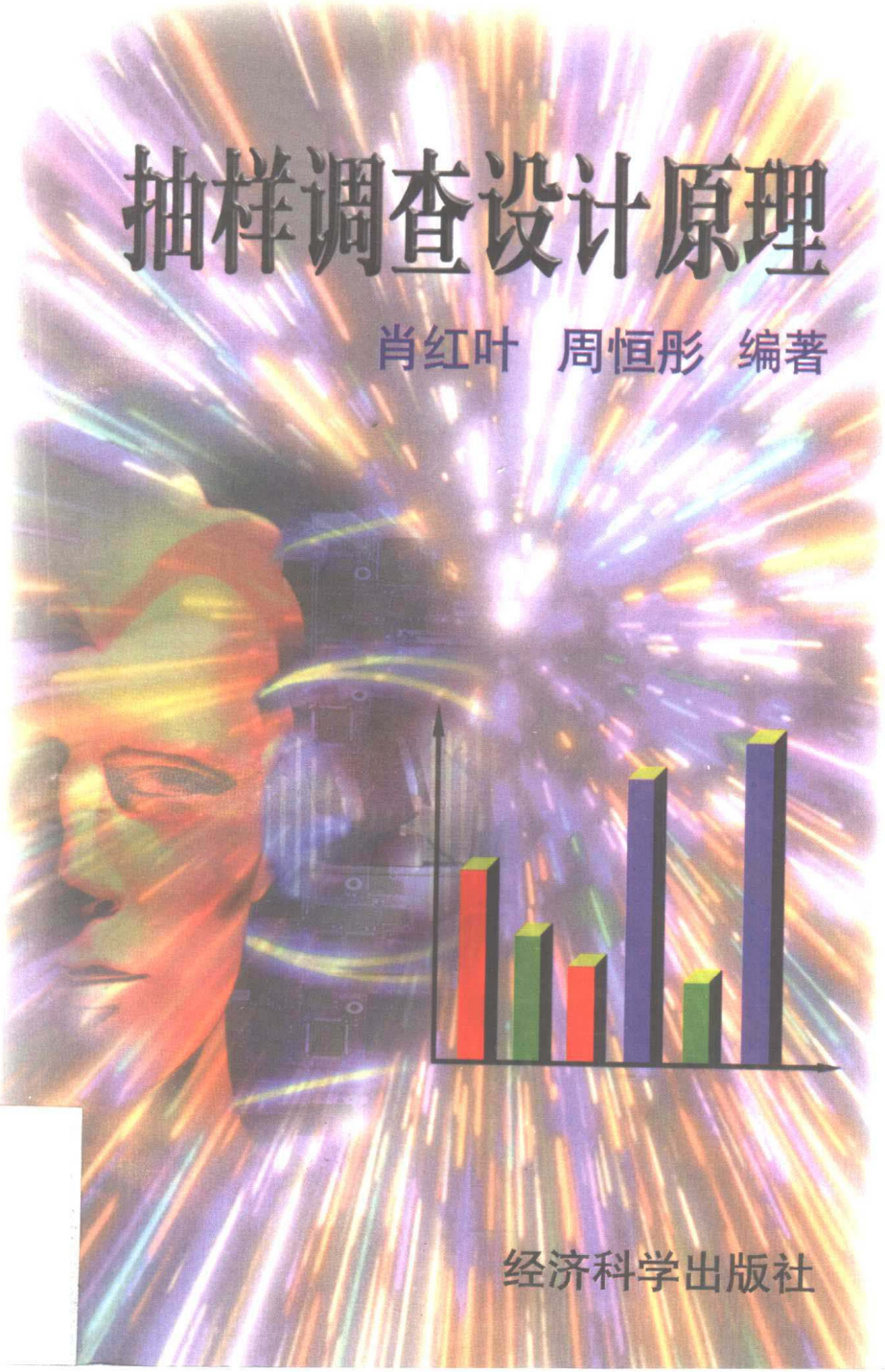
抽样调查设计原理

肖红叶　周恒彤　编著



经济科学出版社

抽样调查设计原理

肖红叶　周恒彤　编著

经济科学出版社

1997年·北京

责任编辑：高续增
责任校对：段健瑛
封面设计：张卫红
版式设计：代小卫
技术编辑：刘　军

抽样调查设计原理

肖红叶　周恒彤　编著

*

经济科学出版社出版、发行　新华书店经销
北京第二外国语学院印刷厂印刷
出版社电话：62541886　发行部电话：62568479
经济科学出版社暨发行部地址　北京海淀区万泉河路 66 号
邮编：100086

*

787×1092 毫米　32 开　8.25 印张　200000 字
1997 年 9 月第一版　1997 年 9 月第一次印刷
印数：0001－3000 册
ISBN 7-5058-1199-1/F・859　定价：10.20 元

前　言

人们在抽样调查实践中深切地感到，面对千差万别的调查任务，要想从调查任务和调查对象的特点以及能够获取的各种有关资料出发，作出科学合理的抽样设计，使其既节省费用，又取得令人满意、可信的估计效果，是一件很不容易的事情。因此，抽样设计问题的研究无疑是十分有意义的。本书试图在这方面提出一点考虑问题的思路。在本书的第二部分，按照思考问题的逻辑顺序，讨论了怎样根据调查任务确认总体、总体基本单位以及怎样建立抽样框（第 6 章）；当调查任务要求不但要估计整个总体，而且要估计总体的各个部分时应该怎么办（第 7 章）；怎样根据对象的特点恰当地选择抽样单位（第 8 章）；怎样尽可能地把辅助资料利用起来；以便提高估计的精度（第 9 章）；以及怎样确定样本量（第 10 章）等五个问题。

考虑到不甚熟悉抽样方法的读者的需要，在本书的第一部分，从抽取单位的几种组织方式（第 2 章）、几种常见的估计量（第 3 章）、不同类型抽样单位的样本（第 4 章），以及总体分层情形下的抽样（第 5 章）等四个不同的角度叙述了各种常用的抽样及估计方法。这一部分可看作第二部分的基础或预备。

应该说，本书所涉及到的仅仅是抽样调查设计中的几个问题而远不是全部问题。把它们从设计的角度集中起来讨论，

这是一种尝试。希望得到同行的各位方家指点，以共同推动抽样调查理论的发展。

有限总体概率抽样理论是数理统计学的一个分支，它有自身的严密的数理逻辑体系。因此，本书给出了大部分结论的证明。但是，考虑到读者的不同需要，本书一律用小号字排印证明过程或是证明性质的文字。假若略过这些内容不谈，不影响叙述的连贯性。

本书主要读者对象是从事抽样调查的方法与应用研究的教师及实际工作者，也可作为本科生和研究生学习抽样调查课程的参考书或教材。

作者感谢经济科学出版社力克学术著作出版工作中的种种困难，对本书的顺利出版和迅速问世所给予的鼎力支持。对这种扶持学术的义举，作者表示十分钦佩并表示深切地感谢。

本书第3、4、5、6、8、9章由周恒彤执笔，第1、2、7、10章由肖红叶执笔。

肖红叶　周恒彤

1997年4月于天津财经学院

统计学系

目　录

第一部分　抽样和估计方法

第 1 章　抽样方法和估计量的一般问题

§1.1　有限总体概率抽样

一、无限总体和有限总体

对无限总体这一术语有不同的用法。这里所说的无限总体，特指在一组不变的条件下重复进行的随机试验定义的总体。把这一随机试验无限重复地进行下去，那么，各次试验结果的集合就是由该随机试验定义的一个无限总体。

例 1.1　在一组不变的条件下加工某种产品。把这一加工过程无限重复地进行下去，所得到的无限的产品集合就是由加工产品的随机试验定义的无限总体。

有限总体是由一定的统计调查任务所规定的，在时间、空间以及若干其他标志上具有共同性质的全体（有限个）客观存在的单位（个体）结合起来的整体。

有限总体总是可以看作适当的无限总体的片断。

例 1.2　在例 1.1 中，把产品加工随机试验在某一报告期内的全部结果作为研究对象。

例 1.3　某时点上的全国人口这一有限总体。它可以被

想象作一个“大的”发生人口的随机试验的一部分结果。它是这个无限重复的随机试验的结果中，在该时点上尚存活着的，并且在空间上恰好聚集在该国的那一部分。

二、无限总体和有限总体的不同研究方法

研究无限总体的方法是统计推断。即，根据对总体的有限次观察结果（样本）去推断这个总体。

研究有限总体的方法是全面调查。不过，对有限总体，有时我们也希望只调查总体的一部分单位去推断总体，藉以完成本该由全面调查去完成的任务。

推断总体所用的样本必须是概率样本。所谓概率样本是要求对总体每进行一次观察（从总体每抽取一个单位），都应该是一次随机试验，并且它们都应该和被观察的总体具有相同的分布。

无限总体本来就是由随机试验来定义的。因而，对总体的任意一些观察，都可组成该总体的概率样本。有限总体是研究人员根据研究任务“聚拢”在一起的一些单位。必须要对从中抽取单位的行为作专门的设计，使之成为观察该有限总体的随机试验，才能获得该有限总体的概率样本。

前面说过，有限总体可以看作适当的无限总体的一个片断。现在进一步说，可以把它看作那个无限总体的概率样本。从有限总体中抽取一些单位，当然也可以把它们看作无限总体的概率样本。但是，对于有限总体而言，如前所说，这些单位必须是经过专门的概率抽样的设计抽取出来的，才能是有限总体的概率样本。

本书所讨论的是有限总体的概率抽样。

三、有限总体概率抽样

如果从总体[①] 中抽取一个单位时，总体中的每个单位都有可能被抽到，并且每个单位都有确定的、不等于零的被抽中的概率，这样的抽样叫作对有限总体的概率抽样。

为满足概率抽样的要求，可以用随机数表具体组织实施。实施的方法在以后的章节中介绍。

有限总体概率抽样有下列性质：当把一个特定的概率抽样方案应用于一个具体的总体时，我们能够确定地列举出能够被抽取到的各个不同的样本(当然，在操作中并不要求把它们一一列举出来)，并且，每个样本都有一个确定的被抽到的概率。因此，有限总体概率抽样也可以理解为：在能够被抽到的各个不同的样本中，依确定的概率抽出其中的一个样本。

四、非概率抽样的若干作法

实践中，人们常常应用某些非概率抽样的作法。由于这些方法也能够得到某种有用的结果，因而在某些场合并不排斥它们的应用。但是，这些方法不满足概率抽样的定义，因而不能应用概率抽样理论，不能用样本推断总体。通常，不要求确切地估计总体时可用非概率抽样；要求确切地估计总体时则要用概率抽样。

不要把非概率样本误当作概率样本去推断总体。为此，下面列举非概率抽样的若干作法。

1. 判断选样

判断选样又叫目的选样。调查者根据对总体的大致了解，选出一小部分“有代表性”的单位作调查。它们可能是标志值较“适中”的单位，也可能是规模特别大的单位，等等。

---

① 今后若无特别说明，“总体”一词都是指有限总体。

2. 方便选样

用某种方便的办法从总体中选样。例如，从敞开的货车的表层取一个煤的样本，在马路上随便找一些行人作调查，对总体中自愿接受调查的人员作调查，等等。选样时一般没有调查人员的主观判断。因此，很容易和概率抽样混淆。二者的区别是：概率抽样在抽取单位时精心设计了随机试验，而方便选样时则没有。

3. 定额选样

把总体分成若干类型组，从每组中按一定比例抽取单位。抽取单位用方便选样的方式。

§1.2 总体和样本

一、总 体

（一）总体的定义

总体（有限总体）的定义已如前节所述。但是，当我们对有限总体实行概率抽样时，可以用另外的观点来给它下定义。在概率抽样时，随机抽取的全体可能结果恰好是总体中每个单位某一个标志（或某几个标志）的标志值，因而此时可以把总体同标志值的集合等同起来，看作随机变量。其分布由于抽样采用等概率抽样或是不等概率抽样而有所不同。

（二）单 位

依前节对有限总体的定义，这种总体是由单位组成的。单位可以被划分成不同的“等级”：承担调查标志的单位叫总体基本单位，或称个体；由总体基本单位结合成的一个个集团叫作群单位，或称群体。

可以对不同等级的单位实施抽样操作。总体中被实施抽

样操作的那一级单位叫抽样单位，或称抽样单元。

在抽样实践中，正确确认总体基本单位非常重要。因为，以总体基本单位为抽样单位时叫个体抽样，以群单位为抽样单位时叫整群抽样，在估计总体时它们有不同的计算公式。

（三）抽样框

在抽样实践中，要靠抽样框来展示总体，以便进行抽样作业。

抽样框是总体抽样单位的一份完整的名单。它可能是总体中一些自然单位的名单（如：住户名单，村庄名单）；也可能是经调查人员划分得到的地理区域名单（如：把一片森林划分为若干区域单位后所建立的名单）；也可能是时间表名单（如：把一个月的时间以小时为单位，形成一个名单）。

（四）目标总体和被抽样总体

我们把作为调查对象的被研究总体称作目标总体。把抽样框所展示的总体称作被抽样总体或作业总体。由于抽样框有可能遗漏目标总体的一些单位，也可能包含一些并不属于目标总体的单位，因此，被抽样总体同目标总体往往并不完全一致。这时应当注意，用样本所作的估计推断，只能说明被抽样总体。至于能否用它也去说明目标总体，这有待于进一步研究其他辅助资料所提供的信息。就遗漏目标总体单位的情形而论，只有当辅助资料提供了下面的信息时，才能把样本的估计结论用于目标总体。即，辅助资料应当能够表明，被遗漏掉的单位内部变异较大，而它们同被抽样总体之间的变异却不大，也就是说，被遗漏掉的那些单位的均值同被抽样总体的均值大体相等。

（五）几种常见的总体指标

设总体有 N 个基本单位，标志值为 y_1，y_2，…，y_N。对

几种常见的总体指标作如下定义：

1. 均值 $\overline{Y}$

$$\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}y_i \tag{1.1}$$

2. 总值 Y

$$Y=\sum_{i=1}^{N}y_i=N\overline{Y} \tag{1.2}$$

3. 比例 P

把 N 个基本单位划分为 C 和 C' 两类，属于 C 类的单位数目为 A，属于 C' 类的单位数目为 A'，$A+A'=N$。定义

$$P=\frac{A}{N} \tag{1.3}$$

4. 比率 R

设总体的 N 个单位各自有成对的标志值构成比率：$r_1=y_1/x_1$，$r_2=y_2/x_2$，……，$r_N=y_N/x_N$。定义

$$R=\frac{Y}{X}=\frac{\overline{Y}}{\overline{X}} \tag{1.4}$$

式中，$\overline{X}$、X 的定义分别与（1.1）式、（1.2）式相同。

5. 方差 σ^2 和 S^2

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(y_i-\overline{Y})^2 \tag{1.5}$$

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{Y})^2 \tag{1.6}$$

显然有 $\sigma^2=\frac{N-1}{N}S^2$ 和 $S^2=\frac{N}{N-1}\sigma^2$

二、样　本

（一）样本的意义

从总体中每次抽取一个抽样单位，抽取 n 次，所抽出的 n 个抽样单位组成容量为 n 的样本。其中的每个单位叫样本

单位，n 叫作样本单位数、样本量或样本容量。在有的文献中把样本中所包含的个体数目叫作样本量。我们约定，在本书中不这样使用“样本量”这个术语。当以个体作抽样单位的时候，样本量和样本中包含的个体数目是一致的；当以群为抽样单位的时候，在本书中，样本量是指抽出的群数（或抽取群的次数），而不是指样本中包含的个体数目。

我们所讨论的是概率抽样。这时，从总体中抽取一个抽样单位是一次随机试验，应当用随机变量来描述。因此，容量为 n 的样本是由 n 个随机变量组成的一个随机变量序列。

（二）样本统计量

样本统计量是样本的函数。假设样本为 $(\xi_1, \xi_2, \cdots\cdots, \xi_n)$，那末对任意函数 f，$T=f(\xi_1, \xi_2, \cdots\cdots, \xi_n)$ 就是统计量，而且不同的函数关系表示不同的统计量。例如，$T=\frac{1}{n}\sum_{i=1}^{n}\xi_i$，$T_*=f(\xi_1, \xi_2, \cdots\cdots, \xi_n)=\min_i \xi_i$，等等都是样本统计量。

样本统计量是由随机变量构造成的，因而它也是随机变量。

研究样本统计量的分布、数学期望和方差，是抽样理论的核心问题之一。为此，须认真弄清这几个概念。

例 1.8 总体有 $N=4$ 个单位，标志值为 $y_1=2$，$y_2=4$，$y_3=6$，$y_4=8$。从中用简单随机不放回抽样方式（见第 2 章）抽取 $n=2$ 个单位。试作出样本均值 $\bar{\xi}=(\xi_1+\xi_2)/2$ 的分布列，并计算其数学期望和方差。

从 4 个单位中无放还地抽 2 个单位，不计它们在样本中先后顺序的不同，共可产生 $C_4^2=4(4-1)/2=6$ 种不同的观

察结果。样本均值的分布，就是每一种观察结果下所计算的样本均值同该种观察结果出现的概率（在第 2 章将证明，这些概率都是 $1/C_N^n$）结合在一起所建立的分布列。见表 1.1。

表 1.1　　样本均值的分布

样本观察结果编号 k	样本观察结果 (ξ_{k1},ξ_{k2})	样本平均数 $\bar{\xi}_k=(\xi_{k1}+\xi_{k2})/2$	各观察结果出现的概率 Π_k	$\Pi_k\bar{\xi}_k$	$\Pi_k\bar{\xi}_k^2$
1	(2,4)	3	1/6	3/6	9/6
2	(2,6)	4	1/6	4/6	16/6
3	(2,8)	5	1/6	5/6	25/6
4	(4,6)	5	1/6	5/6	25/6
5	(4,8)	6	1/6	6/6	36/6
6	(6,8)	7	1/6	7/6	49/6
合计	——	——	——	30/6	160/6

表 1.1 中间两栏是 $\bar{\xi}$ 的分布列。根据分布列计算 $\bar{\xi}$ 的数学期望和方差

$$E(\bar{\xi})=\sum_{k=1}^{6}\bar{\xi}_k\Pi_k=\frac{30}{6}=5$$

$$V(\bar{\xi})=\sum_{k=1}^{6}[\bar{\xi}_k-E(\bar{\xi})]^2\Pi_k$$

$$=\sum_{k=1}^{6}\bar{\xi}_k^2\Pi_k-(\sum_{k=1}^{6}\bar{\xi}_k\Pi_k)^2$$

$$=\frac{160}{6}-(\frac{30}{6})^2=\frac{10}{6}$$

在实际问题中，抽样得到的只是样本观察结果之中的某一种。因此，样本均值的分布列并不列示出来，样本均值的数学期望和方差也不是象上面那样用定义计算。

§1.3 估　计

一、估计的概念

从总体中用某种概率抽样方法抽取样本，构造适当的样本统计量，用这个样本统计量作为某个总体指标的估计量。对所抽得的具体样本作现场调查，取得数据，依据估计量的计算规则算出数值，作为总体指标的估计值。这整个的过程叫作估计。有时不去严格区分估计、估计量、估计值这几个术语，而笼统地说估计。

一个估计量其实就是某一个样本统计量。所以，它也是随机变量。它的分布、数学期望、方差与相应的样本统计量相同。

二、对估计结果的评价

由于样本的随机性，估计的结果也是随机的。一个具体的估计值是随机变量的一个观察值。因此，对于估计的结果，不是就一个具体的估计值，而是在所有可能产生的样本观察结果平均的意义上来进行评价。通常从三个方面来考虑：估计结果的偏斜情况；抽样方案的效果；抽样调查的效果。

（一）估计结果的偏斜情况

估计结果是否偏斜，是看一个抽样方案所有可能产生的估计值的平均值是否等于被估计的总体指标。设总体指标为 θ，它的估计量为 $\hat{\theta}$，建立

$$B=E(\hat{\theta})-\theta \tag{1.7}$$

B 叫作偏差。$B=0$ 时，称 $\hat{\theta}$ 为 θ 的无偏估计量，$|B|>0$ 时，称 $\hat{\theta}$ 为 θ 的有偏估计量。

（二）抽样方案的效果

抽样方案主要是指：所用的抽样组织方式、样本容量的大小、所用的估计量。考察一个抽样方案的效果好坏，是看这个抽样方案所有可能产生的估计值的变异状况（离散状况）。显然，一个抽样方案所有可能产生的估计值较为均匀，则在实际抽样中不论得到其中的哪一个都差不多，这样的抽样方案效果是比较好的。我们知道，描述随机变量离散特征最常用的特征数是方差。现在同样用估计量的方差

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 \qquad (1.8)$$

来描述抽样方案的效果：方差越小的方案效果越好。

为更直观地说明抽样方案的效果，引入效率（或精度）的概念。我们定义估计量方差的倒数为抽样方案的效率（或精度）。所以，二方案效率比（或精度比）等于二方案方差比的倒数。

$V(\hat{\theta})$ 有时又写作 $\sigma_{\hat{\theta}}^2$。它的平方根 $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$ 称作 $\hat{\theta}$ 的标准误。标准误和方差有相同的作用。

如果能作到，一方面，估计量无偏，另一方面，它的方差也小，这当然十分理想。但有时作不到如此理想。这时，我们往往宁愿选择一个虽然估计量有偏（这个偏差是我们能够接受的）但方差较小的方案，而不选择估计量无偏但方差却较大的方案。

（三）抽样调查的效果

考察一次抽样调查的效果好坏（或说准确度高低），是看若在所用抽样方案下反复抽样反复调查，所有可能产生的估计值与被估计的总体指标之间的误差的平均水平。用均方误差（MSE）来表现

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \qquad (1.9)$$

均方误差越小，表明调查效果越好。均方误差的平方根叫均