CORRELATIONS BETWEEN SUMS OF SQUARES AND F STATISTICS IN THE FIXED MODEL FOR UNBALANCED DESIGNS

John M. Jordan & G. Rex Bryce, Brigham Young University

For any analysis of variance model it is well known that the sums of squares of the main effects and the interactions are independent if the design is balanced. However, this independence does not necessarily follow for all pairs of sums of squares when the design is unbalanced, i.e., unequal cell frequencies. Since the latter is more likely to occur in practice, it would be useful to know the consequences of various degrees of imbalance on correlations between the sums of squares.

In the fixed model all the F ratios are correlated to some degree by virtue of a common denominator, the mean square for error. If the design is unbalanced and the sums of squares are fully adjusted, i.e., each sum of squares is obtained when its term is ordered last in the model, then the F ratios are also correlated through their numerators. As a result the tests are not independent and the joint probability of one or more Type I errors is a function of the strength of this correlation.

The Theory

This study depends on the representation of the analysis of variance model as a linear regression model by the use of the dummy variable technique within the cell means format. Define the cell means model in matrix format as

$$\underline{y} = X\underline{\mu} + \underline{e}$$

where $\underline{y}$ is an $n \times 1$ vector of responses, X is an $n \times p$ matrix of ones and zeroes indicating the location of the responses by cell, $\underline{\mu}$ is a $p \times 1$ vector of cell means and $\underline{e}$ is the $n \times 1$ vector of residuals.

We reparameterize this cell means model by the introduction of a full rank $p \times p$ matrix of contrasts, C, chosen to be of interest to the researcher. Write this new model as

$$\underline{y} = XC^{-1}C\underline{\mu} + \underline{e} = Z\underline{\beta} + \underline{e}$$

so that $\underline{\beta} = C\underline{\mu}$ is the set of coefficients dictated by the analyst's chosen set of linear combinations of the cell means and $Z = XC^{-1}$ is a "stretched" version of the inverse of the matrix of contrasts. We will assume throughout this study that all elements of $\underline{\beta}$ are known or specified as measures of the strength of particular contrasts defined on the terms in the model. For a 2×2 model in the usual ordering

$$\underline{\beta}' = \left[\beta_\mu \ \beta_A \ \ \beta_B \ \beta_{AB}\right] .$$

The purpose of this move from one full rank formulation to another should be clear. The parameters in both models are uniquely estimable; however, the original cell means model just estimates the cell means whereas the reparameterized model provides information which allows the kind of comparison of treatment effects needed by the researcher.

Now, if we were interested in estimating $\underline{\beta}$, we could obtain a vector of solutions to this system as

$$\hat{\underline{\beta}} = (Z'Z)^{-1}Z'\underline{y}$$

in the usual manner of multiple linear regression. However, for this study we are interested instead in a factored version of the $Z'Z$ matrix augmented by $Z'\underline{y}$ as a means of obtaining expressions for the sums of squares of the terms in the analysis of variance model. A Cholesky factorization of this augmented matrix can be obtained by premultiplication by a Cholesky operator in the form

$$R\left[Z'Z \mid Z'\underline{y}\right] = R\left[Z'Z \mid \underline{z}\right] = \left[T \mid \underline{t}\right]$$

and it will be useful to us to partition these matrices in order to write exact expressions for the sums of squares. If, for example, the interaction is ordered last in a model with two main effects, then the last partition in the $\underline{t}$ vector, $\underline{t}_{AB}$, will be the product of the elements in the partitioned last row of the R operator matrix with the elements of the $\underline{z}$ vector. The number of rows in $\underline{t}_{AB}$ will depend on the number of degrees of freedom associated with the interaction term.

If we symbolize this partitioned last row of R as R'_{AB} in this ordering and the matrix Z in the same ordering as Z^{AB}, then we have

$$\underline{t}_{AB} = R'_{AB}Z'^{AB}\underline{y}$$

therefore

$$\underline{t}'_{AB}I\underline{t}_{AB} = \underline{y}'Z^{AB}R_{AB}R'_{AB}Z'^{AB}\underline{y} \ .$$

This last formula is a quadratic form for the fully adjusted sum of squares for this interaction (SSAB) for any design, whether balanced or not. The same procedure could be used to find a general expression for the fully adjusted sum of squares for any term in the model by a simple reordering of the matrices. Having these, we can use well known theorems on the variance and covariance of quadratic forms, as in Searle (1971), to obtain the correlation between sums of squares for any design.

If we have $\underline{x} \sim N(\underline{\eta},V)$, then

$$\text{Cov}(\underline{x}'Px,\underline{x}'Qx) = 2\text{tr}(PVQV) + 4\underline{\eta}'PVQ\underline{\eta} \ .$$

Since our response vector is normally and independently distributed with $\underline{\eta} = Z\underline{\beta}$ and $V = \sigma^2 I$, we can use matrix P as defined in the quadratic expression for SSAB and a corresponding Q defined for main effect B ordered last to write

$$\text{Cov}(\text{SSB},\text{SSAB}) = 2\sigma^4\text{tr}(Z^{AB}R_{AB}R'_{AB}Z'^{AB}Z^{B}R_{B}R'_{B}Z'^{B})$$

$$+ 4\sigma^2\underline{\beta}'^{AB}Z'^{AB}Z^{AB}R_{AB}R'_{AB}Z'^{AB}Z^{B}R_{B}R'_{B}Z'^{B}Z^{B}\underline{\beta}^{B}$$

where the superscripts on $\underline{\beta}$, as usual, refer to a particular ordering of the elements in this vector. Observe that the second term in the covariance is the non-centrality portion of the expression since the whole term becomes zero if $\underline{\beta} = \underline{0}$.

Recall that the trace is invariant when cyclic operations are performed on its argument. With this in mind, we have found it convenient to rewrite this portion of the first term in the covariance as

$$\mathrm{tr}(R_{AB}R'_{AB}Z'^{AB}Z^{B}R_{B}R'_{B}Z'^{B}Z^{AB})$$

for which the dimensions are p x p rather than n x n , where p is the number of cells in the design and n is the number of data points in the sample.

Before we can compute the correlation between SSAB and SSB, we need expressions in a similar form for the variance of each sum of squares. Searle (1971, p. 57) shows the general form for the variance of a quadratic form as

$$\mathrm{Var}(\underline{x}'P\underline{x}) = 2\mathrm{tr}(PV)^2 + 4\underline{\eta}'PVP\underline{\eta}$$

where, once again, $\underline{x} \sim N(\underline{\eta},V)$. Proceed as before to write the variance of (say) the interaction sum of squares as

$$\mathrm{Var(SSAB)} = 2\sigma^4\mathrm{tr}(Z^{AB}R_{AB}R'_{AB}Z'^{AB})^2$$

$$+ 4\sigma^2\underline{\beta}'^{AB}Z'^{AB}(Z^{AB}R_{AB}R'_{AB}Z'^{AB})^2Z^{AB}\underline{\beta}^{AB} \quad .$$

Again, the argument of the trace can be cycled for a more manageable matrix product.

The Application to 2 x 2 ECART Designs

Interested readers can verify, using a simple balanced design, that the formulas described above give a well defined zero correlation between any pair of sums of squares, i.e., zero covariance and non-zero variance.

However, our main interest is in unbalanced designs. In order to systematically study the effect of imbalance on correlation between sums of squares, we chose the ECART pattern (equal column and row totals) as a point of departure. This design is attributed to Powers and Herr (1975), who used it to study the partial confounding of row and column effects which occurs in unbalanced designs because of a lack of orthogonality in the design. Using this pattern we describe the cell frequencies in a 2 x 2 model as

	B_1	B_2
A_1	n - k	n + k
A_2	n + k	n - k

Then we can obtain the matrix product $Z = XC^{-1}$ for our reparameterized model, with X a 4n x 4 array of ones and zeroes corresponding to data location and C identified by the term ordered last in this matrix of contrasts.

For example, when interaction is ordered last, we obtain

$$Z^{AB} = X(C^{AB})^{-1} = X\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

from which $\underline{R}'_{AB} = [a \quad 0 \quad 0 \quad b]$

where $a = \frac{k}{2n(n^2-k^2)}$ and $b = \frac{n}{2n(n^2-k^2)}$.

The Z matrix for main effect B ordered last is obtained by a reordering of the columns of the Z^{AB} matrix. If this is done, then the inverse of the transpose of the factored matrix yields a last row

$$\underline{R}'_{B} = [0 \quad a \quad 0 \quad b] \quad .$$

When these results are used to assemble the matrix product needed in the expression for the covariance between SSAB and SSB, the value of this product is zero. Since Var(SSAB) and Var(SSB) can be shown to be non-zero, we conclude that the correlation between SSAB and SSB is zero in the ECART design. The same result is obtained for the correlation between SSAB and SSA for this design.

However, the correlation between the sums of squares for the main effects in the ECART pattern is non-zero. We proceed directly to write the trace in the first term of the covariance between SSA and SSB as

$$2\sigma^4\mathrm{tr}(\underline{R}_A\underline{R}'_A Z'^{A}Z^{B}\underline{R}_B\underline{R}'_B Z'^{B}Z^{A}) = 2\sigma^4 k/n \ .$$

For the second covariance term, the non-centrality portion, we can reduce the matrix product to

$$16k\sigma^2\beta_A\beta_B(1-k^2/n^2)$$

so, as expected, the non-centrality term in the covariance between SSA and SSB is a function, in part, of the contrasts written on these terms. It also depends, as does the first term in the covariance, on the amount of imbalance and on σ^2 .

Before correlation can be computed the variances of the sums of squares must be obtained. The matrix products needed here are easily deduced from our earlier work on the covariance terms. Since the algebra is very much the same as it was in finding the covariance, we will just assert that

$$\mathrm{Var(SSA)} = 2\sigma^4 + 16n\sigma^2\beta^2_A(1-k^2/n^2)$$

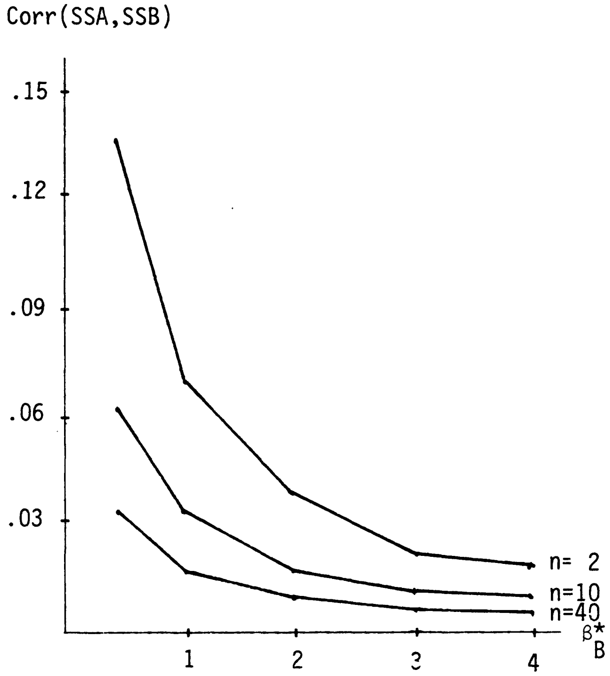
and the parallel expression can be obtained for Var(SSB).

We now have all the ingredients needed to write a general expression for the correlation between SSA and SSB for the ECART design. Before doing so, we will simplify the terms by defining w = k/n as a measure of imbalance in the design. Also, for later ease in handling the contrasts on the main effects, convert them to standardized form as $\beta^*_A = \beta_A/\sigma$ and $\beta^*_B = \beta_B/\sigma$. Then the correlation becomes

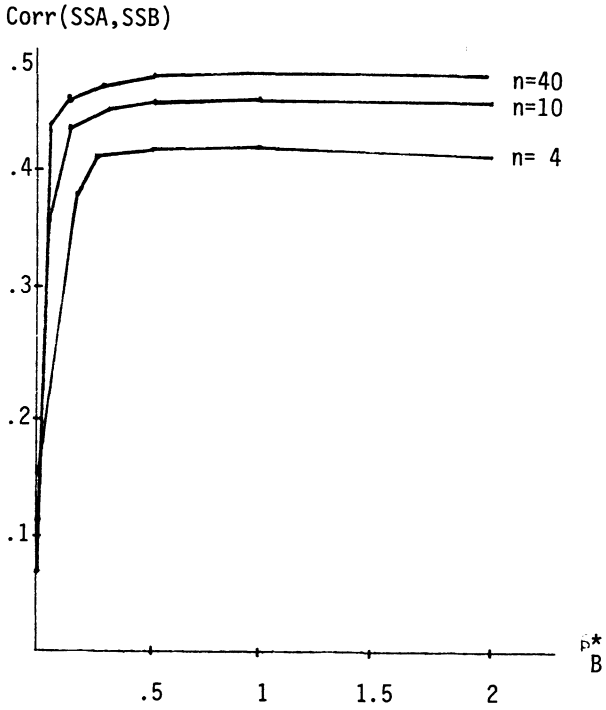
$$\mathrm{Corr(SSA,SSB)} = \frac{w^2 + 8k\beta^*_A\beta^*_B(1-w^2)}{\sqrt{\left[1+8n\beta^{*2}_A(1-w^2)\right]\left[1+8n\beta^{*2}_B(1-w^2)\right]}}$$

Under the null hypothesis, $\beta^*_A = \beta^*_B = 0$, the correlation between the main effects is equal to w^2 and so depends just on the measure of imbalance. Remember that n here is the average number of observations per cell; the total sample size is 4n . We conclude that, under this assumption, a small amount of imbalance in the ECART pattern, perhaps occurring fortuitously in a large sample, causes little correlation between SSA and SSB, whereas any amount of imbalance in a smaller sample will result in enough correlation that we would conclude important dependence exists between the main effects sums of squares and, therefore, between the F ratios used to test the main effects.

When one contrast, say β_A^* , is zero and the other is positive, the correlation between SSA and SSB decreases for increased sample size and for increased strength of the non-zero contrast, given a certain level of imbalance in the design. The values of this correlation are shown for selected cases in the figure below, when w = .5.



For both main effects non-zero, inspection of the general form for correlation between SSA and SSB in the ECART design will confirm that the second terms in the covariance and each of the variances will dominate the expression for most values of the parameters. Written with just these second terms, Corr(SSA,SSB) simplifies to approximately w = k/n , the measure of imbalance for a wide range of values of the parameters. The figure below shows this behavior for $\beta_A^* = .25$ over various levels of β_B^* for several sample sizes given w = .5 . The correlation begins to drop away from w as β_B^* increases but this decrease is small. We note in passing that the correlations are not affected if the interaction is ignored in obtaining main effects sums of squares rather than our fully adjusted approach.



The Application to Other 2 x 2 Designs

If the cell frequencies are identified only by location, i.e.,

	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

Then the Z'Z matrix can still be written in a simple form but the factored matrix is unworkable so it is not possible to write an expression for the correlation between sums of squares as a function of the useful parameters. Instead we must work directly with the matrix products which define covariances and variances for quadratic forms.

To preserve order in our consideration of the generalized 2 x 2 design, we will view new patterns as departures from the ECART design, whose characteristics are known to us. For an example, we will explain the case where $n_{..} = n_{11} + n_{12} + n_{21} + n_{22} = 12$. We will not consider patterns in which missing cells occur so the only two ECART designs are 2,4,4,2 and 1,5,5,1.

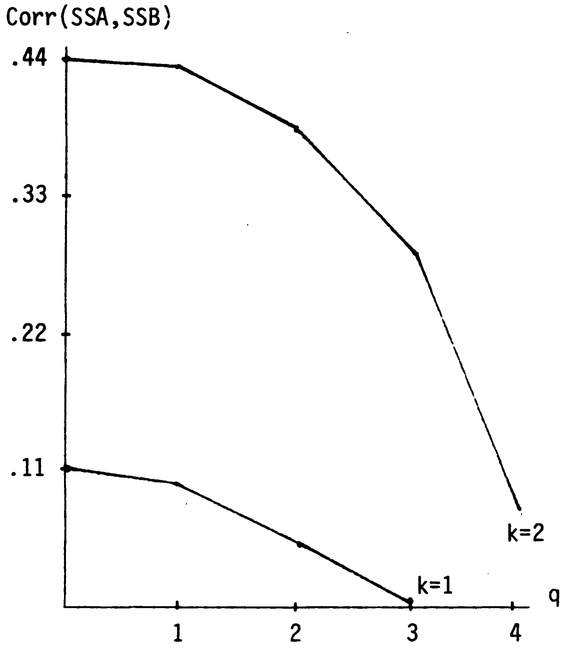
We know that the 2,4,4,2 ECART has n = 3 and k = 1 , therefore Corr(SSA,SSB) is .1111 when we assume the null hypotheses are true. Under the same null assumption, Corr(SSA,SSB) is .4444 for the 1,5,5,1 ECART.

Now, view departures from these patterns as functions of a new parameter q where, in the ECART design, q is used to modify the frequencies in the "n + k" cells:

n-k	n+k-q
n+k+q	n-k

Observe that q = 1 changes a 2,4,4,2 ECART to a 2,3,5,2 pattern but it changes a 1,5,5,1 ECART to a 1,4,6,1 pattern.

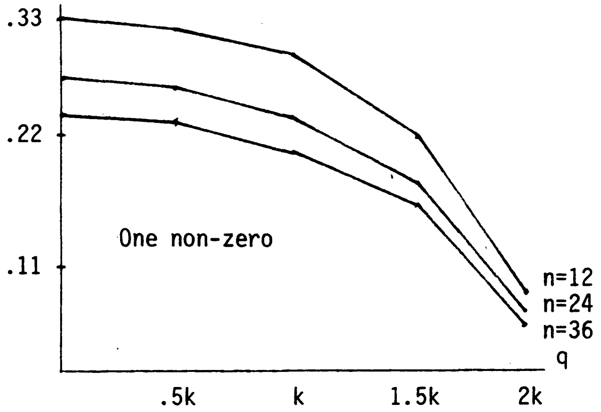
We will consider only the case in which the null hypotheses are true. Under this assumption, i.e., $\underline{\beta}^{*A} = [0\ 0\ 0\ 0]$ and $\underline{\beta}^{*B} = [0\ 0\ 0\ 0]$, Corr(SSA,SSB) decreases for departures from the ECART design and increases for departures from our proportional designs, as seen in the figure below. This agrees with the Powers and Herr (1975) conclusion that the ECART design is the most non-orthogonal design. From these starting points, the correlations for the departures from this design move in predictable directions for increasing values of q .

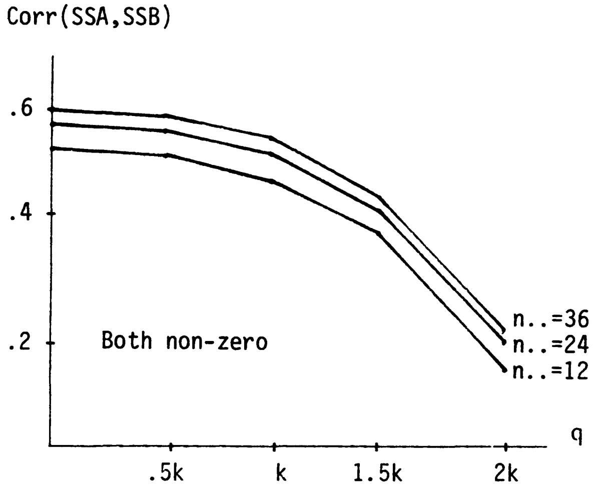


Finally, for the 2 x 2 design, we will briefly consider the change in Corr(SSA,SSB) which occurs as total sample size increases from n.. = 12 to n.. = 24 and n.. = 36 . The number of possible patterns increases greatly with larger sample sizes; we will study only the effect of the larger sample by discussing those patterns in which the ratios of k to n and q to k are the same as they were for the n.. = 12 cases where n = 3 and k = 2 .

Given these conditions on the larger sample sizes, Corr(SSA,SSB) remains the same when the null hypotheses are assumed true. This is the case where the contrasts are all zero so the second terms in the covariance and the variances are also zero. The unchanged correlation means that the change in the variances of the sums of squares as sample size increases is exactly offset by the change in covariance between the sums of squares. Given the fixed k to n ratio, this confirms the result found for the ECART design and extends it to departures caused by non-zero values of q.

However, the behavior of Corr(SSA,SSB) when sample size is increased and one or both of the standardized contrasts on the main effects is non-zero is not as easy to explain, since the rise in n.. causes the correlation to drop in one case and rise in the other. Refer to the figures below to see that the correlation falls for increased n.. over the range of q when one contrast is zero and the other is positive.

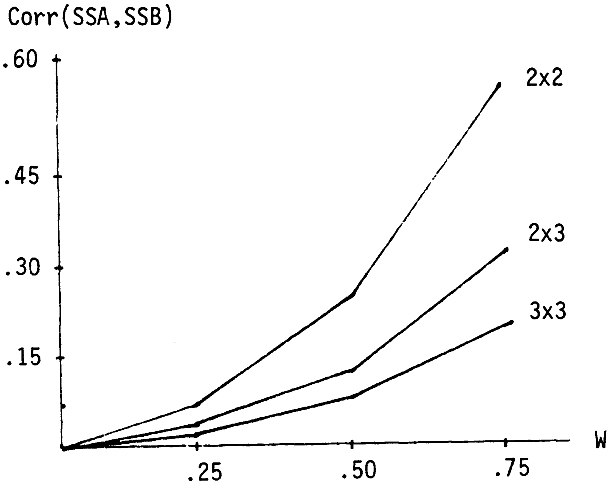




When both contrasts are positive the correlation rises with increased sample size. It would seem that, for both contrasts positive, we are denied the reward of lower correlation for a larger sample. But this assumes that orthogonality is an unmitigated virtue where a little reflection will confirm that it is not. If both null hypotheses about the main effects are, in fact, false then it is better to have high correlation between SSA and SSB so that the chance of rejecting both null hypotheses is enhanced and thus there is a pay-off from the larger sample,just as there is in the other case, where one null is true and the other is not, so the payoff from the larger sample comes in the form of lower correlation between the sums of squares of the main effects.

Application to Larger Designs

Since the determination of the number of levels for a categorical treatment is sometimes arbitrary, e.g., low, medium and high versus just low and high, we have studied to what extent the choice of number of levels affects the correlation between sums of squares. To isolate this influence we studied a fixed sample size, n.. = 36, for fixed amounts of imbalance in the ECART design for three different assumptions about the number of levels per treatment: the 2×2 , the 2×3 and the 3×3 designs. As is shown in the figure below, the level of correlation decreases as the size of design increases, other things being equal, given the null hypotheses true.



This indicates that the larger versions of the ECART design are more orthogonal than the smaller versions with the same nominal level of imbalance. This is probably due to the presence in the larger designs of some cells with frequency of "n" which are unaffected by the parameter k .

Correlation Between F Ratios

When simultaneous F tests are done on the same data set in a fixed analysis of variance model, the F ratios are dependent because they have the same denominator, the error mean square. For a balanced design, the numerator sums of squares are independent, so all the dependence in the F ratios is caused by their common denominator. Hurlburt and Spiegel (1976) evaluated the conditional probability that both F ratios are significant given that one F ratio is significant under the assumption of independent numerator sums of squares. For unbalanced designs, the F ratios are also correlated through their numerators in many cases and our results so far enable us to measure this correlation since it depends, in part, on the covariance between sums of squares. This was done, under the assumption of true null hypotheses, for pairs of F ratios with independent numerators and for pairs with dependent numerators. These correlations will be used to approximate the joint probability that one or other or both F ratios are significant, i.e., that at least one Type I error is committed in the simultaneous tests.

Since the F statistic is a ratio of independent χ^2 statistics divided by their degrees of freedom, let x be distributed as a χ^2 with "a" degrees of freedom, let y be distributed as a χ^2 with "b" degrees of freedom and let z be distributed as a χ^2 with "c" degrees of freedom. We will use z as the denominator χ^2 distribution in defining a pair of F ratios, so that x and y will both be assumed independent of z but not independent of each other. Then, by definition, the general expression for the covariance between two F ratios is

$$\mathrm{Cov}(F_x,F_y) = \mathrm{Cov}\left[\frac{x/a}{z/c}, \frac{y/b}{z/c}\right]$$

$$= E\left[\frac{xy/ab}{(z/c)^2}\right] - E\left[\frac{x/a}{z/c}\right] E\left[\frac{y/b}{z/c}\right]$$

$$= \frac{c^2}{ab}\left[E(xy)E(z^{-2}) - E(x)E(y)E^2(z^{-1})\right]$$

but we can obtain

$$E(xy) = \mathrm{Cov}(x,y) + E(x)E(y)$$

and we know that the expected values of these random variables are just equal to their degrees of freedom if their distributions are centrally χ^2 . This will always be the case for the denominator variable z in our F statistic; we will be concerned with the joint probability of Type I errors made when the null hypotheses are true, so the numerator random variables will also have central χ^2 distributions.

Expected values of powers of these random variables can be obtained using the expectation of the gamma distribution. In this way, it can be shown that

$$E(z^{-2}) = \frac{1}{(c-2)(c-4)}$$

for $c > 4$, and also that

$$E(z^{-1}) = \frac{1}{c-2}$$

for $c > 2$ by the same approach.

Given all of this we have been able to show that

$$\mathrm{Corr}(F_x,F_y) = \frac{(c-2)\mathrm{Cov}(x,y) + 2ab}{2\sqrt{ab(a+c-2)(b+c-2)}}$$

for $c > 4$ and $a,b > -4$. The only operable restriction on this expression is that the number of degrees of freedom for the error sum of squares must be greater than 4 .

In the application to a 2×2 design this expression becomes, in terms of the covariance between main effect sums of squares, for the F ratios used to test these main effects:

$$\mathrm{Corr}(F_A,F_B) = \frac{(4n-6)\mathrm{Cov}(SSA,SSB) + 2\sigma^4}{2(4n-5)\sigma^4}$$

for a design with 4(n-1) degrees of freedom for error, such as the 2 x 2 ECART. When the Cov(SSA,SSB) for this size ECART design is substituted into this formula, under the assumption that the null hypotheses are true, we obtain

$$\mathrm{Corr}(F_A,F_B) = w^2 + \frac{1-w^2}{4n-5}$$

which approaches w^2 assymptotically as sample size becomes large, the same value found earlier for Corr(SSA,SSB) for the 2 x 2 ECART design under the same assumption.

For the generalized 2 x 2 design, the w^2 term is replaced by

$$\mathrm{tr}(Z^A \underline{R}_A \underline{R}'_A Z'^A Z^B \underline{R}_B \underline{R}'_B Z'^B) \quad .$$

The argument of the trace is one that we have seen before in calculations for the correlation between sums of squares.

Equivalent expressions for larger designs can be written as well. For a 2 x 3 size, the error sum of squares will have 6(n-1) degrees of freedom, again using n in our context as an average cell frequency, and the correlation becomes

$$\mathrm{Corr}(F_A,F_B) = \frac{(3n-4)\mathrm{tr}(M) + 1}{\sqrt{3(6n-7)(n-1)}}$$

where the argument M of the trace is the same as used above. If the same is done for the 3 x 3 design, the formula is

$$\mathrm{Corr}(F_A,F_B) = \frac{(9n-11)\mathrm{tr}(M) + 3}{18(n-1)}$$

The Overall Alpha Level for a Pair of F Tests

If the null hypotheses are both true in a pair of F tests, then an alpha error is committed whenever either one or other or both of the null hypotheses are rejected. We are interested in the joint probability that an alpha error will occur in some way, which will be the complement of the event that both decisions are correct, written for the main effects in a 2 x 2 model as

$$1 - \Pr(\text{accept } \beta_A = 0 \text{ and accept } \beta_B = 0) \quad .$$

This probability is described by a central bivariate F distribution for which we have been unable to find a closed form in the literature in the general class when the numerator mean squares are correlated. Hurlburt and Spiegel (1976) evaluated the integral of the bivariate F distribution when the mean squares are independent but extending their results is not straightforward. We have worked with an approximation to probabilities from the bivariate F distribution suggested by Johnson and Kotz (1972, p.242). The results of these calculations, using as input the correlations between pairs of F ratios for tests on the main effects found earlier in this study, indicate that the overall alpha level reacts strongly to imbalance in the design on the order of w = .5 or higher by our measure of imbalance.

We are now attempting to evaluate the bivariate F distribution directly for cases of correlated numerator mean squares in order to obtain precise measures of the overall alpha level. If the series expression for this distribution converges rapidly then accurate values for this probability should be possible.

The foregoing is a part of a much more detailed study found in Jordan (1976).

References

Hurlbert, Russel T. and Douglas K. Spiegel. "Dependence of F Ratios Sharing a Common Denominator Mean Square," The American Statistician, 30 (1976), 74-78.

Johnson, N.L. and S. Kotz. Distribution in Statistics: Continuous Multivariate Distributions. Volume IV. New York City: John Wiley and Sons, Inc., 1972.

Jordan, J.M. (1976). "Correlations Between Sums of Squares and F-Statistics in the Fixed Model for Unbalanced Designs." Unpublished Masters Thesis, Brigham Young University, Provo, Utah.

Powers, William A. and David G. Herr. "A Monte-Carlo Comparison of Error Probabilities of Five ANOVA Methods in Unbalanced, Two Way Designs." Unpublished Paper, University of North Carolina (Greensboro), 1975.

Searle, S.R. Linear Models. New York City: John Wiley & Sons, Inc., 1971.