DOI:10.13546/j.cnki.tjyjc.2016.11.002

敏感问题乘法模型的RRT分层三阶段抽样最优样本量设定

周　霞

（阜阳师范学院 数学与统计学院，安徽 阜阳 236037）

摘　要：文章研究数量特征敏感问题的乘法模型在随机应答技术（RRT）分层三阶段抽样方法下的最优样本量的问题。根据RRT分层三阶段抽样方法给出数量特征敏感乘法模型的调查设计方法，计算出总体均数的估计量及其方差。应用拉格朗日乘数法，给出了两种情况下的最优样本量，一是抽样误差限定而调查费用达到最小情况下的最优样本值，二是调查费用限定而抽样误差达到最小情况下的最优样本值。并计算出抽样误差一定时最小的费用及费用一定时最小的抽样误差。

关键词：数量特征敏感乘法模型；RRT；分层三阶段抽样；拉格朗日乘数法；最优样本量

中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2016)11-0008-05

0　引言

根据调查者的调查目的可以将敏感性问题分为两类，一类是调查敏感问题特征在总体中所占的比例，这类敏感问题也称为属性特征敏感问题；另一类是调查敏感问题特征数值特征大小，这类敏感问题也称数量特征敏感性问题。数量特征敏感问题是估计敏感性问题数值的总体均数，也称为敏感性均值问题，可分为三类：一是数量特征敏感问题的无关联模型，二是加法模型，三是乘法模型[1]。另外，当调查敏感性问题时，倘若调查者采用直接询问的这种传统调查方式，被调查者为了保护个人隐私，通常拒绝配合调查或故意说谎，导致调查数据偏倚，降低调查结果的可靠性。在调查敏感性问题时，不但要保护被调查者的隐私还要确保正确应答率，通常采用随机应答技术（RRT）调查方法。

RRT调查方法既可以保证被调查对象的隐私不被泄露又能避免说谎，被公认为是既可以很好保护被调查对象的个人隐私问题，又能大大提高真实回答率的最有效方法[1]，是Warner在1965年提出的一种敏感问题调查的统计学方法，可有效地在保护被调查者隐私的前提下得到敏感性问题的统计数据[2]。在敏感性问题的大规模抽样调查中，如整群抽样方法、二阶段抽样方法、分层二阶段抽样方法等常常被采用[3-6]。

目前，二项选择敏感问题和数量特征敏感问题研究已经受到学者的广泛关注，而多项选择敏感问题研究相对较少。在二项选择敏感问题和数量特征敏感问题上，所采用的方法也基本上是二阶段抽样或者分层二阶段抽样[3-6]。而关于数量特征敏感乘法模型分层三阶段抽样统计方法这方面的文献少见。本文针对数量特征性敏感问题乘法模型分层三阶段抽样这一方法进行探究，陈述其抽样方法，给出总体均数的估计量及其方差，并计算两种情况下的最优样本量：一是在抽样误差限定而调查费用达到最小情况下的最优样本量，二是在调查费用限定而抽样误差达到最小情况下的最优样本量，并求出此时的最少费用或最小误差。

1　准备工作

引理1[7]：设随机变量 X 与 Y 相互独立，则有：

$E(XY)=E(X)E(Y)$

引理2[8]：设总体 X 具有二阶矩，即 $E(X)=\mu$，$Var(X)=\sigma^2<+\infty$，$x_1,x_2,...,x_n$ 为样本，样本均值 $\bar{x}$，样本方差 s^2，则：

$$Var(\bar{x})=\frac{\sigma^2}{n}=\frac{1}{n}E(s^2)$$

即样本均值的均值与总体均值相同，而样本均值的方差是总体方差的 $\frac{1}{n}$。

引理3[8]：在条件组 $\varphi_k(x_1,x_2,...,x_n)=0$，$k=1,2,...,m$，$(m<n)$ 的限制下，求目标函数 $y=f(x_1,x_2,...,x_n)$ 的极值问题，其中 f 与 $\varphi_k(k=1,2,...,m)$ 在区域 D 内有连续的一阶偏导数。若 D 的内点 $P_0(x_1^{(0)},...,x_n^{(0)})$ 是上述问题的极值点，且雅克比矩阵

$$\begin{bmatrix}\frac{\partial\varphi_1}{\partial x_1} & \cdots & \frac{\partial\varphi_1}{\partial x_n}\\ \vdots & & \vdots\\ \frac{\partial\varphi_m}{\partial x_1} & \cdots & \frac{\partial\varphi_m}{\partial x_n}\end{bmatrix}_{p_0}$$

基金项目：全国统计科学研究(计划)项目(2013LY093)；安徽省高校优秀青年人才支持计划项目(皖教秘人[2014]181号)

作者简介：周　霞(1981—)，女，陕西商南人，博士，副教授，研究方向：概率论与数理统计、随机微分系统。

的秩为 m，则存在 m 个常数 $\lambda_1^{(0)},...,\lambda_m^{(0)}$，使得 $(x_1^{(0)},...,x_n^{(0)},\lambda_1^{(0)},...,\lambda_m^{(0)})$ 为拉格朗日函数 $L(x_1,x_2,...,x_n,\lambda_1,\lambda_2,...,\lambda_m)=f(x_1,x_2,...,x_n)+\sum_{k=1}^{m}\lambda_k\varphi_k(x_1,x_2,...,x_n)$ 的稳定点，即 $(x_1^{(0)},...,x_n^{(0)},\lambda_1^{(0)},...,\lambda_m^{(0)})$ 为下述 $n+m$ 个方程

$$\begin{cases} L_{x_1}=\dfrac{\partial f}{\partial x_1}+\sum\limits_{k=1}^{m}\lambda_k\dfrac{\partial\varphi_k}{\partial x_1}=0 \\ \cdots \\ L_{x_n}=\dfrac{\partial f}{\partial x_n}+\sum\limits_{k=1}^{m}\lambda_k\dfrac{\partial\varphi_k}{\partial x_n}=0 \\ L_{\lambda_1}=\varphi_1(x_1,...,x_n)=0 \\ \cdots \\ L_{\lambda_m}=\varphi_m(x_1,...,x_n)=0 \end{cases}$$

的解。

2 数量特征敏感问题乘法模型分层三阶段抽样的统计方法

2.1 数量特征敏感问题乘法RRT模型

为了保护被调查对象的个人隐私，使用一个黑色且不透明的布袋子，在里面放有完全相同（质感、大小、重量、颜色都相同）的10个小球，分别标有1、2、…、10的数字。首先，被调查对象采取有放回地的方式从布袋子中随机抽取一个小球，然后秘密的把调查敏感问题的特征数值与抽中小球上的数字相乘，最后将结果填入调查表中。被调查对象设敏感性问题的特征数值设为随机变量 X，小球上的数字为随机变量 Y，定义一个新的随机变量 $Z=XY$。随机变量 Y 取值为1、2、…、10，每球被抽中的概率为0.1，其统计性质可以计算。随机变量 Z 通过调查数据可以计算其统计特征的数值，根据 $Z=XY$，可以得到敏感性问题的特征数值。

2.2 数量特征敏感问题乘法RRT模型的分层三阶段抽样方法

设总体中含有 N 个个体，将总体完备的划分成 H 个子总体（子总体也称层），子总体（层）之间相互独立。分层抽样是指在每层中使用概率抽样的方法，相互独立地抽取所需样本，将所抽的样本合起来作为整个总体的样本，这一抽样过程称之为分层抽样[9]。由于分层抽样是在不同的子总体中分别独立进行的，所以各层样本是相互独立的，而且样本的统计量也是相互独立的。

设调查总体为 N，分为 H 层。第 h 层由 N_h 个一级单位组成，h 层第 i 个一级单位是由 N_{hi} 个二级单位组成，h 层第 i 个一级单位内第 j 个二级单位是有 N_{hij} 个被调查对象。h 层平均每个一级单位包含 $\bar{N}_{hi}$ 个二级单位，h 层平均每一个二级单位有 $\bar{N}_{hij}$ 个被调查对象，其中，$h=1,2,...,H$、$i=1,2,..,N_h$、$j=1,2,..,N_{hi}$。

本文采用分层三阶段抽样调查。第一阶段：从 h（$h=1,2,...,H$）层里随机的抽取 n_h 个一级单位。第二阶段：从 h 层第 i（$i=1,2,..,n_h$）个被抽中的一级单位里随机抽取 n_{hi} 个二级单位。第三阶段：从 h 层第 i 个被抽中的一级单位的第 j（$j=1,2,..,n_{hi}$）个二级单位里随机抽取 n_{hij} 个三级单位，也就是被调查对象。平均从 h 层被抽中的一级单位中随机抽取了 $\bar{n}_{hi}$ 个二级单位，平均从 h 层每个被抽中的二级单位里随机抽取了 $\bar{n}_{hij}$ 个三级单位。根据被抽中的 n_{hij} 的被调查对象，使用上文中所描述的调查方法进行调查。

2.3 总体均数的估计量 $\hat{\mu}$ 及总体均数估计量的方差 $Var(\hat{\mu})$

记总体 X 均值为 μ，即 $E(X)=\mu$；总体方差为 σ^2，即 $Var(X)=\sigma^2$；记 x_{hijk} 为第 h 层第 i 个一级单位内第 j 个二级单位里第 k 个被调查对象调查数值，其中 $k=1,2,...,n_{hij}$；μ_{hij} 为 h 层第 i 个一级单位内第 j 个二级单位敏感问题特征 X 的均数，μ_{hi} 为 h 层第 i 个一级单位的均值，μ_h 为第 h 层总体均数，$\hat{\mu}_h$ 为 μ_h 的估计量。记 M_h 为 h 层含有的三级单位个数。$\hat{\mu}$、$\hat{\mu}_{hij}$、$\hat{\mu}_{hi}$、$\hat{\mu}_h$ 分别为 μ、μ_{hij}、μ_{hi}、$\hat{\mu}_h$ 的估计量，$Var(\hat{\mu})$ 为 $\hat{\mu}$ 的方差。根据概率统计相关知识可求出：

$$\hat{\mu}_{hij}=\frac{\sum_{k=1}^{n_{hij}}x_{hijk}}{n_{hij}} \tag{1}$$

$$\hat{\mu}_{hi}=\frac{\sum_{j=1}^{n_{hi}}N_{hij}\hat{\mu}_{hij}}{\sum_{j=1}^{n_{hi}}N_{hij}} \tag{2}$$

$$\hat{\mu}_h=\frac{N_h}{n_hM_h}\sum_{i=1}^{n_h}\frac{N_{hi}}{n_{hi}}\left(\sum_{j=1}^{n_{hi}}N_{hij}\hat{\mu}_{hij}\right) \tag{3}$$

所以有：

$$\hat{\mu}=\frac{\sum_{h=1}^{H}M_h\hat{\mu}_h}{N}=\sum_{h=1}^{H}\frac{M_h}{N}\left[\frac{N_h}{n_hM_h}\sum_{i=1}^{n_h}\frac{N_{hi}}{n_{hi}}\left(\sum_{j=1}^{n_{hi}}N_{hij}\hat{\mu}_{hij}\right)\right] \tag{4}$$

从而，总体均值估计量 $\hat{\mu}$ 的方差 $Var(\hat{\mu})$ 为：

$$Var(\hat{\mu})=\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\left[\frac{\sigma_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)+\frac{\sigma_{hi}^2}{n_h\bar{n}_{hi}}\left(1-\frac{\bar{n}_{hi}}{\bar{N}_{hi}}\right)+\frac{\sigma_{hij}^2}{n_h\bar{n}_{hi}\bar{n}_{hij}}\left(1-\frac{\bar{n}_{hij}}{\bar{N}_{hij}}\right)\right] \tag{5}$$

其中，

$$\sigma_h^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(\mu_{hi}-\mu_h\right)^2$$

$$\sigma_{hi}^2=\frac{1}{N_h}\sum_{i=1}^{N_h}\frac{1}{N_{hi}-1}\sum_{j=1}^{N_{hi}}\left(\mu_{hij}-\mu_{hi}\right)^2$$

$$\sigma_{hij}^2=\frac{1}{N_h}\sum_{i=1}^{N_h}\frac{1}{N_{hi}}\sum_{j=1}^{N_{hi}}\frac{1}{N_{hij}-1}\sum_{k=1}^{N_{hij}}\left(x_{hijk}-\mu_{hij}\right)^2$$

它们的样本估计量分别为：

$$s_h^2=\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{\mu}_{hi}-\hat{\mu}_h\right)^2$$

$$s_{hi}^2=\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{n_{hi}-1}\sum_{j=1}^{n_{hi}}\left(\hat{\mu}_{hij}-\hat{\mu}_{hi}\right)^2$$

$$s_{hij}^2=\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{1}{n_{hi}}\sum_{j=1}^{n_{hi}}\frac{1}{n_{hij}-1}\sum_{k=1}^{n_{hij}}\left(x_{hijk}-\hat{\mu}_{hij}\right)^2$$

2.4 μ_{hij} 的估计量 $\hat{\mu}_{hij}$ 及方差 $\mathrm{Var}(\hat{\mu}_{hij})$

在(1)式给出的 $\hat{\mu}_{hij}$ 计算公式中，x_{hijk} 的值不能直接得到，原因是：为了保护被抽中的每个三级单位对象的隐私，设计了随机装置。随机装置的设计规则是抽中的每个调查对象将敏感问题特征的数值与所抽中小球上的编号数值相乘，将乘积填入相应的调查表中。设 z_{hijk} 为第 h 层被调查对象在调查表中所填写的数值，其中 $k=1,2,...,n_{hij}$ 调查者只能看到随机变量 $Z=XY$ 的取值 z_{hijk}，而随机变量 X 不知，抽中小球上的数值为随机变量 Y，Y 的取值为 1,2,...,10，其均值为：

$$\mu_Y=\frac{1}{10}(1+2+...+10)=5.5 \tag{6}$$

分别记 $\mu_{z_{hij}},\hat{\mu}_{z_{hij}},s^2_{z_{hij}}$ 为 h 层第 i 个一级单位第 j 个二级单位所有调查数值的总体均值、样本均值、样本方差。由于随机变量 X,Y 相互独立，根据引理1可得：

$$\mu_{z_{hij}}=\mu_{hij}\mu_Y \tag{7}$$

$$\text{则 }\hat{\mu}_{hij}=\frac{\hat{\mu}_{z_{hij}}}{\mu_Y}=\frac{1}{\mu_Y}\hat{\mu}_{z_{hij}}=\frac{1}{\mu_Y}\frac{\sum_{k=1}^{n_{hij}}z_{hijk}}{n_{hij}} \tag{8}$$

由(7)式及方差的性质可得：

$$\mathrm{Var}(\hat{\mu}_{hij})=\mathrm{Var}\left(\frac{\hat{\mu}_{z_{hij}}}{\mu_Y}\right)=\frac{1}{\mu_Y^2}\mathrm{Var}\left(\hat{\mu}_{z_{hij}}\right) \tag{9}$$

由引理2可得：

$$\mathrm{Var}(\hat{\mu}_{hij})=\frac{1}{\mu_Y^2}\mathrm{Var}\left(\hat{\mu}_{z_{hij}}\right)=\frac{1}{\mu_Y^2}\frac{1}{n_{hij}}s^2_{z_{hij}} \tag{10}$$

2.5 最优样本量的计算

抽样误差为 $\mathrm{Var}(\hat{\mu})$，抽样调查所需要的费用为[10]：

$$C=\sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}C_{1h}n_h+\sum_{h=1}^{H}C_{2h}n_h\bar{n}_{h1}+\sum_{h=1}^{H}C_{3h}n_h\bar{n}_{h1}\bar{n}_{hij} \tag{11}$$

其中，C 为调查的总费用，C_{0h} 为 h 层调查所需平均费用，C_{1h} 为 h 层每调查一个一级单位的平均费用，C_{2h} 为 h 层每调查一个二级单位的平均费用，C_{3h} 为 h 层用于每个被调查对象的平均费用。

2.5.1 限定抽样误差使调查费用达到最小时最优样本量

限定抽样误差 $\mathrm{Var}(\hat{\mu})$ 为 W，根据引理3，此时最优样本量就是(11)式在约束条件（$\mathrm{Var}(\hat{\mu})$ 为 W）下的极小值点。设 λ 为拉格朗日乘数，则：

$$F=\sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}C_{1h}n_h+\sum_{h=1}^{H}C_{2h}n_h\bar{n}_{h1}+\sum_{h=1}^{H}C_{3h}n_h\bar{n}_{h1}\bar{n}_{hij}+\lambda(\mathrm{Var}(\hat{\mu})-W) \tag{12}$$

由引理3、多元函数求导法则和多元函数微分性质得：

$$\begin{cases}\dfrac{\partial F}{\partial(n_h)}=C_{1h}-\dfrac{\lambda M_h^2(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}{N^2n_h^2}=0\\ \dfrac{\partial F}{\partial(n_h\bar{n}_{h1})}=C_{2h}-\dfrac{\lambda M_h^2(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})}{N^2n_h^2\bar{n}_{h1}^2}=0\\ \dfrac{\partial F}{\partial(n_h\bar{n}_{h1}\bar{n}_{hij})}=C_{3h}-\dfrac{\lambda M_h^2\sigma_{hij}^2}{N^2n_h^2\bar{n}_{h1}^2\bar{n}_{hij}^2}=0\\ \mathrm{Var}(\hat{\mu})-W=0\end{cases} \tag{13}$$

解(13)式的第三个式子可得：

$$n_h=\frac{\sqrt{\lambda}M_h\sigma_{hij}}{\sqrt{C_{3h}}N\bar{n}_{h1}\bar{n}_{hij}} \tag{14}$$

由(13)式的第二个和第三个式子可得：

$$\bar{n}_{h1}=\frac{\sqrt{\lambda}M_h\sqrt{(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}N}\frac{\sqrt{C_{3h}}N\bar{n}_{h1}\bar{n}_{hij}}{\sqrt{\lambda}M_h\sigma_{hij}}$$

整理上式可得：

$$\bar{n}_{hij}=\frac{\sqrt{C_{2h}}}{\sqrt{(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})}}\frac{\sigma_{hij}}{\sqrt{C_{3h}}} \tag{15}$$

由(13)式第一个式子和(14)式可得：

$$\bar{n}_{h1}=\frac{\sigma_{hij}\sqrt{C_{1h}}}{\bar{n}_{hij}\sqrt{C_{3h}}\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}$$

将(15)式代入上式可得：

$$\bar{n}_{h1}=\frac{\sigma_{hij}\sqrt{C_{1h}}\sqrt{C_{3h}}\sqrt{(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sigma_{hij}\sqrt{C_{2h}}\sqrt{C_{3h}}\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}=\frac{\sqrt{C_{1h}}\sqrt{(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}} \tag{16}$$

由(13)式可得：

$$\begin{cases}n_h^2=\dfrac{1}{C_{1h}N^2}\lambda M_h^2(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})\\ n_h^2\bar{n}_{h1}^2=\dfrac{1}{C_{2h}N^2}\lambda M_h^2(\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij})\\ n_h^2\bar{n}_{h1}^2\bar{n}_{hij}^2=\dfrac{1}{C_{3h}N^2}\lambda M_h^2\sigma_{hij}^2\end{cases} \tag{17}$$

由(5)式得：

$$\mathrm{Var}(\hat{\mu})=-\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h}+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\left[\frac{1}{n_h}\left(\sigma_h^2-\frac{\sigma_{h1}^2}{\bar{N}_{h1}}\right)+\frac{1}{n_h\bar{n}_{h1}}\left(\sigma_{h1}^2-\frac{\sigma_{hij}^2}{\bar{N}_{hij}}\right)+\frac{\sigma_{hij}^2}{n_h\bar{n}_{h1}\bar{n}_{hij}}\right]=W$$

将(17)式代入上式得：

$$W+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h}=\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\left[\frac{\sqrt{C_{1h}}N}{\sqrt{\lambda}M_h\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}\left(\sigma_{h1}^2-\frac{\sigma_{h1}^2}{\bar{N}_{h1}}\right)\right]+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\left[\frac{\sqrt{C_{2h}}N}{\sqrt{\lambda}M_h\sqrt{\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij}}}\left(\sigma_{h1}^2-\frac{\sigma_{hij}^2}{\bar{N}_{hij}}\right)+\frac{\sigma_{hij}^2\sqrt{C_{3h}}N}{\sqrt{\lambda}M_h\sqrt{\sigma_{hij}^2}}\right]$$

整理上式可得：

$$\sqrt{\lambda}=\frac{\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)\left[\sqrt{C_{1h}}\sqrt{(\sigma_{h1}^2-\sigma_{h1}^2/\bar{N}_{h1})}+\sqrt{C_{2h}}\sqrt{\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij}}+\sqrt{\sigma_{hij}^2}\sqrt{C_{3h}}\right]}{W+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_{h1}^2}{N_h}} \tag{18}$$

将(15)式和(16)式都代入(14)式得：

$$n_h=\frac{\sqrt{\lambda}M_h\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}{N\sqrt{C_{1h}}} \tag{19}$$

将(18)式代入(19)式得：

$$n_h=\frac{\sqrt{\lambda}M_h\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}{N\sqrt{C_{1h}}}=\sum_{h=1}^{H}\frac{M_h^2\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}}{N\sqrt{C_{1h}}}\left[\sqrt{C_{1h}}\sqrt{(\sigma_h^2-\sigma_{h1}^2/\bar{N}_{h1})}+\sqrt{C_{2h}}\sqrt{\sigma_{h1}^2-\sigma_{hij}^2/\bar{N}_{hij}}+\sqrt{\sigma_{hij}^2}\sqrt{C_{3h}}\right]/\left[W+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h}\right] \tag{20}$$

利用 $n_{hi}=N_{hi}\cdot\frac{\bar{n}_{hi}}{\bar{N}_{hi}}$，将(16)式带入可得：

$$n_{hi}=N_{hi}\cdot\frac{\bar{n}_{hi}}{\bar{N}_{hi}}=\frac{N_{hi}}{\bar{N}_{hi}}\frac{\sqrt{C_{1h}}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}} \tag{21}$$

利用 $n_{hij}=N_{hij}\cdot\frac{\bar{n}_{hij}}{\bar{N}_{hij}}$，将(15)式带入可得：

$$n_{hij}=N_{hij}\cdot\frac{\bar{n}_{hij}}{\bar{N}_{hij}}=\frac{N_{hij}}{\bar{N}_{hij}}\frac{\sqrt{C_{2h}}}{\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}\frac{\sigma_{hij}}{\sqrt{C_{3h}}} \tag{22}$$

(20)、(21)和(22)式即为限定抽样误差为 W 使费用达到最小时的样本量，即为此时的最优样本量。下面计算此时的最小费用。将(17)式带入(12)式得：

$$F_1=\sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}C_{1h}\frac{\sqrt{\lambda}M_h}{N\sqrt{C_{1h}}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sum_{h=1}^{H}C_{2h}\frac{\sqrt{\lambda}M_h}{N\sqrt{C_{2h}}}\sqrt{\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij}}+\sum_{h=1}^{H}C_{3h}\frac{\sqrt{\lambda}M_h}{N\sqrt{C_{3h}}}\sigma_{hij}=\sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}\frac{M_h}{N}\sqrt{\lambda}$$

$$(\sqrt{C_{1h}(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}+\sqrt{C_{3h}}\sigma_{hij})$$

将(18)式带入上式可得最小费用为：

$$F_1=\sum_{h=1}^{H}C_{0h}+\frac{\left\{\sum_{h=1}^{H}\frac{M_h}{N}\left(\sqrt{C_{1h}(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}+\sqrt{C_{3h}}\sigma_{hij}\right)\right\}^2}{W+\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h}} \tag{23}$$

2.5.2　限定调查费用使抽样误差达到最小时最优样本量

当调查总费用为 C 时，根据引理3，得得到抽样误差 $\mathrm{Var}(\hat{\mu})$ 最小值，此时最优样本量为(5)式在约束条件(11)式下的极小值点。设 λ 为拉格朗日乘数，则：

$$F=\mathrm{Var}(\hat{\mu})+\lambda\left(\sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}C_{1h}n_h+\sum_{h=1}^{H}C_{2h}n_h\bar{n}_{hi}+\sum_{h=1}^{H}C_{3h}n_h\bar{n}_{hi}\bar{n}_{hij}-C\right) \tag{24}$$

由引理3、多元函数求导法则和多元函数微分性质得：

$$\begin{cases}\frac{\partial F}{\partial(n_h)}=\lambda C_{1h}-\frac{M_h^2(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}{N^2n_h^2}=0\\ \frac{\partial F}{\partial(n_h\bar{n}_{hi})}=\lambda C_{2h}-\frac{M_h^2(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}{N^2n_h^2\bar{n}_{hi}^2}=0\\ \frac{\partial F}{\partial(n_h\bar{n}_{hi}\bar{n}_{hij})}=\lambda C_{3h}-\frac{M_h^2\sigma_{hij}^2}{N^2n_h^2\bar{n}_{hi}^2\bar{n}_{hij}^2}=0\\ \sum_{h=1}^{H}C_{0h}+\sum_{h=1}^{H}C_{1h}n_h+\sum_{h=1}^{H}C_{2h}n_h\bar{n}_{hi}+\sum_{h=1}^{H}C_{3h}n_h\bar{n}_{hi}\bar{n}_{hij}=C\end{cases} \tag{25}$$

解得：

$$\begin{cases}n_h=\frac{M_h\sigma_{hij}}{\sqrt{\lambda}\sqrt{C_{3h}}N\bar{n}_{hi}\bar{n}_{hij}}\\ \bar{n}_{hij}=\frac{\sqrt{C_{2h}}}{\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}\frac{\sigma_{hij}}{\sqrt{C_{3h}}}\\ \bar{n}_{hi}=\frac{\sqrt{C_{1h}}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}}\end{cases} \tag{26}$$

将(26)式的第二个式子和第三个式子代入到第一个式子可得：

$$n_h\sqrt{\lambda}N\sqrt{C_{1h}}=M_h\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})} \tag{27}$$

将(26)式和(27)式都代入到(11)式，可得：

$$C-\sum_{h=1}^{H}C_{0h}=\sum_{h=1}^{H}C_{1h}\frac{M_h\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}}{\sqrt{\lambda}N\sqrt{C_{1h}}}+\sum_{h=1}^{H}C_{2h}\frac{M_h}{\sqrt{\lambda}N}\frac{\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}}+\sum_{h=1}^{H}C_{3h}\frac{M_h\sigma_{hij}}{\sqrt{\lambda}\sqrt{C_{3h}}N}=\sum_{h=1}^{H}\frac{M_h}{\sqrt{\lambda}N}$$

$$\left(\frac{C_{1h}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}}{\sqrt{C_{1h}}}+\frac{C_{2h}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{C_{2h}}}+\frac{C_{3h}\sigma_{hij}}{\sqrt{C_{3h}}}\right)$$

$$=\sum_{h=1}^{H}\frac{M_h}{\sqrt{\lambda}N}\left(\sqrt{C_{1h}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}+\sqrt{C_{3h}}\sigma_{hij}\right)$$

由上式可得：

$$\frac{1}{\sqrt{\lambda}}=\frac{C-\sum_{h=1}^{H}C_{0h}}{\sum_{h=1}^{H}\frac{M_h}{N}\left(\sqrt{C_{1h}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}+\sqrt{C_{3h}}\sigma_{hij}\right)} \tag{28}$$

由(26)式和(28)式可得：

$$n_h=\frac{M_h\sigma_{hij}}{\sqrt{\lambda}\sqrt{C_{3h}}N\bar{n}_{hi}\bar{n}_{hij}}=\frac{M_h}{\sqrt{\lambda}N}\frac{\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}}{\sqrt{C_{1h}}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}(C-\sum_{h=1}^{H}C_{0h})$$

$$=\frac{}{\sum_{h=1}^{H}\left(\sqrt{C_{1h}}\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}}\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}+\sqrt{C_{3h}}\sigma_{hij}\right)} \tag{29}$$

根据 $n_{hi}=N_{hi}\cdot\frac{\bar{n}_{hi}}{\bar{N}_{hi}}$ 和(26)式可得：

$$n_{hi}=N_{hi}\cdot\frac{\bar{n}_{hi}}{\bar{N}_{hi}}=\frac{N_{hi}\sqrt{C_{1h}}}{\bar{N}_{hi}\sqrt{C_{2h}}}\frac{\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}}{\sqrt{(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}} \tag{30}$$

根据 $n_{hij}=N_{hij}\cdot\frac{\bar{n}_{hij}}{\bar{N}_{hij}}$ 和(26)式可得：

$$n_{hij}=N_{hij}\cdot\frac{\bar{n}_{hij}}{\bar{N}_{hij}}=\frac{N_{hij}\sqrt{C_{2h}}}{\bar{N}_{hij}\sqrt{C_{3h}}}\frac{\sigma_{hij}}{\sqrt{(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}} \tag{31}$$

(29)、(30)和(31)式即为限定抽样费用使抽样误差达到最小时的样本量，即为此时的最优样本量。下面计算此时的最小误差。将(24)、(25)式带入(5)式得：

$$\mathrm{Var}_{min}(\hat{\mu})=\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2[\frac{\sigma_h^2}{n_h}+\frac{\sigma_{hi}^2}{n_h\bar{n}_{hi}}+\frac{\sigma_{hij}^2}{n_h\bar{n}_{hi}\bar{n}_{hij}}-(\frac{\sigma_h^2}{N_h}+\frac{\sigma_{hi}^2}{n_h\bar{N}_{hi}}+\frac{\sigma_{hij}^2}{n_h\bar{n}_{hi}\bar{N}_{hij}})]=\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2[\frac{1}{n_h}\left(\sigma_h^2-\frac{\sigma_{hi}^2}{\bar{N}_{hi}}\right)+\frac{1}{n_h\bar{n}_{hi}}\left(\sigma_{hi}^2-\frac{\sigma_{hij}^2}{\bar{N}_{hij}}\right)+$$

$$\frac{\sigma_{hij}^2}{n_h\bar{n}_{hi}\bar{n}_{hij}}-\frac{\sigma_h^2}{N_h}]=\sum_{h=1}^{H}\frac{M_h\sqrt{\lambda}}{N}[\sqrt{C_{1h}(\sigma_h^2-\sigma_{hi}^2/\bar{N}_{hi})}+\sqrt{C_{2h}(\sigma_{hi}^2-\sigma_{hij}^2/\bar{N}_{hij})}$$

$$+\sqrt{C_{3h}}\sigma_{hij}]-\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h}$$

将(28)式带入上式可得最小抽样误差为：

$$\mathrm{Var}_{min}(\hat{\mu})=\frac{\left\{\sum_{h=1}^{H}\frac{M_h}{N}\left(\sqrt{C_{1h}(\sigma_h^2-\frac{\sigma_{hi}^2}{\bar{N}_{hi}})}+\sqrt{C_{2h}(\sigma_{hi}^2-\frac{\sigma_{hij}^2}{\bar{N}_{hij}})}+\sqrt{C_{3h}}\sigma_{hij}\right)\right\}^2}{C-\sum_{h=1}^{H}C_{0h}}$$

$$-\sum_{h=1}^{H}\left(\frac{M_h}{N}\right)^2\frac{\sigma_h^2}{N_h} \tag{32}$$

3 结论

文中所研究的模型是乘法模型，采用的是RRT分层三阶段抽样的统计方法。因RRT技术的采用使得被调查者的隐私受到很好的保护，真实应答率得到了提高，故该方法可用于敏感性较强的问题抽样调查。

针对数量特征敏感的敏感问题，目前的文献都集中用二阶段抽样调查或者分层二阶段抽样调查方法，采用三阶段抽样调查或者分层三阶段抽样调方法的很少。在文献[11]中，陈科锦等研究了数量特征问题加法模型RRT分层三阶段抽样方法，而没有研究数量特征问题乘法模型。在文献[12]中，靳宗达在其博士论文中研究了数量特征敏感问题乘法模型，但没有给出各阶段最优样本量及最小抽样误差或者最少费用。本文研究的是数量特征敏感问题乘法模型，采用的分层三阶段抽样调查技术，得到在限定抽样误差使费用最少时的最优样本量及此时的费用，同时也给出了在限定费用使抽样误差达到最小时的最优样本量及最小误差。

在已有的文献中对数量特征敏感问题乘法模型的随机装置均为：在不透明深色布袋中放置完全相同的10个小球，小球上分别标有0.1.2.3.4.5.6.7.8.9的数字，在保密的情况下，被调查对象将属于自己的敏感问题特征的数值与抽中小球上的数字相乘，将相乘后的结果填入调查表中。在这种情况下，当某个调查结果为0时，从调查结果无法判断抽到的小球为0还是敏感特征值为0。所以本文将10个小球的数字标签改为1.2.3.4.5.6.7.8.9.10，从而避免了这种情况发生。从这方面来看，本文改进了现有文献中所陈述的随机装置。

在分层抽样调查中，往往调查一级单位所需费用远高于二级单位，而调查二级单位所需费用又远远多于三级单位。如果少抽取一级单位多抽一些二级单位和三级单位，虽然降低了费用，但是由于调查对象缺乏代表性从而增大了误差；如果多抽些一级单位少抽些二级单位和三级单位，这样降低了误差，但是显著的增加了费用。所以，本文使用拉格朗日乘数法构造辅助函数将条件极值转换为无条件极值，给出了各层第一阶段、第二阶段和第三阶段的最优样本量。

参考文献：

[1]Cruyff M J, Van Den Hout A, Van Der Heijden P G, et al. Log-linear Randomized-Response Models Taking Self-Protective Response Behavior Into Account [J]. Sociological methods & research, 2007, 36(2).

[2]Warner S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias[J]. Journal of the American Statistical Association, 1965, 60(309).

[3]王磊，高歌，于明润. 敏感问题双无关问题模型分层二阶段整群抽样的统计方法及应用[J]. 中国卫生统计，2011, 28(1).

[4]刘鹏，高歌，贺志龙等. 数量特征敏感问题加法模型二阶段抽样的统计方法及其应用[J].苏州大学学报(医学版)，2011, 31(3).

[5]Jianfeng Wang, Ge Gao, Yubo Fan, et al. The Estimation of Sample Size in Multi-Stag Esampling and Its Application in Medical Survey [J]. Applied Mathematics and Computation, 2006,178.

[6]周云华，高歌，濮翔科等. 数量特征敏感问题加法模型分层二阶段抽样样本大小研究及其应用[J].中国卫生统计，2014,31(1).

[7]茆诗松，程依明，濮晓龙. 概率论与数理统计教程[M]. 北京：高等教育出版社，2005.

[8]华东师范大学数学系.数学分析(下册)第1版[M]. 北京：高等教育出版社，2002.

[9]刘洋. 关于敏感问题调查技术的研究[D].北京：首都经济贸易大学博士论文，2014.

[10]Christofides T C.A Generalized Randomized Response Technique [J]. Metrika, 2003,57(2).

[11]范玉波.敏感问题RRT模型下(分层)三阶段抽样调查的统计方法及其应用[D]. 苏州：苏州大学博士论文，2013.

[12]靳宗达. 敏感性问题9种RRT模型下(分层)三阶段抽样调查设计的统计方法及其应用[D]. 苏州：苏州大学博士论文，2014.

(责任编辑/亦　民)