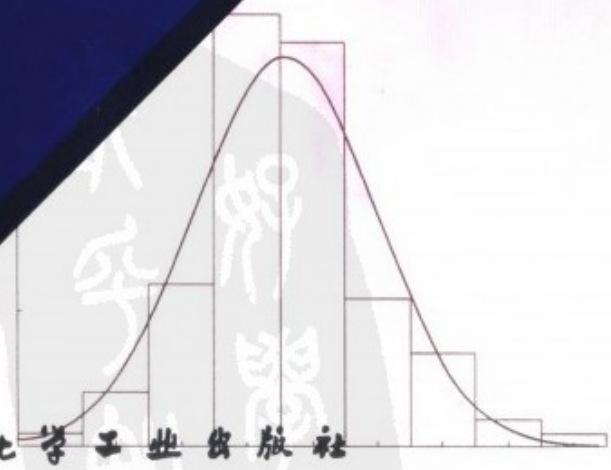
研究生应用数学丛书

应用统计

刘达民　程岩　编



化学工业出版社

研究生应用数学丛书

- 高等应用数学
  ——非线性分析
- 矩阵论及应用
- 应用统计
- 张量分析及应用

ISBN 7-5025-5696-6

9 787502 556969 >

ISBN 7-5025-5696-6/O·54 定价：30.00元

研究生应用数学丛书

应　用　统　计

刘达民　程岩　编

化学工业出版社
·北　京·

（京）新登字 039 号

图书在版编目（CIP）数据

应用统计/刘达民，程岩编. —北京：化学工业出版社，2004.6
（研究生应用数学丛书）
ISBN 7-5025-5696-6

Ⅰ.应… Ⅱ.①刘… ②程… Ⅲ.应用统计学-研究生-教材 Ⅳ.C8

中国版本图书馆 CIP 数据核字（2004）第 070747 号

研究生应用数学丛书
应 用 统 计
刘达民 程岩 编
责任编辑：任文斗
文字编辑：宋 薇
责任校对：李 林
封面设计：蒋艳君
*
化学工业出版社出版发行
（北京市朝阳区惠新里 3 号 邮政编码 100029）
发行电话：（010）64982530
http://www.cip.com.cn
*
新华书店北京发行所经销
北京市彩桥印刷厂印刷
三河市前程装订厂装订
开本 787mm×960mm 1/16 印张 17¾ 字数 315 千字
2004 年 8 月第 1 版 2004 年 8 月北京第 1 次印刷
ISBN 7-5025-5696-6/O·54
定 价：30.00 元

前　言

应用统计对于工科研究生是一门十分重要的课程，它既是许多专业的数学基础，又直接提供实用数学方法，特别是随着计算机的广泛应用，应用统计的理论和方法为处理科学研究和工程技术中众多受随机因素影响的实际问题，提供了有力的工具，日益显示其重要性．

本书编写过程中针对工科研究生的特点，在教材的材料安排、习题的选用上，突出实用性，体现了本课程应用性强的特色．为适应解决实际问题时，处理大量数据的需要，专辟一章介绍统计软件 SPSS 的使用，使读者能应用它去解决诸多的实际问题．全书行文力求通俗易懂，便于读者自学．

本书由刘达民作整体安排并编写了第 1 章至第 6 章，程岩编写了第 7 章．

本书得到北京化工大学“新世纪教学改革项目”及“研究生教育创新基金”的资助；在编写过程中还得到刘慧、王云的大力支持和帮助，在此表示感谢．

限于编者学识水平，书中难免存在不当之处，欢迎读者批评指正．

编者

2004 年 4 月于北京化工大学

内 容 提 要

本书是根据教委颁布的“工学硕士研究生应用统计课程基本要求”编写的，内容除了包括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、正交设计外，还补充了统计软件 SPSS 的应用简介。本书注重阐明统计思想和介绍各种统计方法，强调统计实际应用，并通过材料的选择安排、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。由于是以工学硕士研究生为主要对象，全书论述严谨，不仅追求数学上的严密性和完整性，而且对未给出推导的结论指明出处。本书行文深入浅出，注意启发性，书末附有概率论等基本知识，便于读者自学。

本书可作为工学硕士研究生应用统计课程的教材，也可作为高等院校高年级学生、教师和科研人员、工程技术人员的参考书。

目　　录

第 1 章

数理统计的基本概念

随机性在自然现象和社会现象的普遍存在，促进了数理统计的发展．近几十年来，特别是计算机技术的长足进步，为数理统计的广泛应用提供了广阔的前景．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，正交试验方法在新产品、新工艺、新材料的开发研究过程中得到广泛应用；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，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、保险等各方面得到越来越多的应用．因此，可以说数理统计是一门应用性很强的数学学科．

1.1 数理统计的基本问题

数理统计的研究对象是带随机性影响的数据，任务是如何有效地、合理地收集、整理、分析这些数据，并利用所得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为止．

数理统计不同于工农业生产中各项经济指标完成情况的统计，也不同于政府部门中的各项资料的汇总．它侧重于应用随机现象本身的统计规律性来研究问题，研究如何合理、有效地获得数据资料，并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，因而从理论上讲，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是有限的，甚至是少量的．如在质量管理中，若对产品的检验是破坏性的（如产品是弹药，必须通过击发来检验是否合格），这时要确定该批弹药是否合格，只能从中随机抽取若干件，通过击发

检验其是否是合格的，得出合格品数，并由此对该批弹药的合格率做出推测（即由局部来推断整体）；有时这种检验虽不是破坏性的，但是限于财力、人力和时间，也无法对该批产品中的所有产品逐个检验以确定该批产品的合格率，不得不采用随机抽取一些产品进行检验．

在产品检验中的上述两类情况，均由局部来推断整体，显而易见，这样做的代价是无法得到完全正确的结论，只能得到某种意义下正确的推断．

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：一方面是由“局部性”带来的，这类问题往往是研究的对象数量很大，不可能也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地反映对象的全部信息，但由于只是抽取其中的一部分，所以难免有偶然性；另一方面则是由“不确定性”带来的，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，这是因为实际上总有一些因素无法控制（如生产环境中温度、湿度的微小变化等）或不便控制，这就使质量指标数据具有不确定性．

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方法和试验设计等内容，统计推断则包括估计和检验两类大问题．本书主要介绍统计推断的基本原理和方法．

1.2 总体和样本

1.2.1 总体

通常把研究对象的全体称为**总体**（Population），把总体中的每个元素称为**个体**．如一批产品、一个学校中的全体学生都构成各自的总体，其中每一件产品、每一个学生则是该总体中的个体．

在实际中，人们关心的往往是总体的某个特征指标及其分布状况．如研究显像管的质量，关心的是其寿命（反映质量好坏的一个重要指标）而不是显像管本身．由于各显像管（即使属同一批、同一类型）的寿命不全相同，不可能也没有必要逐个地测出每个显像管的寿命，而只需了解全体显像管的寿命分布状况．由于任一个显像管的寿命测试前是不能确定的，但每一个显像管都确实对应着一个寿命值，所以可以认为显像管的寿命是个随机变量，而人们关心的正是这个随机变量的概率分布．一般说来，都可以认为所考察的总体是用一个随机变量来代表的．由这个观点总体可描述为：**总体就是一**

个具有确定概率分布的随机变量．以后，可以说总体 $F(x)$或总体 X，其含义是总体是一个以$F(x)$为分布函数的随机变量 X. 当然这个随机变量也可能是二维的，这时就可以说二维总体 $F(x,y)$，其含义就是总体是一个以$F(x,y)$为分布函数的二维随机变量（X，Y）．

1.2.2　样本

在数理统计中，总体的分布总是未知的．为了获得总体的分布，就必须对总体进行抽样观察．从总体中抽取一个个体，就是做一次试验，如果进行了 n 次抽样观察，并记录其试验结果，就得到总体的一组数值，记作（x_1，x_2，…，x_n），称之为样本容量为 n 的**样本观测值**，也简称为**样本值**．人们自然希望这一组样本值能很好地反映总体的情况，这就要求对抽取方法加上一定的限制．容易想到，如果总体中每个个体被抽到的机会均等，并且在抽取一个个体后总体的成分不改变，那么，抽得的个体就能得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简单样本观测值．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，便于理论分析处理．简单随机抽样相当于概率论中的有放回抽样，当从总体中抽取容量为 n 的样本（Sample）时，每抽取一个个体作观测后立即放回并搅匀，然后再抽取下一个个体；但实际情形中往往采取的是无放回抽样，特别是当总体所含的个体个数 N 很大，而样本容量 n 相对较小（一般是 n/N 不大于 0.1 时），可以近似地把无放回抽样看做是有放回抽样．

设（$x_1,x_2,\cdots,x_n$）为总体 X 的一组样本值，由于抽样的随机性与独立性，每个 x_i 都可以看做某一个随机变量 $X_i(i=1,2,\cdots,n)$ 所取的观测值，这里 $X_1,X_2,\cdots,X_n$ 相互独立，且皆与总体 X 具有相同分布．因此，（x_1，$x_2,\cdots,x_n$）又可以看做 n 维随机变量（$X_1,X_2,\cdots,X_n$）的观测值，就一次观察结果而言，（$x_1,x_2,\cdots,x_n$）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，这就是说抽样观察结果具有两重性，在不会混淆的情况下，今后，总是将来自总体 X 的容量为 n 的样本记为（$X_1,X_2,\cdots,X_n$）．

以上讨论的概念以定义和定理的形式表述如下．

定义 1.2.1　设（$X_1,X_2,\cdots,X_n$）为来自总体 X 的容量为 n 的样本，如果 $X_1,X_2,\cdots,X_n$ 相互独立且每个都是与总体 X 具有相同分布的随机变量，则称（$X_1,X_2,\cdots,X_n$）为总体 X 的容量为 n 的简单随机样本，简称为样本．它们的观测值（$x_1,x_2,\cdots,x_n$）称为总体 X 的 n 个独立的观测值．

随机向量（$X_1,X_2,\cdots,X_n$）所有可能取值的全体称为**样本空间**，一个

样本值 $(x_1,x_2,\cdots,x_n)$ 就是样本空间中的一个点.

定理 1.2.1 若 $(X_1,X_2,\cdots,X_n)$ 是来自具有分布函数 $F(x)$ 的总体 X 的样本，则 $(X_1,X_2,\cdots,X_n)$ 的联合分布函数为 $F_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i)$.

又若总体 X 具有概率密度函数 $f(x)$，则 $(X_1,X_2,\cdots,X_n)$ 的联合概率密度函数为 $f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)$.

例 1.2.1 总体 $X\sim b(1,p)$，它的概率函数为 $f(x)=p^x(1-p)^{1-x}$，$x=0$，1.

因此，样本 $(X_1,X_2,\cdots,X_n)$ 的联合概率函数为

$$f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}[p^{x_i}(1-p)^{1-x_i}]$$

$$=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i},\ x_i=0,1,i=1,2,\cdots,n$$

例 1.2.2 总体 $X\sim N(\mu,\sigma^2)$，它的概率密度函数为 $f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，$-\infty<x<+\infty$.

因此，样本 $(X_1,X_2,\cdots,X_n)$ 的联合概率密度函数为

$$f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}},\ -\infty<x_i<+\infty,\ i=1,2,\cdots,n$$

1.3 经验分布函数和直方图

总体分布未知时，要用样本对总体进行非参数推断，常用的方法是经验分布函数和直方图.

1.3.1 经验分布函数

总体 X 的分布函数，称为理论分布或总体分布，取自总体的一个容量为 n 的样本的分布函数称为经验分布函数.

定义 1.3.1 从总体 X 中抽取容量为 n 的样本 $(X_1,X_2,\cdots,X_n)$，将其样本值从小到大排列后为 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，对任何实数 x，定义函数

$$F_n^*(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}\leqslant x<x_{k+(1)},\ k=1,2,\cdots,n-1\\ 1, & x\geqslant x_{(n)}\end{cases}\qquad(1.3.1)$$

称 $F_n^*(x)$ 为总体 X 的经验分布函数.

$F_n^*(x)$ 的图形如图 1.1 所示，是呈跳跃上升的一条阶梯形折线，若样本值均不同，则在每一观测值 $x_{(k)}$ 处有一跳跃为 $\frac{1}{n}$，若有相同的观测值，则 $F_n^*(x)$ 在此处按 $\frac{1}{n}$ 的倍数（即相同观测值的个数）跳跃上升.

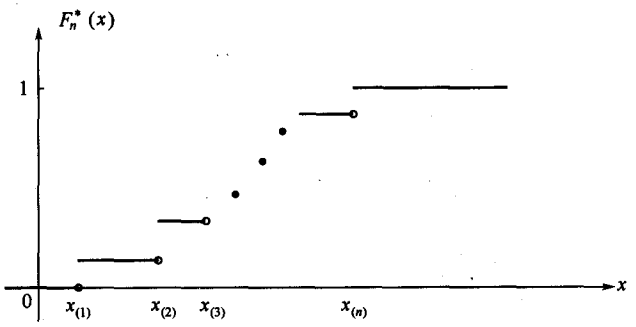


图 1.1　经验分布函数

显然，$F_n^*(x)$ 具有分布函数的一切性质，即有 $0\leqslant F_n^*(x)\leqslant 1$，$F_n^*(x)$ 非降且右连续.

例 1.3.1　对某厂生产的电子仪器作寿命测试，得样本值（单位：100h）为 5，3，7，5，4，试求经验分布函数.

解　样本值（按从小到大顺序排列）的频率分布为

样本值	3	4	5	7
频率	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

将上表视为某个离散型随机变量的概率分布列，与其相应的分布函数便是经验分布函数

$$F_5^*(x)=\begin{cases}0, & x<3\\ \dfrac{1}{5}, & 3\leqslant x<4\\ \dfrac{2}{5}, & 4\leqslant x<5\\ \dfrac{4}{5}, & 5\leqslant x<7\\ 1, & x\geqslant 7\end{cases}$$

自然会想到未知的总体分布函数是否可以用经验分布函数去近似？显然取决于$|F_n^*(x)-F_n(x)|$这个量的大小．下面的理论结果表明，这种做法是合理的．

定理 1.3.1 设 $(X_1,X_2,\cdots,X_n)$ 为来自总体 X 的样本，总体分布函数为 $F(x)$，对任意的 $x\in(-\infty,+\infty)$与任意给定的 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|F_n^*(x)-F(x)|\geqslant\varepsilon\}=0$$

证明 对 $\forall x\in(-\infty,+\infty)$，定义随机变量

$$Y_i=\begin{cases}1 & x_i\leqslant x\\ 0 & x_i>x\end{cases}\quad i=1,\ 2,\ \cdots,\ n$$

显然 $Y_1,Y_2,\cdots,Y_n$ 是独立同分布的随机变量，每个 $Y_i\sim b(1,p)$，其中 $p=P\{Y_i=1\}=P\{X_i\leqslant x\}=F(x)$.

另外由 $F_n^*(x)$的定义得到 $F_n^*(x)=\dfrac{1}{n}\sum_{i=1}^{n}Y_i$.

由于 $E(Y_i)=P\{X_i\leqslant x\}=F(x)$，易得 $E[F_n^*(x)]=F(x)$.

于是，由伯努利（Bernoulli）大数定律即得定理成立．

定理表明，当 n 足够大时，用经验分布函数 $F_n^*(x)$来近似未知的总体分布函数 $F(x)$的效果是比较理想的．

实际上，$F_n^*(x)$还一致地收敛于 $F(x)$．格列汶科在 1933 证明了这一更深刻的结论．

定理 1.3.2 （W. Glivenko 定理）

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

证明请参阅文献［1］．

由定理可知，当 n 足够大时，$F_n^*(x)$与 $F(x)$可以足够地接近，它们的差别的最大值（定理中的 sup 表示最大值），也会随 n 增大而趋于零，这样的结局是以概率 1 发生的事件．因而，当 n 足够大时，就可以用 $F_n^*(x)$来近似代替 $F(x)$，这就是以后可以用样本来推断总体的最基本的理论依据．

1.3.2 直方图

当总体 X 为连续型随机变量时，总体分布可以用总体密度函数 $f(x)$来刻画，当然 $f(x)$是未知的，需要用样本来对它进行推断，直方图是一种简便易行的方法，而且比较直观，特别适合于统计现场使用．下面给出绘制直方图的一般步骤．设 $(x_1,x_2,\cdots,x_n)$ 为取自总体 X 的样本值．

步骤 1 求出 $x_{(1)}=\min_{1\leqslant i\leqslant n}x_i$，$x_{(n)}=\max_{1\leqslant i\leqslant n}x_i$.

步骤 2 取 a 略小于 $x_{(1)}$，b 略大于 $x_{(n)}$，并把区间［a，b］分为 m 个