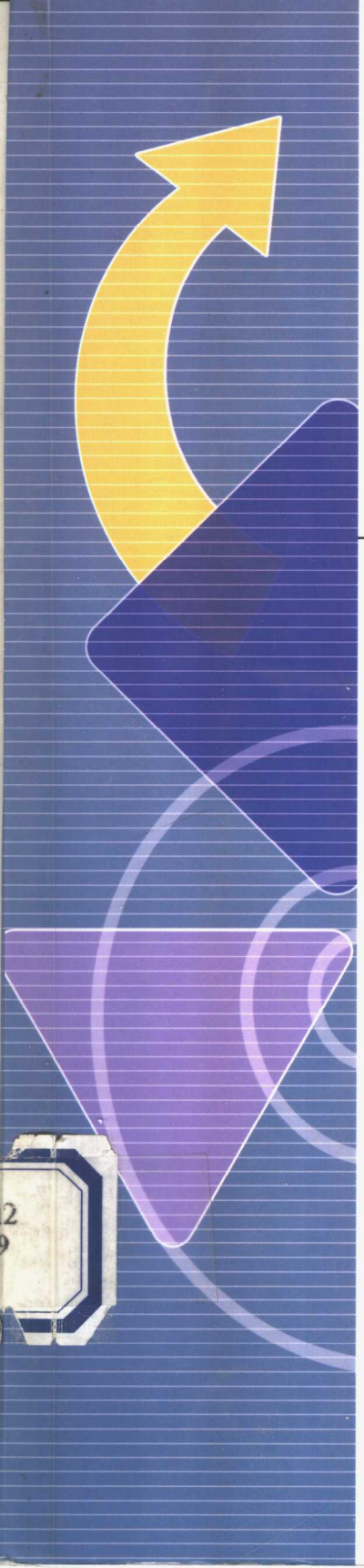

数理统计与多元统计

何 平 编著

西南交通大学出版社
Http://press.swjtu.edu.cn

数理统计与多元统计

何　平　编著

西南交通大学出版社
·成　都·

内容简介

本书以数理统计、多元统计的经典内容为主体，适当涉及其它常用的统计方法. 内容包括数理统计中的参数估计理论、假设检验理论；非参数统计方法初步；多元统计方法中的回归分析、方差分析、判别分析、主成分分析、聚类分析、相关分析，以及有广泛应用的正交试验设计方法.

本书的选材与表述是针对非数学专业硕士研究生的公共课教学，也可作为有关专业本科高年级选修课的教材.

图书在版编目(CIP)数据

数理统计与多元统计/ 何平编著. —成都：西南交通大学出版社，2004.9
ISBN 7-81057-845-6

Ⅰ. 数... Ⅱ. 何... Ⅲ. ①数理统计②多元分析；统计分析 Ⅳ. 0212

中国版本图书馆CIP数据核字(2004)第039849号

数理统计与多元统计

何 平 编著

责任编辑 张宝华
封面设计 何东琳设计工作室
西南交通大学出版社出版发行
新华书店 经销
(成都二环路北一段111号 邮政编码 610031 发行部电话：87600564)
http：//press.swjtu.edu.cn
E-mail：cbsxx@swjtu.edu.cn
四川森林印务有限责任公司印刷
*
开本：787mm×960mm 1/16 印张：15.25
字数：283千字
2004年9月第1版 2004年9月第1次印刷
ISBN 7-81057-845-6/O·062
定价：23.00元

图书如有印装质量问题，本社负责退换
版权所有，盗版必究，举报电话：(028)87600562

前　言

本书是在为我校非数学专业硕士研究生、数学类各专业本科生多年开设“数理统计与多元统计”课程的讲义基础之上，逐渐修改充实而成．包括数理统计的经典内容：参数估计理论、假设检验理论；非参数统计方法初步；多元统计方法中的回归分析、方差分析、判别分析、主成分分析、聚类分析、相关分析，以及有广泛应用的正交试验设计方法．

本书的读者对象是非数学专业硕士研究生，因此在内容选择上以广泛适用的统计推断方法为原则，不局限于学科分支的界限，也不追求方法体系的完整．在内容的叙述上侧重以下四个方面：方法的实际背景、实际问题的数学抽象和数学建模、问题解决中所体现的统计思想和数学技巧、方法运用中的基本步骤．

本书的宗旨是培养学生在将来的工作实践中能用概率统计的眼光去发现问题，对所遇到的实际问题能从概率统计的角度去思考问题，能将实际问题合理地提炼成数学问题，能创造性地运用适当的统计方法去解决实际问题．

本书的写作风格力求主题突出、线条清晰、阐述透彻、内容精练．

在本书的成稿过程中，历届选修本课程的研究生、数学专业本科生对本书的体系结构、内容取舍、写作风格产生了重要影响，他们的建议与批评使本书有了现在的面貌，避免了许多差错．涂汉生教授、赵联文教授对本书的写作给予了热情鼓励和大力帮助．概率统计教研室各位老师对本书的写作给予了全方位的支持．程世娟老师参与了课程的教学活动、协助整理了部分底稿并指出了许多谬误，郝光博士演算了大部分习题．本书的写作过程还得到西南交通大学教材建设基金的资助．本书的完成与出版，一直得到西南交通大学数学系、教务处、研究生院和西南交通大学出版社的大力支持．在此出版之际一并表示衷心的感谢．

在本书的构思与编写过程中，参阅了许多专著与教材，并采用了其中的部分内容、例题与习题，在此表示感谢．也要感谢本书的责任编辑张宝华同志热情细致的工作．

由于编者水平所限，错误、疏漏和不当之处一定还有，敬请读者不吝赐教．

编　者

2004 年 6 月于成都

目　录

第一章　数理统计的基本概念…… 1
第一节　总体与样本…… 1
第二节　样本分布和统计量…… 5
第三节　总体分布的理论分析 …… 10
第四节　统计量的分布 …… 18
习题一 …… 27

第二章　参数估计 …… 29
第一节　点估计 …… 29
第二节　估计量的评价准则 …… 52
第三节　区间估计 …… 57
第四节　样本容量的决定 …… 69
习题二 …… 73

第三章　假设检验 …… 79
第一节　参数的假设检验 …… 79
第二节　构造检验统计量的似然比方法 …… 96
第三节　分布的假设检验 …… 98
习题三…… 105

第四章　非参数统计推断方法…… 110
第一节　非参数统计模型…… 110
第二节　次序统计量和秩统计量…… 110
第三节　估计问题…… 113
第四节　检验问题…… 118
习题四…… 122

试读结束：需要全本请在线购买：　www.ertongbook.com

第五章 回归分析…… 125
第一节 相关关系与回归模型…… 125
第二节 线性回归函数的推断…… 127
第三节 误差方差σ^2的估计 …… 131
第四节 回归方程的显著性检验…… 132
第五节 非线性回归…… 138
习题五…… 140

第六章 方差分析…… 142
第一节 单因素试验的方差分析…… 142
第二节 双因素试验的方差分析…… 149
习题六…… 157

第七章 判别分析…… 159
第一节 距离判别法…… 159
第二节 贝叶斯判别法…… 164
习题七…… 170

第八章 主成分分析…… 172
第一节 客观背景和数学思想…… 172
第二节 主成分的提取…… 173
第三节 主成分的解释与命名…… 177
习题八…… 179

第九章 聚类分析…… 181
习题九…… 187

第十章 相关分析…… 189
第一节 几种相关性度量…… 189
第二节 典型相关分析…… 193
习题十…… 200

第十一章 正交试验设计…… 202
习题十一…… 218

附 录…… 220
参考文献…… 238

第一章　数理统计的基本概念

数理统计学是一门研究随机现象规律性的数学分支. 它以概率论为基础，研究如何有效地收集、整理和分析受到随机影响的数据，并对所关注的问题做出合理的估计或推断，直至为采取决策和行动提供理论依据和建议.

随机现象是多种多样、千姿百态的. 作为研究这类现象内在规律性的一般理论，首先要做的一件事就是透过现象表面的千差万别，抓住其共同具有的本质特征，再通过对这种共同本质特征的研究从而掌握其内在的规律性. 具体地说，就是首先进行科学的抽象，归纳出一些基本的概念，引入一些特定的术语，然后进一步考查它们之间的内在联系.

第一节　总体与样本

一、总体　个体　样本

一般将研究对象的全体所组成的集合称为**总体**，组成总体的每一个成员称为**个体**. 例如，要考查某个城市中居民的住房情况，进一步为该城市将来的规划和建设提供依据，那么，该城市的全体居民就组成了一个总体，每位居民是一个个体.

居民的住房情况可用多种指标来度量. 假如要关心的是其居住面积的大小，由于它是一个数量指标，所以每一个体就对应一个确定的数. 对于这种情况，为了方便，也把这个数量指标的全体所组成的集合称为总体.

由于目的是要研究该城市中居民的居住面积情况，所以全面地普查每位居民的居住面积固然是一个方法，但这往往很不经济，通常采用抽样调查的方法，即在该城市中挑选一定数量的居民(譬如说 1 000 人) 来测量其居住面积，并希望从这 1 000 个数据出发对该城市居民的住房情况做出估计. 譬如人均居住面积是多少?人均居住面积不足 4 m^2 的特困户和人均居住面积超过 20 m^2 的小康户各自所占的比例，等等.

从总体中挑选一部分个体的过程叫做**抽样**. 抽出的这一部分个体叫做**样本**. 样本所含的个体数称为**样本容量**.

数理统计方法的一个显著特点就是由样本去推断总体. 由于样本只包含了总体的一部分个体,依据样本对总体的情况所做出的推断很难达到百分之百的准确,所以只能要求这种推断尽可能地精确可靠,这就需要抽出的这部分样本能够尽可能地反映总体的情况. 因此,面临的首要问题是如何抽样,才能使抽出的样本能很好地代表总体.

二、抽样方法

在统计学中,根据人类在抽样调查方面长期积累的经验,并配合概率论理论分析的需要,对抽样方法提出了一条基本要求:抽样要保证对每一个体"机会均等",即总体中每一个体有同样机会被抽到,谁也不占优先. 凡是满足这个要求的抽样叫做**随机抽样**.

为了实现这种抽样,可以设计一种具体做法:设总体中共有 N 个个体,需要从中抽出 n 个个体作为样本. 把全部 N 个个体分别编号为 $1,2,\cdots,N$,再准备 N 个大小、质地一样的球(或纸条也可以),分别在其上书写数字 $1,2,\cdots,N$,将它们放在一个不透明的口袋中,彻底搅乱后,从中一次性抽出 n 个个体,被抽到的那些数字所对应的个体就作为我们的样本. 显然,从 N 个个体中一次性抽出 n 个个体,不同的组合结果共有 $C_N^n=\dfrac{N!}{(N-n)!n!}$ 种,每一个体出现在样本中的可能性(概率)为 $\dfrac{C_1^1\cdot C_{N-1}^{n-1}}{C_N^n}=\dfrac{n}{N}$. 可见这种做法满足随机抽样的要求.

上述抽球过程中的"一次性抽出 n 个个体"也可以解释为:"抽 n 次,每次抽一个个体,抽出的不再放回". 下面的计算表明,这种抽法上的"解释"不会影响每一个体出现在样本中的概率.

为了表述简洁,用事件 A 表示"个体 A 出现在样本中",用事件 A_k 表示"个体 A 第 k 次被取到",用事件 $\overline{A}_k$ 表示"个体 A 第 k 次没有被取到",则

$$
\begin{aligned}
P(A)&=P\{\bigcup_{k=1}^{n}(A_k)\}=\sum_{k=1}^{n}P(A_k)\\
&=P(A_1)+P(\overline{A}_1)P(A_2|\overline{A}_1)+P(\overline{A}_1\,\overline{A}_2)P(A_3|\overline{A}_1\overline{A}_2)+\cdots+\\
&\quad P(\overline{A}_1\cdots\overline{A}_{n-1})P(A_n|\overline{A}_1\cdots\overline{A}_{n-1})\\
&=\frac{1}{N}+\frac{N-1}{N}\cdot\frac{1}{N-1}+\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\frac{1}{N-2}+\cdots+\\
&\quad\frac{N-1}{N}\cdot\cdots\cdot\frac{N-n+1}{N-n+2}\cdot\frac{1}{N-n+1}=\frac{n}{N}.
\end{aligned}
$$

按随机抽样去获取样本，每一个体至多在样本中出现一次，因而称为“无放回的抽样”.

在实际应用中，抽样几乎都是无放回的，即要求同一个体不能在样本中重复出现. 若因为某种原因需要破除这一限制，也可以一个一个地抽，每次抽出球后，登记其上的数字后放回袋中，彻底搅乱再抽下一个，直到抽出 n 个为止. 这种抽样叫“有放回的抽样”，其样本大小 n 可以超过总体所含个体数 N.

有放回的抽样在理论分析上比无放回抽样简单，且在比值 n/N 很小(例如，不超过 0.05) 时，两种抽样方式的差别从实际观点看并不重要. 因此，有放回抽样在抽样方法中也占有一席之地.

如果总体中所含个体数很大，则按上述“口袋模型”去操作，要准备这么多球(或纸条) 并彻底搅乱，不是一件容易的事. 这时可以使用“随机数表”来代替这只“已经搅乱了的口袋”或者设计某种简易的方法自己产生随机数. 譬如，随意翻开一本书取其页码的尾数；掷几枚骰子取其点数和的尾数；各种彩票摇奖机；等等. 当然，以上这些都不能免除对总体中的个体加以编号这个麻烦.

以上介绍的随机抽样的思想可能一时让人难以接受，人们会认为既然所要达到的目的是使抽出的那些个体(样本) 能够尽可能地代表整个总体的情况(这是正确的)，那么，统筹协调后有计划地筛选，而不委之于随机性，岂不是能更好地实现这一点吗？问题在于后者难于免除人们多少都会存有一些主观意向或考虑不周，特别是当研究者希望得到某种结论时，更是如此. 另外，除非总体所含个体数 N 很小时(这时，抽样调查大概没有必要)，人们是很难掌握总体中个体的分布情况的，而人为地挑选个体可能导致主观偏差. 另一方面，根据概率论中的大数定律，随机抽样的方式保证了当样本容量 n 较大时，总体中具有各种性质的成分各按其比率均衡地出现在样本中，从而在这个无形的“自然调节”中实现了所企求的代表性. 实际应用的经验也证实了这一点.

在统计推断过程中，样本的获取是一项基础性工作. 工作中图方便、考虑不周，往往是破坏抽样随机性的重要原因. 如派调查员去一个县调查农民收入情况，他为图方便，只在县里交通较便利的河流、公路沿线挑选若干农户做了调查. 这种取样方式在无意中就已经带进了系统偏差，这是因为交通方便的地方，农民收入一般也较高，该调查员的抽样调查结果将不能反映全县农民的真实情况. 由于抽样方式的考虑不周导致统计推断失误的一个著名例子是 1936 年美国大选的预测. 当时民主党人罗斯福与共和党人兰登竞选美国总统，美国有一家著名杂志社做了大规模的民意测验，共调查选民千万人以上，做出回答的有 200 余万人. 据其结果，该杂志预言兰登将以压倒优势获胜. 大选结果却完全相反，罗斯福以压倒优势当选. 事后分析，其原因在于该杂志社是从电话号码簿、俱乐部名册、驾驶证编号等中选择被调查者的，这类人多属于富有阶层，倾向共和党者

居多. 另外,大量的"无反应"人群(约 800 万人)也造成了显著的偏差. 这类教训在抽样调查工作中值得记取.

根据总体的情况(指所含个体的多少,个体的分布,取样的难度等),在实际抽样时,上述简单随机抽样方案有时需做些变通. 常用的有以下两种:

一是"集团抽样". 即先把总体中的全部个体,按某种考虑分成一些大集团,每个大集团内又可分为若干个小集团,后者还可以再细分. 抽样时,先用随机的方法抽取若干个大集团,再在抽出的每个大集团内分别抽出若干个小集团,……这样下去,最后在最低一级的集团中随机抽出若干个体. 这样抽出的全部个体构成所需的样本. 这种做法是为了防止样本中的个体在地域上过于分散,从而减少"取样"的工作量. 举例言之,设要通过抽样调查了解某县农户收入状况. 该县农户以 10 万户计,计划从中抽出 400 户,若按简单随机抽样的方式去抽,首先需 把这 10 万农户中的每户都编上号,再随机地去抽,这实际是不可行的. 即便可行,抽出的这 400 户可能遍布全县的每一个角落,逐一调查甚为不便. 为缓解这种不便,可改用变通的集团抽样法:先在全县随机抽出若干个乡,再在抽到的每个乡中分别随机抽出若干个村,最后,从抽出的每个村中随机抽取若干农户,这样,最后抽出的农户相对集中一些,又不甚影响随机性. 在涉及全国规模的抽样调查中,这种做法几乎是必须的.

二是"分层按比例抽样". 举例来说,要了解目前大学教师的收入情况. 在全国,这类人数以 10 万人计,而我们只能调查其中一小部分,例如 1 000 人,在使用随机抽样方法时,由于抽取的量(1 000 人) 不算很大,大数定律的均衡作用未能充分发挥,于是在样本中可能会出现各阶层人员的比例与总体的比例有相当偏离的情况,这就会影响样本的代表性. 例如,若在样本中老教授偏多,则调查结果将偏高. 为了补救这一点,设计了"分层、按比例"的抽法. 把大学教师按现行职称序列分为助教、讲师、副教授、教授四层,根据各层人数的大致比例确定样本中各层人数的比例,然后在各层内用随机抽样的方法抽出所需的人数,这样抽出的属于各层的人即组成样本. 在这一抽样方案中,既有计划的部分,又有随机会而定的部分. 计划的部分(分层、按比例) 对机会的影响作用做了"宏观"的控制,而在"微观"(各层内) 上则让机会起调节作用. 这种做法的目的,不言而喻是为了限制机会的破坏作用,以使样本更具有代表性. 必须指出,这与前面批评过的那种按主观指定样本的做法毫无共同之处. 这里的分层是有客观依据的,并非由人们主观上觉得如何而定. 不过,为保证这种方法有效,必须有两个条件:一是分层的标准应合理. 这主要是指层与层之间确实有较大差异,而每层内各个个体的差异较小. 二是每层所含个体数在总体全部个体数中所占的比例要能够比较确切地知道. 若不然,由人凭主观想像定一个比例,反而会引起系统性的偏差.

另外,分层按比例与集团抽样可以联合使用,每层内可以分集团,集团内也

可以分层，用这种方式可以构造出各种复杂的抽样方案．当然，抽样方案的选定还要考虑实际问题的条件和需要．

以上所举的几个例子中，总体都是由有限个（即使数目可能极大，但仍有限）“看得见、摸得着”的实在个体组成．在这种情况下随机抽样（或任何一种变通的抽样方式）即使不见得易于实施，其意义还是清楚的．但在另一些问题中，总体与个体的关系不像这么清楚，且原则上，总体所含个体数可以为无限，这种总体叫无限总体．现举例如下：

例 1　调查某条河流中的水受污染的情况．以 1 立升河水为单位抽取若干单位做化验．这里“此河中任何 1 立升河水”都是一个个体．由于个体数非常大，因此可认为总体中包含无限个个体．还需要指出，此例的个体是在抽样中“制造”出来的．在抽样前，河中的水并没有自然地分成 1 立升一堆．

例 2　研究某种工艺下所生产的灯泡的寿命．为此，需用在这种工艺下生产的若干只灯泡做试验．在此问题中“每一只在这种工艺下可能生产的灯泡”（更确切地说，每一只这种灯泡的寿命）都是一个个体．由它们所组成的总体不但是无限的（只要不断地生产下去），而且其存在也只能凭想像，若你不去生产，就没有这种灯泡．因而总体中的个体不是早就等在那里让你去抽，而是要随着试验的进行产生出来．

例 3　在天平上称一物件以估计其重量．由于天平的精度以及各种微小偶然因素的影响，每次的读数可能都有一定的偏差．这时，每次称量的结果视为一个个体，所有可以想像的这种称量的结果的全体，构成这个问题的总体．这是一个其存在只能加以想像的无限总体．总体中的个体不是现成摆着的，而是每试验一次，就造出这样一个个体．

在有限总体下，虽然在实践中要保持抽样的随机性并非易事，但在原则上是可行的．而在无限总体下则不然，这时既无法给出随机抽样的确切含义，也不能给出一个一般性的可行实施方法．在无限总体下保证抽样的随机性往往就在于尽力避免各种可能产生的系统偏差．如在河水中抽样的例子中，应尽量多选几个地点、在不同的时间、不要太靠近工厂排水管、不要总在同一水层取样等．

第二节　样本分布和统计量

一、样本分布

样本是按照一定的方法从总体中抽出一部分个体所组成的集合．样本中所含个体数称为**样本容量**或**样本大小**．由于只是对个体的某项指标值感兴趣，因

试读结束：需要全本请在线购买：　www.ertongbook.com

此样本也就具体化为样本中个体的相应指标值(观测数字).

总体中包含很多个体,按随机抽样的方式获取容量为 n 的样本,但具体哪 n 个个体会出现在样本之中,却有一定的偶然性. 换句话说,样本将由哪 n 个观测数字组成是随机的. 因此,为了使有关统计方法的讨论能上升到理论的高度,在数理统计学中将抽样得到的那组具体观测数字$(x_1,\cdots,x_n)$看成是某 n 个抽象的随机变量$(X_1,\cdots,X_n)$的一组取值. 同时,作为概念的提升,将这 n 个抽象的随机变量$(X_1,\cdots,X_n)$称为样本,而把具体的观测数字$(x_1,\cdots,x_n)$看成该样本的一组取值.

这种观念的转变非常重要,它是统计方法理论依据的立足点,也是统计方法与概率论发生联系的纽带. 因为概率论是研究随机现象(抽象为随机变量)的学科,样本既然被视为随机变量,故建立在样本基础上的统计方法就可以借助于概率论这个分析工具来进行讨论,统计方法与概率论发生联系的根源就在这里. 搞应用工作的人往往习惯于样本是一堆具体数据的观点,但为了对统计方法的理论方面有所了解,就必须努力树立"样本$(X_1,\cdots,X_n)$是随机变量,而具体的观测数字$(x_1,\cdots,x_n)$只是样本$(X_1,\cdots,X_n)$的一组观测值"的基本观点.

样本$(X_1,\cdots,X_n)$既然是随机变量,也就有其概率分布. 样本的概率分布称为**样本分布**. 在概率论中已经知道,完整地描述一个随机变量莫过于给出它的概率分布.

下面将在有限总体随机抽样的情况下,讨论样本分布是如何确定的.

假定总体中共有 N 个个体,由于只对个体的某项指标 X 感兴趣,因此,也将指标 X 称为总体. 假定 N 个个体中,指标 X 等于 a_1 的有 N_1 个,等于 a_2 的有 N_2 个,…,等于 a_k 的有 N_k 个,显然 $N_1+N_2+\cdots+N_k=N$,因此,下表构成一个概率分布,称它为**总体分布**. 总体分布也就是所关心的指标 X 的概率分布.

X 的取值	a_1	a_2	…	a_k
取值的概率	$\frac{N_1}{N}$	$\frac{N_2}{N}$	…	$\frac{N_k}{N}$

从总体 X 中抽出的容量为 n 的样本$(X_1,\cdots,X_n)$被视为 n 个随机变量(实际上,是依次抽出并观测 n 个个体的指标值$(x_1,\cdots,x_n)$). 这 n 个随机变量分别服从什么分布?相互之间关系如何?是进一步对样本进行分析和对样本信息加以利用的基础.

首先考查样本分量 X_1 的分布. 由于抽样的随机性,N 个个体中每一个个体都有同等机会被抽到,所以针对不同个体而言,属等可能概型. 但不同个体未必

有不同的指标值，不同的指标值共有 k 个：$a_1,\cdots,a_k$，每一个值都可能成为 X_1 的观测值，由等可能概型中事件概率的计算公式知

$$P(X_1=a_i)=\frac{N_i}{N},\qquad i=1,2,\cdots,k. \tag{1.1}$$

这表明，作为随机变量的样本分量 X_1，其分布与总体 X 的分布（参见表 1）完全一样. 这是随机抽样的特征，也是为什么采用随机抽样的理由之一.

至于样本分量 X_2 的分布，在有限总体无放回抽样的情形下，严格地讲会依赖于第一次抽样的结果，原因是第一次抽出一个个体后，剩下 $N-1$ 个个体中对应 $a_1,\cdots,a_k$ 各种值的个体数的比例会有所变化. 但从实际应用的观点看，对那些需要做统计推断的有限总体而言，它们所含的个体数 N 通常都很大，因此，抽出一个个体后，对应 $a_1,\cdots,a_k$ 各种值的个体数的比例虽有所变化，但微小的变化可以忽略. 忽略这种变化后，采用与 X_1 类似的分析可知，样本分量 X_2 的分布律与 X_1 的分布律完全一样，仍然是

$$P(X_2=a_i)=\frac{N_i}{N},\qquad i=1,2,\cdots,k. \tag{1.2}$$

推而广之，若依次抽取 n 个个体组成一容量为 n 的样本 $X_1,\cdots,X_n$（实际得到的是它们的观测值 $x_1,\cdots,x_n$），则只要 n 相对于 N 而言很小（即 $\frac{n}{N}\approx 0$），样本各分量在分布上的差异都可以忽略，各次抽样的结果以及抽得各种结果的可能性大小可以认为是互不影响的. 因此作为随机变量的 $X_1,\cdots,X_n$，可以认为是相互独立且有相同的分布，每一个 X_i 的分布都与总体 X 的分布相同.

对于有限总体有放回抽样（每抽出一个个体，在观测记录后又放回总体，再随机抽下一个个体），每次抽取时，总体的成分显然保持不变. 样本各分量 $X_1,\cdots,X_n$ 之间也必然是相互独立且都与总体 X 有相同的分布.

对于无限总体的随机抽样，因为总体既然包含了无限个个体，抽出若干个个体后，并不会对总体的成分有任何影响，因而样本 $X_1,\cdots,X_n$ 独立且与总体同分布这一结论必然成立.

通过以上分析注意到，“随机抽样”的要求保证了样本分量 X_1 与总体 X 同分布的性质，“每次抽取时，总体的成分保持不变”的要求保证了样本各分量 $X_1,\cdots,X_n$ 相互独立且都与总体 X 有相同分布的性质.

由于随机抽样在有限总体无放回抽样的情形还不能严格满足“每次抽取时，总体的成分保持不变”的要求，而这个要求对于保证样本分布的优良性质从而使进一步的理论分析得以简化至关重要. 因此，统计学中将“每次抽取时，总体的成分保持不变的随机抽样”称为**简单随机抽样**.

随机抽样在有限总体无放回抽样的情形虽然不能严格满足“每次抽取时，总体的成分保持不变”的要求，但在样本容量 n 相对于总体所含的个体数 N 很小

的情形，每次抽取时，总体成分的微小改变可以忽略. 出于理论分析简化的需要也作为简单随机抽样来对待.

由简单随机抽样得到的样本称为**随机样本**，本课程所涉及的样本都是随机样本，故总是假定从总体 X 中抽得的样本 $X_1,\cdots,X_n$ 是相互独立的且都与总体 X 有相同的分布，并用大写字母 $X_1,\cdots,X_n$ 表示作为随机变量的样本，用小写字母 $x_1,\cdots,x_n$ 表示该样本的观测值.

二、统计量

数理统计方法虽然从样本着手，着眼点却是总体的分布或总体分布的某些特征量的推断. 从样本出发要得到有关总体的某种特征的推断，需要对样本所携带的有关总体的信息进行必要的筛选、浓缩、提炼，以便使与所关心的问题有关的信息更加突出，无关的信息能被剔除或有效抑制. 具体化，就是要对样本进行必要的加工和运算处理. 这种加工和处理的结果称为**统计量**. 从数学形式上看，统计量就是对样本实施某种特定运算的函数表达式 $g(X_1,\cdots,X_n)$，它们能表达一定的加工过程. 例如，样本均值

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

就是一个统计量，它表达了对样本 $X_1,\cdots,X_n$ 加工处理的一种方式，即“取长补短”，它突出表达了总体取值大小的信息. 按照这种加工方式，当样本 $X_1,\cdots,X_n$ 有了具体的观测值 $x_1,\cdots,x_n$ 之后，就可以加工（计算）出一个具体的值 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，将 $\overline{x}$ 称为统计量 $\overline{X}$ 的观测值. 注意 $\overline{x}$ 和 $\overline{X}$ 是有区别的，$\overline{X}$ 作为统计量是随机变量，而 $\overline{x}$ 是一个具体的数，对不同的 $x_1,\cdots,x_n$，$\overline{x}$ 可能取不同的值，而 $\overline{X}$ 表达的是对观测数据的加工方式，不随观测值而改变，仅仅是一个代数表达式.

同理，样本方差

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

也是一个统计量，它突出了总体取值分散程度的信息，相应的观测值为

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

常用的统计量还有：

样本标准差

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

它也刻画取值的分散程度，且与总体有同样的量纲.

样本 k 阶原点矩

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\qquad k=1,2,\cdots;$$

样本 k 阶中心矩

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,\qquad k=1,2,\cdots,$$

等等. 以上这些统计量都只需对样本进行简单的运算就可得到，还有一些统计量则需对样本进行一定的加工处理才能得到. 如，从总体 X 中抽取一容量为 n 的样本 $X_1,\cdots,X_n$，设相应的观测值为 $x_1,\cdots,x_n$，将观测值按由小到大的次序重新排列为

$$x_{(1)}\leqslant\cdots\leqslant x_{(n)},$$

则称 $x_{(1)},\cdots,x_{(n)}$ 为原始样本观测值 $x_1,\cdots,x_n$ 的**次序样本观测值**. $x_{(1)},\cdots,x_{(n)}$ 由 $x_1,\cdots,x_n$ 确定，$x_1,\cdots,x_n$ 的值有一定随机性，导致了 $x_{(1)},\cdots,x_{(n)}$ 的值也有一定随机性. 为此，将次序样本观测值 $x_{(1)},\cdots,x_{(n)}$ 看成(想像成)某 n 个随机变量 $X_{(1)},\cdots,X_{(n)}$(其分布可能与 $X_1,\cdots,X_n$ 的分布截然不同)的观测值，如此定义的 n 维随机变量 $(X_{(1)},\cdots,X_{(n)})$ 被称为原始样本 $X_1,\cdots,X_n$ 的**次序统计量**.

由此派生出的统计量还有：

样本中位数

$$\tilde{X}=\begin{cases}X_{(k+1)}, & \text{若 } n=2k+1,\\ \dfrac{1}{2}(X_{(k)}+X_{(k+1)}), & \text{若 } n=2k.\end{cases}$$

样本极差

$$R=X_{(n)}-X_{(1)},$$

等等. 统计量定义的实质在于：统计量只依赖于样本 $X_1,\cdots,X_n$，而不涉及任何其它未知的量，即它是样本的已知函数 $g(X_1,\cdots,X_n)$，且不能含有任何未知参数. 也就是说当样本 $X_1,\cdots,X_n$ 有了观测值以后，按照统计量 $g(X_1,\cdots,X_n)$ 表达的运算加工方式，一定能算出一个具体的数字 $g(x_1,\cdots,x_n)$(或向量). 例如，对总体 $X\sim N(\mu,\sigma^2)$，若令

$$T=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2,$$

此时，如果 μ 已知，则 T 是统计量；如果 μ 未知，那么 T 就不是统计量.

一个统计量若有约定的名称，那么它的观测值也应用相同的名称. 如，$\overline{X}$ 称为样本均值，其观测值 $\overline{x}$ 也称为样本均值. 其区别是明显的，前者为函数表达式，后者为一数字(若没带入具体数字，则用字母的大、小写来区别).

如上引入的“统计量”是对样本进行加工处理和运算的结果，是统计推断这项“工程”的“预制构件”. 在一个具体推断问题中需引进什么样的统计量，当然要由问题的需要和拟采用的推断方法而定. 在以后各章中将会看到以上统计量以及更多其它统计量的种种应用.

第三节　总体分布的理论分析

统计量是对样本进行处理加工的结果，是用来进行各种统计推断的基本素材. 只要统计量的选择得当，可以使样本中所关心的那些方面的信息更为突出. 因此需要对统计量有更深一步的认识，即了解其特点和性质.

统计量作为样本的函数，本身是一随机变量，而对随机变量的完整描述就是给出它的分布.

统计量取值的随机性源于样本取值的随机性，因此，统计量的分布依赖于样本的分布，而样本与总体的分布相同. 因此，要确定统计量的分布，必须先了解总体的分布. 但是，作为统计推断问题，总体分布总是未知或部分未知的，是统计推断的对象. 在大多数场合，人们对总体分布并非一无所知，而是略知一二.

例如，在涉及随机波动或误差这类总体时，通常认为它服从正态分布 $X \sim N(\mu,\sigma^2)$，只是其中的均值 μ 和方差 σ^2 未知.

又如，在涉及稀有事件出现的频数时，通常认为其服从泊松分布 $\pi(\lambda)$，只是 λ 未知.

再如，在涉及寿命问题时大多认为其服从指数分布，其密度为

$$f(t)=\begin{cases}\lambda e^{-\lambda t}, & t\geqslant 0,\\ 0, & t<0,\end{cases}\quad \text{只是 } \lambda \text{ 未知.}$$

以上这些“认为”都是事先对总体分布类型的猜测或假定. 那么这种猜测或假定有什么依据呢?其实，可以从已积累的经验或对样本数据的直观考查(如直方图、频率分布表等)当中受到启发，即用归纳的方法推测其分布. 但从理论上讲，它们是否具有这种内在的必然性呢?也就是说能否从理论上，在某些基本合理的假定下演绎出这些必然的结论. 下面将从理论上对此做些分析. 鉴于本

教材的性质，不去追求理论上的严谨和完善，只希望通过这种分析能使读者相信，对各类实际问题，其总体分布确实存在着内在的必然性，并了解这种理论上的抽象、建模、简化和演绎的大致思路.

一、正态分布的客观背景

在现实生活中有许多随机现象(变量)，其随机性是由大量相互独立的随机因素的综合影响所造成的，而其中每个因素在总的影响中所起的作用都是很小的.

例 4　某地区成年男性的身高. 对不同的个体，其身高不全相同，因此是一随机变量. 造成这种随机性的因素很多：先天的遗传因素、后天的生长环境、生长发育不同时期的营养状况、健康状况，等等，许许多多的因素共同作用造成了身高的参差不齐. 在正常情况下，可以认为这每一个因素都不能起压倒其它一切的主导作用.

例 5　某批灯泡的使用寿命是一随机变量. 影响每只灯泡寿命的因素有：原材料的质量、生产工艺、使用过程中的电压波动以及环境条件，等等. 正是由于这许多因素的共同作用造成了灯泡寿命的长短不一.

其它类似的例子在实际中还很多，如一个城市的耗电量是大量用户耗电量的总和；一个物理实验的测量误差是由大量无法避免的微小偏差所合成的；炮弹射击的弹着点与目标的偏差是随机的，产生偏差的原因有瞄准时的误差、炮弹或炮身结构所引起的误差、空气阻力产生的误差，等等. 这些例子的共同特点是所感兴趣的那个量是一随机变量，它的随机性是由大量的、每一个又都是微小的随机因素共同作用的结果. 若用 ξ_k 表示其中第 k 个随机因素对 X 影响的大小，则 ξ_k 也是随机变量，X 可表示为 $X=\sum_k \xi_k$. 下面将探讨 X 的分布：

为了理论分析的需要，先简单介绍一个辅助工具 —— 特征函数，以及几个将用到的性质.

定义 1　设 ξ 是一随机变量，ξ 的**特征函数**定义为

$$f_\xi(t)=E[e^{it\xi}],\qquad t\in \mathbf{R}$$

其中，i 是虚数单位；$E[e^{it\xi}]$ 表示对 ξ 的函数 $e^{it\xi}$ 求数学期望，则 $f_\xi(t)$ 是定义在实轴上的一个函数，可以看成是随机变量 ξ(或 ξ 的分布) 的一种变换. 任意的分布都有特征函数存在，且特征函数 $f_\xi(t)$ 与 ξ 的分布函数是一一对应、相互可以唯一确定的.

例如，泊松分布 $\pi(\lambda)$ 随机变量 ξ 的特征函数

试读结束：需要全本请在线购买：www.ertongbook.com