RW Payne

Issue No. 13

1984 March

The GENSTAT Newsletter

EDITORS:

R.W. PAYNE
ROTHAMSTED EXPERIMENTAL STATION
HARPENDEN
HERTFORDSHIRE
AL5 2JQ

M.G. RICHARDSON
NAG CENTRAL OFFICE
MAYFIELD HOUSE
256 BANBURY ROAD
OXFORD
OX2 7DE

Contents

		Page
1.	Editorial	3
2.	Letter to the Editor	3
	Second Genstat Conference	
3.	Interfacing Genstat and a Database Management System	6
4.	Utilisation de Genstat pour le Traitement de Statistique Données Fromageres	8
5.	Some Considerations in Choosing a Package for a Multi-Functional Organisation	13
6.	Teaching Genstat to Non Statisticians	17
7.	Teaching Applied Statistics with Genstat in the University of York	18
8.	Teaching Genstat to Undergraduate Students in applied Mathematics in the University of Genoa, Italy	25
9.	Use of the new Genstat Graph Facilities in the Analysis of Data from Plant Weight – Density Studies	30
10.	Genstat Analysis of Variance and the Distant Client	35
11	Modelling in Genstat	38
12.	Poisson models for the analysis of road traffic accidents	44
13.	Genstat Conference 1985 Order Form	
14.	Genstat Newsletter Order Form	

Published twice yearly by
Rothamsted Experimental Station Statistics Department
and the Numerical Algorithms Group Limited

Printed October 1984

Editorial

The articles in this issue of the Genstat Newsletter comprise the second and final selection of papers from the Second Genstat Conference. The editors would again like to thank the authors for making these papers available to a wider audience through the Newsletter.

There are now well over 250 Genstat sites outside of the ARC: however, few articles are ever received from users at these sites, which omission must introduce a bias in the subject areas covered. We therefore wish to encourage all users to consider submitting material for future issues of the Newsletter.

Our approach to typesetting the Newsletter continues to evolve: in this issue, all Genstat output has been typeset, which, we trust, gives a clearer and more consistent style. (Unfortunately, we have discovered that our typesetting program cannot handle accents in text.) Authors may, in future, wish to submit output, or even entire articles, in machine readable form, to ease the load on our secretarial staff and proofreaders. If so, please first consult the documentation section of NAG regarding acceptable formats. Once again, we should welcome readers comments on the evolving style of the Newsletter.

Letter to the Editor

F. R. House
Department of Pharmacology
Guy's Hospital Medical School
LONDON SE1 9RT
United Kingdom

Dear Sir

The MACRO REPMEAS (Genstat Newsletter No. 11) has proved useful to me already, but many experiments involve more times than subjects in one or more groups, the covariance matrix for any such group cannot be of full rank, causing the MACRO to fail.

The trouble arises from the test of equality of covariance matrices. I have provided a modified version which does not try to be too clever, it just omits the test when any covariance matrix is bound to be singular.

The alterations are marked on the listing by

```
''** MOD1 **''
```

I hope that this proves useful to other readers.

The Macro

```
'MACRO' REPMEAS $
'LOCAL' M,C,SBARJJ,SBAR,DF1,DF2,CHISQ1,CHISQ2,SBAJJ,SBAR,E,DFPOOL,
         UNITY,TWO,THREE,FOUR,SIX,ONE,SBARJ,SSPV,SSV,SO,SSPOOL,DGM,INT,
           V(1...NTIME),VSET,TAB,FULL                     ''** MOD1 **''
'SCALAR' M,C,SBARJJ,SBAR,DF1,DF2,CHISQ1,CHISQ2,SBARJJ,SBAR,E,DFPOOL
 : UNITY=1 : TWO=2 :THREE=3 : FOUR=4 : SIX=6 : FULL=1     ''** MOD1 **''
'VARI' ONE,SBARJ $ NTIME
 :     V(1...NTIME) $ NSUBJ
'SET' VSET=V(1...NTIME)
'TABL' TAB $ TIME,SUBJECT
'PAGE'
'CALC' TAB=RESTAB
'EQUA' VSET=TAB
'INTEGER' INT
'DSSP' SSPV $ VSET
'SYMMAT' SSV,SO,SSPOOL $ NTIME
'DIAGMAT' DGM $ NTIME
'CALC' SSPOOL,DFPOOL,M,C=0
'FOR' I=1..NGRP
'REST' VSET,SUBGROUP $ SUBGROUP=I ; INT
'SSP' SSPV
'EQUAT' SSV=SSPV
'CALC' FULL=FULL*(NVAL(SUBGROUP).GT.NTIME)                ''** MOD1 **''
'CALC' SSV=SSV/(NVAL(SUBGROUP)-UNITY)
'JUMP' .NOT.FULL*SING1                                    ''** MOD1 **''
'CALC' M=M-(NVAL(SUBGROUP)-UNITY)*LOG(DET(SSV))
 :     C=C+UNITY/(NVAL(SUBGROUP)-UNITY)
'LABEL' SING1                                             ''** MOD1 **''
'CALC' SSPOOL=SSPOOL+(NVAL(SUBGROUP)-UNITY)*SSV           ''** MOD1 **''
 :     DFPOOL=DFPOOL+(NVAL(SUBGROUP)-UNITY)
'REPE'
'REST' VSET,SUBGROUP
'CALC' SSV=SSPOOL/DFPOOL
'CAPT' ''
     ** POOLED COVARIANCE MATRIX **''
'PRINT' SSV $ 12.4
'JUMP' .NOT.FULL*SING2                                    ''** MOD1 **''
'CALC' M=M+DFPOOL*LOG(DET(SSV))
 :     C=C-UNITY/DFPOOL
 :     C=C*(TWO*NTIME*NTIME+THREE*NTIME-UNITY)/(SIX*(NTIME+UNITY)*(NGRP-UNITY))
 :     DF1=NTIME*(NTIME+UNITY)*(NGRP-UNITY)/TWO
 :     CHISQ1=(UNITY-C)*M
```

```
'LABEL' SING2                                            ''** MOD1 **''
'CALC' M=(SUM(SSV)-TRACE(SSV))/(NTIME*(NTIME-UNITY)/TWO)  ''** MOD1 **''
 :     C=TRACE(SSV)/NTIME-M
 :     DGM=C
 :     SO=M
 :     SO=SO+DGM
'CAPT' ''
     ** COMPOUND SYMMETRIC COVARIANCE MATRIX **''
'PRINT' SO $ 12.4
'CALC' M=-(NVAL(SUBGROUP)-NGRP)*LOG(DET(SSV)/DET(SO))
 :     C=NTIME*(NTIME+UNITY)*(NTIME+UNITY)*(TWO*NTIME-THREE)
 :     C=C/(SIX*(NVAL(SUBGROUP)-NGRP)*(NTIME-UNITY)*(NTIME*NTIME+NTIME-FOUR))
 :     DF2=(NTIME*NTIME+NTIME-FOUR)/TWO
 :     CHISQ2=(UNITY-C)*M
'JUMP' .NOT.FULL*SING3                                   ''** MOD1 **''
'CAPT' ''
     ** BOX'S TEST OF EQUALITY OF COVARIANCE MATRICES **''
'PRINT' DF1,CHISQ1 $ 5.0,10.4
'LABEL' SING3                                            ''** MOD1 **''
'CAPT' ''
       ** BOX'S TEST OF SYMMETRY OF COVARIANCE MATRICES **''
'PRINT' DF2,CHISQ2 $ 5.0,10.4
'CALC' ONE=UNITY
 :     SBARJJ=TRACE(SSV)/NTIME
 :     SBARJ=PDT(SSV;ONE)/NTIME
 :     SBARJ=MEAN(SBARJ)
 :     E=NTIME*NTIME*(SBARJJ-SBAR)*(SBARJJ-SBAR)
 :     E=E/((NTIME-UNITY)*(SUM(PDT(SSV*SSV;ONE))-TWO*NTIME*SUM(SBARJ*SBARJ)
           +NTIME*NTIME*SBAR*SBAR))
'CAPT' ''
       ** GREENHOUSE-GEISSER EPSILON **''
'PRINT' E $ 10.4
'DEVALUE' M,C,SBARJJ,SBAR,DF1,DF2,CHISQ1,CHISQ2,SBARJJ,SBAR,E,DFPOOL
           UNITY,TWO,THREE,FOUR,SIX,ONE,SBARJ,SSPV,SSV,SO,SSPOOL,DGM,INT,
           V(1...NTIME),VSET,TAB
'ENDMACRO'
```

Second Genstat Conference (Contd)

Interfacing Genstat and a Database Management System

A. Bouvier
I.N.R.A.
Centre National de Recherches Zootechniques
Laboratoire de Biometrie
78350 Jouy-en-Josas
France

Why an interface between Genstat and a Database Management System?

Genstat is a language for statistical analysis and data manipulation but is not very convenient for treating a small amount of data or searching for samples satisfying a logical criterion (filter).

To avoid writing specific formatting programs each time, for each file which we wish to process by Genstat, we modified an existing directive which is not used at INRA, the SYMAP directive, to extract data from a database.

This work was performed on the INRA computer (a CII-HB IRIS 80) at Jouy-en-Josas on an ecopathological database, managed by the database management system "Socrate".

The building of the SYMAP directive

We have re-written the Genstat subroutine SYMAP, which implements the directive of the same name, by using the Fortran–Socrate interface. Socrate has its own inquiry language but also contains some specific subroutines callable from a programming language.

Examples of Socrate subroutines callable from a host-language:

CALL SOPEN (*base-name*, 'I')

opens a database with 'read only' access

CALL SGBD (*base-name, YZO, lg–YZO, sp-name*)

invokes a subroutine named *sp-name* previously written in the Socrate inquiry language. *YZO* is the working memory area address and *lg–YZO* its length.

The SYMAP Fortran subroutine executes these calls according to the arguments the user gives to the SYMAP directive; the number and type of these arguments depend on the Socrate subroutine called and some options have been introduced, mainly to reduce the amount of storage used by the extracted data.

The interfaced database

The interfaced database is an ecopathological database: it contains the results of a continuous survey taken since 1977 on 135 bovine herds in several French districts: the aim of this survey is to study the influence of the environment on the health of the animals.

The data are numerous (about 5 M bytes in 1983) and very diverse:

- occurrences of all diseases and symptoms (300 different codes)
- animal diet and herd management
- blood analysis results
- climatic conditions

Applying the interface system to the ecopathological database

This required the use of several Socrate interface subroutines (about fifteen at present). This number is not limited and we can add more at any time.

Example of a Genstat program which extracts health data

```
'REFE'
```

```
'VARI'  DPT = 29,52
```

29 is the code of the Finistere district and 52 the code of the Haute–Marne

```
    :   EXPL(1)=99
```

By convention, 99 means: ‘all that exists in the database’. Here this avoids naming all the herds of the Finistere district

```
    :   EXPL(2)=-1,10
```

In the Haute–Marne district, the user wants all the herds whose number is between 1 and 10 (this is also a notational convention)

```
    :   CODE (1) = 999
    :   CODE (2) = 420,421
```

The variates CODE are the codes of the diseases of interest: 999 represents the total number of sick animals and 420-421 are the codes of the two sorts of metritis

```
    :   NOCIE = 1
```

Animal category 1 is the ‘milch cow’

```
'HEAD' P=”G-TE”
```

Name of the desired Socrate interface subroutine

```
'SYMAP/MODE=A' P,1,7901,7924,
            DPT,EXPL(1,2),NOCIE,
            CODE(1,2)
            ID,DATE,V(1,2)
```

1 corresponds to the data type ‘Sanitary’

7901 and 7924 are the beginning and ending dates of the period required (year and fortnight)

The MODE option has the same significance as in the READ directive: cumulate all the values concerning the herds of a district

```
'PRIN/P'   ID,DATE,V(1,2) &6
```

the output results must be treated in parallel:

```
'RUN'
 ID      DATE     V(1)    V(2)
2932     7901      61      15
2932     7902      66      17
  .
  .
  .
2932     7924      60      14
5215     7901      40       7
  .
  .
  .
5215     7924      35       5
```

ID is an identifier formed from the code of the district (29) and the number of cumulated herds

DATE is the date of the information

V(1) and V(2) are respectively the total number of sick animals and the number of occurrences of metritis.

Conclusion

At present the SYMAP directive can only be used with one database; moreover the imminent replacement of the IRIS 80 at Jouy-en-Josas by a DPS8 implies the termination of support for Socrate and consequently of this interface. However, similar work is planned using MRDS, the relational Database Management System of the DPS8.

Utilisation de Genstat pour les Traitment de Statistique de Données Fromageres

Eric Derobert
INA PG Laboratoire de Biometrie
16 Rue Claude Bernard
75231 Paris
France

Nous avons pour aider à résoudre des problèmes de fabrication de fromage, coordonné et quelquefois adapté, un certain nombre de méthodes d'analyses de données.

Pour présenter l'apport de Genstat dans les analyses statistiques que nous avons mises en oeuvre et aussi évoquer certains problemes liés à son utilisation, nous commencerons par décrire brièvement le problème posé et les données dont nous disposons. Nous évoquerons ensuite différentes applications de Genstat, puis dans une dernière partie, nous nous focaliserons sur le programme CALIN 2, programme d'analyse discriminante à 2 populations.

Presentation du probleme

Les données fournies par les fromageries, résultent de mesures effectuées durant la fabrication des fromages à partir d'entités repérées. C'est ce qu'on appelle le 'suivi de bassine'. Ces données portent essentiellement sur les caractéristiques des matières premieres utilisées (lait, ferments), sur les dynamiques fondamentales de la fabrication (égouttage, acidification) et sur les leviers de conduite susceptibles de les corriger ou d'en orienter le cours (température ambiante, durée d'égouttage...). Ce premier groupe de données constitue l'ensemble des variables de fabrication. On a disposé souvent de l'ordre d'une centaine de ces variables, quelquefois un peu plus, exceptionnellement beaucoup moins.

Des fromages repérés lors de la fabrication, après emballage, sont conservés dans des conditions simulant le circuit de distribution normal.

Au bout d'un certain nombre de jours, au moment supposé idéal de consommation, les fromages sont dégustés. Un jury d'usine détermine quels fromages sont optimums et pour ceux qui ne le sont pas, de quels défauts ils sont affectés. Sont définies de la sorte une variable optimum et plusieurs variables défauts (entre 10 et 20 réparties en 4 catégories: défaut de coupe, de texture, de goût, d'aspect). Ce deuxième groupe de variables, toutes dichotomiques, constitue l'ensemble des variables de dégustation.

Notre travail a consisté à chercher les méthodes les plus efficientes pour mettre en relation variables de dégustation et variables de fabrication. L'objectif principal est de chercher les variables clefs de la fabrication, les normes qui conduisent à l'optimum et les paramètres qui augmentent les risques d'apparition de tel ou tel défaut.

Nous nous bornerons bien sûr à présenter les analyses pour lesquelles nous avons eu recours à Genstat.

Note: Les analyses statistiques ont été effectuées sur des échantillons d'environ 150 fromages (±50).

Quelques utilisations de Genstat

Analyse Préalable des Variables de Dégustation

Nous verrons dans la dernière partie que nous confrontons par des analyses discriminantes à 2 populations, la population des fromages optimum à différentes populations défaut.

Chaque population défaut comprend l'ensemble des produits possédant le défaut en question. Le problème est de choisir convenablement les défauts de façon à ne pas reproduire à chaque fois la même analyse. En effet, le fait que la population optimum, à laquelle les populations défauts sont confrontées, ne varie pas constitue une source de biais naturel qu'il va s'agir de contourner.

Deux critères se sont imposés pour le choix des défauts à étudier: la quantité et la singularité. Il est souhaitable que les défauts choisis soient suffisamment représentés pour assurer une certaine fiabilité aux analyses et suffisamment différents les uns des autres pour assurer la plus grande originalité possible à chacune des analyses.

Nous procédons donc systématiquement à une approche de l'espace des défauts à la fois quantitative (simples comptages) et analytique (application de méthodes standards telle que l'analyse en composantes principales ou l'analyse des corrélations canoniques).

Avec Genstat les comptages sont très simples à réaliser par le calcul matriciel. On obtient facilement l'ensemble des effectifs des différentes populations défaut, et aussi les effectifs par groupes de 2 défauts. L'ACP renseigne utilement sur les ressemblances entre défauts ce qui évite d'étudier des populations défaut trop proches les unes des autres.

Signalons que, dans les cas épineux, une aide à la décision est obtenue en réalisant une régression linéaire pas à pas de la variable optimum sur les variables défauts.

En supposant –cas le plus courant –qu'un fromage optimum n'a pas de défaut, le premier pas sélectionnera le défaut le plus représenté

Soient N: le nombre de fromages
D: le nombre de fromages possedant le premier défaut sélectionnné
E_x: le nombre de fromages possedant un défaut X
F_x: le nombre de fromages possédant à la fois le premier défaut et le défaut X

Le deuxième pas sélectionnera le défaut de caractéristiques E_x et F_x telles que

$$H(X) = \frac{E_x\,(D+N) - 2N\,F_x}{N - E_x} \quad \text{maximum}$$

(la variance résiduelle de la régression décroit linéairement en fonction de $H\,(X)$)

On constate que H_x est d'autant plus grand que E_x grand (critère de quantité) et F_x petit (critère de singularité) ce qui correspond bien au résultat recherché.

Pour les pas suivants, les calculs se compliquent mais la logique reste la même.

Régression des Variables de Fabrication sur les Variables de Dégustation

En efectuant des régressions linéaires multiples de chacune des variables de fabrication sur l'ensemble des variables de dégustation (sauf la variable optimum), nous construisons implicitement un modèle où nous supposons que chaque variable de fabrication a une valeur optimale et que des écarts (en baisse ou en hausse) à cette valeur sont générateurs de défauts:

Ce modèle s'écrit ainsi:

$$X_i = \mu + \sum_{j \epsilon C_i} \alpha_j + \varepsilon_i$$

$i \;=\; 1 - n$ individus

$C_i \;=$ ensemble des défauts de l'individu i à chaque defaut correspond un effet α_d

Dans un tableau, nous faisons figurer pour chaque variable, l'estimation $\hat{\mu}$ de μ et les défauts d pour lesquels nous jugerons ($\hat{\alpha}_d$ significativement non nul).

La limite de cette méthode est que nous prenons en compte en vrac l'ensemble des variables de dégustation alors que certaines peuvent ne jouer aucun rôle. D'autre part, la corrélation de certains défauts entre eux peut légérement fausser le modèle.

De ce fait, cette méthode n'est pas essentielle dans nos analyses. Néanmoins elle est appréciée des fromagers qui trouvent dans les tableaux un point de référence et des indications utiles et faciles à lire.

Genstat est particulièrement bien adapté pour cette analyse (contrairement par exemple à BMDP). En effet, le calcul de la matrice de régression unique et des coefficients est obtenu simplement par:

```
'SET' FABRICATION = .... 'SET' DEGUSTATION ....
'REGR' FABRICATION, DEGUSTATION 'Y' FABRICATION 'FIT'  DEGUSTATION
```

Modelisation et Simulation

Des problèmes de modélisations partielles et des modélisations globales suivies de simulations ont été traités avec Genstat.

Il serait trop long de donner des exemples détaillés car cela nécessiterait une présentation complète du contexte quelquefois compliqué dans lequel ces analyses ont été menées.

De courbes d'acidification ou de température ont pu être approchées par des méthodes logistiques ou autres à l'aide e l'instruction `'MODEL'` et `'OPTIMISE'`. A notre sens toutefois, la présentation des résultats obtenus gagnerait à être plus claire et plus aérée. En cas de restrictions préalables sur les variables étudiées, le programme n'indique pas le nombre correct de degrés de liberté. Il ne prévoit pas non plus qu'en cas de restrictions successives, la somme des carrés résiduels indiquée est la somme cumulée calculée sur l'ensemble des différentes sous-populations obtenues par restriction.*

*Ces remarques s'appliquent à la version 4.01. Peut être des améliorations ont elles été apportées depuis.

Des simulations simultanées de variables corrélées sont assez simples à effectuer avec Genstat. Des vecteurs multinormaux de loi *N(0,V)* sont obtenus par la multiplication des vecteurs de loi *N(0,I)* par une matrice *A* telle que $AA' = V$. Genstat fournit une solution pour *A:* la matrice triangulaire determinée par la decomposition de Cholesky.

Le programme CALIN 2

(Programme d'analyse discriminante pas à pas à 2 populations avec recherche d'effets linéaires et quadratiques).

Une fois que sont définies les populations défauts à étudier, nous les confrontons une à une à la population optimum. Etant donné le grand nombre de variables en présence, il s'impose de recourir à la méthode du pas à pas.

Il est vite apparu que les écarts types de certaines variables étaient très nettement différents sur la population optimum et sur les populations défauts d'où l'intérêt de procéder à des discriminations quadratiques. Cependent les programmes classiques de discrimination quadratique ne nous ont pas semblé convenir à notre problème, essentiallement pour deux raisons:

- on applique systématiquement à toutes variables un effet quadratique alors que lorsque les écarts-types des variables sur les deux populations sont quasiment éqaux, cet effet quadratique n'a pas de sens.
- sont pris en compte dans l'analyse, outre les carrés des variables (X^2,Y^2) les produits croisés des variables (*XY*). Cette disposition, si elle enrichit théoriquement l'analyse, dans la pratique ne contribue guère à la clarification des résultats surtout quand le nombre des variables est grand. De plus, là encore ces effets croisés apparaissent quelle que soit leur validité réelle.

Nous voulions, pour notre part, répondre aux deux exigences suivantes:

- la prise en compte sélective des effets quandratiques: il est redondant et lourd de mettre en évidence automatiquement des effets quadratiques là où il y en a pas.
- la clarté des résultats et de leur interprétation: il importait, à notre sens, que ces analyses de bases soient perceptibles aisément à leurs utilisateurs ultérieurs. Nous acceptons pour cette analyse une éventuelle (et très relative) réduction de l'information dans la mesure où elle peut en augmenter la lisibilité.

En fonction de ces critères, la méthode que nous utilisons procède de la facon suivante: à chaque pas, nous testons chaque variable en effet linéaire (sélection de *X* test de Ficher à *1* degré de liberté) et en effet quadratique (sélection de *X* et X^2: test de Ficher à *2* degrés de liberté). Nous comparons le milleur Ficher à *1* degré de liberté au meilleur Ficher à *2* degrés de liberté par confrontation de seuils de probabilité correspondants et sélectionnons finalement soit une variable à effet linéaire (indication de '+' ou de '–' suivant que la variable doit prendre des valeurs élevées ou faibles pour augmenter les chances d'optimum) soit une variable en effet quadratique (à partir des coefficients estimés pour *X* et X^2, on calcul une *valeur optimale* qui maximise les chances du produit d'appartenir à la population optimum*). Pour exécuter cette procedure, nous avons écrit en langage Genstat le programme `CALIN 2`.

Comme il s'agit d'une analyse discriminante à 2 populations, nous avons procédé en utilisant les directives de régression.

La première idée qui vient à l'esprit a été de calculer la matrice de régression globale des *n* variables plus les *n* variables au carré. Nous aurions alors effectué une boucle pour la recherche des variables discriminantes. Cette tentative s'est vite heurtée à plusieurs défauts:

* Dans certain cas on trouve une valeur optimale pour appartenir à la population défaut.

- l'utilisation de la directive 'BEST' semble a priori la plus économique pour repérer la meilleure variable linéairement discriminante. Mais à chaque pas on fait intervenir aussi les tentatives de discrimination quadratique. Or une directive 'BEST' n'est pas réversible, c'est à dire qu'on ne peut pas revenir à l'état de la régression précédant le 'BEST' (alors qu'un 'DROP' peut compenser un 'ADD'). Dans le cas où la meilleure discrimination est quadratique, on ne pourra donc annuler l'effet de la directive 'BEST'. Il faut donc utiliser une structure intermédiaire du type 'SET' VARSEL = ensemble des variables déjà sélectionnées pour retravailler à chaque pas en partant d'un 'FIT' VARSEL qu'il s'agit d'améliorer. Mais alors qu'il est possible de conserver – en utilisant la directive 'ASSIGN' – des variables sélectionnées par l'utilisation des 'FIT' ('FOR'V=ENS ; A=RES ; 'FIT' VARSEL,V ; DEV=A .... 'ASSI'Q=V après tests sur A), ce n'est pas le cas avec 'BEST' où l'on perd toute trace de la variable retenue si d'autres calculs de régression indépendants doivent s'intercaler. On pourrait imaginer pour pallier ce problème de disposer d'une option de type:

  ```
  'BEST' ENSVAR ; VRS = Q
  [VRS : variable sélectionnée
     Q : de type  'POINTER' $1 récupérable ensuite par 'SET/POIN = S']
  ```

  Dans l'état actuel des choses, pour notre programme, la directive 'BEST' qui théoriquement permettrait d'exécuter sensiblement les calculs est pratiquement inutilisable.

  Note: la proposition d'option pour 'BEST' aurait aussi des applications pratiques dans des cas beaucoup plus courants (construction de graphes de régressions simples, régressions en chaînes, etc.)
- L'encombrement mémoire provoqué par l'utilisation d'une matrice unique de régression était tel (pour *100* variables: $(100+100)^2 = 40000$) qu'il provoquait sur l'UNIVAC d'Orsay le dépassement d'un seuil tel que le temps de calcul payé etait tout à coup multiplié par 2. L'espoir d'un gain de temps lié à l'utilisation de 'BEST' étant parti en fumée, nous avons dû renoncer à utiliser une matrice de régression unique et avoir recours au calcul à chaque pas et pour chaque variable de la matrice de régression. Cette méthode peu idéale en théorie puisqu'elle oblige à refaire plusieurs fois certains calculs est pourtant la moins coûteuse. Une tentative pour garder en mémoire un corps commun de la matrice et ne calculer à chaque pas que les nouvelles ligbes de cette matrice s'avère plus chère (et de loin à cause des nombreux appels de directives auxquels elle oblige).
- reste enfin le problème de la structure conservant les variables sélectionnées. Nous avons déja noté que nous utilisons la directive 'SET' VARSEL..... Mais le problème du 'SET', c'est qu'il est exécuté à la compilation. Dans une boucle, la valeur du 'SET' sera toujours la première valeur prise. Ce problème a été résolu en effectuant l'ajout des nouvelles variables sélectionnées dans le 'SET' à l'intérieur d'une macro (la boucle est elle-même comprise dans une macro) de ce type:

  ```
  'MACRO' AJOUTLIN
  'SET' VARSEL = VARSEL, Q
  'ENDMACRO'
  ```

  Cette procédure est peut-être un peu lourde et on pourrait tout à fait imaginer une directive équivalente à 'SET' exécutable au run-time.

  (La directive 'ASSIGN' possède cette propriété, mais ne peut contenir qu'une variable).

Conclusion

L'apprentissage de Genstat et l'approfondissement des possibilités qu'il offre constituent des investissements qui rapportent. Tout ce qui est défrichage des données, appréhension des fichiersgruyère, et surtout mise en oeurvre de méthodes un tant soit peu hors des normes et chainage d'analyses classiques est excellemment résolu par Genstat.

En revanche, il faut admettre que pour la plupart des analyses standard, d'autres logiciels sont sourvent plus performants aussi bien en temps de calcul qu'en encombrement mémoire. Et si Genstat a été l'instrument central du travail dont nous rendons compte ici partiellement, d'autres logiciels (notamment BMDP) nous ont aussi beaucoup apporté.

Finalement, il ne s'agit pas tant de mettre à tout prix en concurrence. Les différents logiciels que de tacher d'en user complémentairement suivant les mérites et faiblesses de chacun.

Some Considerations in Choosing a Package for a Multi–Functional Organisation

J Fenlon
Glasshouse Crops Research Institute
Worthing Road
Rustington
Littlehampton
West Sussex

Introduction

Faced with a need to provide some form of statistical software, how should one proceed? Having specified precisely what the requirement is, one might consider first whether it could be met by an in-house provision (e.g. use of NAG Library or other tested algorithms, linked to form a suite). The simple salary cost of one's own time compared with the cost of software from other sources frequently militates against this. The second step is probably to look at the various statistical packages which are available, and choose the one which most nearly meets one's specification.

Comparisons

During the last ten years a small industry has grown up, particularly in the United States, in comparing statistical packages. Whilst a very useful service is provided by these various writers, some of the literature is partisan in that the author is attempting to present his own package in a more favourable light than that of his competitor. Similarly the taxonomies of Francis (1979, 1981), though admirable in their conception and execution, need to be considered carefully in view of the small user sample and the fact that the users were chosen by the developer. Another factor of considerable importance is whether particular comparisons are currently relevant given the process of continually updating most packages.

The literature can be classed broadly as follows:

(a) Desiderata (with critical examples)
(b) Descriptions of Individual Packages
(c) Comparisons of several packages

  (i) consumer reports
  (ii) facilities
  (iii) meeting specification
  (iv) algorithms/accuracy
  (v) educational/documentation.

Table 1 presents a list of some of the papers which compare various packages on the basis of their performance on particular statistical tasks. The majority of these papers are concerned with various aspects of the linear model. Where Genstat is compared it generally fares well, as also does SAS, even though quite old versions are used. Hamer, commenting on Heiberger's 1981 paper in a later edition of the same journal made the point that all the comparisons were made on obsolete versions of the packages – in fact, no version post-dated 1976! And Heiberger's paper is one of the latest cited!

		BMDP	SPSS	GENSTAT	SAS
ANOVA	– Francis (JASA,1973)	X	X		X
	– Heiberger (PSCS, 1976;Am.Stat.,1981)	X	X	X	X
	Speed, Hocking & Hackney (JASA,1978)	X	X	O	X
ANCOVA	– Federer & Henderson (PSCS,1978;1979)	X	X	X	X
	Searle & Hudson (Biom.,1982)	X	X	X	X
REGRESSION	– Heiberger (Int.,1975)				
	Velleman & Francis (Int.,1975)				
MANOVA	– Hohwald & Heiberger (PSCS,1977)	X			X
CLUSTER ANALYSIS	– Blashfield & Aldenerfer (source unknown)	X			
DISCRIMINANT ANALYSIS	– Frazier (PSCS,1979)	X	X		X
	Hohwald & Heiberger (Int., 1978)	X	X	X	X
TIME SERIES	– Davies (PSCS,1977)	X		O	X

Some comparisons of specific procedures

Table 1

Key:		
	PSCS	– Proceedings of the Statistical Computing Section, A.S.A.
	Int.	– Annual Symposia on the Interface
	O	– The comparison is not made explicitly but can be inferred

All this is very important, but from a practical point of view sophisticated linear model analysis may not be of the highest priority. Where observational rather than experimental data are being obtained different approaches to statistical analysis are often needed. Three interesting papers (Nelder, ISR, 1974; Cooper, JRSS(A), 1977; Greenfield and Siday, The Statistician, 1980) look at statistical packages more from the design aspect and consider those features which they regard as essential pre-requisites for any system. All three papers are worthy of attention, but I shall consider that of Cooper for reasons which will become apparent later. Cooper discussed a system in terms of three major facilities: (a) general, (b) data management and (c) data analysis. Most of what Cooper described was not new to Genstat users, though three topics were novel: his proposals for dealing with missing values, the incorporation of hierarchical data sets into the package and the suggestion that the Genstat ANOVA and GLIM algorithms might be incorporated into the system. Here at last was someone who did not want to re-invent the wheel!

The one aspect which I have not discussed so far is ease of use; in some respects this is subjective, though such considerations as documentation, statistical training, help facilities etc. can be scored. Francis' work here is quite useful and the following table (Table 2) is condensed from his COMPSTAT 1980 paper:

	Training		Language		Convenience
	Statistics	Programming	Simplicity	Documentation	of Use
Genstat 4.02	1.0	3.0	2.0	2.0	2.5
SPSS 8.0	0.3	1.7	1.7	2.7	2.3
SAS 79.1	3.0	2.7	2.3	3.0	3.0
BMDP 77	0.7	2.7	2.3	2.7	2.0

[The higher the score, the easier to use]

User Ratings for ease of learning and using

Table 2

Some of these scores are quite surprising; particularly noticeable is the opinion of SPSS users that considerable training in statistics was necessary to use the package – contrast this with the response for SAS, a much more sophisticated system! These comparisons are not entirely fair, and Francis makes the point, but look how uniformly well SAS compares with Genstat! Bernard (COMPSTAT, 1980) looked at some of these topics as a 'sophisticated' user and considered Genstat to be more difficult than SPSS and BMDP in the areas of prior knowledge, documentation and ease of use. He emphasised the point that the quality of documentation and marketing were very important, and stressed that Genstat, in particular, needed selling as it was 'more difficult by its very design'.

The Organisation

So much for the theoretical considerations; what happened in practice? The Severn–Trent Water Authority is a large multi-functional organisation of some 8,000 employees serving the needs of approximately 8 million people. Its essential duties consist of supplying clean, potable water to its consumers and re-cycling waste water, together with other statutory obligations such as controlling pollution, provision and maintenance of fisheries and other amenity areas. The quantity of data collected on a routine quality control basis alone is staggering and various archiving systems have been provided for handling such data. At present, a data base system is being installed to integrate the various discrete data sets, although some of the scientific archives can almost stand alone. Many people are employed in data provision roles by management, though very few have formal statistical training. The availability of statistical software was very limited in 1978, though the water quality archive had routine summary statistics as an optional output (provided via COBOL!). Much of the statistical requirement was of a relatively simple nature: summary and simple descriptive statistics, distribution testing, extreme percentile estimation and tabulation of survey data, one-way MANOVA and simple modelling and simulation. Although there were some designed experiments, much of the data was of an observational nature.

The Outcome

At this time my own experience was mainly with Genstat and GLIM though I also had some experience of SPSS. Canvassing the views of others who might be interested in using such a package, I found that among those who had heard of such packages, SPSS was invariably mentioned. At the outset it was necessary to provide an interface between the water quality archive and any statistical package which was chosen. It was realised at this time that a considerable amount of routine statistical analysis could be effected by incorporating NAG routines and Applied Statistics algorithms into a simple program suite. At about this time we were approached by ICL and asked if we would field test their new 2900 Statistics system: this was essentially the system described by Cooper in his 1977 paper which was now (in 1979) becoming available. Given the circumstances, we agreed, and I, personally, found the system quite rewarding.

Simply, the package has its own command structure which consists of an English-type language. It operates essentially on a data matrix (or data set) though there are facilities for ancillary data structures. Its main features are:

(i) basic descriptive statistics, tabulation and plotting and

(ii) special analyses, comprising regression, ANOVA and various multi-variate techniques.

A particularly interesting feature of the package, at the time, was its interactive (on-line) mode, which enabled teaching sessions to be arranged at the terminal with both terminal and hard-copy outputs.

Two or three such sessions soon had several people using the system. Apart from by myself and another statistician, it was used very occasionally but there was not the same resistance to using it after a long period that I had anticipated. Despite claims that 'English-type' structures can be a disadvantage, my opinion was that the apparent 'friendliness' of the instruction set made up for the occasional inconvenience caused by using the wrong preposition. Advising or assisting users on the telephone was usually a straightforward affair and many problems were solved while users were actually sitting at a terminal with telephone in hand. There were limitations to the system, some of which could not be circumvented, and not a few frustrations, but there were a lot of admirable features, not least its basic simplicity, sensible defaulting and the facility to test and debug a programme on line. It was possible to show a complete, simple program to a beginner and there would be a good chance that he could follow the sense of it and basically understand what it was doing. Earlier and later experiences attempting to do the same thing with Genstat met with considerable resistance. The system was user-friendly and it was possible to provide potential users with a small hand-out (4 sides) of basic instructions which would enable him to solve most of his problems.

Despite Brian Cooper's declared hope, Genstat ANOVA and GLIM were not incorporated into the 2900 package, and one of my early moves was to produce a version of GLIM, which proved invaluable in many ways, not least in its simple data exploration facilities. Finally, Genstat itself was acquired, primarily to mop up the outstanding linear model problems of designed experiments which 2900 STAT made little attempt to cater for.

References

Cooper, B.E. (1977) Advances in statistical system design, JRSS (A), **140**, pp. 166–97.

Frances, I. (1979) A Comparative Review of Statistical Software. Voorburg, The Netherlands: International Association for Statistical Computing.

Frances, I. (1981) Statistical Software: A Comparative Review. New York: Elsevier North-Holland, Inc.

Greenfield, A.A. and Siday, S (1980) Statistical Computing for Business and Industry. The Statistician, **29**, pp. 33–35.

Nelder, J.A. (1974) A User's Guide to the Evaluation of Statistical Pakcages and Systems. Int. Stat. Rev., **42**, pp. 291–8.

Teaching Genstat to Non Statisticians

Elisabeth Lesquoy
Universitė de Paris Sud
Equipe de Statistique ERA-CRNS 532
91405 Orsay Cedex
France
and
INRA-CNRZ
Laboratoire de Biomėtrie
71850 Jouy-en-Josas
France

We will first summarise our practice in teaching Genstat to biologists, or using Genstat with them. We essentially use Genstat in three circumstances:

- lecturing the 4th year students in biology in the module 'Mathematical Methods in Biology' at University of Paris Sud.
- in personal collaboration with non-statisticians
- during one week intensive courses which the Laboratory of Biometry organises for non-statistician colleagues at the Jouy-en–Josas centre (INRA).

The students on the course at the University are undergraduate biologists. They know nothing of Fortran or any other language. I have no time to teach them any programming but Genstat is used to present examples supporting the course and to avoid calculation in the exercises and homework. Each of them is given a copy of the relevant listing and very quickly everybody has ample practice at understanding Genstat output, so that time is available for statistical and biological interpretation. Some students just learn Genstat by themselves with minor assistance and are usually enthusiastic about it.

What I particularly appreciate in Genstat is that:

- programming with Genstat is very fast
- Genstat is versatile enough to have programs which exactly follow the teacher's argument
- statements (without options) are very clear and, even if you do not teach the language, students understand the meaning of the different statements.
- the possibility of performing the same analysis in different ways allows very good exercises (for example: 'ANOVA' or 'REGRE', 'Y' and 'FIT' or direct calculation).

With our colleagues asking for statistical advice, I often need to write a program. I then propose Genstat (mostly out of laziness!). When the collaboration is long enough, it is possible to explain Genstat while we are modifying the original program to adapt it to the developing analysis and, finally, the colleague may start to write some Genstat and read the manual by himself.

The aim of the courses we organised was to introduce Genstat and some statistics at the same time. The program was classical: introduction to the language, tables and graphs, analysis of variance, regressions and an introduction to principal component analysis. The 15 students had a terminal each and one teacher between three during the practical sessions which lasted for half of the course, the other half being devoted to academic lectures.

We encountered some technical problems, in particular we could not use the interactive version, because of environmental reasons and also because the students needed more help than we had presumed, but the main problem was the heterogeneity of the groups. Some of them knew nothing about programming or had never had any contact with a terminal, some of them knew very little statistics (often the same people). For half of the students, the course was exactly what they needed and they are now using Genstat with satisfaction: but we must admit that this group comprises only one or two persons in each laboratory (10 to 20 research workers from the 200 at the Jouy centre).

From a questionnaire which we distributed at the end of the course and from the opinions of four colleagues who gave me detailed answers about the use of Genstat by them and their colleagues, we concluded that:

- Genstat is well adapted, from a statistical point of view, to most of their problems, except for non-parametric statistics which should be handled by normal statements rather than macros.
- The use of Genstat needs a large amount of time and thought: more than is generally possible, except for a minority of students.
- A really conversational version, with 'warnings' and a good HELP statement, might be usable by more non-specialists.

Easier manipulation of MACROs and files in general is required, and also connections to graphical devices, with a simple database management system.

After this first experience (two courses this year) we now have to decide between two options:

- either to continue teaching Genstat to the majority of our colleagues, hoping that Genstat will soon be improved in the sense of more simplicity and 'transparency'
- or to choose another statistical package, able to satisfy the demands of the majority of our colleagues. We would then suggest the use of Genstat or that other product, depending on the time available and the user's knowledge of each. A decision has not yet been reached, although we tend towards the second solution, despite the fact that the system we tested (a sub-system of Multics) is extremely expensive.

Teaching Applied Statistics with Genstat in the University of York

A.J. Weekes
Lecturer in Social and Economic Statistics
Department of Economics and Related Studies
University of York
Heslington
York YO1 5DD
United Kingdom

Background

The teaching of statistics in the University of York is mainly, but not exclusively, in the hands of a small group of the academic staff who form a 'sub Department' of Social and Economic Statistics within the federally structured Department of Economics and Related Studies (hereafter referred to simply as 'the Economics Department'). Although the research and teaching interests of this group are mainly in the area of statistical applications in economics (and, to a lesser extent, in the social sciences generally), we offer courses which provide students with a general background in statistical theory and practice. Nevertheless, what we do has a certain 'flavour', in particular, our examples tend to be drawn from non-experimental sciences.

There are four degree courses in the University of York which carry the word 'Statistics' in their titles. One of these is offered entirely within the Department of Economics; the others involve cooperation with, respectively, the Departments of Politics, Mathematics and Computer Science. All are 'joint' degrees; there is no degree in statistics as a single subject.

We have always placed emphasis on practical work in statistics, feeling that it is important for students to see how the ideas which they learn in theoretical courses are used in practice. We also believe that it is important that students realise, at least at an elementary level, that 'real' data and problems are messy. With these objectives in mind, we require all students registered for one of the degree courses mentioned above to take course in applied statistics in their second and third years. For this purpose, students are not separated according to the degree for which they are registered; we believe that they should see that there is a common thread in all applied statistical work. Typically, we have about 15 students in each year.

Computational Issues

A programme of practical work in statistics inevitably requires that students be introduced to the use of a computer at an early stage. We are fortunate in that we have in the University of York a DEC-System 10 mainframe running under the TOPS–10 operating system. In consequence, we can teach programming in a time sharing environment and one of our resources is a teaching classroom with about 20 terminals. Undergraduates can usually do all their computing without ever needing to use the batch system, and we do not, in fact, tell them about it. This orientation towards time sharing strongly influences our view of how statistical computing should be done.

The operating system is friendly enough and its simple use is easily taught. The more important question which preoccupied us as teachers, for some time after these courses were set up in their present form (around 1976), was how best to proceed from that point. Several papers (for example, the writings of John Tukey and the papers by David Andrews and W J Dixon in the International Statistical Review, 1971 and 1973) on the computational aspects of teaching statistics were then fresh in our minds. These had reminded us that good analysis requires good graphical displays, a flexible approach to the data, the need to examine the data in a variety of ways and to be guided by the results of earlier steps in the analysis. This seemed to point to the idea that interactive working – not necessarily conversational – would be desirable. Furthermore, however the computation was to be tackled, it must not be 'like shelling peas whilst wearing boxing gloves'; the right tools must be used.

The obvious tool was a 'package', and several of these were available, ranging from the University of Chicago's conversational style Ida to SPSS. We considered each of these and found them all wanting in some respects. Inevitably, they reflected what the authors considered important rather than what we wanted to do. Some of them produced vast amounts of unrequired output; others would not allow apparently simple manipulations – presumably because the authors had not considered such tasks as being necessary. We could have done something by using several packages for different parts of the course but this inevitably leads to confusion and waste of time.

The way out of this difficulty seemed to be first to teach students a suitable high level language (although, even here, there was some room for discussion about which). Then, it was thought, they could write programs to do exactly what was required. A further justification for this argument was the feeling that those who call themselves statistics students (and, eventually hoped to call themselves graduates in statistics) ought to be able to program in, say, Fortran. So we proceeded to teach the basic use of the DEC–10, then introduce them to a 'package' – usually Ida – and then launch them on a course on Fortran, meanwhile suspending all statistical work on this course until they had mastered `DO` loops, `FORMAT` statements and so on. Then the real work could begin; we would (we hoped) explore with them the use of various statistical techniques, giving assistance with the programming where necessary by means of a library of Fortran procedures, which we had assembled from various sources, in particular the journal Applied Statistics.

Unfortunately – and this is, I think, by no means an uncommon finding – it doesn't work! Many students gave the programming course relatively low priority in allocating their time, partly because it had no immediate statistical content and partly because one must proceed a fair way with the 'nitty gritty' before worthwhile results start to emerge. In consequence, we spent so much time listening to excuses for failure, debugging students' programs, correcting their misunderstandings of the Fortran course and explaining how to avoid getting into messes, some of which it had not occured to us were possible, that there was little time left to tackle the interesting statistical questions which we had promised. We became unpopular with our colleagues for, as they saw it, apparently wasting students' time with seemingly trivial programming, instead of teaching them some statistics. Our collaboration with the Department of Mathematics suffered because we required the Mathematics/Statistics joint students to learn Fortran while they were still recovering from a dose of Algol! More importantly, the third year programming in applied statistics had usually to be curtailed because several of the students were insufficiently competent to tackle the programming of anything other than fairly routine statistical calculations.

I exaggerate somewhat, of course. Nevertheless, after several years of experimenting with different ways of getting Fortran into students, we remained far from satisfied about what we and the students were achieving at the end of two years of practical classes.

The era of Genstat was about to dawn, however!

The coming of Genstat

I had known about Genstat for some time before becoming an enthusiast for it. The possibility of using it for the purposes described above had been considered but it was felt – wrongly, as I hope to show – far too difficult for undergraduates, except perhaps for those whose courses included a substantial component in Computer Science. Genstat had – and still has – a fearsome reputation; many are said to be stopped by the Manual alone. (Genstat is also said to be dangerous! The opinion quoted by C W Howes in the Genstat Newsletter, September 1981, namely that "...Genstat output in the hands of the non-statisticians (is) like dynamite in the hands of non-mining engineers" has some general currency. This view is not, however, supported by my own experience.)

However, throughout 1980, I was engaged (with R.A. Cooper) in drafting a book entitled 'Data, Models and Statistical Analysis'. (Published by Philip Allan, 1983.) This required a fair amount of statistical computing, with a range of techniques from simple graphing to log linear modelling of contingency tables and some multivariate analysis. This project forced me to learn about Genstat, since this seemed to offer all that I required within a single coherent framework. From this, in turn, I was led to the thought that Genstat might well be the way to fulfil the computational aspects of our practical courses.

The first task was to convince my colleagues; the second was to deal with the question of how we were going to impart enough Genstat to our students to get them started without simply transferring the difficulties which had beset the Fortran course to a new context. The Manual itself was clearly infeasible as a textbook. The alternative, namely presentation of the basic ideas in lecture format, seemed likely to end up as dry and tedious. The main ideas had to be available for reference as the practicals proceeded. What was required was a simple beginner's guide; this could not be written, however, as though it was addressed to those who knew the statistical ideas and now wanted advice on how best to perform the computation. The first taste of Genstat would have to be motivated by the use of a few basic statistical ideas, such as summary statistics and simple regression.

The upshot of these deliberations, doubts and anxieties was a slim booklet of thirty-odd pages entitled 'A Genstat Primer – First Preliminary Edition' and the course in the 1981/82 academic year was based around this. We met the students twice a week; the first meeting would take one of the Chapters of this primer as a text for the sermon, fill out the finer detail and deal with any questions. We then provided an exercise which drew upon both the Genstat and upon the statistical ideas which students had so far encountered. The latter ranged from simple summarisation to elementary model fitting and hypothesis testing.

The second meeting took place in the 'terminal classroom', where students could practise their understanding of Genstat by writing and running programs on line. These sessions were under the supervision of myself and a colleague. Any difficulties of a computational nature – to do with understanding Genstat or the University computer's operating system – were sorted out, usually on the spot. A debriefing on the statistical aspects of what had been done would take place before the next expository Genstat session. In total, this took about 2 hours of the students' time per week, for two terms.

Some impressions of the first two years with Genstat

The experiment was a success. We found that the flexibility of analytical approach which we, as teachers, were trying to encourage was matched by the students' feeling that they were making progress – in some cases excellent progress – with the task of learning Genstat. This first cohort of students entered their third (final) year with skill and confidence in the elementary aspects of Genstat. The better students turned naturally to the Genstat manual to find out more; the less committed (or able) could at least keep their heads above water and produce reports on practical assignments without computational problems adding to their burdens. In their third year, they completed a solid programme of work based upon the use and interpretation of linear models. In terms of Genstat, the early stages of this work could be done by means of standard directives and required only minor extensions of what we had done in the second year. From here we went on to explore the use of 'influence' statistics and adaptively weighted least squares; this was made possible by Genstat's ability to perform a range of standard matrix operations such as singular value decompositions. We concluded the course by demonstrating some aspects of the problem of fitting models to two- and three- way contingency tables.

Credit for this second-level course was obtained from completion of two pieces of work chosen from a list provided by the course teachers. This work was mostly of a high quality – no one had been forced to avoid some difficulty in the analysis because 'the program doesn't allow that' and, equally important, no one could go beyond their statistical depth and attempt to impress us simply by using 'advanced' techniques, since, in my experience, one must understand what one is doing before Genstat will let you do it. A final (unexpected) satisfaction was that students could now add a knowledge of Genstat to their CV's when entering the job market – a useful string to have on one's fiddle.

The analogy between learning a practical skill such as woodwork and learning applied statistics is one which has occurred to me several times during the last two years; it clearly underlies some of the preliminary discussion in this paper. The Genstat system itself is usefully likened to a tool-kit, from which the user can draw a particular tool when required. The use of a simple tool is easily demonstrated, and the beginner can take some pride in having used it successfully; the more advanced tools require some considerable understanding for their use and are inaccessible until that understanding is achieved. Furthermore;, by comparison with certain of the better known statistical 'packages', the tools for displaying results are pleasingly parsimonious and the user is not overwhelmed with exuberant masses of output. Hence the statistical aspects of what had been done can be discussed without having to wade through a sea of paper.

The allegation that Genstat is, somehow, dangerous is now easily disposed of. By comparison with many packages, the more advanced procedures in Genstat typically require the user to have some understanding of what he/she wishes to do before the relevant tool (the appropriate directive) can be used. Choosing what is to be output is similarly limited by one's understanding of the theory of what has been done. There is no way that a user can go beyond his or her 'depth'. (The dangerous/safe distinction is inappropriate anyway.)

Some questions which have been asked

Were there no problems?

Lest all this sounds too eulogistic ('I used to think that statistical computing was a bore until I discovered Genstat', to paraphrase a well-known advertising slogan), it is worth recalling that there are *some* problems with using Genstat in the way described here. Briefly, these can be summed up under three headings: (i) detecting the cause of error messages; (ii) the problems of the interface between Genstat and the Computer's operating system and (iii) the problems caused by the student with the baroque mind.

Perhaps the biggest difficulty which a beginner in Genstat must face is that of deciding exactly what has caused a message such as:

```
LINE 10 STATEMNT 0 FAULT VA 4.
```

When working with students in the terminal classroom, we found that, as our own experience with Genstat grew, we could usually dispose of problems such as these without necessarily knowing exactly what fault `VA 4` is. The inexperienced beginner working alone, however, needs first to decipher the code and then track down the cause. This can be a very discouraging task. Some improvement has been achieved by our Computing Service who have amended Genstat so that it puts out a supplementary error message. (This supplementary message is, I understand, standard in Genstat 4.04. We are using 4.03 at York.) In the case above, the user will also be told:

```
Values not set
```

which is invariably a great help and is well worth having. It points, of course, only to the proximate cause; most faults have more subtle roots. Indeed, in the case whch prompts this example, the user had declared a structure with the identifier '`SYL`' and later referred to it as '`syl`'. Experienced users will immediately spot the problem; Genstat treats these as two different names. The beginner is likely to puzzle long and unsuccessfully over what is wrong, even when it is known what `FAULT VA 4` is. (It can be salutary for more advanced users, particularly those involved in teaching or writing documentation, to think back occasionally to their own first encounters!)

There is no simple answer to this difficulty. It can (and should) be pointed out that much can be learned from careful study of the causes of mistakes and that to be discouraged is not the right reaction.

We also had some difficulties in the beginning with the interface between Genstat and the computer's own operating system. One problem was that the user was required to give the input file a particular name ('`FOR20.DAT`') and to assign the operating system's channels before running Genstat. If the output was to be sent to a file then this too was given a particular name by the operating system. Failure to rename this output file before running Genstat a second time would result in the loss of earlier results, due to their being overwritten. This problem has been partly overcome by means of a 'front end' to Genstat which asks the users about file names for input and output. Other problems of this type remain; they are, I believe, a consequence of using a program designed with a batch operating system in mind in an interactive environment. Collectively, they have the unfortunate consequence of adding to the beginner's problems by requiring a new layer of conceptual understanding at a time when there are many other complexities to be mastered.

The typical consequences of what I call the baroque mind are these: a problem is tackled with such an zeal to use all of some newly acquired (and only partly understood) knowledge that the resulting program becomes excessively complicated. Something then goes wrong – usually the program runs but produces patently absurd answers – and no one, not even the writer, can see why. The problem is not caused by Genstat, of course; the same person would probably make a similar mess using Basic, Fortran or any package. It is very difficult to cope with and the only answer seems to be to find time within the course for some discussion of programming 'style'.

Is Genstat all that is required?

No; we still show students how to use Ida and encourage its use for those tasks which it does well.

Why did we choose Genstat rather than GLIM?

Because we wanted to make use of the wider range of techniques available in Genstat.

The present

For the incoming second year students, the pattern and organisation of practical work will be broadly similar to that followed in the 1981/82 session. 'Data, Models and Statistical Analysis' has now been published and will, we hope, provide a satisfactory reference for the statistical content of the course.

The Genstat Primer itself has been revised twice. It now runs to about 70 pages and, with the benefit of experience, we have been better able to define a minimal subset of Genstat. The following is a list of the chapter headings in the present version:

Chapter	Title
1	Structures, Statements and Summary Statistics
2	Calculations in Genstat
3	Naming and grouping structures; Integer Structures
4	Improving the appearance of output
5	Graphical displays
6	The 'UNIT' and 'COPY' directives
7	Factors
8	Naming factor levels
9	Using Genstat to fit regression models
10	Some other Genstat structures (principally matrix structures)
11	Some notes on multivariate analysis in Genstat
12	A few extensions (mainly things like 'READ' with a read format, reading from a different input channel etc.)

It will be seen that about three quarters of this document is concerned with matters of 'housekeeping' – structures and structure types, data input and output etc. and the rest with the use of the more specialised aspects of Genstat in, for example, regression and multivariate analysis. In selecting the order of material for presentation (and then the level of exposition within a Chapter or section), the guiding concern was to get the beginner started. Formatted printing (and parallel printing), for example, comes under the heading of 'improving the appearance of output'; until then, it is sufficient that calculations are made and the results displayed. Certain topics, such as the definition of table structures and operations thereon, have been omitted entirely. Certain others have been played down, for example a complete survey of the different ways there are for generating factor levels. Some have disappeared and then reappeared in the process of revision, for example the use of the 'FOR' ... 'REPEAT' construct. We now include a discussion of this with an appropriate warning about thinking before using it. (Our doubts about the place of this were initially caused by seeing students using it to do such things as summing the elements of a variate structure, instead of using the SUM function. This is an example of the 'baroque mind' at work!)

For those students now entering their third year, we will probably do much as before. The main text will be Chatterjee and Price: 'Regression Analysis by Example' (Wiley, 1977), although this is now horribly expensive, and Cox and Snell's 'Applied Statistics' (Chapman and Hall, 1981). We now know something about the feasibility of completing the programme of practical work using Genstat and we may vary the pace a little; in particular, we intend to do a little more on generalised linear models if time permits.

Finally, it is worth noting another success, first achieved in the last academic year and hopefully to be repeated in the coming year. In the Spring Term 1983, John Byrne (of the University of York Computing Service) and I collaborated to offer a course on Genstat to a group of postgraduate students (mostly reading for the M.Sc. in Biology and Computation) and some academic staff. We assumed a knowledge of the relevant statistical ideas and concentrated on getting Genstat across. The Genstat Primer was used as the main reference, with some supplementation provided by John on the analysis of designed experiments. The presentation was intensive and took the form of an hour's lecture followed by a practical. We invited participants to comment fully on what we had done when it was all over. Enthusiasm was high and the Primer received generally favourable notice.

The future

We have now sufficient confidence in what has been achieved to feel that, within the next year or so we might actually stop preaching a sermon on each of the chapters of the Primer and simply leave students to read it as directed. The time thereby released could then be used – after dealing with any questions arising from the reading – to talk about statistical issues.

We have also thought about extending the Primer (or writing a follow up), but this is a more ambitious task since, inevitably, the discussion of Genstat must be merged more fully with a discussion of the relevant statistical theory. If time for practical work could be extended, we could see ourselves doing something with Genstat's capabilities for multivariate analysis, using both the relevant directives and the macro library. This development could be a worthwhile complement to our third year optional course which deals with the theory of multivariate analysis.

Finally, it seems that we may have converted our colleagues in other Departments to our enthusiasm. The possibility of teaching Genstat to undergraduate students who are reading for joint honours in Biology and Computation is one which has been mooted, and we have even persuaded one or two quantitatively minded economic historians to take a look at what it has to offer.

Acknowledgements

I am grateful to colleagues in the University of York who have taught with me on the courses described above, especially R.A. Cooper and D.S. Poskitt. Their suggestions have been invaluable. I must also express my thanks to John Byrne who first showed me what Genstat could do and who has greatly encouraged the teaching developments recorded in this note. His efforts to make Genstat more user friendly, particularly by providing enhanced error messages and an interface between the user and the operating system, also deserve gratitude and acknowledgement.

Teaching Genstat to Undergraduate Students in Applied Mathematics in the University of Genoa, Italy

Giovanni Pistone, Ivano Repetto***
Instituto di Matematica
via L.B.Alberti 4
16132 Genova
Italy

* partially supported by the MPI–project 'Matematica Computazionale'
** partially supported by the CNR project 'Informatica'

Computing facilities for students' use at the Instituto di Matematica, Università di Genova

Our students attend a 4–year course for the degree 'Laurea in Matematica'. There are three sections: 1. Pure Mathematics, 2. Mathematics for Teaching, 3. Applied Mathematics.

It is important to note that our University does not provide a degree in Computer Science and most of the students on the Applied Mathematics section have in fact a curriculum midway between Mathematics and Computer Science.

The first two years are the same for all the sections; at this level Computer Science and Statistics are taught as complements to the general courses of Algebra and Calculus. Students learn to use programmable pocket calculators and the Basic programming language on HP9885 desktop computers. Some of the exercises proposed cover the subjects of finite probabilities, Normal distributions, tabulation and simple regression.

The last two years' courses have different lectures which depend on the section. Computer Science, Statistics and Probability are not considered as belonging to 'Pure Mathematics' and they are taught only in the other two sections. Computing facilities consist of two Digital PDP–11 mini computers, one HP9845 desktop computer and three video-terminals connected to the Burroughs B6810 of our Computer Centre.

The PDP machines are used mainly for exercises on assembler language, Pascal, Compilers, Data Bases, etc, and have no mathematical or statistical libraries. The HP9845 is used mainly by students of 'Mathematics for Teaching' section on the grounds that they are more likely to find similar computers in their future employment as secondary school teachers.

The Burroughs B6810 is used by students in Applied Mathematics, programming in Algol and Fortran and using Genstat and the NAG library (in both the Fortran and the Algol versions).

After the completion of all the courses each candidate to the degree gives a short dissertation on a research problem in Mathematics or a report on a stage work. The need of some sort of statistical library comes from the growing importance of the statistical part in the stage work.

A student with some notion of programming and statistics tries to solve his problems by writing Fortran or Pascal code for the statistical method he needs and spends most of his time in writing and debugging his program, even in the case of completely standard algorithms.

In the authors' opinion this situation is not satisfactory and we developed the idea of teaching something about the use of statistical libraries during the third year-course in Probability and Mathematical Statistics. We tried first to state what kind of packages are to be considered standard in our environment but we concluded that – despite some local preference for SPSS and BMDP – no standard package existed.

Colleagues of the IAC/CNR Institute in Rome and of the Faculty at Orsay described the particular features of Genstat to us and we chose this in 1981.

We stress that most of the staff of our Faculties who are more or less concerned with statistical computation do not agree with our choice and the argument is still current.

To use Genstat or write new programs? The teachers' point of view

In September 1983 the following text was submitted to various colleagues in our department, to verify current opinions on statistical software.

'The applied mathematician interested in applied statistics and the teacher willing to introduce some concrete examples in a course in Mathematical Statistics can chóose among three ways of obtaining numerical and graphical answers:

(1) to write, in every case, a new program, using scientific programming languages;

(2) to construct his own library of general programs consulting, when necessary, the published literature and scientific libraries;

(3) to use some general statistical packages such as BMDP, Genstat, GLIM, P-STAT, SAS, SPSS.'

We received the following comments (here freely translated and summarised).

A. Belcastro (Instructor in the mathematical courses for biologist and statistical consultant for the medical research center Istituto Scientifico Tumori): The computers at my disposal are one HP9885 and one HP1000, without any general library, and they have to be used to write interactive programs to fit the needs of biologists and clinicians. Moreover I am not very familiar with the general statistical programs in point (3) and I like to write my programs from the beginning. So my choice is method (1).

P. Boero (Lecturer in 'Complements of Mathematics for Teaching' and co-director of a National Research project on the modelling of meteorological data): I do not like the use of 'general' statistical programs because they cannot cover the variety of interesting applications and encourage the students to a blind use of a set of recipes.

E. Guala (Lecturer in 'Probability and Statistics for Teaching'): my students mainly need training in basic arguments in Probability and Statistics and solutions (1) and (2) fit better with their interest. The use of general programs is better conceived for applied mathematicians and works better in statistics than for future teachers.

In the author's opinion, most of the criticism about the use of 'general libraries', and in particular Genstat, comes from partial knowledge of the subject. In particular, the difference between normal libraries and a 'system' or 'language' is not well understood.

Without trying to give a detailed analysis of special features of Genstat (see, for example, [11]) we recall that many users (see examples [2], [9]) stress the importance of statistically oriented data structures and of the presence of both numerical and statistical subroutines. For example it is very important for the teacher to be able to illustrate the same example – say a regression – with both its own, special commands and with the matrix 'CALCULATE' and to do the analysis in reverse order (say from the factors to the original data).

In our experience, the correct use of Genstat requires a sound knowledge of statistics: for example we cite the knowledge and ingenuity necessary to interpret the output of the 'REGRESS' or 'ANOVA' command. From this point of view, Genstat is not suitable for a novice student. But exposure from the beginning to a small subset of the language could be a good introduction to later, more advanced use.

Finally, it is clear that the research worker cannot solve all of his computational problems within a particular package but our experience is that much of the preliminary work can be done in this way.

To put the preceding comments in their proper perspective, we stress that nobody in our Department is a specialist in statistics but we are all mathematicians with some interest in teaching and some contact with statistics as a tool in applied scientific research.

How we teach Genstat to students in Applied Mathematics

We present our students with a very small subset of Genstat, namely the data structures, the input/output operations, and the 'CALCULATE' command. Some hints are given regarding the statistical commands, such as 'REGRESS' and 'ANOVA' and the interested student is asked to consult the manuals [3], [8] to solve his particular problems. The rationale underlying this choice is the idea that the main pedagogical function of Genstat in our courses comes from the statistical ingenuity incorporated in the definitions of data structures and the possibility of discussing meaningful statistical examples, bypassing all problems in data manipulation, graphical representation, and computation with vectors and matrices. Moreover, the student who will work in a staged program will know of the existence of specialised statistical commands.

Eight to ten hours of classes are used to explain this subset of Genstat by showing simple examples of programs [13]. Most of our students have had two courses of Computer Science and Programming Languages and know Basic, Pascal, Algol, and Fortran, so we do not need to explain terms such as 'declaration' or 'compile time'.

In the practical session, the students are asked to solve one or two very simple but concrete statistical problems: we try to obtain some problems during consultations with colleagues in the experimental sciences and to have them explain to the class the real meaning of these problems.

In our experience, the main difficulties which students meet come from the special features of the Genstat language and from the lack of complete documentation. The use of lists of identifiers instead of indexes is confusing for a beginner, who thinks in terms of loops and recursion, but this problem is quickly solved because most of the students know how a computer works. On the other hand, the problem of documentation is more serious: what we require is a different type of manual with precise information about the numerical algorithms and the general structure of the Genstat Fortran Code. (The new French manual published in 1982 by the INRA will probably solve some of these problems.)

Experience of the use of Genstat during staged work

During the academic years 1981/82 and 1982/83, five reports using Genstat were presented to the Faculty board examiners for the degree. Probably the same number will be presented next year. The subjects treated are parts of applied projects in which people of our department are involved.

(A) New methodologies in Surgery

Our department is collaborating with the department 'Patologia chirurgica' of the School of Medicine, in the statistical analysis of clinical records. These refer mainly to a new operation – the biliopancreatic bypass – developed by N. Scopinaro and co-workers for the treatment of morbid obesity. From the tutorial point of view, it is a very interesting subject involving many statistical methodologies, such as the design of experiments and time series. Also the number of data – nearly 250 patients to date, eventually with multiple records – fits very well with the typical dimension of problems we can treat with our approach. We have also begun to study the project of automatic maintenance of files of clinical records and the interfacing of our data acquisition system with the statistical software. This will be discussed in the next section.

(B) Quality assessment in the food industry

This a new project, the first draft being produced in the summer of 1983. We will investigate the applicability of clustering and classification techniques to comparing the chemical and physical analysis of foods with the response of a panel of professional tasters making sensory evaluations.

(C) Modelling of meteorogical data

A part of the national CNR project concerning the treatment of meteorological data is based on the geophysics department and our own department. The work consists in the production and maintenance of a database of meteorological data of local interest.

As explained before, we encourage the students to refer to Genstat and other software in a systematic way to solve programming problems. However, some situations were met where we could not find how to write simple and efficient algorithms in Genstat.

Examples are:

- the computation of the maximum likelihood statistics in the problem of multiple Markov dependence;
- the manipulation of large time series, classified by year/month/day.

Creation of a data base compatible with Genstat

In the search for a constant supply of real data sets for presenting interesting problems to our students, we were faced with the problem of interfacing a data entry system with the statistical software. It is a natural idea, because most of the laboratories of our University are connected through video terminals to the Burroughs B6810 of our Computer Centre and the operating system allows very simple access to the (public) files of all users.

In 1982 we began to write – in collaboration with the Surgical Department of the School of Medicine – a set of programs to perform the following operations:

(1) input and validation of clinical records
(2) daily output of updated patients' charts
(3) update of files of clinical records organised as data structures and selection of subsets for tutorial use
(4) output of statistical analysis each 2 month.

The first idea was to do (4) with Genstat and to maintain the files (3) as Genstat userfiles. Moreover we tried to write Genstat programs for (1) and (2).

The experience was very interesting in a period when the main problem was to learn in detail how Genstat works. From a practical point of view the result is not satisfactory because:

- on our system the execution of the backing store commands for a big userfile requires times of the order of a minute;
- the time required to update a single clinical record, by running a Genstat job in time-sharing mode, is very variable and this is unacceptable to the users.

We are currently working on a more complex system, using a Burroughs B22 mini-computer connected to the mainframe. The work splits in two parts: operations (1) and (2) are done on the local mini and its disc, while operations (3) and (4) will be done by Genstat and Fortran programs running on the B6810. We expect that the programs running on the mainframe will access the local system as an external file.

Our guess is that the addition to Genstat of some routines for data input and validation could be useful in many analogous situations, where a data set not so big as to require a true data base but is updated considerably more frequently than it is analysed.

References

[1] Alvey, N.G. (1980) Genstat Standard Analysis Forms, The Statistics Department, Rothamsted Experimental Station.

[2] Astier, R. (1982) Post-graduate use of Genstat.
The Genstat Newsletter, 9, March.

[3] Brambilla, C. and Gherardini, P. (1981) Il sistema Genstat.
Quaderno IAC No. 125, Roma.

[4] Di Giorgio, F. and Mong, R. (1982) Burroughs B6810 Genstat Local Guide.
Centro di Çalcolo dell'Università di Genova.

[5] Gambaro, C., Miziti and Pistone, G. (1982) Corso Genstat 4/7 maggio 1982.
Centro di Calcolo dell'università di Genova.

[6] Guala, E., Parenti, L. Repetto, I. and Zappa, A. (1983) Corso di aggiornamento 'Calcolo e calcolatori'.
Istituto di Matematica, Genova.

[7] Moro, M.L., Pesce, G. and Pistone, G. (1982) Programmi per l'analisi delle corrispondenze,
Genova.

[8] Nelder, J.A. e al. (1980) Genstat. A General Statistical Program.
Manual rel. L. 03. Part I: Informal Introduction. Part II: Formal Description, NAG, Oxford.

[9] Payne, N.W. and Nelder, J.A. (1976) Data structures for statistical computing.
Proceedings of the 9th Biometric Conference, Vol. II, 191–208.

[10] Pesce, G. and Pistone, G. (1982) Un sistema di archiviazione di cartelle cliniche orientato all'elaborazione statistica. Programma in linguaggio Genstat, Genova.

[11] Pistone, G. (1982) L'uso del programma Gensta nella didattica della Statistica Matematica, Cagliari 1982 (Quaderno C.N.R. 1983, W. Ravagno ed.)

[12] Pistone, G. (1983) Lezioni di Statistica Matematica, Dispense 1982/83 dello Istituto di Matematica, Genova.

[13] Repetto, I. (1983) Esercizi di Probabilità Statistica, Dispense 1982/83 dello Istituto di Matematica, Genova.

[14] Repetto, I. (1981) Un'esperienza dell'uso dei calcolatori tascabili (CT) in terza media, L'insegnamento della matematica, 4, No. 1, 1981.

Use of the new Genstat Graph Facilities in the Analysis of Data from Plant Weight – Density Studies

G.E.L. Morris
National Vegetable Research Station
Wellesbourne
Warwickshire CV35 9EF
United Kingdom

Many vegetable crops are grown in small modules, in a glasshouse, until they reach an appropriate size and are then transplanted into the field, rather than being sown directly. Increasing costs of materials, fuel and transport have produced a trend towards the use of smaller modules with the result that plants are being raised at higher densities. Plant density is known to affect the rate of plant growth and will consequently affect the time needed to produce plants of a suitable size for transplanting.

A project has started at N.V.R.S. to provide quantitative information, for a range of crops, on the relationship between plant weight (w) and density (d) during plant raising and to study how it varies with time. The relationships, at any given time, have been found to be well represented by inverse linear polynomials of the form:

$$\frac{1}{w} = \frac{1}{w_0} + bd$$

The parameter w_0 has a practical interpretation, the weight of a plant grown in isolation.

Experiments are conducted, using a wide range of densities, to estimate w_0 and b at a range of harvests. The fit of the equation to the data (using Gamma errors with inverse link) from each harvest is usually very good (Frame 001). Linear interpolation between the estimated values of w_0 and b from each harvest (Frame 002) is used to estimate the value of b for any intermediate value of w_0. This is done for a range of values of w_0 and the resulting 'family' of weight-density curves drawn to provide a means of quickly estimating density effects at harvests other than those used in the experiments (Frame 003); the value of w_0 for each curve is written alongside it on the graph – these values can be related, at least qualitatively, to harvest.

Data from each experiment are analysed by a single Genstat program which estimates the parameters, w_0 and b, for each harvest using the generalised linear model facility, produces graphs to illustrate the goodness-of-fit, interpolates parameter values and displays the resulting 'family' of curves using the interpolated values. All graphs are drawn on a Benson 1202 incremental plotter. The program is written to deal with variable numbers of harvests and densities and a range of values of w_0.

The example program, listed below, contains only part of the main program relating to the plotting of various graphs once the inverse polynomials have been fitted to the data from each harvest separately. Graphs involving plant weight are plotted on the untransformed scale so the resulting lines are curved: it has been found that dividing the density scale into 50 intervals, equally spaced on the log scale, and using the smooth line facility to join the corresponding fitted weights produces a good result. The example is set up so that if different values of w_0 are required only one statement (the third in the program) need be altered: in the example there are 9, the minimum and maximum corresponding (with some rounding) to the minimum and maximum obtained over the harvests. The number of harvests has been fixed as *4* (the number normally used in the experiments).

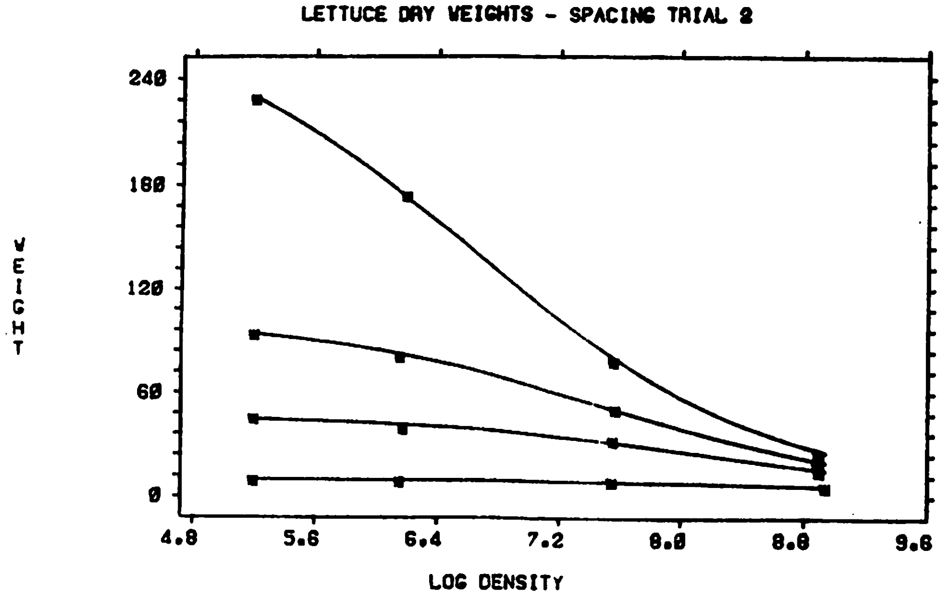
The program contains examples of the following facilities, which became available in Mk 4.04.

(1) Output to a high quality plotter using the options DEVICE and BUFFER.
(2) Setting the symbol size using the option SYMBOL.
(3) Main title to a frame using the option TITLE.
(4) The x-axis titles are allowed to spread over more than one line – in the program they all start with a new line in order to clearly separate them from the x-scale labels.
(5) Insertion of text within a frame: note the use of JOIN to set up the headings which contain the text to be inserted.

Example Program

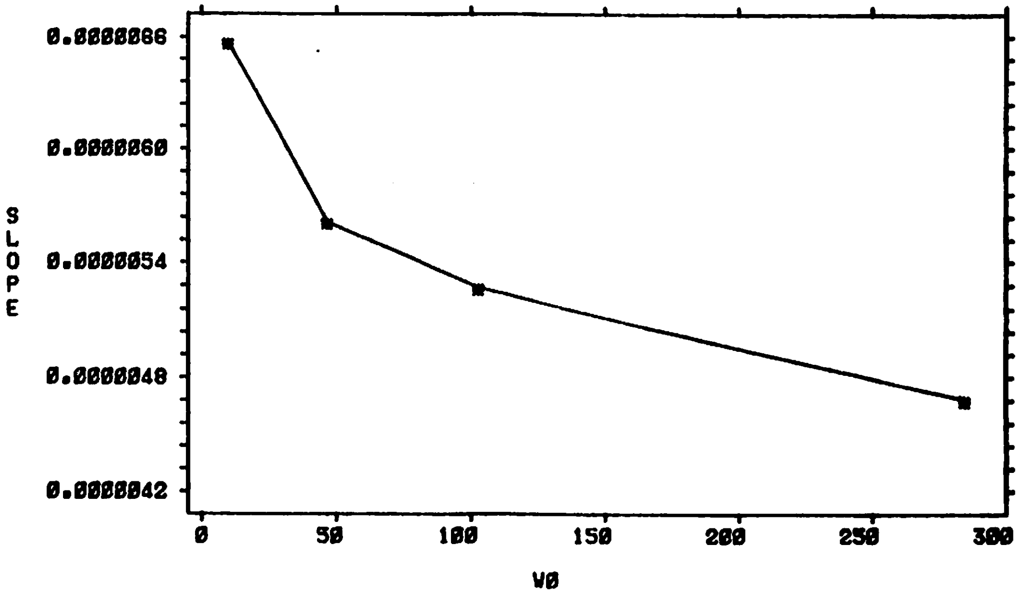
```
'REFE' PLOTS
''  Set up values of WO for interpolation.''
'INTE' INTWO=5,10,25,50,100,150,200,250,300
 'SCAL' NWO,NWOM1
'CALC' NWO=NVAL(INTWO)
 : NWOM1=NWO-1
'RUN'
'INTE' IH=1...4
 : IL=1...NWO
''
  YL(I) & XL(I) hold co-ordinates for position of label 1 in final graph
  (I=1...NWO).
''
'SCAL' YL(IL),XL(IL)
''
  WT(I) holds the observed mean weight at harvest I (I=1...4).
  LOGD holds the observed density.
  BFIT & WOFIT hold the slope & reciprocal of intercept for the
  fitted line at each harvest.
  LOGFITD holds the logs of the densities used to calculate
  fitted values which are joined to produce smooth curves.
  OFIT1(I) holds the corresponding fitted weights at harvest I(I=1...4).
  FIT1(I) holds the corresponding fitted weights calculated from
  the I'th interpolated value of WO (I=1...NWO).
''
'VARI' WT(IH),LOGD $ 4
: BFIT,WOFIT $ 4
: OFIT1(IH),FIT1(IL),LOGFITD $ 51
'HEAD' HO
: H1='' WEIGHT V. DENSITY FOR INTERPOLATED WO''
: H2=''PPPPSSSS''
: H3=''LP''
: H4=''S''
: HS=''S''   : HT=''T''
: HY2,HY4=''WEIGHT''
: HX2,HX4=''LOG DENSITY''
: HY3=''SLOPE''
: HX3='' WO''
: HL1=''*''
: HL(IL)
```

```
'INPUT' 2
'READ' HO
'READ/P' OFIT1(IH),LOGFITD
: WT(IH),LOGD
: BFIT,WOFIT
: FIT1(IL)
: YL(IL),XL(IL)
'INPUT' 1
'RUN'
'INTE' NF $ 1
''
  Set up headings HL(I) (I=1...NWO) which will hold the label to be
  plotted in the frame at the point (XL(I),YL(I)).
  The label is the corresponding value of WO.
''
'FOR' DHL=HL(IL) ; DS=1...NWO
'COPY' NF=INTWO $ DS
'JOIN/VAR=1,LABR=1' DHL=NF $ 4
'REPE'
''
  Set up heading H4 which will define the types of plots in the final
  graph.  It comprises the letter S NWO .times followed by the letter T
  NWO times.
   Since H4 was set equal to ''S'' (to avoid any problems
   which might arise with null headings) the first loop need
   only be traversed NWO-1 times.
''
'FOR' DS=(HS)NWOM1
'JOIN' H4=H4,DS $ 0,0
'REPE'
'FOR' DS=(HT)NWO
'JOIN' H4=H4,DS $ 0,0
'REPE'
'OUTPUT' 2
''
```



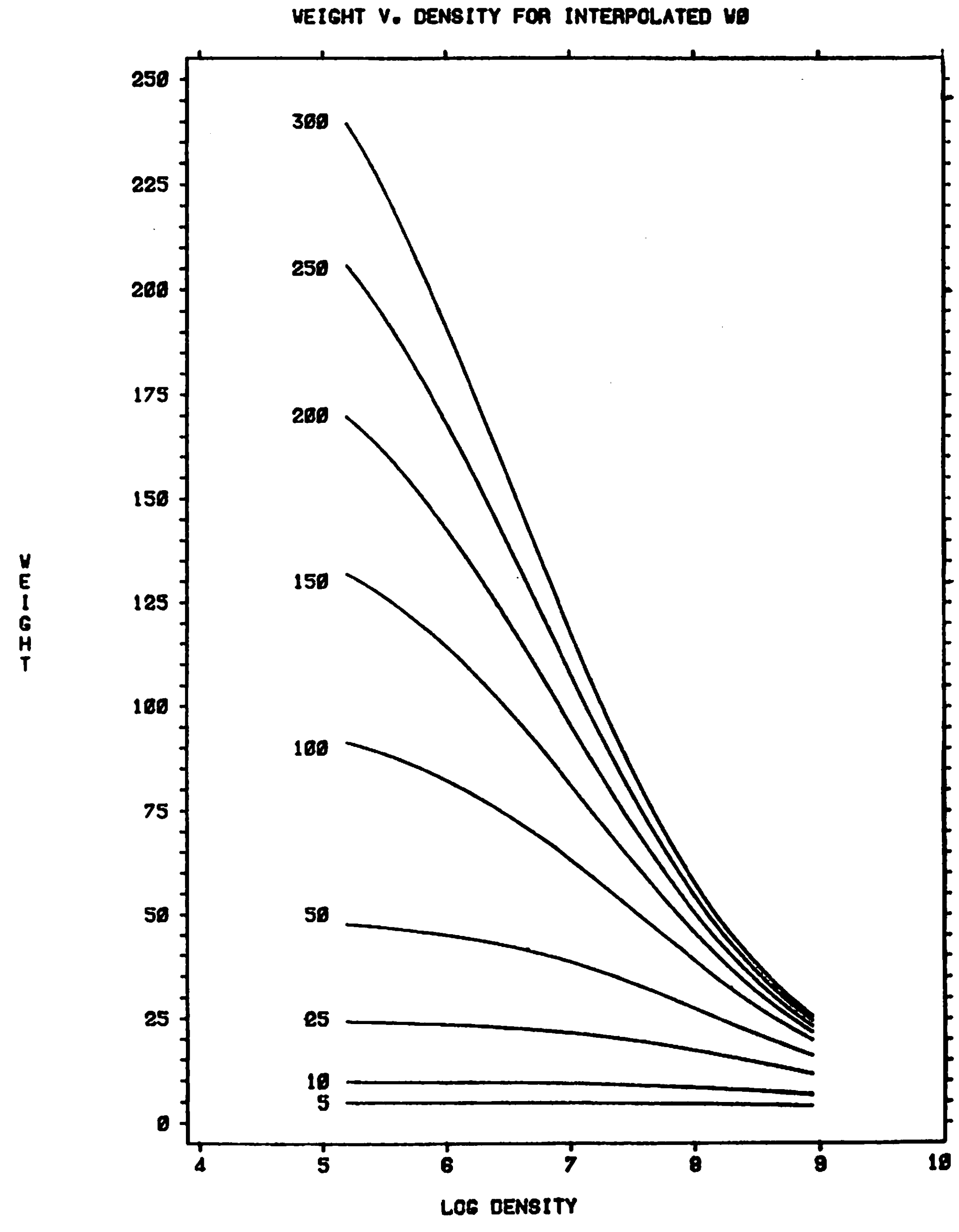
Frame 001

```
..
'GRAPH/DEVICE=1,BUFFER=N,ATY=HY2,ATX=HX2,SYMB=0.25,NRF=21,NCF=60,TITLE=H0'
 WT(IH),OFIT1(IH) ; (LOGD)4,(LOGFITD)4 $ H2
..
```



Frame 002

```
..
'GRAPH/DEVICE=1,BUFFER=N,ATY=HY3,ATX=HX3,SYMB=0.25,NRF=21,NCF=60'
 (BFIT)2 ; (WOFIT)2 $ H3
..
```



Frame 003

```
..
'GRAPH/DEVICE=1,BUFFER=N,ATY=HY4,ATX=HX4,SYMB=0.25,NRF=51,NCF=60,TITLE=H1'
 FIT1(IL),YL(IL) ; (LOGFITD)NWO,XL(IL) $ H4 ; (HL1)NWO,HL(IL)
'OUTPUT' 1
'RUN'
'CLOSE'
'STOP'
```

Genstat Analysis of Variance and the Distant Client

D.A. Preece
Rothamsted Experimental Station
Harpenden
Hertfordshire
England

Abstract

The author is from the Overseas Unit in the Rothamsted Statistics Department. That Unit produces Genstat analyses for agricultural research workers in distant lands. Some of these clients have only an elementary knowledge of statistics and need to be spared any risk of misunderstanding the computer-produced analyses of their experimental results. How nearly self-explanatory, then, is the output from Genstat ANOVA and regression, and how likely is the agronomic client to misinterpret it? Any problems of interpretation are most likely to concern one of the following: (i) strata; (ii) the 'dot' notation as used when A.B means 'B within A'; (iii) coefficients of variation; (iv) covariance analysis; and (v) non-orthogonal analyses. The paper discusses the clients' possible difficulties with these concepts and with output involving them; some of the difficulties are specific to Genstat analyses and some are not. The author discusses whether further automatic annotation of the output is desirable in some of the analyses.

This paper has been prompted by my experience during the last five years, whilst I have been a member of the Overseas Unit in the Rothamsted Statistics Department. That Unit produces Genstat analyses for agricultural research workers in various 'developing' countries. Many of the analyses are sent by post to distant lands, perhaps to remote places. Some of the distant clients are agronomists with only an elementary knowledge of statistics. In these circumstances it is important to try to spare the clients any risk of misunderstanding the computer-produced analyses of their experimental results and to produce output which can readily be understood without resort to time-consuming correspondence.

Most of the analyses with which the Overseas Unit has been concerned are regression analyses and, much more commonly, ANOVA analyses of variance leading to the production of tables of means with standard errors appended. So I restrict myself now to these two types of analysis and ask, "How nearly self-explanatory is the current output from Genstat ANOVA and regression, and how likely is the agronomic client to misinterpret or misuse it?"

I have chosen five concepts which I consider to be the most likely to give my clients difficulties of interpretation: (i) strata; (ii) the 'dot' notation used in 'A.B'; (iii) coefficients of variation; (iv) covariance analysis; and (v) non-orthogonal analyses. I shall consider these five in turn. Some of the points that arise are specific to Genstat; others are not.

Strata

The experimenter whose knowledge of statistics is tied to what can be found in books such as those of Cochran and Cox (1957) and of Little and Hills (1978) cannot be expected to have heard of the strata that appear in Genstat analyses of variance. This matters very little, as output statements about strata can be ignored. An experimenter may be puzzled to find that his blocks sometimes have a stratum of their own and sometimes – following the Manual's comment about simplifying both computation and output – appear in the same single stratum as his treatments. The reason for the discrepancy can however easily be explained. The only appreciable problem to arise when the blocks of a randomised complete block design are coded as treatments within a single stratum is that standard errors are then provided for block means or differences between block means. I have yet to find an instance of such spurious standard errors being copied into a report or published paper but the fact that they may have been 'produced by Rothamsted' makes it only a matter of time before they are so copied or at least puzzled over. I have therefore often tediously spent time crossing such standard errors out. Their promulgation is bad statistical practice (but a practice that I know to be current). A simple mechanism for suppressing them would be welcome, and should be easy to devise. Otherwise, blocks should perhaps always be put in a stratum of their own, whatever the additional computing cost; a new facility for producing block means (which are often wanted) would then be desirable, as non-automatic tabulation and labelling of the block means would be too tiresome and fussy for the user to do frequently.

The 'dot' notation

That the symmetrical-looking notation A.B **can** mean 'B within A' is something that we cannot expect all our clients to know. This matters little, except for the labelling of strata in analyses of variance. Even with well-chosen, informative factor names, the strata names for an experiment with plot-splitting can be daunting, and even something like

```
BLOCKS.WHOLEPLS.SUBPLS
```

is less than satisfactory for an agronomist with little statistical expertise. (Nor would we wish to see

```
BLOCKS.WHOLEPLS.SUBPLS
```

copied straight across to a report or published paper.) Here it seems reasonable to ask that a future Genstat should itself make the translation to

```
SUBPLS WITHIN WHOLEPLS WITHIN BLOCKS
```

Coefficients of variation

That Genstat should provide a coefficient of variation (CV) for each stratum will surprise some of our customers. But the main point to be made about CVs is something else altogether. Programs and packages other than Genstat automatically provide CVs for all variates analysed, including variates consisting of scores and percentages, and including variates transformed to square roots, logarithms and angles. As a consequence, many agricultural research publications of recent years have been printing CVs at every opportunity, often for variates whose CVs are unlikely to be of interest This overuse of the CV goes hand-in-hand with widespread misunderstanding: the CV, rightly recognised as a dimensionless quantity, is seen as an **absolute** quantity, such that a 30% CV is regarded as large and bad, whatever the variate. The experimenter, appalled by a 30% CV for the percentage of plants infected, does not notice that the CV for the percentage **not** infected is, say, 3%! Faced with this situation, I believe that statisticians have a duty to be cautious in their provision of CVs. Whether a Genstat option is required, whose default would suppress CVs, I am not sure. At the very least, I suggest suppressing the CVs for any variate containing a negative value; this would at least suppress the CV for some transformed variates, and would sidestep the recent curious disagreement about whether the CV for a variate with a negative mean should be negative or positive.

Covariance Analysis

Covariance analysis, like the coefficient of variation, is not understood by many who use it. An important difficulty of understanding is associated with analyses where a covariate (i.e. x–variate) is influenced by the treatments – but there seems to be little or nothing that a program-writer can do about this. Another problem arises with factorial experiments, where – as Preece (1980) and Bingham and Fienberg (1982) pointed out – a covariate induces slight non-orthogonality between effects which would otherwise be orthogonal; the present Genstat algorithm takes no account of this non-orthogonality and prints no warning of what is going on, but I have no reason to think that this matters, at least for agricultural experiments. I am happy enough with current Genstat covariance output, except that I recognise a need for other more detailed explanation to be printed. Each covariance analysis is preceded by the name of the *y*–variate but not by those of the covariates; this is a major nuisance for the client but can easily be rectified. The analysis of variance table has COVARIATES in the plural, even for only one *x*–variate; this adds to the difficulties for someone struggling to understand what is going on but this too can easily be changed. Finally, there is the fact that the printed sums of squares for treatment terms are calculated after fitting the covariates, and there is the user's possible uncertainty about how the printed sum of squares for a covariate was calculated. Should explanations of these things not be printed? But this leads me to the last of my five points.

Non-orthogonal analyses

Ever since electronic computers were first used for statistical computing, the research world has been flooded with output from multiple regression analyses and many of these analyses have been ill-motivated or nonsensical. Amongst the many problems associated with all this has been a crucial lack of understanding of non-orthogonality and its implications. Program- and package-writers must take much of the blame for this, in that many of them have produced programs whose inadequate output the user was entitled to believe that he could understand, even though he did not understand it. Here I am referring primarily to analysis of variance tables with sums of squares (and mean squares) printed for several fitted terms. For some such tables, each non-error sum of squares is calculated after **all** the other terms have been fitted; in other such tables, each non-error sum of squares is calculated after all the **preceding** terms have been fitted but before fitting succeeding terms. There is abundant evidence of the distinction between these two possibilities not being understood by many users of multiple regression. In particular, a recent issue of an agricultural research journal had three papers which contained detailed analysis of variance tables for multiple regression analyses; none of the three articles gave a wholly clear account of what was going on and the labelling and description of all the analyses of variance were seriously inadequate. Faced with this sort of thing, I conclude that any computer program providing multiple regression analyses should henceforth automatically print such messages as 'EACH SUM OF SQURES CALCULATED AFTER FITTING ALL TERMS ABOVE IT' with all analysis of variance tables to which they apply.

My comments in this paper have, of course, included what I have intended to be constructive and considered criticisms of a few details of current Genstat output. These criticisms are as nothing compared with what I might have said about various other statistical programs and packages which are, I understand, widely available. Genstat, used sensibly by people who have mastered its philosophy and syntax, has proved itself to be an excellent tool for providing the sorts of analysis needed by clients such as mine. It is because Genstat analysis-of-variance output so nearly satisfies my self-sufficiency criterion that I cannot resist asking for the new modifications that I have mentioned.

Acknowledgement

This paper was written and delivered whilst the author held the post of Biometrics Liaison Officer (Field Crops) of the Overseas Development Administraton of the United Kingdom.

References

Bingham, C. and Fienberg, S.E. (1982) Textbook analysis of covariance – Is it correct? Biometrics, **38**, 747–753.

Cochran, W.G. and Cox, G.M. (1957) Experimental Designs, 2nd Edition. New York: Wiley.

Little, T.M. and Hills, F.J. (1978) Agricultural Experimentation: Design and Analysis. New York: Wiley

Preece, D.A. (1980) Covariance analysis, factorial experiments and marginality. Statistician, **29**, 97–122.

Modelling in Genstat

Bertus Keen
IWIS-TNO
Postbus 100
Wageningen
The Netherlands

Introduction

Problem solving by the scientific method involves the collection of data and modelling, applying statistical methods to discriminate between models, to estimate parameters and to test hypotheses concerning parameters of selected models. Due to developments in statistical theory, a more unified approach to modelling has been achieved and, due to the development of statistical program packages like Genstat, calculations can be carried out for a wide range of models without expert knowledge of numerical procedures. Consequently, more attention can be paid to the experimentation and modelling part of problem solving. In my view, the ideal situation would be to have a minimum number of restrictions in experimentation and in modelling, so that choices concerning design and modelling can be made on the basis of efficiency arguments only and calculations are carried out automatically after specification of the model and the required inference. Genstat seems to be designed to facilitate the calculations in this way. But does the ideal situation exist already, or can the range of models be extended and/or the specification be simplified?

The facilities in Genstat 4.04A for specifying and fitting univariate models of the form $y = f(x) + \varepsilon$ are:

f(x)	ε	Genstat structures allowed	Genstat directives	Type of model
Linear	normally distributed, one or more homogeneous strata	`FACTORs` and `(CO)-` `VARIATEs`	'BLOC' 'TREA' 'COVA' 'ANOV'	ANOVA-model, multiple strata model, variance components model
linear and generalised linear	normal, Poisson, binomial, gamma, inverse normal distribution; one stratum	`FACTORs` and `VARIATEs`	'TERM' 'Y/LINK=..., ERROR=..., 'FIT' etc.	regression model
non-linear	normal, Poisson, binomial, gamma, multinomial distribution; one stratum	`VARIATEs`	'MODE' 'OPTI'	regression model

For ANOVA-models, it will be shown that model specification in non-trivial cases is not always simple but that a minor change in the meaning of specifications will improve it considerably. A generalisation of ANOVA-models, including model specification, will be discussed. Moreover, some syntax and structure definitions will be discussed.

ANOVA–models with Homogeneous Strata

Multi-stratum models are of the form:

$$y = \textit{fixed effects} + \varepsilon_1 + \varepsilon_2 + \dots \tag{1}$$

$\varepsilon_1, \varepsilon_2, \dots$ are random terms, connected to the levels of qualitative factors. Each represents a stratum. The strata are called homogeneous if the εs are independently distributed with equal variances within each of the strata. Fixed terms within a stratum are tested against the residual variance of that stratum. F-tests are exact if the distribution of the random components is normal.

Consider as an example the successive measurement design, with *32* individuals, for whom a characteristic *Y* has been measured at *8* successive time points. The individuals are randomly assigned to *4* treatments, *8* individuals for each of the treatments. The treatments are the levels of a quantitative factor. The Genstat code for specifying the structures involved is:

```
'UNIT' $ 256
'VARI' LEVELS=......
'FACT' TIMES, TIMES2 $ 8 : INDIV $ 32 : TREATM $ LEVELS=64(1...4)
'GENE' INDIV, TIMES
'RUN'
'VALU' TIMES2= TIMES
```

Suppose the model for the observation for treatment *i* at individual *j* at time point τ_k is:

$$Y_{ijk} = \mu + v_i + \varepsilon_{ij} + \beta_{1ij}\tau_k + \beta_{2ij}\tau_k^2 + \varepsilon_k + \varepsilon_{ijk} \tag{2}$$

where μ = the general mean,

v_i = the effect of treatment i, possibly a low degree polynomial in the levels of the treatment factor,

β_{1ij} and β_{2ij} are the regression coefficients of the polynomial in time for individual ij,

ε_{ij}, ε_k and ε_{ijk} are independently distributed random components.

The required ANOVA-table		Deviations of Genstat analysis
INDIV STRATUM		
TREATM	3	
LIN	1	
QUAD	1	
DEV	1	"CUB" instead of "DEV"
RESIDUAL	28	
TOTAL	31	
TIMES STRATUM		"TIMES2" instead of "TIMES"
TIMES	2	
LIN	1	
QUAD	1	extra line
RESIDUAL	5	"ASSIGNED TO ERROR" 5
TOTAL	7	
UNITS STRATUM		"INDIV.TIMES2 STRATUM"
TREATM.TIMES	6	instead of "UNITS STRATUM"
LIN.LIN	1	
QUAD.LIN	1	
DEV.LIN	1	"CUB" instead of " DEV"
LIN.QUAD	1	
QUAD.QUAD	1	
DEV.QUAD	1	"CUB" instead of "DEV"
RESIDUAL	211	extra line: "ASSIGNED TO ERROR" 15
TOTAL	217	

The Genstat code which I found to accomplish the best approximation to the required analysis is:

```
'BLOC'  INDIV +TIMES2
'TREA'  POL( TREATM,3) * POLND( TIMES, 2)
'ANOV/ LIMA=350'
```

The deviations from the required analysis, as indicated alongside the ANOVA-table, are not critical, so the model can indeed be analysed by Genstat. However, it is not a straightforward task to trace the necessary Genstat code. The special difficulties are outlined in the code given above. The following changes in the meaning of directives, options and functions would simplify the model specification appreciably:

(1) 'BLOC' defines the random effects (the strata) and 'TREA' the fixed effects without overruling the effects in 'BLOC' by the effects in 'TREA'.

(2) LIMA-options of 'BLOC' and 'TREA' are introduced and the LIMA-option of 'ANOV' is omitted.

(3) POL(TIMES, 2) means: a second degree polynomial is specified, but it may not be the complete effect of the factor TIMES,

POLND(TIMES) means: a second polynomial describes the effect of the factor TIMES completely.

(4) The third digit of the LIMA-option is redundant.

The specification of model (2) would then be:

```
'BLOC' INDIV + TIMES
'TREA' POL( TREATM ,2) * POLND( TIMES, 2)
'ANOV'
```

ANOVA-Models with Inhomogeneous Strata

If in model (1) variances within a stratum are not constant and/or covariances are not zero then that stratum is called inhomogeneous. Fixed effects within such a stratum should be tested against their own variance estimate. *t*–tests are exact if an independent estimate of the variance exists, *F*-tests usually are not exact and should in the general case be replaced by Hotelling's T^2. An example of inhomogeneous strata is a block experiment when interactions between treatments and blocks exist. Every treatment contrast should then be tested against its own interaction with blocks. A possible model formulation for such an experiment is:

```
'BLOC' BLOCKS + INHOM( TREATM * BLOCKS)
'TREA' POL( TREATM, 2)
```

with the resulting ANOVA-table possibly arranged as:

```
BLOCKS STRATUM

TREATM.BLOCKS INHOMOGENEOUS STRATUM
   TREATM
   RESIDUAL

     LIN.BLOCKS STRATUM
       LIN
       RESIDUAL
     TOTAL

     QUAD.BLOCKS STRATUM
       QUAD
       RESIDUAL
     TOTAL

     DEV.BLOCKS STRATUM
       DEV
       RESIDUAL
     TOTAL

TOTAL
```

Another example of an inhomogeneous stratum is found in the successive measurement design discussed before. Usually, the effects with time are such that interactions between individuals and time exist (for instance, the effect with time may be linear, with different slopes for different individuals). Suppose we have model (2) with inhomogeneous INDIV.TIMES stratum. Consider the model specification:

```
'BLOC' INDIV + TIMES + INHOM( INDIV. TIMES)
'TREA' POL( TREATM, 2) * POLND( TIMES, 2)
```

The resulting ANOVA-table may then be:

```
INDIV STRATUM
   TREATM              3
     LIN               1
     QUAD              1
     DEV               1
   RESIDUAL           28
TOTAL                 31

TIMES STRATUM
   TIMES               2
     LIN               1
     QUAD              1
   RESIDUAL            5
TOTAL                  7

INDIV.TIMES INHOMOGENEOUS STRATUM
   INDIV.POL(TIMES) STRATUM
     TREATM.TIMES      6
     RESIDUAL         56
   TOTAL              62

   INDIV.LIN STRATUM
     TREATM.LIN        1
       LIN.LIN         1
       QUAD.LIN        1
       DEV.LIN         1
     RESIDUAL         28
   TOTAL              31

   INDIV.QUAD STRATUM
     TREATM.QUAD       3
       LIN.QUAD        1
       QUAD.QUAD       1
       DEV.QUAD        1
     RESIDUAL         28
   TOTAL              31

   RESIDUAL          155

TOTAL                217
```

A problem is to specify the model unambiguously. In the examples above this is achieved because the inhomogeneous stratum is the interaction between a fixed and a random factor. In other situations the analysis is not a straightforward generalisation of ordinary analysis of variance.

On Uniform Model Specification

Suppose A, B, C and D are SCALARs,
X1, X2 and X3 are VARIATEs and
F1 and F2 are FACTORs.

A few suggestions for improving uniform model specification for the different types of models within Genstat are:

(a) Specification of the random part and the fixed part of the model and the fitting procedure by separate directives, e.g. 'STRATA', 'EXPECT' and 'FIT' respectively. Changes in the fixed part of the model may be indicated by 'ADD', 'DROP' etc.

(b) The introduction of interactions between VARIATEs in regression models, allowing the specification of product terms, e.g.

```
X1 + X2. X3
```

This means that VARIATEs are considered as first degree polynomials of quantitative factors with many levels.

(c) The ability to specify polynomial models for VARIATESs, e.g.:

```
F1 + POL( F2, 1) + X1 * POLND( X2, 2)
```

(d) The ability to use FACTORs instead of SCALARs as parameters in non-linear models, e.g.

```
C/ (A + EXP( F1 + B * X1))
```

indicating that the parameter F1 depends on the levels of the FACTOR F1. 'DROP' F1 would then reformulate the model to:

```
C/ (A + EXP(D + B * X1))
```

(e) Suggested new vector structures to replace INTEGERs, VARIATEs and FACTORs are: CONTINUOUS, DISCRETE, ORDINAL and NOMINAL. These new structures are more closely related to the real world problems to be solved as they specify population characteristics to be described as well as characteristics of the outside world (factors) in a way which is relevant to the required inference.

The qualitative structures NOMINAL and ORDINAL must have levels specified, the quantitative structures may have levels specified, so all structures can be used as FACTORs. The quantitative structures are treated as VARIATEs, unless indicated otherwise, e.g. by LEV(X1). The qualitative structures are treated as FACTORs, unless indicated otherwise, e.g. by POL(F1, 1).

Poisson models for the analysis of road traffic accidents

Lars Krogsgard Thomsen
Danish Council of Road Safety Research
Ermelundsvej 101
DK – 2820 Gentoffe
Denmark

and

Poul Thyregod
Department of Mathematical Statistics
and Operations Research
Technical University of Denmark
Building 349
DK – 2800 Kgs
Lyngby
Denmark

Quantative data within road safety research and other social science applications often appear as counts.

Thus the analysis of cross-tables (contingency-, homogeneity- or multiplicative-tables) is an important tool. We shall not repeat the theory here but refer to Bishop, Fienberg and Holland (1975), Haberman (1978), Haberman (1979) and Anderson (1980).

This paper contains four examples with Poisson-models related to cross-tables. The first treats the ordinary unweighted case. The last three examples treat what could be called weighted tables, i.e. tables where the raw counts are studied in relation to an externally given structure. In these cases the counts are the accident counts and the external structure is given by the accident-generating traffic-flows.

Literature on the statistical theory of such models is much more sparse, though some guidance can be found in the papers by Andersen (1977), Svensson (1978, 1979a and 1979b), Thomsen (1980) and Thomsen and Thyregod (1981).

Example 1, the 'classical' analysis

This example is an analysis of a *2* by *2* table, of Poisson-counts where the two class-variables are the variable ROADUSER with levels BICYCLE-RIDER and MOPED-RIDER and the variable LIGHTCONDITIONS with the levels DAYLIGHT and DARK.

The number of casualties in August and September is given in Table 1.

	CASUALTY		
LIGHTCON	DAYLIGHT	DARK	MARGIN
ROADUSER			
BICYCLE	369	97	466
MOPED	518	293	811
MARGIN	887	390	1277

Numbers of casualties among bicycle- and moped-riders in Denmark
in August and September 1975. The casualties are classified by category of roaduser
as well as by light conditions at the scene of the accident.

Table 1

A possible log-linear-model for this table is the main-effects-model

$$ln\, m_{ij} = C + R_i + L_j \qquad (1)$$

where

$$R_1 = L_1 = 0.$$

Example 2, the weighted table

The Danish National Institute of Social Research and the Danish Bureau of Statistics performed in 1975 about 5000 interviews with persons constituting a supposedly representative sample of the population in Denmark. From these interviews it is possible to learn something about driving patterns and mileage for Danish bicycle- and moped-riders.

In Table 2 we again give the casualties of Table 1, but additionally 'traffic-flows' are given for the four groups analysed.

```
ROADUSER    LIGHTCON    CASUALTY    KM
BICYCLE     DAYLIGHT    369         1.000
BICYCLE     DARK         97         0.101
MOPED       DAYLIGHT    518         0.436
MOPED       DARK        293         0.102


            CASUALTY
LIGHTCON    DAYLIGHT    DARK     MARGIN
ROADUSER
BICYCLE     369          97       466
MOPED       518         293       811

MARGIN      887         390      1277


            FREQUENC
LIGHTCON    DAYLIGHT    DARK
ROADUSER
BICYCLE      369.00      960.40
MOPED       1183.07     2872.55
```

Casualties as in Table 1 with the corresponding traffic flows
(scaled with kilometres driven by motorcycle riders in daylight as unit).
Also shown is the casualty-frequency as casualties per unit of traffic flow.

Table 2

The main-effects model in this case becomes

$$ln\, m_{ij} = C + R_i + L_j + t_{ij} \qquad (2)$$

where the symbols are as in (2); t_{ij} designates the traffic-flow corresponding to group (i,j).

Summary and Conclusion.

This paper suggests that the analysis of cross classified rates is very simple using the theory of log-linear models. The 'Genstat' program (Alvey et al., 1977) has proven very convenient for this purpose.

The analysed tables are very simple and we have not discussed the important question of choice of variables and levels for the analysis. Very detailed accident analysis demands consideration of many variables and levels and a more appropriate method might be that of Poisson regression and covariance-analysis.

Such analyses have been carried out, e.g. by Bui Quoc, Cambois and Lasarre (1981) with interesting results.

At the moment such models are used for the analysis of accidents in cities in Denmark. A total of about 40 variables including traffic-flow of 11 street-user-categories are included as explanatory variables. We will report on this on a latter occasion.

References

Alvey, N.G. et al. (1977) GENSTAT, A General Statistical Program, Rothamsted Experimental Station, October 1977.

Andersen, E.B. (1980) Discrete Statistical Models with Social Science Applications. North Holland.

Andersen, E.B. (1977) Multiplicative Poisson Models with Unequal Cell Rates. Scandanavian Journal of Statistics, 4, 153–158.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P. (1975) Discrete Multivariate Analysis: Theory and Practice. The MIT Press, London.

Bui Quoc, T., Cambois, M.A. and Lasarre, S. (1981) Etudes Statistique a Caractère Methodologique, phase unique. Organisme National de Sécurité Routiere, 95114 Arcueil Cedex, France.

Edwards, D. and Kreiner, S. (1982) Some Aspects of the Analysis of Large Contingency Tables in Practice. In Nordic Symposium in Applied Statistics and Data Processing, (Hoskuldsson et al eds.), NEUCC, Lyngby, Denmark.

Haberman, S.J. (1978) Analysis of Qualitative Data, Vol. 1. Introductory Topics. Academic Press, London.

Haberman, S.J. (1979) Analysis of Qualitative Date, Vol. 2. New Developments. Academic Press, London.

Knoflacher, H. and Kern, U. (1979) Zusammenhang zwischen stündlicher Verkehrsmenge und Unfallhäufigheit. Kuratorium für Verkehrssicherheit, Kleine Fachbuchsreihe 14, Wien.

Svensson, Å. (1978) A Conditional Limit Theorem. Inst. för Försäkringsmatematik och Matematisk Statistik, Research Report No. 107, University of Stockholm.

Svensson, Å. (1979 a) On a Goodness-of-Fit Test for Multiplicative Poisson Models. Inst. för Försäkringsmatematik och Matematisk Statistik, Research Report No. 115, University of Stockholm.

Svensson, Å. (1979 b) On a Class of Multivariate Models with Linear Structures – A Goodness-of-Fit Test and Estimate. Inst. for Försäkringsmatematik och Matematisk Statistik, Research Report No. 116, University of Stockholm.

Thomsen, L.K. (1980) Statistisk analyse af faerdselsulykker. IMSOR, Danmarks Tekniske Højskole, Lyngby, Denmark.

Thomsen, L.K. and Thyregod, P. (1981) Unweighted and Weighted Poisson-Models for Discrete Data. In Symposium i Anvendt Statistik (Symposium in Applied Statistics), (Hoskuldsson et al eds.), NEUCC, Lyngby, Denmark.

Genstat Conference: 1985

The 1985 Conference will take place at the University of York from 22 to 26 September, 1985. Registration will be available from 3 p.m. on Sunday 22 September (or from 9 a.m. Monday morning for latecomers) and the conference will end mid-afternoon on Thursday 26 September.

The scientific program will include:

- Description of New Genstat Facilities
- Interesting Applications of Genstat
- Use of Genstat in the teaching of statistics
- Lectures explaining the statistical methodology in Genstat
- Descriptions of the use of Macros
- Demonstrations of the latest Genstat release.

Contributions will be selected on the basis of abstracts, which must be submitted to:

R W Payne
Statistics Department
Rothamsted Experimental Station
Harpenden
Hertfordshire AL5 2JQ
United Kingdom

by 1 March 1985.

There will be a full social program, with a Reception on Sunday evening, a Conference Dinner on Monday evening and various scenic and cultural excursions on Tuesday afternoon.

For further information, including detailed costs and registration form when available, please complete and return the form below.

To: R W Payne, Statistics Department, Rothamsted Experimental Station, Harpended, Hertfordshire, AL5 2JQ, United Kingdom.

Please send details of the Genstat Conference 1985 to:

Name: ______________________________

Address: ______________________________

____

____

____

____

____

I intend to submit a paper/poster (abstract by 1 February 1985) entitled

____

____

____

GENSTAT NEWSLETTER ORDER FORM

To order future issues of the Genstat Newsletter, please complete the form below and return it to:

The Genstat Co-ordinator
NAG Central Office
Mayfield House
256 Banbury Road
OXFORD OX2 7DE
United Kingdom

(Each Genstat site representative *automatically* receives one copy of each issue, free of charge.)

Please note that each subscription to the Newsletter costs £ 5.00 per annum (2 issues). This price includes 2nd class/surface postage. Postage at other rates will be charged at cost.

Back issues of the Newsletter are available on microfiche (24×). The first contains issues 1 – 6 and each subsequent fiche contains two issues – 7/8, 9/10 etc. Each fiche costs the same as a year's subscription to the Newletter (£5.00).

To: NAG Ltd., Mayfield House, 256 Banbury Road, OXFORD OX2 7DE, U.K.

Please supply me with ..... copies of each future issue of the Genstat Newsletter

*for ..... years/until further notice beginning with issue number ..... .

(Minimum subscription period: 2 years)

Please supply me with ..... microfiche of each of the following issues ....

□ Enclosed is my remittance for ....

# □ Please invoice me. (An invoice will be sent immediately)

Signature ______________________

Name and address for posting ______________________

(please type or print) ______________________

____

____

____

Special mailing instructions ______________________

Cheques to be made payable to the Numerical Algorithms Group Ltd.

** delete one alternative*

# tick one box