SECOND GENERATION SELECTION WITH PROBABILITY PROPORTIONAL TO PREDICTION

Ray K. Strickland[1/]

Abstract. --Probability proportional to prediction (3-P) sampling has been used increasingly in forestry during the past ten years. In progeny tests, ranking families from best to poorest for the many traits under consideration on the basis of total enumeration, is cost prohibitive; a less expensive system of subsampling is essential. Second generation selection using 3-P subsampling is ideal for this purpose. The generation methods discussed should be applicable to both coniferous and hardwood tree improvement programs.

Additional keywords: Forest tree improvement, family selection, probability sampling.

INTRODUCTION

In the southeast, progeny tests are beginning to be utilized as sources of selection material for our next generation of seed orchards. We often attempt, initially at least, to make inferences concerning relative merits of progeny lines on the basis of total enumeration of all the trees for all the characteristics in which we desire to make genetic gains. It soon becomes obvious, however, that this degree of quantification is cost prohibitive so we begin to consider possible methods of subsampling.

One possibility is to sample every other, every third, etc. tree in each plot in each block (systematic sampling) and use this information to rank our families. Or, perhaps a random sample of a fixed number of trees within each plot would work just as well. Many such sampling schemes of varying intensity and degree of randomness could be employed and might do a satisfactory job of identifying the families from which to select our next generation of breeding trees.

One method utilizing 3-P requires total enumeration of progeny tests for one or perhaps two characteristics, usually disease and/or volume. The remaining traits are sampled only on those lines whose values and/or volumes surpass a certain established minimum, usually check average or some percentage above that. Finally, the individuals within the lines chosen for subsampling are selected with probability proportional to their size or value. Another possibility, which would reduce field work, would require only an estimated population mean and number of trees following a cursory examination of a progeny test.

[1/]Assistant Assistant Professor, University of Florida.

Grosenbaugh (1963) introduced probability proportional to prediction sampling to forestry. Two approaches using the technique for second generation selection are developed here. Other feasible approaches may also be evident for particular needs.

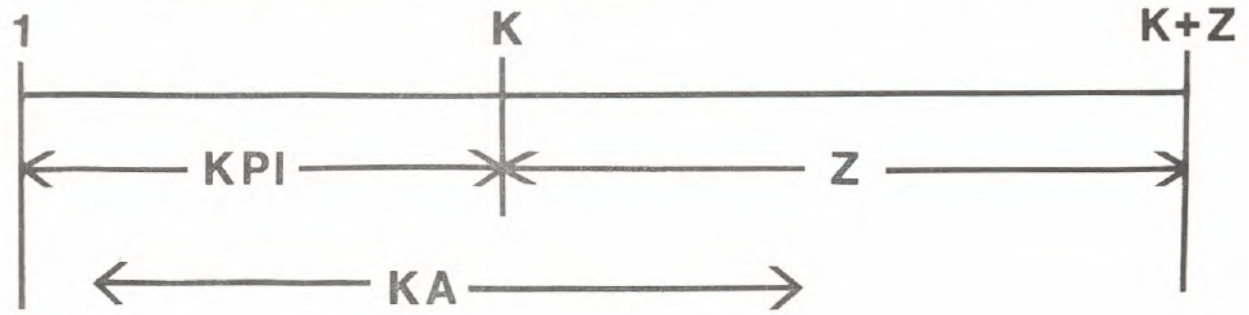
Suppose you have a progeny test that you wish to subsample but have no list or previous measurement information on hand. For simplicity we'll assume we hex(little or no disease but if it is a problem, total enumeration for disease is necessary. A cursory inspection of the progeny test is essential to estimate Y, the average tree size, N, the estimated number of surviving trees and K, the largest expected tree. The size of the subsample, n, needed to properly quantify the families within the progeny test must then be estimated. Usually an eve] age of 15 to 20 trees per family is adequate for a relative rating. A random number generator that is known to be unbiased is needed to produce the 3-P list for field subsamplin,3. The group of random numbers generated, Z, that are in the range from K to K + Z and are defined as "nulls" since they are greater than any of the expected observations in the test. The number K + Z which is put intc the computer to produce the appropriate random array is calculated in the following manner :

$$\hat{K} + Z = \frac{(\hat{N})(\bar{y})}{n} = \frac{\hat{T}}{n} \qquad (1)$$

where $\hat{T}$ is the estimated population total.

A random array dimensioned to the same size and layout as the progeny test is generated using the numbers from 1 through K + Z with replacement (Figure 1). Since the values for K and T may be obtained by estimation or actual measurement, Figure 1 is developed as with actual values. KPI is the random range of sizes or values from 1 through K which corresponds directly to the expected tree sizes or values in the progeny test. The "nulls" are randomly distributed in the array depicting the progeny test and those trees associated with a "null" are not considered for subsampling unless they fall into the "sure to be measured" category. An adjustment, KA, can be made to make K larger or smaller than the actual largest or most valuable tree. If KA is made smaller than K, those individuals in the test that fall in the range between KA and K are "sure to be measured" trees (i.e. they have a probability of 1 of coming into the subsample). On the other hand, if KA is made larger than K, no individual has a probability of 1 of being in the subsample. No adjustment in KA means it is left equal to K.

Figure 1.--The 3-P sampling system depicted as points along a line. Explanation is as detailed in the text (above).



Other than designated "sure to be measured" trees, only individuals in the KPI range are considered for subsampling. If the size or value of a non-"null" tree in the KPI range is equal to or greater than the associated volume or value in the random array, it is designated for subsampling for all the characteristics

The other procedure proposed assumes total enumeration for disease and size or value. With this approach we can eliminate families that do not meet specified standards relative to level of infection and size or value. Also, trees within retained lines that are diseased are eliminated as subsample candidates. The purpose is to rank the remaining families from best to poorest. The same procedure is used except we know T, the sum of the retained individuals, and K, the largest tree, because we have measured it. Therefore, the equation for K + Z is changed slightly to:

$$K + Z = \frac{\sum_{i=1}^{N} Y_i}{n} = \frac{T}{n}$$

where N is the number of individuals within retained families and n is the number of observations desired in the subsample as before.

Subsequent to total enumeration, where the arrays relative to disease and size are in the computer, the generated random array is compared to the retained size array. Those non-"null" individuals that are larger than or equal to the random array KPI's are designated by the number 1 and all others (i.e. "nulls" and individuals smaller than KPI) are designated 0. This gives the position in the field of those individuals to be subsampled for the other characteristics of interest. Actually, the size and disease arrays used for selection purposes can come from previous measurements as well as from current data. Consider as an example slash pine, Pinus elliottii Engelm. var. elliottii, in a high risk fusiform rust, Cronartium fusiforme Hedgc. & Hunt ex Cumm., area. Rockwood and Goddard (1973) found that selection of progeny lines, rather than individual trees, for rust resistance gave the best results. Those lines with highest rust incidence should be eliminated as second generation prospects for utilization in any moderate to high risk fusiform rust area. Also, rarely should a second generation candidate with rust be accepted, even for planting in a low risk area.

DISCUSSIONS AND CONCLUSIONS[1/]

The 3-P subsampling techniques for second generation selection are simple and quite easy to apply. Our purpose is to designate those families within a progeny test that are acceptable as parental families for our next breeding generation. We are generally interested in visiting only the best two or three families within a progeny test to make individual tree selections. The techniques outlined here place major emphasis on freedom from disease and/or size or value. Secondary emphasis is then placed on other traits such as straightness, form, wood quality, branching, etc., as deemed appropriate through a normal selection index procedure.

1/ Further information, including the selection program and an example, may be obtained from the author. The computer program is in APL.

This procedure does not work as well on small progeny tests containing less than about 1,000 trees after diseased trees are eliminated. Actually, the larger the test (within limits of reasonable statistical design) the better the technique works. For example, 300 to 400 subsample trees seem adequate to array the families in a 20 to 25 line test with ten replications (i.e. lines not eliminated on the basis of disease or size) for specific gravity.

If there is relatively little difference among progeny lines for the principle selection variable, size or value, the 3-P approach is at worst a good scheme for drawing a random sample with probability proportional to size of the individual trees within families. When this is the case, little difference in the number of individuals from the various families is expected in the subsample. Some trial runs with little differences among the family means for volume have shown this to be the case. Of course, the greater the differences among family means for the principle variables, the more biased the subsample becomes towards the larger, more valuable families. If desired, additional bias toward the larger, more valuable families can be obtained by assuring a specific number of "sure to be measured" trees. To do this, simply make KA smaller than K which gives the trees in the range from KA to K a probability of 1 of being sampled. Either of the 3-P systems outlined should reduce the quantification work associated with second generation selection to only a fraction of that required for total enumeration. More important, it is placing the subsample where it will do the most good--on the families containing the larger, more valuable trees.

LITERATURE CITED

Grosenbaugh, L. R. 1963. Some suggestions for better sample-tree measurement. Proc. Soc. Amer. Foresters, p. 36-42.

Rockwood, D. L. and R. E. Goddard. 1973. Predicted gains for fusiform rust resistance in slash pine. These proceedings.