Moving on from MSTAT

March 2000

University of Reading
Statistical Services Centre

Biometrics Advisory and
Support Service to DFID

Contents

1. Introduction

This guide has been written as part of our work for DFID to encourage good statistical practice. It is to assist research teams in planning their computing strategy.

MSTAT has been effective in many developing countries, in introducing statistical computing to staff who have had no previous experience. It remains a useful package for the design of trials and for some of its specialist facilities, in particular for breeding programs.

However, MSTAT is limited in its facilities for the analysis of experimental data and relatively cumbersome to use for the Analysis of Variance, compared to the ease of use of modern statistical software under Windows. A recent development is that at least one modern statistics package can read MSTAT files directly. This increases the flexibility of analyses both for staff who have historical data in MSTAT files and for those who would like to combine their use of MSTAT with the transfer to more powerful software when the need arises.

This guide illustrates some advantages in using a more powerful package than MSTAT for the design and analysis of experimental data. We use Genstat for this purpose for two reasons. Firstly, Genstat is an excellent package for the analysis of data from agricultural experiments and secondly Genstat has provided the facility for the transfer of MSTAT files. We believe that it is now much easier to learn a new Windows package, compared to a few years ago and that full exploitation of experimental data needs access to a modern statistical package, such as Genstat.

However, the points of principle in this guide are not limited to Genstat. The program used for the transfer of MSTAT files is available free of charge and can be used to transfer MSTAT data to ASCII or to Excel files, which can then be read into other packages.

We use a set of 3 examples:

Simple one-way Analysis of Variance, using both MSTAT and Genstat.

The Analysis of Variance of an experiment with factorial treatment structure.

Designing an experiment.

This guide is written so users are able to try the analyses themselves if they wish. Sufficient detail has been given for project managers, who have some experience in the use of computers for statistical work, to be able to understand the issues simply by reading this guide.

2. Moving from MSTAT to Genstat

The CUCUMBER data are described on page 9-1 of the MSTAT-C Manual. They are used here to show the steps involved in calculating one-way analysis of variance in both MSTAT and Genstat.

2.1 Analysis in MSTAT

Open the CUCUMBER data file, which has two columns called POPULATION and FRUIT_NO. Perform a one-way analysis of variance (ANOVA-1), by completing the following screens:

Fig. 2.1 MSTAT input screens

```
Select the group variable
+- ANOVA-1 ---------------------------------------------------+
¦                                                             ¦
¦   Enter the number of the GROUP variable (1-2):   1         ¦
¦                                                             ¦
¦ Enter the lowest and highest value in the GROUP variable ¦
¦           Lowest: 1              Highest: 11                ¦
¦                                                             ¦
¦           (Press <F1> for a list of Variables)              ¦
+-------------------------------------------------------------+

Select a sub-set of the data
 +- Get Case Range ---------------------+
 ¦                                      ¦
 ¦  +- Case Range 1 - 66 ----------+s.  ¦
 ¦  ¦                              ¦    ¦
 ¦  ¦  First selected case   1     ¦    ¦
 ¦  ¦                              ¦    ¦
 +--¦  Last selected case    55    ¦----+
    ¦                              ¦
    +------------------------------+

Choose the variable to be analysed
+- Choose up to 1 variables (Press ESC to quit) -+
| 01 (NUMERIC) POPULATION                        ¦
¦⌫02 (NUMERIC) FRUIT NO                          ¦
+------------------------------------------------+
```

```
To perform single DF orthogonal comparisons
 +- ANOVA-1 --------------------------------------+
 ¦                                                ¦
 ¦  Enter the values for the following:           ¦
 ¦   (Press <ESC> to abandon, <F10> to finish)    ¦
 ¦                                                ¦
 ¦        Treatment Number     Coefficient        ¦
 ¦                1                -5.00          ¦
 ¦                2                -4.00          ¦
 ¦                3                -3.00          ¦
 ¦                4                -2.00          ¦
 ¦                5                -1.00          ¦
 ¦                6                 0.00          ¦
 ¦                7                 1.00          ¦
 ¦                8                 2.00          ¦
 ¦                9                 3.00          ¦
 ¦               10                 4.00          ¦
 ¦               11                 5.00          ¦
 ¦                                                ¦
 ¦ (The treatment coefficients must sum to zero.) ¦
 ¦                                                ¦
 +------------------------------------------------+
```

Fig. 2.2 MSTAT output from one-way analysis of variance

```
         Data file: CUCUMBER
     Function: ANOVA-1
     Data case no. 1 to 55

     One way ANOVA grouped over variable 1 (POPULATION)
     with values from 1 to 11.

     Variable 2 (FRUIT NO)

      A N A L Y S I S   O F   V A R I A N C E   T A B L E
         Degrees of    Sum of          Mean
          Freedom      Squares          Square          F-value    Prob.
---------------------------------------------------------------------------
Between      10        967.345          96.735          34.888    0.0000
Within       44        122.000           2.773
--------------------------------------------------------------------------
Total        54       1089.345
```

```
          Coefficient of Variation = 8.21%

       Var.     V A R I A B L E   No. 2
         1      Number        Sum      Average          SD          SE
     ------------------------------------------------------------------
          1      5.00      104.000     20.800         2.17        0.74
          2      5.00       59.000     11.800         1.30        0.74
          3      5.00      113.000     22.600         2.51        0.74
          4      5.00       99.000     19.800         1.64        0.74
          5      5.00       77.000     15.400         1.14        0.74
          6      5.00      127.000     25.400         1.14        0.74
          7      5.00       88.000     17.600         1.14        0.74
          8      5.00      126.000     25.200         1.79        0.74
          9      5.00      111.000     22.200         1.30        0.74
         10      5.00      125.000     25.000         2.00        0.74
         11      5.00       87.000     17.400         1.52        0.74
------------------------------------------------------------------------
     Total      55.00     1116.000     20.291         4.49        0.61
     Within                                           1.67

     Bartlett's test
     ---------------
     Chi-square = 5.889
     Number of Degrees of Freedom = 10
     Approximate significance = 0.825
     -------------------

     Treat.    Coeff.     Treat.    Coeff.
        1      -5.00         7       1.00
        2      -4.00         8       2.00
        3      -3.00         9       3.00
        4      -2.00        10       4.00
        5      -1.00        11       5.00
        6       0.00

     Sum of Squares:    102.989
     Effect:              0.433
     Error:               0.071
     F value:            37.144
     Probability:         0.000
```

The output in Fig. 2.2 is as shown on pages 9.4 and 9.5 of the MSTAT manual.

There are 66 lines of data in the file; however, as shown in Fig. 2.1 only the first 55 lines are used in the analysis, because the final 11 lines contain the means at each of the 11 treatment levels.

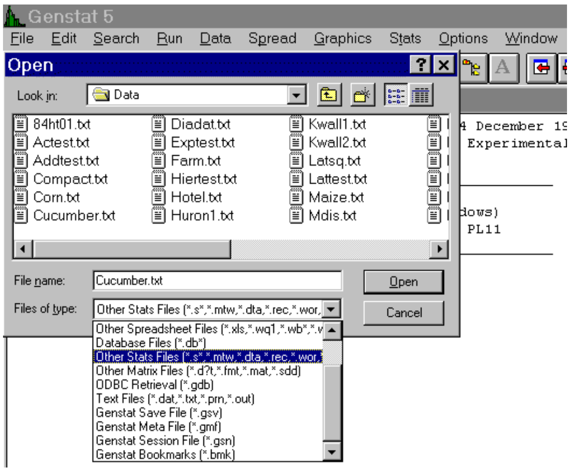
In the analysis, we note that the results from MSTAT give the ANOVA table, plus the coefficient of variation of the experiment, which is 8.21%. The means are also given for each of the 11 levels of the treatment factor; they vary from 11.80 to 25.2. The standard error of each mean is given as 0.74.

2.2 Organising the data in Genstat

We now consider the same analysis in Genstat. We use this example also to introduce Genstat.

The first step is to input the data. Within Genstat, click **Open** on the file menu and move to the folder holding your MSTAT data file (Fig. 2.3). In this example, the cucumber data are loaded into Genstat by highlighting the file **\Mstatc\data\cucumber.txt**.

Fig. 2.3 Opening an MSTAT data file



Look in: Data

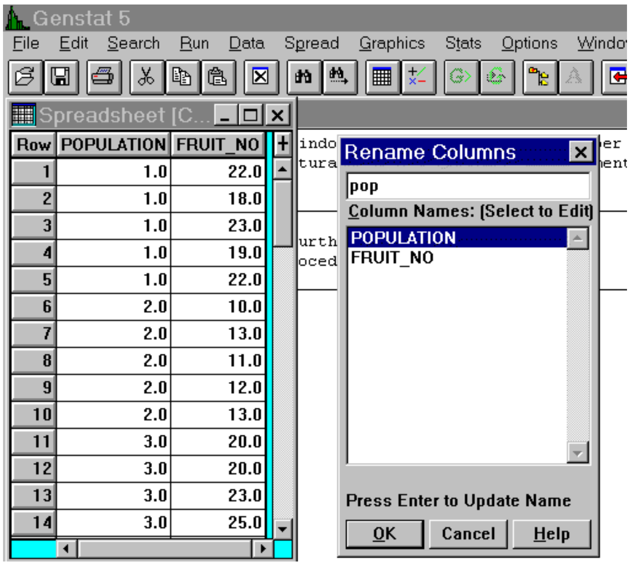
File name: Cucumber.txt

Files of type: Other Stats Files (*.s*,*.mtw,*.dta,*.rec,*.wor,

The remaining preliminary steps are to rename the columns (Fig. 2.4) to delete some rows in the MSTAT file (Fig. 2.5) and to convert the first column to be a factor (Fig. 2.6).

In Genstat, a variable name cannot exceed 8 characters. As a first task we rename "POPULATION" to "pop" and "FRUIT_NO" to "fruit". Fig. 2.4 is obtained from the **Spread** menu, by clicking **Column**, **Rename**. Note that variable names are case-sensitive, i.e. pop *is not* the same as POP.

Fig. 2.4 Renaming columns



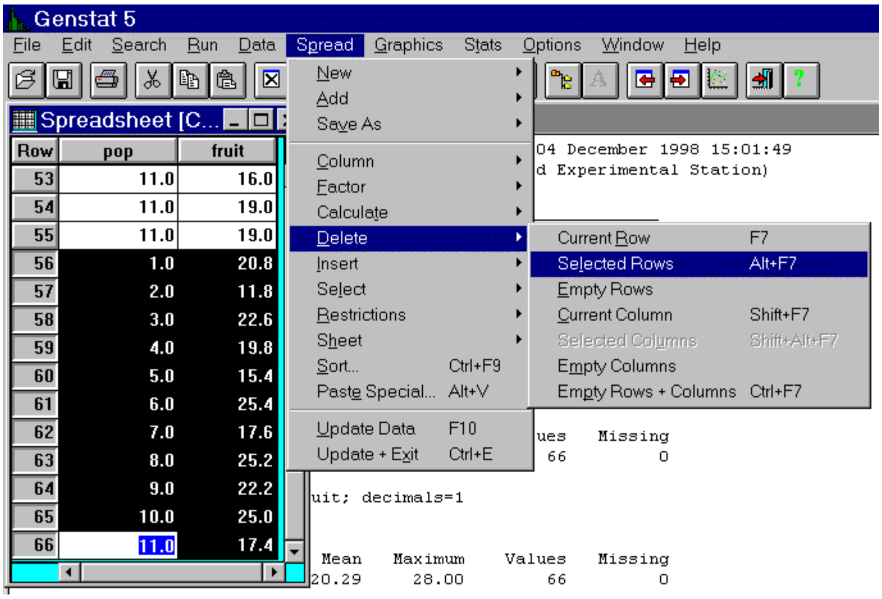
Row	POPULATION	FRUIT_NO
1	1.0	22.0
2	1.0	18.0
3	1.0	23.0
4	1.0	19.0
5	1.0	22.0
6	2.0	10.0
7	2.0	13.0
8	2.0	11.0
9	2.0	12.0
10	2.0	13.0
11	3.0	20.0
12	3.0	20.0
13	3.0	23.0
14	3.0	25.0

pop

POPULATION
FRUIT_NO

As a second task we delete rows 56 to 66 from the spreadsheet. In MSTAT, the cucumber file had 66 rows, but rows 56 to 66 were the means of the 11 levels of the population factor.

Fig. 2.5 Deleting extra rows

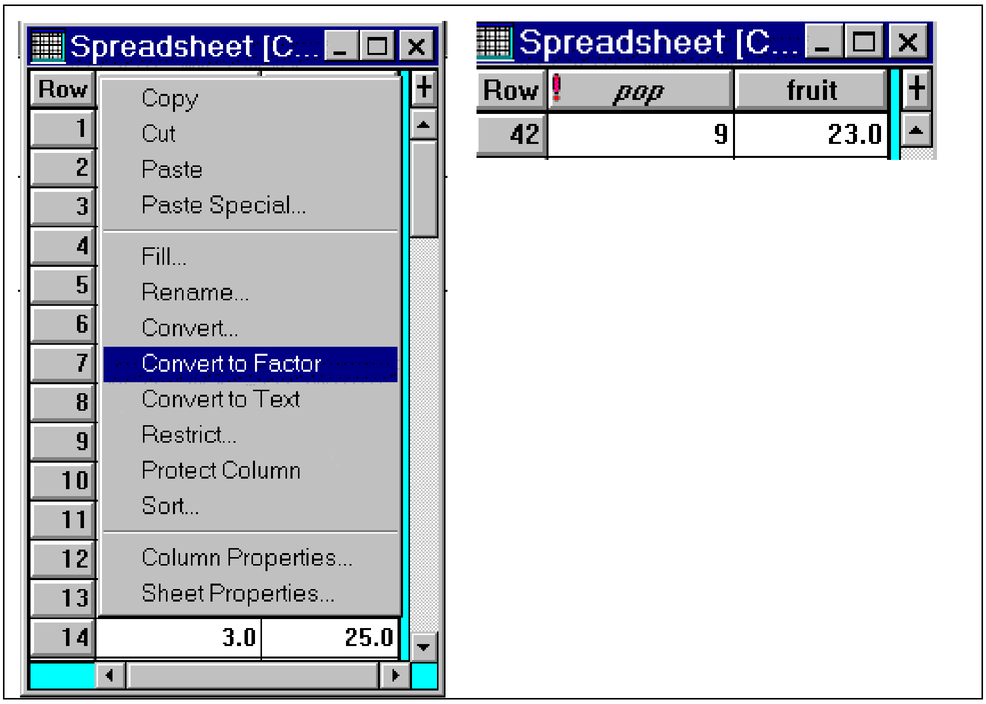


Row	pop	fruit
53	11.0	16.0
54	11.0	19.0
55	11.0	19.0
56	1.0	20.8
57	2.0	11.8
58	3.0	22.6
59	4.0	19.8
60	5.0	15.4
61	6.0	25.4
62	7.0	17.6
63	8.0	25.2
64	9.0	22.2
65	10.0	25.0
66	11.0	17.4

```
04 December 1998 15:01:49
d Experimental Station)

ues    Missing
 66          0

uit; decimals=1

 Mean   Maximum     Values    Missing
20.29     28.00         66          0
```

Before we can do an analysis of variance, we have to convert the column "pop" to be a factor column. To do this, highlight the pop column, click the Right mouse button and select "Convert to factor from menu. Note that to show that it is a factor column, the name, pop, is now preceded by a ! and is in italic.

Fig. 2.6 Converting a column to be a factor



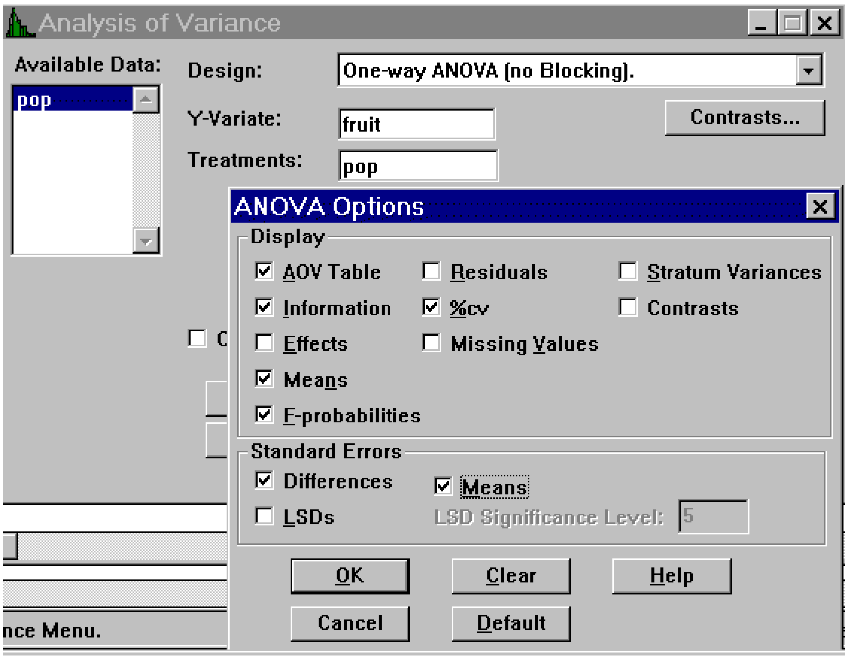
Row		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14	3.0	25.0

Row	! *pop*	fruit
42	9	23.0

2.3 Basic ANOVA in Genstat

We are now ready to perform a one-way analysis of variance. From the Stats menu, select Analysis of Variance, One-way ANOVA, (no Blocking).

Fig. 2.7 Running a one-way ANOVA



Analysis of Variance

Available Data: pop

Design: One-way ANOVA (no Blocking).

Y-Variate: fruit

Treatments: pop

ANOVA Options

Display

☑ AOV Table
☐ Residuals
☐ Stratum Variances
☑ Information
☑ %cv
☐ Contrasts
☐ Effects
☐ Missing Values
☑ Means
☑ F-probabilities

Standard Errors

☑ Differences
☑ Means
☐ LSDs

LSD Significance Level: 5

In Fig. 2.7 we have used the Options button and also ticked %cv and the Standard Errors of the Means boxes. The output is in Fig. 2.8.

Fig. 2.8 Output from one-way analysis of variance

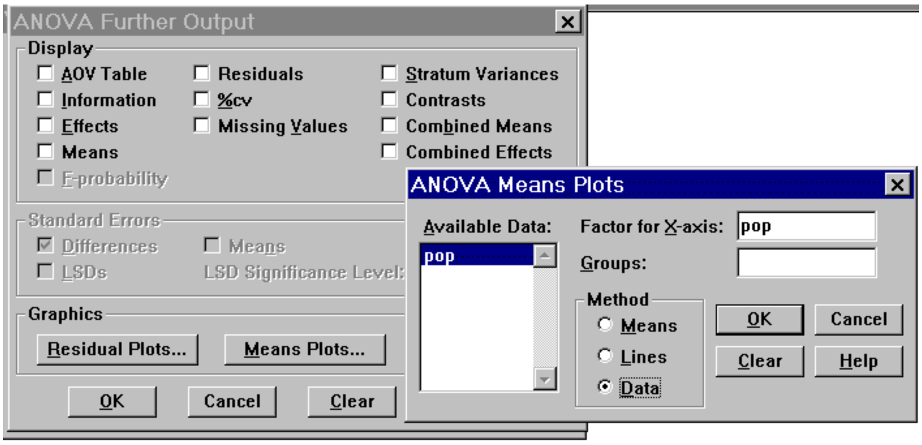
```
***** Analysis of variance *****
Variate: fruit

Source of variation      d.f.      s.s.      m.s.    v.r.  F pr.
pop                        10   967.345    96.735   34.89  <.001
Residual                   44   122.000     2.773
Total                      54  1089.345

***** Tables of means *****

Variate: fruit

Grand mean  20.3

      pop        1        2        3        4        5        6        7
              20.8     11.8     22.6     19.8     15.4     25.4     17.6

      pop        8        9       10       11
              25.2     22.2     25.0     17.4
```

```
*** Standard errors of means ***

Table                  pop
rep.                     5
d.f.                    44
e.s.e.                0.74

*** Standard errors of differences of means ***

Table                  pop
rep.                     5
d.f.                    44
s.e.d.                1.05



***** Stratum standard errors and coefficients of variation *****

Variate: fruit
   d.f.          s.e.         cv%
     44          1.67         8.2
```

2.4 Adding a graph with the fitted means

Once an analysis has been run, **Further Output** can be selected. We select the **Means Plots** … option and choose to include the data in the plot as shown in Fig. 2.9.

Fig. 2.9 Plotting the data and their means

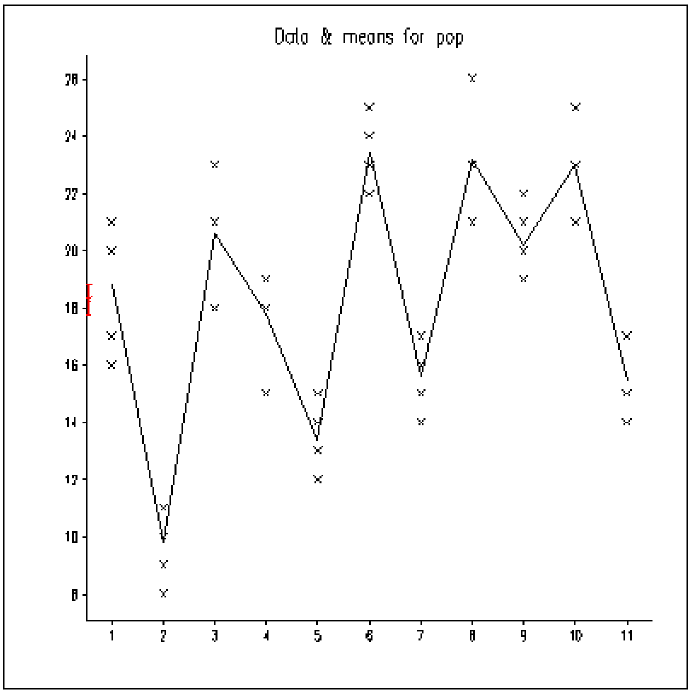


☑ Differences

Factor for X-axis: pop

Method: ● Data

Fig. 2.10 **Plot of Cucumber data and the pop means**

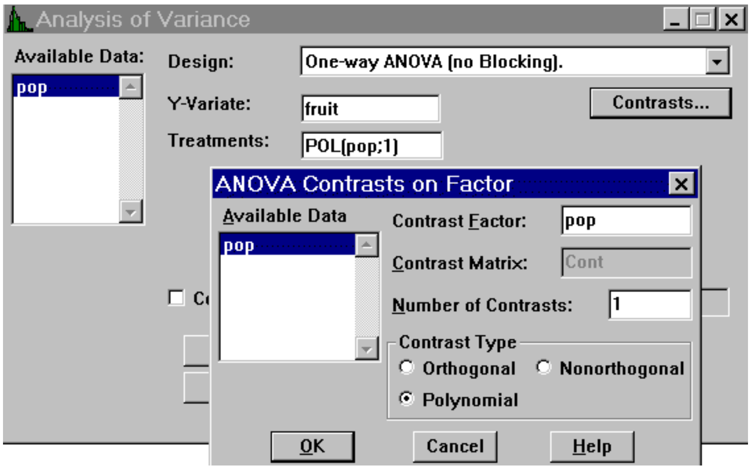


This plot has a “saw-tooth” effect that indicates a linear contrast will not be particularly effective. We produce it, nevertheless, to emulate the results from MSTAT.

2.5 Contrasts

To get the linear contrast in Genstat, we re-run the analysis and use the **Contrasts** button. In the dialogue shown in Fig. 2.11 we specify that we would like a polynomial coefficient for the treatment of factor. We ask for 1 contrast to just give the linear effect, 2 would give quadratic, and so on.

Fig. 2.11 **Input needed to obtain linear contrasts**



Design: One-way ANOVA (no Blocking).

Y-Variate: fruit

Treatments: POL(pop;1)

Contrast Factor: pop

Contrast Matrix: Cont

Number of Contrasts: 1

Contrast Type

○ Orthogonal ○ Nonorthogonal

● Polynomial

Fig. 2.12 **Output showing the ANOVA table with linear contrasts**

```
***** Analysis of variance *****
Variate: fruit
Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.
pop                       10    967.345     96.735   34.89  <.001
  Lin                      1    102.989    102.989   37.14  <.001
  Deviations               9    864.356     96.040   34.64  <.001
Residual                  44    122.000      2.773
Total                     54   1089.345

***** Tables of contrasts *****
Variate: fruit
*** pop contrasts ***
Lin      0.43    s.e. 0.071     ss.div. 550.
Deviations    e.s.e. 0.74    ss.div. 5.00
```

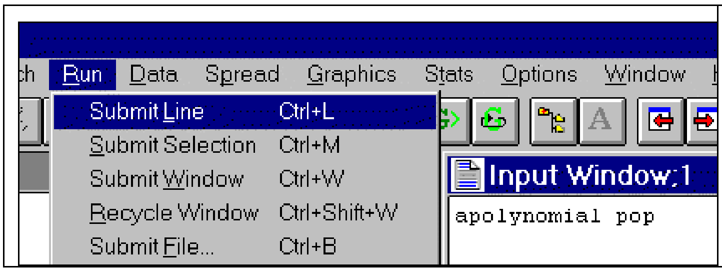
The results are the same as in Fig. 2.2 from MSTAT, but they also show the effect of the contrast in their context in the ANOVA table. The 'Deviations' line confirms the evidence from the graph in Fig. 2.10 that, though the linear contrast is highly significant, it is not an effective way of explaining the treatment effect.

2.6 Submitting a command to Genstat

Occasionally the required output cannot be obtained using the menus and dialogue boxes. It might be useful to give the full equation of the line for the linear contrast (though this is not in the MSTAT output). This can be obtained by typing the command directly into a text window as shown in Fig. 2.13. From the **File menu**, select **New, Text Window** and in the Input window, type **apolynomial pop**.

To run this command, select **Run, Submit Line** as shown in Fig. 2.13.

Fig. 2.13 **Submitting a Genstat command and its output**



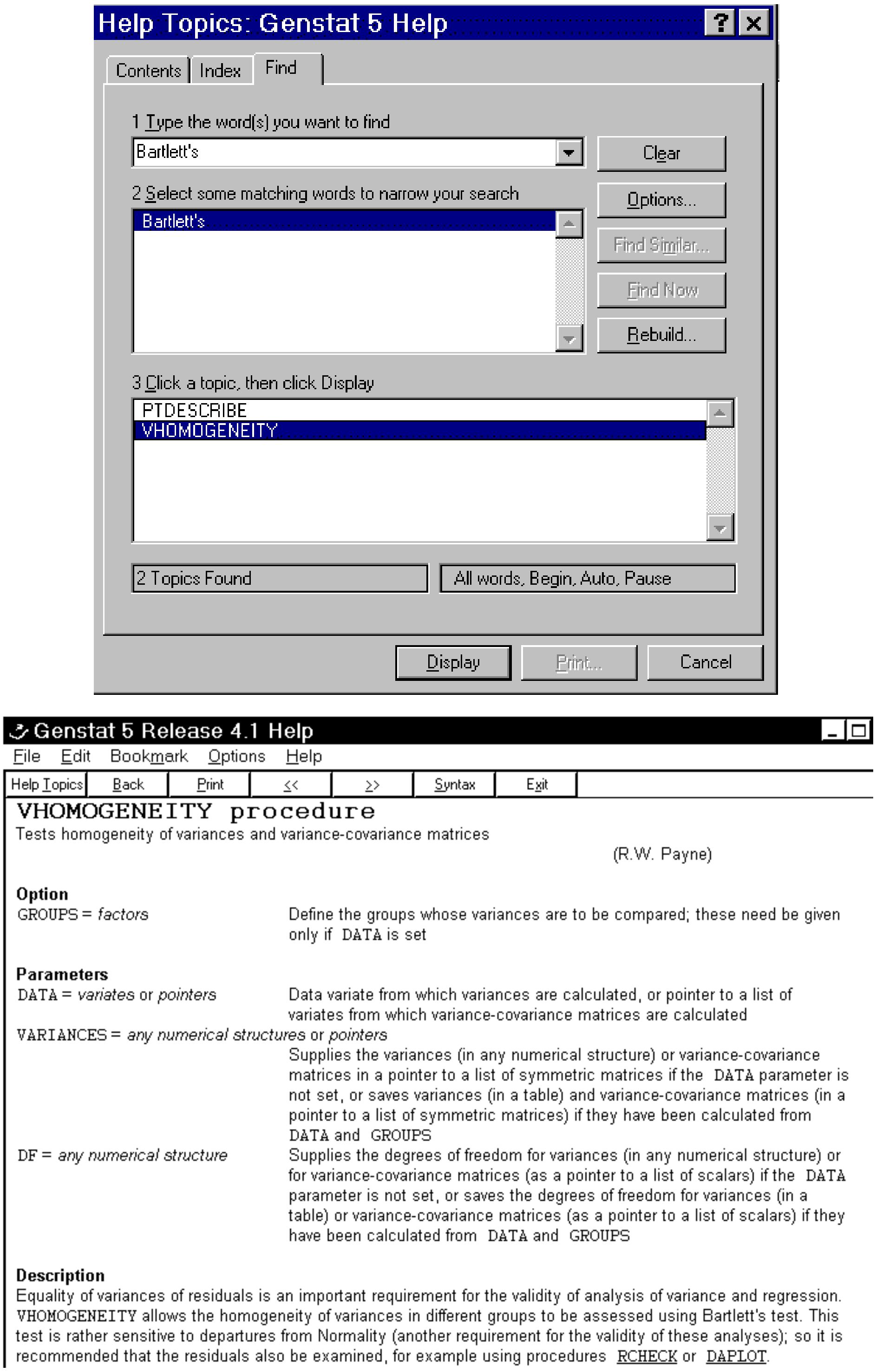
```
apolynomial pop
```

```
46   apolynomial pop

 *** Equation of the polynomial ***

 17.695 + 0.433   * pop
```

2.7 Accessing HELP in Genstat

The MSTAT output also included Bartlett's test. We use this example to show how to use Genstat's Help. From the **Help** menu, select **Search** for help on **Find** and type in Bartlett's. Choose VHOMOGENEITY, as shown in Fig. 2.14 and click **Display**.

Fig. 2.14 **Using Genstat's Help menu**



1 Type the word(s) you want to find: Bartlett's

2 Select some matching words to narrow your search: Bartlett's

3 Click a topic, then click Display:
PTDESCRIBE
VHOMOGENEITY

2 Topics Found

All words, Begin, Auto, Pause

VHOMOGENEITY procedure

Tests homogeneity of variances and variance-covariance matrices

(R.W. Payne)

Option

GROUPS = *factors*	Define the groups whose variances are to be compared; these need be given only if DATA is set

Parameters

DATA = *variates* or *pointers*	Data variate from which variances are calculated, or pointer to a list of variates from which variance-covariance matrices are calculated
VARIANCES = *any numerical structures* or *pointers*	Supplies the variances (in any numerical structure) or variance-covariance matrices in a pointer to a list of symmetric matrices if the DATA parameter is not set, or saves variances (in a table) and variance-covariance matrices (in a pointer to a list of symmetric matrices) if they have been calculated from DATA and GROUPS
DF = *any numerical structure*	Supplies the degrees of freedom for variances (in any numerical structure) or for variance-covariance matrices (as a pointer to a list of scalars) if the DATA parameter is not set, or saves the degrees of freedom for variances (in a table) or variance-covariance matrices (as a pointer to a list of scalars) if they have been calculated from DATA and GROUPS

Description

Equality of variances of residuals is an important requirement for the validity of analysis of variance and regression. VHOMOGENEITY allows the homogeneity of variances in different groups to be assessed using Bartlett's test. This test is rather sensitive to departures from Normality (another requirement for the validity of these analyses); so it is recommended that the residuals also be examined, for example using procedures RCHECK or DAPLOT.

This gives us the syntax for the command, so returning to the Input window, type

vhomogeneity [group=pop] data=fruit

as shown in Fig. 2.15, highlight the line and select **Run, Submit Line**. Note that each time a Genstat command is run, either from the menus or from the Input text window, the commands are displayed in the Output window and given a line number. In Fig. 2.15, this is line 14.

Fig. 2.15 Bartlett's Test

```
14  vhomogeneity [group=pop] data=fruit

 *** Bartlett's Test for homogeneity of variances ***

 Chi-square  5.89   on 10 degrees of freedom: probability 0.8245
```

We have now produced in Genstat all the output that was shown in Fig. 2.2 from MSTAT.

2.8 Good practice

We conclude this section by reviewing some "good-practice" elements in the analysis. We began in Genstat by renaming the columns and then deleted the last 11 lines in the data file (Fig. 2.15). These 11 lines contain the means of each treatment, saved from a previous analysis. It is good-practice that MSTAT permits users to save the treatment means, but unfortunate that they are saved as part of the data. Genstat has a **[SAVE]** button following the ANOVA where all aspects of the analysis can be saved.

The basic presentation of the results by MSTAT (Fig. 2.2) and Genstat (Fig. 2.8) are both acceptable. Genstat offers the useful addition of plotting the means (Fig. 2.10) and also of plotting the residuals.

Both MSTAT (Fig. 2.1) and Genstat (Fig. 2.11) allow the treatment effects to be examined in more detail, using contrasts. For polynomials it is easier with Genstat because the user does not have to type each coefficient. Genstat's facilities are also more powerful and are used in the same way, even if the levels of the factor were unequally spaced.

The display of the results for the contrasts is better in Genstat, (Fig. 2.12), in that they show its effect in the ANOVA table. With MSTAT this would have to be worked out by hand.

3. A Factorial Design

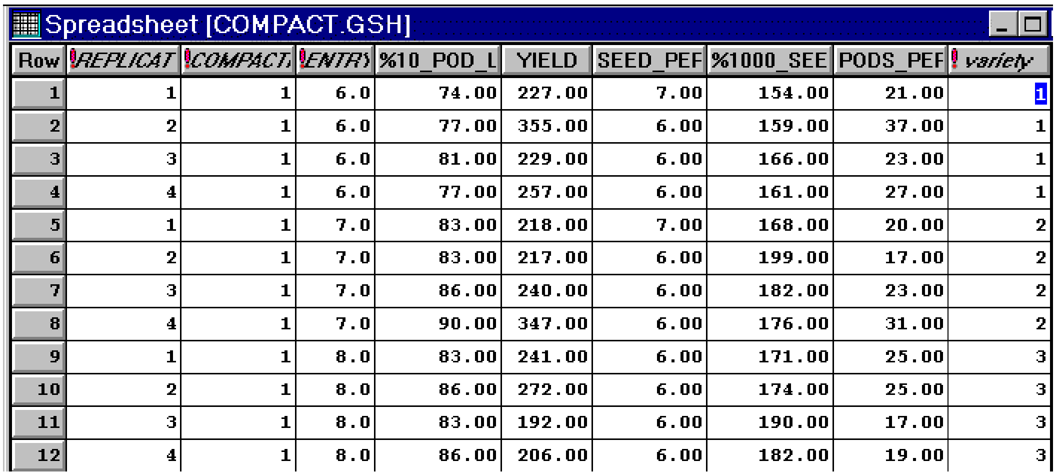
The second example uses the MSTAT data file called COMPACT, from the MSTAT guide, Section 9.4. This is a split plot design with 4 replications. The main plot factor consists of 2 compaction levels, with 15 dry bean varieties as the subplot factor.

We use this example to demonstrate the analysis of an experiment with multiple factors and also to show the importance of using a package that encourages a critical analysis of the data. In the MSTAT guide the analysis is given for variable 8, the number of pods per plant. We choose here to analyse variable 7, the 1000-seed weight.

3.1 Reading the data into Genstat

The data are read directly from the MSTAT file as shown in the previous example, and are as shown below. We have specified the factor columns, but otherwise the data are essentially as imported. One problem with Genstat is that the column names must begin with a letter and cannot contain spaces. Hence the column we will analyse, which is the 1000-grain weight, is given the name %1000_SE

Fig. 3.1 Compact data imported from MSTAT

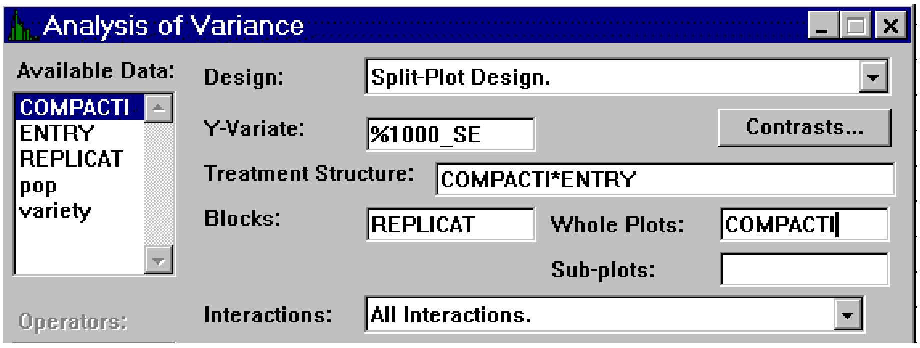


Row	!REPLICAT	!COMPACT	!ENTRY	%10_POD_L	YIELD	SEED_PEF	%1000_SEE	PODS_PEF	! variety
1	1	1	6.0	74.00	227.00	7.00	154.00	21.00	1
2	2	1	6.0	77.00	355.00	6.00	159.00	37.00	1
3	3	1	6.0	81.00	229.00	6.00	166.00	23.00	1
4	4	1	6.0	77.00	257.00	6.00	161.00	27.00	1
5	1	1	7.0	83.00	218.00	7.00	168.00	20.00	2
6	2	1	7.0	83.00	217.00	6.00	199.00	17.00	2
7	3	1	7.0	86.00	240.00	6.00	182.00	23.00	2
8	4	1	7.0	90.00	347.00	6.00	176.00	31.00	2
9	1	1	8.0	83.00	241.00	6.00	171.00	25.00	3
10	2	1	8.0	86.00	272.00	6.00	174.00	25.00	3
11	3	1	8.0	83.00	192.00	6.00	190.00	17.00	3
12	4	1	8.0	86.00	206.00	6.00	182.00	19.00	3

3.2 The initial ANOVA

It is good statistical practice to look at the data before analysis, but we initially ignore this, because our main aim is to demonstrate the use of the ANOVA facilities. The dialogue is completed as shown below.

Fig. 3.2 Split-Plot Design



COMPACTI
ENTRY
REPLICAT
pop
variety

Design: Split-Plot Design.
Y-Variate: %1000_SE
Treatment Structure: COMPACTI*ENTRY
Blocks: REPLICAT
Whole Plots: COMPACTI
Sub-plots:
Interactions: All Interactions.

We do not give the treatment means at this stage, because Genstat draws our attention to a residual that is very large. This is marked in bold in Fig. 3.3. It is a residual of 84.25, with a standard error of 11.98. We normally find that when a residual is more than 4 times the standard error, then something is clearly odd. Here it is about 7 standard errors. Another interpretation is that 84.25^2 is about 7000. The residual sum of squares is given below as 17218, of which about 40% is therefore due solely to this one observation.

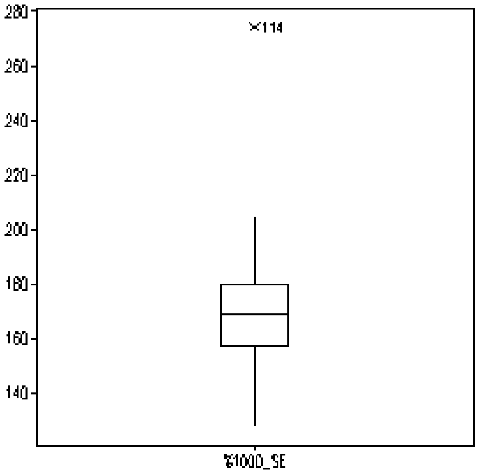
Fig. 3.3 Anova output

```
***** Analysis of variance *****    Variate: %1000_SED WEIGHT

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

REPLICAT stratum           3      755.6      251.9    9.37

REPLICAT.COMPACTI stratum
COMPACTI                   1     4165.4     4165.4  154.99  0.001
Residual                   3       80.6       26.9    0.13

REPLICAT.COMPACTI.*Units* stratum
ENTRY                     14    15973.9     1141.0    5.57  <.001
COMPACTI.ENTRY            14     2904.2      207.4    1.01  0.450
Residual                  84    17217.6      205.0

Total                    119    41097.3

* MESSAGE: the following units have large residuals.

REPLICAT 1    COMPACTI 2    *units* 8          -32.90    s.e. 11.98
REPLICAT 2    COMPACTI 2    *units* 14          84.25    s.e. 11.98
REPLICAT 3    COMPACTI 2    *units* 14         -36.02    s.e. 11.98
```

3.3 Analysis with “odd” observations treated as missing values

It is easy to do boxplots in Genstat, using Graphics ⇒ Boxplots and completing the dialogue. The result is shown in Fig. 3.4 and confirms the message from Fig. 3.3 that observation 114 is indeed odd, by comparison with the other observations. In the absence of information about this observation, it is set to missing. When the analysis is re-run, another observation (row 89) appears "odd", so this is also set to missing.

Fig. 3.4 BoxPlot of %1000_SED WEIGHT



Genstat automatically estimates missing values and the revised results are shown below. We see that the residual mean square is now 65, less than a third of its previous value. We need more explanation about these two observations before simply omitting them, but the precision of the results is dramatically changed by their exclusion.

Fig. 3.5 Output with 2 values set to missing

```
***** Analysis of variance *****

Variate: %1000_SED WEIGHT

Source of variation     d.f.(m.v.)      s.s.       m.s.    v.r.  F pr.

REPLICAT stratum           3          219.73      73.24    0.38

REPLICAT.COMPACTI stratum
COMPACTI                   1         3363.90    3363.90   17.24  0.025
Residual                   3          585.53     195.18    3.00

REPLICAT.COMPACTI.*Units* stratum
ENTRY                     14        18117.36    1294.10   19.89  <.001
COMPACTI.ENTRY            14         1353.25      96.66    1.49  0.135
Residual                  82(2)      5335.88      65.07

Total                    117(2)     28566.37
```

```
***** Tables of means *****

Variate: %1000_SED WEIGHT

Grand mean  188.23

   COMPACTI        1        2
             182.93   193.52

    ENTRY       6.       7.       8.       9.      10.      11.      12.
           163.87   184.37   186.87   194.62   188.62   205.87   170.12

    ENTRY      13.      17.      28.      39.      50.      61.      72.
           189.99   178.62   189.25   203.50   199.50   202.25   171.43

    ENTRY      83.
           194.50

   COMPACTI    ENTRY       6.       7.       8.       9.      10.      11.
          1           160.00   181.25   179.25   187.50   190.50   201.00
          2           167.75   187.50   194.50   201.75   186.75   210.75

   COMPACTI    ENTRY      12.      13.      17.      28.      39.      50.
          1           158.75   181.75   177.50   186.00   194.00   197.00
          2           181.50   198.22   179.75   192.50   213.00   202.00

   COMPACTI    ENTRY      61.      72.      83.
          1           195.50   167.25   186.75
          2           209.00   175.61   202.25
```

```
*** Standard errors of differences of means ***

Table             COMPACTI       ENTRY    COMPACTI
                                             ENTRY
rep.                    60           8           4
s.e.d.               2.551       4.033       6.072
d.f.                     3          82       53.62
Except when comparing means with the same level(s) of
 COMPACTI                                    5.704
 d.f.                                           82

(Not adjusted for missing values)
```

3.4 Good practice

In the results above, Genstat is one of only few statistics packages that gives a good presentation for a split plot experiment. Note that it gives the tables of interaction means as a 2-way table and provides all the standard errors.

In this case there was no evidence of an interaction. If so, then it is easy to save the resulting means and display them as shown in Fig. 3.6 below. Notice that this is not the same as a simple tabulation of the mean values, because of the estimation of the missing values.

Fig. 3.6 Two-way table of means

	Compaction		
Entry Number	**Level 1**	**Level 2**	**Mean**
6	160	168	**164**
7	181	188	**184**
8	179	195	**187**
9	188	202	**195**
10	191	187	**189**
11	201	211	**206**
12	159	182	**170**
13	182	198	**190**
17	178	180	**179**
28	186	193	**189**
39	194	213	**204**
50	197	202	**200**
61	196	209	**202**
72	167	176	**171**
83	187	202	**195**
Mean	**183**	**194**	**188**

In this example we have seen what is meant by "critical analysis" of the data. The boxplot (Fig. 3.4) and the warning on the ANOVA OUTPUT (Fig. 3.3) indicated a problem in the data. It is important to use software that gives users access to the "residuals" after an analysis, and preferably, like Genstat, assists users by indicating potential problems.

4. Designing an experiment

MSTAT has good facilities for randomising experiments that are described in Chapter 4 of the MSTAT-C manual. The first program, called PLAN, permits the randomisation of factorial designs with up to 5 factors. These can be in a randomised block or a split plot design, with up to 4 splits. The results can be presented in a form that is used for data collection and the design displayed as a field plan.

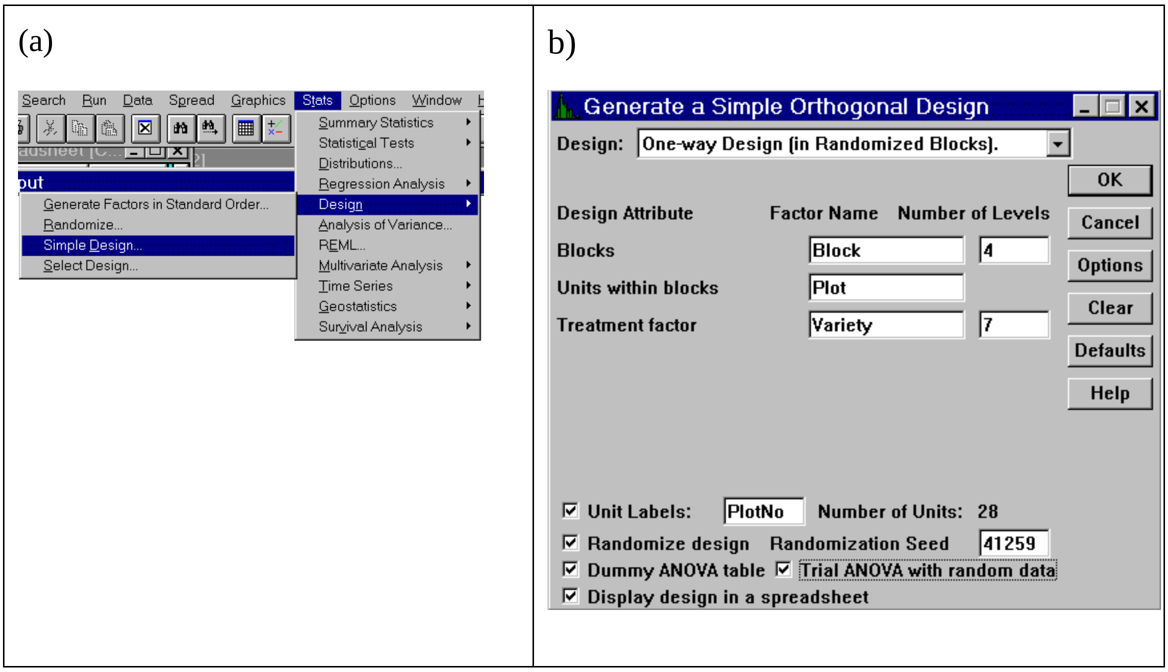
A further possibility that is now available is to use Genstat's Dataload program to export this file to Excel, which can then be used to prepare the data collection forms. This can be used independently of Genstat.

There are however limitations in the designs that can be randomised in MSTAT as we show, by using the equivalent facilities in Genstat. Genstat also has facilities that can be used to compare alternative designs.

4.1 Menu for simple designs

We use Genstat's menu for simple designs. This permits roughly the same range of designs as MSTAT and the main dialogue is shown in Fig. 4.1.

Fig. 4.1 Generating a simple design



(a)

b)

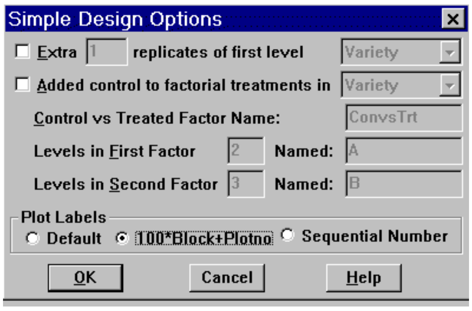
Design: One-way Design (in Randomized Blocks).

Design Attribute	Factor Name	Number of Levels
Blocks	Block	4
Units within blocks	Plot	
Treatment factor	Variety	7

☑ Unit Labels: PlotNo Number of Units: 28
☑ Randomize design Randomization Seed 41259
☑ Dummy ANOVA table ☑ Trial ANOVA with random data
☑ Display design in a spreadsheet

From the Simple Design Options box, Fig. 4.1c, select **100*Blocks+Plots** so that the PlotNo is 101, 102, 103 …407 instead of the Default 1, 2, 3… 28. Running this dialogue gives the randomised design in a Genstat spreadsheet, Fig. 4.1d.

c)



☐ Extra 1 replicates of first level Variety

☐ Added control to factorial treatments in Variety

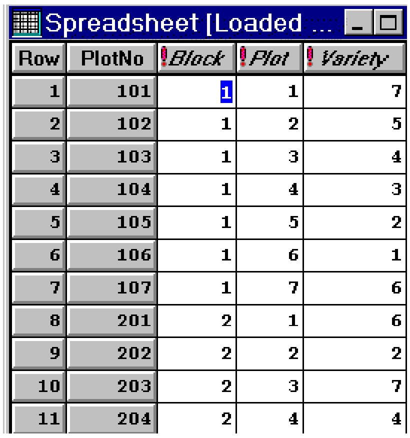
Control vs Treated Factor Name: ConvsTrt

Levels in First Factor 2 Named: A

Levels in Second Factor 3 Named: B

Plot Labels
○ Default ● 100*Block+Plotno ○ Sequential Number

d)



Row	PlotNo	Block	Plot	Variety
1	101	1	1	7
2	102	1	2	5
3	103	1	3	4
4	104	1	4	3
5	105	1	5	2
6	106	1	6	1
7	107	1	7	6
8	201	2	1	6
9	202	2	2	2
10	203	2	3	7
11	204	2	4	4

One possibility is now to save this sheet in Excel, which is then used to prepare the data collection forms. Alternatively the data can be entered directly into the Genstat spreadsheet.

Fig. 4.2 Anova table from generated design

```
     ***** Analysis of variance *****
Source of variation       d.f.

Block stratum                3

Block.Plot stratum
Variety                      6
Residual                    18

Total                       27
```

```
Block     1   2   3   4
  Plot
     1    7   6   7   6
     2    5   2   1   2
     3    4   7   4   5
     4    3   4   2   3
     5    2   5   3   1
     6    1   3   5   4
     7    6   1   6   7
```

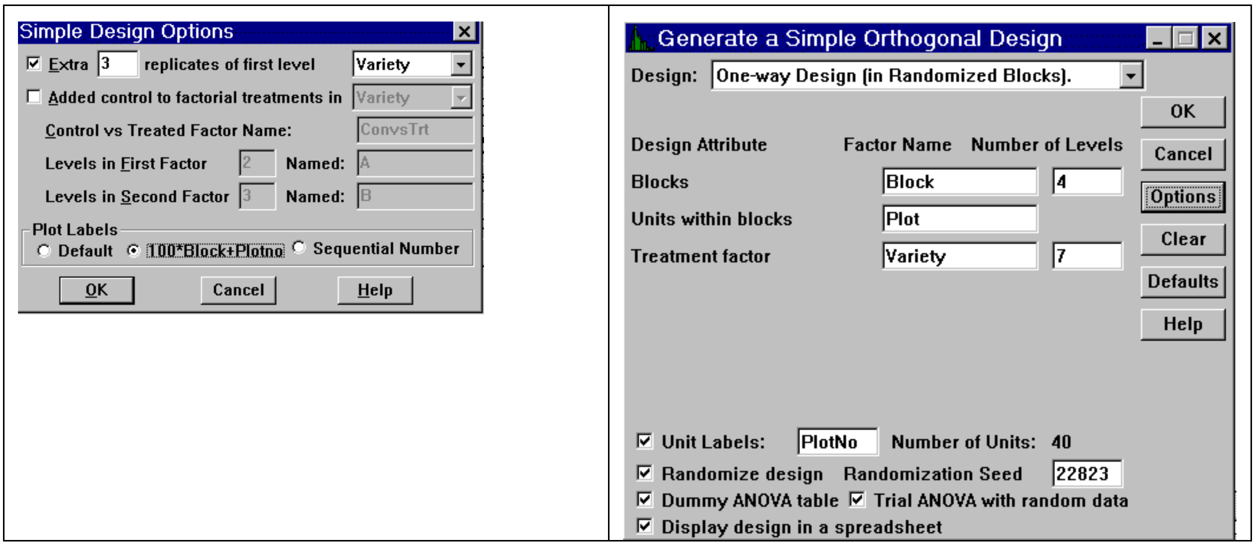
As shown in Fig. 4.2, Genstat also provides a dummy Analysis of Variance. This shows just the terms and the corresponding degrees of freedom. It is very simple here, but is a useful aid in assessing the desirability of more complex designs.

One useful variation on this simple design is to repeat the "Control" variety more often than the others. This is appropriate if the main objective is to compare new varieties with the control.

4.2 Designs with added controls

An example is shown below, where the control is repeated 3 extra times in each block. Together with the 6 test varieties there are therefore now 10 plots in each block. With the 4 blocks this makes a total of 40 plots, as is shown in the dialogue below.

Fig. 4.3 A simple design with added controls



Simple Design Options

☑ Extra 3 replicates of first level Variety

☐ Added control to factorial treatments in Variety

Control vs Treated Factor Name: ConvsTrt

Levels in First Factor 2 Named: A

Levels in Second Factor 3 Named: B

Plot Labels ○ Default ● 100*Block+Plotno ○ Sequential Number

Generate a Simple Orthogonal Design

Design: One-way Design (in Randomized Blocks).

Design Attribute	Factor Name	Number of Levels
Blocks	Block	4
Units within blocks	Plot	
Treatment factor	Variety	7

☑ Unit Labels: PlotNo Number of Units: 40

☑ Randomize design Randomization Seed 22823

☑ Dummy ANOVA table ☑ Trial ANOVA with random data

☑ Display design in a spreadsheet

This is sometimes mistakenly thought to be an "unbalanced" design and therefore difficult to analyse. This is not the case and a similar example is given in the introductory textbook "Statistical Methods in Agriculture and Experimental Biology", by Mead, Curnow and Hasted (1993) page 97. It is, however an example of a simple and useful design that cannot be analysed easily by basic statistics packages, such as MSTAT. To show the form of the analysis we use the option in the dialogue to give a "Trial ANOVA with random data".

Fig. 4.4 Anova output

```
Example Analysis of Variance with Random Data (scaled so RMS=1)

***** Analysis of variance *****

Variate: _Rand_

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Block stratum              3   107.2060    35.7353   35.74

Block.Plot stratum
Variety                    6     6.3382     1.0564    1.06  0.410
Residual                  30    30.0000     1.0000

Total                     39   143.5442
```

```
Variate: _Rand_

Grand mean  14.74

  Variety        1        2        3        4        5        6        7
             15.04    15.08    14.79    14.81    14.33    14.35    13.86
     rep.       16        4        4        4        4        4        4
```

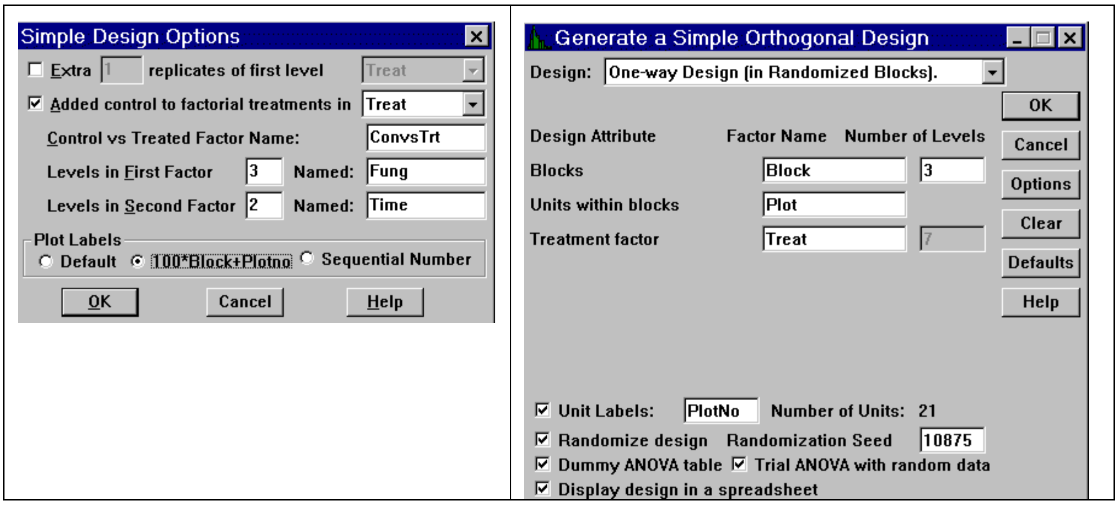
```
*** Least significant differences of means (5% level) ***

Table              Variety
rep.               Unequal
d.f.                    30
l.s.d.               1.444  min.rep
                     1.142  max-min
                     0.722X max.rep

(No comparisons in categories where s.e.d. marked with an X)
```

This shows the way the results will be presented. The table of means shows the unequal replication, with 16 observations for the control and 4 for each of the other varieties.

As shown above, the random numbers used to demonstrate the form of the analysis have been scaled, so the residual mean square is exactly 1. This enables the benefit to be assessed from this change in the design, by comparing the standard errors. As shown above, without the increased replication the lsd is 1.44 and is reduced to 1.14 for comparisons involving the control. The user must now decide whether this increase in precision is worth the extra effort (40 plots, rather than 28) and also whether it is an improvement that will not be modified by the extra block size (10 plots, rather than 7). This is therefore a simple example of the way in which this choice in design can stimulate discussion on the most appropriate experimental structure given a set of objectives.

4.3 Factorial treatment structure plus a control

Fig. 4.5 Factorial treatment structure plus control



Simple Design Options

☐ Extra 1 replicates of first level Treat

☑ Added control to factorial treatments in Treat

Control vs Treated Factor Name: ConvsTrt

Levels in First Factor 3 Named: Fung

Levels in Second Factor 2 Named: Time

Plot Labels: ○ Default ● 100*Block+Plotno ○ Sequential Number

Generate a Simple Orthogonal Design

Design: One-way Design (in Randomized Blocks).

Design Attribute	Factor Name	Number of Levels
Blocks	Block	3
Units within blocks	Plot	
Treatment factor	Treat	7

☑ Unit Labels: PlotNo Number of Units: 21

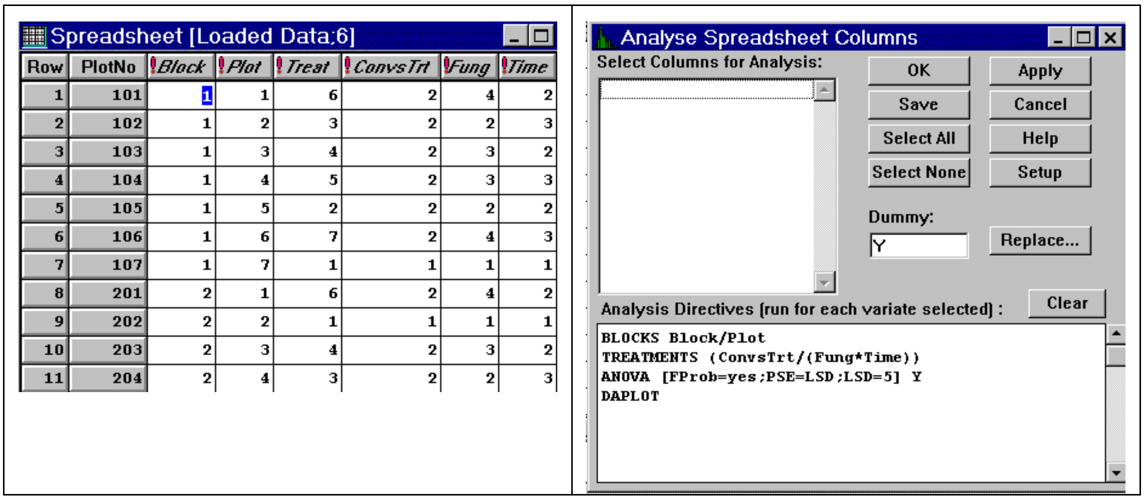
☑ Randomize design Randomization Seed 10875

☑ Dummy ANOVA table ☑ Trial ANOVA with random data

☑ Display design in a spreadsheet

As a final example we consider another common situation that is easily handled in Genstat. This is shown in the dialogues above, where one of the treatment factors is actually a combination of two factors, but there is also an added control. The example above is of 3 fungicides with 2 times of application. There are therefore 3*2=6 treatments. There is also a Control of "No fungicide", so the full structure may be written as 3*2 +1. There are therefore 7 levels to this treatment factor, as shown in the dialogues above.

Fig. 4.6 Factor columns generated, plus commands for an automatic analysis



Spreadsheet [Loaded Data;6]

Row	PlotNo	!Block	!Plot	!Treat	!ConvsTrt	!Fung	!Time
1	101	1	1	6	2	4	2
2	102	1	2	3	2	2	3
3	103	1	3	4	2	3	2
4	104	1	4	5	2	3	3
5	105	1	5	2	2	2	2
6	106	1	6	7	2	4	3
7	107	1	7	1	1	1	1
8	201	2	1	6	2	4	2
9	202	2	2	1	1	1	1
10	203	2	3	4	2	3	2
11	204	2	4	3	2	2	3

Analyse Spreadsheet Columns

Select Columns for Analysis:

Dummy: Y

Analysis Directives (run for each variate selected) :

```
BLOCKS Block/Plot
TREATMENTS (ConvsTrt/(Fung*Time))
ANOVA [FProb=yes;PSE=LSD;LSD=5] Y
DAPLOT
```

Part of the randomisation is shown above. It can be seen that Genstat has included the overall column called Treat and has added 3 further columns. The first is just to compare the control with the other treatments and the remaining factors give the 6 combinations of fungicide and time of application. (Notice above that an extra level has been added to the two factors, Fung and Time, where level 1 represents the control - i.e. "No fungicide" and "No time of application")

One more advanced feature in Genstat's design system is that the spreadsheet also keeps a record of the commands that can be used to give the analysis, once data are available. These commands have effectively been used, through the menu system, to show the form of the analysis below.

Fig. 4.7 Output from analysis

```
Example Analysis of Variance with Random Data (scaled so RMS=1)

***** Analysis of variance *****

Variate: _Rand_

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

Blocks stratum             2     21.362     10.681   10.68

Blocks.Plots stratum
ConvsTrt                   1      0.237      0.237    0.24  0.635
ConvsTrt.Fung              2      1.436      0.718    0.72  0.507
ConvsTrt.Time              1      5.288      5.288    5.29  0.040
ConvsTrt.Fung.Time         2      5.149      2.575    2.57  0.117
Residual                  12     12.000      1.000

Total                     20     45.473

Grand mean  12.37

 ConvsTrt        1        2
             12.11    12.42
     rep.        3       18

 ConvsTrt     Fung        1        2        3        4
        1             12.11
             rep.         3
        2                      12.10    12.79    12.36
             rep.                  6        6        6

 ConvsTrt     Time        1        2        3
        1             12.11
             rep.         3
        2                      11.88    12.96
             rep.                  9        9

 ConvsTrt     Fung     Time        1        2        3
        1        1             12.11
        2        2                      12.04    12.17
                 3                      12.52    13.06
                 4                      11.07    13.65

*** Standard errors of means ***

Table             ConvsTrt    ConvsTrt    ConvsTrt    ConvsTrt
                                  Fung        Time        Fung
                                                          Time
rep.               unequal     unequal     unequal           3
d.f.                    12          12          12          12
e.s.e.               0.577       0.577       0.577       0.577  min.rep
                     0.236       0.408       0.333              max.rep
```

This analysis demonstrates there is no difficulty in analysing such a design. The ANOVA table above shows the split of the 6 degrees of freedom for treatments into the Control v rest, plus the standard subdivision of the factorial components. The treatment means are also presented in a way that is straightforward to interpret.

4.4 Good practice

In terms of ‘good-practice’ on design, we have no problem with the design facilities in the simple packages, such as MSTAT. Our concern is with simple designs that are NOT included, and two examples have been shown here, Fig. 4.3 and Fig. 4.5. These are both designs that were well understood, simple to analyse and used routinely 50 years ago, i.e. before computers were available. They seem rarely to be used now and we wonder whether this is partly because they are not available in the simple statistics packages.

5. Conclusions

Our aim in this document has been to identify good statistical practice. We have also found that most real datasets involve some practical complication and therefore require access to a powerful statistical package for a complete analysis. In the past this has been a dilemma because the powerful packages have needed considerable expertise to be used effectively.

We believe that the situation is now different, partly because of the ease of use of the software under Windows. Users can now use the most appropriate statistics package for the work and can easily change to a different package, when needed, to complete an analysis.

The Statistical Services Centre is attached to the Department of Applied Statistics at The University of Reading, UK, and undertakes training and consultancy work on a non-profit-making basis for clients outside the University.

These statistical guides were originally written as part of a contract with DFID to give guidance to research and support staff working on DFID Natural Resources projects.

The available titles are listed below.

- *Statistical Guidelines for Natural Resources Projects*
- *On-Farm Trials – Some Biometric Guidelines*
- *Data Management Guidelines for Experimental Projects*
- *Guidelines for Planning Effective Surveys*
- *Project Data Archiving – Lessons from a Case Study*
- *Informative Presentation of Tables, Graphs and Statistics*
- *Concepts Underlying the Design of Experiments*
- *One Animal per Farm?*
- *Disciplined Use of Spreadsheets for Data Entry*
- *The Role of a Database Package for Research Projects*
- *Excel for Statistics: Tips and Warnings*
- *The Statistical Background to ANOVA*
- *Moving on from MSTAT (to Genstat)*
- *Some Basic Ideas of Sampling*
- *Modern Methods of Analysis*
- *Confidence & Significance: Key Concepts of Inferential Statistics*
- *Modern Approaches to the Analysis of Experimental Data*
- *Approaches to the Analysis of Survey Data*
- *Mixed Models and Multilevel Data Structures in Agriculture*

The guides are available in both printed and computer-readable form. For copies or for further information about the SSC, please use the contact details given below.

Statistical Services Centre, University of Reading
P.O. Box 240, Reading, RG6 6FN United Kingdom

tel: SSC Administration +44 118 378 8025
fax: +44 118 378 8458
e-mail: statistics@lists.reading.ac.uk
web: http://www.reading.ac.uk/ssc/