This work was supported in part by the U.S. Army Research Office - Durham under Contract No. DAHC04-71-C-0042.

A METHOD FOR SEPARATION OF RESIDUAL AND INTERACTION EFFECTS IN CROSS-CLASSIFICATIONS

N. L. JOHNSON

Institute of Statistics Mimeo Series No. 842

Department of Statistics
University of North Carolina at Chapel Hill

September, 1972

A Method for Separation of Residual and Interaction Effects in Cross-Classifications

N. L. JOHNSON
University of North Carolina at Chapel Hill

1. Description of the Problem

Consider a two-way cross-classification with factors R ("rows") and C ("columns") at r, c levels respectively, and suppose that the expected value of an observation at the i-th level of R and the j-th level of C is represented in the form

$$\xi + \rho_i + \kappa_j + (\rho\kappa)_{ij} \quad .$$

The last term represents the interaction effect between R and C, usually denoted in the abbreviated form R×C.

Variability about this expected value is represented by addition of a random variable with zero expected value. It is usually assumed that these random variables are mutually independent and have a common variance, σ^2 say.

If there is only one observation for each of the rc combinations of levels of R and C (as, for example, in a randomized block experiment) it is not possible to estimate σ^2 separately from the interaction effects $(\rho\kappa)_{ij}$.

It would appear that this will also be the case if we have more than one individual in some cells, but observe only the total (or the arithmetic mean) of the values in each cell. There is then only one *observation* in each cell and the situation appears to be similar to that already described.

However, in certain cases of this kind, it is possible to distinguish between residual and interaction effects. We will consider here cases characterized by the following conditions:

(i) the interaction effects are represented by random variables, mutually independent and also independent of the random variables representing residual variation,

(ii) each of the interaction random variables has expected value zero and the same variance, σ'^2, say, and

(iii) The number of individuals is not the same in all cells.

2. Frequencies Constant Within Rows: Two Subsets

2.1 Estimation In this paper, we will consider only cases in which the number of individuals per cell is the same for a given level of C, whatever be the value of R. We first suppose that

(a) for $r_1(\geq 2)$ levels of R there are n_1 individuals for each of the c levels of C, and

(b) for the other $r_2 = r - r_1(\geq 2)$ levels of R there are n_2 individuals for each of the c levels of C. Without loss of generality we suppose $n_1 > n_2$.

If the two sections of the data are analysed separately, we obtain analyses of variance of the form:

	FIRST r_1 LEVELS OF R		REMAINING r_2 LEVELS OF R	
SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	DEGREES OF FREEDOM	SUM OF SQUARES
R	r_1-1	S_{R1}	r_2-1	S_{R2}
C	c-1	S_{C1}	c-1	S_{C2}
R×C	$(r_1-1)(c-1)$	S_{RC1}	$(r_2-1)(c-1)$	S_{RC2}

The sums of squares can all be expressed in terms of the arithmetic means $\bar{X}_{ij}$ of observations for the i-th level of R combined with the j-th level of C (i=1,2,...,r; j=1,2,...,c). In particular,

$$S_{RC1} = \sum_{i=1}^{r_1} \sum_{j=1}^{c} (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j^{(1)} + \bar{X}^{(1)}_{\cdot\cdot\cdot})^2$$

$$S_{RC2} = \sum_{i=r_1+1}^{r} \sum_{j=1}^{c} (\bar{X}_{ij} - \bar{X}_i - \bar{X}_{\cdot j}^{(2)} + \bar{X}_{\cdot\cdot}^{(2)})^2$$

where

$$\bar{X}_i = c^{-1} \sum_{j=1}^{c} \bar{X}_{ij}; \quad \bar{X}_{\cdot j}^{(1)} = r_1^{-1} \sum_{i=1}^{r_1} \bar{X}_{ij}; \quad \bar{X}_{\cdot j}^{(2)} = r_2^{-1} \sum_{i=r_1+1}^{r} \bar{X}_{ij};$$
$$\bar{X}_{\cdot\cdot}^{(t)} = c^{-1} \sum_{j=1}^{c} X_{ij}^{(t)} \qquad (t=1,2).$$

The expected values of the interaction mean squares are

$$E[(r_1-1)^{-1}(c-1)^{-1} S_{RC1}] = n_1^{-1} \sigma^2 + \sigma'^2 ,$$
$$E[(r_2-1)^{-1}(c-1)^{-1} S_{RC2}] = n_2^{-1} \sigma^2 + \sigma'^2 .$$

Hence

(1) $$V' = (n_1-n_2)^{-1}[n_1(r_1-1)^{-1}(c-1)^{-1}S_{RC1} - n_2(r_2-1)^{-1}(c-1)^{-1} S_{RC2}]$$

is an unbiased estimator of σ'^2, and

(2) $$V = n_1 n_2 (n_1-n_2)^{-1}[(r_1-1)^{-1}(c-1)^{-1} S_{RC1} - (r_2-1)^{-1}(c-1)^{-1} S_{RC2}]$$

is an unbiased estimator of σ^2.

It is, of course, possible that these estimators may take negative values.

It is interesting to note that the variance of V' depends only on the ratio n_1/n_2 and not on the actual values of n_1 and n_2. In fact,

(3) $$\mathrm{var}(V') = \left(\frac{n_1}{n_1-n_2}\right)^2 \mathrm{var}(M_{RC1}) + \left(\frac{n_2}{n_1-n_2}\right)^2 \mathrm{var}(M_{RC2})$$

where $M_{RCt} = (r_t-1)^{-1}(c-1)^{-1}S_{RCt}$ $(t=1,2)$ are the interaction mean squares.

The variance of V, on the other hand, for a fixed value of n_1/n_2, is proportional to the actual values of n_1 and n_2. In fact

$$(4) \qquad \mathrm{var}(V) = \left(\frac{n_1 n_2}{n_1-n_2}\right)^2[\mathrm{var}(M_{RC1}) + \mathrm{var}(M_{RC2})]$$

2.2 Tests of Hypothesis of Zero Interaction

The ratio $n_1M_{RC1}/(n_2M_{RC2})$, may be used as a test of the hypothesis $\sigma' = 0$. If, in addition to the other assumptions, it be assumed that all the random variables have normal distributions, then $n_1M_{RC1}/(n_2M_{RC2})$ is distributed as

$$(\sigma^2+n_1\sigma'^2)(\sigma^2+n_2\sigma'^2)^{-1} F_{(r_1-1)(c-1),(r_2-1)(c-1)} .$$

Since $n_1 > n_2$, large values of the observed ratio are regarded as significant of departure from the hypothesis $\sigma' = 0$. For given values of r_1, r_2 and c the power of the test is an increasing function of $(\sigma^2+n_1\sigma'^2)(\sigma^2+n_2\sigma'^2)^{-1}$, and so of σ'/σ. Evidently the sensitivity of the test will be increased if

(a) the absolute magnitudes of n_1 and n_2 are increased (with n_1/n_2 constant)

or (b) the ratio of n_1 to n_2 is increased, either by increasing n_1, or by decreasing n_2.

Note that in case (a) the power cannot exceed

$$(5) \qquad \Pr[F_{(r_1-1)(c-1),(r_2-1)(c-1)} > (n_2/n_1)F_{(r_1-1)(c-1),(r_2-1)(c-1),1-\alpha}]$$

however large are n_1 and n_2 (with n_1/n_2 fixed) or σ'/σ. (Here $F_{r_1,r_2,1-\alpha}$ denotes the upper $100\alpha\%$ point of F_{r_1,r_2}.)

Table 1 gives a few illustrative values of the power of the test, with a significance level of 5%, and $c=5$, $r_1+r_2=8$ in each case.

Table 1: Power of test ($\alpha=0.05$; $c=5$. **Significance level 0.05**)

$\frac{\sigma'^2}{\sigma^2}$	n_1	n_2	$r_1=2,\ r_2=6$	$r_1=4,\ r_2=4$	$r_1=6,\ r_2=2$
0.5	10	5	0.1959	0.2239	0.1225
	50	25	0.2433	0.2866	0.1468
	†2·∞	∞	0.2604	0.3080	0.1555
	20	5	0.4761	0.6049	0.2930
	100	25	0.5643	0.7180	0.3790
	†4·∞	∞	0.5912	0.7495	0.3922
1.0	10	5	0.2227	0.2590	0.1361
	50	25	0.2512	0.2972	0.1509
	†2·∞	∞	0.2604	0.3080	0.1555
	20	5	0.5282	0.6729	0.3352
	100	25	0.5772	0.7336	0.3859
	†4·∞	∞	0.5912	0.7495	0.3922

(† $n_1=k\cdot\infty$, $n_2=\infty$ case evaluated from (5) with $n_2/n_1 = k^{-1}$

3. Frequencies Constant Within Rows - k Subsets

3.1 Estimation We now consider the more general case when the r levels of R can be divided into k subsets of $r_1, r_2, \ldots, r_k$ levels respectively, with each r_j greater than 1, and $\sum_{t=1}^{k} r_t = r$. In the t-th set, containing r_t levels of R, each of the $r_t c$ cells contain n_t observations, and the numbers $n_1, n_2, \ldots, n_k$ are all different. We will suppose the subsets so ordered that $n_1 > n_2 > \ldots > n_k$.

It is now possible to construct k separate analyses of variance analogous to those in Table 1. With a natural extension of the notation in that Table, we have

$$E[n_t M_{RCt}] = \sigma^2 + n_t \sigma'^2$$

where $M_{RCt} = (r_t-1)^{-1}(c-1)^{-1} S_{RCt}$ is the interaction mean square for the t-th subset of levels of R.

The regression of $n_t M_{RCt}$ on n_t is linear. Fitting this regression provides estimates of σ^2 and σ'^2. In the fitting process we should take account of the conditional variance of $n_t M_{RCt}$, given n_t. If no assumption is made about the distribution of the random variables in the model, we do not have a usable formula for this conditional variance. If it is supposed that all cell means have distributions of the same shape then the conditional variance is proportional to $(\sigma^2+n_t\sigma'^2)^2(r_t-1)^{-1}(c-1)^{-1}$. It seems reasonable to proceed on this assumption - especially since each mean is the arithmetic mean of at least two observed values, and so should have a distribution closer to normality then that of the random variables representing the original observations.

This assumption would lead to seeking the values of σ^2 and σ'^2 ($\tilde{\sigma}^2, \tilde{\sigma}'^2$, say) minimizing

$$(6) \qquad G = \sum_{t=1}^{k} (r_t-1)(\sigma^2+n_t\sigma'^2)^{-2}(n_t M_{RCt}-\sigma^2-n_t\sigma'^2)^2$$

If $\sigma'/\sigma = \omega$, say, were known we would obtain the explicit solution

$$\sigma^2 = \frac{\sum_{t=1}^{k} (r_t-1)(1+n_t\omega)^{-2}(n_t M_{RCt})}{\sum_{t=1}^{k} (r_t-1)(1+n_t\omega)^{-1}(n_t M_{RCt})} .$$

If, as will usually be the case, σ'/σ is not known, G may be minimized by an iterative process.

The following formulae are suggested for estimating the variances of the estimators of σ^2 and σ'^2.

$$(7.1) \qquad \mathrm{var}(\tilde{\sigma}^2) \doteq \frac{2}{c-1}\left[\left\{\sum_{t=1}^{k} \frac{r_t-1}{(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2)^2}\right\}^{-1} + \bar{n}^2\left\{\sum_{t=1}^{k} \frac{r_t-1}{(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2)}(n_t-\bar{n})^2\right\}^{-1}\right]$$

$$(7.2) \qquad \mathrm{var}(\tilde{\sigma}'^2) \doteq \frac{2}{c-1}\left\{\sum_{t=1}^{k} \frac{r_t-1}{(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2)^2}(n_t-\bar{n})^2\right\}^{-1}$$

where $$\bar{n} = \left\{\sum_{t=1}^{k} \frac{r_t-1}{(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2)^2} \cdot n_t\right\} \Big/ \left\{\sum_{t=1}^{k} \frac{r_t-1}{(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2}\right\}$$

(These formulae are based on the assumption that the cell means are normally distributed. For platykurtic (leptokurtic) variation the variances would be expected to be rather less (greater) than indicated by formulae (7.1) and (7.2).)

3.2 Test of Zero Interaction

If $\sigma'^2 = 0$, then the expected values of $n_1M_{RC1}, n_2M_{RC2},\ldots,n_kM_{RCk}$ are all equal, and the slope of the regression of n_tM_{RCt} on n_t is zero.

Assuming normal variation, the hypothesis that $\sigma' = 0$ could be tested by applying the Neyman-Pearson-Bartlett test of equality of variances to the statistics $n_1M_{RC1},\ldots,n_kM_{RCk}$. Apart from the well-known sensitivity of this test to non-normality, it has another drawback in this context (although this does not affect its validity). It neglects the fact that any specific alternative hypothesis $(\sigma'/\sigma \neq 0)$ gives a specific pattern of values for the ratios

$$(\sigma^2+n_1\sigma'^2):(\sigma^2+n_2\sigma'^2): \ \ldots \ :(\sigma^2+n_k\sigma'^2) \ .$$

It seems likely, therefore, that a test of the significance of the slope of the fitted regression line (of n_tM_{RCt} on n_t) will be more powerful. This is, indeed, quite natural, since this slope is, in fact, just σ'^2.

An approximate test can be effected by comparing $\tilde{\sigma}'^2/[\mathrm{var}(\tilde{\sigma}'^2)]^{\frac{1}{2}}$ with a unit normal scale. Formula (7.2) may be used to estimate $\mathrm{var}(\tilde{\sigma}'^2)$.

3.3 Test of Constancy of σ'

There may be reason to suspect that the values of σ and/or σ' may change from one subset of levels of R to another. There are numerous possibilities, and we will not attempt to explore them here. However, a quick and simple test which will detect some types of heterogeneity in σ and/or σ' is provided by testing the regression of $n_t M_{RCt}$ on n_t for linearity. It is possible to construct such a test (on an approximate basis, and assuming normal variation) using the fact that the variance of $n_t M_{RCt}$ is $2(r_t-1)^{-1}(c-1)^{-1}(\sigma^2+n_t\sigma'^2)$. If there is no heterogeneity in σ^2 and/or σ'^2,

$$\tfrac{1}{4}(c-1)\sum_{t=1}^{k}(r_t-1)(\tilde{\sigma}^2+n_t\tilde{\sigma}'^2)^{-2}(n_t M_{RCt}-\tilde{\sigma}^2-n_t\tilde{\sigma}'^2)^2$$

should be approximately distributed as χ^2 with (k-2) degrees of freedom. Large values of the statistic would be regarded as significant of heterogeneity.

4. A More General Case

It is not essential that the subsets of levels of R should each have the same frequency of observations in each cell. Retaining the assumption that any *one* level has the same cell-frequencies for all levels of C, let us denote by n_{th} the number of observations per cell for the h-th level of the t-th subset of levels of R($h=1,2,\ldots,r_t$).

We then use, in place of M_{RCt}, the statistic

$$M'_{RCt} = \sum_{h=1}^{r_t} n_{th} \sum_{j=1}^{c} (\bar{X}^{(t)}_{hj}-\bar{X}^{(t)}_{h\cdot}-\bar{X}^{(t)}_{\cdot j}+\bar{X}^{(t)}_{\cdot\cdot})^2/\{(r_t-1)(c-1)\}$$

where $\bar{X}^{(t)}_{hj}$ is the arithmetic mean of observations for the h-th level of the t-th subset of levels of R, combined with the j-th level of C, and

$$\bar{X}_{h\cdot}^{(t)} = c^{-1} \sum_{j=1}^{c} \bar{X}_{hj}^{(t)}; \quad \bar{X}_{j}^{(t)} = N_t^{-1} \sum_{h=1}^{r_t} n_{th} \bar{X}_{hj}^{(t)}; \quad \bar{X}_{\cdot\cdot}^{(t)} = (cN_t)^{-1} \sum_{h=1}^{r_t} n_{th} \sum_{j=1}^{c} \bar{X}_{hj}^{(t)}$$

with $N_t = \sum_{h=1}^{r_t} n_{th}$.

The expected value of M'_{RCt} is $\sigma^2 + n'_t \sigma'^2$ with

$$(8) \qquad n'_t = (r_t - 1)^{-1}\{N_t - [\sum_{h=1}^{r_t} n_{th}^2]/N_t\}.$$

Similar methods to those described in Section 3 can be used, with n_t replaced by n_t. It must, however, be realized that even assuming normal theory, M'_{RCt} is no longer distributed as a multiple of a χ^2.