WNNR
CSIR

SPECIAL REPORT

PERS. 89

R.S. HALL

STAT
A SYSTEM OF COMPUTER PROGRAMS FOR CARRYING OUT ELEMENTARY STATISTICAL ANALYSIS IN THE CONVERSATIONAL MODE

NATIONAL INSTITUTE FOR PERSONNEL RESEARCH
COUNCIL FOR SCIENTIFIC AND INDUSTRIAL RESEARCH

CSIR Special Report No. 89

UDC 681.3:519.2
Johannesburg, South Africa. August 1966.

ness to
Dr. D. nd
Comput e for
develo e been
possib

for the
advice the
Computi lso my
thanks ns they
have ha

Miss C.M. Elder of the N.I.P.R. who spent many hours at night acting as a "guinea-pig" trying out different programs. Many ●f her suggestions, particularly on the forms of the questions asked have been incorporated in the system.

I would also like to thank Mr. A. Molatedi and Mr. M. Malele for the many times they have repunched and corrected my programs, and Miss P.F. Kabane for typing this report. Thëir assistance greatly lightened the burden of the work.

Introduction

This is probably one of the shortest reports ever emanating from the N.I.P.R., but its size is in inverse ratio to the importance of the developments it forshadows.

Up to the present the machines to which the Institute has had access have been instruments for computing only. Although they have permitted methods of analysis which were previously impracticable, and they have handled a vast amount of work, the overheads in labour and cost in getting work on them, have encouraged the shot-gun approach.

The facilities that will be available to the Institute when the new computer is installed at the University of the Witwatersrand next year are of a very different order. The terminals by which it will be linked to the new machine will permit a worker to probe his data to a degree that can hardly be imagined. With little more effort than is required for tapping a key a psychologist will be able to experiment with whole families of factor structures, a sociologist will be able to follow clues and trace associations through large numbers of responses, and the theoretically-minded will be able to study at first-hand the implications of their models. For all, the stimulus provided by instant responses to questions, will stimulate a degree of concentration not often experienced, because so frequently the process is interrupted by laborious calculations or lengthy delays.

However/......

STAT

A System of programs has been developed for carrying out elementary statistical analysis on the University of the Witwatersrand's 1620 Model II Computer.

These programs have been designed with the following objectives. Firstly to provide experience in programming for conversational mode. Secondly, they have been designed for the user who has little or no knowledge of computers or programs and has difficulty in understanding program write-ups. Once the system has been called into operation by means of the three standard control cards, any further information necessary for processing is supplied by question and answer on the typewriter with, in each case, explicit instructions to the user on the kind of response he has to make. The user's responses are, however, kept to the minimum. For example he does not type in the number of cases to be selected, but merely ends his input of specifications with a standard signal. The program computes the parameters it requires for itself.

If the user knows nothing at all about the system he can by typing "INFO", obtain a list of the call-words, and brief descriptions of the system programs. Thereafter by typing the appropriate call-word when requested, he initiates the computation he requires.

It/.........

It has also been assumed that the user is a poor typist so that not only are all his responses screened to make sure they are of the right form and within the limits prescribed by the program, but after each set of responses he is given an opportunity for correcting any that may have been typed incorrectly.

From the statistical point of view the third aim has been to give the user the greatest possible scope for probing his data. For this purpose, facilities have been provided by means of question and answer, for selecting or rejecting variables, cases, or categories of cases. This allows the user to carry out the statistical computations provided for any conceivable subset of his data. Furthermore several forms of the computing programs have been included which produce amounts of information in inverse proportion to the speed at which they carry out the computations. For example one program will provide a rapid scan of the data by calculating only chi-squares for all pairs of a specified set of variables, while another will, for a single pair of variables, produce the tabulation, calculate any desired table of percentages, compute chi-square with, in addition, a transformation to allow for small expected values, and also provide detailed information on all cases rejected from the table. The two programs operate at very different speeds as the first remains entirely within core while the second is subdivided into 4 sections which overlay each other and have to be called off the disk.

An additional facility for the handling of data is provided by a program which will accept a subroutine written by the user for recoding or transforming data. To make use of this particular facility some knowledge of programming is necessary.

Considerable flexibility with regard to the input of data is provided by the 3 input programs. The first will handle data punched in 6 column fields, 12 to a card. The second allows the user to enter his data from the typewriter, while the third will accept a user supplied subroutine for reading data in any layout. (There is no provision for reading data in free format). Another program will accept a user supplied subroutine for punching data out on cards in any desired format. The last two programs provide an interface with other program systems such as the N.I.P.R. System. Other programs allow the user to list his data or any desired subset of it, or the particular subset specified for a computation so that he may inspect it in detail.

The computations available are the following. Data must be in the form of cases and variables within cases. The system will handle up to 200 variables per case, or a total data volume of cases x variables = 38,000.

1. Means and standard deviations with 95% Confidence limits - for variables.

2. Means and standard deviations with 95% Confidence limits - for cases.

3. Intercorrelation matrix with significance levels - between variables. (maximum size 45x45).

4. Correlation coefficient and both means and s.d.s with 95% confidence limits for any given pair variables.

5. Intercorrelation matrix with significance levels - between cases. (maximum size 40x40).

6. Correlation and significance levels between all pairs of a specified set of variables (not in matrix form, but up to 200 variables may be handled).

7. One way analysis of variance and Bartletts test with significance levels and group means and s.d.s. with 95% confidence limits.

8. Chi-squares and corresponding significance levels for all pairs of variables of a specified set (up to 200 variables).

9. mxn contingency table, chi-square, a table of the contributions to chi-square from each cell, also a transformation allowing for very small expected values, tables of percentages - rowwise, columnwise, or based on the grand total, - and tables of cases rejected from the tabulation, for any given pair of variables.

10./ ....

10. F and t tests between subsets of cases, which may be defined by specified categories on several variables.

11. Fitting orthogonal polynomials.

12. Transformation of data using user supplied subroutine.

13. Input form A - 6 column fields, 12 per card.

14. B - with user supplied subroutine for reading cards.

15. C - from typewriter.

16. Listing of data.

17. Punching of data onto cards using user supplied subroutine.

18. Selecting or rejecting cases, or specified categories of cases.

19. Editing of data.

Appendix I.

A typical sequence of exchanges between user and computer.

These sequences can vary greatly depending on the options selected by the user or by the mistakes he makes.

Appendix II

The Control Cards required are:-

1) Monitor II Call card.

2) Job card ‡ ‡ JOB

3) Execute card ‡ ‡ XEQSSTAT1

4) 2 blank cards

Appendix III

Components of STAT

(as at 1/8/66)

STAT NO.		Callword	Status
0.	System Controller (switching program)	-	X=working
1.	Initialiser	-	X
2.	WILH - subroutine normalising F distribution	-	X
3.	JANEE - " handling yes or no responses	-	X
4.	REJES - " handling selection or rejection cases	-	X
5.	CONAV - " for 95% confidence intervals for means	-	X
6.	CONSD - " for 95% confidence intervals for s.d.s.	-	X
7.	2nd part, 1-way analysis of variance	-	X
8.	GETMAN - subroutine. (dummy for STAT18)	-	dummy
9.	2nd part, intercorrelation matrix for variables	-	X
10.	SIGLV - subroutine for significance level of normal d.f.	-	X
11.	2nd part of tabulation	-	X
12.	Loading from cards in standard format	LOD1	X
13.	QSEL - subroutine for handling input of variable specifications.	-	X
14.	RETR - subroutine for reading data from disk into core	-	X
15.	Loading data from typewriter	TYP1	X
16.	Program for handling user's transformation subroutine	REC1	X
17.	Formation of data subsets	REJ1	X

18./ ....

STAT NO.		Callword	Status
18.	Loading data from cards with user's subroutine	LOD2	X
19.	Punching data on to cards ditto	PUN1	X
20.	Means and std. deviations for variables	MSD1	X
21.	1st part - 1-way analysis of variance	ANV1	X
22.	PUNMAN - subroutine. (dummy for STAT19)	-	dummy
23.			
24.	1st part - intercorrelation matrix for variables	COR1	X
25.	1st part - tabulation	CHI1	X
26.	CONR - subroutine for 95% confidence intervals for correlations.	-	X
27.	RECODE - subroutine (dummy for STAT16)	-	X
28.	Means and std. deviations for cases	MSD2	X
29.	REJSEL - subroutine for handling input of specifications.	-	X
30.	Pairwise correlations	COR2	X
31.	PUTMAN - subroutine	-	dummy
32.	Partial correlations	PAR1	X
33.	F and t tests	TAF1	
34.		-	
35.	Rapid scan of correlations	COR3	X
36.	Rapid scan •f tabulations	CHI2	X
37.	Orthogonal polynomials	REG1	X
38.	3rd part - tabulation	-	X
39.	Intercorrelation matrix for cases	COR4	X

50./ ....

STAT NO.		Callword	Status
50.	4th part - tabulation	-	X
51.	REPT - subroutine for correcting responses	-	X
52.	SELVA - subroutine for handling selection or rejection of variables.		X
53.	LOCNC - subroutine for locating parts of IND array	-	Not used
54.	CATSEL - " for handling input of category specifications.	-	X
55.	LODK - " writing data from core to disk	-	X
56.	Program for editing data	EDIT	
57.	SWITCH - subroutine version of STATO	-	X
58.		-	
59.		-	
60.	Documentation	INFO	X
61.	Listing data subsets	LIST	X
62.	QTES - subroutine for checking numerical responses	-	X
63.		-	
64.	Program for listing data set up for a computation	LOOK	X

Appendix IV

broutines for handling questions and answers and the specification of cases or variables.

size

. JANEE(IRESP) handles Yes or No responses 632
IRESP = 1 if yes, 2 = if no

. QTES(N,M,J,K) checks that numerical responses are within range. 748
N = response, M = maximum, J = 1 if N is OK,
J = 2 if not.
If K = 1, zero is treated as signalling end of input.
= 2, zero is treated as an error.

. REPT(N,J) Offers user opportunity for repeating a question. 660
N = No. of questions covered, J is the no. of the question that is t● be repeated.

. REJSEL(IND,I,MAX) handles input of case or variable numbers, storing them in IND. 2736
I = last location of IND used.
MAX = maximum value that may be used.
Uses QTES and ABSF.

. CATSEL(IND,L,M,N) handles input of catgory specifications, storing them in IND. 1886
L = no elements of IND used on entry, and on exit.
M is the max. value that may be typed in.
N is the number of variables used.
Uses QTES

size

13. QSEL(MAX) handles the input of variable specifications. 1486
MAX is the maximum number that may be typed in.
Uses REJES, REPT, JANEE.

4. REJES(V,N,I) scans cases during input from disk. 1664
V = array holding casa data.
N = 1 if case is to be accepted,
= 2 ditto rejected.
I = case number
where categories are involved, S.R. also handles
OR or AND of category specifications.

52. SELVA(JV,M) scans variables during input from disk. 758
JV = variable no.
M = 1 if variable is to be used.
= 2 if not.

There is a control area in COMMON with the following structure.
NTS = total number of cases.
NTV = total no. of variables per case.
NS = Number of cases used in the current run.
NV = Number of variables used in the current run.
NSI = Number of individual case numbers specified in the IND array.
NVI = " " " variables " " " " "
NCI = " " variable and category numbers specified in the IND array.
NCV = number of variables used in category specifications.
MI = 1 when zeros are proper scores and = 0, when not.
IND = array for holding case, category, or variable specifications in that order.

The/.........

The operation of the subroutines REJES and SELVA are controlled by NSI, NVI, NCI, and NCV which are in turn set by the subroutines CATSEL or QSEL •r by the programs REJ1, or EDIT.

If NSI, NVI or NCI are positive the specifications are used for selection, if negative for rejection. If they are zero then all cases or all variables are to be used.

NCV contains the number of variables whose categories are used for determining subsets of the data. If positive the case must satisfy requirements on all variables to qualify (logical AND), if negative satisfying requirements on one variable only is sufficient (logical OR)

In operation NSI takes precedence over NCI, and NVI is independent of both.

14./ ....

Subroutines for handling transfers of data to and from disk.

size

14. RETR(VAR,NV,ID,IR) handles reading of data from disk to core 950
VAR = array into which data is read
NV = no. of items in VAR
ID = no. of items left in buffer after input of a case.
IR = record count.
Data items are packed 19 to a record regardless of their grouping by cases.

55. LODK(VAR,NV,ID,IR,BUFR) handles reading of data from core to disk. 600
arguments as for RETR.
BUFR is a buffer array into which variables are packed before loading to disk.

5./ ....

Subroutines for indicating statistical significance.

No.			size
5.	CONAV(AV,SD,BN,UCL,BCL)	computes 95% Confidence limits for a mean	1676
6.	CONSD(SD,BN,UCL,BCL)	computes 95% Confidence limits for a s.d.	1754
26.	CONR(R,BN,UCL,BCL)	computes 95% Confidence limits for a correlation.	1460

where AV = mean, SD = standard deviation, R = correlation coefficient
BN = sample size, UCL,BCL = upper and lower 95% confidence limits.

2.	WILH(F,D1,D2,SIG)	normalises an F ratio and computes its significance using SIGLV.	820

where F = a F ratio, D1, D2 = upper and lower degrees of freedom
SIG = significance level returned to calling program.

10.	SIGLV(X,SIG)	computes the one-tailed significance level of a normal deviate.	612

where X is a normal deviate, SIG is the corresponding significance level returned to the main program.

8./ ....

Dummy Subroutines representing user supplied S.R.'s

8. GETMAN(VAR,K) contains a read and a format statement. Used with LOD2.

22. PUTMAN(VAR,K) contains a print with the corresponding format.

27. RECODE(VAR,NTV,K) Recoding S.R. used with REC1.

63. PUNMAN(VAR,K) contains a punch with the corresponding format. Used with PUN1.

VAR is an array for holding the variables belonging to one case
NTV is the number of such variables(must be amended if the number of variables is reduced).

K is an indicator which = 0 on first entry and may be set within the subroutine.

Appendix V.

Rules for Adding New Programs to STAT

1. The program must contain the following DIMENSION and COMMON Statements.

```
DIMENSION IND(200),  VAR(200)
```

If LODK is used BUFR(19) is also required.

```
COMMON NTS, NTV, NS, NV, NSI, NVI, NCI, NCV, MI, IND.
```

(See Appendix IV, page 2 for details).

2. Parameters required by the program should be requested on the typewriter. Responses of the yes/no form should be handled by calling the S.R. JANEE which will return 1 if response is yes, and 2 if it is no.

e.g.

```
      TYPE 30
   30 FORMAT(23HARE ZEROS PROPER SCORES)
      CALL JANEE (IRESP)
      GO TO(40, 50), IRESP
```

3. After the last question, the user should be given the opportunity of revising any of his responses by a call to REPT.

E.G.

```
      CALL REPT(3, IRESP)
      JUMP=2
      GO TO(100, 40, 50, 60), IRESP
  100 CONTINUE
```

where 3 is the number of questions covered by REPT, and the Computed GO TO contains their statement numbers. To avoid repeating all questions following the one being changed, Computed GO TO's which continue when JUMP=1 but revert to REPT when JUMP=2 should be inserted between questions.

4./ ....

4. The location MI in COMMON is 1 if zeros are to be treated as scores and 0, if they indicate missing information. If this affects the operation of the program, MI should be tested.

5. If the program is able to handle batches of variables, or if limitations of capacity necessitate the selection of subsets of variables, the S.R. QSEL should be called to allow the user to select the variables, (QSEL will ask the necessary questicns). The argument of QSEL is NTV, the number of variables loaded.

6. The input of data should be effected by a DO locp over cases in which RETR is called to read in data off the disk, REJES to handle the selection or rejection of subjects. Nested within this loop should be another over variables containing a call to SELVA to handle the selection or rejection of variables.

```
e.g.        DO 100 I=1, NTS
            CALL RETR(VAR, NTV, ID, IR)
            CALL REJES(VAR, L, I)
            GO TO(50, 100), L
         50 CONTINUE
            DO 90 J=1, NTV
            CALL SELVA(J, K)
            GO TO(60, 90), K
         60 CONTINUE
         90 CONTINUE
        100 CONTINUE
```

If data is to be written back to the disk, LODK(VAR, NTV, IDO, IRO, BUFR) should be used within the outer lo•p. IDO should be tested after exit from the loop and if not zero the contents of BUFR should be written to the disk.

7./ ....

7. Two counts should be provided, a Buffer Count ID=0 and a record Count IR=1, for the use of RETR. A further pair is required if LODK is used.

8. There must be a return to the controller by the statement CALL LINK(STATO), or, if there are 3800 characters available, CALL SWITCH, and it is useful to provide an option which allows the program to be repeated without a return to the controller.

9. Modifications to the Controller (STATO)

   a) The program should be given a name e.g. STAT20 and a call-word. The next available location in the array CALLW should be allocated to the new call-word and made equal to its numerical representation.
   e.g. if the call-word is MSD1,
   CALLW(3) = .54624471.

   b) NCALL should then be increased by 1.

   c) the Computed GO TO (statement number 240) should be enlarged to include a jump to a CALL LINK.
   e.g. 280 CALL LINK(STAT20).

10. The modifications described in 7. should also be made to the subroutine SWITCH.

11. A brief description of the program and its call-word should be added to the documentation program STAT60.

WNNR
CSIR

537/66

GA csir 297 s 65