实用多元统计分析

Applied Multivariate Statistical Analysis

管 宇 主编

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

实用多元统计分析

管　宇　主编

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

内容提要

多元统计分析是统计学的一个分支，主要对多个对象和多个指标进行统计学意义上的综合分析，是进行数据处理和分析最重要的方法。本书是针对应用型人员的学习编写的，每章编排流程是基本概念和背景知识、统计方法、SPSS操作、案例分析，其中对案例的深度分析和解读，方便读者的学习和模仿。主要内容包括：多元统计分析概述、常见统计分布、多元数据的图形表示、差异性分析、相关性分析、回归分析、判别分析、聚类分析、主成分分析、因子分析、多维尺度分析、对应分析。

本书适合于作为非概率统计专业的研究生、本科生多元统计分析课程教材或教学参考书，也可作为科技工作者、教师、从事数据分析读者的学习参考书。

图书在版编目（CIP）数据

实用多元统计分析 / 管宇主编. —杭州：浙江大学出版社，2011.10

ISBN 978-7-308-09180-0

Ⅰ.①实… Ⅳ.①管… Ⅲ.①多元分析：统计分析 Ⅳ.①O212.4

中国版本图书馆CIP数据核字（2011）第209437号

实用多元统计分析

管　宇　主编

责任编辑　王元新
封面设计　十木米
出版发行　浙江大学出版社
（杭州市天目山路148号　邮政编码310007）
（网址：http://www.zjupress.com）
排　　版　杭州中大图文设计有限公司
印　　刷　临安市曙光印务有限公司
开　　本　787mm×960mm　1/16
印　　张　17
字　　数　342千
版 印 次　2011年10月第1版　2011年10月第1次印刷
书　　号　ISBN 978-7-308-09180-0
定　　价　35.00元

前　言

多元统计分析起源于20世纪初，1928年Wishart发表的论文《多元正态总体样本协差阵的精确分布》开创了多元分析。20世纪30年代R. A. Fisher、H. Hotelling、S. N. Roy、许宝騄等人作了一系列的奠基性工作，使多元分析在理论上得到了迅速发展。20世纪40年代其在心理、教育、生物等方面有了不少的应用，但由于计算量大，使其发展受到影响，甚至停滞了相当长时间。20世纪50年代中期，随着电子计算机的出现和发展，使多元分析方法在地质、气象、医学、社会学等方面得到了广泛应用。20世纪60年代通过应用和实践其理论得到了完善和发展，而由于新的理论、新的方法不断涌现又促使它的应用范围更加扩大。20世纪70年代初期在我国开始受到各个领域的关注，目前我国在多元统计分析的理论研究和应用上已取得了显著成绩。

进入21世纪后，人们获得的数据正以前所未有的速度急剧增加，产生了很多超大型数据库，遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学以及政府统计等领域，多元统计与人工智能和数据库技术相结合，已在经济、商业、金融、天文等行业得到了成功的应用。“多元统计分析”课程已被越来越多将来需要与大量数据打交道的研究生、本科生列为必修或选修课程。

编者多年从事非统计类研究生、本科生“多元统计分析”课程的教学，深感一本简练但又实用教材的重要性。随着高等教育进一步“大众化”，计算机硬件、软件的发展，学习多元统计分析似乎不太有研究生、本科生的区别了，因为大家学习它多是作为工具在电脑上进行数据的处理和分析。

本书共有12章。第1章是多元统计分析概述，涉及统计学一些基本概念，简单介绍常见统计软件；第2章统计分布，包括常见的一维和多维分布，以及分布检验；第3章可视化分析，介绍五种数据的图形表示；第4章差异性分析，包括假设检验、均值向量的检验、方差分析、协方差阵的检验、协方差分析；第5章相关性分析，对简相关、偏相关、复相关、典型相关都作了介绍；第6章回归分析，重点介绍了线性回归和非线性回归的思想与方法；第7章至第10章分别是最经典的多元统计内容：判别分析、聚类分析、主成分分析、因子分析；第11章多维尺度分析；第12章对应分析。

各章编写结构是：先引言介绍基本概念和背景知识，再进行相应的统计方法介绍，

再 SPSS 操作流程和其中选项的意思解释，最后案例分析。其中案例分析是对输出图表进行详细解读。读者只要能完全按照本书的操作流程操作，就会轻松获得各种统计图表；再模仿案例分析中的解读，就能对自己辛苦做出的实验数据作出深度的统计分析。

与多元统计分析同类书籍相比较，本书有如下特点：

1. 内容更完整。本书编写初衷是想编写一本适合于不需要太多高等数学知识的研究生和高年级大学生的学习和参考用书，希望在实际处理数据时能够方便地在本书中找到想要的统计方法。一般的多元统计分析书籍都不会详细介绍统计分布、差异性检验、回归分析等。

2. 简单但实用。人们学习统计多数是为了用，公式不能不讲，但要在不长的时间里掌握这么多统计方法怎么用，绝大多数的数学演绎就只能完全放弃。个别简单的公式推导被放在练习题中，读者翻阅其他书籍很容易找到这些理论的推导。由于少了大量的数学演绎和演算过程，与同名教材相比，本书内容较多但篇幅却明显地较少。我们的重点是放在对统计软件输出的结果解释上。

3. 直面应用统计中存在的问题。统计方法很有用，人人都来用，但许多情况下统计却被滥用。统计方法的使用都是有前提条件的，有些软件输出结果非常吻合实际经验，但其前提条件却不满足。本书就有两个这方面的案例，作为教学为的是介绍方法因此对输出结果是照讲不误；但在作实际数据分析时必须设法换方法研究。

概率统计和线性代数的知识是学习多元统计分析的基本要求，但不是绝对要求；多模仿多操作，才是学好多元统计分析的关键所在。本书的读者对象是理工农科类、经济类的本科生和研究生，以及其他各个领域中需要进行数据分析处理的实际工作者。本书适用于每周 3～4 学时、每学期约讲授 54～72 学时“多元统计分析”课程或相关课程的教材。若 48 学时或 32 学时，其中有些内容任课教师酌情选用。但编者建议不管学时多少，每一章的实际背景和基本方法都应该介绍。

本书编写过程中参考了国内外相关文献，书后只是列出了其中主要的参考书籍，衷心地感谢为统计学作出贡献的所有前辈和同仁们！本书的出版得到了浙江大学出版社的大力支持，在此表示感谢！编者非常努力地想完成一本令读者满意的书，但难免有这样或那样的错误和欠缺，恳请读者提出宝贵意见，以便进一步修改与完善。

本书的编写得到浙江农林大学研究生部专项经费的资助。

管 宇

2011 年 6 月

目　录

第1章

多元统计分析概述

1.1 引 言

统计学(statistics)作为应用数学的一个分支，主要是通过利用概率论建立数学模型，收集所观测到的数据，进行量化分析、总结，进而进行推断和预测，从而对决策和行动提供依据和建议。它被广泛地应用在各个领域，从自然科学到社会科学，以及生产、生活中的各种决策等。统计学通常包括描述统计学和推断统计学，前者是整个统计学的基础，后者是现代统计学的主要内容。

1.1.1 描述统计学

描述统计学(descriptive statistics)或描述统计是描绘或总结观测量的基本情况的统计总称。

一是对数据资料进行图像化处理，将资料摘要变为图表，以直观了解整体资料分布的情况。通常会使用的工具是频数分布表(frequency distribution table)与图示法，如多边图(polygon)、直方图(histogram)、条形图(barchart)、饼图(piechart)、散点图(scatterplot)等。

二是利用某些指标以了解各变量内的观测值集中与分散的情况，运用的工具有：集中量数(measure of central location)，如平均数(mean)、中位数(median)、众数(mode)、几何平均数(geometric mean)、调和平均数(harmonic mean)；变异量数(measure of variation)，如全距或极差(range)、平均差(average deviation)、标准差(standard deviation)、相对差、四分差(quartiledeviation)等。

1.1.2 推断统计学

推断统计学或统计推断(statistical inference)是研究如何利用样本数据来推断总体特征的统计方法。它既可以用于对总体参数的估计，也可以用作对总体某些分布特

征的假设检验。

多元统计分析(multivariate statistical analysis)是从统计学中发展起来的一个分支,是一种综合分析方法,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。如果每个个体有多个观测数据,或者从数学上说,个体的观测数据能表示为 p 维空间的点,那么这样的数据叫做多元数据(multivariate data),而分析多元数据的统计方法就叫做多元统计分析。它是统计学中的一个重要的分支学科。

20 世纪 30 年代,费希尔(R. A. Fisher)、霍特林(H. Hotelling)、许宝騄以及罗伊(S. N. Roy)等人作出了一系列奠基性的工作,使多元统计分析在理论上得到迅速发展。50 年代中期,随着电子计算机的发展和普及,多元统计分析在地质、气象、生物、医学、图像处理、经济分析等许多领域得到了广泛的应用,同时也促进了理论的发展。各种统计软件包如 SAS、SPSS、S-Plus 等,使实际工作者利用多元统计分析方法解决实际问题更简单方便。重要的多元统计分析方法有多元方差分析、多元回归分析(简称回归分析)、判别分析、聚类分析、主成分分析、对应分析、因子分析、典型相关分析等。多元统计分析的重要基础之一是多元正态分析。它有狭义与广义之分,当假定总体分布是多元正态分布时,称为狭义的,否则称为广义的。

1.2 变量和统计方法选择

1.2.1 变量的概念和分类

变量描述的是变化的量,是运用统计方法所分析的对象。例如,人的身高,我们只能说某人在某一时刻时的身高是多少,如果在另一时间他的身高可能变成另一个值。数据是与变量相关的值,就是统计所要分析的"数据"。统计界有句名言:"统计是用数据说话的",即统计是用来处理数据的。获取数据是统计工作的基础,探索数据的内在数量规律即统计规律是统计的最终目的。

1. 以取值属性来分类

变量按取值属性可分为数值变量和分类变量。

(1)数值变量(numerical variable):也称为定量变量,其变量值用数量表示。数值变量可进一步分为离散变量和连续变量。

离散变量(discrete variable)是指其数值只能取有限个或无限但可数个。这种变量的数值一般用计数方法取得。反之,在一定区间内可以任意取值的变量叫连续变量(continuous variable),其数值是连续不断的,相邻两个数值之间可作无限分割,即可取无限个数值。例如,人体测量的身高、体重、胸围等为连续变量,其数值只能用测量或计量的方法取得。

(2)分类变量(categorical variable):也称为定性变量,其变量值是定性的,表现为互不相容的类别或属性。分类变量可分为无序分类变量和有序分类变量两类。

无序分类变量(unordered categorical variable)是指所分类别或属性之间无程度和顺序的差别,对于无序分类变量的分析,应先按类别分组,清点各组的观测单位数,编制分类变量的频数表,所得资料为无序分类资料,亦称计数资料。

有序分类变量(ordinal categorical variable)各类别之间有程度的差别。如调查结果按非常满意、满意、无所谓、不满意、非常不满意分类。对于有序分类变量,应先按等级顺序分组,清点各组的观测单位个数,编制有序变量(各等级)的频数表,所得资料称为等级资料。

当然变量类型不是一成不变的,根据研究目的的需要,各类变量间可以进行转化。例如血红蛋白量(g/L)原属数值变量,若按血红蛋白正常与偏低分为两类时,可按两项分类资料分析;若按重度贫血、中度贫血、轻度贫血、正常、血红蛋白增高分为五个等级时,可按等级资料分析。有时亦可将分类资料数量化,如可将病人的恶心反应以 0、1、2、3 表示,则可按数值变量资料(定量资料)分析。

2. 以测量尺度来分类

变量按照测量它们的尺度不同,又可以分为三类。

(1)间隔尺度(interval measure):指标度量时用数量来表示,其数值由测量或计数统计得到,如长度、重量、收入、支出等。一般来说,计数得到的数量是离散数量,测量得到的数量是连续数量。在间隔尺度中如果存在绝对零点,又称比例尺度。

(2)顺序尺度(ordinal measure):指标度量时没有明确的数量表示,只有次序关系,或虽用数量表示,但相邻两数值之间的差距并不相等,它只表示一个有序状态序列。如评价酒的味道,分成好、中、次三等,三等有次序关系,但没有数量表示。

(3)名义尺度(nominal measure):指标度量时既没有数量表示也没有次序关系,只有一些特性状态,如眼睛的颜色、化学中催化剂的种类等。在名义尺度中只取两种特性状态的变量是很重要的,如电路的开和关,天气的有雨和无雨,人口性别的男和女,医疗诊断中的“+”和“−”,市场交易中的买和卖等。

1.2.2　统计方法选择

自然界和社会实践中发生的现象是多种多样的。有一类现象在一定条件下必然发生,它的结果总是肯定的,我们称之为确定性现象或必然现象(inevitable phenomenon)。如同性电荷必定互相排斥、能量守恒定律等。研究这类现象的数学工具有数学分析、几何、代数、微分方程等。如果在相同条件下重复进行试验,每次结果未必相同,也就是事前不可预言,则称之为偶然现象或随机现象(random phenomenon)。如:以同样的方式抛置硬币却可能出现正面向上也可能出现反面向上。研究这类现象的数学工具是

概率论和统计。应该说，必然现象是理想的和相对的，随机现象是现实的和绝对的。

从理想角度来讲，只要知道今天的所有气象数据和天气变化规律，人们就能够准确无误地知道明天的天气。但是量子力学的“测不准原理(uncertainty principle)”却说任一粒子的位置和动量不可能同时测准，一个量越确定，另一个量的不确定程度就越大。由于今天的气象数据不可能完全准确观测到，因此天气预报永远只是预报。当然，随着技术条件的进一步提高，人类会让预报的准确率越来越高，但绝对的准确是永远无法实现的。应该说，随着科学技术水平的提高，许多曾经的不确定现象慢慢地变成了确定现象，如现在对许多病症患者通过各种现代医疗仪器可非常准确地诊断，这在没有这些设备的古代要想判断准确身上到底长了什么东西是难以想象的。但是不确定现象是永远不会全部消失的，因为我们避免不了观测误差，只能说是人们判断准确率越来越高。

大千世界无奇不有。由于客观世界和现实生活中有太多的不确定性，作为数据分析的最主要方法——统计，人们也自然而然地设计出数不胜数的统计方法。每一种统计方法都主要针对某些特定背景的问题和变量，表 1-1、表 1-2、表 1-3 列出近似理想状态下的统计方法选择。由于随机误差的普遍存在且预先的不确定，面对一个实际问题很难说有一最优统计处理方法，通常需要几种统计方法分析对比才能得出相对可靠的推断。医生诊断患者疾病时，先进行多种体检化验再会诊且仍存在误诊可能，统计分析与其非常类似。

表 1-1 不同变量类型的数据分析方法选择

因变量	自变量		
	数值变量	分类变量	有序变量
数值变量	相关分析、回归分析	回归分析	相关分析、回归分析
分类变量	logistic 回归分析、聚类分析、判别分析	logistic 回归分析、χ^2 检验	χ^2 检验
有序变量	logistic 回归分析、聚类分析、判别分析	logistic 回归分析、χ^2 检验	相关分析、χ^2 检验

表 1-2 不同研究设计和数据类型的数据分析方法选择

因变量	研究设计类型				
	两组比较	两组以上比较	配对比较	重复测量	两变量间的联系
数值变量	t 检验	方差分析	配对 t 检验	方差分析	回归分析、Pearson 相关系数
分类变量	χ^2 检验	χ^2 检验	配对 χ^2 检验		列联表相关系数
有序变量	Mann-Whitney 秩和检验	Kraskal-Wallis 分析	Wilcoxon 符号秩和检验		Spearman 相关系数

表 1-3　统计方法和研究问题之间的关系

问　题	内　容	方　法
数据或结构性化简	尽可能简单地表示所研究的现象，但不损失很多有用的信息，并希望这种表示能够很容易解释	多元回归分析、聚类分析、主成分分析、因子分析、对应分析、多维尺度法、可视化分析
分类和组合	基于所测量到的一些特征，给出好的分组方法，对相似的对象或变量分组	判别分析、聚类分析、主成分分析、可视化分析
变量之间的相关关系	变量之间是否存在相关关系，相关关系又是怎样体现的	多元回归、典型相关、主成分分析、因子分析、对应分析、多维尺度法、可视化分析
预测与决策	通过统计模型或最优准则，对未来进行预见或判断	多元回归、判别分析、聚类分析、可视化分析
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，能够证实某种假设条件的合理性	多元总体参数估计、假设检验

1.3　随机向量

1.3.1　随机向量的基本概念

实际问题多是非常复杂的，一个随机试验往往要用 s 个指标（变量）整体地讨论其结果。如植物个体生长状态要同时用树高、胸径、树冠、树根等才能完整地加以描述，虽然每个个体的各项指标值会不相同也无法预测，但通常是有一定规律的，即所谓的统计规律。我们就用随机变量或随机向量代表这些指标，变量或向量所取值对应于实际的指标值，这些指标值表现出来的统计规律用统计分布来表达（通常是近似表达）。

设从同一总体中随机抽取 n 个个体，每个个体都观测其 s 个指标，得数据如表 1-4 所示。

表 1-4　变量数据表

样品	指标 1	指标 2	…	指标 j	…	指标 s
1	X_{11}	X_{12}	…	X_{1j}	…	X_{1s}
2	X_{21}	X_{22}	…	X_{2j}	…	X_{2s}
⋮	⋮	⋮		⋮		⋮
i	X_{i1}	X_{i2}	…	X_{ij}	…	X_{is}
⋮	⋮	⋮		⋮		⋮
n	X_{n1}	X_{n2}	…	X_{nj}	…	X_{ns}

在表 1-4 中，X_{ij} 表示第 i 样品的第 j 个指标值。观测之前这些指标值都是未知的，也无法预测，它们都是随机变量。

表 1-4 中的数据通常用矩阵加以表示：

$$\boldsymbol{X}_{n\times s}=\begin{bmatrix}X_{11} & X_{12} & \cdots & X_{1s}\\ X_{21} & X_{22} & \cdots & X_{2s}\\ \vdots & \vdots & & \vdots\\ X_{n1} & X_{n2} & \cdots & X_{ns}\end{bmatrix}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)=\begin{bmatrix}\boldsymbol{X}'_{(1)}\\ \boldsymbol{X}'_{(2)}\\ \vdots\\ \boldsymbol{X}'_{(n)}\end{bmatrix}$$

其中，$\boldsymbol{X}_j=(X_{1j},X_{2j},\cdots,X_{nj})'(j=1,2,\cdots,s)$ 表示第 j 个指标的 n 次观测值；$\boldsymbol{X}_{(i)}=(X_{i1},X_{i2},\cdots,X_{is})'(i=1,2,\cdots,n)$ 表示第 i 个样品的 s 个指标的观测值。本书中矩阵右上角加“′”表示该矩阵的转置矩阵。我们简记 $\boldsymbol{X}_{n\times s}$ 为 $\boldsymbol{X}$，而 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 被称为 s 维(元) 随机向量(random vector)，$\boldsymbol{X}_1$、$\boldsymbol{X}_2$、…、$\boldsymbol{X}_s$ 是 s 个一维随机向量。一维随机向量通常称为随机变量(random variable)。

一般地，使用粗体英文字母表示非一维的矩阵、向量，不加粗的表示单一变量；大写英文字母 X、Y、Z 代表随机变量，小写英文字母 x、y、z 代表相应的观测值。但为了全文书写方便有时会混用，如本书中加粗既表示多元也代表一元。数学字母本身只是一种符号，多数情况下没有硬性的统一规定，除极个别外如通常意义下“π”代表圆周率，但统计学中许多人用 $\pi(\lambda)$ 表示泊松分布，随机过程中的极限分布通常也用 π 表示。

对于随机向量，一般主要讨论离散型和连续型随机向量以及它们的分布。类似于一维随机变量，s 维(元) 随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的分布函数(distribution function) 被定义为

$$F(\boldsymbol{x})=F(x_1,x_2,\cdots,x_s)=P(\boldsymbol{X}_1\leqslant x_1,\boldsymbol{X}_2\leqslant x_2,\cdots,\boldsymbol{X}_s\leqslant X_s)$$

其中，$\boldsymbol{x}=(x_1,x_2,\cdots,x_s)'$ 是 s 维空间中的点，并记 $\boldsymbol{X}\sim F(\boldsymbol{x})$。

定义 1.1 如果 s 维(元) 随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的可能取值是有限组或可列无限组 $(x_{i1},x_{i2},\cdots,x_{is})'$，$i=1,2,\cdots$，则称 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 为 s 维离散型随机向量，将 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 取每组值的概率 $P(\boldsymbol{X}_1=x_{i1},\boldsymbol{X}_2=x_{i2},\cdots,\boldsymbol{X}_s=x_{is})$ $(i=1,2,\cdots)$ 称为 s 维(元) 离散型随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的联合分布律。

定义 1.2 设 s 维(元) 随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的联合分布函数 $F(x_1,x_2,\cdots,x_s)$，如果存在非负函数 $f(x_1,x_2,\cdots,x_s)$，对任意 $\boldsymbol{x}=(x_1,x_2,\cdots,x_s)\in\boldsymbol{R}^s$，有

$$F(x)=F(x_1,x_2,\cdots,x_s)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(x_1,x_2,\cdots,x_s)\mathrm{d}x_1\cdots\mathrm{d}x_s$$

则 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 为称 s 维(元) 连续型随机向量，$f(x_1,x_2,\cdots,x_s)$ 称为 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的联合概率密度或联合密度函数。

离散型随机向量的联合分布律和连续型随机向量的联合密度函数都是非负的且具有归一性。分布是随机现象的一种数学抽象，目的是便于定量地分析处理该随机现象。用

定量的方法主要是数学方法或统计方法解决处理实际问题，称为数学建模(mathematics modeling)或统计建模(staistics modeling)，其中的方法和公式称为模型(model)。

定义 1.3　设 s 维(元)随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$，由它的 $q(q<s)$ 个分量组成的子向量 $(X_{k1},X_{k2},\cdots,X_{kq})'$ 的分布称为 $\boldsymbol{X}$ 的边缘分布(marginal distribution)，只要将 $(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$ 的分布函数 $F(x_1,x_2,\cdots,x_s)$ 中非 $(X_{k1},X_{k2},\cdots,X_{kq})'$ 项的变量值都取为 $+\infty$ 即可。

1.3.2　数字特征

随机向量的分布函数，能完整地描述随机向量的统计规律。在实际问题中，除个别简单的随机向量外，通常其分布函数非常复杂甚至难以显式表达，但是它们的某些特征却往往容易知道。如在自然和社会中，对称或近似对称是非常普遍的现象，它们的平均值就是对称中心或对称轴。最重要的数字特征是数学期望和协方差。

定义 1.4　设 s 维(元)随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$，若对每个分量都存在有限的数学期望 $E(X_i)=\mu_i(i=1,2,\cdots,s)$，则称 $E(\boldsymbol{X})=\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_s)'$ 为 $\boldsymbol{X}$ 的数学期望或均值(向量)。

均值向量具有以下运算性质：

(1) $E(\boldsymbol{AX})=\boldsymbol{A}E(\boldsymbol{X})$；

(2) $E(\boldsymbol{AXB})=\boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$；

(3) $E(\boldsymbol{AX}+\boldsymbol{BY})=\boldsymbol{A}E(\boldsymbol{X})+\boldsymbol{B}E(\boldsymbol{Y})$。

其中，随机向量 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_s)'$，$\boldsymbol{Y}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_s)'$，$\boldsymbol{A}$、$\boldsymbol{B}$ 为大小适合运算的常数矩阵。

需要特别强调的，我们学习数学和统计的目的是为了计算和处理数据，尽可能客观地、定量地解决实际问题。因此，从本质上讲数学和统计书籍的编写格式基本都是千篇一律的：通常是先用式子(许多时候这些式子被称为公理)给出概念或定义(什么是什么，往往对应于某一类实际问题的抽象)，再从定义公理出发演绎出一些主要的运算规律(以性质、定理、公式形式出现)。本书是为统计应用学习编写的，许多严格的数学推演都被略去而只列出结论，学习时关键是理解这些结果，清楚它们在数学和统计学意义上的含义和实际应用背景，对证明推理感兴趣的读者可参阅书后列出的相关参考文献。

譬如矩阵理论，先定义什么是矩阵和一些特殊形状的矩阵，再规定常数与矩阵、矩阵与矩阵之间的加减乘除运算法则，然后推出一系列运算公式。其中逆矩阵的引入就是为了矩阵间的相除运算。只要知道了数学和统计的这一本质，也就明白了为什么会有那么多性质定理公式，才会有“从薄到厚，再从厚到薄”的学习过程。

定义 1.5 设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_s)',\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_s)'$，称

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=E[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{Y}-E(\boldsymbol{Y}))]'$$

$$=\begin{bmatrix}\operatorname{cov}(X_1,Y_1) & \operatorname{cov}(X_1,Y_2) & \cdots & \operatorname{cov}(X_1,Y_s)\\ \operatorname{cov}(X_2,Y_1) & \operatorname{cov}(X_2,Y_2) & \cdots & \operatorname{cov}(X_2,Y_s)\\ \vdots & \vdots & \vdots & \vdots\\ \operatorname{cov}(X_s,Y_1) & \operatorname{cov}(X_s,Y_2) & \cdots & \operatorname{cov}(X_s,Y_s)\end{bmatrix}$$

为 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的协方差阵。当 $\boldsymbol{X}=\boldsymbol{Y}$ 时，记 $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=\operatorname{cov}(\boldsymbol{X})=D(\boldsymbol{X})$ 或 $\sum$。称

$$\boldsymbol{R}_{XY}=(r_{ij})_{s\times s}$$

$$=\begin{bmatrix}\dfrac{\operatorname{cov}(X_1,Y_1)}{\sqrt{D(X_1)}\sqrt{D(Y_1)}} & \dfrac{\operatorname{cov}(X_1,Y_2)}{\sqrt{D(X_1)}\sqrt{D(Y_2)}} & \cdots & \dfrac{\operatorname{cov}(X_1,Y_s)}{\sqrt{D(X_1)}\sqrt{D(Y_s)}}\\ \dfrac{\operatorname{cov}(X_2,Y_1)}{\sqrt{D(X_2)}\sqrt{D(Y_1)}} & \dfrac{\operatorname{cov}(X_2,Y_2)}{\sqrt{D(X_2)}\sqrt{D(Y_2)}} & \cdots & \dfrac{\operatorname{cov}(X_2,Y_s)}{\sqrt{D(X_2)}\sqrt{D(Y_s)}}\\ \vdots & \vdots & \vdots & \vdots\\ \dfrac{\operatorname{cov}(X_s,Y_1)}{\sqrt{D(X_s)}\sqrt{D(Y_1)}} & \dfrac{\operatorname{cov}(X_s,Y_2)}{\sqrt{D(X_s)}\sqrt{D(Y_2)}} & \cdots & \dfrac{\operatorname{cov}(X_s,Y_s)}{\sqrt{D(X_s)}\sqrt{D(Y_s)}}\end{bmatrix}$$

为 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的相关系数矩阵，简称相关阵。

若 $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=0$，则称 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关。注意，由 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 相互独立可推得 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关；反之，由 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关不能推出 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 相互独立。

协方差阵具有以下性质：

(1) 对于常数向量 $\boldsymbol{a}$，$D(\boldsymbol{X}+\boldsymbol{a})=D(\boldsymbol{X})$；

(2) $D(\boldsymbol{AX})=\boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}'$；

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

(4) $E(\boldsymbol{X}'\boldsymbol{AX})=tr(\boldsymbol{A}D(\boldsymbol{X}))+E(\boldsymbol{X})'\boldsymbol{A}E(\boldsymbol{X})$。

其中，$tr(\boldsymbol{A}D(\boldsymbol{X}))$ 是矩阵 $\boldsymbol{A}D(\boldsymbol{X})$ 的迹，$\boldsymbol{X}$、$\boldsymbol{Y}$ 为随机向量，$\boldsymbol{A}$、$\boldsymbol{B}$ 为大小适合运算的常数矩阵。

类似于一维随机变量，多元随机向量也可定义一般的原点矩、中心矩和混合矩等，此处略。

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造成的影响，通常要先进行所谓的标准化变换：

$$\boldsymbol{X}_j^*=[\boldsymbol{X}_j-E(\boldsymbol{X}_j)]/[D(\boldsymbol{X}_j)]^{1/2},\boldsymbol{Y}_j^*=[\boldsymbol{Y}_j-E(\boldsymbol{Y}_j)]/[D(\boldsymbol{Y}_j)]^{1/2},j=1,2,\cdots,s$$

则

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

$$D(\boldsymbol{X}_j^*)=D(\boldsymbol{Y}_j^*)=1$$