人口流迁

流动人口调查抽样的实践与思考*

庄亚儿 李伯华

【内容摘要】文章对流动人口调查中抽样框的编制、重复抽选及抽选后合并、目标与抽样总体关系的复杂化等问题进行了描述和讨论。实践表明,在编制流动人口调查抽样框时,研究人员不仅需要得到相应数据库的支持,而且还应特别注意偶然性因素的影响;对规模过大或过小的抽样单位,重复抽选或抽选后合并是比较简便的方式;此外,在以"登记户籍人口"作为目标总体的调查中,受"跨界流出人口"的影响,特定地区目标总体与抽样总体的关系将表现为"一对多"的形式,数据处理时必须进行"事后加权"。为保证"事后加权"有效性,需要考虑不同省(区、市)人口流动的特点,特别是样本中"省际"和"省内"流动人口的规模。"2013 年全国生育意愿调查"数据处理的过程表明,不同省(区、市)"事后加权"的分组变量及分组数量并不完全相同,具有一定的灵活性。

【关键词】流动人口;抽样框;重复抽选及抽选后合并;目标总体与抽样总体的关系;事后加权

【作者简介】庄亚儿,中国人口与发展研究中心副研究员;李伯华,中国人口与发展研究中心研究员。北京:100081

Reflections on Sampling of Migrants Surveys

Zhuang Yaer Li Bohua

Abstract: This paper describes and discusses issues on the establishment of sampling frame, replicate selection and linking after selection, the relation between target population and frame population in migrants surveys. Experience shows that when preparing the sampling frame of migrants surveys, researchers not only need the support of relevant database, but also should pay particular attention to the effects of accidental factors. For oversized or undersized sample units, a relatively simple approach is to replicate selection, or link sample units after selection. In addition, in a survey whose target population is the "Registered-household Population", the relation between target population and frame population in specific areas will appear as "one-to-many" as a result of impacts of cross-border migrant outflows. Thus, post-weighting is necessary when processing data. To ensure the effectiveness of the post-weighting, the characteristics of population flows in different provinces should be taken into account, especially the size of floating population both across and within province in the survey sample. Data processing of 2013 National Fertility Desires Survey indicates that the grouping variables and number of groups of post-weighting in different provinces are not exactly the same, but subject to some flexibility.

Keywords: Migrants, Sampling Frame, Replicating Selection and Linking after Selection, Relation between Target Population and Frame Population, Post-weighting

Authors: Zhuang Yaer is Associate Research Fellow, China Population and Development Research Center; Li Bohua is Research Fellow, China Population and Development Research Center.

* 感谢国家科技支撑计划项目"人口与发展数学模型与综合决策支持系统"(编号:2012BAI40B01)对本研究的资助。

中华人民共和国 2012 年国民经济和社会发展统计公报显示，截止 2012 年末，中国大陆的流动人口总数已经达到 2. 36 亿人(不包括市辖区内的“人户分离”人口) ，占同一时间全国大陆总人口的 17. 4% ，约相当于每 6 个人中就有 1 个流动人口。

由于流动人口在总人口中占有相当的比例，以及作为流动人口的主体，即从农村流入城镇的这部分人群兼有城乡居民的某些特点，因此，在人口社会领域所进行的调查中，为有效代表各类群体，将流动人口作为一个独立的“层”明确划分出来，对其比例进行更精确的控制是十分必要的。

在人口社会调查中，以下三种调查均涉及到流动人口: 一是“现住流动人口调查”，其目标人群包括现住特定区域范围内的“区域内流动人口”、“跨区域流入人口”，这种调查即通常所说的“流动人口调查”。另一种是“现住人口调查”，其目标人群包括现住特定区域范围内的“户籍人口”、“区域内流动人口”、“跨区域流入人口”。第三种是“登记户籍人口调查”，其目标人群为特定区域范围内所有登记的户籍人口，共包括三组人群，即现住特定区域范围内的“户籍人口”、“区域内流动人口”、现住该区域范围之外的“跨区域流出人口”。

流动人口空间分布具有易变和不均匀性的特点，其抽样框的编制和样本抽选的难度较大。此外，在上述“登记户籍人口调查”中，由于“跨区域流出人口”只能从本区域之外的各个“分抽样框”中抽选，因此对于一个特定范围的目标总体来说，其与抽样总体在空间上将出现“一对多”的复杂关系，数据处理时必须进行“事后加权”。鉴于 31 个省(区、市) 人口流动的特点存在明显差异，“事后加权”的分组很难做到完全一致。本文对抽样框的编制、重复抽选及抽选后合并、目标总体与抽样总体关系的复杂化问题进行了描述和讨论。

1　抽样框的编制

在抽样领域，随机的概念被表述为“总体中的每个单位均以一个已知的非零概率进入样本”(M. Bulmer，1983) 。实际上，由于人口流动的影响，这一“已知的非零概率”经常处于一种不稳定状态，因此在所有涉及流动人口的调查中，如何尽量减少这种不稳定性的影响，使得所选择的样本基本能够满足调查目标的需要，是调查组织者在整个调查过程中始终必须面对的重要问题。

抽样框是保证样本随机性的重要基础。世界生育率调查组织(WFS，1975) 曾提出过关于抽样框的 5 条基本标准，即无遗漏、无重复、最新、单位的界线(地理边界，社会标准) 清楚、现场易识别。此外，还提出了 2 条附加标准: 单位规模及与调查变量有关的特征。

对流动人口调查来讲，在上述 7 条标准中，“无遗漏”、“无重复、“最新”这几条较为关键，其中尤以“最新”更加重要。只有使用“最新”的抽样框，才能尽量缩小抽样框与实际情况之间的差距，否则有可能陷入抽样难以进行或者现场调查无法实施的困境。

为了深入了解流动人口的生存发展状况，及时掌握流动人口特点和变动趋势，密切关注流动人口基本公共服务政策落实情况，2009 ～2013 年，国家卫生计生委(原国家人口计生委) 连续 5 年进行流动人口动态监测调查。为了满足不仅对全国、而且对各省(区、市) 具有代表性的要求，近年的流动人口动态调查在全国 31 个省和新疆生产建设兵团进行，采取了分层、多阶段、与规模成比例的 PPS 抽样方法(即 Probability Proportionate to Size Sampling) 。实践表明，为了获得质量较高的抽样框，除了对全国流动人口数据库及时进行更新之外，还应当注意两个方面的问题: 一是需要选择恰当的调查时间，在流动人口处于“相对稳定”状态时启动包括编制抽样框在内的各项调查程序; 二是应当尽量缩短抽样框的编制时间和样本抽取时间，尽可能地降低偶发性因素对流动人口分布的影响。

1. 1　调查时间的选择

影响流动人口分布的因素可分为长期性和阶段性两类。长期性因素包括经济、政策、环境、交通状况等方面，这些因素由于相对稳定，或者即使有变化，一般也属于缓和渐变，因而基本上不会对流动

人口调查抽样框的编制产生太大影响。阶段性因素又可分为规律性和偶发性两类,前者与调查时间的选择有关,后者因具有突发性特点,对抽样框的编制和现场调查的影响最大。

法定假日、季节变化、学校寒暑假属于具有一定规律的因素,虽然对流动人口的分布会产生较大影响,但可以设法避开。

法定假日对流动人口分布的影响是全国性的。季节变化的影响主要发生在冬季,受影响的主要是从事户外劳动的流动人口,具有明显的区域性特点。至于学校寒暑假对流动人口分布的影响,则既是全国性的,又呈现出局部性特征:假期时,学校周边的流动人口减少,开学后则回升。原国家人口计生委,现国家卫生和计划生育委员会组织的"全国流动人口动态监测调查"以当年 3 月底的流动人口数据库为依据编制抽样框,5 月份开始现场调查。实践表明,这样的时间安排是比较恰当的,避开了法定假日、季节变化、学校寒暑假对流动人口分布的影响,是开展大型流动人口调查的"黄金时段"。

1.2 影响流动人口分布的偶发性因素

影响流动人口分布的阶段性偶发性因素很多,以下列举其中比较常见的几种情况:

(1)自然灾害。地震、洪水等自然灾害有可能导致局部地区的调查工作无法进行。调查的组织者在调查进行过程中应密切注意相关信息,并根据灾害的严重程度,决定是否可以通过调整抽样框继续进行调查。如果受灾害影响的地区范围较大,无法通过调整抽样框继续进行调查,可按照世界生育率调查组织的相关建议,缩小原定目标总体的范围(WFS,1977)。

(2)集体性搬迁。在城镇地区,集体性搬迁是导致流动人口数量和分布发生变化较常见的偶发性因素,但除了少数例外,其影响的范围一般比较局限。

(3)企业停工。企业因经营不善、污染环境或其他原因而暂时或永久停工,原有的流动人口数量显著减少。

(4)工程的开工与竣工。某项工程开工可以导致流动人口数量增加,而某项工程竣工(如修筑道路、建造房屋)则可引起流动人口数量的减少。为了尽可能地降低偶发性因素对流动人口分布的影响,不仅在启动抽样程序之前,而且在多阶段抽样的各个阶段均需要对流动人口的现况进行核查。这项工作虽然艰苦、耗时,但却是不可或缺的重要环节。因为现况数据是基础,只有基础数据的质量提高了,进一步了解和掌握受偶然性因素影响所发生的变化才有意义。有一个例子可以说明核查工作的重要性。某市在核查村(居)委会一级抽样框时,发现原来填报的流动人口数与实际居住在本地的流动人口数出入较大,通过核查发现,这种差错与一个大型物流公司有关。该公司的雇员主要由流动人口组成,公司的注册登记地虽然属于该居委会,但由于工作性质关系,大部分人并不居住在这个居委会,而是分散在全国各地,属于流动人口中的"流动人口"。对这种情况,应当只填报目前居住在本居委会的流动人口,而不是该公司招聘注册的全部流动人口。

除了对流动人口的现况进行核查外,为及时掌握偶然性因素对流动人口的影响,还需要对未来一段时间人口流动趋势有所了解。集体搬迁、企业停工、工程开工或竣工等情况虽具有一定突然性,但也不可能"毫无征兆"。调查组织者应当和相关部门或单位保持联系,以便对核查时尚未发生,有可能在核查后或现场调查期间发生的情况事前就做到"心中有数",并采取相应措施,包括在最小范围内对抽样框进行必要调整,以保证抽样和调查工作顺利进行。

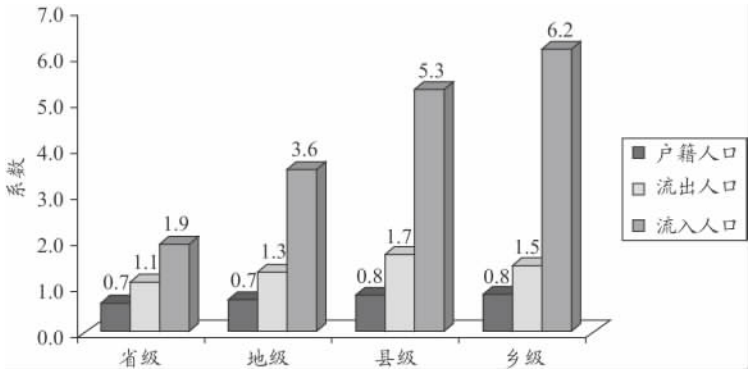
2 重复抽选及抽选后合并

社会调查领域的多阶段抽样有多种不同的组合形式,就其与人口规模的关系来说,大致可分为三类:一类是在各个阶段均考虑人口规模,另一类是仅在某个阶段考虑人口规模,第三类是在各个阶段均不考虑人口规模。流动人口调查,特别是大型流动人口调查,经常采用第一种类型的多阶段抽样,即通常意义上的 PPS 抽样。

在流动人口调查中采用 PPS 抽样主要基于两个方面的考虑，一是为了获得一个自动加权的样本，另一方面则是由流动人口分布高度的不均匀性所决定的。这种高度的不均匀性使得一般的名录抽样很难实现对样本量的有效控制，而且那些没有流动人口的单位还将导致无效抽样。采用 PPS 抽样，虽然抽样框的编制比较复杂，但由于各抽样单位的规模是已知的，因而有利于实现对样本量的有效控制，并避免无效抽样。

图 1　不同级别行政区域户籍、流出、流入人口变异系数(全国)

Figure 1　Coefficient of Variation of Registered-household Population, Out-migrant Population and In-migrant Population at Different Levels of Administrative Region



数据来源: 根据原国家人口计生委发规信息司 2011 年全国乡(镇、街道) 级抽样框数据计算结果。

图 1 显示，与户籍人口相比，由于流入人口对工作、居住地点的选择目的明确、更加灵活，因而在省(区、市) 、市(地、州) 、县(市、区) 、乡(镇、街道) 四个级别上，其变异系数(此处指人口规模的标准差与其平均数的比值) 均显著大于户籍人口。另一方面，越是在较低的行政区划级别上，流入人口的变异系数越大，全国乡(镇、街道) 一级流入人口的变异系数甚至超过了 6，说明不同乡(镇、街道) 流入人口的规模差别巨大(村民和居民委员会的情况也类似) ，因而在抽样过程中需要对规模过大或过小的单位采取一些特别的方法。在这方面，有关文献提供了多种选择(Leslie Kish，1965) ，重复抽选或抽选后合并是其中比较简便有效的方式。

抽样单位的“重复抽选”，指的是对那些规模较大的抽样单位允许其被多次抽选。一般来说，当这种规模较大的群的数量较少时，“重复抽选”是较好的选择。与农村地区相比，城镇地区出现“重复抽选”情况的可能性更大; 与主城区相比，城乡结合部出现“重复抽选”的可能性更大。

抽样单位的“抽选后合并”，指的是对那些虽然已经被抽中，但因其规模较小，不能满足 PPS 抽样对样本量的要求，因此需要将其与相邻的其他单位加以“合并”再作为一个单位入选样本的情况。一般来说，当规模较小的群的数量较少时，“抽选后合并”是较好的选择。从实际情况来看，“抽选后合并”主要发生在流入人口较少的农村地区，而且主要出现在村委会这一级。

与重复抽选相比，“抽选后合并”的操作过程稍显复杂，因为前者所面对的仅仅是规模较大的某个单位本身，而后者则除了需要面对已经被抽中的某个规模较小单位之外，还需要考虑与哪一个其他单位相互“合并”的问题。

具体操作过程中，“抽选后合并”应遵循“近邻原则”。所谓“近邻原则”，指的是应以抽中的规模较小单位为中心，与其相邻的未入选单位相合并，以达到所要求的样本规模，而不是跳过“近邻”，随意寻找一个“远亲”进行合并。

需要指出的是，某些情况下，如果与相邻的某个单位“合并”后仍不能满足对样本量需要，就必须

扩大近邻的数量和范围，将几个单位“合并”为能满足样本量需求的一个有效样本单位。

表 1 按地区划分的乡(镇、街道)级流入人口变异系数

Table 1 Coefficient of Variation of In-migrant Population at Township Level

地区	流入人口变异系数	地区	流入人口变异系数
北京	1.6	湖北	3.8
天津	2.5	湖南	4.2
河北	2.5	广东	3.1
山西	2.6	广西	4.5
内蒙古	3.7	海南	3.7
辽宁	4.0	重庆	3.4
吉林	2.4	四川	5.5
黑龙江	5.1	贵州	4.5
上海	1.0	云南	3.8
江苏	3.7	西藏	7.5
浙江	2.0	陕西	4.1
安徽	3.4	甘肃	4.0
福建	3.1	青海	3.5
江西	4.6	宁夏	3.0
山东	3.6	新疆	2.5
河南	5.7	新疆兵团	1.7

数据来源：根据原国家人口计生委发规信息司 2011 年全国乡(镇、街道)级抽样框数据计算结果。

经验表明，在对那些流入人口较多、且变异系数较小的地区抽样时，往往不需要经过“抽选后合并”即可一次完成抽样；而在对那些流入人口较少、且变异系数较大的地区抽样时，通常需要通过“抽选后合并”的方式才能完成抽样过程。表 1 是按省(区、市)划分的乡(镇、街道)一级流入人口变异系数分布。一般来说，在流入人口变异较小的地区，如上海、北京抽取乡(镇、街道)时，由于各乡(镇、街道)的流入人口分布相对均匀，因而基本上不存在“重复抽选及抽选后合并”问题，抽样过程相对简单；而在那些流入人口变异较大的地区，如河南、四川、黑龙江抽取乡(镇、街道)时，遇到“重复抽选及抽选后合并”的可能性明显增加，抽样过程相对复杂(出于工作实际的考虑，西藏的流动人口调查仅在拉萨市进行，拉萨市乡(镇、街道)流入人口的变异系数为 5.1)。

3 目标总体与抽样总体关系的复杂化

抽样文献在论述总体的分类时，一般将总体分为推论总体、目标总体、抽样(范围)总体、调查总体共四类(Leslie Kish，1987)，指出这四类总体的大小关系类似于一个倒金字塔，推论总体最大而调查总体最小。在谈到目标总体与抽样总体的关系时，认为两者的差别主要受“未覆盖”的影响，而抽样总体与调查总体之间的差别，则主要受“无应答”的影响。图 2 为目标、抽样、调查三类总体之间关系示意图。在上文列举的“现住流动人口调查”、“现住人口调查”中，虽然它们的目标人群不尽相同，但两者有一个共同点，即目标总体与抽样总体的区域范围是一致的，与图 2 所表明的关系相同。

2013 年，国家卫计委进行了一次全国性生育意愿调查，调查范围为大陆除西藏、新疆之外的 29 个省(区、市)，要求调查结果不仅对全国，而且对调查省(区、市)也具有代表性。该调查的目标总体为“登记户籍人口”，其调查抽样的复杂性在于，虽然某个省(区、市)“登记户籍人口”中的“现住户籍人口”和“省内流动人口”样本可以从该省(区、市)的相应抽样框抽选，但“登记户籍人口”中的“跨省流出人口”却必须从其他 30 个省(区、市)的“流入人口抽样框”抽选。

图 2　目标、抽样、调查三类总体关系示意图

Figure 2　Schematic Illustration of the Relationship between the Target, Frame and Survey Populations

以安徽省为例，图 3 为该调查中目标总体与抽样总体之间对应关系示意图。

图 3　登记户籍人口调查中目标、抽样、调查三类总体关系示意图

Figure 3　Schematic Illustration of the Relationship between the Target, Frame and Survey Populations for Registered-household Population

在图 3 中，安徽省“登记户籍人口”这一目标总体除了与包括本省“现住户籍人口”和“省内流动人口”的抽样总体 A 对应外，还与安徽省流动至其他 30 个省（区、市）中的“跨省流出人口”抽样总体 Bx 相联系（x 代表其他 30 个省、区、市），从而形成了“一对多”的状况。

在目标总体与抽样总体关系复杂化，出现“一对多”的状况下，利用其他 30 个省（区、市）的流动人口“抽样比”，是无法将安徽省流动至该省（区、市）的“跨省流出人口”直接换算为安徽省“登记户籍人口”中的“跨省流出人口”的。在此情况下，需要先将分散在其他 30 个省（区、市）的安徽省“跨省流出人口”汇总，然后再按照第六次人口普查时该省“登记户籍人口”中的“流动与非流动人口比例”，对该省样本数据（流入安徽省的“跨省流入人口”除外）进行“事后加权”，才能进行相关指标的计算。

在进行上述“事后加权”时，必须考虑到不同地区人口流动的特点，特别是各省（区、市）“省际”和“省内”流动人口的规模。由于该项调查中各省（区、市）所调查的流动人口样本量相同，因此在进行“事后加权”时，为保证加权结果的有效性，所有分组的绝对数均不为零，也不能过小，各分组的权数应尽量控制在 2 以内。为满足这些条件，各省（区、市）的“事后加权”分组变量和分组数很难完全相同，实际操作中具有一定的灵活性。生意意愿调查具体的“事后加权”过程共分为以下几个步骤：

（1）根据 2013 年全国生育意愿调查以及 2010 年第六次人口普查的问卷内容，将生育意愿调查对象的性别、年龄（按 15～29、30～44 岁划分为两组）、户口性质（农业、非农业）、是否流动人口（流动人口、非流动人口）作为事后加权的基本分组变量，交叉分组数共 16 个。

（2）根据各省（区、市）的人口流动特点确定各自的事后加权分组变量和相应的分组。

按上述所有 16 个分组划分，能够保证“事后加权”的有效性。在所调查的 29 个省（区、市）当中，

安徽、河南等 20 个省(区)属于这种类型。

按 16 个分组划分,不能保证"事后加权"的有效性,分组权数的差别也显著加大。对于这些地区,采取了舍去"性别"变量的做法(以往和本次调查的结果表明,在其他条件相同的情况下,不同性别调查对象生育意愿的差别很小),实际的事后加权分组数减少到 8 个。在所调查的 29 个省(区、市)当中,吉林、黑龙江等 6 个省(区)属于这一类型。

北京、天津、上海属于另外一种情况。由于这三个直辖市"跨市流出人口"数量很少,而且按照"人户分离"的规定,样本中"市内流动人口"的数量也很少。在此情况下,舍去了"是否流动人口"变量,按性别、年龄、户口性质共 8 个交叉分组进行了"事后加权"。

(3)分别计算出各省(区、市)的权数以及全国权数。

在按抽样比计算权数时,权数等于抽样比的倒数。但在"事后加权"中,使用的是"比例法",即调查样本各分组在全部样本中所占比例乘以权数之后应当等于普查数据中的同一比例。

在将各省(区、市)的数据加权为全国数据时,需要将本次调查中各省(区、市)的"登记户籍人口"在全部样本中的比例逐一转换为第六次人口普查中的相应比例。

(4)不论是各省(区、市)还是全国,未加权之前的实际调查人数必须与加权之后的相应人数相同,为此就需要对有关的权数进行"标准化",即在未加权状态下将原有的权数逐一除以其平均数,并将这一结果作为最终权数。

4 结语

以上对流动人口调查中抽样框的编制、重复抽选及抽选后合并、目标与抽样总体关系的复杂化等问题进行了描述和讨论。毋庸讳言,目前的流动人口调查抽样方法尚有待改进,比如,针对不同省(区、市)人口流动的特点如何采取不同的抽样方法?在村(居)委会一级如何将地图抽样和名录抽样相结合,以保证"非正规居所"中流动人口入选样本的机会?这些问题均需要在今后的实践中继续进行探索。

参考文献/References:

1 国家统计局. 中华人民共和国 2012 年国民经济和社会发展统计公报. www.stats.gov.cn/tjgb/ndtjgb/qgndtjgb/t20130221_402874525.htm

National Bureau of Statistics of the People's Republic of China. Statistical Communiqué of the People's Republic of China on the 2012 National Economic and Social Development. www.stats.gov.cn/tjgb/ndtjgb/qgndtjgb/t20130221_402874525.htm

2 M. Bulmer and D. P. Warwick. 1983. Social Research in Developing Countries. JOHN WILEY & SONS: 91.

3 World Fertility Survey(WFS). 1975. Manual on Sample Design. The Hague, Netherlands, International Statistical Institute (World Fertility Survey Basic Documentation No. 3): 23.

4 World Fertility Survey (WFS). 1977. Guidelines for Country Report No. 1. The Hague, Netherlands, International Statistical Institute(World Fertility Survey Basic Documentation No. 8): 175-186.

5 Leslie Kish. 1965. Survey Sampling. John Wiley& Sons: 43-246.

6 Leslie Kish. 1987. Statistical Design for Research. John Wiley & Sons: 28-29.

(责任编辑:沈 铭 收稿时间:2013-10)