A labor-saving sampling method for estimating a normally distributed population by using a local median

Takashi WADA,1,* Satoru URANO1 and Kohji YAMAMURA2
1 National Agricultural Research Center for Kyushu Okinawa Region; Koshi 861–1192, Japan
2 National Institute for Agro-Environmental Science; Tsukuba 305–8604, Japan
(Received 9 December 2005; Accepted 11 August 2006)

Abstract
We propose a sampling method using a median (random-median sampling) for estimating the mean and variance of a normally distributed population. First, a sampling unit is selected at random, and then two sampling units adjacent to the selected sampling unit are drawn. These three units are compared and only the unit with the median quantity among the three units, which we call the local median, is adopted. This procedure is repeated for a given sample size. A computer simulation was conducted to compare the required sample size for random-median sampling with that for simple random sampling for homogeneous populations. The sample size required to attain a given precision of estimates by random-median sampling relative to that by random sampling (=100) was shown to be constant, i.e. about 45%, irrespective of the mean and standard deviation of the population; therefore, this principle is generally applicable for any normally distributed populations. The discrimination limit (DL) is defined as 10% of the mean of the population when considering practical field selection of the median by eye. Random-median sampling with a DL in which samples are randomly selected among candidates when the difference between three or two, candidates is smaller than the DL, also decreased the sample size significantly. The sample sizes required by random-median sampling were 45–60% of those required by simple random sampling when the DL ratio (=standard deviation of the population/DL) was greater than 1.5. Therefore, in field sampling, when selection of a median is not time-consuming, for example, when examining soybean yield, random-median sampling with a DL saves labor in comparison with random sampling, at least for homogeneous populations. Unbiased variances of samples from random-median sampling and random-median sampling with a DL were about 45% and 45–60% of those from random sampling, respectively. Thus, it is possible to estimate the variance of the population using both random-median methods.

Key words: Labor-saving sampling; normal distribution; median

INTRODUCTION
Samplings from a population are frequently taken in order to estimate, in particular, the mean and the variance of the parent population in various fields of industry and research. In crop protection, samplings are often implemented at considerable cost in terms of both labor and expense, to assess crop pest densities or crop yields. If it were possible to estimate a population from a smaller sample size, this would save a considerable amount of labor. The merit of reducing the sample size is more evident when samples require certain treatments after sampling, or when the space for storing samples is insufficient. In this paper, we propose a sampling method using a median as a labor-saving approach, when the difference in the quantity of individuals can be easily discriminated. As an example of such an assessment, we adopted the estimation of soybean yield in a field. In this situation, it is possible to discriminate plants that have a higher yield from a few plants selected by eye, when the difference between the plants is sufficiently large to be noticeable. Here we report the results of a simulation demonstrating that the new method can markedly reduce the required sample size in comparison with ordinary random sampling.

* To whom correspondence should be addressed at: E-mail: twada@affrc.go.jp
DOI: 10.13031/aez.2007.9
METHODS

Proposed sampling method. We propose a sampling method referred to as random-median sampling (RM sampling). This method is a modification of two-stage sampling; it comprises a random sampling step to yield a primary sampling unit, and the selection of a median, which we call the local median, among three secondary sampling units. In conventional two-stage sampling procedures, secondary sampling units are usually selected at random from each of the secondary units; however, random sampling is laborious in most cases. Hence, we further propose another simplified procedure. We first selected a secondary unit simultaneously with the selection of a primary unit. Then, we selected two secondary units that are adjacent to the selected secondary unit. In the case of the examination of soybean yields, for example, we first selected a plant by random sampling. Then, we selected the two plants flanking it in the same row. The reason for selecting the two adjacent plants is simply to minimize the cost in obtaining a secondary sampling unit. Next, we sampled the plant with the second largest yield (local median) among the three candidates, by judging from the appearance of the plants by eye; thereafter, we continued with the next sampling until a given sample size was reached.

We also propose another related sampling method referred to as random-median sampling with a discrimination limit (RM-DL sampling). This method is defined as random-median sampling with discrimination errors. When choosing the median among three candidates, this method permits discrimination errors when the difference in the quantity among the candidates is small. In this method we assume a discrimination limit (DL). If the difference among the candidates is smaller than the DL, we choose a sample randomly from among the candidates. In the present study, the DL was fixed at 10% of the population mean.

In the case of soybean field sampling, it is not possible to discriminate the median plant from the three candidates by eye when the difference in the plant yields between two or three candidates is small; therefore, we fixed the DL at 10% of the population mean. If the difference in plant yields between either of the two candidates or among three candidates is smaller than the DL, then the sampling unit is selected randomly. We consider this method to be practical for field sampling because it is easy to select the median plant from among three candidates by eye when the difference in plant yield is large enough. When the difference is small, we can choose a sample that looks "apparently" median without hesitating about discrimination error. The merit of this approach is that sampling can be performed with minimal extra labor in comparison with ordinary random sampling.

Simulation study. We conducted a sampling simulation to estimate yield in a soybean field. We considered a normally distributed population of 5,000 soybean plants with a mean (±standard deviation) of 58.6 (±12.8). We assumed a homogeneous soybean field of 10 a; the yield follows independent identical normal distribution at any place in the field. The mean and standard deviation of the population were derived from data on grain weight per plant for a specific soybean cultivar (Kyushu143) obtained in 2002 in a soybean field in Kumamoto Prefecture, Japan (Wada et al., 2006). Individual plant grain data for this finite normally distributed population were generated using Mathematica Ver. 3 (Wolfram Research, Inc.). The virtual soybean plants were numbered from 1 to 5,000. We selected a plant at random. Then, we selected two adjacent plants. Three these plants were compared and the plant of median yield was adopted.

We simulated samplings from the virtual population and estimated the mean and variance of the population. The samplings applied were RM sampling, RM-DL sampling and ordinary random sampling. Sampling with a replacement was permitted for the sake of simplicity. We selected sample sizes from 1 to 300 and estimated the mean of the population. If the estimated mean of the samples ranged from 95% to 105% of the actual population mean, then we considered the estimation "successful". We performed simulated samplings 2,000 or 9,000 times for each sample size in each sampling method. We defined a "successful estimation rate (\%)" as the rate of samplings per 2,000 or 9,000 trials in which the means of the samples were successfully estimated. By using successful estimation rates, we can determine the required sample size at any specific level of precision for each sampling method. We defined "relative sample size (RSS)" as the percentage of the sample size required by
each respective method relative to the sample size (=100) required for random sampling. We also defined the "DL ratio" as the value of DL relative to the standard deviation of the population (standard deviation/DL). Since the required sample size at any specific level in random sampling can be obtained theoretically using population variance (Cochran, 1977), we calculated the required sample size at the 70% and 90% probability level by random sampling and the values were compared with the results obtained from the simulation. All simulations were performed by F-BASIC ver. 6.3 (Fujitsu Middleware Co. Ltd.).

RESULTS

Comparisons of the successful estimation rates among the three sampling methods after 9,000 simulated samplings from the population at each sample size are shown in Fig. 1. For random sampling, the curve for the estimation rate against sample size rose less steeply than for the other two methods, indicating that estimated population means were less accurate. The theoretical values of the sample size at which the means of the samples attained successful estimation (population mean ±5%) at 70% and 90% probability were about 22 and 54, respectively. These theoretical values coincided with the results of the simulation. For RM sampling, the curve rose more steeply, indicating more accurate estimation for each sample size. Figure 1 shows that for RM sampling the sample sizes required for successful estimation at 70% or 90% probability were roughly half of those for random sampling. The curve for RM-DL sampling almost overlapped that for RM sampling, indicating a more accurate estimation of the population mean than for random sampling at each sample size.

The relative sample size (RSS) at various levels of precision is shown in Fig. 2. Comparisons of required sample sizes between each of the two RM methods and random sampling were conducted 15 times after 2,000 simulated samplings at each sample size. The values of RSS for RM sampling were almost constant at about 45%, and those for RM-DL sampling were also almost constant at about 50%. These results indicate that the population mean can be estimated with the same accuracy by RM sampling or RM-DL sampling using a sample size that is about half of that required for random

Fig. 1. Proportion of successful estimation of the mean of a population from samples obtained with three sampling methods (random, random-median and random-median with DL). The three methods used are described in the text. A normally distributed population with a mean of 58.6 and a standard deviation of 12.8 was used for simulation. Simulated samplings were performed 9,000 times for each sample size. If the sample mean is 95–105% of the population mean, this sampling is regarded as a successful estimation. The vertical axis indicates the proportion of successful estimations of 9,000 trials. The numerals indicate theoretical values for a sample size attaining 70% and 90% successful estimation in random sampling. The values coincided with the simulation results.

Fig. 2. Relative sample size (RSS) in two random-median samplings. RSS is defined as the value (%) of the required sample size attaining each level of precision in each random-median sampling relative to that (=100) obtained by random sampling. Simulated samplings were performed 2,000 times for each sample size. The required sample size at each level of precision was obtained from three sampling methods, and subsequently the relative value (%) was calculated. These procedures were repeated 15 times and the averages (± S.D.) were plotted in the figure.
sampling, irrespective of the proportion of successful estimates. The values of RSS at 50% successful estimation were somewhat different from the others and had wider standard deviations in both RM and RM-DL methods. This is because the values of the required sample sizes at this rate were small and discrete (integral number) in all of the sampling methods, resulting in greater variance of the RSS values.

Figure 3 shows RSS at 90% successful estimation when the standard deviation of the population moved from 3.2 to 25.6 (eight steps) with a constant population mean (58.6). The values of RSS for RM sampling were almost constant at about 45%, irrespective of the values of the standard deviations. The values of RSS for RM-DL sampling approached those for RM sampling as the standard deviation of the population increased, and gradually approximated 100% when the standard deviation became smaller. When the standard deviation of the population increases, the probability of a difference among the three candidates in secondary sampling lying within DL decreases, and thus RM-DL sampling approximates RM sampling. In the same way, when the standard deviation becomes smaller, samplings among the candidates within DL increase, and thus RM-DL sampling substantially approaches random sampling. Thus, in RM-DL sampling, the values of RSS depended on the DL ratio, and were less than 60% when the DL ratio exceeded 1.5. At other successful estimation rates, the trend for RSS should be similar to the case at 90% shown in Fig. 3, because RSS is constant, irrespective of the successful estimation rates, as indicated in Fig. 2.

Figure 4 shows the values of RSS at 90% successful estimation when the populations mean moves from 29.3 to 117.2 (four steps) and the value of the standard deviation remains constant (12.8). The value of RSS in RM sampling is constant at about 45%, irrespective of the change in the population mean. The values for RM-DL sampling approach those for RM sampling with a decrease in the population mean, and approximate 100% as the mean becomes larger. This is because DL is fixed at 10% of the population mean. When the mean decreases, the value of DL declines, and thus RM-DL sampling substantially approximates RM sampling. In the same way, when the mean increases, RM-DL sampling approaches random sampling. When the mean moves, it may not be practical to fix DL at 10% of the population mean because the DL changes substantially. We can therefore consider the case for fixing the value of DL. In such a case, however, the shape of the normal distribution of the population does not change. Therefore, the trends of RSS in RM sampling and
RM-DL sampling should be similar to those in Fig. 2.

From the simulations shown in Figs. 1 to 4, we can conclude that RSS, which indicates the proportion of the required sample size relative to that for random sampling, is constant at about 45% in RM sampling, irrespective of the target precision and the values of the mean and standard deviation of the population. In RM-DL sampling, RSS is also below 60% when the DL ratio is sufficiently large (more than about 1.5). These results indicate that both RM sampling and RM-DL sampling effectively reduce the sample size when estimating the mean of a normally distributed population at any level of precision.

We can then consider estimation of the variance of a population. Figure 5 shows the relationship between unbiased variances of samples from RM sampling and random sampling. There is a linear relationship between estimated variances; the variances for RM sampling are about 45% of those for simple random sampling. Simple random sampling yields unbiased estimates of population variance. Thus, we can estimate the variance of a population by multiplying the variance estimated from RM sampling by 100/45.

The relationship between the unbiased variances from RM-DL sampling and from random sampling depends on the DL ratio. As has already been described, RM-DL sampling approximates RM sampling when the DL ratio is large. Accordingly, unbiased variances of the samples from RM-DL sampling gradually approach 45% of the unbiased variance of samples from random sampling (Fig. 6). When the DL ratio becomes smaller, the unbiased variances from RM-DL sampling approximate those from random sampling (relative variance=100). When the DL ratio is larger than 1.5, the relative variances range from 45% to 60%, and thus we can roughly estimate the population variance by RM-DL sampling.

**DISCUSSION**

The results of the present simulation revealed that RM sampling can effectively reduce the sample size to about 45% of that required by random sampling. In other words, RM sampling enhances the precision of population estimates if we fix the sample size. This principal is generally applicable in any normal-distributed population, irrespective of its mean or standard deviation. Therefore, RM sampling is labor-saving if the cost of selecting the median is small. The balance between the cost required for sampling and the cost required for treatment of the samples determines whether RM sam-
sampling should be adopted. Often it may be difficult to select the median by eye. In such cases, use of appropriate apparatus may be necessary in order to apply the RM sampling method.

RM-DL sampling also effectively reduces the sample size when DL is 10% of the population mean and the DL ratio is greater than 1.5. Using soybean yield as an example in which actual observed data of the mean and standard deviation were used, the reduction of the sample size facilitated by RM-DL sampling was almost comparable to that for RM sampling; therefore, we can recommend this method for practical sampling of crops in the field. In soybean sampling to estimate yield, the extra cost of secondary sampling is quite small; the selection of a median among three candidates by eye can be performed immediately because errors in selecting the median are permissible when the differences in plant yields are small. On the other hand, the cost of treatment after sampling (storage and drying of plants, removal and weighing of seeds, etc.) is generally great. Therefore, reduction of the sample size performed with RM-DL sampling greatly contributes to labor-saving for population estimation.

Further reduction of sample size can be expected if we increase the numbers of candidates in the secondary sampling unit. Again, the balance of the cost of each step determines the proper candidate size for secondary sampling. For soybeans, the number of candidates in secondary sampling should be three because the cost of secondary sampling increases appreciably when the number of candidates exceeds five.

The amount of the reduction in sample size in both MR sampling methods, however, may decrease in actual field samplings. A field is usually more or less heterogeneous due to various causes. Correlations among the candidates will be unavoidable when we select adjacent candidates. The effectiveness of sampling decreases with increasing the correlation. We can reduce this inefficiency by selecting candidates randomly in determining the local median although the sampling cost increases. This procedure is especially recommended if the samples require complex treatments after sampling; i.e., the cost of the treatment is higher than the additional cost of secondary sampling.

Sampling procedures are often performed for populations in various fields. Both RM sampling methods will be widely applicable in field research and industrial inspection; for example, the measurement of organs of specimens drawn from a large numbers of collections, sampling for quality inspection in industrial goods. These methods are especially recommended, as mentioned above, if sampling units have fewer spatial correlations. In this paper we have examined the effect of RM sampling in a population that follows a normal distribution; however, animals and plants often do not show a normal distribution. We can expect a significant reduction of the sample size and estimation of the population parameters not only in normally distributed populations but also in other symmetrically distributed populations (Yamamura et al., unpublished data). We will focus on the theoretical foundation of random-median sampling, including various types of distribution, in a separate paper.

ACKNOWLEDGEMENTS

We thank the anonymous referees for their comments that improved the manuscript.

REFERENCES
