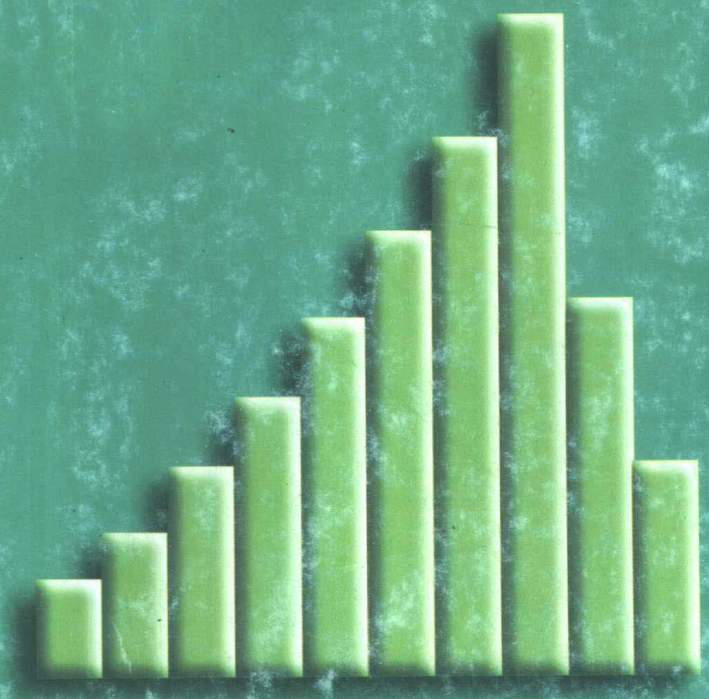
高等学校教材

数理统计

数理统计编写组　编



西北工业大学出版社

高等学校教材

数理统计

数理统计编写组　编

西北工业大学出版社

1999年7月　西安

(陕)新登字 009 号

【内容简介】 本书共分九章，主要内容包括：数理统计的基本概念、抽样分布、参数估计、假设检验、方差分析、回归分析、试验设计、多元分析初步、随机模拟与统计软件介绍等。

本书可作为高等学校工学、经济学硕士研究生数理统计课程的教材，也可作为理、工、农、医、师范、财经、统计、管理等专业本科生的教材或教学参考书，亦可供工程技术人员参考。

图书在版编目(CIP)数据

数理统计/赵选民等编．—西安：西北工业大学出版社，1999.7

ISBN 7-5612-1145-7

Ⅰ．数… Ⅱ．赵… Ⅲ．数理统计-研究生-教材 Ⅳ．O212

中国版本图书馆 CIP 数据核字(1999)第 27975 号

1999 西北工业大学出版社出版发行

(邮编：710072 西安市友谊西路 127 号 电话：8491147)

全国各地新华书店经销

陕西富平县印刷厂印装

ISBN 7-5612-1145-7/O·152(课)

*

开本：787 毫米×1 092 毫米 1/16 印张：17.125 字数：418 千字

1999 年 7 月第 1 版 1999 年 7 月第 1 次印刷

印数：1—3 000 册 定价：18.50 元

前　言

本书是根据全国工科院校硕士研究生“数理统计”课程的教学基本要求而编写的。全书共分九章，前五章介绍数理统计的基本理论与基本方法，内容包括：数理统计的基本概念，抽样分布、参数估计、假设检验、方差分析和回归分析。考虑到面向21世纪工科研究生数理统计课程教学改革和实际应用的需要，第六章介绍了正交试验设计、SN比及其试验设计和三次设计等方法，第七章、第八章、第九章分别介绍了多元分析初步、随机模拟方法和常用统计软件。这些方法在工、农业生产，社会、经济、工程技术和自然科学等领域都具有广泛的应用。本书各章均配有适量的习题，书末附有习题答案或提示。

在本书的编写过程中，考虑到工科硕士研究生的数学基础和教学特点，对数理统计学的基础与核心内容，尽量做到循序渐近，由浅入深，叙述严谨，分析透彻。而对应用方法部分，通过对典型实例的分析来介绍方法，培养学生应用所学知识解决工程实际问题的能力。本书的后四章内容基本独立，教师可根据学时和不同的教学要求选讲有关内容，或留给学生自学。

本书可作为工学、经济学硕士研究生“数理统计”课程48～70学时的教材，也可作为理学、农学、医学、师范、财经、统计、管理等专业本科生、研究生的教材或教学参考书，亦可供工程技术人员参考。

本书的第一章和第二章由秦超英编写，第三章和第六章由赵选民编写，第四章和第七章由师义民编写，第五章、第八章和第九章由徐伟编写，全书由赵选民统稿整理。本书的初稿曾作为讲义在西北工业大学95级、96级、97级和98级四届研究生教学中使用，并几经修改得以完善。西北工业大学应用数学系概率统计教研室、西北工业大学出版社和研究生院对本书的编写、出版给予了大力支持和帮助，西安交通大学范金城教授仔细地审阅了全稿，并提出了许多宝贵意见与建议，在此，一并致以衷心的谢忱。

由于编者水平有限，书中存在的不妥之处，敬请读者指正。

编　者

1998年12月于西北工业大学

目　录

第一章　数理统计的基本概念与抽样分布

数理统计学是研究随机现象规律性的一门学科，它以概率论为理论基础，研究如何以有效的方式收集、整理和分析受到随机因素影响的数据，并对所考察的问题作出推理和预测，直至为采取某种决策提供依据和建议．数理统计研究的内容非常广泛，概括起来可分为两大类：一是试验设计，即研究如何对随机现象进行观察和试验，以便更合理更有效地获得试验数据；二是统计推断，即研究如何对所获得的有限数据进行整理和加工，并对所考察的对象的某些性质作出尽可能精确可靠的判断．

数理统计是一门应用性很强的数学学科，已被广泛地应用到自然科学和工程技术的各个领域．数理统计方法已成为各学科从事科学研究以及在生产、管理、经济等部门进行有效工作的必不可少的数学工具．本章主要介绍数理统计中的一些基本概念，如总体、样本、统计量以及一些重要的统计量的分布等．

§1.1　基本概念

一、总体和样本

总体和样本是数理统计的两个重要的基本概念，读者须对此有较透彻的理解．

1. 总体

在数理统计学中，把所研究对象的全体元素组成的集合称为总体(或称母体)，而把组成总体的每个元素称为个体．例如，在考察某批灯炮的质量时，该批灯泡的全体就组成一个总体，而其中每个灯泡就是个体．又如，在考察某校学生的身体素质时，该校学生的全体组成一个总体，该校的每个学生是一个个体．

但是，在实际应用中，人们所关心的并不是总体中个体的一切方面，而所研究的往往是总体中个体的某一项或某几项数量指标．例如，考察灯泡质量时，我们并不关心灯泡的形状、式样等特征，而只研究灯泡的寿命、亮度等数量指标特征．如果只考察灯泡寿命这一项指标时，由于一批灯泡中每个灯泡都有一个确定的寿命值，因此，自然地把这批灯泡寿命值的全体视为总体，而其中每个灯泡的寿命值就是个体．又如，考察一批钢筋这一总体，当我们只关心它的强度这一指标(当然也可以关心其长度、重量、某化学成分等数量指标)时，那么，这批钢筋的强度值的全体就是总体，每个钢筋的强度值就是个体．由于具有各种不同强度值的钢筋比例数是按一定规律分布的，即任取一根钢筋其强度为某一可能值是有一定概率的，也就是说，这批钢筋强度是一个随机变量．同样就一批灯泡这个总体而言，如果关心的是这批灯泡的寿命这个数量指标 X，它也是一个随机变量．假定 X 的分布函数为 $F(x)$，并把表示这个数量指标

的随机变量 X 的可能取值的全体看作总体，且称这一总体 X 为具有分布函数 $F(x)$ 的总体，这样就把总体与随机变量联系起来了，因而，任何一个总体，都可用一个相应的随机变量来描述．总体便视为一个带有确定概率分布的随机变量．对总体的研究就归结为对表示总体某个数量指标的随机变量的研究，所谓总体的分布及数字特征，就是指表示总体某个数量指标的随机变量的分布及数字特征．例如，正态总体即指表示总体某个数量指标的随机变量服从正态分布．为了方便，今后常用大写字母 X、Y、Z 等来表示总体．

2. 样本

为了对总体 X 的分布规律或某些特征进行研究，就必须对总体进行抽样观察，根据抽样观察所得到的结果来推断总体的性质．这种从总体 X 中抽取若干个个体来观察某种数量指标 X 的取值过程，称为抽样（又称采样），这种做法称为抽样法．抽样法的基本思想是从所研究对象的全体中抽取一小部分进行观察和研究，从而对整体进行推断．

从一个总体 X 中，随机地抽取 n 个个体 $X_1, X_2, \cdots, X_n$（例如，在 1 万件产品中随机抽取 50 件），通常记为 $(X_1, X_2, \cdots, X_n)$．这样取得的 $X_1, X_2, \cdots, X_n$ 称为总体 X 的一个样本（又称子样），样本中个体的数目 n 称为样本容量．

由于每个 $X_i (i = 1, 2, \cdots, n)$ 都是从总体 X 中随机抽取的，它的取值就在总体可能取值范围内随机取得，当然每个 X_i 都是一个随机变量，而样本 $(X_1, X_2, \cdots, X_n)$ 自然就是一个 n 维随机变量．在一次抽取之后，它们是 n 个具体的数据 $(x_1, x_2, \cdots, x_n)$，称之为样本 $(X_1, X_2, \cdots, X_n)$ 的一个观测值，简称样本值．一般来说，两次不同的抽取（每次取 n 个）得到的样本值（两批 n 个数据）是不相同的．我们把样本本身是随机变量，而一经抽取又是一组确定的具体值的这种特性，称为样本的两重性．

样本 $(X_1, X_2, \cdots, X_n)$ 所可能取值的全体，称为样本空间，记为 Ω，一个样本值 $(x_1, x_2, \cdots, x_n)$ 就是样本空间 Ω 中的一个点．

我们的目的是要根据从总体 X 中抽取的一个样本值 $(x_1, x_2, \cdots, x_n)$，对总体 X 的分布或某些特征进行分析推断，因而要求抽取的样本能很好地反映总体的特性，这就需要对如何抽取样本提出一些要求，通常提出下面两条：

(1) 独立性 —— 因为独立观察是一种最简单而实用的观察方法，这就要求 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量．也就是说，每个观察结果既不影响其它观察结果，也不受到其它观察结果的影响．

(2) 代表性 —— 因抽取的样本要能尽可能地代表总体的特性，所以要求每个 $X_i (i = 1, 2, \cdots, n)$ 必须与总体 X 具有相同的分布．

满足上述两条性质的样本称为简单随机样本．获得简单随机样本的方法称为简单随机抽样．

在实际中抽取简单随机样本的方法很简单．例如当抽取的样本容量 n 相对总体中的个体来说是很小时（如总体为 10 000 件，抽取 $n = 50$ 件），则连续抽取的 n 个个体就可以近似认为是一个简单随机样本，因为此时抽取的个数相对总体而言是很小的，不放回抽取对总体的影响很小．如果是有放回抽取，则不必要求 n 相对很小，这样抽取的 n 个个体也是一个简单随机样本．又如，对一个物体重复测量其长度，测量值是一个随机变量，重复测量 n 次得到的 $X_1, X_2, \cdots, X_n$ 也是简单随机样本．

为使读者对以上讨论的总体和样本有一个明确的数学概念，给出如下定义．

定义 1.1 一个随机变量 X 或其相应的分布函数 $F(x)$ 称为一个总体.

定义 1.2 若随机变量 $X_1, X_2, \cdots, X_n$ 相互独立且每个 $X_i (i = 1, 2, \cdots, n)$ 与总体 X 具有相同的分布，则称 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的简单随机样本，简称为样本.

今后，如无特别说明，凡提到样本都指简单随机样本. 关于样本的分布有如下性质.

定理 1.1 设总体 X 的分布函数为 $F(x)$（或分布密度为 $\varphi(x)$ 或分布律为 $P\{X = x^{(i)}\} = P(x^{(i)}), i = 1, 2, \cdots$），则来自总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为 $\prod_{i=1}^{n} F(x_i)$（或联合分布密度为 $\prod_{i=1}^{n} \varphi(x_i)$ 或联合分布律为 $\prod_{i=1}^{n} P(x_i)$）.

证明 样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为

$$F(x_1, x_2, \cdots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n) =$$

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

式中第一个等号利用了样本的独立性，$F_{X_i}(x_i)$ 表示 X_i 的分布函数，而第二个等号利用了 $X_i (i = 1, 2, \cdots, n)$ 与总体 X 同分布.

例 1.1 设总体 X 服从参数为 p 的两点分布，即

$$P\{X = 1\} = p, \quad P\{X = 0\} = 1 - p, \quad 0 < p < 1$$

试求样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布律.

解 由于总体 X 的分布律可以写成

$$P(x) = P\{X = x\} = p^x (1 - p)^{1 - x}, \quad x = 0, 1.$$

故由定理 1.1，样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布律为

$$\prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1 - x_i} = p^{\sum_{i=1}^{n} x_i} (1 - p)^{n - \sum_{i=1}^{n} x_i}$$

例 1.2 设总体 X 服从正态 $N(\mu, \sigma^2)$ 分布，试求样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布密度.

解 总体 X 的分布密度为

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

故样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布密度为

$$\prod_{i=1}^{n} \varphi(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

二、统计量和样本矩

1. 统计量

样本是总体的代表和反映，但在抽取样本之后，并不能直接利用样本进行推断，而需要对样本进行“加工”和“提炼”，把样本中关于总体的信息集中起来，这便是针对不同的问题构造出样本的某种函数. 为此，引进统计量的概念.

定义 1.3 设 $(X_1, X_2, \cdots, X_n)$ 为总体 X 的一个样本，若 $f(X_1, X_2, \cdots, X_n)$ 为一个函数，且 f 中不含任何有关总体分布的未知参数，则称 $f(X_1, X_2, \cdots, X_n)$ 为一个统计量.

例如，设总体 $X \sim N(\mu, \sigma^2)$，μ 已知，σ^2 未知，$(X_1, X_2, \cdots, X_n)$ 是总体 X 有一个样本，则

$\frac{1}{n}\sum_{i=1}^{n}X_i$，$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ 都是统计量，而 $\frac{1}{\sigma}\sum_{i=1}^{n}X_i$，$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 都不是统计量，因为它们含有未知参数 σ.

由于样本 $X_1,X_2,\cdots,X_n$ 是随机变量，统计量 $f(X_1,X_2,\cdots,X_n)$ 也是随机变量，它们应有确定的概率分布，又因样本具有两重性，故统计量也具有两重性.

2. 常用统计量 —— 样本矩

定义 1.4 设 $(X_1,X_2,\cdots,X_n)$ 是从总体 X 中抽取的样本，称统计量：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad \text{为样本均值}$$

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2 \qquad \text{为样本方差}$$

$$S_n^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \qquad \text{为修正样本方差(简称样本方差)}$$

$$S_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2} \qquad \text{为样本标准差}$$

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \quad (k=1,2,\cdots) \qquad \text{为样本 } k \text{ 阶原点矩}$$

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k \quad (k=1,2,\cdots) \qquad \text{为样本 } k \text{ 阶中心矩}$$

由定义 1.4 可见，$A_1=\overline{X}$，$B_2=S_n^2$，$S_n^{*2}=\frac{n}{n-1}S_n^2$.

用 $\bar{x},s_n^2,a_k,b_k$ 分别表示 $\overline{X},S_n^2,A_k,B_k$ 的值，此时只要把定义 1.4 中 X_i 改为 x_i 即可.

由大数定律可以证明，只要总体 X 的 k 阶矩存在，则样本的 k 阶矩依概率收敛于总体 X 的 k 阶矩. 即对任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|\overline{X}-\mu|<\varepsilon\}=1$$

$$\lim_{n\to\infty}P\{|S_n^2-\sigma^2|<\varepsilon\}=1$$

式中 $\mu=EX,\sigma^2=DX$. 此结论表明，n 很大时可用一次抽样后所得的样本均值 $\overline{X}$ 和样本方差 S_n^2 分别作为总体 X 的均值 μ 和方差 σ^2 的近似值.

定理 1.2 设总体 X 具有 $2k$ 阶矩，则来自总体 X 的样本 k 阶原点矩 A_k 的数学期望和方差分别为

$$EA_k=\alpha_k$$

$$DA_k=\frac{\alpha_{2k}-\alpha_k^2}{n}$$

其中 $\alpha_k=EX^k(k=1,2,\cdots)$ 表示总体 X 的 k 阶原点矩.

证明

$$EA_k=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^k\right)=\frac{1}{n}\sum_{i=1}^{n}EX_i^k=\frac{1}{n}\sum_{i=1}^{n}EX^k=\alpha_k$$

$$DA_k=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i^k\right)=\frac{1}{n^2}\sum_{i=1}^{n}DX_i^k=\frac{1}{n^2}\sum_{i=1}^{n}DX^k=$$

$$\frac{1}{n}(EX^{2k}-(EX^k)^2)=\frac{\alpha_{2k}-\alpha_k^2}{n}$$

设总体 X 的数学期望 $EX=\mu$，方差 $DX=\sigma^2$，则有下列推论．

推论 $E\overline{X}=\mu$，$D\overline{X}=\frac{1}{n}\sigma^2$，$ES_n^2=\frac{n-1}{n}\sigma^2$，$ES_n^{*2}=\sigma^2$．

证明 由于 $A_1=\overline{X}$，在定理 1.2 中令 $k=1$，则有

$$E\overline{X}=\alpha_1=\mu$$

$$D\overline{X}=\frac{\alpha_2-\alpha_1^2}{n}=\frac{EX^2-(EX)^2}{n}=\frac{DX}{n}=\frac{\sigma^2}{n}$$

令 $k=2$ 时，有

$$EA_2=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)=\alpha_2=\mu^2+\sigma^2$$

又

$$E\overline{X}^2=D\overline{X}+(E\overline{X})^2=\frac{\sigma^2}{n}+\mu^2$$

故

$$ES_n^2=E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2\right)=$$

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-E\overline{X}^2=\mu^2+\sigma^2-\left(\frac{1}{n}\sigma^2+\mu^2\right)=\frac{n-1}{n}\sigma^2$$

由推论可见，样本均值 $\overline{X}$ 具有与总体 X 相同的数学期望，但是 $\overline{X}$ 的方差却是总体 X 的方差的 n 分之一，因而它更向期望值 μ 集中，n 越大，$\overline{X}$ 越向 μ 集中．

三、次序统计量和经验分布函数

1. 次序统计量

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由小到大的递增序列重新排列：

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$$

当 $(X_1,X_2,\cdots,X_n)$ 取值为 $(x_1,x_2,\cdots,x_n)$ 时，定义 $X_{(k)}$ 取值为 $x_{(k)}(k=1,2,\cdots,n)$，由此得到的 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$，称为样本 $(X_1,X_2,\cdots,X_n)$ 的次序统计量．显然有

$$X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$$

其中 $X_{(1)}=\min\limits_{1\leqslant i\leqslant n}X_i$ 称为最小次序统计量，它的值 $x_{(1)}$ 是样本值中最小的一个；而 $X_{(n)}=\max\limits_{1\leqslant i\leqslant n}X_i$ 称为最大次序统计量；它的值 $x_{(n)}$ 是样本值中最大的一个．由于次序统计量的每个分量 $X_{(k)}$ 都是样本 $(X_1,X_2,\cdots,X_n)$ 的函数，所以 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 也都是随机变量．样本 $(X_1,X_2,\cdots,X_n)$ 是相互独立的，但其次序统计量 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 一般不是相互独立的，因为次序统计量的任一值均按由小到大次序排列．下面不加证明给出最小、最大次序统计量的分布以及次序统计量的联合分布．

定理 1.3 设总体 X 的分布密度为 $\varphi(x)$（或分布函数为 $F(x)$），$(X_1,X_2,\cdots,X_n)$ 是 X 的样本，$(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 为其次序统计量，那么

(1) 最小次序统计量 $X_{(1)}$ 的分布密度为

$$\varphi_{X_{(1)}}(x)=n[1-F(x)]^{n-1}\varphi(x)$$

(2) 最大次序统计量 $X_{(n)}$ 的分布密度为

$$\varphi_{X_{(n)}}(x)=n[F(x)]^{n-1}\varphi(x)$$

(3) $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 的联合分布密度为

$$\varphi_{(X_{(1)},X_{(2)},\cdots,X_{(n)})}(x_{(1)},x_{(2)},\cdots,x_{(n)})=\begin{cases}n!\prod_{k=1}^{n}\varphi(x_{(k)}), & x_{(1)}<x_{(2)}<\cdots<x_{(n)}\\ 0, & \text{其它}\end{cases}$$

例 1.3 设总体 X 服从区间(0,1)上的均匀分布，$(X_1,X_2,\cdots,X_n)$ 是来自总体 X 的一个样本．易知，X 的分布密度为

$$\varphi(x)=\begin{cases}1, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

X 的分布函数为

$$F(x)=\begin{cases}0, & x\leqslant 0\\ x, & 0<x\leqslant 1\\ 1, & x>1\end{cases}$$

由定理 1.3 得，最小次序统计量 $X_{(1)}$ 的分布密度为

$$\varphi_{X_{(1)}}(x)=\begin{cases}n(1-x)^{n-1}, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

最大次序统计量 $X_{(n)}$ 的分布密度为

$$\varphi_{X_{(n)}}(x)=\begin{cases}nx^{n-1}, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

次序统计量 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 的联合分布密度为

$$\varphi_{(X_{(1)},\cdots,X_{(n)})}(y_1,y_2,\cdots,y_n)=\begin{cases}n!, & 0<y_1<y_2<\cdots<y_n<1\\ 0, & \text{其它}\end{cases}$$

下面介绍样本中位数和样本极差，它们是由次序统计量的函数给出的统计量．

样本中位数定义为

$$\tilde{X}=\begin{cases}X_{(\frac{n+1}{2})}, & n\text{ 为奇数}\\ \frac{1}{2}\left[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}\right], & n\text{ 为偶数}\end{cases}$$

它的值为

$$\tilde{x}=\begin{cases}x_{(\frac{n+1}{2})}, & n\text{ 为奇数}\\ \frac{1}{2}\left[x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right], & n\text{ 为偶数}\end{cases}$$

由定义可知，当 n 为奇数时，样本中位数取 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 的正中间那个数；当 n 为偶数时，样本中位数取正中间两个数的算术平均值．样本中位数与样本均值一样是刻划样本的位置特征的量，而且它计算方便且不受样本中异常值的影响，有时比样本均值更具有代表性．

样本极差定义为

$$R=X_{(n)}-X_{(1)}=\max_{1\leqslant i\leqslant n}X_i-\min_{1\leqslant i\leqslant n}X_i$$

它的值为

$$r=x_{(n)}-x_{(1)}=\max_{1\leqslant i\leqslant n}x_i-\min_{1\leqslant i\leqslant n}x_i$$

即样本极差是样本中最大值与最小值之差，它与样本方差一样是反映样本值的变化幅度或离

散程度的数字特征，而且计算方便，所以在实际中有广泛的应用．

例 1.4 从总体中抽取容量为 6 的样本，测得样本值为

32，65，28，35，30，29

试求样本中位数，样本均值，样本极差，样本方差，样本标准差．

解 将样本值按由小到大次序排列如下：

$x_{(i)}$：28，29，30，32，35，65

样本中位数 $\tilde{x}=\frac{1}{2}[x_{(3)}+x_{(4)}]=31$

样本均值 $\bar{x}=\frac{1}{6}\sum_{i=1}^{6}x_i=36.5$

样本极差 $r=\max\limits_{1\leqslant i\leqslant 6}x_i-\min\limits_{1\leqslant i\leqslant 6}x_i=37$

样本方差 $s_n^2=\frac{1}{6}\sum_{i=1}^{6}x_i^2-\bar{x}^2=167.583$

样本标准差 $s_n=\sqrt{\frac{1}{6}\sum_{i=1}^{6}x_i^2-\bar{x}^2}=12.945$

由上例可见，样本均值 $\bar{x}$ 大于样本值 6 个数中的 5 个数，这是因为有一个特别大的数 65 的缘故．样本均值对异常值或极端值较为敏感，而样本中位数 $\tilde{x}$ 则不受异常值的影响，因此，有时估计总体均值用样本中位数比用样本均值效果更好．

2. 经验分布函数

根据样本值求总体的分布函数，是数理统计要解决的一个重要问题．为此，引进经验分布函数的概念．

定义 1.5 设 $(X_1,X_2,\cdots,X_n)$ 是来自总体 X 的样本，样本的次序统计量为 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$，当给定一组次序统计量的值 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$ 时，对任意实数 x，称下列函数

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

为总体 X 的经验分布函数．换句话说，对任何实数 x，$F_n(x)$ 等于样本值中不超过 x 的个数再除以 n.

经验分布函数的性质：

(1) 当给定 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 时，$F_n(x)$ 是一个分布函数，因为它具有通常分布函数的特征，即具有以下性质：

(i) $0\leqslant F_n(x)\leqslant 1$；

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

(iii) 非减右连续．

$F_n(x)$ 是变量 x 的一个阶梯函数，其图形呈跳跃上升的一条阶梯形折线，只在 $x_{(k)}(k=1,2,\cdots,n)$ 处有间断点，跃度是 $\frac{1}{n}$ 的倍数，即样本值不重复，其跃度为 $\frac{1}{n}$，若有 l 次重复，其跃度为 $\frac{l}{n}$.

(2) $F_n(x)$ 是随机变量，且服从二项分布．事实上，对于不同的样本值，一般来说，将得到不同的 $F_n(x)$，由于 $(x_1,x_2,\cdots,x_n)$ 是样本 $(X_1,X_2,\cdots,X_n)$ 所取的值，因而 $F_n(x)$ 又是依赖于样

本的函数，所以经验分布函数 $F_n(x)$ 是随机变量，且该随机变量所有可能取值为 $0,\frac{1}{n},\frac{2}{n},\cdots,\frac{n-1}{n},1$，即 $F_n(x)$ 是离散型随机变量．由定义 1.5 可见

$$F_n(x)=\frac{(x_1,x_2,\cdots,x_n\text{ 中不超过 }x\text{ 的个数})}{n}\qquad(-\infty<x<+\infty)$$

即等于在 n 次独立重复试验中，事件 $\{X\leqslant x\}$ 发生的频率，因而 $nF_n(x)$ 表示事件 $\{X\leqslant x\}$ 在 n 次独立重复试验中出现的次数，所以 $nF_n(x)$ 服从二项分布 $B(n,p)$，即

$$P\left\{F_n(x)=\frac{k}{n}\right\}=C_n^k p^k(1-p)^{n-k}\qquad(k=0,1,\cdots,n)$$

其中 $p=P\{X\leqslant x\}=F(x)$．从而有

$$E[F_n(x)]=F(x),\quad D[F_n(x)]=\frac{1}{n}F(x)[1-F(x)]$$

(3) 经验分布函数 $F_n(x)$ 与总体分布函数 $F(x)$（又称理论分布函数）之间的关系．

由于对固定的 x，$F_n(x)$ 表示 n 次独立重复试验中事件 $\{X\leqslant x\}$ 发生的频率，而 $F(x)$ 表示事件 $\{X\leqslant x\}$ 发生的概率，则根据贝努里大数定律可知，对任意 $\varepsilon>0$ 及 $x\in(-\infty,+\infty)$，有

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

即 $F_n(x)$ 依概率收敛于 $F(x)$．但是，这里 n 的大小依赖于 x，因为 $F_n(x)$ 依概率收敛于 $F(x)$ 是对每一固定的 x 而言，带有很大的局限性．格列汶科(Glivenko) 于 1953 年给出了一个更为深刻的结果．

定理 1.4 （格列汶科定理）当 $n\to\infty$ 时，$F_n(x)$ 以概率 1 关于 x 均匀地收敛于 $F(x)$，即

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

此定理表明，当 n 很大时，对一切 x 而言，$\{|F_n(x)-F(x)|<\varepsilon\}$ 都是大概率事件，这里 ε 是任意给定的很小正数，因而可利用一次抽样后得到的 $F_{(n)}(x)$ 来一致逼近理论分布函数 $F(x)$．换句话说，当 n 很大时，由每一组样本值得到的经验分布函数 $F_n(x)$ 都是总体分布函数 $F(x)$ 的一个良好的近似．格列汶科定理不仅提供了总体分布函数 $F(x)$ 的一个良好的估计，而且也可以作为利用样本的性质来推断总体具有相应性质的理论依据．

§1.2 抽样分布

所谓抽样分布是指统计量的概率分布．确定统计量的分布是数理统计学的基本问题之一．关于统计量的分布，我们关心两类问题：(1) 当总体 X 的分布已知时，对于任一自然数 n，求出给定的统计量 $U_n=f(X_1,X_2,\cdots,X_n)$ 的分布，这个分布称为统计量的精确分布．它对数理统计中的所谓小样问题（即样本容量 n 较小时的统计问题）的研究是很重要的．(2) 当 $n\to\infty$ 时，求统计量 U_n 的极限分布，统计量的极限分布对于数理统计中的所谓大样问题（即样本容量 n 较大时的统计问题）的研究很有用处．

一、X^2 分布

定义 1.6 设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且同服从于标准正态分布 $N(0,1)$，则称随

机变量

$$\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \tag{1.1}$$

所服从的分布为自由度为 n 的 χ^2 分布，记为 $\chi_n^2 \sim \chi^2(n)$. 这里自由度 n 表示式(1.1)中独立变量的个数. 随机变量 χ_n^2 亦称为 χ^2 变量.

如果平方和 $\sum_{i=1}^{n} X_i^2$ 中，$X_1, X_2, \cdots, X_n$ 之间存在着 k 个独立的线性约束条件，则称 $\sum_{i=1}^{n} X_i^2$ 的自由度为 $n-k$(即自由变量的个数). 由于式(1.1)中 $X_1, X_2, \cdots, X_n$ 之间没有线性约束条件，即 $k=0$，所以 χ_n^2 的自由度为 n.

定理 1.5 由式(1.1)定义的随机变量 χ_n^2 的分布密度为

$$\varphi(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, & x > 0 \\ 0, & x \leqslant 0 \end{cases} \tag{1.2}$$

其中 $\Gamma\left(\frac{n}{2}\right)$ 是伽玛函数 $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$ 在 $\alpha = \frac{n}{2}$ 处的值.

证明 采用数学归纳法证明.

当 $n=1$ 时，$\chi_1^2 = X_1^2$，而 $X_1 \sim N(0,1)$，即 X_1 的分布密度是

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad -\infty < x < +\infty$$

由于 χ_1^2 只取非负值，当 $y \leqslant 0$ 时，显然，它的分布密度 $\varphi(y) = 0$. 又因为 $y = x^2$ 在 $x \leqslant 0$ 和 $x > 0$ 时分别是单调降与单调增的，所以当 $y > 0$ 时，χ_1^2 的分布密度为

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} |(\sqrt{y})'| + \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} |(-\sqrt{y})'| = \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}$$

所以当 $n=1$ 时(1.2)式成立.

假设 $n=k$ 时(1.2)式成立，即 $\chi_k^2 = X_1^2 + X_2^2 + \cdots + X_k^2$ 的分布密度为

$$\varphi(x) = \begin{cases} \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

当 $n = k+1$ 时，$\chi_{k+1}^2 = (X_1^2 + X_2^2 + \cdots + X_k^2) + X_{k+1}^2 = \chi_k^2 + X_{k+1}^2$. 由于 χ_{k+1}^2 取非负值，当 $x \leqslant 0$ 时，它的分布密度 $\varphi(x) = 0$.

当 $x > 0$ 时，利用两个独立随机变量和的分布密度的卷积公式，有

$$\varphi(x) = \int_0^x \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} t^{\frac{k}{2}-1} e^{-\frac{t}{2}} \frac{1}{2^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)} (x-t)^{\frac{1}{2}-1} e^{-\frac{x-t}{2}} dt =$$

$$\frac{e^{-\frac{x}{2}}}{2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)} \int_0^x t^{\frac{k}{2}-1} (x-t)^{\frac{1}{2}-1} dt \xlongequal{\text{令 } u = t/x}$$

$$\frac{e^{-\frac{x}{2}} x^{\frac{k+1}{2}-1}}{2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)} \int_0^1 u^{\frac{k}{2}-1} (1-u)^{\frac{1}{2}-1} du =$$

$$\frac{e^{-\frac{x}{2}}x^{\frac{k+1}{2}-1}}{2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{1}{2}\right)}B\left(\frac{k}{2},\frac{1}{2}\right)=\frac{1}{2^{\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right)}x^{\frac{k+1}{2}-1}e^{-\frac{x}{2}}$$

其中 $B(\frac{k}{2},\frac{1}{2})$ 是贝塔函数在 $(\frac{k}{2},\frac{1}{2})$ 处的值. 由此可见式(1.2)对 $n=k+1$ 时成立.

χ_n^2 分布密度函数曲线如图 1.1 所示,它随 n 取不同的值而不同.

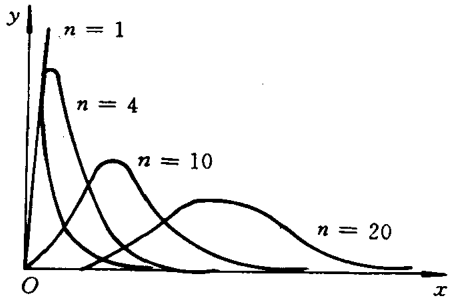


图 1.1 χ^2-分布的密度函数

例 1.5 设 $X_1,X_2,\cdots,X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的一个样本,求随机变量

$$Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

的概率分布.

解 因为 $X_1,X_2,\cdots,X_n$ 相互独立,且 $X_i\sim N(\mu,\sigma^2)$ $(i=1,2,\cdots,n)$.

作变换

$$Y_i=\frac{X_i-\mu}{\sigma}\quad(i=1,2,\cdots,n)$$

显然 $Y_1,Y_2,\cdots,Y_n$ 相互独立,且 $Y_i\sim N(0,1)$ $(i=1,2,\cdots,n)$. 因此由定义 1.6 得

$$Y=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}Y_i^2$$

服从自由度为 n 的 χ^2 分布.

χ^2 分布具有下列性质:

性质 1 $E\chi_n^2=n$, $\quad D\chi_n^2=2n$.

证明 由定义 1.6,并注意到 $X_1,X_2,\cdots,X_n$ 相互独立,且 $EX_i=0,DX_i=1(i=1,2,\cdots,n)$. 有

$$E\chi_n^2=E\Big(\sum_{i=1}^{n}X_i^2\Big)=\sum_{i=1}^{n}EX_i^2=\sum_{i=1}^{n}[DX_i+(EX_i)^2]=n$$

$$D\chi_n^2=D\Big(\sum_{i=1}^{n}X_i^2\Big)=\sum_{i=1}^{n}DX_i^2=2n$$

上式最后一个等号用到

$$DX_i^2=EX_i^4-(EX_i^2)^2=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}x^4e^{-\frac{x^2}{2}}dx-1=3-1=2$$

性质 2 若 $\chi_1^2\sim\chi^2(n),\chi_2^2\sim\chi^2(m)$,且 χ_1^2 与 χ_2^2 相互独立,则

$$\chi_1^2+\chi_2^2\sim\chi^2(n+m)$$

证明 令 $Z=\chi_1^2+\chi_2^2$. 由于 Z 只取非负值,当 $z\leqslant 0$ 时,Z 的分布密度

$$\varphi(z)=0$$

当 $z>0$ 时,利用求独立随机变量和的分布密度的卷积公式,有

$$\varphi(z)=\int_0^z\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}(z-x)^{\frac{m}{2}-1}e^{-\frac{z-x}{2}}dx=$$

$$\frac{1}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}e^{-\frac{z}{2}}\int_0^z x^{\frac{n}{2}-1}(z-x)^{\frac{m}{2}-1}dx\xrightarrow{令\,u=x/z}$$

$$\frac{1}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}e^{-\frac{z}{2}}z^{\frac{n+m}{2}-1}\int_0^1 u^{\frac{n}{2}-1}(1-u)^{\frac{m}{2}-1}du=$$

$$\frac{1}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}e^{-\frac{z}{2}}z^{\frac{n+m}{2}-1}B\left(\frac{n}{2},\frac{m}{2}\right)=$$

$$\frac{1}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n+m}{2}\right)}z^{\frac{n+m}{2}-1}e^{-\frac{z}{2}}$$

即 $$Z=\chi_1^2+\chi_2^2\sim\chi^2(n+m).$$

性质2称为χ^2分布的可加性．这个性质还可以推广到多个变量的情形，即n个相互独立的χ^2变量之和亦是χ^2变量，且它的自由度等于各个χ^2变量相应自由度之和．

性质3 设$\chi_n^2\sim\chi^2(n)$，则对任意x，有

$$\lim_{n\to\infty}P\left\{\frac{\chi_n^2-n}{\sqrt{2n}}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

证明 由假设χ_n^2可表示成n个相互独立的标准正态变量$X_1,X_2,\cdots,X_n$的平方和，即

$$\chi_n^2=X_1^2+X_2^2+\cdots+X_n^2$$

显然，$X_1^2,X_2^2,\cdots,X_n^2$独立同分布，且$\mu=EX_i^2=1,\sigma^2=DX_i^2=2(i=1,2,\cdots,n)$，由中心极限定理得

$$\lim_{n\to\infty}P\left\{\frac{\chi_n^2-n}{\sqrt{2n}}\leqslant x\right\}=\lim_{n\to\infty}P\left\{\frac{\sum_{i=1}^{n}X_i^2-n\mu}{\sqrt{n}\,\sigma}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

此性质说明χ^2变量的极限分布是正态分布，因而，当n很大时，$\frac{\chi_n^2-n}{\sqrt{2n}}$近似服从标准正态分布$N(0,1)$，亦即$n$很大时$\chi_n^2$近似服从正态分布$N(n,2n)$.

下面介绍一个比性质2更为深刻的结论 —— 柯赫伦(Cochran)分解定理．

定理1.6（柯赫伦分解定理） 设$X_1,X_2,\cdots,X_n$相互独立且$X_i\sim N(0,1)$ $(i=1,2,\cdots,n)$．令$Q=\sum_{i=1}^{n}X_i^2$，Q是自由度为n的χ^2变量．若Q可以分解成

$$Q=Q_1+Q_2+\cdots+Q_k$$

其中$Q_i(i=1,2,\cdots,k)$是秩为n_i的关于$(X_1,X_2,\cdots,X_n)$的非负二次型．则$Q_i(i=1,2,\cdots,k)$，相互独立且$Q_i\sim\chi^2(n_i)$ $(i=1,2,\cdots,k)$的充要条件是

$$n_1+n_2+\cdots+n_k=n$$

定理的必要性依χ^2变量的可加性是显然的，充分性的证明需要用到较多的线性代数知识，故从略．该定理在第四章方差分析中起着重要的作用．它将被这样应用：如果由$(X_1,X_2,\cdots,X_n)$构成的自由度为n的χ^2变量Q能够分解成若干个关于$(X_1,X_2,\cdots,X_n)$的非负二次型，那么只要这若干个二次型的秩之和为n，则每个二次型均服从χ^2分布，且分布的自由度等于相应于该二次型的秩．

二、t分布

定义1.7 设$X\sim N(0,1),Y\sim\chi^2(n)$，且$X$与$Y$相互独立，则称随机变量