The Student's SEMSTAT

Author: Jack Prins

Table of Contents

Chapter 1

The Student's SEMSTAT

Introduction
The SEMSTAT is a collection of statistical routines intended for the use of faculty and students at undergraduate and graduate curriculums. The topics covered are:

Descriptive Statistics and Plotting including
- Line Plots and Histograms
- Box Plots
- Stem and Leaf Plots
- Measures of central tendency and dispersions
- Random Number Generators

Inferential Statistics including
- Numerical Integration of Probability Distributions and their Inverses
- Hypothesis Testing and Confidence Intervals
- Analysis of Variance and Contrast Testing
- Simple and Multiple Linear Regression

Linear Programming and Applications
- Simplex method
- Transportation Problems
- Assignment Problems
- Two Person Zero Sum Games

Statistical Quality Control charts including
- Shewhart $\bar{x}$ and S control charts
- Shewhart $\bar{x}$ and R control charts
- Shewhart $\bar{x}$ control charts
- Individual X and Moving Range control charts
- Run sum and Moving Average Control Charts
- Cumulative Control Charts for means
- Horizontal Cusum Charts
- Exponential Weighted Moving Average control charts
- Multivariate Control Charts (MEWMA, HOTELLING, PCA)
- Average Run Length calculations
- Double and Single Sampling Plans
- Sequential Sampling and Skip lot Sampling

Time Series Analysis, including

Box-Jenkins identification routines
Box-Jenkins estimation/diagnostics/forecasting
Stepwise Autoregression (Box-Jenkins AR model)
Multivariate Autoregressive Models
Exponential Smoothing by Holt-Winters (single, double and triple)
Single and multiple Linear Regression

Nonparametric Statistics, including

Two-sample Sign Test for Ordinal Levels
Wilcoxon matched-pairs signed-ranks test
Mann-Whitney U test
Two-way Friedman's ANOVA for related samples
Kruskal-Wallis 1-way ANOVA for independent data
Spearman Rank correlation coefficient
Kendall Rank correlation coefficient
Kendall partial rank correlation coefficient
Kendall coefficient of concordance
Kolmogorov-Smirnov, goodness of fit test
Anderson-Darling and Cramer Von Mises goodness of fit tests
Fisher's Exact Test of Proportions
McNemar Test for significance of changes

Multivariate Statistics, including the following techniques

Multiple Analysis of Variance (MANOVA)
Multiple Analysis of Covariance (MANCOVA)
Classification Analysis
Principal Components followed by Rotation
Eigen System Solutions

Reliability , including the following routines

Parameter estimation for censored and uncensored samples for the Exponential and the two parameter Weibull , Lognormal and Uncensored Gamma Distributions. Inverse Distribution.
Duane Plots and other analyses for the Exponential Distribution
Hazard Rates, Confidence Intervals, OC Curves , more …

Design of Experiments, including the following design families

Full 2^n factorials.
Fractional 2^{n-p} factorials.
Full 3^n factorials.
Plackett-Burman Designs.
CCD Designs

Utility Programs, including:

- Factorials of integers and half integers
- Combinations
- Matrix Inversion
- Determinants
- Chi-square Contingency Tables
- Goodness of Fit Test
- Box-Cox Transformation
- Solution for Linear Systems
- Gramm-Schmidt Orthogonalization
- Generalized Trapezoidal Rule

The system is optionally menu driven. To start it one has to be in the directory that contains the SEMSTAT system, and then type **GO**. The main menu will appear.
You can use the mouse or the arrow keys. The Shift-F1 toggles between the methods.
If you select the arrow keys method, the following message appears:

To select an item on any menu, use the arrow keys to move the cursor to the position of the line with the desired item and press the enter key.

Main Menu

SEMSTAT, a statistical library for Students
Descriptive Statistics Probability Distributions Hypothesis Testing and Confidence Intervals Analysis of Variance and Contrast Testing Linear Regression Statistical and Mathematical Utilities Linear Programming Statistical Quality Control Time Series Analysis Non Parametric Statistics Multivariate Statistics Reliability Design of Experiments Exit

The SEMSTAT package is distributed on one CD.

The installation procedure wants you to insert the CD in the appropriate drive. Then click START, RUN from the desk top, and select D: \setup, where D is the drive letter of the CD-ROM. Then follow the prompts.

The files are compressed so that you cannot copy them directly onto your hard disk
The SETUP routine expands and decompresses.

Storage requirement on the hard-disk.

The approximate storage requirements on your hard disk are:

The Statistics Basic Package (STATLIB)	1250 K bytes
The Statistical Quality Control (SQC)	1100 K bytes
The Non Parametric component (NONPARAM)	367 K bytes
The Time Series component (TIMESTAT)	540 K bytes
The Linear programming component (LP)	85 K bytes
The Multivariate component (MVAR)	525 K bytes
The Reliability component (RELIABLT)	496 K bytes
The Design of Experiments component (DOE)	650 K bytes
Total	5013 K bytes

Part I: The Statistical Components

Part 1 of this manual deals with the statistical portion of the software. Let us begin with following its installation.

Installation

The installation procedure is straight forward.
To install the SEMSTAT system insert the CD-ROM drive and click
START, RUN and then type D:**SETUP (D is here the CD ROM drive)** and answer the prompts.

The setup procedure will first create one main directory and six sub-directories. This is done to make the overall book-keeping easier. The default name for the main directory is: **C:\STATS** You may choose any name or path you wish. In fact that is the intention of the very first prompt that is issued. It then creates another main directory and 4 sub-directories for the D.O.E. component. The default name for this main directory is **C:\DOESTUD** (for student)

```
*******************************************************
*   This installs SEMSTAT on the hard disk and creates six      *
*   sub directories containing:                                 *
*   Linear Programming, Quality Control,                        *
*   Timeseries Analysis, Nonparametric Statistics,              *
*   Mulitvariate Statistics and Reliability.                    *
*******************************************************

The default directory to store the SEMSTAT is          C:\STATS
The default directory to store the LP programs is      C:\STATS\SIMPLEX
The default directory to store Quality Control is      C:\STATS\SQC
The default directory to store Time Series Analysis    C:\STATS\TIMESTAT
The default directory to store Nonparametric Stats     C:\STATS\NONPARAM
The default directory to store Multivariate Stats      C:\STATS\MVAR
The default directory to store Reliability is          C:\STATS\RELIABLT

Default is accomplished by pressing Enter as answer to a prompt.

 Name the directory to store the main program  or press Enter

OK?  y/n: y
```

CREATING DIRECTORIES AND COPYING FILES

Now the D.O.E. component will be installed...

The default directory to store the D.O.E. system is C:\DOESTUD.
The default directory to store data files is C:\DOESTUD\DATA.
The default directory to store output reports is C:\DOESTUD\REPORTS.
The default directory to store generated designs is C:\DOESTUD\DESIGNS.
Default is accomplished by pressing Enter as answer to a prompt.

Name the directory to store the main program or press Enter:
OK? y/n: y

CREATING DIRECTORIES AND COPYING FILES ...

If all goes well, the following remark will be displayed:

The D.O.E. system is successfully installed.

Press the Enter key to return.:

The SEMSTAT system is successfully installed
To start the system, be sure to be in the directory
where the system resides and then type GO.

Press the Enter key to return to the DOS prompt:

****
* Running the SEMSTAT Programs from Windows 9x/NT *
****

1 Move the cursor to an empty slot on the screen.
2 Click the right mouse button.
3 A little screen pops up, click NEW and then click SHORTCUT.
4 A new screen appears. Fill in the blank line with:
5 C:/STATS/GO.BAT (C:\STATS or whatever you selected
as main directory at installation).
6 Click NEXT, another little screen overwrites the present one.
7 Fill in the blank line with a name for the Shortcut icon.
8 Click on FINISH

You now have created a shortcut icon. Clicking it will start SEMSTAT.
To ensure that you operate in the full-screen mode, instead of a window
proceed as follows:
1 Click the shortcut item with the right mouse button.
2 Click Properties.
3 Click Options.
4 Click Full Screen.
5 Click OK.

****
* Running the SEMSTAT Programs from Windows 3.x *
****

1 Start Windows
2 Click on the Applications Icon. The Application window is displayed
3 Click on File, then New
4 Select Program Item, then OK. This displays :Program Items ...
5 In the Description Box put SEMSTAT (or whatever title you want)
6 In the command line put GO.BAT
7 In the working Directory box put C:\ STATS (or wherever you installed)
A new icon SEMSTAT is now added to the Applications group
You can change the icon by clicking the Change icon button.
8 click OK, then close.

You can now execute the SEMSTAT programs from Windows.

The rest of this part of the user's guide involves computer output on sample terminal sessions. This enables you to understand the syntax of the routines. The guide does not pretend to be a text book. There are many excellent texts that should function as the background material. A list of a few will be found in the section on References.

Data Input

How does one input the data?

There are a few ways to prepare input files in ASCII format. This format is required by SEMSTAT.

1. Use the program INPUT.
   This prompts for the number of rows and columns.
   Just follow the prompts for input and editing (if necessary)

2. Use any editor or spreadsheet package and create a raw date file. There is one catch: the raw data file MUST be an ASCII file. So if you use Lotus 1-2-3 as the vehicle, be sure to output the file in the .PRN form. (use /PRINT, then RANGE, then GO) or if you work in Windows, follow the instructions.

3. Use any word processor such as WordPerfect or Microsoft Word and use the asscociated Text Output option.

4. If Host Systems are involved, use the appropriate download routines.

```
INPUT (this is the routine  in SEMSTAT)

 This program forms  data files
 How many rows (enter 0 to exit): ? 6
 How many columns: ? 3

 Start inputting. Press the enter key after each entry...
    1  :1    2    3
    2  :4    5    6
    3  :7    8    9
    4  :9    8    7
    5  :6    5    4
    6  :3    2    1

 Wish to see the input file (for eventual corrections)? y/N:  n

 Enter a file name or press the ENTER key (<-┘ ) to name it DATA.FIL: my.fil
 The data are stored in file my.fil
```

Switches in selected programs

1) **Mouse-Arrow-Keys-Switch**

The menus are mouse driven by default. Some users don't like the mouse. One can switch to arrow keys that position a large cursor to the desired line in the menu by pressing the Shift + F1 combination simultaneously. The Shift-F1 key toggles between the two methods. If you press them twice in a row, you are back to the mouse.

2) **Mono-Color-Switch**

The menus and program prompts appear in color by default. If you prefer black and white (or don't have a color monitor), you can switch to monochrome by pressing the Shift + F2 keys simultaneously. In general, the rest of the session, inclusive of the plots, will appear in black and white or their equivalents. as is the case for the Shift + F1 keys, the mono-color switch toggles between the two display colors. If you press while a prompt expects some answer. you must press the enter key in order to have the switch operational.

3) **End-of-Program-Switch**

If you wish to abort you current session, press the F4 key at any prompt, and then press the Enter key. This will, in general, return you to the menu. of origin. (However, there are some cases when this will not work)

4} **Data-File-Edit-Switch**

In the Statistical Quality Control component, you can edit your input files "on the fly" by pressing the F6 key. This will prompt for information in the form of line editing. The input lines are displayed with numerical prefixes. This indicate the order of the line in the file. You edit desired by prefix, line(s)

Chapter 2

Descriptive Statistics

The menu for descriptive statistics is shown below:

Use mouse or up and down arrow keys to position the cursor, then press Enter

Descriptive Statistics and Plotting
Mean, Median, Mode, Variance, Range, Skewness, Kurtosis, etc. General Plotting Routine and Histograms Box-Plots Stem and Leaf Plots File Generation Routine Prints Saved Plots Random Number Generators Exit.

Example of the 1st line on the menu

DESCRIBE

DESCRIPTIVE STATISTICS

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you press the enter key (↵), ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): To return, press the F10 key. ? demo1.fil

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry)
c) or press the enter key (↵) for all data. ?

DESCRIPTIVE STATISTICS FOR FILE :demo1.fil

MAX	MIN	MEAN	VARIANCE	NO.DATA
80.0000	23.0000	51.1286	141.8238	70

STD.DEVIATION : 11.9090
RANGE : 57.0000

THE MEDIAN IS: 51.5
THE MODE IS: 38 Number of items is: 3
THE MODE IS: 64 Number of items is: 3
THE MODE IS: 71 Number of items is: 3
Multi-Modal ...

SKEWNESS: -0.0722 Z VALUE : -0.2468
KURTOSIS: -0.0607 Z VALUE : -0.1037

QUADRATIC MEAN : 52.4779
GEOMETRIC MEAN : 49.6306
HARMONIC MEAN : 47.9437

Example of Line Plot

Plot

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵), ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): To exit, press F10.
? demo1.fil

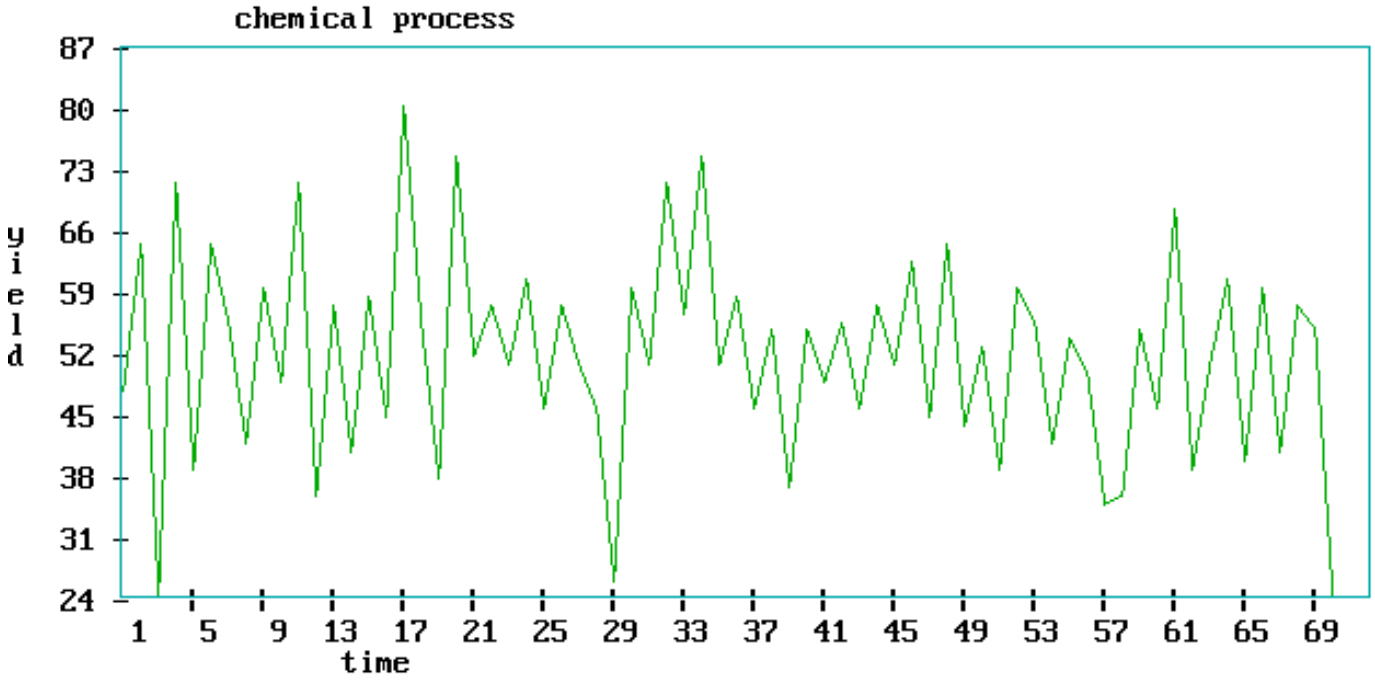
Enter 0 (or press the Enter key) to plot against the INDICES,
or 1 to plot against a selected data column,
or h for help...

Want to see the first 10 lines of the file ? y/N:

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number
e.g. 12-46 (the hyphen is a MUST) analyzes rows 12 to 46.
b) or just the first sequence number
e.g. 12 analyzes from row 12 till the last row.
c) or press the enter key (↵) for all data rows. ?

MAX	MIN	MEAN	VARIANCE	column
80.0000	23.0000	51.1286	141.8239	1

NUMBER OF OBSERVATIONS: 70


chemical process
87
80
73
66
59
52
45
38
31
24
yield
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
time

Use up and down arrow keys to position the cursor, then press Enter

OPTIONS MENU

GENERATE DATES
TRANSFORMATIONS
SEASONAL ADJUSTMENT
DIFFERENCING
STORE PLOT ON DISK
LEGENDS
GRIDS
TITLES
HISTOGRAM

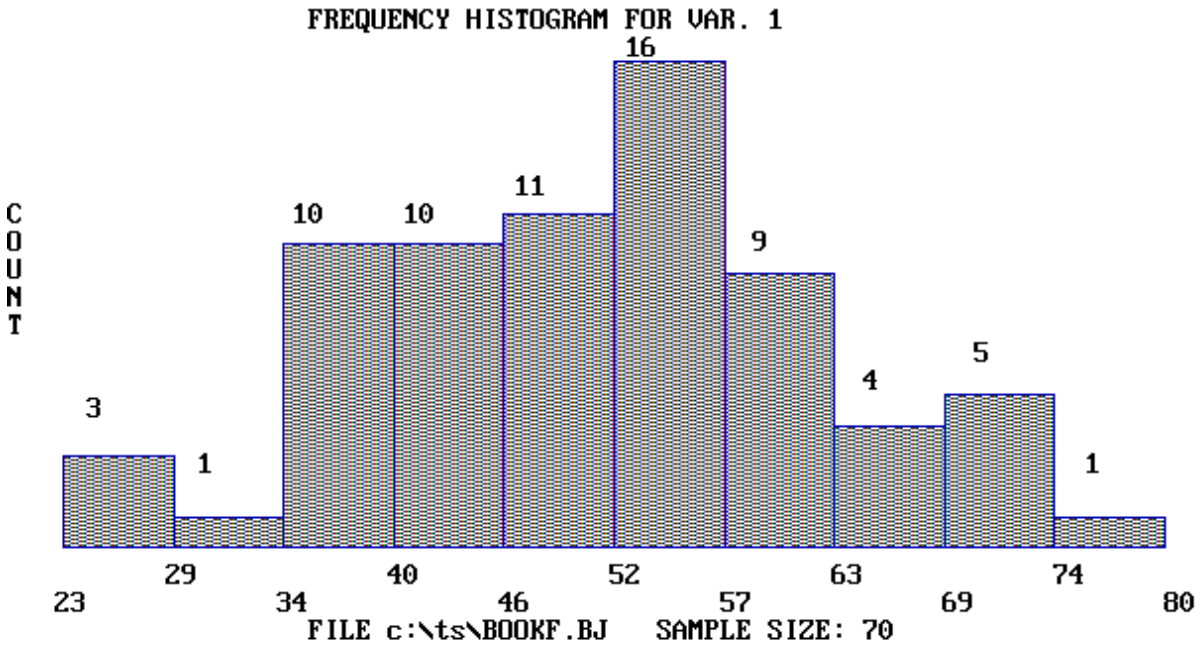
PLOT

The HISTOGRAM was selected

Want a frequency histogram? Y/n: y
Enter 1 to construct it from a given number of cells,
Enter 2 to construct it from a given width per cell: ? 1
Enter 1 for absolute or 2 for relative frequency: ? 1

HISTOGRAM FOR VARIABLE : 1

How many cells? Press enter for 18: 10



BOXPLOT

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵) , ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10. demo1.fil
You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry),
c) or press the enter key (↵) for all data. ?

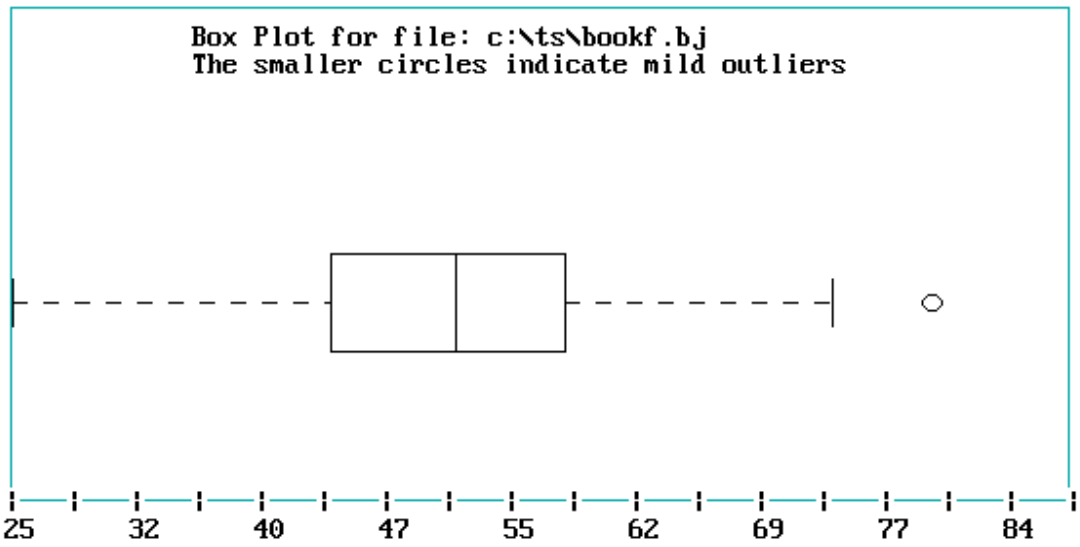
Statistics for file demo1.fil

MAX	MIN	MEAN	VARIANCE	NO.DATA
80.0000	23.0000	51.1286	141.8238	70

Median Location 35.5	Hinge Location 18		
Median	51.500		
Hinges	44.000	58.000	H-spread: 14.000
Adjacent Values	25.000	74.000	
Inner Fences	23.000	79.000	
Outer Fences	2.000	100.000	

Mild Outliers
80

Extreme Outliers



STEMLEAF

```
Enter FILESPEC: To return, press the enter key.  demo1.fil

For Series demo1.fil

     MAX           MIN           MEAN          STD.DEV.      NO.DATA
    80.0000       23.0000       51.1286       11.9090         70

                          Stem and Leaf Plot for Series demo1.fil
                            SCALE FACTOR FOR STEM =  10
2   |3 3 5
3   |4 5 5 6 7 8 8 8 9
4   |0 0 1 1 3 4 4 5 5 5 5 5 7 8 8 9
5   |0 0 0 0 0 0 1 2 3 4 4 4 4 5 5 5 5 6 7 7 7 7 7 8 8 9 9 9 9
6   |0 0 2 4 4 4 8
7   |1 1 1 4 4
8   |0
```

Probability Distributions

Use the mouse or the up and down arrow keys to position the cursor, then press Enter

SAMPLING DISTRIBUTIONS		
1 Normal	Input z	output prob(z)
2 Student's t	Input df and t	output prob(t)
3 Chi-square	Input df and chi-sq	output pr(chi-sq)
4 Snedecor's f	Input df's and f	output prob(f)
5 Inverse Normal	Input prob(z)	output z
6 Inverse t	Input prob(t) and df	output t
7 Inverse chisq	Input prob(chi-sq) and df	output chi-sq
8 Inverse f	Input prob(f) and df's	output f
9 Binomial	Input n and p	output pdf, cdf
10 Poisson	Input x and np	output pdf, cdf
11 Hypergeometric	Input N, D, n, x	output pdf, cdf
12 Non Central t	Same for t, chi-square and	output prob(t)
13 Non Central c	f, plus a non-centrality	output pr(chi-sq)
14 Non Central f	parameter.	output prob(f)
Exit.		

Enter the normal z value: *1.65* selected 1

Probability = 0.9505

More? y/n: *n* back to the menu
Enter the Student's t value: *1.65* selected 2
Enter the degrees of freedom: *5*
Probability = 0.9201

More? y/n: *n* back to the menu
Enter the Chi-square value: *12* selected 3
Enter the degrees of freedom: *14*
Probability = 0.3930

More? y/n: *n* back to the menu
Enter the Snedecor's f value: *4.8* selected 4
Enter the degrees of freedom of numerator : *8*
Enter the degrees of freedom of denominator: *12*

Probability = 0.9923

More? y/n: n back to the menu
Enter the normal probability: *.8* selected 5
z = 0.8415

More? y/n: y
Enter the normal probability: *.997*
z = 2.7482

More? y/n: *n* back to the menu
Enter the Student's t probability *.789* selected 6
Enter the degrees of freedom: *12*
t = 0.8312

More? y/n: *n* back to the menu
Enter the Chi-square probability: *.9* selected 7
Enter the degrees of freedom: *15*
chi-square = 22.2930

More? y/n: *n* back to the menu
Enter the Snedecor's f probability: *.9* selected 8
Enter the degrees of freedom of numerator : *8*
Enter the degrees of freedom of denominator: *16*
f = 2.0877

More? y/n: *n* (selected 0)
Binomial
Individual and cumulative terms of the Binomial Distribution

Enter N, the number of trials or press Enter to quit: 12
Enter P, the probability in a single success: .16

X	PDF	CDF
0	0.123410	0.123410
1	0.282081	0.405491
2	0.295500	0.701004
3	0.187627	0.888632
4	0.080412	0.969043
5	0.024506	0.993550
6	0.005446	0.998996
7	0.000889	0.999885
8	0.000106	0.999991
9	0.000009	0.999999

HYPOTHESIS TESTING

The menu for hypothesis testing is shown below:

```
HYPOTHESIS TESTING AND CONFIDENCE BOUNDS

One population mean, large sample size.  ( n > 30)
Difference between 2 populations means, large sample size
One population mean, small sample size.
Difference between 2 population means, small sample size.
One proportion (binomial parameter). (n > 30)
Difference between 2 proportions (binomial parameters)
One population variance
Ratio of two population variances
Homogeneity of variances
Exit.
```

Example 1

```
One population mean, large sample size.  ( n > 30)
   Hypothesis testing. H0: μ1 = μ0.  Variance is known.
   Z test for one population.

  Enter the population mean: 120
  Enter the sample mean:     117
  Enter the sample std.dev.: 12
  Enter the sample size:     56

  Enter value for α (press Enter for .05):
  Enter 1 or 2  for One or Two sided test: 1

    The Z-test statistic   =  1.870829      The critical value =  1.645211
    Probability of Z-test  =  0.969316      Right Tail Area   =  0.030684
    Reject the null hypothesis.

    A two-sided confidence about the population mean is given by:
    Prob {   113.8564  <   U0 <     120.1436 } =      0.9500
```

Example 2

Hypothesis testing. H0:U1 = U2. Variance is known.
Z test for two populations.

Enter sample mean 1: 6.70
Enter sample std.dev.1: .60
Enter sample size 1: 100

Enter sample mean 2: 6.54
Enter sample std.dev.2: .63
Enter sample size 2: 100

Enter value for α (press Enter for .05): .10
Enter 1 or 2 for One or Two sided test: 2

The Z-test statistic = 1.839080 The critical value = 1.645211
Probability of Z-test = 0.967048 Right Tail Area = 0.032952
Reject the null hypothesis.

A two-sided confidence about the population mean is given by:
Prob { 0.0169 < DELTA < 0.3031 } = 0.9000

DELTA is the absolute difference between μ1 and μ2.

Example 3

Hypothesis testing. H0:U1 = U0. Variance is unknown.
T Test for one population.

Enter the population mean: 120
Enter the sample mean: 117
Enter the sample std.dev.: 12
Enter the sample size: 15

Enter value for α (press Enter for .05):
Enter 1 or 2 for One or Two sided test: 1

The t-test statistic = 0.968246 The critical value = 1.761732
Probability of t-test = 0.825324 Right Tail Area = 0.174676
Accept the null hypothesis.

A two-sided confidence about the population mean is given by:
Prob { 110.3529 < U0 < 123.6471 } = 0.9500

Example 4

Difference between 2 population means, small sample size.

Hypothesis testing. H0:U1 = U2. Variance is unknown.

T test for two populations.

Enter sample mean 1: 495
Enter sample std.dev.1: 55
Enter sample size 1: 15

Enter sample mean 2: 545
Enter sample std.dev.2: 50
Enter sample size 2: 15

Enter value for α (press Enter for .05):
Enter 1 or 2 for One or Two sided test: 2

The t-test statistic = 2.605251 The critical value = 2.048890
Probability of t-test = 0.992732 Right Tail Area = 0.007268
Reject the null hypothesis.

A two-sided confidence about the population mean is given by:
Prob { 10.6777 < DELTA < 89.3223 } = 0.9500

DELTA is the absolute difference between μ1 and μ2.

Example 5

```
 Hypothesis testing. H0:p-hat = p0.  n > 30.
 Z test for one proportion.

 Enter n, the number of trials: 200
 Enter the number of successes: 26
 Enter the hypothesized P     : .10

 Enter value for α (press Enter for .05):
 Enter 1 or 2  for One or Two sided test: 1

The Z-test statistic =  1.414 The critical value =  1.645
Probability of Z-test = 0.921 Right Tail Area    =  0.078
Accept the null hypothesis.

   A two-sided confidence for P is given by:
Prob {  0.0951  <   P <    0.1649 } =   0.9500
```

Example 6

```
Hypothesis testing. H0:p1 - p2 = 0. n1 and n2 > 30.
Z test for absolute difference between two proportions.

      FOR SAMPLE 1
 Enter n, the number of trials: 1000
 Enter the number of successes: 52

      FOR  SAMPLE 2
 Enter n, the number of trials: 1000
 Enter the number of successes: 23

 Enter value for α (press Enter for .05):
 Enter 1 or 2  for One or Two sided test: 1

The Z-test statistic  =  3.41    The critical value =  1.645
Probability of Z-test  =  0.9997 Right Tail Area =  0.000321
Reject the null hypothesis.

   A two-sided confidence for P1 - P2  is given by:
Prob {  0.0150  <   DELTA  <    0.0430 } =   0.9500
```

Example 7

```
Hypothesis testing. H0:VAR1 = VAR0.
Chi-square test.

Enter the population variance: 100
Enter the sample variance: 195
Enter the sample size:     10

Enter value for α (press Enter for .05):
Enter 1 or 2  for One or Two sided test: 1

The Chi.sq-test statistic   = 17.550
The critical value = 16.9047
Probability of Chi.sq-test  =  0.9594
Right Tail Area    =  0.040590
Reject the null hypothesis.

A two-sided confidence about the population variance is given by:

Prob {103.81 < Population Variance < 529.49} = 0.90
```

Example 8

```
Hypothesis testing. H0:VARIANCE 1 = VARIANCE 2.
F Test.

Enter sample variance 1: 1.04
Enter sample size 1:     25

Enter sample variance 2: .51
Enter sample size 2:     25

Enter value for α (press Enter for .05):
The following F test is one sided; Ha:var 1 > var 2

The F-test statistic =  2.039  The critical value =  1.984
Probability of F-test=  0.956  Right Tail Area    =  0.043
Reject the null hypothesis.    Variance 1 > Variance 2.

A two-sided confidence about the F ratio is given by:
Prob { 0.8983  < F RATIO <   4.6275 } =  0.9500
```

Homogeneity of Variances Tests

When a usual one way Analysis of Variance is performed, it is assumed that the group variances are statistically equal. If this assumption is not valid, then the resulting F-test in Anova is invalid. There is a technique , called the Welch One Way ANOVA, that allows for unequal variances. This section describes four tests for equality of group variances and the Welch ANOVA. The tests for homogeneity of variances are:

- O'Brien's Test
- Brown-Forsythe Test
- Levene's Test
- Bartlett's Test

The first three tests are based on the creation of or transformation to a new response variable and then to perform an analysis of variance on this new variable. These new response variables are constructed to measure the spread in each group.
The fourth test (Bartlett's test) is derived from the likelihood ratio test under the normal distribution.

- O'Brien's test computes a new response so that its cell or group means are equal to the variances of the original response. The new response is computed as:

$$z_{ij} = \frac{(n_j + w - 2)n_j \left(y_{ij} - \bar{y}_{j.}\right)^2 - w s_j^2 \left(n_j - 1\right)}{\left(n_j - 1\right)\left(n_j - 2\right)}$$

where
y_{ij} is the original observation in row i and group j
$\bar{y}_j$ is the mean of group j
s_j^2 is the variance of group j
n_j is the number of responses in group j
w is set to .5

- Brown-Forsythe is the model F statistic from an one way Anova on

$$z_{ij} = \left|y_{ij} - \tilde{y}_j\right| \quad \text{where } \tilde{y}_j \text{ is the median of group } j$$

If any z_{ij} is zero then it is replaced by the next smallest z_{ij} in group j. This is introduced to correct for the artificial zeros that come about with odd numbers of observations in a group.

- The Levene F is the model F statistic in the one way Anova from

$$z_{ij} = \left|y_{ij} - \bar{y}_j\right| \quad \text{where } \bar{y}_j \text{ is the mean of group } j$$

The model F or test statistic for the above three tests is the ratio of $MS_{\text{treatment}}/MS_{\text{error}}$ which, after some algebra, is

$$F = \frac{(N-p)\sum_{j=1}^{p} n_j \left(\bar{z}_{j.} - \bar{\bar{z}}_{..}\right)^2}{(p-1)\sum_{j=1}^{p}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}_{j.}\right)^2}$$

where
N = total number of observations, n_j = number of observations per group and
p = number of groups

The degrees of freedom for the first three tests are given by:
DF numerator = p - 1
DF denominator = N - p, where N is the total number of responses

- Bartlett's test statistic is computed as

$$T = \frac{v \ln\left(\sum_{j=1}^{p} \frac{v_j}{v} s_j^2\right) - \sum_{j=1}^{p} v_j \ln s_j^2}{1 + \left[\frac{\left(\sum_{j=1}^{p} \frac{1}{v_j}\right) - \frac{1}{v}}{3(p-1)}\right]}$$

where

$v_j = n_j$, $n_j =$ the number of responses in group j

$v = \sum_{j=1}^{p} v_j$, $p =$ the number of groups

The Bartlett test statistic follows a Chi Square distribution with p-1 degrees of freedom.
Division of the test statistic by the degrees of freedom yields an F value.
A modification of Bartlett's test statistic that leads to a different F value is

$$F = \frac{v_2 M}{v_1 (b - M)}$$

where

$$M = (N - p) \ln s^2_{pooled} - \sum_{j=1}^{p} (n_j - 1) \ln (s_j^2)$$

$$A = \frac{1}{3(p-1)} \left[\sum_{j=1}^{p} \left(\frac{1}{n_j - 1} \right) - \left(\frac{1}{N - p} \right) \right]$$

$$s^2_{pooled} = \frac{\sum_{j=1}^{p} (n_j - 1) s_j^2}{N - p}$$

$$v_1 = p - 1 \qquad v_2 = \frac{p+1}{A^2}$$

$$b = \frac{v_2}{1 - A + 2/v_2}$$

This F has a sampling distribution better approximated by the $F(v_1, v_2)$ distribution. The values of v_2 are not necessary integers, so one may have to interpolate in the F table.

The reported p values are the probability of exceeding the reported F values. That is, the probability of obtaining by chance alone an F value larger then the computed F, if in fact the variances are equal for all groups.

The Welch approach (an Anova allowing for unequal variances) computes the following F statistic

$$F = \frac{\dfrac{\sum_{j=1}^{p} w_j \left(\bar{y}_j - \tilde{y}_{..} \right)}{p-1}}{1 + \dfrac{2(p-2)}{k^2-1} \left[\sum_{j=1}^{p} \dfrac{\left(1 - \dfrac{w_j}{u} \right)^2}{n_j - 1} \right]}$$

where

$$w_j = \frac{n_j}{s_j^2} \qquad u = \sum_{j=1}^{p} w_j \qquad \tilde{y}_{..} = \sum_{j=1}^{p} \frac{w_j \bar{y}_{j.}}{u}$$

An example follows:

```
              Test for Homogeneity of Variances

  Number of Observations  =  40
  Number of Groups        =  6

  Group Statistics:
  Group      Median     Mean         Std.Dev      N
   1        89.5000   99.0000        29.3355      8
   2        98.0000   94.7143        16.2349      7
   3        96.0000  100.8333        17.7551     12
   4       106.0000  108.2857        10.8737      7
   5       115.0000  118.3333         8.5049      3
   6       134.0000  140.6667        28.5890      3

 The p-value for rejecting = .05
 Test                       F-Ratio   p-value
Levene                       3.0079    0.0236     Reject H0
Brown-Forsythe               2.1035    0.0890     Accept H0
O'Brien                      2.0498    0.0963     Accept H0
Bartlett  Chi2 = 8.1337      1.6267    0.1490     Accept H0

Bartlett's test is chi-square with df =  5
All F tests are done with df =  5 and  34


Welch 1 way ANOVA            2.4737    0.1120     Accept H0

Here, the variances may be unequal. The df are 5, 10.112
```

The input file is:

Grp 1	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6
95	112	81	92	112	116
123	107	91	112	115	134
74	67	142	128	128	172
145	98	84	111	.	.
64	105	85	105	.	.
84	95	93	104	.	.
128	79	99	106	.	.
79	.	119	.	.	.
.	.	92	.	.	.
.	.	112	.	.	.
.	.	99	.	.	.
.	.	113	.	.	.

Chapter 3

ANALYSIS OF VARIANCE

The menu for the various routines for the Analysis of Variance is displayed below:

Analysis of Variance (ANOVA) and Contrast Testing
One-Way ANOVA Randomized Block ANOVA Two-Way ANOVA Three-Way ANOVA Four-Way ANOVA Five-Way ANOVA Data File Generation Multiple Comparisons Exit.

Examples from the program:

One-Way ANOVA FOR EQUAL SAMPLE SIZES PER GROUP

The input to this program is an multicolumn ASCII file. You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵), ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): To exit, press F10.
? anova1.fil

MAX	MIN	MEAN	VARIANCE	SERIES
4.1000	1.9000	3.0000	0.5720	1
3.9000	2.2000	2.9167	0.3977	2
4.3000	2.3000	3.1000	0.5960	3

NUMBER OF OBSERVATIONS: 6

```
                    **********************************
                    *  ANALYSIS OF VARIANCE TABLE    *
                    **********************************
   Source                    S.S          D.F.          M.S          F-RATIO
Between Groups             0.1011          2          0.0506   0.0969
Within Groups         7.8283          15          0.5219
Total (corrected)     7.9294          17
-------------------------------------------------------------------------------
Mean                      162.6006    1
Total                     170.5300    18

Probability of F =      0.091763  Right Tail Area  =  0.908237

Want contrast testing ?  (y/N): y

Input α or press Enter for default of .05: .1

        Means
 3.000    2.917    3.100

          ************************************************************
          *                      CONTRAST TESTING                    *
          ************************************************************
A normalized contrast is the contrast divided by its standard error.  There is a maximum value
for the normalized contrast.  The contrast coeffcients for the maximum normalized contrast are:

  -0.105      -1.677       1.782

You will now be prompted for contrast coefficents.  If you do NOT wish to select a treatment in
the test, PRESS THE ENTER KEY!  Remember, the sum of the coefficients must be zero!
Enter contrast coefficent for treatment  1 :  1
Enter contrast coefficent for treatment  2 :  0
Enter contrast coefficent for treatment  3 : -1

The computed contrast =    -0.100                    Its standard error    =     0.417
The test statistic (normalized contrast)   =              -0.240
The critical value  =                                      2.322
Prob {    -1.068 < True Contrast <       0.868 } =       0.9000

For hypothesis testing the absolute value is used ...
```

he null hypothesis is that there is NO difference among treatments. Do not reject the null hypothesis.

Two-Way ANOVA

An evaluation of an encapsulation applied to 3 different materials was conducted at 2 different laboratories. Each laboratory tested 3 samples from each of the treated materials. The data represents the time till leakage of each sample.

	Materials (B)		
Lab (A)	1	2	3
1	4.1	3.1	3.5
	3.9	2.8	3.2
	4.3	3.3	3.6
2	2.7	1.9	2.7
	3.1	2.2	2.3
	2.6	2.3	2.5

Questions:

a. Is there a difference between the materials?
b. Is there a difference between the labs?
c. Is there interaction between laboratory and material?

This is a 2x3 factorial experiment.
Labs (factor A) is at a = 2 levels.
Materials (factor B) is at b = 3 levels.
Each factor combination occurs r = 3 times.
The total number of observations = n = rab = 3(2)(3) = 18.

Form the input matrix to the SEMSTAT ANOVA Program:
There are ab = 2(3) = 6 rows.
There are r = 3 columns.

Let us input the data such that the LEFT most factor (e.g A) varies first:

$A_1 B_1$	4.1	3.9	4.3
$A_2 B_1$	2.7	3.1	2.6
$A_1 B_2$	3.1	2.8	3.3
$A_2 B_2$	1.9	2.2	2.3
$A_1 B_3$	3.5	3.2	3.6
$A_2 B_3$	2.7	2.3	2.5

The actual file, created by using the INPUT program, omits the Ai Bj headings.

TWO-WAY ANOVA FOR EQUAL SAMPLE SIZES PER CELL

The input to this program is a tabular ASCII file. The rows are the cells with the replicates. The columns are the factors, arranged with the FIRST factor varying first. Enter name of the data file, or press F10 to exit. ?ANOVA2.fil

Number of levels for Factor A: ? 3
Number of levels for Factor B: ? 2
Number of Replicates : ? 3

****
* ANALYSIS OF VARIANCE TABLE *
****

Source	SS	DF	MS	F-RATIO
Factor A	0.4444	2	0.2222	4.4444
Factor B	1.8050	1	1.8050	36.1000
Interaction AB	5.0800	2	2.5400	50.8000
Error	0.6000	12	0.0500	
Total (corrected)	7.9294	17		

F AND P VALUES FOR FIXED MODEL

Probability of F(A) =	0.964059	Right Tail Area =	0.035941
Probability of F(B) =	0.999939	Right Tail Area =	0.000061
Probability of F(AB) =	0.999999	Right Tail Area =	0.000001

Want contrast testing ? (y/N): y
Enter desired factor/interaction (A, B, AB, etc) or press Enter to quit: a
Input α or press Enter for default of .05:
Means
2.078 2.078 1.856

****
* CONTRAST TESTING *
****

A normalized contrast is the contrast divided by its standard error. There is a maximum value for the normalized contrast. The contrast coeffcients for the maximum normalized contrast are:
1.225 1.225 -2.449
You will now be prompted for contrast coefficents. If you do NOT wish to select a treatment in the test, PRESS THE ENTER KEY! Remember, the sum of the coefficients must be zero!
Enter contrast coefficent for treatment 1 : 1
Enter contrast coefficent for treatment 2 : 0
Enter contrast coefficent for treatment 3 : -1

The computed contrast = 0.222 Its standard error = 0.105
The test statistic (normalized contrast) = 2.108
The critical value = 2.788
Prob { -0.072 < True Contrast < 0.516 } = 0.9500

The null hypothesis is that there is NO difference among treatments.
Do not reject the null hypothesis

Three-Way ANOVA FOR EQUAL SAMPLE SIZES PER CELL

The input to this program is a tabular ASCII file.
The rows are the cells with the replicates.
The columns are the factors, arranged with the FIRST factor varying first.
Enter name of the data file, or press F10 to exit. ? *anova3.fil*

Number of levels for Factor A: ? 2
Number of levels for Factor B: ? 2
Number of levels for Factor C: ? 2
Number of Replicates : ? 3

****
* ANALYSIS OF VARIANCE TABLE *
****

Source	SS	DF	MS	F-RATIO
Factor A	181.5000	1	181.5000	20.7429
Factor B	253.5000	1	253.5000	28.9714
Interaction AB	13.5000	1	13.5000	1.5429
Factor C	73.5000	1	73.5000	8.4000
Interaction AC	13.5000	1	13.5000	1.5429
Interaction BC	73.5000	1	73.5000	8.4000
Interaction ABC	1.5000	1	1.5000	0.1714
Error	140.0000	16	8.7500	
Total (corrected)	750.5000	23		

F AND P VALUES FOR FIXED MODEL

Probability of F(A)	= 0.999675	Right Tail Area	= 0.000325
Probability of F(B)	= 0.999939	Right Tail Area	= 0.000061
Probability of F(AB)	= 0.767908	Right Tail Area	= 0.232092
Probability of F(C)	= 0.989522	Right Tail Area	= 0.010478
Probability of F(AC)	= 0.767908	Right Tail Area	= 0.232092
Probability of F(BC)	= 0.989522	Right Tail Area	= 0.010478
Probability of F(ABC)	= 0.315658	Right Tail Area	= 0.684342

Multiple Comparisons

The input file to this routine is automatically generated by the
One Way Anova program under the name of "ANOVA.MNS"
For n-way Anova's select the desired factor or interaction and
then construct your own file with the means.
For this example the ANOVA.MNS file consist of one column and 3 rows:

3.5167
5.4667
4.5667

```
*********************************************
*    Confidence Limits on all Pairwise Differences        *
*    by Tukey's , Scheffe's and Bonferroni's methods      *
*********************************************

 Enter filename of the column file with the means: anova.mns
 Enter the MSE from the ANOVA table (0 to quit)  : 4.8
 Enter the family confidence coefficient, (default=.95) :
 Equal sample sizes?  y/n :
 Enter the total sample size (0 to quit)  : 18

                Means          Abs. Dif
 1 - 2    3.5167 -    5.4667 =    1.950
 1 - 3    3.5167 -    4.5667 =    1.050
 2 - 3    5.4667 -    4.5667 =    0.900

 Tukey's        T statistic:    2.5986
 Scheffe's      S statistic:    2.7134
 Bonferroni'    B statistic:    2.6943
 Sidak's        t statistic:    2.3347
 Student's      t statistic:    2.1320

  0.95  confidence limits for all pairwise comparisons.

                Tukey              Scheffe            Bonferroni
 1 - 2    -1.3370   5.2370    -1.4822   5.3822    -1.4581   5.3581
 1 - 3    -2.2370   4.3370    -2.3822   4.4822    -2.3581   4.4581
 2 - 3    -2.3870   4.1870    -2.5322   4.3322    -2.5081   4.3081
```

Simple Linear Regression

LEAST SQUARES LINEAR REGRESSION

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵), ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? regress1.fil

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized. In other prompts, the default is 0 (zero) unless indicated otherwise

The first 10 lines of the file...

```
20.0   89.5
14.8   79.9
20.5   83.1
12.5   56.9
18.0   66.6
14.3   82.5
27.5  126.3
16.5   79.3
24.3  119.9
20.2   87.6
```

The number of columns in this file = 2

You can extract a selected number of columns from this file. The selection is made by entering the numbers of the desired columns, separated by at least one blank. Consider a file of 5 columns. The file can be re-arranged by changing the order of these numbers. For example, 1 4 constructs a new file consisting of columns 1 and 4, and 1 5 3 constructs a new file consisting of columns 1, 5, and 3.

Enter the selected columns, (press Enter to retain all):
If the next prompt is answered by pressing the enter key, or 0, the system will generate the X variable: 1,2,...n.

How many independent variables ? 1
Want to name the variables? (1-8 letters) y/N: n
Degree of fit ? (1 for linear, 2 for quadratic, etc) : 1

If you know the number of data points (maybe from previous analysis), you can analyze all or part of the data. Enter one of the following:

a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry)
c) or press the enter key (↵) for all data. ?

Which column contains the DEPENDENT variable? Press Enter for 1: 2
To verify, do you want to see the first 10 records again? y/N : n

CORRELATION MATRIX OF THE VARIABLES

1.000 0.805
0.805 1.000

DEPENDENT VARIABLE 2

MAX	MIN	MEAN	STD.DEV	NO.DATA
126.3000	56.9000	88.8400	21.0954	15

SET OF INDEPENDENT VARIABLES

MAX	MIN	MEAN	STD.DEV	VARIABLE
27.5000	12.3000	18.1733	4.3768	1

SLCT	OPTION MENU	Help	Now
1	Transformations	F1	OFF
2	Seasonal Adjustment for Dependent Variable	F2	OFF
3	Discounted Least Squares	F3	OFF
4	Backwards Elimination	F4	OFF
5	Listing Residuals, Hat Diagonal and Cook's D	F5	OFF
6	Plot of Residuals vs Actuals or/and Fitted	F6	OFF
7	Plot of Actuals versus Computed	F7	OFF
8	Durbin-Watson and Box-Ljung Statistics	F8	OFF
9	Cochran-Orcut Transformation	F9	OFF
0	Plot of Forecasts and Prediction Limits	F10	OFF
Esc	Exit from this menu and continue the analysis		

Enter your selection (Do NOT press the Enter key) 5
Enter 1 for listing of residuals
Enter 2 for saving residuals in a file
Enter 3 for both listing and saving
Enter 0 for none of the above: 1
Enter name of output file to store the analysis or press Enter for ANALYSIS.REG

REGRESSION ANALYSIS FOR FILE regress1.fil

		estimate	std.error	t	p-value
Constant	:	18.3541	14.8077	1.2395	0.2371
B1	:	3.8785	0.7936	4.8872	0.0003

The Model is: 18.3541 + 3.8785*X1

REGRESSION ANALYSIS TABLE

Source	SS	DF	MS	F	p-value
Total (corrected)	6230.2188	14	445.0156		
Due to Regression	4034.3967	1	4034.3967	23.8850	59.35E-05
Due to Residuals	2195.8220	13	168.9094		
Correction Factor	118388.2031	1			

Standard Error of Estimate	:	12.9965
Coefficient of Multiple Determination, R^2	:	0.6476
Adjusted Coefficient of Determination, Ra^2	:	0.6204
Coefficient of Multiple Correlation , R	:	0.8047

CORRELATION MATRIX OF THE REGRESSION COEFFICIENTS

B0	B1
1.0000	-0.9740
-0.9740	1.000

Want listing of residuals? y/N: y

Data	Y observed	Y computed	Residual	Standardized Residual
1	89.5000	95.9248	-6.4248	-0.4943
2	79.9000	75.7564	4.1436	0.3188
3	83.1000	97.8641	-14.7641	-1.1360
4	56.9000	66.8358	-9.9358	-0.7645
5	66.6000	88.1677	-21.5677	-1.6595
6	82.5000	73.8172	8.6828	0.6681
7	126.3000	125.0138	1.2862	0.0990
8	79.3000	82.3499	-3.0499	-0.2347
9	119.9000	112.6025	7.2975	0.5615
10	87.6000	96.7005	-9.1005	-0.7002
11	112.6000	103.6819	8.9181	0.6862
12	120.8000	92.0463	28.7537	2.2124
13	78.5000	66.0601	12.4399	0.9572
14	74.3000	72.6536	1.6464	0.1267
15	74.8000	83.1256	-8.3256	-0.6406

Multiple Linear Regression

LEAST SQUARES LINEAR REGRESSION

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵), ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? regress3.fil

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized. In other prompts, the default is 0 (zero) unless indicated otherwise

The first 10 lines of the file...

89.5	20.0	5	4.1
79.9	14.8	10	6.8
83.1	20.5	8	6.3
56.9	12.5	7	5.1
66.6	18.0	8	4.2
82.5	14.3	12	8.6
126.3	27.5	1	4.9
79.3	16.5	10	6.2
119.9	24.3	2	7.5
87.6	20.2	8	5.1

The number of columns in this file = 4

You can extract a selected number of columns from this file. The selection is made by entering the numbers of the desired columns, separated by at least one blank. Consider a file of 5 columns. The file can be re-arranged by changing the order of these numbers. For example, 1 4 constructs a new file consisting of columns 1 and 4, and 1 5 3 constructs a new file consisting of columns 1, 5, and 3.

Enter the selected columns, (press Enter to return):

If the next prompt is answered by pressing the enter key, or 0, the system will generate the X variable: 1,2,...n.

How many independent variables ? 3

Want to name the variables? (1-8 letters) y/N:

Degree of fit ? (1 for linear, 2 for quadratic, etc) : 1

Enter 1 to generate all 1st order crossterms. Enter 2 to generate your own crossterms conversationally. Or press Enter for no crossterms
?

If you know the number of data points (maybe from previous analysis), you can analyze all or part of the data. Enter one of the following:

a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST)
b) or just the first sequence number, e.g. 12, (last number is last entry)
c) or press the enter key (↵) for all data. ?

Which column contains the DEPENDENT variable? Press Enter for 1 :
To verify, do you want to see the first 10 records again? y/N :

CORRELATION MATRIX OF THE VARIABLES

1.000	0.805	-0.521	0.372
0.805	1.000	-0.809	-0.171
-0.521	-0.809	1.000	0.410
0.372	-0.171	0.410	1.000

DEPENDENT VARIABLE 1

MAX	MIN	MEAN	STD.DEV	NO.DATA
126.3000	56.9000	88.8400	21.0954	15

SET OF INDEPENDENT VARIABLES

MAX	MIN	MEAN	STD.DEV	VARIABLE
27.5000	12.3000	18.1733	4.3768	2
16.0000	1.0000	8.6667	4.0297	3
12.9000	4.1000	6.5400	2.3494	4

The same option menu was presented, in the interest of space the output is omitted...
Enter your selection (Do NOT press the Enter key) 4 and 5 were selected.

Enter name of output file to store the analysis or press Enter for ANALYSIS.REG

REGRESSION ANALYSIS FOR FILE regress3.fil

		estimate	std.error	t	p-value	
Constant	:	-16.0568	19.0711	-0.8419	0.4177	
B1	:	4.1461	0.7512	5.5195	0.0002	1
B2	:	-0.2361	0.8812	-0.2680	0.7937	2
B3	:	4.8309	0.9011	5.3613	0.0002	3

The Model is: -16.0567 + 4.1461*X1 - 0.2361*X2 + 4.8309*X3

REGRESSION ANALYSIS TABLE

Source	SS	DF	MS	F	p-value
Total (corrected)	6230.2188	14	445.0156		
Due to Regression	5707.4209	3	1902.4736	40.0293	33.38E-07
Due to Residuals	522.7976	11	47.5271		
Correction Factor	118388.2031	1			

Standard Error of Estimate : 6.8940
Coefficient of Multiple Determination, R^2: 0.9161
Adjusted Coefficient of Determination, Ra^2: 0.8932
Coefficient of Multiple Correlation , R : 0.9571

Want to perform the Backwards Elimination Procedure? y/N: y
Input alpha level or press Enter for default of .1:

Leaving variable is : 2 : B2
Computed F value is : .0718026 with DF: 1 11
Computed F level is : .7936882

PRESENT SET

		estimate	std.error	t	p-value	
Constant	:	-20.3718	9.8139	-2.0758	0.0601	
B1	:	4.3117	0.4104	10.5059	0.0000	1
B3	:	4.7177	0.7646	6.1705	0.0000	2

Present Residual Variance : 43.85073
Previous Residual Variance : 47.52705
Significance of difference : .3633695

END OF SEARCH.

CORRELATION MATRIX OF THE REGRESSION COEFFICIENTS

B0	B1	B3
1.0000	-0.8471	-0.6395
-0.8471	1.0000	0.1710
-0.6395	0.1710	1.0000

Want listing of residuals? y/N: y

Data	Y observed	Y computed	Residual	Standardized Residual
1	89.5000	85.2048	4.2952	0.6486
2	79.9000	75.5218	4.3782	0.6612
3	83.1000	97.7396	-14.6396	-2.2108
4	56.9000	57.5848	-0.6848	-0.1034
5	66.6000	77.0532	-10.4532	-1.5786
6	82.5000	81.8579	0.6421	0.0970
7	126.3000	121.3167	4.9833	0.7525
8	79.3000	80.0211	-0.7211	-0.1089
9	119.9000	119.7854	0.1146	0.0173
10	87.6000	90.7849	-3.1849	-0.4810
11	112.6000	104.2072	8.3928	1.2674
12	120.8000	122.4090	-1.6090	-0.2430
13	78.5000	77.9522	0.5478	0.0827
14	74.3000	66.8830	7.4170	1.1201
15	74.8000	74.2786	0.5214	0.0787

Want to see the Box-Meyer Method to study dispersion effects? y/N:

Want listing of residuals with Cook's D statistics? y/N:

FORECASTING SECTION

Defaults are obtained by pressing the enter key, without input.
Default for number of forecasts = 6.
Default for the prediction band around the forecast = 90%.

How many points ahead to forecast? (F3 or 9999 to quit...): 9999

Chapter 4

Utility Routines

Factorials

```
Enter the number you wish to factorialize: (Enter to quit) 6
FACTORIAL OF 6 =            720

Enter the number you wish to factorialize: (Enter to quit) 10.5
FACTORIAL OF 10.5 =         11899423.08396225

Enter the number you wish to factorialize: (Enter to quit) 0
FACTORIAL OF 0 =            1

Enter the number you wish to factorialize: (Enter to quit) 170
FACTORIAL OF 170 =          7.257D+306

Enter the number you wish to factorialize: (Enter to quit)
```

Contingency Tables

The following table consists of 2 rows and 3 columns. Our task is to test whether the row and column classifications are independent from each other.

249	494	201
26	26	4

```
                    CHI-SQUARE CONTINGENCY TABLE
The input to this program is a tabular ASCII file.
Enter name of the data file, or press F10 to exit.
? chisqr.fil

Enter alpha or press Enter for .05:

The test statistic =                13.2459
The number of degrees of freedom = 2
The critical value =                 5.9383

H0 is that the two classifications are independent.
REJECT THE NULL HYPOTHESIS
```

Example of Combinations
COMB

```
Combinations of N items take X at a time

Enter N (Enter to quit)          12
Enter X                           4
The number of combinations of 12  taken 4 at a time = 495

Combinations of N items taken X at a time

Enter N (Enter to quit)          52
Enter X                          13
The number of combinations of 52  taken 13 at a time =  6.350136D+011
```

Example of Matrix Inversion
MATINV

```
The input matrix must be an ASCII file and could be one of the following:
   1.  A square matrix, ready to be inverted.
   2.  A columnar matrix, that will be squared by the X'X operation.
Enter 1 or 2: 1
Enter name of matrix or press Enter to exit: MATRIX.DAT
Enter dimension: 5

 INPUT MATRIX
        -1.0000        -5.0000        -6.0000        -1.0000        -1.0000
        -8.0000        -1.0000         5.0000         2.0000        11.0000
        -7.0000        13.0000         1.0000         2.0000        -4.0000
         1.0000         6.0000         1.0000        -2.0000        -3.0000
        -3.0000         5.0000        -5.0000         4.0000         6.0000
 INVERSE
        -0.1060        -0.0614        -0.0777         0.0240         0.0550
        -0.0239        -0.0003        -0.0020         0.1095         0.0499
        -0.1002         0.0189         0.0145        -0.0543        -0.0689
        -0.1708        -0.1080         0.0787        -0.3641         0.0399
        -0.0027         0.0573        -0.0776         0.1183         0.0686
Enter filename to store the inverse or press Enter: MATRIX.INV
```

Example of the Determinant
DETERMINANT

The input matrix must be an ASCII file and could be one of the following:
1. A square matrix, ready to be inverted.
2. A columnar matrix, that will be squared by the X'X operation.

Enter 1 or 2: 1
Enter name of matrix or press Enter to exit: DETER.DAT
Enter dimension: 5

INPUT MATRIX

-1.0000	-5.0000	-6.0000	-1.0000	-1.0000
-8.0000	-1.0000	5.0000	2.0000	11.0000
-7.0000	13.0000	1.0000	2.0000	-4.0000
1.0000	6.0000	1.0000	-2.0000	-3.0000
-3.0000	5.0000	-5.0000	4.0000	6.0000

THE DETERMINANT IS: -29192

Goodness of Fit Test Using the Chi Square Criteria

```
*********************************************************
*          Goodness of Fit  for a Normal or Poisson distribution      *
*********************************************************
```

The input to this program is a single or two-column ASCII file.
If the file is two-columns it is a frequency table:
The LEFT column contains the midpoints or values of the cells,
The RIGHT column contains the corresponding frequencies.
Enter the FILESPEC: To return, press F10.
? bookf.bj

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry),
c) or press the enter key (↵) for all data. ?

For Data File bookf.bj

MAX	MIN	MEAN	STD.DEV.	NO.DATA
80.0000	23.0000	51.1286	11.9090	70

Enter value for alpha or press Enter for .05:
Enter 1 for a Poisson fit or press Enter for Normal:
How many cells for the frequency table? 6

Enter lower bound, press Enter for the minimum 23: 20
Enter upper bound, press Enter for the maximum 80:

CELL NO.	CLASS CELL BOUND	OBSERVED FREQUENCY	EXPECTED CUMULATIVE FREQUENCY	EXPECTED CLASS FREQUENCY	CHI-SQUARED TERMS
1	30.000	3	2.661	2.348	0.181
2	40.000	9	12.252	9.591	0.036
3	50.000	16	32.358	20.105	0.838
4	60.000	29	54.029	21.672	2.478
5	70.000	7	66.043	12.014	2.093
6	80.000	6	69.463	3.420	1.946
TOTALS		70		69.150	7.573

The test statistic = 7.57
The critical value = 7.78
Accept the null hypothesis

This test requires that each expected class frequency is ≥ 5.
If this is grossly violated the test is only approximate.
Rerun and reduce the number of cells .

The Box-Cox Transformation

Very often, we assume the normal distribution when we make statistical inferences. Unfortunately, in many cases this assumption is not substantiated. What can we do when the normality assumption is violated? One strategy is to attempt to make non-normal data resemble normal data by using a transformation. There is no dearth of transformations in statistics, the question is which one to select for the situation at hand.

Sometimes, knowing the subject matter is helpful in applying the appropriate transformation. But most of the time the choice of the transformation is not obvious. This was recognized in 1964 by two celebrated statisticians, G.E.P. Box and D.R. Cox. They teamed up and wrote a paper in which a useful family of power transformations was suggested. These transformations are only defined for positive values. This is not a restriction, because a single constant can always be added if the set of observations contains one or more negative values. The power Transformations are given by:

$$x(\lambda) = \frac{(x^{\lambda} - 1)}{\lambda} \quad if\ \lambda \neq 0$$

$$x(\lambda) = \ln(x) \quad if\ \lambda = 0$$

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x(\lambda)$$

$$f(x) = -\frac{n}{2}\ln\left[\frac{\sum_{i=1}^{n}(x_i(\lambda) - \bar{x}(\lambda))^2}{n}\right] + (\lambda - 1)\sum_{i=1}^{n}\ln(x_i)$$

where $\bar{x}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$ is the arithmetic mean of the transformed data.

Given the observations x_1, x_2, ... x_n, the Box-Cox solution for the maximum likelihood estimator of the power λ is the one that maximizes the logarithm of the likelihood function:

In addition a confidence interval can be constructed for λ. A set of λ values that represent an approximate 100(1-α)% confidence interval for λ is determined by those λ that satisfy:

$$l(\lambda) \geq \dot{\lambda} - .5\chi^2(1-\alpha, 1)$$

where $\dot{\lambda}$ denotes the maximum likelihood for λ.

To illustrate the procedure, the data given by Johnson and Wichern's textbook, "Applied Multivariate Statistical Analysis", (Prentice Hall, 1988) in Example 4.14 were used. The observation are microwave radiation measurements.

.15	.09	.18	.10	.05	.12	.08
.05	.08	.10	.07	.02	.01	.10
.10	.10	.02	.10	.01	.40	.10
.05	.03	.05	.15	.10	.15	.09
.08	.18	.10	.20	.11	.30	.02
.20	.20	.30	.30	.40	.30	.05

Values of Lambda and the log likelihood function

LAMBDA	LLF	LAMBDA	LLF
-2.0	7.1146	0.0	104.8276
-1.9	14.1877	0.1	105.8406
-1.8	21.1356	0.2	106.3947
-1.7	27.9468	0.3	106.5069
-1.6	34.6082	0.4	106.1994
-1.5	41.1054	0.5	105.4985
-1.4	47.4229	0.6	104.4330
-1.3	53.5432	0.7	103.0322
-1.2	59.4474	0.8	101.3254
-1.1	65.1147	0.9	99.3403
-1.0	70.5226	1.0	97.1030
-0.9	75.6471	1.1	94.6372
-0.8	80.4625	1.2	91.9643
-0.7	84.9421	1.3	89.1034
-0.6	89.0587	1.4	86.0714
-0.5	92.7855	1.5	82.8832
-0.4	96.0974	1.6	79.5521
-0.3	98.9722	1.7	76.0896
-0.2	101.3923	1.8	72.5061
-0.1	103.3457	1.9	68.8106
		2.0	65.0113

This table shows that lambda = .3 maximizes the log likelihood function.
This becomes .28 when a second digit of accuracy was calculated.

Linear Programming Using The Simplex Method

The Simplex Method in this program is used to solve small to medium scale Linear Programming problems. The maximum number of constraints is 100. The maximum number of variables depends on the available memory for applications. It is calculated and printed by the program.

This includes slack, surplus and artificial variables. In considering the number of variables, note that each '≤' constraint uses one slack variable and each '≥' constraint uses one surplus AND one artificial variable, and that each '=' constraint uses one artificial variable. At 550K bytes of memory the maximum problem is 100 constraints and 250 variables.

INPUT AND NAMING CONVENTION

The objective function is entered algebraically as follows: Z = aX1 + bX2 + ... or: -Z = aX1 + bX2 + ... (- to minimize). Names of variables can be up to 16 characters. Long names are not recommended because of the danger of typos. The objective function can be spread over ten lines of 255 characters each. To proceed to the next line, enter a slash (/) between the name of the variable and the + or - symbol. The last line CANNOT end in a slash.

The constraints are entered algebraically, as follows:

aX1 + bX2 +... < RHS for 'less than or equal to constraints'
aX1 + bX2 +... > RHS for 'greater than or equal to constraints'
aX1 + bX2 +... = RHS for 'equal to constraints'

The constraints can also be named, (via prompts in the program) and these names can be up to 16 characters. Constraints occupy one line only. The file can be constructed during execution of the program, or it can be prepared in advance, using any ASCII producing wordprocessor or editor. To add a name, enter a colon (:) to the right of the RHS, then the name. When you name your file, be sure to have '.LP' as the extension.

For assignment/transportation problems, enter only the objective function. When the program asks for the constraints, just hit the ENTER key. After that you can choose the type of problem you have from a menu.

The Simplex Method for Linear Programming	
	Enter Data, Solve Problem
	Perform Sensitivity Analysis
	Change an Objective Function Coefficient
	Change a Right-Hand Side Constant
	Add a New Constraint
	Transportation/Assignment/Games Demos
	End Application (can use 0 or Enter)

Manual Input.
Blanks between entries are optional. The objective function is entered algebraically as follows: Z = aX1 + bX2 + ... or: -Z = aX1 + bX2 + ... (- to minimize). You have 255 columns to work with. Keep entering, after 80 columns, the next line automatically appears. If you need more, enter a slash (/) AFTER the last name but BEFORE the + or - and press Enter, for another 255. Then continue entering. You have up to 10 slashes. The maximum name length = 16.

The constraints are entered algebraically, as follows:
aX1 + bX2 +... < RHS for 'less than or equal to constraints'
aX1 + bX2 +... > RHS for 'greater than or equal to constraints'
aX1 + bX2 +... = RHS for 'equal to constraints'
To stop entering, press the Enter key.

EXAMPLE
Maximize the following objective function
Z = .03X1 + .035X2 + .04X3 + .045X4 + .05X5 + .055X6

Subject to: 1X3 + 1X4 < 350
1X5 + 1X6 < 350
1X1 + 1X2 > 400
1X1 + 1X2 + 1X3 + 1X4 + 1X5 + 1X6 = 1000

SIMPLEX (the syntax to execute the LP program)
Max number of constraints = 100
Max number of variables (inclusive of slack, surplus and artificial) = 199

```
                    ***  FINDING THE OPTIMAL SOLUTION  ***
CREATE NEW   DATA SET ====> y/N ?
The following files are available
═══════════════════════════════════════════════════════════════════
AGG1        .LP   AGG2        .LP   ASSIGNA     .LP   ASSIGNB    .LP
GAMES       .LP   HEXNUT      .LP   INVEST      .LP   MULTIPLE   .LP
NETWORK     .LP   PRODUCT     .LP   TESTCASE    .LP   XPORTA     .LP
XPORTB      .LP   XPORTC      .LP   XPORTD      .LP
 10985472 Bytes free
═══════════════════════════════════════════════════════════════════
 If you want another directory or drive, enter a $,
 otherwise enter file name (NO extension) or F3 to return to the menu: invest

Enter 'Y' or 'y' for constraint names, or press Enter for defaults:

Display the set up ? y/N:

Type  Y (or press Enter) to continue or type N to edit the data:

Wish to print tableau after each iteration ?  y/N:

                              *** OPTIMAL  SOLUTION ***
PRIMAL BASIS OR ACTIVITY LEVELS
 VARIABLE          SOLUTION
    Z    =          44.5000
 2  X2   =         400.0000
 4  X4   =         250.0000
 6  X6   =         350.0000
 7  ≤SLK 1=        100.0000

REDUCED COST OF THE DECISION VARIABLES
 1  X1          =          0.0050
 3  X3          =          0.0050
 5  X5          =          0.0050
```

*** SENSITIVITY ANALYSIS ***

OBJECTIVE FUNCTION RANGES

VARIABLE	SOLUTION	LOWER	GIVEN	UPPER
1 X1	0.0000	-INFINITY	0.0300	0.0350
2 X2	400.0000	0.0300	0.0350	0.0450
3 X3	0.0000	-INFINITY	0.0400	0.0450
4 X4	250.0000	0.0400	0.0450	0.0550
5 X5	0.0000	-INFINITY	0.0500	0.0550
6 X6	350.0000	0.0500	0.0550	INFINITY

SHADOW RIGHT-HAND SIDE RANGES

CONSTRAINT	PRICE	LOWER	GIVEN	UPPER
1 ≤SLK 1	0.0000	250.0000	350.0000	INFINITY
2 ≤SLK 2	0.0100	250.0000	350.0000	600.0000
3 ≥SUR 3	-0.0100	300.0000	400.0000	650.0000
4 =	0.0450	750.0000	1000.0000	1100.0000

Applications of Linear Programming	
1	Transportation Problem A
2	Transportation Problem B (Dual of A)
3	Transportation Problem C (Dummy Source)
4	Transportation Problem D (Restriction of Supply)
5	Assignment Problem A (Minimize Cost)
6	Assignment Problem B (Maximize Profit)
7	Two Person Zero-Sum Games Problem
0	Return to the Main Menu. Esc: Return to DOS.

TRANSPORTATION PROBLEM A

The following matrix represents the per unit cost for items shipped from source R(i) to destination C(j).

	C1	C2	C3
R1	100	60	150
R2	50	120	110

The following matrix represents the quantities shippe, X(i,j), the variables and the capacitities of the source of supply as well as the requirements of destination of the demands. The capacities are of the 'at most' variety, and the requirements of the 'at least' variety.;

		DESTINATION		
SOURCE	C1	C2	C3	SUPPLY
R1	X1	X2	X3	≥ 3000
R2	X4	X5	X6	≥ 1000
DEMANDS	≤1500	≤1700	≤600	

The objective is to minimize the total shipping cost. Surplus is allowed. The destination will sell at cost.

To formulate the above situation as a linear programming problem :
Minimize Z=100X1 + 60X2 + 150X3 + 50X4 + 120X5 + 110X6 subject to

$$
\begin{array}{rrrrrrr}
X1 & & & +X4 & & & \leq 1500 \\
 & X2 & & & +X5 & \leq 1700 & \\
 & & X3 & & & +X6 & \leq 600 \\
X1 + & X2 + & X3 & & & & \geq 3000 \\
 & & & X4 + & X5 + & X6 & \geq 1000
\end{array}
$$

Note in general:

1) If supply ≥ demand, there is always a solution
2) If demand > supply, 'dummy' supplies must be introduced, for example, a row of zeros.
3) If certain cells are not allowed, assign a very large value.

These data are stored in the file: XPORTA.LP. The minimum shipping cost = $292,000.
The solution is:

X1 = 500
X2 = 1700
X3 = 600
X4 = 1000

TRANSPORTATION PROBLEM B

This is the dual of problem A. The problem becomes now:

Find non-negative numbers X1, X2, X3, X4, and X5, which maximize: Z=1500X1 + 1700X2 + 600X3 - 3000X4 - 1000X5 subject to:

$$
\begin{array}{rrrrrr}
X1 & & & -X4 & & \leq 100 \\
 & X2 & & -X4 & & \leq 60 \\
 & & X3 & -X4 & & \leq 150 \\
X1 & & & & -X5 & \leq 50 \\
 & X2 & & & -X5 & \leq 120 \\
 & & X3 & & -X5 & \leq 110
\end{array}
$$

These data are stored in the file: XPORTB.LP. With the dual solutions, the same results as in problem A are obtained.

TRANSPORTATION PROBLEM C

This is a modification of problem A. Source R1 is now limited to a capacacity of at most 2000, instead of the original 3000. The total demands remain unaltered, namely 1500, 1700, 600 respectively, for a total of 3800. Since source R2 still has the 1000 ceiling, demand exceeds supply. To handle this situation, a 'dummy' source is called upon. This source will supply ficticious quantities, which are of course the amounts NOT shipped to the designated destination The per unit cost matrix becomes now:

	C1	C2	C3
R1	100	60	150
R2	50	120	110
R dummy	0	0	0

And the quantities-shipped/supply/demand matrix becomes now:

	DESTINATION			
SOURCE	C1	C2	C3	SUPPLY
R1	X1	X2	X3	≤ 2000
R2	X4	X5	X6	≤ 1000
R dummy	X7	X8	X9	≤ 800
demands	≥1500	≥1700	≥600	

As before the total cost of shipping is to be minimized,
e.g., Minimize Z=100X1 +60X2 +150X3 +50X4 +120X5 +110X6 +0X71 +0X8 +0X9 subject to:

$$
\begin{aligned}
X1 + X4 + X7 &\geq 1500 \\
X2 + X5 + X8 &\geq 1700 \\
X3 + X6 + X9 &\geq 600 \\
X1 + X2 + X3 &\leq 2000 \\
X4 + X5 + X6 &\leq 1000 \\
X7 + X8 + X9 &\leq 800
\end{aligned}
$$

Or, another way analogus to problem B, the dual can be solved.

These data are stored in the file: XPORTC.LP. The minimum shipping cost = 182,000. The final quantities-shipped matrix, including dummy source is:

	C1	C2	C3
R1	300	1700	0
R2	1000	0	0
R dummy	200	0	600

Hence, C1 will only obtain 1300 and C3 won't receive the goods at all

TRANSPORTATION PROBLEM D

This is a modification of problem C Source R2 is not allowed to supply destination C1. The only change in the setup is that in the per unit cost matrix an excessively large number in the disputed slot is entered. This is done to prevent this entry to appear in the final solution so the per unit cost matrix becomes now:

	C1	C2	C3
R1	100	60	150
R2	1000000	120	110
R dummy	0	0	0

and the objective function reads now:
minimize Z= 100X1 +60X2 +150X3 +1000000X4 +120X5 +110X6 +0X7 +0X8 +0X9
subject to the same constraints as in problem C.

These data are stored in the file: XPORTD.LP

The minimum shipping cost = 262,000

The quantities-shipped matrix ,including dummy source is:

	C1	C2	C3
R1	700	1300	0
R2	0	400	600
R dummy	800	0	0

Here, C1 only receives 700!

ASSIGNMENT PROBLEM A

A major college has rented 3 terminals of different types, A,B and C. Rental is a function of connect and CPU time. There are 4 locations available, some of which are more desirable then others due to workload and proximity of certain users. As a result, the objective is to assign these terminals to the available locations such that the total rental cost is minimized. The following matrix represents the expected yearly cost in 1000 dollars.

	LOCATION			
	1	2	3	4
A	13	10	12	11
B	15	13	20	
C	5	7	10	6

Location 2 is for political reasons off limits for terminal B!

To formulate this problem as an assignment problem, two tasks must be performed:

1) a dummy location must be assigned for the extra location
2) an extremely large value,say 1,000,000, should be used to assign B to 2, in order to prevent this from entering in the optimal solution.

The resulting assignment problem cost matrix becomes:

	1	2	3	4
A	13	10	12	11
B	15	1000000	13	20
C	5	7	10	6
D	0 0	0	0	

To formulate the above as a linear program:

Minimize Z= 13X1 + 10X2 + 12X3 + 11X4 +15X5 + 1000000X6 + 13X7 + 20X8
+ 5X9 + 7X10 + 10X11 + 6X12 + 0X13 + 0X14 + 0X15 + 0X16

subject to:

X1 + X5 + X9 + X13 = 1
X2 + X6 + X10 + X14 = 1
X3 + X7 + X11 + X15 = 1
X4 + X8 + X12 + X16 = 1
X1 + X2 + X3 + X4 = 1
X5 + X6 + X7 + X8 = 1
X9 + X10 + X11 + X12 = 1
X13 + X14 + X15 + X16 = 1

These data are stored in the file: ASSIGNA.LP

The solution is

	1	2	3	4
A		1		
B			1	
C	1			
D				1

That is, X2, X7, and X9 are in the final basis. The minimum cost is: 28

ASSIGNMENT PROBLEM B

The same college plans to expand its computer science curriculum. Three openings are to be filled, one to teach user oriented languages, (such as FORTRAN or C++), the second to teach Assembler and the third to teach Math/Stat.

Four candidates are interviewed by the rest of the faculty of the department. Each candidate has distinct qualifications. A rating scheme is used:

1) Each interviewer assigns a number between 1-10 as a measure of the candidate's potential.
2) The final entry is the pooled score of all interviewers per candidate, for each opening

The table below summarizes the job ratings.

		JOB		
		FORTRAN	ASSEMBLER	MATH
	A	5.3	7.9	8.3
CANDIDATE	B	9.0	8.3	7.2
	C	7.6	4.3	6.0
	D	3.4	7.5	8.1

The department wishes of course to maximize potential subject to:

1) every candidate is assigned to at most 1 opening
2) each opening is filled by at most 1 person

To formulate above as an assignment problem:

Maximize Z= $5.3X1 + 7.9X2 + 8.3X3 + 9.0X4 + 8.3X5 + 7.2X6 +$
$7.6X7 + 4.3X8 + 6.0X9 + 3.4X10 + 7.5X11 + 8.1X12$

subject to:

$X1 + X4 + X7 + X10 \leq 1$
$X2 + X5 + X8 + X11 \leq 1$
$X3 + X6 + X9 + X12 \leq 1$
$X1 + X2 + X3 \leq 1$
$X4 + X5 + X6 \leq 1$
$X7 + X8 + X9 \leq 1$
$X10 + X11 + X12 \leq 1$

These data are stored in the file: ASSIGNB.LP

The solution is:

		JOB		
		FORTRAN	ASSEMBLER	MATH
	A		yes	
CANDIDATE	B	yes		
	C			
	D			yes

The optimum is : 25, e.g. an average rating of 8.333

Solution of 2 Person Zero-Sum Games

Let the matrix A be the payoff matrix from person 2 to person 1

		2	
	A_{11}	A_{12}	A_{13}
1	A_{21}	A_{22}	A_{23}
	A_{31}	A_{32}	A_{33}

This means: Person 1 receives Aij from Person 2, when 1 selects row i and 2 column j. The problem is to determine the best strategy for each player in the selection of rows and columns Let Person 1 select the 3 rows with probabilities x1, x2, x3. Let Person 2 select the 3 columns with probabilities y1, y2, y3. Of course each probability must be ≥ 0, and their sum = 1.

Now consider Person 2's point of view: Depending on Person 1's choice of row, he or she has one of the following 3 quantities as expected loss:

$L1 = A_{11}y_1 + A_{12}y_2 + A_{13}y_3$
$L2 = A_{21}y_1 + A_{22}y_2 + A_{23}y_3$
$L3 = A_{31}y_1 + A_{32}y_2 + A_{33}y_3$

Let L be the maximum loss. The objective is to minimize this loss.

But if L > 0, minimizing L is the same as maximizing 1/L. L can be forced to become > 0, by adding the largest negative entry to each element in A.

This must be later subtracted from the computed optimal value. The computation of probabilities is not affected by this device. Since y1 + y2 + y3 = 1, it follows:
maximize z = y1/L + y2/L + y3/L =1/L.
Let g = y/L, then: maximize 1/L = g1 + g2 + g3 becomes the objective function subject to the following constraints:

$A_{11}g1 + A_{12}g2 + A_{13}g3 \leq 1$
$A_{21}g1 + A_{22}g2 + A_{23}g3 \leq 1$
$A_{31}g1 + A_{32}g2 + A_{33}g3 \leq 1$

Example: find optimal stratigies for both players, and optimal payoff for the game with matrix:

		2	
	0	1	1
1	1	0	2
	2	1	0

then, the objective function is: z = g1+g2+g3 and the constraints are:

g2 + g3 ≤1
g1 + 2g3 ≤1
2g1 + g2 ≤1

The data are stored in the file GAME.LP
The solutions are as follows:
Optimum value for 1/L = 1.2. Hence L = 5/6

g1 = 2/10. Hence y1 = L*g1 = 5/6*2/10) = 1/6.
g2 = 6/10. Hence y2 = L*g2 = 3/6
g3 = 4/10. Hence y3 = L*g3 = 2/6

Let p(i) denote x(i)/L. Then the probabilities are obtained from the dual solution.

p1 = 6/10. Hence x1 = L*p1 = 3/6
p2 = 2/10. Hence x2 = L*p2 = 1/6
p3 = 4/10. Hence x3 = L*p3 = 2/6

Chapter 5

Statistical Quality Control Charts

The main menu is given below:
Use the mouse or the arrow keys to position the cursor. <– is for programs, –> for help.

SLCT Keys	SQC CONTROL CHARTS		Help Keys
1	SHEWHART	Shewhart XBAR and S control charts	F1
2	SHEWRANG	Shewhart XBAR and R control charts	F2
3	SHEWXBAR	Shewhart XBAR control charts	F3
4	SHEW1	Individual X and Moving Range control charts	F4
5	CUSUM	Cumulative Control Charts for means	F5
6	PRTCUM	Tabular and Graphical Cusum Charts	F6
7	EWMA	Exponential Weighted Moving Average ctl charts	F7
8	ACCEPT	Acceptance Sampling Plan and OC curves	F8
	MOREMENU	More Control Charts	
	ARLMENU	ARL's for the above control charts	
	MULTMENU	Multiple Control Charts	
		Return	

The Shewhart $\bar{x}$, s control chart

The Shewhart $\bar{x}$, s control chart scheme consists of two charts: The $\bar{x}$ chart plots the averages of the samples.

The center line is the overall average or a target that ser can select.
The upper control limit, UCL, is defined as:

$$UCL = T + 3\,\bar{s} / (c_4 \sqrt{n})$$

and the lower control limit, LCL as:

$$LCL = T - 3\,\bar{s} / (c_4 \sqrt{n})$$ where:

T is the grand mean (overall average) or Target,
$\bar{s}$ is the average of the sample standard deviations,
n is the sample or subgroup size, and
c_4 is a constant, to correct for the bias of S, the sample standard deviation. S estimates $c_4\sigma$. c_4 is defined as:

$$c_4 = \sqrt{(2/n-1)}\ [\ \Gamma(n/2) / \Gamma(n-1)/2\ ]$$

$\Gamma(x) = (x-1)!$, x is a positive integer or a multiple of 1/2.

The s-chart plots the standard deviations of the samples.

The center line is the average of the sample standard deviations,
$CL = \bar{s}$.
The $UCL = \bar{s} + 3\,\bar{s} / c_4\ [\sqrt{(1 - c_4^2)}]$
The $LCL = \bar{s} - 3\,\bar{s} / c_4\ [\sqrt{(1 - c_4^2)}]$

Here is an example:

```
**************************************************************
*          SHEWHART XBAR AND S CONTROL CHARTS                *
**************************************************************

 You can enter a valid filespec, as long as it has an extension, or you
 can select a file extension to search for files of particular interest.
 If you merely press the enter key (<┘), ALL file names are displayed.
 Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? thick.sqc
```

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized.
In other prompts, the default is 0 (zero) unless indicated otherwise.

To view and/or correct the file press the F6 key, followed by Enter.

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry),
c) or press the enter key (↵) for all data. ?

For Data File thick.sqc

	MAX	MIN	MEAN	STD.DEV.	NO.DATA
XBAR	6.5000	3.5000	4.7625	0.7884	20
S	3.1623	0.8165	1.7993	0.7134	

Subgroup size? *4*

C4	B3	B4
0.9213	0.0000	2.2660

Type the target or press Enter for the mean:

Type your own Sigma or press Enter to use the program's estimate:

XBAR CONTROL LIMITS
3-SIGMA UPPER CONTROL LIMIT = 5.9945
3-SIGMA LOWER CONTROL LIMIT = 3.5305

S CONTROL LIMITS
3-SIGMA UPPER CONTROL LIMIT = 4.0772
3-SIGMA LOWER CONTROL LIMIT = 0.0000

The $\bar{x}$ chart
The $\bar{x}$ chart is the same as in the $\bar{x}$, s program, except that it appears on a full screen page, instead of a half. In addition the 1, 2, and 3 sigma upper and lower control limit lines are drawn.
There is no s chart

The Control Chart for Individuals
The Control Chart for Individual Units consists of two charts:

The $\bar{x}$ chart plots the individual measurements.

The center line is the average or a target selected by the user.
The UCL = $T + 3\bar{r} / d2$
The LCL = $T - 3\bar{r} / d2$
where:

T is the average or Target,
$\bar{r}$ is the average of the moving ranges of 2 observations
d2 is a constant, 1.128

The Shewhart $\bar{x}$, R control chart
The Shewhart $\bar{x}$, R control chart scheme consists of two charts: The $\bar{x}$ chart plots the averages of the samples. The center line is the overall average or a target that the user can select. The control limits for the subgroup averages are:
UCL: $T + A_2 \bar{r}$
LCL: $T - A_2 \bar{r}$

where:

T is the grand mean (overall average) or Target,
$\bar{r}$ is the average of the sample ranges ,
A_2 is a tabulated factor depending on the the sample size of the subgroup

The R chart plots the ranges of the subgroups.
The center line is the Average Range. The control limits are:
UCL = $D_4 * \bar{r}$
LCL = $D_3 * \bar{r}$
D_3 and D_4 are tabulated factors, depending on the sample size.

The Moving Range chart
The Moving Range chart plots the moving ranges.

The center line = $\bar{r}$
The UCL = $\bar{r}(3.267)$ (The D_4 factor for n = 2)
The LCL = 0

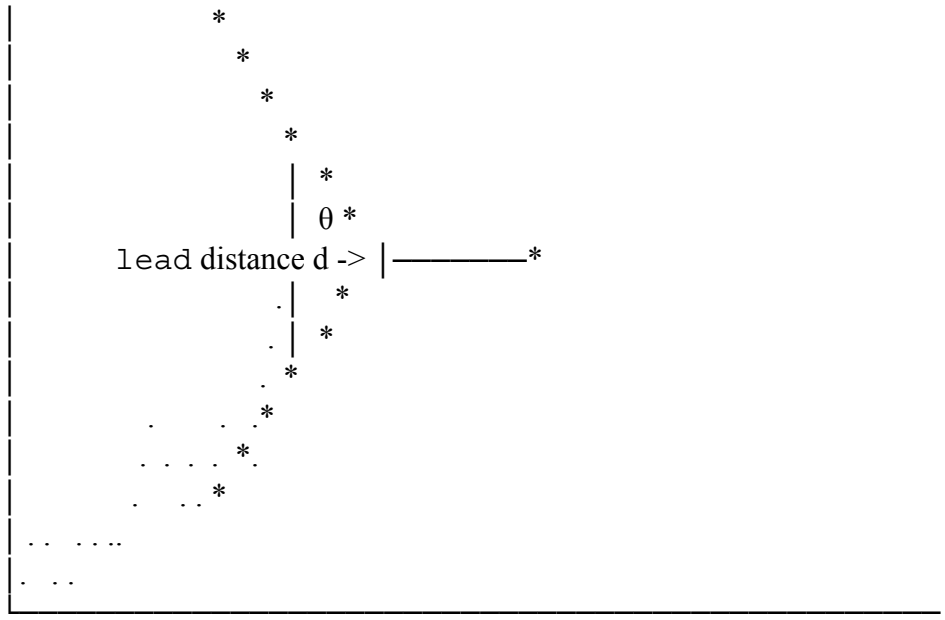
The Cusum Control Chart
The cumulative (or cusum) control chart is an alternative to the Shewhart chart. It incorporates all of the information from the sample values by plotting the cumulative sums as follows:

$S_m = \sum_{i}^{m} (\bar{x}_i - \mu_0)$ $\bar{x}$ is the average of sample i against the sample number m.

μ_0 is the target for the process mean. S_m is the cusum for m samples.

Cusum charts are more effective that Shewhart charts for detecting small process shifts. They are particularly applicable for samples of n = 1. The *V mask* by Barnard(1959) is applied to determine whether the process is out of control. It is placed with the vertical bar on the last point. The process is out of control if at least one of the points lies outside the arms of the V mask.

VMASK



The Cusum (Y-scale) is divided by the standard deviation of the mean, $\sigma_{\bar{x}}$ in order to avoid scaling problems.
The performance of the cusum control chart is determined by the parameters of the V mask. Quite often one uses the lead distance, d, and the angle θ to define the V mask.
Let α be the probability of a false alarm, that is: concluding that a shift has occurred while in fact it did not. Let ß the probability of missing to detect a shift in the mean. Let delta be the shift in the process mean that we wish to detect, in terms of standard deviations of xbar. Then a widely used design for the cumulative control chart is:

$$d = (2/\delta^2)\ln[(1-ß)/\alpha]$$

and

$$\theta = \arctan(\delta/2) \ .$$

where

$$\delta = \text{delta} / (\sigma \text{ xbar})$$

If ß is small we can use the following equation for d,

$$d = -2 \ln(\alpha) / \delta^2$$

Example:

```
                CUMULATIVE SUM CONTROL CHARTS
  You can enter a valid filespec, as long as it has an extension, or you
+  can select a file extension to search for files of particular interest.
  If you press the enter key (<-┘), ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? volts.sqc

  NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized.
       In other prompts, the default is 0 (zero) unless indicated otherwise.

  To view and/or correct the file press the F6 key, followed by Enter.

  You can analyze all or part of the data. Enter one of the following:
  a) First AND last sequence number, e.g.  12-46 (the hyphen is a MUST),
  b) or just the first sequence number, e.g. 12, (last number is last entry),
  c) or press the enter key (<-┘) for all data.  ?


     MAX          MIN          MEAN         STD.DEV.     NO.DATA
   328.5000     324.1250      325.9537       1.5483          20

  Group Sample Size? 4


  Enter Target or Press Enter for the Mean   : 325
  Enter Sigma or Press Enter for the Std.Dev :

  The size of the shift you wish to detect is given in standard deviations.
  Enter size of shift, (default = 1 std.dev):
  Enter alpha risk,    (default = .00135)   :


  Save the plot on disk?  y/N:
  TO RETURN TO THE PROGRAM, AFTER THE PLOT IS DISPLAYED, PRESS ENTER.

  Enter X coordinate for V-Mask, or press Enter for 20:
```

PRTCUM
This is an extension of the regular CUSUM program. It does NOT divide the cusum by $\sigma_{\bar{x}}$ and calculates θ by :

$$\theta = \arctan (\text{delta} / 2s)$$

s is a scale factor relating the verticale scale to the horizontal. It is set to $2\sigma_{\bar{x}}$ but can be changed by the user. delta is the shift (in units of σ) that we wish to detect. δ = delta / σ (σ means the standard deviation of xbar)

$$d = (2/\delta^2)\ln[(1-ß)/\alpha] \quad \text{as before}$$

PRTCUM also offers a tabular form of the V mask as follows: Let h = s d tan (Θ) and k = h/d

Then calculate for i = 1 to n, the number of samples:
$S_{hi(i)} = \max [\ 0, S_{hi(i-1)} + x_i - \text{target} - k]$
$S_{lo(i)} = \min [\ 0, S_{lo(i-1)} + x_i - \text{target} + k]$

where $S_{lo(0)}$ an$S_{hi(0)} = 0$. If $S_{hi(i)} > h$ or $S_{lo(i)} < -h$, the process is considered out of control.

An example of this type of chart, using the same data as for the "regular" cusum chart is given on the next page.

CUMULATIVE SUM CONTROL CHARTS
You can enter a valid filespec, as long as it has an extension, or you
can select a file extension to search for files of particular interest.
If you merely press the enter key (↵), ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? *volts.sqc*

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized.
In other prompts, the default is 0 (zero) unless indicated otherwise.
To view and/or correct the file press the F6 key, followed by Enter.

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b) or just the first sequence number, e.g. 12, (last number is last entry),
c) or press the enter key (↵) for all data. ?

MAX MIN MEAN STD.DEV. NO.DATA

328.5000 324.1250 325.9537 1.5483 20

Group Sample Size? 4

Enter Target or press Enter for the Mean : 325
Enter Std.dev or press Enter for the program's :

The size of the shift you wish to detect is given in standard deviations.
Enter size of shift, (default = 1 std.dev):
Enter alpha risk, (default = .00135) :
Enter beta risk, (default = .10) :
Enter Scale factor or press Enter for 1.548:

Tabular Output? y/N: y

d	Θ (deg)	U0 (TARGET)	H	K (H/d)
13.0046	14.0362	325.0000	5.0338	0.3871

		Increase in mean		decrease in mean		
SAMPLE	X	X-U0	X-U0-K	S hi	X-U0+K	S lo	CUSUM
1	324.92	-0.08	-0.46	0.00	0.31	0.00	-0.08
2	324.67	-0.33	-0.71	0.00	0.06	0.00	-0.40
3	324.73	-0.27	-0.66	0.00	0.11	0.00	-0.68
4	324.35	-0.65	-1.04	0.00	-0.26	-0.26	-1.33
5	325.35	0.35	-0.04	0.00	0.74	0.00	-0.98
6	325.23	0.23	-0.16	0.00	0.61	0.00	-0.75
7	324.13	-0.88	-1.26	0.00	-0.49	-0.49	-1.63
12	325.15	0.15	-0.24	0.00	0.54	0.00	-2.50
13	328.33	3.33	2.94	2.94	3.71	0.00	0.83
14	327.25	2.25	1.86	4.80	2.64	0.00	3.08
15	327.83	2.83	2.44	7.24*	3.21	0.00	5.90
16	328.50	3.50	3.11	10.35*	3.89	0.00	9.40

* = out of control

Press Enter to continue

d	Θ (deg)	U0 (TARGET)	H	K (H/d)
13.0046	14.0362	325.0000	5.0338	0.3871

		Increase in mean		decrease in mean		
SAMPLE	X	X-U0	X-U0-K	S hi	X-U0+K	S lo	CUSUM

17	326.67	1.67	1.29	11.64*	2.06	0.00	11.08
18	327.77	2.77	2.39	14.03*	3.16	0.00	13.85
19	326.88	1.88	1.49	15.52*	2.26	0.00	15.73
20	328.35	3.35	2.96	18.48*	3.74	0.00	19.08

* = out of control

→

	Increase in mean			decrease in mean			
SAMPLE	X	X-U0	X-U0-K	S hi	X-U0+K	S lo	CUSUM
8	324.52	-0.48	-0.86	0.00	-0.09	-0.58	-2.10
9	325.23	0.23	-0.16	0.00	0.61	0.00	-1.88
10	324.60	-0.40	-0.79	0.00	-0.01	-0.01	-2.27
11	324.63	-0.38	-0.76	0.00	0.01	0.00	-2.65
12	325.15	0.15	-0.24	0.00	0.54	0.00	-2.50
13	328.33	3.33	2.94	2.94	3.71	0.00	0.83
14	327.25	2.25	1.86	4.80	2.64	0.00	3.08
15	327.83	2.83	2.44	7.24*	3.21	0.00	5.90
16	328.50	3.50	3.11	10.35*	3.89	0.00	9.40
17	326.67	1.67	1.29	11.64*	2.06	0.00	11.08
18	327.77	2.77	2.39	14.03*	3.16	0.00	13.85
19	326.88	1.88	1.49	15.52*	2.26	0.00	15.73
20	328.35	3.35	2.96	18.48*	3.74	0.00	19.08

* = out of control

The Exponential Weighted Moving Average (EWMA) Control Chart

The underlying model is the Integrated Moving Average Process of order (0, 1, 1) , described by Box and Jenkins in their Time Series Analysis book (1976). This model is expressed as follows:

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t = \bar{z}\ \pi\ + a_t$$

where $z(\pi)_{t-1}$ is a weighted moving average of the previous values of the process, and 'a' is the shock or disturbance at time t. It can be shown that $\Sigma\, \pi = 1$, so that the above expression is indeed the weighted sum divided by the sum of the weights.

It can also be shown that the $\pi_j = \delta(1-\delta)^{j-1}$ for $j > 0$ and $0 < \delta < 2$

The weighted moving average has exponentially decreasing values:

$\delta,\ \delta(1-\delta),\ \delta(1-\delta)^2,\ \delta(1-\delta)^3$ ...

hence the name EWMA.
Stu Hunter (1986) showed that the variance of the EWMA is: $[\delta/(1-\delta)]\sigma^2$. An estimate of σ^2 can be obtained from the minimum sum of squares that exists while estimating δ from the observed values. We still need is a TARGET, associated with a process under control. This also serves as the initial predicted value of the EWMA. Once we have the target, and computed the δ and its variance, it is simple to construct the EWMA control chart.

Example

```
**************************************************************
* EXPONENTIALLY WEIGHTED MOVING AVERAGE CONTROL CHART *
**************************************************************
You can enter a valid filespec, as long as it has an extension, or you can select a file extension to
search for files of particular interest. If you merely press the enter key (<—┘), ALL file names
are displayed. Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.
? dougewma.sqc

For Time Series dougewma.sqc
      MAX          MIN         MEAN     VARIANCE   NO.DATA
    15.0000       6.0000      10.3158      5.3435       38

Type the target or press Enter for the mean: 10.5
Type Lamda or press Enter to use the program's estimate:

Smallest value allowed for Lamda is .05...

LAMDA =  0.0500
RESIDUAL VARIANCE  =     5.6174
EWMA SIGMA         =     0.3795
Type your own Sigma or press Enter to use the program's estimate:

3-SIGMA UPPER CONTROL LIMIT =    11.6386
3-SIGMA LOWER CONTROL LIMIT =     9.3614

Press Enter to view the EWMA plot
Save the plot on disk?  y/N:  n
```

SLCT Keys		**SQC OTHER QUALITY CONTROL CHARTS**	Help Keys
1	NP	P and NP control charts	F1
2	CU	C and U control Charts	F2
3	STAR	Stepwise Autoregression and AR control charts	F3
4	RUNSUM	Run Sum Control Chart	F4
5	CIMTUKEY	Cimera-Tukey Modified Median Range ctl charts	F5
6	MACS	Moving Average control charts	F6
7	DOUBLE	Double Samplingplans	F7
8	SEQPLAN	Sequential Sampling Plans	F8
9	SKIPLOT	Skip Lot Sampling	F9
10	ASN	Average Sampling Number	F10
		Return to the main menu.	

Select by moving the cursor to the desired line and pressing Enter

P and NP control charts

The control chart for the fraction or number of nonconforming or defective product deals with the concept of a *binomial* distribution of D, the number of defectives in a sample of n.
The sample fraction of nonconforming, p-hat = D/n.

The p chart is for the fraction of nonconforming items.

The center line is the average of the sample p-hat's, $\bar{P}$

$$UCL = \bar{P} + 3\sqrt{\bar{P}(1 - \bar{P})/n}$$

$$UCL = \bar{P} - 3\sqrt{\bar{P}(1 - \bar{P})/n}$$

The number of inspection units may vary in size. In this case the input concists of two columns, the number of nonconforming and the corresponding sample sizes. The control limits are now:

$$UCL = \bar{P} + 3\sqrt{\bar{P}(1 - \bar{P})/n_i}$$

$$UCL = \bar{P} - 3\sqrt{\bar{P}(1 - \bar{P})/n_i}$$

It is also possible to base a control chart on the number of defectives, rather than the fraction.

The center line is $n\bar{P}$, where $\bar{P}$ is the average of the sample p-hat's

$$UCL = n\bar{P} + 3\sqrt{n\bar{P}(1 - \bar{P})}$$

$$UCL = n\bar{P} - 3\sqrt{n\bar{P}(1 - \bar{P})}$$

This is called the np control chart.

CU Charts

The control chart for the nonconformities or defects deals with the concept of a Poisson distribution of x, the number of defects or nonconformities in one inspection unit.
The Poisson parameter may be estimated by $C = \Sigma x / m$, that is, the average of the number of defects in m inspection units.

The C chart is for the number of nonconformities or defects.

The center line is C
The UCL = $C + 3\sqrt{C}$
The LCL = $C - 3\sqrt{C}$

It is also possible to base a control chart on the number of defects in more than one inspection unit, say n.
This is called the u control chart, where $U = C / n$.

The center line is U, where U is the average no. defects/unit.

The UCL = $U + 3\sqrt{U/n}$
The LCL = $U - 3\sqrt{U/n}$

The number of inspection units may vary in size. In this case the input concists of two columns, the number of nonconformities and the corresponding sample sizes. The control limits are now:

The UCL = $U + 3\sqrt{U/n_i}$
The LCL = $U - 3\sqrt{U/n_i}$

STAR

There are many techniques to analyze time series. To mention a few: Exponential Smoothing, by Holt and Winters. Exponential smoothing by Brown and the Box-Jenkins methods. One of these Box-Jenkins methods is the autoregressive (AR) model.
The form is :

$$Z_t - \mu = C + \phi_1 z_{t-1} + \phi_2 Z_{t-2} \ldots + \phi_n Z_{t-n} + a_t$$

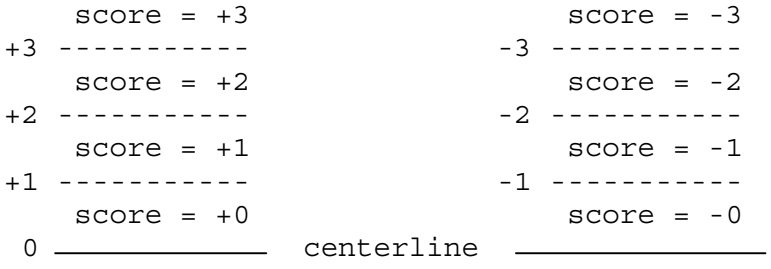
where C is a constant, μ is the mean, Z_t the observation at time t, and a_t is an error term at time t.

The ϕ parameters are estimated by the program, using stepwise regression. The program uses the 'F test' to determine if an added ϕ is necessary.

The program performs the *Box-Ljung Chi-Square test* to find out if the data should not have been de-seasonalized prior to analysis.
The program can detrend your data, using the 'trend-seasonal' method.
The forecasts and their confidence limits are then 're-trended'.

The Run Sum Control Chart

A simple way to control a process average is to use a decison rule based on the sum of scores assigned to the distance from the target of an observation. Here is how this works:
The process standard deviation σ is estimated, based on either the average of *n* sample standard deviations, or from another source.
Divide the control chart into zones one standard deviation wide. This is measured in either direction from the centerline of the chart. This centerline is usually the target or process average.



When a point (sample mean) falls above the centerline, and no higher than the limit of the first zone above the centerline, it gets a score of +0. If it falls in the second zone above the centerline, but below or on the upper limit of zone 2, it receives a score of +1; and so forth. Negative scores are assigned to pints falling below the center line according to the same scheme. The scores are accumulated until there is a change in the sign of the scores. If the accumulated score is greater in absolute value than four we conclude that the process is out of control.

Modified Median Range Control Chart

The control chart for individuals is based on the mean and moving

range. The limits are obtained from either the average moving range or the median moving range. Another word for moving range is first absolute difference between succesive observations. These first absolute differences are occassionaly poor candidates to yield meaningful estimates for å, because there may be many consecutive zeros among them.
A control chart to circumvent this was proposed by J. Ciminera and J. Tukey in a paper entitled 'Control - Charting Automated Laboratory Instruments when many Successive Differences may be Zero', in the Journal of Quality Technology, Vol; 21, No. 1, January 1989.
The procedure consists of the following 6 steps:

1. Find the differences between observations and take their absolute values.

2. Find their frequency distribution.

3. Move a cut-off up this distribution, starting from zero. go up until at least 50% has been accumulated. The upper tail should then be equal to or less than 50%.
   The cut-off value is the average of the bordering values.

4. Calculate *p*: $$\frac{\text{Remaining Upper Tail Count} + 1/6}{2(\text{Total Count}) + 2/3}$$

   This *p* is the area under the normal curve from z_p to infinity

5. Find z_p and then calculate:

$$s = \frac{\text{Cut-off Value}}{z_p \sqrt{2}}$$

6. The control limits are now: target +/- 3s

MACS

The Moving Average Control Chart works as follows:

Samples of size n have been collected and XBAR(1), ... XBAR(t) are the corresponding sample means.
The Moving Average of span *w* at time *t* is defined as:

$$M(t) = \frac{\bar{x}(t) + \bar{x}(t-1) + \ldots + \bar{x}(t - w + 1)}{w}$$

That is, at time t, the oldest sample mean is dropped and the newest is added.
The variance of the moving average, $V[M(t)] = \sigma^2/nw$
Then the parameters of the Moving Average control chart are:

The center line is T,where T is the total average or a target.

The UCL = $T + 3\sigma/\sqrt{wn}$
The LCL = $T - 3\sigma/\sqrt{wn}$

An estimator for σ^2 is the residual variance

DOUBLE SAMPLING PLANS

Double and multiple sampling plans were invented to give a questionable lot another chance. For example, if in double sampling the results of the first sample are not conclusive with regard to accepting or rejecting, a second sample is taken. Application of double sampling requires that a first sample of size n1 is taken at random from the (large) lot. The number of defectives is then counted and compared to the first sample's acceptance number a1 and rejection number r1. Denote the number of defectives in sample 1 by d1 and in sample 2 by d2, then:

if d1 < a1, the lot is accepted
if d1 > r1, the lot is rejected
if a1 < d1 < r1, the second sample is taken

If a second sample of size n2 is taken, the number of defectives, d2, is counted. The total number of defectives is D2 = d1 + d2. Now this is compared to the acceptance number a2, and the rejection number r2, of sample 2. In double sampling r2 = a2 + 1, to insure a decision on the sample.

if D2 < a2, the lot is accepted
if D2 > r2, the lot is rejected

DESIGN OF A DOUBLE SAMPLING PLAN

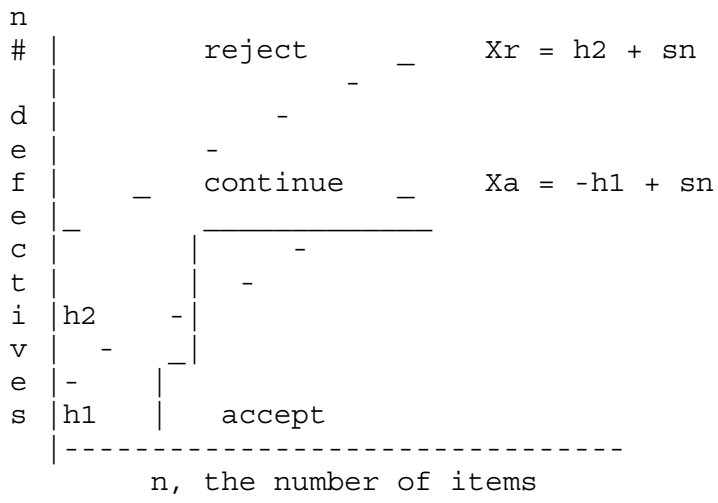
The required parameters to construct the OC curve are similar to the single sample case. The two points of interest are $(p_1, 1-\alpha)$ and (p_2, β), where p_1 is the lot fraction defective for plan 1 and p_2 is the lot fraction defective for plan 2. As far as the respective sample sizes is concerned, the second sample size must be equal to or an even multiple of the first sample size. There exist a variety of tables

that assist the user in constructing double and multiple sampling plans. The index to these tables is the p_2/p_1 ratio, where $p_2 > p_1$. A table, taken from the Chemical Corps Engineering Agency for $\alpha = .05$ and $\beta = .10$ and $n_1 = n_2$ is given below:

R= p_2/p_1	accept numbers a1	a2	approx. values of pn_1 for p=.95	p=.10	R= p_2/p_1	accept numbers a1	a2	approx. values of pn_1 for p=.95	p=.10
11.90	0	1	.21	2.50	3.21	3	7	2.15	6.91
7.54	1	2	.52	3.92	3.09	4	8	2.62	8.10
6.79	0	2	.43	2.96	2.85	4	9	2.90	8.26
5.39	1	3	.76	4.11	2.60	5	11	3.68	9.56
4.65	2	4	1.16	5.39	2.44	5	12	4.00	9.77
4.25	1	4	1.04	4.42	2.32	5	13	4.35	10.08
3.88	2	5	1.43	5.55	2.22	5	14	4.70	10.45
3.63	3	6	1.87	6.78	2.12	5	16	5.39	11.41
3.38	2	6	1.72	5.82					

Sequential Sampling

Sequential sampling is different from single, double or multiple sampling. Here one takes a sequence of samples from a lot. How many samples is a function of the results of the sampling process. The sequence can be one sample at a time, and then the sampling process is usually called item-by-item sequential sampling. One can also use sample sizes greater than one, in which case the process is referred to as group sequential sampling. Item-by-item is more popular so we concentrate on it. The operation of such a plan is illustrated below:



The cumulative observed number of defectives is plotted on the graph. For each point the x-axis is the total number of items thus

far selected, and the y-axis is the total number of observed defectives. If the plotted point falls within the parallel lines the process continues by drawing another sample. As soon a point falls on or above the upper line, the lot is rejected. And when a point falls on or below the lower line, the lot is accepted. The process can theoretically last until the lot is 100% inspected. However, as a rule of thumb, sequential sampling plans are truncated after the number inspected reaches three times the number that would have been inspected using a corresponding single sampling plan.

The equations for the two limit lines are functions of the parameters p_1, α, p_2, and β as follows:

$$X_a = -h1 + sn \text{ (acceptance line)}$$
$$X_r = h2 + sn \text{ (rejection line)}$$

where $k(h1) = \ln (1-\alpha)/\beta)$, $k(h2) = \ln(1-\beta)/\alpha$,

$$k = \ln \{p_2(1-p_1)/p_1(1-p_2)\},$$
and $$s = \{\ln [(1-p_1)/(1-p_2)]\}/k$$

For $n = 24$, $p_1 = .01$, $p_2 = .10$, $\alpha = .05$ and $\beta = .10$, the acceptance number is 0 and the rejection number is 3. The corresponding single sampling plan is (52,2) and double sampling plan is (21,0), (21,1).

Skip Lot Sampling

Skip Lot sampling means that only a fraction of the submitted lots are inspected. This mode of sampling is of the cost-saving variety in terms of time and effort. However skip-lot sampling should only be used when it has been demonstrated that the quality of the submitted product is very good.. A skip-lot sampling plan is implemented as follows:

1. Design a single sampling plan, by specifying the alpha and beta risks, and the consumer/producer's risks. This plan is called the *reference sampling plan*.
2. Start with normal lot-by-lot inspection, using the reference plan.
3. When a pre-specified number, i, of consecutive lots are accepted, switch to skipping inspection. Now a fraction f of the lots are inspected. The selection of the members of that fraction is done at random.
4. When a lot is rejected return to normal inspection.

The parameters f and i are essential to calculate the probability of

acceptance for a skip-lot sampling plan. In this scheme, i , called the clearance number is a positive integer and the sampling fraction *f* is such that 0 < *f* < 1. Hence, when *f* = 1 there is no longer skip-lot sampling. The calculation of the acceptance probability for the skip-lot sampling plan is performed via the following formula

$$Pa(f,i) = [\ fP + (1-f)\ p\hat{}i\ ]\ /\ [\ f+(1-f)\ p\hat{}i\ ]$$

The following relationships hold:
for a given i, the smaller is *f*, the greater is P_a
for a given *f*, the smaller is i, the greater is P_a

for an example, select SKIPLOT from the menu and enter *f*=.25 and i=5.

An important property of skip-lot sampling plans is the average sample number (ASN). The ASN of a skip-lot sampling plan is:

$$ASN\ (skiplot) = F\ ASN(reference)$$

where $F = f\ /\ \{\ [(1-f)P^i]\ +f\ \}$

Therefore, since 0 < F < 1, it follows that the ASN of skip-lot sampling is smaller than the ASN of the reference sampling plan.
In summary, skip-lot sampling is preferred when the quality of the submitted lots is excellent and the supplier can demonstrate an proven track record.

The Average Sample Number (ASN)

In single sampling, the sample size remains constant, while in double sampling the sample size depends on whether or not a second sample is required. The probability of a second sample will vary with p', the true fraction defective in an incoming lot.

Consider a double-sampling plan n1 = 50, c1 = 2, n2 = 100, c2 = 6, where n1 is the sample size for plan 1, with reject number c1, and n2, c2, are the sample size and reject number for plan 2.

Let p' = .06. Then the chance of acceptance on the first sample which is the chance of getting two or less defectives = .416 (using binomial tables)
The chance of rejection on the first sample, which is the chance of getting more than six defectives = 1-.971 = .029. The probability of making a decision on the first sample is .445, equal to the sum of .416 and .029
With complete inspection of the second sample the average size sample is equal to the size of the first sample times the probability that

there will only be one sample plus the size of the combined samples times the probability that a second sample will be necessary. For the sampling plan under consideration the average sample number (ASN) with complete inspection of the second sample for a p' of .06 is

$$50(.445) + 100(.555) = 106$$

Key	Average Run Lengths (ARLs)
	SHEWHART ARL CUSUM ARL (integral method) CUSUM ARL (Siegmund method) EWMA ARL (integral method) EWMA ARL (Markov chain method) DESIRED ARL for all (lambda, L) 's DESIGN for optimal (lambda, L) ARL for in-control MEWMA (integral) ARL for out-of-control MEWMA (Markov) SHOW replays stored plots EXIT

Some background on EWMA/MEWMA

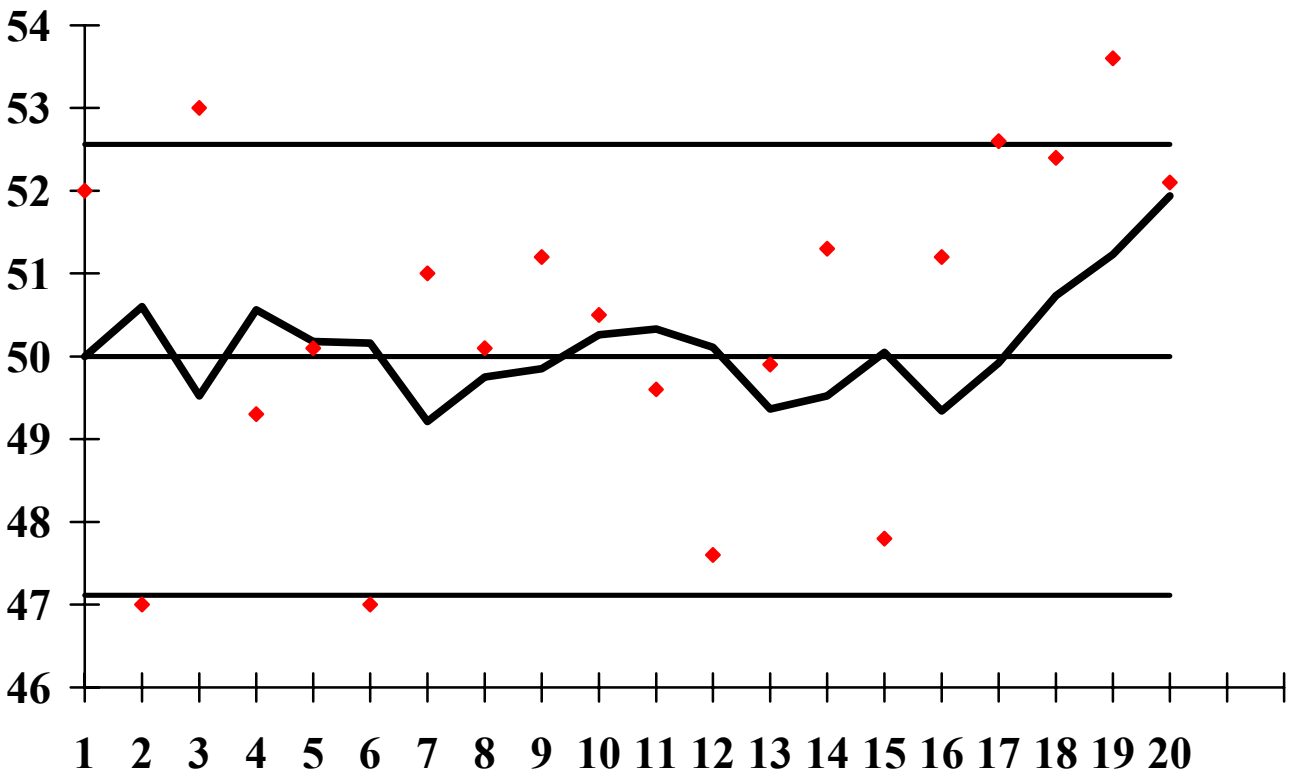
Univariate EWMA

Control schemes have been used by the majority of manufacturing organizations for the purpose of monitoring the process. Shewhart charts are the most popular and have demonstrated to be excellent in detected large shifts. Other control schemes such as the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) are better (faster) than the Shewhart method in detecting small shifts. The EWMA schemes are defined as follows:

Roberts (1959) gave us: $z_t = \lambda y_t + (1-\lambda) z_{t-1}$
Hunter (1986) issued this version: $z_{t+1} = \lambda y_t + (1-\lambda) z_t$
He used a Box-Jenkins IMA (1,1) model, where $\lambda = 1-\theta$.

In both schemes z_t is the current EWMA, y_t is the current observation and z_0 is the starting value, usually the target, and $0 < \lambda < 1$.

Example of an EWMA plot



The variance of the EWMA is given by $(\lambda/(2-\lambda))\sigma^2$. σ^2 is the population variance. σ^2 can be estimated from the computer program that estimates λ. The variance is needed to construct the control limits.

To start the computation, we need a TARGET, which also serves as the initial predicted value of the EWMA.

Consider the following data set:

52.0	47.0	53.0	49.3	50.1	47.0	51.0	50.1	51.2	50.5
49.6	47.6	49.9	51.3	47.8	51.2	52.6	52.4	53.6	52.1

Let λ = .3. Then using Hunter's formula, we obtain the following results, rounded to 2 decimal places.

50.00	50.60	49.52	50.56	50.18	50.16	49.21	49.75	49.85	50.26
50.33	50.11	49.36	49.52	50.05	49.34	49.92	50.73	51.23	51.94

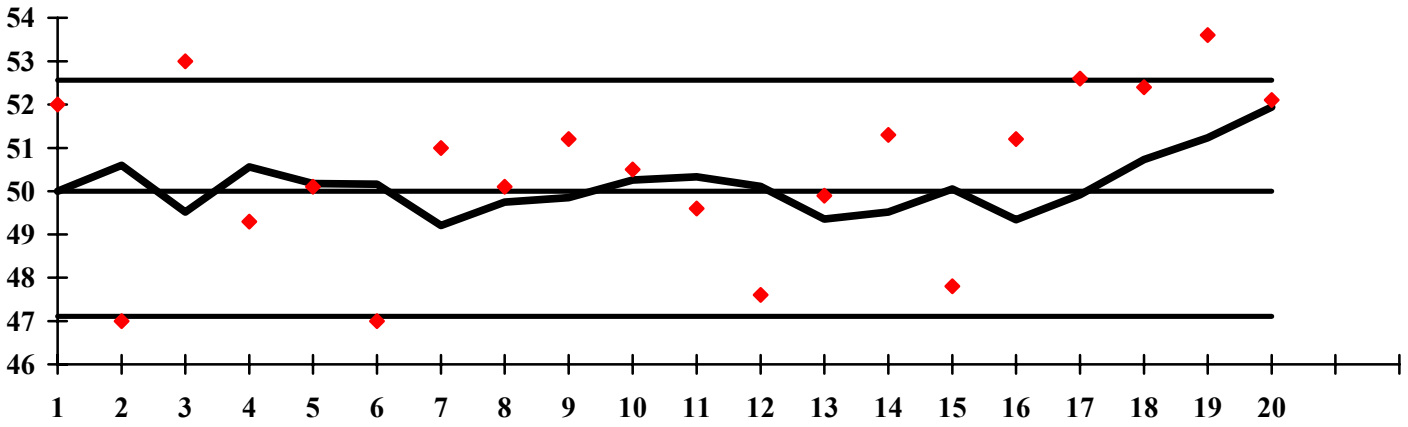
The calculated value for the residual variance is 4.2187. Therefore, the estimated standard deviation of the EWMA is

$$\hat{\sigma}_{ewma} = \sqrt{\frac{\lambda}{2-\lambda}}\hat{\sigma}^2 = \sqrt{\frac{.3}{1.7}}4.2187 = .8628$$

Setting the target at 50, we compute the control limits:

UCL = 50 + (3)(.8628) = 52.5884
LCL = 50 – (3)(.8628) = 47.4115

The resulting control chart appears below.



The dots are the actual observations and the connecting line is the EWMA
The properties of a control scheme can be described by considering their run length distribution. The run length is the number of samples taken until an out of control

signal is observed. The amount of production is proportional to the average run length (ARL) and therefore the ARL is often used to evaluate control schemes.

The ARL should be large when the process is in control and should be small when the process is out of control, that is, when it is operating off target.

The process is assumed to be operating on target as long as the control statistic (the EWMA) lies within the bounds of the UCL and LCL.

The control limits for the EWMA are usually chosen symmetrically around the target value as:

UCL = target + $L\ s_z$
LCL = target – $L\ s_z$,

where

$s^2_z = (\lambda/(2-\lambda))\ s^2_y$.

The two parameters, L and λ are chosen to give a specific in control ARL. This process is called "the design of an EWMA scheme".

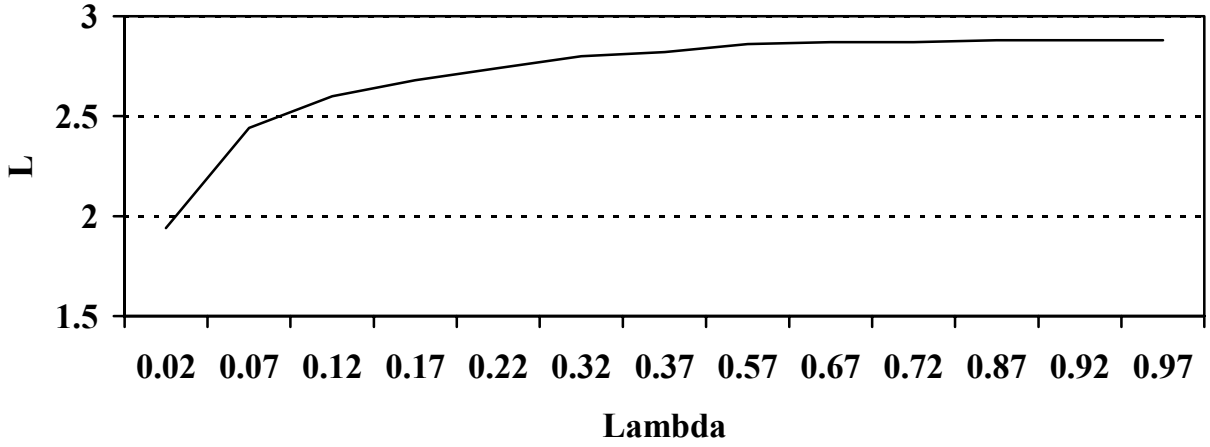
There are two ways to select λ.

Hunter advises to pick that λ which minimizes the sum of squares of the fit. In this case the ARL plays no role.

Other authors advise to calculate all λ and L which give a desired in control ARL and then choose the (λ , L) which yield the smallest out of control ARL for a specified mean shift.

If you wish to detect small shifts, a small λ is required. If you wish to detect bigger shifts, use a larger λ.

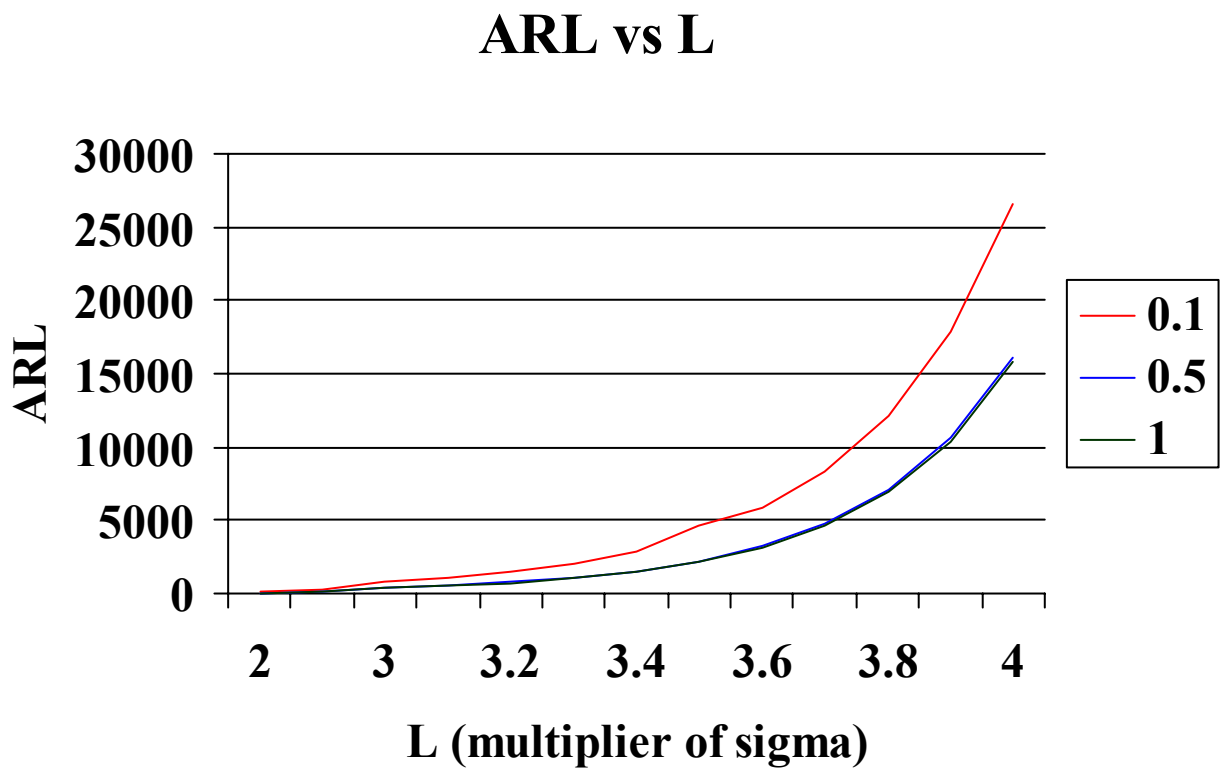
The graph below shows values of λ and L for a specific in control ARL.



The following table illustrates how to select the best λ , *L* combinations.

	1	2	3	
λ ->	0.070	0.370	0.970	Column no.
L ->	2.440	2.820	2.880	of min ARL
Shift				
0.00	252.711	249.740	251.504	
0.25	65.174	120.538	191.824	1
0.50	23.793	42.756	107.047	1
0.75	13.557	19.029	56.808	1
1.00	9.389	10.506	31.254	1
1.25	7.189	6.799	18.158	2
1.50	5.843	4.911	11.169	2
1.75	4.939	3.824	7.268	2
2.00	4.292	3.137	4.991	2
2.25	3.807	2.672	3.609	2
2.50	3.431	2.338	2.737	2
2.75	3.131	2.087	2.170	2
3.00	2.883	1.890	1.791	3
3.25	2.673	1.729	1.533	3
3.50	2.491	1.591	1.356	3
3.75	2.337	1.469	1.233	3
4.00	2.212	1.363	1.149	3

The following figure illustrates the behavior of the ARL for various values of *L*.



Multivariate EWMA

The Multivariate EWMA is an extension of the univariate EWMA as follows:

The input data matrix for *p* variables and *n* observations per variable looks like

$$Z_t = \Lambda Y_t + (1 - \Lambda) Y_{t-1}$$

where

Z_t is the t^{th} EWMA vector and

Y_t is the t^{th} observation vector

t = 1, 2, ..n is the number of observation vectors

Z_0 is a target vector, usually supplied by the user

Λ is the diagonal $(\lambda_1, \lambda_2, \lambda_3)$

$$\begin{matrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{matrix}$$

The control statistic is

$$T_i^2 = Z_i' \Sigma_{z_i}^{-1} Z_i$$

It has been shown (Lowry etal, 1992) that the (k,l) the elememt of the covariance matrix of the ith EWMA , Σ_{z_i}, is

$$\Sigma_{z_i}(k,l) = \lambda_k \lambda_l \frac{\left[1 - (1-\lambda_k)^i (1-\lambda_l)^i\right]}{\left[\lambda_k + \lambda_l - \lambda_k \lambda_l\right]} \Gamma_{k,l}$$

where $\Gamma_{k,l}$ is the (k,l) th element of Σ, the covariance matrix of the Y's.

If $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$, then the above expression simplifies to

$$\Sigma_{z_i} = \frac{\lambda}{2 - \lambda}\left[1 - (1-\lambda)^{2i}\right]\Sigma$$

where Σ is the covariance matrix of the input data.

If a single λ is selected then remember that small λ are good to detect small shifts and large λ are for large shifts.

Specify an in control ARL and find from tables in the listed reference the appropriate λ, *L* combination. The accompanying software will also perform this feat.

If several λ are used, the following procedure is suggested:

1) Compute the λ that results in the smallest sum of squares of the residuals for each of the corresponding series.
2) Select the smallest of these λ.
3) Compute the *L* factor that corresponds to a desired ARL and this λ.

The upper control limit is found be either Hotelling's UCL, which is $\chi^2(\alpha, p)$. For example let p , the number of variates be 2, and $\alpha = .01$. Then χ^2 at $1-.01 = 9.21$.

The other option is to find the best λ , L combination for a selected in-control ARL. There are many tables and graphs available in the literature , see the reference list.
Software can generate such combinations, as is illustrated in the following computer output.
This program solves an integral equation furnished by Rigdon (1995).

In what was presented thus far, the assumption was made that the data are uncorrelated. But this assumption is (probably) often violated.

A pronounced consequence is that the true ARL is shorter then was believed and as a result we arrive at wrong conclusions about the state of process control.

Fortunately it is possible to remedy this situation by fitting a time series model to the data and then apply EWMA/MEWMA control charts to the *residuals* from the fitted model.

The modeling can be performed by ARIMA models (from Box-Jenkins). The EWMA with $\lambda = 1-\theta$ is in fact an ARIMA (0, 1, 1) = IMA (1,1) model.

It turns out that by using the Hunter EWMA equation, the EWMA is the one-step-ahead forecast and the resulting prediction errors are independent.

Here are some general ARIMA models:

$$z_t = m + a_t - q a_{t-1}$$

This is the first-order-moving average (MA) process, where μ is the process mean , a_t is a random shock , z_t is the observation at time t and $|\theta| < 1$.

$$z_t = m + f z_{t-1} + a_t$$

This is the first-order autoregressive (AR) process where $|\phi| < 1$

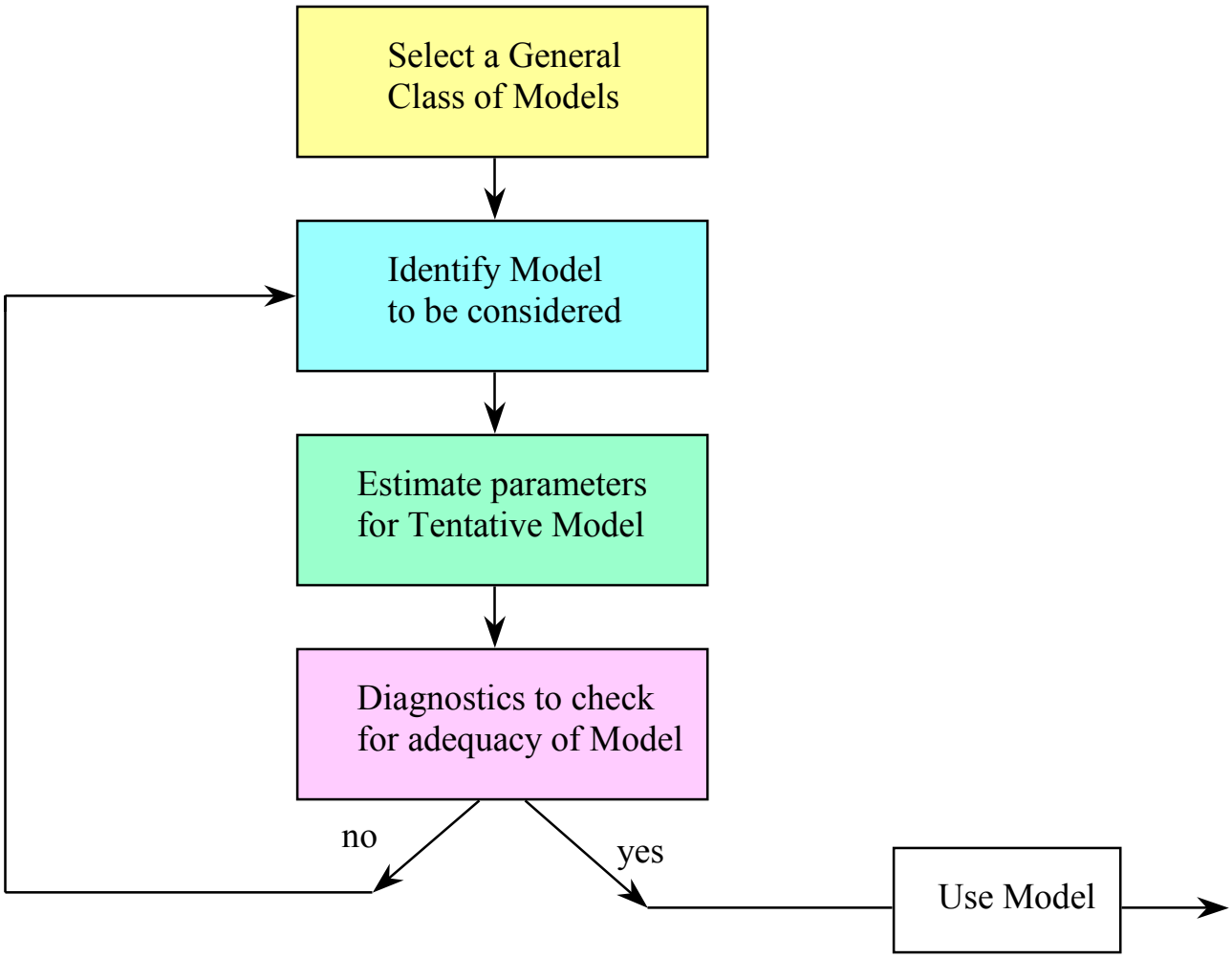
$$z_t - \phi z_{t-1} = a_t - \theta a_{t-1}$$

This is the first order AR-MA process.

$$w_t - \phi_1 w_{t-1} = a_t - \theta_1 a_{t-1}$$

This is the first order ARIMA (AutoRegressive Intergrated Moving Average) process. $w_t = z_t - z_{t-1}$ is the first difference. The I stands for *intergrated* (summed).

ARIMA model building is an iterative process according to the diagram below:



Markov Chains

A popular method used in computing the ARL is the application of Markov Chains..
A Markov Chain is a stochastic process with discrete time space as well as discrete state space. The change from one state to another is governed by the transition probability matrix. The Markov property holds: that is, the state the process is in at the present, depends on the process history only via the state the process was in just prior to the present.

An EWMA (or MEWMA) control scheme can be represented by a Markov Chain. Consider the formula for EWMA:
$EWMA_t = \lambda X_t + (1-\lambda)\, EWMA_{t-1}$ (using Robert's equation)
This verifies the Markov property that the current EWMA depends on the past only through the previous EWMA.
Furthermore, the EWMA is a continuous state MARKOV process and when we discretize the distance between the Upper and Lower Control Limits, it becomes a Markov Chain.

The procedure to obtain the ARL using Markov Chains is :

- Divide [LCL, UCL] into $2m+1$ subintervals.
- Set the value of the EWMA at time t to the midpoint of the interval into which it falls.
- The transition probability matrix has the form

$$\mathbf{P} = \begin{pmatrix} \mathbf{R} & (\mathbf{I}-\mathbf{R})\mathbf{1} \\ \mathbf{0}' & 1 \end{pmatrix}, \quad \text{where}$$

$$\mathbf{R} = (p_{ij}), \quad \text{and}$$

$$p_{ij} = P(E_t \text{ in state } j \,|\, E_{t-1} \text{ in state } i).$$

- Let **p'** be the initial probability vector. The probability that the process starts in control and goes out of control at or before time t is

$P(N \le t) = \mathbf{p'(I - R)1}$

From this we get

$P(N = t) = \mathbf{p'(R^{t-1} - R^{t})1.}$

Then the Average Run Length is calculated as:

$$ARL(n) = \sum_{t=1}^{\infty} tP(N = t)$$
$$= \sum_{t=1}^{\infty} t\mathbf{p'(R^{t-1} - R^{t})1.}$$
$$= \mathbf{p'(I - R)1}$$

The main steps in the calculations are

· Break up the distance between the UCL and LCL into N = 2m+1 subintervals of width 2δ. Denote the midpoints of these intervals by $x_j, j = 1, 2, ...N$

· Construct the **p** vector of length N, with all zeros and a 1 in the middle position, indicating that the process starts on target.

· Construct the R matrix, where each element is $p_{ij} = P(E_t = X_j \,|\, E_{t-1} = X_i)$

$$p_{ij} = \Phi\left[\frac{\left(X_j + \delta\right) - (1 - \lambda)X_i}{\lambda s}\right] - \Phi\left[\frac{\left(X_j - \delta\right) - (1 - \lambda)X_i}{\lambda s}\right]$$

where Φ [ ] is the cdf of the standard Normal distribution.

A numerical example, taken from Lucas and Saccuci (1990) may illustrate the above principles.

Consider a scheme with λ = .25 and L = 3 for a N(0,1) process. The target =0.

What is the in-control ARL?

The variance of the EWMA is :

$$s^2_{ewma} = \left[\frac{\lambda}{(2-\lambda)}\right] s^2_{population} = (.25/1.75)(1) = .142857$$

$$s_{ewma} = .378$$

Then the control limits for the EWMA are obtained from

$$\text{UCL} = \text{Target} + L\sigma_z = 1.134$$
$$\text{LCL} = \text{Target} - L\sigma_z = -1.134$$

Using 5 in-control states (m = 2) we find that $\delta = (\text{UCL} - \text{LCL}) / 10 = .22678$
and the corresponding midpoints are -.907 -.454 0 .454 .907

The first element of the R matrix is computed as follows:

$$A = \left[\frac{\left(X_j + \delta\right) - (1-\lambda)X_i}{\sigma_s}\right] = \left[\frac{(-.907 + .2268) - (.75)(-.907)}{(.25)(1)}\right] \approx 0$$

$$B = \left[\frac{\left(X_j + \delta\right) - (1-\lambda)X_i}{\sigma_s}\right] = \left[\frac{(-.907 - .2268) - (.75)(-.907)}{(.25)(1)}\right] = -1.81$$

$\Phi(0) = .500$ $\Phi(-1.81) = .0348$ and $p_{11} = \Phi(0) - \Phi(-1.81) = .4652$

By repeating this calculation for all the elements of **R** we get

$$R = \begin{bmatrix} .4652 & .4652 & .0347 & .0001 & .0000 \\ .0861 & .5881 & .3134 & .0116 & .0000 \\ .0032 & .1789 & .6357 & .1789 & .0032 \\ .0000 & .0116 & .3134 & .5881 & .0861 \\ .0000 & .0001 & .0347 & .4652 & .4652 \end{bmatrix}$$

The initial probability vector for zero state is
$\boldsymbol{P}' = (0 \quad 0 \quad 1 \quad 0 \quad 0)$

And

$\mathbf{(I - R)}^{-1}\ \mathbf{1}$ = (287 304 307 304 287) '

Then

ARL = $\boldsymbol{p}'\ (\mathbf{(I - R)}^{-1}\ \mathbf{1}$ = 307

The initial probability for steady state is formed by
$\Phi(LCL + j\ \delta) / \sigma_{ewma} - \Phi\ (LCL + (j-1)\ \delta)/ \sigma_{ewma}$ for j = 1, 2, …N
This is for a N(0,1) process.

The result is:
$\boldsymbol{P}'$ = (.04 .24 .43 .24 .04)
And the corresponding ARL = 304

Note: If we would have been interested in a shift in terms of σ, the procedure
is identical except that the shift has to be subtracted from the argument for Φ () that was used in the above example.
For example if the shift of interest would have been .5 the resulting ARL = 43.8

How good is the use of $N = 5$? As you may perhaps expect, 5 states does not yield precise results. Had we used, for example $N = 101$ (N must be odd) the computer program would have given:

N	Shift	ARL
101	0	502
101	.5	48

User Guide for the EWMA/MEWMA programs.

This part contains the program input and output. Programs can be run individually (while in DOS) or through a menu and mouse. The menu looks as follows:

key	ARLs and Control Charts
1	SHEWHART ARL
2	CUSUM ARL (integral method)
3	EWMA ARL (integral method)
4	EWMA ARL (Markov chain method)
5	DESIRED ARL for all (lambda, L) pairs
6	DESIGN for optimal (lambda, L) pairs
7	ARL for in-control MEWMA (integral)
8	ARL for out-of-control MEWMA (Markov)
9	Multivariate (MEWMA) Control Charts
	SHOW displays stored plots
Esc	EXIT

To run the routines on an individual basis type its name as is illustrated below.

ARL_SHEW

```
  ************************************************************
  * ARL for Shewhart Xbar Charts and 3 sigma control limits *
  * The ARL is 1/p, where p is the probability for a point  *
  * to fall outside the control limit.                      *
  ************************************************************

  Enter mean-shift in terms of sigma: 1

  The ARL =    43.89
```

ARL_GW

```
 ************************************************************
 *GOEL-WU's ARL for CUSUM Control charts by solving a       *
 * System of Linear Algebraic Equations from an integral    *
 * using a 16 point Gauss-Legendre Quadrature evaluation.   *
 ************************************************************

  Enter the standardized h: 5
  Enter the standardized k .5
  The Average Run Length (ARL) =     10.38
  The shift in the mean = (u-k)+.5 =  1.00
```

```
------------------------------------------------------------
```

ARL_EWMA

```
**********************************************************
*  Crowder's method for the ARL of EWMA Control charts,   *
*  solving a System of Linear Algebraic Equations and     *
*  using a 24 point Gauss-Legendre Quadrature evaluation. *
**********************************************************

  Enter the value of Lambda, or press Enter for .5      :.5
  Enter the shift in mean, in terms of σ of sample mean : 0
  Enter control limit multiple, k, or press Enter for 3 : 3
  Iterate to a shift of 4, in increments of .25?  y/n)  : y

  Lambda      k     shift   ARL
    0.50    3.00    0.00   397.46
    0.50    3.00    0.25   208.54
    0.50    3.00    0.50    75.35
    0.50    3.00    0.75    31.46
    0.50    3.00    1.00    15.74
    0.50    3.00    1.25     9.21
    0.50    3.00    1.50     6.11
    0.50    3.00    1.75     4.45
    0.50    3.00    2.00     3.47
    0.50    3.00    2.25     2.84
    0.50    3.00    2.50     2.41
    0.50    3.00    2.75     2.10
    0.50    3.00    3.00     1.87
    0.50    3.00    3.25     1.69
    0.50    3.00    3.50     1.53
    0.50    3.00    3.75     1.41
    0.50    3.00    4.00     1.31

------------------------------------------------------------
```

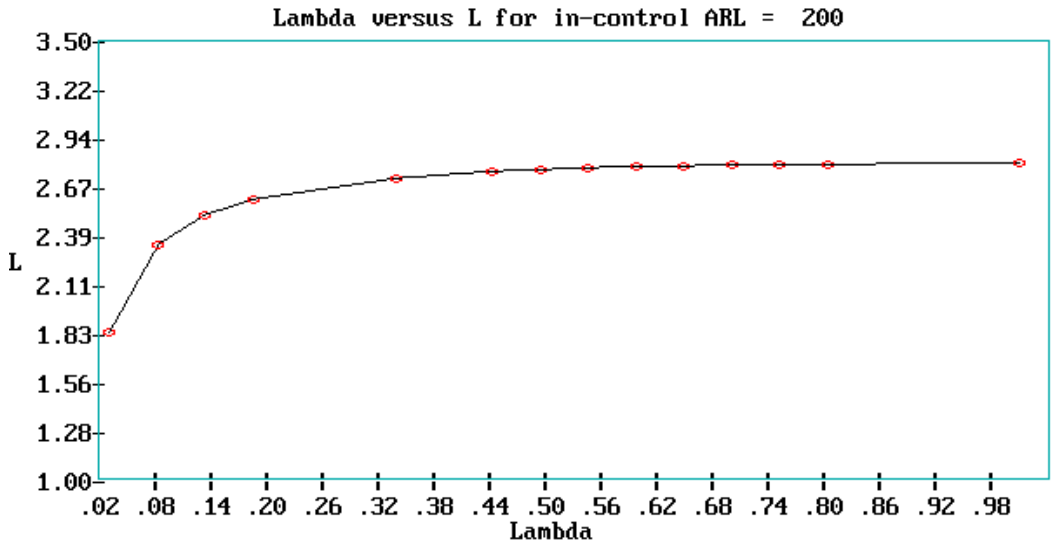
MARKOV

```
**************************************************************
* Computes zero state ARL for EWMA at N control states using  *
* Markov Chains                                               *
**************************************************************

Defaults for N, Lambda, L and shift are 83, .25, 3, 0 respectively
You invoke a default by pressing the Enter key at a given prompt.

  Input N, the number of control states : 101
  Input Lambda: .5
  Input L     : 3
  Input shift : 1
  The zero state ARL at 101 control states is :    15.738
------------------------------------------------------------
```

DESIRED

```
**********************************************************
* Finding all combinations of Lambda and L for which    *
* an in-control ARL is a desired number plus or minus   *
* a given delta, for example 370 +/- 5                  *
**********************************************************

 Enter desired in-control ARL: 200
 Enter the distance from this ARL, default =  2 : 2
 Enter a digit from 1 - 10 to indicate steps in the search
 1 means steps of .01, 2 means steps of .02, etc: 5

 Enter name for output file or press Enter for L200.fil
 The output is stored in file: L200.fil
```

Lambda	L	ARL
0.02	1.83	201.17
0.07	2.34	200.86
0.12	2.51	201.24
0.17	2.60	200.89
0.32	2.72	198.42
0.42	2.76	200.39
0.47	2.77	199.21
0.52	2.78	199.66
0.57	2.79	201.40
0.62	2.79	198.09
0.67	2.80	201.62
0.72	2.80	199.67
0.77	2.80	198.22
0.97	2.81	201.89



DESIGN

```
*********************************************
* Design of an EWMA control chart.          *
* The results are stored in file DESIGN.ARL  *
*********************************************

  Enter number of (lambda, L) pairs: 3

  Note! separate the lambda and L by a comma...
  Enter (Lambda,L) pair number  1: .2,3
  Enter (Lambda,L) pair number  2: .5,3
  Enter (Lambda,L) pair number  3: .9,3

                      1        2        3
       LAMBDA ->    0.200    0.500    0.900    Column no.
         L    ->    3.000    3.000    3.000    of min ARL

   Shift
-----------------------------------------------------------------
    0.00           559.878  397.463  370.954
    0.25           163.120  208.544  267.372      1
    0.50            44.127   75.354  136.551      1
    0.75            18.843   31.459   67.605      1
    1.00            10.836   15.738   35.309      1
    1.25             7.407    9.215   19.744      1
    1.50             5.605    6.111   11.831      1
    1.75             4.519    4.450    7.580      2
    2.00             3.801    3.468    5.170      2
    2.25             3.294    2.841    3.735      2
    2.50             2.919    2.413    2.840      2
    2.75             2.632    2.103    2.259      2
    3.00             2.408    1.870    1.868      3
    3.25             2.233    1.685    1.599      3
    3.50             2.094    1.535    1.410      3
    3.75             1.982    1.410    1.277      3
    4.00             1.885    1.305    1.182      3

-----------------------------------------------------------------
Conclusion:
Use lambda = .2 to detect shifts up to 1.5 sigma
Use lambda = .5 to detect shifts from 1.5 to 3.0 sigma
Use lambda = .9 to detect shifts from 3.0 to 4.0 sigma
-----------------------------------------------------------------
```

MEWMIN

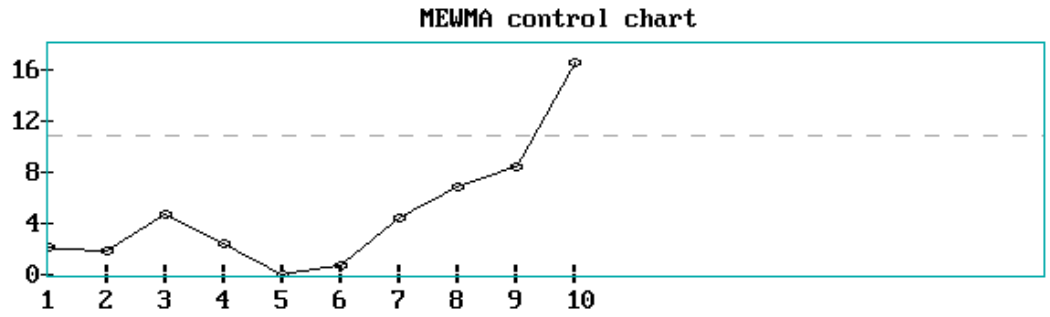
```
    ***************************************************
    * This routine computes the upper control limit    *
    * (h4) for the MEWMA control chart when the        *
    * desired in-control ARL is given and the number   *
    * of series (np) and the smoothing constant (r).   *
    * adapted from a paper by Brodden and Rigdon (1999)*
    ***************************************************

  Input smoothing constant, r : .5
  Input number of variates, np: 4
  Want the ARL for a fixed value of h4? y/n: n
  Input desired ARL,          : 200

   iteration     ARL          h4
        1     950.0030      18.2800     r =   0.500
        2     394.0500      16.2800
        3     291.1130      15.5819
        4     223.1652      14.9640
        5     203.9341      14.7534
        6     200.2156      14.7103
        7     200.0022      14.7078
        8     200.0000      14.7078
        9     200.0000      14.7078
       10     200.0000      14.7078
-----------------------------------------------------------------
```

MEWMA

```
         ****************************************************
         *      Multivariate EWMA Control Chart             *
         ****************************************************

You can enter a valid filespec, as long as it has an extension
If you  press the enter key ALL file names are displayed.
Enter FILESPEC: F10 to return to the menu.
? mewma.dat

  Type Lambda or press Enter to use the program's estimate: .3
  Enter target for series 1 or press Enter for the mean: 0
  Enter target for series 2 or press Enter for the mean: 0

   DATA SERIES                  EWMA
     1         2           1         2
 -1.190     0.590      -0.357     0.177
  0.120     0.900      -0.214     0.394
 -1.690     0.400      -0.657     0.396
  0.300     0.460      -0.370     0.415
  0.890    -0.750       0.008     0.066
  0.820     0.980       0.252     0.340
 -0.300     2.280       0.086     0.922
  0.630     1.750       0.249     1.170
  1.560     1.580       0.643     1.293
  1.460     3.050       0.888     1.820

 VEC  MSE      Lamda
 1   1.108     0.300
 2   1.197     0.300

     COVARIANCE MATRIX OF THE INPUT DATA
  1.135     0.380
  0.380     1.164


   Enter h4 or press Enter for a 99% Chisquare limit: 10.81

OBSERVATIONS and MEWMA CONTROL STATISTIC (last column)
           1          2
    1 -1.190D+00  5.900D-01    2.1886
    2  1.200D-01  9.000D-01    1.8581
    3 -1.690D+00  4.000D-01    4.7849
    4  3.000D-01  4.600D-01    2.4063
    5  8.900D-01 -7.500D-01    0.0225
    6  8.200D-01  9.800D-01    0.6828
    7 -3.000D-01  2.280D+00    4.4242
```

```
    8  6.300D-01  1.750D+00    6.7870
    9  1.560D+00  1.580D+00    8.4266
   10  1.460D+00  3.050D+00   16.6240

 MEAN  2.600D-01  1.124D+00

 The UCL =    10.810
```

```
*********************************************************************
                  MEWMA Control Statistics
*********************************************************************

For Data File mewma.dat

 MAX             MIN             MEAN          STD.DEV.          N
16.6240          0.0225          4.8205          4.9219         10
```



p= 2 n= 10 The broken line is the UCL at alpha = 1.000

References

1. Bodden, Kevin and Rigdon, Steven, 1999, "A Program for Approximating the In-Control ARL for the MEWMA Chart", *Journal of Quality Technology*, 31

2. Champ, Charles and Rigdon, Steven, 1991, "A Comparison of the Markov Chain and the Integral Equation Approaches for Evaluating the Run Length Distribution of Quality Control Charts", *Communications in Statistical Simulation*, 20

3. Crowder, Stephen, 1987, "Average Run Lengths of Exponentially Weighted Moving Average Control Charts", *Journal of Quality Technology*, 19

4. Crowder, Stephen, 1989, "Design of Exponentially Weighted Moving Average Schemes", *Journal of Quality Technology*, 21

5. Harris, Thomas and Ross, William, 1991, "Statistical Process Control for Correlated Observations", *The Canadian Journal of Chemical Engineering*, 69.

6. Hawkins, Douglas, 1991, "Multivariate Quality Control Based on Regression-Adjusted Variables", *Technometrics*, 33

7. Hunter, J. Stuart, 1986, "The Exponentially Weighted Moving Average", *Journal of Quality Technology*, 18

8. Kramer, H.G. and Schmid, W., 1997, "EWMA Charts for Multivariate Time Series", *Sequential Analysis*, 16

9. Michelson, Diane, 1994, "Statistical Process Control for Correlated Data", unpublished Ph.D. dissertation, Department of Statistics, Texas A&M University, College Station, TX

10. Lowry, Cynthia and Montgomery, Douglas, 1995, "A Review of Multivariate Control Charts", HE Transactions, 27.

11. Lowry, Cynthia; Woodall, William; Champ, Charles and Rigdon, Steven, 1992, "A Multivariate Exponentially Weighted Moving Average Control Chart", *Technometrics*, 34.

12. Lucas, James and Saccucci, Michael, 1990, "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements", *Technometrics*, 32

13. Montgomery, Douglas and Mastrangelo, Christina, 1991, "Some Statistical Process Control Methods for Autocorrelated Data", *Journal of Quality Technology*, 23

14. Prabhu, Sharad and Runger, George, 1997, “Designing a Multivariate EWMA Control Chart”, *Journal of Quality Technology*, 29

15. Prins, Jack and Mader, Doug, 1997, “Multivariate Control Charts for Grouped and Individual Observations”, *Quality Engineering*, 10

16. Rigdon, Steven, 1994, “A Double-Integral Equation for the Average Run Length of a Multivariate Exponentially Weighted Moving Average Control Chart”, *Statistics and Probability Letters*, 24

17. Runger, George and Prabhu, Sharad, 1996, “A Markov Chain Model for the Multivariate Exponentially Weighted Moving Averages Control Chart”, *Journal of the American Statistical Association*, 91

18. Saccucci, Michael and Lucas, James, 1990, “Average Run Lengths for Exponentially Weighted Moving Average Control Schemes Using the Markov Chain Approach”, *Journal of Quality Technology*, 22

19. Scranton, Richard; Runger, George; Keats, J. Bert and Montgomery, Douglas, 1996, “Efficient Shift Detection Using Multivariate Exponentially Weighted Moving Average Control Charts and Principal Components”, *Quality and Reliability Engineering International*, 12

20. Wold, Svante, 1994, “Exponentially Weighted Moving Principal Components Analysis and Projection to Latent Structures”, *Chemometrics and Intelligent Laboratory Systems*, 23

21. Yumin, Liu, 1996, “An Improvement for MEWMA in Multivariate Process Control”, *Computers ind. Engineering*, 31

MULTIVARIATE CONTROL CHARTS FOR GROUPED AND INDIVIDUAL OBSERVATIONS

This software package analyzes and plots multivariate data using the following methods:

Hotelling's T^2 for grouped data

Hotelling's T^2 for individual data

Principal Components

Multivariate EWMA charts

The following data management features are incorporated:

A data input routine. (although any editor, spreadsheet or wordprocessor can prepare the input files)

A program that can display or/and print files

A program that can display or/and print saved plots

The theoretical background material for this package can be found in a paper **"MULTIVARIATE CONTROL CHARTS FOR GROUPED AND INDIVIDUAL OBSERVATIONS" by** J.Prins and D.Mader in Quality Engineering **September 1997 Vol. 10, No. 1, pp. 49-57**

The rest of this mini user's guide consists of computer sessions featuring the high-lights of each program. User's input is in bold-italic.

The software is IBM based and requires a laser printer to output the plots.

MENU

Use up / down arrow keys to position the cursor.

key	MULTIVARIATE CONTROL CHARTS
1	Analyze Data with multivariate EWMA
2	Analyze Grouped Data with Hotelling
3	Analyze Single Data with Hotelling T-sqr
4	Analyze Data with multivariate CUSUM
5	Analyze Data with Principal Components
6	Computation of Covariance Matrix
7	Data Input Routine
8	Display and Printout of Files
9	Display and Printout of Saved Plots
Esc	EXIT

MULTI

```
**************************************************
*   Multi-Variate Control Chart for Grouped Data  *
*       Using the Hotelling T Square Statistic    *
**************************************************
```

What is the file-id of the data? (Press Enter to exit) ***TEST.DAT***
Input p, the number of variables: ***4***
Input k, the number of subgroups: ***14***
Input n, the number of observations : ***5***
Enter alpha, the level of significance for the UCL**: .01***

	MEANS				HOTELLING T SQR	
					MEAN	DISPERSION
	1	2	3	4		
1	9.968E+00	1.497E+01	4.988E+01	6.004E+01	28.049	21.996
2	9.978E+00	1.498E+01	4.991E+01	6.009E+01	55.161	17.168
3	9.970E+00	1.497E+01	4.991E+01	6.004E+01	6.347	11.129
4	9.966E+00	1.497E+01	4.994E+01	6.006E+01	10.872	14.165
5	9.974E+00	1.498E+01	4.991E+01	6.004E+01	3.192	14.592
6	9.990E+00	1.499E+01	4.989E+01	6.002E+01	2.884	5.595
7	9.990E+00	1.499E+01	4.992E+01	6.002E+01	6.053	3.124
8	9.990E+00	1.500E+01	4.991E+01	6.000E+01	13.379	7.000
9	9.992E+00	1.499E+01	4.991E+01	5.999E+01	33.087	6.778
10	9.998E+00	1.500E+01	4.991E+01	6.000E+01	21.412	24.631
11	1.000E+01	1.500E+01	4.987E+01	5.996E+01	31.692	19.529
12	9.986E+00	1.499E+01	4.990E+01	6.002E+01	0.907	9.821
13	9.992E+00	1.499E+01	4.991E+01	6.005E+01	7.847	23.351
14	9.976E+00	1.497E+01	4.993E+01	6.006E+01	10.380	45.120
GRAND	9.984E+00	1.498E+01	4.991E+01	6.003E+01		

The UCL for T sqr of means is: 14.502
The UCL for T sqr of dispersions is: 23.243

VARIANCE-COVARIANCE MATRIX

```
 1.907E-04  2.089E-04 -9.643E-05 -9.607E-05
 2.089E-04  2.929E-04 -9.750E-05 -1.032E-04
-9.643E-05 -9.750E-05  1.936E-03  1.525E-03
-9.607E-05 -1.032E-04  1.525E-03  1.524E-03
```

INVERSE OF VARIANCE-COVARIANCE MATRIX

```
 2.423E+04 -1.716E+04  2.595E+02  1.054E+02
-1.716E+04  1.565E+04 -2.334E+02  2.120E+02
 2.595E+02 -2.334E+02  2.447E+03 -2.449E+03
 1.054E+02  2.120E+02 -2.449E+03  3.128E+03
```

Wish to display the individual covariance matrices? y/N*: **N***

Enter file-id to save the output or press the Enter key to quit...***MULTI.OUT***

```
**************************************************
*                 Hotelling T Squares            *
**************************************************
```

For Data File TEST.DAT

MAX	MIN	MEAN	STD.DEV.	NO.DATA
55.1611	0.9065	16.5187	15.5160	14

Press Enter to view the plot

Save the plot on disk? y/N: **Y**
The plot will be saved as: TEST.PCX on the default disk.
Press Enter to keep this name or enter a different name:
This can be redisplayed at later stages by typing: SHOW.

PRESS ENTER TO RETURN TO THE PROGRAM, AFTER THE PLOT IS DISPLAYED

SINGLE

```
**************************************************
*          Multi-Variate Control Chart for Individuals          *
*            Using the Hotelling T Square Statisitc             *
**************************************************
```

What is the file-id of the data? (Press Enter to exit) ***TEST.DAT***
What is the file-id of the covariance matrix? ***SINGLE.COV***
Input p, the number of variables: **4**
Input k, the number of individals: **70**

Enter alpha, the level of significance for the UCL: **.01**
OBSERVATIONS and HOTELLING T-SQUARE (last column)

	1	2	3	4	
1	9.960E+00	1.497E+01	4.989E+01	6.002E+01	4.4151
2	9.950E+00	1.494E+01	4.984E+01	6.002E+01	9.7387
3	9.950E+00	1.495E+01	4.985E+01	6.000E+01	7.3756
4	9.990E+00	1.499E+01	4.989E+01	6.006E+01	3.2641
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
67	9.990E+00	1.500E+01	4.991E+01	6.004E+01	1.5722
68	9.990E+00	1.498E+01	4.992E+01	6.004E+01	2.4405
69	1.000E+01	1.500E+01	4.988E+01	6.000E+01	1.2707
70	9.990E+00	1.499E+01	4.995E+01	6.010E+01	3.7590
$\bar{X}$	9.984E+00	1.498E+01	4.991E+01	6.003E+01	

The UCL = 13.308

VARIANCE-COVARIANCE MATRIX

2.755E-04	2.863E-04	-1.444E-04	-3.328E-04
2.863E-04	3.614E-04	-1.379E-04	-3.219E-04
-1.444E-04	-1.379E-04	1.907E-03	1.546E-03
-3.328E-04	-3.219E-04	1.546E-03	2.307E-03

INVERSE OF VARIANCE-COVARIANCE MATRIX

```
 2.210E+04 -1.663E+04 -5.097E+02  1.210E+03
-1.663E+04  1.574E+04  9.475E+01 -2.660E+02
-5.097E+02  9.475E+01  1.186E+03 -8.549E+02
 1.210E+03 -2.660E+02 -8.549E+02  1.144E+03
```

Enter file-id to save the output or press the Enter key to quit.***.SINGLE.OUT***

```
**********************************************
*                Hotelling T Squares                *
**********************************************

                  For Data File TEST.DAT

  MAX         MIN         MEAN        STD.DEV.     NO.DATA
 33.3403     0.2106      3.9408      4.7509          70
```

Press Enter to view the plot

Save the plot on disk? y/N: **N**

Note:
As you may have noticed, the individuals multivariate control chart routine requires the input of a covariance matrix. This can either be generated by pooling all vector observations, or by constructing a matrix that consists of successive differences between two observations. The program COVAR outputs the covariance using method 1 (pooling) and COVARDIF outputs the covariance matrix using method 2(differences). The menu will ask you which method you wish and selects the appropriate program accordingly

PRINCO

```
**************************************************
*                Principal Components Analysis                *
**************************************************

 What is the file-id of the data? (Press Enter to exit) test.dat
 Input N, the number of rows   :   70
 Input M, the number of columns :  4

 EIGENVALUES     PROPORTION
  1   2.408          0.602
  2   1.266          0.919
  3   0.237          0.978
  4   0.089          1.000

 How many factors to retain? (0 to quit): 2

CORRELATION MATRIX OF THE VARIABLES

  1.000  0.907 -0.199 -0.417
  0.907  1.000 -0.166 -0.353
 -0.199 -0.166  1.000  0.737
 -0.417 -0.353  0.737  1.000


 FACTOR STRUCTURE
  0.850  0.481
  0.819  0.529
 -0.629  0.708
 -0.786  0.505


 FINAL COMMUNALITY
   1   0.954
   2   0.951
   3   0.896
   4   0.873
```

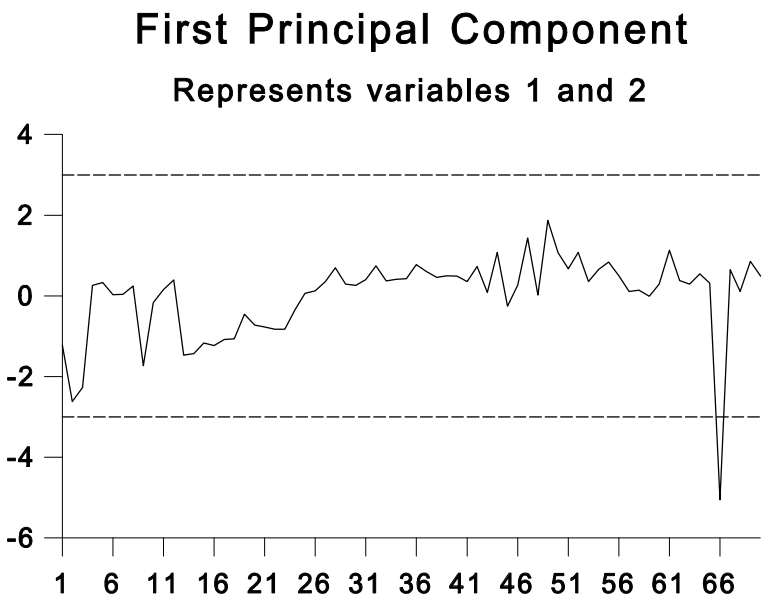
COEFFICIENT MATRIX

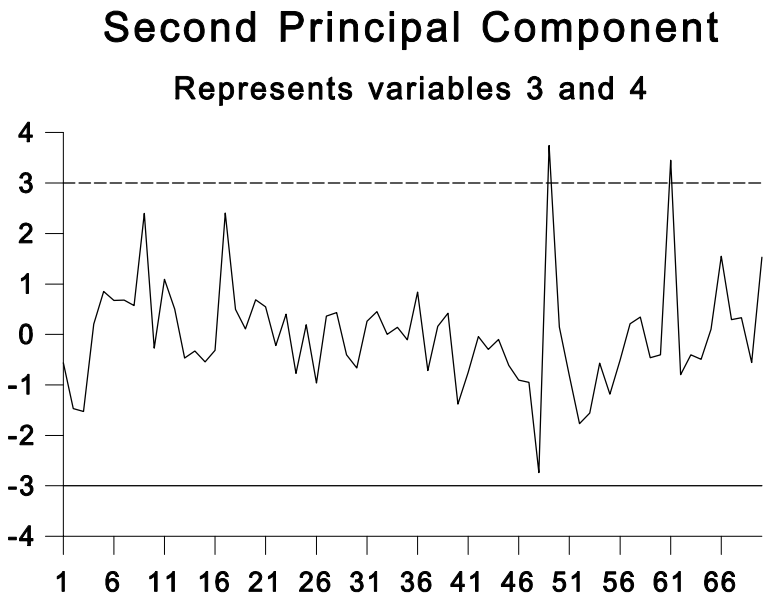
0.353	0.380
0.340	0.418
-0.261	0.559
-0.327	0.399

FACTOR SCORES

1	-0.606	-1.161
2	-1.060	-2.700
3	-0.803	-2.517
4	0.123	0.307
.	.	.
.	.	.
.	.	.
68	-0.101	0.305
69	0.989	0.128
70	-0.512	1.414

Enter filename to store output, or press Enter to quit: ***PRINCO.OUT***


First Principal Component
Represents variables 1 and 2
4
2
0
-2
-4
-6
1 6 11 16 21 26 31 36 41 46 51 56 61 66


Second Principal Component
Represents variables 3 and 4
4
3
2
1
0
-1
-2
-3
-4
1 6 11 16 21 26 31 36 41 46 51 56 61 66

COVAR

```
**************************************************
*          Computes Covariance Matrix and its Inverse, and       *
*           the Mean of an  n rows by p  columns Data File        *
**************************************************
```

What is the file-id of the data? ***TEST.DAT***
Input p, the number of variables: ***4***
Input n, the number of observations: ***70***

MEANS

9.984E+00 1.498E+01 4.991E+01 6.003E+01

VARIANCE-COVARIANCE MATRIX

```
 2.755E-04  2.863E-04 -1.444E-04 -3.328E-04
 2.863E-04  3.614E-04 -1.379E-04 -3.219E-04
-1.444E-04 -1.379E-04  1.907E-03  1.546E-03
-3.328E-04 -3.219E-04  1.546E-03  2.307E-03
```

INVERSE OF VARIANCE-COVARIANCE MATRIX

```
 2.214E+04 -1.666E+04 -5.098E+02  1.211E+03
-1.666E+04  1.576E+04  9.462E+01 -2.668E+02
-5.098E+02  9.462E+01  1.186E+03 -8.551E+02
 1.211E+03 -2.668E+02 -8.551E+02  1.144E+03
```

Enter file-id to save the output or press the Enter key to quit. ***SINGLE.OUT***

INPUT

This program forms data files
How many rows (enter 0 to exit): ? ***3***
How many columns: ? ***2***

Start inputting. **Press the enter key after each entry...**
1 : ***1*** ***4***
2 : ***5*** ***3***
3 : ***7*** ***8***

Wish to see the input file (for eventual corrections)? y/N: ***N***

Enter a file name or press the ENTER key () to name it DATA.FIL:
MY.DAT
The data are stored in file MY.DAT
Press the ENTER key () to return.

Multivariate Cusum Control Charts

The software is based on a paper by Crosier in Technometrics, August 1988 , Vol. 30, no3 , *"Multivariate Generalizations of Cumulative Sum Control Schemes*".
He suggested 2 methods, the first one is to compute the Hotelling T^2 statistic and then do a Cusum on its square root. The second method is the replacement of the scalar results of a univariate Cusum by vectors.
The following program session uses the first method, which Crosier named COT, (cusum of T).

The input to the program is a *p* column file, where *p* is the number of variables. A row may be a single vector observation or part of a group of *n* observations. The program will prompt for the number of groups and determines from this whether we deal with single or group observations.

The best way to illustrate the MCUSUM program is by way of an example. The data came from Crosier's paper.

MCUSUM

```
      ****************************************
      * Multi-Variate Control Cusum Chart using the     *
      * square root of the Hotelling T Square Statistic     *
      ****************************************

 You can enter a valid filespec, as long as it has an extension, or you
 can select a file extension to search for files of particular interest.
 If you  press the enter key  ALL file names are displayed.
 Enter FILESPEC or EXTENSION (1-3 letters): F10 to quit.
? crosier.dat

 The input file consists of  the columns labeled X1 and X2
 (on the next page)


 Input k, the number of subgroups:    10
 Enter the name of the covariance matrix or press Enter
 to have the program compute it from the data:  crosier.cov
```

The covariance matrix from the Crosier (88) paper is

1	.5
.5	1

```
 sample   X1      X2     T2    SQRT(T2)
  1     -1.19    .59    3.288   1.813
  2       .12    .90    0.955   0.977
  3     -1.69    .40    4.923   2.219
  4       .30    .46    0.218   0.467
  5       .89   -.75    2.696   1.642
  6       .82    .98    1.106   1.051
  7      -.30   2.28    7.963   2.822
  8       .63   1.75    3.143   1.773
  9      1.56   1.58    3.287   1.813
 10      1.46   3.05    9.308   3.051


     MAX            MIN            MEAN          STD.DEV.     NO.DATA
      3.0509         0.4670         1.7629         0.8035        10


 Enter Target or press Enter for the Mean       : 0
 Enter  Process Std.dev or press Enter for the program's s(xbar) : 1

 The size of the shift you wish to detect is given in standard deviations.
 Enter size of shift, (default = 1 std.dev): 1
 Enter alpha risk,    (default = .00135)   :
 Enter value for h (default =  5): 4.04
 Enter value for k (default = .5): 1.41

 Tabular Output? y/N: y
  U0 (TARGET)      H       K (H/d)
     0.0000     4.0400     1.4100

                  Increase in mean
SAMPLE    T      T-U0  T-U0-K    S hi
-------------------------------------
   1   1.813   1.813   0.403   0.403
   2   0.977   0.977  -0.433   0.000
   3   2.219   2.219   0.809   0.809
   4   0.467   0.467  -0.943   0.000
   5   1.642   1.642   0.232   0.232
   6   1.051   1.051  -0.359   0.000
   7   2.822   2.822   1.412   1.412
   8   1.773   1.773   0.363   1.775
   9   1.813   1.813   0.403   2.178
  10   3.051   3.051   1.641   3.819
```

The out of control criteria is 4.04 applied to S hi.
In this case there was no out of control situation

Method 2:
Replacement of the scalar quantities of a univariate CUSUM scheme by vectors

Crosier suggested a multivariate CUSUM of the following form:

$$Y_n = \left(\mathbf{S_n} \sum^{-1} \mathbf{S_n} \right)^{1/2}$$

where

$$\mathrm{S_n} = \begin{cases} 0 & \text{if } C_n \le k \\ (\mathbf{S_{n-1}} + \mathbf{X_n} - \hat{\mu})(1 - k / C_n) & \text{if } C_n > k \end{cases}$$

and

$$C_n = \left[(\mathbf{S_{n-1}} + \mathbf{X_n} - \hat{\mu}) \sum^{-1} (\mathbf{S_{n-1}} + \mathbf{X_n} - \hat{\mu})' \right]^{1/2}$$

A recommended choice for k is .5

An example, using the same data as in the previous example now follows:

VCUSUM

```
       ************************************************
       * Multi-Variate Control Cusum Chart using the     *
       * vector approach suggested by Crosier (1988)     *
       ************************************************

 You can enter a valid filespec, as long as it has an extension, or you
 can select a file extension to search for files of particular interest.
 If you  press the enter key, ALL file names are displayed.
 Enter FILESPEC or EXTENSION (1-3 letters): F10 to quit.
? crosier.dat


 Enter value for h (default =  5): 5.5
 Enter value for k (default = .5): .5

 Input the number of subgroups:    10

 Enter the name of the covariance matrix or press Enter
 to have the program compute it from the data:  crosier.cov

 Enter target for series 1 or press Enter for the mean: 0
 Enter target for series 2 or press Enter for the mean: 0
```

OUTPUT

```
          S VEC                                                Multivariate Cusum
         1          2
  1 -8.619D-01  4.273D-01                                            1.313
  2 -5.650D-01  1.011D+00                                            1.597
  3 -1.950D+00  1.220D+00                                            3.198
  4 -1.402D+00  1.428D+00                                            2.830
  5 -2.978D-01  3.939D-01                                            0.694
  6  3.340D-01  8.787D-01                                            0.887
  7  2.929D-02  2.723D+00                                            3.128
  8  5.910D-01  4.010D+00                                            4.330
  9  1.960D+00  5.095D+00                                            5.140
 10  3.211D+00  7.647D+00                                            7.679 *

 H =   5.50  K =   0.50

  Enter file-id to save the output or press Enter key to quit...
```

The ARL approximation by Siegmund (1985) for a one-sided cusum

$$ARL = \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2} \qquad \text{for } \Delta \neq 0$$

where $\Delta = \delta^* - k$ for the upper one-sided cusum
and $\Delta = -\delta^* - k$ for the lower one-sided cusum
and δ^* = the shift of the mean in terms of σ
$b = h + 1.166$

if $\Delta = 0$ then set $ARL = b^2$
if $\delta^* = 0$ then ARL_0 is calculated from the above equation
To obtain the ARL of the two-sided cusum , compute the ARL for both sides, call them ARL^+ and ARL^- and then use

$$1/ARL = 1/ARL^+ + 1/ARL^-$$

Example: (from Montgomery, 1996, page 324)

Consider a two-sided cusum with $k = .5$ and $h = 5$. We wish to find ARL_0
We use Siegmund's equation for the upper side:
Since $\delta^* = 0$, $\Delta = \delta^* - k = -.5$. $b = 1+h = 5+1.166 = 6.166$.
Then $ARL_0 = \exp[-2(-.5)(6.166) - 2(-.5)(6.166) - 1] / 2(-.5)^2 = 9382$.
By symmetry the lower side has the same ARL_0 and we obtain

1/ARL = 1/938.2 + 1/938.2
ARL = 469.1 (The ARL computed using Markov Chains = 465)

Here is a computer run, using the Siegmund approximation:

SIEGMUND

```
**********************************************************
*   Siegmund's approximation for the ARL of the CUSUM    *
*   Good for shifts up to 2 sigma at k = .5 and h = 4 or 5  *
**********************************************************

Enter value for h: 5
Enter value for k: .5
Enter shift in the mean (in terms of units of sigma): 2

The ARL for the upper one sided cusum is:      3.89
The ARL for the two sided cusum is:            3.89

The contribution of the lower one sided cusum is negligible.
```

SHOW

DISPLAY AND PRINT-OUT OF SAVED PLOTS

Want a screen dump to a graphics printer ? y/N *:* ***N***
The above files reside on the disk:

TEST1.PCX TEST.PCX

Use arrow keys to move cursor.
Press the enter key (↵) to select.
Press Esc to exit.
Type $ to search another disk or directory.

SHOWFILS

Displays filenames in a selected sub-directory.

The above * files reside on the *.* disk: (default disk)

SINGLE.COV SINGLE2.COV TEST.DAT TEST2.DAT
SINGLE.FIL PRINC.OUT SINGLE.OUT SINGLE1.T2
SINGLE2.T2

Use arrow keys to move cursor. Press the enter key (↵) to select.
Press Esc to exit. Type $ to search another disk or directory.

The cursor was moved to TEST.DAT and the Enter key was pressed.

TEST.DAT

(S)creen or (P)rinter? S

The test.dat file was printed to the screen…

More? y/N: ***n***

Chapter 6

Time Series Analysis

SLCT	TIMESTAT Statistical Timeseries Analysis		Help
1	BJID	Box-Jenkins identification routines	F1
2	BJ	Box-Jenkins estimation/diagnostics/forecasting	F2
3	STAR	Stepwise Autoregression (Box Jenkins AR model)	F3
4	MVAR	Multivariate Auto Regression	F4
5	HW	Exponential Smoothing by Holt-Winters	F5
6	LR	Single and multiple linear regression	F6
7	ROWINPUT	Input routine	F7
8	ROWEDIT	Edit routine	F8
9	DISPLAY	Displays graphs that are stored on disk	F9
10	PLOT	Plot of raw data (to 8 series)	F10
		Box-Jenkins Tutorial and Data Input Menus. 0 or ⏎ to Exit	

The Box-Jenkins Methods

The Box and Jenkins methods are explained in their book, 'TIME SERIES ANALYSIS, Forecasting and Control', Holden-Day 1976. It would be an insult to try to condense their work in a few lines of code, instead the user is invited to buy this excellent book! The main approach is to fit stochastic models to stationary series as follows:

IDENTIFICATION Here stationarity is investigated by means of the auto correlation function and a tentative assessment of the order of the model is made.

ESTIMATION This produces estimators for the model parameters. For ARMA processes non linear estimation is used, according to the Marquardt algorithm. For AR models regular linear estimation is employed.

DIAGNOSTICS This is an examination of the residuals for randomness and periodicity. Also the standard errors of estimators are studied to eliminate over estimation.

FOREASTING If all is well the model equation can be used for minimum squared error forecasting, with confidence bands that are a function of the residual variance.

The model for all time series is given by:

$$z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + ...\varphi_p z_{t-p} + a_t$$
$$-\theta_1 a_{t-1} - \theta_2 a_{t-2} - ...\theta_q a_{t-q}$$

where: z is the observation at time 't' minus its mean. a is a random variable from a normal distribution with mean = 0 and variance = a finite value. The φ's are weights or coefficients that multiply the z's and the θ's are weights that multiply the a's (or 'shocks).

The ojective of the analysis is to estimate the φ's and θ's and the variance of the a's, σ. The estimates for the a's turn out to be the residuals.

When there are no θ terms in the model, we get this form:

$$z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + ...\varphi_p z_{t-p} + a_t$$

This is known as the *Auto Regressive* (AR) model of order 'p'.

And when the φ terms are absent, we obtain:

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ...\theta_q a_{t-q}$$

This is known as the *Moving Average* (MA) model of order 'q'.

Stepwise Auto Regression

There are many techniques to analyze time series. To mention a few: Exponential Smoothing, by Holt and Winters. Exponential smoothing by Brown. The Box-Jenkins methods. One of these Box-Jenkins methods is the *autoregressive* (AR) model. The form is :

$$Z_t - \mu = C + \varphi_1 z_{t-1} + \varphi_2 Z_{t-2} ...+ \varphi_n Z_{t-n} + a_t$$

Where
C is a constant, μ is the mean, Z_t is the observation at time t, and a_t is an error term at time t.

The φ parameters are estimated by the program, using stepwise regression. The program uses the 'F test' to determine if an added φ is necessary.

The program performs the Box-Pierce and Box Ljung Chi-Square test to find out if the data should not have been de-seasonalized prior to analysis. The program can detrend your data, using the 'trend-seasonal' method. The forecasts and their confidence limits are then 're-trended'.

Triple Exponential Smoothing by Holt and Winters

The idea here is to fit a model by using 3 updating equations, which need 3 weights or smoothing constants. The equations are:

1. $sa_t = \alpha s_t / f_{t-p} + (1-\alpha) sa_{t-1} + r_{t-1}$
2. $f_t = ß s_t / sa_t + (1-ß) f_{t-p}$
3. $rt = \gamma(sa_t - sa_{t-1}) + (1-\gamma) r_{t-1}$

t is the current time period
p is the period of seasonality (e.g 12 for monthly)
s_t is the latest observation
sa_t is the current mean in time period t
f_t is the estimated seasonal factor for period t
r_t is the estimated trend term in period t
$\alpha, ß, \gamma$ are weights. Their magnitude range from 0 to 1.

These weights are computed by the program using a nonlinear least squares method, known as 'the Marquardt Algorithm'.

The forecasts from time t are calculated by:

$$s_t,h = (sa_t + hr_t) f_{t-p+h} \qquad h=1,2,...p$$

Single Exponential Smoothing

You could use equation 1 only, as follows:

$$sa_t = \alpha s_t + (1-\alpha) sa_{t-1}$$

This is called SINGLE exponential smoothing.

Double Exponential Smoothing

You could use equations 1 and 2 or 1 and 3. This is DOUBLE exponential smoothing. Using equations 1 and 2 is for demand and seasonal variations, and equations 1 and 3 is for demand trend effects. In the seasonal case, the quantity r_{t-1} in equation 1 is set to 0, since there is no trend assumed.

The seasonal variation in the above set of equations is known as 'multiplicative'. There is also an 'additive' seasonal variation. To account for the additive seasonal variation, the updating equations are then:

1. $sa_t = \alpha(s_t - f_{t-p}) + (1-\alpha) sa_{t-1} + r_{t-1}$
2. $f_t = ß(s_t - sa_t) + (1-ß) f_{t-p}$
3. $r_t = \gamma(sa_t - sa_{t-1}) + (1-\gamma) r_{t-1}$

Program Example of the HOLT-WINTERS' TRIPLE EXPONENTIAL SMOOTHING

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you press the ENTER key (↵), ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): To return to DOS, press F10. ? *bookg.bj*

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized. In other prompts, the default is 0 (zero) unless indicated otherwise

MAX	MIN	MEAN	VARIANCE	NO.DATA
622.0000	104.0000	280.2986	14391.9229	144

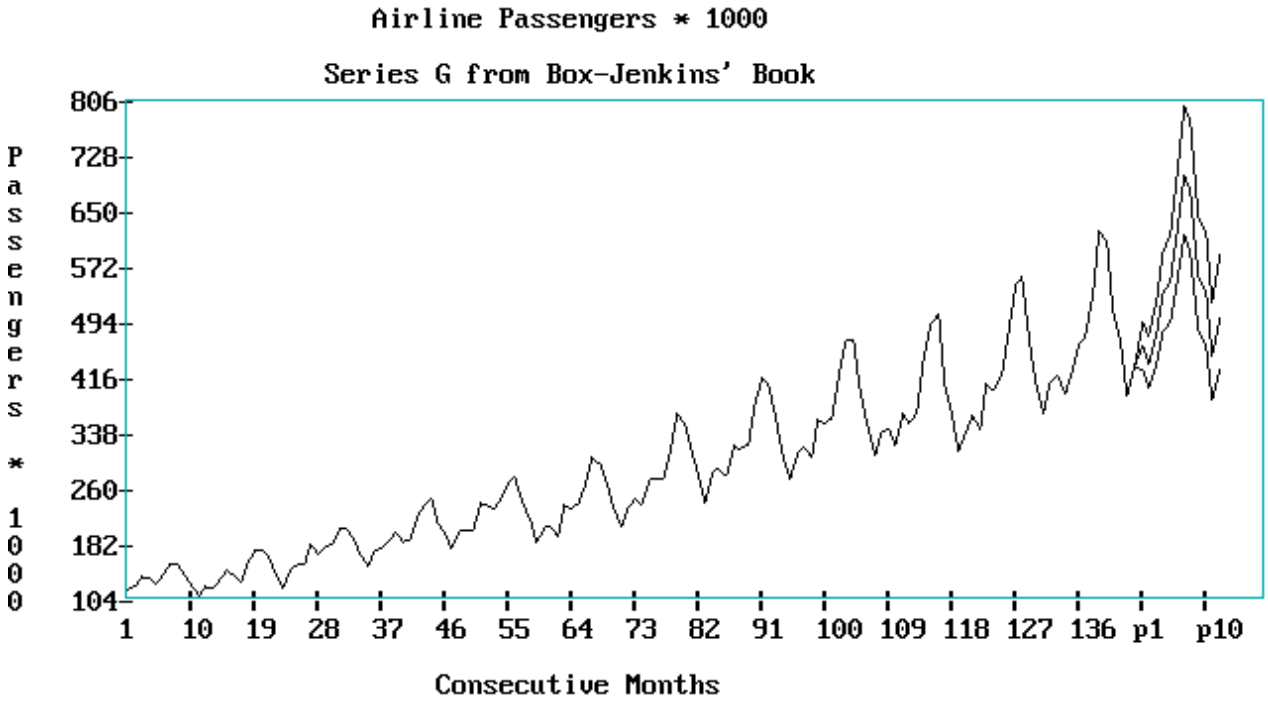
:

```
Enter M for manual or press the Enter key for automatic fitting:


 ENTERING THE MARQUARDT NON LINEAR FITTING ROUTINE

                   mean  seasonality trend
Iteration RES.VAR   ALPHA    BETA     GAMMA
 0        236.079  0.30000  0.30000  0.30000
 1        153.454  0.32679  0.43836  0.00345
 2        118.978  0.35861  0.72755  0.02346
 3        115.427  0.34204  0.89529  0.00888
 4        114.023  0.33923  0.89025  0.02698
 5        115.487  0.33923  0.89025  0.02698
 6        115.412  0.33923  0.89025  0.02698
 7        114.493  0.33923  0.89025  0.02698
 8        113.947  0.33935  0.89171  0.02445

  Seasonal period =  12

    RESIDUAL              UPDATING COEFFICIENTS
    VARIANCE          mean       seasonality     trend
    113.94687       0.33935        0.89171       0.02445

 Original Variance           :    14391.91699
 Residual Variance           :      113.94687
 Coefficient of Determination:       99.20826
```

Would you like to plot the forecasts? Y/n : y
.

The plot consists of the last 'n' data points, followed by the forecasts.

```
The next 12 forecasts from observation 144

        145      446.720
        146      419.750
        147      466.696
        148      495.688
        149      507.718
        150      576.545
        151      667.290
        152      659.219
        153      550.499
        154      492.148
        155      419.834
        156      464.934
```



Linear Regression applied to Time Series Analysis
The program LR is dual functioned. It can be used in the same form as the Box-Jenkins and Exponential smoothing programs, or it can be applied to any garden variety linear regression analysis. The time series approach is invoked by answering the prompt: ' Number of independent variables? ' with 0 (or just pressing the ENTER key). LR will generate a fake X variable, consisting of the integers 1,2,...n , n = the number of observations.

If the prompt is answered with 1,2, ... p, you must have the input prepared with the corresponding number of independent variables. The file format is similar to the BJID, BJ, STAR and HW format, e.i each line or row consists of an optional 4 digit identifier, followed by the data, but here you m ust continue on the SAME line with the X or independent variable(s). The format is not too rigid, there must be at least one blank between the fields, and the data must be LEFT justified. The maximum number of X variables is 16. This includes polynomial and/or crossterms. The maximum number of data points is 500. The regression models generated by LR are of form:

$$Y = B_0 + B_1X_1 + B_2X_2 + ...B_pX_p$$

Polynomial terms are added: $B_{11}X_1^2 + B_{22}X_2^2 + ...B_{pp}X_p^2$. Similarly for higher order terms. Crossterms are added last: $B_{12}X_1X_2 + ...$

Plotting
There exist an old Chinese proverb:
'One graph is worth at least 1000 words...'

By plotting the original time series you may swiftly observe existence of trends, outliers, typos, odd patterns, and other phenomena. The input to the PLOT routine is described in the third menu, titled: 'Input and Output Menu '.

It is the same input file used for all the programs in TIMESTAT. Briefly, each line of the file consists of an optional 4 digit identifier, followed by at least one blank, followed by the observation. A handy way to form the file is using ***ROWINPUT***, described in the I/O menu. ROWINPUT inputs the data conversationally, line by line and lets you name the resulting file.

TIMESTAT Box Jenkins Mini Tutorial	
1	Input of data and output or results
2	Differencing and stationarity.
3	Auto correlation function. (The ACF).
4	Partial correlation function (The PACF).
5	Seasonality.
6	How many terms should one use?
7	Sample problem 1, Series F of Box-Jenkins, model (2,0,0)
8	Sample problem 2, Series G, (1 0 1)x(1 0 1) at period 12
0	Return to the main menu. Esc: Return to DOS

Tutorials and Examples
The Box and Jenkins approach is essentially to find a probability model for times series. These models are sometimes referred to as stochastic processes. In particular the class of these processes that are analysable are those that are stationary.

Stationarity can be defined in mathematical terms, but for our purpose we mean a flat looking series. That is there is no trend, no change in variance, and no periodic variations. Like a smooth flowing river. Of course in practice we do not start out with such a series. But with regular and seasonal differencing we go a long way. In addition some 'well chosen' transformation often helps to reduce or eliminate changes in variance over time. We then fit a model to the (hopefully) stationary series. Later on, at the forecasting stage, the programs automatically account for these actions and compute the forecasts in the original mode.

We know from the general model equation that the z(t) are not the original observations, but that they are the mean-corrected observations. In general when the series are differenced at least once, the mean is equal to zero. But when we do not have to difference, we subtract the mean from the observations, perform the analysis, and later on, during forecasting, add this mean back in.

The sample autocorrelation function (acf) plays an important role in the identification stage of the Box-Jenkins procedure. What does it do? It computes the correlation between a given set of

observations that were measured over time, and the SAME set, but lagged back one or more time periods. The FUNCTION is the set of correlations at lags 1,2,...k.

The shape of this function provides insight into the model that generated the data. Each model has a unique theoretical acf. The following interpretations may be of help:

SHAPE	MODEL
Exponential decaying to zero	AR (autoregressive)
Alternating, decaying to zero	AR
One or more 'spikes', rest 0	MA (moving average)
Decay starting after a few lags	Mixed ARMA
All zero, or close to zero	Random
High values at fixed intervals	Seasonality
NO decaying to zero at all	Non-Stationarity, take the next order of difference.

Despite its imposing name, the PACF is a simple concept. It is computed as follows: An autoregressive (AR) model of order 1 is run. Then an AR of order 2 is run. At this point the PACF is the AR(1) and the SECOND (or last) coefficient of the AR(2) model. The process is repeated for succesive higher order AR models. The PACF will consist of the last coefficient of each AR(n) model, for n=1 to p. The BJID program uses p=10.

The 'partial' autocorrelation function is sort of a mirror image of the autocorrelation function. The shape of the PACF for AR processes displays high values (or 'spikes') for the number of AR terms, and then drops down to zero. For example, an AR process (or model) of order 2 has an PACF with its first two terms much higher in value than the rest.

You are not obligated to calculate this function. You could use the program STAR, (Stepwise Autoregression), which determines the proper order of the model, without having to generate the same order AR model as the BJID program. Try both methods and compare.

If you observe very large correlations at lags spaced n periods apart, for example at lags 12 and 24, then there is evidence of periodicity. That effect should be removed, since the objective of the identification stage is to reduce the correlations throughout. So if simple differencing was not enough, try seasonal differencing at a selected period. In the above case, the period is 12. It could of course be any value, such as 4 or 6.

The number of seasonal terms is rarely more than 1. If you know the shape of your forecast function, or you wish to assign a particular shape to the forecast function, you can select the appropriate number of terms for seasonal AR or seasonal MA models.

The book by Box and Jenkins, Time Series Analysis Forecasts and Control has a discussion on these forecast functions on pages 326 - 328. Again, if you have only a faint notion, but you do know that there was a trend upwards before differencing, pick a seasonal MA term and see what comes out in the diagnostics.

Using Seasonal Indices
Another technique to adjust a time series for seasonality is to compute seasonal indices and divide them into the time series.

$$X_i = Z_i/P_s$$

where P is the period of seasonality, Z is the time series, and X is the seasonally adjusted series.

One method used is the *trend-adjusted relative percent* approach. This first removes the trend via a regression line, and then computes the indices as follows:

Let p= 12. Then the expected percentage of a full year of data, for a given month = 1/12, or 8.333 percent. The actual percent for each month is divided by 8.333 to yield the monthly seasonal index. If several years are available, the corresponding monthly indices are averaged. The forecasts based on the deseasonalized series must be multiplied by the indices. This is automatically done by the forecasting portion. The seasonal index scheme is used by the autoregression routines.

Order of Model
What is the order of the model you propose? That is, how many AR or MA terms should the program estimate? Box and Jenkins stress the principal of parsimony. That is, as few terms as possible. The inspection of the ACF and its plot may have given you a clue. In addition the PACF plot would have told you how many AR terms should be in the model.
If you are still not sure, assign 2 AR and 2 MA terms. Later on during the diagnostic section, redundancies will be shown up. Then rerun the program.

Example 1

BOX-JENKINS ARIMA MODEL IDENTIFICATION SECTION

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you press the enter key, ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): To return to DOS, press F10

? *bookf.bj*

MAX	MIN	MEAN	VARIANCE	NO.DATA
80.0000	23.0000	51.7086	141.8238	70

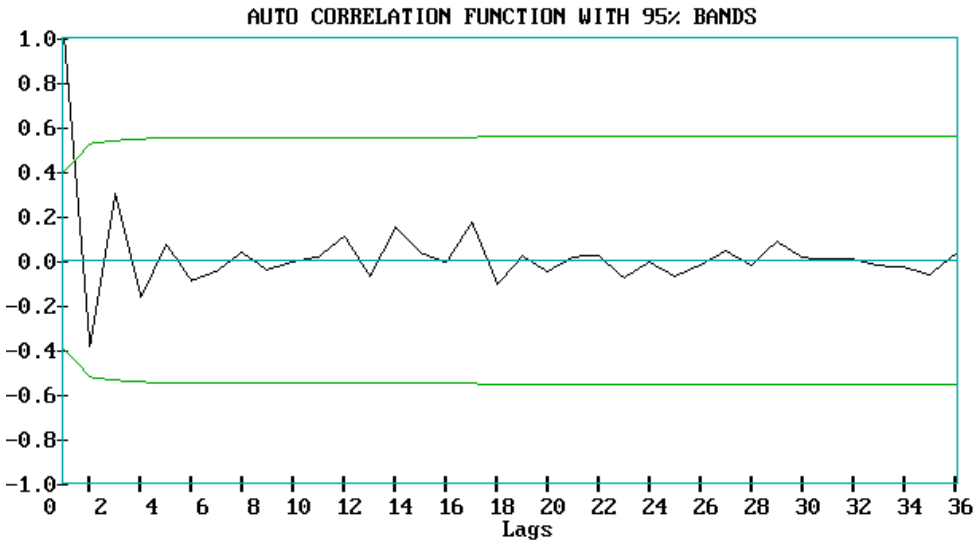
Do you wish to make transformations? y/n n
Input order of difference or 0: 0
Input period of seasonality (2-12) or 0: 0

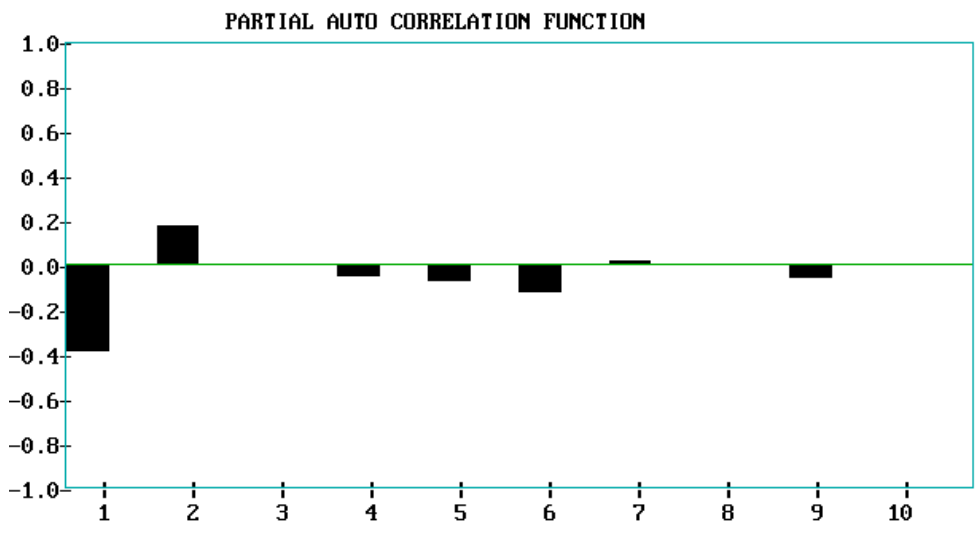
These are defaults you can get them by merely pressing the ENTER key.

Time Series: bookf.bj. Regular difference: 0 Seasonal Difference: 0

Auto Correlation Function for the first 35 lags

0	1.0000	12	-0.0688	24	-0.0731
1	-0.3899	13	0.1480	25	-0.0195
2	0.3044	14	0.0358	26	0.0415
3	-0.1656	15	-0.0067	27	-0.0221
4	0.0707	16	0.1730	28	0.0889
5	-0.0970	17	-0.7013	29	0.0162
6	-0.0471	18	0.0200	30	0.0039
7	0.0354	19	-0.0473	31	0.0046
8	-0.0435	20	0.0161	32	-0.0248
9	-0.0048	21	0.0223	33	-0.0259
10	0.0144	22	-0.0787	34	-0.0629
11	0.1099	23	-0.0096	35	0.0261





BOX-JENKINS ARIMA MODEL

Enter FILESPEC or EXTENSION (1-3 letters): To return to DOS, press F10.
? *bookf.bj*

MAX	MIN	MEAN	VARIANCE	NO.DATA
80.0000	23.0000	51.7086	141.8238	70

Do you wish to make transformations? y/n n
Input order of difference or 0: 0
Input NUMBER of AR terms: 2
Input NUMBER of MA terms: 0
Input period of seasonality (2-12) or 0: 0

Proceed directly to Forecasting? y(es) / h(elp) or press the ENTER key:

********** OUTPUT SECTION **********
AR estimates with Standard Errors
Phi 1 : -0.3397 0.1224
Phi 2 : 0.1904 0.1223

Original Variance : 141.8238
Residual Variance : 110.8236
Coefficient of Determination: 21.8582

***** Test on randomness of Residuals *****
The Chi-Square value = 11.7034
with degrees of freedom = 23
The 95th percentile = 35.16596

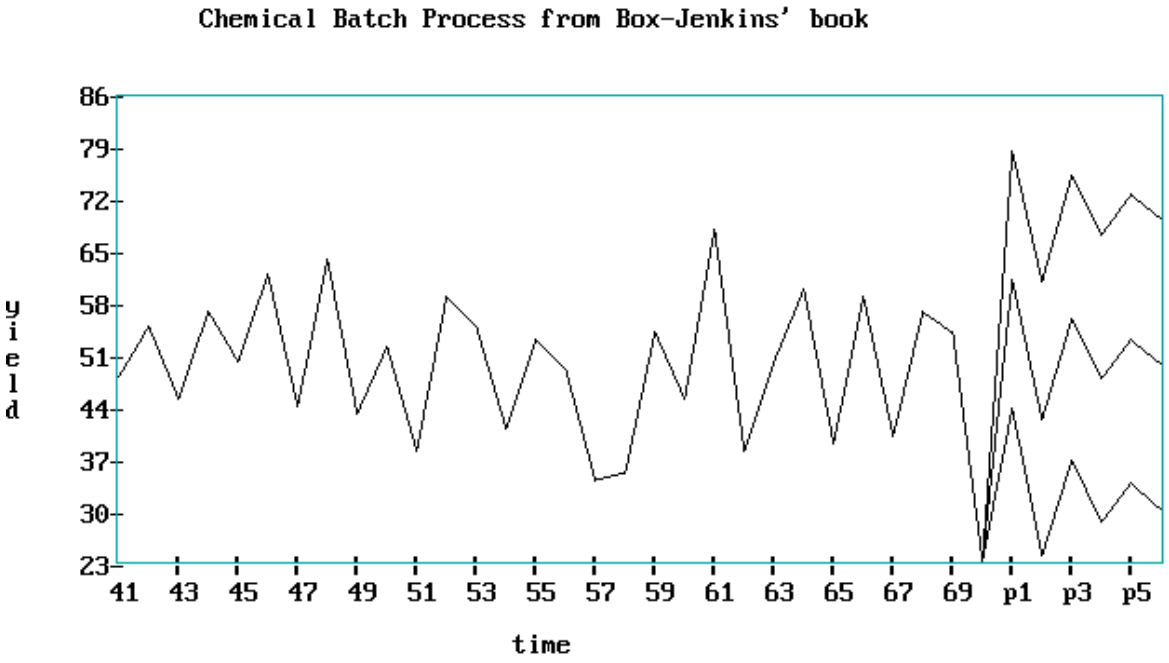
Hypothesis of randomness accepted.
Press any key to proceed to the forecasting section

FORECASTING SECTION

Defaults are obtained by pressing the enter key, without input.
Default for number of periods ahead from last period = 6.
Default for the confidence band around the forecast = 90%.

How many periods ahead to forecast? (9999 to quit...) :
Enter confidence level for the forecast limits :

	90 Percent Confidence limits		
Next Period	Lower	Forecast	Upper
71	43.8734	61.1930	78.5706
72	24.0239	42.3156	60.6074
73	36.9575	56.0006	75.0438
74	28.4916	47.7573	67.0229
75	33.7942	53.1634	72.5326
76	30.3487	49.7573	69.1658



Example2, incorporating seasonality

BOX-JENKINS ARIMA MODEL

You can enter a valid filespec, as long as it has an extension, or you. If you merely press the enter key, ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters):
To return to DOS, press F10.
? *bookg.bj*

MAX	MIN	MEAN	VARIANCE	NO.DATA
622.0000	104.0000	280.2986	14391.9229	144

Do you wish to make transformations? y/n y

The following transformations are available:
1 Square root 2 Exponential
3 Natural log 4 Reciprocal
5 Normalizing, (X-Xbar)/Standard deviation
6 Coding, (X-Constant 1)/Constant 2

ENTER YOUR SELECTION, BY NUMBER: 3

Statistics of Transformed Series:
Mean: 5.542 Variance 0.195

Input order of difference or 0: 1
Input NUMBER of AR terms: 0
Input NUMBER of MA terms: 1 .

Input period of seasonality (2-12) or 0: 12

Input order of seasonal difference or 0: 1
Input NUMBER of seasonal AR terms: 0
Input NUMBER of seasonal MA terms: 1

Statistics of Differenced Series:
Mean: 0.000 Variance 0.002

OK to continue? y/n :
Proceed directly to Forecasting? y(es) / h(elp) or press the ENTER key:

Incorporate back forecasting ? y(es) / h(elp) or press the ENTER key :

Estimation is Finished! Press the ENTER key for display of results...

********** OUTPUT SECTION **********

MA estimates with Standard Errors
Theta 1 : 0.3965 0.0811

Seasonal MA estimates with Standard Errors
Theta 1 : 0.5699 0.0995

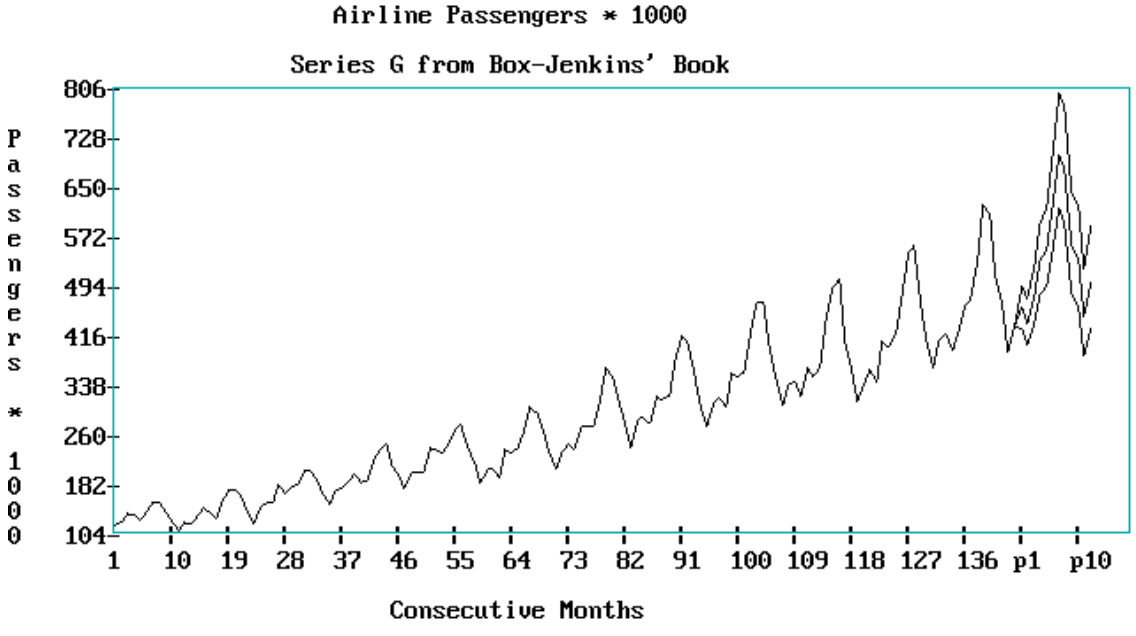
Original Variance	:	0.0021
Residual Variance	:	0.0014
Coefficient of Determination	:	33.9383

FORECASTING SECTION

Defaults are obtained by pressing the enter key, without input.
Default for number of periods ahead from last period = 6.
Default for the confidence band around the forecast = 90%.

How many periods ahead to forecast? (9999 to quit...):
Enter confidence level for the forecast limits : .95

	95 Percent Confidence limits		
Next Period	Lower	Forecast	Upper
145	418.7390	450.4743	484.6147
146	390.9401	426.0882	464.3964
147	435.4600	480.0083	529.1144
148	442.7423	493.0024	548.9684
149	453.5897	509.7755	572.9210
150	515.7007	584.5687	662.6335



Example 3

STEPWISE AUTOREGRESSION

You can enter a valid filespec, as long as it has an extension, or you can select a file extension to search for files of particular interest. If you merely press the enter key (↵) , ALL file names are displayed. Enter FILESPEC or EXTENSION (1-3 letters):
To return, press F10.
? *bookf.bj*

NOTE! In the y/n prompts, the default (pressing 'enter') is capitalized. In other prompts, the default is 0 (zero) unless indicated otherwise

You can analyze all or part of the data. Enter one of the following:
a) First AND last sequence number, e.g. 12-46 (the hyphen is a MUST),
b)
b) or just the first sequence number, e.g. 12, (last number is last entry)
c) or press the enter key (↵) for all data. ?

MAX	MIN	MEAN	VARIANCE	NO.DATA
80.0000	23.0000	51.1286	141.8238	70

Enter seasonal period or 0 or H for help: 0

FITTING ORDER: 1

			std.error
Constant	:	73.086	6.096
Phi 1	:	-0.425	
Res.Var	:	119.805	

Not a very good model...
Final order : AR(1)
Original Variance : 141.8238
Residual Variance : 119.8046
Coefficient of Determination : 15.5258

***** Test on randomness of Residuals *****
The Box-Ljung value = 18.2228
The Box-Pierce value = 14.3342
with degrees of freedom = 23
The 95th percentile = 35.16596

Hypothesis of randomness accepted.

FORECASTING SECTION

Defaults are obtained by pressing the enter key, without input.
Default for number of periods ahead from last period = 6.
Default for the confidence band around the forecast = 90%.

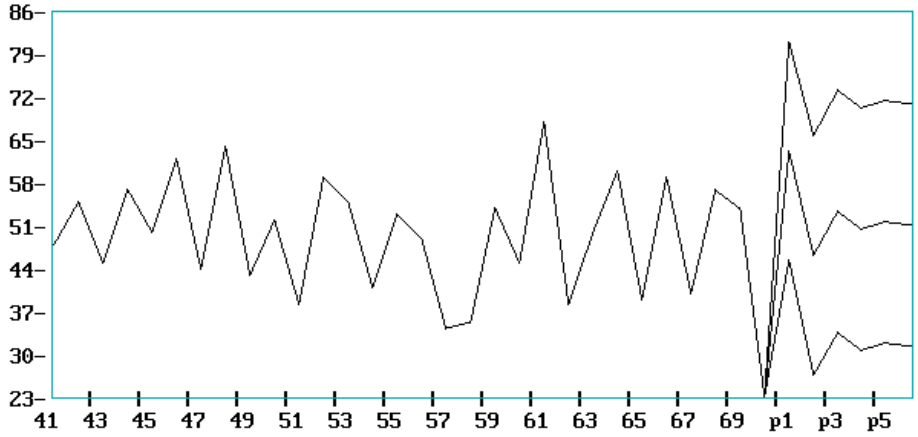
How many periods ahead to forecast? (F3 or 9999 to quit...): 6
Enter confidence level for the forecast limits : .95

95 Percent Confidence limits

Next Period	Lower	Forecast	Upper
1	41.8560	63.3136	84.7711
2	22.8699	46.1842	69.4984
3	29.8286	53.4625	77.0965
4	26.6787	50.3699	74.0611
5	27.9825	51.6840	75.3855
6	27.4223	51.1256	74.8290

Would you like to plot the forecasts and their limits? Y/n : y

The plot consists of the last 'n' data points, followed by the forecasts.
Enter the number of preceding data points or press enter for 30: (the enter key was pressed)
Total number of points plotted = 36
You wish to apply titles ? y/N n



TIMESTAT Input and Output	
1	Input of data from diskette or fixed disk.
2	Output of analysis to file or printer.
3	Printout of graphs.
4	Row Input.
5	Row Edit.
0	Return to the tutorial menu. (Can also use the ↵ key)
Home	Return to the main menu.
Esc	Return to DOS.

Data Input

Input to all programs is prepared by you in advance and stored in an ASCII file. For SINGLE time series you need one or two columns. That is, each row or line consists of one or two pieces of information. The leftmost, or first, is a 4 digit identifier and the rightmost, or second, is the data itself. The identifier is optional, its only purpose is to supply values to the X-axis when the forecasts are plotted. The number of lines is the number of data points in the series.

For example, consider the 'famous' timeseries, the Wolfer Sunspot Numbers. The left column is the year, the right column the number of sunspots that was observed that year. The first few entries are:

1770	101	Two variations of the identifier
1771	82	are allowed: mm/dd or mm-dd, e.g.
1772	66	03/24 or 03-24. It should start
.	.	in column 1.

The file can be prepared using any of your favorite editors, word processors or spreadsheet packages, or you could use ROWINPUT, which comes with TIMESTAT.

There is also ROWEDIT, for editing files created by ROWINPUT. You may give the file any name you wish, for example: WOLF.SUN.

Input to regression analysis is similar to the single variable case, except you need two or more columns. That is, each row or line consists of two or more data elements. The left is the

optional identifier. The others are the Y (dependent) followed by the X (independent) variable(s) for linear, polynomial or multiple regression.

For example, consider the following file: The left column is the year, the middle column is dependent variable, and the right column is the independent variable.

```
1960   -.23   53.6   Two variations of the identifier
1961   .45    47.9   are allowed: mm/dd or mm-dd, e.g.
1962   .01    48.3   03/24 or 03-24. It should start in column 1      .      .
```

The format is not rigid, there must be at least ONE blank space between the fields. ROWINPUT is an excellent tool to input this file.

MULTI-VARIATE AUTOREGRESSION

The multivariate form of the Box-Jenkins' univariate models is sometimes called the ARMAV model, for AutoRegressive Moving Average Vector or plainly vector ARMA process.
The ARMAV model for a stationary multivariate time series, with a zero mean vector, represented by

$$\mathbf{x}_t = \left(x_{1,t}, x_{2,t}, \cdots, x_{n,t}\right)^T \qquad -\infty < t < \infty$$

is of the form:

$$\mathbf{x_t} = \boldsymbol{\varphi}_1\mathbf{x}_{t-1} + \boldsymbol{\varphi}_2\mathbf{x}_{t-2} + \cdots + \boldsymbol{\varphi}_p\mathbf{x}_{t-p} + \mathbf{a}_t$$
$$-\boldsymbol{\theta}_1\mathbf{a}_{t-1} - \boldsymbol{\theta}_2\mathbf{a}_{t-2} - \cdots - \boldsymbol{\theta}_q\mathbf{a}_{t-q}$$

where

$\mathbf{x_t}$ and $\mathbf{a_t}$ are n x 1 column vectors

and

$$\boldsymbol{\varphi}_k = \left\{f_{k,ij}\right\}, k = 1, 2, \cdots, p$$

$$\boldsymbol{\theta}_k = \left\{q_{k,ij}\right\}, k = 1, 2, \cdots, q$$

n x n matrices for the autoregressive and moving average parameters.

$$E\left[\mathbf{a}_t\right] = 0$$

$$E\left[\mathbf{a}_t, \mathbf{a}_{t+k}\right] = D,$$ the dispersion or covariance matrix.

As an example, for a bivariate series with $n = 2$, p=2 and $q = 1$, the ARMAV (2,1) model is:

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} f_{1.11} & f_{1.12} \\ f_{1.21} & f_{1.22} \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} f_{2.11} & f_{2.12} \\ f_{2.21} & f_{2.22} \end{pmatrix} \begin{pmatrix} x_{1,t-2} \\ x_{2,t-2} \end{pmatrix}$$
$$+ \begin{pmatrix} a_{1,t} \\ a_{2,t} \end{pmatrix} - \begin{pmatrix} q_{1.11} & q_{1.12} \\ q_{1.21} & q_{1.22} \end{pmatrix} \begin{pmatrix} a_{1,t-1} \\ a_{2,t-1} \end{pmatrix}$$

The estimation of the matrix parameters and covariance matrix is complicated and virtually impossible without computer software. Especially the estimation of the Moving Average matrices is an ordeal. If we opt to ignore the MA component(s) we are left with the ARV model given by:

$$\mathbf{x}_t = \phi_1 \mathbf{x}_{t-1} + \phi_2 \mathbf{x}_{t-2} + \cdots + \phi_p \mathbf{x}_{t-p} + a_t$$

where

$\mathbf{x}_t$ is a vector of observations, $x_1, x_2, \cdots, x_n$ at time t

$\mathbf{a}_t$ is a vector of white noise , $a_1, a_2, \cdots, a_n$ at time t

$\varphi_k = \{\phi_{k,ij}\}, k = 1, 2, \cdots, p$ is a nxn matrix of AR parameters.

$E[\mathbf{a}_t] = 0$

$E[\mathbf{a}_t, \mathbf{a}_{t-k}] = D,$ the dispersion or covariance matrix.

A model with p autoregressive matrix parameters is an ARV(p) model., or a vector AR model.

The parameter matrices may be estimated by multivariate least squares, but there are other methods such as maximum likelihood estimation.

There are some interesting properties associated with the phi (ϕ) or AR parameter matrices. Consider the following ARV(2) model with $p = 2$:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \phi_{1.11} & \phi_{1.12} \\ \phi_{1.21} & \phi_{1.22} \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + \begin{pmatrix} \phi_{2.11} & \phi_{2.12} \\ \phi_{2.21} & \phi_{2.22} \end{pmatrix} \begin{pmatrix} x_{t-2} \\ y_{t-2} \end{pmatrix} + \begin{pmatrix} a_{1t} \\ a_{2t} \end{pmatrix}$$

Without loss of generality assume that the X series are input and the Y series are output and that the mean vector = (0,0). Therefore transform the observation by subtracting their respective averages. The diagonal terms of each Phi matrix are the scalar estimates for each series, in this case:

$\phi_{1.11}, \phi_{2.11}$ for the input series X

$\phi_{2.11}, \phi_{2.22}$ for the output series Y

The **lower** off-diagonal elements represent the influence of the input on the output.
This is called the "transfer" mechanism or transfer-function model as discussed by Box and Jenkins in chapter 11 of their book.
The ϕ terms in the bivariate ARV(2) model correspond to the δ terms in their transfer model.

The **upper** off-diagonal terms represent the influence of the output on the input. This is called "feedback". The presence of feedback can also be seen as a high value for a coefficient in the correlation matrix of the residuals. A "true" transfer model exists when there is no feedback. This can be seen by expressing the matrix form into scalar form:

$$x_t = f_{1.11}x_{t-1} + f_{2.11}x_{t-2} + f_{1.12}y_{t-1} + f_{2.12}y_{t-2} + a_{1t}$$

$$y_t = f_{1.22}y_{t-1} + f_{2.22}y_{t-2} + f_{1.21}x_{t-1} + f_{2.21}x_{t-2} + a_{2t}$$

.When $\phi_{1.12}$ and $\phi_{2.12}$ are equal to zero there is no contribution of the y terms (e.i. the output) on the x terms (the input).

Finally, delay or "dead' time can be measured by studying the lower off-diagonal elements again. If, for example, $\phi_{1.21}$ is non-significant, the delay is 1 time period.

For an example of multivariate autoregression we will analyze the gas furnace data from the Box-Jenkins text book. In this example gas furnace air and methane were combined in order to obtain a mixture of gases which contained CO2 (carbon dioxide). The methane gas feed rate constituted the input series and followed the process

Methane Gas Input Feed = .60 - .04 X(t)

the CO2 concentration was the output , Y(t).. In this experiment 296 successive pairs of observations (Xt, Yt) were read off from the continuous records at 9 second intervals. For the example described below the first 60 pairs were used. It was decided to fit a bivariate model as described in the previous section and to study the results.
The plots of the input and output series are displayed at the end of the example, together with their forecasts.

MULTIVARIATE AUTOREGRESSION

Enter FILESPEC GAS.BJ

How many series? : 2 the input and the output series
Which order? : 2 this means that we consider times t-1, and t-2 in the model , which is a special case of the general ARV model.

MAX	MIN	MEAN	VARIANCE	SERIES
56.8000	45.6000	50.8650	9.0375	1

2.8340 -1.5200 0.7673 1.0565 2

NUMBER OF OBSERVATIONS: 60 .
THESE WILL BE MEAN CORRECTED. so we don't have to fit the means

---

OPTION TO TRANSFORM DATA
Transformations? : y/N

OPTION TO DETREND DATA
Seasonal adjusting? : y/N

FITTING ORDER: 2
OUTPUT SECTION
the notation of the output follows the notation of the previous section

MATRIX FORM OF ESTIMATES

ϕ 1

1.2265	0.2295
-0.0755	1.6823

ϕ 2

-0.4095	-0.8057
0.0442	-0.8589

Statistics on the Residuals

MEANS

-0.0000	0.0000

COVARIANCE MATRIX

0.01307	-0.00118
-0.00118	0.06444

CORRELATION MATRIX

1.0000	-0.0407
-0.0407	1.0000

ORIGINAL VARIANCE	RESIDUAL VARIANCE	COEFFICIENT OF DETERMINATION	SERIES
9.03746	0.01307	99.85542	1

1.05651 0.06444 93.90084 2

This illustrates excellent univariate fits for the individual series.

This portion of the computer output lists the results of testing for independence (randomness) of each of the series.

Theoretical Chi-Square Value:
The 95th percentile = 35.16595
for degrees of freedom = 23

Test on randomness of Residuals for Series: 1
The Box-Ljung value = 20.7039
The Box-Pierce value = 16.7785
Hypothesis of randomness accepted.

Both Box-Ljung and Box-Pierce tests for randomness of residuals using the chi-square test on the sum of the squared residuals.

Test on randomness of Residuals for Series: 2
The Box-Ljung value = 16.9871
The Box-Pierce value = 13.3958
Hypothesis of randomness accepted.

For example, 16.98 < 35.17 and 13.40 < 35.17

FORECASTING SECTION

The forecasting method is an extension of the model and follows the theory outlined in the previous section. Based on the estimated variances and number of forecasts we can comput the forecasts and their confidence limits. The user, in this software, is able to choose how many forecasts at what confidence level.

Defaults are obtained by pressing the enter key, without input.
Default for number of periods ahead from last period = 6.
Default for the confidence band around the forecast = 90%.

How many periods ahead to forecast? 6
Enter confidence level for the forecast limits : .90:

SERIES: 1

90 Percent Confidence limits

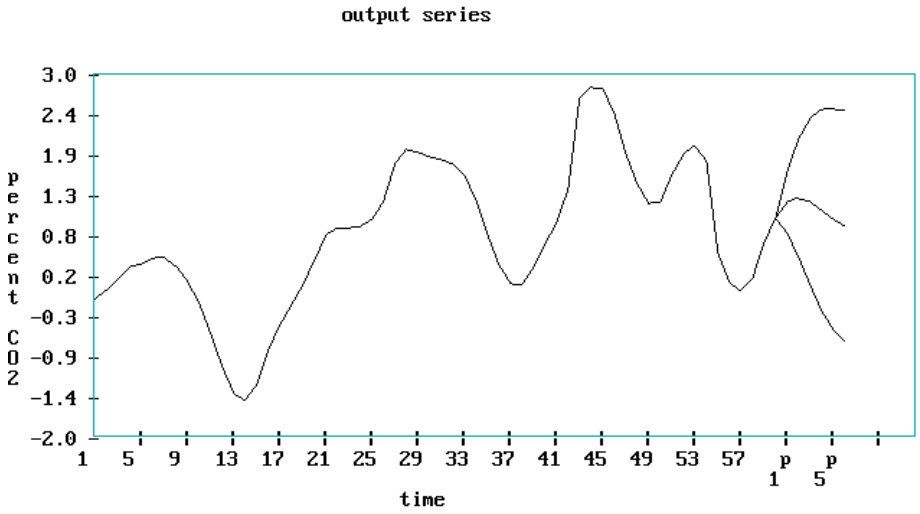
Next Period	Lower	Forecast	Upper
61	51.0534	51.2415	51.4295
62	50.9955	51.3053	51.6151
63	50.5882	50.9641	51.3400
64	49.8146	50.4561	51.0976
65	48.7431	49.9886	51.2341
66	47.6727	49.6864	51.7001

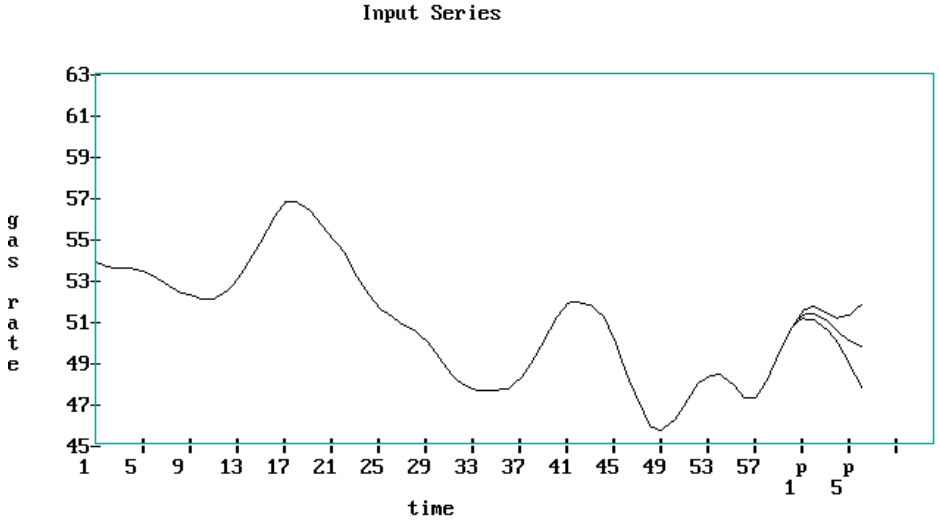
SERIES: 2

90 Percent Confidence limits

Next Period	Lower	Forecast	Upper
61	0.8142	1.2319	1.6495
62	0.4777	1.2957	2.1136
63	0.0868	1.2437	2.4005
64	-0.2661	1.1300	2.5260
65	-0.5321	1.0066	2.5453
66	-0.7010	0.9096	2.5202

In the plots below, the letter p stands for "prediction" These predictions follow the observation, and appear with their confidence bands.


output series
3.0
2.4
1.9
1.3
0.8
0.2
-0.3
-0.9
-1.4
-2.0
percent CO2
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 p 1 p 5
time


Input Series
63
61
59
57
55
53
51
49
47
45
gas rate
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 p 1 p 5
time

Chapter 7

Non Parametric Statistics

The following nonparametric routines are available with the SEMSTAT software:

	NONPARAMETRIC ROUTINES		Help Keys
	SIGNTEST	Two-sample case for Ordinal Levels	H1
	WILCOX	Wilcoxon matched-pairs signed-ranks test	H2
	UTEST	Mann-Whitney U test	H3
	FRIEDMAN	Two-way ANOVA for related sample	H4
	KRUSWAL	Kruskal-Wallis 1-way ANOVA	H5
	SPEARMAN	Rank correlation coefficient	H6
	KENDALL	Rank correlation coefficient	H7
	PARTIAL	Kendall partial rank correlation coefficient	H8
	CONCORD	Kendall coefficent of	H9
	KS	Kolmogorov-Smirnov goodness of fit test	H10
	FISHER	Fisher's exact probability test	H11
	MCNEMAR	MCNemar's test for difference of 2 proportions	H12
	CMH	Cochran-Mantel-Haenzsel Odds ratio test	H13
	TOLERANC	Sample size for nonparametric Tolerance Limits.	H14
		Return to the main menu.	

Select by moving the cursor to the desired routine and clicking the right mouse

Sign Test

The Sign Test compares two population distributions. It is the non-parametric counterpart for the paired t test. The experiment to gather the data is a randomized block design. For example: Five students from the Computer Information Systems department are randomly selected and each is given two PC's, one from brand A and one from brand B. These micro computers have similar configurations. The students are asked to use the systems for an equal amount of time over a period of one semester. A series of tasks is prepared (called benchmark) in addition to general usage. (akin to compulsory and freestyle figure skating in competitions). At the end of the semester the students are asked their preferences. The results can be presented in several ways:

	I		II		III
	PC		Some form of Scaling		The short form
Student	A	B	A	B	A
1	+	-	10	8	+
2	+	-	8	7	+
3	-	+	8	9	-
4	+	-	7	5	+
5	+	-	9	6	+

A + sign means that brand A is preferred, a - sign favors brand B. The null hypothesis is: the 2 population distributions are equal, that is, $P(A > B) = p = .5$ the alternative hypothesis is ; $P(A > B)$ is not .5 (two-sided), Or: $P(A > B) = p > .5$. for the one sided case. The test statistic X, is computed by counting the number of times A exceeded B.

The critical value for a given value of α are the lower and upper values of a binomial distribution with $p = .5$ that satisfies for a two-tailed test: $\{P(X \leq \text{lower})\} \leq \alpha/2$ and $\{P(X \geq \text{upper}\} \leq \alpha/2$. for a 1-tailed test: $\{P(X \geq \text{upper}\} \leq \alpha$. For large samples, ($n \geq 25$), we can use the normal approximation,

There are 5 pairs, four favor A over B. Tables for the Sign Test, available in any textbook that offers Nonparametric Methods, show the following:

Lower rejection region	Upper rejection region	Value of alpha
0	5	0.0625
1	4	0.3750

The null hypothesis is that the 2 distributions are identical. If we reject the null hypothesis we take an alpha risk of 0.3750, since the value of 4 appears in line 2 of the table. This is too much, hence we accept the null.

The Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test for a Paired Experiment is another way to analyze paired differences by using the ranks of the data. It is an improvement of the sign test, since this only looks at the sign of the differences, but does not consider the magnitude. The Signed Rank Test does, and is a better nonparametric equivalent of the paired t test. The procedure is:

1: Using a Likert scale or equivalent to assign a scaled response to each of the n pairs, calculate the difference (Xa - Xb). Differences of 0 are eliminated and n is reduced accordingly.
2: Rank the absolute difference in ascending order. Assign the average of the ranks that are tied to each of the tied group.
3: Calculate the rank sum for the negative differences, Tn, and for the positive differences, Tp. T can be either Tp or Tn.
4: The null hypothesis is that the two populations have the same relative frequncy distribution. The alternative hypothesis for a two-tailed test is that they have different distributions.

For large samples (n > 25), the test statistic is: $z = [T - E(T)] / \sigma$

where:
$E(T) = n(n + 1) / 4$ and $\sigma = n(n + 1)(2n + 1) / 24$
z is approximately normally distributed.

5: The critical value is the z at a given value for $\alpha / 2$ for a two tailed test or z at α for a one tail test.

Let's use the PC evaluation study from the sign test:

	Likert Score			Absolute	
Student	A	B	Difference	Difference	Rank
1	10	8	2	2	3.5
2	8	7	1	1	1.5
3	8	9	-1	1	1.5
4	7	5	2	2	3.5
5	9	6	3	3	5.0

The sum of positive ranks = 13.5, and of negative ranks = 1.5.
T = 13.5, E(T) = 5(6)/4 = 7.5,
σ = 5(6)(11)/24 = 3.708.
Then z = (13.5 - 7.5) / 3.708 = 1.628.

The critical value at $\alpha = .05$ for a 2-tail test = 1.96.
Since 1.628 < 1.96 we CANNOT reject the null hypothesis.

HOWEVER WE USED HERE THE LARGE SAMPLE APPROXIMATION! and n is only 5. For small samples the test statistic is the smaller of Tp and Tn, 1.5. The critical value Tc, is given in tables, presented in any textbook that features nonparametric statistics. Tc is 1 for n = 5 and $\alpha = .05$. The null hypothesis is rejected when T < Tc. But, this is NOT the case.

The Mann-Whitney U-Test

The Mann-Whitney U-Test compares two population distributions. It is the nonparametric equivalent of the t test based on independent random samples. The U-Test uses the rank sums of the two samples. The procedure is:

1. Rank all (n1 + n2) observations in ascending order. Ties are handled by averaging the tied ranks.

2. Calculate the sum of the ranks, Ta and Tb.

3. Calculate the U statistic: $Ua = n1(n2) + .5(n1)(n1 + 1) - Ta$
   or: $Ub = n1(n2) + .5(n2)(n1 + 1) - Tb$
   where: $Ua + Ub = n1(n2)$.

The null hypothesis is: The population relative frequency distributions for A and B are identical. The alternative hypothesis is: They are not the same (2-tailed). The test statistic U, is the smaller of Ua and Ub. For sample sizes exceeding 9 we can use the normal z as follows:

$$z = [U - E(U)] / \sigma$$

where: $E(U) = .5(n1)(n2)$ and $\sigma = [n1(n2)(n1 + n2 + 1)] / 12$

The critical value is the z at a given value for $\alpha / 2$ for a two tailed test or z at α for a one tail test.

For small samples, assuming a 2 tail test and a given value for α, reject the null, if $U \leq Uo$, where $P(U \leq Uo) = \alpha/2$. For example: Consider the following exam data, where 4 students were trained under program A and 4 different students under program B.

Exam	Data	Ranks			
A	B	A	B		
8	33	3	7	Ta = 12	Tb = 24
1	29	6	4	Ua = 14	Ub = 2

7	35	2	8	The test statistic is:
5	30	1	5	U = 2

Using the appropriate table; for n1 = n2 = 4, $P(U_o \leq 1) = .0286$.
Using $U_o \leq 1$ as the rejection region, $\alpha = 2(.0286) = .057$. Since U, (the observed value) is 2, it does not fall in the rejection region, and the null hypothesis is NOT rejected. Had we used the large sample approximation, the results are:
U = 2; E(U) = 8; $\sigma = 3.46$; z = 1.73; z at $\alpha/2$ = z at .025 = 1.96.
Since the test statistic (1.73) < the critical value (1.96) we cannot reject the null hypothesis.

The Friedman Test

The Friedman Test is the nonparametric counterpart of the randomized block design. This design extends the paired t test to k (more than 2) population means. The k populations are called 'treatments'. In order to isolate the experimental (random) error, the treatments are randomly assigned within 'blocks', which are units of relatively homogenous material. In the two-way Friedman set up, the rows are the blocks, and the treatments, which are assigned in a random manner, are the columns. Thus, the table consists of N rows and k columns. Within each row (block) the data are ranked. Then:

Denote the sum of the ranks in each column by R(j) Then the test statistic is:

$$\chi_r^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1)$$

The degrees of freedom for this value of chi-square = (k - 1).
Consider the following table of ranks in each row:

	A	B	C	D
	1	2	4	3
	2	1	3	4
	1	3	4	2
	1	3	2	4
	2	3	4	1
SUM	7	12	17	14

Then the Friedman test statistic is:

(12)/(5)(4)(5)] [(49 + 144 + 289 + 256)] - 3(5)(5) = 81.36 - 75 = 6.36

The null hypothesis is: The probability distributions for all treatments are identical.
The alternative hypothesis is: at least two are different.
The critical value is the chi-square at α with c - 1 degrees of freedom.
At α = .05 and 4 - 1 = 3 df, this value = 7.81.
Since 6.36 < 7.81 we cannot reject the null.hypothesis.

The Kruskal-Wallis test-
The Kruskal-Wallis test compares k > 2 Population distributions. It is an extension of the Mann-Whitney U test. It is the nonparametric counterpart of the randomized design for the one-way ANOVA. Of course as in most of the nonparametric procedures, it is based on ranked data. The table containing the ranks of the data consists of k columns. The columns do not have to be of the same length. Denote the total number of ranks (observations) by N. Denote the sum of the ranks in each column by R(j). The test statistic is:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

n_j = number of cases in the jth sample
The degrees of freedom for this value of chi-square = (k - 1).

The null hypothesis is: The relative frequency distributions of all of the k > 2 populations are identical.
The alternative hypothesis is: At least two of these distributions differ.

Example Data				Ranks of the observations			
A	B	C	D	A	B	C	D
4.2	3.3	1.9	3.5	17	10	2	11
4.6	2.4	2.4	3.1	19	4.5	4.5	9
3.9	2.6	2.1	3.7	14	6	3	12
4.0	3.0	1.7	4.4	15	8	1	18
	3.8	2.7	4.1	13	7	16	
				SUM 65	41.5	17.5	66

$$H = \frac{12}{20(21)}\left(\frac{65^2}{5} + \frac{41.5^2}{5} + \frac{17.5^2}{5} + \frac{66^2}{5}\right) - 3(20) = 13.678$$

The critical value for chi-square for $\alpha = .05$ with df = (c - 1) = 3 = 7.812.
Since 13.678 > 7.812 reject the null hypothesis.

The Spearman's Rank Correlation Coefficient

The Spearman's Rank Correlation Coefficient is a statistic that indicates the degree of linearity between the ranks of two variables. It is the nonparametric counter part of the 'regular' sample correlation computed on the observations instead of the ranks. A rank correlation coefficient is often referred to as a coefficient of agreement for preference data. When there are no ties, the the computation for the Spearman's Rank Coefficient is simplified to:

$$r = 1 - \frac{6\Sigma (d_i)^2}{n(n^2 - 1)}$$

$d_i = x_i - y_i$
x_i and y_i are the ranks of the ith pair of observations

Example:

Rank x_i	Rank y_i	d_i	$(d_i)^2$
7	1	6	36
4	5	-1	1
2	3	-1	1
6	4	2	4
1	8	-7	49
3	7	-4	16
8	2	6	36
5	6	-1	1
		Sum:	144

$$r = 1 - \frac{6(144)}{8(64 - 1)} = -.714$$

There is a significance test for r when n > 9:

$$t = r\sqrt{\frac{N-2}{1-r^2}}, \qquad \text{with } df = N-2$$

Reject the null when the absolute value for t ≥ t at 1 - α /2.
Using this for the example when n = 8 we obtain:

$$t = -.714\sqrt{\frac{6}{1-.714^2}} = -.714(3.4985) = -2.498$$

at α = .05 and 6 degrees of freedom, the critical value = 2.447.
Since the absolute value of the test statistic = 2.498, reject the H_o (that r = 0) at the .025 level.

The Kendall Rank Correlation coefficient

The Kendall rank correlation coefficient r_t also computes the correlation between two sets of ranks as the Spearman counterpart but uses a different scale. This is best illustrated with an example. Let judges A and B rank the top 4 figure skaters in the world. Here are the rankings on skaters a, b, c, and d.

Skater	a	b	c	d
Judge X	3		4	2	1
Judge Y	3		1	4	2

BUT NOW REARRANGE THE ORDER OF THE RANKS OF JUDGE X APPEARS IN THE NATURAL ORDER, (i.e. 1, 2, ...N)

We obtain:

Skater	d	c	a	b
Judge X	1	2	3	4
Judge Y	2	4	3	1

We count how many pairs of ranks in Judge Y's set are in their correct (natural) order with respect to each order. We see that to the right of 2 is 4. That's ok, score a 1. Now 3 is also the right of 2. Score another 1. BUT, the last rank to the right is 1, WHICH IS NOT IN THE CORRECT ORDER. Now score a -1. Starting from 4, the scores become -1, -1.

Finally, starting from 3, the score is -1. Adding all scores we get S = -2. Let N the number of skaters ranked by both X and Y, which is here 4. Then the value for r_t is computed by:

$$r_t = \frac{S}{.5N(N-1)} = \frac{-2}{.5(4)(4-1)} = -.33$$

When N > 10, we consider r_t to be normally distributed with:

$$\mu = 0 \text{ and } \sigma = \sqrt{\frac{2(2N+5)}{9N(N-1)}}$$

which gives: $z = r_t / \sigma$

Using our example with N = 4, we get: $z = -.33 / \sqrt{(26 / 108)} = -.673$.
The absolute value = .673. The critical value for a two-tailed test at $\alpha = .05 = 1.96$.
Since .673 is smaller then 1.96, accept the null hypothesis.(that is there is no evidence of correspondence between the two judges).

The Kendall Partial Rank Correlation Coeficient

Correlation between two variables is sometimes due to the association between each of the variables and a third variable. The Kendall Partial Rank Correlation Coeficient eliminates this effect by keeping the third variable constant.

Kendall has shown that

$$r(xy.z) = \frac{r(xy) - r(xz)r(yz)}{\sqrt{(1 - r(xy)^2)(1 - r(xz)^2}}$$

where r(xy.z) is the partial rank correlation coeffcient between x and y when z is kept constant.
r(xy) is the unadjusted correlation between the ranks of x and y
r(xz) is the unadjusted correlation between the ranks of x and z
r(yz) is the unadjusted correlation between the ranks of y and z

Example:

Subject	a	b	c	d

Rank on Z	1	2	3	4
Rank on X	3	1	2	4
Rank on Y	2	1	3	4

$r(xy) = .67,\ r(yz) = .67,\ r(xz) = .33$

$$r(xy.z) = \frac{.67 - (.67)(.33)}{\sqrt{(1 - .67^2)(1 - .33^2)}} = .63$$

The Kendall Coefficient of Concordance

The Kendall Coefficient of Concordance, W, measures the relation among k (several) rankings of N objects or individuals. Recall that the Spearman and Kendall Coefficients of Correlation deal with two sets of rankings. The procedure to compute W is as follows:

1 find the sum of the ranks, R_j in each column of a k x N table.
2 sum the R_j and divide that sum by N to obtain the mean value.
3 subtract the mean from each of the R_j .
4 compute s, the sum of the squares of these deviations.

5 then $$W = \frac{12\,(s)}{k^2\,(N)\,(N^2 - 1)}$$

Example:
A search committee of 3 are asked to rank 6 candidates for a position. Question: Are they in agreement amongs each other?

	Applicant					
	a	b	c	d	e	f
Member X	1	6	3	2	5	4
Member Y	1	5	6	4	2	3
Member Z	6	3	2	5	4	1

If the committee had been in perfect agreement, they would have ranked the six applicants in the same order. One applicant would have received 3 ranks of 1, and the sum of his ranks would have been 3. The runner up would have received 3 ranks of 2, and the sum = (3)(2) = 6. The least

promising applicant would have received 3 ranks of 6. The sum = Nk = 18. In general, with perfect agreement we get for R_j , the sum per row, the series: k, 2k, 3k, ..., Nk (though not necessarily in that order). In our example: 3, 6, 9, 12, 15, 18.

On the other hand, if there had been no agreement, the various R_j 's would be approximately equal. Hence the larger are the deviations from the mean, the greater is the degree of association. In the example, the R_j were 8, 14, 11, 11, 11, and 8. The mean = 10.5

Then $s = (8-10.5)^2 + (14-10.5)^2 + (3)(11-10.5)^2 + (8-10.5)^2 = 25.5$

$$\text{And} \quad W = \frac{12(25.5)}{(3)^2(6)(6^2-1)} = .16$$

W = .16 expresses the degree of agreement among the three members in ranking the six applicants.

The Kolmogorov-Smirnov Test

The Kolmogorov-Smirrnov test is a better alternative to the chi-square goodness of fit test. The K-S test does not require any minimum value for expected frequencies and can be used with relatively small sample sizes. The data are randomly taken from some population. H_o (the null) is that the sample was drawn from the specified distribution. H_a (the alternative) is: it is not.

The procedure is as follows:

1 Rank the data in ascending order. For large samples one may use a frequency distribution, but this option is not (yet) available in this offering.

2 Compute F_i, the observed cumulative relative frequency for the ith value, (or, if using a frequency distribution, the ith cell).

3 Compute E_i , the expected cumulative relative frequency for the ith value, (or, if using a frequency distribution, the ith cell).

4 Compute the test statistic, $D = \max |F_i - E_i|$, the maximum of the absolute values of the differences.

5 The critical value is a function of the distribution of the max in order statistics, and is tabulated in many text books.

Example:
A random sample of only the following 5 data values was drawn: 288, 231, 249, 146, and 291.

Could these have drawn from a normal distribution with $\mu = 200$ and $\sigma = 50$?

> Note that the relative frequency for each of the 5 data = 1/5 or .2. Then the observed cumulative relative frequency is .2, .4, ...1. The expected cumulative frequencies or computed by conversion to z score, e.g. z = (288 - 200) / 50 = 1.76. Using normal tables, the area under the curve (the cumulative frequency) = .9608.

The results are presenting in the following table:

Data Value	Relative Frequency	Cumulative Relative Frequencies Observed	Expected	Absolute Difference
146	0.2000	0.2000	0.1401	0.0599
231	0.2000	0.4000	0.7324	0.3324
249	0.2000	0.6000	0.8365	0.2365
288	0.2000	0.8000	0.9608	0.1608
291	0.2000	1.0000	0.9656	0.0344

At $\alpha = .10$ and n = 5 the table, or program, gives a critical value of 0.510.
The maximum value of D = 0.3324.
Since 0.3324 < 0.510, we cannot reject Ho, and conclude that the data could have come from the hypothesized normal distribution.

Fisher's Exact Probability Test

The Fisher Exact Probability Test is an excellent nonparametric technique for analyzing discrete data (either nominal or ordinal), when the two independent samples are small in size. It is used when the results from two independent random samples fall all into one or the other of two mutually exclusive classes (i.e. defects vs good, or successes vs failures). In other words, every subject in both groups obtains one of two possible scores. These scores are represented by frequencies in a 2x2 contingency table. The following discussion, using a 2x2 contingency table illustrates how the test operates.

We are working with two independent groups, such as experiments and controls, males and females, the Chicago Bulls and the New York Knicks, etc. This situation is shown in the following table:

	−	+	Total
Group I	A	B	A + B
Group II	C	D	C + D
Total	A + C	B + D	N

The column headings, here arbitrarily indicated as plus and minus, maybe of any two classifications, such as: above and below the median, passed and failed, Democrat and Republican, agree and disagree, and so on.
Fisher's test determines whether the two groups differ in proportion with which they fall into the two classifications. For the table above, the test would determine whether Group I and Group II differ significantly in the proportion of plusses and minuses attributed to them.

The method proceeds as follows:

$$p = \frac{\dbinom{A+C}{A}\dbinom{B+D}{B}}{\dbinom{N}{A+B}} = \frac{\dbinom{A+C}{A}\dbinom{B+D}{B}}{\dfrac{N!}{(A+B)!(C+D)!}} = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

The exact probability of observing a particular set of frequencies in a 2 X 2 table, when the marginal totals are regarded as fixed, is given by the hypergeometric distribution. That is, the exact probability of the observed situation is computed by taking the ratio of the product of the factorials of the marginal totals to the product of the cell frequencies, multiplied by N (the total) factorial. But the test does not just look at the observed case. If needed, it also computes the probability of more extreme outcomes, with the same marginal totals. This will become clear in the next illustrative example.

Consider the following set of 2 X 2 contingency tables:

observed data — More extreme outcomes with same marginals

a

2	5	7
3	2	5
5	7	12

b

1	6	7
4	1	5
5	7	12

c

0	7	7
5	0	5
5	7	12

Table (a) shows some observed frequencies and tables (b, c) show the two more extreme distributions of frequencies, which could occur with the same marginal totals 7, 5. Given the observed data in table (a) , we wish to test the null hypothesis at, say, $\alpha = .05$. Applying the previous formula to tables (a), (b) and (c) we obtain

$$p_a = \frac{7!5!5!7!}{12!2!5!3!2!} = .26515$$

$$p_b = \frac{7!5!5!7!}{12!1!6!4!1!} = .04399$$

$$p_c = \frac{7!5!5!7!}{12!0!7!5!0!} = .00126$$

The probability associated with the occurrence of values as extreme as the observed results under Ho. is given by adding these three p's:

$$.26515 + .04399 + .00126 = .31040$$

So p = .31040 is the probability that we get from Fisher's test.
Since .31040 is larger than a we cannot reject the null hypothesis.

Tocher's Modification

Tocher (1950) showed that a slight modification of the Fisher test makes it an even stronger test. He starts with isolating the probability of all cases more extreme than the observed one. In this example that is

pa + pb = .04399 + .00126 = .04525.

Now, if this probability is larger than α, we cannot reject Ho But if this probability is less than α, while the probability that we got from Fisher's test is greater than α (as is the case in our example) then Tocher advises the computes the following ratio:

$$\frac{\alpha - p_{\text{more extreme cases}}}{p_{\text{observed data}}}$$

for the data in the example, that would be

$$\frac{a - (p_b + p_c)}{p_a} = \frac{.05 - .0425}{.26515} = .0179$$

Next we go to a table of random numbers and at random draw a number between 0 and 1. If this random number is smaller than the ratio above of .0179, we reject Ho. If it is larger we cannot reject H_o. This added small probability of rejecting H_o makes the Fisher test a little bit less conservative. The test is a one-tailed test. For a two-tailed test, the value of p obtained from the formula must be doubled,

McNemar's test for difference in two proportions

Consider the following layout:

		condition (group 2) yes	condition (group 2) no
Condition (group 2)	yes	A	C
	no	B	D

Where:

A = no. of respondents answering yes to cond. 1 and yes to cond. 2
B = no. of respondents answering yes to cond. 1 and no to cond. 2
C = no. of respondents answering no to cond. 1 and yes to cond. 2
D = no. of respondents answering no to cond. 1 and no to cond. 2

A Computer Session:
Enter A, B, C, D separated by blanks or press Enter to quit: 14 4 3 4
The null hypothesis is: Pa = Pd. Let $\alpha = .05$ for a 1 tail test.

Computed Chisquare: 4.500
Theoretical Chisquare: 3.748
Computed p value: 0.032
Reject the null hypothesis, since .032 < .05

The Cochran-Mantel-Haenszel (CMH) methods

The Cochran-Mantel-Haenszel (CMH) methods test the null hypothesis that the variables X and Y are conditionally independent given the variable Z. This means that the conditional odds ratio between X and Y in each partial table = 1. The CHM methods first test the null hy[othesis Ho: X and Y are conditionally independent and then estimate the Common Odds Ratio with an confidence interval. To test Ho they generate a test statistic, which follows for large samples, a chi-square distribution with 1 degree of freedom.

The following example studies the effect of a weight reducing diet in three different regions. The people that went on the diet are the levels of group classification X, the two possible outcomes (yes, no) for weight reduction are the levels of a response variable, Y and the different groups (regions) are levels of a control variable, Z. The regions may vary in climate or in socioeconomic status etc. Thus we wish to study the association between X and Y while controlling for Z.

		Yes	No
DIET	Yes	60	30
	No	35	65

Group 1

	Yes	No
Yes	82	60
No	25	47

Group 2

	Yes	No
Yes	52	12
No	22	62

Group 3

Here is the computer output:

The CMH test-statistic = 64.029 with a p-value < .001, hence reject Ho.
The CMH odds ratio = 4.212 with a 95% Confidence Band of 2.962 - 5.991.

This means that the odds of weight reduction when following the prescribed diet equal about 4 times the odds when not following it.

To execute the program select the CMH line from the menu,
When the routine prompts for the name of the input file, type:CMH.FIL

The input was formed as follows :and stored in a file named CMH.FIL

60 30 35 65 corresponding to group 1
82 60 25 47 corresponding to group 2
52 12 22 62 corresponding to group 3

Sample size fo Nonparametric Tolerance Limits

Nonparametric Tolerance Limits is a method for constructing intervals which have a specified chance ϕ (probability) of covering a certain proportion, P, of any population. The range between the observed maximum and minimum of a random sample of size N can be expected to include a certain proportion P of a population. From the distribution of the range we can find N by the equation:

$$1-\phi = NP^{N-1} - (N-1)P^{N}$$

This is solved iteratively. For example, if we wish to find N for a 90% chance of including 99% of the population, we set (as requested by the program) ϕ to .9 and P to .99 and obtain N = 387.

The program session is listed below:

```
**************************************************
 *         Nonparametric two-sided tolerance limits          *
*          Find the sample size N, to cover a proportion P    *
*          of the population with probability í                *
**************************************************

 Enter the proportion of the population to be covered:    .99
 Enter the probability of coverage or press Enter to quit: .9

 Take a sample of N =  387
 The probability is .9 that at least a proportion .99
 of the population falls between the sample minimum and maximum.

 More? y/n:   n
```

Chapter 8

Multivariate Statistics

The MANOVA Model

In the one-way analysis of variance, the differences among populations are studied on the basis of observations drawn from different groups.
The groups can be considered as the independent variable and the observations as the dependent variable. When the observations become vectors, the analysis of variance becomes *multivariate.* The *dependent* vector variable is assumed to follow a multivariate normal distribution, with the same dispersion, or variance-covariance matrix, for each population.
The linear model for MANOVA is:

$$\mathbf{X}_{ki} = \mathbf{m} + (\mathbf{m}_k - \mathbf{m}) + (\mathbf{X}_{ki} - \mathbf{m}_k)$$

where $\mathbf{X}_{ki}$ is the dependent vector variable for the ith subject in the kth sample.
k = 1, 2....g where g is the number of groups. Groups is the multivariate equivalent of treatments.
m is the vector of total **sample means**, also called common or grand centroid. $\mathbf{m}_k$ is the centroid for sample k. Centroid is the vector equivalent for mean in the univariate case.

Analoqous to the univariate case we compute the "among groups" and the "within groups" matrices , defined as:

$$\mathbf{A} = \sum_{k=1}^{g} \sum_{i=1}^{N_k} (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})'$$

and

$$\mathbf{W} = \sum_{k=1}^{g} \sum_{i=1}^{N_k} (\mathbf{X}_{ki} - \mathbf{m}_k)(\mathbf{X}_{ki} - \mathbf{m}_k)'$$

The "total" matrix is defind as: **T = A + W**

The Centroid

The estimator of the common population dispersion, based on the group means vector is:

$$\mathbf{D}_A = \left(\frac{1}{g-1}\right)\mathbf{A}$$

The null hypothesis is: H_0: $\boldsymbol{\mu}_\mathbf{k} = \boldsymbol{\mu}$ for k = 1,2, ..., g.

If it is true, than the best estimator for the common populations centroid, $\boldsymbol{\mu}$, is **m**, the grand centroid, defined by:

$$\mathbf{m} = \frac{1}{N}\sum_{k=1}^{g}\sum_{i=1}^{Ng} X_{ki}$$

When the null hypothesis is rejected, the treatment effects for further testing are contained in the matrix of deviations of group means from the grand means. Each column of this matrix is formed as $\mathbf{m}_k$-**m**.
In order to test the null hypothesis of equality of group centroids, one has to come up with a test-statistic and a critical value. Wilks (1932) devised a test-statistic that is a ratio of determinants:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

This ratio is known as Wilks' Lambda. It is a family of three parameter curves, the parameters are derived from the number of groups, the number of observations, and the number of elements in a vector variable. Although many efforts have been undertaken to tabulate Lambda for a set of specific values for the parameters, Lambda was difficult to apply. Rao (1952) finally derived an *F* transformation that fitted very closely to the Lambda cumulants and can be successfully used for testing the null hypothesis.
To implement Rao's *F* approximation it is necessary to compute a set of functions of the three design parameters. These design parameters are *p*, (the number of elements in the vector), *g*, the number of groups), and *N* (the total number of observations in all he groups). See the appendix for the details.

The Dispersion

The estimator for ***W*** based on the pooled within-groups deviations is:

$$\mathbf{D}_w = \left(\frac{1}{N-g} \right) \mathbf{W}$$

N is the total number of observations and *g* is the number of groups.

The estimator for each individual groups is:

$$\mathbf{D}_k = \left(\frac{1}{N_k - g} \right) \mathbf{W}_k$$

where

$$W_k = \sum_{k=1}^{N_k} \left(X_{ik} - m_k \right) \left(X_{ik} - m_k \right)'$$

and

$\mathbf{m}_k$ is the group centroid, N_k is the group sample size.

George Box defines a test criterion M for the null hypothesis on dispersion,
H_o: $\Delta_k = \Delta$, $k = 1,2,...,g$.

$$M = \left(N - g \right) \ln \left| \mathbf{D}_w \right| - \sum_{k=1}^{g} \left(N_k - 1 \right) \ln \left| \mathbf{D}_k \right|$$

where

$\left| \mathbf{D}_w \right|$ and $\left| \mathbf{D}_k \right|$ are the determinants of the respective dispersions.

From M and the design parameters, we obtain a test-statistic and critical value based o n the F distribution. The functions needed appear in the appendix.

Univariate F Ratios

If the null hypothesis of equality of centroids is *rejected* , further investigation is possible by inspecting the univariate F ratios.

1. The mean squares of among-groups are obtained by dividing the diagonal elements of **A** by the degrees of freedom of the respective groups. The degrees of freedom are g - 1.

2. The mean squares of within-groups are obtained by dividing the diagonal elements of **W** by the associated degrees of freedom, which are N - g.

To be sure, these are *not independent* tests. Univariate anova in this sense only serves to find out which variable may have contributed to rejection of the Lambda test.

The menu for the multivariate routines is displayed below:

Move the mouse to the desired line and click the left button

SLCT Keys		MULTIVARIATE ROUTINES
1	MANOVA	Multivariate ANOVA and ANOCOVA
2	CLASSIF	Classification Analysis
3	PRINCO	Principal Components
4	EIGEN	Eigenvalues
		Return to main menu

If you press the Shift and F1 keys the mouse will dissapear and you can then use the arrow keys to move the cursor to the desired line and then press Enter or type 1, 2 or 3 followed by pressing the Enter key.

Example

The hypothetical sample data consists of six test scores obtained from 30 individuals in four different groups. Each group is taught by a different instructor.
Multiple analysis of variance (manova) is used to test equality of centroids.

Group 1						Group 2						Group 3						Group 4					
3	11	9	15	20	10	3	10	8	8	23	8	3	10	9	8	24	8	9	10	27	8	28	16
4	12	3	8	22	7	11	7	8	9	8	15	9	4	10	7	9	9	4	12	3	8	23	7
9	3	2	8	9	8	8	10	2	8	27	16	4	13	10	7	21	15	9	3	2	8	21	7
16	2	2	2	7	2	1	6	8	14	14	13	8	5	16	16	16	7	15	2	2	2	7	2
5	10	5	8	23	9	7	8	9	6	18	2	6	9	10	5	23	11	9	10	26	8	27	16
17	3	2	8	6	3	7	9	8	2	19	9	8	10	5	8	27	16	8	9	2	8	26	16
2	10	9	8	29	16	7	10	5	8	27	17	17	3	2	7	6	3	7	8	6	9	18	2
7	10	5	8	28	18													7	10	5	8	26	16

NOTE: The computer input file is formed by stacking the groups vertically, with a blank line between groups.

```
              ******************************************
              *  Multi-Variate Analysis of Variance (MANOVA)   *
              ******************************************

 You can enter a valid filespec, as long as it has an extension, or you
 can select a file extension to search for files of particular interest.
 If you merely press the enter key (<-┘), ALL file names are displayed.
 Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? test.dat

 NO. CONTROL VARIABLES FOR ANALYSIS OF COVARIANCE?  (0 TO SKIP): 0

  BOX TEST FOR EQUALITY OF DISPERSIONS
  Box M value:  152.818
  Box F value:    1.331  with DF =  63  1522
  The probability of F is:   0.955
  Reject the null hypothesis at the .05 level ...

  RAO TEST FOR EQUALITY OF CENTROIDS
  Wilks Lamda value:    0.647
```

Rao F value: 0.546 with DF = 18 59
The probability of F is: 0.078
Accept the null hypothesis at the .05 level ...

UNIVARIATE ANOVA. DF ARE: 3 26

VAR	AMONG MS	WITHIN MS	F RATIO	PROBABILITY
1	0.859	19.619	0.044	0.018
2	1.465	11.945	0.123	0.058
3	37.573	39.459	0.952	0.570
4	6.346	9.833	0.645	0.407
5	22.918	62.786	0.365	0.219
6	20.173	29.575	0.682	0.429

MEANS

	1	2	3	4	5	6
1	7.875D+00	7.500D+00	4.625D+00	7.250D+00	1.850D+01	8.875D+00
2	7.143D+00	8.571D+00	9.571D+00	7.857D+00	2.014D+01	1.257D+01
3	7.857D+00	7.857D+00	8.857D+00	9.286D+00	1.743D+01	1.014D+01
4	7.750D+00	8.000D+00	6.750D+00	7.375D+00	2.138D+01	9.250D+00
GRAND	7.667D+00	7.967D+00	7.333D+00	7.900D+00	1.940D+01	1.013D+01

VARIANCE-COVARIANCE MATRIX

1.962D+01	-1.116D+01	-5.215D+00	-6.099D+00	-2.275D+01	-9.541D+00
-1.116D+01	1.195D+01	5.618D+00	1.918D+00	2.261D+01	1.067D+01
-5.215D+00	5.618D+00	3.946D+01	3.937D+00	1.623D+01	9.345D+00
-6.099D+00	1.918D+00	3.937D+00	9.833D+00	4.622D+00	3.838D+00
-2.275D+01	2.261D+01	1.623D+01	4.622D+00	6.279D+01	3.018D+01
-9.541D+00	1.067D+01	9.345D+00	3.838D+00	3.018D+01	2.957D+01

CORRELATION MATRIX

1.000	-0.729	-0.187	-0.439	-0.648	-0.396
-0.729	1.000	0.259	0.177	0.826	0.568
-0.187	0.259	1.000	0.200	0.326	0.274
-0.439	0.177	0.200	1.000	0.186	0.225
-0.648	0.826	0.326	0.186	1.000	0.700
-0.396	0.568	0.274	0.225	0.700	1.000

The following results, in addition to the displayed results,
can be captured in an optionally saved file:
the Total, Among Groups and Within Groups sum of squares matrices,
and the individual groups covariance matrices.

Enter file-id to save the output or press the Enter key to quit...

Appendix

To obtain the F based test-statistic for testing the null hypothesis on equality of centroids:

$$s = \sqrt{\frac{p^2(g-1)^2 - 4}{p^2 + (g-1)^2 - 5}} \qquad p^2 + (g-1)^2 > 0 \text{ else } s = 1$$

$$n_1 = p(g-1)$$

$$n_2 = s\left[(N-1) - \frac{p + (g-1) + 1}{2}\right] - \frac{p(g-1) - 2}{2}$$

Now let $y = \Lambda^{1/s}$ and $F_{n_2}^{n_1} = \left(\frac{1-y}{y}\right)\left(\frac{n_2}{n_1}\right)$

where n_1 is the number of degrees of freedom for the numerator and n_2 the denominator for the above variance ratio test-statistic.
It has been shown by Rulon and Brooks (1968) how this test-statistic can be applied to:

Hotelling's T^2 statistic for g = 2

The univariate anova *F* for p = 1

Student's *t* for g = 2 and p = 1

The Lambda test of the null hypothesis (equality of mean vectors)
assumes the g group covariance (dispersion) matrices are based on samples drawn from
g multivariate normal populations, with the same dispersion matrix, **Δ**.

Functions used in testing the null hypothesis on equality of dispersion matrices

$$A_1 = \left(\sum_{k=1}^{g} \frac{1}{N_k - 1} - \frac{1}{N - g} \right) \frac{2p^2 + 3p - 1}{6(g-1)(p+1)}$$

$$A_2 = \left(\sum_{k=1}^{g} \frac{1}{(N_k - 1)^2} - \frac{1}{(N - g)^2} \right) \frac{(p-1)(p+2)}{6(g-1)}$$

for $A_2 > A_1^2$

$$n_1 = \frac{(g - 1)p(p+1)}{2}$$

$$n_2 = \frac{n_1 + 2}{A_2 - A_1^2}$$

$$b = \frac{n_1}{1 - A_1 - (n_1 / n_2)}$$

$$F_{n_2}^{n_1} = \frac{M}{b}$$

for $A_2 < A_1^2$

$$n_1 = \frac{(g - 1)p(p+1)}{2}$$

$$n_2 = \frac{n_1 + 2}{A_2 - A_1^2}$$

$$b = \frac{n_1}{1 - A_1 - (2 / n_2)}$$

$$F_{n_2}^{n_1} = \frac{n_2 M}{n_1 (b - M)}$$

So now we have the test-statistics M/b or $n_2M/n_1(b-M)$.
The critical value is a function of n_1, n_2, and α, and can be found from F tables.

Classification Analysis

This is a technique to obtain information from data that originate from multiple groups for the purpose of classifying individuals into one of these groups.
Broadly speaking, we are estimating the probability that subject i is a member of population j. We are examining a set of hypotheses pertaining to the group membership of subject i, given g groups. Each subject (observation) consists of p measurements.
There are several methods in the multivariate literature, this program brings two of them, Anderson's method and Geisser' method.

The Anderson method first evaluates each of a calculated set o*f linear functions*, one corresponding to each group, and then assigns the subject to the group that exhibits the largest probability, which in turn is associated with the largest linear function.
He employs a test statistic which is due to Mahalonobis, which is similar to the well known Hotelling T statistic, calculated as follows:

$$\mathbf{M} = \sum_{\mathbf{i=1}}^{\mathbf{g}} n_i \left(\mathbf{m_i} - \mathbf{m}\right)' \mathbf{D}^{-1} \left(\mathbf{m_i} - \mathbf{m}\right)$$

where

$\boldsymbol{m}$ is the grand centroid
$\boldsymbol{m_i}$ is the sample centroid of group j, j = 1, 2, ... g
$\boldsymbol{D}$ is the sample dispersion matrix
n_i is the sample size per group

$\boldsymbol{M}$ follows a chi-square distribution with p(g-1) degrees of freedom
g is the number of groups, p is the number of elements per subject..

It can be shown that each group has a constant term:

$$\boldsymbol{C_{0j}} = -.5\ \boldsymbol{m_j'} \boldsymbol{D^{-1}} \boldsymbol{m_j} \qquad j = 1,2,\cdots,g$$ and p coefficients:

$$\boldsymbol{C_{ij}} = \boldsymbol{D^{-1}} \boldsymbol{m_j} \qquad i = 1,\ldots p \qquad j = 1,\ldots g$$

To determine the maximum likelihood of membership for subject i, these constants and coefficient vectors of each of the g groups is used with $\boldsymbol{X}_{ij}$
That is, for each ith subject in each jth group the following calculations are performed:

the *discrimant function* is computed by:

$$\boldsymbol{f_j} = \boldsymbol{C_{0j}} + \boldsymbol{C_{ij}} \boldsymbol{X_{ij}}$$

The assumption is made that the distribution for each group population is multivariate normal, with equal dispersion.

The probability associated with the largest discriminant function is

$$p_i = 1 / \sum_{k=1}^{g} e^{(f_k - f_l)}$$

where f_l = the value of the largest discriminant function.

The Anderson method works best for reasonably large sample sizes, N > 30.

The Geisser method is based on small sample theory.

For each ith subject in each jth group, these calculations are performed:

$$f_{ij} = q_j \left| \frac{n_j}{n_j + 1} \right|^{\frac{N-g}{2}} \left[1 + \frac{n_j (\mathbf{x_j} - \mathbf{m_j})' \mathbf{D}^{-1} (\mathbf{x_j} - \mathbf{m_j})}{n_j + 1} \right]^{-\frac{N-g+1}{2}}$$

where

$\boldsymbol{m}_j$ is the sample centroid of group j
$\boldsymbol{D}$ is the sample dispersion matrix
n_j is the group sample size
N is the total sample size
g is the number of groups
p is the number of elements per subject
q_j is the prior probability of assigment to group j

From the f_{ij} the probabilities are computed as follows:

$$P_{ij} = \frac{f_{ij}}{\sum_{k=1}^{g} f_{ik}} \qquad i = 1, \ldots p \qquad j = 1, \ldots g$$

The subject is assigned to the group with the largest probability

Example 1

```
         ****************************************************
         *          Classification Analysis                 *
         *            Anderson's Method                     *
         ****************************************************

  You can enter a valid filespec, as long as it has an extension, or you
  can select a file extension to search for files of particular interest.
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? test.dat

  The Mahalonobis Chi-Square Statistic is:   12.781
  The degrees of freedom are:                18
  The level of significance (p value) is:     0.805

                                   MEANS
          1          2          3          4          5          6
    1  7.875D+00  7.500D+00  4.625D+00  7.250D+00  1.850D+01  8.875D+00
    2  7.143D+00  8.571D+00  9.571D+00  7.857D+00  2.014D+01  1.257D+01
    3  7.857D+00  7.857D+00  8.857D+00  9.286D+00  1.743D+01  1.014D+01
    4  7.750D+00  8.000D+00  6.750D+00  7.375D+00  2.138D+01  9.250D+00

GRAND  7.667D+00  7.967D+00  7.333D+00  7.900D+00  1.940D+01  1.013D+01


      VARIANCE-COVARIANCE MATRIX

  1.962D+01 -1.116D+01 -5.215D+00 -6.099D+00 -2.275D+01 -9.541D+00
 -1.116D+01  1.195D+01  5.618D+00  1.918D+00  2.261D+01  1.067D+01
 -5.215D+00  5.618D+00  3.946D+01  3.937D+00  1.623D+01  9.345D+00
 -6.099D+00  1.918D+00  3.937D+00  9.833D+00  4.622D+00  3.838D+00
 -2.275D+01  2.261D+01  1.623D+01  4.622D+00  6.279D+01  3.018D+01
 -9.541D+00  1.067D+01  9.345D+00  3.838D+00  3.018D+01  2.957D+01

      CONSTANT AND COEFFICIENTS

 -2.849D+01  2.639D+00  2.122D+00 -1.717D-01  1.912D+00  5.848D-01 -4.048D-01
 -2.921D+01  2.619D+00  2.252D+00 -4.816D-02  1.883D+00  4.373D-01 -2.178D-01
 -3.186D+01  2.744D+00  2.396D+00 -6.457D-02  2.133D+00  4.262D-01 -3.272D-01
 -3.082D+01  2.719D+00  2.039D+00 -1.335D-01  1.945D+00  7.168D-01 -4.876D-01

  CASE      LARGEST      ASSOCIATED   GROUP IT
   NO.      FUNCTION     PROBABILITY  BELONGS TO

  GROUP  1
    1        25.392         0.381        4
    2        32.337         0.370        1
    3        18.597         0.363        1
    4        24.733         0.442        1
    5        30.164         0.345        1
    6        40.487         0.442        3
    7        22.380         0.318        2
```

8	34.796	0.293	2

```
GROUP  2
  1         39.411          0.510         2
  2         32.275          0.501         3
  3         38.169          0.348         4
  4         16.308          0.431         3
  5         26.922          0.443         4
  6         19.125          0.364         2
  7         34.577          0.285         2

GROUP  3
  1         39.383          0.676         3
  2         18.049          0.466         2
  3         29.163          0.546         2
  4         39.688          0.667         3
  5         23.373          0.306         2
  6         37.769          0.330         4
  7         38.354          0.390         3

GROUP  4
  1         24.809          0.337         4
  2         32.922          0.375         1
  3         26.700          0.623         4
  4         22.094          0.457         1
  5         39.022          0.522         2
  6         35.413          0.341         4
  7         33.159          0.431         4
  8         34.362          0.278         1
Number of correct classifications =  14
This is  46.667 percent
```

```
  Example 2
        ***************************************************
        *          Classification Analysis                *
        *                Geisser's Method                 *
        ***************************************************

  You can enter a valid filespec, as long as it has an extension, or you
  can select a file extension to search for files of particular interest.
  If you merely press the enter key (□ÄÙ), ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? test.dat

  The Mahalonobis Chi-Square Statistic is:   12.781
  The degrees of freedom are:                18
  The level of significance (p value) is:     0.805
```

MEANS

	1	2	3	4	5	6
1	7.875D+00	7.500D+00	4.625D+00	7.250D+00	1.850D+01	8.875D+00
2	7.143D+00	8.571D+00	9.571D+00	7.857D+00	2.014D+01	1.257D+01
3	7.857D+00	7.857D+00	8.857D+00	9.286D+00	1.743D+01	1.014D+01
4	7.750D+00	8.000D+00	6.750D+00	7.375D+00	2.138D+01	9.250D+00
GRAND	7.667D+00	7.967D+00	7.333D+00	7.900D+00	1.940D+01	1.013D+01

VARIANCE-COVARIANCE MATRIX

1.962D+01	-1.116D+01	-5.215D+00	-6.099D+00	-2.275D+01	-9.541D+00
-1.116D+01	1.195D+01	5.618D+00	1.918D+00	2.261D+01	1.067D+01
-5.215D+00	5.618D+00	3.946D+01	3.937D+00	1.623D+01	9.345D+00
-6.099D+00	1.918D+00	3.937D+00	9.833D+00	4.622D+00	3.838D+00
-2.275D+01	2.261D+01	1.623D+01	4.622D+00	6.279D+01	3.018D+01
-9.541D+00	1.067D+01	9.345D+00	3.838D+00	3.018D+01	2.957D+01

CASE NO.	LARGEST PROBABILITY	GROUP IT BELONGS TO
GROUP 1		
1	0.360	4
2	0.343	1
3	0.337	1
4	0.395	1
5	0.334	1
6	0.395	3
7	0.308	2
8	0.286	2
GROUP 2		
1	0.432	2
2	0.457	3
3	0.328	4

4	0.384	3
5	0.404	4
6	0.342	2
7	0.280	2

```
GROUP  3
  1          0.571         3
  2          0.422         2
  3          0.481         2
  4          0.536         3
  5          0.298         2
  6          0.316         4
  7          0.359         3

GROUP  4
  1          0.324         4
  2          0.348         1
  3          0.529         4
  4          0.407         1
  5          0.445         2
  6          0.323         4
  7          0.396         4
  8          0.274         1
Number of correct classifications =  14
This is  46.667 percent
```

Principal Components

Principal components approach consists of transforming the *p*-dimensional data into a lower dimensional set of data (sometimes bivariate or even univariate). This is accomplished by setting up meaningful weighted linear combinations of the *p*-dimensions.
Those new variables are called principal components.These were derived by Harold Hotelling in 1933. It works as follows:
Let $\left(X_i^1, X_i^2, \cdots, X_i^p\right)$ be the *i*-th *p*-dimensional observation in the original data.

Now we create a new *p*-dimensional observation $\left(Z_i^1, Z_i^2, \cdots, Z_i^p\right)$ such that the *i*-th

Variable in the **Z**'s is a linear combination of the deviations of the original *p* dimensions from their targets.

$$\mathbf{Z}_i^{'} = \sum_{j=1}^{p} c_{ij}\left(\mathbf{X}_i^j - \mathbf{m}^j\right)$$

For a process with multidimensional data, its overall variance is defined as the sum of the variances of the *p* variables, i.e. the sum of the diagonal elements of the covariance matrix **Σ**.
In the interest of clarity and completeness a brief overview of the principal components methodology will be given below. This method will reduce the number of parameters that has to be estimated in a *p*-element vector. In general, the number of estimators is:

p	means
p	variances
$(p^2-p)/2$	covariances

So for p = 4 there are 14 estimators, for p = 6 there are 27 and for p = 10 there are 65 estimators. Would not it be nice if this number can be reduced? The answer is YES, and it can be done. The method of principal components will accomplish this reduction. It begins with the standardization of the vector variable., that is subtract the mean and divide each element by the standard deviation The vector means will all be zero.. For the standardized vector **Z**, the dispersion matrix becomes the correlation matrix

$$\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^{'}$$

Next an uncorrelated vector is constructed. This is accomplished by transforming the data. To produce a transformation vector for **y**, is saying that we want a **V** matrix such that its dispersion matrix $\mathbf{D}_y$ is diagonal. This means that

$$\mathbf{y = V'Z}$$

where **V** is a *p x n* coeffcient matrix that carries the *p*-element vector into the *n*-element derived variable **y.**

The centroid of y is

$$\mathbf{m_y = V'm_z = 0}$$

and its dispersion is

$$\mathbf{D_y = V'D_z = V'RV}$$

where **R** is the correlation matrix for **Z**.
The transformation "find **V** such that $\mathbf{D}_y$ is a diagonal matrx" is called an orthogonalizing transformation. There exists an infinity of values for **V** that will yield a diagonal $\mathbf{D}_y$ for any correlation matrix **R**, so a restriction is imposed on the problem. The first element of **y** is called the *first principal component* and is defined by the coefficients in the first column of **V**, denoted by $\mathbf{v}_1$. The variance of $\mathbf{y}_1$ will be maximized. The restriction on the quantities in $\mathbf{v}_1$ is that the sum of the squares of the coefficients be equal to unity.
So the problem can be stated as: "maximize $\mathbf{v}'_1\ \mathbf{R}\ \mathbf{v}_1$ subject to $\mathbf{v}'_1\ \mathbf{v}_1 = 1$"
Incorporating Lagrange multipliers, taking partial derivatives with respect to $\mathbf{v}_1$, setting them to zero and performing some arduous algebra, yields

$$(\mathbf{R}-\lambda\mathbf{I}) = 0$$

This is known as the "problem of the eigenstructure" of **R**.
The characteristic equation of **R** is a polynomial of degree *p*, which is obtained by expansion of the determinant of

$$|\mathbf{R}-\lambda\mathbf{I}| = 0$$

and solving for the roots of λ.
Of special interest is the largest eigenvalue λ_1 and its associated eigenvector $\mathbf{v}_1$
λ_1 is the variance of the normalized linear component **Z** that has maximum variance.

There exist some interesting relationships:

1 $\displaystyle\sum_{i=1}^{p} \lambda_i = trace(\mathbf{R}) = p$

2 $\displaystyle\prod_{i=1}^{p} |\lambda_i| = |\mathbf{R}|$

3 Let **L** be a diagonal matrix with λ_1 in the jth position on the diagonal. Then the full eigenstructure of **R** is

$$\mathbf{RV = VL}$$

where

$$\mathbf{V'V = V\ V' = I}$$

and $\mathbf{V'RV = L = D}_j$

The primary interpretative device in principal component analysis is the *factor structure*

$$\mathbf{S = VL}^{1/2}$$

S is a matrix whose elements are the correlations between the principal components and the variables. If we retain, for example, two eigen values, meaning there are two principal components, then the S matrix consists of two columns and p (number of variables) rows. Consider for example the following table:

	Principal	Component
Var	1	2
1	r_{11}	r_{12}
2	r_{21}	r_{22}
3	r_{31}	r_{32}
4	r_{41}	r_{42}

If this correlation matrix,(i.e. the *factor structure matrix*), does not help much in the interpretation, it is possible to rotate the axis of the principal components. This may result in a polarization of the correlation coefficients. A detailed explanation of principal components and rotation can be found in Harman (1967) or Cooley and Lohnes (1972).

A measure of how well the selected factors (principal components) "explained" the variance of each of the variables is given by a statistic called *communality.*
This is defined by:

$$h_k^2 = \sum_{i=}^{k} \mathbf{S}_{ki}^2$$

That is: the square of the correlation of variable k with factor i gives the part of the variance of the variable accounted for by that factor. The sum of these squares for n factors is the communality, or explained variance for that variable (row).

The primary device that enables us to plot the principal factors the matrix of *factor score coefficients*

$$\mathbf{B} = \mathbf{V}\mathbf{L}^{-1/2}$$

Finally, the factors scores are calculated from:

$$\mathbf{F} = \mathbf{Z}\mathbf{B}$$

In summary, **Z** is the matrix of the standardized original data matrix. **L** is a diagonal matrix, where the diagonal elements are the eigenvalues of **R**, the correlation matrix of **Z**. **V** is the matrix of eigen vectors, F are the transformed data that can be plotted

Example

The datafile is *test.dat* which was also used in the MANOVA example.

```
 ***************************************************
  *       Principal Components Analysis          *
  ***************************************************

CORRELATION MATRIX OF THE VARIABLES

  1.000 -0.730 -0.192 -0.421 -0.639 -0.398
 -0.730  1.000  0.274  0.173  0.815  0.572
 -0.192  0.274  1.000  0.238  0.304  0.322
 -0.421  0.173  0.238  1.000  0.142  0.229
 -0.639  0.815  0.304  0.142  1.000  0.667
 -0.398  0.572  0.322  0.229  0.667  1.000

   EIGENVALUES     PERCENT   CUM.PCT  MULTIPLE CORRELATION
   1   3.194        53.228    53.228          0.646
   2   1.016        16.934    70.162          0.748
   3   0.870        14.498    84.660          0.158
   4   0.549         9.153    93.813          0.296
   5   0.208         3.459    97.271          0.738
   6   0.164         2.729   100.000          0.490

  How many factors should be retained? (enter 0 to quit):   4

  The sum of the eigenvalues     =   6.000
  The product of the eigenvalues =   0.053

  BARTLETT'S SPHERICITY TEST
  It tests if the population correlation matrix is an identity matrix.
  A high Chisquare probability rejects this hypothesis.
  A low (< .9) probability means further factoring is not needed..

 Chisquare          DF      Probability
  77.030            15         1.000
  31.037            10         0.999
  22.479            6          0.999
  10.556            3          0.986
   0.329            1          0.427

     FACTOR STRUCTURE
   1   0.811 -0.054  0.411  0.282
   2  -0.888 -0.263 -0.090 -0.175
   3  -0.461  0.431  0.721 -0.282
   4  -0.413  0.812 -0.338  0.188
   5  -0.888 -0.296  0.044  0.005
   6  -0.759 -0.107  0.238  0.569

  The number of factors =  4

     FINAL COMMUNALITY
```

1	0.909
2	0.896
3	0.999
4	0.979
5	0.878
6	0.968

COEFFICIENT MATRIX

0.568	-0.194	0.088	0.430
-0.467	-0.176	0.009	-0.092
0.061	-0.070	1.056	-0.169
0.147	0.945	-0.070	0.056
-0.279	-0.196	-0.006	0.220
0.332	0.083	-0.147	1.038

VARIMAX ROTATION

Initial Varimax Criterion: 0.1825
Final Varimax Criterion: 0.4871
Convergence is achieved at cycle: 5

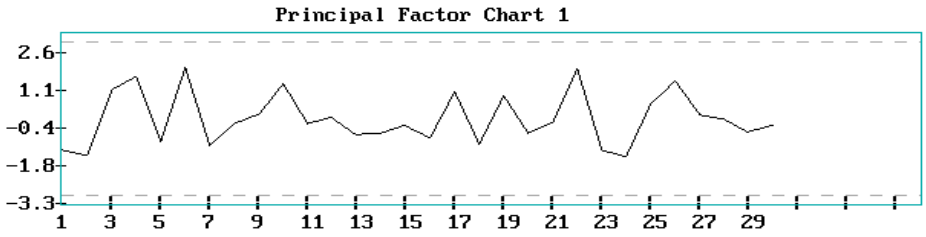
FINAL FACTOR STRUCTURE

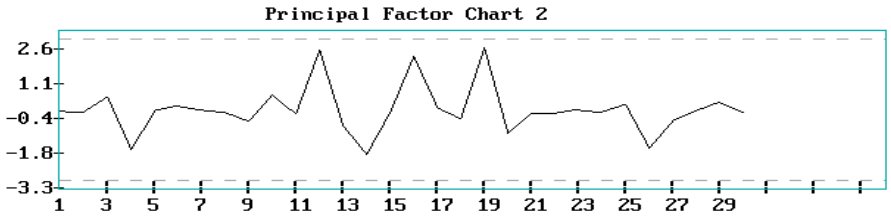
	1	2	3	4	
1	0.8816	-0.3589	-0.0294	-0.0479	factor 1 represents variables 1, 2
2	-0.8771	-0.0011	0.1336	0.3301	factor 2 represents variable 4
3	-0.1194	0.1115	0.9761	0.1378	factor 3 represents variable 3
4	-0.1223	0.9716	0.1110	0.0879	factor 4 represents variable 6
5	-0.7620	-0.0456	0.1582	0.5194	
6	-0.2943	0.1242	0.1430	0.9198	

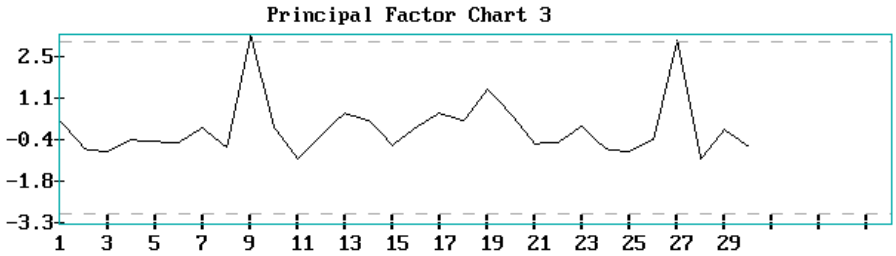
FACTOR SCORES

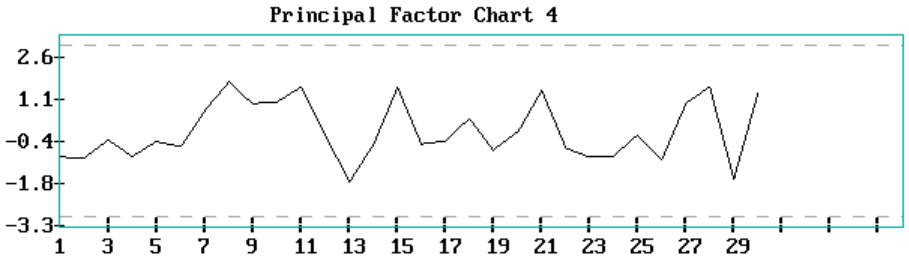
1	-1.218	-0.032	0.246	-0.874
2	-1.417	-0.083	-0.727	-0.918
3	1.102	0.535	-0.834	-0.298
4	1.612	-1.658	-0.380	-0.895
5	-0.882	-0.039	-0.425	-0.388
6	1.997	0.157	-0.522	-0.540
7	-1.036	0.012	-0.003	0.748
8	-0.226	-0.120	-0.637	1.743
9	0.141	-0.495	3.233	0.955
10	1.387	0.605	0.030	1.004
11	-0.207	-0.138	-1.074	1.505
12	0.051	2.504	-0.251	-0.135
13	-0.633	-0.676	0.544	-1.798
14	-0.577	-1.883	0.272	-0.461
15	-0.251	-0.110	-0.609	1.517
16	-0.748	2.237	0.033	-0.495
17	1.052	0.093	0.535	-0.366
18	-0.998	-0.376	0.276	0.394
19	0.880	2.616	1.387	-0.687
20	-0.571	-0.994	0.465	-0.053

21	-0.177	-0.172	-0.559	1.423
22	1.948	-0.155	-0.499	-0.559
23	-1.191	0.005	0.075	-0.876
24	-1.454	-0.109	-0.728	-0.889
25	0.593	0.207	-0.816	-0.144
26	1.475	-1.611	-0.401	-0.999
27	0.168	-0.458	3.062	0.953
28	-0.026	-0.058	-1.076	1.504
29	-0.517	0.296	-0.040	-1.661
30	-0.277	-0.099	-0.580	1.290


Principal Factor Chart 1
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 2
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 3
2.5
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 4
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29

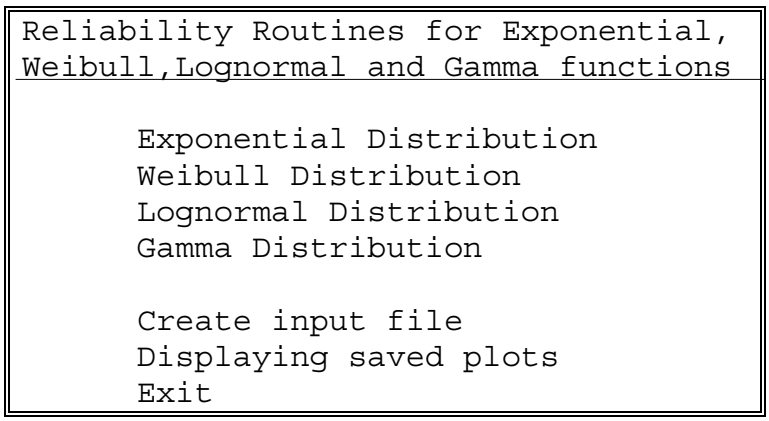
Chapter 9

RELIABILITY

This chapter deals with reliability issues.

MAIN MENU

Move the mouse to the desired line and click the left button



EXPONENTIAL DISTRIBUTION MENU

Move the mouse to the desired line and click the left button

Reliability Routines, using the Exponential Distribution
Duane Analysis Planning Experiments using the Exponential Distribution Analyzing Experiments using the Exponential Distribution Operating Curve (OC) for Exponential Test Plans Ratio of two exponential Distributions Bayesian estimates Create input file Displaying saved plots Exit.

Properties of the Exponential Distribution

CDF: $F(t) = 1 - e^{-\lambda t}$

RELIABILITY: $R(t) = e^{-\lambda t}$

PDF: $f(t) = \lambda e^{-\lambda t}$

MEAN: $1/\lambda$

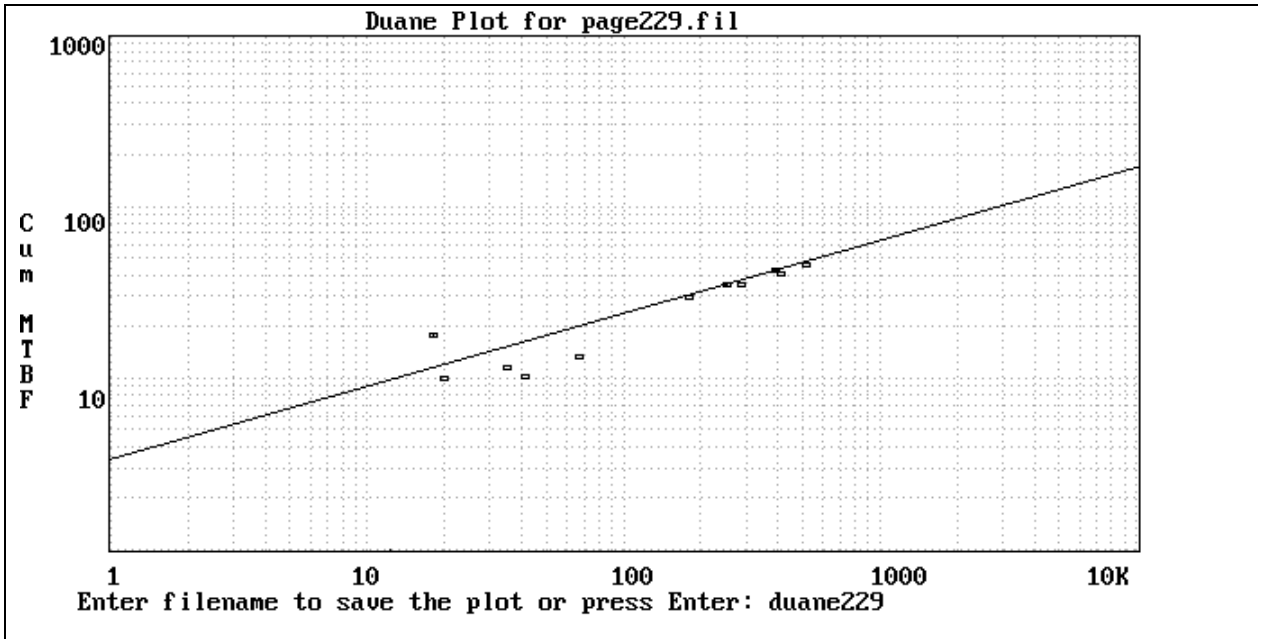
MEDIAN: $\frac{\ln 2}{\lambda} \cong \frac{.693}{\lambda}$

VARIANCE: $\frac{1}{\lambda^2}$

FAILURE RATE: $h(t) = \lambda$

Examples of the Exponential Routines:

```
**********************************************************************
*                                                                    *
*                          DUANE ANALYSIS                            *
*                                                                    *
*  This will output the following:                                   *
*                                                                    *
*  1 Plots:                                                          *
*     Cumulative Fails versus Time                                   *
*     Interarrival Times versus fail numbers                         *
*     Reciprocals of interarrival Times versus fail numbers          *
*     Duane's MTBF vesus Time on log-log scale                       *
*                                                                    *
*  2 Significance tests of trends, using the Reverse Arrangements   *
*                                                                    *
*  3 Estimate of growth slope b and intercept a                      *
*                                                                    *
*  4 The end of test MTBF with confidence bounds                     *
*                                                                    *
**********************************************************************

 Enter filename or press Enter to quit: page229.fil

       This file contains 11 times to failure:
   18  20  35  41  67  180  252  287  390  410  511

Enter test duration or press Enter for the last failure time :
550
Enter α, for a 100(1-α) 1-sided confidence bound : .2

     Some meaningful statistics:

  BETA:  0.4269        ALPHA: 0.2957

  The test was stopped at Time:  550
  The number of failures is   :   11

  The MTBF at end of test = :   87.244
  The approximate  .8  lower bound =     :   55.43
  The approximate  .8  upper bound =     :  117.78

  The total possible number of reversals =  55
  The observed number of reversals =        40
  Evidence of improvement at a confidence level of:   0.9875
```


Duane Plot for page229.fil
1000
100
10
Cum MTBF
1
10
100
1000
10K
Enter filename to save the plot or press Enter: duane229

```
**********************************************************
*  Planning Experiments using the Exponential Distribution *
**********************************************************

           The following cases are presented:

         A  One Repairable System on Test

  1  Determine Minimum Test Length, (0 failures allowed).
  2  Determine T, the duration of the test in hours.
  3  Determine R, the number of allowed failures.

         B  Multiple  Units on Test

  4  Determine Minimum Test Length, (0 failures allowed).
  5  Determine Minimum Sample Size. (0 failures allowed).
  6  Determine N, the sample size.
  7  Determine T, the duration of the test in hours.
  8  Determine R, the number of allowed failures.

 Make your selection by typing the appropriate number (1-8): 1

 Enter the specified MTBF in hours:     500
   Enter α, the risk level:  .2

   The Minimum Test Length in hours is:      805
```

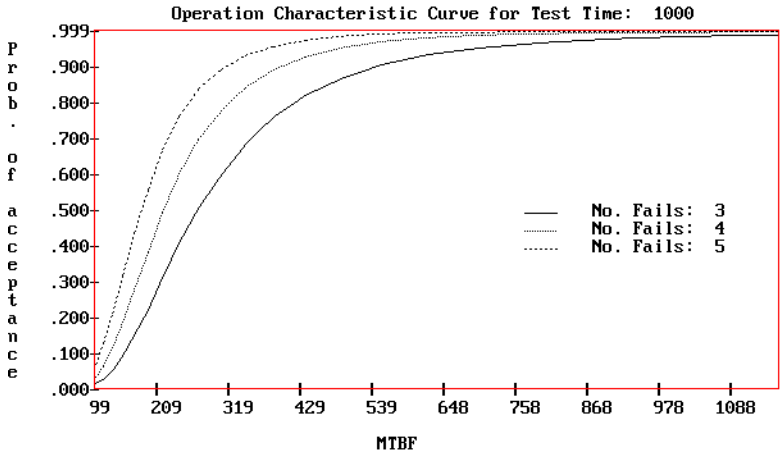
```
 Make your selection by typing the appropriate number (1-8): 2

 Determine T, the duration of the test in hours.

 Enter the specified MTBF in hours:     500
   Enter the number of expected failures: 1
   Enter α, the risk level:   .2

   The Test Length, T, is               1492.40
   The estimated MTBF (N*T/#fails) is   1492.40

 Make your selection by typing the appropriate number (1-8): 7
 Determine T, the duration of the test in hours.

 Enter number of samples:                10
 Enter the specified MTBF in hours:     500
 Enter the number of expected failures:   2
```

```
  Enter α, the risk level:  .2

  The Test Length, T, is                 213.52
  The estimated MTBF (N*T/#fails) is   1067.59
```

```
*****************************************************************
*   Analyzing Experiments using the Exponential Distribution    *
*****************************************************************

  Enter the observed number of failures  : 1
  Enter the duration of the test in hours: 1200
  Enter α, the risk level:   .2


  The estimated MTBF =    1200.000

  For Time Censored Data:
  The One Sided 80 pct Lower Bound for the MTBF =     402.037

  For Failure Censored Data:
  The One Sided 80 pct Lower Bound for the MTBF =     750.448

  For Time and Failure Censored Data:
  The One Sided 80 pct Upper Bound for the MTBF =    5328.520

  The confidence coefficient for a Two-sided Confidence Interval
  using these limits =  60 pct
```

```
     ********************************************************
     *  SINGLE ACCEPTANCE PLANS FOR EXPONENTIAL MTBF        *
     ********************************************************
You can either design an acceptance plan or plot the OC curve
directly.
Enter time and accept number for OC curve, separated by a comma
or press Enter to design:

   ******************************************************
   *   Calculate an Acceptance plan for exponential MTBF *
   ******************************************************

   Enter the 'good' MTBF : 600
   Enter the 'bad'  MTBF : 200
   The default α and β risks are .2
   Enter α , the producer's risk:.2
   Enter β , the consumer's risk:.2


  A suggested testplan is:
  Test Time in Hours:    1000   Allowed no. of Fails:  3

  Calculated 'good'  MTBF =   434 at α =  0.201
  Calculated 'bad '  MTBF =   182 at β =  0.201
```

```
     ****************************************************
     *  SINGLE ACCEPTANCE PLANS FOR EXPONTIAL MTBF       *
     ****************************************************
 You can either design an acceptance plan or plot the OC curve
directly.
   Enter time  and accept number for OC curve, separated by a comma
   or press Enter to design:    1000,3

   You can generate up to 5 additional OC curves, incremented by 1
fail.
   Howmany extras do you want, press Enter for none: 2
   Wish to see OC coordinates in tabular form ?  y/n:
```


Operation Characteristic Curve for Test Time: 1000
Prob. of acceptance
.999
.900
.800
.700
.600
.500
.400
.300
.200
.100
.000
99
209
319
429
539
648
758
868
978
1088
MTBF
No. Fails: 3
No. Fails: 4
No. Fails: 5

```
     ***********************************************************
     *   Comparison of two Exponential Distributions   *
     ***********************************************************

Enter filename of sample 1 (press Enter to quit) exp1.fil
Enter filename of sample 2                       exp2.fil

  Enter original no. put on test for sample 1 : 10
  Enter original no. put on test for sample 2 : 10

  Theta 1     Theta 2   Calculated F   p-value
  619.714     767.200      0.808        0.348

  A two-sided .95 confidence interval:   0.393  3.896
  accept the null hypothesis of equal means at the .05 % level
```

WEIBULL DISTRIBUTION MENU

Reliability Routines, using the Weibull Distribution
M.L.E two parameter estimation, and Hazard rate with C.I.
Rectification (Regression) method of 2 parameter estimation
Estimation of μ, the shape parameter, given θ, the scale
Estimation of θ, the scale parameter, given μ, the shape
Calculation of PDF, CDF, Reliability and Hazard rate
Goodness of fit test, using the Kolmogorov-Smirnov Method
Random Number Generation Routine

Move the mouse to the desired line and click the left button

Properties of the Weibull Distribution

CDF: $F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$

RELIABILITY: $e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$

PDF: $f(t) = \frac{\gamma}{t}\left(\frac{t}{\alpha}\right)^{\gamma} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}$

FAILURE RATE: $\frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1}$

MEAN: $\alpha\Gamma\left(1+\frac{1}{\gamma}\right)$

MEDIAN: $\alpha(\ln 2)^{\frac{1}{\gamma}}$

VARIANCE: $\alpha^2\Gamma\left(1+\frac{2}{\gamma}\right) - \left[\alpha\Gamma\left(1+\frac{1}{\gamma}\right)\right]^2$

Examples of the Weibull Routines

A number of the terminal examples that follow use the file TEST.DAT. This is a file from "Applied Reliability" by Tobias and Trindade, publisher Chapman and Hall, New York, NY 1995.

Here is the file:

```
.7, 52.7, 129.4, 187.8, 264.4, 272.8, 304.2, 305.1, 309.8,1945.0, 2419.5,
310.5, 404.8, 434.9, 479.2, 525.3, 620.3, 782.8, 1122.0, 1200.8, 1224.1,
1322.7,
2894, 2920.1
```

Notice that input files to the SEMSTAT routines are *column* files. The above row of data is only to save space in this documentation.

```
     ****************************************************************
     *     Weibull Parameter Estimation Using M.L.E Method          *
     ****************************************************************

  You can enter a valid filespec, as long as it has an extension
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? test.dat

     ****************************************************************
     *        Weibull Parameter Estimation, MLE Method              *
     *        The Weibull Model: F(x) = 1 - exp[ - (x^m) /θ]        *
     *        The data are assumed to be of type II censoring       *
     ****************************************************************

For Data File test.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    2920.1000          0.7000        834.7200        858.6343         25

  Enter total sample size (at beginning of test): 50

  M (SHAPE)    =        0.6207
  é (SCALE)    =      199.5052
  C (CHARACTERISTIC LIFE, θ^1/m) =  5078.3002

  Approximate 95% Confidence Limits
  m :       0.3916       0.8497
  é :       0.0000     558.6559
```

```
     ****************************************************************
     *     Weibull Parameter Estimation Using Linear Regression     *
     ****************************************************************
For Data File test.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    2920.1000          0.7000        834.7200        858.6343         25

  Enter total sample size (at beginning of test): 50

  m (SHAPE)    =        0.5502
  θ (SCALE)    =      124.9164
  C (CHARACTERISTIC LIFE, θ^1/m) =  6470.6608
  B (-M LN C) =     -4.8276
```

```
The following analysis is on readout data.
The file in transposed form is:

TIME      24  168  200  400  600  1000  1500  2000  2500  3000
# FAILS    1    2    1    6    5     2     4     1     1     2

     ****************************************************************
     *     Weibull Parameter Estimation Using Linear Regression     *
     ****************************************************************

  You can enter a valid filespec, as long as it has an extension
  A ONE column file contains times-to-failure.
  A TWO column file contains readout times in column 1, and number of failures in column 2
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? readout.dat
For Data File readout.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    3000.0000         24.0000       1139.2000       1059.9284         10
```

```
  Enter total sample size (at beginning of test): 50

  M (SHAPE)    =       0.7807
  θ (SCALE)    =     617.8957
  C (CHARACTERISTIC LIFE, θ^1/m) =  3759.0143
  B (-M LN C) =    -6.4263
```

```
     ****************************************************************
     *     Weibull Shape Parameter Estimation given θ, the scale    *
     ****************************************************************

  You can enter a valid filespec, as long as it has an extension
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? test.dat

     ****************************************************************
     *        The data are assumed to be of type II censoring       *
     ****************************************************************

For Data File test.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    2920.1000          0.7000        834.7200        858.6343         25

  Enter total sample size (at beginning of test): 50
  Enter the value of θ, the known scale parameter: 200

  m (SHAPE)    =       0.6210
  θ (SCALE)    =     200.0000
  C (CHARACTERISTIC LIFE, θ^1/m) =  5074.1378

  Approximate 95% Confidence Limits for m
     0.5701 <  0.6210 <  0.6719
```

```
     ****************************************************************
     *     Weibull Scale Parameter Estimation given m, the shape    *
```

```
     ****************************************************************


For Data File test.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    2920.1000          0.7000        834.7200        858.6343         25

  Enter total sample size (at beginning of test): 50
  Enter value of m, the shape parameter: .65

  Exact 95% Confidence Limits for θ given M
    174.8952 <     249.8507 <     386.2240
```

```
*****************************************************************
     *   Weibull Calculation of PDF, CDF, Hazard and Reliability    *
     *****************************************************************

  You can enter a valid filespec, as long as it has an extension
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? test.dat


     *****************************************************************
     *        The data are assumed to be of type II censoring        *
     *****************************************************************

For Data File test.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    2920.1000          0.7000        834.7200        858.6343         25

  Enter m, the shape parameter: .6
  Enter θ, the scale parameter: 200


   VALUE           PDF          CDF          HAZARD    RELIABILITY

     0.70       0.00345       0.00403       0.00346       0.99597
    52.70       0.00058       0.05253       0.00061       0.94747
   129.40       0.00039       0.08835       0.00043       0.91165
      .            .             .             .             .
      .            .             .             .             .
      .            .             .             .             .
  2419.50       0.00008       0.41495       0.00013       0.58505
  2894.50       0.00007       0.44949       0.00012       0.55051
  2920.10       0.00007       0.45123       0.00012       0.54877
```

```
     ************************************************
     *   N RANDOM MUMBERS FROM WEIBULL DISTRIBUTION   *
     *   WITH PARAMETERS m (SHAPE) AND theta (SCALE)  *
     ************************************************
  How many Random Numbers?     100
  Input shape parameter, m:     .6
  Input scale parameter, θ:    200
  Sorted (S) or Unsorted (Press Enter)? : s
  Type 1 to see screen output, or press Enter Enter was pressed)

  The set of Weibul numbers are stored in file WEI.RND
```

```
  If you want to display this file type : T WEI.RND
  The T.EXE routine should be copied from DISK 2 of the
  SEMSTAT software.   Or you could return to the main menu
  of SEMSTAT and click on DATA MANAGEMENT
```

```
     ************************************************************
     *             Kolmogorov-Smirnov Goodness of Fit test      *
     ************************************************************

  You can enter a valid filespec, as long as it has an extension
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the
menu.
? wei.rnd


  Enter value for m,     the shape parameter:  .6
  Enter value for theta, the scale parameter: 20


                       Cumulative
  Data     Relative  Observed  Expected  Absolute
  Value    Frequency Frequency Frequency Difference
-------------------------------------------------
     0.15   0.0200    0.0200    0.0016    0.0184
     1.88   0.0200    0.0400    0.0073    0.0327
    22.82   0.0200    0.0600    0.0321    0.0279
      .       .         .         .         .
      .       .         .         .         .
      .       .         .         .         .
 61483.53   0.0200    0.9600    0.9761    0.0161
 93070.54   0.0200    0.9800    0.9917    0.0117
108879.20   0.0200    1.0000    0.9948    0.0052

         Maximum absolute difference =    0.0744
        (at observation 5762.90, not shown here...)

  The p value >  .2 : ACCEPT H0
  That is, the data could conceivably follow a Weibull distribution.
  with shape parameter:  .6 and scale parameter: 200
```

```
  CONVERSION FROM CHARACTERISTIC LIFE TO THETA (SCALE)

  ENTER M, THE SHAPE PARAMETER: .6
  ENTER CHARACTERISTIC LIFE: 4000
```

```
  THETA =  144.956
```

LOG NORMAL DISTRIBUTION MENU

Reliability Routines, using the Log Normal Distribution
M.L.E two parameter estimation, and Hazard rate with C.I. Rectification (Regression) method of 2 parameter estimation Double Censoring or Truncation Estimation of μ given σ, or of σ given μ Calculation of PDF, CDF, Reliability and Hazard rate Goodness of fit test, using the Kolmogorov-Smirnov Method Random Number Generation Routine Exit

Properties of the Lognormal Distribution

PDF: $f(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln \mathbf{T_{50}})^2}$

CDF: $F(T) = \int_0^T \frac{1}{\sigma t\sqrt{2\pi}} e^{-\left(\frac{1}{2\sigma^2}\right)(\ln t - \ln \mathbf{T_{50}})^2} = \Phi\left(\frac{\ln t / \ln \mathbf{T_{50}}}{\sigma}\right)$

where $\Phi(z)$ is the standard Normal CDF

RELIABILITY: $R(T) = 1 - F(T)$

FAILURE RATE: $h(t) = \frac{f(t)}{R(t)}$

MEAN: $\mathbf{T_{50}} e^{\frac{\sigma^2}{2}}$

MEDIAN: $\mathbf{T_{50}}$

VARIANCE: $\mathbf{T_{50}^2}\, e^{\sigma^2}\,(e^{\sigma^2} - 1)$

Examples of the Lognormal Routines

```
************************************************************************
* MU AND SIGMA MLE ESTIMATES FOR A CENSORED OR TRUNCATED LOGNORMAL SAMPLE *
************************************************************************

  Enter file id (or press Enter to quit): test.dat
  Type 1 for censoring (default) or 2 for truncation:     1
  Type 1 for on the right (default) or 2 for on the left: 1
  Enter total number put on test: 50
  Enter the censoring/truncation point or press Enter failure no. 25

   Type 2 Censoring
  MU =       3165.379
  SIGMA =       2.576

  Approximate 95% Confidence Bands:
  MU    :    1797.343   5574.685
  SIGMA :       2.060      3.093
```

```
  ****************************************************************
  *     Lognormal Parameter Estimation Using Linear Regression   *
  ****************************************************************

You can enter a valid filespec, as long as it has an extension.
A ONE column file contains times-to-failure. A TWO column file
contains readout times in column 1, and number of failures in column 2
If you press the enter key, ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? test.dat


     ****************************************************************
     *        Lognormal Estimation of mu and s via regression       *
     *        The data are assumed to be of type II censoring       *
     ****************************************************************

For Data File test.dat

       MAX            MIN            MEAN         STD.DEV.
NO.ITEMS
    2920.1000         0.7000       834.7200       858.6343        25
```

```
  Enter total sample size (at beginning of test): 50

  MU,   (the median parameter) =   3189.9027
  SIGMA, (the shape parameter) =      2.5936
```

```
     ************************************************************
     *     Lognormal Parameter Estimation Using Linear Regression   *
     ************************************************************

You can enter a valid filespec, as long as it has an extension.
A ONE column file contains times-to-failure. A TWO column file
contains readout times in column 1, and number of failures in column 2
If you press the enter key, ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? readout.dat

     ************************************************************
     *      Lognormal Estimation of mu and sigma via regression     *
     *        The data are assumed to be of type II censoring       *
     ************************************************************

For Data File readout.dat

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
    3000.0000        24.0000      1139.2000      1059.9284        10

  Enter total sample size (at beginning of test): 50

  MU,   (the median parameter) =   2608.2364
  SIGMA, (the shape parameter) =      2.0738
```

```
     ************************************************************
     *   MU AND SIGMA FOR DOUBLY CENSORED LOGNORMAL SAMPLES    *
     ************************************************************

  Enter file id (or press Enter to quit): test.dat
  Enter  left censoring point (hours) : 300
  Enter right censoring point (hours) : 2000
  Enter no. of  left-censored items   : 6
  Enter no. of right-censored items   : 3

Statistics for file test.dat
```

```
       XBAR     STD.DEV  NO.FAILS
     608.5989    0.6153    16

 Total put on test:    25
 MU =       2000.000
 SIGMA =       0.903
```

```
   ************************************************************
   *  MU OR SIGMA MLE ESTIMATES FOR A RIGHT CENSORED OR TRUNCATED  *
   *       LOGNORMAL SAMPLE WHEN ONE OR THE OTHER IS KNOWN         *
   ************************************************************

Enter file id (or press Enter to quit): test.dat
Type 1 if MU is known (default) or 2 for known SIGMA: 2
Enter the value of the known SIGMA: 2.58

MU =       3195.762

Approximate 95% Confidence Bands for the median:
 1764.367 <    3195.762 <    5788.420
```

```
   ************************************************
   *       RANDOM LOGNORMAL NUMBER GENERATOR        *
   ************************************************

How many Random Numbers? 50
Enter the MEDIAN (default = 1) 3000
Enter SIGMA (default = 1)      2.58
Sorted (S) or Unsorted (Press Enter)? : s
Type 1 to see screen output, or press Enter to skip this...


The set of lognormal numbers are stored in file LNORM.RND
```

```
    ************************************************************
    * Kolmogorov-Smirnov Goodness of Fit test for a Lognormal Distn*
    ************************************************************

  You can enter a valid filespec, as long as it has an extension.
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the
menu.
? lnorm.rnd

  Enter value for MU or press Enter for XBAR:  Enter was pressed
  Enter value for SIGMA, or press Enter for S: Enter was pressed

   Log                  Cumulative
   Data    Relative  Observed  Expected  Absolute
   Value   Frequency Frequency Frequency Difference

    -0.17   0.0200    0.0200    0.0030    0.0170
     2.79   0.0200    0.0400    0.0372    0.0028
     3.31   0.0200    0.0600    0.0533    0.0067
      .       .         .         .         .
      .       .         .         .         .
      .       .         .         .         .
    13.17   0.0200    0.9600    0.9432    0.0168
    13.35   0.0200    0.9800    0.9497    0.0303
    15.05   0.0200    1.0000    0.9858    0.0142

         Maximum absolute difference =    0.0680
         (at observation 9.01, not shown here.)

  The p value >  .2 : ACCEPT HO
  That is, the data could conceivably follow a Lognormal distribution.
  with μ:   3983.4439
  and σ:      3.0827
```

GAMMA DISTRIBUTION MENU

Reliability Routines, using the Gamma Distribution
M.L.E and Moments for a and b and an approximate C.I. for a Incomplete Gamma function P(α, β, x) Inverse Gamma for α, β, and probability p Estimation of α and b given two percentiles with their MTBF's Calculation of PDF, CDF, Reliability and Hazard rate Goodness of fit test, using the Kolmogorov-Smirnov Method Random Number Generation Routine Exit.

Properties of the Gamma Distribution

PDF: $f(t,a,b) = \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt}$ **or,**

$$f(t,\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(a)} t^{\alpha-1} e^{\frac{-t}{\beta}}$$ **or,**

$$f(t,\gamma,\beta) = \frac{1}{\beta^\gamma \Gamma(\gamma)} t^{\gamma-1} e^{\frac{-t}{\beta}}$$

CDF: $F(t) = \int_0^t f(t)dt$

RELIABILITY: $R(t) = 1 - F(t)$

FAILURE RATE: $h(t) = \frac{f(t)}{R(t)}$

MEAN: $\frac{a}{b}$ **or** $\alpha\beta$ **or** $\gamma\beta$

VARIANCE: $\frac{a}{b^2}$ **or** $\alpha\beta^2$ **or** $\gamma\beta^2$

Examples of the Gamma Routines

Much of the background for these programs come from the book by
J.F. Lawless, *Statistical Models and methods for Lifetime Data*
John Wiley and Sons, New York (1982)

The input file named LAWLESS.FIL is in column form.
Its horizontal representation is
152 152 115 109 137 88 94 77 160 165 125 40 128 123 136 101 62 153 83 69

```
*************************************************************
*  PARAMETER ESTIMATES FOR THE UNCENSORED GAMMA DISTRIBUTION *
*************************************************************
  Enter filename (or press Enter to quit): lawless.fil

 Approximate values for the estimators of the 2 parameter Gamma
Distribution
  α:     12.8932
  k:      8.7992
  β:      0.0776   note: α = 1/β.  k is sometimes denoted as β

  Moment estimates:
  α:     11.2904
  k:     10.0484
  β:      0.0886


  Approximate 95% Confidence Limits for k
   4.0057 <  k  < 14.3609

  Approximate 95% Confidence Limits for α
   4.6560 < α <   21.1305
```

```
************************************
   Incomplete Gamma Function P(α,β,x)
 ************************************

 Enter upper limit, x: 15
 Enter α (0 to quit) : 5
 Enter β : 2

 Incomplete Gamma:      0.867939
 Ln Γ(α)          :     3.178053
```

```
*************************************************************
*   Inverse Gamma for α,β and probability p                 *
*   The form is: f(x) = [1/(β ^α)(α-1)!] [x ^ (α-1)] exp(-x/β) *
*************************************************************

  Input p : .868
  Input α : 5
  Input β : 2

    X   =   15.002
    difference between computed and inputted p's: 0.11925D-04
```

```
*************************************************************
*   Estimates a and b of the incomplete Gamma distribution  *
*   when two percentiles with their MTBF values are given,  *
*   The form is: f(x) = [1/(β ^α)( α-1)!] [x ^ (α-1)] exp(-x/β) *
*************************************************************

  Input Upper Percentile  : 90
  Input Upper MTBF        : 500
  Input Lower Percentile  : 10
  Input Lower MTBF        : 200

  ALPHA =      8.1890
  BETA  =   2401.0000

   OBSERVED   COMPUTED
     RATIO      RATIO
     2.5000     2.4996
```

```
************************************************
*         RANDOM GAMMA NUMBER GENERATOR        *
************************************************

  How many Random Numbers? 100
```

```
  Enter the ALPHA parameter 5
  Enter the BETA  parameter 2
  Sorted (S) or Unsorted (Press Enter)? : s
  Type 1 to see screen output, or press Enter to skip this...


  The set of gamma numbers are stored in file GAMMA.RND

     ****************************************************************
     *    Gamma Calculation of PDF, CDF, Hazard and Reliability     *
     ****************************************************************

  You can enter a valid filespec (containing times-to-failure).
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): To return, press F10.

? gamma.rnd


For Data File gamma.rnd

       MAX             MIN             MEAN          STD.DEV.
NO.ITEMS
      24.5885          4.7801         12.2809           4.6477       100

 Enter α (0 to quit) : 5
 Enter β : 2

OBS  VALUE         PDF          CDF          HAZARD    RELIABILITY

1    4.78       0.06229      0.09462       0.06880       0.90538
2    5.00       0.06681      0.10885       0.07497       0.89115
3    5.17       0.07017      0.12060       0.07980       0.87940
      .           .            .             .             .
      .           .            .             .             .
      .           .            .             .             .
99  21.99       0.00511      0.98484       0.33721       0.01516
99  22.73       0.00403      0.98820       0.34183       0.01180
100 24.59       0.00218      0.99382       0.35243       0.00618
```

```
     ************************************************************
     *             Kolmogorov-Smirnov Goodness of Fit test      *
     ************************************************************

  You can enter a valid filespec, as long as it has an extension.
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the
menu.
? lawless.fil


  Enter value for α, or press Enter to quit... : 12.9
  Enter value for β: 8.8
-------------------------------------------------
                       Cumulative
  Data     Relative  Observed  Expected  Absolute
  Value    Frequency Frequency Frequency Difference
-------------------------------------------------
    40.00   0.0500    0.0500    0.0010    0.0490
    62.00   0.0500    0.1000    0.0303    0.0697
    69.00   0.0500    0.1500    0.0600    0.0900
    77.00   0.0500    0.2000    0.1124    0.0876
    83.00   0.0500    0.2500    0.1651    0.0849
    88.00   0.0500    0.3000    0.2170    0.0830
    94.00   0.0500    0.3500    0.2868    0.0632
   101.00   0.0500    0.4000    0.3753    0.0247
   109.00   0.0500    0.4500    0.4794    0.0294
   115.00   0.0500    0.5000    0.5555    0.0555
   123.00   0.0500    0.5500    0.6496    0.0996
   125.00   0.0500    0.6000    0.6714    0.0714
   128.00   0.0500    0.6500    0.7026    0.0526
   136.00   0.0500    0.7000    0.7766    0.0766
   137.00   0.0500    0.7500    0.7848    0.0348
   152.00   0.0500    0.8000    0.8835    0.0835
   152.00   0.0500    0.8500    0.8835    0.0335
   153.00   0.0500    0.9000    0.8885    0.0115
   160.00   0.0500    0.9500    0.9189    0.0311
```

```
  165.00    0.0500    1.0000    0.9360    0.0640

        Maximum absolute difference =    0.0996
        at data value = 123

 The p value >  .2 : ACCEPT HO
 That is, the data could conceivably follow a Gamma distribution.
 with α = 12.9 and β = 8.8
```

Chapter 9A.

Classification Analysis

This is a technique to obtain information from data that originate from multiple groups for the purpose of classifying individuals into one of these groups.
Broadly speaking, we are estimating the probability that subject i is a member of population j. We are examining a set of hypotheses pertaining to the group membership of subject i, given g groups. Each subject (observation) consists of p measurements.
There are several methods in the multivariate literature, this program brings two of them, Anderson's method and Geisser' method.

The Anderson method first evaluates each of a calculated set o*f linear functions*, one corresponding to each group, and then assigns the subject to the group that exhibits the largest probability, which in turn is associated with the largest linear function.
He employs a test statistic which is due to Mahalonobis, which is similar to the well known Hotelling T statistic, calculated as follows:

$$\mathbf{M} = \sum_{\mathbf{i=1}}^{\mathbf{g}} n_i \left(\mathbf{m_i} - \mathbf{m}\right)' \mathbf{D}^{-1} \left(\mathbf{m_i} - \mathbf{m}\right)$$

where

m is the grand centroid
$\boldsymbol{m_i}$ is the sample centroid of group j, j = 1, 2, ... g
D is the sample dispersion matrix
n_i is the sample size per group

M follows a chi-square distribution with p(g-1) degrees of freedom
g is the number of groups, p is the number of elements per subject..

It can be shown that each group has a constant term:

$$C_{0j} = -.5\ \boldsymbol{m}_j' \boldsymbol{D}^{-1} \boldsymbol{m}_j \qquad j = 1,2,\cdots,g$$ and p coefficients:

$$C_{ij} = \boldsymbol{D}^{-1} \boldsymbol{m}_j \qquad i = 1,\ldots p \qquad j = 1,\ldots g$$

To determine the maximum likelihood of membership for subject i, these constants and coefficient vectors of each of the g groups is used with $\boldsymbol{X}_{ij}$
That is, for each ith subject in each jth group the following calculations are performed:

the *discrimant function* is computed by:

$$f_j = C_{0j} + C_{ij} X_{ij}$$

The assumption is made that the distribution for each group population is multivariate normal, with equal dispersion.

The probability associated with the largest discriminant function is

$$p_i = 1 / \sum_{k=1}^{g} e^{(f_k - f_l)}$$

where f_l = the value of the largest discriminant function.

The Anderson method works best for reasonably large sample sizes, N > 30.

The Geisser method is based on small sample theory.

For each ith subject in each jth group, these calculations are performed:

$$f_{ij} = q_j \left[\frac{n_j}{n_j + 1} \frac{n_j (\mathbf{x}_j - \mathbf{m}_j)' \mathbf{D}^{-1} (\mathbf{x}_j - \mathbf{m}_j)}{n_j + 1} \right]^{-.5(N-g+1)}$$

where

$\boldsymbol{m}_j$ is the sample centroid of group j
$\boldsymbol{D}$ is the sample dispersion matrix
n_j is the group sample size
N is the total sample size
g is the number of groups

p is the number of elements per subject
q_j is the prior probability of assigment to group j

From the f_{ij} the probabilities are computed as follows:

$$P_{ij} = \frac{f_{ij}}{\sum_{k=1}^{g} f_{ik}} \qquad i = 1,\ldots p \qquad j = 1,\ldots g$$

The subject is assigned to the group with the largest probability

Example 1

```
        ****************************************************
        *          Classification Analysis                 *
        *            Anderson's Method                     *
        ****************************************************

  You can enter a valid filespec, as long as it has an extension, or you
  can select a file extension to search for files of particular interest.
  If you press the enter key, ALL file names are displayed.
  Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? test.dat

  The Mahalonobis Chi-Square Statistic is:   12.781
  The degrees of freedom are:                18
  The level of significance (p value) is:     0.805

                                   MEANS
          1          2          3          4          5          6
    1  7.875D+00  7.500D+00  4.625D+00  7.250D+00  1.850D+01  8.875D+00
    2  7.143D+00  8.571D+00  9.571D+00  7.857D+00  2.014D+01  1.257D+01
    3  7.857D+00  7.857D+00  8.857D+00  9.286D+00  1.743D+01  1.014D+01
    4  7.750D+00  8.000D+00  6.750D+00  7.375D+00  2.138D+01  9.250D+00

GRAND  7.667D+00  7.967D+00  7.333D+00  7.900D+00  1.940D+01  1.013D+01


      VARIANCE-COVARIANCE MATRIX

  1.962D+01 -1.116D+01 -5.215D+00 -6.099D+00 -2.275D+01 -9.541D+00
 -1.116D+01  1.195D+01  5.618D+00  1.918D+00  2.261D+01  1.067D+01
 -5.215D+00  5.618D+00  3.946D+01  3.937D+00  1.623D+01  9.345D+00
 -6.099D+00  1.918D+00  3.937D+00  9.833D+00  4.622D+00  3.838D+00
 -2.275D+01  2.261D+01  1.623D+01  4.622D+00  6.279D+01  3.018D+01
 -9.541D+00  1.067D+01  9.345D+00  3.838D+00  3.018D+01  2.957D+01

      CONSTANT AND COEFFICIENTS

 -2.849D+01  2.639D+00  2.122D+00 -1.717D-01  1.912D+00  5.848D-01 -4.048D-01
 -2.921D+01  2.619D+00  2.252D+00 -4.816D-02  1.883D+00  4.373D-01 -2.178D-01
 -3.186D+01  2.744D+00  2.396D+00 -6.457D-02  2.133D+00  4.262D-01 -3.272D-01
 -3.082D+01  2.719D+00  2.039D+00 -1.335D-01  1.945D+00  7.168D-01 -4.876D-01

  CASE      LARGEST      ASSOCIATED   GROUP IT
   NO.      FUNCTION     PROBABILITY  BELONGS TO

  GROUP  1
    1        25.392         0.381        4
    2        32.337         0.370        1
    3        18.597         0.363        1
    4        24.733         0.442        1
    5        30.164         0.345        1
    6        40.487         0.442        3
    7        22.380         0.318        2
```

8	34.796	0.293	2

```
GROUP  2
  1         39.411          0.510         2
  2         32.275          0.501         3
  3         38.169          0.348         4
  4         16.308          0.431         3
  5         26.922          0.443         4
  6         19.125          0.364         2
  7         34.577          0.285         2

GROUP  3
  1         39.383          0.676         3
  2         18.049          0.466         2
  3         29.163          0.546         2
  4         39.688          0.667         3
  5         23.373          0.306         2
  6         37.769          0.330         4
  7         38.354          0.390         3

GROUP  4
  1         24.809          0.337         4
  2         32.922          0.375         1
  3         26.700          0.623         4
  4         22.094          0.457         1
  5         39.022          0.522         2
  6         35.413          0.341         4
  7         33.159          0.431         4
  8         34.362          0.278         1
Number of correct classifications =  14
This is  46.667 percent
```

Example 2

****
* **Classification Analysis** *
* **Geisser's Method** *
****

You can enter a valid filespec, as long as it has an extension, or you
can select a file extension to search for files of particular interest.
If you merely press the enter key (□ÄÙ), ALL file names are displayed.
Enter FILESPEC or EXTENSION (1-3 letters): F10 to return to the menu.
? test.dat

The Mahalonobis Chi-Square Statistic is: 12.781
The degrees of freedom are: 18
The level of significance (p value) is: 0.805

MEANS

	1	2	3	4	5	6
1	7.875D+00	7.500D+00	4.625D+00	7.250D+00	1.850D+01	8.875D+00
2	7.143D+00	8.571D+00	9.571D+00	7.857D+00	2.014D+01	1.257D+01
3	7.857D+00	7.857D+00	8.857D+00	9.286D+00	1.743D+01	1.014D+01
4	7.750D+00	8.000D+00	6.750D+00	7.375D+00	2.138D+01	9.250D+00
GRAND	7.667D+00	7.967D+00	7.333D+00	7.900D+00	1.940D+01	1.013D+01

VARIANCE-COVARIANCE MATRIX

1.962D+01	-1.116D+01	-5.215D+00	-6.099D+00	-2.275D+01	-9.541D+00
-1.116D+01	1.195D+01	5.618D+00	1.918D+00	2.261D+01	1.067D+01
-5.215D+00	5.618D+00	3.946D+01	3.937D+00	1.623D+01	9.345D+00
-6.099D+00	1.918D+00	3.937D+00	9.833D+00	4.622D+00	3.838D+00
-2.275D+01	2.261D+01	1.623D+01	4.622D+00	6.279D+01	3.018D+01
-9.541D+00	1.067D+01	9.345D+00	3.838D+00	3.018D+01	2.957D+01

CASE NO.	LARGEST PROBABILITY	GROUP IT BELONGS TO
GROUP 1		
1	0.360	4
2	0.343	1
3	0.337	1
4	0.395	1
5	0.334	1
6	0.395	3
7	0.308	2
8	0.286	2
GROUP 2		
1	0.432	2
2	0.457	3
3	0.328	4

4	0.384	3
5	0.404	4
6	0.342	2
7	0.280	2

```
GROUP  3
  1          0.571         3
  2          0.422         2
  3          0.481         2
  4          0.536         3
  5          0.298         2
  6          0.316         4
  7          0.359         3

GROUP  4
  1          0.324         4
  2          0.348         1
  3          0.529         4
  4          0.407         1
  5          0.445         2
  6          0.323         4
  7          0.396         4
  8          0.274         1
Number of correct classifications =  14
This is  46.667 percent
```

Principal Components

Principal components approach consists of transforming the *p*-dimensional data into a lower dimensional set of data (sometimes bivariate or even univariate). This is accomplished by setting up meaningful weighted linear combinations of the *p*-dimensions.
Those new variables are called principal components.These were derived by Harold Hotelling in 1933. It works as follows:
Let $\left(X_i^1, X_i^2, \cdots, X_i^p\right)$ be the *i*-th *p*-dimensional observation in the original data.

Now we create a new *p*-dimensional observation $\left(Z_i^1, Z_i^2, \cdots, Z_i^p\right)$ such that the *i*-th

Variable in the **Z**'s is a linear combination of the deviations of the original *p* dimensions from their targets.

$$\mathbf{Z}_i^{'} = \sum_{j=1}^{p} c_{ij}\left(\mathbf{X}_i^j - \mathbf{m}^j\right)$$

For a process with multidimensional data, its overall variance is defined as the sum of the variances of the *p* variables, i.e. the sum of the diagonal elements of the covariance matrix **Σ**.
In the interest of clarity and completeness a brief overview of the principal components methodology will be given below. This method will reduce the number of parameters that has to be estimated in a *p*-element vector. In general, the number of estimators is:

p	means
p	variances
$(p^2-p)/2$	covariances

So for p = 4 there are 14 estimators, for p = 6 there are 27 and for p = 10 there are 65 estimators. Would not it be nice if this number can be reduced? The answer is YES, and it can be done. The method of principal components will accomplish this reduction. It begins with the standardization of the vector variable., that is subtract the mean and divide each element by the standard deviation The vector means will all be zero.. For the standardized vector **Z**, the dispersion matrix becomes the correlation matrix

$$\mathbf{R} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^{'}$$

Next an uncorrelated vector is constructed. This is accomplished by transforming the data. To produce a transformation vector for **y**, is saying that we want a **V** matrix such that its dispersion matrix $\mathbf{D}_y$ is diagonal. This means that

$$\mathbf{y = V'Z}$$

where **V** is a *p x n* coeffcient matrix that carries the *p*-element vector into the *n*-element derived variable **y.**

The centroid of y is

$$\mathbf{m_y = V'm_z = 0}$$

and its dispersion is

$$\mathbf{D_y = V'D_z = V'RV}$$

where **R** is the correlation matrix for **Z**.
The transformation "find **V** such that $\mathbf{D}_y$ is a diagonal matrx" is called an orthogonalizing transformation. There exists an infinity of values for **V** that will yield a diagonal $\mathbf{D}_y$ for any correlation matrix **R**, so a restriction is imposed on the problem. The first element of **y** is called the *first principal component* and is defined by the coefficients in the first column of **V**, denoted by $\mathbf{v}_1$. The variance of $\mathbf{y}_1$ will be maximized. The restriction on the quantities in $\mathbf{v}_1$ is that the sum of the squares of the coefficients be equal to unity.
So the problem can be stated as: "maximize $\mathbf{v}'_1\ \mathbf{R}\ \mathbf{v}_1$ subject to $\mathbf{v}'_1\ \mathbf{v}_1 = 1$"
Incorporating Lagrange multipliers, taking partial derivatives with respect to $\mathbf{v}_1$, setting them to zero and performing some arduous algebra, yields

$$(\mathbf{R}-\lambda\mathbf{I}) = 0$$

This is known as the "problem of the eigenstructure" of **R**.
The characteristic equation of **R** is a polynomial of degree *p*, which is obtained by expansion of the determinant of

$$|\mathbf{R}-\lambda\mathbf{I}| = 0$$

and solving for the roots of λ.
Of special interest is the largest eigenvalue λ_1 and its associated eigenvector $\mathbf{v}_1$
λ_1 is the variance of the normalized linear component **Z** that has maximum variance.

There exist some interesting relationships:

1 $\displaystyle\sum_{i=1}^{p} \lambda_i = trace(\mathbf{R}) = p$

2 $\displaystyle\prod_{i=1}^{p} |\lambda_i| = |\mathbf{R}|$

3 Let **L** be a diagonal matrix with λ_1 in the jth position on the diagonal. Then the full eigenstructure of **R** is

$$\mathbf{RV = VL}$$

where

$$\mathbf{V'V = V\ V' = I}$$

and $\mathbf{V'RV = L = D}_j$

The primary interpretative device in principal component analysis is the *factor structure*

$$\mathbf{S = VL}^{1/2}$$

S is a matrix whose elements are the correlations between the principal components and the variables. If we retain, for example, two eigen values, meaning there are two principal components, then the S matrix consists of two columns and p (number of variables) rows. Consider for example the following table:

	Principal	Component
Var	1	2
1	r_{11}	r_{12}
2	r_{21}	r_{22}
3	r_{31}	r_{32}
4	r_{41}	r_{42}

If this correlation matrix,(i.e. the *factor structure matrix*), does not help much in the interpretation, it is possible to rotate the axis of the principal components. This may result in a polarization of the correlation coefficients. A detailed explanation of principal components and rotation can be found in Harman (1967) or Cooley and Lohnes (1972).

A measure of how well the selected factors (principal components) "explained" the variance of each of the variables is given by a statistic called *communality.*
This is defined by:

$$h_k^2 = \sum_{i=}^{k} \mathbf{S}_{ki}^2$$

That is: the square of the correlation of variable k with factor i gives the part of the variance of the variable accounted for by that factor. The sum of these squares for n factors is the communality, or explained variance for that variable (row).

The primary device that enables us to plot the principal factors the matrix of *factor score coefficients*

$$\mathbf{B} = \mathbf{V}\mathbf{L}^{-1/2}$$

Finally, the factors scores are calculated from:

$$\mathbf{F} = \mathbf{Z}\mathbf{B}$$

In summary, **Z** is the matrix of the standardized original data matrix. **L** is a diagonal matrix, where the diagonal elements are the eigenvalues of **R**, the correlation matrix of **Z**. **V** is the matrix of eigen vectors, F are the transformed data that can be plotted

Example

The datafile is *test.dat* which was also used in the MANOVA example.

```
 ***************************************************
  *       Principal Components Analysis          *
  ***************************************************

CORRELATION MATRIX OF THE VARIABLES

  1.000 -0.730 -0.192 -0.421 -0.639 -0.398
 -0.730  1.000  0.274  0.173  0.815  0.572
 -0.192  0.274  1.000  0.238  0.304  0.322
 -0.421  0.173  0.238  1.000  0.142  0.229
 -0.639  0.815  0.304  0.142  1.000  0.667
 -0.398  0.572  0.322  0.229  0.667  1.000

   EIGENVALUES     PERCENT   CUM.PCT  MULTIPLE CORRELATION
   1   3.194        53.228    53.228          0.646
   2   1.016        16.934    70.162          0.748
   3   0.870        14.498    84.660          0.158
   4   0.549         9.153    93.813          0.296
   5   0.208         3.459    97.271          0.738
   6   0.164         2.729   100.000          0.490

  How many factors should be retained? (enter 0 to quit):   4

  The sum of the eigenvalues     =   6.000
  The product of the eigenvalues =   0.053

  BARTLETT'S SPHERICITY TEST
  It tests if the population correlation matrix is an identity matrix.
  A high Chisquare probability rejects this hypothesis.
  A low (< .9) probability means further factoring is not needed..

 Chisquare          DF      Probability
  77.030            15         1.000
  31.037            10         0.999
  22.479            6          0.999
  10.556            3          0.986
   0.329            1          0.427

     FACTOR STRUCTURE
   1   0.811 -0.054  0.411  0.282
   2  -0.888 -0.263 -0.090 -0.175
   3  -0.461  0.431  0.721 -0.282
   4  -0.413  0.812 -0.338  0.188
   5  -0.888 -0.296  0.044  0.005
   6  -0.759 -0.107  0.238  0.569

  The number of factors =  4

     FINAL COMMUNALITY
```

1	0.909
2	0.896
3	0.999
4	0.979
5	0.878
6	0.968

COEFFICIENT MATRIX

0.568	-0.194	0.088	0.430
-0.467	-0.176	0.009	-0.092
0.061	-0.070	1.056	-0.169
0.147	0.945	-0.070	0.056
-0.279	-0.196	-0.006	0.220
0.332	0.083	-0.147	1.038

VARIMAX ROTATION

Initial Varimax Criterion: 0.1825
Final Varimax Criterion: 0.4871
Convergence is achieved at cycle: 5

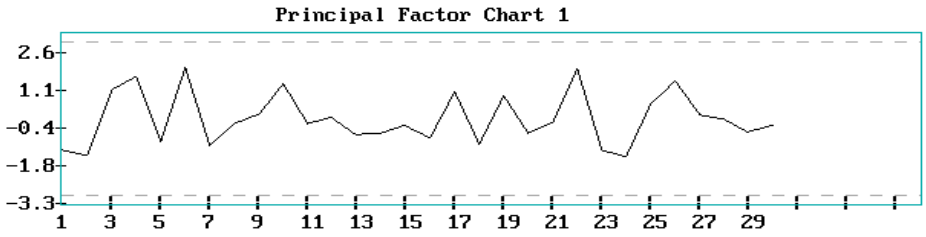
FINAL FACTOR STRUCTURE

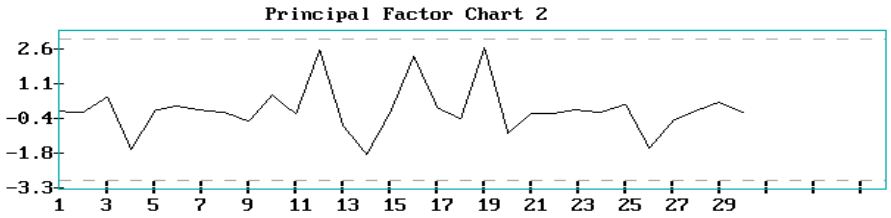
	1	2	3	4	
1	0.8816	-0.3589	-0.0294	-0.0479	factor 1 represents variables 1, 2
2	-0.8771	-0.0011	0.1336	0.3301	factor 2 represents variable 4
3	-0.1194	0.1115	0.9761	0.1378	factor 3 represents variable 3
4	-0.1223	0.9716	0.1110	0.0879	factor 4 represents variable 6
5	-0.7620	-0.0456	0.1582	0.5194	
6	-0.2943	0.1242	0.1430	0.9198	

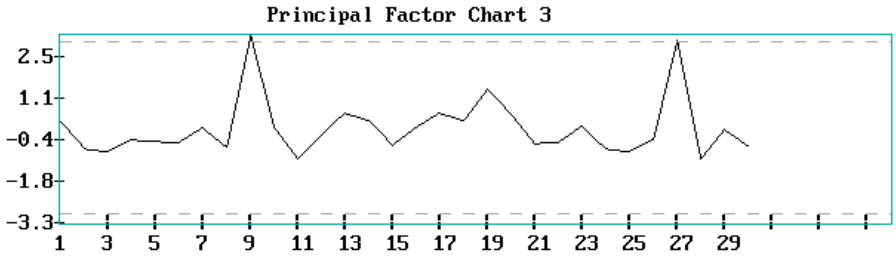
FACTOR SCORES

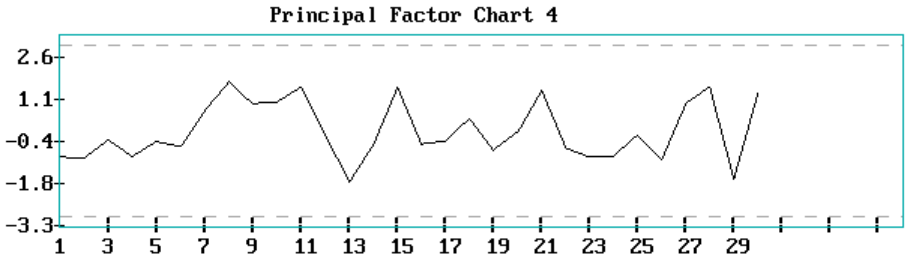
1	-1.218	-0.032	0.246	-0.874
2	-1.417	-0.083	-0.727	-0.918
3	1.102	0.535	-0.834	-0.298
4	1.612	-1.658	-0.380	-0.895
5	-0.882	-0.039	-0.425	-0.388
6	1.997	0.157	-0.522	-0.540
7	-1.036	0.012	-0.003	0.748
8	-0.226	-0.120	-0.637	1.743
9	0.141	-0.495	3.233	0.955
10	1.387	0.605	0.030	1.004
11	-0.207	-0.138	-1.074	1.505
12	0.051	2.504	-0.251	-0.135
13	-0.633	-0.676	0.544	-1.798
14	-0.577	-1.883	0.272	-0.461
15	-0.251	-0.110	-0.609	1.517
16	-0.748	2.237	0.033	-0.495
17	1.052	0.093	0.535	-0.366
18	-0.998	-0.376	0.276	0.394
19	0.880	2.616	1.387	-0.687
20	-0.571	-0.994	0.465	-0.053

21	-0.177	-0.172	-0.559	1.423
22	1.948	-0.155	-0.499	-0.559
23	-1.191	0.005	0.075	-0.876
24	-1.454	-0.109	-0.728	-0.889
25	0.593	0.207	-0.816	-0.144
26	1.475	-1.611	-0.401	-0.999
27	0.168	-0.458	3.062	0.953
28	-0.026	-0.058	-1.076	1.504
29	-0.517	0.296	-0.040	-1.661
30	-0.277	-0.099	-0.580	1.290


Principal Factor Chart 1
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 2
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 3
2.5
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29


Principal Factor Chart 4
2.6
1.1
-0.4
-1.8
-3.3
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29

Chapter 10

Part II: The Design of Experiment

Synopsis

D.O.E. for Students is a program package that features some well known statistical Designs of Experiments. As is the case of the SEMSTAT component, the D.O.E. package only accepts ASCII files as input. The Design Families handled by D.O.E. for Students are:

Full 2^n factorials.
Fractional 2^{n-p} factorials.
Full 3^n factorials.
Plackett-Burman Designs.
Central Composite Designs

The system is optionally menu driven. To start D.O.E. one has to be in the directory that contains the D.O.E. system, and then type **GO**. The main menu will appear.

Almost all response analysis programs have options to plot a normal probability plot and/or a pareto chart.

Installation
The installation procedure is described in the beginning of this document
The D.O.E. system follows the installation of the SEMSTAT components.

The installation procedure will create one main directory and four sub-directories. This is done to make the overall bookkeeping easier. The default name for the main directory is: **C:\DOE.**
You may choose any name or path you wish. In fact that is the intention of the very first prompt that is issued.

Let us assume that you have selected the default path of C:\DOE.
Then the sub-directories are named as follows:

C:\DOE\INPUTS This sub-directory contains a number of hard-coded designs.

C:\DOE\DATA This sub-directory is for storing of the response files, e.g. the data-files.

C:\DOE\DESIGNS This sub-directory is to hold the design files.

C:\DOE\REPORTS This sub-directory is for storing the output reports that are spawned from the analysis programs.

Although you could work with one big directory, it will prove handy to maintain a system with one operations headquarters (such as C:\DOE) and four sub stations.

Starting the D.O.E. system.

To start the system in the menu-driven mode, type **GO.**
This will display the main-menu as follows:

Use mouse or up / down arrow keys to position the cursor.

key	DOE MAIN MENU
1	2^N Full Factorial Designs
2	2^N-P Fractional Factorial Designs
3	3^N Full Factorial Designs
4	Central Composite Designs
5	Placket-Burman Designs
6	Other Programs
Esc	EXIT

By default selection is made by positioning the mouse on the desired line and then click the left hand mouse button. If you prefer arrow keys, press the Shift-F1 keys, which will toggle between mouse and arrow-keys.

If the arrow-keys system is selected the program types:

You may select the desired program by one of the following options:

Move the cursor to the selected line in the menu and press the Enter key.
Press the associated number key.
Press the associated F key.

For example, if you wish to work with a Central Composite design, you can move the cursor, by using the arrow keys (or mouse), till it reaches the line with the name Central Composite, and then press the Enter key.
The same could be accomplished by pressing the 4 or F4 keys.

Some of us don't like menus. In that case you could forego the menus and start a desired program by typing its name. In the above case type **CCD** and answer the ensuing prompts.

Main Programs

Name	Function
FULLFAC	2^n Full Factorial Analysis
YATES	2^n Full Factorial Designs
FRACTION	2^{n-p} Fractional Factorial Analysis
FDESIGN	2^{n-p} Fractional Factorial Designs
FULLFAC3	3^n Full Factorial Analysis
THREE	3^n Full Factorial Designs
CCD	Central Composite Designs and Analysis
PLACKBUR	Placket-Burman Designs

Design or Analysis?

You can either generate a selected design or you can analyze the responses that were gathered during the running of the design.Say, for example, that you opt to work with full factorials of the 2^n variety. If you employ the menu-driven option, type GO.
You will discover that the cursor is already positioned on the full-factorial option, so all you have to do here is to press the Enter key. (Some key boards has this key labelled as Return Key).

The following announcement will appear:

Type 1 to generate a design
or Type 2 to analyze responses.

Output of designs are meant to be a basis for reproducing printed outputs, **however, the order of the runs should be randomized.** There is a routine named RANDOM that will produce the random sequence for any design. An example is furnished in the section on Terminal Examples, later in this user's guide. The designs are stored in sub-directory C:\DOE\DESIGNS,
(or in place of DOE it will be the path name that you assigned during the installation procedure.)

Analysis of responses consists of computing effects and their associated sums of squares. If an estimate of the standard error of the effect is available, the associated t and p value will be included in the analysis. Significance at the .05 level is indicated by a star (*).

Example

A DESIGN MATRIX FOR A FULL FACTORIAL.

Enter number of variables: ? 3
(D)esign matrix or (A)nalysis matrix? d

Enter name of output file to store the design
or press Enter for f:\doe\DESIGNS\FULLFAC.3

Design matrix for a 2^3 full factorial.

Run	A	B	C
1	-1	-1	-1
2	1	-1	-1
3	-1	1	-1
4	1	1	-1
5	-1	-1	1
6	1	-1	1
7	-1	1	1
8	1	1	1

Would you want to use the real (uncoded) values? y/N: y
Enter high level for variable 1: 80
Enter low level for variable 1: 20

Enter high level for variable 2: 100
Enter low level for variable 2: 50

Enter high level for variable 3: 10
Enter low level for variable 3: 5

The matrix is stored on disk in file: c:\:doe\DESIGNS\FULLFAC.3

Let us print the file named FULLFAC.3

Design matrix for a 2^3 full factorial.

Run	A	B	C	Random order
1	-1	-1	-1	8
2	1	-1	-1	3
3	-1	1	-1	2
4	1	1	-1	5
5	-1	-1	1	4
6	1	-1	1	7
7	-1	1	1	6
8	1	1	1	1

Run the experiment as follows:

8	80.00	100.00	10.00	1
3	20.00	100.00	5.00	2
2	80.00	50.00	5.00	3
5	20.00	50.00	10.00	4
4	80.00	100.00	5.00	5
7	20.00	100.00	10.00	6
6	80.00	50.00	10.00	7
1	20.00	50.00	5.00	8

Example

AN ANALYSIS MATRIX FOR A FULL FACTORIAL.

```
 Enter number of variables: ? 3
 (D)esign matrix or (A)nalysis matrix? A

 Enter name of output file to store the design
 or press Enter for f:doe\DESIGNS\FULLFAC.3A

 Analysis matrix for a 2^3 full factorial.

run A  B  C  A  A  B  A
              B  C  C  B
                    C
 1  -1 -1 -1   1  1  1 -1
 2   1 -1 -1  -1 -1  1  1
 3  -1  1 -1  -1  1 -1  1
 4   1  1 -1   1 -1-1 -1
 5  -1 -1  1   1 -1 -1  1
 6   1 -1  1  -1  1 -1 -1
 7  -1  1  1  -1 -1  1 -1
 8   1  1 1    1  1  1  1

The matrix is stored on disk in file: c:\doe\DESIGNS\FULLFAC.3A
```

Data Input

How does one input the data?

There are a few ways to prepare input files in ASCII format.
This format is required by D.O.E.

1 Use the program DOEINP. This prompts for the number of rows and columns.
If there are no replications, the number of columns is 1.

2 Use any editor or spreadsheet package and create a raw date file. There is one catch:
the raw data file MUST be an ASCII file. So if you use Lotus 1-2-3 as the vehicle, be
sure to output in text format.

3 Use any word processor such as WordPerfect or Microsoft Word and use
the asscociated Text Output option.

4 If Host Systems are involved, use the appropriate download routines.

Output Reports on Disk.

All the output results from any analysis are automatically stored
on disk under a naming convention. For example the analysis on a full factorial 2^n is named FULLFAC.n, where n is the power of 2, that indicates the number of factors or variables in the experiment.
More specifically, if we performed an analysis on a 2^4 factorial, the name of the output report is: FULLFAC.4
Its DOS path name is: C:\DOE\REPORTS. Hence the full file id of the report is:
C:\DOE\REPORTS\ANALYSIS.4.

The date of the analysis and the name of the response data file are attached to the reports.
In almost all cases, D.O.E. suggests the default file-id for disk storage and gives you the option to give it a different name.

The general output reports in D.O.E. are arranged in the following sequence: First all the main effects, then the 2nd order interactions, then the third order interactions, etc.

For example for a 2^4 full factorial the rows in the reporting tables
are labeled: A, B, C, D, AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, ABCD.

Display the file names in a selected Directory

Type **DISPLAY**
The next screen will be displayed:

```
*******************************************************
* DISPLAY OF FILE NAMES IN A SELECTED DIRECTORY.   *
* THESE CAN BE PRINTED ON SCREEN OR PRINTER.       *
*******************************************************

Type 1 to display files in directory C:\DOE\INPUTS
Type 2 to display files in directory C:\DOE\DATA
Type 3 to display files in directory C:\DOE\REPORTS
Type 4 to display files in directory C:\DOE\DESIGNS

Your selection ?  3

 ANALYSIS.2   ANALYSIS.3   ANALYSIS.3T  ANALYSIS.4   ANALYSIS.5
 CCD.2        CCD.3        DESIGN3.2    FULLFAC.3A   PLACKMAN.15


The above  files reside on the C:\DOE\REPORTS\ disk:
Use arrow keys to  move cursor. Press the Enter key to select.
Press Esc to exit. Type $ to search another disk or directory.

 The cursor was moved to ANALYSIS.3
 (S)creen or (P)rinter?         S

 ANALYSIS ON FILE: bhh.321

Variable     Effects    t value    p value      SS
MEAN          5.667     37.720      0.000    770.667 *
A            -2.167     -7.211      0.000     28.167 *
B             2.500      8.321      0.000     37.500 *
C            -2.000     -6.656      0.000     24.000 *
AB           -0.333     -1.109      0.142      0.667
AC           -0.167     -0.555      0.293      0.167
BC            0.167      0.555      0.293      0.167
ABC           0.000      0.000      0.500      0.000

The estimated standard error =    0.300,
 based on degrees of freedom  =  16
```

Significant contrasts at the .05 level: MEAN A B C

Plotting of Effects

There are three plots available:

Normal Probability Plots
Bar graph of Effects
Pareto Plots

These plots can be saved on disk and later be displayed and/or printed to a laser or dot-matrix printer, by using the routine named: **SHOW**.

The naming convention for stored plots is the file name of the responses, and extension of GRA for EGA/VGA or .PLT for CGA quality graphs. The DOS path is the same as for reports. Let us ponder about the following situation:

We are analyzing a response vector (data-file) the file name is BHH326.DAT.
Now the normal probability plot will be named something like **C:\DOE\REPORTS\BHH326.GRA** or **C:\DOE\REPORTS\BHH326.PLT.** depending on the resolution of the monitor.

As you may be aware of, normal probability plots plot the ascendingly ranked effects against their cumulative probabilities. The scaling of the Y axis or ordinate is such that the cumulative normal probability plots as a straight line. The statistical significance of an effect is indicated by a marked deviation from the straight line. More on this willbe discussed in the section on terminal examples later in this work.

Ranking Tables.

The key to the correspondence between value and rank of the effects is stored on disk in a file called:
C:\DOE\REPORTS\BHH326.RNK.

It can be displayed by using the D.O.E editor named ROWEDIT or using any editor or word processor or the DOS command TYPE. You can also print it using the DOS command PRINT.
For example: PRINT c:\doe\reports\bhh326.rnk

Pareto Charts

If you have selected to obtain a normal probability plot, you will also have the opportunity to create and optionally store a pareto plot of the effects.

The names of pareto charts are almost the same as for the normal probability plot, except that the last letter for the name is now an E for the bargraph of the effects or an P for the Pareto plot.
For example: **C:\DOE\REPORTS\BHH32E.GRA** for an effects plot or
C:\DOE\REPORTS\BHH32P.GRA for a pareto plot.

Name Conventions and other attributes.
The following table contains some of the attribute of the main programs. As we stated in the Introduction, programs can be executed by way of menu, cursor and cursor keys, or by typing the name of the program.
The routines that perform the analysis require a ASCII file with the responses (or results, as some people prefer to call them). For example, the program FULLFAC performs the analysis on the 2^n full factorial designs, and needs the responses of each of the runs. On the other hands, the program named YATES only generates the designs and does not need a response file. The names of the programs are in column 1 of the following table and the need or lack of need is in column 2.

Next, most of the programs utilize hard coded information, which is distributed by way of "zip" files.
That is, all the information files are condensed and packed in **one** file. The INSTALL2 program will "unzip" them in their native form. They are housed in the sub-directory [d:XXX]/INPUTS, where d is the drive and XXX is the name of the main directory that you selected during the installation procedure.For example, the routine FRACTION calls upon two information files. It uses the file FRACTION.FIL to obtain generators for a selected design and the file ALABELS.FIL to get a label scheme that is used to complete the design. The names of the information files appear in column 3.

Some of the programs avail themselves of hardcoded designs. This is a time savings device, but we pay the price in terms of storage! The names of the design files are in column 5.
The routines that generate and output designs affix a descriptive name to the design output files. For example, YATES, the routine that generates designs for full 2^n factorials, offers two options: either the design matrix or the analysis matrix. The analysis file is named FULLFAC.nA,where n denotes the number of variables or factors, and A stands for analysis. Absence of the A indicates design only. The names of the output files that contain the designs are in column 5.

The programs that perform and output analyses likewise name the output files in a descriptive manner. For example the output of FULLFAC, the program that analyzes full 2^n factorials, is named ANALYSIS.n, where n is the number of factors or variables. FRACTION is the program to analyze fractional 2^{n-p} factorials. Its output file is named ANALYSIS.nF, n is as above and F indicates fractional. FULLFAC3 is the program to analyze full 3^n factorials. Its

output file is named ANALYSIS.nT, n is as above and T indicates three for 3^n. The names of the output files that contain the analysis are in column 6. The ranges for n are in the last column , column 7 of the following table.

Name of Program	Res-ponse File	Name of Info File	Name of Input Design	Name of Output Design	Name of Output Analysis	Range of n
FULLFAC	YES				ANALYSIS.n	2 ≤ n ≤ 7
YATES	NO			FULLFAC.n FULLFAC.NA		2 ≤ n ≤ 7
FRACTION	YES	FRACTION.FIL ALABELS.FIL			ANALYSIS.n	2 ≤ n ≤ 2
FDESIGN	NO	FRACTION.FIL ALABELS.FIL DESIGN.GEN		FRACTION.n FRACTION.n		2 ≤ n ≤ 7
FULLFAC3	YES	LABELS.n	THREE.		ANALYSIS.n	2 ≤ n ≤ 4
THREE	NO	THREE.DES		DESIGN3.n DESIGN3.nA		2 ≤ n ≤ 4
CCD	YES	CCD.FIL	CCDn	CCDn.DES	CCD.n	2 ≤ n ≤ 7
PLACKBUR	YES	PLACKBUR.FIL	PB.n	PB.n	PLACKMAN.n	1 ≤ n ≤ 16 1 ≤ c ≤ 20

Factorial Experimental Designs

An experiment can be thought of as "trying out something" and compare the results of that what we tried.
For example there may be two slightly different routes from home to get to "the office". We may want to find out which is the shorter one in time. So here is a suggested experiment: We take route A for a week, and record the travel time. The next week we take route B. Then we compare the recorded times.
End of experiment.
There are an infinite amount of such examples of experiments. But it would make for very dull reading, and this is not really the intention of this work. Let us delve a little bit deeper in the route study.
The week for route A may have been ideal from the weather perspective. The B week may have been awful.
Snow, Ice, Sleet, etc ...Or, the A week may have witnessed some sort of political demonstrations with road blocks, police actions and the like. The B week may have been tame with smooth flowing traffic.
Another situation could have muddied the waters: You drove your own car during the A week, but had to default to the spouse's vehicle during the B week, because Spouse liked your car better and took it for a spin to a far away place ...The point is that this experiment may not have been conclusive since one or more factors were introduced that influenced the final outcome. What can we do to improve the experiment?
The very **thought** of organizing is the basis of the concept of **designing** an experiment. In other words a designed experiment is a planned aproach to conduct an experiment. In the case of the route study, we could have driven on alternate days instead of alternate weeks. Or discounted days on which demonstrations had to be conducted. Or/and make sure that the same vehicle would serve us during the experiment. And so on ...

Example of a series of Random Numbers

```
*****************************************************************************
*                RANDOM NUMBER GENERATOR                                    *
*****************************************************************************

How Many Random Numbers?      16

1  8  11  10  5   12   15  14   9   16   3   2   13   4   7   6

More?  Y/N    n
```

Another syntax is: **RANDOM** 16
This yields the same result, without the prompt.

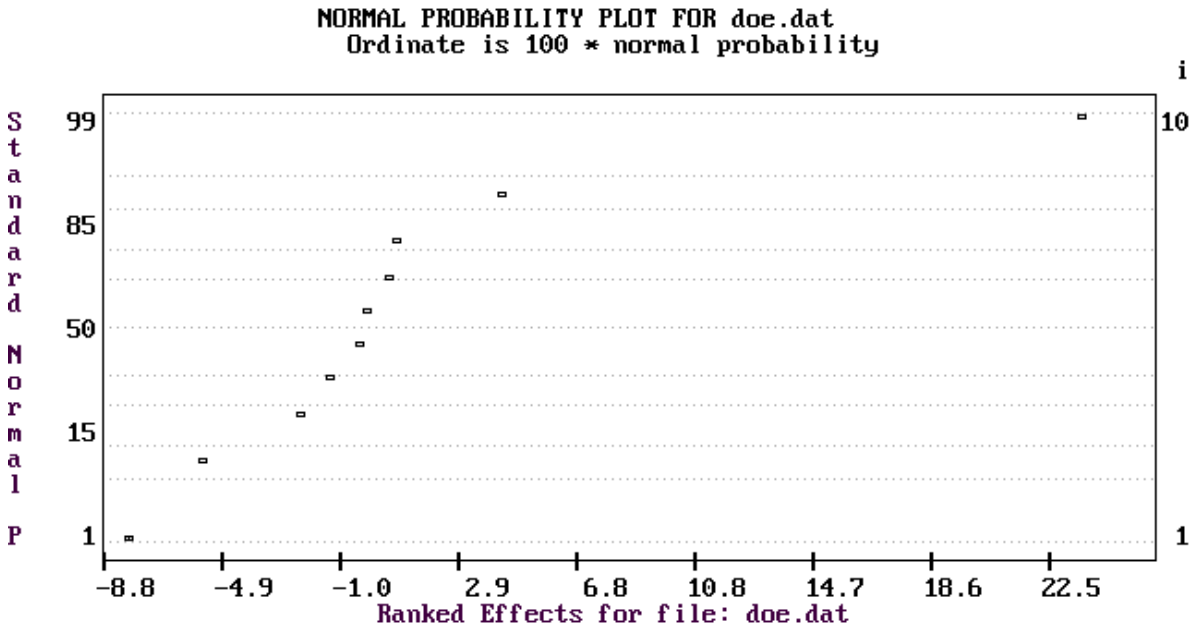
Example of a 2^n Factorial

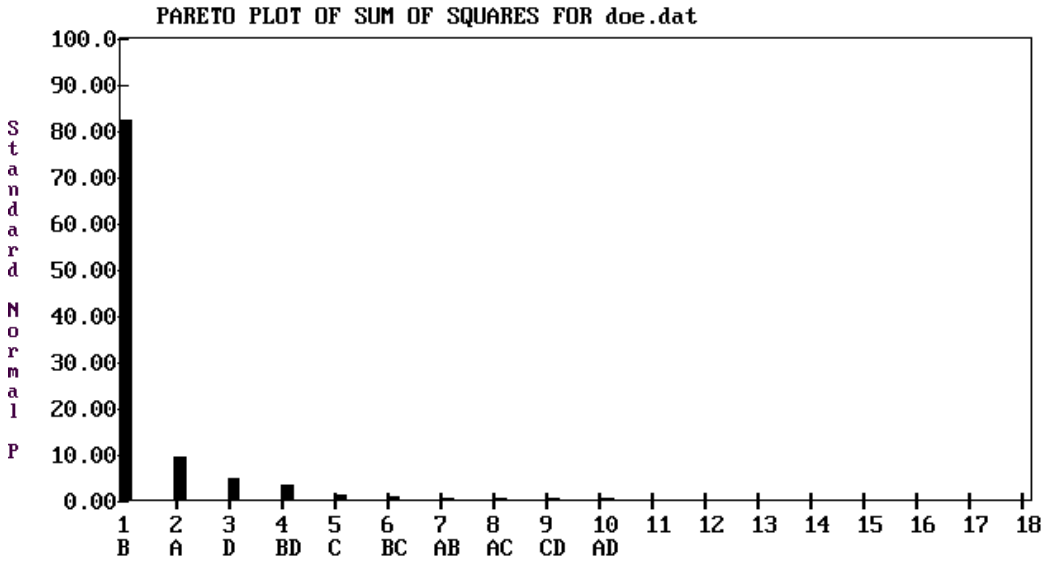
The data consists of four variables and 16 runs
We can analyze this experiment as a 2^4 full factorial.
The response vector is:
71 61 90 82 68 61 87 80 61 50 89 83 59 51 85 78
Note, that is *all* the information that the program FULLFAC needs
to perform the analysis.
We could invoke the program by simple typing: **FULLFAC** or use the
menu approach by typing **GO** and work through the selection process.
You notice that when the program prompts for the ascii (text) esponse input file, we typed the rather lengthy file id, C:\DOE\DATA\BHH326.DAT.
That is not really necessary. We could have simply pressed the Enter key, which would show all the files in the directory DOE\DATA. Then we would move the cursor to the appropriate name, in this case BHH326.DAT, and press the Enter key again. That would have accomplished the goal. Furthermore, if you have your data files in another directory or on a floppy, you could select that by first pressing the $ key, (the Shift and 4 keys) and then answer the ensuing prompt for another directory by the choosen one.

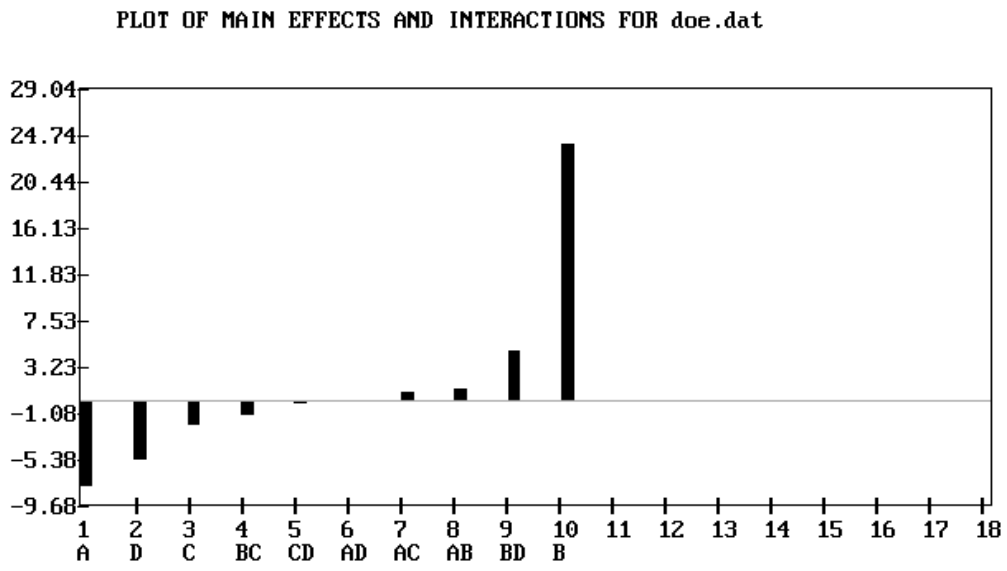
```
FULLFAC
THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A 2^k
EXPERIMENT.
Enter file id of responses or press Enter to see a list:
c:\doe\data\bhh326.dat
 Want to name the variables? (1-8 letters) y/N:
Type 1 to enter your own standard error or
type 2 to use higher order interactions or
press the Enter key for no standard error: 2
 Enter beginning order of the interactions to be used: 3
    Enter name of output file to store the analysis
or press Enter for C:\DOE\REPORTS\ANALYSIS.4
 ANALYSIS ON FILE: c:\doe\data\bhh326.dat
 Variable           Effects    t value    p value      SS
MEAN              72.250    263.820      0.000  83521.000 *
A                 -8.000    -14.606      0.000    256.000 *
B                 24.000     43.818      0.000   2304.000 *
C                 -2.250     -4.108      0.005     20.250 *
D                 -5.500    -10.042      0.000    121.000 *
AB                 1.000      1.826      0.064      4.000
AC                 0.750      1.369      0.115      2.250
AD                 0.000      0.000      0.500      0.000
BC                -1.250     -2.282      0.036      6.250 *
BD                 4.500      8.216      0.000     81.000 *
CD                -0.250     -0.456      0.667      0.250

The estimated standard error =    0.548, based on degrees of
freedom  =  5
Significant contrasts at the .05 level:
MEAN  A  B  C  D  BC  BD
```

For file: c:\doe\data\bhh326.dat

MAX	MIN	MEAN	STD.DEV.	NO.DATA
24.0000	-8.0000	0.7667	7.0277	15


NORMAL PROBABILITY PLOT FOR doe.dat
Ordinate is 100 * normal probability
i
Standard Normal P
99
85
50
15
1
10
1
-8.8
-4.9
-1.0
2.9
6.8
10.8
14.7
18.6
22.5
Ranked Effects for file: doe.dat


PARETO PLOT OF SUM OF SQUARES FOR doe.dat
Standard Normal P
100.0
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1 B
2 A
3 D
4 BD
5 C
6 BC
7 AB
8 AC
9 CD
10 AD
11
12
13
14
15
16
17
18



The correspondence between the ranks, effects and labels for every response file is written to a little help file named RRRRR.RNK where RRRRR is the name of the response data file.
It is stored in the sub-directory REPORTS.
For our file the full path and file id is: C:\DOE\REPORTS\BHH326.RNK.
This of courses assumes that the drive is C and the main directory for the program package is DOE. You may have picked different drives and name during the installation program.

Ranked Effects for file C:\DOE\DATA\BHH326.DAT

RANK	EFFECT	LABEL
1	-8.000	A
2	-5.500	D
3	-2.250	C
4	-1.250	BC
5	-0.750	ABC
6	-0.750	BCD
7	-0.250	CD
8	-0.250	ACD
9	-0.250	ABCD
10	0.000	AD
11	0.500	ABD
12	0.750	AC
13	1.000	AB
14	4.500	BD

```
   15   24.000     B
```

Example of replicated data

THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A 2^N EXPERIMENT.

Enter file id of responses or press Enter to see a list: C\DOE\DATA\BHH321.DAT

```
        6    7      6
        4    5      5
       10    9      8
        7    7      6
        4    5      4
        3    3      1
        8    7      7
        5    5      4
```

This file with 3 replicates can be formed using DOEINPUT or any word processor of your choice. You can also use spreadsheets as long as you save in ascii (text) format.

name the variables? *(*1-8 letters*)* y/N*: n*

Enter name of output file to store the analysis
or press Enter for c:\doe\REPORTS\ANALYSIS.3

```
Effects    t value    p value      SS
 MEAN       5.667     37.720      0.000    770.667 *
  A        -2.167     -7.211      0.000     28.167 *
  B         2.500      8.321      0.000     37.500 *
  C        -2.000     -6.656      0.000     24.000 *
  AB       -0.333     -1.109      0.142      0.667
  AC       -0.167     -0.555      0.293      0.167
  BC        0.167      0.555      0.293      0.167
  ABC       0.000      0.000      0.500      0.000

The estimated standard error =    0.300
based on degrees of freedom  =    16
```

Significant contrasts at the .05 level: Mean A B C

Example of missing data

```
THIS PROGRAM ANALYZES THE RESPONSES OF A 2^N EXPERIMENT.

Enter file id of responses or press Enter to see a list:
c:\doe\data\bhh321l.dat


Want to name the variables? (1-8 letters) y/N:


   6       7       6
   4       5       5
   10              8
   7       7       6
   4       5       4
   3       3
   8       7       7
           5       4



Enter name of output file to store the analysis
or press Enter for c:\doe\REPORTS\ANALYSIS.3

ANALYSIS ON FILE: c:\doe\data\bhh321l.dat     07-22-1992

Variable      Effects    t value    p value      SS
MEAN            5.762     39.417      0.000    697.190 *
A              -2.190     -7.493      0.000     25.190 *
B               1.619      5.538      0.000     13.762 *
C              -2.000     -6.841      0.000     21.000 *
AB              0.095      0.326      0.375      0.048
AC             -1.619     -5.538      0.000     13.762 *
BC              0.667      2.280      0.020      2.333 *
ABC            -1.238     -4.235      0.000      8.048 *

The estimated standard error = 0.292
based on degrees of freedom  =  13

Significant contrasts at the .05 level:
 MEAN  A  B  C  AC  BC  ABC
```

Example of a 2^{n-p} Fractional Factorial

This works with the same data as above, but here we would like to analyze **five** variables in stead of four, still with 16 runs. Since we would need 32 runs to analyze a 2^5 experiment but we have only 16 runs at our disposition we resort to **fractional factorials.** The usual notation is:

2^{n-p} where n is the number of variables or factors that we wish to study and p is the order of fractionality. For example if p = 1, we have fractionalized 2^{-1}, which is ½. If p = 2 then we fractionalized 2^{-2} which is $½^2$ or ¼.

In the case of analyzing 5 variables with 16 runs, the fraction is 16/32 or ½ which means that n = 5 and p = 1.

FRACTION

THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A FRACTIONAL 2^K-P FACTORIAL EXPERIMENT.

Enter id of responses or press Enter for a list:

c:\doe\data\bhh326.dat

```
Enter by number the desired design or press Enter for a list:
#     Factors   Fraction   Resolution   Runs
1       3       2^(3-1)       III        4
2       4       2^(4-1)       IV         8
3       5       2^(5-1)       V         16
4       5       2^(5-2)       III        8
5       6       2^(6-1)       VI        32
6       6       2^(6-2)       IV        16
7       6       2^(6-3)       III        8
8       7       2^(7-1)       VIII      64
9       7       2^(7-2)       IV        32
10      7       2^(7-3)       IV        16
11      7       2^(7-4)       III        8
12      8       2^(8-2)       V         64
13      8       2^(8-3)       IV        32
14      8       2^(8-4)       IV        16
15      9       2^(9-3)       IV        64
16      9       2^(9-4)       IV        32
17      9       2^(9-5)       III       16
18      10      2^(10-4)      V         64
19      10      2^(10-5)      IV        32
20      10      2^(10-6)      III       16
```

Enter number of desired design or press Enter for a list: ? 3
Want to name the variables? (1-8 letters) y/N:

Type 1 to enter your own standard error or
type 2 to use higher order interactions or
press the Enter key for no standard error: 2

Enter beginning order of the interactions to be used: 3

Enter name of output file to store the analysis
or press Enter for C:\DOE\REPORTS\ANALYSIS.4

```
        ANALYSIS ON FILE: c:\doe\data\bhh326.dat
Variable      Effects    t value    p value      SS
MEAN           72.250     70.196      0.000  83521.000 *
A              -8.000     -3.886      0.006    256.000 *
B              24.000     11.659      0.000   2304.000 *
C              -2.250     -1.093      0.162     20.250
D              -5.500     -2.672      0.022    121.000 *
E              -0.250     -0.121      0.454      0.250
AB              1.000      0.486      0.324      4.000
AC              0.750      0.364      0.365      2.250
AD              0.000      0.000      0.500      0.000
AE             -0.750     -0.364      0.365      2.250
BC             -1.250     -0.607      0.285      6.250
BD              4.500      2.186      0.040     81.000 *
BE             -0.250     -0.121      0.454      0.250
CD             -0.250     -0.121      0.454      0.250
CE              0.500      0.243      0.409      1.000
DE             -0.750     -0.364      0.365      2.250

The estimated standard error =    2.059, based on degrees of freedom  =  5
 Significant contrasts at the .05 level:
 MEAN  A    B    D    BD

 The analysis is stored on disk in file: C:\DOE\REPORTS\ANALYSIS.4
 Want a normal probability plot of the effects? y/n:
 More?    y/n:
```

Example of Placket-Burman Designs

In this example we will only output the design. The main purpose of this example is to show the table of available designs. The analysis, had we selected that option, proceeds identically to fractional factorials.

```
 PLACKMAN
*****************************************
               *        Placket-Burman Designs           *
               *****************************************

 THIS PROGRAM ANALYZES THE RESPONSES OF A PLACKET-BURMAN EXPERIMENT.
  The following designs are available:

   #    NAME    NO.OF  NO. OF
                RUNS  VARIABLES
   1    PB.8       8       7
   2    PB.12     12      11
   3    PB.16     16      15
   4    PB.20     20      19
   5    PB.24     24      23
   6    PB.32     32      31
   7    PB.36     36      35
   8    PB.44     44      43
   9    PB.48     48      47
  10    PB.60     60      59
  11    PB.64     64      63
  12    PB.68     68      67
  13    PB.72     72      71
  14    PB.80     80      79
  15    PB.84     84      83
  16    PB.128   128     127

  Enter number of desired design: ? 2

 Type 'D' if all you wish is to generate the design, otherwise press Enter: d
        1 file(s) copied

  Design PB.12  is stored in C:\DOE\DESIGNS
```

This is the design:

```
 NT   A  B  C  D  E  F  G  H  I  J  K
```

1	1	-1	1	-1	-1	-1	1	1	1	-1	1
2	1	1	-1	1	-1	-1	-1	1	1	1	-1
3	-1	1	1	-1	1	-1	-1	-1	1	1	1
4	1	-1	1	1	-1	1	-1	-1	-1	1	1
5	1	1	-1	1	1	-1	1	-1	-1	-1	1
6	1	1	1	-1	1	1	-1	1	-1	-1	-1
7	-1	1	1	1	-1	1	1	-1	1	-1	-1
8	-1	-1	1	1	1	-1	1	1	-1	1	-1
9	-1	-1	-1	1	1	1	-1	1	1	-1	1
10	1	-1	-1	-1	1	1	1	-1	1	1	-1
11	-1	1	-1	-1	-1	1	1	1	-1	1	1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Example of a 3^n design

The following data set are responses of a 3^3 design.

```
159   395   149    25   255   251   184   363   378
260   454   112    98   422   270   237   362   363
146   417   150   103   455   172   195   492   278
```

```
FULLFAC3
             ******************************************
             *              3^n Designs                *
             ******************************************

THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A 3^N EXPERIMENT.
Enter file id of responses or press Enter for a list:

c:\doe\data\d333.dat

Want to name the variables? (1-8 letters) y/N:

Type 1 to enter your own error variance or
type 2 to use higher order interactions or
press the Enter key for no standard error: 2


Enter beginning order of the interactions to be used: 3

Enter name of output file to store the analysis
or press Enter for C:\DOE\REPORTS\ANALYSIS.3T
```

```
ANALYSIS ON FILE: c:\doe\data\d333.dat

Variable      Effects    t value    p value      ss

MEAN          264.630     18.777      0.000 1890778.750 *
A              39.778      2.305      0.016   28480.889 *
B              33.889      1.963      0.032   20672.223 *
C              13.833      0.801      0.216    3444.500
a             -68.519     -6.876      0.000  253518.516 *
b              18.370      1.843      0.040   18223.408 *
c             -10.907     -1.095      0.143    6424.463
AB             46.417      2.196      0.020   25854.084 *
AC            -21.167     -1.001      0.164    5376.333
Ab            -19.028     -1.559      0.067   13034.027
Ac              7.389      0.605      0.276    1965.444
BC              2.500      0.118      0.454      75.000
Ba             21.028      1.723      0.050   15918.027
Bc              5.611      0.460      0.325    1133.444
Ca            -22.333     -1.830      0.041   17956.000 *
Cb             -9.667     -0.792      0.219    3364.000
ab              3.102      0.440      0.332    1039.120
```

Example of a CCD analysis

Consider the following set of responses:

76.5 78.0 77.0 79.5 75.6 78.4 77.0 78.5 79.9 80.3 80.0 79.7 79.8

Central Composite Designs are often used in Response Surface Methodology (RSM). They will tell us if we have reached an optimum, what is the value of that optimum, whether it is a maximum, minimum or neither (called a saddle point) and if we should have used a CCD in the first place.

We used the NAMING option in this example.

In this example the regression analysis determined that the quadratic model was acceptable, and subsequently continued with more analyses.

```
CCD
              ******************************************
              *  Analysis of Central Composite Designs  *
              ******************************************

   The following Uniform-Precision Central Composite Designs are available:

 #  Number of  Number of  Number of Number of Total        α
    Variables  factorial  star      center    number of
               trials     trials    trials    trials
-------------------------------------------------------------------
 2     2          4          4         5         13       1.4142
 3     3          8          6         6         20       1.6820
 4     4         16          8         7         31       2.0000
 5     5         16         10         6         32       2.0000
 6     6         32         12         9         53       2.3780
 7     7         64         14        14         92       2.8280
 8     8        128         16        20        164       3.3640

 Enter number of desired Central Composite Design: 2
 Enter D to output the design or press Enter to continue:

 THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A CCD EXPERIMENT.
 Enter file id of responses or press Enter to see a list:
 c:\doe\data\doug534.dat


 Want to name the variables? (1-8 letters) y/N: y

 Type K to enter from key board (default)
 or   F to enter from file
 or   H for help

 Enter name of variable: 1  time
```

```
 Enter name of variable: 2  temp
 Save the names in file XXXXXXXX.NAM ?  y/N:


 Enter name of output file to store the analysis
 or press Enter for C:\DOE\REPORTS\CCD.2
```

```
 ANALYSIS ON FILE: c:\doe\data\doug534.dat

 Variable      Effects    std.err    t value      p value
 MEAN           79.940      0.103    775.586      0.000
 A               0.995      0.081     12.211      0.000 *    time
 B               0.515      0.081      6.322      0.000 *    temp
 AB              0.250      0.115      2.169      0.025 *    time*temp
 AA             -1.376      0.087    -15.749      0.000 *    time*time
 BB             -1.001      0.087    -11.457      0.000 *    temp*temp
 The estimated std.dev. using center points only =    0.230, with df =  4
 The estimated std.dev. using residuals =    0.266, with df =  7

The center points were used in computing the std.errors of the estimates.

 Significant contrasts at the .05 level:
  A  B  AB  AA  BB


 ****************************************************************
 *                      Lack-of-Fit Test                        *
 ****************************************************************

 Source             SS        DF             MS             F
 Regression       28.235       5            5.647         79.637
 Residuals         0.496       7            0.071
 Lack of Fit       0.284       3            0.095          1.782
 Pure Error        0.212       4            0.053

 Probability of F Lack-of-Fit =     0.7103

 Probability of F Regression  =     1.0000
```

```
 Do not reject the hypothesis of model adequacy
```

```
*****************************************************************
*                       Canonical Analysis                      *
*****************************************************************
The stationary point is:
  0.3892    0.3058

The distance from the design center is:     0.3892

The predicted response at the stationary point is:    80.2124

EIGENVALUES OF B
    -1.414      -0.964
The response surface is a maximum

Canonical form of the quadratic model
Y =  80.2124 +  -1.4143 W1^2  +  -0.9635 W2^2

The analysis is stored on disk in file: C:\DOE\REPORTS\CCD.2

Want a normal probability plot of the effects? y/n:

More?    y/n:
```

References

1 Douglas C. Montgomery
"Design and Analysis of Experiments"
John Wiley & Sons, 1991

2 William J Diamond
"Practical Experiment Designs"
Van Nostrand-Rheinhold, 1989

3 Stephen R. Schmidt and Robert G. Launsby
"Understanding Industrial Experiments"
Air Academy Press, 1989

4 G.E.P Box and Norman R. Draper
"Emperical Model-Building and Response Surfaces"
John Wiley & Sons, 1987

5 G.E.P Box, William G. Hunter, and J. Stuart Hunter
"Statistics for Experimenters"
John Wiley & Sons, 1978

6 Owen Davies, Editor
"The Design and Analysis of Industrial Experiments
Hafner Publishing, 1963

7 Forrest W. Breyfogle III
"Statistical Methods"
John Wiley and Sons, 1992

Chapter 11

Example of a series of Random Numbers

RANDOM

```
**************************************************************************
*              RANDOM NUMBER GENERATOR                                   *
**************************************************************************
```

How Many Random Numbers? *16*

1 8 11 10 5 12 15 14 9 16 3 2 13 4 7 6

More? Y/N *n*

Another syntax is: **RANDOM** 16
This yields the same result, without the prompt.

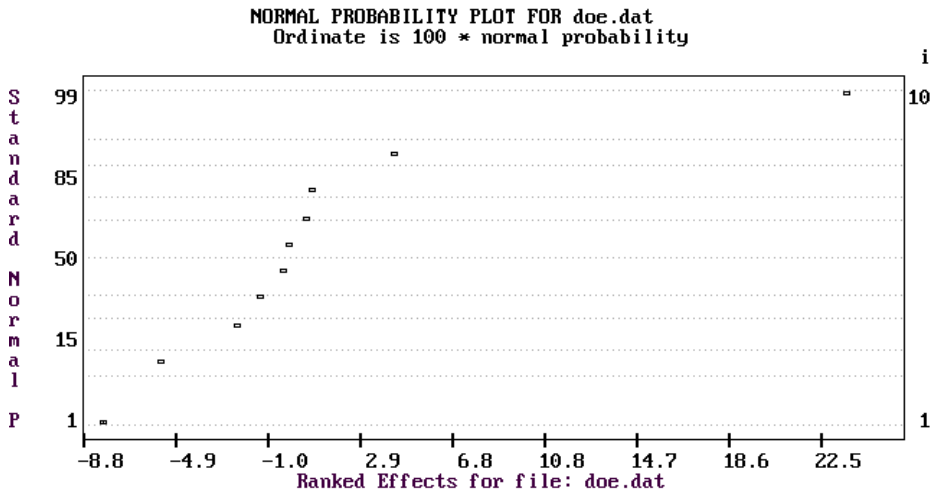
Example of a 2^n Factorial

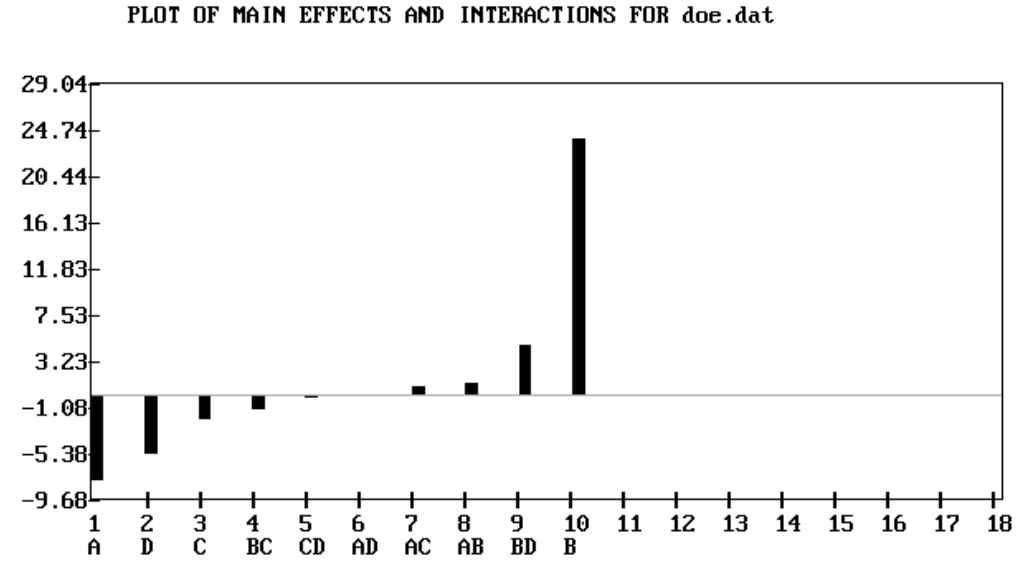
The data consists of four variables and 16 runs
We can analyze this experiment as a 2^4 full factorial.
The response vector is:
71 61 90 82 68 61 87 80 61 50 89 83 59 51 85 78
Note, that is *all* the information that the program FULLFAC needs
to perform the analysis.
We could invoke the program by simple typing: **FULLFAC** or use the
menu approach by typing **GO** and work through the selection process.
You notice that when the program prompts for the ascii (text) esponse input file, we typed the rather lengthy file id, C:\DOE\DATA\BHH326.DAT.
That is not really necessary. We could have simply pressed the Enter key, which would show all the files in the directory DOE\DATA. Then we would move the cursor to the appropriate name, in this case BHH326.DAT, and press the Enter key again. That would have accomplished the goal. Furthermore, if you have your data files in another directory or on a floppy, you could select that by first pressing the $ key, (the Shift and 4 keys) and then answer the ensuing prompt for another directory by the choosen one.

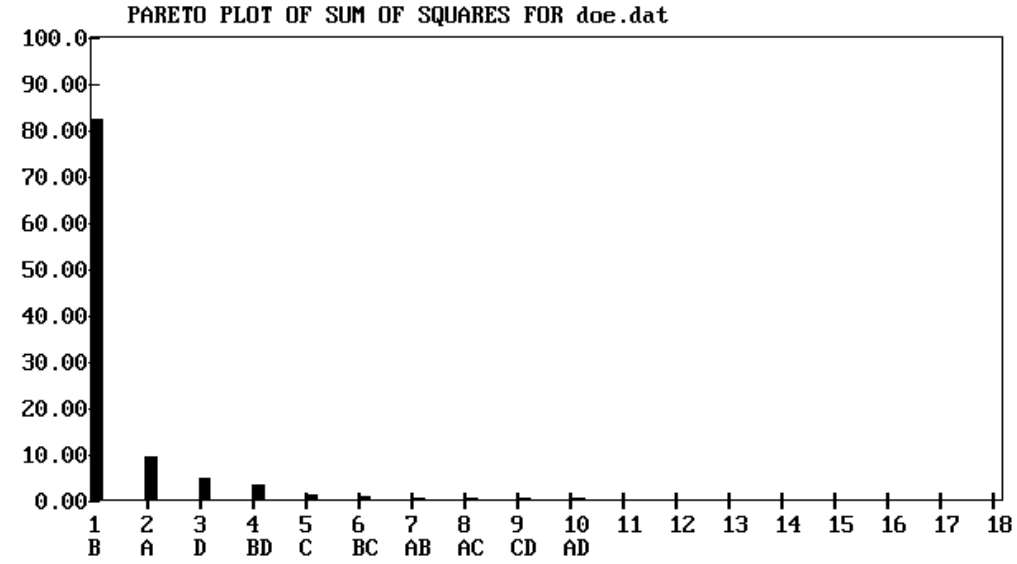
FULLFAC
THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A 2^k EXPERIMENT.
Enter file id of responses or press Enter to see a list:
c:\doe\data\bhh326.dat
 Want to name the variables? (1-8 letters) y/N:
Type 1 to enter your own standard error or
type 2 to use higher order interactions or
press the Enter key for no standard error: 2
 Enter beginning order of the interactions to be used: 3
    Enter name of output file to store the analysis
or press Enter for C:\DOE\REPORTS\ANALYSIS.4
 ANALYSIS ON FILE: c:\doe\data\bhh326.dat

Variable	Effects	t value	p value	SS	
MEAN	72.250	263.820	0.000	83521.000	*
A	-8.000	-14.606	0.000	256.000	*
B	24.000	43.818	0.000	2304.000	*
C	-2.250	-4.108	0.005	20.250	*
D	-5.500	-10.042	0.000	121.000	*
AB	1.000	1.826	0.064	4.000	
AC	0.750	1.369	0.115	2.250	
AD	0.000	0.000	0.500	0.000	
BC	-1.250	-2.282	0.036	6.250	*
BD	4.500	8.216	0.000	81.000	*
CD	-0.250	-0.456	0.667	0.250	

The estimated standard error = 0.548, based on degrees of freedom = 5
Significant contrasts at the .05 level:
MEAN A B C D BC BD


NORMAL PROBABILITY PLOT FOR doe.dat
Ordinate is 100 * normal probability
i
Standard Normal P
99
85
50
15
1
10
1
-8.8
-4.9
-1.0
2.9
6.8
10.8
14.7
18.6
22.5
Ranked Effects for file: doe.dat


PLOT OF MAIN EFFECTS AND INTERACTIONS FOR doe.dat
29.04
24.74
20.44
16.13
11.83
7.53
3.23
-1.08
-5.38
-9.68
1 A
2 D
3 C
4 BC
5 CD
6 AD
7 AC
8 AB
9 BD
10 B
11
12
13
14
15
16
17
18


PARETO PLOT OF SUM OF SQUARES FOR doe.dat
100.0
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1 B
2 A
3 D
4 BD
5 C
6 BC
7 AB
8 AC
9 CD
10 AD
11
12
13
14
15
16
17
18

For C:\DOE\REPORTS\BHH326.DAT

MAX	MIN	MEAN	STD.DEV.	NO.DATA
24.0000	-8.0000	0.7667	7.0277	15

The correspondence between the ranks, effects and labels for every response file is written to a little help file named RRRRR.RNK where RRRRR is the name of the response data file.
It is stored in the sub-directory REPORTS.
For our file the full path and file id is: C:\DOE\REPORTS\BHH326.RNK.
This of courses assumes that the drive is C and the main directory for the program package is DOE. You may have picked different drives and name during the installation program.

Ranked Effects for file C:\DOE\DATA\BHH326.DAT

RANK	EFFECT	LABEL
1	-8.000	A
2	-5.500	D
3	-2.250	C
4	-1.250	BC
5	-0.750	ABC
6	-0.750	BCD
7	-0.250	CD
8	-0.250	ACD
9	-0.250	ABCD
10	0.000	AD
11	0.500	ABD
12	0.750	AC
13	1.000	AB
14	4.500	BD
15	24.000	B

Example of replicated data

THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A 2^N EXPERIMENT.

Enter file id of responses or press Enter to see a list: C\DOE\DATA\BHH321.DAT

```
 6    7    6          This file with 3 replicates can be
 4    5    5          formed using DOEINPUT or any  word
10    9    8          processor of your choice. You can also
 7    7    6          use spreadsheets as long as you save
 4    5    4          in ascii (text) format.
 3    3    1
 8    7    7
 5    5    4
```

name the variables? *(*1-8 letters*)* y/N*: n*

Enter name of output file to store the analysis

or press Enter for c:\doe\REPORTS\ANALYSIS.3

Effects	t value	p value	SS		
MEAN	5.667	37.720	0.000	770.667	*
A	-2.167	-7.211	0.000	28.167	*
B	2.500	8.321	0.000	37.500	*
C	-2.000	-6.656	0.000	24.000	*
AB	-0.333	-1.109	0.142	0.667	
AC	-0.167	-0.555	0.293	0.167	
BC	0.167	0.555	0.293	0.167	
ABC	0.000	0.000	0.500	0.000	

```
The estimated standard error =    0.300
based on degrees of freedom  =    16
```

Significant contrasts at the .05 level: Mean A B C

Example of missing data

```
THIS PROGRAM ANALYZES THE RESPONSES OF A 2^N EXPERIMENT.

Enter file id of responses or press Enter to see a list:
c:\doe\data\bhh321l.dat


Want to name the variables? (1-8 letters) y/N:


   6       7       6
   4       5       5
   10              8
   7       7       6
   4       5       4
   3       3
   8       7       7
           5       4



Enter name of output file to store the analysis
or press Enter for c:\doe\REPORTS\ANALYSIS.3

ANALYSIS ON FILE: c:\doe\data\bhh321l.dat

Variable      Effects    t value    p value      SS
MEAN            5.762     39.417      0.000    697.190 *
A              -2.190     -7.493      0.000     25.190 *
B               1.619      5.538      0.000     13.762 *
C              -2.000     -6.841      0.000     21.000 *
AB              0.095      0.326      0.375      0.048
AC             -1.619     -5.538      0.000     13.762 *
BC              0.667      2.280      0.020      2.333 *
ABC            -1.238     -4.235      0.000      8.048 *

The estimated standard error = 0.292
based on degrees of freedom  =  13
```

```
Significant contrasts at the .05 level:
 MEAN  A  B  C  AC  BC  ABC
```

Example of a 2^{n-p} Fractional Factorial

This works with the same data as above, but here we would like to analyze **five** variables in stead of four, still with 16 runs. Since we would need 32 runs to analyze a 2^5 experiment but we have only 16 runs at our disposition we resort to **fractional factorials.** The usual notation is:

2^{n-p} where n is the number of variables or factors that we wish to study and p is the order of fractionality. For example if p = 1, we have fractionalized 2^{-1}, which is ½. If p = 2 then we fractionalized 2^{-2} which is $½^2$ or ¼.

In the case of analyzing 5 variables with 16 runs, the fraction is 16/32 or ½ which means that n = 5 and p = 1.

FRACTION

THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A FRACTIONAL

2^K-P FACTORIAL EXPERIMENT.

Enter id of responses or press Enter for a list:

c:\doe\data\bhh326.dat

Enter by number the desired design or press Enter for a list:

#	Factors	Fraction	Resolution	Runs
1	3	2^(3-1)	III	4
2	4	2^(4-1)	IV	8
3	5	2^(5-1)	V	16
4	5	2^(5-2)	III	8
5	6	2^(6-1)	VI	32
6	6	2^(6-2)	IV	16
7	6	2^(6-3)	III	8
8	7	2^(7-1)	VIII	64
9	7	2^(7-2)	IV	32
10	7	2^(7-3)	IV	16
11	7	2^(7-4)	III	8
12	8	2^(8-2)	V	64
13	8	2^(8-3)	IV	32
14	8	2^(8-4)	IV	16
15	9	2^(9-3)	IV	64
16	9	2^(9-4)	IV	32
17	9	2^(9-5)	III	16

```
 18       10       2^(10-4)       V          64
 19       10       2^(10-5)       IV         32
 20       10       2^(10-6)       III        16
```

Enter number of desired design or press Enter for a list: ? 3
Want to name the variables? (1-8 letters) y/N:

Type 1 to enter your own standard error or
type 2 to use higher order interactions or
press the Enter key for no standard error: 2

Enter beginning order of the interactions to be used: 3
Enter name of output file to store the analysis
or press Enter for C:\DOE\REPORTS\ANALYSIS.4

```
ANALYSIS ON FILE: c:\doe\data\bhh326.dat
Variable      Effects    t value    p value       SS
MEAN           72.250     70.196      0.000  83521.000 *
A              -8.000     -3.886      0.006    256.000 *
B              24.000     11.659      0.000   2304.000 *
C              -2.250     -1.093      0.162     20.250
D              -5.500     -2.672      0.022    121.000 *
E              -0.250     -0.121      0.454      0.250
AB              1.000      0.486      0.324      4.000
AC              0.750      0.364      0.365      2.250
AD              0.000      0.000      0.500      0.000
AE             -0.750     -0.364      0.365      2.250
BC             -1.250     -0.607      0.285      6.250
BD              4.500      2.186      0.040     81.000 *
BE             -0.250     -0.121      0.454      0.250
CD             -0.250     -0.121      0.454      0.250
CE              0.500      0.243      0.409      1.000
DE             -0.750     -0.364      0.365      2.250

The estimated standard error = 2.059, based on 5 df
Significant contrasts at the .05 level:
MEAN  A    B    D    BD
The analysis is stored on disk in file:
C:\DOE\REPORTS\ANALYSIS4
Want a normal probability plot of the effects? y/n:   n
More?    y/n:     n
```

Example of Placket-Burman Designs

In this example we will only output the design. The main purpose of this example is to show the table of available designs. The analysis, had we selected that option, proceeds identically to fractional factorials.

```
PLACKMAN
              ******************************************
*        Placket-Burman Designs           *
******************************************
THIS PROGRAM ANALYZES THE RESPONSES OF A PLACKET-BURMAN
EXPERIMENT.
 The following designs are available:
  #    NAME    NO.OF  NO. OF
               RUNS  VARIABLES
  1    PB.8       8      7
  2    PB.12     12      11
  3    PB.16     16      15
  4    PB.20     20      19
  5    PB.24     24      23
  6    PB.32     32      31
  7    PB.36     36      35
  8    PB.44     44      43
  9    PB.48     48      47
 10    PB.60     60      59
 11    PB.64     64      63
 12    PB.68     68      67
 13    PB.72     72      71
 14    PB.80     80      79
 15    PB.84     84      83
 16    PB.128   128     127

 Enter number of desired design: ? 2
```

Type ‘D’ if all you wish is to generate the design, otherwise press Enter: d

```
1 file(s) copied
 Design PB.12  is stored in C:\DOE\DESIGNS
```

This is the design:

NT	A	B	C	D	E	F	G	H	I	J	K
1	1	-1	1	-1	-1	-1	1	1	1	-1	1
2	1	1	-1	1	-1	-1	-1	1	1	1	-1
3	-1	1	1	-1	1	-1	-1	-1	1	1	1
4	1	-1	1	1	-1	1	-1	-1	-1	1	1
5	1	1	-1	1	1	-1	1	-1	-1	-1	1
6	1	1	1	-1	1	1	-1	1	-1	-1	-1
7	-1	1	1	1	-1	1	1	-1	1	-1	-1
8	-1	-1	1	1	1	-1	1	1	-1	1	-1
9	-1	-1	-1	1	1	1	-1	1	1	-1	1
10	1	-1	-1	-1	1	1	1	-1	1	1	-1
11	-1	1	-1	-1	-1	1	1	1	-1	1	1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Example of a 3^n design

The following data set are repsonses of a 3^3 design.

```
159   395   149    25   255   251   184   363   378
260   454   112    98   422   270   237   362   363
146   417   150   103   455   172   195   492   278
```

```
   FULLFAC3
                ******************************************
*               3^n Designs                      *
******************************************
     THIS PROGRAM ANALYZES THE RESPONSES OF A 3^N EXPERIMENT.
     Enter file id of responses or press Enter for a list:
     c:\doe\data\d333.dat

     Want to name the variables? (1-8 letters) y/N
     Type 1 to enter your own error variance or
type 2 to use higher order interactions or
     press the Enter key for no standard error: 2

     Enter beginning order of the interactions to be used: 3
     Enter name of output file to store the analysis
     or press Enter for C:\DOE\REPORTS\ANALYSIS.3T

      ANALYSIS ON FILE: c:\doe\data\d333.dat

     Variable      Effects    t value    p value Sum of Squares
     MEAN          264.630     18.777      0.000 1890778.750 *
     A              39.778      2.305      0.016   28480.889 *
     B              33.889      1.963      0.032   20672.223 *
     C              13.833      0.801      0.216    3444.500
     a             -68.519     -6.876      0.000  253518.516 *
     b              18.370      1.843      0.040   18223.408 *
     c             -10.907     -1.095      0.143    6424.463
     AB             46.417      2.196      0.020   25854.084 *
     AC            -21.167     -1.001      0.164    5376.333
     Ab            -19.028     -1.559      0.067   13034.027
     Ac              7.389      0.605      0.276    1965.444
     BC              2.500      0.118      0.454      75.000
     Ba             21.028      1.723      0.050   15918.027
     Bc              5.611      0.460      0.325    1133.444
     Ca            -22.333     -1.830      0.041   17956.000 *
     Cb             -9.667     -0.792      0.219    3364.000
     ab              3.102      0.440      0.332    1039.120
```

Example of a CCD analysis

Consider the following ser of responses:

76.5 78.0 77.0 79.5 75.6 78.4 77.0 78.5 79.9 80.3 80.0 79.7 79.8

Central Composite Designs are often used in Response Surface Methodology (RSM). They will tell us if we have reached an optimum, what is the value of that optimum, whether it is a maximum, minimum or neither (called a saddle point) and if we should have used a CCD in the first place.

We used the NAMING option in this example.
In this example the regression analysis determined that the quadratic model was acceptable, and subsequently
continued with more analyses.

CCD

```
              ******************************************
              *  Analysis of Central Composite Designs  *
              ******************************************
The following Uniform-Precision Central Composite Designs are
available:

#  Number of  Number of  Number of Number of Total        α
   Variables  factorial  star      center    number of
              trials     trials    trials    trials
_____________________________________________________________
_
 2     2         4          4         5         13       1.4142
 3     3         8          6         6         20       1.6820
 4     4        16          8         7         31       2.0000
 5     5        16         10         6         32       2.0000
 6     6        32         12         9         53       2.3780
 7     7        64         14        14         92       2.8280
 8     8       128         16        20        164       3.3640

 Enter number of desired Central Composite Design: 2
 Enter D to output the design or press Enter to continue:
 THIS PROGRAM ANALYZES THE RESPONSES (RESULTS) OF A CCD
EXPERIMENT.
 Enter file id of responses or press Enter to see a list:
 c:\doe\data\doug534.dat
 Want to name the variables? (1-8 letters) y/N: y
 Type K to enter from key board (default)
 or   F to enter from file
 or   H for help

 Enter name of variable: 1  time
 Enter name of variable: 2  temp
 Save the names in file XXXXXXXX.NAM ?  y/N: n

 Enter name of output file to store the analysis
 or press Enter for C:\DOE\REPORTS\CCD.2
 ANALYSIS ON FILE: c:\doe\data\doug534.dat

 Variable      Effects    std.err    t value      p value
 MEAN           79.940      0.103    775.586      0.000
 A               0.995      0.081     12.211      0.000 * time
 B               0.515      0.081      6.322      0.000 * temp
 AB              0.250      0.115      2.169      0.025 *
time*temp     AA             -1.376      0.087    -15.749
```

```
0.000 * time*time     BB              -1.001      0.087
-11.457      0.000 * temp*temp
```

The estimated std.dev. using center points only = 0.230, with df =4
The estimated std.dev. using residuals = 0.266, with df = 7
The center points were used in computing the std.errors of the estimates.

Significant contrasts at the .05 level:
A B AB AA BB

****
* Lack-of-Fit Test *
****

Source	SS	DF	MS	F
Regression	28.235	5	5.647	79.637
Residuals	0.496	7	0.071	
Lack of Fit	0.284	3	0.095	1.782
Pure Error	0.212	4	0.053	

Probability of F Lack-of-Fit = 0.7103

Probability of F Regression = 1.0000

Do not reject the hypothesis of model adequacy

****
* Canonical Analysis *
****

The stationary point is:
0.3892 0.3058

The distance from the design center is: 0.3892

The predicted response at the stationary point is:
80.2124

```
EIGENVALUES OF B
 -1.414      -0.964
The response surface is a maximum

Canonical form of the quadratic model
Y =  80.2124 +  -1.4143 W1^2  +  -0.9635 W2^2

The analysis is stored on disk in file: C:\DOE\REPORTS\CCD.2

Want a normal probability plot of the effects? y/n:    n

More?    y/n:  n
```

Chapter 12

References

1 Douglas C. Montgomery
"Design and Analysis of Experiments"
John Wiley & Sons, 1991

2 William J Diamond
"Practical Experiment Designs"
Van Nostrand-Rheinhold, 1989

3 Stephen R. Schmidt and Robert G. Launsby
"Understanding Industrial Experiments"
Air Academy Press, 1989

4 G.E.P Box and Norman R. Draper
"Emperical Model-Building and Response Surfaces"
John Wiley & Sons, 1987

5 G.E.P Box, William G. Hunter, and J. Stuart Hunter
"Statistics for Experimenters"
John Wiley & Sons, 1978

6 Owen Davies, Editor
"The Design and Analysis of Industrial Experiments
Hafner Publishing, 1963

7 Forrest W. Breyfogle III
"Statistical Methods"
John Wiley and Sons, 1992

Data Files

Here are the listings of the data files used in the examples.

ANOVA1.FIL

4.1	3.9	4.3
2.7	3.1	2.6
3.1	2.8	3.3
1.9	2.2	2.3
3.5	3.2	3.6
2.7	2.3	2.5

ANOVA3.FIL

24	29	25
20	22	18
15	15	12
16	9	11
18	19	23
15	10	11
15	20	13
10	14	6

BOOKF.FIL (also named DEMO1.FIL)

47	71	51	50	48	38	68
64	35	57	71	55	59	38
23	57	50	56	45	55	50
71	40	60	74	57	41	60
38	58	45	50	50	53	39
64	44	57	58	62	49	59
55	80	50	45	44	34	40
41	55	45	54	64	35	57
59	37	25	36	43	54	54
48	74	59	54	52	45	23

BOOKG.FIL

112	115	145	171	196	204	242	284	315	340	360	417
118	126	150	180	196	188	233	277	301	318	342	391
132	141	178	193	236	235	267	317	356	362	406	419
129	135	163	181	235	227	269	313	348	348	396	461
121	125	172	183	229	234	270	318	355	363	420	472
135	149	178	218	243	264	315	374	422	435	472	535
148	170	199	230	264	302	364	413	465	491	548	622
148	170	199	242	272	293	347	405	467	505	559	606
136	158	184	209	237	259	312	355	404	404	463	508
119	133	162	191	211	229	274	306	347	359	407	461
104	114	146	172	180	203	237	271	305	310	362	390
118	140	166	194	201	229	278	306	336	337	405	432

REGRESS1.FIL

20.0	89.5
14.8	79.9
20.5	83.1
12.5	56.9
18.0	66.6
14.3	82.5
27.5	126.3
16.5	79.3
24.3	119.9
20.2	87.6
22.0	112.6
19.0	120.8
12.3	78.5
14.0	74.3
16.7	74.8

REGRESS3.FIL

89.5	20.0	5	4.1
79.9	14.8	10	6.8
83.1	20.5	8	6.3
56.9	12.5	7	5.1
66.6	18.0	8	4.2
82.5	14.3	12	8.6
126.3	27.5	1	4.9
79.3	16.5	10	6.2
119.9	24.3	2	7.5
87.6	20.2	8	5.1
112.6	22.0	7	6.3
120.8	19.0	11	12.9
78.5	12.3	16	9.6
74.3	14.0	12	5.7
74.8	16.7	13	4.8

DOUGEWMA.SQC

10.5	14.5	13.0	10.0
6.0	9.5	9.0	12.0
10.0	12.0	12.0	8.0
11.0	12.5	6.0	9.0
12.5	10.5	12.0	13.0
9.5	8.0	15.0	11.0
6.0	9.5	11.0	9.0
10.0	8.0	7.0	10.0
10.5	10.0	9.5	15.0