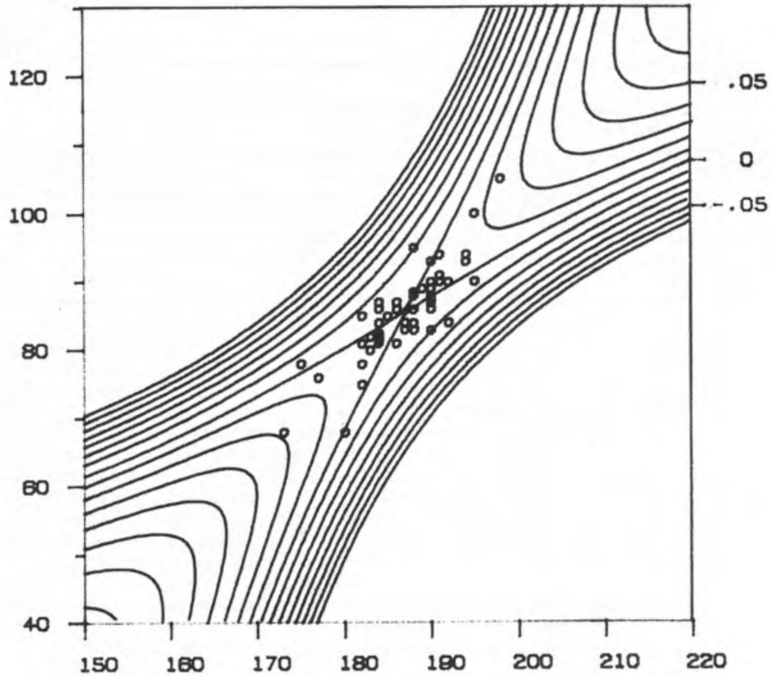
SURVO 76

Communications 1/1979



SURVO 76
Communications 1/1979
****

Editor: Seppo Mustonen, Department of Statistics
University of Helsinki, Hallituskatu 11-13
SF-00100 Helsinki 10, Finland

Contents:

RECENT DEVELOPMENTS IN THE SURVO 76 SYSTEM

In the beginning of 1979 the statistical data processing system SURVO 76 has been extended and updated in several ways.

The most important new contributions are:

1. New SURVO 76 version for Wang 2200VP

2. New SURVO 76 modules (DATAM, CORRM, CORRML) for treatment of missing data values
   and modification of basic SURVO 76 modules to work despite incomplete data,

3. New SURVO 76 modules for nonparametric methods,

   TABTEST for contigency tables,
   COMPARE for comparing two ore more independent samples (t test, Mann-Whitney test, Fisher-Pitman randomization test, Kolmogorov-Smirnov test, Cramer-von Mises test, one way analysis of variance, Kruskal-Wallis test)
   These modules compute both the values of the test statistics and the significance levels. In case of small samples the exact critical levels are obtained by simulation.

4. Means for computing with large partitioned matrices up to the capacity of the disk memory in the MATRI subsystem (MAXMA operations),

5. New special modules:
   STEPCLU for stepwise cluster analysis using Wilks' lambda, Hotelling's trace and minimum variance criteria according to new algorithms presented in P.Korhonen (1979), A stepwise procedure for multivariate clustering, Research report no.7, Computing Centre, University of Helsinki,
   MNONLIN for estimating of multivariate nonlinear regression models by least squares (a modification of NONLIN),
   CORROBU for interactive detecting of multivariate outliers using Mahalanobis' distances and for robust estimation of means, standard deviations and correlations,

6. Program descriptions for the basic SURVO 76 modules.

In the present form the SURVO 76 system consists of ca. 60 different statistical programs (SURVO 76 modules). The volume is almost 1000 kbytes; 5 diskettes are needed for SURVO 76.

In addition the following demonstration diskettes containing ready made conversations with SURVO 76 are available:

SURVO 76/D1 general demonstration (8 topics),
SURVO 76/D2 graphic demonstration (20 different visual representations made by SURVO 76 modules and displayed on the graphic CRT),
SURVO 76/D3 matrix demonstration (10 topics, showing possibilities of the MATRI subsystem in statistical matrix operations).

New 2200VP version of SURVO 76

Although the original SURVO 76 can be run on Wang 2200VP without any modifications, some alterations and extensions have been made to take the best advantage of the efficiency and improvements in Wang 2200VP.
The further development of SURVO 76 will now be concentrated upon the new VP version and the most suitable configuration is

2200VP with 32K memory or more,
dual diskette drive,
2282 graphic CRT plotter with a hard copy printer.

The SURVO 76 VP-version can be easily modified for various configurations of Wang 2200VP and MVP. For instance, the type of the CRT (24x80 or 16x64) and printer (line length 80,112 or 132) can be preselected. Also some SURVO 76 system constants (number of variables in a data file, number of data values in one page etc.) have been changed into a parametric form and can be easily adjusted according to the memory size and the needs of the user.

The most SURVO 76 modules employ dynamic storage allocation in the VP version and select automatically the maximum dimensions according to the central memory space which is available.

The speed of 2200VP makes also possible to provide SURVO 76 with many advanced statistical techniques. We are going to direct the development of SURVO 76 towards more sophisticated and computationally demanding statistical methods. The items 2,3,4 and 5 listed on page 3 are indications of this trend.

Treatment of missing data values in SURVO 76

The open structure of SURVO 76 makes possible for the user to cope with the missing observation values according to one's own preferences by using data transformations and means for conditional processing. Hence no special automatic control of missing valus is absolutely necessary.
However, in order to make handling of incomplete data more convenient, some refinements have been added.

SURVO 76 provides now modules (DATAM, CORRM, CORRML) for substituting missing values according to various techniques and for estimation of parameters (especially means, standard deviations and correlations) in case of incomplete data.

Also most of the SURVO 76 basic modules are working correctly despite missing values and do what they can in such situations.

Indicating missing values

Missing values in SURVO 76 data files are indicated normally by the value 9E99. When entering data values in DATA module (F2- and F5-starts) the answer '?' means entering 9E99 as a data value.
Since 9E99 is almost the maximum numeric value in Wang 2200 careless working with these values eventually leads to overflow and an error message will be displayed. However, all basic SURVO 76 have now been adjusted so that missing values will be noticed before they are processed.
In some modules it is also possible to indicate a restricted range for each variable separately and values outside these specific ranges are treated like missing values.

Special modules for incomplete data files

Module DATAM

This new SURVO 76 module can be used for substituting missing data values according to various techniques. DATAM can also be employed for complicated nonlinear data transformations. In DATAM missing values can be indicated in addition to 9E99 also in some other ways. The indicated missing values can then be substituted using any of the following methods:

1. selected value,
2. mean value of existing observations
3. linear least squares predictor
4. given function of other variables
5. classification by a given function

The substituted values together with the existing values of each variable examined by DATAM can be saved as values of another variable (image variable) in the same data file so that the original missing values will not be "destroyed".
Using DATAM (F5-start) all existing 9E99 values will be easily identified and listed.
By F6-start a missing value indicator variable can be defined for conditional processing of other SURVO 76 modules.

Modules CORRM and CORRML

CORRM computes means, standard deviations and correlations from incomplete data using a "maximum information" principle so that each estimate is computed from all available data values. When a considerable portion of data values are missing this technique may give biased results and the correlation matrix can be indefinite. To see what is the situation also a simple eigenvalue analysis is included.

CORRML does the same thing as CORRM but using the maximum likelihood principle along the proposal of Orchard and Woodbury (1972).
When the data is an incomplete sample from a multivariate normal distribution, CORRML gives maximum likelihood estimates of means, standard deviations and correlations (correlation matrix will now be non-negative definite) by an iterative procedure. As a by-product the 'best' regression estimates of the missing values will be obtained and saved in 'image' variables.

The results of CORRM and CORRML can be saved in correlation files and used as a starting point in LINREG, STEPREG, PCOMP, DISCRI and other modules for linear models and multivariate analysis.

How SURVO 76 acts in case of incomplete data?

When a 9E99 value is encountered the SURVO 76 modules take various actions according to their nature.

1. Working despite missing values

Many SURVO 76 modules control the missing values and produce the results using the available information. If the data to be processed contains 9E99 values, number of legal observations used to produce each result will be displayed.
Modules working along this principle are for instance, UNI, SORT, TABLE, DIAGRAM, HISTO, CORRM, CORRML.

2. Refusing to work when a 9E99 value is encountered

Usually the more advanced statistical methods are so sensitive to missing data values that it is wise to prevent their use in case of incomplete data.
Thus some of the SURVO 76 modules operating directly on SURVO 76 data files simply break their working when the first 9E99 value is encountered and an error message is displayed.
In connection of this error message often some hints about a possible continuation will be given.

Reference:

Orchard,T. and Woodbury,M.A.(1972): A missing information principle: Theory and applications, Proceedings 6th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 697-715.

Computing critical levels of test statistics by simulation

In the new SURVO 76 modules TABTEST and COMPARE for nonparametric statistical methods a new approach in presenting the results has been employed.
It is a well-known fact that the exact distributions of many test statistics used in nonparametric methods may be difficult to compute and using the asymptotic theory often leads to wrong results. There are also some fine testing principles as Fisher's method of randomization which "are almost impossible to apply unless the sample sizes are very small." (Conover,1971,p.357)
Since the value of a test criterion without any information of its critical level is almost worthless and consulting statistical tables may be tedious or even impossible (when there are no tables) we have tried to provide the signicicance levels in a very "SURVO 76-like" fashion which is heavily based on the true interactive method of use.

In our approach after the computing and displaying the value of the test statistic and other pertinent information, the testing module starts immediately estimating the critical level by simulation and displays continuously on the CRT the approximate critical level and information on its accuracy. So usually within a few seconds or at least some minutes the user will have a reasonable solution for the testing problem.

This 'instant testing procedure' is now in use for X↑2-test (TABTEST), Mann-Whitney, Fisher-Pitman randomization, Kolmogorov-Smirnov and Kruskal-Wallis comparison tests (COMPARE). The method works especially well when the sample sizes are small (typically n<50) and it helps just in those cases where the asymptotic theory is not valid.

As an illustration, let us see how the module TABTEST is used for testing contigency tables. When TABTEST has been selected we'll have the following display on the CRT:

```
SURVO 76: 'TABTEST'/SM
F1: BASIC START (TABLE INPUT)
F2: RESTART OF CRITICAL LEVEL SIMULATION
F3: PRINTOUT OF RESULTS

'TABTEST' COMPUTES BY SIMULATION THE CRITICAL LEVEL OF
THE PEARSON'S X↑2 STATISTIC IN A FREQUENCY TABLE.
THE CRITICAL LEVELS ACCORDING TO THE FOLLOWING ASSUMPTIONS
MAY BE OBTAINED:
  1. BOTH ROW AND COLUMN TOTALS ARE FIXED (FISHER'S TEST)
  2. ONLY ROW TOTALS FIXED (CONSTANT SAMPLE SIZES)
  3. NO FIXED MARGINAL TOTALS.
```

By F1-start the frequency table can be entered by "filling a form" on the CRT element by element. There are no essential restrictions in the dimensions of the table and the total number of observations, but of course, the fourfold tables and other small contigency tables are the most interesting and suitable cases.
After the input the table will be displayed with the X↑2-value and its critical level P according to the usual CHI↑2-approximation:

```
FREQUENCY TABLE:  N= 12
  0  1  2  3
  4  2  0  0
X↑2=   9.33  DF=  3  P=0.0248 (CHI↑2-APPROXIMATION)

COMPUTING CRITICAL LEVELS BY SIMULATION:
   1. BOTH ROW AND COLUMN TOTALS FIXED (FISHER'S EXACT TEST)
   2. ONLY ROW TOTALS FIXED (CONSTANT SAMPLE SIZES)
   3. NO FIXED MARGINAL TOTALS
SELECT 1,2 OR 3? 2
```

Thereafter the user can select one of the 3 alternatives for specifying the null hypothesis. Alternative 1 means proceeding along the principle of the Fisher's exact test where both the row totals and the column totals are fixed.
In alternative 2 only the row totals are fixed and this means comparing two or more independent samples of fixed sizes.
In alternative 3 no marginals, but only the grand total is fixed.
For alternatives 2 and 3 practically no ready made tables are available and even in alternative 1 only the case 2xn is covered by standard tables.
To be accurate one must know the origin of the table to be analyzed in order to select the right alternative which corresponds to the sampling scheme used. Although the same CHI↑2-approximation is valid for all three cases, these alternatives may give different results especially when n is small.

When the user has selected one of the three alternatives TABTEST starts generating of new random tables which all have the same properties as the original table from the point of view of the null hypothesis.
When a random table is ready TABTEST computes the X↑2-value and compares this value to the original X↑2.
The only recorded information will be the total number of random tables generated (N) and and the proportion p of tables having a X↑2-value >= the original value.
Then p will be a simulated critical level of the test and its accuracy will be measured by the standard error
s=SQR(p*(1-p)/N). Since the distribution of p is approximately normal with mean = the true critical level and standard deviation = s, it is also possible to compare the value of p to the nearest standard level e (e=0.001,0.01 or 0.05) and estimate the probability (q) that the null hypothesis will be rejected on that level.

Hence we have presented all the constituents for the display which is kept up to date during the whole simulation process:

```
FREQUENCY TABLE:  N= 12
  0  1  2  3
  4  2  0  0
X↑2=  9.33  DF=  3  P=0.0248 (CHI↑2-APPROXIMATION)

CASE 2: ONLY ROW TOTALS FIXED

REPLICATES  CRITICAL LEVEL P    S.E. OF P
      500         0.00800        0.00398
X↑2 IS SIGNIFICANT AT THE 1 % LEVEL WITH PROBABILITY 0.69217

TO STOP THE SIMULATION, PRESS RETURN(EXEC)
```

The underlined figures are changing after each 10 iterations. In this display we have formed N=500 random tables along the alternative 2 and we have p=0.008, s=0.00398, e=0.01, q=0.69. On 2200VP it took 35 seconds to reach this situation and the following table illustrates the continuation of the same simulation:

N	p	s	e	q
500	0.00800	0.00398	0.01	0.69217
1000	0.00600	0.00244	0.01	0.94928
1500	0.00600	0.00199	0.01	0.97757
2000	0.00500	0.00157	0.01	0.99923
2500	0.00640	0.00159	0.01	0.98800
3000	0.00633	0.00144	0.01	0.99432
3500	0.00685	0.00139	0.01	0.98787
4000	0.00675	0.00129	0.01	0.99396
4500	0.00711	0.00125	0.01	0.98945
5000	0.00680	0.00116	0.01	0.99705
5500	0.00690	0.00111	0.01	0.99717
6000	0.00666	0.00105	0.01	0.99924
6500	0.00707	0.00103	0.01	0.99753
7000	0.00714	0.00100	0.01	0.99773

So it is obvious that in this case X↑2=9.33 is significant at the 1 % level and in fact, the approximate true critical level is p=0.007, if we assume that the row totals are fixed. Observe also that for this table the usual CHI↑2-approximation seems to be rather conservative.

When this approach is used even Fisher's randomization principle becomes applicable for quite reasonable sample sizes (on 2200VP for n=50 or even more). For instance, the new SURVO 76 module COMPARE contains the Fisher-Pitman randomization test for comparing two independent samples, (for the definition of this test, see, for instance, Conover, 1971,p.357-364.)
The exhaustive enumeration of critical combinations needed for the traditional approach is really a monumental task for sample sizes of 15 and 20, but the 'instant simulation' gives satisfactory results usually without any delay.

Reference:

Conover,W.J.(1971): Practical Nonparametric Statistics, John Wiley, New York.

LIST OF SURVO 76 MODULES 4.5.1979

GUIDE:	SURVO 76 TEACHER
DATA:	DATA INPUT, SAVING, EDITING AND TRANSFORMATIONS
DATA2:	TRANSFERRING AND COMBINING DATA FILES
UNI:	UNIVARIATE STATISTICS
CORR:	MEANS, STANDARD DEVIATIONS AND CORRELATIONS
SORT:	DATA SORTING AND ORDER STATISTICS
TABLE:	2-DIMENSIONAL CLASSIFIED FREQUENCY TABLES, TABLES FOR MEANS AND STANDARD DEVIATIONS, TABLE EDITING ON THE CRT, CHI↑2 AND T TESTS, 1-AND 2-WAY ANALYSIS OF VARIANCE
HISTO:	UNIVARIATE CLASSIFIED FREQUENCY DISTRIBUTIONS, HISTOGRAMS
PLOT:	PLOTTING A TIME SERIES OR SCATTER DIAGRAM (MAX 170 OBSERVATIONS, AUTOMATIC SCALING)
DIAGRAM:	PLOTTING A TIME SERIES OR SCATTER DIAGRAM (UNLIMITED NUMBER OF OBSERVATIONS, SCALING IS AUTOMATIC OR DETERMINED BY THE USER, ALSO ANY NONLINEAR SCALE CAN BE SPECIFIED)
CURVE:	CURVE PLOTTING
SURFACE:	SURFACE PLOTTING IN CENTRAL PROJECTION
CHANCE:	RANDOM DATA GENERATOR, SIMULATION OF VARIOUS DISTRIBUTIONS ON THE CRT
FRAME:	HALF PREPARED SURVO MODULE FOR INTERACTIVE COMPOSING OF NEW SURVO MODULES
LINREG:	MULTIPLE LINEAR REGRESSION ANALYSIS
STEPREG:	STEPWISE LINEAR REGRESSION ANALYSIS WITH LINEAR PARAMETER CONSTRAINTS
NONLIN:	NONLINEAR REGRESSION ANALYSIS AND NONLINEAR OPTIMIZATION
PCOMP:	ANALYSIS OF PRINCIPAL COMPONENTS, PRINCIPAL AXES SOLUTION FOR FACTOR ANALYSIS
FACTA:	ORTHOGONAL ROTATIONS IN FACTOR ANALYSIS ON THE CRT, GRAPHICAL, VARIMAX AND QUARTIMAX ROTATIONS
SPECTRUM:	AUTO- AND CROSS-CORRELATIONS, SPECTRAL ANALYSIS
MATRI:	MATRIX OPERATIONS ON MATRICES IN SURVO FILES OR MATRICES GIVEN BY THE USER
DISTRIBS:	VALUES OF THEORETICAL DENSITY AND DISTRIBUTION FUNCTIONS
DISCRI:	MULTIPLE DISCRIMINANT ANALYSIS
CLASSI:	CLASSIFICATION OF OBSERVATIONS USING MAHALANOBIS D↑2 AND BAYES PROBABILITIES
LINCO:	LINEAR COMBINATIONS OF VARIABLES, PRINCIPAL COMPONENT, FACTOR AND DISCRIMINANT SCORES
DATASORT:	SORTING A DATA FILE AND TRANSFERRING THE SORTED DATA IN ANOTHER FILE
PRINT:	'NEAT' PRINTOUT OF SURVO 76 DATA FILES

DATAM:	SUBSTITUTING MISSING VALUES OF DATA BY LINEAR LEAST SQUARES PREDICTORS AND OTHER CRITERIA
CORRM:	MEANS, STANDARD DEVIATIONS AND CORRELATIONS FROM INCOMPLETE DATA (MAXIMUM INFORMATION PRINCIPLE)
CORRML:	MEANS, STANDARD DEVIATIONS AND CORRELATIONS FROM INCOMPLETE DATA (MAXIMUM LIKELIHOOD ESTIMATION)
CORROBU:	DETECTING OUTLIERS FROM A MULTIVARIATE NORMAL SAMPLE ACCORDING TO MAHALANOBIS' D↑2, ROBUST CORRELATIONS
TABTEST:	COMPUTING BY SIMULATION THE CRITICAL LEVEL OF THE CHI↑2-STATISTIC IN A FREQUENCY TABLE
COMPARE:	COMPARING TWO OR MORE INDEPENDENT SAMPLES USING VARIOUS PARAMETRIC AND NON-PARAMETRIC TESTS
N-TEST:	TESTS OF NORMALITY (SHAPIRO-WILK ETC.)
MN-TEST:	TESTS OF MULTINORMALITY
NORMA:	IMPROVING THE (MULTI)NORMALITY OF THE DATA USING THE POWER TRANSFORMATIONS
DEPEND:	TESTS FOR INDEPENDENCE OF VARIABLES
STEPCLU:	CLUSTERING OF OBSERVATIONS USING WILKS' LAMBDA, HOTELLING'S TRACE AND (MINIMUM) VARIANCE CRITERIA
COPY:	RAPID TRANSFERS OF DATA FILES
TDATA:	AS 'DATA' BUT AUTOMATIC LABELLING FOR TIME SERIES OBSERVATIONS
MATDATA:	TRANSFERS A MATRIX SAVED ON DISK IN A SURVO 76 DATA FILE
AGGRE:	AGGREGATION OF OBSERVATIONS
HALEY:	SEEKS ALL THE ROOTS OF AN ALGEBRAIC EQUATION
BINORM:	SIMULATION OF BIVARIATE NORMAL DISTRIBUTION ON THE CRT
CURVE2:	AS "CURVE", BUT ALSO FOR IMPLICIT FUNCTIONS
SCURVE:	THE FUNCTION PLOTS OF MULTIDIMENSIONAL DATA BY THE METHOD OF ANDREWS
CLUSTER:	CLUSTERING OF OBSERVATIONS (ACCORDING TO ISODATA)
RESTREG:	LINEAR REGRESSION ANALYSIS WITH LINEAR PARAMETER CONSTRAINTS
PARTCORR:	PARTIAL CORRELATIONS, CONDITIONAL MEANS AND STANDARD DEVIATIONS
FOSS:	EXPONENTIAL CURVE FITTING BY NUMERICAL INTEGRATION
MULTGEN:	GENERATING SAMPLES FROM MULTIVARIATE NORMAL DISTRIBUTION
MNONLIN:	MULTIVARIATE NONLINEAR REGRESSION ANALYSIS, ORDINARY LEAST SQUARES METHOD

On the front page:

Influence curves for correlation coefficient

The kernel of this graph is a correlation diagram (r=0.85) for the height x (cm) and the weight y (kg) of n=48 athletes.

The influence curve with parameter a describes the location of such new observations x,y which produce an increment of a to r.
The equation of this influence curve can be shown to be

```
        (r(1-z)-uv)/z=a,
```

where

```
        u=sqr(n/(n↑2-1))(x-m(x))/s(x),
        v=sqr(n/(n↑2-1))(y-m(y))/s(y),
        z=sqr((1+u↑2)(1+v↑2))
```

and m(),s() are notations for mean and standard deviation, respectively.

The influence curves have been plotted for

```
        a=-0.05,-0.04,...,0.04,0.05
```

by using the SURVO 76 module CURVE2 intended for plotting of contour curves and implicit functions in general.

It can be seen that, for instance, a new observation x=175, y=89 would decrease r from 0.85 to 0.78.