Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 738097, 6 pages
http://dx.doi.org/10.1155/2014/738097

Research Article

Sample Size Determination for the Polychotomous Randomized Response Model for Sensitive Questions in a Stratified Two-Stage Sampling Survey

Zongda Jin,[1] Bo Yu,[2] Xiangke Pu,[3] and Ge Gao[1]

[1] *School of Public Health, Medical College of Soochow University, Suzhou 215123, China*
[2] *Department of Mathematics, Dezhou University, Dezhou 253023, China*
[3] *Changzhou No. 2 People's Hospital, Changzhou 213003, China*

Correspondence should be addressed to Ge Gao; gaoge01@163.com

Received 1 August 2013; Revised 12 March 2014; Accepted 18 March 2014; Published 22 April 2014

Academic Editor: Ch. Tsitouras

Copyright © 2014 Zongda Jin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Methods of finding the minimum value and the Lagrange function were applied to deduce the formulae for the optimum sample sizes for polychotomous randomized response technique (RRT) model in stratified two-stage sampling, so as to minimize the cost for specified sampling errors and to minimize the sampling errors under the constraint of a fixed budget. These formulae were successfully applied to sensitive topics survey among men who have sex with men (MSM) in Beijing, China.

1. Introduction

Surveys are an important source of collecting information about the characteristics of a population, from matters of medical and public health study. Their accuracy depends on ample participation and an unbiased sample [1]. However, the validity of survey on sensitive attitudes and behaviours suffers from the tendency of individuals to distort their response towards their perception of what is socially desirable [2]. As a consequence, the established conventional and routine methods like direct questioning have their own limitation in some epidemiological investigations [3]. Direct enquiring often leads to refusals or untruthful replies.

To encourage respondent's cooperation and to procure reliable data, the randomized response technique (RRT) was first introduced by Warner in 1965, which allowed respondent to elicit trustful response to the sensitive question without revealing anything definite to the interviewer in the course of the survey [4].

Sample size estimation, like all design issues, is a critical part of the design of a public health survey. For each study, an acceptable sample size needs to be chosen that balances the likelihood of a statistically significant result with the expense and cost involved in conducting the sampling survey [5].

Our previous studies involved the estimators for the proportion of population carrying the sensitive characteristic in the qualitative case or the estimators for the population mean in the quantitative case, which had been obtained with the implementation of the RRT model under complex survey on sensitive topics [6–8].

Based on the premise that the estimators of the population parameters for polychotomous RRT model in the stratified two-stage sampling survey were given, an attempt is made in this paper to provide sample sizes formulae for stratified two-stage sampling survey. These formulae have minimized the cost of survey implementation for a specified level of precision and meanwhile provided reasonably precise estimates under the constraint of a fixed budget. What is more, an example about preliminary study in Beijing is presented to determine the optimum sample size for a formal field investigation in Beijing which will be carried out in the future.

2. Survey Method

2.1. Randomized Response Designs for Polychotomous Characteristics. The RRT for dichotomous polling can be generalized to polychotomous RRT model [9]. A respondent can

belong to one of K mutually exclusive groups. All groups consist of a set of sensitive categories. Suppose that P_K is the proportion of respondent who belongs to group K. Randomization device is chosen to be a pack of $K+1$ cards identical in all respects number, labeled by the integers from 0 to K. Fix the probabilities P_0 and $P_1, \ldots, P_K$, such that $P_0 + P_1 + \ldots + P_K = 1$. Each respondent is instructed to pick out one card. If the card labeled by 0 is chosen, the respondent reveals his/her true response. If the others are chosen, the respondent discloses this figure on the card.

2.2. Stratified Two-Stage Sampling Design. Suppose that a population was subdivided into L non-overlapping strata. The hth stratum was subdivided into N_{1h} primary sampling units (PSUs). The ith PSU in stratum h comprised N_{i2h} secondary sampling units (SSUs). On average, each PSU contained $\overline{N}_{2h}$ SSUs in stratum h ($h = 1, 2, \ldots, L$, and $i = 1, 2, \ldots, N_{1h}$). The population was comprised of N SSUs (population elements).

In the first stage, n_{1h} PSUs were randomly drawn in the hth stratum. In the second stage, n_{i2h} SSUs were randomly drawn within each of n_{1h} selected PSUs from stratum h. On average, $\overline{n}_{2h}$ SSUs were randomly drawn from each chosen PSU from stratum h ($h = 1, 2, \ldots, L$, and $i = 1, 2, \ldots, n_{1h}$). The polychotomous RRT model was employed to investigate all the chosen SSUs.

3. The Formula Deduction

3.1. Estimation for the Population Proportions of the Sensitive Polychotomous Attribute and Their Estimator's Variance. Note that p_j represents the estimator of the population proportion in the jth sensitive category, p_{h-j} stands for the estimator of the population proportion in the jth sensitive categories from stratum h, p_{ih-j} denotes the estimator of the population proportion in the jth sensitive category in the ith PSU from stratum h. Then by Gao and Wang [10], it is shown that

$$p_j = \sum_{h=1}^{L} W_h p_{h-j}, \tag{1}$$

where $j = 1, 2, \ldots, K$ and $W_h = N_h/N$. Consider the following:

$$p_{h-j} = \frac{\sum_{j=1}^{n_{1h}} N_{i2h} p_{ih-j}}{\sum_{j=1}^{n_{1h}} N_{i2h}}. \tag{2}$$

The variance of p_j is expressed as

$$V\left(p_j\right) = \sum_{h=1}^{L} W_h^2 \left[\frac{\sigma_{1h-j}^2}{n_{1h}} \left(1 - \frac{n_{1h}}{N_{1h}}\right) + \frac{\sigma_{2h-j}^2}{n_{1h}\overline{n}_{2h}} \left(1 - \frac{\overline{n}_{2h}}{\overline{N}_{2h}}\right) \right]. \tag{3}$$

The sample estimator of σ_{1h-j}^2 is as follows:

$$s_{1h-j}^2 = \frac{1}{n_{1h} - 1} \sum_{i=1}^{n_{1h}} \left(\frac{N_{i2h}}{\overline{N}_{2h}} \right)^2 (p_{ih-j} - p_{h-j})^2, \tag{4}$$

where $j = 1, 2, \ldots, K$, and $h = 1, 2, \ldots, L$.

The sample estimator of σ_{2h-j}^2 is as follows:

$$s_{2h-j}^2 = \frac{1}{\sum_{i=1}^{n_{1h}} N_{i2h}} \sum_{i=1}^{n_{1h}} \frac{N_{i2h} p_{ih-j} \left(1 - p_{ih-j}\right)}{n_{i2h} P_0^2}, \tag{5}$$

where $j = 1, 2, \ldots, K$ and $h = 1, 2, \ldots, L$.

3.2. The Formulae for P_{ih-j}. Suppose again that P_{ih-j} is the estimator of the population proportion in the jth sensitive category from the ith PSU in the hth stratum, m_{ih-j} denotes the frequency of people who answer j in the ith PSU from stratum h, and λ_{ih-j} stands for the probability of people who answer j in the ith PSU from stratum h. $\widehat{\lambda}_{ih-j}$ is estimated by

$$\widehat{\lambda}_{ih-j} = \frac{m_{ih-j}}{n_{i2h}}, \tag{6}$$

under total probability formula; we could get $\lambda_{ih-j} = P_{ih-j} P_0 + P_j$, for all $i = 1, 2, \ldots$, and $j = 1, 2, \ldots, K$, provided that

$$P_{ih-j} = \frac{\lambda_{ih-j} - P_j}{P_0}. \tag{7}$$

An unbiased estimator for P_{ih-j} is as follows:

$$P_{ih-j} = \frac{\widehat{\lambda}_{ih-j} - P_j}{P_0}, \tag{8}$$

where $i = 1, 2, \ldots$, $h = 1, 2, \ldots, L$, and $j = 1, 2, \ldots, K$.

3.3. Sample Size Formulae. Let the overall survey cost C be

$$C = \sum_{h=1}^{L} C_{0h} + \sum_{h=1}^{L} C_{1h} n_{1h} + \sum_{h=1}^{L} C_{2h} n_{1h} \overline{n}_{2h}, \tag{9}$$

where C_{0h} equals the fixed costs of initiating the survey in hth stratum, C_{1h} represents the average cost of approaching to one PSU within stratum h, and C_{2h} is the average cost of interviewing an SSU in stratum h ($h = 1, 2, \ldots, L$).

The variance of p_j can also be written in the following alternative form:

$$V\left(p_j\right) = \sum_{h=1}^{L} W_h^2 \left[\frac{1}{n_{1h}} \left(\sigma_{1h-j}^2 - \frac{\sigma_{2h-j}^2}{\overline{N}_{2h}} \right) + \frac{\sigma_{2h-j}^2}{n_{1h}\overline{n}_{2h}} - \frac{\sigma_{1h-j}^2}{N_{1h}} \right]. \tag{10}$$

To minimize the sampling cost C under a given variance ($V(p_j) = V$), the optimum sampling size can be considered as the minimal values of function (9) subject to the constraint (10). The Lagrange function F is defined as

$$\begin{aligned} F\left(n_{1h}, n_{1h}\overline{n}_{2h}, \lambda\right) &= \sum_{h=1}^{L} C_{0h} + \sum_{h=1}^{L} C_{1h} n_{1h} \\ &\quad + \sum_{h=1}^{L} C_{2h} n_{1h} \overline{n}_{2h} + \lambda \left(V\left(p_j\right) - V \right), \end{aligned} \tag{11}$$

where λ is a Lagrange multiplier.

The necessary conditions for the solution of the problem are

$$\frac{\partial F}{\partial n_{1h}} = C_{1h} - \frac{\lambda W_h^2}{n_{1h}^2}\left(\sigma_{1h-j}^2 - \frac{\sigma_{2h-j}^2}{\overline{N}_{2h}}\right) = 0,$$
$$\frac{\partial F}{\partial\left(n_{1h}\overline{n}_{2h}\right)} = C_{2h} - \frac{\lambda W_h^2 \sigma_{2h-j}^2}{n_{1h}^2 \overline{n}_{2h}^2} = 0 \tag{12}$$

for $h = 1, 2, \ldots, L$.

Equation (12) gives

$$n_{1h} = \frac{\sqrt{\lambda} W_h \sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}}}{\sqrt{C_{1h}}}, \tag{13}$$

$$n_{1h}\overline{n}_{2h} = \frac{\sqrt{\lambda} W_h \sigma_{2h-j}}{\sqrt{C_{2h}}}. \tag{14}$$

Substituting the values of n_{1h} from expression (13) in (14), the $\overline{n}_{2h}$ is obtained as

$$\overline{n}_{2h} = \frac{\sigma_{2h-j}}{\sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}}} \cdot \sqrt{\frac{C_{1h}}{C_{2h}}}, \quad \text{for } h = 1, 2, \ldots, L. \tag{15}$$

And from (14), the n_{1h} is obtained as

$$n_{1h} = \frac{\sqrt{\lambda} W_h \sigma_{2h-j}}{\overline{n}_{2h}\sqrt{C_{2h}}}, \quad \text{for } h = 1, 2, \ldots, L. \tag{16}$$

Substituting the values of n_{1h} and $\overline{n}_{2h}$ from (15) and (16), respectively, formula (10) gives, when $V(p_j) = V$ (V is a given variance of p_j),

$$V = \sum_{h=1}^{L} W_h^2 \times \left[\frac{\overline{n}_{2h}\sqrt{C_{2h}}}{\sqrt{\lambda} W_h \sigma_{2h-j}}\left(\sigma_{1h-j}^2 - \frac{\sigma_{2h-j}^2}{\overline{N}_{2h}}\right) + \frac{\sqrt{C_{2h}}\sigma_{2h-j}^2}{\sqrt{\lambda} W_h \sigma_{2h-j}} - \frac{\sigma_{1h-j}^2}{N_{1h}}\right]. \tag{17}$$

Hence,

$$\sqrt{\lambda} = \left\{\sum_{h=1}^{L} W_h \sqrt{C_{2h}} \times \left[\frac{\overline{n}_{2h}}{\sigma_{2h-j}}\left(\sigma_{1h-j}^2 - \frac{\sigma_{2h-j}^2}{\overline{N}_{2h}}\right) + \sigma_{2h-j}\right]\right\} \times \left(V + \sum_{h=1}^{L} \frac{W_h^2 \sigma_{1h-j}^2}{N_{1h}}\right)^{-1}. \tag{18}$$

The minimum value of $V(p_j)$ under a cost function (fixed survey cost C), the optimum sampling size is obtained as the minimum values of function (10) subject to the constraint (9). Consider the following Lagrange function F:

$$F\left(n_{1h}, n_{1h}\overline{n}_{2h}, \lambda\right) = V\left(p_j\right) + \lambda\left(\sum_{h=1}^{L} C_{0h} + \sum_{h=1}^{L} C_{1h} n_{1h} + \sum_{h=1}^{L} C_{2h} n_{1h}\overline{n}_{2h} - C\right), \tag{19}$$

where λ is a Lagrange multiplier.

The optimums n_{1h} and $\overline{n}_{2h}$ are the solution of the following numerical problem:

$$\frac{\partial F}{\partial n_{1h}} = \lambda C_{1h} - \frac{W_h^2\left(\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}\right)}{n_{1h}^2} = 0,$$
$$\frac{\partial F}{\partial\left(n_{1h}\overline{n}_{2h}\right)} = \lambda C_{2h} - \frac{W_h^2 \sigma_{2h-j}^2}{n_{1h}^2 \overline{n}_{2h}^2} = 0. \tag{20}$$

Results are presented as follows:

$$n_{1h} = \frac{W_h \sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}}}{\sqrt{\lambda}\sqrt{C_{1h}}},$$
$$n_{1h}\overline{n}_{2h} = \frac{W_h \sigma_{2h-j}}{\sqrt{\lambda}\sqrt{C_{2h}}}. \tag{21}$$

We have the approximate optimal sample sizes given by

$$\overline{n}_{2h} = \frac{\sigma_{2h-j}}{\sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}}} \cdot \sqrt{\frac{C_{1h}}{C_{2h}}}, \quad \text{for } h = 1, 2, \ldots, L. \tag{22}$$

$$n_{1h} = \frac{W_h \sigma_{2h-j}}{\overline{n}_{2h}\sqrt{C_{2h}}} \cdot \frac{1}{\sqrt{\lambda}}, \quad \text{for } h = 1, 2, \ldots, L. \tag{23}$$

Define C as the value of the survey cost, from (21); the formula of the overall survey cost is expressed as

$$C - \sum_{h=1}^{L} C_{0h} = \sum_{h=1}^{L}\left(\frac{C_{1h} W_h \sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}}}{\sqrt{\lambda}\sqrt{C_{1h}}} + \frac{C_{2h} W_h \sigma_{2h-j}}{\sqrt{\lambda}\sqrt{C_{2h}}}\right). \tag{24}$$

Hence,

$$\frac{1}{\sqrt{\lambda}} = \frac{C - \sum_{h=1}^{L} C_{0h}}{\sum_{h=1}^{L} W_h\left(\sqrt{C_{1h}} \cdot \sqrt{\sigma_{1h-j}^2 - \sigma_{2h-j}^2/\overline{N}_{2h}} + \sqrt{C_{2h}} \cdot \sigma_{2h-j}\right)}. \tag{25}$$

For the h stratum, the optimum size of the sample of SSUs in each selected PSU is given by

$$n_{i2h} = N_{i2h} \cdot \frac{\overline{n}_{2h}}{\overline{N}_{2h}}, \quad \text{for } h = 1, 2, \ldots, L. \tag{26}$$

It is noted that the value of j may need to be considered in the process of estimating $\overline{n}_2$ and n_1. Difference of j value leads to difference of $\overline{n}_2$ and n_1. Taking the maximum value of $\overline{n}_2$ and n_1 is necessary to be ensured.

4. Applications

4.1. Preliminary Survey. Homosexual behaviour features were investigated in stratified two-stage sampling study of MSM living in Beijing from August to October 2010. The information was used to characterize high-risk sexual behaviours among MSM. All the respondents were arranged in two strata, the first consisting of MSM aged 15 to 29 years ($h = 1$) and the second consisting of MSM aged 30 to 49 years ($h = 2$). Districts/counties in Beijing were defined as PSUs. Beijing currently comprises 16 county-level subdivisions including 14 districts and 2 rural counties. Each stratum contained 16 districts/counties ($N_{11} = N_{12} = 16$). The MSM were considered as SSUs. We took this figure of 2.5% as being the proportion of adult males who were homosexually active in the city of Beijing. This suggested that there were 67750 MSM aged 15 to 49 years living in Beijing. An average of 2466 MSM and 1768 MSM were indicated in each district/county within stratum 1 and stratum 2, respectively ($\overline{N}_{21} = 2466$ and $\overline{N}_{22} = 1768$). In the first sampling stage, 13 districts/counties were randomly drawn within each stratum ($n_{11} = n_{12} = 13$), while in the second sampling stage, 1523 MSM were randomly selected from all the chosen subdivisions. In the first and second strata, the average of MSM was 68 and 49 drawn from each selected subdivision, respectively ($\overline{n}_{21} = 68$ and $\overline{n}_{22} = 49$).

The participants underwent an interview using polychotomous RRT model focusing on male-to-male sexual behaviour. The detailed information pertained to use condoms, each commercial same-sex behavioural cost, the proportion to engage in commercial same-sex services, HIV testing status, STD testing status, the preference for sexual behaviours, and latex condom failure. Sensitive quantitative variable closely followed a normal distribution in MSM population. And sensitive qualitative variable was associated with discrete probability distribution.

Take condom use, for example, which was particularly important for combatting the spread of HIV. This typical sensitive question seemed like "Did you use a new condom with every act of anal intercourse?" with answers "1—Never use," "2—Occasionally use," "3—Consistently use," and "4—Say no to anal sex." By these answers, respondents were classified into four mutual exclusive groups. Randomizing device was given to be a deck of cards identical in all respects number, labelled by the integers from 0 to 4. Fix the probabilities P_0, P_1, P_2, P_3, and P_4, so that $P_0 : P_1 : P_2 : P_3 : P_4 = 0.6 : 0.1 : 0.1 : 0.1 : 0.1$ ($P_0 + P_1 + P_2 + P_3 + P_4 = 1$). Each SSU (the selected MSM) was instructed to draw one card from the deck with replacement randomly. Drawing the card labelled with the number 0, the respondent revealed his true response whether he used a new condom during anal intercourse. Drawing the others, he disclosed the value of the chosen card.

In the first stratum, 66 MSM were randomly drawn in the district/county one ($n_{i2h} = n_{121} = 66$), 12 of those who gave answer 1 (when $j = 1$, $m_{ih-j} = m_{11-1} = 12$). And so the probability of answering 1 was 0.1818 ($\widehat{\lambda}_{11-1} = m_{11-1}/n_{121} = 12/66$). Randomizing device was set as follows. A participant either revealed his true type with probability P_0 ($P_0 = 0.6$) or answered 1 with probability $P_1 (P_1 = 0.1)$. From formula (8), therefore, we could approximately get the percentage of MSM who had never used condom for each act of anal intercourse in the district/county one from stratum 1: $p_{ih-j} = p_{11-1} = (\widehat{\lambda}_{11-1} - P_1)/P_0 = (12/66 - 0.1)/0.6 \doteq 0.1364$.

In a similar way, the proportions of MSM who had never used condom for each act of anal intercourse in other districts/counties within each stratum were obtained. Furthermore, s^2_{1h-j} and s^2_{2h-j} were given by the formulae (2), (4), and (5). Table 1 showed both these variances which were needed in the determination of optimum sample size.

4.2. Optimum Sample Size Estimation. We plan to conduct a formal investigation of stratified two-stage sampling design among the population of MSM in Beijing by the end of 2014. The way to guarantee confidentiality is to apply polychotomous RRT. Survey sample size, including the number of participants and districts/counties in the formal investigation, can be determined based on every response category of polychotomous sensitive question. Accordingly, both different sensitive topics and different response categories with respect to the same sensitive topic lead to variation in optimum sample sizes. It is proper to take the maximum value as the final optimum sample size. Taking the case of condom use, sample size determination is presented as follows.

Based on the preliminary investigation, the formal investigation's budget was given. The average cost of initiating the survey within each stratum was fifty thousand Yuan ($C_{01} = C_{02} = 100000$). And then the average cost of approaching to one district/county within each stratum was a hundred thousand Yuan ($C_{11} = C_{12} = 100000$). Also, the average cost of obtaining information on sensitive characteristics in one respondent from each stratum is fifteen Yuan ($C_{21} = C_{22} = 15$).

Table 1 indicated that related estimators of sample variance within each stratum, s^2_{11-1}, s^2_{21-1}, s^2_{12-1}, and s^2_{22-1}, were 0.0058, 0.0152, 0.0075, and 0.0154, respectively. From expressions (15) and (22), an average size of MSM who were needed to be recruited in each chosen district/county from stratum 1 and stratum 2, respectively, was given by

$$\begin{aligned} \overline{n}_{21} &= \frac{\sqrt{0.0152}}{\sqrt{0.0058 - 0.0152/2466}} \times \sqrt{\frac{100000}{15}} \doteq 132, \\ \overline{n}_{22} &= \frac{\sqrt{0.0154}}{\sqrt{0.0075 - 0.0154/1768}} \times \sqrt{\frac{100000}{15}} \doteq 117. \end{aligned} \tag{27}$$

Table 1: Variances necessary for sample size formulae.

Condom use	Stratum (h)	s_{1h-j}^2	s_{2h-j}^2
Never use ($j = 1$)	$h = 1$, MSM aged 15 to 29 years	0.0058	0.0152
	$h = 2$, MSM aged 30 to 49 years	0.0075	0.0154
Occasionally use ($j = 2$)	$h = 1$, MSM aged 15 to 29 years	0.0641	0.0206
	$h = 2$, MSM aged 30 to 49 years	0.0383	0.0208
Consistently use ($j = 3$)	$h = 1$, MSM aged 15 to 29 years	0.0573	0.0141
	$h = 2$, MSM aged 30 to 49 years	0.0378	0.0172
Say no to anal sex ($j = 4$)	$h = 1$, MSM aged 15 to 29 years	0.0135	0.0081
	$h = 2$, MSM aged 30 to 49 years	0.0127	0.0054

To minimize the survey cost under the constraint of sampling error, where the value of sampling error $V(p_1)$ was 0.000057 ($V = 0.000057$). From formula (18), we can get

$$\begin{aligned}\sqrt{\lambda} = &\left\{0.5824 \times \sqrt{15}\right.\\ &\times \left[\frac{132}{\sqrt{0.0152}}\left(0.0058 - \frac{0.0152}{2466}\right) + \sqrt{0.0152}\right]\\ &+ 0.4176 \times \sqrt{15}\\ &\left.\times \left[\frac{117}{\sqrt{0.0154}}\left(0.0075 - \frac{0.0154}{1768}\right) + \sqrt{0.0154}\right]\right\}\\ &\times \left(0.000057 + 0.5824^2 \times \frac{0.0058}{16}\right.\\ &\left. + 0.4176^2 \times \frac{0.0075}{16}\right)^{-1}\\ &\doteq 99072.0981.\end{aligned} \tag{28}$$

The number of districts/counties which were needed to be chosen within each stratum was given by formula (16):

$$\begin{aligned}n_{11} &= \frac{0.5824 \times \sqrt{0.0152} \times 99072.0981}{132 \times \sqrt{15}} \doteq 13,\\ n_{12} &= \frac{0.4176 \times \sqrt{0.0154} \times 99072.0981}{117 \times \sqrt{15}} \doteq 11.\end{aligned} \tag{29}$$

To minimize the sampling error for the fixed overall survey cost, where the value of the fixed overall survey cost C was 1000000 ($C = 1000000$), from formula (25), we can get

$$\begin{aligned}\frac{1}{\sqrt{\lambda}} = &(1000000 - 100000 - 100000)\\ &\times \left[0.5824\right.\\ &\times \left(\sqrt{100000}\right.\\ &\left.\times \sqrt{0.0058 - \frac{0.0152}{2466}} + \sqrt{15} \times \sqrt{0.0152}\right)\\ &+ 0.4176 \times \left(\sqrt{100000} \times \sqrt{0.0075 - \frac{0.0154}{1768}}\right.\\ &\left.\left. + \sqrt{15} \times \sqrt{0.0154}\right)\right]^{-1}\\ &\doteq 30855.4891.\end{aligned} \tag{30}$$

The number of districts/counties which need to be sampled from each stratum was given by formula (23):

$$\begin{aligned}n_{11} &= \frac{0.5824 \times \sqrt{0.0152}}{132 \times \sqrt{15}} \times 30855.4891 \doteq 4,\\ n_{12} &= \frac{0.4176 \times \sqrt{0.0154}}{117 \times \sqrt{15}} \times 30855.4891 \doteq 3.\end{aligned} \tag{31}$$

Table 2 summarized the $\overline{n}_{21}$, $\overline{n}_{22}$, n_{11}, and n_{12} in the other different extent of condom usage among MSM discussed in this research.

The determination of sample size for sampling survey may vary with different categories related to polychotomous sensitive topics. And so the maximum sample size is necessary to be ensured. According to the sampling survey on condom use among MSM, an average of 132 MSM and 117 MSM should be sampled in each chosen district/county in the first stratum and second stratum, respectively ($\overline{n}_{21} = 132$ and $\overline{n}_{22} = 117$). When $\overline{n}_{2h}$ was gotten, we could determine the number of MSM drawn from the ith district/county in the hth stratum by formula (26). For example, if a certain chosen district/county had 3342 MSM in the first stratum, the number of MSM drawn from this district/county in the first stratum should be $3342 \times 132/2466 \doteq 179$.

5. Discussion

We have earlier reported that sample size formulae associated with (stratified) multistage sampling survey on nonsensitive topics were derived [11]. However, sample size formulae for multistage sampling survey on sensitive characteristics are not yet available. The main purpose of this paper is to provide sample size determination for polychotomous RRT model for sensitive characteristics in a stratified two-stage sampling design. We extend the application of sample size formulae

Table 2: Sample size for occasional, consistent condom use, and never having anal sex among MSM in Beijing.

Condom use	$\overline{n}_{21}$	$\overline{n}_{22}$	Given sampling error $V(P_j)$			Given cost of survey C	
			$V(P_j)$	n_{11}	n_{12}	n_{11}	n_{12}
Occasionally use ($j = 2$)	46	60	0.000423	15	9	5	3
Consistently use ($j = 3$)	40	55	0.000385	15	9	5	3
Say no to anal sex ($j = 4$)	63	53	0.000102	15	10	4	3

for multistage sampling design from nonsensitive questions to sensitive questions.

China is currently undergoing a serious HIV epidemic [12]. Male-to-male sexual contact is one of the leading modes of HIV transmission [13]. There seems to be a trend of increasing HIV prevalence among MSM. MSM in China might have an important role in spreading the HIV-1 epidemic. The proposed method in this study seems to be an effective technique for obtaining more accurate population ratio estimates for sensitive qualitative characteristics among HIV-related high risk groups. What is more, sampling survey schemes under the project 81273188 which will commence in 2014 to estimate the quantities of HIV-related high risk groups have been completed on the basis of sample size formulae deduced in this study.

The principles of validity and reliability are fundamental cornerstones of the scientific method. A good way to assess a survey is in terms of its validity and reliability. Both high validity and reliability can be arguably considered as the most important criteria for good quality of survey. Treating validity and reliability in the RRT model for sensitive quantitative/qualitative characteristics under a complex survey is the recourse to correlation analysis of repeated survey data and Monte Carlo simulation in our previous studies [6, 14, 15]. These survey methods and statistical formulae showed high validity and reliability.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by a Grant (81273188) from National Natural Science Foundation and a Grant (CXLX13_839) from Postgraduate Research and Innovation Program of Jiangsu. The authors are grateful to Wei Li, Xiangyu Chen, Qiaoqiao Du, Mingrun Yu, and Xudong Li for their invaluable help in field investigations. The authors wish to thank reviewers for comments and suggestions.

References

[1] F. Esponda, "Negative surveys," http://arxiv.org/abs/math/0608176.

[2] M. Moshagen, *Multinomial randomized response models*, 2008.

[3] G. L. Tian, M. L. Tang, Z. Liu, M. Tan, and N. S. Tang, "Sample size determination for the non-randomised triangular model for sensitive questions in a survey," *Statistical Methods in Medical Research*, vol. 20, no. 3, pp. 159–173, 2011.

[4] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–66, 1965.

[5] W. Brannath, "Book Review: S. C. Chow and M. Chang 2007: adaptive design methods in clinical trials," *Clinical Trials*, vol. 6, no. 1, pp. 102–103, 2009.

[6] Q. Q. Du, G. Gao, W. Li, and X. Y. Chen, "Application of monte carlo simulation in reliability and validity evaluation of two-stage cluster sampling on multinomial sensitive question," in *Information Computing and Applications*, pp. 261–268, Springer, 2012.

[7] W. Li, G. Gao, Y. H. Ruan, X. Y. Chen, and Q. Q. Du, "Analysis of sensitive questions of MSM based on RRT," in *Information Computing and Applications*, pp. 273–279, Springer, 2012.

[8] X. K. Pu, G. Gao, Y. B. Fan, and M. Wang, "Stratified cluster sampling under multiplicative model for quantitative sensitive question survey," *Interciencia*, vol. 37, pp. 833–837, 2012.

[9] A. Ambainis, M. Jakobsson, and H. Lipmaa, "Cryptographic randomized response techniques," in *Public Key Cryptography—PKC 2004*, vol. 2947 of *Lecture Notes in Computer Science*, pp. 425–438, Springer, Berlin, Germany, 2004.

[10] G. Gao and S. G. Wang, "The estimation of sample size in stratified two-stage sampling," *Chinese Journal of Health Statistics*, vol. 15, pp. 51–53, 1998.

[11] J. F. Wang, G. Gao, Y. B. Fan et al., "The estimation of sample size in multi-stage sampling and its application in medical survey," *Applied Mathematics and Computation*, vol. 178, no. 2, pp. 239–249, 2006.

[12] K. H. Choi, H. Liu, Y. Guo, L. Han, J. S. Mandel, and G. W. Rutherford, "Emerging HIV-1 epidemic in China in men who have sex with men," *The Lancet*, vol. 361, no. 9375, pp. 2125–2126, 2003.

[13] G. Mumtaz, N. Hilmi, W. McFarland et al., "Are HIV epidemics among men who have sex with men emerging in the Middle East and North Africa?: a systematic review and data synthesis," *PLoS Medicine*, vol. 8, no. 8, Article ID e1000444, 2011.

[14] Z. D. Jin, H. R. Zhu, B. Yu, and G. Gao, "A monte-carlo simulation investigating the validity and reliability of two-stage cluster sampling survey with sensitive topics," *Computer Modelling and New Technologies*, vol. 17, pp. 65–79, 2013.

[15] M. Wang and G. Gao, "Cluster sampling and its application on quantitative sensitive questions," *Chinese Journal of Health Statistics*, vol. 25, pp. 586–598, 2008.