全国高等林业院校试用教材

抽样技术

宋新民　编著

中国林业出版社

全国高等林业院校试用教材

抽样技术

宋新民 编著

中国林业出版社

图书在版编目(CIP)数据

抽样技术／宋新民主编.—北京：中国林业出版社，1995.12
全国高等林业院校试用教材
ISBN 7-5038-1429-2

Ⅰ.抽… Ⅱ.宋… Ⅲ.森林抽样调查-基本知识-高等学校：专业学校-教材 Ⅳ.S757.2

中国版本图书馆CIP数据核字（95）第03408号

中国林业出版社出版
（100009 北京西城区刘海胡同7号）
河北昌黎县印刷厂印刷 新华书店北京发行所发行
1995年10月第1版 1995年10月第1次印刷
开本：787×1092毫米1/16 印张：14.25
数字：343 千字 印数：1—1000 册
定价：11.10元

前　言

本书是为全国高等林业院校林学专业编写的试用教材。随着我国教育体制和专业方向的不断改革，近几年许多林业院校都感到《测树学》中所讲授的抽样调查内容的不足，要求开设《抽样技术》这门课。为了满足教学的急需，受林学专业指导委员会的委托，我们编写了《抽样技术》一书。

本书比较全面、系统地论述了常用抽样方法的理论和方法，包括抽样调查方法步骤、样本组织、基本估计原理与效率分析。在内容次序安排上大体分为两个层次，抽样方法、统计估计方法和算例等实用部分排在每章的前面；把估计原理，公式证明、效率分析评价放在后面。这样做能突出重点，便于教学。使用时可根据教学对象不同，学时多少不同供教师选择。书中凡涉及到林业院校《数理统计》已讲过的一些方法、公式，本书基本上直接应用，对超出《数理统计》教学计划或未讲过的部分内容，比如：比估计有偏性、系统内与群内相关系数，两阶抽样、整群抽样、联合估计、不等概抽样等估计值及其方差公式的证明，难度较大，又不是本课程所要研究的主要问题，鉴于这方面的参考文献又很少，我们用注记的形式列于各章后面，供师生参考。此外，书中虽然应用不少数学方法，但数学方法在这里只是工具，是为实际应用服务的，因此，对一些数学理论未作严格证明。

本教材所讲授时间为40～60学时，各院校可根据具体情况对其内容自行取舍。课内实习可作5～7次，由教师给出图、表及调查资料，如能结合综合教学实习、生产实践及社会调查则更有利于提高教学效果。

在编写中，我们立足于实际应用，从实际出发，引进适合我国国情的国外抽样方法；同时，特别着重总结了我国广大林业调查工作者30多年来应用抽样技术的丰富经验，使这本书具有中国特色。可以说，这本《抽样技术》的出版是我国农林科技工作者、调查员集体智慧的结晶。

本书的主要内容是参考我校举办的历届森林调查进修班所用讲义编写成的，考虑到其他专业要求，增添了运用抽样方法进行农业估产、社会经济、森林病虫害等方面的调查内容。其中大部分内容曾在林学专业专科生、本科生及研究生中作为教材讲授，以后不断加以改进和充实。第十三章森林连续清查，采用《测树学》中的全部内容。

在编写过程中，得到符伍儒、周沛村、贾乃光、黄用廉等教授的帮助和鼓励，最后由贾乃光教授悉心对全书作了审定，在此，表示诚挚的谢意。

本书第一次作为教材出版，加之编者水平所限，错误和疏漏在所难免，敬希读者批评指正。

宋新民

1994年7月于北京林业大学

目　录

第一章　抽样调查的基础知识

第一节　抽样调查的目的与应用

抽样技术是一门应用广泛的学科，它是以概率论和数理统计为基础，专门研究抽样方法、抽样理论及其应用的学科。抽样技术是现代统计学的重要组成部分，它既是统计调查的方法又是统计分析的方法，并将两者结合起来，成为整个统计理论中不可缺少的而又成熟的一个分支，为世界各国所重视。

人们的认识和行动很大程度上依赖于掌握信息的多少，这一点在日常生活和科学研究中都是正确的。当今，社会已步入信息时代，无疑信息量多少直接影响着政府、部门、企业及个人的认识与行为决策。信息的采集方法，有下列几种：全面调查（普查）、典型调查、重点调查、定期报表汇总和抽样调查等。

全面调查可以获得调查对象的实际全面信息，比如我国10年一次的全国人口普查；政府各级管理部门的定期报表制度，它类似普查性质，也是常用的一种方法，因为统计报表在逐级填报、累计过程中往往受工作人员素质的影响，有时发生虚假现象，使数据严重失真。如长期以来我国农作物种植面积和产量的统计数据；各年造林面积及成活率的统计，都有过虚假现象，所以1982年后，政府不再采用累计报表数据、代之以专业队伍的抽样调查。

典型调查、重点调查都属于有意抽样（purposive selection）的一种形式，它是以调查者主观判断取样，即凭调查者对现象的了解和自己的判断能力，从总体中选取具有平均水平的典型单位作为调查对象。这种方法的优点在于可以发挥调查者的主观能动性，充分利用被调查对象已有的信息、避免发生很大的偏差。这种方法多应用于为某种特殊目的进行的专业调查，比如，林业上用标准地资料编制林分生长过程表、标准表等。但是由于这种方法受人为主观因素影响，不仅常常发生评价标准不同，意见不统一，而且难于避免因调查者的主观意图所造成的偏差。

统计方法作为认识的方法，已经历了较长的发展时期。抽样调查方法则是本世纪才发展起来的，而它的自身发展又经历了若干阶段，直至1925年在罗马举行的第16次国际统计学会上“抽样方法应用研究委员会”才从理论和实践上充分肯定了抽样方法的科学性。1940年后，抽样方法被世界各国普遍采用。目前，世界各国政府除了对基本国情、国力调查采用全面调查和统计报表外，其他大量的社会经济调查、自然资源调查，则都采用抽样方法。例如，我国的农产量和种植面积，考察市场物价变化，社会公众的民意测验等，以及对自然环境的评估，森林生态效益的调查，乃至于在实验室中所做的科学实验，工厂产品的检验等都离不开抽样方法。

抽样调查的基本内涵，是根据非全面调查资料，来推断（估计）全面的情况。抽样可以

是有意抽样，即凭主观认识选取样本；也可以是概率抽样，即按规定方式进行的等概或不等概抽样。本书所讨论的抽样理论和方法是指概率抽样，从全部所研究对象之中，抽取一部分单位，进行实际调查，并依据所获得的数据，对全部研究对象的数量特征作出有一定可靠程度的估计和判断，以达到对现象总体的认识。

现在抽样技术的应用范围还在不断扩大，它的抽样方法和估计理论，已成为统计学中发展最快、最活跃的分支之一；抽样技术所提供的各种方法还构成其他应用科学的基础，如计量经济学，管理会计等。

第二节　抽样方法的优点

一、费 用 较 低

抽样调查是非全面调查，是由部分推断整体的一种方法。它只对部分单位进行实际调查，但研究的目的是对全部对象的数量特征，如总体平均水平、总体规模、结构等作出估计。由于抽样调查既能省时、省力，又能达到认识总体的目的，这就表明它的科学价值。例如，根据百万分之五的城市居民家庭收入，可以推算全国上亿户城市居民的消费水平；根据不足万分之一的农作物收获面积的实际产量，来推算一个县、一个省乃至全国的农产量；用千分之几的森林面积，可以估计一个几十万公顷林场(林业局)的森林蓄积量等。对于大范围的抽样调查，其经济效益更为明显。

二、速　度　快

抽样调查成果的时效性是普查、统计报表不可比的。抽样调查的工作量小，组织专业队伍直接取样，减少中间环节，提高了时效，特别适宜于时间性要求很强的调查项目。以农作物产量调查而论，依靠全面报表制度，一个地区从收割、打晒到称重入仓要花很长时间。另外，像市场物价水平、自然灾害预测，抽样调查的时效性更是其他方法无法比拟的。

三、精度高，有概率保证

由专业人员实施抽样调查，不仅便于组织，而实地调查工作容易受到有关人员的指导和监督，能保证数据的准确。更重要的是取样可以按照随机原则，排除人为主观因素的影响，使样本有较好的代表性，可以计算抽样误差，并可通过抽样过程控制误差。对有经验的调查者，典型选样有可能达到非常准确的结果，但是它的精度及可靠性却无法给出。抽样方法则不仅能估计总体特征的数量指标，还能指出在不同概率保证下的误差限，这是其他部分调查方法无法做到的。

四、抽样方法的灵活性

抽样调查的内容可多可少，调查范围可大可小。既适用于专项性质的研究，也适于经常性的调查，如政策评估、市场信息、环境监测等只要需要随时都可因地制宜实施抽样调查。

五、应用范围广

与全面调查比较，除了上述各项优点外，抽样方法还解决那些无法全面调查或很难调查

的一些问题。无法全面调查主要有以下几方面：

1．无限总体。例如，气象因子调查、新材料、新设备、新工艺的检查等。

2．包括未来时间序列的总体。如生产过程稳定性检查等。

3．破坏性的产品质量检验。例如，灯泡寿命、木材抗折力检验、轮胎里程试验等。这些只能用抽样方法对总体作出判断。

还有一些很难全面调查的现象，有如下几方面：

1．非常大的有限总体。虽然是有限总体，因数量、范围太大，进行全面调查实际上不可能。如一个林场的林木总株数、水库的鱼苗数等。

2．有些调查对象，根据调查任务要求，也没必要全面调查。如民意测验等。

3．有些调查受时间和条件限制，不允许进行全面调查。

此外，随着抽样理论的发展，目前，世界上许多国家还用抽样方法进行生产过程的质量控制，将事后的调查、检验、估计推广到生产过程的控制，用抽样方法提供有关信息，分析各种有利和不利的因素，以便采取措施，使生产过程保持稳定运转。利用抽样结果进行风险预测，为人们的行为决策提供依据，是抽样方法的又一新的发展。用抽样方法对总体特征的某种假设进行检验，判断假设的真伪，人们可做出是接受假设还是拒绝这一假设，以期达到在冒最小的风险下取得最佳效果。例如，对某一种新药物是否推广，当然首先要取决于它的疗效是否显著，但疗效对每个人都受随机因素的影响，所以我们需要对药的疗效是否显著或不显著作出一定假设，然后根据试验结果及检验所作假设是否成立，从而对能否推广做出抉择。

第三节　总体与样本

一、总体及其有关概念

1．总体　调查对象的全体称总体。由于它是产生样本的基础故也称母体。组成总体的每个基本单位称总体单元。例如，若调查某个地区的人口情况，该地区的全体居民便构成总体，该区每个人便是总体单元；又如我们要研究某林场的森林总蓄积量，则林场全部立木的材积就构成总体。总体单元可以是自然单位，如林场内每一棵树，也可是人为划分的单位，如林场内以 $1000m^2$ 面积上的立木材积为单元。实际中应用最多的是人为区划的单元，如村庄、街道、居委会等。

在抽样中，必须弄清目标总体和抽样总体两个既有区别又有联系的总体。目标总体是指研究对象的总体单元之集合。抽样总体是按某一标志排列，供抽取样本的那部分单元的集合。例如，研究对象是整个林场的蓄积量，而林场内有一些农田、荒山，如在有林地中抽样，所得到的森林蓄积量只适用于有林地总体；要想适合于全林场，必须有其他信息，如农田和荒山的面积，否则就会导致偏差。通常解决的办法就是使目标总体单元和抽样总体单元一致。

总体单元既然可以是自然单元，也可以人为划分，于是总体就有有限总体与无限总体之分，我们把含有限单元数的总体称为有限总体；包含无限单元数的总体称为无限总体。

2．标志、标志值　为说明总体单元在某一方面的特征而采用的名称即为标志。

如果总体单元的特征是用数量表示的，如年龄、树高、胸径、海拔高等，称为数量标志。如果标志是用属性表示的，如树种、坡向、健康或不健康等，称为非数量标志或品质标志。

每个总体单元在数量标志上所观察到的数值即称单元标志值，如树高 15m，胸径 12cm 等。对于品质标志也可以转换为数量标志值，通常的做法是将具有某种品质标志的记为 1，不具该品质标志的记为零。如健康者取值 1，不健康者取值为零。

3．总体特征数　总体特征数是指描述总体所有单元在某标志上数量特征的数值。包括总体平均数、总体总量、总体成数、总体方差和标准差等，又称总体参数。显然，对于一个总体来说，这些总体特征数是唯一的、确定的数值，而且这些总体特征数通常是未知的。进行抽样调查的目的在于通过抽取部分总体单元，对总体某些特征数作出估计，就是说，即使进行抽样调查和推断，也只能给出这些总体参数的估计值。

二、样本及其有关概念

1．样本　从全部总体单元中，按照预先规定的方法抽取一部分单元，则被抽出的这部分单元之集合称为样本，又称子样。组成样本的每个单元称样本单元，样本单元数又称样本容量。通常总体单元数用N表示，样本单元数用 n 表示。

2．样本的抽取　从含有N个单元的总体中抽取 n 个单元，抽样的方法有等概抽样和不等概抽样两种；如果抽样是按照随机原则，即能保证使每个总体单元都有同样的机会被抽中，称为等概抽样。这种抽样方法是最常用、最基本的方法。不等概抽样是指总体各单元被抽中的概率与各该单元大小成比例的抽样。本书将在第十二章专门讨论不等概抽样问题。

不论是等概抽样还是不等概抽样，都可以用重复抽样或不重复抽样两种方式进行。在应用中，一般常用不重复抽样，这是因为通常总体单元数很大，即使有限总体中，抽样比 $f=\frac{n}{N}$ 也很小，所以常把不重复抽样视为重复抽样，如在森林资源调查中，当 $f<0.05$ 时，即按重复抽样对待，其结果会使抽样误差略大些。

3．样本单元的形状和大小　样本单元是总体单元的一部分，因为我们实际调查观测的是样本单元，它的形状和大小直接影响着调查的工作量、质量和成本。这个问题在社会经济抽样调查中似乎不那么重要，但对大面积的自然资源调查却是十分重要的。下面介绍森林资源调查中有关这方面的一些技术问题。

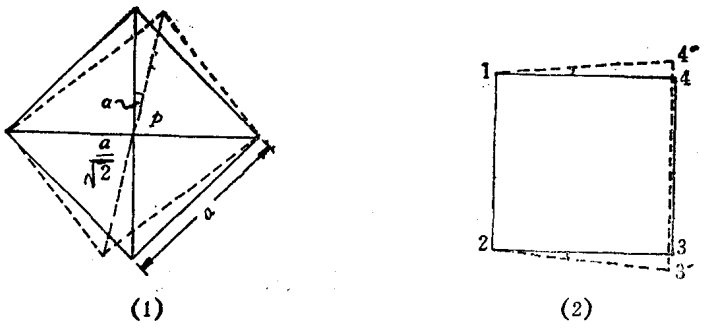
从理论上讲，凡是按照随机原则去抽取样本，不论单元的形状和大小如何，都可以获得总体特征数的无偏估计值。在既定的精度或总费用条件下，单元形状和面积大小，对抽样效率影响则是显著的。

森林资源调查中，依单元形状可分为：

① 面积（样地、样方）抽样；② 样点（角规点、成数点）抽样；③ 线段(截距)抽样；④ 样木（单株树）抽样。其中应用最多的是样地和样点两种。

(1) 样地的形状：样地有带状、矩形、圆形和方形。带状样地包含的地形复杂，变动小。但带状样地周界长、边界木多，易产生误差。一般认为矩形较好，便于设置，其长与宽之比以 2:1～5:1 合适。方形样地测量边界容易，边界木少，一般以样点为样地中心，分别量取对角线一半，如图1-1(1)。方形样地灵活性大，但比圆形样地周界长。若用测边法设置，角度闭合差大时，面积误差大，会影响估计精度。以图1-1(2) 而言，假如量测 1、2 两点产生5°的外偏，其面积误差达到8.5%；对角线法，用$a/\sqrt{2}$距离确定四个角点位，同样偏 5°，其面积误差只有－0.4%。

圆形样地也称样圆，是以样点为圆心，用半径r确定，方法比较简单，一般面积小于0.04 ha，测设最快。在凹凸不平的坡地设置样圆，应注意由八个方向实测坡度，改算半径，如图1-2。



(1)　　(2)

图 1-1

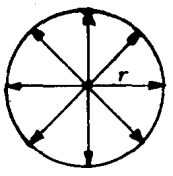


图 1-2　圆形样地设置

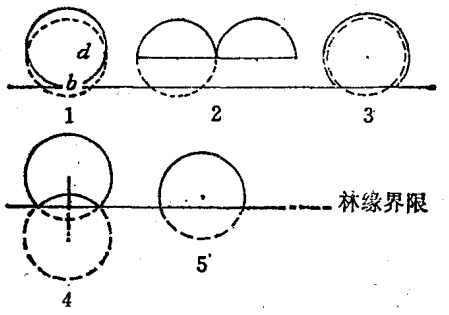


图 1-3　林缘样圆处理示意图

样地边界应清晰，检尺前对恰恰处在周界线上的树木取舍应有统一规定。

样地往往落于林缘，为使估计值无偏，必须在样本单元值上正确表达林缘情况。有两种情况：一种是样地中心落在林分界外，虽然经过分别测定、计算可以用不完整的和完整的样地材料，对总体做出无偏估计，但确定这些样地各地段面积很复杂，故一般舍去。另一种是样地中心落在界内，一般有5种处理方法，图1-3中的①对样地部分落在边界外的样圆，向林内移动，直到样圆圆周与边界相切；②改为相邻的两个半圆；③扩大半径增加林内样圆面积；④折影法，先测出样点至林缘的距离d，再在林外用d得第二个样圆，把两个圆在林内重叠部分测两次；⑤样圆不变，用比例法求出整个样圆值。

(2) 样地的大小：样地面积越大，变动越小，单元大小与变动系数之间的关系是，变动系数随单元面积增大而减小，当增大到一定程度时，变动系数趋于稳定，面积增大到0.05ha以上，变动系数开始稳定。当样本单元数相同时，面积大的样地估计精度高于面积小的样地。但是，面积大的样地耗费人力多，成本较高，因此，最优样地面积的确定以变动系数开始稳定的面积为宜。

样地面积我国一般采用 0.06～0.08ha，在林分变动较大的林区用0.1ha，幼龄林用0.01 a较适宜。

第四节　抽样调查的主要工作步骤

由于调查总体的复杂程度和要求不同，制定一个抽样调查计划和实施的若干步骤是非常必要的。在人口稠密、交通方便、经济发达的地区进行调查，与在交通不便的深山老林、边远山区，甚至没有地形图的地区进行调查两者区别很大。同样的调查内容，在某个地区难以解决，而在另一地区可能很容易调查。

一次抽样调查主要步骤如下：

一、明确调查目的

调查目的是制定计划的基础，主要应包括调查成果的要求、精度、详细程度，这是各项调查不可缺少的。

二、确定总体范围

划清总体的范围界限，如系地域性的调查，应准备的地形图、平面图、航空像片等资料。将总体界限勾绘在图上，并在图上求出所需的面积，为使目标总体和抽样总体一致起来，可将非抽样对象单独区划出来，这样做可以防止偏差又能提高工效。

三、划分总体单元

单元数量越大总体变动越小，但从抽样效率来看，单元数量大，调查费用增加，比如要调查一个乡的四旁树株数要比调查几十户的四旁树株数困难得多。另外，测量一个 1000m² 的方形样地要比量测 5 个 200m² 的圆形样地费工。在抽样比相同的情况下，单元小，样本单元数多的抽样精度高，这个问题在抽样前应慎重考虑。

四、收集资料

与本次调查有关的前人留下的历史资料都应尽量利用，如调查报告、文件、统计报表、专业报告及图面资料等。

五、设计抽样调查方案

它主要体现在每次调查所制定的原则方案和抽样调查技术细则之中，内容包括调查目的任务、精度要求、经费预算、抽样方法、计量方法及标准等一系列调查人员必须遵守的统一规定。

六、预先试验

当调查项目多，对象比较复杂的情况下，可制定不同抽样方案，先在小范围内进行试验，目的是依据试验结果，对不同方案的误差和成本进行评价，同时对方案的技术问题进行改进，达到选择最佳技术方案。

七、制定抽样框，抽取样本

抽样前应编制总体单元清单——抽样框。要求抽样框中单元不应有遗缺或重复。依据规定的抽样方法和确定的样本单元数进行样本抽取。

八、实地调查工作的组织

包括调查队伍的组织、分工，人员技术培训、质量检查制度，后勤保障等。

九、数据综合分析

首先是对外业调查数据、图面区划进行检查，弃舍和订正；待资料完整无误后进行统计分析，这些工作主要由计算机完成。

十、调查文件的编写

调查成果应于调查完后迅速提供调查说明书或调查报告、专题报告。其主要内容是调查最终结果、数量和估计精度以及采用的主要技术方法。必要时还应有各种附件：如统计表、图件、实物像片等。

第五节 抽样误差

一、误差的概念和种类

一般说待测物在每种标志上都有其固有值，这个值称真值。测定或估计的目的就是要了解这个真值。在实际中所得到的观察值和样本统计量都是测定值或估计值，用它们来估计真值不可避免地会产生误差，因此我们可定义误差为测定值或估计值与真值之差。即

$$\text{误差}=\text{测定值或估计值}-\text{真值}$$

抽样调查以样本统计量估计总体参数，中间要通过许多环节，从抽取样本、调查、测定、记录、统计计算至估计方法，都可能出现误差。这些误差有的可能相互抵销，有的会累积，但最终反映在估计值与真值（总体参数）之差上。我们把从样本单元调查测定以及估计过程中产生的各种误差的综合量称为抽样总误差。

误差的分类可从各种角度划分。根据误差的性质和来源可把总误差分解为非抽样误差、偏差和抽样误差三类。弄清它们的意义、性质、排除措施以及它们之间的关系都是抽样估计中需要着重解决的问题。

二、非抽样误差

非抽样误差是指不是由于抽样和估计方法引起的误差。它不是抽样调查固有的，即使进行全面调查也会存在。其来源很多，例如，过失性错误，调查人员错测、错记、被调者无回答、虚报瞒报等等；又如，测量误差，任何仪器在无偏差的情况下，也不可能获得标志值的真值，不过这两种误差，前者可以通过做好调查人员的培训、教育、宣传、检查等措施来排除。测量误差也可视为随机误差，实际上无法避免，只是随着样本单元数的增大，其误差值逐渐减小罢了。

在总误差中，一般不包括非抽样误差，但并非它不重要，而是因为这种误差的来源复杂，无法用抽样理论去估计它的量。应认真研究非抽样误差可能的来源及量的大小，并注意采取预防措施使之减少到最低限度。

三、偏　差

数学期望这一概念可用于由样本提供的估计值，如果 $\hat{\mu}$ 是由某个样本 n 提供的总体参数估计值，它的数学期望值等于总体平均数 μ，即 $E(\hat{\mu})=\mu$，则样本估计值是无偏的。

如果不是这样，估计值 $\hat{\mu}$ 则是有偏的，偏差（Bias）等于估计值的数学期望与总体参数实际值之差。

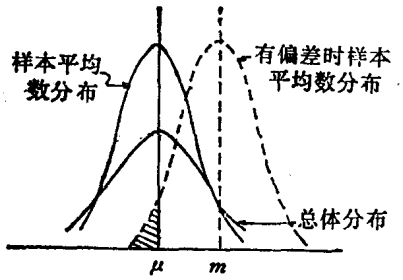


图 1-4　偏差对样本平均数分布的影响

在有偏估计情况下：$E(\hat{\mu})=m\neq\mu$，其偏差：$B=m-\mu$。图 1-4，说明总体为正态分布，总体平均数为 μ，在无偏估计时，样本平均数的分布以 μ 为中心。在发生偏差情况下，各样本平均数的分布以 m 为中心。其偏差 $B=m-\mu$。这时以有偏的样本平均数估计总体平均数，使估计误差限失去原来的意义，因为图 1-4 中小于总体 μ 的概率只是斜线面积部分，而大部分估计值落于 μ 的右边，所以估计值偏高。

偏差也称系统误差或恒定误差。其来源可归纳为下列三个方面：

1．测量仪器　用没有校正或不合格的仪器测定样本单元，会带来方向一致的偏差。例如，用偏小的轮尺测定一株实际胸径为 20cm 的林木，检尺结果为 21cm，所产生断面积偏大值 $B=1.1025-1=0.1025$（即 10.25%）。除此之外，有时还会因使用经验数表、数据分组（直径整化）、观测者或航片判读者的视觉等因素引起一定偏差。

2．抽样过程　如采用系统抽样，遇到总体周期性变动的影响；典型选样；沿交通方便地带多取样地代替不可及地带；通过电话号码簿抽取居民户进行收支调查等等，也都会造成有偏估计。

3．估计方法　采用有偏的估计方法，例如，对随机抽取的样本，用比估计方法估计总体参数，通常估计值是有偏的。不过这种偏差量是可以用数学理论估计的。

偏差的性质与随机误差不同，它具有确定的方向性，不能相互抵销，即是说它不随着样本单元数的增加而减小，同样也不因量测次数多而减小。

为了研究偏差的存在如何影响概率，下面引用W·G·科克伦著《抽样技术》中的计算结果（表 1-1），表中计算了估计值的误差大于 1.96σ 的真正概率，这个误差是与实际 μ 值相比较而计算的。

从表中清楚地看到：干扰量完全取决于偏差对标准差的比值。对于大于 1.96σ 的全部概率，如果偏差小于$\frac{1}{10}$的标准差，那么偏差的影响就很小。当偏差等于标准差的$\frac{1}{10}$时，则总的概率是 0.0511，而不是我们原以为的 0.05。当偏差进一步增大时，干扰量变得更严重。当 $B=\sigma$ 时，总的误差的概率为 0.17，约大于原来假想的 0.05 的 3 倍。

两端受到的影响是不同的。对一个正偏差（如图 1-4），当 $B=\sigma$ 时，低估值超过 1.96σ

表 1-1 偏差B对大于1.96σ的概率影响

B/σ	误差为下列数值时的概率		总　和
	$<-1.96\sigma$	$>1.96\sigma$	
0.02	0.0238	0.0262	0.0500
0.04	0.0228	0.0274	0.0502
0.06	0.0217	0.0287	0.0504
0.08	0.0207	0.0301	0.0508
0.10	0.0197	0.0314	0.0511
0.20	0.0154	0.0392	0.0546
0.40	0.0091	0.0594	0.0685
0.60	0.0052	0.0869	0.0921
0.80	0.0029	0.1230	0.1259
1.00	0.0015	0.1685	0.1700
1.50	0.0003	0.3228	0.3231

的概率很快从预定的 0.025 缩小到 0.0015，这时相应的高估值的概率却稳步增大。在大多数的应用中，只注意总的误差，但偶尔会特别关注到一个方向上的误差。

在应用中，作为一条规则，若偏差小于估计值标准差的1/10，则此偏差对估计值的准确度（accuracy）的影响可以略而不计。例如，在比估计中，当 $B/\sigma<0.1$ 时，偏差就可忽略不计。应用这个结论时，要分清偏差的来源，在比估计方法中，可以从数学上找到比值B/σ的上限，当样本 n 足够大时，可以相信 B/σ 不会大于 0.1 。相反地，对来自其他方面的偏差，想找到一个小的 B/σ 的可靠上限，通常是不可能的。这样说并不意味着有偏估计方法完全不能用，在偏差比较小的情况下，有时比估计所给出的误差限比简单随机抽样估计值的误差限还小。

为了比较有偏估计量与无偏估计量，或比较两个偏差量大小不同的估计值，一个有用的比较标准是用均方误差（mean square error），缩写为 MSE，它是与要估计的总体值相比较而计算得出的。用公式表示。

设 $E(\hat{\mu})=m$，则

$$\begin{aligned} MSE&=E(\hat{\mu}-\mu)^2=E[(\hat{\mu}-m)+(m-\mu)]^2\\ &=E(\hat{\mu}-m)^2+(m-\mu)^2\\ &=(\hat{\mu}\text{ 的方差})+(\text{偏差})^2 \end{aligned} \tag{1-1}$$

由于 $E(\hat{\mu}-m)=0$，交叉乘积项为 0 。

使用 MSE 作为估计值精度的标准，就是认为两个估计值有相同的MSE。严格说来并不对，因为当两个估计值有大小不同的偏差时，大小不同的误差 $(\hat{\mu}-\mu)$ 的频率分布不会是一样的。然而汉森·赫维茨和麦多（1953）证明，当 B/σ 小于 0.5 时，对大小不同的绝对误差$|\hat{\mu}-\mu|$，两个分布几乎相同，甚至当 $B/\sigma=0.6$ 时，与 $B/\sigma=0$ 时相比，概率的变化也很小。

四、抽样误差

通过上面的分析，在抽样过程中，即使完全排除了偏差的影响，以样本统计量（如平均数、总体总量）估计总体参数，不可避免地还会产生误差，这种由于只测样本单元而没有观测全部总体单元而产生的误差称抽样误差。

抽样的总误差与偏差和抽样误差三者的关系可表述如下：

$$(\text{总误差})^2=(\text{抽样误差})^2+(\text{偏差})^2 \tag{1-2}$$

要减少抽样估计的总误差，就必须同时考虑减少抽样误差和偏差这两个方面。

在无偏估计情况下：总误差＝抽样误差。

抽样误差可以用抽样精度来表示。

为了说明上面两种情况，弄清准确度（accuracy）和精度（precision）这两个既相互联系又有区别的概念是有益的。准确度是指当有偏差干扰时，使样本平均数偏离总体平均数的概念。偏差越大准确度越小(见图 1-4)；精度是指各样本平均数以m为中心分布的变动程度，m是反复使用同一抽样方法所获得的平均数。

由此可见，在抽样估计中，有时精度高但准确度低，不能认为是有效的估计结果。只有当精度高、无偏或精度高、准确度也很高的情况下，才能认为是有效的估计结果。

抽样误差是由抽样方法本身引起的误差。本书旨在介绍和讨论各种抽样方法，因而对各种方法所导致的抽样误差给予分别论述。下面用简单随机、重复抽样为例，就抽样误差的性质和理论计算方法作几点分析。

在数理统计中，把样本统计量（如平均数）的标准差称之为标准误。抽样误差是通过标准误的估计实现的，因此可以说标准误就是抽样误差的计量尺度。

总体标准误：

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{1-3}$$

用样本估计：

$$S_{\bar{x}}=\frac{S}{\sqrt{n}} \tag{1-4}$$

式中：$\bar{x}$ 和 S 分别是总体$\bar{X}$和 σ 的无偏估计值。

1．通常所说的抽样误差（标准误)是平均抽样误差　这是因为从含N个单元的总体中，随机地抽取 n 个样本单元，全部可能抽取的样本数为 C_N^n 的组合数。然而当N很大时，不可能列出所有各样本平均数 $\bar{x}$ 与$\bar{X}$的实际抽样误差，因而需要从理论上认识各$\bar{x}$与$\bar{X}$误差的平均水平，即以样本抽样误差来描述各样本平均数与总体平均数的实际抽样误差的平均状况。标准误 $\sigma_{\bar{x}}$ 与 $S_{\bar{x}}$ 就描述了全部可能样本的 $\bar{x}$ 与$\bar{X}$的平均误差。

2．标准误是一个确定的数值　尽管每个样本都有它的$|\bar{x}-\bar{X}|$实际抽样误差，但对一个总体来说，每次抽取 n 个，样本个数总是有限多个，因而对这些样本平均数与$\bar{X}$的离差可以求得平均值。根据抽样方差定义有：

$$\sigma_{\bar{x}}^2=E(\bar{x}-\bar{X})^2 \tag{1-5}$$

$$\sigma_{\bar{x}}=\sqrt{\frac{1}{\Omega}\sum^{\Omega}(\bar{x}-\bar{X})^2} \tag{1-6}$$

式中：Ω 为总体中全部可能抽取的样本个数。

$\sigma_{\bar{x}}$ 值是唯一的，但在实际抽样估计中，人们通常是用样本的 S^2 代替 σ^2 来估计抽样误差。由于样本统计量 S^2 是随机变量，故用它计算出的抽样误差 $\sigma_{\bar{x}}$ 的估计值 $S_{\bar{x}}$ 也仍是随机变量。

3．$\sigma_{\bar{x}}$ (1-6 式）是抽样误差的理论公式　在实际中难以应用，这是因为使用 (1-6)式，必须已知 $\bar{X}$。此外，还必须计算出 Ω 个数以及每个样本的实际抽样误差，这都是在通常情况下不可能的。因而，在实际应用中，只能用数据统计推导出的误差公式去估计。