Thailand Statistician
January 2012; 10(1) : 129-140
http://statassoc.or.th
Contributed paper

Estimation of Proportion of a Finite Population

Yongyuth Chaiyapong

Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand.

Email: yongchai@chiangmai.ac.th

Received: 13 February 2012
Accepted: 2 April 2012

Abstract

This paper discusses and presents maximum likelihood estimators of proportion for finite population as an alternative to the standard estimator. The properties of the obtained random sample under simple random sampling are investigated. The likelihood functions for both with and without replacement sampling schemes are constructed and the Maximum Likelihood estimators for proportion are derived. The properties of estimators are also investigated analytically. It was found that, under simple random sampling without replacement, the Maximum Likelihood estimator is biased with a larger variance comparing to that of standard estimator, sample proportion. The bias adjusted estimator may be a potential candidate in estimating population proportion, and may also lead to more accurate inferences in term of interval estimation or hypothesis testing. Moreover, the concrete link between theory of sample survey and theory of statistical inference is elaborated. The discussion given would enable for in depth knowledge in statistics, i.e. understanding in concepts, principles, and theories of statistics to be rigorously developed. This, in turn, allows for cognitive skills, one of the keys learning outcomes in statistics education at tertiary level, be achieved.

Keywords: finite population, maximum likelihood estimation, proportion, simple random sampling, statistics education.

1. Introduction

In conducting a sample survey, when simple random sampling is employed, the estimation of population proportion, $P = A/N$, is intuitively based on the sample proportion, $p = a/n$, as long been suggested in Murthy [1], Cochran [2], Sampford [3], and Thompson [4]. Indeed, intuition is always valuable, and in many regards, intuition or even common sense plays a vital role in statistical inference both for finite and infinite populations. Obviously, the very distinct examples are the estimation of population characteristics in theory of sample survey, population average and proportion, in particular.

However, these estimators solely depend on personal capability and experiences which could be learned and appreciated. It would be more beneficial if the development of these estimators could be connected to the standard inferential methods such as Maximum Likelihood Estimation (MLE). Throughout the years, many attempts on formal estimation procedures have been pursued, as discussed in Smith [5], Hansen [6], Rao [7,8]. These are, in fact, demonstrated the rigorous links among theory of sample survey and theory of standard statistical inference and theory of probability. As a statistician, clear understanding in the development of estimation procedures together with the key properties of those derived estimators is essential, not only for choosing the most appropriate one but also for developing a new estimator when needed, or for conducting research works so as to improve the efficiency of those existing ones.

In this paper, formal estimation procedure for population based on simple random sampling is discussed. The fundamental objective of the discussion is to demonstrate and reveal the extension on the applications of method of Maximum Likelihood Estimation for population proportion in finite population domain. In addition, the highlights on the links among theory of sample survey to the theory of standard statistical inference and theory of probability especially in the case of the discrete probability is also a main focus of the discussion. Ultimately, it is aimed at establishing a more concrete foundation on theory of sample survey so as to improve or enhance key learning outcomes i.e. knowledge and cognitive skills, Allen [9], for statistics education and also to develop the theoretical grounds for survey statisticians of the country.

2. Maximum Likelihood Estimation for Proportion of a Finite Population

Maximum Likelihood Estimation, one of the most classical methods of finding estimator which has been well accepted to yield estimators with required basic properties in statistics science. Based on the joint probability function of sampled data, the relevant likelihood function is constructed and consequently, when likelihood principle is valid, the required estimator is derived. Maximum Likelihood estimators could be either in explicit forms, or may be obtained directly from the relevant likelihood function through appropriate computational techniques.

In general, when the sample data are independent identically distributed random sample, construction of likelihood function is simple and straight forwards, as the product of the underlying distribution of the data serving as the joint probability function of the sampled data yields thus required likelihood function. However, for sampled data from a finite population, particularly, when without replacement sampling scheme is employed, the analysis of the properties of the obtained random sample is essential, so that the likelihood function of the data is correctly constructed.

Under simple random sampling, when a finite population of interest containing N population units of fixed values $y_1, y_2, \ldots, y_N$ is sampled one at a time, until n of population units are achieved to be sampled data. Therefore, at the end of the sampling process, n sampled units are obtained. Thus, random samples are in fact a collection of random variables $Y_1^*, Y_2^*, \ldots, Y_n^*$ with specific properties. The relevant probability distribution(s) of each of the random variable depends on thus employed sampling scheme, as well as the population characteristics to be estimated.

To estimate the population proportion, a population of interest could be viewed as a collection of N units of $y_1, y_2, \ldots, y_N$ that each of these units could take only two values, either 1 or 0, for those being under the category of interest and otherwise respectively. Consequently, with replacement sampling scheme, each of the random sample $Y_1^*, Y_2^*, \ldots, Y_n^*$ is in fact a Bernoulli random variable, taking the value of “1” when a unit under the category of interest is sampled or otherwise “0”. On each of the sampling occasions, as a previously sampled population unit is returned back into the population and ready to be sampled again, as at the original stage; therefore, it is

dramatic that $Y_1^*, Y_2^*, \ldots, Y_n^*$ is independent and identically distributed random sample under $Bernoulli\,(P)$, in accordance with the random experiment performed.

Hence, the likelihood function is

$$L(P, y^*) = \prod_{i=1}^{n} P^{y_i^*} (1-P)^{1-y_i^*} \quad (1)$$

From the likelihood function (1), the Maximum Likelihood estimator of the parameter P, as in standard theory of statistical inference, is typically in the closed form

$$\hat{P}_{wr} = \frac{1}{n} \sum_{i=1}^{n} Y_i^* . \quad (2)$$

Thus, the estimator as obtained in (2) is simply the sample proportion, $p = a/n$, which is the same as the long been suggested and used estimator as discussed above. In addition, as had already been proved in all the standard text books in theory of sample survey, the estimator is in fact unbiased, with the variance

$$Var(\hat{P}_{wr}) = \frac{P(1-P)}{n} \text{ or } \frac{A}{nN}\frac{N-A}{N}$$

Under simple random sampling without replacement, however, the construction of the likelihood function needs further analysis on the properties of the random sample $Y_1^*, Y_2^*, \ldots, Y_n^*$. Firstly, it is dramatic that the set of these random variables are not independent, as the occurrence of each Y_i^* indeed affects the others. Obviously, the random variable Y_i^* is Bernoulli distributed, however, with changing different value of the probability of success, as a unit from a category of interest is obtained with a probability solely relies on previous sampling history. The likelihood function of the random sample $Y_1^*, Y_2^*, \ldots, Y_n^*$ for estimating the population proportion may be constructed based on the verification of the conditional probabilities relating to the sampled data.

On the contrary, it is more efficient to treat the probability function of the sum of those random variables $Y_1^*, Y_2^*, \ldots, Y_n^*$ as an alternative to their joint probability

function, and therefore, the likelihood function. The sum of $Y_1^*, Y_2^*, \ldots, Y_n^*$ is actually a number of units from a category of interest that are obtained from sampling without replacement of n units out of the population of N units which are classified into two group of categories, i.e. A units belonging to the category of interest, and $N - A$ in the other.

The probability function of the sum, in fact, the total number of successes from sampling n units without replacement from the population of size N containing A units of interest, is known to be Hypergeometric distribution. Unlike the case of sampling with replacement, the sum is binomially distributed. Thus, the likelihood function of the data with respect to A as the parameter of interest could be obtained as follow:

$$L(P, y^*) = \frac{\binom{A}{\Sigma y_i^*}\binom{N-A}{n-\Sigma y_i^*}}{\binom{N}{n}}. \qquad (3)$$

The estimation of the parameter A would therefore lead to the estimation of population proportion $P = \frac{A}{N}$, which is the required population characteristic.

For the estimation of parameter A, the likelihood function (3) is a discrete function in A. Therefore, the Maximum Likelihood estimate could be obtained by directly searching from the likelihood function, computationally or graphical display. A more efficient method in obtaining the estimate is based on the ratio of the likelihood function

$$D(A) = L(A) / L(A+1).$$

As discussed in Zhang [10], subject to certain criteria, the maximum likelihood estimator for the parameter A is of the form;

$$\hat{A} = \begin{cases} \frac{\sum_{i=1}^{n} Y_i^*(N+1)}{n} - 1, \text{ or } \frac{\sum_{i=1}^{n} Y_i^*(N+1)}{n}; \text{ if } \frac{\sum_{i=1}^{n} y_i^*(N+1)}{n} \text{ is an integer} \\ \frac{\sum_{i=1}^{n} Y_i^*(N+1)}{n} \quad ; \text{ if } \frac{\sum_{i=1}^{n} y_i^*(N+1)}{n} \text{ is not an integer} \end{cases} \qquad (4)$$

From (4), based on invariance principle, the estimation of population proportion derived from the MLE of $\hat{A}$ is therefore

$$\hat{P}_{wor} = \frac{\hat{A}}{N}. \quad (5)$$

3. Properties of the Maximum Likelihood Estimator for Proportion

Clearly, this estimator as given in (5) is no longer the sample proportion, $p = a/n$, as traditionally recommended. The properties of thus obtained estimator could also be verified based on the properties of the sum of the random sample $Y_1^*, Y_2^*, \ldots, Y_n^*$, the random variable under Hypergeometric distribution as above discussed, with the expectation and variance

$$E(\sum_{i=1}^{n} Y_i^*) = n\frac{A}{N}, \text{ and}$$

$$Var(\sum_{i=1}^{n} Y_i^*) = n\frac{A}{N}\frac{(N-A)}{N}\frac{N-n}{N-1} \text{ respectively.}$$

To investigate the properties of the estimator $\hat{P}_{wor}$, the properties of the MLE $\hat{A}$ is examined. As described in (4), it could be seen that $\hat{A}$ is based on a function of the random sample $Y_1^*, Y_2^*, \ldots, Y_n^*$, or a statistic

$$T(Y_i^*) = \frac{\sum_{i=1}^{n} Y_i^*(N+1)}{n}. \quad (6)$$

Therefore, initially, the properties of the statistic $T(Y_i^*)$ as described in (6) are to be examined.

Based on the expectation of the random variable under Hypegeometric distribution, the expectation and variance of the statistic $T(Y_i^*)$ are:

$$E\left(T(Y_i^*)\right)=\frac{(N+1)}{N}A, \text{ and}$$

$$Var\left(T(Y_i^*)\right)=\frac{(N+1)^2}{n}\frac{A}{N}\frac{(N-A)}{N}\frac{N-n}{N-1}.$$

To evaluate the overall performance of the estimator $\hat{P}_{wor}$, the probability function of the statistic $T(Y_i^*)$ must be evaluated. However, the evalutions for each of those conditions could be done as follows. When the value of the statistic $T(Y_i^*)$ is an integer, the MLE $\hat{A}$ could either be $T(Y_i^*)-1$, or $T(Y_i^*)$, which give different values of the estimate for A, and therefore lead to different values of estimate for P. In fact, when the value of $T(Y_i^*)$ is an integer, the estimator $\hat{P}_{wor}=\frac{\hat{A}}{N}$ could be either of the forms

$$\hat{P}_{wor1}=\frac{1}{N}\left(T(Y_i^*)-1\right), \quad (7)$$

or $$\hat{P}_{wor2}=\frac{1}{N}\left(T(Y_i^*)\right). \quad (8)$$

Both of the estimators as in (7) and (8) are biased, since their expectations are not equivalent to the population proportion P,

$$E(\hat{P}_{wor1})=\frac{(N+1)}{N}P-\frac{1}{N},$$

and $$E(\hat{P}_{wor2})=\frac{(N+1)}{N}P.$$

Therefore the biases of the estimators are respectively:

$$B(\hat{P}_{wor1})=\frac{1}{N}(P-1); \quad (9)$$

and $$B(\hat{P}_{wor2}) = \frac{1}{N}P \ . \qquad (10)$$

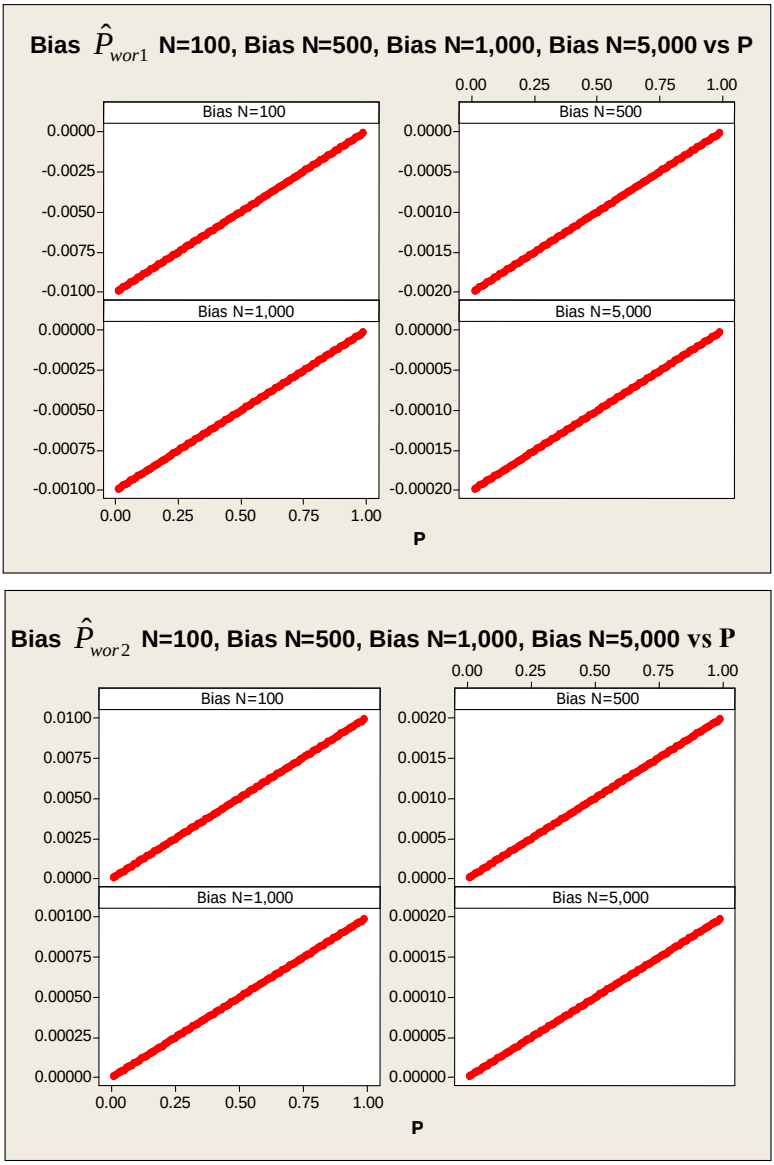

Figure 1: Biases of $\hat{P}_{wor1}$ and $\hat{P}_{wor2}$, for $0.01 \le P \le 0.99$, with $N = 100, 500, 1{,}000, 5{,}000$

The bias of the estimator $\hat{P}_{wor1}$ as given in (9) is always negative in contrast to that of the estimator $\hat{P}_{wor2}$, as given in (10). This means that the former underestimates and the latter overestimate the true value of the population proportion accordingly. For large N, the biases of both estimators tend to zero. The values and behaviour of these bias functions for various population sizes could be observed from Figure 1 and 2 as given above and below.

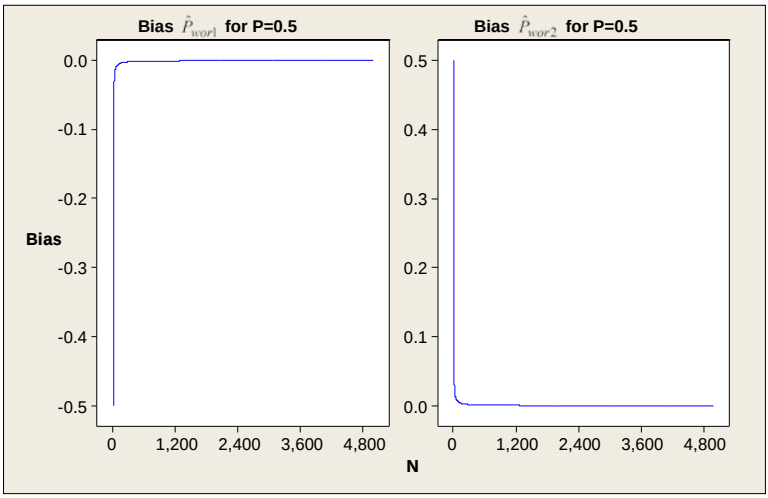

Figure 2: Biases of $\hat{P}_{wor1}$ and $\hat{P}_{wor2}$, for $P = 0.50$, with $1 \le N \le 5{,}000$

It could also verify that the variances of the estimators are the same as

$$Var(\hat{P}_{wor1}) = Var(\hat{P}_{wor2}) = \left(\frac{N+1}{N}\right)^2 \frac{A}{nN}\frac{(N-A)}{N}\left(\frac{N-n}{N-1}\right). \qquad (11)$$

Due to the contribution of the first term of $\left(\frac{N+1}{N}\right)^2$ which is always greater than one when N is an interger, then the variance estimators in (11), is always greater than that

of the sample proportion $Var(p) = \frac{A}{nN}\frac{N-A}{N}\left(\frac{N-n}{N-1}\right)$. For large N, the value of this term tends to one. Hence, assymptotically, the variance of the estimators $\hat{P}_{wor1}$ and $\hat{P}_{wor2}$ is as of the variance of the sample proportion.

When the statistic $T(Y_i^*)$ is not an integer, the value of the estimator is obtained as the round down of the value of the statistic to the nearest integer, this could also be considered as another form of the estimator for population proportion, therefore

$$\hat{P}_{wor3} = \frac{1}{N}\left[T(Y_i^*)\right]. \quad (12)$$

The estimator $\hat{P}_{wor3}$ as given in (12) is also biased as $\hat{P}_{wor2}$ since they are both in the form of the statistic $T(Y_i^*)$. However, the bias function as well as the variance of the estimator must take into account the contributions of the round off decimal values of the statistic $T(Y_i^*)$. In fact, the decimal values of this statistics depend solely on the term $\frac{N+1}{n}$, which is for a certain population and fixed sample size is a constant and could be computed when the sample size is already determined.

Obviously, the bias and variance of the estimator $\hat{P}_{wor3}$ depends solely on the term $\frac{N+1}{n}$ of the statistic $T(Y_i^*)$. It is clear that the round off decimal values is always less than one so the bias of the estimator $\hat{P}_{wor3}$ must be in between the bias of the estimators $\hat{P}_{wor1}$, and $\hat{P}_{wor2}$. The investigation on the bias as well as the variance functions of the estimator $\hat{P}_{wor3}$ could be preceded by investigating the occurrence of the decimal values due to the term $\frac{N+1}{n}$ of the statistic $T(Y_i^*)$.

The investigation on the overall performances of the estimator $\hat{P}_{wor}$ needs the properties of those respective estimators $\hat{P}_{wor1}$ or $\hat{P}_{wor2}$, together with $\hat{P}_{wor3}$. However in practice we can always choose the sample size that gives the value of the statistic $T(Y_i^*)$ to be an integer. The knowledge on the properties of the estimators $\hat{P}_{wor1}$ or $\hat{P}_{wor2}$ may be adequate.

4. Conclusion

The distinction of population proportion from the probability of success, the parameter of Bernoulli distribution, needs to be clearly identified so that the estimation procedure could be developed correctly. The analysis on the properties of random sample is essential as the relevant likelihood function is then constructed correctly. This requires deep appreciation in theory of probability, under the domain of finite sample space and discrete probability, in particular.

The link of the theory of standard inference for finite population and the theory of sample survey, especially their common foundation of the theory of probability, provides in depth understanding in the theory of sample survey and would therefore enhance the skills in situation analysis as well as creative problems solving in both practical applications and theoretical domains of statistics science. In practice, based on the properties of the estimators as discussed, when simple random sampling without replacement is employed, for a large population, the sample proportion is still recommended, and for a relatively large population, the estimator $\hat{P}_{wor}$ with bias adjusted may be considered as a better alternative in estimating population proportion.

To establish clear knowledge in theory of sample survey which then leads to and establish cognitive skills in statistics education, the discussion on the foundation on theory of probability and theory of standard statistical inference together with the links among them must be emphasized. Statisticians need to be trained and possess all the knowledge as above discussed and provided. It is essential that statistics curriculum should be revised at all levels on these aspects.

References

[1] Murthy, M.N., Sampling Theory and Methods, Statistical Publishing Company, 1967.

[2] Cochran, W.G., Sampling Techniques, 3rd edition, John Wiley & Sons, 1977.

[3] Sampford, M.R., An Introduction to Sampling Theory, Oliver and Boyd, 1962.

[4] Thompson, M.E., Theory of Sample Survey, Chapman & Hall, 2002.

[5] Smith, T.M.F., The Foundations of Survey Sampling: a Review, J. R. Statist. Soc. A, 1976.

[6] Hansen, M.H., Some History and Reminiscences on Survey Sampling, Statistical Science, 1987; 2: 180-190.

[7] Rao, J.N.K., Interplay between Sample Survey Theory and Practice: An Appraisal, Statistics Canada, 2005; 31: 117-138.

[8] Rao, J.N.K., Empirical Likelihood Methods for Sample Survey Data: An Overview, Austrian Journal of Statistics, 2006; 35: 191-196.

[9] Allen, I. National Qualifications Framework for Higher Education in Thailand, Implementation Handbook. 2006.

[10] Zhang, H., A Note about the Maximum Likelihood Estimator in the Hypergeometric Distribution, Comunicaciones en Estadı´stica, Diciembre 2009.