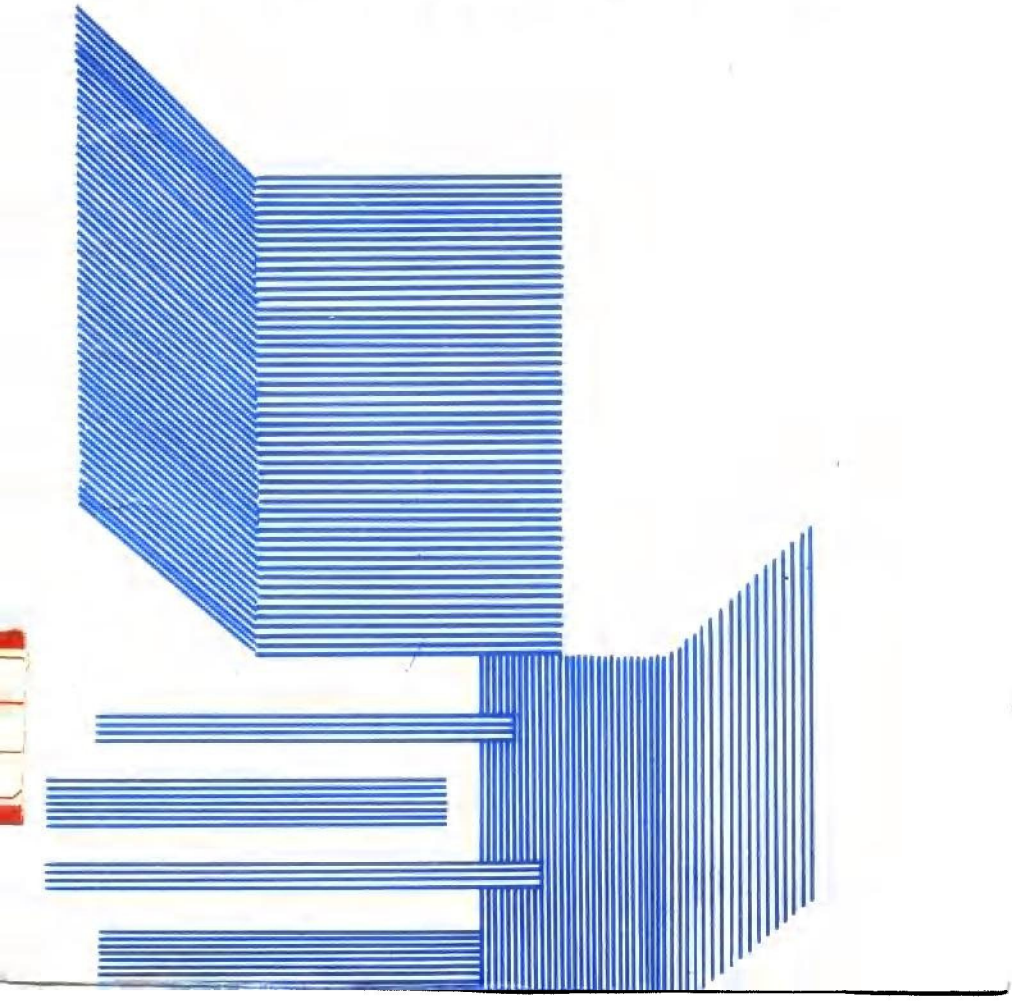
● 研究生教材 ● 研究生教材

应用数理统计

吴　翊　李永乐　胡庆军　编著



■ 研究生教材 ■

吴 翊 李永乐 胡庆军 编著

应用数理统计

国防科技大学出版社

内 容 提 要

本书是根据教委新颁布的“工学硕士研究生应用统计课程基本要求”编写的。除包括“基本要求”所规定的内容，即基本概念，参数估计，假设检验，回归分析，方差分析，正交设计以外，还补充了实用多元统计分析的若干内容。注重阐明统计思想，突出统计方法介绍，强调统计实际应用，是本书的特色所在；并通过诸如材料的选择安排，问题的引入，内容的阐述，例题和习题的配置等环节体现上述特色。全书论述严谨，行文深入浅出，富有启发性。为便于读者自学，书末还以附录形式给出了有关概率论等基础知识。本书适宜作为工学硕士研究生应用统计课程的教材，也可作为大学本科生及教师的教学用书，对于广大工程技术人员亦是一本不可多得的参考书。

应 用 数 理 统 计

吴 翊等 编著

责任编辑 张建军

*

国防科技大学出版社出版发行

邮编：410073 电话：4555681

新华书店总店北京发行所经销

国防科技大学印刷厂印装

*

850×1168 毫米 1/32 印张：14.1875 字数：356 千字

1995 年 8 月第 1 版第 1 次印刷 印数：2700 册

ISBN 7—81024—334—9

O·43 定价 14.00 元

序

数理统计对于许多工科专业的研究生是一门十分重要的课程，它既是众多专业的数学基础，又能直接提供某些实用数学方法。这门课教学质量的提高，能使相当多数的学生受益匪浅。为此，我们感到有必要编写一本适合工科研究生的数理统计教材。

编写这本教材的主要指导思想有三点：第一，它应覆盖国家教委工科研究生数学课程指导小组制定的“工学硕士研究生应用统计课程教学基本要求”（以下简称“基本要求”）所规定的内容。特别对于像贝叶斯估计、双因素实验数据的方差分析、正交试验设计等许多同类教材未编入或未全编入的章节，我们按照“基本要求”作了安排。第二，除体现“基本要求”的精神外，还应适当包括近代数理统计中应用性强的内容，以满足各类学生特别是在本科阶段已学过概率统计课程的这部分学生的需要。为此，我们编入了应用面较广的多元统计分析一章，并在其它各章节也适当地充实了一些应用性较强的内容，如估计量评价准则的补充，样本容量问题，非参数检验的若干方法等。第三，整个教材的材料安排，行文风格应适应工科研究生的特点。对于工科学生主要是掌握数理统计的基本概念、基本原理和基本方法，特别是能在实际问题中灵活应用数理统计知识。因此我们在阐述某一统计概念方法时，一般先提出问题的实际背景，并较多地采用例题的形式，以使学生一开始便带着实际问题学习思考。这些例题除了展示通常的数理统计方法以外，还为统计建模做出了示范。在行文上，我们力求生动，以增加其可读性。为避免枯燥地罗列定义、定理，我们把许多定义融合在叙述中，这并不失数学本身的严密性。对于

大多数数学推导，都给得较为详尽，以方便自学之用。

书末给出三个附录，给未学过概率论等基础课程的学生备查。每章配有习题，其中部分习题是需要上机计算的，书末给出了部分参考答案。

全书共分七章，由吴翊主持编写，具体分工是：第一、二章及第三章 3.1～3.4 节由吴翊撰写；第三章 3.5 节及第四、五章由李永乐撰写；第六、七章由胡庆军撰写。时间仓促，加之作者才疏学浅，纰误之处，在所难免，诚望读者提出宝贵意见。

本书始终是在我校研究生院和我系大力支持、资助下完成的。特别是骆志刚同志对此书出版做了许多工作，在此深表谢意。

作　者

1994 年 9 月于长沙

目　录

第一章　数理统计的基本概念

1.1　导　言

什么是数理统计？它的主要研究对象和任务是什么？这大概是每位初次接触这门课程的读者所关心的。一般地说数理统计这一学科研究对象是带随机性影响的数据，任务是如何有效地收集、整理、分析这些数据，对所观察的现象作出推断或预测，直到为采取决策提供依据。下面我们将通过例子，对此加以逐步解释。

说到“统计”这个名词，大家在日常工作和生活中并不陌生：工厂里，每月的产量、产值、成本、利润等各项经济指标完成情况要用到统计；政府部门中，本地区的产业、财金、文教、人口等各项资料的汇总要用到统计……。事实上，“统计学”一词的英文 statisics 源于拉丁文的 status（国家），意思为国情资料的收集或国情学，这似乎与我们通常说的统计是一致的。但要指出的是，“数理统计”并不完全是上述通常意义的统计。尽管“数理统计”的应用对象也可以是国情资料或者是工厂的生产情况的收集处理等，但就其学科实质来说是有不同的。什么是数理统计的研究范畴呢？先考察下面的例子。

[**例 1.1**]　试完成下列工作：

(i) 汇总某车间当月生产的某产品的件数、废品数，并算出当月的废品率。

(ii) 从车间当月生产的产品中随机抽取（今后再具体解释“随机抽取”的含义）出几件产品，查出其中废品数，并由此对车间当月的废品率做个推测。

[例 1.2] 试完成下列工作：

(i) 普查某地区的总人数及患某种疾病的人数，算出该地区该疾病的患病率，并与全国平均患病率（设有已知资料）比较。

(ii) 从某地区中随机抽查 n 个人，查出其中患某种疾病的人数，并推断出该地区该疾病的患病率是否明显高于全国平均水平。

上述两个例子中，每个例子的第 (i) 项工作都是通常的统计工作，第 (ii) 项工作则都属于数理统计的研究范畴。那么每个例子的两件工作究竟有什么异同呢？首先共同的是两件工作都是对数据的采集和处理，不同的是在第 (i) 项工作中，数据是确定无误的，处理方法是既定的，其结果也是唯一准确的；而在第 (ii) 项工作中，数据则要受到偶然因素即随机性的干扰，处理的结果或结论既与数据的采集方法有关，又与处理方法有关，因而结论也未必是完全准确的。如在例 1.2 (ii) 中，要在该地区随机地抽取 n 个人，抽哪 n 个人所得到的患者人数不可能都一样，由此来推断全地区的患病率当然也会不一样，况且还与推断方法有关，因此结果并不是完全准确的，它毕竟不是该地区真正的患病率，从而最终得出的患病率是否明显高于全国平均水平的结论也不是完全准确的。

为什么需要数理统计？例如在前述两例中，是否有必要做第 (ii) 项工作？有必要，因为在实际问题中，我们可能无法完成类似于第 (i) 项的工作，这里的原因是多方面的。比方说在例 1.2 中，限于财力、人力和时间，我们无法做第 (i) 项普查，不得不用第 (ii) 项工作来取代。当然这样做的代价是无法得到完全准确的结论，只能得到在某种意义下正确的推断，幸而这在很多场合下已经很充分了。又比方说在例 1.1 中，若对产品的检验是破坏性的，(如产品是弹药，必须通过击发检验是否是废品)，这在实际上第 (i) 项工作是没有意义的，也只能通过第 (ii) 项工作来对其废品率做个估算。

在实际问题中，数据的随机性通常是无法避免的。它的来源大致有两个方面：一是由“偶然性”带来的。这类问题往往是研究的对象数量很大，不可能或者没有必要对它们全部加以考察，只能抽取一部分来加以研究；尽管从抽取方式来说应力求能较全面反映全部对象的信息，但由于只是抽取其中的一部分，就难免有偶然性。如例 1.2 中的第 (ii) 项工作中的数据就是属于带有偶然性的数据。另一方面，则是由“不确定性”带来的，例如在产品生产中即使使用同样的材料、设备、工艺流程，所生产的产品质量仍然有差异，这是因为上述条件只是看起来完全一样，实际上总是有一些因素无法控制或不便控制，这就使质量指标数据具有不确定性。在实际问题中这两类随机性常常交织在一起，如例 1.1 中，假设产品是在同一生产条件下生产出来的，如果出现了废品，这表明了不确定性对质量数据的影响；而随机抽取 n 件产品后，查得的废品件数则与抽取的偶然性有关。

随机性的普遍存在，促进了数理统计的发展，也为数理统计的应用提供了一个广阔的用武之地。数理统计首先就是因为生物学、遗传学和农业科学研究的需要而兴起的，在近一个世纪的发展中，数理统计几乎不同程度地渗到所有人类活动的领域。在农业方面，方差分析已经是农业试验的常规手段；在工业生产中，正交试验设计方法在新产品、新工艺、新材料的开发研究过程中得到广泛运用；在医学中，显著性检验是说明一些药物和治疗手段疗效的典型方法；在国防尖端武器的研制中，精度分析主要也是用数理统计的方法。特别是随着电子计算机的发展，计算机大批量、高速度处理数据的能力给数理统计的应用提供了新的条件。至今诸如回归分析、多元统计分析等数理统计方法更是在测量、通信、质量控制、气象、水文、地震预报、地质探矿、市场预测乃至考古研究，刑事鉴别等各方面都得到愈来愈多的应用。因此可以说数理统计是一门应用性很强的数学。

数理统计的基本内容大致包括数据采集和统计推断两个方面，两者在应用中都很重要，而且关系密切。数据采集主要包含抽样理论和试验设计等内容，统计推断则包括估计和检验两类大问题，如前述例子中，例 1.1 中第 (ii) 项工作属于估计问题，例 1.2 中第 (ii) 项工作属于检验问题。本书将花主要篇幅介绍这两类统计推断问题的常用方法及它们的应用，另外也适当介绍一下试验设计的方法。

需要强调的是既然如前所说统计推断的结果往往有赖于方法，因此需要尽可能地采用“有效的”方法。什么是有效的方法？这就有一个衡量问题的基点和标准问题。一般来讲由于数理统计方法处理的是带随机性影响的数据，因此对统计方法进行评价的基点往往是从概率出发，即一个统计方法的效果好在实际中总是含有诸如“大量重复使用该方法总体效果好”这类意思。这里也可看出数理统计和概率论的关系，即概率论是数理统计的基础，这是毫无疑义的。但两者的研究对象是不同的，在概率论中，往往是已知随机变量 X 的概率分布，需要我们去研究 X 的种种数学性质，而数理统计则往往是对研究对象 X 的概率分布不全知道或全不知道，需要从有关数据中去推断 X 的分布。如在例 1.1 (ii) 中，我们知道如果把产品是否为废品定义为取值为 1，0 的随机变量 X，设出废品的概率为 p，则 X 的分布可写成

$$\mathrm{P}\{X=i\}=p^{i}(1-p)^{1-i},\ i=0,1$$

在概率论中，人们的兴趣往往是在 p 已知的前提下进一步研究诸如 X 的期望 $\mathrm{E}(X)$，方差 $\mathrm{D}(X)$，或 X 的已知函数 $f(X)$ 的分布……等性质。而数理统计则往往是考虑当 p 未知时，如何从抽检的产品来推断产品的废品率 p，即在此问题中，是用数据来推断 X 的概率分布，这实际上也是统计推断的基本任务，我们后面还要提到。

1.2 样本和总体

1.2.1 样本

数理统计的研究对象是受随机性影响的数据，这些**通过观察**或试验得到的数据称为**样本或子样**，这些观察或试验**过程称为抽样**。例如用同一架天平称某重物 n 次，得到一组 n 个**数据**

$$X_1, X_2, \cdots, X_n \tag{1.1}$$

就称它们是一个样本，其中 n 称为**样本容量**。每个容量为 n 的样本都可称为 n 维空间的一个点，样本所有可能的取值构成了 n 维空间的一个子集，称为**样本空间**，记作 $\mathscr{X}$. 注意“数据”一词在这里是广义的。它可以是实数值，例如 X_i 表示称得某**重物的重量**；也可以是事物的属性：例如 X_i =“正品”，(或“废品”) 等等，通常为了方便研究，也常将这些属性数量化，例如用“1”表示“废品”，“0”表示“正品”，当然这不是本质的问题。有时**数据也可**以是一组向量，例如武器试验中给出一组弹着点的坐标

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$$

即为二维向量的一组样本，在多元统计分析中，将专门研究这种情形。对于样本需要强调两点：

a) 样本并非一堆杂乱无章无规律可循的数据，它是**受随机性**影响的一组数据，因此用概率论的话说就是每个样本既可以视为一组数据又可视为一组随机变量，这就是所谓样本的**二重性**。当通过一次具体的试验，得到一组观测值，这时样本表现为一组数据；但这组数据的出现并非是必然的，它只能以一定的概率（或概率密度）出现，这就是说当考察一个统计方法是否具有某种普遍意义下的效果时，又需要将其样本视为随机变量，而一次具体试验得到的数据则可视为随机变量的一个实现值。今后为行文方便，我们常交替使用上述两种观点来看待样本，而不去每次声明此处样本是指随机变量还是其实现值，同时一律采用记号(1.1)来

表示它。

b）样本(1.1)也不是任意一组随机变量，我们要求它是一组独立同分布的随机变量，同分布就是要求样本具有代表性，独立是要求样本中各数据的出现互不影响，就是说抽取样本时应该是在相同条件下独立重复地进行。如

[**例 1.3**] 设一组抽奖券共 10000 张，其中有 5 张有奖。问连续抽取 3 张均有奖的概率为多少？

为了讨论这个问题，不妨设

$$X_i=\begin{cases}1,\text{第 } i \text{ 次抽到奖}\\0,\text{第 } i \text{ 次未抽到奖}\end{cases}$$

要求该事件的概率，实际上即是求联合概率分布

$$P\{X_1=x_1,X_2=x_2,X_3=x_3\}(x_i=0 \text{ 或 } 1)$$

在 $x_1=x_2=x_3=1$ 处的值。但题中没有说明“连续抽取”是“有放回的”还是“无放回的”，我们不妨都计算一下：

(i) 无放回时：

$$P\{X_1=1,X_2=1,X_3=1\}=\frac{5}{10000}\,\frac{4}{9999}\,\frac{3}{9998}$$

(ii) 有放回时：

$$P\{X_1=1,X_2=1,X_3=1\}=\frac{5}{10000}\,\frac{5}{10000}\,\frac{5}{10000}=\left(\frac{5}{10000}\right)^3$$

显然 (i) 中的抽样方式不是独立的，每次抽样的结果都将影响下一次取样的分布，这种抽样不是我们通常研究的抽样。而 (ii) 中的抽样，则是多次独立的抽样，它们是同分布的，即我们通常称为的**随机抽样**。这样得到的数据，即是我们常研究的**简单样本**，或就直接称为样本。由此可以看出，对于样本(1.1)，如果每个 X_i 的共同分布为 F，则样本(1.1)的分布为

$$F(X_1)F(X_2)\cdots F(X_n) \qquad (1.2)$$

相应地，若 X_i 有共同概率密度 f，则(1.1)的概率密度为

$$f(X_1)f(X_2)\cdots f(X_n) \tag{1.3}$$

1.2.2 总体

总体或**母体**在许多教科书上通常被定义为研究对象全体的集合。其含义是我们观察到的样本总是由某个具体事物产生并反映该事物的特征，这时可以把样本视为一些被抽取的该事物的个体，而将该事物本身视为所有个体的集合即总体。但这样说多少有点模糊。如在例 1.3 中，我们自然可以将 10000 张抽奖券视为总体，但如果是用一架天平去重复称同一重物得到重物的重量，在这种事中什么是研究对象的全体呢?因此我们宁愿采用另一种说法，即说总体是一个随机变量，它的分布即为(1.1)中每个 X_i 的共同分布，或者可以看作样本容量 $n=1$ 时的样本 X_1 的分布 F，用这个观点叙述一些问题就显得很方便，例如样本(1.1)就可视为，由总体 X 独立“拷贝”出来的同分布的 n 个随机变量。又如

[**例 1.4**] 用两台车床车同一批产品，分别车 m 及 n 件，尺寸为 $X_1, X_2, \cdots, X_m$ 及 $Y_1, Y_2, \cdots, Y_n$. 这时，我们得到的样本是

$$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n \tag{1.4}$$

它们显然通常不会是同分布的，但这种样本在我们的研究中经常出现。为此我们用总体的观点，可以很方便地视它为出自两个总体 X, Y 的样本。有了总体这个概念，我们就可以将统计推断的基本任务概括为由样本推断总体。例如在例 1.4 中，我们就可以从样本(1.4)中推断出总体 X 与 Y 是否有显著差别。关于这一基本任务，我们今后可以慢慢体会到。由于推断总体实质上是推断总体的分布，即解决一个实际统计问题往往归结为总体分布的确定，所以我们也常称总体的分布是该问题的**统计模型**。

1.2.3 参数与参数空间

如前所述，在数理统计问题的分布一般来说是未知的，需要通过样本来推断。但如果对总体绝对地一无所知，那么所能做出