对敏感性问题调查之随机化回答技术的思考

□文/ 李 炜

（暨南大学经济学院，广东广州 510500）

〔摘 要〕 抽样调查在人们生活、工作的方方面面中起着越发重要的作用，但数据的无回答现象降低了数据的质量，使调查的精度大打折扣。针对于此，本文介绍了在敏感性问题调查中使用频率颇高的沃纳模型和西蒙斯模型，阐述了随机化回答模型的基本原理，并比较了两种模型，指出了沃纳模型应用的局限性及西蒙斯模型在此基础上进行的改进。最后提出了在我国应用随机化回答技术应注意的几个问题。

〔关键词〕 敏感性问题；随机化回答技术；沃纳模型；西蒙斯模型

〔中图分类号〕 F224.7 〔文献标识码〕A 〔文章编号〕1004－1516(2000)12－0027－02

在日常生活中，抽样调查的应用是十分广泛的。人们使用抽样调查方法的目的是以样本推断总体，获得对总体的一定认识。但如在抽取到的样本中，数据产生无回答（即无响应）情况，则必会减少有效样本量，使数据失真或不准，降低了数据的质量，违背了调查的初衷。

其中，引起抽样的数据产生无回答现象的原因很多，就被调查对象而言，经常会发生因调查内容涉及敏感性问题而导致被调查对象有意识不回答的现象。何谓敏感性问题？敏感性问题是指所调查的内容涉及私人机密而不愿或不便于公开表态或陈述的问题，例如在考试中的学生作弊现象，工商业者的偷税漏税问题，司机的违例驾驶等问题就是一般人不愿正面回答的敏感性问题。可以预见，在调查这一类问题时阻力很大，结果要么是引起被调查者的反感——他们直接拒绝回答，要么是被调查者给出并不真实的回答。有鉴于此，需要有一种经过特别设计的回答技术去消除被调查者的种种可能的顾虑，而本文所探讨的正是针对于此的随机化回答技术。随机化回答技术是专门针对敏感性问题的调查而设计的，它的功能在于使被调查者能对所调查的问题采取随机回答的方式，避免在没有任何“掩护”的情形下直接回答敏感性问题。这种技术更多地考虑了人的心理因素，减弱了被调查者的抵御心理，争取了他们最大限度的合作可能。总的来说，随机化回答技术都是试图通过随机回答的方式，使被调查者的担忧、顾虑心理得以解除，降低回答偏差，并相应提高回答率。下面，我们就通过应用这种技术的两种模型具体了解它的原理。

一、沃纳随机化回答模型。

随机化回答模型最早是由沃纳于1965年首先提出的。它的设计是向被调查者显示两个与敏感性问题（具有特征A）有关，但完全对立的问题：一个问题是“你具有特征A吗？”，另一个问题是“你具有特征$\bar{A}$吗？”（$\bar{A}$表示不具有特征A）。对两个问题的答案都只有“肯定（是）”与“否定（否）”两种。关键是设计一种随机化装置，使被调查者以概率P回答第一个问题而以概率（1－P）回答第二个问题，重要的是只有被被查者本人知道他究竟是回答哪一个问题，而调查员却并不知道，因此便能取得被调查者的合作。

例如某出租汽车公司欲估计本公司司机在工作期间发生违例驾驶行为的比例π，则随机抽取了n个司机进行调查，对每个司机显示了两个问题——问题1：你曾有违例驾驶行为，对吗？；问题2：你不曾有违例驾驶行为，对吗？然后再将一个事先设计好的密闭容器交给司机，该容器中装有两种颜色（例如红与白）但大小、形状与重量完全一样的球，红球与白球的个数比例是P/（1－P）。令被调查的司机在容器中随机的摸一个球（注意抽球时不向任何人显示，即只有被调查的司机本人才知道他摸到的是什么颜色的球），若抽到红球则如实回答问题1，若抽到白球则如实回答问题2。由于两个问题的答案均只有“是”与“否”两个答案，因此当司机回答时，调查员根据答案并不能断定该司机回答的是哪一个问题，由此就确保了被调查司机的隐私安全性。

容易了解，有两种情况都会导致回答“是”：一种是抽到红球，而被调查的司机确曾违例驾驶；另一种是抽到白球，而被调查的司机并不曾违例驾驶。抽到红球的概率是P，抽到白球的概率是（1－P）；不管是抽到红球还是抽到白球，曾有违例驾驶行为的概率都是π，不同的是回答问题1答案为“是”的概率为π，而回答问题2答案是“是”的概率为（1－π），所以对任意一个被调查的司机，他的回答为“是”的概率：$Pr(是)=P\pi+(1-P)(1-\pi)$，我们设n个被调查的司机中共有m个回答“是”，所以Pr(是)的一个合理估计是m/n，于是π的一个估计$\hat{\pi}$满足：$m/n=P\times\hat{\pi}+(1-P)\times(1-\hat{\pi})$。由此得：$\hat{\pi}=1/(2p-1)\times[m/n-(1-P)]$，（其中$P\neq 1/2$）。沃纳指出$\hat{\pi}$是π的极大似然估计，且是无偏的，它的方差为：

$$V(\hat{\pi})=\pi\times(1-\pi)/n+P\times(1-P)/[n(2p-1)^2]。$$

上式表示，π的方差由两个部分组成：前一项是直接回答敏感性问题的方差（若忽略有限总体修正系数），后一项是由于采用随机化回答而引起的方差增加。据计算，除非π接近1/2，$P>0.85$，否则第二项方差常大于第一项，而且经常是大许多的。事实上，随机化回答方法是用牺牲精度的代价去换取被调查者的真实回答，因此，如果调查问题的敏感度不太高，采用随机化回答技术就没有什么实际意义了。

在沃纳模型中，问题1与问题2是相关的，它们是同一敏感问题的两

〔收稿日期〕 2000－10－08

个方面，这仍然不能根本打消被调查者的顾虑，因为此时被调查者认为他们回答的还是同一敏感问题，所选择的只是回答这个敏感问题的哪一个方面而已，还是担心易被别人猜出自己回答的是哪一个问题，不能保护自己的隐私。而且，在该模型中，有一个限制条件，$P\neq 1/2$。这就使该模型的开展有了束缚性。关于这一点，在下面的西蒙斯模型中将得到改善。

二、西蒙斯随机化回答模型。

从沃纳模型可看到，该模型对被调查者的隐私保密程度还不够高或不彻底，有鉴于此，西蒙斯模型作了以下改进：首先，所提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，因此西蒙斯模型也称为无关问题的随机化回答模型；其次，沃纳模型中有一限制性条件，$P\neq 1/2$，而西蒙斯模型中没有这个限定，P 可等于 1/2，事实上，当 $P=1/2$，意味着被调查者抽中两种球的可能性相同，那么会有更大的掩护性，令被调查者感到更安全。

例如，两个问题为——问题 1：你曾有违例驾驶行为，对吗？问题 2：你的生肖属牛，对吗？在此，我们的目的是要估计曾有违例驾驶行为的司机的比例 π_A，而问题 2 的比例 π_B 在设计时要求已知（此例中 $\pi_B=1/12$）。仍应用在上文中提到的密闭容器装置，因此两个问题在随机化回答中出现的比例仍为 $P/(1-P)$，其中 P 已知。而调查结果仍是在 n 个被调查的司机中，有 m 个人回答“是”。同理，对每个被调查者来说，他回答“是”的概率为：$Pr(是)=P\pi_A+(1-P)\pi_B$。因此 π_A 的估计 $\hat{\pi}_A$ 满足：$m/n=P\times\hat{\pi}_A+(1-P)\pi_B$。由此得，$\hat{\pi}_A=1/P\times[m/n-(1-P)\pi_B]$。$\hat{\pi}_A$ 也是无偏的，且它的方差的一个无偏估计为：

$$V\hat{\pi}_A=1/(n-1)P^2\times(m/n)\times(1-m/n)$$

在此例中，如 $P=1/2$，共有 600 个司机参加调查（即 $n=600$），其中有 127 个回答“是”（即 $m=127$），则我们可估计司机发生违例驾驶行为的比例：

$$\hat{\pi}_A=1/0.5[127/600-(1-0.5)\times 1/12]=34\%$$

$$V(\hat{\pi}_A)=1/(599\times 0.25)\times(127/600)\times(1-127/600)=0.001114,$$

$$S(\hat{\pi}_A)=0.0034。$$

因此可估计司机发生违例驾驶行为的比例的 90% 的置信区间为：（$0.34-1.64\times 0.0334$，$0.34+1.64\times 0.0334$）即（28.52%，39.48%）。

正如上提到的，西蒙斯模型中多了一个无关问题，增强了保护作用，更彻底的打消了被调查者的顾虑：且应用时，P 并无任何限制条件，在推广应用上更有实际意义。因此笔者认为，西蒙斯模型在理论上比沃纳模型更完善。

三、关于应用随机化回答技术中应注意的几个方面。

1、如何才能将该技术实施得更好？——应从调查员与被调查者两方面分别着手。采用该技术的主要目的是为了消除被调查者的顾虑使其愿意配合，并对所抽到的问题给予正确的回答。我们要知道，随机化回答模型不仅是一个理论问题，而且还是一个实践问题。对于采用的模型，调查员应能充分掌握操作的整个过程、步骤，这就要求搞好对调查员的培训工作，使他能对被调查者进行有说服力的必要的解释和说明；同对还要使被调查者充分理解这种方法的特点，让他知道他所回答的究竟是哪一个问题，别人是不知道的，使他们能打消不必要的顾虑。要在这两方面达到较好的效果，对调查员及被调查对象的素质要求就会很高，这一点同现实也很吻合——据统计在较低素质的被调查者群体中，随机化回答技术的效果不明显，反之在较高素质的被调查者群体中，随机化回答技术能发挥一定作用。而从我国的实际国情出发，虽然当前调查员及被调查者的总体素质尚不算高，但是有逐渐提高的趋势，因此随机化回答技术在我国有着开阔、乐观的前景。

2、在调查中所提的问题必须简单明了，防止产主歧义，这也是一般调查中的基本要求。比如说问题“你晚上经常夜归吗？”，就必须把“夜归”的时间跨度事先解释清楚。

3、随机化回答技术有其特定的运用范围——敏感性问题，调查问题的敏感性越强，被调查者对回答的保密性要求越高，随机化回答的效果就越突出。而敏感性不强的问题，随机化回答的效果并不太好。如在沃纳模型中曾提到，随机化回答方法是用牺牲一定精度的代价去换取被调查者的真实回答，此时如调查的是非敏感性问题，则不适于采用随机化回答技术。

4、最后必须强调的是，关于敏感性问题的调查，应更注意对人们心理情况的研究，如在调查中由于人的逆反心理作祟，人们往往会讲反话，这时就收不到效果。这启示我们可在统计理论中结合心理学、社会学进行研究，力求产生更完美的随机化回答技术模型。

（责任编辑：薛金龙）