A COMPUTER LANGUAGE FOR THE ANALYSIS OF VARIANCE

David J. Armor, Harvard University

Computer processing methods in the social sciences have undergone radical change during the sixties. The change has been along two major dimensions. First, increased computing capability has led to considerable development in numerical analysis and statistics. It has become possible to handle such complicated computational problems as matrix inversion and characteristic equation solutions. The consequence is increased stress on multivariate methods which have been relatively ignored to this point. Among social scientists, interest has grown in the potential applications of such methods as multiple regression, factor analysis, and discriminant function analysis. Aside from their conceptual relevance, for the first time it is practical for him to consider such methods since "canned" programs remove the requirement of understanding the necessary mathematical operations. Several programs for the multivariate analysis of variance, one of the more complicated methods, have become widespread.[1]

Second, at the beginning of this decade the few social scientists who analysed their numerical data by computer relied on fairly simple "canned" programs, such as those supplied in Dixon's BMD series.[2] While many of these programs provided the social scientist with considerable computing power, they were very often quite restricted in their assumptions about data format, transformations, the number of parameters, and the like. As more social scientists began utilizing the computer, the demands for flexibility and generality spurred efforts to develop more comprehensive computer systems for data analysis. Some of these systems have taken on the character of "natural" languages which make it possible for the researcher to request many different and complex analysis with a minimum of knowledge about computer technology.[3]

I would like to describe the development of a data analysis system called Data-Text which takes advantage of these two trends.[4] In particular, I would like to describe the way in which the Data-Text is a natural computer language with instructions expressed in words or terms which can be understood by the average social science researcher. The language allows for a wide variety of data input and transformation. Moreover, the statistical analysis routines, particularly those for the analysis of variance, contain many options and combinations which are ordinarily found only in separate canned programs. The existence of these features makes it relatively easy for social scientists to request quite complicated analyses of variance on a great variety of data types.

Before presenting the analysis of variance language, it will be necessary to describe briefly the overall Data-Text language. All of the statistical routines in Data-Text take advantage of the preliminary Data-Text features for the definition and labeling of variables. Rather than describe the whole language in detail, an example will be used to illustrate the relevant features. Let us assume a study involving any number of subjects and several sets of variables: background variables, such as sex, age, education and an ability test score; treatment variables, such as drug and stress conditions; and dependent variables, such as blood pressure, pulse rate, and a test battery of 10 yes-no questions measuring anxiety symptoms.

We assume that the data is punched with two cards per subject (any number would be possible). If UNIT refers to a subject identification field, and if COL refers to column number (and COL.../n)

```
*DECK       SAMPLE DATA FOR ANOVA
*CARD(1)/  UNIT = COL(1-4), CARD = COL(5)
*CARD(2)/  UNIT = COL(2-5), CARD = COL(79)
*SEX = COL(9/1) = SEX OF RESPONDENT(MALE/FEMALE)
*EDUC = RECODE(A)  COL(36-37/2) = EDUCATION(GRAM/HS/COLL)
*AGE = COL(10-11/1)
*CODE(A) = (1-4=1/5-7,9=2/8,10-12=3/OTHERS=BLANK)
*DRUG = COL(15/2)=DRUG CONDITION(PLACEBO/ASPIRIN/CODEINE)
*BLP(1-3) = COL(41-43,45-47,49-51/2) = BLOODPRESSURES
*VAR(4) = COL(6/1)+1 = STRESS(LOW/HIGH)
*VAR(6) = COL(71-72/1)=ABILITY TEST SCORE
*VAR(7) = (VAR(6)/AGE) *100 = IQ
*PULSE(1-3) =COL (62 63,64-65,66-67/2) IF DRUG = 2, 3 = PULSE RATES
*ITEMS(1-10) = COL( 51(10)/1) = ANXIETY ITEMS
*ANXIETY = SUM ITEMS(1-10) = ANXIETY INDEX
*PRINT UNIT IF VAR(7) GREATER THAN 200
*COMPUTE FREQUENCIES
*COMPUTE CORRELATIONS(6,7,ANXIETY BY BLP(1-5), PULSE(1-3)) TEST
*COMPUTE FACTORS (ITEMS(1-10)), MAX = 3, NOSCORES
*COMPUTE REGRESSION (4, ANXIETY ON SEX, EDUC, 6, 7,), RESIDUALS
*COMPUTE CROSSTABS (4, DRUG BY EDUC BY SEX), TEST, GAMMA
*COMPUTE PLOTS (7, ANXIETY BY BLP(1-3))
(data Deck or a READ TAPE instruction)
*END
```

refers to columns in the nth card), then the following Data-Text instructions are sufficient to define the variables, with the transformations indicated, and to compute a variety of statistical analyses (each line might represent either a punched card or a line typed at a console):

Each of the statistical analyses requested by the *COMPUTE instructions is a separate routine loaded by the Data-Text system. This makes it possible to perform several quite distinct analyses with a single computer run. On a console, it would be possible to get the results of each analysis before requesting another. The results are printed out making full use of the variable numbers, variable names, and category names.

For example, one face of the three-way crosstab appears as:

One of the keys to the simplicity of the *COMPUTE instructions is the choice of default conditions which correspond to the most common usage of a given analysis. For example, in all routines missing observations are always assumed by default; in regression, stepwise is assumed; in factor analysis, principle components is assumed. The options available are specified separately by users who desire them. Aside from its simplicity, this approach avoids cluttering up a print-out with information not useful or not meaningful to a user (e.g., a full inverted correlation matrix of the independent variables in regression analysis).

The *COMPUTE instructions for the analysis of variance routine in Data-Text are somewhat more complicated than the instructions presented so far. The main reason is that the routine is fairly generalized so that a wide variety of de-

CONTINGENCY TABLE 1

CELL PERCENTS BASED ON COLUMN SUMS
SUBTABLE OF UNITS WITH MALE ON SEX OF RESPONDENT

		EDUCATION			TOTAL	PERCENT
		GRAM	HS	COLL		
VAR(4) STRESS	LOW	22.2 4	48.6 18	68.2 15	37	48.1
	HIGH	77.8 14	51.4 19	31.8 7	40	51.9
TOTAL		18	37	22	77	
PERCENT		23.4	48.1	28.6		100.0

CHISQUARE = 8.388 WITH 2 DF (SIGNIFICANT AT THE .015 LEVEL)
GAMMA = -.529

signs can be handled. The routine is designed to handle a large number of factors (or classificatory variables) with any number of levels on each, the main restriction being the available computer memory.

The routine handles factors which are crossed or nested, or any combination of such. In these designs, UNIT's (or subjects) are assumed to be nested within the cells generated by the classification structure. For example, using the variables derived in our example,

*COMPUTE ANOVA (SEX BY EDUC), ANXIETY

takes the anxiety score as the dependent variable and sex and education as the factors. The operator "BY" indicates a factorial design, and the number of levels is taken from the number of categories given in the variable definition of each. Thus, this would be a 2 X 3 factorial design with replications.

If several univariate analyses with the same design are desired, the instruction would be

```
*COMPUTE ANOVA(SEX BY EDUC),VAR(4,7),ANXIETY
```

Three separate anovas would be carried out, one at a time, for stress, IQ, and anxiety.

There are two general default assumptions in these examples which apply to certain other examples as well. As we said, UNIT's (subjects) are assumed to be nested within each cell. There may be, however, unequal numbers of observations per cell, so that non-orthogonal factorial designs can be analyzed. The method used is that of unweighted means.[5]

The second default is fixed effects; i.e., the levels of each factor are assumed to be a universe of factor levels. If the levels are sampled, the sampling fraction -- or the option R (random) if sampling from an infinite universe -- can be placed after the VAR number:

```
*COMPUTE ANOVA(SEX BY EDUC(R)),ANXIETY
```

would cause education to be treated as a random effect, and this analysis would be handled as a mixed model.

Factors which are nested are indicated by the operator "within". Assume we defined two additional variables as follows:

```
*STATE=COL(77/1)=(NY/MASS)
*CITY=COL(78/1)=(ALBANY/NYC/BOSTON/SPRING)
```

City is nested within state, and the ANOVA request

```
*COMPUTE ANOVA(CITY WITHIN STATE),ANXIETY
```

would cause the appropriate nested design analysis. In these designs UNIT's are assumed to be nested within cells.

Any number of factors (up to 10), either fixed or random, can be combined in an expression and the correct analysis will be carried out. Parentheses are used for clarity of the nesting relationships:

```
*COMPUTE ANOVA(EDUC BY(CITY WITHIN STATE)),ANXIETY
```

would be a three-factor design, with education crossed by both city and state.

The testing of effects is made possible by implementing the Tukey-Cornfield rules for finding the correct denominators for F-tests.[6] These rules cover most combinations of fixed or random and crossed or nested factors. Special options are available in the event that a denominator cannot be found for a given test.

In many behavioral science applications, subjects are measured several times, with a subject becoming his own control. Examples are survey panel studies and learning experiments. These set-ups are often termed "repeated measure" designs. They present special problems for the Data-Test system. For example, assume that the pulse rate variables, PULSE(1-3), were actually measured at three different times, and it is desired to test for changes over time. The usual approach is to assume a factorial design with UNIT's crossed by time and with one observation per cell. However, the time factor is not an explicit Data-Text variable, as were the factors in our previous examples; the levels of time are implicit in the existence of three pulse rate measures. Moreover, the dependent variable, pulse, is not a single Data-Text variable. To indicate this type of design, we provide special *FACTOR and *MEASURE definition instructions.

```
*COMPUTE ANOVA(UNIT BY A),MEASURE(1)
*FACTOR(A)=TIME(TIME1/TIME2/TIME3)
*MEASURE(1)=PULSE(1=A1/2=A2/3=A3)=PULSE RATES
```

The factor instruction gives the structure and labels for the factor, and the measure instruction relates the dependent variable to the levels of the factor (A1,A2,A3). In these designs, the default is that UNIT's are a random effect, so that this example represents a mixed model.

If we assume that there was a second repeated measure factor, say a treatment condition of some kind, and additional pulse rate measurements, then the following instructions would specify a three factor design with UNIT's crossed by time and by treatment:

```
*COMPUTE ANOVA(A BY B),REPEATED MEASURE(1)
*FACTOR(A)=TIME(TIME1/TIME2/TIME3)
*FACTOR(B)=TREATMENT(COND1/COND2)
*MEASURE(1)=Pulse(1=A1,B1/2=A2,B1/3=A3,B1/    *
*   4=A1,B2/5=A2,B2/6=A3,B2)=PULSE RATES
```

The option REPEATED on the *COMPUTE instruction has the effect of crossing UNIT by every factor within the parenthetical design specification.

More complex designs can be requested which have some factors crossed by UNIT and other factors within which UNIT's are nested. An example might be

```
*COMPUTE ANOVA((UNIT WITHIN SEX)BY A),   *
*                MEASURE(1)
*FACTOR(A)=TIME, etc.
```

In this case we have the repeated measure assessment -- UNIT's by time -- carried out on both males and females. Thus, UNIT's are nested within sex, but time is crossed by sex and by UNIT.

All of the designs discussed can have covariates specified, and the appropriate analysis of covariance will be computed.

```
*COMPUTE ANOVA(SEX BY EDUC),ANXIETY/VAR(6)
```

will treat VAR(6), ability score, as a covariate. The results include the regular tests for anxiety, the covariance or regression test, and tests for the anxiety effects after adjusting for VAR(6). A covariate (or any number of covariates) can be specified on any of the designs discussed earlier.

The routine can also handle the generalized multivariate case. If one has several dependent variables which are to be tested simultaneously, (the 10 anxiety items, for example), then the option MANOVA on the following instruction will cause a multivariate analysis:

*COMPUTE MANOVA(SEX BY EDUC),ITEMS(1-10)

The results include a multivariate test for each effect implied in the design using the likelihood ratio criterion.[7] The univariate tests are also given. The MANOVA option can be used with any of the designs discussed so far, including the covariance case.

The output display in both the covariance and the MANOVA cases include the appropriate vectors of cell and marginal means, and the within-cell standard deviations and correlation matrices. The instruction

*COMPUTE CROSSTAT(SEX BY EDUC),ITEMS(1-10)

will produce just this display part without the univariate and multivariate testing. The only difference is that all marginal means will be weighted if cell N's are not equal. This option makes it easy to get basic statistics and correlation matrices within a complex grouping structure.

As in the other statistical routines in Data-Text, considerable attention is given to the problems of missing observations in the various analysis of variance designs. Missing observations on a single dependent variable are handled by treating the design as non-orthogonal; i.e., unequal numbers of cases per cell. If the problem is multivariate, as in MANOVA, CROSSTAT, or covariance, missing observation cross-products matrices are accumulated within cells and pooled to form an estimate of the population correlation matrix.

For the repeated measures case, the problem is somewhat more complicated since a missing observation is tantamount to a missing cell. The default procedure adopted is one of iterative least squares estimation of missing values for a given UNIT using the marginal means for that UNIT.[8] The user may select an option to omit UNIT's with missing observations for both the repeated measure and the multivariate cases.

There is not sufficient space to show a detailed example of the planned printed output of the results; instead, I shall summarize the major contents of the output for the various designs.

1) For non-repeated measure designs, the standard deviations, and counts will be displayed in a tabular form similar to the CROSSTAB table shown earlier. Full use will be made of variable and category labels. All possible marginal means will also be displayed.

2) If requested, effect estimates will be shown in tabular form similar to the display of means.

3) In the MANOVA, CROSSTAT, and covariance cases, the within-cell correlation matrix of the dependent variables and covariates will be printed. In the covariance case, adjusted effect estimates will also be displayed.

4) An analysis of variance table will be produced showing source, degrees of freedom, sums of squares, variance components, F-ratios (where possible), and significance levels. In the covariance case, both the original and adjusted anova tables will be shown. In the MANOVA case, the likelihood ratio criterion will be printed.

5) If an F-test cannot be found for a given effect using the Tukey-Cornfield rules, a table of expected mean squares will be produced to aid the researcher in making pseudo-F ratios.

Our present plans do not call for handling more complex designs. For example, the routine will not provide a solution for factorial designs with missing cells or nested designs with unequal numbers of nests. Moreover, there is no provision handling such special designs as Latin squares. Future plans do call for the addition of an option for pooling mean squares for the purpose of combining or deleting various effects in the model, and options for testing special comparisons among main effects.

Obviously, the goal of a simplified language means some sacrifice in the scope of the routine, although the present routine will handle the most common designs. The main purpose of the Data-Text system is to make complex methods available to the average social science researcher. Many researchers avoid analysis of variance because of the complexities involved in learning the computer procedures or because they must learn complex statistical terminology. Hopefully, the gain should be increased utilization of an extremely powerful technique.

Since the Data-Text project and the analysis of variance routine are still in the developmental stages, we welcome critical comments and suggestions. The analysis of variance procedures adopted are sufficiently complicated to deserve continued scrutiny and revisions when necessary.

NOTES

* The Data-Text system was developed originally under the direction of Dr. A.S. Couch. Principle associates were David Peizer and Mary Hyde. Peizer also designed the original plans for the analysis of variance routine. Principle programmers for the routine have been Rod Montgomery, Frank Benford, and Karl Deirup. Donald Rubin has given further statistical assistance with the help of staff members in the Department of Statistics, Harvard University. The current version runs on the IBM 7090/94. A project to revise the current version for the IBM 360 and other computers is being supported by an NIMH grant (MH-15884-01), with the author as principle investigator.

1. For example, the MANOVA program from Dean J. Clyde, et al, "Multivariate Statistical Programs," Biometric Laboratory, University of Miami, 1966; also Jeremy D. Finn, "Univariate and Multivariate Analysis of Variance and Covariance," Statistical Laboratory, Department of Education, University of Chicago, 1966.

2. W.J. Dixon, ed., "BMD -- Biomedical Computer Programs," Berkeley: University of California Press, 1967.

3. Examples are Jeffrey W. Bean, et al, "The Beast," Washington, D.C., The Brookings Institution, 1968; Norman Nie, "Statistical Package for the Social Sciences," NORC, University of Chicago, 1968.

4. A$S. Couch, "The Data-Text System," Departmental of Social Relations, Harvard University, 1967.

5. Henry Scheffe, The Analysis of Variance, New York: John Wiley, 1959, p. 362-363.

6. A description can be found in B.J. Winer, Statistical Principles in Experimental Design, New York: McGraw-Hill, 1962, 195-199.

7. See T.W. Anderson, An Introduction to Multivariate Statistical Analysis, New York: John Wiley, 1958, Chapter 8. The criterion is also known as "Wilks lambda."

8. George W. Snedecor and William G. Cochran, Statistical Methods, sixth ed., Ames, Iowa: Iowa State University Press, 1967, p. 317-321.