# 应用 CRM 估计敏感性问题调查中的偏差

檀亦丽 万星火 纪 楠
（河北理工大学理学院，河北唐山，063009）

【摘 要】本文探索用捕获-再捕获方法（Capture-Recapture Methods，CRM）估计敏感性问题抽样调查中产生的偏差。利用沃纳（Warner）模型和西蒙斯（Simmons）模型及格林伯格双无关问题模型的信息依靠 CRM 校正因被调查者回答失真而造成的数值误差，获得较准确的调查结果，并将结论用于对大学生考试作弊这一属性特征的敏感性问题的抽样调查得到了较客观的调查结果。

【关键词】捕获-再捕获方法；随机化回答技术；敏感性问题

## 0 引 言

在当今的社会经济调查等各种统计调查中，经常会遇到各种各样的敏感性问题。所谓敏感性问题，是指与个人或单位的隐私或私人利益有关而不便向外界透露的问题。比如，个人或单位是否偷税漏税及数额的多少；考生在考试中是否有作弊行为，吸毒、赌博；是否有犯罪行为及类似的为社会所不赞成的各种事件。对于这类敏感性问题，抽样调查一般采用一种科学可行的技术 —— 随机化回答技术（Randomized Response Technique 简记为 RRT）。随机化回答是指在调查中使用特定的随机化装置，使得被调查者以预定的概率 $p$ 来回答敏感性问题。这一技术的宗旨就是最大限度地为被调查者保守秘密，从而取得被调查者的信任。尽管如此，和其他一般统计调查的问题一样，调查结果的可靠性仍然是调查者关心的问题。本文探索用捕获-再捕获方法（CRM）估计敏感性问题抽样调查中产生的偏差，并利用沃纳模型和西蒙斯模型及格林伯格双无关问题模型现有的信息依靠 CRM 校正因漏报而造成的数值误差，提高敏感性问题抽样调查结果的可靠性。

## 1 原理与方法

（1）随机化回答技术

沃纳模型（Warner Model）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。

其模型的设计及参数的估计的思想为设总体可分为互不相容的两类：具有敏感性特征的

一类 $A$ 与不具敏感性特征的一类 $\overline{A}$。即总体中的每一个体或者具有敏感性特征（属于 $A$），或者不具有敏感性特征（属于 $\overline{A}$）。我们的目的是估计具有敏感性特征（属于 $A$）的人在总体中所占的比例 $\pi_A$。

设总体容量为 $N$，在简单随机无放回抽样下从总体中抽得 $n$ 个样本，然后对这 $n$ 个样本进行随机化回答调查。所使用的随机装置描述如下，外形相同的卡片上分别写有问题："你属于 $A$ 吗？"与"你属于 $\overline{A}$ 吗？"如（"你在考试中作弊了吗？"与"你在考试中没有作弊吗？"）以预定的比例 $p$ 混合后放入一盒子中。调查时，被调查者从盒子中任拿出一卡片，根据卡片上的问题进行回答。回答完后仍把卡片放回盒子，供其他被调查者使用。设 $\pi_A$ 是具有敏感性特征的人所占的比例，$p$ 是写有问题"你属于 $A$ 吗？"的卡片所占的比例。设调查结果中有 $n_1$ 个人回答"是"，有 $n-n_1$ 个人回答"否"，则 $\pi_A$ 的极大似然估计为：

$$\hat{\pi}_A=\frac{\hat{\lambda}-(1-p)}{2p-1} \tag{①}$$

式中，$\hat{\lambda}=\frac{n_1}{n}$，$n_1$ 是回答"是"的人数。$f=\frac{n}{N}$ 方差的无偏估计量为：

$$\hat{\mathrm{Var}}(\hat{\pi}_A)\approx(1-f)\cdot\frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n}+\frac{p(1-p)}{n(2p-1)^2}\leqslant a$$

$$n=\left[\frac{1}{4\left(a+\frac{1}{4N}\right)}+\frac{p(1-p)}{(2p-1)^2\left(a+\frac{1}{4N}\right)}\right]+1 \tag{②}$$

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？"

（2）模型的设计与参数的估计

模型的基本设计为：制作一个能产生两种实验结果的随机化装置，如两套外形一样的卡片，一套卡上写有敏感性问题，如"你属于 A 吗？"（比如"你在考试中作弊了吗？"）不妨称为 1 号卡片。另一套卡片写有无关问题"你属于 Y 吗？"其中 Y 是与 A 无关的非敏感性问题，如"你是四月份出生的吗？"称此卡片为 2 号卡片。将 1 号卡片和 2 号卡片按预定比例混合后，放入一盒子中，调查时，被调查者只需从盒子中任意抽取一张卡片，根据卡片上的问题作出真实的回答，当然调查员无权知道卡片上写的究竟是哪一个问题。

在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 $n_1$，$n_2$。对于第一个样本，随机化装置出现 1 号卡片的概率为 $p_1$，2 号卡片出现的概率为 $1-p_1$；第二个样本，1 号卡片出现的概率为 $p_2(p_1\neq p_2)$，2 号卡片出现的概率为 $1-p_2$。$\pi_{AU}$ 是具有敏感性特征 A 的人所占比例，$\pi_y$（未知）是具有无关特征 Y 的人所占比例。$\pi_{AU}$ 的一个无偏估计量为：

$$\hat{\pi}_{AU}=\frac{\hat{\lambda}_1(1-p_2)-\hat{\lambda}_2(1-p_1)}{p_1-p_2} \tag{③}$$

其中 $$\hat{\lambda}_i = \frac{n_i}{n}$$

格林伯格双无关问题模型是针对西蒙斯模型 $\pi_y$ 未知的情况提出的。它更好地利用了原来基本上用于估计 $\pi_y$ 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 $\pi_{y1}$，$\pi_{y2}$ 分别表示 $Y_1$，$Y_2$ 在总体中所占的真实比例，且 $\pi_{y1}$，$\pi_{y2}$ 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 $p$，$\lambda_i^r$，$\lambda_i^d$ 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 $\pi_A$ 的估计量（有偏但具有较好的大样本性质）

$$\hat{\pi}_{AF} = \hat{\omega}\hat{\pi}_A(1) + (1-\hat{\omega})\hat{\pi}_A(2) \qquad ④$$

式中 $$\hat{\omega} = \frac{\hat{\sigma}_{22} - \hat{\sigma}_{12}}{\hat{\sigma}_{11} + \hat{\sigma}_{22} - 2\hat{\sigma}_{12}}$$

其中

$$\hat{\sigma}_{11} = p^{-2}\left[\frac{\hat{\lambda}_1^r(1-\hat{\lambda}_1^r)}{n_1} + \frac{(1-p^2)\hat{\lambda}_2^d(1-\hat{\lambda}_2^d)}{n_2}\right]$$

$$\hat{\sigma}_{22} = p^{-2}\left[\frac{\hat{\lambda}_2^r(1-\hat{\lambda}_2^r)}{n_2} + \frac{(1-p^2)\hat{\lambda}_1^d(1-\hat{\lambda}_1^d)}{n_1}\right]$$

$$\hat{\sigma}_{12} = -(1-p)p^{-2}\left[\frac{\hat{\lambda}_1^{rd} - \hat{\lambda}_1^r\hat{\lambda}_1^d}{n_1} + \frac{\hat{\lambda}_2^{rd} - \hat{\lambda}_2^r\hat{\lambda}_2^d}{n_2}\right]$$

$$\hat{\pi}_A(1) = \frac{\hat{\lambda}_1^r - (1-p)\hat{\lambda}_2^d}{p}$$

$$\hat{\pi}_A(2) = \frac{\hat{\lambda}_2^r - (1-p)\hat{\lambda}_1^d}{p}$$

其中 $$\hat{\lambda}_1^r = \frac{n_{11}^r}{n_1}, \qquad \hat{\lambda}_2^r = \frac{n_{21}^r}{n_2}$$

其中，$n_{i1}^r$ 是第 $i$ 个样本中随机化回答得到的回答“是”的人数，$n_{i1}^d$ 是第 $i$ 个样本中直接回答得到的回答“是”的人数，$\hat{\lambda}_1^r$，$\hat{\lambda}_1^d$，$\hat{\lambda}_2^r$，$\hat{\lambda}_2^d$ 分别是 $\lambda_1^r$，$\lambda_1^d$，$\lambda_2^r$，$\lambda_2^d$ 的无偏估计，$\hat{\lambda}_i^{rd}$ 是样本（$i=1,2$）中的两个问题都回答“是”的概率。

（3）捕获-再捕获法（Capture-Recapture Methods，CRM）的应用

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 $m = \min\{N_w, N_s, N_L\}$，其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：

$$M=\frac{(m_1+1)(m_2+1)}{m+1}-1 \qquad ⑤$$

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

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：

$$\frac{M-m_1}{N}\times 100\%$$

第二来源样本的失真率为：

$$\frac{M-m_2}{N}\times 100\%$$

两来源样本合并后的失真率为：

$$\frac{M-(m_1+m_2-m)}{N}\times 100\% \qquad ⑥$$

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。

## 2 实例与结论

目前，一些大学里有相当一部分大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，有些学生甚至有厌学情绪，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。为了确切了解现在大学生的考试作弊情况，我们在我校利用上述三个模型对作弊这一属性特征的敏感性问题进行了抽样调查。并利用捕获-再捕获方法校正因漏报而造成的数值误差，获得较准确的调查结果。过程如下：

在调查学生考试作弊的问题中，由②式计算得（设学校学生总数为 $N=12\,485$）$n=100$。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①可得

$$\hat{\pi}_A=\frac{\hat{\lambda}-(1-p)}{2p-1}=\frac{\frac{28}{100}-(1-0.75)}{2\times 0.75-1}=0.06$$

同理，设计西蒙斯装置及格林伯格双无关问题装置由公式③、④分别得：

$$\hat{\pi}_{AU}=\frac{\hat{\lambda}_1(1-p_2)-\hat{\lambda}_2(1-p_1)}{p_1-p_2}=0.065$$

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

因此，三个模型估计出我校大学生的考试作弊人数分别为：

$$N_w=12\,485\times 0.06\approx 749,\quad N_S=12\,485\times 0.064\approx 799,\quad N_L=12\,485\times 0.068\approx 849$$

于是捕获-再捕获法技术中，$m_1 = 799$，$m_2 = 849$，$m = 749$，由公式⑤、⑥得

$$M = \frac{(m_1+1)(m_2+1)}{m+1} - 1 = 906$$

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

总漏报率

$$\frac{M-(m_1+m_2-m)}{M} \times 100\% = 0.77\%$$

由结果可知，通过应用 CRM，调整了调查数据的偏差，获得较客观的调查结果，与学校管理部门掌握的情况基本相符。但是应该注意 CRM 是建立在一定假设的条件下：（1）所有的记录，目标群体必须有相同的解释与定义。（2）两样本中共同的个体能被鉴别。（3）研究期间研究群体总数（在校学生数）近似一个常数（即封闭性假设）。（4）两样本是独立的，本研究三个模型是独立抽样收集资料的，所有个体都有同等概率被不同样本所捕获，如果每个样本对不同个体的捕获概率不同，可以按影响捕获概率不等的主要因素进行分层再汇总等方法进行估计。（5）重复信息不足会导致估计结果的不准确，本研究两个独立来源样本重复符合率分别为 93.7%和 88.2%，满足符合率应大于 60%的要求。

敏感性问题的统计调查是抽样调查经常会遇到问题，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者关心的问题。通过本研究发现利用捕获-再捕获方法（CRM）估计敏感性问题抽样调查中产生的偏差。并利用沃纳模型和西蒙斯模型及格林伯格双无关问题模型现有的信息依靠 CRM 校正因漏报而造成的数值误差，可以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[M]. 北京：北京大学出版社，2004

4 施锡铨. 抽样调查的理论和方法[M]. 上海：上海财经大学出版社，1999

5 孙山泽等. 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）


# Estimate the Deviation in the Sampling Survey Aiming at Sensitive Question by CRM


Tan Yi-li，Wan Xing-huo，Ji Nan

（School of Science，Hebei Polytechnic University，Tangshan 063009，P.R.China）



【Abstract】 This paper estimate the deviation in the sampling survey aiming at sensitive question

by CRM (Capture-Recapture Methods). We correct the value error which informant distort the answer by using the CRM which information comes from the Warner model, Simmons model and Lind-beg model. The conclusion is applied for the sampling survey aiming at sensitive question of attribute character for university students' cheating in a test.

【**Key words**】capture-recapture methods; randomized response technique; sensitive question