

An Alternative Approach to Selecting a 3P sample¹

P.W. West

SciWest Consulting, 67 Gahans Rd, Meerschaum Vale, NSW 2477, Australia and School of Environmental Science and Management, Southern Cross University, Lismore, NSW 2480, Australia

Web: www.nor.com.au/users/pwest/index.htm

Abstract

Sampling with probability proportional to prediction (3P sampling) is a well developed method of sample selection, used quite extensively in some types of forest inventory. In forest mensuration texts, a more or less ‘standard’ protocol has been developed as to how the 3P sample should be collected and how the estimates of the population total and its variance (hence confidence limit) should be determined from the sample. This standard protocol requires that *each and every individual sampling unit in the entire population* be visited during the sampling process and an estimate made, by eye, of the value of the characteristic which is being measured in the population. This restricts the use of 3P sampling to populations which are sufficiently small in size that it is feasible practically to visit each and every sampling unit in it. This paper suggests an ‘alternative’ protocol, both for collection of the 3P sample and determination from it of the estimates for the population. The alternative requires that only as many sampling units be visited and estimated as are necessary to obtain the required size of the 3P sample. Once that has been achieved, the sampling process may stop. However, the method still requires that the total number of sampling units in the population be determined.

Introduction

In any inventory, where the characteristics of a population are to be estimated by measuring a sample drawn from the population, it is usually desired to use the most ‘efficient’ technique to collect the sample. In mathematical statistical terms, this means using a technique which minimises the sampling effort, whilst still achieving the desired precision of the result. In simple forest inventories, ‘stratified random sampling’, ‘model-based sampling’ or ‘sampling with varying probability of selection of sampling units’ are three sampling techniques used. All are more efficient than ‘simple random sampling’ (West 2004).

To use these more efficient techniques, some prior information must be available about the forest before sampling starts. For stratified random sampling, strata must have been identified which divide the forest into areas which are more uniform with respect to the forest characteristic which is being estimated in the inventory. For model-based sampling or sampling with varying probability of selection, some measure of the size of each and every sampling unit in the entire population must be available before sampling starts.

One method of sampling with varying probability of selection is known as ‘sampling with probability proportional to prediction’, commonly referred to as 3P sampling. The

¹ Paper presented to the 2005 meeting of the Western Forest Mensurationists, Naniloa Resort, Hilo, Hawaii, USA, 4th-7th July 2005.

method was invented by L.R. Grosenbaugh. Shiver and Borders (1996) and Iles (2003) give some of the background of how the method developed. In America particularly it seems to be used quite widely for forest inventory.

The great idea behind 3P sampling is that it can be used to carry out sampling with varying probability of selection *without* there being available information about the size of each and every sampling unit in the entire population before sampling starts. Instead, the person collecting the sample generates that information by estimating, by eye, the size of a sampling unit as it is encountered during the inventory. Strict rules are then applied to determine whether or not that sampling unit will indeed be included in the sample. If it is included, the desired characteristic is formally and properly measured on it.

Over the years since 3P sampling was introduced, a more or less ‘standard’ protocol has developed as to how it should be done. This is described in texts on forest measurement (e.g. Shiver and Borders 1996, Avery and Burkhart 2002, Iles 2003), together with details of how the results are used to obtain the estimate from the sample of the total of the desired characteristic over the whole population and its variance, hence, its confidence limit.

Part of the standard protocol is that it is necessary for the sampler to visit and estimate, by eye, the value of the desired characteristic of *each and every sampling unit in the entire population*. If the population is very large, say with many thousands of sampling units, this becomes an impossibly large task. In practice, it means that 3P sampling must be restricted to relatively small populations, of such a size that it is reasonable for the sampler to visit all the sampling units.

This paper proposes an alternative protocol for collecting a 3P sample and determining the results from it. This alternative does not require that all the sampling units in the population be visited during the sampling process.

Standard Protocol for Selection of a 3P sample

The standard protocol for selecting a 3P sample from a population involves the following four basic steps. For this paper, I have adapted the steps and the mathematical terminology from Shivers and Borders (1996, pp 303-5) and Avery and Burkhart (2002, pp 262-263):

- 1) From a preliminary survey of the population, estimate the total, summed over the whole population, of the characteristic being measured (T'_x) and the maximum value of the characteristic which will occur in any individual sampling unit (K).
- 2) Decide on the sample size required (n_e).
- 3) Determine a value κ (which is referred to somewhat confusingly as $K+Z$ or KZ in other works) as

$$\kappa = \max(T'_x/n_e, K) \quad . \quad (1)$$

This value serves simply to set an upper limit in the sampling process which follows and helps to ensure that the sample size achieved finally is close to n_e .

- 4) Set out to visit each and every sampling unit in the entire population. As each sampling unit is visited, estimate by eye the value of the characteristic being considered in the

inventory. Generate (by whatever method is available) a random value in the range 0- κ . If the estimate of the characteristic for that sampling unit is greater than or equal to the random value, include that sampling unit as part of the 3P sample and actually measure the characteristic on it. Otherwise, move on to the next sampling unit.

After these four steps have been followed and the 3P sample selected, suppose there were found to be a total of N sampling units in the whole population and that n of them were included in the 3P sample. The value of κ determined with Eq. (1) will ensure the value of n is close to n_e .

For the sampling units included in the 3P sample, there will then be n values ($X_i, i=1\dots n$) of the estimates of the characteristic made on them and n values ($Y_i, i=1\dots n$) of the actual measured value of the characteristic taken from them. There will also be $N-n$ values of the estimates ($x_i, i=1\dots N-n$) of the characteristic on the $N-n$ sampling units which were visited, but not actually measured.

One problem that may arise during this process is that an individual sampling unit is encountered for which the estimate by eye exceeds κ . For any such sampling units, the value of the characteristic must be actually measured. At the end of the process, those sampling units will be considered as a quite separate sub-population of the whole population and the information collected from them will be dealt with separately, as discussed later. This sub-population has been referred to as the ‘sure-to-be-measured’ population (eg Wiant 1976). Suppose there are N_s such individuals, then N will be the number of sampling units in the remainder of the population.

Calculating the Results

Once a 3P sample has been selected collected using this standard protocol, the results may be used to estimate the population total (Y_T) of the characteristic and its variance (V_T) as follows.

First, the population total of the estimates made for every sampling unit in the population (Y_X) is determined as

$$Y_X = \sum_{i=1\dots n} (X_i) + \sum_{i=1\dots N-n} (x_i) \quad (2)$$

The ratios of measured to estimated values in the 3P sample are then used to adjust T_X to give the final estimate of the population total (Y_T) as

$$Y_T = Y_X \sum_{i=1\dots n} (Y_i/X_i)/n \quad (3)$$

This is the equation shown near the bottom of p 264 of Avery and Burkhart (2002) and is also Eq. (10.1) of Shivers and Borders (1996).

There seem to have been many attempts to determine an expression to give the variance of Y_T , though none seems to have been entirely successful. On their p 265, Avery and Burkhart (2002) suggest that an approximate formula, which they feel gives satisfactory results, is

$$V_T = \{ \sum_{i=1\dots n} [(Y_i Y_X / X_i) - Y_T]^2 \} / \{ n(n-1) \} \quad (4)$$

This formula was developed by Grosenbaugh and is repeated as Eq. (10.2) of Shivers and Borders (1996), in a somewhat different mathematical formulation.

The confidence limit about the estimate of the total (C_T) may be determined as

$$C_T = t\sqrt{V_T} \quad (5)$$

where t is the value of Students t for the required probability and with $(n-1)$ degrees of freedom. Since Eq. (4) is only an approximate estimator for V_T , the confidence limit may be estimated alternatively using the ‘bootstrap’ technique².

If desired, the estimate of the mean of the population (Y_M), its variance (V_M) and its confidence limit (C_M) may be determined as

$$Y_M = Y_T/N \quad (6)$$

$$V_M = V_Y/N^2 \quad (7)$$

and

$$C_M = t\sqrt{V_M} \quad (8)$$

When selecting the 3P sample, if a ‘sure-to-be-measured’ population was identified, the measured values from those sampling units are summed and added to the value of Y_T . Their mean is added to the value of Y_M . Neither the variances nor the confidence limits of either of these are altered from the values determined above. Because each individual in the ‘sure-to-be-measured’ population was measured, a complete inventory was made of them and so has no variance, hence, contributes nothing to the estimates of the variance or confidence limit of the total or mean.

A Different Method of Calculating the Results

West (2004) suggested that the formally established estimators for any form of sampling with varying probability of selection could be used to estimate the population total (Y_T) and its variance (V_T). These estimators are given in Eqs. (3.7) and (3.9) of Schreuder et al. (1993). In the mathematical terminology used here, they are

$$Y_T = \sum_{i=1..n} (Y_i/p_i) \quad (9)$$

and

$$V_T = (1/2)\sum_{i,j=1..n, i \neq j} \{[(p_i p_j - p_{ij})/p_{ij}][Y_i/p_i - Y_j/p_j]^2\} \quad (10)$$

where

$$p_{ij} = p_i p_j N(n-1)/[n(N-1)] \quad (11)$$

In his Eq. (10.8), West determined that that the probability of selection of the selected sampling units in the 3P sample ($p_i, i=1..n$) was

$$p_i = (n_v/N)X_i/\kappa \quad (12)$$

² Bootstrapping involves re-sampling, at random and with replacement, from the original sample to make a new sample of the same size. The estimate of the population total is then determined from this new sample. This process is then repeated a large number (say 1,000) times, to give 1,000 new estimates of the population total. The estimates are then arranged in order from smallest to largest. The particular estimates within that 1,000, which are spaced equally above and below the estimate of the population total from the original sample, and within which 95% (say) of the 1,000 new estimates lie, can then be considered the upper and lower limits of the 95% confidence interval. Probability levels other than 95% can be determined similarly with this process.

where n_v is the total number of sampling units that were visited and estimated for the desired characteristic, before the final n sampling units were selected and measured³.

Alternative Protocol to Selection of a 3P Sample

If the method outlined in Eqs. (9-12) is used to calculate results from a 3P sample, it leads, in effect, to an alternative protocol for collecting the sample. The differences from the standard protocol are as follow:

- 1) Because the total of the estimates made on each and every sampling unit in the population (Y_x) does not appear in Eqs. (9-12), it is no longer necessary to visit each and every sampling unit and estimate, by eye, the characteristic being measured. However, to obtain the value of N , it is necessary to count the sampling units. If simply counting them is much quicker and easier than estimating them, time would be saved.
- 2) Visiting sampling units, estimating their values and, if including the sampling unit in the 3P sample, measuring it, would cease as soon as the required number of sampling units had been included in the sample (n_e). The total number of sampling units visited up to that point would then be n_v , as required in Eq. (12).
- 3) Because sampling stops when n_e sampling units have been included in the sample, there is no longer any need to estimate initially the total of the characteristic summed over the whole population (T'_x). However, the maximum value of the characteristic which will occur in any individual sampling unit (K) must still be estimated. The value used for κ will then be the value of K , rather than the value defined in Eq. (1).
- 4) An important limitation of this alternative protocol is that it fails if a sampling unit is encountered for which the estimate made of it exceeds κ (which equals K). This means that it will be necessary to deliberately over-estimate the largest value which will be encountered in the population. However, it follows from the random value method, by which the selection of the 3P sampling units is made, that the larger the over-estimate, the larger will be the number of sampling units which have to be visited before the required sample size is achieved.

An Example Using the Standard and Alternative Protocols

The population considered in this example was the 107 trees in an 0.25 ha plot of *Eucalyptus maculata* regrowth forest in northern New South Wales. The objective of the sampling exercise was to estimate the total stem wood volume of all the trees in the population.

Suppose that the standard 3P sampling protocol was used. Suppose that, before setting out to take the sample, it was estimated that the total stem wood volume on the plot was $T'_x = 48 \text{ m}^3$ and that the largest individual tree stem wood volume that would be found in

³ West's Eq. (10.8) differs slightly from Eq. (12) because he suggested that, before sampling started, an estimate be made of the smallest value, as well as the largest (K), of the desired characteristic which would be encountered in the entire population. This has ramifications for the range from which random values are selected in choosing which sampling units are to be included in the 3P sample. In the description of 3P sampling given here, it is assumed that the smallest value is zero. Avery and Burkhart (2002) gave the same equation for the probability of selection of 3P sampling units near the bottom of their p 262. In their case, every sampling unit in the population was assumed to have been visited, so n_v/N in Eq. (12) has a value of 1.

the population was $K = 2.1 \text{ m}^3$. Suppose it was desired to collect a 3P sample of size $n_e = 15$. Using Eq. (1) then gives $\kappa = 3.2 \text{ m}^3$. Each and every one of the 107 sampling units in the plot (the individual trees) would then be visited and its stem wood volume estimated by eye, to give the results in the 'Estimate' column in Table 1. As each was visited, a random value from the range 0-3.2 m^3 would be selected. If the estimated volume of the tree exceeded the random value, the tree would be included in the 3P sample and its stem wood volume actually measured, using whatever measurement technique was considered appropriate. The measured volumes of the trees so selected are shown in the 'Standard' column of the table. After completion of this procedure, a total of $n = 13$ trees were actually selected in the 3P sample. This will be termed here the 'standard' 3P sample.

Suppose then that the alternative sampling protocol was used also to collect an 'alternative' 3P sample. The same value for $K = 2.1 \text{ m}^3$ would be used as in the standard sample, giving a value of $\kappa = 2.1 \text{ m}^3$ for the alternative. To be consistent with the standard sample, suppose it was desired to collect a 3P sample of size $n_e = 13$. The sampler would then set out to visit trees in the population, selecting or rejecting them for inclusion in the alternative 3P sample based on random values selected in the range 0-2.1 m^3 . The measured volumes of the trees so selected are shown in the 'Alternative' column of the Table 1. After $n_v = 67$ trees had been visited, it was found that $n = 13$ trees had been included in the alternative 3P sample. Sampling then stopped. The remaining 40 trees which were not visited are marked with an asterisk in the 'Alternative' column of the table.

For the standard 3P sample, the computations to determine the estimates of the total stem wood volume of the trees in the population and its confidence limit were done in several ways. Firstly, the standard method given by Eqs. (2-5) was used. Secondly, since there may be some uncertainty about the use of Eq. (4) as an unbiased estimator of the variance of the total, the confidence limit was estimated using the bootstrap technique. Thirdly, the alternative method, given by Eqs. (9-12 and 5), was used with $n_v = N = 107$. For the alternative 3P sample, only the alternative method of computation was used, with $n_v = 67$.

The results of these computations are given in Table 2. They show that there was little practical difference between the results obtained with either of the samples and any of the computation methods

Table 1. Stem wood volume (m³) data collected when selecting a 3P sample from the trees growing in an 0.25 ha plot of *Eucalyptus maculata* regrowth forest in northern New South Wales, Australia. The plot contained 107 trees. The column marked ‘Estimate’ gives the estimates made by eye of the volumes of all the trees in the plot. The columns marked ‘Standard’ and ‘Alternative’ show the measured volume of those trees included in the 3P sample when it was selected using the standard or alternative protocols, respectively, as discussed in the text. In the alternative column, trees marked with an asterisk (*) were not visited during sample selection.

Estimate	Standard	Alternative	Estimate	Standard	Alternative	Estimate	Standard	Alternative
			0.635			0.279		
			0.590			0.277		
			0.564		0.552	0.246		*
1.846	*		0.563	0.562		0.233		
		1.52				0.233		*
1.641		9				0.212		
1.416		*						
1.310		*						
1.308		1.45						
		7						
1.276			0.549		*			
			0.525					
1.198	1.312		0.492		*	0.210		*
		1.07				0.206		
1.094		4	0.489			0.199		
		1.19				0.194		*
1.091		4	0.488			0.189		
1.053	0.958		0.469		*	0.173		
			0.468		0.520	0.169		*
0.974			0.431			0.168		
0.931	0.993	*	0.431	0.484		0.166		
		*				0.158		*
0.925		0.91	0.407			0.151		*
0.858		3	0.393		*	0.149		*
0.811		*	0.377			0.148		*
0.804		*	0.376	0.333	*	0.145		
0.794	0.707	*	0.347			0.140		
		*			0.30	0.124		0.12
0.781	0.851	*	0.339		1	0.123		*
		0.66	0.335			0.120		*
0.740		2	0.328	0.369	*	0.118		*
		0.67				0.109		*
0.740	0.675	5	0.327			0.108		
			0.327			0.101		
0.739			0.327			0.100		*
0.715	0.726	*	0.326	0.323				
			0.319					
0.676		0.68	0.311					
		0	0.308					
0.666			0.304		*			*
0.641		0.56	0.283					
0.641		5						

0.099	*	0.092		0.079	0.093
0.096	*	0.092	*	0.078	*
0.094	*	0.086		<hr/>	

Table 1. (Continued)

Estimate	Standard	Alternative
0.074		
0.070		*
0.069		
0.061		
0.055		
0.055		*
0.054		
0.052		
0.050		
0.048		
0.032		
0.027		*
0.025		*
0.015		

Table 2. For the standard and alternative 3P samples Results of various calculation methods to estimate the total stem wood volume and its confidence limit of the trees in a eucalypt forest plot.

3P sample	Calculation method	Estimate of plot total volume (m³)	95% confidence limit (m³)
Standard	Eq. (2-5)	45.9	2.7
Standard	Eq. (2) and Bootstrap	45.9	2.3
Standard	Eq. (9-12 and 5)	42.5	2.3
Alternative	Eq. (9-12 and 5)	43.2	2.1

Conclusion

In the example given here, 3P samples selected using the standard protocol or the alternative protocol suggested here gave similar results. The most important difference between the standard and alternative protocols is that the standard requires that each and every sampling unit in the entire population be visited and an estimate made by eye of the value of the characteristic which is being measured on the population. In practice, this restricts the use of 3P sampling to smaller populations. The alternative protocol avoids this problem. Only as many sampling units need to be visited and estimated as is necessary to obtain the sample size required. However, the total number of sampling units in the population must still be determined. The alternative protocol fails if, during sampling, a sampling unit is encountered for which the estimate of its size exceeds the maximum value it was assumed, at the outset, would be encountered.

Acknowledgements

I am indebted to Prof Harry Wiant for discussion on some of the concepts considered in this paper.

References

- Avery TE, Burkhardt HE (2002) *Forest Measurements*. 5th Edition. McGraw-Hill, New York
- Iles K (2003) *A Sampler of Inventory Topics*. Kim Iles & Associates Ltd, Nanaimo, British Columbia
- Schreuder HT, Gregoire TG, Wood GB (1993) *Sampling Methods for Multiresource Forest Inventory*. Wiley, New York
- Shiver BD, Borders BE (1996) *Sampling Techniques for Forest Resource Inventory*. Wiley, New York
- West PW (2004) *Tree and Forest Measurement*. Springer, Berlin
- Wiant HV (1976) *Elementary 3P Sampling*. Bulletin 650T, West Virginia University Agricultural and Forestry Experiment Station, Morgantown, West Virginia