AN OPERATIONAL APPROACH TO THE r-WAY CROSSED CLASSIFICATION[1]

By J. D. Bankier

McMaster University

1. Summary. An operational method is used to obtain known formulas [2] for the expected values of the mean squares and the variances of estimates of variance components obtained from the analysis of variance of an r-way crossed classification. The results are independent of normality assumptions.

2. Introduction. We begin with the results and notation of a book by Mann [4], considering an r-way classification with replication. An observation is denoted by $x_a = x_{a_1 \cdots a_{r+1}}$, $a_i = 1, \cdots, t_i$, where the last subscript is used to indicate replications. Main effects and interactions are represented by $\mu(I, a_I) = \mu(i_1, \cdots, i_k; a_{i_1}, \cdots, a_{i_k})$, where $I = (i_1, \cdots, i_k)$ is a subset of $R = (1, \cdots, r)$, and $\mu(I, a_I) = \mu$ if I is the null set.

We now introduce two operators, $D_i = D_{a_i}$ which drops i and a_i from $\mu(R, a_R)$ and $M_i = M_{a_i}$ which averages a function over a_i, if a_i appears, and otherwise leaves the function unchanged. These operators are commutative with respect to themselves and each other and all the ordinary laws of algebra, excluding division, hold. It is assumed that

$$E(x_a) = (1 + D)_R \mu(R, a_R), \tag{2.1}$$

$$M_{i_j} \mu(I, a_I) = 0, \qquad j = 1, \cdots, k, \tag{2.2}$$

where $(1 + D)_R = (1 + D_1) \cdots (1 + D_r)$. It is easy to establish

Lemma 2.1. *Independent of condition* (2.2),

$$\begin{gathered} M_i D_i = D_i, \qquad (1 - M_i) D_i = 0, \qquad M_i(1 + D_i) = M_i + D_i, \\ (1 - M_i)(1 + D_i) = 1 - M_i. \end{gathered} \tag{2.3}$$

Subject to condition (2.2),

$$(M_i + D_i)\mu(R, a_R) = D_i \mu(R, a_R), \qquad (1 - M_i)\mu(R, a_R) = \mu(R, a_R), \tag{2.4}$$

provided that these expressions are not multiplied by D_i.

Mann establishes an identity [4, p. 52]

$$tMx_a^2 = (1 + D)\mathrm{SS}(R + 1) \tag{2.5}$$

Received July 17, 1957; revised June 17, 1959.

[1] This investigation was supported (in part) by a research grant from the National Research Council of Canada.

where $R + 1 = (1, \cdots, r + 1), t = t_{h+1} = t_1 \cdots t_{r+1}, (1 + D) = (1 + D)_{R+1}$,

$$\text{SS}(I) = tQ(I)/t_I = tM_I A^2(I, a_I), \tag{2.6}$$

$$A(I, a_I) = M_{R+1-I}(1 - M)_I x_a, \tag{2.7}$$

$$t_I = t_{i_1} \cdots t_{i_k}, \qquad M_{R+1-I} = M_1 \cdots M_{r+1}/M_{i_1} \cdots M_{i_k}, \qquad \text{and}$$

$$(1 - M)_I = (1 - M_{i_1}) \cdots (1 - M_{i_k}).$$

The following lemma is an immediate consequence of definition (2.7).

LEMMA 2.2 $A(I, a_I) = M_b(t\delta_{ab} - 1)_I x_b$ where

$$(t\delta_{ab} - 1)_I = (t_{i_1}\delta_{a_{i_1}b_{i_1}} - 1) \cdots (t_{i_k}\delta_{a_{i_k}b_{i_k}} - 1).$$

The above notation and operators can be employed to simplify considerably the usual derivations of the analysis of variance for the r-way crossed classification. As an example, we prove the identity (2.5). Making use of the equations (2.6), (2.7), and Lemma 2.2, we have

$$\begin{aligned} (1 + D)\text{SS}(R + 1) &= t(1 + D)_{a_{R+1}} M_a A^2(R + 1, a_{R+1}) \\ &= t(1 + D)_{a_{R+1}} M_a M_b M_c (t\delta_{ab} - 1)_{R+1}(t\delta_{ac} - 1)_{R+1} x_b x_c \\ &= t(1 + D)_{a_{R+1}} M_b M_c (t\delta_{bc} - 1)_{R+1} x_b x_c \\ &= tM_b M_c(1 + t\delta_{bc} - 1)_{R+1} x_b x_c = tM_b x_b^2 . \end{aligned}$$

3. The Type I model. We define ϵ_a by the equation $x_a = E(x_a) + \epsilon_a$ and make the usual assumptions for the Type I model [1, p. 348] save that we do not assume that the ϵ_a are normally distributed. It will be sufficient if they are independently distributed with zero means, a common variance, σ^2, and such other moments as we require. In addition to the sum of squares SS(I), we will be interested in the error sum of squares, SSE, which is obtained by summing all SS(I) for which the last index is $r + 1$. It is easy to verify, using the operators, that

$$\text{SSE} = tM[(1 - M_{r+1})x_a]^2. \tag{3.1}$$

The corresponding mean sums of squares, MS(I) and MSE, are obtained by dividing the above sums of squares by their degrees of freedom which are $(t - 1)_I$ and $t_R(t_{r+1} - 1)$, respectively. In the case where $t_{r+1} = 1$, it is necessary to assume that the $\mu(R, a_R)$ are zero. In this case the value of SS(I) is unchanged, but SSE is SS(R) and the corresponding degrees of freedom are $(t - 1)_R$.

We now state and prove two lemmas.

LEMMA 3.1. *For the Type I model*

$$A(I, a_I) = \mu(I, a_I) + M_{R+1-I}(1 - M)_I \epsilon_a . \tag{3.2}$$

PROOF. We deduce from (2.7), and (2.1) that we must prove

$$\mu(I, a_I) = M_{R+1-I}(1 - M)_I(1 + D)_R \mu(R, a_R).$$

By (2.3), (2.4), and (2.2)

$$M_{R+1-I}(1 - M)_I(1 + D)_R\mu(R, a_R) = M_{R-I}(1 - M)_I(1 + D)_{R-I}\mu(R, a_R)$$
$$= (1 - M)_I(M + D)_{R-I}\mu(R, a_R)$$
$$= (1 - M)_I D_{R-I}\mu(R, a_R) = (1 - M)_I\mu(I, a_I) = \mu(I, a_I).$$

LEMMA 3.2. *For the Type I model*

$$E[M_{R+1-I}(1 - M)_I\epsilon_a]^2 = (t - 1)_I\sigma^2/t. \tag{3.3}$$

PROOF. The desired expected value is necessarily of the form $c\sigma^2$ where c is a constant which does not depend upon the form of the distribution of the ϵ_a, which we may assume to be NID(0, σ^2). Under these conditions, SS(I) is distributed as $\chi^2\sigma^2$ when the $\mu(I, a_I) = 0$ and the result (3.3) is easily established.

As an almost immediate consequence of these lemmas we obtain:

THEOREM 3.1. *For the Type I model*

$$E[\text{MS}(I)] = \sigma^2 + t\sigma^2(I)/t_I \tag{3.4}$$

where $\sigma^2(I) = t_I M_I \mu^2(I, a_I)/(t - 1)_I$. *When* $t_{r+1} = 1$, $E(\text{MSE}) = \sigma^2 + \sigma^2(R)$, *and, otherwise,* $E(\text{MSE}) = \sigma^2$.

Assuming that $t_{r+1} > 1$ and equating mean squares to their expected values, we obtain the unbiased estimates of the variance components

$$\widehat{\sigma^2}(I) = t_I[\text{MS}(I) - \text{MSE}]/t, \qquad \widehat{\sigma^2} = \text{MSE}.$$

Examination of $(\text{SSE})^2$ indicates that $E(\text{SSE})^2 = k_1\mu_4 + k_2\mu_2^2$ where k_1 and k_2 are independent of the distribution of the ϵ_a. Accordingly, we may assume that the ϵ_a are NID(0, σ^2) and obtain a linear relationship between k_1 and k_2 since SSE is then distributed as $\chi^2\sigma^2$. A little computation determines k_1 and leads to the conclusion that $\text{var}(\widehat{\sigma^2}) = (\mu_4 - 3\mu_2^2)/t + 2\mu_2^2/t_R(t_{r+1} - 1)$.

Making use of (3.2), we find that $\widehat{\sigma^2}(I) = \sigma^2(I) + D + F$ where

$$D = 2t_I M_b \mu(I, b_I)\epsilon_b/(t - 1)_I,$$

$$F = t_I M_b M_c[(t\delta_{bc} - 1)/(t - 1)_I - (t_{r+1}\delta_{b_{r+1}c_{r+1}} - 1)\delta_{bc_R}/(t_{r+1} - 1)]\epsilon_b\epsilon_c.$$

We note that the quadratic form F contains no squared terms, $E(D) = E(DF) = E(F) = 0$, $E(D^2) = 4t_I^2 M_b M_c \mu(I, b_I)\mu(I, c_I)\delta_{bc}\sigma^2/(t - 1)_I^2 = 4t_I\sigma^2\sigma^2(I)/t(t - 1)_I$, and a similar, but longer, calculation gives

$$E(F^2) = 2t_I^2[1/(t - 1)_I + 1/t_R(t_{r+1} - 1)]\mu_2^2/t^2.$$

These results lead to the conclusion that

$$\text{var}[\widehat{\sigma^2}(I)] = 4t_I\sigma^2\sigma^2(I)/t(t - 1)_I + 2t_I^2[1/(t - 1)_I + 1/t_R(t_{r+1} - 1)]\sigma^4/t^2 \tag{3.5}$$

4. The Type II model. The usual assumptions are made for a Type II model [3] save that we make no normality assumptions. We recall that equation (2.2) no longer holds. We state

LEMMA 4.1. *For any model,*

$$A(I, a_I) = M_{R-I}(1 - M)_I(1 + D)_{R-I}\mu(R, a_R) + M_{R+1-I}(1 - M)_I\epsilon_a.$$

PROOF. This relation was established in the proof of Lemma 3.1.

This lemma and the method used to prove Theorem 3.1 leads at once to:

THEOREM 4.1. *For the Type II model,*

$$E[\mathrm{MS}(I)] = \sigma^2 + t(1 + D)_{R-I}[\sigma^2(R)/t_R], \tag{4.1}$$

where $\sigma^2(R)$ *is the variance of the* $\mu(R, a_R)$.

Equating mean squares to their expected values and solving we obtain:

LEMMA 4.2. *For a Type II model, without normality assumptions,*

$$\widehat{\sigma^2}(I) = (-1)^{r-k}(1 - D)_{R-I}[\mathrm{MS}(R)]/t_{R+1-I} \qquad (k < r),$$

and

$$\widehat{\sigma^2}(R) = [\mathrm{MS}(R) - \mathrm{MSE}]/t_{r+1}.$$

The estimates are given in a more convenient form in

LEMMA 4.3. *For a Type II model, without normality assumptions,* $\widehat{\sigma^2}(I) = A + B + C$ $(k < r)$, *where*

$$\begin{aligned}
A &= t_R M_{b_R} M_{c_R}(t\delta_{bc} - 1)_I(1 - \delta_{bc})_{R-I} w_{b_R} w_{c_R}/(t - 1)_R,\\
B &= 2t_R M_{b_R} M_c(t\delta_{bc} - 1)_I(1 - \delta_{bc})_{R-I} w_{b_R}\epsilon_c/(t - 1)_R,\\
C &= t_R M_b M_c(t\delta_{bc} - 1)_I(1 - \delta_{bc})_{R-I}\epsilon_b\epsilon_c/(t - 1)_R,\\
w_{b_R} &= (1 + D)_{R-I}\mu(R, b_R).
\end{aligned}$$

It will be noted that the coefficients of $w_{b_R}^2$ and ϵ_b^2 are equal to zero and it follows that $E[\widehat{\sigma^2}(I)]^2 = E(A^2 + B^2 + C^2)$, and computation gives us

$$\begin{aligned}
\mathrm{var}[\widehat{\sigma^2}(I)] &= [\mu_4(I) - 3\sigma^4(I)]/t_I\\
&\quad + 2t_I \sum_{X,Y \subset R-I} \sigma^2(I + X)\sigma^2(I + Y)/t_I(t - 1)_I t_{XY}(t - 1)_{XY} t_{X-XY} t_{Y-XY}\\
&\quad + 4t_I\sigma^2(1 + D)_{R-I}[\sigma^2(R)/(t - 1)_R]/t + 2t_I t_R\sigma^4/t^2(t - 1)_R \qquad (k < r),
\end{aligned} \tag{4.2}$$

where XY is the set of numbers common to X and Y.

If $t_{r+1} > 1$,

$$\begin{aligned}
\mathrm{var}[\widehat{\sigma^2}(R)] = [\mu_4(R) - 3\sigma^4(R)]/t_R + 2\sigma^4(R)/(t - 1)_R + 4\sigma^2\sigma^2(R)/t_{r+1}(t - 1)_R\\
+ 2[1/(t - 1)_R + 1/t_R(t_{r+1} - 1)]\sigma^4/t_{r+1}^2.
\end{aligned}$$

5. The Type III model. We assume that the t_I expressions of the form $\mu(I, a_I)$ are a sample from a finite population, $P(I)$, consisting of T_I members with zero mean and variance

$$\sigma^2(I) = T_I \mathfrak{M}_I \mu^2(I, a_I)/(T-1)_I \tag{5.1}$$

where $\mathfrak{M}_i = \mathfrak{M}_{a_i}$ is an operator which averages a function over a_i when the range of a_i is from 1 to T_i. We also assume that

$$\mathfrak{M}_{i_j}\mu(I, a_I) = 0, \qquad j = 1, \cdots, k, \tag{5.2}$$

and that random variables from different populations are independent. We draw our sample from $P(I)$ by selecting t_{i_j} numbers at random from the set $1, \cdots, T_{i_j} (j = 1, \cdots, k)$. Lemma 4.1 holds under these conditions, so our problem reduces to finding the expected value of expressions of the form

$$[M_{R-I}(1-M)_I \mu(R, a_R)]^2.$$

We will require the following lemmas.

LEMMA 5.1. *For a Type III model*

$$E[\mu(I, a_I)\mu(I, b_I)] = (\delta_{ab} - 1/T)_I \sigma^2(I)$$

where $\delta_{a_{i_j} b_{i_j}} = 1$ *if* $a_{i_j} = b_{i_j}$ *and is zero otherwise.*

LEMMA 5.2. *For the Type III model*

$$E[M_{R-I}(1-M)_I \mu(R, a_R)]^2 = (1-1/t)_I (1/t - 1/T)_{R-I} \sigma^2(R) \quad (k \leqq r).$$

The application of the above lemmas leads to the following theorem which has also been stated by Bennett and Franklin [1].

THEOREM 5.1. *For the Type III model*

$$E\,MS(I) = \sigma^2 + t(1+D)_{R-I}(1-t/T)_{R-I}\sigma^2(R)/t_R.$$

Equating mean squares to their expected values and solving we obtain

THEOREM 5.2. *For the Type III model*

$$\widehat{\sigma^2}(I) = [(D-1)_{R-I}(1-t/T)_{R-I}\text{MS}(R) - t_{R-I}\text{MSE}/T_{R-I}]/t_{R+1-I} \quad (k \leqq r).$$

Computations which are too long to be included in this paper lead to the conclusion that

$$E[\widehat{\sigma^2}(I)]^2 = A + B + C + D \tag{5.3}$$

where

$$A = 4(t_I/t)\sigma^2(1+D)_{R-I}[(1-1/T)_{R-I}(1-t/T)_{R-I}\sigma^2(R)/(t-1)_R],$$

$$B = 2t_R t_I\{[(t-1)/T^2 + (1-1/T)^2]_{R-I} + (t-1)_R/t_I(t_{r+1}-1)T^2_{R-I}\}\sigma^4/t^2,$$

$$C = 2\sum_{\substack{X,Y \subset R-I \\ X \neq Y}} (1-t/T)_{X+Y}\sigma^2(I+X)\sigma^2(I+Y)/(t-1)_I(t-1)_{XY}t_{X+Y-XY},$$

and

$$t_R(t-1)_R D = \sum_{X \subset R-I} t_{R-I-X}(t-1)_{R-I-X}(1-t/T)_X /$$
$$[T(T-1)(T-2)(T-3)]_{I+X} \times \sum_{f_{I+X}} \sum_{g_{I+X}} \sum_{h_{I+X}} \sum_{m_{I+X}}$$
$$[\{(t-1)^2T^2 - (3t-1)(t-1)T + t^2 + t\}\delta_{fg}\delta_{hm}$$
$$+ (T-t)(T-t-1)(\delta_{fh}\delta_{gm} + \delta_{fm}\delta_{gh})$$
$$+ T(T-t)\{(t-1)T - t - 1\}\delta_{fg}\delta_{gh}\delta_{hm}]_I$$
$$\times [(3 - t - t/T - 1/T)\delta_{fg}\delta_{hm} + (T-1)^2(T-t-1)$$
$$(\delta_{fh}\delta_{gm} + \delta_{fm}\delta_{gh})/T + \{-2T^2 + (3t+1)T - t - 1\}$$
$$\delta_{fg}\delta_{gh}\delta_{hm}]_X \times \mu(I+X, f_{I+X})$$
$$\mu(I+X, g_{I+X})\mu(I+X, h_{I+X})\mu(I+X, m_{I+X}).$$

The following formula which is used in the derivative of the expression for D may be of interest:

$$[T(T-1)(T-2)(T-3)]_R E[\mu(R, b_R)\mu(R, c_R)\mu(R, d_R)\mu(R, e_R)]$$
$$= \sum_{f_R} \sum_{g_R} \sum_{h_R} \sum_{m_R} [(A_1T^2 + A_2T + A_3)\delta_{fg}\delta_{hm} + (B_1T^2 + B_2T + B_3)\delta_{fh}\delta_{gm}$$
$$+ (C_1T^2 + C_2T + C_3)\delta_{fm}\delta_{gh} + (D_1T^3 + D_2T^2 + D_3T - 6)\delta_{fg}\delta_{gh}\delta_{hm}]_R$$
$$\cdot \mu(R, f_R)\mu(R, g_R)\mu(R, h_R)\mu(R, m_R)$$

where

$A_1 = \delta_{bc}\delta_{de} - C(3, 4)$, $A_2 = C(2, 3) + C(3, 4) - \delta_{bc} - \delta_{de} - 3\delta_{bc}\delta_{de}$,

$A_3 = 1 + 3\delta_{bc} + 3\delta_{de} - C(1, 2) - C(2, 3) + C(2, 4)$, $B_1 = \delta_{bd}\delta_{ce} - C(3, 4)$,

$B_2 = C(2, 3) + C(3, 4) - \delta_{bd} - \delta_{ce} - 3\delta_{bd}\delta_{ce}$,

$B_3 = 1 + 3\delta_{bd} + 3\delta_{ce} - C(1, 2) - C(2, 3) + C(2, 4)$, $C_1 = \delta_{be}\delta_{cd} - C(3, 4)$

$C_2 = C(2, 3) + C(3, 4) - \delta_{be} - \delta_{cd} - 3\delta_{be}\delta_{cd}$,

$C_3 = 1 + 3\delta_{be} + 3\delta_{cd} - C(1, 2) - C(2, 3) + C(2, 4)$, $D_1 = C(3, 4)$,

$D_2 = C(3, 4) - C(2, 3) - C(2, 4)$, $D_3 = 2C(1, 2) - C(2, 3) + C(2, 4)$,

$$C(1, 2) = \delta_{bc} + \delta_{bd} + \delta_{be} + \delta_{cd} + \delta_{ce} + \delta_{de},$$

$$C(2, 3) = \delta_{bc}\delta_{cd} + \delta_{bc}\delta_{ce} + \delta_{bd}\delta_{de} + \delta_{cd}\delta_{de}.$$

$$C(2, 4) = \delta_{bc}\delta_{de} + \delta_{bd}\delta_{ce} + \delta_{be}\delta_{cd},$$

$$C(3, 4) = \delta_{bc}\delta_{cd}\delta_{de}.$$

6. Acknowledgement. I am indebted to the referees for helpful comments and suggestions.

REFERENCES

[1] C. A. Bennett and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, John Wiley & Sons, New York, 1954.
[2] Jerome Cornfield and John W. Tukey, "Average values of mean squares in factorials," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 907–949.
[3] S. L. Crump, "The estimation of variance components in analysis of variance," *Biometrics Bulletin*, Vol. 2 (1946), pp. 7–11.
[4] H. B. Mann, *Analysis and Design of Experiments*, Dover Publications, New York, 1949.