INSTITUTE FOR DEVELOPMENT STUDIES
UNIVERSITY OF NAIROBI

Discussion Paper No. 101

ONE WAY ANALYSIS OF VARIANCE
(with sub-group averages):
DESCRIPTION OF A FORTRAN COMPUTER PROGRAMME

b y

D.S. Shepard

December, 1 9 7 0

ONE WAY ANALYSIS OF VARIANCE
(with Sub-Group Averages)[1]

This program provides a way of using knowledge of an explanatory variable to make predictions about a dependent variable. It also tests the existence of and measures the strength of the relationship between the variables.

In Appendix II, for example, we are given information of a sample of about a thousand Form IV male school leavers. We wish to investigate the relation between the explanatory variables (characteristics of the boys' families) and the dependent variable, their aggregate School Certificate score.

Theory of Analysis of Variance.

The essence of analysis of variance is to see how much of the variance in the dependent variable can be explained by each of the classificatory variables. This program is a one-way analysis because the cases are classified by only one explanatory variable at a time. (To test for interaction effects between two explanatory variables, a more elaborate procedure, a two-way analysis of variance, would be required.)

---

1 The need for this program, as well as the style of this description, were suggested by James Morgan, who recently left IDS. It is expected that similar programs for other types of analysis will be written or adapted at the IDS in the future.

Another program issued by the ICL computer company is also available for use. It is part of the 'XDS2' package, which also does regressions. The limitation of 'XDS2' Analysis of variance is that it does not give sub-group means, frequencies, or variances, has rigid input requirements, and does not calculate as many summary statistics. On the other hand, the XDS2 program is useful for a two or three-way analysis of variance not handled by the program written at IDS. See ICL Statistical Analysis Mark 2, 1900 Series, International Computers Limited, 1966, pp. 65-73.

Each explanatory variable, such as fathers' occupation, can classify the population into groups. The program calculates the mean of the dependent variable for each group, the number of cases in the group, and the variance within each group about its mean. The technique of analysis of variance can also measure the significance of an explanatory variable, and its power (the fraction of variance explained). The technique has been widely applied to data from experiments in agriculture and psychology. Many texts on statistics discuss the technique.[2] Analysis of variance differs from correlation and regression in that the coded values of the explanatory variables do not necessarily have any ordinal significance, and are a small number of discrete values.

Data Preparation

To use the program, a list of all the variables of interest (both dependent and explanatory) should be made in the order they appear on the data cards. The dependent variable must be a numerical variable for which an average is meaningful - i.e. an income, family size, test score, size or weight, agricultural yield, profit, a period of time; it can also be a fraction or percentage.[3]

---

2 Morgan, J. "Survey Methods - Addendum to Outline sheet 6" mimeographed. (Available from IDS computer programmers. Read this first)

Snedecor, G.W., Statistical Methods, Iowa State College Press, Ames, Iowa U.S.A. (1956). pp 237-290 (Excellent and thorough).

Finney, D.J., An Introduction to the Theory of Experimental Design, University of Chicago Press, Chicago (1960). pp. 15-21

Suits, D.B. Statistics, Rand McNally and Co., Chicago (1963) pp. 124-142. (Easily understood).

Hoel, P.G., Introduction to Mathematical Statistics, John Wiley and Sons, London (1962) pp. 299-313

Several of these may be obtained from D. Shepard or Miss B. Crockett, IDS.

3 However, an average of percentages may not be meaningful if the group sizes from which the percentages are calculated differ widely.

Number the variables sequentially, beginning with "1". (The variables are referred to by these sequence numbers.) Indicate on your list the card columns occupied by each variable and its maximum "regular" value.[4] Your data may be in any form provided that it is entirely numerical. (For this program, columns containing letters must be skipped.) On your list of variables note which one is the dependent variable. All the other variables are taken as explanatory. (If you have additional variables which will be dependent variables for a later set of analyses, they are probably not meaningful as explanatory variables. Hence you should not include them in your list of variables for the first set of analyses.) Studying the explanatory variables note their highest regular value.

Program Control Cards.

Since each run of the program requires only three control cards, it is expremely easy to use. Instructions are given on the annotated coding sheet (Appendix I) with comments in the following paragraphs.

(1) Storage limitation. For each set of analyses in this program, the following limitations must be observed:

(a) The total number of variables must be less than or equal to 201.

i.e. element 1(b) $\leq$ 201

(b) The number of cells used must be less than or oqual to 1000.

i.e. (element 1(d) + 2) x (element 1(b) - 1) $\leq$ 1000.[5]

4 By "regular" value is meant the highest coded value excluding special codes for "don't know" or "not applicable", etc., e.g. A variable describing a person's training is coded "0" to "12" according to the level of course completed, and "99" for respondents whose training is not known. For this variable the maximum "regular" value is 12. It is good practice to code "don't know" or "not applicable" as "99" for one-column variables, as "99" for two column variables, etc.

5 The first factor in the inequality is the number of categories per classification, i.e. one for each value between 0 and the maximum specified, and one for values above the maximum. The second factor is the number of classifications, which is the number of explanatory variables.

If either of these storage limitations is exceeded, the program will print an error message and stop. Therefore if your set of analyses is too big, before running it you must shorten it by:

(a) Deleting some of the explanatory variables, or putting them in a separate run. (This is particularly easy if the input unit is tape, described below.).

(b) Compressing categories on the explanatory variables with the largest number of categories, or removing these variables and placing them in a separate run.

(2) Filtering. There may be some data cases which would distort your analysis and you wish to delete entirely. The program allows for a simple filtering based on the value of the dependent variable. For example, suppose a dependent variable were coded such that "0" meant "not applicable" and "999" meant "not available". (A blank is interpreted as a zero.) To exclude these extraneous cases from the calculation of means and variance, the minimum valid value, element 2(a) would be 0.01, and the maximum valid value, element 2(b) would be 998.99.

(3) Format Statement. For the computer to know the columns for each variable, you must write a "FORMAT" statement of the type used in FORTRAN programs. Begin with an opening parenthesis. Use the letter "X" for "skip columns", the letter "F" for "read a variable", a slash "/" for "go to the next card for more data on the same record", and a comma "," for separating fields. Precede an "X" by the number of columns to be skipped; follow an "F" by the number of columns for the variable, then by ".0". You may precede the "F" with a number to indicate the number of consecutive variables of the column width specified. End the Format statement with a closing parenthesis. For example, the Format Statement

(6X,F1.0/10X,10F1.0,20X,F5.0)

means: Skip six columns, read a variable of one column, and read a variable of two columns; go to the next card, skip ten columns, read ten variables each of one column, skip twenty columns, read a variable of five columns.

(4) Data cards and Multiple analyses. Insert data cards after control card 3 if the input unit is cards. You may analyse a new set of data, or do further analysis on the same data, in one submission. Simply repeat cards 1 to 4 for each analysis as many times as required. If the input is cards, you must always provide a set of data cards following control card 3. (Data cards used for a previous analysis have not been stored by the computer.) If the data is on magnetic tape, then the tape is automatically rewound following each set of analyses, and may be used again for the next set. If the data is on paper tape, a paper tape must be set by the operator each time before control card 3 is encountered. If the input unit is not cards, control card 3 is followed by the next card 1, or by card 5.

Multiple runs are especially advantageous is a single run would exceed the storage limitations, or if one desires to analyse the same data with different variables considered as the dependent variable. (Each analysis allows only one dependent variable.)

If you have a lot of data, or want multiple runs, it is best to have your data transferred to tape. This is easily accomplished with a Library Program available from Mrs. Eveline Caldwell at the Computing Centre. It is "SFPA" - Tape Creation Program.

(5) Final Card: At the end of the last run place a blank card. This signifies that no further analyses follow, and causes the computer to stop.

Program Output

(1) Sub-group Averages. The first piece of output is sub-group averages. Taking the explanatory variables in order, the program gives for each sub-group (those observations with the same value of the explanatory variable) the "mean" (mean of the dependent variable), the "variance" (the sub-group sample variance of the dependent variable)[6]

---

6 Sample variance is defined by:

$$\mathrm{Var}(X) = \left[\left(\sum_{i=1}^{n} x_i^2 \right) - n\bar{x}^2 \right] / (n-1)$$

and the "frequency" (the number of cases in the sub-group). If the words "AND ABOVE" are printed below the last value of the explanatory variable, then the last sub-group refers to cases for which the explanatory variable is of the stated value or higher. This category typically contains the "don't know" or "not applicable" responses. The line labelled "population" gives population statistics based on all cases. Values of the explanatory variable containing no cases are not listed.

(2) Analysis of variance summary. It can be proved that the sum of the squares of the values of the dependent variable from its grand mean can be partitioned into the sum of squares "between" - the weighted sum of squared differences of the sub-group means from the grand mean, and the sum of squares "within" - the sum of squared differences of the values of the dependent variable from their sub-group mean. The sum of squares "between" is also termed "explained" because it is explained by the group means; the balance, or sum of squares "within", is termed "unexplained". Variance within may also be obtained as a weighted average of the sub-group sample variances. The stronger the relation between the dependent and explanatory variables, the more distinct will be the sub-group means and the smaller the variance within the sub-groups.

(3) F - Test. To see whether a given explanatory variable is "significant" (i.e. whether the sub-group means are significantly different) one uses the F-test.[7]

7 Dividing a sum of squares by its degrees of freedom ("D of F") yields the mean square. Under the null hypothesis that the sub-groups are not different, then both the "MEAN SQ BETWEEN" and the "MEAN SQ WITHIN" are unbiased estimates of the population variance. The value calculated by the computer is:

$$F = \frac{\text{mean square between}}{\text{mean square within}}$$

If there is little relation between the explanatory and dependent variables, then F is near unity.

The higher the value of the F ratio, the greater the probability that the explanatory variable is significant. Refer to a table of the F-Distribution[8] with the numbers of degrees of freedom printed in the output in the output. The first number, (the numerator) is generally at the top of the table and the second (the denominator) is generally down the side of the table. If the F value printed by the computor exceeds the value in the table for 5%, then the explanatory variable is significant at 5%. Similarly, if the value for 1% is exceeded, the variable is significant at 1%. F-values less than unity are never significant. Some typical values of F (from a table) are:
$F(4/100) = 2.46$ (at 5%);
$= 3.51$ (at 1%).

(4) Intraclass Correlation. If a certain variable is significant (or at least if its the F ratio exceeds unity) then one would like to know the explanatory power of the variable. This is measured by the intra-class correlation coefficient "RI" (ρ_I).

We assume a population model in which the variance among individuals (σ_T^2) is due to variance between groups (σ_B^2) plus variance within groups (σ_W^2).[9] Unbiased estimates of these variances are printed in the analysis of variance table under "POP VAR", and of the corresponding standard deviations under "POP SD". The intraclass

8 Tables of F may be found in:
Hoel, op cit., pp. 404 - 407
Suits, op cit., pp. 254 -257
Snedecor, op cit., pp. 246 - 249

9 See Morgan, op cit., and Snedecor, op.cit., pp. 282-285. The mean square within in an unbiased estimate of σ_W^2. But the mean square "between" contains a component of variance because of inaccuracy in the estimation of sample means. Therefore an unbiased estimate of variance "between" is:

$$\sigma_B^2 = \frac{\text{mean SQ Between} - \text{mean SQ within}}{N_0}$$

The denominator is the number of cases in each sub-group. If the sub-group sizes differ then N_0 is a kind of average group size, Ganguli's N, printed by the computer.

correlation is defined by:

$$RI = \frac{\text{variance between}}{\text{variance (total)}} = \frac{\sigma_B^2}{\sigma_T^2}$$

The value of RI is zero when F is less than or equal to one. RI can never exceed unity, and typically is below 0.50. The similarity of Snedecor's term "intraclass correlation" to the usual "correlation coefficient" is slightly misleading. Although the "intraclass correlation" itself is a ratio of variances, it is the square of the usual "correlation coefficient" which is the ratio of variances of expected and actual values.

Although often not presented, the intraclass correlation is very important. In a large sample, even very weak effects (relationships) become significant. Thus it is not sufficient merely to test for the existence of a relationship, but also to try to measure its strength.

(5) R - Squared. Two other summary statistics of traditional interest but of less precise statistical meaning are also supplied. The R - squared, analogous to the R - squared obtained by multiple regression, is defined by:

$$R - \text{Squared} = \frac{\text{Sum of squares between}}{\text{Sum of squared (total)}}.$$

This ratio is between 0 and 1, inclusive. If the explanatory variable has an ordinal meaning so that its correlation with the dependent variable is meaningful, then "R - squared" is the same as r^2, where r is the simple correlation coefficient between the explanatory and dependent variables.

A variant is the "Adjusted R-squared" in which downward adjustment for varying degrees of freedom has been made, defined by:

$$\text{Adjusted R-squared} = 1 - \frac{\text{mean square within}}{\text{mean square (total)}}$$

If the F-ratio is less than unity, then the Adjusted R-Squared is negative. Otherwise the Adjusted R-squared is positive, but less than or equal to one. In general, the higher the R-squared or Adjusted R-squared, the more information the explanatory variable furnishes about the dependent variable.

Cost of Runs

This program operates successfully on the University of Nairobi's computer, ICL model 1902. At present the charge on the University's computer is Shs. 140/= per hour (or 2/35 per minute). Although experience with this program is still too limited for a careful evaluation of running times, a rough estimate is as follows:

	Step	time (minutes)	Cost (Shs.)
1.	Compilation of FORTRAN program (overhead prior to each run of program)	1	2/50
2.	Processing of data cards or tape - per 1000 (pro-rated)	8	18/50
3.	Printing results of one set of analysis with about 20 explanatory variables	3	5/00

For example, consider the exercise in Appendix II. The first set of analysis had 993 cases, requiring 1 + 8 + 3 = 12 minutes. The second set had 260 cases, requiring 1 + 2 - 3 = 6 minutes. Thus the total computer time is 18 minutes, or Shs. 42/=. To do this exercise with a counter-sorter and desk calculator would undoubtedly require several days, and the chances of error would be increased.

Appendix I. Annotated coding sheet.

80 COL. DATA SHEET JOB: ______ DATE: ______

1 a b c d e f

(a) Number of cases. (Note: All values in line 1 must be right-justified without a decimal point.)
(b) Number of variables (explanatory and dependent; maximum allowed is 201).
(c) Sequence number of the dependent variable.
(d) Maximum "regular" value of any explanatory variable.
(e) Input unit: 0 = cards, 1 = magnetic tape, 3 = paper tape.
(f) Print data: 0 = no, 1 = yes.

2 a b c d e

(a) Minimum valid value of the dependent variable (may be entered with decimal point).
(b) Maximum valid value of the dependent variable (may be entered with decimal point).
(c) Program name: "SO4A".
(d) Identification of this set of analyses (e.g. user's initials, number, or an abbreviation).
(e) Name of magnetic data tape (blank if input unit not magnetic tape).

3 (a

(a) Format Statement for data. (Number of "F" fields must agree with item 1(b). Format statement must begin with and end with parentheses.)

4 INSERT DATA CARDS HERE if item 1(e) specifies cards.

(Number of cases must agree with item 1(a). Format statement must agree with item 3(a).)

REPEAT NUMBERS 1 TO 4 for each additional set of analyses, as many times as required.

5 ∇∇∇∇∇

Blank card signifies end of job. ("∇" indicates blank.)

PUNCHED VERIFIED

APPENDIX II

ILLUSTRATION

To illustrate the use of the one-way analysis of variance program, it has been applied to some of the data collected under the School Leavers' Tracer Project, conducted by Mr. Tony Somerset and Mr. John Anderson of the Institute. On the following pages are listed some of the variables for which data was collected, and the computer analyses for four explanatory variables (nos. 9, 4, 12, and 7).

The purpose of the analysis of variance exercise was to see which factors in a student's background affect his performance on School Certificate, and how much. The data for boys and for girls were treated separately; only result from the analysis of boys' data have been reproduced here.

The results, like those of any statistical program, must be interpreted with understanding. Because this program looks at the effect of only one explanatory variable on the dependent variable at a time, there is the possibility of "spurious correlation" i.e. the explanatory and dependent variables may both be related to some third variable.

Tracer Project
(Form IV School Leavers)

Var No.	Cols.	Variable Name	Maximum Value	Format skip var.
1	8	Age	9	7X, F1.0,
2	10	Tribe	8	1X, F1.0,
3	27-28	School Certificate result	Dep. Var.*	16X, F2.0,
4	29	CSC Category	5	F1.0
5	46	Where pupil lives	8	16X
6	47	Religion	4	
7	48	Father alive?	6	
8	49	Mother alive?	6	
9	50	Father's present occupation	8	
10	51	Father's previous occupation	8	
11	52	Father's business	7	
12	53	Father's positions of authority	8	18F1.0
13	54	Father's participation in local activities	8	
14	55	Father's land	8	
15	56	Farm labour	2	
16	57	Father's cattle	4	
17	58	Cash crops	8	
18	59	Father's education	4	
19	60	Father's English	2	
20	61	Mother's education	4	
21	62	Mother's English	2	
22	63	Mother's occupation	8	

Maximum Regular value: 9 (for Explanatory variables)

* Range of valid values for dependent variables: 6 to 54 inclusive

Number of Variables: 22

Dependent variable's number 3

Format Statement: (7X,F1.0,1X,F1.0,16X,F2.0,F1.0,16X. 18F1.0)

Number of cases: For boys (first set of analyses) 993
For girls (second set of analyses) 260.

Coding sheet with control cards for example of Tracer Project.

First set of analyses.	993 22 3 9 0 0
	6 54 SHEI BOY
	(7X, F1.0, 1X, F1.0, 16X, F2.0, F1.0, 16X, 18F1.0)
Insert 993 Data cards	
Second set of analyses.	260 22 3 9 0 0
	6 54 SHEI GRL
	(7X, F1.0, 1X, F1.0, 16X, F2.0, F1.0, 16X, 18F1.0)
260 Data cards	
Blank Card.	▽▽▽▽▽

Selected portions of computer output.

(Explained in text which follows, according to typewritten line and item numbers.)

I. Input Parameters.

```
1) CASES  VARS  DEP  MAX  TAPE  LIST
   a 993  b 22  c 3.  d 9  e 0  f 0
2)   a 6.000   (b 54.000  c SHEI  d BOY
3) (7X,F1.0,1X,F1.0,16X,F2.0,F1.0,16X,18F1.0)
```

II. Sub Group Averages.

(4) VARIABLE NO. [e]9	MEAN	VARIANCE	FREQUENCY	Father's present occupation
[a]0	[b]39.2326	[c]130.3732	[d]43	Retired, died, no occupation
1	30.9167	146.5563	72	Professional, managerial
2	34.3725	100.3984	51	Teacher
3	29.1034	120.5246	29	Clerical occupation
4	34.6154	98.4231	13	Armed forces, police
5	32.5469	81.9660	64	Skilled, semi-skilled manual
6	36.9429	133.6890	105	Unskilled manual
7	35.3704	144.0861	81	Entrepreneurs, traders
8	36.5988	121.9183	491	Farmers
9	35.0455	120.9746	44	N.R. (No Response)
(5) POPULATION	35.5478	126.7580	993	All occupations

III. Analysis of Variance Summary.

[a]SOURCE	[b]SUM OF SQ	[c]D OF F	[d]MEAN SQ	[e]POP VAR	[f]POP SD
BETWEEN	4750.84	9.	527.8713	5.0901	2.2561
WITHIN	120993.14	983.	123.0856	123.0856	11.0944
TOTAL	125743.98	992.	126.7580	128.1757	11.3215

```
(8) R-SQUARED= 0.0378     ADJUSTED R-SQUARED= 0.0290     F( a 9. / b 983.) = c 4.289
(9) SANGULI'S N = a 79.52            RI (INTRACLASS CORRELATION) = b 0.0397
```

Explanation of Program Output

(I) Input Parameters. The first three lines are the contents of the three input control cards forthe first set of analyses. The meanings of parameters on each card follow.

Line number and item:

1. (a) Number of cases (observations) in data for for this set of analyses.

   (b) Number of variables.

   (c) Number of the dependent variable. (i.e. School Certificate Result)

   (d) Maximum "regular" value for explanatory variables.

   (e) Input unit. ("0" indicates card input.)

   (f) Print data? ("0" indicates "no".)

2. (a) Minimum valid value of the dependent variable.

   (b) Maximum valid value of the dependent variable. (i.e. The range of valid School Certificate Results is 6 to 54 points, inclusive.)

   (c) Former program name. Now "SO4A".

   (d) User's identification. (i.e. "BOY" indicates that this set of analyses refers to boys only.)

   (e) Name of data tape. (Blank because no tape used).

3. Format Statement for data. (To check: note that the number of "F" fields agrees with the number of variables in 1(b), i.e. 22).

If the printing of data were requested in item 1(f), it would follow between output lines 3 and 4. Printing of data for each case requires two lines or more. The first line gives the case number, and the value of the dependent variable. The second line gives the values of the first thirty variables. If required, the third line gives the value of the next thirty variables, etc. The values are given in the order the variables appear on the data card. The value of the dependent variable is given in its normal sequence.

Since this set has 993 cases and each case requires two lines a complete listing of the data would require about 40 pages. This would be costly in terms of computer time.

(II) Sub-Group Averages. Lines 4 to 9 are an example of the explanatory variable is a tabulation of sub-group averages and analysis of variance summary provided for each explanatory variable within each set of analyses. In this discussion one explanatory variable (no. 9) is discussed in detail, and three others briefly.

Line number and item:

4. (a) Number of explanatory variable. i.e. 9 = Father's present occupation.

5. (a) Coded values of explanatory variables. (Meanings taken from researcher's coding instructions are typewritten alongside.)

(b) Mean of dependent variable. i.e. Var 3, School Certificate Result.

Note: School Certificate results are scored such that the lowest result is the best score. Thus boys whose fathers had died or retired (coded "0") scored nine points worse than boys whose fathers were professionals (coded "1").

(c) Variance of dependent variable in each sub-group.

Note: The standard deviation, or square root of the variance, is in the same units as the means.

(d) Frequency of each value of explanatory variable.

6. "POPULATION" - estimates for population based on entire sample.

This line is the same for each explanatory variable.

General Comments: Analysis of variance should be an appropriate technique for examining the relationship between the father's present occupation and the son's School Certificate result. It would be difficult to rank occupations a priori, and almost impossible to place them along a numerical scale. In general the pattern above is as expected; the higher the income and the status of the father's occupation, the better the son's score.

(III) Analysis of Variance Summary.

Meanings of abbreviations and calculation of statistics by line number and item.

7. (a) "SOURCE" - Gives the three components of sums of squares.

(b) Sum of Squares (of values of dependent variable about its grand mean.)

Let N = number of cases, i.e. 993

Let C = number of sub-groups, i.e. 10

Let y_j = value of dependent variable for case j. i.e. (in data)

Let $\bar{y}$ = grand mean of dependent variable, i.e. 35.55

Let $\bar{y}_i$ = mean of dependent variable in sub-group i, i.e. 35.41, etc

Let N_i = number of cases in sub-group i (from FREQUENCY) i.e. 807, etc

Calculation of SUM OF SQ:

$$\text{BETWEEN} = \sum_{i=1}^{C} N_i(\bar{y}_i)^2 - N(\bar{y})^2$$

i.e. $130{,}230.37 - 993(35.55)^2 = 4{,}750.84$

WITHIN = TOTAL SUM OF SQ - BETWEEN SUM OF SQ

i.e. $125{,}743.98 - 4{,}750.84 = 120{,}993.14$

$$\text{TOTAL} = \sum_{j=1}^{N} (y_j)^2 - N(\bar{y})^2$$

i.e. $251{,}223.51 - 993(35.55)^2 = 125{,}743.98$

(c) Degree of Freedom. Calculation:

BETWEEN = C - 1.	i.e. 10 - 1	=	9
WITHIN = N - C.	i.e. 993 - 10	=	983
TOTAL = N - 1.	i.e. 993 - 1	=	992

(d) Mean Square. Calculation:

$$\text{MEAN SQ} = \frac{\text{SUM OF SQ}}{\text{D OF F}} \text{ i.e.}$$

$$\text{i.e. "BETWEEN"} = \frac{4750.84}{9} = 527.87$$

(e) Population Variance estimate. Calculation:

$$\text{BETWEEN} = \frac{(\text{MEAN SQ BETWEEN} - \text{MEAN SQ WITHIN})}{\text{GANGULI'S N}}$$

$$\text{i.e. } \frac{527.87 - 123.09}{79.52} = 5.09$$

WITHIN = MEAN SQ WITHIN i.e. 123.09 = 123.09

TOTAL = POP VAR WITHIN + POP VAR BETWEEN

i.e. 5.09 + 123.09 = 128.18

(f) Population Standard Deviation. Calculation:

$$\text{POP SD} = \sqrt{\text{POP VAR.}} \text{ i.e. "BETWEEN"} = \sqrt{5.09} = 2.25$$

8. F-test of significance.

(a) D OF F BETWEEN (in greater mean square) i.e. 9

(b) D of F WITHIN (in lesser mean square) i.e. 983

(c) Calculated F-Ratio. Calculation:

$$F = \frac{\text{MEAN SQ BETWEEN}}{\text{MEAN SQ WITHIN}} \text{ i.e. } \frac{527.87}{123.09} = 4.289$$

9. (a) Ganguli's N, N_o. Calculation (see definitions above):

$$N_o = \frac{1}{(C - 1)} \left(N - \frac{\sum_{i=1}^{C} N_i^2}{N}\right)$$

$$\text{i.e. } \frac{1}{9} \left(993 - \frac{275,343}{993}\right) = 79.52$$

(b) Intraclass correlation. Measures the fraction of variance explained.

Calculation:

$$RI = \frac{\text{POP VAR BETWEEN}}{\text{POP VAR TOTAL}}$$

$$\text{i.e. } \frac{5.09}{128.17} = .0397$$

<u>Further program output.</u> (Brief explanation on following page.)

VARIABLE NO. 4	MEAN	VARIANCE	FREQUENCY	CSO Category
1	18.2000	12.6935	155	Grade I Pass
2	28.0640	11.9075	250	Grade II Pass
3	37.9008	10.1561	242	Grade III Pass
4	45.0200	25.4092	250	G.C.E.
5	52.4479	2.2288	96	Fail
POPULATION	35.5478	126.7580	993	

SOURCE	SUM OF SQ	D OF F	MEAN SQ	POP VAR	POP SD
BETWEEN	111837.94	4.	27959.4856	144.2964	12.0123
WITHIN	13906.04	988.	14.0749	14.0749	3.7517
TOTAL	125743.98	992.	126.7580	158.3714	12.5846

R-SQUARED= 0.8894 ADJUSTED R-SQUARED= 0.8890 F(4. / 988.) =1986.474
GANGULI'S N = 193.67 RI (INTRACLASS CORRELATION) = 0.9111

VARIABLE NO. 12	MEAN	VARIANCE	FREQUENCY	Father's positions of authority
0	35.5204	126.8704	636	None
1	38.0000	211.2941	18	Political
2	37.0252	114.3158	159	Self-help, educational, Coop
3	32.9423	132.1731	52	Religious
4	37.9000	133.7789	20	Political & Self-help/Coop
5	37.0000	0.0000	1	Political & Religious
6	34.7234	146.7697	47	Self-help & Religious
7	33.8333	124.9667	6	Political & Self-help & Rel.
9	33.7222	105.7610	54	N.R.
POPULATION	35.5478	126.7580	993	

SOURCE	SUM OF SQ	D OF F	MEAN SQ	POP VAR	POP SD
BETWEEN	1263.15	8.	157.8933	0.4553	0.6747
WITHIN	124480.83	984.	126.5040	126.5040	11.2474
TOTAL	125743.98	992.	126.7580	126.9602	11.2677

R-SQUARED= 0.0100 ADJUSTED R-SQUARED= 0.0020 F(8. / 984.) = 1.244
GANGULI'S N = 68.94 RI (INTRACLASS CORRELATION) = 0.0036

VARIABLE NO. 7	MEAN	VARIANCE	FREQUENCY	Father alive?
1	35.4139	131.0493	807	Yes: Father alive
2	35.9189	130.8544	37	F. died when R was 0-4 years old
3	34.9189	69.7988	37	F. died when R was 5-9 years old
4	36.6889	101.0374	45	F. died when R was 10-14 years
5	38.9487	107.2605	39	F. died when R was 15- years old
6	33.2857	120.1143	21	F. died, but R's age N. S.
9	32.8571	211.4762	7	N. R.
POPULATION	35.5478	126.7580	993	

SOURCE	SUM OF SQ	D OF F	MEAN SQ	POP VAR	POP SD
BETWEEN	702.02	6.	117.0025	0.0000	0.0000
WITHIN	125041.96	986.	126.8174	126.8174	11.2613
TOTAL	125743.98	992.	126.7580	126.8174	11.2613

R-SQUARED= 0.0056 ADJUSTED R-SQUARED=-0.0005 F(6. / 986.) = 0.923
GANGULI'S N = 55.06 RI (INTRACLASS CORRELATION) = 0.0000

(IV) Explanatory Variable No. 4 - School Certificate Category. Since the CSC Category is based upon the student's aggregate score, the very strong relationship between the two variables is to be expected. For truly different variables, values of F and RI as high as those obtained here are virtually unheard of.

(V) Explanatory Variable No. 12 - Father's Positions of Authority. This variable indicates the importance of printing the sub-group means. Although the F value is not significant even at the 10% level, the means nevertheless indicate a significant pattern. Boys whose fathers hold positions in religious organizations (coded 3) or in both religious and other types of organizations (coded 5,6 and 7) score relatively well.

(VI) Explanatory Variable No. 7 - Father Alive? This variable is an example of one with virtually no relationship to the dependent variable. Not only is the F-value below one (thus not significant at any level), and also the sub-group means are close together and follow no consistent pattern according to the respondent's age when his father died.