TEACHING ESTIMATION ABOUT THE FAMILY OF SRS, PPS AND MPPS SAMPLING DESIGNS

Zhang Yong, Xie Qingquan and Niu Junjun
National Bureau of Statisitcs, China
niujj@stats.gov.cn

This paper introduces the MPPS sampling design developed by the National Agricultural Statistics Service of Agricultural Department in the U. S. A. to solve the problems of multi-purpose survey. Then it gives a uniform estimator formula that is suitable to the family of SRS, PPS sampling and MPPS sampling designs under some conditions. The application of the MPPS sample design and how to give this course to the students are discussed also.

INTRODUCTION

There are a lot of discussions concerning simple random sampling (SRS) and probability proportional to size (PPS) sampling in statistics textbooks. Cochran (1977), Wolter (1985) and Feng Shiyong (1998) give details. There are few discussions about the estimators for more complex sampling designs in statistics teaching and learning. The National Agricultural Statistics Service, Department of Agriculture, USA (NASS) put forward multivariate probability proportional to size (MPPS) sampling at the end of the 20th century for multi-purpose surveys, (see the Food and Agriculture Organization of United Nations (FAO), 1998). The Chinese version was published in 2000. The systematic sampling method is usually applied in practice for these three kinds of sampling designs. Under some conditions, a uniform estimator can be used to express the estimators of these three kinds of sampling designs.

China is a large country in both area and population, so the sample survey takes a very important role in management and administration to get data. There are statistics departments that recruit students to specialise in statistics in more than one hundred colleges and universities. These students study sample survey design and some of them will be employed as statisticians to design sample surveys after they graduate. Up to now they have studied SRS and PPS sampling designs in sample survey courses in schools. There are some demands for MPPS sampling design in practice, so these students have to study MPPS sampling design by themselves after they graduate from schools, specifically for the graduate students. We find that there is a good and easy way to give a course about MPPS design for students to understand. We would like to share our experience with colleagues and hope more teachers will know the uniform estimator and give this course to students of statistics specialty.

DEVELOPMENT OF THE MPPS SAMPLING DESIGN

We first introduce MPPS sample design briefly. Each sampling unit owns only one selected probability P_i that is determined by current auxiliary data and expected sample size of characteristics for MPPS sampling in a population.

$$P_i = \min\left[1, \max\left[n_1 \frac{x_{1,i}^{3/4}}{\sum_{i=1}^{N} x_{1,i}^{3/4}}, \ldots, n_K \frac{x_{K,i}^{3/4}}{\sum_{i=1}^{N} x_{K,i}^{3/4}}\right]\right], i = 1, 2, \cdots, N$$

Here P_i is the probability of the selected unit i; $x_{k,i}$ is the value of characteristic k for the unit i; n_k is the expected sample size of characteristic k (the number of units with characteristic k in the sample, $k = 1,2,\cdots,K$); K is the number of interested characteristics; N is the total number of units in the population. The power of 3/4 is applied to calculate auxiliary data, since the auxiliary data are generally collected a year before, and we do not want P_i to be too large for the upper level unit. We certainly can also choose another power. It will make the number of selected units smaller when $P_i \le 1$.

In theory, sampling with replacement has good statistical characteristics and simpler formula for the estimator, so it is usually used to deduce the estimator. In practice, we can choose a sample using a systematic PPS sampling design. Now $\{P_i\}$ is a set of selected probabilities, $i = 1,2,\cdots,N$. Let *n* be the integer part of $\sum_{i=1}^{N} P_i$, so *n* is the fixed sample size. Let r be a random number in the interval (0, 1), then when $\sum_{j=1}^{i_k-1} p_j < r+l$, $\sum_{j=1}^{i_k} p_j \geq r+l, l = 0,1,\cdots,n-1$, the units $i_1, i_2, \cdots, i_n$ will be selected as sample units from the population. Because $P_i \leq 1$, the sampling is strictly without replacement, no unit can be chosen more than one time. If we allow a unit to be chosen more than one time, there is no need to put $P_i \leq 1$.

The permanent random number method (PRN), also called Poisson sampling, can also be used to select the sample. There are a lot of advantages if Poisson sampling is used from FAO (1998), although we find some problems need to be further researched if the Poisson sampling method is applied so we do not discuss it here.

One of the special features of PPS sampling for one variable is that it has a simple estimator that is more easily understood. In MPPS sampling, the weight $w_i = 1/p_i$ comes from many variables, so that the sum of $w_i = 1/p_i$ is meaningless to estimate anything. Therefore, the ratio estimator should be used to adjust the weight. Let $\hat{Y}_{k,i}$ be the estimator of characteristic *k* for unit i. In the formula below, $x_{k,i}$ is the auxiliary data for characteristic *k* that is used to calculate the selected probability for unit i. The adjustment to characteristic *k* is $\sum_{j=1}^{N} x_{k,j} / \sum_{i=1}^{n} w_i x_{k,i}$. The estimators of the total and the variance are given below.

$$\hat{Y}_k = \sum_{j=1}^{N} x_{k,j} \frac{\sum_{i=1}^{n} w_i \hat{Y}_{k,i}}{\sum_{i=1}^{n} w_i x_{k,i}}, \quad v(\hat{Y}_k) = \frac{\left(\sum_{j=1}^{N} x_{k,j}\right)^2}{\left(\sum_{i=1}^{n} w_i x_{k,i}\right)^2} \sum_{i=1}^{n} w_i^2 \hat{e}_{k,i}^2, \text{ where } \hat{e}_{k,i} = \hat{Y}_{k,i} - x_{k,i} \frac{\sum_{j=1}^{n} w_j \hat{Y}_{k,j}}{\sum_{j=1}^{n} w_j x_{k,j}}$$

The above formula is for sampling with replacement to ratio estimator. The true variance estimator is:

$$v(\hat{Y}_k) = \frac{n}{n-1} \frac{\left(\sum_{j=1}^{N} x_{k,j}\right)^2}{\left(\sum_{i=1}^{n} w_i x_{k,i}\right)^2} \sum_{i=1}^{n} w_i^2 \hat{e}_{k,i}^2 \text{ , where we can ignore } \frac{n}{n-1} \text{ when } \frac{n}{n-1} \approx 1.$$

NASS uses the formula below to solve the problem of sampling without replacement for the crops survey in Guangdong province of China, adding an adjustment coefficient 1- p_i in the formula.

$$v(\hat{Y}_k) = \frac{\left(\sum_{j=1}^{N} x_{k,j}\right)^2}{\left(\sum_{i=1}^{n} w_i x_{k,i}\right)^2} \sum_{i=1}^{n} (1 - p_i) w_i^2 \hat{e}_{k,i}^2$$

There are many variables when we need to collect statistical data in Chinese agricultural surveys, such as for example, area of cultivated land, area of permanent crops, area of woodland, area of grassland; rice, wheat, corn, leguminous crops, rape seed, peanut, cotton, hemp-crops, sugar crops, tobacco, medicinal crops, vegetables, melons, fodder crops, etc. For this reason the MPPS design should be used to draw the sample and calculate the estimates, since otherwise it is difficult for us to design the survey. The National Bureau of Statistics of China uses the MPPS

method to design Chinese agricultural surveys with the help of NASS. The MPPS sampling design is also currently applied to study the Chinese enterprise survey. We believe that this design has variety of applications in China and around the world.

THE ESTIMATORS OF THREE KINDS OF SAMPLING DESIGNS

We have obtained the estimators for simple random sampling and probability proportional to size sampling as below.

SRS: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $v(\bar{y}) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, where $f = \frac{n}{N}$

PPS: $\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$, $v(\hat{Y}) = \frac{1-\hat{f}}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\frac{y_i}{z_i} - \hat{Y})^2$, where $\hat{f} = \sum_{i=1}^{n}\pi_i / n$

The estimator of SRS sampling is well-known. Cochran (1977), Wolter (1985) and Feng Shiyong (1998) gave the formulas of SRS in their textbooks. Wolter (1985) and Feng Shiyong (1998) also gave the estimator of PPS sampling. It is easy for us to use the estimator of SRS sampling to get the estimator of PPS sampling. We can get the result by using $\frac{y_i}{z_i}$ in PPS sampling to substitute y_i in SRS sampling and by changing the adjustment coefficient. When the sample size is big enough, if the ratio estimator of MPPS sampling is treated as a simple estimator, we can get the estimator of MPPS design with the method given below.

MPPS: $\hat{Y}_k = \sum_{j=1}^{N} x_{k,j}\frac{\sum_{i=1}^{n} w_i \hat{Y}_{k,i}}{\sum_{i=1}^{n} w_i x_{k,i}} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i \hat{Y}_{k,i}$, where $\alpha_i = \frac{n\sum_{j=1}^{N} x_{k,j}}{\sum_{i=1}^{n} w_i x_{k,i}} w_i$

$$v(\hat{Y}_k) = \frac{1-\hat{f}}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\alpha_i \hat{Y}_{k,i} - \hat{Y}_k)^2, \text{ where } \hat{f} = \sum_{i=1}^{n} p_i / \sum_{i=1}^{N} P_i \approx \sum_{i=1}^{n} p_i / n$$

A uniform estimator for SRS, PPS and MPPS sampling can be obtained as below when we summarize the above three kinds of estimators.

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\alpha_i y_i,\ v(\hat{Y}) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\alpha_i y_i - \hat{Y})^2, \text{ where } f = \frac{\sum_{i=1}^{n} q_i}{\sum_{i=1}^{N} Q_i}$$

We can discuss this uniform estimator according whether the sampling method is with or without replacement.

Sampling with Replacement

When sampling with replacement is applied, we have $f = 0$. If the sampling design is a SRS, then $\alpha_i = N, i = 1, \cdots, n$. If the sampling design is a PPS, then $\alpha_i = \frac{n}{\pi_i} = \frac{1}{z_i}, i = 1, \cdots, n$.

If the sampling design is MPPS, then $\alpha_i = \alpha_{k,i} = \frac{n\sum_{j=1}^{N} x_{k,j}}{\sum_{i=1}^{n} w_i x_{k,i}} w_i, i = 1, \cdots, n$ for characteristic *k*.

Sampling without Replacement

If the sampling design is a SRS, then:

$\alpha_i = N$, $q_i = 1$, $Q_i = 1$, $i = 1, \cdots, n$, $f = \frac{n}{N}$

If the sampling design is a PPS, then:

$$\alpha_i = \frac{n}{\pi_i} = \frac{1}{z_i}, q_i = \pi_i, \quad Q_i = \Pi_i, \quad i = 1, \cdots, n, \quad f = \frac{\sum_{i=1}^{n} \pi_i}{\sum_{i=1}^{N} \Pi_i} = \frac{\sum_{i=1}^{n} \pi_i}{n} = \sum_{i=1}^{n} z_i$$

If the sampling design is a MPPS, then

$$\alpha_i = \alpha_{k,i} = \frac{n \sum_{j=1}^{N} x_{k,j}}{\sum_{i=1}^{n} w_i x_{k,i}} w_i \text{ for characteristic } k$$

$$q_i = p_i, \quad Q_i = P_i, \quad i = 1, \cdots, n, \quad f = \sum_{i=1}^{n} p_i \Big/ \sum_{i=1}^{N} P_i .$$

The students of statistics specialty study SRS design first, then PPS design in their sample survey course. We have no idea if there is a topic of MPPS design in textbooks of statistics up to now. The Food and Agriculture Organization of United Nations published the book called Multiple Frame Agricultural Surveys in 1998; this discusses agricultural survey programmes based on area frame or dual frame (area and list) sample designs. As a matter of fact, the sampling survey course is only for the students of statistics specialty in colleges and universities in China. Some Chinese professors are going to give lectures about MPPS design and write this content in textbooks. The students can understand SRS, PPS and MPPS designs better if we tell them that the three kinds of designs have a uniform estimator as a summary after they study these three kinds of designs. We even can simply use Excel to choose sample and calculate estimates for these three kinds of designs in practical survey.

DISCUSSION

There are some research papers about MPPS sampling design published in Chinese magazines, for example Statistical Research. We do not know if there is a school which offers the course on MPPS sample design in China, but we do know some teachers and students are studying this design and writing relevant research papers. It is not difficulty for the students of statistics specialty to study MPPS design after they have studied PPS design in schools for both PPS and MPPS designs have the same principle to be implemented in practice. We do not think that it would take a lot of time to give the lectures about MPPS design for these students. It is acceptable for both teachers and students of statistics specialty to have this course.

This paper briefly introduces MPPS sampling design developed by NASS, and then gives a uniform estimator of SRS, PPS and MPPS sampling designs. The uniform estimator is easy for us to know the relationship among these three kinds of sampling designs and gives us an overall view about them. We hope that there will be more sampling designs to join this family and the students can understand sample designs better from it.

In practical survey design, the ratio estimator can get more precise results. We use practical data to simulate the model and get the result also. Further more, we have studied the precision of MPPS sampling design with replacement and without replacement using different adjustment coefficients. We have compared the results of three different adjustment coefficients and obtained some significant results. It tells us that there are still some problems that exist in theory and practice for us to study.

REFERENCES

Cochran, W.G. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Shiyong, F., Jiaxun, N., and Guohua, Z. (1998). *Theory and method of sampling survey*. Beijing: China Statistics Publishing House.

The Food and Agriculture Organization of United Nations. (2000). *Multiple Frame Agricultural Survey*. Beijing: China Statistics Publishing House.