中国农民工抽样调查方案的设计

段景辉，陈珍珍
（厦门大学 经济学院，福建 厦门 361005）

摘 要：中国农民工是一个流动的群体，对其进行抽样调查的方案设计对估计总体结果的影响很大。文章系统介绍了中国农民工抽样调查中采用的多目标分层复合抽样设计方案，并给出了多目标下总和、均值和比率的估计公式。

关键词：农民工统计；抽样设计；多目标分层次抽样
中图分类号：F222.1 文献标识码：A 文章编号：1002-6487(2008)24-0025-02

自改革开放以来，中国的社会分层演变出一个新的阶层—农民工。农民工是指具有农村户口，却在城镇务工的劳动者，是中国现代化进程加快和现行户籍制度严重冲突所产生的客观结果。他们离开土地，但又没有完全融入城市，是区别于传统市民和农民的新生群体。近年来，农民工的收入水平、工作稳定性、工作安全性、子女教育、看病就医、精神文化生活等方面已经引起社会广泛的关注。为了了解目前中国农民工的工作和生活情况，我们拟设计出适当的抽样调查方案，对大量吸引农民工的华南地区、长江三角洲地区、东北地区、西北地区和华北地区进行抽样调查，力图能够反映中国农民工的总体情况。

1 抽样调查组织形式

为了使调查具有充分的代表性，在固定经费下保证最大的准确度，我们首先对大量吸引农民工的五个地区分别进行了初步调查，发现广东省是华南地区中最大的农民工流向地，该省吸纳全国最多的农民工；在长江三角洲地区，上海吸引农民工的能力最强，拥有相当数量的农民工；东北地区以辽宁省农民工人数最多；北京是中国的政治和经济文化中心，劳动力需求旺盛，自然吸引大量的农民工流入，是华北地区吸收农民工的最大市场；随着西部大开发政策的实施，农民工慢慢流向西部，新疆省在西北地区拥有不可忽视的农民工数量。

鉴于以上初步调查结果，我们设定广东省、上海市、北京市、辽宁省和新疆区的农民工作为调查总体，其中每个省市又可以看作一个子总体，拟分别进行抽样调查。每个农民工都是一个调查单元，按照科学、效率和便利的原则，对总体采用多目标分层复合抽样方法。具体步骤有三：首先，取得各个省市有关农民工的姓名、性别、户籍等信息，形成完备的抽样框，利用均匀分布给抽样总体中的每一个农民工配上一个分布在(0,1)之间的永久随机数，在随后的抽样过程中该随机数将不再变化。考虑到农民工的流动性，我们仅取有固定工作且在工作地居住满一年的农民工入框。其次，根据中国现行的行政区划，对各省市以地区为标志进行复合分层，实现层内指标差异减小，层间差异增大，提高调查精度，使目标估计量的抽样误差尽可能小。最后，在各省市所划分的层内实行两阶段不等概率随机抽样。

层内两阶段不等概率抽样的过程稍显复杂，其具体抽取方法为：

第一阶段是在层内按多项抽样抽取县级单位或街道居委会。我们采用汉森-赫维茨(Hansen-Hurwitz)随机方法进行抽取。以广东省为例，其具体实施方法如下：首先，我们赋予广东省每个地区级单位 M_i 个代码，即地区内的每个县级单位都拥有一个代码。例如东莞地区拥有代 $1\sim M_1$，则依次潮州地区拥有代码 $M_1+1\sim M_1+m_2$，第 i 个地区拥有代码 $\sum_{j=1}^{i-1}M_j+1\sim\sum_{j=1}^{i-1}M_j\cdots\cdots$，最后一个地区拥有代码 $\sum_{j=1}^{N-1}M_j+1\sim M_0(=\sum_{j=1}^{N-1}M_j)$。其次，利用计算机程序产生$[1,M_0]$之间的离散均匀分布随机数，与所产生的随机数代码相应的县级单位就作为抽中的样本单位，若一个县级单位被抽到两次或两次以上，则仍作为一个采样单位处理，继续抽取下一个县级单位，直到所需的县级单位样本数满足为止。最终，在入样概率 $Z_i=\frac{M_i}{M_0}$ 下，每层抽取的县级单位个数为 $n_i=\frac{M_i}{M_0}\cdot n$，其中，$n_i$ 为第 i 层应抽取的县级单位个数，n 为样本中应抽取的全部县级单位个数，M_i 为第 i 层所拥有的全部县级单位个数，M 为广东省所拥有的全部县级单位个数。

第二阶段是在每个被抽到的县级单位中抽取 m 个农民工进行实地调查。所用方法是将把各县级单位抽到的农民工，再进行简单随机抽样。其具体实施方法如下：利用均匀分布在(0,1)之间的永久随机数进行简单随机抽样，与所产生的永久随机数一致的农民工作为采样样本，若同一个永久随机数被抽到两次或两次以上，则视为一个随机数，继续抽取下一个，直到所需的农民工人数满足为止，这一过程也可以看成具有某一特征随机数的单位入选。最终，第 i 层抽取的农民工人数为 $m_i=n_im_{ij}$。若广东省市共有 N 个县级单位和街道居委会，则广东省抽取到的全部农民工人数为 $m=\sum_{i=1}^{N}m_i$。

2 总体总和、均值和比率估计

在中国农民工抽样调查当中，所涉及的调查指标不止一个，可以多达数百个。我们对样本资料需要从不同的角度进行综合对比和估计，所以，本文分别就调查指标的总和、均值和比率分别给出以下的估计结果。

2.1 估计子总体总和 Y 和方差

基金项目：国家统计局全国统计科学研究计划重点项目(2007LZ011)

对于二阶段抽样中子总体总和的估计，一般是先对每个被抽中的初级单元 i，利用第二阶段抽到的样本，估计初级单元的总和 Y_i，然后再利用单阶段的结果进一步估计 Y。

利用汉森-赫维估计量给出 Y 的无偏估计量：

$$\hat{Y}_{HH}=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{z_i}$$

其中，z_i 是第 i 个县级单位相应的入样概率 Z_i。

其方差的无偏估计量为：

$$v(\hat{Y}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{z_i}-\hat{Y}_{HH}\right)^2$$

因为在本方案中，第二阶段是简单随机抽样，所以 $\hat{Y}_i=M_i\bar{y}_i$ 是 Y_i 的无偏估计量，于是有

$$V(\hat{Y}_{HH})=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{z_i}$$

$$\hat{Y}_{HH}=\frac{1}{n}\sum_{i=1}^{N}z_i\left(\frac{Y_i}{z_i}-Y\right)^2+\frac{1}{n}\sum_{i=1}^{N}\frac{M_i^2(1-f_{zi})s_{2i}^2}{m_iz_i}$$

其中：$f_{2i}=\frac{m_i}{M_i}$

$$v(\hat{Y}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^{N}\left(\frac{M_i\bar{y}_i}{z_i}-\hat{Y}_{HH}\right)^2$$

2.2 估计子总体均值和方差

子总体的均值估计为 $\hat{\bar{Y}}=\frac{M_i}{M_0}\frac{\bar{y}_i}{z_i}$，其中 $\bar{y}_i=\frac{y_i}{m_i}$ 表示第 i 个县级单位中样本均值。

均值方差为：

$$\begin{aligned}V(\hat{\bar{Y}})&=V_1[E_2(\bar{y}_i)]+E_1[v_2(\bar{y}_i)]\\&=V_1(\bar{Y}_i)+E_1\left[\frac{(M_i-m_i)S_{2i}^2}{M_im_i}\right]\\&=\sum_{i=1}^{N}\frac{M_i}{M_0}(\bar{Y}_i-\bar{\bar{Y}}_i)^2+\sum_{i=1}^{N}\frac{(M_i-m_i)S_{2i}^2}{M_0m_i}\\&=\frac{1}{M_0}\left[\sum_{i=1}^{N}M_i(\bar{Y}_i-\bar{\bar{Y}}_i)^2+\sum_{i=1}^{N}\frac{(M_i-m_i)S_{2i}^2}{m_i}\right]\end{aligned}$$

其中 $S_{2i}^2=\frac{1}{M_i-1}\sum_{j=1}^{M_i}(Y_i-\bar{Y}_i)^2$ 表示子总体第 i 个县级单位内农民工之间的方差

对于不等概率的两阶段抽样，可以采用 $v(\hat{\bar{Y}})=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{Y}_i-\bar{\bar{Y}}_i)^2$ 作为均值方差的估计量。

2.3 计算子总体比率估计

设某地区有 N 个县级单位，第 i 县拥有农民工 M_i 个人，此县其中第 j 个农民工的月收入为 Y_{ij}，其工作时间为 X_{ij}，则单位时间内平均收入为 $R=\sum_{i=1}^{N}\sum_{j=1}^{M_i}Y_{ij}/\sum_{i=1}^{N}\sum_{j=1}^{M_i}X_{ij}$，此时 Y_{ij}，X_{ij} 都需要调查，其单位时间内平均收入 R 有总体比率的形式，这就要求对总体比率进行估计。以下我们采用联合比估计量。

子总体比估计 R 的无偏估计量为：$\hat{R}=\hat{Y}/\hat{X}$

总体比率的方差为：

$$V(\hat{R})=\frac{M_0}{nX^2}\left[\sum_{i=1}^{N}M_i(\bar{Y}_i-\bar{\bar{Y}})^2+\sum_{i=1}^{N}\frac{M_i(1-f_{2i})}{m_i}S_{2i}^2\right]$$

对于不等概率的两阶段抽样，可以采用 $v(\hat{R})=\frac{1}{n(n-1)\bar{x}^2}\sum_{i=1}^{n}(\bar{y}_i-\bar{\bar{y}})^2$ 作为 $V(\hat{R})$ 的估计量。

2.4 对总体进行估计

以上我们分别给出了每个子总体的总和、均值和比率估计，总体估计仅仅是对各子总体的简单加权平均。可用公式 $Q=\sum_{i=1}^{5}W_iQ_i$ 计算。其中 Q 表示对总体指标的估计；Q_i 表示对子总体的指标估计；W_i 表示各省市所占权重

3 多目标分层复合抽样的设计效应

中国农民工调查中的总体很大，抽样比相对较小，因此总体方差可以近似看作两阶段方差之和[4]，即 $S^2\approx S_1^2+S_2^2$。所以在二阶段抽样中，第一阶段抽取 n 个县级单位，第二阶段在被抽中的县级单位中抽取 m 个农民工，样本量为 nm。

在相同样本量下，简单随机抽样的样本均值方差为：

$$V_{srs}(\bar{\bar{Y}})=\frac{S^2}{nm}\approx\frac{S^2+S_2^2}{nm}$$

在相同样本量下，分层复合抽样的样本均值方差为：

$$V(\bar{\bar{Y}})=\frac{S_1^2}{n}+\frac{S_2^2}{nm}=\frac{mS_1^2+S_2^2}{nm}$$

所以，可以得到多目标分层复合抽样的设计效应：

$$deff=\frac{V(\bar{\bar{Y}})}{V_{srs}(\bar{\bar{Y}})}=\frac{mS_1^2+S_2^2}{S_1^2+S_2^2}$$

通常情况下，M>1，此时设计效应大于 1，说明多目标分层复合抽样的效率要低于简单随机抽样。但是由于多目标分层复合抽样有着样本集中，可省时、省力和省费用的优点，因此此抽样方法的效率是远远高于简单随机抽样的。

4 抽样调查中的相关问题

4.1 样本容量的确定问题

样本容量既要兼顾各个目标，又要尽可能小[5]。在大型的抽样设计中，兼顾多目标容易产生样本容量过大的现象，造成不必要的浪费。尽量使每个目标都在给定条件下确定最小样本容量，使样本容量比较经济。

4.2 样本轮换问题

多目标的分层复合抽样设计由于使用了永久随机数，调查单元与永久随机数有唯一确定性，新的调查单元会与其永久随机数一起不断补充到抽样框中，而消亡的调查单元也会与其永久随机数一起从抽样框中消失。随着抽样框的变动，调查单元的入样概率会发生变化，此时需要重新计算入样概率。

4.3 辅助变量问题

样本容量的确定和对总体进行分层都涉及到辅助变量。在农民工调查中可以以农民工的工作地域、工作行业为、户籍等为辅助变量。不同的辅助变量，就有不同的抽样框，相应的抽样方式也不同。因此，辅助变量的选择要顾及尽可能多的目标要求，要相对固定，才有利于样本的轮换。

参考文献：

[1]L.基什.抽样调查[M].北京：中国统计出版社，1997.

[2]冯士雍，倪加勋，邹国华.抽样调查理论与方法[M].北京:中国统计出版社，2002.

[3]俞纯权.二阶抽样总体比率的估计[J].应用概率统计，2007,23(8).

[4]杜子芳.抽样技术及其应用[M].北京:清华大学出版社，2005.

[5]罗纳德.扎加，约翰尼著.布莱尔.抽样调查设计导论[M].重庆:重庆大学出版社，2007.

（责任编辑/亦 民）