

中国1986年74城镇人口迁移抽样调查 目标量的估计方法与精度分析

高嘉陵 冯士雍

“中国1986年74城镇人口迁移抽样调查”(以下简称“迁移调查”)是由中国社会科学院人口研究所承担,得到联合国人口活动基金资助,被列为国家‘七五’期间哲学和社会科学重点研究项目,联合十六省(市)人口研究单位共同合作研究“中国城镇人口迁移与城镇化”课题的组成部分。此项“迁移调查”填补了我国城镇人口迁移资料的空白,提供了我国自1949年以来城镇人口迁移的流量、流向、结构、原因和后果的主要数据,它们不仅是人口学、经济学、社会学、地理学、生态学等学科所需的基本数据资料,也是国家决策部门制定改革政策的参考依据。

现已公布的计算机汇总数据资料^①是按城市规模汇总的实际样本数据。为进一步将这些宝贵的调查数据进行开发利用,我们针对此项调查的抽样设计以及实际需要,运用抽样调查的理论和方法提出了74城镇人口迁移有关目标量的估计方法以及对全国相应目标量的推总估计方法,并用随机分组方法对74城镇上述目标量估计的精度(方差)进行了估计和分析,同时还对全国指标的推算值做了评估。

一 抽样设计简介

“迁移调查”在1986年7月开始进行,同年年底先后完成。其中43个城市的调查范围是居住在城市地区的人口,即居住在城市市区、近郊区、工业区的人口,它包括了城市中绝大部分非农业人口和一部分农业人口;31个镇的调查范围是镇的总人口。以上城市的调查范围人口都已有明确的统计,我们以此做为“迁移调查”的目标总体。“迁移调查”确定的调查样本总量为25 000户。各省(市)遵照大城市多抽,小城市和镇少抽的原则,以及根据本单位的工作条件与经费情况确定本省(市)城镇的样本量,从而决定了各城镇的抽样比(样本量与调查范围人口的比)。

“迁移调查”的抽样方案采用四级整群抽样。第一级抽样是从全国抽省(市)。16个样本省(市)即16个人口研究单位所在省(市)是指定的,是根据研究单位的条件与可能自愿参加的。第二级抽样从上述样本省内抽城镇,并采用典型选取和随机抽取相结合的方法,在典型选取时按城市规模的大小,把城镇分为特大城市(100万以上人口)、大城市(50~100万人口)、中等城市(20~50万人口)、小城市(20万以下人口)和镇五类,选取中兼顾各种功能的城镇。第三级抽样是在城镇内抽取街道,抽取的方法是按比例分配分层,如城区、近郊区、工业区、商业区等层。对某些较小的城镇也有不分层情况。层内采用等概率或不等

^① 《中国1986年74城镇人口迁移抽样调查资料》,《中国人口科学》专刊Ⅱ,1988年。

概率按地址编码系统抽样或简单随机抽样抽取街道。最后一级抽样是在被抽中的街道内用等概率系统抽样抽取家庭户（集体户划分为四人一群相当一户，集体户与家庭户的抽取比例按人口比例分配）。其中每一个街道抽取的户数也按该街道的总户数比例分配。对抽中的户则进行整户调查，即调查户内所有成员。

二 数据处理的基本思想和目标量的确定

以上“迁移调查”的抽样方案，从整体上说不是一个严格的概率抽样，特别是在省（市）一级和城、镇一级均未按概率抽样方法抽取，因而无法用抽样调查的一般方法处理，如根据样本对目标量做推总估计和精度估计。然而，我们注意到，这16省（市）已超过大陆当时26省（市、自治区）的半数，且东北、华北、华东、西北、中南、西南各地区内至少有二个省（市）。假若我们取消省（市）一级，直接观察74城、镇，从城镇的数量和地域分布来看对全国还有一定的代表性；并且各省（市）抽取的城镇是按同一原则典型选取的，因此若对调查的74城镇进行合理的“事后分层”，则可以利用分层抽样的计算方法对全国性指标进行数据处理。

鉴于以上理由，我们将全国居住在城市地区的人口做为推论总体，而将调查的74城镇中城市地区的人口作为目标总体^①。

据此我们分二步进行数据处理。第一步对每个调查的城镇计算有关目标量的估计和相应的方差估计，然后汇总为目标总体相应的估计。第二步将推论总体和目标总体按同一原则分层，由74城镇目标总体主要目标量的估计分层加权得到全国城镇人口（推论总体）迁移指标的推算值。由于74城镇不是按概率抽样从全国抽取的，因此，对于推论值的偏差和精度则根据74城镇目标量估计与方差估计，做经验的定性分析。

在数据处理的第一步中，我们首先给出各城镇目标总体各类目标量的估计公式。对每个未知的目标量 θ ，用调查所得的样本数据对它进行估计，得到估计量 $\hat{\theta}$ 。由于 $\hat{\theta}$ 随样本而异，故有必要对它的精度加以讨论。描述一个估计量精度的准则之一是它的方差。方差表示估计量偏离其均值（对无偏估计量也就是目标量 θ 的真值）的大小的衡量，这种偏离在抽样调查中是不可避免的。如果我们用同一种抽样方法重复多次，即可得出方差的估计。当然在实际中重复抽样是不大可能的，因而根据样本数据作方差估计是十分重要的。在“迁移调查”中，由于采用的抽样方案比较复杂，且在各城镇中方法也不尽相同，因此在进行方差估计时，没有直接的公式可用，在本文中我们采用了随机分组法^②。

随机分组法亦称交叉子样本法，它的基本思路是将含有 N 个单元的样本（母样本）按一定方式划分为 b 个（ $b > 2$ ）子样本（随机组），先分别求得每个子样本以及母样本目标量的估计，用不同于样本估计量之间的差异估计总体目标量的方差。随机分组方法的基本要求是这些子样本（随机组）的构成一般要求与母样本的抽样方法相一致，也就是说子样本的抽样结构与母样本的结构基本相同。

为了达到上述目的，我们将每个城镇中每个样本街道中的所有调查户按一定方法（详见下节）划分为 b 组。特大城市一般分为15组（上海分50组），大城市、中等城市、小城市分10

① Kish, L. Survey Sampling, John Wiley Sons, 1965.

② Wolter K.M, Introduction to Variance Estimation, Springer Verlag 1985.

组, 镇分5组, 以保证每一个街道小组中有5至10户的样本量。城镇中所有样本街道的第一组组成城镇的第一个子样本, 所有的第二组组成第二个子样本, 以此类推, 这样将城镇母样本划分为b个子样本(随机组)。分别对母样本及b个子样本进行数据处理, 然后对每个城镇进行目标总体目标量的

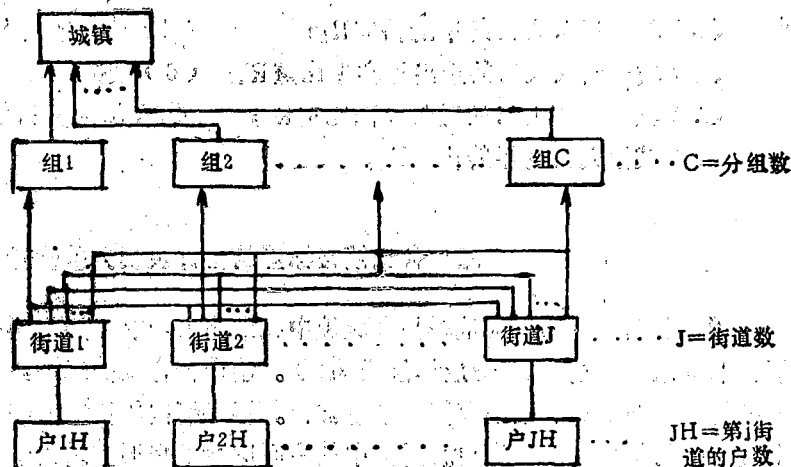


图1 城镇随机分组示意图(图中未表示城镇内街道分层的情况)

表1 全国74城、镇分层表①

	特大城市	大城市	中等城市	小城市	镇
沿海	石家庄 沈阳 上海 杭州 济南 广州	宁波②	承德 绍兴 烟台② 肇庆	泊头② 威海 珠海	辛集 峡石 宁海 德庆
内地	南昌 郑州 武汉 长沙 成都 西安	平顶山 株洲 渡口	九江 赣州 宜昌 随州② 宜宾 铜川	周口 老河口② 津市 汉中②	上栗 凤仪 樟树 永平(湖北) 新安(五镇) 社旗 浏阳
边远	包头 哈尔滨 贵阳	呼和浩特 大庆	银川 遵义	黑河 安顺 集宁	卓资山 新包力格 (黑龙江六镇) (贵州六镇)

- ① 辽宁、河北、山东、江苏、浙江、福建、广东、广西为沿海地区, 吉林、山西、陕西、河南、湖北、湖南、四川、安徽为内地, 黑龙江、宁夏、甘肃、新疆、青海、西藏、贵州、云南为边远地区
② 城市非农业人口小于总人口的50%, 城市规模大小降一级, 随州降二级

估计及其方差估计。城镇中随机分组示意图如图1。

数据处理的第二步, 首先将推论总体按目标总体的原则分层, 然后计算推论总体目标量估计值。

我们将74城镇按地理分布(沿海、内地、边远地区)及城市规模大小划分为15个层,(见表1)。

推论总体全国城镇的分层将在下面“全国及各种规模城市和镇人口迁移指标的推算”中介绍。

“迁移调查”的指标项目共有62个之多, 从数据处理方法角度上讲, 这些指标的目标量可分为二类。第一类是某个指标如 y 的总量 Y , 例如迁入人口总数, 迁入人口中男性总数等; 第二类是两个总量 Y 与 X 的比值 R , 例如迁入人口的性别构成, 即迁入人口男(女)性总数与迁入人口总数之比, 其中二者人口数都需要通过样本进行估计。其他如平均值或凡是在总人口 Z (调查时是已知的, 不需要估计)中所占的比例 $P=Y/Z$ 则可归为第一类处理。我们参照“迁移调查”研究报告①, 选择了以下指标为主要目标量, 它们包括了 $Y(P)$ 、 R 这二类目标量;

- (1) 城镇迁入人口占总人口的比例 P ; (2) 城镇迁入人口的性别构成 R_1 ;

① 《中国1986年74城镇人口迁移抽样调查资料》,《中国人口科学》专刊Ⅱ, 1988年

- (3) 城镇迁入人口的年龄构成 R_2 ; (4) 城镇迁入人口的文化构成 R_3 ;
 (5) 城镇迁入人口的迁出地类型比重 R_4 ; (6) 城镇迁入人口的迁出年代比重 R_5 ;
 (7) 城镇迁入人口的迁入原因比重 R_6 ; (8) 城镇人口的分性别年龄构成 P_1 ;
 (9) 城镇人口的年龄构成 P_2 。

三 各城镇目标量的估计及其方差估计

“迁移调查”第三级抽样是在城镇中抽取街道，有分层和不分层抽取二种情况。现按分层抽取介绍计算公式(不分层即层数为1)。层内抽样又分二级，第一级一般按等概率系统抽样抽街道，在上海采用不等概率系统抽样。第二级按等概率系统抽样，也即等距抽样，从每个被抽中的街道中抽家庭户。对抽中的户则进行整户调查。由于街道和户的排列顺序是按地址编码和户籍顺序排列的，与迁移情况无关，也即与调查的指标量不相关，故可看作是“随机顺序”。在此情形，系统抽样与简单随机抽样可看成是等价的，因此我们可将这样的系统抽样按简单随机抽样公式处理。

(一) 符号介绍。为介绍目标量的估计和方差估计方式，首先引进若干记号： h, i, j, k 分别为层、街道、户、人的编号； X, Y, \dots 为调查指标；

Z_{hij} 为 h 层第 i 街道第 j 户中的被调查人数； m_{hi} 为 h 层第 i 街道抽中的户数；

n_h 为 h 层内抽中的街道数；

M_i 为第 i 街道的总户数；

N_h 为 h 层内街道总数；

Z_{hi} 为 h 层第 i 街道的人数；

Z_h 为 h 层内总人数；

Z 为城镇人口总数；

(二) 目标量的估计。首先我们讨论第 h 层某个指标 y 的总量 Y_h 的估计。由于在每个街道内的抽样都是等概率的系统整群(户)抽样。正如前面所指出的那样，我们可用简单随机抽样的有关公式。对于 h 层内第 i 街道的指标 y 的平均数 Y_{hi} 可用以下简单估计：

$$\hat{Y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{Z_{hij}} Y_{hijk} \quad (1)$$

其中 Y_{hijk} 是第 h 层第 i 街道第 j 户第 k 人的指标。

因此 h 层内总量 Y_h 可按以下公式估计：

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \hat{Y}_{hi} \quad (2)$$

若层内抽样是按街道人口数成比例的不等概率系统抽样，则按照不放回不等概率抽样的 Thompson-Horvitz 估计：

$$Y_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi} Y_{hi}}{\Pi_i} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{Z_{hi}}{Z_h} M_{hi} \hat{Y}_{hi} \quad (3)$$

其中 Π_i 是第 i 街道被抽中的概率。

获得层内总量 Y_h 的估计后，比例 $P_h = Y_h / Z_h$ 的估计即可随之得到。

$$\hat{P}_h = \frac{\hat{Y}_h}{Z_h} \quad (4)$$

而两个量的比 $R_h = Y_h/X_h$ 的估计则由下式给出:

$$\hat{R}_h = \frac{\hat{Y}_h}{\hat{X}_h} \quad (5)$$

其中 X_h 是指标 X 的层总量的估计, 可以从 X_{hijk} 按公式 (2) 或 (3) 同样处理。

当在层内抽中街道的总户数以及在不同街道抽中的户数都是按户数比例分配时, 则样本是自加权的, 此时目标量的估计可简化为

$$\hat{Y}_h = \frac{1}{f} \sum_i \sum_j \sum_k Y_{hijk} \quad (6)$$

其中 f 为层内总抽样比。

$$\hat{P}_h = \frac{1}{f} \sum_i \sum_j \sum_k Y_{hijk} / Z_h \quad (7)$$

$$\hat{R}_h = \sum_i \sum_j \sum_k Y_{hijk} / X_{hijk} \quad (8)$$

即目标总量的估计等于样本指标总和除以抽样比, 比例型估计为样本总和与层内总人数之比, 比值型估计为二个指标的样本总和之比。

根据分层抽样公式, 城镇目标总量的估计为

$$\hat{Y} = \sum \hat{Y}_h \quad h=1, 2, \dots, L \quad (9)$$

其中 L 是城镇中划分的层数。

比例型目标量 P 及比值型目标量 R 分别可估计为

$$\hat{P} = \sum W_h \hat{P}_h \quad h=1, 2, \dots, L \quad (10)$$

$$\hat{R} = \sum W_h \hat{R}_h \quad h=1, 2, \dots, L \quad (11)$$

其中 $W_h = Z_h/Z$ 是层数。

(三) 估计量方差的估计。正如在第三节中所述, 我们采用随机分组法对估计量的精度 (用方差表示) 进行估计。我们首先介绍随机组的组成方法, 然后根据随机组给出方差的估计。若在城镇中不分层抽取街道, 则将每个被抽中的街道中的 (设为 m 个) 家庭户用系统抽样方法划分为 b 个随机组。在整数 1 至 b 中, 抽取一个随机整数 r , 将第一个样本户划为第 r 组, 第二个样本户为第 $r+1$ 组, 以此类推, 直到某一样本户为第 b 组, 以下的样本户顺序为第 1 组, 第 2 组... 第 r 组... 第 b 组, 再从第 1 组顺序排下去。如果街道的样本量 m 不是 b 的整倍数, 令 $m = bC + q$ (C 为整数), 则余下的 q ($q < b$) 个样本分别划为 r_1, r_2, \dots, r_q 组, r_1, \dots, r_q 为从 1 至 b 个整数中抽取的 q 个不放回的随机整数。第 a 个随机组则由所有 n 个样本街道的第 a 组的家庭户组成。

对于第 a 个随机组, 采用上述目标量估计公式计算某目标量 θ (Y, P 或 R) 的估计值 $\hat{\theta}_a$, 另外采用未分组的母样本按照上述公式求得的 θ 的估计量为 $\hat{\theta}$, 则 $\hat{\theta}$ 的方差的随机分组估计量为:

$$V(\hat{\theta}) = \frac{1}{b(b-1)} \sum_{a=1}^b (\hat{\theta}_a - \hat{\theta})^2 \quad (12)$$

若城镇采用分层抽取街道,则将L个层按以上方法每层分为b组,对城镇和每个随机组都进行总量估计有:

$$\hat{Y} = \sum_{h=1}^L Y_h \quad (13)$$

$$\hat{Y}_a = \sum_{h=1}^L \hat{Y}_{ha} \quad (14)$$

则目标量Y的方差估计为: $V(\hat{Y}) = \frac{1}{b(b-1)} \sum_{a=1}^b (\hat{Y}_a - \hat{Y})^2$ (15)

对比例型估计P,我们用下式估计 P_a 及 P :

$$\hat{P}_a = Y_a/Z, \quad \hat{P} = \hat{Y}/Z \quad (16)$$

于是 $V(P)$ 可用下式估计: $V(\hat{P}) = \frac{1}{b(b-1)} \sum_{a=1}^b (\hat{P}_a - \hat{P})^2$ (17)

比值型估计 \hat{R} ,我们仍用同样的估计量: $\hat{R} = \hat{Y}/\hat{X}$ (18)
这里 \hat{Y} 与 \hat{X} 根据(13)式计算。

为了估计 $V(\hat{R})$,我们利用泰勒级数,可以得到 $V(\hat{R})$ 的以下近似公式:

$$V(\hat{R}) = \hat{R}^2 \frac{V(\hat{Y})}{\hat{Y}} + \frac{V(\hat{X})}{\hat{X}} - \frac{2\text{cov}(\hat{Y}, \hat{X})}{\hat{Y}\hat{X}}$$

它的一个估计是:

$$V(\hat{R}) = \hat{R}^2 \left[\frac{V(\hat{Y})}{\hat{Y}} + \frac{V(\hat{X})}{\hat{X}} - \frac{V(\hat{U}) - V(\hat{Y}) - V(\hat{X})}{\hat{Y}\hat{X}} \right] \quad (19)$$

其中 $V(\hat{Y})$ 及 $V(\hat{X})$ 用(15)式计算,而 $V(\hat{U})$ 则对新指标U,即 $U_{ijk} = Y_{ijk} + X_{ijk}$ 用公式(15)式计算而得。

在求得34城镇目标量估计和方差估计之后,我们按照前述的分层方法,用分层抽样公式得到目标总体74城镇主要目标量的估计和方差估计。

四 全国及各种规模城市和镇人口迁移指标的推算

全国及各种规模城镇目标量的估计是通过计算了74城镇的目标量估计之后,对它们进行“事后分层”用分层抽样公式求得的。为此,我们需对全国城镇分层,并需已知各层中的城市和居住在这些城市地区的层人口数。将全国城镇分层的原则必须与74城镇分层原则一致。因为居住在城市地区的人口包括了城市中绝大部分的非农业人口,所以在对全国城市按规模分层时,我们使用公安部所编1986年度全国分县市人口统计资料^①中市非农业人口一览表。

^① 见《中华人民共和国全国分县市人口统计资料》1986年度中华人民共和国公安部编,地图出版社,1987年。

在确定某规模层的城市时，从大到小取到在本规模层中被调查城市是非农业人口最少的城市为止。比如全国的大城市，我们从福州市取到株洲市，全国的中等城市从双鸭山市取到肇庆市。也因为居住在城市地区的人口包括了城市中绝大多数的非农业人口，所以我们以43城市的调查范围人口与城市中非农业人口的比例为权数，来计算各层居住在城市地区的人口数。经分层加权计算得推论总体全国居住在城市地区的总人口为14.963千万人。镇的调查范围与它的总人口一致为20.37千万人。

据统计，1986年全国城市总人口为2亿3千万人，非农业人口为1亿2千万人，由“迁移调查”推算的全国居住在城市地区的人口为1亿5千万人。那么全国城市地区的农业与非农业人口的比例约为20:80，我国城市中农业与非农业人口比例的变化与城市建制原则和城乡划分标准的变动有关。“迁移调查”的调查范围是在城市地区的实际人口，在一定时期内其农业与非农业人口的比例是相对稳定的，因此由“迁移调查”所推算的全国城市地区一亿五千万人的人口数在一定程度上反映了中国城市化的实际水平。（考虑到一些特大城市没有调查郊区，城市化真实水平农业人口的比例要略高些）

全国各层目标量的估计由层内抽中城镇的目标量估计值按居住在城镇地区人口数加权求得：

$$\hat{\theta} = \sum_{d=1}^c W_d \hat{\theta}_d = \sum_{d=1}^c \frac{Z_d}{Z} \hat{\theta}_d \quad (20)$$

其中：C为层内抽中城市数；

Z_d 为层内抽中第d个城市居住在城市地区的人口。

Z为层内居住在城市地区的总人口；

全国各地区及各种规模城镇主要人口迁移指标的推算由各层目标量估计按分层抽样公式求得（公式略）各分层的权数列表如下（见表2、表3、表4）。

表2 全国分层城镇权数 (%)

地 区	全国市	特	大	中	小	镇
全国市	100	47.11	12.23	29.42	11.24	100
沿 海	48.49	27.57	7.05	10.41	3.46	51.38
内 地	34.35	12.70	2.21	15.09	4.34	35.34
边 远	17.17	6.85	2.96	3.92	3.44	13.28

表3 地区分层城市权数表 (%)

地区	合计	特	大	中	小
沿海	100	56.85	14.54	21.47	7.14
内地	100	36.98	6.43	43.94	12.56
边远	100	39.90	17.27	22.82	20.01

表4 城市类型分层城市权数表 (%)

地区	特	大	中	小
合 计	100	100	100	100
沿 海	58.50	57.68	35.39	30.80
内 地	26.97	18.07	51.30	38.64
边 远	14.54	24.25	13.31	30.56

五 主要目标量估计结果和精度分析

我们用以上方法计算出了做为目标总体的43个城市和各种规模城市及所有镇的主要目标量的估计及其方差估计。

每个目标量的估计值为 $\hat{\theta}$ ，估计值的标准差估计为 $S(\hat{\theta})$ ： $S(\hat{\theta}) = \sqrt{V(\hat{\theta})}$ (21)

变异系数的估计值为 $CV(\hat{\theta})$ ： $CV(\hat{\theta}) = S(\hat{\theta})/\hat{\theta}$ (22)

还可以计算某一目标量的置信区间和估计量的绝对误差和相对误差。例如：43城市迁入人口占居住在城市地区总人口的比例中的95%的置信区间为

$$\hat{P} \pm 1.96 \times S(\hat{P}) = 0.3082 \pm 1.96 \times 0.0021 \quad (23)$$

即(30.41%，31.23%)。上式中的1.96是标准正态分布的双侧分位数。

估计值 $\hat{P} = 30.82\%$ 在95%的置信水平下的绝对误差为

$$d(\hat{p}) = 1.96 \times S(\hat{p}) = 0.41\% \quad (24)$$

相对误差为 $r(\hat{p}) = 1.96 \times CV(\hat{p}) = 1.33\%$ (25)

对于不同指标的项目，调查的精度有所差别，反映在估计量的变异系数上。这是由于不同项目的有效样本有较大的差异造成的。例如迁入人口性别构成的精度低于总人口性别构成的精度。

对于不同规模城市的目标量估计，其精度也有所差别。结果表明，大城市的目标量估计的精度最差。这是因为大城市所抽中的城市数量少，且沿海地区这一层只有宁波一个城市所致。中小城市的目标量估计的精度明显低于特大城市，也是由于样本量较少的缘故。

从这些目标量的估计计算结果表明，“迁移调查”目标总体中绝大多数指标项目的估计量的变异系统在10%以下，其精度是可以满足要求的。

同时，我们又计算了作为推论总体的全国城镇的估计值，结果表明，推论总体与目标总体的目标量估计值是不同的，有的还有明显的差异，比如大城市的迁入人口比例，在目标总体中为46%，在推论总体中为34%，这是因为所抽中的各类城市的人口在不同总体中所占的比重不同，我们说74城镇对全国还有代表性，但并不能简单地使用74城镇目标量的估计值说明全国性的问题，必须根据这74城镇在全国各种规模城市和地区中（或其他分层方法）所占的比重加权，才能得到偏差较小的全国估计值。

全国推论总体的估计值与实际值之间是有偏差的，这种偏差体现在74城镇分层分布中，“迁移调查”抽中的城镇偏重于沿海和内地，西北边远地区抽中城镇样本较少。在大城市中，抽中的城市大多为新兴工业城市，自然迁入人口的比例较大，这种偏差对目标量的影响，在目标总体中非常明显，而在推论总体中经分层加权后便有所改善。

对于全国城镇目标量估计的精度，我们不进行确切的计算，只通过74城镇目标量估计的精度进行定性分析。（本文责任编辑：郭汉英）

（作者单位：高嘉陵 中国社会科学院人口研究所

冯士雍 中国科学院系统科学研究所）