抽样信息在复杂调查数据中的应用研究*

吕 萍

内容提要:随着社会经济的发展,人们越来越多地使用调查数据,尤其是利用复杂调查数据进行研究。复杂调查数据源于复杂抽样设计,是指在抽样调查过程中使用分层、多阶段、整群和不等概率等复杂抽样设计获得调查数据。在数据分析中,若忽视层、群等抽样设计的复杂性,直接利用调查数据按照传统数据分析方法,容易得出错误的结论,尤其是涉及标准误的估计。本文主要介绍复杂抽样设计和复杂调查数据的特征,对抽样信息在复杂抽样数据中的应用进行研究,并以中国家庭追踪调查(China Family Panel Studies, CFPS)为例详细说明抽样信息在复杂调查中的应用,说明在复杂抽样调查数据中使用抽样信息的重要性。

关键词:复杂调查数据;复杂抽样设计;层;群;权数
DOI: 10.19343/j.cnki.11-1302/c.2017.01.011
中图分类号: C811 文献标识码: A 文章编号: 1002-4565(2017)01-0108-11

The Research on the application of Sampling Information in Complex survey data

Lv Ping

Abstract: With the development of society and economy, more and more people do some research using survey data especially the complex survey data. Complex survey data refers to sample design in which samples have been sampled in a way that is multi-stage, stratified, clustered and unequal probability sampling design. In data analysis, it will get the wrong conclusions if use the traditional data analysis method and neglect the complex sampling design, especially for the standard error estimators. This paper mainly introduces the characters in complex sampling design and complex survey data and studies the application of sampling information in complex sample survey data. Basing on the data of China family panel studies, it shows how to use the information in complex survey data and explains the importance of sampling information in complex sample survey data.

Key words: complex survey design; complex survey data; stratum; cluster; weight

一、复杂抽样设计中的抽样信息及其特征

在实际调查中,通常需要根据实际调查的要求,例如调查的目的、估计精度、调查经费和可操作

* 本文为国家社会科学基金“追踪调查中小域估计的方法及其应用研究”(11CTJ005)的阶段性成果;获国家自然科学基金项目“基于涵盖误差的我国周期性普查数据质量评估方法:理论与应用研究”(71301033)、国家社会科学基金“平衡抽样设计及其在政府统计调查中的应用研究”(5BTJ008)的资助。

性,结合各种抽样方法的特点,扬长避短,采用分层、多阶段、不等概率和整群的抽样设计,提高抽样效率和样本的代表性。在抽样设计中,将分层、整群、不等概率和多阶段等多种抽样方法结合的抽样设计称为复杂抽样设计。

在抽样调查中,常用设计效应(Kish L,1965)[1]度量复杂抽样设计相对于简单随机抽样的效率或相对精确程度。设计效应是复杂抽样设计与具有相同样本量的简单随机抽样设计的估计量的方差之比。当设计效应大于 1,代表该抽样设计的变异性大,需要更大的样本量才能达到简单随机抽样的效率; 反之若小于 1 说明只需要少量的样本量就能达到简单随机抽样的效率。不同的抽样设计方法,其设计效应不同。复杂抽样设计综合了各种抽样方法,所以复杂抽样设计的标准误和设计效应的计算十分复杂,需要考虑抽样信息等因素的影响。复杂抽样设计的抽样信息主要表现在:

(1) 层: 充分利用辅助信息的有效分层可以大大提高抽样效率,忽视层一般会高估抽样误差;

(2) 群: 虽然可以节省调查费用,方便调查的实施,但是由于群内的相关系数往往较高,群内有更多的相似性,所以群内相关性和群规模对抽样效率的影响较大,忽视群一般会低估抽样误差;

(3) 权数: 权数主要是用来弥补不等概率、样本的结构性偏差和无应答偏差,利用权数可以得到无偏估计量但会增加抽样误差;

(4) 多阶段: 在多阶段抽样方法中,多阶段的抽样设计主要是便于调查和抽样的组织实施,会增加抽样误差。但是,当初级单元的抽样比例较小时,初级抽样单元是影响抽样误差的最主要因素。因此在一个大型抽样设计中(初级单元的抽样比例较小),通常只考虑初级抽样单元对抽样效率的影响; 否则需要考虑各个阶段的层、群和权数等信息对抽样效率的影响。

因此,在复杂抽样设计中,首先需要根据抽样设计方案获取复杂抽样信息,例如中国家庭追踪调查(CFPS),需要考虑复杂抽样设计中包含的层、群、权数和多阶段等抽样信息及其影响。

二、抽样信息在复杂抽样数据中的应用

(一) 复杂调查数据的特征

复杂抽样设计下得到的调查数据称为复杂抽样数据,由于其包含层、群和不等概率等因素的影响,使其不再满足独立同分布的假定,因此其分析方法与传统的基于简单随机抽样的数据分析方法有所不同。由于复杂抽样数据中个体样本之间的相似性大于简单随机抽样数据中的个体样本之间的相似性,若使用传统的基于独立同分布假定的统计方法,往往导致错误的结论。这是因为基于复杂抽样设计的调查数据其变异数和自由度都发生了变化,而大部分统计分析(参数估计和假设检验)都是以变异数、自由度为基础。因为复杂抽样设计的自由度小于简单随机抽样的自由度,方差会大于基于简单随机抽样的方差。所以,若用传统的方法分析复杂调查数据会低估变异数,高估统计显著性,得到错误的假设检验结果。因此需要对传统的统计方法进行调整。例如,标准误的计算、卡方独立检验、F 和 T 检验、回归分析的最小平方法及逻辑回归分析的最大似然法等,都无法直接应用在复杂抽样调查的数据分析上,而必须根据复杂抽样信息有所调整。

(二) 抽样信息在复杂数据的标准误计算中的应用

在复杂抽样调查数据中,抽样信息在标准误计算中的应用是十分重要的,若忽视抽样信息,会严重低估抽样误差,从而得到错误的结论,其应用主要包含以下三种方法。

1. 忽视抽样设计。

若忽略复杂抽样信息,直接计算方差,然后再乘以设计效应,其中设计效应一般来自公开发布的比较权威的数据或研究文件。但是这种方法比较粗糙,不同的研究变量和研究群体有不同的设计效应,并且不同时间或不同调查项目的设计效应也不同,因此该方法使用较少。

2. 在拟合模型时加入抽样设计信息。

在对调查数据的分析中,有基于设计和基于模型两种方法。前者要求在分析调查数据时考虑样本的选择过程,即无论是描述性研究还是分析性研究都需要考虑抽样权数等抽样信息;后者认为样本的选择过程在某个具体的模型中是无关紧要的,当模型的假定正确时,由于因变量(目标变量)可以通过模型中的自变量(辅助变量)拟合,因此样本的选择过程不受影响。但是在实际应用中,由于辅助变量无法获得、辅助变量虽然可以获得但是缺失严重无法使用或无法获得理论上正确的模型,此时利用调查数据得到的结论是有偏的,对数据的解释也是错误的。而如果使用抽样信息,在模型假定错误时至少可以得到稳健的估计量。所以,在拟合统计模型时,将层、群和加权时使用的设计信息(例如各个抽样阶段信息、无应答调整和事后分层调整时划分的调整层的信息)等抽样信息加入统计模型可以获得稳健的估计量。但是,对于复杂的抽样设计,首先往往包含较多的层和群变量,无法将所有的复杂抽样设计信息加入统计模型。其次,加入较多的抽样设计信息,不仅容易导致模型的过度拟合等统计模型问题,而且将所有抽样信息加入模型中可能与模型的理论相违背,例如设计信息与目标变量无关、设计信息与部分自变量相关等。最后,在实际调查中,由于保密性考虑,次级单元的抽样信息往往是不公布的,也无法将所有的抽样信息包含在统计模型中。

3. 复杂方差计算方法。

在数据分析中,即使是简单随机抽样下的调查数据,对于某些估计量,例如中位数或比例估计量,其方差的计算也是比较复杂的,传统的统计方法并没有较为灵活的方差估计方法。在抽样调查中,对于分层、整群、系统抽样和多阶段抽样设计都有相应的方差计算公式,但是当遇到复杂抽样设计时,方差的计算公式就更为复杂了,这就限制了公式法的使用。在实际应用中,应用较多的是泰勒级数展开法和重抽样方法,其中重抽样方法包含平衡半样本方法(Balanced Repeated Replication)、刀切法(Jackknife Repeated Replication)和自助法(Bootstrap Repeated Replication)。这几种方法各有优劣,其中泰勒级数展开法可以估计非线性的函数,方法简单且计算时间短,但是无法估计中位数、分位数等无法函数化的估计量;平衡半样本方法主要针对层数较多,且每个层中包含两个初级样本单元的情况;刀切法主要针对每个层中包含至少两个初级样本单元;自助法需要复制大量与原始抽样设计相同样本量和样本结构的伪样本,且误差较大,因此使用相对较少。但是,不论是哪一种方法都需要根据复杂抽样设计获取其群、层和权数等抽样信息。

由上可知,在复杂调查数据的抽样误差的计算中多使用第三种复杂抽样方差的计算方法,即泰勒级数展开法和重抽样方法。例如,在 CFPS 中,同样使用重抽样方法计算标准误,由于其异常复杂的抽样设计,若忽视抽样信息,容易低估抽样误差。但是,若采用拟合模型的方法,过多的层和群的变量会影响模型的拟合效果。目前,部分统计软件,例如 R、SAS、SPSS 和 STATA 软件可以使用泰勒级数展开法和重抽样方法。

(三)复杂抽样数据中抽样信息的提取

1. 误差层和误差群。

在分析复杂调查数据时,需要根据复杂数据分析的要求和复杂抽样设计的特征获得正确的层、群和权数等抽样信息。此处的层和群是指误差层和误差群,而传统的层和群是抽样设计中的层和群,误差层和误差群是使用复杂抽样方差计算方法根据抽样设计和权数的计算方法对原始的层和群进行修正后得到的伪层和伪群。例如在计算误差时,需要满足每个层中至少有两个抽样单元,且每个层和群需要满足一定的样本量。但是,有些大型调查的抽样设计无法满足这个要求,例如某些层中的初级抽样单元的数目不足两个,或为了保护受访者的隐私使某些层中的群的数量太多且每个群中的抽样单元较少。此时,需要根据抽样设计方案、抽样实施过程和样本结构对层或群进行调

整,以获得使用泰勒级数展开法和重抽样方法的误差层和误差群。一般合并的原则是: ①考虑各层的特性,尤其是实际抽样时样本的排列顺序,将相似的层合并在一起。②合并时,需要考虑各个层的大小,即大层和大层合并,小层和小层合并,使其在重复抽样中误差层的差异不大,若层和层的差异较大,则需要利用加权方法调整,以避免高估抽样误差。③某些层中的群数量较多,且观察值很少,此时将层中的群进行合并,使每个误差群的个数满足一定的样本量,也有助于保护受访者的隐私。最终,经过调整得到误差层和误差群。因此,使用泰勒级数展开法和重抽样方法最关键的是根据复杂抽样设计信息获取其误差层和误差群。

2. 多阶段的抽样信息。

在多阶段抽样设计中,抽样误差的计算非常复杂,需要考虑各个阶段对抽样误差的影响,即需要考虑各个阶段的误差层、误差群和权数对抽样误差的影响。但是,在实际调查中,一方面出于对调查和受访者隐私的保护,调查只给出初级单元的信息。另一方面,对于一个大型的抽样设计,由于初级单元的抽样比例较小(例如小于 10%) ,根据多阶段抽样误差的计算公式[2] ,抽样误差主要体现在初级单元间的误差。虽然这种方法会低估抽样误差,但是,在实际调查中,初级单元一般是无放回的抽样设计,为简化抽样误差的计算,往往将初级单元视为有放回的抽样设计,导致高估抽样误差,因此一定程度上可以抵消部分初级单元后的误差。所以,在一个大型的多阶段抽样设计中,一般仅考虑初级单元的误差层、误差群和权数对抽样误差的影响。对于多阶段的复杂抽样调查数据,常用的统计软件 SAS、R、SPSS 和 STATA 都可以使用第一阶段的误差层和误差群来估计抽样误差。但是,到目前为止,权威的 SAS 软件只能使用第一阶段的误差层和误差群计算抽样误差,R、SPSS 和 STATA 可以使用所有阶段的误差层和误差群计算抽样误差。在 CFPS 中,由于数据保密性等因素,同样只给出了第一阶段的抽样信息。

在实际调查中,有时为了保护调查或受访者的隐私,调查数据中不包含层、群的信息,而是使用重权数的方法,即将复杂设计的群、层和权数等抽样信息整合在重权数[3] 中,在数据分析时只需要利用重权数进行相关的数据分析。

实际上,复杂抽样设计方法在抽样设计的教材中均有较为详细的理论介绍,但是研究者仍旧习惯使用传统的统计数据分析方法,究其原因,一方面是由于统计学中的估计和统计推断都是建立在数据独立同分布的假定条件下,人们习惯这些理论和方法,同时,大多数统计软件都是默认采用传统的统计方法进行数据分析。而复杂数据的分析较为复杂,人们往往忽视数据结构的复杂性,软件中复杂数据统计方法的发展也相对滞后。另一方面,国内的大多数调查数据没有规范的文件详细介绍其抽样设计方法,即使有详细的抽样设计方案,非抽样设计者往往难以从复杂的抽样设计方案中得到所需要的抽样设计信息,这就限制了对复杂抽样调查数据的研究与正确使用。但是,随着国内一些大型调查研究机构的兴起,例如北京大学的中国社会科学调查中心、中国人民大学的中国调查与数据中心和西南财经大学的中国家庭金融调查与研究中心等研究机构,不仅免费公布其代表性的大型调查数据,如 CFPS、中国养老与健康追踪调查(China Health and Retirement Longitudinal Study,CHARLS) 、中国综合社会调查(Chinese General Social Survey,CGSS) 等,而且提供了大量抽样、调查执行方面的技术报告,以帮助数据使用者对复杂调查数据进行深入研究。

下面以 CFPS 为例,详细说明其复杂抽样设计和抽样信息及其在复杂抽样调查数据中的应用。

三、中国家庭追踪调查的复杂数据分析

(一) CFPS 的复杂抽样设计

CFPS 是北京大学中国社会科学调查中心主持的追踪调查项目之一,调查对象是中国(除香

港、澳门、台湾、新疆维吾尔自治区、西藏自治区、青海省、内蒙古自治区、宁夏回族自治区和海南省等省区之外) 的 25 个省市自治区的家庭户和家庭户中的所有满足调查条件的家庭成员。调查重点关注的是中国社会经济、教育、家庭、人口和健康等方面的变迁。在抽样设计上,首先将 25 个省市分成两类: 一类为省级层次的推断总体,用以满足省级推断的要求,包含辽宁省、上海市、河南省、广东省和甘肃省 5 个省市,也称为“大省”; 其余 20 个省市为非省级推断总体。两类样本数据加权得到 25 个省市总体的有效估计,用来推断全国①。抽样设计是分别在 5 个“大省”和 1 个“小省”中采用三阶段、不等概率的整群抽样设计[4]。

第一阶段: 分别在广东省、甘肃省、辽宁省、河南省 4 个“大省”和 1 个“小省”的区县抽样框中,以区县的人口数为辅助变量,按照与区县的人口数成比例的系统 PPS 抽样方式抽取 16 个或 80 个区县样本。

上海市只有 19 个区县,为了提高抽样的效率,将街道乡镇确定为初级抽样单元。即在上海市的街道乡镇抽样框中,以街道乡镇的人口数为辅助变量,按照与街道乡镇的人口数成比例的系统 PPS 抽样方法抽取 32 个街道乡镇样本。

第二阶段: 在区县样本或街道乡镇样本的村居抽样框中,以村居的人口数为辅助变量,按照与村居人口数成比例的系统 PPS 抽样方式,抽取 4 或 2 个村居样本。

第三阶段: 在村居样本的家庭户抽样框中,按照循环等距抽样方式,抽取 28 ~ 42 不等的家庭户样本,对家庭户样本和家庭户样本中所有满足调查要求的家庭成员进行调查。

由于 CFPS 抽样设计的复杂性,调查数据清理和加权调整也异常复杂,耗时较长,无法满足研究者对调查数据进行快速分析的需求。因此,在中国家庭追踪调查的数据库中包含了一个再整合数据库,即对 5 个“大省”进行再抽样调整样本,使 5 个“大省”的抽样比与“小省”的抽样比近似相同,以便在没有及时获得权数的情况下利用调查数据对总体推断。但是,由于其抽样设计的复杂性,后期仍旧对其进行了加权调整,以方便研究者使用。

由上可知,CFPS 是复杂抽样设计,估计总体包含 5 个自我代表省(“大省”) 、1 个全国总体和 1 个再整合全国总体(下文也称为全国再整合) 。

CFPS 是追踪调查,目前已经完成 2010 年的基线调查、2012 年和 2014 年两轮追踪调查。还要对每个总体的追踪数据和各年截面数据分析,表 1 是 CFPS 抽样设计中各阶段的抽样信息表。

表 1　　CFPS 基线调查抽样设计中各阶段的抽样信息表

省市	甘肃省	广东省	河南省	辽宁省	上海市	全国	全国再整合
估计总体	自代表省	自代表省	自代表省	自代表省	自代表省	全国总体	全国总体
初级抽样单元	区县	区县	区县	区县	街道乡镇	区县或街道乡镇	区县或街道乡镇
抽样方法	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS
初级单元数	16	16	16	16	32	176	106
二级抽样单元	村居	村居	村居	村居	村居	村居	村居
抽样方法	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS	系统 PPS
二级单元数	64	64	64	64	64	640	416
三级抽样单元	家户	家户	家户	家户	家户	家户	家户
抽样方法	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样	系统抽样
三级单元数	1872	2038	1888	1972	2564	19986	12756
三级有效数	1600	1600	1600	1600	1600	8000	10400

(二) CFPS 的复杂抽样设计信息

CFPS2010 和 CFPS2012 的数据库中都给出了最终的权数、6 个大层和 162 个区县,并给出了抽

① CFPS 样本虽然仅包含 25 个省市自治区,但是对全国有一定推断能力。

样设计和加权调整方案的技术报告。目前,国内外期刊网上可以查到数篇使用 CFPS 调查数据的文章、研究报告或论文,但是这些文献或者没有使用任何抽样信息或仅使用了权数或使用了抽样信息但错误地将 6 个大层和 162 个区县作为 CFPS 的误差层和误差群,笔者认为这是不准确的。因此,本文根据 CFPS 的抽样设计方案重新得到 CFPS 的误差层和误差群,以便为研究者使用抽样信息进行复杂抽样调查数据分析时提供参考。下面给出误差层、误差群和权数的说明。

1. 误差层。

CFPS 包含 5 个"大省"和一个"小省",数据库中共包含 6 个抽样框,每个抽样框使用多阶段、不等概率的整群抽样设计。由表 1 可知,初级单元(PSU)和二级抽样单元(SSU)的抽样都充分利用辅助信息、采用与人口成比例的系统 PPS 抽样方法,以提高抽样效率;三级抽样单元(TSU)的抽样采用系统抽样方法。换言之,在每个抽样框中的每个抽样阶段均采用了某种系统抽样的方法。根据抽样设计原理,系统抽样设计可视为每个层包含 1 个抽样单元的分层抽样设计。但是,抽样误差的计算要求每个层中至少包含两个抽样单元,所以需要将抽样单元数量不足的层合并。在 CFPS 中需要按照抽样过程中抽样单元的排列顺序邻近合并,合并方法是:

(1) 5 个"大省":第一阶段有 16 个初级抽样单元,按照抽样时初级单元的排列顺序两两合并,共有 8 个误差层。第二阶段,在每个 PSU 误差层中有 8 个二级抽样单元,按照抽样时二级抽样单元的排列顺序两合并,每个 PSU 误差层中包含 4 个 SSU 误差层,共 32 个 SSU 误差层。第三阶段,在每个 SSU 误差层中,抽取三级抽样单元(TSU),同样两两合并,得到相应的 TSU 误差层。

(2) 在全国样本中,5 个"大省"的误差层与(1)相同,1 个"小省"的误差层的获得方法与(1)相似,共 88 个 PSU 误差层和 320 个 SSU 误差层。

(3) 在全国再整合样本中,由于对 5 个"大省"进行再抽样,所以 5 个"大省"的误差层需要重新按照(1)的方法获得,"小省"的误差层与(2)相同。

最终,得到 CFPS 的误差层,如表 2 所示。

表 2　　CFPS 抽样设计的层

省市	估计总体	PSU 层	PSU 误差层	SSU 层	SSU 误差层	TSU 层	TSU 误差层
甘肃省	自代表省	16	8	64	32	1872	936
广东省	自代表省	16	8	64	32	2038	1019
河南省	自代表省	16	8	64	32	1888	944
辽宁省	自代表省	16	8	64	32	1972	986
上海市	自代表省	32	8	64	32	2564	1282
全国	全国	176	88	640	320	19986	9993
全国再整合	全国	106	53	416	208	12756	6378

由表 2 可知,按照 CFPS 的抽样设计,误差层的数量较多,尤其是第三阶段(末端抽样)。在实际分析中,由于末端抽样采用的是系统抽样,村居内样本差异较小,抽样过程中也没有使用辅助信息,因此 CFPS 的末端抽样可视为简单随机抽样。在多阶段计算中,每个阶段采用的都是无放回的抽样设计,但是由于第一和第二阶段多采用与人口成比例的系统 PPS 抽样方法,抽样比例较小,一般可以省略,可视为有放回的抽样。由于在多阶段抽样设计中抽样误差主要集中在第一阶段,同时 CFPS 由于数据保密性等原因,也只给出了第一阶段的抽样信息,因此只考虑第一阶段的抽样误差和误差层。

2. 群。

由表 1 可知,CFPS 的各个估计总体的抽样单元即为群,即第一阶段的群是区县(上海市是街道乡镇),第二阶段是村居,第三阶段是家庭户,因此群分别为区县、村居和家庭户。由表 2 得到 CFPS

的误差层，且每个误差层中包含两个群，满足抽样误差计算中要求每个误差层中至少有两个误差群的要求。但是，在实际应用中需要注意每个误差群中的样本量不能太少，若某个群的样本数量太少则需要对群进行临近合并。在 CFPS 中，因为只考虑第一阶段的抽样误差，而且每个初级单元的样本量基本相等且满足一定的规模，所以每个误差群中的两个 PSU 即为 CFPS 的误差群。

3. 权数。

目前，CFPS2010 和 CFPS2012 的权数调整[4]已经完成且包含在相应的家庭和个人数据库中，但是 CFPS2012 的权数还未公布。因此，本文以 CFPS2010 和 CFPS2012 的复杂数据为例进行分析。

CFPS2010 是基线调查，主要针对家庭和个人数据库进行权数调整，共包含抽样设计权数、无应答权数、事后分层权数和极值调整权数四部分。由于 CFPS 包含 5 个子总体、1 个全国总体和 1 个再整合全国总体，抽样设计十分复杂，其权数调整也异常复杂，因此在 CFPS2010 数据库中，没有给出各阶段的抽样设计权数、无应答调整权数和事后分层调整权数，仅给出最终的极值调整权数。

CFPS2012 是第一期的追踪调查，权数调整包含两部分：即 CFPS2012 的截面权数和追踪权数，其中截面权数用来分析 2012 年各个总体的情况；追踪权数分析 2010—2012 年的变化情况。由于 CFPS 仅对 2010 年完成家庭成员问卷的基因成员进行追踪，是对个人的追踪，因此 2012 年的截面权数是在 2010 年的家庭成员问卷的无应答调整权数的基础上完成的，其样本包含 2010 年的基因成员和 2012 年新进的基因成员，权数调整过程同样包含 CFPS2012 的无应答调整权数、事后分层权数和极值调整权数，最终的极值调整权数为截面权数，用来对 2012 年的总体情况进行分析。追踪权数仅针对 2012 年追踪到的 2010 年的基因成员加权调整，同样包含无应答调整权数、事后分层权数和极值调整权数，最终极值调整权数为追踪权数。

各个总体的家庭和个人数据库中的权数变量如表 3 所示。

表 3　　CFPS 的估计总体及其权数

总体	估计总体	2010 家庭权数	2010 个人权数	2012 家庭截面权数	2012 家庭追踪权数	2012 个人截面权数	2012 个人追踪权数
甘肃省	自代表省	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
广东省	自代表省	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
河南省	自代表省	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
辽宁省	自代表省	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
上海市	自代表省	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
全国	全国	Fswt_Nat	Rswt_Nat	fswt_natcs12	fswt_natpn1012	rswt_natcs12	rswt_natpn1012
全国再整合	全国	Fswt_Res	Rswt_Res	fswt_rescs12	fswt_respn1012	rswt_rescs12	rswt_respn1012

（三）CFPS 复杂调查数据分析

由上，得到 CFPS 的误差层、误差群和权数等抽样信息后，下面以 CFPS 的家庭人均纯收入为例，介绍如何使用抽样信息进行复杂数据的数据分析。家庭人均纯收入不可避免地存在缺失数据（项目无应答），缺失数据的处理有多种方法[5]（删除法、插补法和模型法等），取决于缺失数据的数量、缺失模式和缺失机制，本文主要研究抽样信息对复杂抽样调查数据分析的影响，所以不对缺失数据进行插补，假定缺失数据为随机缺失，采用删除法。但是，在实际数据分析中，需要根据研究目的对缺失数据进行插补。

1. CFPS 2010 家庭人均收入的描述性统计分析。

家庭人均纯收入常用均值和中位数描述，本文使用 5 种不同方法（不加入抽样信息、仅加入权数、加入权数和层、加入权数和群，以及加入所有抽样信息）对 CFPS 2010 数据各个总体进行估计。

估计结果显示，从不加任何抽样信息到逐渐加入所有抽样信息，各个总体的估计量发生了变化，尤其是标准误和自由度，由此可以得出：

(1) 不加入抽样信息,自由度是总样本量减 1,估计量的标准差最小,说明忽视复杂抽样设计,视为简单随机样本,导致标准误差的低估。

(2) 仅仅使用权数,估计量减小,自由度不变,由于权数的变异性,估计量的标准差增加,设计效应大于 1。加权调整过程中考虑了分层、整群和不等概率的复杂设计信息,得到无偏估计量。虽然权数增加了估计量的标准误,但是由表 4,权数的变异并没有增加较多的标准误,说明权数是有效的。但是,最终的权数只是一个数值,说明每个样本代表的总体数量不同,无法代表复杂抽样设计中的群和层等抽样信息,因此还需要进一步考虑层、群等抽样信息。

表 4　　CFPS2010 全国分城乡收入组的人均家庭收入的均值估计量

	收入组	收入均值(加入抽样信息)			收入均值(不加入抽样信息)		
		样本数量	均值	标准误	样本数量	均值	标准误
全国	0 ~ 25%	3571	1714	22	3422	1675	14
	25% ~ 50%	3262	4415	17	3411	4338	14
	50% ~ 75%	3348	8170	32	3555	8385	28
	75% ~ 100%	3670	24963	995	3463	26476	514
城镇	0 ~ 25%	875	1818	39	818	1733	28
	25% ~ 50%	1193	4457	37	1250	4380	24
	50% ~ 75%	1758	8303	43	1886	8536	39
	75% ~ 100%	2840	26037	1168	2712	27910	613
农村	0 ~ 25%	2696	1674	26	2604	1657	16
	25% ~ 50%	2069	4388	21	2161	4313	18
	50% ~ 75%	1590	8030	47	1669	8214	40
	75% ~ 100%	830	21639	1425	751	21294	819

(3) 当在权数的基础上加入层时,因为自由度是总样本量减去层数,所以自由度减小,估计量的标准误和设计效应有所下降。这说明,使用权数和层,相当于在估计时使用复杂抽样设计中的层和不等概率的抽样信息,不仅得到了无偏估计量,而且良好的分层增加了抽样效率,提高了估计精度,所以标准差比单独使用权数有所减少,也进一步说明了 CFPS 的分层是有效的。

(4) 当在权数的基础上仅使用群时,由于群内样本的相似性,自由度是群的数量减 1,CFPS 群的数量相对是非常小的,即自由度较小,群内的同质性较强,因此仅使用群和权数,抽样误差增加很大,设计效应也较大。本次调查,群的影响较大,这进一步说明,整群抽样虽然降低了实施的难度,但是却降低了抽样效率和估计精度。

(5) 在使用所有的抽样信息时,权数得到无偏估计量,层增加估计精度,群降低估计精度,自由度是群的数量减去层的数量,自由度减少,因此相对于直接使用权数和群的信息,抽样误差有所下降,相应的设计效应也有所下降。但是由于 CFPS 群的数量较小,群内样本量较大,同质性较高,标准误变化较大。进一步说明群和权数增加了抽样误差,降低了估计精度,层增加了估计精度,只有包含所有的抽样信息,才能既得到无偏估计量,又得到正确的标准误。

由上可知,在对复杂数据进行分析时,使用权数可以得到无偏估计量。但是,只有使用所有的抽样信息(层、群和权数),才能得到正确的标准误。此外,加权前后的标准误差异并不是很大,说明权数并没有导致增加较大的方差,而层和群,尤其是群对标准误差的影响较大。对于 5 个“大省”,广东省、上海市和辽宁省的标准误差较大,可能是由于完访样本数量较少,一定程度上说明这 3 个省市的初始样本量估计不足。

在使用 CFPS 2010 数据进行估计时,有两个全国总体,即全国总体和全国再整合总体,在统计分析时如何使用这两个总体呢? 由于全国样本中包含 5 个“大省”,5 个“大省”是过度抽样,所以在没有权数时,全国样本是无法单独分析的,因此全国样本的家庭人均纯收入的均值估计量是没有

意义的。但是，加权后的均值估计量是无偏估计量，是有意义的。加入所有抽样信息后的标准误差是对均值估计量标准误的正确估计。而全国再整合总体，由于数据清理和权数调整的滞后性，可以满足研究者在没有权数时对 CFPS 数据及时分析的需求，因此其不使用抽样信息的均值估计量与全国总体的加权均值估计量比较接近。由于中位数受极值的影响较小，因此全国样本的加权和不加权的中位数估计量差异较小。为方便对再整合样本的分析，我们对 CFPS 的再整合样本也进行了加权调整，两个总体都可以对全国总体估计，在数据分析时选用哪个总体，需要根据所研究的目标变量确定。以家庭人均纯收入为例，首先由全国和全国再整合总体的设计效应可知，全国总体的抽样设计的复杂性要高于全国再整合总体。其次，加入抽样信息后全国总体的家庭人均收入的均值估计量的标准误低于全国再整合总体的标准误，因此，在对家庭人均收入的分析中使用全国总体样本的估计精度要高于全国再整合总体样本。最后，由两个总体的设计效应和样本量可知，全国总体的有效样本量较全国再整合总体的有效样本量多，也说明了全国样本的精度高于全国再整合样本的精度。由上分析，对于家庭人均收入这个研究变量，应该使用全国总体样本。

2. CFPS2010 家庭人均纯收入的收入组统计分析。

在使用收入做数据分析时，研究者常用收入分位数划分的收入组进行相关研究，但是在复杂抽样调查数据中，需要使用含有抽样信息的分位数划分收入组。比较全国样本的家庭纯收入的收入组的划分和统计结果。可知，加入抽样信息后，家庭人均纯收入的 25%、50% 和 75% 的分位数估计量与不加入抽样信息的结果是不同的，其中 25% 的分位数估计量加权前后有所提高，75% 的分位数估计量加权前后有所降低，说明低收入人群过抽样，而高收入人群样本量少，某种程度上也与高收入人群的拒访率高于低收入人群有关。加权前后标准误差异较大，说明抽样信息对误差的影响较大。由于全国总样本中包含 5 个“大省”，为了更清楚地了解加入抽样信息的家庭人均纯收入的分位数统计量，下面给出加入抽样信息后全国总体以及分城乡的收入组的估计量统计表。

由表 4 可知，加入抽样信息后，低和高收入组，尤其是高收入组的样本量有所增加，而中间收入组的样本量有所减少。加权后，0 ~ 50% 的收入组的均值增加，而 50% ~ 100% 的收入组的均值减少，说明权数调整对收入组人群的数量和均值估计量都有较好的调整作用。加入抽样信息后，抽样误差变化较大，尤其是低和高收入人群中，反映了这两部分人群样本量不足，抽样信息对收入组的误差估计量的作用很大。分城乡后，无论是否加权，城镇的收入均值都要高于农村，但是加权后低和高收入人群的收入差异有所减少。进一步使用列联表对收入组进行描述性统计分析。相比于传统的列联表，加入抽样信息的列联表需要考虑权数等抽样信息的影响，如表 5 所示。

表 5　　CFPS 2010 加权前后的收入组估计量

收入组	收入比例估计量(使用抽样信息)				收入比例估计量(不使用抽样信息)		
	百分比	加权百分比	标准误	设计效应	百分比	标准误	设计效应
0 ~ 25%	25.8	25	1.5	16	24.7	0.4	1
25% ~ 50%	23.5	24.4	0.7	4	24.6	0.4	1
50% ~ 75%	24.2	25.6	0.8	5	25.7	0.4	1
75% ~ 100%	26.5	25	1.6	19	25	0.4	1

由表 5，加入抽样信息后，权数对收入组尤其是高收入组的影响较大，说明高收入组样本观察值不足，标准误差也较大。加入抽样信息前，各个收入组的标准误约为 0.4，但是加入抽样信息后，标准误差差异较大，尤其是低和高收入组标准误差差异较大，设计效应增加较大，说明抽样信息对其影响显著。通过独立性卡方检验方法检验 4 个收入组之间是否有差异，传统的 Pearsons 卡方检验方法假定样本满足独立同分布，在复杂抽样设计下，由于层、群和权数的影响，需要对这个方法进

行调整。调整的方法有两种,一种是由 Fellegi(1980) [6] 提出的传统卡方值的平均设计效应修正,另一种是使用 Rao、Scott(1984) [7] 提出的 Rao-Scott 修正方法。此处使用 Rao-Scott 修正方法检验,当不使用抽样信息时,收入组的卡方检验的 P 值是 0. 29,而加入抽样信息后,Rao-Scott 检验的 P 值是 0. 93(>0. 29) ,说明不使用抽样信息,容易犯第一类错误。同样的方法,检验城乡变量的差异,不加入抽样信息时,城乡的卡方检验的 P 值小于 0. 001,说明城乡之间存在差异。但是,在使用抽样信息时,Rao-Scott 检验的 P 值是 0. 63,城乡之间是没有差异的。可见,在进行假设检验时,抽样信息是十分重要的。

3. CFPS 2012 家庭人均收入的数据分析。

CFPS 2012 包含截面数据和追踪数据的分析,根据分析目的不同,权数的使用如表 3 所示。由于 2012 年是对 2010 年样本的追踪,不涉及抽样,因此其层和群与 2010 年相同。

表 6 是对 2010 年和 2012 年的家庭人均纯收入(对通货膨胀进行了调整之后的、与 2010 年可比的家庭人均纯收入) 的均值和中位数进行复杂数据分析。在分析时,将 2010 年和 2012 年的家庭人均纯收入视为单独的截面数据,分别使用的是 2010 年和 2012 年的全国截面权数。

表 6　　CFPS 2010 和 CFPS 2012 家庭人均纯收入的截面数据估计量比较

	加入抽样信息			不加入抽样信息		
	2010 估计量(标准误)	2012 估计量(标准误)	比率(2012/2010)	2010 估计量(标准误)	2012 估计量(标准误)	比率(2012/2010)
均值	9842(517)	11726(416)	1. 19	10254(153)	12022(190)	1. 17
25% 的分位数	3059(148)	3441(165)	1. 12	2997(31)	3562(67)	1. 19
50% 的分位数	5998(268)	8111(252)	1. 35	5998(52)	8162(86)	1. 36
75% 的分位数	11222(603)	14937(466)	1. 33	11931(124)	14999(151)	1. 26

由表 6 可知,加入抽样信息前后,2010—2012 年的家庭人均纯收入都有所增加,总体均值加权前后增加幅度由 17% 变为 19% ,其中 25% 分位数均值的增加幅度由 19% 下降到 12% ,50% 分位数的增加幅度基本不变,但是 75% 分位数的增加幅度由 26% 变为 33% ,增加幅度最大,说明加权后对低收入和高收入人群的影响最大。从标准误看出,加入抽样信息后标准误的变化较大,尤其是对 25% 和 75% 人群的影响较大,说明这两部分人群的异质性较大,进一步也说明抽样信息的影响较大,忽视抽样信息是不可取的。

为了进一步了解 2010—2012 年收入组人群的变化情况,在研究中使用 2012 年的追踪权数对 2010—2012 年收入组变动情况进行分析,如表 7 所示。

表 7　　CFPS 2010 至 CFPS 2012 加入抽样信息前后收入组变动统计表　　(%)

2010 年收入组	2012 年收入组(加入抽样信息)				2012 年收入组(不加抽样信息)			
	0 ~25%	25% ~50%	50% ~75%	75% ~100%	0 ~25%	25% ~50%	50% ~75%	75% ~100%
0 ~25%	45(1. 23)	26(0. 81)	18(0. 86)	11(0. 72)	42(0. 77)	28(0. 70)	19(0. 62)	11(0. 49)
25% ~50%	31(1. 29)	32(1. 11)	25(1. 17)	12(0. 92)	30(0. 84)	31(0. 84)	26(0. 80)	12(0. 60)
50% ~75%	21(1. 10)	25(1. 09)	31(1. 06)	23(1. 12)	20(0. 72)	25(0. 78)	31(0. 83)	25(0. 78)
75% ~100%	12(1. 14)	14(0. 89)	22(1. 20)	52(2. 05)	11(0. 58)	13(0. 61)	21(0. 75)	55(0. 92)

注: 由于 2010 年和 2012 年的家庭人均纯收入数据有缺失,此处保留 2010 年和 2012 年同时存在的数据,在实际分析中需要进行插补。表中括号外的数字表示 2010 年各个收入组人群在 2012 年的收入组的比例估计量,括号内的数字表示比例估计量的标准误。

由表 7 可知无论是否加入抽样信息,仍旧保留在原收入组的比例最高,但是加权后低收入组的比例从 42% 增加到 45% ,而高收入组的比例从 55% 下降到 52% 。不同收入组的变动,加入抽样信息后高收入组向低收入组变动的趋势增加,其中变化最大的收入组(25% ~50%) 变动到 2012 年的低收入组(0 ~25%) 的比例从 30% 增加到 31% ,向高收入组变动的趋势有所减少,其中低收入

组(0～25%)变动到2012年的低收入组(25%～50%)的比例从28%下降到26%。加入抽样信息后,标准误差都有所增加,其中高收入组的误差增加最高,说明使用抽样信息是重要的。

在模型分析中,由于传统模型中多假定样本满足独立同分布,这个假定在复杂调查数据中通常不成立,因此需要对传统的模型估计和模型假定方法进行修正。否则,由于复杂抽样带来的变异性和自由度的改变,会导致模型的参数估计量有偏,统计结论错误。对于复杂调查数据的模型分析,一方面要结合复杂抽样设计的特征,将抽样信息加入模型中,这里的抽样信息不仅包含抽样设计中的层、群和设计权数,还包含无应答和事后分层调整中的层和权数。另一方面,需要充分考虑抽样信息,在模型的参数估计中考虑加权估计量,在模型的参数检验中使用 Wald 检验。

四、结论

本文主要对复杂抽样设计的特征及其抽样信息在复杂调查数据中的应用进行研究,说明分层、多阶段、整群和不等概率的抽样设计对调查数据影响很大,进行复杂调查数据分析时应考虑层、群和权数等抽样信息,并以 CFPS 为例,说明抽样信息在复杂调查数据中的应用。主要结论如下:

(1)在对复杂抽样进行数据分析时,需要充分了解复杂抽样设计信息,权数有助于得到无偏估计量,群和权数增加抽样误差、降低估计精度,层增加估计精度,只有包含所有抽样信息,才能既得到无偏估计量,又得到准确的标准误估计量。

(2)国内的抽样调查数据需要给出规范的文件介绍抽样设计过程,及复杂数据分析需要的层、群和权数等信息,以便数据使用者正确分析复杂调查数据。值得注意的是,复杂抽样数据中的层和群是指误差层和误差群,需要根据抽样设计和实际抽样过程得到,一般由抽样设计人员给出。

(3)在复杂抽样设计中,抽样设计人员需要考虑到复杂数据分析的需求,采用尽可能简单的抽样设计,方便数据使用者了解设计方案、数据结构和提取抽样信息。必要时,对数据使用者和潜在的数据使用人群进行培训。

(4)复杂抽样中样本数据通常不满足独立同分布的假定,在模型分析时,同样需要将抽样设计中的层、群和设计权数,及包含无应答和事后分层调整中的层和权数加入模型,在参数检验中,采用 Wald 检验。

参考文献

[1] Kish L. Survey Sampling [M]. New York: Wiley,1965.

[2] 金勇进,杜子芳,蒋妍. 抽样技术 [M]. 北京:中国人民大学出版社,2012.

[3] 吕萍. 重权数在复杂调查的方差估计中的应用 [J]. 统计研究,2011(2):93－99.

[4] Yu X, Lu P. The sampling design of the China Family Panel Studies [J]. Chinese Journal of Sociology, 2015, 1(4): 471－484.

[5] 金勇进,邵军. 缺失数据的统计处理 [M]. 北京:中国统计出版社,2009.

[6] Fellegi I P. On Adjusting the Pearson Chi-Squared Statistic for Cluster Sampling [J]. Journal of the American Statistical Association, 1980(71): 665－670.

[7] Rao J N K, Scott A J. On Chi-Squared Tests for Multi-Way Contingency Tables with Cell Proportions Estimated from Survey Data [J]. Annals of Statistics, 1984(12):46－60.

作者简介

吕萍,女,2009 年毕业于中国人民大学统计学院,获经济学博士学位,现为北京大学中国社会科学调查中心副研究员。研究方向为统计调查技术和数据分析。

(责任编辑:曹　麦)