Scientific and Technical Articles/ Articles scientifiques et techniques

A Microcomputer Program for Assisting in the Design of Simple Random Samples

by

P.L. Marshall¹

Abstract

An interactive microcomputer program was developed to aid the design of simple random sampling with or without replacement. The program determines: (1) sample size for a set of given conditions for up to 20 variables; (2) combinations of conditions that will yield a given sample size; and (3) the relationship between percentage error and sample size for a given set of conditions. Potential applications are illustrated with three simple examples.

Key Words: sample size, simple random sampling

Résumé

Un programme interactif de micro-ordinateur a été développé pour aider l'élaboration d'un plan d'échantillonnage aléatoire simple avec ou sans remise. Le programme détermine: (1) la taille de l'échantillon pour un ensemble de conditions données comprenant jusqu'à 20 variables; (2) des combinaisons de conditions qui produiront un échantillon d'une taille donnée, et (3) la relation entre le pourcentage d'erreur et la taille de l'échantillon pour un ensemble de conditions données. Des applications possibles sont abordées à l'aide de trois exemples simples.

Mots-clés: Taille de l'échantillon, échantillonnage aléatoire simple.

Introduction

Simple random sampling (SRS) (either with or without replacement) is widely used in virtually all facets of forestry. Determination of sample size (n) for SRS is straightforward if an estimate of population variability is available and desired sampling error and significance levels are known. In many instances, however, the designer is initially unclear as to what levels of sampling error and significance are appropriate. Thus it is often beneficial to examine the impact of several combinations of sampling error and significance levels on n . It is also useful to examine the relationship between sampling error and n for given significance levels. The formulas for calculating n require t or F values which depend on the value of n . Exact solution requires an iterative procedure that adjusts the t or F values for each new estimate of n . When n is to be calculated for a single variable under a variety of different combinations of sampling error and significance levels or for a number of variables that are to be observed on each sample unit, the number of calculations required increases materially.

The potentially large number of calculations required can make the determination of an appropriate sample size time

consuming and laborious even for a simple sampling design like SRS. However, the process can be simplified by automating the calculation of n , and allowing the sample designer to examine the implications of several design criteria on n at one time. This paper reports on an interactive microcomputer program (SASCAL — Sample Size Calculation) designed for this purpose.² Examples are included to illustrate simple applications of this program.

Program Features

SASCAL is written in BASIC for use on IBM or IBM compatible personal computers. The present version accepts only SRS designs (with or without replacement). The program consists of three options: (1) calculate n for a given set of conditions (coefficient of variation (CV), desirable percentage error (PE) and significance levels) for up to 20 variables; (2) calculate sets of conditions that yield a given n ; and (3) determine the relationship between PE and n for given CV and significance levels. The user has the choice of specifying "beta" significance levels in addition to the traditional "alpha" significance levels for any of these options. The program is menu-driven and the options can be selected in any order and

¹Assistant Professor, Department of Forest Resources Management, Faculty of Forestry, University of British Columbia, 2357 Main Mall, Vancouver, B.C. V6T 1W5

²Copies of a compiled version of this program and a user's manual can be obtained by sending a formatted, double-sized, double-density diskette and \$10.00 to cover copying and postage costs to the author.

as frequently as desired. Results can be output to the screen, the printer or a file.

The Use of Beta Significance Levels

Most users specify only alpha significance levels when calculating sample size. Alpha significance levels represent the proportion of all possible samples of size n that would yield confidence intervals encompassing the population parameter being estimated (Li 1964). If n is calculated using only alpha significance, each confidence interval will have a 50:50 chance of exceeding the width implied by the sampling error or PE component in the sample size formula (Harris *et al.* 1948, Payandeh and Beilhartz 1978). This is a result of estimates of population variability being sample-based, thus subject to variability themselves.

Beta significance represents the probability that the confidence interval will not exceed the length implied by the sampling error or PE component in the formula used to calculate sample size (Harris *et al.* 1948). Ensuring that the confidence interval will not exceed a specified length more than 50% of the time requires a larger sample size than when alpha significance levels alone are specified (Li 1964).

One situation where a sampler may wish to consider using an explicit beta significance level involves decisions based on the results of composited samples (e.g., foliar analyses and soil analyses). When samples are composited before analysis, it is impossible to construct a confidence interval for an estimate of the population mean or total. If the sampler wishes to compare an estimate based on such a sample with some standard, it may be advantageous to control the probability that the confidence limit of the estimate will exceed some specified value. Ballard (1985) and Marshall and Jahraus (1987) have discussed this point in relation to foliar nutrient analyses.

A second situation arises when the sampler wishes some assurance against the results of a "bad" example. A sample may be considered "bad" if the confidence interval of the estimate exceeds the desired width. This can happen as a result of the chance selection of non-representative sample units being drawn from the population. "Bad" samples can occur even if a good estimate of the population variability (or CV) is used for calculating sample size. The probability of a large deviation from the expected confidence interval width will increase as the precision of the estimate of population variability decreases. Many samplers address the possibility of a "bad" sample in an *ad hoc* fashion by using "conservative" estimates of population variability for calculating n or by adding some arbitrary number of sample units to a calculated n . The explicit use of beta significance levels removes this arbitrariness. It allows the user to explore the impact on sample size (hence sampling cost) of achieving a particular level of confidence that some maximum error level will not be exceeded. It is worth noting, however, that it is necessary to estimate population variability from a pilot sample if sample size calculations incorporating beta significance levels are to be properly employed.

Information Required

SASCAL requires option information and program information to be provided by the user. Option information consists of Yes/No responses to program prompts and option numbers chosen from menus. Option information governs the order of branching within the program and determines the

form of the appropriate equation. Program information is the information necessary to solve this equation.

Most of the required program information consists of sample design criteria (i.e., PE and significance levels). Choice of appropriate values requires some thought on the part of the user, but seldom requires any expense. Estimation of CV, on the other hand, may require a pilot survey of the population to be sampled. In fact, the size of the pilot survey is required for calculating sample size (equation 3 or 4) when beta significance levels are explicitly used.

Option 1 — Calculating Sample Size

Sample size is affected by three factors: 1) variation of the variable of interest (estimated as the sample variance or CV); 2) the desired sampling error (expressed in absolute terms as a confidence interval half-width or as a PE); and 3) the desired alpha and beta significance levels. Sample size for SRS with replacement when only alpha significance is specified may be calculated as (Freese 1962);

$$n = \frac{t^2 (1 - \alpha/2, n-1) * CV^2}{PE^2} \quad [1]$$

The corresponding formula for sampling without replacement is:

$$n = \frac{1}{\frac{1}{N} + \frac{CV^2 * t^2 (1 - \alpha/2, n-1)}{PE^2}} \quad [2]$$

where N is the population size. If a beta significance level greater than 50% is desired, the following equation (Harris *et al.* 1948, Li 1964) is appropriate for SRS with replacement:

$$n = \frac{CV^2 * F (1 - \alpha) (1, n-1) * F (1 - \beta) (n-1, \Theta)}{PE^2} \quad [3]$$

where Θ is the degrees of freedom associated with the estimate of population variance (one less than the pilot sample size), $1 - \alpha$ equals the alpha significance level and $1 - \beta$ equals the beta significance level. The corresponding formula for sampling without replacement is:

$$n = \frac{1}{\frac{1}{N} + \frac{CV^2 * F (1 - \alpha) (1, n-1) * F (1 - \beta) (n-1, \Theta)}{PE^2}} \quad [4]$$

Equations 1 through 4 require t or F percentile values for solution. Since the appropriate value of t or F depends on the sample size, an iterative procedure is necessary to solve these equations. If the user was required to enter appropriate values for each iteration, the routine would quickly become tedious. In order to avoid this difficulty algorithms were used to approximate percentile values from the t - and F -distributions. These values are not exact, but their use offers considerable savings in computational time over cumbersome table-lookup procedures. Tests of accuracy in the literature (e.g. Sahai and Thompson 1974) and tests carried out by the author (see User's Manual) indicated that the approximations work very well except for situations where degrees of freedom are very small (e.g. < 5).

Example of Option 1

Consider the simple problem of determining the number of 0.05 ha plots required to estimate the average volume per ha of a 100 ha block of timber ($N=2000$). The sampling procedure is to be SRS without replacement. Ten plots were randomly located in the block as a pilot sample and the CV was estimated to be 15%. The sample size will depend on the levels of significance and PE selected.

SASCAL can handle up to 20 sample size calculations in each run. Each calculation may be for a different variable (i.e. involve different CV's) or for the same variable with different specifications (i.e. different levels of PE, alpha and beta significance). Three levels of PE (5, 10 and 15%), three levels of alpha significance (90, 95 and 99%) and two levels of beta significance (50 and 90%) were examined. Eighteen sample size calculations were required. SASCAL took 13 minutes to produce the sample sizes given in Figure 1 on an IBM-PC/XT. More than 11 minutes of this time was occupied by data entry.

Option 2 — Calculation of Conditions That Yield a Given n

If n is set at some arbitrary value, equations 1 and 2 can be solved for various combinations of PE and alpha significance, and equations 3 and 4 can be solved for various combinations of PE, alpha and beta significance. Algorithms documented in Park *et al.* (1982) were used for approximating alpha and beta significance levels. No indication of the accuracy of these algorithms was found in the literature, but tests made by the author (see User's Manual) indicated that the accuracy was quite acceptable for sample design purposes.

Example of Option 2

Sometimes an idea of sample size can be obtained from preliminary budgets. Whether that size is actually suitable for the intended use of the sample depends on the expected PE, which in turn is related to significance levels if n is fixed. The same size of sample can be a product of various combinations of PE and significance levels. Examining some of these combinations can help the sampler decide whether a particular sample size is suitable for the intended use.

Option 2 of SASCAL performs the necessary calculations to produce these combinations. As an illustration of this

Figure 1. Sample sizes for different combinations of PE, alpha and beta significance levels.

capability, some calculations leading to a sample size of 40 were calculated for the same forest block employed in the example for Option 1. Beta significance was fixed at 50% and allowed to assume higher values. The results of this analysis are displayed in Figures 2a and 2b. The production of these results required only a few minutes, including the time spent entering the necessary information.

Option 3 — Calculation of the Relationship between PE and n

Equations 1 through 4 can also be reformulated to express PE as a function of n , CV and values of t , F and N where appropriate. These formulations allow easy computation of the relationship between PE and n for given CV and significance levels. The routine included in SASCAL requires the user to specify CV and significance levels (as well as N when appropriate), and solves the reformulated equations for values of n ranging from 10 to 500 in increments of 10. The algorithms employed for estimating t and F values in the sample size formulas are also used in these calculations.

Figure 2a. Combinations of PE and alpha significance yielding a sample size of 40 (beta significance = 50%).

Figure 2b. Combinations of PE and alpha and beta significance yielding a sample size of 40 (beta significance = 50%).