一种定量敏感性问题的调查方法及其估计

孙慧玲，何秀梅
（华中师范大学 数统学院，湖北 武汉 430079）

[摘 要]讨论了定量敏感性问题的一种调查方法，在 Eichhorn · Hayre 的随机化调查模型的基础上，提出了一种更为可行的模型设计，并通过效率比较证明该模型比 Hayre 的随机化调查模型精度更高；最后利用比例估计，得出定量敏感变量高于（或低于）某一值 a 的调查者在总体中所占的比例.

[关键词]敏感性问题；简单随机抽样；比例估计
[中图分类号]O212 [文献标识码]A [文章编号]1673 - 8012(2007)05 - 0018 - 03

所谓敏感性问题，是指个人或单位因某种原因不便向外界透露的问题. 这些问题中，可以用“是”或“否”回答的敏感性问题，我们称之为定性敏感性问题. 但现实中存在诸如某行业的利润百分比为多少、某单位员工的年收入是多少等敏感性问题，因此我们有必要引入定量敏感性问题.

对于敏感性问题，调查中若采用直接问答的方式，被调查者为了保护自己的隐私或出于其它目的往往会拒绝回答，这样就破坏了我们收集证据的真实性，而且破坏程度的大小我们也无法度量.

自 Warner 于 1965 年提出使用随机化问答技术实施对敏感性问题的抽样调查之后，Simmons（1967 年）、N · S · Mangat（1990 年）等对敏感性问题提出了多种调查方法. 对于定性敏感性问题，Warner（1965 年）、Simmons（1967 年）、N · S · Mangat（1990 年）以及文献[1]已经讨论得比较清楚了，定量敏感性问题是对定性问题的直接推广. 本文主要讨论定量敏感性问题.

1 Eichhorn · Hayre 随机化调查模型

为估计定量敏感变量的均值，Eichhorn · Hayre 提出一个随机化调查方法，即被调查者回答一个随机扰动变量值与其自身敏感变量的乘积，扰动变量的一个随机数，其分布是已知的. 由于调查者不知道在调查中被调查者使用的随机数，这样就起到为被调查者保密的作用.

设 X 是所要调查的定量敏感问题的特征量，Y 是一个与敏感问题无关的扰动随机变量，Y 的分布已知. 调查方法为：

第 1 步：产生一概率密度为 $f(y)$ 的随机数 Y；

第 2 步：被调查者回答 X 、Y 的积，其中 X 与 Y 独立，记 $Z = XY$，记 $\mu_x = E(X)$，$\mu_y = E(Y)$，$\sigma_x^2 = V(X)$，$\sigma_y^2 = V(Y)$. 其中，μ_y 、σ_y^2 是已知的，μ_x 、σ_x^2 是未知的.

从 N 个总体中使用有放回的简单随机抽样抽取容量为 n 的样本，n 个个体提供随机回答 $Z_1, Z_2, \cdots, Z_n$. 样本均值为：$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$；$Z_i$ 的均值为：$E(Z_i) = E(X_i)E(Y_i) = \mu_x\mu_y$；$Z_i$ 的方差为：$V(Z_i) = \mu_y^2\sigma_x^2 + (\sigma_x^2 + \sigma_y^2)\sigma_y^2$.

第 3 步：给出敏感指标均值 μ_x 的一个无偏估计量为：$\widehat{\mu_x} = \frac{\overline{Z}}{\mu_y}$；

方差为：$V(\widehat{\mu_x}) = \frac{1}{n}\left\{\sigma_x^2 + (\sigma_x^2 + \mu_x^2)\frac{\sigma_y^2}{\mu_y^2}\right\}$.

* [收稿日期]2007 - 07 - 09
[作者简介]孙慧玲（1982 - ），女，湖北孝感人，在读硕士研究生.

特别地，如果 $Y = 1$，则 Eichhorn · Hayre 提出的随机化调查方法就是直接调查，此时估计量的方差为 σ_x^2/n. Hayre 提出的方法简单，容易操作，但每个被调查者给出的回答在含有敏感信息问题的同时，又引入了与调查问题无关的扰动信息，使调查精度下降. 因此，我们需要在尽可能多地包含有用信息的同时，减少无关信息的引入，既使每个被调查者的回答都含有敏感性问题的信息，又使随机数仅以一定的概率影响被调查者的最终回答.

2 提出改进方法

设 X 为所要调查的定量敏感问题，为提高被调查者的合作程度，得到真实、有效的数据，使用随机化装置独立产生一随机数 Y.

其调查方法为：

第 1 步：产生一概率密度为 $f(y)$ 的随机数 Y；

第 2 步：产生一(0,1)分布的随机数 ε，使 $P(\varepsilon = 1) = P$，$P(\varepsilon = 0) = 1 - P$；

第 3 步：如果 $\varepsilon = 1$，则要求被调查者回答 $X\mu_y$；如果 $\varepsilon = 0$，则回答 XY. X、Y 与 ε 相互独立，研究者只能看到被调查者给出的最终回答 Z，则得到：

$$Z = \varepsilon X\mu_y + (1 - \varepsilon)XY.$$

其中，μ_y、σ_y^2 是已知的，μ_x、σ_x^2 是未知的，$\mu_x = E(X)$，$\mu_y = E(Y)$，$\sigma_x^2 = V(X)$，$\sigma_y^2 = V(Y)$.

从容量为 N 的总体中使用简单随机有效有放回抽样，抽取容量为 n 的样本，n 个样本个体提供的样本数据为 $Z_1, Z_2, \cdots, Z_n$，样本均值为 $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Z_i 的均值为：$E(Z_i) = \mu_x\mu_y$. 令 $\bar{Z} = E(Z_i)$，得到敏感问题特征量均值 X 的一个无偏估计：$\hat{\mu}_{xm} = \frac{\bar{Z}}{\mu_y}$.

定理 1 估计量 $\hat{\mu}_{xm}$ 的方差为：$V(\hat{\mu}_{xm}) = \frac{1}{n}\left\{\sigma_x^2 + (1 - P)(\sigma_x^2 + \mu_x^2)\frac{\sigma_y^2}{\mu_y^2}\right\}$.

证明 由 $\hat{\mu}_{xm} = \frac{\bar{Z}}{\mu_y}$，有 $V(\hat{\mu}_{xm}) = \frac{1}{n\mu_y^2}V(Z_i)$. 下面只需计算 $V(Z_i)$.

$$V(Z_i) = EZ_i^2 - (EZ_i)^2 = E(\varepsilon X_i\mu_y + (1 - \varepsilon)X_iY)^2 - (\mu_x\mu_y)^2 =$$

$$P\mu_y^2EX_i^2 + (1 - P)E(x_i^2)E(y_i^2) - (\mu_x\mu_y)^2 =$$

$$P\mu_y^2(\mu_x^2 + \sigma_x^2) + (1 - P)(\mu_x^2 + \sigma_x^2)(\mu_y^2 + \sigma_y^2) - (\mu_x\mu_y)^2 = \mu_y^2\sigma_x^2 + (1 - P)\sigma_y^2(\mu_x^2 + \sigma_x^2).$$

3 模型的比较与选择

在采用随机化模型时，一般要从无偏性和有效性两个方面进行比较.

(1) 关于无偏性，从上面的结论可以分析出，原 Hayre 模型和改进模型敏感性变量 X 的无偏估计都为 $\hat{\mu}_{xm} = \frac{\bar{Z}}{\mu_y}$.

(2) 关于有效性，我们选择方差最小的为“最优”的无偏估计.

定理 2 对所有的 $0 < P < 1$，都有 $V(\hat{\mu}_{xm}) < V(\hat{\mu}_x)$ 成立.

证明 由于 $V(\hat{\mu}_x) = \frac{1}{n}\left\{\sigma_x^2 + (\sigma_x^2 + \mu_x^2)\frac{\sigma_y^2}{\mu_y^2}\right\}$，$V(\hat{\mu}_{xm}) = \frac{1}{n}\left\{\sigma_x^2 + (1 - P)(\sigma_x^2 + \mu_x^2)\frac{\sigma_y^2}{\mu_y^2}\right\}$，

$V(\hat{\mu}_x) - V(\hat{\mu}_{xm}) = \frac{1}{n}P(\sigma_x^2 + \mu_x^2)\frac{\sigma_y^2}{\mu_y^2}$. 当 $0 < P < 1$ 时，$(\hat{\mu}_x) - V(\hat{\mu}_{xm}) > 0$ 恒成立. 因此，改进模型比原有模型的精确度提高了.

4 对改进模型比例的估计

对定量敏感性问题，我们除了想知道估计均值外，还想考虑具有某一特征的比例估计. 如在调查化妆品行业的利润时，我们会想到利润高于某一值 a 的被调查者在总体中占的比例. 下面给出 $\theta = P(x > a)$ 的估计：由于 $Z = \varepsilon(X\mu_y) + (1 - \varepsilon)XY$，所以可得：$X = \frac{Z}{\varepsilon\mu_y + (1 - \varepsilon)Y}$.

首先把变量进行改造,把满足条件的 Z 作为1,否则为0. 这样,利用二项分布的定义,我们很容易得到 θ 的估计: $\theta = qP(\frac{Z}{\mu_y} > \alpha) + (1-q)P(\frac{Z}{Y} > \alpha) = qP(Z > \alpha\mu_y) + (1-q)P(Z > \alpha Y)$.

(1)若 Y 为离散型随机变量,假设取值为 $Y_j(j = 1,2\cdots)$,则:

$$\theta = qP(Z > \alpha\mu_y) + (1-q)\sum_{j=1}^{\infty}[P(Z > \alpha Y_j) \cdot P(Y = Y_j)] =$$

$$\frac{1}{n}[q\sum_{i=1}^{\infty} I(Z_i > \alpha\mu_y) + (1-q)\sum_{i=1}^{n}\sum_{j=1}^{\infty} P(Y = Y_j) \times I(Z_i > \alpha Y_j)] =$$

$$\frac{1}{n}[q\sum_{i=1}^{\infty} I(Z_i > \alpha\mu_y) + (1-q)\sum_{i=1}^{n}\sum_{\substack{j=1,\cdots,n \\ Z_j > \alpha Y_i}}^{\infty} P(Y = Y_j)].$$

(2)若为连续型变量,则:

$$\theta = qP(Z > \alpha\mu_y) + (1-q)\int_y [P(Z > \alpha Y) \cdot f(y)]\mathrm{d}y =$$

$$\frac{1}{n}[q\sum_{i=1}^{\infty} I(Z_i > \alpha\mu_y) + (1-q)\sum_{i=1}^{n}\int_y I(Z_i > \alpha Y) \cdot f(y)\mathrm{d}y] =$$

$$\frac{1}{n}[q\sum_{i=1}^{\infty} I(Z_i > \alpha\mu_y) + (1-q)\sum_{i=1}^{n}\int_{y<\frac{Z_i}{\alpha}} f(y)\mathrm{d}y].$$

5 结语

根据对定量敏感性变量的均值、方差进行估计,本文提出的改进方法比 Eichhorn · Hayre 的乘积模型更有效. 虽然在调查过程中,本文的方法略显复杂,但在实施过程中,若费用一定,本文提出的参数 p 使调查方法更具可调节性,精度也提高了. 另外,本文提出的比例估计计算简单,对于估计敏感变量高于(或低于)某一值 a 是一种行之有效的方法.

[参考文献]

[1]Yan Zaizai. An alternative randomized response device[J]. 应用概率统计,2005.

[2]Eihhorn B H, Hayre Ls. Scrambled randomized response methods for obtaining. Sensitive quantitative data[J]. J of Statistical Planning and Inference,1983.

[3]冯士雍,倪加勋,邹国华. 抽样调查理论与方法[M]. 北京:中国统计出版社,1998.

An Investigating Method of Quantitative and Sensitive Questions with its Estimation

SUN Hui – ling, HE Xiu – mei

(*Mathematics and Statistics College, Huazhong Normal University, Wuhan Hubei* 430079, *China*)

Abstract: This paper discussed an investigating method of quantitative and sensitive questions. Based on the random model of Eichhorn · Hayre, a more feasible model is put forward, which proved to be more precise than Model Hayer after the comparison of efficiency. In the end, the conclusion can be reached that the sensitive variable is higher or lower than the proportions of a certain value a.

Key words: sensitive questions; simple and random sample; estimation of proportions