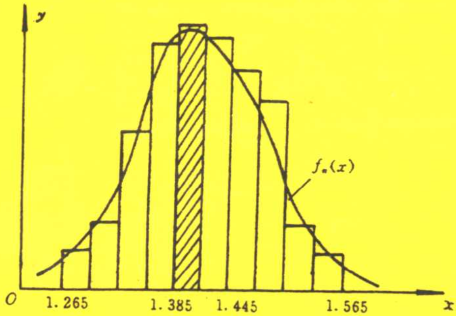
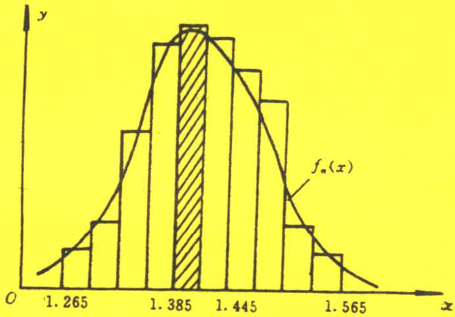
应用统计

何灿芝主编

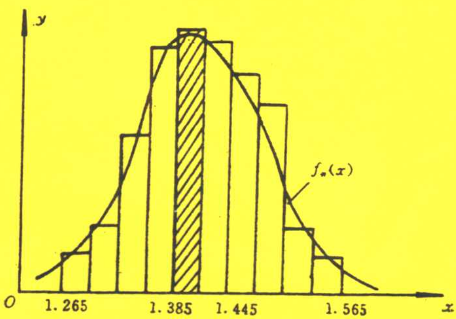
湖南科学技术出版社



YINGYONG TONGJI yingyong tongji



YINGYONG TONGJI yingyong tongji



YINGYONG TONGJI yingyong tongji

应用统计

主编　何灿芝
参编　罗　汉　喻胜华
　　　许和连　封建强

湖南科学技术出版社

应用统计

主　　编：何灿芝
责任编辑：胡海清
出版发行：湖南科学技术出版社
社　　址：长沙市展览馆路 11 号
印　　刷：湖南省新华印刷二厂
厂　　址：邵阳市双坡岭
邮　　编：422001
（印装质量问题请直接与本厂联系）
经　　销：湖南省新华书店
出版日期：1997 年 1 月第 1 版第 1 次
开　　本：787mm×1092mm　1/32
印　　张：10.75
字　　数：243,000
印　　数：1—2,310
征订期号：地科 215—02
书　　号：ISBN 7—5357—2135—4/O·153
定　　价：14.00 元

内容简介

本书系统地阐述了应用统计的基本概念，基本理论和基本方法。内容包括抽样分布、参数估计、假设检验、回归分析、方差分析、正交试验和多元统计分析等。各章节配有典型例题，各章之后配有适量习题，且在书后附有答案。

读者只要具备大学微积分、初等概率论和线性代数知识就可由浅入深地学习此书。

本书可作为工科硕士生的教材，也可作为工科高等院校高年级学生，特别是考研学生的参考书，还可以作为广大科学工作者和工程技术人员的参考书。

目录
CONTENTS

前言

PREFACE

概率论和应用统计（数理统计）是数学的两个密切联系的分支学科。概率论是研究各种描述随机现象的数学模型的概率特征，着重于理论上的探讨；应用统计是研究如何有效地收集、整理和分析受到随机影响的数据，从而对所考察的实际应用问题作出统计推断，提供解决问题的各种方法。它的理论基础是概率论。

应用统计的内容十分广泛，大致分为抽样理论和统计推断两大类。属于前者的内容有抽样技术，试验设计等；属于后者的内容有参数估计和假设检验，还有非参数估计，回归分析，方差分析，多元分析等。

应用统计已广泛地应用在工业、农业、军事、医学、公共事业、尖端科学等各个领域，并发挥着越来越重要的作用。因此，应用统计是每一个科学工作者和工程技术人员的必修课。

考虑到目前工科硕士生的数学基础，本书着重介绍基本概念，基本理论和基本应用方法，对于繁琐或涉及过深数学基础知识的证明从略。内容叙述力求做到简明易懂，例题和习题的选择做到具有典型性，便于教学与自学。

本书是在上几届工科研究生用的《数理统计》讲义的基础上修改、补充而形成的。该书的问世得到了湖南大学应用数学系和研究生部的大力支持和热情帮助，还得到了经济数学教授周怀生、蔡海涛和朱秀娟等各位的关心指导，在此一并表示衷心的感谢。

由于编者水平有限，不妥之处恳请读者批评指正。

编　者

1996 年 12 月于岳麓山

第一章　样本及其分布

§1.1　总体和样本

一　什么是总体和个体

数理统计中把“研究对象的全体”称为总体，又叫母体，将“组成总体的每一个元素”称为个体。

例如要考察一批灯泡，若这批灯泡有 1 万个，这就是总体，而每个灯泡都是个体。不过我们现在关心的不是灯泡本身，而是其“寿命”这个量的大小的分布情况。因此，在应用上总体是指研究对象的某个数量指标 X（如灯泡的寿命）的取值的全体。显然，这里的 X 是个随机变量。

定义 1.1　一个随机变量 X 或其相应的分布叫做一个总体，X 的每一个可能取值叫做一个个体。

在实际问题中，X 的分布未知或部分未知，故它正是统计推断的对象。

二　抽样和样本

1. 简单抽样

当我们研究某种总体的某个特性（如灯泡的寿命）时，若将总体中每一个个体都进行试验，这在实际中一般是不可能的，这是因为：第一，试验往往带有破坏性，如寿命测试就是破坏性试验；第二，即使试验无破坏性，但由于总体数量大，限于人力和物力也不可能进行全部试验。因此，需要采用由局部推

断总体的方法，即从总体 X 中抽取部分个体，如 n 个：X_1，X_2，…，X_n，由这 n 个个体所获得的信息来推断总体的情况。

由于 X_1，X_2，…，X_n 是从 X 中抽取的，故每个 X_i 都是一个随机变量，我们称 X_1，X_2，…，X_n 为来自总体的样本。从总体中抽取样本的过程称为抽样，最实用的抽样是简单抽样，它必须满足以下两个要求：

(1) 样本具有代表性：要求 X_1，X_2，…，X_n 的分布与总体 X 的分布相同。这就要求总体中每个个体被抽到的机会是均等的。

(2) 样本具有独立性：要求 X_1，X_2，…，X_n 相互独立。这就要求每抽出一个个体之后，总体的元素不变。（或近似不变）

实际上，简单随机抽样就是独立地，重复地从总体中抽取个体进行随机试验。

由简单抽样得到的样本，称为简单样本。今后，如无特别说明，样本都是指简单样本，于是我们可以给样本一个严格的数学定义。

2. 样本

定义 1.2 若随机变量 X_1，X_2，…，X_n 相互独立，且每个 X_i（$i=1$，2，…，n）与总体 X 有相同的分布，则称随机向量（X_1，X_2，…，X_n）为总体 X 的一个样本，n 称为样本容量，（X_1，X_2，…，X_n）的所有可能取值的集合称为样本空间。在一次具体的抽样中所得到的数值（x_1，x_2，…，x_n）称为样本（X_1，X_2，…，X_n）的一个观察值，简称样本观察值，它是样本空间中的一个点，故又称样本点。一般说来，不同次数的抽样，所得到的样本点是不同的。

3. 样本的联合分布

设（X_1，X_2，…，X_n）是总体 X 的一个样本，若 X 的分布

函数为 $F(x)$，概率密度为 $f(x)$，则根据样本的定义，因个体 X_i $(i=1, 2, \cdots, n)$ 与总体同分布，即 X_i 具有分布函数 $F(x_i)$ 和概率密度 $f(x_i)$，又 X_1，X_2，$\cdots$，X_n 是相互独立的，故它们的联合分布函数和联合概率密度分别为：

$$F^*(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i); \tag{1.1}$$

$$f^*(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i). \tag{1.2}$$

〔注意〕(i) 样本具有两重性：抽样试验之前，它是随机变量，抽样试验之后，得到的是确定的观察值。

(ii) 样本是相互独立的，与总体同分布的一组随机变量或一个随机向量。

§1.2 统计量

一 统计量

在应用统计中，样本是对总体进行估计和推断的依据，而经常用到的又是样本所构成的函数，假若这些函数中不含未知参数，则称为统计量。

定义 1.3 设 X_1，X_2，$\cdots$，X_n 是总体 X 的样本，$\hat{g}(X_1, X_2, \cdots, X_n)$ 是定义在样本空间上的不含未知参数的连续函数，则称 $\hat{g}(X_1, X_2, \cdots, X_n)$ 是一个统计量。若 x_1，x_2，$\cdots$，x_n 是 X_1，X_2，$\cdots$，X_n 的观察值，则 $\hat{g}(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值。显然，统计量是随机变量。

例 1.1 判断下列随机变量哪些是统计量？

设总体 $X\sim N(\mu, \sigma^2)$，其中 μ 已知，σ^2 未知；又 X_1，X_2，$\cdots$，X_n 是 X 的样本，随机变量为：

(1) $Y_1=\sum_{i=1}^{n}(X_i-\mu)^2$;

(2) $Y_2=\left(\sum_{i=1}^{n}X_i\right)/\sigma^2$.

解 Y_1 是不含未知参数的样本的连续函数，故 Y_1 是一个统计量；Y_2 中含未知参数，故不是统计量。

二 样本矩——常用统计量

设 X_1，X_2，…，X_n 是总体 X 的样本，x_1，x_2，…，x_n 为样本观察值，则有下列定义。

定义 1.4

(1) 样本均值：统计量

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

称为样本均值，$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 为其观察值。

(2) 样本方差：统计量

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

称为样本方差，它的观察值是 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$；其中 S 称为样本均方差。

(3) 样本 k 阶原点矩，统计量

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \qquad (k=1,2,\cdots).$$

称为样本 k 阶原点矩，其观察值为 $a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k(k=1,2,\cdots,n)$

(4) 样本 k 阶中心矩：统计量

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k \qquad (k=1,2,\cdots,n)$$

称为样本 k 阶中心矩，它的观察值为 $b_k = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

〔注意〕(i) 样本均值是一阶原点矩。

(ii) 样本方差本应是二阶中心矩，但为了今后使用的方便，我们把二阶中心矩的修正值 $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 定义为样本方差。

例 1.2 设总体 X 有有限的期望与方差：

$$EX = \mu, DX = \sigma^2.$$

(1) 求 $E\overline{X}, D\overline{X}$.

(2) 求 ES^2，EB_2.

解 (1) $E\overline{X} = E\left(\dfrac{1}{n}\sum_{i=1}^{n}X_i\right) = \dfrac{1}{n}\sum_{i=1}^{n}EX_i = \dfrac{1}{n}\sum_{i=1}^{n}\mu = \mu$,

$$D\overline{X} = D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}DX_i = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \sigma^2/n.$$

(2)
$$\begin{aligned}
ES^2 &= E\left[\frac{1}{n-1}\Sigma(X_i - \overline{X})^2\right] \\
&= \frac{1}{n-1}E\{\Sigma[(X_i - \mu) - (\overline{X} - \mu)]^2\} \\
&= \frac{1}{n-1}E\{\Sigma[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) \\
&\quad + (\overline{X} - \mu)^2]\} \\
&= \frac{1}{n-1}E\{\Sigma(X_i - \mu)^2 - 2n(\overline{X} - \mu)^2 \\
&\quad + n(\overline{X} - \mu)^2\} \\
&= \frac{1}{n-1}E\{\Sigma(X_i - \mu)^2 - n(\overline{X} - \mu)^2\} \\
&= \frac{1}{n-1}\{\Sigma DX_i - nD\overline{X}\} \\
&= \frac{1}{n-1}\left\{\Sigma\sigma^2 - n\,\frac{\sigma^2}{n}\right\} = \sigma^2.
\end{aligned}$$

$\because \qquad B_2 = \frac{1}{n}\Sigma(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$,(其中$\Sigma = \sum_{i=1}^{n}$,以下同)

$\therefore \qquad EB_2 = \frac{n-1}{n}ES^2 = \frac{n-1}{n}\sigma^2.$

可见，σ^2 是 S^2 取值的集中位置，而不是 B_2 取值的集中位置。

§1.3 经验分布函数

一 经验分布函数

定义 1.5 设总体 X 的样本（X_1，X_2，…，X_n）的取值（x_1，…，x_n），将其按大小排列：

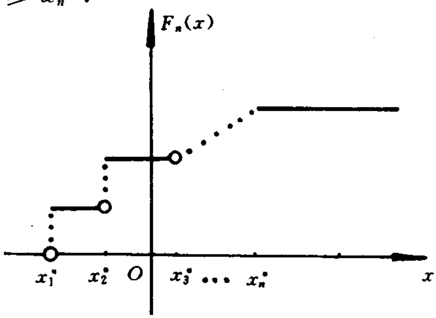
$$x_1{}^* \leqslant x_2{}^* \leqslant \cdots \leqslant x_n{}^*,$$

则称函数

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

为总体 X 的经验分布函数（或样本函数）。

它是一个阶梯函数，跃度为$\frac{1}{n}$（重合点跃度合并）。见图 1.1。



（图 1.1）

1. 经验分布函数的性质

1° $0 \leqslant F_n(x) \leqslant 1$；

2° $F_n(-\infty) = 0$，$F_n(+\infty) = 1$；

3° $F_n(x)$ 单调不减；

4° $F_n(x)$ 右连续。

2. $F_n(x)$ **的极限分布**

$F_n(x)$ 是样本观察值 $x_i(i=1,2,\cdots,n)$ 的函数，观察值不同，$F_n(x)$ 也不同，故 $F_n(x)$ 是一个随机变量(x 为参数)，当 $x_k^* \leqslant x < x_{k+1}^*$ 时，则不大于 x 的观察值出现的频率是 $\frac{k}{n}$，因为 $F_n(x)$ 等于 n 重贝努利试验中，事件 $A=\{X \leqslant x\}$ 所发生的频率，即

$$F_n(x)=\frac{k}{n}=\frac{(x_1\cdots x_n\text{ 中不大于 } x \text{ 的个数})}{n},$$

因而 $nF_n(x)$ 表示事件 A 在 n 重贝氏试验中出现的次数，所以 $nF_n(x)$ 服从二项分布，即

$$nF_n(x)\sim B(n,p),$$

其中 $p=P(A)=P(X\leqslant x)=F(x)$，$F(x)$ 是总体 X 的分布函数。

根据贝努利大数定律：对$\forall\ \varepsilon>0$，$|x|<+\infty$，有

$$\lim_{n\to\infty}P\{|F_n(x)-F(x)|<\varepsilon\}=1, \tag{1.3}$$

即 $F_n(x)$依概率 1 收敛于 $F(x)$，可见对每个固定的 x，当 n 充分大时，$\{|F_n(x)-F(x)|<\varepsilon\}$是大概率事件，按照实际推断原理，在一次抽样中，此事件几乎必然会发生，即可利用一次抽样所得到的经验分布函数 $F_n(x)$ 值来近似 $F(x)$，故 $F_n(x)$ 也是一个统计量，常用它来估计 X 的理论分布函数 $F(x)$。

由于 $F_n(x)$ 依赖于 x，故(1.3) 式有局限性，格利文科在 1953 年给出了一个更深入的结果。

定理 1.1（格利文科定理） 设 $F(x)$ 是总体 X 的理论分布函数，$F_n(x)$ 是经验分布函数，则

$$P\{\lim_{n\to\infty}\ \mathop{\mathrm{Sup}}_{|x|<+\infty}|F_n(x)-F(x)|=0\}=1. \tag{1.4}$$

该定理说明，当 n 很大时，$F_n(x)$ 非常接近 $F(x)$，这就说明

用样本可以推断总体。

二 频率直方图

前面讲了经验分布函数 $F_n(x)$ 可用来近似求理论分布函数 $F(x)$。下面介绍一种对连续型随机变量概率密度函数的近似求法——直方图法。

用频率直方图求概率密度函数的步骤：

1. 设 $f(x)$ 是总体 X 的概率密度函数，$X_1,\cdots,X_n$ 是样本，x_1，…，x_n 是样本值。

（1）将 x_1，…，x_n 按大小排列：

$x_1^* \leqslant x_2^* \leqslant \cdots \leqslant x_n^*$，找出 x_1^* 和 x_n^*，

$x_1^* = \min\{x_1,\cdots,x_n\}$；$x_n^* = \max\{x_1,\cdots,x_n\}$，

$R = x_n^* - x_1^*$.（称为样本极差）

选取 a（它略小于 x_1^*）和 b（它略大于 x_n^*），

使样本值全部落入 $[a, b]$ 区间内。

（2）将 $[a, b]$ m 等分，得分点

$$a = c_1 \leqslant c_2 \leqslant \cdots \leqslant c_{m+1} = b.$$

区间 $[c_i, c_{i+1}]$ 的长度 $h=(b-a)/m$ 称为组距。

（3）用唱票的办法，求出样本值落在每个小区间 $[c_i, c_{i+1}]$ 中的个数，称为频数，记作 n_i，再求出频率：$f_i=n_i/n$（$i=1, 2, \cdots, m$），并作频数分布表：

表 1.1 频数分布表

分组 $c_i \sim c_{i+1}$	频数 n_i	频率 $f_i=n_i/n$
$c_1 \sim c_2$	n_1	n_1/n
$c_2 \sim c_3$	n_2	n_2/n
…	…	…
$c_m \sim c_{m+1}$	n_m	n_m/n
Σ	n	1

（4）列出频率、组距比例表