全国统计教材编审委员会“十二五”规划教材

应用多元统计分析

第二版

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“十二五”规划教材

应用多元统计分析

第二版

何晓群 编著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

应用多元统计分析 / 何晓群编著. —— 2 版. —— 北京：中国统计出版社，2015.8

ISBN 978－7－5037－7449－2

Ⅰ. ①应… Ⅱ. ①何… Ⅲ. ①多元分析－统计分析－研究生－教材 Ⅳ. ①O212.4

中国版本图书馆 CIP 数据核字(2015)第 131788 号

应用多元统计分析(第二版)

作　　者/何晓群
责任编辑/张　赏
封面设计/上智博文
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编号/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://www.zgtjcbs.com
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/787×1092mm　1/16
字　　数/470 千字
印　　张/19.5
版　　别/2015 年 8 月第 2 版
版　　次/2015 年 8 月第 1 次印刷
定　　价/41.00 元

出版说明

全国统计教材编审委员会是国家统计局领导下的、全国统计教材建设工作的最高指导机构和咨询机构，自1988年成立以来，分别组织编写和出版了“七五”至“十一五”全国统计规划教材。

“十二五”时期，是我国全面实施素质教育，全面提高高等教育质量，深化教育体制改革，推动教育事业科学发展，提高教育现代化水平的时期。“十二五”伊始，统计学迎来了历史性的重大变革和飞跃。2011年2月，在国务院学位委员会第28次会议通过的新的《学位授予和人才培养学科目录(2011)》(以下简称“学科目录”)中，统计学从数学和经济学中独立出来，成为一级学科。这一变革和飞跃将对中国统计教育事业产生巨大而深远的影响，中国统计教育事业将在“十二五”时期发生积极变化。

正是在这一背景下，全国统计教材编审委员会制定了《“十二五”全国统计教材建设规划》(以下简称“规划”)。根据“学科目录”在统计学下设有数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科的构架，“规划”对“十二五”全国统计规划教材建设作了全面部署，具有以下特点：

第一，打破以往统计规划教材出版学科单一的格局。全面发展数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科规划教材的出版，使“十二五”全国统计规划教材涵盖5个二级学科，形成学科全面并平衡发展的出版局面。

第二，打破以往统计规划教材出版层次单一的格局。在编写出版好各学科本科生教材的基础上，对研究生教材出版进行深入研究，出版一批高水平高层次的研究生教材，为我国研究生教育、尤其是应用统计研究生教育提供教学服务。同时，积极重视统计专科教材出版，联合各专科院校，组织编写和出版适应统计专科教学和学习的优秀教材。

第三，打破以往统计规划教材出版品种单一的格局。鼓励内容创新，联系统计实践，具有教学内容和教学方法特色的、各高校自编的相同内容选题的精品教材出版，促进统计教学向创新性、创造性和多样性发展。

第四，重视非统计专业的统计教材出版。探讨对非统计专业学生的统计教学问题，为非统计专业学生组织编写和出版概念准确、叙述简练、深入浅出、表

达方式活泼、练习题贴近社会生活的统计教材，使统计思想和统计理念深入非统计专业学生，以达到统计教学的最大效果。

第五，重视配合教师教学使用的电子课件和辅助学生学习使用的电子产品的配套出版，促进高校统计教学电子化建设，以期最后能形成系统，提高统计教育现代化水平。

第六，重视对已经出版的统计规划教材的培育和提高，本着去粗存精、去旧加新、与时俱进的原则，继续优化已经出版的统计教材的内容和写作，强化配套课件和习题解答，使它们成为精品，最后锤炼成为经典。

“十二五”期间，编审委员会将本着“重质量，求创新，出精品，育经典”的宗旨，组织我国统计教育界专家学者，编写和编辑出版好本轮教材。本轮教材出版后，将能够形成学科齐全、层次分明、品种多样、配套系统的高质量立体式结构，使我国统计规划教材建设再上新台阶，这将对推动我国统计教育和统计教材改革，推动我国统计教育事业科学发展，提高我国统计教育现代化水平产生积极意义。

让教师的教学和学生的学习事半功倍，并使学生在毕业之后能够学以致用的统计教材，是本轮教材的追求。编审委员会将努力使本轮教材好教、好学、好用，尽力使它们在内容上和形式上都向国外先进统计教材看齐。限于水平和经验，在教材的编写和编辑出版过程中仍会有不足，恳请广大师生和社会读者提出批评和建议，我们将虚心接受，并诚挚感谢！

国家统计局

全国统计教材编审委员会

2012 年 7 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、财经管理类学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，为学生进行量化分析起一定示范作用。

本书也可作为应用统计专业硕士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21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、财经管理类学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、财经管理类学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、财经管理类学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，为应用统计专业硕士学位和非统计专业的人文社会科学、财经管理类学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80余位学者获奖。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大数据时代对我们的基本要求。

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，本着提高学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，为非统计专业学生进行量化分析起一定示范作用。需要注意的是，教师不必拘泥于哪一种软件，是用SPSS、SAS、MINITAB、R，还是用S—Plus完全由教师和读者自由选择，哪种软件用着方便，能解决你的问题就是好软件。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届研究生和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。

本书可作为应用统计学专业本科生多元统计分析课程的教材。

本书的大部分内容都给非统计专业学生讲授过，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。

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。我的博士生马学俊、胡小宁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。本书修订过程中得到西京学院院长任芳博士、副院长张辉教授、教务处长肖建军先生的大力支持。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。

何晓群

2015 年春于西京学院应用统计科学研究中心

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就是我们研究的总体，而每个家庭就是个体。

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、偏度与峰度

偏度和峰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。

峰度用峰度系数 V_2 表示：

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

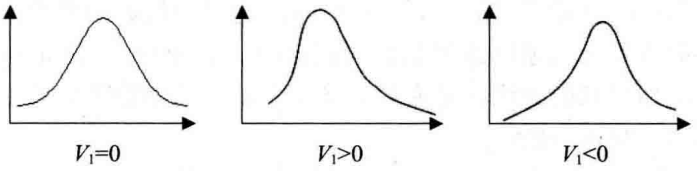

图 1.1

当峰度系数 $V_2=3$ 时，一般为标准正态分布。

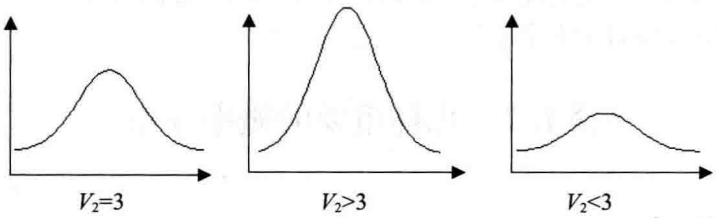

图 1.2

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

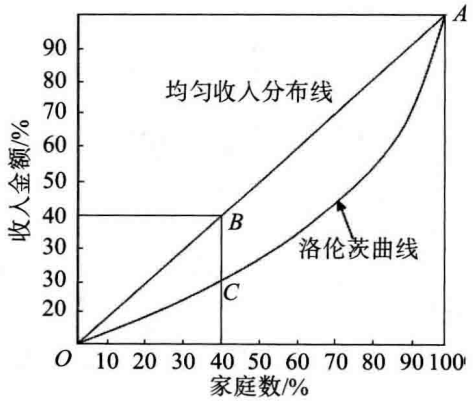

图 1.3

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，所得到的曲线就是宏观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%。

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。

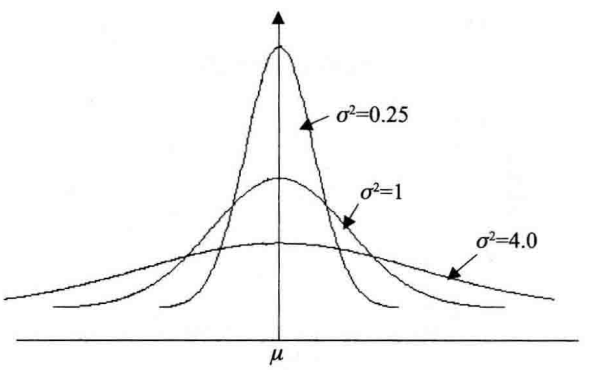

图 1.4

正态分布曲线下，位于 $\mu\pm\sigma$，$\mu\pm2\sigma$，$\mu\pm3\sigma$ 之间的面积分别约占总面积的 68.26%，95.45%和 99.73%，如图 1.5 所示。

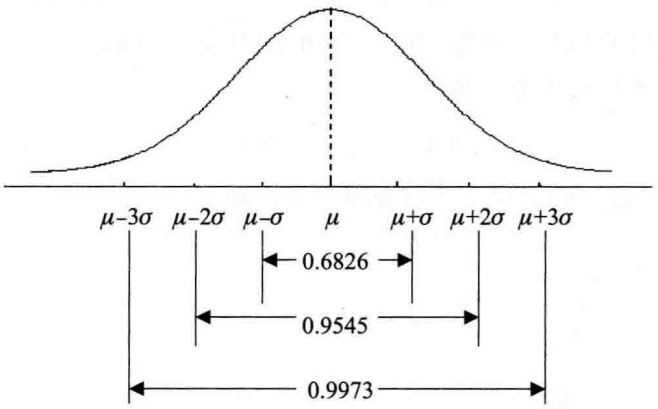

图 1.5

在正态分布的概率密度中，当 $\mu=0$，$\sigma=1$ 时，称随机变量 X 遵从标准正态分布，记为$X\sim N(0,1)$。

在质量管理中，我们知道

$$P(\mu-3\sigma<X<\mu+3\sigma)=\Phi(3)-\Phi(-3)=0.9973$$

这表明当某质量特性 $X\sim N(\mu,\sigma^2)$时，其特性值落在区间$(\mu-3\sigma,\mu+3\sigma)$外的概率仅为 0.27%。这是一个小概率事件，通常在一次试验中是不会发生的，一旦发生就认为质量发生了异常，在质量检验和过程控制中就用这一思想。

当上下公差不变时，6σ 的质量水准就意味着产品合格率达到 99.9999998%，即

$$P(\mu-6\sigma<X<\mu+6\sigma)=\Phi(6)-\Phi(-6)=0.999999998$$

其特性值落在区间$(\mu-6\sigma,\mu+6\sigma)$外的概率仅为十亿分之二。

由于种种随机因素的影响，任何流程在实际运行中都会产生偏离目标值或者期望值的情况，我们通常把这种偏移称为漂移。

我们知道传统的“3σ 原则”下，质量标准的合格率为 99.73%。即使在没有任何漂移的情况下，也意味着 2700ppm(Part Per million)的不合格率。考虑到漂移时是 66807ppm。

通常当考虑到 1.5σ 漂移后，$6\sigma<\neq>$十亿分之二次品率，不合格率为百万分之 3.4。即在某生产流程或服务系统中有 100 万个出现缺陷的机会，而 6σ 质量水准出现的缺陷不到 4 个。更为详细的内容见文献[1]。

关于正态分布的理论已很完善，数学上也易于处理。此外，当一个经济问题的模型误差是由很多因素构成的时候，总体的分布与正态分布近似。所以，在计量经济学和一些经济问题的建模研究中常假定为正态分布。另外，当总体概率分布为正态分布时，作为从中抽出的样本，其统计量的样本概率分布有 χ^2 分布、t 分布、F 分布等。因此，正态分布成为计量经济学乃至统计学中最重要的概念。