多项选择敏感问题 RRT二阶段抽样的统计方法及应用*

苏州大学公共卫生学院流行病与卫生统计教研室 (215123)　贺志龙　高　歌△　王　冕　朱宏儒　于明润　李旭东

【提　要】 **目的**　为多项选择敏感性问题提供科学的抽样调查方法及其统计量的计算公式。**方法**　本文中,多项选择敏感问题 RRT模型、两阶段抽样调查方法、抽样理论、全概率公式、方差的基本性质等理论与方法被应用。**结果**　推导出多项选择敏感问题随机应答模型在两阶段抽样下总体比例的估计量及其估计方差的计算公式;应用此调查方法及相关公式调查得苏州大学学生近两个学期考试作弊 0次、1～2次、＞2次的比例分别为 68.066%, 18.072%, 13.861%。**结论**　多项选择敏感问题随机应答模型下的二阶段抽样调查方法及相应的统计计算公式科学可行。

【关键词】 敏感问题　多项选择　随机应答技术　二阶段抽样　考试作弊

所谓敏感性问题,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外部知晓的问题,如未婚流产、吸毒、考试作弊、艾滋病、性取向等[1]。对于属于个人的隐私或秘密,即敏感问题,被调查者可能不回答或有意高报和低报,造成特殊的系统误差——敏感问题误差。为了控制这类误差,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 (randomized 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[2]。它避免了被调查者在没有任何保护的情况下直接回答敏感问题,从而能取得被调查者的信任,获得较为真实的资料。在沃纳 (Warner)模型中,总体总是被划分成互相排斥的两类,但在实际中,常会碰到总体可划分为多于两类的情况,即多项选择敏感问题,如调查大学生最常用的自慰方式,可分为"手淫","借助工具","其他方式","从未有过自慰行为"四种互斥的情况。

本文对多项选择敏感问题随机应答模型的二阶段抽样调查的统计方法进行了研究,并结合苏州大学学生考试作弊严重程度的调查实例,取得了良好的应用效果。

调查方法

1. 多项选择敏感问题的 RRT模型

设一敏感问题分为 k种互斥的类别 1, 2, …, k,为了估计 1, 2, …, k类在总体所占的比例 p_1, p_2, …, p_k,设计一随机化装置,如:将分别写有 0, 1, 2, …, k的 k+1种按数量比例 P_0: P_1: P_2 … P_k ($P_0+P_1+P_2+\cdots+P_k=1$) 的若干卡片混合放入袋中。每个抽中的人有放回地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敏感问题的那一类的序号;若卡片上写有 0以外的某个数则回答该数。调查的整个过程都是在调查者无法知道具体情况下进行,这样调查者只知道被调查者给出的一个数字,无法知道被调查者的真实情况,从而保护了被调查者的隐私。

2. 多项选择敏感问题 RRT模型的二阶段抽样

假定总体由 N_1个一级单位组成,第 i个一级单位由 N_{i2}个二级单位组成, i=1, 2, …, N_1,平均每个一级单位包含 $\bar{N}_2$个二级单位。又假定第一阶段随机抽取 n_1个一级单位,第二阶段从第 i个被抽中的一级单位内随机抽取 n_{i2}个二级单位, i=1, 2, …, n_1,平均从每个被抽中的一级单位内抽取了 $\bar{n}_2$个二级单位。对每个被抽中的二级单位 (被调查的个人)采用多项选择敏感问题 RRT模型进行调查。

公式推导

1. 总体比例的估计量及其估计方差

虽然原始分类多于两类,但当调查的目的是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类得到的,所以两分类的理论均适合于此种情况[3]。

假设第 i(i=1, 2, …, n)个抽中一级单位属于类别 j的比例为 π_{ij},总体中属于类别 j的比例为 π_j。p_{ij}记第 i个抽中一级单位第 j类的样本比例,根据 Jianfeng Wang Ge Gao给出的结果[4],第 j类的总体比例的估计量 p_j为:

$$p_j=\frac{\sum_{i=1}^{n_1}N_{i2}p_{ij}}{\sum_{i=1}^{n_1}N_{i2}} \tag{1}$$

p_j的估计方差 $v(p_j)$为:

$$v(p_j)=\frac{s_{1j}^2}{n_1}\left(1-\frac{n_1}{N_1}\right)+\frac{s_{2j}^2}{n_1\bar{n}_2}\left(1-\frac{\bar{n}_2}{\bar{N}_2}\right) \tag{2}$$

其中, $S_{1j}^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(\frac{N_{i2}}{\bar{N}_2}\right)^2(p_{ij}-p_j)^2$ (3)

*:国家自然科学基金项目 (项目编号: 30571620; 30972548)
△通讯作者:高歌, E-mail: gaoge@suda.edu.cn

$$S_{2j}^2=\frac{1}{\sum_{i=1}^{n_1}N_{2i}}\sum_{i=1}^{n_1}N_{2i}p_{ij}(1-p_{ij}) \quad (4)$$

2. p_{ij}的估计

设 π_{ij}为第 i个一级单位第 j类敏感问题的比例，以 m_{ij}记样本中第 i个一级单位回答数字 j的频数，λ_{ij}（$\lambda_{ij}=m_{ij}/n_{2i}$）表示样本中第 i个一级单位回答数字 j的概率，根据全概率公式[5]：

$\lambda_{ij}=\pi_{ij}P_0+P_j$，则有：$\pi_{ij}=\frac{\lambda_{ij}-P_j}{P_0}$，于是：

$$p_{ij}=\frac{\lambda_{ij}-P_j}{P_0} \quad i=1,2,\cdots,n_1;\ j=1,2,\cdots,k \quad (5)$$

应用实例

1. 调查设计

研究总体为苏州大学独墅湖校区在校学生，共 12个学院（N_1），总计 7 274人，平均每个学院约 1 139人（$\bar{N}_2$）。调查指标为多项选择敏感问题——上两学期考试作弊的严重程度 j，j=1(作弊 0次)，j=2(作弊 1～2次)，j=3(作弊 >2次)。以学院为一级单位，以学生为二级单位，采用二阶段抽样：第一阶段随机抽取 5个学院（n_1），第二阶段分别从被抽中的学院内随机抽取共 1 738人，平均从每个学院抽取了约 347人（$\bar{n}_2$）。使用多项选择敏感问题 RRT模型：设计一套随机装置，将 10个大小、重量、触感完全相同小球，4个写上数字 0，2个写上数字 1，2个写上数字 2，2个写上数字 3，即 P_0：P_1：P_2：P_3 为 0.4：0.2：0.2：0.2（$P_0+P_1+P_2+P_3=1$），混合放入袋中；每个学生有放回地从袋中随机抽中一个小球，若小球上写有 0则真实回答敏感问题（自己上两学期考试作弊的次数）；若小球上写有 0以外的数字则回答该数字。

2. 数据管理与计算

对收集的问卷进行仔细检查，必须是独立完成的完整问卷，无漏填项目。本次调查问卷回收率 100%、回收问卷合格率 100%。用 Excel 2003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 Excel 2003及 SAS 9.13完成。

3. 各学院考试作弊严重程度的调查计算结果

按（5）式计算得各学院考试作弊严重程度三类的发生比例 p_{ij}，结果见表 1。

4. 校区（总体）考试作弊各严重程度比例的估计及其估计方差

按（1）式计算得整个校区考试作弊严重程度中第 1类的样本比例 p_1 为：$p_1=\frac{\sum_{i=1}^{5}N_{2i}p_{i1}}{\sum_{i=1}^{5}N_{2i}}=\frac{1098\times73.352\%+\cdots+1819\times72.146\%}{(1098+2422+\cdots+1819)}=68.066\%$

表 1 多项选择敏感问题 RRT模型的两阶段抽样调查苏大学生考试作弊严重程度比例（%）

学院	p_{i1}	p_{i2}	p_{i3}
1	73.352	11.676	14.972
2	60.987	20.628	18.386
3	73.428	18.396	8.176
4	67.188	15.848	16.964
5	72.146	19.635	8.219
（校区）p_j	68.066	18.072	13.861

按（3）式计算得：

$$S_{11}^2=\frac{1}{5-1}\sum_{i=1}^{5}\left(\frac{N_{2i}}{\bar{N}_2}\right)^2(p_{i1}-p_1)^2=0.00784$$

按（4）式计算得：

$$S_{21}^2=\frac{1}{\sum_{i=1}^{5}N_{2i}}\sum_{i=1}^{5}N_{2i}p_{i1}(1-p_{i1})=0.21449$$

按（2）式计算得 p_1 的估计方差为：

$$v(p_1)=\frac{s_{11}^2}{n_1}\left(1-\frac{n_1}{N_1}\right)+\frac{s_{21}^2}{n_1\bar{n}_2}\left(1-\frac{\bar{n}_2}{\bar{N}_2}\right)=\frac{0.00784}{5}\left(1-\frac{5}{12}\right)+\frac{0.21449}{5\times347.6}\left(1-\frac{347.6}{1139.42}\right)=0.00099$$

由此，可得考试作弊严重程度中第 1类的总体比例的 95%可信区间为：

$$p_1\pm1.96\times\sqrt{v(p_1)}=0.61868\sim0.74264$$

同理，$p_2=\frac{\sum_{i=1}^{5}N_{2i}p_{i2}}{\sum_{i=1}^{5}N_{2i}}=18.072\%$

$$v(p_2)=\frac{s_{12}^2}{n_1}\left(1-\frac{n_1}{N_1}\right)+\frac{s_{22}^2}{n_1\bar{n}_2}\left(1-\frac{\bar{n}_2}{\bar{N}_2}\right)=0.00029$$

$$p_2\pm1.96\times\sqrt{v(p_2)}=0.14757\sim0.21387$$

$$p_3=\frac{\sum_{i=1}^{5}N_{2i}p_{i3}}{\sum_{i=1}^{5}N_{2i}}=13.861\%$$

$$v(p_3)=\frac{s_{13}^2}{n_1}\left(1-\frac{n_1}{N_1}\right)+\frac{s_{23}^2}{n_1\bar{n}_2}\left(1-\frac{\bar{n}_2}{\bar{N}_2}\right)=0.00064$$

$$p_3\pm1.96\times\sqrt{v(p_3)}=0.08906\sim0.18816$$

讨　论

1. 由于敏感性问题的隐秘性特点，用一般的观察方法和调查技术难以获得调查对象可靠的信息数据。1965年沃纳首先创立了一个二项属性特征的敏感性问题的随机化回答模型，它提供了对诸如考试作弊、漏税、吸毒等敏感性问题进行调查的一种方法，被命名为沃纳模型。1967年由西蒙斯（Simmons）对沃纳模型进行了改进，形成了西蒙斯模型。由此形成了专门调

查敏感性问题的随机化回答技术。在沃纳模型和西蒙斯模型中,总体是被划分成互斥的两类,但在实际工作中,常会碰到总体可划分多于两类的情况,即多项选择的敏感问题[6]。本文对多项选择敏感问题 RRT模型二阶段抽样调查方法进行了研究,推导出多项选择敏感问题 RRT在二阶段抽样下总体比例估计量及其估计方差的计算公式,并成功地应用于大学生考试作弊的调查中,此前未见国内外文献报道。

2. 本文是通讯作者主持的国家自然科学基金项目——“敏感性问题的抽样调查设计”的一部分。该国家项目分别在多种 RRT模型多种抽样调查方法的多种组合下,设计了调查方法及推导出相应的统计公式,并对调查方法及其公式进行了信度评价,说明本文介绍的多项选择敏感问题 RRT模型下二阶段抽样的调查方法和统计公式具有很高的信度[7,8]。

3. 对于敏感问题,若像对于非敏感性问题那样,按通常的回答方式要求被调查者一一回答每一个问题,被调查者必然会拒绝回答或做出不真实的回答[9]。在多项选择敏感问题随机回答方案的抽样调查设计中,加入了随机化装置,调查者并不知道被调查者回答的数值是在随机化装置中抽到的随机数字,还是回答敏感性问题的具体数值,再加上问卷调查的匿名性,更能得到被调查者的配合。但是这种随机化装置也有缺点:即采用了一套较为复杂的随机化装置,可能会增加被调查者的反感情绪和产生一定程度的不合作,且调查范围有限制,一般只能由调查者实地调查。

4. 在调查敏感问题时为降低非抽样误差,提高数据质量,在应用二阶段抽样调查方法时需要注意以下问题:(1)调查员要完全掌握 RRT的原理及实施方法,并对被调查者作详细的解释,让他们懂得 RRT之所以能保密的原理,从而取得被调查者的信任。(2)样本的选择一定要遵循随机化的原则,并且要求有足够的样本量,否则很难保证调查结果的可靠性。

5. 随着我国经济体制的改革和市场化进程的加快,改革开放过程中西方文化、价值观念和道德标准的侵入,在社会生活的方方面面新产生出大量的影响社会稳定和经济健康发展的敏感性问题。这些敏感性问题往往反映的正是社会经济发展过程中产生或者存在着的各种深层次矛盾,需要政府和社会共同努力,设法了解、揭示或揭露这类问题。故敏感性问题的调查日益突出,提出一个好的解决方案非常重要,以抑制、避免此类问题的产生或打击由此引发的各种犯罪活动,以维护社会秩序,促进社会发展。

6. 统计工作要求调查数据实现“快”、“精”、“准”,数据不“准”会失去其应用价值,甚至会对决策者形成误导。尽管上述方法有一定的局限性,但在敏感问题的实际工作中仍不失为较好的解决方法。

Statistical Methods of Two-stage Sampling on the Randomized Response Technique for Sensitive Question Survey with Multiple Choices and its Application　He Zhilong, Gao Ge, Wang Mian, et al. Public Health School, Soochow University (215123), Suzhou

【Abstract】 Objective To explore scientific sampling methods and corresponding formulas for multiple choices sensitive questions survey with the sample selected by two-stage sampling. Methods Randomized response technique of multiple choices sensitive questions and two-stage sampling were showed in this paper. Moreover, sampling theories, total probability formulas and properties of variance were used. Results Formulas for the estimator of the population proportion and its variance on randomized response model of multiple choices sensitive questions in two-stage sampling were deduced. Our survey methods and formulas on the RRT model may have a successful application for multiple choices sensitive questions survey, what students of Soochow University cheated in exams. The severity of cheating in exams was classified into three types and the proportion was 68.066%, 18.072% and 13.861% respectively. Conclusion The methods and corresponding formulas on multiple choices sensitive questions survey with the sample selected by two-stage sampling were feasible.

【Key words】 Sensitive questions; Multiple choices; Randomized response technique; Two-stage sampling; Cheating in exams

参 考 文 献

1. 张泮洲. 敏感问题调查技术新探. 统计研究, 2001, 11: 48-50.
2. Maarten JLF, Cruyff, Ardo van den Hout, et al. Log-linear randomized-response models taking self-protective response behavior into account. Sociological Methods & Research, 2007, 36(2), 266-282.
3. Cochran WG. Sampling Techniques 3rd Edition. Wiley, New York, 1977: 60-61.
4. Wang JF, Gao G. The estimation of sampling size in multi-stage sampling and its application in medical survey. Applied Mathematics and Computation, 2006, (178), 239-249.
5. 苏良军, 高等数理统计. 北京: 北京大学出版社, 2007: 3.
6. 孙明举, 段钢, 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, 2004, 19(2): 61-63.
7. 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. 苏州大学学报(医学版), 2008, 5: 750-754.
8. Wang Mian, Gao G. Quantitative sensitive question survey in cluster sampling and its application. Recent Advance in Statistics Application and Related Areas, 2008, 648-652.
9. 王春平, 王志锋, 张光成. 属性特征敏感性问题的设计、分析及评价. 中国卫生统计, 2006, 23(1): 60-62.