

对敏感性问题调查的随机化回答方法

曹 健

在社会学调查中,常常有些涉及到个人秘密的问题。例如在生育力调查中,被调查者对于象“你过去有过流产吗?”这类问题往往感到窘迫而不愿意回答或者不真实回答,这就给调查带来了困难,产生调查偏差。如何消除由于逃避回答敏感性问题而带来的调查偏差呢?近二十年来,已经发展了一种新的方法,它采用一种随机化的装置,不用被调查者向调查者泄露其在所提问题上属于什么情况,就可以估计出所有被调查者中属于某种情况的比例。因为采用了一种既能鼓励作出真实回答同时又能保守个人秘密的随机化装置(如一支旋转的箭或一个装有不同颜色小球的匣子),所以,就把这种调查技术称为“随机化回答”方法。目前,在美国等许多国家,随机化回答方法得到了广泛应用,曾被用来估计私生子数量、人工流产数、吸毒者、与犯罪组织有联系的人的比例。我国台湾在1979年开展的妇女生育力调查中也采用了这种方法,效果较好。

随机化回答方法首先是由美国的沃纳于1965年在一篇论文中提出来的。

假若我们要调查在一个由几个被调查者组成的样本中属于A类情况人所占的比例 R_A 。沃纳建议,向被调查者提出两个对立的陈述,如“我是A类中的成员”和“我不是A类的成员”,每个陈述要求回答“是”或“不是”。调查者通过一个随机化的装置分别以概率(可能性) P 和 $(1-P)$ 向被调查者提出上面两个陈述,这时调查者并不知道任何一个被调查者已回答了哪一个问题,他所知道的仅仅是答复为“是”的比例,记为 ϕ 。如果问题是回答得真实的,则 ϕ 与被调查者中 R_A 之间的关系是:

$$\phi = P \cdot R_A + (1 - P)(1 - R_A) = (2P - 1)R_A + (1 - P)$$

容易推得几个被调查者中属于A类的比例 R_A 的估计式是:

$$R_A = \frac{[\phi - (1 - P)]}{(2P - 1)} \quad (P \neq 1/2)$$

俗以及社会的精神文明和体育运动的关系,还有学校与体育运动的关系等;个人与社会的关系,如个人社会化与体育运动、体育教师和教练员的社会地位以及与学生和运动员的关系等);研究的重点应当是探讨体育运动的社会性质、地位和功能等问题;研究的方法基本上需采用社会学的方法,同时应具有吸收和使用新的研究方法的特点,尤其是自然科学的一些方法,如数学方法、计量方法和电子计算机方法的应用等,重点要采用

统计分析的方法。

体育运动社会学是一门新兴的综合边缘学科,但它在体育运动事业的发展过程中所体现的作用已经表明了是一门富有生命力的科学,它所具有的功能和特点决定这是一门永恒的学科,只要体育运动在整个人类社会中存在,只要人类社会的有效生存离不开体育运动,体育运动社会学的研究就不会中断,而且在整个研究过程中,促使体育运动和社会有效地向一体化方向发展。

这个公式就是沃纳在其论文中提出来的。显然，他的方法比一个直接的敏感性问题引起的误差要来得小。

受沃纳方法的启发，如果我们对第二个陈述作改进，使之毫无敏感性，与第一个陈述无关，例如，“我是在五月出生的”，第一个陈述不变，那么可以预料，被调查者的合作会有所改进。如果所有的被调查者都作出真实的回答，则现在总体中作出“是”的回答比例是：

$$\phi = P \cdot R_A + (1 - P) \cdot R_u$$

这里 R_u 是被调查者中在五月份出生的人所占的比例， R_u 往往是已知的或者容易知道的。则 R_A 的估计式就表示为：

$$R_A = \frac{[\phi - (1 - P)R_u]}{P}$$

统计学可以证明，这种方法作出的 R_A 的方差要比第一种方法来小。一个直观解释是，后面的 R_A 估计式中利用了被调查者的已知信息 (R_u)。

下面，我们举一个例子来具体说明如何应用这两种随机化回答方法（例中的数字是笔者假设的）。

假定在一次生育力调查中，我们抽取了一个由 500 名被调查者组成的随机样本，欲调查他们采取避孕措施的比例 R_A 。用第一种方法，向他们提出两个对立的陈述：“我使用避孕器具”和“我不使用避孕器具”。使用的随机化装置是一个内装有许多黑、白两色小球的匣子，两球的比例各为 60%（记为 P ）和 40%（记为 $1-P$ ）。调查要求被调查者随机地从匣子中摸取一球，摸球结果只是被调查者本人知道，当摸到的是黑色球时，对第一个问题作出答复：“是”或“不是”，当为白球时，回答第二个问题。由于匣中小球是均匀拌和的，由概率论知识知道，黑、白小球各占的比例 P 和 $1-P$ 就分别代表了两个陈述被提出的概率。当全部答复中“是”为 275 次时，调查者就可以运用前面的第一个公式计算出被调查者中使用避孕器具的比例 R_A 。计算步骤如下：

因为 $\phi = 275/500 = 0.55$ ， $P = 0.60$ ，

$$\text{所以 } R_A = \frac{[\phi - (1 - P)]}{(2P - 1)} = \frac{0.55 - (1 - 0.60)}{(2 \times 0.60 - 1)} = \frac{0.15}{0.2} = 0.75 = 75\%$$

我们也可以用第二种方法来进行调查。现在第二个陈述改为：“我是南方人”，第一个陈述不变。显然，改动的第二个陈述与第一个陈述是毫无关系的，被调查者是否属于南方人事先可以知道（如从行政记录上得到），这样一来，由于现在的第二个陈述毫无敏感性，则有可能使被调查者乐于回答。接下来的具体做法同上， R_A 的公式前面也已给出了，这里不多谈了。

上面介绍的两种随机化回答方法是基本的，自沃纳提出随机化回答方法以来，引起了许多社会学家和统计学家的兴趣，他们从理论与方法上讨论了处理敏感问题的随机化回答模型，提出了许多方法。例如，使用两个或两个以上不相关的问题； R_u 也是未知的；以及多个敏感性问题等等方法，并且还证明了按某种随机化回答模型的理论结构，具有许多优良的统计性质（如无偏性、方差最小）。但是，必须指出，所有这些方法只不过是本文所介绍的两种最基本方法的变形而已。