Statistical Computing Based on Text Editing

S. Mustonen, University of Helsinki, Finland

SUMMARY

Editorial mode is a new form of interactive computing enabling the user to perform mathematical, statistical and graphical operations in connection with normal text editing. Various aspects of the editorial approach are considered through examples.

Keywords: interactive analysis, statistical graphics, text processing, editorial mode, touch mode

1. INTRODUCTION

Interactivity in statistical computing is often considered almost synonymous with conversational use of the computer. Many of the friendly interactive statistical systems are controlled by a dialogue between the user and the system. Originally also SURVO 76 (Mustonen 1977,1980) had all the functions provided in conversational mode. When developing this system we soon realized that there are other forms of interactivity which on certain occasions may be still 'more interactive'.

In some areas of statistical data management and computing the conversational mode is surpassed by a new approach which we call editorial mode. Especially tasks closely related to report writing, such as editing of results, making of statistical graphs, manipulation with multiway tables, arithmetical and statistical calculations etc. seem to be very conveniently performed in editorial mode.

Our editorial approach is based on a general text editing program SURVO 76 EDITOR which is one of the SURVO 76 modules. It is planned primarily for report generating tasks in co-operation with the conversational programs.

All the operations of this editor are performed in an edit field which typically has 100-250 lines and 70-120 columns. The edit field is always partially visible on the screen, which acts like a window. The user may easily scroll that window in any direction and he may enter text, data and operations in any part of the edit field. Various data files and results may be loaded to the edit field and the contents of the field may be stored for subsequent use.

Simple text editing is done using the soft keys. These are labelled for typical word processing tasks like insertion and deletion of characters and lines, moving of the cursor etc. The leading principle in all SURVO 76 EDITOR operations is that a minimal effort, i.e. the least possible number of touches on the keyboard is needed. In this sense this editor is comparable with the most efficient text processing systems.

In our editor, however, standard text processing is merely a solid background for advanced work related to data management and to mathematical, statistical, and graphical applications.

In more complicated tasks the general idea is to enable the user to specify the job by typing some control information among the text and data and then let the system continue the work automatically. Thus the SURVO 76 EDITOR is normally in typing mode and it accepts and displays any text which is entered from the keyboard. The system may, however, be activated by pressing the key CONTINUE. What then happens, depends solely on the situation in the edit field and on the position of the cursor.

2. ARITHMETICAL AND STATISTICAL CALCULATIONS

For example, when the user has typed (as we are doing just now when preparing this manuscript)

```
p=0.837 log(p/(1-p))=_
```

and presses CONTINUE, the system responds by giving the value

```
p=0.837 log(p/(1-p))=1.6360738697
```

and it immediately returns to the typing mode thus permitting the user, for example, to round the resulting value, to alter the expression, to take a printout of specific lines or to continue typing etc.

Similarly, activation of

```
DER (X+A)↑X X_
```

by pressing CONTINUE leads to the display

```
DER (X+A)↑X X
 Derivative of (X+A)↑X with respect to X is
   (X+A)↑X*(LOG(X+A)+X/(X+A))
```

Here DER is an editing operation which produces analytic derivatives of elementary functions.

More extensive mathematical and statistical calculations may be presented as computation schemes. For example, when comparing means of two samples by doing a standard t test we could have the following situation in the edit field:

```
35    *
36    *   Comparing means of two normal populations
37    *                  size    mean      std.dev.
38    *   Sample 1:      N1=16   M1=38.5   S1=13.3
39    *   Sample 2:      N2=20   M2=30.0   S2=15.4
40    *
41    *   The test statistics is t=(M1-M2)/(S*sqr(1/N1+1/N2)),
42    *   where S=sqr((F1*S1↑2+F2*S2↑2)/(F1+F2) and F1=N1-1, F2=N2-1.
43    *
44    *   Assuming that the samples are from the same normal distribution,
45    *   the probability that t exceeds the observed value
46    *   t:=                  is 1-t.F(F1+F2,t):=
47    *   where t.F(n,t) is the cdf of the t distribution.
48    *
```

Now, if this scheme is activated by keeping the cursor either in the position after 't:=' or after '1-t.F(F1+F2,t):=' as shown above and CONTINUE is pressed, the last lines will be redisplayed in the form:

```
44    *   Assuming that the samples are from the same normal distribution,
45    *   the probability that t exceeds the observed value
46    *   t:=1.710791283   is 1-t.F(F1+F2,t):=0.0482531105
47    *   where t.F(n,t) is the cdf of the t distribution.
```

Observe, that to make a computation scheme of this kind corresponds to programming, but the form of presentation is more natural. It is like teaching a human being how to do a t test from the data presented. When the scheme has been activated the system itself has to find all the constituents needed for evaluating the expressions. To facilitate the specification of the computation schemes several mathematical and statistical functions, like 'sqr' and 't.F' above, are available. The user may also define more functions and use them in the computation schemes (Mustonen 1981a,c).

Another alternative for arithmetic operations with single numbers and numerical arrays is provided by a technique, which we call the touch mode. While working with a certain edit field the user can pass from normal typing mode to touch mode by pressing a specific key. In touch mode the cursor may be moved as in typing mode, but when the cursor touches a number in the edit field, this number may be taken into calculation simply by pressing any of the keys +,-,*,/ etc. and the expression formed and evaluated so far will be displayed on the last line of the screen. The touched numbers will be displayed in inverse video and the expression evaluated may be put to any place in the field by moving the cursor to that place and by pressing =.

In touch mode the operations defined as chains of consecutive key strokes may automatically be repeated from various starting points. This feature enables the user to perform rather complicated calculations in a very natural fashion.
For example, to compute a five term moving average from a time series in the edit field, we enter the touch mode and move the cursor to the first observation:

```
TOUCH MODE                                                SURVO 76 EDITOR
 28    *
 29    * Mean temperature (C°) in July, Helsinki 1944-1978
 30    * year      temperature
 31    * 1944      18.7
 32    * 1945      19.5
 33    * 1946      18.8
 34    * 1947      17.9
 35    * 1948      17.9
 36    * 1949      17.3
 ..    . ....      ....
```

Now the chain for computing the desired average is defined by first calculating the mean of the first 5 observations as a sequence of key strokes

```
key          comments
BEGIN        start of a definition
+            +18.7
↓            cursor one step downwards
+            +18.7+19.5
↓
+            +18.7+19.5+18.8
↓
+            +18.7+19.5+18.8+17.9
↓
+            +18.7+19.5+18.8+17.9+17.9
S            the present expression is evaluated (92.8)
C            enter a constant
5 RETURN     constant=5
/            92.8/5
↑↑→→→→→→     cursor 2 steps upwards and 7 steps to the right
=            display 18.56 (=92.8/5) in the current position
```

After these key strokes (an experienced user needs 20 seconds at most to perform them) we shall have the following display:

```
TOUCH MODE                                                SURVO 76 EDITOR
 28    *
 29    * Mean temperature (C°) in July, Helsinki 1944-1978
 30    * year      temperature
 31    * 1944      18.7
 32    * 1945      19.5
 33    * 1946      18.8    18.56
 34    * 1947      17.9
 35    * 1948      17.9
 36    * 1949      17.3
 ..    . ....      ....
```

To activate an automatic repetition, we merely move the cursor to touch the second value (19.5) and press CONTINUE. Then, the moving average will immediately be computed for all successive observations and we shall have the display

```
TOUCH MODE                                                    SURVO 76 EDITOR
 28    *
 29    * Mean temperature (C°) in July, Helsinki 1944-1978
 30    * year      temperature
 31    * 1944      18.7
 32    * 1945      19.5
 33    * 1946      18.8    18.56
 34    * 1947      17.9    18.28
 35    * 1948      17.9    17.54
 36    * 1949      17.3    16.9
 ..    . ....      ....    .....
```

3. OPERATIONS ON DATA MATRICES

Various statistical data sets may be loaded from SURVO 76 data files into the edit field. Small data sets can also be entered directly in the field. In addition to various editing operations for data matrices and multiway tables (which can be presented in a natural nested form), several statistical operations (for analysis of variance, log-linear models, linear and nonlinear regression etc.) are available.

Each data matrix to be handled by statistical operations has to be labelled with a DATA specification which indicates the lines used for the observation values and the line of the column labels. In the following display we have a small data set consisting of 12 observations and 5 variables presented in the edit field:

```
 1    *                  Consumption of various beverages
 2    *DATA COUNTRIES,A,B,C
 3    C                  Coffee  Tea  Beer   Wine Spirits
 4    A Finland            12.5 0.15  54.7   7.6  2.7
 5    * Sweden             12.9 0.30  58.3   7.9  2.9
 6    * Denmark            11.8 0.41 113.9  10.4  1.7
 7    * Norway              9.4 0.19  43.5   3.1  1.8
 8    * France              5.2 0.10  44.5 104.3  2.5
 9    * Ireland             0.2 3.73 124.5   3.8  1.9
10    * Italy               3.6 0.06  13.6 106.6  2.0
11    * Holland             9.2 0.58  75.5   9.7  2.7
12    * Portugal            2.2 0.03  27.5  89.3  0.9
13    * Switzerland         9.1 0.25  73.5  44.9  2.1
14    * Spain               2.5 0.03  43.6  73.2  2.7
15    B England             1.8 3.49 113.7   5.1  1.4
16    *                            11111
17    *SORT COUNTRIES,16_
18    *
```

To sort the data according to beer consumption, we have typed a SORT operation on line 17. SORT refers to the data set (COUNTRIES) and to an image line (16) having a mask '11111' indicating the sort key. By now activating line 17 by keeping the cursor on that line and pressing CONTINUE, the data set will be sorted and we have the display:

```
 1    *                  Consumption of various beverages
 2    *DATA COUNTRIES,A,B,C
 3    C                  Coffee  Tea  Beer   Wine Spirits
 4    A Italy              3.6 0.06  13.6 106.6  2.0
 5    * Portugal           2.2 0.03  27.5  89.3  0.9
 ..    .........           .... .... ..... .....  ...
14    * Denmark           11.8 0.41 113.9  10.4  1.7
15    B Ireland            0.2 3.73 124.5   3.8  1.9
16    *                            11111
17    *SORT COUNTRIES,16_
```

To estimate linear and nonlinear regression models, an ESTIMATE operation is available. In the following display the preceding example has been continued by typing a regression model and an ESTIMATE operation on lines 17 to 20.

```
 14    * Denmark         11.8  0.41 113.9  10.4  1.7
 15    B Ireland         0.2  3.73 124.5   3.8  1.9
 16    *
 17    *MODEL Beer1
 18    *log(Beer)=constant+coeff*log(Tea)
 19    *
 20    *ESTIMATE COUNTRIES,Beer1,21_
```

The ESTIMATE operation on line 20 refers to the data set (COUNTRIES) and to the model (Beer1) defined by the MODEL specification. The third parameter (21) is the first line for the results. When ESTIMATE is activated by pressing CONTINUE, the model 'Beer1' will at first be formally analyzed. It is converted into standard notation LOG(X(3))=A(1)+A(2)*LOG(X(4)) where A(1) and A(2) stand for those words 'constant' and 'coeff' which are not recognized as names of variables and are thus interpreted as parameters to be estimated. Also, the first and second partial derivatives of this model function will be formed analytically and this information is employed in selecting a proper computation algorithm. In this case the second derivatives vanish indicating a linear model with respect to the parameters, and so one iteration of the Newton-Raphson method (which is one of the standard alternatives) will be performed. Finally, the results are listed from line 21 onwards as follows:

```
 20    *ESTIMATE COUNTRIES,Beer1,21
 21    * constant=4.488964       (0.1565749)
 22    * coeff=0.3276288         (0.0752709)
 23    * RSS=1.553654  R↑2=0.6545
```

The estimation process may be controlled in various ways by extra specifications typed in the edit field. For example, the estimation criterion may be specified by CRITERION=Lp, where CRITERION=L2 (i.e. ordinary least squares) is default. Similarly the observations may be weighted by entering WEIGHT='weight function', where the weight function is any function of variables in the data set, etc. (Mustonen 1981b).

4. STATISTICAL GRAPHICS

A similar approach is used when making plots of data sets, curves and surfaces (Mustonen 1982). Each extra specification has a default value depending on the data and on other specifications. Thus, a PLOT operation with a few additional keywords usually creates a decent graph. Since all the information needed for a certain plot remains in the edit field, it is easy to improve the result by editing the extra specifications. An extensive inquiry facility is also available for prompt information about details.

For instance, to make a scatter diagram for 'Beer' vs 'Tea' in the preceding example the following PLOT operation, assisted by an image line (23) and one extra specification (YSCALE=LOG,...) implying a logarithmic scale on the Y axis, will be sufficient:

```
 21    * Denmark         11.8  0.41 113.9  10.4  1.7
 22    B Ireland         0.2  3.73 124.5   3.8  1.9
 23    * SSS                  YYYY XXXXX
 24    *PLOT COUNTRIES,23
 25    *YSCALE=LOG,0.01,0.02,0.05,0.1,0.2,0.5,1,2,5
 26    *
```

On the image line 23 the mask XXXXX refers to the X variable, the mask YYYY to the Y variable and SSS indicates the labels of the observations in the diagram. Thus, activation of PLOT on line 24 will produce the graph shown in Fig.1.

Similarly, to make a pie chart of the following table, a PLOT operation with two extra specifications (HEADER, TYPE) is used as follows

```
 31 *
 32 *HEADER=Number of males (1000) in various age groups
 33 *
 34 *DATA MALES,A,B,M
 35 M          0-14 15-24 25-44 45-64  65-
 36 A Sweden    841   571  1188   930   585
 37 * Denmark   564   385   731   537   308
 38 * Finland   506   399   727   468   202
 39 * Iceland    32    22    29    20    10
 40 B Norway    474   316   534   442   251
 41 *
 42 * PLOT MALES_
 43 * TYPE=%PIE
 44 *
```

and activation of PLOT on line 42 produces Fig.2.

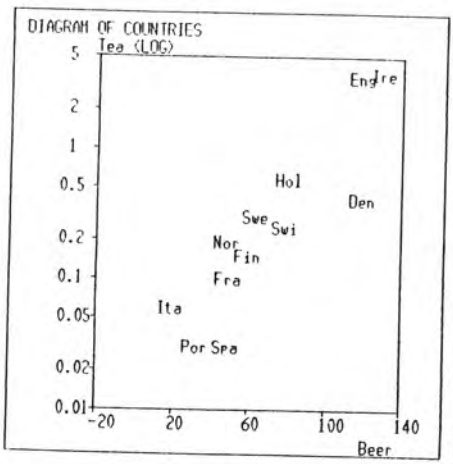

Fig.1

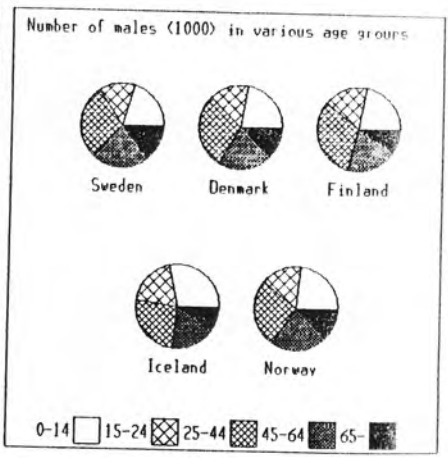

Fig.2

REFERENCES

Mustonen, S., SURVO 76, a statistical data processing system, Research Report No.6, Dept. of Statistics, University of Helsinki, 1977

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

Mustonen, S., SURVO 76 EDITOR, a new tool for interactive statistical computing, text and data management, (RELEASE 2), Research Report No.24, Dept.of Statistics, University of Helsinki, 1981a

Mustonen, S., SURVO 76 EDITOR, Estimation of regression models, Research Report No.29, Dept. of Statistics, University of Helsinki, 1981b

Mustonen, S., Statistical computing with a text editor, Computational Statistics, ed.by Herbert Büning and Peter Naeve, 327-348, Walter de Gruyter, Berlin, 1981c

Mustonen, S., Statistical graphics in SURVO 76 EDITOR, Research Report No.33, Dept. of Statistics, University of Helsinki, 1982.