第35卷第4期 数学的实践与认识 Vol. 35 No. 4
2005年4月 MATHEMATICS IN PRACTICE AND THEORY April, 2005

数量特征敏感性问题调查的两个随机化回答模型

俞纯权

(山东经济学院统计系，山东 济南 250014)

摘要： 研究数量特征敏感性问题的抽样调查，设计了基于离散均匀分布和均匀分布的两个随机化回答模型.

关键词： 数量特征；敏感性问题；随机化回答模型；均匀分布

1 引 言

在社会经济调查中，有些调查项目具有很强的敏感性，被调查者往往不如实提供资料甚至拒绝调查. 对敏感性问题采用直接调查的方法不可能获得真实的数据资料，从而不能保证调查结论的准确性、可靠性，而且所得结论与事实的出入大小也无法度量，因此必须采取特殊的调查方法.

1965年Warner提出了随机化回答模型，开创敏感性问题调查之先河. 这种调查方法的基本特征是在调查中引入随机化装置，使被调查者在保证真实的前提下采取随机化回答的方式，因而既能为被调查者保守个人机密，同时也使调查者获得所需真实资料.

敏感性问题调查从统计上看可以分成两类. 一类是属性特征敏感性问题调查，它要解决的是估计总体的各种比例. 另一类是数量特征敏感性问题调查，它要解决的是估计总体某项指标的均值或总值. 对于这两种类型的敏感性问题调查，从Warner起已经有许多随机化回答模型问世. 本文基于离散均匀分布和均匀分布设计了两个随机化回答模型，以期为数量特征敏感性问题调查提供简便实用的调查方法.

2 模型设计的思路

设数量特征敏感性问题调查指标为Y，某有限总体含N个单元，其指标值记为Y_1，Y_2，…，Y_N，该指标的总体均值为

$$\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$$

总体方差为

$$S_Y^2=\frac{1}{N-1}\sum_{i-1}^{N}(Y_i-\overline{Y})^2$$

设调查的目标量是$\overline{Y}$. 由于Y_i是被调查者不愿公开暴露的数值，为得到估计$\overline{Y}$的可靠的样本数据，我们设计随机化回答模型. 设计的原则是：

1) 能为被调查者保密，只有被调查者确信这种调查方法能为其保密，他才肯做出真实的回答.

收稿日期：2003-09-14

2) 能获得估计 $\overline{Y}$ 的可靠的样本数据，不仅可以给出 $\overline{Y}$ 的无偏估计量，而且也能给出其方差的无偏估计量，同时为使估计达到一定精度，模型应具有可调节性.

3) 所设计的随机化回答模型不能太复杂，易于操作，有利于被调查者合作.

基于上述原则，引入具有离散均匀分布或均匀分布的随机变量 X，记

$$Z = XY \tag{1}$$

则 Z 是随机变量 X 与调查指标 Y 的乘积. 对于总体中第 i 个被调查者，设其在该模型设计下的一次随机试验中观察到 X 的一个随机观测值 X_i，此值仅让该被调查者观察，调查者不能观察，那么该被调查者就按式(1)计算得

$$Z_i = X_iY_i$$

然后该被调查者将数据 Z_i 提供给调查者. 由于调查者最终得到的数据是 Z_i 而非 X_i 及 Y_i，特别是调查者观察不到 X_i，因而他在获得数据 Z_i 后绝不可能推算出 Y_i 的值，这样就起到了为被调查者保密的作用. 式(1)是两项之积，计算极简单，易于操作，因而容易取得被调查者的合作.

为估计 $\overline{Y}$，在总体中进行简单随机抽样，记 n 个样本观测值为

$$z_i = x_iy_i, \quad i = 1,2,\cdots,n$$

则

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

即是指标 Z 的样本均值.

3　基于离散均匀分布的随机化回答模型

1) 模型

设随机变量 X 以等概率取[0, 2]上的有限个离散值. 首先设 X 取 0.6, 0.8, 1, 1.2, 1.4 五个值，

$$P\{X = m\} = \frac{1}{5}, \quad m = 0.6, 0.8, 1, 1.2, 1.4$$

这很容易实现. 在一个盒子里放 5 张同样大小的纸条，每张纸条上分别写上 0.6, 0.8, 1, 1.2, 1.4 中的一个数，充分混合后随机抽取一张，则抽到上述五个数中任何一个数的概率均为 $\frac{1}{5}$.

分别以 $\overline{X}$、$E(X^2)$ 表示 X 的均值和二阶原点矩，易知

$$\overline{X} = E(X) = 1, \quad E(X^2) = 1.08$$

由于 X 与 Y 相互独立，故若分别以 $\overline{Z}$、S_Z^2 表示 Z 的总体均值和总体方差，则

$$\overline{Z} = E(Z) = E(XY) = E(X)\cdot E(Y) = \overline{X}\cdot\overline{Y} = \overline{Y} \tag{2}$$

$$S_Z^2 = V(Z) = E(Z^2) - [E(Z)]^2 = 1.08(S_Y^2 + \overline{Y}^2) - \overline{Y}^2 = 1.08S_Y^2 + 0.08\overline{Y}^2$$

因为 $z_1, z_2, \cdots, z_n$ 是总体 Z 的简单随机样本，故

$$E(\bar{z}) = \overline{Z} = \overline{Y}$$

$$V(\bar{z}) = \frac{1-f}{n}S_Z^2 = \frac{1-f}{n}(1.08S_Y^2 + 0.08\overline{Y}^2)$$

式中 $f = \frac{n}{N}$ 是抽样比，记

$$v(\bar{z}) = \frac{1-f}{n}s_z^2$$

其中 $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$ 为样本方差，则

$$E\{v(\bar{z})\} = V(\bar{z})$$

由此我们便得到 $\bar{Y}$ 的一个无偏估计为

$$\hat{\bar{Y}}_1 = \bar{z} \tag{3}$$

其方差及方差的无偏估计分别为

$$V(\hat{\bar{Y}}_1) = \frac{1-f}{n}(1.08S_Y^2 + 0.08\bar{Y}^2) \tag{4}$$

$$v(\hat{\bar{Y}}_1) = \frac{1-f}{n}s_z^2$$

其次设 X 取 0.4，0.6，0.8，1，1.2，1.4，1.6 七个值，

$$P\{X = m\} = \frac{1}{7}, \quad m = 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6$$

这很容易实现. 在一个盒子里放7张同样大小的纸条，每张纸条上分别写上0.4，0.6，0.8，1，1.2，1.4，1.6 中的一个数，充分混合后随机抽取一张，则抽到上述七个数中任一个数的概率均为 $\frac{1}{7}$.

类似于前述过程可得

$$\begin{aligned}&\bar{X} = E(X) = 1, E(X^2) = 1.16\\ &\bar{Z} = E(Z) = E(XY) = E(X)\cdot E(Y) = \bar{X}\cdot\bar{Y} = \bar{Y}\\ &S_Z^2 = V(Z) = E(Z^2) - [E(Z)]^2 = 1.16S_Y^2 + 0.16\bar{Y}^2\end{aligned} \tag{5}$$

此时 $\bar{Y}$ 的无偏估计量为

$$\hat{\bar{Y}}_2 = \bar{z} \tag{6}$$

其方差及方差的无偏估计分别为

$$V(\hat{\bar{Y}}_2) = \frac{1-f}{n}(1.16S_Y^2 + 0.16\bar{Y}^2) \tag{7}$$

$$v(\hat{\bar{Y}}) = \frac{1-f}{n}s_Z^2$$

2) 关于模型的讨论

① 在给定 X 的分布时，我们选择以1为均值且在[0，2]上取得奇数个等间隔数值的离散均匀分布，其目的一是使式(2)式(5)成立，从而使 $\bar{Z}$ 的无偏估计 $\bar{z}$ 直接成为 $\bar{Y}$ 的无偏估计，而且估计量构造简单，同时容易得到估计量方差的无偏估计量. 二是容易给出 S_Z^2 与 S_Y^2 之间的关系从而易于得到估计量的方差.

② 考虑一个特殊情形. 若 X 恒取1，由式(1)则 $Z = Y$，此时对应于直接调查法，无保密性可言. 假设被调查者都能给出实事求是的回答，以 $\hat{\bar{Y}}_0$ 记 $\bar{Y}$ 的无偏估计，则

$$V(\hat{\bar{Y}}_0) = \frac{1-f}{n}S_Y^2 \tag{8}$$

将式(4)式(7)与式(8)比较得

$$V(\hat{\bar{Y}}_0) < V(\hat{\bar{Y}}_1) < V(\hat{\bar{Y}}_2)$$

由此可知使用随机化回答 Z 由于引入随机变量 X，增加了不确定信息，因而使 $S_Z^2 > S_Y^2$，导致随机化回答的精度比基于正确回答的直接调查法低. 但实际上直接调查法或者根本得不到回答或者得到带有系统偏差的数据，无精度可言，因而两相比较，随机化回答毕竟给出具有一定精度的无偏估计.

③ $V(\hat{\bar{Y}}_1) < V(\hat{\bar{Y}}_2)$ 表明当 X 取 0.6，0.8，1，1.2，1.4 五值时随机化回答的精度比 X 取 0.4，0.6，0.8，1，1.2，1.4，1.6 七值时的精度高. 这说明该模型在精度上具有可调节性. 理论上可以证明，X 取值越密集于1 附近，精度就越高，但保密性也就越差，因为 Z_i 与 Y_i 值比较接近. 在极端情形，当 X 恒取1 时，精度达到最高，但毫无保密性可言，因为此时 $Z_i = Y_i$. X 取值离散程度越大，则保密性就越好，因为 Z_i 值与 Y_i 值差异比较大，但精度比较低. 因此兼顾保密性和精度同时达到最理想水平的随机化回答模型设计是不可能的，X 取值范围的选择应在精度和保密性两者之间权衡. 根据调查问题的性质，若对保密性要求不是太高，为能得到精度比较高的估计应令 X 取值较集中于 1 的附近，离散程度小一些；若问题是高度敏感性的，对保密性要求很高，则取值范围就必须大一些，否则被调查将不乐于合作. 但取值范围也不能太大，一般情况下取以 0 和 2 为两个顶端值的离散均匀分布即可.

4　基于均匀分布的随机化回答模型

1) 模型

设随机变量 X 服从 $[1-a, 1+a]$ 上的均匀分布，其中 a 事先给定，$0 < a \leq 1$，则 X 的密度函数为

$$f(x) = \begin{cases} \dfrac{1}{2a}, & 1-a \leq x \leq 1+a \\ 0, & \text{其它} \end{cases}$$

得到服从该均匀分布的随机数很容易在计算机上实现. 分别以 $\bar{X}$、$E(X^2)$ 表示 X 的均值和二阶原点矩，易知

$$\bar{X} = E(X) = 1, \quad E(X^2) = \frac{3+a^2}{3}$$

由于 X 与 Y 相互独立，故若分别以 $\bar{Z}$、S_Z^2 表示 Z 的总体均值和总体方差，则

$$\bar{Z} = E(Z) = E(XY) = E(X) \cdot E(Y) = \bar{X} \cdot \bar{Y} = \bar{Y} \tag{9}$$

$$S_Z^2 = V(Z) = E(Z^2) - [E(Z)]^2 = \frac{3+a^2}{3} S_Y^2 + \frac{a^2}{3} \bar{Y}^2$$

因为 $z_1, z_2, \cdots, z_n$ 是总体 Z 的简单随机样本，故

$$E(\bar{z}) = \bar{Z} = \bar{Y}$$

$$V(\bar{z}) = \frac{1-f}{n} S_Z^2 = \frac{1-f}{n}\left(\frac{3+a^2}{3} S_Y^2 + \frac{a^2}{3} \bar{Y}^2\right)$$

式中 $f = \dfrac{n}{N}$ 是抽样比. 记

$$v(\bar{z}) = \frac{1-f}{n} s_z^2$$

其中 $s_z^2=\frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2$ 是样本方差，则

$$E\{v(\bar{z})\}=V(\bar{z})$$

由此我们便得 $\bar{Y}$ 的一个无偏估计为

$$\hat{\bar{Y}}_3=\bar{z} \tag{10}$$

其方差及方差的无偏估计分别为

$$V(\hat{\bar{Y}}_3)=\frac{1-f}{n}\left(\frac{3+a^2}{3}S_Y^2+\frac{a^2}{3}\bar{Y}^2\right) \tag{11}$$

$$v(\hat{\bar{Y}}_3)=\frac{1-f}{n}s_z^2$$

2) 关于模型的讨论

① 我们取 X 的分布为$[1-a,1+a]$上的均匀分布，其中 $0<a\leq 1$，其目的一是使X的均值为1使式(9)成立，从而使 $\bar{Z}$ 的无偏估计 $\bar{z}$ 直接成为 $\bar{Y}$ 的无偏估计，而且估计量构造简单，同时容易得到估计量方差的无偏估计量. 二是容易给出 S_Z^2 与 S_Y^2 之间的关系，从而易于得到估计量的方差.

② 设计模型时我们取 $0<a\leq 1$. 考虑一个特殊情形. 若 $a=0$，即 X 恒取1，由式(1)则 $Z=Y$，此时对应于直接调查法，无保密性可言. 假设被调查者都能给出实事求是的回答，以 $\hat{\bar{Y}}_0$ 记 $\bar{Y}$ 的无偏估计，则有式(8). 将式(11)与式(8)比较得

$$V(\hat{\bar{Y}}_0)<V(\hat{\bar{Y}}_3)$$

由此可知，使用随机化回答Z由于引入随机变量 X，增加了不确定信息，因而使 $S_Z^2>S_Y^2$，导致随机化回答的精度比基于正确回答的直接调查法低. 但实际上直接调查法或者根本得不到回答，或者得到的是带有系统偏差的数据，无精确度可言，因此两相比较，随机化回答毕竟给出具有一定精度的无偏估计.

③ 由式(11)知，a 取不同值时估计量方差也不同，这说明该模型在精度上具有可调节性. 当a趋于0时，由式(11)知 $V(\hat{\bar{Y}}_3)$ 将趋于 $V(\hat{\bar{Y}}_0)$，精度比较高，但保密性比较差，因为 Z_i 值与 Y_i 值比较近. 当 $a=0$ 时精度达最高，但毫无保密性可言，因为此时 X 恒为1，$Z_i=Y_i$. 当 a 取值越大时，保密性就越好，因为 Z_i 值与 Y_i 值差异比较大，但精度比较低. 因此 a 的选择不能同时兼顾保密性和精度同时达到最理想的水平，a 取值的选择应在精度和保密性两者之间权衡. 根据调查问题的性质，若对保密性要求不是太高，为能得到精度较高的估计 a 取值应较小些；若问题是高度敏感性的，对保密性要求很高，则 a 取值就应大一些，否则被调查者将不乐于合作. 但 a 取值也不能太大，一般情况下 a 最大取到1即可.

参考文献：

[1] 科克伦 W G. 抽样技术[M]. 张尧庭，吴辉译，北京：中国统计出版社，1985.

[2] 冯士雍，施锡铨著. 抽样调查——理论、方法与实践[M]. 上海：上海科学技术出版社，1996.

[3] 金莹，梁小筠. 对定量的敏感性问题的一种改进调查法及其估计量[J]. 统计研究，2000，(11)：58—61.

[4] 俞纯权. 连续变量隐私问题的随机化回答模型[J]. 数理统计与管理，1999，(1)：39—44.

The Two Randomized Response Models of Survey on Sensitive Question with Quantitative Characteristic

YU Chun-quan

(Shandong Economics University, Jinan Shandong 250014, China)

Abstract: This paper studies the survey sampling on sensitive question with quantitative characteristic, designs two randomized response models in view of discrete uniform distribution and uniform distribution.

Keywords: quantitative characteristic; sensitive question; randomized response models; uniform distribution

期 刊 简 介

本刊主要刊登数学的最新的理论成果,及其在工业、农业、环境保护、军事、教育、科研、经济、金融、管理、决策等工程技术、自然科学和社会科学中的应用成果、方法和经验. 主要任务是沟通数学工作者与其他科技工作者之间的联系,推动应用数学在我国的发展,为四化建设作贡献.

主要栏目:数学建模、管理科学、工程、问题研究、知识与进展、学科介绍、方法介绍、高等数学园地、数学史、研究简报、书刊评介、简讯.

注:① 投稿一式二份,原稿自己保存,编辑部不退还投稿的稿件.

② 编辑部不接收网上投稿.