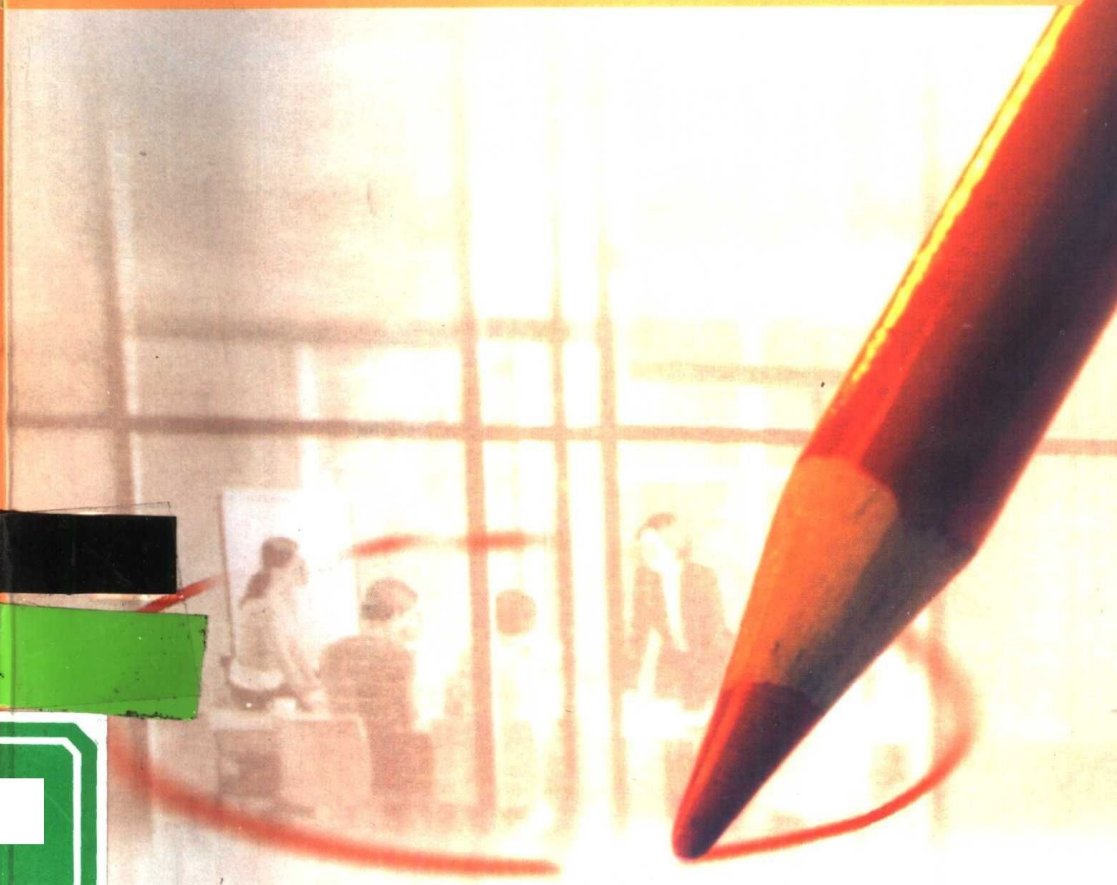
国家“九五”社科基金重点研究项目

抽样调查的理论方法和应用

胡健颖　孙山泽　主编



北京大学出版社

抽样调查的理论、方法和应用

主　编　胡健颖
　　孙山泽
副主编　雷　明

北京大学出版社
北　京

图书在版编目(CIP)数据

抽样调查的理论、方法和应用/胡健颖，孙山泽主编．北京：北京大学出版社，2000．6

ISBN 7-301-04547-6

Ⅰ．抽… Ⅱ．①胡… ②孙… Ⅲ．社会经济统计-抽样调查 Ⅵ．C811

中国版本图书馆 CIP 数据核字(2000)第 09589 号

书　　名：抽样调查的理论、方法和应用

著作责任者：胡健颖　孙山泽

责任编辑：符　丹　刘灵群

标准书号：ISBN 7-301-04547-6/F．0341

出 版 者：北京大学出版社

地　　址：北京市海淀区中关村北京大学校内　100871

网　　址：http://cbs.pku.edu.cn/cbs.htm

电　　话：出版部 62752015　发行部 62754140　编辑部 62752027

电子信箱：zpup@pku.pku.edu.cn

排 版 者：兴盛达激光照排中心

印 刷 者：

发 行 者：北京大学出版社

经 销 者：新华书店

850 毫米×1168 毫米　32 开本　9.125 印张　220 千字

2000 年 6 月第 1 版　2000 年 6 月第 1 次印刷

定　　价：16.00 元

前　言

本书是“九五”国家社会科学基金重点研究项目“抽样技术在社会经济统计调查中应用问题的研究”的最终研究成果。项目主持人为北京大学光华管理学院胡健颖教授和北京大学概率统计系孙山泽教授，协助主持项目研究工作的有：北京大学光华管理学院雷明副教授和崔兆鸣副教授，北京大学概率统计系杨宝慧讲师和孙明举硕士研究生。相应地，本书主编由胡健颖和孙山泽担任，副主编由雷明担任。参加本项目调查研究的还有北京大学光华管理学院、北京大学概率统计系的一些研究生。

本书除前言和后记外，共分为三部分十三章。第一部分“社会经济统计调查抽样方法基础”（第1—6章）；第二部分“社会经济统计抽样调查中一些特殊问题研究”（第7—9章）；第三部分“抽样技术在社会经济统计调查中的案例分析”（第10—13章）。

本书的主要观点概述如下：

1. 社会主义市场经济条件下，在社会经济统计调查中，需大量运用抽样技术，但如何科学地应用抽样技术，不仅具有重要的理论意义，而且具有重大的现实意义。要弄清楚抽样技术在社会经济统计调查中的应用问题，首先要科学的理解抽样调查的特点以及常用的几种抽样调查方法的适用范围。

2. 在中国统计调查中，采用抽样技术虽然起步较早，而且也取得了有益的经验和较大的成绩，但仍存在一些问题。诸如，目前绝大多数社会经济调查中采取忽略不响应样本的处理方法，从而

使调查结果往往产生偏差。其原因在于，响应调查的样本和不响应调查的样本两个群体之间常存在差异，如要获得全面正确的调查结果，必需探求不响应样本的状况，用概率统计方法作出正确的分析。国外许多先进国家均非常重视研究处理不响应样本的方法，并设有专题研究小组和大量的这类调查文献发表。国内概率统计学界也有一些人对这一问题从数理统计的理论方面作过一些研究，但与社会经济调查结合的研究极少见到，甚至出现抽样调查中错误地处理不响应样本，导致决策失误。为此，本书着重就下列问题进行了研究：

第一，在社会经济抽样调查中，当个体不响应或个体对某些项目不响应，亦或有意、无意错误响应时，应如何进行处理，以免信息失真；

第二，在社会经济抽样调查中，遇到敏感性问题要求回答，应如何设计调查方案？如何结合中国现实生活中存在的案例，探讨新的抽样调查方法和统计分析方法，以便消除被调查者的疑虑，获得真实的信息；

第三，如何结合中国实际情况，借鉴国外抽样调查的有益经验，探讨用随机选择回答，设立相关问题，准许不回答等科学方法，来提高中国抽样调查资料的实用性，为决策部门提供可靠信息。

3. 现实中抽样调查应用的领域极其广泛，如人口抽样调查、农村抽样调查、城市住户抽样调查等等。在这多种领域并存的情况下，不仅每个领域内部由于受内外部影响因素的变化而变化，从而要求抽样调查的方法有所创新，而且随着社会经济发展人们发现过去被忽视的领域被提上了重要的研究历程。例如，环境抽样调查问题已成为当前全世界统计界面临的新课题。可以说，有关环境统计调查和环境指标体系设立的研究和实务工作，尤其是就环境问题开展统计抽样调查还处于初级阶段，要使其完善并与现有国民经济统计体系融通尚需要时日。本书对环境生态统计调查，特别是

环境生态抽样调查，结合中国实际进行了初步论述。

在此需要指出的，本书提出的一些新观点，由于作者水平有限，不足之处，我们殷切期待有关专家及广大读者的指正。

作　者

1999 年 12 月于北京大学

目　录

第一章 引 言

一、抽样调查的特点与作用

抽样调查是一种非全面调查，是从调查对象的总体中随机抽取一部分单位进行观察，并依据所获得的数据对总体的数量特征得出具有一定可靠性的估计判断，从而达到对总体的认识。由于抽样调查是针对总体中的一部分单位进行的，所以，与全面调查相比它具有费用低、速度快的特点，特别是对于资料信息的时效性很强的现象进行调查时，这一优点尤为重要。另外，抽样调查能够处理全面调查所无法解决或很难解决的问题，如罐头食品的质量检验、水库中的鱼苗数、森林区的木材蓄积量的调查，以及在社会经济抽样调查中，当个体不响应或个体对某些项目不响应，或有意、无意错误响应时，如何进行调查，等等，这些只能采取抽样调查的方法来推断其总体特征。再有，抽样调查还有可能取得比全面调查更为准确的结果，这一方面是由于在工作量减少以后，可以对调查人员进行更严格、更细致的训练以提高其素质，同时，在抽样调查中还可使调查工作受到更严谨的监督和控制，从而使获得的数据在一定条件下可比全面调查所获得的相应数据更为准确。鉴于上述特点，抽样调查方法在实际工作中，尤其是在市场经济的条件下得到愈来愈广泛的研究和应用。

二、总体与样本

总体就是所要调查研究的全体，如要研究某城市职工的生活水平，则该市全部职工就构成总体。

总体又有被抽样总体或作业总体与目标总体之别，被抽样总体即从中进行抽样的总体，是抽样取本的依据。目标总体就是要从中得到信息对之进行说明的总体。被抽样总体应与目标总体一致，有时为了实用与方便，被抽样总体在范围上比目标总体要受到较多的限制，若这样的话，则从样本中得出的结论只适用于被抽样总体。

在抽样之前，总体必须划分成称为抽样单位的各部分，这些单位必须互不重叠并且能合成总体，也就是说，总体中的每个个体属于且只属于一个单位，比如，在农作物的抽样中，单位可以是一块田、一个农场或是形状，大小都由我们决定的一片土地。编制的抽样单位的名单称为抽样框。

样本就是从被抽样总体中抽取的一部分单位，样本又称子样。样本是总体的缩影是总体的代表，我们正是依据样本的调查结果来推断总体的特征的，样本作为总体的子集所含有的单位数称为样本容量。

从总体中可能抽取的全部样本的数目称为可能样本数目。可能样本个数的多少不但与样本容量的大小有关，而且也与抽取样本的方法有关。抽取样本的方法大致可分为两种：一种是概率抽样，另一种是非概率抽样。概率抽样也称随机抽样，就是在依据一种抽样方法所形成的所有可能的样本中，每一个样本被抽中的机会都等于某一与自己相对应的概率值，所有样本的概率之和为 1。不是概率抽样的样本抽取方法就是非概率抽样。一种常用的非概率抽样是指所谓的判断抽样，或称经验抽样。这种抽样是根据抽样

者的主观经验和判断，从总体中选择认为有代表性的同时又容易取得的个体作为样本单位。

三、抽 样 误 差

抽样调查中的误差来源主要有两个。一种是非抽样误差也称调查误差，它是调查过程中由于观察、测量、登记上的差错以及被调查者不真实回答等原因使在调查中获得的原始数据不准确而引起的误差。这种误差非抽样调查所特有，而是所有统计调查都有可能存在。这种非抽样误差的减少，只能是通过改进调查表的设计或加强组织管理等手段才能予以实现。比如，对于不易获得被调查者真实情况的诸如敏感性问题的调查必须通过设计特殊的调查方法进行处理。

抽样调查中的另一种误差是用样本数据对总体参数作出估计所引起的误差，是由抽样方法本身所引起的误差，这种误差称为抽样误差。本书中主要考虑这种误差。

我们用估计量这个词表示根据样本结果来计算某个总体参数的估计值的规则或公式，用估计值这个词表示依据一个具体的样本所估算得的该总体参数的数值。设总体参数为 θ，$\hat{\theta}$ 为它的估计量，则抽样误差一般用以下的均方误差来表示：

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

均方误差又可进一步改写成：

$$\begin{aligned}MSE(\hat{\theta})&=E(\hat{\theta}-\theta)^2=E[\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta]^2\\&=E[(\hat{\theta}-E(\hat{\theta}))^2+2(\hat{\theta}-E(\hat{\theta}))E(\hat{\theta}-\theta)\\&\quad+(E(\hat{\theta})-\theta)^2]\\&=E[\hat{\theta}-E(\hat{\theta})]^2+[E(\hat{\theta})-\theta]^2\\&=V(\hat{\theta})+B^2(\hat{\theta})\end{aligned}$$

其中 $V(\hat{\theta})=E[\hat{\theta}-E(\hat{\theta})]^2$ 为 $\hat{\theta}$ 的方差，$B^2(\hat{\theta})=[E(\hat{\theta})-\hat{\theta}]^2$ 为 $\hat{\theta}$

的偏差$|E(\hat{\theta})-\theta|$的平方。若偏差为零，即$E(\hat{\theta})=\theta$，则$\hat{\theta}$称为θ的无偏估计量。对于无偏估计量，它的均方误差就是它的方差。

四、抽样调查问题的再提出

与发达国家相比，我国统计调查中采用抽样技术虽然起步较早，而且也取得了有益的经济和较大的成绩，例如对农产量、居民收支、科技投入、工业产品质量等各类的抽样调查。但仍存在一些问题，诸如，目前绝大多数社会经济调查中采取忽略不响应样本的处理方法，从而使调查结果往往产生偏差，其原因在于响应调查的样本和不响应调查的样本的两个群体之间常存在差异，如果获得全面的正确调查结果，必须探究不响应样本的状况，用概率统计方法做出正确的分析。国外许多先进国家均非常重视处理不响应样本的方法，设有研究专题小组，并有大量的这类调查文献发表。国内概率统计学界也有一些人对这一问题从数理统计的理论等方面做过一些研究，但与社会经济调查结合的研究极少见到，甚至出现抽样调查中错误地处理不响应样本，导致决策失误。

第二章　简单随机抽样

一、简单随机抽样(纯随机抽样)

设总体由 N 个样本单位组成,从其中抽取 n 个单位,使得 C_N^n 个不同的样本每一个被抽中的机会都相等,即每个样本被抽中的概率都为 $1/C_N^n$,这种抽样方法就是简单随机抽样。按简单随机抽样,抽到的样本称为简单随机样本。实际上,一个简单随机样本可以采取逐个样本单位不放回抽样得到,即从总体中的 N 个单位中逐个不放回地抽取单位,每次抽取到尚未在样本中的任何一个单位的机会都相等。采用这个办法,则所有的 C_N^n 个不同的样本都有相同的概率被抽中。为此让我们来看一下一个特定的样本,就是 n 个已确定的单位的一个集合。在第一次抽取时,抽出这 n 个确定的单位中某一个单位的概率是 $\frac{n}{N}$,第二次抽取时,抽中剩下的 $n-1$ 个单位中的某一个的概率是 $\frac{n-1}{N-1}$,依此下去,在 n 次抽取中,这 n 个确定的单位全部被抽中的概率是

$$\frac{n}{N}\cdot\frac{(n-1)}{N-1}\cdot\frac{(n-2)}{N-2}\cdots\frac{1}{N-n+1}=\frac{n!\ (N-n)!}{N!}=\frac{1}{C_N^n}$$

二、定义和有关符号

在抽样调查中,我们要对抽取的样本中的每个单位的某些特征进行测量和记录,这些被测量的单位的特征就称为标志,通常用

大写字母与小写字母来分别表示有关总体与样本的标志值。例如一含有 N 个单位的总体，其标志值可记为 $Y_1, Y_2, \cdots, Y_N$，而 $Y = \sum_{i=1}^{N} Y_i$ 及 $\overline{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ 分别表示总体总和及总体均值。

而一样本容量为 n 的样本，其中各个单位的标志值通常用 $y_1, y_2, \cdots, y_n$ 来表示。我们将

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n \text{ 及}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

分别称为样本和及样本均值，用符号 ^ 表示从一个样本所得到的总体标志的一个估计量。

抽样调查是为了推断总体的某些特征或性质，但总体的特征是多种多样的，而我们的兴趣大都集中于总体的以下四项标志。

1. 均值 $\overline{Y}$（例如平均每个居民小区的人数）
2. 总值 Y（例如一个地区内小麦的总产量）
3. 两个总值的比率或两个均值的比率

$$R = Y/X = \overline{Y}/\overline{X}$$

（例如一组家庭中食物支出与其收入之比）

4. 具有某种特征的单位所占的比例

例如一城市下岗人员所占的比例

在本章中，对简单随机抽样，对总体均值 $\overline{Y}$，总体总值 Y 分别采取如下的估计

$\hat{\overline{Y}} = \bar{y}$ 即总体均值估计量为样本均值

$\hat{Y} = N\bar{y} = \frac{N}{n} \sum_{i=1}^{n} y_i$ 即总值的估计量为总体总数乘以样本均值。

n/N 是样本含量与总体含量之比，称为抽样比，用字母 f 表示。

三、估计量的性质

定理 2.1 1. 样本均值 $\bar{y}$ 是 $\bar{Y}$ 的无偏估计量。

2. $\hat{Y}=N\bar{y}$ 是总体总值 Y 的无偏估计量。

证明

1. 在全部可能的 C_N^n 个简单随机样本中含有总体中每个单位的个数都相等，所以有 $E(y_1+y_2+\cdots+y_n)$ 一定是 $Y_1+Y_2+\cdots+Y_N$ 的倍数，根据求和中单位个数的计算，这个倍数恰是 n/N，所以有

$$E\bar{y}=\frac{1}{n}\cdot\frac{n}{N}\sum_{i=1}^{N}Y_i=\frac{1}{N}\sum_{i=1}^{N}Y_i=\bar{Y}$$

2. $E\hat{Y}=E(N\bar{y})=NE\bar{y}=N\bar{Y}=Y$

按一般的定义，有限总体的方差为

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

我们用另一个符号 S^2 来表示总体方差的形式稍加变动后的结果，即

$$S^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

这样做的目的就是为了使大多数结果有一个稍为简洁一些的表达式。

定理 2.2 对于简单随机抽样，$\bar{y}$ 的方差为

$$V(\bar{y})=\frac{S^2}{n}(1-f)$$

其中 $f=n/N$ 为抽样比。

证明

利用对称性可知

$$E[(y_1-\bar{Y})^2+\cdots+(y_n-\bar{Y})^2]=\frac{n}{N}[(Y_1-\bar{Y})^2+\cdots+(Y_N-\bar{Y})^2]$$

以及 $E[(y_1-\bar{Y})(y_2-\bar{Y})+(y_1-\bar{Y})(y_3-\bar{Y})+\cdots+(y_{n-1}-\bar{Y})(y_n-\bar{Y})]=\frac{n(n-1)}{N(N-1)}[(Y_1-\bar{Y})(Y_2-\bar{Y})+(Y_1-\bar{Y})(Y_3-\bar{Y})+\cdots+(Y_{N-1}-\bar{Y})(Y_N-\bar{Y})]$

又由于有 $n(\bar{y}-\bar{Y})=(y_1-\bar{Y})+(y_2-\bar{Y})+\cdots+(y_n-\bar{Y})$

可推出

$$\begin{aligned}
&n^2E(\bar{y}-\bar{Y})^2\\
&=\frac{n}{N}\{(Y_1-\bar{Y})^2+\cdots+(Y_N-\bar{Y})^2+\frac{2(n-1)}{N-1}\\
&\quad[(Y_1-\bar{Y})(Y_2-\bar{Y})]+\cdots+(Y_{N-1}-\bar{Y})(Y_N-\bar{Y})]\}\\
&=\frac{n}{N}\left\{\left(1-\frac{n-1}{N-1}\right)[(Y_1-\bar{Y})^2+\cdots+(Y_N-\bar{Y})^2]+\frac{n-1}{N-1}\right.\\
&\quad\left.[(Y_1-\bar{Y})+\cdots+(Y_N-\bar{Y})]^2\right\}\\
&=\frac{n(N-n)}{N(N-1)}\sum_{i=1}^{N}(Y_i-\bar{Y})^2
\end{aligned}$$

故 $V(\bar{y})=E(\bar{y}-\bar{Y})^2=\frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$

$$=\frac{S^2}{n}\cdot\frac{(N-n)}{N}=\frac{S^2}{n}(1-f)$$

作为总体总值的估计量 $\hat{Y}=N\bar{y}$ 的方差为

$$V(\hat{Y})=\frac{S^2}{n}N(N-n)=\frac{N^2S^2}{n}(1-f)$$

当从一个无限总体中抽取一个含量为 n 的随机样本或从一个有限总体再放回地抽取 n 个单位，我们知道其均值方差为 $\frac{\sigma^2}{n}$，当 N 很大时 $\frac{\sigma^2}{n}\approx\frac{S^2}{n}$，因此，从有限总体中抽得的简单随机样本均值的方差要比从无限总体中抽取的样本均值的方差小，两者相差 1