Comparative study on within-household sampling methods in household survey

Lv Ping*
Peking University, Beijing, China, isslvp@pku.edu.cn

Sampling survey is a primary mean to obtain the socio-economic survey data, which usually adopts stratified multi-stage unequal probability sampling design. However, in the practical survey, the people ignore the individual sampling in household. This article covers a comparative study on the ultimate sampling methods based on the data of Chinese family panel study.

Key Words: Sampling survey, within-household sampling, the Kish Grid, the Next-Birthday method

1. Introduction

Sampling survey is a primary mean to obtain the socio-economic survey data. Probability sampling is the premise of ensuring the sample representativeness. In actual process of survey, stratified multistage non-equal sampling is usually adopted to improve the efficiency and save survey funding. For instance, when conducting a nationwide sample. Sample districts are firstly selected based on valid stratification, in which sample villages are selected. Finally, valid sampled individuals can be selected. However, in practice, ultimate personal sampling frame is difficult to be achieved. Because the information should be achieved by the residents committees, it is difficult to acquire and to maintain, besides it can’t meet the concern for the floating population survey, especially with the development of society, migrant workers, empty-nest families are becoming more and more serious. Besides, the different definitions of target individual in different surveys led to the difficulties in achieving the ultimate samples. Therefore, in actual process, the households or the addresses are sampled, because this information is relatively stable and more accessible, and then the individuals can be sampled in the selected households. Thus, it is necessary to sample individuals change to sample households.

There are several methods to achieve individual samples in households. One method is to sample every individual that meets the requirement of the survey in the household, in other words, to use cluster sampling. This method ensures the equal probability of selection in sample households, individuals and households have equal probabilities to be selected. However, there are certain drawbacks of this method. First, the sample size is difficult to control, we can’t make the survey design and budget before the survey. Second, individuals from a same household may have high homogeneity, which increases the variance of sampling and decrease the efficiency of sampling. In addition, the duration of interviewing many individuals is long, which can be a burden for the interviewees, so this method is rarely adopted. In actual survey, one individual in the household is sampled. If in a surveys, the objective is to examine the economic conditions in the households, interviewing anyone who understands this information would be sufficient. Some surveys decide one individual (head of the household, for example) in the sampled household in advance, but more need to select one individual from several individuals with qualifications in the sample household. So, we must study how to sample a individual in the sampled household, this is the problem of within-household sampling, which can affect the data quality significantly. This paper mainly studies the problems of within-household sampling.

2. The Methods of ultimate sampling

As mentioned above, in actual surveys, the households are firstly sampled, and then individual is sampled in selected households. The methods of within-household sampling mainly divided into probability sampling and non-probability sampling.

Probability sampling ensures every individual has the probability to be sampled. The Kish Grid is a method that is widely used in probability sampling. The Kish Grid was proposed by Leslie Kish in 1949. It is a method using 8 tables to ensure the probability samples. In the first, we uses a simple procedure for ordering the individuals who meet the requirements in the household. First, the males are numbered in order of decreasing age, followed by the females in the same order, sometimes the individuals can be ordered by age or names.

Some people think that this order can lead to the similarity between the sample and the population in age and gender. But as Kish’s design, this order can only help interviewer to sort the individuals in selected household to sample the individual and help to check the data. The Kish Grid method guarantees the randomness of samples, which is a probability sampling. Therefore, The Kish Grid method is widely used in surveys. However, there are some drawbacks of the Kish Grid method. For example, the probability of seventh individual equals zero when the household has more than 6 individuals and the probabilities are not equal when the household has 5 individuals. In actual situation, the sex, age and names of every individual in the household should be interviewed, this thing not only time-consuming but also disturb the privacy of interviewees, which may affect interviewees’ resentment and refusal to the survey. Deming (1960) designed 12 tables, it had similar effect to the Kish Grid, yet the weakness of the Kish Grid was not overcome.

Semi-probabilistic method，Troldahl-Carter method(1964) was developed on the basis of the Kish Grid according to its actual implement problems. Salmon and Nichols(1983) developed birthday method. The interviewer asks the individual in the sampled household who has the next birthday. These methods were developed later and individuals with last, next and most recent birthdays will be interviewed respectively. This method is easy to conduct. It is adopted mostly in telephone interview. But some interviewees cannot remember all the individuals’ birthdays in the sampled household in the village.

Non-probabilistic method, which means select the sample voluntary in household or quota sampling, the people who at home for long time or willing to accept survey lead, so the sample can’t have well representativeness and we can’t estimate the population. Thus the sampling method would not be probabilistic sampling, and the population cannot be inferred effectively. However, in actual process of surveying, non-probabilistic and semi-probabilistic method can increase the response rate and cooperation rate, reduced the burdens of interviewees and curtail the expenditure. Therefore these two methods are widely adopted in surveys, especially commercial surveys. Yet, as neither of them being probabilistic sampling, in other words, the surveys not being probabilistic, these two methods are frequently challenged theoretically.

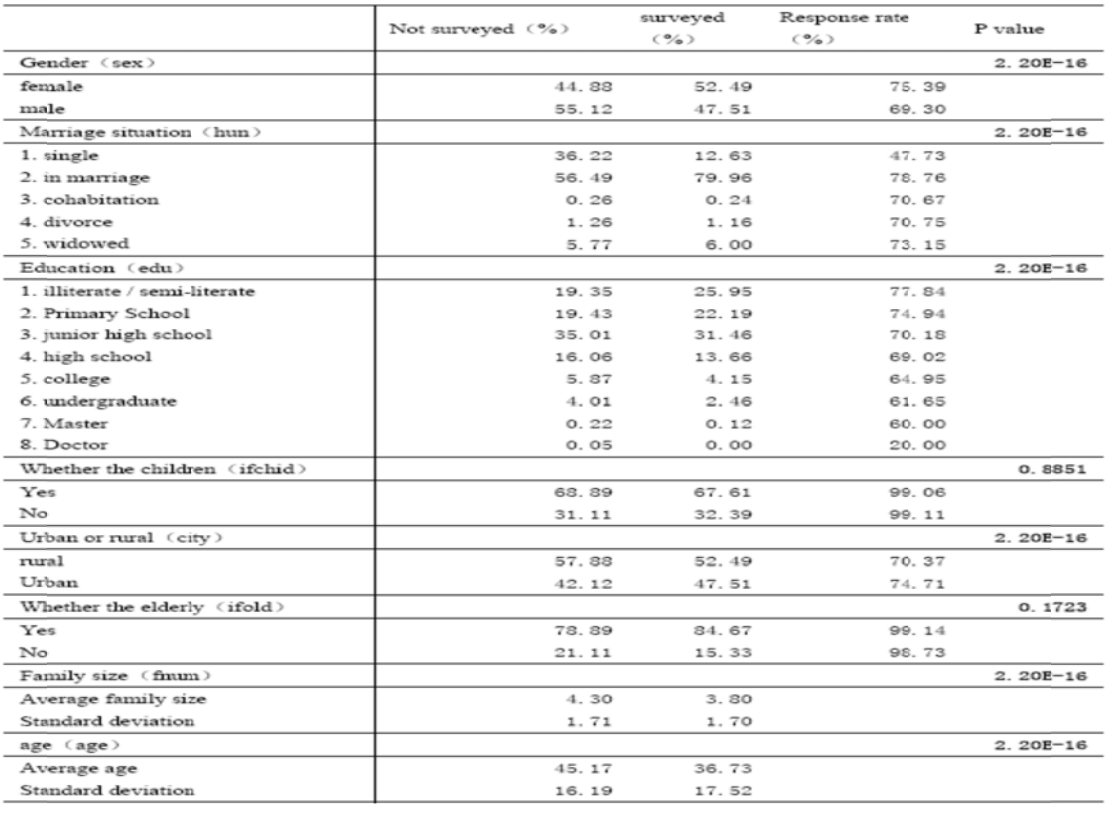
In order to obtain the probability samples in survey and used in the actual process of surveys, no-response rate should be avoided. Kish has also mentioned that the Kish Grid has its applicability at that time, but it could be changed accordingly to the change of location. This article studies the sampling method of extreme sample in sampling survey based on the follow-up statistics of CFPS conducted by institution of social science survey, Peking University, hoping to find out the applicable method to China.

3 Data Description

Chinese Family Panel Studies (CFPS) is the Long study conducted by institution of social science survey, Peking University. The survey started in the early 2010, the data of which has already been published on its website. The article conducts a stimulation analysis of the statistics in CFPS resampling.

The Chinese Family Panel Studies (CFPS) are designed to collect individual, family-, and community -level longitudinal data in China. The studies focus on the economic, as well as the non-economic, wellbeing of the Chinese population, with a wealth of information covering such topics as economic activities, education outcomes, family dynamics and relationships, migration, and health. The data will be made available worldwide for academic research and policy making. The national survey started in the early 2010, the data of which has already been published on its website. The article conducts a stimulation analysis of the statistics in CFPS resampling. The CFPS2010 sample design is a multi-stage probability proportional to size (PPS) sampling in six frames. In this paper, I use the adjusted sample. There are 9751 households who finished family members' questionnaire, in which the 9661 households finished family questionnaire, the response rate is 99%; the 30148 individuals finished adult questionnaire, the response rate is 72.3%; child questionnaire has 6816, of which 5944 completed the adult questionnaire, completion rate is 87.2%. The response rate is still a concern factors in sampling survey, the response tendency analysis based on family questionnaire data. In this data, I used median or mean imputation to deal with the item nonresponse.

Table 1 the nonresponse analysis of CFPS2010 data



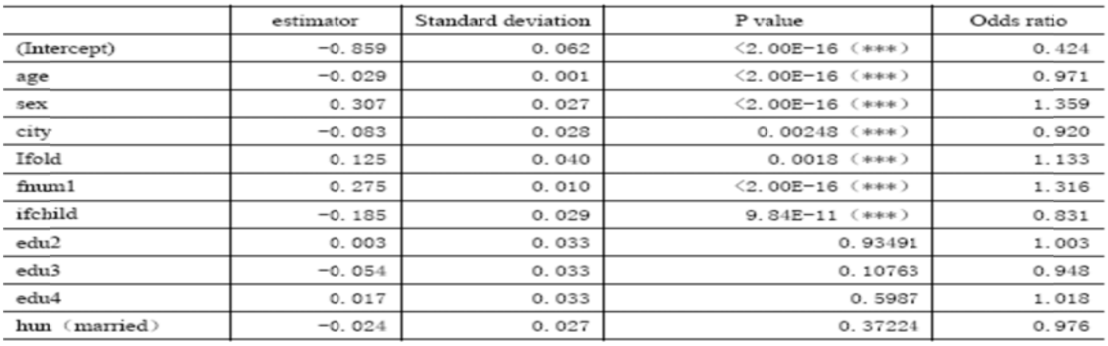
	Not surveyed（%）	surveyed（%）	Response rate（%）	P value
Gender（sex）				2.20E−16
female	44.88	52.49	75.39	
male	55.12	47.51	69.30	
Marriage situation（hun）				2.20E−16
1. single	36.22	12.63	47.73	
2. in marriage	56.49	79.96	78.76	
3. cohabitation	0.26	0.24	70.67	
4. divorce	1.26	1.16	70.75	
5. widowed	5.77	6.00	73.15	
Education（edu）				2.20E−16
1. illiterate / semi-literate	19.35	25.95	77.84	
2. Primary School	19.43	22.19	74.94	
3. junior high school	35.01	31.46	70.18	
4. high school	16.06	13.66	69.02	
5. college	5.87	4.15	64.95	
6. undergraduate	4.01	2.46	61.65	
7. Master	0.22	0.12	60.00	
8. Doctor	0.05	0.00	20.00	
Whether the children（ifchid）				0.8851
Yes	68.89	67.61	99.06	
No	31.11	32.39	99.11	
Urban or rural（city）				2.20E−16
rural	57.88	52.49	70.37	
Urban	42.12	47.51	74.71	
Whether the elderly（ifold）				0.1723
Yes	78.89	84.67	99.14	
No	21.11	15.33	98.73	
Family size（fnum）				2.20E−16
Average family size	4.30	3.80		
Standard deviation	1.71	1.70		
age（age）				2.20E−16
Average age	45.17	36.73		
Standard deviation	16.19	17.52		

Through the above table 1, the age, gender, family size, whether the elderly in household, whether the children in household variable are influence factors in nonresponse. In this paper, we use logic model to analysis nonresponse tendency. In which, P is people whether accepted the survey or not, that is, P=1 is the person who didn't accepted the visit, P=1 is the person who accepted the survey. So, the model is

$$\begin{aligned}\log_e P &= \beta_0 + \beta_1\text{age} + \beta_2\text{sex} + \beta_3\text{city} + \beta_4 ifold + \beta_5\text{ifchid}\\ &\quad + \beta_6\, fnum1 + \beta_7\text{ifchid} + \beta_8 edu2 + \beta_9 edu2 + \beta_{10} edu2 + \beta_{11} hun\end{aligned}$$

The result is,

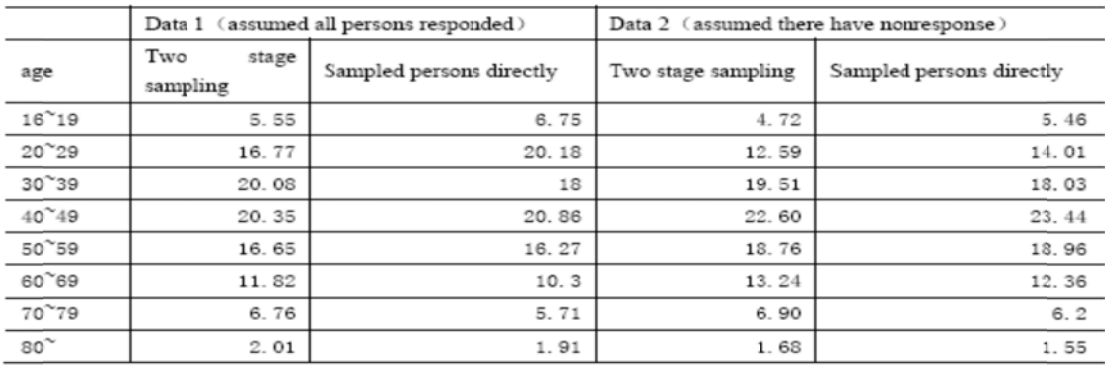
Table 2 the result of the model



	estimator	Standard deviation	P value	Odds ratio
(Intercept)	−0. 859	0. 062	<2. 00E−16 （***）	0. 424
age	−0. 029	0. 001	<2. 00E−16 （***）	0. 971
sex	0. 307	0. 027	<2. 00E−16 （***）	1. 359
city	−0. 083	0. 028	0. 00248 （***）	0. 920
Ifold	0. 125	0. 040	0. 0018 （***）	1. 133
fnum1	0. 275	0. 010	<2. 00E−16 （***）	1. 316
ifchild	−0. 185	0. 029	9. 84E−11 （***）	0. 831
edu2	0. 003	0. 033	0. 93491	1. 003
edu3	−0. 054	0. 033	0. 10763	0. 948
edu4	0. 017	0. 033	0. 5987	1. 018
hun （married）	−0. 024	0. 027	0. 37224	0. 976

Through the above the table 2, the age, gender, urban or rural, family size, whether the elderly in household, whether the children in household variable are influence factors in nonresponse. In the social and economic survey, age, gender, the general are important factors affecting the social investigation target variables. In this paper, we compared the proportion in different age in full data and responded data. In every situation, we think about two methods, one method is two stage sampling, we sampled households firstly, and then sampled person; another method, sampled person directly. The proportion is as follows:

Table 3 the age data in CFPS2010



	Data 1 （assumed all persons responded）		Data 2 （assumed there have nonresponse）	
age	Two stage sampling	Sampled persons directly	Two stage sampling	Sampled persons directly
16˜19	5. 55	6. 75	4. 72	5. 46
20˜29	16. 77	20. 18	12. 59	14. 01
30˜39	20. 08	18	19. 51	18. 03
40˜49	20. 35	20. 86	22. 60	23. 44
50˜59	16. 65	16. 27	18. 76	18. 96
60˜69	11. 82	10. 3	13. 24	12. 36
70˜79	6. 76	5. 71	6. 90	6. 2
80˜	2. 01	1. 91	1. 68	1. 55

From the table above, the low proportion of people aged 16~39 using first method, but and more than 40 people high proportion in full data. Assuming there have nonresponse samples, the proportion of people aged 16~39 continue to decrease, while the proportion of people over the age of 40~70 further increased, the proportion of people aged more than 70 decreased. Maybe, the youth people people go out to work and go to school lead to the low proportion, and the population over the age of 70 cannot answer the questionnaire because of health reasons.

4 Simulation analysis

In this paper, I use the data for further simulation study. There are 419 villages, we assume N is the total population, we sampled n_1 villages using systematic probability proportional to village's population size, there are H_i households, N_i persons in sampled village, and we sampled n_2 households , in which N_{ij} persons in the households. In this person, we sampled 320 villages, and then sampled 7 households in selected villages. There are several methods in within-household sampling. In this paper, we use six methods:

（1） Sampled persons in the persons frame of 320 villages.

（2） Sampled all persons in sampled household.

（3） Sampled one person using Kish grid sampling method.

（4） Sampled one person using random sampling method in the selected household in which the persons ordered by decreasing age.

（5） Sampled one person using random sampling method in the selected household in which the persons ordered by sample id.

（6） Sampled one person using random sampling method in the selected household in which the persons ordered by sex and age like Kish grid method, but the age in increasing order.

In this paper, we use simulation method to repeat 100 times, we using average bias and MSE indicator to evaluate the above methods, that is

$$Bias_j = \sum_{r=1}^{k} (\hat{\bar{Y}}_{jr} - \bar{Y}_{jr}) / k \text{，} \quad RMSE_j = \sqrt{\sum_{r=1}^{k} (\hat{\bar{Y}}_{jr} - \bar{Y}_{jr})^2 / k} \text{，} \quad (k \text{ is repeat times})$$

Firstly, we assumed the sampling weights are W_1, W_2, W_4, W_3, W_3, W_4 :

$$w_1 = \frac{N}{n_1 N_i} \times \frac{N_i}{n_2} = \frac{N}{n_1 n_2}$$

$$w_2 = \frac{N}{n_1 N_i} \times \frac{H_i}{n_2} = \frac{N}{n_1 n_2 N_{ij}}$$

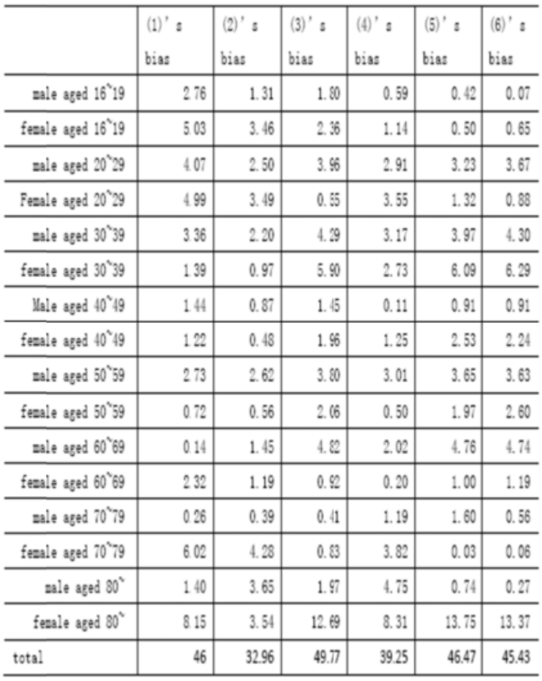
$$w_3 = \frac{N}{n_1 N_i} \times \frac{H_i}{n_2} \times \frac{N_{ij}}{1} = \frac{N H_i N_{ij}}{n_1 n_2 N_i}$$

$$w_4 = \frac{N}{n_1 N_i} \times \frac{H_i}{n_2} \times \frac{N_{ij}{}'}{1} = \frac{N H_i N_{ij}{}'}{n_1 n_2 N_i}$$

So, the 1,4,5 are equal probability sampling methods, the 3,6 are approximate equal probability methods, but the 2 is a non-equal probability sampling methods. In this paper, we can’t think out weight.

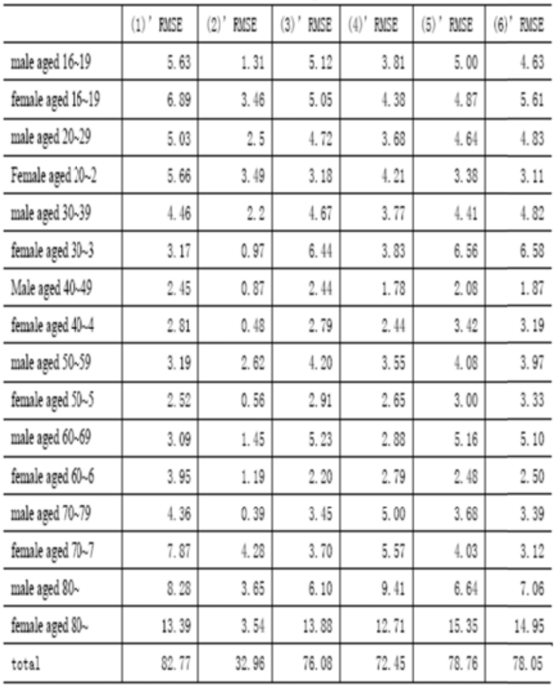
The simulation results are shown in the following table after 100 times by the simulation method, the bias and RMSE of response rate in data1 are:

Table 4 the bias of response rate in data1



	(1)' s bias	(2)' s bias	(3)' s bias	(4)' s bias	(5)' s bias	(6)' s bias
male aged 16~19	2.76	1.31	1.80	0.59	0.42	0.07
female aged 16~19	5.03	3.46	2.36	1.14	0.50	0.65
male aged 20~29	4.07	2.50	3.96	2.91	3.23	3.67
Female aged 20~29	4.99	3.49	0.55	3.55	1.32	0.88
male aged 30~39	3.36	2.20	4.29	3.17	3.97	4.30
female aged 30~39	1.39	0.97	5.90	2.73	6.09	6.29
Male aged 40~49	1.44	0.87	1.45	0.11	0.91	0.91
female aged 40~49	1.22	0.48	1.96	1.25	2.53	2.24
male aged 50~59	2.73	2.62	3.80	3.01	3.65	3.63
female aged 50~59	0.72	0.56	2.06	0.50	1.97	2.60
male aged 60~69	0.14	1.45	4.82	2.02	4.76	4.74
female aged 60~69	2.32	1.19	0.92	0.20	1.00	1.19
male aged 70~79	0.26	0.39	0.41	1.19	1.60	0.56
female aged 70~79	6.02	4.28	0.83	3.82	0.03	0.06
male aged 80~	1.40	3.65	1.97	4.75	0.74	0.27
female aged 80~	8.15	3.54	12.69	8.31	13.75	13.37
total	46	32.96	49.77	39.25	46.47	45.43

Table 5 the RMSE of response rate



	(1)' RMSE	(2)' RMSE	(3)' RMSE	(4)' RMSE	(5)' RMSE	(6)' RMSE
male aged 16-19	5.63	1.31	5.12	3.81	5.00	4.63
female aged 16~19	6.89	3.46	5.05	4.38	4.87	5.61
male aged 20-29	5.03	2.5	4.72	3.68	4.64	4.83
Female aged 20~2	5.66	3.49	3.18	4.21	3.38	3.11
male aged 30-39	4.46	2.2	4.67	3.77	4.41	4.82
female aged 30~3	3.17	0.97	6.44	3.83	6.56	6.58
Male aged 40-49	2.45	0.87	2.44	1.78	2.08	1.87
female aged 40~4	2.81	0.48	2.79	2.44	3.42	3.19
male aged 50-59	3.19	2.62	4.20	3.55	4.08	3.97
female aged 50~5	2.52	0.56	2.91	2.65	3.00	3.33
male aged 60-69	3.09	1.45	5.23	2.88	5.16	5.10
female aged 60~6	3.95	1.19	2.20	2.79	2.48	2.50
male aged 70-79	4.36	0.39	3.45	5.00	3.68	3.39
female aged 70~7	7.87	4.28	3.70	5.57	4.03	3.12
male aged 80~	8.28	3.65	6.10	9.41	6.64	7.06
female aged 80~	13.39	3.54	13.88	12.71	15.35	14.95
total	82.77	32.96	76.08	72.45	78.76	78.05

From the table 4， the result of method 2 is best, another is method 1. The method 3 is worst. From the table 5， the result of method 2 is best, another is method 4. The method 1 is worst. As the same way, in the table of gender-age in full data, the bias of method 2 is best, another is method 4. The method 1 is worst; the RMSE of method 2 is best, another is method 1, the third is method 4 the method 5 is worst. The bias of gender-age in responded data， the bias of method 2 is best, another is method 1, the method 3 is worst; the RMSE of method 2 is best, another is method 1. The method 5 is worst.

5 Conclusion

This paper covers a comparative study on the ultimate sampling methods based on the 2010 data of Chinese family panel study. It explained fatherly, sampled all persons in the selected household have better representation, but this method need large sample size, so we can't control sample size and the data have correlation within household, so in practical the people use this method less. The method 4, that is, sampled one person using random sampling method in the selected household in which the persons ordered by decreasing age, this method has better result. The Kish grid method can be improved if the persons in the household according to age in ascending order. In this paper, I didn't consider the influence of sampling weights and non-probabilistic sampling methods, these will be the next research contents.

References

Binson, D. , J. A. Canchola& J. A. Catania. (2000), Random Select ion in a Telephone Survey: AComparison of the Kish, Next birthday, and Last birthday Methods. Journal of Official Statistics 16.

Bryant , B. E. (1975), Respondent Selection in a Time of Changing Household Composition. Journal ofMarketing Research 12.

Deming,W. E. (1960), Sample Design in Business Research . New York: John Wiley and Sons, IncHagan,D. E. & C. M. Collier (1982), Must Respondent Selection Procedures for Telephone Surveys BeInvasive Public Opinion Quarterly 47.

Kish, Leslie (1949), A Procedure for Objective Respondent Selection within the Household. Journal oftheAmerican Statistical Association 44.

Lavrakas, P. J. (1993), Telephone Survey Methods: Sampling, Selection and Supervision. AppliedSocial Research Methods Series 7.

Troldahl, V. C & R. E Carter, Jr. (1964), Random Selection of Respondents Within Households in PhoneSurvey. Journal of Marketing Research .

作者简介：

Lv Ping, institute of Social Science Survey, Peking University, Beijing, China, Email: issslvp@pku.edu.cn.