DOI:10.13546/j.cnki.tjyjc.2010.01.015

基于双无关问题的敏感性问题随机化回答模型

饶贤清

(上饶师范学院 数学与计算机系,江西 上饶 334001)

摘 要:研究数量特征敏感性问题的抽样调查,设计了双无关问题双样本随机化回答模型,给出了总体均值的无偏估计、估计量的方差, 并得出新的模型具有较好的精度。

关键词:敏感性问题;随机化回答;Greenberg 模型;无偏估计

中图分类号:O212;C8 文献标识码:A 文章编号:1002-6487(2010)01-0161-02

0 引言

1965 年 Warner 提出了敏感性问题的随机化回答调查方法[1]; 1969 年 Greenberg 提出了数量特征敏感性问题的无关问题随机化回答方法[2],在 Greenberg 模型中由于有的概率回答是与被调查者无关的问题,使得信息大量浪费, 从而导致敏感性问题估计量的估计精度不高;2000 年孙山泽等人提出了数量特征敏感性问题的随机变量和模型,在使用随机变量和模型调查时,要求每个被调查者回答看到的计算机产生的来自密度函数为 g(y)的随机数与自己隐性收入的和,记被调查者的最终回答为 Z;对敏感性问题的回答为 X;对随机数的回答为 Y,则 Z=X+Y。记 X,Y,Z 的总体均值分别为 $\bar{X},\bar{Y},\bar{Z}$,总体方差分别为 S_X^2,S_Y^2,S_Z^2;设 $z_1,z_2,\cdots,z_n$ 是从总体中抽取容量 n 为的简单随机样本。记 Z 的样本均值为 $\bar{Z}$,样本方差为 s_z^2,其中 $s_z^2=\frac{1}{n-1}\sum_{i=1}^{n}(z_1-\bar{z})^2$。则 $\bar{X}$ 的无偏估计量为 $\hat{\bar{X}}^{(1)}=\bar{z}-\bar{Y}$,估计量的方差为

$$Var(\hat{\bar{X}}^{(1)})=\frac{1-f}{n}S_Z^2=\frac{1-f}{n}(S_X^2+S_Y^2)$$

其中 $f=\frac{n}{N}$。又由于 $Var(\hat{\bar{X}}^{(1)})-Var(\hat{\bar{X}}^{(0)})=\frac{1-f}{n}(S_X^2+S_Y^2)-\frac{1-f}{n}S_X^2\geqslant 0$,其中 $Var(\hat{\bar{X}}^{(0)})=\frac{1-f}{n}S_X^2$ 为直接提问时均值估计量的方差,可见随机变量和模型与直接回答模型相比精度差,要使估计量精度提高我们提出以下的改进方法。

1 双无关问题的随机化回答模型

双无关问题的数量特征敏感性问题随机化回答调查法由敏感性问题和两个非敏感性问题组成,使用该模型进行调查时,将被调查者分成两组,要求两组被调查者(两组样本)都要给出两个回答:随机化回答和直接回答,具体的调查法可由表 1 表示。

表 1

调查采用技术	样本 1	样本 2
随机化回答	敏感性问题敏感性问题+无关问题 1	敏感性问题敏感性问题+无关问题 2
直接回答	无关问题 2	无关问题 1

假设每组样本中随机化回答部分被调查者对敏感性问题回答的概率均为 p。记两组被调查者的最终回答 $Z^{(1)},Z^{(2)}$;被调查者对敏感性问题的回答为 X;对两个无关问题的回答分别为 Y_1,Y_2.从大小为 N 的总体中按简单随机抽样抽取样本容量分别 n_1,n_2 的两个独立样本.第个样本中被调查者对随机化装置的回答为 $Z_1^{(i)}$,对直接问题的回答为 $Z_1^{(i)}$,i=1,2,.X,Y_1,Y_2,$Z_1^{(i)},Z_2^{(i)}$,的总体均值分别为 $\bar{X},\bar{Y}_1,\bar{Y}_2,\bar{Z}_1^{(i)},\bar{Z}_2^{(i)}$;总体方差分别为 $S_X^2,S_{Y_1}^2,S_{Y_2}^2,Z_{Z_1^{(i)}}^2,Z_{Z_2^{(i)}}^2$;其协方差分别为 $S_{XY_1},S_{XY_2},S_{Y_1Y_2},S_{Z_1^{(i)}Z_2^{(i)}},S_{Z_1^{(2)}Z_2^{(2)}}$。$Z_1^{(i)},Z_2^{(i)}$的样本均值为 $\bar{Z}_1^{(i)},\bar{Z}_2^{(i)}$,样本方差为 $S_{Z_1^{(i)}}^2,S_{Z_2^{(i)}}^2$,样本协方差为 $s_{Z_1^{(i)}},s_{Z_2^{(i)}},s_{Z_1^{(2)}Z_2^{(2)}}$其中 $s_{Z_1^{(i)}}^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(z_{1j}^{(i)}-\bar{z}_1^{(i)})^2$,$s_{Z_2^{(i)}}^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(z_{2j}^{(i)}-\bar{z}_2^{(i)})^2$,i=1,2。将上述方法归纳为如下模型

$$\begin{cases}Z_1^{(1)}=\varepsilon X+(1-\varepsilon)(X+Y_1)\\ Z_2^{(2)}=Y_2\end{cases}\text{与}\begin{cases}Z_1^{(2)}=\varepsilon X+(1-\varepsilon)(X+Y_2)\\ Z_2^{(2)}=Y_1\end{cases}$$

其中 $p(\varepsilon=1)=p$,且 ε 分别与 X,Y_1,Y_2 独立,由数学期望公式得:

$$\begin{cases}\bar{Z}_1^{(1)}=p\bar{x}+(1-p)(X+\bar{Y}_1)\\ \bar{Z}_2^{(2)}=\bar{Y}_2\end{cases}\text{与}\begin{cases}\bar{Z}_1^{(2)}=p\bar{x}+(1-p)(X+\bar{Y}_2)\\ \bar{Z}_2^{(2)}=\bar{Y}_1\end{cases}$$

我们取估计量 $\hat{\bar{Z}}_1^{(1)}=\bar{z}_1^{(1)},\hat{\bar{Z}}_2^{(1)}=\bar{z}_2^{(1)},\hat{\bar{Z}}_1^{(2)}=\bar{z}_1^{(2)},\hat{\bar{Z}}_2^{(2)}=\bar{z}_2^{(2)}$ 该组估计量是无偏的,且 $s_{Z_1^{(i)}}^2,s_{Z_2^{(i)}}^2,s_{Z_1^{(2)}}^2,s_{Z_1^{(2)}}^2$ 及 $s_{Z_1^{(i)}Z_2^{(i)}},s_{Z_1^{(2)}Z_2^{(2)}}$分别为 $s_{Z_1^{(i)}}^2$ $s_{Z_2^{(i)}}^2,s_{Z_1^{(2)}}^2s_{Z_2^{(2)}}^2$ 及 $S_{Z_1^{(i)}Z_2^{(i)}},S_{Z_1^{(2)}Z_2^{(2)}}$的无偏估计,可得 $\bar{X}$ 的两个无偏估计

$$\hat{\bar{X}}^{(21)}=\bar{z}_1^{(1)}-(1-p)\bar{z}_2^{(2)}$$

及

作者简介:饶贤清(1970-),男,江西上饶人,硕士,研究方向:概率论与数理统计。

$\hat{\overline{X}}^{(22)}=\bar{z}_1^{(2)}-(1-p)\bar{z}_2^{(1)}$

我们可取两者的加权平均为其估计量，即

$\hat{\overline{X}}^{(2)}=\omega\hat{\overline{X}}^{(21)}+(1-\omega)\hat{\overline{X}}^{(22)},0\leqslant\omega\leqslant1$

从使方差最小的理论的角度出发，在给定 n_1 和 n_2 时，可取 ω 使 $\mathrm{Var}(\hat{\overline{X}}^{(2)})$取最小值。 令

$$\sum\nolimits_1^2=\mathrm{Var}(\hat{\overline{X}}^{(21)})=\frac{1-f_1}{n_1}s_{Z_1^{(1)}}^2+(1-p)^2\times\frac{1-f_2}{n_2}s_{Z_2^{(2)}}^2$$

$$\sum\nolimits_2^2=\mathrm{Var}(\hat{\overline{X}}^{(22)})=\frac{1-f_2}{n_2}s_{Z_1^{(2)}}^2+(1-p)^2\times\frac{1-f_1}{n_1}s_{Z_2^{(1)}}^2$$

$$\sum\nolimits_{12}=\mathrm{Cov}(\hat{\overline{X}}^{(21)},\hat{\overline{X}}^{(22)})=-(1-p)[\frac{1-f_1}{n_1}s_{Z_1^{(1)}Z_2^{(1)}}+\frac{1-f_2}{n_2}s_{Z_2^{(2)}Z_2^{(2)}}]$$

可得

$$\mathrm{Var}(\hat{\overline{X}})=\omega^2(\sum\nolimits_1^2+\sum\nolimits_2^2-2\sum\nolimits_{12})-2\omega(\sum\nolimits_2^2-2\sum\nolimits_{12})+\sum\nolimits_2^2$$

要使 $\mathrm{Var}(\hat{\overline{X}}^{(2)})=\mathrm{Min}$，令 $\frac{\partial\mathrm{Var}(\hat{\overline{X}}^{(2)})}{\partial\omega}=0$，可得

$$\omega_{opt}=(\sum\nolimits_2^2-\sum\nolimits_{12})/(\sum\nolimits_1^2+\sum\nolimits_2^2-2\sum\nolimits_{12})$$

在给定 n_1 和 n_2 的情况下，得

$$\mathrm{MinVar}(\hat{\overline{X}}^{(2)})=(\sum\nolimits_1^2\sum\nolimits_2^2-\sum\nolimits_{12})/(\sum\nolimits_1^2+\sum\nolimits_2^2-2\sum\nolimits_{12})$$

其方差估计量可取为

$$\mathrm{Var}(\hat{\overline{X}}^{(2)})=(\hat{\sum}\nolimits_1^2\hat{\sum}\nolimits_2^2-\hat{\sum}\nolimits_{12})/(\hat{\sum}\nolimits_1^2+\hat{\sum}\nolimits_2^2-2\hat{\sum}\nolimits_{12})$$

该估计量是无偏的，其中

$$\hat{\sum}\nolimits_1^2=\hat{\mathrm{V}}\mathrm{ar}(\hat{\overline{X}}^{(21)})=\frac{1-f_1}{n_1}s_{Z_1^{(1)}}^2+(1-p)^2\frac{1-f_2}{n_2}s_{Z_2^{(2)}}^2$$

$$\hat{\sum}\nolimits_2^2=\hat{\mathrm{V}}\mathrm{ar}(\hat{\overline{X}}^{(22)})=\frac{1-f_2}{n_2}s_{Z_1^{(2)}}^2+(1-p)^2\frac{1-f_1}{n_1}s_{Z_2^{(1)}}^2$$

$$\hat{\sum}\nolimits_{12}=\hat{\mathrm{C}}\mathrm{ov}(\hat{\overline{X}}^{(21)},\hat{\overline{X}}^{(22)})=-(1-p)[\frac{1-f_1}{n_1}s_{Z_1^{(1)}Z_2^{(1)}}+\frac{1-f_2}{n_2}s_{Z_2^{(2)}Z_2^{(2)}}]$$

2 关于模型的讨论

如果引用的两个无关问题独立同分布且我们抽取的两组样本容量相等，即，$\overline{Y}_1=\overline{Y}_2=\overline{Y}$，$S_{Y_1}^2=S_{Y_2}^2=S_Y^2$，$\sum_{12}=0$ 和 $n_1=n_2=\frac{n}{2}$，此时 $\omega_{opt}=\frac{1}{2}$，估计量的最小方差为

$$\mathrm{MinVar}(\hat{\overline{X}}^{(2)})=\sum\nolimits_1^2/2=\sum\nolimits_2^2/2=\frac{1-f}{n}[S_X^2+(2+p)(1-p)S_Y^2+p(1-p)\overline{Y}_2]$$

与随机变量和模型估计量的方差 $\mathrm{Var}(\hat{\overline{X}}^{(1)})=\frac{1-f}{n}S_Z^2=\frac{1-f}{n}(S_X^2+S_Y^2)$相比，有

$$\mathrm{Var}(\hat{\overline{X}}^{(2)})-\mathrm{Var}(\hat{\overline{X}}^{(1)})=\frac{1-f}{n}[(1-p-p^2)S_Y^2+p(1-p)\overline{Y}^2]$$

当 $1-p-p^2<0$，即 $\frac{\sqrt{5}-1}{2}<p<1$ 时，$(1-p-p^2)S_Y^2<0$，因此，在给定的 p 下，我们可以选择合适的无关问题，就可以使 $\mathrm{Var}(\hat{\overline{X}}^{(2)})-\mathrm{Var}(\hat{\overline{X}}^{(1)})<1$，即使用双无关问题随机化回答模型比随机变量和模型具有更小的方差，而且，要求 $p>\frac{\sqrt{5}-1}{2}$ 是与尽可能地应用被调查者对敏感性问题回答的信息相吻合的。若取 p=0.7，则

$$\mathrm{Var}(\hat{\overline{X}}^{(2)})-\mathrm{Var}(\hat{\overline{X}}^{(1)})=\frac{1-f}{n}(-0.19S_Y^2+0.21\overline{Y}^2)$$

所以，我们在选择无关问题时，只要无关问题的分布满足 $19S_Y^2>21\overline{Y}^2$，就有

$$\mathrm{Var}(\hat{\overline{X}}^{(2)})<\mathrm{Var}(\hat{\overline{X}}^{(1)})$$

注意到当 $\frac{\sqrt{5}-1}{2}<p<1$ 时，$1-p-p^2<0$ 而且严格单调递减，$p(1-p)>0$ 也严格单调递减，当 $p=\frac{\sqrt{2}}{2}$ 时，$1-p-p^2=-p(1-p)>0$，所以，当 $\frac{\sqrt{5}-1}{2}<p<\frac{\sqrt{2}}{2}$ 时，无关问题的分布满足 $S_Y^2>\overline{Y}^2$；当 $\frac{\sqrt{2}}{2}<P<1$ 时，无关问题的分布要满足 $S_Y^2<\overline{Y}^2$，都有使用独立同分布的双无关问题随机化回答模型比随机变量和模型有更小的方差。

参考文献：

[1]Warner, S. L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias[J].Journal of the American Statistical Association,1965,(60).

[2]Greenberg, B. G, Abul-Ela, A., Simmons, W. R., Horvitz, D. G. The Unrelated Question Randomized Response Model:Theoretical Framework[J].Journal of the American Statistical Association.1969,(64).

[3]W. G 科克伦著.抽样技术[M].张尧庭、吴光军译.北京:中国统计出版社,1985.

（责任编辑/李友平）