人口抽样调查数据分析中的加权方法

亓 昕
(首都经济贸易大学 人口经济研究所, 北京 100026)

摘要: 本文介绍了人口抽样调查中经常遇到的三种情况: 不等概率抽样、无应答和抽取的样本与总体不符的问题。结合实际例子论述了如何利用已知信息对获得样本和变量加权的方法以解决这三种问题, 最终达到样本对总体的无偏估计。

关键词: 抽样调查; 加权方法
中图分类号: C924.25 文献标识码: A 文章编号: 1000-4149 (2003) 01-0040-04

Weighting Method Used in Sample Survey Analysis

QI Xin
(Institute of Population Economics, Capital University of Economics and Business, Beijing 100026)

Abstract: This paper introduced three common situations in population survey: unequal probability sample, no response and hotgenious between population and sample. Combined with concrete examples, this paper expatiate how to use knowable information to weight sample cases and variables, and then give an unbiased population estimation.

Keywords: sample survey; weighting method

一、问题的提出

抽样调查作为非全面调查已被广泛的应用。它是依照一定的程序从抽样总体中抽取部分元素 (即样本) 进行调查, 根据样本数据, 对总体目标量进行估计和统计推断。由于从样本估计总体存在抽样误差, 人们总希望这种误差越小越好, 估计越精确越好。但调查时也会经常遇到这样的问题: (1) 被抽元素的入选机会是不等概率的; (2) 有些问题无应答; (3) 获得的样本结构 (或分布) 与总体不相符合。忽略这些问题会使样本对总体的代表性受到影响, 样本特征不再反映总体特征, 样本估计不再是总体的无偏估计。

问题 (1) 属于不等概率抽样。不等概率抽取一般有两种情况, 一是因某种需要 (如等概率抽取时样本量偏少, 不能进行地域间的比较) 而采取这种方式, 这种加权方法较为简单 (这将在“无应答时的修正方法 (1)”中介绍)。二是在实际中由于无法直接面对较小的基本单元抽样, 产生了元素入选的不等概率性。例如, 在人口抽样中, 一般抽样设计是按市、街道、居委会, 最后落到户, 这样一级一级抽取。入户后才随机选取符合要求的 (如年龄 18 岁以上) 家庭成员一或两人 (根据样本

收稿日期: 2002-07-03
作者简介: 亓 昕 (1963-), 女, 黑龙江木兰县人, 首都经济贸易大学人口经济研究所副研究员, 主要研究方向: 抽样设计、人口统计分析。

量而定)，不可能将总体中所有符合要求的人员一一列出，直接抽取。但各个家庭户中，符合要求的（如年龄 18 岁以上）人数是不完全相等的，这就使得各户中符合要求的人（如年龄 18 岁以上）被选中的概率不同。这种情况一般用加权方法解决，在加权方法的应用“1”中专门讨论。

问题（2）是调查中经常遇到的。如被调查者对某些问题特别敏感，或者认为问题涉及到个人隐私等，被调查者会拒绝回答。拒绝回答使得涉及这些问题的有效样本量小于设计的样本量，而且不同层内情况会大不相同，不加修正的去估计总体时，必然带来偏误。

问题（3）是分层抽样中经常遇到的情况。例如，抽样设计一般是根据以前的调查，对总体按一定的标志分层（如经济类别或城乡类别），在每一层内按一定比例抽取样本。但是在抽样框中，每一样本的特征并不十分清楚，有些可能已经发生变化。只有到调查后样本的特征才会清楚。抽取后，往往会发现总的样本特征（如经济类别或城乡类别比重）偏离总体特征，这就需要采用事后分层的方法对样本进行修正，即加权处理后才能使样本对总体的估计是无偏的。

二、加权方法的应用

以下通过具体例子说明如何加权修正样本，减少偏误，以期达到样本对总体的无偏估计的方法。

1. 不等概率抽样时的调整

如在一个拥有 240 户的小区内，按抽样比 f= 1/40 抽取样本，即选取其中的 6 户进行调查，在每一被选中户内，随机抽选一位成年人男性，测试他的健康状况，指标之一是肺活量 u_i。

假如被选户是第 α，该户有 B_α 位成年人男性，第 β 位男性被选中，测得肺活量值为 u_α。若他在业 z_α= 1，不在业 z_α= 0

那么，成年男性 β 被选中的概率应该是:

$$p(\alpha\beta) = p(\alpha) * p(\beta \mid \alpha)$$

$$p(\alpha\beta) = (1/40) * (1/B_\alpha) = 1/w_\alpha,\ (w_\alpha = 40 * B_\alpha)$$

表 1　抽取户、调查登记和加权调整

户序号 (α)	户内成年男性数 B_α	权数 w_α	变量 u_α	变量 z_α	加权值 $u_\alpha w_\alpha$	加权值 $z_\alpha w_\alpha$
1	2	80	7	1	560	80
2	1	40	9	1	360	40
3	3	120	6	0	720	0
4	2	80	8	1	640	80
5	3	120	4	0	480	0
6	4	160	3	0	480	0
总计	15	600	37	3	3240	200

倘若仅以抽样比的倒数加权，会使得不同户（成年男性人数不等）的，代表性不同的入选的成年男性被同等对待。因此，每一被抽中的成年男性应按照被选中的概率倒数加权（1/p（αβ））。那么，变量值总和应该为 $\sum u_\alpha w_\alpha$、$\sum z_\alpha w_\alpha$，总测量数应该为 $\sum w_\alpha$。

加权平均数 $\bar{u}_w$、比例值 $\bar{z}_w$ 计算分别为:

$$\bar{u}_w = (\sum u_\alpha w_\alpha)/(\sum w_\alpha) = 3240/600 = 5.4$$

$$\bar{z}_w = (\sum z_\alpha w_\alpha)/(\sum w_\alpha) = 200/600 = 0.33$$

若不加权:

$$\bar{u} = (\sum u_\alpha)/6 = 6.2$$

$$\bar{z} = (\sum z_\alpha)/6 = 0.5$$

显然，加权与不加权对总体的估计相差很大，特别是成年男性的在业比重，相差 17 个百分点。样本加权平均数是对总体的无偏估计，即 E（$\bar{u}_w$）= Ū，E（$\bar{z}_w$）= Z̄。不加权的样本均值对总体估计是有偏的。

表 1 列出被选户、每户成年男性人数 B_α、变量值 u_α、z_α、权数 w_α 和加权后变量值 $u_\alpha w_\alpha$、$z_\alpha w_\alpha$。

2. 无应答时的修正方法

当某些问题的回答率不是 100%，而且在不同层内的回答率不相等时，需要加权调整。

例如在一个分层的多阶段抽样中，设计抽取 950 户，其中 400 户选自北部地区，550 户选自南部地区。出于某种考虑，北部地区的抽样率设计为南部的 2 倍。每一地区又分为城市和农村两部分。就某一问题（r）的回答率各不相同。设：变量 r 为被调查户拥有某种高档商品的数量，设：n_h，n'_h，n'_h 分别表示设计样

表 2　设计样本数及调查结果

地区		设计样本 n_h	获得样本 n_h'	变量 r_h'
北部	城市	100	70	60
	乡村	300	270	140
南部	城市	150	120	110
	乡村	400	380	210
总计		950	840	—

本数，获得的（有应答的）样本数和有应答户的拥有某种高档商品的数量，具体数值如表 2。

在这一例子中，应考虑两方面的问题：一是否是等概率抽取，二是回答率不同对总体估计的影响。

(1) 考虑不等概率抽取的加权调整

因已经知道北部地区的抽样率是南部地区的 2 倍，那么相对于北部地区，南部地区的入选概率应为 1/2。则不等概率抽取的加权因子 w_{1h}：北部为 1，南部为 2。这是第一步的加权。

(2) 对无应答的调整

获得的样本数 n_h' 与设计样本数 n_h 之比是回答率，而其倒数则表明每一获得样本对设计样本的代表程度，此数值应作为权数 w_{2h} 对获得样本数和相应变量的加权，修正无应答对总体估计带来的偏差。

权数应为：$w_{2h} = n_h / n_h'$。这是第二步的加权因子。

合并 (1) 和 (2)，最终权数为 $w_h = w_{1h} * w_{2h}$。用 w_h 对变量 r' 和获得样本数 n_h' 加权，如表 3 所示。则由样本估计总体中拥有某些高档商品的数量的比例应为：

$$(\sum w_h r_h') / (\sum w_h n_h') = 957.2 / 1497.8 = 0.639$$

表 3 列出各权数值及加权后的样本值和变量值。

表 3　多阶段分层抽样的加权调整

地区		权数因子 w_{1h}	权数因子 w_{2h}	最终权数 w_h	加权值 $w_h n_h'$	加权值 $w_h r_h'$
北部	城市	1	1.43	1.43	100.1	85.8
	乡村	1	1.11	1.11	299.8	155.4
南部	城市	2	1.25	2.50	300.0	275.0
	乡村	2	1.05	2.10	798.0	441.0
总计		—	—	—	1497.8	957.2

3. 事后分层方法[1]

在分层多阶段抽样中，当所获得的样本结构与总体有偏差时，应该用事后分层方法对样本和相应的变量进行修正。例如已经知道总体的城乡人口比例，但当所获得的样本的城乡人口比例与总体不符时，样本必须调整，使其与总体一致。相应的指标如城镇与乡村的文盲人数等指标都应该相应调整。以下以实例说明。

假设以等概率对某地区抽样，获得样本 23.8 万人，其中城镇人口 4.8 万，农村人口 19 万。抽样比为 0.0495。用样本估计总体，以抽样比的倒数加权得：总人口为 478 万，城镇人口 94.01 万，农村人口 383.99 万，城镇人口比重为 19.77%。而实际调查总体人口总量为 481 万，其中，城镇人口 125.9 万，农村人口 355.1 万。城镇人口比重为 26.17%。显然，样本构成与总体不一致，估计上有很大偏差。

修正方法是利用总体分布对样本加权，使得样本分布与总体一致，同时也对其他变量加权。

(1) 事后分层方法及计算结果

设：n_1，n_2 分别表示城镇、农村样本数，y_1，y_2 分别表示样本中城镇和乡村文盲人数，p_1，p_2 分别表示样本中城镇人口比例和农村人口比例，P_1，P_2 分别表示总体中城镇人口比例和农村人口比例，W_1，W_2 分别表示事后分层对城镇和农村的加权。

权数的计算公式为：$W_1 = P_1 / p_1$，$W_2 = P_2 / p_2$。

利用权数 w_i 重新计算城镇、乡村样本数。

事后分层的城镇样本数：$n_1^* = w_1 * n_1$，

事后分层的乡村样本数：$n_2^* = w_2 * n_2$，

事后分层的城镇文盲数：$y_1^* = W_1 * y_1$，

事后分层的乡村文盲数：$y_2^* = W_2 * y_2$，

利用事后分层的样本重新估计总体分城乡人数、总数及文盲人数、文盲率。

表 4　事后分层方法的应用

层	样本 n_I (I= 1, 2)	总体 N_I (I= 1, 2)	样本中 p_I (I= 1, 2)	总体中 P_I (I= 1, 2)	权数 w_I (I= 1, 2)	$n_I^* = w_I \cdot n_I$ (I= 1, 2)	文盲数 y_I (I= 1, 2)	$y_I^* = w_I \cdot y_I$ (I= 1, 2)
城镇	4.8	125.9	0.2017	0.2617	1.2976	6.23	0.6925	0.8986
农村	19	355.1	0.7983	0.7383	0.9248	17.57	4.8374	4.4736
合计	23.8	481	1	1	–	23.8	5.5300	5.3722

计算方法如下:

城镇人口数: $N_1 = n_1^* / f$

乡村人口数: $N_2 = n_2^* / f$

城乡人口总数: $N = N_1 + N_1$

其中 f 为抽样比, $f = n/N$

事后分层的文盲率:

事后分层的城镇文盲率:

$\bar{y_1} = (w_1 * y_1) / n_1^*$

事后分层的乡村文盲率:

$\bar{y_2} = (w_2 * y_2) / n_2^*$

事后分层的城乡文盲率:

$\bar{y} = (w_1 * y_1 + w_2 * y_2) / (w_1 * n_1 + w_2 * n_2)$

城镇人口比重= N_1/N

具体计算列于表 4。

(2) 无事后分层的指标估计

若不进行事后分层处理, 计算方法如下:

城镇人口数: $\hat{N_1} = n_1/f$

乡村人口数: $\hat{N_2} = n_2/f$

城乡人口总数: $\hat{N} = \hat{N_1} + \hat{N_1}$

城镇文盲率: $\hat{y_1} = y_1/n_1$

乡村文盲率: $\hat{y_2} = y_2/n_2$

事后分层的城乡文盲率:

$\hat{y} = (y_1 + y_2) / (n_1 + n_2)$

城镇人口比重= $\hat{N_1}/\hat{N}$

(3) 结果对比

如表 5 所示。(2) 与 (1) 比较, 无事后分层的计算结果, 城镇人口数、城镇人口比重、城乡人口总数被低估了, 农村人口数、城乡文盲率被高估了。显然, 事后分层方法运用, 修正了偏误, 提高了样本对总体估计的精确度。

表 5　无事后分层与事后分层结果对比

指　标	无事后分层	事后分层
城镇人口数 (万人)	94.01	125.88
乡村人口数 (万人)	383.99	355.12
城乡人口总数 (万人)	478	481
城镇文盲率	14.4%	14.4%
乡村文盲率	25.5%	25.5%
城乡文盲率	23.2%	22.6%
城镇人口比重	19.7%	26.2%

三、结论

加权方法的使用很好的解决了样本对总体代表性的问题, 这也是抽样调查中经常面临的亟待解决的问题。三种方法的使用应视具体情况而定。在复杂多阶段抽样中, 三种方法有时需同时被应用。

除以上所讨论的问题之外, 抽样调查还会遇到许多情况涉及样本对总体代表性的问题, 经常使用的方法还有: 对抽样数据的逻辑纠错, "热校" 方法, 对缺失值修正的回归方法等, 这些方法与本文所述的方法常常结合使用。

参考文献:

[1] 胡英. 事后分层方法在抽样调查中应用. 人口研究, 1997, (6).

[责任编辑　黄荣清]