OPTIMUM ALLOCATION AND VARIANCE COMPONENTS IN NESTED SAMPLING WITH AN APPLICATION TO CHEMICAL ANALYSIS

SOPHIE MARCUSE

U. S. Naval Research Laboratory, Washington, D. C.

INTRODUCTION

A SAMPLING TECHNIQUE frequently used in chemical and physical analyses for estimating the mean of a population is that of multiple random subsampling, called nested sampling by P. C. Mahalanobis.[1] For instance, when determining the moisture content of cheese, a food chemist might wish to select his samples randomly from different lots, and again from different cheeses of each lot, and finally make duplicate determinations on each cheese. A primary objective in the statistical design of such a sampling procedure is to minimize the cost of obtaining the sample estimate if the desired degree of precision is fixed, or conversely, to maximize the precision of the estimate obtained from a given amount of expenditure including personnel, time, and equipment. The question arises as to how the number of sampling units at each level should be determined to meet these optimum requirements assuming equal frequencies in the subclasses.

It is assumed in this paper that at each classification level, the cost is proportional to the number of units sampled at this level, and that the cost per sampling unit is known. Thus the total cost is a linear function of the numbers of sampling units at the various levels, with coefficients representing the (known) costs per sampling unit at these levels. On the other hand, the precision of the mean yielded by the experiment can be expressed in terms of the variance of this sample mean; it will then also be a linear function of the variances corresponding to each level, with coefficients involving the reciprocals of the number of units at the various levels. If the variances at the various levels are not known, they should be estimated from a preliminary experiment. The present paper discusses optimum allocation of the sampling units in nested sampling in terms of 3 levels. As an illustration of an experimental situation, a numerical example is given involving the estimation of variance components. In the appendix, the formulas for optimum allocation in nested sampling with k levels are derived.

[1]For reference see M. Ganguli's paper on Nested Sampling [7].

For concreteness, we consider the above mentioned specific problem of planning in the most economical way an experiment in food chemistry designed to determine the moisture content of cheese, the subsampling levels involving lots, cheeses, and determinations. Clearly, the principles elucidated in terms of this particular problem for 3 levels are applicable to a wider class of problems involving more levels in subsampling, as, for instance, by expanding this simplified experiment to more than one factory. Also, they may be applied to other than chemical investigations involving nested sampling, for instance: in the determination of the breaking strength of a certain type of bronze, a metallurgist may wish to choose random samples from different ladles, then again from different molds of each ladle, and make duplicate determinations on the samples from each mold; in a manufacturing process, the subsampling categories may be lots, bags, and batches; in a gunnery experiment, test shooting may be done by different operators taking a number of observations on different runs; in agricultural investigations, the entire area under survey may be subdivided into a large number of zones, these in turn into a large number of smaller zones, and so on; in studies of spray deposit in insect work, plots, trees, and apple samples have been used as subsampling levels [2]. Examples of nested sampling in biological and industrial work together with analyses of variance components may be found in G. W. Snedecor's [10] and L. H. C. Tippett's [12] books. In designing a sample survey for estimating the jute crop in India, P. C. Mahalanobis [9] has used the cost function for considerations of optimum allocation and discussed their general application to large scale sample surveys; principles of optimum allocation in nested sampling have been used by M. H. Hansen et al. [8] in a sample survey of business involving 2-fold nested sampling from finite populations (countries, stores), and by L. H. C. Tippett [12] who describes an experiment where in obtaining soil samples from counts of cysts, a number of "borings" of soil were taken and then several counts made on each boring.

DEFINITION OF NESTED SAMPLING

The problem considered is one in which the total population is subdivided into primary sampling units (lots); these in turn are subdivided into secondary sampling units (cheeses) on which several measurements (determinations) are made representing the tertiary sampling units. The nested sample is obtained by selecting at random first n_1 primary (lots), then n_2 secondary (cheeses), and finally n_3 tertiary sampling units (determinations) from each of the preceding units, where n_1, n_2,

n_3 represent the class frequencies. A measure of the variance of the sample mean in terms of the class frequencies is desired. Before deriving it, the structure of the mathematical model will be explained.

Let x_{hij} denote the j-th determination from the i-th cheese of the h-th lot. Assuming that the effects of the sampling units at the different levels are additive, we may describe an individual observation x_{hij} in nested sampling [7] as:

$$x_{hij} = \mu + \xi_h + \eta_{hi} + \zeta_{hij} \tag{1}$$

$h = 1, 2, \cdots, n_1$ where h refers to the lot of cheese
$i = 1, 2, \cdots, n_2$ where i refers to the cheese in each lot
$j = 1, 2, \cdots, n_3$ where j refers to the determination on each cheese.
The value μ represents the general population mean and is thus a fixed constant. The components ξ_h, η_{hi}, ζ_{hij} are random variables with means and covariances equal to zero and with variances equal to σ_1^2, σ_2^2, σ_3^2, respectively, called variance components. Thus the components ξ_h, η_{hi}, ζ_{hij} represent the effects peculiar to the lots, cheeses, and determinations, and the variance components the variabilities at the different levels.

VARIANCE OF SAMPLE MEAN AND ESTIMATION OF VARIANCE COMPONENTS IN NESTED SAMPLING

From the definition of an individual observation x_{hij} in nested sampling, given by equation (1), we have for the sample mean

$$\bar{x} = \mu + \frac{\sum_{h=1}^{n_1} \xi_h}{n_1} + \frac{\sum_{h=1}^{n_1} \sum_{i=1}^{n_2} \eta_{hi}}{n_1 n_2} + \frac{\sum_{h=1}^{n_1} \sum_{i=1}^{n_2} \sum_{j=1}^{n_3} \zeta_{hij}}{n_1 n_2 n_3} \tag{2}$$

Then because of the assumptions made for the random variables ξ_h, η_{hi}, ζ_{hij} we obtain for the variance of the sample mean

$$\sigma_{\bar{x}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} + \frac{\sigma_3^2}{n_1 n_2 n_3} \tag{3}$$

This expression gives the variance or precision of the sample mean as a linear function of the reciprocals of n_1, $n_1 n_2$, and $n_1 n_2 n_3$ representing the total number of lots, cheeses, and determinations used. The coefficients are the variance components σ_1^2, σ_2^2, σ_3^2, being the variances encountered at the 3 subsampling levels.

As long as the parameter values σ_1^2, σ_2^2, σ_3^2 are unknown, the variance function $\sigma_{\bar{x}}^2$ in (3) cannot be used for solving the problem to determine the optimum values of the class frequencies. On the other hand, if a set of class frequencies were given and used in performing an experiment in nested sampling, then the unknown parameters σ_1^2, σ_2^2, σ_3^2 could be estimated from an analysis of variance of the experimental data. This dilemma[2] may be evaded by first carrying out a preliminary experiment in nested sampling[3] using a set of arbitrarily chosen class

TABLE 1
ANALYSIS OF VARIANCE IN 3-FOLD NESTED SAMPLING

Source of Variation	Degrees of freedom	Mean Square	Expected Mean Square
Primary sampling units	$n_1^* - 1$	MS_1	$\sigma_3^2 + n_3^*\sigma_2^2 + n_3^* n_2^* \sigma_1^2$
Secondary sampling units within primary units	$n_1^*(n_2^* - 1)$	MS_2	$\sigma_3^2 + n_3^*\sigma_2^2$
Tertiary sampling units within secondary units	$n_1^* n_2^*(n_3^* - 1)$	MS_3	σ_3^2

frequencies. We will show how the data obtained from such a preliminary experiment give advance estimates of σ_1^2, σ_2^2, σ_3^2, say s_1^2, s_2^2, s_3^2, to be used for estimating the coefficients of the variance function.

Denote by n_1^*, n_2^*, n_3^* the given class frequencies of the preliminary experiment in nested sampling. Perform a customary analysis of variance on the observed data, as shown in the first 3 columns of table 1, where MS_1, MS_2, and MS_3 denote the mean squares corresponding to the primary, secondary, and tertiary sampling units. It can be shown that the expected values of the mean squares MS_1, MS_2, and MS_3 are the expressions shown in the last column of table 1[4]. Considering the estimates of these expressions by substituting the estimated variance components s_1^2, s_2^2, s_3^2, we obtain the equations

$$
\begin{aligned}
MS_1 &= s_3^2 + n_3^* s_2^2 + n_3^* n_2^* s_1^2 \\
MS_2 &= s_3^2 + n_3^* s_2^2 \qquad (4) \\
MS_3 &= s_3^2
\end{aligned}
$$

[2]See M. Friedman's discussion of a similar situation in planning an experiment ([11], p. 345).

[3]Or a mixed model design of experiment (e.g. randomized blocks or split plot) which includes the subsampling categories under consideration. Note that such a design might involve more degrees of freedom thus increasing the reliability of the estimated variance components ([3], [4]).

[4]Results for any number of sub-samplings and unequal frequencies are given by M. Ganguli [7].

Whence we have the solutions

$$s_3^2 = MS_3$$

$$s_2^2 = \frac{MS_2 - MS_3}{n_3^*} \tag{5}$$

$$s_1^2 = \frac{MS_1 - MS_2}{n_2^* n_3^*}$$

in which the estimated variance components are expressed in terms of the mean squares calculated in the analysis of variance table of the experimental data from nested sampling.[5] These equations can be extended from three to k subsamplings by the same reasoning.

OPTIMUM ALLOCATION IN 3-FOLD NESTED SAMPLING

The variance of the sample mean and the total cost expenditure for determining it, expressed in terms of the class frequencies, are the two functions needed for solving the optimum allocation problem under consideration. Considering the case of 3 levels, let $C(n_1, n_2, n_3)$ be the cost function and $V(n_1, n_2, n_3)$ the variance function, the variables n_1, n_2, n_3 representing the class frequencies. As given by equation (6), the cost function $C(n_1, n_2, n_3)$ is assumed to be an additive function of the costs at the three levels, that is the costs of n_1 primary, n_1n_2 secondary, and $n_1n_2n_3$ tertiary sampling units altogether, the cost per primary, secondary, and tertiary sampling unit being c_1, c_2, and c_3 respectively. The variance function $V(n_1, n_2, n_3)$ is given by equation (3) showing the variance of the sample mean, $\sigma_{\bar{x}}^2$, in 3-fold nested sampling; its parameters may be estimated from the data of a preliminary experiment by the analysis of variance procedure for estimating variance components as described above. Thus we have:

$$C(n_1, n_2, n_3) = c_1n_1 + c_2n_1n_2 + c_3n_1n_2n_3 \tag{6}$$

$$V(n_1, n_2, n_3) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1n_2} + \frac{\sigma_3^2}{n_1n_2n_3} \tag{3}$$

The problem of optimum allocation is to minimize $C(n_1, n_2, n_3)$ by proper choice of n_1, n_2, n_3 subject to the constraint that the allowable

[5]This analysis of the variance components was performed on data from nested sampling, which is a special case of Model II analysis of variance as shown below. If a similar analysis of variance components is routinely carried out on data belonging to Model I, the interpretation differs. In Model II, the computed variance components estimate the variances σ_1^2, σ_2^2, σ_3^2 associated with random factors, whereas in Model I, these are dummy symbols representing sums of squares of differences related to the variation of systematic (or fixed) factors ([1], [5]).

amount of variance is preassigned, say v, or to minimize $V(n_1, n_2, n_3)$ by proper choice of n_1, n_2, n_3 subject to the constraint that the total amount of cost is fixed, say c. Let n_{C1}, n_{C2}, n_{C3} and n_{V1}, n_{V2}, n_{V3} be the optimum solutions of the two problems respectively. By applying Lagrange multipliers it can be shown[6] that these optimum values of n_1, n_2, n_3 are

$$
\begin{aligned}
n_{C1} &= \frac{\sigma_1}{v} \frac{\sum_{i=1}^{3} (\sigma_i \sqrt{c_i})}{\sqrt{c_1}} \\
n_{C2} &= \frac{\sigma_2}{\sigma_1} \sqrt{\frac{c_1}{c_2}} \\
n_{C3} &= \frac{\sigma_3}{\sigma_2} \sqrt{\frac{c_2}{c_3}}
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
n_{V1} &= \frac{\sigma_1}{\sum_{i=1}^{3} (\sigma_i \sqrt{c_i})} \frac{c}{\sqrt{c_1}} \\
n_{V2} &= \frac{\sigma_2}{\sigma_1} \sqrt{\frac{c_1}{c_2}} \\
n_{V3} &= \frac{\sigma_3}{\sigma_2} \sqrt{\frac{c_2}{c_3}}
\end{aligned}
\tag{8}
$$

The sets of equations (7) and (8) show similar features. Except for the first level, the optimum combination of the number of sampling units is independent of the given degree of precision or the fixed total cost, being the same whether the precision or the amount of cost is assigned beforehand. Therefore, when planning an experiment in nested sampling the analyst need be concerned with the given cost or precision only in selecting the number of primary sampling units. Clearly, an increase in funds would be utilized most efficiently, that is resulting in the highest possible precision, by a proportional increase in the number of primary sampling units, and similarly, the most economical way for attaining a higher degree of precision would consist in choosing a correspondingly greater number of primary sampling units.

In many instances, the research analyst might not wish to depend

[6]See appendix for development of these formulas.

on considerations of optimum allocation in the choice of the frequencies at all levels, but might prefer to take, for instance, duplicate or triplicate determinations from each cheese for check purposes, thus preassigning the class frequency associated to the tertiary sampling unit, n_3. If n_3 is prefixed, the corresponding optimum allocation formulas[7] are

$$
\begin{aligned}
n'_{C1} &= \frac{\sigma_1}{v} \frac{\left[\sigma_1 \sqrt{c_1} + \sqrt{\left(\sigma_2^2 + \frac{\sigma_3^2}{n_3}\right)(c_2 + c_3 n_3)}\right]}{\sqrt{c_1}} \\
n'_{C2} &= \frac{\sqrt{\sigma_2^2 + \frac{\sigma_3^2}{n_3}}}{\sigma_1} \sqrt{\frac{c_1}{c_2 + c_3 n_3}}
\end{aligned}
\tag{9}
$$

in the case that the variance v is given; and

$$
\begin{aligned}
n'_{V1} &= \frac{\sigma_1}{\left[\sigma_1 \sqrt{c_1} + \sqrt{\left(\sigma_2^2 + \frac{\sigma_3^2}{n_3}\right)(c_2 + c_3 n_3)}\right]} \frac{c}{\sqrt{c_1}} \\
n'_{V2} &= \frac{\sqrt{\sigma_2^2 + \frac{\sigma_3^2}{n_3}}}{\sigma_1} \sqrt{\frac{c_1}{c_2 + c_3 n_3}}
\end{aligned}
\tag{10}
$$

in the case that the total cost c is given.

NUMERICAL EXAMPLE

The figures shown in table 2 are results from analyses of samples of cheese for the determination of moisture content.[8] They will serve as the preliminary data for obtaining estimates of the variance components. The experimental set-up in nested sampling involves duplicate determinations made on 2 cheeses from each of 3 lots, the different cheeses and the different lots being randomly selected ($n_1^* = 3$, $n_2^* = 2$, $n_3^* = 2$).

The first 4 columns of table 3 show the results of an analysis of variance of these data. In nested sampling the sums of squares may be calculated as follows: Consider first table 2 (in which there are 3 factors: duplicates, cheeses, and lots) and refer to the figures, representing 1 determination, as "totals." Subsequently, obtain the totals

---

[7]See appendix for development of formulas in which all but the first k' are fixed.

[8]The data are drawn from "Report on Sampling Fat and Moisture in Cheese" by William Horwitz and Lila F. Knudsen, J. Ass. Off. Agr. Chem., vol. 31 (1948), pp. 300-306; slight modifications have been made for illustrative purposes. The author acknowledges the suggestions of Lila F. Knudsen;

TABLE 2
MOISTURE CONTENT OF 2 CHEESES FROM EACH OF 3 DIFFERENT LOTS, DETERMINED 2 TIMES

Cheese	Lot I	Lot II	Lot III
1	39.02 38.79	35.74 35.41	37.02 36.00
2	38.96 39.01	35.58 35.52	35.70 36.04

for the duplicates on each cheese (there remain 2 factors: cheeses and lots), and also the totals of the 4 determinations on each lot (there remains 1 factor: lots), in addition to the total for the entire table (no

TABLE 3
ANALYSIS OF VARIANCE OF DATA ON MOISTURE CONTENT OF CHEESE GIVEN IN TABLE 2

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square	Estimated Variance Components
Lots	2	$SS_1 = 25.9001$	$MS_1 = 12.9501$	$\sigma_3^2 + 2\sigma_2^2 + 4\sigma_1^2$	$s_1^2 = 3.2028$
Cheeses within lots	3	$SS_2 = .4166$	$MS_2 = .1389$	$\sigma_3^2 + 2\sigma_2^2$	$s_2^2 = .0143$
Determinations within cheeses	6	$SS_3 = .6620$	$MS_3 = .1103$	σ_3^2	$s_3^2 = .1103$

factor remains). Denote by Q_3, Q_2, Q_1, and Q_0 the sum of squares of these corresponding totals divided by the number of determinations making up each total:

$$Q_3 = 39.02^2 + 38.79^2 + \cdots + 35.70^2 + 36.04^2 = 16{,}365.5607$$

$$Q_2 = \frac{77.81^2 + 77.97^2 + 71.15^2 + 71.10^2 + 73.02^2 + 71.74^2}{2}$$

$$= 16{,}364.8988$$

$$Q_1 = \frac{155.78^2 + 142.25^2 + 144.76^2}{4} = 16{,}364.4821$$

$$Q_0 = \frac{442.79^2}{12} = 16{,}338.5820$$

Then the sums of squares in analysis of variance, SS_1, SS_2, SS_3, are the successive differences of these expressions:

$$SS_1 = Q_1 - Q_0 = 25.9001$$

$$SS_2 = Q_2 - Q_1 = 0.4166$$

$$SS_3 = Q_3 - Q_2 = 0.6620^{9}$$

The sums of squares and the corresponding mean squares are shown in columns 3 and 4 of table 3. The estimated variance components s_1^2, s_2^2, s_3^2, shown in the last column of table 3, follow from equations (5). These values represent the advance estimates from the preliminary data to be used in the planning of the experiment.

The problem of designing an experiment with optimum allocation may arise in chemical laboratory work, e.g., when it is desired to set up in the most economical way routine analyses of samples of cheese for the determination of moisture content. In the example under consideration we assume that the chemist wants to spend not more than 60 dollars altogether to be allocated in such a way that the highest precision results; that he requires duplicate determinations for check purposes; and that the cost factors per lot, cheese, and determination are 10, 3, and 1 dollar respectively. Since these requirements prefix the class frequency n_3 and the total cost C, formulas (10) are appropriate. Substituting $n_3 = 2$, $c = 60$, $c_1 = 10$, $c_2 = 3$, and $c_3 = 1$, and for the variances σ_1^2, σ_2^2, σ_3^2 their estimates $s_1^2 = 3.2028$, $s_2^2 = 0.0143$, $s_3^2 = 0.1103$, we obtain:

$$n'_{V1} = 5.43 \qquad n'_{V2} = 0.21$$

The corresponding integer values have to be chosen in accordance with the conditions of the experiment. Since n_2, the number of cheeses selected from each lot, must be at least one, the number of lots, n_1, may be reduced. An examination of the integers smaller than n'_{V1} shows that $n_1 = 4$ together with $n_2 = 1$ fulfill the required conditions. Thus 4 lots and 1 cheese give the optimum solution for the problem under consideration.

The merit of this optimum combination may be judged by comparing it to other combinations of class frequencies. In table 4 a number of various combinations (columns 1 and 2) are presented together with the precision of the sample mean (columns 5 and 6) and

[9] Using the figures given for Q_2, Q_3 above, we have $Q_3 - Q_2 = .6619$ instead of .6620. Such a difference in the last decimal place is due to rounding off results, intermediate computations being carried out to more decimal places.

TABLE 4

ESTIMATED PRECISION AND COST OF DETERMINING MOISTURE CONTENT OF CHEESE WHEN A SPECIFIED NUMBER OF LOTS (n_1) AND A SPECIFIED NUMBER OF CHEESES FROM EACH LOT (n_2) ARE USED AND TWO DETERMINATIONS ($n_3 = 2$) ARE MADE ON EACH CHEESE. CONSTANTS USED ARE ADVANCE ESTIMATES CALCULATED FROM PRELIMINARY DATA (TABLES 2 AND 3).

Formulas used:	Constants used:
$N = n_1 n_2 n_3$	$n_3 = 2$
$C = c_1 n_1 + c_2 n_1 n_2 + c_3 n_1 n_2 n_3$	$c_1 = 10, c_2 = 3, c_3 = 1$
$V = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} + \frac{s_3^2}{n_1 n_2 n_3}$	$s_1^2 = 3.2028, s_2^2 = .0143, s_3^2 = .1103$
$CV = \frac{\sqrt{v}}{\bar{x}} \times 100$	$\bar{x} = 36.90$

Number of—		Expenditure		Estimated Precision	
Lots n_1 (1)	Cheeses n_2 (2)	Number of Determinations N (3)	Total Cost in dollars C (4)	Variance of mean V (5)	Coefficient of Variation CV (6)
5	3	30	125	0.6452	2.18
5	2	20	100	0.6475	2.18
5	1	10	75	0.6544	2.19
4	3	24	100	0.8065	2.43
4	2	16	80	0.8094	2.44
4	1	8	60	0.8181	2.45
3	3	18	75	1.0753	2.81
3	2	12	60	1.0792	2.82
3	1	6	45	1.0907	2.83
2	3	12	50	1.6130	3.44
2	2	8	40	1.6188	3.45
2	1	4	30	1.6361	3.47
1	3	6	25	3.2259	4.87
1	2	4	20	3.2375	4.88
1	1	2	15	3.2722	4.90

the expenditure involved in determining it (columns 3 and 4). Column 3 shows the total number of determinations made, the total cost is given in column 4, and column 6 compares the relative precision of the sample mean, indicated by its coefficient of variation, to the absolute precision in terms of the variance (column 5). Duplicate determinations are used throughout. It can be seen that the 4-1-2 combination is more economical than the 3-2-2 combination—the one used in the preliminary experiment—since it obtains a higher precision but requires the same cost (60 dollars). Also, the combination 3-2-2 is less efficient than the combination 3-1-2 since, for the same precision, the latter combination needs half the number of determinations and requires only 45 dollars instead of 60 dollars. In general, it pays to increase the number of lots instead of the number of cheeses since the former are more variable.

REMARKS ON NESTED SAMPLING AS A SPECIAL CASE OF MODEL II ANALYSIS OF VARIANCE

The mathematical model of nested sampling as given by the fundamental equation (1) and its assumptions, is closely related to one specific mathematical model used in analysis of variance. Two models of analysis of variance, usually referred to as Model I and Model II, have been discussed recently by S. L. Crump [3] and C. Eisenhart [5]. It seems worthwhile to show that, in virtue of the underlying assumptions, nested sampling represents a special case of Model II of analysis of variance.

The two different models of analysis of variance involve the analysis of two different types of factors: systematic factors in Model I and random factors in Model II. A factor such as "treatment" or "lot" is a random or a systematic factor depending on the way its variants are chosen. Here the term "variant" of a factor is used based on Fisher's terminology [6], for instance, the variants of the factor "treatment" may be e.g. "nitrogen" and "phosphate" and different lots the variants of the factor "lot." When an experimenter selects the two treatments "nitrogen" and "phosphate," he selects them systematically from a population of possible treatments on the basis of subject matter judgment; on the other hand, when selecting different lots of material for studying the effects of the treatments, he generally bases his choice on random selection ([5], [10] Chapter 8). Since systematically chosen variants produce systematic variation and randomly chosen variants random variation, the type of factor may be determined according to the issue: systematic or random variation. Usually, "methods" and "treatments" represent systematic factors, "blocks" and "lots" random factors, whereas factors such as "days" or "animals" or "locations" may represent either systematic or random factors; both types of factor will often occur in the same experiment; then the model is a mixed one.

Now the factors encountered in nested sampling are the primary, secondary, tertiary sampling units (lots, cheeses, determinations). Under the assumptions made, the variants of these factors, i.e. the units selected at each level, were chosen randomly. These factors, therefore, are random factors and thus nested sampling belongs to Model II.

In order to describe more accurately the relationship of nested sampling to Model II of analysis of variance, we subdivide the random factors of Model II into two categories: cross classified[10] with respect to another factor or not. For instance, in the 2 factor "day-animal" experiment discussed by C. Eisenhart [5] as an example of Model II, the random factor "animal" is cross classified with respect to the factor "days," each of the randomly chosen animals being tested on all days (the analysis of variance table contains: "Between days", "Between animals," and "Residual" with $d - 1$, and $a - 1$, and $(a - 1)(d - 1)$ degrees of freedom respectively). On the other hand, there would be no cross classification, if on each day a number of animals were randomly chosen for testing, as for instance in an inoculation experiment affecting the sensitivity of the animal (the analysis of variance contains: "Between days," and "Between animals within days" with $d - 1$, and $d(a - 1)$ degrees of freedom respectively). Likewise, no cross classification would be involved for the random factor "animal" if each animal would be tested on a couple of days which were randomly selected, as e.g. if only one animal could be tested per day (the analysis of variance contains: "Between animals," and "Between days within animals" with $(a - 1)$, and $a(d - 1)$ degrees of freedom respectively). Nested sampling represents the second category of Model II in which the random factors involved are not cross classified since for each primary sampling unit a number of secondary sampling units is selected randomly, and so on. The question as to which order of subsampling should be adopted in the nested sampling procedure, as, for instance, whether to use "animals" as primary sampling units and "days" as secondary sampling units, or conversely, is a decision to be made on the basis of subject matter judgment.

APPENDIX

We shall now derive the optimum values of the class frequencies, given for the three-fold level by formulas (7), (8), (9), and (10), for the general case of k-fold nested sampling. Instead of solving the prob-

[10]This term is not synonymous with "ordered". Note that items in table 2 below are ordered for purely designative reasons there being neither a cross classification nor an element of "sequence" involved.

lem directly by introducing the Lagrange multiplier, we will apply this procedure to a pair of generalized functions. We then obtain as special cases the solution formulas for optimum allocation in

i. k-fold nested sampling
ii. k-fold nested sampling in which some class frequencies are fixed beforehand
iii. stratified sampling from finite population (k strata, 2 levels).

a. *Minimum Problem for 2 Generalized Functions*

Let the two generalized functions be

$$F_1(N_1, \cdots, N_k) = \sum_{i=1}^{k} a_{1i}N_i + a_1 \tag{11}$$

$$F_2(N_1, \cdots, N_k) = \sum_{i=1}^{k} \frac{a_{2i}}{N_i} + a_2 \tag{12}$$

where $N_1, \cdots, N_k$ denote variables and a_1, a_2, and a_{1i}, a_{2i} ($i = 1, \cdots, k$) are constants.

Consider first the problem to minimize $F_1(N_1, \cdots, N_k)$ subject to the side condition

$$F_2(N_1, \cdots, N_k) = b_2 \tag{13}$$

where b_2 is a constant. Using the Lagrange multiplier λ in the usual way, we let the derivatives of $F_1 + \lambda F_2$ with respect to N_i ($i = 1, \cdots, k$) be zero, and obtain

$$a_{1i} - (\lambda a_{2i}/N_i^2) = 0$$

or

$$N_i = \sqrt{\lambda}\sqrt{a_{2i}/a_{1i}}$$

Substituting these values of N_i in (13), where F_2 is given by (12), we have

$$F_2(N_1, \cdots, N_k) = (1/\sqrt{\lambda}) \sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}} + a_2 = b_2$$

Therefore

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2}$$

Hence we obtain the optimum values

$$N_{1i} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2} \sqrt{\frac{a_{2i}}{a_{1i}}} \tag{14}$$

Similarly, we obtain the solution of the problem to minimize $F_2(N_1, \cdots, N_k)$ subject to the side condition

$$F_1(N_1, \cdots, N_k) = b_1 \tag{15}$$

where b_1 is a constant:

$$N_{2i} = \frac{b_1 - a_1}{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}} \sqrt{\frac{a_{2i}}{a_{1i}}} \tag{16}$$

Now introduce the variables

$$n_1 = N_1, \qquad n_i = N_i/N_{i-1} \qquad (i = 2, \cdots, k) \tag{17}$$

then $N_i = n_1 \cdots n_i (i = 1, \cdots, k)$. Substituting the new variables in (11) and (12), we obtain the functions

$$f_1(n_1, \cdots, n_k) = \sum_{i=1}^{k} a_{1i}n_1 \cdots n_i + a_1 \tag{18}$$

$$f_2(n_1, \cdots, n_k) = \sum_{i=1}^{k} \frac{a_{2i}}{n_1 \cdots n_i} + a_2 \tag{19}$$

Substituting (14) in (17), we find that the minimum solutions of $f_1(n_1, \cdots, n_k)$ under the side condition $f_2(n_1, \cdots, n_k) = b_2$ are:

$$n_{11} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2} \sqrt{\frac{a_{21}}{a_{11}}}$$

and (20)

$$n_{1i} = \sqrt{\frac{a_{2i}a_{1,i-1}}{a_{1i}a_{2,i-1}}} \qquad (i = 2, \cdots, k)$$

Similarly, substituting (16) in (17), we find the minimum solutions of $f_2(n_1, \cdots, n_k)$ under the side condition $f_1(n_1, \cdots, n_k) = b_1$:

$$n_{21} = \frac{b_1 - a_1}{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}} \sqrt{\frac{a_{21}}{a_{11}}}$$

and (21)

$$n_{2i} = \sqrt{\frac{a_{2i}a_{1,i-1}}{a_{1i}a_{2,i-1}}} \qquad (i = 2, \cdots, k)$$

Note that $n_{1i} = n_{2i} (i = 2, \cdots, k)$.

b. *Application to Optimum Allocation Problems in Sampling*

i. *Nested Sampling*

Substituting $a_{1i} = c_i$, $a_{2i} = \sigma_i^2$ and $a_1 = a_2 = 0$ in (18) and (19), we obtain the 2 functions

$$g_1(n_1, \cdots, n_k) = \sum_{i=1}^{k} c_i n_1 \cdots n_i \tag{22}$$

$$g_2(n_1, \cdots, n_k) = \sum_{i=1}^{k} \frac{\sigma_i^2}{n_1 \cdots n_i} \tag{23}$$

These functions represent the general case of the cost function $C(n_1, n_2, n_3)$ and the variance function $V(n_1, n_2, n_3)$ used above in section 4. Setting $b_1 = c$ and $b_2 = v$ yields the corresponding side conditions. Therefore applying formulas (20) and (21), we have as the minimum solutions of $g_1(n_1, \cdots, n_k)$ under the side condition $g_2(n_1, \cdots, n_k) = v$

$$n_{11} = \frac{\sigma_1}{v} \frac{\sum_{i=1}^{k} (\sigma_i \sqrt{c_i})}{\sqrt{c_1}}$$

and (24)

$$n_{1i} = \frac{\sigma_i}{\sigma_{i-1}} \sqrt{\frac{c_{i-1}}{c_i}} \qquad (i = 2, \cdots, k)$$

and as the minimum solutions of $g_2(n_1, \cdots, n_k)$ under the side condition $g_1(n_1, \cdots, n_k) = c$

$$n_{21} = \frac{\sigma_1}{\sum_{i=1}^{k} (\sigma_i \sqrt{c_i})} \frac{c}{\sqrt{c_1}}$$

and (25)

$$n_{2i} = \frac{\sigma_i}{\sigma_{i-1}} \sqrt{\frac{c_{i-1}}{c_i}} \qquad (i = 2, \cdots, k)$$

Specializing equations (24) and (25) to the case $k = 3$ yields equations (7) and (8). Specializing equation (25) to the case $k = 2$ and letting cost be expressed in terms of time, $c_1 = kt$, $c_2 = t$, gives equation 10.32 in L. H. C. Tippett's book [12].

ii. *Nested Sampling with Some Prefixed Class Frequencies*

Let $n_1', \cdots, n_{k'}'$ be the unknown frequencies and $n_{k'+1}, \cdots, n_k$ be fixed beforehand. The equations (22) and (23) may then be rewritten in terms of $n'_1, \cdots, n'_{k'}$ as follows:

$$h_1(n'_1, \cdots, n'_{k'}) = \sum_{j=1}^{k'} c_j n'_1 \cdots n'_j + n'_1 \cdots n'_{k'} \sum_{l=1}^{k-k'} c_{k'+l} n_{k'+1} \cdots n_{k'+l} \tag{26}$$

$$= \sum_{j=1}^{k'} c'_j n'_1 \cdots n'_j$$

where $c'_j = c_j \; (j = 1, \cdots, k' - 1)$

and

$$c'_{k'} = c_{k'} + \sum_{l=1}^{k-k'} c_{k'+l} n_{k'+1} \cdots n_{k'+l} \tag{27}$$

$$h_2(n'_1, \cdots, n'_{k'}) = \sum_{j=1}^{k'} \frac{\sigma_j^2}{n'_1 \cdots n'_j} + \frac{1}{n'_1 \cdots n'_{k'}} \sum_{l=1}^{k-k'} \frac{\sigma_{k'+l}^2}{n_{k'+1} \cdots n_{k'+l}} \tag{28}$$

$$= \sum_{j=1}^{k'} \frac{\sigma_j'^2}{n'_1 \cdots n'_j}$$

where $\sigma'_j = \sigma_j \; (j = 1, \cdots, k' - 1)$

and

$$\sigma_{k'}'^2 = \sigma_{k'}^2 + \sum_{l=1}^{k-k'} \frac{\sigma_{k'+l}^2}{n_{k'+1} \cdots n_{k'+l}} \tag{29}$$

Thus the functions h_1 and h_2 of the variables $n'_1, \cdots, n'_{k'}$, given by (26) and (28), represent the same types of function as the functions g_1 and g_2 of the variables $n_1, \cdots, n_k$ given by (22) and (23). Therefore the minimum solutions of $h_1(n'_1, \cdots, n'_{k'})$ and $h_2(n'_1, \cdots, n'_{k'})$ under the side conditions $h_2(n'_1, \cdots, n'_{k'}) = v$ and $h_1(n'_1, \cdots, n'_{k'}) = c$ respectively, may be obtained from equations (24) and (25) by replacing k by k', σ by σ', and c by c', and then substituting back σ'_j and c'_j $(j = 1, \cdots, k')$ from equations (27) and (29).

For $k = 3$, $k' = 2$ we obtain from (27) and (29)

$$c'_1 = c_1 \qquad c'_2 = c_2 + c_3 n_3$$

$$\sigma'_1 = \sigma_1 \qquad \sigma_2'^2 = \sigma_2^2 + \frac{\sigma_3^2}{n_3}$$

The substitution of these values into (24) and (25) after replacement of k, c, σ by k', c', σ' gives the formulas (9) and (10) used above.

Note that the results of b. ii. may also be obtained from a. and then b. i. be considered as the special case $k' = k$.

iii. *Stratified Sampling from Finite Populations*

We will indicate briefly the applicability of the above used generalized functions to stratified sampling involving two levels.

Let there be k strata in the population with M_i elements x_{ij} in the i-th stratum ($i = 1, \cdots, k$; $j = 1, \cdots, M_i$). Assume that the N_i sample elements x_{ij} ($i = 1, \cdots, k$; $j = 1, \cdots, N_i$) are independently drawn at random from the k finite strata. Then the sample mean

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{k} M_i \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i}$$

has the variance

$$\sigma_{\bar{x}}^2 = \frac{1}{M^2} \sum_{i=1}^{k} M_i^2 \frac{\sigma_i^2}{N_i} \frac{M_i - N_i}{M_i - 1}$$

where $M = \sum_{i=1}^{k} M_i$ and σ_i^2 denotes the variance between elements in the i-th stratum. Thus we have

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{k} \frac{a_{2i}}{N_i} + a_2$$

where $\quad a_{2i} = \dfrac{M_i^3 \sigma_i^2}{M^2(M_i - 1)} \quad$ and $\quad a_2 = -\dfrac{1}{M^2} \sum_{i=1}^{k} \dfrac{M_i^2 \sigma_i^2}{M_i - 1}$

Let c_i be the cost per element in the i-th stratum and $c = \sum_{i=1}^{k} c_i N_i$ the total cost, then c may be written $c = \sum_{i=1}^{k} a_{1i} N_i + a_1$ where $a_{1i} = c_i$ and $a_1 = 0$. Thus c and $\sigma_{\bar{x}}^2$ correspond to the functions $F_1(N_1, \cdots, N_k)$ and $F_2(N_1, \cdots, N_k)$ respectively in (11) and (12). Therefore equations (14) and (16) give the desired minimum solutions where b_1 and b_2 determine the side conditions corresponding to (13) and (15). In case the populations in the strata are large ($M_i \sim M_i - 1$), we obtain the well known optimum allocation formulas:

$$N_{1i} = \frac{\sum_{i=1}^{k} (M_i \sigma_i \sqrt{c_i})}{M^2 b_2 + \sum_{i=1}^{k} (M_i \sigma_i^2)} \frac{M_i \sigma_i}{\sqrt{c_i}}$$

$$N_{2i} = \frac{b_1}{\sum_{i=1}^{k} M_i \sigma_i \sqrt{c_i}} \frac{M_i \sigma_i}{\sqrt{c_i}}$$

LITERATURE CITED

[1] Anderson, R. L. Use of Variance Components in the Analysis of Hog Prices in Two Markets, *J. Am. Stat. Ass.*, 42: 612–634, 1947.

[2] Cassil, C. C., Wadley, F. M., and Dean, F. P. Sampling Studies on Orchard Spray Residues in the Pacific Northwest, *J. of Econ. Entom.*, 36: 227–231, 1943.

[3] Crump, S. Lee. The Estimation of Variance Components in Analysis of Variance, *Biometrics Bulletin*, 2: 7–11, 1946.

[4] Daniels, H. E. The Estimation of Components of Variance, *Supplement to the Journal of the Royal Statistical Society*, 6: 186–197, 1939.

[5] Eisenhart, Churchill. The Assumptions Underlying the Analysis of Variance, *Biometrics Bulletin*, 3: 1–21, 1947.

[6] Fisher, R. A. *The Design of Experiments*, 3rd Edition. Oliver and Boyd, Ltd., Edinburgh and London, 1942.

[7] Ganguli, M. A Note on Nested Sampling, *Sankhya*, 5: 449–452, 1941.

[8] Hansen, M. H., Hurwitz, W. N., and Gurney, M. Problems and Methods in a Sample Survey of Business, *J. Am. Stat. Ass.*, 41: 173–189, 1946.

[9] Mahalanobis, P. C. On Large Scale Sample Surveys, *Philos, Transactions of the Royal Society, Series B, Biolog. Sciences*, 23: 329–451, 1944.

[10] Snedecor, G. W. *Statistical Methods Applied to Experiments in Agriculture and Biology*, 4th Edition. The Collegiate Press, Inc., Ames, Iowa, 1946.

[11] Statistical Research Group, Columbia University. *Selected Techniques of Statistical Analysis: For Scientific and Industrial Research and Production and Management Engineering*. McGraw-Hill Book Company, Inc., New York, 1947.

[12] Tippett, L. H. C. *The Methods in Statistics*, 3rd Edition. Williams and Norgate, Ltd., London, 1940.