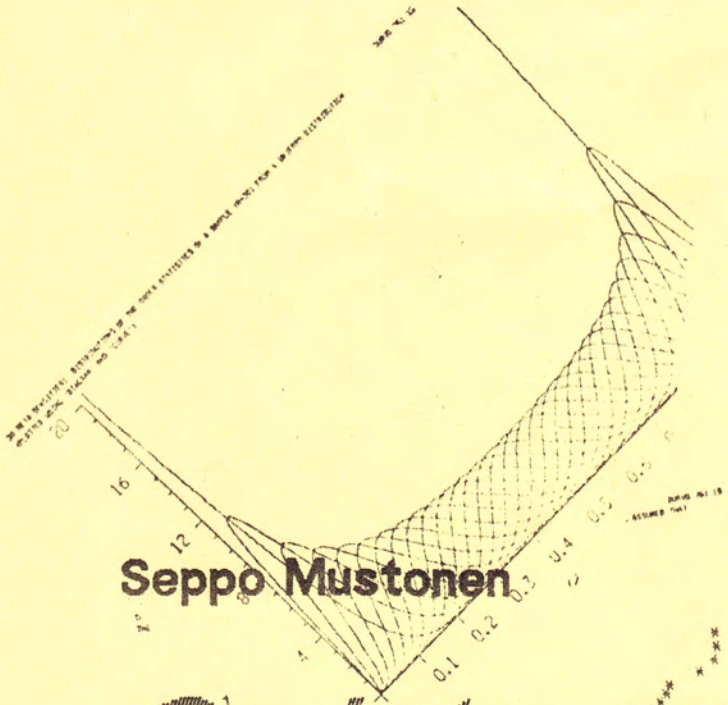
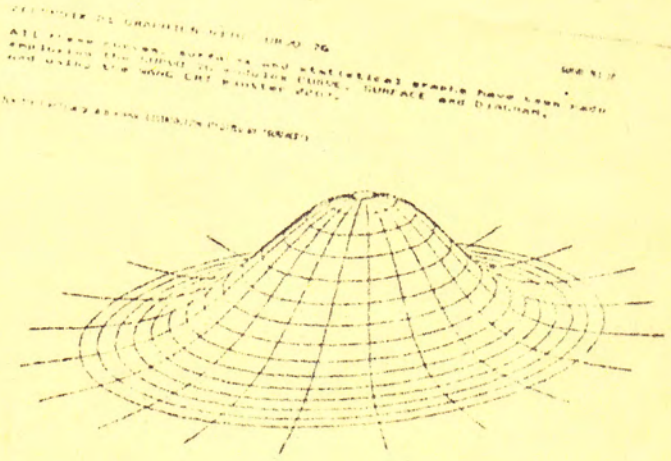
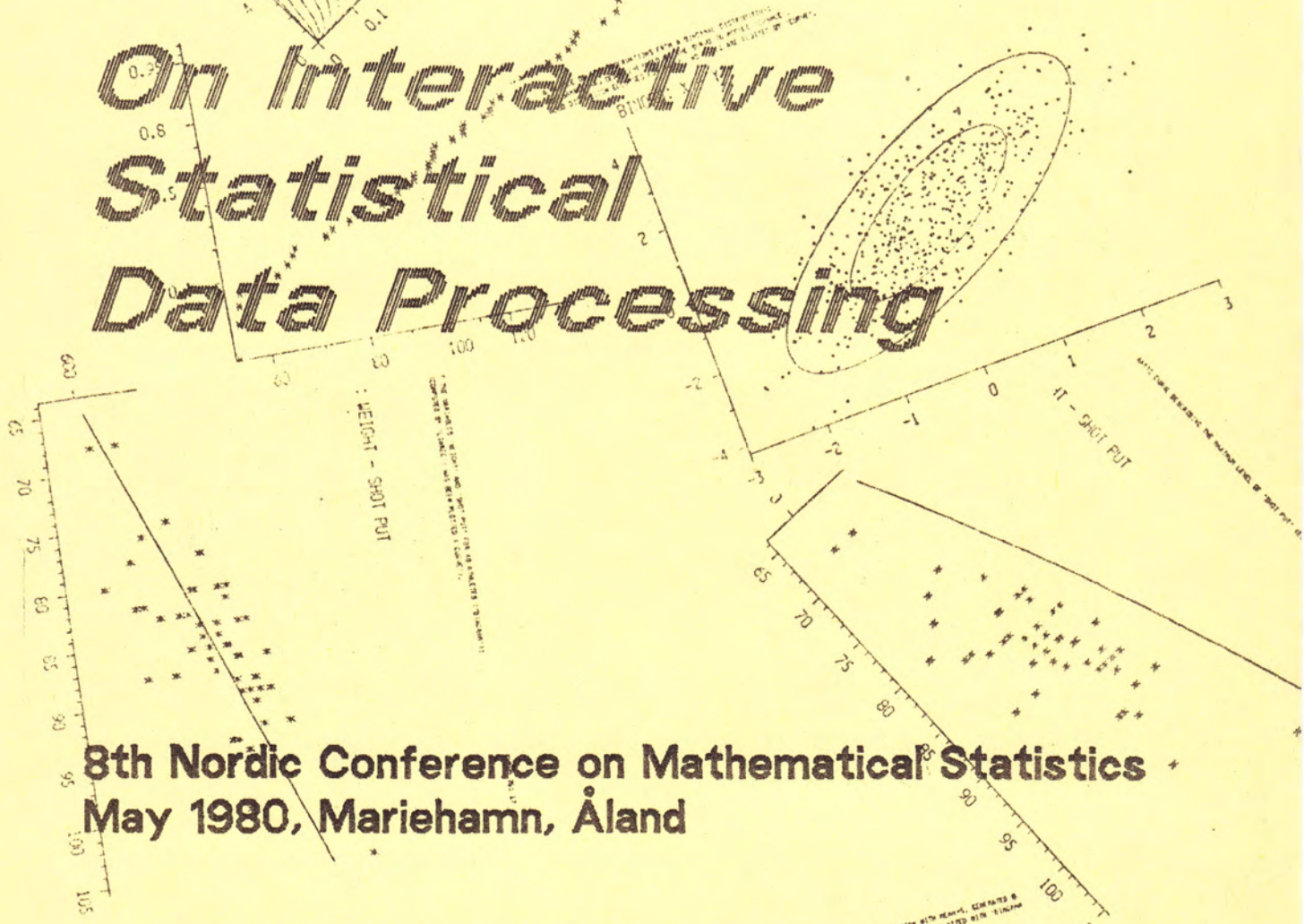


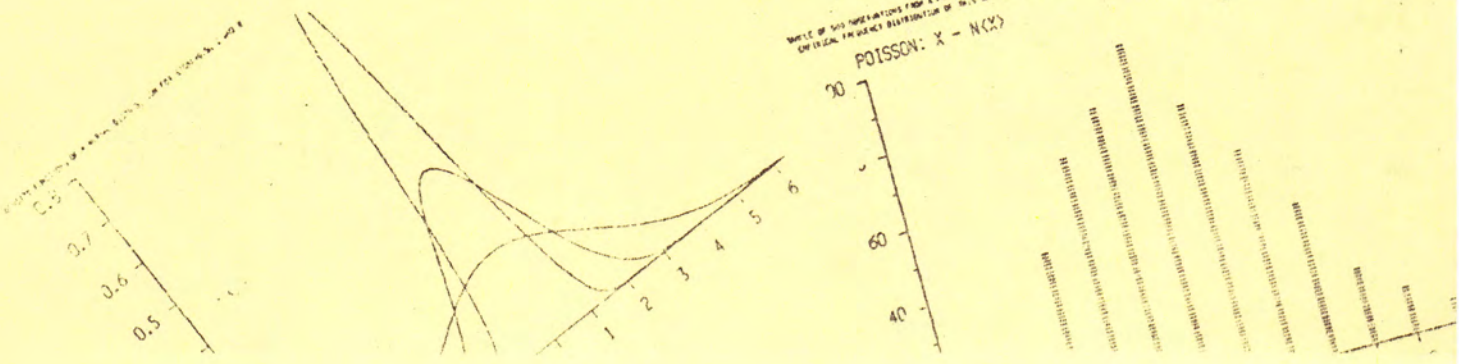


Seppo Mustonen

On Interactive Statistical Data Processing



8th Nordic Conference on Mathematical Statistics
May 1980, Mariehamn, Åland



S.Mustonen, Department of Statistics, University of Helsinki

ON INTERACTIVE STATISTICAL DATA PROCESSING

1. Introduction

The impact of automatic data processing has in recent years been enormous also in the field of statistics. In statistical research it is not enough to solve purely theoretical problems. Clear computational results are equally important.

Although complicated mathematical models are applied in statistical analysis it does not imply that the problems of statistical data processing are only mathematical and related to statistical theory and numerical analysis. In statistical computing the knowledge of various fields of computer science and system analysis is essential as well.

The major efforts in statistical data processing have been concerned with the problems of data analysis. There are many large collections of programs available for the purpose.

We feel, however, that in spite of these ingenious program packages many statisticians are not quite satisfied with the present situation.

There are several reasons for this dissatisfaction:

1)Many of the program collections are really like "canned packages"; they are simple to use in standard applications, but it is almost impossible to see what is really inside this "package" and how to open it. Thus it is difficult to study the internal structure of the programs and make alterations whenever needed. The rigidity of these packages also restricts their use for teaching purposes.

2)Many statistical programs are often too automatic or they are automatic in the wrong places. After entering some initial information into the computer the user cannot but wait for the final results without any possibility to intervene. Thus even when there is a slight error in the initial information the whole process goes through and must then be restarted. It is also typical, for instance, that a program for linear regression analysis selects the regressors automatically, but when the residuals are to be plotted the user has to declare each tiny detail in order to have an decent graph. So in the worst cases the roles of the statistician and the computer have changed and the user seems to be controlled by the system and not vice versa.

3)There are situations where a statistical program may be quite satisfactory, but everything is spoiled by an inadequate operating system. For instance, in a time sharing environment strongly varying response times of the computer system may totally ruin a well designed interactive approach.

4)Many statistical packages are good for their special task, but they are too restrictive. A multistage research process cannot be carried out as a whole, but some steps in the process must be done by other means. Working with several badly synchronized programs may be very frustrating.

5)It is common when writing a report containing numerical tables that the computer printout 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. Only when the statistical system includes text editing facilities this offers no problems.

In general, there should be a tendency to move away from isolated packages and individual programs towards statistical operating systems which cover all the activities in the field of statistical computing in a unified form. A statistical operating system can be considered an enlargement of a normal operating system having the typical statistical operations among its constituents. In this way the user has total support from the computer to the various needs in statistical computing.

There have also been proposals for statistical programming langua-

ges. Special languages and codes are useful in restricted areas of statistical computing (as in simulation). In general, however, it is hardly possible to proceed in this direction, since there do not exist simple ways of expressing statistical operations in unified form.

One practical solution is an interactive statistical operating system where a natural language like English forms the essential link between the system and the statistician.

Some statisticians and computer specialists seem to be rather suspicious of the possibilities of interactive computing. There prevails, however, a strong agreement about the merits of interactivity in exploratory data analysis especially with small data sets. But when working with large samples and using more sophisticated techniques the opinions are shared. For instance, Nelder (1978) says that "For larger problems interactive working may be less important, because the response time of the user to his intermediate results becomes the limiting factor in the analytical process". This is an interesting statement, since usually people tell about the experiences that it is the response time of the computer which actually is limiting interactive working.

We feel, however, that, in principle, there are no restrictions in using interactive approach to all kinds of problems. In practice the limit of profitableness is continuously moving in favour of interactivity.

Although everything is not yet so paying at this stage with interactive means it is worthwhile to study this alternative also in more complicated tasks, since the rules of making good interactive software are not the same as in batch processing and it takes time to learn this new attitude. In recent years, it has been quite common to modify existing program packages into more interactive form, but we feel that this is not the best way to proceed. True interactivity needs and deserves another starting point.

This does not imply that in the future everything should be solved by interactive systems. The real needs, tastes and working habits of statisticians are extremely varying. So it is impossible to think that the progress will lead to some unique solution. We must have continuously several alternatives for different tasks.

In the Department of Statistics at the University of Helsinki we have studied various forms of interactive computing. Therefore this presentation will be devoted mainly to the direction we have chosen.

2. Principles of SURVO 76

In order to give a more precise account of the possibilities of an interactive statistical operating system we shall describe SURVO 76 which has been developed for the small desk top computer Wang 2200.

This 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,Tienari,Mustonen 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.

It cannot be claimed that SURVO 76 is a statistical operating system in the true sense, since it is not a part of the basic operating system of the computer. We think, however, that many important aspects of such a statistical system can be illustrated using SURVO 76 as an example. In SURVO 76 we have tried to test various approaches covering a wide range of activities in statistical computing as a "laboratory experiment" in order to learn more about the rules of interactive work.

The SURVO 76 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 at present a rather large system consisting 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.

This may be a surprise for those who have been told that BASIC is a elementary language meant for simple tasks and short programs only. This is of course true for the original BASIC, but the various extensions in BASIC-2 have removed many of the drawbacks and there are no severe obstacles for making large programs. Even in this extended form BASIC is lacking many features existing in more sophisticated languages, but they need not be so important. We have a feeling that the importance of the programming language can be exaggereted by "computer specialists" who do not actually know the practical needs of programmers. Discussion about the relative merits of various languages in statistical computing seems often to be on a wrong basis.

An interpretative language like BASIC is, of course, inefficient with respect to computing time, but in an interactive mode of working this is seldom a real harm. On the other hand the possibility to make alterations in the programs rapidly without extra system commands and program compiling improves and speeds up both advanced use of the system and system development. We believe that the future technical progress will still increase the relative merits of the interpretative languages in statistical computing.

Portability (i.e. the possibility to use a program package easily in different machines) is another feature which has been emphasized when evaluating statistical programs. It is easy to agree, but we think again that even this property has been exaggereted. The truth is still at the moment that when one likes to create a universal solution working in all important computers very little can be realized without huge extra labor, time and costs.

It is also too restrictive to think in terms of an intersection of all the available alternatives. If we like to make progress in the area of statistical computing we must start from rather specialized computers having properties which we hope are common in the nearest future. It is the ideas which are portable and the computers which should be portable.

SURVO 76 is an interactive system and no special job describing language or code is needed. Using this system 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 (speed is almost 5000 characters/sec.) and from the user to the system by a keyboard having also "soft keys" for various control tasks.

For a more precise and detailed output a line printer, a graphic CRT and/or a plotter are available.

The possibility for rapid interchange of information between the user and the system is one cornerstone in a true interactive statistical system. It is also important that this property has been adopted in such a way that the user can instantly reach any part of the data to be analyzed for inspection. Equally important is a rapid access to the different modules of the statistical system to get an idea of how the system works and to make temporary modifications and enlargements to the modules.

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 main construction principles of SURVO 76 is essential.

Even interactive systems are sometimes frustrating since they may in their own gentle way compel the user to a long unproductive conversation without a natural exit. In SURVO 76 this dependence is avoided by splitting the programs into a lot of small modules. When the user becomes exhausted with a certain module he can interrupt the conversation and call any of the neighbouring modules by pressing one single key on the keyboard, without losing contact with the previous stages of the job.

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 consideration in the SURVO 76 system which can in many respects be operated like a desk calculator with very powerful keys.

On the Wang 2200 keyboard there are 32 special function keys (denoted by F0,F1,...,F31) which can be defined as starting points for different parts of the program. In SURVO 76 the functions of these "soft keys" vary depending on the module in use. The user not knowing which F-key to press next, can always resort to key F0 which in SURVO 76 displays on the CRT the functions of other F-keys operative in the present situation.

Each F-start leads typically to a sequence of questions made by the system and these have to be answered by the user. The whole dialogue is displayed on the screen and this procedure allows the system to give the user many comments and hints relevant in the context without any waste of time and paper.

In order to speed up the conversation SURVO 76 itself volunteers with a suggestion for an answer which is displayed after the question. To give reasonable suggestions SURVO 76 tries to remember the previous actions of the user or even to guess what he might attempt next. If the user agrees with the suggestion of SURVO 76 it is enough to press the RETURN key. Otherwise he must type his own answer.

Each interchange of questions and answers leads eventually to a series of different actions and computations. The results are printed on the CRT. When the computations are finished the user can select another F-start or another module. Certain F-starts are reserved for moving the results just obtained from the screen to the printer or for saving them on disk as intermediate results for subsequent analysis with other modules.

The modules performing various statistical analyses can co-operate and use the same original data files or intermediate results without any modifications whenever this is statistically reasonable.

Each statistical method in SURVO 76 has been split into small submodules and the various computations and data manipulations can be carried out by combining the corresponding F-starts properly.

Hence it is the user's responsibility to make good choices. It would, of course, be easy to connect different submodules in a fixed "right" order, but then the user would be at the mercy of the system which is the undesirable feature of some statistical packages.

The possibility to select different combinations of actions quite freely means that the user can employ the system in a creative manner and not only by repeating traditional computation chains. It also means that the user must know in advance a great deal of the method he likes to use, but not much of data processing in general. We think that easy use in connection with statistical programs must not imply that they could be used "easily" without any knowledge of statistics. There are nowadays plenty of rigid "automatic" statistical programs which can be mechanically operated by anybody, but this at the same time is a source for uncritical application of statistical methods.

3. Data management and analysis in SURVO 76

An ideal configuration for SURVO 76 is at the moment a Wang 2200VP having a central processing unit with a memory of at least 32K, a CRT display 24x80, a dual floppy disk drive, a printer, a graphic CRT and a plotter. Observe that the the BASIC-2 interpreter and the operating system are in a separate control memory of ca. 50K.

When the SURVO 76 system is in use one of the disk drives is reserved for the SURVO 76 program disks and another is for the user's data and possible additional programs. Any of the disks can be changed in a few seconds whenever necessary.

The system consists of a central module and various statistical and special modules, one of which at a time can be in use together with the central module. The central module takes care of the co-operation between the different statistical modules and it contains system subroutines, e.g. for data transfers between the central and the disk memory. Thus the user needs never worry about the location of the data during the computations.

The number of SURVO 76 modules is not in any way limited. New modules for simple data analysis can be generated even in an interactive mode by consulting a half prepared module FRAME. Employing FRAME to build up a new module guarantees that the module will be compatible with the requirements of the SURVO 76 system.

SURVO 76 contains several modules for statistical data analysis. When beginning to develop the system the most traditional and elementary forms of analysis were emphasized and they gave a natural basis for the the system. Now the development has been directed towards more sophisticated and computationally demanding methods.

The system includes modules, e.g. for following activities:

-basic statistics,
-frequency distributions and tables,
-data sorting, order statistics,
-statistical tests and tables,
-linear and nonlinear regression analysis,
-multivariate methods,
-cluster analysis
-time series analysis.

Several non-standard methods are also available. Samples with missing values can be treated and techniques for detecting outliers and for robust estimation are included.

The problems of data input, editing and transformation have received special attention. There are standard modules to cover the activities in this field and they make the system self-contained. The newest contribution to data management in SURVO 76 is a general purpose editing program. It is connected to the statistical modules and makes possible text editing and various report generating activities with numeric and alphanumeric data and results.

One of the basic principles in SURVO 76 is that any potentially important observations and intermediate results can be used in subsequent computations without extra modifications of the system and the data. We thus have uniform representations for various data structures.

SURVO 76 allows both variables and observations to be labelled with alphanumeric names. This makes the results more readable and the monitoring of the computations easier. Each module is supposed to record continuously on the CRT what it is doing. For example, when observations are processed the system displays the names of the observations.

It is not necessary that the user has time to read all that is shown on the CRT; usually a crude impression is enough for monitoring. But when something unexpected seems to happen it is possible to stop the information flow on the screen and see what really is going on. If

necessary the output rate can be slowed down to a normal reading level.

4. Special forms of interactivity

Some interactive approaches used in the SURVO 76 system will now be described, although we know that it is rather difficult to explain these dynamic properties without actual working with the system.

4.1. 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. Special graphs like Andrews' function plots and Chernoff's faces are also 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 or by selecting it from certain standard alternatives. For example, various probability papers may be specified in this way.

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 the 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 "?".

The same search procedure applies in the display of the Mahalanobis' distance distribution when using the module CORROBU, intended for robust estimation of means, standard deviations and correlations along a modification of the technique presented in Gnanadesikan (1977). 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.

4.2. 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 standard 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. The columns and rows in matrices can also 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.

4.3. Random data simulation

In methodological work and in 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.

The user has to type the statements needed to generate a typical observation according to the advice given by CHANCE. For this task, 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.

4.4. Testing of statistical hypotheses

As an example of 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=0.0249 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 as 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 delay.

4.5. Program modifications in advanced use

Interactivity offers many benefits for those users who like to modify existing programs temporarily for their special tasks. When the programming language is interpretative this is especially profitable, since alterations can be made as a part of the conversation even when running the program.

In SURVO 76 this approach is already adopted in some standard operations. For instance, specification of new transformed variables is carried out by inserting the transformation statements in the program according to instructions given by the system. Although this procedure presupposes rudimentary programming skills we have found it powerful compared with the normal convention of presenting lists or codes for specific standard alternatives.

In some other activities in SURVO 76 we do have such a list, but at the same time there is an option for a general user-defined approach. For example, in the module HISTO for plotting histograms and fitting univariate frequency distributions by theoretical models, the theoretical distribution can be selected among 8 alternatives or defined by the user quite freely by entering the equation of the corresponding density. (In fact, the kernel of the density up to a constant factor is sufficient, since HISTO takes care of scaling the integral to 1). The density function may include unknown parameters and before the fitted density is plotted on the histogram and the goodness-of-fit tests are performed, these parameters will be automatically estimated by HISTO using the maximum likelihood method. This procedure has proved to be useful even in estimating truncated and mixed distributions.

4.6. Text processing in connection with data analysis

It has earlier been pointed out that it may be frustrating for a statistician to retype the computer output manually to reach a form suitable for final reporting. We can, of course, have highly specialized systems for text processing, but usually they are not directly connected to statistical programs.

To lessen the burden for a statistician in the report writing stage we have tried to develop an editor program as an integrated part of our system. This editor can be used not only for normal text processing purposes, but also for input of data in an unformatted form, for

transferring data into SURVO 76 files and for editing SURVO 76 files and results together with normal text by using powerful editing operations. These operations are for instance:

-to make up the text to a certain line length,
-to transform and edit numeric tables
(new columns and rows can also be inserted by 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.

4.7. Documentation

Documentation is not only important for the results of a statistical analysis; it is equally important for the statistical programs, since a program without a decent description is often rather worthless.

In interactive systems the program text itself contains so much information concerning the discussion with the user that mere lists of the programs are helpful. Thus a user knowing the main constructing principles of SURVO 76 can find much information just by listing parts of programs on the CRT or on paper. In addition, non-standard activities will be declared to the user during the conversation. For some more comprehensive topics special interactive teaching programs are included.

It is assumed that in ambivalent situations the user has courage to find his way by trial and error. SURVO 76 is not an easy system in that it does everything automatically for the user. On the contrary, it assumes that the statistician makes his own decisions and takes initiatives. On the other hand, this type of system offers information and suggestions to support the decisions. There are statisticians who love to work on this basis, but there are also people who find it difficult or too vague.

Although we have normal program descriptions of SURVO 76, they cannot tell all essentials, since paper is too rigid a medium for the dynamic aspects. Therefore we have tried to compose automatic demonstration programs which contain ready made SURVO 76 conversations between the system and a fictitious user. The user can watch these conversations like a TV program, but he can also break the conversation and continue in his own fashion.

This dynamic documentation approach seems to be fruitful also in teaching statistical methods. In theoretical and applied research work this type of documentation will obviously be of considerable support and it could even offer an alternative to a traditional research paper.

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.

Nelder J.A.(1978), The future of statistical software, Proceedings in Computational Statistics, Physica-Verlag, Wien

APPENDICES:

1. Graphics with SURVO 76
2. List of SURVO 76 modules

Fig.1 Density function of a two-dimensional normal distribution (plotted by module SURFACE)

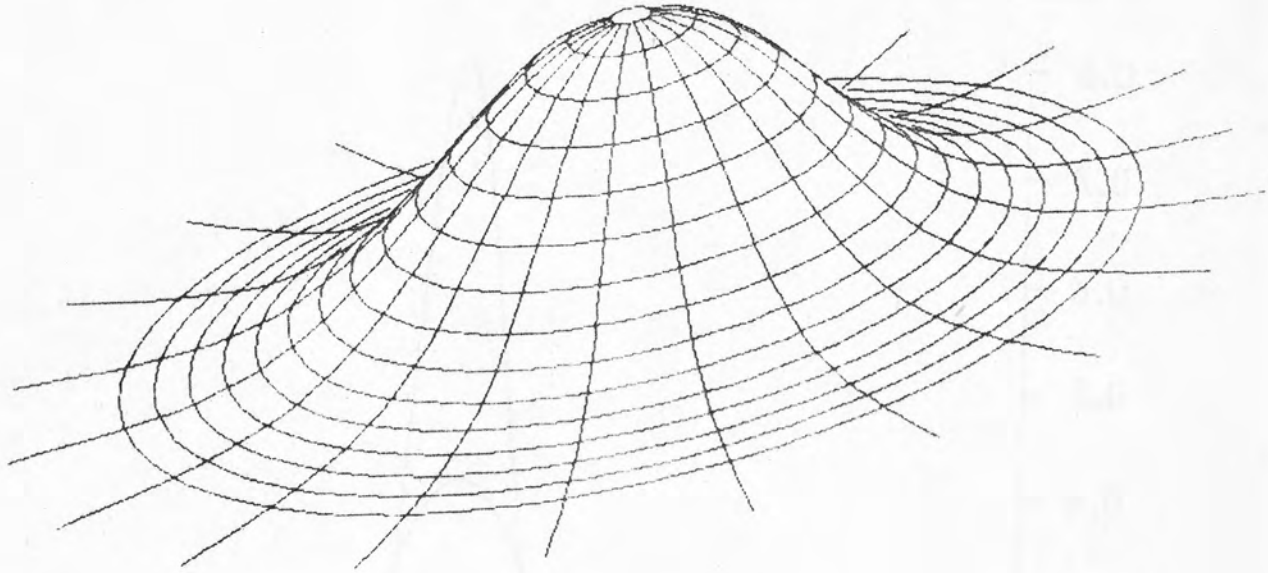


Fig.2 A sample of 500 observations from a two-dimensional normal distribution
(sample generated by CHANCE and plotted by DIAGRAM)
Confidence ellipses for p=0.5 and p=0.1 are plotted by CURVE

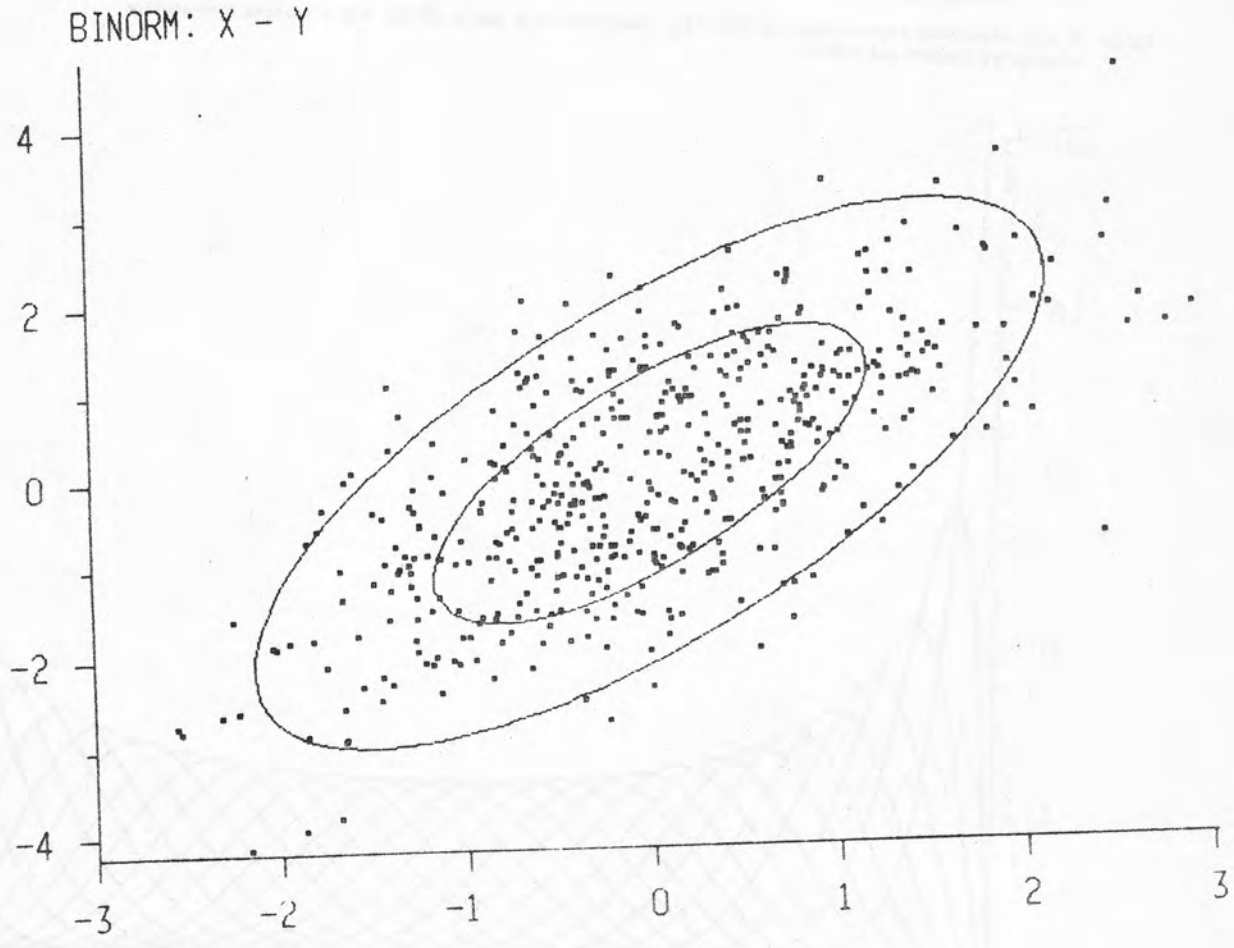


Fig.3 Density functions of a normal distribution for sigma=0.5,1,2
(plotted by DIAGRAM and CURVE)

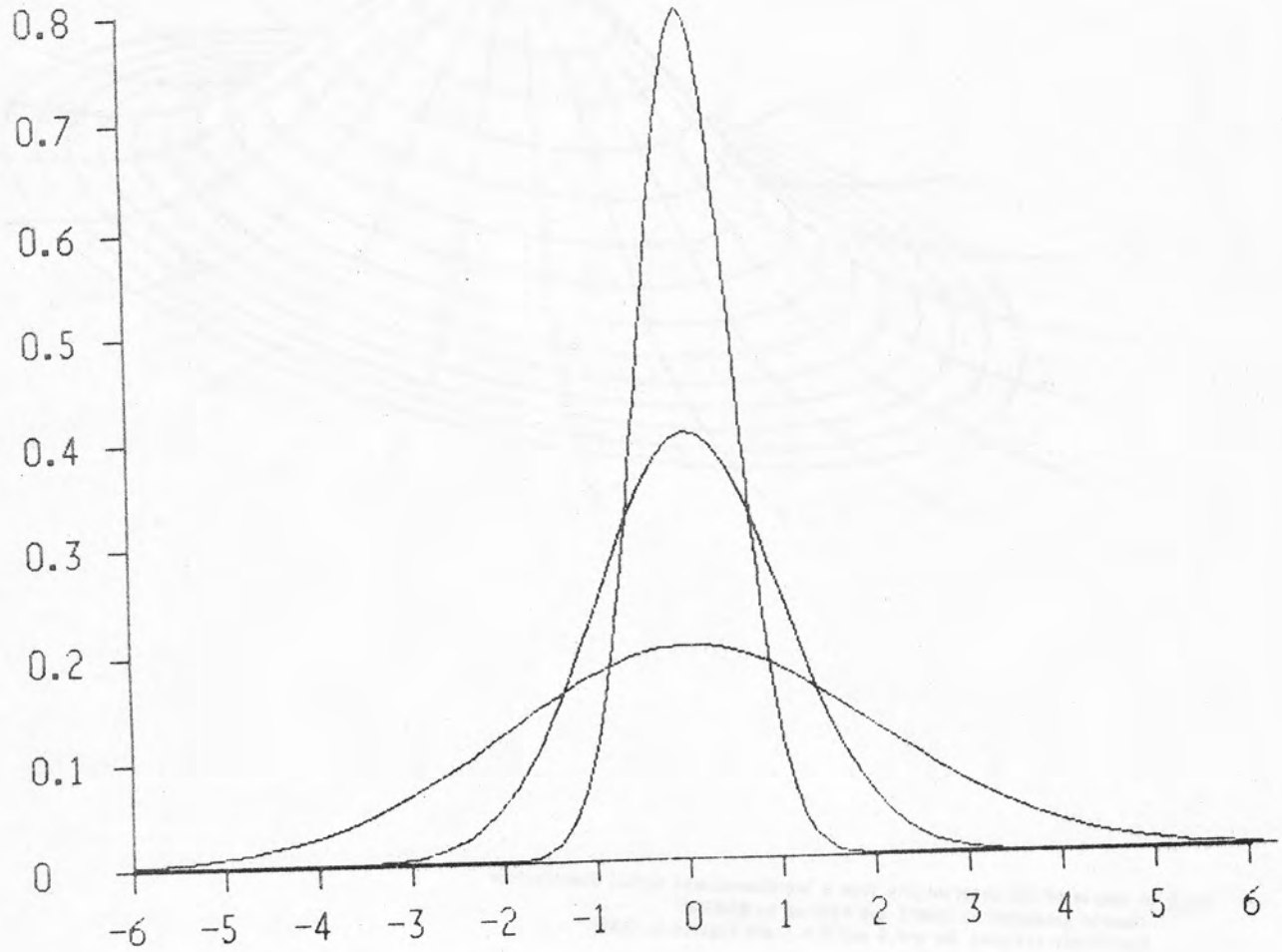


Fig.4 30 beta densities: Distributions of the order statistics of a sample (N=30) from a uniform distribution
(plotted by DIAGRAM and CURVE)

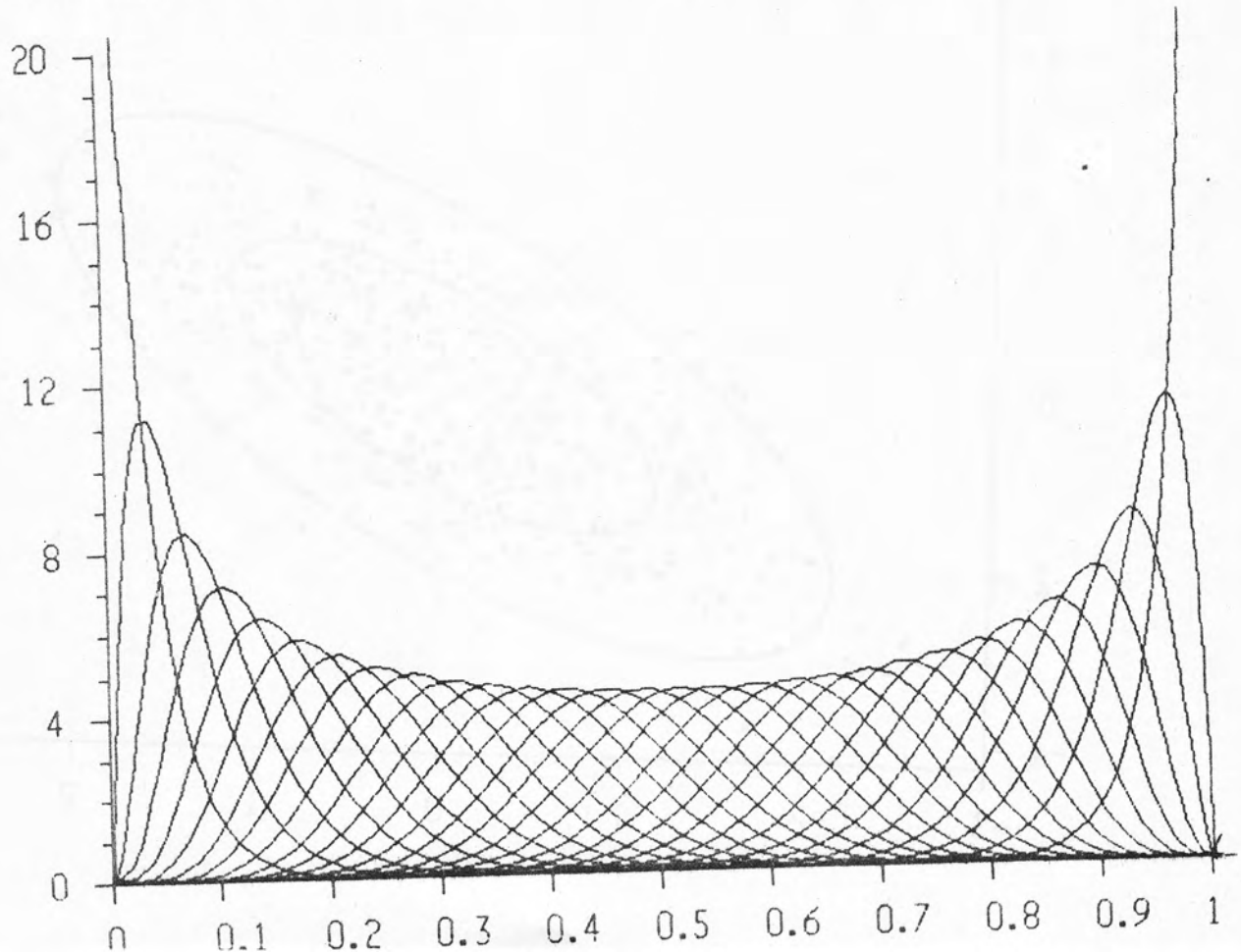


Fig.5 A histogram with a fitted normal distribution (plotted by HISTO)

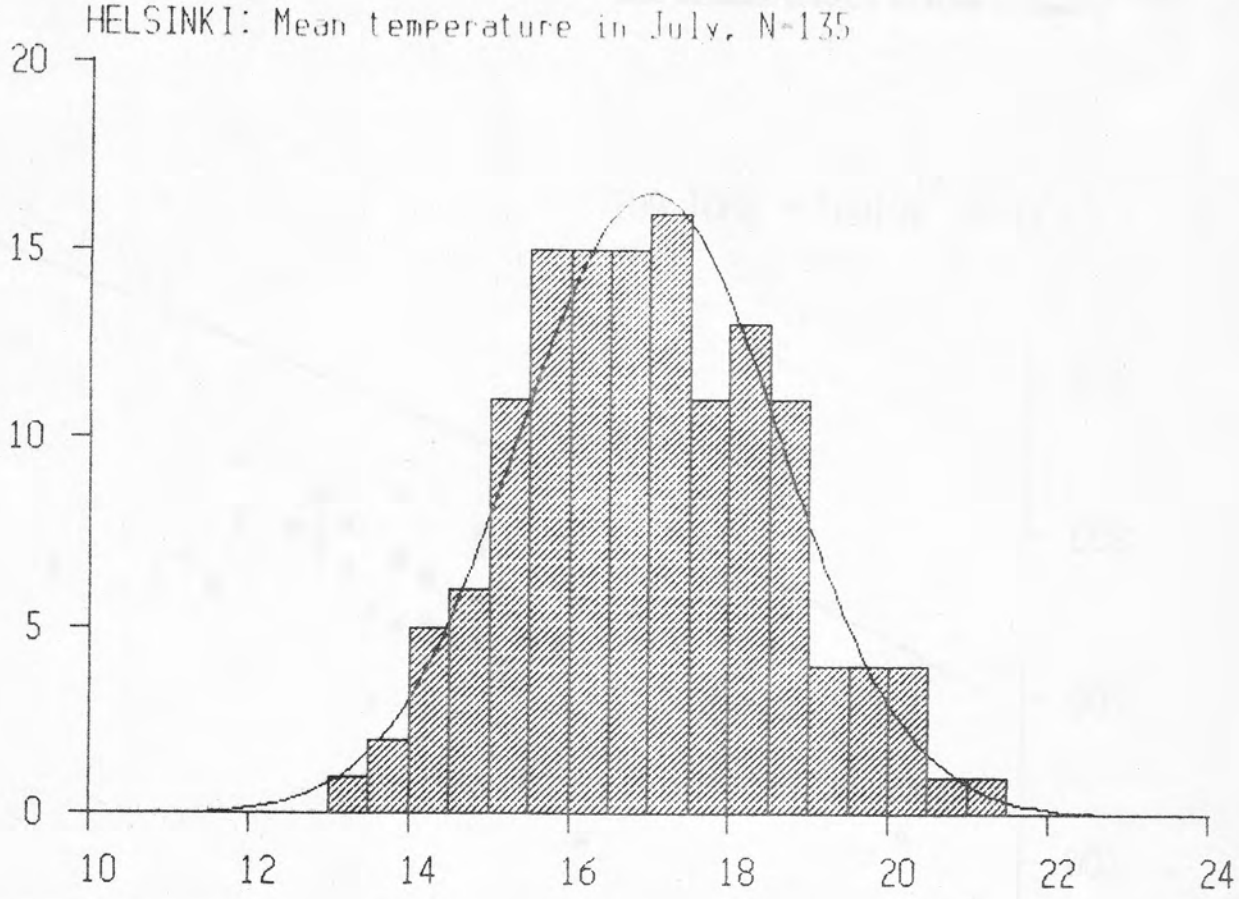


Fig.6 A correlation diagram of the weight and the result of shot put for 48 athletes (DIAGRAM)
Also a regression line (computed by LINREG) is plotted (CURVE)

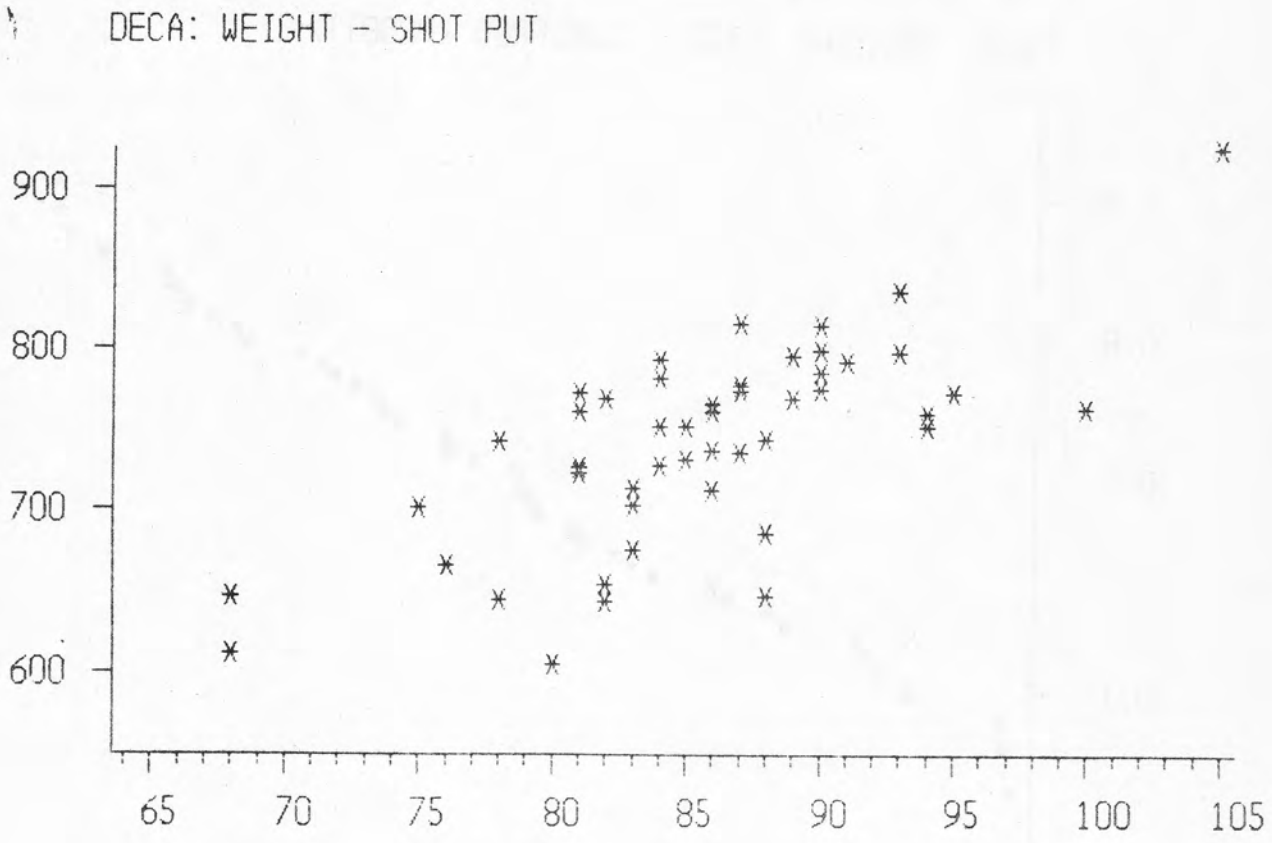


Fig.7 The same correlation diagram, but now with a quadratic curve for the maximum level in shot put (estimated by NONLIN and plotted by DIAGRAM and CURVE)

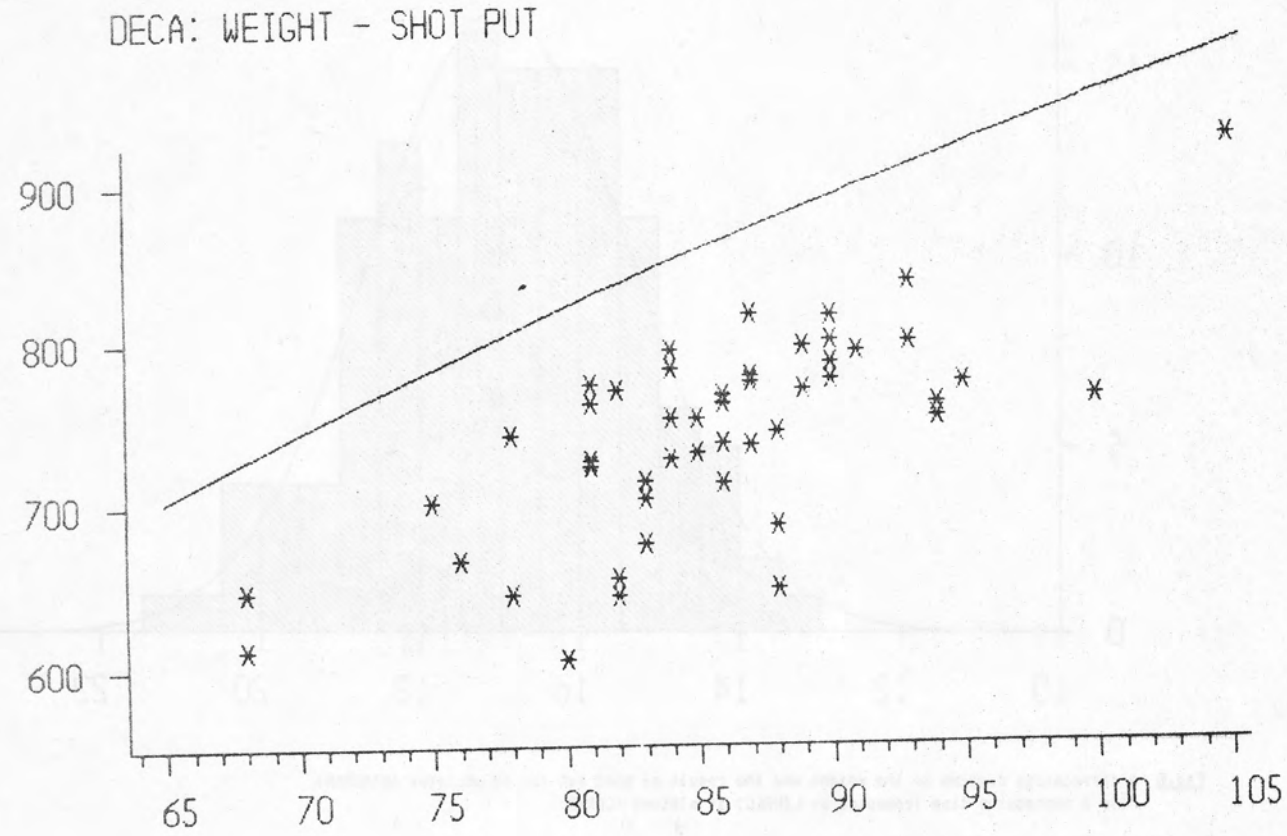


Fig.8 In the previous display it was assumed in estimation of the maximum level that the residuals of the model have a lognormal distribution.
Estimated residuals are now plotted on a lognormal probability paper (DIAGRAM)

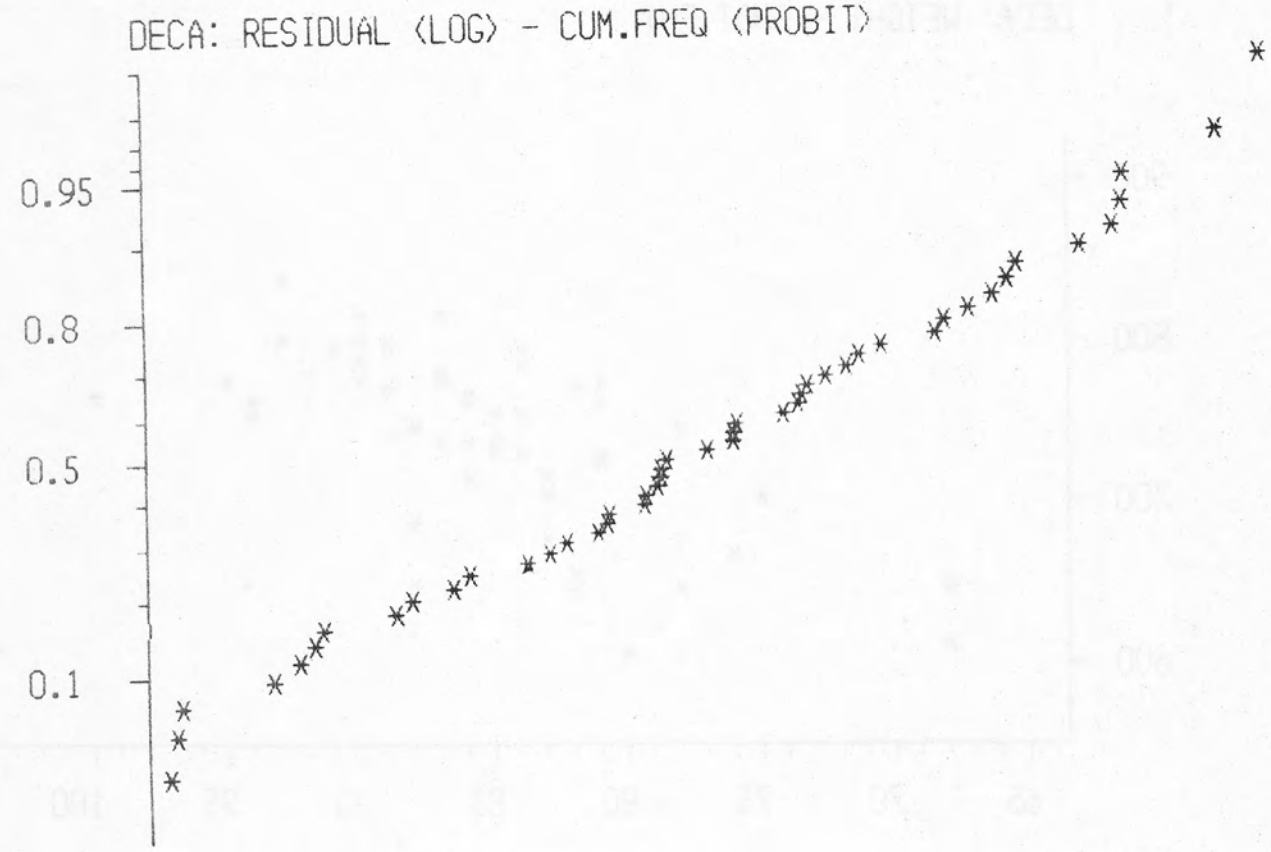


Fig.9 A coin has been tossed 500 times and the relative frequency of heads N(H)/N is recorded (CHANCE). Now the stochastic convergence of N(H)/N to 1/2 is illustrated in the following graph plotted by CURVE. (A logarithmic scale for N is employed)

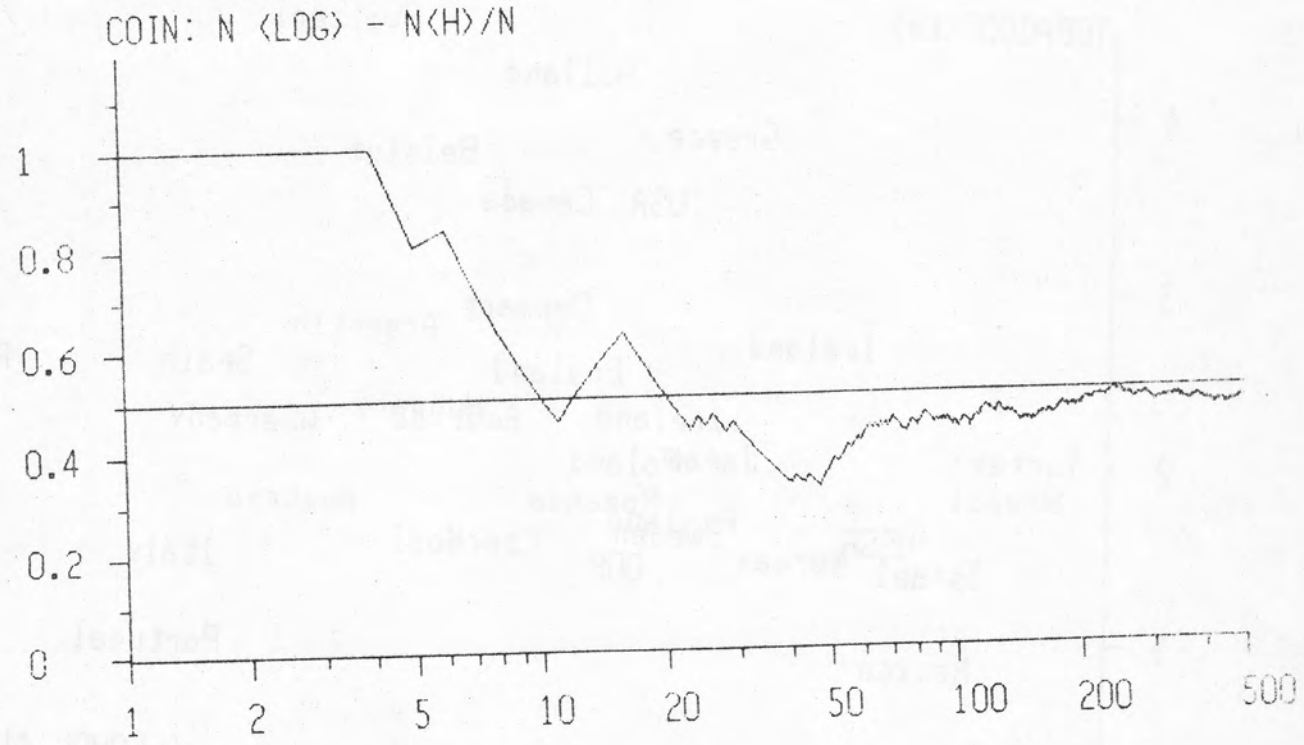


Fig.10 Two monthly time series (plotted by DIAGRAM)

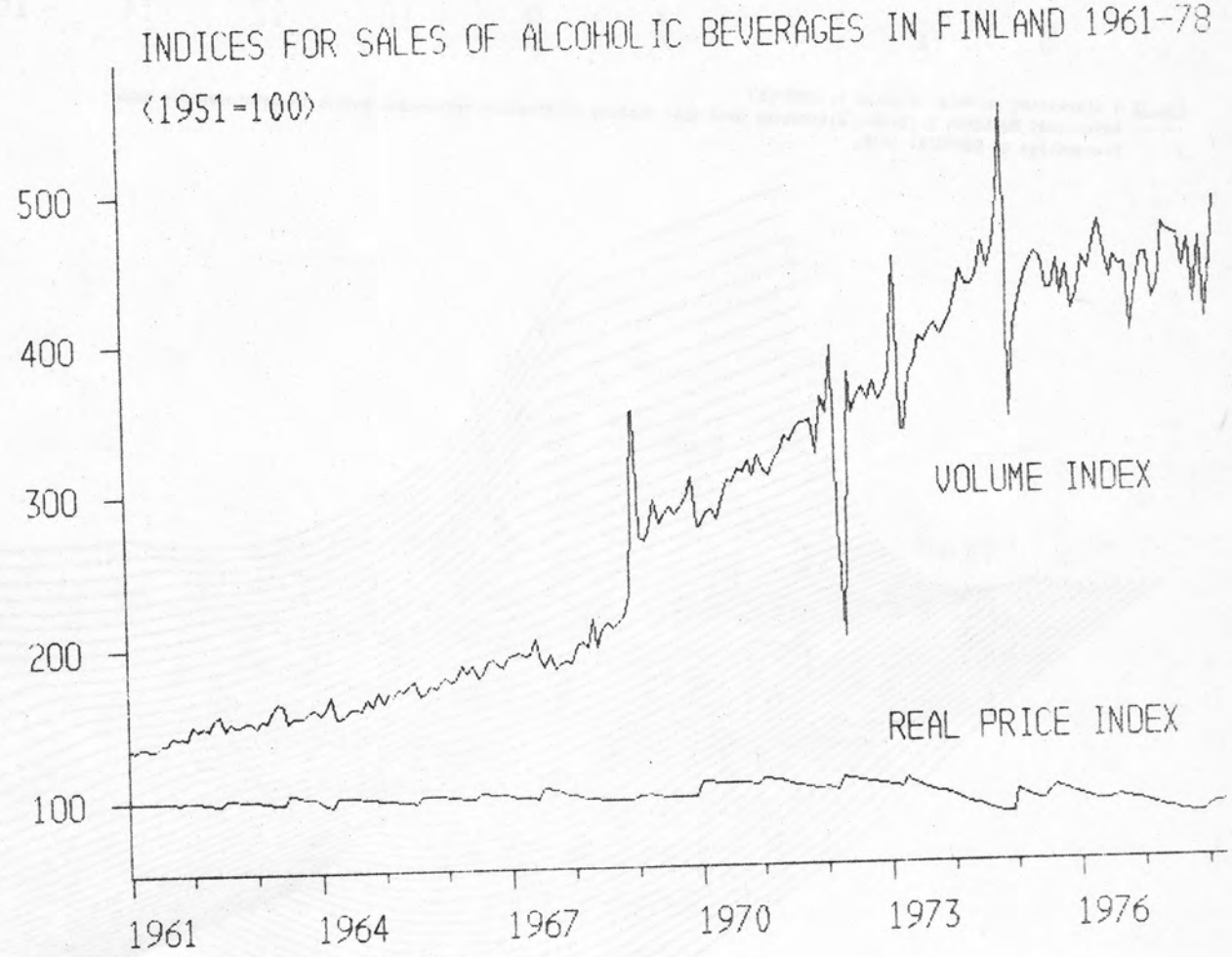


Fig.11 Yearly consumption of alcoholic beverages and tobacco in various countries (plotted by DIAGRAM)

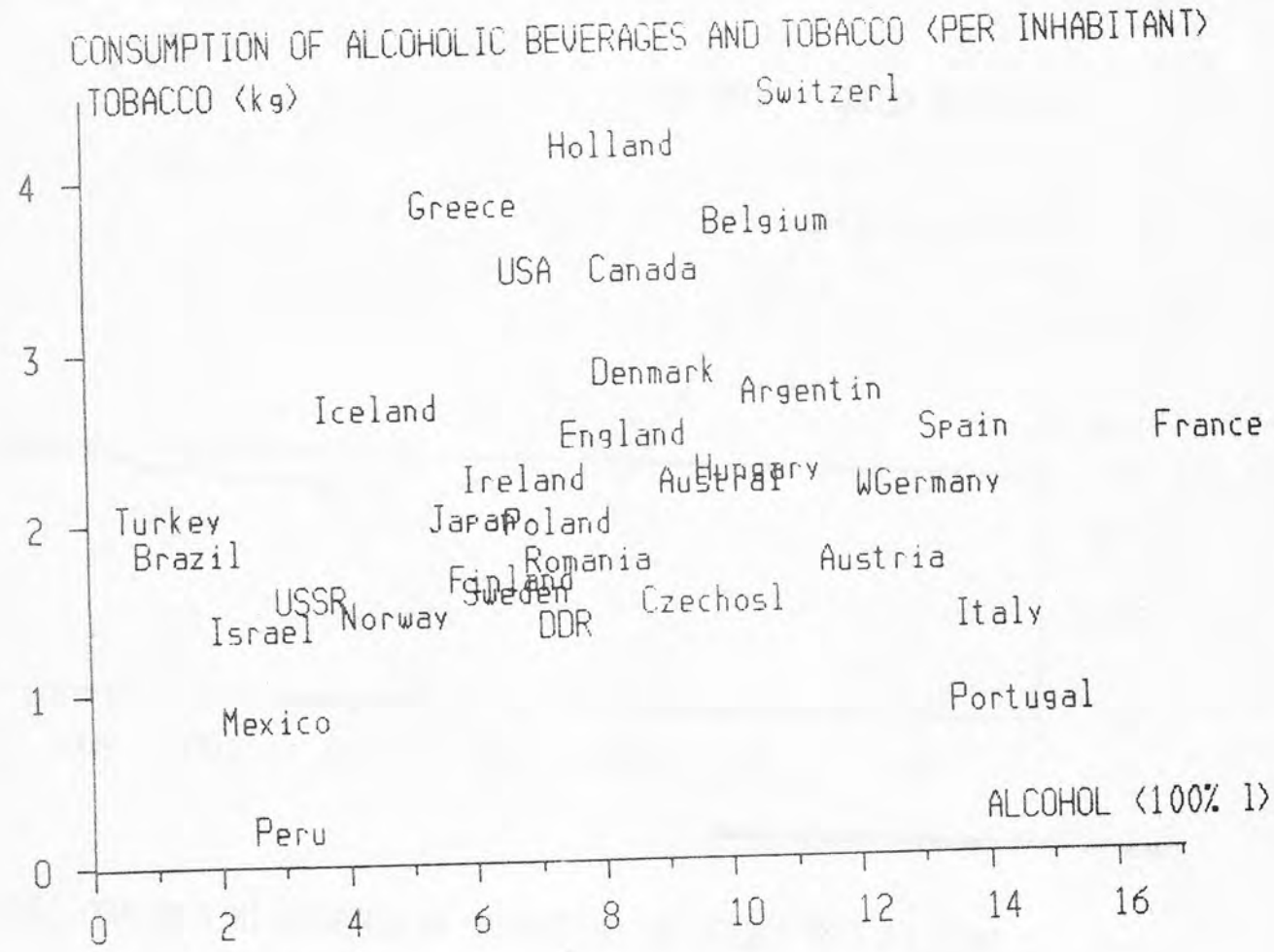


Fig.12 A digression surface (plotted by SURFACE)
Reference: Mustonen S.(1978), Digression analysis, fitting alternative regression models to heterogeneous data, Proceedings of COMPSTAT 1978.

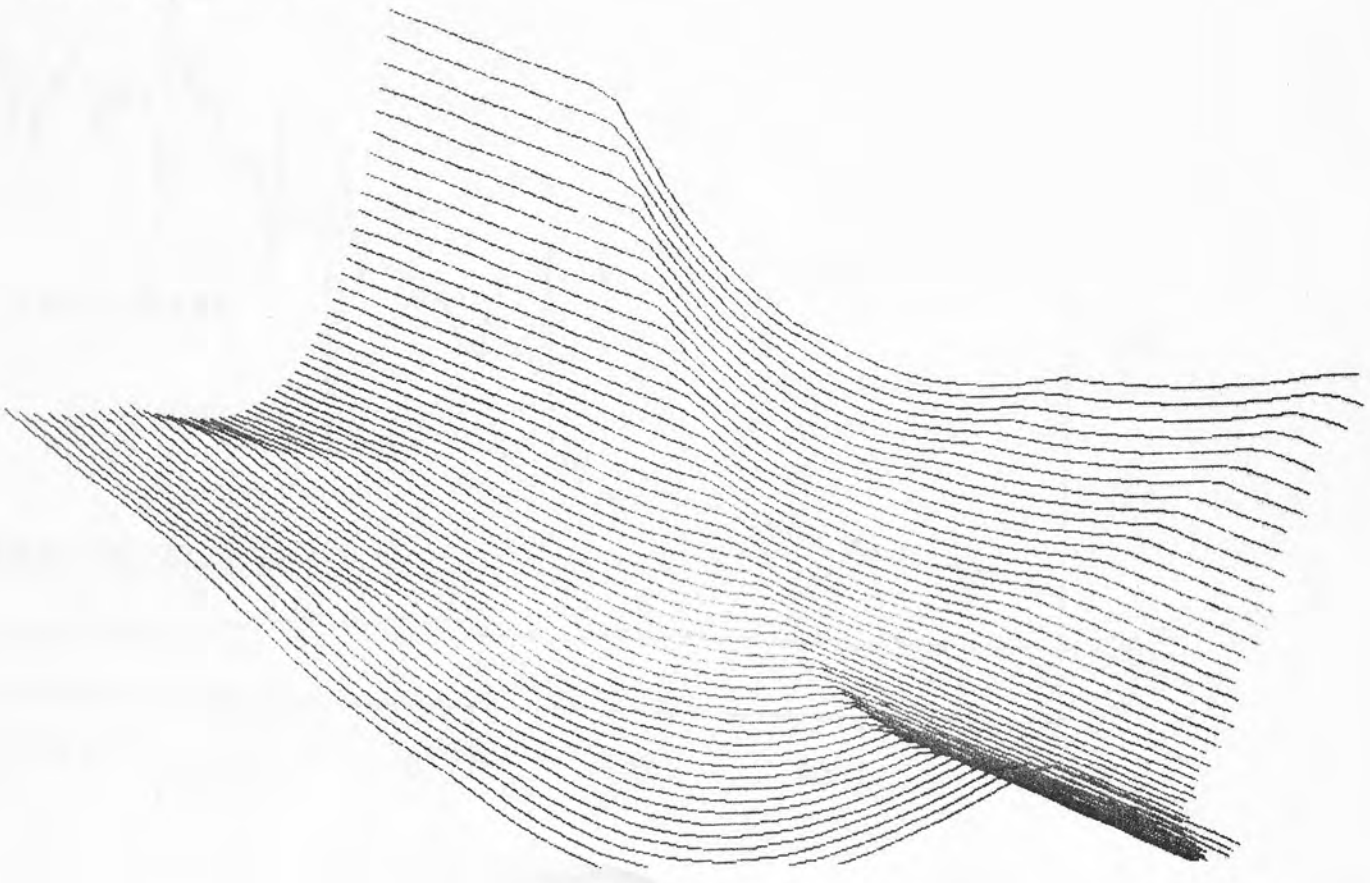


Fig.13 Representatitives of human races, apes and fossiles as Andrews' function plots (SCURVE)
The actual variables are measurements on teeth.
Reference: Andrews D.F.(1972), Plots of high-dimensional data, Biometrics 28,125-36.

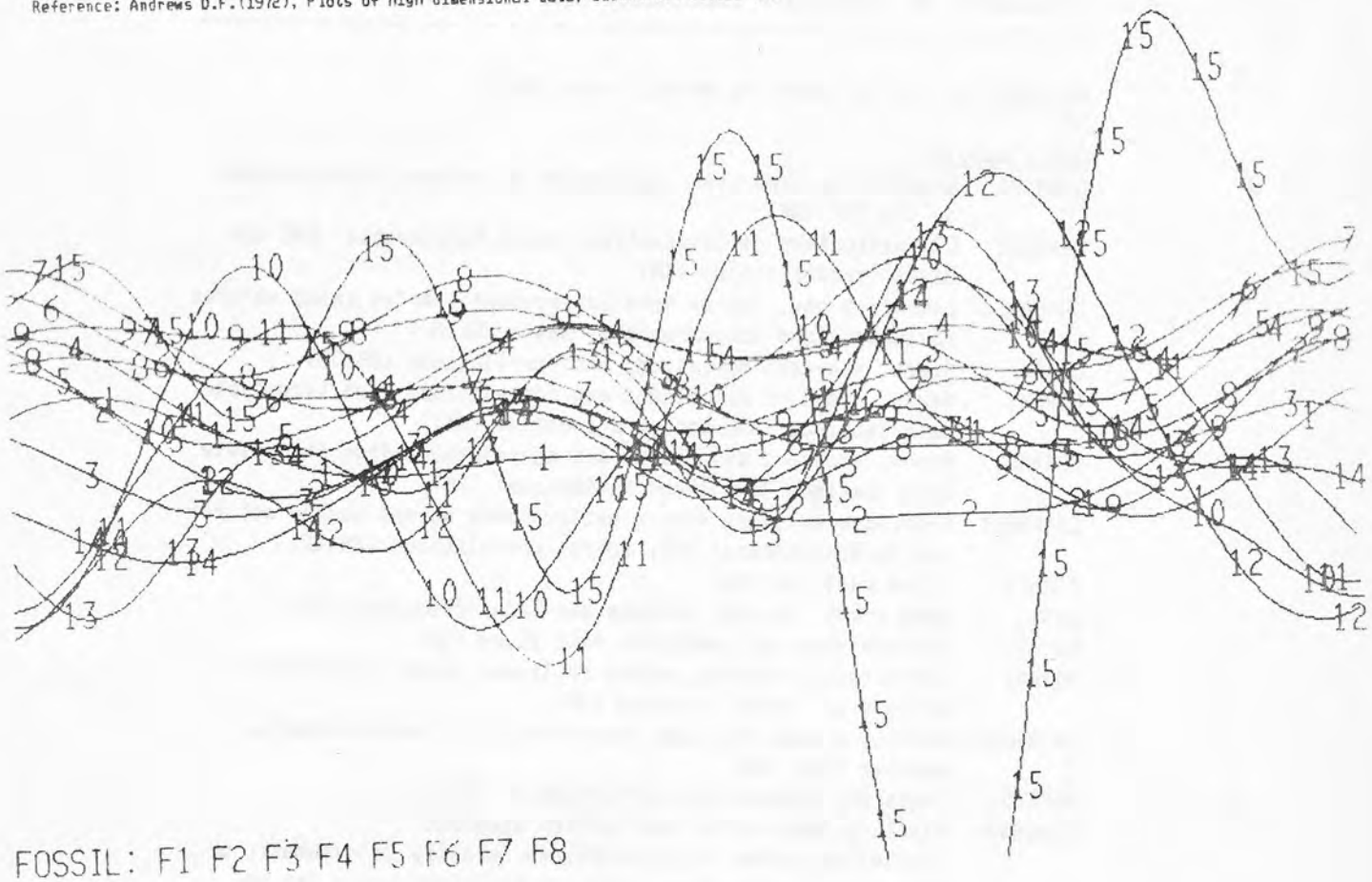
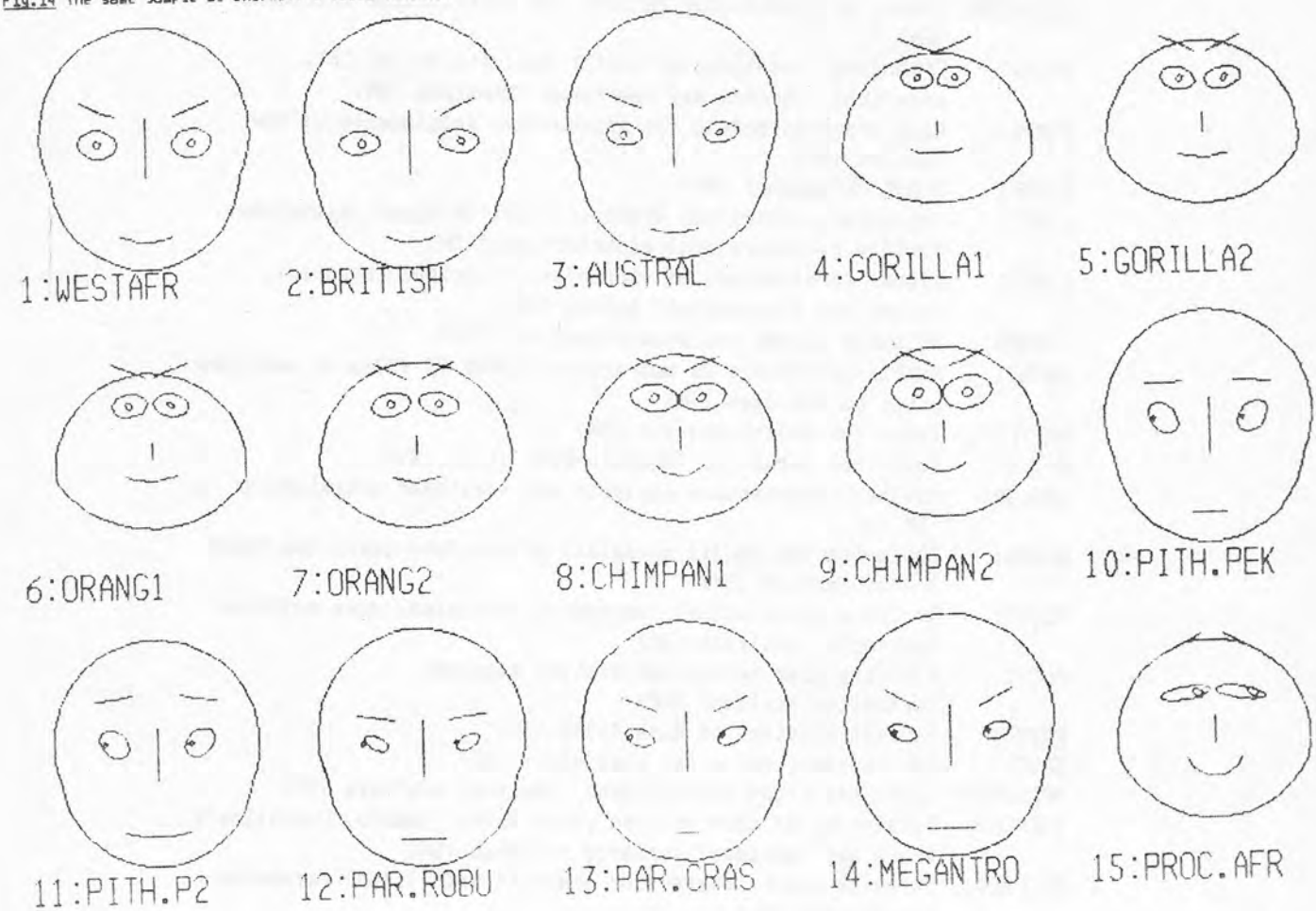


Fig.14 The same sample as Chernoff's faces (SCURVE)



APPENDIX 2: LIST OF SURVO 76 MODULES (May 1980)

Basic modules

CHANCE: Random data generator, simulation of various distributions on the CRT (SM)

CLASSI: Classification of observations using Mahalanobis' D↑2 and Bayes' probabilities (SM)

COMPARE: Comparing one, two or more independent samples using various parametric and non-parametric tests (SM,RL)

CORR: Means, standard deviations and correlations (SM,OS)

CORRM: Means, standard deviations and correlations from incomplete data (maximum information principle) (SM)

CORRML: Means, standard deviations and correlations from incomplete data (maximum likelihood estimation) (SM)

CORROBU: Detecting outliers from a multivariate normal sample according to Mahalanobis' D↑2, robust correlations (SM,RS)

CURVE: Curve plotting (SM)

DATA: Data input, saving, editing and transformations (SM)

DATA2: Transferring and combining data files (SM)

DATAM: Substituting missing values by linear least squares predictors and other criteria (SM)

DATASORT: Sorting a data file and transferring the sorted data in another file (SM)

DEPEND: Tests for indepencence of variables (IM)

DIAGRAM: Plotting time series and scatter diagrams (unlimited number of observations, scaling is automatic or determined by the user, also any nonlinear scale can be specified) (SM)

DISCRI: Multiple discriminant analysis (PH)

DISTRIBS: Values of theoretical density and distribution functions (EN)

FACTA: Orthogonal rotations of factor analysis on the CRT, graphical, varimax and quartimax rotations (SM)

FRAME: Half prepared module for interactive programming of new modules (SM)

GUIDE: SURVO 76 teacher (SM)

HISTO: Univariate classified frequency distributions, histograms, fitting by theoretical distributions (SM)

LINCO: Linear combinations of variables, principal component, factor and discriminant scores (SM)

LINREG: Multiple linear regression analysis (SM)

MATRI: Matrix operations on matrices in SURVO 76 files or matrices given bu the user (SM)

MN-TEST: Tests for multinormality (SM)

N-TEST: Tests for normality (Shapiro-Wilk etc.) (SM)

NONLIN: Nonlinear regression analysis and nonlinear optimization (SM,OS)

NORMA: Improving the (multi)normality of the data using the power transformation (SM)

PCOMP: Analysis of principal components, principal axes solution for factor analysis (SM)

PLOT: Plotting time series and scatter diagrams (automatic scaling) (MR)

PRINT: Tabular printout of data files (SM)

SORT: Data sorting and order statistics (SM)

SPECTRUM: Auto- and cross-correlations, spectral analysis (MR)

STEPCLU: Clustering of observations using Wilks' lambda, Hotelling's trace and (minimum) variance criteria (SM)

STEPREG: Stepwise linear regression analysis with linear parameter constraints (RL)

SURFACE: Surface plotting in general projection (SM)

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 (SM)
TABTEST: Computing the critical level of the X↑2-statistics in a frequency table by simulation (SM)
UNI: Univariate statistics (SM,JP)

Supplementary modules

AGGRE: Aggregation of observations (SM)
BIASREG: Biased linear regression analysis (ridge, latent root, and principal component regression, Stein's estimation) (IM)
BINORM: Simulation of bivariate normal distribution on the CRT (SM)
CLUSTER: Clustering of observations (according to ISODATA) (TM)
CLUSTREG: Linear regression analysis in heterogeneous data, estimation of two alternative models (JT)
CURVE2: As CURVE, but also for implicit functions (PT)
E: Editor for free text, data files and SURVO 76 results (SM)
FACES: Plotting multivariate observations as faces along the proposal of Chernoff (JASA 1973) (SM)
FOSS: Exponential curve fitting by numerical integration (OH)
HALEY: Seeks all the roots of an algebraic equation (SM)
LINTEST: Testing multivariate normality by testing linearity of regressions (Cox&Small 1978) (TA)
MATDATA: Transfers a matrix saved on disk in a data file (SM)
MIXNORM: Estimation of a mixture of two normal distributions by the method of moments (EN)
MNONLIN: Multivariate nonlinear regression analysis, ordinary least squares method (IM)
MULTGEN: Generating samples from a multivariate normal distribution (IM)
PARTCORR: Partial correlations, conditional means and standard deviations (SM)
RESTREG: Linear regression analysis with linear parameter constraints (OS)
SCURVE: The function plots of multidimensional data by the method of Andrews (SM)
TDATA: As DATA, but automatic labelling for time series observations (SM)

SURVO 76 contributors

TA	Timo Alanko	EN	Erkki Nykyri
OH	Olli Hämäläinen	JP	Juni Palmgren
PH	Pekka Hakkarainen	MR	Markku Rahiala
RL	Ritva Luukkonen	OS	Osmo Soininvaara
IM	Ilkka Mellin	RS	Reino Siren
SM	Seppo Mustonen	JT	Juha Tarkka