STRATIFIED CLUSTER SAMPLING UNDER MULTIPLICATIVE MODEL FOR QUANTITATIVE SENSITIVE QUESTION SURVEY

Pu Xiangke, Gao Ge, Fan Yubo and Wang Mian

SUMMARY

Advanced sampling methods and corresponding formulas are seldom applied to quantitative sensitive question survey. This paper develops a series of formulas for parameter estimation in cluster sampling and stratified cluster sampling under multiplicative model on the basis of classic sampling theories and total probability formulas. The performance of the sampling methods and formulas in survey of cheating on exams on Dushu Lake Campus of Soochow University, China, are provided. The reliability of the survey methods and formulas for quantitative sensitive question survey is found to be high.

Introduction

So-called sensitive questions are some private issues such as acquired immune deficiency syndrome (AIDS), drug addiction, gambling, prostitution, driving while intoxicated, personal income, tax evasion, premarital sex, venereal diseases, homosexual tendencies, and so on. It is difficult to get honest answers by asking direct questions on sensitive issues in survey research (Tourangeau and Yan, 2007). To reduce the response error, randomized response models have been developed for collecting sensitive information after the pioneering work of Warner (1965). The multiplicative model was designed for quantitative sensitive question survey, with advantages of simple design, small sample size and small sampling error (Wang, 2003; Raghunath and Georg, 2006).

Simple random sampling has been considered a useful sampling method under multiplicative model for quantitative sensitive question survey. However, this method is vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population and it is cumbersome when sampling from a large target population. Stratified sampling is superior to simple random sampling in reducing sampling error and cluster sampling is less costly, but these methods are not so popular because they are not too simple (Ding and Gao, 2008). Furthermore, formulas for simple random sampling may also be wrongly used when a relatively complicated sampling procedure is conducted in sensitive question survey and evaluation of reliability of sampling methods and corresponding formulas under randomized response models for sensitive question survey has seldom been set.

In this paper, designs for cluster sampling and stratified cluster sampling under multiplicative model and corresponding formulas for parameter estima-

KEYWORDS / Multiplicative Model / Parameter Estimation / Quantitative Sensitive Question / Stratified Cluster Sampling /

Received: 09/01/2010. Modified: 09/05/2011. Accepted: 09/07/2011.

Pu Xiangke. Ph.D. candidate in Epidemiology and Health Statistics, Soochow University, China. Technologist-in-charge, Institute of Hepatology, Third People's Hospital, Changzhou, China. e-mail: puxk@live.cn

Gao Ge. M.Sc. in Biostatistics, Soochow University, China. Professor, Soochow University, China. Address: Department of Health Statistics, School of Radiation Medicine and Public Health, Suzhou, 215123, P.R.China. e-mail: gaoge@suda.edu.cn

Fan Yubo. Ph.D. candidate in Epidemiology and Health Statistics. Soochow University, China.

Wang Mian. M.Sc. in Epidemiology and Health Statistics. Soochow University, China.

MUESTREO POR CONGLOMERADOS ESTRATIFICADOS BAJO UN MODELO MULTIPLICATIVO PARA PREGUNTAS SENSIBLES EN ENCUESTAS CUANTITATIVAS

Pu Xiangke, Gao Ge, Fan Yubo y Wang Mian

RESUMEN

Los métodos de muestreo avanzados y las correspondientes fórmulas son raramente aplicados en preguntas sensibles de encuestas cuantitativas. Este trabajo desarrolla una serie de fórmulas para parámetros de estimación en muestreo de conglomerados estratificados bajo un modelo multiplicativo basadas en teorías clásicas de muestreo y fórmulas de probabilidad total. Se ilustra el desempeño de los métodos de muestreo y las fórmulas en una encuesta de engaños en exámenes en el Campus Dushu Lake de la Universidad de Soochow, China. La confiabilidad de los métodos de encuesta y las fórmulas para encuestas cuantitativas de preguntas sensibles resulta ser alta.

AMOSTRAGEM ESTRATIFICADA POR CONGLOMERADOS SOB UM MODELO MULTIPLICATIVO PARA PERGUNTAS SENSÍVEIS EM PESQUISAS QUANTITATIVAS

Pu Xiangke, Gao Ge, Fan Yubo e Wang Mian

RESUMO

Os métodos avançados de amostragem e as correspondentes fórmulas são raramente aplicados em perguntas sensíveis de pesquisas quantitativas. Este trabalho desenvolve uma série de fórmulas para estimação de parâmetros em amostragem de conglomerados estratificados sob um modelo multiplicativo baseadas em teorias clássicas de amostragem e fórmulas de probabilidade total. Ilustra-se o desempenho dos métodos de amostragem e as fórmulas em uma pesquisa de enganos em exames no Campus Dushu Lake da Universidade de Soochow, China. A confiabilidade dos métodos de pesquisa e as fórmulas para pesquisas de perguntas sensitivas quantitativas resultam ser alta.

tion are provided. These complex sampling methods have been employed in survey of cheating on Dushu Lake Campus of Soochow University, China, and may be applicable to a large population.

Methods for Quantitative Sensitive Question Survey

Multiplicative model

In the multiplicative model (Wang, 2003), a randomized device is designed to randomly generate an integer between 0 and 9. In the device, ten balls of identical size are respectively labeled by the ten integers. By a randomization process, each respondent takes a ball labeled by an integer and multiplies the integer by the numerical value of his response to the quantitative sensitive question so as to get a final result.

Cluster sampling under the multiplicative model on quantitative sensitive questions

It is convenient to use a cluster sampling method. The population is divided into several clusters (primary units), and each cluster is composed of secondary units. Some clusters are randomly selected from the population. Then, the multiplicative model (above) is applied to all the secondary units of selected clusters for the quantitative sensitive question survey.

Stratified cluster sampling under the multiplicative model

Before cluster sampling, the population is divided into different strata by characters. Then, cluster sampling is applied to each stratum. In each stratum, different clusters (primary units) are also composed of secondary units. And then, the multiplicative model is employed to those secondary units of selected clusters for the quantitative sensitive question survey.

Deduction of Formulas

Formulas for cluster sampling

Suppose the population is divided into N clusters, and the ith cluster contains M_i subunits. Then, n clusters are drawn randomly from the population.

Estimation of the population mean and its variance for quantitative sensitive question survey. Suppose the mean of the response values in the ith cluster is μ_i, y_i denotes the sum of values in the ith cluster and the population mean is μ. By Cochran (1977) and Wang and Gao (2006) the estimator of the population mean can be stated as

$$\hat{\mu}=\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i},\quad i=1,2,\ldots,n \tag{1}$$

and when Mi=M,

$$\hat{\mu}=\frac{1}{nM}\sum_{i=1}^{n} y_i=\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{M}=\frac{1}{n}\sum_{i=1}^{n}\mu_i \tag{2}$$

Also following Cochran (1977), the estimator of the variance of $\hat{\mu}$ can be obtained as

$$v(\hat{\mu})=\frac{1-f}{n\bar{M}^2}\frac{\sum_{i=1}^{n} y_i^2-2\hat{\mu}\sum_{i=1}^{n} y_i M_i+\hat{\mu}^2\sum_{i=1}^{n} M_i^2}{n-1} \tag{3}$$

where $\bar{M}=\sum_{i=1}^{n} M_i/n$ is the mean of the subunits in each cluster, and

$f=\frac{\sum_{i=1}^{n} M_i}{\sum_{i=1}^{N} M_i}$ is the sampling ratio.

When Mi=M,

$$v(\hat{\mu})=\frac{1-f}{n(n-1)}\sum_{i=1}^{n}(\mu_i-\hat{\mu})^2 \tag{4}$$

where f=nM / NM=n/N is the sampling ratio.

Calculation of μ_i *and* y_i. Suppose μ_i denotes the mean of variables with sensitive character in the ith cluster, μ_{iz} denotes the mean of numerical values of the answers in the ith cluster, and μ_y denotes the mean of all the random numbers in the randomized device. Then, from characteristics of means (Wang *et al*., 2006),

$$\mu_{iz}=\mu_i\,\mu_y,\ i=1,2,\ldots,n \tag{5}$$

$$\mu_i=\mu_{iz}/\mu_y,\ i=1,2,\ldots,n \tag{6}$$

and $y_i=M_i\mu_i$, i= 1,2,...,n

Formulas for stratified cluster sampling

Suppose the population is composed of L strata, and

the hth stratum contains N_h clusters (primary units), and the ith cluster contains M_{ih} secondary units. The population contains N secondary units and n_h clusters are randomly drawn from the hth stratum.

Estimation of the population mean and its variance of the hth stratum. From Eq. 1, the estimator of the population mean (μ_h) of the hth stratum can be obtained as

$$\hat{\mu}_h = \frac{\sum_{i=1}^{n_h} y_{ih}}{\sum_{i=1}^{n_h} M_{ih}}, \quad h= 1,2,\ldots, L \qquad (7)$$

and from Eq. 3, we can get the estimator of the variance of $\hat{\mu}_h$:

$$v(\hat{\mu}) = \frac{1-f_h}{n_h \bar{M}_h^2} \frac{\sum_{i=1}^{n_h} y_{ih}^2 - 2\hat{\mu}_h \sum_{i=1}^{n_h} y_{ih} M_{ih} + \hat{\mu}_h^2 \sum_{i=1}^{n_h} M_{ih}^2}{n_h - 1}, \quad h= 1,2,\ldots,L \qquad (8)$$

where $\bar{M}_h = \sum_{i=1}^{n_h} M_{ih}/n_h$ is the mean of subunits of each cluster in the hth stratum, and $f_h = \frac{\sum_{i=1}^{n_h} M_{ih}}{\sum_{i=1}^{N_h} M_{ih}}$, which is the sampling ratio in the hth stratum.

When $M_{ih}=M_h$, from Eq. 2, we can also get the estimator of the population mean:

$$\hat{\mu}_h = \frac{1}{n_h M_h} \sum_{i=1}^{n_h} y_{ih} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{y_{ih}}{M_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} \mu_{ih}, \quad h= 1,2,\ldots,L \qquad (9)$$

and from Eq. 4, the estimator of the variance of $\hat{\mu}_h$ can be obtained as

$$v(\hat{\mu}_h) = \frac{1-f_h}{n_h (n_h-1)} \sum_{i=1}^{n_h} (\mu_{ih} - \hat{\mu}_h)^2, \quad h= 1,2,\ldots,L \qquad (10)$$

Estimation of the population mean and its variance. According to Cochran (1977), the estimator of the population mean is

$$\hat{\mu} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{ih} \hat{\mu}_h}{N} = \sum_{h=1}^{L} W_h \hat{\mu}_h \qquad (11)$$

where $W_h = \sum_{i=1}^{N_h} M_{ih}/N$ is the relative size of the hth stratum according to the number of subunits.

As samples of each stratum are independent, from Eq. 11, its variance is

$$V(\hat{\mu}) = \sum_{h=1}^{L} W_h^2 V(\hat{\mu}_h) \qquad (12)$$

According to Eqs. 8 and 10, the estimator of $V(\hat{\mu}_h)$ is $v(\hat{\mu}_h)$. And from Eq. 12, $v(\hat{\mu})$ will be the estimator of $V(\hat{\mu})$.

Calculation of μ_{ih} *and* y_{ih}

Let μ_{ih} denote the mean of variables with sensitive character in the ith cluster of the hth stratum, μ_{iz} denote the average value of all the answers in the ith cluster of the hth stratum, and μ_y be the average value of all the random numbers in the randomized device. Then, from characteristics of means (Wang *et al*., 2006; Su, 2007), we can get

$$\mu_{ihZ} = \mu_{ih}\, \mu_Y \qquad (13)$$

Thus,

$$\mu_{ih} = \mu_{ihZ} / \mu_Y, \; i= 1,2,L,n_h; \quad i= 1,2,L,L \qquad (14)$$

and we obtain $y_{ih} = M_{ih}\, \mu_{ih}$.

Survey of Cheating on Exams on Dushu Lake Campus of Soochow University

Let the students on Dushu Lake Campus of Soochow University be the target population, which is divided into two strata. Define undergraduates as the first stratum which contains 9689 students and graduates as the second stratum which contains 1890 students, and we could get $W_1= 9689/(9689+1890)\approx 0.84, W_2\approx 0.16$

Let every class be a cluster, and clusters in each stratum are approximately the same size. Twenty clusters containing 1080 students were randomly drawn from undergraduates and 18 clusters that contain 818 students were drawn from graduates, there being 1898 students in the selected 38 clusters. Each student was repeatedly surveyed twice at different times, and 3796 surveys were conducted in total. All the questionnaires were recovered and the passing rate of the questionnaires was 100%. A Data bank was established IN Excel2003 and all the data was analyzed by SAS9.13.

In the randomized device, there were 10 balls of identical size in a bag and the balls were respectively tagged by integers from 0 to 9. Each chosen student was asked to select a ball from the bag to get the corresponding integer, multiplied it by the number to times of his cheating on exams in the last two semesters, and wrote down the final result.

TABLE I
THE MEAN OF TIMES OF CHEATING OF STUDENTS IN 38 CLASSES IN LAST TWO SEMESTERS BY TWICE REPEATED STRATIFIED CLUSTER SAMPLING UNDER MULTIPLICATIVE MODEL

Serial numbers of undergraduate classes	μ_{i1}	μ'_{i1}	Serial numbers of graduate classes	μ_{i2}	μ'_{i2}
1	0.8085	0.9787	1	1.3990	1.3788
2	1.9222	1.7333	2	2.3308	2.6925
3	1.2278	1.4722	3	2.3457	2.2370
4	1.5934	1.5177	4	0.8322	1.1111
5	1.0505	0.9748	5	0.9660	0.9669
6	0.6333	0.4858	6	0.8133	0.9622
7	1.1358	0.9926	7	0.9235	0.8889
8	1.0922	1.0118	8	0.7488	0.9220
9	0.9827	0.8494	9	0.5383	0.5333
10	1.0431	1.1066	10	0.8009	0.7824
11	1.1376	1.4233	11	0.7565	0.7488
12	0.7727	0.6010	12	0.7546	0.7593
13	1.2691	1.1012	13	0.7565	0.7234
14	0.8636	0.8232	14	0.4921	0.4815
15	0.9082	0.9034	15	0.9975	0.9728
16	1.0000	0.8485	16	0.8931	0.8637
17	0.1401	0.0821	17	0.7677	0.6717
18	0.9312	0.9206	18	0.9333	0.8404
19	0.8360	0.7196			
20	0.4840	0.5284			

Survey on the mean of cheating times in each class

Through twice repeated surveys on cheating on exams in the last two semesters by stratified cluster sampling under the multiplicative model and using Eq. 13, it is also possible to get (Table I) the mean of times of cheating on exams of undergraduates in 20 classes in the first survey (μ_{i1} (i= 1, 2, …, 20)), the mean of times of cheating on exams of undergraduates in 20 classes in the second survey (μ'_{i1} (i= 1, 2, ….20)), the mean of times of cheating on exams of graduates in 18 classes in the first survey (μ_{i2} (i= 1, 2,…, 18)), and the mean of times of cheating on exams of graduates in 18 classes in the second survey (μ'_{i2} (i= 1, 2,…, 18).

Estimation of the population mean and variance of cheating times in each stratum

Population mean of cheating times in each stratum. By the first survey on undergraduates and Eq. 9, we can get the es-

timator of the population mean of cheating times of undergraduates in last two semesters:

$$\hat{\mu}_1=\frac{1}{20}(0.8085+1.9222+L+0.4840)$$
$$=0.9916$$

By the first survey on graduates and Eq. 9, we can also get the estimator of the population mean of cheating times of graduates in last two semesters:

$$\hat{\mu}_2=\frac{1}{18}(1.3990+2.3308+L+0.9333)$$
$$=1.0028$$

Variance of the population mean of cheating times in each stratum. By the first survey on undergraduates and Eq. 10, we can get the estimator of the variance of the population mean of cheating times of undergraduates in last two semesters:

$$v(\hat{\mu}_1)=\frac{1-f_1}{n_1(n_1-1)}\sum_{i=1}^{n_1}(\mu_{i1}-\hat{\mu}_1)^2$$
$$=\frac{1-1080/9689}{20(20-1)}\left[(0.8085-0.9916)^2+(1.9222-0.9916)^2+\cdots+(0.4840-0.9916)^2\right]$$
$$=0.0062$$

By the first survey on graduates and Eq. 10, we can also get the estimator of the variance of the population mean of cheating times of graduates in last two semesters:

$$v(\hat{\mu}_2)=\frac{1-f_2}{n_2(n_2-1)}\sum_{i=1}^{n_2}(\mu_{i2}-\hat{\mu}_2)^2$$
$$=\frac{1-818/1890}{18(18-1)}\left[(1.3990-1.0028)^2+(2.3308-1.0028)^2+\cdots+(0.9333-1.0028)^2\right]$$
$$=0.0086$$

Population mean and variance of cheating times of students on Dushu Lake Campus

From Eq. 11 the estimator of the population mean of cheating times of students on Dushu Lake Campus is

$$\hat{\mu}=\sum_{h=1}^{2}W_h\hat{\mu}_h=0.84\times0.9916+0.16\times1.0028=0.9934$$

And from Eqs. 10 and 12, we can get the estimator of the variance of the population mean of cheating times of students on Dushu Lake Campus as

$$v(\hat{\mu})=\sum_{h=1}^{2}W_h^2v(\hat{\mu}_h)=0.84^2\times0.0062+0.16^2\times0.0086=0.0046$$

Thus, the 95% confidence interval of the population mean of cheating times of students on Dushu Lake Campus of Soochow University is given by

$$\hat{\mu}\pm1.96\sqrt{v(\hat{\mu})}=0.9934\pm1.96\sqrt{0.0046}=0.8605\sim1.1263$$

Reliability evaluation

Correlative analysis was applied to the data of two repeated surveys under the multiplicative model from 20 undergraduate classes by using SAS9.13. The Shapiro-Wilks' W test was applied to μ_{i1} and μ'_{i1}, and the corresponding W values were 0.951357 and 0.966911 respectively; the P values were 0.3882 and 0.6888, respectively, and were normally distributed. This analysis showed that the coincidence between the results of the two repeat cluster samplings in the first stratum was high (coefficient of product-moment correlation r= 0.9351, P<0.0001). Rank correlation analysis was applied to the data of two repeated surveys under the multiplicative model from 18 graduate classes by using SAS9.13 and showed that the coincidence between results of two repeat cluster samplings in the second stratum was high (the Spearman rank correlation coefficient r_s= 0.83634, P<0.0001, not normal distribution). Finally, rank correlation analysis was applied to the data of two repeated surveys under the multiplicative model from all the 38 classes by using SAS9.13 and showed that the coincidence of results between two repeat stratified cluster samplings was high (the Spearman rank correlation coefficient r_s= 0.90311, P<0.0001, not normal distribution), which also indicate that the reliability of our survey methods and its formulas was quite high.

Discussion

Sensitive question survey is very popular and important in social and medical research, especially for the prevention of AIDS. Through the introduction period and the growth period of AIDS, China is now facing the threat of an AIDS outbreak. To prevent the disease accurate data about it are needed, although many people may be unwilling to give honest answers to such a sensitive question. Survey methods and formulas for parameter estimation proposed in this article will be helpful to get reliable data to prevent sexually transmitted diseases and improve public health.

Randomized response models have been widely used to make people cooperative in sensitive question survey. Recently, a meta-analysis was applied to 38 relevant papers published from 1965 to 2000 and showed that the application of randomized response models has a significant advantage of accuracy and reliability compared with other traditional survey methods (Lensvelt-Mulders *et al*., 2005). And the multiplicative model is generally superior to the previously used models of randomizing questions for quantitative data (Pollock and Bek, 1976; Yan and Nie, 2005). As to the sampling design for sensitive question survey, statisticians have provided several sampling methods. But so far, researches of sampling design have been limited to simple random sampling, and studies on the evaluation of the reliability and validity of sample survey on sensitive questions are rare (Wang and Gao, 2008; Gao and Fan, 2008).

In this paper, formulas for parameter estimation in cluster sampling and stratified cluster sampling under multiplicative model were deduced and quantitative data about sensitive issues could be easily acquired. Under the multiplicative model, cluster sampling and stratified cluster sampling were successfully applied to the survey of cheating on Dushu Lake Campus of Soochow University. And in the evaluation of the test-retest reliability, the coincidence of results between two repeated surveys was high, showing that the survey methods and statistical formulas are highly reliable.

For cluster and stratified cluster sampling, the sample size is generally large and the sample mean usually follows the normal distribution. With the formulas herein deduced, one can get estimators of population means and their variances, estimate the intervals of those population means and, further, compare the mean of each stratum by using either t-test, Z test, analysis of variance or rank test. The study of cluster sampling and stratified cluster sampling methods and relevant formulas under multiplicative model will likely be helpful for survey of quantitative sensitive issues.

ACKNOWLEDGEMENTS

The authors thank Yongzhong Wang, Shuangrong Hang, Min Chen and Hongyu Shen and the Third People's Hospital of Changzhou, for their encouragement, and the referees and editors for their suggestions and help.

This work was supported by grants from the National Natural Science Foundation of China (Nº 81273188, to Gao Ge), the Preventive Medicine Research Project of Jiangsu Province (Y2012072, to Pu Xiangke) and the Applied Basic Research Program of Changzhou (CJ20112013, to Pu Xiangke).

REFERENCES

Cochran WG. (1977) *Sampling Techniques*. 3rd ed. Wiley. New York; USA. 288-289 pp.

Ding Y, Gao G (2008) *Health Statistics*. Science Press. Beijing, China. 13-20 pp.

Gao G, Fan Y (2008) The research of stratified cluster sampling on simmons model for sensitive question survey. *Chin. J. Health Stat. 25*: 562-569.

Lensvelt-Mulders GJLM, Hox JJ, van der Heijden PG.M, Maas CJM (2005) Meta-analysis of randomized response research, thirty-five years of validation. *Sociol. Meth. Res. 33*: 319-348.

Pollock KH, Bek Yuksel (1976) A comparison of three randomized response models for quantitative data. *J. Am. Stat. Assoc. 71*: 884-886.

Raghunath A, Georg D (2006) Randomized response techniques for complex survey designs. *Stat. Pap. 48*: 131-141.

Su L (2007) *Advanced Mathematical Statistics*. Peking University Press. Beijing, China. 3 pp.

Tourangeau R, Yan T (2007) Sensitive questions in surveys. *Psychol. Bull. 133*: 859-883.

Wang J, Gao G (2006) The estimation of sampling size in multi-stage sampling and its application in medical survey. *Appl. Math. Comput. 178*: 239-249.

Wang J (2003) *Practical Medical Research Methods*. People's Medical Publishing House. Beijing, China. 442-450 pp.

Wang M, Gao G (2008). Cluster Sampling and its Application on Quantitative Sensitive Questions. *Chin. J. Health Stat. 25*: 586-589.

Wang Y, Sui S, Wang A (2006) *Mathematical Statistics and Engineering Data Analysis by MATLAB*. Tsinghua University Press. Beijing, China. 9-11 pp.

Warner S L (1965) Randomized response: A survey technique for eliminating answer bias. *J. Am. Stat. Assoc. 60*: 63-69.

Yan Z, Nie Z (2005) An alternative technique of sensitive questions survey. *Or Trans. 9*: 30-34.