21世纪高校规划教材·经济数学类

概率论与数理统计新编

严 忠 肖彰仁 岳朝龙 编著

中国科学技术大学出版社

21世纪高校规划教材·经济数学类

概率论与数理统计新编

严　忠　肖彰仁　岳朝龙　编著

中国科学技术大学出版社

合　肥

内 容 简 介

本书是面向21世纪高等院校经济学学科门类和管理学学科门类本科数学基础课教材之一。本书依据经济管理类专业人才教育培养目标的要求，以经济问题提出，从数据初步处理入手，介绍了数据处理、随机变量、母函数、大数定律与中心极限定理、抽样分布、参数估计、假设检验、方差分析、回归分析等概率论与数理统计的基本内容，重点介绍了数理统计方法及其在经济、管理领域中的应用。多数章节均配有适量的习题，书末附有参考答案，以方便读者学习与研究。

本书除可供经济管理类各专业本科教育教学使用外，还可作为其他相关人员的学习参考书。

图书在版编目(CIP)数据

概率论与数理统计新编/严忠，肖彰仁，岳朝龙编著. —合肥：中国科学技术大学出版社，2003.2

21世纪高校规划教材·经济数学类

ISBN 7-312-01495-X

Ⅰ.概… Ⅱ.①严…②肖…③岳… Ⅲ.概率论—数理统计—高等学校—教材 Ⅳ.0212

中国版本图书馆CIP数据核字(2002)第076490号

凡购买中国科大版图书，如有白页、缺页、倒页者，由承印厂负责调换。

中国科学技术大学出版社出版发行

(安徽省合肥市金寨路96号，邮编：230026，发行电话：0551-3602905，3602906)

中国科学技术大学印刷厂印刷

全国新华书店经销

开本：787mm×960mm 1/16 印张：15 字数：336千

2003年2月第1版 2003年2月第1次印刷

印数：1—4000册 **定价：19.00元**

ISBN 7-312-01495-X/O·268

总 前 言

随着我国教育体制改革的不断深入，特别是高等教育的改革和发展，对21世纪人才培养提出了更高、更新的要求。不同类型的高等院校对人才培养的目标要体现各自的特色。作为理工科普通高等学校应侧重于应用型人才的培养。在体现宽专业、厚基础、重应用的基础上，要充分体现21世纪经济建设主战场对人才的需求。因此，需要在教学计划、教材、教学大纲以及教育教学方法等方面得到反映。对此，我校经济学院在教育部"新世纪高等教育教学改革工程"本科教育教学改革课题《一般工科院校经济学类本科生人才培养模式研究》的研究取得成果之后，组织编写了这套系列教材。

该系列教材包括《政治经济学新编》、《现代西方经济学》、《管理经济学》、《国际贸易》、《国际贸易实务》、《国际金融学》、《现代货币银行学》、《金融投资学》、《统计学》、《期货与期权交易》、《概率论与数理统计新编》、《SAS在经济分析中的运用》等。

该系列教材的宗旨：面对21世纪经济社会发展对人才的需求，适应我国加入WTO及世界经济一体化的需求，体现一般院校经济管理类专业应用型人才培养目标。力求内容完整、重点突出、深入浅出；重点、难点内容叙述详细；注重应用和案例教学。同时，力求能够反映当今国内、国际经济科学发展前沿。

参加这套教材编写工作的作者都是长期在教学岗位上从事教学和研究工作的教授、学者，他们具有较高的学术造诣和宽厚的专业基础知识以及丰富的教学和科研经验。更为可贵的是他们思路开阔，接受新知识、新事物的能力强，热爱本专业，具有远大的志向。当然，由于水平和经验的局限，这套教材也难免存在这样或那样的不足或不妥之处，希望广大读者提出宝贵的意见和建议，以便他们在今后的教学和研究工作中总结、提高，从而使本书在以后再版时渐臻完美，为造就新一代高素质经济管理复合型人才做出更多的贡献。

安徽工业大学校长 董元儒

2002年1月18日

前　言

目前经济管理类本科数学基础课教材一般使用龚得恩先生所编写的经济数学基础——《概率论与数理统计基础》，我们认为，该教材在前一时期对经济管理类本科教学质量的提高起到了很大的作用，为经济管理学科的发展和人才培养做出了积极的贡献。随着我国高等教育事业的快速发展，特别是教育部提出的“新世纪高等教育教学改革工程”的全面实施，对该学科教学内容也提出了适应改革的需要的新要求，为此，我们编写了这本《概率论与数理统计新编》。

在本教材的编写中我们着重考虑了下列指导思想：

(1) 教材建设应紧密结合“新世纪高等教育教学改革工程”中的本科教育教学改革立项项目的研究，这是因为教材建设也是高等教育教学改革的内容之一，而在有关的本科教育教学改革立项项目的研究目标中，也常含有教材建设。因此，将教材建设纳入本科教育教学改革立项项目的研究，会进一步深化教改成果。本教材在正式出版前，曾以初稿形式在我校承担的教育部“新世纪高等教育教学改革工程”课题研究试点班教学中使用。

(2) 随着各学科的发展和教育教学改革的发展，应以体现时代性的新观点、新方法构建教材体系的内容结构。例如，随着知识经济时代的到来和目前对金融风险研究的迅速发展，如何加强数理统计这一有力支撑的基础教学，就必须重构教材体系的内容结构。

(3)随着现代化计算技术的不断提高，教材建设应紧密结合计算机的应用，将有关应用软件如 Excel，SAS 的教学运用融入教材之中。

根据上述指导思想，我们力求新教材体现下列特色：

(1) 基础性——以经济问题的提出，从数据初步处理开始，通俗地导入经济管理类专业的基础课程“概率论与数理统计”内容。

(2) 应用性——针对概率论与数理统计在经济管理学科的实际应用背景，构建本书内容体系。例如，淡化概率论部分，增加回归分析部分；考虑到金融风险

研究的需要，加写了“母函数”一章，等等。

（3）现代性——结合现代计算技术的使用，介绍 Excel，SAS 等应用软件在本书中例题和习题中的计算和解题。

（4）不但适用经济管理类各专业的教学，也可作为其他教学，如数学建模的参考书。

本书共九章。第一章，数据处理，介绍了数据处理的基本方法。第二章，随机变量，实际上涵盖了概率论基础的基本内容。如上所述，考虑到金融风险研究的需要，我们加写了第三章，母函数，供教材使用者根据教学对象酌情选用。这三章由严忠编写。第四章，大数定律和中心极限定理，将概率论与数理统计加以连接。第五章，介绍抽样分布，以体现“通过样本，估计总体”这一基本思想。第六章为参数估计。这三章由肖彰仁编写。第七章为假设检验。针对经济管理学科的实际应用背景，我们注意到了第八章方差分析和第九章回归分析的重要性。这三章由岳朝龙编写。

限于作者的水平，书中疏漏和不当之处在所难免，恳请同行专家、学者、读者不吝赐教。

此外，本书在编写过程中曾得到高等教育出版社的关注，中国科技大学有关专家、教授对本书的出版也给予了大力的支持，在此一并表示衷心感谢！

严 忠

2002 年 6 月

目　次

第一章　数 据 处 理

△知识经济时代也是信息经济时代，作为信息的主要载体之一的数据，在经济生活和社会生活中有着重要作用。

△实际统计的主要过程：收集数据⟶ 数据处理⟶ 统计推断⟶ 分析报告。

1.1　数 据 描 述

1.1.1　数据的基本概念

1. 批数据（batch data）

批数据是一组相关的观察数据，例如：

2001 年我国各主要日用品的零售价格的一组数据；

安徽工业大学 2001 级学生“概率统计”课程成绩；

……

2. 样本数据（sample data）

样本数据是一组从研究对象总体中抽出的，可用以推断总体性质的一组个体的观察数据，例如：

【例 1.1】 某银行从 2001 年中抽出 60 天的 11：00 —12：00 期间窗口等待服务的人数所组成的数据：

2 1 3 1 1 4 5 2 2 1 4 5 4 2 2 0 3 2 2 2 1 2 2 2 3 2 3 1 1 4

3 2 1 3 0 3 1 6 2 2 3 1 7 4 0 2 2 1 3 5 6 1 2 3 1 0 4 6 1 1

【例 1.2】 某保险公司从火险索赔案中抽出 100 个案例中的索赔额所组成的数据；

……

3. 截面数据（Cross data）

截面数据是同一个时间点上的一组相关的观察数据，例如：

【例 1.3】 某企业某日按产量（件）分组，统计出的完成相应产量的工人人数数据

（见表 1.1）。

表 1.1

按产量（件）分组	40－50	50－60	60－70	70－80	80－90
工人数（人）	12	30	52	27	9

……

4．时序数据（Time data）

时序数据是在一定时间范围内每个时间点上的一组相关的观察数据，例如：

【例 1.4】 表 1.2 是某国 1985—1988 年的 GNP（单位：亿美元）时序资料：

表 1.2

年度	1985	1986	1987	1988	1989
GNP	450	470	465	471	478

……

5．混合数据或称平行数据（Panel data）

混合数据，或称平行数据（panel data）,是截面数据和时序数据的合成体，例如：

【例 1.5】 表 1.3 是某种商品 12 年的需求量（y，单位：万件）、价格（x_1，单位：元/件）和广告支出（x_2，单位：万元）时序资料：

表 1.3

序号	1	2	3	4	5	6	7	8	9	10	11	12
y	36	48	45	40	30	56	63	53	61	68	66	65
x_1	56.7	63.9	62.7	59.7	55.9	68.7	69.2	65.5	69.4	73.4	74.1	74.4
x_2	12	9	11	13	14.5	9.5	7.5	11	9	8.5	10	9.5

对于有些数据我们可直观地以点图表示出来，如例 1.5 中，我们可以用点、圈和叉号在 noy 坐标中（其中 n 坐标轴为序号坐标轴，y 坐标轴为数据值坐标轴）分别将 y，x_1 和 x_2 表示出来。

1.1.2 频数分布与条形图

对例 1.1 的数据可采用简明的描述方式即频数分布加以描述。通常，将一个变量 x（例 1.1 中人数）的所有可能取值为 0，1，2，3，……，而把这些数值出现的次数即称为**频数**，用 m 表示，则表 1.4 给出的数据描述方式即称为**频数分布**。

表 1.4

x	0	1	2	3	4	5	6	7	8 及以上
m	4	15	18	10	6	3	3	1	0

进一步，我们可以用条形图将此频数分布更直观地表达出来（图 1-1）。

1.1.3　组频数与直方图

组频数又称群频数或分组频数，例 1.3 给出的表 1.1 便是这种数据描述方式。通常，将一个变量 x（表 1-1 中产量）的所有可能取值分为区间长相等的连续区间：40－50，50－60，60－70，70－80，80－90，……而对应于这些区间的工人人数也就是这些区间出现的次数即称为**群频数**或**分组频数**，仍用 m 表示。同样，我们可以用条形图将此组频数分布更直观地表达出来（图 1-2）。

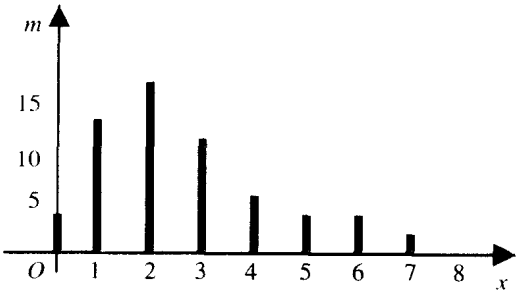

图 1-1

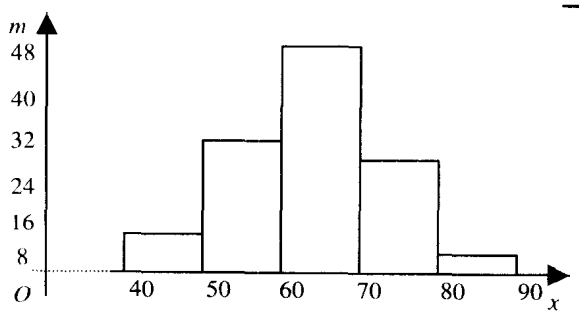

图 1-2

1.1.4　数据的其他描述方式

1. 相对频数

相对频数即频数与数据总数的比值。例如,例 1.1 中的相对频数如表 1.5 所示。

表 1.5

x	0	1	2	3	4	5	6	7	8 及以上
相对频数	0.067	0.250	0.300	0.167	0.100	0.050	0.050	0.017	0

在第二章中，我们将会看到这是一个对应于随机变量概率的数据，被视为从数据中得到的经验概率。

2. 累积频数

累积频数是指数据小于和等于某值的某数据组的频数累加之和（包括取其自身和比其更小值的频数）。例如,例 1.1 中的累积频数如表 1.6 所示。

表 1.6

x	0	1	2	3	4	5	6	7	8 及以上
累积频数	4	19	37	47	53	56	59	60	60

3. 相对累积频数

相对累积频数即累积频数与数据总数的比值。例如,例 1.1 中的相对累积频数如表 1.7 所示。

表 1.7

x	0	1	2	3	4	5	6	7	8 及以上
相对累积频数	0.067	0.317	0.617	0.783	0.883	0.933	0.983	1	1

在第二章中,我们将会看到，相对累积频数分布相似于随机变量的分布函数。

1.2 数据分布的基本度量——均值、标准差与偏斜度

人们对数据分布情况通常关心的主要是：数据分布的位置、密集或离散程度、偏斜程度等等，本节将讨论对数据分布情况的基本度量。

1.2.1 数据分布位置的度量

1. 均值（mean）

我们这里所指的是通常所说的算术平均数。对于其他均值我们再冠以特别表示，如调和均值、几何均值等。

对一个有 n 个数据的数列 $x_1, x_2, \cdots x_n$(或$x_i, i=1,2,\cdots,n$)

其均值为：

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

例如,例 1.5 中：

$\bar{y}=(36+48+45+40+30+56+63+53+61+68+66+65)/12=52.583\doteq 52.6$

$\bar{x}_1=(56.7+63.9+62.7+59.7+55.9+68.7+69.2+65.5+69.4+73.4+74.1+74.4)/12$
$\doteq 66.13$

$\bar{x}_2=(12+9+11+13+14.5+9.5+7.5+11+9+8.5+10+9.5)/12\doteq 10.38$

对于取值 $x_1, x_2, \cdots, x_k$ 的频数分布，其对应的频数为 $m_1, m_2, \cdots, m_k$，其中,$\sum m_i=n$，其均值为：

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

例如，例 1.1 中：

$\bar{x}=(0\times 4+1\times 15+2\times 18+3\times 10+4\times 6+5\times 3+6\times 3+7\times 1)/60=2.417\doteq 2.4$

而对于组频数分布，其各组频数为 $m_1, m_2, \cdots, m_k$，其中，$\sum m_i = n$，取各组的中间值 $x_1, x_2, \cdots, x_k$，则其均值为：

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

例如,例 1.3 中：

$$\bar{x} = (\frac{40+50}{2}\times 12 + \frac{50+60}{2}\times 30 + \frac{60+70}{2}\times 52 + \frac{70+80}{2}\times 27 + \frac{80+90}{2}\times 9)/130 \doteq 64.3$$

需要指出的是，我们在计算中所设的精确度取比原始数据多一位小数。

2. 中位数（median）

中位数是指一个按大小顺序排列的数据列中的中间那个数。如果有奇数个数据，则中位数就是中间那个数，如果有偶数个数据，则中位数就是中间两个数的平均数。

中位数与均值是两个不同的概念。例如，对 5 个观察值：1，2，3，4，5，其中位数是 3，均值也是 3；但对于另 5 个观察值：1，2，3，4，100，其中位数还是 3，而均值是 22，相差很大。即一些极端数据可能对均值影响很大，而对中位数来说，这个影响就消除了。

下面我们将会看到，中位数还可以更好地描述那些数据分布高度偏斜的数列。

3. 众数（mode）

众数是指数据列中频数最多或最典型的那个数。例如，例 1.1 中的众数为 2，它是 8 个数之一,但它出现的次数占总次数的 3/10，即频数最多。

1.2.2 数据分布密集或离散程度的度量

1. 标准差（standard deviation）

标准差是反映数据偏离均值远近程度的度量。

对一个均值为 $\bar{x}$ 的数列 $x_i, i = 1,2,\cdots,n$，取 $(x_i - \bar{x})$ 为 x_i 到均值的距离，即与均值的偏差。显然，这些偏差的均值为零，这是因为：

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x} = \bar{x} - \frac{1}{n}\cdot n\bar{x} = \bar{x} - \bar{x} = 0$$

因此，用数列到均值的偏差之和无法度量数据的密集或离散程度。为此我们采用偏差平方均值的根（使度量单位与初始单位保持一致）$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ 来度量数据的密集或离散程度，再由于统计推断的技术原因，以 $n-1$ 来代替 n，于是，我们可定义标准差为：

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

记为 s，有：

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

称为**方差**。对于方差的计算公式，我们将其变换为一个宜于通常计算的公式：

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

或者

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

例如，例 1.1 中，由频数分布可得：

$$\sum m_i x_i=145,\sum m_i x_i^2=505,$$

$$\therefore\quad s^2=\frac{1}{59}(505-\frac{1}{60}145^2)=2.62$$

$$s=1.62\doteq 1.6\quad\text{(取与均值一样的精度)}$$

而对于组频数分布，各组取中间值代入计算，例如,例 1.3 中：

$$\sum m_i x_i=8360,\sum m_i x_i^2=551650$$

$$\therefore\quad s^2=\frac{1}{129}(551650-\frac{1}{130}8360^2)=108.82$$

$$s=10.43\doteq 10.4\quad\text{(取与均值一样的精度)}$$

关于数据均值与标准差的关系，我们将在第五章里研究，这里不再赘述。

2. 极差

极差是关于数据密集或分散的一个简单度量，被定义为最大值与最小值之间的差，即

$$R=\max(x_i)-\min(x_i)$$

显然极差越小数据越密集。

例如,例 1.1 中，极差为：R=7–0=7。

1.2.3 数据分布偏斜程度的度量

偏斜度的度量可以通过后续内容所要学习的三阶矩来进行。这里我们以均值、众数、中位数的相对位置来表明偏斜度。一般地来说，如果一个数列的均值、众数、中位数非常