第二军医大学学报 2009年2月第30卷第2期 http://www.ajsmmu.cn
Academic Journal of Second Military Medical University, Feb. 2009, Vol. 30, No. 2

DOI:10.3724/SP.J.1008.2009.00170

·论 著·

# 敏感问题随机应答技术模型分层整群抽样下参数的估计

高 歌*,范玉波,王 冕
苏州大学公共卫生学院卫生统计学教研室,苏州 215123

[摘要] **目的**:为敏感性问题提供科学的较复杂抽样调查方法及其统计量的计算公式。**方法**:使用Cochran W.G.的经典抽样理论,2种随机应答技术(RRT)模型,全概率公式,均数、方差的性质等理论与方法进行公式推导。**结果**:推导出二分类敏感问题双无关问题RRT模型分别在整群抽样、分层整群抽样下总体比例的估计量及其估计方差的计算公式;推导出数量特征敏感问题加法RRT模型分别在整群抽样、分层整群抽样下总体均数的估计量及其估计方差的计算公式;并在苏州大学学生婚前性行为、考试作弊次数的调查中取得了成功的应用效果。**结论**:本研究提供的敏感性问题2种RRT模型的整群抽样、分层整群抽样调查方法与统计量的计算公式信度较高。

[关键词] 敏感问题;RRT模型;分层整群抽样;参数估计;估计方差
[中图分类号] R 195 [文献标志码] A [文章编号] 0258-879X(2009)02-0170-08

## Estimation of parameters in stratified cluster sampling on randomized response technique for sensitive question survey

GAO Ge*, FAN Yu-bo, WANG Mian
Department of Health Statistics, Public Health School, Soochow University, Suzhou 215123, China

[**ABSTRACT**] **Objective**: To explore a scientific, complicated sampling method and corresponding formulas for sensitive question survey. **Methods**: Cochran W. G.'s classic sampling theories, two randomized response technique (RRT) models, total probability formulas, and properties of mean and variance were used in this study. **Results**: Formulas for the estimation of the population proportion and its variance for dichotomous sensitive questions on Greenberg model in cluster sampling and stratified cluster sampling were deduced. Formulas for the estimation of the population mean and its variance for quantitative sensitive question on the additive constant model in cluster sampling and stratified cluster sampling were deduced. Our survey methods and formulas were successfully applied in surveying the pre-marriage sex and cheating in examination in Soochow University. **Conclusion**: Our survey methods and formulas on two RRT models for sensitive question survey in cluster sampling and stratified cluster sampling have high reliablility.

[**KEY WORDS**] sensitive question; RRT model; stratified cluster sampling; estimation of parameters; estimated variance
[Acad J Sec Mil Med Univ, 2009, 30(2): 170-177]

所谓敏感性问题是指高度私人机密性或大多数人认为不便在公开场合表态及陈述的问题,例如吸毒、赌博、卖淫、酒后驾驶、个人收入、逃税、婚前性行为、性病、艾滋病、同性恋倾向等。敏感问题可分为分类特征敏感问题和数量特征敏感问题,分类特征敏感问题可进一步分为二分类敏感问题(实际中较多发生)和多分类敏感问题。对于敏感性问题的调查,若采用直接提问的方式,被调查者为了保护自己的隐私或出于其他目的,往往会拒绝回答或故意说谎,使调查结果产生偏倚[1]。为了防止偏倚,随机应答技术(randomized response technique, RRT)被认为是能有效保护被调查者隐私,提高其真实回答率的一种方法[2]。RRT由美国社会学家Warner于1965年首先提出,并设计了一个与敏感性问题A相对立问题的两个相关问题模型[3];为了进一步消除被调查者对隐私暴露的担心,1967年Simmons提出一个(与敏感问题)无关非敏感问题的模型[4];1971年Greenberg等为了解决Simmons模型中非敏感问题特征比例未知的问题,考虑了两个(与敏感问题A)无关的非敏感问题B、C,称双无关问题模型(Greenberg模型)[3];以上3种RRT模型均是对二分类敏感问题而设计。对数量特征敏感问题,较好

[收稿日期] 2008-07-06 [接受日期] 2008-12-03
[基金项目] 国家自然科学基金(30571620). Supported by National Natural Science Foundation of China(30571620).
[作者简介] 高 歌,教授,博士生导师.
* 通讯作者(Corresponding author). Tel:0512-65880078, E-mail:gaoge@suda.edu.cn

的 RRT 模型是加法和乘法模型,具有设计简单、所需样本较小、抽样误差较小等优点[5]。

目前国内外对敏感问题 RRT 的抽样调查设计研究,仅局限于简单随机抽样,实际应用也局限于小范围特殊人群小样本的简单随机抽样,或将敏感问题 RRT 模型的复杂抽样方法调查资料误用 RRT 简单随机抽样调查的有关公式来统计分析,而且也极少对敏感问题 RRT 模型抽样调查的信度与效度进行评价。

本文对的二分类敏感问题双无关问题 RRT 模型下较复杂的常用整群抽样、分层整群抽样调查方法进行了设计,推导出二分类敏感问题双无关问题 RRT 模型在整群抽样、分层整群抽样下总体比例的估计量及其估计方差的计算公式;对数量特征敏感问题加法 RRT 模型下较复杂的常用整群抽样、分层整群抽样调查方法进行了设计,推导出数量特征敏感问题加法 RRT 模型在整群抽样、分层整群抽样下总体均数的估计量及其估计方差的计算公式;结合苏州大学学生婚前性行为、考试作弊的调查实例,对二分类敏感问题双无关问题 RRT 模型、数量特征敏感问题加法 RRT 模型的整群抽样、分层整群抽样取得了成功的应用效果。

## 1　调查方法

### 1.1　敏感问题总体比例的调查

1.1.1　*双无关问题的 RRT 模型*　双无关问题的 RRT 模型(Greenberg 模型)[3]:随机抽取 2 个相互独立且互不相交的样本,称为样本 1 和样本 2;设计一随机化装置,如:将大小、重量、质感相同的若干数量的红球和白球按一定比例混合放入袋中,每个被抽中的人有放回地从袋中随机抽出一球。对样本 1 中的每个人:抽得红球时回答是否具有敏感特性 A,抽得白球时回答是否具有无关非敏感特性 B,全部回答是否具有无关非敏感特性 C。对样本 2 中的每个人:抽得红球时回答是否具有敏感特性 A,抽得白球时回答是否具有无关非敏感特性 C,全部回答是否具有无关非敏感特性 B。

1.1.2　*双无关问题 RRT 模型的整群抽样*　整群抽样的优点是抽样框要求简单,调查单元比较集中,调查工作的组织和进行比较方便,调查每个基本单元的费用降低,使得同样的费用可调查更多的基本单元。作为一种经济实用、实施方便的抽样方法,在卫生工作与医学科研中被广泛应用。双无关问题 RRT 模型的整群抽样可分为 3 个步骤:第一步将总体划分为群(一级单元),各群由二级单元组成;第二步以群为抽样单元,从总体中随机抽取一部分群;第三步分别对每个抽中群,将其全部二级单元随机分成样本含量相等的两部分(样本 1 和样本 2),对两样本中全部二级单元采用双无关问题 RRT 模型(见 1.1.1)进行二分类敏感问题的调查。

1.1.3　*双无关问题 RRT 模型的分层整群抽样*　分层抽样的主要优点是减少抽样误差。双无关问题 RRT 模型分层整群抽样可分为 4 个步骤:第一步将总体根据某项或某几项特征划分成若干层;第二步分别将各层划分为群(一级单元),各群由二级单元组成;第三步以群为抽样单元,分别从各层随机抽取一部分群;第四步分别对每个抽中群,将其全部二级单元随机分成样本含量相等的两部分(样本 1 和样本 2),对两样本中全部二级单元采用双无关问题 RRT 模型(见 1.1.1)进行二分类敏感问题的调查。

### 1.2　敏感问题总体均数的调查

1.2.1　*数量特征敏感问题加法 RRT 模型*　数量特征敏感问题加法 RRT 模型[3]:设计一套随机装置,如在小布袋中放置大小、重量、质感相同的小球 10 个,分别贴有 0、1、2、…、9 的数字标签各 1 个。被调查者从小布袋中有放回地随机抽取 1 个小球,将自己数量特征敏感问题的数值与抽中小球上的数值相加的结果填入调查表中。

1.2.2　*数量特征敏感问题加法 RRT 模型的整群抽样*　数量特征敏感问题加法 RRT 模型的整群抽样可分为 3 个步骤:第一步将总体划分为群(一级单元),各群由二级单元组成;第二步以群为抽样单元,从总体中随机抽取一部分群;第三步分别对各抽中群的每个二级单元,采用加法 RRT 模型(见 1.2.1)进行数量特征敏感问题的调查。

1.2.3　*数量特征敏感问题加法 RRT 模型的分层整群抽样*　数量特征敏感问题加法 RRT 模型的分层整群抽样可分为 4 个步骤:第一步将总体根据某项或某几项特征划分成若干层;第二步分别将各层划分为群(一级单元),各群由二级单元组成;第三步以群为抽样单元,分别从各层随机抽取一部分群;第四步分别对各抽中群的每个二级单元,采用加法 RRT 模型(见 1.2.1)进行数量特征敏感问题的调查。

## 2　公式推导

### 2.1　敏感问题总体比例的调查

2.1.1　*整群抽样*　假定总体划分成 $N$ 个群,第 $i$ 个群包含 $M_i$ 个二级单元,随机抽取 $n$ 个群。

总体比例的估计量及其估计方差:假定第 $i$ 个样本群具有敏感特性 A 的二级单元数,比例分别为

$a_i$、$\pi_i$($i=1,2,\cdots,n$)，总体中具有特性 A 的比例为 $\pi$。由 Wang 等[6]给出的结果，整群抽样总体比例 $\pi$ 的估计量为：

$$\hat{\pi}=\frac{\sum_{i=1}^{n}a_i}{\sum_{i=1}^{n}M_i}\overset{M_i=M}{=}\frac{1}{n}\sum_{i=1}^{n}\pi_i \tag{1}$$

由 Cochran[7]给出的结果，统计量 $\hat{\pi}$ 的估计方差为：

$$v(\hat{\pi})=\frac{1-f}{n\bar{M}^2}\cdot\frac{\sum_{i=1}^{n}M_i^2(\pi_i-\hat{\pi})^2}{n-1}$$

$$\overset{M_i=M}{=}\frac{1-f}{n(n-1)}\sum_{i=1}^{n}(\pi_i-\hat{\pi})^2 \tag{2}$$

其中 $\bar{M}=\sum_{i=1}^{n}M_i/n$ 是样本中每群包含的平均二级单元数，$f=\frac{\sum_{i=1}^{n}M_i}{\sum_{i=1}^{N}M_i}\overset{M_i=M}{=}\frac{n}{N}$ 是抽样比。

$\pi_{Bi}$(由样本 2 的调查估计)、$\pi_{Ci}$(由样本 1 的调查估计)分别表示第 $i$ 群 2 个无关问题 $B_i$、$C_i$ 比例，第 $i$ 群敏感问题的发生比例为 $\pi_i$，红球的比例为 $P$。假定第 $i$ 群样本 1 中回答“是”的人数为 $m_{1i}$，样本 2 中回答“是”的人数为 $m_{2i}$。

则根据全概率公式[8]，第 $i$ 群样本 1 中随机回答“是”的比例：

$$\lambda_{1ri}=P\cdot\pi_i+(1-P)\cdot\pi_{Bi}\qquad(\hat{\lambda}_{1ri}=2m_{1i}/M_i)$$

则由第 $i$ 群样本 1 得 $\pi_i$ 的估计量 $\hat{\pi}_{1i}$ 为：

$$\hat{\pi}_{1i}=\frac{\hat{\lambda}_{1ri}-(1-P)\pi_{Bi}}{P},\ i=1,2,\cdots,n$$

当 $\pi_{Bi}$ 为第 $i$ 群的参数(在整群抽样中对非敏感问题的调查较容易获取)时，根据二项分布的方差计算公式及方差的性质[9]，可得：

$$Var(\hat{\pi}_{1i})=\frac{\lambda_{1ri}(1-\lambda_{1ri})}{P^2M_i/2},\ i=1,2,\cdots,n$$

同理，由第 $i$ 群样本 2 得 $\pi_i$ 的估计量 $\hat{\pi}_{2i}$ 及其方差分别为：

$$\hat{\pi}_{2i}=\frac{\hat{\lambda}_{2ri}-(1-P)\pi_{Ci}}{P}$$

$$\hat{\lambda}_{2ri}=2m_{2i}/M_i,\ i=1,2,\cdots,n$$

$$Var(\hat{\pi}_{2i})=\frac{\lambda_{2ri}(1-\lambda_{2ri})}{P^2M_i/2},\ i=1,2,\cdots,n$$

第 $i$ 群敏感问题发生比例 $\pi_i$ 的计算公式[3]为：

$$\pi_i=\omega_i\times\hat{\pi}_{1i}+(1-\omega_i)\times\hat{\pi}_{2i},\ i=1,2,\cdots,n \tag{3}$$

其中：

$$\omega_i=\frac{Var(\hat{\pi}_{2i})-Var(\hat{\pi}_{1i},\hat{\pi}_{2i})}{Var(\hat{\pi}_{1i})+Var(\hat{\pi}_{2i})-2Var(\hat{\pi}_{1i},\hat{\pi}_{2i})}$$

式中 $Var(\hat{\pi}_{1i},\hat{\pi}_{2i})$ 是 $\hat{\pi}_{1i}$、$\hat{\pi}_{2i}$ 的协方差。实际中很容易选择发生比例相等的两个非敏感问题，使 $\pi_{Bi}=\pi_{Ci}$，从而 $\lambda_{1ri}=\lambda_{2ri}$，进一步有 $Var(\hat{\pi}_{1i})=Var(\hat{\pi}_{2i})$，得 $\omega_i=0.5$。

于是：$\alpha_i=M_i\pi_i$，$i=1,2,\cdots,n$

2.1.2　分层整群抽样　假定总体划分成 $L$ 层，第 $h$ 层包含 $N_h$ 个群(一级单元)，$h$ 层第 $i$ 群包含 $M_{ih}$ 个二级单元，总体共包含 $N$ 个二级单元，从 $h$ 层随机抽取 $n_h$ 个群。

$h$ 层总体比例的估计及其估计方差：假定 $h$ 层第 $i$ 个群二级单元中敏感问题的发生比例为 $\pi_{ih}$，有 $a_{ih}$ 个二级单元具有特性 A。由(1)式得 $h$ 层总体比例的估计量为：

$$\hat{\pi}_h=\frac{\sum_{i=1}^{n_h}a_{ih}}{\sum_{i=1}^{n_h}M_{ih}}\overset{M_{ih}=M_h}{=}\frac{1}{n_h}\sum_{i=1}^{n_h}\pi_{ih},\ h=1,2,\cdots,L \tag{4}$$

由式(2)得 $h$ 层 $\hat{\pi}_h$ 的方差 $V(\hat{\pi}_h)$ 估计量：

$$v(\hat{\pi}_h)=\frac{1-f_h}{n_h\bar{M}_h^2}\cdot\frac{\sum_{i=1}^{n_h}M_{ih}^2(\pi_{ih}-\hat{\pi}_h)^2}{n_h-1}$$

$$\overset{M_{ih}=M_h}{=}\frac{1-f_h}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\pi_{ih}-\hat{\pi}_h)^2$$

$$h=1,2,\cdots,L \tag{5}$$

式(4)、(5)中各符号均是在式(1)、(2)中各符号添加下标 $h$ 而成，代表 $h$ 层式(1)、(2)中各符号相对应的意义。

总体比例的估计量为[7]：

$$\hat{\pi}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}M_{ih}\hat{\pi}_h}{N}=\sum_{h=1}^{L}W_h\hat{\pi}_h \tag{6}$$

其中 $W_h=\sum_{i=1}^{N_h}M_{ih}/N$ 是按二级单元计算的 $h$ 层的相对大小。因各层的样本是独立的，对(6)式根据方差的性质有[9]：

$$V(\hat{\pi})=\sum_{h=1}^{L}W_h^2V(\hat{\pi})_h \tag{7}$$

按(5)式估计(7)式中的 $V(\hat{\pi}_h)$。

$\hat{\pi}_{1ih}$、$\hat{\pi}_{2ih}$ 分别代表 $h$ 层第 $i$ 群样本 1、样本 2 对 $h$ 层第 $i$ 群比例 $\pi_{ih}$ 的估计量；$\hat{\lambda}_{1rih}$、$\hat{\lambda}_{2rih}$ 分别代表 $h$ 层第 $i$ 群样本 1、样本 2 中随机回答“是”的样本比例；$\pi_{Bih}$、$\pi_{Cih}$ 分别代表 $h$ 层第 $i$ 群非敏感问题 $B_{ih}$、$C_{ih}$ 的比例。由(3)式得 $h$ 层第 $i$ 群总体比例的计算公式：

$$\pi_{ih}=\omega_{ih}\times\hat{\pi}_{1ih}+(1-\omega_{ih})\times\hat{\pi}_{2ih}$$

$$i=1,2,\cdots,n_h;h=1,2,\cdots,L \tag{8}$$

其中：$\hat{\pi}_{1ih}=\frac{\hat{\lambda}_{1rih}-(1-P)\pi_{Bih}}{P}$

$$\hat{\pi}_{2ih}=\frac{\hat{\lambda}_{2rih}-(1-P)\pi_{Cih}}{P}$$

当两样本人数相等、$\lambda_{1rih}$ 与 $\lambda_{2rih}$ 相等(实际中很容易选择发生比例相等的两个非敏感问题 $B_{ih}$、$C_{ih}$)时,$\omega_{ih}=0.5$。

于是:$a_{ih}=M_{ih}\pi_{ih}$, $i=1,2,\cdots,n_h$;$h=1,2,\cdots,L$

2.2　敏感问题总体均数的调查

2.2.1　整群抽样　由 Wang 等[6]给出的结果,整群抽样的总体均数 $\mu$ 的估计量为:

$$\hat{\mu}=\frac{\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}M_i}\overset{M_i=M}{=}\frac{1}{n}\sum_{i=1}^{n}\mu_i\ (y_i\text{ 为 }i\text{ 群数值之和})\quad(9)$$

$$v(\hat{\mu})=\frac{1-f}{n\bar{M}^2}\frac{\sum_{i=1}^{n}M_i^2(\mu_i-\hat{\mu})^2}{n-1}$$

$$\overset{M_i=M}{=}\frac{1-f}{n(n-1)}\sum_{i=1}^{n}(\mu_i-\hat{\mu})^2\quad(10)$$

用 $\mu_i$ 表示第 $i$ 群敏感问题特征变量的均数,$\mu_{iZ}$ 表示第 $i$ 群所有回答值的均数。$\mu_Y$ 表示随机化装置中所有随机数的均数,由均数的性质[9]可得:

$\mu_{iZ}=\mu_i+\mu_Y$, $i=1,2,\cdots,n$

则:$\mu_i=\mu_{iZ}-\mu_Y$, $i=1,2,\cdots,n$　(11)

于是:$y_i=M_i\mu_i$, $i=1,2,\cdots,n$

2.2.2　分层整群抽样　将 $h$ 层作为一个(子)总体,由式(9)、(10)式得 $h$ 层总体均数 $\mu_h$ 的估计量及其估计方差为:

$$\hat{\mu}_h=\frac{\sum_{i=1}^{n_h}y_{ih}}{\sum_{i=1}^{n_h}M_{ih}}\overset{M_{ih}=M_h}{=}\frac{1}{n_h}\sum_{i=1}^{n_h}\mu_{ih},\ h=1,2,\cdots,L\quad(12)$$

$$v(\hat{\mu}_h)=\frac{1-f_h}{n_h\bar{M}_h^2}\frac{\sum_{i=1}^{n_h}M_{ih}^2(\mu_{ih}-\hat{\mu}_h)^2}{n_h-1}$$

$$\overset{M_{ih}=M_h}{=}\frac{1-f_h}{n_h(n_h-1)}\sum_{i=1}^{n_h}(\mu_{ih}-\hat{\mu}_h)^2,$$

$$h=1,2,\cdots,L\quad(13)$$

式(12)、(13)是在式(9)、(10)中各符号添加下标 $h$ 而成,表示 $h$ 层内式(9)、(10)中各符号相对应的意义。

根据 Cochran[7]给出的结果,在式(6)中将 $\hat{\mu}_h$ 代替 $\hat{\pi}_h$,得总体均数 $\mu$ 的估计量:

$$\hat{\mu}=\frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}M_{ih}\hat{\mu}_h}{N}=\sum_{h=1}^{L}W_h\hat{\mu}_h\quad(14)$$

因各层的样本是独立的,对(14)式根据方差的性质[9]有:

$$V(\hat{\mu})=\sum_{h=1}^{L}W_h^2V(\hat{\mu}_h)\ [V(\hat{\mu}_h)\text{为 }\hat{\mu}_h\text{ 的方差}]\quad(15)$$

按(13)式估计(15)式中的 $V(\hat{\mu}_h)$,得 $v(\hat{\mu})$ 的对应估计方差 $v(\hat{\mu})$。

用 $\mu_{ih}$ 表示 $h$ 层第 $i$ 群敏感问题特征变量的均数,$\mu_{ihZ}$ 表示 $h$ 层第 $i$ 群所有回答值的均数。$\mu_Y$ 表示随机化装置中所有随机数的均数,由均数的性质[9]可得:

$\mu_{ihZ}=\mu_{ih}+\mu_Y$

则:$\mu_{ih}=\mu_{ihZ}-\mu_Y$, $i=1,2,\cdots,n_h$;$h=1,2,\cdots,L$　(16)

于是:$y_{ih}=M_{ih}\mu_{ih}$

## 3　应用实例

以 2007 年苏州大学新校区全体在校学生为调查总体,划分为本科生(1 层)、研究生(2 层)两个层,本科生共 9 689 人,研究生共 1 890 人,$W_1=9\,689/(9\,689+1\,890)\doteq0.84$,$W_2\doteq0.16$。以班为群,使用大班拆小班、小班并大班的做法,使各层内各班学生数近似相等。采用双无关问题 RRT 模型的分层整群抽样(各层内即为整群抽样),分别随机抽取本科班 20 个共 1 080 人、研究生班 18 个共 818 人,总计 38 个班 1 898 人。每人重复调查 2 次,总计调查 3 796 人次。本次调查问卷回收率达 100%,无漏填项目,回收问卷的合格率达 100%。用 Excel2003 建立数据库录入数据,对所有资料进行手工及计算机纠错,数据分析通过 SAS9.13 完成。

3.1　总体比例的调查　分别将 38 个样本班随机分为样本含量相等的样本 1 和样本 2 两部分。设计一随机化装置:一口袋中放置大小、重量、触感完全相同的 6 个红球、4 个白球,即 $P=0.6$。在没有旁人在场时,每个被抽中班的每个学生有放回地从袋中随机抽出一球,只需回答"是"或"否"。对样本 1 中的每个人:抽到红球时回答敏感问题 A"你有过婚前性行为吗?";抽到白球时回答非敏感问题 B,如:"你是男生吗?";全部回答非敏感问题 C,如:"你的学号是单号吗?"。对样本 2 中的每个人:抽到红球时回答敏感问题 A;抽到白球时回答非敏感问题 C;全部回答非敏感问题 B。

3.1.1　各班婚前性行为发生比例的调查计算结果　对双无关模型分层整群抽样重复 2 次调查苏州大学新校区 38 个班学生婚前性行为数据,按(8)式计算得:20 个本科班第一次调查的婚前性行为发生比例 $\pi_{i1}$($i=1,2,\cdots,20$)及第二次调查的婚前性行为发生比例 $\pi'_{i1}$($i=1,2,\cdots,20$);18 个研究生班第一次调查的婚前性行为发生比例 $\pi_{i2}$($i=1,2,\cdots,18$)及第二次调查的婚前性行为发生比例 $\pi'_{i2}$($i=1,2,\cdots,18$)。结果见表 1。

**表 1　双无关问题模型分层整群抽样重复 2 次调查各班婚前性行为的发生比例**

**Tab 1　Proportions of pre-marriage sex behavior in each class for two times investigation on Greenberg model in stratified cluster sampling**

| Students | Means (First time ) | Means (Second time) | Students | Means (First time ) | Means (Second time) |
|---|---|---|---|---|---|
| Undergraduate (Class) | $\pi_{i1}$ | $\pi'_{i1}$ | Postgraduate (Class) | $\pi_{i2}$ | $\pi'_{i2}$ |
| 1 | 0.237 0 | 0.238 9 | 1 | 0.250 6 | 0.230 8 |
| 2 | 0.140 1 | 0.123 0 | 2 | 0.225 3 | 0.215 5 |
| 3 | 0.239 4 | 0.208 7 | 3 | 0.214 5 | 0.224 6 |
| 4 | 0.215 8 | 0.254 8 | 4 | 0.242 9 | 0.238 4 |
| 5 | 0.179 6 | 0.217 0 | 5 | 0.233 3 | 0.248 3 |
| 6 | 0.197 7 | 0.227 6 | 6 | 0.262 4 | 0.271 9 |
| 7 | 0.152 7 | 0.192 6 | 7 | 0.251 4 | 0.237 9 |
| 8 | 0.234 8 | 0.203 8 | 8 | 0.238 9 | 0.202 7 |
| 9 | 0.055 4 | 0.055 3 | 9 | 0.210 5 | 0.210 0 |
| 10 | 0.073 1 | 0.110 0 | 10 | 0.252 7 | 0.235 2 |
| 11 | 0.107 0 | 0.073 8 | 11 | 0.205 5 | 0.226 9 |
| 12 | 0.012 4 | 0.000 0 | 12 | 0.208 3 | 0.211 1 |
| 13 | 0.112 1 | 0.135 5 | 13 | 0.248 5 | 0.213 2 |
| 14 | 0.212 2 | 0.212 2 | 14 | 0.269 4 | 0.239 7 |
| 15 | 0.138 1 | 0.127 4 | 15 | 0.360 1 | 0.348 7 |
| 16 | 0.170 4 | 0.170 4 | 16 | 0.212 4 | 0.212 4 |
| 17 | 0.149 5 | 0.149 5 | 17 | 0.219 0 | 0.233 4 |
| 18 | 0.144 8 | 0.151 2 | 18 | 0.219 0 | 0.212 4 |
| 19 | 0.158 9 | 0.192 4 | | | |
| 20 | 0.169 3 | 0.132 3 | | | |

3.1.2　各层婚前性行为发生比例的估计及其估计方差　以本科生第一次调查的数据，按(4)式计算得本科生婚前性行为发生比例的估计值为：

$$\hat{\pi}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}\pi_{i1}=0.155\ 0$$

按(5)式计算得 $\hat{\pi}_1$ 的估计方差为：

$$v(\hat{\pi}_1)=\frac{1-f_1}{n_1(n_1-1)}\sum_{i=1}^{n_1}(\pi_{i1}-\hat{\pi}_1)^2=0.000\ 2$$

以研究生第一次调查的数据，按(4)式计算得研究生婚前性行为发生比例的估计值为：

$$\hat{\pi}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}\pi_{i2}=0.240\ 3$$

按(5)式计算得 $\hat{\pi}_2$ 的估计方差为：

$$v(\hat{\pi}_2)=\frac{1-f_2}{n_2(n_2-1)}\sum_{i=1}^{n_2}(\pi_{i2}-\hat{\pi}_2)^2=0.000\ 04$$

3.1.3　苏州大学新校区学生婚前性行为发生比例的估计量及其方差　按(6)式苏州大学新校区学生婚前性行为发生比例的估计值为：

$$\hat{\pi}=\sum_{h=1}^{L}W_h\cdot\hat{\pi}_h=W_1\hat{\pi}_1+W_2\hat{\pi}_2=0.168\ 6$$

由(7)式得 $\hat{\pi}$ 的估计方差为：

$$v(\hat{\pi})=\sum_{h=1}^{L}W_h{}^2\cdot v(\hat{\pi}_h)=W_1{}^2\cdot v(\hat{\pi}_1)+W_2{}^2\cdot v(\hat{\pi}_2)=0.000\ 1$$

$\pi$ 的 95%的可信区间为：

$$\hat{\pi}\pm1.96\sqrt{v(\hat{\pi})}=0.148\ 8\sim0.188\ 0$$

3.1.4　调查的信度评价　采用 SAS 9.13 将 20 个本科班重复 2 次调查计算的比例数据进行相关分析($\pi_{i1}$、$\pi'_{i1}$ 分别经正态性 $W$ 检验，$W$ 值分别为 0.951 999、0.951 926，$P$ 值分别为 0.398 4、0.397 3，符合正态分布)，Pearson 积差相关系数 $r=0.917\ 34$，$P<0.000\ 1$，显示第一层整群抽样重复 2 次调查结果间存在较高程度的相关性；采用 SAS 9.13将 18 个研究生班重复 2 次调查计算的比例数据进行平方根反正弦变换，对变换后的数据进行相关分析，Pearson 积差相关系数 $r=0.866\ 88$，$P<0.000\ 1$，显示第二层整群抽样 2 次调查结果间存在较高程度的相关性；采用 SAS 9.13 将全部调查的 38 个班重复 2 次调查计算的比例数据进行平方根反正弦变换，对变换后的数据进行相关分析，Pearson 积差相关系数 $r=0.944\ 16$，$P<0.000\ 1$，显示分层整群抽样重复 2 次调查结果间相关程度较高；说明本研究的调查方法与计算公式的(重测)信度较高。

3.2　总体均数的调查　设计一套随机装置：在小布袋中放置大小、重量、质感相同的小球 10 个，分别贴有 0、1、…、9 数字标签各 1 个。38 个样本班的每个学生从小布袋中有放回地随机抽取 1 个小球，将自

己上两个学期考试作弊的次数与抽中小球上的数值相加的结果填入调查表中。

3.2.1　*各班考试作弊平均次数的调查计算结果*　对加法 RRT 模型分层整群抽样重复 2 次调查苏州大学新校区 38 个班学生上两个学期考试作弊次数数据,按(16)式计算得:20 个本科班第一次调查的考试作弊平均次数 $\mu_{i1}(i=1,2,\cdots,20)$ 及第二次调查的考试作弊平均次数 $\mu'_{i1}(i=1,2,\cdots,20)$;18 个研究生班第一次调查的考试作弊平均次数 $\mu_{i2}(i=1,2,\cdots,18)$ 及第二次调查的考试作弊平均次数 $\mu'_{i2}(i=1,2,\cdots,18)$。计算结果见表 2。

**表 2　加法 RRT 模型分层整群抽样重复 2 次调查各班学生上两个学期考试作弊平均次数**

**Tab 2　Mean of cheating frequency on exams in recent two semesters in each class for two times investigation on Additive constant model in stratified cluster sampling**

| Students | Means (First time ) | Means (Second time) | Students | Means (First time) | Means (Second time) |
|---|---|---|---|---|---|
| Undergraduate (Class) | $\mu_{i1}$ | $\mu'_{i1}$ | Postgraduate (Class) | $\mu_{i2}$ | $\mu'_{i2}$ |
| 1 | 1.053 2 | 0.989 4 | 1 | 1.318 2 | 1.431 8 |
| 2 | 2.025 0 | 2.350 0 | 2 | 2.430 2 | 2.779 1 |
| 3 | 1.600 0 | 1.500 0 | 3 | 2.655 6 | 2.477 8 |
| 4 | 1.670 2 | 1.861 7 | 4 | 1.117 0 | 1.414 9 |
| 5 | 1.068 2 | 1.113 4 | 5 | 0.969 4 | 0.949 0 |
| 6 | 0.650 0 | 0.593 0 | 6 | 0.920 0 | 0.920 0 |
| 7 | 1.166 7 | 1.011 1 | 7 | 0.855 6 | 1.055 6 |
| 8 | 1.202 1 | 1.074 5 | 8 | 0.565 2 | 0.840 4 |
| 9 | 1.166 7 | 0.744 4 | 9 | 0.566 7 | 0.544 4 |
| 10 | 1.112 2 | 1.153 1 | 10 | 0.770 8 | 0.750 0 |
| 11 | 1.214 3 | 1.452 4 | 11 | 0.776 6 | 0.755 3 |
| 12 | 0.522 7 | 0.500 0 | 12 | 0.750 0 | 0.791 7 |
| 13 | 1.100 0 | 1.011 1 | 13 | 0.755 3 | 0.734 0 |
| 14 | 0.795 5 | 0.772 7 | 14 | 0.523 8 | 0.500 0 |
| 15 | 0.934 8 | 0.913 0 | 15 | 1.033 3 | 1.055 6 |
| 16 | 0.931 8 | 0.863 6 | 16 | 0.934 0 | 0.887 4 |
| 17 | 0.152 2 | 0.108 7 | 17 | 0.795 5 | 0.772 7 |
| 18 | 0.952 4 | 0.881 0 | 18 | 0.900 0 | 0.918 1 |
| 19 | 0.833 3 | 0.785 7 | | | |
| 20 | 0.522 2 | 0.500 0 | | | |

3.2.2　*各层考试作弊次数总体均数的估计及其估计方差*　以本科生第一次调查的数据,按(12)式计算得本科生上两个学期考试作弊次数总体均数的估计量为:$\hat{\mu}_1=1.033\ 7$

以研究生第一次调查的数据,按(12)式计算得研究生上两个学期考试作弊次数总体均数的估计量为:

$$\hat{\mu}_2=1.035\ 4$$

以本科生第一次调查的数据,按(13)式计算得本科生上两个学期考试作弊次数总体均数估计量的方差为:

$$v(\hat{\mu}_1)=\frac{1-f_1}{n_1(n_1-1)}\sum_{i=1}^{n_1}(\mu_{i1}-\hat{\mu}_1)^2=0.007\ 9$$

以研究生第一次调查的数据,按(13)式计算得研究生上两个学期考试作弊次数总体均数估计量的方差为:

$$v(\hat{\mu}_2)=\frac{1-f_2}{n_2(n_2-1)}\sum_{i=1}^{n_2}(\mu_{i2}-\hat{\mu}_2)^2=0.010\ 7$$

3.2.3　*苏大新校区学生考试作弊次数总体均数的估计量及其方差*　按式(14)苏大新校区学生考试作弊次数总体均数的估计量:

$$\hat{\mu}=\sum_{h=1}^{2}W_h\hat{\mu}_h=1.034\ 0$$

按式(15)苏大新校区学生考试作弊次数总体均数估计量的估计方差为:

$$v(\hat{\mu})=\sum_{h=1}^{2}W_h^2v(\hat{\mu}_h)=0.005\ 8$$

苏大新校区学生上两学期考试作弊次数总体均数的 95%可信区间为:

$$\hat{\mu}\pm1.96\sqrt{v(\hat{\mu})}=0.884\ 7\sim1.183\ 3$$

3.2.4　*调查的信度评价*　采用 SAS 9.13,将 20 个本科班加法 RRT 模型重复 2 次调查计算的均数数据进行相关分析($\mu_{i1}$、$\mu'_{i1}$ 分别经正态性 W 检验,W

值分别为 0.961 6、0.928 8，$P$ 值分别为 0.576 8、0.146 4，符合正态分布），Pearson 积差相关系数 $r=0.957\ 55$，$P<0.000\ 1$，显示第一层整群抽样重复两次调查结果间存在较高程度的相关性；采用 SAS 9.13，将 18 个研究生班加法 RRT 模型重复两次调查计算的均数数据（不服从正态分布）进行秩相关分析，Spearman 等级相关系数 $r_s=0.902\ 43$，$P<0.000\ 1$，显示第二层整群抽样两次调查结果间存在较高程度的相关性；采用SAS 9.13，将全部调查的 38 个班加法 RRT 模型重复 2 次调查计算的均数数据（不服从正态分布）进行秩相关分析，Spearman 等级相关系数 $r_s=0.903\ 11$，$P<0.000\ 1$，显示分层整群抽样重复两次调查结果间相关程度较高；说明加法 RRT 模型分层整群抽样调查方法与计算公式的（重测）信度较高。

## 4　讨　论

4.1　*本研究的实用性*　敏感性问题的调查在卫生工作与医学科研中非常普遍和十分重要，特别在我国艾滋病防治工作中尤为重要。我国 HIV/AIDS 的流行经过传入期（1985～1988 年）、扩散期（1989～1994 年）、增长期（1995～2001 年）和较快增长期（2002～2007 年），目前正面临着特定人群和局部地区高流行的威胁！我国目前艾滋病病毒感染率究竟是多少？我国目前男、女同性恋的人口数究竟有多少？我国目前究竟有多少妓女？我国目前嫖客的数量及其年均嫖娼的次数是多少？我国目前多性伴侣人员年均性伴侣人数是多少？我国目前吸毒人数究竟有多少？我国目前国产安全套阴道交、肛交、口交的使用破损率各是多少？国家有关 HIV/AIDS 防治政策、规划的制定需要准确的数据！准确的数据呼唤对敏感问题的科学调查方法与统计公式。本研究结果为国家及地方各级卫生主管部门和有关单位制定相关规划与政策提供科学可靠的数据，对防病治病，尤其对艾滋病、性病的防治，对提高人民群众的健康水平，对发展社会主义经济具有重要的实际应用价值。

4.2　*本研究的创新性*　近期国外学者对所收集的 1965～2000 年间 38 篇相关文献进行 Meta 分析的结果显示，应用随机应答技术调查敏感性问题在结果的准确性、可靠性方面较传统调查法有着显著的优势[11]。关于敏感性问题的抽样调查设计，国内外不少统计学者进行了研究并提出了不少抽样调查方法。但到目前为止，国内外对敏感问题抽样调查的设计研究，只局限于简单随机抽样，而且对敏感问题抽样调查的信度与效度评价也极少研究。

本研究采用二分类敏感问题双无关问题 RRT 模型、数量特征敏感问题加法 RRT 模型，对较复杂的常用整群抽样、分层整群抽样调查方法，在国内外首次推导出敏感问题总体比例、均数的估计量及其估计方差的计算公式，填补了国内外卫生统计学、生物统计学、人口统计学、经济统计学、科学技术统计学、社会统计学、环境与生态统计学等各统计学科敏感性问题抽样调查设计的空白，具有较大的创新意义。

4.3　*本研究的可靠性*　本研究结合苏州大学学生婚前性行为、考试作弊次数的调查实例，对二分类敏感问题双无关问题 RRT 模型、数量特征敏感问题加法 RRT 模型的整群抽样、分层整群抽样，取得了成功的应用效果。并对调查的重测信度进行了评价，重复两次调查的结果之间相关程度很高。本研究并对本文应用实例中同样的 38 个样本班，采用同样的分层整群抽样方法，分别使用加法 RRT 模型、乘法 RRT 模型调查近两学期考试作弊的次数，两 RRT 模型的调查结果高度相关[12]。很好地说明本研究的调查方法与统计量计算公式的信度较高，即说明本研究的调查方法与统计量计算公式的可靠性较高。

4.4　*本研究公式的潜作用*　本研究对 2 种随机应答技术在整群抽样、分层整群抽样下推导出敏感问题总体比例、总体均数的估计量及其估计方差的计算公式。当敏感问题各层总体比例、各层总体均数、总体比例、总体均数的估计量及其估计方差按本研究提供的公式计算出来以后，可进一步进行（层）总体比例、（层）总体均数的区间估计（因整群抽样、分层整群抽样的样本含量一般较大，所以样本比例、样本均数一般近似服从正态分布）；进一步进行各层比例（均数）间比较的 $t$ 检验、$Z$ 检验、方差分析或秩和检验。

4.5　*本文的相关研究内容*　本文是作者 2006～2008 年主持的国家自然科学基金项目——“敏感性问题的抽样调查设计”（项目编号 30571620）的主要研究结果之一。本项目对多种 RRT 在多种复杂抽样方法下的调查技术及统计量计算公式进行了系统研究，并对调查的信度、效度作了评价。本研究并发现，当个体隐私保护度（装置中敏感性问题的配置比例）不同时，估计误差也不相同，这一结果与洪志敏等[13]得到的研究结论基本一致。

## [参 考 文 献]

[1]　Tourangeau R，Yan T. Sensitive questions in surveys[J]. Psy-

chol Bull，2007，133：859-883.

[2] 李　鲁. 社会医学[M]. 3 版. 北京：人民卫生出版社，2007：86-87.

[3] 王建华. 实用医学科研方法[M]. 北京：人民卫生出版社，2003：444-450.

[4] 丁元林，高　歌. 卫生统计学[M]. 北京：科学出版社，2008：13.

[5] Raghunath A，Georg D. Randomized response techniques for complex survey designs[J]. Statistical Papers，2006，48：131-141.

[6] Wang J F，Gao G，Fan Y B，Chen L L，Liu S X，Jin Y L. The estimation of sampling size in multi-stage sampling and its application in medical survey[J]. Appl Mathemat Comput，2006，178：239-249.

[7] Cochran W G. 抽样技术[M]. 张尧庭，吴　辉 译. 北京：中国统计出版社，1987：93-95，432，491.

[8] 苏良军. 高等数理统计[M]. 北京：北京大学出版社，2007：3.

[9] 王　岩，隋思涟，王爱青. 数理统计与 MATLAB 工程数据分析[M]. 北京：清华大学出版社，2006：10-11.

[10] Huang K C. Estimation for sensitive characteristics using optional randomized response technique[J]. Qual Quant，2008，42：679-686.

[11] Jlm G，Jjh L M，Gm P，Jmm C. Meta-analysis of randomized response research：thirty-five years of validation[J]. Sociol Meth Res，2005，33：319-348.

[12] Zhang H，Zhu K L，Han C L. Recent advance in statistics application and related areas[M]. Sydney：Aussino Academic Publishing House，2008：648-652.

[13] 洪志敏，闰在在. 基于相同保护度的随机化装置效率比较[J]. 工程数学学报，2008，25：97-102.

[本文编辑]　尹　茶

· 读者 作者 编者 ·

## 中草药名称中文、拉丁文及英文对照表(十三)

| 汉语拼音名 | 中文名 | 拉丁名 | 英文名 |
|---|---|---|---|
| Kuzhuye | 苦竹叶 | *Folium Pleioblasti* | Bitter Bamboo Leaf |
| Laifuzi | 莱菔子 | *Semen Raphani* | Radish Seed |
| Lajiao | 辣椒 | *Fructus Capsici* | Hot Pepper |
| Laliao | 辣蓼 | *Herba Polygonie Hydropiperis* | Red-Knees Herb |
| Lameihua | 腊梅花 | *Flos Chimonanthi Praecocis* | Wintersweet Flower |
| Langbacao | 狼把草 | *Herba Bidentis Tripartitae* | Bur Beggarticks Herb |
| Langdangye | 莨菪叶 | *Folium Hyoscyami* | Henbane Leaf |
| Langdu | 狼毒 | *Radix Euphorbiae Fischerianae/Radix Euphorbiae Ebracteolatae* | Langdu Root |
| Langyupi | 榔榆皮 | *Cortex Ulmi Parvifoliae* | Chinese Elm Bark |
| Lanxiangcao | 兰香草 | *Herba Caryopteridis Incanae* | Common Bluebeard Herb |
| Laohuyu | 老虎芋 | *Rhizoma Alocasiae Cucullatae* | Hoodshaped Alocasia Rhizome |
| Leigongteng | 雷公藤 | *Radix Tripterygii Wilfordii* | Common Threewingnut Root |
| Leiwan | 雷丸 | *Omphalia* | Stone-like Omphalia |
| Lianfang | 莲房 | *Receptaculum Nelumbinis* | Lotus Seed Pot |
| Liangmianzhen | 两面针 | *Radix Zanthoxyli* | Shinyleaf Pricklyash Root |
| Lianqiancao | 连钱草 | *Herba Glechomae* | Longtube Ground Ivy Herb |
| Lianxu | 莲须 | *Stamen Nelumbinis* | Lotus Stamen |
| Lianzi | 莲子 | *Lotus Seed* | Semen Nelumbinis |
| Lianzixin | 莲子心 | *Plumula Nelumbinis* | Lotus Plumule |
| Liaogewang | 了哥王 | *Radix Wikstroemae* | Indian Stringbush Root |
| Liaoqiao | 连翘 | *Fructus Forsythiae* | Weeping Forsythiae Capsule |
| Liedang | 列当 | *Herba Orobanches* | Skyblue Broomrape Herb |
| Liexiangdujuan | 烈香杜鹃 | *Folium Rhododendri Anthopogonoidis* | Savoury Rhododendron Leaf |
| Lilu | 藜芦 | *Radix et Rhizoma Veratri* | Falsehellebore Root and Rhizome |
| Ling | 菱 | *Pedicellus etPericarpiumTrapae* | Water Calptrop Base and Peel |
| Linglan | 铃兰 | *Herba Convallariae* | Lilyofthevalley Herb |