人口普查中多报人口数的估计*

胡桂华　武 洁　安 军

【摘　要】文章针对目前许多国家在人口普查质量评估工作中，估计人口普查中多报人口所用估计量存在的缺陷，构造多报比率估计量替代现行的估计量。文章用指标体系梳理与抽样估计相结合的研究方法，首先厘清普查多报估计所需人口普查登记的人口状态指标，并构建这些指标的平衡关系，然后讨论如何依据样本数据建立新的估计量，以及估计其方差。最后，文章利用中国某地区64个社区及行政村“六普”复查和19个样本调查小区的资料，演示人口普查的多报比率估计量及其方差估计量的计算过程，从而有助于对该估计量的理解，以期在中国2020年人口普查的质量评估中推广应用。

【关键词】多报　比率估计量　方差估计量

【作　者】胡桂华　重庆工商大学数学与统计学院，教授；武 洁　国家统计局人口和就业统计司，高级统计师；安 军　重庆工商大学数学与统计学院，副教授。

一、引　言

人口普查质量评估的一个重要工作是用抽样方法估计其多报、漏报人口数与净误差。人口普查中的多报是指登记了不应登记的人口。比如，登记了普查标准时点之后的出生人口，以及之前的死亡人口等。人口普查中的漏报是指应登而未登的人口。

绝大多数国家在估计人口普查净误差的同时，还强调对普查多报人口的估计，这样做是由于普查多报估计比净误差估计荷载的信息要丰富，而探究信息线索，则可进一步查找人口普查登记中的失误。这些失误往往是这次人口普查登记工作的难点，也是下一次人口普查登记可能发生差错的根源。鉴于此，美国普查局在2010年人口普查质量评估中，将工作重心转移到了多报估计上来，并提供了诸种多报人口估计值及其方差估计值。当然，美国普查局对外只发布了全国人口普查总的多报率，至于普查的各种多报率，仅限于供内部参考使用。

* 本文为国家社科基金项目“人口普查净误差估计中的三系统估计量研究”（编号：15BTJ011）的阶段性成果。

除了加拿大和中国等少数国家或地区外，绝大多数国家着力构造覆盖普查中全部多报类型的多报人口数估计量及其方差估计量。中国和加拿大等少数国家，除了人口普查中的重报外，其他人口多报微乎其微，因此可以忽略不计，于是把多报人口等同于重报。然而，中国第六次人口普查的复查结果表明，重报之外的其他多报是客观存在的。国务院第六次全国人口普查领导小组办公室及其下属机构，尽管事先对普查员进行了普查操作程序培训，但有些普查员或居民户却未严格按照普查操作细则填报普查短表或长表。

目前，包括中国、美国、新西兰、英国、澳大利亚和日本在内的所有国家均在人口普查质量评估中使用比较法估计多报人口（United Nations Secretariat，2010）。比较法适用于各类口径的人口普查多报人口估计。中国国家统计局在 2010 年人口普查质量评估中，从中国人口国情、社会背景与绝大多数国家不同的实际出发，使用比较法分别估计了现有人口、户籍人口、常住人口的普查重报率（国务院人口普查办公室、国家统计局人口和就业统计司，2013；陈培培、金勇进，2014），至于重报之外的其他多报人口，因远小于漏报人口而被忽略不计。

中国人口学学者所做人口普查质量评估研究工作（张二力、路磊，1992；张为民、崔红艳，2003；翟振武、陶涛，2010），呈以下 4 个特点：（1）在假设本次人口普查某个时点的某个年龄组（10～19 岁等）的人口，不存在普查漏报或多报的前提下，使用生命周期表法或其他方法估计上次人口普查同一时点的某个年龄组（0～9 岁）的人口（王金营、戈艳霞，2013；靳永爱、赵梦晗，2013）。如果这个估计的人口数大于上次普查登记人口数，就视为上次普查人口漏报，反之视为上次普查人口多报。（2）主要评估低年龄组人口或分性别人口组的人口普查登记质量。（3）把普查人口多报等同于普查重复登记（陶涛、张现苓，2013）。换句话说，他们根据中国实际情况，未估计除普查重复登记之外的其他普查多报人口数。（4）用本次人口普查结果评估以往人口普查部分结果的登记质量（郭志刚，2011）。

然而，上述研究存在一些明显的缺陷：（1）研究设定的本次人口普查不存在普查多报或漏报假设不成立。在每次人口普查中，总有些人漏报，也有些人多报。在不成立的假设条件下得出的任何研究结论都是不可靠的。（2）评估对象局限在部分年龄组，偏离了人口普查质量评估的基本目标。包括美国、中国在内的所有国家都是将人口普查质量评估的基本目标定在全国总人口的人口普查登记质量上。（3）混淆了普查中的人口多报、漏报和净误差概念，认为用估计的上次人口普查数据与上次人口普查登记数据计算的结果就是上次的人口普查多报率或漏报率。然而，美国及其他国家都是把估计的总体实际人口数与已知的普查登记人口数之差，再除以前者所得到的结果，称之为人口普查净多报率或净漏报率。中国国家统计局对外公布的 2010 年人口普查的 0.12%净漏报率被错误解读为普查漏报率。（4）使用的评估方法违背了独立性原则，背离了世界各国人口普查质量评估的主流方法。人口普查质量评估工作从产生的那一天起，世界上所有进行人口普查质量评估国家的国家统计局在制订的人口普查质量评估方案中都明确规定，只能通过在本次人口普查之后组织的质量评估调查，才能对本次（不能对上次或以往历次）人口普查的登记质量进行评估，而且还要

设法使这两项调查保持实质性独立。例如，使用不同工作人员、不同工作手段、不同工作机构、不同数据采集及处理方法等。为统一各国人口普查质量评估方法，以使各国人口普查评估结果具有可比性，联合国统计司于 2010 年邀请世界知名人口普查质量评估专家撰写了人口普查质量评估著作《事后计数调查——操作指南——技术报告》。由于每个国家人口普查质量评估工作是由各国国家统计局开展的，因此人口普查质量评估学者的研究应该围绕其制订的工作方案进行，在研究方法上保持基本一致。

鉴于此，本文试图通过引入普查登记中个人信息完整人数这一辅助变量，构造具有可操作性的“普查多报比率估计量”。该估计量能够较好地解决现行比较法由于样本中可能只观察到很少，甚至观察不到普查多报人口，从而导致普查多报人口数估计量难以构造的困难，并避免了比较法复杂的比较程序和可能产生的比较误差。这为各国构造普查多报人口数估计量提供了新思路、新途径，同时也有助于中国 2020 年人口普查多报人口估计精度的提高。

二、有关总体人口数指标及指标间的数量关系

为了说明如何根据观察到的样本数据构造用来估计普查多报人口数的“普查多报比率估计量”，首先要了解与人口普查登记质量有关的指标及这些指标之间的数量关系。本文讨论的普查登记中的各种指标之间的关系，是指将各种指标的全部调查小区值在汇总后得到的全国总值之间的关系。这些数量关系对现有人口、常住人口和户籍人口 3 种统计口径的普查登记人口数均适用。

按不同登记状态，对人口普查登记结果进行划分，如普查正确计数、各种不同类型的普查多报、是否登记全部普查项目等。它们共同构成了普查登记结果的平衡关系。普查多报人口需通过平衡关系来推算（见图）。

图中的指标包括：（1）普查登记人口数 C，是指直至人口普查质量评估调查工作开始时质量评估调查工作中心所掌握的人口普查登记人数。（2）普查正确计数人数 CE，指所进行的计数属于被调查人口总体成员，并且被计数者是在正确的地点进行登记，所做的登记完整地填写了应该填写的各项人口统计特征项目，而且只登记了一次。（3）地点错误的普查计数人数 WL，指虽然所进行的计数属于被调查人口总体成员，但被计数者的登记地点不正确，而所做的登记完整地填写了应该填写的各项人口统计特征项目。（4）不完整信息的普查计数人数 II，指虽然属于被调查人口总体成员的计数，但被计数者未能完整地

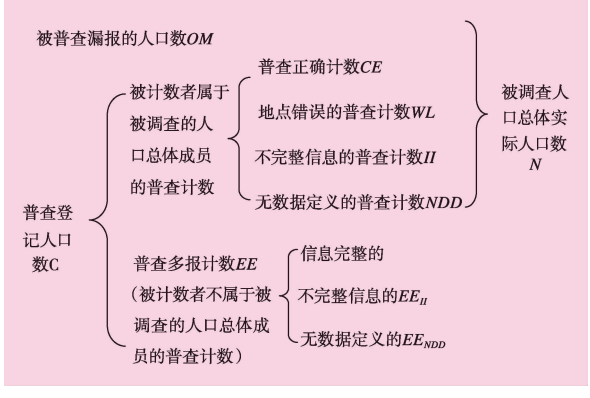

图　按人口普查登记结果不同状态分别统计的各种人数指标

填写应该填写的各项人口统计特征项目，不过尚能辨别出该人，至于登记地点是否正确在此处不进行区分。(5)无数据定义的普查计数人数 *NDD*，指虽然是对被调查人口总体成员的计数，但被计数者未能完整地填写应该填写的各项人口统计特征项目，以致无法辨别出该人，至于登记地点是否正确在此处不进行区分。(6)普查多报人口数 *EE*，指所进行的计数不属于被调查人口总体成员。在此本文对属于被调查人口总体成员但登记地点错误的普查计数进行一点说明。在美国2000年人口普查质量评估工作中，只估计净误差，规定把地点错误(未计数在常住地所在街区群或其周围环形区域)的普查计数视为多报；2010年在估计多报人数时，规定只要是对被调查人口总体成员的计数，即使被计数者的登记地点错误也不算多报。为了与2000年口径一致，本研究在估计净误差时仍然把地点错误的普查计数视为多报。(7)被普查漏报的人口数 *OM*，指被调查人口总体成员中应该进行普查登记却未登记的人口数。主要指对普查漠不关心、故意躲避普查、无固定住所、不相信政府和居住在偏远地区者，以及被监护的儿童等。(8)被调查人口总体实际人口数 *N*。为讨论问题方便，本文把人口普查结果视为一个常量。在人口普查质量评估工作中用普查结果与使用双系统估计量估计的总体实际人口数相减所得之差并不是人口普查结果与 *N* 之间的差距，它只是用来描述人口普查结果的分布特征的一种统计量，不能描述人口普查结果的准确度。

与人口普查多报人数估计有关的普查登记人数指标及若干关系式为：

$$EE=C-(CE+WL+II+NDD) \tag{1}$$

$$Y=CE+WL+II+NDD \tag{2}$$

$$X=C-(II+NDD)-(EE_{II}+EE_{NDD}) \tag{3}$$

式(1)至式(3)中，*EE* 为普查多报人口数，*Y* 为普查登记人数中属于被调查人口总体成员的人数，*X* 为个人信息完整的普查登记人口数。

三、估计普查多报人口数的程序

在人口普查质量评估中，与估计普查多报人口数有关的指标需要用抽取的调查小区的样本人口的数目及权数来估计。

(一)抽样方法

为便于样本的抽取，以全国各个行政区(省份)为抽样范围，采取等概率、分层、两步、整群抽样方法抽取样本(胡桂华、吴东晟，2014；胡桂华、武洁，2015)。具体步骤为：(1)第一步抽样。为了使样本尽可能覆盖总体，按城乡分层标志将每一个行政区内的所有调查小区分别划分在城市调查小区层或乡村调查小区层中，其中每一层用 *h* 表示，*h*=1，2。在每一抽样层中，从有关行政管理单位获取或自行编制调查小区抽样框。以调查小区为抽样单位，采取等概率纯随机不重复抽样方法从抽样框抽取调查小区样本。第一步抽样抽取的样本用于普查多报人口数方差估计值的计算。(2)第二步抽样。在抽取第二步样本之前，先对抽取的第一步样本调查小区按调查难度分为调查难度小、难度一般、难度大3层(*g* 为层标，*g*=1，2，3)。

在每一个 g 层，通过现场调查核实的方法重新编制调查小区抽样框，依照事先确定的抽样比率，仍然以调查小区为抽样单位，采取等概率纯随机不重复抽样方法从重新编制的抽样框中抽取第二步样本调查小区。第二步抽样抽取的样本用于普查多报人口数估计值的计算。

（二）获取样本数据

从式（1）可以看出，为了估计普查多报人口数，关键是要估计（$CE+WL+II+DD$）中的每一项。为了获得估计普查正确计数人数 CE 的样本数据，需要做 4 项工作。（1）审查普查表，看是否登记了某人的姓名、性别和年龄等所规定填写的全部项目。（2）在不同普查表之间核查，确定某人是否只登记 1 次。（3）向样本调查小区相关负责人了解该小区是否有人在普查日前后出生、死亡，或迁入、迁出。（4）检查普查表中填写的某人登记地点是否在研究区域范围内。只有同时具备普查项目信息登记完整、未重复登记、在普查总体内、登记地点正确，才能判定样本人口为普查正确计数者。不完整信息的普查计数人数 II 和无数据定义的普查计数人数 NDD 的样本数据，通过查阅样本调查小区普查表即可进行判断。为了获得估计登记地点错误的普查计数人数 WL 的样本数据，需要做两项工作：（1）确定某样本个人是否在普查日有 2 个及以上的住所。（2）确认在普查日有 2 个及以上住所者在普查中的“应该计数地点”和“实际计数地点”。

（三）“普查多报比率估计量”的构造

下面的公式可以分别代表常住人口、户籍人口和现有人口的样本调查小区的人口数据。$y_{hi}=CE_{hi}+WL_{hi}+II_{hi}+NDD_{hi}$ 为第一步抽样 h 层中 i 调查小区在人口普查中登记的属于人口总体成员的人数。其中，CE_{hi}、WL_{hi}、II_{hi}、NDD_{hi} 分别表示 h 层中 i 调查小区的个人信息填写完整并在正确地点计数的人数、个人信息填写完整但在错误地点计数的人数、不完整信息的登记人数、不符合普查数据定义的登记人数。$x_{hi}=C_{hi}-(II_{hi}+EE_{IIhi})-(NDD_{hi}+EE_{NDDhi})$ 为第一步抽样 h 层中 i 调查小区在人口普查中登记的个人信息完整的人数（已知）。其中，EE_{IIhi} 是该调查小区不属于人口总体的不完整信息的登记人数，EE_{NDDhi} 是该调查小区不属于人口总体的不符合普查数据定义的登记人数。

显然，从普查登记人数 C 中减去在人口普查中登记的属于人口总体成员的人数 Y 的估计量便得到普查多报人数估计量。至于 Y 的估计量，这里采用比率估计量来构造。在这个比率估计量中，引入普查登记中个人信息完整的人数 x 作为辅助变量。为构造 Y 的比率估计量，需要先构造 Y 的线性估计量和 X 的线性估计量。普查登记人数中属于人口总体成员的人数 $Y=CE+WL+II+NDD$ 和个人信息完整的普查登记人数 $X=C-(II+NDD)-(EE_{II}+EE_{NDD})$ 的线性估计量为：

$$\hat{Y}_u=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\frac{N_h}{n_h}\frac{n_{hg}}{r_{hg}}b_{hgi}I_{hgi}y_{hi} \qquad (4)$$

$$\hat{X}_u=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\frac{N_h}{n_h}\frac{n_{hg}}{r_{hg}}b_{hgi}I_{hgi}x_{hi} \quad (5)$$

在式(4)和式(5)中,下标 u 表示这里所构造的是线性估计量(用于与下面将要构造的以 R 为下标的比率估计量进行区别);y_{hi} 和 x_{hi} 的定义如前所述;b_{hgi} 表示对 h 层的第一步样本进行次级分层的示性函数,如果层 h 中的调查小区i 属于第二步抽样层 g,则 b_{hgi} 取值为 1,否则 b_{hgi} 取值为 0。由于层 h 中的每一个调查小区一定属于且仅属于某一个 g,换句话说,层 h 中的一个调查小区 i 的 b_{hgi},一定对且只能对某一个 g 取值为 1,而对其他的 g 则必然取值为 0。因此,用这个示性函数就把层 h 中的各个调查小区分到不同的 g 层;I_{hgi} 表示从 hg 层中抽取出来的第二步样本的示性函数,如果层 hg 中的调查小区 i 进入第二步样本,则 I_{hgi} 取值为 1,否则 I_{hgi} 取值为 0;$(N_h/n_h)(n_{hg}/r_{hg})$是 y_{hi} 的抽样权数,它是 h 层 i 调查小区经过两步抽样的样本被抽中概率的倒数。下面以$\hat{Y}_u$ 和$\hat{X}_u$ 为基础构造 Y 的比率估计量为:

$$\hat{Y}_R=X\frac{\hat{Y}_u}{\hat{X}_u} \quad (6)$$

在式(6)中,下标 R 表示这里所构造的是比率估计量;式中的 Y 由式(2)定义,X 由式(3)定义,$\hat{Y}_u$ 用式(4)计算,$\hat{X}_u$ 用式(5)计算。从普查登记人数 C 中减去由式(6)构造的估计量$\hat{Y}_R$,就得到人口普查多报人口数估计量。即:

$$\hat{EE}=C-\hat{Y}_R \quad (7)$$

(四)"普查多报比率估计量"的方差估计

现在考虑有限总体概率抽样产生的方差。本文中普查登记人口数 C 被视为常数,$\hat{EE}$ 的方差也就是$\hat{Y}_R$ 的方差。$\hat{Y}_R$ 的方差可以用"大折刀"方法来估计。

1."大折刀"方法简介及调查小区复制权数的计算

$\hat{Y}_R$ 属于总体参数复杂估计量。对于复杂估计量的方差计算,通常使用国外学者发明的分层刀切(Jackknife)方差估计量。在逐一切掉某一个第一步样本调查小区后,把重新计算的所有第一步样本调查小区的抽样权数称之为复制权数。被切掉的那个调查小区的抽样权数也叫复制权数,只不过它等于零。复制权数所要传递的意思是,切除某个调查小区后,原来调查小区的抽样权数变为多少(金勇进、张喆,2014)。

"大折刀"方差计算法的难点在于调查小区复制权数的计算。下面分别针对 h 层 g 子层 i 调查小区与被切的调查小区 t 之间关系的 5 种情况,具体分析如何确定"刀切"调查小区 t 后的切断后复制权数。(1)h 层 g 子层 i 调查小区与现在被切的调查小区 t 不在同一个 h 层。这时,调查小区所在的 h 层不受此次"刀切"的影响,所以调查小区 hgi 的抽样权数此时仍然是原来的权数$[(N_h/n_h)(n_{hg}/r_{hg})]$。(2)$h$ 层 g 子层 i 调查小区与现在被切的调查小区 t 在同一个 h 层但不在同一个g 子层。这时,hgi 调查小区所在的 h 层中的第一步样本受

此次“刀切”的影响减少了 1 个调查小区，也就是调查小区 hgi 的抽样权数此时应为 $[(N_h/n_h)(n_{hg}/r_{hg})][n_h/(n_h-1)]$。(3)$h$ 层 g 子层 i 调查小区与现在被切的调查小区 t 在同一个 h 层并且也在同一个 g 子层，此时，被切的调查小区 t 没有进入第二步样本，h 层 g 子层 i 调查小区与现在被切的调查小区 t 不是同一个调查小区。另外，hgi 调查小区所在的 g 子层中的第一步样本受此次“刀切”的影响也减少了 1 个调查小区，相应的在作为产生第二步样本的“总体”的 h 层 g 子层第一步样本的调查小区数 n_{hg} 应当换成 $(n_{hg}-1)$。为了实现这一要求，需要在原来权数 $[(N_h/n_h)(n_{hg}/r_{hg})]$ 的基础上乘以 $(n_{hg}-1)/n_{hg}$。此次“刀切”过程对 hgi 调查小区所在的 g 子层的第二步样本没有影响。这时，调查小区 hgi 的抽样权数为 $[(N_h/n_h)(n_{hg}/r_{hg})][n_h/(n_h-1)][(n_{hg}-1)/n_{hg}]$。(4)$h$ 层 g 子层 i 调查小区与现在被切的调查小区 t 在同一个 h 层也在同一个 g 子层并且被切的调查小区 t 进入了第二步样本，h 层 g 子层 i 调查小区与现在被切的调查小区 t 不是同一个调查小区。这时，hgi 调查小区所在的 h 层中的第一步样本受此次“刀切”的影响减少了 1 个调查小区。另外，hgi 调查小区所在的 g 子层中的第一步样本受此次“刀切”的影响也减少了 1 个调查小区。再有，hgi 调查小区所在的 g 子层中的第二步样本受此次“刀切”的影响也减少了 1 个调查小区。这时，调查小区 hgi 的抽样权数此时应为 $[(N_h/n_h)(n_{hg}/r_{hg})][n_h/(n_h-1)][(n_{hg}-1)/n_{hg}][r_{hg}/(r_{hg}-1)]$。(5)$h$ 层 g 子层 i 调查小区就是现在被切的调查小区 t。这时，应当将调查小区 hgi 的观察值从对 y_{vhi} 的求和中删除。显然，这只要将调查小区 hgi 的抽样权数由 $[(N_h/n_h)(n_{hg}/r_{hg})]$ 改为 0 便可实现。所以，调查小区 hgi 的抽样权数此时应为 0。

根据上述讨论，h 层 g 子层 i 调查小区的切断调查小区 t 后复制权数为 $\alpha_{hgi}^{(st)}$：

$$\alpha_{hgi}^{(st)}=\begin{cases}\dfrac{N_h}{n_h}\dfrac{n_{hg}}{r_{hg}}, \text{当 } h\neq s \text{ 时}\\ \dfrac{n_h}{n_h-1}\dfrac{N_h}{n_h}\dfrac{n_{hg}}{r_{hg}}, \text{当 } h=s, b_{sgt}=0 \text{ 时}\\ \dfrac{n_{hg}-1}{n_{hg}}\dfrac{n_h}{n_h-1}\dfrac{N_h}{n_h}\dfrac{n_{hg}}{r_{hg}}, \text{当 } h=s, b_{sgt}=1, I_{sgt}=0, i\neq t \text{ 时}\\ \dfrac{r_{hg}}{r_{hg}-1}\dfrac{n_{hg}-1}{n_{hg}}\dfrac{n_h}{n_h-1}\dfrac{N_h}{n_h}\dfrac{n_{hg}}{r_{hg}}, \text{当 } h=s, b_{sgt}=1, I_{sgt}=1, i\neq t \text{ 时}\\ 0, \text{当 } h=s, i=t \text{ 时}\end{cases} \tag{8}$$

式(8)中，s 表示被切掉的调查小区 t 所在的层。

2. $\hat{Y}_u$ 和 $\hat{X}_u$ 的切断后复制

依据式(8)复制权数构造的式(4)、式(5)和式(6)称之为复制估计量。下面分别就切掉的第一步样本调查小区 t，给出 $\hat{Y}_u$、$\hat{X}_u$ 和 $\hat{Y}_R$ 的切断后复制计算公式。切掉第一步样本调查小区 t 后，$\hat{Y}_u$、$\hat{X}_u$ 的切断后复制值 $\hat{Y}_u^{(st)}$、$\hat{X}_u^{(st)}$ 的计算公式分别为：

$$\hat{Y}_u^{(st)}=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)}\,b_{hgi}I_{hgi}y_{hi} \tag{9}$$

$$\hat{X}_u^{(st)}=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)}\,b_{hgi}I_{hgi}x_{hi} \tag{10}$$

式(9)和式(10)中的复制权数 $\alpha_{hgi}^{(st)}$ 由式(8)定义。

3. $\hat{Y}_R$ 的切断后复制

将式(9)和式(10)代入式(6),得到切掉调查小区 t 时,$\hat{Y}_R$ 的切断后复制值 $\hat{Y}_R^{(st)}$ 为:

$$\hat{Y}_R^{(st)}=X\frac{\hat{Y}_u^{(st)}}{\hat{X}_u^{(st)}} \tag{11}$$

4. $\hat{Y}_R$ 的"大折刀"方差与普查多报人口数估计量的方差

$\hat{Y}_R$ 的方差的"大折刀"估计量为:

$$v(\hat{Y}_R)=\sum_{h=1}^{H}\sum_{t=1}^{n_h}\frac{n_h-1}{n_h}(\hat{Y}_R^{(st)}-\hat{Y}_R)^2 \tag{12}$$

普查多报人口数估计量的方差为:

$$v(\hat{EE})=v(\hat{Y}_R) \tag{13}$$

四、实证分析

为节省篇幅和避免不必要的重复,本文只对人口普查中常住人口多报估计进行实证分析。

(一)有关情况和资料来源

中国第六次人口普查登记工作结束后,普查指导员组织普查员按照规定的方法全面复查了以家庭为单位填写的普查表。在复查中发现,虽然在普查前进行了培训,但有的普查员并未严格按照培训手册进行普查登记,登记了不应该登记的人口。虽然复查规定,对发现的问题重新入户核对,并经过确认后予以更正。但实际中并未真正做到这一点。因此,经过复查后,有些普查表仍然未能达到填写质量要求。如果复查发现了普查登记过程中的全部多报与漏报,并全部予以更正,那就没有必要单独进行人口普查质量评估。正是由于普查及复查工作存在的诸多问题,国家统计局对经过复查的第六次人口普查登记人口数进行了质量评估。

本文实证资料一部分来自广西南宁市西乡塘区人口普查办公室提供的其管辖范围内的64个社区及行政村的人口普查登记项目信息完整常住人口数(502 882人)及普查登记常住人口数(513 500人);另一部分来自从这64个社区及行政村按分层(社区层和行政村层)、两步、整群抽样方法抽取的第一步19个调查小区(含12个进入第二步样本和7个未进入第二步样本的调查小区)的常住人口数资料。这里所获得的每一个样本调查小区及每一个家

庭的常住人口数是通过现场调查得到的第一手原始资料，是实际常住人口数。

（二）样本的抽取及有关计算

1. 样本的抽取及抽样权数的计算

用 h 表示第一步样本抽取前对总体所有调查小区划分的层；N_h 为层 h 调查小区数；n_h 为从层 h 抽取的调查小区数；g 为对抽取的第一步样本进一步划分的层；n_{hg} 为 hg 层调查小区数；r_{hg} 为从 hg 层抽取的调查小区数。相关结果如表 1 所示。

表 1　样本的抽取与抽样权数的计算

h	N_h	n_h	N_h/n_h	g	i	n_{hg}	r_{hg}	n_{hg}/r_{hg}	$(N_h/n_h)\times(n_{hg}/r_{hg})$
h=1	1000	10	100	g=1	1√	3	2	1.5	150
h=1	1000	10	100	g=1	2√	3	2	1.5	150
h=1	1000	10	100	g=1	3	3	2	1.5	150
h=1	1000	10	100	g=2	4√	3	2	1.5	150
h=1	1000	10	100	g=2	5√	3	2	1.5	150
h=1	1000	10	100	g=2	6	3	2	1.5	150
h=1	1000	10	100	g=3	7√	4	2	1.5	200
h=1	1000	10	100	g=3	8√	4	2	1.5	200
h=1	1000	10	100	g=3	9	4	2	1.5	200
h=1	1000	10	100	g=3	10	4	2	1.5	200
h=2	1100	9	122	g=1	11√	3	2	1.5	183
h=2	1100	9	122	g=1	12√	3	2	1.5	183
h=2	1100	9	122	g=1	13	3	2	1.5	183
h=2	1100	9	122	g=2	14√	3	2	1.5	183
h=2	1100	9	122	g=2	15√	3	2	1.5	183
h=2	1100	9	122	g=2	16	3	2	1.5	183
h=2	1100	9	122	g=3	17√	3	2	1.5	183
h=2	1100	9	122	g=3	18√	2	2	1.5	183
h=2	1100	9	122	g=3	19	3	2	1.5	183

注：√记号为第一步样本进入第二步样本的调查小区。

表 1 显示，通过分层（2 层）、两步、整群（调查小区）、等概率简单随机抽样方法，从 h=1，2 层的 2 100 个调查小区中抽取第二步样本调查小区 12 个，其编号分别为 1、2、4、5、7、8、11、12、14、15、17、18，其抽样权数分别为 150、150、150、150、200、200、183、183、183、183、183、183。由于对第一、第二步样本调查小区中的常住人口是 100%抽样，所以在每一步不存在无应答的情况下，这些样本调查小区的抽样权数也就是其中每一个家庭或家庭成员的抽样权数。换句话说，样本调查小区抽样权数可以应用到样本常住人口，将样本人口数扩展到总体常住人口数，满足抽样调查总体参数估计理论要求。

对抽取的第二步样本调查小区，在社区负责人及其工作人员的协助下，由调查员现场获取该小区以家庭为登记范围的普查表，详细审查普查表中登记的常住人口信息，判断表

表2 第二步样本调查小区常住人口数据

h	g	i	y_{hi}	x_{hi}
h=1	g=1	1√	250	248
h=1	g=1	2√	240	238
h=1	g=2	4√	243	241
h=1	g=2	5√	230	230
h=1	g=3	7√	220	218
h=1	g=3	8√	190	190
h=2	g=1	11√	265	264
h=2	g=1	12√	300	297
h=2	g=2	14√	230	227
h=2	g=2	15√	250	250
h=2	g=3	17√	230	228
h=2	g=3	18√	240	239

注：同表1。

中登记的常住人口是否属于人口普查目标总体，以及普查项目是否登记完整。在此基础上，以每一个样本调查小区为范围，汇总该小区在人口普查中登记的，并且属于普查目标人口总体成员的常住人数 y_{hi}，以及人口普查项目登记信息完整的常住人数 x_{hi}。计算结果如表2所示。

2. 普查多报常住人口数的估计

利用表1和表2的数据，使用式(4)和式(5)，计算普查登记人数中属于人口总体成员的常住人数 $\hat{Y}_u$ 和个人信息完整的普查常住人数 $\hat{X}_u$ 的线性估计值为：

$$\hat{Y}_u=250\times150+240\times150+\cdots+240\times183=503695\text{ 人}$$

$$\hat{X}_u=248\times150+238\times150+\cdots+239\times183=500565\text{ 人}$$

使用式(6)估计的2010年广西南宁西乡塘区64个社区及行政村的在普查中登记且属于该地区的常住人口数为：$\hat{Y}_R=502882\times\frac{503695}{500565}\approx506000$ 人，其中的502 882人为2010年广西南宁西乡塘区64个社区及行政村的人口普查登记项目信息完整的常住人口数。使用式(7)估计的2010年广西南宁西乡塘区64个社区及行政村的人口普查多报常住人口数为：$\hat{EE}=513500-506000=7500$ 人，其中的常住人口数513 500人为2010年广西南宁西乡塘区64个社区及行政村的人口普查登记常住人口数。

3. 普查多报常住人口数的方差估计

普查多报常住人口数方差估计的具体过程是：(1)对第一步样本的19个调查小区进行轮换“刀切”，按式(8)和表1数据计算每切掉1个第一步样本调查小区后，其他18个第一步样本调查小区的抽样权数，被切掉的调查小区的抽样权数为零。例如，切掉 $s=1$ 及第一步样本中的第一个样本调查小区 $t=1$ 后，第一步19个样本调查小区的复制抽样权数为：第一个样本调查小区抽样权数为零；第二和第三个样本调查小区分别为222和111；第4～6个样本调查小区均为167；第7～10个样本调查小区均为222；第11～19个样本调查小区均为183。(2)按式(9)和式(10)计算切掉 t 调查小区条件下的 $\hat{Y}_u^{(st)}$ 和 $\hat{X}_u^{(st)}$ 的切断后复制值，进一步把这2个结果合并在一起使用式(11)计算切掉 t 调查小区条件下的 $\hat{Y}_R^{(st)}$ 切断后复制值。(3)把经过轮换19次“刀切”的19个 $\hat{Y}_R^{(st)}$ 切断后复制值，根据式(12)合并计算 $\hat{Y}_R$ 的方差估计值 $v(\hat{Y}_R)$，即为普查多报常住人口数的方差估计值 $v(\hat{EE})$(见表3)。

表 3　切断后复制值及刀切法普查多报常住人口数方差估计值计算结果

切掉第一步样本调查小区 t	$\hat{Y}_u^{(st)}$	$\hat{X}_u^{(st)}$	$\hat{Y}_R^{(st)}$	$(\hat{Y}_R^{(st)}-\hat{Y}_R)^2$	$(n_h-1)/n_h(\hat{Y}_R^{(st)}-\hat{Y}_R)^2$
1	500536	497484	505941	3481	3132
2	502756	499704	505928	5184	4665
3	556036	552540	506038	1444	1299
4	501155	498213	505826	30276	27248
5	504041	500655	506257	66049	59444
6	555101	551715	505943	3249	2924
7	501336	498504	505713	82369	74132
8	511326	507828	506320	102400	92160
9	574596	571098	505937	3969	3572
10	574596	571098	505937	3969	3572
11	504650	501289	506219	47961	4263
12	495025	492214	505728	73984	65763
13	576960	573328	506042	1764	1568
14	508410	505668	505583	173889	154568
15	502910	499343	506449	201601	179200
16	571180	567616	506014	196	174
17	507720	504703	505862	19044	16928
18	504970	501678	506156	24336	21632
19	570500	566936	506018	324	288
合计	—	—	—	—	716532

表 3 中的$\hat{Y}_R$为 506 000 人。从表 3 可以看出，估计的总体普查多报常住人口数 7 500 人的抽样方差为 716 532，抽样标准差为 848 人。这表明，在所有可能的样本中，平均每个样本估计的总体普查多报常住人口数为 7 500 人，相应的抽样平均标准误差为 848 人，即每个样本估计总体普查多报常住人口数与总体实际普查多报常住人口数的平均标准差异为 848 人。

五、结　语

本文通过对人口普查登记指标体系的构建及其平衡关系状态的梳理，使用两步整群抽样法及样本调查小区人口的抽样权数构造估计普查多报人口数的比率估计量，并用实际案例演示了该估计量计算的全过程。在此基础上，本文得到以下主要结论。

第一，为了使用“普查多报比率估计量”估计普查多报人口数，需要按照普查人口的登记状况对普查登记结果进行分类，并设法获得总体普查信息登记完整人口数这个辅助信息。利用与普查多报人口数高度相关的辅助信息构造的“普查多报比率估计量”，能明显提高总体普查多报人口数估计的精度。目前中国尚未在普查多报估计中对普查登记人口进行分类。因此，本文建议 2020 年全国人口普查时，对登记人口进行分类，尝试使用“普查多报比率估计量”估计普查多报人口数。

第二,估计人口普查多报人口数,需要使用广义的普查登记位置正确定义。当研究范围是某一个镇时,那么,这个镇内的任何一个人在人口普查中只要在该镇的任何一个地方登记,均认为是普查登记位置正确。同样,如果研究的范围是全国,那么该国内的任何一个人只要在这个国家的任何一个地方登记,就是普查登记位置正确。自1982年起,中国每次在普查多报人口数估计中,均使用狭义上的普查登记位置正确定义。因此本文建议在2020年普查多报人口数估计中使用广义上的普查登记位置正确定义,以避免普查多报人口数被不适当地虚高估计。

第三,不能把普查重复登记人口视为普查多报人口的全部。事实上,前者只是后者的一部分。在人口普查登记工作中,有些普查登记员不适当地把普查标准日之后出生的婴儿或之前死亡的人口错误登记为本次普查人口。中国在历次人口普查质量评估中只是估计了普查重复登记人口数,而未考虑估计其他多报人口数,这就低估了普查多报人口数,以致掩盖了本次普查操作程序的缺陷,不利于下一次普查方案的科学制订。因此本文建议在未来人口普查质量评估工作中严格按照普查多报人口的定义,构造覆盖全部普查多报人口的普查多报人口数估计量。

第四,在估计了普查多报人口数之后,还要计算其方差估计值。抽样理论告诉我们,用样本资料构造的总体参数估计量能否使用及使用效果如何的一个判断标准是抽样误差的大小。抽样误差过大,说明样本无法代表总体,用这样的样本估计的总体参数必然以较大幅度偏离其实际值,甚至使估计失去意义。中国虽然每次均计算了普查多报人口数的抽样误差估计值,但使用的方差计算方法需要改进。因此本文建议中国在2020年普查多报人口数估计中使用分层"刀切"方差估计量。

参考文献:

1. 陈培培、金勇进(2014):《对我国人口普查数据质量评估的若干思考》,《现代管理科学》,第9期。
2. 国务院人口普查办公室、国家统计局人口和就业统计司(2013):《第六次全国人口普查科学讨论会论文集》,中国统计出版社。
3. 郭志刚(2011):《"六普"结果表明以往人口估计和预测严重失误》,《中国人口科学》,第6期。
4. 胡桂华、吴东晟(2014):《人口普查质量评估调查的抽样设计》,《数量经济技术经济研究》,第4期。
5. 胡桂华、武洁(2015):《人口普查质量评估中Logistic回归模型的应用》,《数量经济技术经济研究》,第4期。
6. 金勇进、张喆(2014):《.抽样调查中的权数问题研究》,《统计研究》,第9期。
7. 靳永爱、赵梦晗(2013):《第六次人口普查数据中的年龄误报与分析》,《人口研究》,第1期。
8. 陶涛、张现苓(2013):《六普人口数据的漏报与重报》,《人口研究》,第1期。
9. 王金营、戈艳霞(2013):《2010年人口普查数据质量评估以及对以往人口变动分析校正》,《人口研究》,第1期。
10. Mule T.(2012),*2010 Census Coverage Measurement Estimation Report*:*Summary of Estimates of Coverage for Persons in the United States*. U.S. Census Bureau.
11. National Research Council(2009),*Coverage Measurement in the 2010 Census*,The National Academies Press.
12. United Nations Secretariat(2010),*Post Enumeration Survey*,United Nations Press.

(责任编辑:朱 萍)