SURVO 76, an interactive statistical data processing system for a desk computer

S.Mustonen, Helsinki

1. PRINCIPLES OF SURVO 76

SURVO 76 is a statistical data processing system intended to cover a wide range of activities in computational statistics.

The SURVO 76 system has been designed especially for 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.

SURVO 76 is an interactive system and no special job describing language or code is needed. In its present form it has been implemented on a desk computer WANG 2200 which provides suitable means for rapid interchange of information between the system and the user.

The user can instantly reach any part of his data for inspection. Equally important is a rapid access to the different modules (programs and subroutines) 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.

SURVO 76 employs BASIC as a source language in a fairly extended form including many additional matrix, sort and alphanumeric statements.
Using the editing facilities of WANG 2200 the system modifications can be made easily and instantly.

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 dependance 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 previous stages of his job.

It is evident that many statisticians do not like to think in terms of computer programs. They prefer carrying out their computations 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 modules. In SURVO 76 the functions of these keys vary depending on the SURVO 76 module in use. The user not knowing which F-key to press next, can always resort to key F0 which displays on the CRT the functions of all other F-keys operative in the present situation.

Each F-start leads to a sequence of questions made by SURVO 76 which have to be answered by the user. The whole dialogue is displayed on the CRT and 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 printed after the question.
To give good 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 he merely presses the RETURN key. Otherwise he must write his own answer.

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

The SURVO 76 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 included in SURVO 76 has been splitted into small modules and it is the user's responsibility to combine them in a reasonable way employing F-keys. It would, of course, be easy to connect different submodules in a fixed order, but then the user would be at the mercy of the system - an undesirable but unfortunately typical feature of many statistical computing packages.

2. OPERATING SURVO 76

SURVO 76 can be run in WANG 2200 installations having a central processing unit with a memory of at least 20K, a CRT display, a dual removable floppy disk drive, a printer and a plotter. 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.

SURVO 76 consists of a central module and various statistical modules, of which one 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 disk memory. Thus the user needs not worry about the location of his data during computations.

The number of SURVO 76 modules is not limited in any way. New modules 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.

When SURVO 76 is loaded from disk a list of alternative modules is displayed on the CRT and the user has to select the module he wants to use next:

```
STATISTICAL DATA PROCESSING SYSTEM SURVO 76
 1= MENU       2= GUIDE      3= DATA       4= DATA2
 5= UNI        6= CORR       7= SORT       8= CHANCE
 9= TABLE     10= PLOT      11= DIAGRAM   12= FRAME
13= LINREG    14= NONLIN    15= PCOMP     16= SPECTRUM
17= ANOVA     18= HISTO     19= FACTA     20= RESTREG
21= CURVE     22= MATRI     23= DISTRIBS  24= MANOVA
25= DISCRI    26= CLASSI    27= LINCO     28= SURFACE
29= DATASORT

SELECT A SURVO 76 MODULE (NUMBER OR NAME)?
```

If the user selects module no.1 (MENU), he gets on the CRT a list of all SURVO 76 modules with a short description of their functions.
This list is reproduced in Appendix 1.

A more thorough description of each module can be obtained on the CRT by selecting the module in question. The module presents its activities including a list of operative F-starts and eventually some special instructions.

Module no.2 is a SURVO 76 teacher GUIDE which is an interactive program in itself and gives information pertinent to the SURVO 76 system during the conversation. After consulting GUIDE the user will be ready to call any SURVO module to suit his present needs.

3. SPECIAL FEATURES OF SURVO 76

3.1. DATA ANALYSIS

SURVO 76 contains several modules for statistical data analysis which is the main field of any statistical data processing system. Until now the emphasis has been on the most traditional and elementary forms of analysis (cf. Appendix 1) which give a natural basis for the development of the system.

In SURVO 76 the problems of feeding, editing and transforming data have received special attention. The modules DATA and DATA2 have been planned to cover various activities in this field to make the system self-contained in this respect.

One of the basic principles in SURVO 76 is that any potentially important observations or intermediate results can be used in subsequent computations without extra modifications of the system.

SURVO 76 allows both variables and observations to be labelled with alphanumeric names. This makes the results more readable and monitoring the computations easier.

Each SURVO 76 module is supposed to tell continuously on the CRT what it is doing. For instance, when observations are processed the system prints on the CRT the name of the observation to be dealt with next.

3.2. GRAPHIC REPRESENTATIONS

Both the plotter and the CRT are valuable devices for visualizing statistical data and theoretical models.

SURVO 76 can plot, for instance, scatter diagrams, time series, analytical curves and surfaces. To begin with, the graphs are usually plotted rapidly but inaccurately on the CRT. When the user finds that a certain graph is interesting and deserves closer examination he can easily transfer it on paper in a more accurate form using the plotter or the printer.

When making graphs many small nuisances like placing proper marks and indications on co-ordinate axes may annoy the user. Therefore it is valuable that the system can come to the rescue of the user in problems of that kind. Thus SURVO 76 takes care of scaling of the variables if desired and it also selects "nice" marking points on the axes according to the size of the graph (determined by the user) and the range of the variables to be plotted.

It is also 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, a time series and some of its components in the same picture. Likewise, after making a scatter diagram the user might wish to decide what kind of model should be fitted, estimate the model and return to plot a linear or nonlinear regression curve on the same graph.

3.3. MATRIX OPERATIONS

One attractive feature of the WANG 2200 computer is that 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 operations.

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 with SURVO 76 modules or other programs. For this purpose the system contains a module called MATRI.

With MATRI the user can perform a wide range of matrix operations using the computer like a calculator. In MATRI all the F-keys are defined for various matrix operations including matrix inversion, eigenvalues and vectors for symmetric matrices, partitioning and combining matrices.
The matrices required as an input can be keyed in manually or transferred from different SURVO 76 files. Results can be saved in special matrix files for later operations.

In the central memory there is space for three matrices only (two operands X,Y and one result Z). An essential feature of MATRI is that the module does a lot of book-keeping and labels each result with a name corresponding to the ordinary matrix notation. Hence although the user has to split matrix formulae into basic operations which are carried out by single F-keys he always has on the CRT labels of the latest operands and results in almost normal matrix notation.

3.4. RANDOM DATA SIMULATION

In many methodological considerations and teaching situations it is useful to analyse artificial random data whose origin is perfectly known.
Planning of such experiments can be substantially facilitated by employing a SURVO 76 module called CHANCE which is a random data generator.

The user has to write the statements needed to generate a typical observation which is done according to the instructions given by CHANCE. Several subroutines are immediately available in CHANCE to generate pseudo random variates from various distributions.

Thus it is easy even for an unexperienced "programmer" to construct random data according to a given statistical model. The simulated files can subsequently be treated as ordinary data files by means of SURVO 76.

Employing CHANCE the form of different sample distributions can also be demonstrated on the CRT. The user selects the distribution and its parameters and CHANCE starts to generate observations from that distribution. Observations are plotted on the CRT one after another as constantly growing histogram.

APPENDICES:

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

REFERENCES:

Mustonen S., SURVO 76: A statistical data processing system, Research Report No.6, Department of Statistics, University of Helsinki, 1977

Also two demonstration diskettes containing ready made conversations with SURVO 76 are available.
This paper has been prepared using an editing program developed in connection with SURVO 76.

APPENDIX 1: LIST OF SURVO 76 MODULES

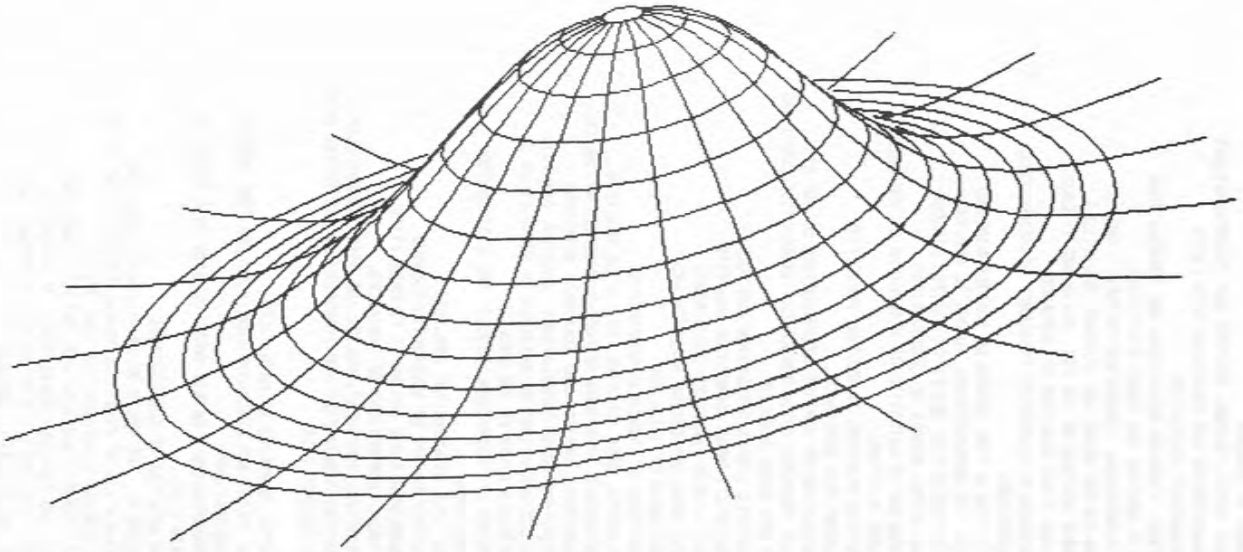
GUIDE: SURVO 76 TEACHER
DATA: DATA INPUT, SAVING, EDITING AND TRANSFORMATIONS
DATA2: TRANSFERRING AND COMBINING DATA FILES
UNI: UNIVARIATE STATISTICS
CORR: MEANS, STANDARD DEVIATIONS AND CORRELATIONS
SORT: DATA SORTING AND ORDER STATISTICS
TABLE: 2-DIMENSIONAL CLASSIFIED FREQUENCY TABLES,
TABLES FOR MEANS AND STANDARD DEVIATIONS,
TABLE EDITING ON THE CRT, CHI↑2 AND T TESTS,
1-AND 2-WAY ANALYSIS OF VARIANCE
HISTO: UNIVARIATE CLASSIFIED FREQUENCY DISTRIBUTIONS,
HISTOGRAMS
PLOT: PLOTTING A TIME SERIES OR SCATTER DIAGRAM
(MAX 170 OBSERVATIONS, AUTOMATIC SCALING)
DIAGRAM: PLOTTING A TIME SERIES OR SCATTER DIAGRAM
(UNLIMITED NUMBER OF OBSERVATIONS,
SCALING IS AUTOMATIC OR DETERMINED BY THE USER)
CURVE: CURVE PLOTTING
SURFACE: SURFACE PLOTTING IN CENTRAL PROJECTION
CHANCE: RANDOM DATA GENERATOR,
SIMULATION OF VARIOUS DISTRIBUTIONS ON THE CRT
FRAME: HALF PREPARED SURVO MODULE FOR INTERACTIVE COMPOSING
OF NEW SURVO MODULES
LINREG: MULTIPLE LINEAR REGRESSION ANALYSIS
RESTREG: MULTIPLE LINEAR REGRESSION ANALYSIS
WITH LINEAR PARAMETER CONSTRAINTS
NONLIN: NONLINEAR REGRESSION ANALYSIS AND
NONLINEAR OPTIMIZATION
PCOMP: ANALYSIS OF PRINCIPAL COMPONENTS,
PRINCIPAL AXES SOLUTION FOR FACTOR ANALYSIS
FACTA: ORTHOGONAL ROTATIONS IN FACTOR ANALYSIS ON THE CRT,
GRAPHICAL, VARIMAX AND QUARTIMAX ROTATIONS
SPECTRUM: AUTO- AND CROSS-CORRELATIONS, SPECTRAL ANALYSIS
MATRI: MATRIX OPERATIONS ON MATRICES IN SURVO FILES OR
MATRICES GIVEN BY THE USER
DISTRIBS: VALUES OF THEORETICAL DENSITY AND DISTRIBUTION
FUNCTIONS
DISCRI: MULTIPLE DISCRIMINANT ANALYSIS
CLASSI: CLASSIFICATION OF OBSERVATIONS USING
MAHALANOBIS D↑2 AND BAYES PROBABILITIES
LINCO: LINEAR COMBINATIONS OF VARIABLES,
PRINCIPAL COMPONENT, FACTOR AND DISCRIMINANT SCORES
DATASORT: SORTING A DATA FILE AND TRANSFERRING THE SORTED DATA
IN ANOTHER FILE

COPY: RAPID TRANSFERS OF DATA FILES
TDATA: AS 'DATA' BUT AUTOMATIC LABELLING FOR TIME SERIES
OBSERVATIONS.
MATDATA: TRANSFERRING A MATRIX SAVED ON DISK IN A SURVO 76
DATA FILE
AGGRE: AGGREGATION OF OBSERVATIONS
HALEY: SEEKS ALL THE ROOTS OF AN ALGEBRAIC EQUATION.
BINORM: SIMULATION OF BIVARIATE NORMAL DISTRIBUTION ON THE CRT
SIMALCO: A GENERAL SIMULATOR FOR TIME SERIES
CURVE2: AS "CURVE", BUT ALSO FOR IMPLICIT FUNCTIONS
DEPEND: TESTING FOR THE INDEPENDENCE OF VARIABLES
SCURVE: THE FUNCTION PLOTS OF MULTIDIMENSIONAL DATA
BY THE METHOD OF ANDREWS
N-TEST: TESTS OF NORMALITY (SHAPIRO-WILK ETC.)
MN-TEST: TESTS OF MULTINORMALITY
PRINT: 'NEAT' PRINTOUT OF SURVO 76 DATA FILES
GROUPS: CLUSTERING OF OBSERVATIONS (ACCORDING TO ISODATA)
PARTCORR: PARTIAL CORRELATIONS,
CONDITIONAL MEANS AND STANDARD DEVIATIONS

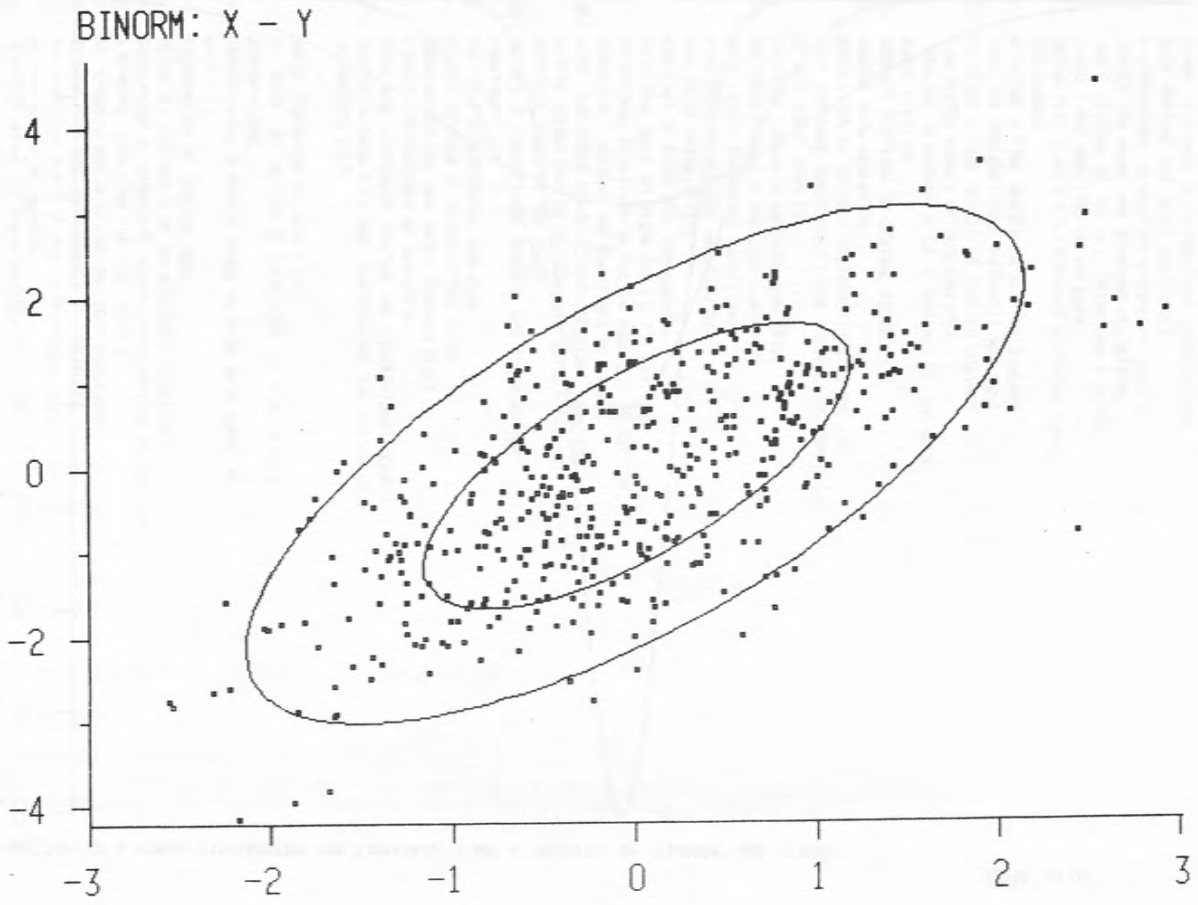
APPENDIX 2: GRAPHICS WITH SURVO 76

All these curves, surfaces and statistical graphs have been made employing the SURVO 76 modules CURVE, SURFACE and DIAGRAM, and using the WANG CRT plotter 2282.

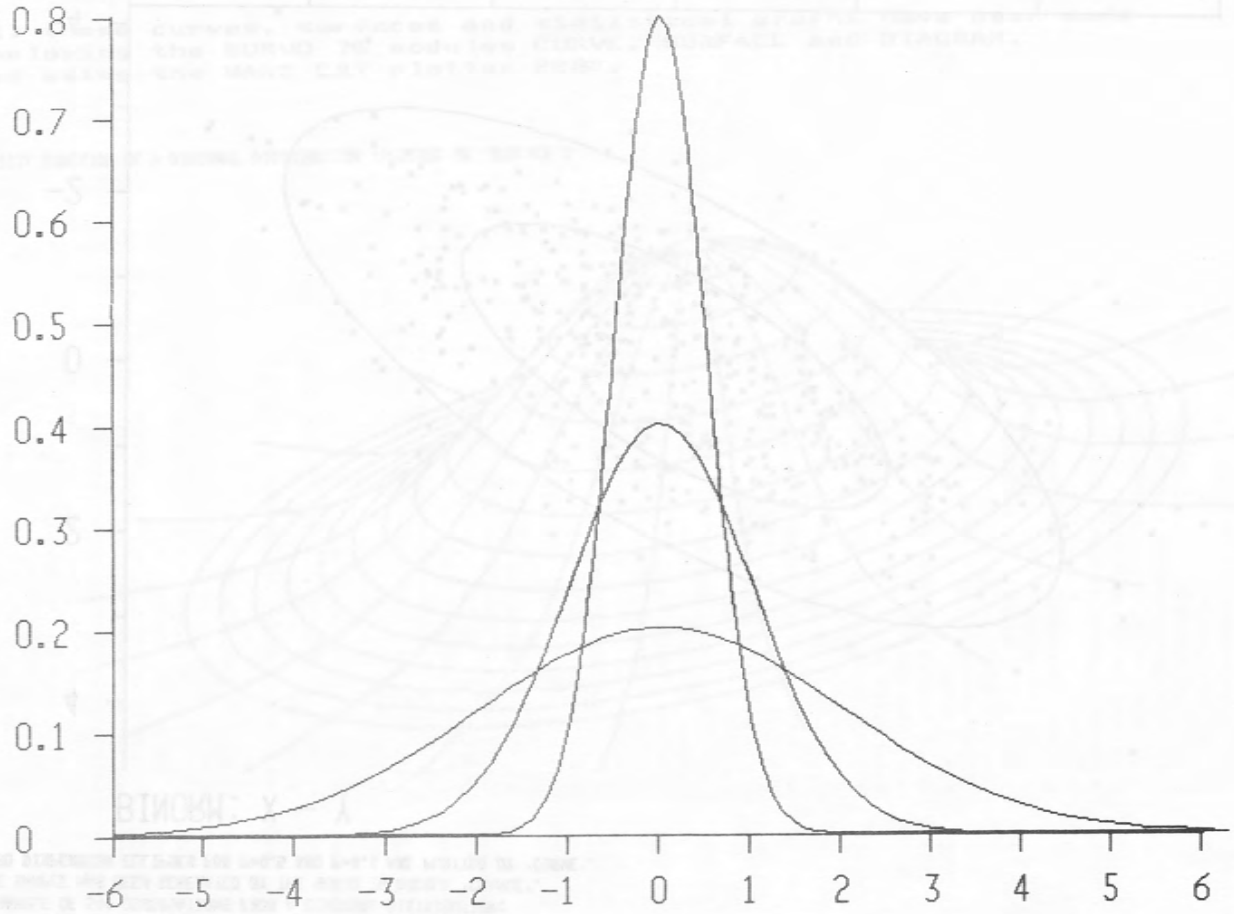
DENSITY FUNCTION OF A BINORMAL DISTRIBUTION (PLOTTED BY 'SURFACE')



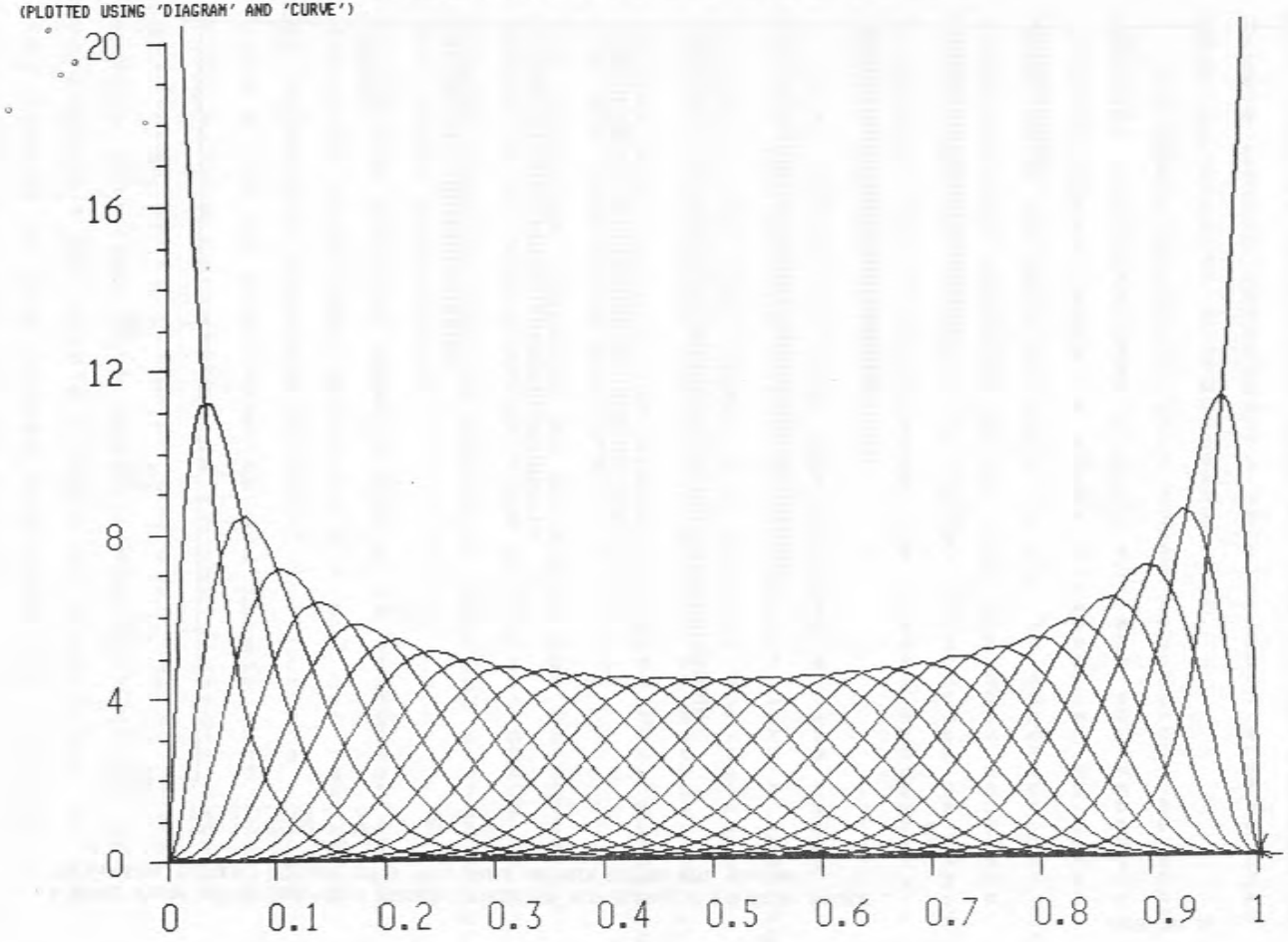
A SAMPLE OF 500 OBSERVATIONS FROM A BINORMAL DISTRIBUTION:
THE SAMPLE HAS BEEN GENERATED BY THE SURVO 76 MODULE 'CHANCE'.
ALSO DISPERSION ELLIPSES FOR P=0.5 AND P=0.1 ARE PLOTTED BY 'CURVE'.



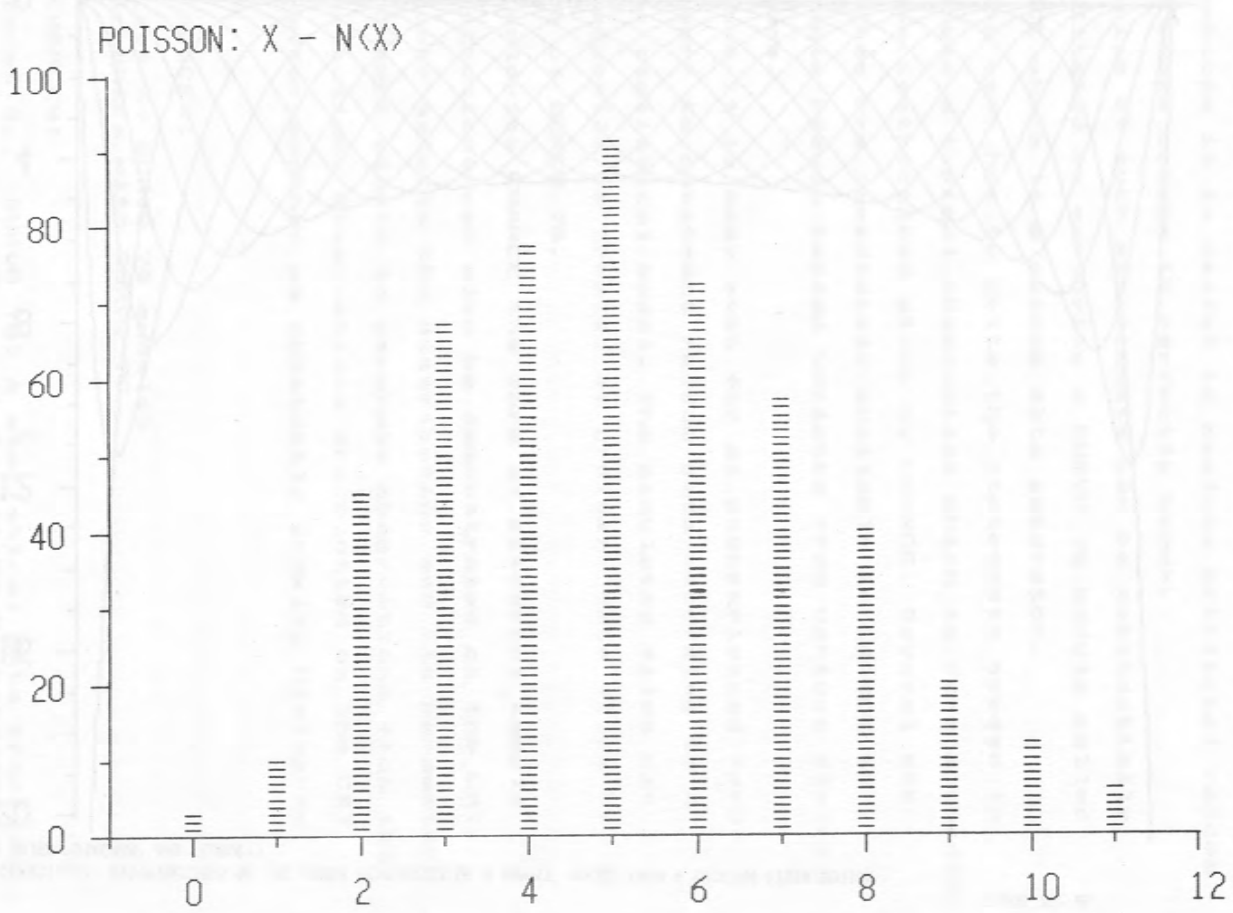
DENSITY FUNCTIONS OF A NORMAL DISTRIBUTION FOR SIGMA=0.5, 1 AND 2 (PLOTTED BY 'DIAGRAM' AND 'CURVE')



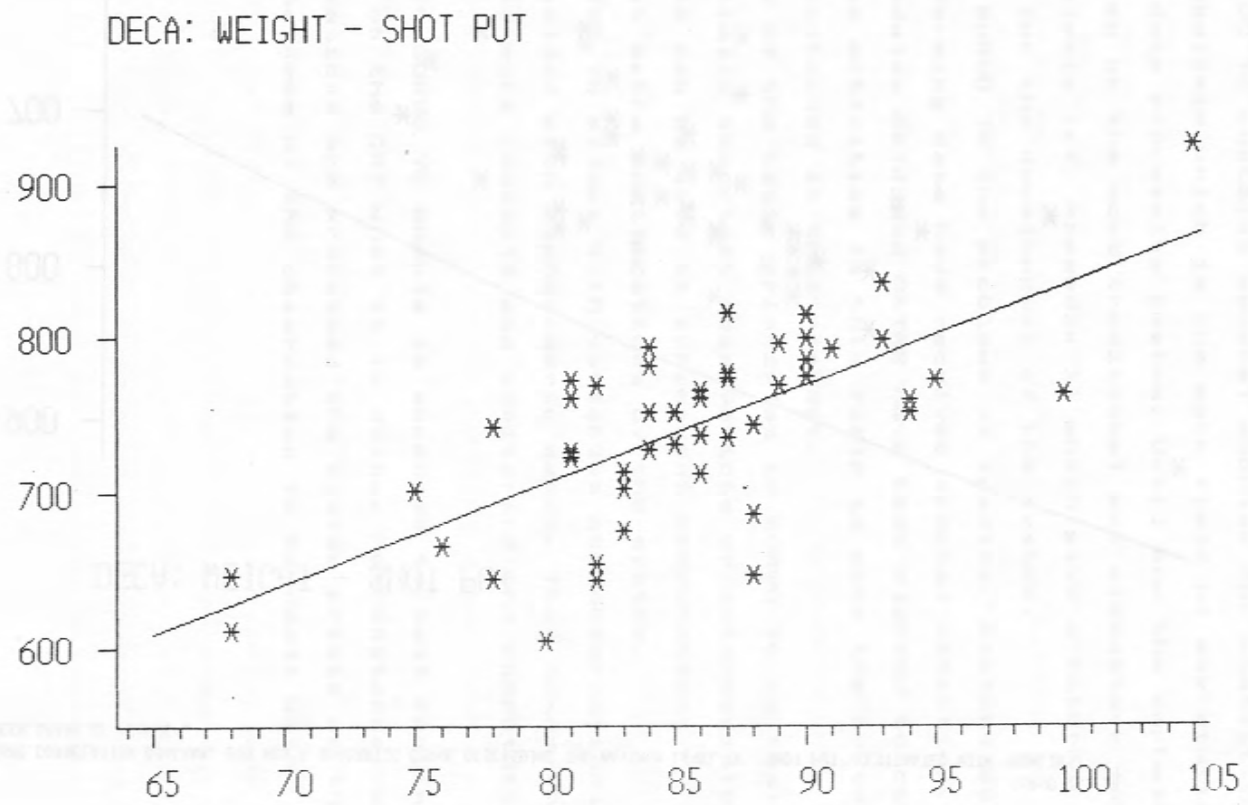
30 BETA DENSITIES: DISTRIBUTIONS OF THE ORDER STATISTICS OF A SAMPLE (N=30) FROM A UNIFORM DISTRIBUTION
(PLOTTED USING 'DIAGRAM' AND 'CURVE')



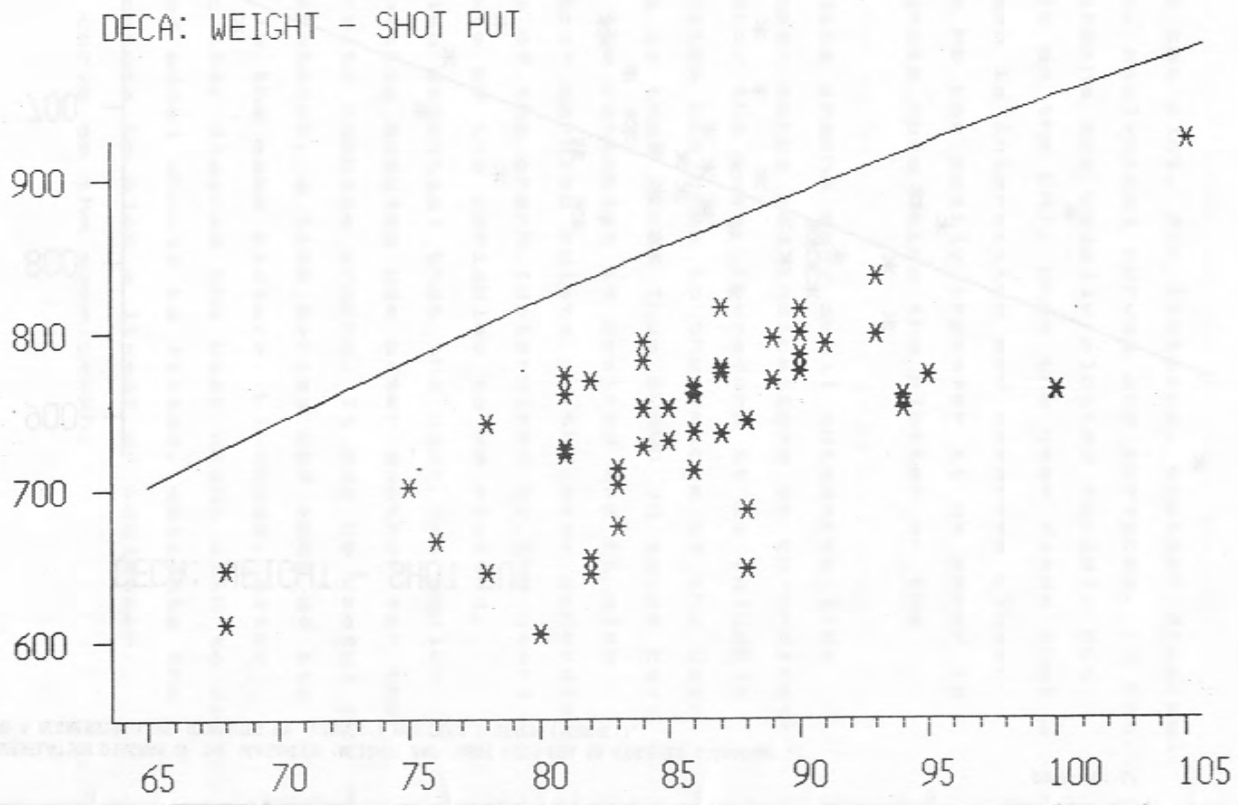
A SAMPLE OF 500 OBSERVATIONS FROM A POISSON DISTRIBUTION WITH MEAN=5, GENERATED BY 'CHANCE'.
THE EMPIRICAL FREQUENCY DISTRIBUTION OF THIS SAMPLE HAS BEEN PLOTTED WITH 'DIAGRAM'.



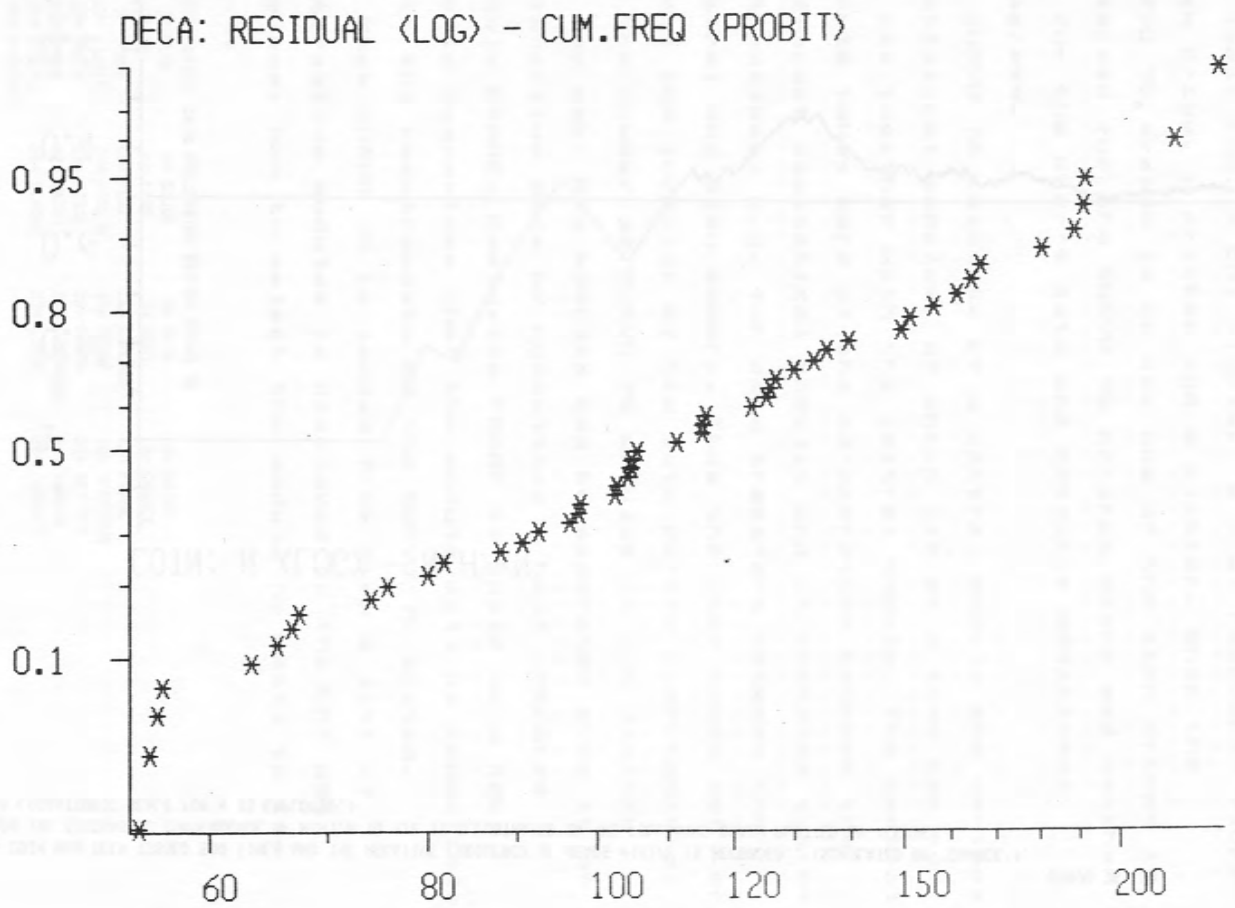
A CORRELATION DIAGRAM OF THE VARIABLES 'WEIGHT' AND 'SHOT PUT' FOR 48 ATHLETES ('DIAGRAM'):
ALSO A REGRESSION LINE (COMPUTED BY 'LINREG') HAS BEEN PLOTTED ('CURVE').



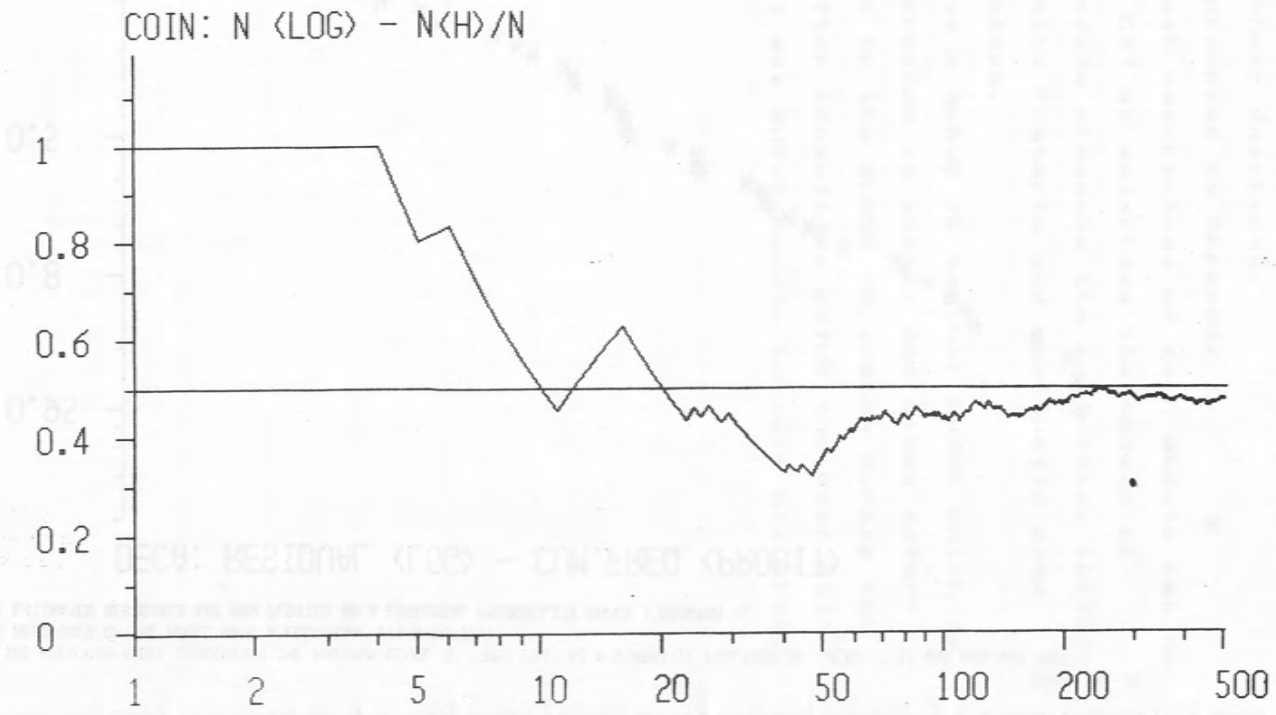
THE SAME CORRELATION DIAGRAM, BUT NOW A QUADRATIC CURVE DESCRIBING THE MAXIMUM LEVEL OF 'SHOT PUT' (ESTIMATED WITH 'NONLIN')
HAS BEEN DRAWN BY 'CURVE'.



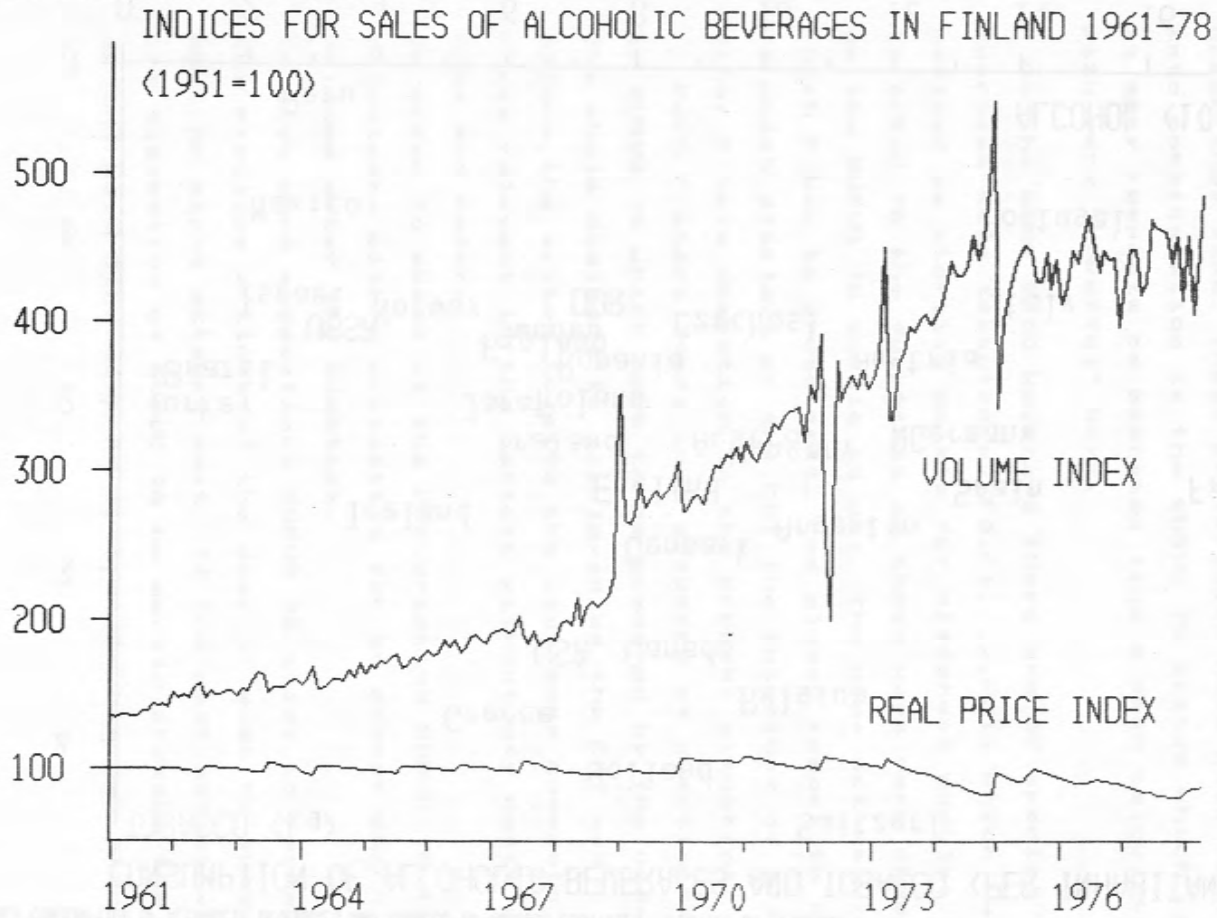
IN THE PREVIOUS MODEL CONCERNING THE MAXIMUM LEVEL OF 'SHOT PUT' AS A QUADRATIC FUNCTION OF 'WEIGHT' IT WAS ASSUMED THAT
THE RESIDUALS OF THE MODEL HAVE A LOGNORMAL DISTRIBUTION.
THE ESTIMATED RESIDUALS ARE NOW PLOTTED ON A LOGNORMAL PROBABILITY PAPER ('DIAGRAM').



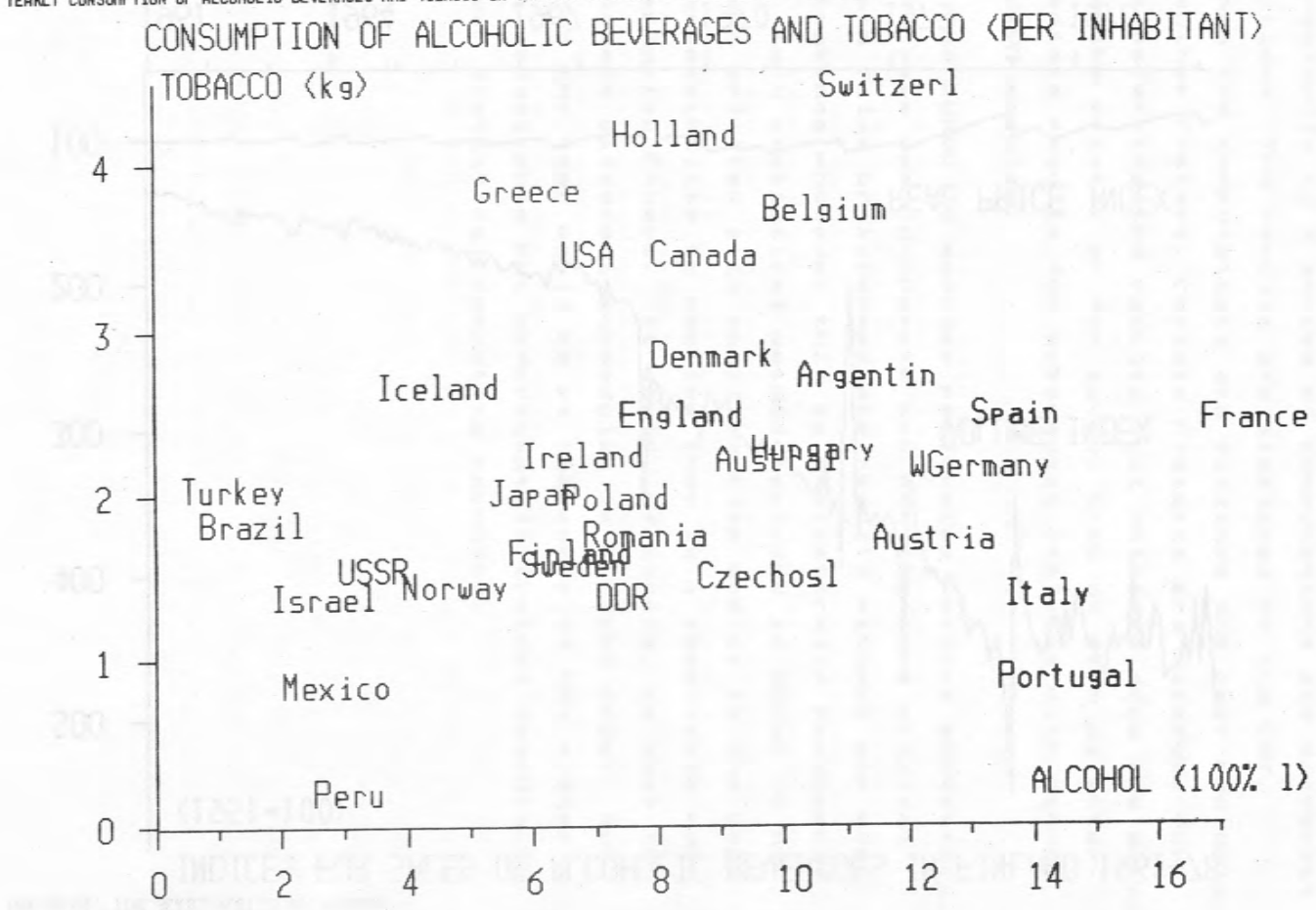
A COIN HAS BEEN TOSSED 500 TIMES AND THE RELATIVE FREQUENCY OF HEADS N(H)/N IS RECORDED., (SIMULATED BY 'CHANCE').
NOW THE STOCHASTIC CONVERGENCE OF N(H)/N TO 1/2 IS ILLUSTRATED IN THE FOLLOWING GRAPH PLOTTED BY 'CURVE'.
(A LOGARITHMIC SCALE FOR N IS EMPLOYED.)



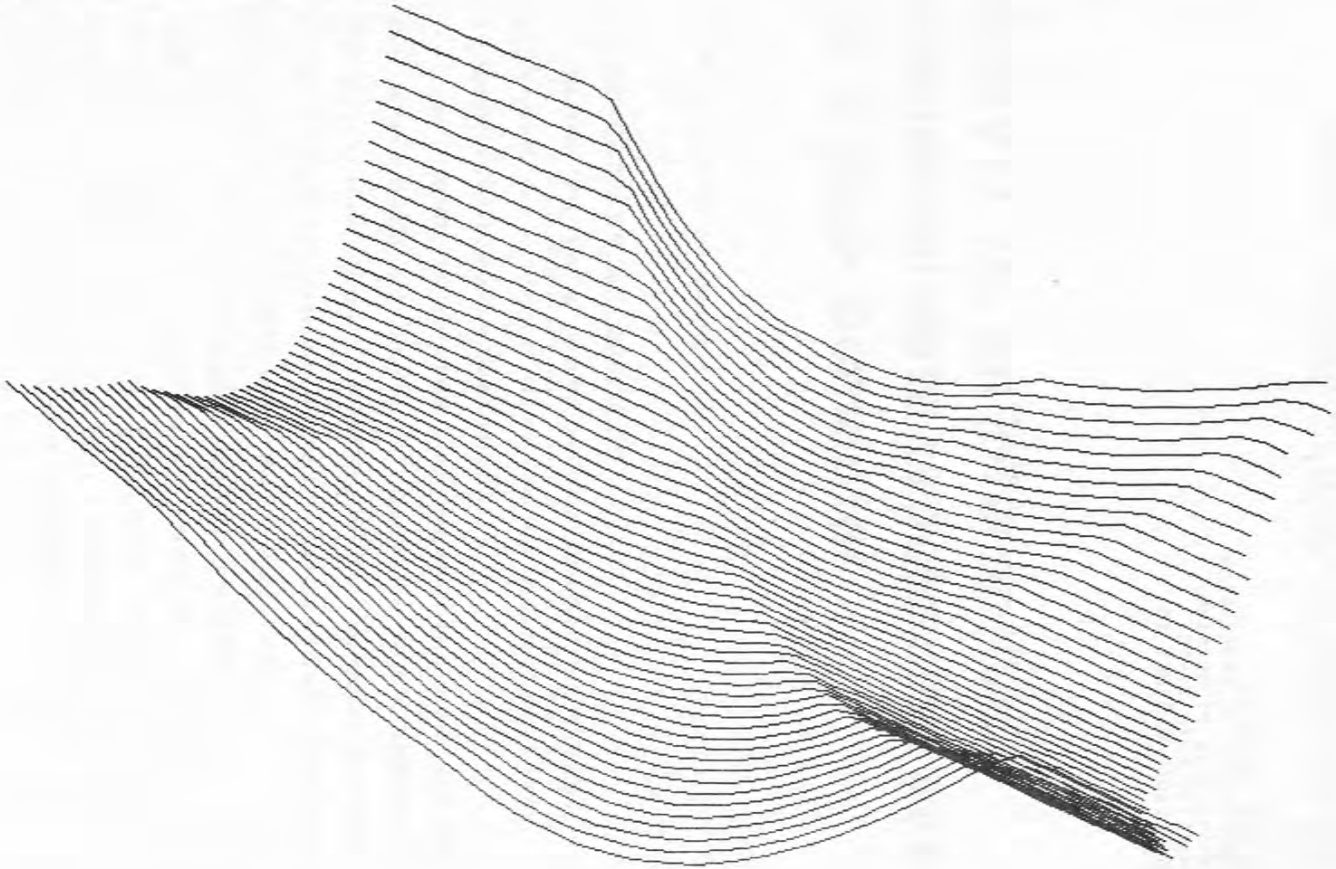
TWO MONTHLY TIME SERIES PLOTTED BY 'DIAGRAM'



YEARLY CONSUMPTION OF ALCOHOLIC BEVERAGES AND TOBACCO IN VARIOUS COUNTRIES (PLOTTED BY 'DIAGRAM')



A 'DIGRESSION' SURFACE (PLOTTED BY 'SURFACE')



This work has been supported by the Academy of Finland and
ALKO (the State Alcohol Monopoly of Finland).

Seppo Mustonen
Department of Statistics
University of Helsinki
Hallituskatu 11-13
SF-00100 HELSINKI 10