21世纪高等院校优秀教材

概率论与数理统计（第2版）

The Theory of Probability and Statistics

刘海燕 编　　赵联文 审

下

国防工业出版社

21世纪高等院校优秀教材

概率论与数理统计(第2版)

(下册)

刘海燕　编
赵联文　审

国防工业出版社
·北京·

图书在版编目(CIP)数据

概率论与数理统计/李裕奇,刘海燕编.—2版.
北京:国防工业出版社,2004.8
ISBN 7-118-03542-4

Ⅰ.概...　Ⅱ.①李...②刘...　Ⅲ.①概率论②数理统计　Ⅳ.021

中国版本图书馆CIP数据核字(2004)第065349号

国防工業出版社出版发行
(北京市海淀区紫竹院南路23号)
(邮政编码　100044)
北京奥隆印刷厂印刷
新华书店经售
*
开本710×960　1/16　印张29¾　565千字
2004年8月第2版　2004年8月北京第4次印刷
印数:15001—22000册　总定价:38.00元·上册20.00元 下册18.00元

前　言

本书是依据高等学校工科数学教材编审委员会审定的《概率论与数理统计教学大纲》，及编者多年来从事概率论与数理统计课程的教学经验，专为高等院校学生编写的教材。为方便读者使用，全书分为上、下册，上册主要介绍概率论的基础理论知识及简单应用，由李裕奇负责编写；下册主要介绍数理统计的基础知识与应用，以及部分多元统计知识介绍，由刘海燕负责编写。

概率论与数理统计是一门专门研究和探索客观世界中随机现象统计规律性的科学，它在自然科学与社会科学的许多领域中得到广泛的应用；它在金融、保险、经济与企业管理、工农业生产、军事、医学、地质学、空间技术、气象与灾害预报，以及许多新兴学科与边缘学科，如信息论、排对论，决策论、生物统计、统计物理、人工智能、控制论、金融数学及其它工程技术的发展中都作出了非常重要的贡献。反过来，这些领域中提出的许多新的课题进一步促进了概率论与数理统计的发展，因而概率论与数理统计是近代数学最活跃的分支之一。鉴于概率论与数理统计的重要性，目前研究生入学数学统考试题中，概率论与数理统计所占比例已达到 20%～25%。各高等院校的理、工、医、农、金融、管理与许多文科专业都设置了概率论与数理统计课程，因此，学习与掌握概率论与数理统计的基本理论与应用，不仅是将来从事科学研究与工程实际工作的需要，也是继续学习现代科学技术与个人深造的需要，也是高度发展的现代科学技术对现代化人才提出的要求。

本书系统地阐述了概率论与数理统计的基本概念、基本思想、基本原理与基本方法，注重理论联系实际，突出解题思路，详尽介绍各种概率模型的掌握与应用，介绍多种解题方法与方法之间的联系，并使解题方法条理化，使读者便于学习与记忆，并受到很好的基本训练。在每一章末都注明了对学生的基本学习要求，配备了大量的思考题，基本练习、综合练习与自测题，使学生通过学习容易了解自己学习的成绩。章节安排循序渐进，层次分明，前后呼应，便于教学，便于读者更好、更快地学习掌握及运用概率论与数理统计知识。

概率论与数理统计讲授各需 34 学时左右，读者只需具备工科高等数学的基本知识，就可顺利阅读全书。

本教材是西南交通大学应用数学系概率统计教研室与西南交通大学统计咨询中心全体同仁通力合作的结晶，是工科数学概率论与数理统计教材改革的成果之

一。该教材在编写与出版过程中，得到西南交通大学应用数学系与概率统计教研室、教务处、教材科与国防工业出版社的鼎力支持与帮助，编者谨此一并表示由衷的感谢。

由于编者水平有限，书中难免会有错误与不妥之处，恳请同行及读者批评指正。

编　者

2004年6月于成都

内容简介

本书是专为高等院校学生学习概率论与数理统计课程编写的教材,也可作为有关专业的参考书与从事概率论与数理统计相关工作的科研与工程技术人员的参考书。

本书分为上、下册,共 10 章,上册包括概率论的基本概念;随机变量及其分布;多维随机变量及其分布;随机变量的数字特征;大数定律与中心极限定说及概率论的简单应用等知识。下册包括数理统计的基本概念;样本分布;参数估计;假设检验;线性统计推断以及最常用的多元统计方法。

本书每章节末都配有大量的思考题、基本练习,综合练习与自测题,帮助读者循序渐进地牢固地掌握概率论与数理统计知识。

目　录

（下册）

第六章　样本分布

§6.1　总体与样本

为了更好地介绍数理统计所研究的问题，先引入一些常用的概念和术语。

一、总体与样本

定义 6.1.1　研究对象的全部元素组成的集合，称为总体，组成总体的每一个元素称为个体。

例 6.1.1　考察某地区全体居民的身高情况，则该地区所有人的身高便构成一个总体，而每个人的身高就是一个个体。

例 6.1.2　检查一批电视机的质量，则该批电视机的全体就可看成一个总体，其中一个电视机就是一个个体。

首先总体可以是有限的，也可以是无限的，其个数称为总体容量，通常用 n 表示。

在研究总体性质时，我们关心的往往是总体的一个或几个数量指标，如检查电视机质量时往往是检查“电视机的寿命”。考察人时往往是身高，体重等。这些数量指标是通过每一个个体共同表现出来的，而总体中的每个个体不尽相同，因此，刻画总体的数量指标 X 可看作一个随机变量。为了研究方便，我们通常把随机变量 X 取值的全体当作总体，称为总体 X。而把每一个体的数值看成是 X 的一个观测值。这样我们可以借助随机变量来研究总体，或者说总体就是一个随机变量。如果要研究的数量指标不止一个，那么可分为几个总体来研究。

对全部个体即总体的检验是困难的或者是不可能做到的，一个可能达到的，最恰当的替代办法是在这一总体中检验一小部分。利用这部分个体提供的信息，对总体特征作出某种程度的推断。例如检验一批电视机的寿命，不可能把所有电视机都用来做试验。但我们可以从这批电视机中随机抽取 n 个，获得总体 X 的 n 个观察值 $x_1,x_2,\cdots,x_n$，然后根据这组观察值对整体电视机的寿命（总体 X）作出分析和判断。

从总体中随机抽取的 n 个个体称为容量为 n 的子样。我们要从子样的观察或试验结果的特性来对总体的特性作出估计与推断，一方面自然要研究应该怎样从

总体中抽取子样，使得子样尽可能大的反映总体的特性；同时必须建立一整套的方法，使能根据所选取的子样的性质，来对总体的特性进行估计与推断。因此，我们在抽取子样来对总体作出估计与推断时，从总体中抽取子样必须是随机的，即每一个个体都有同等概率被抽取（当总体中的个体是有限时，要采用有返回抽取方式）。其具体要求为两个方面：独立性，是指 $X_1,X_2,\cdots,X_n$ 相互独立；代表性，是指 $X_1,X_2,\cdots,X_n$ 中每一个都与总体 X 有相同分布。

定义 6.1.2 （简单随机子样）设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的容量为 n 的子样，如果 $X_1,X_2,\cdots,X_n$ 相互独立，且每一个都是与总体 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 的简单随机子样，简称为简单子样或子样。

在以后的章节中，我们所讲的子样 $X_1,X_2,\cdots,X_n$，无特别声明的话，都是指简单子样，即 $X_1,X_2,\cdots,X_n$ 相互独立，且每个都与总体 X 有相同的分布。引入简单随机子样，是基于要从子样所获得的统计特性来对总体的统计特性作出估计与推断。

二、样本分布

由上所述，样本是一 n 维随机向量，依概率论知，它必有其概率分布函数。所谓样本分布即指样本的概率分布函数，对于简单随机样本 $(X_1,X_2,\cdots,X_n)$，它的分布函数（联合分布函数）可由总体 X 的分布函数 $F(x)$ 完全决定，其表达式为

$$P\{X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n\}$$

$$\overset{①}{=}P\{X_1\leqslant x_1\}\cdot P\{X_2\leqslant x_2\}\cdot\cdots\cdot P\{X_n\leqslant x_n\}$$

$$F_n(x_1,x_2,\cdots,x_n)\overset{②}{=}F_X(x_1)\cdot F_X(x_2)\cdot\cdots\cdot F_X(x_n)=\prod_{i=1}^{n}F_X(x_i)$$

① 处等号是因为 $X_1,X_2,\cdots,X_n$ 独立，② 处因为 $X_1,X_2,\cdots,X_n$ 同分布。

特别地，如果总体 X 是连续型随机变量，且具有概率密度函数 $f(x)$，则样本 $(X_1,X_2,\cdots,X_n)$ 的概率密度函数（联合分布密度）为

$$f_n(x_1,x_2,\cdots,x_n)=f(x_1)\cdot f(x_2)\cdot\cdots\cdot f(x_n)=\prod_{i=1}^{n}f(x_i)$$

如果总体 X 是离散型随机变量，则其分布律为

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}=\prod_{i=1}^{n}P\{X_i=x_i\}=\prod_{i=1}^{n}p(x_i)$$

其中 $P\{X=x\}=p(x)$ 为离散型随机变量 X 的概率分布。

例 6.1.3 某厂生产的电容器的使用寿命服从指数分布，但参数 λ 未知，为此任意抽查 n 只电容器，测其实际使用寿命，试问此例中什么是总体，样本及其分布。

解：总体 X 表示一个电容器的使用寿命，服从参数为 λ 的指数分布，其概率密度为

$$f_X(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\leqslant 0\end{cases}$$

样本$(X_1,X_2,\cdots,X_n)$表示所抽取的n只电容器中各只电容器的使用寿命，由于$X_1,X_2,\cdots,X_n$相互独立，并且与X一样服从参数为λ的指数分布，所以样本的联合概率密度函数为

$$f_X(x_1,x_2,\cdots,x_n)=\begin{cases}\lambda^n e^{-\lambda(x_1+x_2+\cdots,x_n)} & x_1,x_2,\cdots,x_n>0\\ 0 & \text{其它}\end{cases}$$

三、统计量

子样$X_1,X_2,\cdots,X_n$也可看作为n维随机向量$(X_1,X_2,\cdots,X_n)$。记x_i为X_i一次观察值，并称$(x_1,x_2,\cdots,x_n)$为子样的一次观察值。

定义6.1.3　设$X_1,X_2,\cdots,X_n$为总体X的样本，若子样的函数$T=T(X_1,X_2,\cdots,X_n)$且不含任何未知参数，则$T(X_1,X_2,\cdots,X_n)$称为统计量。

若$(x_1,x_2,\cdots,x_n)$为子样的一次观察值，则T的取值$t=T(x_1,x_2,\cdots,x_n)$称为统计值。

我们在用子样$X_1,X_2,\cdots,X_n$获得的信息来对总体X作出估计与推断时，按不同的要求规定子样的各种函数。

例6.1.4　设$X_1,X_2,\cdots,X_n$是来自正态总体$N(\mu,\sigma^2)$的容量为n的样本，其中μ已知，σ未知，则$\sum\limits_{i=1}^{n}(X_i-\mu)^2$，$X_1+X_2$，都是统计量；而$\frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$不是统计量，因为它包含未知参数$\sigma$。

例6.1.5　设$X_1,X_2,\cdots,X_n$为总体X的容量为n的子样，记作

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

则$\bar{X}$及S^2都是统计量，称$\bar{X}$及S^2分别为子样$X_1,X_2,\cdots,X_n$的平均值及方差。子样的观察值为$x_1,x_2,\cdots,x_n$，$\bar{X}$及S^2的观察值分别记作

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\qquad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$C_r=\dfrac{S}{\bar{X}}$称为变异系数，

$$c_r=\frac{s}{\bar{x}}$$

今后，大写的S^2表示统计量，小写的s^2表示统计量S^2的观察值。

$A_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^k\ (k=1,2,\cdots)$为样本$k$阶原点矩；

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

且它们相应的观察值依次为

$$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k\quad k=1,2,\cdots$$

$$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k\quad k=1,2,\cdots$$

定义 6.1.4（顺序统计量） 设 $X_1,X_2,\cdots,X_n$ 为总体 X 的子样，今由子样建立 n 个函数：

$$X_k^*=X_k^*(X_1,X_2,\cdots,X_n)\quad k=1,2,\cdots,n$$

其中 X_k^* 为这样的统计量，它的观察值为 x_k^*，x_k^* 为子样 $X_1,X_2,\cdots,X_n$ 的观察值 $x_1,x_2,\cdots,x_n$ 中按大小由小至大排列：

$$x_1^*\leqslant x_2^*\leqslant\cdots\leqslant x_k^*\leqslant\cdots\leqslant x_n^*$$

后的第 k 个数值，称 $X_1^*,X_2^*,\cdots,X_n^*$ 为顺序统计量。

易见，$X_1^*=\min\{X_1,X_2,\cdots,X_n\}$，$X_n^*=\max\{X_1,X_2,\cdots,X_n\}$。若 n 为奇数，称 $X_{\frac{n+1}{2}}^*$ 为子样的中位数；若 n 为偶数，则称 $\frac{1}{2}\left(X_{\frac{n}{2}}^*+X_{\frac{n}{2}+1}^*\right)$ 为子样的中位数。即

$$M_n=\begin{cases}X_{\frac{n+1}{2}}^* & n\text{ 为奇数}\\ \frac{1}{2}\left[X_{\frac{n}{2}}^*+X_{\frac{n}{2}+1}^*\right] & n\text{ 为偶数}\end{cases}$$

定义 6.1.5（极差） 设 $X_1,X_2,\cdots,X_n$ 为总体 X 的子样，则称统计量 $D_n^*=X_n^*-X_1^*$ 为子样的极差。它是子样中最大值与最小值之差，反映了子样观察值的波动幅度。它同方差一样是反映观察值离散程度的数量指标，而且计算最方便。

例 6.1.6 设 $X_1,X_2,\cdots,X_5$ 为 X 的容量为 5 的子样，今对这个子样作了 3 次观察，其值如下表所列，试求 $\bar{X}$，S^2 及 D_n^* 的观察值。

X / x	X_1	X_2	X_3	X_4	X_5
1	3	1	10	5	6
2	2	6	7	2	8
3	8	3	9	10	5

解：得

X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	$\bar{x}$	s^2	D_n^*	中位数 M_n
1	3	5	6	10	5	11.5	9	5
2	2	6	7	8	5	8	6	6
3	5	8	9	10	7	8.5	7	8

例 6.1.7 某工厂制作一种线圈，为了控制生产过程保持稳定，从产品中任取

10 件,测定其电阻阻抗值 X(单位:Ω) 所得数据如下:

15.3, 13.0, 16.7, 14.2, 14.5, 14.5, 15.9, 15.0, 15.1, 16.4

试求:(1) 样本中位数 M_n 的值;

(2) 若取第 11 件数据为 15.2,此时 M_n 又为何值;

(3) 极差。

解:先将所得数据从小到大顺序排列为

13.0, 14.2, 14.5, 14.5, 15.0, 15.1, 15.3, 15.9, 16.4, 16.7

(1)$n = 10$ 为偶数,故

$$M_n = \frac{1}{2}\left[X^*_{\left(\frac{n}{2}\right)} + X^*_{\left(\frac{n}{2}+1\right)}\right] = \frac{1}{2}(15.0 + 15.1) = 15.05,$$

(2) 当 $X_{11} = 15.2$ 时,$n = 11$ 为奇数,故

$$M_n = X^*_{\left(\frac{n+1}{2}\right)} = X^*_6 = 15.1$$

(3) $\nabla^*_n = 16.7 - 13.0 = 3.7$

众数(mod):数据中最常出现的值为众数,即样本中出现可能性最大的值,众数可能不惟一。

例 6.1.8　现有一数据集合:{2,3,3,3,3,4,4,5,6,6,6,6,6,7,7,8},其中每一个值出现的次数如下:

数值	2	3	4	5	6	7	8
出现次数	1	4	2	1	5	2	1

解:可见,众数为 6。

四、样本平均数和样本方差的简算公式

当样本值很复杂时,样本平均数和样本方差的计算常采用以下简算公式。

设 $x_1, x_2, \cdots, x_n$ 为样本的 n 个观察值。

(1) 选择适当常数 a,令 $y_i = x_i - a(i = 1,2,\cdots,n)$ 获得一组较简单数据 y_1, $y_2, \cdots, y_n$。

由
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(y_i + a) = \frac{1}{n}\sum_{i=1}^{n} y_i + a$$

则
$$\bar{x} = \bar{y} + a$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(y_i + a) - (\bar{y} + a)]^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

即
$$S_x^2 = S_y^2$$

例 6.1.9 取某型号火箭 8 枚进行射程试验，测得数据(单位:km) 如下：

$$54,\quad 52,\quad 49,\quad 57,\quad 43,\quad 47,\quad 50,\quad 51$$

试计算样本平均数和样本方差。

解:(1) 取 $a=50$,由 $y_i = x_i - 50$ 把样本值变为

$$y_i: 4,\quad 2,\quad -1,\quad 7,\quad -7,\quad -3,\quad 0,\quad 1,$$

则
$$\bar{y} = \frac{1}{8}\times(4+2-1+7-7-3+0+1) = 0.375$$

$$S_y^2 = \frac{1}{8-1}\left(\sum_{i=1}^{8} y_i^2 - 8\times 0.375^2\right) = 18.27$$

所以
$$\bar{x} = \bar{y} + 50 = 0.375 + 50 = 50.375$$

$$S_x^2 = S_y^2 = 18.27$$

(2) 选择适当常数 a 及非零常数 c,作变换,使

$$y_i = \frac{x_i - a}{c} \quad i = 1,2,\cdots,n$$

获得一组简单数据 $y_1, y_2, \cdots, y_n$。容易计算

$$\bar{x} = c\bar{y} + a,\quad S_x^2 = c^2 S_y^2$$

简化计算的思路是通过常数 a、c 的选择,使得样本值经过变换后尽量简单。用简单数据算出 $\bar{y}$ 和 S_y^2 后,再计算 $\bar{x}$ 和 S_x^2。

例 6.1.10 计算下列样本值的平均数和方差：

$$0.806,\quad 0.882,\quad 0.794,\quad 0.800,\quad 0.785,$$
$$0.793,\quad 0.817,\quad 0.820,\quad 0.771$$

解:取 $a = 0.800, c = 0.001$,通过 $y_i = \dfrac{x_i - 0.800}{0.001}$ 把样本值变换为

$$y_i: 6,\quad 82,\quad -6,\quad 0,\quad -15,\quad -7,\quad 17,\quad 20,\quad -29$$

容易计算：

$$\bar{y} = \frac{1}{9}\times(6+82-6+0-15-7+17+20-29) \approx 7.56$$

$$S_y^2 = \frac{1}{9-1}\left(\sum_{i=1}^{9} y_i^2 - 9\times 7.56^2\right) \approx 1010.7$$

则

$$\bar{x} = 0.001\times 7.56 + 0.800 = 0.80756$$

$$S_x^2 = 0.001^2 \times 1010.7 = 0.0010107$$

例 6.1.11 某厂某种悬式绝缘子机电的破坏负荷数值分组列表如下：

组限	5.5 ~ 6	6 ~ 6.5	6.5 ~ 7	7 ~ 7.5	7.5 ~ 8	8 ~ 8.5	8.5 ~ 9	9 ~ 9.5	9.5 ~ 10
频数	4	3	15	42	49	78	50	31	5

若各组以该组中位数作为此样本的数值,近似计算样本均值和样本方差。

解:首先计算各组的组中值,得如下分布表:

组中位数	5.75	6.25	6.75	7.25	7.75	8.25	8.75	9.25	9.75
频数	4	3	15	42	49	78	50	31	5

容易计算:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{9} m_i x_i = \frac{1}{277}(4 \times 5.75 + 3 \times 6.25 + 15 \times 6.75 + \cdots + 5 \times 9.75) = 8.1$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{9} m_i (x_i - \bar{x})^2 = \frac{1}{276}[4 \times (5.75 - 8.1)^2 + 3 \times (6.25 - 8.1)^2 + \cdots + 5 \times (9.75 - 8.1)^2] = 0.79$$

基本练习 6.1

1. 设总体 $X \sim N(\mu,\sigma^2)$,$X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本,试写出 $X_1, X_2, \cdots, X_n$ 的联合概率分布密度函数。

2. 设总体 $X \sim U(a,b)$(均匀分布),$X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本,试写出 $X_1, X_2, \cdots, X_n$ 的联合概率分布密度函数。

3. 已知总体 $X \sim N(\mu,\sigma^2)$,其中 μ 未知,$\sigma > 0$ 为已知,样本 $X_1, X_2, \cdots, X_n$ 的下列函数是否为统计量?

(1) $\frac{1}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$;　　(2) $\frac{1}{\sigma^2}(X_1^2 + X_2^2 + \cdots + X_n^2)$;

(3)$(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots (X_n - \mu)^2$。

4. 设总体 $X \sim U(a,b)$,其中 a 已知,b 未知,设 X_1, X_2, X_3 是取自总体 X 中的一个样本,试指出:

(1)$X_1 + X_2, + X_3$;　　(2)$X_1 + 2a$;

(3)X_1;　　(4)$\max\{X_1, X_2, X_3\}$;

(5)X_3^2;　　(6)aX_1;

(7)$\sum_{i=1}^{3}\frac{X_i^2}{b^2}$;　　(8)$\frac{X_3 - X_1}{2}$;

(9)$X_2 + b$。

中哪些是统计量,哪些不是统计量?并说明理由。

5. 为调查土壤中的营养成分,随机采取土壤样品 8 个,测得其重量(单位:g)分别为

230,243,185,240,228,196,246,200

(1) 写出总体,样本,样本值,样本容量;

(2) 求样本均值、样本方差和标准差;

(3) 求顺序统计量、极差和中位数。

6. 从一批产品中随机抽取8件，测得它们的质量(单位:kg)如下：

143,100,146,130,185,140,128,196

试计算样本均值、样本方差、二阶原点矩、极差和中位数。

7. 某大学抽样调查100名男生的身高，资料如下：

身高 /cm	150 ~ 160	160 ~ 170	170 ~ 180	180 ~ 190
人数	5	36	44	15

求 $\bar{X}, S_n^2$。

8. 某地区抽样调整200个居民户的人均月收入，统计资料如下：

月人均收入 / 百元	5 ~ 6	6 ~ 7	7 ~ 8	8 ~ 9	9 ~ 10	10 ~ 11	11 ~ 12
户数	18	35	76	24	19	14	14

求样本容量 n，样本均值 $\bar{X}$，样本方差 S^2。

§6.2 样本分布函数

一、样本分布函数

定义6.2.1 设从总体 X 中抽取容量为 n 的子样 $X_1, X_2, \cdots, X_n$，当顺序统计量 $X_1^*, X_2^*, \cdots, X_n^*$ 的值固定时，对任何实数 x，我们定义函数 $F_n^*(x)$：

$$F_n^*(x) = \begin{cases} 0 & x < X_1^* \\ \dfrac{k}{n} & X_k^* \leqslant x < X_{k+1}^* \quad k = 1,2,\cdots,n-1 \\ 1 & x \geqslant X_n^* \end{cases}$$

称 $F_n^*(x)$ 为总体 X 样本分布函数或经验分布函数(见图 6.2.1)。

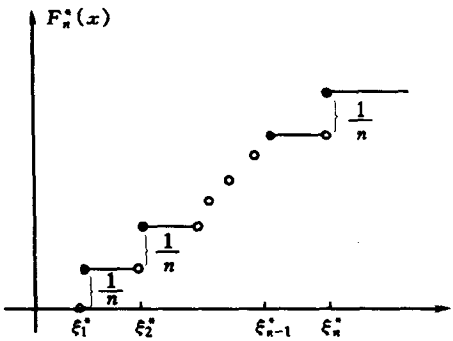

图 6.2.1

易见，$F_n^*(x)$ 是单调、非降、右连续且在 $x = X_k^*$ 处有间断点，在每个间断点上跳跃量都是$\frac{1}{n}$。显然，$0 \leqslant F_n^*(x) \leqslant 1$，并具有分布函数的其它性质。

由定义知,对于 x 的每一数值而言,经验分布函数 $F_n^*(x)$ 为子样 $X_1,X_2,\cdots,X_n$ 的函数,它是一统计量,即为一随机变数,其可能取值为 $0,\frac{1}{n},\cdots,\frac{n-1}{n},1$。事件"$F_n^*(x)=\frac{k}{n}$"发生的概率,由于 $X_1,X_2,\cdots,X_n$ 相互独立且有相同的分布函数 $F(x)$,因而它等价于 n 次独立重复试验的贝努利概型中事件"$X\leqslant x$"发生 k 次而其余 $n-k$ 次不发生的概率,即有

$$P\{F_n^*(x)=\frac{k}{n}\}=C_n^k\{F(x)\}^k\{1-F(x)\}^{n-k}$$

其中 $F(x)=P\{X\leqslant x\}$,它是总体 X 的分布函数。

对于任一实数 x,$F_n^*(x)$ 是事件"$X\leqslant x$"发生的频率,即在 n 个样本值中小于或等于 x 的个数与样本容量 n 的比值,这与分布函数的概念是一致的。从频率与概率的关系知道,$F_n^*(x)$ 可以作为未知总体分布函数 $F(x)$ 的一个近似,且由格列汶科定理知,对任何实数 x,记

$$D_n=\sup_{-\infty<x<+\infty}|F_n^*(x)-F_n(x)|$$

则有

$$P\{\lim_{n\to\infty}D_n=0\}=1$$

这也是为什么可以依据样本的性质来推断总体特征的依据。

例 6.2.1　从一批标准重量为500g的罐头中,随机抽取8听,测得误差(单位:g)如下:

$$8,-4,6,-7,-2,1,0,1$$

求经验分布函数,并作出图形。

解:将样本值按大小顺序排列为

$$-7<-4<-2<0<1=1<6<8$$

则其样本分布函数为

$$F_8^*(x)=\begin{cases}0 & x<-7\text{时}\\ \frac{1}{8} & -7\leqslant x<-4\\ \frac{2}{8} & -4\leqslant x<-2\\ \frac{3}{8} & -2\leqslant x<0\\ \frac{4}{8} & 0\leqslant x<1\\ \frac{6}{8} & 1\leqslant x<6\\ \frac{7}{8} & 6\leqslant x<8\\ 1 & x\geqslant 8\end{cases}$$