CSCanada
Progress in Applied Mathematics
Vol. 2, No. 1, 2011, pp. 67-72
DOI: 10.3968/j.pam.1925252820110201.Z55

ISSN 1925-251X [Print]
ISSN 1925-2528 [Online]
www.cscanada.net
www.cscanada.org

Two-stage Sampling on Additive Model for Quantitative Sensitive Question Survey and Its Application

LIU Peng[1,*]; GAO Ge[1]; HE Zhilong[1]; RUAN Yuhua[2]; LI Xudong[1]; YU Mingrun[1]

[1]School of Radiation Medicine and Public Health, Soochow University, Suzhou, China 215123
[2]STD and AIDS Prevention and Control Center, Beijing, China 100050
Supported by "National Natural Science Foundation of China" (No: 30972548)
*Corresponding author.
Address: School of Radiation Medicine and Public Health, Soochow University, Suzhou, China 215123. Email: cpulp@163.com

Received 24 June 2011; accepted 27 July 2011

Abstract
Objective To explore scientific sampling methods and corresponding formulas for quantitative sensitive question survey on two-stage random sampling. To provide scientific data for the prevention and control of high risk AIDS population in Beijing. **Methods** Additive model for quantitative sensitive question survey, two-stage random sampling, properties of variance and mean were used. **Results** Formulas for the estimation of the population proportions and its variance on additive model for quantitative sensitive question survey were deduced. The survey methods and formulas were employed successfully in the survey of the age of the first time when MSM having sex with men and the result was 21.9747. **Conclusion** The methods and corresponding formulas for two-stage sampling on additive model for quantitative sensitive question survey are feasible.
Key words
Sensitive questions; Additive model for randomized response technique; Two-stage sampling; MSM

LIU Peng, GAO Ge, HE Zhilong, RUAN Yuhua, LI Xudong, & YU Mingrun (2011). Two-stage Sampling on Additive Model for Quantitative Sensitive Question Survey and Its Application. *Progress in Applied Mathematics, 2*(1), 67-72. Available from: URL: http://www.cscanada.net/index.php/pam/article/view/j.pam.1925252820110201.Z55
DOI: http://dx.doi.org/10.3968/j.pam.1925252820110201.Z55

INTRODUCTION

The investigation of sensitive question often encountered in various areas of sample surveys. The sensitive problem refer to the problem that involve the privacy or interest of individuals and units, and most people think is not convenient to stand in public statement .Even in some case ,including criminal acts. Such as homosexuality, prostitution, illegitimate children, extramarital sex, property status, and drug abuse [1]. If the direct form was used in such kind of investigation, the respondent often refused to answer or deliberately lying, resulting in the exceptional system error: error of sensitive issues [2].

So scientific method must be used in the survey of sensitive question to ensure the authenticity and reliability. And so far, randomized-response technique (RRT) was considered as the most effective way to protect the privacy and increase the real rate. A meta-analysis by Gerty [3] shows that RRT yields the most valid prevalence estimates than other methods for sensitive questions. Before our research group start this project, the research at home and abroad are mainly restricted to simple random sampling technique. In this paper, formulas for the estimation of the population proportions and its variance on additive model for quantitative sensitive question survey were deduced and employed successfully in the survey of MSM in Beijing (Men who have sex with men [4], MSM) investigation.

1. SURVEY METHODS

1.1 Additive Model for Quantitative Sensitive Question

In this model, we design a random device: put 10 balls into a bag, each ball was same and numbered from 0 to 9. Every respondent (second-stage units) selected from the primary unit pick a ball from the bag with replacement and produce a number, then add the number to the value of his/her own quantitative sensitive questions. Then we will get the final result z. Fill the number z into his questionnaire. The whole procedure was carried out without the knowledge of investigator. So the investigator just gets the number provided by the respondents. So the privacy of respondent get protected because nobody else know the truth except themselves.

1.2 Two-stage Sampling on Additive Question Model for Quantitative Sensitive Question

Suppose the population contains N_1 primary units; the ith primary unit contains N_{i2} second- stage units, $i = 1, 2, \ldots, N_1$ and every primary unit contains $\bar{N}_2$ second-stage units on average; Then in the first stage of sampling, we randomly select n_1 primary units from the population; In the second stage, n_{i2} second-stage units are randomly selected from the ith selected primary unit; On average, $\bar{n}_2$ second-stage units are randomly selected from every selected primary unit. Every respondent are surveyed by additive model for quantitative sensitive questions.

2. FORMULAS DEDUCTION

Let y_{ijk} be the value of the jth second-stage unit of the ith primary unit, μ_i be the mean value of the population of the jth second-stage unit of the ith primary unit, and $\hat{\mu}_i$ is the estimate of μ_i. Then let $V(\hat{\mu}_i)$ be the variance of population, $v(\hat{\mu}_i)$ be the estimate of $V(\hat{\mu}_i)$. According to the result of Wang Jianfeng, Gao Ge [5], the estimate of μ

$$\hat{\mu} = \frac{\sum_{i=1}^{n_1} N_{i2}\hat{\mu}_i}{\sum_{i=1}^{n_1} N_{i2}} \tag{1}$$

The variance of the estimator $V(\hat{\mu})$ is:

$$V(\hat{\mu}) \doteq \frac{\sigma_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \frac{\sigma_2^2}{n_1\bar{n}_2}\left(1 - \frac{\bar{n}_2}{\overline{N}_2}\right) \tag{2}$$

Where

$$\sigma_2^2 = \frac{1}{\sum_{i=1}^{N_1} N_{i2}} \sum_{i=1}^{N_{i2}} \frac{N_{i2}}{N_{i2}-1} \sum_{j=1}^{N_{i2}} \left(y_{ij} - \mu_i\right)^2 = \frac{1}{\sum_{i=1}^{N_1} N_{i2}} \sum_{i=1}^{N_{i2}} N_{i2} V(\hat{\mu}_i)$$

The estimator of $\sigma_1^2$is shown to be

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{N_{i2}}{\overline{N}_2}\right)^2 (\hat{\mu}_i - \hat{\mu})^2 \tag{3}$$

The estimator of σ_2^2 is shown to be

$$S_2^2 = \frac{1}{\sum_{i=1}^{n_1} N_{i2}} \sum_{i=1}^{n_1} \frac{N_{i2}}{n_{i2}-1} \sum_{j=1}^{n_{i2}} \left(y_{ij} - \hat{\mu}_i\right)^2 = \frac{1}{\sum_{i=1}^{n_1} N_{i2}} \sum_{i=1}^{n_1} N_{i2} v(\hat{\mu}_i) \tag{4}$$

Let μ_{iZ}, $\hat{\mu}_{iZ}$ and s_{iz}^2 denote the population mean, the sample mean and the sample variance of all the values of answers in the ith first-stage units respectively. Let μ_Y be the mean of all the random numbers of the randomization device. According to the basic character of mean, we could obtain:

$$\mu_i = \mu_{iZ} - \mu_Y$$

So:

$$\hat{\mu}_i = \hat{\mu}_{iZ} - \mu_Y \tag{5}$$

According to the basic character of variance, the estimation variance of $\hat{\mu}_i$ is given by [6]:

$$v(\hat{\mu}_i) = s_{iz}^2 / n_{i2} \tag{6}$$

3. APPLICATIONS

3.1 Survey Design

The population of research is MSM aged from 15 to 50 of Beijing. According to the proportion that the MSM aged from 15 to 50 of Beijing accounted for about 1.0% of local male from 15 to 50 [7], we can estimate that the total number of MSM in Beijing aged from 15 to 50 is 57,213 .Let the district be the first unit ,there are 16 districts in Beijing (N_1=16); Let the MSM be the second-stage unit, there are about 3576 people of each district in average. ($\overline{N}_2$=3576). Two-stage random sampling was carried out. According to the formula of sample size given by Wang Jianfeng, Gao Ge [5]. At the first stage, we select 9 districts from16. (n_1=9); at the second stage, we pick 620 MSM from the chosen district of the first stage. The mean of each district is 69. Apply the survey on the individual of MSM with quantitative sensitive questions RRT models, the survey index is the age when they first have sex with man. We designed a randomization set for the respondents: a bag with 10 coins in it and every coin is exclusively numbered from 0 to 9. Every chosen respondent pick out a coin and add the number of the coin to the age that he have sex with man first time. Then fill the result in the questionnaire.

During the survey, the Soochow University, China CDC, CCAVG and the CDC of each district cooperate successfully, and we also get power support from MSM volunteers, MSM sites and MSM organizations. All of these ensure the representativeness of the sample and survey quality. All the questionnaires were reclaimed and up to grade at the ratio of 100%. And the data was processed and analyzed by Excel2003 and SAS9.13.

3.2 Results

In district 1, we pick 49 MSM as respondents The sample mean of all the answer values is$\hat{\mu}_{iz} = \hat{\mu}_{1z} = 25.2653$; The sample variance of all the answer values is $s_{iz}^2 = s_{1z}^2 = 38.4521$; The mean of all the random numbers in the randomization device is $\mu_Y = (0 + 1 + \cdots + 9)/10 = 4.5$. Thus we can calculate the average age by formula (5) is $\hat{\mu}_i = \hat{\mu}_1 = \hat{\mu}_{1Z} - \mu_Y = 25.2653 - 4.5 = 20.7653$. The estimate variance calculated by formula (6) is $v(\hat{\mu}_i) = v(\hat{\mu}_1) = s_{1z}^2/n_{12} = 38.4521/49 = 0.7847$

Similarly, we can also calculate the means and estimate variance of other districts. (Table. 1.)

Table 1
The Means and Estimate Variance of Age that MSM Have Sex with Man First Time in Two-stage Sampling on Additive Question Model RRT

District	$\hat{\mu}_{iZ}$	μ_Y	$\hat{\mu}_i$	$v(\hat{\mu}_i)$	$\hat{\mu}$
District 1	25.2653	4.5	20.7653	0.7847	21.9747
District 2	26.1569	4.5	21.6569	0.2005	
District 3	30.1667	4.5	25.6667	3.2108	
District 4	27.6000	4.5	23.1000	1.1626	
District 5	26.9744	4.5	22.4744	0.6100	
District 6	25.4967	4.5	20.9967	0.1865	
District 7	25.8333	4.5	21.3333	0.8140	
District 8	27.8462	4.5	23.3462	2.5037	
District 9	24.8571	4.5	20.3571	1.5416	

By formula (1), the estimate mean $\hat{\mu}$ is shown to be:

$$\hat{\mu} = \frac{\sum_{i=1}^{9} N_{i2}\hat{\mu}_i}{\sum_{i=1}^{9} N_{i2}} = \frac{(3328 \times 20.7653 + \cdots + 4062 \times 20.3571)}{(3328 + 10364 + \cdots + 4062)} = 21.9747$$

By formula (3) and (4), the sample estimator of σ_1^2 and σ_2^2 is shown to be:

$$\begin{aligned} S_1^2 &= \frac{1}{9-1}\sum_{i=1}^{9}\left(\frac{N_{i2}}{\overline{N}_2}\right)^2(\hat{\mu}_i - \hat{\mu})^2 \\ &= \frac{1}{9-1}\left[\left(\frac{3328}{3576}\right)^2 \times (20.7653 - 21.9747)^2 + \cdots + \left(\frac{4062}{3576}\right)^2 \times (20.3571 - 21.9747)^2\right] \\ &= 2.1152 \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{\sum_{i=1}^{9} N_{i2}}\sum_{i=1}^{9} N_{i2}v(\hat{\mu}_i) \\ &= \frac{1}{(3328 + \cdots + 4062)} \times [3328 \times 0.7847 + \cdots + 4062 \times 1.5416] \\ &= 0.4858 \end{aligned}$$

Put S_1^2 and S_2^2 into formula (2), we get the estimate of $V(\hat{\mu})$:

$$\begin{aligned} v(\hat{\mu}) &= \frac{S_1^2}{n_1}\left(1-\frac{n_1}{N_1}\right)+\frac{S_2^2}{n_1\bar{n}_2}\left(1-\frac{\bar{n}_2}{\overline{N}_2}\right) \\ &= \frac{2.1152}{9}(1-\frac{9}{16})+\frac{0.4858}{9\times 69}(1-\frac{69}{3576}) = 0.2482 \end{aligned}$$

Thus, we get the 95% confidence interval of the population mean of the age that the first time MSM have sex with man.

$$\hat{\mu} \pm 1.96 \times \sqrt{v(\hat{\mu})} = 20.9983 \sim 22.9510$$

CONCLUSION

Additive model has the advantage of smaller sample size requirements. The design is relatively simple. Regardless of the distribution of X is discrete or continuous type, their means can be estimated [8]. However, the application should be noted that it should according to the specific questions of investigation to determine the range of random variable Y. If the scope is too large will increase the s_{iz}^2, too small will reduce the protective effect on the respondents.

Validity refers to measure or observations in the extent to which the affairs of authenticity, pertaining to accuracy of data. Reliability is to point to in the same condition, the same objective thing of repetitive measure several times, the consistent extent with each other measurement results, pertaining to the reliability of data. The project members of this topic have carried out surface validity, content validity, criterion validity, reliability evaluation and harmony, test-retest reliability evaluation on various statistical methods that combine several RRT models with multiple sample methods [9, 10, 11]. The results indicate that the sample survey methods and statistics of our research have a high validity and reliability.

With the rapid development of China's economy, there come large numbers of sensitive question that affect social stability and economic health in all aspects of the social life. These sensitive questions often reflect the deep contradictions that exist in social and economic development. The government and society need to work together to reveal or expose these problems. So the investigation of sensitive question is becoming more and more important. The two-stage sampling for quantitative sensitive question survey introduced in this paper is scientific and feasible, and can be applicable to a wide range of sensitive question survey. It can provide scientific and accurate basis to state and local governments, and also has great significance to the prevention and control of AIDS high-risk groups.

REFERENCES

[1] LIU Wen, GAO Ge, & LI Xudong. (2010). Stratified Random Sampling on Simmons Model for Sensitive Question Survey. *Suzhou University Journal of Medical Science, 30*(4), 759-762.

[2] LI Xudong, GAO Ge, HE Zhilong, et al. (2009). Stratified Random Sampling on the Randomized Response Technique for Multi-class Sensitive Question Survey. *Suzhou University Journal of Medical Science, 29*(4), 668-670.

[3] Gerty J.L.M., Lensvelt-Mulders, Joop J.Hox, Peter G.M, & van der Heijden. (2005). Meta-analysis of Randomized Response Research: 35 Years of Validation Studies. *Sociological Methods & Research, 33*(3), 319-348.

[4] QIAN Yuesheng, FU Jihua, & BI Zhenqiang. (2006). MSM and HIV. *STD*, 6, 583-584.

[5] WANG Jianfeng, GAO Ge, FAN Yubo, CHEN Lilin, LIU Shengxue, JIN Yali, & YU Jinguo. (2006). The Estimation of Sampling Size in Multistage Sampling and Its Application in Medical Survey. *Applied Mathematics and Computation, 178*, 239-249.

[6] WANG Jianhua. (2003). *Practical Medical Research Methods*. Beijing: People's Medical Publishing House.

[7] WANG Liyan, XIA Dongyan, WU Yuhua, et al. (2006). Application of a Multiplier Method to Estimate the Population Size of Men Who Have Sex with Men. *South China J Prev Med, 32*(3), 9-15.

[8] ZHANG Yongqing, LU Wei, & YE Dongqing. (2003). Investigation Technique on Sensitive Problem of Quantity Characteristic. *Chinese Journal of Disease Control & Prevention, 7*(6), 542-544.

[9] WANG Mian, & GAO Ge. (2008). Quantitative Sensitive Question Survey in Cluster Sampling and Its Application. *Recent Advance in Statistics Application and Related Areas, 8*, 648-652.

[10] LIU Wen, GAO Ge, & WANG Lei. Stratified Random Sampling on the Simmons Model for Sensitive Question Survey. *Data Processing and Quantitative Economy Modeling, 10*, 22-26.

[11] YU Mingrun, GAO Ge, & LI Xudong. (2008). Strafed Two-stage Cluster Sampling on the Simmons Model for Sensitive Question Survey. *Recent Advance in Statistics Application and Related Areas, 8*, 801-805.