ANALYSIS OF VARIANCE WITH UNBALANCED DATA (INCLUDING EMPTY CELLS): COMMENTS ON A PAPER

BU-593-M October, 1976

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, N. Y. 14853

Abstract

Misleading features in Golhar and Skillings (1976) are corrected.

Paper No. BU-593-M in the Biometrics Unit Mimeo Series, Cornell University.

ANALYSIS OF VARIANCE WITH UNBALANCED DATA (INCLUDING EMPTY CELLS): COMMENTS ON A PAPER

Carrying out analyses of variance of unbalanced (unequal-subclass numbers) data, including the possibility of empty cells, can be difficult on two counts. One is knowing what computations are needed, and the other is finding a computer program that carries them out. The Golhar and Skillings (1976) paper is welcome on the second count, in providing information about the different outputs produced by several programs, but unfortunately it contains several misleading features concerning description of just what those outputs are. The following comments seem pertinent.

1. The paper indicates that it explains how to treat the "analysis of variance of unbalanced data". This is not strictly true. It deals only with unbalanced data having all cells filled and never mentions the case of empty cells. This distinction is most important. For example, the analysis of Table 1 of the paper is nonsense for data that have empty cells. And the authors give no warning of this.

2. The heading "adjusted for" in Table 1 is misleading, because there is no exact description of its meaning. Description is essential for entries such as the 2.24 which, in terms of the familiar $R(\cdot|\cdot)$ notation, would appear to be $R(A|\mu,B,AB)$. Indeed, Section 4 of the paper describes it as such, but is vague because it gives no explanation of how the computation has been made. One use of the phrase "adjusted for" as applied to a sum of squares labeled $R(A|\mu,B,AB)$ gives the value as identically zero [e.g., Searle (1971)]. Other uses [Speed and Hocking (1976)] do not.

Golhar and Skillings (1976) fail to recognize these different uses of "adjusted for" and thus perpetuate the confusion that has surrounded them for too long.

3. A prime object of calculating sums of squares in analyses of variance is for testing hypotheses, and it is of no help to say, as is done on page 50, that sums of squares do "not test hypotheses (Y) in a strict sense". They do not do so in any sense. The exact form of each test is known and should be given; e.g., $R(A|\mu,B) = 50.63$ of example 1 is used for testing [see Searle (1971, p. 308)] the hypothesis H: $\varphi_i = 0$ for all i, in this case i = 1, 2, where $\varphi_i = \Sigma_{i'}\Sigma_j\lambda_{ii',j}(\alpha_{i'} + \tau_{i'j})$ for $\lambda_{ii',j} = \delta_{ii'}n_{ij} - n_{ij}n_{i'j}/n_{.j}$ with $\delta_{ii'}$ being the Kronecker delta. And for each of at least three different quantities that might be represented by the symbol $R(A|\mu,B,AB)$ there is correspondingly three different hypotheses that can be tested — as has recently been described with great clarity by Speed and Hocking (1976).

4. The absence of an entry under "adjusted for" alongside the AB entry of Table V (e.g., 250.92 in example 1) may puzzle some readers. The description $R(\mu,A,B,AB) - R(\mu,A) - R(\mu,B) + R(\mu)$ provides clarification.

5. Explanation is also lacking as to why the sums of squares add up to $\Sigma\Sigma\Sigma x^2_{ijk} - n_{..}\bar{x}^2_{...}$ in Tables III, IV and V whereas in Tables I and II they do not. Mathematical statisticians know why but many users of computer packages do not.

6. No indication is given of how the computer programs handle more classifications than those of the 2-way crossed classification dealt with in the two examples. This is, of course, a large task — but at least the reader needs warning that extrapolation to higher-order classifications is not necessarily obvious.

7. Finally, a small point: SPSS MANOVA (option = 9) is mentioned at the end of the paper, but nowhere else.

Computer users are often confused by the matters raised in the preceding comments because many of them wrongly think that analyses of unbalanced data involve only small variations on those

of balanced data. It is therefore important that statistical writing directed toward computer users be completely accurate in these matters.

BIBLIOGRAPHY

Golhar, M. B. and Skillings, J. H. (1976). A comparison of several analysis of variance programs with unequal cell size. Commun. Statist. – Simulation and Computation B5, 43-54.

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

Searle, S. R. (1973). How little computing need we teach to statistics majors? Proceedings Computer Science & Statistics: 7th Annual Symposium on the Interface, 204-209, Statistical Laboratory, Iowa State University.

Speed, F. M. and Hocking, R. R. (1976). The use of the R( ) notation with unbalanced data. The American Statistician 30(1), 30-33.

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