41

某地区高校贫困生现状调查的抽样设计[①]

侯志强[②③]，李扬[③]

一、引言

近年来，高等院校学生的学费及生活费支出持续维持在高位，这无形中给家庭经济状况困难的学生增加了较大的物质和精神压力，为了切实摸清某地区高等院校贫困生整体的生活状况，特采用概率抽样的方法从某地区高等院校的所有学生中抽取部分学生进行调查，本文对此给出了相应的抽样及估计方法。

二、调查范围的界定

某地区除了公办高等院校外还有部分民办院校，由于公办高等院校的持续扩招，民办院校的招生越来越困难，学生的绝对数量不大，所以本调查将其排除在外。其余学生按照身份可以区分为全日制学生和业余学生，由于业余学生大都有工作，所以本调查也将其排除在外。另外，由于研究生与本科生在学习性质等方面存在明显不同，所以本调查也将其排除在外。确切地说，本调查的范围为某地区公办高等院校全日制本科生。

三、分层

依据招生范围，将某地区所有高等院校分为两层，一层为部属院校，另一层为省

① 本文得到中国人民大学研究生“创新性研究活动激励计划”项目的资助。

② 北方工业大学 统计系，北京 100041。

③ 中国人民大学 统计学院，北京 100872。

(市)属院校。部属院校主要面向全国招生,省(市)属院校主要面向某省(市)招生。根据预调查结果,部属院校的贫困生比例要高于省(市)属院校。

四、样本量的确定

(一)简单随机抽样下的样本量

简单随机抽样下,对于较大的总体比例,若要求其估计量的标准差上限为0.02,即

$$\sqrt{\frac{p(1-p)}{n}} \leqslant 0.02 \tag{1}$$

则样本量满足

$$n \geqslant \frac{p(1-p)}{0.02^2}$$

易知,$p(1-p)$ 的最大值为0.25,将其代入上式可得最低的样本量为625人。

对于较小的总体比例,若要求其估计量的相对标准差上限为0.2,即

$$\frac{\sqrt{\frac{p(1-p)}{n}}}{p} \leqslant 0.2 \tag{2}$$

则样本量满足

$$n \geqslant \frac{1-p}{0.2^2 p}$$

根据预调查结果,最小的总体比例估计为3%,则将 $p = 0.03$ 代入上式可得最低的样本量为809人。

综上所述,根据文献[1],可得保守的样本量为809人。因为贫困生的比例约占学生总数的1/6弱,所以,在简单随机抽样下,应该调查的学生人数最低大约为5 000人。

(二)复杂抽样下的样本量

由于本调查采用了复杂抽样设计,所以,本调查的样本量应该为简单随机抽样下样本量的某个倍数。这个倍数即是本抽样设计的设计效果系数。由于采用了分层*PPS*抽样技术,根据文献[2],估计本抽样设计的设计效果系数为2。这样,本调查的最终样本量确定为10 000人,即10 000个学生。

五、抽样方法

本调查采用分层三阶段抽样设计。抽样分别在部属院校层和省(市)属院校层中进行。第一阶段抽取高等院校,第二阶段从每个中选的高等院校中抽取自然班,第三阶段从每个中选的自然班中抽取学生。

第一阶段采用*PPS*抽样,辅助变量为各高等院校的学生人数(即属于调查范围的人数,不包括研究生和业余学生等,下文同)。比如,部属院校层中某高等院校的学生人数占该层总人数的比例为0.05,则其在每次抽取中入样概率皆为0.05。第二

阶段也采用 *PPS* 抽样，辅助变量为中选院校内各个自然班的人数。第三阶段采用随机起点等距抽样。

综合考虑精度与费用，确定抽取 25 所高等院校。高等院校的样本量在各层间按照各层的学生人数比例分配。从每所中选的高等院校中抽取 40 个自然班。从每个中选自然班中抽取 10 个学生。

根据文献[3]，上述抽样设计是自加权的，即从整体上看，该抽样设计类似于简单随机抽样。

六、参数估计

参数估计是指对反映贫困生现状的指标的估计，如对生活费支出的估计等。

如果界定前半年平均生活费在某数额以下的学生为贫困生，则根据文献[4]，可采用子总体估计的技术对反映贫困生现状的指标做出估计。

（一）均值的估计

设某地区高等院校共有 N 个学生，从中抽取了 n 个学生，并定义新的变量

$$\alpha_{ijkl}=\begin{cases}1,\text{若样本第 } i \text{ 层第 } j \text{ 所高等院校第 } k \text{ 个自然班第 } l \text{ 个人属于贫困生}\\0,\text{否则}\end{cases}$$

又设样本第 i 层第 j 所高等院校第 k 个自然班第 l 个人的观测值为 y_{ijkl}，则均值的估计为

$$\bar{y}=\frac{1}{n_0}\sum_{i=1}^{2}\sum_{j=1}^{m_i}\sum_{k=1}^{40}\sum_{l=1}^{10}\alpha_{ijkl}y_{ijkl} \tag{3}$$

其中，n_0 为样本中贫困生的数量，m_i 为第 i 层高等院校的样本量。

其方差估计为

$$v(\bar{y})=\frac{1-f}{n_0}\frac{1}{n_0-1}\sum_{i=1}^{2}\sum_{j=1}^{m_i}\sum_{k=1}^{40}\sum_{l=1}^{10}(\alpha_{ijkl}y_{ijkl}-\bar{y})^2 \tag{4}$$

其中，$f=\frac{n}{N}$。

（二）总量的估计

总量的估计为

$$\hat{Y}=\frac{N}{n}\sum_{i=1}^{2}\sum_{j=1}^{m_i}\sum_{k=1}^{40}\sum_{l=1}^{10}\alpha_{ijkl}y_{ijkl} \tag{5}$$

其方差估计为

$$v(\hat{Y})=\frac{N^2(1-f)}{n(n-1)}\left[\sum_{i=1}^{2}\sum_{j=1}^{m_i}\sum_{k=1}^{40}\sum_{l=1}^{10}(\alpha_{ijkl}y_{ijkl}-\bar{y})^2+np_0q_0\bar{y}^2\right] \tag{6}$$

其中，$f=\frac{n}{N}$，$p_0=\frac{n_0}{n}$，$q_0=1-p_0$。

比例指标为两个总量指标之比，因此，求出总量指标的估计后不难得出其估计。此处不再赘述。

七、无回答的处理

无回答有两种情形，一种是项目无回答，另一种是单元无回答。所谓项目无回答，指某个被调查者对调查问卷中的某个或某些问题没有回答，此时，相应的变量值缺失。对于缺失值，可以采用许多方法进行插补调整，比如，可以使用 hot deck 法。所谓单元无回答指某个被调查者未回答任何问题。单元无回答的存在不仅减少了有效样本量，而且还可能导致估计偏差。对于某个变量，若无回答单元与回答单元本质上并无很大差异，则可采用对估计量加权的技术得到总体参数的无偏估计量；若差异很大，则需要采取某些措施取得其数据。

八、结束语

本文给出的自加权分层三阶段抽样设计使得总体参数的估计量具有较简单的形式。由于调查问卷中包含某些敏感性问题，所以可能导致相应的变量值失真，进而导致估计失去意义。为了尽量避免这种情况的出现，可以采用某种随机化回答技术，但这已经超出了本文的范围，有兴趣的读者可以参考相关文献。

参考文献

[1]侯志强，吴启富. 抽样调查样本量的确定[J]. 全国商情·经济理论研究，2007，3(3)：108～109.

[2]金勇进. 设计效应应用中的若干问题[J]. 统计教育，2006，14(1)：6～7.

[3]侯志强，刘喜波. 分层三阶段及以上抽样的自加权抽样设计[J]. 数学的实践与认识，2007，37(6)：113～116.

[4]冯士雍，倪加勋，邹国华. 抽样调查理论与方法[M]. 北京：中国统计出版社，1998：63～67.