21世纪经济管理精品教材·管理科学与工程系列

应用多元统计分析

党耀国　米传民　钱吴永　编　著

清华大学出版社

21世纪经济管理精品教材·管理科学与工程系列

应用多元统计分析

党耀国　米传民　钱吴永　编　著

清华大学出版社
北京

内 容 简 介

本书系统地介绍了多元统计分析中的经典理论和方法，重点讲解多元正态总体的参数估计和假设检验、聚类分析、判别分析、主成分分析、因子分析、对应分析及典型相关分析。力求以统计思想为主线，以 SPSS 软件为工具，深入浅出地介绍各种多元统计方法的理论和应用；以大量实际问题为背景，介绍多元统计分析的基本概念和方法，具有很强的实用性；在基本原理和方法的介绍方面，尽量避免复杂的理论证明，通过大量通俗易懂的例子进行理论方法的讲解，具有较强的趣味性，又不失理论性，理论难度由浅入深，适合不同层次的读者。

本书将 SPSS 软件的学习和案例分析有机结合，体现了多元统计分析方法的应用，并配备有多媒体教学课件，既可作为经济类、管理类等有关专业的高年级本科生或研究生教材，也适合自学多元统计分析的读者阅读参考。同时，也可作为市场研究、数据分析等领域实际工作者的多维数据分析参考书。

图书在版编目(CIP)数据

应用多元统计分析/党耀国，米传民，钱吴永编著. --北京：清华大学出版社，2012.5
(21 世纪经济管理精品教材·管理科学与工程系列)
ISBN 978-7-302-28356-0

Ⅰ. ①应… Ⅱ. ①党… ②米… ③钱… 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(2012)第 046865 号

责任编辑：贺　岩
封面设计：汉风唐韵
责任校对：王凤芝
责任印制：张雪娇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：12.25　　**字　　数**：276 千字
版　　次：2012 年 5 月第 1 版　　**印　　次**：2012 年 5 月第1 次印刷
印　　数：1～5000
定　　价：22.00 元

产品编号：039700-01

前言

多元统计分析是近几十年来迅速发展起来的一门学科。随着计算机的普及和软件的发展，信息储存手段以及数据信息的成倍增长，使得多元统计分析在自然科学和社会科学的各个领域得到了越来越广泛的应用，成为一种非常重要和实用的多元数据处理方法。在实践中我们还面临复杂数据的处理问题，特别是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，它的重要理论基础之一就是多元统计分析。多元统计分析是统计学中一个非常重要的分支，是一套非常有用的数据处理方法。

到目前为止，世界上已经出版的多元统计分析教材不下数百种，各有特色。实际上，对以解决实际问题为目标的经济管理类专业学生而言，最重要的是通过本门课程的学习，培养系统解决问题的能力，促成其运用多元统计分析的知识解决实际问题。为此，在教材编写过程中，力求以统计思想为主线，深入浅出地介绍各种多元统计方法，并以经济管理中的实际问题为基本素材，强调实践中的理念和悟性，通过大量实际案例的分析和讲解，以求加深读者对实际问题的认识，增强其学习兴趣；深入浅出地讲解多元统计分析的基本概念、基本方法和求解思路，尽力避开纯粹数学上的复杂推导，易于学生理解和自学；教材体系结构清晰，涵盖了多元统计分析的经典理论和方法，内容选择安排合理，简单实用。本书适用于高等院校经济管理类专业的本科生和研究生，以及面向实际应用的工程类、管理类和各类管理干部进修班的学员。

全书共分10章，其基本框架为：第1章为多元统计分析概述；第2章和第3章介绍多元正态总体的参数估计和假设检验；第4章～第9章介绍常用的多元统计方法，包括聚类分析、判别分析、主成分分析、因子分析、对应分析及典型相关分析；第10章介绍SPSS在多元统计分析中的应用。

本书的撰写分工如下：第1章～第6章由党耀国执笔，第7章、第8章由钱吴永执笔，第9章、第10章由米传民执笔。

本书在编写过程中参考了一些国内外相关文献资料，书后列出了主要参考文献。本书的出版得到了清华大学出版社和南京航空航天大学研究生院的大力支持，在此一并表示衷心感谢。由于作者水平有限，本书的缺点甚至错误在所难免，敬请专家、学者及读者不吝指正，以便今后进一步修改与完善。

党耀国

2012年4月

目录

第1章 多元统计分析概述

1.1 引　言

多元统计分析是运用数理统计的方法来研究解决多变量(多指标)问题的理论和方法,它是一元统计学的推广。

客观世界中任何事物的形成、变化和发展都受多种因素的影响,并且各种因素之间又存在着广泛而又错综复杂的联系。例如疾病的产生就受到多种因素的支配,各种病因之间也常存在着一定的内在联系和相互制约。要了解一个国家、省、市经济发展的类型需要观测很多指标,如人均国民收入、人均工农业产值、R&D 经费支出占 GDP 比重、万人科技活动人员数等;要衡量一个地区的经济发展水平,需要观测的指标有社会消费品零售总额、城镇居民人均可支配收入、农村居民人均纯收入、劳动生产率、万元产值能耗、财政收入等。对于这些指标,我们需要分析哪些指标是主要的、本质的,哪些指标是次要的、片面的,它们之间的相互关系等诸多问题。多元统计分析正是为了解决这些问题而产生的。

多元统计分析起源于 20 世纪初,1928 年 Wishart 发表了论文《多元正态总体样本协方差阵的精确分析》,可以说是多元统计分析的开端。随后多元统计分析得到了迅速发展,40 年代多元统计分析在心理、教育、生物等方面有不少应用,但由于计算量大,其发展受到一定的影响。50 年代中期,随着电子计算机的出现和发展,多元统计分析在地质、气象、医学、社会学等方面得到应用。60 年代通过应用和实践,新的理论和方法不断涌现,使它的应用范围更加扩大。70 年代初期,多元统计分析在我国才得到关注,并在理论研究和应用上取得了显著成绩,有些研究工作已达到了国际水平,并形成了一支科技队伍,活跃在各条战线上。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,产生了许多超大型数据库,这为多元统计分析与其他学科融合提供了重要的平台。

近几十年来,随着计算机应用技术的发展和科研生产的迫切需要,多元统计分析已被广泛地应用于工业、农业、医学、地质、气象、水文、环境以及经济、管理等诸多领域,成为解决实际问题的有效方法。多元统计学在广泛吸收和融合相关学科的新理论的基础上,不断开发应用新技术和新方法,深化和丰富了统计学传统领域的理论与方法研究,并拓展了统计学研究的新领域。具体表现在:

1. 统计学和计算机科学相互促进

在统计信息搜集、存储和传递过程中利用计算机提高工作效率,使统计信息时空结构有了新发展;在网络推断、统计软件包、统计建模中的计算机诊断等方面,提出了统计思想直接转化为计算机软件,通过软件对统计过程实行控制,以及利用计算机程序识别模型改善统计量性质的新方法。这些研究成果使人们认识到计算机技术正在促使统计研究发

生革命性的变化。在软件质量评估和统计程序及方法对软件可靠性的检验等方面也有了新的发展。

2. 统计理论与分析方法不断发展

近年来，统计方法研究成果累累，在贝叶斯方法、非线性时间序列、多元分析、统计计算、线性模型、极值统计、稳健统计、混沌理论和统计检验等方面取得了大量研究成果。而且不同方法之间相互渗透、交叉融合，衍生出许多新的分析方法。如马尔科夫链在贝叶斯似然计算中的应用，参数估计方法的非参数校正等。

3. 统计调查方法的创新

调查方法是统计学的重要组成部分，近年来，在抽样理论与方法、抽样调查、实验设计等方面进行了大量探索，对于如何改进调查技术、减少抽样误差进行了研究。在调查过程的综合管理、不等概论抽样设计、分层总体的样本分配、抽样比例的回归分析和实验设计正交数组的构造方法等方面也有新的突破。

1.2 多元统计分析的应用背景

通过阐述多元统计分析方法与研究目的、研究内容之间的关系，可以了解多元统计分析中每种方法能够解决的具体问题。它们之间的关系见表 1.1。

表 1.1 多元统计分析方法与研究内容之间的关系

问 题	内 容	方 法
数据或结构性简化	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解释所研究问题的现象	聚类分析、主成分分析、因子分析
分类和组合	基于研究问题，对测量到的一些现象特征，给出好的分组方法，对相似的对象或变量分组	聚类分析、判别分析、主成分分析、因子分析
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	典型相关分析、多元回归分析、主成分分析、因子分析
预测与决策	通过统计模型或最优准则，对未来进行预测或判断	多元回归分析
假设的提出与检验	检验多元总体参数的某种假设，并验证该假设的合理性	多元总体参数估计、假设检验

为了让读者从感性上加深对多元统计分析的认识，下面我们列举一些实际问题来说明多元统计分析的应用领域。

1. 经济学

(1) 在社会经济领域中存在着大量分类问题。如对我国 31 个省(市、自治区)城镇居民收支分布规律进行分析，一般不是逐个省(市、自治区)去分析，而较好的做法是选取能反映城镇居民收支分布规律的代表性指标，如城镇居民收入来源及支出指标(在收入方面，如工资性收入、财产性收入等；在支出方面，如食品、住房、生活用品、文化教育等)，根据这些指标对全国各省(市、自治区)城镇居民收支分布情况进行分类，然后根据分类结果对城镇居民收支状况进行综合评价。

(2) 研究国民收入(工农业国民收入、运输业国民收入等)与投资(生产建设投资、劳动者人数等)之间的相关关系。研究经济效益与资金、利税等主要财务指标之间的关系。这些可以使用相关分析,也可以利用典型相关分析法。

(3) 对我国 31 个省(市、自治区)经济效益进行综合评价,需要选择很多指标,如固定资产投资完成额、工业全员劳动生产率、工业销售利税率、万元工业产值能耗、职工工资总额等。要将这些有错综复杂关系的指标综合成几个较少的指标来分析和解释问题,又不至于使所研究的问题信息丢失过多,可利用主成分分析和因子分析方法。

(4) 研究国民收入的生产、分配与最终使用的关系。如研究我国财政收入与国民收入、工农业总产值、人口、就业、固定资产投资等因素的关系,可利用回归分析方法建立预测模型,对今后的财政收入进行预测。

2. 工业

(1) 如对我国 31 个省(市、自治区)独立核算工业企业经济效益进行分析时,选取能反映企业经济效益的代表性指标,如百元固定资产实现利税、资金利税、产值利税率等,根据这些指标对全国各省(市、自治区)进行分类,然后根据分类结果对企业经济效益进行综合评价,就易于得出科学的分析。

(2) 考察某产品质量指标(多个)与影响产品质量的因素(多个)之间的关系。在商品需求研究中,考察商品销售量与商品价格、消费者收入等之间的关系,可以利用回归分析方法建立数学模型进行分析。

(3) 研究某产品使用不同原料进行生产时,原料对产品质量有无显著影响,研究某商场今年与以前年份经营状况在经营指标方面有没有显著性的差异等,可以利用多元正态总体均值向量和协方差阵的假设检验进行分析。

3. 农业

(1) 某地区种植某种农作物,有多种种子在该地区播种,使用多种化肥,判断各种种子与化肥对该农作物产量的影响。

(2) 有 n 个地区,有 m 种农作物,每个地区可以种植多种农作物,每种农作物在不同的地区的产出不同,可以通过比较分析每个地区适合种植哪些农作物,以使生产效益最高。

4. 教育学

(1) 某高中对参加高考的考生成绩进行预测分析。根据以往大量的资料,分析考生高考成绩与高中学习期间成绩之间的相关关系,并通过考生在高中学习期间的成绩预测考生的综合成绩。

(2) 某大学对该校在校学生的学习成绩与该生高考的各门课程成绩之间的关系进行分析;还可以分析该校新生录取成绩次序的排队的最佳方案;还可以分析该校高考入学成绩的排队问题,可以按录取总成绩排队,也可以按其他方式进行排队。比如某工科院校,直接按总成绩进行排队并不是很合适,可以根据某些要求,对数学、物理、化学、英语等课程进行加权求和排队,有些课程权重可能大一些,有些可能小一些,需要研究它们之间的权重如何确定问题。

(3) 某高校根据 n 个学生在一学年的 m 门课程成绩,对学生学习成绩进行分类,以便

确定该校学生奖学金类别。

5. 医学

(1) 由于疾病的产生会受到多种因素的支配，各种病因之间也常存在着一定的内在联系和相互制约，这就需要分析哪些因素是主要的、本质的，哪些因素是次要的、片面的，它们之间的相互关系怎样等问题。

(2) 有了患胃炎的病人和健康人的一些化验指标，就可以从这些化验指标发现两类人的区别。把这种区别资料利用判别分析方法建立诊断的准则，然后对怀疑患胃炎的人就可以根据其化验指标用判别公式进行诊断。

(3) 可以根据病人的多种症状(体温、恶心、呕吐、腹部压疼感等)，来判断该病人患何种疾病。

6. 社会学

(1) 某公司对招聘人员的知识和能力进行测评，主要测评 6 个方面的内容：语言表达能力、逻辑思维能力、判断事物的敏捷和果断程度、思想修养、兴趣爱好、生活常识等，根据这 6 个方面的内容对招聘人员进行综合评价，决定是否录取。

(2) 某调查公司从一个大型零售公司随机调查了 n 人，测量了 5 个职业特性指标和 7 个职业满意变量。职业特性指标如用户反馈、任务重要性、任务多样性、任务特殊性和自主权，职业满意变量如主管满意度、事业前景满意度、财政满意度、工作强度满意度、公司地位满意度、工作满意度和总体满意度，讨论两组指标之间是否存在关联性。

7. 体育学

(1) 如何对影响运动员成绩的多项心理、生理测试指标(简单反映、时间知觉、综合反映等)进行主要因素分析。

(2) 研究运动员体能指标(反复横向跳、立定体前屈、俯卧上体后仰等)与运动能力测试指标(耐力跑、跳远、投球等)之间的相关关系。

8. 气象学

根据气象站资料，研究某地降雨量与前一天的气温、气压、湿度、风速、风向等之间的关系；有 n 个地区的降雨量、气温、湿度等指标，根据这些指标判断这 n 个地区所属的气候类型。

9. 其他

多元统计分析方法在其他很多领域也有广泛的应用，比如环境保护、地质学、考古学、地震预报、军事科学、生态学、文学、心理学等。

第2章 多元正态分布及其参数估计

在多元统计分析中，多元正态分布占有重要的地位，这是因为许多实际问题涉及的随机变量大都服从正态分布或近似服从正态分布。因此我们首先介绍多元正态分布的基本概念与性质。

2.1 基本概念

1. 随机向量及其概率分布

我们所讨论的是多个变量的总体，所研究的数据由多个指标构成，而且又是观测 n 次得到的，常常把它们看成一个整体进行研究。

定义 2.1 将 p 个随机变量 $X_1, X_2, \cdots, X_p$ 的整体称为 p 维随机向量，记为 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$。

在多元统计中，仍将所研究的对象称为总体。它是由许多个个体构成的集合。如果构成总体的个体是具有 p 个需要观测指标的个体，我们称这样的总体为 p 维总体。由于是从 p 维总体中随机抽取一个个体，而 p 个指标观测值依赖于被抽到的个体，因此 p 维总体可用一个 p 维随机向量来表示。若从 p 维总体中观测了 n 个个体，称每一个个体的 p 个变量组成为一个样品，而全体 n 个样品组成一个样本。常把 n 个样品排成一个 $n\times p$ 矩阵，记为

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

多维随机向量的统计特性可用它的分布函数来完整地描述。

定义 2.2 设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 是 p 维随机向量，称

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

为 $\boldsymbol{X}$ 的联合分布函数，简称为分布函数，记为 $\boldsymbol{X}\sim F(\boldsymbol{x})$，其中 $\boldsymbol{x}=(x_1, x_2, \cdots, x_p)^{\mathrm{T}}\in \boldsymbol{R}^p$，$\boldsymbol{R}^p$ 表示 p 维欧氏空间。

定义 2.3 设 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 是 p 维随机向量，若存在有限个或可列个 p 维向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$，记为 $P(\boldsymbol{X}=\boldsymbol{x}_k)=p_k(k=1,2,\cdots)$，且满足 $p_1+p_2+\cdots=1$，则称 $\boldsymbol{X}$ 为离散型随机向量，称 $P(\boldsymbol{X}=\boldsymbol{x}_k)=p_k(k=1,2,\cdots)$ 为 $\boldsymbol{X}$ 的概率分布。

若存在一个非负函数 $f(x_1, x_2, \cdots, x_p)$，使得对于一切 $\boldsymbol{x}=(x_1, x_2, \cdots, x_p)^{\mathrm{T}}\in \mathbf{R}^p$ 有

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

则称 $\boldsymbol{X}$ 为连续型随机向量，称 $f(x_1, x_2, \cdots, x_p)$ 为 $\boldsymbol{X}$ 的联合分布密度函数，简称密度函数

或分布密度。

一个 p 元函数 $f(x_1,x_2,\cdots,x_p)$ 能作为 $\boldsymbol{R}^p$ 中某个随机向量的密度函数，必须满足以下条件：

(1) $f(x_1,x_2,\cdots,x_p)\geqslant 0$，$\forall (x_1,x_2,\cdots,x_p)^{\mathrm{T}}\in \boldsymbol{R}^p$

(2) $\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(t_1,t_2,\cdots,t_p)\mathrm{d}t_1\mathrm{d}t_2\cdots\mathrm{d}t_p=1$

离散型随机向量的统计性质可由它的概率分布完全确定，连续型随机向量的统计性质可由它的分布密度完全确定。

例 2.1 试证函数

$$f(x_1,x_2)=\begin{cases}\mathrm{e}^{-(x_1+x_2)}, & x_1\geqslant 0,\quad x_2\geqslant 0\\ 0, & \text{其他}\end{cases}$$

为随机向量 $\boldsymbol{X}=(x_1,x_2)^{\mathrm{T}}$ 的密度函数。

证明：只要证明它满足密度函数的两个条件即可。

(1) 显然，当 $x_1\geqslant 0, x_2\geqslant 0$ 时，有 $f(x_1,x_2)\geqslant 0$

(2) $$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2=\int_0^{+\infty}\int_0^{+\infty}\mathrm{e}^{-(x_1+x_2)}\mathrm{d}x_1\mathrm{d}x_2=\int_0^{+\infty}\left[\int_0^{+\infty}\mathrm{e}^{-(x_1+x_2)}\mathrm{d}x_1\right]\mathrm{d}x_2$$
$$=\int_0^{+\infty}\mathrm{e}^{-x_2}\mathrm{d}x_2=1$$

定义 2.4 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 是 p 维随机向量，它的分布函数为 $F(x_1,x_2,\cdots,x_p)$，则 $F_{(X_1,X_2,\cdots,X_r)}(x_1,x_2,\cdots,x_r)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_r\leqslant x_r)=F(x_1,x_2,\cdots,x_r,+\infty,+\infty,\cdots,+\infty)$

称为随机向量 $(X_1,X_2,\cdots,X_r)^{\mathrm{T}}$ 的边缘分布函数；

$$f(x_1,x_2,\cdots,x_r)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(t_1,t_2,\cdots,t_p)\mathrm{d}t_{r+1}\mathrm{d}t_{r+2}\cdots\mathrm{d}t_p$$

称为随机向量 $(X_1,X_2,\cdots,X_r)^{\mathrm{T}}$ 的边缘分布密度函数，简称边缘密度函数。

例 2.2 对例 2.1 中的 $X=(x_1,x_2)^{\mathrm{T}}$，求其边缘密度函数。

解：$$f_1(x_1)=\int_0^{+\infty}\mathrm{e}^{-(x_1+x_2)}\mathrm{d}x_2=\begin{cases}\int_0^{+\infty}\mathrm{e}^{-(x_1+x_2)}\mathrm{d}x_2=\mathrm{e}^{-x_1}, & x_1\geqslant 0\\ 0, & \text{其他}\end{cases}$$

同理可得

$$f_2(x_2)=\begin{cases}\mathrm{e}^{-x_2}, & x_2\geqslant 0\\ 0, & \text{其他}\end{cases}$$

定义 2.5 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 是 p 维随机向量，若 p 个随机变量 $X_1,X_2,\cdots,X_p$ 的联合分布函数 $F(x_1,x_2,\cdots,x_p)$ 等于各自的边缘分布函数的乘积，则称随机变量 $X_1,X_2,\cdots,X_p$ 是相互独立的。即

$$F(x_1,x_2,\cdots,x_p)=\prod_{i=1}^{p}F_{x_i}(x_i)$$

若 p 个随机变量 $X_1,X_2,\cdots,X_p$ 的联合密度函数 $f(x_1,x_2,\cdots,x_p)$ 等于各自的边缘密度函数的乘积，则称随机变量 $X_1,X_2,\cdots,X_p$ 是相互独立的。即

$$f(x_1,x_2,\cdots,x_p)=\prod_{i=1}^{p}f_{x_i}(x_i)$$

需要注意的是：若 $X_1,X_2,\cdots,X_p$ 相互独立，可以推知 X_i 与 X_j 独立($i\neq j$)，但反之不成立。

例 2.3 试问例 2.2 中的 X_1 与 X_2 是否相互独立？

解：由于 $f(x_1,x_2)=\begin{cases}e^{-(x_1+x_2)}, & x_1\geqslant 0,x_2\geqslant 0\\ 0, & \text{其他}\end{cases}$

$$f_1(x_1)=\begin{cases}e^{-x_1}, & x_1\geqslant 0\\ 0, & \text{其他}\end{cases}$$

$$f_2(x_2)=\begin{cases}e^{-x_2}, & x_2\geqslant 0\\ 0, & \text{其他}\end{cases}$$

所以 $$f(x_1,x_2)=f_1(x_1)f_2(x_2)$$

故 X_1 与 X_2 相互独立。

2. 随机向量的数字特征

定义 2.6 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 是 p 维随机向量，若 $E(X_i)(i=1,2,\cdots,p)$存在且有限，则称 $E(\boldsymbol{X})=(E(X_1),E(X_2),\cdots,E(X_p))$为 $\boldsymbol{X}$ 的数学期望，通常把 $E(\boldsymbol{X})$和 $E(X_i)$分别记为$\boldsymbol{\mu}$ 和μ_i，即

$$E(\boldsymbol{X})=\boldsymbol{\mu}=(\mu_1,\mu_2,\cdots,\mu_p)$$

随机向量的数学期望具有下列性质：

(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 $\boldsymbol{X},\boldsymbol{Y}$ 为随机向量，$\boldsymbol{A},\boldsymbol{B}$ 为大小适合运算的常数矩阵。

定义 2.7 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 具有数学期望，则称

$$D(\boldsymbol{X})=E(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^{\mathrm{T}}$$

$$=\begin{bmatrix}\operatorname{cov}(X_1,X_1) & \operatorname{cov}(X_1,X_2) & \cdots & \operatorname{cov}(X_1,X_p)\\ \operatorname{cov}(X_2,X_1) & \operatorname{cov}(X_2,X_2) & \cdots & \operatorname{cov}(X_2,X_p)\\ \vdots & \vdots & \ddots & \vdots\\ \operatorname{cov}(X_p,X_1) & \operatorname{cov}(X_p,X_2) & \cdots & \operatorname{cov}(X_p,X_p)\end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 的协方差阵。

$D(\boldsymbol{X})$简记为$\boldsymbol{\Sigma}$，$\operatorname{cov}(X_i,X_j)$简记为 σ_{ij}^2。从而有

$$\boldsymbol{\Sigma}=(\sigma_{ij}^2)_{p\times p}$$

$\boldsymbol{\Sigma}$ 为对称矩阵。当 $p=1$ 时，$\boldsymbol{\Sigma}$ 就是一元统计分析中的方差。

定义 2.8 设随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$，$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_p)^{\mathrm{T}}$，则称

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=E(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{Y}-E(\boldsymbol{Y}))^{\mathrm{T}}$$

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

为随机向量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的协方差阵。

两个随机向量的协方差阵一般不是对称的。

当 $\boldsymbol{X}=\boldsymbol{Y}$ 时，则有

$$\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y}) = E(\boldsymbol{X}-E\boldsymbol{X})(\boldsymbol{Y}-E\boldsymbol{Y})^{\mathrm{T}} = D(\boldsymbol{X}) = \boldsymbol{\Sigma}$$

若 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的协方差阵存在，且每个分量的方差大于零，则称

$$\rho_{ij} = \frac{\sigma_{ij}^2}{\sqrt{\sigma_{ii}^2}\sqrt{\sigma_{jj}^2}} = \frac{\operatorname{cov}(X_i,X_j)}{\sqrt{\operatorname{var}(X_i)}\sqrt{\operatorname{var}(X_j)}}$$

为 X_i 与 X_j 的相关系数。

由相关系数 ρ_{ij} 组成的矩阵

$$\boldsymbol{R} = (\rho_{ij})_{p\times p} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix}$$

称为随机向量 $\boldsymbol{X}$ 的相关阵。

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带来的影响，往往在使用各种统计分析之前，常常将每个指标“标准化”，即进行如下变换：

$$X_j^* = \frac{X_j - E(X_j)}{\sqrt{D(X_j)}}, \quad j = 1,2,\cdots,p$$

由上式构成的随机向量记为

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

设

$$\boldsymbol{C} = \begin{bmatrix} \sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{pp} \end{bmatrix}$$

有

$$X_j^* = \boldsymbol{C}^{-1}(\boldsymbol{X}-E(\boldsymbol{X}))$$

$$E(X_j^*) = E[\boldsymbol{C}^{-1}(\boldsymbol{X}-E(\boldsymbol{X}))] = \boldsymbol{C}^{-1}E[(\boldsymbol{X}-E(\boldsymbol{X}))] = 0$$

$$D(X_j^*) = D[\boldsymbol{C}^{-1}(\boldsymbol{X}-E(\boldsymbol{X}))] = \boldsymbol{C}^{-1}D[(\boldsymbol{X}-E(\boldsymbol{X}))]\boldsymbol{C}^{-1} = \boldsymbol{C}^{-1}D(\boldsymbol{X})\boldsymbol{C}^{-1} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}\boldsymbol{C}^{-1} = \boldsymbol{R}$$

则有

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{R}\boldsymbol{C}$$

这说明由 $\boldsymbol{\Sigma},\boldsymbol{C}$ 可以得到 $\boldsymbol{R}$，也可以由 $\boldsymbol{C},\boldsymbol{R}$ 得到 $\boldsymbol{\Sigma}$，并且由于 $\boldsymbol{\Sigma}\geqslant 0$，可知 $\boldsymbol{R}\geqslant 0$。

上式还说明标准化数据的协方差阵正好是原指标的相关矩阵。

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

由 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 相互独立，可推知 $\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})=0$，即 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 不相关，但反过来当 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 不相关时，一般不能推得它们相互独立。

协方差阵具有以下性质：

(1) $D(\boldsymbol{X})\geqslant 0$，即 $\boldsymbol{X}$ 的协方差阵是对称非负定阵；

(2) $D(\boldsymbol{X}+\boldsymbol{a})=D(\boldsymbol{X})$，对任意的常数向量 $\boldsymbol{a}$；

(3) 设 $\boldsymbol{A},\boldsymbol{B}$ 为常数矩阵，则 $\mathrm{cov}(\boldsymbol{AX},\boldsymbol{BY})=\boldsymbol{A}\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}^{\mathrm{T}}$；

(4) 设 $\boldsymbol{A}$ 为常数矩阵，则 $D(\boldsymbol{AX})=\boldsymbol{A}D(\boldsymbol{X})\boldsymbol{A}^{\mathrm{T}}$。

其中 $\boldsymbol{a},\boldsymbol{A},\boldsymbol{B}$ 为大小适合运算的常数向量和矩阵。

2.2 多元正态分布

1. 多元正态分布的定义

多元正态分布在多元统计分析中的重要地位，与一元统计分析中一元正态分布所占的地位一样。多元统计分析中的许多重要理论与方法都是直接和间接建立在正态分布的基础上，多元正态分布是多元统计分析的基础。此外在实用中遇到的随机向量常常是服从正态分布或近似服从正态分布，因此现实世界中许多实际问题的解决办法都是以总体服从正态分布或近似服从正态分布为前提的。

定义 2.9 若 p 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的密度函数为

$$f(x_1,x_2,\cdots,x_p)=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

其中 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$，$\boldsymbol{\mu}$ 是 p 维向量，$\boldsymbol{\Sigma}$ 是 p 阶正定矩阵，则称 $\boldsymbol{X}$ 服从 p 元正态分布，简记为 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{\mu}$ 是随机向量 $\boldsymbol{X}$ 的数学期望，$\boldsymbol{\Sigma}$ 是随机向量 $\boldsymbol{X}$ 的协方差阵。

当 $|\boldsymbol{\Sigma}|=0$ 时，$\boldsymbol{\Sigma}^{-1}$ 不存在，随机向量 $\boldsymbol{X}$ 也就不存在通常意义下的密度函数，然而可以形式地给出一个表达式，使得有些问题可以利用这一形式对 $|\boldsymbol{\Sigma}|\neq 0$ 及 $|\boldsymbol{\Sigma}|=0$ 的情况给出一个统一的处理。在此不再详述，有兴趣的读者可参考相应的参考书。

当 $p=2$ 时，设 $\boldsymbol{X}=(X_1,X_2)$ 服从二元正态分布，则

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_2\sigma_1\rho & \sigma_2^2\end{bmatrix},\quad \rho\neq\pm 1$$

其中 σ_1^2,σ_2^2 分别是 X_1,X_2 的方差，ρ 是 X_1,X_2 的相关系数。即有

$$|\boldsymbol{\Sigma}|=\sigma_1^2\sigma_2^2(1-\rho^2)$$

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix}\sigma_2^2 & -\sigma_1\sigma_2\rho\\ -\sigma_2\sigma_1\rho & \sigma_1^2\end{bmatrix}$$

故 X_1,X_2 的密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

对于 $\rho=0$，那么 X_1 与 X_2 是相互独立的；

$\rho>0$，则 X_1 与 X_2 正相关，$\rho<0$，则 X_1 与 X_2 负相关。

定理 2.1 设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，则有

$$E(\boldsymbol{X})=\boldsymbol{\mu},\quad D(\boldsymbol{X})=\boldsymbol{\Sigma}$$

这里需要说明的是，多元正态分布的定义有多种，可以采用特征函数来定义，也可以采用线性组合的方式来定义。

2. 多元正态分布的基本性质

在讨论多元统计分析的理论和方法时，经常会用到多元正态变量的某些性质，利用这些性质可使得正态分布的处理变得更容易一些。

(1) 若 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{\Sigma}$ 是对角阵，则 $X_1,X_2,\cdots,X_p$ 相互独立。

(2) 若 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，$\boldsymbol{A}$ 是 $s\times p$ 阶常数矩阵，$\boldsymbol{d}$ 为 s 维常数向量，则

$$\boldsymbol{AX}+\boldsymbol{d}\sim N_s(\boldsymbol{A\mu}+\boldsymbol{d},\boldsymbol{A\Sigma A}^{\mathrm{T}})$$

即正态随机向量的任意线性组合仍然服从正态分布。

(3) 若 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$，将 $\boldsymbol{X},\boldsymbol{\mu},\boldsymbol{\Sigma}$ 作如下划分：

$$\boldsymbol{X}=\begin{bmatrix}Y_1\\Y_2\end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\quad \boldsymbol{\Sigma}=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}$$

则

$$Y_1\sim N_q(\mu_1,\Sigma_{11}),Y_2\sim N_{p-q}(\mu_2,\Sigma_{22})$$

注意：

(1) 多元正态分布的边缘分布仍为正态分布，但反之不成立。

(2) 由于 $\Sigma_{12}=\mathrm{cov}(X_1,X_2)$，故 $\Sigma_{12}=0$ 表示 X_1 与 X_2 不相关。对于正态分布而言，X_1 与 X_2 不相关与它们独立是等价的。

顺便指出，多元分析中的很多统计方法，大都假定数据来自多元正态总体。但是要判断已有的一批数据是否来自多元正态总体，并不是一件容易的事情。但是反过来要肯定数据不是来自于正态总体，倒是有一些简易的方法，其依据是：如果 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 服从 p 元正态分布，则每个分量必须服从一元正态分布，因此把某个分量的 n 个样本值作成直方图，如果断定其不呈正态分布，就可以断定随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 也不可能服从 p 元正态分布。

2.3 多元正态分布的参数估计

在实际应用中，多元正态分布中的均值向量$\boldsymbol{\mu}$与协方差阵$\boldsymbol{\Sigma}$通常是未知的，需要由样本来估计，参数估计的方法很多，这里只介绍最常用的最大似然估计法。

1. 多元样本的概念

从多元总体中随机抽取 n 个个体 $X_{(1)},X_{(2)},\cdots,X_{(n)}$，若 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 相互独立且与总体 $\boldsymbol{X}$ 同分布，则称 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为该总体的一个多元随机样本，简称为简单样本。样本中的每一个个体称为样品，样本中含有的样品的个数称为样本容量。

多元总体的样本观测数据常是观测 n 个样品的 p 个变量的取值，因此多元总体的样本观测数据用一个 $n\times p$ 阶矩阵 $\boldsymbol{X}$ 表示：

$$\boldsymbol{X}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X}_{(1)}^{\mathrm{T}}\\\boldsymbol{X}_{(2)}^{\mathrm{T}}\\\vdots\\\boldsymbol{X}_{(n)}^{\mathrm{T}}\end{bmatrix}$$