 ξ 是 p 维随机向量，τ 是 q 维随机向量，称 $p\times q$ 矩阵

$$\mathrm{cov}(\xi,\tau)=\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{pq} \\ C_{21} & C_{22} & \cdots & C_{2q} \\ \cdots\cdots & & & \\ C_{p1} & C_{p2} & \cdots & C_{pq}\end{bmatrix} \qquad (1.1.15)$$

为 ξ 和 τ 的协方差矩阵，其中 $C_{\alpha\beta}$ 为 ξ_α 与 τ_β 的协方差，即

$$C_{\alpha\beta}=\mathrm{cov}\ (\xi_\alpha,\ \tau_\beta) \qquad (1.1.16)$$

二个随机向量的协方差矩阵一般不是对称矩阵。

注意，若令 $p+q$ 维向量 $\zeta=\begin{bmatrix}\xi\\ \tau\end{bmatrix}$ 则：

$$D\zeta=\begin{bmatrix} D\xi & \mathrm{cov}(\xi,\tau) \\ \mathrm{cov}(\tau,\xi) & D(\tau)\end{bmatrix}$$

与 1.1.13 类似，对于线性变换，协方差公式为：

$$\mathrm{cov}(A\xi,\ B\tau)=A\mathrm{cov}(\xi,\ \tau)B' \qquad (1.1.17)$$

特别 $\mathrm{cov}(a'\xi,\ b'\xi)=a'D(\xi)b$ (1.1.18)

9. 多元正态分布

如果一个 p 元随机向量 ξ 的分布密度为：

$$f(x_1,\cdots,x_p)=\frac{1}{(2\pi)^{\frac{1}{2}p}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-a)'\Sigma^{-1}(x-a)\right] \qquad (1.1.19)$$

就称 ξ 服从 p 维正态分布，它的数学期望和协方差矩阵分别为：

$$E\xi = a$$

$$D\xi = \Sigma \qquad (1.1.20)$$

多元正态分布在多元统计中占有重要地位，在第六章中将详细讨论。

二维正态分布的图形大致如图 1.1 所示。图形最高点的坐标 (a_1, a_2) 为它的数学期望，它的每一个水平截面为一个椭圆形。

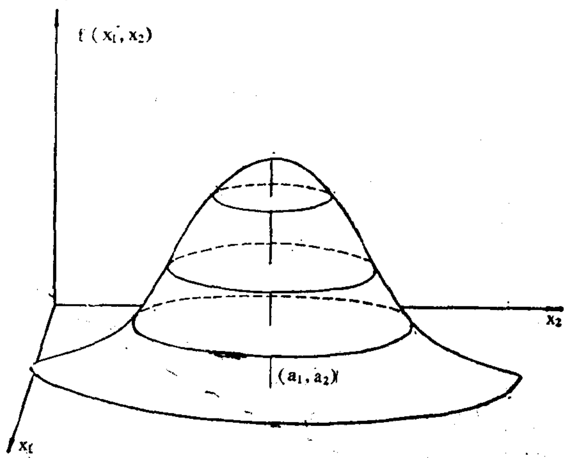

图 1.1　二元正态分布密度示意图

如果我们从二元正态总体中抽取足够多的 n 个点，则这些点

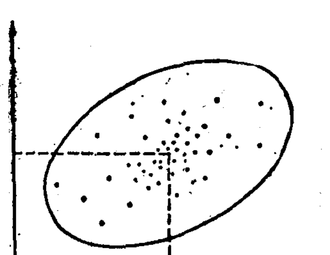

图 1.2　正态分布样本点分布示意图

大致对称地分布于一个椭圆内，越近椭圆中心，样本点分布越密；越向外则越稀疏。

可以在多维空间中类似地想象多维正态分布的样子。例如，p维正态分布的样本点大致对称地分布于一个p维椭球中，越近中心样本点的密度越大，见图1.2。

§1.2 多元样本

1. 概　　述

一般我们无法得到多元总体的各项特征数字，只能从我们所测定的样本出发。

假定在p元总体x中，抽取n个样本单元进行观测，得到多元样本数据

$$X=[x_1,x_2,\cdots,x_n]=\begin{bmatrix} x_{11} & x_{12}\cdots x_{1n} \\ x_{21} & x_{22}\cdots x_{2n} \\ \cdots\cdots & \\ x_{p1} & x_{p2}\cdots x_{pn} \end{bmatrix} \qquad (1.2.1)$$

既然p元总体x的每一个分量x_α是一个一元总体，这个一元总体在所抽取的n个样本单元上的取值，就是矩阵X中的第α行。也就是说多元样本数据中的每一行是一组一元样本，因此可以定义相应的特征数字，例如样本平均数，样本标准差，协方差，相关系数等。

作为一元样本特征数字的直接推广，可以定义多元样本的特征数字（向量或矩阵），对于每一种定义，采用二种符号写出来，一种是一般记号，一种用矩阵记号，以便对照。

2. 样本平均值（向量）

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (1.2.2)$$