应用数理统计

韓於羹 编

北京航空航天大學出版社

应用数理统计

韓於羹 編

北京航空航天大學出版社

内 容 提 要

本书是为工科研究生编写的教材。内容包括数理统计的基本概念、估计理论、假设检验、回归分析、方差分析、正交设计、多元正态分布的估计与检验、判别分析、多元相关。本书对数理统计中的重要内容及近代成果都作了扼要论述。每章都有适量例题及习题并附有解答或提示，以帮助读者对概念及方法的理解和应用。

应用数理统计

YINGYONG SHULI TONGJI

韓於羹 编

責任編輯 郭維烈

北京航空航天大学出版社出版

新华书店总店科技发行所发行 各地新华书店经售

北京航空航天大学印刷厂印装

797×1092 1/32 印张：13.75 字数：320千字

1989年3月第一版 1989年3月第一次印刷 印数：4500册

ISBN 7-81012-094-8/O·009 定价：2.45 元

前　　言

本书是根据我过去几年在北航研究生院为工科研究生讲授《应用数理统计》的讲稿修改扩充而成的。由于是以工科研究生为对象，本书着重介绍数理统计的各种方法，应用条件以及结果的含义，而不追求数学上的严密性与完整性。书中给出了一些必要的数学推导，同时对未给出推导的结论往往指明出处。全书叙述力求通俗易懂，便于读者自学。

由于目前尚未有正式出版的适合工科研究生用的数理统计教材，为解燃眉之急，匆匆成章，一定存在着许多疏漏或错误，敬請批评指正。

在编写过程中，参阅了国内外有关书籍，引用了一些例子，恕不一一指明出处，在此一并向有关作者致谢！

本书得以完成，要感谢北航研究生院的大力支持和鼓励以及北航出版社的通力协作，特别是郭维烈同志为此书付出了大量的辛勤劳动，在此一并致谢。

韩於羹

一九八八年七月

目　　录

第四章　回归分析

第五章　方差分析

第六章　正交设计

第七章　多元正态分布的参数估计与检验

第八章　判别分析

第一章 概 论

§1 数理統計学的基本問題

数理统计学是应用概率论的基本理论，根据试验或观察得到的数据，对研究对象的客观规律性作出种种合理的估计和判断。

数理统计学不同于一般的资料统计，它侧重于应用随机现象本身的规律性来研究问题，研究如何合理地获得数据资料，研究如何利用有限资料对所关心的问题做出尽可能精确、可靠的结论。

由于随机现象的统计规律性必然在大量重复试验中呈现出来，因而就理论上讲，只要对随机现象进行足够多次的观察便能揭示出规律性。然而，实际上所允许的观察或试验往往只能是有限的，甚至是少量的。例如，我们不能把所有生产出的炮弹都进行试射，以研究该批炮弹的性能，而只能抽取其中一部分或几枚炮弹来进行试验，获取数据，并通过这些数据来对所研究的全部炮弹的性能做出估计和推断。通常将从所研究对象的全体中抽取一部分进行试验或观察，并获取试验数据的工作称为抽样或采样。由于抽样所得的数据不能包含研究对象的全部信息，因而所得结论必然含有不确定性，对这种不确定性的度量就需要以概率论作为工具。数理统计学的基本任务就是：研究以有效的方式收集、整理和分析受到随机性影响的数据，并对所观察的问题作出推测和判

断，对采取的决策提供依据或建议。

数理统计研究的内容在日益扩大，应用领域也愈来愈广泛，几乎在人类活动的一切领域中都能程度不同地找到它的应用。数理统计方法可以涉及到几乎所有科学技术领域，工农业生产和国民经济的各个部门，诸如在气象预报、水文预报、地震预报、质量控制、抽样验收、可靠性分析、遗传工程、以及机器制造、土木建筑、国防、地质、交通、化工、纺织、冶金、医药卫生等领域中，数理统计都得到了广泛地应用。就数理统计的内容来说，可以概括地分为试验设计和统计推断两个方面，二者有密切的联系，本讲义限于篇幅着重介绍统计推断的有关问题。

§2 基本概念

一、总体、样本、简单随机样本

实践中经常会遇到统计推断的问题。例如，有一批灯泡，要以“使用寿命”这个指标来衡量其质量。若规定，寿命低于1000小时者为次品，问如何确定这批灯泡的次品率？显然，不能通过对每只灯泡进行寿命试验来计算次品率，否则一旦试验结束，产品也就不复存在了。因此，只能从整批灯泡中选取一些灯泡做寿命试验，并记录下试验结果，然后根据这组数据来推断整批灯泡的寿命情况，回答上面提出的问题。

在统计学中，总体与样本是两个重要概念。通常把研究对象的全体称为**总体**（或母体），而把组成总体的每个基本单元称为**个体**。例如上面提到的那批灯泡的全部就组成了总

体，而其中每个灯泡就是个体。不过，在实际中，我们关心的常常是研究对象的某个指标X（例如灯泡寿命、晶体管的直流放大系数、钢筋的强度等等），它是一个随机变量。因此，总体通常又是指某个随机变量X取值的全体，其中的每一个个体皆是实数。当所研究的指标不止一个时（例如，对于某校的学生，我们既要研究其身高的情况，又要研究其体重的情况）那末又可以分成几个总体来研究。如果表征总体的随机变量X的分布函数为$F(x)$，我们也称总体$F(x)$。

总体的类型随研究的问题不同而改变。例如，研究某灯泡厂1月份生产的某型号灯泡的次品率，总体即是1月份所生产的该种灯泡的全体，是一个有限总体，而灯泡的寿命X是离散型的。但有时为了研究方便，又通常把相同条件下所有可能生产的灯泡寿命全体，看成一个总体，这样理想化了的灯泡寿命X则成了一个连续型随机变量。而处理连续型随机变量就比处理离散型随机变量更加方便。今后，凡是提到总体就是指一个随机变量，一个带有确定概率分布的随机变量。通常用大写字母X，Y，Z等表示总体。

为了获得总体的分布，就必须对总体进行抽样观察。所谓抽取一个个体，就是做一个随机试验，并记录其试验结果。如果进行了n次抽样观察，就得到总体X的一组数值$(x_1, x_2, \cdots, x_n)$，其中x_i是第i次抽样观测的结果。我们自然希望这一组抽样观察结果能很好地反映总体的情况，这就要对抽取个体的方法加上一定限制。容易想到，如果总体中每个个体被抽到的机会是均等的，并且在抽取一个个体后总体的成分不改变，那么，抽得的个体就能很好地反映总体的情况。基于这种想法的抽取个体的方法称为**简单随机抽样**。抽得的一组个体称为一组**样本观察值**。换句话说，

简单随机抽样就是独立地、重复地做随机试验，从而能使试验结果既具有独立性，又有代表性，十分便于理论分析处理。通常把简单随机抽样的结果（x_1，x_2，…，x_n）称为总体X的一组样本观察值。

抽样，就具体方法而论，好比在袋中摸球，有放回抽样与不放回抽样之分。对有限总体，使用放回抽样方法，即：将每个个体看成一个球，放在一个"袋中"，每次随机地摸取一只，并记录其结果，得到一个样本观察值，然后放回，将球搅匀，再摸，再记录观察结果。如此重复摸n次就得到一组样本观察值（x_1，x_2，…，x_n）。对于无限总体，放回与否并不改变总体成分，因而可使用不放回抽样方法，这样会更方便些。在实际中，即使总体个数N有限，但只要被抽取的个体数n与N之比小于0.01，仍然采取不放回抽样，所得样本性质近似于简单随机抽样的结果。还可以借助"随机数表"来抽样，使所得的个体具有随机性。今后，如无特别说明，抽样皆指简单随机抽样。

设（x_1，x_2，…，x_n）为总体X的一组样本观察值，由于抽样的随机性与独立性，每个x_i都可以看作某一个随机变量X_i（$i=1, 2, \cdots, n$）所取的观察值。这里X_1，X_2，…，X_n相互独立，且皆与X具有相同分布。因此，（x_1，x_2，…，x_n）又可以看作n维向量(X_1，X_2，…，X_n）的观察值或一个实现。就一次观察结果而论，（x_1，x_2，…，x_n）是完全确定的一组值，但它又随每次抽样观察而改变，具有随机性。这就是说抽样结果具有两重性。它既有随机性的一面，而对具体的一次抽样又是一组确定的值。下面我们将以上讨论的概念用定义和定理的形式表述如下。

定义2.1 设X为具有分布函数$F(x)$的随机变量，若X_1，X_2，…，X_n为具有相同分布函数$F(x)$的相互独立的随机变量，则称（X_1，X_2，…，X_n）为来自总体X的容量为n的**简单随机样本**，简称**样本**（或子样），它们的观察值（x_1，x_2，…，x_n）又称为X的n个**独立的观察值**。

随机向量（X_1，X_2，…，X_n）所可能取值的全体（这儿是n维空间或其中一个子集）称为**样本空间**，一个样本观察值（x_1，x_2，…，x_n）就是样本空间中的一个点。

定理2.1 若（X_1，X_2，…，X_n）是来自总体X（具有分布函数$F(x)$）的样本，则（X_1，X_2，…，X_n）的联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n)=\prod_{i=1}^{n} F(x_i)$$

又若X具有概率密度函数$f(x)$，则（X_1，X_2，…，X_n）的联合概率密度函数为

$$f^*(x_1, x_2, \cdots, x_n)=\prod_{i=1}^{n} f(x_i)。$$

二、统计量和样本矩

样本是总体的代表和反映，但要想使样本能推断我们所关心的问题，还必需对所抽取的样本进行“加工”和“提炼”，把所需的信息集中起来。为此，针对不同的问题对样本需要做一些运算，即构造样本的某种函数。这种函数被称为统计量。

定义2.2 设（X_1，X_2，…，X_n）为总体X的一个样本，$g(X_1,X_2,\cdots,X_n)$为一连续函数，如果g中不包含任

何未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 为一个统计量。

例如，$X \sim N(\mu, \sigma^2)$，其中 μ 已知，σ^2 未知。而 (X_1, X_2) 是从 X 中抽取的一个样本，则 $\frac{1}{2}(X_1+X_2)$ 是统计量。$\frac{1}{2}\sum_{i=1}^{2}(X_i-\mu)^2$ 是统计量。但 $\frac{X_1-\mu}{\sigma}$ 却不是统计量。

由于样本 $(X_1, X_2, \cdots, X_n)$ 是随机向量，故统计量是一个随机变量。常用的统计量为样本矩。诸如：

样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$；

样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$；

样本 k 阶原点矩 $M_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$，$k=1, 2, \cdots$；

样本 k 阶中心矩 $M_k'=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$，

$k=1, 2, \cdots$；

其中 $\bar{X}$ 和 S^2 是两个特别重要的统计量。显然 $M_1=\bar{X}$，$M_2'=\frac{n-1}{n}S^2$，即样本二阶中心矩与样本方差略有不同。通常又记 M_2' 为 $\tilde{S}^2$。

三、顺序统计量和经验分布函数

设 $(X_1, X_2, \cdots, X_n)$ 是从总体 X 中抽取的一个样本。记 $(x_1, x_2, \cdots, x_n)$ 为样本的一个观察值，将观察值的各个分量按递增次序重新排列为：

$$x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$$

当 $(X_1, X_2, \cdots, X_n)$ 取值为 $(x_1, x_2, \cdots, x_n)$ 时，我们定义 $X_k^{(n)}$ 取值为 x_k^*，称由此得到的 $X_1^{(n)}, X_2^{(n)}, \cdots, X_n^{(n)}$ 为 $(X_1, X_2, \cdots, X_n)$ 的一组顺序统计量。显然，$X_1^{(n)} \leqslant X_2^{(n)} \leqslant \cdots \leqslant X_n^{(n)}$，且 $X_1^{(n)} = \min\limits_{1 \leqslant i \leqslant n} X_i$，$X_n^{(n)} = \max\limits_{1 \leqslant i \leqslant n} X_i$。

记

$$F_n^*(x) = \begin{cases} 0, & \text{当} x < x_1^*, \\ \dfrac{k}{n}, & \text{当} x_k^* \leqslant x < x_{k+1}^*, \ k = 1, 2, \cdots, n-1. \\ 1, & \text{当} x \geqslant x_n^*. \end{cases}$$

称 $F_n^*(x)$ 为总体 X 的经验分布函数。显然 $0 \leqslant F_n^*(x) \leqslant 1$，且是单调、非减及右连续函数，在 $x = x_k^*$ 处间断，且在每个间断点上跳跃都是 $\frac{1}{n}$。亦即 $F_n^*(x)$ 具有分布函数的性质。事实上，$F_n^*(x)$ 是在 n 次重复独立试验中，事件 $\{X \leqslant x\}$ 的频率。显然，对于不同的样本观察值，得到的经验分布函数也不相同，所以对于 x 的每一个数值，$F_n^*(x)$ 是一个随机变量。

由大数定律知道，在满足一定的条件下，事件发生的频率依概率收敛于这个事件发生的概率。人们自然要问，总体 X 的经验分布函数 $F_n^*(x)$，即事件 $\{X \leqslant x\}$ 发生的频率，当 n 足够大时，是否也渐近于事件 $\{X \leqslant x\}$ 发生的概率（即总体 X 的分布函数 $F(x)$）呢？格里文科（В.И. Гливенко）于1933年作了肯定的回答。

定理2.2（格里文科定理） 当 $n \to \infty$ 时 $F_n^*(x)$ 依概率1关于 x 均匀地收敛于 $F(x)$，即

$$P\{\lim_{n\to\infty}\sup_{-\infty<x<+\infty}|F_n^*(x)-F(x)|=0\}=1$$

(证明参看[波兰]M·费史著，王福保译《概率论与数理统计》一书345页)

格里汶科定理揭示了随机变量X的经验分布函数$F_n^*(x)$与理论分布函数$F(x)$之间的内在联系，并且指出，当样本容量n足够大时，从样本算得的经验分布函数$F_n^*(x)$与理论分布函数$F(x)$之间的差别可以任意地小，也就是说，当n足够大时，样本分布函数$F_n^*(x)$与总体分布函数$F(x)$相差最大处也会足够地小，这就是我们之所以可以用样本推断总体的基本理论依据。

经验分布函数$F_n^*(x)$与样本矩之间有下列关系：设$(x_1, x_2, \cdots, x_n)$为样本观察值，则有

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\int_{-\infty}^{+\infty}x\,dF_n^*(x),$$

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\int_{-\infty}^{+\infty}(x-\bar{x})^2dF_n^*(x),$$

$$m_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k=\int_{-\infty}^{+\infty}x^k dF_n^*(x),\quad k=2,3,\cdots,$$

$$m_k'=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k=\int_{-\infty}^{+\infty}(x-\bar{x})^k dF_n^*(x),$$

$k=2, 3, \cdots$。

§3 抽样分布

统计量是我们对总体X的分布函数或数字特征进行估计与推断的最主要的基本概念，求出统计量$g(X_1, X_2, \cdots,$

X_n）的分布函数是数理统计学的基本问题之一。统计量的分布，称为**抽样分布**。

设总体X的分布函数$F(x)$已知，对于任一自然数n，如能求出给定的统计量$g(X_1, X_2, \cdots, X_n)$的分布函数，就称这分布为统计量g的**精确分布**，这对于数理统计学中的所谓**小样问题**（即在样本容量n比较小的情况下所讨论的各种统计问题）的研究是很重要的。

一般说来，要确定一个统计量的精确分布是比较困难的。然而，对于一些重要的特殊情形，如总体X服从正态分布时，这个问题已取得了很好的结果。

若统计量$g(X_1, X_2, \cdots, X_n)$的精确分布求不出来，或其表达式非常复杂而难于应用，通常就设法求出它在$n\to\infty$时的极限分布，这对于数理统计学中的所谓**大样问题**（即在样本容量n比较大的情况下讨论的各种统计问题）的研究很有用。自然在用极限分布时，样本的容量n要比较大才行。不过大和小也是一个相对概念，并没有一般的标准来规定何为比较大或比较小。

下面我们介绍n个正态总体的统计量的精确分布，它们在估计理论，假设检验及方差分析等数理统计学的基本内容中都有重要作用。

一、正态总体样本的线性函数的分布

定理3.1 设总体$X\sim N(\mu, \sigma^2)$，$(X_1, X_2, \cdots, X_n)$是总体的容量为n的简单随机样本，令：

$$U=a_1X_1+a_2X_2+\cdots+a_nX_n \qquad (3.1)$$

其中$a_1, a_2, \cdots, a_n$是已知常数，则U也是正态随机变量，其均值、方差分别为