5-1979

A μ-Model Approach on the Cell Means: The Analysis of Full, Design Models with Non-Orthogonal Data

Richard Van Koningsveld
Utah State University

A μ-MODEL APPROACH ON THE CELL MEANS: THE ANALYSIS OF FULL, DESIGN MODELS WITH NON-ORTHOGONAL DATA

by

Richard Van Koningsveld

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Applied Statistics

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

1979

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

A μ-Model Approach on the Cell Means:
the Analysis of Full, Design Models
With Non-Orthogonal Data

by

Richard Van Koningsveld, Master of Science
Utah State University, 1979

Major Professor: Dr. Ron Canfield
Department: Applied Statistics and Computer Science

This work considers the application of a μ-model approach on the cell means to a special yet important class of experimental designs. These include full factorial, completely nested, and mixed models with one or more observations per cell. By limiting attention to full models, an approach to the general data situation is developed which is both conceptually simple and computationally advantageous.

Conceptually, the method is simple because the design related effects are defined as if the cell means are single observations. This leads to a rather simple algorithm for generating main effect contrasts, from which associated interaction contrasts can also be formed. While the sums of squares found from these contrasts are not additive with non-orthogonal data, they do lead to the class of design related hypotheses with the clearest interpretation in terms

of the cells.

The computational method is advantageous because the sum of squares for each source of variation is evaluated separately. This avoids the storage and inversion of a potentially large matrix associated with alternative methods, and allows the user to evaluate only those sources of interest.

The methodology outlined in this work is programmed into a user-easy, interactive terminal version for the analysis of these n-factor design models.

(52 pages)

INTRODUCTION

The analysis of variance technique is widely applied to data arising from experiments. Its utility as a statistical technique comes not only as a method to estimate 'treatment' effects, but because it provides a means of testing the likelihood such effects are distinguishable.

The computational methods for the analysis of variance are simplest for orthogonal data. This occurs when the cell frequencies are either equal, or can be considered proportional to their respective populations (A cell is defined by a specific value (level) from each factor in an experiment). With orthogonal data the sums of squares (SS) for component sources of variation are mutually independent and together add up to the total SS. In practice, various submodels may also be investigated simply by adding the SS of appropriate sources of variation (SV) together. This is because, for orthogonal data, the various component SS and effects remain invariant to the model.

Frequently, however, the cell frequencies are both unequal and disproportionate, and the above situation does not hold. The effects and SS obtained for a particular SV depends on the model assumed in the analysis, and the additivity property for SS is lost. Concurrently, the computational methods for non-orthogonal data are inherently more difficult. This arises directly from the unequal and disaproportionate nature of the cell frequencies themselves, causing the various effects and SS to become entangled.

In certain instances, observation balance can be restored by

randomly deleting observations or by generating them through missing data formulae. When this is undesirable or impossible one must be satisfied with an approximate analysis or resort to the more complicated computational procedures.

Approximate methods closely parallel those for orthogonal data. Because of this, they are easy to apply and remain conceptually simple. While these alternatives give good results for 'nearly' orthogonal data, it is difficult to establish the effect of observation imbalance on the degree of approximation in the analysis.

Subsequently, regression on dummy (0,1) or (1,0,-1) variables has become the standard exact method to handle the general data situation. This least squares approach is not easily seen as a natural extension of the techniques used for orthogonal data. In part, this is because the focus is shifted away from the cell means (or totals) to the individual parameters of a regression model.

In the past decade, a method for the general data situation has been formalized in terms of the cells of the experiment. This is known as the μ-model approach. While stastically equivalent to dummy variable regression, conceptually it is somewhat easier to understand because the cells represent the 'objects' on which the data were initially collected.

While the μ-model approach is conceptually easier, in general there is no real computational advantage over the regression approach. The only exception for the general data situation arises when the cell parameters of the μ-model are estimated by the observed cell means. Such models are called full (or saturated) models.

This work considers the application of a μ-model approach to a special and important class of experimental designs, which include full factorial, completely nested, and mixed models with one or more observations per cell. By limiting attention to full models, an approach to the general data situation is developed which is both conceptually simple and computationally advantageous. To accomplish this, an algorithm for computing the effects and SS associated with any SV for these n-factor ANOVA models is derived.

LITERATURE REVIEW

The analysis of variance for multi-factor experiments is easiest to compute and interprete when the data are orthogonal. Despite this fact, and for various reasons, the need to analyze non-orthogonal data frequently arises. Several methods, both approximate and exact, are available in this situation; each approach having its own advantages and disadvantages.

Early Methods

The first approximate solutions to the problem of non-orthogonal data in multi-factor experiments were the method of unweighted means (Yates, 1933) and the method of expected subclass numbers (Snedecor, 1934). Respectively, they are recommended when the cell frequencies 'nearly' satisfy the equal and proportional numbers case. With only minor adjustments, each defines treatment effects in terms of the cell means, treated as single observations; and proceed as if the data were balanced. Bancroft (1968, p.35), in describing these methods, points out that "both make use of the addition theorem for the analysis of variance which holds for orthogonal data and thus make the SS calculations less complicated." While these methods are exact for orthogonal data, their utility as a quick alternative is diminished by the degree of observation imbalance in the cells. However, when justified, these methods are especially advantageous in that a wide range of experimental designs can be handled.

Yates (1934) published another procedure known as the method of weighted squares of means. Its utility is generally restricted to finding exact main effect SS for full factorial (fixed effects) models. Like his other method, the cell means are treated as single observations; but as noted by Searle (1971), each component SS is inversely weighted by the variance of that component. The weights are a function of the cell frequencies so as to account for the different sampling distributions of the cell means involved. This fact, ignored by the approximate methods mentioned, provides for its exactness. Each main effect SS derived by this method is adjusted for the remaining main effects and interactions in the full model. Their corresponding mean squares, divided by the (pooled) within cell variance, provide F statistics upon which differences in the effects are detected.

Least Squares

The methods mentioned were developed to avoid the available but time consuming method of least squares. Computers made a least squares solution routinely practical and the technique rapidly evolved. Unlike the methods mentioned so far, the least squares approach is not framed in terms of the cells of the experiment. In this way, it departs from the traditional ways of conceptualizing the analysis of variance. However, as an exact method it provides a unified approach to handle the whole gamit of linear models. For comparative purposes, it is worthwhile to quickly review this approach.

In a contemporary form, an analysis of variance (ANOVA) model written,

$$Y_{i..jm} = u + A_i + \ldots + e_{i..jm} \qquad [1]$$

is parameterized through dummy variable (0,1) concepts to the regression form:

$$Y = XB + E. \qquad [2]$$

Here, B is the vector of unknown paramters and X is an incidence matrix which governs the selection of parameters pertinent to each observation in the vector Y. The vector E is a vector of observational errors or residuals. Under the usual ANOVA assumptions each element of E is NID $(0, \sigma_e^2)$.

To obtain a unique solution for B a variety of constraints may be imposed. The constraints normally associated with ANOVA models are the 'usual' constraints, and one of the following of these is imposed on each SV depicted in the model given by [1].

a) Crossed Main Effect

$$\Sigma A_i = 0$$

where the summation extends over the I levels of the single factor A.

b) Nested Main Effect

$$\Sigma B_{(.j\ldots k.)1} = 0$$

where the summation extends over the L levels of B existing within each particular set of nesting subscripts (.j..k.). These terms can be viewed as a main effect within each unique set of nesting subscripts.

c) Crossed Interaction

$$\Sigma A...B_{i...j} = 0$$

where the summation extends over the existing levels defined for any particular crossed factor appearing in the interaction.

d) Nested Interaction

$$\Sigma A...B_{(.j..k.)i...l} = 0$$

where the summation extends over the levels existing for any particular crossed factor within each unique set of nesting subscripts, (.j..k.). This latter term may be viewed like a crossed interaction existing within each unique set of nesting subscripts.

By imposing constraints, such as the ones above, the X matrix of [2] is brought to full column rank and a unique solution for the unknown parameters in B may be found. This is given by:

$$B = (X'X)^{-1} X'Y, \qquad [3]$$

where now, the column rank of X is equal to the degrees of freedom of the model assumed in the analysis.

After a statistical solution is found by this least squares procedure, the SS associated with an individual subset of parameters may be computed. The SS for a SV is found when the subset is composed of the parameters the source represents. Specifically, the SS for testing the multivariate hypothesis:

$$H_o : b = 0, \qquad [4]$$

with K contiguous parameters from the solution vector B will have k degrees of freedom.

The SS for this SV is found by:

$$SS\ (H_o) = b\ S^{-1}\ b. \qquad [5]$$

where b is a vector, and S is a (k x k) sub-matrix of its related elements in the $(X'X)^{-1}$ matrix. Repeated use of the proceeding expression generates the SS for each SV in the ANOVA model. A more detailed discussion of linear models can be found in Searle (1971).

μ-Model

General Discussion

Speed (1969) formalized what amounts to a reparameterization of the regression form just discussed, known as the μ-model approach. Like least squares, this approach provides a unified method for the

analysis of variance. While in general the μ-model approach has no real computational advantage over least squares, the μ-model is framed in terms of the cells of the experiment, giving it greater conceptual appeal.

Instead of expressing the observations directly in terms of the parameters of a model written:

$$Y_{i...jm} = u + A_i + ... + e_{i..jm};$$

the μ-model begins by expressing the observations more simply by the model:

$$Y_{i..jm} = u_{i..j} + e_{i..jm}. \qquad [6]$$

That is, each observation is divided into two components where:

$u_{i..j}$ represents a parameter characteristic of the cell (i..j);

$e_{i..jm}$ is the random error associated with the m^{th} observation collected on the cell (i..j).

The generalization of expression [6] to all observations is given in matrix form by:

$$Y = WU + E \qquad [7]$$

where:

Y is a vector of the observations

W is a matrix of weights

U is a vector of the cell parameters, and

E is a vector of observational errors.

The equivalence of the regression and μ-model is demonstrated by introducing the identity matrix, I, into equation [7] and partitioning it into two parts. That is:

$$Y = WU + E$$

$$Y = WIU + E$$

$$Y = W(K^{-1}K)\ U + E$$

$$Y = (WK^{-1})\ (KU) + E \qquad [8]$$

$$Y = ZB + E \qquad [9]$$

Equation [9] is recognized as the regression form.

The key observation in examining equations [8] and [9] is that the effects (parameters) in B are represented by a set of linear functions of the vector U. This set is denoted by K. As a consequence of this, the SS due to a subset of k contiguous parameters in B is statistically equivalent to the SS corresponding to k linear functions of U. That is, if H represents these k functions, then in direct analogy to equation [5] the SS is given by:

$$SS(H) = (HU)'\ (H(W'W)^{-1}\ H')^{-1}\ (HU) \qquad [10]$$

In general, the matrix W is unknown, so the matrix solution

$$U = (W'W)^{-1}\ W'Y \qquad [11]$$

cannot be used to evaluate U directly. Instead, a completely different computational method is required. The general procedure for doing so is outlined by Speed (1969) and this has been programmed by Bryce (1974).

Applied to Cell Means

Searle (1971) mentions the special case of the μ-model $Y = WU + E$ where U is a vector of the cell means. In this case W is a (0,1) incidence matrix governing the selection of the cell mean in U appropriate to each observation in Y. And as before, E is a vector of errors associated with corresponding observations in the vector Y.

The nature of W is such that $(W'W)^{-1}$ is a diagonal matrix with the reciprocals of the cell frequencies along the diagonal. This result can be seen directly from first principles. Since the within cell (observational) errors are assumed to be NID $(0,\sigma_e^2)$, the distribution of the mean of, say the p^{th} cell is NID $(u_p, \sigma_e^2/n_p)$. That is, each mean has a variance proportional to $1/n_p$. The independence of the cell means implies the off-diagonal (covariance) elements are zero. Thus the SS associated with H (the linear combinations of these cell means) follows from [10] and is given by:

$$SS(H) = (HU)' \, (HD(1/n_{pp})H')^{-1} \, (HU) \qquad [12]$$

$D(1/n_{pp})$ represents the diagonal matrix of the reciprocals of the cell frequencies; the p^{th} element along its diagonal corresponding to the p^{th} mean is U.

In practice, U is replaced by a corresponding vector of observed cell means as the best linear unbiased estimates of the cell population

means. Upon making this substitution and writing:

$$SS(H) = (HY)' \, (HD(1/n_{pp}) \, H')^{-1} \, (H\bar{Y}) \qquad [13]$$

it is apparent that its evaluation only awaits framing the appropriate linear functions of interest in terms of the cell means. And, as Searle (1971) points out this is done within the context of what the data represent. In any event the estimated variance-covariance matrix for any set of linear functions of the cell means, estimated by $H\bar{Y}$, is given by:

$$V(H\bar{Y}) = (HD(1/n_{pp})H')s_e^2 \qquad [14]$$

where s_e^2 is the estimator of σ_e^2 pooled from the cells containing data.

This work incorporates expression [13] as the basis for an analysis of variance procedure applied to common experimental designs with no missing cell. In the analysis of multi-factor experiments with unbalanced data, the SS obtained from the use of [13] are those derived in considering the full (or saturated) model. Full models are those which include all possible interactions among the factors it contains, whereas those lacking one or more interactions are called restricted. This distinction is unnecessary with balanced data as the SS obtained for a SV, say a main effect, is the same whether the model is full or restricted. This is untrue for unbalanced data. The SS for a given SV depends on the model assumed in the analysis. Subsequently, the hypotheses actually tested by their associated mean squares must also

differ when expressed in terms of the cells of the experiment.

Hypothesis Testing

The SS derived in the analysis of variance are the fundamental quantities used in estimating components of variance and for testing hypotheses. Searle (1971) discusses methods for estimating components of variance and are not considered further in this paper. However, some discussion of hypotheses tested with non-orthogonal data seems worthwhile, particularly when considered in terms of the cells in the experiment.

Kutner (1974) and Searle (1971) discuss the hypotheses tested in the two-way model. For clarity, the reduction notation used by Greybill (1961) accompanies each type of hypotheses as an aid to the model assumed in each case. The two-way model is considered in its μ-model form:

$$Y_{ijk} = u_{ij} + e_{ijk}$$

with

$$i = 1,2,\ldots,I$$

$$j = 1,2,\ldots,J$$

$$k = 1,2,\ldots,n_{ij}$$

$$N = \sum_{ij}\sum n_{ij}$$

and where u_{ij} is the mean of the cell (i,j).

The levels associated with the factors A and B are respectively denoted by the subscripts i and j. By adopting the convention that a dot indicates that a subscript has been summed over, the unique types of hypotheses can then be enumerated in terms of the cells of the experiment. These are:

a) $R(A/u,B,AB)$

$H_o: u_{1.} = u_{2.} = \ldots = u_{I.}$ where $u_{i.} = (1/I)\sum_j u_{ij}$

b) $R(AB/u,A,B)$

$H_o: u_{ij} - u_{i'j} - u_{ij'} + u_{i'j'} = 0$

for all i,i',j,j' provided $i \neq i', j \neq j'$

c) $R(A/u)$

$H_o: \sum_j (n_{ij}/n_{i.})u_{ij}$ equal for all i

d) $R(A/u,B)$

$H_o: \sum_j (n_{ij} - n_{ij}^2/n_{.j})u_{ij} - \sum_{i \neq i'} \sum_j (n_{ij}n_{i'j}/n_{.j})u_{i'j'}$

equal for all i.

Several points about these hypotheses, framed in terms of the cell means, deserve discussion. Hypotheses a) and b) result from the use of the full model considered here and these appear as the same simple functions of the cell means found with orthogonal data. Hence, Francis (1973) contends these tests reflect conceptually what most people desire from their experiments. The hypothesis given in a) implies a test

for the equality of the marginal means evenly weighted over the remaining factor(s). In general, this is equivalent to the manner main effects are defined in the balanced case. The interaction hypothesis given in b) also indicates a test for no interaction effects in the same way as for balanced data.

The interpretation of the hypotheses c) and d) are not as simple, because they involve the cell frequencies. Kutner (1974) contends the hypothesis c) is "appropriate only when the n_{ij}/N are good estimates of the population proportions." And this is suggested for a proportional analysis only when the interaction is negligible. For unbalanced data, the hypothesis given in d) is impossible to interprete in terms of the cell means. As Searle (1971) notes, this hypothesis reflects less about the relative magnitudes of the cell means than the number of observations the cell contain. This is clearly undesirable as the cell means themselves are the 'objects' of interest in the experiments, not the cell frequencies. Of course, this is not to say that one cannot assume this 'no interaction' model with non-orthogonal data, but rather that this hypothesis has no clear meaning in terms of the cell means themselves.

Summary

Various computational methods for the analysis of variance are available for non-orthogonal data. The methods of unweighted means and expected subclass numbers provide a quick, approximate analysis for a wide variety of experimental designs. The method of weighted squares of means, by contrast, yields exact sums of squares but only

for main effects in factorial models. Computers made a least squares solution to the analysis of variance a practical exact method. However, it is not a generalization of the methods used for orthogonal data. The μ-model approach is conceptually advantageous in this respect, as it is framed in terms of the cells of the experiment. But, in general it has no computational advantage over least squares. An exception to this arises in the case of full design models considered in this thesis. Additional justification for developing this approach to these commonly encountered designs comes from the interpretation given full model hypotheses.

METHODOLOGY FOR THE CELL MEANS MODEL

The purpose of this thesis is to develop a computer program for the analysis of variance of full model, multi-factor experiments having one or more observations per cell. The basis of this program is a μ-model approach applied to cell means. This allows the sums of squares (SS) of individual sources of variation (SV) in the model to be evaluated. Prior to the discussion of the development of this procedure, it is necessary to define some basic notation.

Basic Model Notation

The general framework of this approach begins by considering an experiment with M factors described by the μ-model:

$$Y_{ij*lmn} = \bar{y}_{ij*lm} + e_{ij*lmn} \qquad [15]$$

where:

Y_{ij*lmn} is the n^{th} observation taken on the cell with factor level subscripts (ij*lm).

$\bar{y}_{ij*lm}$ is the (observed) mean of the cell with factor level subscripts (ij*lm).

e_{ij*lmn} is the error associated with the n^{th} observation taken on the cell with factor level subscripts

(ij*lm). These are assumed to be NID (o, σ_e^2).

The asterisk is used to denote the subscripts of other possible factors in the experiment.

Since no structure (design) has been given the experiment, a general notation to describe the levels for each of the M factor is necessary. This follows the notation describing a nested classification.

factor	subscript and levels
1	i = 1,2,...,I
2	j = 1,2,...,J(i)
.	.
.	.
.	.
M-1	l = 1,2,...,L(ij..)
M	m = 1,2,...,M(ij..l)

When the number of levels for a factor is the same, as it is for example with crossed factors, the subscripts defining the levels may be dropped.

For much of the following discussion it is easiest to reparameterize the model so as to have a way of addressing a cell by a single subscript. This can be done by letting p(p = 1,2,...,P) be uniquely associated with an existing combination of the M factor levels. The number of observations in the p^{th} cell is then denoted simply by n_p, and the total number of observations, N, by $N = \sum_p^P n_p$. Using this notation, the model given by [15] is rewritten:

$$y_{pn} = \bar{y}_p + e_{pn}, \qquad [16]$$

with the corresponding terms defined as before.

Preliminary One-Way Analysis

Before proceeding with the analysis of multi-factor experiments, it is often recommended to first determine if the (whole) model shows significance. For full models this amounts to a one-way analysis of variance conducted on the cells; and this usually requires sufficient observations to provide an estimate of the within cell error, s_e^2. The basic quantities for this preliminary analysis are calculated in the usual manner as given below:

a) $SS_{total} = \sum_p \sum_n^{n_p} y_{pn}^2 - CT$

where $CT = (\sum_{pn} \sum y_{pn})^2/N = y_{..}^2/N$

b) $SS_{among} = \sum_p n_p \bar{y}_p^2 - CT = \sum_p y_{p.}^2/np - CT$

c) $SS_{error} = SS_{total} - SS_{among}$

These results are summarized in an ANOVA table with their degrees of freedom (df) as follows:

TABLE 1. ANOVA table for one-way analysis.

Source of Variation	df	MS	F-ratio
Total	N - 1	$SS_{total}/(N-1)$	
Among cells	P - 1	$SS_{among}/(P-1)$	MS_{among}/MS_{within}
Within cell (error)	N - P	$SS_{within}/(N-P)$	

This F-ratio with (P-1) and (N-P) degrees of freedom used to test the (full) model, is at the same time a test for the equality of the cell means.

Evaluation of Sums of Squares

General Discussion

A μ-model of the cell means can always be equated to a full model reflecting the SV associated with its design. That is to say,

$$\bar{y}_p = u + A_i + \ldots \qquad [17]$$

provided the model on the right includes all possible interactions of the factors it contains.

These individual SV do not generally produce an analysis of variance in the sense of SS which add up to the total SS. But, they do provide an analysis with the same degree of freedom partitions. Hence, any full model, with P cells, has P degrees of freedom; one of which is lost estimating the overall location parameter, u. The

remaining (P - 1) degrees of freedom (among cells) are partitioned in accordance with the SV in the experimental design.

The SS for these SV can be calculated once the effects (parameters) associated with these sources are defined. It is a feature of full models that these effects are defined in terms of the same marginal means found with orthogonal data. The cell means in this sense are treated as single observations, and the various design related marginal means are computed as if the data were balanced. A 'usual' constraint imposed on a SV then defines the effects as straight-forward functions of these marginal means.

The utility of the SS expression:

$$SS(H) = (H\bar{Y})' \, (HD(1/n_{pp})H')^{-1} \, (H\bar{Y}) \qquad [18]$$

reproduced from equation [13] hinges on the ability to unravel these effects and re-express them directly in terms of the cells.

At this point, it is worthwhile to review the terms in equation [18] and discuss their dimensionality.

a) As just mentioned, H is a matrix specifying the linear combinations (contrasts) of the cell means which define the effects associated with a SV. In general, this is a (k x P) matrix where k is the df in the source, and P is the number of cells.

b) $\bar{Y}$ is the (P x 1) vector of the (observed) cell means.

c) $D(1/n_{pp})$ is a diagonal (P x P) matrix whose p^{th} element on the diagonal is the reciprocal of the number of observations in the p^{th} cell.

As already noted, it is a property of full models that the 'usual' constraints define the effects for a SV in terms of straight-forward functions of the marginal means. Before attempting to unravel these functions, it is critical to have a way to express these marginals in terms of the cells.

Definitions of Marginal Means

There are four design related marginals which need to be discussed; these are involved with main effects. But first, some further notational conventions need to be established.

In keeping with an earlier convention, an asterisk (*) denotes subscripts of other possible factors in the experiment. However, another symbol is needed to indicate that these additional subscripts have been appropriately averaged over. The symbol used for this is a hash mark (#). Each dot (.) associated with a mean also implies an averaged over factor, but these relate to factors of explicit interest.

The overall mean, $\bar{y}_{\#}$. The overall mean is an estimator for the location parameter, μ; and is one of the marginals involved in defining crossed main effects. It is not necessarily a simple average where each cell mean is given the same weight, although this is the case with cell balanced experiments like factorials.

In general the overall mean is given by:

$$\bar{y}_{..\#.} = \sum_{i}^{I} \sum_{j}^{J(i)} \cdots \sum_{m}^{M(ij*1)} y_{ij*m} / IJ(i) \cdots M(ij*1). \qquad [19]$$

Examination of [19] reveals that the weight (divisor) applied to each cell mean is a function of the subscripts of that cell. In designs which may lack cell balance, like completely nested designs, these weights are not necessarily equal.

To facilitate the unraveling of the marginals into tracable functions of the cells, it is prudent to express the various marginals as weighted averages. As before, let p(p = 1,2,...,P) be uniquely associated with a specific combination of its M factor levels (ij*m) so that:

$\bar{y}_p = \bar{y}_{ij*m}$ is the mean of the p^{th} cell with associated subscripts (ij*m), and

$c_p = c(ij*m)$

$= 1/IJ(i) \ldots M(ij*1)$

is the corresponding cell weight.

The overall mean can now be re-written from [19] more simply by:

$$\bar{y}\# = \sum_p c_p \bar{y}_p \qquad [20]$$

$$= \sum_p z_p .$$

Like any weighted average, these weights have the property that:

$$\sum_p c_p = 1$$

Crossed factor marginal, $\bar{y}_{\#k'\#}$. A crossed factor with K levels has K distinct marginal means denoted:

$$\bar{y}_{\#k\#} \text{ with } k = 1,2,\ldots,K.$$

These are involved in the definition of main effects for crossed factors.

The overall mean, just discussed, can always be expressed as a simple average of these marginals. That is:

$$\bar{y}_{\#.\#} = (1/K)\ (\bar{y}_{\#1\#} + \bar{y}_{\#2\#} + \ldots + \bar{y}_{\#K\#}) \qquad [21]$$

Each marginal, say $\bar{y}_{\#k'\#}$, has been averaged over only those cells for which the subscript $k = k'$. This suggests that the overall mean given by [20] can be segregated into K terms related to those of the marginal means:

$$\bar{y}_{\#.\#} = \sum_p z_p$$

$$= \left(\sum_{p:\ k=1} z_p + \sum_{p:\ k=2} z_p + \ldots + \sum_{p:\ k=K} z_p \right). \qquad [22]$$

Equating [21] and [22] and multiplying through by K leads to:

$$\bar{y}_{\#1\#} + \bar{y}_{\#2\#} + \ldots + \bar{y}_{\#K\#} = K(\sum_{p:\ k=1} z_p) + K(\sum_{p:\ k=2} z_p)$$

$$+ \ldots + K(\sum_{p:\ k=K} z_p).$$

From this expression, it is clear that corresponding terms on the left and right hand sides are equivalent. Thus, a marginal mean of a crossed factor is given in terms of the cells by:

$$\bar{y}_{\#k'\#} = K(\sum_{p:\ k=k'} z_p). \qquad [23]$$

Deriving the crossed factor marginals in this manner serves to illustrate their relationship to the overall mean. However, this result can be obtained in a more general way. Any marginal mean can be expressed as a weighted average of the cell means; subject to the condition that the weights sum to one. This can be guaranteed by expressing a weighted average in its general form:

$$(1/\Sigma c_p)\Sigma c_p \bar{y}_p = (1/\Sigma c_p)\Sigma z_p$$

where the summations range over the same values of p.

In this particular case, only the cells of p for which k = k' were included in the sum. However, the multiplicative factor (1/K) was included in the definition of the c_p's, reflecting that this factor had already been averaged over. Subsequently, the result of [24]:

$$\bar{y}_{\#k'\#} = (1/(1/K)) \sum_{p:\ k=k'} z_p$$

$$= K \sum_{p:\ k=k'} z_p$$

is identical to that previously obtained in [23].

Nested 'overall mean', $\bar{y}_{(k'*n')\#}$. The 'overall mean' for a nested main effect parallels the role played by the overall mean when defining crossed main effects. However, in this case there is one such marginal associated with each unique set of nesting subscripts, (k'*n').

From the principles established earlier, a specific nested "overall mean' can be expressed as a weighted average of the cell means:

$$\bar{y}_{(k'*n')\#} = \left(1/\sum_{p:(k'*n')} c_p\right) \sum_{p:(k'*n')} c_p \bar{y}_p$$

$$= \left(1/\sum_{p:(k'*n')} c_p\right) \sum_{p:(k'*n')} z_p \qquad [25]$$

where the summations extend only over the cells of p which have the same corresponding subscript values as the nesting subscripts, (k'*n'). The multiplicative constant, $(1/\sum_{p:(k'*n')} c_p)$ is equal to the product of the levels associated with the subscripts (k'*n'), but for computational reasons it is convenient to leave it in its present form.

Nested factor marginal, $\bar{y}_{(k'*n')\#1'\#}$. There is a set of nested marginals associated with each unique set of nesting subscripts, (k'*n').

The number of marginals depends on the number of levels defined within each nesting set. In general, these marginals and their levels may be represented:

$$\bar{y}_{(k'*n')\#1\#} \text{ with } 1 = 1,2,\ldots,L(k'*n').$$

A simple average of nested marginals yields their 'overall mean'. That is:

$$\bar{y}_{(k'*n')\#} = (1/L)\ (\bar{y}_{(k'*n')\#1\#} + \bar{y}_{(k'*n')\#2\#} + \ldots + \bar{y}_{(k'*n')\#L\#}) \qquad [26]$$

where $L = L(k'*n')$.

Since any nested marginal, denoted $\bar{y}_{(k'*n')\#1'\#}$, involves only those cells with nesting subscripts (k'*n') and 1 = 1', expression [25] can be segregated into L terms related to those of each marginal. That is:

$$\bar{y}_{(k'*n')\#} = \left(1/\sum_{p:(k'*n')} c_p\right)\left(\sum_{p:(k'*n')\#1\#} z_p + \sum_{p:(k'*n')\#2\#} z_p + \ldots + \sum_{p:(k'*n')\#L\#} z_p\right) \qquad [27]$$

By equating [26] and [27] and multiplying both sides by L, it can be

seen that corresponding terms are equivalent. It follows that a general nested factor marginal may be represented:

$$\bar{y}_{(k'*n')\#1'\#} = \left(1/ \sum_{p:\,(k'*n')} c_p\right) L(k'*n') \sum_{p:\,(k'*n')1'} z_p \qquad [28]$$

Definition of Main Effects

It was stated earlier that the effects in full models can be expressed as straight-forward functions of the marginal means. Now that the design marginals have been derived, it is a simple matter to express the various effects in terms of the cells. The one remaining step is to manipulate the main effect expressions so that a single weight can be applied to each cell involved in the effect.

Crossed main effect, $\bar{y}_{\#k'\#} - \bar{y}_{\#}$. Each main effect for a crossed factor is estimated by the difference between a specific marginal and the overall mean. These may be denoted, say for factor A with K levels, by:

$$A_k = \bar{y}_{\#k\#} - \bar{y}_{\#}. \quad k = 1,2,\ldots,K$$

Substituting expressions [23] and [20] for a specific effect, $A_{k'}$, yields:

$$A_{k'} = \bar{y}_{\#k'\#} - \bar{y}_{\#} = K \sum_{p:\,k=k'} z_p - \sum_{p} z_p \qquad [29]$$

The right most term involves all P cells, including those for which

k=k'. Segregating these cells into the term on the left divides the cells into two groups; those cells for which k=k' and those where k≠k'. Thus from [29]:

$$A_{k'} = (K-1) \sum_{p:\ k=k'} z_p + (-1) \sum_{p:\ k\neq k'} z_p \qquad [30]$$

Nested main effect, $\bar{y}_{(k'*n')\#1\#} - \bar{y}_{(k'*n')\#}$. Deriving the weights for a nested factor based on the subscripts of the cells closely follows the procedure for crossed factors. Each nested main effect is estimated by the difference between a specific nested marginal and its corresponding 'overall' mean. These are denoted, say for factor B, by:

$$B_{(k'*n')l} = \bar{y}_{(k'*n')\#1\#} - \bar{y}_{(k'*n')\#} \quad l = 1,2,\ldots,L(k'*n') \qquad [31]$$

The number of levels for B depends on the factor(s) which nest it. Letting L = L(k'*n') and substituting expressions [28] and [25] into [31], a specific nested effect becomes:

$$B_{(k'*n')1'} = \bar{y}_{(k'*n')\#1'\#} - \bar{y}_{(k'*n')}$$

$$\left(1/\sum_{p:(k'*n')} c_p\right)\left(L\sum_{p:(k'*n')1'} z_p - \sum_{p:(k'*n')} z_p\right) \qquad [32]$$

Only those cells with the nesting subscripts (k'*n') are involved in this effect. However, the right-most summation in [32] includes all

L levels whereas the adjacent term involves only those for which l=l'. So that each cell appears only once, those cells in the right-most term for which l=l' are segregated into the adjacent one. This gives:

$$B_{(k'*n')l'} = \left(1/\sum_{p:(k'*n')} c_p\right)\left((L-1)\sum_{\substack{p:(k'*m')\\ l=l'}} z_p + (-1)\sum_{\substack{p:(k'*n')\\ l\neq l'}} z_p\right).$$

General main effect. Separate expressions were derived for crossed and nested main effects to facilitate the discussion. However, it is possible to treat a crossed factor as a special case of a nested one. Later it is seen that the cell weights defining an interaction effect can be obtained from the product of its corresponding main effect weights. This leads to the desirability of having a single expression (algorithm) to handle either type of main effect. With slight notational changes, this is now done.

Let all cells having the same value(s) for its nesting subscript(s) be organized into Q distinct sets. Let these sets be represented by S(q), q=1,2,...,Q. And let l=1,2,...,L(q) be the factor levels in the q^{th} such set. Then any specific main effect is given by:

$$B_{(q')l'} = \left(1/\sum_{p:S(q')} c_p\right)\left((L(q')-1)\sum_{\substack{p:S(q')\\ l=l'}} z_p + (-1)\sum_{\substack{p:S(q')\\ l\neq l'}} z_p\right) \quad [34]$$

This is the general form given in [33] for a nested main effect. When Q=1, S(q') includes all P cells. And since $\sum_{p}^{P} c_p = 1$, [34] reduces to [30] as given for a crossed main effect.

Earlier, it was mentioned that the utility of the SS expression [18] was tied to expressing an effect in terms of the cell means. A cursory examination of [34] reveals this was done in terms of a transformed variate, $z_p = c_p \bar{y}_p$. However, from [34] it is also apparent this would not be difficult to do. Before H is displayed, a slight problem has to be resolved. The summations in [34] extend only over those cells involved in the effect. However, in [18]:

$$SS(H) = (H\bar{Y})' \, (HD(1/n_{pp})H')^{-1} \, (H\bar{Y}) \qquad [35]$$

H was defined as a (df x P) matrix involving all P cells. This difficulty is rectified by letting the weights for these excluded cells to be zero. Thus, it is fairly obvious from [34] that an element of H for a main effect, say $h_{l'p}$, is given by:

$$h_{l'p} = \begin{cases} w(L(q')-1)\ c_p & \text{for } p\text{: } S(q'),\ l=l' \\ w\ (-1)\ c_p & \text{for } p\text{: } S(q'),\ l\neq l' \\ w\ (0)\ c_p & \text{for } p\text{: } S(q) \neq S(q') \end{cases}$$

where $w = (1/\ \sum_{p:\ S(q')} c_p\ )$, and $l' = 1,2,\ldots,$ df are the first $(L(q')-1)$ linearly independent effects.

Definition of General Interactions Effects

Because the models considered here are all full models, the means are viewed as single observations in the cells of the design. The effects are all subsequently defined as if the data were orthogonal. In this connection, the weighted averages defining the effects are contrasts of the cell means. That is, any specific effect given by $\sum_{p:\, S(q')} h_{l'p}\bar{y}_p$ has the property that $\sum_{p:\, S(q')} h_{l'p} = 0$. And further, pairwise contrasts for distinct sources of variation are orthogonal in the sense that $\sum_{p:\, S(q)} h_{l'p}h_{m'p} = 0$. This leads to generating the contrasts for interaction effects through their specific main effect contrasts just as is possible for orthogonal data.

To show how this is done, it is easiest to begin by letting the three components of a main effect element given by [36] to be represented by:

$$h_{l'p} = w\, g_{l'p}\, c_p$$

This can be generalized to all such elements by:

$$H = wGC \qquad [37]$$

where: w is the scalar, $(1/\sum_{p:\, S(q')} c_p)$

G is a (df(q') x P) matrix whose element, $g_{l'p}$, is given by:

$$g_{l'p} = \begin{cases} df(q') = (L(q')-1) & \text{for } p\text{: } S(q'),\ l=l' \\ (-1) & \text{for } p\text{: } S(q'),\ l\neq l' \\ (0) & \text{for } p\text{: } S(q) \neq S(q') \end{cases}$$

C is a (P x P) diagonal matrix where the p^{th} element on the diagonal is c_p; defined as before.

The elements of C are a set of proportional weights related to the cell structure of the design which put the cells on an equal basis with one another. And the scalar w is the reciprocal of the sum of the weights associated with the cells involved in the specific effect. Hence, w is a scale factor to guarantee these already proportional weights sum to unity.

Within a design, the only component of H which makes an effect unique is the matrix G. And it is only this component which is necessary to generate the G matrix for interaction effects.

The procedure for obtaining the matrix G for a general two-way interaction can be illustrated in the following way. Let:

> ${}_A G$ be the $({}_A df(q') \times P)$ matrix for the main effects of factor A in the q'^{th} set from the P cells. And let $i = 1,2,\ldots,{}_A df_{(q')}$ correspond to the effect associated with a row in ${}_A G$.

Let ${}_B G$ be the similarly defined $({}_B df_{(q')} \times P)$ matrix for the factor B with its rows denoted by $j = 1,2,\ldots,{}_B df_{(q')}$. The interaction matrix ${}_{AB} G$ is then the $(({}_A df_{(q')})({}_B df_{(q')}) \times P)$ matrix obtained by miltiplying out the rows of ${}_A G$ and ${}_B G$. That is, the row of ${}_{AB} G$ associated with

the interaction effect $AG_{(q')i'j'}$ is given by:

$$AB^{g}(1',j')p = ({}_{A}g_{i'p})\ ({}_{B}g_{j'p}).$$

Repeating this for all i',j' pairs for which $i' \geq j'$ gives ${}_{AB}G$ for the set of linearly independent interaction effects.

This illustrates the procedure for two-way interactions, however this same rationale can be generalized to higher order interactions. Each row of ${}_{AB}G$ forms a contrast of the cell means in that

$w \sum_{p:\ S(q')} g_{(i',j')p}\ C_p = 0$ as before. Another factor composed with ${}_{AB}G$ yields the G matrix for a three-way interaction, again defining a set of contrasts; and so on.

In summary, there is a matrix G for each factor appearing in an interaction. A particular row in each of these matrices is associated with a specific main effect, and can be found by applying the definition of G in [37]. When one row from each of these matricies is selected and their corresponding column elements multiplied together, a single row of the interaction matrix is found. (The specific interaction row so determined is identified by the specific main effect rows selected.) The whole interaction matrix G can be build up by repeating this basic procedure until all unique rows have been generated. Once this has been done, the matrix H is completely determined by H = wGC, where w and C are defined as in [37].

Computational Methodology

Applying the SS expression

$$SS(H) = (H\bar{Y})' \, (HD(1/n_{pp})H')^{-1} \, (H\bar{Y}) \qquad [38]$$

to a μ-model approach on the cell means requires the matrix H to be specified. An algorithm to generate this matrix has been derived for both main effects and interactions in full design models.

The primary difficulty in deriving H is ultimately tied to the manner the marginal means are defined in designs which may lack cell balance. Specifically, this can arise in designs which involve nested factors. The source of this difficulty comes from the fact that the contribution of each cell mean in a marginal is not the same. Subsequently, it is more tracable to express the marginals, and the effects derived from them, in terms of the transformed variables, $z_p = c_p \bar{y}_p$, than the cell means themselves. The c_p's are the cell specific and proportional contributions of the cell means; and depend solely on the cell structure of the design. It is not surprising, therefore, that certain computational advantages accrue from the use of this transformation. This is now done.

The matrix H can be decomposed into three separable components:

$$H = wGC \qquad [39]$$

where from [37]:

C is a (P x P) diagonal matrix of these proportional weights.

the p^{th} element on the diagonal is given by $c_p = 1/IJ(i) \ldots M(ij \ldots 1)$ and corresponds to the p^{th} mean, $\bar{y}_p$, in $\bar{Y}$.

w is a scalar which guarantees these weights sum to unity.

and G is a (df x P) matrix of generated weights associated with the linearly independent effects in a source of variation.

The effects, given by $H\bar{Y}$ can then be written:

$$H\bar{Y} = (wGC)\bar{Y}$$

$$= wG(C\bar{Y})$$

$$= wGZ \qquad [40]$$

where Z is a (P x 1) vector of the transformed variates ($z_p = c_p\bar{y}_p$). The SS associated with these effects is found by substituting [40] and [39] into [38]. That is:

$$SS(H) = (H\bar{Y})' \, (HD(1/n_p)H')^{-1} \, (H\bar{Y})$$

$$= (wGZ)' \, ((wGC)D(wGC)')^{-1} \, (wGZ)$$

$$= (w^2/w^2) \, (GZ)' \, ((GC)D(C'G'))^{-1} \, (GZ)$$

$$= (GZ)' \, (G(CDC')G')^{-1} \, (GZ)$$

$$= (GZ)' \, (GS(c_p^2/n_p)G')^{-1} \, (GZ) \qquad [41]$$

where S is a diagonal (P x P) matrix whose p^{th} element on the diagonal is (c_p^2/n_p).

The corresponding variance of each effect estimated in $H\bar{Y} = w(GZ)$ is found on the diagonal of the variance-covariance matrix. This

matrix follows from [14] and is given by:

$$\text{Var-Cov } (H\bar{Y}) = w^2(GSG')s_e^2$$

Degrees of Freedom

In full design models the degree of freedom for the various sources of variation are well known. But reviewing this topic serves to clarify a point about the computational methodology for nested terms.

A crossed factor with K levels has (K - 1) degrees of freedom; and those for a crossed interaction are found by taking the product of the degrees of freedom in the individual factors composing it. In either case, the degrees of freedom are reflected by the number of linearly independent effects, or rows in G.

For nested terms, the effects involve only those cells which have the same value(s) for its nesting subscript(s); and hence, the cells can be divided into Q distinct sets. The sums of squares for the effects within each set are, of course, not independent. However, the SS between these Q sets are independent because they involve completely different cells. This means that the sums of squares for each of the Q sets can be computed separately and simply added together to obtain the sums of squares for the term. Within each of the Q sets, the degrees of freedom are computed like a crossed term; and when added together give the degrees of freedom in the source.

Summary

The basic methodology for the analysis of full, design models was developed in this chapter. This procedure is framed as a μ-model approach on the cell means. Because the models are full, the design related marginals are defined as if the cell means are single observations. Subsequently, the effects from one source of variation form contrasts which are orthogonal to the contrasts of any other source, just as if the cells were observation balanced. This leads to a simple algorithm for generating the main effect contrasts from which the interaction contrasts can also be derived.

A definite computational advantage accrues by applying a transformation on the cell means which put the cells on a proportionally equivalent basis with one another. This methodology is incorporated in an interactive terminal version of a computer program to analyze these design models.

SUMMARY AND CONCLUSION

The objective of this thesis was to develop a method for the analysis of full, design models and to incorporate this procedure into a computer program. The methodology was developed in the conceptual context of a μ-model approach on the cell means. A Fortran IV program makes tangible the algorithm devised in the previous chapter. The user documentation for this program is found in Appendix A, and the program listing is on file with the Applied Statistics/Computer Science Department at Utah State University.

The program is interactive and is designed to be executed from a remote terminal. The program was validated by analyzing a variety of models and data situations. The results agreed with those obtained from a multiple regression analysis using Statpac, a statistical package available at Utah State University.

The application of this program is primarily intended for the exact analysis of non-orthogonal data from full factorial or completely nested designs. However, the same basic algorithm allows the analysis of full, mixed models, which characteristically involve interactions between crossed and nested factors. Exact tests of hypotheses are unavailable for mixed models with unbalanced data because the distribution of the various sums of squares are unknown. Despite this, the extension to mixed models provides the fundamental quantities for estimating components of variance, and is potentially useful for analyzing designs with orthogonal data.

With orthogonal data, the analysis obtained for experiments handled by the program is not restricted to cases where each source of variation is specifically delineated. Other analyses can be obtained from pooling appropriate sums of squares. Anderson (1974) gives a good account of how this is done for the split-plot and other designs. Application of this program to orthogonal data would, of course, be less efficient computationally than one using orthogonal data methods for the same design. However, to an individual well versed in these procedures, there are advantages in having a single program which is easy to use and requires little or no set-up time.

For designs with non-orthogonal data, the sums of squares derived from this procedure are limited to those obtained from a model containing all possible interactions. Thus, the sum of squares for each source of variation found by this algorithm is adjusted for every other source.

Computing these same sums of squares using regression (or general μ-model) procedures requires the inversion of a matrix whose rank equals the degrees of freedom in the model. This can become a practical difficulty in large multi-factor experiments. Since the approach taken here directly evaluates sums of squares for individual sources, and this only requires the inverstion of matrices whose rank equals the source degrees of freedom, this problem is not as limiting. Thus, this program can be applied when a full model analysis is desired, or as a check to determine if certain interactions are indeed negligible, thereby reducing the model degrees of freedom for subsequent analysis.

BIBLIOGRAPHY

Anderson, V.L. and McLean, R.A. (1974), Design of Experiments, New York: Marcel Dekker.

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

Bryce, G.R., and M.W. Carter. (1974), Compstat 1974 - Proceedings in Computational Statistics, Vienna: Physica-Verlag.

Francis, I. (1973), "A Comparison of Several Analysis of Variance Programs," Journal of American Statistical Association, 68, 860-865.

Greybill, F.A. (1961), An Introduction to Linear Statistical Models, Vol. I, New York: McGraw-Hill Book Co.

Kutner, M.H. (1974), "Hypothesis Testing in Linear Models (Eisenhart Model I)," The American Statistician, 28, 98-100.

Morrison, D.F. (1967), Multivariate Statistical Methods, New York: McGraw-Hill Book Co.

Neter, J., and Wasserman, W. (1974), Applied Linear Statistical Models, Homewood: Richard D. Irwin.

Searle, S.R. (1971), Linear Models, New York: John Wiley and Sons.

Snedecor, G.W. (1934), "The Method of Expected Numbers for Table of Multiple Classification with Disproportionate Subclass Numbers," Journal of American Statistical Association, 29, 389-393.

Speed, F.M. (1969), "A New Approach to the Analysis of Linear Models," Ph.D. Thesis, Texas A and M University.

Yates, F. (1933), "The Principle of Orthogonality and Confounding in Replicated Experiments," Journal of Agricultural Science, 23, 108.

Yates, F. (1934), "The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes," Journal of American Statistical Association, 29, 51-66.

APPENDICIES

APPENDIX A

User Documentation for ANOVA

A. Description

This program may be used to compute the analysis of variance for full, design models with non-orthogonal data. This includes factorial, completely nested, and mixed models with one or more observations per cell.

It will on control print the mean and standard error for each cell, along with its associated factor level subscripts. The program automatically computes and prints the between and within cell mean squares. Their ratio is generally used to test the overall significance of the model.

The degrees of freedom and mean square of specific sources of variation are then computed for model terms the user supplies. The effect coefficients for each term and its single degree of freedom partitions are under a print control option left to the user.

The program is writtein in Fortran IV. Logical units 5 and 6 are used as input and output files in the terminal version. Logical unit 15 is reserved as an alternate input file when the data is read from disk. For this option, a file equate is required prior to program execution. A program listing is on file with the Applied Statistics/Computer Science Department at Utah State University.

B. Methodology

The program applies the 'usual' constraints to the cell means, where each mean is treated as a single observation in a cell of the experiment. However, unlike the method of unweighted means, the pooled within cell variance is weighted by the individual cell frequencies. This procedure provides an exact analysis for full models.

A general outline of the procedure begins with the 'usual' constraints for main effects, like

$$\sum_i \alpha_i = 0 \qquad \text{where } \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...} \qquad [1]$$

$$\sum_l \beta_{(i..k)l.} = 0 \quad \text{for all } (i..k)$$

$$\text{where } \hat{\beta}_{(i..k)l.} = \bar{y}_{(i..k)l.} - \bar{y}_{(i..k)..}$$

are respectively defined in terms of means which are 'evenly' weighted over the remaining factors. For design models these expressions can be unraveled and expressed as linear combinations (contrasts) of the observed cell means. Interactions, similarly defined, can be found from its main effect contrasts. The sum of squares for each source of variation can subsequently be computed separately.

The sum of squares (SS) associated with a particular hypothesis, say

$$H_o: \alpha_i = 0 \qquad \text{for all } i,$$

is written as

$$H_o: H\bar{Y} = 0$$

and can be computed by:

$$SS(H) = (H\bar{Y})' (HD(1/n_p)H')^{-1} (H\bar{Y}).$$

D is a diagonal matrix of reciprocals of the cell frequencies. The vector $(H\bar{Y})$ with k degrees of freedom has k corresponding elements in the solution vector of the dummy variable regression approach. The degrees of freedom and sum of squares for a source of variation derived in this manner are subsequently the same as those found due to subsets using regression. In each case the SS obtained by this method is adjusted for every other source of variation in the full (design) model.

C. Input

1. CONTROL

The program is called for by one of the following commands:

(1) E$ANOVA

(2) E$ANOVA;FILE FILE 15 (KIND = DISK, TITLE = , ...)

depending on whether the formatted input data is to be read from the terminal or disk.

The program first requests the user to enter the number of factors. This may vary from 1 to 6 in the current version, but may be increased by changing dimension statements.

The program will then request in turn the maximum number of levels for each factor, starting with the first or i^{th} subscript, the second or j^{th} subscript, etc.; until all such maximums have been specified. Each is supplied free format and while none may exceed 40 levels, their product may not exceed 200. After the data input device is given, the data format is required. The format may not exceed 80 characters and should begin with a left parenthesis in column one.

2. DATA RECORDS:

The input data can be one of two types. In either case, the input source can originate from the terminal or disk. The first option takes the raw data from which the cell frequencies and means are tabulated. The second option allows the user to input the cell frequencies and means directly. This latter option requires an estimate of the within cell (error) variance. The expected input form for each of these options is discussed in turn.

a. Raw Data

The data read statement expects the codes for the cell subscripts to be followed by an observed value of the dependent variable. The Fortran read statement for a data record is of the form:

```
READ(IRD,FMT,END=14) (NT(L),L=1,NFACT),YT.
```

The codes, representing the levels for each factor, are required to be one of the consecutive integers beginning with one for the first or lowest level; and extending sequentially to the last or highest level.

It is unnecessary to sort the data records as the program uses these cell subscripts as they are read in to compute the proper cell address for each observation. Because of the way each address is computed, the factor code associated with the first or i^{th} subscript must be read first, followed by the second or j^{th} subscript code, etc.; until all cell subscripts have been read. It is usually possible to properly arrange this read order by using 'T' format specifiers in the data input format. If the data is entered from the terminal, the last record is followed by a ?END.

b. Cell Data

For this option, the data read statement expects the codes for the cell subscript(s) to be followed by the cell frequency, then the cell mean. The Fortran read statement is of the form:

```
READ(IRD,FMT,END=14)(NT(L),L=1,NFACT),KN,YT.
```

The cell subscripts are subject to the same conditions as outlined above for raw data.

3. ENTERING MODEL TERMS:

The user may terminate the program by a '?END' anytime after the basic ANOVA table is printed. At this point the program is in a loop which reads in a model term, decodes it, and performs the necessary computations. After the results are printed, the next model term may be entered or the program terminated.

The expected input form closely resembles terms as they ordinarily

appear in analysis of variance models. In general, this form is given by:

TEXT/(NESTING SUBSCRIPTS: IF ANY) CROSSED SUBSCRIPTS/P.

The text may consist of any string of alphanumeric characters and is followed by a slash. A slash is used as a delimiter. Any subscripts enclosed within parentheses are considered nesting subscripts by the program, hence parentheses appear only for nested model terms. The non-nested or crossed subscripts are then given, and are followed by another slash. The program scans for the subscript information between the two right-most delimiters, so it is permissible for the text to contain slashes. The text itself is ignored. The letter P, immediately following the right-most delimiter, causes the estimated effect coefficients and their associated mean squares to be printed.

The model terms may be entered in any order and each may consist of up to 24 characters. A string of 24 dashes is provided as a guide to accommodate this constraint.

To illustrate the input expected for various model terms, consider the hypothetical model:

$$y_{ijklm} = u + A_i + B_j + C_{(ij)k} + D_{(ijk)l} + AB_{ij} + e_{ijklm}$$

The format appropriate to each source of variation in the above model is given below:

1) A/I/

2) B/J/P

3) C/(IJ)K/P

4) D/(IJK)L/

5) AB/IJ/

Note that the effects will be printed in 2) and 3).

Current dimensioning restricts the degrees of freedom a term may have. For terms involving only non-nested factors this limit is 40. Because of the way nested factors are handled, there can be up to 40 degrees of freedom within any specific set of nesting subscripts. The subscript set associated with the effects are printed to the right of the F ratio column in the output for nested effects. The flow diagram (Attached) summarizes the various input-output options available in the program.

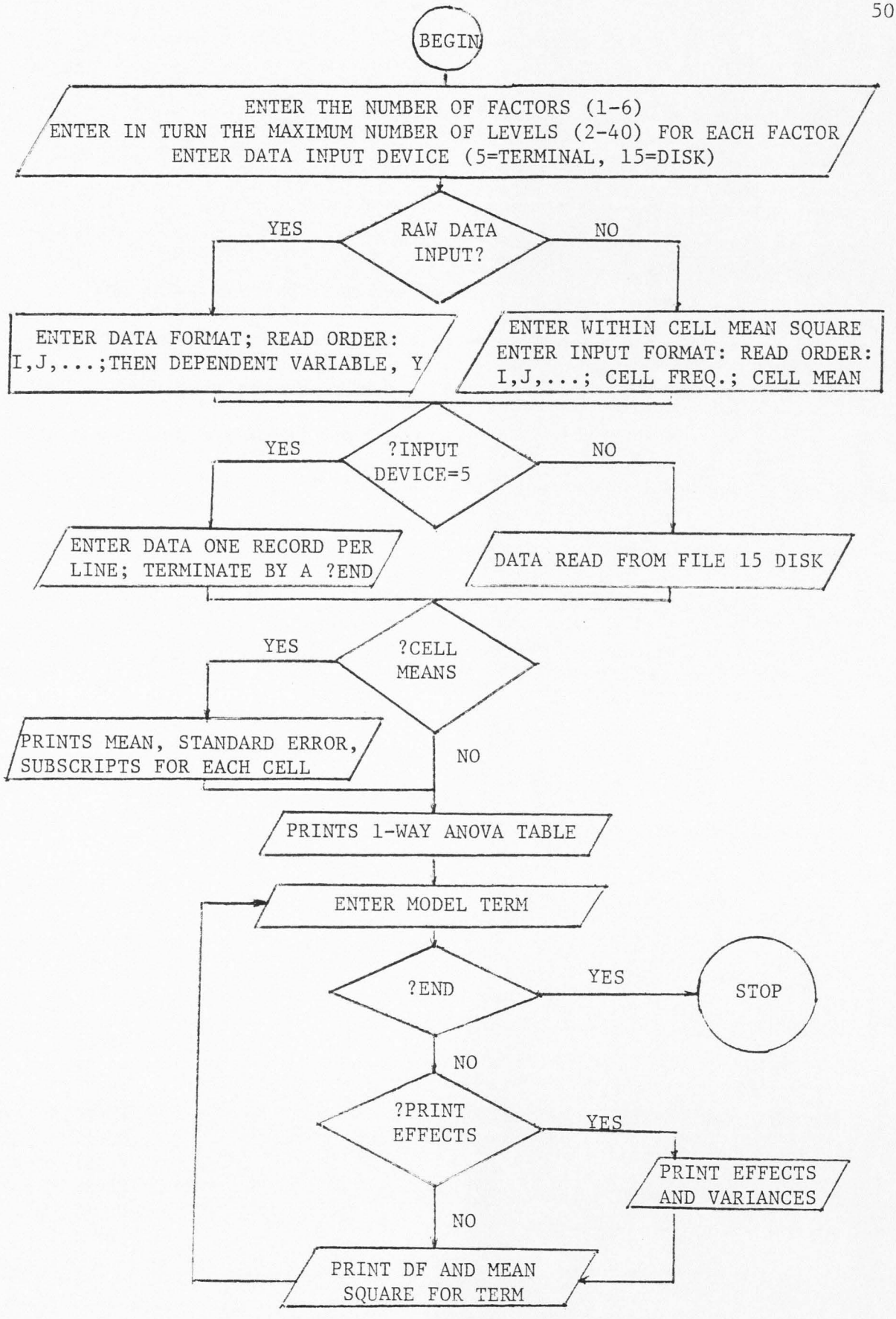
D. Output

The output is labeled and should present no difficulties to persons familiar with analysis of variance. Particular care should be taken, especially with unbalanced data, to ensure proper F tests. F-ratios are computed only for the single degree of freedom partitions. In each case the denominator is the within cell mean square.

E. Sample Problem

This is an example of a 2 X 3 factorial represented by the model:

$$y_{ijk} = u + A_i + B_j + AB_{ij} + e_{ijk}$$



Flow Diagram for Input/Output Options

where

$i = 1,2$

$j = 1,2,3.$

The data, entered from the terminal one record per line, is organized so that the level of A is given first followed by the level of B. Each pair of subscripts is followed by an observed value of the dependent variable. The data records are terminated by a ?END.

```
E$ANOVA
#RUNNING 0015

        ENTER THE NUMBER OF FACTORS (1-6)
#?
2

        ENTER THE MAXIMUM CODE FOR EACH (1-40)
  I=2
  J=3
         ENTER DATA INPUT DEVICE(5=TERMINAL, 15=DISK)5
         RAW DATA INPUT? (Y OR N)Y
         ENTER DATA FORMAT: ORDER I,J,...; THEN Y
(2I1,F3.0)

 ENTER DATA ONE RECORD PER LINE.  TERMINATE BY A ?END
11 26
11 16
12 18
13 04
13 14
21 39
21 26
21 28
22 19
22 05
23 29
?END

#
```

```
 CELL MEANS? (Y OR N)
#?
Y
```

NO.	OBS	CELL MEAN	CELL S. ERROR	SUBSCRIPT
1	2	.210000E+02	.544059E+01	1 1
2	1	.180000E+02	.769415E+01	1 2
3	2	.900000E+01	.544059E+01	1 3
4	3	.310000E+02	.444222E+01	2 1
5	2	.120000E+02	.544059E+01	2 2
6	1	.290000E+02	.769415E+01	2 3

SOURCE			DF	MEAN SQUARE	F
TOTAL			10	111.45	
AMONG			5	163.71	2.7654
WITHIN			5	59.200	
	COEFFICIENT	VAR(COEFF)			
FACTOR A /I/P					
1	-4.00000	0.106481	1	150.261	2.5382
			1	150.261	
FACTOR B /J/P					
1	6.00000	0.175926	1	204.632	3.4566
2	-5.00000	0.231481	1	108.000	1.8243
			2	112.982	
A*B INTERACTION/IJ/P					
1	-.100000E+01	0.175926	1	5.68421	.96017E-01
2	7.00000	0.231481	1	211.680	3.5757
			2	115.509	

```
?END

#
#ET=4:27.2 PT=1.1 IO=0.4
```