Power Comparison of the Weighted Squares of Means Method (Type III) and the Method of Fitting Constants (Type II) in Unbalanced ANOVA.

by

Ramon C. Littell and Richard O. Lynch
University of Florida

1. Introduction. This paper deals with the comparison of the so-called Types II and III sums of squares (SS's) that are computed by the GLM procedure in SAS. We specifically address the two-way classification with factors A and B, both fixed, with a levels of A and b levels of B. It is assumed that we have n_{ij} observation y_{ijk}, $k=1,\ldots,n_{ij} > 0$, from the population corresponding to levels i of A and j of B, i=1,,..,a and j=1,...,b. Let μ_{ij} denote the mean of the ij population. We assume $V(y_{ijk}) = \sigma^2$, independent of i and j, and that all y_{ijk} are mutually independent and normally distributed. We may then write $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$, where the ϵ_{ijk} are independent and normally distributed with variance σ^2. Defining $\mu = \bar{\mu}_{..}$, $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$, $\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}$, and $\gamma_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$ we have

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where $\alpha_. = \beta_. = \gamma_{i.} = \gamma_{.j} = 0$, i=1,...,a, j=1,...,b, $k=1,\ldots,n_{ij}$. The parameters α_i, β_j, and γ_{ij} measures A main effect, B main effect, and A*B interaction, respectively.

In general terms, we are concerned with testing whether there is a main effect of factor A. Specifically, we wish to test H_0^A: $\alpha_1 = \ldots = \alpha_a = 0$, or equivalently H_0^A: $\bar{\mu}_{1.} = \ldots = \bar{\mu}_{a.}$. Note that we are considering a main effect test in the possible presence of interaction. In many applications this may be irrelevant; it may be more meaningful to analyze simple effects. There are, however, situations in which main effects are important even with interaction, e.g., if b is too large to feasibly separate all simple effects, if practical considerations require a single overall judgement, or if interaction is small relative to main effects.

Several papers have addressed the topic of hypothesis testing in unbalanced data, notably Yates (1934) in the early years. Textbook discussions are offered by Bancroft (1968), Searle (1971), Steel and Torrie (1960, 1980), and elsewhere. The topic has been of considerable recent interest, due largely to the variety of computational procedures available in computer packages such as SAS. Many papers have been presented at past SUGI meetings, including, among others, Speed and Hocking (1979), Searle (1979, 1981), and Hocking (1981). Survey papers are given by Speed, Hocking, and Hackney (1978) and Searle, Speed, and Henderson (1981). A primary focus of these writings has concerned the concept of the "hypotheses tested" by the F-tests based on the various sums of squares in an analysis of variance. The term "hypotheses tested" can be explained as follows: An F-statistic in a fixed effects model has a noncentral F distribution. The "hypothesis tested" by an F-test is a set of equations, linear in the parameters, that are true if and only if the noncentrality parameter is equal to zero. Thus the F-statistic has a central chi-square distribution when the

"hypothesis tested" is true. Hence a critical value F_α can be determined so that the rejection probability is equal to the prescribed α when the hypothesis tested is true and greater than α when the hypothesis tested is false. Put another way, an F-statistic tests H_0 if the F-test is unbiased for H_0. Conversely, an F-statistic does not test H_0 if it is biased for H_0; that is, if the rejection probability can exceed the prescribed α when H_0 is true.

Speed, Hocking, and Hackney (1978) presented the linear functions of cell means that are tested by the Types II and III F-tests in SAS, as well as several other types. In particular, they stated that the Type III F-test does test H_0^A: $\bar{\mu}_{1.} = \ldots = \bar{\mu}_{a.}$, but that the Type II F-test does not test H_0^A. This same point is made in somewhat different terminology by Bancroft (1968) and Searle (1971).

We now turn to discussion of the types II and III SS's in more detail. To do so, consider the four models:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \alpha_. = \beta_. = \gamma_{.j} = \gamma_{i.} = 0 \qquad (1)$$

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad \alpha_. = \beta_. = 0 \qquad (2)$$

$$y_{ijk} = \mu + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \beta_. = \gamma_{.j} = \gamma_{i.} = 0 \qquad (3)$$

$$y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}, \quad \beta_. = 0 \qquad (4)$$

Let $SS_{E,\ell}$ and $df_{E,\ell}$ denote, respectively, the residual sums of squares and degrees of freedom for the four models, ℓ=1,2,3,4. The Types II and III SS's for factor A are then expressible in the usual "R" notation, as

$$SS_A(II) = R(\alpha|\mu,\beta) = SS_{E,4} - SS_{E,2} \qquad (5)$$

and

$$SS_A(III) = R(\alpha|\mu,\beta,\gamma) = SS_{E,3} - SS_{E,1} \qquad (6)$$

(Note: Our "R" notation here corresponds to the Σ-restricted reduction R* notation of Searle (1981).)

The type II SS is sometimes referred to as being obtained from the "method of fitting constants". The type III SS is known to be equivalent to the "weighted squares of means" SS computing procedure. (See Searle et al., 1981.) If all n_{ij} are equal, then $SS_A(II) = SS_A(III)$; otherwise inequality holds, generally.

Referring to the R notation, we derive the following guidelines:

1) In the absence of interaction, use Type II

2) In the presence of interaction and assuming H_0^A is of interest, use Type III.

The choice between Types II and III, therefore, hinges on the presence or absence of interaction. In application, however, the data analyst rarely knows whether interaction is present. In fact, interaction is probably always present, at least in minute amount. A common practice, which is suggested by Bancroft (1968), is to first of all test for significance of interaction at, say, the α = .25 level. If interaction is not significant, use the Type II test. If interaction is significant (and main effects are still of interest), use the Type III test.

As noted, if interaction is present, then the type II test is biased for testing

H_0^A: $\bar{\mu}_{1.} = \ldots = \bar{\mu}_{a.}$ in the sense that the probability of rejecting H_0^A will be larger than the prescribed α; i.e. the type 1 error rate is inflated, and the amount of inflation depends on the magnitude of the interaction. For this reason it is said that the type II F does not test H_A. Further, the probability of rejecting H_0^A when H_0^A is false (the power of the test) depends on the interaction, and this dependency can be either an increase or a decrease in power.

The type III test always gives an unbiased test of H_0^A, in that the rejection probability is equal to α under H_0^A regardless of the amount of interaction. Therefore, it is said that the type III F does test H_0^A. However, as noted e.g. by Steel and Torrie (1960) and Bancroft (1968), if interaction is absent, then the type III test is inefficient. Specifically, the power of the type III test is less than the power of the type II test in the absence of interaction, due to the superfluous estimation of zero-valued interaction parameters. The practical issue when small amounts of interaction are known or suspected to exist becomes a trade-off between a small bias in the type II test and a small power deficiency in the type III test.

Whereas the issue of what hypothesis is being tested by a given SS has been extensively discussed, relatively much less has been written about the amount of bias (error rate inflation) in the type II test when interaction is present. Relatively little has also been reported on the degree of inefficiency of the type III test when interaction is absent. Gosslee and Lucas (1965) make limited power computations for both the types II and III tests. Overall, Lee, and Hornick (1981) and Cramer and Appelbaum (1980) debate related issues primarily regarding estimation.

The purpose of the present paper is to assess rejection probabilities of the types II and III F-tests in SAS. Specifically, we make computations for:

i. The amount of inflation in the type 1 error rate of the type II F-test in the presence of interaction.

ii. The loss in power (inefficiency) of the type III F-test in the absence of interaction.

iii. The effect that interaction has on the power of the type II F-test.

In order to reduce the problem to a manageable size, we discuss primarily the 2x2 case. Then we have $\alpha_2 = -\alpha_1$, $\beta_2 = -\beta_1$, and $\gamma_{22} = -\gamma_{21} = -\gamma_{12} = \gamma_{11}$.

2. Power functions. In the notation above, the two F-tests we consider for an A main effect are

$$F_A(II) = MS_A(II)/MS_E$$

and

$$F_A(III) = MS_A(III)/MS_E$$

where $SS_E = \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2$. Note that, like GLM, we do not pool SS_{A*B} with SS_E in $F_A(II)$.

The sums of squares $SS_A(II)$ and $SS_A(III)$ can be computed according to the "reduction" method in (5) and (6). A more direct method for $SS_A(III)$ is

$$SS_A(III) = \sum_i H_i (\tilde{y}_{i..} - \tilde{y}_{...})^2 ,$$

where

$$H_i = b^2 \left(\sum_{j=1}^{b} (1/n_{ij}) \right)^{-1} ,$$

$$\tilde{y}_{i..} = \left(\sum_{i=1}^{a} y_{ij.}\right)/a$$

and $$\tilde{y}_{...} = \left(\sum_{i=1}^{a} H_i \tilde{y}_{i..}\right)/\sum_{i=1}^{a} H_i .$$

In the 2*2 case that we consider here, the SS's can be obtained from the contrasts

$$C_A(II) = h_1(\bar{y}_{11.} - \bar{y}_{21.}) + h_2(\bar{y}_{12.} - \bar{y}_{22.})$$

and

$$C_A(III) = \bar{y}_{11.} - \bar{y}_{21.} + \bar{y}_{12.} - \bar{y}_{22.}$$

where $h_j = 2(n_{1j}^{-1} + n_{2j}^{-1})^{-1}$ is the harmonic mean of n_{1j} and n_{2j}. Specifically,

$$SS_A(II) = \frac{[h_1(\bar{y}_{11.} - \bar{y}_{21.}) + h_2(\bar{y}_{12} - \bar{y}_{22.})]^2}{2(h_1 + h_2)}$$

and

$$SS_A(III) = \frac{[\bar{y}_{11.} - \bar{y}_{21.} + \bar{y}_{12.} - \bar{y}_{22.}]^2}{2(h_1^{-1} + h_2^{-1})}$$

Note that proportional data structure $n_{11}n_{22} = n_{12}n_{21}$ is not sufficient for $SS_A(II) = SS_A(III)$ to hold; a common misconception.

The F statistics corresponding to $SS_A(II)$ and $SS_A(III)$ have non-central F distributions with noncentrality parameters

$$\lambda_A(II) = \frac{[h_1(\mu_{11} - \mu_{21}) + h_2(\mu_{12} - \mu_{22})]^2}{2(h_1+h_2)\sigma^2}$$

$$= \frac{[(h_1+h_2)(\alpha_1-\alpha_2)+h_1(\gamma_{11}-\gamma_{21})+h_2(\gamma_{12}-\gamma_{22})]^2}{2(h_1+h_2)\sigma^2}$$

$$= 2[(h_1+h_2)\alpha_1 + (h_1-h_2)\gamma_{11}]^2/(h_1+h_2)\sigma^2$$

and

$$\lambda_A(III) = \frac{[\mu_{11} - \mu_{21} + \mu_{12} - \mu_{22}]^2}{2(h_1^{-1} + h_2^{-1})\sigma^2}$$

$$= 8\alpha_1^2/(h_1^{-1} + h_2^{-1})\sigma^2$$

Note that $\lambda_A(II)$ is a function of γ_{11} as well as α_1 and thus $\lambda_A(II)$ is not necessarily equal to zero when H_0^A is true. Note also that $\lambda_A(III)$ does not depend on γ_{11}, so that $\lambda_A(III) = 0$ if H_0^A is true.

The power of the F-tests is directly related to the noncentrality parameters. For even denominator degrees of freedom, the power $1-\beta$ can be computed using equations (9), (10), and (11) of Johnson and Kotz (1970).

Note that, in addition to the model parameters, the noncentrality parameters (and hence the powers) depend on the values of h_1 and h_2. We shall discuss specific aspects of the dependency in the context of a fixed total sample size $n_{11}+n_{12}+n_{21}+n_{22}=40$. Now $0 < h_1+h_2 < 20$, so the possible points (h_1, h_2) lie in the region as depicted in Figure 1. Selected points in Figure 1 are identified with the actual cell frequency distribution.

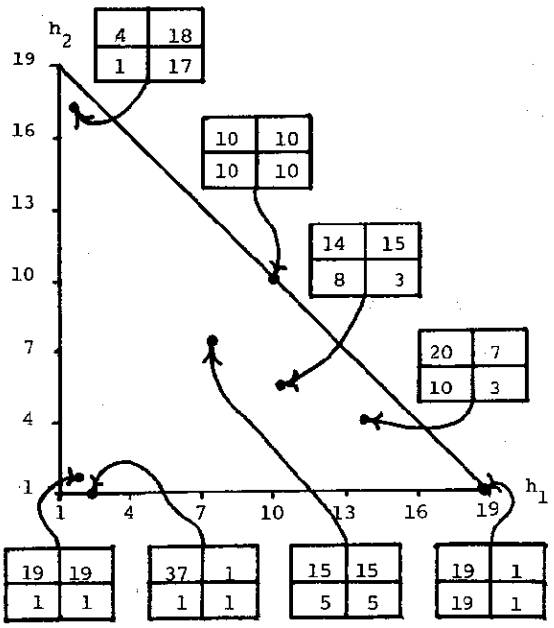

Figure 1. Region of possible values of (h_1,h_2), with points identified corresponding to certain cell frequency patterns.

Table 1. Noncentrality parameters (λ) and rejection probabilities (Prob.) for various tests and selected sample sizes. All tests are at the .05 level of significance, except the interaction test which is at the .25 level.

						$\gamma_{11}=.5\sigma$		$\alpha_1=5\sigma$		$\alpha_1=5\sigma$		$\alpha_1=1\sigma$, $\gamma_{11}=.5\sigma$		$\gamma_{11}=.5\sigma$	
						Type II Bias		Type II Power		Type III Power		Interaction Effect on Type II Power		Power of Interaction	
n_{11}	n_{21}	h_1	n_{12}	n_{22}	h_2	λ	Prob.	λ	Prob.	λ	Prob.	λ	Prob.	λ	Prob.
4	1	1.60	18	17	17.49	3.31	.42	4.77	.57	1.47	.22	6.51	.70	1.47	.53
10	10	10.00	10	10	10.00	0.00	.05	5.00	.59	5.00	.59	20.00	.99	5.00	.86
14	8	10.18	15	3	5.00	0.44	.10	3.80	.47	3.35	.43	21.81	.99	3.35	.75
20	10	13.33	7	3	4.20	1.19	.19	4.38	.53	3.19	.41	27.86	.99	3.19	.73
19	19	19.00	1	1	1.00	4.05	.50	5.00	.59	0.95	.16	42.05	.99	0.95	.44
15	5	7.50	15	5	7.50	0.00	.05	3.75	.47	3.75	.47	15.00	.96	3.75	.78
9	1	1.90	1	9	1.90	0.00	.05	0.95	.16	0.95	.16	3.80	.47	0.95	.44
37	1	1.95	1	1	1.00	0.08	.06	0.74	.13	0.66	.12	3.97	.49	0.66	.39

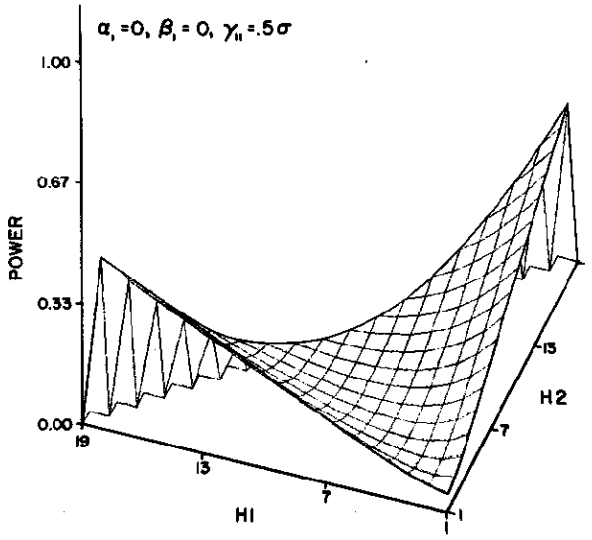

Figure 2. Rejection probability (bias) of F_A(II).

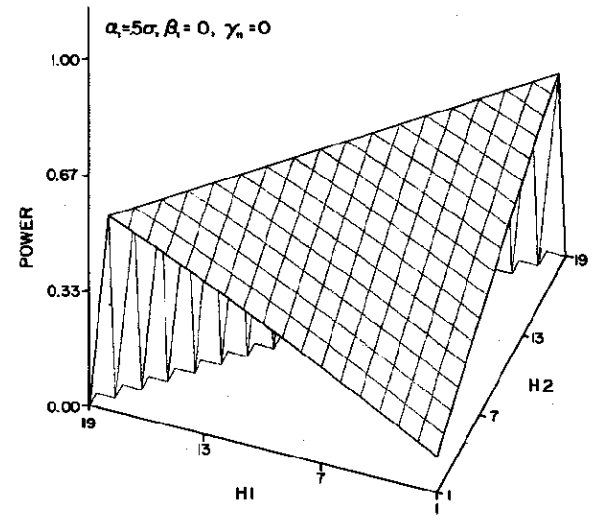

Figure 3. Rejection probability of F_A(II).

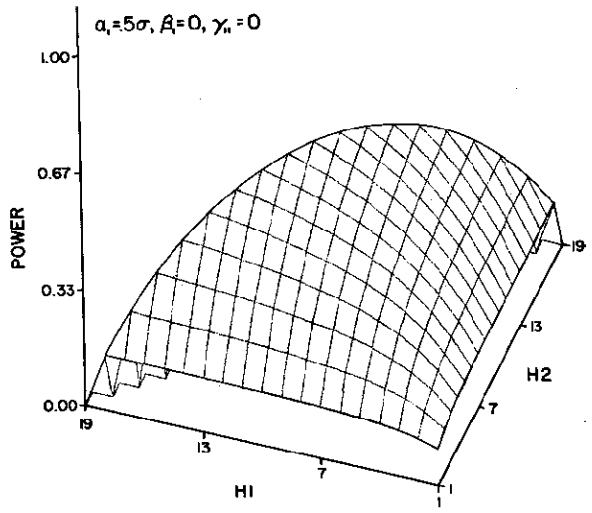

Figure 4. Rejection probability of F_A(III).

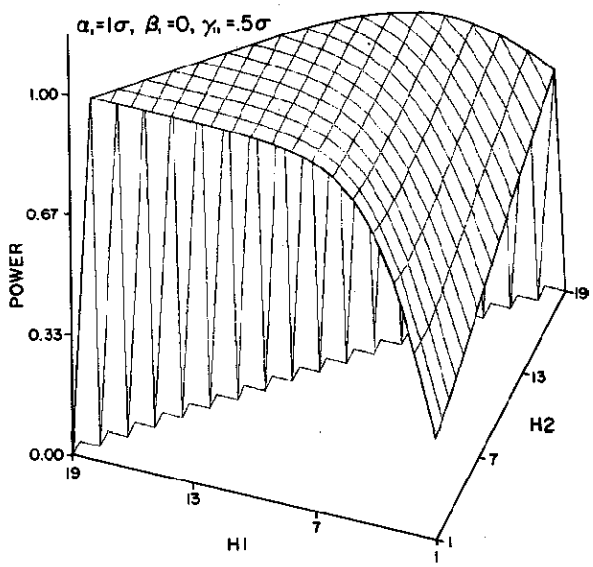

Figure 5. Rejection probability of F_A(II).

2.1 Bias in Type II test due to interaction. It was pointed out earlier that the Type II test does not test H_0^A: $\bar{\mu}_{1.} = \ldots = \bar{\mu}_{a.}$. Therefore, the rejection probability (type I error rate) will be greater than the prescribed α, depending on the magnitude of $(h_1-h_2)^2\gamma_{11}^2/(h_1+h_2)$. The true rejection rate for a nominal size $\alpha = .05$ test with $\gamma_{11} = .5\sigma$ is shown in Figure 2 as a function of (h_1,h_2). Calculations for selected cell frequency patterns are shown in Table 1. Note that the true rejection rate can be large as .2 for cell frequency distributions that are not uncommonly disparate.

2.2 Power of tests. The bias in F_A(II) noted above is due to the interaction parameter γ_{11}. If $\gamma_{11} = 0$, then both F_A(II) and F_A(III) are unbiased. Comparison of λ_A(II) and λ_A(III) when $\gamma_{11} = 0$ shows λ_A(II) > λ_A(III) with equality only if $h_1 = h_2$. Powers of F_A(II) and F_A(III) for $\alpha_1 = .5\sigma$ are shown in Figures 3 and 4, respectively, as functions of h_1 and h_2 for .05 level tests. The inefficiency of F_A(III) noted by Steel and Torrie (1960) and Bancroft (1968) is most pronounced when h_1 and h_2 differ greatly. Note that the power of F_A(II) is constant for constant $h_1 + h_2$. See Table 1.

The presence of γ_{11} in the noncentrality parameter for F_A(II) affects the power as well as the type one error rate. For given α_1, the rejection probability can (theoretically) range anywhere between the nominal α to 1 depending on the value of γ_{11}. The minimum rejection probability (= nominal α) occurs when $(h_2 - h_1)\gamma_{11} = (h_1 + h_2)\alpha_1$. Large rejection probabilities occur for large positive values of $(h_1 - h_2)\gamma_{11}$.

Figure 5 shows the dependence of the rejection probability on (h_1, h_2) for $\alpha_1 = 1.0\sigma$ and $\gamma_{11} = .5\sigma$, for .05 level tests.

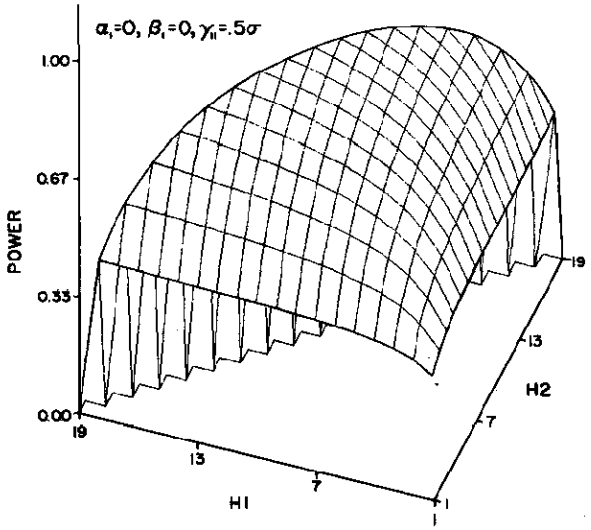

Figure 6. Rejection probability of interaction test.

As noted, Bancroft (1968) suggests testing H_0^{AB}: $\gamma_{11} = 0$ (Comparing models (1) and (2)) at α=.25, and recommends: 1.) testing H_0^A with F_A(III) if the test of H_0^{AB} is nonsignificant, or, 2.) testing H_0^A with F_A(III) if the test of H_0^{AB} is significant, provided the main effect hypothesis H_0^A is still of interest. This recommendation attempts to utilize preliminary information about interaction to aid in the selection of the main effect test statistic. Figure 6 shows the power of the α=.25 level interaction test with $\gamma_{11} = .5\sigma$ as a function of h_1 and h_2. Selected power computations are given in Table 1. The power of the interaction test has essentially the same form as the power of F_A(III), and is thus most adversely affected when h_1 and h_2

differ greatly. Note that for the case n_{11}= 20, n_{21}= 10, n_{12}= 7, and n_{22}= 3, the power is only $1-\beta(\lambda)$ = .73. Thus, for this degree of unbalance, even using Bancroft's suggested procedure, the probability is $\beta(\lambda)$ = .27 of incorrectly assuming no interaction and thus being led to the use of F_A(II). And this in turn could lead to excessive rejection probability if H_0^A: α_1= 0 is true.

2.3 The 2xb case. The results for the 2x2 case extend directly to the 2xb case if one recognizes that SS_A(II) and SS_A(III) derive from the contrasts

$$C_A(II) = \sum_j h_j(\bar{y}_{1j.} - \bar{y}_{2j.})$$

and

$$C_A(III) = \sum_j (\bar{y}_{1j.} - \bar{y}_{2j.}) ,$$

where $h_j = 2(n_{1j}^{-1} + n_{2j}^{-1})^{-1}$ is the harmonic mean of n_{1j} and n_{2j}, j=1,...,b. The corresponding noncentrality parameters are

$$\lambda_A(II) = 2\left[\alpha_1 \sum_j h_j + \sum_j h_j \gamma_{1j}\right]^2 / \sum_j h_j$$

and

$$\lambda_A(III) = 2b^2 \alpha_1^2 / \sum_j h_j^{-1} .$$

Inspection of λ_A(II) and λ_A(III) reveals that the conclusion drawn for the b=2 case apply, as general phenomena, to the b > 2 case as well. That is, highly disparate h_j, resulting from small cell frequencies for some levels of b, create large bias in F_A(II) and inefficiency in F_A(III).

3. Conclusion. Unbalanced cell frequencies in a 2x2 classification affect the power of various tests in the analysis of variance. Structures for which the harmonic means h_1 and h_2 differ greatly are particularly disruptive. This condition typically occurs when there is a small cell frequency in one of the columns (levels of B) but not both, and results in bias in F_A(II), and inefficiency in F_A(III) and in the test for interaction. We are not able to provide any easy solution to the problem. However, we hope to have at least made the point that a dogmatic rule such as "always use type II" or "always use type III" is not without pitfalls. Based on the limited calculations we have made here, it appears that bias in F_A(II) may be a more severe problem than inefficiency in F_A(III). Caution is urged in generalizing, however.

References

Bancroft, T.A. (1968). Topics in Intermediate Statistical Methods, Iowa State University Press, Ames.

Cramer, E.M. and Appelbaum, M.I. (1980). "Nonorthogonal Analysis of Variance-Once Again," Psychological Bulletin, 87, 51-57.

Gosslee, D.G. and Lucas, H.L. (1965). "Analysis of Variance of Disproportionate Data When Interaction is Present," Biometrics, 21, 115-133.

Hocking, R.R. (1981). "An Analysis of Methods Used in ANOVA with Missing Cells," Proc. 6th Ann. SUGI Conf., 94-99.

Johnson, N.L. and Kotz, Samuel, (1970). Continuous Univariate Distributions-2, Houghton Miffin, New York.

Overall, J.E., Lee, D.M. and Hornick, C.W. (1981). "Comparison of Two Strategies for Analysis of Variance in Nonorthogonal Designs," Psychological Bulletin, 90, 367-375.

Searle, S.R. (1971). Linear Models. Wiley, New York.

Searle, S.R. (1979). "Relationships between the Estimable Functions of SAS GLM Output for Unbalanced Data and the Hypotheses Tested by Traditional-style F-statistics," Proc. 4th Ann. SUGI Conf., 196-207.

Searle, S.R. (1981). "Quicks in Linear Model Computations for Unbalanced Data," Proc. 6th Ann. SUGI Conf., 38-45.

Searle, S.R., Speed, F.M. and Henderson, H.V. (1981). "Some Computational and Model Equivalences in Analysis of Variance of Unequal-Subclass-Numbers Data," The American Statistician, 35, 16-33.

Speed, P.M. and Hocking, R.R. (1980). "A Characterization of the GLM Sums of Squares," Proc. 5th Ann. SUGI Conf., 215-218.

Speed, P.M., Hocking, R.R. and Hackney, O.P. (1978). "Methods of Analysis of Unbalanced Data," J. American Statistical Association, 73, 105-112.

Steel, R.G.D. and Torrie, J.H. (1960). Principles and Procedures of Statistics, McGraw-Hill, New York.

Steel, R.G.D. and Torrie, J.H. (1980). Principles and Procedures of Statistics, 2nd ed., McGraw-Hill, New York.