研究生应用数学教材系列

数理统计方法

SHULI TONGJI FANGFA

⊙ 陆元鸿 编著

华东理工大学出版社
EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS

研究生应用数学教材系列

数理统计方法

陆元鸿　编著

华东理工大学出版社

图书在版编目(CIP)数据

数理统计方法/陆元鸿编著. —上海：华东理工大学出版社，2005.8
（研究生应用数学教材系列）
ISBN 7-5628-1778-2

Ⅰ.数… Ⅱ.陆… Ⅲ.数理统计—研究生—教材 Ⅳ.O212

中国版本图书馆 CIP 数据核字(2005)第 087809 号

研究生应用数学教材系列

数理统计方法

编　　著/ 陆元鸿
责任编辑/ 徐知今
封面设计/ 王晓迪
责任校对/ 张　波

出版发行/ 华东理工大学出版社
地　　址：上海市梅陇路 130 号，200237
电　　话：(021)64250306(营销部)
传　　真：(021)64252707
网　　址：www.hdlgpress.com.cn
印　　刷/ 上海展强印刷有限公司
开　　本/ 787×960　1/16
印　　张/ 16.75
字　　数/ 307 千字
版　　次/ 2005 年 8 月第 1 版
印　　次/ 2005 年 8 月第 1 次
印　　数/ 1-4 100 册
书　　号/ ISBN 7-5628-1778-2/O·150
定　　价/ 25.00 元

内 容 提 要

本书是按国家教育部“工学硕士研究生应用统计课程教学基本要求”，并结合作者多年的教学经验，为非数学类专业，特别是工科研究生编写的一本数理统计教材，也可以作为数理统计课学时较少的本科数学专业的教材，还可供从事实际工作的科技工作者和工程技术人员阅读、参考.

本书共分为9章，内容包括：概率论的基础知识，数理统计的基本概念，参数估计，假设检验，回归分析，方差分析和正交试验设计，逐步回归分析，主成分分析，判别分析和聚类分析.

本书以介绍方法为主，内容力求做到简明扼要、清晰易懂，着重讲清数理统计的基本概念、基本原理和计算方法. 在介绍各种基本的数理统计方法的同时，也介绍了一些常用的多元统计分析方法. 在介绍数理统计方法的各章的后面，都设置了一定数量的习题，并在书后附有习题答案.

前 言

数理统计是一门研究如何有效地收集、整理和分析受随机因素影响的数据，对所考察的问题作出推断，进而为制定决策和采取行动提供科学依据的学科. 现在，数理统计方法已经在工业、农业、国防、科研、经济、管理、社会、医学、生物、考古、地质、气象等领域得到了广泛的应用，而且随着计算机的普及和发展，数理统计越来越受到各行各业的普遍重视，应用范围日益扩大，应用水平不断提高，应用成果也层出不穷.

本书是作者根据多年的教学经验，参考国家教育部制定的“工学硕士研究生应用统计课程教学基本要求”，为非数学类专业特别是工科研究生编写的一本数理统计教材. 也可以作为数理统计课学时较少的本科数学专业的教材，还可供从事实际工作的科技工作者和工程技术人员阅读、参考.

考虑到数理统计是一门应用性很强的学科以及非数学专业研究生的特点，本书以介绍方法为主，着重讲清数理统计的基本概念、基本原理和计算方法，省略了一些与实际应用无关的统计理论(如统计量的充分性、完备性、一致最小方差无偏估计、一致最优势检验等)的内容，增加了应用方面的实例，使学生学习后能将统计方法熟练地应用到实际问题中去. 本书在内容叙述、推导证明中，力求做到文字通畅、简明扼要、清晰易懂，便于学生通过阅读，自学掌握所学的内容. 考虑到有一部分研究生未曾学习过作为数理统计基础的概率论，特地增加了第1章“概率论的基础知识”，可以供这一部分学生补习、参考. 在第1章到第6章的后面，都设置了一定数量的习题，书后还附有习题答案或解答提示.

多元统计分析是数理统计中近年来发展特别迅速的一个重要分支，从自然科学到社会科学的许多方面，都日益证实它是一种有效的数据处理方法，在应用中取得了很大的成绩，引起了广泛的注意. 本书的第7章到第9章，简要介绍了4种最常用的多元统计方法——逐步回归分析、主成分分析、判别分析和聚类分

析. 由于多元统计分析需要大量的计算,实际应用中都是靠现成的计算机软件在电脑上完成计算的,所以本书没有详细介绍这些方法的计算步骤和编写计算机程序中的一些处理的细节,而是着重介绍这些方法的基本思想和基本概念,和在实际问题中如何应用,以及如何解释所得到的结果,便于学生和实际工作者学习、使用.

由于编写时间比较仓促,本书内容中难免有疏漏差错之处,欢迎读者指正,以便再版时补充和修正. 作者的电子信箱是: lu _ yuanhong@163. com.

作 者

2005. 7

目 录

1 概率论的基础知识

1.1 随机变量、频率与概率

1.1.1 随机变量

在自然界和人类活动中，我们经常会遇到各种各样的变量. 其中有一些变量，只要在相同的条件下进行同样的观测或实验，它的取值就是完全确定的. 例如，在一个电路中，只要电压和电阻固定不变，电流就是一个确定的值；在真空中，一个从高处落下的自由落体，只要高度和重力加速度保持不变，落地所用的时间就是一个确定的值.

但是，我们有时还会遇到另一种变量. 例如，掷一颗骰子，看掷出的点数，虽然是在同样的条件下掷同一颗骰子，掷出的点数却是不一定的，可以是 1、2、3、4、5 或 6；用一台车床加工零件，测量加工出来零件上的某个尺寸，由于不可避免地存在加工误差，虽然是同一台车床在同样的条件下加工同一种零件，加工出来的尺寸却可能有大有小，并不是确定的.

像这种在同样条件下进行同样的试验，其取值在试验之前无法预先确定的变量，称为**随机变量**. 换句话说，随机变量就是随着试验结果的不同而随机地取各种不同值的变量.

随机变量是概率论和数理统计的主要研究对象. 在概率论和数理统计中，通常将随机变量记为 ξ, η, …(或 X, Y, …).

1.1.2 频率与概率

既然随机变量的取值是随机的、偶然的、无法预知的，那么，是不是它就没有什么规律可以研究了呢？事实上，并不是这样. 如果我们在相同的条件下进行多

次重复试验，随着试验次数的不断增多，就会发现，在它的后面隐藏着某种必然的、规律性的东西.

下面看一个例子.

例 同时扔 2 枚硬币，设 ξ 是出现正面向上的硬币的个数，ξ 的取值事先无法确定，可能是 0、1 或 2，显然，它是一个随机变量.

我们把这种扔硬币的试验重复进行 n 次，设在这 n 次试验中，出现 0、1、2 个正面向上的情形分别为 n_1、n_2、n_3 次. 称 $n_k(k=1,2,3)$ 为**频数**，称 $\frac{n_k}{n}$ $(k=1,2,3)$ 为**频率**. 下表中给出了在不同的试验次数下对频率的统计结果.

n	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\frac{n_3}{n}$
10	0.2	0.7	0.1
100	0.27	0.54	0.19
1 000	0.263	0.492	0.245
10 000	0.247 2	0.504 7	0.248 1
100 000	0.249 48	0.500 24	0.250 28
1 000 000	0.250 001	0.500 012	0.249 987

从表中数据不难看出，随着试验次数 n 的不断增多，频率越来越明显地稳定在一个常数值附近，$\frac{n_1}{n}$ 的稳定值为 $p_1=0.25=\frac{1}{4}$，$\frac{n_2}{n}$ 的稳定值为 $p_2=0.50=\frac{1}{2}$，$\frac{n_3}{n}$ 的稳定值为 $p_3=0.25=\frac{1}{4}$.

其实，对于任何其他的随机变量，我们发现也总是有这样的现象，这种现象称为**频率的稳定性**. 因此，我们有下列定义.

定义 1.1 **设随机变量 ξ 的取值范围可以分成 r（r 可以是可列无穷大）种互不重复的情形，在 n 次试验中，出现第 k 种情形的次数（频数）为 n_k，频率为 $\frac{n_k}{n}$，$k=1,2,\cdots,r$. 随着试验次数 n 的无限增大，频率 $\frac{n_k}{n}$ 会越来越明显地稳定在一个常数值 p_k 附近，这个常数值反映了随机变量 ξ 取值为这种情形的可能性的大小，我们称它为概率.**

在概率统计中，通常用 $P\{\xi=k\}$ 表示“ξ 的值等于 k 的概率”，用

$P\{a<\xi\leqslant b\}$ 表示“ξ的值大于 a 小于等于 b 的概率”,…….

1.2　概率的基本性质

下面不加证明地列出概率的一些重要的基本性质.

性质 1　对任何概率 p,必有 $0\leqslant p\leqslant 1$.

性质 2　设随机变量 ξ 的取值范围可以分成 r(r 可以是可列无穷大)种互不重复的情形,ξ 的值属于第 k 种情形的概率为 p_k, $k=1,2,\cdots,r$,则必有 $\sum\limits_{k=1}^{r}p_k=1$.

性质 3　互逆(即正好相反)的两个事件的概率加起来等于 1.

例如,有 $P\{\xi>a\}=1-P\{\xi\leqslant a\}$, $P\{|\xi|\leqslant a\}=1-P\{|\xi|>a\}$,….

性质 4　设 ξ, η 是两个相互独立(即它们的取值互不影响)的随机变量,则它们(在点或在区间上)同时取值的概率,等于它们各自取这些值的概率的乘积,例如

$$P\{\xi=i,\ \eta=j\}=P\{\xi=i\}P\{\eta=j\},$$

$$P\{\xi\leqslant x,\ \eta\leqslant y\}=P\{\xi\leqslant x\}P\{\eta\leqslant y\},$$

$$P\{\xi>x,\ \eta>y\}=P\{\xi>x\}P\{\eta>y\},$$

……,对于多个相互独立的随机变量,这个性质同样也是成立的.

1.3　离散型随机变量的概率分布

1.3.1　离散型随机变量的概率分布的定义

有些随机变量,只能在离散点上取值,例如,掷一个骰子掷出的点数,同时扔两枚硬币出现正面向上的硬币个数,这种随机变量称为**离散型随机变量**.

定义 1.2　设 ξ 是离散型随机变量,将 ξ 可能取的所有的值以及它取这些值的概率一一列举出来,这样得到的一组概率,称为 ξ 的概率分布(或分布列,分布律).

1.3.2　常见的离散型随机变量的概率分布

下面介绍几个常见的离散型随机变量的概率分布.

1. 离散均匀分布

如果 ξ 可能取的值为 1, 2, …, r,其中 r 是一个正整数,ξ 的概率分布为

$$P\{\xi = k\} = \frac{1}{r},\ k = 1, 2, \cdots, r,$$

则称 ξ 服从参数为 r 的**离散均匀分布**.

例 1 一颗均匀的骰子掷出的点数 ξ,它的概率分布为

$$P\{\xi = k\} = \frac{1}{6},\ k = 1, 2, \cdots, 6,$$

ξ 服从的就是 $r = 6$ 的离散均匀分布.

2. 二项分布

如果 ξ 可能取的值为 0, 1, 2, …, n,其中 n 是一个正整数,$0 < p < 1$ 是一个常数,ξ 的概率分布为

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \cdots, n,$$

则称 ξ 为服从参数为 n, p 的**二项分布**,记为 $\xi \sim b(n, p)$.

例 2 同时扔 n 枚均匀的硬币,ξ 为这 n 个硬币中出现正面向上的硬币个数,ξ 的分布就是参数为 n 和 $p = \frac{1}{2}$ 的二项分布,即有 $\xi \sim b\left(n, \frac{1}{2}\right)$,$\xi$ 的概率分布为

$$P\{\xi = k\} = C_n^k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} = \frac{C_n^k}{2^n},\ k = 0, 1, \cdots, n.$$

在二项分布中,如果 $n = 1$,即当 $\xi \sim b(1, p)$ 时,ξ 的概率分布简化为

$$P\{\xi = k\} = p^k (1-p)^{1-k},\ k = 0, 1.$$

这时,ξ 的值只能取 0 或 1,$P\{\xi = 0\} = 1 - p$,$P\{\xi = 1\} = p$,我们将 ξ 的分布称为 **0—1 分布**(或**两点分布**,或 **Bernoulli 分布**).

3. 几何分布

如果 ξ 可能取的值为 1, 2, 3, …, $0 < p < 1$ 是一个常数,ξ 的概率分布为

$$P\{\xi = k\} = (1-p)^{k-1} p,\ k = 1, 2, 3, \cdots,$$

则称 ξ 服从参数为 p 的**几何分布**,记为 $\xi \sim g(p)$.

例 3 向同一目标连续射击,直到击中目标为止.设每次射击的命中率为

p，ξ是到击中目标为止所需要的射击次数，ξ服从的就是参数为 p 的几何分布，即有 $\xi\sim g(p)$，

$$P\{\xi=k\}=(1-p)^{k-1}p,\ k=1,2,3,\cdots.$$

这从直观上很容易理解：到第 k 次射击才击中，一定是前 $k-1$ 次都没有击中，这样的概率为 $(1-p)^{k-1}$，最后一次击中，概率为 p，把它们全部乘起来就是 $(1-p)^{k-1}p$.

例 4 设袋中有 10 个球，其中 3 个是红球，7 个是白球. 从中每次任意取一个球，取后仍放回. 设 ξ 是取到红球为止所需的取球次数，则有

$$P\{\xi=k\}=\left(\frac{7}{10}\right)^{k-1}\times\frac{3}{10},\ k=1,2,3,\cdots,$$

即 $\xi\sim g\left(\frac{3}{10}\right)$.

4. Poisson(普阿松，泊松)分布

如果 ξ 可能取的值为 $0,1,2,\cdots,\lambda>0$ 是一个常数，ξ 的概率分布为

$$P\{\xi=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,1,2,\cdots,$$

则称 ξ 服从参数为 λ 的 **Poisson 分布**，记为 $\xi\sim P(\lambda)$.

例如，单位时间内打来的电话数，单位时间内来到的顾客数，单位时间内观测到的放射性粒子数，一年内发生的交通事故数，每页书中的排版印刷错误数，……，都服从或近似服从 Poisson 分布，参数 λ 就是 ξ 的平均值.

实际上，还可以有其他各种不同形式的概率分布.

有时，我们会遇到这样的问题：只知道一个概率分布的形式，但分布中有些常数却不知道，需要我们来确定.

例 5 设 ξ 的概率分布为

$$P\{\xi=k\}=\frac{k}{A},\ k=1,2,\cdots,r,$$

其中，r 是一个已知的正整数，A 是未知常数，求 A.

解 根据前面介绍过的概率的性质 $\sum\limits_{k=1}^{r}p_k=1$，得

$$1=\sum_{k=1}^{r}P\{\xi=k\}=\sum_{k=1}^{r}\frac{k}{A}=\frac{1+2+\cdots+r}{A}=\frac{\frac{r(r+1)}{2}}{A},$$

所以，$A=\dfrac{r(r+1)}{2}$，即

$$P\{\xi=k\}=\frac{k}{A}=\frac{2k}{r(r+1)},\ k=1,2,\cdots,r.$$

1.4 连续型随机变量的分布函数和概率密度

1.4.1 连续型随机变量的分布函数和概率密度的定义

有些随机变量，它们的取值范围是实数轴上的连续区间，例如，加工零件时的加工误差，炮弹落点到目标的距离，两次电话打来之间的时间间隔，它们都在连续区间上取值，这种随机变量称为**连续型随机变量**.

连续型随机变量的取值不可能一一列举出来，所以，不能用概率分布的形式给出它的分布. 要表达它的分布，必须采取其他的形式.

定义 1.3 设 ξ 是一个连续型随机变量，称

$$F(x)=P\{\xi\leqslant x\},\ -\infty<x<+\infty,$$

为 ξ **的分布函数**(有些书上，将分布函数定义为 $F(x)=P\{\xi<x\},\ -\infty<x<+\infty$).

如果存在一个函数 $\varphi(x)$，使得

$$F(x)=\int_{-\infty}^{x}\varphi(t)\mathrm{d}t,\ -\infty<x<+\infty,$$

则称 $\varphi(x)$ **是** ξ **的概率分布密度函数**，简称**概率密度**(或**分布密度**，或**密度函数**).

概率密度 $\varphi(x)$ 的大小反映了 ξ 在 x 的邻域内取值的概率的大小.

下面不加证明地给出连续型随机变量的分布函数 $F(x)$ 和概率密度 $\varphi(x)$ 的一些性质.

性质 1 $F(x)$ 是单调非降连续函数，$F(-\infty)=0$，$F(+\infty)=1$.

性质 2 在 $\varphi(x)$ 的连续点上，有 $\varphi(x)=\dfrac{\mathrm{d}}{\mathrm{d}x}F(x)$.

性质 3 $\varphi(x)\geqslant 0,\ -\infty<x<+\infty$.

性质 4 $\displaystyle\int_{-\infty}^{+\infty}\varphi(x)\mathrm{d}x=1$.

性质 5 对于任何实数 $a\leqslant b$(a，b 也可以是无穷大)，有

$$P\{a<\xi<b\}=P\{a\leqslant\xi<b\}=P\{a\leqslant\xi\leqslant b\}=P\{a<\xi\leqslant b\}$$

$$= \int_a^b \varphi(x)\mathrm{d}x = F(b) - F(a).$$

例 1 设 ξ 的概率密度为 $\varphi(x) = \begin{cases} \dfrac{A}{(1+2x)^2}, & \text{当 } x > 0 \text{ 时;} \\ 0, & \text{当 } x \leqslant 0 \text{ 时.} \end{cases}$ 其中 A 是未知常数.

求：(1) 常数 A；

(2) 概率 $P\{\xi \geqslant 2\}$.

解 (1) 根据概率密度的性质，得

$$1 = \int_{-\infty}^{+\infty} \varphi(x)\mathrm{d}x = \int_{-\infty}^{0} 0\mathrm{d}x + \int_0^{+\infty} \frac{A}{(1+2x)^2}\mathrm{d}x$$

$$= 0 + \frac{A}{2}\int_0^{+\infty} \frac{1}{(1+2x)^2}\mathrm{d}(1+2x) = \frac{A}{2}\left(-\frac{1}{1+2x}\right)\Big|_0^{+\infty}$$

$$= \frac{A}{2}\left(-0 + \frac{1}{1+0}\right) = \frac{A}{2}.$$

所以 $A = 2$，即有 $\varphi(x) = \begin{cases} \dfrac{2}{(1+2x)^2}, & \text{当 } x > 0 \text{ 时;} \\ 0, & \text{当 } x \leqslant 0 \text{ 时.} \end{cases}$

(2) $P\{\xi \geqslant 2\} = P\{2 \leqslant \xi < +\infty\} = \int_2^{+\infty} \varphi(x)\mathrm{d}x = \int_2^{+\infty} \frac{2}{(1+2x)^2}\mathrm{d}x$

$$= \int_2^{+\infty} \frac{1}{(1+2x)^2}\mathrm{d}(1+2x) = -\frac{1}{1+2x}\Big|_2^{+\infty} = -0 + \frac{1}{1+4}$$

$$= \frac{1}{5}.$$

1.4.2 常见的连续型随机变量的概率密度

下面介绍几种常见的连续型随机变量的概率密度.

1. 均匀分布

如果 ξ 的概率密度为(其中 $a < b$ 为常数)

$$\varphi(x) = \begin{cases} \dfrac{1}{b-a}, & \text{当 } a \leqslant x \leqslant b \text{ 时;} \\ 0, & \text{其他.} \end{cases}$$

则称 ξ 服从区间 $[a, b]$ 上的(即参数为 a, b)的**均匀分布**,记为 $\xi \sim U(a, b)$.

均匀分布 $U(a, b)$ 的分布函数为

$$F(x)=\begin{cases}0, & \text{当 } x \leqslant a \text{ 时;}\\ \dfrac{x-a}{b-a}, & \text{当 } a<x \leqslant b \text{ 时;}\\ 1, & \text{当 } x>b \text{ 时.}\end{cases}$$

例如,公交车固定每隔 5 min 来一辆,乘客可能在这 5 min 内的任一时刻到达车站,乘客的等车时间就服从区间 $[0, 5]$ 上的均匀分布.

2. 指数分布

如果 ξ 的概率密度为(其中 $\lambda>0$ 为常数)

$$\varphi(x)=\begin{cases}\lambda e^{-\lambda x}, & \text{当 } x>0 \text{ 时;}\\ 0, & \text{当 } x \leqslant 0 \text{ 时.}\end{cases}$$

则称 ξ 服从参数为 λ 的**指数分布**,记为 $\xi \sim E(\lambda)$.

指数分布 $E(\lambda)$ 的分布函数为

$$F(x)=\begin{cases}1-e^{-\lambda x}, & \text{当 } x>0 \text{ 时;}\\ 0, & \text{当 } x \leqslant 0 \text{ 时.}\end{cases}$$

例如,两个电话打来之间的时间间隔长度,两次交通事故发生之间的时间间隔长度,都服从指数分布.

3. 正态分布

如果 ξ 的概率密度为(其中 $\mu, \sigma>0$ 为常数)

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty<x<+\infty,$$

则称 ξ 服从参数为 μ, σ 的**正态分布**,记为 $\xi \sim N(\mu, \sigma^2)$.

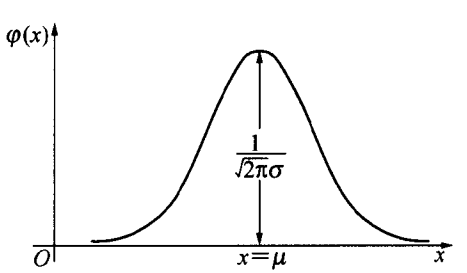

图 1-1

正态分布 $N(\mu, \sigma^2)$ 的概率密度的图像如图 1-1 所示,且有以下性质:

(1) $\varphi(x)$ 关于 $x=\mu$ 左右对称.

(2) 当 $x \to \pm\infty$ 时, $\varphi(x) \to 0$.

(3) 当 $x=\mu$ 时, $\varphi(x)$ 取到最大值,最大值为 $\dfrac{1}{\sqrt{2\pi}\sigma}$.