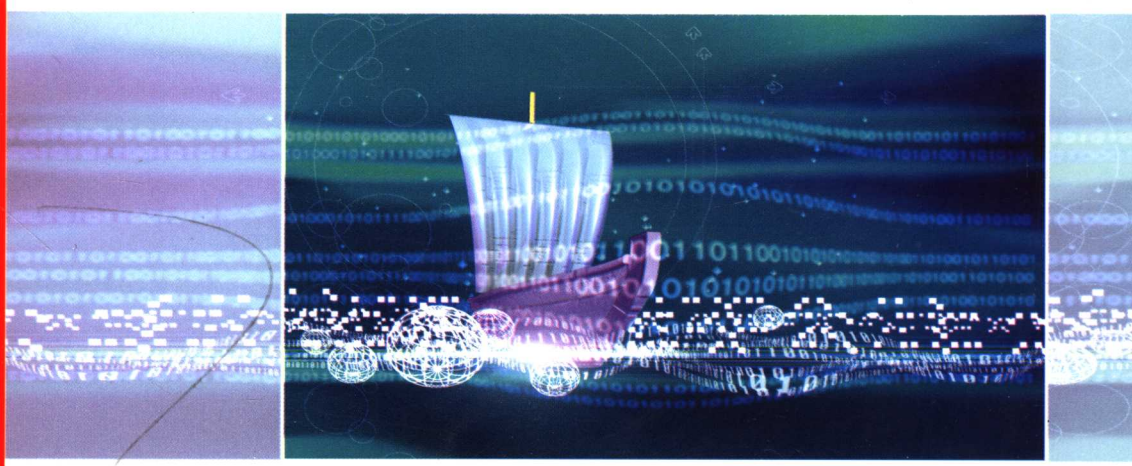
普通高等教育“十一五”国家级规划教材

Applied Mathematical Statistics

应用数理统计

史道济　张玉环　主编



天津大学出版社
TIANJIN UNIVERSITY PRESS

普通高等教育“十一五”国家级规划教材

应用数理统计

史道济　张玉环　主编

天津大学出版社
TIANJIN UNIVERSITY PRESS

内 容 提 要

本书是普通高等教育“十一五”国家级规划教材.共分7章,系统介绍数理统计的基本内容.第1章阐述数理统计的基本概念,第2~4章的参数估计、假设检验和线性模型是数理统计最基本内容,第5、6章是非参数统计和统计判决函数,第7章是选学内容,包括异常值、统计诊断及自助法、刀切法等数据处理方法.其他各章也有一些供选学的内容,如广义最小二乘估计、广义线性模型、多重比较等.

本书的主要特点是突出统计方法与统计软件包R的结合.R语言简单易学,R软件免费使用,源代码完全开放,是培养学生创新能力的工具之一,附录是对R的简单介绍.此外,构造置信区间的差异度函数也是国内同类教材中不多见的.

本书可作为数学与应用数学专业本科生的数理统计教材,但不拘泥于数学上的细节,因此也可作为非数学专业研究生的应用统计学教材使用,对广大科研技术人员也是一本合适的参考书.

图书在版编目(CIP)数据

应用数理统计/史道济,张玉环主编.—天津:天津大学出版社,2008.4

ISBN 978-7-5618-2655-3

Ⅰ.应… Ⅱ.①史…②张… Ⅲ.数理统计 Ⅳ.0212

中国版本图书馆CIP数据核字(2008)第036502号

出版发行	天津大学出版社
出 版 人	杨欢
地　　址	天津市卫津路92号天津大学内(邮编:300072)
电　　话	发行部:022-27403647　邮购部:022-27402742
网　　址	www.tjup.com
短信网址	发送“天大”至916088
印　　刷	天津泰宇印务有限公司
经　　销	全国各地新华书店
开　　本	185mm×260mm
印　　张	19
字　　数	490千
版　　次	2008年4月第1版
印　　次	2008年4月第1次
印　　数	1-3 000
定　　价	29.00元

前　言

本书主要是为工科院校数学与应用数学专业“数理统计”课程而编写，也可作为综合性大学、师范院校和工程院校数学系以及需要较多数理统计知识的其他各专业教材，更为广大自然科学工作者、社会科学工作者提供一个入门的基础性读物.

本书作者讲授数理统计已有二十几年，早在1988年在参考国内外有关教材的基础上编写了《数理统计》油印讲义，作为天津大学数学系应用数学专业的教材.2002年再次整理、修改、印刷，今年又在此基础上做了较大的修改，删除了一些较烦琐的内容，补充了统计软件包中用得最多的区间估计大样本方法，增加了使用R语言或R软件包做统计分析的内容.

为适应数字化信息社会对统计的要求，利用计算机及相应的软件进行统计分析已经成为一种趋势.R软件是目前唯一的源代码完全开放的专用统计软件包.SAS软件也许是公认的最优秀软件之一，但它是商业化的，必须支付高额的使用费。其他如SPSS、MINITAB、MATLAB、EXCEL等也是如此.且上述软件(除了MATLAB)是傻瓜式的，不利于培养学生的动手能力和创新精神.虽然R软件是免费使用，但它有专门程序员负责日常的维护和不断的升级，确保内容的先进性及多样性.

R软件作为当今世界上最好的统计分析工具之一，是一种功能强、效率高、便于进行科学计算的交互式软件包，已经深受广大用户的喜爱.我们将在正文的适当地方介绍有关R函数的使用方法，并在书末的附录中对R作简单介绍，以此培养读者自己动手进行统计分析的能力.这样的安排在国内还是一种尝试，欢迎广大读者提出宝贵意见.

有别于传统的教学，我们的目的在于增加学习兴趣.另一方面，也是基于国内大多数学校并不开设统计计算课程，为弥补这方面的不足而所作的.培养用计算机及统计软件解决实际问题的能力，自己动手编写程序的能力，在知识创新时代是必需的，对学习过某种计算机语言的理工科大学生也是完全可能的，因为R的函数表达式跟普通的数学表示几乎没有区别.

本书强调统计思想、统计观点、统计方法、统计概念，我们不想把数理统计教材写成“纯数学”教材，而着重于告诉读者如何运用数理统计知识分析、研究实际问题，特别强调在应用统计方法时必须注意的事项、应用的条件、适用的范围.希望通过本书的学习，不仅能帮助学生掌握基本的统计方法，还能用统计思想考虑、解决一些实际问题.

本书对每个统计方法的介绍，首先强调统计模型，提出统计问题，再对具体的统计问题给出具体的统计方法，有助于读者理解并掌握这些统计方法的统计思想.本书对同一问题的不同统计方法给出了优良性准则，并按这些原则进行比较.本书同时注意到数理统计基本理论的完整性，使读者能从理论上认识数理统计，并为进一步学习打下良好的基础.让读者了解某些统计历史可能是有益的，在不增加多少篇幅的前提下，本书适当地介绍一些史料.

数理统计内容极其广泛，本书着重介绍数理统计的基本原理及重要的统计方法.全书共分7章.第1章讲述了数理统计的基本概念，这些概念将贯穿于本书的始终.由于参数估计和假

设检验是统计推断的最基本形式,把它们放在第2、3章,其中,利用差异度函数构造置信区间,是国内同类教材中不多见的内容.第4章讨论的线性模型是实际中最经常遇到的统计模型之一.如果没有对随机变量所服从的分布作进一步的假定,此时的统计推断应该用非参数方法,这是第5章的内容.统计决策理论已不同程度地在数理统计的各个分支中出现,因此掌握一些基本内容是必须的,我们将它们作为第6章.第7章是数据与模型,讲述二者应该是一个怎样的关系.每章末附有一定数量的习题,都是经精心选择而编入的,适当地选做一部分,可以加深对正文内容的理解.有星号的章节内容,包括广义最小二乘估计、广义线性模型、多重比较以及异常值、统计诊断、自助法、刀切法等,并不是必学的.

考虑到国内统计名词符号在某些方面的不一致,本书采用《统计学名词术语国家标准GB3358—82》规定,并同时给出名词的英文拼写.

本书曾得到天津大学"十五"规划教材建设项目的支持,同时感谢天津大学理学院数学系领导、全体教师和学生,特别感谢胡飞、关静、梁冯珍老师在多年使用中提出的许多宝贵意见,梁冯珍、关静、韩月丽老师在百忙中阅读了书稿,并提出修改意见,研究生蔡霞、贺广婷、吴新荣编写了R程序,熊红霞、唐爱丽、蔡霞、贺广婷输入了书稿.最后感谢天津大学出版社对本书的支持,在他们的指导和帮助下,本书的出版才最终成为现实.由于水平所限,不当之处在所难免,恳请使用本书的教师、学生和广大读者批评指正.

编者

2008年1月

目　录

第 1 章　数理统计的基本知识

§1.1　引论

一、数理统计的基本任务

在工农业生产、科学实验以及社会、经济及管理等各个领域，我们经常要接触许多数据.这些数据提供了非常有用的信息，它可以帮助人们认识事物的内在规律.但是，这些信息并非一目了然，而是蕴藏在数据之中，特别是这些数据可能受到随机性影响.因此必须对数据进行整理和分析，才能有效地利用所获得的资料，尽可能可靠地、正确地提取信息.

下面举几个例子加以说明.

例 1.1　某厂生产了一批产品，共有 N 个，需要检查这批产品的不合格品率 p.逐一检查每个产品的质量显然是不合适的，这样既费时、费力，又会提高产品的成本，有时甚至是不可能的(如检查是破坏性的)，所以只能随机地从中抽取一部分产品检查，希望从这一部分产品的质量情况了解到整批产品的不合格品率.由于被检查产品抽取的偶然性，必须用数理统计方法估计这批产品的不合格品率.

例 1.2　为测定一个物理常数 μ，做了 n 次试验.由于各种随机因素的干扰产生了试验误差，得到的 n 个数据不可能完全相同，它们应该是这个物理常数 μ 及随机误差之和.如何由这 n 个试验数据估计物理常数 μ?

例 1.3　为了减少对环境的污染，或研究药物对某种疾病的治疗效果，或了解生产同一种产品的机器之间的不同性能等实际问题，希望比较在相同条件下不同处理方式间的差异.譬如用甲种药给随机挑选的 m 个患者服用，乙种药让另外 n 个患者服用，观测这两种药的疗效.但由于患者的体质、病情等等的不同，即使服用同一种药也可能产生不同的效果.因此如何排除患者的因素，分辨出这两种药的好坏，是一个应该考虑的问题.

和概率论一样，数理统计也是研究大量随机现象统计规律性的数学学科.要研究一种随机现象，首先要知道它的概率分布.在概率论的许多问题中，这种概率分布通常是已知的，在此基础上通过计算与推理去研究随机现象的性质、特点和规律.但在实际中，情况并非都是如此.我们可能完全不知道一种随机现象所服从的分布，或者由于物理上的、技术上的原因，可以知道随机现象是什么分布，但不知道分布中的参数.从理论上说，只要对随机现象进行足够多次观测或试验，规律性一定能清楚地呈现出来.但实际上所允许的观测永远只能是有限的，有时甚至是少量的.因此，就要有效地利用有限的由观测或试验所得的资料，去掉由于资料不足所引起的随机干扰，对所研究的问题作出尽可能精确和可靠的推断.这种推断必然有一定程度的不确定性.我们可以用概率表示这种不确定性.这种以一定概率表明其可靠程度的推断称为**统计推断**(Statistical Inference).

数理统计就是研究如何以有效的方法收集、整理和分析受到随机性影响的数据，以对所考察的问题作出推断或者预测，直至为采取决策和行动提供依据和建议.

二、数理统计的基本内容

数理统计的研究范围随着科学技术和生产实际的不断发展而逐步扩大，但概括起来大致可分为“收集数据”和“统计推断”两部分.

收集数据就是根据一个统计问题的目的和要求，选择合理有效的方法，科学地安排试验，通过对所考察的统计特征的观测，最经济、最有效地取得进行统计推断所必需的数据资料；并把这些常常看来是杂乱无章的原始数据进行初步整理和加工，集中地表示为一些便于应用的数学形式(图、表、公式等).

数据收集的方法有全面观测、抽样观测及安排特定的试验等.

人口普查是全面观测的一个例子，有时这是必要的.如果普查过程准确无误，利用所得数据可以把感兴趣的指标计算出来，如男女性别数等，则无须数理统计方法.但由于普查过程中可能发生遗漏、重复等人为的错误，以及费时、耗力的原因，全面观测方法并不一定是最好的.在第二次世界大战期间，战局急剧变化，为及时而有效地收集有关的情况，只能进行抽样调查，这就促进了对数理统计理论和方法的研究.现在，抽样调查已成为一种重要的调查方法.经验表明，精心设计的抽样调查的结果在精度上可以超过全面观测。这种研究抽样方法的技术叫做**抽样技术**(Sampling Techniques).

安排特定的试验来收集数据也是很常见的.特别是在农业生产中，对种子品种、肥料和耕作方法等的适当选定，都需要通过试验.由于农业试验周期长、环境因素难以控制，更需要对试验方案仔细设计，并使用有效的分析方法.在工业产品生产中，如果有几种原材料和设备可以使用，生产的各种工艺因素，如温度、压力、反应时间等，又都可以取各种不同的状态，那么，这些条件怎样搭配才能使得此种产品的生产达到优质高产呢？由于人力、物力、时间上的原因，一般不可能对所有可能情况逐一试验，即不能全面试验，而只能挑选一部分搭配进行试验.所以，应要求它们有代表性，所得数据便于分析，试验规模也恰当.这就构成数理统计学中的一个分支——**试验设计**(Design of Experiments).

统计推断是数理统计的主要内容.所谓推断就是由部分而及整体，统计推断就是由受到随机性影响的观测或试验数据推断总体的性质.在统计推断中，只考虑在已给定数据所服从的概率模型条件下，如何通过数据检验选定的模型与实际是否符合、确定模型中某些未知的成分，而不考虑怎样获得这些数据.推断的基本问题可以分为**参数估计**(Parameter Estimation)和**假设检验**(Hypothesis Testing)两大类，主要方法又分为**参数方法**(Parametric Method)和**非参数方法**(Nonparametric Method).而对每一种统计推断方法，也有使用精确分布还是极限分布或渐近性质的区别.

本教材所介绍的内容属于统计推断的各个分支，但限于篇幅、学时，像多元统计分析、时间序列分析等重要内容，我们都没有涉及.

三、数理统计的基本应用

数理统计是一个应用性很强的数学学科，数理统计方法只是一个辅助性工具，要成功应用还需依赖于一定的专业知识.

数理统计对工农业生产的发展起了重要作用.试验设计的基本思想方法，就是在田间试验中发展起来的.在工业生产中需要决定一组最优生产条件，正交设计、回归设计与分析、方差分析以及多元分析等方法是处理这类问题的有力工具.在现代工业生产中，由于生产批量大，对产品的可靠性要求高，必须考虑连续生产过程中的工序控制，制定一批产品的抽样验收方案，

对元件进行寿命试验以确定其寿命分布，进而估计采用这些元件的设备的可靠性等问题.质量控制图、抽样检验、可靠性分析等一系列统计方法就因此而产生，统计质量管理已得到许多部门的高度重视.

在医药卫生与生物学方面，统计方法已越来越重要.例如，比较治疗某种疾病的不同药物和不同治疗方法，分析某种疾病的发生是否与特定因素有关(如矽肺与工作环境的关系)，在污染大气的许多成分中确定哪些成分对哪方面有影响等等问题，都离不开数理统计.在数量遗传学的研究上，数理统计也有着影响，对统计学作出了重大贡献的 R.A.Fisher 就是著名的遗传学家.

数理统计在自然科学中发挥着很大的作用.一条理论上的规律或者由初步观测而提出的某种学说，究竟是否正确或在多大程度上正确，最终还得由实验来验证.例如物理学中的 Boyle 定律(对一定质量的气体，pV = 常数)最初只是一个由观测得出的经验规律，经过统计分析证明这个规律在一定限度内正确，进而提出了改进形式.又如 Mendel 遗传定律的确定，曾用数理统计学中的“拟合优度检验法”仔细检验过.在基础和应用研究中，对大量数据的处理，数理统计更提供了必需的方法，如最常用的误差分析及建立经验公式的方法.在地质勘探、地震、水文和气象预报这些应用性很强且有基础理论意义的领域中，数理统计都发挥了重要的作用，目前已出版了一些阐述数理统计在这些领域中应用的专著.

数理统计方法对社会、经济及管理领域也有重要意义.统计方法在社会领域中应用的一个重要方法就是抽样调查.社会现象研究的定量化趋势，势必会利用包括统计方法在内的一些科学分析方法.在经济学中，早在 20 世纪 20～30 年代，时间序列分析方法就曾用于市场预测.现在一系列的统计方法，从回归分析到随机过程统计，都在数量经济中有了应用.质量管理从统计质量管理，到全面质量管理，再到现在的六西格玛管理，已经从开始时的质量改进工具到成为打造组织企业核心竞争力的有效经营战略，始终离不开统计方法的应用.

近半个世纪以来，数理统计在理论、方法和应用上都有较大的发展，内容异常丰富，应用范围越来越广.计算机的广泛应用对数理统计的发展产生了重要影响，没有现代计算机，就没有现代的统计应用，许多重要统计方法的应用都涉及大量的计算.通过计算机模拟，可以使某些复杂的精确分布得到有实用意义的解.而且，计算机在短时间内处理海量数据的能力，使人们有可能对数据进行更透彻的分析，从中提取更多信息.计算机的广泛应用为数理统计提供了巨大机会，也提出了一些新的研究课题.统计计算就是一门包括数理统计、计算数学以及计算机科学的交叉学科，近年来已得到迅速发展，各种统计软件包的编制及广泛使用，大大推动了应用统计方法的普及.另外，随着信息技术的发展产生了许多超大型数据库，因此而提出的**数据挖掘**(Data Mining)技术得到快速发展，这是一个介于统计学、模式识别、人工智能、机器学习、数据库技术以及高性能并行计算等领域的新学科.

§1.2 数理统计的基本概念

一、总体和样本

数理统计的一个基本问题就是依据观测或试验所取得的有限信息，如何对整体进行统计推断的问题.例如我们要研究某批灯泡的平均寿命，由于测试灯泡的寿命具有破坏性，所以只能从这批产品中抽取一部分进行寿命试验，希望由这部分灯泡的寿命数据对整批灯泡的平均

寿命作出统计推断.

在数理统计中,把一个统计问题所研究的全部元素组成的集合称为**总体**(Population),总体中的每个元素称为**个体**(Individuality).例如一批灯泡的全体就组成一个总体,而其中的每一只灯泡就是个体.但是我们仅关心个体的某个或某几个数量指标,以及数量指标(一维的或多维的)在总体中的分布情况.在上面例子中,如果我们以"使用寿命 X"这个指标来衡量一批灯泡的质量,那么我们只关心灯泡的使用寿命,而不考虑灯泡的其他性能.由于每个灯泡的寿命是不同的,抽取了若干个个体而得到了寿命 X 的不同数值,因此这个 X 是一个随机变量,而 X 的分布 $F(x)$就完全描述了数量指标的分布状况.由于我们关心的正是这个数量指标,因此以后就把总体和数量指标 X 可能取值的全体等同起来(有时称前者为有形总体),并且称这一总体为具有分布函数 $F(x)$的总体.这样,就把总体与随机变量 X 联系起来了,可以用 X 或它的分布函数 $F(x)$来表示总体.在具体问题中,$F(x)$常常未知或部分未知,这正是统计推断的对象.以后,我们常用"总体 X 服从正态分布"这样的术语,表示总体的某个数量指标 X 服从正态分布,或简言之为"正态总体"等等.

为了研究总体的分布规律,我们不可能对整个总体进行观测,而只能对总体中随机抽出的一些个体进行观测,譬如抽取了 n 个个体,且得到这些个体的指标值 $X_1, X_2, \cdots, X_n$,我们称 $(X_1, X_2, \cdots, X_n)$为一个**样本**(Sample),n 是样本中所包含的个体数,称为**样本大小**或**样本容量**或**样本量**(Sample Size).显然,从总体中随机抽出一个个体,并且测量相应的指标值 X_i,就是一个随机试验,而指标值 X_i是一维随机变量.因而大小为 n 的样本$(X_1, X_2, \cdots, X_n)$可以看成一个 n 维随机向量.对某次抽样观测得到$(X_1, X_2, \cdots, X_n)$的一组确定值$(x_1, x_2, \cdots, x_n)$称作**样本观测值**,样本$(X_1, X_2, \cdots, X_n)$可能取值的全体称为**样本空间**(Sample Space),记为 $\mathscr{X}$.它可以是整个 n 维空间,也可以是其中的一个子集,样本的一次观测值$(x_1, x_2, \cdots, x_n)$就是样本空间 $\mathscr{X}$ 中的一个点,即$(x_1, x_2, \cdots, x_n) \in \mathscr{X}$.

实际上,从总体中抽取样本可以有各种不同的方法.但为了使样本尽可能地反映总体的特性,不仅要使所得数据含有最大的信息,从而减少样本大小,而且要使数据分析具有一些较好的性质,因此需要对抽样方法提出一定的要求.这里提出两点:

(1)代表性,对每个个体的观测应在完全相同条件下进行,即 $X_1, X_2, \cdots, X_n$中的每一个 X_i 都应该与总体 X 有相同的分布;

(2)独立性,每个个体的观测应是独立进行的,即 $X_1, X_2, \cdots, X_n$是相互独立的随机变量.

我们把满足以上两条性质的样本称为**简单随机样本**(Simple Random Sample),简称为样本.把抽得简单随机样本的抽样方法称为**随机抽样**(Random Sampling).以后,如不作特别说明都是指简单随机样本.对于简单随机样本,我们可以应用概率论中关于独立随机变量情况的许多重要定理,这就为数理统计提供了必要的理论基础.

设总体 X 具有分布函数 $F(x)$,$(X_1, X_2, \cdots, X_n)$为取自这一总体的大小为 n 的样本,则 $(X_1, X_2, \cdots, X_n)$的联合分布函数为 $\prod_{i=1}^{n} F(x_i)$.

例 1.4 检查某批产品的不合格品率 p.产品质量用数量指标 X 来反映,$X=1$ 表示某件产品是不合格品;$X=0$ 表示某件产品是合格品.如此研究这批产品的质量就归结为讨论随机变量 X 的分布及其主要数字特征,此时总体的分布为两点分布.设从这批产品中任取 n 件,每

取一件检查后立即放回，混合后再取下一件.这样从 n 件产品中观测得到 X 的值($x_1, x_2, \cdots, x_n$)，是 n 维随机向量($X_1, X_2, \cdots, X_n$)的一组观测值.每个 X_i 都与 X 有相同的分布.样本空间由一切可能的 n 维向量($x_1, x_2, \cdots, x_n$)组成，其中每个 x_i 只取 1 或 0，因此样本空间 $\mathscr{X}$ 是 n 维空间中的 2^n 个点的集合.($X_1, X_2, \cdots, X_n$)就是大小为 n 的简单随机样本.

实际的抽样检查常常不是如上所述的有放回抽样，而是无放回抽样.在无放回抽样时，前一次抽到的产品是否为合格品就要影响后一次抽到不合格品的概率.设一批产品的总数是 N，在第一次抽到合格品条件下，第二次抽到不合格品的概率为 $\Pr\{X_2 = 1 \mid X_1 = 0\} = \dfrac{N_p}{N-1}$，在第一次抽到不合格品条件下，第二次抽到还是不合格品的概率为 $\Pr\{X_2 = 1 \mid X_1 = 1\} = \dfrac{N_p - 1}{N-1}$.显然，在此时所得的样本就不是简单随机样本了.但当 N 很大，而 n 不大(如 $n/N \leqslant 0.1$)时，可以将所得的样本近似地看成一个简单随机样本.

二、直方图

在概率论研究中，认为随机变量的分布是给定的，这是研究一个概率问题的出发点.但在数理统计研究中，总体的分布是未知的，根据观测结果估计总体的分布密度函数或分布函数，正是数理统计要解决的一个重要问题.下面简单介绍近似的分布密度函数曲线——频率直方图.

首先找出样本观测值的最小值和最大值，并把包含它们的区间$[a, b]$分成 m 等分，记 $h = (b-a)/m$，称为**组距**(Class Interval)，各分点为 $a = c_0 < c_1 < \cdots < c_m = b$.分组的多少应与样本大小 n 相适应，分组过少会使结果太粗而丧失了一些有用的信息，分组过多会突出随机性的影响而降低稳定性.分组多少还与总体的分布性质有关.一般以 7 ~ 18 组为宜.有人建议一个经验法则，以 $m = 1 + 3.32 \lg n$ 作为组数.数出样本观测值落在各区间$(c_{i-1}, c_i]$中的个数 n_i，称为第 i 组的**组频数**(Class Absolute Frequency)，$f_i = n_i/n$ 称为第 i 组的**组频率**(Class Relative Frequency)，有时也称前者为绝对频数，后者为相对频率.经分组后，同一组的数据都看成是相同的，它们都等于**组中值**(Mid-point of Class)$(c_{i-1} + c_i)/2$，如此即得分组整理表.如果进一步在 x 轴上标出点 c_i，$i = 0, 1, \cdots, m$，以各区间$(c_{i-1}, c_i]$为底，组频率与组距之比 $y_i = f_i/h = n_i/(nh)$为高作矩形，这种图称为**频率直方图**(Frequency Histogram)，它是总体密度曲线的一种近似.

例 1.5 表 1.1 中的 125 个数据表示某高炉所炼生铁中锰的含量，每天测得 5 个数据.为了作频率直方图，首先找出其中的最小值 1.06%，最大值 1.80%.为了方便，取 $a = 0.99\%$，$b = 1.89\%$，使全部数据落在区间(a, b)内.然后决定组数，譬如分为 9 组，得组距 $h = (1.89 - 0.99)/9 = 0.10(\%)$，如此分点自然取为 0.99，1.09，…，1.79，1.89.最后数出这 125 个观测值落在各组的频数 n_i，所得结果列成表 1.2，并描出频率直方图 1.1.

表 1.1　生铁中锰的含量(%)

1.40	1.28	1.36	1.38	1.44	1.40	1.34	1.54	1.44	1.46
1.80*	1.44	1.46	1.50	1.38	1.54	1.50	1.48	1.52	1.58
1.52	1.46	1.42	1.58	1.70	1.62	1.58	1.62	1.76	1.68
1.68	1.66	1.62	1.72	1.60	1.62	1.46	1.38	1.42	1.38
1.60	1.44	1.46	1.38	1.34	1.38	1.34	1.36	1.58	1.38
1.34	1.28	1.08	1.08	1.36	1.50	1.46	1.28	1.18	1.28
1.26	1.50	1.52	1.38	1.50	1.52	1.50	1.46	1.34	1.40
1.50	1.42	1.38	1.36	1.38	1.42	1.34	1.48	1.36	1.36
1.32	1.40	1.40	1.26	1.26	1.16	1.34	1.40	1.16	1.54
1.24	1.22	1.20	1.30	1.36	1.30	1.48	1.28	1.18	1.28
1.30	1.52	1.76	1.16	1.28	1.48	1.46	1.48	1.42	1.36
1.32	1.22	1.72	1.18	1.36	1.44	1.28	1.10	1.06*	1.10
1.16	1.22	1.24	1.22	1.34					

表 1.2　生铁含锰量频数表

各组分点(%)	组中值	频数	各组分点(%)	组中值	频数
0.99 ~ 1.09	1.04	3	1.39 ~ 1.49	1.44	29
1.09 ~ 1.19	1.14	9	1.49 ~ 1.59	1.54	19
1.19 ~ 1.29	1.24	18	1.59 ~ 1.69	1.64	9
1.29 ~ 1.39	1.34	32	1.69 ~ 1.79	1.74	5
			1.79 ~ 1.89	1.84	1

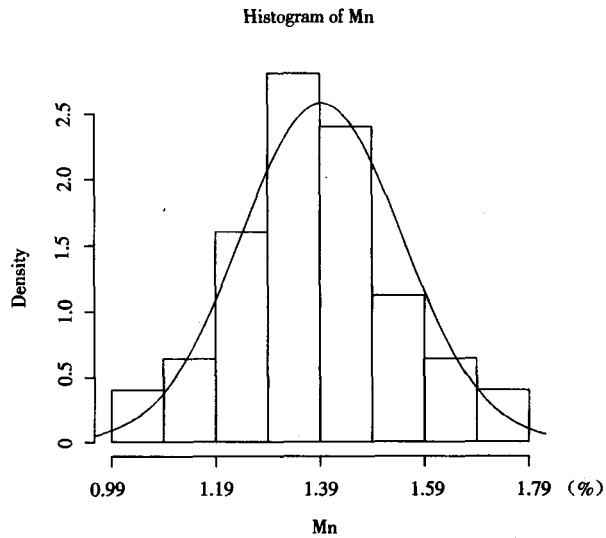


图 1.1　生铁含锰量的频率直方图

由频率直方图可以看出数据分布的三个重要特征.

(1)数据的平均值.平均值常有三种不同的意义.第一是数据最集中的取值,也就是最大频数所对应的组中值,这个值称为**众数**(Mode).本例中最大频数为 32,其所对应的组中值为 1.34%,即众数 1.34%可作为数据的一种平均值.第二是**算术平均**(Arithmetic Mean),也是最常

用的.经简单计算(用 R 函数 mean(Mn)),可得到算术平均为 1.404%.第三种平均值是**中位数**(Median),即将数据按大小次序排列后,居于中间的那个数值,这里为 1.40%.关于算术平均和中位数的详细讨论将在下一小节及 § 1.3 进行.

(2)数据的变异性.最大观测值与最小观测值之差是反映数据变异性的一个数量指标,由表 1.1 可知最大值为 1.80%,最小值为 1.06%,它们的差称为**极差**(Range),本例中为 1.80 - 1.06 = 0.74(%),但也可由最大组中值与最小组中值之差来表示,这里 1.84 - 1.04 = 0.80(%).

(3)曲线的形式.将图 1.1 中各个长方形上边的中点用一条光滑的曲线连接起来,就得到近似的分布密度曲线.样本大小 n 越大,分组越细,频率直方图就越接近分布密度曲线.本例中分布密度曲线的形状是单峰、对称的.但并不总是这样,有些分布的密度曲线呈多峰,有些则不对称.

R 函数 hist 用于画频率直方图,详细说明见附录 5.1 节或帮助文件.命令

```
> a<-mean(Mn)
> b<-sd(Mn)
> hist(Mn,axes=F,prob=T)
> axis(1,labels=c(0.99,1.19,1.39,1.59,1.79))
> axis(2,labels=c(0,0.5,1.0,1.5,2.0,2.5))
> curve(dnorm(x,a,b),add=T)
```

给出如图 1.1 的生铁含锰量(Mn)频率直方图,如果不希望有英文标题,只需添上参数项 main = NULL,而 main = "生铁含锰量(Mn)频率直方图"给出中文标题.

频率直方图的缺点在于分组区间及组数 m 的多少因人而异,且只对连续型随机变量才适用.另外,在样本大小 $n < 60$ 时,直方图的意义就不大了.

三、统计量

抽样是为了通过取得的样本对总体中某些未知因素作出推断.既然样本来自总体,自然包含了总体分布的信息.但样本常常是一堆杂乱无章的数据,不经过一定的整理难于提取出有用的信息.整理数据的方法主要有两种:一是用图、表等把它们表成直观、醒目的形式,如上段所提的直方图,还有茎叶图、箱线图等;二是针对不同问题,构造样本的某种函数,它应该汇集样本中与总体有关的主要信息,而舍弃无关的次要部分,且不包含任何未知参数,这个函数称为统计量.寻找一个合适的统计量是数理统计的中心问题之一.

定义 1.1 设$(X_1, X_2, \cdots, X_n)$是来自总体 X 的一个样本,$T = T(x_1, x_2, \cdots, x_n)$是样本空间 $\mathscr{X}$ 上的实值函数,若 $T(X_1, X_2, \cdots, X_n)$也是随机变量,且不依赖于任何未知参数,则称 $T(X_1, X_2, \cdots, X_n)$为**统计量**(Statistics).

以后,我们涉及的样本函数一般都是连续的,因此都是随机变量.尽管一个统计量不依赖于任何未知参数,但是它的分布可能依赖于总体 X 分布中的未知参数.如果$(x_1, x_2, \cdots, x_n)$是样本$(X_1, X_2, \cdots, X_n)$的一个观测值,则 $T(x_1, x_2, \cdots, x_n)$是统计量 $T(X_1, X_2, \cdots, X_n)$的一个观测值.

按照定义，随机变量 $T=\sum_{i=1}^{n}X_i$ 是一个统计量，$\chi^2=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2$ 是样本 $(X_1,X_2,\cdots,X_n)$ 的连续函数，但只有当 σ^2 已知时，它才是统计量.

由样本构造统计量，实际上是对样本包含的总体信息按某种要求进行加工，把分散在样本中的信息集中到统计量的取值上，不同的统计推断要求构造不同的统计量.统计量在数理统计中，就如随机变量在概率论中那样，占有非常重要的地位.

以下我们讨论一些常用的统计量.

定义 1.2 设 $(X_1,X_2,\cdots,X_n)$ 是取自总体 X 的大小为 n 的样本，记

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

它们都是统计量，分别称 $\bar{X}$ 和 S^2 为**样本均值**(Sample Mean)和**样本方差**(Sample Variance).一般分别称统计量

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\quad B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$$

为**样本的 k 阶(原点)矩**(k-th Moment of Sample)和**样本的 k 阶中心距**(k-th Central Moment of Sample).二阶中心矩 B_2 有时记为 $\tilde{S}^2$，即

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

特别地，$A_1=\bar{X}$，$B_2=\tilde{S}^2=\frac{n-1}{n}S^2$.容易得到

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

分别称

$$b_s=\frac{B_3}{B_2^{3/2}},\quad b_k=\frac{B_4}{B_2^2}-3$$

为样本的**偏度**(Skewness)和**峰度**(Kurtosis).称 $V=S/\bar{X}$ 为样本的**变异系数**(Coefficient of Variation).

对例 1.5 中给出的生铁中锰含量数据，利用 R 函数 mean，var，median 分别可以计算出样本均值，样本方差，样本中位数.函数 max，min 分别给出样本最大值和最小值，而函数 summary 给出更多，其中 1st Qu.和 3rd Qu.分别表示 1/4 与 3/4 分位数.

```
> mean(Mn)
[1] 1.40 368
> var(Mn)
[1] 0.02 397 022
> summary(Mn)
   Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
  1.060    1.300    1.400   1.404    1.500   1.800
```

样本均值 $\bar{X}$ 和样本方差 S^2 在数理统计学中有着重要的作用.下面给出它们的一些基本性