高 等 学 校 教 材

9

应用数理统计

APPLICATION OF MATHEMATICAL STATISTICS

曹 莉 文海玉 主编
王 勇 田波平 主审

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

应用数理统计

APPLICATION OF MATHEMATICAL STATISTICS

曹 莉 文海玉 主编
王 勇 田波平 主审

哈爾濱工業大學出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

内 容 简 介

本书内容主要涉及数理统计的基本概念、参数估计、假设检验、回归分析、方差分析、正交试验、多元统计分析等。本书的知识体系结构与国内主流的数理统计教材基本一致，但例题的编排比较新颖，增加了一些实用而且比较先进的模拟方法。

本书可作为高等院校工科、经济类、财经、统计、管理等非数学专业的硕士研究生和博士研究生以及高年级本科生学习数理统计课程的教科书，亦可作为高等学校教师及工程技术人员的参考书。

图书在版编目(CIP)数据

应用数理统计/曹莉，文海玉主编. —哈尔滨：
哈尔滨工业大学出版社，2012.12
ISBN 978-7-5603-3890-3

Ⅰ.①应… Ⅱ.①曹… ②文… Ⅲ.①数理统计—研究生—教材 Ⅳ.①O212

中国版本图书馆 CIP 数据核字(2012)第 298286 号

策划编辑　刘培杰　张永芹
责任编辑　张永芹　齐新宇
封面设计　孙茵艾
出版发行　哈尔滨工业大学出版社
社　　址　哈尔滨市南岗区复华四道街 10 号　邮编 150006
传　　真　0451-86414749
网　　址　http://hitpress.hit.edu.cn
印　　刷　哈尔滨工业大学印刷厂
开　　本　787mm×960mm　1/16　印张 14.75　字数 300 千字
版　　次　2012 年 12 月第 1 版　2012 年 12 月第 1 次印刷
书　　号　ISBN 978-7-5603-3890-3
定　　价　48.00 元

前　言

数理统计作为应用数学中最重要、最活跃的学科之一，它在自然科学和社会科学中的应用越来越广泛深入，在国民经济和科学技术中的作用也越来越重要。作为工科研究生，理应具备数理统计的基础知识、掌握其思想方法。为了适应21世纪工科研究生教学改革和实际应用的需要，编者根据多年教学经验和我校研究生相关学科的特点，编写了这本教材。

本教材综合了编者近几年的教学笔记及在应用数理统计等书的基础上结合我校工科硕士生的教学内容编写而成。此课程是哈工大研究生院立项的学位课程，是工科院校本科生所学的工科数学《概率论与数理统计》的后续课程。在编写过程中我们注意了结合我校学生动手能力的培养，注重思想方法的介绍，注重突出数理统计学科的特点，注重它的应用性和注重与《概率论与数理统计》教材的有机衔接。

本教材共分为6章。前5章是我们讲授的主要内容，包括了数理统计的基本概念、数据分析、估计问题、假设检验、回归分析、方差分析等。这些内容涵盖了一般工科硕士研究生学习的基本要求，对于讲授的学时（36学时）来说，任务还是很重的，在实际讲授时可以根据具体情况适当删减。最后1章增加了多元统计分析的内容，在这1章中首先介绍了学习多元统计所需要的基本概念，又针对我校工科硕士学科的特点，增加了主成分分析的理论与应用。这些内容读者可以选择性地阅读。

初稿完成后，王勇、田波平教授审阅了全书，提出了许多宝贵意见，在此深深地感谢王勇教授和田波平教授。本书的出版得到了哈尔滨工业大学研究生院、哈尔滨工业大学数学系、哈尔滨工业大学出版社的大力支持，尤其得到刘培杰老师和张永芹老师给予的帮助。在此向所有协助本书出版的老师表示衷心的感谢。

在编写本书的过程中，我们参考了较多的相关文献，但是由于篇幅有限未在参考文献中一一列出，在此对文献作者表示衷心的感谢。

由于我们学识有限，虽经多次纠错和修改，书中难免有疏漏不当之处，敬请读者批评指正。

编　者

2012 年 12 月

于哈尔滨工业大学

目　录

第 1 章　基本概念及数据汇总

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. 该学科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理论，必须掌握不断更新的数理统计知识.

1.1　数理统计简介

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，以便对所要研究的问题下某种形式的结论. 在这两个步骤中，都将遇到许多数学问题，为了解决这些问题，人们发展了许多理论和方法并以此构成了数理统计学的主体内容.

数理统计是研究怎样用有效的方法去收集、分析和使用受随机性影响的数据. 数理统计学研究的对象是受随机性影响的数据. 是否假定数据有随机性，这是区别数理统计方法和其他数据处理方法的根本点. 数据的随机性来源有二：一是抽样的随机性，出于经济原因的考虑或时间的限制或问题性质决定，不可能或没有必要得到研究对象的全部资料，而只能用“一定的方式”抽取其中一部分进行考察，这样所得到的数据的随机性就是来自抽样的随机性；二是试验过程中的随机误差，即在试验过程中未加控制或无法控制或不便控制，甚至是不了解的因素所引起的误差. 在实际问题中这两类随机性常常交织在一起. 例如某工厂生产出大量的电视机显像管，为了检测显像管的寿命，推断寿命的分布类型、相关参数的具体数值以及是否达到生产要求等，必须对显像管的寿命进行测试，由于寿命试验具有破坏性，所以只能抽取少量显像管以一定的方式进行加速老化试验而得到部分数据，这里，抽样的随机性对数据便有影响. 另外，产品即使是在同一条件下生产出来的，

但各台显像管的寿命仍会有差异,这就是随机误差对数据的影响.

数理统计学研究的内容随着科学技术和生产实践的不断进步而逐步扩大,概括起来可以分为两大类:(1)用有效的方法去收集数据. 这里"有效"一词有两方面的含义:一是可以建立一个在数学上便于处理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. 对该问题的研究构成了数理统计学中的两个分支,即抽样理论和试验设计,这些不是本书的主要内容. (2)有效地使用数据. 获取数据以后,必须使用有效的方法去集中和提取数据中的相关信息,以对所研究的问题作出尽可能精确和可靠的结论,这种"结论"在统计学中叫做"推断". 有效地使用数据是比有效地收集数据更为复杂的问题,这一问题的研究构成了数理统计学的中心内容——统计推断. 上面提到的推断显像管寿命的分布类型、相关参数的具体数值以及是否达到生产要求等都是统计推断所要解决的问题.

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. 这主要是因为试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处不在的;反过来,应用上的需要又是统计方法发展的动力.

数理统计方法在社会、经济领域中有很多应用,如抽样调查,经验表明经过精心设计和组织的抽样调查其效果可以达到甚至超过全面调查的水平;另外,对社会现象的研究也有向定量化发展的趋势. 在经济学中,早在 20 世纪二三十年代,时间序列的统计分析方法就应用于市场预测,发展到今天,各种统计方法,从简单的到深奥的,都可以在数量经济学和数理经济学中找到应用.

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,人们已经开发出了一些非常实用的统计软件和数学软件. 这里简单介绍几种常见的统计软件:

(1)SPSS:这是一种非常常见的软件,在欧洲各研究机构中得到广泛应用. 它操作简单、界面十分友好、功能齐全、输出结果美观而且输出的表格和图形可以编辑修改,可以复制插入 Word 文档中,非常方便. 本书涉及的所有统计计算都可以通过 SPSS 完成,数据计算可以简单地通过点击相应的菜单和对话框来完成(菜单方式),也可以通过编程的方式完成(程序方式),还可以二者同时使用完成(混合方式). 以上特点使得 SPSS 深受专业统计和非专业统计工作者的欢迎.

(2)SAS:这是一种功能非常齐全的巨无霸统计软件,被誉为国际上的标准统计软件和最权威的组合式优秀统计软件. 美国很多大公司(主要是制药公司)都使用该软件. 该软件人机对话界面不太友好、图形操作界面不方便、一切围绕编程设计,初学者学习起来较困难(编程),该软件的说明书非常难懂,价格也很昂贵,因此

不适合基本统计课程的教学使用.

(3)S-Plus:这是 Insightful 公司的标志性产品,是 S 语言(AT&T 贝尔实验室)的后续发展,它有极为强大的统计功能和绘图能力,应用上以理论研究、统计建模为主,它需要使用者有较好的数理统计背景,对使用者编程能力的要求极高,它在北美和欧洲都有市场,价格比 SAS 便宜,但也不太适合基本统计课程的教学使用.

(4)R 软件:这是一种免费软件,是基于 S 语言的统计软件包. 它可以从网上免费下载,是发展最快的软件,与 S-Plus 很相似,但它由志愿者管理,它的运行的稳定性缺乏保证.

(5)Minitab:这是一种和 SPSS 非常相似的傻瓜式软件,它的操作界面很友好、使用方便、功能齐全,是北美大学教学中的常用软件,但在中国不如 SPSS 普遍.

(6)Eviews:这是一种计量经济学软件,由 TSP 发展而来,它主要针对时间序列分析,也可以对截面数据进行分析,该软件小巧实用,但功能不够强大.

(7)Matlab:这是一种计算软件,在工程计算方面应用很广,它以编程为主,有一些统计函数可供调用,但不如专门的统计软件使用方便.

(8)Excel:这是一种数据表格处理软件,有一些统计函数可供调用,对于简单分析,Excel 还算方便,但对于多数统计推断问题还需要其他专门的统计软件来处理.

1.2 总体、样本与统计量

1.2.1 总体与样本

总体、个体、样本是数理统计中三个最基本的概念. 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(即使试验是非破坏性的,由于试验要花费人力、物力、时间),故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,$x_1, x_2, \cdots, x_n$ 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

$$\begin{aligned} F(x_1, x_2, \cdots, x_n) &= P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_n \leqslant x_n) \\ &= \prod_{i=1}^{n} P(X_i \leqslant x_i) = \prod_{i=1}^{n} F(x_i) \quad (x_i \in \mathbf{R}; i=1,\cdots,n) \end{aligned}$$

当总体 X 是连续型随机变量且具有密度函数 $f(x)$ 时,样本的联合密度函数 $f(x_1, \cdots, x_n)$ 为 $\prod_{i=1}^{n} f(x_i)$.

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

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

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

样本分布 $F(x_1, \cdots, x_n)$,$f(x_1, \cdots, x_n)$ 或 $P(x_1, \cdots, x_n)$ 是统计推断的基础.

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

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

则样本 $X_1,X_2,\cdots,X_n$ 的联合概率分布为

$$\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

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

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

则样本的联合分布密度为

$$f(x_1,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}=\lambda^n e^{-n\bar{x}\lambda}\quad (x_i\geqslant 0;i=1,2,\cdots,n)$$

1.2.2 统计量

样本是对总体进行统计分析和推断的依据,虽然样本含有总体的信息,但比较分散,必须经过一定的加工、提炼,把分散在样本中有用的信息集中起来. 具体地说,就是针对不同问题构造样本的各种函数,再利用这些函数去推断总体的性质,在数理统计学中称这种函数为统计量.

定义 1.2.1 设 $(X_1,X_2,\cdots,X_n)$ 为取自总体 X 的一个样本,$T(x_1,x_2,\cdots,x_n)$ 为 $(x_1,x_2,\cdots,x_n)$ 的一个实值连续函数,且 T 中不包含任何未知参数,则称 $T=T(X_1,X_2,\cdots,X_n)$ 为一个统计量.

作为统计量必须不含任何未知参数,这一点是非常重要的. 因此在有些情形,统计量 T 是作为未知参数 θ 的估计量而构造的,若 T 中含有未知参数 θ,就无法作为 θ 的估计了. 注意到样本的二重性,作为样本的函数的统计量也就具有二重性,即统计量 $T(X_1,X_2,\cdots,X_n)$ 为随机变量,它应有确定的概率分布,称之为抽样分布. 而对于样本的一个观测值 $(x_1,x_2,\cdots,x_n)$,统计量 $T(X_1,X_2,\cdots,X_n)$ 也有一个相应的值 $T(x_1,x_2,\cdots,x_n)$.

下面介绍几个常用的重要统计量.

定义 1.2.2 设 $(X_1,X_2,\cdots,X_n)$ 是从总体 X 中抽取的一个样本,我们定义下列统计量:

样本均值

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1.1}$$

样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right) \tag{1.2}$$

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

$$M_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\quad(k=1,2,\cdots) \tag{1.3}$$

样本 k 阶中心矩

$$M_k^*=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k\quad(k=2,3,\cdots) \tag{1.4}$$

这些统计量统称为总体的样本矩.

显然 $M_1=\bar{X}$,$\bar{X}$ 是样本的算术平均值,$M_2^*=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$. 本书中常将 M_2^* 用 S^{*2} 表示,S^{*2} 与 S^2 略有不同,但它们都是样本平均偏差平方和. $\bar{X}$ 和 S^2 是以后用得最多的统计量,由下面的性质可以看出,$\bar{X}$ 集中反映了总体均值的信息,S^2 集中反映了总体方差的信息.

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,$\bar{X}\xrightarrow{p}\mu$.

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

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

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

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

证明:(1)由样本方差公式知

$$\begin{aligned}ES^2&=E\left(\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\frac{n}{n-1}\bar{X}^2\right)=\frac{1}{n-1}\sum_{i=1}^{n}EX_i^2-\frac{n}{n-1}E\bar{X}^2\\&=\frac{n}{n-1}EX^2-\frac{n}{n-1}E\bar{X}^2=\frac{n}{n-1}(DX+(EX)^2-D\bar{X}-(E\bar{X})^2)\\&=\frac{n}{n-1}(DX+(EX)^2-\frac{DX}{n}-(EX)^2)=DX\end{aligned}$$

再由公式(1.4)得

$$EM_2^*=\frac{n-1}{n}ES^2=\frac{n-1}{n}DX$$

(2)由已知,有

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2&=\sum_{i=1}^{n}((x_i-a)+(a-\bar{x}))^2\\&=\sum_{i=1}^{n}(x_i-a)^2+n(a-\bar{x})^2+2(a-\bar{x})\sum_{i=1}^{n}(x_i-a)\\&=\sum_{i=1}^{n}(x_i-a)^2+n(a-\bar{x})^2-2(a-\bar{x})\sum_{i=1}^{n}(a-x_i)\\&=\sum_{i=1}^{n}(x_i-a)^2-n(a-\bar{x})^2\leqslant\sum_{i=1}^{n}(x_i-a)^2\end{aligned}$$

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

$$E\bar{X}=EX=\frac{\theta}{2}$$

$$D\bar{X}=\frac{1}{n}DX=\frac{1}{n}\cdot\frac{(\theta-0)^2}{12}=\frac{\theta^2}{12n}$$

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

需要指出的是,若总体 X 的 k 阶矩存在,则样本的 k 阶矩必依概率收敛于总体的 k 阶矩. 例如,$M_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$ 为样本 k 阶原点矩,$\mu_k=E(X^k)$ 为总体 k 阶原点矩. 因为 $X_1,X_2,\cdots,X_n$ 相互独立且与 X 同分布,所以 $X_1^k,X_2^k,\cdots,X_n^k$ 相互独立且与 X^k 同分布,再注意到

$$E(M_k)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^k\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i^k)=\frac{1}{n}\sum_{i=1}^{n}E(X^k)=\mu_k$$

故由独立同分布的辛钦(Хинчин)大数定律可知,当 $n\to\infty$ 时,M_k 依概率收敛于 μ_k.

定义 1.2.3　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 $X_{(1)},X_{(2)},\cdots,X_{(n)}$,称其为样本 $X_1,X_2,\cdots,X_n$ 的顺序统计量.

特别的,称 $X_{(1)}$ 为最小顺序统计量,$X_{(n)}$ 为最大顺序统计量,称 $R=X_{(n)}-X_{(1)}$ 为极差,称

$$\tilde{X}=\begin{cases}X_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数}\\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right), & n\text{ 为偶数}\end{cases}\tag{1.5}$$

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,而极差反映了随机变量 X 取值的分散程度. 由于在计算上它们比 $\bar{X},S^2$ 容易,因此更适于现场使用,但它们的理论研究较为困难,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

$$\begin{aligned}F_{(1)}(x)&=P(X_{(1)}\leqslant x)=1-P(X_{(1)}>x)\\&=1-P(X_1>x,X_2>x,\cdots,X_n>x)\\&=1-\prod_{i=1}^{n}P(X_i>x)=1-(P(X>x))^n\\&=1-(1-F(x))^n\end{aligned}\tag{1.6}$$

$$F_{(n)}(x)=P(X_{(n)}\leqslant x)=P(X_1\leqslant x,X_2\leqslant x,\cdots,X_n\leqslant x)$$

$$= \prod_{i=1}^{n} P(X_i \leqslant x) = \prod_{i=1}^{n} P(X \leqslant x) = F^n(x) \tag{1.7}$$

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

$$f_{(1)}(x) = \frac{\mathrm{d}F_{(1)}(x)}{\mathrm{d}x} = n(1 - F(x))^{n-1} f(x) \tag{1.8}$$

$$f_{(n)}(x) = \frac{\mathrm{d}F_{(n)}(x)}{\mathrm{d}x} = n(F(x))^{n-1} f(x) \tag{1.9}$$

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

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

因此，由式(1.8)和式(1.9)得

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

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？

1.3 数据汇总

数据汇总是处理数据的描述和汇总方法，其中的数据都是以单个样本、多个样本或成批样本形式出现的. 这些方法大部分以图形的方式展示数据，可以用其揭示

数据结构,而原始数据要么列示在纸张上,要么作为计算机文档记录在磁带或磁盘中. 在不使用随机模型的情况下,这些方法完全可以达到描述性分析的目的. 如果适当考虑随机模型,那么关注点也是集中在方法模型的内涵上.

1.3.1 经验累积分布函数

假设 $x_1,x_2,\cdots,x_n$ 是一组数据(单词样本通常用作 x_i 独立同分布地来自某个分布函数的情形,单词暗含着没有假定随机模型). 经验累积分布函数(empirical cumulative distribution function,ECDF)定义如下

定义 1.3.1 设 $x_1,x_2,\cdots,x_n$ 为来自总体 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

$$F_n(x)=\begin{cases}0, x<x_{(1)}\\ \dfrac{k}{n}, x_{(k)}\leqslant x<x_{(k+1)} \quad (k=1,2,\cdots,n-1)\\ 1, x\geqslant x_{(n)}\end{cases} \tag{1.10}$$

称 $F_n(x)$ 为总体 X 的经验累积分布函数.

ECDF 是随机变量累积分布函数在数据形式下的对应类似函数:$F(x)$ 给出了 $X\leqslant x$ 的概率,$F_n(x)$ 给出了小于或等于 x 的数据比例.

例 1.3.1 作为使用 ECDF 的例子,我们考虑取自怀特(White),Riethof 和库什尼尔(Kushnir)(1960)的蜂蜡化学性质的研究数据. 这个研究的目的是通过一些化学试验,探测蜂蜡中人造蜡的存在性. 例如,添加微晶蜡可以提高蜂蜡的熔点. 如果所有的纯蜂蜡具有相同的熔点,那么确定熔点可以探测蜂蜡的稀释性. 然而,熔点和蜂蜡的其他化学性质随着蜂巢的不同而不同. 作者得到 59 个纯蜂蜡的样本,测量几个化学性质,检验测量值的变异性. 这 59 个熔点(℃)如表 1.1 所示. 作为这些测量值的汇总,图 1.1 画出了它们的 ECDF 图象.

表 1.1

63.78	63.45	63.58	63.08	63.40	64.42	63.27	63.10
63.34	63.50	63.83	63.63	63.27	63.30	63.83	63.50
63.36	63.86	63.34	63.92	63.88	63.36	63.36	63.51
63.51	63.84	64.27	63.50	63.56	63.39	63.78	63.92
63.92	63.56	63.43	64.21	64.24	64.12	63.92	63.53
63.50	63.30	63.86	63.93	63.43	64.40	63.61	63.03
63.68	63.13	63.41	63.60	63.13	63.69	63.05	62.85
63.31	63.66	63.60					