关于人口变动情况抽样调查方案的再探讨

胡 英　贾同金

《人口与经济》1991年第1期刊登了周祖根、梁小筠同志关于“人口变动情况抽样调查方案探讨”（以下简称“方案探讨”）一文，针对1989年人口变动情况抽样方案的设计，提出了几点分析意见。本文就全国人口变动情况抽样方案和“方案探讨”一文提出的问题进行再探讨。

一、人口变动情况抽样调查方案的设计思想

为了准确、及时地掌握全国人口变动情况，自1982年以来，国家统计局每年都进行一次人口变动情况抽样调查，以为国家制定国民经济计划和社会发展规划提供依据。从1982年到1988年，每年人口变动情况调查的样本量是50万人，大约占全国总人口的0.45‰，即抽样比为0.45‰，调查的样本量仅对全国人口变动情况有代表性。为了解决抽样调查对各省人口变动指标的代表性，从1989年起，全国调查的样本量由50万扩大到180万，抽样比为1.6‰左右。

在扩大调查样本量后，抽样方案的设计思想是以全国为总体，以省为次总体，使调查的结果不仅对全国人口变动情况有代表性，而且对各省也有较好的代表性。全国多数省级单位是采用分层、三级整群等概率的抽样方法。具体抽法是省抽县、市、区（以下简称县），县抽乡、镇、街道（以下简称乡），乡抽村民或居民小组（以下简称村民小组）。同时要求对抽样框内第一级县级单位进行分层。分层的原则是尽可能使层内各单位之间人口变动指标的差异减少，层与层单位之间的差异增大，以便降低抽样误差。

省级单位分层后，各层抽取的样本量是按比例分配的，各省抽样比f等于省内各层抽样比f_h。同时确定三级的抽样比f_1、f_2、f_3为常数。根据几年来的调查经验，一般要求第一级抽样比f_1为25％左右，第二级抽样比f_2为10％，第三级抽样比由公式$f_3 = f/(f_1 \times f_2)$计算，总抽样比$f = f_1 \times f_2 \times f_3 = 1.6‰$。

当各级抽样比f_1，f_2，f_3为常数时，虽然每个区域以同等的概率被抽中，但最终调查的样本量是一个随机变量。为了解决这个问题，我们在各地区第一级单位分层时，不仅以经济、地理条件为分层指标，还考虑每个县级单位的人口规模，并在编制每级抽样框时，按各级单位人口规模降序排列，按排列的序号，随机等距地抽取f_1，f_2，f_3比例的单位。由于调查样本规模之间的差异能够相互抵消，调查的样本量与原设计的样本量基本一致。1989年全国设计的样本量是180万，调查的样本量为186万，达到了设计的要求。

二、省级样本量的确定

根据抽样原理，总体调查样本量的多少主要是取决于总体内各整群之间变异系数，而不是总体规模。在确定样本量的同时，还要考虑调查费用以及费用效率。1989年人口变动调查的186万样本分布在30个省、自治区、直辖市的916个县级单位中。除西藏调查5000人外，一般省级单位调查5—7万人，最多的为10万人。

样本量设计的主要参数是人口出生率(CBR)，允许误差（Δ）、把握程度（t）和设计效率（deff）。由于全国30个省地区间人口变动指标差别大，我们要求调查人口出生率的相对误差控制在10%以下，允许误差控制在1‰～1.5‰的范围内，抽样调查的把握程度为95%（t=2），设计效率估计为1.4。计算样本量的公式为

$$n=\frac{t^2\times CBR\times(1-CBR)}{\Delta^2}\times deff \qquad (1)$$

$$\text{设计效率（deff）}=\frac{\text{复杂抽样方法所得样本的出生率估计量方差}}{\text{相同样本量的简单随机抽样方法样本出生率的估计量方差}}$$

$$=\frac{\hat{V}\ (CBR)\ \text{复杂}}{\hat{V}\ (CBR)\ \text{简单}}$$

由于各省的抽样方差不同，deff的值也就不同。全国设计效率1.4是根据几个省的数据计算后估计的，仅作为各省样本设计的参考值。各省可利用已掌握人口变动调查数据计算deff值。

从公式（1）看出，在允许误差（Δ）一定的情况下，总体方差$\sigma=CBR\times(1-CBR)$愈大，所需样本量愈多。CBR=Y/X是两个估计量之比，均需从调查中获得。

“方案探讨”一文中提出总体方差估计式

$$V_{RAX}(R)=\frac{N-n}{N\cdot n\bar{X}^2}\cdot\frac{\sum_{i=1}^{n}(Y_i-RX_i)^2}{n-1}$$

能够对总体方差给予估计，但该式需要各省前一年调查各村民小组的数据。为了便于计算，通常可以采用调查指标成数的方差公式作为V_{RAX}的估计值：

$$\sigma=\frac{P\times(1-P)}{n}=\frac{CBR\times(1-CBR)}{n}$$

三、省级人口指标的估计方法

根据上述设计的抽样方案，省级人口指标的计算方法为：

$$\text{人口出生率估计值}\hat{R}=\frac{\hat{Y}}{\hat{X}} \qquad (2)$$

$\hat{Y}$表示调查所得的年出生人数，$\hat{X}$表示调查年初人数和年末人数的平均值。

$\hat{R}$的方差估计值$V(\hat{R})\approx[V(\hat{y})+\hat{R}^2\cdot V(\hat{x})-2\cdot\hat{R}\cdot C(\hat{x},\hat{y})]/\hat{X}^2$ （3）

其中：

$$V(\hat{y})=\sum_h a_h S_{yh}^2$$

$$V(\hat{x})=\sum_h a_h S_{xh}^2$$

$$C(\hat{x},\hat{y})=\sum_h a_h S_{xyh}$$

$$S_{yh}^2=\sum_\alpha[y_{h\alpha}-(y_h/a_h)]^2/(a_h-1)$$

$$S_{xh}^2=\sum_\alpha[X_{h\alpha}-(X_{ah}/h)]^2/(a_h-1)$$

$$S_{xyh}=\sum_{\alpha}\left[X_{h\alpha}-(X_h/_{ah})\right]\left[Y_{h\alpha}-(Y_h/ah)\right]/(ah-1)$$

（3）式中C（$\hat{x}$，$\hat{y}$）是x和y的样本协方差。$Y_{h\alpha}$表示第h层内第一级抽样单位中第α个单位调查的出生人数，$X_{h\alpha}$是第h层内的第一级抽样单位第α个单位的样本规模。

“方案探讨”一文提出人口出生率估计值为下式：

$$\hat{R}=\frac{\sum_{h=1}^{L}\frac{Nh}{nh}\cdot\sum_{i=1}^{nh}\frac{M_i(h)}{m_i(h)}\cdot\sum_{j=1}^{m_i(h)}\frac{K_{ij}(h)}{k_{ij}(h)}\cdot\sum_{\mu=1}^{k_{ij}(h)}Y_{iju}(h)}{\sum_{h=1}^{L}\frac{Nh}{nh}\cdot\sum_{i=1}^{nh}\frac{M_i(h)}{m_i(h)}\cdot\sum_{j=1}^{m_i(h)}\frac{K_{ij}(h)}{k_{ij}(h)}\cdot\sum_{\mu=1}^{k_{ij}(h)}X_{iju}(h)} \qquad (4)$$

我们认为（4）式与（2）式是不矛盾的。

我们知道，由样本变量推算总体的变量时，一般表达式为：

$$\hat{Y}=\Sigma W_iY_i$$

Y_i是第i个抽中的最终样本单位值，W_i表示权数。如果抽样单位的权数不同，由样本推算总体的各个调查指标时，需要耗费大量的时间确定权数。因此，为了减少汇总阶段的工作量，需要设计一种能使样本单位都有相同权数的抽样方法，也称自行加权抽样方法。各省人口变动方案中采用相同的抽样比$f=f_1\times f_2\times f_3$是给每个抽中的单位相同的权数。因此（4）式可写为以下形式：

$$R=\frac{\sum_{h=1}^{L}\frac{1}{f_1}\sum_{i=1}^{nh}\frac{1}{f_2}\sum_{j=1}^{m_i(h)}\frac{1}{f_3}\sum_{\mu_i=1}^{k_{ij}(h)}Y_{iju}(h)}{\sum_{h=1}^{L}\frac{1}{f_1}\sum_{i=1}^{nh}\frac{1}{f_2}\sum_{j=1}^{m_i(h)}\frac{1}{f_3}\sum_{\mu_i=1}^{k_{ij}(h)}X_{iju}(h)}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{nh}\sum_{j=1}^{m_i(h)}\sum_{\mu=1}^{k_{ij}(h)}Y_{iju}(h)}{\sum_{h=1}^{L}\sum_{i=1}^{nh}\sum_{j=1}^{m_i(h)}\sum_{\mu=1}^{k_{ij}(h)}X_{iju}(h)}=\frac{\hat{Y}}{\hat{X}}$$

可见，在样本单位等权的条件下，（4）式与（2）式是相同的。

“方案探讨”一文中方差估计公式也是在考虑不等权的条件下提出的。在多级分层整群抽样时，我们通常用（3）式估计抽样方差。在使用（2）式时，要求调查的整群规模（$X_{h\alpha}$）的变异系数$C(x)=\frac{S(x)}{\bar{X}}$小于0.2，以便在使用样本统计量推断总体值时发生的偏差可忽略。由于1989年人口变动情况在抽取第一级样本单位时，考虑了按人口规模分层，基本上保证了群规模的变异系数小于0.2。

三、全国人口变动主要指标的推算

人口变动情况抽样调查的主要目的之一是估计全国年末人口数和该年度的人口出生率、死亡率和自然增长率。1989年全国调查的样本量是186万人。为了及时、准确地公布全国和分地区的人口数据，全国首先采用加权手工汇总的计算方法，利用调查的186万样本，估计全国人口自然变动指标。现以全国人口出生率估计方法为例。

$$\text{全国人口出生率}\widehat{CBR}=\sum_{h=1}^{30}W_h\hat{R}_h$$

$\widehat{CBR}$表示该年全国人口出生率的估计值，$\hat{R}_h$表示该年第h省人口出生率的估计值。$W_h=X_h/X$是该年度第h省平均人口（X_h）与全国年平均人口（X）的比例，由于这一比例在两个相近年度相对稳定，W_h可用1988年末第h省人口数占全国人口的比例来代替。

全国人口出生率的方差估计量 $V(\widehat{CBR}) = \sum_{h=1}^{30} W_h^2 V(\hat{R}_h)$，

调查数据把握程度为95%（t=1.96），允许误差（Δ）$= 1.96 \times \sqrt{V(\widehat{CBR})}$

相对误差（CV）$= \frac{\Delta}{\widehat{CBR}}$

总体人口出生率估计值的允许误差范围为$\widehat{CBR} \pm \Delta$

根据上式计算，1989年全国人口出生率允许误差为0.35%，相对误差控制在2%，达到了设计的要求。

为了进一步取得详细资料，要对调查数据进行机器汇总。机器汇总数据通常采用两种加权方法，一种是不等权数的汇总方法，按各省不同的抽样比汇总，这种方法能够利用调查的全部数据。一种是等权的汇总方法，按指定同一抽样比汇总，这种方法能够给数据使用者带来许多方便。1989年国家机器汇总的样本是从各省上报的调查样本中，根据随机抽样的原理，按相同比例抽取的117万样本。我们认为今后人口变动抽样调查数据汇总，采用哪种权数汇总的问题，还有待进一步的研究。

近几年来，全国人口变动抽样调查方案在实践中不断改进，获得了很大的成功。为了使人口变动抽样调查方法更具有科学性、可靠性和统一性，我们还需要对方法进行不断的探讨。特别是第四次全国人口普查资料给人口变动调查的样本设计、抽样方法和调查指标确定提供了可靠的依据，将会使我们今后每年度的人口变动情况抽样调查工作进一步提高和完善。

主要参考资料：

1.国家统计局1989年人口变动情况抽样调查办法，1989年8月。

2.《Intorduction to survey sampling》, GRAHAM KACTON, University of Michigan.

3.《抽样技术》，W.G.科克伦，中国统计出版社，1985年4月。

（作者工作单位：国家统计局人口司）

《婚姻经济学》一书即将出版

［本刊讯］武汉大学人口研究所青年学者谭仁杰经数年研究而写就的著作——《婚姻经济学》一书即将由河南人民出版社出版。该书由上、下两篇组成，17章，34万言。作者在书中依据对现实问题的考察，就婚姻经济问题作了理论上的探索和分析。学术界认为，由于此书是以经济为背景开展对婚姻问题的研究，故具有相当的理论和实际意义。

（舒涛）