“十二五”普通高等教育本科国家级规划教材

21世纪统计学系列教材

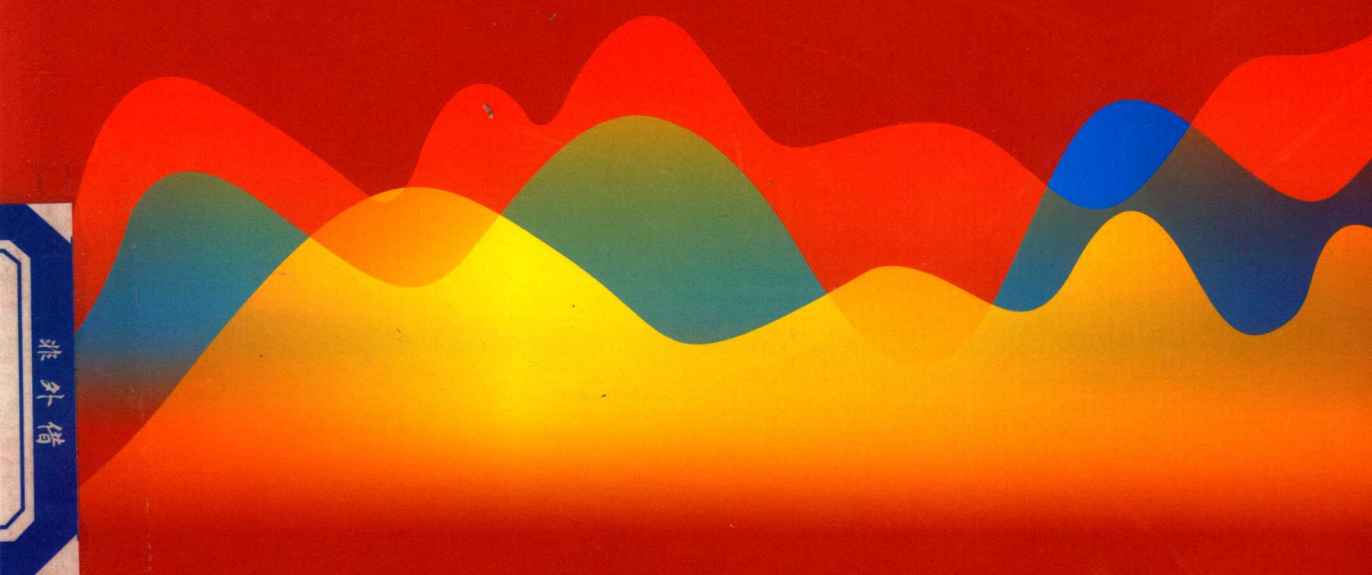
多元统计分析

（第5版）

何晓群　编著

Multivariate Statistical Analysis

(Fifth Edition)



中国人民大学出版社

“十二五”普通高等教育本科国家级规划教材

21世纪统计学系列教材

多元统计分析

（第5版）

何晓群　编著

Multivariate Statistical Analysis

(Fifth Edition)

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

多元统计分析/何晓群编著. —5版. —北京：中国人民大学出版社，2019.6
21世纪统计学系列教材
ISBN 978-7-300-26858-3

Ⅰ.①多… Ⅱ.①何… Ⅲ.①多元分析-统计分析-高等学校-教材 Ⅳ.①O212.4

中国版本图书馆CIP数据核字（2019）第066914号

“十二五”普通高等教育本科国家级规划教材
21世纪统计学系列教材
多元统计分析（第5版）
何晓群 编著
Duoyuan Tongji Fenxi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
经　　销	新华书店		
印　　刷	北京鑫丰华彩印有限公司	**版　　次**	2004年4月第1版
规　　格	185 mm×260 mm　16开本		2019年6月第5版
印　　张	17插页1	**印　　次**	2019年7月第2次印刷
字　　数	393 000	**定　　价**	39.00元

总序

Preface

改革开放以来，高等统计教育有了很大的发展。随着课程设置的不断调整，有不少教材出版，同时也翻译引进了一些国外优秀教材。作为培养我国统计专门人才的摇篮，中国人民大学统计学系自1952年创建以来，走过了风风雨雨，一直坚持着理论与应用相结合的办学方向，培养能够理论联系实际、解决实际问题的高层次人才。随着新知识经济和网络时代的到来，我们在教学科研的实践中，深切地感受到，无论是自然科学领域、社会科学领域的研究，还是国家宏观管理和企业生产经营管理，甚至人们的日常生活，信息需求量日益增多，信息处理技术更加复杂，作为信息技术支柱的统计方法，越来越广泛地应用于各个领域。

面对新的形势，我们一直在思索，课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要。在反复酝酿、不断尝试的基础上，我们决定与统计学界的同仁，共同编写、出版一套面向21世纪的统计学系列教材。

这套系列教材聘请了中科院院士、中国科学技术大学陈希孺教授，上海财经大学数量经济研究院张尧庭教授，中国科学院数学与系统科学研究所冯士雍研究员等作为编委。他们长期任中国人民大学的兼职教授，一直关心、支持着统计学系的学科建设和应用统计的发展。中国人民大学应用统计科学研究中心2000年已成为国家级研究基地，这些专家是首批专职或兼职研究人员。这一开放性研究基地的运作，将有利于提升我国应用统计科学研究的水平，也必将进一步促进高等统计教育的发展。

这套教材是我们奉献给新世纪的，希望它能促进应用统计教育水平的提高。这套教材力求体现以下特点：

第一，在教材选择上，主要面向经济类统计学专业。选材既包括统计教材也包括风险管理与精算方面的教材。尽管名为统计学系列教材，但并不求大、求全，而是力求精选。对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材，并未列入本次出版计划。

第二，每部教材的内容和写作，注意广泛吸收国内外优秀教材的成果。教材力求简明易懂、内容系统和实用，注重对统计方法思想的阐述，并结合大量实际数据和实例说明统计方法的特点及应用条件。

第三，强调与计算机的结合。为着力提高学生运用统计方法分析解决问题的能力，教材所涉及的统计计算，要求运用目前已有的统计软件。根据教材内容，选择使用SAS、SPSS、TSP、STATISTICA 、EViews、MINITAB、Excel、R等。

感谢中国人民大学出版社的同志们，他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望，经过反复论证，使这套教材得以出版。感谢参与教材编写的同行专家、统计学系的教师。愿大家的辛勤劳动能够结出丰硕的果实。我们期待着与统计学界的同仁，共同创造应用统计辉煌的明天。

易丹辉

前言

Foreword

多元统计分析是统计学中一个非常重要的分支。在国外，从20世纪30年代开始在自然科学、管理科学和社会、经济等领域广泛应用多元统计分析。我国自20世纪80年代起在多个领域拉开了多元统计分析应用的帷幕。

本书写作的指导思想是：在不失严谨的前提下，明显不同于纯数理类教材，努力突出实际案例的应用和统计思想的渗透，结合统计软件全面系统地介绍多元分析的实用方法。为了贯彻这一思想，本书参考了国内外大量书籍及文献，在系统介绍多元分析基本理论和方法的同时，尽力结合社会、经济、自然科学等领域的研究实例，把多元分析的方法与实际应用结合起来，注意定性分析与定量分析的紧密结合，努力把同行以及我们在实践中应用多元分析的经验和体会融入其中。几乎每种方法都强调它们各自的优缺点和实际运用中应注意的问题。为使读者掌握本书内容，同时考虑到这门课程的应用性和实践性，每章末给出了简单的思考与练习题。我们鼓励读者自己利用实际数据去实践这些方法。多元分析的应用离不开计算机，本书的案例主要运用在我国广泛流行的SPSS 23.0软件实现，部分方法用R软件完成。本书一个显著的特点是，在讲解每种方法后，结合实例概要介绍SPSS或R软件的操作实现过程。需要注意的是，读者不必拘泥于哪一软件，哪种软件使用方便就用哪种。在每章末还注明了参考文献，有兴趣的读者可进一步阅读。

全书共11章。主要内容包括多元正态分布、均值向量和协方差阵的检验、聚类分析、判别分析、主成分分析、因子分析、对应分析、典型相关分析等常见的主流方法，还参考国内外大量文献系统介绍了近年来在市场研究、顾客满意度研究、金融研究、环境研究等领域应用颇广的较新方法，包括定性数据的建模分析、对数线性模型、Logistic回归、多变量的图表示法、多维标度法等。

本书可作为统计学专业本科生的多元分析课程教材。由于本书的内容较多，教师在选用本书为教材时可以灵活选讲。本书还可作为非统计专业研究生量化分析教材。根据我们多年的教学实践，本书讲授48课时较为合适。

自2004年第1版出版以来，承蒙数万读者的厚爱，许多高校都采用其作为教材。有许多教师和学生给予我热情的鼓励，并且对书中某些地方提出中肯批评。这都是读者对我及本书的无私关心和奉献。在此我谨表衷心感谢。

本次再版充分考虑了广大读者的批评建议，在不失严谨的前提下仍然保持强调应用的风格。我对本书做了如下修订：

(1) 调整和更换了大部分例题；

(2) 对上一版中的一些错误进行了修正，并且斟酌了其中的一些文字；

（3）对于应用 SPSS 软件不方便的方法，采用了 R 软件；

（4）考虑到各个学校对课时的压缩，本次修订强调多元统计分析的主流方法，删去了路径分析、结构方程模型和联合分析等内容。

本书在写作过程中，始终得到中国人民大学 21 世纪统计学系列教材编审委员会和中国人民大学出版社的支持。编写大纲经过教材编审委员会的认真讨论，教材初稿得到吴喜之教授的认真审阅，吴教授还提出不少中肯意见。在此基础上，我对教材做了认真修改。若书中仍有不妥之处，责任当属笔者自负。本书的大部分案例是我们多年教学和科研工作的积累，有部分案例为体现其典型性引用他人著作。我还要特别感谢长期鼓励我进行应用研究的几位导师，他们是方开泰、陈希孺、张尧庭先生。在此，谨向对本书出版有过帮助的师长和朋友表示衷心的感谢。

本书的完成可以说是我们师生合作的共同成果。我的博士研究生王作成、付韶军、李因果、耿贵珍、王惠惠、胡小宁、马学俊、刘赛可，硕士研究生陈少杰、李强对本书的案例计算及输入做了大量工作。在我们的合作中，我不仅仅是他们的老师，还常常从他们的研究和提问中得到重要启发，教学相长在合作过程中得到真正体现。西北农林科技大学应用数学专业教授郭满才博士在我校访问时也为此书做了许多具体工作。本次再版修订得到安康学院数学与统计学院领导的理解和支持。由于水平有限，书中难免有不足之处，尤其是在一些应用研究的体会性讨论中恐有偏颇之处，恳切希望读者批评指正。

本书所提供的附表和案例数据，请读者到中国人民大学出版社网站（www. crup. com. cn）免费下载。

何晓群

目录 Contents

第1章 多元正态分布

学习目标

1. 掌握多元分布的有关概念；
2. 掌握统计距离的概念；
3. 理解多元正态分布的定义及其有关性质；
4. 了解常用多元分布及其抽样分布的定义和基本性质。

在基础统计学中，随机变量的正态分布在理论和实际应用中都有着重要的地位。同样，在多元统计学中，多元正态分布也占有相当重要的位置。原因是许多实际问题研究中的随机向量确实遵从或近似遵从多元正态分布；对于多元正态分布，已有一整套统计推断方法，并且可以得到许多完整的结果。

多元正态分布是最常用的一种多元概率分布。此外，还有多元对数正态分布、多项式分布、多元超几何分布、多元 β 分布、多元 χ^2 分布、多元指数分布等。本章从多维变量及多元分布的基本概念开始，着重介绍多元正态分布的定义及一些重要性质，以及常用多元分布及其抽样分布的定义和基本性质。

1.1 多元分布的基本概念

在研究社会、经济现象和许多实际问题时，经常遇到的是多指标的问题。例如研究职工薪酬构成情况时，计时工资、基础工资与职务工资、各种奖金、各种津贴等都是同时需要考察的指标；又如研究公司的运营情况时，要涉及公司的资金周转能力、偿债能力、获

利能力及竞争能力等财务指标，这些都是多指标研究的问题。显然，由于这些指标之间往往不独立，仅研究某个指标或者将这些指标割裂开来分别研究，都不能从整体上把握所研究问题的实质。一般，假设所研究的问题涉及 p 个指标，进行了 n 次独立观测，将得到 np 个数据，我们的目的就是对观测对象进行分组、分类，或分析这 p 个变量之间的相互关联程度，或找出内在规律等。下面简要介绍多元分析中涉及的一些基本概念。

1.1.1 随机向量

假定所讨论的是多个变量的总体，所研究的数据是同时观测 p 个指标（即变量），进行了 n 次观测得到的，我们把这 p 个指标表示为 X_1，X_2，…，X_p，常用向量

$$\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$$

表示对同一个体观测的 p 个变量。若观测了 n 个个体，则可得到如表 1-1 所示的数据，称每一个体的 p 个变量为一个样品，而全体 n 个样品形成一个样本。

表 1-1

序号	变量			
	X_1	X_2	…	X_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮		⋮
n	x_{n1}	x_{n2}	…	x_{np}

横看表 1-1，记

$$\boldsymbol{X}_{(\alpha)}=(x_{\alpha 1},x_{\alpha 2},\cdots,x_{\alpha p})',\quad \alpha=1,2,\cdots,n$$

它表示第 α 个样品的观测值。竖看表 1-1，第 j 列的元素：

$$\boldsymbol{X}_j=(x_{1j},x_{2j},\cdots,x_{nj})',\quad j=1,2,\cdots,p$$

表示对第 j 个变量 X_j 的 n 次观测数值。

因此，样本资料矩阵可用矩阵语言表示为：

$$\boldsymbol{X}=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p)=\begin{bmatrix}\boldsymbol{X}'_{(1)}\\ \boldsymbol{X}'_{(2)}\\ \vdots\\ \boldsymbol{X}'_{(n)}\end{bmatrix}$$

若无特别说明，本书所称向量均指列向量。

定义 1.1 设 X_1，X_2，…，X_p 为 p 个随机变量，由它们组成的向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 称为随机向量。

1.1.2 分布函数与密度函数

描述随机变量的最基本工具是分布函数。类似地，描述随机向量的最基本工具还是分

布函数。

定义 1.2 设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 是一随机向量，它的多元分布函数是

$$F(\boldsymbol{x})=F(x_1,x_2,\cdots,x_p)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_p\leqslant x_p) \tag{1.1}$$

式中，$\boldsymbol{x}=(x_1, x_2, \cdots, x_p)\in R^p$，并记成 $\boldsymbol{X}\sim F$。

多元分布函数的有关性质此处从略。

定义 1.3 设 $\boldsymbol{X}\sim F(\boldsymbol{x})=F(x_1, x_2, \cdots, x_p)$，若存在一个非负的函数 $f(\cdot)$，使得

$$F(\boldsymbol{x})=\int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p} f(t_1,t_2,\cdots,t_p)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_p \tag{1.2}$$

对一切 $\boldsymbol{x}\in R^p$ 成立，则称 $\boldsymbol{X}$（或 $F(\boldsymbol{x})$）有分布密度 $f(\cdot)$，并称 $\boldsymbol{X}$ 为连续型随机向量。

一个 p 维变量的函数 $f(\cdot)$ 能作为 R^p 中某个随机向量的分布密度，当且仅当

(i) $f(\boldsymbol{x})\geqslant 0, \quad \forall \boldsymbol{x}\in R^p$；

(ii) $\int_{R^p} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$。

例 1-1

若随机向量 (X_1, X_2, X_3) 有密度函数

$$f(x_1,x_2,x_3)=x_1^2+6x_3^2+\frac{1}{3}x_1x_2$$

$$0<x_1<1, \quad 0<x_2<2, \quad 0<x_3<\frac{1}{2}$$

容易验证它符合分布密度函数的两个条件 (i) 和 (ii)。

最重要的连续型多元分布——多元正态分布将留在 1.3 节讨论。

1.1.3 多元变量的独立性

定义 1.4 两个随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 称为相互独立的，若

$$P(\boldsymbol{X}\leqslant\boldsymbol{x},\boldsymbol{Y}\leqslant\boldsymbol{y})=P(\boldsymbol{X}\leqslant\boldsymbol{x})P(\boldsymbol{Y}\leqslant\boldsymbol{y}) \tag{1.3}$$

对一切 $\boldsymbol{x}$, $\boldsymbol{y}$ 成立。若 $F(\boldsymbol{x}, \boldsymbol{y})$ 为 $(\boldsymbol{X}, \boldsymbol{Y})$ 的联合分布函数，$G(\boldsymbol{x})$ 和 $H(\boldsymbol{y})$ 分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的分布函数，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立当且仅当

$$F(\boldsymbol{x},\boldsymbol{y})=G(\boldsymbol{x})H(\boldsymbol{y}) \tag{1.4}$$

若 $(\boldsymbol{X}, \boldsymbol{Y})$ 有密度 $f(\boldsymbol{x}, \boldsymbol{y})$，用 $g(\boldsymbol{x})$ 和 $h(\boldsymbol{y})$ 分别表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的分布密度，则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 独立当且仅当

$$f(\boldsymbol{x},\boldsymbol{y})=g(\boldsymbol{x})h(\boldsymbol{y}) \tag{1.5}$$

注意在上述定义中，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的维数一般是不同的。

类似地，若它们的联合分布等于各自分布的乘积，则称 p 个随机向量 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立。由 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立可以推知任何 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j(i\neq j)$ 独立，但是，若已知任何 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j(i\neq j)$ 独立，并不能推出 $\boldsymbol{X}_1$，$\boldsymbol{X}_2$，…，$\boldsymbol{X}_p$ 相互独立。

1.1.4 随机向量的数字特征

1. 随机向量 $\boldsymbol{X}$ 的均值

设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 有 p 个分量。若 $E(X_i)=\mu_i$ $(i=1, 2, \cdots, p)$ 存在，定义随机向量 $\boldsymbol{X}$ 的均值为：

$$E(\boldsymbol{X})=\begin{bmatrix}E(X_1)\\E(X_2)\\\vdots\\E(X_p)\end{bmatrix}=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix}=\boldsymbol{\mu} \tag{1.6}$$

$\boldsymbol{\mu}$ 是一个 p 维向量，称为均值向量。

当 $\boldsymbol{A}$，$\boldsymbol{B}$ 为常数矩阵时，由定义可立即推出如下性质：

$$(1)\quad E(\boldsymbol{AX})=\boldsymbol{A}E(\boldsymbol{X}) \tag{1.7}$$

$$(2)\quad E(\boldsymbol{AXB})=\boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B} \tag{1.8}$$

2. 随机向量 $\boldsymbol{X}$ 的协方差阵

$$\begin{aligned}\boldsymbol{\Sigma}&=\operatorname{cov}(\boldsymbol{X},\boldsymbol{X})=E(\boldsymbol{X}-\boldsymbol{EX})(\boldsymbol{X}-\boldsymbol{EX})'=D(\boldsymbol{X})\\&=\begin{bmatrix}D(X_1)&\operatorname{cov}(X_1,X_2)&\cdots&\operatorname{cov}(X_1,X_p)\\\operatorname{cov}(X_2,X_1)&D(X_2)&\cdots&\operatorname{cov}(X_2,X_p)\\\vdots&\vdots&&\vdots\\\operatorname{cov}(X_p,X_1)&\operatorname{cov}(X_p,X_2)&\cdots&D(X_p)\end{bmatrix}\\&=(\sigma_{ij})_{p\times p}\end{aligned} \tag{1.9}$$

称它为 p 维随机向量 $\boldsymbol{X}$ 的协方差阵，简称为 $\boldsymbol{X}$ 的协方差阵。

称 $|\operatorname{cov}(\boldsymbol{X}, \boldsymbol{X})|$ 为 $\boldsymbol{X}$ 的广义方差，它是协方差阵的行列式之值。

3. 随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差阵

设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 和 $\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_q)'$ 分别为 p 维和 q 维随机向量，它们之间的协方差阵定义为一个 $p\times q$ 矩阵，其元素是 $\operatorname{cov}(X_i, Y_j)$，即

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=(\operatorname{cov}(X_i,Y_j)),\quad i=1,2,\cdots,p;j=1,2,\cdots,q \tag{1.10}$$

若 $\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{0}$，称 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是不相关的。

当 $\boldsymbol{A}$，$\boldsymbol{B}$ 为常数矩阵时，由定义可推出协方差阵有如下性质：

(1) $D(\boldsymbol{AX})=\boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}'=\boldsymbol{A\Sigma A}'$

(2) $\operatorname{cov}(\boldsymbol{AX}, \boldsymbol{BY})=\boldsymbol{A}\operatorname{cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}'$

(3) 设 $\boldsymbol{X}$ 为 p 维随机向量，期望和协方差存在，记 $\boldsymbol{\mu}=E(\boldsymbol{X})$，$\boldsymbol{\Sigma}=D(\boldsymbol{X})$，$\boldsymbol{A}$ 为 $p\times p$

常数阵，则

$$E(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})=\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma})+\boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$$

对于任何随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 来说，其协方差阵 $\boldsymbol{\Sigma}$ 都是对称阵，同时总是非负定（也称半正定）的。大多数情形下是正定的。

4. 随机向量 $\boldsymbol{X}$ 的相关阵

若随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$ 的协方差阵存在，且每个分量的方差大于零，则 $\boldsymbol{X}$ 的相关阵定义为：

$$\boldsymbol{R}=(\mathrm{corr}(X_i, X_j))=(r_{ij})_{p\times p}$$

$$r_{ij}=\frac{\mathrm{cov}(X_i, X_j)}{\sqrt{D(X_i)}\sqrt{D(X_j)}},\quad i,j=1,2,\cdots,p \tag{1.11}$$

r_{ij} 也称为分量 X_i 与 X_j 之间的（线性）相关系数。

对于两组不同的随机向量 $\boldsymbol{X}$ 及 $\boldsymbol{Y}$，它们之间的相关问题将在典型相关分析的章节中详细讨论。

在数据处理时，为了克服由于指标的量纲不同对统计分析结果的影响，往往在使用某种统计分析方法之前，将每个指标“标准化”，即做如下变换：

$$X_j^*=\frac{X_j-E(X_j)}{[D(X_j)]^{1/2}},\quad j=1,2,\cdots,p \tag{1.12}$$

$$\boldsymbol{X}^*=(X_1^*, X_2^*, \cdots, X_p^*)'$$

于是

$$E(\boldsymbol{X}^*)=\boldsymbol{0}$$

$$D(\boldsymbol{X}^*)=\mathrm{corr}(\boldsymbol{X})=\boldsymbol{R}$$

即标准化数据的协方差阵正好是原指标的相关阵：

$$\boldsymbol{R}=D(\boldsymbol{X}^*)=E(\boldsymbol{X}^*\boldsymbol{X}^{*\prime}) \tag{1.13}$$

1.2 统计距离

在多指标统计分析中，距离的概念十分重要，样品间的不少特征都可用距离来描述。大部分多元方法是建立在简单的距离概念基础上的，即平时人们熟悉的欧氏距离，或称直线距离。如几何平面上的点 $P=(x_1, x_2)$ 到原点 $O=(0, 0)$ 的欧氏距离，依勾股定理有

$$d(O,P)=(x_1^2+x_2^2)^{1/2} \tag{1.14}$$

一般，若点 P 的坐标 $P=(x_1, x_2, \cdots, x_p)$，则它到原点 $O=(0, 0, \cdots, 0)$ 的欧氏距离，依勾股定理有

$$d(O,P)=\sqrt{x_1^2+x_2^2+\cdots+x_p^2} \tag{1.15}$$

所有与原点距离为 C 的点满足方程

$$d^2(O,P)=x_1^2+x_2^2+\cdots+x_p^2=C^2 \tag{1.16}$$

因为这是一个球面方程（$p=2$ 时是圆），所以，与原点等距离的点构成一个球面，任意两个点 $P=(x_1,\ x_2,\ \cdots,\ x_p)$ 与 $Q=(y_1,\ y_2,\ \cdots,\ y_p)$ 之间的欧氏距离为：

$$d(P,Q)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_p-y_p)^2} \tag{1.17}$$

但就大部分统计问题而言，欧氏距离是不能令人满意的。这是因为每个坐标对欧氏距离的贡献是同等的。当坐标轴表示测量值时，它们往往带有大小不等的随机波动，在这种情况下，合理的办法是对坐标加权，使变化大的坐标比变化小的坐标有较小的权系数，这就产生了各种距离。

欧氏距离还有一个缺点，那就是当各个分量为不同性质的量时，“距离”的大小竟然与指标的单位有关。例如，横轴 x_1 代表重量（以 kg 为单位），纵轴 x_2 代表长度（以 cm 为单位）。有四个点 A，B，C，D，它们的坐标如图 1-1 所示。

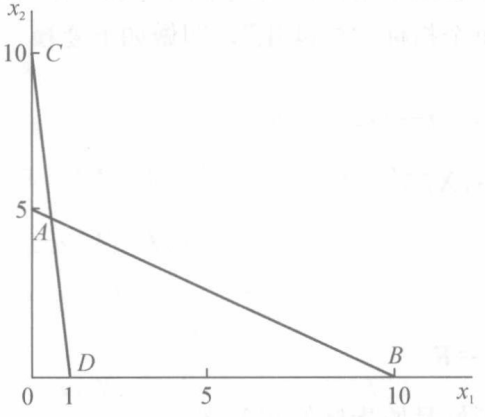


图 1-1

这时

$$AB=\sqrt{5^2+10^2}=\sqrt{125}$$

$$CD=\sqrt{10^2+1^2}=\sqrt{101}$$

显然，AB 要比 CD 长。

现在，如果 x_2 用 mm 作单位，x_1 单位保持不变，此时点 A 坐标为（0，50），点 C 坐标为（0，100），则

$$AB=\sqrt{50^2+10^2}=\sqrt{2\,600}$$

$$CD=\sqrt{100^2+1^2}=\sqrt{10\,001}$$

结果 CD 反而比 AB 长！这显然是不够合理的。因此，有必要建立一种距离，这种距离应能够体现各个变量在变差大小上的不同，以及有时存在的相关性，还要求距离与各变量所

用的单位无关。看来，我们选择的距离要依赖于样本方差和协方差。因此，采用“统计距离”这个术语，以区别通常习惯用的欧氏距离。

下面先介绍统计距离。

设 $P=(x_1, x_2, \cdots, x_p)$，$Q=(y_1, y_2, \cdots, y_p)$，且 Q 的坐标是固定的，P 的坐标相互独立地变化。用 S_{11}，S_{22}，…，S_{pp} 表示 p 个变量 X_1，X_2，…，X_p 的 n 次观测的样本方差。将坐标的各维度除以相应变量的样本标准差 $\sqrt{S_{ii}}$，得到标准化的坐标，其中各变量的样本标准差的倒数可以看作坐标各维度的权重系数，则 P 到 Q 的统计距离为：

$$d(P,Q)=\sqrt{\frac{(x_1-y_1)^2}{S_{11}}+\frac{(x_2-y_2)^2}{S_{22}}+\cdots+\frac{(x_p-y_p)^2}{S_{pp}}} \tag{1.18}$$

所有与点 Q 的距离平方为常数的点 P 构成一个椭球，其中心在点 Q，其长短轴平行于坐标轴。容易看到：

(1) 在式 (1.18) 中，令 $y_1=y_2=\cdots=y_p=0$，得到点 P 到原点 O 的距离。

(2) 如果 $S_{11}=S_{22}=\cdots=S_{pp}$，则用欧氏距离式 (1.17) 是方便可行的。

还可以利用旋转变换的方法得到合理的距离。考虑点 $P=(x_1, x_2, \cdots, x_p)$ 和点 $Q=(y_1, y_2, \cdots, y_p)$，这里 Q 为固定点，而 P 的坐标是变化的，且彼此相关，$O=(0, 0, \cdots, 0)$ 为坐标原点，则 P 到 O 和 Q 的距离分别为：

$$\begin{aligned} d(O,P)&=(a_{11}x_1^2+a_{22}x_2^2+\cdots+a_{pp}x_p^2+2a_{12}x_1x_2+\cdots+2a_{p-1,p}x_{p-1}x_p)^{1/2}\\ &=(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})^{1/2} \end{aligned} \tag{1.19}$$

和

$$\begin{aligned} d(P,Q)=&[a_{11}(x_1-y_1)^2+a_{22}(x_2-y_2)^2+\cdots+a_{pp}(x_p-y_p)^2\\ &+2a_{12}(x_1-y_1)(x_2-y_2)+\cdots\\ &+2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)]^{1/2}\\ =&[(\boldsymbol{X}-\boldsymbol{Y})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{Y})]^{1/2} \end{aligned} \tag{1.20}$$

这里

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p}\\ a_{21} & a_{22} & \cdots & a_{2p}\\ \vdots & \vdots & & \vdots\\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix},\quad \boldsymbol{X}=\begin{bmatrix} x_1\\ x_2\\ \vdots\\ x_p \end{bmatrix},\quad \boldsymbol{Y}=\begin{bmatrix} y_1\\ y_2\\ \vdots\\ y_p \end{bmatrix}$$

且 $\boldsymbol{A}$ 为对称阵，满足条件：对任意的 $\boldsymbol{X}$，恒有 $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\geqslant\boldsymbol{0}$，且等号成立当且仅当 $\boldsymbol{X}=\boldsymbol{0}$，即 $\boldsymbol{A}$ 为正定方阵。

最常用的一种统计距离是印度统计学家马哈拉诺比斯 (Mahalanobis) 于 1936 年引入的，称为马氏距离。下面先用一个一维的例子说明欧氏距离与马氏距离在概率上的差异。设有两个一维正态总体 G_1：$N(\mu_1, \sigma_1^2)$ 和 G_2：$N(\mu_2, \sigma_2^2)$。若有一个样品，其值在点 A 处，点 A 距离哪个总体近些呢？如图 1-2 所示。