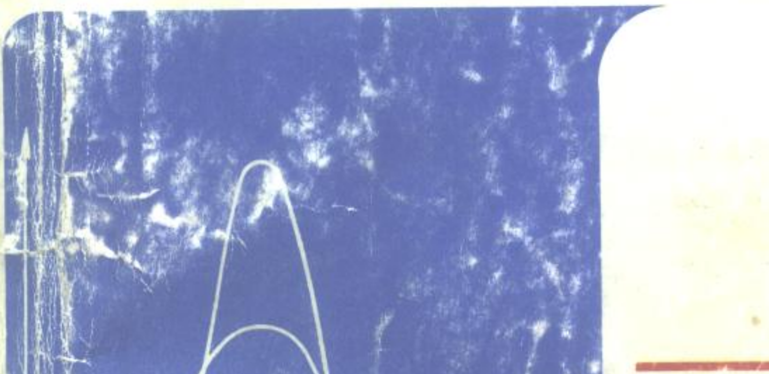
数理统计

SHU LI TONG JI



王式安 编

北京理工大学出版社

62

数 理 统 计

王式安　编

北京理工大学出版社

内 容 提 要

本书系统地阐述数理统计的思想、原理及方法。内容包括数理统计的基本概念、参数估计、假设检验、方差分析、回归分析、正交试验、多元分析、抽样调查和统计质量管理等。各章都有适量例题或习题。

本书可作为工科研究生的教材，或工科高等院校高年级学生和有关科技人员的参考书。

图书在版编目(CIP)数据

数理统计/王式安编．—北京：北京理工大学出版社，1995

ISBN 7-81045-029-8

Ⅰ．数… Ⅱ．王… Ⅲ．数理统计-高等学校-教材 Ⅳ．021

中国版本图书馆 CIP 数据核字(95)第 06673 号

北京理工大学出版社出版发行

(北京市海淀区白石桥路 7 号)

(邮政编码 100081)

各地新华书店经售

北京房山先锋印刷厂印刷

*

787×1092 毫米　32 开本　12.125 印张　282 千字

1995 年 7 月第一版　　1995 年 7 月第一次印刷

印数：1～2500 册　　定价：12.00 元

前　　言

数理统计是工科研究生的一门重要基础课。本书的前五章基本上是根据工科研究生数理统计课程基本要求编写的，这些内容可以在40学时内讲完。后三章是针对数理统计要求较多的专业，补充有关知识而编写的。各章具有相对独立性，大致可以用20学时选讲。

本书着重介绍数理统计中各种常见的方法及其应用，而不追求推导过程中数学上的严密性与完整性。叙述力求通俗易懂，便于读者自学。在阅读本书时，要求读者具备大学的微积分、初等概率论和线性代数知识。

本书编写过程中得到雷红旗、吴佃水同志的热情帮助，特别承秦明达教授审阅，在此表示感谢。在编写过程中，参阅了国内外有关书籍，引用其中某些例子，恕不一一指明出处，在此一并向有关作者致谢。

限于编者水平，书中不妥之处在所难免，真诚希望广大读者提出批评和指正。

王式安

1994年2月

目　　录

第一章　数理统计基本概念

本章主要介绍数理统计中的一些基本术语、基本概念、重要的统计量及抽样分布，它们是后面各章的基础。

§1-1　总体与子样

一、总体与个体

我们把所研究对象的全体所组成的集合称为**总体**，或**母体**。组成总体的每个元素称为**个体**。例如，在研究某批灯泡的质量时，该批灯泡全体就组成总体，而其中每个灯泡即为个体。

数理统计中所研究的常常是总体中每个个体的某一项（或某几项）数量指标和该数量指标在总体中分布的情况。用 X 表示数量指标，如果研究的是个体的 n 项数量指标，则 X 表示一个 n 维向量。今后在涉及“总体”这一概念时就用这些数量指标的全体组成的集合来代替原来的集合进行讨论。

例 1.1　研究某厂生产的一批灯泡的使用寿命，这时每个灯泡的使用寿命成为一个体，全部灯泡的寿命组成了总体。

例 1.2　某批产品 1000 个，每个产品等级有一等品、二等品和次品三类。我们将一等品用数字“1”表示，二等品用数字“2”表示，次品用数字“3”表示。如果只研究产品等级，这时每个产品的等级数量指标成为个体，1000 个产品的等级数量

指标的全体则构成总体。

例 1.3 研究一批炮弹的质量，每发炮弹有重量、穿透率、射程等方面的数量指标。如果只研究这批炮弹的重量指标时，各发炮弹的重量的全体就是一个总体，每发炮弹的重量是这个总体的个体。如果只考虑这批炮弹的穿透率指标时，各发炮弹的穿透率的全体就是现在的总体，每发炮弹的穿透率都是它的个体。如果要同时考虑炮弹的重量、穿透率和射程指标，则各发炮弹的重量、穿透率、射程这三项共同构成一个三维向量指标，仍可用 X 表示，这三维向量指标的全体组成一个总体。

在上述几个例子中，就数量指标 X 而言，每个个体所取的数值是不同的，如从总体中抽取一个个体就会观察到数量指标 X 的一次取值。由于在抽取时事先无法确定抽到的是总体中的哪一个个体，因而在还未具体抽定时可以认为 X 是一个随机变量。X 的分布就完全描写了总体中所研究的数量指标的分布状况。今后我们把总体和数量指标 X 可能取值全体所组成的集合等同起来。例如研究某厂生产的灯泡寿命 X 时，尽管工厂生产的灯泡数量不可能无限，因而所研究的灯炮寿命只可能取有限个值，但我们认为每次抽到的值 X 可能为任意一个非负的正数。这里总体被想象为所生产的灯泡的寿命的可能取值全体所组成的集合，是一个无限集。于是，X 可作为连续型随机变量处理。

总体与个体是相对的，要随研究目的而定。当所研究的问题变化时，总体和个体也要相应改变。如研究某地区各工厂生产灯泡的平均使用寿命时，则往往把该地区每个工厂生产灯泡的平均寿命看作个体，而将各工厂生产灯泡平均寿命的全体视为总体；但如果只研究某一工厂内生产的灯泡平均寿命，

此时，可将该厂生产的每个灯泡寿命作个体，而该厂生产的灯泡寿命的全体就形成总体。

二、子样

要了解总体的分布规律，就应对总体中的个体进行测定。但要把总体中所有个体一一加以测定，在实际中常常是不可能的或者不必要的。例如当测定本身具有破坏性或者总体中个体个数很多，或者测定很费时耗资等时，就不可能或者不必要把总体中所有个体一一加以测定。在统计分析工作中往往是从总体中进行抽样观察，而且还不止进行一次抽样观察。如果记第 i 次抽样观察的结果为 x_i，即第 i 次被观察到的个体的指标 X 的值为 x_i。观察了 n 次得到总体指标 X 的一组数值 $(x_1,x_2,\cdots,x_n)$，$(x_1,x_2,\cdots,x_n)$ 称为**容量为 n 的子样的观察值**。我们在抽样观察时从总体中进行抽取个体必须是随机的，只有这样才能经过多次观察比较全面地了解总体。对某一次具体抽样结果而言，x_i 是完全确定的一个值，但每次抽样观察它的值又是不一样，而我们就是根据这些观察结果来进行分析，因此在进行分析过程中不能把 x_i 看成确定的值，而是一个随机变量 X_i。n 次抽样观察得到的结果看作一个 n 维随机向量 $X=(X_1,X_2,\cdots,X_n)$，称 $(X_1,X_2,\cdots,X_n)$ 为**容量为 n 的子样**，简称子样或样本。取得子样的过程称为**抽样**。今后，用大写字母 $X_1,X_2,\cdots,X_n$ 表示随机变量，用小写字母 $x_1,x_2,\cdots,x_n$ 表示观察值。为方便起见，有时在书中某些地方并不严格区别这两者的记号。但可以从所涉及的内容判断是随机变量还是观察值。

由于抽取子样的目的是为了研究总体的性质，因而要求抽取的子样能很好地反映总体的特性，这就必须对抽样方法

提出一定的要求。该要求有两个方面:①代表性,即子样$(X_1,X_2,\cdots,X_n)$的每个分量X_i与总体X具有相同的概率分布。②独立性,即$X_1,X_2,\cdots,X_n$必须是相互独立的随机变量,也就是每次抽样的结果既不影响其余各次抽样的结果,也不受其它各次抽样结果的影响。满足上述两点要求的子样称为**简单随机子样**。获得简单随机子样的抽样方法叫简单随机抽样。今后,如果不作特殊声明,我们所说的子样都是指简单随机子样。

三、子样的分布

若总体X的分布函数已知是$F(x)$,根据抽样的代表性和独立性得出子样$(X_1,X_2,\cdots,X_n)$的联合分布函数是

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

若总体X具有分布密度$f(x)$,则子样$(X_1,X_2,\cdots,X_n)$的联合分布密度是$f_n(x_1,x_2,\cdots,x_n)=\prod\limits_{i=1}^{n}f(x_i)$。

若总体X是离散型,分布律为$p_i=P\{X=x_i\},i=1,2,\cdots$,则子样的分布律为

$$\begin{aligned}p_{i_1,i_2,\cdots,i_n}&=P\{X_1=x_{i_1},X_2=x_{i_2},\cdots,X_n=x_{i_n}\}\\&=\prod_{k=1}^{n}P\{X=x_{i_k}\}=\prod_{k=1}^{n}p_{i_k}\end{aligned}$$

其中$i_k=1,2,\cdots;\quad k=1,2,\cdots,n$

§1-2 统计量及其分布

一、统计量

1. 统计量定义

子样$(X_1,X_2,\cdots,X_n)$是从总体中随机抽取的,它是总体

的代表和反映，它应含有总体的各种信息。在实际工作中，为了要从子样中抽取所需信息，往往对子样进行必要的处理和计算，把子样中所包含的我们所关心的信息反映出来，针对不同的问题来构造子样的不同函数，而把这种函数称为统计量。

定义 设$(X_1,X_2,\cdots,X_n)$为总体 X 的一个子样，$\varphi(X_1,X_2,\cdots,X_n)$为子样的一个不含有任何未知参数的连续函数，则称 $\varphi(X_1,X_2,\cdots,X_n)$为一个**统计量**。

如果$(x_1,x_2,\cdots,x_n)$是子样$(X_1,X_2,\cdots,X_n)$的一个观察值，则称 $\varphi(x_1,x_2,\cdots,x_n)$是统计量 $\varphi(X_1,X_2,\cdots,X_n)$的一个观察值。

例1.4 设总体 X 服从正态分布 $N(\mu,\sigma^2)$，其中 μ 已知，但 σ^2 为未知，$(X_1,X_2,\cdots,X_n)$为 X 的一个子样，则 $\sum\limits_{i=1}^{n}\frac{1}{n}(X_i-\mu)^2$ 和 $\sum\limits_{i=1}^{n}2X_i^2$ 都是统计量。但是 $\sum\limits_{i=1}^{n}\frac{X_i^2}{\sigma}$和 $\sum\limits_{i=1}^{n}\frac{1}{\sigma^2}(X_i-\mu)^2$ 都不是统计量，因为它们包含了未知参数 σ。

2. 常用统计量

(1) 子样矩

设$(X_1,X_2,\cdots,X_n)$是总体 X 的子样，则其子样 k 阶原点矩为

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

特别地，对 $k=1$，称 A_1 为**子样均值**，记为 $\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$，它反映了总体的平均值。

子样 k 阶中心矩为

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

特别地，对 $k=1, B_1=0$，对 $k=2$，称 B_2 为子样二阶中心矩，记为

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

而我们称 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 为**子样方差**，它反映了子样中各分量对其平均值的离散程度，而且称 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 为**子样均方差**。

(2) 顺序统计量

设 $(x_1,x_2,\cdots,x_n)$ 是子样 $(X_1,X_2,\cdots,X_n)$ 的一个观察值，将它们按从小到大的次序排列起来，得到 $x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_n^*$。定义第 i 个顺序统计量 X_i^* 是 $X_1,X_2,\cdots,X_n$ 的这样一个函数，不论 $(X_1,X_2,\cdots,X_n)$ 取怎样的一组观察值 $(x_1,x_2,\cdots,x_n)$，它总是取其中的 x_i^* 为其观察值，由此得到的 $X_1^*,X_2^*,\cdots,X_n^*$ 为 $(X_1,X_2,\cdots,X_n)$ 的一组顺序统计量。显然 $X_1^*\leqslant X_2^*\leqslant\cdots\leqslant X_n^*$。

例 1.5 设 (X_1,X_2,X_3) 是容量为 3 的一个子样，将它们的三次观察值及 X_1^*,X_2^*,X_3^* 相应观察值列表如下：

表 1-1

观察值 \ 子样 / 观察序	X_1	X_2	X_3	X_1^*	X_2^*	X_3^*
1	1	0	1	0	1	1
2	3	1	0	0	1	3
3	1	2	0	0	1	2

从而可以看出，随机变量 X_1^* 一般不恒等于某个 X_i。显

然，$X_1^*=\min\{X_1,X_2,\cdots,X_n\}$，称为最小顺序统计量；$X_n^*=\max\{X_1,X_2,\cdots,X_n\}$，称为最大顺序统计量。

子样中位数 $\tilde{X}$ 定义为

$$\tilde{X}=\begin{cases}X_{m+1}^*, n=2m+1\\ \frac{1}{2}(X_m^*+X_{m+1}^*)\ , n=2m\end{cases}$$

称 $R=X_n^*-X_1^*$ 为**子样极差**，它在质量控制中起重要作用。在例 1.5 中，R 的三次观测值分别为 $R_1=1-0=1, R_2=3-0=3, R_3=2-0=2$。

二、抽样分布

由于统计量是子样$(X_1,X_2,\cdots,X_n)$的函数，而子样又是随机变量，因此，统计量也是随机变量，称它的分布为“抽样分布”。

下面介绍几种重要的抽样分布。由于这些抽样分布的论证过程需要用到较多的数学知识。这里，我们主要给出有关结论，以供应用。

(1) 设总体 X 服从正态分布 $N(\mu,\sigma^2)$，记为 $X\sim N(\mu,\sigma^2)$，$(X_1,X_2,\cdots,X_n)$是它的一个子样。由概率论知识知道，$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 必服从正态分布，其数学期望

$$E\overline{X}=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

方差

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

可以看出，n 越大，$\overline{X}$ 越向总体均值 μ 集中。

(2) χ^2-分布

设 $X_1, X_2, \cdots, X_n$ 为互相独立且都服从分布 $N(0,1)$ 的随机变量，把它们的平方和记作 χ^2，即 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$，则称随机变量 χ^2 所服从的分布为自由度 n 的 χ^2-分布，记作 $\chi^2 \sim \chi^2(n)$。

$\chi^2(n)$ 的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{2^{2/n}\Gamma(n/2)} \chi^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{1-1}$$

其中 $\Gamma(\frac{n}{2})$ 是 Γ 函数 $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ 在 $x = \frac{n}{2}$ 处的值。

$f(x)$ 的图形如图 1-1 所示。

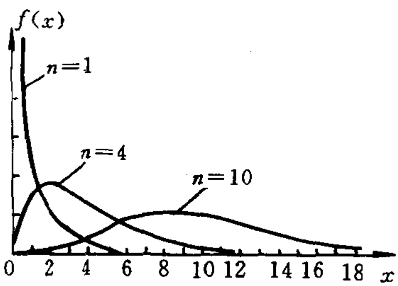


图 1-1　$\chi^2(n)$ 的概率密度

在图 1-1 中画出 $n=1, n=4, n=10$ 时 $\chi^2(n)$ 概率密度函数，其证明可参见其他有关参考书。

χ^2-分布具有下列性质：

① $E\chi^2 = n, D\chi^2 = 2n$，因为

$$E\chi^2 = E[\sum_{i=1}^{n} X_i^2] = \sum_{i=1}^{n} EX_i^2 = \sum_{i=1}^{n} DX_i = n$$

$$D\chi^2 = D[\sum_{i=1}^{n} X_i^2] = \sum_{i=1}^{n} DX_i^2 = \sum_{i=1}^{n} 2 = 2n$$

其中 $DX_i^2=EX_i^4-(EX_i^2)^2$

$$=\frac{1}{2\pi}\int_{-\infty}^{+\infty}x^4e^{-x^2/2}dx-1=3-1=2$$

② χ^2-分布对参数 n 具有可加性。

设 $\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，且它们相互独立，则 $\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$。

该性质可以推广成：k 个相互独立的 χ^2-分布的随机变量之和仍然是服从 χ^2-分布，它的自由度等于各个 χ^2-分布相应的自由度之和。

自由度可理解为表示平方和中独立变量的个数。由于随机变量 χ^2 是 n 个独立的随机变量 X_i 的平方和，因此 χ^2 的自由度为 n。

由于 n 个变量 $X_1-\overline{X},X_2-\overline{X},\cdots,X_n-\overline{X}$ 之间存在唯一的线性约束条件：

$(X_1-\overline{X})+(X_2-\overline{X})+\cdots+(X_n-\overline{X})=X_1+X_2+\cdots+X_n-n\overline{X}=0$

因此，平方和 $(n-1)S^2=\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$（或 $S^2=\frac{1}{n-1}\cdot\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$）中独立变量个数只有 $n-1$ 个。

对于给定的 $\alpha(0<\alpha<1)$，称满足条件：

$$P\{\chi^2>\chi_\alpha^2(n)\}=\alpha$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$-分布的上 100α 百分位点。对于不同的 α 及 n，上 100α 百分位点 $\chi_\alpha^2(n)$ 的值已制成表格，见附表Ⅳ。

当 $n\leqslant45$ 时，$\chi_\alpha^2(n)$ 的值可以直接查表得到。例如，对于 $\alpha=0.25$，$n=20$，查得 $\chi_{0.25}^2(20)=23.828$，即有 $P\{\chi^2(20)>23.828\}=0.25$。