·学术讨论·

利用小样本预处理技术提高敏感性问题调查精度*

河北理工大学理学院(063009) 万星火 檀亦丽 王 宏

本文首先尝试用小样本预处理技术——熵判别方法处理和判别敏感性问题调查中不同调查方法的调查结果中的粗大误差,然后用捕获—再捕获方法(CRM)校正调查结果的数值误差,提高敏感性问题抽样调查结果的可靠性。

原理与方法

(一)熵判别法

熵是信息论中的一个基本概念,在信息论中信息量 $I(A_k)$ 表示观测到一个以概率 p_k 发生的事件A_k 的信息。

定义 $I(x_k)=\ln(\frac{1}{p_k})=-\ln p_k$,定义熵为 $H(x_k)=E\{I(x_k)\}=E[-\ln p_k]$,熵与方差之间存在一定的对应关系

$$H(x)=\ln(\alpha\sigma)$$

其中 α 为与分布有关的常数,σ 为标准差。

由上式可以看出,x 取值的分散程度越大,其熵越大,因为熵是对不确定性程度的一种度量,熵越大不确定性越大。由正态法,当置信水平为95%时,确定包含因子为2,相应的不确定度为 $\Delta_x=\pm\frac{1}{2}e^{H(x)}$,由于计算结果往往比真实值小,当样本数据较少时,将上式乘以调整系数 k_1(经验取1.5),所以 Δ_x 的范围调整为$\pm\frac{3}{4}e^{H(x)}$。

由于采样得到的数据是离散的,均信息量的熵也应该是离散熵

$$H(x)=-\sum_{k=1}^{N}p_k\ln p_k$$

小样本样本容量较小,不能用统计频数代替概率估计,此时应采用秩估计的方法进行熵估计。具体方法如下:

(1)将采样数据 $x_1,x_2,\cdots,x_N$ 按由小到大的顺序排成序列 $x(1),x(2),\cdots,x(N)$;

(2)定义秩为

$$r_k=\int_{-\infty}^{x(k)}p(x)dx=\int_{-\infty}^{x(k)}dP(x)=P(x(k))$$

$P(x)$ 为 x 的概率分布函数,$P(x(k))$ 的估计为

$$P(x(k))=\hat{r}_k=\frac{k}{N+1}$$

(3) $H(x)$的估计

$$H(x)=-\sum_{k=1}^{N-1}\ln\left[\frac{\Delta P(x(k))}{\Delta x(k)}\right]\Delta P(x(k))=-\sum_{k=1}^{N-1}\ln\left[\frac{\hat{r}_{k+1}-\hat{r}_k}{x(k+1)-x(k)}\right](\hat{r}_{k+1}-\hat{r}_k)$$

(4)$\Delta_x=\pm\frac{3}{4}e^{H(x)}$

如果$\Delta x_i=x_i-x_{中位数}$超过了Δx 的范围,则判定 x_i 含有粗大误差。

(二)常用的随机化回答技术

1.沃纳模型(Warner model)是1965年由warner提出的,它的提出开创了随机化回答的先河。其设计原则是根据敏感性特征设计两个相互对立的问题,让被调查者按预定的概率从中选一个回答,调查者无权过问被调查者究竟回答的是哪一个问题,从而起到了为被调查者保密的效果。

2.西蒙斯模型的设计思想仍是基于沃纳的随机化回答思想,只是在设计中用无关的问题 Y 代替了沃纳模型中的敏感性问题 A 的对立问题。比如敏感性问题为"你在考试中作弊了吗",沃纳模型中的对立问题是"你在考试中没有作弊吗",在西蒙斯模型中,用一个与敏感性问题无关的问题来代替这一问题,比如"你是四月份出生的吗?"

3.格林伯格双无关问题模型是1973年针对西蒙斯模型 π_y 未知的情况提出的。它更好地利用了原来基本上用于估计 π_y 的样本,与一个敏感性特征 A 相联系,他们考虑了两个非敏感性特征 Y_1,Y_2。设 π_{y1},π_{y2} 分别表示 Y_1,Y_2 在总体中所占的真实比例,且 π_{y1},π_{y2} 是未知的。从总体中用简单随机有放回抽样方式下抽取两个相互独立的有放回的而且是互不相交的简单随机样本,样本容量分别为 n_1,n_2。每一个样本中的被调查者均需回答两个问题,一个是调查者直接询问的无关的非敏感性问题,另一个是被调查者自己使用随机化装置选择的问题,在这两个样本中,设被调查者随机选到敏感性问题的概率均为 p,λ_i^r,λ_i^d,分别表示第 i 个样本中通过随机化回答和直接回答所得到的回答"是"的概率,则得 π_A 的估计量(有偏但具有较好的大样本性质)

(三)捕获—再捕获法(Capture—Recapture Methods,CRM)

* 河北理工大学科学研究基金项目(200629)

本研究假设目标群体(人数已知)具有敏感性特征的人数为 M,将沃纳模型和西蒙斯模型及格林伯格双无关问题模型估计出的具有敏感性特征的人数分别记为 N_W, N_S, N_L。令 $m=\min\{N_W, N_S, N_L\}$,剩余两个模型估计出的具有敏感性特征的人数分别记为 m_1, m_2,两个估计重复的人数即为 m,依照 Chapman 等提出的无偏估计公式估计目标群体(人数已知)具有敏感性特征的人总数为:

$$M=[(m_1+1)(m_2+1)/(m+1)]-1$$

$$\mathrm{Var}(M)=(m_1+1)(m_2+1)(m_1-m)(m_2-m)/(m+1)^2(m+2)$$

回答失真率等于估计的群体总数和具有敏感性特征的人数的差值与估计的群体总数的百分比:

第一来源样本的失真率为

$(M-m_1)/N\times100\%$

第二来源样本的失真率为

$(M-m_2)/N\times100\%$

两来源样本合并后的失真率为:

$[M-(m_1+m_2-m)]/N\times100\%$

符合率等于具有敏感性特征的人数与估计的群体总数的比值,符合率与漏报率的关系是:符合率=1-失真率。

(四)实例

当前大学里有相当一部分学生的学习状况并不理想,基础不够扎实、不能刻苦学习、学习动力不足,再加上就业压力使有些学生产生了厌学情绪,导致考试作弊问题较为突出。这些情况的出现引起了学校、教师的忧虑。为确切了解现在大学生的考试作弊情况,我们对抽样调查结果进行了以下处理。

1.首先在我校利用教务处考试记录分析(方法1)、委婉询问法(方法2)、沃纳模型(方法3)、西蒙斯模型(方法4)、格林伯格双无关问题模型(方法5),分别对作弊这一属性特征的敏感性问题进行了抽样调查。把调查结果用小样本预处理技术——熵判别方法处理和判别考试作弊问题调查中不同方法的结果中的粗大误差。过程如下:

在调查学生考试作弊的问题中,学校学生总数为 $N=12\,485$。我们设计了外形完全一样的卡片80个,其中60个卡片上写上"你考试是否作过弊?",20个卡片上写上"你在考试中没有作弊吗?"。然后放在一盒子里。调查时,由被调查者从盒子里任抽一卡片,根据卡片上的问题做出是或否的回答,回答完毕再把卡片放回盒子。结果为 $n_1=28$

由沃纳模型

$$\hat{\pi}_{AW}=\frac{\lambda-(1-p)}{2p-1}=\frac{\frac{28}{100}-(1-0.75)}{2\times0.75-1}=0.06$$

同理,设计西蒙斯装置及格林伯格双无关问题装置由公式分别得

$$\hat{\pi}_{AS}=[\lambda_1(1-p_2)-\lambda_2(1-p_1)]/(p_1-p_2)=0.0639$$

$$\hat{\pi}_{AL}=\hat{\omega}\hat{\pi}_A(1)+(1-\hat{\omega})\hat{\pi}_A(2)=0.068$$

方法1的结果为0.0379,方法2的调查结果为0.0301,于是五种方法得到的结果可以看成来自某一总体的样本

0.0379, 0.0301, 0.06, 0.0639, 0.068

将所有数据按由小到大的方式排序,且每项都减去最小的数据,并扩大一定的倍数,得到一个新序列

0, 0.78, 2.99, 3.38, 3.79

由熵判别法,计算熵估计值为

$$H(x)=-\sum_{k=1}^{4}\ln\left[\frac{\hat{r}_{k+1}-\hat{r}_k}{x(k+1)-x(k)}\right](\hat{r}_{k+1}-\hat{r}_k)=0.5207$$

计算 $\Delta X=\pm\frac{3}{4}e^{H(x)}=\pm\frac{3}{4}e^{0.5207}=\pm1.2624$。

又因为 $x_{中位数}=2.99$,这样判断出方法1中的0.0379和方法2中的0.0301含有粗大误差。将方法1、2结果剔除。

2.用捕获一再捕获方法校正此问题抽样调查中产生的偏差。把剔出粗大误差后的三个结果用捕获一再捕获法技术处理,具体过程如下:

在捕获一再捕获法中,$m_1=0.0639\times12485\approx798$, $m_2=0.068\times12485\approx849$, $m=0.06\times12485\approx749$,由公式

$$M=[(m_1+1)(m_2+1)/(m+1)]-1=906$$

$$\mathrm{Var}(M)=(m_1+1)(m_2+1)(m_1-m)(m_2-m)/(m+1)^2(m+2)=8.049$$

总漏报率$=[M-(m_1+m_2-m)]/M\times100\%=0.064\%$

结　　论

敏感性问题的统计调查是抽样调查经常会遇到的问题,和其他一般统计调查的问题一样,如何提高敏感性问题抽样调查结果的可靠性是调查2倍者关心的问题。通过本研究发现用小样本预处理技术一熵判别方法处理和判别敏感性问题调查中不同调查方法的调查结果中的粗大误差。然后用捕获一再捕获方法校正调查结果的数值误差,提高敏感性问题抽样调查结果的可靠性。该方法设计合理,简便易行,具有较广泛的实用价值且不必花费较多的人力物力,但在使用时必须注意其使用的前提条件,不能盲目地套用公式。

参 考 文 献

1. Chaudhuri A. Randomized Response Theory and Technique Marcel Pekker, 1988.
2. Hook EB, Regal RR. Capture-recapture Methods. Lancet, 1992.
3. 孙山泽. 抽样调查. 北京大学出版社, 2004.
4. 施锡铨. 抽样调查的理论和方法. 上海财经大学出版社, 1999.