DESCRIPTION OF TWO FORTRAN PROGRAMMES FOR SPEARMAN'S RANK CORRELATION COEFFICIENT

by

B. Wasow
and
M. Cowen

TECHNICAL PAPER NO. 7

INSTITUTE FOR DEVELOPMENT STUDIES
UNIVERSITY OF NAIROBI
P.O. Box 30197,
Nairobi, Kenya

APRIL 1974

I.D.S. TECHNICAL PAPER NO. 7

DESCRIPTION OF TWO FORTRAN PROGRAMMES FOR SPEARMAN'S RANK CORRELATION COEFFICIENT. by B. Wasow and M. Cowen

Addendum to page 2:

The increased size of memory now enables the handling of up to 9345 units in the variable-observation constraint equation:

$$2\theta + 6\theta + 2v^2 - 2v \leq 9345$$

For up to 10460 units (a doubling of the previous capacity) on the right hand side of the inequality, advise the computer operators to run the programme under <u>EXECUTIVE</u>.

DESCRIPTION OF TWO FORTRAN PROGRAMMES FOR SPEARMAN'S RANK CORRELATION COEFFICIENTS*

INTRODUCTION.

Correlation coefficients between two variables measure the extent to which these variables rise and falls together or independently. The coefficients can vary from +1, which indicates perfect direct variation between the two variables (a perfectly monotonic increasing relationship in the case of the rank correlation coefficient), to -1, which indicates perfect inverse variation. A rank correlation coefficient near to 0 indicates that no systematic monotonic relationship exists.

Spearman's rank correlation coefficient is the simple correlation coefficient between the rankings of variables rather than between their values. It has two basic advantages over the ordinary correlation coefficient.

a) It is non-parametric. This means that one can draw conclusions about the relationships between the variables without making any prior assumptions about their distribution. To infer conclusions from the ordinary correlation coefficient, it is necessary to assume a bivariate normal distribution for the two variables in question. This assumption is often invalid in social science research, as distribution of wealth, power, ownership etc. are often highly skewed, with a long "tail" to the right. (The rich get richer while the population gets skewed.)

b) It is not influenced unduly by extreme values. An ordinary simple correlation coefficient is heavily influenced by any observations which lie many standard deviations from the mean. Since the rank correlation is devived from ordinal data only, it is unaffected by the in interval between values of observations.

The disadvantages of the rank correlation coefficient are these:

c) If two variables do in fact have a bivariate normal distribution, then the simple correlation coefficient is a more efficient statistic than the rank correlation coefficient. The loss of efficiency is not great, however.

d) If observation values are tightly clustered, the ranking may indicate large differences where these do not seem, intuitively, to exist.

In fact, it will be seen that points (c) and (d) are just the "other side of the coin" from (a) and (b).

Aside from being interested in the rank correlation coefficient, researchers may be interested simply in the rankings of observations by various variables. This information is also made available by these programmes.

The Two I.D.S. Programmes

On file at the I.D.S. are two decks of computer cards which can generate rank correlation information from data. The programme named TAKA TAKA digests a number of variables on each of a number of observations, and generates a rank correlation matrix for these variables. The programme named MILK GETS IN YOUR TEA considers only two variables at a time, but it can group observations and calculate rank correlation coefficients for each sub-group before calculating the over-all rank correlation coefficient.

Both programmes are limited in their capacity to store information. This constraint should be relaxed considerably after May 1974, when the computer memory is to be augmented, but until then, MILK GETS IN YOUR TEA can handle only 350 observations. TAKA TAKA must work within the following constraint: 1

$$2\,\theta . V + 6\theta + 2V^2 - 2V \leq 6230$$

where $\theta \equiv$ the number of data observations

$V \equiv$ the number of variables per observation.

The right hand side of the inequality reflects the size of the memory and should increase very substantially after May, 1974. A graph in Appendix III shows the approximate frontier of the inequality under current conditions. At the present, TAKA TAKA can handle up to 620 observations if only two variables are used. With 10 variables, 230 observations can be used, with 20 variables about 118 observations and so on.

1. TAKA TAKA

This programme ranks a number of variables for anumber of observations, prints out an observation matrix with values and ranks, and then calculates and prints a matrix of correlation coefficients between all pairs of variables. At present, the programme is set up to accomodate up to 70 observations of up to 29 variables each, but these values can be changed (See Section A, c) below). We refer to this 70, 29 mix as the "set capacity."

A. Input.

(a) Programme cards

The Programme requires three control cards.

IV = M)
IΘ = N) Integers (no decimal)
C = X Real (decimal)

The first two cards give the number of variables and observations respectively (designated by "M" and "N" above). The last card gives the value of a constant, C, which will be used to convert any non-integer input data to integer form. Each data input is simply multiplied by C to inflate its value and thereby get rid of decimal values. (If all data are already integers, then C should be set equal to one)

Example: If there are 43 observations of 24 variables each, then IΘ = 43 and IV = 24.

If the smallest data value read into the programme is $.0035 = 35 \times 10^{-4}$, then $10^4 = 10{,}000$ would be an appropriate value for C, and the third card would read

C = 10000.

Note however, that the largest data value times C must not exceed 8×10^6 (eight million) or the computer will give an error message.

The value stored in the computer memory is simply the portion of the product of C times the values of the data on the card which lies to the left of the decimal.

Example: If C = 1000 and a certain variable is fed into the computer as .03976, then the value 39 will be stored in the memory. 39 is simply the portion to the left of the decimal of 39.76 = .03976 X 1000.

For purposes of ranking, only the order must be preserved in a transformation of data values, so lost decimal figures are generally of no significance.

In addition to these three required control cards, one of two optional cards may be added to the deck. If only the correlation matrix is desired, then the following card should be included:

NØØBS = 1

If only the observation matrix, with rankings, is desired (no correlation matrix) then this card is included:

NØCØR = 1

If neither of these cards is included, then both matrices will be printed.

All control cards should be punched from column 7 onwards. They should be inserted after the tenth card in the deck.

b) Data Cards.

Data may be entered in any format desired, provided the appropriate FORMAT instruction is included. Before we discuss procedures for doing this, however, we will describe the input format already provided for in the programme.

i) Prescribed Data Format (if this format is followed, no changes need be made in the programme cards). Column 1 should be left blank Each data card should have a name of up to six characters in columns 2 -7. In the remaining columns, six places have been reserved for each variable. One card can take a name and twelve variables. If more than twelve variables are used for each observation, simply continue on as many cards as needed (up to a maximum of 29 variables, three cards, in the set capacity). The observation name should be repeated on each card. Card sequence may be indicated in column 80, if desired. Each variable must occupy no more than six columns including a minus sign and a decimal point, if needed. If no decimal point is included, then the variable value must be "right justified," that is, the last digit must lie in a column which is numbered a multiple of six plus one.

Note: Each six column field can contain information for one variable only. One data card must be filled with twelve variables before the next is begun. The same variable must appear in the same field for all observations.

(ii) Variable Names:

After the data cards, a list of variable names must be included. Each variable should have a name of up to eight characters, and a space should be left between each six character field. Thus, the first name should begin in column 1, the second in column 9, the third in column 18, and so on. Each card will accomodate up to eight names. Use as many cards as needed (again, upto 29 variable names, three cards, in the set capacity.)

(iii) Other Data Formats

If another data input format is desired, a FORMAT card must be changed in the program. This is not difficult, and a University Programming Advisor from the Computer Centre will show you quickly how to do this. The format card to be changed is labled 42 in columns one and two. (If more than 44 variables are to be used, then format card 850 must also be changed.)

c) Changing the Set Capacity for Observations and Variables.

This section is relevant only if more than 70 observations or more than 29 variables are to be used as input. Otherwise it can be skipped. TAKA TAKA has three segments, a MASTER segment and two subroutines named SORTY and AVRANK. Each segment has cards which specify the maximum capacity of certain storage areas in the computer's memory. These capacities may be changed as long as they satisfy the inequality presented earlier. In particular the following cards must be prepared:

In the MASTER TAKA TAKA segment

```
INTERGER ØBS (Θ, 2.V), IR (Θ), AM (Θ), TIE (Θ),
                          MFRIG (V), RM (Θ)
REAL NAME (Θ), UAR (V), CØR (V-1, V-1)
```

Example: If the program is to be changed to accomodate up to 10 variables and upto 230 observations, then the cards above must read

```
INTEGER ØBS (230, 20), IR (230), AM(230), TIE(230),
                          MFRIG (10), RM(225)
REAL  NAME(230) VAR(10), CØR(9,9).
```

In the subroutines, the corresponding cards must be changed, namely:-

```
SUBROUTINE SORTY (.............)
INTEGER AM(Θ), IRM (Θ), RM(Θ)
REAL NAME (Θ)
and
SUBROUTINE AVRANK (..........)
INTEGER AM(Θ), TIE (Θ), MFRIG(V), RM (Θ)
```

B. OUTPUT.

TAKA TAKA prints out two matrices.

(a) The Observation Matrix.

This matrix shows the value of each variable for each observation, and its rank. At the top of each pair of columns is the variable name. The value for each observation of the variable is printed in the left hand column and the rank is printed in the right hand column. On the far left are the observation names. It should be noted that variable values are multiplied by the constant "C" discussed above, while ranks are multiplied by 10. Thus the rank "10" is really 1.0 while the rank 215 is really 21.5 — a tied ranking, (This has been done to save storage.)

(b) The Correlation Matrix.

The correlation matrix contains the rank correlation coefficients for all pairs of variables. Since this matrix is symmetrical about the diagonal, only one side is printed.

In addition, the first column and the last row have been ommitted as these would contain only zeros in any case. All off-diagonal coefficients are printed out. If you do not find the coefficient you are looking for in column (J), row (I), it will appear in row (J) and column (I). The significance of these coefficients can be assessed using the table in Appendix II.

II. MILK GETS IN YOUR TEA.

This programme sub-divides the set of observations into upto 14 mutually exclusive groups and then calculates the rank correlation coefficient for each sub-set and for all observations together. It also calculates a table of 't' statistics for assessing whether the differences between the correlation coefficients of the various groups are significant. In addition, it gives the mean and standard deviation of each variable for each group.

Grouping of data can be accomplished either by including a group number with each observation, or by letting the computer divide up the observations according to the value of any specified variable. For example, a set of observations dealing with agricultural output may be grouped by size of holdings, total income, location, ethnic group and so on. If catagorical variables (such as location) are used for grouping purposes, then the group number must be coded onto the observation cards. If a continuous or ordinal variable is used (such as size of holding), then the computer can generate groups from group limits provided on a separate group limits card.

The grouping facility permits analysis along three dimensions rather than only two, albeit in limited form. Thus, for example, it is possible to find out not only whether income is related to family size, but whether this relationship differs between urban and rural families or by hectares of land owned, etc.

A. INPUT.

a) Programme Cards.

Two programme parameter cards are needed for every run.

IT = M (integer)

C = X (real number)

The first card specifies the number of groups used in the run (designated by "M" above), while the second gives a real number ("X" above) to be used in converting all input to integer values in the program. The role of this "C" is explained in section I.A.(a) above. If the data are already in integer form, then the value C = 1. should be used. (The data are integers if no decimal points are used.)

Example: if the data are to be divided into nine groups, and if the smallest value of a variable is 0.3 then the cards should read

IT = 9
C = 10

In addition to these two cards, a third card must be added if the computer is to group the observations according to group limits provided with the data.

IGR = 1

If this card is included in the deck, the computer will look for a data card on which group limits are provided. Then the computer will group the data according to any specified variable, using the group limits provided on the group limits card. If the IGR = 1 card is omitted, the computer will read the group number directly from the observation card.

The final programme card is a FORMAT statement which specifies which columns on the data card are to be read. Since this package deals with only two variables (plus a grouping variable) in any one run, it is necessary to specify which variables are to be read each time a run is undertaken. The FORMAT card in question is labled 100 (in columns one through three) and can be prepared with the help of a Programming Advisor at the University Computing Centre. We suggest that you find the fields (columns) containing the two variables to be ranked and the third variable used in grouping, and then that you seek the help of the Programming Advisor. The card is very simple, and anyone can learn how to prepare it in a quarter of an hour with a good teacher.[2]

The programme cards can be inserted after the 14th card in the MILK GETS IN YOUR TEA deck. (They may be inserted in any order.)

(b) Data Cards

Three types of data cards must be provided.

i) Group Limits. If the computer is to undertake the task of grouping the observations a card with group limits must be provided.

Example: if grouping is to be done by the value of the variable "Growth Rate of Population" then the following group limits might be provided. -.02 0.0 .02 .05 .10 These five class limits would divide the data into six groups as follows:

Group	Group limits
1	$X < -.02$
2	$-.02 \leq X$ 0.0
3	0.0 $X \leq .02$
4	$.02 < X \leq .05$
5	$.05 < X \leq .10$
6	$.10 < X$

The group limits may be punched from column 1, with each group limit occupying up to six columns. Thus the first group limit will end in column 6, the second in column 12, the third in column 18 and so on. A maximum of 13 group limits (using 78 columns) may be specified. This will provide for up to 14 groups.

NOTE. If the group limits card is included then the card IGR = 1 must appear with the programme cards. If this programme card is excluded, then the computer will look for group numbers on the observation cards. Thus the card IGR = 1 must appear in the programme segment if and only if a group limits card is also provided with the data.

ii) Observation Cards: These cards contain the data to be analysed. They may be organized in any format with the following constraints.

(a) The first column must be left blank on all but the final observation card. The final observation should have a 1 purched in the first column, before the observation name.

(b) Each observation must have a name of up to six characters as the first entry.

(c) Variable values must occupy the same fields for all observations (i.e. if the variable "age at birth" is punched into columns 8-10 for the observation KIHARA, it must be in these same columns for all other observations as well.)

(d) Variables which are likely to be used in grouping should follow the variables to be ranked and correlated, on the observation card. This is because the variable to be used for grouping must be read in after the two variables to be ranked.

(e) If group numbers are to be punched directly on the observation cards, these should also follow the variables to be ranked. Group numbers should begin with 1 and may go to 14.

(f) As many cards as necessary may be prepared for each observation. All 80 columns may be used on each card (as long as the first column on the first card of each observation but the last one is left blank).

(g) Variables may be coded with or without decimal points. (However, the format card must read in all variables in "F" format.)

(iii) Variable Names. The final card in the data set must contain two names, of up to eight characters each, for the two variables being ranked. These should occupy columns 1 - 8 and 9 - 16 respectively.

The data cards should be inserted before the final card in the I.D.S. deck.

B. OUTPUT

(a) Group Information. For each group, the value and rank of each observation for each of the two variables is printed. The observation values are multiplied by the constant "C" described above, while the rank is multiplied by 10. (Thus a rank of 10 should be read as 1.0 — the smallest value in the group — while a rank of 415 should be read as 41.5 — a tied ranking.)

After the observations and rankings are printed, the number of observations in the group are printed, followed by the value of Spearman's rank correlation coefficient for the group. Finally the group mean and standard deveiation is presented for each variable.

b) The "t" Test Matrix. After group results are printed for each group; a matrix is printed which may be used to determine whether differences between the rank correlation coefficients for the various groups are statistically significant. This matrix contains the "t" test of the difference between the rank correlation coefficients of groups i and j in row i and column j. In row j and column i, the degrees of freedom for this "t" test is given. Appendix II gives the critical values for the "t" test.

Example: To check whether the coefficient of group 5 is significantly different from the coefficient of group 9 look at the 5th row and 9th column of the "t" test matrix for the value of the "t" statistic, and then look at the 9th row and the 5th column to get the degrees of freedom for the "t" test.

c) The Pooled Run. After the "t" test matrix is printed, all the observations will be pooled into a single group and all the group information described above will be printed.

Footnotes

* These programmes were developed from a programme written for M. Cowen by Dr. F. Childe, formerly of the U.N. Computing Centre.

1. A graph of this constraint appears at the end of this paper.

2. It should be noted that the FORMAT card must always begin 100
FORMAT (I1,..........)
Also, all data must be read in as real numbers (i.e. in "F" or "E" format).

Appendix I Formulae Used

Spearman's Rank Correlation Coefficient

$$P = 1 - \frac{6\,(D^2 + T)}{N\,(N^2 - 1)}$$

where

$$D^2 = \sum_{i=1}^{N} (R_{1i} - R_{2i})^2$$ Differences squared

$$T = \frac{1}{12} \sum_{i=2}^{N} i(i^2-1) \quad \gamma(i)$$ Correction for ties

and R_{ni} is the rank of observation i by variable n

$\gamma(i)$ is the number of times i observations are tied on either ranking. (Example, if two ties of three observations each occur, then i=3 and $\gamma(i) = 2$)

N is the number of observations

"T" Test for Differences Between RHOs.

$$t = \frac{P_1 - P_2}{\sqrt{\frac{1-P_1^2}{N_1-2} + \frac{1-P_2^2}{N_2-2}}}$$

Note that no pooled variance was used.

$$DF = N_1 + N_2 - 2$$ Degrees of freedom

Appendix II Critical Values of the Correlation Coefficient

DEGREES * OF FREEDON	5% LEVEL	1% LEVEL	DEGREES OF FREEDOM	5% LEVEL	1% LEVEL
5	.750	.893	26	.374	.478
6	.714	.857	27	.367	.470
7	.683	.833	28	.361	.463
8	.648	.794	29	.355	.456
9	.602	.735	30	.349	.449
10	.576	.708	35	.325	.418
11	.553	.684	40	.304	.393
12	.532	.661	45	.288	.372
13	.514	.641	50	.273	.354
14	.497	.623	60	.250	.325
15	.482	.606	70	.232	.302
16	.468	.590	80	.217	.283
17	.456	.575	90	.205	.267
18	.444	.561	100	.195	.254
19	.433	.549	125	.174	.228
20	.423	.537	150	.159	.208
21	.413	.526	200	.138	.181
22	.404	.515	300	.113	.148
23	.396	.505	400	.098	.128
24	.388	.496	500	.088	.115
25	.381	.487	1000	.062	.081

* The degrees of fredon are two fewer than the number of observations.

Appendix II "t" Statistic Values

DEGREES OF FREEDOM	T95 (two tailed "t" test at 95% cl.)	T99 (2 tailed "t" test at 99% cl.)	DEGREES OF FREEDOM	T95	T99
8	2.306	3.355	25	2.060	2.787
9	2.262	3.250	26	2.056	2.779
10	2.228	3.169	27	2.052	2.771
11	2.201	3.106	28	2.048	2.763
12	2.179	3.055	29	2.045	2.756
13	2.160	3.012	30	2.042	2.750
14	2.145	2.977	35	2.020	2.724
15	2.131	2.947	40	2.021	2.704
16	2.120	2.921	45	2.014	2.690
17	2.110	2.898	50	2.008	2.678
18	2.101	2.878	55	2.004	2.669
19	2.093	2.861	60	2.000	2.660
20	2.086	2.845	70	1.994	2.648
21	2.080	2.831	80	1.989	2.631
22	2.074	2.819	90	1.986	2.631
23	2.069	2.807	100	1.982	2.625
24	2.064	2.797	120	1.970	2.617
			00	1.960	2.576

UNIVERSITY OF NAIROBI **COMPUTING CENTRE**

80 COL. DATA SHEET JOB: DATE PAGE No. OF

EXAMPLES OF INPUT FOR TAKA TAKA

Programme Cards

IO=29	Twenty nine observations
IY=18	Eighteen variables per observation
C=100.	Constant used to convert all data to integer form
NAOBS=1 NECER=1	Optional. One or neither of these cards may be included. If the first is included, no observation matrix will be printed; if the second is included no correlation analysis will be undertaken.

Data Cards

Observation Cards

```
ARINS2  210 -101    3.240 12.6  -.03    .16   492    0  0.0    81 10019   -571
  -       2   14      51  963    47      .2   5.4  99.  98.6   - 5  -5.         2
```

Above: Twenty Three Variables on two cards. Note sequencing in column 80.

```
CASE46 998.1  62    18
```

Above: Three variables on one card.

Variable Names

```
OUTPUT POPULA TEMPO  LIBRTY EQLITY FRTRNY OXPLOW CHANGE MONEY  TIME  VAR065
TAXREV CHILD# BYEBYE BABA
```

PUNCHED VERIFIED

UNIVERSITY OF NAIROBI — COMPUTING CENTRE

80 COL. DATA SHEET   JOB: DATE   PAGE No.   OF

E X A M P L E S O F I N P U T F O R M I L K G E T S I N Y O U R T E A

Programme Cards

```
      IT=6            Six groups
      C=10.           Constant used to convert all data to integer form
      IGR=1           Optional.  When this card is included, the computer will use the group  limits below to group the data.
100   FORMAT(I1,A6,14X,F4.0,F4.1,F5.3)  Note: All variables must be read in
                                        "F" format.
```

Data Cards

Group Limits Card

Five group limits are presented for six groups

```
 -.02  0.0   .02   .05   .10  This card should only be included when IGR=1, above
```

Observation Cards

```
KIAMBU 36.4   6214  1962  -.1   .026 }  Two observations from the same set.  The computer will group
1TURKAN 4.1    906  1970  .6    .008 }  the data.  TURKAN is the last observation
VARSTY 10064   91 16 984  -1   10.6   -.26   3  The last variable for this observation is its
                                                group number
```

Variable Names

```
MANUF QAG  GROW
```

PUNCHED   VERIFIED

Output from TAKA TAKA

From OBSERVATION MATRIX

Observation Names --->	D1025B			D105FB
POPULA	1966	130	14200	155
INCOMEPC	7708	260	14100	340
AG PROD	2963	250	33800	340
MAIZEUSE	7557	460	-1	100
LABINCOM	-142	320	11070	50
LANDUSE	22070	80	4505	20
EXPORTS	2820	240	0	200
12 MONTH	-10480	10	184300	480
↑ Variable Names	↑ Variable values (X "C")	↑ Ranks(x10)		

From CORRELATION MATRIX

	INCOMEPC	AG PROD	MAIZEUSE	LABINCOM
POPULA	0.8468	-0.1234	-0.6113	-0.1064
INCOMEPC	0.0000	-0.1116	-0.7759	0.0286
AG PROD	0.0000	0.0000	0.5145	0.3015
MAIZEUSE	0.0000	0.0000	0.0000	0.1611
LABINCOM	0.0000	0.0000	0.0000	0.0000
LAND USE	0.0000	0.0000	0.0000	0.0000
EXPORTS	0.0000	0.0000	0.0000	0.0000

Note that the value of Spearman's Rank Correlation Coefficient for AG PROD and INCOMEPC is located in the INCOMEPC row and the AG PROD column (the value is -0.1116). The AG PROD row and INCOMEPC column is left blank. Note also that the last row (12 MONTH) and the first column (POPULA) are left off the matrix. All coefficients are included, however.

OUTPUT FROM MILK GETS IN YOUR TEA

Group Results (Sample)

GROUP 3

NAME	PROF TAX VALUE	PROF TAX RANK	EXPORTS RANK	EXPORTS VALUE
OBS 12	110	40	40	91
OBS 19	90	30	10	0
OBS 115	2	10	30	50
OBS 182	63	20	20	16

4 OBSERVATIONS IN GROUP

GROUPED RHO = 0.2000 ← Spearman's Rank Correlation Coefficient for this group

PROF TAX

GROUP MEAN = 66.2500

GROUP 50 = 46.9950

EXPORTS

GROUP MEAN = 39.2500

GROUP 50 = 40.3474

"t" Test Matrix (Sample)

TTEST MATRIX

	1	2	3	4	5 ← Group Numbers
1	0.0000	2.6185	1.9031	-3.3119	-0.1236
2	55.0000	0.0000	0.8679	-2.6167	0.3651
3	21.0000	38.0000	0.0000	-2.4010	2.0010 ← "t" test of difference between RHO of groups 3 & 5.
4	59.0000	76.0000	42.0000	0.0000	1.5632
5	80.0000	97.0000	63.0000	101.0000	0.0000

↑ (63.0000) degrees of freedom for "t" test between groups 3 and 5.

APPENDIX III

FEASIBLE COMBINATIONS OF OBSERVATIONS AND VARIABLES FOR TAKA TAKA