VARIANCE COMPONENTS IN THE UNBALANCED 2-WAY NESTED CLASSIFICATION

By S. R. Searle

New Zealand Dairy Board, Wellington, New Zealand

Introduction. Sampling variances of estimates of components of variance obtained from data that are balanced (having the same number of observations in all subclasses) are easily derived because the mean squares in the analysis of variance are independent and distributed as χ^2. The variance component estimates are linear functions of the mean squares and their variances can be derived accordingly, although their distributions are, in general, unknown. When the data are not balanced, however, and there are unequal numbers of observations in the subclasses the mean squares are no longer independent and they do not have χ^2-distributions. Methods of deriving expressions for the sampling variances of the variance component estimates are developed for these situations in an earlier paper [3] and applied to the 1-way classification. A second paper [4] gives these expressions for the 2-way factorial classification, and extension to the 2-way hierarchical (nested) classification is presented here.

Model and analysis of variance. The earlier work discussed sampling variances of variance component estimates obtained by Henderson's Method 1 [2] from data having unequal subclass numbers, based on the completely random model, namely Eisenhart's Model II, [1]. The same situation is considered here for the 2-way nested classification.

The linear model for an observation x_{ijk} is taken as

$$x_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$$

where μ is the general mean, α_i is the effect due to the ith main classification, β_{ij} is the effect due to the jth sub-class within the ith main classification, and e_{ijk} is the residual error term peculiar to x_{ijk}. We suppose the number of classes in the main classification is a, so that $i = 1, \cdots, a$; and that there are c_i sub-classes within each of these so that $j = 1, \cdots, c_i$. The total number of such sub-classes will be represented by b, giving $b = \sum_{i=1}^{a} c_i$. The number of observations in the jth subclass of the ith class is taken as n_{ij}. All terms of the model (except μ) are assumed to be normally distributed random variables with zero means and variances σ_α^2, σ_β^2 and σ_e^2. These are the variance components to be estimated, along with the sampling variances of their estimates.

Received December 4, 1960; revised June 30, 1961.

The customary analysis of variance can be written as

Analysis of Variance

Term	d.f.	Sums of Squares
Between main classes	$a - 1$	$T_a - T_f$
Between subclasses within main classes	$b - a$	$T_{ab} - T_a$
Within subclasses	$N - b$	$T_o - T_{ab}$
Total	$N - 1$	$T_o - T_f$

where, with the customary notation for totals and means, namely

$$x_{i..} = \sum_j \sum_k x_{ijk}\,, \qquad n_{i.} = \sum_j n_{ij} \qquad \text{and} \qquad \bar{x}_{i..} = x_{i..}/n_{i.}$$

we have the uncorrected sums of squares

$$T_a = \sum_i n_{i.}\bar{x}^2_{i..}\,,$$

$$T_{ab} = \sum_i \sum_j n_{ij}\bar{x}^2_{ij.}\,,$$

$$T_o = \sum_i \sum_j \sum_k x^2_{ijk}$$

and

$$T_f = N\bar{x}^2_{...}\,,$$

N being the total number of observations, $N = \sum_i \sum_j n_{ij}$.

The variance components can be estimated by equating each line of the above analysis of variance (except that for "total") to its expected value. Denoting the resulting estimates as $\hat{\sigma}^2_\alpha$, $\hat{\sigma}^2_\beta$ and $\hat{\sigma}^2_e$, the equations for obtaining them are, as given in Section 10.17 of [5]

$$\begin{aligned} T_a - T_f &= (N - k_1)\hat{\sigma}^2_\alpha + (k_{12} - k_3)\hat{\sigma}^2_\beta + (a - 1)\hat{\sigma}^2_e \\ (1)\qquad T_{ab} - T_a &= (N - k_{12})\hat{\sigma}^2_\beta + (b - a)\hat{\sigma}^2_e \\ T_o - T_{ab} &= (N - b)\hat{\sigma}^2_e\,. \end{aligned}$$

The k's are functions of the n_{ij}'s, namely

$$k_1 = \sum_i n^2_{i.}/N$$

$$k_3 = \sum_i \sum_j n^2_{ij}/N$$

and

$$k_{12} = \sum_i \left(\sum_j n^2_{ij}\right)/n_{i..}$$

The notation here follows that used previously in [4].

Variances and covariances required. The within sub-classes sum of squares, $T_o - T_{ab}$, has a χ^2-distribution with $(N - b)$ degrees of freedom and hence the variance of $\hat{\sigma}_e^2$ is

$$\text{var}\,(\hat{\sigma}_e^2) = 2\sigma_e^4/(N - b). \tag{2}$$

Furthermore, $T_o - T_{ab}$ is distributed independently of T_a, T_{ab} and T_f so that the covariances of $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$ with $\hat{\sigma}_e^2$ are obtained directly as

$$\text{cov}\,(\hat{\sigma}_\beta^2, \hat{\sigma}_e^2) = -(b - a)\,\text{var}\,(\hat{\sigma}_e^2)/(N - k_{12}) \tag{3}$$

and

$$\begin{aligned}\text{cov}\,(\hat{\sigma}_\alpha^2, \hat{\sigma}_e^2) &= -[(k_{12} - k_3)\,\text{cov}\,(\hat{\sigma}_\beta^2, \hat{\sigma}_e^2) + (a - 1)\,\text{var}\,(\hat{\sigma}_e^2)]/(N - k_1) \\ &= [(k_{12} - k_3)(b - a)/(N - k_{12}) - (a - 1)]\,\text{var}\,(\hat{\sigma}_e^2)/(N - k_1). \qquad (4)\end{aligned}$$

This independence property is also used for obtaining the variances of $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$ and the covariance between them as linear functions of var $(\hat{\sigma}_e^2)$ and the variances and covariances of T_a, T_{ab} and T_f. Thus

$$\text{var}(\hat{\sigma}_\beta^2) = \frac{\text{var}(T_{ab} - T_a) + (b - a)^2\,\text{var}\,(\hat{\sigma}_e^2)}{(N - k_{12})^2} \tag{5}$$

and

$$\begin{aligned}&(N - k_1)^2(N - k_{12})^2\,\text{var}\,(\hat{\sigma}_\alpha^2) \\ &\quad = \text{var}\,[(N - k_3)T_a - (k_{12} - k_3)T_{ab} - (N - k_{12})T_f] \\ &\qquad + [(N - k_3)a - (k_{12} - k_3)b - (N - k_{12})]^2\,\text{var}\,(\hat{\sigma}_e^2) \qquad (6)\end{aligned}$$

and

$$\begin{aligned}&(N - k_1)(N - k_{12})\,\text{cov}\,(\hat{\sigma}_\alpha^2, \hat{\sigma}_\beta^2) = \\ &\quad \text{cov}\,(T_a - T_f)(T_{ab} - T_a) + (a - 1)(b - a)\,\text{var}\,(\hat{\sigma}_e^2) \qquad (7) \\ &\qquad - (N - k_{12})(k_{12} - k_3)\,\text{var}\,(\hat{\sigma}_\beta^2).\end{aligned}$$

The second term in each of these expressions can be obtained from equation (2) and the first can be found as a linear function of the variances and covariances of T_a, T_{ab} and T_f. These we now proceed to find.

Matrix methods. The sampling variance of a quadratic function, $\mathbf{x}'F\mathbf{x}$, of normally-distributed random variables represented by the vector $\mathbf{x}$ is $2\text{tr}(VF)^2$ where V is the variance-covariance matrix appropriate to the variables in $\mathbf{x}$. The covariance between two quadratics $\mathbf{x}'F\mathbf{x}$ and $\mathbf{x}'G\mathbf{x}$ is $2\text{tr}(\text{VFVG})$. These results can be applied to obtain the terms needed for equations (5) through (7) using matrices similar to those employed in Searle, [4]. First we define square matrices U_{ij}, $U_{i\cdot}$ and U_N of order n_{ij}, $n_{i\cdot}$ and N respectively, with all elements equal to one. Square matrices of order N with U-matrices in the diagonal and zeros elsewhere are defined as D-matrices; thus D_{ab} has the matrices U_{ij} in its

diagonal, for all values of i and j, and D_a has the matrices $U_{i\cdot}$ in its diagonal, for all values of i. The variance-covariance matrix of the N observations arrayed in order $k = 1 \cdots n_{ij}$ within j-classes within each i-class can now be expressed as

$$V = \sigma_\alpha^2 D_a + \sigma_\beta^2 D_{ab} + \sigma_e^2 I, \tag{8}$$

I being an identity matrix.

Defining C_{ab} and C_a similar to D_{ab} and D_a only with matrices $(n_{ij})^{-1}U_{ij}$ and $(n_{i\cdot})^{-1}U_{i\cdot}$ in the diagonal enables the quadratics in the analysis of variance to be written as

$$T_a = \mathbf{x}'C_a\mathbf{x}$$

$$T_{ab} = \mathbf{x}'C_{ab}\mathbf{x}$$

and

$$T_f = \mathbf{x}'U_N\mathbf{x}.$$

Thus

$$\begin{aligned} \text{var}\,(T_a) &= 2\,\text{tr}\,(VC_a)^2 \\ &= 2\,\text{tr}\,(\sigma_\alpha^2 D_a + \sigma_\beta^2 D_{ab}C_a + \sigma_e^2 C_a)^2 \end{aligned} \tag{9}$$

after substitution from (8). This is a quadratic in the variance components which can be expanded, through the special form of the matrices, in terms of the n_{ij}'s using the expressions

$$k_4 = \sum_i \sum_j n_{ij}^3 \qquad k_5 = \sum_i \left(\sum_j n_{ij}^3\right)/n_{i\cdot}$$

$$k_6 = \sum_i \left(\sum_j n_{ij}^2\right)^2/n_{i\cdot} \qquad k_7 = \sum_i \left(\sum_j n_{ij}^2\right)^2/n_{i\cdot}^2$$

$$k_8 = \sum_i n_{i\cdot}\left(\sum_j n_{ij}^2\right) \qquad k_9 = \sum_i n_{i\cdot}^3$$

Thus

$$\text{var}\,(T_a) = 2(Nk_1\sigma_\alpha^4 + k_7\sigma_\beta^4 + a\sigma_e^4 + 2Nk_3\sigma_\alpha^2\sigma_\beta^2 + 2N\sigma_\alpha^2\sigma_e^2 + 2k_{12}\sigma_\beta^2\sigma_e^2).$$

A similar procedure for the other terms in (5), (6) and (7) leads to the following results:

$$\begin{aligned} \text{var}\,(T_{ab}) &= 2\,\text{tr}\,(VC_{ab})^2 \\ &= \text{var}\,(T_a) + 2[(Nk_3 - k_7)\sigma_\beta^4 + (b-a)\sigma_e^4 + 2(N - k_{12})\sigma_\beta^2\sigma_e^2] \\ \text{var}\,(T_f) &= 2\,\text{tr}\,(VU_N)^2/N^2 \\ &= 2(k_1\sigma_\alpha^2 + k_3\sigma_\beta^2 + \sigma_e^2)^2 \\ \text{cov}\,(T_aT_{ab}) &= 2\,\text{tr}\,(VC_aVC_{ab}) \\ &= \text{var}\,(T_a) + 2(k_5 - k_7)\sigma_\beta^4 \end{aligned}$$

$$\begin{aligned}\operatorname{cov}(T_aT_f) &= 2\ \operatorname{tr}(VC_aVU_N)^2/N \\ &= 2[(k_9/N)\sigma_\alpha^4 + (k_6/N)\sigma_\beta^4 + \sigma_e^4 \\ &\quad + 2(k_8/N)\sigma_\alpha^2\sigma_\beta^2 + 2k_1\sigma_\alpha^2\sigma_e^2 + 2k_3\sigma_\beta^2\sigma_e^2] \\ \operatorname{cov}(T_{ab}T_f) &= 2\ \operatorname{tr}(VC_{ab}VU_N)/N \\ &= \operatorname{cov}(T_aT_f) + 2\sigma_\beta^4(k_4 - k_6)/N.\end{aligned}$$

Results. Substituting the above expressions into (5) leads, after simplification, to

$$\operatorname{var}(\hat{\sigma}_\alpha^2) = \frac{2(\lambda_1\sigma_\alpha^4 + \lambda_2\sigma_\beta^4 + \lambda_3\sigma_e^4 + 2\lambda_4\sigma_\alpha^2\sigma_\beta^2 + 2\lambda_5\sigma_\alpha^2\sigma_e^2 + 2\lambda_6\sigma_\beta^2\sigma_e^2)}{(N - k_1)^2(N - k_{12})^2}$$

where

$$\begin{aligned}\lambda_1 &= (N - k_{12})^2[k_1(N + k_1) - 2k_9/N], \\ \lambda_2 &= k_3[N(k_{12} - k_3)^2 + k_3(N - k_{12})^2] + (N - k_3)^2k_7 \\ &\quad - 2(N - k_3)[(k_{12} - k_3)k_5 + (N - k_{12})k_6/N] \\ &\quad + 2(N - k_{12})(k_{12} - k_3)k_4/N, \\ \lambda_3 &= [(N - k_{12})^2(N - 1)(a - 1) - (N - k_3)^2(a - 1)(b - a) \\ &\quad + (k_{12} - k_3)^2(N - 1)(b - a)]/(N - b), \\ \lambda_4 &= (N - k_{12})^2[k_3(N + k_1) - 2k_8/N], \\ \lambda_5 &= (N - k_{12})^2(N - k_1),\end{aligned}$$

and

$$\lambda_6 = (N - k_{12})(N - k_3)(k_{12} - k_3).$$

Similarly, expression (6) becomes

$$\operatorname{var}(\hat{\sigma}_\beta^2) = \frac{2(k_7 + Nk_3 - 2k_5)\sigma_\beta^4 + 4(N - k_{12})\sigma_\beta^2\sigma_e^2 + 2(b - a)(N - a)\sigma_e^4/(N - b)}{(N - k_{12})^2}$$

and (7) reduces to

$$\begin{aligned}(N - k_1)(N - k_{12})\operatorname{cov}(\hat{\sigma}_\alpha^2\hat{\sigma}_\beta^2) &= 2[k_5 - k_7 + (k_6 - k_4)/N]\sigma_\beta^4 \\ &\quad + 2(a - 1)(b - a)\sigma_e^4/(N - b) - (N - k_{12})(k_{12} - k_3)\operatorname{var}(\hat{\sigma}_\beta^2).\end{aligned}$$

These variances are in terms of the unknown variance components σ_α^2, σ_β^2 and σ_e^2 so that estimation of the variances in any particular case is only possible by replacing the components in these formulae by their estimates.

Balanced data. The above formulae reduce to the well-known results for balanced data when all the n_{ij} are put equal, to n, say. Suppose that all levels of the main classification have c sub-classes so that $b = ac$. Then, for example,

$$\text{var}(\hat{\sigma}^2_\beta) = \frac{2(an^2 + acn^2 - 2an^2)\sigma^4_\beta + 4an(c-1)\sigma^2_\beta\sigma^2_e + 2a(c-1)\sigma^4_e/ac(n-1)}{a^2n^2(c-1)^2}$$

which reduces to

$$\text{var}(\hat{\sigma}^2_\beta) = \frac{2}{n^2}\left[\frac{n\sigma^2_\beta + \sigma^2_e)^2}{a(c-1)} + \frac{\sigma^4_e}{ac(n-1)}\right].$$

This is the result obtained directly for the balanced case when $T_{ab} - T_a$ and $T_o - T_{ab}$ are distributed independently as χ^2 with $a(c-1)$ and $ac(n-1)$ degrees of freedom respectively. Their expectations, obtained from equation (1), are

$$E(T_{ab} - T_a) = a(c-1)(n\sigma^2_\beta + \sigma^2_e)$$

and

$$E(T_o - T_{ab}) = ac(n-1)\sigma^2_e$$

and their variances equal twice the square of their expectations divided by their degrees of freedom. The variance of the estimate of σ^2_β, namely

$$\hat{\sigma}^2_\beta = \frac{1}{n}\left[\frac{T_{ab} - T_a}{a(c-1)} - \frac{T_o - T_{ab}}{ac(n-1)}\right],$$

is accordingly as shown above.

Acknowledgment. Sincere thanks are given to Dr. C. R. Henderson of Cornell University, Ithaca, N. Y., for support of this work from funds of a research grant entitled "Estimation of genetic parameters" from the National Science Foundation of the U. S. A.

REFERENCES

[1] EISENHART, CHURCHILL, "The assumptions underlying the analysis of variance," *Biometrics*, Vol. 3 (1947), pp. 1–21.

[2] HENDERSON, C. R., "Estimation of variance and covariance components," *Biometrics*, Vol. 9 (1953), pp. 226–252.

[3] SEARLE, S. R., "Matrix methods in variance and covariance components analysis," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 737–748.

[4] SEARLE, S. R., "Sampling variances of estimates of components of variance," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 167–178.

[5] SNEDECOR, G. W., *Statistical Methods*, 5th Ed., Iowa State College Press, Ames, Iowa, 1957.