NOTES

contents dd. 4-11-1964

INTRODUCTION

This binder will contain information on programs and procedures, written in ALGOL 60, in use at the Department of Mathematical Statistics of the Mathematical Centre, Amsterdam.

Since these programs are run with the Electrologica X-1 computer of the Mathematical Centre, they are subject to restrictions imposed by the M.C. Algol-system. On the other hand they make use of the facilities of this system. If necessary reference will be made to publications of the Computational Department.

The contents of this binder will consist of four parts.

part 1

A series of notes, in which general remarks will be made on conventions for input-output and organisation of programs. These notes will be numbered 1,2,..... etc. preceded by "note". The pages of each note will be numbered separately. If a note is replaced by a new version a letter will be added to its number, at a following replacement this letter will be changed.
So "note 1.a.2" indicates the second version of note 1, page 2.

part 2

In this part description will be given of programs and procedures that are available and may be of interest. For a program this description will contain: (at least)

In brief the purpose of the program;
rules for preparing an input tape;
formulae according to which the input data are processed;
meaning of the output data.

The description of a procedure will at least consist of:

Purpose of the procedure;
restrictions of the parameters;
method or reference to method.

The pages of this part will be numbered as those of part 1, but preceded by "descr" instead of "note".

<u>part 3</u>

A series of Algol programs, numbered STAT 000 etc. Of each program in this series a description will be given in part 2, but the description of a certain program in part 2 does not imply that the Algol text will be presented in part 3. A program presented in this series will have to satisfy high standards of effectiveness, organisation and lay-out of input-output, use of Algol and lay-out of the text.

<u>part 4</u>

A series of Algol procedures, numbered SP 100 etc. The same as said for the series of programs, holds true for this series. Unlike the programs these procedures will not make use of special facilities of the M.C. Algol-system. Due to testing the restrictions cannot be avoided.

Each of the four parts will be preceded by a list of contents, and the "starting calendar date" of these contents.

Also a list of replacements will be given. It is possible that somewhere reference is made to old pages, so they should not be discarded.

The responsibility for the contents of this binder is divided between:

Jac. Anthonisse,	for Algol,
B. Dorhout ,	for Operations Research,
J. Oosterhoff ,	for Statistics.

GENERAL REMARKS

Although it is possible to deduce conventions by reading the texts of the programs, some will be noted here.

Input of each program is via tape. An input tape is supposed to have a code-number and the first number on the tape should be equal to its code number.

Output is also via tape. The results of a program are preceded by the code-number of the program and the code-number(s) of the input-tape(s).

The typewriter is only used to give instructions to the operator. Most instructions, like "put input tape into reader" have an obvious meaning. The instruction "set XEENtal" however, needs some explanation. Generally a program will be performed on one set of input-data, and can be restarted to process the next data. If a program has to be performed p times this implies at least 2 p stops, as the program stops also after typing an instruction "put (next) input tape into reader". A program that types the instruction "set XEENtal" avoids these stops, if XEEN(-0)=p and all sets of input data are on one tape.

The output of a program consists of text and numbers. In the output-description these texts will be given between "≮" and "≯". These "string quotes" are not given in the output.

INPUT OF SETS OF OBSERVATIONS

The input of a program often contains one or more sets of observations. If a single set is punched in a standard-form, the tape can serve as part of the input of several programs. The standard-form of a set of N observations, each consisting of d numbers, is a lexicographical sorted frequency table, punched in the following way:

cod	n	d					
i_1	i_2	i_3	-	-	-	i_d	
$x_{1,1}$	$x_{1,2}$	-	-	-	-	$x_{1,d}$	f_1
.							
.							
.							
$x_{n,1}$	$x_{n,2}$					$x_{n,d}$	f_n .

Meaning:

cod = code-number of the set

n = number of different observations

d = number of variables

i_j = index of the j-th variable, $i_j > 0$, to be omitted if d = 1

$x_{r,k}$ = value of the k-th variable in the r-th observation

f_r = frequency of the r-th observation.

The observations are said to be sorted lexicographical if for $r = 2,3,\ldots,n$:

there is a k(r) $1 \leq k \leq d$ with:

a. $x_{r-1,m} = x_{r,m}$ $m = 1,2,\ldots,k-1$

b. $x_{r-1,k} < x_{r,k}$

Non-sorted observations should be given in this way:

cod	-n	d				
i_1	i_2	i_3	-	-	-	i_d
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	-	-	-	$x_{1,d}$
.						
.						
.						
$x_{n,1}$	$x_{n,2}$	$x_{n,3}$				$x_{n,d}$.

Where n = number of observations, the other symbols have the same meaning as above.

It will be impossible, in general, to use a non-sorted frequency table of observations as input for a program.

If the input of a program contains more than one, or a non-prefixed number of sets of observations they should be preceded by some numbers to make the declaration of such arrays possible that only a minimum of storage space will be used by these observations.

MATRICES OF CROSSPRODUCT SUMS, COVARIANCES, CORRELATIONS

In note 3 the standard-form of a set of observations has been defined. That form is used to define the standard-form of:

I. Matrix of crossproduct sums:

cod			d+1	
i_1	$a_{1,1}$			
i_2	$a_{2,1}$	$a_{2,2}$		
.				
.				
.				
i_d	$a_{d,1}$	$a_{d,2}$	- - - - -	$a_{d,d}$
i_{d+1}	$a_{d+1,1}$	$a_{d+1,2}$	- - - - - - - -	$a_{d+1,d+1}$

In which:

cod = code-number of the matrix,

d+1 = number of variables + 1,

i_j = code-number of the j-th variable, $i_j > 0$,

$a_{r,k} = \sum_{l=1}^{n} x_{l,r} \cdot x_{l,k} \cdot f_l \; , \; 1 \leq k \leq r \leq d,$

$a_{d+1,k} = \sum_{l=1}^{n} x_{l,k} \cdot f_l \; , \; 1 \leq k \leq d,$

$a_{d+1,d+1} = \sum_{l=1}^{n} f_l \; .$

II. Matrix of covariances:

cod			d+1	
i_1	$b_{1,1}$			
i_2	$b_{2,1}$	$b_{2,2}$		
.				
.				
.				
i_d	$b_{d,1}$	$b_{d,2}$	- - - - -	$b_{d,d}$
i_{d+1}	$b_{d+1,1}$	$b_{d+1,2}$	- - - - - - - -	$b_{d+1,d+1}$

Meaning:

cod = code-number of the matrix,

d+1 = number of variables + 1,

i_j = code-number of the j-th variable, $i_j > 0$,

$b_{r,k} = (a_{r,k} - a_{d+1,r} \cdot a_{d+1,k} / a_{d+1,d+1}) / (a_{d+1,d+1} - 1)$, $1 \leq k \leq r \leq d$,

$b_{d+1,k} = a_{d+1,k} / a_{d+1,d+1}$, $1 \leq k \leq d$,

$b_{d+1,d+1} = a_{d+1,d+1}$,

$a_{r,k}$ as defined in I.

III. Matrix of correlations:

cod			d+1	
i_1	$c_{1,1}$			
i_2	$c_{2,1}$	$c_{2,2}$		
.				
.				
.				
i_d	$c_{d,1}$	$c_{d,2}$	-----	$c_{d,d}$
i_{d+1}	$c_{d+1,1}$	$c_{d+1,2}$	-------	$c_{d+1,d+1}$

In which:

cod = code-number of the matrix,

d+1 = number of variables + 1,

i_j = code of the j-th variable, $i_j > 0$,

$c_{r,k} = b_{r,k} / \sqrt{b_{r,r} \cdot b_{k,k}}$, $1 \leq k \leq r \leq d$,

$c_{d+1,k} = b_{d+1,k}$, $1 \leq k \leq d + 1$,

$b_{r,k}$ as defined in II.

DESCRIPTIONS
contents dd. 30-3-65

DESCRIPTIONS
contents dd. 30-3-65
Continued

SORTING and RANKING

Using the procedures presented in descr 18 this program produces a sorted frequency table of observations or a sorted frequency table of rank numbers.

A. Sorted observations

Input is a non-sorted set of observations (see note 3):

$$\begin{array}{lcl} \text{cod} & -\text{n} & \text{d} \\ i_1 & \text{-----} & i_d \\ x_{11} & \text{-----} & x_{1d} \\ \vdots & & \\ x_{n1} & \text{-----} & x_{nd} \end{array} \qquad (1)$$

The output is the lexicographically sorted frequency table of these observations:

$$\begin{array}{llll} & \text{code} & \text{m} & \text{d} \\ i_1 & \text{-----} & i_d & \\ y_{11} & \text{-----} & y_{1d} & f_1 \\ \vdots & & & \\ y_{m1} & \text{-----} & y_{md} & f_m \end{array} \qquad (2)$$

Here code = 100 × cod + 6

m = number of different observations in (1)

f_i = number of times $(y_{i1}, \dots, y_{id})$ appears in (1)

B. Rank numbers

Input consists of a lexicographically sorted frequency table (2).

The output is the lexicographically sorted frequency table of rank-numbers:

$$
\begin{array}{llll}
\text{code nr} & \text{m} & & \text{d} \\
i_1 & - - - - - & i_d & \\
z_{11} & - - - - - & z_{1d} & f_1 \\
\vdots & & & \vdots \\
z_{m1} & - - - - - & z_{md} & f_m
\end{array}
\qquad (3)
$$

Here: $z_{ij} = s_{ij} + \frac{t_{ij} + 1}{2}$

with $s_{ij} = \sum_{k=1}^{m} \delta(y_{ij}, y_{kj})$ $\quad \delta(y_{ij}, y_{kj}) = \begin{cases} f_k & \text{if } y_{ij} > y_{kj} \\ 0 & \text{otherwise} \end{cases}$

$t_{ij} = \sum_{k=1}^{m} \gamma(y_{ij}, y_{kj})$ $\quad \gamma(y_{ij}, y_{kj}) = \begin{cases} f_k & \text{if } y_{ij} = y_{kj} \\ 0 & \text{otherwise} \end{cases}$

code nr = code + 1

The other symbols have the same meaning as in (2).

CROSSPRODUCTS SUMS, COVARIANCES, CORRELATIONS

The program described here may be used to compute the matrices that are defined in note 4.

The input consist of a set of observations, punched in one of the two forms given in note 3 , followed by four integers: j_1, j_2, j_3 and j_4.

First the matrix of Cross product sums is computed, in accordance with the formulae of note 3. If the set of observations is a large one this will be done in several stages: In each stage a sub-set of observations is read into the computer and its matrix of inner products added to the matrix that has been formed in previous stages.
If j_1 = 1 the matrix of Cross products sums is punched. Its code-number = 100 . (code-number of input-tape) + 1.

If j_2 = 1 the following list of means and variances is punched:
{ Result < code number of program > }
{ see descr 2 }
{ Means and Variances }

cod	d		
{ index	N	mean	variance }
i_1	$b_{d+1,d+1}$	$b_{d+1,1}$	$b_{1,1}$
.			
.			
.			
i_d	$b_{d+1,d+1}$	$b_{d+1,d}$	$b_{d,d}$.

Where cod = 100 . (... input-tape) + 2, if d = 1 the column of i_j's is omitted, the other symbols have the same meaning as under "Matrix of covariances" in note 4.

If j_3 = 1 the matrix of Covariances is punched, its code-number = 100 . (... input-tape) + 3.

If j_4 = 1 the matrix of Correlations is punched, its code-number = 100 . (... input-tape) + 4.

Each matrix is punched in accordance with note 4, preceded by some text:

‡ Result < code-number of program > ‡

‡ see descr 2. ‡

‡ matrix of <   > ‡

The program defines $i_{d+1} = 11111$, and, if $d = 1$, $i_1 = 1$.

STEPWISE REGRESSION

A set of N independent (n + 1)-dimensional observations $(\underline{x}_{j,0}, x_{j,1}, \ldots, x_{j,n})$, $j = 1, 2, \ldots, N$, is supposed to be given. In the sequel we assume that the random variable $\underline{x}_0$ has linear regression on the n independent fixed variables $x_1, x_2, \ldots, x_n$. A polynomial regression equation can, of course, be reduced to a linear equation by introducing new independent variables for different powers of the same original variable.

If it is desirable to reduce the number of independent variables in the regression function, one can determine the subset of m independent variables associated with the largest sum of squares due to regression, for $m = 1, 2, \ldots, n$. However, such an optimal subset of m variables can only be computed by comparing all $\binom{n}{m}$ different subsets of m variables. In this program a less cumbersome method is presented, called the method of "stepwise regression" or "forward selection". Full details concerning this method are given in [1]. We remark, that the subsets of m variables produced by this method are not always optimal (cf. [1] and [2]). An easily evaluated criterion guaranteeing optimality of the produced subsets is as yet unknown.

The computations are performed in several steps. At the first step the sums of squares due to regression on each of the n independent variables separately $R_i^{(1)}$ $(i = 1, \ldots, n)$ are computed, and the variable associated with the largest sum of squares, say x_{i_1}, is included in the regression function. At the r-th step $(r = 2, \ldots, n)$ the difference $R_i^{(r)}$ of the sum of squares due to regression on the set of independent variables $\{x_{i_1}, \ldots, x_{i_{r-1}}, x_i\}$ and the sum of squares due to regression on the set $\{x_{i_1}, \ldots, x_{i_{r-1}}\}$ is computed for all $i = 1, \ldots, n$; $i \neq i_1, \ldots, i_{r-1}$, and the variable x_i associated with the largest difference, say x_{i_r}, is added to the variables already included in the regression function. Having chosen the variable to be included in the regression function, the regression coefficients are computed at each step.

This procedure can be carried out on a matrix of (crude) cross-product sums or on a matrix of covariances (see note 4). If a constant is admitted in the regression function, a variable with constant value 1 should be among the n independent variables and the crude crossproduct sums must be used.

The explicit computations performed follow the procedure sketched in [1] and will not be restated here. The output not only provides the differences of the sums of squares $R_{i_r}^{(r)}$, but also the sums of squares $\sum_{l=1}^{r} R_{i_l}^{(l)}$ itself and these quantities as percentages of the total sum of squares.

Some modifications of this procedure are possible:

1. In the first k steps (k = 1, ..., n) k given independent variables can forcibly be included in the regression function by proper input indications.
2. A set of k independent variables (k = 1, ..., n-1) can be given with the effect, that none of the variables will be included in any regression equation by proper input indications.
3. The procedure can be stopped after the r-th step by proper input indications.

Literature

[1] H.C. HAMAKER, On multiple regression analyses, Statistica neerlandica 16 (1962), 31-56.

[2] J. OOSTERHOFF, On the selection of independent variables in a regression equation, to be published.

Input:

a matrix of crossproduct sums or covariances as defined in note 4,
code-number of the problem,
index of the independent variable,
index of the constant variable (11111) or 0, (if no constant variable is present),
maximum number of independent variables to be included in a regression equation,
-1 or a row of indices followed by -1, with the effect that none of the variables with one of these indices will be included in any equation,
*) -1 or a row of indices followed by -1, with the effect that in the r-th step of the procedure the variable with the r-th of these indices will be added to the regression function. If this row contains k indices, the criterion given above is applied from step k+1 on.
The program contains XEENtal.

Output:

{Result <code number of program>}
{Stepwise Regression}
{see descr 3}
{input} code number of the matrix code number of the problem

{Sum of squares} $a = \sum_{l=1}^{N} \{x_{l,0}^{(1)}\}^2$

{step	variable	R	Rcum	Rper	Rcumper }
1	i_1	$R_{i_1}^{(1)}$	$R_{i_1}^{(1)}$	$100.R_{i_1}^{(1)}/a$	$100.R_{i_1}^{(1)}/a$
r	i_r	$R_{i_r}^{(r)}$	$\sum_{l=1}^{r} R_{i_l}^{(l)}$	$100.R_{i_r}^{(r)}/a$	$100.\sum_{l=1}^{r} R_{i_l}^{(l)}/a$

If, in step r, the constant variable is included, or a variable is forced into the equation via *) this is mentioned after the corresponding line in this scheme.

{survey of R }

i_1 $R_{i_1}^{(1)}$

i_r $R_{i_r}^{(1)}$ $R_{i_r}^{(2)}$ - - - - - - $R_{i_r}^{(r)}$

{regression coefficients}

1 β_{1,i_1}

r β_{r,i_1} β_{r,i_2} - - - - - - β_{r,i_r}

LINEAR PROGRAMMING (Explicit inverse algorithm)

Consider the following problem:

Find the minimum of

(1) $c' x$,

where x is restricted by

(2) $A_k x \leqq b_k$, $(b_k \geqq 0)$

(3) $A_g x \geqq b_g$, $(b_g \geqq 0)$

(4) $A_e x = b_e$, $(b_e \geqq 0)$

(5) $x \geqq 0$,

with

$$c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad b_k = \begin{bmatrix} b_1 \\ \vdots \\ b_{m_1} \end{bmatrix}, \quad b_g = \begin{bmatrix} b_{m_1+1} \\ \vdots \\ b_{m_2} \end{bmatrix}, \quad b_e = \begin{bmatrix} b_{m_2+1} \\ \vdots \\ b_m \end{bmatrix},$$

$$A_k = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m_1,1} & \cdots & a_{m_1,n} \end{bmatrix}, \quad A_g = \begin{bmatrix} a_{m_1+1,1} & \cdots & a_{m_1+1,n} \\ \vdots & & \vdots \\ a_{m_2,1} & \cdots & a_{m_2,n} \end{bmatrix},$$

$$A_e = \begin{bmatrix} a_{m_2+1,1} & \cdots & a_{m_2+1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix}.$$

A_k and b_k, A_g and b_g, A_e and b_e are allowed to be empty.

It is assumed that $a_{m_2,0} \geqq a_{j,0}$, $j = m_1+1, \ldots, m_2-1$.

The program, described here, solves the problem in the following way. Let the $(m_2 - m_1) \times (m_2 - m_1)$-matrix F and the $(m_2 - m_1)$-vector e_{m_2} be defined by

$$F = \begin{bmatrix} -1 & 0 & \circ & \circ & \circ & 0 & 1 \\ 0 & -1 & \circ & & & \circ & \circ \\ 0 & 0 & \circ & \circ & & \circ & \circ \\ \circ & & \circ & \circ & \circ & \circ & \circ \\ \circ & & & \circ & \circ & 0 & 1 \\ \circ & & & & \circ & -1 & 1 \\ 0 & 0 & \circ & \circ & \circ & 0 & 1 \end{bmatrix} \qquad e_{m_2} = \begin{bmatrix} 0 \\ \circ \\ \circ \\ \circ \\ \\ 0 \\ 1 \end{bmatrix}$$

Then the problem is transformed into:

Find the minimum of

(6) $\bar{c}' \bar{x}$,

where $\bar{x}$ is restricted by

(7) $A \bar{x} = b$,

(8) $\bar{x} \geq 0$,

with $\bar{c}' = \left[\bar{c}_1, \ldots, \bar{c}_{n+m+1}\right] \overset{\text{def}}{=} \left[c_1 \ldots c_n \; 0 \ldots 0\right]$,

$\bar{x}' = \left[x_1 \ldots x_{n+m+1}\right]$,

$$A = \begin{bmatrix} A_k & I_1 & 0 & 0 & 0 \\ FA_g & 0 & -F & e_{m_2} & 0 \\ A_e & 0 & 0 & 0 & I_2 \end{bmatrix} , \quad b = \begin{bmatrix} b_k \\ Fb_g \\ b_e \end{bmatrix}$$

(I_1 and I_2 are unit matrices of appropriate size);

$x_{n+1} \ldots x_{n+m_2}$ are slack variables, $x_{n+m_2+1} \ldots x_{n+m+1}$ are artificial variables.

The columns of A are denoted by $a_{\circ j}$, $j = 1 \ldots n+m+1$.

This problem is treated in 3 phases, each of them consisting of 0 or more steps. In each step the inverse of the basis is transformed. The solution of (7) and (8), corresponding to the current basis

$B = \left[a_{\cdot j_1} \ldots a_{\cdot j_m}\right]$ is

$\bar{x}_B = \left[\bar{x}_{j_1} \ldots \bar{x}_{j_m}\right]' = B^{-1} b.$

Defining $\bar{c}_B' = \left[\bar{c}_{j_1} \ldots \bar{c}_{j_m}\right]$ and

$d_j = \bar{c}_j - \bar{c}_B' B^{-1} a_{\cdot j}, \; j = 1 \ldots n+m_2,$

the vector $a_{\cdot s}$ to be inserted in the new basis is found in phases 2 and 3 by taking for s the lowest index of the most negative d_j. If all the d_j are non-negative, then the optimal solution is obtained.

Computations start in phase 3, if there are only restrictions (2), otherwise in phase 1.

Phase 1: $\sum_{j=n+m_2+1}^{n+m+1} x_j$ is minimized in the way described above, but with this difference, that $\bar{c}' \bar{x}$ is replaced by the appropriate objective function. Phase 2 starts, if the minimum is 0. If 0 cannot be reached the problem is infeasible: Computations are terminated.

Phase 2: (6) is minimized. If possible an artificial vector is eliminated from the basis. Otherwise the step is performed in the usual way. Phase 3 starts, when a basis without artificials is obtained. (So in non-degenerate problems phase 2 is deleted).

Phase 3: (6) is minimized.

It is possible to use the inverse basis and x of the optimal solution to solve a new problem. This problem must be almost the same as the original one: only c and the non-basic $\bar{a}_{\cdot j}$ of the optimal solution may be replaced.

Input for the program should contain an indicator to specify whether the inverse basis of a previous problem will be used, and whether the output should contain the inverse basis of the present problem. The use of this indicator is shown in the table:

indicator	former result is used	inverse basis required
0	no	no
1	no	yes
2	yes	no
3	yes	yes

A number t is considered to be 0, if $|t| < \varepsilon$.
This ε should be given in the input.
It is necessary to introduce this ε because of the inexactness of the real arithmetic. The results of the program may be disturbed by too small or too large values of ε.
In general $\varepsilon = 10^{-8}$ turns out to be satisfying.

The indices of x may be arbitrarily chosen integers $k_1, \ldots, k_n$, instead of $1, \ldots, n$.

Literature

1. G.B. DANTZIG, Linear Programming and Extensions, Princeton 1963;

2. S.I. GASS, Linear Programming: Methods and Applications, New York 1958;

3. G. HADLEY, Linear Programming, Reading (Mass. U.S.A.) 1962;

4. J. KRIENS, Leergang Mathematische Besliskunde, hoofdstuk XII, Mathematisch Centrrum, rapport S 265 (C 13), Amsterdam 1962;

5. G. ZOUTENDIJK, Methods of Feasible Directions, Amsterdam 1960.

Input:

code number of the input tape

$n, \quad m_1, \quad m_2, \quad m,$

$$\begin{array}{llllll} k_1 & c_1 & a_{1,1} & \ldots & a_{m,1} \\ \vdots & \vdots & \vdots & & \vdots \\ k_n & c_n & a_{1,n} & \ldots & a_{m,n} \\ & & b_1 & \ldots & b_m \end{array}$$

indicator, ε

If indicator $\geq$ 2: output VI of a previous problem.

The program contains set XEENtal.

Output:

I. {Result <code number of the program>}
{Linear Programming,}
{see descr 4}
{input} code number of the problem

II. If indicator $\geq$ 2:
{code number of feasible solution} code number of output VI of a previous problem

III. If the problem has a finite solution:
{solution}
$k_{j_i} \qquad x_{k_{j_i}} \qquad$ for all j_i with $1 \leq j_i \leq n$

‡slack variables‡

$j_i - n$ $\qquad x_{j_i}$ $\qquad$ for all j_i with $j_i > n$

‡d_j‡

$d_1 \ldots d_{n+m_2}$

‡number of steps‡ number of steps

‡minimum‡ minimum of (6)

IV. If the problem is infeasible:

‡problem infeasible‡

V. If the problem has an infinite solution:

‡variable‡ k_j ‡infinite‡ or

‡slack in row‡ $j - n$ ‡infinite‡

VI. If the indicator = 0 or 2, if the problem is infeasible or if the problem has an infinite solution:

‡<code number of the program> input for further computations‡

(code number of the problem) × 100 + 5

$$
\begin{array}{c}
\begin{array}{l} j_1 \ldots j_m \end{array} \\
\left[\begin{array}{c|c}
B^{-1} b & B^{-1} \\
\hline
-\bar{c}'_B B^{-1} b & -\bar{c}'_B B^{-1}
\end{array}\right]
\end{array}
$$

Here $B = \left[a_{\cdot j_1} \ldots a_{\cdot j_m} \right]$ denotes the basis of the last step accomplished.

TESTS FOR A CONTINGENCY TABLE

Suppose that in each of N identical and independent experiments one of the events $A_1, A_2, \ldots, A_r$ and one of the events $B_1, B_2, \ldots, B_k$ will occur. Let p_{ij} denote the probability that A_i and B_j will occur together in any given experiment, and a_{ij} the total number of occurences of this event A_iB_j ($i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, k$, $\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} = 1$, $\sum_{i=1}^{r} \sum_{j=1}^{k} a_{ij} = N$). Furthermore the following partial sums are defined:

$$R_{ij} = \sum_{l=1}^{j} a_{il} \text{ , } K_{ij} = \sum_{l=1}^{i} a_{lj} \text{ , } T_{ij} = \sum_{m=1}^{i} \sum_{n=1}^{j} a_{mn} \text{ ;}$$

obviously

$$T_{ij} = \sum_{m=1}^{i} R_{mj} = \sum_{n=1}^{j} K_{in} \quad \text{and} \quad T_{rk} = N.$$

The numbers a_{ij} may be given in the form of a kxr-contingency table

	B_1	B_2	B_3 - - - - - -	B_k	Total
A_1	a_{11}	a_{12}	a_{13} - - - - - -	a_{1k}	$R_{1,k}$
A_2	a_{21}	a_{22}	a_{23} - - - - - -	a_{2k}	$R_{2,k}$
A_3	a_{31}	a_{32}	a_{33} - - - - - -	a_{3k}	$R_{3,k}$
⋮	⋮			⋮	⋮
A_r	a_{r1}	a_{r2}	a_{r3} - - - - - -	a_{rk}	$R_{r,k}$
Total	$K_{r,1}$	$K_{r,2}$	$K_{r,3}$	$K_{r,k}$	$T_{r,k} = N.$

In this program the hypothesis of independence of the A_i and B_j is tested, i.e. $p_{ij} = p_{i.}\, p_{.j}$. If this hypothesis is rejected, two different methods are available to obtain information concerning the type of dependence. With each chi-squared test given, the number of a_{ij} having expectations < 5 and < 1 are given.

A. Chi-squared test for independence

Hypothesis H_0: $p_{ij} = p_{i\circ}\, p_{\circ j}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, k$.
The test-statistic for testing H_0 is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(a_{ij} - b_{ij})^2}{b_{ij}}, \text{ where } b_{ij} = \frac{R_{ik} K_{rj}}{T_{rk}} ;$$

under H_0 it has approximately a chi-squared distribution with $(r - 1)(k - 1)$ degrees of freedom.

B. Components of chi-squared [2]

The total chi-squared given under A may be split up into $(r - 1)(k - 1)$ additive components

$$\chi^2_{i+1,j+1} =$$

$$= \frac{T_{r,k}\left[K_{r,j+1}(R_{i+1,k}T_{i,j} - R_{i+1,j}T_{i,k}) - T_{rj}(R_{i+1,k}K_{i,j+1} - a_{i+1,j+1}T_{i,k})\right]^2}{R_{i+1,k}K_{r,j+1}T_{rj}T_{r,j+1}T_{i,k}T_{i+1,k}} ,$$

$i = 1, 2, \ldots, r-1$, $j = 1, 2, \ldots, k-1$, each of which has under H_0 approximately a chi-squared distribution with one degree of freedom. $\chi^2_{i+1,j+1}$ is associated with the 2x2-table

	$B_1 v B_2 v \ldots v B_j$	B_{j+1}	Total
$A_1 v A_2 v \ldots v A_i$	$T_{i,j}$	$K_{i,j+1}$	$T_{i,j+1}$
A_{i+1}	$R_{i+1,j}$	$a_{i+1,j+1}$	$R_{i+1,j+1}$
Total	$T_{i+1,j}$	$K_{i+1,j+1}$	$T_{i+1,j+1}$

and $$\chi^2 = \sum_{i=1}^{r-1} \sum_{j=1}^{k-1} \chi^2_{i+1,j+1}.$$

C. <u>Tests for outlying columns and rows</u>

For j = 1, 2, ..., k consider the 2xr-table

	B_j	$\bar{B}_j$	Total
A_1	a_{1j}	$R_{1k}-a_{1j}$	R_{1k}
A_2	a_{2j}	$R_{2k}-a_{2j}$	R_{2k}
⋮	⋮	⋮	⋮
A_r	a_{rj}	$R_{rk}-a_{rj}$	R_{rk}
Total	K_{rj}	$T_{rk}-K_{rj}$	T_{rk}

The test-statistic for testing independence of B_j and A_1, A_2, ..., A_r is

$$\chi^2_{0j} = \frac{1}{K_{rj}(T_{rk} - K_{rj})} \sum_{i=1}^{r} \frac{(a_{ij}T_{rk} - R_{ik}K_{rj})^2}{R_{ik}}$$

which, in the null case, has approximately a chi-sguared distribution with (r-1) degrees of freedom. When testing for onlying columns

$$\max_{1 \le j \le k} \chi^2_{0j} = \chi^2_{0j_1}$$

is computed, the j_1-th column of the contingency table is dropped and the procedure is repeated for the resulting (k-1)xr-table.
The critical value for the l-th step is given by c_l, l = 1, 2, ..., k-1, for which one may choose (cf [3]):

$$c_l = \chi^2_{r-1,\frac{\alpha}{k+1-l}} ,$$

which is the upper $\frac{\alpha}{k+1-l}$ - point of the chi-squared distribution with (r-1) degrees of freedom, where α provides an upper bound for the overall significance level of this procedure.

To test for outlying rows consider the kx2-table

	B_1	B_2	B_k	Total
A_i	a_{i1}	a_{i2}	a_{ik}	R_{ik}
$\bar{A}_i$	$K_{r1}-a_{i1}$	$K_{r2}-a_{i2}$	$K_{rk}-a_{ik}$	$T_{rk}-R_{ik}$
Total	K_{r1}	K_{r2}	K_{rk}	T_{rk}

The test-statistic for testing independence of A_i and $B_1, B_2, \ldots, B_k$ is

$$\chi^2_{i0} = \frac{1}{R_{ik}(T_{rk} - R_{ik})} \sum_{j=1}^{k} \frac{(a_{ij}T_{rk} - R_{ik}K_{rj})^2}{K_{rj}}$$

which, in the null case, has approximately a chi-squared distribution with (k-1) degrees of freedom. When testing for outlying rows

$$\chi^2_{i_1 0} = \max_{1 \le i \le r} \chi^2_{i0}$$

is computed, the i_1-th row of the table is dropped and the procedure is repeated for the resulting kx(r-1)-table. The critical value for the l-th step is given by c'_l, l = 1, 2, ..., r-1, for which one may choose

$$c'_l = \chi^2_{k-1,\ \frac{\alpha}{r+1-l}} .$$

<u>Literature</u>

[1] C.R. RAO, Advanced statistical methods in biometric research, Wiley & Sons, New York 1952;

[2] A.W. KIMBALL, Short-cut formulas for the exact partition of χ^2 in contingency tables, Biometrics <u>10</u> (1954), 452-458;

[3] R. DOORNBOS, Statistische methoden voor het aanwijzen van uitbijters, Statist.Neerl. <u>13</u> (1959), 453-462.

Input of the program:

codenumber of the table

r

k

a_{11} - - - - - - a_{1k}

$\vdots$

a_{r1} - - - - - - a_{rk}

the program contains setXEENtal (see note 2).

Output:

{Result <code number of the program>}

{Tests for a contingency table}

{see descr 6}

{input } code number of the table

{chitotal df}

χ^2 $(r-1)(k-1)$

{components of chi - **squared**}

$\chi^2_{2,2}$ - - - $\chi^2_{2,k}$

$\vdots$

$\chi^2_{r,2}$ - - - $\chi^2_{r,k}$

{Outlying columns}

j_1 χ^2_{0,j_1}

$\vdots$

j_{k-1} $\chi^2_{0,j_{k-1}}$

{Outlying rows}

i_1 $\chi^2_{i_1,0}$

$\vdots$

i_{r-1} $\chi^2_{i_{r-1},0}$

TESTS FOR A CONTINGENCY TABLE

Suppose that in each of N identical and independent experiments one of the events $A_1, A_2, \ldots, A_r$ and one of the events $B_1, B_2, \ldots, B_k$ will occur. Let p_{ij} denote the probability that A_i and B_j will occur together in any given experiment, and a_{ij} the total number of occurences of this event A_iB_j ($i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, k$, $\sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} = 1$, $\sum_{i=1}^{r} \sum_{j=1}^{k} a_{ij} = N$). Furthermore the following partial sums are defined:

$$R_{ij} = \sum_{l=1}^{j} a_{il},\quad K_{ij} = \sum_{l=1}^{i} a_{lj},\quad T_{ij} = \sum_{m+1}^{i} \sum_{n=1}^{j} a_{mn};$$

obviously

$$T_{ij} = \sum_{m=1}^{i} R_{mj} = \sum_{n=1}^{j} K_{in} \quad \text{and} \quad T_{rk} = N.$$

The numbers a_{ij} may be given in the form of a k×r-contingency table

	B_1	B_2	B_3		B_k	Total
A_1	a_{11}	a_{12}	a_{13}	----------	a_{1k}	$R_{1,k}$
A_2	a_{21}	a_{22}	a_{23}	----------	a_{2k}	$R_{2,k}$
A_3	a_{31}	a_{32}	a_{33}	----------	a_{3k}	$R_{3,k}$
¦	¦				¦	¦
A_r	a_{r1}	a_{r2}	a_{r3}	----------	a_{rk}	$R_{r,k}$
Total	$K_{r,1}$	$K_{r,2}$	$K_{r,3}$		$K_{r,k}$	$T_{r,k} = N.$

In this program the hypothesis of independence of the A_i and B_j is tested, i.e. $p_{ij} = p_{i.}\, p_{.j}$. If this hypothesis is rejected, two different methods are available to obtain information concerning the type of dependence. With each chi-squared test given, the numbers e5 and e1 of a_{ij} having expectations < 5 and < 1 are given.

A. Chi-squared test for independence

Hypothesis H_0: $p_{ij} = p_{i.} \, p_{.j}$, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, k$.
The test-statistic for testing H_0 is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(a_{ij} - b_{ij})^2}{b_{ij}} \text{, where } b_{ij} = \frac{R_{ik} K_{rj}}{T_{rk}} ;$$

under H_0 it has approximately a chi-squared distribution with $(r - 1)(k - 1)$ degrees of freedom. With χ^2 and $(r - 1)(k - 1)$ the corresponding right-hand tail probability is given in the output.

B. Components of chi-squared [2]

The total chi-squared given under A may be split up into $(r - 1)(k - 1)$ additive components

$$\chi^2_{i+1,j+1} =$$

$$= \frac{T_{r,k}\left[K_{r,j+1}(R_{i+1,k}T_{i,j} - R_{i+1,j}T_{i,k}) - T_{rj}(R_{i+1,k}K_{i,j+1} - a_{i+1,j+1}T_{i,k})\right]^2}{R_{i+1,k}K_{r,j+1}T_{r,j}T_{r,j+1}T_{i,k}T_{i+1,k}} ,$$

$i = 1, 2, \ldots, r-1$, $j = 1, 2, \ldots, k-1$, each of which has under H_0 approximately a chi-squared distribution with one degree of freedom. $\chi^2_{i+1,j+1}$ is associated with the 2×2-table

	$B_1 v B_2 v \ldots v B_j$	B_{j+1}	Total
$A_1 v A_2 v \ldots v A_i$	$T_{i,j}$	$K_{i,j+1}$	$T_{i,j+1}$
A_{i+1}	$R_{i+1,j}$	$a_{i+1,j+1}$	$R_{i+1,j+1}$
Total	$T_{i+1,j}$	$K_{i+1,j+1}$	$T_{i+1,j+1}$

$$\text{and} \quad \chi^2 = \sum_{i=1}^{r-1} \sum_{j=1}^{k-1} \chi^2_{i+1,j+1} .$$

C. Tests for outlying columns and rows

For $j = 1, 2, \ldots, k$ consider the 2×r-table

	B_j	$\bar{B}_j$	Total
A_1	a_{1j}	$R_{1k}-a_{1j}$	R_{1k}
A_2	a_{2j}	$R_{2k}-a_{2j}$	R_{2k}
⋮	⋮	⋮	⋮
A_r	a_{rj}	$R_{rk}-a_{rj}$	R_{rk}
Total	K_{rj}	$T_{rk}-K_{rj}$	T_{rk}

The test-statistic for testing independence of B_j and $A_1, A_2, \ldots, A_r$ is

$$\chi^2_{0j} = \frac{1}{K_{rj}(T_{rk} - K_{rj})} \sum_{i=1}^{r} \frac{(a_{ij}T_{rk} - R_{ik}K_{rj})^2}{R_{ik}}$$

which, in the null case, has approximately a chi-squared distribution with (r − 1) degrees of freedom. When testing for outlying columns

$$\max_{1 \leq j \leq k} \chi^2_{0j} = \chi^2_{0j_1}$$

is computed.

Let α be the given significance level. Then for not too small sample sizes

$$P\{\chi^2_{0j_1} > x_{r-1}\left(\frac{\alpha}{k}\right)\} \leq \alpha$$

where $x_{r-1}\left(\frac{\alpha}{k}\right)$ is the upper $\frac{\alpha}{k}$ percentage point of the χ^2 distribution with r−1 degrees of freedom.

Dropping the j_1-th column of the contingency table the procedure is repeated for the resulting $(k - 1)\times r$ table.

In the l-th step $\chi^2_{0j_l}$ and the right-hand tail probability

$$P_l = P\{\underline{\chi}^2_{r-1} \geq \chi^2_{0j_l}\}$$

are given in the output, where $\underline{\chi}^2_{r-1}$ has a χ^2 distribution with r-1 degrees of freedom.

This procedure is continued until $P_l \geq \frac{\alpha}{k - l + 1}$.

To test for outlying rows consider the k×2-table

	B_1	B_2	B_k	Total
A_i	a_{i1}	a_{i2}	a_{ik}	R_{ik}
$\bar{A}_i$	$K_{r1}-a_{i1}$	$K_{r2}-a_{i2}$	$K_{rk}-a_{ik}$	$T_{rk}-R_{ik}$
Total	K_{r1}	K_{r2}	K_{rk}	T_{rk}

The test-statistic for testing independence of A_i and $B_1, B_2, \ldots, B_k$ is

$$\chi^2_{i0} = \frac{1}{R_{ik}(T_{rk}-R_{ik})} \sum_{j=1}^{k} \frac{(a_{ij}T_{rk}-R_{ik}K_{rj})^2}{K_{rj}}$$

which, in the null case, has approximately a chi-squared distribution with $(k - 1)$ degrees of freedom. When testing for outlying rows

$$\chi^2_{i_1 0} = \max_{1 \leq i \leq r} \chi^2_{i0}$$

is computed.
Let α be the given significance level again. Then for not too small sample sizes

$$P\{\chi^2_{i,0} > x_{k-1}(\frac{\alpha}{r})\} \leq \alpha$$

where $x_{k-1}(\frac{\alpha}{r})$ is the upper $\frac{\alpha}{r}$ percentage point of the χ^2 distribution with k-1 degrees of freedom.

Dropping the i_1-th row of the contingency table the procedure is repeated for the resulting k×(r-1) table.

In the l-th step $\chi^2_{i_l 0}$ and the right-hand tail probability

$$P_l = P\{\underline{\chi}^2_{k-1} \geq \chi^2_{i_l 0}\}$$

are given in the output, where $\underline{\chi}^2_{k-1}$ has a χ^2 distribution with k-1 degrees of freedom.
This procedure is continued until $P_l \geq \frac{\alpha}{r - l + 1}$.

Literature

[1] C.R. RAO, Advanced statistical methods in biometric research, Wiley & Sons, New York 1952;

[2] A.W. KIMBALL, Short-cut formulas for the exact partition of χ^2 in contingency tables, Biometrics 10 (1954), 452-458;

[3] R. DOORNBOS, Statistische methoden voor het aanwijzen van uitbijters, Statist. Neerl. 13 (1959), 453-462.

Input of the program:

codenumber of the table

r

k

a_{11} - - - - - - - - a_{1k}

|

|

|

a_{r1} - - - - - - - - a_{rk}

α

the program contains setXEENtal (see note 2).

Output:

{Result <code number of the program>}

{Tests for a contingency table}

{see descr 6a}

{input }   code number of the table

{chitotal   df   P }

χ^2 $(r-1)(k-1)$ P_{χ^2}

{e1   e5}

e1   e5

{components of chi-squared}

$\chi^2_{2,2}$ $\chi^2_{2,k}$

$\chi^2_{r,2}$ $\chi^2_{r,k}$

{Outlying columns}

j_1 χ^2_{0,j_1} P_1

⋮   ⋮   ⋮

{Outlying rows}

i_1 $\chi^2_{i_1,0}$ P_1

⋮   ⋮   ⋮

ONE-SAMPLE TESTS FOR THE MEAN (ONE-DIMENSIONAL CASE)

A sample of n independent observations $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ is supposed to be given. These observations may be given directly or arise as differences between pairs of observations $\underline{x}_i$ and $\underline{y}_i$; in the latter case $\underline{z}_i = \underline{x}_i - \underline{y}_i$ $(i = 1, \ldots, n)$.

In this program the hypothesis is tested, that the observations $\underline{z}_i$ have distributions symmetric with respect to μ_0. Two tests are presented. In both cases the tests are especially powerful against shift-alternatives: $\mathcal{E}\underline{z}_i = \mu_i \neq \mu_0$.

A. STUDENT's t-test

Hypothesis H_0: the observations $\underline{z}_i$ are normally distributed with mean value μ_0 and unknown but equal variance.
The test statistic for testing H_0 is:

$$\underline{t} = \left[\sum_{i=1}^{n} (\underline{z}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{z}_j)^2 \right]^{-\frac{1}{2}} \cdot \sqrt{n(n-1)} \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \underline{z}_i - \mu_0 \right] , \qquad (1)$$

which under H_0 has a t-distribution with n-1 degrees of freedom.

B. WILCOXON's test for symmetry (signed rank test)

Hypothesis H_0: the observations $\underline{z}_i$ are symmetrically distributed with respect to μ_0. The test statistic $\underline{T}$ for testing H_0 is defined as follows. Omit all observations $\underline{z}_i$ equal to μ_0; let n_1 be the number of remaining observations. Rank the absolute values of the $\underline{z}_i - \mu_0$. Then $\underline{T}$ is the sum of the signed ranks of the $\underline{z}_i - \mu_0$.

For small values of n_1, $n_1 = 3(1)20$, one-sided tail-probabilities of $\underline{T}$ under H_0 have been tabulated, if no ties are present. For large n_1 $\underline{T}$ is approximately normally distributed under H_0 with expected value 0 and variance

$$\sigma_T^2 = \frac{1}{6} n_1(n_1 + 1)(2n_1 + 1) - \frac{1}{12} (D - n_1) , \tag{2}$$

where $D = \sum_j t_j^3$ and t_j is the size of the j^{th} tie of absolute values of the $z_i - \mu_0$ ($t_j = 1$ inclusive).

For large values of n_1, the statistics

$$T_l = \frac{1}{\sigma_T} (T + 1) \qquad \text{(one-sided test: left hand tail)} \tag{3}$$

$$T_r = \frac{1}{\sigma_T} (T - 1) \qquad \text{(one-sided test: right hand tail)} \tag{4}$$

$$T_2 = \frac{1}{\sigma_T} (|T| - 1) \qquad \text{(two-sided test)} \tag{5}$$

can be considered as random normal deviates and can be used to find approximate tail-probabilities in the cases of one-sided or two-sided tests respectively.

Large positive (negative) values of t and T indicate that $\mu_i > \mu_0$ ($\mu_i < \mu_0$).

Remark: In this program the test statistic T is computed as follows.

Split the $z_i - \mu_0$ unequal to zero in two groups: the positive $z_i - \mu_0$ and the negative $z_i - \mu_0$. Let N_1 be the number of positive $z_i - \mu_0$ and N_2 the number of negative $z_i - \mu_0$. Then for $N_1 \neq 0$ and $N_2 \neq 0$ we have

$$T = W - N_1 N_2 + \frac{1}{2} (N_1 + N_2 + 1)(N_1 - N_2) , \tag{6}$$

where W is the test statistic of WILCOXON's two-sampled test applied to the samples of positive $z_i - \mu_0$ (sample I) and the absolute value of the negative $z_i - \mu_0$ (sample II).

Literature

In most books on elementary statistics details of STUDENT'- t-test can be found, e.g.

W.J. DIXON & F.A. MASSEY, Jr., Introduction to statistical analysis (Chapter 9), Mc Graw-Hill, New York 1951;

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 12), Mc Graw-Hill, New York, 1963.

Discussions of WILCOXON's test for symmetry are given in

A. BENARD & C. VAN EEDEN, Handleiding voor de symmetrietoets van Wilcoxon, Report S 208 of the Statistical Department of the Mathematical Centre, Amsterdam, 1956;

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden, 1963;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 5), Mc Graw-Hill, New York, 1956.

Input:

a   1-dimensional set of observations, given as a sorted frequency table (see note 3)

μ_0

Output:

{Result <code number of program>}
{One-sample test for the mean}
{see descr 7 }

{input	t	df	T	sigT	T1	Tr	T2 }
code of sample	t	n - 1	T	σ_T	T_1	T_r	T_2

The program contains set XEENtal, the text of the output is given once, the numerical results of the several samples are tabulated below this text.

ONE-SAMPLE TESTS FOR THE MEAN (ONE-DIMENSIONAL CASE)

A sample of n independent observations $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ is supposed to be given. These observations may be given directly or arise as differences between pairs of observations $\underline{x}_i$ and $\underline{y}_i$; in the latter case $\underline{z}_i = \underline{x}_i - \underline{y}_i$ $(i = 1, \ldots, n)$.

In this program the hypothesis is tested, that the observations $\underline{z}_i$ have distributions symmetric with respect to μ_0. Two tests are presented. In both cases the tests are especially powerful against shift-alternatives: $\mathcal{E}\, \underline{z}_i = \mu_i \neq \mu_0$.

A. STUDENT's t-test

Hypothesis H_0: the observations $\underline{z}_i$ are normally distributed with mean value μ_0 and unknown but equal variance.
The test statistic for testing H_0 is:

$$\underline{t} = \left[\sum_{i=1}^{n} \left(\underline{z}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{z}_j\right)^2 \right]^{-\frac{1}{2}} \cdot \sqrt{n(n-1)} \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \underline{z}_i - \mu_0 \right], \qquad (1)$$

which under H_0 has a t-distribution with n-1 degrees of freedom.
With t and n-1 the corresponding right-hand tail probability is given in the output.

B. WILCOXON's test for symmetry (signed rank test)

Hypothesis H_0: the observations $\underline{z}_i$ are symmetrically distributed with respect to μ_0. The test statistic $\underline{T}$ for testing H_0 is defined as follows. Omit all observations $\underline{z}_i$ equal to μ_0; let n_1 be the number of remaining observations. Rank the absolute values of the $\underline{z}_i - \mu_0$. Then $\underline{T}$ is the sum of the signed ranks of the $\underline{z}_i - \mu_0$.

For small values of n_1, $n_1 = 3(1)20$, one-sided tail probabilities of $\underline{T}$ under H_0 have been tabulated, if no ties are present. For large

n_1 $\underline{T}$ is approximately normally distributed under H_0 with expected value 0 and variance

$$\sigma_T^2 = \frac{1}{6} n_1(n_1 + 1)(2n_1 + 1) - \frac{1}{12} (D - n_1), \qquad (2)$$

where $D = \sum_j t_j^3$ and t_j is the size of the j^{th} tie of absolute values of the $\underline{z}_i - \mu_0$ ($t_j = 1$ inclusive).

For large values of n_1, the statistics

$$\underline{T}_l = \frac{1}{\sigma_T} (\underline{T} + 1) \qquad \text{(one-sided test: left hand tail)} \qquad (3)$$

$$\underline{T}_r = \frac{1}{\sigma_T} (\underline{T} - 1) \qquad \text{(one-sided test: right hand tail)} \qquad (4)$$

$$\underline{T}_2 = \frac{1}{\sigma_T} (|\underline{T}| - 1) \qquad \text{(two-sided test)} \qquad (5)$$

can be considered as random normal deviates and can be used to find approximate tail-probabilities in the cases of one-sided or two-sided tests respectively. With T_l, T_r and T_2 the corresponding normally approximated probabilities are given in the output.

Large positive (negative) values of $\underline{t}$ and $\underline{T}$ indicate that $\mu_i > \mu_0$ ($\mu_i < \mu_0$).

Remark: In this program the test statistic $\underline{T}$ is computed as follows. Split the $\underline{z}_i - \mu_0$ unequal to zero in two groups: the positive $\underline{z}_i - \mu_0$ and the negative $\underline{z}_i - \mu_0$. Let $\underline{N}_1$ be the number of positive $\underline{z}_i - \mu_0$ and $\underline{N}_2$ the number of negative $\underline{z}_i - \mu_0$. Then for $\underline{N}_1 \neq 0$ and $\underline{N}_2 \neq 0$ we have

$$\underline{T} = \underline{W} - \underline{N}_1\underline{N}_2 + \frac{1}{2} (\underline{N}_1 + \underline{N}_2 + 1)(\underline{N}_1 - \underline{N}_2), \qquad (6)$$

where $\underline{W}$ is the test statistic of WILCOXON's two-sampled test applied to the samples of positive $\underline{z}_i - \mu_0$ (sample I) and the absolute value of the negative $\underline{z}_i - \mu_0$ (sample II).

Literature

In most books on elementary statistics details of STUDENT's t-test can be found, e.g.

W.J. DIXON & F.A. MASSEY, Jr., Introduction to statistical analysis (Chapter 9), Mc Graw-Hill, New York 1951;

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 12), Mc Graw-Hill, New York, 1963.

Discussions of WILCOXON's test for symmetry are given in

A. BENARD & C. VAN EEDEN, Handleiding voor de symmetrietoets van Wilcoxon, Report S 208 of the Statistical Department of the Mathematical Centre, Amsterdam, 1956;

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden, 1963;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 5), Mc Graw-Hill, New York, 1956.

Input:

a 1-dimensional set of observations, given in one of the two forms described in note 3.

μ_0

tests = { 1: A performed
          2: A and B performed

The program contains set XEENtal.

Output:

```
⊰Result <code number of program>⊱
⊰One-sample tests for the mean, see descr 7a⊱
⊰sample⊱  code number of the sample
⊰mu0   t   df   P   T   sigT   T1   P   Tr   P   T2   P ⊱
```

mu0	t	df	P	T	sigT	T1	P	Tr	P	T2	P
μ_0	t	$n-1$	P_t	T	σ_T	T_1	P_{T_1}	T_r	P_{T_r}	T_2	P_{T_2}

TESTS FOR INDEPENDENCE (TWO-DIMENSIONAL CASE)

A two-dimensional sample of pairs of observations $(\underline{x}_i, \underline{y}_i)$, $i = 1, \ldots, n$, is supposed to be given.

In this program the hypothesis is tested, that the observations $\underline{x}_i$ and $\underline{y}_i$ are stochastically independent. Three tests are presented. They are especially powerful against alternatives where the correlation coefficient $\rho(\underline{x}, \underline{y}) \neq 0$.

In the first two tests it is assumed, that the pairs of observations are independent. In the last test no such assumption is necessary.

A. Normal-distribution test for independence

Hypothesis H_0: the pairs of observations $(\underline{x}_i, \underline{y}_i)$ have identical two-dimensional normal distributions, and the components are independent. Equivalent test statistics for testing H_0 are

$$\underline{r} = \left[\sum_{i=1}^{n} (\underline{x}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{x}_j)^2 \right]^{-\frac{1}{2}} \cdot \left[\sum_{i=1}^{n} (\underline{y}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{y}_j)^2 \right]^{-\frac{1}{2}} \cdot$$

$$\cdot \sum_{i=1}^{n} (\underline{x}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{x}_j)(\underline{y}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{y}_j) \qquad (1)$$

and

$$\underline{t} = \left[1 - \underline{r}^2 \right]^{-\frac{1}{2}} \cdot \sqrt{n - 2} \cdot \underline{r} \,. \qquad (2)$$

Critical values of the sample correlation coefficient $\underline{r}$ under H_0 are tabulated for small n. The statistic $\underline{t}$ has under H_0 a t-distribution with n-2 degrees of freedom.

B. SPEARMAN's rank correlation test

Hypothesis H_0: the pairs of observations $(\underline{x}_i, \underline{y}_i)$ have identical two-dimensional distributions, and the components are independent. Equivalent test statistics $\underline{r}_s$ and $\underline{t}_s$ for testing H_0 are defined as follows.

Rank the $\underline{x}_i$ and $\underline{y}_i$ separately, and let $\underline{X}_i$ be the rank of $\underline{x}_i$, and $\underline{Y}_i$ the rank of $\underline{y}_i$. Define $\underline{d}_i = \underline{X}_i - \underline{Y}_i$, i = 1, ..., n. Let $t_\nu(x)$ = number of x-observations in the ν^{th} tie of the x's, and define $T_\nu(x) = \frac{1}{12}(t_\nu^3(x) - t_\nu(x))$; $T_\lambda(y)$ is defined similarly for the y-observations. Then

$$\underline{r}_s = \frac{1}{2}\left[\sum_{i=1}^{n} \underline{X}_i^{*2} - \sum_\nu T_\nu(x)\right]^{-\frac{1}{2}} \cdot \left[\sum_{i=1}^{n} \underline{Y}_i^{*2} - \sum_\lambda T_\lambda(y)\right]^{-\frac{1}{2}} \cdot \left[\sum_{i=1}^{n} \underline{X}_i^{*2} + \sum_{i=1}^{n} \underline{Y}_i^{*2} - \sum_\nu T_\nu(x) - \sum_\lambda T_\lambda(y) - \sum_{i=1}^{n} \underline{d}_i^2\right] \quad (3)$$

and

$$\underline{t}_s = \left[1 - \underline{r}_s^2\right]^{-\frac{1}{2}} \cdot \sqrt{n-2} \cdot \underline{r}_s \quad , \quad (4)$$

where $\underline{X}_i^* = \underline{X}_i - \frac{1}{n}\sum_{j=1}^{n} \underline{X}_j$ and $\underline{Y}_i^* = \underline{Y}_i - \frac{1}{n}\sum_{j=1}^{n} \underline{Y}_j$.

For small n, n = 2(1)30, critical values of SPEARMAN's rank correlation coefficient $\underline{r}_s$ are tabulated (if no ties are present). For large n the statistic $\underline{t}_s$ has under H_0 approximately a t-distribution with n-2 degrees of freedom.

C. KENDALL's rank correlation test

Hypothesis H_0: for a given ranking of the x-observations all permutations of the y-observations are equally probable.

If the pairs of observations $(\underline{x}_i, \underline{y}_i)$, i = 1, ..., n, are independent, the hypothesis H_0 is implied by the hypothesis H_0'.

Hypothesis H_0': the pairs of observations $(\underline{x}_i, \underline{y}_i)$ are identically distributed, and the components are independent.

The test statistic for testing H_0 (or H_0') is:

$$\underline{S} = \sum_{i<j} \underline{s}_{i,j} \qquad (5)$$

where

$$s_{i,j} = \begin{cases} +1 & \text{if} \quad (x_i - x_j)(y_i - y_j) > 0 \\ 0 & \text{if} \quad (x_i - x_j)(y_i - y_j) = 0 \\ -1 & \text{if} \quad (x_i - x_j)(y_i - y_j) < 0. \end{cases}$$

For small n, n = 4(1)40, one-sided critical values of $\underline{S}$ under H_0 are tabulated (if no ties are present). For large n, $\underline{S}$ is approximately normally distributed under H_0 with expected value 0 and variance

$$\sigma_S^2 = \left[18n\ (n-1)(n-2) \right]^{-1} \cdot \left[2\{n^3 - D_1 - 3(n^2 - C_1)\} \cdot \right.$$
$$\left. \cdot \{n^3 - D_2 - 3(n^2 - C_2)\} + 9(n-2)(n^2 - C_1)(n^2 - C_2) \right], \qquad (6)$$

where

$C_1 = \sum_{\nu} t_\nu^2(x)$, $D_1 = \sum_{\nu} t_\nu^3(x)$, $t_\nu(x)$ = number of x-observations in the ν^{th} tie of the x's, and

$C_2 = \sum_{\lambda} t_\lambda^2(y)$, $D_2 = \sum_{\lambda} t_\lambda^3(y)$, $t_\lambda(y)$ = number of y-observations in the λ^{th} tie of the y's.

For large values of n the statistics

$$\underline{S}_l = \frac{1}{\sigma_s} (\underline{S} - 1) \qquad \text{(one-sided test: left hand tail)} \quad (7)$$

$$\underline{S}_r = \frac{1}{\sigma_s} (\underline{S} + 1) \qquad \text{(one-sided test: right hand tail)} \quad (8)$$

$$\underline{S}_2 = \frac{1}{\sigma_s} (|\underline{S}| - 1) \qquad \text{(two-sided test)} \quad (9)$$

can be considered as random normal deviates and can be used to find approximate tail-probabilities in the cases of one-sided or two-sided tests respectively.

KENDALL's rank correlation coefficient $\underline{\tau}$ is defined as

$$\underline{\tau} = 2 \left[(n^2 - C_1)(n^2 - C_2) \right]^{-\frac{1}{2}} \cdot \underline{S}. \qquad (10)$$

Large positive (negative) values of $\underline{r}$, $\underline{t}$, $\underline{r}_s$, $\underline{t}_s$ and $\underline{S}$ (in case of H_0') indicate, that $\underline{x}$ and $\underline{y}$ are positively (negatively) correlated.

Remarks: 1) The statistic $\underline{r}_s$ is computed using the fact that

$$\sum_{i=1}^{n} X_i^{*2} = \sum_{i=1}^{n} Y_i^{*2} = \frac{n^3 - n}{12} .$$

2) In this program the statistic $\underline{S}$ is computed from the relation

$$\underline{S} = \sum_{q<p} \underline{W}_{pq}, \qquad (11)$$

where $\underline{W}_{pq}$ is the test statistic of WILCOXON's two-sample test applied on the samples "p" (as sample I) and "q" (as sample II); here p and q denote the ranks of y-observations (not necessarily integers), the sample "p" consists of all x-observations corresponding to y-observations with rank p, and the sample "q" consists of all x-observations corresponding to y-observations with rank q.

Literature

In many books on mathematical statistics the distribution of the sample correlation coefficient is derived, e.g.

H. CRAMÉR, Mathematical methods of statistics, Chapter 29, Princeton University Press, Princeton 1946.

The rank correlation tests of SPEARMAN and KENDALL are discussed in

C. VAN EEDEN, Handleiding voor de rangcorrelatietoets van Kendall, Report S 262 of the Statistical Department of the Mathematical Centre, Amsterdam 1959;

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden 1963;

M.G. KENDALL, Rank correlation methods, C. Griffin & Company, London 1955;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 9), Mc Graw-Hill, New York 1956.

Input:

there are two ways of input, not yet in accordance with note .3.

1) code number of the sample

n

x_1 y_1

$\vdots$

x_n y_n

2) code number of the sample

$-m$

x_1 y_1 f_1

$\vdots$

x_m y_m f_n

where $x_i \leq x_j$ if $i < j$

$y_i < y_j$ if $x_i = x_j$ $i < j$

f_i = the frequency of observation (x_i, y_i)

m = the number of different observations

TESTS FOR INDEPENDENCE (TWO-DIMENSIONAL CASE)

A two-dimensional sample of pairs of observations $(\underline{x}_i, \underline{y}_i)$ $i = 1, \ldots, n$, is supposed to be given.

In this program the hypothesis is tested, that the observations $\underline{x}_i$ and $\underline{y}_i$ are stochastically independent. Three tests are presented. They are especially powerful against alternatives where the correlation coefficient $\rho(\underline{x}, \underline{y}) \neq 0$.

In the first two tests it is assumed, that the pairs of observations are independent. In the last test no such assumption is necessary.

A. Normal-distribution test for independence

Hypothesis H_0: the pairs of observations $(\underline{x}_i, \underline{y}_i)$ have identical two-dimensional normal distributions, and the components are independent. Equivalent test statistics for testing H_0 are

$$\underline{r} = \left[\sum_{i=1}^{n} \left(\underline{x}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{x}_j\right)^2 \right]^{-\frac{1}{2}} \cdot \left[\sum_{i=1}^{n} \left(\underline{y}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{y}_j\right)^2 \right]^{-\frac{1}{2}} \cdot \sum_{i=1}^{n} \left(\underline{x}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{x}_j\right)\left(\underline{y}_i - \frac{1}{n} \sum_{j=1}^{n} \underline{y}_j\right) \quad (1)$$

and

$$\underline{t} = \left[1 - \underline{r}^2 \right]^{-\frac{1}{2}} \cdot \sqrt{n - 2} \cdot \underline{r}. \quad (2)$$

Critical values of the sample correlation coefficient $\underline{r}$ under H_0 are tabulated for small n. The statistic $\underline{t}$ has under H_0 a t-distribution with n-2 degrees of freedom.
With t and n-2 the corresponding right-hand tail probability is given in the output.

B. SPEARMAN's rank correlation test

Hypothesis H_0: the pairs of observations $(\underline{x}_i, \underline{y}_i)$ have identical two-dimensional distributions, and the components are independent.
Equivalent test statistics $\underline{r}_s$ and $\underline{t}_s$ for testing H_0 are defined as follows.
Rank the $\underline{x}_i$ and $\underline{y}_i$ separately, and let $\underline{X}_i$ be the rank of $\underline{x}_i$, and $\underline{Y}_i$ the rank of $\underline{y}_i$. Define $\underline{d}_i = \underline{X}_i - \underline{Y}_i$, $i = 1, \ldots, n$. Let $t_\nu(x)$ = number of x-observations in the ν^{th} tie of the x's, and define
$T_\nu(x) = \frac{1}{12}(t_\nu^3(x) - t_\nu(x))$; $T_\lambda(y)$ is defined similarly for the y-observations. Then

$$\underline{r}_s = \frac{1}{2}\left[\sum_{i=1}^{n} \underline{X}_i^{*2} - \sum_\nu T_\nu(x)\right]^{-\frac{1}{2}} \cdot \left[\sum_{i=1}^{n} \underline{Y}_i^{*2} - \sum_\lambda T_\lambda(y)\right]^{-\frac{1}{2}} \cdot$$

$$\cdot \left[\sum_{i=1}^{n} \underline{X}_i^{*2} + \sum_{i=1}^{n} \underline{Y}_i^{*2} - \sum_\nu T_\nu(x) - \sum_\lambda T_\lambda(y) - \sum_{i=1}^{n} \underline{d}_i^2\right] \qquad (3)$$

and

$$\underline{t}_s = \left[1 - \underline{r}_s^2\right]^{-\frac{1}{2}} \cdot \sqrt{n-2} \cdot \underline{r}_s, \qquad (4)$$

where $\underline{X}_i^{*} = \underline{X}_i - \frac{1}{n}\sum_{j=1}^{n} \underline{X}_j$ and $\underline{Y}_i^{*} = \underline{Y}_i - \frac{1}{n}\sum_{j=1}^{n} \underline{Y}_j$.

For small n, n = 2(1)30, critical values of SPEARMAN's rank correlation coefficient $\underline{r}_s$ are tabulated (if no ties are present). For large n the statistic $\underline{t}_s$ has under H_0 approximately a t-distribution with n-2 degrees of freedom.
With t_s and n-2 the corresponding approximate right-hand tail probability is given in the output.

C. KENDALL's rank correlation test

Hypothesis H_0: for a given ranking of the x-observations all permutations of the y-observations are equally probable.
If the pairs of observations $(\underline{x}_i, \underline{y}_i)$, $i = 1, \ldots, n$, are independent, the

hypothesis H_0 is implied by the hypothesis H_0'.
Hypothesis H_0': the pairs of observations $(\underline{x}_i, \underline{y}_i)$ are identically distributed, and the components are independent.
The test statistic for testing H_0(or H_0') is:

$$\underline{S} = \sum_{i<j} \underline{s}_{i,j} \tag{5}$$

where

$$s_{i,j} = \begin{cases} +1 & \text{if} \quad (x_i - x_j)(y_i - y_j) > 0 \\ 0 & \text{if} \quad (x_i - x_j)(y_i - y_j) = 0 \\ -1 & \text{if} \quad (x_i - x_j)(y_i - y_j) < 0. \end{cases}$$

For small n, n = 4(1)40, one-sided critical values of $\underline{S}$ under H_0 are tabulated (if no ties are present). For large n, $\underline{S}$ is approximately normally distributed under H_0 with expected value 0 and variance

$$\sigma_S^2 = \left[18n\ (n-1)(n-2) \right]^{-1} \cdot \left[2\{n^3 - D_1 - 3(n^2 - C_1)\} \cdot \right.$$

$$\left. \cdot \{n^3 - D_2 - 3(n^2 - C_2)\} + 9(n-2)(n^2 - C_1)(n^2 - C_2) \right], \tag{6}$$

where

$C_1 = \sum_\nu t_\nu^2(x)$, $D_1 = \sum_\nu t_\nu^3(x)$, $t_\nu(x)$ = number of x-observations in the ν^{th} tie of the x's, and

$C_2 = \sum_\lambda t_\lambda^2(y)$, $D_2 = \sum_\lambda t_\lambda^3(y)$, $t_\lambda(y)$ = number of y-observations in the λ^{th} tie of the y's.

For large values of n the statistics

$$\underline{S}_l = \frac{1}{\sigma_s}(\underline{S} - 1) \qquad \text{(one-sided test: left hand tail)} \tag{7}$$

$$\underline{S}_r = \frac{1}{\sigma_s}(\underline{S} + 1) \qquad \text{(one-sided test: right hand tail)} \tag{8}$$

$$\underline{S}_2 = \frac{1}{\sigma_s}(\ \underline{S} - 1) \qquad \text{(two-sided test)} \tag{9}$$

can be considered as random normal deviates and can be used to find approximate tail-probabilities in the cases of one-sided or two-sided tests respectively.

With S_1, S_r and S_2 the corresponding approximate tail probabilities are given in the output.

KENDALL's rank correlation coefficient τ is defined as

$$\tau = 2 \left[(n^2 - C_1)(n^2 - C_2) \right]^{-\frac{1}{2}} \cdot S. \qquad (10)$$

Large positive (negative) values of r, t, r_s, t_s and S (in case of H_0') indicate, that x and y are positively (negatively) correlated.

Remarks: 1) The statistic r_s is computed using the fact that

$$\sum_{i=1}^{n} X_i^{*2} = \sum_{i=1}^{n} Y_i^{*2} = \frac{n^3 - n}{12} \quad .$$

2) In this program the statistic S is computed from the relation

$$S = \sum_{q<p} \{W_{p,q} - t_p t_q\} \quad , \qquad (11)$$

where $W_{p,q}$ is the test statistic of WILCOXON's two sample test applied to the samples "p" (as sample I) and "q" (as sample II).

The sample "p" ("q") consists of the y_i's determined by the p-th (q-th) tie of the x_i's. t_p (t_q) is the size of the p-th (q-th) tie of the x_i's.

Literature

In many books on mathematical statistics the distribution of the sample correlation coefficient is derived, e.g.

H. CRAMÉR, Mathematical Methods of statistics, Chapter 29, Princeton University Press, Princeton 1946.

The rank correlation tests of SPEARMAN and KENDALL are discussed in

C. VAN EEDEN, Handleiding voor de rangcorrelatietoets van Kendall, Report S 262 of the Statistical Department of the Mathematical Centre, Amsterdam 1959;

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden 1963;

M.G. KENDALL, Rank correlation methods, C. Griffin & Company, London 1955;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 9), Mc Graw-Hill, New York 1956.

Input:

a 2-dimensional set of observations, in one of the two forms described in note 3,

tests = { 1: A performed; 2: A, B and C performed

The program contains XEENtal, see note 2.

Output:

{Result <code number of program>}
{Tests for independence, see descr 8a}
{sample} code number of the sample

{r	t	df	P	rS	tS	df	P	S	sigS	tau}
r	t	$n-2$	P_t	r_S	t_S	$n-2$	P_{t_S}	S	σ_S	τ

{S1	P	Sr	P	S2	P }
S_1	P_{s_1}	S_r	P_{s_r}	S_2	P_{s_2}

Output:

⊀<code number of program > ingelezen ⊁ code of sample

⊀Normale toets ⊁

⊀ r ⊁	r
⊀ t ⊁	t
⊀ vg ⊁	n - 2

⊀toets van Spearman ⊁

⊀ rhos ⊁	r_s
⊀ t ⊁	t_s
⊀ vg ⊁	n - 2

⊀toets van Kendall ⊁

⊀ S ⊁	S
⊀ tau ⊁	τ
⊀ sigma ⊁	σ_T
⊀ Srechts ⊁	S_r
⊀ Slinks ⊁	S_1
⊀ Stweez ⊁	S_2

the program contains set XEENtal.
To compute the ranks of the observations a very slow sort procedure is used.

THE METHOD OF m RANKINGS (ONE k-DIMENSIONAL SAMPLE)

A sample of m independent k-dimensional observations $(\underline{x}_{11}, \ldots, \underline{x}_{1k})$, ..., $(\underline{x}_{m1}, \ldots, \underline{x}_{mk})$, not necessarily identically distributed, is supposed to be given. Such a sample can arise if m judges independently rank an ordered set of k elements, or if m subjects are studied under k conditions (in this latter case one often uses k matched samples of m subjects each). In most applications the situation is analogous to two-way analysis of variance (without interaction).

In this program the hypothesis is tested, that the simultaneous distribution of the k components is symmetrical in the components. We present three rank tests, of which the first has reasonable power against alternatives where the marginal distributions of the components are shifted with respect to each other, and the last two are especially powerful against trend-alternatives.

A. FRIEDMAN's test

The components of each k-dimensional observation are ranked separately.
Hypothesis H_0: in each observation all permutations of the ranks are equally probable.
If the k components of each observation are independent, the hypothesis H_0 is implied by the hypothesis H_0'.
Hypothesis H_0': in each observation the k components are identically distributed (their distributions may differ for different observations).
Several closely connected statistics for testing H_0 (or H_0') are commonly used.
FRIEDMAN's test statistic is:

$$\underline{X}^2 = 12 \left[m(k^2 + k) \right]^{-1} \circ \left[\sum_{j=1}^{k} \underline{R}_j^2 - \frac{1}{4} m^2 k(k+1)^2 \right], \qquad (1)$$

where

$\underline{R}_j$ = sum of the ranks of the j^{th} component over the m observations.
If no ties are present, the exact one-sided tail probabilities of χ^2 under H_0 are tabulated for k = 3, m = 2(1)9 and k = 4, m = 2(1)4.
KENDALL and SMITH considered the test statistic

$$\underline{S} = \sum_{j=1}^{k} \underline{R}_j^2 - \frac{1}{4} m^2 k(k+1)^2 \qquad (2)$$

and tabulated one-sided tail probabilities for k = 3, m = 2(1)10; k = 4, m = 2(1)6 and k = 5, m = 3 (if no ties are present).
KENDALL introduced a coefficient of concordance "between the judges", which is defined as (with tie correction)

$$\underline{W} = \left[\frac{1}{12} m^2 (k^3 - k) - m \sum_i T_i\right]^{-1} \cdot \underline{S}\ , \qquad (3)$$

where

$T_i = \frac{1}{12} \sum_l t_{il} (t_{il}^2 - 1)$, i = 1, ..., m, and t_{il} is the size of the l^{th} tie in the components of the i^{th} observation.
For large m the test statistic

$$\underline{\chi}^2 = \begin{cases} \left[\frac{1}{12} m^2 (k^3 - k) + 2\right]^{-1} \cdot m(k-1)(\underline{S} - 1) & \text{if no ties are present} \\ \left[\frac{1}{12} m^2 (k^3 - k) - m \sum_i T_i\right]^{-1} \cdot m(k-1)\, \underline{S} & \text{otherwise} \end{cases} \qquad (4)$$

has under H_0 approximately a χ^2-distribution with k-1 degrees of freedom.
Also for large m, the test statistic

$$\underline{F} = \left[1 - \frac{1}{m(k-1)} \underline{\chi}^2\right]^{-1} \circ \frac{m-1}{m(k-1)} \underline{\chi}^2\ , \qquad (5)$$

where $\underline{\chi}^2$ is defined in (4),
has under H_0 approximately an F-distribution with ν_1 and ν_2 degrees of freedom; ν_1 and ν_2 are given by

$$\nu_1 = \frac{1}{2}(k-1)(1-\frac{1}{m}) \circ \left[\sum_{i<l} \mu_i \mu_l \right]^{-1} \circ \left[\sum_{i=1}^{m} \mu_i \right]^2 - \frac{2}{m} \qquad (6)$$

$$\nu_2 = (m-1)\nu_1 \; , \qquad (7)$$

where $\mu_i = \frac{1}{12}(k^2-1) - \frac{1}{k} T_i$.

For moderate values of m and k the F-approximation is reputed to be somewhat better than the χ^2-approximation.

B. VAN ELTEREN's test for a trend

Hypotheses H_0 and H_0': as in FRIEDMAN's test.

If a test is wanted with large power against the class of alternative hypotheses characterized by $\mathcal{E}\underline{x}_{ij_1} < \mathcal{E}\underline{x}_{ij_2} < \ldots < \mathcal{E}\underline{x}_{ij_k}$, $i = 1, \ldots, m$, where $(j_1, \ldots, j_k)$ is a given permutation of $(1, 2, \ldots, k)$, an extra pseudo-observation $(j_1, j_2, \ldots, j_k)$ is added to the m observations.

The test statistics for testing H_0 (or H_0') are the test statistics of FRIEDMAN's test applied to the m+1 observations. However, to reject H_0 and accept the trend alternative, not only the test statistic must be significantly large, but also the one-sided tail probability of the statistic must be smaller than the one-sided tail probability of the corresponding test statistic derived from the m original observations.

C. The L-test of E.B. PAGE

Hypotheses H_0 and H_0': as in FRIEDMAN's test.

The same class of alternative hypotheses is considered as in VAN

ELTEREN's test, and the same pseudo-observation $(j_1, j_2, \ldots, j_k)$ is added to the m observations.
The test statistic for testing H_0 (or H_0') is:

$$\underline{L} = \sum_{i=1}^{k} j_i \underline{R}_i , \qquad (8)$$

where $\underline{R}_i$ is again the sum of the ranks of the i^{th} component over the m original observations.
Critical values of $\underline{L}$ under H_0 are tabulated for k = 3(1)8, m = 2(1)12 and k = 3, m = 13(1)20.

For large values of m, the test statistic

$$\underline{U}_1 = \sqrt{m(k-1)} \cdot \left[12\{mk(k^2-1)\}^{-1} \cdot \underline{L} - 3(1 + 2/(k-1)) \right] \qquad (9)$$

has approximately a normal N(0,1) distribution under H_0.
Large positive values of $\underline{\chi}^2$, $\underline{S}$, $\underline{\chi}^2$ and $\underline{F}$ indicate, that the permutations of the ranks are not equally probable; if the components of the k-dimensional observations are independent, this indicates that the marginal distributions of the components are shifted with respect to each other. Large positive values of $\underline{L}$ and $\underline{U}_1$ indicate, that
$\mathcal{E}\underline{x}_{ij_1} < \mathcal{E}\underline{x}_{ij_2} < \ldots < \mathcal{E}\underline{x}_{ij_k}$ (i = 1, ..., m).

Literature

A discussion of FRIEDMAN's test is given in
H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden 1963;

M.G. KENDALL, Rank correlation methods, C. Griffin & Company, London 1955;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 7), Mc Graw-Hill, New York 1956.

A detailed description of the L-test can be found in
E.B. PAGE, Ordered hypotheses for multiple treatments: a significance test for linear ranks, J.A.S.A. 58 (1963) 216-230.

Input: two ways of input are possible, not yet in accordance with note .3.

1) code number of tape

k

n

x_{11} x_{12} - - - - - x_{1k}

⋮

x_{n1} x_{n2} - - - - - x_{nk}

code number of second tape

x = 1: Friedman to be performed, 0: not performed

y = 1: v.Elteren to be performed, 0: not performed

z = 1: L-test to be performed, 0: not performed

if $y = 1$ or $z = 1$: j_1 ..... j_k

2) code number of tape

k

$-m$

x_{11} - - - - - x_{1k} f_1

⋮

x_{m1} x_{mk} f_n

code number of second tape

x, y, z } see 1)

if $y = 1$ or $z = 1$: j_1 ..... j_k

Here sorted observations are used,

f_i = frequency of observation $(x_{i1}, x_{i2}, \ldots, x_{ik})$

m = number of different observations

Output:

{<code number of program >}	
{ingelezen getal band }	code number of tape
{besturingsband }	code of second tape
{Friedman-toets }	
{ S }	S
{ w }	W
{ chi}	χ^2
{ vg }	$k - 1$
{ F }	F
{ nu1}	ν_1
{ nu2}	ν_2
{TREND-toets }	
the same as given under Friedman	
{L-toets }	
{L }	L
{Ul }	U_1

To rank the components of the observations a very slow sort procedure is used.

THE METHOD OF m RANKINGS (ONE d-DIMENSIONAL SAMPLE)

A sample of m independent d-dimensional observations $(\underline{x}_{11}, \ldots, \underline{x}_{1d}), \ldots, (\underline{x}_{m1}, \ldots, \underline{x}_{md})$, not necessarily identically distributed, is supposed to be given. Such a sample can arise if m judges independently rank an ordered set of d elements, or if m subjects are studied under d conditions (in this latter case one often uses d matched samples of m subjects each). In most applications the situation is analogous to two-way analysis of variance (without interaction).

In this program the hypothesis is tested, that the simultaneous distribution of the d components is symmetrical in the components. We present two rank-tests, of which the first has reasonable power against alternatives where the marginal distributions of the components are shifted with respect to each other, and the last is especially powerful against trend-alternatives.

A. FRIEDMAN's test

The components of each d-dimensional observation are ranked separately. Hypothesis H_0: in each observation all permutations of the ranks are equally probable.
If the d components of each observation are independent, the hypothesis H_0 is implied by the hypothesis H_0'.
Hypothesis H_0': in each observation the d components are identically distributed (their distributions may differ for different observations).
Several closely connected statistics for testing H_0 (or H_0') are commonly used.
FRIEDMAN's test statistic is:

$$\underline{X}^2 = 12 \left[m(d^2 + d) \right]^{-1} \cdot \left[\sum_{j=1}^{d} \underline{R}_j^2 - \frac{1}{4} m^2 d(d+1)^2 \right] , \qquad (1)$$

where

$\underline{R}_j$ = sum of the ranks of the j^{th} component over the m observations.
If no ties are present, the exact one-sided tail probabilities of χ^2 under H_0 are tabulated for d = 3, m = 2(1)9 and d = 4, m = 2(1)4.
KENDALL and SMITH considered the test statistic

$$\underline{S} = \sum_{j=1}^{d} \underline{R}_j^2 - \frac{1}{4} m^2 d(d+1)^2 \qquad (2)$$

and tabulated one-sided tail probabilities for d = 3, m = 2(1)10; d = 4, m = 2(1)6 and d = 5, m = 3 (if no ties are present).
KENDALL introduced a coefficient of concordance "between the judges", which is defined as (with tie correction)

$$\underline{W} = \left[\frac{1}{12} m^2 (d^3 - d) - m \sum_i T_i \right]^{-1} \underline{S}, \qquad (3)$$

where

$T_i = \frac{1}{12} \sum_l t_{il} (t_{il}^2 - 1)$, i = 1, ..., m, and t_{il} is the size of the l^{th} tie in the components of the i^{th} observation.
For large m the test statistic

$$\underline{\chi}^2 = \begin{cases} \left[\frac{1}{12} m^2 (d^3 - d) + 2 \right]^{-1} \cdot m(d-1)(\underline{S} - 1) & \text{if no ties are present} \\ \left[\frac{1}{12} m^2 (d^3 - d) - m \sum_i T_i \right]^{-1} \cdot m(d-1)\, \underline{S} & \text{otherwise} \end{cases} \qquad (4)$$

has under H_0 approximately a χ^2-distribution with d-1 degrees of freedom.
With χ^2 and d-1 the corresponding approximate right-hand tail probability is given in the output.
Also for large m, the test statistic

$$\underline{F} = \left[1 - \frac{1}{m(d-1)} \underline{\chi}^2 \right]^{-1} \cdot \frac{m-1}{m(d-1)} \underline{\chi}^2 , \qquad (5)$$

where $\underline{\chi}^2$ is defined in (4),
has under H_0 approximately an F-distribution with ν_1 and ν_2 degrees of freedom; ν_1 and ν_2 are given by

$$\nu_1 = \frac{1}{2}(d - 1)(1 - \frac{1}{m}) \cdot \left[\sum_{i<l} \mu_i \mu_l \right]^{-1} \cdot \left[\sum_{i=1}^{m} \mu_i \right]^2 - \frac{2}{m} \qquad (6)$$

$$\nu_2 = (m - 1)\nu_1, \qquad (7)$$

where $\mu_i = \frac{1}{12}(d^2 - 1) - \frac{1}{d} T_i$.

For moderate values of m and d the F-approximation is reputed to be somewhat better than the χ^2-approximation. With F, ν_1 and ν_2 the corresponding right-hand tail probability is given in the output.

B. The L-test of E.B. PAGE

Hypotheses H_0 and H_0': as in FRIEDMAN's test.

If a test is wanted with large power against the class of alternative hypotheses characterized by $\mathcal{E}\underline{x}_{ij_1} < \mathcal{E}\underline{x}_{ij_2} < \ldots < \mathcal{E}\underline{x}_{ij_d}$, $i = 1, \ldots, m$, where $(j_1, \ldots, j_d)$ is a given permutation of $(1, 2, \ldots, d)$, the L-test of E.B. PAGE can be used.

The test statistic for testing H_0 (or H_0') is:

$$\underline{L} = \sum_{i=1}^{d} j_i \underline{R}_i , \qquad (8)$$

where $\underline{R}_i$ is again the sum of the ranks of the i^{th} component over the m original observations.

Critical values of $\underline{L}$ under H_0 are tabulated for d = 3(1)8, m = 2(1)12 and d = 3, m = 13(1)20.

For large values of m, the test statistic

$$\underline{U}_L = \sqrt{m(d - 1)} \cdot \left[12 \{md(d^2 - 1)\}^{-1} \cdot \underline{L} - 3(1 + 2/(d - 1)) \right] \qquad (9)$$

has approximately a normal N(0,1) distribution under H_0.

With U_L the corresponding approximate right-hand tail probability is given in the output.

Large positive values of $\underline{X}^2$, $\underline{S}$, $\underline{\chi}^2$ and $\underline{F}$ indicate, that the permutations of the ranks are not equally probable; if the components of the d-dimensional observations are independent, this indicates that the marginal distributions of the components are shifted with respect to each other. Large positive values of $\underline{L}$ and $\underline{U}_L$ indicate, that

$\mathcal{E}\,\underline{x}_{ij_1} < \mathcal{E}\,\underline{x}_{ij_2} < \ldots < \mathcal{E}\,\underline{x}_{ij_d}$ $(i = 1, \ldots, m)$.

Literature

A discussion of FRIEDMAN's test is given in

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden 1963;

M.G. KENDALL, Rank correlation methods, C. Griffin & Company, London 1955;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 7), Mc Graw-Hill, New York 1956.

A detailed description of the L-test can be found in

E.B. PAGE, Ordered hypotheses for multiple treatments: a significance test for linear ranks, J.A.S.A. 58 (1963) 216-230.

Input:

a d-dimensional set of observations, in one of the two forms described in note 3.

tests = { 1: A performed
2: A and B performed

code number of second tape
$j_1, j_2, \ldots, j_d$ } (only if tests = 2)

The program contains XEENtal.

Output:

{Result <code number of program>}
{The method of m rankings, see descr 9a}
{input} code numbers of sample and second tape

{X2	S	W	Chi	df	P	F	df	df	P	L	UL	P }
X^2	S	W	χ^2	d-1	P_{χ^2}	F	ν_1	ν_2	P_F	L	U_L	P_{U_L}

TWO-SAMPLE TESTS (ONE-DIMENSIONAL CASE)

Two samples of independent observations are supposed to be given:

sample I (size m): $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m$, all observations identically distributed,

sample II (size n): $\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_n$, all observations identically distributed.

In this program the hypothesis is tested, that $\mathcal{E}\underline{x} - \mathcal{E}\underline{y} = d$. Two tests are presented. Both tests are especially powerful against shift-alternatives: $\mathcal{E}\underline{x} - \mathcal{E}\underline{y} \neq d$.

In addition, a test is given to test the hypothesis, that the variances of both distributions are equal.

A. The STUDENT t-test for two samples

Hypothesis H_0: the observations $\underline{x}_i$ and $\underline{y}_i$ are normally distributed, $\sigma^2(\underline{x}) = \frac{1}{c}\sigma^2(\underline{y})$ $(c > 0)$ and $\mathcal{E}\underline{x} - \mathcal{E}\underline{y} = d$.

The second condition is imposed on H_0 for the rare cases, where the ratio of the variances is known to be $\neq 1$. Usually one supposes $c = 1$.

The test statistic for testing H_0 is:

$$\underline{t} = \left[\underline{S}_1^2 + \frac{1}{c}\underline{S}_2^2\right]^{-\frac{1}{2}} \cdot \sqrt{\frac{(m+n-2)mn}{n+cm}} \cdot \left[\frac{1}{m}\sum_{i=1}^{m}\underline{x}_i - \frac{1}{n}\sum_{i=1}^{n}\underline{y}_i - d\right] \quad (1)$$

where

$$\underline{S}_1^2 = \sum_{i=1}^{m}(\underline{x}_i - \frac{1}{m}\sum_{j=1}^{m}\underline{x}_j)^2 \text{ and } \underline{S}_2^2 = \sum_{i=1}^{n}(\underline{y}_i - \frac{1}{n}\sum_{j=1}^{n}\underline{y}_j)^2.$$

Then $\underline{t}$ has under H_0 a t-distribution with m+n-2 degrees of freedom. With t and m+n-2 the corresponding right-hand tail probability is given in the output.

B. The F-test for equality of variances

Hypothesis H_0: the observations $\underline{x}_i$ and $\underline{y}_i$ are normally distributed with equal but unknown variances $\sigma^2(\underline{x}) = \sigma^2(\underline{y})$.

The test statistic for testing H_0 is:

$$\underline{F} = \frac{n-1}{m-1} \underline{S}_1^2 / \underline{S}_2^2 , \qquad (2)$$

which under H_0 has an F-distribution with m-1 and n-1 degrees of freedom.
With F, m-1 and n-1 the corresponding right-hand tail probability is given in the output.

C. WILCOXON's two-sample test

Hypothesis H_0: the observations $\underline{x}_i - d$ and $\underline{y}_i$ are identically distributed.

The test statistic for testing H_0 is:

$$\underline{W} = \sum_i \sum_j \delta(\underline{x}_i - d, \underline{y}_j) , \qquad (3)$$

where

$$\delta(\underline{x}_i - d, \underline{y}_j) = \begin{cases} 2 & \text{if} \quad \underline{x}_i - d > \underline{y}_j \\ 1 & \text{if} \quad \underline{x}_i - d = \underline{y}_j \\ 0 & \text{if} \quad \underline{x}_i - d < \underline{y}_j . \end{cases}$$

For n and m up to 10 one-sided tail-probabilities of $\underline{W}$ under H_0 have been tabulated, if no ties are present. For large n and m $\underline{W}$ is approximately normally distributed with

$$\text{mean value} \quad \mu_w = mn \qquad (4)$$

$$\text{variance} \quad \sigma_w^2 = \frac{1}{3} \left[(m+n)(m+n-1) \right]^{-1} \cdot \left[(m+n)^3 - D \right] \cdot mn , \qquad (5)$$

where $D = \sum_j t_j^3$ and t_j is the size of the j^{th} tie in the pooled sample ($t_j = 1$ inclusive). Thus the standardized $\underline{W}$ is

$$\underline{W}^* = \frac{\underline{W} - \mu_w}{\sigma_w} . \qquad (6)$$

For large values of m and n the statistics

$$\underline{W}_1 = \frac{1}{\sigma_w} (\underline{W} - \mu_w + 1) \quad \text{(one-sided test: left hand tail)} \quad (7)$$

$$\underline{W}_r = \frac{1}{\sigma_w} (\underline{W} - \mu_w - 1) \quad \text{(one-sided test: right hand tail)} \quad (8)$$

$$\underline{W}_2 = \frac{1}{\sigma_w} (|\underline{W} - \mu_w| - 1) \quad \text{(two-sided test)} \quad (9)$$

can be considered as random normal deviates and can be used to find approximate tail-probabilities in the cases of one-sided or two-sided tests respectively.
With W_1, W_r and W_2 the corresponding approximate tail probabilities are given in the output.

Large positive (negative) values of $\underline{t}$ and $\underline{W}^*$ indicate that $\mathcal{E}\,\underline{x} - \mathcal{E}\,\underline{y} > d$ ($\mathcal{E}\underline{x} - \mathcal{E}\underline{y} < d$).
Large (small) positive values of $\underline{F}$ indicate that $\sigma^2(\underline{x}) > \sigma^2(\underline{y})$ ($\sigma^2(\underline{x}) < \sigma^2(\underline{y})$).

Literature

In most books on elementary statistics details of STUDENT's two-sample t-test can be found, e.g.

W.J. DIXON & F.J. MASSEY, Jr., Introduction to statistical analysis (Chapter 9), Mc Graw-Hill, New York 1951.

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 12), Mc Graw-Hill, New York 1963.

In the first book the F-test for equality of two variances is also described (Chapter 8). Discussions of WILCOXON's test are given in

C. VAN EEDEN & D. WABEKE, Handleiding voor de toets van Wilcoxon, Report S 176 of the Statistical Department of the Mathematical Centre, Amsterdam 1955.

H. DE JONGE, Inleiding tot de medische statistiek, Part I, Leiden 1963.

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 16), Mc Graw-Hill, New York 1963.

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 6), Mc Graw-Hill, New York 1956.

Input:

code number of the tape

2 = number of samples

1 = dimension of each sample

p $\geq l_1 + l_2$

l_1 = the number of different observations in sample I if this sample is given as a sorted frequency table or tests = 2, otherwise l_1 = number of observations in sample I; l_2 is defined for sample II in the same way

tests = $\begin{cases} 1: & \text{A and B performed} \\ 2: & \text{A, B and C performed} \end{cases}$

sample I, given in one of the two forms described in note 3

sample II, given in one of the two forms described in note 3

c

d

The program contains XEENtal.

Output:

{Result<code number of program>}
{Two-sample tests, see descr 10,}
{input} code number of the tape

{c	d	t	df	P	F	df	df	P	muW	sigW}
c	d	t	$m+n-2$	P_t	F	$m-1$	$n-1$	P_F	μ_w	σ_w
{Ws	Wl	P	Wr	P	Wz	P}				
W	W_1	P_{W_1}	W_r	P_{W_r}	W_2	P_{W_2}				

k-SAMPLE TESTS (ONE-DIMENSIONAL CASE)

k samples of independent observations $\underline{x}_1^{(1)}, \ldots, \underline{x}_{n_1}^{(1)}; \ldots; \underline{x}_1^{(k)}, \ldots, \underline{x}_{n_k}^{(k)}$ are supposed to be given, $k \geq 3$. The observations in a sample are assumed to have identical distributions.

In this program some tests are presented for the k-sample problem. Under A and C two tests are given to test the hypothesis, that the k samples are from identically distributed populations; these tests are especially powerful against shift-alternatives: $\mathcal{E}\underline{x}^{(i)} \neq \mathcal{E}\underline{x}^{(j)}$ for at least one pair (i, j). Under B a test is given to test the hypothesis, that the $\underline{x}^{(i)}$ ($i = 1, \ldots, k$) have equal variances. Under D a test is given for trend in the samples; this test is consistent for all weak-trend alternatives.

A. The F-test for one-way analysis of variance

Hypothesis H_0: the observations $\underline{x}_j^{(i)}$ have identical normal distributions.

The test statistic for testing H_0 is:

$$\underline{F} = \frac{n-k}{k-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\underline{x}_j^{(i)} - \overline{\underline{x}}^{(i)})^2\right]^{-1} \cdot \left[\sum_{i=1}^{k} n_i(\overline{\underline{x}}^{(i)} - \overline{\underline{x}})^2\right] ,(1)$$

where

$$\overline{\underline{x}}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i}\underline{x}_j^{(i)} \quad (i = 1, \ldots, k),\ n = \sum_{i=1}^{k} n_i \text{ and}$$

$$\overline{\underline{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i\,\overline{\underline{x}}^{(i)}.$$

Then $\underline{F}$ has under H_0 an F-distribution with k-1 and n-k degrees of freedom. With F, k-1 and n-k the corresponding right-hand tail probability is given in the output.

B. BARTLETT's test for equality of variances

Hypothesis H_0: the observations $\underline{x}_j^{(i)}$ are normally distributed and have equal but unknown variances.

The test statistic for testing H_0 is:

$$\underline{L} = \left[1 - \frac{1}{3(k-1)}\left\{\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right\}\right] \cdot$$

$$\cdot \left[(n-k)\log\left\{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\underline{x}_j^{(i)} - \underline{\bar{x}}^{(i)})^2\right\} + \right.$$

$$\left. - \sum_{i=1}^{k}(n_i - 1)\log\left\{\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\underline{x}_j^{(i)} - \underline{\bar{x}}^{(i)})^2\right\}\right]. \qquad (2)$$

Then for large n_i the statistic $\underline{L}$ has under H_0 approximately a χ^2-distribution with k-1 degrees of freedom.
With L and k-1 the corresponding approximate right-hand tail probability is given in the output.

C. The KRUSKAL-WALLIS test

Hypothesis H_0: the observations $\underline{x}_j^{(i)}$ have identical distributions.

The test statistic for testing H_0 is:

$$\underline{H} = 3(n-1)\left[n^3 - n - \sum_l T_l\right]^{-1} \cdot \left[\sum_{i=1}^{k}\frac{1}{n_i}\left\{\sum_{\substack{j=1\\ j\neq i}}^{k}(\underline{W}_{ij} - n_i n_j)\right\}^2\right], \qquad (3)$$

where

$T_l = t_l^3 - t_l$ and t_l is the size of the j^{th} tie of observations in the pooled sample,

$\underline{W}_{ij}$ = the test statistic of WILCOXON's two-sample test applied to the i^{th} and j^{th} sample, as samples I and II respectively.

For k = 3 and n $\leq$ 5 one-sided tail probabilities are tabulated (if no ties are present). For large sample sizes $\underline{H}$ has under H_0 approximately a χ^2-distribution with k-1 degrees of freedom.

With H and k-1 the corresponding approximate right-hand tail probability is given in the output.

Starting from the KRUSKAL-WALLIS test, a test for outlying samples is constructed in the following way.
Let α be the given significance level. If the approximate tail probability of H is smaller than α, the statistics $V_i^{(1)}$ $(i = 1, \ldots, k)$ are computed, defined by

$$\underline{V}_i^{(1)} = \left[n_i(n - n_i)(n^3 - D_1)\right]^{-\frac{1}{2}} \cdot \frac{1}{3n(n-1)} \left\{ \sum_{\substack{j=1 \\ j\neq i}}^{k} \underline{W}_{ij} - n_i(n - n_i)\right\} , \qquad (4)$$

where $D_1 = \sum_l t_l^3$.
Define i_1 as the index i for which $|\underline{V}_i^{(1)}|$ obtains its maximum value.
Then for not too small sample sizes

$$P \left\{\underline{V}_{i_1}^{(1)} > U_{\alpha/k}\right\} \leq \alpha, \qquad (5)$$

where $U_{\alpha/k}$ is the upper $\frac{\alpha}{k}$ percentage point of the normal N(0,1) distribution.

If the approximate two-sided tail probability of $V_{i_1}^{(1)}$ is smaller than α, the statistics $\underline{V}_i^{(2)}$ $(i = 1, \ldots, k; i \neq i_1)$ are computed in a similar fashion as the $\underline{V}_i^{(1)}$ from (4), but omitting the i_1th sample from all considerations. Then define i_2 as the index i for which $|\underline{V}_i^{(2)}|$ obtains its maximum value. Analogous to (5) we have

$$P \left\{\underline{V}_i^{(2)} > U_{\alpha/(k-1)}\right\} \leq \alpha. \qquad (6)$$

This procedure is continued until the approximate two-sided tail probability of $V_{i_l}^{(l)}$ is larger than $\alpha/(k-l+1)$ or until only two samples are left. At each step $V_{i_l}^{(l)}$ and the corresponding approximate two-sided tail probability are given in the output.

D. TERPSTRA's test for a trend

Hypothesis H_0: the observations $\underline{x}_j^{(i)}$ are identically distributed.

The test statistic for testing H_0 is:

$$\underline{W} = \sum_{i<j}\sum \underline{W}_{ij} / n_i n_j ,$$

where $\underline{W}_{ij}$ is defined in C.
The test statistic $\underline{W}$ has under H_0 approximately a normal distribution with expected value $\mu_w = \frac{1}{2} k(k - 1)$ and variance

$$\sigma_w^2 = \left[3n(n-1)(n-2) \right]^{-1} \cdot \left[\{n^3 - D_1 - 3(n^2 - \sum_l t_l^2)\} \cdot \sum_{i=1}^{k} \frac{1}{n_i} (k + 1 - 2i)^2 - \{2(n^3 - D_1) - 3n(n^2 - \sum_l t_l^2)\} \cdot \sum_{i<j}\sum \frac{1}{n_i n_j} \right] , \qquad (7)$$

where t_l and D_1 have been defined in C.
The standardized $\underline{W}$ is thus

$$\underline{W}^* = \frac{1}{\sigma_w} (\underline{W} - \frac{1}{2} k(k - 1)). \qquad (8)$$

With W^* the corresponding right-hand tail probability is given in the output.

Large positive values of $\underline{F}$ and $\underline{H}$ indicate, that at least two of the mean values $\mathcal{E}\underline{x}^{(1)}, \ldots, \mathcal{E}\underline{x}^{(k)}$ are different. Large positive values of $\underline{L}$ indicate, that at least two of the variances $\sigma^2(\underline{x}^{(1)}), \ldots, \sigma^2(\underline{x}^{(k)})$ are different. Large positive (negative) values of $\underline{V}_{i_1}^{(1)}$ indicate, that the mean value of $\underline{x}^{(i_1)}$ is larger (smaller) than the means of the remaining samples. Large positive values of $\underline{W}$ indicate, that $\mathcal{E}\underline{x}^{(1)} \geq \mathcal{E}\underline{x}^{(2)} \geq \ldots \geq \mathcal{E}\underline{x}^{(k)}$ with at least one strict inequality.
It is not necessary to give the samples in the order in which a trend is supposed to exist. If Terpstra's test is to be performed, the input must contain a second tape with a permutation $j_1, \ldots, j_k$ of the numbers $1, \ldots, k$ indicating the order in which a trend is supposed to exist.

Remark: The test statistic $\underline{H}$ of the KRUSKAL-WALLIS test is usually defined in a somewhat different manner. Rank the observations in the pooled sample, and let $\underline{R}_i$ be the sum of the ranks of the observations in the i^{th} sample. Then

$$\underline{H} = \left[\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{n_i} R_i^2 - 3(n+1) \right] \cdot \left[1 - \frac{1}{n^3 - n} \sum_{l} T_l \right]^{-1}. \qquad (9)$$

However, the definition is fully equivalent with definition (3).

Literature

In most books on elementary statistics details of the F-test of one-way analysis of variance can be found, e.g.

W.G. DIXON & F.J. MASSEY, Introduction to statistical analysis (Chapter 10), Mc Graw-Hill, New York 1951;

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 14), Mc Graw-Hill, New York 1963.

In the first book BARTLETT's test is described, but with an F-approximation instead of a χ^2-approximation. A description of the test where the χ^2-approximation is used is given in the book of H. DE JONGE mentioned below (Part II, Chapter 14).
Discussions of the KRUSKAL-WALLIS test and TERPSTRA's test for trend are given in

H. DE JONGE, Inleiding tot de medische statistiek, Part I (Chapter 10), Leiden 1963;

W.H. KRUSKAL & W.A. WALLIS, Use of ranks in one-criterion variance analysis, J. Amer. Statist. Ass. 47 (1952) 583-621;

S. SIEGEL, Nonparametric statistics for the behavioral sciences (Chapter 8), Mc Graw-Hill, New York 1956.

The test for outlying samples is based on an idea of R. DOORNBOS. An exposition of his technique (with several examples) is given in

R. DOORNBOS, Statistische methoden voor het aanwijzen van uitbijters, Statistica Neerlandica, 13 (1959) 453-462.

Input:

code-number of the first tape

k = number of samples

l = dimension of each sample

p $\geq l_1 + l_2 + \ldots + l_k$

l_i = number of different observations in the i^{th} sample if this sample is given as a sorted frequency table or if tests ≠ 1, otherwise l_i = number of observations in the i^{th} sample;

tests = { 1: A and B performed
2: A, B and C performed
3: A, B, C and D performed }

k times: a l-dimensional sample, given in one of the two forms described in note 3.

α (only if tests ≠ 1)

code-number of second tape
$j_1, j_2, \ldots, j_k$
} (only if tests = 3)

The program contains XEENtal.

Output:

{Result <code number of program>}

{k-Sample tests, see descr 11}

{input} code numbers of first and second tape

{F	df	df	P	L	df	P	W	muW	sigW}
F	k-1	n-k	P_F	L	k-1	P_L	W	μ_W	σ_W
{Ws	P	H	df	P }					
W^*	P_{W^*}	H	k-1	P_H					

{Outlying samples}

code number of sample i_1 $V_{i_1}^{(1)}$ $P(V_{i_1}^{(1)})$

ANALYSIS OF VARIANCE (FIXED EFFECTS; GENERAL MODEL)

Let $\underline{y}_{ij}$, $j = 1, 2, \ldots, n_i$, $n_i \geq 1$, $i = 1, 2, \ldots, k$, denote k independent random samples from normal distributions with equal variances but possibly different expectations:

$\underline{y}_{ij}$ independent, $\underline{y}_{ij} \in N(\eta_i, \sigma^2)$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$.

We write

$$n = \sum_{i=1}^{k} n_i, \quad \underline{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{y}_{ij}, \quad \underline{y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \underline{y}_{ij}.$$

A model is given by the sets of equations

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j, \quad i = 1, 2, \ldots, k, \quad x_{ij} = 0, \text{ or } 1;$$

$$0 = \sum_{j=1}^{p} h_{ij} \beta_j, \quad i = 1, 2, \ldots, l, \quad h_{ij} = 0, \text{ or } 1.$$

This means that the expectation of an observation in the i^{th} sample (or "cell") is obtained by adding a number of the "effects" $\beta_1, \beta_2, \ldots, \beta_p$, which are themselves subject to l linear constraints. We define

- η : the column-vector $(\eta_1, \eta_2, \ldots, \eta_k)$;
- $\mathring{y}$: the column-vector $(y_{1\cdot}, y_{2\cdot}, \ldots, y_{k\cdot})$;
- β : the column-vector $(\beta_1, \beta_2, \ldots, \beta_p)$;
- N : the kxk diagonal matrix with elements $n_{ij} = \delta_{ij} n_i$, when $\delta_{ij} = 1$ if $i = j$ and 0 otherwise;
- X : the matrix (x_{ij}) with k rows and p columns;
- H : the matrix (h_{ij}) with l rows and p columns;
- $\binom{X}{H}$: the $(k + l)$xp matrix of which the first k rows are the rows of X and the last l rows are the rows of H.

The transpose of a matrix will be denoted by adding a prime, e.g. X' is the pxk matrix (x_{ij}).

In matrix notation the model becomes

$$\eta = X\beta$$
$$0 = H\beta.$$

The model is consistent, determines β uniquely and β is estimable, if rank $X = r$, rank $H = p - r$, rank $\binom{X}{H} = p$.

The following quantities are computed (with each sum of squares the number of degrees of freedom (df.) is given):

1. The total sum of squares

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{..})^2 \qquad \text{df } (n - 1).$$

2. The residual sum of squares

$$R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 \qquad \text{df } (n - k)$$

3. The determinant

$$\det (X'NX + H'H).$$

This is meant to provide a check on the model, since the ranks of X, H and $\binom{X}{H}$ are not determined in the program. If rank $\binom{X}{H} < p$ the matrix X'NX + H'H is singular and the remainder of the program will produce nonsensical results. If the determinant is zero the program is stopped accordingly.

4. The column vector $b = (b_1, b_2, \ldots, b_p)$ of least squares estimates of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$

$$b = (X'NX + H'H)^{-1} x'N\mathring{y}.$$

5. The column-vector $a = (a_1, a_2, \ldots, a_k)$ of least squares estimates of $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ for the model given

$$a = Xb.$$

6. The sum of squares

$$S - R = \sum_{i=1}^{k} n_i (y_{i.} - a_i)^2 \qquad \text{df } (k - r),$$

where $$S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - a_i)^2$$

denotes the minimal sum of squares for the model given.

7. The test statistic for the model against the underlying assumption

$$F = \frac{S - R}{R} \cdot \frac{n - k}{k - r} ;$$

under the assumptions of the model $\underline{F}$ has an F-distribution with $k - r$ and $n - k$ df.

8. The estimated covariance-matrix $C(b)$ of b

$$C(b) = \frac{S}{n - r} \cdot ZNZ'$$

where $$Z = (X'NX + H'H)^{-1} X'.$$

The special feature of this program in its great generality: it processes analysis of variance results with unequal numbers n_i of observations per cell as well as missing plots. Since the matrix inversion in the program can be rather time-consuming, it should preferably not be used for routine designs with equal n_i.
When using the program one should realize that generally, for unequal numbers of observations per cell, the F-test for the existence of a given set of effects depends upon the model as a whole, i.e. on what other effects one wishes to retain in the model.

If missing plots occur ($n_i = 0$) the corresponding cells are simply emitted from the model. One often has to be extremely careful in this case that β is uniquely determined by the model and estimable (e.g. interactions cannot be separated from main effects, etc.).

It may be necessary to try a number of different models of increasing complexity before a satisfactory one is found. The estimated covariance matrix C(b) may thereafter be used to carry out multiple comparison procedures. The program provides this covariance matrix only if the test statistic F for the model satisfies

$$F \leq cref$$

where cref is some preassigned critical value for F. For F-values exceding cref the model is not supposed to be satisfactory anyway. It is also possible to have the computer print only parts of the matrix C(b).

Literature

H. SCHEFFÉ, The analysis of variance, Wiley New York, 1959.

Note that the matrices X and H are denoted by X' and H' by SCHEFFÉ.

The input data should be given in the following way:

code number of the tape

k

the k samples, each given in one of the two forms described in note .3.

cref

the number of models to be tried

the models, each consisting of:

code number of the model

l

p

r

$\sum_{i=1}^{k} m_i$

$$\begin{matrix} h_{11} & \cdots & h_{1p} \\ \vdots & & \\ h_{l1} & \cdots & h_{lp} \end{matrix}$$

$$\begin{matrix} m_1 & z_1 & \cdots & z_{m1} \\ \vdots & & & \\ m_k & z_1 & \cdots & z_{mk} \end{matrix}$$

where: $m_i = \sum_{j=1}^{p} x_{ij}$

z_j such that $x_{iz_j} = 1$ and $z_{j-1} < z_j$

s = number of parts of C(b) to be computed

$$\left.\begin{matrix} l_1 & u_1 \\ \vdots & \\ l_s & u_s \end{matrix}\right\}$$ see description of output

Output:

{Result <code number of program>}
{Analysis of variance}
{see descr 12}
{input} code number of input tape
{T } T { df } n-1
{R } R { df } n-k
of each model:
{model} code number of model
{det } det(X'NX + H'H)
{vector b vector a}

b_1 a_1
⋮ ⋮
b_p a_k

{S - R} S - R { df } k-r
{F } F
{C(b) }
the s parts of C(b), t = 1, 2, ..., s :

$$\begin{matrix} c_{1_t 1_t} & & \\ \vdots & \ddots & \\ c_{u_t 1_t} & \cdots & c_{u_t u_t} \end{matrix}$$

where c_{ij} = covariance (b_i, b_j)

The program contains setXEENtal (see note .2.)

MULTIPLE POLYNOMIAL REGRESSION

In this program the regression coëfficiënts of a multiple polynomial regression function are estimated from the given sample and some tests are presented for testing the model and certain assumptions about the regression coëfficiënts.

The following sample is supposed to be given:

$$\{ \underline{y}_{ij}, z_{1i}, \ldots, z_{ki} \}, \; j = 1, \ldots, m_i; \; i = 1, \ldots, n; \; \sum_{i=1}^{n} m_i = m,$$

where the $z_{1i}, \ldots, z_{ki}$ are the values of k independent (fixed) variables $z_1, \ldots, z_k$ and the $\underline{y}_{ij}$ are independent observations of the dependent (random) variable $\underline{y}$.

The underlying model to be investigated is:
"the $\underline{y}_{ij}$ normally distributed with expectations $\mathcal{E}\underline{y}_{ij} = P(z_{1i}, \ldots, z_{ki})$ and variances $a_i^2\sigma^2$,
where $P(z_{1i}, \ldots, z_{ki})$ is a polynomial in $z_{1i}, \ldots z_{ki}$ with coëfficiënts $\beta_1, \ldots, \beta_p$ of the form

$$P(z_{1i}, \ldots, z_{ki}) = \sum_{r=1}^{p} \beta_r \prod_{l=1}^{k} (z_{li})^{t_{rl}} \text{"}. \qquad (1)$$

In each instance the appropriate regression function can be obtained by choosing the right exponents t_{rl} for each term (r = 1, ..., p). In the case of linear regression without constant term the matrix (t_{rl}) is the kxk identity-matrix, in the case of linear regression with a constant term the matrix (t_{rl}) is the (k+1)xk matrix of the form

$$(t_{rl}) = \begin{pmatrix} 0 \; 0 \ldots 0 \\ \hline I \end{pmatrix},$$

where I is the kxk identity-matrix.
The a_i^2 are introduced to allow for different variances of the $\underline{y}_{ij}$ for different values of the independent variables; if a constant variance is thought to be adequate, all a_i^2 should be taken equal to one.

Under A the regression coëfficiënts are estimated, estimated standard deviations of the estimators $\underline{b}_r$ of β_r (r = 1, ..., p) are given and separate confidence intervals for the β_r (r = 1, ..., p) are derived. Under B a test for the assumed model is given in case of replicated observations of $\underline{y}$ for constant values of the independent variables; this part of the program is (not) evaluated if the input-variable B = 1(0). Under C a test for the assumed model against a model with a regression function containing more (higher degree) terms is given; this part of the program is (not) evaluated if the input-variable C = 1(0). Under D a test is presented to test the hypothesis that some of the regression coëfficiënts have given values; this part of the program is (not) evaluated if the input-variable D = 1(0).

A. Estimation of the regression coëfficiënts

The (column) vector of least squares estimators $\underline{b} = (\underline{b}_1, \ldots, \underline{b}_p)'$ of the vector of regression coëfficiënts $(\beta_1, \ldots, \beta_p)'$ is given by

$$\underline{b} = K^{-1} X \underline{\tilde{y}} , \qquad (2)$$

where

$K = XX'$,

$X = (x_{ri})$ is a pxn matrix of rank p with elements

$$x_{ri} = (a_i)^{-1}\sqrt{m_i} \cdot \prod_{l=1}^{k} (z_{li})^{t_{rl}} \quad (r = 1, \ldots, p;\ i = 1, \ldots, n) \qquad (3)$$

and

$\underline{\tilde{y}}$ is an n-dimensional (column) vector with elements

$$\underline{\tilde{y}}_i = (a_i\sqrt{m_i})^{-1} \sum_{j=1}^{m_i} \underline{y}_{ij} \quad (i = 1, \ldots, n); \qquad (4)$$

a prime denotes the transpose of a matrix or vector.

Define

$$\underline{R}^2 = \sum_{i=1}^{n} \frac{1}{a_i^2} \{ \sum_{j=1}^{m_i} \underline{y}_{ij}^2 - \frac{1}{m_i} (\sum_{j=1}^{m_i} \underline{y}_{ij})^2 \} \tag{5}$$

$$\underline{S}_\Omega = \sum_{i=1}^{n} \underline{\tilde{y}}_i^2 - \underline{b}' X \underline{\tilde{y}} , \tag{6}$$

then $\underline{R}^2/\sigma^2$ and $\underline{S}_\Omega/\sigma^2$ are independently distributed as chi squared with m-n and n-p degrees of freedom respectively. Thus $\underline{R}^2/(m-n)$ and $\underline{S}_\Omega/(n-p)$ are independent estimators of the error variance σ^2.

The (conditional) error variance $\sigma^2 = \text{var}(\underline{y}|z_1, \ldots, z_k)$ as a fraction of the unconditional variance var $\underline{y}$ is estimated by

$$1 - \underline{C}^2 = \frac{n-1}{m-p} \left[\underline{R}^2 + \sum_{i=1}^{n} \underline{\tilde{y}}_i^2 - \frac{1}{n} (\sum_{i=1}^{n} \underline{\tilde{y}}_i)^2 \right]^{-1} \cdot \left[\underline{R}^2 + \underline{S}_\Omega \right] ; \tag{7}$$

if the regression function is linear, $\underline{C}$ is the sample multiple correlation coëfficiënt between $\underline{y}$ and the set $z_1, \ldots, z_k$.

Estimates of the standard deviations $\sigma(\underline{b}_r)$ are provided by

$$\underline{\hat{\sigma}}(\underline{b}_r) = \left[(K^{-1})_{rr} \cdot (\underline{S}_\Omega + \underline{R}^2)/(m - p) \right]^{\frac{1}{2}} . \tag{8}$$

Exact two-sided confidence intervals with confidence level $1-\alpha$ are given by

$$\underline{b}_r - \underline{\hat{\sigma}}(\underline{b}_r) \cdot t_{\frac{\alpha}{2}} < \beta_r < \underline{b}_r + \hat{\sigma}(\underline{b}_r) \cdot t_{\frac{\alpha}{2}} \quad (r = 1, \ldots, p) , \tag{9}$$

where

$t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ - point of the t-distribution with m-p degrees of freedom.

B. Testing the model in case of replicated observations

If some of the m_i are larger than one, the fit of the data to the general form of the regressior function $P(z_1, \ldots, z_k)$ can be tested with an F-test. The test statistic is

$$\underline{F}_1 = \frac{m-n}{n-p} \underline{S}_\Omega / \underline{R}^2 , \tag{10}$$

which has an F-distribution with n-p and m-n degrees of freedom if $P(z_1, \ldots, z_k)$ is the true regression function.

C. Testing the regression function against a higher degree polynomial

If m-n is zero or very small, a direct test like (9) of the adequacy of the regression function is not available. However, the regression function can be tested against a higher degree polynomial, or more generally, against a polynomial regression function $P^*(z_1, \ldots, z_k)$ with more terms. Let this more general alternative regression function be defined by

$$P^*(z_1, \ldots, z_k) = \sum_{r=1}^{p+q} \beta_r^* \prod_{l=1}^{k} (z_{li})^{t_{rl}} . \tag{11}$$

The function P^* can simply be obtained by adding q extra rows to the matrix (t_{rl}) of original exponents $(r = p+1, \ldots, p+q)$.

The (column) vector of least squares estimators $\underline{b}^* = (\underline{b}_1^*, \ldots, \underline{b}_{p+q}^*)'$ of the vector of new regression coëfficiënts $(\beta_1^*, \ldots, \beta_{p+q}^*)$ is now given by

$$\underline{b}^* = (K_2)^{-1} \cdot \begin{pmatrix} X \\ X^* \end{pmatrix} \cdot \underline{\tilde{y}} , \tag{12}$$

where

$X^* = (x_{ri})$ is a qxn matrix with elements

$$x_{ri} = (a_i)^{-1}\sqrt{m_i} \cdot \prod_{l=1}^{k} (z_{li})^{t_{rl}} \quad (r = p+1, \ldots, p+q;\ i = 1, \ldots, n) \tag{13}$$

and

$$K_2 = \begin{pmatrix} X \\ X^* \end{pmatrix} (X' \; X^{*\prime}) .$$

Let

$$S_{\omega_1} = \sum_{i=1}^{n} \tilde{y}_i^2 - \underline{b}^{*\prime} \begin{pmatrix} X \\ X^* \end{pmatrix} \tilde{\underline{y}} , \tag{14}$$

then the test statistic for testing P against P^* is

$$\underline{F}_2 = \frac{m - p - q}{q} (\underline{S}_\Omega - \underline{S}_{\omega_1})/(\underline{S}_{\omega_1} + \underline{R}^2) , \tag{15}$$

which has an F-distribution with q and m-p-q degrees of freedom if $P(z_1, \ldots, z_k)$ is the true regression function.

D. Testing hypotheses about the regression coëfficiënts

In this part of the program the following hypothesis is tested:

$$H_0:\ \beta_{r_1} = \beta_{r_1}^{**}, \ldots, \beta_{r_\nu} = \beta_{r_\nu}^{**} , \tag{16}$$

where $\{r_1, \ldots, r_\nu\}$ is a subset of $\{1, 2, \ldots, p\}$ and $\beta_{r_1}^{**}, \ldots, \beta_{r_\nu}^{**}$ are given constants.

The (column) vector of least squares estimators $\underline{b}^{**} = (\underline{b}_1^{**}, \ldots, \underline{b}_{p-\nu}^{**})'$ of the vector of the remaining p-ν regression coëfficiënts is given by

$$\underline{b}^{**} = (K_3)^{-1} X^{**} \tilde{\underline{y}}^{**} , \tag{17}$$

where

X^{**} is a (p-ν)xn matrix of rank p-ν which is obtained from the matrix X by **deleting** the rows with row-indices $r_1, \ldots, r_\nu$,

$$K_3 = X^{**} (X^{**})' ,$$

$\underline{\tilde{y}}^{**}$ is an n-dimensional (column) vector with elements

$$\underline{\tilde{y}}_i^{**} = \underline{\tilde{y}}_i - \sum_{\alpha=1}^{\nu} \beta_{r_\nu}^{**} x_{r_\nu,i} . \qquad (18)$$

Let

$$\underline{S}_{\omega_2} = \sum_{i=1}^{n} (\underline{\tilde{y}}_i^{**})^2 - \underline{b}^{**\prime} X^{**} \underline{\tilde{y}}^{**} , \qquad (19)$$

then the test statistic for testing H_0 is

$$\underline{F}_3 = \frac{m - p}{\nu} (\underline{S}_{\omega_2} - \underline{S}_\Omega)/(\underline{S}_\Omega + \underline{R}^2) , \qquad (20)$$

which under the hypothesis H_0 has an F-distribution with ν and m-p degrees of freedom.

Large values of $\underline{F}_1$, $\underline{F}_2$ and $\underline{F}_3$ indicate, that the hypotheses tested are not correct.

Literature

In most introductory texts on statistics the above tests are discussed, e.g.

F.A. GRAYBILL, An introduction to linear statistical models, Part I (Chapters 6 and 8), Mc Graw-Hill, New York 1961;

A.M. MOOD & F.A. GRAYBILL, Introduction to the theory of statistics (Chapter 13), Mc Graw-Hill, New York 1963.

Input:

code number of the tape

$$\begin{array}{lllll} k & m & n & p & q \\ B & C & D & & \\ t_{\alpha/2} & & & & \end{array}$$

$$\begin{array}{ccc} z_{1,1} & \cdots & z_{1,n} \\ \vdots & & \vdots \\ z_{k,1} & \cdots & z_{k,n} \end{array}$$

$$\begin{array}{cccc} m_1 & y_{1,1} & \cdots & y_{1,m_1} \\ \vdots & \vdots & & \\ m_n & y_{n,1} & \cdots & y_{n,m_n} \end{array}$$

$$\begin{array}{ccc} t_{1,1} & \cdots & t_{1,k} \\ \vdots & & \vdots \\ t_{p,1} & \cdots & t_{p,k} \\ \vdots & & \vdots \\ t_{p+q,1} & \cdots & t_{p+q,k} \end{array}$$

$$a_1 \cdots a_n$$

$$\left.\begin{array}{ll} \nu & \\ r_1 & \beta_{r_1}^{**} \\ \vdots & \vdots \\ r_\nu & \beta_{r_\nu}^{**} \end{array}\right\} \text{only if } D = 1$$

the program contains setXEENtal

Output:

{Result<code number of program>}
{Multiple Polynomial Regression}
{see descr 13}
{input}code number of input tape

{detK} det K

{r	b	sigma	t × sigma}
1	b_1	$\hat{\sigma}(b_1)$	$t_{\alpha/2} \times \hat{\sigma}(b_1)$
⋮	⋮	⋮	⋮
p	b_p	$\hat{\sigma}(b_p)$	$t_{\alpha/2} \times \hat{\sigma}(b_p)$

{SStot	S	df	R-squared	df	rho-squared}
$\sum_{i=1}^{n} \tilde{y}_i^2$	S_Ω	n-p	R^2	m-n	ρ^2

{F1	df	df	detK2	F2	df	df	detK3	F3	df	df}
F_1	n-p	m-n	$detK_2$	F_2	q	m-p-q	$detK_3$	F_3	ν	m-p

SORTING and RANKING

The procedures described here may be used to:

1) sort the sequence a_p, a_{p+1}, ..., a_r by calculating a row of integers k_p, k_{p+1}, ..., k_r for which $a_{k_j} \leq a_{k_{j+1}}$ $j = p, p+1, \ldots, r$;

2) obtain a lexicographically sorted frequency table of the numbers:

$$\begin{matrix} a_{p,q} & ,\text{--------}, & a_{p,k} \\ \vdots & & \\ a_{r,q} & ,\text{--------}, & a_{r,k} \end{matrix} \qquad (1)$$

by

a) calculating that sequence of integers k_p, ..., k_s for which the rows

$$\begin{matrix} a_{k_p,q} & ,\text{------}, & a_{k_p,k} \\ \vdots & & \\ a_{k_s,q} & ,\text{------}, & a_{k_s,k} \end{matrix}$$

are different and lexicographically sorted;

b) assigning such values to f_{k_p}, ..., f_{k_s} that f_{k_j} = the number of times the row $a_{k_j,q}$, ..., $a_{k_j,k}$ appears in (1);

3) assign to b_i ($i = p, p+1, \ldots, r$) the average rank-number of a_i in the sequence a_p, a_{p+1}, ..., a_r;
in this procedure a non-sorted or sorted frequency table of a_i is also permitted.

$$b_i = s_i + \frac{t_i + 1}{2} \qquad \text{where}$$

$$s_i = \sum_{j=p}^{r} \delta(a_j, a_i) \qquad \delta(a_j, a_i) = \begin{cases} 1 \text{ if } a_j < a_i \\ 0 \text{ otherwise} \end{cases}$$

$$t_i = \sum_{j=p}^{r} \gamma(a_j, a_i) \qquad \gamma(a_j, a_i) = \begin{cases} 1 \text{ if } a_j = a_i \\ 0 \text{ otherwise} \end{cases}$$

The method underlying these procedures is SHELL's sort procedure, which in turn is an application of the "sinking" method.
This "sinking" method interchanges the values of two adjacent elements a_i and a_{i+1} of a sequence $a_1, \ldots, a_n$, if $a_i > a_{i+1}$.

For example:

5	6	1	2	4
5	1	6	2	4
1	5	6	2	4
1	5	2	6	4
1	2	5	6	4
1	2	5	4	6
1	2	4	5	6

In SHELL's procedure this method is applied to sub-sequences $s_{i,j}$ of the sequence $a_1, a_2, \ldots, a_n$.

$$s_{i,j} = \{ a_j, a_{j+d_i}, a_{j+2d_i}, \ldots, a_{j+m_{ij}d_i} \}.$$

Here:

$$i = 1, 2, \ldots, t ,$$
$$j = 1, 2, \ldots, d_i ,$$
$$m_{ij} = \text{entier}\left(\frac{n - j}{d_i} \right) ,$$

t depends on the sequence d_1, d_2, ...
It seems to be difficult to determine the optimal sequence d_1, d_2, ..., d_t; for the present procedures this sequence is defined by $d_1 = \max(1, 2^d - 1)$ $d = \text{entier}(^2\log n) - 1$ and

$d_i = \text{entier}\left(\frac{d_{i-1}}{2}\right)$. SHELL's method is useful if the number of elements to be sorted does not exceed 200.
Now the headings of the procedures are given, the effect of each procedure is described with the help of the notation used in 1), 2) or 3).

procedure set1K (A, K, i, p, r);
value p, r; **integer** K, i, p, r; **real** A;

The values of A should depend on i, for i = p, p+1, ..., r should A be equal to a_i (see 1).
K must be a subscribted variable, also depending on i; the procedure assigns to K the numbers i if the correlation between a_i and i is positive and p+r-i if the correlation is negative (i = p, p+1, ..., r).

procedure sort1 (A, K, i, p, r);
value p, r; **integer** K, i, p, r; **real** A;

The values of A should depend on i and be equal to a_i (i = p, ..., r).
K must be a subscribted variable, also depending on i.
A call of sort1 is only permitted if the elements of K (i = p, ..., r) contain a permutation of the integers p, p+1, ..., r; for instance produced by set1K.
The effect of sort1 is: K:= k_i (i = p, ..., r) (see 1).

integer **procedure** SORT1 (A, i, p, r, j, q, k, K, F);
value p, r, q, k; **integer** i, j, p, r, q, k, K, F; **real** A;

The values of A should depend on i and j, and be equal to a_{ij}

(i = p, ..., r ; j = q, ..., k).
K and F must be subscribted variables, both depending on i.
A call of SORT1 is only permitted if the elements of K (i = p, ..., r) contain a permutation of the integers p, p+1, ..., r.
The effect of SORT1 is:
SORT1:= s , K:= k_i , F:= f_i (i = p, ..., s) (see 2)

SORT1 uses sort 1.

__procedure__ RANK1 (A, B, F, i, p, r);
__value__ p, r; __integer__ F, i, p, r; __real__ A, B;

The values of A and F should depend on i and be equal to a_i or the frequency of a_i respectively (i = p, ..., r).

B must be a subscribted variable, also depending on i.
The effect of RANK1 is: B:= b_i (i = p, ..., r).

RANK1 uses set1K and sort1.

If A is a subscribted variable, the call RANK1 (A, A, F, i, p, r) is permitted, in this case a_i is replaced by b_i (see 3).

Literature

Communications of the ACM, 6 (1963), number 5.

WILCOXON's W

The procedure described here may be used to compute Wilcoxon's W and some related quantities.

Given sorted frequency tables of two one-dimensional sets of observations:

$$\begin{array}{llll} x_m & f^{(x)}_m & y_p & f^{(y)}_p \\ \vdots & \vdots & \vdots & \vdots \\ x_n & f^{(x)}_n & y_q & f^{(y)}_q \\ x_i < x_{i+1} & & y_j < y_{j+1} & \end{array}$$

these quantities are defined:

$$W_1 = \sum_{j=p}^{q} \sum_{i=m}^{n} f^{(y)}_j \cdot f^{(x)}_i \cdot \delta(x_i, y_j)$$

where

$$\delta(x_i, y_j) = \begin{cases} 2 \text{ if } x_i > y_j \\ 1 \text{ if } x_i = y_j \\ 0 \text{ if } x_i < y_j \end{cases}$$

$$W_2 = W_5 \cdot W_6$$

$$W_3 = \left\{\frac{1}{3} \cdot \left[(W_5 + W_6)(W_5 + W_6 - 1)\right]^{-1} \cdot \left[(W_5 + W_6)^3 - W_4\right] \cdot W_2\right\}^{\frac{1}{2}}$$

$$W_4 = \sum_{j=p}^{q} \left\{f^{(y)}_j\right\}^3 + \sum_{i=m}^{n} \left\{f^{(x)}_i\right\}^3 + \sum_{j=p}^{q} \sum_{i=m}^{n} \gamma(x_i, y_j)$$

where

$$\gamma(x_i, y_j) = \begin{cases} 3 \cdot f^{(y)}_j \cdot \left\{f^{(x)}_i\right\}^3 + 3 \cdot f^{(x)}_i \cdot \left\{f^{(y)}_j\right\}^3 & \text{if } x_i = y_j \\ 0 \quad \text{if } x_i \neq y_j \end{cases}$$

$$W_5 = \sum_{i=m}^{n} f_i^{(x)}$$

$$W_6 = \sum_{j=p}^{q} f_j^{(y)} \quad .$$

```
procedure WILCOX (X, FX, i, m, n, Y, FY, j, p, q, W);
value m, n, p, q; integer i, m, n, j, p, q, FX, FY;
real X, Y; real array W;
```

X and FX are a real and an integer variable, which for i = m, m+1,..., n are x_i and $f_i^{(x)}$ respectively.
Y and FY are a real and an integer variable, which for j = p, p+1,..., q are y_j and $f_j^{(y)}$ respectively.
W is an output array: W[k]: = W_k k = 1,2,3,4,5,6.

Legitimate calls of WILCOX are

```
    WILCOX (X[H[k]], G[I[k]], k, p, r, X[1], G[1-10], 1, 10, 38, Q)
and WILCOX (A(j), B(j), j, i, h, C(d), D(d), d, h+1, k, R) .
```

COMBINATIONS

This procedure may be used to compute all $\binom{n}{t}$ combinations of t from n objects.

```
procedure COMBINA1 (T, t, n, first, READY);
value t, n;
integer t, n;
Boolean first;
integer array T;
label READY;
```

n different objects may be labeled 1, ..., n.
Each subset of t ($1 \leq t \leq n$) objects from this set of n objects is determined by the labels of these t objects.

There are $\binom{n}{t}$ different subsets of t from n objects, the corresponding sets of t labels may be sorted lexicographically:

```
1 , 2 , 3 , ...... , t-2 , t-1 , t
1 , 2 , 3 , ...... , t-2 , t-1 , t+1
'
'
'
1 , 2 , 3 , ...... , t-2 , t-1 , n
1 , 2 , 3 , ...... , t-2 ,  t  , t+1          (1)
'
'
'
1 , 2 , 3 , ...... , t-2 ,  t  , n
1 , 2 , 3 , ...... , t-2 , t+1 , t+2
'
'
'
n-t+1, n-t+2, ....... , n-2 , n-1 , n .
```

Each call of COMBINA1 generates the next line of (1) into the output array T: T[i] is made equal to the i-th number of the next line (i = 1, 2, ..., t).
If first = **true** this is the first line of (1), and first is made **false**.
If no next line is defined the call results in **goto** READY, in this case the contents of T are meaningless.
Between two calls of COMBINA1 the elements of T should not be altered.

The use of COMBINA1 may be demonstrated by the following parts of two programs:

```
       start: = true;
again: COMBINA1 (LABELS, sub, set, start, exit);                    (2)
       use LABELS; goto again;
exit:

back:  newset:= true;
again: COMBINA1 (objects, t from, n, newset, back);                 (3)
       use objects; goto again;
```

In (2) something is computed once for each of the different subsets, whereas in (3) after the last line of (1) the first is generated again.

GENERAL COMBINATIONS

The procedure presented here generates a combination of t from N objects, where the set S of N objects consists of the n different objects 1, 2, ..., n , the object h (h = 1, 2, ..., n) appearing F_h times.

All possible combinations of t from this set S can be arranged in a lexicographically sorted scheme. The combination generated is the line of the scheme following the last one generated.

For example: t = 2 n = 3 N = 6

$F_1 = 1$ $F_2 = 2$ $F_3 = 4$

or: S = { 1,2,2,3,3,3,3 } .

The possible combinations of 2 objects are:

1	2
1	3
2	2
2	3
3	3 .

<u>procedure</u> COMBINA2 (T, g, t, F, f, h, n, first, READY);
<u>value</u> t, n; <u>integer</u> T, g, t, F, f, h, n; <u>Boolean</u> first; <u>label</u> READY;

T is understood to be a subscribted integer variable, the value of which depends on g; T_g is made equal to the g-th element of the present combination, where g = 1, 2, ..., t.
F should be an integer variable, depending on h, $F_h > 0$ (h = 1, 2, ..., n), describing the set of $N = \sum_{h=1}^{n} F_h$ objects.
f must be a subscribted integer variable, depending on h (h = 1, 2, ..., n); f_h is made equal to the number of times h appears in the present combination.

If first = <u>true</u> the first combination is generated, and first is made <u>false</u>.

If no next combination is defined, the call of COMBINA2 results in <u>goto</u> READY, in this case the contents of T are meaningless.

Between two calls of COMBINA2 the elements of T and f should not be altered.

The use of COMBINA2 may be demonstrated by the following part of a program:

```
      eerste: = true;
next: COMBINA2 (R[1,i], i, 8, V[j,j], S[j,2], 10, eerste, out);
      use R and S; goto next;
out:
```

STANDARD NORMAL PROBABILITY FUNCTION

The standard normal probability function

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ can be approximated with the help of the

approximation H(x) of the function

$$G(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \qquad 0 \leq x.$$

$$H(x) = 1 - (1 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + a_6 x^6)^{-16}$$

where: a_1 = .0705230784 a_4 = .0001520143
a_2 = .0422820123 a_5 = .0002765672
a_3 = .0092705272 a_6 = .0000430638 ,

$|H(x) - G(x)| < .0000003.$

real procedure PHI(x); value x; real x;

$$PHI: = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Literature:

Dr. Hastings, Approximations for digital computers, sheet 63.

RANDOM NORMAL DEVIATES

Random normal deviates can be obtained in the following way:

Generate a random number x from the rectangular (0,1) distribution. Compute r such that

$$x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} e^{-\frac{t^2}{2}} dt \qquad \text{or} \qquad r = \Phi^{-1}(x)$$

An approximation of r can be computed with the help of the approximation H(q) of G(q):

$$q = \frac{1}{\sqrt{2\pi}} \int_{G(q)}^{\infty} e^{-\frac{t^2}{2}} dt \qquad 0 < q \leq .5$$

$$H(q) = z - \left\{\frac{a_0 + a_1 z + a_2 z^2}{1 + b_1 z + b_2 z^2 + b_3 z^3}\right\}$$

where: $z = \sqrt{\ln\left(\frac{1}{q^2}\right)}$

$a_0 = 2.515517$ $\qquad b_1 = 1.432788$

$a_1 = .802853$ $\qquad b_2 = .189269$

$a_2 = .010328$ $\qquad b_3 = .001308$

$|H(q) - G(q)| < .00045$

<u>real procedure</u> PHINV (x); <u>value</u> x; <u>real</u> x;

PHINV: = $\Phi^{-1}(x)$

Literature:

Dr. Hastings, Approximations for digital computers, sheet 68.

GAMMA and BETA FUNCTION (integer values of parameters)

The four procedures described here compute:
$\Gamma(n)$, $\Gamma_x(n) / \Gamma(n)$, $B(p,q)$, $B_x(p,q) / B(p,q)$ for positive integer values of the parameters.

The incomplete functions may be written as finite sums, these are computed with the help of the ratio between two subsequent terms:

$$\Gamma_x(n) = \int_0^x e^{-t}\, t^{n-1}\, dt = \left\{1 - e^{-x} \sum_{i=0}^{n-1} \frac{x^i}{i!}\right\} \cdot (n-1)! =$$

$$= \left\{1 - e^{-x} \sum_{i=0}^{n-1} t_i\right\} \cdot (n-1)!$$

where $t_0 = 1$, $t_i = \frac{x}{i} \cdot t_{i-1}$

$$B_x(p,q) = \int_0^x t^{p-1}\, (1-t)^{q-1}\, dt = x^p \sum_{i=0}^{q-1} (-1)^i \binom{q-1}{i} \frac{x^i}{i+p} =$$

$$= x^p \sum_{i=0}^{q-1} t_i$$

where $t_0 = \frac{1}{p}$, $t_i = \frac{x}{i} \cdot \frac{1-i-p}{i+p} \cdot t_{i-1}$.

real procedure C gamma 1 (n); **value** n; **integer** n;

C gamma 1: $= \Gamma(n)$

real procedure Q gamma 1 (n,x); **value** n,x; **integer** n; **real** x;

Q gamma 1: $= \Gamma_x(n) / \Gamma(n)$

real procedure C beta 1 (p,q); **value** p,q; **integer** p,q;

C beta 1: $= B(p,q)$

real procedure Q beta 1 (p,q,x); **value** p,q,x; **integer** p,q; **real** x;

Q beta 1: $= B_x(p,q) / B(p,q)$

DISCRETE DISTRIBUTION FUNCTIONS

The program presented here tabulates the distribution functions of the

Poisson,
Binomial,
Negative Binomial and
Hypergeometric distribution.

The probabilities are computed with the help of the following relations:

1. <u>Poisson distribution</u>

$$p_i = \frac{e^{-p}\, p^{2i}}{i!} \qquad i = 0, 1, \ldots \quad p > 0$$

$$\frac{p_i}{p_{i-1}} = \frac{p}{i}$$

2. <u>Binomial distribution</u>

$$p_i = \binom{n}{i} p^i (1 - p)^{n-i} \qquad i = 0, 1, \ldots, n \quad n = \text{positive integer}$$

$$\frac{p_i}{p_{i-1}} = \frac{p(n - i + 1)}{i(1 - p)}$$

3. <u>Negative Binomial distribution</u>

$$p_i = \binom{i-1}{n-1} p^n (1 - p)^{i-n} \qquad i = n, n+1, \ldots \quad n = \text{positive integer}$$

$$\frac{p_i}{p_{i-1}} = \frac{(1 - p)(i - 1)}{(i - n)}$$

4. <u>Hypergeometric distribution</u>

$$p_i = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \qquad i = i_1, i_1+1, \ldots, i_2 \text{ where}$$

$i_1 = \max(0, M - N + n)$ $\quad$ $i_2 = \min(M, n)$

N, n, M = positive integers

$M \leq N$, $n \leq N$

$$\frac{p_i}{p_{i-1}} = \frac{(M - i + 1)(n - i + 1)}{(N - M - n + i)i}$$

The distribution functions are tabulated for the j's for which

$$.000001 \leq \sum_{i=s}^{j-1} p_i \leq .999999,$$

where s stands for the starting point of the distribution.

Input:

code number of the tape

number of functions to be tabulated

for each function one of the following lines:

1	p			(Poisson distribution)
2	n	p		(Binomial distribution)
3	n	p		(Negative Binomial distribution)
4	N	n	M	(Hypergeometric distribution)

Output

{Result <code number of the program>}
{Discrete distribution functions, see descr 26, input} code number of tape for each function:

{<name of the function>} <parameters>
{ 0 1 2 3 4 5 6 7 8 9}
Table of the function,
each line preceded by an integer k_l (an integermultiple of 10) such that the number in the j-th column and l-th row of the table is equal to

$$\sum_{i=s}^{k_l+j} p_i \, .$$

GAMMA II

The procedures presented in this description compute:

$$\Gamma\left(\frac{n}{2}\right) = \int_0^\infty e^{-t}\, t^{\frac{n}{2}-1}\, dt, \quad n = \text{positive integer}$$

$$\frac{\Gamma_x\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{\int_0^x e^{-t}\, t^{\frac{n}{2}-1}\, dt}{\Gamma\left(\frac{n}{2}\right)}, \quad n = \text{positive integer}, \quad x \geq 0$$

The following formulae are used:

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} \sqrt{\pi} & \text{if } n = 1 \\ 1 & \text{if } n = 2 \\ \left(\frac{n}{2} - 1\right) \cdot \Gamma\left(\frac{n}{2} - 1\right) & \text{if } 3 \leq n \leq 6 \\ e^{-\frac{n}{2}} \cdot \left(\frac{n}{2}\right)^{\frac{n}{2} - \frac{1}{2}} \cdot \sqrt{2\pi} \cdot \left[1 + \frac{1}{6n} + \frac{1}{72n^2}\right] & \text{if } n \geq 7 \end{cases}$$

This asymptotic formula has, for $n \geq 7$, a relative error $\leq .0002$

$$\frac{\Gamma_x(\frac{n}{2})}{\Gamma(\frac{n}{2})} = \begin{cases} 1 - 2\Phi(-\sqrt{2x}) & \text{if } n = 1; \\[1ex] 1 - e^{-x} & \text{if } n = 2; \\[1ex] \dfrac{(\frac{n}{2}-1)\Gamma_x(\frac{n-2}{2}) - e^{-x}\,x^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})} & \text{if } 3 \le n \le 6; \\[2ex] \dfrac{x^{\frac{n}{2}}e^{-x}}{\Gamma(\frac{n}{2}+1)}\left[1 + \dfrac{x}{(\frac{n}{2}+1)} + \ldots + \dfrac{x^{k-1}}{(\frac{n}{2}+1)\ldots(\frac{n}{2}+k-1)}\right] & \text{if } \begin{array}{l} x \le \frac{n}{2} \\ 7 \le n < 100 \end{array} \\[1ex] \text{here the summation is stopped if} & \\ \left(\dfrac{x}{\frac{n}{2}+k}\right)^k \le .001\text{, this ensures a relative error in} & \\ \dfrac{\Gamma_x(\frac{n}{2})}{\Gamma(\frac{n}{2})} \text{ of at most } .001; & \\[2ex] 1 - \dfrac{x^{\frac{n}{2}-1}e^{-x}}{\Gamma(\frac{n}{2})}\left[1 + \dfrac{\frac{n}{2}-1}{x} + \ldots + \dfrac{(\frac{n}{2}-1)\ldots(\frac{n}{2}-k+1)}{x^{k-1}}\right] & \text{if } \begin{array}{l} x > \frac{n}{2} \\ 7 \le n < 100 \end{array} \\[1ex] \text{here the summation is stopped if} & \\ \dfrac{(\frac{n}{2}-1)\ \ldots\ (\frac{n}{2}-k)}{x^k} \le .001\text{, ensuring a relative error} & \\ \le .001 \text{ in } 1 - \dfrac{\Gamma_x(\frac{n}{2})}{\Gamma(\frac{n}{2})}\ ; & \\[2ex] \Phi\left(\left\{\left(\dfrac{2x}{n}\right)^{\frac{1}{3}} + \dfrac{2}{9n} - 1\right\} . \sqrt{\dfrac{9n}{2}}\right) & \text{if } n \ge 100. \end{cases}$$

Here $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{u^2}{2}}\,du$ (see descr 23).

real procedure Gamma2(n); **value** n; **integer** n;

$$Gamma2 := \Gamma\left(\frac{n}{2}\right)$$

real procedure Qgamma2(n, x); **value** n, x; **integer** n; **real** x;

$$Qgamma2 := \frac{\Gamma_x\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

Qgamma2 uses Cgamma2 and PHI.

Literature

Handbook of Mathematical Functions
N.B.S. - A.M.S. 55 (1964) Ch. 6 and 26.

STUDENT's t-DISTRIBUTION

The probability function of Student's t, m degrees of freedom, is defined as

$$F_m(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{\pi m}\,\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^{t} \left(1 + \frac{u^2}{m}\right)^{-\frac{m+1}{2}} du.$$

For $t \geq 0$ the following formulae may be used to compute $F_m(t)$:

$$F_m(t) = \frac{1}{2} + \frac{1}{2}\sin\alpha \left\{ 1 + \frac{1}{2}\cos^2\alpha + \frac{1.3}{2.4}\cos^4\alpha + \ldots + \frac{1.3\ldots(m-3)}{2.4\ldots(m-2)}\cos^{m-2}\alpha \right\}$$

for m even

$$F_m(t) = \frac{1}{2} + \frac{\alpha}{\pi} + \frac{\sin\alpha}{\pi} \left\{ \cos\alpha + \frac{2}{3}\cos^3\alpha + \ldots + \frac{2.4\ldots(m-3)}{3.5\ldots(m-2)}\cos^{m-2}\alpha \right\}$$

for m odd

where $\alpha = \arctan\left(\frac{t}{\sqrt{m}}\right)$.

The finite sums are computed with the help of the ratio between two subsequent terms.

For $t < 0$ the relation $F_m(t) = 1 - F_m(-t)$ is used.

<u>real</u> <u>procedure</u> tDistr (m,t); <u>value</u> m,t; <u>integer</u> m; <u>real</u> t;

tDistr: = $F_m(t)$.

Corrections for Ties

In several distribution-free techniques the following problem arises: given the numbers

$$\begin{array}{cc} x_p & f_p \\ \vdots & \vdots \\ x_r & f_r \end{array}$$

where f_i denotes the frequency of the (1-dimensional) observation x_i, but these x_i's are neither different nor given in ascending order, the quantities

$$T_1 = \sum t_j \qquad T_2 = \sum t_j^2 \qquad T_3 = \sum t_j^3$$

are to be computed.
Here t_j denotes the size of the j-th tie in the x_i's.
Clearly $T_1 = \sum_{i=p}^{r} f_i$, as each t_j is the sum of some f_i's.

T_1, T_2 and T_3 may be computed by the

<u>procedure</u> TIECORR (A, F, i, p, r, T); <u>value</u> p, r;
<u>integer</u> F, i, p, r; <u>real</u> A; <u>real</u> <u>array</u> T;

The value of A and F are supposed to depend on i (i = p, ..., r) and be equal to x_i and f_i respectively.
TIECORR assigns to T[j] the value T_j (j = 1, 2, 3).

It has proved to be useful to combine the computing of T_1, T_2 and T_3 with the ranking of the observations.

This is possible with the

procedure RANK2 (A, B, F, i, p, r, T); **value** p, r;
integer F, i, p, r; **real** A, B; **real** **array** T;

This procedure has the same effect on A, B, F, i, p, r as RANK1 (A, B, F, i, p, r) (see descr 18) and as TIECORR (A, F, i, p, r, T) on the output array T.

Both TIECORR and RANK2 use set1K and sort1 (see descr 18).

ANALYSIS OF VARIANCE (FIXED EFFECTS; GENERAL MODEL)

Let $\underline{y}_{ij}$, $j = 1, 2, \ldots, n_i$, $n_i \geq 1$, $i = 1, 2, \ldots, k$, denote k independent random samples from normal distributions with equal variances but possibly different expectations:

$\underline{y}_{ij}$ independent, $\underline{y}_{ij} \in N(\eta_i, \sigma^2)$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$.

We write

$$n = \sum_{i=1}^{k} n_i, \quad \underline{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{y}_{ij}, \quad \underline{y}_{..} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \underline{y}_{ij}.$$

A model is given by the sets of equations

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j, \quad i = 1, 2, \ldots, k, \quad x_{ij} = 0 \text{ or } 1;$$

$$0 = \sum_{j=1}^{p} h_{ij} \beta_j, \quad i = 1, 2, \ldots, l.$$

This means that the expectation of an observation in the i^{th} sample (or "cell") is obtained by adding a number of the "effects" $\beta_1, \beta_2, \ldots, \beta_p$, which are themselves subject to l linear constraints.
We define

η : the column vector $(\eta_1, \eta_2, \ldots, \eta_k)$;

$\underline{\mathring{y}}$: the column vector $(y_{1.}, y_{2.}, \ldots, y_{k.})$;

β : the column vector $(\beta_1, \beta_2, \ldots, \beta_p)$;

N : the kxk diagonal matrix with elements $n_{ij} = \delta_{ij} n_i$, when $\delta_{ij} = 1$ if $i = j$ and 0 otherwise;

X : the matrix (x_{ij}) with k rows and p columns;

H : the matrix (h_{ij}) with l rows and p columns;

$\begin{pmatrix} X \\ H \end{pmatrix}$: the $(k + l)$xp matrix of which the first k rows are the rows of X and the last l rows are the rows of H.

The transpose of a matrix will be denoted by adding a prime, e.g. X' is the pxk matrix (x_{ij}).

In matrix notation the model becomes

$$\eta = X\beta$$
$$0 = H\beta.$$

The model is consistent, determines β uniquely and β is estimable, if rank $X = r$, rank $H = p - r$, rank $\begin{pmatrix} X \\ H \end{pmatrix} = p$.

The following quantities are computed (with each sum of squares the number of degrees of freedom (df.) is given):

1. The total sum of squares

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{..})^2 \qquad \text{df } (n - 1)$$

2. The residual sum of squares

$$R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 \qquad \text{df } (n - k)$$

3. The determinant

$$\det (X'NX + H'H).$$

This is meant to provide a check on the model, since the ranks of X, H and $\begin{pmatrix} X \\ H \end{pmatrix}$ are not determined in the program. If rank $\begin{pmatrix} X \\ H \end{pmatrix} < p$ the matrix $X'NX + H'H$ is singular and the remainder of the program will produce nonsensical results. If the determinant in absolute value is smaller than 10^{-10} the program is stopped accordingly.

4. The column vector $b = (b_1, b_2, \ldots, b_p)$ of least squares estimates of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$

$$b = (X'NX + H'H)^{-1} X'N\mathring{y}.$$

5. The column vector $a = (a_1, a_2, \ldots, a_k)$ of least squares estimates of $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ for the model given

$$a = Xb.$$

6. The sum of squares

$$S - R = \sum_{i=1}^{k} n_i (y_{i.} - a_i)^2 \qquad \text{df } (k - r)$$

where

$$S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - a_i)^2$$

denotes the minimal sum of squares for the model given.

7. The test statistic for the model against the underlying assumption

$$F = \frac{S - R}{R} \cdot \frac{n - k}{k - r};$$

under the assumptions of the model F has an F-distribution with $k - r$ and $n - k$ df. With F, $k - r$ and $n - k$ the corresponding right hand tail probability P is given in the output.

8. The estimated covariance-matrix $C(b)$ of b

$$C(b) = \frac{S}{n - r} \cdot ZNZ'$$

where $$Z = (X'NX + H'H)^{-1} X'.$$

The special feature of this program in its great generality: it processes analysis of variance results with unequal numbers n_i of observations per cell as well as missing plots. Since the matrix inversion in the program can be rather time-consuming, it should preferably not be used for routine designs with equal n_i.
When using the program one should realize that generally, for unequal numbers of observations per cell, the F-test for the existence of a given set of effects depends upon the model as a whole, i.e. on what other effects one wishes to retain in the model.

If missing plots occur ($n_i = 0$) the corresponding cells are simply emitted from the model. One often has to be extremely careful in this case that β is uniquely determined by the model and estimable (e.g. interactions cannot be separated from main effects, etc.).

It may be necessary to try a number of different models of increasing complexity before a satisfactory one is found. The estimated covariance matrix C(b) may thereafter be used to carry out multiple comparison procedures. The program provides this covariance matrix only if the test statistic F for the model satisfies

$$P \geq \alpha$$

where α is some preassigned confidence level.
For P-values smaller than α the model is not supposed to be satisfactory anyway. It is also possible to have the computer print only parts of the matrix C(b) in case $P \geq \alpha$.

Literature

H. SCHEFFÉ, The analysis of variance, Wiley New York, 1959.

Note that the matrices X and H are denoted by X' and H' by SCHEFFÉ.

Input:

code number

k

the k samples, each given in one of the two forms described in note 3

α

the number of models to be tried

the models, each consisting of:

code number of the model

l

p

r

$\sum_{j=1}^{p} m_j$

h_{11} - - - - - - h_{1p}

⋮

h_{l1} - - - - - - h_{lp}

m_1 z_{11} - - - - - - - z_{1m_1}

⋮

m_p z_{p1} - - - - - - - z_{pm_p}

where:

$m_j = \sum_{i=1}^{k} x_{ij}$ = number of 1's in the j^{th} column of X

z_{ji} such that $x_{z_{ji}j} = 1$ and $z_{j,i-1} < z_{j,i}$

so $z_{j1}, - - ---, z_{jm_j}$ are the row-indices of the elements of the j-th column of X which are equal to 1.

s = number of parts of C(b) to be computed

l_1 ... l_s, u_1 ... u_s see description of output

The program contains XEENtal.

Output:

⊰Result<code number of program>⊱
⊰Analysis of Variance, see descr 40⊱
⊰samples⊱ code number

⊰T	df	R	df ⊱
T	n-1	R	n-k

of each model:

⊰model nr⊱ code number of model
⊰determinant⊱ det (X'NX + H'H)
⊰vector b⊱
b_1, — — — — — , b_p
⊰vector a⊱
a_1, — — — — — , a_k

⊰S - R	df	F	df	df	P⊱
S - R	k-r	F	k-r	n-k	P

⊰parts of matrix C(b)⊱
the s parts of C(b), t = 1, ..., s:

l_t u_t

$c_{l_t\, l_t}$ — — — —
⋮
$c_{u_t\, l_t}$ — — — — — — — — $c_{u_t\, u_t}$

when c_{ij} = covariance (b_i, b_j).

STAT programs
contents dd. 10-3-65

STAT

```
begin   comment  program JMA 100364 10693, STAT 001,
                 Matrices of Crossproduct-sums, Covariances and Correlations,
                 see descr 2;
        integer p,n,m,tal,band,n1;
        boolean nfq;

        NLCR;PRINTTEXT(⟨JMA 100364 10693 put input tape into reader, set XEENtal, press BVA ⟩);stop;
        tal:=XEEN(-0);
PPP:    band:=read;n:=read;m:=read;p:=(m+1)×(m+2):2;nfq:=n<0;if nfq then n:=-n;

        begin   integer k1,k2,k3,k4,k5;
                real a;
                boolean een,twee,drie,vier;
                integer array N[1:m+1];
                real array M[1:p];
                procedure OUTPUT(naam,code);integer code;string naam;
                begin   integer i,j,k;

                        RUNOUT;PUNLCR;PUTEXT(⟨Result JMA 100364 10693,  STAT 001⟩);PUNLCR;
                        PUTEXT(⟨see descr 2⟩);PUNLCR;PUTEXT(⟨Matrix of ⟩);PUTEXT(naam);
                        FIXP(10,0,code);FIXP(10,0,m+1);k:=0;
                        for i:= 1 step 1 until m+1 do
                        begin
                                PUNLCR;PUNLCR;FIXP(6,0,N[i]);
                                for j:= 1 step 1 until i do
                                begin
                                        k:=k+1;FLOP(6,2,M[k]);
                                        if j:10×10=j then begin  PUNLCR;PUSPACE(8)  end
                                end
                        end;
                        TAPEND
                end;
```

STAT 001

```
        for k1:= 1 step 1 until p do M[k1]:=0;
        if m=1 then N[1]:= 1 else for k1:= 1 step 1 until m do N[k1]:=read;N[m+1]:=11111;
        k5:=0;n1:=(available-100):(2×m+(if nfq then 0 else 1));
LAB:    k5:=k5+n1;if k5>n then n1:=n-k5+n1;

        begin   integer array F[1:(if nfq then 1 else n1)];
                real array X[1:n1,1:m];

                for k1:= 1 step 1 until n1 do
                begin
                        for k2:= 1 step 1 until m do X[k1,k2]:=read;if ¬nfq then F[k1]:=read
                end;
                k3:=0;
                for k1:= 1 step 1 until m do
                for k2:= 1 step 1 until k1 do
                begin
                        k3:=k3+1;M[k3]:=M[k3]+(if nfq then INPROD(k4,1,n1,X[k4,k1],X[k4,k2])
                                                         else SUM(k4,1,n1,X[k4,k1]×X[k4,k2]×F[k4]))
                end;
                for k1:= 1 step 1 until m do
                begin
                        k3:=k3+1;M[k3]:=M[k3] (if nfq then SUM(k4,1,n1,X[k4,k1])
                                                         else INPROD(k4,1,n1,X[k4,k1],F[k4]))
                end;
                M[p]:=M[p]+(if nfq then n1 else SUM(k4,1,n1,F[k4]))
        end;
        if k5<n then goto LAB;een:= read=1;twee:= read=1;drie:= read=1;vier:= read=1;
        if een then OUTPUT(⟨Crossproduct sums⟩,100×band+1);
        k3:=0;k4:=m×(m+1):2;k5:=M[p];if ¬twee then goto LAC;

        RUNOUT;PUNLCR;PUTEXT(⟨Result JMA 100364 10693,  STAT 001⟩);PUNLCR;
        PUTEXT(⟨see descr 2⟩);PUNLCR;PUTEXT(⟨Means and Variances⟩);
        FIXP(10,0,100×band+2);FIXP(10,0,m);PUNLCR;
```

```
        PUTEXT(⊰index      N         mean           variance    ⊱);
        for k1:= 1 step 1 until m do
        begin
                PUNLCR;if m≠1 then FIXP(6,0,N[k1]) else PUSPACE(8);FIXP(6,0,k5);a:=M[k4+k1]/k5;
                FLOP(6,2,a);FLOP(6,2,(M[k1×(k1+1):2]-a×a×k5)/(k5-1))
        end;
        TAPEND;
LAC:    if ¬drie ∧ ¬vier then goto OUT;

        for k1:= 1 step 1 until m do
        begin
                a:=M[k4+k1]/k5;
                for k2:= 1 step 1 until k1 do
                begin
                        k3:=k3+1;M[k3]:=(M[k3]-a×M[k4+k2])/(k5-1)
                end
        end;
        for k1:= 1 step 1 until m do M[k4+k1]:=M[k4+k1]/k5;
        if drie then OUTPUT(⊰Covariances⊱,100×band+3);if ¬vier then goto OUT;k3:=0;

        for k1:= 1 step 1 until m do
        begin
                a:=sqrt(M[k1×(k1+1):2]);
                for k2:= 1 step 1 until k1-1 do
                begin
                        k3:=k3+1;M[k3]:=M[k3]/a/sqrt(M[k2×(k2+1):2])
                end;
                k3:=k3+1
        end;
        for k1:= 1 step 1 until m do M[k1×(k1+1):2]:= 1;OUTPUT(⊰Correlations⊱,100×band+4);
    end;
OUT: tal:=tal-1;if tal>0 then goto PPP;NLCR;PRINTTEXT(⊰ READY ⊱)
end
```

STAT 001
continued

```
begin   comment program JMA 270563 9603, STAT 002,
                Stepwise Regression,
                see descr 3;
        integer tal,k;

        NLCR;PRINTTEXT(⊀JMA 270563 9603, put input tape into reader, set XEENtal, press BVA⊁);stop;
        tal:=XEEN(-0);
PPP:    RUNOUT;PUTEXT1(⊀

Result JMA 270563 9603 STAT 002
Stepwise Regression
see descr 3

input⊁);ABSFIXP(10,0,read);k:=read;

        begin   integer k1,k2,k3,te,con,stap,m;
                real a,b,c;
                boolean v;
                integer array N,M[1:k];
                real array A[1:k,1:k],C[1:k];
                procedure WSL(een,twee);integer een,twee;
                begin   integer r;
                        real t;

                        for r:= 1 step 1 until een-1 do
                        begin
                                t:=A[r,een];A[r,een]:=A[r,twee];A[r,twee]:=t
                        end;
                        for r:= een+1 step 1 until twee-1 do
                        begin
                                t:=A[een,r];A[een,r]:=A[r,twee];A[r,twee]:=t
                        end;
                        t:=A[een,een];A[een,een]:=A[twee,twee];A[twee,twee]:=t;
```

```
                for r:= twee+1 step 1 until k do
                begin
                        t:=A[een,r];A[een,r]:=A[twee,r];A[twee,r]:=t
                end;
                for r:= 1 step 1 until stap do
                begin
                        t:=A[een+1,r];A[een+1,r]:=A[twee+1,r];A[twee+1,r]:=t
                end;
                r:=M[een];M[een]:=M[twee];M[twee]:=r
        end WSL;

        for k1:= 1 step 1 until k do
        begin
                N[k1]:=read;for k2:= 1 step 1 until k1 do A[k2,k1]:=read
        end;
        ABSFIXP(10,0,read);te:=read;con:=read;m:=read;
        for k1:= 1 step 1 until k do M[k1]:=k1;v:=true;stap:=0;
L1:     k1:=read;if k1=-1 then goto L3;k2:=0;
L2:     if k2=k then goto L41;k2:=k2+1;if k1≠N[M[k2]] then goto L2;WSL(k2,k);k:=k-1;goto L1;
L3:     if te=N[M[k]] then goto HIER;k2:=0;
L4:     if k2=k then goto L41;k2:=k2+1;if te≠N[M[k2]] then goto L4;WSL(k2,k);goto HIER;
L41:    NLCR;PRINTTEXT(⟨input contradictory⟩);stop;
```

```
HIER:   stap:=stap+1;te:=0;a:=0;
        for k1:= stap step 1 until k-1 do
        begin
                if A[k1,k1]<⏨-11 then goto NIET;
                b:=A[k1+1,stap]:=A[k1,k]↑2/A[k1,k1];
                if b>a then begin  a:=b;te:=k1  end;
        NIET:
        end;
        if te=0 then goto DAAR;
        if ¬v then goto L6;k1:=read;
        if k1=-1 then v:=false else te:=k1;if ¬v then goto L6;k1:=stap-1;
L5:     if k1=k-1 then goto L41;k1:=k1+1;
        if te≠N[M[k1]] then goto L5;te:=k1;N[M[k1]]:=-N[M[k1]];
L6:     if te≠stap then WSL(stap,te);a:=A[stap,stap];
        for k1:= stap step 1 until k do A[stap,k1]:=A[stap,k1]/a;
        for k1:= stap+1 step 1 until k-1 do
        for k2:= k1 step 1 until k do
        A[k1,k2]:=A[k1,k2]-A[stap,k1]×A[stap,k2]×a;
        if stap<m then goto HIER;

DAAR:   if te=0 then stap:=stap-1;c:=0;a:=A[k,k];
        PUNLCR;PUTEXT1(‹Sum of squares›);FLOP(6,2,a);
        PUNLCR;PUTEXT1(‹
step    variable      R          Rcum          Rper       Rcumper›);
        for k1:= 1 step 1 until stap do
        begin
                PUNLCR;ABSFIXP(3,0,k1);ABSFIXP(8,0,N[M[k1]]);b:=A[k1+1,k1];c:=c+b;
                FLOP(6,2,b);FLOP(6,2,c);ABSFIXP(3,2,100×b/a);ABSFIXP(3,2,100×c/a);
                if con=abs(N[M[k1]]) then PUTEXT1(‹constant  ›);
                if N[M[k1]]<0 then PUTEXT1(‹forced›)
        end;
```

```
                  PUNLCR;PUTEXT1(≮
survey of R≯);
                  for k1:= 2 step 1 until k do
                  begin
                           k3:= if k1<stap+1 then k1-1 else stap;PUNLCR;if k1=stap+2 then PUNLCR;
                           ABSFIXP(6,0,N[M[k1-1]]);
                           for k2:= 1 step 1 until k3 do
                           begin
                                    FLOP(6,2,A[k1,k2]);if k2:10×10=k2 then begin  PUNLCR;PUSPACE(8)  end
                           end
                  end;
                  PUNLCR;PUTEXT1(≮
regression coefficients≯);
                  for k1:= 1 step 1 until stap do
                  begin
                           C[1]:=A[k1,k];A[k1,k1]:=0;
                           for k2:= 2 step 1 until k1 do C[k2]:=-SUM(k3,1,k2-1,A[k1-k2+1,k1-k3+1]×C[k3]);
                           PUNLCR;ABSFIXP(6,0,k1);
                           for k2:= 1 step 1 until k1 do
                           begin
                                    a:=A[k1+1,k2]:=A[k1,k2]+C[k1-k2+1];FLOP(6,2,a);
                                    if k2:10×10=k2 then begin  PUNLCR;PUSPACE(8)  end
                           end
                  end;
                  STOPCODE
         end;
         tal:=tal-1;if tal>0 then goto PPP;NLCR;PRINTTEXT(≮READY≯)
end
```

STAT 002
continued

SP procedures

contents dd. 10-3-65

SP		
100		
101		
102		
103		
104		
105		
106		
107		
108		
109		
110		
111		
112	replaced by SP 142	
113	sort1	see descr 18
114	SORT1	see descr 18
115	RANK1	see descr 18
116	C gamma 1	see descr 25
117	Q gamma 1	see descr 25
118	C beta 1	see descr 25
119	Q beta 1	see descr 25
120	replaced by SP 139	
121		
122		
123	COMBINA 1	see descr 20
124		
125		
126		
127		
128		
129	PHINV	see descr 24
130	WILCOX	see descr 19a

SP procedures
contents dd. 30-3-65
Continued

```
comment        SP 112, see descr 18;
procedure set1K(A,K,i,p,r);value p,r;integer K,i,p,r;real A;
begin   integer h;
        real a,b,c;
        Boolean g;

        a:=b:=0;i:=p-1;
L0:     i:=i+1;c:=A;a:=a+c;b:=b+c↑2;if i<r then go to L0;h:=p+r;g:= b>a×h/2;i:=p-1;
L1:     i:=i+1;K:=if g then i else h-i;if i<r then go to L1
end set1K;
```

```
comment        SP 113, see descr 18;
procedure sort1(A,K,j,p,r);value p,r;integer K,j,p,r;real A;
begin   integer e,d,h,i,m,k;
        real a,b;

        d:=entier(ln(r-p+1)/ln(2))-1;if d≤0 then d:=1;d:=2↑d-1;
L:      m:=p-1;
L0:     j:=i:=m:=m+1;if m≥p+d then go to L4;j:=K;a:=A;
L1:     j:=i:=i+d;if i>r then go to L0;k:=j:=K;b:=A;if a>b then go to L11;a:=b;go to L1;
L11:    h:=i;j:=i:=h-d;e:=K;j:=h;K:=e;
L2:     j:=i:=i-d;if i<m then go to L3;e:=j:=K;if b≥A then go to L3;j:=i+d;K:=e;go to L2;
L3:     j:=i+d;K:=k;i:=h;go to L1;
L4:     d:=d:2;if d>0 then go to L
end sort1;
```

```
comment        SP 114, see descr 18;
integer procedure SORT1(A,i,p,r,j,q,k,K,F);value p,r,q,k;integer i,p,r,j,q,k,K,F;real A;
begin   integer e,g,h;
        real t;

        j:=q;sort1(A,K,i,p,r);
L0:     i:=g:=p;if p=r then go to L3;i:=K;t:=A;
L1:     i:=g+1;i:=K;if A≠t then go to L2;g:=g+1;if g<r then go to L1;
L2:     if g=p∨j=k then go to L3;p:=SORT1(A,i,p,g,j,q+1,k,K,F);j:=q;go to L6;
L3:     i:=p;i:=K;F:=g-p+1;
L6:     r:=r-g+p;if p≠g then for h:=p+1 step 1 until r do begin  i:=h+g-p;e:=K;i:=h;K:=e  end;
L5:     if p=r then go to L4;p:=p+1;go to L0;
L4:     SORT1:=p
end SORT1;
```

```
comment        SP 115, see descr 18;
procedure RANK1(A,B,F,i,p,r);value p,r;integer F,i,p,r;real A,B;
begin   integer g,h,m,n,k;
        real a,b;
        integer array K[p:r];

        set1K(A,K[i],i,p,r);sort1(A,K[i],i,p,r);i:=K[p];a:=A;m:=g:=p;h:=F;n:=0;
L1:     if g=r then go to L3;g:=g+1;i:=K[g];b:=A;if b≠a then go to L2;h:=h+F;go to L1;
L2:     g:=g-1;
L3:     a:=b;b:=n+(h+1)/2;for k:=m step 1 until g do begin  i:=K[k];B:=b  end ;
        if g<r then begin  n:=n+h;m:=g:=g+1;i:=K[g];h:=F;go to L1  end
end RANK1;
```

SP 114-115

```
comment        SP 116, see descr 25;
real procedure Cgamma1(n);value n;integer n;
begin   integer i;
        real fac;

        fac:= 1;i:= 0;
LL:     if i = n-1 then go to MM;i:=i+1;fac:=fac × i;go to LL;
MM:     Cgamma1:= fac
end Cgamma1;


comment        SP 117, see descr 25;
real procedure Qgamma1(n,x);value n,x;integer n;real x;
begin   integer i;
        real s,t;

        s:=t:= 1;i:= 0;
LL:     if i = n-1 then go to MM;i:=i+1;t:=t×x/i;s:=s+t;go to LL;
MM:     Qgamma1:= 1-s×exp(-x)
end Qgamma1;


comment        SP 118, see descr 25;
real procedure Cbeta1(p,q);value p,q;integer p,q;
begin   integer i;
        real a;

        a:= 1/(p+q-1);i:= 0;
LL:     if i = q-1 then go to MM;i:=i+1;a:=a×i/(i+p-1);go to LL;
MM:     Cbeta1:= a
end Cbeta1;


comment        SP 119, see descr 25;
real procedure Qbeta1(p,q,x);value p,q,x;integer p,q;real x;
begin   integer i;
        real a,s,t;

        s:=t:= x↑p/p;i:= 0;a:= 1/(p+q-1);
LL:     if i = q-1 then go to MM;
        i:=i+1;t:=t×x×(q-i)×(1-i-p)/(i+p)/i;
        s:=s+t;a:=a×i/(i+p-1);go to LL;
MM:     Qbeta1:= s/a
end Qbeta1;
```

```
comment         SP 120, see descr 23;
real procedure PHI(x);value x;real x;
begin    real procedure psi(x);value x;real x;

         psi:= 1 -
         (1+.0705230784×x   +.0422820123×x↑2
           +.0092705272×x↑3+.0001520143×x↑4
           +.0002765672×x↑5+.0000430638×x↑6)↑(-16);

         PHI:=
         if x>0 then .5 + psi(  x/1.4142135623731 )/2
                else .5 - psi( -x/1.4142135623731 )/2

end PHI;
```

```
comment        SP 123, see descr 20;
procedure COMBINA1(T,t,n,first,READY);value t,n;integer t,n;Boolean first;
integer array T;label READY;
begin    integer k1;

         k1:=t+1;
FIRST:   if ¬first then go to NEXT;
         for k1:= 1 step 1 until t do T[k1]:=k1;first:=false;go to OUT;
NEXT:    k1:=k1-1;if k1=0 then go to READY;if T[k1]=n-t+k1 then go to NEXT;
         T[k1]:=T[k1]+1;for k1:= k1+1 step 1 until t do T[k1]:=T[k1-1]+1;
OUT:
end COMBINA1;
```

```
comment        SP 129, see descr 24;
real procedure PHINV(x);value x;real x;
begin    real procedure psi(r);value r;real r;
         begin
                  r:= sqrt(ln(1/r/r));
                  psi:= r - (2.515517  + .802853×r  + .010328×r↑2)/
                            (1         +1.432788×r  + .189269×r↑2  + .001308×r↑3)
         end;

         PHINV:= if x≤.5 then -psi(x) else psi(1-x)

end PHINV;
```

```
comment        SP 130, see descr 19;
procedure WILCOX(X,FX,i,m,n,Y,FY,j,p,q,W);value m,n,p,q;integer i,m,n,j,p,q,FX,FY;real X,Y; real array W;
begin   integer fx,fy,f,xf,yf;
        real x,y,w,d;

        i:=m-1;f:=0;d:=0;
cfx:    i:=i+1;fx:=FX;f:=f+fx;d:=d+fx↑3;if i<n then go to cfx;xf:=f;yf:=0;w:=0;i:=m-1;j:=p-1;
L0:     j:=j+1;y:=Y;fy:=FY;d:=d+fy↑3;yf:=yf+fy;
L1:     i:=i+1;x:=X;fx:=FX;if x>y then go to L2;if x≠y then go to L3;w:=w+fx×fy;d:=d+3×fx×fy↑2+3×fy×fx↑2;
L3:     f:=f-fx;if i<n then go to L1;go to L4;
L2:     w:=w+2×fx×fy;i:=i-1;if j<q then go to L0;
L4:     j:=j+1;if j>q then go to exit;fy:=FY;yf:=yf+fy;d:=d+fy↑3;go to L4;
exit:   f:=xf+yf;y:=xf×yf;x:=(f↑3-d)×y/f/(f-1)/3;
        W[1]:=w;W[2]:=y;W[3]:=sqrt(x);W[4]:=d;W[5]:=xf;W[6]:=yf
end  WILCOX;
```

```
comment         SP 131, see descr 28;
real procedure tDistr(m,t);value m,t;integer m;real t;
begin   integer i;
        real x,s,u,v;

        x:=arctan(abs(t)/sqrt(m));v:=cos(x)↑2;
        if      m:2×2=m
        then    begin   s:=u:=sin(x)/2; i:=0          end
        else    begin
                        u:=4×arctan(1);s:=x/u;u:=sin(2×x)/u/2;
                        if m>3 then s:=s+u;i:=1
                end;
        m:=m-2;
L:      if i>m then go to out;i:=i+2;u:=u×v×(i-1)/i;s:=s+u;go to L;
out:    tDistr:= if t<0 then .5-s else .5+s
end tDistr;
```

```
comment        SP 132, see descr 22;
procedure COMBINA2(T,g,t,F,f,h,n,first,READY);
value t,n;integer T,g,t,F,f,h,n;Boolean first;label READY;
begin    integer i,j,k;

         g:=t;i:=j:=0;if first then begin   first:=false;g:=0;h:=1;go to next   end;
NEXT:    g:=g-i;if g=0 then go to READY;h:=T;i:=f;if h=n then go to NEXT;
         h:=h+1;if f=F then go to NEXT;h:=h-1;f:=f-1;h:=h+1;j:=g:=g-1;
next:    i:=F;j:=j+i;if j>t then i:=i-j+t;f:=i;k:=0;
rrrr:    k:=k+1;g:=g+1;T:=h;if k<i then go to rrrr;h:=h+1;if j<t then go to next;
         for h:= h step 1 until n do f:=0
end      COMBINA2;
```

```
comment        SP 137, see descr 32;
procedure TIECORR(A,F,i,p,r,T);value p,r;integer F,i,p,r;real A;real array T;
begin    integer g,h,k;
         real a,b,d,e,f;
         integer array K[p:r];

         set1K(A,K[i],i,p,r);sort1(A,K[i],i,p,r);
         i:=K[p];a:=A;f:=h:=F;g:=p;d:=h↑2;e:=h↑3;
L1:      g:=g+1;i:=K[g];b:=A;k:=F;f:=f+k;if a=b ∧ g<r then go to L2;
         if a=b then h:=h+k else g:=g-1;a:=b;d:=d+h↑2;e:=e+h↑3;g:=g+1;h:=0;
L2:      h:=h+k;if g<r then go to L1;
         T[1]:=f;T[2]:=d;T[3]:=e
end TIECORR;
```

```
comment        SP 138, see descr 32;
procedure RANK2(A,B,F,i,p,r,T);value p,r;integer F,i,p,r;real A,B;real array T;
begin    integer g,h,m,n,k;
         real a,b,d,e,f;
         integer array K[p:r];

         set1K(A,K[i],i,p,r);sort1(A,K[i],i,p,r);
         i:=K[p];a:=A;f:=h:=F;m:=g:=p;n:=0;d:=h↑2;e:=h↑3;
L1:      g:=g+1;i:=K[g];b:=A;k:=F;f:=f+k;if a=b ∧ g<r then go to L2;
         if a=b then h:=h+k else g:=g-1;a:=b;b:=n+(h+1)/2;
         d:=d+h↑2;e:=e+h↑3;n:=n+h;
         for h:=m step 1 until g do begin  i:=K[h];B:=b  end;
         m:=g:=g+1;h:=0;
L2:      h:=h+k;if g<r then go to L1;
         T[1]:=f;T[2]:=d;T[3]:=e
end RANK2;
```

```
comment        SP 139, see descr 23;
real procedure PHI(x);value x;real x;
begin   real z;

        z:=abs(x)/sqrt(2);
        z:=(1+z×(.0705230784+z×(.0422820123+z×(.0092705272+
             z×(.0001520143+z×(.0002765672+z× .0000430638))))))↑(-16);

        PHI:=if x>0 then 1-z/2 else z/2

end PHI;
```

SP 139

```
comment          SP 140, see descr 32;
procedure TIECORR(A,F,i,p,r,T);value p,r;integer F,i,p,r;real A;real array T;
begin     integer g,h,k;
          real a,b,d,e,f;
          integer array K[p:r];

          set1K(A,K[i],i,p,r);sort1(A,K[i],i,p,r);
          i:=K[p];a:=A;g:=p;h:=F;d:=e:=f:=0;
L1:       if g=r then go to L3;g:=g+1;i:=K[g];b:=A;
          if b≠a then go to L2;h:=h+F;go to L1;
L2:       g:=g-1;
L3:       a:=b;f:=f+h;d:=d+h↑2;e:=e+h↑3;
          if g<r then begin  g:=g+1;i:=K[g];h:=F;go to L1  end;
          T[1]:=f;T[2]:=d;T[3]:=e
end TIECORR;


comment          SP 141, see descr 32;
procedure RANK2(A,B,F,i,p,r,T);value p,r;integer F,i,p,r;real A,B;real array T;
begin     integer g,h,m,n,k;
          real a,b,d,e,f;
          integer array K[p:r];

          set1K(A,K[i],i,p,r);sort1(A,K[i],i,p,r);
          i:=K[p];a:=A;m:=g:=p;h:=F;n:=0;d:=e:=f:=0;
L1:       if g=r then go to L3;g:=g+1;i:=K[g];b:=A;if b≠a then go to L2;h:=h+F;go to L1;
L2:       g:=g-1;
L3:       a:=b;b:=n+(h+1)/2;f:=f+h;d:=d+h↑2;e:=e+h↑3;
          for k:=m step 1 until g do begin  i:=K[k];B:=b  end;
          if g<r then begin  n:=n+h;m:=g:=g+1;i:=K[g];h:=F;go to L1  end;
          T[1]:=f;T[2]:=d;T[3]:=e
end RANK2;
```

```
comment         SP 142, see descr 18;
procedure set1K(A,K,i,p,r);value p,r;integer K,i,p,r;real A;
begin    integer h;
         real a,b,c;

         a:=b:=0;
         for i:=p step 1 until r do begin  c:=A;a:=a+c;b:=b+i×c   end;
         h:=p+r;
         if      b>a×h/2
         then    for i:=p step 1 until r do K:=i
         else    for i:=p step 1 until r do K:=h-i

end set1K;
```

```
comment          SP 143, see descr 27;
real procedure Cgamma2(n);value n;integer n;
begin    real g;

         g:=1.77245;if n≥7 then go to L;if n:2×2=n then g:=1;
K:       n:=n-2;if n<1 then go to M;g:=g×n/2;go to K;
L:       g:=g×exp(-n/2)×n↑(n/2-.5)×(1+(.166667+.0138889/n)/n)/2↑(n/2-1);
M:       Cgamma2:=g
end  Cgamma2;


comment          SP 144, see descr 27;
real procedure Qgamma2(n,x);value n,x;integer n;real x;
begin    integer k;
         real a,s,t,u;

         s:=0;if x≤0 then go to out;a:=n/2;if n≥100 then go to WH;
         if n≤6 then go to LE;if x≤a then go to I else go to II;
LE:      if n:2×2≠n then go to L1;t:=exp(-x);s:=1-t;if n=2 then go to out;
         t:=t×x;s:=s-t;if n=4 then go to out;s:=s-t×x/2;go to out;
L1:      t:=sqrt(x);s:=1-2×PHI(-t×1.4142135624);if n=1 then go to out;
         t:=exp(-x)×t/.8862269255;s:=s-t;if n=3 then go to out;
         s:=s-t×x×2/3;go to out;
I :      s:=t:=exp(-x)×x↑a/a/Cgamma2(n);u:=x/(a+1);k:=1;
L2:      if u≤.001 then go to out;t:=t×x/(a+k);s:=s+t;
         k:=k+1;u:=(x/(a+k))↑k;go to L2;
II:      t:=exp(-x)×x↑(a-1)/Cgamma2(n);s:=1-t;u:=(a-1)/x;k:=1;
L3:      if u≤.001 then go to out;t:=t×(a-k)/x;s:=s-t;
         k:=k+1;u:=u×(a-k)/x;go to L3;
WH:      s:=PHI(((x/a)↑(1/3)+1/a/9-1)×sqrt(9×a));
out:     Qgamma2:=s
end Qgamma2;
```