Sampling Design for the Primary Farm Household Survey of the Taiwanese Agriculture Study

Chang-Tai Chao*[1], Chien-Min Huang†[1], Chiu-Yen Lee‡[2], Shiow-Ing Lin§[2] and Yu-wen Liu¶[2]

[1]Department of Statistics, National Cheng Kung University, Taiwan
[2]Council of Agriculture, Executive Yuan, Taiwan

Abstract
The census of Agriculture, Forestry, Fishery and Animal Husbandry is an important official survey in Taiwan. However, enormous survey cost and effort for the data processing are required for such a nationwide census, hence this census can only be conducted every five years. Therefore, certain annual agriculture survey is necessary for the realization of the current related industry information, so that proper policy can be formulated timely. A sampling strategy for the Taiwanese primary farm household survey is constructed in this research. A stratified random sampling design, in which the optimal stratum boundary and allocation are carried out based on the 2010 census data, is proposed for the purpose to enhance the estimation precision and investigate certain subpopulations of interest. The result indicates that the performance of the proposed stratified sampling is much more advantageous than simple random sampling without replacement and other stratified design with optimal allocation and equal within-stratum size stratification boundary under a comparable total sample size.

Key Words: Agriculture survey, Primary farm household, Sampling strategy, Stratified sampling, Stratum boundary, Optimal allocation

1. Introduction

The census of Agriculture, Forestry, Fishery and Animal Husbandry is one of the important official surveys conducted regularly in Taiwan. Implemented every five years, the main purpose of the survey is to provide the government with information about the production structure, characteristic of labor force and operating behavior. Hence, the related authorities can make use of the census data to revise production structure so that a proper policy can be evaluated accordingly and then properly revised. However, taking a nationwide census requires a great deal of survey cost. Therefore, sampling survey, which can provide

*National Cheng Kung University
†National Cheng Kung University
‡Council of Agriculture Executive Yuan
§Council of Agriculture Executive Yuan
¶Council of Agriculture Executive Yuan

population information at a much lower cost than a census is often used. In this article, we focus on how to choose for a survey to study the primary agriculture households, including both of the sampling design and the associated inference based on the census data at the year of 2010. In addition, the sampling strategy to be proposed is required to provide statistical inference which can meet certain precision level, and also feasible in practice.

To establish a proper sampling strategy which can provide a representative information, the property of the population variable of primary interest has to be studied first. Based on the census data of 2010, total population size of agriculture and animal husbandry in Taiwan area is 781,518 and the agricultural gross income, which is referred as gross income below, is chosen to be the population variable of primary interest. Since some of the farm households are not active, such the households with small agriculture scale and/or all the household agriculture workers are older than 65 years old. Therefore, in order to comprehend the households which are able to dedicate in the further development of agriculture industry, we define the target population to be the farm households whose annual gross income are between 200,000 and 50 millions New Taiwan dollars (NTD) and at least one household member under the age of 65 is currently engaged in the agriculture work. Such a target population is referred as “primary farm households”. The basic population characteristic of primary farm households is shown in Table 1.1. The average annual gross income of the primary farm households is 846 thousand dollars and 95% of the incomes are between 200 and 4,500 thousand dollars. The distributions of target population tends to be highly right skewed.

Table 1.1: Population characteristic of primary farm households (thousand NTD)

Number of households	150,456
mean	846.01
standard deviation	2120.54
median	400
95% interval	(200,4500)

A proper sampling strategy should provide samples with the similar structure of the target population so the population can be better represented. The well-known stratified sampling is able to ensure a representative sample if a proper stratified variable is used. Therefore, the scale of a agriculture household, which is believed to have a great impact on the population variable of interest, is used as one of the stratified variables. In addition, it is also of interest to study the subpopulations of different types of main production, hence the main production type is used as another stratified variable in stratified sampling. Due to the highly skewed property of the target population, stratification that follows the principle of stratification (Thompson 2012) results in large improvements to the variance of the estimation of population mean or total. That is, the precision of the estimation can be enhanced when the strata are comprised of units which are as similar to each other as possible. Such stratification can be achieved by finding proper stratum boundary. Several stratification algorithms for the optimal set of stratification boundaries are reviewed in Section 2, and how to apply a proper stratification algorithm is also described. For a better estimation result, the optimal allocation based on within-stratum variance and within-stratum size is also given in Section 2.

In Section 3, the estimation precision of the proposed design is calculated based on the 2010 census data, and compared with other sampling designs in terms of the relative estimation error and relative efficiency. Finally, conclusion about the sampling design and related discussion are given in Section 4.

2. Stratified Sampling

2.1 Stratification Boundaries

There are three main issues that a stratification design need to address, the number of strata, the placement of stratum boundaries and the allocation of sample among the strata. In order to decrease the estimation variance, stratification should follow the principal of stratification (Thompson 2012), that is, population should be stratified such that the units in the same stratum are as similar as possible. To achieve this objective, we need to find a proper placement of stratum boundaries and some notable contributions to this problem have been discussed before. Given the number of strata, equations for determining the best stratum boundaries under proportional and Neyman allocation have been worked out by Dalenius (1957). However, the equations have considerable dependencies among the components. Therefore, a number of approximate methods have been devised, like the first approximation suggested by Dalenius and Hodges (1959) constructed the strata by taking equal intervals on the cumulative function of the square root of the frequencies. Besides, Lavallée & Hidiroglou (1988) proposed an iterative procedure to find the optimal stratum boundaries and the method is particularly suitable for highly skewed population. The application of the Lavallée & Hidiroglou algorithm was found slow to or even did not coverage. Consequently, Kozak (2004) suggested an alternative random search algorithm based on the work of Lednicki and Wieczorkowski (2003) for more efficient way to find stratum boundaries. Moreover, Gunning & Horgan (2004) also suggested using a geometric progression algorithm for the construction of stratum boundaries in positively skewed populations.

Both Lavallée & Hidiroglou algorithm and geometric progression algorithm seem reasonable of the primary agriculture households in Taiwan due to the skewed property. However, stratum boundaries just based on the primary variable of interest may not have practical implication. Also, one of the purposes of the stratified sampling is to investigate the subgroups of interest; hence, auxiliary variable instead of variable of interest is often used as a stratification variable. Rivest (2002) constructed a generalization of the Lavallée & Hidiroglou algorithm for the discrepancy between the stratification variable and survey variable in terms of statistical models. Therefore, we employ Lavallée & Hidiroglou (1988) algorithm to find the proper stratification boundaries to decrease the estimation variance.

The gross income depends on the type of agriculture type at certain level; besides, the gross income of the subpopulations defined by different types of production are also one of the main research interest of this survey. Therefore, we divide the primary farm household into eight subpopulations by the production types of rice, vegetables, fruits, coarse grain and special crops, other crops, hog farms, chicken farms and other livestock farms. The first five productions are crop farms and the other are livestock farms. In each production strata, instead of using the survey variable as the stratification variable, we should attempt to make use of other auxiliary variable as stratification variable in order to observe the different scale in the main production. For example, cultivated land area is a criterion for different scale of crop farms. On the other hand, the heads on a farm can represent the scale of livestock farms, nevertheless, the census data includes the heads on farms at the end of year only, which may mislead the scale. Therefore, we consider the yearly sale to define the scale of a livestock farm, and this variable is available in the census data in terms of the gross income itself. As a result, we utilize the cultivated land area as a stratification variable for crop farms and gross income as a stratification variable for livestock farms. For livestock farms, we use thousand unit of gross income as a stratification variable and implement Lavallée & Hidiroglou algorithm and random search method of Kozak to stratify each main industry into 3 strata. And for crop farms, because the stratification variable is not equal to the survey variable, we refer to the statistical model suggested by Rivest (2002) and construct a log-linear model between gross income and cultivated land area first, and then make use of the Lavallée & Hidiroglou and random search method to stratify each main industry into 3 strata. The resulting strata associated with the numbers of agriculture households in each stratum are concluded in Table 2.1.

The stratification rule shown in Table 2.1 satisfies the general requirement of the stratification, but the stratum boundaries obtained from the algorithm do not possess the functional value, that is, the value of the boundaries in each main industry are meaningless in practice. Hence, we try different stratum boundaries which are multiples of 0.5 or 5 based on the boundaries presented in Table 2.1. By considering the practical meaning and the precision of estimate, we obtain the stratification rule which is to be put in practice for the primary farm households. Result is shown in Table 2.2.

2.2 Optimal Allocation

After forming the stratification boundaries, how to allocate the total sample size into strata is another issue in stratified sampling. To improve the precision of estimation, we use the

Table 2.1: Stratum boundaries associated with the numbers of agriculture households in stratum of the primary farm households I

Crop Farms	**Boundary (land area)**			
		Unit		**Subtotal**
Rice	below 1.77 ha	1.77–5.31 ha	above 5.31 ha	
	17939	8193	1142	27274
Vegetables	below 0.8 ha	0.84–2.67 ha	above 2.67 ha	
	16122	13737	2321	32180
Fruits	below 0.86 ha	0.86–2.27 ha	above 2.27 ha	
	23848	27281	8889	60018
Coarse Grain and	below 0.96 ha	0.96–2.62 ha	above 2.62 ha	
Special Crops	5945	6162	1494	13601
Other Crops	below 0.08 ha	0.08–0.8 ha	above 0.8 ha	
	251	4163	2447	6861
Livestock Farms	**Boundary (gross income in thousand NTD)**			
		Unit		**Subtotal**
Hog Farms	below 3550	3550–11750	above 11750	
	2783	1352	434	4569
Chicken Farms	below 4250	4250–14300	above 14300	
	1903	850	289	3042
Other Livestock Farms	below 2650	2650–11750	above 11750	
	2148	540	223	2911
Population total				150456

Table 2.2: Stratum boundaries associated with the numbers of agriculture households in stratum of the primary farm households II

Crop Farms	**Boundary (land area)**			
		Unit		**Subtotal**
Rice	below 1.75 ha	1.75–5 ha	above 5 ha	
	17722	8194	1358	27274
Vegetables	below 1 ha	1–2.5 ha	above 2.5 ha	
	18812	10674	2694	32180
Fruits	below 1 ha	1–2.5 ha	above 2.5 ha	
	27930	24377	7711	60018
Coarse Grain and	below 1 ha	1–2.5 ha	above 2.5 ha	
Special Crops	6206	5672	1723	13601
Other Crops	below 0.5 ha	0.5–1 ha	above 1 ha	
	2754	2160	1947	6861
Livestock Farms	**Boundary (gross income in thousand NTD)**			
		Unit		**Subtotal**
Hog Farms	below 5000	5000–10000	above 10000	
	3215	809	545	4569
Chicken Farms	below 5000	5000–15000	above 15000	
	2012	759	271	3042
Other Livestock Farms	below 2000	2000–10000	above 10000	
	1988	647	276	2911
Population total				150456

optimal allocation described by Neyman, which is concerned with the minimization of the variance of estimator, to allocate fixed total sample size into stratum. The optimal result can be derived by Lagrange multiplier based on the within-stratum size and within-stratum variance and the optimal allocation in each stratum h for a fixed sample size n is

$$n_h = \frac{nN_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h} \tag{1}$$

where N_h is the unit in each stratum and σ_h is the population standard deviation in the stratum. In the practice, overall sample size for the primary farm households survey is 1000. Within-stratum standard deviation associated with the stratification rule in Table 2.2 and the optimal sample allocation for the primary farm households is demonstrated in Table 2.3 and 2.4. Table 2.4 indicates that more samples are allocated into the stratum with larger standard deviation and/or within-stratum size.

Table 2.3: Stratum standard deviation of the primary farm households (thousand NTD)

Crop Farms	**Boundary (land area)**		
	Standard deviation		
Rice	below 1.75 ha	1.75–5 ha	above 5 ha
	113.29	253.58	1354.23
Vegetables	below 1 ha	1–2.5 ha	above 2.5 ha
	196.23	310.61	1936.61
Fruits	below 1 ha	1–2.5 ha	above 2.5 ha
	199.32	370.83	1256.28
Coarse Grain and	below 1 ha	1–2.5 ha	above 2.5 ha
Special Crops	372.32	622.64	1819.46
Other Crops	below 0.5 ha	0.5–1 ha	above 1 ha
	2217.91	1673.22	2870.30
Livestock Farms	**Boundary (gross income in thousand NTD)**		
	Standard deviation		
Hog Farms	below 5000	5000–10000	above 10000
	1283.52	1379.57	8172.53
Chicken Farms	below 5000	5000–15000	above 15000
	1264.46	2662.21	8030.82
Other Livestock Farms	below 2000	2000–10000	above 10000
	464.03	2105.83	8082.54

Table 2.4: Optimal sample allocation of the primary farm households

Crop Farms	**Boundary (land area)**			
	Sample size			**Subtotal**
Rice	below 1.75 ha	1.75–5 ha	above 5 ha	
	23	24	21	68
Vegetables	below 1 ha	1–2.5 ha	above 2.5 ha	
	42	38	59	139
Fruits	below 1 ha	1–2.5 ha	above 2.5 ha	
	63	103	110	276
Coarse Grain and Special Crops	below 1 ha	1–2.5 ha	above 2.5 ha	
	26	40	36	102
Other Crops	below 0.5 ha	0.5–1 ha	above 1 ha	
	70	41	64	175
Livestock Farms	**Boundary (gross income in thousand NTD)**			
	Sample size			**Subtotal**
Hog Farms	below 5000	5000–10000	above 10000	
	47	13	51	111
Chicken Farms	below 5000	5000–15000	above 15000	
	29	23	25	77
Other Livestock Farms	below 2000	2000–10000	above 10000	
	11	16	25	52
Total sample size				1000

3. Precision of Estimation

To evaluate the performance of the sampling design for the average gross income of the target population, we calculate the related estimation precision by the proposed stratified design with the simple random sampling without replacement (SRSWOR) as the within-stratum design. Under the stratification rule and sample allocation given in Table 2.4 for the primary farm households, the overall theoretical maximum absolute estimation error under 95% confidence level of the average gross income for the primary farm households is 35.72 thousand dollars and the maximum relative estimation error under 95% confidence level is 4.22% based on the census data of 2010. Maximum relative estimation error under 95% confidence level is calculated as (maximum estimation error / population mean) × 100%, where the maximum estimation error is calculated based on the finite population central limit theorem. Also, the overall absolute average error of simulated 5,000 samples is 14.87 thousand dollars and the relative average error is 1.75%.

From the viewpoint of the design-based sampling design, no population model is assumed. Hence, due to the skewed property of the target population and small sample sizes in some strata, which may influence the maximum estimation error under 95% confidence level calculated based on finite population central limit theorem, we also simulate 5,000 samples to check the probability that the estimation error of the simulated sample greater than the theoretical maximum estimation error. The probability that the estimation error for the primary farm households average gross income greater than the theoretical maximum error is 0.052, which follows the the theoretical value 0.05. Besides the error, the coverage probability of 95% confidence interval of simulated 5,000 samples is 0.942, which is close to 95%.

3.1 Comparison between Different Sampling Design

In order to show that the proposed stratified sampling is an appropriate sampling design, estimation error of simple random sampling without replacement (SRSWOR) and stratified sampling without employing stratification boundary algorithm but the optimal allocation, which is referred as stratified sampling II below, are calculated in comparison with the proposed stratified sampling, which is referred as stratified sampling I below, under the same total sample size. In stratified sampling II, we also divided the target population into eight productions and sorted the population by stratification variable and divided the production into three strata by taking equal within-stratum size in each stratum. Table 3.1 summarizes the comparison of estimation error between SRSWOR, stratified sampling I and stratified sampling II. The table indicates that stratified sampling I has the lowest estimation error and SRSWOR has the highest.

Table 3.1: Estimation error of different sampling designs for the primary farm households

	SRSWOR	Stratified sampling I	Stratified sampling II
Maximum absolute estimation error under 95% confidence level	130.99	35.72	43.34
Simulated absolute average error of 5,000 samples	53.84	14.87	17.60
Maximum relative estimation error under 95% confidence level	15.48%	4.22%	5.12%
Simulated relative average error of 5,000 samples	6.36%	1.75%	2.08%
Pr (sampling error $>$ theoretical maximum estimation error)	0.049	0.052	0.049

3.2 Simulation Study

Besides estimation error, we also make use of Relative Efficiency, which is often used to compare the performances of two estimators to evaluate the performances of three estimators. The definition of Relative Efficiency of estimator E_1 to E_2 is

$$\text{RE} = \frac{\text{MSE}(E_2)}{\text{MSE}(E_1)}$$

If RE is greater than 1, then E_1 is more favorable than E_2 assuming equal sample size for two estimators.

We simulate MSE of three sampling designs for different sample sizes and calculate the relative efficiency of stratified sampling to SRSWOR. The result is shown in Figure 3.1. It is clear that stratified sampling I and II is significantly better than SRSWOR. And for the two stratified sampling, the result is shown in Figure 3.2. From Figure 3.2, the performance of stratified sampling I is superior to stratified sampling II since the relative efficiency of stratified sampling I to II is always greater than 1 under different total sample sizes.

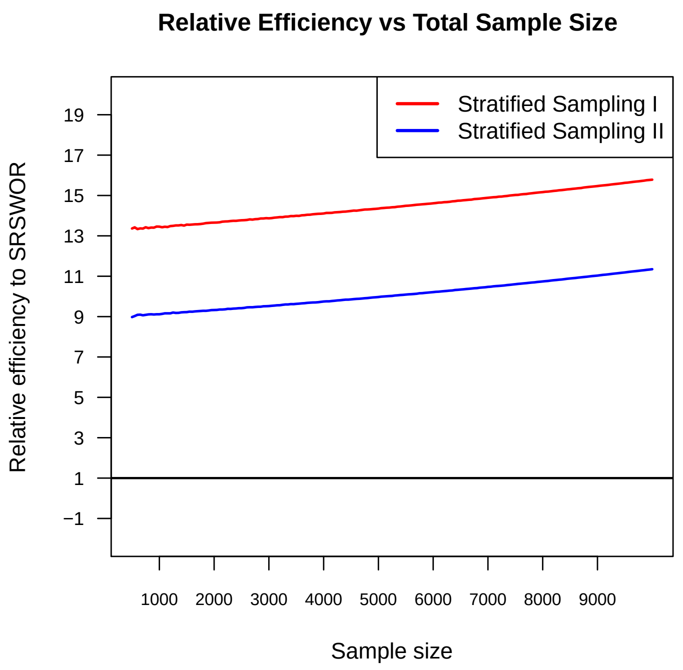

Figure 3.1: Relative Efficiency of stratified sampling to SRSWOR under different sample size

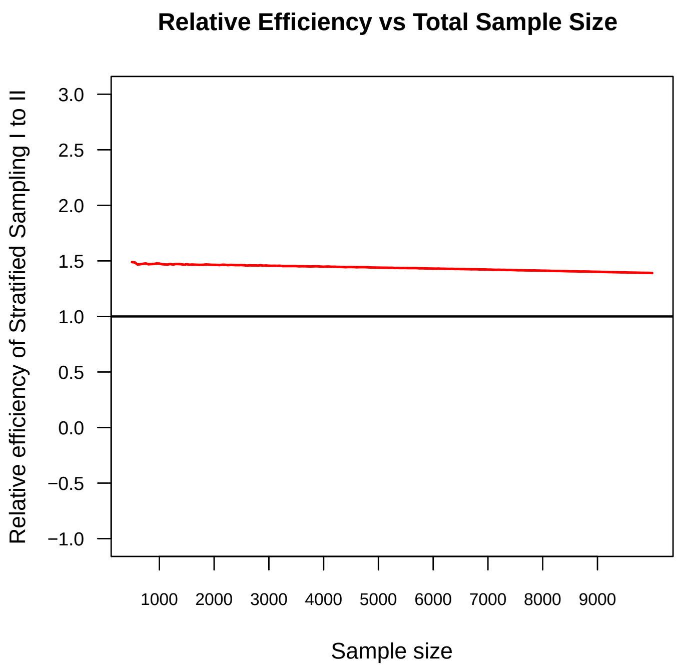

Figure 3.2: Relative Efficiency of stratified sampling I to stratified sampling II under different total sample size

4. Final Comments

A stratified sampling design, in which the stratification algorithm for the optimal stratification boundary and the optimal allocation of the within-strata sample sizes are utilized, for the primary agriculture household survey is described in this article. A set of stratification boundaries with more practical meaning are determined based on the optimal ones, hence the subpopulations of interest can be studied without further post-stratification. The simulation result indicates that the proposed design can be much better than the SRSWOR as expected. In addition, it is better than the stratified design in which the boundaries are determined by equal within-stratum size. Hence the advantage of the stratification algorithm is also illustrated in this research.

REFERENCES

Dalenius, T. (1957), *Sampling in Sweden: Contributions to the methods and theories of sample survey practice*, Almqvist & Wicksell, Stockholm, Sweden

Dalenius, T. and Hodges, J. (1959), "Minimum variance stratification," *Journal of American Statistical Association*, Vol.54, No.285, pp.88-101.

Gunning, P. and Horgan, J. M. (2004), "A new algorithm for the construction of stratum boundaries in skewed populations," *Survey Methodology*, Vol.30, pp.159-166.

Kozak, M. (2004), "Optimal stratification using random search method in agricultural surveys," *Statistics in Transitions*, Vol.6, pp.797-806.

Lavallée, P. and Hidiroglou, M. (1988), "On the stratification of skewed population," *Survey Methodology*, Vol.14, pp.33-43.

Lednicki, B. and Wieczorkowski, R. (2003), "Optimal stratification and sample allocation between subpopulations and strata," *Statistics in Transitions*, Vol.6, pp.287-306.

Rivest, L.-P. (2002), "A generalization of the Lavallée and Hidiroglou Algorithm for stratification in business surveys," *Survey Methodology*, Vol.28, pp.191-198.

Thompson, S.K. (2012), *Sampling*, 3rd edition, John Wiley & Sons, Hoboken, NJ.