GENSTAT NEWSLETTER NO. 1

DECEMBER 1975

This is the first Genstat newsletter containing hints and warnings on the use of the program. Whether it is also the last will depend on the contributions received from users. So, if you have any generally useful devices or even harrowing experiences, please submit them to the Genstat Secretary well before a release is due i.e. one month before June 1st or December 1st.

A list of manual amendments is appended.

'Between and Within Groups' Analysis of Variance

There seem to be two ways in Genstat to obtain particular contrasts from treatment effects. Either one can use REG (8.4.1) to take out single degrees of freedom from a factor or one can set up extra factors and do a nested analysis as illustrated in Users Guide No. 2 (2.4) to deal with control treatments. The latter method can be usefully extended to deal with factors which naturally fall into groups and so require a between and within group analysis. To do this it is necessary to set up one factor to take out the between group sums of squares and one factor for each of the within group sums of squares. A nested treatment structure is then used to obtain the required analysis.

For example, experiments in the plant-breeding section at NVRS involved eight varieties of carrot consisting of two types, Chantenay and Autumn King.

Chantenay			Autumn King		
1.	Red Cored	(RC)	5.	Red Giant Improved	(RG)
2.	Red Cored 36	(RC36)	6.	Vita Larga	(VL)
3.	Long Chantenay	(LC)	7.	Rialto Improved	(RI)
4.	Royal	(R)	8.	Autumn King Original	(AKO)

The form of analysis of variance required was

source	df
Between Types	1
Within Chantenay	3
Within Autumn King	3

i.e. the tabular structure to which the tables of means correspond is

TYPE	CHANT		AUKING		
	x		x		
TYPE. CHANT	RC	RC36	LC	R	AUKING
CHANT	x	x	x	x	
AUKING					x
TYPE. AUKING	RG	VL	RI	AKO	CHANT
CHANT					x
AUKING	x	x	x	x	

This can be obtained as follows, assuming for simplicity that there is no replication and that no other factors are involved and that the data is in the order varieties 1 to 8.

```
'NAME'    NT = CHANT, AUKING
    :     NCH = RC, RC36, LC, R, AUKING
    :     NAK = RG, VL, RI, AKO, CHANT
'FACT'    TYPE ⧸ NT = 4 (1, 2)
    :     CHANT ⧸ NCH = 1...4, 4(5)
    :     AUKING ⧸ NAK = 4 (5), 1, 2, 3, 4
'TREAT'   TYPE / (CHANT + AUKING)
```

In data sets where the varieties are in random order the use of the 'GROUPS' directive facilitates setting up the order of the factors. By specifying which variety belongs to which type the actual randomisation need only be fed in once.

In the previous example if the order of the data had been as follows

Variety	R	LC	RC	RI	AKO	RC36	RG	VL
Type	1	1	1	2	2	1	2	2
Chantenay	4	3	1	5	5	2	5	5
Autumn King	5	5	5	3	4	5	1	2

the randomisations for each factor could be set up thus

```
'NAME'     NT = CHANT, AUKING
    :      NCH = 1, 2, 3, 4, AUKING
    :      NAK = 1, 2, 3, 4, CHANT
'FACT'     VARIETY ⧸ 8 = 4, 3, 1, 7, 8, 2, 5, 6
    :      TYPE ⧸ NT : CHANT ⧸ NCH : AUKING ⧸ NAK
'INTEGER'  IT = 1, 2, 3, - 4, 5, 6, 7, - 8
    :      IC = -1, -2, -3, -4, 5, 6, 7, -8
    :      IA = 5, -6, -7, -8, 1, 2, 3, -4
'GROUPS'   TYPE = GROUP (VARIETY; IT)
    :      CHANT = GROUP (VARIETY; IC)
    :      AUKING = GROUP (VARIETY; IA)
'TREAT'    TYPE/(CHANT + AUKING)
```

This is long-winded for a small amount of data but very useful for larger experiments.

A similar method can also be used to split up interaction terms or residual terms into components, although this may not be practical because the number of factors involved creates space problems.

Kathleen Phelps
NVRS

The Beauties of 'JOIN'

Genstats from this locality have suddenly become much better annotated because of the long-overdue arrival of the directive 'JOIN'. Instead of laboriously typing

```
'HEAD' H(1) = "WEIGHT TONS/AC 1ST ASSESSMENT"
     : H(2) = "WEIGHT TONS/AC 2ND ASSESSMENT"
     : H(3) = "NUMBER 1000's/AC 1ST ASSESSMENT"
       etc.
```

one can now type

```
'HEAD' J(1) = "WEIGHT TONS/AC"
     : J(2) = "NUMBER 1000's/AC"
     : J(3) = "1ST ASSESSMENT"
     : J(4) = "2ND ASSESSMENT"
'JOIN' H(1) = J(1, 3): H(2) = J(1, 4): H(3) = J(2, 3) etc.
```

and obtain any combinations that might be required.

As it is only possible to put one heading on the left hand side it is frequently necessary to form the headings required in a loop. This gives a lot of scope for mistakes and games of the sort where you guess which person's top half fits which bottom half. So if we're joining up our headings we could end up with

"Number of / rotten / flies / (angular transformation)"

"Weight of / red-eyed / cabbage / (1000's/ha)"

It could add a new dimension to biological research.

Kathleen helps
NVRS

P.S.

'JOIN' is the way of producing headings with more than 80 characters per line.

Howard Simpson
RES

Converting Days of the Month to days of the Year.

This is a common requirement as days of the year can be operated on like any other variate. In this example it is assumed that dates of the month have been recorded as six digit integers where the first pair of digits represent the day of the month, the second pair the month and the third pair the year (e.g. 210456 represents 21st April 1956). It is also assumed that they have been punched in fixed format in the first 60 columns of a record i.e. 10 dates of the month per record. In this example the days of the month are assumed to start from 1st January, but the coding can be adapted to start from any date.

```
'REFE'      DAYMTH
'INTE'      I = 1...12
'FACT'      F $ I
'READ/P' DAYMTH, F $ F, (2, 2, 2X) 10. $/
'RUN'
   (Data)
```

'EOD'

'INTE' I = 0, 31, 59, 89, 120, 150, 181, 212, 242, 273, 303, 334

'CALC' DAYEAR = DAYMNTH + VARFAC (F)

Norman Alvey
RES

The Secondary Output Channel in Genstat

This facility was originally created to allow users to produce files containing only relevant parts of the full output for publication by some direct form of duplication, possibly after editing. On machines other than the ICL 4-70 the user can usually specify the nature of this file (i.e. record length, block size, etc.). On the 4-70 these details have to be built into the program and cannot be changed by the user. For the 4-70 therefore the record size was set at 132 characters (plus the carriage control character) - this seemed a reasonable maximum, corresponding to the record length acceptable by the line printer; a smaller size would be unnecessarily restrictive.

For the original purpose this is quite satisfactory: the file can be output to paper tape and edited on paper tape equipment if necessary. It can also be used as input to Genstat, since (on the 4-70) Genstat does not use standard Fortran I/O. However, such files cannot be read by other Fortran programs, and from recent enquiries it appears that there is some demand for such a facility.

The solution is to use DELIC (RES Program Guide PG/117) to convert the Genstat output file into a suitable form. Clearly it is sensible to ensure that no line contains more than 80 characters and the secondary output channel should be selected by

'OUTPUT' 2,80

All Genstat output should then be restricted to 80 characters per line automatically; if any transgressions of this rule are observed please send the evidence to me.

Howard Simpson
RES

Lagging Variates

Lagged variates are often useful (e.g. in time series analysis). They can be formed easily by the use of 'EQUATE'. First set up two scalars 'SCAL' S1, S2

Then to lag a variate by one unit write
'EQUATE' S1, LAG1 = VARIATE, S2

To lag it by n units write
'EQUATE' N/(S1), LAGN = VARIATE, N/(S2)

Both these instructions will place 1 or n missing values in the last locations of the lagged variate.

For a cyclic shift write
'EQUATE' S1 = VARIATE,

Norman Alvey
RES

Partial Aliasing and Confounding

Recent queries have indicated that ANOVA's limitations with partial aliasing and confounding are not as well known as they might be so in this note I shall attempt to explain how they occur and how they can be overcome.

Most of these problems arise from the fact the Anova algorithm cannot distinguish between the individual degrees of freedom of a model term. Thus, in the dummy analysis, if a non zero sum of squares is found for a term, it is assumed that all its effects are estimable.

As an example of partial confounding consider a split plot experiment in which treatment A is applied on the whole plots and treatment B on the sub plots. If A within B effects are to be estimated, the treatment formula should be B/A i.e. B + A.B. (The meaning of a term like A.B depends upon context - it represents all possible effects remaining in an A by B table after all other effects before A.B have been removed. Thus, since here the A main effect has not been taken out, A.B represents A within B rather than the AB interaction. See 8.2p1 or user guide 2 chapter 2.)

Parts of A.B are estimated both in the whole plots stratum (the A main effect), and in the whole plot-sub plot stratum (the AB interaction). The sums of squares in the analysis will be correct but the degrees of freedom will not be partitioned so that the df of the stratum residuals will be incorrect thus, the variance ratios, standard errors etc. will also be wrong. This may seem another trivial (and stupid) example but it illustrates the general principle that if different effects in a term are estimated in different strata, the degrees of freedom for the term in each stratum will be the total degrees of freedom for the term.

To overcome this, pseudo factors should be used. The term is then partitioned into a separate term for each pseudo factor and a term for the remaining effects if any. Each pseudo factor term is examined separately in the analysis to see in which stratum it is estimable and to calculate the appropriate efficiency factor (see 8.2.4). Thus pseudo factors should be used to group together effects estimated in the same stratum with the same efficiency factor (see also for example the partially balanced lattice, Anova example AOV 6).

Thus in the partial confounding example above the treatment formula should be B + A.B //A so that the effects which are estimated in the whole plots stratum are distinguished from those which are estimated in the sub-plots stratum. The Generate directive can be used to set up the pseudo factors if necessary.

Partial aliasing occurs when some of the effects belong to more than one model term and the model terms are both in either the block formula or the treatment formula. Anova can recognise partial aliasing between interaction terms and their corresponding marginal terms like A.B and B above, but in other cases the sum of squares for the aliased effects will be taken out in the first term and not in subsequent terms so that their degrees of freedom will be incorrect. (In fact partial aliasing really shows that the model formula has not been correctly specified, since it is not clear which term should contain the aliased effects.)

For example, Anova example AOV 4, the factors FUMIGANT, with five levels representing four types of fumigant and no fumigant, and LEVEL, with three levels representing no fumigant and two doses, each contain the comparison between no fumigant and any fumigant at any dose.

In AOV 4 aliasing is avoided by including a factor CONTROL in the treatment formula to take out the no fumigant versus any fumigant at any level effect, with fumigant and level nested within the level of control in which any fumigant is applied.

Similar methods can be used in more general cases or else the aliased effects can be indicated by a pseudo factor in second and subsequent terms so that the algorithm is able to detect that these effects have already been removed.

Roger Payne
RES

ADDENDUM TO NEWSLETTER NO. 1

On page 3

The example at the bottom should start;

```
'REFE'    DAYMONTH
'UNIT'    $ 10
'INTE'    I = 1...12
```

On page 4

The fifth line from the bottom should read,

```
'EQUATE'    N! (S1), LAGN = VARIATE, N! (S1)
```

and the bottom line should read

```
'EQUATE'    S1, LAG1 = VARIATE, S1
```