Interactive analysis in SURVO 76

Mustonen, S., Helsinki, SF

Session I3/second paper
Interactive computing

SUMMARY: SURVO 76 is a statistical system covering a wide range of activities in computational statistics. The system is interactive in a real sense and no special job describing language or code is needed. In its present form SURVO 76 has been implemented on the desktop computer Wang 2200VP which provides suitable means for rapid interchange of information between the system and the user. In this paper some features of SURVO 76 related to interactive analysis are described.

KEYWORDS: interactive analysis, statistical operating systems, graphical analysis, randomization tests, text processing.

1. PRINCIPLES OF SURVO 76

In an interactive environment it is natural to expect that the system can do more than a pure statistical package. Many users like to have all the services their computer can offer within the same system frame. Thus when planning interactive programs for statistical computing there should be a tendency to move from isolated packages and individual programs towards "statistical operating systems" which besides the normal statistical data processing activities also provide various supporting features for data management and text processing.

The SURVO 76 system has an early predecessor SURVO 66 which was the first general purpose statistical package in Finland and had many of the features now common in statistical systems (Alanko, Mustonen,Tienari 1968). However, in order to achieve true interactivity, only a minor part of the properties of this first SURVO has been accepted in SURVO 76.

The new system has been intended to meet especially the needs of statisticians in both teaching and research work and its aims are slightly different from those of conventional statistical packages generally available for data analysis. In a certain sense the scope of SURVO 76 is wider permitting extended possibilities for data and text editing, simulation, matrix computations and graphical analysis.

Our main goal has been to provide suitable tools for a statistician who likes to have a quick test of his research ideas by making a computational experiment. Usually such an experiment reveals that the idea was silly, but when we learn this fact in a few minutes or hours instead of wasting several days, our whole research process will be speeded up considerably.

SURVO 76 is a rather large system consisting at present of about 60 statistical programs and subsystems (SURVO 76 modules) and the total volume is almost 1 million bytes of program text. Formally SURVO 76 is a single program written in the extended BASIC language (BASIC-2) of Wang 2200VP.

Using SURVO 76 is like discussing with the computer; we speak about SURVO 76 conversations. The discussion is transmitted from the system to the user by a CRT display and from the user to the system by a keyboard having also "soft keys" (special function keys) for various control tasks. For a more precise and detailed output a line printer, a graphic CRT and a plotter are available.

Due to interactivity a user knowing the main principles of statistical computing can learn to use SURVO 76 by just starting to use it without any detailed instructions. No programming experience is necessary in standard application of SURVO 76, but in more advanced use command of BASIC and the main construction principles of SURVO 76 are essential.

It is evident that many statisticians do not like to think in terms of computer programs. They prefer carrying out their computations and data manipulations in minor steps in the order they like. These preferences have been taken into account in the SURVO 76 system which can in many respects be operated like a desk calculator with very powerful keys.

2. SPECIAL FORMS OF INTERACTIVITY

2.1. Graphical analysis

In SURVO 76 typical statistical graphs like histograms, scatter diagrams and plots of time series combined with analytical curves and surfaces can be produced interactively with the graphic CRT and plotter. Also some special graphs like Andrews' function plots and Chernoff's faces are available.

SURVO 76 takes care of the scaling of the variables if desired and selects appropriate notations on the co-ordinate axes thus relieving the user of those nuisances. On the other hand the user has a free choice in many really important matters. For instance, when plotting scatter diagrams any nonlinear scale on the axes can be defined by entering the equation of the corresponding scale transformation.

It is essential that the user can employ various plotting modules one after another for the same picture to combine graphs. It may be useful to have, for instance, several related time series in the same picture. Likewise, after making a scatter diagram the user may estimate various models and return to plot the fitted curves on the same graph.

The graphs also have an important role in preliminary investigation of the data. In SURVO 76 interactive techniques are available for detecting outliers by graphical means. It is typical that when, for instance, a scatter diagram is displayed on the CRT the user can point at any observation with the cursor and find the name of the observation simply by pressing key "?".

In the module CORROBU, intended for robust estimation of means, standard deviations and correlations along a modification of the technique presented in Gnanadesikan (1977) the same procedure applies in the display of the Mahalanobis' distance distribution. In addition, the user can point at the rejection treshold for the outliers with the cursor. Using this interactive technique iteratively we have reached promising results.

In an interactive environment it is possible to revive techniques which have been difficult to computerize before. The problem of rotation in factor analysis is a good example. When the rotation is carried out with a computer without the possibility of instant graphical displays the criteria for suitable rotation have to be modified to a blind analytic form. Many analytic rotation programs give good results in standard applications, but they are rather insensible to the special needs of the user. In our system the factor rotations are performed graphically and stepwise on the CRT, but the user can also employ some analytic criteria as advice for each step.

2.2. Matrix operations

In many desk computers various arithmetic operations can be performed and results displayed just by operating the machine like a normal calculator. To a certain extent this also applies to matrix computations.

We feel, however, that these operations as such are not sophisticated enough for the multifarious computational needs of statisticians. It is often desirable to have an opportunity to continue certain computations manually after the standard routines have been performed. For this purpose SURVO 76 contains a special subsystem called MATRI.

With MATRI the typical matrix operations needed in statistics can be performed using the computer like a calculator. In MATRI the "soft keys" are defined for various matrix operations. The matrices required as an input can be keyed in manually (usually by filling a form with proper dimensions and labels on the CRT) or transferred from different SURVO 76 files. Results can be saved in special matrix files for later operations.

An essential feature of MATRI is that it does a lot of bookkeeping and labels each result with a name corresponding to the ordinary matrix notation. Also the columns and rows in matrices can be labelled with names and these names will be moved in MATRI operations along certain rules.

The user can also define extra operations and make simple matrix programs (MATRI chains) by just carrying out a sequence of matrix operations and this sequence can be repeated automatically with other input matrices. These MATRI chains can be saved on disk and used in connection with other MATRI operations when needed.

2.3. Random data simulation

In methodological considerations and teaching situations it is useful to analyze artificial random data whose origin is perfectly known. The planning of such experiments can be substantially facilitated by employing the module CHANCE which is a random data generator.

Several subroutines are immediately available to generate pseudo random variates from various distributions. Thus it is

easy to construct random data according to a given statistical model. The simulated files can subsequently be treated as ordinary data files in SURVO 76.

Using CHANCE the behaviour of different sample distributions can also be demonstrated on the CRT. The user selects the distribution and its parameters and CHANCE starts to generate and plot observations on the CRT one after another as a constantly growing histogram.

2.4. Testing of statistical hypotheses

As an example on the use of interactivity in simple statistical inference let us consider the technique used in the SURVO 76 module TABTEST. A typical display on the CRT during a TABTEST run is the following:

```
FREQUENCY TABLE:  N= 12
  0  1  3  2
  4  2  0  0
X↑2=   9.33  DF=  3   P=0.02489 (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 user has started this job by entering 2 samples of 6 observations in the form of a 2x4 frequency table and the goal of this analysis is to decide whether these samples are from the same population. For this purpose TABTEST has computed the common X^2 -value 9.33 and indicates that its critical level is P=0.02489 according to the chi-squared approximation. We know, however, that in case of few observations this approximation may be rather poor and the exact distribution of X^2-statistic should be used instead.

Nowadays it is typical to construct tables for complicated tests by numerical methods and simulation. Here, however, we are using simulation in a slightly different way.

TABTEST does not consult any ready made tables, but tries to find the true critical level just for the case presented. After

the user has specified the null hypothesis (here CASE 2: ONLY ROW TOTALS FIXED) TABTEST immediately starts to estimate the critical level by generating random samples according to the null hypothesis, forms the corresponding tables, computes the X^2-value and the proportion of those tables for which X^2 exceeds the value 9.33 in our case. This proportion P will then approximate the true critical level. The underlined numbers in the display are changing during the simulation experiment and the user can watch the process as long he likes. Since P is approximately normal with mean equal to the true critical value, TABTEST displays also the probability for this estimate to go below the nearest standard level (1% in this case).

Usually it is not necessary to know the exact P-value, but a crude approximation is sufficient for practical purposes. Here it takes only a few seconds to obtain the display above and it reveals that the original chi-squared approximation seems to be rather conservative.

In SURVO 76 this "instant simulation" approach has been used for various nonparametric tests and even Fisher's randomization principle becomes applicable for quite reasonable sample sizes. For instance, the SURVO 76 module COMPARE includes the Fisher-Pitman randomization test for comparing two independent samples. (For the definition of this test see, for instance, Conover 1971, pp.357-364). The exhaustive enumeration of critical combinations needed for the traditional approach is formidable already for sample sizes 15 and 20, but "instant simulation" usually gives satisfactory results without any delay.

2.5. SURVO 76 and text processing

It is quite common that when writing a research report containing numerical tables the output from the computer cannot be used as such, but the results have to be retyped manually. This may happen even if the computer output is well designed, since the needs of the user may change during the reporting phase. In an interactive environment a good way of avoiding those editorial problems is to have text processing facilities in connection with the statistical operating system.

As an extensive new option in SURVO 76 we have developed an

editor module. It can be used not only for normal text processing purposes, but also for input of data in unformatted form, for transferring data into SURVO 76 files and for editing SURVO 76 files and results together with normal text using powerful editing operations. These operations are, for example:

-to make up the text to a certain line length,
-to transform and edit numeric tables,
(new columns and rows can be inserted also using numeric transformations),
-to numeric and alphanumeric sorting of data,
-to print out selected parts of the text on the printer.

All the information is represented in an 'edit field' which consists, for example, of 100 columns and 250 rows. The field is always partially visible on the CRT. The editing operations are also typed in this field and they can be treated as normal text. Any operation can be activated by moving the cursor to the corresponding line and by pressing key CONTINUE. Whenever needed the contents of the edit field (tables, text and operations) can be saved in an edit file.

It seems quite natural to extend editing operations towards normal statistical operations and this will be a new form of interactive statistical computing which covers the final documentation as well.

REFERENCES:

Alanko T.,Mustonen S.,Tienari M.(1968), A statistical programming language SURVO 66, BIT 8,69-85.
Conover W.J.(1971), Practical Nonparametric Statistics, John Wiley, New York.
Gnanadesikan R.(1977), Statistical Data Analysis of Multivariate Observations, John Wiley, New York.
Mustonen S.(1977), SURVO 76, A statistical data processing system, Research report No.6, Dept.of Statistics, University of Helsinki.
Mustonen S., Mellin I.(1980), SURVO 76 program descriptions, Dept.of Statistics, University of Helsinki.