

人口普查质量评估中的三系统估计量研究^①

胡桂华¹ 吴 婷² 范署姗²

(1. 广西师范大学数学与统计学院; 2. 重庆工商大学数学与统计学院)

研究目标: 用三系统估计量替代双系统估计量, 以解决后者由于普查与质量评估调查不独立引起的交互作用偏差问题。**研究方法:** 采用数理模型与实地调研相结合的方法研究总体人口的等概率分层, 三次捕获模型, 三系统估计量及其抽样方差估计量。**研究发现:** 对数线性模型、最大似然估计、不完整三维列联表和对数似然比是建立和选择三次捕获模型或三系统估计量的重要工具; 三系统估计量不受三份人口名单是否独立的限制。**研究创新:** 阐释三次捕获模型与三系统估计量的内在逻辑关系; 提出抽样登记的三系统估计量。**研究价值:** 三系统估计量应用于中国2020年人口普查净误差估计, 开创世界人口普查净误差估计领域应用三系统估计量的先河。

关键词 三次捕获模型 条件最大似然估计 总体规模估计量 不完整三维列联表 对数线性模型

中图分类号 F061 **文献标识码** A

引 言

2020年, 中国、美国和其他许多国家将进行人口普查。按照惯例, 也将组织人口普查质量评估。各国政府统计部门把净误差作为评估的主要目标(胡桂华, 2018), 并试图用估计的净误差率修正普查登记人口数。联合国统计司把净误差率作为评价各国人口普查登记质量的核心指标。人口普查净误差通常定义为未知的普查目标总体实际人口数与已知的普查登记人口数之差。不难看出, 净误差估计的关键是获得总体实际人口数的一个估计量。依据普查人口名单及其质量评估调查人口名单构造的双系统估计量, 是目前各国估计总体实际人口数的主要方法(杨贵军等, 2016; 胡桂华等, 2016、2017)。

本文拟在现行做法的基础上增加行政记录人口名单, 与前两个名单并列在一起构造三系统估计量替代现行的双系统估计量, 以解决后者的交互作用偏差问题(孟杰, 2019)。双系统估计量是由生物学家构建的最初用来估计野生动物总体规模的捕获一再捕获模型移植而来, 而三系统估计量则是由三次捕获模型移植而来。在人口普查质量评估中, 把人口普查的人口名单看作第一次捕获, 质量评估调查的人口名单看作第二次捕获, 用以建立估计目标总体人口数的双系统估计量; 如果再进一步把行政记录人口名单看作第三次捕获, 则可用这三份人口名单建立起估计目标总体人口数的三系统估计量。三系统估计量尚未应用于任何国家

^① 本文获得全国统计科学研究重点项目“人口普查登记误差估计研究”(2019LZ28)和教育部人文社会科学研究规划基金项目“人口普查漏报估计研究”(20YJA910002)的资助。

的人口普查净误差估计,属于当今人口普查质量评估领域的前沿方法。

根据对总体人口等概率分层方法的不同,三系统估计量有两种。一是基于对总体人口采用变量直接分层的三系统估计量;二是基于对总体人口采取 Logistic 回归模型间接分层的三系统估计量。国外学者构造了三份人口名单对总体全面登记的第一种三系统估计量。本文构造人口普查质量评估所需要的三份人口名单对总体抽样登记的第一种三系统估计量。相比第一种三系统估计量,第二种三系统估计量无须受样本规模所限而减少对总体人口等概率分层的变量数目。然而,现有的人口普查质量评估理论及概率抽样理论,还满足不了第二种三系统估计量构造的要求。

三系统估计量经历了三个发展研究阶段:

第一阶段,三次捕获模型。为构造三次捕获模型,需要获得同一总体在三次全面捕获中各次捕获的个体数日、同时在两次捕获中捕到的个体数日、同时在三次捕获中出现的个体数目。将这些数据填写在三维列联表中。该表由 8 个单元构成。其中一个单元的个体数目缺失,称其为缺失单元,其他 7 个单元的个体数日可以观察得到。列联表的个体总数目为 8 个单元的个体数目之和。由于缺失单元个体数目未知,所以总体的个体总数目也未知。三次捕获模型就是要估计总体中的个体数目。显然,建立三次捕获模型的关键是要构造缺失单元估计量。这个估计量依据 7 个单元的已知个体数目估计。估计方法依据三次捕获之间的统计关系而定(Bartlett, 1935; Birch, 1963; Fienberg, 1972)。统计关系不同,缺失单元的估计量也不同。三次捕获共有 8 种可能的统计关系(三次捕获之间两两独立,在给定第一次捕获的条件下第二与第三次捕获独立,等等),相应就有 8 种缺失单元估计量,8 种三次捕获模型。为了判断既定数据下,三次捕获之间究竟属于何种统计关系,需要引入对数线性模型,用其拟合三维列联表 7 个个体数目已知单元的期望频数,求得其期望频数的最大似然估计量。采用 χ^2 统计量或对数似然比值 G^2 (这两个检验统计量与 7 个单元的期望频数估计量及其观察值有关),检验对数线性模型对既定数据拟合的适当性,从中选择最适合于既定数据拟合的对数线性模型。相应地,就判断出既定数据属于哪种统计关系,就采用这种统计关系的三次捕获模型估计总体中的个体数目。三次捕获模型,也可以通过另外一种途径构造。那就是,把三次捕获中的两次捕获合并在一起,并且看作一次捕获。这样,三次捕获就变为两次捕获,根据两次捕获模型推出三次捕获模型公式(Marks 等, 1974)。

三次捕获模型在鱼和鸟类数目估计中有广泛的应用(Smith, 1988)。经过动物专家、概率论学者多年努力,三次捕获模型逐步由动物总体移植到人类总体建立三系统估计量。在移植的初始阶段,直接把三次捕获模型当作三系统估计量,用对总体人口的三份全面登记资料及其匹配资料构造三系统估计量。这样的三系统估计量不符合人口普查质量评估的要求。人口普查质量评估调查实际得到的是样本普查小区的普查人口名单、质量评估调查人口名单及行政记录人口名单。将三次捕获模型移植到人类总体构造三系统估计量需要解决一系列问题。例如,人口移动问题、构成元素的抽样估计问题、对总体人口的等概率分层问题。

第二阶段,三系统估计量在特殊群体人口数目估计中的应用。这方面的例子包括:Smith 依据报纸、调查和电话三个来源的数据,使用三系统估计量估计美国马萨诸塞州小城镇的志愿者机构数目(Bishop 等, 1975)。中国学者应用三系统估计量理论评估中国户籍登记系统的完整率(胡桂华和薛婷, 2018)。在特殊群体人口数目估计中应用的成功,表明三系统估计量可以应用到整个人群,例如全国、全国以下行政区总人口的数目估计。人口普查质量评估的对象是总人口,而不是特殊群体人口。

第三阶段,在人口普查质量评估领域应用的试点研究。这主要是借助已有的人口普查资料,辅以其他资料在小范围进行用三系统估计量估计人口数目的试验。Zaslavsky (1993) 利用三系统估计量估计美国黑人成年男性的人口普查净误差。Griffin (2014) 研究三份人口名单不同统计关系的三系统估计量,测算比对误差对三系统估计量的影响。本阶段尚未构造人口普查净误差估计所需要的抽样登记的三系统估计量。

据悉,美国普查局打算在 2020 年人口普查质量评估工作中,用三系统估计量估计总体实际人口数及人口普查净误差^①。我国国家统计局尚未开始三系统估计量研究。本文研究有助于我国赶上本领域国际先进水平的步伐。

笔者认为,用三系统估计量取代双系统估计量是人口普查质量评估领域学术发展的必然趋势。这一预言主要基于三点理由。理由之一,这是充分利用辅助信息的需要。充分利用辅助信息是一个十分重要的统计推断原则。目前正在使用的双系统估计量,就是为了引进人口普查登记资料这一辅助信息,因而才取代了仅仅依据质量评估调查资料构造的单系统估计量。但是,双系统估计量对辅助信息的利用仍不充分,未能把人口行政记录这一重要辅助信息利用起来,而三系统估计量同时利用了这两种辅助信息。理由之二,这是摆脱双系统估计量系统性偏误困扰的需要。系统性偏误是双系统估计量的一个致命的缺陷。它是由现实世界无法满足该估计量关于两个资料系统独立的要求造成的。三系统估计方法给出了三个资料系统各种不同统计关系下三系统估计量的构造方式,这使得它对普查人口名单、质量评估调查人口名单及行政记录人口名单之间统计关系的要求非常宽松,从而摆脱了独立性的严格限制。理由之三,三系统估计量的研究已经达到了投入使用的水平。三系统估计量已经具有了成熟的理论体系,并且积累了在特殊人群,乃至区域人口中应用的试点经验。

本文创新体现在以下三个方面。第一,截至目前,笔者所见到的三系统估计量文献,还主要是对总体全面调查理想条件下的讨论。在有限总体概率抽样条件下三系统估计量及方差的构造策略要独立思考地提出。通过研究,我们率先提出三个系统对总体抽样登记的、具有国际领先水平的三系统估计量,使三系统估计量应用于人口普查净误差估计成为现实可能。第二,三系统估计量研究的一个疑难问题是,怎样对总体人口等概率分层,以满足三系统估计量在等概率人口层构造及使用的要求。有些研究人员直接在非等概率人口总体中使用三系统估计量,也有些科研人员凭经验感觉主观选择分层变量或避免选择分层变量,甚至还有的科研机构省去同时未登记在普查人口名单、质量评估调查人口名单及行政记录人口名单的人数。本文基于加权优比排序法(胡桂华,2015),同时考虑其他因素,综合确定对总体人口等概率分层的变量,在等概率人口层建立三系统估计量。为避免三系统估计量低估总体人口数,将同时遗漏于三份人口名单的三重漏报人口纳入其中。第三,解读三维列联表单元期望频数最大似然估计量及三次捕获模型的构造原理,厘清估计动物总体规模的三次捕获模型与估计人类总体规模的三系统估计量之间的数量关系,从三次捕获模型推出三系统估计量公式。

本文对我国即将到来的 2020 年人口普查质量评估工作的借鉴指导意义表现在以下几个方面。第一,为我国构造适合于人口普查净误差估计的三系统估计量及其抽样方差估计量提

^① 美国普查局原计划在 2020 年人口普查净误差估计中,用基于 Logistic 回归模型的双系统估计量替代对总体人口直接分层的双系统估计量,以解决后者等概率人口层登记概率异质性引起的交互作用偏差而低估总体实际人口数的问题。由于基于 Logistic 回归模型的双系统估计量过于复杂,于是计划采用三系统估计量。

供了理论公式。第二,为我国对总体人口等概率分层提供了方法。第三,为我国采用分层刀切法近似计算三系统估计量或净误差的抽样方差提供了程序。这些无疑有助于我国在2020年构建以三系统估计量为核心的人口普查质量评估方案。

一、三次捕获模型

1. 描述三次捕获结果的不完整三维列联表

三次捕获模型的一类典型案例是:对同质动物封闭总体进行三次重复全面捕获,用所获得的有关数据来估计总体中动物的数目。“同质动物总体”指的是,具有相同活动能力的同种动物,它们具有相同的不等于0的被捕到的概率;“封闭总体”指的是,总体中的动物在三次捕获期间不会发生个体的增加和减少;“重复捕获”指的是,每次捕获完成之后把被捕到的动物放回总体;“全面捕获”指的是,每次捕获的目标是“一网”把总体中的动物全部捕到。由于众多偶然因素的影响,总体中的动物有一些会没有捕到。

三次捕获的操作中有一个重要的技术环节:第一次捕获后把该次被捕到的所有动物用某一种颜色(例如红色)涂上市号;第二次捕获后把该次被捕到的所有动物用不同于第一次的另一种颜色(例如绿色)涂上市号;第三次捕获后把该次被捕到的所有动物用不同于第一次与第二次的第三种颜色(例如蓝色)涂上市号。

经过三次捕获,一方面获得第一次、第二次、第三次被捕到的动物的数目数据;另一方面还可以获得只涂有某一种颜色的动物的数目、同时涂有两种颜色的动物的数目、同时涂有三种颜色的动物的数目数据。把三次捕获数据列示于表1。

表1 三次捕获的不完整三维列联表

三次捕获结果	在第三次捕获中 ($k=1$)		不在第三次捕获中 ($k=2$)	
	在第二次捕获 ($j=1$)	不在第二次捕获 ($j=2$)	在第二次捕获 ($j=1$)	不在第二次捕获 ($j=2$)
在第一次捕获中 ($i=1$)	x_{111}	x_{121}	x_{112}	x_{122}
不在第一次捕获中 ($i=2$)	x_{211}	x_{221}	x_{212}	x_{222}

表1中有一个缺失单元,其数据 x_{222} 未知。另外7个单元的数据是可以观察到的,用 x_{ijk} 表示每个单元(ijk)的数据, i 、 j 、 k (或者说,它们所示的下角标位置)分别表示第一次捕获、第二次捕获、第三次捕获;每一个下角标位置分别都只取1、2两个值,取1时表示动物在该次捕获中出现,取2时表示动物在该次捕获中未出现。单元(222)对应于总体中的个体在三次捕获中都没有出现。

用 x 表示总体中观察到的个体的总数目:

$$x = \sum^* x_{ijk} = x_{111} + x_{121} + x_{211} + x_{221} + x_{112} + x_{122} + x_{212} \quad (1)$$

其中,加 $*$ 的求和号表示,这个求和不包括单元(222)。另外,我们用 m_{ijk} 表示单元(ijk)的动物数目 x_{ijk} 的期望频数,即 $E(x_{ijk}) = m_{ijk}$ 。

我们的目的是估计总体中的动物总数 X 。 $X = x + m_{222}$ 。由于 x 已知,所以也可以把日的说成估计 m_{222} 。

2. 不完整三维列联表单元期望频数及总体规模的最大似然估计

为了求得不完整三维列联表单元期望频数及总体规模的最大似然估计,需要引入对数线

性模型。对数线性模型是估计三维列联表单元期望频数及总体规模的重要工具。对数线性模型依据三次捕获之间的统计关系建立。对每种统计关系,建立一个与之相应的对数线性模型。每个单元的期望频数 m_{ijk} 用对数线性模型 $\ln m_{ijk}$ 拟合。对数线性模型分为两类:一类是基于三次捕获为两两相关的饱和对数线性模型;另一类是基于三次捕获为其他统计关系的非饱和对数线性模型。通常是,先给出饱和对数线性模型的缺失单元频数的最大似然估计量,再将非饱和对数线性模型 7 个单元频数的最大似然估计量代入其中,得到后者缺失单元频数的最大似然估计量。

为书写便利,用 a 、 b 、 c 分别表示第一次捕获、第二次捕获、第三次捕获。三次捕获的统计关系共包括四类:Ⅰ类,即三次捕获之间两两相关;Ⅱ类,即其中一次捕获分别与另外两次捕获独立;Ⅲ类,即在给定其中一次捕获的条件下,另外两次捕获独立;Ⅳ类,即三次捕获相互独立。Darroch (1958) 给出了这 4 类统计关系的对数线性模型。

(1) Ⅰ类统计关系(即三次捕获之间两两相关)对数线性模型的最大似然估计。

$$\ln m_{ijk} = \lambda + \lambda_i^a + \lambda_j^b + \lambda_k^c + \lambda_{ij}^{ab} + \lambda_{jk}^{bc} + \lambda_{ik}^{ac} \quad (2)$$

其中, $\ln m_{ijk}$ 是单元 (i, j, k) 理论频数(单元频数的期望值) m_{ijk} 的对数,描述了 $\ln m_{ijk}$ 的数据结构。等号右边各项的意义如下: λ 是各单元 m_{ijk} 的几何平均的对数,把它作为各单元 $\ln m_{ijk}$ 的基础水平; λ_i^a 是第一次捕获的第 i 种情况对 $\ln m_{ijk}$ 的影响量; λ_j^b 是第二次捕获的第 j 种情况对 $\ln m_{ijk}$ 的影响量; λ_k^c 是第三次捕获的第 k 种情况对 $\ln m_{ijk}$ 的影响量; λ_{ij}^{ab} 是第一次捕获的第 i 种情况和第二次捕获的第 j 种情况对 $\ln m_{ijk}$ 的交互影响量; λ_{jk}^{bc} 是第二次捕获的第 j 种情况和第三次捕获的第 k 种情况对 $\ln m_{ijk}$ 的交互影响量; λ_{ik}^{ac} 是第一次捕获的第 i 种情况和第三次捕获的第 k 种情况对 $\ln m_{ijk}$ 的交互影响量。

式(2)允许两次捕获之间相关,但不允许三次捕获之间相关。由于三维列联表只有 7 个单元的数据可以观察到,因此这 7 个单元的期望频数的最大似然估计量都为各自的观察值(Bishop 等, 1975), 即 $\hat{m}_{ijk} = x_{ijk}$ 。总共 8 个单元期望频数的最大似然估计量为:

$$\begin{aligned} \hat{m}_{111} &= x_{111} & \hat{m}_{221} &= x_{221} & \hat{m}_{122} &= x_{122} & \hat{m}_{212} &= x_{212} \\ \hat{m}_{121} &= x_{121} & \hat{m}_{211} &= x_{211} & \hat{m}_{112} &= x_{112} \end{aligned} \quad (3)$$

$$\hat{m}_{222} = \frac{\hat{m}_{111}\hat{m}_{221}\hat{m}_{122}\hat{m}_{212}}{\hat{m}_{121}\hat{m}_{211}\hat{m}_{112}} \quad (4)$$

总体个体数目 X 的估计量为:

$$\hat{X} = x + \hat{m}_{222} \quad (5)$$

(2) Ⅱ类(其中一次捕获分别与另外两次捕获独立)对数线性模型的最大似然估计。这包括三种情形:① c 与 a 、 b 独立;② b 与 a 、 c 独立;③ a 与 b 、 c 独立。我们给出在第①种情形下(其他情形可以类似写出)的对数线性模型及其最大似然估计。

$$\ln m_{ijk} = \lambda + \lambda_i^a + \lambda_j^b + \lambda_k^c + \lambda_{ij}^{ab} \quad (6)$$

Birch (1963) 证实,非饱和对数线性模型的单元期望频数的最大似然估计量的边际之和,为相应单元观察值边际之和。对式(6),它就是:

$$\hat{m}_{ij-} = x_{ij+} \quad \hat{m}_{++k} = x_{++k} \quad (7)$$

由于 a 、 b 相关,它们与 c 独立。这时,可以把 a 、 b 合并在一起当作一次捕获, c 当作

另外一次捕获, 这两次捕获独立。在两次捕获独立情况下, 依据独立双系统估计量公式可以得到:

$$\hat{m}_{ijk} = \frac{\hat{m}_{ij+} \hat{m}_{++k}}{\hat{m}_{++}} = \frac{x_{ij+} x_{++k}}{x} \quad (8)$$

式(8)不适合单元(221)和单元(222)期望频数估计量的构造。

单元(221)的期望频数估计量 $\hat{m}_{221}$, 这个单元的频数估计量只与该单元的观察值有关, 与其他单元观察值无关, 即 $\hat{m}_{221} = x_{221}$ 。其他6个单元的期望频数估计量依据式(8)构造。科学研究的一个重要原则是, 用过了的信息最好不要重复使用, 如 x_{221} ; 对无法获取的信息, 从公式中删除, 以使公式具有可用性, 如 x_{222} 。现在分析式(8)分子和分母的每一项。 x_{ij+} 共有四种情况: $x_{11+} = x_{111} + x_{112}$, $x_{12+} = x_{121} + x_{122}$, $x_{21+} = x_{211} + x_{212}$, $x_{22+} = x_{221} + x_{222}$ 。由于最后一种包括了已经用过的 x_{221} 和无法获取的 x_{222} , 因此 x_{ij+} 实际上只包括前面的3种。 x_{++k} 共有两种情形, x_{++1} 和 x_{++2} 。 $x_{++1} = x_{111} + x_{121} + x_{211} + x_{221}$, 由于 x_{221} 已经使用, 因此应该从 x_{++1} 中剔除, 使用 $x'_{++1} = x_{111} + x_{121} + x_{211}$ 。 $x_{++2} = x_{112} + x_{122} + x_{212} + x_{222}$ 。由于 x_{222} 无法得到, 因此用 $x'_{++2} = x_{112} + x_{122} + x_{212}$ 替代 x_{++2} 。由于 x 包括了 x_{221} , 因此应该用 $x' = x - x_{221}$ 替代 x 。基于上述分析, 7个单元的期望频数估计量为:

$$\begin{aligned} \hat{m}_{221} &= x_{221} & \hat{m}_{111} &= \frac{x_{111} x'_{++1}}{x'} & \hat{m}_{121} &= \frac{x_{121} x'_{++1}}{x'} & \hat{m}_{211} &= \frac{x_{211} x'_{++1}}{x'} \\ \hat{m}_{112} &= \frac{x_{111} x'_{++2}}{x'} & \hat{m}_{122} &= \frac{x_{121} x'_{++2}}{x'} & \hat{m}_{212} &= \frac{x_{211} x'_{++2}}{x'} \end{aligned} \quad (9)$$

将式(9)代入式(4)得到:

$$\hat{m}_{222} = \frac{x_{221} x'_{++2}}{x'_{++1}} \quad (10)$$

缺失单元估计量 $\hat{m}_{222}$, 也可以通过二维列联表缺失单元估计量公式推导出来。这时把 a 和 b 合并在一起当作一次捕获, 把 c 当作另外一次捕获。 a 和 b 分别与 c 独立, 合并后依然与 c 独立。其中在前一次捕获但未在另外一次捕获的结果为 $x_{112} + x_{122} + x_{212}$, 即 x'_{++2} , 未在前一次捕获但在另外一次捕获的结果为 x_{221} , 在这两次捕获的结果为 $x_{111} + x_{121} + x_{211}$, 即 x'_{++1} 。在两次捕获独立条件下, 缺失单元估计量 $\hat{m}_{222}$ 与式(10)相同。

将式(10)代入式(5)得到:

$$\hat{X} = x + \frac{x_{221} x'_{++2}}{x'_{++1}} \quad (11)$$

(3) III类(在给定其中一次捕获的条件下, 另外两次捕获独立)的对数线性模型最大似然估计。这也包括三种情形: ①给定 c 条件下, a 与 b 独立; ②给定 b 条件下, a 与 c 独立; ③给定 a 条件下, b 与 c 独立。我们给出第②情形下(其他情形可以类似写出)的对数线性模型及其最大似然估计。

$$\ln m_{ijk} = \lambda + \lambda_i^a + \lambda_j^b + \lambda_k^c + \lambda_{ij}^{ab} + \lambda_{jk}^{bc} \quad (12)$$

式(12)的单元 (ijk) 期望边际之和的最大似然估计量设定为:

$$\hat{m}_{+jk} = x_{-jk} \quad \hat{m}_{ij+} = x_{ij-} \quad \hat{m}_{++} = x_{++} \quad (13)$$

由于 b 和 a 相关, b 和 c 相关, a 和 c 独立, 所以 b 和 a 合并及 b 和 c 合并后还是独立, 依据独立条件下的双系统估计量公式得到式 (14):

$$\hat{m}_{ijk} = \frac{\hat{m}_{+jk}\hat{m}_{ij+}}{\hat{m}_{+j+}} \quad (14)$$

式 (14) 不适合于单元 (221)、单元 (122)、单元 (121) 及单元 (222) 期望频数估计量的构造。其中前 3 个单元期望频数估计量只与各自单元的观察值有关, 而与其他单元的观察值无关, 即它们的期望频数估计量分别等于其观察值。另外, 单元 (111)、单元 (211)、单元 (112) 和单元 (212) 的期望频数估计量, 依据式 (14) 构造。最终得到:

$$\begin{aligned} \hat{m}_{221} &= x_{221} \quad \hat{m}_{122} = x_{122} \quad \hat{m}_{121} = x_{121} \quad \hat{m}_{111} = \frac{x_{11+}x_{+11}}{x_{+1+}} \\ \hat{m}_{211} &= \frac{x_{21+}x_{+11}}{x_{+1+}} \quad \hat{m}_{112} = \frac{x_{11+}x_{+12}}{x_{+1+}} \quad \hat{m}_{212} = \frac{x_{21+}x_{+12}}{x_{+1+}} \end{aligned} \quad (15)$$

将式 (15) 代入式 (4) 得到:

$$\hat{m}_{222} = \frac{x_{221}x_{122}}{x_{121}} \quad (16)$$

将式 (16) 代入式 (5) 得到:

$$\hat{X} = x + \frac{x_{221}x_{122}}{x_{121}} \quad (17)$$

(4) IV 类 (三次捕获相互独立) 的对数线性模型最大似然估计。

$$\ln m_{ijk} = \lambda + \lambda_i^a + \lambda_j^b + \lambda_k^c \quad (18)$$

式 (18) 单元 (ijk) 期望频数的最大似然估计量为:

$$\hat{m}_{i++} = x_{i++} \quad \hat{m}_{+j+} = x_{+j+} \quad \hat{m}_{++k} = x_{++k} \quad (19)$$

由于 x_{2++} 、 x_{+2+} 、 x_{++2} 各自都包括 x_{222} , 而 x_{222} 未知, 因此单元期望频数最大似然估计量无法从式 (19) 直接获得。Bishop 等 (1975) 建议使用 Deming-Stephen 的比例迭代拟合法来间接求各个 $\hat{m}_{ijk}$ 。可以证明, 各个 (ijk) 单元的迭代结果收敛于该单元的期望频数 m_{ijk} 。注意, 不含缺失单元。

迭代的操作过程大致如下:

各单元的迭代需要由初始值来启动。为此, 首先要指定各个 (ijk) 单元的迭代初始值。一般各单元的初始值均令其等于 1, 即 $\hat{m}_{ijk}^{(0)} = 1$ 。等号左边记号右上角标带括号的数字表示第几次迭代的结果。“(0)”表示初始值, “(1)”表示第一次迭代, 等等。初始值指定后, 在各个单元, 同时以初始值为基础进行第一次迭代, 求出各个单元的第一次迭代结果 $\hat{m}_{ijk}^{(1)}$ 。接下来, 同时在该单元的 $\hat{m}_{ijk}^{(1)}$ 的基础上进行第二次迭代, 求出各个单元的第二次迭代结果 $\hat{m}_{ijk}^{(2)}$ 。再接下来, 同时在各单元 $\hat{m}_{ijk}^{(2)}$ 的基础上进行第三次迭代, 求出各个单元的第三次迭代结果 $\hat{m}_{ijk}^{(3)}$ 。至此, 完成了第一轮迭代 (由上面的叙述可以看出, 完整的一轮迭代中, 要做三次迭代), $\hat{m}_{ijk}^{(3)}$ 便是第一轮迭代的最后结果。为了与第二轮迭代结果相区别, 把它记作 $\hat{m}_{ijk}^{(1-3)}$, 上角标短线前的 1 表示第一轮迭代的结果。第一轮迭代结束之后, 继续同时在各单元启动第二轮迭代。这时, 各单元第一轮迭代结果 $\hat{m}_{ijk}^{(1-3)}$, 是该单元第二轮迭代的初始

值,由它启动第二轮迭代,经过与第一轮相仿的三次迭代,得到各单元第二轮迭代结果 $\hat{m}_{ijk}^{(2-3)}$ 。

是否进行第三轮迭代?迭代何时停止? Bishop 等(1975)介绍了一个简便的停止规则。使用这个规则时,要事先定出一个量 δ ,规定它的值(例如 $\delta=0.1$ 或 $\delta=0.01$)。然后,对每一个 (ijk) 单元,计算第二轮迭代结果 $\hat{m}_{ijk}^{(2-3)}$ 与第一轮迭代结果 $\hat{m}_{ijk}^{(1-3)}$ 之差的绝对值。如果所有 7 个单元的上述差的绝对值都小于 δ ,则迭代到此(第二轮完成后)停止;如果 7 个单元的上述计算结果中有的不小于 δ ,则还需要继续进行第三轮迭代。第三轮完成后,用所得结果与第二轮的结果相减,看各单元相减结果的绝对值是否全都小于 δ 。如果是,迭代停止;如果不是,则还要进行第四轮迭代。如此循环往复地进行下去,什么时候 7 个单元的本次迭代结果与上次迭代结果差的绝对值全都小于 δ ,迭代就停止。

读者可能会问, δ 的取值应如何确定呢?从前面介绍的停止规则,不难看出 δ 实际上是控制各单元估计量 $\hat{m}_{ijk}$ 误差大小的一个工具。对 $\hat{m}_{ijk}$ 的要求越高,越是希望它的误差小, δ 的值就应当定得越小。

在前面关于迭代过程的叙述中遗留了一个问题:在每一轮迭代中所进行的三次迭代的计算公式是什么样的? Bishop 等(1975)以第一轮迭代为例给出了其中的三次迭代的公式。

第一次迭代:

$$\hat{m}_{ijk}^{(1)} = \hat{m}_{ijk}^{(0)} \frac{x_{ij+}}{\hat{m}_{ij+}^{(0)}} \quad (20)$$

第二次迭代:

$$\hat{m}_{ijk}^{(2)} = \hat{m}_{ijk}^{(1)} \frac{x_{i+k}}{\hat{m}_{i+k}^{(1)}} \quad (21)$$

第三次迭代:

$$\hat{m}_{ijk}^{(3)} = \hat{m}_{ijk}^{(2)} \frac{x_{+jk}}{\hat{m}_{+jk}^{(2)}} \quad (22)$$

式(20)~式(22)中,各个比式分子分母下角标中的十号,表示对相应位置的下角标求和。例如 $x_{11+}=x_{111}+x_{112}$, $x_{1-1}=x_{111}+x_{121}$, $x_{+11}=x_{111}+x_{211}$ 。

通过式(20)~式(22)得到不完整三维列联表 7 个可观察单元期望频数的最大似然估计。但无法得到缺失单元期望频数及总体规模的最大似然估计。下面解决这个问题。由于三次捕获相互独立,所以单元 (ijk) 的概率为其边际概率的乘积,即:

$$\pi_{ijk} = \pi_{i++}\pi_{+j-}\pi_{-+k} \quad (23)$$

$$\hat{\pi}_{1+-} = \frac{x_{1++}}{\hat{X}} \quad \hat{\pi}_{+1+} = \frac{x_{+1-}}{\hat{X}} \quad \hat{\pi}_{++1} = \frac{x_{+1+}}{\hat{X}} \quad (24)$$

$$\frac{\hat{m}_{222}}{\hat{X}} = (1 - \hat{\pi}_{1+-})(1 - \hat{\pi}_{+1+})(1 - \hat{\pi}_{++1}) = \left(1 - \frac{x_{1++}}{\hat{X}}\right) \left(1 - \frac{x_{+1+}}{\hat{X}}\right) \left(1 - \frac{x_{+1+}}{\hat{X}}\right) \quad (25)$$

$$(\hat{X} - x_{1+-})(\hat{X} - x_{+1+})(\hat{X} - x_{++1}) = \hat{X}^2(\hat{X} - x) \quad (26)$$

从式(26)可以求得 $\hat{X}$ 。在得到 $\hat{X}$ 的基础上,根据 $\hat{X}=x+\hat{m}_{222}$,可以推出 $\hat{m}_{222}$ 。

3. 对数线性模型的选择或三次捕获统计关系的判定

这里,我们先把上述对数线性模型式(2)、式(6)、式(12)和式(18)简称为模型

2、模型 6、模型 12 和模型 18。

为了找到最好的对数线性模型来拟合既定数据,用上述方法分别求得 4 个对数线性模型各自的 7 个单元期望频数的最大似然估计量,然后进行与观察数据的拟合优度检验。检验的工具是 χ^2 分布统计量或对数似然比统计量 G^2 ^①。

$$\chi^2 = \sum_i \sum_j \sum_k \frac{(\hat{m}_{ijk} - x_{ijk})^2}{\hat{m}_{ijk}} \quad (27)$$

$$G^2 = \sum_i \sum_j \sum_k x_{ijk} \ln \left(\frac{x_{ijk}}{\hat{m}_{ijk}} \right) \quad (28)$$

模型 2、模型 6、模型 12 和模型 18,对应的式 (27) 或式 (28) 的 χ^2 分布统计量或对数似然比统计量 G^2 的自由度分别是 0、2、1、3。

选择对数线性模型其实是进行下面的三个原假设:

第一个检验 H_0 : 模型 6 与模型 2 没有差异;

第二个检验 H_0 : 模型 12 与模型 2 没有差异;

第三个检验 H_0 : 模型 18 与模型 2 没有差异。

在这里,我们没有对第 II 类模型的三种情形以及第 III 类模型的三种情形分别进行检验。这是因为,每一类下面的三种情形所用的检验统计量是一样的,所以没有必要分开单独检验。

这几个检验的前提是,模型 2 与数据完全拟合,而这几个检验的用意是,想要看一看,能否用较简约的模型(模型 6、模型 12、模型 18)取代模型 2 去拟合数据。若原假设被拒绝,表明不能用作为该检验目标的简约模型取代模型 2,换句话说,该简约模型被拒绝;若原假设不能被拒绝,表明能够用作为该检验目标的简约模型能够取代模型 2,即该简约模型不能被拒绝。

上述 3 个检验所用的检验统计量分别是:第一个检验:模型 6 的 χ^2 —模型 2 的 χ^2 ;第二个检验:模型 12 的 χ^2 —模型 2 的 χ^2 ;第三个检验:模型 18 的 χ^2 —模型 2 的 χ^2 。所得之差仍服从 χ^2 分布,自由度是相减的两个 χ^2 统计量自由度之差。

倘若检验结果表明,有一个原假设不能被拒绝(例如第二个),这里我们提出“尽量使用简约模型”原则,依照这个原则,应选用模型 12,而舍弃模型 2。

假若有两个原假设不能被拒绝(例如第一个和第二个),依照“尽量使用简约模型”原则,应选用模型 6 和模型 12 而舍弃模型 2。这时,还要进一步在模型 6 和模型 12 之间进行选择。为此,应建立下列原假设:

H_0 : 模型 6 与模型 12 没有差异。

这时的检验统计量是前面第一个检验的检验统计量与第二个检验的检验统计量之差,它仍然服从 χ^2 分布,自由度是相减的两个检验统计量的自由度之差。

假定经过一系列检验之后选定了模型 6。这时,总体中动物总数使用这种统计关系的三次捕获模型来估计。

二、三系统估计量

三次捕获模型并不是三系统估计量。在将三次捕获模型移植到人类总体构造三系统估计

① 式 (27) 或式 (28) 是三次捕获对同一总体全面捕获的公式。将三次捕获模型移植到人类总体构造三系统估计量时,须将这两个公式的 χ^2 或 G^2 分别更改为 $\hat{\chi}^2$ 或 $\hat{G}^2$, x_{ijk} 更改为 $\hat{x}_{ijk}$, $\hat{m}_{ijk}$ 更改为 $\hat{\hat{m}}_{ijk}$ 。

量时,有许多理论与实践问题需要解决。例如,对总体人口等概率分层,在等概率人口层建立三系统估计量。三次捕获模型中的第一次、第二次、第三次捕获,分别对应于三系统估计量中的人口普查、人口普查的质量评估调查及人口行政记录。

1. 怎样构造基于三次捕获模型的三系统估计量

(1) 满足三次捕获模型的基本假设。三系统估计量来源于三次捕获模型。三次捕获模型基于三个基本假设。构造三系统估计量时,应遵循这些基本假设,做好满足这些假设的相关性技术工作。

第一,三次捕获模型的封闭总体假设。依此假设,构造三系统估计量所用的普查人口名单、质量评估调查人口名单、行政记录人口名单,都应该是对普查时点同一人口总体的登记。可是,实际情况并不满足这个要求。质量评估调查,通常在普查登记工作结束后一段时间才开始。在这段时间内,会发生小区之间的人口移动。有些人从其他小区迁移到本小区(向内移动人口),也有些人从本小区迁移到其他小区(向外移动人口),还有些人一直居住在本小区(无移动人口)。在这种情况下,质量评估调查人口有两种构成方法。一是由无移动人口和向外移动人口构成。二是由无移动人口和向内移动人口组成。由于无移动人口和向外移动人口在普查时点都居住在本小区,为满足该假设,质量评估调查人口采取第一种方法构造更为合理。行政记录人口名单,可能包括实际上不存在的人口,也可能包括重复人口,还可能包括滞后登记人口。为满足该假设,应该剔除重复人口和实际上不存在的人口,补充滞后普查时点登记的人口,以使行政记录人口名单尽可能包括普查时点实际存在的全部人口。

第二,三次捕获模型的总体动物的同质性假设。在每次捕获中,假定总体中的动物有同样的捕获概率。依此假设,总体中的人口,应该有同样的概率在每份人口名单登记。但总体中的人,在生活习性、对待统计调查的态度等诸多方面存在较大差异。这就使得不同的人在进行人口普查登记时、进行人口普查的质量评估调查登记时、进行人口行政记录登记时,登记的概率不相同。为解决这个问题,通行的对策是,分别针对每份名单,找出影响人口参与该名单登记概率的变量,把三个名单的影响变量集合在一起,形成变量群。然后,对这个变量群进行筛选,再后,用筛选后的变量群对人口总体进行交叉分层。这样,同一层内的人口在进行人口登记时便会具有大致相同的登记概率,在这样的层构造三系统估计量,然后再把所有层的三系统估计量加总,得到整个总体人口数的三系统估计量。把这样的层简称为等概率层^①。在实际工作中,等概率分层是在抽取质量评估调查样本之后进行的。在每一个样本普查小区,把其中的人分别划入各个等概率人口层。

第三,三次捕获模型的无错误捕获假设。在每次捕获中,动物总体中的有些动物可能未捕获到,但总体中的动物不会在一次捕获中多次捕获,或者捕获总体之外的动物。可是,在每次人口登记中,有些人可能多次重复登记,也可能登记不应该登记的人,例如,登记普查日前的死亡者或普查日之后的出生者。因此,在构造三系统估计量时,应该确保普查人口名单、质量评估调查人口名单和行政记录人口名单不包括普查目标总体之外的人,即这三份人口名单无错误登记人口。

(2) 进行匹配性比对。三次捕获模型采用在动物身体上涂色的方法对同时出现在某两次捕获中的动物、同时出现在三次捕获中的动物进行计数。构造三系统估计量也需要这种数

^① 关于等概率分层变量群的组建和筛选问题请参见胡桂华(2015)。

据。它们是采用匹配性比对的办法得到的。把普查小区中的三份名单拿来进行比较,找出两份名单以及三份名单匹配的部分(匹配者,就是同时在这两份或三份名单中出现的人)。如果某人在两份或三份名单的姓名、性别、年龄、婚姻状况、文化程度、居住地等相同,我们就称他或她为匹配者。

(3) 用样本资料构造三系统估计量各个构成部分的总体总量指标。三次捕获模型的试验背景是无限重复随机试验。每一次捕获的结果是样本,所要估计的动物总数是试验总次数。如果普查人口名单、质量评估调查人口名单、行政记录人口名单,是对普查时点人口总体全面调查的登记结果,那么,用这三份名单构造的总体人口数目的三系统估计量相当于三次捕获模型构造的动物总数估计量。可是,在实际工作中,人口普查质量评估调查是从总体中以普查小区为单位抽取概率样本来进行的^①(胡桂华和吴东晟,2014)。这时,上面的三系统估计量中用到的各种总体总量指标就要用样本来估计,所得到的总体人口数的估计量是上面的三系统估计量的估计量。

(4) 采用刀切法计算三系统估计量的抽样误差。三次捕获模型的各个构成元素是子总体指标,可以直接得到,无须估计,抽样方差使用 Delta 方法近似计算。但三系统估计量的构成元素,以估计量形式表现,而且是复杂估计量,使用分层刀切方差估计量或泰勒线性方差估计量近似计算其抽样误差。泰勒线性方差估计量虽然在计算上优于分层刀切方差估计量,但过于复杂,一般很少采用。在从目标总体抽取概率样本的情况下,三系统估计量中所使用的具有特征(总体中的人口在三份名单登记、在其中的两份名单登记,在其中的一份名单登记)的试验结果出现的次数具有双重身份:一方面是多项分布随机试验的样本数据,另一方面又是目标人口有限总体的总体指标(需要用有限总体概率样本来估计)。经过这样的估计以后,所得到的三系统估计量,应当称作用目标人口总体全面调查结果构造的三系统估计量的估计量。有限总体概率抽样的操作会产生方差。所以,这个估计量的方差便包含了两个层次:一是用目标人口总体全面调查结果构造三系统估计量时所产生的方差;二是有限总体概率抽样时所产生的方差。现在的任务是要把二者结合起来计算一个完整的方差。显然,这个任务太复杂了,无法用一个数学解析式子来表达,只好用分层刀切方差估计量之类的数据计算方法近似估计^②。

2. 等概率人口层的三系统估计量

按照是否纳入人口移动因素,可以构造两种形式的三系统估计量:一是无人人口移动的三系统估计量;二是人口移动的三系统估计量。考虑到迄今所有国家的政府统计部门尚未研究或使用三系统估计量,本着由易到难,循序渐进的原则,本文只研究无人人口移动的三系统估计量。人口移动三系统估计量另文论述。由于前者是后者的基础,所以只要研究好前者,研究后者就不会遇到太大的困难。

用 v 表示任意一个等概率人口层。等概率人口层的总层数用 V 表示。在每一个层 v 建立三系统估计量,用 $\hat{X}_v$ 表示。汇总所有层 v , 得到总体的三系统估计量 $\sum_{v=1}^V \hat{X}_v$ 。理论上,应该依据每一种统计关系,建立与这种统计关系相适应的三系统估计量。实际上看,可

① 在人口普查质量评估调查中,有些国家采取分层整群抽样,例如中国、乌干达和南非。也有些国家采取分层多重抽样,例如美国和瑞士。还有些国家采取分层多阶段抽样,例如新西兰。

② 美国、瑞士等少数国家的政府统计部门,在人口普查质量评估中,采用分层刀切抽样方差估计量近似计算净误差的抽样方差。包括我国在内的许多国家的政府统计部门尚未采用这一抽样方差计算公式。

在对三份名单可能发生的统计关系分析的基础上,构造实际上最可能发生的统计关系的三系统估计量。根据现实情况的分析,本文构造最可能发生的3种统计关系的三系统估计量^①。

为节省篇幅,我们忽略三份人口名单对总体全面登记的三系统估计量,直接构造三份名单对总体抽样登记的三系统估计量的估计量。

(1) 基于模型2的三系统估计量,即普查、质量评估调查和行政记录两两相关的三系统估计量的估计量:

$$\hat{X}_v = \hat{x}_{111v} + \hat{x}_{112v} + \hat{x}_{121v} + \hat{x}_{122v} + \hat{x}_{211v} + \hat{x}_{221v} + \hat{x}_{212v} + \frac{\hat{x}_{111v}\hat{x}_{221v}\hat{x}_{122v}\hat{x}_{212v}}{\hat{x}_{121v}\hat{x}_{211v}\hat{x}_{112v}} \quad (29)$$

(2) 基于模型6的三系统估计量,即人口行政记录与普查和质量评估调查独立,后两者相关的三系统估计量的估计量:

$$\hat{X}_v = \hat{x}_{111v} + \hat{x}_{112v} + \hat{x}_{121v} + \hat{x}_{122v} + \hat{x}_{211v} + \hat{x}_{221v} + \hat{x}_{212v} + \frac{(\hat{x}_{112v} + \hat{x}_{122v} + \hat{x}_{212v})}{(\hat{x}_{111v} + \hat{x}_{121v} + \hat{x}_{211v})} \quad (30)$$

(3) 基于模型12的三系统估计量,即既定普查下,质量评估调查和人口行政记录独立的三系统估计量的估计量:

$$\hat{X}_v = \hat{x}_{111v} + \hat{x}_{112v} + \hat{x}_{121v} + \hat{x}_{122v} + \hat{x}_{211v} + \hat{x}_{221v} + \hat{x}_{212v} + \frac{\hat{x}_{212v}\hat{x}_{221v}}{\hat{x}_{211v}} \quad (31)$$

式(29)~式(31)等号右边的估计量采用有限总体概率样本来构造。质量评估调查属于大规模抽样调查,通常采用分层二重抽样,也称双重抽样。它是一项节约成本且效率高的抽样技术。在第一重抽样中抽取一个大样本,获得辅助变量,整合这些变量,作为第二重抽样的分层变量。在同样成本下,以这种抽样方式构造的估计量,比一重抽样的抽样方差小。在二重抽样中,常见的一种方式,是第一重抽样并不分层,第二重抽样分层,分层变量为第一重抽样得到的辅助信息。本文采取这种二重抽样方式。

不难看出,构造三系统估计量的关键是获得上述等号右边各个估计量的计算公式。它们中的每一个所要估计的都可以看作总体各个普查小区在 v 层的观察值的总体总值。把第 i 普查小区在层 v 的观察值记为 y_{iv} ,则层 v 的总体总值为:

$$Y_v = \sum_{i=1}^N y_{iv} \quad (32)$$

其中, N 为总体的普查小区总数目。上述所有估计量可以统一看作由式(32)所写出的总体总值 Y_v 的估计量。下面只需要考虑怎样由样本普查小区的观察值 y_{iv} 得到估计量 $\hat{Y}_v$ 。如果采取一重简单随机抽样,那么 Y_v 的一个无偏估计量为:

$$\hat{Y}_v = \sum_{i \in A_1} w_i y_{iv} \quad (33)$$

其中,第一重样本普查小区 i 的抽样权数 $w_i = N/n$, n 为第一重样本规模, A_1 是第一重样本普查小区的集合。为压缩最终样本规模及提高样本代表性,对第一重样本分层。第一重样本中的第 i 普查小区可能落在 G 个层中的某一层。对所有 $i \in A_1$,如果 i 属于层 g ,那

^① 三份名单相互独立这种统计关系在现实的人口普查质量评估中是不可能发生的。首先,人口普查和其质量评估调查由政府统计部门统一组织实施,为了加快后者工作进度,往往从本次普查员中挑选质量评估调查员。其次,中国有些住户户口簿填写普查短表和长表。

么 $x_{ig}=1$, 否则 $x_{ig}=0$ 。不难看出, $\sum_{g=1}^G x_{ig}=1$ 。从层 g 采取简单随机抽样方法抽取第二重样本。用 $n_g = \sum_{i \in A_1} x_{ig}$ 表示第一重样本在层 g 的普查小区数日, $r_g = \sum_{i \in A_2} x_{ig}$ 表示层 g 第二重样本普查小区的数目, $w_g = n_g / r_g$ 表示第二重样本普查小区的抽样权数。 A_2 为第二重样本小区的集合。对第二重样本, Y_v 的无偏估计量为:

$$\hat{Y}_v = \sum_{i \in A_2} \alpha_{gi} y_i \quad (34)$$

$$\alpha_{gi} = w_i w_g = (N/n)(n_g / r_g) \quad (35)$$

式 (34) 为双重扩张估计量。 Y_v 的再加权扩张估计量是:

$$\hat{Y}_v = \sum_{g=1}^G \left(\sum_{i \in A_1} w_i x_{ig} \right) \frac{\sum_{i \in A_1} w_i x_{ig} y_i}{\sum_{i \in A_2} w_i x_{ig}} = \sum_{i \in A_2} \alpha_{gi} y_i \quad (36)$$

$$\alpha_{gi} = \sum_{g=1}^G \left[\frac{\sum_{j \in A_1} w_j x_{jg}}{\sum_{j \in A_2} w_j x_{jg}} \right] w_i x_{ig} \quad (37)$$

3. 等概率人口层三系统估计量的抽样方差

现在计算式 (29) ~ 式 (31) 三系统估计量的抽样方差。这样的三系统估计量十分复杂, 其方差难以精确计算。泰勒线性方差估计量虽然可以精确计算其抽样方差, 但十分复杂。这使得复制方法成为唯一可行的方差估计法。分层 Jackknife 估计法属于复制方法之一 (胡桂华等, 2016)。它是通过将一系列复制值和三系统估计量估计的结果之间的差异平方和来计算的。其中每一个复制值是在剔除第一重样本中的一个普查小区后, 用新的复制权数重新计算的值。复制次数等于第一重样本规模。分层 Jackknife 法分三步进行。

第一步是计算进入第二重样本的普查小区的复制权数。复制权数是指剔除层 s 的小区 t 后计算的层 g 的普查小区 i 的抽样权数。小区 gi 的原始抽样权数为 $\alpha_{gi} = (N/n)(n_g / r_g)$ 。现在讨论小区 gi 各种可能情况的复制权数计算。如果层 g 不是层 s , 此时剔除层 s 的小区 t 对小区 gi 的抽样权数没有影响, 其复制权数 $\alpha_{gi}^{(s)}$ 等于原始抽样权数。如果 $g=s$, $i \in g$, 但 $t \in A_1$, 此时 A_1 减少了 1 个小区, 变为 $(n-1)$, 而层 g 小区数没有变化, 第二重样本小区 gi 的复制权数变为 $\alpha_{gi}^{(s)} = [N / (n-1)]$ 。如果 $g=s$, $i \in g$, $t \in g$, $t \in A_1$, 但 $t \notin A_2$, 此时 A_1 和层 g 的小区数各减少一个, 但 A_2 小区数无改变, $\alpha_{gi}^{(s)} = [N / (n-1)] [(n_g - 1) / r_g]$ 。如果 $g=s$, $i \in g$, $t \in g$, $t \in A_1$, $t \in A_2$, 那么 A_1 , g , A_2 小区数各减少一个, 小区 gi 的复制权数变为 $\alpha_{gi}^{(s)} = [N / (n-1)] [(n_g - 1) / (r_g - 1)]$ 。如果 $i=t$, 那么不难理解 $\alpha_{gi}^{(s)} = 0$ 。将这些分析结果用式 (38) 表示:

$$\alpha_{gi}^{(s)} = \begin{cases} \frac{N}{n} \frac{n_g}{r_g} & (g \neq s) \\ \frac{N}{n-1} \frac{n_g}{r_g} & (g=s, i \in g, t \notin g, t \in A_1) \\ \frac{N}{n-1} \frac{n_g-1}{r_g} & (g=s, i \in g, t \in g, t \in A_1, t \notin A_2) \\ \frac{N}{n-1} \frac{n_g-1}{r_g-1} & (g=s, i \in g, t \in g, t \in A_1, t \in A_2) \\ 0 & (i=t) \end{cases} \quad (38)$$

第二步计算复制值。用式(38)的 $\alpha_{gi}^{(s)}$ 替代式(36)的 α_{gi} 得到其复制估计量:

$$\hat{Y}_v^{(s)} = \sum_{i \in \Lambda_g} \alpha_{gi}^{(s)} y_i \quad (39)$$

式(29)~式(31)中的每一项用式(39)表示,得到其三系统估计量的复制估计量:

$$\hat{X}_v^{(s)} = \hat{x}_{111v}^{(s)} + \hat{x}_{112v}^{(s)} + \hat{x}_{121v}^{(s)} + \hat{x}_{122v}^{(s)} + \hat{x}_{211v}^{(s)} + \hat{x}_{221v}^{(s)} + \hat{x}_{212v}^{(s)} + \frac{\hat{x}_{111v}^{(s)} \hat{x}_{221v}^{(s)} \hat{x}_{122v}^{(s)} \hat{x}_{212v}^{(s)}}{\hat{x}_{121v}^{(s)} \hat{x}_{211v}^{(s)} \hat{x}_{112v}^{(s)}} \quad (40)$$

$$\hat{X}_v^{(s)} = \hat{x}_{111v}^{(s)} + \hat{x}_{112v}^{(s)} + \hat{x}_{121v}^{(s)} + \hat{x}_{122v}^{(s)} + \hat{x}_{211v}^{(s)} + \hat{x}_{221v}^{(s)} + \hat{x}_{212v}^{(s)} + \hat{x}_{221v}^{(s)} \frac{(\hat{x}_{112v}^{(s)} + \hat{x}_{122v}^{(s)} + \hat{x}_{212v}^{(s)})}{(\hat{x}_{111v}^{(s)} + \hat{x}_{121v}^{(s)} + \hat{x}_{211v}^{(s)})} \quad (41)$$

$$\hat{X}_v^{(s)} = \hat{x}_{111v}^{(s)} + \hat{x}_{112v}^{(s)} + \hat{x}_{121v}^{(s)} + \hat{x}_{122v}^{(s)} + \hat{x}_{211v}^{(s)} + \hat{x}_{221v}^{(s)} + \hat{x}_{212v}^{(s)} + \frac{\hat{x}_{212v}^{(s)} \hat{x}_{221v}^{(s)}}{\hat{x}_{211v}^{(s)}} \quad (42)$$

第三步计算 $\hat{X}_v$ 的抽样方差估计量。式(29)~式(31)的抽样方差可统一表示为:

$$v(\hat{X}_v) = \sum_l \left(1 - \frac{n}{N}\right) \frac{n-1}{n} (\hat{X}_v^{(s)} - \hat{X}_v)^2 \quad (43)$$

4. 总体的三系统估计量及抽样方差估计量

$$\hat{X} = \sum_{v=1}^V \hat{X}_v \quad (44)$$

$$\text{var}(\hat{X}) = \sum_v \sum_{v'} \text{Cov}(\hat{X}_v, \hat{X}_{v'}) \quad (45)$$

$$\text{Cov}(\hat{X}_v, \hat{X}_v) = \text{var}(\hat{X}_v) \quad (46)$$

$$\text{Cov}(\hat{X}_v, \hat{X}_{v'}) = \sum_l \left(1 - \frac{n}{N}\right) \frac{n-1}{n} (\hat{X}_v^{(s)} - \hat{X}_v) (\hat{X}_{v'}^{(s)} - \hat{X}_{v'}) \quad (47)$$

5. 总体的净误差估计量及抽样方差估计量

用 $N\hat{C}E$ 表示总体的普查净误差估计量,其公式为:

$$N\hat{C}E = \hat{X} - C\hat{C} \quad (48)$$

式(48)中, CC 表示总体的普查登记人口数。

$$\text{var}(N\hat{C}E) = \text{var}(\hat{X}) \quad (49)$$

三、实证分析

1. 基本情况及样本资料

通过自行组织调查,在与社区负责人签订数据使用保密协议,以及支付数据使用费的前提下,我们获得了重庆市南岸区南坪街道的7个社区的6个样本小区的普查人口名单、质量评估调查人口名单及行政记录人口名单。南坪街道共有13个社区,其中7个社区的负责人提供数据。以7个社区为实证对象。这7个社区共有200个普查小区。在第一重抽样中,以普查小区为抽样单位,从200个普查小区中简单随机抽取10个。在第二重抽样中,首先对第一重样本普查小区,按照住房单元规模分为两层:中型层,每个小区含住房单元70~90个;其他层,每个小区含住房单元70个以下或90个以上。然后,仍然以普查小区为抽样单

位, 分别从每一新层简单随机抽取 4 个和 2 个普查小区。为便于叙述, 这 6 个样本普查小区分别用 1、2、3、4、5、6 表示。抽样结果及样本普查小区的抽样权数, 请见表 2。

在三系统估计量的实证分析中, 需要选择恰当变量对总体人口分层(我国迄今尚未在人口普查净误差估计中对总体人口等概率分层)。基于加权优比排序法(胡桂华, 2015), 参照美国最近三次人口普查质量评估等概率分层方案, 同时结合我国人口实际状况, 本文理应选择房屋所有权、文化程度、婚姻状况、普查小区人口的流动性、年龄和性别, 对 7 个社区的 200 个普查小区的全部人口分层。这些变量交叉分层, 形成若干等概率人口层。然而受样本规模所限, 一些分层变量不得不舍弃, 或合并一些交叉层, 使每个等概率人口层的样本人口达到一定规模, 减少实际人口数估计值的抽样方差。但为便于全面演示三系统估计量及其抽样方差估计量的计算过程, 我们只是使用性别和年龄对 200 个小区的人口进行等概率分层。共分为 6 层: 男 0~14 岁; 男 15~59 岁; 男 60 岁及以上; 女 0~14 岁; 女 15~59 岁; 女 60 岁及以上。每个等概率人口层及各个单元的样本常住人口数分别见表 3 和表 4。

表 2 样本形成及抽样权数

N	n	w_i	g	n_g	r_g	w_g	α_{ig}
200	10	20	$g=1$	6	4	1.5	30
200	10	20	$g=2$	4	2	2	40

表 3 各个等概率人口层样本小区的未加权常住人口数 (单位: 人)

等概率人口层 v	样本普查小区					
	1	2	3	4	5	6
男 0~14 岁	22	24	33	18	20	20
男 15~59 岁	91	98	96	77	89	85
男 60 岁及以上	17	18	16	15	19	18
女 0~14 岁	19	22	21	15	19	19
女 15~59 岁	84	90	92	73	81	80
女 60 岁及以上	17	18	19	12	16	16
合计	250	270	277	210	244	238

表 4 样本小区的等概率人口层各个单元的未加权样本常住人口数 (单位: 人)

样本小区及等概率人口层	$\mathcal{X}_{111v}$	$\mathcal{X}_{112v}$	$\mathcal{X}_{122v}$	$\mathcal{X}_{121v}$	$\mathcal{X}_{211v}$	$\mathcal{X}_{212v}$	$\mathcal{X}_{221v}$
1 男 0~14 岁	8	4	1	3	3	1	2
1 男 15~59 岁	25	16	7	14	14	7	8
1 男 60 岁及以上	6	4	1	2	1	1	2
1 女 0~14 岁	6	4	0	3	3	1	2
1 女 15~59 岁	24	15	6	13	13	6	7
1 女 60 岁及以上	6	3	1	2	2	1	2
2 男 0~14 岁	7	5	2	4	3	1	2
2 男 15~59 岁	25	18	7	16	15	8	9
2 男 60 岁及以上	6	4	1	2	2	1	2
2 女 0~14 岁	8	4	1	3	3	1	2

(续)

样本小区及等概率人口层	x_{111v}	x_{112v}	x_{122v}	x_{121v}	x_{211v}	x_{212v}	x_{221v}
2女15~59岁	25	16	7	13	14	7	8
2女60岁及以上	6	4	1	3	2	0	2
3男0~14岁	7	4	2	3	3	2	2
3男15~59岁	25	18	7	16	15	7	8
3男60岁及以上	5	3	1	2	2	1	2
3女0~14岁	7	4	1	3	3	1	2
3女15~59岁	25	17	7	14	14	7	8
3女60岁及以上	6	3	2	3	2	1	2
4男0~14岁	6	3	1	3	2	1	2
4男15~59岁	22	14	5	12	13	5	6
4男60岁及以上	4	3	1	2	3	1	1
4女0~14岁	5	3	1	2	2	1	1
4女15~59岁	21	13	5	12	11	5	6
4女60岁及以上	4	2	1	1	2	1	1
5男0~14岁	6	4	1	3	3	1	2
5男15~59岁	24	16	6	15	15	6	7
5男60岁及以上	6	4	1	3	2	1	2
5女0~14岁	6	4	0	3	3	1	2
5女15~59岁	24	14	6	13	13	5	6
5女60岁及以上	5	3	1	3	2	1	1
6男0~14岁	6	4	1	3	3	1	2
6男15~59岁	24	15	5	14	14	6	7
6男60岁及以上	6	4	1	3	2	1	1
6女0~14岁	6	4	1	2	3	1	2
6女15~59岁	23	14	6	13	12	5	7
6女60岁及以上	5	3	1	2	3	1	1

2. 三系统估计量的选择

依据表2~表4数据,以及式(27)得到各个等概率人口层的 χ^2 值,见表5。

表5 各个等概率人口层的 χ^2 值

等概率人口层	基于模型2	基于模型6	基于模型12
男0~14岁	0	32.13	419.06
男15~59岁	0	169.90	1461.14
男60岁及以上	0	27.13	416.97
女0~14岁	0	91.83	450.75
女15~59岁	0	79.41	1277.49
女60岁及以上	0	13.95	298.26

给定显著性水平 $\alpha=0.05$ ，基于模型 6 的 $\chi^2_{2,0.05}=5.991$ ，基于模型 10 的 $\chi^2_{1,0.05}=3.841$ 。从表 5 可以看出，基于模型 6 和模型 12 的每个等概率人口层的 χ^2 值与基于模型 2 的 χ^2 值的差都大于显著性水平，因而模型 6 和模型 12 都不适合于拟合既定数据，因而选择模型 2 拟合既定数据。相应地，选择这种模型的三系统估计量，即用式 (29) 的三系统估计量估计总体实际人口数。事实上，模型 2 是饱和对数线性模型，自然适合于既定数据拟合。

3. 实际人口数估计值及净误差估计值

根据式 (29) 和式 (44)，以及表 2~表 4 数据，得到各个等概率人口层及总体的实际人口数、普查净误差及净误差率的估计值，见表 6。表 6 中的普查人数来源于重庆市南岸区南坪街道的 7 个社区组织的住户全面调查。

表 6 等概率人口层及总体净误差估计值

等概率人口层	普查人数 (人)	实际人数 (人)	净误差 (人)	净误差率 (%)
男 0~14 岁	4700	4855	155	3.19
男 15~59 岁	19753	20208	455	2.25
男 60 岁及以上	3990	4054	64	1.58
女 0~14 岁	4549	4696	147	3.13
女 15~59 岁	18315	18726	411	2.19
女 60 岁及以上	3616	3665	49	1.34
总体	54923	56204	1281	2.28

表 6 表明，总体净误差为 1281 人，它等于 6 个等概率人口层净误差的总和；总体净误差率为 2.28% (1281/56204)，表明每 100 人中平均有 2.28 人应该在普查中登记却未登记；各等概率人口层净误差率存在较大差异，反映出不同年龄、不同性别人口在普查中的登记概率存在差异。男性比女性的净误差率高，年龄小的比年龄大的净误差率高。在 6 个等概率人口层中，“男 0~14 岁”的净误差率最高，而“女 60 岁及以上”净误差率最低。这源于有些父母在普查登记中漏报孩子，尤其是超生婴儿。60 岁及以上女性相比男性和其他年龄的人口流动性小，在普查中漏报的可能性小；年龄和性别对净误差率有显著影响，表明用性别和年龄对总体人口进行等概率分层是合理的选择。

4. 抽样方差及协方差计算

依据表 2~表 4 数据，以及式 (43)、式 (45)、式 (47) 和式 (49) 得到各个等概率人口层及总体人口普查净误差估计值的抽样方差，见表 7。

表 7 等概率人口层及总体的抽样方差及协方差

等概率层	男 0~14 岁	男 15~59 岁	男 60 岁及以上	女 0~14 岁	女 15~59 岁	女 60 岁及以上	合计
男 0~14 岁	2496966	9166660	2267143	3772357	7664668	1452065	26819859
男 15~59 岁	9166660	33697805	8348824	13855001	28144991	5331435	98544716
男 60 岁及以上	2267143	8348824	2082548	3438259	6960015	1312740	24409529
女 0~14 岁	3772357	13855001	3438259	5712249	11581405	2188341	40547612
女 15~59 岁	7664668	28144991	6960015	11581405	23538942	4164231	82354252
女 60 岁及以上	1452065	5331435	1312740	2188341	4464231	855381	15604193
合计	26819859	98544716	24409529	40547612	82354252	15604193	288280161