国内大型社会调查数据中的流动人口结构比较分析*

周 皓　雷琳旋

【摘　要】数据是社会科学研究的基础，决定着研究结论的可靠性。文章比较了国内6个大型社会调查中“流动人口”的定义，探讨了各流动人口样本结构与全国总体的差异，并基于相同的模型设置呈现了样本结构对统计结果的影响。研究结果表明：(1)不同社会调查对流动人口的定义在诸多方面存在较大差异，对应不同的研究总体；(2)不同调查估计的流动人口比例差异较大，且省际和省内县际流动的比例远低于小普查；(3)各调查中的流动人口样本，在性别、年龄、受教育程度、城乡等基础变量上存在较大的结构性差异；(4)在相同模型设定下，分别以受教育年限和是否为流动人口为因变量构建回归模型，基础变量在显著性和方向上存在显著的差异；统计结果同时反映了社会现实的共性和样本结构性偏差。文章还进一步讨论了测量对研究对象、测量结果和统计结果的重要性。文章建议，各调查实行统一的流动人口测量方式以统一研究对象。

【关键词】抽样调查　流动人口　测量　样本结构

【作　者】周 皓　北京大学中国社会与发展研究中心、社会学系，教授；雷琳旋　北京大学社会学系，博士研究生。

一、引　言

经验知识的完备性和真实性是认识纯粹知识的基础，是因果推断的重要基础。抽样调查作为社会科学研究收集数据的重要方法之一，是经验性材料与知识的重要来源；调查数据的质量直接影响研究的可复制性、结论的可比性和统计推断的可靠性。在社会科学研究中，调查数据扮演着连接理论与经验、局部与整体的桥梁角色，在认识论上涉及从局部知识构建对整体的理解。测量和代表性是影响抽样调查结果完备性和真实性的两个重要方面。只有在测量工具有效、样本具有代表性的前提下，社会科学研究才能为理论构建和政策制定提供可靠的知识基础。

* 本文为教育部人文社会科学重点研究基地重大项目“中国人口长期均衡发展关键问题研究”(编号：22JJD840001)的阶段性成果。

当前，大型抽样调查已成为社会科学研究的重要手段，为描述和分析社会现实提供了重要的信息源。在以调查数据为基础的定量研究中，测量和样本代表性综合体现在样本结构中，样本结构的偏差会影响分析结果，导致因果推论中实验效应估计量及相关变量的系数估计量存在偏差（周皓，2023）。因此，认识大型社会抽样调查的样本结构是正确认识与分析社会现实的基础。

抽样调查中的流动人口数据不仅涉及流动人口的定义和测量（即谁是流动人口），也涉及流动人口样本的代表性。流动人口样本结构的偏差不仅会影响关于人口迁移流动的方向、规模及分布等社会现实的判断，还会影响从原因到后果的各类研究结论乃至后续政策制定的有效性。

目前已有众多研究讨论了流动人口抽样调查方法中存在的问题，如抽样设计方案（沈明明、李磊，2007）、抽样框的构建过程（齐嘉楠等，2014；庄亚儿、李伯华，2014）、调查地点的选择（吕利丹、段成荣，2012；梁玉成等，2015）等。这些讨论有助于深入理解流动人口抽样调查的复杂性和可能的方法论。但到目前为止，鲜有研究从抽样调查结果的角度对比分析国内各大型社会调查中的流动人口样本结构及其差异，这使得基于这些调查数据的不少研究存在严重的隐患。

本文以国内 6 个重要的大型社会调查数据为样本，先梳理比较各调查中流动人口的定义，再比较分析各调查中流动人口的样本结构，以期揭示各调查中流动人口样本之间的差异及可能的统计偏误。本文主要关注以下研究问题：（1）不同大型社会调查中流动人口的定义有何异同？（2）不同社会调查的样本结构与全国流动人口总体结构之间是否存在差异？（3）样本结构差异是否影响统计结果？本文的核心关切不在于质量评价，而在于强调研究者在数据使用和分析中应重视测量与样本代表性的重要性，关注结构性偏差的可能影响，以免被可能存在的偏差所误导。

二、不同调查中流动人口的定义和比例

本文选取当下国内最具代表性、使用最为广泛的 6 个全国性社会调查数据：中国家庭追踪调查（CFPS）、中国综合社会调查（CGSS）、中国家庭金融调查（CHFS）、中国健康与营养追踪调查（CHNS）、中国社会状况综合调查（CSS）和中国劳动力动态调查（CLDS）。为使各调查具有更好的可比性，本文的数据选择与处理考虑以下几点。（1）由于各调查进行的年份各不相同，本文选用 2015 年全国 1%人口抽样调查（以下简称 2015 年小普查）数据为参照数据。（2）各调查均选用 2015 年或相邻年份的调查轮次数据。其中，2015 年的数据包括：CGSS、CSS、CHNS、CHFS；CLDS 和 CFPS 分别选用 2014 和 2016 年的数据。考虑到跟踪调查项目（CFPS、CHNS 和 CLDS）可能因调查过程中样本损耗与流失等问题而导致样本偏差，本文未选用这些调查的最新数据。（3）为使样本能更好地反映对总体

的代表性，本文后续分析均为加权后的结果。

（一）各调查中流动人口定义

本文研究的流动人口特指户籍口径下的流动人口，即离开户籍登记地半年以上、且跨越乡镇街道的人口，但不包括市内人户分离人口。这一定义涉及户口登记状况、空间范围和流动时间 3 个维度。总的来看，这 6 个大型社会调查对流动人口的定义虽然同为户籍口径，但 3 个维度的具体标准并不统一。

首先是户口登记状况。虽然多数调查以“户口登记地”为题项收集信息，将“人户分离”作为判断流动人口的首要标准，但具体的提问方式和选项存在一定的差异。多数调查直接询问“现在的户口所在地”，但 CHNS 问卷中并没有户口所在地信息，只询问了“仍在家住吗？”，即只询问了是否在本地或本户居住，而非户籍登记信息。基于 CHNS 数据的已有文献中识别流动人口的常用方式是将“因外出打工而不在家住”（Tong 等，2012；李琴、宋月萍，2009）或“户口为农业但居住地在城市”（易龙飞、亓迪，2014；樊敏杰等，2023）的样本定义为流动人口。CHNS 的题项设置与普查口径有较大差异，分析时需特别注意。CHFS 则依次询问“户口是否在他 / 她目前所居住的乡镇 / 街道？（或区 / 县，或省区）”，把问题与具体选项结合起来，通过 3 个问题来收集具体的户口登记状况信息。

其次是空间范围。人口普查定义中，只有跨越一定行政区域范围（通常为乡、镇、街道）且非市内人户分离人口者才能被认定为流动人口；行政边界的选取标准会直接影响流动人口的识别范围。CFPS、CGSS、CLDS 和 CSS 中，行政边界均以乡镇街道为基本单位，与人口普查的标准一致；CHFS 中的基本空间单位仍然是乡镇街道，只是它是通过“是 / 否”题项来直接询问、而非作为选项之一；CHNS 的提问形式（是否在家居住）缺乏明确的行政边界范围，无法准确判断人口的空间移动范围，也就无法以空间来区分流动人口类型（如县内、省内县际或省际）。不过，CHNS 在家庭问卷中询问了居住地址，因此，将个体与家庭信息相联系才能够识别具体空间范围。同时，CFPS 在技术文件中说明了地址码只到省级，但会根据出生地和户口所在地的情况合成并公布“相对的地址变量，以说明该地区是否与调查所在地是同一个省 / 市、区 / 县、乡镇 / 街道、村 / 居”。

再次，流动时间。这是流动人口登记的重要维度之一，也是数据分析过程中识别流动人口、确定迁移流动登记对象的重要标准。时间维度的题项既可以针对最近一次迁移提问，也可以对离开户口登记地的时间提问。这两种不同的题项设计具有不同的含义，对应不同的流动时间。这些具体流动时间信息不仅有助于按照流入时间识别流动人口，还可以为后续分析（如在本地居住时间对流动人口社会融合等迁移后果研究）提供基础信息。

在本文考察的 6 项调查中，CFPS 和 CLDS 登记了离开户籍登记地的时间，但具体提问方式不同。CFPS 利用跟踪调查的特点，询问了上一次调查至今居住地址是否发生过变化，同样也询问了最近一次的迁移时间；CGSS 不仅登记了离开户籍登记地的时

间，还登记了来本地居住的时间（对应于最近一次迁移口径，且登记了流出地信息）和户口迁移时间。与之相比，CSS 询问了来本地居住的时间（类似于最近一次迁移的定义）；CHNS 询问了离家时间；CHFS 则询问了是否有过离开户籍省份去其他地方工作的经历，以及返回的时间。各调查中有关流动时间的具体信息也不尽相同。CFPS、CGSS 和 CSS 均询问了流动的具体时间（CFPS 记录了年和月、CGSS 和 CSS 只记录了年份）；CHNS 要求被访者填报具体时长（以月为单位）；CLDS 只记录了是否离开户口登记地半年及以上，未登记具体的流动时间；CHFS 则未登记流动时长。综合来看，上述调查中只有 CGSS 和 CFPS 同时记录离开户口登记地的时间和最近一次迁移时间；CLDS 登记了离开户口登记地的时间，CSS 只登记了最近一次迁移时间；CHFS 和 CHNS 中登记的时间与普查有较大的差异。各调查在时间维度的定义、题项设计等测量方式的不同，导致其流动人口样本所对应的总体各不相同，因此，使用各调查样本中的流动人口规模与比例来推断总体时需要谨慎。

此外，在实际操作中，剔除市内人户分离人口是准确测量流动人口的重要前提。在上述大型社会调查中，CFPS 问卷要求填写“当前居住省份的其他县区”信息，但需要申请更细致的地址信息才能识别市内人户分离人口。CSS 和 CGSS 也需要根据地址码来剔除市内人户分离人口。CLDS 在调查设计中设置程序以区分市内人户分离人口，其调查手册中明确提到“问卷部分专门针对流动人口的问题，属于市区内人户分离的人，系统也会跳过这部分题目，这时候可以填答 99998（不适用）”。相比之下，CHNS 和 CHFS 没有明确的户籍登记地信息，在没有详细地址码的情况下，这两个调查可能会因无法剔除市内人户分离人口而高估省内流动的规模和强度。

最后，迁移史信息很难根据普查数据获得，但在人口迁移流动历程研究中极为重要。限于调查目的，各大型调查在登记迁移史信息时采用的策略各不相同。CFPS 作为追踪调查，直接登记两次调查之间的居住地是否发生变化；CGSS 是多轮截面调查，历次调查登记了离开户口登记地和迁到本地居住的时间（以及对应的迁出地），从而有可能识别出“出生地—户口登记地—最近一次迁出地—本地”4 个地点的 3 次迁移及相应的时间空间范围。CLDS 则是以户口迁移为标准，登记了户口是否曾迁移及迁移的次数，并询问了 14 岁以来的迁移发生时间、迁出地和原因等；由于 CLDS 仅针对户口迁移，调查信息无法反映“流动”过程。类似地，CSS 也只登记了户口迁移的时间，无法反映“流动”过程。与之相比，CHNS 并没有调查迁移流动经历；CHFS 则是从农村流出地的角度调查工作及返乡经历，且调查对象是省际劳动力流动。因此，总体来看，CFPS 和 CGSS 登记了迁移流动史，CLDS 和 CSS 登记了户籍迁移史，CHFS 则含有省际流动返乡的情况；这些信息可以为中国人口迁移转变研究提供丰富的研究素材。

综上所述，本文考察的 6 个大型社会调查中流动人口的定义虽然都遵循了“人户分

离”标准(户籍口径),但各自对于跨越行政边界的范围界定、具体时间信息和识别市内人户分离人口等方面存在很大差异，也与人口普查统计口径之间有一定的区别。在使用这些数据进行实证研究时,特别是多个调查数据之间的比较分析,必须充分考虑这些定义以及样本所代表的总体的差异,以确保比较分析的准确性和一致性。

(二) 各调查中各类流动人口的比例

调查样本中的流动人口比例是推断全国流动人口规模的基础。按照各调查的口径,本文计算了各调查中流动人口的规模与比例(见表 1)。结果显示,不同调查之间差异明显。以流动人口比例为例,CGSS(25.63%)和 CSS(25.76%)高于小普查(21.33%),CHFS(17.71%)和 CHNS(14.41%)略低于小普查,CFPS(10.15%)、CLDS(8.93%)则偏差较大。

从流动距离看,各调查的省际流动人口比例均远低于小普查(7.07%),如 CGSS为 4.91%,CSS 为 5.47%,其他调查中相应比例更低。从省内县际流动的比例看,除 CSS(8.09%)和 CHFS(7.47%)与小普查(8.32%)较为接近外,其他调查均低于小普查。可见,各调查中流动人口的流动距离及地域结构可能与总体存在不同的偏差。

表 1　各调查流动人口样本规模及比例比较

调查名称	总样本(人)	流动人口(人)	流动人口比例(%)	省际流动比例(%)	省内流动比例(%)	省内跨县流动比例(%)	县内流动比例(%)
2015 小普查	21312241	4545865	21.33	7.07	14.26	8.32	5.94
CFPS	38915	3951	10.15	1.44	8.72	2.33	6.38
CGSS	10866	2785	25.63	4.91	20.72	4.35	16.37
CLDS	41227	5676	13.77	3.39	10.38	4.34	6.04
CLDS(个人)	22310	1994	8.93	3.13	5.81	3.90	1.91
CSS	10243	2638	25.76	5.47	20.29	8.09	12.19
CHNS	23937	3449	14.41	-	-	-	-
CHFS	36730	6504	17.71	4.72	12.99	7.47	5.52

注:表中内容根据各调查官网的数据计算得到,且根据个体权重作加权处理。“CLDS(个人)”指仅包括 CLDS 调查的个人库数据;“CLDS”除个人库外,还结合了 CLDS 家庭数据库中的流出地调查。

这种结构偏差既与流动人口的定义有关(如某些调查未包括县内流动人口或无法区分市内人户分离人口),也可能与各调查的抽样过程(特别是抽样框的设定)有关。CFPS 和 CLDS 等跟踪调查还可能因流动人口的“流动性”(样本损耗)而导致后续跟踪调查轮次中相应比例降低。总之,流动人口的定义以及在问卷中的题项与选项等设计均会影响对全国流动人口规模与比例的推断结果,其内部结构性差异也会使各种研究结果存在统计偏差。

三、样本的结构对比

本文重点比较各调查中流动人口的年龄结构、性别结构、受教育结构和城乡结构的

差异。需要说明的是，分析过程的加权处理方法如下：CGSS 和 CSS 分别按照各自数据中的调查设计权重进行加权；CLDS 数据中提供了多种权重，本文选择无回答调整权重进行加权[①]。参照官方使用手册的方法，CFPS 仅对全国总样本进行加权，全国再抽样样本不做加权处理；CHFS 参照《中国家庭金融调查报告》中提供的方法，在个体层次比较时不做加权处理；CHNS 则因数据中未提供权重而无法进行加权处理。

（一）年龄结构

各调查中流动人口的年龄金字塔如图 1 所示。各调查覆盖的年龄范围差异较大：CHNS、CHFS 和 CFPS 收集了 0 岁起的全年龄段人口；CGSS 仅针对 18 岁及以上人口；CSS 针对 18～70 岁人口。CLDS 个人问卷调查的是 15～64 岁劳动年龄人口，本文结合

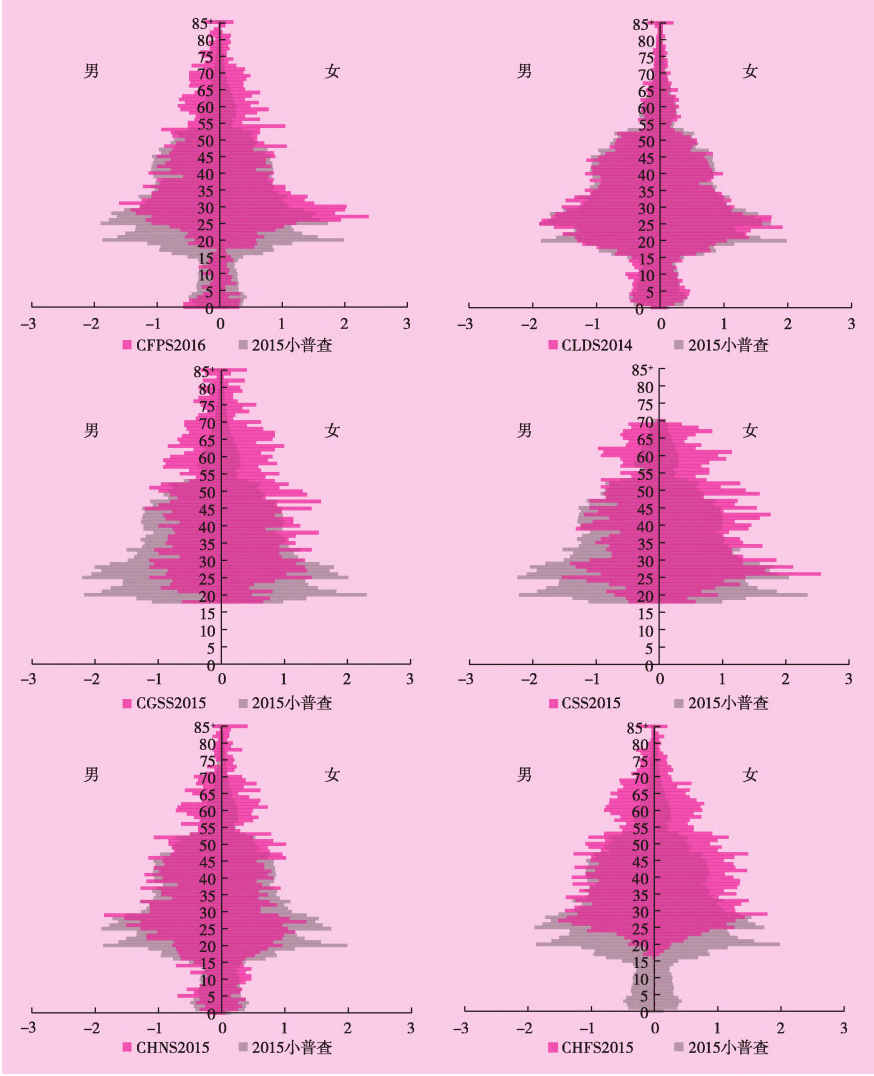

图 1　各调查中流动人口的年龄金字塔

① CLDS 中的无回答权重“wpr”根据受访者的拒答情况计算得到，可以在一定程度上重新察看数据的代表性。数据还提供了以 2010 年普查数据为基础计算得出的事后调整权重“wpp”，此处在与 2015 年小普查数据进行对比时不宜采用。此外，“个体数据中，因劳动力调查的抽样设计为 15～64 岁年龄人群，所以给出的权数为 15～64 岁劳动力个体的权数，此年龄范围之外，无权数”（《CLDS2014 数据 20160520 版本使用说明》）。因此，对于个人库中的流动人口，仅对 15～64 岁个体进行加权；对于在家庭库而不在个人库的个体，按照家庭权重除以家庭人数进行加权，且同样仅对 15～64 岁个体进行加权。

CLDS 家庭数据库中的流出地调查，重构了全年龄人口结构，旨在更全面地考察流动人口特征。由于 85 岁及以上人口较少，本文将 85 岁及以上人群合并为“85+”年龄段。

从图 1 可以得出以下结论。(1)CLDS 中流动人口的性别年龄结构与小普查最接近，仅在 15～19 岁和 20～24 岁组略低，其他年龄组则与小普查较为接近，这可能反映了流出地调查的优势。(2)CFPS 中 0～25 岁流动人口比例低于小普查，26～34 岁女性比例高于小普查，男性则偏低；35～50 岁女性与小普查接近，仅男性略低；50 岁以上男性和女性流动人口所占比例均高于小普查。(3)CHNS 在 0～15 岁年龄段的流动人口比例与小普查基本一致；在 16～30 岁这一流动人口最活跃的阶段，其比例明显低于小普查，在 20 岁左右的差异尤为明显；30～45 岁女性比例显著低于小普查，而男性比例与小普查接近，仅有小幅差异；45 岁以上流动人口比例明显高于小普查，尤其是 55～70 岁年龄组。(4)CGSS 和 CSS 呈现出相似的特点，均表现出 0～25 岁年龄段人口比例大幅低于小普查、55 岁及以上年龄人口大幅高于小普查，以及 30～39 岁的男性流动人口比例明显低于女性的特点；其中，CSS 在 40～59 岁年龄组还呈现出女性人口比例显著高于男性的特点。(5)CHFS 与小普查的主要差异在于 0～23 岁流动人口的严重缺失、30～53 岁年龄段呈现柱状分布特征，以及 30～55 岁女性和 55 岁以上全人群比例显著高于小普查。

为控制各调查覆盖的年龄差异，本文将所有调查的流动人口年龄统一限定在 18～70 岁重新分析，结果显示：(1)这些调查中，年轻段(18～30 岁)的比例均低于小普查；即使是 CLDS，18～22 岁的比例亦相对较低；(2)各调查均呈现出老年段比例相对较高的特征；(3)除 CHNS 和 CLDS 外，其他调查中女性所占比例均高于小普查。总体而言，6 个社会调查中流动人口样本的年龄结构与小普查之间存在较大的差异。

(二) 性别结构

图 2 展示了各调查中流动人口的年龄别(5 岁组)性别比。经过 5 岁组合并后，各调查中不同年龄组的性别比依然波动较大，且差异显著。(1)在 15 岁以下的 3 个年龄组中，除 CFPS 调查的 5～9 岁组、CHNS 和 CHFS 的 10～14 岁组性别比低于小普查外，其他调查在各年龄组均高于小普查。如 CFPS 0～4 岁组的性别比为 159.38，远超小普查的 115.55；CFPS 和 CLDS 在 10～14 岁组性别比分别达 154.55 和 173.56，远高于小普查的 121.05。(2)在 15～64 岁年龄段，除 CLDS 略高于小普查、CHNS 在 20～44 岁组[①]高于小普查以外，其他调查均低于小普查。(3)在 65 岁及以上的老年段，除 CHNS 和 CGSS 相对较低以外，其他各调查均高于小普查。

(三) 受教育结构

图 3 给出了各调查中流动人口的受教育结构。可以看出，各调查中流动人口的受

① CHNS 中流动人口的年龄别性别比在 30～34 岁组形成一个尖峰。

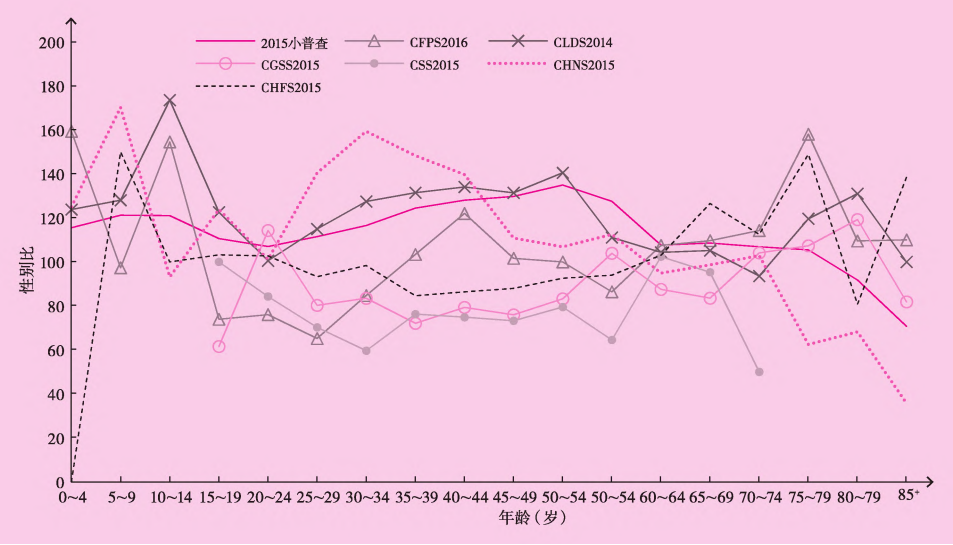

图 2　2015 年小普查与各调查流动人口年龄别性别比

教育结构也与小普查存在一定差异。小普查中未上过学的比例为 2.07%，CFPS 中相应比例高达 20.20%，其他调查则在 3.63% ~ 7.23%之间。CFPS 中未上过学的比例偏高，可能与该调查以家庭为调查对象、涉及更多的老年人口有关。

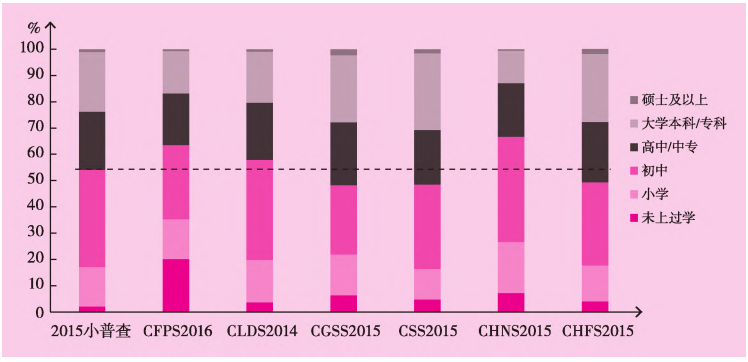

图 3　2015 年小普查与各调查的流动人口受教育结构

总体来看，CLDS 与小普查中流动人口的受教育结构比较接近，仅高中及以上比例略低于小普查。CHFS 中流动人口初中学历比例比小普查低 5.23 个百分点外，其余类别的差异均在 2 个百分点以内。由于 CGSS 中流动人口初中学历比例（26.30%）远低于小普查的 36.91%，导致 CGSS 的其他比例均相对高于小普查。CSS 和 CHFS 表现为小学、初中相对较低，大学和研究生比例略高于小普查。但 CSS 高中比例略低，而 CHFS 的高中比例则略高。

（四）户口属性

各调查中，流动人口样本的户口属性（农业 / 非农）构成同样存在明显差异。CHFS 的户口属性分布与小普查最为接近，虽然其农业人口比例超过 50%（与小普查相反），但农业和非农人口的分布较为均衡，非农人口比例接近 50%（48.08%）。相较之下，CFPS 和 CGSS 的流动人口样本中非农人口比例显著高于小普查，分别达到 62.74%和 60.53%。

与之相反，CLDS、CSS 和 CHNS 则呈现出农业人口比例偏高的特点。其中，CSS 流动人口中农业人口超过 60%，CLDS 中这一比例超过 70%，CHNS 则高达 99.01%。可见，流动人口户口属性在各样本之间同样存在较大差异。

（五）各数据与小普查结构的差异总结

上述比较表明，各调查数据在年龄、性别、受教育、城乡等结构上均与小普查数据存在不同程度的差异。（1）年龄结构：多个调查显示出低年龄段人口比例偏低，高年龄段人口比例偏高、且高年龄段中的女性人口相对较多的特点。（2）性别结构：大体上，低龄段的性别比相对远高于小普查、15 ~ 64 岁人口的性别比相对低于小普查、老年段的性别比则略高于小普查。（这个性别比的差异可能导致在流动儿童或随迁儿童的研究中得到男孩偏好的有偏结果。）（3）受教育结构："低龄组偏低、高龄组偏高"的年龄结构会影响到各调查样本中的教育结构，CFPS、CLDS、CHNS 调查中受教育水平为初中及以下的比例相对较高，而高中及以上的比例相对较低，尤其是大学及以上的流动人口比例；而 CGSS 和 CSS 中的流动人口呈现出偏高受教育水平的结构分布。（4）户口属性：CFPS 和 CGSS 中非农户口的流动人口比例偏高，与之相反，CLDS、CSS、CHFS 和 CHNS 流动人口样本中农业户口比例偏高。

四、样本结构差异对统计分析结果的影响

为进一步探讨上述样本结构差异可能带来的统计偏差，本文选用各调查共有的基础（人口学特征）变量，以统一的模型设置考察各调查中样本结构的多维差异。需要强调的是：（1）模型的分析目的不是探讨因果关系，而是展现不同数据之间样本结构的综合差异及其可能的统计偏差；（2）样本结构偏离是多变量联合分布之偏离，只有同时纳入相同变量才有可能体现这种联合分布的偏离。当然联合分布的基础仍然是单个变量的样本结构。

（一）变量选择与模型设定

本文构建如下两个模型。第一个模型以"受教育年限"为因变量，以性别、年龄、户口和婚姻状况为自变量，讨论各样本内部结构的差异对统计结果的可能影响。第二个模型以各调查全样本为分析对象，以"是否为流动人口"为因变量，以性别、年龄、户口、婚姻状况和受教育年限为自变量，考察流动人口样本在全样本中的相对结构差异及其对统计结果的影响。

选择上述变量的理由包括：（1）共有性，即所有社会调查均采集了上述基础特征变量；（2）测量的一致性，相比于其他变量（如收入），包括受教育年限在内的各基础变量的测量在各调查间具有一致性；（3）避免缺失案例的影响。分析过程中的权重处理与上文相同。

（二）模型结果

以"受教育年限"为因变量的回归结果如表 2 所示。总体上，除了性别与户口这两个变

量在各调查中呈现出相同的作用以外，其他变量有不同程度的差异。性别与户口属性的回归系数表明，流动人口中男性受教育水平高于女性，非农户口者的受教育水平高于农业户口流动者。

表 2　各调查中受教育年限的线性回归模型

	2015 小普查	CFPS	CGSS	CLDS	CSS	CHNS	CHFS
性别（女 =0）	0.158***	1.215***	0.768***	0.679***	0.603***	0.535**	0.876***
	（0.008）	（0.146）	（0.130）	（0.059）	（0.142）	（0.174）	（0.082）
年龄	0.385***	0.171***	-0.145***	0.036*	-0.079	-0.145***	-0.223***
	（0.001）	（0.026）	（0.025）	（0.018）	（0.042）	（0.037）	（0.018）
年龄平方	-0.005***	-0.003***	0.000	-0.002***	-0.001	0.000	0.001***
	（0.000）	（0.000）	（0.000）	（0.000）	（0.000）	（0.000）	（0.000）
户口（农业 =0）	1.155***	3.541***	3.727***	3.794***	3.437***	0.662	3.779***
	（0.008）	（0.159）	（0.134）	（0.067）	（0.145）	（0.924）	（0.084）
婚姻（未婚 =0）							
已婚	-1.810***	-1.194***	-0.321	-1.235***	-1.270***	-0.626	-0.511***
	（0.014）	（0.329）	（0.220）	（0.099）	（0.231）	（0.425）	（0.154）
同居	-2.172***	-2.276	-0.601	-1.375***	0.785		-0.644
	（0.038）	（1.185）	（0.556）	（0.374）	（0.798）		（0.536）
离婚	-1.373***	-1.562**	0.028	-1.618***	-0.983	-0.589	-0.921**
	（0.034）	（0.579）	（0.373）	（0.271）	（0.574）	（0.827）	（0.290）
丧偶	5.868***	-1.272**	-1.467***	-1.083**	-3.106***	-0.801	-1.898***
	（0.017）	（0.443）	（0.337）	（0.415）	（0.643）	（0.608）	（0.276）
截距	0.158***	6.992***	14.669***	10.807***	13.977***	14.985***	16.540***
	（0.008）	（0.471）	（0.518）	（0.296）	（0.725）	（0.771）	（0.370）
样本量	305376	3130	2694	11266	2255	1246	6185
R^2	0.360	0.339	0.437	0.326	0.369	0.234	0.383

注：括号内为标准误，*、**、*** 分别表示在 5%、1%、1‰的水平上显著。

年龄和婚姻状况估计结果的差异可能体现了样本结构的偏差问题。首先，年龄对教育年限的影响在各调查中差异显著。从显著性看，CSS 中年龄的作用并不显著；2015 小普查和其余 5 个调查的样本中则均显著。从系数的方向看，小普查、CFPS 和 CLDS 中年龄与受教育程度正相关，即年龄越大，受教育年限越长；而 CGSS、CHNS 和 CHFS 中呈现显著的负相关。年龄的平方项亦是如此：小普查、CFPS 和CLDS 中呈负相关，CHFS 中为正相关，CGSS、CSS 和 CHNS 中则不显著。因此，年龄在各调查中呈现出完全相反的结果，这与各调查的年龄结构和教育结构的联合分布存在一定关联。其次，婚姻状况与受教育年限的相关性也存在一定差异。以未婚为参照组，已婚变量的作用大多为负，在 CGSS 和 CHNS 中并不显著；同居变量只在 CLDS 中显著为负，其他调查中均不显著；离婚变量在小普查、CFPS、CLDS 和 CHFS 中呈现显著的负向作用，在 CGSS、CSS 和 CHNS 中不显著；丧偶变量在 CHNS 中不显著，在其他调查中均显著为负。可见，在各调查中婚姻变量与教育的联合分布也存在显著差异。

以“是否为流动人口”为因变量的模型结果如表 3 所示。模型结果同样反映了共性与样本结构性偏差。

在 6 个调查样本中，CFPS 和 CLDS 中性别对是否为流动人口没有显著影响，这与小普

表 3　各调查中受访者是否为流动人口的 logistic 回归结果

	2015 小普查	CFPS	CGSS	CLDS	CSS	CHNS	CHFS
性别(女 =0)	- 0.006	- 0.012	- 0.169**	- 0.097	- 0.145**	- 0.190**	- 0.163***
	(0.004)	(0.056)	(0.054)	(0.059)	(0.055)	(0.068)	(0.030)
年龄	- 0.018***	- 0.012***	- 0.017***	- 0.029***	- 0.032***	0.001	- 0.040***
	(0.000)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.001)
受教育程度(未上过学 =0)							
小学 / 私塾	0.239***	- 0.092	0.354**	0.704***	- 0.101	0.238	0.124
	(0.009)	(0.097)	(0.113)	(0.152)	(0.113)	(0.131)	(0.077)
初中	0.476***	0.130	0.656***	1.059***	0.392***	0.392**	0.566***
	(0.008)	(0.090)	(0.113)	(0.151)	(0.109)	(0.131)	(0.074)
高中 / 中专	0.777***	0.045	0.908***	1.293***	0.508***	0.714***	0.749***
	(0.009)	(0.102)	(0.121)	(0.161)	(0.122)	(0.142)	(0.079)
大学本科 / 专科	1.014***	0.305**	0.997***	1.492***	0.927***	1.235***	0.941***
	(0.009)	(0.111)	(0.131)	(0.175)	(0.132)	(0.183)	(0.084)
硕士及以上	1.149***	0.591	1.276***	1.471***	0.972**	2.131*	1.073***
	(0.025)	(0.310)	(0.255)	(0.319)	(0.299)	(0.845)	(0.149)
户口(农业 =0)	0.693***	1.674***	0.597***	- 0.133	- 0.011	- 5.158***	- 0.155***
	(0.004)	(0.062)	(0.063)	(0.080)	(0.067)	(0.298)	(0.036)
婚姻(未婚 =0)							
已婚	0.625***	0.027	- 0.178	0.629***	0.330***	- 0.446**	- 0.207***
	(0.007)	(0.097)	(0.093)	(0.105)	(0.096)	(0.148)	(0.058)
离婚	0.280***	0.065	0.190	0.715**	0.442*	0.358	0.023
	(0.018)	(0.198)	(0.183)	(0.233)	(0.209)	(0.304)	(0.105)
丧偶	0.461***	0.050	- 0.298*	0.018	0.262	- 0.439*	- 0.035
	(0.016)	(0.177)	(0.149)	(0.376)	(0.191)	(0.216)	(0.098)
截距	- 2.279***	- 2.144***	- 0.910***	- 2.620***	- 0.433**	- 1.358***	0.191
	(0.008)	(0.129)	(0.150)	(0.182)	(0.153)	(0.208)	(0.100)
样本量	1950636	28356	10710	22020	10145	11666	34677
Log- Likelihood	- 804700	- 3.150e+08	- 5588	- 2.690e+08	- 5436	- 3026	- 14815
Pseudo R^2	0.0493	0.103	0.0785	0.0370	0.0566	0.218	0.0857

注：同表 2。

查的结果相同；但在其他调查中则存在负相关，即相对于女性，样本中男性是流动人口的可能性相对更低。年龄的回归系数在CLDS、CSS 和 CHNS 中不显著，而在小普查、CFPS、CGSS 和 CHFS 中显著为负。受教育水平的系数显示，相对于未上过学者，大学和研究生这两个类别更有可能是流动人口；初中学历的系数在所有调查中同样也显著为正。相比之下，小学与高中学历的系数在各调查中并不一致。在 CFPS、CSS 和 CHNS 中，小学学历不显著；在 CFPS 中，高中也不显著。这些差异既与各调查中的受教育水平的结构相关，也反映了变量的联合分布对单个变量在模型中作用的差异。

不同调查中户口变量的系数也存在较大差异。在小普查、CFPS、CGSS 中，户口呈现为显著的正向作用，即非农户口的受访者更可能是流动人口；在 CLDS 和 CSS 中户口的影响不显著；而在 CHNS 和 CHFS 中则呈现为显著的负向作用，即农业户口的受访者更可能是流动人口。户籍制度作为中国的主要社会经济制度之一，在人们生活的各个方面均有重要作用，流动人口户口类型的选择性到底如何，是农业人口还是非农人口更可能成为流动人口，既涉及学理上关于流动人口形成机制的解释，更涉及现实的制度安排，如农业人口流出地的相关政策制定等。婚姻变量的表现亦是如此。以已婚为例，CFPS 和 CGSS 中不显著，在小普查、CLDS 和 CSS 中显著为正；而在 CHNS 和 CHFS 中则显著为负。

上述两个统计模型的结果既揭示了共性规律，也反映了不同社会调查中样本结构所导致的统计结果差异性。一方面，未婚、拥有非农户口的男性流动人口往往具有更高的受教育水平，高学历（大学或研究生）群体更有可能是流动人口等，这些共性是社会现实的反映。另一方面，两个模型中各调查统计结果之间的差异，从根本上体现了样本结构的差异。

五、结论与讨论

（一）结论

本文通过详细对比国内 6 个常用的大型社会调查中流动人口的定义及样本结构，定量估计并讨论了社会调查中样本结构偏差对统计结果的可能影响。主要研究结论如下。

首先，目前大型社会调查对流动人口的定义各不相同。以本文考察的 6 项调查为例，尽管各调查均以“人户分离”作为判断流动人口的首要标准，但这些调查关于流动人口定义中的人户分离标准、时间和空间界定标准、市内人户分离人口的标准等具体测量方式均各不相同。由此导致，各调查中流动人口样本与总体存在不同程度偏离，使用这些调查数据估计全国流动人口规模与分布特征势必出现偏误。

其次，各调查收集的流动人口样本在相对规模和结构等方面有不同程度的偏离。若以 2015 年小普查为参照，CGSS 和 CSS 估计的流动人口比例相对偏高，其他调查则偏低；各调查估计的省际流动人口比例均大幅偏低，多数调查中省内县际流动人口比例偏低。多数调查中流动人口样本呈现“低龄组偏低、高龄组偏高”的年龄结构特征，其性别结构呈现“低龄组与老年组性别比偏高、劳动力年龄人口性别比偏低”的特征；除个别调查（CGSS）外，多数调查中流动人口的非农户口比例明显偏低。另外，部分调查（如 CFPS、CLDS、CHNS）呈现低学历流动人口比例偏高、高学历流动人口比例偏低；其他调查（CGSS 和 CSS）则相反。

最后，不同调查中，流动人口样本在不同维度特征的联合分布既有共性规律也有明显差异，反映了与各调查设计相关的样本选择性差异。由于各调查中流动人口的样本结

构存在不同程度、不同方向的偏差，简单基于这些调查数据中的流动人口样本进行统计推断不仅可能导致研究结果的偏误、阻碍对社会现实的准确认识，而且可能误导实践决策和政策制定。这些问题和研究误区需要在学术界引起高度重视。

（二）进一步的讨论

本文的研究结果表明，常用的大型社会调查中流动人口样本在相对规模及结构特征等方面与流动人口总体存在较大差异。其可能的原因既包括流动人口的统计口径与操作化等测量的差异，也与各调查自身的调查目的、抽样设计与实施等有关。这些研究结论为定量研究提出了诸多值得延伸讨论的问题。

1. 测量的重要性

社会科学研究中，测量的准确性既关系着研究对象在不同调查之间的一致性（如流动人口的操作化定义），也影响相关变量统计描述的准确性与统计推断的有效性（如测量偏差可能引发内生性问题）。因此，测量在整个社会科学研究中起着举足轻重的作用。本文以“流动人口”为例得出的研究发现，初步揭示了实证研究与测量相关的问题。

测量的重要性表现在以下方面，其一，研究对象识别的准确性。本文研究发现，各调查有关流动人口的定义和识别标准不一致，这意味着不同调查中的流动人口对应不同总体。部分调查使用的测量仅能识别“流动人口”中的一部分，而其他调查中的测量则可能难以识别和剔除非“流动人口”。由此导致，这些调查中关于流动人口的测量可能形成对总体的覆盖误差，并影响后续分析结论的正确性。其二，样本结构的无偏性。变量的测量可能会由于非随机的选择性应答模式造成样本结构的偏离。例如年龄、收入等变量在不同人群中会有不同的误报偏差，主观态度类变量的误报可能更为明显；由此导致变量的结构出现不同程度的偏差，成为导致统计结果有偏的重要原因之一。其三，统计结果的有效性。测量误差是导致内生性问题的重要原因之一（陈云松、范晓光，2010），进而影响到因果推断（Imai 等，2010；Kuroki 等，2014）。然而，目前大量实证研究忽略了测量的准确性及其可能带来的影响。

2. 调查设计的影响

不同的调查设计及与之相关的应答者差异，也是调查误差的重要来源（周皓，2019）。就流动人口研究而言，社会调查中的抽样设计及实施过程会影响其流动人口的样本结构。

就调查设计而言，目前，国内各大型社会抽样调查大多以了解转型时期中国社会变迁为主要目的，这些调查多采用家庭户（或户内人员）为基本抽样单位的抽样设计。因而，在区域覆盖上，不少调查并未考虑人口流动的区域差异，除 CGSS 和 CSS 以外的几个大型社会调查均未包括部分重要的人口流入省份（如新疆等）。加之，由于流动人口与户籍常住人口的居住空间及居住密集程度往往存在差异，大型社会调查抽样常用的住户或地图等抽样框容易出现对流动人口居住空间覆盖度低、代表性不足的问题（梁玉成等，

2015）；不少以家庭住户为主的调查在创建抽样框阶段已经剔除了一些流动人口聚集的建筑群，如工厂集体宿舍等，导致流动人口存在大量选择性遗漏和偏离。

从无应答情况来看，调查对象及选项的选择性无应答均是影响有效样本结构的关键因素。随着越来越多的举家迁移流动人口，流出地的流动人口比例由于单位无应答而可能被低估，从而引起样本结构的偏离；与流动人口相关的选择性应答拒访，可能导致流动人口的无应答率提高，进一步强化样本结构偏离。加之，流动者本人以外的其他人代答，可能因缺乏对被调查流动者实际情况的了解，增加数据中由选项无应答而导致非随机性缺失的可能。这些无应答导致的样本或变量缺失，均可能带来样本结构偏误。

3. 样本结构偏误的可能危害

流动人口样本的结构偏离可能导致两方面的危害。一方面，无法正确估计流动人口的总量（如比例与规模）及结构特征（如来源地、流入地等）。有偏的估计结果不利于准确了解人口迁移流动的总体状况及其具体特征，对相关科学研究和政策制定带来困扰。另一方面，可能危害对理论的实证检验。有偏的样本结构可能导致统计结果与研究结论出现偏误，无法达成研究的可复制性要求，从而阻碍对相关理论的有效检验和论证。

总之，关注测量、理解抽样过程与调查数据，识辨可能影响研究对象反映总体各方面特征的结构性偏差，是确保研究成果科学有效的前提和基础。

（三）相关建议

基于上述讨论，本文提出以下两点建议。一是研究者应正确认识并合理使用数据。二是建议各社会调查设计者尽可能使用统一的测量方式和统计口径，以便提高研究对象之间的可比性和不同调查结果之间的可交叉检验性。具体而言，研究者在使用数据之前应尽可能地了解相应调查的研究设计与抽样过程，以正确全面地认识数据、理解其适用性。具有全国代表性的样本可能对特定群体存在结构性偏离，忽略相应问题可能导致统计结果的未知偏误。只有在正确认识研究设计与抽样实施的基础上，才能真正地认识数据、合理地使用数据。以流动人口研究为例，关于流动人口总量（包括流量、强度与流向等）的估计，建议使用人口普查数据。在理解与解释各调查间统计结果的差异时，建议着重总结其中的共性规律；例如，在不同调查数据中反复被印证的结果更有可能是社会现实的真实反映。对于不同调查间不同的统计结果，应当审慎对待，深入探讨其可能的原因，如样本结构性偏差及其有效性问题等，而不是简单将其视为异质性。

参考文献：

1. 陈云松、范晓光（2010）：《社会学定量分析中的内生性问题　测估社会互动的因果效应研究综述》，《社会》，第 4 期。
2. 樊敏杰等（2023）：《农村劳动力迁移与城乡居民健康人力资本差异》，《开发研究》，第 4 期。
3. 李琴、宋月萍（2009）：《劳动力流动对农村老年人农业劳动时间的影响以及地区差异》，《中国农村经

济》,第 5 期。
4. 梁玉成等(2015):《流出地调查法:农村流动人口调查的理论与实践》,《华中科技大学学报(社会科学版)》,第 4 期。
5. 吕利丹、段成荣(2012):《对我国流动人口统计调查的总结与思考》,《南方人口》,第 3 期。
6. 齐嘉楠等(2014):《流动人口监测调查抽样设计的思考》,《统计与决策》,第 3 期。
7. 沈明明、李磊(2007):《流动人口、覆盖偏差和 GPS 辅助的区域抽样方法》,《理论月刊》,第 6 期。
8. 易龙飞、亓迪(2014):《流动人口健康移民现象再检验:基于 2006～2011 年 CHNS 数据的分析》,《西北人口》,第 6 期。
9. 周皓(2019):《两种调查视角下流动人口结构的对比分析》,《人口研究》,第 5 期。
10. 周皓(2023):《样本结构性偏差与因果推论——基于实验数据的分析》,《社会研究方法评论》,第 2 期。
11. 庄亚儿、李伯华(2014):《流动人口调查抽样的实践与思考》,《人口研究》,第 1 期。
12. Imai K., Yamamoto T.(2010), Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis. *American Journal of Political Science*. 54(2):543-560.
13. Kuroki M., Pearl J.(2014), Measurement Bias and Effect Restoration in Causal Inference. *Biometrika*. 101(2):423-437.
14. Tong Y., Piotrowski M.(2012), Migration and Health Selectivity in the Context of Internal Migration in China, 1997-2009. *Population Research and Policy Review*. 31(4):497-543.

Comparative Analysis of the Sample Structure of Migrants in Large-Scale Social Survey Data in China

Zhou Hao　Lei Linxuan

Abstract: Data forms the foundation of social science research. This paper compares definitions and data collection methods of "floating population" across six large-scale social surveys in China, highlighting structural differences between survey samples and the national floating population. Using a consistent model, it demonstrates how sample structure impacts analytical outcomes. Key findings include: (1) Significant definitional differences across surveys yield varied study population; (2) Estimated floating population proportions differ by survey, with both inter-provincial and intra-county migration rates generally lower than 2015 census data; (3) Floating population samples vary widely in gender, age, education, and urban-rural distribution; and (4) Basic demographic characteristics differ significantly in both significance and direction, revealing common social patterns and sample-specific biases. The study underscores the importance of measurement in defining research population, variable measurement, and statistical outcomes. It recommends using census or 2015 census data for estimating total floating population, emphasizing common patterns across samples to reflect social realities, and standardizing floating population metrics across surveys.

Keywords: Sample Survey; Floating Population; Measurement; Sample Structure

(责任编辑:李玉柱)