21世纪高等院校选用教材

经济、管理类

山东省“九五”立项教材

# 概率论与数理统计

山东经济学院 山东大学数学学院 组编

刘锦萼 杨喜寿 俞纯权 房俊岭 编著

科学出版社

21世纪高等院校选用教材(经济、管理类)
山东省"九五"立项教材

# 概率论与数理统计

山东经济学院、山东大学数学学院　组编

刘锦萼　杨喜寿
俞纯权　房俊岭　编著

科学出版社
2001

## 内 容 简 介

本书系山东省"九五"立项教材.全书的特点是改变传统的编写模式,从数据分析入手,先讲描述统计,后讲推断统计,再到现代统计方法.全书内容由四个模块有机组成:一是描述性统计,二是基础概率,三是统计推断基础,四是常用数理统计方法(包括多元统计分析方法).为适应市场经济对统计的需求,书中增加了投资分析、股票指数、保险精算、Bayes 预测等内容和实例.各章后面均配有习题,书末附有习题答案.

本书可作为高等学校经济、统计、管理类专业的教材,也可作为其他专业学生和各类经济管理人员的参考书.

**图书在版编目(CIP)数据**

概率论与数理统计/刘锦萼,杨喜寿,俞纯权,房俊岭编著 —北京:科学出版社,2001.8

(21 世纪高等院校选用教材(经济、管理类))

ISBN 7-03-009576-6

Ⅰ.概… Ⅱ.①刘… ②杨… ③俞… ④房… Ⅲ.①概率论-高等学校-教材 ②数理统计-高等学校-教材 Ⅳ.O21

中国版本图书馆 CIP 数据核字(2001)第 041399 号

科学出版社 出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2001 年 8 月第 一 版 开本:720×1000 1/16

2001 年 8 月第一次印刷 印张:27

印数:1—5 000 字数:484 000

**定价:35.00 元**

(如有印装质量问题,我社负责调换〈环伟〉)

# 前　言

概率论与数理统计是一门研究现实世界中随机现象统计规律的科学.在社会、经济和科学技术中广泛存在着随机现象,需要用概率统计方法去分析和处理各种带有随机干扰的数据,直至做出科学的决策.因此本课程不仅是统计学专业的主干课程,而且也是经济管理类专业的必修课程.随着社会经济的发展和科学技术的进步,概率统计的研究领域不断拓展,研究内容也日益更新.为了"全面适应现代化建设对经济类人才培养的需求",对原有的课程体系和教学内容进行改革,剔除陈旧的内容,吸收先进的成果,编写一本具有时代特色和较高水平的概率统计教材,是高等教育改革的要求,也是我们追求的目标.

就笔者所涉及的范围,本课程目前在国内已出版的教材大致有两类,一类理论性较强,数学论证严谨冗长,不适宜经济管理类专业教学;一类内容陈旧,仍沿用20～30年前的教学大纲,已远落后于社会经济发展和科学技术进步的步伐.本书是山东省教育厅立项的"九五"时期重点教材,笔者力图在教材的体系、内容等方面进行较为深入的改革和探索,以适应新时期经济管理类专业的教学要求.在市场经济时代,自然科学和社会科学的相互渗透和优势互补成为科学发展的必然要求和显著特征.我们认为,定性分析与定量分析相结合,广泛运用数理统计方法研究市场经济,是经济管理类专业概率论与数理统计课程改革的最根本特征.为体现上述意图,我们在体系和内容上采用了带有创新意识的设计.全书由四个模块有机组成.第一模块是描述性统计,包括数据的整理分析、指数理论和时间数列分析;第二模块是基础概率,包括概率的定义、性质、随机变量与分布函数,以二项分布、泊松分布、正态分布为代表的分布族,数字特征与极限定理;第三模块是统计推断基础,包括数理统计的基本概念、常用抽样分布、统计推断的两大基本方式——估计与检验;第四模块是常用数理统计方法,包括方差分析、回归分析、聚类分析、主成分分析、因子分析.在教材中,结合当前经济发展中人们关注较多的金融、保险知识,增加了物价指数、道·琼斯股票价格指数、保险精算、福利奖券、投资决策分析等内容或例题.对于在经济科学中应用日益广泛的贝叶斯学派的统计思想,本书也有介绍和评注.对于一些初学者不易理解的概念,书中都介绍得十分详细和清楚.

此外,笔者从伯努利试验模型、泊松流和随机误差等实例出发,分别导出以二项分布、泊松分布、正态分布为代表的分布族,力图使读者能从机制上来

理解这些分布的来源及适用情形，而不仅仅是面对一系列抽象的数学表达式.统计方法与技术是现代社会经济研究中的基本方法.现代社会经济的复杂性决定了统计描述是多指标的，即多元的.在多指标场合如何有效地使用和分析统计信息以显示社会经济现象的统计特征，是多元统计分析的任务.对于有较高实用价值的聚类分析、主成分分析和因子分析等多元统计分析方法，本书侧重于直观背景及案例来进行介绍.

就全书内容而言，本书比一般常见的同类教材要丰富一些，书中既有概率统计的基本内容和基本训练，也有为报考硕士研究生的读者准备的较为深刻的内容.我们的意图是想让读者和使用本书的教师有较大的选择空间.由于不同专业之间不可避免地存在差异，同时读者的基础和要求也不可能完全一致，因此在使用本书时可根据各专业需要来决定内容的取舍.书中有些如 * 的内容，如概率空间、条件分布、条件数学期望、矩母函数、依概率收敛、次序统计量、贝叶斯估计、估计的有效性和克拉美-罗不等式、非参数检验、主成分分析、因子分析等是较为深入难懂的内容，初次学习时可以略而不学.好在即便跳过这些内容，仍可学到概率论和数理统计的最基本的知识.鉴于这些内容在统计学中的重要性，建议读者在具备了较好的概率统计基础以后再去补学它们.

笔者在高等学校从事概率统计的教研工作都已有数十年，也积累了不少的经验，但真的要写好一本具有特色和新意的书也非易事.我们的想法和做法都是一种尝试，效果如何有待于实践的检验及读者和使用本书的教师的评论.我们将不断努力，使本书日臻完善，使之成为一本跟上时代步伐的较高水平的教材.我们深信除了笔者自身的努力之外，还必须得到广大师生和读者的支持，特别是广泛听取对本书的批评和建议，进行精益求精的修改，才能使这一目标得以实现.我们由衷地期待着各种指正和建议.

编写本书的分工如下：大纲和体系由四人集体商讨而定.其中第一、二、三、十章由杨喜寿撰写，第四至九章由刘锦萼撰写，第十一至十五章由俞纯权撰写，习题和附图的配备由房俊岭负责，最后由刘锦萼审阅全书.

本书的撰写和出版得到山东省教育厅、山东经济学院、山东大学数学学院的大力支持.科学出版社的吕虹编审给了很多鼓励和帮助.特别是郑骏、徐侃、张凤祥诸位教授在百忙中拨冗对初稿进行了评阅，花费了很多的时间和心血，提出了许多宝贵意见.对于上述机构和同志，笔者借此机会表示衷心的感谢.

编　者

2001.1

# 目　录

# 第一章　数据整理与分析

在分析处理经济与管理问题时，经常使用大量的数据.例如要了解全国农民在实行联产承包责任制之后生活改善的状况，就需要对大量的农户的收支情况作调查.再如电视机厂要了解某批显像管的质量，就要对相当多的显像管的某些指标进行测试.今后我们称上述类型的调查或测试为“观察”；称所要考察研究的对象的全体为“总体”；总体中每个元素称为“个体”.观察可对每个个体进行，也可以只观察一部分个体.所观察的部分个体，称为“样本”.一个样本所含有的个体的数目，称为这个样本的“容量”.

观察的结果，常常是一些杂乱无章的数，称为原始数据，或简称为“数据”.管理者要从数据中提取有用的信息，用来反映所观察总体的某些特性，必须先对数据加以整理.根据对数据不同的使用目的，常用的几种整理数据的方法是分组法、图示法以及特征数计算法等.这三种类型的数据整理方法又各自包括许多具体的方法.本章着重介绍这些方法的基本要求和步骤.读者学习和使用这些方法，可以根据面对的实际问题自行改进.评价这类方法的好坏，主要是看是否简便、适用，并没有严格的理论上的要求.

在 Windows 系统上运行的电子表格软件“Microsoft Excel”是目前世上最为流行的处理繁杂数据的工具，本章所介绍的内容都不难使用该软件来完成.

## §1.1　数据分组方法

数据的分布情况，即在各种大小区段所占的百分比的情况，能够反映出所考察的总体的某些性质.原始数据一般是杂乱无章的.为了较清楚地看出数据的分布情况，最基本的方法是将数据按一定的规则分组.下面我们将结合例子来讨论数据分组的一般要求、原则和步骤.

**例 1.1.1**　某装配车间要考察一批轴承的质量，从该批轴承中抽取 56 个，测量其内径，其值为(单位：毫米)：

73.93　74.05　73.98　74.02　73.99　74.02　73.99　74.02
74.08　73.94　74.09　73.97　74.05　73.99　74.00　74.02
74.02　73.95　74.05　73.96　74.04　74.00　74.02　74.01.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 74.10 | 73.94 | 73.96 | 74.00 | 74.01 | 74.06 | 74.04 | 74.03 |
| 73.97 | 74.00 | 74.06 | 74.00 | 73.04 | 74.03 | 74.07 | 73.96 |
| 74.01 | 74.04 | 74.00 | 74.00 | 74.01 | 73.98 | 74.01 | 73.99 |
| 74.00 | 74.04 | 73.97 | 74.03 | 73.99 | 74.00 | 74.06 | 73.97 |

试通过对数据的分组整理，反映出数据的分布情况.

第 1 步：确定分组个数 $k$. 分组个数 $k$ 是根据样本的容量 $N$ 来确定，一般 $N$ 越大，相应分组个数 $k$ 也大. 经验表明，当 $N$ 在 30 至 100 时，$k$ 取 5 到 10 比较适宜；当 $N$ 在 100 至 300 时，$k$ 取 7 到 15 比较适宜. 分组太少，会损失较多的信息，分组过多，又不易反映出数据的概貌.

本例样本容量为 56，确定分为 6 组.

第 2 步：找出数据中的最大值 $L$，最小值 $S$，并计算极差 $R$（定义极差 $R=L-S$）.

本例中，$L=74.10, S=73.93, R=0.17$

第 3 步：确定组间距 $h$. 其公式为

$$h \approx \frac{R}{k}$$

一般应取 $h$ 稍大于 $\frac{R}{k}$，以便在分组时能将最小值 $S$ 包含在第一组，将最大值 $L$ 包含在最后一组. 为了简便，又要求 $h$ 取位数尽可能少一些.

本例中，$\frac{R}{k}=0.028$，取 $h=0.03$.

第 4 步：确定组界. 关于组界的确定，针对不同的实际问题，有几种不同类型的方法，后面我们还要讨论. 这里，我们要求所划分的各组的组界比原始数据多取一位数，以使得在将原始数据分组时，数据不至于落到组界上.

本例中，原始数据的最小数为 73.93，我们将第一组的左界取为 73.925，那么第一组的右界即为

$$73.925+h=73.925+0.03=73.955$$

第 $j$ 组的左、右界分别为

$$73.925+(j-1)h, 73.925+jh$$

如表 1-1-1 的第(2)列所示.

第 5 步：计算各组的频数、频率、累计频数、累计频率.

第 4 步确定了组界之后，每个组就是一个确定的区间. 称每个组（即对应的区间）所包含的原始数据的个数为该组的频数. 假定以 $f_j$ 表示第 $j$ 组的频数，则称 $\frac{f_j}{N}$ 为该组的频率，一般用百分数表示. 这里 $N$ 表示原始数据的个数. 一个组的频率，就是原始数据落在这个组的比数. 假定将第 $j$ 组的频率记为

$f_j^*$,则第 $i$ 组的累计频数和累计频率分别定义为

$$F_i = f_1 + f_2 + \cdots + f_i (i = 1,2,\cdots,k)$$

$$F_i^* = f_1^* + f_2^* + \cdots + f_i^* (i = 1,2,\cdots,k)$$

本例中,各组对应的频数、频率、累计频数、累计频率如表 1-1-1 所示.

**表 1-1-1 数据(频数、频率)分组表**

| 组号 (1) | 组界 (2) | 组中值 (3) | 频数 (4) | 累计频数 (5) | 频率(%) (6) | 累计频率(%) (7) |
|---|---|---|---|---|---|---|
| 1 | 73.925 73.955 | 73.94 | 4 | 4 | 7.1 | 7.1 |
| 2 | 73.955 73.985 | 73.97 | 8 | 12 | 14.3 | 21.4 |
| 3 | 73.985 74.015 | 74.00 | 16 | 28 | 28.6 | 50.0 |
| 4 | 74.015 74.045 | 74.03 | 17 | 45 | 30.4 | 80.4 |
| 5 | 74.045 74.075 | 74.06 | 8 | 53 | 14.3 | 94.7 |
| 6 | 74.075 74.105 | 74.09 | 3 | 56 | 5.3 | 100 |

第 6 步:计算各组的组中值.假定以 $a_j$ 和 $b_j$ 分别表示第 $j$ 组的左界和右界,则称

$$C_j = \frac{a_j + b_j}{2}$$

为第 $j$ 组的组中值.在处理实际问题时,常用组中值作为该组的代表值.

本例中.各组的组中值如表 1-1-1 中第(3)列所示.

第 7 步:将第 4,5,6 步所得结果列表.本例列表如表 1-1-1.

例 1.1.1 中,原始数据为 56 个,利用上述分组方法分为 6 个组.观察各组界及各组的频率,可以清楚地看出数据的分布情况.

下面我们将进一步讨论分组法中的几个问题.

数据分组方法是一种经验性整理数据的方法.要根据实际情况灵活地选取、调整分组法中的几个参数,如分组个数 $k$、组间距 $h$、各组的组界等.

在例 1.1.1 中,为了不使原始数据落到组界上,我们在确定组界时多取了一位小数.有时为了简便,也可以使用和原始数据有相同精确位数的组界,将各个组表示为形如$(a,b]$或形如$[a,b)$的半开半闭的区间.特别对于那些只可能取整数的量,这种确定组界的方法是合适的.例如为考察某家电器门市部日销售彩色电视机台数,现有 60 天的销售记录,其中日销售量最少为 2 台,最多为 19 台.取组间距为 3,共分 6 组,其分组区间表示为$[2,5)$,$[5,8)$,

[8,11),[11,14),[14,17),[17,20),这是左闭右开形区间.区间包含其左界,而不包含其右界,所以60个数据中的每一个必有确定的组别归属.类似地,也可以表示为左开右闭形区间.

在分组时,每个组的组间距也可以取不同的值.例如上面提到的日销售彩色电视机台数的例子,如销售12台以上的天数是很少的,而销售8台左右的天数特别多,根据实际情况,我们也可以将数据分组为[2,5),[5,7),[7,8),[8,9),[9,12),[12,20).

**例1.1.2**　某年对某地区农户人均年纯收入作了抽样调查,调查户数为4200户,其分组表如下表所示.

**表1-1-2　农户人均年纯收入分组表**(单位:元)

| 组号 | 组　界 | | 频数 | 累计频数 | 频率(%) | 累计频率(%) |
|---|---|---|---|---|---|---|
| 1 | 1000以下 | | 15 | 15 | 0.36 | 0.36 |
| 2 | 1000 | 1500 | 57 | 72 | 1.36 | 1.72 |
| 3 | 1500 | 2000 | 86 | 158 | 2.05 | 3.77 |
| 4 | 2000 | 3000 | 369 | 527 | 8.79 | 12.56 |
| 5 | 3000 | 4000 | 606 | 1133 | 14.43 | 26.99 |
| 6 | 4000 | 5000 | 674 | 1807 | 16.05 | 43.04 |
| 7 | 5000 | 6000 | 710 | 2517 | 16.90 | 59.94 |
| 8 | 6000 | 8000 | 865 | 3382 | 20.59 | 80.53 |
| 9 | 8000 | 10000 | 411 | 3793 | 9.87 | 90.31 |
| 10 | 10000 | 15000 | 327 | 4120 | 7.78 | 98.09 |
| 11 | 15000 | 20000 | 57 | 4177 | 1.36 | 99.45 |
| 12 | 20000以上 | | 23 | 4200 | 0.55 | 100 |

表1-1-2就是一个不等组间距分组表实例.特别是第一组和最末一组,各有一段是无界的,也称为"开口"的.

上面所介绍的分组方法是按事物的数量标志进行的.分组法也可按事物性质进行.

**例1.1.3**　消防部门为了分析、比较火灾的成因,以便加强消防管理,将火灾按引起的原因分组,见表1-1-3.

通过表1-1-3,人们可以清楚地看出火因的分布情况,例如生活用火不慎时引起火灾的很重要的因素,他所引起的火灾占全部火灾起数的36.9%.

表 1-1-3 某年全国火因分组表

| 组号 | 火 因 | 频数(年) | 频率(%) | 指数* |
|---|---|---|---|---|
| A | 吸烟 | 3161 | 9.0 | 81.3 |
| B | 小孩玩火 | 3294 | 9.4 | 84.7 |
| C | 放火 | 1770 | 5.1 | 45.5 |
| D | 生活用火不慎 | 12919 | 36.9 | 322.2 |
| E | 违反安全操作规定 | 3541 | 10.1 | 91.1 |
| F | 违反电器安装使用规定 | 5214 | 14.9 | 134.1 |
| G | 自燃 | 703 | 2.0 | 18.1 |
| H | 其他 | 1944 | 5.6 | 50.0 |
| I | 未明 | 2450 | 7.6 | 63.0 |
| | 合计 | 34996 | 100 | 900 |

* 参考 §2.1.

## §1.2 数据中心趋向的度量

数据的中心趋向是表示总体特征的一个很重要的特征数.例如人们通常讲的某县小麦平均亩产量,某球队的平均身高,一批电子元件的平均使用寿命,农业总产值平均每年递增百分之几等等.经常使用的表示数据中心趋向的特征数有:算术平均数、加权平均数、几何平均数、中位数和众数等.

### 1.2.1 算术平均数

设 $x_1, x_2, \cdots, x_n$ 是某个变量的 $n$ 个观察值,即容量为 $n$ 的一组数据,称

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

为"算术平均数",也常简称"平均数".这是众所周知的在经济与管理以及日常生活中,经常使用的表达数据中心趋向的特征数.

### 1.2.2 加权平均数

设有数据 $x_1, x_2, \cdots, x_n$,则称

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} (w_i > 0, i = 1,2,\cdots,n)$$

为加权平均数,其中 $w_1, w_2, \cdots, w_n$ 称为权数.