ANALYSES OF VARIANCE OF UNBALANCED DATA FROM 3-WAY AND HIGHER-ORDER CLASSIFICATIONS

BU-606-M March, 1977

Shayle R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

Abstract

Answers are given to three questions asked of a panel* at a workshop session "Computing Approaches to the Analysis of Variance for Unbalanced Data" held at the Tenth Annual Symposium* on the Interface of Computer Science and Statistics.

Question 1 Is there a statistically valid default decision short of fitting all possible orders of main effects followed by all meaningful orders of interactions? On what features of the design structure does it depend? Is there a default strategy for simultaneously determining several interesting sets of hypotheses and computing their sums of squares?

Answer - It seems hards to believe that there could ever be a statistically valid default decision – particularly just one such decision, unique for all purposes. And the phrase begs the question as to what is meant by "statistically valid". Any hypothesis $H: \underset{\sim}{K}'\underset{\sim}{\beta} = \underset{\sim}{m}$, for which $r(\underset{\sim}{K}')_{s \times p} = s$ with $\underset{\sim}{K}'\underset{\sim}{\beta}$ estimable and $s \leq r(\underset{\sim}{X})$ for $E(\underset{\sim}{y}) = \underset{\sim}{X}\underset{\sim}{\beta}$, can be validly tested under normality using $F(H) = Q/s\hat{\sigma}^2$ for $Q = (K'b^\circ - m)'(K'GK)^{-1}(K'b^\circ - m)$ where $b^\circ = GX'y$ and $X'XGX'X = X'X$; then, under H the distribution of F(H) is F on s and N - r(X) degrees of freedom. Any default

---

* Panelists J. W. Frane, J. H. Goodnight, R. R. Hocking, S. R. Searle and G. Wilkinson. The meeting was held in Washington, D.C., April 14-15, 1972.

decision that leads to an H of this sort can be validly tested. But since there are many such H's, with boundless ideas for being interested in some rather than others, it is difficult to see how any computer program can contain unique specifications for a choice that will be suitable for all possible kinds of data.

The suggestion that a computer program could choose "interesting sets of hypotheses" strikes an odd chord. "Interesting" to whom? To the person whose data are being analyzed, presumably. But isn't the choice of interesting hypotheses part of the scientific method, indeed that very part which so often involves human conjecture? This, then, is the bailiwick of the experimenter, the data gatherer, the survey analyst, of the person who wants to make a step forward in his understanding of nature. It is not even the statistician's job, let alone that of an inhuman, non-thinking, automaton computer. Certainly a statistician can help, not as an automaton but as a clear thinking scientist discussing nature with the researcher, helping him formulate, i.e. put into formal terms, the hypotheses or conjectures about nature that he has in mind. One large aspect of the statistician's help is to confine the scientists' hypotheses to ones that are testable – i.e., to those involving estimable functions.

Question 2 If interactions are found significant is an automatic procedure for splitting the design into more homogeneous subdesigns feasible?

Answer Any answer to this question must be preceded by considering a more fundamental question such as "what is the meaning of interactions in high-order classifications and how can they be tested, especially when unbalancedness of data includes many empty cells?" For example, can one give a useful, practical meaning to a 4-way interaction; and if 30% of the sub-most cells of the data set have no data, what is the meaning of interactions being "found significant"?

The complexities of trying to understand interactions in 3-, 4-, 5-way and

higher order classifications do, I believe, overpower any consequences of what should be done "if interactions are found significant" – especially for unbalanced data in which there are many empty cells. Even suggesting that a computer program could be planned to split the "design into more homogeneous sub-designs" therefore seems somewhat absurd. To heighten the absurdity, what would it do if 5'th order and 3'rd-order interactions were significant but 4'th order ones were not?

It seems clear to me that contemplating interactions in high-order classifications having unbalanced data including empty cells highlights the absolute necessity to abandon overparameterized models and to fall back on cell means. This is, of course, what Hocking and co-workers (1,2) have been advocating for years and indeed is precisely what Fisher did when he started this whole analysis of variance business anyway (see 4). The model is then $E(\underset{\sim}{y}) = \underset{\sim}{\mu}$ with each element of $\underset{\sim}{\mu}$ being a population cell mean, $\mu_{ijk\ell m}$, say, for a 5-way classification. Then $\hat{\mu}_{ijk\ell m} = \bar{y}_{ijk\ell m.}$ is the b.l.u.e. of $\mu_{ijk\ell m}$ with variance $\sigma^2/n_{ijk\ell m}$. A hypothesis about any number of linear combinations of the μ's is then testable, $\underset{\sim}{K}'\underset{\sim}{\mu} = \underset{\sim}{m}$ say, with its F-statistic being $Q/s\hat{\sigma}^2$ for $Q=(\underset{\sim}{\bar{y}}'\underset{\sim}{K} - \underset{\sim}{m}')(\underset{\sim}{K}'\underset{\sim}{G}\underset{\sim}{K})^{-1}(\underset{\sim}{K}'\underset{\sim}{\bar{y}} - \underset{\sim}{m})$ where $\underset{\sim}{\bar{y}}$ is the vector of cell means and $\underset{\sim}{G}$ is the diagonal matrix of reciprocals of cell numbers, $1/n_{ijk\ell m}$. Under these circumstances the model is simple to learn, to understand and to use: and the task of what hypotheses are to be tested is laid fairly and squarely where it should be: at the foot of the researcher. However, his task is now easy, compared to his task in overparameterized models. Any hypothesis about the value of any linear combinations of the population cell means (the $\mu_{ijk\ell m}$'s) can be tested. He has only to state his conjectures in this form, without any limitation at all on what sort of linear combination (because each and every one of them is an estimable function) can be the basis of a hypothesis. No statistician need persuade him to be confined to just certain (estimable) kinds of linear combinations; they are all permissible.

Question 3 What is the appropriate criterion for the testing of hypotheses? The hypotheses tested by the F statistics are orthogonal even with unbalanced data if the sums of squares for each line of the ANOVA table is computed by adjusting for all lines above it and ignoring all lines below it. The set of contrasts associated with almost any other set of sums of squares is not orthogonal and therefore open to ambiguity of interpretation.

Answer Ambiguity of interpretation is built into the analysis of unbalanced data. Furthermore in many kinds of data, empty cells are a virtual certainty. In family surveys, for example, the 72-year-old father with 4 children under 5, on welfare, living in Georgetown in a 1-room house with 5 cars, 2 yachts and a Lear Jet simply does not exist. The Howard Hughes of this world seldom get caught in survey data.

So what do we do, insofar as hypotheses are concerned? Fall back on cell means is undoubtedly the only rational thing to do; and, thankfully, it is an easy route to take.

The concern implicit in this third question is that of orthogonal hypotheses and/or orthogonal sums of squares. Traditionally, hypotheses $H_1: \underset{\sim}{k}_1'\underset{\sim}{\beta} = 0$ and $H_2: \underset{\sim}{k}_2'\underset{\sim}{\beta} = 0$ (for $\underset{\sim}{k}_1'$ and $\underset{\sim}{k}_2'$ being row vectors) would be considered orthogonal when $\underset{\sim}{k}_1'\underset{\sim}{k}_2 = 0$. However, the numerator sums of squares for the corresponding F-statistics are independent (under normality) if and only if $\underset{\sim}{k}_1'\underset{\sim\sim}{G}\underset{\sim}{k}_2 = 0$; in which case they sum to that used for testing H_1 and H_2 simultaneously (3, Sec. 5.5g). Therefore $\underset{\sim}{k}_1'\underset{\sim\sim}{G}\underset{\sim}{k}_2 = 0$ seems an appropriate generalization of the orthogonality concept for unbalanced data – and it reduces, of course, to $\underset{\sim}{k}_1'\underset{\sim}{k}_2 = 0$ when $\underset{\sim\sim}{G}$ is a scalar matrix (or when appropriate principal submatrices of $\underset{\sim\sim}{G}$ are). In this sense, orthogonal hypotheses have independent numerator sums of squares – but not independent F-statistics (both denominators contain $\hat{\sigma}^2$). As a result, the concept of hypotheses being orthogonal seems to deserve less importance than is implied by this question. The criteria for testing a

hypothesis should be (i) that it is testable and (ii) that it is meaningful and of interest to the experimenter.

Final Comment An overriding comment is to tell computer jocks not to write fully general programs. They are too difficult to explain and are so fraught with dangers for possible erroneous use that they frequently do get used erroneously -- and often without the user knowing of the errors perpetuated. Complementary advice for statisticians would be to encourage data gatherers to set up their own hypotheses, and to assist them by relying entirely upon the cell means model. It is straightforward, requires no computers, is easy to understand and is in direct line with the way in which most experimenters think about their data.

References

1. Hocking, R. R. and Speed, F. M. (1975). A full rank analysis of some linear model problems. J. Amer. Stat. Assoc. 70, 706-712.

2. Speed, F. M. and Hocking, R. R. (1976). The use of the R( )-notation with unbalanced data. The American Statistician 30, 30-34.

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

4. Urquhart, N. S., Weeks, D. L. and Henderson, C. R. (1973). Estimation associated with linear models: a revisitation. Communications in Statistics 1, 303-330.