文章编号:1001－5167(2004)02-0086-05

分层抽样下敏感问题研究*

洪志敏,闫在在
(内蒙古工业大学理学院,呼和浩特 010062)

摘要:在已有的敏感问题调查方法中,大多数讨论的是通过对随机化装置的改进以提高调查的效率.本文从理论和数值模拟两方面说明了抽样设计的恰当选择同样也能提高调查的精度.

关键词:敏感性问题;分层抽样;比例分配;沃纳模型
中图分类号:O212.4 **文献标识码:**A

0 引 言

在抽样调查中,有些调查项目具有很强的敏感性,如被调查者是否有吸毒行为,是否偷税漏税等.如果对这些敏感性问题采用直接调查的方法,调查者将难以控制样本信息,得不到可靠的样本数据.为了得到敏感性问题的可靠的样本数据,使被调查者能够很好地配合调查.Warner 在 1965 年开创性地提出了随机化回答调查法.这种调查方法的原理是在调查中引入随机化装置,使被调查者在保证真实回答的前提下,采用随机化回答装置,既能为被调查者保护个人隐私,也能使调查者获得所需的真实信息.之后,Horvitz,Shah & Simmons(1967)、N·S·Mangat(1990)、Anthony Y·C·KUK(1990)、Mangat(1994)等提出了各种敏感问题调查方法,但这些方法均是讨论如何改进随机化装置而没有考虑抽样设计的选择.这就激发我们考虑在敏感性问题调查中抽样方法对调查精度的影响.本文讨论了在分层抽样设计下使用随机化回答模型,并在费用一定的条件下精确比较了分层随机化方法与经典的随机化方法,同时给出了分层抽样优于简单抽样的条件.

在直接调查中,如果总体分成若干子总体,在每一子总体内单元间差异较小,这时只需在子总体中抽取少量样本单元就能很好地代表子总体的特征,从而通过分层抽样可以提高调查的精度,这也是实际中使用分层抽样设计的前提条件.如果敏感问题调查总体分成若干子总体,具有层间差异大,层内差异小的特点,则调查采用分层抽样应该有高的精度.本文从理论和数值模拟两方面给出了这方面的相关理论.

1 沃纳模型及分层抽样

沃纳模型是 1965 年由 Warner 提出的,其设计原则是提供给被调查者一包外形完全相同的卡片,卡片上分别写有问题"你有敏感属性 A 吗?(你属于 A 吗?)"和该问题的对立问题"你没有敏感属性 A 吗?(你属于 A^C 吗?)".两项问题卡片按预定比例配置,让被调查者从中抽取一张,根据抽到的问题和自

* 收稿日期:2004-06-03
作者简介:洪志敏(1975～),女,赤峰市人,内蒙古工业大学理学院助教.

身属性的匹配情况作出真实回答.这一过程不被调查者观察到,从而起到了为被调查者保密的效果.

1.1　模型设计及参数估计

设总体中的第一个体或者有敏感属性 A,或者没有敏感属性 A,调查目的是估计具有敏感属性(属于 A)的个体在总体中所占的比例 π

假定根据所掌握的有关调查指标的信息将总体分成 L 层,N_h 代表第 h 层的总体容量,在每一层使用简单随机放回抽样的方法抽取容量为 n_h 的样本($n=\sum_{h=1}^{L}n_h$),然后对每一层的样本进行随机化回答调查.W_h 代表第 h 层总体的层权且它是已知的,且 $W_h=N_h/N$.

设 π_h 代表第 h 层总体中的敏感比例,π是分层总体中具有敏感属性个体所占的比例,且 $\pi=\sum_{h=1}^{L}W_h\pi_h$. 在每层中使用 Warner 随机化装置

(i) 你有敏感属性 A 吗? P

(ii) 你没有敏感属性 A 吗? $1-P$

设 λ_h 表示调查中第 h 层个体回答"是"的概率,则

$$\lambda_h=\pi_hP+(1-\pi_h)(1-P),\quad h=1,2,\cdots,L$$

设调查中在第 h 层有 m_h 个人回答"是",记 $\hat{\lambda}_h=\frac{m_h}{n_h}$. 由 $\lambda_h=\hat{\lambda}_h$ 得到 π_h 的一个矩估计且是极大似然估计:

$$\hat{\pi}_h=\frac{\hat{\lambda}_h-(1-P)}{2P-1}=\frac{\frac{m_h}{n_h}-(1-P)}{2P-1}\qquad(P\neq 0.5)$$

进而得到分层总体敏感比例 π的一个估计量:

$$\hat{\pi}=\sum_{h=1}^{L}W_h\hat{\pi}_h=\frac{\sum_{h=1}^{L}W_h\hat{\lambda}_h-(1-P)}{2P-1}\tag{1}$$

估计量的无偏性

$$E(\hat{\pi})=E(\sum_{h=1}^{L}W_h\hat{\pi}_h)=\sum_{h=1}^{L}W_hE(\hat{\pi}_h)=\sum_{h=1}^{L}W_h\frac{E(\hat{\lambda}_h)-(1-P)}{2P-1}$$

$$=\sum_{h=1}^{L}W_h\frac{\frac{E(m_h)}{n_h}-(1-P)}{2P-1}=\sum_{h=1}^{L}W_h\frac{\lambda_h-(1-P)}{2P-1}=\pi$$

即 $\hat{\pi}$是 π的一个无偏估计量,其中 $E(m_h)=n_h\lambda_h$.

定理 1　估计量 $\hat{\pi}$的方差为

$$V_1(\hat{\pi})=\sum_{h=1}^{L}W_h^2\left[\frac{1}{n_h}\cdot\pi_h(1-\pi_h)+\frac{P(1-P)}{n_h(2P-1)^2}\right]\tag{2}$$

证明

$$V_1(\hat{\pi})=V(\sum_{h=1}^{L}W_h\hat{\pi}_h)=\sum_{h=1}^{L}W_h^2V(\hat{\pi}_h)$$

$$=\sum_{h=1}^{L}W_h^2\frac{V(\hat{\lambda}_h)}{(2P-1)^2}$$

$$=\sum_{h=1}^{L}W_h^2\frac{V(m_h)}{n_h^2(2P-1)^2}$$

$$=\sum_{h=1}^{L}W_h^2\left[\frac{1}{n_h}\cdot\pi_h(1-\pi_h)+\frac{P(1-P)}{n_h(2P-1)^2}\right]$$

1.2　无放回抽样方式下的沃纳模型

将容量为 N 的总体分成 L 层,每层独立使用简单随机无放回抽样抽取容量为 n_h 的样本,则沃纳模

型的估计量：

$$\hat{\pi}=\sum_{h=1}^{L}W_h\hat{\pi}_h=\frac{\sum_{h=1}^{L}W_h\hat{\lambda}_h-(1-P)}{2P-1} \tag{3}$$

即 $\hat{\pi}$是 π的一个无偏估计.

定理 2　估计量 $\hat{\pi}$的方差为

$$V_2(\hat{\pi})\approx\sum_{h=1}^{L}W_h^2\left[\frac{1-f_h}{n_h}\cdot\pi_h(1-\pi_h)+\frac{P(1-P)}{n_h(2P-1)^2}\right] \tag{4}$$

证明

$$V_2(\hat{\pi})=\sum_{h=1}^{L}W_h^2\left[\frac{N_h-n_h}{N_h-1}\cdot\frac{1}{n_h}\cdot\pi_h(1-\pi_h)+\frac{P(1-P)}{n_h(2P-1)^2}\right]$$

$$\approx\sum_{h=1}^{L}W_h^2\left[\frac{1-f_h}{n_h}\cdot\pi_h(1-\pi_h)+\frac{P(1-P)}{n_h(2P-1)^2}\right]$$

其中 $\hat{\lambda}_h=\frac{m_h}{n_h}$, m_h 是第 h 层回答"是"的人数, $f_h=\frac{n_h}{N_h}$是第 h 层的抽样比.

2　效率比较

本节比较比例分配分层抽样与相同样本量下简单抽样的精度.

比例分配是指 $n_h/N_h=n/N=f_n=f$ 即每一层的抽样比相同，此时也有 $n_h/n=N_h/N=W_h$. 这时在分层放回抽样下估计量的方差为：

$$V_{prop1}(\hat{\pi})=\frac{1}{n}\cdot\sum_{h=1}^{L}W_h\pi_h(1-\pi_h)+\frac{P(1-P)}{n(2P-1)^2} \tag{5}$$

分层无放回抽样下估计量的方差为：

$$V_{prop2}(\hat{\pi})\approx\frac{1-f}{n}\cdot\sum_{h=1}^{L}W_h\pi_h(1-\pi_h)+\frac{P(1-P)}{n(2P-1)^2} \tag{6}$$

简单放回和不放回抽样下方差的表达式分别为：

$$V_{srs1}(\hat{\pi})=\frac{1}{n}\cdot\pi(1-\pi)+\frac{P(1-P)}{n(2P-1)^2} \tag{7}$$

$$V_{srs2}(\hat{\pi})\approx\frac{1-f}{n}\cdot\pi(1-\pi)+\frac{P(1-P)}{n(2P-1)^2} \tag{8}$$

定理 3　简单抽样与分层抽样下估计量 $\hat{\pi}$的方差之间的关系式为

$$V_{prop1}(\hat{\pi})=V_{srs1}(\hat{\pi})-\frac{1}{n}\sum_{h=1}^{L}W_h(\pi_h-\pi)^2 \tag{9}$$

$$V_{prop2}(\hat{\pi})\approx V_{srs2}(\hat{\pi})-\frac{1-f}{n}\sum_{h=1}^{L}W_h(\pi_h-\pi)^2 \tag{10}$$

证明　从表达式(5)与(7)及(6)与(8)可以看出要比较两种抽样下估计量的方差大小只需比较 $\sum_{h=1}^{L}W_h\pi_h(1-\pi_h)$ 与 $\pi(1-\pi)$ 即可. 为此我们来求两个式子的差：

$$\pi(1-\pi)-\sum_{h=1}^{L}W_h\pi_h(1-\pi_h)=\pi-\pi^2-\sum_{h=1}^{L}W_h\pi_h+\sum_{h=1}^{L}W_h\pi_h^2$$

$$=-\pi^2+\sum_{h=1}^{L}W_h(\pi_h-\pi+\pi)^2=-\pi^2+\sum_{h=1}^{L}W_h(\pi_h-\pi)^2+\pi^2=\sum_{h=1}^{L}W_h(\pi_h-\pi)^2$$

即有

$$V_{prop1}(\hat{\pi})=V_{srs1}(\hat{\pi})-\frac{1}{n}\cdot\sum_{h=1}^{L}W_h(\pi_h-\pi)^2$$

$$V_{prop2}(\hat{\pi}) \approx V_{srs2}(\hat{\pi}) - \frac{1-f}{n} \cdot \sum_{h=1}^{L} W_h(\pi_h - \pi)^2$$

结论得证.

由表达式(9),(10)可以看出,当每一层的敏感比例 π_h 相等或相差不大时,$\sum_{h=1}^{L} W_h(\pi_h - \pi)^2$ 的值就等于零或很小,则此时分层抽样的效率与简单抽样的效率相等或差别不大,这时分层抽样就无优势可言.

但是当 $\pi_1, \pi_2, \cdots, \pi_L$ 有显著差异时,$\sum_{h=1}^{L} W_h(\pi_h - \pi)^2$ 的值就相对大,从而分层抽样随机化方法显著地优于经典的 Warner 模型,又因为两种抽样是在相同样本量下进行的,所以它们的调查费用是一样的.由此可见如果在调查时根据已掌握的有关调查指标的一些信息或它的相关信息能将总体进行分层,那么分层抽样随机化方法在费用一定的情况下精度上将会有所改进.

3　数值模拟

这一部分我们将构造三个例子从数值上验证上面的比较.

设总体 $N=3000$,分成 5 层,即 $L=5$,采用等容量分法,$N_1=N_2=N_3=N_4=N_5=600$,抽取的各层样本量为 $n_1=n_2=n_3=n_4=n_5=100$,则分层总体的样本量为 $n=500$,层权为 $W=1/5$,沃纳模型装置中敏感问题"你有属性 A 吗?"的比例设置为 $P=0.8$.

3.1　π_h 无差异

$\pi_1=\pi_2=\pi_3=\pi_4=\pi_5=0.2$ 将以上相关数值分别代入相应的方差表达式中得

$$V_{prop1}(\hat{\pi}) = 0.001209 \qquad V_{prop2}(\hat{\pi}) = 0.001156$$
$$V_{srs1}(\hat{\pi}) = 0.001209 \qquad V_{srs2}(\hat{\pi}) = 0.001156$$

即当各层的敏感比例无差异时分层抽样与简单抽样效率相同.

3.2　π_h 有较小差异

$\pi_1=0.1$, $\pi_2=0.2$, $\pi_3=0.3$, $\pi_4=0.15$, $\pi_5=0.25$ 计算得

$$V_{prop1}(\hat{\pi}) = 0.001199 \qquad V_{prop2}(\hat{\pi}) = 0.001147$$
$$V_{srs1}(\hat{\pi}) = 0.001209 \qquad V_{srs2}(\hat{\pi}) = 0.001156$$

3.3　π_h 有显著差异

$\pi_1=0.001$, $\pi_2=0.001$, $\pi_3=0.003$, $\pi_4=0.195$, $\pi_5=0.8$ 计算得

$$V_{prop1}(\hat{\pi}) = 0.001018 \qquad V_{prop2}(\hat{\pi}) = 0.000996$$
$$V_{srs1}(\hat{\pi}) = 0.001209 \qquad V_{srs2}(\hat{\pi}) = 0.001156$$

由以上计算,我们可以看出当 π_h 有较小差异时,$V_{srs}(\hat{\pi})/V_{prop}(\hat{\pi}) \approx 1.008$,即分层抽样与简单抽样的效率差别不大,而当 π_h 有显著差异时,$V_{srs}(\hat{\pi})/V_{prop}(\hat{\pi}) \approx 1.188$,由此可见,分层抽样在精度上比简单抽样有显著提高.

参考文献:

[1]　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.

[2]　孙山泽,孙明举,段钢.二项选择敏感性问题调查的基本方法 [J].数理统计与管理,2000,1～2.

[3]　陈雪如.敏感性问题中的抽样调查方法与均方误差 [J].南京师大学报,1997.12～16.

[4]　Arijit Chaudhuri. Using Randomized Respouse from a Complex Survey to Estimate a Sensitive Proportion in Dichotomous Finite Population [J]. Journal of Statistical Planning and Inference, 2001, 94: 37～42.

[5]　Anthony Y, C, Kuk Asking Sensitive Questions Indirectly [J]. Biometrika, 1990, 77, 2. 436～8.

[6]　Warner S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias [J]. Stat. Assoc.

1965,60:63～69.

A STUDY ON STRATIFIED SAMPLING FOR SURVEYING SENSITIVE QUESTIONS

HONG Zhi-min, YAN Zai-zai

(*School of Basic Sciences,*

Inner Mongolia University of Technology, Huhhot 010062, *PRC*)

Abstract: There have been quite a few surveying methods for sensitive questions. Most of these methods try to increase the surveying efficiency by use of the technique that improves randomized response. In this paper, it is proved that the surveying efficiency can also be improved by reasonable sampling design.

Keywords: sensitive question; stratified sampling; proportional allocation; warner model