分层抽样下多项选择敏感问题随机抽样调查方法及应用

邱赛兵，唐　波
（湖南城市学院 数学与计算科学学院，湖南 益阳 413000）

[摘　要] 目的是为多项选择敏感性问题提供科学的、精度更高的随机抽样调查方法及其统计量的计算公式，设计出多项选择敏感问题分层抽样下的随机抽样调查模型，并推导出在此模型下总体比例的估计量及其估计方差的计算公式，计算出敏感属性比例 95% 的置信区间。并在湖南城市学院本科学生考试作弊情况的实例调查中取得了信誉度较高的应用效果。

[关键词] 敏感问题；多项选择；随机化调查；分层抽样
[中图分类号] O212.4　　[文献标识码] A　　[文章编号] 1673－0712(2013)02－0039－03

敏感性问题是指机构、组织或个人由于经济、安全、形象等原因不宜或拒绝让外部知晓的问题，如政府机密、企业商务秘密、个人隐私等。对于这类敏感性问题，调查中若采用直接回答的方式，被调查者为了保护自己的隐私或出于其他的目的，往往会拒绝回答或故意做出错误的回答。这样就破坏了我们收集数据的真实性。因而为了得到敏感性问题的可靠的样本数据，美国社会学家 S. L. Warner 在 1965 年首次提出了敏感问题的调查与统计处理技术，也称随机应答技术(Ranomized Response Technique，RRT)。RRT 使用特定的随机化装置，根据概率论知识计算出敏感问题特征在人群中的分布[1]。在沃纳模型中，总体总是被划分为互相排斥的两类，如“考试中作弊的学生”与“考试中没有作弊的学生”。但在实际中，常会碰到总体可划分多于两类的情况，如调查某厂职工对领导的满意程度，职工可分为“满意”、“一般”、“不满意”三种互斥情况。本文对多项选择敏感问题分层抽样下的随机应答模型进行了研究，并推导出了每种敏感属性的总体比例和方差估计值计算公式及敏感属性比例 95% 的置信区间。

一　调查方法

(一) 多项选择敏感问题的随机化调查模型

设一敏感问题可分为 k 种互相排斥的类别 $A_1, A_2, \ldots A_k$，为了估计 $A_1, A_2, \ldots A_k$ 在总体中所占的比例 $\pi_1, \pi_2, \ldots \pi_k$，制作 m 张外形完全一样的卡片，上面分别标上号码 0，1，2，$\cdots k$。其张数分别为 $m_0, m_1, m_2, \ldots m_k$，$\sum_{i=0}^{k} m_i = m$。把卡片放入盒子中，被调查者有放回的从袋中随机抽出一张卡片。若卡片上写有 0 则真实回答自己属于哪一类，若他摸到的卡片是 0 以外的数字 i，则回答数字 i[2]。整个过程都是在调查者无法确切知道具体情况下进行的。这样，调查者只知道一个数字，并不知道被调查者是属于哪一类型的人，从而较好的保护了被调查者的隐私。

(二) 多项选择敏感问题分层抽样下的 RRT 模型

对调查的总体若按与所调查敏感性问题相关的标志分层，则可提高其精度。假设总体分为 L 层，π_{hi} 为第 h 层中具有第 i 类敏感性特征的人在总体中所占的真实比例，第 h 层容量为 N_h，层权为 w_h，抽样的样本容量为 n_h，设 n_{hi} 表示抽取的 n_h 个人中回答 i 的人数，λ_{hi} 为此方案下每层样本中每个人回答 i 的概率。

二　公式推导

虽然原始分类多于两类，但当调查的目的是

[收稿日期] 2012－07－17.
[基金项目] 湖南省教育厅项目基金资助(1200578)。
[作者简介] 邱赛兵(1977—)，女，湖南益阳人，湖南城市学院数学与计算机学院讲师，硕士，研究方向：数理统计。

要估计总体中任何一类的个数占总数的比例时，实质上这些比例是按两分类得到的，所以两分类的理论均适合与此种情况[3]。λ_{hi} 分别服从参数为 (λ_{hi}, n_{hi}) 的二项分布。

设摸到卡片号码为 0 的概率为 $p_0 = \frac{m_0}{m}$，p_i 为卡片号码不为 0 的情况下，号码为 i（$i = 1,2,\ldots k$）的概率，则它是个条件概率，其值为：

$$p_i = \frac{m_i}{m - m_0} = \frac{m_i}{m} \cdot \frac{1}{1 - p_0} \tag{1}$$

由全概率公式可得：

$$\lambda_{hi} = p_0\pi_{hi} + (1 - p_0)p_i \tag{2}$$

则有：

$$\pi_{hi} = \frac{\lambda_{hi} - (1 - p_0)p_i}{p_0} \tag{3}$$

显然，$\hat{\lambda}_{hi} = \frac{n_{hi}}{n_h}$ 为 λ_{hi} 的极大似然无偏估计量，$\mathrm{var}(\hat{\lambda}_{hi}) = \frac{\lambda_{hi}(1 - \lambda_{hi})}{n_h}$，所以 π_{hi} 的极大似然无偏估计量为：

$$\hat{\pi}_{hi} = \frac{\frac{n_{hi}}{n_h} - (1 - p_0)p_i}{p_0} \tag{4}$$

方差为：$\mathrm{var}(\hat{\pi}_{hi}) = \frac{1}{n_h p_0^2}\lambda_{hi}(1 - \lambda_{hi})$ (5)

于是，总体比例的估计量及其方差为为：

$$\hat{\pi}_i = \sum_{h=1}^{L} w_h \hat{\pi}_{hi} = \sum_{h=1}^{L} w_h \left[\frac{n_{hi}}{n_h} - (1 - p_0)p_i\right] \cdot \frac{1}{p_0} \tag{6}$$

$$\mathrm{var}(\hat{\pi}_i) = \sum_{h=1}^{L} w_h^2 \mathrm{var}(\hat{\pi}_{hi}) = \sum_{h=1}^{L} w_h^2 \frac{1}{n_h p_0^2}\lambda_{hi}(1 - \lambda_{hi}) \tag{7}$$

定理 $\hat{\pi}_i$ 是 π_i 的无偏估计量。

证明 $E(\hat{\pi}_i) = \sum_{h=1}^{L} E(w_h \hat{\pi}_{hi})$

$$= \sum_{h=1}^{L} \frac{w_h}{p_0} E\left[\hat{\lambda}_{hi} - (1 - p_0)p_i\right]$$

$$= \sum_{h=1}^{L} \frac{w_h}{p_0} \left[\lambda_{hi} - (1 - p_0)p_i\right]$$

$$= \sum_{h=1}^{L} w_h \pi_{hi} = \pi_i$$

对于等比例分配有[4]：

$\frac{n}{N} = \frac{n_h}{N_h}$，$n = \sum_{h=1}^{L} n_h$，$N = \sum_{h=1}^{L} N_h$，$w_h = \frac{N_h}{N} = \frac{n_h}{n}$

则总体比例的估计量和方差为：

$$\hat{\pi}_i = \frac{\frac{\sum_{h=1}^{L} n_{hi}}{n} - (1 - p_0)p_i}{p_0} \tag{8}$$

$$\mathrm{var}(\hat{\pi}_i) = \sum_{h=1}^{L} \frac{n_h}{n^2 p_0^2}\left[p_0\pi_{hi} + (1 - p_0)p_i\right]\left[1 - p_0\pi_{hi} - (1 - p_0)p_i\right] \tag{9}$$

在 Neyman 最优分配情形下有：

$$n_h = n \cdot \frac{N_h\sqrt{\lambda_{hi}(1 - \lambda_{hi})}}{\sum_{h=1}^{L} N_h\sqrt{\lambda_{hi}(1 - \lambda_{hi})}}$$

所能达到的最小方差为：

$$min\{\mathrm{var}\}(\hat{\pi}_i) = \frac{\left[\sum_{h=1}^{L} w_h\sqrt{\lambda_{hi}(1 - \lambda_{hi})}\right]^2}{np_0^2} - \frac{1}{N}\sum_{h=1}^{L} w_h\pi_{hi}(1 - \pi_{hi}) \tag{10}$$

π_i 的 95% 的置信区间为：

$$\left(\hat{\pi}_i - \mu_{0.05}\sqrt{Var(\hat{\pi}_i)}, \hat{\pi}_i + \mu_{0.05}\sqrt{Var(\hat{\pi}_i)}\right) \tag{11}$$

三 应用实例

以湖南城市学院朝阳校区全体在校大学二、三、四年级学生为总体，调查指标为多项选择敏感问题：上学年考试作弊的严重程度 k，$k = 1,2,3,4$ 分别表示作弊次数为 0，1，2 次和大于 2 次。$\hat{\pi}_{ik}$ 表示 i 年级学生作弊次数为第 k 类属性所占的比例，$i = 2,3,4$；$k = 1,2,3,4$。划分总体为三层，大二为第一层，共 1100 人，大三为第二层共 1280 人，大四为第三层共 1080 人，分别随机抽取 110 人，128 人，108 人进行调查。设计 20 张外形相同的卡片，12 张写上数字 0，其它分别写上 1，2，3，4 其张数均为 1 张，混合均匀放入盒内。被调查的学生有放回地随机从盒内抽取一张卡片，若卡片上写的是数字 0 则真实回答自己上学年曾作弊的次数，若是 0 以外的数字则回答该数字。本次调查问卷回收率达 100%，回收问卷合格率为 100%. 用 Excel2000 建立数据库录入数据，对所有资料进行手工及计算机纠错. 数据管理与计算通过 Excel2000 及 SAS9.13 完成.

表 1　各年级考试作弊严重程度比例

年级	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$	$\hat{\pi}_{i4}$
2	0.7019	0.1512	0.0926	0.0543
3	0.6521	0.1783	0.1065	0.0640
4	0.6023	0.2071	0.1226	0.068

按(6)式计算得大二、大三、大四学生在考试作弊严重程度的的估计值 $\hat{\pi}_1 = 0.6523$，$\hat{\pi}_2 = 0.1791$，$\hat{\pi}_3 = 0.1071$，$\hat{\pi}_3 = 0.0615$。其中 $w_1 = 0.3180$，$w_1 = 0.3699$，$w_1 = 0.3121$。

按(7)式计算 $\hat{\pi}_1$，$\hat{\pi}_2$，$\hat{\pi}_3$，$\hat{\pi}_4$ 估计方差分别为:

$Var(\hat{\pi}_1) = 0.00056$，$Var(\hat{\pi}_2) = 0.000067$，$Var(\hat{\pi}_3) = 0.000087$，$Var(\hat{\pi}_4) = 0.000034$

四　结论

分层随机抽样的特点是在获得总体参数估计的同时，也能获得有关各层的参数估计。在分层抽样过程中，可将差异不太大的对象归为一类，从而使分层抽样的样本比纯随机抽样的样本更具有代表性，也使得抽样调查中的数据的收集汇总和处理更为方便，操作性更强[5]。由前面的公式(10)和(11)，我们容易得出，对分层随机抽样，当 $\frac{1}{N_h} \to 0^+$，则 Neyman 最优分配下估计量方差小于等比例分配下的方差，且这两种抽样下的精度均高于纯随机简单抽样。

参考文献:

[1] WARNER S L. Randomized Response: a Survey Technique for Eliminating Evasive Answer Bias [J]. American statistical Association, 1965(60): 63 – 69.

[2] ARIJIT CHAUDHURI. Christofides' Rondomized Response Technique in Complex Sample Surveys [J]. Metrika, 2004(60): 223 – 228.

[3] 贺志龙，高歌. 多项选择敏感问题 RRT 二阶段抽样的统计方法及应用 [J]. 中国卫生统计，2009(26): 580 – 582.

[4] 孙明举，段刚，孙山泽. 多项选择随机化调查的多样本模型. 数理统计与管理 [J]. 2004，19(2): 61 – 63.

[5] 高歌，范玉波. 敏感问题改进的随机应答技术模型分层整群抽样研究及应用 [J]. 苏州大学学报: 医学版，2008(5): 750 – 754.

Random Sampling Survey Method of Sensitive Issues and Application of Multiple Choice under Stratified Sampling

QIU Sai-bing, *TANG Bo*

(Institute of Mathmatics and Computer Science, Hunan City University, Yiyang 413000, China)

Abstract: Objective is to provide scientific precision calculating formula of higher random sampling investigation method and statistics for multiple choices sensitive problems, design a multiple choice random sampling model for sensitive question under stratified sampling, calculation formula of estimator and its variance and deduced the overall proportion, to compute confidence intervals for the sensitive attribute ratio 95%. Investigation on cheating in the undergraduate student exam of Hunan City University has been achieved high reputation.

Key words: sensitive questions; multiple choices; randermized survey; statified sampling

(责任编校: 李传熹)