【文章编号】 1004-1540(2006)03-0251-03

是非型敏感问题抽样调查的统计研究

刘春雨[1]，王宏宇[2]
(1.中国计量学院 管理学院，浙江 杭州 310018；2.大连海洋学校 基础部，辽宁 大连 116000)

【摘　要】 抽样调查的方法已经在社会生活的各个领域得到了普遍应用，但对于敏感性问题的调查由于问题的特殊性，一般方法效果不佳.运用沃纳模型(Warner model)随机化回答技术解决这一问题，并就该方法给出了无偏估计的证明和样本量的计算公式.

【关键词】 敏感性问题；沃纳模型；无偏估计；样本量
【中图分类号】 C811　　【文献标识码】 A

In sample survey statistical research on Y/N-sensitive questions

LIU Chun-yu[1], WANG Hong-yu[2]
(1.College of Management, China Jiliang University, Hangzhou 310018, China; 2.Dalian Ocean School, Dalian 11600, China)

Abstract: Nowadays, sample survey enjoys universal applications. However, because of the sensitive and privateness of the question, the general method is not effective. This article plans to use the Warner model randomization technology to solve this problem. Proof and the formula on unbiased estimate and the sample size are given.

Key words: sensitive questions; Warner model; unbiased estimate; sample size

抽样调查现在已被广泛应用于社会经济生活的各个方面.它省时省力，能比较迅速获得较为可靠的结果.在各种社会调查中，经常会遇到诸如吸毒、走私、赌博、考试作弊、收入财产、性行为取向等敏感性私密问题.调查中若采用直接问答方式，被调查者为保护自己的隐私或出于其他目的，往往会拒绝回答或故意做出错误的回答.这样就破坏了数据的真实性.运用随机化回答技术设计问题进行调查可以解决此问题.本文就是非型敏感性问题调查加以分析.

在市场调查教学使用的教材中发现这样一个问题.下面是原文内容：这个方法是沃纳1965年提出来的.它向被调查者提出两个问题.假如要调查对改革的看法，这两个问题可以是：

A：您赞成改革吗?　　1—是　　2—不是

B：您不赞成改革吗?　　1—是　　2—不是

被访者随机抽一个问题回答.调查员不知道每个人回答哪个问题，但回答A类问题的人占的比例是他事先确定的.例如向100个人调查，取 $P=0.7$(注意 p 不能取0.5).可以制作100张卡

【收稿日期】 2006-03-10

【作者简介】 刘春雨(1971—)，男，辽宁黑山人，硕士.主要研究方向为统计学及数量经济分析.

片，其中 70 张上印有问题 A，30 张上印有问题 B．让被访者随意抽取问题卡片[1]．

当时就有疑问为什么 p 不能取 0.5？简单地以(0－1)分布方差当 $P=0.5$ 时最大作为解释．

1 关于沃纳模型

它是 1965 年沃纳提出了沃纳模型(Warner model)后才发展起来的．其基本思想是所论及的总体为一个简单的二元总体，即总体中每个单位或属于 A，或不属于 $\bar{a}$，除此之外，别无它属．根据敏感性特征设计两个相互对立的问题，让被调查者按预定的概率从中选一个回答，调查者无权过问被调查者究竟回答的是哪一个问题，从而起到了为被调查者保密的效果[3,6]．

1) 抽样方法及比例的估计：

制作两类卡片：

卡片 1："具有特征 A"

卡片 2："不具有特征 A"

卡片 1 在其中所占的比重以 p 表示；卡片 2 所占的比重相应为 $1-p$．

在简单随机有放回抽样下从总体中抽得 n 个人作样本，然后对这 n 个样本单位进行调查．让其从混合均匀的两类卡片中任取一张卡片，根据卡片上的问题作答．则此人回答"是"的概率是：

$$\begin{aligned}\lambda &= p(\text{具有特征 A 的人，并回答是的}) + \\ &\quad p(\text{不具有特征的 A 的人，并且回答是}) \\ &= p(\text{具有特征 A 的人}) \times \\ &\quad p(\text{回答是} \mid \text{具有特征 A 的人}) + \\ &\quad p(\text{不具有特征 A 的人}) \times \\ &\quad p(\text{回答是} \mid \text{不具有特征 A 的人}) \\ &= \pi_A \cdot p + (1-\pi_A)(1-p) \\ &= (1-p) + (2p-1)\pi_A \qquad (1)\end{aligned}$$

$$\lambda_1 = \pi_A(1-p) + (1-\pi_A)p \qquad (\text{回答否的概率})$$

于是由式(1)可解出：$\pi_A = \dfrac{\lambda-(1-p)}{2p-1}(p \neq 1/2)$．因此如果能得到 λ 的估计量 $\hat{\lambda}$，就可用它来估计 λ，进而可估计 π_A，即具有特征 A 的比例．

为了得到 λ 的估计量 $\hat{\lambda}$，从 N 中用重复抽样方法取一个容量为 n 的简单随机样本，令 n 个人分别用重复抽样方法从两类卡片中各取一张，并根据抽中卡片上的问题作出真实回答，则可得到 λ 估计量：$\hat{\lambda} = \dfrac{n_A}{n}$，其中 n_A 为样本中回答"是"的人数．$\hat{\lambda}$ 满足：

①$E(\hat{\lambda}) = \lambda$ ②$D(\hat{\lambda}) = \dfrac{1}{n}\Lambda(1-\Lambda)$，其无偏估计量为 $\dfrac{1}{n}\hat{\lambda}(1-\hat{\lambda})$．

证明：记 $x_i = 1$(回答"是")或 $x_i = 0$ 时(回答"不是")．由于 x_i 和 x_j 相互独立，于是：

$$E(x_i) = \lambda,$$

$$D(x_i) = E(x_i^2) - (Ex_i)^2 = \lambda - \lambda^2 = \lambda(1-\lambda),$$

$$E(x_i x_j) = E(x_i)E(x_j) = \lambda^2.$$

令计数函数 $n_A = \sum_{i=1}^{n} Ex_i$，于是

$$E(n_A) = \sum_{i=1}^{n} Ex_i = \sum_{i=1}^{n} \lambda = n\lambda,$$

$$D(n_A) = D\left(\sum_{i=1}^{n} x^i\right) = \sum_{i=1}^{n} D(x_i) = n\lambda(1-\lambda).$$

$$\begin{aligned}E[\hat{\lambda}(1-\hat{\lambda})] &= E(\hat{\lambda} - \hat{\lambda}^2) = E\left(\frac{n_A}{n} - \left(\frac{n_A}{n}\right)^2\right) \\ &= \frac{n\lambda}{n} - \frac{1}{n^2}E(n_A^2) \\ &= \lambda - \frac{1}{n^2}(D(n_A) + E^2(n_A)) \\ &= \lambda - \frac{1}{n^2}(n\lambda(1-\lambda) + (n\lambda)^2) \\ &= \frac{n-1}{n}\lambda(1-\lambda)\end{aligned}$$

所以：

$$\begin{aligned}E(\hat{\lambda}) &= \frac{1}{n}E(n_A) = \lambda, D(\hat{\lambda}) = \frac{1}{n^2}D(n_A) \\ &= \frac{1}{n}\lambda(1-\lambda), E\left[\frac{1}{n}\hat{\lambda}(1-\hat{\lambda})\right] \\ &= \frac{1}{n} \cdot E[\hat{\lambda}(1-\hat{\lambda})] = \frac{1}{n-1} \cdot \frac{n-1}{n} \cdot \lambda(1-\lambda) \\ &= \frac{1}{n}\lambda(1-\lambda).\end{aligned}$$

以 $\hat{\lambda}$ 作为 λ 的估计量，可得 π_A 的无偏估计量：

$\hat{\pi}_A = \dfrac{\hat{\lambda}-(1-p)}{2p-1}(p \neq 1/2)$，$\hat{\pi}_A$ 满足：

(a) $E(\hat{\pi}_A = \hat{\pi}_A$，

(b) $D(\hat{\pi}_A) = \dfrac{1}{(2p-1)^2}D(\hat{\lambda})$

$= \dfrac{1}{(2p-1)^2} \cdot \dfrac{1}{n} \cdot \lambda(1-\lambda)$，

(c) $D(\hat{\pi}_A) = \dfrac{1}{(2p-1)^2} \cdot \dfrac{1}{n-1} \cdot \lambda(1-\lambda)$．

在 Warner 方法当中，p 不能取 1/2；其抽样误差：

将 $\lambda=(2p-1)(1-\pi_A)+(1-p)$ 代入方差公式，得：

$$D(\hat{\pi}_A)=\frac{\pi_A\cdot(1-\pi_A)}{n}+\frac{p(1-p)}{n(2p-1)^2}.$$

上面等式右端第一项是直接回答时估计量的方差；第二项是随机化回答引起的误差增量．当 $p=1$ 或 $p=0$ 时，第二项为 0，此时相当于直接回答，但不能保护被调查者的隐私，难于得到真实资料；当 $p\to 1/2$ 时，对被调查者的保护程度加强，但是方差将增大．这是 $p\neq 1/2$ 的真正原因．对此问题的解决可以采用西蒙模型(1967)、二次随机化模型(赵俊康 1994)、改进的随机化模型．

2) 卡片比例 p 与样本量的确定

由上式可以看出，$p\to 1/2$，则 $D(\hat{\pi}_A)$ 的值越大．当 p 越靠近 0 或 1 时，$D(\hat{\pi}_A)$ 就越小．但当 p 越接近 1 或 0 时，对被调查者的保护程度就会降低，从而降低被调查者的合作程度，使随机化回答的作用降低，增加收集真实数据的难度．p 的取值一般应根据实际调查问题的敏感程度适当选取．

由于 $D(\hat{\pi}_A)=\frac{\pi_A(1-\pi_A)}{n}+\frac{p(1-p)}{(2p-1)^2n}\leqslant\frac{1}{4n}+\frac{p(1-p)}{(2p-1)^2n}$，所以预先给定精度要求方差不超过 a，那么一般只要样本量 n 的取值大于 $\left[\frac{1}{4}+\frac{p(1-p)}{(2p-1)^2}\right]/2$ 即可．

3) 对于无放回方式抽样下的沃纳模型．

假设总体容量为 N，采用无放回简单随机抽样方法抽取容量为 n 的样本，则 π_A 估计为：

$\hat{\pi}_A=(\lambda-(1-p))/(2p-1)$，其方差为：

$$D(\hat{\pi}_A)=\frac{N-n}{N-1}\cdot\frac{\pi_A(1-\pi_A)}{n}+\frac{p(1-p)}{n(2p-1)^2}$$

$$\approx(1-f)\cdot\frac{\pi_A(1-\pi_A)}{n}+\frac{p(1-p)}{n(2p-1)^2}$$

2　改进的随机化回答模型

模型的设计如下：制作一套卡片，由三类卡片组成，第一类卡片上写有“如果你具有特性 A，请回答数字 1；如果你具有特性 $\bar{A}$，请回答数字 0”．第二类卡片上写上“请直接回答数字 1”．第三类卡片上写有“请直接加答数字 0”．将这三类卡片各若干混合均匀放入袋中，其比例分别为 p_1，p_2 和 p_3，且 $p_1+p_2+p_3=1$．从总体中有放回地抽取容量为 n 的简单随机样本，由被抽中的人采用有放回的方法随机地从中抽去一张作答，回答结果只有“0”和“1”根据回答“1”的数目 m，可估计出 π_A 的值．

设总体中具有特性 A 的比例为 π_A，样本中回答“1”的概率 λ，则 $\lambda=p(x=1)=\pi_A p_1+p_2$，$\pi_A$ 的无偏估计量为：$\hat{\pi}_A=\frac{\frac{m}{n}-p_2}{p_1}$，其方差为：

$$D(\hat{\pi}_A)=\frac{\pi_A(1-\pi_A}{n}+\frac{p_1(1-p_1)\pi_A}{np_1^2}+\frac{p_2(1-p_2)-2\pi_A p_1 p_2}{np_1^2}$$

又由于 $E(\hat{\pi}_A)=E(\frac{\frac{m}{n}-p_2}{p_1})=\frac{1}{p_1}E(\frac{m}{n})-\frac{p_2}{p_1}=\frac{1}{p_1}(\pi_A p_1+p_2)-\frac{p_2}{p_1}=\pi_A$，即 $\hat{\pi}_A$ 为 π_A 的无偏估计量．此方法克服了沃纳模型中不论抽中几号卡片都必须回答敏感性问题的缺点，消除了被调查者的顾虑；且设计简便，利于实际操作．

关于敏感性问题的研究除了是非属性问题外，还有数量型的问题．本文在这方面还没有来得及讨论，有待于今后进一步加以解决和完善．

【参　考　文　献】

[1] 柯惠新，丁立宏．市场调查与分析[M]．北京：中国统计出版社，2000：132．

[2] 科克伦 W G．抽样技术[M]．张尧庭，译．北京：中国统计出版社，1985：543－607．

[3] KISH L．抽样调查[M]．倪加勋，译．北京：中国统计出版社，1997：597－635．

[4] 冯士雍，倪加勋．抽样调查理论与方法[M]．北京：中国统计出版社，1998：305－324．

[5] 梁小筠．祝大平抽样调查的方法与原理[M]．北京：华东师范大学出版社，1994：184－235．

[6] 赵俊康．统计调查中的抽样设计理论与方法[M]．北京：中国统计出版社，2002：230－290．

[7] 郑　俊．敏感性问题调查方法的探讨[J]．数理统计与管理，1994(1)：29－38．

[8] 孔圣元．敏感问题的问卷调查模型研究[J]．统计研究，1997(3)：34－38．

[9] WARNER S L，RESPOND R．A survey technique for eliminating evasive answers bias，[J]．JASA，1965，60，63－69．

[10] GREENBERG B G．The unrelated questions randomizied response model：Theoretical framework[J]．JASA，1969，64，520－539．