# 人口变动情况抽样调查方案探讨

周祖根　梁小筠

为了准确、及时地掌握全国人口变动情况，为政府制定国民经济和社会发展计划等项工作提供可靠的人口资料，自1983年以来，全国每年都进行一次年度人口变动情况抽样调查。在当前严峻的人口形势下，为了对各地人口增长情况进行监督，根据国务院领导的要求，有必要每年公布分省人口数字和人口变动数字。为此，国家统计局1989年提出了以省为总体的分层、多级整群抽样调查方案。

1989年人口变动的抽样调查，在国家统计局领导下，全国30个省、自治区、直辖市动员了大量人力、物力和财力，从1990年1月1日起，累计调查了179万人，经过汇总和推算，已于1990年2月2日公布了1989年中国大陆人口出生率、死亡率和自然增长率，以及1989年末中国大陆总人口数。

从总体上讲，这次全国人口变动情况抽样调查是很成功的。不过，仔细推敲，仍感到1989年的以省为总体的人口变动情况抽样调查方案尚有可商榷之处。本文对该方案提出一些看法以供讨论，希望能使今后的年度人口变动抽样调查工作更为合理，所耗费的大量人力、物力、财力能产生更好的效益。

## 一　省级人口指标的计算方法

按照国家统计局方案①，全国多数省级单位采用分层、三级整群抽样方案。具体抽样方法是：省抽县、市、区（以下统称县），县抽乡、镇、街道（以下统称乡），乡抽村、居民小组（以下统称村民小组）。并要求对第一级抽样框内的县进行分层，分层的原则是尽可能使层内各单位之间人口变动指标差异减少，各层单位之间人口变动变异增加，以便降低抽样误差。在此抽样方案基础上，国家统计局提出了如下省级人口指标的计算方法。人口出生率为

$$R=\frac{Y}{X} \qquad (1)$$

其中Y为调查所得的年出生人数，X为调查所得的年初人口数与年末人口数的平均值。R的方差为

$$V(R)=[V(y)+R^2\cdot V(x)-2R\cdot C(x,y)]/X^2 \qquad (2)$$

其中$V(y)=\sum_{i=1}^{h}\frac{u_i}{u_i-1}\cdot\sum_{j=1}^{u_i}(y_{ij}-\bar{y}_i)^2$，

$V(x)=\sum_{i=1}^{h}\frac{u_i}{u_i-1}\cdot\sum_{j=1}^{u_i}(x_{ij}-\bar{x}_i)^2$，$C(x,y)$

$=\sum_{i=1}^{h}\frac{u_i}{u_i-1}\cdot\sum_{j=1}^{u_i}(y_{ij}-\bar{y})(x_{ij}-\bar{x}_i)$；

这里h表示层数，$u_i$表示第i层的整群个数，$X_{ij}$表示第i层中第j个整群的年均人数，$y_{ij}$表示第i层中第j个整群内的出生人数，$\bar{X}_i=\frac{1}{u_i}\cdot\sum_{i=1}^{u_i}x_{ij}$，

$\bar{y}_i=\frac{1}{u_i}\sum_{j=1}^{u_i}y_{ij}$

由此可以看到，这里的出生率的方差估计公式（2），只能说是在一定程度上体现了分层随机抽样的比率估计的计算，并没有反映出实际上各省实行的分层、多级整群抽样的特点。人口出生率的点估计公式（1）也只是在要求保证各阶段抽样比一致的情况下才能使用。但如果要保证这一条件的实现，就有可能产生大量小数现象。比如第三级抽样比$f_3=4\%$，某乡有40个村民小组，则按比例应抽取1.6个村民小组，这在实际工作中是不现实的。

因此，我们认为，立足于分层、多级整群抽样的实际状况，计算人口出生率采用下式更为适宜：

$$R=\frac{Y}{X} \qquad (3)$$

① 参阅国家统计局《1989年人口变动情况抽样调查办法》，1989年8月。

其中$Y=\sum_{h=1}^{L}\frac{N_h}{n_h}\cdot\sum_{i=1}^{n_h}\frac{M_i(h)}{m_i(h)}\cdot\sum_{j=1}^{m_i(h)}\frac{K_{ij}(h)}{k_{ij}(h)}\cdot\sum_{u=1}^{k_{ij}(h)}y_{iju}(h)$，$y_{iju}(h)$为抽中的第h层第i个县第j个乡的第u个村民小组的出生人数，故Y为全省的出生人数估计值，$X=\sum_{h=1}^{L}\frac{N_h}{n_h}\cdot\sum_{i=1}^{n_h}\frac{M_i(h)}{m_i(h)}\cdot\sum_{j=1}^{m_i(h)}\frac{K_{ij}(h)}{k_{ij}(h)}\cdot\sum_{u=1}^{k_{ij}(h)}X_{iju}(h)$，$X_{iju}(h)$为抽中的第h层第i个县第j个乡的第u个村民小组的年均人数，故X为全省的年均人数估计值。这儿的$N_h$为全省第h层的县的个数，全省县的总数$N=\sum_{h=1}^{L}N_h$，$n_h$为第h层的抽中县的个数，全省抽中县的总数$n=\sum_{h=1}^{L}n_h$，$M_i(h)$为第h层抽中的第i个县的乡的个数，$m_i(h)$为第h层抽中的第i个县的抽中乡的个数，$K_{ij}(h)$为第h层抽中的第i个县的第j个乡的村民小组数，$k_{ij}(h)$为第h层抽中的第i个县的第j个乡的抽中村民小组数。这些数据在抽样过程中是必然会得到的。R的方差为

$$V(R)=[(R+1)\cdot V(Y)+R\cdot(R+1)\cdot V(X)-R\cdot V(x+y)]/X^2 \qquad (4)$$

其中V(y)、V(x)及V(x+y)的计算只要分别以$y_{iju}(h)$、$x_{iju}(h)$及$x_{iju}(h)+y_{iju}(h)$代替下列公式(5)中的$y_{iju}(h)$即可。

根据类似的推导过程，也可得到估算省级年末人数的方法。如果$y_{iju}(h)$表示第h层抽中的第i个县第j个乡第u个村民小组的1989年年末人数，那么公式(3)的分子Y就是该省1989年年末人数的估计，而它的方差估计量为各层年末人数的方差估计量之和，即

$$V(Y)=\sum_{h=1}^{L}V(y_h) \qquad (5)$$

各层的$V(y_h)$为(从表达方便考虑，下面公式已略去层标记h)

$$V(y_h)=\frac{N(N-n)}{n}S_1^2+\frac{N}{n}\sum_{i=1}^{n}\frac{M_i(M_i-m_i)}{m_i}S_{2i}^2+\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{K_{ij}(K_{ij}-k_{ij})}{k_{ij}}S_{8ij}^2$$

其中$S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{K_{ij}}{k_{ij}}\sum_{u=1}^{k_{ij}}y_{iju}-\bar{y}_n\right)^2$，$\bar{y}_n=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}\frac{K_{ij}}{k_{ij}}\sum_{u=1}^{k_{ij}}y_{iju}$；

$S_{2i}^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(\frac{K_{ij}}{k_{ij}}\sum_{u=1}^{k_{ij}}y_{iju}-\bar{y}_{m_i}\right)^2$，$\bar{y}_{m_i}=\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{K_{ij}}{k_{ij}}\sum_{u=1}^{k_{ig}}y_{iju}$；

$S_{8ij}^2=\frac{1}{k_{ij}-1}\sum_{u=1}^{k_{ij}}\left(y_{iju}-\bar{y}_{k_{ij}}\right)^2$，$\bar{y}_{k_{ij}}=\frac{1}{k_{ij}}\sum_{u=1}^{k_{ij}}y_{iju}$

由此可见，省级年末人口总数的点估计和区间估计问题得到了解决。

北京、天津、上海、海南及宁夏采用分层、二级整群抽样方法，它们的人口出生率估计及年末人口总数估计完全与此类似，只是更简单而已。

这一具体计算过程若用计算器来计算的确比较麻烦。但现在微机已经相当普及，这一计算步骤的程序编制是很方便的。此外，各层的N、n、$M_i$、$m_i$、$K_{ij}$、$k_{ij}$在抽样过程中已经得到。因此，只要从人口变动抽样调查的手工汇总中得到各层的$x_{iju}$及$y_{iju}$等指标，就立即可得到该省人口的各主要指标的点估计及区间估计。

利用上海市1989年人口变动抽样调查资料，根据公式（3），计算得到上海市1989年人口出生率的点估计为12.79‰；根据公式（4），计算得到置信度为95%情况下的上海市1989年人口出生率的区间估计为（11.61‰，13.97‰）。而根据公式(1)，计算得到上海市1989年人口出生率的点估计则为12.11‰。

当然，如果能在抽样过程中，很好地保证对于所有h、所有i及所有j，$\frac{k_{ij(h)}}{K_{ij(h)}}$为常数；对所有h及所有i，$\frac{m_{i(h)}}{M_{i(h)}}$为常数；对所有h，$\frac{n_h}{N_h}$也为常数，那么公式（3）就可简化为公式（1）。

## 二 全国人口主要指标的推算

人口变动情况抽样调查主要目的之一是要推算全国年末人口数及该年度全国人口出生率、死亡率和自然增长率。根据国家统计局1989年抽样调查方案，全国调查样本是179万人，以解决省级人口指标推算的代表性问题。国家样本采取事后抽样办法，从各省上报的调查样本中按比例根据随机等距原则抽取部分样本，全国累计117万，使得各省事后抽样样本量在117万中所占比例与1988年末该省人口占全人口比例一致，并由此利用117万人的样本信息来推算全国人口主要指标。

这就必然带来两个问题。首先，推算全国人口主要指标仅利用117万个样本，其余62万个调查样本在全国人口主要指标的推算中未予利用，这不能不说是一个极大的损失。任何已经得到的样本都是来之不易的。其次，各省人口主要指标的推算立足于179万人的样本量，而全国人口主要指标的推算仅限于其中的117万人的样本量，这必然造成推算得到的各省人口指标与全国人口指标的差异。比如推算得到的全国人口总数与各省人口总数之和可能会有较大差距。

现在的问题是如何直接从179万人的样本量来推算全国人口主要指标。实际上，若令$y_h$表示该年第h省人口出生数估计，则全国人口出生数估计为

$$Y=\sum_{h=1}^{30} y_h$$

如果$x_h$表示该年第h省年均人数估计，x表示该年全国年均人数估计，于是从上式可得

$$\frac{Y}{X}=\sum_{h=1}^{30}\frac{x_h}{X}\cdot\frac{y_h}{x_h}$$

也即
$$R=\sum_{h=1}^{30} W_h\cdot R_h \qquad (6)$$

其中R表示该年全国人口出生率的估计值，$R_h$表示该年第h省人口出生率的估计值；$W_h=\frac{x_h}{X}$，是该年第h省年均人口占全国年均人口的比例。由于邻近年份这一比例是相对稳定的，因此在计算1989年度的人口出生率时，$W_h$可以用1988年末第h省人口数占全国人口数的比例来代替。又由于各省的样本是独立抽取的，从而全国人口出生率方差的估计量为

$$V(R)=\sum_{h=1}^{30} W_h^2\cdot V(R_h) \qquad (7)$$

式（6）与式（7）中的$R_h$及$V(R_h)$属省级人口指标，如前所述，他们的推算都已解决，因此，R与$V(R)$也就很方便的得到了。

同样道理，在已经解决各省年末总人口估计$Z_h$及其方差估计$V(Z_h)$的情况下，可以得到全国年末总人口估计

$$Z=\sum_{h=1}^{30} Z_h \qquad (8)$$

及其方差估计
$$V(Z)=\sum_{h=1}^{30} V(Z_h) \qquad (9)$$

根据式（6）、（7）及式（8）、（9）计算全国人口主要指标，就可充分利用179万人的样本信息，也保证了各省人口指标与全国人口指标相匹配。

## 三 省级样本量的确定

为了保证调查结果对人口出生率有较好的代表性，国家统计局1989年抽样调查方案提出了把握程度为95%情况下的省级样本量的计算公式：

$$n=\frac{t^2\cdot CBR\cdot(1-CBR)}{\Delta^2}\cdot deff \qquad (10)$$

其中t=2；CBR为人口出生率，用上年的人口出生率代替；△为人口出生率的允许误差，取值范围为1‰～1.5‰；deff为设计效率，规定为1.4。

由此得到的上海、浙江、辽宁、黑龙江、新疆等省、自治区、直辖市所需样本数都是5万人。首先，从直观上看，这似乎也不尽合理。合适的样本数应该同各省人口出生率估计量的方差有密切关系。上海作为一个大城市，管辖范围相当小，所属各地区的经济生活水平基本相同，计划生育政策及其执行程度大体一致，从近几年的资料来看，各地区的出生率、一孩率等都相当接近。因此，上海市各地区人

口出生状况的一致性很好，内部方差较小。但是，在辽宁、新疆等省、自治区，城市与农村的经济生活水平有相当大的差距。即使同样是农村，平原地区和山区差距也很大。此外，计划生育政策的贯彻程度也不尽相同，汉族地区和少数民族地区的计划生育政策本身也并非一致。因此，省内各地区的出生率差异较大。为了在同样的可信度下，得到出生率估计的相同精确度，那些省或自治区的样本数应该比上海大得多。

当然，式（10）用于计算省级样本量的不合理性主要是在于它的前半部分，它是在简单随机抽样背景下，当CBR是一种比例时，用来确定省级样本量的。比如CBR = Y/N，这里N是一个确定数，只有Y需给以估计。但现在CBR = Y/X是一种比率，是两个估计量之比，分子与分母均需从抽样调查中估计。因此，从理论上看，由式（10）出发来估计省级样本量是欠妥的。并且在实际使用中，不区别各省具体情况，一律取设计效率为1.4；必然得出样本数与各省人口出生率估计量的方差毫无关系的结论。这当然令人难以接受。

为了能得到较为合理的省级样本数，方法之一是提供更加妥当的省级样本量计算公式（这有待于我们进一步研究）；方法之二是，如果利用公式（10）作为一种近似计算，那么根据各省的具体情况，应使用不同的设计效率。按照科克伦在《抽样技术》中的论述，设计效率应为①

$$deff = \frac{\text{复杂抽样方案所得样本的出生率估计量方差}}{\text{相同样本量的简单随机样本的出生率估计量方差}}$$

式中的分子部分为公式（4）的计算结果，分母部分则为

$$V_{RAX}(R)=\frac{N-n}{Nn\cdot\bar{x}^2}\cdot\frac{\sum_{i=1}^{n}(y_i-R\bar{x}_i)^2}{n-1}$$

这里的N是全省的最后一级抽样单元数，即村民小组数，n是抽中的村民小组数，$y_i$是第i个村民小组在1989年的出生人数，$x_i$是第i个村民小组在1989年的年均人数，$\bar{x}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i$，R为该省样本按简单随机抽样原则下的人口出生率估计，即

$$R=\sum_{i=1}^{n}y_i\Big/\sum_{i=1}^{n}x_i \qquad (13)$$

对一个省来讲，近一二年的生育状况变化相对较小，从而可以充分利用1989年人口变动情况抽样调查得到的资料，来估计每个省人口出生率的设计效率。这对第二年的人口变动抽样调查中省级样本量的确定有重要作用。（本文责任编辑：洪　映）

（作者工作单位：周祖根　上海市人口普查办公室、梁小筠　华东师范大学数理统计系）

① 参阅W·G·科克伦：《抽样技术》4.11和6.3节，中国统计出版社，1985年4月。

（上接第64页）

所部分同志认为：沿海地区的人口素质是影响我国沿海地区外向型经济发展的关键因素。专题调查的资料表明，沿海地区的人口素质普遍高于内地，但结构性问题十分突出。技师和技工不足，普通工人比例过大。这种技术断层将降低沿海地区对国外资金、技术的吸收能力，也将拖延新技术的传播周期。此外，分配体制和用人体制恶化了人力资本的投资环境，造成对人才、技能的需求“疲软”。人不能尽其才，物必不能尽其用，扭曲的体制降低了资源配置效率，从而降低了沿海经济的发展速度。

会议期间，与会代表到大连对外经济开发区进行了考察，金州区大魏家乡后石村的阎克振同志介绍了该村人口与经济实现良性循环的经验。

最后由田雪原教授作了会议总结，他在总结发言中指出，第一，会议开的适时，沿海人口与经济方面存在的值得研究的问题正在暴露出来，太早太晚都难以处理。第二，会议开出了特色，突出了理论与实际相结合，绝大多数论文都在调研的基础上写成，参加会议的有理论工作者也有实际工作者。第三，会议交流了成果，也提出了一些新的研究课题。田雪原教授在总结中还概括了沿海人口与经济变化呈现的7个趋势，即，(1)伴随经济起飞，沿海人口数量继续增长的趋势；(2)人口自然增长的双向变动趋势；(3)人口向沿海迁移迅速增加的趋势；(4)城市人口加速增长的趋势；(5)生产年龄人口增长和就业结构转型发展趋势；(6)经济技术进步和人口文化素质不断提高的趋势；(7)人口加速老龄化趋势。

与会代表认为这次会议很有收获，并希望中国人口学会能组织力量，对有关问题进行更深入的探讨。（本文责任编辑：汪正鸣）

（作者工作单位：中国社会科学院人口研究所）