

# 敏感问题的调查技术

章扬熙

在公共卫生监测中,有时需要了解一些隐私问题(包括心理、行为等方面)。当询问这类问题时,对方往往不愿意回答,这类问题就叫做敏感问题。对于敏感问题,被调查者往往拒绝回答或者编造假话来应付,这就会导致无应答率的增高和正确应答率的下降。为了解决这个问题,而产生了敏感问题的调查技术,即随机应答技术(randomized response technique, RRT)。

一、RRT 方法的相关问题模式: RRT 方法的原理是让被调查者根据随机的原则,对一个敏感问题提出的两个相关联却又相反的描述句(有敏感问题与没有敏感问题)中的一个做出回答,虽然调查者不知道被调查者回答的是哪个描述句,但已知抽到第①句的概率  $P$  和抽到第②句的概率  $1 - P$ , 并知道在抽到第①句中回答“是”的人数比例为  $\pi$ , 从而也知道在抽到第②句中回答“是”的人数比例  $1 - \pi$  (因为抽到第①名回答“不是”的人数比例为  $1 - \pi$ , 与抽到第②句回答“是”的人数比例应该相同), 再从调查中可知道总的回答“是”的比例  $\lambda$ , 依据概率的有关法则可以计算出在被调查者中, 具有该敏感问题特征的人所占的比例, 从而达到既打消了被调查者的顾虑, 又了解到敏感问题客观真相的目的。

比如, 想了解婚外性行为情况, 则第①句为“我有过婚外性行为”, 第②句为“我没有过婚外性行为”。然后让被调查者从一个装有红球和白球的袋子里摸出一个球(其中红球的比例为  $P$ , 白球的比例为  $1 - P$ ), 摸到红球则回答第①句, 摸到白球则回答第②句, 被调查者抽球与答题情况是在无人条件下不记名进行的, 这样可完全解除被调查者的顾虑, 据实回答。调查结束后, 调查员只能统计出有多少人回答“是”, 有多少人回答“不是”, 这些是有多少是回

答第①句, 有多少是回答第②句不清楚, 更不知道这些“是”是谁回答的。可计算出总的回答“是”的比例  $\lambda$  值。设在抽到第①句中, 回答“是”的人所占比例为  $\pi$ , 则在抽到第②句中, 回答“是”的人所占比例应为  $1 - \pi$ , 根据概率法则可得到

$$\lambda = P\pi + (1 - P)(1 - \pi) \quad (1)$$

$$\hat{\pi} = \frac{\lambda - (1 - P)}{2P - 1} \quad (2)$$

$\pi$  的方差为

$$\begin{aligned} \text{Var}(\hat{\pi}) &= \frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{P(1 - P)}{n(2P - 1)^2} \quad \text{或} \\ &= \frac{\lambda(1 - \lambda)}{n(2P - 1)^2} \end{aligned} \quad (3)$$

总体  $\pi$  的 95% 可信区间为

$$\pi \pm 1.96 \sqrt{\text{Var}(\pi)} \quad (4)$$

[例 1] 对 600 名已婚妇女进行调查, 想了解婚外性行为, 应用 RRT 法进行调查, 设红球比例  $P = 0.8$ , 白球比例  $1 - P = 0.2$ , 通过调查得到: 总的回答“是”的比例  $\lambda = 0.4$ , 求有婚外性行为者占百分之多少。

应用公式(2)、(3)、(4), 求得

$$\hat{\pi} = \frac{0.4 - (1 - 0.8)}{2 \times 0.8 - 1} = 0.33$$

$$\begin{aligned} \text{Var}(\hat{\pi}) &= \frac{0.33 \times (1 - 0.33)}{600} + \\ &\frac{0.8 \times (1 - 0.8)}{600 \times (2 \times 0.8 - 1)^2} = 0.0011 \end{aligned}$$

总体  $\pi$  的 95% 可信区间为

$$0.33 \pm 1.96 \sqrt{0.0011} = 0.26 \sim 0.40$$

即有 33% 的已婚妇女有婚外性行为, 总体已婚妇女婚外性行为发生率的 95% 可信区间为 26% ~ 40%。

应当指出, 应用 RRT 的相关问题模式不可以设  $P = 0.5$ , 从公式(2)可知  $P = 0.5, \pi = \infty$ , 通常设  $P = 0.8$  或  $0.9$ 。同时, 从公式(3)可以看出, 应用 RRT 方法所得的  $\text{Var}(\pi)$  比用直接询问法要大, 多了  $P(1 - P)/n(2P - 1)^2$  这个部分, 方差大则可信区间

作者单位: 中国医学科学院皮肤病研究所 南京 210042

大,说明效能较低。为了弥补这个缺点,可让被调查重复回答2~3个回合,使例数n增大2~3倍,从而减小Var( $\pi$ )的数值。

[例2] 对例1重复2次与重复3次,计算Var( $\pi$ )及总体 $\pi$ 的95%可信区间。若重复2次,n=1 200,将有关值分别代入公式(3)、(4)得

$$\text{Var}(\hat{\pi}) = \frac{0.33 \times (1 - 0.33)}{1200} + \frac{0.8 \times (1 - 0.8)}{1200 \times (2 \times 0.8 - 1)^2} = 0.00055$$

总体 $\pi$ 的95%可信区间为

$$0.33 \pm 1.96 \sqrt{0.00055} = 0.28 \sim 0.38$$

若重复3次,n=1 800,将有关值分别代入公式(3)、(4)得

$$\text{Var}(\hat{\pi}) = \frac{0.33 \times (1 - 0.33)}{1800} + \frac{0.8 \times (1 - 0.8)}{1800 \times (2 \times 0.8 - 1)^2} = 0.00037$$

总体 $\pi$ 的95%可信区间为

$$0.33 \pm 1.96 \sqrt{0.00037} = 0.29 \sim 0.37$$

从例2可见,重复2~3次,Var( $\pi$ )值不断缩小,95%可信区间也相应地缩小了范围,从而提高了对总体 $\pi$ 区间估计的精确度。

**二、RRT方法的不相关问题模式:**RRT方法的不相关问题模式是提出两个不相关的问题,第①题是敏感问题,第②题是不敏感问题。具体的做法与相关问题模式相仿。

设P为抽到第①题的概率,1-P为抽到第②题的概率; $\pi_1$ 为在抽到第①题的人中,回答“是”的比例, $\pi_2$ 为在抽到第②题的人中,回答“是”的比例; $\lambda$ 为总的回答“是”的比例。则

$$\lambda = P\pi_1 + (1 - P)\pi_2 \quad (5)$$

或 
$$\hat{\pi}_1 = \frac{\lambda - (1 - P)\pi_2}{P} \quad (6)$$

$$\text{Var}(\hat{\pi}_1) = \frac{\lambda(1 - \lambda)}{nP^2} \quad (7)$$

式中 $\pi_1$ 为研究的变量, $\pi_2$ 有时是已知的,P是设定的,从一个样本调查可以得到 $\lambda$ ,则 $\pi_1$ 可求,总体 $\pi_1$ 的95%可信区间也可用公式(4)求得。有时, $\pi_2$ 是未知的,这时则要做两个P值不同的样本调查,得到两个 $\lambda$ ,即 $\lambda_1$ 和 $\lambda_2$ ,建立下列联立方程式

$$\begin{cases} \lambda_1 = P_1\pi_1 + (1 - P_1)\pi_2 \\ \lambda_2 = P_2\pi_1 + (1 - P_2)\pi_2 \end{cases} \quad (8)$$

式中 $\lambda_1, \lambda_2, P_1, P_2$ 已知, $P_1 \neq P_2 \neq 0.5$ ,求出 $\pi_1, \pi_2$ 。

[例3] 对500名未婚18岁以上女性进行调查,想了解其中有过性关系的情况,提出两个问题:①你是否有过性关系?②你是否是龙年出生的?调查用RRT方法进行,结果总的回答“是”的比例 $\lambda = 0.3$ ,红球比例 $P = 0.8$ ,白球比例 $1 - P = 0.2$ ,在一般人群中,龙年出生的比例 $\pi_2 = \frac{1}{12}$ ,求抽到敏感问题①人中,回答“是”的比例 $\pi_1$ 。

应用公式(6),将本例有关数值代入得

$$\hat{\pi}_1 = \frac{0.3 - 0.2 \times \frac{1}{12}}{0.8} = 0.3542$$

这说明未婚女性有35.42%的人有过性关系。

再应用公式(7)、(4)求得

$$\text{Var}(\hat{\pi}_1) = \frac{0.3 \times 0.7}{500 \times 0.8^2} = 0.00066$$

总体 $\pi_1$ 的95%可信区间为

$$0.3542 \pm 1.96 \sqrt{0.00066} = 0.3040 \sim 0.4044$$

这说明总人群中,未婚女性有过性关系的比例在30.40%~40.44%范围内的可能性为95%,也就是说,做这样判断100次,有95次是对的。

**三、RRT方法的定量模式:**上述两种模式都只能解决具有敏感问题的人占多大的比例,是个定性的率,不能回答具有敏感问题的程度,比如,例3只说明了有35.42%的未婚女性有过性关系,但不能说明有过多少性伴,发生过多少次性关系。RRT方法的定量模式,就是为了解决这类定量的敏感问题而设计的。具体做法是只设立一个定量的敏感问题,比如,你有几个婚外性伴侣?在一个袋中放两种颜色的球(比如,为红、白色球),红球上没有任何数,白球上有一个数码,可为0,1,2,...,K,K是估计人群有敏感问题的人的最高程度,比如,婚外性伴侣最多为5个,则K=5。被调查者在无人的条件下,进行抽球,然后不记名地回答。若抽到红球,应答者就根据自己的实际情况回答,比如,被调查者有2个婚外性伴侣则回答“2”,若没有婚外性伴侣则回答“0”,如果抽到白球,则回答白球的数码。

设白球总数为 $W = \sum_{i=0}^k w_i$ ,红球总数为r,袋中球的总数为r+w,n为样本数,人群中有i个婚外性伴侣的比例为 $\pi_i$ ,则有 $\sum_{i=0}^k \pi_i = 1$ 。显见,人群中总的回答“i”的比例 $\lambda_i$ 为

$$\lambda_i = \pi_i \left( \frac{r}{r+w} \right) + \left( \frac{w_i}{r+w} \right) \quad (9)$$

或 
$$\hat{\pi}_i = \frac{\lambda_i(r+w) - w_i}{r} \quad (10)$$

$$\text{Var}(\hat{\pi}_i) = \left(\frac{r+w}{r}\right)^2 \left(\frac{\lambda_i(1-\lambda_i)}{n}\right) \quad (11)$$

[例 4] 某地对 300 名已婚成年男子调查其中有婚外性伴侣的数量情况,估计最多婚外性伴侣的数量为 5 个。在袋中装入红球 20 个,写有 0, 1, 2, 3 的白球各 2 个,写有 4, 5 的白球各 1 个。在无人的条件下,让被调查者从袋中摸球,凡摸到红球的人,则无记名的回答“你有几个婚外的性伴侣?”,有几个即回答“几个”,没有即回答“零”。若摸到白球,则回答白球上的数码即可。结果为  $\lambda_0 = 0.6, \lambda_1 = 0.12, \lambda_2 = 0.10, \lambda_3 = 0.09, \lambda_4 = 0.05, \lambda_5 = 0.04$ , 求  $\pi_i (i = 0,$

1, 2, 3, 4, 5) 及总体  $\pi_i$  的 95% 可信区间。

已知  $n = 300, r = 20, w = 10$  (其中  $w_0 \sim w_3$  皆为 2,  $w_4, w_5$  皆为 1), 将有关数值代入公式(10), 得

$$\hat{\pi}_0 = \frac{0.6 \times (20 + 10) - 2}{20} = 0.8$$

余类推, 结果见附表。

再将有关数值代入公式(11)、(4), 得

$$\text{Var}(\hat{\pi}_0) = \left(\frac{20+10}{20}\right)^2 \times \left(\frac{0.6 \times (1-0.6)}{500}\right) = 0.00108$$

总体  $\pi_0$  的 95% 可信区间为

$$0.8 \pm 1.96 \sqrt{0.00108} = 0.74 \sim 0.86$$

余类推, 结果见附表。

附表 某地已婚成年男子婚外性伴侣数量的比例

	婚外性伴侣数 i					
	0	1	2	3	4	5
有 i 个婚外性伴侣的比例 $\pi_i$	0.80	0.08	0.05	0.035	0.025	0.01
总体 $\pi_i$ 的 95% 可信区间	0.74~0.86	0.0373~0.1227	0.0106~0.0894	0~0.0726	0.0045~0.0455	0~0.0358

这个结果说明, 无婚外性伴侣的已婚男子占 80%, 从这个样本来推断总体无婚外性伴侣的已婚男子占该人群比例的 95% 可信区间为 74%~86%。

四、几点注意事项:

1. 当进行敏感问题调查时, 事先要对被调查者

说明该调查的意义, 该调查方法的绝对保密性, 使调查者消除顾虑, 如实回答问题。

2. 被调查者摸球、答题的过程要在无人的条件下进行, 环境安静舒适, 没有干扰。

3. 调查人态度和蔼可亲, 热情礼貌, 服务周到。

## 英国医学杂志中文版创刊

英国医学杂志中文版(BMJ Chinese Edition)于 1998 年 5 月创刊, 由英国医学杂志中文版编辑委员会编辑, 中华医学会主办并出版。本刊为高级综合性医学翻译期刊, 它以快捷、易读的方式将世界先进医学成果及动态及时介绍给中国读者。它如同一扇窗户, 使中国读者更加了解世界; 它如同一座桥梁, 促进国内外医学界的沟通与合作。本刊适合于医学院校、研究机构、县以上医院的各级医师, 重点为主治医师以上的临床工作者。主要栏目有: 述评、医学新闻、论著、综述、教育与争鸣、观点与评论、读者来信等。

本刊编辑部地址: 北京东四西大街 42 号; 邮政编码: 100710; 电话: (010)65273367; 传真: (010)65232552。

本刊为季刊, 大 16 开本, 48 页, 图文并茂, 每期定价 5.00 元。1998 年出版 3 期, 自办发行。需要者可向中华医学会杂志社出版发行部邮购, 1999 年起将由邮局发行。地址: 北京东四西大街 42 号; 邮政编码: 100710; 电话: (010)65251918; 传真: (010)65221454。请于汇款单上写明杂志名称及所需期号、册数。

英国医学杂志中文版编辑部