doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 013

基于正确登记概率的普查直接多报估计

吴 婷[1], 胡桂华[2]
(1. 重庆市万州区疾病预防控制中心, 重庆 404000;
2. 重庆工商大学 数学与统计学院, 重庆 400067)

摘 要:基于现有的普查直接多报估计量,在二重抽样调查下,对样本小区登记人口数梳理分类,构建基于正确登记概率的普查直接多报估计量,以解决前者由于计数对象不确定而造成的多报虚增问题。创新之处在于:构建样本小区下登记人口数的指标体系,提出基于正确登记概率的普查直接多报估计量、发生重报估计量和重报估计量,并将普查直接多报估计量与基于正确概率的普查直接多报估计量进行比较研究。理论研究和实证研究得出:普查直接多报估计值及抽样标准误差都高于基于正确登记概率的普查直接多报估计值及抽样标准误差,普查直接多报估计量虚增了普查多报人口数,基于正确登记概率的普查直接多报估计量精度更高。

关键词:抽样调查;质量评估调查;普查多报
中图分类号:C924. 25 **文献标志码**:A **文章编号**:1672-058X(2022)02-0090-09

0 引 言

人口普查目标是不重不漏登记普查目标总体内的每一个人,然而由于人口流动等原因,这一目标难以实现。在每次人口普查中,总是会登记普查目标总体之外的人,即普查多报人口,包括重报及其他普查多报。近几十年随着时代的变迁,重报问题变得越来越刻不容缓,其中美国、加拿大、英国、澳大利亚最近几次人口普查都十分重视重报的估计[1-4]。张广宇等[5]提到:在中国的人口普查中,户籍登记和身份证制度纳入了人口普查工作,并且将流动人口在现住地和常住地都进行登记,这样的工作可能会降低普查漏报的概率,但是却加大了重报的可能性。

2020 年,我国进行了第七次全国人口普查,将对普查登记质量进行评估,多报属于人口普查质量评估的一部分,本文将深入研究普查多报。胡桂华等[6]提到普查多报又可分为重报与其他多报,在进行人口普查时,由于有些被调查者有多个住所,且对于普查的不重视或理解不到位,从而导致在多个地方进行了重复登记。对于重复登记的人,不仅要找到该人的重复登记记录,还要找到他的正确登记地址。对于普查日在多个住所活动,并进行了普查重复登记的人口,如何进行检测和解决重复登记是一大难题。

普查重报有两种统计口径,一是发生重报人数,即属于目标总体的某人不管登记了几次,只要是登记了 1 次以上,都看作 1。该人登记了多个地址,发生重报人口数统计时,统计的是应该登记地址的发生重报人数,属于真正计数的人数且属于目标总体。二是普查重报人数,即属于目标总体的某人登记次数减 1,统计的是应该登记地址的重报人数,属于重复登记的次数。而其他多报指普查员登记普查时点实际上不存在的人,例如,在普查表中登记宠物、普

收稿日期:2021-01-04;**修回日期**:2021-02-28.
基金项目:重庆工商大学 2020 年研究生创新型科研项目(YJSCXX2020-094-77).
作者简介:吴婷(1995—),女,重庆云阳人,硕士研究生,从事人口普查质量评估研究.

查时点之前死亡的人口或普查时点之后出生的人口。同一人在人口普查中登记不止一次时,被认为是重复的。重报不止发生在目标总体内,也会发生在目标总体外,但若不属于目标总体的人进行了重复登记,视为其他多报。

由文献[7]可知,目前估计多报的主要方法是利用样本普查小区多报人口及其抽样权数构造“普查直接多报估计量”。但是构造这种估计量存在两个缺陷,一是若将研究范围进行限定,在研究范围外进行了登记或多次登记,则研究范围内的多报、重报及发生重报该如何统计;二是人口流动性大,实际工作中会出现重复者应该登记位置不易确定,多报、重报及发生重报该如何统计的问题。本文目标是在有限总体概率抽样及普查多报估计理论基础上,构造基于正确登记概率的普查直接多报估计量、基于正确登记概率的普查直接重报估计量、基于正确登记概率的普查直接发生重报估计量。

基于正确登记概率的普查直接多报估计量是对于多报中的计数对象不确定哪次登记属于目标总体人口的计数而建立的,本估计量将解决上述普查直接多报估计量的两个缺陷问题。如某人在研究范围内的不同地方登记了 3 次,但是正确登记地址不确定,对于普查直接发生重报,是在 3 个地方都算作发生重报数,就将发生重报数算作了 3,实际该人发生重报数为 1;普查直接重报,是在 3 个地方都算了重报数,每个地方算的重报数都为 2,总共就算作了 6,实际该人重报数为 2。所以,普查直接重报和普查直接发生重报分别造成了重报数和发生重报数的虚增。基于正确登记概率的普查直接多报估计量和普查直接多报估计量都属于复杂估计量,将采用分层刀切法计算其抽样方差[8-9],将两个估计量的精度进行比较研究,为人口普查多报估计提出更符合实际的方法。

1　普查多报估计指标的构建

采取分层二重抽样进行多报研究,第一重样本采取分层整群抽样抽取,抽样单位为普查小区。用 H 表示第一重样本抽样层的总层数,h 为任意层,N_h 为层 h 的普查小区总数,n_h 为层 h 抽取的样本普查小区总数;g 为第二重抽样层,α_{hg} 为样本普查小区集合,M_{hg} 为层 hg 普查小区总数,m_{hg} 为层 hg 样本普查小区总数。

进行分层二重抽样后得到样本小区 i,将样本小区 i 的人口普查名单和事后质量抽查名单进行数据处理和数据比对后得到表 1 中的指标,为构造正确登记概率的普查直接多报估计量提供基础。

表 1　层 *hg* 第 *i* 样本小区登记人口数的分类

Table 1　Classification of registered population of the *i*-th sample communities in layer *hg*

总　体	分　类
计数对象属于目标总体人口的计数确定	发生重报人数 c_{1hgi}
	不发生重报人数 c_{0hgi}
计数对象不属于目标总体人口的计数确定	重报人数 d_{1hgi}
	其他多报人数 d_{0hgi}
计数对象不确定哪次登记属于目标总体人口的计数	发生重报概率人数 p_{c1hgi}
	重报概率人数 p_{c2hgi}

根据表 1,将普查登记人口数分为三大类,计数对象属于目标总体人口的计数确定、计数对象不属于目标总体人口的计数确定和计数对象不确定哪次登记属于目标总体人口的计数。

将计数对象属于目标总体人口的计数确定进行分类得到发生重报人数 c_{1hgi} 和不发生重报人数 c_{0hgi},这两部分是属于目标总体样本小区的确定人口数。发生重报的人口数 c_{1hgi},指在本小区进行了一次登记且在本小区外研究范围内的其他地方进行了登记,但是正确登记地址属于本小区的人数。不发生重报人口数 c_{0hgi},指被计数者在本小区登记一次且未在研究范围内其他地方进行登记,正确登记地址在本小区的人数。

将计数对象不属于目标总体人口的计数确定进行分类得到重报人数 d_{1hgi} 和其他多报人数 d_{0hgi},这两部分是不属于目标总体样本小区的确定人口数。重报人数 d_{1hgi},指在本小区进行了一次登记且在本小区外研究范围内的其他地方进行了登记,但是正确登记地址不属于本小区的人数。其他多报人数 d_{0hgi},指普查员将研究范围内普查时点前死亡的人口、普查时点后出生的人口或宠物进行了登记。

计数对象不确定哪次登记属于目标总体人口的计数,根据重报的两种统计口径分为发生重报概率人数 p_{c1hgi} 和重报概率人数 p_{c2hgi}。发生重报概率人数 p_{c1hgi},指在本小区进行了一次登记且在本小区外研究

范围内的其他地方也进行了登记,但是正确登记地址不确定是否属于本小区的人数,将根据式(1)进行概率登记,统计的是发生重报的概率。发生重报概率人数 p_{c2hgi},指在本小区进行了一次登记且在本小区外研究范围内的其他地方也进行了登记,但是正确登记地址不确定是否属于本小区的人数,将根据式(2)进行概率登记,统计的是重报的概率。

对于登记地址不确定者,如某人进行了多次普查登记,但不确定哪一次登记才算做应该登记,若在每个地方都算作重报或发生重报,将虚增多报人口数。对于这种情况,将在每个地址都进行概率登记处理。

重报者 A 每个登记地址发生重报应该登记的概率为 $P_{A,Duplicate1}$,*Duplicate*1 表示发生重报,对于重报者 A,n 表示对于重报者 A 普查登记次数,发生重报概率:

$$P_{A,Duplicate1}=1/n\ (n>1) \tag{1}$$

重报者 A 每个登记地址重报概率为 $P_{A,Duplicate2}$,*Duplicate*2 表示重报,对于重报者 A,n 表示对于重报者普查登记次数,重报概率:

$$P_{A,Duplicate2}=(n-1)/n\ (n>1) \tag{2}$$

当限定了研究范围时,重报者 A 发生重报应该登记的概率为 $P^{loc}_{A,Duplicate1}$,*loc* 表示研究的范围,*Duplicate*1 表示发生重报,对于重报者 A,$n(n>1)$ 表示重报者普查登记次数,$m(m>1)$ 表示重报者研究范围内普查登记次数,研究范围内发生重报概率:

$$P^{loc}_{A,Duplicate1}=\begin{cases}1/n,\text{如果普查登记了 } n \text{ 次,}\\ \quad \text{每次都在研究范围内}\\ 1/m,\text{ 如果普查登记了 } n \text{ 次,}\\ \quad \text{有 } m \text{ 次都在研究范围内}\\ 0,\text{ 如果普查登记了 } n \text{ 次,}\\ \quad \text{有 } n-1 \text{ 次都在研究范围外}\end{cases} \tag{3}$$

在限定了研究范围时,重报者 A 应该登记的概率为 $P^{loc}_{A,Duplicate2}$,*loc* 表示研究的范围,*Duplicate*2 表示重报,对于重报者 A,$n(n>1)$ 表示对于重报者普查登记次数,$m(m>1)$ 表示重报者研究范围内普查登记次数,研究范围内重报概率:

$$P^{loc}_{A,D}=\begin{cases}(n-1)/n,\text{如果普查登记了 } n \text{ 次,}\\ \quad \text{每次都在研究范围内}\\ (m-1)/m,\text{如果普查登记了 } n \text{ 次,}\\ \quad \text{有 } m \text{ 次都在研究范围内}\\ 0,\text{ 如果普查登记了 } n \text{ 次,}\\ \quad \text{有 } n-1 \text{ 次都在研究范围外}\end{cases} \tag{4}$$

其中,D 表示 *Duplicate*2。

2 普查多报估计量的构造

2.1 普查直接多报估计量的构造

普查直接多报估计量是目前大多数国家进行多报估计中采用的估计量,用 $\widehat{DE}_3$ 表示,将 $\widehat{DE}_1$ 和 $\widehat{DE}_2$ 分别表示普查直接发生重报估计量及普查直接重报估计量。

在普查直接多报估计中,表 1 的计数对象不确定哪次登记属于目标总体人口的计数时,根据重报的两种统计口径,分为直接发生重报 e_{1hgi} 和直接重报 e_{2hgi}。根据表 1 建立的相关指标得到普查直接发生重报估计量、普查直接重报估计量、普查直接其他多报估计量和普查直接多报估计量。

二重抽样后,用 b_{hgi} 表示示性函数,如果第一重样本普查小区 i 属于层 g,则 $b_{hgi}=1$,否则 $b_{hgi}=0$。用 I_{hgi} 表示另外一个示性函数,如果第一重样本普查小区 i 进入 α_{hg},则 $I_{hgi}=1$,否则 $I_{hgi}=0$。样本小区 i 的抽样权数 α_{hgi} 为 $(N_h/n_h)(M_{hg}/m_{hg})$。

$$\widehat{DE}_1=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(c_{1hgi}+e_{1hgi}) \tag{5}$$

式(5)是构造的普查直接发生重报估计量,是重报的口径之一,统计人口普查中有多少人口发生了重复登记。

$$\widehat{DE}_2=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(d_{1hgi}+e_{2hgi}) \tag{6}$$

式(6)是构造的普查直接重报估计量,体现哪些重复登记导致实际普查人口数的增加,每一个重复登记者重复登记的次数不一致,并且重复登记的原因是不同的,该估计量可以获得因为重复登记导致的实际人口虚增的人口数。

$$\widehat{DE}_3=\sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}d_{0hgi} \tag{7}$$

式(7)构造的是普查直接其他多报估计量,目的是估计总体中的其他多报人口数,该估计量也导致实际人口数的增加,但不是由于重复登记导致的增加,而是登记了不属于普查时点的人口数,如宠物、普查时点前死亡的人口、普查时点后出生的人口。有些国家觉得其他多报人口数少,于是没有估计或者单独估计。但其他多报属于多报的一部分,在多报估计工作中也不应忽视它的存在。

$$\widehat{DE}_4 = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(d_{1hgi} + e_{2hgi} + d_{0hgi}) \tag{8}$$

式(8)是式(6)和式(7)的总和,代表的是总体中人口普查重报与其他多报的人口数,该指标能获知普查人口相对于实际人口虚增的人口数。

2.2　基于正确登记概率的普查直接多报估计量的构造

基于正确登记概率的普查直接多报估计量,用 $\widehat{PE}_3$ 表示,将 $\widehat{PE}_1$ 和 $\widehat{PE}_2$ 分别表示基于正确登记概率的普查直接发生重报估计量及基于正确登记概率的普查直接重报估计量。根据表 1 建立的相关指标得到如下估计量:

$$\widehat{PE}_1 = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(c_{1hgi} + p_{c1hgi}) \tag{9}$$

式(9)构造的是基于正确登记概率的普查直接发生重报估计量,该估计量相对于式(5)构造的普查直接发生重报估计量,对计数对象不确定哪次登记属于目标总体的人口,进行发生重报概率登记,以免造成发生重报的估计增多。

$$\widehat{PE}_2 = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(d_{1hgi} + p_{c2hgi}) \tag{10}$$

式(10)构造的是基于正确登记概率的普查直接重报估计量,该估计量相对于式(6)构造的普查直接重报估计量,对计数对象不确定哪次登记属于目标总体的人口,进行重报概率登记,以免造成重报的估计增多。

$$\widehat{PE}_3 = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}b_{hgi}I_{hgi}(d_{1hgi} + p_{c2hgi} + d_{0hgi}) \tag{11}$$

式(11)构造的是基于正确登记概率的普查直接多报估计量,该估计量相对于式(8)构造的普查直接多报估计量,对重复计数对象不确定哪次登记属于目标总体的人口,进行重报概率登记,以免造成多报的估计增多。

3　普查多报估计量的方差估计

在构造式(5)——式(11)之后,所要做的工作是计算它们的抽样方差。下面将对普查直接多报估计量及其相关估计量和基于正确登记概率的普查直接多报估计量及其相关估计量的抽样方差估计量进行构造。

先计算样本普查小区的复制抽样权数 $\alpha_{hgi}^{(st)}$,计算如式(12)。复制权数是指剔除第一重抽样层 s 的 t 小区后重新计算的第一重样本普查小区,其中包括进入第二重样本普查小区的抽样权数,被剔除的第一重样本小区的复制权数为零。h 层 g 子层 i 普查小区切断普查小区 t 后复制权数 $\alpha_{hgi}^{(st)}$ 为以下几种情况(式(12)):

$$\begin{cases}\dfrac{N_h}{n_h}\dfrac{M_{hg}}{m_{hg}}, h\neq s\\ \dfrac{N_h}{n_h}\dfrac{M_{hg}}{m_{hg}}\dfrac{n_h}{(n_h-1)}, h=s, b_{sgt}=0\\ \dfrac{N_h}{n_h}\dfrac{M_{hg}}{m_{hg}}\dfrac{n_h}{(n_h-1)}\dfrac{(M_{hg}-1)}{M_{hg}}, h=s, b_{sgt}=1, I_{sgt}=0, i\neq t\\ \dfrac{N_h}{n_h}\dfrac{M_{hg}}{m_{hg}}\dfrac{n_h}{(n_h-1)}\dfrac{(M_{hg}-1)}{M_{hg}}\dfrac{m_{hg}}{(m_{hg}-1)}, h=s, b_{sgt}=1, I_{sgt}=1, i\neq t\\ 0, h=s, i=t\end{cases} \tag{12}$$

3.1　普查直接多报估计量的方差估计

下面构造普查直接多报估计量及相关估计量的方差,先构造各估计量 $\widehat{DE}_1, \widehat{DE}_2, \widehat{DE}_3, \widehat{DE}_4$ 的复制估计量,即 $\widehat{DE}_1^{(st)}, \widehat{DE}_2^{(st)}, \widehat{DE}_3^{(st)}, \widehat{DE}_4^{(st)}$;然后计算各估计量的抽样方差即 $var(\widehat{DE}_1)$, $var(\widehat{DE}_2)$, $var(\widehat{DE}_3)$, $var(\widehat{DE}_4)$。

$\widehat{DE}_1$ 的复制估计量和抽样方差分别为式(13)和式(14):

$$\widehat{DE}_1^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)}b_{hgi}I_{hgi}(c_{1hgi} + e_{1hgi}) \tag{13}$$

$$var(\widehat{DE}_1) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{n_h - 1}{n_h} \times (\widehat{DE}_1^{(st)} - \widehat{DE}_1)^2 \tag{14}$$

$\widehat{DE}_2$ 的复制估计量和抽样方差分别为式(15)和式(16):

$$\widehat{DE}_2^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)}b_{hgi}I_{hgi}(d_{1hgi} + e_{2hgi}) \tag{15}$$

$$var(\widehat{DE}_2) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{n_h - 1}{n_h} \times$$

$$(\widehat{DE}_2^{(st)} - \widehat{DE}_2)^2 \tag{16}$$

$\widehat{DE}_3$ 的复制估计量和抽样方差分别为式(17)和式(18):

$$\widehat{DE}_3^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)} b_{hgi} I_{hgi} d_{0hgi} \tag{17}$$

$$var(\widehat{DE}_3) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\times (\widehat{DE}_3^{(st)} - \widehat{DE}_3)^2 \tag{18}$$

$\widehat{DE}_4$ 的复制估计量和抽样方差分别为式(19)和式(20):

$$\widehat{DE}_4^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)} b_{hgi} I_{hgi}(d_{1hgi} + e_{2hgi} + d_{0hgi}) \tag{19}$$

$$var(\widehat{DE}_4) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\times (\widehat{DE}_4^{(st)} - \widehat{DE}_4)^2 \tag{20}$$

3.2 基于正确登记概率的普查直接多报估计量的方差估计

下面估计基于正确登记概率的普查直接多报估计量及相关估计量的方差,先构造各估计量 $\widehat{PE}_1$, $\widehat{PE}_2$,$\widehat{PE}_3$ 的复制估计量,即 $\widehat{PE}_1^{(st)}$,$\widehat{PE}_2^{(st)}$,$\widehat{PE}_3^{(st)}$;然后计算各估计量的抽样方差即 $var(\widehat{PE}_1)$, $var(\widehat{PE}_2)$,$var(\widehat{PE}_3)$。

$\widehat{PE}_1$ 的复制估计量和抽样方差分别为式(21)和式(22):

$$\widehat{PE}_1^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)} b_{hgi} I_{hgi}(c_{1hgi} + p_{c1hgi}) \tag{21}$$

$$var(\widehat{PE}_1) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\times (\widehat{PE}_1^{(st)} - \widehat{PE}_1)^2 \tag{22}$$

$\widehat{PE}_2$ 的复制估计量和抽样方差分别为式(23)和式(24):

$$\widehat{PE}_2^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)} b_{hgi} I_{hgi}(d_{1hgi} + p_{c2hgi}) \tag{23}$$

$$var(\widehat{PE}_2) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\times (\widehat{PE}_2^{(st)} - \widehat{PE}_2)^2 \tag{24}$$

$\widehat{PE}_3$ 的复制估计量和抽样方差分别为式(25)和式(26):

$$\widehat{DE}_3^{(st)} = \sum_{h=1}^{H}\sum_{g=1}^{G}\sum_{i=1}^{n_h}\alpha_{hgi}^{(st)} b_{hgi} I_{hgi}(d_{1hgi} + p_{c2hgi} + d_{0hgi}) \tag{25}$$

$$var(\widehat{PE}_3) = \sum_{h=1}^{H}\sum_{t=1}^{n_h}\left(1-\frac{n_h}{N_h}\right)\frac{n_h-1}{n_h}\times (\widehat{PE}_3^{(st)} - \widehat{PE}_3)^2 \tag{26}$$

4 实证分析

4.1 样本数据

本文的实证数据主要来源于调查数据,其调查的实证对象是广西南宁市西乡塘区的某行政区,采取了分层二重抽样方法,所抽取的样本见表 2。

在表 2 中,该行政区分为两层:社区层和行政村层,分别用 $h=1$ 和 $h=2$ 表示。社区层共有普查小区 1 000 个,行政村层共有普查小区 1 100 个。从社区层和行政村层分别简单随机抽取 10 个和 9 个普查小区。按照调查难度,将第一重样本普查小区分为 3 层,即容易调查层、中等难度调查层和调查难度大层,分别用符号 $g=1$,$g=2$ 和 $g=3$ 表示。所有样本普查小区及其个人 100%提供答复,此时样本个人抽样权数等于样本普查小区抽样权数。抽样层、抽样权数及样本数据见表 2 和表 3。其中,N_h 和 n_h 分别表示层 h 的普查小区总数及样本普查小区数,M_{hg} 和 m_{hg} 分别表示层 hg 的普查小区数及从中抽取的第二重样本普查小区数,i 表示样本普查小区,表 3 中的(2)表示第一重样本普查小区进入到第二重样本。

表 2 二重抽样样本分布情况表

Table 2 Distribution of double sampling samples

h	N_h	n_h	ω_{hi}	g	M_{hg}	m_{hg}	α_{hgi}
				1	3	2	150
1	1 000	10	100	2	3	2	150
				3	4	2	200
				1	3	2	183
2	1 100	9	122	2	3	2	183
				3	3	2	183

表 3　抽样层及第二重样本数据

Table 3　Sampling layer and second sample data

h	g	i	计数对象属于目标总体人口的计数		计数对象不属于目标总体人口的计数	
			发生重报人数 c_{1hgi}	不发生重报人数 c_{0hgi}	重报人数 d_{1hgi}	其他多报人数 d_{0hgi}
1	1	1 (2)	1	249	0	0
1	1	2 (2)	0	263	0	0
1	2	4 (2)	0	241	1	1
1	2	5 (2)	0	251	0	0
1	3	7 (2)	1	231	0	1
1	3	8 (2)	0	262	1	0
2	1	11 (2)	1	251	0	0
2	1	12 (2)	0	250	0	1
2	2	14 (2)	1	260	0	0
2	2	15 (2)	0	250	1	0
2	3	17 (2)	0	248	0	0
2	3	18 (2)	0	254	1	0

在抽取的样本小区中,只有样本小区 1、样本小区 5、样本小区 14、样本小区 17 有人存在重复登记的计数对象不确定是否属于目标总体,他们属于在普查研究范围内的登记次数超过 1,并且不确定哪次登记属于目标总体人口。根据上文中指标解释,基于式(3)和式(4),得出普查直接多报估计量和基于正确登记概率的普查直接多报估计量所需要的样本数据资料(表 4)。

表 4　计数对象不确定哪次登记属于目标总体的名单

Table 4　List of counting objects not sure which registration belongs to the target population

样本小区 i	样本人	普查研究范围内登记的次数	发生重报概率 p_{c1hgi}	重报概率 p_{c2hgi}	普查直接发生重报 e_{1hgi}	普查直接重报 e_{2hgi}
1	A	3	1/3+1/2	2/3+1/2	2	5
1	B	2				
5	C	3	1/3	2/3	1	3
14	D	2	1/2+1/2	1/2+1/2	2	4
14	E	2				
17	F	3	1/3+1/2	2/3+1/2	2	5
17	G	2				

4.2　基于正确登记概率的普查直接多报估计值

根据表 2—表 5,式(5)—式(26),得到基于正确登记概率的普查直接多报估计量及其相关估计量的多报估计值和抽样方差估计值。本文由于表格的篇幅过长,只列出 $\widehat{DE}_1$ 的抽样方差计算表格,如表 5 所示。

表 5 $\widehat{DE}_1$ 的抽样方差计算

Table 5 Sampling variance calculation of $\widehat{DE}_1$

被剔除的层 s 的小区 t	N_h	n_h	$\widehat{DE}_1^{(st)}$	$\left(1-\frac{n_h}{N_h}\right)\left(\frac{n_h-1}{n_h}\right)(\widehat{DE}_1^{(st)}-\widehat{DE}_1)^2$
0	—	—	1 898	—
1	1 000	10	1 489	149 047
2	1 000	10	2 155	58 850
3	1 000	10	1 822	5 146
4	1 000	10	2 045	19 254
5	1 000	10	1 823	5 012
6	1 000	10	1 934	1 155
7	1 000	10	1 766	15 525
8	1 000	10	2 099	35 997
9	1 000	10	1 933	1 091
10	1 000	10	1 933	1 091
11	1 100	9	1 830	4 077
12	1 100	9	2 105	37 776
13	1 100	9	1 967	4 197
14	1 100	9	1 418	203 124
15	1 100	9	2 243	104 934
16	1 100	9	1 829	4 197
17	1 100	9	1 624	66 188
18	1 100	9	2 174	67 158
19	1 100	9	1 898	0
总和	—	—	—	783 821

表 5 是普查直接发生重报估计量 $\widehat{DE}_1$ 的抽样方差计算。首先,根据表 3 和表 4 的数据,利用式(5)计算普查直接发生重报估计量的估计值。

计算 $\widehat{DE}_1$ 的抽样方差时,使用式(12)计算复制权数,并计算切掉第一重抽样层 s 的 t 小区后普查小区条件下的 $\widehat{DE}_1^{(st)}$,根据式(13),经过轮换 19 次刀切,每次刀切后计算 $\widehat{DE}_1^{(st)}$ 的值,最后根据式(14)计算 $\widehat{DE}_1$ 的方差估计值,即普查发生重报人口数的方差估计值 $var(\widehat{DE}_1)$。

从表 5 的数据可以看出:总体普查发生重报人口数 1 898 人的抽样方差为 783 821,抽样标准差为 885 人。这表明,平均每个样本估计的总体普查发生重报人口数为 1 898 人,相应的抽样平均标准误差为 885 人,即每个样本估计总体普查发生重报人口数与总体实际普查发生重报人口数的平均差异为 885 人。

4.3 两个多报及多报率估计量的比较

现在进行这两个多报估计量数据精度上的比较。在这之前,把普查直接多报估计值和基于正确登记概率的普查直接多报估计值及其抽样方差估计值统一列在表 6 中。

表 6 普查多报估计值

Table 6 Estimates of overcoverage in census

多报指标	估计人数	抽样方差	抽样标准误差
普查直接发生重报人数	1 898	783 821	885
普查直接重报人数	3 563	1 859 541	1 364
普查直接其他多报人数	533	118 707	345
普查直接多报人数	4 096	1 670 040	1 292
基于正确登记概率的普查直接发生重报人数	1 227	334 882	579
基于正确登记概率的普查直接重报人数	1 388	120 920	348
基于正确登记概率的普查直接多报人数	1 921	122 657	350

从表 6 可以得出:普查直接发生重报估计量估计的发生重报人口数为 1 898 人,抽样标准误差为 885 人,而基于正确登记概率的普查直接发生重报估计量估计的发生重报人口数为 1 227 人,抽样标准误差为 579 人,普查直接发生重报估计值及抽样标准误差都高于基于正确登记概率的普查直接发生重报估计值及抽样标准误差,普查直接发生重报人数虚增了发生重报人口数;普查直接重报估计量估计的重报人口数为 3 563 人,抽样标准误差为 1 364 人,而基于正确登记概率的普查直接重报估计量估计的重报人口数为 1 388 人,抽样标准误差为 348 人,普查直接重报估计值及抽样标准误差都高于基于正确登记概率的普查直接重报估计值及抽样标准误差,前者虚增了重报人口数;普查直接其他多报估计量估计的其他多报人数为 533 人,抽样标准误差为 345 人,普查直接其他多报人数估计的是在普查中将普查时点前死亡人口、普查时点后出生人口或宠物登记的数目,且包含这些登记的重复登记数目;普查直接多报估计量估计的多报总人口数为 4 096 人,抽样标准误差为 1 292 人,而基于正确登记概率的普查直接多报估计量提供的多报人口总数为 1 921 人,抽样标准误差为 350 人,它们都是各自统计的重报人口与其他多报人口之和,普查直接多报估计值及抽样标准误差都高于基于正确登记概率的普查直接多报估计值及抽样标准误差,普查直接多报人数虚增了多报人口数。

5　结　论

本文以广西南宁市西乡塘区的一个行政区为观测对象,从现有的国际多报研究基础上,通过对多报、重报及其他多报定义的明确,构造直接普查多报估计量的模型,并构造直接其他多报估计量,及从重报的两种口径构造的重报估计量。本文新建的另一种多报模型是构造基于正确登记概率的普查直接多报估计量,并从重报的两种口径构造基于正确登记概率的重报估计量及发生重报估计量。本文对两种普查多报模型进行理论和实证研究后,最终得出如下结论:

基于正确登记概率的普查直接多报估计量 $\widehat{PE}_3$ 提供的普查多报估计值为 1 921 人,普查直接多报估计量 $\widehat{DE}_4$ 提供的普查多报估计值为 4 096 人。实证研究表明,基于正确登记概率的直接多报估计量提供的精度更高,提供的估计值接近于实际值,选择基于正确登记概率的普查直接多报估计量更合理,普查直接多报估计量虚增了多报人口数。

重报的两种口径:发生重报和重报。 $\widehat{PE}_1$ 提供在普查中发生重报的有 1 227 人,不影响总体普查登记人口数。表明在本次普查中,总共有 1 227 人在普查中登记一次以上。通过这个指标的估计,能够查明哪些人容易发生重报。普查多报估计的主要目的就是查明重报者和非重报者特征。 $\widehat{PE}_2$ 提供在普查中总体重报人口数为 1 388 人,相当于在重复登记中虚增普查登记人口数 1 388 人。发生重报估计值小于重报估计值,原因在于后者统计的是重报的次数,而前者统计的是发生重报的人数,只有当样本中每个重报人口的重报次数为 1,两者才相等。例如,某人在普查中填写普查表 5 次,那么重报次数为 4,发生重报的人数为 1。而普查直接发生重报估计值 $\widehat{DE}_1$ 虚增了发生重报人数,普查直接重报估计值 $\widehat{DE}_2$ 虚增了重报人数。所以在进行重报估计时,使用基于正确登记概率的普查直接多报估计量估计总体普查重报和发生重报人口数。

基于正确登记概率的重报估计量 $\widehat{PE}_2$ 提供在本次普查中重报的次数为 1 388 人,$\widehat{DE}_3$ 提供的本次普查其他多报人口数为 533 人。这一方面表明,相比重报,其他多报人口数少很多,重报是普查多报的主要来源。另外一方面也说明,其他多报是客观存在的,忽视其他多报会低估总体普查多报人口数。在普查多报估计中,应该分别估计重报和其他多报人口数,以及普查多报人口总数。也就是说,在普查多报估计研究报告中,分别提供这 3 种普查多报人口数。值得注意的是,重报只是针对普查目标总体内的普查登记。虽然普查目标总体外的普查登记也可能登记一次以上,但归于其他普查多报。

为了比较普查直接多报估计量和基于正确登记概率的普查直接多报估计量的估计精度,并且为了使得这种比较具有可比性,统一使用刀切法近似计算其抽样方差。

参考文献(References):

[1] MULE T. Census coverage measurement estimation report: summary of estimates of coverage for persons in the United States[R]. Washington, DC: US Census Bureau, 2012.

[2] STATISTICS CANADA. 2011 census technical report:

coverage[R]. Ottawa: Statistics Canada, 2015.
[3] OFFICE FOR NATIONAL STATISTICS. 2011 census: methods and quality report: overcount estimation and adjustment[R]. London: Office for National Statistics, 2012.
[4] AUSTRALIAN BUREAU OF STATISTICS. Census of population and housing: details of overcount and undercount[R]. Canberra: Australian Bureau of Statistics, 2017.
[5] 张广宇, 顾宝昌. 人口重报:人口普查面临的新挑战[J]. 人口与经济, 2018(3): 1—12.
ZHANG Guang-yu, GU Bao-chang. Overcoverage in census: a new challenge [J]. Population & Economics, 2018(3): 1—12.
[6] 胡桂华, 吴婷, 廖金盆, 等. 人口普查多报及重报估计[J]. 统计与信息论坛, 2019, 34(8): 104—112.
HU Gui-hua, WU Ting, LIAO Jin-pen, et al. Estimation for erroneous census enumerations and duplicates[J]. Statistics & Information Forum, 2019, 34 (8): 104—112.
[7] 冯乃林, 李希如, 武洁, 等. 人口普查的事后质量抽查[R]. 北京:国家统计局人口和就业统计司, 2012.
FENG Nai-lin, LI Xi-ru, WU Jie,et al. Post census quality sampling[R]. Beijing: Department of Population and Employment Statistics, National Bureau of Statistics, 2012.
[8] EFRON B. The jackknife, the bootstrap, and other resampling plans[M]. Philadelphia:Society for Industrial & Applied Mathematics, 1982.
[9] HAGAN A, MURPHY T B, SCRRUCCA L. Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap, and weighted likelihood bootstrap[J]. Computational Statistics, 2019, 34(4): 1779—1813.

Estimation of Erroneous Enumerations in Population Census Based on Correct Registration Probability

WU Ting[1], HU Gui-hua[2]

(1. Wanzhou District Center for Disease Control and Prevention, Chongqing 404000, China;
2. School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: Based on the existing census direct estimator of erroneous enumerations, this paper classifies the registered population in the sample area under the double sampling survey, and constructs the census direct erroneous enumerations based on the correct registration probability, so as to solve the problem of erroneous enumerations and false increase caused by the uncertainty of the target population. The innovation lies in constructing the index system of the number of registered populations in the sample area, putting forward the direct erroneous enumerations, occurrence duplicate enumerations and duplicate enumerations of census based on the correct registration probability, and comparing the direct erroneous enumerations estimate of the census with the direct erroneous enumerations estimate based on the correct probability. Theoretical and empirical studies show that the direct overestimation and sampling standard error of census are higher than the direct overestimation and sampling standard error of census based on correct registration probability. The results show that the direct erroneous enumerations falsely increase the number of erroneous enumerations, and the accuracy of the direct erroneous enumerations based on the correct registration probability is higher.

Key words: sampling survey; post-enumeration survey; erroneous enumerations

责任编辑:李翠薇

引用本文/Cite this paper:
吴婷,胡桂华.基于正确登记概率的普查直接多报估计 .[J].重庆工商大学学报(自然科学版),2022,39(2):90—98.
WU Ting,HU Gui-hua. Estimation of erroneous enumerations in population census based on correct registration probability[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):90—98.