DOI:10.13546/j.cnki.tjyjc.2005.21.002

整群抽样总体均值的

在经济学研究中，数据分析是很重要的。可是如何获得数据，用哪种抽样方法得到数据是需要精心选择的。本文提出了运用整群抽样的方法，来减少数据分析时的误差。

整群抽样的样本是按一定方式，由某些群中所有的单元组成的。使用整群抽样的主要原因是为了节约成本，因为群内单元都比较集中，所以比分散的单位更便于调查。举例来说，人口调查中先取户，再调查户内人口，比对人直接用简单随机抽样成本小，因此整群抽样在社会经济抽样调查中被广泛应用。由于整群抽样的精度与群的性质有很大关系，我们通常的分群原则是：尽可能使群间方差小，群内方差大。但在实际问题中常基于某种方便的原则，按自然形成的单位定义群，如按行政区、部门行业等自然形成的群划分。这就导致了群内方差过小，群间方差过大。对于这种整群抽样，我们就要利用回归的方法来提高整群的估计精度。

本文就是在群内所含单元数目已知的条件下，利用抽样中的两种信息，一为主要特征信息，另一为辅助特征信息，建立它们之间的回归方程来进行估计。

设总体的主要统计特征为 Y，与之相关的辅助统计特征为 X，总体中含有 N 个群，样本中有 n 个群，总体第 i 个群中含有 M 个次级单元（i=1,2,…N），共有 $M_0=\sum_{i=1}^{N}M_i$ 个次级单元，样本中第 i 个群的次级单元数为 $m_i(i=1,2,\cdots,n)$，对群进行简单随机抽样。

一、平均群的不加权回归估计

根据定义，总体均值回归估计量可如下表述：$\bar{\bar{y}}_{lr}=\bar{\bar{y}}+b(\bar{\bar{X}}-\bar{\bar{x}})$ （）

其中，$\bar{\bar{X}}$已知，b 为参数，可以为事先设定的常数 b_0，也可以是某个特定的统计量 b_1。

一般取 $b_1=\sum_{i=1}^{n}(\bar{y}_i-\bar{\bar{y}})(\bar{x}_i-\bar{\bar{x}})/\sum_{i=1}^{n}(\bar{x}_i-\bar{\bar{x}})^2$

则 $E(\bar{\bar{y}}_{lr})=E(\bar{\bar{y}}+b(\bar{\bar{X}}-\bar{\bar{x}}))=\bar{\bar{Y}}+E(b(\bar{\bar{X}}-\bar{\bar{x}}))$，所以$\bar{\bar{y}}_{lr}$ 是有偏估计量。

若 $b=b_0$ 已知，则$\bar{\bar{y}}_{lr}$ 作为$\bar{\bar{Y}}$的估计量是无偏的。

令：$\bar{M}=M_0/N$，$M_0=\sum_{i=1}^{N}M_i$

$$\bar{\bar{y}}'=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\bar{M}}\sum_{j=1}^{m_i}y_{ij}=\frac{1}{n\bar{M}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}y_{ij}$$

$$\bar{\bar{x}}'=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\bar{M}}\sum_{j=1}^{m_i}x_{ij}=\frac{1}{n\bar{M}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}y_{ij}$$

$$\bar{y}'_i=\frac{1}{\bar{M}}\sum_{j=1}^{m_i}y_{ij},\quad \bar{x}'_i=\frac{1}{\bar{M}}\sum_{j=1}^{m_i}x_{ij}$$

$$\bar{y}'_i=\frac{1}{\bar{M}}\sum_{j=1}^{M_i}y_{ij},\quad \bar{X}'_i=\frac{1}{\bar{M}}\sum_{j=1}^{M_i}x_{ij}$$

$$\bar{\bar{y}}'=\frac{1}{n}\sum_{i=1}^{n}\bar{y}'_i,\quad \bar{\bar{x}}'=\frac{1}{n}\sum_{i=1}^{n}\bar{x}'_i$$

$$\bar{\bar{X}}=\frac{1}{N}\sum_{i=1}^{N}\bar{X}'_i,\quad \bar{\bar{Y}}=\frac{1}{N}\sum_{i=1}^{N}\bar{Y}'_i$$

$$\bar{\bar{y}}'_{lr}=\bar{\bar{y}}'+b'(\bar{\bar{X}}-\bar{\bar{x}}')$$

定理 1：$\bar{\bar{x}}'$,$\bar{\bar{y}}'$ 是$\bar{\bar{X}}$, $\bar{\bar{Y}}$的无偏估计量。

证：将$\bar{\bar{Y}}=\frac{1}{N}\sum_{i=1}^{N}\bar{Y}'_i$ 看作总体均值

$\bar{\bar{y}}'=\frac{1}{n}\sum_{i=1}^{n}\bar{y}'_i$ 看作样本均值

由于对群的抽取是简单随机的，所以$\bar{\bar{y}}'$ 是$\bar{\bar{Y}}$的无偏估计量。

同理可证$\bar{\bar{x}}'$ 是$\bar{\bar{X}}$的无偏估计量。

引理 1：$\bar{\bar{x}}'$,$\bar{\bar{y}}'$ 分别是抽自总体均值为$\bar{\bar{X}}$, $\bar{\bar{Y}}$ 的有限总体的简单随机样本的均值，样本容量为 n，则对非负整数 k,m 有：

$$E(\bar{\bar{y}}'-\bar{\bar{Y}})^k=O(n^{-[(k+1)/2]})$$

$$E(\bar{\bar{y}}'-\bar{\bar{Y}})^k(\bar{\bar{x}}'-\bar{\bar{X}})^m=O(n^{[-k+m+1)/2]})$$

引理 2：设 $B=\frac{\sum_{i=1}^{N}(\bar{Y}'_i-\bar{\bar{Y}}')(\bar{X}'_i-\bar{\bar{X}}')}{\sum_{i=1}^{N}(\bar{X}'_i-\bar{\bar{X}}')^2}$是有限总体回归系数，$\bar{\bar{x}}'$ 是简单随机样本$\bar{x}'_i$ 的均值，记 $\varepsilon_i=(\bar{Y}'_i-\bar{\bar{Y}})-B(\bar{X}'_i-\bar{\bar{X}})$

则 $\bar{\varepsilon}=0$

$$\sum_{i=1}^{N}\varepsilon_i(\bar{X}'_i-\bar{\bar{X}})=0$$

$$E[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\bar{x}'_i-\bar{\bar{x}})]^2=O(\frac{1}{n})$$

$$E[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\bar{x}'_i-\bar{\bar{x}})]^4=O(\frac{1}{n^2})$$

引理 3：在简单随机抽样下，记

$$b'=\frac{\sum_{i=1}^{n}(\bar{y}'_i-\bar{\bar{y}}')(\bar{x}'_i-\bar{\bar{x}}')}{\sum_{i=1}^{N}(\bar{x}'_i-\bar{\bar{x}}')^2}$$，则

$$b'=B+\frac{\sum_{i=1}^{n}\varepsilon_i(\bar{x}'_i-\bar{\bar{x}}')}{\sum_{i=1}^{n}(\bar{x}'_i-\bar{\bar{x}}')^2}$$

$$E(b'-B)=O(\frac{1}{\sqrt{n}})$$

$$E(b'-B)^2=O(\frac{1}{n})$$

$$E(b'-B)^4=O(\frac{1}{n^2})$$

定理 2：在简单随机抽样下，由以上引理可以得到，整群回归估计量$\bar{\bar{y}}'$ $\bar{\bar{X}}$有以下性质：

$$(1)B(\bar{\bar{y}}'_{lr})=E(\bar{\bar{y}}'_{lr}-\bar{\bar{Y}})=E(\bar{\bar{y}}'_{lr})-\bar{\bar{Y}}$$

$$=\frac{1-f}{(n-1)S_x'^2}\frac{\sum_{i=1}^{N}\varepsilon_i(\bar{X}'_i-\bar{\bar{X}})}{N-1}+O(\frac{1}{n^{3/2}})=O(\frac{1}{n})$$

$$(2)MSE(\bar{\bar{y}}'_{lr})=\frac{1-f}{n}S_x'^2(1-\rho'^2)+O(\frac{1}{n^{3/2}})=O(\frac{1}{n})$$

$$(3)E(\frac{1}{n-2}\sum_{i=1}^{n}[(\bar{y}'_i-\bar{\bar{y}}')-b'(\bar{x}'_i-\bar{\bar{x}}')]^2)$$

$$=S_x'^2(1-\rho'^2)+O(\frac{1}{\sqrt{n}})$$

其中，$S_Y'^2=\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}'_i-\bar{\bar{Y}})^2$，$S_x'^2=\frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}'_i-\bar{\bar{X}})^2$，$S_{YX}'^2=\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}'_i-\bar{\bar{Y}})(\bar{X}'_i-\bar{\bar{X}})$，

基金项目：国家自然科学基金项目（0271079）。

回归估计

■金 莹 牛美玲 汤银才

$$\rho'^2=\frac{S'^2_{YX}}{S'^2_X S'^2_Y}$$

由 ()知 $\bar{\bar{y}}'_{lr}=\bar{y}'_i+b'(\bar{\bar{X}}-\bar{\bar{x}}')$ 是 $\bar{\bar{Y}}$ 的有偏估计量。

推论 1：当 n 很大时，$V(\bar{\bar{y}}'_{lr})\approx MSE(\bar{\bar{y}}'_{lr})\approx\frac{1-f}{n}S'^2_Y(1-\rho'^2)$

证：$MSE(\bar{\bar{y}}'_{lr})=V(\bar{\bar{y}}'_{lr})+[B(\bar{\bar{y}}'_{lr})]^2=V(\bar{\bar{y}}'_{lr})+O(\frac{1}{n^2})$

$V(\bar{\bar{y}}'_{lr})=MSE(\bar{\bar{y}}'_{lr})+O(\frac{1}{n^2})=\frac{1-f}{n}S'^2_Y(1-\rho'^2)+O(\frac{1}{n^{3/2}})+O(\frac{1}{n^2})$

$=\frac{1-f}{n}S'^2_Y(1-\rho'^2)+O(\frac{1}{n^{3/2}})$

故 $V(\bar{\bar{y}}'_{lr})\approx MSE(\bar{\bar{y}}'_{lr})\approx\frac{1-f}{n}S'^2_Y(1-\rho'^2)$

推论 2：$\hat{V}(\bar{\bar{y}}'_{lr})=\frac{1-f}{n}\frac{1}{n-2}\sum_{i=1}^{n}[(\bar{y}'_i-\bar{\bar{y}}')-b'(\bar{x}'_i-\bar{\bar{x}}')]^2$ 是 $V(\bar{\bar{y}}'_{lr})$ 的近似无偏估计量。

证：$E(\frac{1-f}{n}\frac{1}{n-2}\sum_{i=1}^{n}[(\bar{y}'_i-\bar{\bar{y}}')-b'(\bar{x}'_i-\bar{\bar{x}}')]^2)$

$=\frac{1-f}{n}[S'^2_Y(1-\rho^2)+O(\frac{1}{\sqrt{n}})]\approx V(\bar{\bar{y}}'_{lr})$

因此 $\hat{V}(\bar{\bar{y}}'_{lr})=\frac{1-f}{n}\frac{1}{n-2}\sum_{i=1}^{n}[(\bar{y}'_i-\bar{\bar{y}}')-b'(\bar{x}'_i-\bar{\bar{x}}')]^2$

推论 3：$\bar{\bar{y}}'_{lr}$ 是可用的。

证：$S\frac{B(\bar{\bar{y}}'_{lr})}{\sqrt{MSE(\bar{\bar{y}}'_{lr})}}=\frac{O(\frac{1}{n})}{\sqrt{O(\frac{1}{n})}}=O(\frac{1}{\sqrt{n}})$

因此 $\bar{\bar{y}}'_{lr}$ 是可用的

二、平均群的加权回归估计

在群大小不等的情况下，群的样本量大小不同，在整个实验中所处的地位不平等，这样用这些数据作回归时，就不能把他们同等看待，用不加权的估计量精度可能会不高，必须有所侧重，这就需要用到加权回归来计算。

根据加权回归理论，取

$$b^*=\frac{\sum_{i=1}^{n}m_i\bar{y}'_i\bar{x}'_i-\frac{1}{M_n}\sum_{i=1}^{n}m_i\bar{y}'_i\sum_{i=1}^{n}m_i\bar{x}'_i}{\sum_{i=1}^{n}m_i\bar{x}'^2_i-\frac{1}{M_n}(\sum_{i=1}^{n}m_i\bar{x}'_i)^2}$$

$\bar{\bar{x}}^*=\frac{\sum_{i=1}^{n}m_i\bar{x}'_i}{M_n}$，$\bar{\bar{y}}^*=\frac{\sum_{i=1}^{n}m_i\bar{y}'_i}{M_n}$，

其中 $M_n=\sum_{i=1}^{n}m_i$

则 $\bar{\bar{Y}}$ 的加权回归估计量为

$\bar{\bar{y}}^*_{lr}=\bar{\bar{y}}^*+b^*(\bar{\bar{X}}-\bar{\bar{x}}^*)$，同样它也是有偏估计。方差的估计为

$\hat{V}(\bar{\bar{y}}^*_{lr})=\frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\frac{m_i}{M_n}[(\bar{y}_i-\bar{\bar{y}}^*)-b^*(\bar{x}'_i-\bar{\bar{x}}^*)]$ 群大小不等时，整群抽样下简单估计的估计量为 $\bar{\bar{y}}=\frac{\bar{y}}{\bar{M}}=\frac{\sum_{i=1}^{n}y_i/n}{\sum_{i=1}^{N}M_i/N}$，其抽样方差为

$V(\bar{\bar{y}})=\frac{1}{\bar{M}^2}\frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}'_i-\bar{\bar{Y}})^2=\frac{1-f}{n}S'^2_Y$

其中，$Y_i=\sum_{j=1}^{M_i}Y_{ij}$，$\bar{Y}=\sum_{i=1}^{N}Y_i/N$

与不加权的估计量方差相比

$$\frac{1-f}{n}S'^2_Y(1-\rho'^2)-\frac{1-f}{n}S'^2_Y=\frac{1-f}{n}S'^2_Y\cdot\rho'^2<0$$

所以，不加权的回归估计量方差小，其估计精度优于整群抽样的简单估计量。

例：以 1999 年全国工资与消费情况的抽样调查为例，比较群大小不等情况下四种估计量。

把全国 31 个省作为不同群体，从中简单随机抽取 8 个省的资料 见下表）。

已知 $N=31,n=8$，$\sum_{i=1}^{N}M_i=11278$，$\sum_{i=1}^{n}M_i=4429.6$，$\sum_{i=1}^{n}X_i=11473$，$\bar{\bar{X}}=1.0173$

全国 8 省职工工资、消费调查表

m_i	x_i	y_i
职工人数 万人）	工资总额 万元）	消费总额 万元）
403	566.3762	302.1887
379.4	230.1061	132.5237
352.6	252.3911	129.1108
327.1	544.3271	269.7819
408.9	266.4392	159.5450
320.4	304.0596	168.7447
305.9	206.4519	106.5245
335.4	232.4657	132.5920

()不加权估计

平均群回归估计系数 $b'=0.4889$

平均群回归估计量 $\bar{\bar{y}}'_{lr}=0.4212$ 万元）

平均群回归估计量方差 $v_1=7.8309\times10^{-5}$

(2)加权估计

平均群加权回归估计系数 $b^*=0.4902$

平均群加权回归估计量 $\bar{\bar{y}}^*_{lr}=0.5435$ 万元）

平均群加权回归估计量方差 $v_2=6.1718\times10^{-5}$

(3)简单估计

估计量 $\bar{\bar{y}}_{简单}=0.4814$ 万元）

估计量方差 $v_3=0.0036$

从这个例子我们可以看出，在群大小不等的情况下，本文提出的新的回归估计量及新的加权回归估计量比简单估计量的方差小，即精度高，估计效果好。因此，在群大小不等的情况下，我们可以用本文提出的平均群回归估计量及其加权回归估计量，估计效果会更好。

作者单位/上海师范大学数理信息学院，华东师范大学）

责任编辑/李友平）