敏感性问题抽样调查方法新设计

朱　宏　　吕　恕

〔内容提要〕对于敏感性问题的抽样调查，沃纳的随机化回答方法精确性不够，西蒙斯的“无关的第二个问题”方法保密性不好。对西蒙斯的方法进行改进，提出敏感性问题抽样调查方法的新设计，既保持了该方法具有较高精确性的优点，又避免了其不能完全为被调查者保密的缺陷，可以使调查获得满意的效果。

一、引　　言

敏感性问题是指高度私人机密性的问题以及大多数人认为不便于在公开场合表态或陈述的问题。在抽样调查中，经常会遇到关于敏感性问题的调查，比如：考试中考生作弊情况的调查；个体户偷漏税情况的调查；私人财产调查；民意测验；性问题调查等等。进行这类问题调查时，如果采用直接调查的方法，很难得到被调查者的合作而获得真实情况；采用匿名答卷方法，情况也许稍有改观，但被调查者会担心留下笔迹被辨认出来，也常常采取拒绝回答或做不真实回答的态度。

关于敏感性问题的调查，不少统计学者研究并提出了很多调查方法，其中最著名的是沃纳（Warner）提出的随机化回答方法①和西蒙斯（Simmons）等提出的“无关的第二个问题”方法②。著名统计学家科克伦（Cochran）③将上述方法做了归纳。近年来，我国的一些统计工作者应用上述方法对一些实际问题进行了调查和统计分析④，取得了一定的成果。

本文针对抽样调查和敏感性问题的特点，从调查结果的精确性和保密性两个方面，对上述方法进行分析和比较，提出改进方案，设计新的敏感性问题抽样调查方法。

二、对现有方法的分析和比较

首先，我们通过实例叙述上述两种主要调查方法。

例如，调查某地区个体户中有过偷税漏税行为者所占的比例。如前所述，若采用直接询问方法：“你曾有过偷税漏税行为吗？”是难以获得真实情况的。

按沃纳（Warner）设计的随机化回答方法进行调查。让被调查者在如下两个问题中随机选择一个做出肯定（是）或否定（否）的回答。

A：你曾有过偷税漏税行为吗？

B：你没有过偷税漏税行为吗?

调查者事先设计好一个随机试验，比如在一个盒子中装有红球和白球，从中任取一球为红球和白球的概率分别是P和（1－P），被调查者回答问题前在盒子中随机地取出一球（球的颜色只有他自己知道），如果取到红球则回答问题A，取到白球则回答问题B。

我们要调查的是曾有过偷税漏税行为个体户所占的比例，不妨将此比例记为π_A，并记调查的回答中“是”的比例为ϕ，则实际调查得到的回答中“是”的比例$\hat{\phi}$是ϕ的一个估计量。ϕ与π_A之间的关系是：

$$\begin{aligned}\phi &= P\cdot\pi_A + (1-P)\cdot(1-\pi_A)\\ &= (2P-1)\cdot\pi_A + (1-P)\end{aligned} \qquad (1)$$

由于P可以事先确定，故当P已知时由（1）式得到π_A的极大似然估计$\hat{\pi}_{AW}$（下标W表示Warner估计）：

$$\hat{\pi}_{AW} = \frac{\hat{\phi}-(1-P)}{(2P-1)} \qquad (P\neq\tfrac{1}{2}) \qquad (2)$$

$\hat{\pi}_{AW}$是π_A的无偏估计量，方差是：

$$D(\hat{\pi}_{AW}) = \frac{\phi(1-\phi)}{n(2P-1)^2} \qquad (3)$$

按西蒙斯（Simmons）提出的“无关的第二个问题”方法进行调查。把向被调查者提出的第二个问题B换成一个与第一个问题无关的、毫无敏感性的问题，如让被调查者随机地回答下述问题：

A：你曾有过偷税漏税行为吗?

B：你是4月份出生的吗?

记被调查者中4月份出生的人所占比例为π_B，则ϕ、π_A、π_B之间的关系是：

$$\phi = P\cdot\pi_A + (1-P)\cdot\pi_B \qquad (4)$$

当π_B已知时，由（4）式得到π_A的极大似然估计$\hat{\pi}_{AS}$(下标S表示Simmons估计)：

$$\hat{\pi}_{AS} = \frac{\hat{\phi}-(1-P)\pi_B}{P} \qquad (5)$$

$\hat{\pi}_{AS}$也是π_A的无偏估计量，方差是：

$$D(\hat{\pi}_{AS}) = \frac{\phi(1-\phi)}{np^2} \qquad (6)$$

下面，我们对上述两种调查方法进行分析和比较。

先分析两种方法的精确性。因为$\hat{\pi}_{AW}$与$\hat{\pi}_{AS}$都是π_A的无偏估计量，故其方差越小者，估计的精确性越好。为比较$D(\hat{\pi}_{AW})$与$D(\hat{\pi}_{AS})$，由（3）、（6）两式，我们比较$(2P-1)^2$与P^2的大小：

$$\begin{aligned}P^2-(2P-1)^2 &= P^2-4P^2+4P-1\\ &= -3P^2+4P-1\end{aligned}$$

$$= (1-P)(3P-1)$$

由于$1-P>0$，故当$3P-1\geqslant 0$即$P\geqslant\frac{1}{3}$时，$P^*\geqslant(2P-1)^2$。

于是由（3）、（6）两式有：

当$P\geqslant\frac{1}{3}$时，$D(\hat{\pi}_{AS})\leqslant D(\hat{\pi}_{AS})$。等号当且仅当$P=\frac{1}{3}$时成立。

在采用沃纳的随机化回答方法进行调查时，P值通常不会取得太小，否则，被调查者将很少有机会对我们关心的敏感性问题提供回答。这样，通过上面的证明及分析可见，通常情况下，西蒙斯方法给出的结果比沃纳方法给出的结果更精确一些。

另一方面，分析两种方法的保密性。仍用前例说明。假定个体户甲，不是4月份出生，而在调查中他的回答为“是”，则可以肯定甲有偷税漏税行为。因为甲如果回答的是问题B，必是否定回答“否”，现在他的回答为“是”，则可以判断他回答的是问题A，并且做了肯定回答。因此，在回答“是”的被调查者中有$100(1-\pi_B)\%$的人能被确定具有不轨行为。可见，西蒙斯方法不能为一部分被调查者保密。

综上所述，沃纳的方法精确性不够，西蒙斯的方法保密性不好。西蒙斯改进沃纳的方法，虽然较大程度地增加了精确性，但却损失了一些保密性。下面我们在西蒙斯方法的基础上，提出改进后的新设计。

三、敏感性问题抽样调查方法的新设计

西蒙斯方法不能为一部分被调查者保密的关键在于他设计的与第一个敏感性问题无关的第二个问题上。由于被调查者在第二个问题上的属性是公开的，这就使得该种调查方法不能对一部分被调查者保密。因此，我们考虑改进第二个问题的设计，既能使调查者掌握被调查者在这个问题上的属性的总的比例π_B，又能使被调查者的属性只有其本人知道。

针对通常的调查，我们给出两种情况下的改进方案：

1、被调查者总数为确定时 当我们的调查是在某个确定范围内进行时，比如在一个学生班内调查学生考试作弊情况，被调查者总数即学生班的人数是确定的，记为N。调查前做N个阄放在一起，其中M个阄有记号，其余的没有，让每个被调查者随机地取一个阄，然后按西蒙斯方法让被调查者随机地回答下述两个问题：

A：你在考试中作过弊吗？

B：你抓到的阄有记号吗？

这样，π_B为已知，等于$\frac{M}{N}$，我们可以根据（5）式得到π_A的估计值。对于具体的某一个被调查者，由于调查者不知道他抓到的阄是否有记号也不知道他回答的究竟是问题A还是问题B，因而不能判断被调查者本人是否有过考试作弊行为。从而，既改进了西蒙斯方法不能为一部分被调查者保密的缺陷，又保持了西蒙斯方法具有较好精确性的优点，能比较准确地估计π_A之值。

2、被调查者总数不确定时 在抽样调查研究中，常常也会遇到被调查者总数不确定的情况，比如一个大范围的民意测验等。下面以在某企业中调查职工们对现任领导班

子是否信任为例说明我们的方法。设计一个装有外形完全相同的小球的盒子，小球上分别写有A、B（蓝色）、B（白色）字母，其所占比例已知分别为P_1、P_2、P_3（$P_1+P_2+P_3=1$）。调查前让每一个被调查者随机抽取一个小球，按小球上的字母对应回答下述二问题之一：

A：你对本厂现任领导班子信任吗？

B：你抽到蓝色字母B的小球吗？

记号ϕ的意义同上，π_A表示职工中对现任领导班子信任的人的比例，则有ϕ与π_A之间的关系式：

$$\phi=P_1\pi_A+\frac{P_2'}{P_2+P_3}(1-P_1)$$

$$=P_1\pi_A+P_2 \qquad (7)$$

于是实际调查中得到的回答“是”的比例$\hat{\phi}$是ϕ的二项式估计量，由（7）式我们可以得到π_A的估计值（亦是π_A的极大似然估计）$\hat{\pi}_A$：

$$\hat{\pi}_A=\frac{\hat{\phi}-P_2}{P_1} \qquad (8)$$

不难看出，由（8）式得到的π_A的估计量$\hat{\pi}_A$与西蒙斯方法的估计量$\hat{\pi}_{AS}$具有相同的功效，但我们的方法能为每个被调查者保密。

四、结束语

本文给出了关于敏感性问题抽样调查的一种新方法，该方法排除了逻辑推断出被调查者情况的可能，因而能为被调查者保守秘密。在这个前提下，利用该方法能得到较为精确的调查结果。需要说明的是，使用这个方法进行调查，必须让被调查者相信方法的保密性，使其密切配合，真实地回答问题。此外，还需说明的是，随机试验设计中，参数P及P_1、P_2、P_3的具体取值，可根据具体问题对保密性和精确性的要求确定。从（6）式可知，当P在区间（0，1）内增加时，估计的精确性提高；从贝叶斯公式可知，P越接近$\frac{1}{2}$，对被调查者的情况越不易猜测出，因而保密性越好。因此，根据需要可以在区间（$\frac{1}{2}$，1）内确定P的一个取值。至于P_1及P_2、P_3的确定，类似进行，不再赘述。

（责任编辑：武震）

参考文献

①Warne, S·L·Jour·Amer·Stat·Assoc·, 66, P63—69, 1965

②Horvitz, D·G·, Shah, B·V·and Simmons, W·R·Proc·Soc·Stat·Sect·Amer·Stat·Assoc·, P65—72, 1967

③Cochran, W·G·Sampling Techniques (Third edition)·John Wiley and Sons, 1977

④《数理统计与管理》1988年，总第33期第31—35页，总第35期，第35—37页