

抽样调查理论基础

CHOUYANG DIAOCHA LILUN JICHU



抽样调查理论基础

chouyang diaocha lilun jichu

联合国统计局编

Lian He Guo Tong Ji Ju
Bian

谢 嘉
郑德如
黄树颜译

Xie Jia
Zheng De Ru
Huang Shu Yan
Yi

贾宏宇
赵章甫校

Jia Hong Yu
Zhao Zhang Fu
Jiao

上海人民出版社

Shang Hai
Ren Min Chu Ban She

Elements of Sample Survey Theory
Statistical Office of the United Nations
New York 1972

责任编辑 黄明辉
封面装帧 王建纲

抽样调查理论基础

联合国统计局编

谢 嘉 郑德如 黄树颜 译

贾宏宇 赵章甫 校

上海人民出版社出版

(上海绍兴路54号)

新华书店上海发行所发行 上海商务印刷厂印刷

开本 787×1092 1/32 印张 9 插页 2 字数 177,000

1984年3月第1版 1984年3月第1次印刷

印数 1—10,000

书号 4074·528 定价(五) 0.67 元

译 者 的 话

《抽样调查理论基础》是联合国统计局 1976 年发行的《统计抽样方法简要手册》第一卷。这本手册主要是为政府统计工作人员编写的，也可作为统计人员培训机构和高等院校统计专业的参考用书。

七十年代初期，抽样调查的理论和方法有了迅速发展，特别是在社会经济调查中，得到了广泛的应用。本书是在总结当前各国对于抽样调查方法的实际应用的基础上，经过系统研究而编写成的。第一编论述抽样原理，结合人口、农业、家计、工业等抽样调查，举例详细说明；第二编论述三十二种抽样调查方法，每种方法都从问题的要求、数据、计算估计平均数和估计方差的公式、例题计算和注释等方面进行了比较分析。全书内容的特色是：全面和详细地分析了各种抽样方法；对各种方法的数效率与费用效率作了详尽研究；对于估计中发生的偏差和除偏的方法作了深入探讨；提出了设计样本的新技术。

社会主义统计是制订国民经济计划和政策的重要依据，是对国民经济实行科学管理和检查监督的重要手段。在实现我国社会主义现代化建设的过程中，迫切需要多种多样的统

62.1/10

计资料。根据我国国情,应当灵活运用多种统计方法,抽样调查就是其中最主要的一种方法。本书对于抽样调查的理论和实践,作了较全面和深入的阐述,可供参考。

本书原文版中凡是我们发现错误的地方,均已改正。译文也经过多次校阅,但限于水平,缺点、错误在所难免,请读者批评指正。

序 言

这本文件是 1960 年联合国统计局发行的方法研究丛刊 F, 第 9 号,《统计抽样方法简要手册》第一卷,《抽样调查理论基础》的修订本。六十年代在抽样领域内取得许多重要的进展,尤其在应用于经济和社会调查方面。现在这个修订本主要是增加了十种抽样方法,其中有一些被认为对于发展中国家的抽样调查专业人员是新颖的,可能会引起广泛的兴趣,其他一些,虽然抽样专家们以前就熟悉的,但也包括在这里,希望全部三十二种方法能提高许多国家的统计部门、训练机关和大学使用本手册的效果。新增加的十种方法是:关于比例和百分数抽样;样本容量的估计;由于分层而得益的估计;分层随机抽样中的比率和回归估计程序;整群抽样中的比例估计;多阶段抽样中不放回的不等概率;系统抽样(等距抽样);多相抽样中的比率估计,以及从函询无回答者中的最适当的抽访部分。

手册原版中的二十二种方法的计算机程序,统计局最近出版了一本书可供参考,这本书的书名是:《统计抽样方法简要手册》第二卷,《抽样设计的计算机程序》。这本修订本中新增加的十种方法的计算机程序亦已备有文件,不久即可发行^①。

^① 原著业已出版——译者。

关于符号记法的说明

对于内容如此广泛的一门学科，要有一套完全令人满意的数学符号记法是困难的。人们只得一方面力求公式和代数运算符号的简化，另一方面保持全体一致但却用复杂的符号体系（包括原字母缀有下标和上标），而力求在两者之间采取折衷的办法。本手册的方针偏向于记法的简化和大体的一致，至于完全一致则仅限于在第一编各节及第二编各种方法之内。希望大体一致的程度，整个说来，足可使读者能理解各公式中的符号体系，而无需不适当地或重复地依靠定义。

以下是若干具体要点：

(1) 总体和样本单位数 单一阶段和第一阶段，总体单位数是 N ，样本单位数是 n 。在多阶段方法中，第二阶段总体和样本的单位数分别是 (M_i, m_i) ，第三阶段是 (Q_{ij}, q_{ij}) 。

(2) 测度 符号 y 表示受检查中的特征的测度值， x 表示辅助变量的测度值。基本单位的测度值以小 y_i 表示，而在单一阶段方法中， i 一般依据上下文表示样本中或总体中的第 i 个单位。在多阶段方法中， $y_{ijk} \cdots$ 中的 i 表示第一阶段的第 i 个， j 表示第二阶段的第 j 个， k 表示第三阶段的第 k 个基本单位的测度值…并有辅助测度值的相应符号。

(3) 总数和平均数的真值和估计值 总数的真值用大写字母表示,例如用 Y 表示总体的真实总测度值,用 Y_i 表示多阶段方法的第一阶段中第 i 个单位的总和等;用加撇号的字母 Y' 、 Y'_i 表示样本估计量(此外,当在同一节内遇到用不同方法取得总数的估计量时,用 Y'' 、 Y''_i 等等表示)。平均数总是在字母上加横杠表示,因此, $\bar{y}_N$ 是总体平均数,下标 N 表示这种平均数是根据总体所有 N 个单位得来的,又 $\bar{M}_N$ 表示从有 N 个第一阶段单位的总体得出的第二阶段单位的平均数。当估计的数量是平均数时不管用什么方法估计,通常用 $\bar{y}$ 表示(没有下标)。当总体算术平均数的估计量用作其他方法的伴随数字时,即加记一个下标,如 $\bar{y}_N$ 、 $\bar{x}_N$ 。

(4) 总体数值与参数估计量 总体数值用希腊字母表示,而参数的样本估计量,则用相应的拉丁字母表示,因此, σ 、 s 分别表示总体的和样本的标准差; α 、 β 和 a 、 b 分别表示总体的和样本的回归系数; ρ 、 r 分别表示总体的和样本的相关系数等等。偶尔用到弯曲符号记法,如用 σ^2 表示总体方差 σ^2 的估计量。

(5) 不等概率的抽样与分层 在不等概率方法中选择总体第 i 个单位的概率,用 P_i 表示。有 k 层的总体,其第 j 层的第一阶段单位所占的比例,用 π_j 表示。

(6) 其他符号 以上是大多数符号的定义。其他符号,则必须参照正文。

导 论

这本手册,正如它的标题的含意那样,不是为抽样专家们编写的;也不是要代替列在参考书目中的那些极好的教科书,而要正确体会和理解这里说明的技术方法,必须参考这些书籍。这本手册是经过对实际使用的方法,主要是政府统计机关实际使用的方法,作了一次严密的研究后编成的。这些方法只依据少数模式,并假定这些主要由环境造成的模式大体上将继续存在。虽然费用问题和当地的行政安排通常对于方法的选择起着制约的作用,但是经过研究许多调查访问,看来只要更加密切注意抽样理论和实践的原则,就有可能达到更高的效率,而只需稍加一点费用,甚至不增加费用。

再者,这本手册是为统计主管人员编写的,我们考虑到这些人不是数学家,但是他们深知本国统计的缺点,而这些缺点是能用抽样调查方法加以补救的,并且他们的责任就是要为达到这种目的作出决策。这里,列出了估计公式和与通用方法结合的抽样误差方差公式,但一般避免讲解数学。不过,某些数学方面的问题还是在附录中加以论述。

第一编的要旨在于以简单的措辞描述、讨论和比较理论性的概念和方法,强调的是常识而不是问题的数学方面,尽管

一般也给出估计公式。不过要把意思表示得既简单而又有适当的严谨程度,将两者调和起来是一个不易克服的困难。即使在基本原理上,也有一些固有的复杂性,不依靠数学是不能解决的。希望这本手册对于过去缺少抽样理论知识的统计人员,至少大部分能够看懂,而对理论一向了解但已荒疏的其他人员则能起到复习的作用。

第二编,大部分是说明各种各样方法的一批例子,开头是估计平均数(或总数)及其联带的抽样误差方差所使用的公式,并较详细地说明数据的运用。这些数据绝大部分包括可用于一个欧洲国家有关一个地区的一宗实际的农业统计。它象兵站的巡防军一样,已经经历过许多检查分析的过程。不言而喻,数字结果本身是没有多大意义的。样本小到没有代表性,而参数(如相关系数之类)虽常为特定方法的数效率所依据,但这里的参数,即使对于农业,也未必是典型。这并无意于暗示,用来自这样一种可变总体的这些小样本,即能得出充分精确的结果。总而言之,第二编的用意主要在于说明怎样进行计算。而对各种方法所加的附注,希望也是有用的。

说明这本手册是为政府统计人员而编写,或许是限定其适用范围的最好办法。本手册只打算用在经济和社会调查方面,这和实行起来多少要用点特殊技术的民意测验是不同的。由于在科学技术方面的抽样方法有大量的文献,因而本手册也不涉及抽样法对科学技术的实地实验和其他应用。

附带讲到的只是抽样实践、访问人员的采访、训练和控制的一般指导、数据的加工、不回答邮寄调查的处理以及大量的这类问题；只有解决这些问题，才能使调查获得成功。我们充分体会到抽样的数学论证是比较容易的，而使数学所根据的假设生效的实际困难是较大的。

目 录

序言

关于符号记法的说明	1
-----------------	---

导论	1
----------	---

第一编 理论	1
--------------	---

第一节 抽样的作用	2
-----------------	---

第二节 抽样调查的目标和特征的研究	6
-------------------------	---

第三节 等概率与不等概率抽样	9
----------------------	---

第四节 精确度和费用	15
------------------	----

第五节 样本单位	27
----------------	----

第六节 分层抽样	37
----------------	----

第七节 比率法	43
---------------	----

第八节 回归法	48
---------------	----

第九节 多阶段情况下的估计	53
---------------------	----

第十节 多相抽样	60
----------------	----

第十一节 穿插样本	67
-----------------	----

第十二节 设计样本单位的技术	75
----------------------	----

第十三节 偏差和除偏	82
------------------	----

第二编 抽样方法、计算公式和实例	95
------------------------	----

方法一 等概率抽样	97
-----------------	----

方法二 比例和百分率抽样	101
--------------------	-----

方法三 样本容量估计	104
------------------	-----

方法四	分层抽样	108
方法五	由于分层而得益的估计	114
方法六	不等概率抽样	118
方法七	比率法	121
方法八	分层随机抽样中的比率法	128
方法九	回归法	133
方法十	分层随机抽样中的回归法	137
方法十一	穿插样本	140
方法十二	整群抽样: 等概率、等群体、各个平均数的平均数	143
方法十三	整群抽样: 等概率、不等群体、各个平均数的加权 平均数	147
方法十四	整群抽样: 等概率、不等群体、各个平均数的未加 权平均数	151
方法十五	整群抽样: 比率、辅助资料: 群体中的基本单位数	154
方法十六	整群抽样: 比率、辅助资料: 群体的面积	157
方法十七	整群抽样: 不等概率, 根据: 群体中的基本单位数	160
方法十八	整群抽样: 不等概率, 根据: 群体的面积	164
方法十九	整群抽样中比例的估计	169
方法二十	多阶段抽样: 第一阶段等概率, 各个平均数的加权 平均数	172
方法二十一	多阶段抽样: 第一阶段等概率, 各个平均数的未 加权平均数	178
方法二十二	多阶段抽样: 第一阶段等概率、比率、辅助资料: 第一阶段单位中基本单位数	181
方法二十三	多阶段抽样: 第一阶段不等概率, 根据: 第一阶 段单位中基本单位数	184
方法二十四	多阶段抽样: 第一阶段不等概率, 根据: 第一阶 段单位的面积	189
方法二十五	多阶段抽样: 不放回的不等概率	193

方法二十六	系统抽样	197
方法二十七	设计样本单位的技术	201
方法二十八	多相抽样: 分层法	205
方法二十九	多相抽样: 比率法	208
方法三十	多相抽样: 回归法	212
方法三十一	用回归法由一个有偏的样本作无偏的估计	215
方法三十二	样本调查中无回答的: 无回答者中最佳的抽样 百分率	221
	对各种方法的概述: 比较费用的观点	225
附录		247
本书统计名词英汉对照		266
参考书目		276

第一编 理 论

这部分论述抽样原理, 主要根据常识而不是数学观点, 尽管数学是不能完全避免的。经过一般地论述问题的实质之后(这些问题如抽样框和样本单位, 等概率(EP)和不等概率(UP)抽样, 及估计量的精确度和效率所依靠的因素), 再较详细地分析比较通用的设计。对于较为特殊的处理(包括一些例子), 凡属可能, 都在第二编中随时相互引证。手册特别着重通过预先试验或对过去调查结果的分析, 宜于事前查明关于每个特殊问题的最有效率的设计。应用辅助数据来提高估计的精确度的方式亦有说明。一般假定所有样本都是大的, 所以应当认为适用“大样本”的概率理论。

在前几节中顺便提到的若干概念, 将会在后几节中加以定义并予以讨论。这样, 前面说到的, 就可得到相互参证。

第一节 抽样的作用

1.1 抽样调查几乎是与统计学本身同时发展起来的,而且它的应用正在日益扩大。它在政府统计工作中的重要作用已属无可争议。现在根据对近年来在不同领域里的许多应用实例所作的研究来总的探讨一下抽样方法,也许更有用处。

1.2 某些调查历来总是抽样的,并且仍可能一直要使用抽样调查。突出的例子是物价(它不同于单位价值)统计。人们也不会设想在全面进行的基础上作家计调查。虽然,对某些现金销售的谷物的实际产量和耕地面积可以定期得到全面统计数字(从而计算单位面积的产量),但是,一般地说,采用现代化技术的亩产量统计还必需在抽样区域内进行。在所要研究的总数中抽样若干部分的方法是常用的:例如,在估算平常不包括在工业生产调查以内的小企业工业产品时,就是抽样的。另一种典型地用于人口普查的很值得推荐的估计某些数据的方法,不论对于某些次要资料,或者按问题的性质认为不宜对每一调查对象都进行调查的资料,都可从全体中抽取一小部分(如百分之一、百分之五或百分之十)来进行估算。用这样的方法,可以肯定,虽然样本只占总体的一小部分,但样本单位的个数是很多的。并且,如果对分类细节不要求太详

尽的话,该估计将是准确的。因此,对一些不需要非常精确——可能只是为了管理上的或其他用处,有时只要找出其大小的次序的一些次要项目,抽样方法有着广泛的用处。

1.3 在统计工作还处于初级发展阶段的国家里,即使是主要的统计调查也必需依靠抽样。这些国家一般不具备全面调查的设备。各国所需的基本统计资料是人口统计,而许多国家至今还缺乏进行这种工作所必需的先决条件——大型的处理机器。他们必须依靠抽样,因为这种方法非常有利于形成一个培训熟练调查人员的核心。每个国家于一定时期必须进行一次全面的人口普查。即使在可见的将来,其他统计资料的搜集仍将继续使用抽样方法。事实上,即使在统计工作发达的国家里,全面调查的最重要的作用之一,是对各种统计如有关小业主和手工业工人等等,提供一个样本框(参阅第2.1小节)。抽样调查在检验全面调查或大样本抽样调查的准确性方面,也大有用处,这里还涉及到邮寄调查的检查和多相抽样法问题(详见第十节)。

1.4 与此同时,也可能会夸大抽样调查的用处,正象已经指出的那样,抽样调查通常并不能在耕地亩数、个人年龄组、职业等方面提供足够准确的详细统计数字;例如,假定设计的某一抽样调查对全国实际总数(如全国牛的总数)的可能误差(见第4.10小节)为百分之二;譬如说,全国由25个县组成,那末每个县的相应误差约达到百分之十,这个误差便可能太大,不能应用了。

1.5 抽样调查人员所遭遇的最大困难,可说是对于变化