普通高等教育"十五"国家级规划教材
北京大学数学教学系列丛书

本科生
数学基础课教材

抽样调查

孙山泽　编著

北京大学出版社

北京大学数学教学系列丛书

抽 样 调 查

孙山泽 编著

北京大学出版社
·北 京·

图书在版编目(CIP)数据

抽样调查/孙山泽编著. —北京：北京大学出版社，2004.2
(普通高等教育"十五"国家级规划教材)
(北京市高等教育精品教材立项项目)
(北京大学数学教学系列丛书)
ISBN 7-301-06857-3

Ⅰ. 抽… Ⅱ. 孙… Ⅲ. 抽样调查-高等学校-教材 Ⅳ. C811

中国版本图书馆 CIP 数据核字(2003)第 119250 号

书　　名：抽样调查
著作责任者：孙山泽　编著
责任编辑：王国义　刘　勇
标准书号：ISBN 7-301-06857-3/O・0584
出版发行：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　电子信箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　理科编辑部 62752021
印 刷 者：北京大学印刷厂
经 销 者：新华书店
890 mm×1240 mm　A5　6.75 印张　188 千字
2004 年 2 月第 1 版　2004 年 2 月第 1 次印刷
印　　数：0001—4000 册
定　　价：13.50 元

普通高等教育“十五”国家级规划教材

北京市高等教育精品教材立项项目

内容简介

本书是高等院校概率统计系本科生“抽样调查课”的教材。主要讲述抽样调查的基本理论和方法。全书共分八章，内容包括引言、简单随机抽样、不等概抽样、分层抽样、多阶抽样、整群抽样与系统抽样、二相抽样以及抽样实践中常见的几个问题的讨论。本书沿袭许宝騄先生在《抽样论》(北京大学出版社，1982)中所用的处理方法，并扩充了实践内容，增加了具体案例。本书对一些最基本的调查方法理论作了统一处理，并吸收了国内外抽样调查前沿研究的理论和实践，使读者能在短时间内掌握抽样调查的基本方法。作者在编写本书时特别注意结合我国当前调查的实际经验和需求，给出许多调查实例，使读者参照这些实例便可设计出较好的调查方案。本书叙述简明清晰，定理证明浅显易懂，凡是具有高等代数、微积分和初等概率统计知识的读者都可以读懂本书内容。

本书可作为综合大学、理工科院校概率统计系、社会学系和经济管理系本科生教材，也可作为应用统计工作者设计抽样调查方案的参考书。

作者简介

孙山泽 北京大学数学科学学院教授，1962 年毕业于北京大学数学力学系，长期从事概率论与数理统计的教学、科研工作，主要研究方向是抽样调查与应用统计，参加过国家多项抽样调查与应用统计的研究项目。

序　言

自 1995 年以来，在姜伯驹院士的主持下，北京大学数学科学学院根据国际数学发展的要求和北京大学数学教育的实际，创造性地贯彻教育部“加强基础，淡化专业，因材施教，分流培养”的办学方针，全面发挥我院学科门类齐全和师资力量雄厚的综合优势，在培养模式的转变、教学计划的修订、教学内容与方法的革新，以及教材建设等方面进行了全方位、大力度的改革，取得了显著的成效。2001 年，北京大学数学科学学院的这项改革成果荣获全国教学成果特等奖，在国内外产生很大反响。

在本科教育改革方面，我们按照加强基础、淡化专业的要求，对教学各主要环节进行了调整，使数学科学学院的全体学生在数学分析、高等代数、几何学、计算机等主干基础课程上，接受学时充分、强度足够的严格训练；在对学生分流培养阶段，我们在课程内容上坚决贯彻“少而精”的原则，大力压缩后续课程中多年逐步形成的过窄、过深和过繁的教学内容，为新的培养方向、实践性教学环节，以及为培养学生的创新能力所进行的基础科研训练争取到了必要的学时和空间。这样既使学生打下宽广、坚实的基础，又充分照顾到每个人的不同特长、爱好和发展取向。与上述改革相适应，积极而慎重地进行教学计划的修订，适当压缩常微、复变、偏微、实变、微分几何、抽象代数、泛函分析等后续课程的周学时。并增加了数学模型和计算机的相关课程，使学生有更大的选课余地。

在研究生教育中，在注重专题课程的同时，我们制定了 30 多门研究生普选基础课程（其中数学系 18 门），重点拓宽学生的专业基础和加强学生对数学整体发展及最新进展的了解。

教材建设是教学成果的一个重要体现。与修订的教学计划相配合，我们进行了有组织的教材建设。计划自 1999 年起用 8 年的时间修订、编写和出版 40 余种教材。这就是将陆续呈现在大家面前的《北京大学数学教学系列丛书》。这套丛书凝聚了我们近十年在人才培养方面的思考，记录了我们教学实践的足迹，体现了我们教学改革的成果，反映了我们对新世纪人才培养的理念，代表了我们新时期的数学教学水平。

经过 20 世纪的空前发展，数学的基本理论更加深入和完善，而计算机技术的发展使得数学的应用更加直接和广泛，而且活跃于生产第一线，促进着技术和经济的发展，所有这些都正在改变着人们对数学的传统认识。同时也促使数学研究的方式发生巨大变化。作为整个科学技术基础的数学，正突破传统的范围而向人类一切知识领域渗透。作为一种文化，数学科学已成为推动人类文明进化、知识创新的重要因素，将更深刻地改变着客观现实的面貌和人们对世界的认识。数学素质已成为今天培养高层次创新人才的重要基础。数学的理论和应用的巨大发展必然引起数学教育的深刻变革。我们现在的改革还是初步的。教学改革无禁区，但要十分稳重和积极；人才培养无止境，既要遵循基本规律，更要不断创新。我们现在推出这套丛书，目的是向大家学习。让我们大家携起手来，为提高中国数学教育水平和建设世界一流数学强国而共同努力。

张继平

2002 年 5 月 18 日

于北京大学蓝旗营

前　言

调查是现代社会进程的一个重要组成部分，其中大量进行的是抽样调查。抽样调查是一门应用性很强的学科，在实施抽样时应注意强调以下几点：第一是要获得正确的信息，第二是要省时省力省钱，第三是在制定抽样方案时，调查的抽样方法要与后续的数据分析配套，第四是要充分利用已知信息来提高抽样精度。这些都是抽样工作的要点。本书力求贯彻这些要点。目前高等院校中的许多专业，如统计学、社会学、经济管理等都设置了抽样调查课程，但这一课程的适用教材甚为缺乏，迫切需要一本适用的教材和抽样调查指导手册。

1982 年北京大学出版社曾出版过我国已故统计学家许宝騄先生的《抽样论》，我们为了教学的需要，在许先生的《抽样论》的基础上扩充内容编写了抽样调查讲义，在北京大学概率统计系本科生的教学中多次作为教材使用。本书将该讲义又扩充了实践内容，增加了具体的案例。本书沿袭许宝騄先生的《抽样论》中的处理方法，对一些最基本的调查方法理论作了统一处理，并吸取了国内外抽样调查前沿研究的理论和实践，使读者能在短时间内掌握最新的抽样调查理论，为读者对抽样理论进一步的学习和研究打下基础。本书在编写时特别注意结合我国当前调查的实际经验和需求，给出许多调查示例，使读者即使不阅读该书中各个定理的证明过程，也可以参照书中实例设计出精确度高、省时省力又节省经费的调查方案。

作为高等学校的抽样理论教材，通过本书学生可以掌握基本

的抽样方法，能够理解各种抽样方法的数理统计原理和获得实施具体项目的处理能力和分析能力。对于从事调查工作的实际工作者也是一本很好的参考书。调查者根据要实施的实际调查情况，只要对照书中实例即可以方便地制定出自己的抽样策略。

本书适用于高等学校的学生作为教材，同时也为具体调查方案的实际设计者提供抽样策略。书中内容简明清晰，定理的证明浅显易懂。凡具有一般的高等代数、微积分和初等概率统计知识的读者都可以读懂本书内容。

本书承蒙北京大学概率统计系郑忠国教授审阅了全书，北京大学出版社刘勇同志积极推动了本书的出版，在此一并表示衷心的感谢。

书中疏漏和不当之处，敬请读者批评指正。

孙山泽

2004 年 2 月

目　录

第一章　引言 …… (1)
§1.1　大规模抽样调查 …… (1)
§1.2　有限总体抽样的样本分布 …… (5)
§1.3　概率抽样的几种基本的抽样方法 …… (8)
习题一 …… (11)

第二章　简单随机抽样 …… (13)
§2.1　简单随机抽样的几个基本定理 …… (13)
§2.2　简单随机抽样的实现 …… (16)
§2.3　简单估值法 …… (19)
§2.4　区间估计与样本量的确定 …… (26)
§2.5　比估计 …… (35)
§2.6　差估计与回归估计 …… (42)
习题二 …… (50)

第三章　不等概抽样 …… (53)
§3.1　PPS 抽样 …… (53)
§3.2　不等概 πPS 抽样 …… (60)
§3.3　Rao-Hartley-Cochran 随机分群抽样 …… (65)
习题三 …… (69)

第四章　分层抽样 …… (73)
§4.1　估值法(一) …… (74)
§4.2　估值法(二)——组合比估计和回归估计 …… (77)
§4.3　样本量的分配 …… (81)
§4.4　与简单随机抽样之比较 …… (88)

§4.5 如何适当分层…… (90)
§4.6 后分层估计和定额抽样…… (95)
习题四 …… (97)

第五章 多阶抽样…… (101)
§5.1 二阶抽样的提法 …… (101)
§5.2 二阶抽样之估值法 …… (103)
§5.3 二阶抽样的效率 …… (113)
习题五…… (116)

第六章 整群抽样与系统抽样…… (121)
§6.1 整群抽样 …… (121)
§6.2 群内相关系数 …… (123)
§6.3 系统抽样 …… (127)
§6.4 个体指标具有特殊结构时的系统抽样 …… (131)
§6.5 系统抽样估计量方差的估计 …… (136)
习题六…… (138)

第七章 二相抽样…… (142)
§7.1 为分层的二相抽样 …… (142)
§7.2 二相分层抽样的最优分配问题 …… (147)
§7.3 为 PPS 抽样的二相抽样 …… (151)
§7.4 为比估计的二相抽样 …… (156)
习题七…… (159)

第八章 抽样实践中常见的几个问题的讨论…… (163)
§8.1 定期连续抽样调查中使用历史数据的技术 …… (163)
§8.2 敏感性问题的调查方法 …… (177)
§8.3 不完善抽样框的处理 …… (189)

附录 随机数表…… (196)
参考文献…… (204)

第一章 引 言

§1.1 大规模抽样调查

本书所述抽样调查亦称**大规模抽样调查**. 在社会经济的诸多领域,如国家资源状况、人口现状、农业产量及虫害估计、疾病医疗等许多方面,需要新的信息时往往要进行大规模的抽样调查.

大规模抽样调查一般说来大致有三类,即**普查**、**概率抽样调查**和**典型调查**. 普查也就是对研究对象的全体进行全面调查,如我国进行的人口普查、工业企业普查等等. 普查需投入大量的人力、物力,且需较长的时间,调查的规模庞大,组织工作艰巨. 这就决定了这类调查不能频繁进行,像人口普查都为十年或五年进行一次. 普查可获得全面的资料,如人口普查,不但可以了解全国的状况,而且可以了解各省乃至县乡的状况. 概率抽样调查是在非全面调查中运用概率统计理论指导的抽样调查方法. 比起全面调查,概率抽样调查可以节约人力、物力,节省时间. 概率抽样调查要根据研究对象总体的一些已知情况,设计适宜的抽样方案,充分用好已知的辅助信息,获得有代表性的好样本,从而对总体的特征指标做出好的估计. 按概率统计原理设计的抽样还会对每一个特征指标的估计,给出一个估计的误差. 概率抽样调查是目前许多领域获取调查信息时推荐使用的抽样调查方法. 本书将在后面对常用的各种基本方法,介绍它们的概率统计原理、应用的计算公式及一些调查实例. 典型调查是一种完全依靠先验知识的抽样调查. 所抽取的样本是根据调查者掌握的先验信息,认为能很好地反映总体的特征指标的一些个体. 这种抽样一般取样很少,但样本能否正确反映所需的总体特征,完全依赖调查者的主观信息,无法获得客观的误差评价. 因而典型调查往往要以普查或概率抽样调查为基础,确定典型样本,典型调查由于样本量一般很少,可以经

常进行.这三类调查相互配合,能够获得正确而时效性很强的总体信息.

1978年统计学家Jessen曾做过一个典型抽样与概率抽样比较的实验.他用126块大小不一的石头组成总体,由16个学生取典型样本.由于学生可直观地看到石头的大小和形状,可有"典型"的先验知识.他让每一位学生取样本量为1,2,5,10的样本各3个,每种样本量的样本共48个,样本量为20的样本每人一个,共16个,求石头总体的平均重量.另外按概率的随机数表,取样本量为1的随机样本126个,样本量为2的30个,样本量为5的90个,样本量为10的60个,样本量为20的10个.两类抽样估出的总体平均重量的平均绝对偏差如表1.1所示(单位:克).虽然在这一试验中,典型抽样与随机抽样在样本量相同的样本的个数没有设计成相同的值,是一个缺陷,但从表中仍可看出样本量的增加对典型抽样的精度改变不大.而随机抽样的精度随着样本量的增加有明显的改善.样本量很小时,依靠较充足的先验信息获取的典型样本为佳,而样本量较大时,随机样本的估计则更好些.

表 1.1

样本量	1	2	5	10	20
典型抽样	40.0	44.9	35.3	38.5	31.0
随机抽样	80.6	71.4	43.3	34.1	26.2

一个完整的实际问题的概率抽样调查,不仅要依据概率统计的理论作出适宜的抽样方案及数据分析方案,同时包含大量的现场调查的实践活动,这涉及组织管理、测量技术、人的心理反映等诸多方面.因此,有人说抽样调查既是一门科学、也是一门艺术.

进行一个抽样调查的操作流程大致包括以下几个方面.

(一) 建立课题,明确调查的目的

进行一项工作要有一个明确的目的.开展一项调查,想要得到什么信息,自然是应该明确的.但现实中提出一个调查项目时,往往目

的比较朦胧.通过项目提出单位、抽样设计者和执行者的研讨,会使目的逐渐明确.有了明确的目的才能在总体的确定及调查方案具体细节方面有针对性.

(二)调查的准备阶段

(1)总体及目标量的确定.总体即为要想达到调查的目的所关注的那个集合.应该有明确的条件去判定一个个体是否属于这个集合.这些条件必须是在实际工作中可用的,调查工作者应能明确判定一个可疑个体是否属于调查的总体.

(2)抽样框.抽样框是制定抽样调查方案时的一项重要内容.制定抽样方案时,总体必须划分成一些抽样单位,这些单位是相互不重叠的,并且能完全覆盖总体.划分出的抽样单位应该是可识别的,也就是想调查哪一个单位,就可以将这个单位找出来进行调查.编制抽样框可以说是实际调查工作中一项很重要、很艰难的工作.抽样框除包含有抽样单位的编号及抽样单位与总体、个体单元的联系外,还应包含一些有用的辅助信息.这些辅助信息可用于抽样方案的设计和数据处理,有益于提高调查的质量.实际工作中,抽样框非常完美地覆盖调查总体的情况是稀有的,一般需要调查的总体与抽样框所包含的个体会有一些差异,应该注意这些差异,必须使两者很接近,使差异在可容忍的范围内.

(3)收集数据的方法.调查的环境以及人力、物力、时间都会影响收集数据的方法.常用的方法有电话采访、书信邮寄和派员面访等等.

(4)抽样设计.抽样设计是抽样调查的一项主要工作,要决定从抽样框中抽出哪些单位来进行调查.既要考虑到人力、物力、对获得的信息的精度要求,也要考虑到实际现场调查工作的可行性.抽样调查方案的设计者要利用抽样框的辅助信息,综合各种基本的概率抽样方法,制定出一个可行的、精度满足要求而且费用最省的抽样方案.抽样方案不但包括调查哪些抽样单位,以及调查失败时的补救措施,还应包括调查数据获得后主要信息量的计算公式.

(5) 问卷设计. 收集数据无论采用哪些方式,都应有一份调查问卷. 根据调查的目的设计若干问题,从问题的回答中提取所需的信息. 问题要简单明确,不会产生歧义. 多数调查采用选择题的方式. 问题的设计要考虑到被访者的心理,使回答没有障碍,能得到真实的回答. 整个问卷也不宜冗长. 问卷设计是一个需要专心研究的课题.

(三) 现场工作阶段

进行一项大规模抽样调查,会遇到许多管理问题. 要进行调查员的培训,使他们了解调查的目的以及如何使用问卷获得回答. 要建立调查工作的监督机制,检查采集到的数据的质量,以及数据的保管等等. 这一阶段有大量的实际技术工作和组织管理工作.

(四) 数据处理阶段

(1) 数据验收、编辑. 数据处理的第一步是对回收的调查表进行审查,以便订正填报的错误,把明显的错误数据删去;了解调查表回收的情况,不响应是否严重,不响应的机制如何,等等. 这些情况对下一步的分析、估算有直接影响. 另外通常要进行数据的整理、编辑、计算机录入,制成数据文件,便于使用计算机处理.

(2) 估计、分析. 对调查总体特征指标的估计要按抽样设计时的既定计算公式进行,但要考虑到对数据审查时发现的种种情况,进行必要的调整. 对算出的估计值,特别是一些重要的估计值要按概率抽样调查的方案算出预期的误差大小. 当调查获得结果后,除了进行特征量的估算外,还要充分利用数据,进行其他的统计分析.

(五) 写出报告、结论

一个调查之后,写出一篇调查总结报告是必不可少的. 报告应列举获得的各项估计值,也应陈述从数据反映出的问题及相应的建议. 作为一个好的调查报告应对此次调查的得失作出总结,为今后类似的调查提供经验,提供有价值的数据信息. 这些经验、信息对将来的抽样有指导作用,它提供的特征量的均值、方差、测量值的变异性质

等，都会在未来的抽样调查中成为重要的参考数据.

本书主要讨论抽样调查的概率统计理论，集中研究抽样设计和相应的调查数据分析.

§1.2 有限总体抽样的样本分布

大规模抽样调查研究的对象是一个有有限个单元的总体. 总体中的单元是可识别的，因此我们可以将一个有 N 个个体单元的有限总体记为

$$\mathscr{U}(N) = \{U_1, U_2, \cdots, U_N\},$$

N 一般是已知的，称为总体的大小. 我们要研究的是这些单元的某项特定指标量 Y，每一个个体单元有一个对应的数据，研究的指标量集合为

$$\{Y_1, Y_2, \cdots, Y_N\},$$

我们不妨直接将指标集合 $\{Y_1, Y_2, \cdots, Y_N\}$ 作为总体. 按某一抽样方案从总体 $\mathscr{U}(N)$ 中取出 n 个单元作为样本，观测各样本单元的数量指标，样本记为

$$y_1, y_2, \cdots, y_n,$$

$\{y_1, y_2, \cdots, y_n\}$ 系 $\{Y_1, Y_2, \cdots, Y_N\}$ 的一部分. 抽取的方案使每一可能的样本有一个确定的出现概率. 这就构成一个由抽样设计形成的样本概率分布. 依据这个分布我们可以计算一些样本统计量(比如样本平均值)的期望、方差等等. 这些期望、方差是基于设计产生的概率分布计算的，因而称为**基于设计的期望、方差**.

例 1 对有限总体 $\{Y_1, Y_2, \cdots, Y_N\}$ 作有放回抽样，每次随机抽出一个单元观测后放回再抽下一个单元，得样本 $(y_1, y_2, \cdots, y_n)$，则由古典概率方法可以算出，一切可能的样本总个数为 N^n，每一具体样本出现的概率为 $1/N^n$.

例 2 对有限总体 $\{Y_1, Y_2, \cdots, Y_N\}$ 作无放回抽样，每次随机抽出一个单元，不再放回，继续抽出下一个，抽取 n 次得样本 $(y_1, y_2, \cdots, y_n)$. 对样本 $(y_1, y_2, \cdots, y_n)$ 的样本单元不计它们出现的顺序. 由古典