ANALYSIS OF COVARIANCE IN TWO-WAY CLASSIFICATION WITH DISPROPORTIONATE CELL FREQUENCIES

By M. N. Das
Indian Council of Agricultural Research, New Delhi

1. Introduction

Wilks (1938, 1950) has dealt with the problem of analysis of covariance with disproportionate cell frequencies. He has indicated the solutions determinantly such that the analysis of 2×2 table, for example, involves the writing of determinants of the 9th order. The purpose of this paper is to give an explicit solution in a simple form suitable for systematic computations.

2. Notation

The two factors under study will be termed factors A and B. Factor A has p classes with effects $a_1, a_2, \ldots a_p$ and factor B has q classes with effects $b_1, b_2, \ldots b_q$. The number of observations in the cell defined by the i-th class of A and j-th class of B will be denoted by n_{ij}. The k-th observation in the (i, j) cell will be denoted by y_{ijk} for the dependent variate y and by x_{ijk} for the concomitant variate x.

Besides, the following notations will be used:—

Grand mean $= m$.

Regression of y on $x = a$.

$$\sum_i n_{ij} = n_{.j},\ \sum_j n_{ij} = n_{i.},\ \sum_{ij} n_{ij} = \sum_i n_{i.} = \sum_j n_{.j} = n.$$

$$\sum_k y_{ijk} = Y_{ij},\ \sum_{kj} y_{ijk} = \sum_j Y_{ij} = Y_{i.},\ \sum_{ik} y_{ijk} = \sum_i Y_{ij} = Y_{.j},$$

$$\sum_k x_{ijk} = X_{ij},\ \sum_{jk} x_{ijk} = \sum_j X_{ij} = X_{i.},\ \sum_{ik} x_{ijk} = \sum_i X_{ij} = X_{.j},$$

$$\sum_{ijk} y_{ijk} = Y,\ \sum_{ijk} x_{ijk} = X$$

$$Q_i = Y_{i.} - \sum_j \frac{Y_{.j}}{n_{.j}} n_{ij} \text{ and } Q_{i(x)} = X_{i.} - \sum_j \frac{X_{.j}}{n_{.j}} n_{ij}$$

The total sum of products $\sum_{ijk} x_{ijk}\, y_{ijk}$ will be denoted by $T.S.P.$

The total sum of squares for x, *i.e.*, $\sum_{ijk} x^2_{ijk}$ by $(T.S.S.)_x$;

The sum of products $\sum_j \frac{Y_{.j} \times X_{.j}}{n_{.j}}$ by $B.S.P.$;

$\sum \frac{X_{.j}^2}{n_{.j}}$ by $(B.S.S.)_x$;

$\sum \frac{Y_{ij} X_{ij}}{n_{ij}}$ by P_{xy} and $\sum \frac{X_{ij}^2}{n_{ij}}$ by T_x^2

3. Normal Equations

On the postulate that $y_{ijk} = m + a_i + b_j + ax_{ijk} + \epsilon_{ijk}$, where ϵ_{ijk} is a random variate with zero mean and variance σ^2, it follows from the least squares theory that the best estimates of a_i, b_j, a and m or any of their linear functions are obtainable from the following normal equations:—

$$Y_{i.} = n_{i.} m + n_{i.} a_i + \sum_j n_{ij} b_j + aX_{i.}$$

$$(i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, q) \quad (1)$$

$$Y_{.j} = n_{.j} m + n_{.j} b_j + \sum_i n_{ij} a_i + aX_{.j}$$

$$(i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, q) \quad (2)$$

$$Y = nm + \sum_i n_{i.} a_i + \sum_j n_{.j} b_j + aX$$

$$T.S.P. = Xm + \sum_i X_{i.} a_i + \sum_j X_{.j} b_j + a\,(T.S.S.)_x.$$

Eliminating b_j's these become

$$Q_i = a_i \left(n_{i.} - \sum_j \frac{n_{ij}^2}{n_{.j}}\right) - \sum_{k \neq i} a_k \left\{\sum_j \frac{n_{kj} \times n_{ij}}{n_{.j}}\right\} + a \cdot Q_{i(x)}$$

$$(i = 1, 2, \ldots, p) \quad (3)$$

and

$$(T.S.P. - B.S.P.) = \sum_i a_i Q_{i(x)} + a\,\{(T.S.S.)_x - (B.S.S.)_x\} \quad (4)$$

Denoting $(T.S.P. - B.S.P.)$ by β; $\{(T.S.S.)_x - (B.S.S.)_x\}$ by α and β/α by b, the last equation becomes

$$b - \frac{\sum a_i Q_{i(x)}}{\alpha} = a$$

Hence from (3)

$$Q_i - bQ_{i(x)} = a_i\left(n_{i.} - \sum_j \frac{n_{ij}^2}{n_{.j}} - \frac{Q_{i(x)}^2}{a}\right)$$

$$-\sum_{k\neq i} a_k \left(\sum_j \frac{n_{kj}\times n_{ij}}{n_{.j}} + \frac{Q_{i(x)}\times Q_{k(x)}}{a}\right)$$

Putting

$$Q_i - bQ_{i(x)} = Q_i'; \left(n_{i.} - \sum_j \frac{n_{ij}^2}{n_{.j}} - \frac{Q_{i(x)}^2}{a}\right) = c_{ii}$$

and

$$-\left(\sum_j \frac{n_{kj}\times n_{ij}}{n_{.j}} + \frac{Q_{i(x)}\times Q_{k(x)}}{a}\right) = c_{ik},$$

the normal equations in a_i alone become

$$Q_i' = \sum_k c_{ik}\, a_k \qquad (i, k = 1, 2, \ldots, p)$$

As only $p - 1$ of the above equations are independent, the restriction $\Sigma a_i = 0$ will be taken together with them to make the solution unique.

Imposing this restriction and eliminating any treatment say, a_p, the equations become

$$Q_i' = a_i P_{ii}' - \sum_{k\neq i,p} a_k P_{ik}' \qquad (i = 1, 2, \ldots, p-1)$$

where

$$P_{ii}' = P_{ii} - \frac{Q_{i(x)}}{a}\,(Q_{i(x)} - Q_{p(x)})$$

and

$$P_{ii} = n_{i.} - \sum_j \frac{n_{ij}\,(n_{ij} - n_{pj})}{n_{.j}}$$

and

$$P_{ik}' = P_{ik} + \frac{Q_{i(x)}}{a}\,(Q_{k(x)} - Q_{p(x)})$$

and

$$P_{ik} = \sum_j \frac{n_{ij}\,(n_{kj} - n_{pj})}{n_{.j}}$$

4. Sum of Squares

The total sum of squares $\sum_{ijk} y_{ijk}^2$ splits up into two parts, *viz*. (i) that due to the estimates of all the constants including a and m

and (ii) that due to the deviations from the regression involving all the constants. The sum of squares due to the estimates is equal to

$$Y.m + \Sigma\, a_i Y_{i.} + \Sigma\, b_j Y_{.j} + a\,(T.S.P.)$$

Eliminating b_j's, a and m this becomes

$$\sum_i a_i Q_i' + \sum_j \frac{Y_{.j}^2}{n_{.j}} + b\beta$$

Again the sum of squares due to the estimates of all the constants on the hypothesis $a_1 = a_2 = \ldots = a_p$ is equal to

$$\sum_j \frac{Y_{.j}^2}{n_{.j}} + b\beta$$

Hence the sum of squares due to the estimates of a_i's alone, *i.e.*, the sum of squares adjusted for all other effects, is equal to $\Sigma\, a_i Q_i'$ based on $(p - 1)$ *d.f.* Let (A) denote this adjusted sum of squares and A, the sum of squares due to the estimates of all the constants on the hypothesis $b_1 = b_2 = \ldots = b_q$, to be hereafter called the unadjusted *S.S.* due to A. Also let (B) and B denote the corresponding sums of squares for the factor B. Then the sum of squares due to the estimates of all the constants is equal to $(A) + B = A + (B)$ from symmetry. Hence $A - (A) = B - (B) = \delta$ (say) where δ may be called an adjustment factor for non-orthogonality.

Thus once the adjusted *S.S.* for one of the factors (conveniently for that one which has the smaller number of classes) is obtained, that for the other factor can be obtained from A and B, the calculation of which is straightforward.

The sum of squares due to the deviations is equal to

$$\sum_{ijk} y_{ijk}^2 - (A) - B$$

As is evident from the postulate this method of analysis is based on the assumption that there is no interaction between the two factors. If it is assumed that there is interaction the postulate will change, *i.e.*,

$$y_{ijk} = m + a_i + b_j + h_{ij} + a x_{ijk} + \epsilon_{ijk}$$

where h_{ij}'s are the interaction effects. The sum of squares due to the estimates of all the constants including h_{ij}'s is equal to

$$\sum_{ij} \frac{Y_{ij}^2}{n_{ij}} + \frac{(T.S.P. - P_{xy})^2}{(T.S.S.)_x - T_x^2} \text{ with } pq + 1 \; d.f.$$

Again the sum of squares due to the estimates of all the constants on the hypothesis that $h_{ij} = 0$ is evidently equal to the sum of squares due to the estimates of m, a, a_i and b_j, *i.e.*, $(A) + B$ on $(p + q)$ *d.f.*

Hence the sum of squares due to the interaction effects is equal to

$$\sum_{ij} \frac{Y_{ij}^2}{n_{ij}} + \frac{(T.S.P. - P_{xy})^2}{(T.S.S.)_x - T_x^2} - (A) - B \text{ on } (p-1)(q-1) \ d.f.$$

The sum of squares due to the deviations in this case is equal to

$$\sum_{ijk} y_{ijk}^2 - \sum \frac{Y_{ij}^2}{n_{ij}} - \frac{(T.S.P. - P_{xy})^2}{(T.S.S.)_x - T_x^2} \text{ on } (n - pq - 1) \ d.f.$$

5. Comparison of Class Effects

Let $\Sigma\, l_i\, a_i$ where $\Sigma\, l_i = 0$ be any comparison of the effects of A. As the best estimate of $\Sigma\, l_i\, a_i$ is obtained by substituting in it the values of a_i obtained from the solution of the normal equations, it will be certain function of Q_i''s.

So let

$$\Sigma\, l_i\, a_i = \sum_i q_i\, Q_i'$$

$$= \Sigma\, q_i\, (c_{ii}\, a_i + \sum_{k \neq i} c_{ik}\, a_k)$$

$$= \Sigma\, a_i\, (c_{ii}\, q_i + \sum_{k \neq i} c_{ki}\, q_k)$$

Hence q_i's can be obtained by equating the coefficients of a_i's on both sides, *i.e.*, from

$$l_i = c_{ii}\, q_i + \sum_{k \neq i} c_{ki}\, q_k \qquad (i = 1, 2, \ldots, p)$$

It will be seen that these equations are the same as the normal equations in a_i with Q_i''s replaced by l_i's.

Variance of $\Sigma\, q_i\, Q_i' = \sum_{ik} q_i q_k \operatorname{cov}(Q_i', Q_k')$

$= \sum_{ik} q_i q_k c_{ik} \sigma^2$ as cov (Q_i', Q_k') can be shown to be equal to $c_{ik}\, \sigma^2$

$= \sigma^2 \sum_i q_i\, (q_1 c_{i1} + q_2 c_{i2} + \ldots + q_p c_{ip})$

$= \sigma^2 \sum_i q_i\, l_i$

The variance of the estimate of $a_i - a_k$, in particular, comes out to be $\sigma^2 \{(R_{ii} - R_{ik}) - (R_{ki} - R_{kk})\}$ where R_{ii} and R_{ik} are respectively the coefficients of Q_i' and Q_k' in the solution of a_i. Also R_{ki} and R_{kk} are the coefficients of Q_i' and Q_k' in the solution of a_k. Actually R_{ik}, etc., are the elements of the inverse of the information matrix $[P_{ik}']$. R_{ik} need not be equal to R_{ki} as the information matrix is not symmetrical after the elimination of a_p. The variance of the contrast $a_i - a_k$ can also be obtained from $\sigma^2 (\bar{a}_i - \bar{a}_k)$ where $\bar{a}_i$ and $\bar{a}_k$ are the solutions of the normal equations obtained by replacing Q_i' by 1 and Q_k' by -1 and the rest of the Q''s by zero.

The estimate of the comparison $(b_j - b_m)$ is obtained from

$$\bar{y}_j - \bar{y}_m - b(\bar{x}_j - \bar{x}_m) - \sum_{i=1}^{p-1} a_i (M_i - M_p)$$

where

$$\bar{y}_j = \frac{Y_{.j}}{n_{.j}}, \ \bar{x}_j = \frac{X_{.j}}{n_{.j}} \text{ and } M_i = \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{im}}{n_{.m}}\right) - Q_{i(x)} \frac{(\bar{x}_j - \bar{x}_m)}{\alpha}$$

Variance of the estimate of $(b_j - b_m)$ is given by

$$\sigma^2 \left\{\left(\frac{1}{n_{.j}} + \frac{1}{n_{.m}}\right) + \frac{(\bar{x}_j - \bar{x}_m)^2}{\alpha} + \sum_{i=1}^{p-1} (M_i - M_p) \sum_{k=1}^{p-1} R_{ik} M_k\right\}$$

where R_{ik} is as before the coefficient of Q_k' in the solution for a_i.

6. Particular Cases of $2xq$ and $3xq$

Case I.—When any one of the factors, say A, has only two classes, we have

$$a_1 = -a_2 = \frac{Q_1'}{2\left(\sum \frac{n_{1j} \times n_{2j}}{n_{.j}} - \frac{Q_{1\,(x)}^2}{\alpha}\right)} = \frac{Q_1'}{2P'} \text{ (say)}$$

The adjusted $S.S.$ due to the effects of A is given by $(A) = \dfrac{Q_1'^2}{P'}$

Variance of the estimate of $a_1 - a_2 = \dfrac{\sigma^2}{P'}$

Estimate of $b_j - b_m = (\bar{y}_j - \bar{y}_m) - b(\bar{x}_j - \bar{x}_m) - \dfrac{Q_1' M_1}{P'}$

Variance of the estimate of $(b_j - b_m)$

$$= \sigma^2 \left\{\left(\frac{1}{n_{.j}} + \frac{1}{n_{.m}}\right) + \frac{(\bar{x}_j - \bar{x}_m)^2}{\alpha} + \frac{M_1^2}{P'}\right\}$$

Case II.—When A has 3 classes, we have

$$a_1 = \frac{Q_1'P_{22}' + Q_2'P_{12}'}{c}$$

$$a_2 = \frac{Q_1'P_{21}' + Q_2'P_{11}'}{c}$$

$$a_3 = -(a_1 + a_2) = -\frac{Q_1'(P_{22}' + P_{21}') + Q_2'(P_{11}' + P_{12}')}{c}$$

where

$$c = P_{11}'P_{22}' - P_{12}'P_{21}'$$

The adjusted $S.S.$ (A) is given by

$$\frac{1}{c}\{Q_1'^2(2P_{22}' + P_{21}') + Q_1'Q_2'(P_{11}' + P_{22}' + 2P_{12}' + 2P_{21}') + Q_2'^2(2P_{11}' + P_{12}')\}$$

Estimate of $a_1 - a_2$

$$= \frac{1}{c}\{Q_1'(P_{22}' - P_{21}') - Q_2'(P_{11}' - P_{12}')\}$$

Variance of the estimate of $(a_1 - a_2)$

$$= \frac{\sigma^2}{c}\{P_{11}' + P_{22}' - P_{12}' - P_{21}'\}$$

Estimate of $a_1 - a_3$

$$= \frac{1}{c}\{Q_1'(2P_{22}' + P_{21}') + Q_2'(P_{11}' + 2P_{12}')\}$$

Variance of $(a_1 - a_3)$

$$= \frac{\sigma^2}{c}(2P_{22}' + P_{21}')$$

Estimate of $(b_j - b_m)$

$$= (\bar{y}_j - \bar{y}_m) - b(\bar{x}_j - \bar{x}_m) - a_1(2M_1 + M_2) - a_2(M_1 + 2M_2)$$

Variance of $(b_j - b_m)$

$$= \sigma^2\left[\left(\frac{1}{n_{.j}} + \frac{1}{n_{.m}}\right) + \frac{(\bar{x}_j - \bar{x}_m)^2}{a} + \frac{1}{c}\{M_1^2(2P_{22}' + P_{21}') + M_1M_2(P_{11}' + P_{22}' + 2P_{12}' + 2P_{21}') + M_2^2(2P_{11}' + P_{12}')\}\right]$$

In cases where the number of classes for either of the factors is greater than three the results do not assume any simple form. Two difficulties are mainly encountered in such cases. One is the writing of the normal equations and the other is their solution. A convenient way of writing down the normal equations has been shown in Section 8.

7. Case of Two Concomitant Variates x and z

In this case

$$Q_i' = Q_i - b_x Q_{i(x)} - b_z Q_{i(z)}$$

where b_x and b_z are the partial regression coefficients of y on x and z respectively calculated from the sums of squares and products within B classes ignoring A classification.

The normal equations in a_i's come out to be

$$Q_i' = P_{ii}' a_i - \sum_{k \neq i, p} P_{ik}' a_k$$

where

$$P_{ii}' = P_{ii} - \frac{\Sigma z^2}{c} Q_{i(x)} (Q_{i(x)} - Q_{p(x)}) - \frac{\Sigma x^2}{c} Q_{i(z)} (Q_{i(z)} - Q_{p(z)})$$
$$+ \frac{\Sigma xz}{c} \{Q_{i(x)} (Q_{i(z)} - Q_{p(z)}) + Q_{i(z)} (Q_{i(x)} - Q_{p(x)})\}$$

$$P_{ik}' = P_{ik} + \frac{\Sigma z^2}{c} Q_{i(x)} (Q_{k(x)} - Q_{p(x)}) + \frac{\Sigma x^2}{c} Q_{i(z)} (Q_{k(z)} - Q_{p(z)})$$
$$- \frac{\Sigma xz}{c} \{Q_{i(x)} (Q_{k(z)} - Q_{p(z)}) + Q_{i(z)} (Q_{k(x)} - Q_{p(x)})\}$$

and

$$c = \Sigma x^2 \Sigma z^2 - (\Sigma xz)^2$$

and Σx^2, Σz^2 and Σxz are the sums of squares and products within B classes ignoring A classification. The sum of squares due to the estimates of all the constants

$$= \Sigma a_i Q_i' + \sum_j \frac{Y_{.j}^2}{n_{.j}} + b_x \Sigma xy + b_z \Sigma yz$$

The adjusted $S.S.$ due to the estimates of A class effects, *i.e.*,

$$(A) = \Sigma a_i Q_i'$$

In this case

$$B = \sum_j \frac{Y_{.j}^2}{n_{.j}} + b_x \Sigma xy + b_z \Sigma yz$$

If b_x', b_z', $\Sigma' xy$, $\Sigma' yz$ and $\Sigma' xz$ are obtained from the sums of squares and products within the cells then the sum of squares due to the interaction between the factors is

$$\sum \frac{Y_{ij}^2}{n_{ij}} + b'_x \Sigma' xy + b_z' \Sigma' yz - (A) - B$$

on $(p-1)(q-1)$ $d.f.$

The error $S.S.$ will be

$$\sum_{ijk} y_{ijk}^2 - \sum \frac{Y_{ij}^2}{n_{ij}} - b'_x \Sigma' xy - b'_z \Sigma' yz \text{ on } (n - pq - 2)\ d.f.$$

All other results which are functions of Q_i' and P_{ik}' will remain the same but for their new meaning. The estimate of $(b_j - b_m)$

$$= (\bar{y}_j - \bar{y}_m) - b_x(\bar{x}_j - \bar{x}_m) - b_z(\bar{z}_j - \bar{z}_m) - \sum_{i \neq p} (M_i - M_p)\, a_i$$

Variance of $(b_j - b_m)$

$$= \sigma^2 \left\{ \left(\frac{1}{n_{.j}} + \frac{1}{n_{.m}} \right) + \frac{\Sigma z^2}{c} (\bar{x}_j - \bar{x}_m)^2 + \frac{\Sigma x^2}{c} (\bar{z}_j - \bar{z}_m)^2 \right.$$

$$\left. - 2 \frac{\Sigma xz}{c} (\bar{x}_j - \bar{x}_m)(\bar{z}_j - \bar{z}_m) + \sum_{i \neq p} (M_i - M_p) \sum_k R_{ik} M_k \right\}$$

where

$$M_i = \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{im}}{n_{.m}} \right) - (\bar{x}_j - \bar{x}_m) \left(\frac{\Sigma z^2}{c} Q_{i(x)} - \frac{\Sigma xz}{c} Q_{i(z)} \right)$$

$$- (\bar{z}_j - \bar{z}_m) \left(\frac{\Sigma x^2}{c} Q_{i(z)} - \frac{\Sigma xz}{c} Q_{i(x)} \right)$$

and R_{ik} is as before the coefficient of Q_k' in the solution of a_i.

8. WRITING OF THE NORMAL EQUATIONS AFTER ELIMINATING b_j'S, m AND a

The information matrix of the normal equations in a_i's may be obtained from the following operations on the number of observations in the different cells:—

No. of observations in the cells (n_{ij})			Total $(n_{.j})$	$(d_{ij})=(n_{ij}-n_{pj})$ $i=1, 2, \cdots p-1$ $j=1, 2, \cdots q$					$\left(\frac{n_{1j}d_{ij}}{n_{.j}}\right)$			$\left(\frac{n_{2j}d_{ij}}{n_{.j}}\right)$			$\left(\frac{n_{p-1,j}d_{ij}}{n_{.j}}\right)$	
n_{11}	n_{21}	n_{p1}	$n_{\cdot 1}$	d_{11}	$\cdot\cdot$	d_{21}	$\cdots$	$d_{p-1,1}$	$\frac{n_{11}d_{11}}{n_{.1}}$	$\frac{n_{11}d_{21}}{n_{.1}}$	$\frac{n_{11}d_{p-1,1}}{n_{.1}}$	$\frac{n_{21}d_{11}}{n_{.1}}$	$\frac{n_{21}d_{21}}{n_{.1}}$	$\cdots$	$\frac{n_{p-1,}d_{11}}{n_{.1}}$	$\cdots$
n_{12}	n_{22}	n_{p2}	$n_{\cdot 2}$	d_{12}	$\cdot\cdot$	d_{22}	$\cdots$	$d_{p-1,2}$	$\frac{n_{12}d_{12}}{n_{\cdot 2}}$	$\frac{n_{12}d_{22}}{n_{\cdot 2}}$						
.	.	.	.	.	.	.		.	.	.	.	.	.	...	.	...
.	.	.	.	.	.	.		.	.	.	.	.	.	...	.	...
.	.	.	.	.	.	.		.	.	.	.	.	.	...	.	...
n_{1q}	n_{2q}	n_{pq}	$n_{\cdot q}$	d_{1q}	$\cdot\cdot$	d_{2q}	$\cdots$	$d_{p-1,q}$	$\frac{n_{1q}d_{1q}}{n_{\cdot q}}$	$\frac{n_{1q}d_{2q}}{n4q}$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		
$Q_{1(x)}$	$Q_{2(x)}$	$Q_{p(x)}$	α	$\{Q_{1(x)}-Q_{p(x)}\}$		$\{Q_{2(x)}-Q_{p(x)}\}$		$\{Q_{p-1(x)}-Q_{p(x)}\}$	$Q_{1(x)}(Q_{1(x)}-Q_{p(x)})$	$Q_{1(x)}(Q_{2(x)}-Q_{p(x)}$						
								TOTALS ...	$n_{1.}-P_{11}'$	P_{12}'	$P_{1,p-1}'$	P_{21}'	$n_{2.}-P_{22}'$		$P'_{p-1,1}$	$n_{p-1.}-P'_{p-1,p-1}$

Now the normal equations may be written as

$$
\begin{array}{l}
P_{11}'a_1 \quad - P_{12}'a_2 \quad - P_{13}'a_3 \ldots - P'_{1,p-1}a_{p-1} = Q_1' \\
- P_{21}'a_1 \quad + P_{22}'a_2 \quad - P_{23}'a_3 \ldots - P'_{2,p-1}a_{p-1} = Q_2' \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
- P'_{p-1,\,1}a_1 - P'_{p-1,\,2}a_2 \qquad \ldots + P'_{p-1,\,p-1}a_{p-1} = Q'_{p-1}
\end{array}
$$

9. An Example

The data analysed are the body weight records of graded up ewes of age one-and-a-half year collected in a sheep breeding scheme in Madras. The ewes were born in three generations and possessed four different colours, *viz.*, (i) completely white, (ii) white body with other face colour, (iii) mixed colour, and (iv) black colour. The object of the analysis is to test the variance in growth of these ewes due to generations, colours and their interaction independently of each other after correcting for the inequality in weight at birth. The cell frequencies together with the variate totals for each cell are shown below:—

Colour	GENERATIONS													
	FIRST			SECOND			THIRD			TOTAL				
	No. of observations	Total weight: When mature lb. (y)	Total weight: At birth oz. (x)	No. of observations	Total weight: When mature lb. (y)	Total weight: At birth oz. (x)	No. of observations	Total weight: When mature lb. (y)	Total weight: At birth oz. (x)	No. of observations	Total weight: When mature lb. (y)	Total weight: At birth oz. (x)	$\bar{y}$	$\bar{x}$
	1	2	3	1	2	3	1	2	3	1	2	3		
(i) ..	3	160·2	310	7	348·5	772	4	232·0	448	14	740·7	1530	52·91	109·29
(ii) ..	4	188·1	480	24	1271·7	2620	9	456·0	966	37	1915·8	4066	51·78	109·89
(iii) ..	40	2394·0	4311	5	239·6	500	1	53·2	96	46	2686·8	4907	58·41	106·67
(iv) ..	20	1285·3	2100	2	128·2	266	1	43·0	126	23	1456·5	2492	63·33	108·35
Total ..	67	4027·6	7201	38	1988·0	4158	15	784·2	1636	120	6799·8	12995		
Adjusted Totals	Q_i	58·8951			−43·7270			−15·1688						
	$Q_{i(x)}$		−·3367			5·5294			−5·1916					

As generation is having the smaller number of levels the adjusted *S.S.* due to generations has been obtained to get the adjustment factor.

The following table gives the unadjusted $S.S.$:—

	$\Sigma\, y^2$	$\Sigma\, xy$	$\Sigma\, x^2$	Unadjusted $S.S.$ $\left(\equiv S+\frac{\beta^2}{\alpha}\right)$
Between colours ..	387552·31 (S)	735898·97	1407479·13	$387552{\cdot}31+\frac{2445{\cdot}63^2}{22257{\cdot}87}=387821{\cdot}03$
Within colours ..		2445·63 (β)	22257·87 (α)	
Between generations ..	387114·52	735935·77	1407352·25	$387114{\cdot}52+\frac{2408{\cdot}83^2}{22384{\cdot}75}=387373{\cdot}73$
Within generations ..		2408·83	22384·75	
Between cells ..	388953·81	735821·49	1410238·63	$388953{\cdot}81+\frac{2523{\cdot}11^2}{19498{\cdot}37}=389280{\cdot}30$
Within cells ..		2523·11	19498·37	
TOTAL ..	401294·36	738344·60	1429737·00	

$$\text{Error } S.S. = 401294{\cdot}36 - 389280{\cdot}30$$
$$= 12014{\cdot}06$$

The combined adjusted totals Q_i' which will be required in the normal equations may now be obtained as $Q_i' = Q_i - bQ_{i(x)}$ where $b = \beta/\alpha = {\cdot}1099$ as obtained from the sum of squares and products 'within colours'.

Thus $Q_1' = 58{\cdot}9321$, $Q_2' = -44{\cdot}3347$ and $Q_3' = -14{\cdot}5982$.

Operations on the table of observations for writing the normal equations

Colours	GENERATIONS									
	n_{1j}	n_{2j}	n_{3j}	Totals $n_{.j}$	$d_{1j} = n_{1j} - n_{3j}$	$d_{2j} = n_{2j} - n_{3j}$	$\frac{n_{1j} \times d_{1j}}{n_{.j}}$	$\frac{n_{1j} d_{2j}}{n_{.j}}$	$\frac{n_{2j} \times d_{1j}}{n_{.j}}$	$\frac{n_{2j} \times d_{2j}}{n_{.j}}$
(i) ..	3	7	4	14	−1	3	−·2143	·6429	− ·5000	1·5000
(ii) ..	4	24	9	37	−5	15	−·5403	1·6216	−3·2432	9·7297
(iii) ..	40	5	1	46	39	4	33·9130	3·4783	4·2391	·4348
(iv) ..	20	2	1	23	19	1	16·5217	·8696	1·6522	·0870
$Q_{i(\alpha)}$ ..	−0·3367	5·5294	−5·1916	22257·87(α)	4·8549	10·7210	−·0001	−·0002	·0012	·0027
						TOTALS ..	49·6798	6·6122	2·1493	11·7542

The last four totals may be checked as follows:—

If the totals are denoted by T_1, T_2, T_3, ..., then

$$\Sigma\, T_i = \text{total no. of obs.} - (p+1)\,\Sigma\, n_{pj} + p\left(\sum \frac{n_{pj}^2}{n_{.j}} + \frac{Q_{p(x)}^2}{a}\right)$$

In the present case

$$\Sigma\, T_i = 70{\cdot}1955$$

The right-hand side $= 120 - 4\times 15 + 3\times 3{\cdot}3984 = 70{\cdot}1952$

Now

$$P_{11}' = 67 - 49{\cdot}6798 = 17{\cdot}3202$$

$$P_{12}' = 6{\cdot}6122$$

$$P_{21}' = 2{\cdot}1493$$

$$P_{22}' = 38 - 11{\cdot}7542 = 26{\cdot}2458$$

So the normal equations are

$$17{\cdot}3202\, g_1 - 6{\cdot}6122\, g_2 = 58{\cdot}9321$$

$$-2{\cdot}1493\, g_1 + 26{\cdot}2458\, g_2 = -44{\cdot}3347$$

As there are only two unknowns, the equations have been solved by the usual method of elimination.

The solutions have been obtained as

$$g_1 = 2{\cdot}8466$$

$$g_2 = -1{\cdot}4561$$

Now the adjusted *S.S.* due to generations

$$= g_1Q_1' + g_2Q_2' + (g_1 + g_2)(Q_1' + Q_2')$$

$$= 252{\cdot}6107$$

Hence the adjustment factor

$$= 387373{\cdot}73 - 252{\cdot}61$$

$$= 387121{\cdot}12$$

The analysis of variance table may now be written by subtracting the adjustment factor from the unadjusted *S.S.* due to (*a*) generations, (*b*) colours and (*c*) total *S.S.* and then finding the interaction *S.S.* by subtraction, keeping the error unchanged.

Table of analysis of variance

Sources of variation	*S.S.* unadjusted	d.f.	*S.S.* (adjusted)	M.S.
Between generations ..	387373·73	2	252·61	126·30
Between Colours ..	387821·03	3	699·91	233·30
Int., generations and colours ..	..	6	1206·66*	201·11
Error ..	12014·06	107	12014·06	112·28
TOTAL ..	401294·36	118	14173·24	

* By subtraction.

As it happens none of the mean squares have come out significant. However, to illustrate the computational procedure the estimate of the difference between (i) the first two generations and (ii) the second and third colours and their *S.E.* have been worked out.

The difference between the first two generation effects is evidently

$$g_1 - g_2 = 2 \cdot 8466 - (-1 \cdot 4561)$$

$$= 4 \cdot 3027$$

$$\text{var}\,(g_1 - g_2) = \frac{\sigma^2}{c}\,(P_{11}' + P_{22}' - P_{12}' - P_{21}')$$

where $c = P_{11}'P_{22}' - P_{12}'P_{21}'$

$$= \frac{112 \cdot 28}{440 \cdot 37}\,(17 \cdot 3202 + 26 \cdot 2458 - 6 \cdot 6122 - 2 \cdot 1493)$$

$$= 8 \cdot 8740$$

$$S.E.\,(g_1 - g_2) = 2 \cdot 98$$

To estimate the difference between the effects of the second and the third colours, *i.e.*, $c_2 - c_3$ (say) M_1 and M_2 need be calculated.

Here

$$M_1 = \frac{4}{37} - \frac{40}{46} - (-\,.3367)\,\frac{(109 \cdot 89 - 106 \cdot 67)}{22257 \cdot 87} = -\cdot 7615$$

$$M_2 = \frac{24}{37} - \frac{5}{46} - 5 \cdot 5294\,\frac{(109 \cdot 89 - 106 \cdot 67)}{22257 \cdot 87} = \cdot 5391$$

stimate of $c_2 - c_3 = (\bar{y}_2 - \bar{y}_3) - b(\bar{x}_2 - \bar{x}_3) - g_1(2M_1 + M_2)$

$$- g_2(M_1 + 2M_2)$$

$$= (51 \cdot 78 - 58 \cdot 41) - \cdot 1099 (109 \cdot 89 - 106 \cdot 67)$$

$$- 2 \cdot 8466 (2 \times - \cdot 7615 + \cdot 5391)$$

$$+ 1 \cdot 4561 (- \cdot 7615 + 2 \times \cdot 5391)$$

$$= - 3 \cdot 72$$

$$\text{ar}(c_2 - c_3) = \sigma^2 \left[\frac{1}{n_{\cdot 2}} + \frac{1}{n_{\cdot 3}} + \frac{(\bar{x}_2 - \bar{x}_3)^2}{a} + \frac{1}{c} \{ M_1^2 (2P_{22}' + P_{21}') \right.$$

$$+ M_1 M_2 (P_{11}' + P_{22}' + 2P_{12}' + 2P_{21}')$$

$$\left. + M_2^2 (2P_{11}' + P_{12}') \} \right]$$

After substituting the values and simplification the variance comes to $10 \cdot 2468$ and $S.E. = 3 \cdot 20$.

10. Summary

The paper describes the method of conducting an analysis of covariance with disproportionate cell frequencies in two-way classification. The whole analysis divides itself into two parts, *viz.*, first obtaining the unadjusted sum of squares due to the effects of the factors and their interaction and then calculating an adjustment factor which plays the part of correction factor of the ordinary analysis of variance in getting the adjusted sum of squares. It has been shown in this paper how the adjustment factor can be obtained by solving a set of equations with $p - 1$ unknowns. It has also been described how the estimate and variance of any contrast of the class effects of either of the factors may be obtained without solving any further equation.

Simplified formulæ for the particular cases of $2 \times q$ and $3 \times q$ have been given. The latter has been numerically illustrated. Extension of the results to the case of two concomitant variates has been indicated.

It may also be mentioned that the results of the analysis of variance may be obtained from the corresponding results of analysis

of covariance by putting each of $Q_{i(x)}$, $X_{i\cdot}$, $X_{\cdot j}$ and b equal to zero wherever they occur in the results.

REFERENCES

1. Wilks, S. S. .. *Mathematical Statistics* (Sec. 9.7), Princeton University Press, 1950.
2. ———— .. "Analysis of variance and co-variance of non-orthogonal data," *Metron*, 1938, 13: 141–54.