

梁小筠 祝大平 编著

抽样调查的方法和原理



抽样调查的方法和原理

梁小筠 祝太平 编著

华东师范大学出版社

(沪)新登字第 201 号

抽样调查的方法和原理

梁小筠 祝大平 编著

华东师范大学出版社出版发行

(上海中山北路 3663 号)

邮政编码: 200062

新华书店上海发行所经销 江苏省句容县排印厂印刷

开本: 850×1168 1/32 印张: 8 字数: 200 千字

1994 年 12 月第一版

1994 年 12 月第一次印刷

印数: 001—2,000 本

ISBN 7-5617-1217-0/O·042 定价: 9.60 元

前 言

抽样调查是获取统计资料的重要手段。广泛应用抽样方法是当今国际统计的发展趋势。在改革开放的形势下,抽样调查已经在我国许多领域中得到应用,并引起各级政府部门和有关人士的关注。随着社会主义市场经济体制的逐步建立,市场调查业蓬勃兴起,抽样调查日益成为企业家投资决策前了解市场和进行经营管理的重要步骤。

一个抽样调查项目,由好几个环节组成。怎样科学地设计抽样方案?如何按照调查目的编制问卷?有哪些调查方法?如何处理数据?最后,怎样撰写调查报告?这每一步都有学问,掌握并运用其中的科学知识是成功的关键。

我系从1986年开始开设《抽样调查》课。不久,校方又把这门课作为全校学生的公共选修课。成人教育学院也开设了这门课。本书就是将修改了多次的讲义再次加工而成的。其中,“问卷的设计”及“调查和调查报告的撰写”这两章是请上海城市社会经济调查队祝大平高级统计师撰写的。书中附有不少例题。例题和习题中有一定数量的实例。对理论部分既注意科学性,又结合多年教学中的体会,尽可能使其深入浅出。同类问题的计算公式有的列表表示,以便读者查阅。大部分定理的证明和理论性较强的内容被安排在打“*”号的章节或段落里。具有概率统计基础的读者不难看懂它们。

本书可以作为大专院校文、理科《抽样调查》课的教材,也可以供实际工作者在从事抽样调查工作时参考。

衷心感谢我系魏宗舒教授、茆诗松教授以及同行专家张尧庭教授、冯士雍教授对我从事抽样调查、编著本书的鼓励和支持。

我系的谢鹰和濮宗韦老师对本书的编写提供了许多宝贵的意见和建议。不少实际部门的同志提供了实例。有些实例来自我和协作单位的同志及我系历届同学搞过的课题。在此，向有关同志一并致谢。

由于作者水平有限，书中不妥之处在所难免。恳请同行专家和广大读者指正。

梁 小 筠

1993 年 12 月于华东师范大学数理统计系

目 录

第一章 序言	17
§ 1.1 什么是抽样调查	1
§ 1.2 总体和样本	3
§ 1.3 抽样方法概述	8
§ 1.4 抽样调查的一般步骤	4
第二章 问卷的设计	10
§ 2.1 问卷设计的原则和步骤	10
§ 2.2 问题的设立	12
§ 2.3 问卷的格式	22
§ 2.4 实例	24
第三章 简单随机抽样	30
§ 3.1 什么是简单随机抽样	30
§ 3.2 总体均值及总体总和的估计	32
§ 3.3* 有关的定理	40
§ 3.4 比例的估计	44
§ 3.5 样本容量的确定	46
习题	54
第四章 分层抽样	59
§ 4.1 什么是分层抽样	59
§ 4.2 估计量	60
§ 4.3 比例配置和最优配置	66
§ 4.4 总样本量的确定	76
§ 4.5 怎样分层能提高精度	78
§ 4.6 多指标时样本量在各层的分配	86
习题	90
第五章 比估计与回归估计	95
§ 5.1 比率的估计	95

§ 5.2 比估计	102
§ 5.3 回归估计	116
§ 5.4* 若干定理的证明及有关问题的分析	124
§ 5.5* 估计量的偏倚及均方误差的阶的估计	130
§ 5.6* 消除或减少比估计偏倚的方法	133
习题	137
第六章 整群抽样	140
§ 6.1 什么是整群抽样	140
§ 6.2 估计量	140
§ 6.3* 整群抽样的设计效应和样本容量的确定	146
习题	154
第七章 多阶抽样	157
§ 7.1 什么是多阶抽样	157
§ 7.2 初级单元大小相等的情形	158
§ 7.3 初级单元大小不等的情形	165
§ 7.4* 几个定理的证明	173
§ 7.5* 样本容量的确定	179
习题	184
第八章 不等概率抽样	187
§ 8.1 放回的不等概率抽样	187
§ 8.2 不放回的不等概率抽样	196
习题	208
第九章 等距抽样	212
§ 9.1 什么是等距抽样	212
§ 9.2 估计量及其方差	213
§ 9.3 等距抽样的精度与总体单元排列次序的关系	217
§ 9.4 具有线性趋势的总体的处理方法	221
§ 9.5 方差估计量	224
习题	226
第十章 抽样调查中的非抽样误差	228
§ 10.1 非抽样误差及其来源	228
§ 10.2 无回答的影响及处理方法	229

§ 10.3 敏感性问题的抽样调查	234
第十一章 调查和调查报告的撰写	237
§ 11.1 调查	237
§ 11.2 调查报告的撰写	238
附表 随机数表	244
参考书目	246

第一章 序 言

§ 1.1 什么是抽样调查

一、什么是抽样调查(sampling survey)

今天,“抽样调查”在人们心目中已经不是一个陌生的名词了。通过报刊、电视和广播可以经常获得有关的消息。走在大街上也会遇到调查员向过往行人征询对某些问题的看法。即使坐在家里,也可能收到寄来的或是送上门的调查表。那么,什么是抽样调查呢?

抽样调查就是按一定的程序从全体研究对象(总体)中,抽取一部分(样本)进行调查(或测试),根据样本估计总体的特征数。它是数理统计的一个重要的分支。

二、抽样调查的特点

调查可以分为全面调查和非全面调查。抽样调查是非全面调查的一种重要方式。全面调查能全面、详细地获取信息,它一直是我国有关部门了解和掌握某些基本情况的重要途径。与全面调查相比,抽样调查具有费用较低、速度较快、应用面较广的特点。由于可以较严格地挑选和培训调查员,调查和数据处理的质量比较容易控制,而且被调查者的心理压力较小。因此,与全面调查相比,抽样调查可能获得较准确的结果。此外,当调查对象非常多甚至无限,测试方法具有破坏性时,根本无法进行全面调查,这就必须进行抽样调查了。

三、抽样调查应用广泛

抽样调查是获取统计资料的重要手段。广泛应用抽样方法是当今国际统计的发展趋势之一,已经成为世界各国进行社会经济基本情况调查和各项专题调查的一种主要方法。

在我国,1953年第一次人口普查后,就在全中国抽了343个县和城镇,进行人口抽样调查。1954年开始进行职工家计调查。接着,又成立了农村抽样调查队。可是,由于历史的原因,“抽样”受到歧视。党的十一届三中全会以来,在改革开放的形势下,抽样调查已在我国许多领域中获得应用,并引起政府部门和有关部门的重视。

1984年成立的全国及各地的城、乡抽样调查队(现为城市社会经济调查队和农村社会经济调查队)已经是一支收集社会经济信息,为党政机关及企业家提供决策咨询的轻骑兵。他们定期进行农户家庭生活调查、农产量调查、农村经济调查、城市住户调查、物价调查,以了解人民的生活水平,计算各类价格指数;进行电视、广播视听率调查,定期向有关部门提供收视、收听的意向资料;承担各级党政领导布置的规模较大的各项专题调查、有关部门委托的市场调查和其它社会经济调查,并提供咨询服务。

全国和各地的人口普查办公室除了负责人口普查外,也组织人口方面的抽样调查。1987年组织了1%人口抽样调查。从1983年开始,我国每年进行人口变动情况抽样调查,估计人口出生率、死亡率及人口总数等指标。为了提前取得普查中主要项目的信息,尽快为各级领导制订政策和规划提供依据,还在普查资料中先抽10%进行数据处理。

十多年来,我国进行了一些全国性的抽样调查。如1979、1985、1990年的全国结核病流行病学调查,1982和1988年的全国妇女生育力调查,1984年的全国粮食农药污染情况调查,1986年的全国成年人(工业人口)人体尺寸调查,1987年的残疾人调查,1987和1992年的儿童情况调查,1989年的全国回国留学人员情况调查,1991年的工人阶级状况调查等。

各地区、各单位进行的抽样调查已不计其数。特别值得提出的是:抽样调查在质量管理方面也大有可为。它可以用来估计产品的合格率、平均指标(如平均强度、平均寿命),了解已售出的产

品的质量、使用情况等。

随着社会主义市场经济的建立和发展,市场调查业蓬勃兴起,抽样调查技术必将得到更广泛的应用。

§ 1.2 总体和样本

一、总体(population)

研究对象的全体称为总体。进行抽样调查,必须明确什么是这次调查的总体。例如残疾人调查,怎样的人算“残疾人”?应该有明确的标准。儿童调查中规定:在标准时点14岁以下(含14岁)的人为儿童。

与经典统计不同,抽样调查中的总体一般是有限的,对其分布一般不作假定,其中的个体都是具体的、可辨的。

在抽样以前,必须根据实际情况把总体划分成抽样单元(sampling unit)。这些单元应该互不重迭并且能合成总体。抽样单元不一定是组成总体的最小单位——个体。例如,家计调查中把户作为抽样单元。流动人口抽样调查中,可以以居(村)民委员会为抽样单元。

抽样单元又有大小之分。一个大的抽样单元可以分成若干个小的抽样单元。例如,一项全国性的调查,如果把省作为一级(初级)单元,可以把县作为二级(次级)单元,乡作为三级单元,村作为四级单元。

抽样调查需估计的总体的特征数,一般有以下几个:

(1) 总和。如全国耕地面积、上海市人口总数。

(2) 均值。如某单位职工平均收入、某县水稻亩产量。

(3) 比例(成数),即具有某种特性的单元在总体中占的比例。如某大学中参加勤工俭学的学生在全校学生中占的比例。

(4) 比率,即两个总体总和之比或均值之比。例如,上海市流动人口(总数需估计)中,居住时间超过半年者(总数也需估计)所占的比例。

(5) 中位数或其它分位数。如成年男子身高的 5%、50% 及 90% 分位数。

二、样本(sample)

按一定的程序从总体中抽取的一组抽样单元, 称为样本。样本中包含的抽样单元个数称为样本容量(sampling size), 又称样本含量或样本大小。

§ 1.3 抽样方法概述

如何抽取样本, 大有讲究。

例如 1936 年, 美国总统竞选。当时美国影响最大的民意测验机构——《文学文摘》杂志, 耗资五十万美元, 按全国各地的电话号码簿发出一千万张问卷(调查表), 要求被调查者问答: 在十一月份的全国大选中, 你将投谁的票? 该杂志社根据二百万张寄回来的问卷汇总的结果, 预测共和党候选人兰敦(A. M. Landon)将以压倒多数的票数当选总统, 而民主党候选人罗斯福(F. D. Roosevelt)将遭到惨败。结果, 惨败者却是这家预测机构, 它所预计的当选票数与实际选举的票数相差 20%。《文学文摘》杂志不得不在第二年宣告停刊。这主要由于这家机构没有考虑到 1936 年美国经济萧条, 能拥有电话机的多数是富人。投谁的票往往与经济地位有关。所以, 他们所抽到的样本不具有代表性。

与此同时, 由盖洛普(G. H. Gallup)领导的美国舆论研究所等三家民意测验机构, 事先根据人口分布的特点设计抽样方案, 并派调查员访问调查。只调查了三千个选民, 预测罗斯福将当选, 结果居然在意料之中。盖洛普极风趣地形容自己这次成功的秘诀是: “用两头马可以拉的马车, 却用五十头来拉是无用的”。由此可见, 科学地设计抽样方案, 可以事半功倍。

抽样方法很多, 归根结底可以分成两类。

一、非随机抽样(非概率抽样)

非随机抽样方法, 主要有以下几种:

1. 随意抽样

随意抽样就是按抽样者的方便,随意地抽取样本。例如,调查员在大街上,商店里随意地寻找行人或顾客进行调查。上面提到的《文学文摘》杂志社采用的也是随意抽样。这种抽样方法的优点是简便易行,但是,它受抽样者主观意志支配,很难保证样本的代表性。

2. 定额抽样

定额抽样就是依据一定的标志将总体分成若干层(若干部分),确定各层在总体中所占的比例。并按这些比例分配样本总量在各层的数额,让调查员抽到每一层所需的“定额”为止。常用的标志是地理区域、年龄、性别、民族、经济水平等。这种方法在民意测验和市场调查中用得较多。盖洛普在美国总统竞选的预测中采用的就是定额抽样法。

定额抽样法以了解总体的性质为基础。如果总体性质有了变化,而抽样者没有掌握有关的最新信息,定额抽样将会产生大的偏差。例如,1948年盖洛普的民意测验所预测共和党候选人杜威(T. E. Dewey)将战胜民主党候选人杜鲁门(H. S. Truman),选举结果正相反。预测失败的主要原因是定额抽样的样本出了偏差。当时城市居民倾向于投民主党的票,农民倾向于投共和党的票。第二次世界大战后,美国大量人口从农村流向城市。而盖洛普对人口总体的认识却仍停留在1940年的状况,因而样本中农民的比例过大,造成了对共和党得票人数过高的估计。

3. 经验抽样(判断抽样)

经验抽样就是抽样者根据自己的经验抽取具有平均水平的、典型的单元作为样本。这种方法可以利用抽样者所掌握的信息,避免发生极端的偏误。如果总体的变异程度较大,又不希望样本量太大,而抽样者是富有经验的,经验抽样可能获得较高的精度。但是,这种方法受主观因素的影响,不但由于评价标准不同,意见难以统一,而且无法避免因抽样者的主观爱好和认识水平造成的

严重差错。

4. 滚雪球抽样

抽样者先找出少数个体,通过这些个体了解其它个体,再由已了解到的个体去发现更多的个体,以此类推,这就象滚雪球一样,了解到的个体越来越多,越接近总体。例如,对吸毒者的调查就可以采用这种方法。容易看出,只有当个体之间能够并且愿意联系时,这种方法才能实行。

抽样调查是在总体中抽取一部分,以这一部分的数据来估计总体的特征数。也就是以局部推断整体,因此必然产生误差,这种误差称为抽样误差。抽样调查中抽样误差是不可避免的。如果抽样误差太大,调查结果也就失去意义了。

非随机抽样方法,有一个共同的缺点,就是无法估计抽样误差,当然更不能控制抽样误差。唯有以概率论为基础的随机抽样方法能解决这个问题。

二、随机抽样(概率抽样)

通俗地说,随机抽样就是样本的选取完全是“随机而定”,不受抽样者主观意志的影响。随机抽样具有以下特点:

(1) 能够确切地定义(或区分样本),即可以明确地指出每一个样本中包含哪些单元;

(2) 每个样本都有一个事先确定的被抽到的概率,称为入样概率;

(3) 通过与入样概率一致的随机抽样方法抽取样本;

(4) 利用样本估计总体时,要考虑入样概率。

为了进行随机抽样,必须先列出总体中所有的抽样单元,并编上号码。这就是编制抽样框。抽样框可以是一份名单(名单抽样框)、一张地图(区域抽样框),在与时间有关的调查中,也可以按时间先后排列总体中的单元,这样得到的抽样框称为时序抽样框。

在有些实际问题中,编制抽样框会遇到很大的困难,需要化费一定的人力、物力、财力。

随机抽样是一种科学的抽样方法，只要可能，就应该尽量采用。随机抽样的基本方法如下：

1. 简单随机抽样(simple random sampling)

从总体的 N 个单元中抽取 n 个，使得每个由 n 个单元组成的样本被抽到的概率都相等。由此可见，简单随机抽样是一种不放回的等概率抽样。它是其它抽样方法的基础。

2. 分层抽样(stratified sampling)

根据实际问题的需要，把总体分成若干个子总体(称为“层”)，在每层中独立地抽样，这就是分层抽样。分层抽样既可以对总体进行估计，又可以对子总体进行估计，样本的代表性较好，实施和管理也很方便。如果分层分得适当，还可以提高估计量的精度。

3. 整群抽样(cluster sampling)

设总体中每个初级单元(群)可以分成若干个次级单元。在抽中的初级单元中调查其所有的次级单元，这就是整群抽样。初级单元可以等概率地抽取，当初级单元大小不等时，也可以按不等概率抽取。由于调查单元相对集中，整群抽样实施方便，可以节省费用。

4. 多阶抽样(multi-stage sampling, subsampling)

在抽中的初级单元中再抽一部分次级单元进行调查，这就是二阶抽样。由此不难理解三阶抽样及更高阶的抽样。多阶抽样实施方便，效率一般高于整群抽样。

5. 等距抽样(systematic sampling)

把 N 个单元排成一行或一个圆圈。根据 N 和 n 的大小，在一定范围内随机确定第一个入样单元，以后每隔相等的距离抽取一个单元，这就是等距抽样。等距抽样实施方便，我国统计工作者乐于使用。如果抽样者了解总体的单元指标值的排列规律并加以利用，有可能提高估计量的精度。

上面介绍的是基本的抽样方法。在设计抽样方案时，可以根据实际情况，将这些基本方法加以组合。

值得提醒的是,抽样调查中的样本是复杂的,就连简单随机样本也不是独立同分布的。在抽样调查理论中,对复杂样本有独特的处理方法。这一点,学习时必须注意。

§ 1.4 抽样调查的一般步骤

一、明确调查目的、要求,安排调查计划,进行经费预算

明确调查目的、要求是很重要的,它关系到整个调查工作的成败。这一步也包括弄清调查的总体。如果是委托调查,应与委托单位签订合同。在需要估计的总体特征数中要分清主次,如有可能,对主要特征数的估计量给出精度要求,以便进一步确定样本量。

在此基础上,可以制订调查计划,进行经费预算。

二、设计抽样方案

设计抽样方案(如何抽取样本?),确定样本量(抽多少?),将需估计的总体特征数分类,对每一类给出估计量及其精度的有关计算公式。这一步很重要,它是抽样调查能否建立在科学基础上的关键。

三、抽样

编制抽样框,根据抽样方案利用随机数进行抽样。为了便于调查和管理,应列出样本单元名册(包括编号、姓名、地址、调查情况等)。

四、设计问卷

问卷应根据最终需提出的调查报告或汇总表的内容来设计。为了提高调查质量,可以做个试点调查,看看调查是否可行,问题是否提得恰当,并对初拟的问卷进行修改。设计好的问卷要请委托单位有关同志过目,必要时再作修改。

五、调查

调查方式很多,可以邮寄调查表、上门访谈,也可以把被调查对象请来进行调查。在调查前要培训调查员,使他们明确调查目

的、内容和要求。对问卷中的问题进行充分的讨论,避免对一个问题产生不同的理解,还要弄清问题的逻辑关系。

六、数据处理

首先要对填写好的问卷进行检查,找出有异常值或有逻辑错误(如十岁的小孩已退休)的问卷,并加以处理,例如重新调查或作为废卷。

对比较复杂的问卷可以编码,但需加以检查,避免产生新的错误。

对录入的数据,要作值域检查和逻辑检查,以保证录入质量。

数据检查完以后,可按已给出的公式进行计算或打印汇总表,还可以利用其它统计方法进一步提取数据中蕴含的信息。

七、分析数据处理的结果,撰写调查报告