



普通高等教育“十五”国家级规划教材

应用数理统计

(第二版)

孙荣恒 编著

 科学出版社
www.sciencepress.com

普通高等教育“十五”国家级规划教材

应用数理统计

(第二版)

孙荣恒 编著

科学出版社

北京

内 容 简 介

本书是为应用数学专业、数学专业、概率统计专业、信息与计算科学专业本科大学生和非数学专业的硕士生学习数理统计而编写的教材。主要内容有：抽样分布、参数估计、假设检验、方差分析与正交试验设计、线性回归模型。本书每章末附有习题，书后附有答案。

本书可作为应用数学专业、数学专业、概率统计专业、信息与计算科学专业大学生和非数学专业的研究生，以及教师和科技工作者使用。

图书在版编目 (CIP) 数据

应用数理统计/孙荣恒编著. —2 版. —北京: 科学出版社, 2003

(普通高等教育“十五”国家级规划教材)

ISBN 7-03-011160-5

I. 应… II. 孙… III. 数理统计—高等学校—教材 IV. O212

中国版本图书馆 CIP 数据核字 (2003) 第 009086 号

责任编辑: 林 鹏 陈玉琢/责任校对: 刘小梅

责任印制: 安春生/封面设计: 王 浩

科 学 出 版 社 出 版

北京东黄城根北街16号

邮政编码: 100717

<http://www.sciencep.com>

深 海 印 刷 有 限 责 任 公 司 印 刷

科学出版社发行 各地新华书店经销

*

1998年9月第一版 开本: B5 (720×1000)

2003年5月第二版 印张: 17 1/2

2003年5月第三次印刷 字数: 330 000

印数: 6 001—10 000

定价: 21.00 元

(如有印装质量问题, 我社负责调换〈环伟〉)

第二版序

本书是 1998 年出版的《应用数理统计》的修改本. 除改正了原书的一些印刷错误外, 在第一章中, 对“总体”和定义 1.1.1 补充了一些引入说明, 对 χ^2 分布、 t 分布和 F 分布补充了一些(显而易见的)性质, 还增加了两个习题; 在第二章中, 增加了例 2.1.2、例 2.1.8, 引入“均方误差最小估计量”的理由、§ 2.6 和两个习题; 在第三章中, 增加了例 3.1.2 和 3.2.7 与 3.2.8 两小节; 在第四章中, 增加了例 4.3.1; 在第五章中补充了引入“线性模型”的说明; 其他基本照旧. 所增补的内容不仅使本书更加完整, 而且在实际中有重要的应用.

本书自 1992 年起就以讲义的形式在重庆大学使用. 1998 年由科学出版社出版后曾被国内多所大学选用. 2002 年被教育部选定为“普通高等教育‘十五’国家级规划教材”. 对上述单位作者表示衷心感谢!

孙荣恒
2002 年 7 月

序 言

数理统计起源于人口调查.早在公元前 3000 年古代的巴比伦、中国和埃及就已进行过人口调查,由此可知数理统计的历史源远流长.但是现代数理统计大规模发展始于 19 世纪末 20 世纪初.在 1856 年到 1863 年之间,孟德尔(Gregor Mendel)从一个科学实验中发现了遗传学的统计规律.1889 年左右高尔登(F. Galton)受达尔文(Charle Darwin)的《物种起源》一书的刺激,研究了平均值的偏差问题与回归问题,对生物统计学做出重要的贡献.1890 年卡·皮尔逊(Karl Pearson)受高尔登工作的激发,开始把数学与概率论应用于达尔文的进化论,从而开创了现代数理统计的时代.他一生致力于统计方法的研究,今天的描述性统计学的大部分内容是他整理出来的,大部分数理统计用语也是他命名的,这使得他赢得了“统计学之父”的称号.现代数理统计发展历史大致可分为两个阶段.第一阶段大致到第二次世界大战结束为止.在这一阶段中,对数理统计有重大影响的学者除卡·皮尔逊外,还有费希尔(R. A. Fisher)、奈曼(J. Neyman)、伊·皮尔逊(E. S. Pearson)等.他们从实际出发,推动了一些主要数理统计分支的建立,逐步为之建立了一套系统而严格的理论.第二个阶段是从第二次世界大战结束至今.在这个阶段中数理统计的研究向纵深发展,除把第一阶段中的不足和粗糙之处弥补外,还提出了许多新问题、新理论和新的研究方向.此外,数理统计在各个领域中的实际应用不仅加快了速度,而且也越来越普遍了,研究队伍也越来越壮大了.目前,我国几乎所有大学都有数理统计课,而且不少大学建立了统计系或统计专业.

本书是为应用数学专业、数学专业、概率统计专业、信息与计算科学专业本科生和非数学专业的硕士生学习数理统计而编写的教材,是在 1992 年第一次印刷的基础上经较大规模的修改而成的,并曾在重庆大学应用数学系使用过多次.本书主要介绍抽样分布、参数估计、假设检验、方差分析与正交试验设计、线性回归模型.内容系统丰富、推导简明严谨,强调应用是本书的另一特色.

本书共五章,每章后附有适量的习题,书末附有答案.经适当选择后,本书也可作为其他理工科本科大学生的教材.

初稿完成后,伊亨云教授审阅全书,提出了许多宝贵意见.李幼英同志为初稿的打印付出了辛勤劳动,作者在此向他们表示衷心感谢!

由于作者水平所限,虽经多次使用和修改,书中一定还存在不少缺点和错误,恳请读者批评指正!

孙荣恒

1997.11

目 录

第二版序

序言

第一章 抽样分布	1
§ 1.1 基本概念、顺序统计量与经验分布函数	1
1.1.1 基本概念	1
1.1.2 顺序统计量	3
1.1.3 经验分布函数	6
1.1.4 几个重要分布	8
§ 1.2 多元正态分布与正态二次型	11
§ 1.3 抽样分布定理	19
§ 1.4 分位数	22
习题一	24
第二章 参数估计	29
§ 2.1 点估计常用方法	29
2.1.1 矩法	29
2.1.2 极大似然法	32
§ 2.2 评价估计量好坏的标准	36
2.2.1 无偏性与有效性	36
2.2.2 一致最小方差无偏估计量	44
2.2.3 一致性(相合性)	47
§ 2.3* 充分性与完备性	48
2.3.1 充分性	49
2.3.2 完备性	52
§ 2.4 区间估计	57
2.4.1 一个正态总体的情况	58
2.4.2 两个正态总体的情况	61
2.4.3 指数分布与0—1分布参数的区间估计	65
§ 2.5 贝叶斯(Bayes)估计	67
2.5.1 决策论的基本概念	67
2.5.2 最大风险最小化估计	68

2.5.3 后验分布	70
2.5.4 贝叶斯估计	73
2.5.5 先验分布的选取	78
2.5.6 最大后验估计	82
2.5.7 贝叶斯区间估计	83
§ 2.6 截尾寿命试验中指数分布的参数估计	85
2.6.1 指数分布中参数的点估计	86
2.6.2 指数分布中参数的区间估计	89
习题二	90
第三章 假设检验	97
§ 3.1 假设检验的基本思想与基本概念	97
§ 3.2 参数假设检验	101
3.2.1 单个正态总体均值的假设检验	102
3.2.2 单个正态总体方差的假设检验	108
3.2.3 两个正态总体均值的假设检验	113
3.2.4 两个正态总体方差的假设检验	116
3.2.5 广义似然比检验	124
3.2.6 似然比检验	127
3.2.7 指数分布中参数 λ 的假设检验	128
3.2.8 截尾试验中指数分布参数的假设检验	129
§ 3.3 非参数假设检验	130
3.3.1 分布函数的拟合检验	131
3.3.2 两总体之间关系的假设检验	141
§ 3.4* 一致最优势检验	149
3.4.1 势函数	149
3.4.2 奈曼-皮尔逊基本引理	151
§ 3.5* 质量控制	157
3.5.1 验收抽样方案的制订	158
3.5.2 计量控制	161
3.5.3 计件控制与计点控制	164
习题三	166
第四章 方差分析与正交试验设计	171
§ 4.1 单因素方差分析	171
4.1.1 数学模型	171
4.1.2 方差分析	172

§ 4.2* 双因素方差分析	177
4.2.1 数学模型	177
4.2.2 方差分析	178
§ 4.3 正交试验设计	184
4.3.1 正交表	184
4.3.2 正交表的分析	187
习题四	191
第五章 线性回归模型	193
§ 5.1 线性模型	193
§ 5.2 最小二乘法估计	196
5.2.1 β 的最小二乘法估计	196
5.2.2 最小二乘法估计量的性质	198
5.2.3 例子	204
§ 5.3 检验、预测与控制	209
5.3.1 线性模型与回归系数的检验	209
5.3.2 预测与控制	214
§ 5.4 带有线性约束的线性回归模型	220
5.4.1 拉格朗日乘子法	220
5.4.2 β_H 的性质	221
5.4.3 对假设 $H_0: H\beta = d$ 的检验	223
§ 5.5* 多项式回归与线性回归正交设计	227
5.5.1 正交多项式的应用	228
5.5.2 回归模型与回归系数的检验	232
5.5.3 多元正交多项式回归	234
5.5.4 线性回归正交设计	236
习题五	239
附录 常用数理统计表	243
答案	263
参考文献	269

第一章 抽样分布

与概率论一样,数理统计也是研究随机现象统计规律性的一门数学学科,概率论研究的基本内容是:在已知随机变量分布的情况下,着重讨论了随机变量的性质.但是对一个具体的随机变量来说,如何判断它服从某种分布.如果知道它服从某种分布又该如何确定它的各个参数.对于这些问题在概率论中都没有涉及到,它们都是数理统计所要研究的内容,并且这些问题的研究都直接或间接建立在试验的基础上.数理统计学是利用概率论的理论对所研究的随机现象进行多次的观察或试验,研究如何合理地获得数据,如何对所获得的数据进行整理、分析,如何对所关心的问题作出估计或判断的一门数学学科.其内容非常丰富.一般可分为两大类:一类是试验的设计与研究,一类是统计推断.我们着重讨论统计推断.

本章首先介绍数理统计的基本概念,然后介绍多元正态分布与正态二次型,最后介绍有关抽样分布的几个定理,为以后各章作必要的准备.

§ 1.1 基本概念、顺序统计量与经验分布函数

1.1.1 基本概念

总体、个体、样本是数理统计学中三个最基本的概念.我们称研究对象的全体为总体或母体.称组成总体的每个单元为个体.从总体中随机抽取 n 个个体,称这 n 个个体为容量是 n 的样本.

例如,为了研究某厂生产的一批灯泡质量的好坏,规定使用寿命低于 1000 小时的为次品.则该批灯泡的全体就为总体,每个灯泡就是个体.实际上,数理统计学中的总体是指与总体相联系的某个(或某几个)数量指标 ξ 取值的全体.比如,该批灯泡的使用寿命 ξ 的取值全体就是研究对象的总体.由于对不同的个体, ξ 取不同的值,且事先无法准确预言,所以 ξ 是随机变量,这时,我们就称 ξ 的概率分布或更简单地就称 ξ 为总体.为了判断该批灯泡的次品率,最精确的办法是把每个灯泡的寿命都测出来.然而,寿命试验是破坏性试验(即使试验是非破坏性的,由于试验要花费人力、物力和时间),我们只能从总体中抽取一部分,比如说, n 个个体进行试验.试验结果可得一组数值 (x_1, x_2, \dots, x_n) , 其中每个 x_i 是一次抽样观察的结果.由于我们要根据这些观察结果对总体进行推断,所以对每次抽取就有一定的要求,要求每次抽取必须是随机的、独立的,这样才能较好地反映总体情况.所谓随机的是指每个个体被抽到的机会是均等的,这样抽到的个体才具有代表性.所谓独

立的是指每次抽取之后不能改变总体的成分.这就要求:如果试验是非破坏性的且总体是有限的,抽取应该是有放回的;如果试验是破坏性的总体应该是无限的或是很大的.基于上述思想的抽样方法称为简单随机抽样.用简单随机抽样方法抽取 n 个个体进行试验,其结果是确定的一组数值 (x_1, \dots, x_n) ,但是这组数值 (x_1, \dots, x_n) 是随着每次抽样而改变的.因此 (x_1, \dots, x_n) 实际上是一个 n 维随机向量 (ξ_1, \dots, ξ_n) 的一次观察值.即在试验之前, (x_1, x_2, \dots, x_n) 实际上是随机向量 $(\xi_1, \xi_2, \dots, \xi_n)$.又因抽样是随机的、独立的,所以 ξ_1, \dots, ξ_n 是相互独立的 n 个随机变量,且每个都与总体 ξ 同分布.我们称 (ξ_1, \dots, ξ_n) 或 ξ_1, \dots, ξ_n 为总体 ξ 的容量为 n 的简单随机样本,简称为样本(如无特别说明我们今后只讨论简单随机样本),称每个 ξ_i 为样品.样本 (ξ_1, \dots, ξ_n) 的所有可能的观察值组成的集合 \mathcal{X} 称为样本空间.它是 n 维空间或其一个子集.这样样本 (ξ_1, \dots, ξ_n) 的一次观察值 (x_1, \dots, x_n) 就是样本空间 \mathcal{X} 中的一个点.即 $(x_1, \dots, x_n) \in \mathcal{X}$.

由于对总体进行统计推断的依据是样本提供的信息,然而样本是 n 维随机变量或 n 个随机变量,讨论起来很不方便.人们自然会想到能否用样本的函数代替样本对总体进行统计推断.当然,这个函数不能太任意了,最好是一个随机变量,这样使用起来才方便;同时这个函数中不能含有任何未知参数.由此,我们引入如下定义.

定义 1.1.1 设 (ξ_1, \dots, ξ_n) 为总体 ξ 的样本, $T(x_1, \dots, x_n)$ 为样本空间 \mathcal{X} 上的实值波雷尔可测函数.如果 $T(\xi_1, \dots, \xi_n)$ 中不包含任何未知参数,则称 $T(\xi_1, \dots, \xi_n)$ 为一个统计量.

例 1.1.1 设 (ξ_1, \dots, ξ_n) 为总体 ξ 的样本,记

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i, \quad S^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2, \quad (1.1.1)$$

则称 $\bar{\xi}, S^2$ 与 S 分别为样本 (ξ_1, \dots, ξ_n) 的均值、方差与标准.它们都是统计量.当 $D(\xi)$ 存在有限时,显然有

$$E(\bar{\xi}) = E(\xi), D(\bar{\xi}) = \frac{D(\xi)}{n}, E(S^2) = \frac{n-1}{n} D(\xi), \quad (1.1.2)$$

因此可用 $\bar{\xi}$ 来估计 $E(\xi)$, 用 S^2 估计 $D(\xi)$.

当 $(\xi_1, \xi_2, \dots, \xi_n)$ 的观察值为 (x_1, x_2, \dots, x_n) 时,记 $\bar{\xi}, S^2$ 的观察值分别为 \bar{x}, s^2 , 即

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1.1.3)$$

注意:统计量中不能包含任何未知参数.例如,设总体 $\xi \sim N(a, \sigma^2)$, 其中 a, σ^2 都是未知参数.设 (ξ_1, ξ_2) 为 ξ 的样本,则 $\frac{1}{2}(\xi_1 + \xi_2) - a$ 与 $\frac{\xi_1}{\sigma}$ 都不是统计量,而

ξ_2 与 $\xi_1 + \xi_2$ 都是统计量.

定义 1.1.2 设 (ξ_1, \dots, ξ_n) 为总体 ξ 的样本, 记

$$A_r = \frac{1}{n} \sum_{i=1}^n \xi_i^r, \quad B_r = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^r, \quad (1.1.4)$$

则称 A_r, B_r 分别为样本 (ξ_1, \dots, ξ_n) 的 r 阶原点矩与 r 阶中心矩. 显然 $A_1 = \bar{\xi}, B_2 = S^2$, 且 A_r, B_r 都是统计量.

定义 1.1.3 设 $(\xi_1, \eta_1), (\xi_2, \eta_2), \dots, (\xi_n, \eta_n)$ 为二维总体 (ξ, η) 的样本, 记

$$\begin{aligned} \bar{\xi} &= \frac{1}{n} \sum_{i=1}^n \xi_i, & S_1^2 &= \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2, \\ \bar{\eta} &= \frac{1}{n} \sum_{i=1}^n \eta_i, & S_2^2 &= \frac{1}{n} \sum_{i=1}^n (\eta_i - \bar{\eta})^2, \\ S_{12} &= \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}), \end{aligned} \quad (1.1.5)$$

$$R = \frac{S_{12}}{S_1 S_2}, \quad (1.1.6)$$

则称 S_{12}, R 分别为二维样本的协方差与二维样本的相关系数.

显然有

$$S_{12} = \frac{1}{n} \sum_{i=1}^n \xi_i \eta_i - \bar{\xi} \bar{\eta}. \quad (1.1.7)$$

设 $\varphi_\xi(t)$ 为总体 ξ 的特征函数, ξ_1, \dots, ξ_n 为总体 ξ 的样本, 则样本均值 $\bar{\xi}$ 的特征函数为

$$\varphi_{\bar{\xi}}(t) = E(e^{jt\bar{\xi}}) = E\left(e^{jt \sum_{i=1}^n \xi_i / n}\right) = \left\{ \varphi_\xi\left(\frac{t}{n}\right) \right\}^n, \quad j^2 = -1. \quad (1.1.8)$$

利用上式与特征函数的惟一性定理, 由总体 ξ 的分布, 常可求得 $\bar{\xi}$ 的分布. 例如, 设 ξ_1, \dots, ξ_n 为总体 $\xi \sim N(a, \sigma^2)$ 的样本, 因为有

$$\varphi_\xi(t) = e^{jta - \frac{1}{2}t^2\sigma^2},$$

从而样本均值 $\bar{\xi}$ 的特征函数为

$$\varphi_{\bar{\xi}}(t) = \left\{ \varphi_\xi\left(\frac{t}{n}\right) \right\}^n = e^{jta - \frac{1}{2}t^2\left(\frac{\sigma}{\sqrt{n}}\right)^2},$$

此为正态分布特征函数, 由特征函数惟一性定理知

$$\bar{\xi} \sim N\left(a, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) = N\left(a, \frac{\sigma^2}{n}\right).$$

1.1.2 顺序统计量

定义 1.1.4 设 $\xi_1, \xi_2, \dots, \xi_n$ 为总体 ξ 的样本, 现由样本 $\xi_1, \xi_2, \dots, \xi_n$ 建立 n

个函数:

$$\hat{\xi}_{(k)} = \xi_{(k)}(\xi_1, \xi_2, \dots, \xi_n), k = 1, 2, \dots, n,$$

其中 $\xi_{(k)}$ 为这样的统计量, 其观察值为 $x_{(k)}$, 而 $x_{(k)}$ 为样本的观察值 x_1, x_2, \dots, x_n 按递增次序排列成

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(k)} \leq \dots \leq x_{(n)}$$

后的第 k 个数值. 则称 $\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(n)}$ 为样本 $\xi_1, \xi_2, \dots, \xi_n$ 的顺序统计量或次序统计量. 称 $\xi_{(k)}$ 为样本 $\xi_1, \xi_2, \dots, \xi_n$ 的第 k 个顺序统计量 ($1 \leq k \leq n$). 实际上 $\xi_{(k)}$ 是样本 $\xi_1, \xi_2, \dots, \xi_n$ 中第 k 个最小的样品, $1 \leq k \leq n$. 显然有

$$\xi_{(1)} \leq \xi_{(2)} \leq \dots \leq \xi_{(k)} \leq \dots \leq \xi_{(n)}, \quad (1.1.9)$$

$$\xi_{(1)} = \min_{1 \leq k \leq n} \xi_{(k)}, \quad \xi_{(n)} = \max_{1 \leq k \leq n} \xi_{(k)}. \quad (1.1.10)$$

记

$$\tilde{\xi} = \begin{cases} \xi_{(\frac{n+1}{2})}, & \text{当 } n \text{ 为奇数,} \\ \frac{1}{2}(\xi_{(\frac{n}{2})} + \xi_{(\frac{n}{2}+1)}), & \text{当 } n \text{ 为偶数,} \end{cases} \quad (1.1.11)$$

则称 $\tilde{\xi}$ 为样本中位数(中值). 称 $R_n \equiv \xi_{(n)} - \xi_{(1)}$ 为样本极差. 样本中位数的基本思想是把样本分为两个相等的部分(大数部分与小数值部分), 而样本中位数是分界线. 样本极差是样本中最大值与最小值之差, 它反映样本观察值波动程度, 它与样本标准差 S 一样反映观察值的离散程度.

设 $F(x)$ 为总体 ξ 的分布函数, 由文献[21]中定理 2.7.6 的推论 1 知, $\xi_{(1)}$, $\xi_{(n)}$ 的分布函数分别为

$$F_{\xi_{(1)}}(x) = 1 - [1 - F(x)]^n,$$

$$F_{\xi_{(n)}}(x) = [F(x)]^n. \quad (1.1.12)$$

如果总体 ξ 为连续型随机变量, 且有密度函数 $f(x)$, 则 $\xi_{(1)}$, $\xi_{(n)}$ 为连续型随机变量, 其密度函数分别为

$$f_{\xi_{(1)}}(x) = nf(x)[1 - F(x)]^{n-1},$$

$$f_{\xi_{(n)}}(x) = nf(x)[F(x)]^{n-1}. \quad (1.1.13)$$

由文献[21]中定理 2.7.7 知 $(\xi_{(1)}, \xi_{(n)})$ 的联合分布函数为

$$F_{1,n}(x, y) = \begin{cases} [F(y)]^n - [F(y) - F(x)]^n, & x < y, \\ [F(y)]^n, & x \geq y. \end{cases} \quad (1.1.14)$$

如果总体 ξ 为连续型的且有密度 $f(x)$, 则 $(\xi_{(1)}, \xi_{(n)})$ 为二维连续型随机向量, 其密度为

$$f_{1,n}(x, y) = \begin{cases} n(n-1)f(x)f(y)[F(y) - F(x)]^{n-2}, & x < y, \\ 0, & x \geq y. \end{cases} \quad (1.1.15)$$

从而极差 $R_n = \xi_{(n)} - \xi_{(1)}$ 有密度

$$f_{R_n}(z) = \begin{cases} n(n-1) \int_{-\infty}^{\infty} f(x)f(z+x)[F(z+x) \\ - F(x)]^{n-2} dx, & z > 0, \\ 0, & z \leq 0. \end{cases} \quad (1.1.16)$$

当总体 $\xi' \sim N(0, 1)$ 时, 记 R'_n 为相应样本极差, 且记 $C_n = E(R'_n)$, $v_n^2 = D(R'_n)$. 由(1.1.16)式可计算出 C_n 与 v_n^2 的值(见表 1.1.1). 当总体 $\xi \sim N(a, \sigma^2)$ 时, 设 R_n 为其样本极差, 令

$$\xi' = \frac{\xi - a}{\sigma}, \quad \xi'_i = \frac{\xi_i - a}{\sigma}, \quad i = 1, 2, \dots, n,$$

表 1.1.1

n	C_n	$\frac{1}{C_n}$	v_n
1	1.12838	0.8862	0.853
3	1.69257	0.5908	0.888
4	2.05875	0.4857	0.880
5	2.32593	0.4299	0.864
6	2.53441	0.3946	0.848
7	2.70436	0.3698	0.833
8	2.84720	0.3512	0.820
9	2.97003	0.3367	0.808
10	3.07751	0.3249	0.797

则 $\xi'_1, \xi'_2, \dots, \xi'_n$ 为总体 $\xi' \sim N(0, 1)$ 的样本. 记 R'_n 为样本 $\xi'_1, \xi'_2, \dots, \xi'_n$ 的极差, 则

$$R'_n = \xi'_{(n)} - \xi'_{(1)} = \max_{1 \leq i \leq n} \left(\frac{\xi_i - a}{\sigma} \right) - \min_{1 \leq i \leq n} \left(\frac{\xi_i - a}{\sigma} \right) = \frac{R_n}{\sigma},$$

故

$$C_n = E(R'_n) = \frac{1}{\sigma} E(R_n),$$

从而

$$\sigma = E\left(\frac{R_n}{C_n}\right), \quad D\left(\frac{R_n}{C_n}\right) = D\left(\frac{\sigma R'_n}{C_n}\right) = \frac{v_n^2}{C_n^2} \sigma^2.$$

所以, 我们可用 $\frac{R_n}{C_n}$ 来估计正态总体标准差 σ . 将用 $\frac{R_n}{C_n}$ 估计 σ 与用 S 来估计 σ 进行比较, 当 $n \leq 10$ 时, 其效率相当高; 然而当 $n > 10$ 时, 其效率迅速下降. 为提高效

率,当 $n > 10$ 时,可随机地把样本观察值分成每组只有少数几个(最好 5 个)样品的若干组,然后由各组分别估计 σ ,最后再取平均值.

1.1.3 经验分布函数

定义 1.1.5 设 $\xi_1, \xi_2, \dots, \xi_n$ 为总体 ξ 的样本, $\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(n)}$ 为样本 $\xi_1, \xi_2, \dots, \xi_n$ 的顺序统计量. 对任意实数 x , 记

$$F_n(x) = \begin{cases} 0, & x \leq \xi_{(1)}, \\ \frac{k}{n}, & \xi_{(k)} < x \leq \xi_{(k+1)}, k = 1, 2, \dots, n-1, \\ 1, & x > \xi_{(n)}, \end{cases} \quad (1.1.17)$$

则称 $F_n(x)$ 为总体 ξ 的经验分布函数.

记

$$\mu(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0, \end{cases} \quad (1.1.18)$$

则式(1.1.17)可改写为

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mu(x - \xi_k), \quad x \in R, \quad (1.1.19)$$

$\sum_{k=1}^n \mu(x - \xi_k)$ 表示小于 x 的那些样品 ξ_k 的个数. 因为对样本 $\xi_1, \xi_2, \dots, \xi_n$ 的任一观察值 x_1, x_2, \dots, x_n , $F_n(x)$ 是 x 的单调不减、左连续函数, 且

$$0 \leq F_n(x) \leq 1, \quad F_n(-\infty) = 0, \quad F_n(+\infty) = 1.$$

所以 $F_n(x)$ 是分布函数, 其图形如图 1-1 所示.

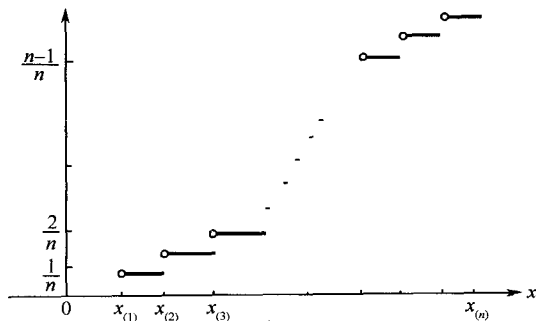


图 1-1

由式(1.1.19)知,对固定的 x , $F_n(x)$ 是样本 $\xi_1, \xi_2, \dots, \xi_n$ 的函数. 又因 $\mu(x - \xi_1), \mu(x - \xi_2), \dots, \mu(x - \xi_n)$ 独立同分布且

$$P\{\mu(x - \xi_1) = 1\} = P\{x - \xi_1 > 0\} = F(x),$$

$$P\{\mu(x - \xi_1) = 0\} = 1 - F(x),$$

其中 $F(x)$ 为总体 ξ 的分布函数. 所以

$$\mu(x - \xi_k) \sim B(1, F(x)),$$

从而

$$\sum_{k=1}^n \mu(x - \xi_k) \sim B(n, F(x)),$$

即

$$nF_n(x) \sim B(n, F(x)), \quad (1.1.20)$$

所以

$$E[F_n(x)] = \frac{1}{n}E[nF_n(x)] = \frac{1}{n}nF(x) = F(x), \quad (1.1.21)$$

$$\begin{aligned} D[F_n(x)] &= D\left[\frac{1}{n}nF_n(x)\right] = \frac{1}{n^2}D[nF_n(x)] \\ &= \frac{F(x)[1 - F(x)]}{n}, \end{aligned} \quad (1.1.22)$$

从而

$$\lim_{n \rightarrow \infty} E|F_n(x) - F(x)|^2 = \lim_{n \rightarrow \infty} \frac{F(x)[1 - F(x)]}{n} = 0. \quad (1.1.23)$$

即

$$F_n(x) \xrightarrow{2} F(x) \text{ (当 } n \rightarrow \infty \text{ 时)}, \quad (1.1.24)$$

故

$$F_n(x) \xrightarrow{P} F(x) \text{ (当 } n \rightarrow \infty \text{ 时)}, \quad (1.1.25)$$

记 $\eta_k = \mu(x - \xi_k)$, 则对任意固定的 x , $\{\eta_k\}$ 为独立同服从 $0-1$ 分布随机变量序列, 且 $E(\eta_k) = F(x)$. 由科尔莫戈罗夫 Kolmogorov 强大数定理知, $\{\eta_k\}$ 服从强大数定律, 即

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mu(x - \xi_k) = \frac{1}{n} \sum_{k=1}^n \eta_k \xrightarrow{\text{a.s.}} F(x), \quad (1.1.26)$$

格列汶科于 1933 年证明了比上式更深刻的结果:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0. \quad (1.1.27)$$

因此当 n 充分大时经验分布函数是总体分布函数较好的近似.

由棣莫弗-拉普拉斯中心极限定理知, 随机变量序列 $\{\mu(x - \xi_k)\}$ 还服从中心极限定理, 即

$$\frac{\sqrt{n}[F_n(x) - F(x)]}{\sqrt{F(x)[1 - F(x)]}} \xrightarrow{L} \zeta \sim N(0, 1) \quad (\text{当 } n \rightarrow \infty \text{ 时)},$$

所以当 n 充分大时有

$$F_n(x) \approx \frac{\sqrt{F(x)[1-F(x)]}}{\sqrt{n}} \zeta + F(x). \quad (1.1.28)$$

故渐近地有

$$F_n(x) \sim N\left(F(x), \frac{F(x)[1-F(x)]}{n}\right). \quad (1.1.29)$$

1.1.4 几个重要分布

一、 Γ 分布

如果连续型随机变量 ξ 的密度函数为

$$f(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0 \end{cases}, \alpha > 0, \lambda > 0,$$

其中 $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$, 称为 Γ 函数. 则称 ξ 服从参数为 α, λ 的 Γ 分布. 记为 $\xi \sim \Gamma(\alpha, \lambda)$.

Γ 函数 $\Gamma(\alpha)$ 有如下公式:

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha), \Gamma(1) = 1, \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \quad (1.1.30)$$

$$\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = B(p, q) \equiv \int_0^1 x^{p-1}(1-x)^{q-1} dx, \quad (1.1.31)$$

称 $B(p, q)$ 为 β (贝塔) 函数.

定理 1.1.1 设 $\xi_i \sim \Gamma(\alpha_i, \lambda), i = 1, 2, \dots, N$, 且 $\xi_1, \xi_2, \dots, \xi_N$ 相互独立, 则

$$\sum_{i=1}^N \xi_i \sim \Gamma\left(\sum_{i=1}^N \alpha_i, \lambda\right).$$

证明 当 $N=2$ 时, $\xi_1 + \xi_2$ 的密度函数为

$$\begin{aligned} f_{\xi_1 + \xi_2}(z) &= \int_{-\infty}^{\infty} f_{\xi_1}(x) f_{\xi_2}(z-x) dx \\ &= \begin{cases} \frac{\lambda^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} \int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx, & z > 0, \\ 0, & z \leq 0, \end{cases} \\ &\stackrel{\text{令 } x=zt}{=} \begin{cases} \frac{\lambda^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} z^{\alpha_1 + \alpha_2 - 1} B(\alpha_1, \alpha_2), & z > 0, \\ 0, & z \leq 0, \end{cases} \\ &= \begin{cases} \frac{\lambda^{\alpha_1 + \alpha_2} z^{\alpha_1 + \alpha_2 - 1}}{\Gamma(\alpha_1 + \alpha_2)} e^{-\lambda z}, & z > 0, \\ 0, & z \leq 0. \end{cases} \end{aligned}$$

此示当 $N=2$ 时结论成立, 现设 $N=k$ 时结论成立, 往证 $N=k+1$ 时结论也成立.

因为 $\xi_1, \xi_2, \dots, \xi_{k+1}$ 相互独立, 所以, $\sum_{i=1}^k \xi_i$ 与 ξ_{k+1} 独立, 由上证明与假设得

$$\sum_{i=1}^{k+1} \xi_i = \sum_{i=1}^k \xi_i + \xi_{k+1} \sim \Gamma\left(\sum_{i=1}^k \alpha_i + \alpha_{k+1}, \lambda\right) = \Gamma\left(\sum_{i=1}^{k+1} \alpha_i, \lambda\right) \text{ 证毕.}$$

设 $\xi \sim \Gamma(\alpha, \lambda)$, 则对任意正整数 k 有

$$\begin{aligned} E(\xi^k) &= \int_0^{\infty} x^k \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} dx \quad (\text{令 } \lambda x = t) \\ &= \int_0^{\infty} \frac{1}{\lambda^k \Gamma(\alpha)} t^{k+\alpha-1} e^{-t} dt = \frac{\Gamma(k+\alpha)}{\lambda^k \Gamma(\alpha)}, \end{aligned} \quad (1.1.32)$$

所以

$$\begin{aligned} E(\xi) &= \frac{\alpha}{\lambda}, \\ D(\xi) &= E(\xi^2) - [E(\xi)]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}. \end{aligned} \quad (1.1.33)$$

二、 χ^2 分布

定理 1.1.2 设 $\xi_i \sim N(0, 1), i=1, 2, \dots, N$, 且 $\xi_1, \xi_2, \dots, \xi_N$ 相互独立, 记 $\chi^2 = \sum_{i=1}^N \xi_i^2$, 则 χ^2 有密度函数:

$$f_{\chi^2}(x) = \begin{cases} \frac{1}{2^{\frac{N}{2}} \Gamma\left(\frac{N}{2}\right)} x^{\frac{N}{2}-1} e^{-\frac{x}{2}}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

并称 χ^2 服从自由度为 N 的 χ^2 分布, 记为 $\chi^2 \sim \chi^2(N)$.

证明 易见 $\xi_i^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right), i=1, 2, \dots, N$, 又因 $\xi_1, \xi_2, \dots, \xi_N$ 相互独立, 所以 $\xi_1^2, \xi_2^2, \dots, \xi_N^2$ 相互独立, 由定理 1.1.1 知 $\chi^2 \sim \Gamma\left(\frac{N}{2}, \frac{1}{2}\right)$, 从而定理 1.1.2 得证.

因为

$$\chi^2 \sim \chi^2(N) = \Gamma\left(\frac{N}{2}, \frac{1}{2}\right),$$

所以

$$E(\chi^2) = N, D(\chi^2) = 2N. \quad (1.1.34)$$

设 $\xi \sim \chi^2(n_1), \eta \sim \chi^2(n_2)$, 且 ξ 与 η 独立. 易证:

$$\xi + \eta \sim \chi^2(n_1 + n_2). \quad (1.1.35)$$

设 $\xi \sim N(0, 1)$, 则 $\xi^2 \sim \chi^2(1)$, 所以, 由式(1.1.34)得