Proc. Second International Tampere Conference in Statistics
(Tampere, Finland, 1-4 June 1987)
Tarmo Pukkila and Simo Puntanen, *Editors*

pp. 205 - 224

INVITED PAPER

Editorial approach in statistical computing

Seppo MUSTONEN
University of Helsinki, Finland

Key words and phrases: Interactive analysis, work station, operating system, SURVO 84.

ABSTRACT

This paper describes an environment for statistical computing, data management, graphics and report generating on a micro computer. The main approach is *editorial mode* which permits the statistician to control all stages of the work by a general text editor. Even the statistical data sets can be written in the edit field which is a visible work sheet on the screen. For large data sets special data files are provided. Information from other sources, like text files, can be easily processed.

The editorial approach is characterized mostly by examples taken from the SURVO 84C system which is entirely based on the editorial approach.

1. INTRODUCTION

SURVO 84 is an integrated interactive system for statistical analysis, computing, graphics and report generation. It also includes features related to spread sheet computing, matrix algebra and computer aided teaching. It provides tools for making application programs in various special areas. All functions are based on the **editorial approach** developed by the author in 1979. The center of the activities is an **edit field** that at all times is partially visible on the screen. The edit field is maintained by the **SURVO 84 Editor.**

The user works with SURVO 84 by typing text in the edit field and by activating various operations and commands written among the text. In many applications it is convenient to create **work schemes** including several extra **specifications,** also written in the text and in arbitrary order.

The data and the results of various operations and application schemes (like plotting schemes and matrix programs) are displayed in the same edit field when required. For more extensive data sets and tables of results SURVO 84 provides its own file representations. SURVO 84 can also communicate with text (ASCII) files.

From the user's point of view SURVO 84 is one huge program which is controlled along certain general principles. The truth is, however, that SURVO 84 is a collection of several technically independent programs (modules) which are called by the SURVO 84 editor according to the user's activations. The user hardly notices the shifting of programs, but sees the system as one integrated world without any need to know its internal structure.

As a collection of programs, SURVO 84 is open for additional modules made by experienced users according to certain rules. These rules and different tools for making modules are described in a separate document *"Programming SURVO 84 modules in C"*. After a new module has been programmed and compiled, the commands and operations defined in it can be used as any standard SURVO 84 operation.

The open structure of SURVO 84 allows calling any other program and using it while staying in SURVO 84. After finishing the job with the other program we shall be back in our current SURVO 84 session again. Because the commands of the operating system can also be employed in this way, SURVO 84 can be considered an extension of the operating system.

The SURVO 84 system may be compared to any extensive text processing program. However, when using SURVO 84 as a word processor we readily have all other activities available, too.

SURVO 84 is also a tool for making new application programs. It provides several ready-made structures and user-friendly "languages" for such tasks. The SURVO 84 **matrix interpreter** and working modes like **tutorial** and **touch mode** are examples of such an approach.

Basically SURVO 84 is intended for professional users, but it is an easy system even for a beginner, since everything is based on simple text editing. Speaking about "ease" in this context may be misleading. If a system is made easy and friendly just for a beginner, after a short learning period it may turn out to be very frustrating for a user who already knows its characteristics.

A good system should be like a musical instrument that requires a lot from its player before yielding its best. If, for example, the violin were invented in recent days, many people would object to its poor "user interface". However, the violin is far more advanced than mechanical, simple musical instruments, since it gives scope for true skills and even for virtuosity.

If one knows the main ideas and working methods of SURVO 84, there is no need to read manuals and user's guides. The best and always up-to-date source of information is the system's own inquiry and help facility, which is readily available during any SURVO 84 session.

Another way to get acquainted with the system is to watch tutorials recorded during normal SURVO 84 sessions. The users can produce such teaching programs on any topic during the work by turning on **tutorial mode.** This permits saving of all actions selected by the user.

The purpose of this paper is to present some typical working methods of SURVO 84. We feel that it is impossible to transmit ideas of an interactive system by structural and theoretical considerations only. Therefore we try to give many practical examples by presenting displays from SURVO 84 sessions. We know, however, that even in this form, a paper is too rigid a medium to give a true picture of a dynamic system.

A historical note

Many of the ideas and principles appearing in SURVO 84C have been adopted from the earlier versions. The first in line was SURVO 66 originated by the author in 1966 and implemented on Elliott 803. One explanation to the name SURVO is the word "survey", since the first SURVO was primarily planned for analysis of survey data. It can also be derived from the Finnish verb *"survoa"* which means *"compress"*. The SURVO 66 jobs were controlled by a simple command language. The original SURVO 66 was further developed in the University of Tampere and is now known under the name of SURVO/71.

In 1976 the first interactive version SURVO 76 was initiated by the author. It was completed in 1984 by him and his research group. Originally SURVO 76 was made in conversational (menu-based) form. The editorial approach was introduced in 1979. SURVO 76 runs on the Wang 2200 minicomputer.

The work on SURVO 84 started in 1984 on the basis of SURVO 76 by using the interpretative Basic language. This was the the first microcomputer version and could be run on the Wang PC only.

The current SURVO 84C system was originated in 1985. From the user's viewpoint it is much like SURVO 84 and it is also highly compatible with SURVO 76. But the latest version is far more efficient and it allows wider applications, since it is programmed in the C language. SURVO 84C can be run on the MS-DOS microcomputers.

2. SURVO 84 EDITOR

The work area for the SURVO 84 EDITOR is an *edit field* which is entirely located in the central memory of the computer. The edit field has typically 100 lines and 100 columns. Each line is preceded by a control symbol for special notations; this control symbol is always initially '*' on each line.

During the work the user can maintain any number of edit fields. However, only one of them is active at a time (in central memory and partially visible on the screen). The others are in edit files on disks, but they can be scanned in a temporary window (by a SHOW operation) and/or loaded partially to the current edit field.

Usually one is working with various edit fields one after another by saving the current field after editing (SAVE operation) and loading another as a whole (LOAD). Thus it is simple to change the active field when necessary.

In its basic form one edit field corresponds to about two pages of normal text. However, in no application it is necessary to identify one edit field with a page or two in some report. When printing documents consisting of several pages the general PRINT operation of SURVO 84 will automatically take care about proper page division (obeying the wishes given by the user, of course). When defining the printout the user simply tells what are the edit fields and chapters in them which belong to the document. Also pictures made earlier by PLOT schemes may be included and positioned automatically.

For example, all the 20 pages of this paper have been produced by a single PRINT activation.

The edit field (as well as the edit files) normally contains text and tables written by the user, various SURVO 84 operations (commands, work schemes etc.) and their results. Representation of various data structures in the same space formed by the edit field is essential and gives exciting possibilities for combining different activities in a creative manner.

When working with larger data sets the space given in the edit field is not enough for the data itself. Although the dimensions of the edit field may be expanded (up to 600 lines with 100 columns, for example, by the REDIM command), it is not wise to create very large edit fields. For big data sets and tables SURVO 84 supplies special data files.

SURVO 84 also supports some other data representations and permits information from other files to be loaded to the edit field. Even text files created by other systems can be processed in SURVO 84. Furthermore SURVO 84 data and results can be easily moved back to text files.

2.1 Control of the edit field

When starting a new job the upper left corner of an empty edit field is displayed on the screen:

```
  1  1 SURVO 84C EDITOR Sat Dec 06 17:18:01 1986                A: 100 100 0
  1 *_
  2 *
  3 *
  4 *
  5 *
  6 *
  7 *
  8 *
  9 *
 10 *
 11 *
 12 *
 13 *
 14 *
 15 *
 16 *
 17 *
 18 *
 19 *
 20 *
 21 *
 22 *
 23 *
```

On the header line some basic information is given like date and time, the data disk drive designation (A:) and the size of the edit field ('100 100' means 100 lines and 100 columns).

The user can now start writing text as on a standard typewriter. The ENTER key moves the cursor to the next line. A new line is initialized automatically when the visible line becomes full. Correspondingly, when the last visible line has been filled, the visible part of the edit field automatically scrolls upwards giving space for a new line at the bottom.

It is always possible to move the cursor in the field and among the text by using the arrow keys or the PgUp (previous page) and PgDn (next page) keys.

Simple editing takes place by typing over the previous text and by using INSERT and DELETE keys for inserting and deleting texts, respectively. To make room for new empty lines between current text lines, the INSERT key in upper shift has to be pressed. DELETE in upper shift deletes the current line entirely.

In all editing functions the foremost principle is to keep them as simple as possible. The best way to learn these simple actions is to practice them at the computer.

2.2 Operations and commands

Because SURVO 84 includes hundreds of activities that all are invoked from the editor, it would be too heavy to do everything by special keys and key combinations. Therefore each more advanced operation is carried out by writing on any free line some command words that are activated by the ESC key.

Assume that we have written in the edit field:

```
 24  1 SURVO 84C EDITOR Mon Jan 19 17:48:14 1987                A: 100 100 0
  1 *
  2 *(Conover: Practical nonparametric Statistics, Wiley 1971, p.208)
  3 *Twelve sets of identical twins were given psychological tests to
  4 *determine whether the first-born of the twins tends to be more
  5 *aggressive than the other. The results were as follows, where the
  6 *higher score indicates more aggressiveness.
  7 *
  8 *              1  2  3  4  5  6  7  8  9 10 11 12
  9 *DATA First:  86 71 77 68 91 72 77 91 70 71 88 87 END
 10 *DATA Second: 88 77 76 64 96 72 65 90 65 80 81 72 END
 11 *
 12 *COMPARE First,Second,14  / TEST=Pairwise
 13 *
 14 *
 15 *
 16 *
 17 *
 18 *
 19 *
 20 *
 21 *
 22 *
 23 *
```

When writing this text our aim is to make a comparison between two paired samples which are typed on lines 9 and 10. The test statistics will be computed by a COMPARE operation (line 12) referring to samples (First,Second) and giving a line number for the results (14). TEST=Pairwise on the same line is an extra specification which determines the nature of comparison.

The COMPARE operation has now been activated and after completing its task the edit field will be changed into form:

```
24  1 SURVO 84C EDITOR Mon Jan 19 17:50:17 1987                A: 100 100 0
 1 *
 2 *(Conover: Practical nonparametric Statistics, Wiley 1971, p.208)
 3 *Twelve sets of identical twins were given psychological tests to
 4 *determine whether the first-born of the twins tends to be more
 5 *aggressive than the other. The results were as follows, where the
 6 *higher score indicates more aggressiveness.
 7 *
 8 *              1  2  3  4  5  6  7  8  9 10 11 12
 9 *DATA First:  86 71 77 68 91 72 77 91 70 71 88 87 END
10 *DATA Second: 88 77 76 64 96 72 65 90 65 80 81 72 END
11 *
12 *COMPARE First,Second,14_ / TEST=Pairwise
13 *
14 *Paired comparisons:
15 *Samples: N=12            First          Second          Difference
16 *Mean                     79.08333       77.16667       -1.916667
17 *Standard deviation       8.887768       10.37333        7.153617
18 *Paired t=-0.928 (P=0.1866 one-sided t test df=11)
19 *Wilcoxon signed ranks test=-0.756 (P=0.2247 normal approximation)
20 *Critical levels by simulation:
21 *              Differences Signed rank
22 *Critical level 0.19684     0.23807  N=17100
23 *Standard error 0.00304     0.00326
```

Before yielding the final results (on lines 14-23) the COMPARE operation displays various temporary information on the screen. For example, it tries to estimate the critical level of certain tests by Fisher's randomization principle by simulation. N=17100 (on line 22) gives the number of replicates until the user has interrupted the process.

Many commands and operations refer to edit lines (lines of the edit field) like 14 in the previous COMPARE operation. Instead of line numbers also line labels (of one character like A,B,x,y,+,-) in the control column can be used.

Because the operations (commands) are written into the text like any other information, it is easy to edit them and activate again. Thus the commands do not disappear from the the screen (edit field) after they have been completed. The user can scratch them like any text by the ERASE key, for example.

2.3 Work schemes

In demanding SURVO 84 applications various commands and operations together with various extra **specifications** are written as **work schemes** that to some extent resemble programs.

In a typical work scheme not only the activated operation but also the specifications appearing in the edit field may have particular influence. Each operation has its own specification words. If a specification is not given in the edit field, when the operation (work scheme) is activated by ESC, a default value is automatically entered. By using the specification the user can change the default values.

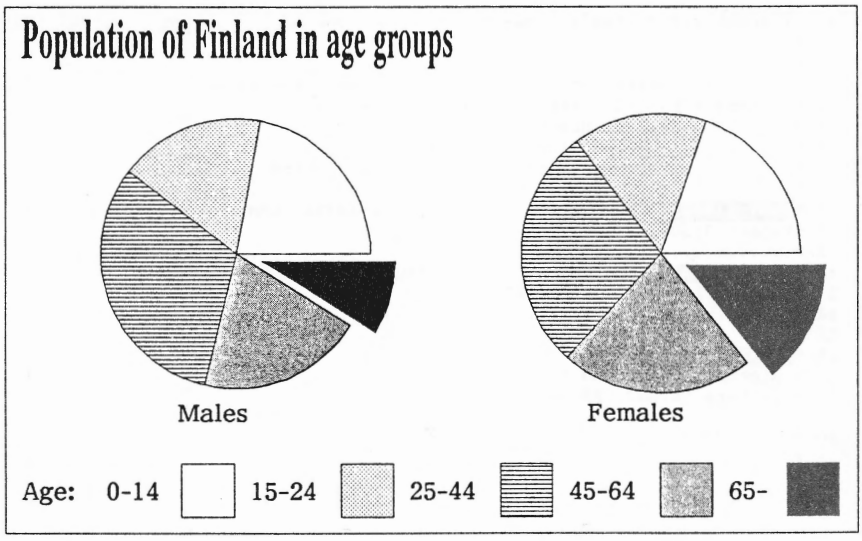
When planning SURVO 84 operations and program modules much attention has been paid to the selection of specifications and their default values. Good solutions in this respect lessen the burden of the user and make the system more intelligent. They form an essential part of the interface between the user and SURVO 84.

All specifications have the form <specification word>=<values>. For example, when making pictures by a PLOT operation the size of the graph is controlled by the SIZE specification. SIZE=1300,800 indicates that the width of the graph will be 1300 units and the height 800 units. If a laser printer is used, one unit equals 0.1 mm and in this case the dimensions are 13 cm by 8 cm.

Below a typical work scheme with the result is displayed. Note the free setting of specifications (no strict order) and the possibility to alter any detail in the scheme before a new activation.

```
 13  1 SURVO 84C EDITOR Sat Dec 06 19:03:02 1986                A: 100 100 0
  1 *
  2 *
  3 *   Population in Finland (1000)
  4 *
  5 *DATA FINLAND
  6 *Sex       0-14 15-24 25-44 45-64 65-
  7 *Males      506   399   727   458   202
  8 *Females    484   381   693   547   353
  9 *
 10 *HEADER=([HIGH]),Population_of_Finland_in_age_groups
 11 *PLOT FINLAND
 12 *SIZE=1300,800       size of the picture (13*8 cm)
 13 *TYPE=PIE            pie chart
 14 *LEGEND=Age:         text before the legend
 15 *SHADING=0,1,5,8,9P shadings (colors), P=pull out sector
 16 *XDIV=0,1,0          no vertical margins
 17 *
 18 *
 19 *
 20 *
 21 *
 22 *
 23 *
```

When the PLOT operation on line 11 is activated, the following graph is produced on the laser printer:



2.4 Subfields

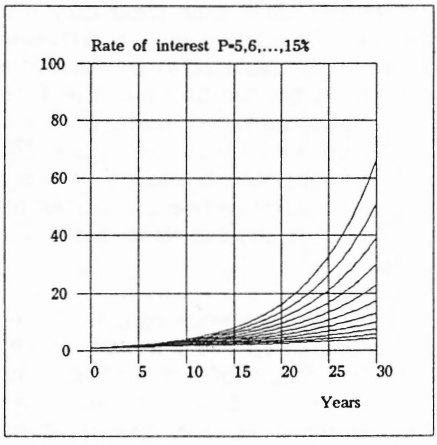
Work schemes which are placed in the same edit field may disturb each other if they are using the same specifications in different ways. To avoid confusions, the schemes may be isolated from each other by typing a border line between them. A border line has the form *..... (i.e. '*' in the control column followed by at least 10 dots.) If the whole border line (as usual) is filled with dots, the editor displays it as a thicker stripe that clearly reveals the boundary between the work schemes. Then all operations activated between two border lines will adhere to the specifications in this limited area only. The area is called a *subfield.*

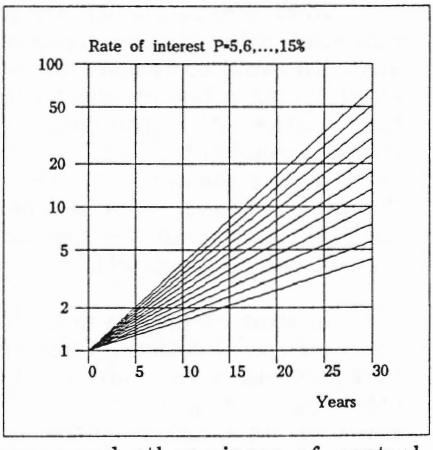
In some cases it is desirable to have several work schemes using the same specifications. Those specifications have to be written in the first subfield in the current edit field and this subfield must contain the keyword *GLOBAL* arbitrarily positioned in that subfield. When now some work scheme in the edit field is activated, the specifications are primarily searched for in the local subfield and secondarily in the *GLOBAL* subfield. If neither local nor global specification is found, a default value is maintained.

Below two families of curves are plotted with separate PLOT schemes but using some common specifications given in the *GLOBAL* subfield.

```
12  1 SURVO 84C EDITOR Sun Dec 07 14:03:40 1986              A: 100 100 0
 1 *
 2 *
 3 *Compound interest on linear and logarithmic scales
 4 * *GLOBAL* P=5,15,1 SIZE=650,650 PEN=[INDEX]
 5 *          GRID=XY HEADER=
 6 *          XSCALE=0(5)30 XLABEL=Years
 7 *          YLABEL=Rate_of_interest_P=5,6,...,15%
 8 *.....................................................................
 9 *PLOT Y(X)=(1+P/100)^X          / DEVICE=INTER1.CAN
10 *YSCALE=0(20)100
11 *.....................................................................
12 *PLOT Y(X)=(1+P/100)^X          / DEVICE=INTER2.CAN
13 *YSCALE=*log(y),1,2,5,10,20,50,100
14 *.....................................................................
15 *
16 *PRINT 17,18
17 - picture INTER1.CAN,250,100
18 - picture INTER2.CAN,940,100
19 *
20 *
21 *
22 *
23 *
```

The PLOT schemes on lines 9 and 12 save their results in files INTER1.CAN and INTER2.CAN, respectively. The PRINT operation on line 16 yields the following pair of graphs:





Since the commands, work schemes and other pieces of control information reside in the edit field and are so easy to edit, there is no need to abbreviate keywords as happens in many command languages. By using various copying and editing facilities offered by the editor it is possible to avoid writing of same words several times. It pays to learn how to avoid unnecessary work and how to use old material for new applications. Furthermore, by saving pertinent edit fields the user may create work schemes (or programs) that form a basis for new altered and improved applications.

2.5 Operation sequences

A series of operations and work schemes can be activated by one touch by placing the operations needed on consecutive edit lines. If the first operation is activated by pressing F2:PREFIX and then ESC, then after the execution of the first operation the cursor will automatically be moved to the next line and the operation on that line is activated. The operations are carried out as long as there are feasible operation lines. Usually an empty line is used to interrupt the operation sequence.

It is not necessary that the operations belonging to the operation sequence are written on consecutive lines, since jumps to other lines may be performed by a GOTO command. For example, GOTO 24,30,15 changes the display in the edit field so that edit line 24 will be the first visible one and the cursor will be located on the 15th position of the 30th line. Thus any place in the current edit field can be reached during a sequence of operations.

Likewise jumps to other edit fields may be done by using a LOAD operation. For example, LOAD PART2,1,10 replaces the current edit field by another (PART2) from the data disk and places the cursor there on the 10th line so that the first edit line is the first visible line.

The next display shows a operation sequence that saves a 3x4 matrix A in a matrix file, computes AA' and its inverse matrix and finally displays the result in the edit field.

```
  1  1 SURVO 84C EDITOR Sun Dec 07 14:26:33 1986                A: 100 100 0
 24 *
 25 *Computing the inverse matrix of AA'
 26 *
 27 *MATRIX A ///
 28 *       12.5   4.2  11.0  -8.1
 29 *        0.0  -1.3   5.6   2.4
 30 *        5.2  10.4  -9.3   0.3
 31 *
 32 *MAT SAVE A
 33 *MAT B=MMT(A)          / *B~A*A' S3*3
 34 *MAT B=INV(B)          / *B~INV(A*A') 3*3
 35 *MAT LOAD B,37
 36 *
 37 *MATRIX B
 38 *INV(A*A')
 39 *///              1         2         3
 40 *  1        0.00347 -0.00662 -0.00200
 41 *  2       -0.00662  0.06308  0.01857
 42 *  3       -0.00200  0.01857  0.00998
 43 *
 44 *
 45 *
 46 *
```

Our operation sequence is on edit lines 32-35 and contains MAT operations only. The situation after the completion of the sequence is displayed. The changes and results due to the operations are indicated in a gray shading. The parameter 37 of the MAT LOAD operation on line 35 refers to the first line of the results.

3. ARITHMETICS AND SPREADSHEET COMPUTING

SURVO 84 provides two different working modes for computing with numbers and tables. Furthermore the statistical operations and general mathematical tools (like the matrix interpreter) give support in more advanced applications. In this section, however, we are mainly considering typical calculations needed in normal work.

Editorial computing permits calculating values for various expressions which have been typed in the edit field. Even more complicated formulae or *computation schemes* (consisting of formulae with instructions) may be entered and activated with given starting values.

In addition to basic arithmetics, various mathematical and statistical functions are readily available. New functions may also be defined by the user either very easily in the edit field or by programming them in the C language.

The functions for editorial computing are also available in the VAR operation (described later) for tasks related to spreadsheet computing. The main applications of VAR are the transformations of variables in statistical data sets.

Touch mode is another unique approach in SURVO 84 for calculations. In this mode various computations are carried out by moving the cursor to touch any number in the edit field and the number is activated by pressing any of the keys +,-,* or / which correspond to standard arithmetical operations. Several numbers can be activated in this way and the resulting numerical expression will appear at each stage on the bottom line of the screen.

To print the current result in the edit field, the cursor is moved to indicate the desired position for output and the key = is pressed.

Touch mode enables very powerful tools for spreadsheet computing, too. Any systematic computing sequence consisting of several steps may be first defined by simply making the sequence in touch mode for the first case. After this definition stage the sequence may be automatically repeated for remaining cases by a single activation. These computation sequences, *touch chains*, can also be saved and used later when needed.

Touch mode provides the same mathematical and statistical library functions as editorial computing.

3.1 Examples on editorial computing

Arithmetic expressions are typed in the edit field according to normal mathematical notation.

For example, to calculate the arithmetic mean of numbers 12, 17 and 25 we enter:

```
 22  1 SURVO 84C EDITOR Sat Jan 03 17:08:54 1987                A: 100 100 0
  1 *
  2 *
  3 *        (12+17+25)/3=_
  4 *
  5 *
```

Now, when the cursor is blinking immediately after =, we press the activation key ESC. Because there is no command on the current line the SURVO 84 Editor studies the character just before the cursor position. If it is = as in this case, the editor assumes that the user wants to calculate something and calls the editorial computing module which analyzes the current expression, computes its value and writes the result in the edit field. Finally the control is transferred back to the editor and we may continue the work.

In this case we obtain immediately the following display:

```
 22  1 SURVO 84C EDITOR Sat Jan 03 17:08:54 1987                A: 100 100 0
  1 *
  2 *
  3 *         (12+17+25)/3=18
  4 *
  5 *
```

When same numbers are used for several computations or when more general expressions are wanted, we may type also

```
 31  1 SURVO 84C EDITOR Sat Jan 03 17:45:43 1987                A: 100 100 0
  1 *
  2 *         X=12 Y=17 Z=25
  3 * Arithmetic mean is (X+Y+Z)/3=
  4 * Geometric mean is (X*Y*Z)^(1/3)=
  5 *
```

After activating both expressions we get the following display:

```
 34  1 SURVO 84C EDITOR Sat Jan 03 17:45:43 1987                A: 100 100 0
  1 *
  2 *         X=12 Y=17 Z=25
  3 * Arithmetic mean is (X+Y+Z)/3=18
  4 * Geometric mean is (X*Y*Z)^(1/3)=17.2130062073
  5 *
```

We can also define temporary functions like AM and GM functions in the next display and activate several expressions simultaneously by having .= instead of = at the end of expressions:

```
 21  1 SURVO 84C EDITOR Sun Jan 04 18:27:00 1987                A: 100 100 0
  1 *
  2 *
  3 * Arithmetic mean  AM(X,Y,Z):=(X+Y+Z)/3
  4 * Geometric mean   GM(X,Y,Z):=(X*Y*Z)^(1/3)
  5 *
  6 *   AM(12,17,25).=
  7 *   GM(12,17,25).=
  8 *
  9 *   A=11.5 B=14.7 C=16.1
 10 *   AM(A,B,C).=
 11 *   GM(A,B,C).=
 12 *   AM(A+1,B+1,C+1).=
 13 *   GM(A+1,B+1,C+1).=
 14 *
```

The definitions of function AM and GM appear on lines 3 and 4. On lines 6-13 several expressions using these functions have been written. To evaluate these expressions, it is enough (since all of them are tailed by .=) to activate just one of them and we shall have

```
21  1 SURVO 84C EDITOR Sun Jan 04 18:27:00 1987                A: 100 100 0
 1 *
 2 *
 3 * Arithmetic mean  AM(X,Y,Z):=(X+Y+Z)/3
 4 * Geometric mean   GM(X,Y,Z):=(X*Y*Z)^(1/3)
 5 *
 6 *   AM(12,17,25).=18
 7 *   GM(12,17,25).=17.2130062073
 8 *
 9 *   A=11.5 B=14.7 C=16.1
10 *   AM(A,B,C).=14.1
11 *   GM(A,B,C).=13.9619801763
12 *   AM(A+1,B+1,C+1).=15.1
13 *   GM(A+1,B+1,C+1).=14.971612979
14 *
```

3.2 Computation schemes

It is always up to the user how he/she organizes the computations when working with the SURVO 84 Editor. In many applications all the formulas and operations needed for reaching a specific goal can be expressed as a **computation scheme** which to some extent resembles a computer program. One clear distinction is, however, that in a computation scheme there is no specific order of statements.

Each activation in a SURVO 84 work scheme leads always to a search process where the editor and other programs called for help are looking for the information needed for carrying out the task activated by the user.

In fact all the previous examples of editorial computing have been computation schemes in a modest sense. A real computation scheme, however, usually contains instructions and comments to help the user in applying the scheme.

For example, testing of correlation coefficient by using Fisher's z transformation could be represented as a computation scheme as follows:

```
 33  1 SURVO 84C EDITOR Sun Jan 11 12:22:39 1987                A: 100 100 0
 47 *  Testing the correlation coefficient
 48 *  The sample correlation coefficient is r and the sample size n.
 49 *  To test the hypothesis that in the population the unknown
 50 *  correlation coefficient rho is r0 against the alternative rho>r0,
 51 *  we form the test statistic
 52 *         U=sqrt(n-3)*(Fisher(r)-Fisher(r0))
 53 *  where
 54 *         Fisher(r):=0.5*log((1+r)/(1-r))
 55 *  is Fisher's transformation of the correlation coefficient.
 56 *
 57 *  If the null hypothesis is true, U is approximately N(0,1).
 58 *  Hence we reject the hypothesis, if P=1-N.F(0,1,U)
 59 *  is less than the risk level (say 0.05).
 60 *
 61 *  Assume now that n=25,  r=0.85 and r0=0.7
 62 *
 63 *  Then U.=1.8238788825   and P.=0.03408519244644
 64 *
 65 *  Thus if P<0.05, we reject the hypothesis that correlation
 66 *  coefficient in the population is r0.=0.7
 67 *.....................................................................
 68 *  Instructions: Insert your own values on line 61 and
 69 *                activate P.= on line 63, for example.
```

Above the notation P=1-N.F(0,1,U) on line 58 refers to a library function N.F() which is the standard normal distribution function.

4. DATA ANALYSIS

Tasks related to statistical analysis and statistical computing can be performed in SURVO 84C by various statistical operations. Certainly there is no clear-cut difference between genuine statistical operations and others. For example, the general calculating techniques (editorial computing and touch mode) and the VAR operation for making transformed variables are helpful in many tasks. Similarly SURVO 84C graphics and matrix operations are essential tools in statistical applications.

There are, however, certain structural factors which bind actual statistical procedures together in SURVO 84C.

The primary data values (samples etc.) are always represented as data lists or tables in the edit field or as data files on disk. Data from other sources, like text files, may be easily converted into SURVO 84C representation by special FILE operations.

In some statistical methods, however, the computations include several steps where the output of one step should be studied carefully before continuing with the next step. The common trend in the successive steps of the analysis is to compress the original data to sufficient statistics which often can be represented as matrices with lower dimensions than the raw data.

For example, in standard multivariate analysis based on multinormality of original variables the means, standard deviations and correlations form sufficient statistics. Then matrix files consisting of those statistics can be used as a basis for computations as well. On the other hand it is good to remember that there are often better compu-

tational techniques (based, for example, on orthogonalization of data) which do not rely on correlations at all.

In any case a statistical system should provide tools for operating with various intermediate results in matrix form as easily as with primary data. In SURVO 84C we have a large subsystem, the matrix interpreter, for these tasks. Many SURVO 84 operations related to linear models and multivariate analysis are making their matrix computations through the matrix interpreter. An advanced user may as well use the matrix interpreter directly by the MAT commands and chains.

In addition to standard data representation there are other traits which connect various statistical methods in SURVO 84C. Although any statistical operation may have its own special requirements, it is important that the user can expect each operation to work according to the same style as the neighbouring operations at least to some extent.

In sequel we shall present some typical applications. All these examples are based on artificial data sets (files) generated by the VAR operation.

4.1 Correlations

In the next exhibit 200 observations from a first-order autoregressive process are generated and then autocorrelations for lags 1,2,3,4,5 are computed.

The FILE CREATE scheme on lines 2-6 creates a data file AUTO1 for 200 observations of variable X. The VAR operation on line 8 generates the observations. Another VAR scheme on lines 11 and 12 computes the lagged variables and finally CORR AUTO1,15 on line 14 computes the means, standard deviations and correlations. The results are displayed from line 15 onwards. Furthermore they are saved in full precision in matrix files for subsequent analysis.

```
14  1 SURVO 84C EDITOR Sun Feb 22 11:06:33 1987          D:\STAT\ 120  80 0
 1 *
 2 *FILE CREATE AUTO1,24,6,64,7,200
 3 *  200 observations from X(t)=0.8*X(t-1)+eps
 4 *FIELDS:
 5 *1 N 4 X
 6 *END
 7 *
 8 *VAR X=if(ORDER=1)then(eps)else(0.8*X[-1]+eps) TO AUTO1
 9 *      eps=probit(rnd(1))
10 *.....................................................................
11 *VAR X1,X2,X3,X4,X5 TO AUTO1
12 *X1=X[-1] X2=X[-2] X3=X[-3] X4=X[-4] X5=X[-5]
13 *.....................................................................
14 *CORR AUTO1,15
15 *Means, std.devs and correlations of AUTO1  N=200
16 *# of missing observations =5
17 *Variable  Mean        Std.dev.
18 *X         -0.106722    1.756378
19 *X1        -0.102504    1.753710
20 *X2        -0.102143    1.753462
21 *X3        -0.104000    1.756002
22 *X4        -0.106121    1.758365
23 *X5        -0.110815    1.764946
24 *Correlations:
25 *              X       X1      X2      X3      X4      X5
26 * X           1.0000  0.8092  0.6211  0.4683  0.3556  0.2878
27 * X1          0.8092  1.0000  0.8088  0.6185  0.4665  0.3531
28 * X2          0.6211  0.8088  1.0000  0.8082  0.6184  0.4664
29 * X3          0.4683  0.6185  0.8082  1.0000  0.8087  0.6203
30 * X4          0.3556  0.4665  0.6184  0.8087  1.0000  0.8093
31 * X5          0.2878  0.3531  0.4664  0.6203  0.8093  1.0000
32 *
```

4.2 Fitting a univariate distribution

We shall study a mixture of two normal distribution of the form p*N($m1$,$s1$^2)+(1-p)*N($m2$,$s2$^2). We generate 10000 observations with parameters p=0.7, $m1$=0, $s1$=1, $m2$=2, $s2$=0.5 and try to re-estimate them from starting values INIT=0.5,-1,1.5,3,1.

A file SIMUDATA is first created for 10000 observations of variable X and values from the mixture are then computed by the VAR scheme on lines 8-11.

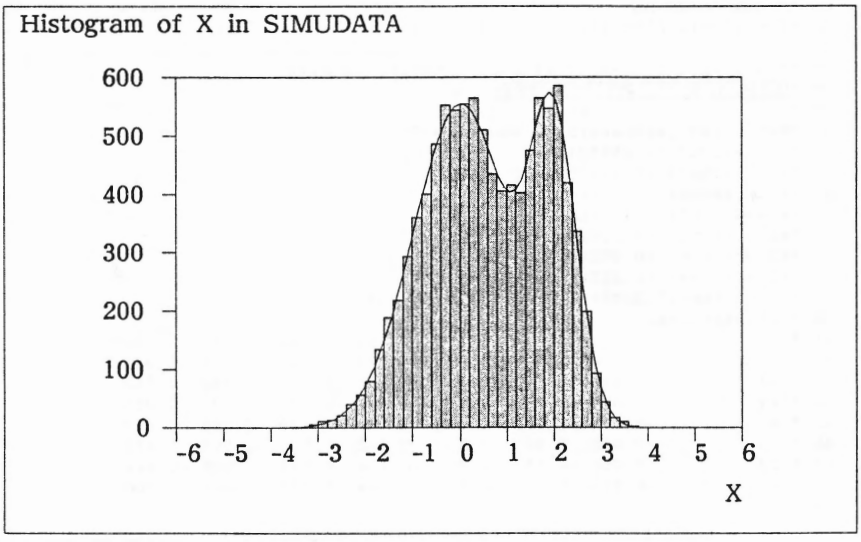
To estimate the parameters from the sample a frequency distribution of X is formed by a HISTO operation on line 17. The FIT=MIXNORM specification on line 19 implies HISTO to fit the MIXNORM distribution defined on lines 14-16 to SIMUDATA. The results of estimation are displayed on lines 22-35.

```
 20  1 SURVO 84C EDITOR Fri Feb 27 15:35:41 1987          D:\STAT\ 120  80 0
  1 *
  2 *FILE CREATE SIMUDATA,4,1,64,7,10000
  3 *   Sample (N=10000) from a mixture of two normal distributions
  4 *FIELDS:
  5 *1 N 4 X
  6 *END
  7 *
  8 *VAR X TO SIMUDATA
  9 *   X=if(rnd(1)<0.7)then(X1)else(X2)
 10 *   X1=probit(rnd(1))
 11 *   X2=0.5*probit(rnd(1))+2
 12 *.....................................................................
 13 *
 14 *DENSITY MIXNORM(p,m1,s1,m2,s2)
 15 *y(x)=c*(p/s1*exp(-0.5*((x-m1)/s1)^2)+(1-p)/s2*exp(-0.5*((x-m2)/s2)^2))
 16 *      c=0.39894226
 17 *HISTO SIMUDATA,X,22
 18 *X=-6(0.2)6 XSCALE=-6(1)6 YSCALE=0(100)600
 19 *FIT=MIXNORM INIT=0.5,-1,1.5,3,1
 20 *SIZE=1300,800
 21 *.....................................................................
 22 *HISTO: Estimated parameters of MIXNORM:
 23 *p=0.7003 (0.0125)
 24 *m1=0.0301 (0.0296)
 25 *s1=1.0077 (0.0193)
 26 *m2=2.0095 (0.0180)
 27 *s2=0.4906 (0.0137)
 28 *logL=16043.335886  # of function evaluations =311
 29 *Correlations:
 30 *                  p     m1     s1     m2     s2
 31 * p            1.000  0.845  0.796  0.728 -0.707
 32 * m1           0.845  1.000  0.802  0.673 -0.661
 33 * s1           0.796  0.802  1.000  0.572 -0.591
 34 * m2           0.728  0.673  0.572  1.000 -0.700
 35 * s2          -0.707 -0.661 -0.591 -0.700  1.000
```

HISTO also lists the frequency distribution and various related statistics as follows

```
 20  1 SURVO 84C EDITOR Fri Feb 27 15:35:41 1987          D:\STAT\ 120  80 0
 36 *Frequency distribution of X in SIMUDATA: N=10000
 37 *
 38 *Class midpoint  f     %   Sum     %     e      e     f     X^2
 39 *       <=-3.6    0   0.0     0   0.0   1.1
 40 *         -3.5    2   0.0     2   0.0   1.2
 41 *         -3.3    1   0.0     3   0.0   2.4
 42 *         -3.1    6   0.1     9   0.1   4.5    9.3    9     0.0
 43 *         -2.9   12   0.1    21   0.2   8.2    8.2   12     1.7
 44 *         -2.7   14   0.1    35   0.4  14.3   14.3   14     0.0
-------------
 71 *          2.7  199   2.0  9829  98.3 199.7  199.7  199     0.0
 72 *          2.9   93   0.9  9922  99.2 105.9  105.9   93     1.6
 73 *          3.1   44   0.4  9966  99.7  48.4   48.4   44     0.4
 74 *          3.3   18   0.2  9984  99.8  19.2   19.2   18     0.1
 75 *          3.5   11   0.1  9995 100.0   6.8
 76 *          3.7    4   0.0  9999 100.0   2.2
 77 *          3.9    1   0.0 10000 100.0   0.7
 78 *        > 4.0    0   0.0 10000 100.0   0.4   10.0   16     3.6
 79 *Mean=0.623320 Std.dev.=1.267161
 80 *Fitted by MIXNORM(0.7003,0.0301,1.0077,2.0095,0.4906) distribution
 81 *Chi-square=27.48 df=28 P=0.4921
 82 *
```

and plots the graph:



4.3 Nonlinear regression

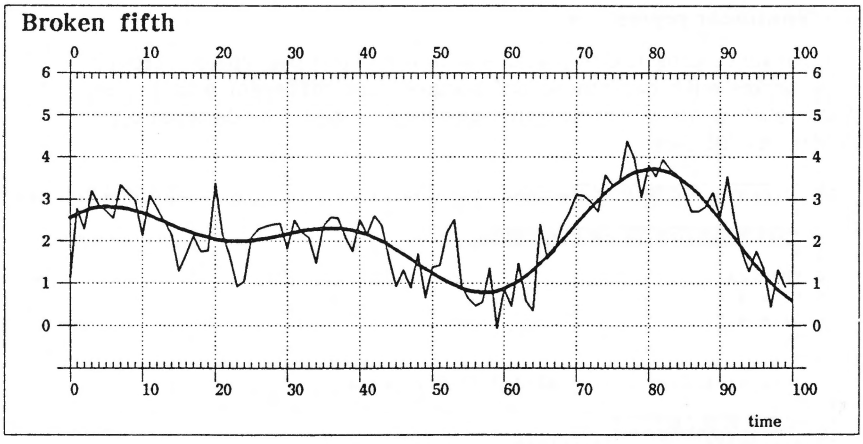
Our last example deals with a sample of 100 observations from a time series with two sinusoidal components, different amplitudes and phase shifts. Also a noteworthy noise is included. The data set is generated as follows:

```
 15  1 SURVO 84C EDITOR Sun Mar 01 13:39:27 1987          C:\S\ 100 100 0
  1 *
  2 *FILE CREATE SOUND,8,2,64,7,100
  3 *
  4 *FIELDS:
  5 *1 N 4 t
  6 *2 N 4 Y
  7 *END
  8 *
  9 *c=2 a1=1 f1=0.1 s1=0 a2=0.7 f2=0.15 s2=2
 10 *
 11 *VAR t,Y TO SOUND
 12 * t=ORDER-1 Y=c+a1*sin(f1*t+s1)+a2*sin(f2*t+s2)+eps
 13 * eps=0.5*probit(rnd(1))
 14 *
```

We shall smoothen this time series by using the ESTIMATE operation which can be applied to various nonlinear estimation problems. It is the user's task to specify the model (see lines 25-26 below). The default method in ESTIMATE is the ordinary least squares technique. ESTIMATE also forms analytically the first and second partial derivatives of the model function and this information is employed for selecting a proper optimization algorithm as well as for computing the gradient and the Hessian matrix.

Starting from crude initial values (on line 28) ESTIMATE gives the following results:

```
 27  1 SURVO 84C EDITOR Sun Mar 01 14:13:08 1987                C:\S\ 100 100 0
 24
 25 *MODEL INTERVAL
 26 *Y=c+a1*sin(f1*t+s1)+a2*sin(f2*t+s2)
 27 *.....................................................................
 28 *c=2.1 a1=1.1 f1=0.07 s1=1 a2=0.6 f2=0.18 s2=1
 29 *ESTIMATE SOUND,INTERVAL,31
 30 *.....................................................................
 31 *Estimated parameters of model INTERVAL:
 32 *c=1.993927 (0.058301)
 33 *a1=0.922475 (0.116705)
 34 *f1=0.104486 (0.004641)
 35 *s1=-0.300844 (0.241594)
 36 *a2=0.846823 (0.107104)
 37 *f2=0.154264 (0.006873)
 38 *s2=1.540947 (0.333908)
 39 *n=100 rss=27.238543 R^2=0.70012 nf=424
 40 *Correlations:
 41 *                 c      a1      f1      s1      a2      f2      s2
 42 * c           1.000 -0.243   0.054 -0.072 -0.060   0.064 -0.012
 43 * a1         -0.243  1.000   0.519 -0.192   0.509 -0.582   0.342
 44 * f1          0.054  0.519   1.000 -0.772   0.530 -0.390   0.035
 45 * s1         -0.072 -0.192  -0.772  1.000 -0.140 -0.194   0.498
 46 * a2         -0.060  0.509   0.530 -0.140   1.000 -0.589   0.412
 47 * f2          0.064 -0.582  -0.390 -0.194  -0.589  1.000 -0.894
 48 * s2         -0.012  0.342   0.035  0.498   0.412 -0.894   1.000
 49 *
```



REFERENCES

Mustonen, S. (1980). Interactive analysis in SURVO 76. *Proceedings* in *Computational Statistics*, ed.by M.M.Barritt and D.Wishart, 253-259. Physica Verlag, Wien.

Mustonen, S. (1981). Statistical computing with a text editor. *Computational Statistics*, ed.by H.Büning and P.Naeve, 327-348. Walter de Gruyter, Berlin.

Mustonen, S. (1982). Statistical computing based on text editing. *Proceedings in Computational Statistics*, ed.by H.Caussinus, P.Ettinger and R.Tomassone, 353-358. Physica Verlag, Wien.

Received 10 March 1987

Department of Statistics
University of Helsinki
Aleksanterinkatu 7
SF-00100 Helsinki
Finland