全国统计教材编审委员会“十一五”规划教材

应用多元统计分析

★ 何晓群 编著

中国统计出版社
China Statistics Press

全国统计教材编审委员会“十一五”规划教材

应用多元统计分析

★ 何晓群 编著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

应用多元统计分析/何晓群编著. --北京:中国统计出版社,2010.6

ISBN 978-7-5037-5953-6

Ⅰ.①应… Ⅱ.①何… 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(2010)第 103861 号

应用多元统计分析

作　　者/何晓群
责任编辑/张　赏
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/18
字　　数/330 千字
印　　张/24
印　　数/1—3000 册
版　　别/2010 年 6 月第 1 版
版　　次/2010 年 6 月第 1 次印刷
书　　号/ISBN 978-7-5037-5953-6/O·74
定　　价/42.00 元

出版说明

“十一五”时期是继续深化教育改革，加强素质教育，努力建设有利于创新型科技人才成长的教育体系的关键时期。为了更好地培育统计创新型科技人才，适应统计教育发展的新形势，全国统计教材编审委员会制定了《“十一五”全国统计教材建设规划》（以下简称规划）。规划坚持“以人为本”的科学发展观，坚持统计教育与实践相结合，坚持统计教育同国际接轨，坚持培养创新型的统计人才的指导思想，编写符合国民经济发展需要和统计事业发展需要的统计教材。

这批教材是在深入分析统计教育形势和统计教材建设发展状况，总结多年来统计教材建设经验的基础上，本着以建设本科统计教材为主的方针，积极探索研究生层次的统计教材，力争使规划统计教材的编写做到层次分明，有针对性和实用性。建设精品教材，是编委会自成立以来就孜孜以求的目标。考虑到统计教材建设的实际情况，“十一五”期间，本科教材主要以修订为主，对以往规划统计教材中使用面广，得到广大教师和学生普遍认可的教材组织了修订。修订后的教材，淘汰了过时的内容和例子，增加了计算机操作和大量的案例，编写手法也做了一定的调整，在实用性、可操作性等方面有了较大的改进。

近年来，我国现代化建设快速发展，高等教育规模持续扩大，尤其是研究生教育规模的扩大，使得高等学校研究生统计教学工作面临着许多新情况、新问题，任务艰巨。因此，必须坚持科学发展观，在规模持续发展的同时，把提

高研究生统计教学质量放在突出的位置，培养全面发展的创新型的统计人才。教材是统计教学的载体，建设高质量的研究生层次的统计教材是统计教育发展的需要。因此，编委会在“十一五”期间对研究生的统计基础课教材做了些有益的探索。根据《规划》的要求，这批教材主要采取招标和邀请的方式组织有关院校的专家、学者编写。

值得特别提出的是，在这批教材中，有《非参数统计》、《概率论与数理统计》、《经济计量学教程》、《医学统计》、《应用时间序列分析》、《多元统计分析》、《统计学》、《统计指数理论及应用》、《现代金融投资统计分析》9种教材入选国家教育部组织编写的“普通高等教育‘十一五’国家级规划教材”，更加充实和完善了“十一五”期间统计教材的建设。

为了便于教学和学习，这批教材里面包含了与之相配套的《学习指导与习题》，使得这批教材在编辑出版上形成了比较完整的体系。我们相信，这批教材的出版和发行，对于推动我国统计教育改革，加快我国统计教材体系和教材内容更新、改造的步伐，打造精品教材，都将起到积极的促进作用。

限于水平和经验，这批教材的编审、出版工作还会有缺点和不足，诚恳欢迎教材的使用单位、广大教师和同学们提出批评和建议。

全国统计教材编审委员会

2006年6月

内容提要

该书假定学生已具有线性代数、概率论与数理统计的基础知识，本着提高人文社会科学、财经管理类研究生量化分析能力的宗旨，在不失理论严密性的前提下，力求将多元统计分析主流方法的背景、思想、具体的步骤、分析的技巧讲清楚。为重点突出方法的思想和应用，每种方法尽可能结合中国社会、经济、管理方面的实际问题，以案例研究为导向，为研究生进行量化分析起一定示范作用。

本书也可作为应用统计硕士专业学位和统计学专业本科生多元统计分析课程的教材。此书还可作为从事社会、经济、管理等研究和实际工作的同志进行量化研究的参考书。

序

进入新世纪以来，现代统计分析方法在我国的应用方兴未艾，尤其令人欣喜的是我国的人文社会科学、财经管理类研究越来越多地运用多变量统计分析的定量方法。

作为人文社会科学、财经管理类研究生学习一些现代统计分析方法，掌握定性与定量有机结合的研究技能是十分必要的。

何晓群编著的《应用多元统计分析》一书为非统计专业的人文社会科学、财经管理类研究生学习现代统计分析方法提供了一本较好的教材。

该书在众多统计方法中选择了一些主流的实用多元统计分析方法，假定学生已具备一些基础数理统计知识，在不失理论严谨性的前提下，略去了令非数学专业学生头疼的许多证明。该书的显著特点是除个别典型案例外，尽可能结合中国社会、经济、管理方面的实际问题，以案例研究为导向，主要运用 SPSS 软件来实现计算，力求将问题的背景、方法的思想、具体的步骤、分析的技巧讲清楚。为非统计专业研究生进行定性与定量结合分析起了一定的示范作用。

本书也可作为统计学专业本科生多元统计分析课程的教材，还可作为从事社会、经济、管理等研究和实际工作的同志进行量化研究的参考书。

相信该书为推动现代统计分析方法在我国的深入应用一定会起到积极作用。

方开泰

2010 年 3 月 28 日于珠海

前 言

面对21世纪，深刻的社会变革、迅猛的经济发展，使我国的人文社会科学、财经管理类研究生面临严峻的挑战和难得的机遇。时代呼唤我们精通定量分析的研究方法，掌握定性与定量有机结合的研究技能。《应用多元统计分析》一书正是适应这一需要，为应用统计硕士专业学位和非统计专业的人文社会科学、财经管理类研究生学习现代统计分析方法而编著的。

统计理论与方法是现代社会、经济、管理类研究运用的基本方法。自1969年设立诺贝尔经济学奖以来，已有50多位学者获奖。这些获奖者大都精通现代统计方法，对统计方法的运用极为娴熟，在社会、经济研究中取得了举世瞩目的成就。学习和运用统计方法已成为时代对我们的要求。

作者假定学生已具有线性代数、概率论与数理统计的基础知识，本着提高研究生量化分析能力的宗旨。在不失理论严密性的前提下，力求将问题的背景、方法的思想、具体的步骤、分析的技巧讲清楚。为重点突出方法的思想和应用，每种方法尽可能结合中国社会、经济、管理方面的实际问题，以案例研究为导向，主要运用SPSS软件来实现计算，为非统计专业研究生进行量化分析起一定示范作用。为了节省篇幅，本书的例题和习题数据大都放在中国人民大学六西格玛质量管理研究中心的网站，需要的读者请点击 www.ruc-6sigma.com 或 www.stats.gov.cn/tjshujia/zjxs/t20100613_402650165.htm 即可获得。

1996年，中国人民大学率先在非统计专业的人文社会

科学、财经管理类研究生中开设“统计方法与技术”必修课，作者有幸从1996年以来给中国人民大学的多级研究生和MBA主讲此课。在教学实践中，学生们给了我许多启发和鼓励，因为他们结合自己的专业，对统计方法的学习产生了浓厚的兴趣，看到了统计方法的用武之地，清楚哪些方法最有用；他们在学习的过程中也渴望拥有一本合适的教材。

本书的大部分内容都给非统计专业研究生讲授过，根据笔者的经验，如有计算机配合，学生掌握这些基本方法和技能并不困难。选用本书的教师可有一定的灵活性，根据不同专业有选择地讲授该书内容。本书参考教学课时为54学时。

本书也可作为统计学专业本科生多元统计分析课程的教材。此书还可作为从事社会、经济、管理等研究和实际工作的同志进行量化研究的参考书。

本书在写作过程中，我的导师香港浸会大学数学系讲座教授方开泰先生对本书的写作给予许多悉心指点。我的博士生付韶军、耿贵珍、李因果、王惠惠等为本书的部分案例做过一些计算验证和补充。中国统计出版社的总编辑刘科和教材编辑部主任陈悟朝博士对本书编写做过精心策划和一些具体建议。在此，我谨向对本书出版给予支持的师长和朋友表示衷心的感谢。

由于本人学识有限，书中谬误之处在所难免，恳请读者批评指正。

中国人民大学应用统计科学研究中心
中国人民大学六西格玛质量管理研究中心
何晓群
庚寅年三月初六于长安

目录

目录

目录

第 1 章

统计学基础回顾

《统计学》是经济、管理类专业本科阶段的必修课程，其中有些概念是学习应用多元统计分析的重要基础。为了更顺利地学习该课程的内容，本章将对统计学中的一些基本概念和术语作一简要回顾。

§1.1 统计数据的整理与描述

统计学是研究实际问题变量数据规律性的方法论学科，统计数据是统计学研究的主要内容。借助统计学方法研究任何实际问题，首先要做的工作就是收集数据，收集数据是一项很重要的基础工作。收集数据的一般方法是查阅各种统计年鉴和报表，再就是运用某种调查方法获取欲研究问题的有关数据。抽样调查获取数据的方式在我国方兴未艾，抽样调查的方法很多，有一定的专业性，需要利用抽样方法获取数据的学者，还需很好地学习有关抽样技术的课程。

一、总体与样本

在一个统计问题中，通常把所要调查研究的事物或现象的全体称为总体，而把组成总体的每个元素（成员）称为个体，一个总体中所含的个体的数量称为总体的容量。例如要研究某城市居民的家庭收入状况，那么这个城市所有家庭的收入状况就是我们研究的总体，而每个家庭的收入状况就是个体。

为了推断总体的某些特征，需要从总体中按一定的抽样技术抽取若干个体，将这一抽取过程称为抽样。所抽取的部分个体称为样本，样本中所含个体的数量称为样本容量。如在研究居民家庭收入时，随机抽取 3000 户来进行调查，这 3000 户就是一个样本，样本容量就是 3000。

二、统计量

通过抽样或查统计年鉴得到的原始数据，一般是杂乱无章的，很难从中直接看出有价值的东西。因此，对获取的原始数据一般需要加以整理，以便把我们感兴趣的信息提取出来，并用简明醒目的方式加以表述。画原始数据的散点图、饼图、直方图等方法是直观表达数据的常见方式。统计学中最主要的提取信息方式就是对原始数据进行一定的运算，以算出某些代表性的数字，足以反映出数据某些方面的特征，这种数字被称为统计量。用统计学语言表述就是：统计量是样本的函数，它不依赖于任何未知参数。

例如均值和方差就是最重要的常用统计量。

均值是对数据集中特征的描述，方差是对数据波动特征的描述。

设 $x_1, x_2, \cdots, x_n$ 是一组独立的随机样本，则样本均值为：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

样本方差为：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

样本标准差为：

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

例如有两组数据：(4,6,8,10,12)

(6,7,8,9,10)

他们的均值 $\bar{x}$ 都是 8，这说明两组数据都以 8 为中心。读者可计算他们的方差，第一组数据的方差比第二组的要大，说明第一组数据相对均值 8 来说比较分散，第二组数据相对均值 8 来说比较集中。这两组数据可很直观地看出均值及方差的意义。

需要注意的是，方差带单位是没有意义的，只有标准差带上单位才有实际意义。

三、变异系数

如果两组数据的计量单位相同，且均值一样，可以利用标准差来比较两组数据的离散程度。但当两组数据的计量单位不同或均值不同时，就不能直接比

较两组数据的标准差来分析两组数据的离散程度。由此引入变异系数 V，

$$V=\frac{S}{\bar{x}}$$

例如下面两组数据(4,5,6,7,8)与(40,50,60,70,80)的标准差分别是1.58和15.8,如果仅从标准差来看显然第二组数据的分散程度大的多。但是由于两组数据的均值不同,分别为6和60,单纯由标准差来判断数据的分散程度就不合适。实际上,当我们算出两组数据的变异系数时,得到V都是0.26。比较而言,两组数据的分散程度就是相同的了。

四、偏度与峭度

偏度和峭度是描述统计数据分布偏斜和陡峭程度的统计量。

偏度用偏度系数 V_1 来描述：

$$V_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{S^3(n-1)}$$

其中 S 为样本标准差。

偏度系数 V_1 的意义由图 1－1 可表示出来。

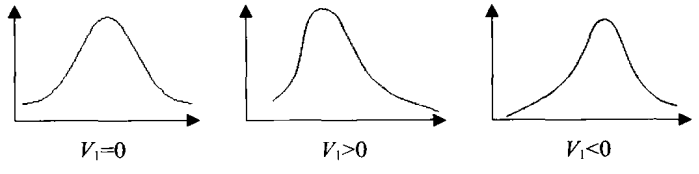

图 1－1

峭度用峭度系数 V_2 表示：

$$V_2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{S^4(n-1)}$$

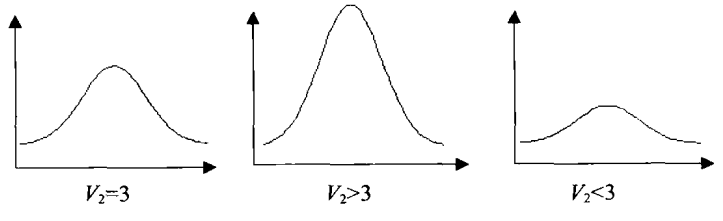

图 1－2

当峭度系数 $V_2=3$ 时，一般为标准正态分布。

五、累积频数分布

在社会经济调查中，经常得到的数据是频数。例如家庭月收入按等级划分时，我们就会得到每个等级的家庭数，常常将这些数据列在表中或画成直方图。

读者可依收入等级从低到高画出累积频数的直方图。

表 1.1　累积频数分布表

收入等级(元)	家庭数	
	频数	累积频数
5000～6000	800	800
6001～7000	700	1500
7001～8000	500	2000
8001～9000	300	2300

在社会经济研究中，洛伦茨(M. E. Lorentz)曲线是累积频数的典型应用。如果按收入从低到高排列，各收入等级的家庭的累积数(百分比)为横坐标，与之相对应的收入的累计(百分比)为纵坐标，所得到的曲线就是西方经济学中著名的洛伦茨曲线。在宏观经济的收入差距研究中，就可运用这一描述方法。

图 1－3 中对角线 OA 是均匀收入分布线。图中 B 点表明在数量上占全体 40%的家庭在收入上也占 40%。收入分布不大可能绝对平均，所以洛伦茨曲线一般并不是一条直线。图中 C 点表示从最低收入开始的 40%的家庭收入的合计还占不到总收入的 30%。

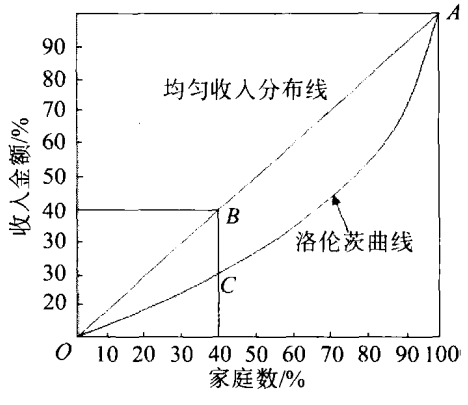

图 1－3

关于累积频数的百分比曲线可拓宽到衡量贫富差距的基尼(Gini)系数。基尼系数理论在中国当今的宏观经济研究中非常有用。

§1.2　几种重要的概率分布

一、正态分布

在经济研究和工商管理中，有许多随机变量的概率分布都可用正态分布来描述。例如一个城市居民的家庭收入、消费支出，某种股票月收益的百分比，某种商品的年销售等都可近似用正态分布来描述。在实际问题的研究中，可以通过该随机变量的抽样数据的频数直方图与正态概率分布的钟形曲线相比较，来判断该随机变量是否为正态随机变量。

正态随机变量 X 的概率密度函数的形式如下：

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty<x<\infty$$

其中，μ 为随机变量 X 的均值，σ^2 为随机变量 X 的方差。

通常对具有均值为 μ，方差为 σ^2 的正态概率分布，记为 $N(\mu,\sigma^2)$。于是有正态随机变量 $X\sim N(\mu,\sigma^2)$。

一般来说，正态分布的密度曲线是以 μ 为中心，在 μ 的两侧呈对称的形状，曲线的形状像一个钟的剖面，故称为钟形曲线。σ 越大，密度曲线的峰度越低；σ 越小，密度曲线的峰度越高。无论参数 μ 和 σ 取何值，密度曲线下所覆盖的面积均等于 1。正态分布的密度曲线见图 1－4。

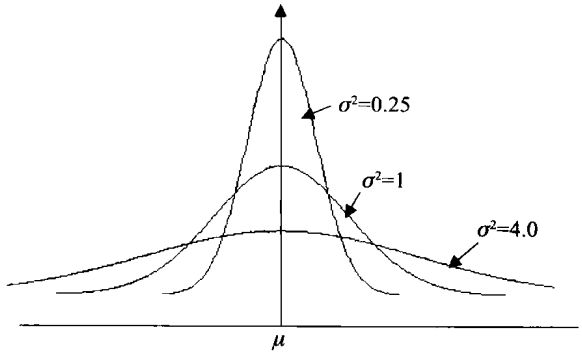

图 1－4