CSCanada
Studies in Mathematical Sciences
Vol. 3, No. 1, 2011, pp. 46-51
DOI: 10.3968/j.sms.1923845220110301.4Z095

ISSN 1923-8444 [Print]
ISSN 1923-8452 [Online]
www.cscanada.net
www.cscanada.org

Statistical Methods of Two-Stage Sampling on Simmons Model for Sensitive Question Survey with and Its Application

LI Wei[1]; GAO Ge[2,*]; HE Zhilong[1]

[1]School of Public Health, Soochow University, P.R. China, 215123
[2] School of Public Health, Soochow University, P.R. China, 215123
*Corresponding author. E-mail: tgliwei124@163.com
Received 10 July 2011; accepted 16 August 2011

Supported by (in part) National Natural Science Foundation of China (No. 30972548)

Abstract: To explore scientific survey methods and corresponding formulas for sensitive question survey on two-stage sampling. We use Simmons model for dichotomous sensitive questions, two-stage sampling, total probability formulas and properties of variance to deduce corresponding formulas. Then the formulas and its variance on Simmons model for dichotomous sensitive questions on two-stage sampling were designed and applied for the survey of the using rate of condoms among the Men who have sex with men in Beijing, the rate is 78.65% and its 95% confidence limit is 71.10% to 82.60%.
Keywords: Sensitive question; Simmons model; Two-stage sampling; MSM

LI Wei, GAO Ge and HE Zhilong (2011). Statistical Methods of Two-Stage Sampling on Simmons Model for Sensitive Question Survey with and Its Application. *Studies in Mathematical Sciences, 3*(1), 46-51. Available from: URL: http://www.cscanada.net/index.php/sms/article/view/j.sms.1923845220110301.4Z095. DOI: http://dx.doi.org/10.3968/j.sms.1923845220110301.4Z095.

INTRODUCTION

When we conduct a survey we may encounter some sensitive questions, which refers to questions with high personal confidentiality and can't be answered publically[1] .Sensitive questions include homo-sexuality, AIDS, drug abuse, corrupt practice etc. If we use traditional methods (such as trying to get information through telephone or asking questions in a face-to-face way) to direct an interview, many respondents may refuse to answer or tell lies because of the fear of revealing their privacy .It is obvious that traditional methods will make the survey exist response bias[2]when the survey involves sensitive questions. To solve the response bias, Warner initiated Randomized Response Technique(RRT)In 1965.RRT is considered as the most effective method to protect the privacy of respondents .It is also the best way to raise the rate of honest answer and get a reliable estimator of population[3]. However, sampling on sensitive questions is usually confined to simple random sampling. In this paper, we deduced Formulas for the estimation of the population proportions and its variance on Simmons model for dichotomous sensitive questions on two-stage sampling. And corresponding survey methods and formulas were successfully designed and applied in the survey of the using rate of condoms among the Men who have sex with men in Beijing of China.

1. SURVEY METHODS

1.1 Simmons Model

Simmons model which based on Warner's randomized response technique was put forward by Simmons in 1967[4]. In this design, respondents are given two unrelated questions. One is a sensitive question and the other is a non-sensitive question. Which question will be answered by respondents is depending on the outcome of a randomization set. The probability ($P, P \neq 0.5$) of selecting the sensitive question are determined during designing the randomization set. Depending on the outcome of the randomization process, the respondent answers "Yes" or "no" to the sensitive question or non-sensitive question.

1.2 Simmons Model in Two-Stage Sampling

Suppose the population contains N_1 first-stage units, the ith first-stage unit contains N_{i2} second-stage units ($i = 1, 2, \ldots, N_1$), and each of first-stage units has $\overline{N}_2$ second-stage units on average. In the first stage we randomly select n_1 first-stage units, then we draw n_{i2} second-stage units from the ith first-stage units ($i = 1, 2, \ldots, n_1$), and each selected first-stage unit has $\bar{n}_2$ second-stage units on average. Simmons model are used on each selected second-stage unit in the survey of dichotomous sensitive questions.

2. FORMULAS DEDUCTION

2.1 The Estimator of the Population Mean and Its Variance

We use p_i to represent the sample proportion of the ith first-stage unit of sensitive characteristic variables, From the formula and results given by Jianfeng Wang and Ge Gao[5], the estimator proportion (P) of the population is shown to be:

$$p = \frac{\sum_{i=1}^{n_1} N_{i2}\, p_i}{\sum_{i=1}^{n_1} N_{i2}} \tag{1}$$

The variance of P is given by

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

Here the estimator of σ_1^2 is given by

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

Also the estimator of σ_2^2 is shown to be

$$S_2^2 = \frac{1}{\sum_{i=1}^{n_1} N_{i2}}\sum_{i=1}^{n_1} N_{i2}\, p_i(1 - p_i) \tag{4}$$

And then the estimator of $V(p)$ is

$$v(p) = \frac{S_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \frac{S_2^2}{n_1 \bar{n}_2}\left(1 - \frac{\bar{n}_2}{\overline{N}_2}\right) \tag{5}$$

2.2 The Estimator of p_i

In the randomization set of Simmons model, the respondent will answer the sensitive question A when he selected the red ball, and when he selected the white ball, he will answer the non-sensitive question B. Suppose the probability of selecting the red ball is P', and the rate of people who have character B in ith first-stage unit is R_i(R_i can be got by references or by special investigation). We use π_i and p_i to represent the sample proportion and population proportion of the ith first-stage unit of sensitive characteristic variables. Let $\hat{\lambda}_i$ be the population proportion of people who answer "Yes" in ith first-stage unit, and the number who answer "Yes" in ith first-stage unit is m_i. Let $\hat{\lambda}_i$ be the estimator of λ_i.Then: $\hat{\lambda}_i = m_i/n_{i2}$. According to the total probability formulas[6]:

$$\lambda_i = \pi_i \cdot P' + (1 - P)R_i \quad \pi_i = \frac{\lambda_i - (1 - P')R_i}{P'}$$

$$p_i = \frac{\hat{\lambda}_i - (1 - P')R_i}{P'} = \frac{m_i/n_{i2} - (1 - P')R_i}{P'} i = 1, 2, \ldots, n_1 \tag{6}$$

3. APPLICATION

3.1 Survey Design

The study population is the men who have sex with men (MSM) aged from 15 to 50 in Beijing in 2010.The percent of MSM in all the male people aged from 15 to 50 in Beijing is 1.0%[7],thus we can calculate the total number of MSM aged from 15 to 50 in Beijing is 57213. We took the 16 districts in Beijing as the first-stage units ($N_1 = 16$). and took the MSM as the second-stage units. The average of MSM in each district is 3576($\overline{N}_2$=3576).Between August 1,2010, and October31, 2010, we conducted a two-stage sampling toward MSM in Beijing. According the method of estimating sampling size given by Jianfeng Wang and Ge Gao [5],we randomly select 9 districts (n_1 =9) in first-stage and randomly select 620 people from these 9 districts who are MSM in second stage, The average of MSM in each district is 69($\overline{n}_2$=69).

The randomization set is as following: the respondent was asked to draw a ball at random from a pack containing two colors (red and white) of balls. The red ball (with proportion P=0.6) and white ball (with proportion 0.4) have no other difference except color. When the respondent selected the red ball he should answer the sensitive question: "Did you use the condom during your latest anal sex?" When he selected the white ball, he will answer the non-sensitive question: "Are your birthday odd number?"

All the investigator were strictly trained before investigating to make sure the information they got is accurate enough. The recovery rate of the survey was 100%, with no failure questionnaire. The data was inputted twice using EpiData3.1 and was analysed by SAS9.13.

3.2 Results of Each District

There is 35 MSM who answered "Yes" in the first district (m_1=35), and the number of MSM randomly selected in this district is 45, thus the estimator of the population proportion of people who answer "Yes" in the first district is: $\hat{\lambda}_i = m_i/n_{i2} = 35/49$. The proportion MSM whose birthday is odd number is 0.5(R_i=0.5).

The probability of answering the sensitive question (or selecting the red ball) is 0.6(P '=0.6). Combining with formula (6) , we can calculate the using rate of condom during the latest anal sex of MSM in the first district as following:

$$p_i = \frac{\hat{\lambda}_i - (1-P)R_i}{P'} = \frac{m_i/n_{i2} - (1-P)R_i}{P'} = \frac{35/49 - (1-0.6)\times 0.5}{0.6} = 0.8571$$

Thus, we can get the using rate of condom during the latest anal sex of MSM in other districts (table 1).

Table 1
Condom Using Rate During the Latest Anal Sex of MSM in Nine Districts in Beijing

district	m_i	n_{i2}	$\hat{\lambda}_i$	p_i	p
1	35	49	0.7143	0.8571	
2	138	204	0.6765	0.7941	
3	11	18	0.6111	0.6852	
4	28	40	0.7000	0.8333	
5	51	78	0.6538	0.7564	0.7865
6	114	153	0.7451	0.9085	
7	17	24	0.7083	0.8472	
8	14	26	0.5385	0.5641	
9	17	28	0.6071	0.6786	

3.3 Results of Population

Combining with formula (1), we can calculate the using rate of condom during the latest anal sex of MSM in Beijing.

$$p = \frac{\sum_{i=1}^{9} N_{i2}\, p_i}{\sum_{i=1}^{9} N_{i2}} = \frac{(3328\times 0.8571 + \cdots + 4062\times 0.6786)}{(3328 + 10364 + \cdots + 4062)} = 0.7865$$

Combining with formula (3) and (4), the estimator of σ_1^2andσ_2^2 can be obtained:

$$\begin{aligned} S_1^2 &= \frac{1}{n_1 - 1}\sum_{i=1}^{9}\left(\frac{N_{i2}}{\overline{N}_2}\right)^2 (p_i - p)^2 \\ &= \frac{1}{9-1}\left[\left(\frac{3328}{3576}\right)^2 \times (0.8571 - 0.7865)^2 + \cdots + \left(\frac{4062}{3576}\right)^2 \times (0.6786 - 0.7865)^2\right] \\ &= 0.0254 \end{aligned}$$

$$\begin{aligned} S_2^2 &= \frac{1}{\sum_{i=1}^{9} N_{i2}}\sum_{i=1}^{9} N_{i2}\, p_i(1-p_i) \\ &= \frac{1}{(3328 + \cdots + 4062)} \times [3328\times 0.8571\times(1-0.8571) + \cdots + 4062\times 0.6786\times(1-0.6786)] \\ &= 0.1582 \end{aligned}$$

And then the estimator of $V(p)$is:

$$v(p) = \frac{S_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \frac{S_2^2}{n_1 \bar{n}_2}\left(1 - \frac{\bar{n}_2}{\overline{N}_2}\right) = \frac{0.0254}{9} \times (1 - \frac{9}{16}) + \frac{0.1582}{9\times 69} \times (1 - \frac{69}{3576}) = 0.0015$$

Thus, we can calculate the 95% confidence limit of the condom using rate during the latest anal sex of MSM in Beijing:

$$p \pm 1.96 \times \sqrt{v(p)} = 0.7865 \pm 1.96 \times \sqrt{0.0015} = 0.7110 \sim 0.8620$$

DISCUSSION

Sensitive questions are everywhere and usually you will not get honest reply if you try to ask someone these questions in a direct way. Some sensitive questions is highly important to a country to establish public policies. One of these questions is AIDS. It is necessary for us to study some feasible methods to acquire accurate information when conducting a survey on sensitive questions. This paper supplies a method and corresponding formulas on Simmons model in two-stage sampling, it have some meaningful in statistic technique.

We use reliability and validity to evaluate whether a measurement methods can reflect the character of objective things reliably and truly, This paper is one parts of the National Natural Science Foundation project of China. Before this paper, a certain number related methods were studied which were combined several RRT models with complicated sampling technique. Formulas were deduced and all these methods have highly reliability and validity[8,9,10].Which indicated that our methods and corresponding formulas are feasible to obtain real data of sensitive issues in a wide range of area.

In order to decrease the bias and get accurate data, two points should be pay attention to when conducting the two-stage sampling. One is training investigator so that each investigator has an overall understanding on two-stage sampling; another is that the sample size is feasible to make sure the sample can represent the population reliably.

Presently there are about 5-10million MSM in China, these MSM usually have several sexual partners and have sexual intercourse without protection[11]. The MSM has become the second high risk group of infection of AIDS in China. The infection rate of AIDS among MSM has increased from 3% to 10% for the past 3years in some city of China[12]. It is important for us to acquire the characters of sexual behavior among MSM so that we can make usable measurement to control the infection of AIDS. We selected the Beijing as the survey area because Beijing is one of the chief city where amount of MSM gather there. Our study shows that the using rate of condoms among the MSM in Beijing is 78.65% and its 95% confidence limit is 71.10% to 82.60%, Thus it can be seen there are still lots of MSM (about 21.8%) have unsafe sexual behavior. We have known using condoms can prevent the infection of AIDS evidently[13], so we should enhance the education on safe sexual behavior towards MSM, preventing the AIDS spreading in China even in the world.

REFERENCES

[1] DING Yuanlin, GAO Ge (2008). *Health statistics*. Beijing: Science Press, 13. (in Chinese)

[2] WANG Chunping, WANG Zhifeng and ZHANG Guangcheng (2006). Design, Analysis, Evaluation for Nature Character Sensitive Questions. *Chinese Journal of Health Statistics, 23* (1), 60-62. (in Chinese)

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

[4] Simmons WR, Horvitz DG, Shah BV, et al (1969). The Unrelated Question Randomized Response Model. *Proceedings in the Social Statistics Section, American Statistical Association, 64* (326), 520-

539.

[5] WANG Jianfeng, GAO Ge, FAN Yubo et al (2006). The Estimation of Sampling Size in Multi-Stage Sampling and Its Application in Medical Survey. *Applied Mathematics and Comptation, 178* (2), 239-249.

[6] SU Liangjun (2007). *Advanced Mathematical Statistic*. Beijing: Beijing University Press, 3. (in Chinese)

[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 (MSM). *South China Journal of Preventive Medicine, 32* (3), 9-15.

[8] GAO Ge, FAN Yubo (2008). Stratified Cluster Sampling and Its Application on the Simmons RRT Model for Sensitive Question Survey. *Chinese Journal of Health Statistics, 25*(6), 562-565, 569. (in Chinese)

[9] WANG Mian, GAO Ge (2008). *Quantitative Sensitive Question Survey in Cluster Sampling and Its Application. Recent Advance in Statistics Application and Related Areas*. Sydeny: Aussino Academic Publishing House, 648-652.

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

[11] CAO Gan, GUAN Wenhu, WU Xiaogang, et al (2007). Study on Infection Rate of HIV/SYPHILIS among Men Who Have Sex with Men in a Balneary. *Acta Universitatis Medicinalis Nanjing, 27* (6), 637-640. (in Chinese)

[12] Ministry of Health of the People's Republic of China, UNAIDS, WHO. Epidemic Situation and Preventing Progress of AIDS in China in 2005[R].2006.

[13] Z. Mukandavire, K. Bowa and W. Garira (2007). Modelling Circumcision and Condom Use as HIV/AIDS Preventive Control Strategies. *Mathematical and Computer Modelling, 46*(11-12), 1353-1372.