CBU Statistical Resources

Compiled by Ian Nimmo-Smith
(Last revised December 4, 2002)

1 Maintained Multipurpose Packages

Bmdp

Bmdp is a suite of programs to implement a wide range of statistical procedures. The facilities range from simple data description to advanced techniques. The interface is rather old-fashioned, with e.g. the powerful repeated measures analysis of variance module invoked under the name **`BMPD4V`**. **bmdp** is available both under Solaris and Windows.

Under Solaris there is a rather grotty graphical interface invoked by the command

```
unix% xbmdp
```

Alternatively the command line version

```
unix% bmdp
```

does the business.

Genstat

a statistical programming language, good for more complicated ANOVAs, balanced or with unequal-n, missing data; analysis of covariance, with linear and non-linear regression; multivariate data procedures; transformation and tabulation; probit and logit analyses; time series modelling; multidimensional scaling. There is a primitive menu-driven front-end to some of the procedures.

Available under Solaris.

Invoked interactively by

```
unix% genstat
```

there advantages to running **Genstat** offline e.g.

```
unix% genstat <[IN=]infile> <[OUT=]outfile>
```

or using more parameters from the following list:

```
IN, OUT
     specify the primary input and output files

IN2, OUT2, ... IN4, OUT4, OUT5
     attach secondary input and output files

BS1 ... BS6
     attach files  for  unformatted  I/O,  using  STORE  and
     RETRIEVE, to channels 1...6

UF1 ... UF4
     attach files for unformatted I/O, using  READ/PRINT  or
     RECORD/RESUME, to channels 1...4

PL1 ... PL3
     attach user Procedure Library files to channels 1...3

D    specifies  the  graphical   device   number,   D=_?n   is
     equivalent to using a DEVICE _?n command within Genstat.

GR   specifies a graphical output file associated  with  the
     device specified using the D parameter.
```

Ian Nimmo-Smith is the local expert.

Glim

a program for generalised linear interactive modelling, embracing conventional multiple regression; log-linear modelling of contingency tables; signal detection theory; probit and logit analysis.

Available under Solaris.

Invoked interactively by

unix% glim

the full range of file options is as follows:

```
glim [pip=primary-input] [pop=primary-output]
     [sip=secondary-input] [sop=secondary-output]
     [log=log-file] [dmp=dump-file]
```

Minitab

interactive smaller-scale statistical analysis. **Minitab** is available both on Solaris and Windows machines. Peter Watson is the local expert.

Sas

a large statistical package for data analysis. Uses an X-windows menus and buttons interface.

Available on Solaris and Windows machines.

Invoked under Solaris by

unix% sas

Splus

a statistical programming language, with many state-of-the-art statistical procedures being implemented by leading researchers in techniques of data analysis and modelling; excellent (interactive) graphing capabilities.

Available on Solaris machines. Invoked by typing **`splus`** to the **`unix%`** prompt.

Ian Nimmo-Smith is the local expert. He has books and manuals if you want to find out more.

Here is a minimal session:

```
unix% splus
...
... startup banner etc.
...
> options(gui='motif')        # to tell splus what window system to use
> help.start()                # to start up a help window
...
... your commands
...
> q()                         # to quit
```

SPSS

a very comprehensive package of procedures for cross-tabulation; correlations and scatter plots; multivariate analyses such as factor analysis, discriminant analysis, cluster and scaling; ANOVA and T-tests. There is a good graphical interface which serves most purposes, though occasionally needed features may require editing a SPSS syntax file directly.

Be aware that when analysing unbalanced (unequal-n) data, or analysis of covariance the default type of sum of squares 'Type III' may not be the option you require. As usual, consulting a friendly neighbourhood statistician can save later embarrassment. A further web page on Sums of Squares is under preparation.

SPSS is available on Windows and MAC machines.

Peter Watson is the local expert.

SuperAnova

a useful application for doing analysis of variance (ANOVA), including graphing of means, multiple comparisons, and simple main effects.

Be aware that when analysing unbalanced (unequal-n) data, or analysis of covariance the default type of sum of squares 'Type III' may not be the option you require. As usual, consulting a friendly neighbourhood statistician can save later embarrassment. A further web page on Sums of Squares is under preparation.

SuperAnova has adopted a cautious approach to issues arising in repeated measures ANOVA, which deprecates various traditional forms of analysis of simple main effects as advocated by **Winer et al. (1991)**. The approach is close to that in **Howell (1997)** or **Maxwell and Delaney (1990)**.

SuperAnova is available on MAC machines.

StatView

a good range of statistical and graphical procedures, including non-parametric tests, and regression.

StatView is available on MAC machines.

Excel

Don't forget you may be able to meet your statistical needs within **Excel**, a powerful spreadsheet. However, there is need for caution in using the built-in statistical procedures: for instance Regression with zero intercept can produce erroneous results.

Excel is available on Windows and MAC machines.

2 Special Purpose Packages

Eqs

Eqs implements Structural Equation Modelling (SEM), which embraces conventional (exploratory) factor analysis as well as more theory driven modelling where the relationship between observed measurements is represented graphically in terms of unobserved latent variables.

Eqs is at present solely available on Peter Watson's Windows machine.

3 Public Domain ANOVA Programs

alice

A program for handling large multi-way tables, doing ANOVAs and regression analyses. Popular with psycholinguists.

Dennis Norris is the local guru.

bw

does repeated measures ANOVA. Complete balanced designs with up to 8 between-subject and 8 within-subject factors.

The user interface could do with a revamp. Suppose you have a '1-between 2-within' design with 4 subjects per group, that there are two levels (groups) of the between subject factor, and that there are respectively 3 and 2 levels of the (within-subject) repeated measures factors, and that you have the measurement data from the following table in a file called (arbitrarily) `expt4a.data`.

		a1 b1	a1 b2	a2 b1	a2 b2	a3 b1	a3 b2
c1	s1	1	3	2	4	3	5
	s2						
	s3						
	s4						
c2	s5	5	5	9	0	4	2
	s6						
	s7						
	s8						

Table 1.1: Dummy dataset for BW example

4 | STAT or PIPE STAT

Pronounced 'pipestat', Gary Perlman's golden oldie suite of statistical tools is again available, in response to overwhelming demand. If you want documentation here are Gary's pages.

- `maketrix` creates a data matrix
- `abut` concatenates data matrices sideways
- `transpose` flips rows and columns
- `colex` extracts columns
- `sort` sorts rows according to specified column keys
- `dm` performs a variety of column oriented data manipulation
- `corr` and `regress` do correlation and regression analyses
- `desc` gives descriptive statistics
- `pair` does paired-data analysis
- `contab` contingency tables and chi-square statistics
- `oneway` and `anova` perform analysis-of-variance

5 Local MRC-CBU statistical utilities

The following programs, written by Ian Nimmo-Smith and Peter Watson, have been developed to meet various needs at CBU over the years. Most of them are interactive, to the extent that just typing the name of the program will prompt you for the necessary information. Most of them can be improved – suggestions are welcome.

Working out significance levels

pvalues

can save you looking up the significance levels of a variety of commonly encountered statistics.

minf

calculates Clarke's minimum F' statistic for generalization from subjects and item separately to subjects and items jointly.

You have (suppose) a 'by subjects' F_1 statistic on (n, d_1) degrees-of-freedom and a 'by items' F_2 on (n, d_2) degrees-of-freedom.

At the unix prompt, type **minf**.

Reply to
Give An F-Value And Its Degrees Of Freedom:-
with F_1 n d_1 (three numbers separated by spaces), and to
And The Second One:-
with F_2 n d_2.

Your $F'_{\min}$ is presented to you with its associated degrees of freedom. These are typically fractional as far as the second (denominator) degrees of freedom is concerned. This comes from the underlying distributional approximations. If reviewers are puzzled, mutter the name 'Satterthwaite' at them and they may become calmer!

signtest

works out the one- and two-sided significance level for a Sign Test. Equivalent to selecting the 'Binomial' option with $p = 0.5$ in **pvalues**.

Discrete distributions

hyper

gives information on the hypergeometic distribution.

Handling Correlations

avcor

calculates average correlations based on Fisher's z-transform.

Here is a sample session:

```
unix% avcor
Average correlation by Fisher's Z transform
Input correlations, one to a line,
terminating with CTRL-D
.56
.87
```

```
.23
Average of  3 correlations =  0.6251
```

equalcor

test for the equality of two correlation coefficients.

partial

calculates partial correlation coefficients.

tetra

calculates tetrachoric correlation coefficients.

Contingency tables

chisq

does two-way contingency table analyses, including the χ^2 statistic and calculation of residuals.

fisher.exact

performs a Fisher's Exact Test analysis of a 2×2 table.

fishrc

performs a Fisher's Exact Test analysis of $r \times c$ table.

kappa

yields Cohen's Kappa measure of agreement between a pair of categorical raters.

According to Fleiss (1981, p218), Landis and Koch (1977) have characterized different ranges of values for kappa:

Greater than 0.75 = "excellent agreement beyond chance"

Below 0.4 = "poor agreement beyond chance" Between 0.4 amd 0.75 = "fair to good agreement beyond chance"

Fleiss, J.L. (1981) Statistical Methods for Rates and Proportions (2nd ed.) New York:Wiley

Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data, Biometics, 33, 159-174

Information theory stuff

choose

calculates some information-theoretic measures of the complexity of a combinatorial search problem.

h_

h_digits and **h_letters** perform information-theoretic analyses of data arising form random digit/letter generation tasks.

Multiple comparison aids

mc

assists in the performance of multiple-comparison tests, such as the Newman-Keuls procedure, based on the studentized range statistic.

This version assumes that there are an equal number of observations per mean.

For unequal n data try **mcneq**.

mcneq

assists in the performance of multiple-comparison tests, such as the Newman-Keuls procedure, based on the studentized range statistic.

This version allows for unequal numbers of observations per mean.

For equal n data try **mc** for simplicity, though the programs will agree in this case.

Simple Main Effects assisted by mc

Simple Main Effects

(in SPSS use results from 'sphericity assumed')

Suppose you have a significant interaction between a factor F and a bunch of stuff H. H may be one or more factors. Let B be the Between subject factors in H, and let W be the Within Subject factors in H.

We want to look at the Simple Main Effect (SME) of F at some specified combination of levels of the factors in H.

Examples ...

If F is a Within Subjects factor:

Construct a Pooled Error Sum of Squares (PESS) by adding together all the error sums of squares involving F and any one or more of the terms in B. Construct a Pooled Error Degrees of Freedom (PEDF) in the same way.

Identify the means corresponding to the SME that you are wanting to examine. Work out the number ND of raw data values that have been used in these means. If there are Within Subject factors not included in F or H then you need to take account of collapsing over these.

In a window on a unix machine, type '**mc**', and you will get

```
>  mc: assists with multiple comparisons analyses
>  Input the number of means for comparison
```

Input the number of levels of F

```
>  Input the means, in ANY order
>        1:  mean for condition 1
>        2:  mean for condition 2
>        ...
```

When you have finished, mc prints the means out in ascending order.
To the question

```
>  How many observations per mean ?
```

you reply with ND (see above).
To the question

```
>  Input the Error Sum of Squares (possibly pooled)
```

you reply with PESS
To the question

```
>  Input its Degrees of freedom (d.f.)
```

you reply with PEDF
Here is a sample session and its output

```
 mc
mc: assists with multiple comparisons analyses


Input the number of means for comparison
4
Input the means, in ANY order
  1: 8.5
  2: 8.1
  3: 2.5
  4: 2.9
Multiple comparisons
Means in ascending order
[   3]   2.500
[   4]   2.900
[   2]   8.100
[   1]   8.500
How many observations per mean ? 115
```

```
Input the Error Sum of Squares (possibly pooled) 630
Input its Degrees of freedom (d.f.)342
Studentised range statistics (q) and (APPROXIMATE p-values)


 [ 3, 4] [ 3, 2] [ 3, 1]
   3.160  44.247  47.407
  (0.025) (0.000) (0.000)

          [ 4, 2] [ 4, 1]
           41.086  44.247
           (0.000) (0.000)

                  [ 2, 1]
                    3.160
                   (0.025)
>
```

If F is a Between Subjects factor then:

peritz

performs the Peritz multiple comparison procedure.

Signal Detection Measures

ratings

performs Signal Detection Theory analysis for the case of 2 or more ordered responses (e.g. Yes, Unsure, No).

sigdet

gives d' and β statistics from hit-rate and false-alarm-rate data.

Miscellany

mortal

assess repeated binary (e.g. success/fail survive/die) data for constant failure rate for independent of length of phase of continuous successes.

spread

applies a gap test for the detection of outliers.

slink

does single-linkage cluster analysis of similarity data. Now more easily handled in **SPSS**, **Genstat** or **Splus**.

6 Using standard Unix tools

If you are more adventurous and like a more programming-orientated way of doing data manipulation and analysis, you might like to get to know **`awk`**, **`sort`** and/or **`perl`**.

awk

A Unix system for the processing of files of information. Programs are written in terms of a grammar of regular expressions with a C-like syntax. Useful for extraction of information from files containing output of other programs, or for textual, string-oriented manipulation of data, as well as writing 'quickie' programs for data processing. The book *The **`AWK`** Programming Language* by Aho, Kernighan and Weinberger shows the scope of this tool.

sort

The **`sort`** utility can be used to perform complex reordering and restructuring of data.

perl

In recent years, **`perl`** has taken over from **`awk`** as the Unix guru's language for writing scripts to manipulate textual databases.

7 Graphing your Data

This section is in preparation.

Many of the statistical packages mentioned in this booklet have some way of graphing various aspects of your data using *e.g.* line graphs, bar charts, scatter plots etc. These tend to be rather inflexible and may not allow you to represent produce the graphical representation that you seek.

Excel can be used to graph data. There are versions for both Windows and MAC machines.

Kaleidograph combines the resources of a spreadsheet with a very flexible graphing package. It runs only on Windows machines.

The most commonly used specialist graphing applications on the MAC are . . .

Under Solaris, **Splus** has very flexible graphing facilities allowing the user to build up a plot either from scratch or by tailoring the defaults in an off-the-peg routine.

8 Text Processing with Formulae and Equations

This section is in preparation.

Word on MAC and Windows machines have a method of creating and incorporating mathematical symbols into a document.

By contrast, TEX (and its popular dialect LATEX) are 'markup' text processing systems, where a 'markup language' has to be learnt. The commands of this language are embedded in the text of your document. Editing, processing and viewing the document are three distinct steps in the development of your final version. The cost of learning the language is compensated by producing professional quality complicated maths typography. An increasing number of publishing houses (OUP, CUP, Springer) are publishing books and journals directly from LATEX. They supply the author with the **.sty** files which contain the macros and defaults for making your document appear in the appropriate 'house style'.

Versions of TEXare available on all platforms.

9 Mathematical Modelling Tools

Mathematica

A MAC or Solaris graphical front-end give access to a kernel which 'knows' a vast amount of maths, and can be taught more. Can explore mathematically described formulae with powerful tools for manipulating and representing logical, symbolic, algebraic and numerical structures. Excellent graphing abilities.

MatLab

LabView

A MAC-based package for the manipulation of models of linear systems, with powerful facilities for handling vectors and matrices.

10 On-line documentation

This section is under development.

Many of the Solaris, MAC and Windows programs have their own **help** systems, once you have got the application running.

On the Solaris the **`man`** (short for 'manual') command followed by the name of the program will produce available information on that program. For instance **`man genstat`** will give information about running **Genstat** package under Solaris.

11 Porting Data between Application and Platforms

This section is in preparation.

12 What More?

Is there anything you would like to have found mentioned here but didn't?
Let Ian Nimmo-Smith or Peter Watson know.