DOI:10.13546/j.cnki.tjyjc.2007.03.012

随机化回答技术在敏感性问题调查中的一种新应用

■陈根龙

最近，在我们开展的一次关于深圳市居民生活水平调查时遇到这样一个问题：我们需要知道深圳居民的银行存款状况，所以拟了以下一个调查问题："您现在的银行存款是（ ）？以下有六个备选项，(1)3 万以下；(2)3-10 万；(3)10-30 万；(4)30-50 万；(5)50-80 万；(6)80 万以上。"很显然该问题属于强敏感性问题，因此被调查者极有可能出于保护个人隐私而拒绝回答或作虚假回答。解决敏感性调查问题比较科学和常用的方法就是随机化回答技术。随机化回答技术最早由沃纳（Warner）于 1965 提出，之后就有很多人提出改进的随机化回答模型，如双无关问题模型、隐含的随机化回答模型、随机截尾的 Warner 与 Simmons 模型等。从二项选择到多项选择、从定性到定量，随机化回答技术在解决敏感性问题调查方面日渐完善。理论的可行性和操作的简便性是设计随机化回答模型的基本要求也是设计的难点。目前，很多的模型在理论的可行性方面无可非议，但其过于繁琐的操作规则使模型的价值大打折扣。本文旨在通过对西蒙斯随机化回答模型作一定延伸，从而获得解决多项选择的敏感性调查问题的一种易于操作的方法。

一、随机化回答技术及其基本模型介绍

随机化回答模型有两种模型：沃纳模型和西蒙斯模型。

（一）沃纳模型

沃纳（Warner）于 1965 年首先提出随机化回答模型。这种模型设计的思想是：为了调查某个敏感性问题，同时列出两个对立问题，被调查者随机从两个问题中抽取一个，但并不告诉调查人员抽到的是哪一个问题，然后根据自己的实际情况回答，调查人员只记录回答的结果。由于调查人员不知道被调查者回答的是哪个问题，从而实现了被调查者提供真实情况的自我保护。具体模型的建立和实施本文不再赘述。

（二）西蒙斯模型

1.模型的改进之处

与沃纳模型不同的是，西蒙斯模型所提出的两个问题是无关的，其中第二个问题是与所调查的敏感性问题完全不相关的非敏感性问题；而且，在设计卡片时可以使 P=1/2。

2.模型的设计及参数估计

使用外形一样的两套卡片 1 号卡片和 2 号卡片，1 号卡片上写上敏感性问题"你具有特征 A 吗？"2 号卡片上写有完全不相关的非敏感性问题"你有特征 B 吗？"将一定数量的 1 号卡片和 2 号卡片按预定比例混合后放入一盒子中。调查时被调查者只需从盒子中任意抽取一张卡片，根据卡片上的问题做出真实的回答，回答完毕后将卡片放回盒子，摇匀后由下一个被调查者继续。在此过程中调查人员只能获得"是"或"不是"的回答，而不知道被调查者回答的是哪个问题。

设抽样方式是简单随机抽样，1 号卡片的比例为 P（可以为 1/2），2 号卡片的比例为 1-P，X 为总体中有特征 A 人数的比例，Y 为总体中具有特征 B 人数的比例（这是已知的）。N 为样本容量，M 为回答"是"的人数，则调查中回答"是"的人数的比例为：

$$\frac{M}{N}=PX+(1-P)Y$$

X 的极大似然估计为：

$$\hat{X}=\frac{\frac{M}{N}-(1-P)Y}{P}$$

其方差的无偏估计为：

$$\widehat{Var}(\hat{X})=\frac{1}{N\cdot P^2}\left(\frac{M}{N}\right)\left(1-\frac{M}{N}\right)$$

二、对西蒙斯模型进行延伸

（一）问题的提出

无论是沃纳模型还是西蒙斯模型，它们都有一个明显的缺陷，即被调查者只能回答"是"或"不是"，这就极大地限制了该模型的适用范围。对于备选项超过两项的敏感性调查问题，沃纳模型和西蒙斯模型都无能为力。如本文开头所提到的调查居民的存款问题，其备选项有 6 项之多。我们还能举出更多的例子，如在市场调查中我们需要了解消费者的月收入状况，调查者会将月收入划分成若干区间，形成多个备选项让被调查者选择；对单位的领导进行评价，有"满意"、"一般"、"不满意"三个备选项等等。

（二）构模思路

保留西蒙斯模型中要被调查者回答的两类问题，非敏感性问题仍作为保护被调查者的"道具"。使非敏感性问题和敏感性问题有相同数目的备选项，且要求非敏感问题的各备选项出现的概率是已知的。被调查者随机抽取调查问题并选择一个真实的备选项。

（三）延伸后的模型

预先告诉被调查者两个问题及每个问题的多个备选项，问题 A 是调查者所要调查的敏感性问题，问题 B 是与调查的敏感性问题完全无关的非敏感性问题。依然使用外形一样的两套卡片 1 号卡片和 2 号卡片，1 号卡片上写着"请回答问题 A"，2 号卡片上写着"请回答问题 B"。将一定数量的 1 号卡片和 2 号卡片按预定比例混合后放入盒子中，调查时被调查者只须从盒子中任意抽取一张卡片，根据卡片上的提示做出真实的回答，回答完毕后将卡片放回盒子，摇匀后由下一个被调查者抽取并回答。此时调查人员获得的将不再是"是"或"不是"的回答，而是"1"、"2"、"3"、"4"……或者"a"、"b"、"c"、"d"……。同样，在延伸后的模型中，调查者也不知道被调查者回答的是哪个问题。

现在,设抽样方式是随机抽样,1 号卡片的比例为 P 可以为 1/2),2 号卡片的比例为 1- P,a_1、a_2、a_3、a_4……分别是敏感性问题 A 的多个备选项,b_1、b_2、b_3、b_4……分别是非敏感性问题 B 的多个备选项。敏感性问题 A 和非敏感性问题 B 有同样多个的备选项,且(1)、(2)、(3)、(4)……为它们的序号 调查时被调查者只需报出其序号即可)。X_1、X_2、X_3、X_4……分别是总体中具有特征 a_1、a_2、a_3、a_4……人数的比例,Y_1、Y_2、Y_3、Y_4……分别是总体中具有特征 b_1、b_2、b_3、b_4……人数的比例 这是已知的)。N 为样本容量,M_1、M_2、M_3、M_4……分别是回答为 "1"、"2"、"3"、"4"……的人数。由全概率公式,调查中回答为 "1" 的人数的比例为:

$$\frac{M_1}{N}=P\cdot X_1+(1-P)\cdot Y_1$$

同理可得到回答为 "2"、"3"、"4"……的人数的比例为:

$$\frac{M_2}{N}=PX_2+(1-P)Y_2$$

$$\frac{M_3}{N}=PX_3+(1-P)Y_3$$

$$\frac{M_4}{N}=PX_4+(1-P)Y_4$$

……

X_1 的极大似然估计为:

$$\hat{X}_1=\frac{\frac{M_1}{N}-(1-P)Y_1}{P}$$

同理,X_2、X_3、X_4……的极大似然估计为:

$$\hat{X}_2=\frac{\frac{M_2}{N}-(1-P)Y_2}{P}$$

$$\hat{X}_3=\frac{\frac{M_3}{N}-(1-P)Y_3}{P}$$

$$\hat{X}_4=\frac{\frac{M_4}{N}-(1-P)Y_4}{P}$$

……

其方差的无偏估计为:

$$\widehat{Var}(\hat{X}_1)=\frac{1}{NP^2}\left(\frac{M_1}{N}\right)\left(1-\frac{M_1}{N}\right)$$

$$\widehat{Var}(\hat{X}_2)=\frac{1}{NP^2}\left(\frac{M_2}{N}\right)\left(1-\frac{M_2}{N}\right)$$

$$\widehat{Var}(\hat{X}_3)=\frac{1}{NP^2}\left(\frac{M_3}{N}\right)\left(1-\frac{M_3}{N}\right)$$

$$\widehat{Var}(\hat{X}_4)=\frac{1}{NP^2}\left(\frac{M_4}{N}\right)\left(1-\frac{M_4}{N}\right)$$

……

(四)案例分析

以本文开头所提出的居民存款的调查为例,调查问题是:"您现在的银行存款是( )?有六个选择,(1)3 万以下,(2)3-10 万,(3)10-30 万,(4)30-50 万,(5)50-80 万,(6)80 万以上。"采用延伸后的西蒙斯随机化回答模型。设 "您现在的银行存款是?"为问题 A,同时设计一个完全不相关的非敏感性问题 B—— "您是哪月出生的?",问题 B 也有六个备选项:(1)1-2 月,(2)3-4 月,(3)5-6 月,(4)7-8 月,(5)9-10 月,(6)11-12 月。所用的随机化装置是一个装有两种卡片的盒子,1 号卡片上写着 "请回答问题 A",2 号卡片上写着 "请回答问题 B",两种卡片共 20 张各占一半。通过配额分层抽样,在深圳抽取 500 户家庭。调查结果是 为计算方便,已对数据进行了处理):回答为 "1" 的有 150 户,回答为 "2" 的有 100 户,回答为 "3" 的有 100 户,回答为 "4" 的有 50 户,回答为 "5" 的有 50 户,回答为 "6" 的 50 户。以 95%的置信度对居民的银行存款额在各个区间所占的比例进行区间估计。

由题意得:

N=500,P=1/2;

M_1=150,M_2=100,M_3=100,M_4=50,M_5=50,M_6=50;

$Y_1=Y_2=Y_3=Y_4=Y_5=Y_6=1/6$。

由延伸后西蒙斯随机化回答模型,银行存款额属于(1)——3 万以下的户数的比例为:

$$\hat{X}_1=\frac{\frac{M_1}{N}-(1-P)Y_1}{P}$$

$$=\frac{\frac{150}{500}-\left(1-\frac{1}{2}\right)\frac{1}{6}}{\frac{1}{2}}=\frac{13}{30}=0.4333$$

X_1 的方差估计量为

$$=\frac{1}{NP^2}\left(\frac{M_1}{N}\right)\left(1-\frac{M_1}{N}\right)$$

$$=\frac{1}{500\left(\frac{1}{2}\right)^2}\cdot\frac{150}{500}\left(1-\frac{150}{500}\right)$$

=0.00168

置信区间为

$$(\hat{X}_1)\pm\sqrt{(\hat{X}_1)}=0.4333\pm1.96\times\sqrt{0.00168}=0.4333\pm0.0804$$

即 (0.3529,0.5137)。

同理可以得到 X_2,X_3,X_4,X_5,X_6 的估计值:

$\hat{X}_2$=0.2333,$\hat{X}_3$=0.2333,$\hat{X}_4$=0.0333

$\hat{X}_5$=0.0333,$\hat{X}_6$=0.0333

X_2,X_3,X_4,X_5,X_6 的方差估计量为

$\widehat{Var}(\hat{X}_2)$=0.00128,$\widehat{Var}(\hat{X}_3)$=0.00128

$\widehat{Var}(\hat{X}_4)$=0.00027,$\widehat{Var}(\hat{X}_5)$=0.00027

$\widehat{Var}(\hat{X}_6)$=0.00027

相应的在 95%的置信度下的置信区间分别为

(0.1631,0.3035)、(0.1631,0.3035)、(0.0011,0.0654)、(0.0011,0.0654)、(0.0011,0.0654)。

由此可以得出银行存款数额在前述 6 个档次的分别约占:33.33%、23.33%、23.33%、3.33%、3.33%、3.33%;有 95%的把握程度说明银行存款在 3 万以下的居民占 35.29%—51.37%等结论。

三、实际应用过程中的建议

(一)要明确延伸后的西蒙斯模型的适用范围

延伸后的西蒙斯模型适用于多项选择的敏感性调查问题,若是二项选择可直接套用沃纳模型或西蒙斯模型。

(二)做好调查问题的设计

在设计调查问题时很重要的一步就是选择一个合适的非敏感性问题。这个非敏感性问题要能够分解出和要调查的敏感性问题有相同数量的备选项,且每个备选项出现的概率是可以确定的。例如,本文的案例中以被调查者的出生月份为相应的非敏感性问题。应用时可根据实际情况找出恰当的非敏感性问题。

(三)做好对调查人员的培训

理解该模型具有一定的难度,调查人员只有对调查方法有深刻的了解,才能向被调查者解释清楚该方法是如何保护被调查者的个人隐私的。还有诸如问卷的记录、随机化装置 例如卡片)的使用都应在对调查人员的培训中介绍清楚。

(四)改善细节,提高效率

在调查中使用随机化回答技术的缺点之一就是比较耗时,时间主要消耗在对被调查者的解释和随机化装置的使用上。因此,在调查时应力求使用最简短的语言向被调查者传递 "游戏规则",根据调查的实际情况设计出便于携带和使用的随机化装置。这些都是调查过程中的细节,只要对这些细节进行改善,调查的效率就能提高。

(作者单位/深圳大学)

(责任编辑/李友平)