非数学类专业
研究生教学用书

数理统计

Mathematical Statistics

杨 虎　刘琼荪　钟 波　编著

高等教育出版社

非数学类专业
研究生教学用书

数理统计

Mathematical Statistics

杨虎　刘琼荪　钟波　编著

高等教育出版社

内容简介

本书根据非数学类硕士研究生数理统计课程的基本要求，从数理统计的基本概念出发，较系统地介绍了数理统计的原理和方法。内容主要包括统计的基本概念、参数估计、假设检验、回归分析、方差分析和正交设计，还补充了回归诊断、均匀设计、多元分析与数据挖掘等若干内容。本书注重统计思想和方法介绍，强调统计的实际应用。全书论述深入浅出，富有启发性。为方便读者自学，附录给出了概率知识的简单总结。每章配有习题，书后附有习题答案。

读者对象为非数学类各专业研究生和数学类本科高年级学生，也可供教师、科技工作者和工程技术人员参考。

图书在版编目(CIP)数据

数理统计/杨虎，刘琼荪，钟波编著. —北京：高等教育出版社，2004.10

ISBN 7-04-015481-1

Ⅰ.数… Ⅱ.①杨…②刘…③钟… Ⅲ.数理统计-研究生-教材 Ⅳ.O212

中国版本图书馆 CIP 数据核字(2004)第 085065 号

策划编辑	李艳馥	**责任编辑**	张耀明	**封面设计**	李卫青
责任绘图	吴文信	**版式设计**	史新薇	**责任校对**	金 辉
责任印制	杨 明				

出版发行	高等教育出版社	**购书热线**	010-64054588
社 址	北京市西城区德外大街 4 号	**免费咨询**	800-810-0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
经 销	新华书店北京发行所		
排 版	高等教育出版社照排中心		
印 刷	中国农业出版社印刷厂		
开 本	787×960 1/16	**版 次**	2004 年 10 月 第 1 版
印 张	19.25	**印 次**	2004 年 10 月 第 1 次印刷
字 数	320 000	**定 价**	26.80 元

物料号：15481-00

前　言

数理统计是很多学科不可缺少的理论基础和数据分析工具，它广泛地应用于理、工、农、医、管、经、文等各个学科的理论和应用研究领域。长期以来，由于各个层次、各个环节的学生都开设了数理统计课程，造成了一定的资源和时间的重叠和浪费，也给教材编写带来一定的困难。

非数学类研究生（工科为主）究竟需要什么样的数理统计教材？数十年来国内同类教材大都限于传统的教学内容，诸如：参数估计、假设检验、方差分析、回归分析，并以数学推导为主，忽略了更多统计新方法的介绍，这种现状一方面可能是教学时数的限制，另外也不排除部分编写者对统计新方法了解的限制。固然，一本全新的教材会给教师带来新的教学上的困难，因为对有些比较新颖的统计方法的融会贯通需要对统计学研究前沿的了解，毕竟数理统计的教学，尤其是统计思想和统计方法的讲授和数学完全是两码事。没有一定的应用经验和对方法的深入了解是无法上好这门课的。

本书的编写正是希望进行这方面的尝试，作为非数学专业研究生的一门基础课程，没有必要进行过多和过于复杂的数学推导，而是尽可能地让学生体会统计思想并掌握统计方法，因此在参数估计和假设检验作了局部的改动，主要是舍弃了理论上的严格阐述和论证；应用上增加了很多新的内容，如：假设检验增加了质量管理，在试验设计部分除了介绍传统的正交设计外，介绍了均匀设计，在回归分析里增加了回归诊断，最后一章讲述了多元分析的主要方法并介绍了数据挖掘。随着计算机技术的飞速发展，数据挖掘成为很多工科和管理学科方向的研究热门，介绍其中重要的统计方法——多元分析是很有必要的。

对于如此丰富的内容，我们在编排上尽量用浅显的描述和文字说明，避免过多符号的演算，因此，阅读本书不需要太多的数学知识，学生必备的知识仅限于高等数学和线性代数，加 * 号的内容需要熟练掌握矩阵运算工具，供教学取舍和学有余力的学生自学之用。当然，为了弥补本书在数学理论上的不足，书末附有部分参考书籍，供研究生进一步学习和科研之需。因此本书更多的角色是充当统计思想的传播、统计方法的学习和应用统计研究的工具书和入门读物，以适应不同学科（非数学类）研究生的教学和研究需要。

本书需要48学时的课堂讲授。全书各章节的内容，编者均多次在重庆大学各类研究生课程中讲授过，本书的完成得到重庆大学研究生院的专项资助，特在此表示衷心的感谢！重庆大学数理学院部分研究生和教师对部分书稿的有关环节和相关素材收集给予了帮助，也一并在此致谢！最后需要说明的是本书在正式出版前以内部讲义的形式在重庆大学2003级研究生和2004级工程硕士中大范围使用过，授课教师们发现了不少正文和习题中的错误，特此致谢！

本书属于全新的尝试，效果如何尚待检验。限于编者水平，全书错谬之处一定不少，欢迎读者批评指正！

编 者

2004年8月于重庆大学数理学院

本书符号说明

样本空间：Ω

参数空间：Θ

集合、随机事件：采用大括号{}

概率：$P(\cdots)$或 p

随机变量：$X, Y, \cdots$

分布函数：$F(x)$，标准正态分布函数 $\Phi(x)$

密度函数：$f(x)$或 $f(x;a,b)$，其中 a,b 为参数

条件密度函数：$f(x|y)$，表示随机变量 $Y=y$ 时，X 的密度函数

常数：a,b,c,d,l,m,n,k

变量：x,y,z,t,u,v,w

样本：$X_1, X_2, \cdots, X_n$

矩阵：$\boldsymbol{X}$ 或 $\boldsymbol{A}$

矩阵 $\boldsymbol{\Sigma}$ 正定记为 $\boldsymbol{\Sigma}>0$

样本观测值：$x_1, x_2, \cdots, x_n$

样本均值：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

样本中位数：$\tilde{X}$

样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

样本标准差：$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

样本 k 阶原点矩：$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1,2,\cdots$

样本 k 阶中心矩：$M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, k = 2,3,\cdots$

总体均值：EX

总体方差：DX ，$\mathrm{Var}X$

协方差：$\mathrm{Cov}(X,Y)$

当 $\boldsymbol{X}, \boldsymbol{Y}$ 分别表示随机向量时，$E\boldsymbol{X}$，$D\boldsymbol{X}$，$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})$分别表示数学期望、方差向量和协方差矩阵，$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{X})$可简单表示为 $\mathrm{Cov}(\boldsymbol{X})$.

相关系数矩阵：$\boldsymbol{R}$

相关系数：r

n 阶单位矩阵：$\boldsymbol{I}_n$（常略去下标 n，可根据前后文判断阶数）

目　录

第一章　基本概念 …… (1)
§1.1　数理统计简介 …… (1)
§1.2　总体、样本与统计量 …… (3)
§1.3　顺序统计量、经验分布函数和直方图 …… (7)
§1.4　抽样分布 …… (12)
习题一 …… (19)
第二章　参数估计 …… (23)
§2.1　点估计和区间估计的概念 …… (23)
§2.2　矩估计和最大似然估计 …… (23)
§2.3　点估计的优良性准则 …… (29)
§2.4　区间估计 …… (37)
§2.5* Bayes 估计 …… (47)
习题二 …… (53)
第三章　假设检验 …… (58)
§3.1　问题的提法和基本概念 …… (58)
§3.2　参数假设检验 …… (63)
§3.3　非参数假设检验 …… (79)
§3.4* 质量控制 …… (90)
习题三 …… (98)
第四章　回归分析 …… (103)
§4.1　回归分析概述 …… (103)
§4.2　一元线性回归 …… (105)
§4.3　一元非线性回归 …… (118)
§4.4　多元线性回归 …… (125)
§4.5* 回归诊断 …… (135)
习题四 …… (150)
第五章　方差分析与试验设计 …… (153)
§5.1　方差分析的基本原理 …… (153)
§5.2　单因素方差分析 …… (154)
§5.3* 双因素方差分析 …… (162)
§5.4　正交设计 …… (168)

§5.5* 均匀设计 …… (178)
习题五 …… (187)
第六章 多元分析与数据挖掘 …… (190)
§6.1 聚类分析 …… (190)
§6.2 主成分分析 …… (199)
§6.3 因子分析 …… (207)
§6.4* 判别分析 …… (214)
§6.5* 数据挖掘 …… (226)
习题六 …… (235)
附录 A 随机变量、概率分布、数字特征 …… (238)
附录 B 协方差矩阵与多元正态分布 …… (255)
附录 C 常用数理统计表 …… (258)
习题提示与解答 …… (287)
参考文献 …… (295)

第一章　基本概念

数理统计学是一门应用性很强的学科，其方法被广泛应用于现实社会的信息、经济、工程等各个领域，学习和运用数理统计方法已成为当今技术领域里的一种时尚，面对信息时代，为了处理大量的数据以及从中得出有助于决策的量化结论，必须掌握不断更新的数理统计知识．本着提高非数学专业硕士生统计分析能力的宗旨，使他们了解随机现象中蕴涵的带有普遍性的统计规律及其深刻的统计思想、掌握丰富多彩的数理统计方法，在不失理论严密性的前提下，力求将问题的背景、思想和方法讲解清楚，使学生能体验该门课程对于实际数据分析的重要性和具体应用情况，为今后的研究和应用提供新的思路和有效解决方案．

§1.1　数理统计简介

虽然数理统计在今天的社会已经被广泛的了解，但到目前为止，用少量的文字对“数理统计学”这个学科下一个正式的定义也很困难，很难找到无懈可击的定义．任何定义都必须加上大量的解释，否则就难以理解．

当用观察和试验的方法去研究一个问题时，第一步需要通过观察或试验收集必要的数据．这些数据会受到偶然性(随机性)因素的影响，因此第二步需要对所收集的数据进行分析，以对所要研究的问题下某种形式的结论．在这两个步骤中，都将碰到许多数学问题，为了解决这些问题，发展了许多理论和方法并以此构成了数理统计学的内容主体．

数理统计是研究怎样用有效的方法去收集和使用带随机性影响的数据的学科．

1．数据必须带有随机性的影响，才能成为数理统计学的研究对象．考虑一个国家的人口普查，如人力、物力、时间允许对每个人的状况进行调查，而这种调查又是准确无误的，则我们可利用普查所得数据，通过预先确定的方法，计算出需要的指标．例如，男性人口占全体人口的百分比，在所作假定之下这是准确无误的，这里就不需要用到数理统计方法．

2．数据随机性的来源：一是所研究的对象为数很多，不可能一一加以研究，而只能采用“一定的方式”挑选一部分加以考察．一般地，社会调查一类的问题规定了调查的范围，比如要研究某一地区内以农户为单位

的经济状况,则该地区的全体农户都是调查对象.若这个数目太大,我们只能挑一部分作深入调查.这时,所得数据的随机性就来自被挑出的农户的随机性.对这种数据作分析,就必须使用数理统计方法;二是试验的随机误差,这是由存在于试验过程中的,无法控制、甚至不了解的因素所引起的误差.

3."用有效的方式收集数据"中"有效"一词的解释:一是可以建立一个在数学上可以处理并尽可能方便的模型来描述所得数据;二是数据中要包含尽可能多的、与所研究的问题有关的信息.

研究如何用有效的方式收集数据的问题构成了数理统计学的两个分支,一是抽样理论;二是试验设计.

4. 如何"有效地使用数据"? 获取数据的目的是提供与所研究的问题有关的信息.但这种信息的获取却不是一目了然的,需要用"有效"的方式去集中、提取进而加以利用,并在此基础上作出结论.这种"结论"在统计上就称为"推断".有效地使用数据,就是使用有效的方法去集中和提取试验数据中的有关信息,对所研究的问题作出尽可能精确和可靠的推断.之所以只能做到"尽可能"而非绝对地精确的原因是由于数据本身受到随机性因素的影响.这种影响可以通过统计方法去估计或缩小其干扰作用,但不可能完全消除.

数理统计方法应用极其广泛,可以说,几乎人类活动的一切领域中都能不同程度地找到它的应用,如产品的质量控制和检验、新产品的评价、气象(地震)预报、自动控制等.这主要是因为实验是科学研究的根本方法,而随机性因素对试验结果的影响是无所不在的;反过来,应用上的需要又是统计方法发展的动力.

数理统计方法是科学研究的重要工具,为了便于处理各种统计问题的计算,已经开发出了一些非常实用的统计软件和数学软件,如 SAS、SPSS、SYSTAT、S-Plus、Eviews、Mathematica、MathCAD、Matlab 等.

数理统计学是一门非常年轻的学科,它主要的发展是从 20 世纪初开始的,在早期发展中,起领导作用的是以 R. A. Fisher 和 K. Pearson 为首的英国学派.特别是 Fisher,在本学科的发展中起着独特的作用.目前许多常用的统计方法以及教科书中的内容都与他的名字有关.其他一些著名的学者如 W. S. Gosset(Student)、J. Neyman、E. S. Pearson、A. Wald以及我国的许宝騄等,都作出了根本性的贡献.他们的工作奠定了许多统计分支的基础,提出了一系列有重要应用价值的统计方法和一系列的基本概念和重要理论问题.瑞典统计学家 H. Cramer 在 1946 年发表的著作《Mathematical Methods of Statistics》标志着数理统计学科已达

到成熟的地步.

20世纪前40年是数理统计学辉煌发展的时期.但第二次世界大战后,许多在战前开始成形的统计分支得到飞速的发展,数学上的深度比以前大大加强了,也出现了若干带根本性的新发展,如Wald的统计判决理论与Bayes学派的兴起.20世纪末,由于电子计算机这一有力工具的迅速普及,统计理论和方法开始孕育全新的形象,伴随着数据挖掘技术和方法的全面推广和实施,统计学也开始面临全新的应用层面和学科本身的现代化问题.

§1.2 总体、样本与统计量

总体、个体、样本是数理统计中三个最基本的概念.我们称研究对象的全体为总体(population).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 个个体为容量为 n 的样本(sample).

例1.2.1 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1 000 h的灯泡为次品.则该批灯泡的全体就是总体,每个灯泡就是个体.实际上,数理统计中的总体是灯泡的使用寿命 X 的取值全体,称随机变量 X 为总体,它的分布称为总体分布,记为 $F(x)$,即 $F(x)=P(X\leqslant x)$,$x\in\mathbf{R}$.

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试出来.然而,寿命试验是破坏性试验,我们只能从总体中抽取一部分,比如,抽取 n 个个体进行试验,试验结果可得一组数值 $x_1,x_2,\cdots,x_n$,由于这组数值是随着每次抽样而变化的,所以,$(x_1,x_2,\cdots,x_n)$是一个 n 维随机变量$(X_1,X_2,\cdots,X_n)$的一个观察值.

我们称 $X_1,X_2,\cdots,X_n$ 为总体 X 的一组样本,称 n 为样本容量,称 $x_1,x_2,\cdots,x_n$ 为样本的一组观测值.

为了保证所得到的样本能够客观地反映总体的统计特征,设计随机抽样方案是非常重要的.实际使用的抽样方法有很多种,要使抽取的样本能对总体作出尽可能好的推断,需要对抽样方法提出一些要求,这些要求需要满足以下两点:

1) 独立性:要求样本 $X_1,X_2,\cdots,X_n$ 为相互独立的随机变量;

2) 代表性:要求每个样本 X_i $(i=1,2,\cdots,n)$ 与总体 X 具有相同分布.

称满足以上要求抽取的样本 $X_1,X_2,\cdots,X_n$ 为简单样本(simple sample).

本书今后提到的样本都是指简单样本. 由所有样本值组成的集合 $\Omega=\{(x_1,x_2,\cdots,x_n)\mid x_i\in\mathbf{R},i=1,2,\cdots,n\}$ 称为样本空间.

在无放回抽样情况下得到的样本,从理论上说就不再是简单样本,但当总体中个体的数目很大或可以认为很大时,从总体中抽取一些个体对总体成分没有太大的影响,因此,即使是无放回抽样也可近似地看成是有放回抽样,其样本仍可看成是独立同分布的.

本节最后讨论样本的分布.

设总体 X 的分布函数为 $F(x)$, $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本,则该样本的联合分布函数为

$$\begin{aligned}F(x_1,x_2,\cdots,x_n)&=P(X_1\leqslant x_1,\cdots,X_n\leqslant x_n)\\&=\prod_{i=1}^{n}P(X_i\leqslant x_i)=\prod_{i=1}^{n}F(x_i)\\&\quad(x_i\in\mathbf{R},i=1,2,\cdots,n).\end{aligned}$$

若总体 X 是连续型随机变量且具有密度函数 $f(x)$,则样本的联合密度函数为

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

当总体 X 是离散型随机变量且具有分布律 $P(X=x_i)$ $(i=1,2,\cdots)$ 时,为今后叙述上方便起见,采用记号

$$f(x)=\begin{cases}P(X=x),\text{当 } x=x_i(i=1,2,\cdots),\\0,\text{其他},\end{cases}$$

从而样本 $X_1,X_2,\cdots,X_n$ 的概率分布仍为 $\prod_{i=1}^{n}f(x_i)$.

例 1.2.2 设总体 X 服从 0-1 分布,即 $X\sim B(1,p)$, $X_1,X_2,\cdots,X_n$ 为该总体的样本,记

$$f(x)=\begin{cases}p^x(1-p)^{1-x},x=0,1\text{ 且 }0<p<1,\\0,\text{其他},\end{cases}$$

则样本 $X_1,X_2,\cdots,X_n$ 的联合概率分布为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{n\bar{x}}(1-p)^{n-n\bar{x}},$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

例 1.2.3 假设灯泡的使用寿命 X 服从指数分布,密度函数为

$$f(x)=\begin{cases}\lambda\,\mathrm{e}^{-\lambda x},x\geqslant 0,\\0,x<0,\end{cases}$$

则样本的联合分布密度为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}(\lambda e^{-\lambda x_i})=\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}=\lambda^n e^{-n\bar{x}\lambda},$$

其中 $x_i\geqslant 0, i=1,2,\cdots,n$.

样本是总体 X 进行估计或推断的依据.由于样本是 n 个随机变量或 n 维随机向量,使用起来很不方便,我们通常是将样本提供的信息集中起来,这就是针对不同的问题构造出样本的适当函数,在统计学中称这种样本的函数为统计量.

设 $X_1,X_2,\cdots,X_n$ 为总体 X 的一个样本,$G(x_1,x_2,\cdots,x_n)$为关于 n 维变量 $x_1,x_2,\cdots,x_n$ 的连续函数,且该函数中不含任何未知参数,则称$G(X_1,X_2,\cdots,X_n)$为统计量,很明显,统计量是一个随机变量.下面介绍几个常用的统计量

样本均值:$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, (1.2.1)

样本方差:$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$, (1.2.2)

样本 k 阶原点矩:$M_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\quad(k=1,2,\cdots)$,

样本 k 阶中心矩:$M_k^*=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k\quad(k=2,3,\cdots)$,

样本标准差:$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$,

显然 $\bar{X},S^2,M_k,M_k^*,S$ 是统计量,且都是随机变量,并且有如下关系:

$$M_1=\bar{X}, \tag{1.2.3}$$

$$M_2^*=\frac{n-1}{n}S^2. \tag{1.2.4}$$

另外,常用的样本方差有如下计算公式

$$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right). \tag{1.2.5}$$

样本均值 $\bar{X}$ 有如下性质

1) $\sum_{i=1}^{n}(X_i-\bar{X})=0$;

2) 若总体 X 的均值、方差存在,且 $EX=\mu, DX=\sigma^2$,

则 $$E\bar{X}=\mu, D\bar{X}=\frac{\sigma^2}{n};$$

3) 当 $n\to\infty$ 时，$\bar{X}\xrightarrow{P}\mu$.

证 1) $\sum_{i=1}^{n}(X_i-\bar{X})=\sum_{i=1}^{n}X_i-n\bar{X}=n\bar{X}-n\bar{X}=0$;

2) $E\bar{X}=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\sum_{i=1}^{n}EX=\mu$,

$D\bar{X}=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}DX_i=\frac{1}{n^2}\sum_{i=1}^{n}DX=\frac{1}{n^2}\times n\times\sigma^2=\frac{\sigma^2}{n}$;

3) 由概率论中的大数定律知(见附录(A.29)式)，当 $n\to\infty$ 时，$\bar{X}\xrightarrow{P}\mu$.

上述性质3)表明，随着样本容量 n 的逐渐增大，样本均值 $\bar{X}$ 依概率收敛于总体均值 μ. 因此，样本均值常用于估计总体均值，或用它来检验关于总体均值 μ 的各种假设.

样本方差 S^2 的性质

1) 如果 DX 存在，则 $ES^2=DX$, $EM_2^*=\frac{n-1}{n}DX$;

2) 对任意实数 a，有 $\sum_{i=1}^{n}(x_i-\bar{x})^2\leqslant\sum_{i=1}^{n}(x_i-a)^2$.

证 1) 由公式(1.2.5)知，

$$ES^2=E\left(\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\frac{n}{n-1}\bar{X}^2\right)=\frac{1}{n-1}\sum_{i=1}^{n}EX_i^2-\frac{n}{n-1}E\bar{X}^2$$

$$=\frac{n}{n-1}EX^2-\frac{n}{n-1}E\bar{X}^2=\frac{n}{n-1}(DX+(EX)^2-D\bar{X}-(E\bar{X})^2)$$

$$=\frac{n}{n-1}\left(DX+(EX)^2-\frac{DX}{n}-(EX)^2\right)=DX.$$

再由公式(1.2.4)得 $EM_2^*=\frac{n-1}{n}DX$.

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}((x_i-a)+(a-\bar{x}))^2$$

$$=\sum_{i=1}^{n}(x_i-a)^2+n(a-\bar{x})^2+2(a-\bar{x})\sum_{i=1}^{n}(x_i-a)$$

$$=\sum_{i=1}^{n}(x_i-a)^2+n(a-\bar{x})^2-2(a-\bar{x})\sum_{i=1}^{n}(a-x_i)$$

$$=\sum_{i=1}^{n}(x_i-a)^2-n(a-\bar{x})^2\leqslant\sum_{i=1}^{n}(x_i-a)^2.$$

例 1.2.4 设总体 $X\sim U[0,\theta]$ $(\theta>0)$, $X_1,X_2,\cdots,X_n$ 为 X 的样本. 求 $E\bar{X}$, $D\bar{X}$, EM_2^*.

解 $E\bar{X}=EX=\dfrac{\theta}{2}$; $D\bar{X}=\dfrac{1}{n}DX=\dfrac{1}{n}\cdot\dfrac{(\theta-0)^2}{12}=\dfrac{\theta^2}{12n}$;

$EM_2^*=\dfrac{n-1}{n}DX=\dfrac{(n-1)\theta^2}{12n}$.

§1.3 顺序统计量、经验分布函数和直方图

一、顺序统计量

定义 1.3.1 设 $X_1,X_2,\cdots,X_n$ 为总体 X 的样本, $x_1,x_2,\cdots,x_n$ 为样本观测值, 将 $x_1,x_2,\cdots,x_n$ 按从小到大的递增顺序进行排序: $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$. 当样本 $X_1,X_2,\cdots,X_n$ 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 n 个统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$, 称为样本 $X_1,X_2,\cdots,X_n$ 的顺序统计量.

显然有 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$, 特别地, 称 $X_{(1)}$ 为最小顺序统计量, $X_{(n)}$ 为最大顺序统计量, 称 $R_n^*=X_{(n)}-X_{(1)}$ 为极差, 称

$$\tilde{X}=\begin{cases}X_{\left(\frac{n+1}{2}\right)}, & \text{当 } n \text{ 为奇数时},\\ \dfrac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right), & \text{当 } n \text{ 为偶数时}\end{cases}\tag{1.3.1}$$

为样本中位数. 样本中位数反映了随机变量 X 在实轴上分布的位置特征, 而 R_n^* 反映了随机变量 X 取值的分散程度. 由于在计算上它们比 $\bar{X}$, S^2 容易, 更适于现场使用, 但理论研究较为困难, 特别是研究极差和样本中位数的分布特征有一定的难度.

设 $F(x)$ 是总体 X 的分布函数, $X_1,X_2,\cdots,X_n$ 为 X 的样本, $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为顺序统计量, $F_{(1)}(x),F_{(n)}(x)$ 分别表示随机变量 $X_{(1)},X_{(n)}$ 的分布函数. 则对任意的实数 x, 有

$$\begin{aligned}F_{(n)}(x)&=P(X_{(n)}\leqslant x)=P(X_1\leqslant x,\cdots,X_n\leqslant x)\\&=\prod_{i=1}^{n}P(X_i\leqslant x)=\prod_{i=1}^{n}P(X\leqslant x)=F^n(x),\end{aligned}\tag{1.3.2}$$

$$F_{(1)}(x)=P(X_{(1)}\leqslant x)=1-P(X_{(1)}>x)=1-P(X_1>x,\cdots,X_n>x)$$

$$= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \prod_{i=1}^{n} P(X > x) = 1 - (P(X > x))^n$$

$$= 1 - (1 - F(x))^n. \tag{1.3.3}$$

当 X 为连续型随机变量且有密度函数 $f(x)$ 时，则 $X_{(1)}, X_{(n)}$ 也是连续型随机变量，且它们的密度函数分别为

$$f_{(n)}(x) = \frac{\mathrm{d}F_{(n)}(x)}{\mathrm{d}x} = n(F(x))^{n-1} f(x), \tag{1.3.4}$$

$$f_{(1)}(x) = \frac{\mathrm{d}F_{(1)}(x)}{\mathrm{d}x} = n(1 - F(x))^{n-1} f(x). \tag{1.3.5}$$

以上公式在统计分析中经常遇到，如何应用它们呢？下面给出一个例子.

例 1.3.1 设总体 $X \sim U[0,\theta]$ $(\theta > 0)$，$X_1, X_2, \cdots, X_n$ 为 X 的样本. 分别求 $X_{(1)}, X_{(n)}$ 的密度函数 $f_{(1)}(x), f_{(n)}(x)$.

解 因为 $X \sim U[0,\theta]$，所以 X 的密度函数与分布函数分别为

$$f(x) = \begin{cases} \dfrac{1}{\theta}, x \in [0,\theta], \\ 0, x \notin [0,\theta], \end{cases} \quad F(x) = \begin{cases} 0, x \leqslant 0, \\ \dfrac{x}{\theta}, 0 < x \leqslant \theta, \\ 1, x > \theta. \end{cases}$$

因此，由公式(1.3.4)和(1.3.5)得

$$f_{(1)}(x) = n(1 - F(x))^{n-1} f(x)$$

$$= \begin{cases} n\left(1 - \dfrac{x}{\theta}\right)^{n-1} \dfrac{1}{\theta}, x \in [0,\theta] \\ 0, \qquad\qquad x \notin [0,\theta], \end{cases}$$

$$f_{(n)}(x) = n(F(x))^{n-1} f(x)$$

$$= \begin{cases} n\left(\dfrac{x}{\theta}\right)^{n-1} \dfrac{1}{\theta}, x \in [0,\theta], \\ 0, \qquad\qquad x \notin [0,\theta]. \end{cases}$$

思考：样本 $X_1, X_2, \cdots, X_n$ 是一组独立同分布的随机变量，那么顺序统计量 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 是否是一组独立同分布的随机变量？

二、经验分布函数

样本是总体的代表. 总体 X 的分布函数 $F(x)$ 称为理论分布，往往是未知的，如何由样本对总体的分布进行推断呢？一般可用经验分布函数

去描述(推断)总体的分布,用直方图去描述(推断)总体 X(连续)的密度函数.

定义 1.3.2 设 $x_1, x_2, \cdots, x_n$ 为来自于总体 X 的样本的观测值,将这些值由小到大排序:$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$,对任意实数 x,记

$$F_n(x)=\begin{cases}0, & x<x_{(1)},\\ \dfrac{k}{n}, & x_{(k)} \leqslant x<x_{(k+1)}, k=1,2,\cdots,n-1,\\ 1, & x \geqslant x_{(n)},\end{cases} \tag{1.3.6}$$

称 $F_n(x)$ 为总体 X 的经验分布函数(empirical distribution function).

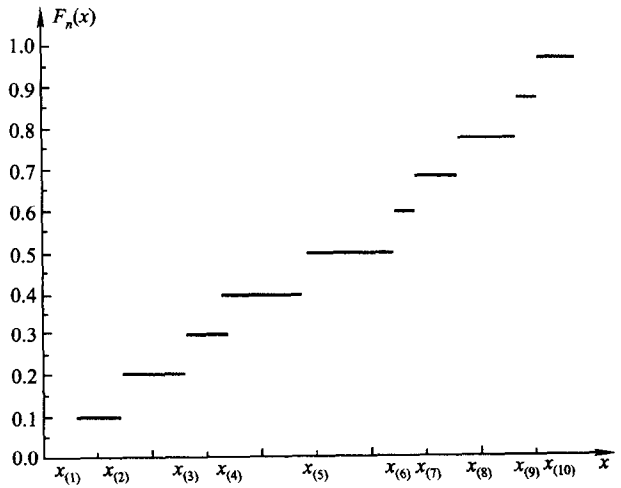

图 1.3.1 $n=10$ 的经验分布曲线 $F_n(x)$

由公式(1.3.6)及图 1.3.1 知,$F_n(x)$ 是 x 的单调不减函数,且具有如下性质:

1) $0 \leqslant F_n(x) \leqslant 1$;

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

3) $F_n(x+0)=F_n(x)$ (右连续性).

即 $F_n(x)$ 满足分布函数 $F(x)$ 的三个基本性质.值得注意的是:对于样本的不同观测值 $x_1, x_2, \cdots, x_n$,得到的经验分布函数 $F_n(x)$ 是不同的.因此,在试验之前,对应每个固定的 x 值,$F_n(x)$ 是样本 $X_1, X_2, \cdots, X_n$ 的函数.从而 $F_n(x)$ 是一个随机变量,即它是一个统计量.