Issue No. 21

1988(A)

The GENSTAT Newsletter

NAG NUMERICAL ALGORITHMS GROUP

Editors

P W Lane
Rothamsted Experimental Station
Harpenden
Hertfordshire
United Kingdom AL5 2JQ

K I Trinder
NAG Central Office
Mayfield House
256 Banbury Road
Oxford
United Kingdom OX2 7DE

Printed and produced by the Numerical Algorithms Group

©The Numerical Algorithms Group Limited 1988
All rights reserved.

NAG is a registered trademark of The Numerical Algorithms Group Ltd

ISSN 0269-0764

The views expressed in contributed articles are not necessarily those of the publishers.

GENSTAT NEWSLETTER
Issue No. 21

NP1711

1988 March

Contents

Published Twice Yearly by
Rothamsted Experimental Station Statistics Department
and the Numerical Algorithms Group Ltd

Editorial

This issue contains more articles from the Fifth Genstat Conference. There may be further articles from the Conference still to come, but we hope to see new material as well: perhaps you have an item that would be of interest to other Genstat users? You do not have to write a long article in the style of the Conference-based articles in order to contribute. The editors are just as keen to see short pieces about any aspect of Genstat: particular applications, practical details of individual commands, comparisons of Genstat with other systems or between versions of Genstat. Following the redesign of the command language, we do not expect that there will be many articles about the special tricks needed to get Genstat to do some operations, as appeared in early issues of the Newsletter. Nevertheless, there are bound to be useful combinations of options, parameters or statements that are worth bringing to the attention of other users, particularly in the context of writing procedures.

The issue concludes with a complete index of all Genstat Newsletter articles from Issue No. 1 onward, with articles classified by author and subject. All macros and procedures which have been described in articles are also listed. It is hoped that this will provide a useful reference guide and it will be updated periodically for future issues of the Newsletter. We would like to thank Sue Welham for bringing this material together.

Implementations of Genstat 5 are steadily being released; the systems on which Genstat 5 is now available are VAX VMS, Prime and IBM compatible PC's, with a number of other implementations close to release. Demand for the IBM compatible PC version, perhaps not surprisingly, is already proving to be considerable, even with little opportunity for promotion, as yet. The system requirements are 640 Kb memory (of which Genstat uses about 582 Kb), a mathematical coprocessor, a hard disk with about 4 Mb of free space and PC-DOS or MS-DOS 3.1 or higher. Regrettably, it is impractical to provide the Fortran extension facilities in the PC version and the high resolution graphics facility has also been omitted at this stage, although this will be incorporated as soon as possible.

In addition to the implementation becoming available, a much extended Genstat Procedure Library has been formed. The new version of the Library has 51 procedures, covering many of the ares in the Genstat 4 Macro Library.

In order to improve the Genstat service for users, one-day conferences and courses have been introduced. The first of the one-day conferences was on 28th April 1988 at Rothamsted. The subject was 'Extending Genstat in Fortran' and it was very well attended. The next will be on 6th October 1988, again at Rothamsted, and the topic will be 'The Analysis of Repeated Measures in Genstat'. Further information can be obtained from Roger Payne at Rothamsted or Keith Trinder at NAG. Any suggestions of topics for future meetings are welcomed. Note that the next full Genstat Conference will be in Edinburgh, from 11th-15th September 1989. Also, two types of training course are being run; a three day 'Introduction to Genstat 5' course and a two day 'Conversion from Genstat 4 to Genstat 5' course. Each of these have been run once and the next will be an introductory course which it is intended will be held in December 1988 at Birmingham (U.K.); further details will be sent to sites when available. Tailored on-site courses can also be arranged, both in the United Kingdom and elsewhere; please contact NAG for more information.

Finally, Oxford University Press report that the Genstat 5 Reference Manual has sold out ifs first printing and has been reprinted (with corrections) in paperback, price £22.50, as well as in hardback. 'Genstat 5: an Introduction' is also available, and 'Genstat 5: a Second Course' is in the last stages of production.

The Use of Genstat in Estimating the Expected Numbers of Cases of AIDS Adjusted for Reporting Delays

S H Heisterkamp
J C Jager

Centre for Mathematical Methods
National Institute of Public Health and Environmental Protection
Bilthoven
The Netherlands

J A M van Druten
Dept. of Mathematical Statistical Consultation
University of Nijmegen
Nijmegen
The Netherlands

A M Downs
WHO Collaborating Centre on AIDS
Paris
France

and

Unite de Recherches Biomathematiques et Biostatistiques
(INSERM U 263)
Universite Paris 7
Paris
France

This paper is the result of a collaboration between The National Institute of Public Health and Environmental Protection, the WHO Collaborating Centre on AIDS and the University of Nijmegen. It was made possible by a grant from the European Economic Community.

1. Introduction

This paper is mainly concerned with the use of Genstat in the estimation of numbers of AIDS cases, but first we have to clarify the ideas of the estimation procedure we propose. The problem with data from the AIDS epidemic, and probably with all data concerning new epidemics which are of public interest, is that they are rarely up-to-date. However, decision makers want to have the data as soon as they are available. They need the data for three things: to describe the past, to describe the present status of the epidemic, and even worse, to predict the future.

One of the main problems is the fact that there is a delay between the date of reporting and the date of diagnosis of the AIDS cases. This is usually the case with epidemiological data, but as decision makers want to act on the most recent data, it is in this case worthwhile to try to overcome this problem. Data received from various sources in twelve European countries sometimes show delays of two or three years. In a study by Brookmeyer *et al.* [3] this problem was solved by deletion of the last reported cases. Downs *et al.* [5,6] and Morgan and Curran [14] adjusted the incidence for reporting delays, though without stating an explicit statistical model. As we reported earlier [9], it is possible to formulate an explicit model which is flexible enough to make use of existing epidemiological models.

Our first aim is to give – as accurately as possible – a description of the past and present numbers of cases, and secondly to provide short-term predictions by means of extrapolation of a past trend. Description of the epidemic is also needed for the development and evaluation of mathematical-epidemiological models, which is done in collaboration with the University of

Nijmegen. It is clear that long-term prediction can only be achieved by means of sound epidemiological models. However, as long as there are many uncertainties about the underlying mechanism of the epidemic and the parameters involved, short-term predictions by means of extrapolation of fitted curves are necessary, but should be treated with caution.

2. The Data

In what follows we explain how we make use of the reported cases, classified according to both period of diagnosis and period of first reporting, in order to estimate the essentially *unobserved* numbers of AIDS cases, the so-called 'adjusted numbers'. Once this is done, one could proceed by fitting some curve to those adjusted cases, ignoring the statistical dependency. However, making use of the OPTIMIZE directive in Genstat 4.04, we directly estimated the appropriate trend parameters from our original model, and in doing so, we circumvented the need to estimate the adjusted numbers of cases. We fitted the exponential model and found that it still fits very well. The data seem to indicate that the three-parameter logistic model is, although theoretically better, as yet somewhat premature.

Tables 1(a)–1(c) display data for The Netherlands, the United Kingdom and Italy respectively. Each row represents the reporting period of six months, and each column the period of diagnosis. The figures represent the numbers of cases being reported for the first time for a certain period of diagnosis, with the exception of the first row of each table. Reporting to the WHO collaborating centre started in the first six months of 1984 and the exact reporting period of data before this time was unknown (in Paris). This was another problem to solve.

		Diagnosis Period										
		1	2	3	4	5	6	7	8	9	10	11
	5	1	2	6	7	5						
	6	0	0	0	1	1	19					
	7	0	0	0	1	0	4	19				
Reporting	8	0	0	0	0	0	0	6	26			
Period	9	0	0	0	0	0	1	3	4	39		
	10	0	0	0	0	0	0	0	0	11	62	
	11	0	0	0	0	1	0	0	1	0	13	75
	Total	1	2	6	9	7	24	28	31	50	75	75

Table 1(a)

Cases in The Netherlands, reported by June 30 1987

		Diagnosis Period										
		1	2	3	4	5	6	7	8	9	10	11
Reporting Period	9	3	2	8	8	18	29	31	43	20		
	10	0	0	1	1	3	0	2	19	39	32	
	11	0	0	0	1	2	0	3	2	17	23	37
	Total	3	2	9	10	23	29	36	64	76	55	37

Table 1(b)

Cases in The United Kingdom, reported from June 1985 until June 1986

		Diagnosis Period										
		1	2	3	4	5	6	7	8	9	10	11
Reporting Period	5	0	0	2	2	3						
	6	0	0	0	0	1	6					
	7	1	0	0	0	0	5	32				
	8	0	0	0	0	2	1	29	56			
	9	0	0	0	0	2	1	6	41	110		
	10	0	0	0	0	0	2	4	0	48	169	
	11	0	0	0	0	0	0	3	3	14	54	273
	Total	1	0	2	2	8	15	74	97	172	223	273

Table 1(c)

Cases in Italy, reported from June 1984 until June 1987

3. Basic Model

In this section we describe the estimators for the expected numbers of AIDS cases from reported incidence without assumptions about a trend in time. The problem addressed here is related to estimation of the number of missing values in a multinomial distribution [2] or estimating the number of unseen species [8]. The difference is that we have such a missing-value problem for each period of diagnosis. Rather than looking at the total reported incidences, thus disregarding the cases not yet reported, we find it more useful to look at the patterns of how cases are reported in the past. The easiest way is to think about the reporting delay as constant over time, e.g. 75% of the 'real' cases are reported without any delay, 20% are reported with a delay of six months, and so on. This is in fact what we assumed, and something more as will be explained below.

We consider an observation n_{ij} as the number of cases first reported at period i for the period of diagnosis j. We assume that for each period of diagnosis the total numbers of AIDS cases, denoted by $N_{.1}, N_{.2}, \ldots N_{.t}$, are fixed, but not observable. We will refer to them throughout the paper as the expected incidence of AIDS, as opposed to the reported incidence of AIDS, which we denote as $n_{.j}$, for each period of diagnosis j, up to the given reporting time t. Our basic assumptions will be that in principle all the cases could be known, i.e. we define our sources as the ultimate, although incomplete, truth. We also assume that all reported cases are correctly diagnosed, i.e. a false diagnosis or a recording error with respect to time period is supposed to be

non-existent. Furthermore we assume that reporting proportions are the same, given a similar delay between date of report and date of diagnosis. However, this assumption is not really essential for the model, although it would slightly complicate the manner of estimation. One could argue, for example, that directives from decision makers could change the delay time between diagnosis and reporting negatively or positively, depending on the effect of the measures. In principle it is not impossible to adapt the estimation procedure for such changes. As for the probabilistic assumptions, we assume $n_{ij} \sim \text{Poisson}(\theta_{ij})$ and independent. According to our basic assumptions, which will be modified later on, the parameters θ_{ij} have the following relations to each other:

$$\theta_{ij} = N_{ij}p_k \text{ with } k = i - j, \quad j \le i \le t, \quad 1 \le j \le t, \quad i \ge t_0 \quad p_k > 0 \tag{1}$$

We also assumed that after a fixed period of time, T, no new cases would be reported. So:

$$\sum_{k=0}^{T} p_k = 1 \qquad \text{fixed } T \tag{2}$$

Note that at the start of the reporting, t_0, only the reported incidences are available for each period of diagnosis. We therefore assumed that the delay proportions remained unchanged in the period before t_0, and moreover were the same as after t_0. According to the original assumptions the incidence at t_0, $m_{.1}, m_{.2} \ldots m_{.t_0-1}$ are now independently Poisson distributed with parameters given in (3).

$$E(m_{.l}) = N_{.l} \sum_{k=0}^{t_0-1} p_k, \qquad l = 1,2,\ldots,t_0-1 \tag{3}$$

In Table 2 the expected numbers of reported cases for an arbitrarily chosen range of time are given, together with the expected numbers of AIDS cases, in which we are primarily interested.

		Period of Diagnosis					
		83b	84a	84b	85a	85b	86a
Reporting time	84a	$\Sigma p_k N_{.6}$	$p_0 N_{.7}$				
	84b	$p_2 N_{.6}$	$p_1 N_{.7}$	$p_0 N_{.8}$			
	85a	$p_3 N_{.6}$	$p_2 N_{.7}$	$p_1 N_{.8}$	$p_0 N_{.9}$		
	85b	$p_4 N_{.6}$	$p_3 N_{.7}$	$p_2 N_{.8}$	$p_1 N_{.9}$	$p_0 N_{.10}$	
	86a	$p_5 N_{.6}$	$p_4 N_{.7}$	$p_3 N_{.8}$	$p_2 N_{.9}$	$p_1 N_{.10}$	$p_0 N_{.11}$
Expected incidence		$N_{.6}$	$N_{.7}$	$N_{.8}$	$N_{.9}$	$N_{.10}$	$N_{.11}$

Table 2

Expected numbers of first reported cases in the basic model
(Expected incidence is not observed)

Although it is not necessary for Genstat that the actual solutions of the maximum likelihood equations are known, it would of course be much better to have them, and it could be of great help to supply starting values to OPTIMIZE. The solution of the equations seems somewhat tiresome but is in fact very straightforward, if one changes the order in which the equations are to be solved.

First compute the $N_{.1} \ldots N_{.t-T}$ expected cases (5), which are simply the reported ones. After that, compute p_T with (6), and $N_{.t-T+1}$ by noting that the sum in the denominator in (4) is simply one minus p_T, and proceed with p_{T-1} (6). Cycle through these steps and note that the denominator in

(4) is just one minus the sum of the proportions p_k computed so far. Finally, compute p_0 as one minus the sum of the reporting proportions.

$$\hat{N}_{.j} = \frac{\sum_{i=i_0}^{t} n_{ij}}{\sum_{k=0}^{t-j} \hat{p}_k} \qquad j = t-T+1,\ldots,t \tag{4}$$

$i_0 = \max(j,t_0)$ while $m_{.j} \equiv n_{i_0 j}$ $j \le t_0$ for $t - T > 0$

$$\hat{N}_{.j} = \sum_{i=i_0}^{t} n_{ij} \qquad j = 1 \ \ldots \ t-T \tag{5}$$

$$\hat{P}_k = \frac{\sum_{l=l_0}^{t-k} n_{l+k,l}}{\sum_{l=t-T+1}^{t-k} \hat{N}_{.l} - \sum_{j=t_0-T+1}^{t_{min}} \frac{m_j}{\sum_{i=0}^{t_0-j} \hat{p}_i} + \sum_{i=t_{max}}^{t} \frac{n_{i,i-T}}{\hat{p}_T}} \tag{6}$$

$k = 0,1,2\ldots T-1 \qquad l_0 = \max(1,t_0-k+1) \qquad$ if $k = 0 \quad l_0 = t_0$

$t_{min} = \min(t_0-k,t_0-1) \quad t_{max} = \max(t_0+1,T+1)$

4. Implementation of the Basic Model in Genstat 4

Although the basic model could easily be programmed with the linear regression directives, we chose to use the OPTIMIZE directive, mainly because we were thinking about computing a nonlinear model for the expected number of AIDS cases as well. We discovered, by the way, that direct optimization, with the LIK=1 option, does not work with more than five parameters, in spite of the use of the option METHOD=GN.

The problem for Genstat is basically this: the MODEL directive (still referring to Genstat 4.04) only accepts relatively simple computations, and one is not allowed to use FOR loops, as would be convenient with factor structures. However, the solution is rather simple. Suppose the observations are in variate *y* of length *n*, $n = (t(t+1)-t_0(t_0-1))/2$. The expected numbers of reported cases are supposed to be in a variate mu of length *n*, which are calculated in the MODEL statement. The parameters to be estimated are scalar arrays: p(0...T), for the reporting proportions, and N(1...t) for the expected numbers of cases.

We now define some dummy variates each of length *n*. Declare variates nn(1...t), corresponding to the *t* columns of Table 2. Each individual element of variate nn(*j*), $j = 1...t$ contains the value 1 at the appropriate position referring to column *j*, and the value 0 otherwise. Possible lines of the MODEL statement are then:

```
'MODEL'  basic $ aux1(1...t) = nn(1...t)*N(1...t)
               $ mu1         = VSUM(aux1(1...t))
```

where aux1(1...t) and mu1 are also variates of length *n*. The last line compresses the *t* variates, containing the value 0 or parameter values N(*j*), in one variate, with the value of N(*j*) at the appropriate position, and FOR loops are not actually necessary.

A similar trick was carried out to transfer the values p(0...T). Define an array of variates np(0...T) of length *n*. Each element of the variate np(*k*) now corresponds with an entry in Table 2 parallel to the main diagonal (corresponding with np(0)). So each element of np(*k*) is 1 or 0 depending on whether a value of the parameter p_k is to be used or not. Note that for the positions of the elements of np(*k*) coinciding with the first row of Table 1, 1 also had to be put in, in order to get the summation right for cases reported before June 1984. As only *T*–1

parameters are free, we constrain T by 1 minus the sum of the remaining parameters. The corresponding lines in the MODEL statement are:

```
$ p(T)          = 1.0-SUM(VSUM(p(0...T_1)))
$ aux2(0...T)   = np(0...T)*p(0...T)
$ mu2           = VSUM(aux2(0...T))
$ mu            = mu1 * mu2
```

The last line calculates the expected numbers of reported cases which must be used in the actual OPTIMIZE statement, as in

```
'OPTIMIZE/LIK=4,METHOD=GN' MODEL=basic;
   PARAM=N(1...t),p(0...T_1); Y=y; Z=mu
```

where T_1 is short for T minus 1. Of course the last line could be, or rather should be, extended with nameable lists for residuals, variance covariance matrix and so on, and the option list should be extended with proper settings for printed output.

Implementation in Genstat turns out to be rather straightforward, the only real trouble in fact being with initialising the dummy variables to 1 and 0. This was done with nested FOR loops, and, as the dummy variables of the inner loop depend on the outer loop variables, this could only be achieved with the use of macros for the inner loops. Initialising the variables actually took most of the computer time: at least that is our impression.

In the foregoing lines of the MODEL statement, the values for p(0...T) were not constrained to lie between 0 and 1. Making use of the nameable lists in OPTIMIZE for upper and lower bounds for the parameters does not produce the wanted results: the program just terminates if the upper or lower bounds are reached by accident. We therefore used the (anti–)logit transformation for the parameters p(0...T), i.e. the actual parameters used in OPTIMIZE were the logits of p(0...T), with no constraints, and in the MODEL statement a back-transformation was performed in order to compute the expected values mu. For reasons of clarity those statements are not included here. As starting values for OPTIMIZE, we used the values computed according to the explicit equations above, which were also computed with Genstat. The same comment with regard to the use of loops for initialising applies for the computation of the starting values.

5. Model with Trend Parameters

Decision makers are interested in future developments, but it is not easy making predictions, especially those concerning the future! If a function can be used to predict future trends, it is in fact quite easy to plug this function into the maximum likelihood equations, i.e. to use Genstat to compute the relevant parameters. Instead of using the estimated expected numbers of AIDS cases to fit a curve, with all the complicated problems of correlation and unequal variance of the estimators which should be accounted for, it is much more elegant to use direct estimators of the parameters of the prediction function. Instead of the use of the t parameters N_j, we substitute the function $N_j = f(\theta,t_j)$, where $\theta = (\theta_1,\theta_2...\theta_r)$. This may result in a drastic reduction of the numbers of parameters. In fact we used the exponential model with parameters α and β, or alternatively α and γ, the latter being the doubling time. Solutions for the total system of equations are now no longer explicitly computable as they involve nonlinear equations. The model used was $f(t,\alpha,\beta) = \alpha\exp(\beta t)$ or $f(t,\alpha,\beta) = \alpha\exp((\log(2)/\gamma)t)$ if the doubling time is the parameter of interest. We also tried to fit the logistic model, which gave no success at all, as the epidemic is still in its first stages.

6. Implementation of the Nonlinear Model in Genstat

Considering the above lines of the MODEL statement, the implementation of a nonlinear model in Genstat is easy enough. For example, a model for the exponential curve can be made by inserting a new line before the first line of the basic model:

```
'MODEL' exponential $ N(1...t)    = a * EXP(b * time(1...t))
                    $ aux1(1...t) = ..... etc.
```

where `time(1...t)` is an array of scalars containing the time of report, and with `a` and `b` being the estimates for α and β. The `OPTIMIZE` statement is changed with regard to the model name and parameter list accordingly. Starting values for `a` and `b` were computed by ordinary regression of `log(N(1...t))` on `time`, and the estimates `N(1...t)` from the basic model. The calculations in the `MODEL` statement should protect against exponential overflow (error message: `invalid value of a function`), by restricting the exponent to the largest number for the Vax 750 (something like 80). These lines have been omitted here for reasons of simplicity.

7. Some Implementation Problems and Their Solutions

Although the implementation was rather straightforward, we encountered some problems in the actual process of computing the estimates for the reporting proportions, using the Genstat `OPTIMIZE` directive. Note that, if in some period of diagnosis no cases were reported – this occurs especially at the start of the epidemic – N_j will be estimated as zero by the maximum likelihood equations, the Poisson distribution being degenerate. For `OPTIMIZE` this is a problem, as the minimum of the deviance is then attained at the border of the parameter space, and at least one error occurs as the variance covariance matrix is not semi-positive. The same holds if for some 'delay' periods no cases are reported, i.e. if the sum of one or more of the observed diagonals in Table 2 is 0, which happens for example in the case of extremely long delays. This problem was overcome by using the `SET` directive to provide a reference to those parameters for each country that were not zero, subsequently used in the `OPTIMIZE` and `MODEL` statements. This, in our opinion, again demonstrates the flexibility of Genstat, as for each country both data and the program code concerning the `SET` directive for these parameters were read from one file.

8. Results and Discussion

In the above section we used a reporting model based on the rather rigid assumption concerning the time T after which no case could be reported. In actual data it seems that after some time a rare event of the report of one patient takes place, which is not related to the model of the reporting proportions. Insisting on that model will produce strange results, i.e. the expected numbers of cases are much inflated for the last periods as compared with data without the 'one' rare reported patient. The model does not seem very robust against this kind of aberration. A much more robust model is achieved if one allows for rare events after some delay time T–1 say. We introduced an extra parameter, h, which represents the proportion of cases to be reported after time T–1 which is the same for each period of diagnosis. The effect is that the numbers of cases to be reported after some time T–1 are spread over the whole lower triangle of Table 2. It actually means that in Genstat the values of 1 in dummy variable `np(T)` were replaced by weights $1/(t–j)$, and that of course also some values of 0 were replaced by those weights, in the appropriate places where new observations were allowed to occur.

Tables 3(a) and 3(b) show the results of fitting the basic model and the exponential model to data from The Netherlands.

Tables 4 and 5 show the results for the exponential model fitted to data from the United Kingdom and Italy. The data for the United Kingdom were not updated for June 1986, as it seemed that problems in the definition of AIDS resulted in a totally new set of data, i.e. t_0 is actually June 1987 for the U.K., and the model was not applicable for this new set, as only the top row of Table 2 was reported.

Reporting proportions

p	0.76	0.17	0.03	0.03	0.01
se	0.057	0.034	0.014	0.023	0.013

Expected Incidences with confidence intervals calculated by log transformation

N	1.0	2.0	6.0	9.0	7.0	24.0	28.0	31.4	52.2	80.5	98.9
se	0.52	0.93	2.15	3.19	2.15	4.71	5.15	5.68	7.43	9.46	13.45
upper	2.78	4.98	12.13	18.05	12.79	35.26	40.17	44.78	68.97	101.33	129.13
lower	0.36	0.80	2.97	4.49	3.83	16.33	19.52	22.03	39.45	63.90	75.79

Expected reported incidences from the basic model

	1	2	3	4	5	6	7	8	9	10	11
5	1.0	2.0	5.8	8.4	5.3						
6	0.0	0.0	0.2	0.2	1.2	18.2					
7	0.0	0.0	0.0	0.3	0.2	4.2	21.2				
8	0.0	0.0	0.0	0.0	0.2	0.6	4.9	23.8			
9	0.0	0.0	0.0	0.0	0.0	0.7	0.7	5.5	39.5		
10	0.0	0.0	0.0	0.0	0.0	0.2	0.8	0.8	9.1	61.0	
11	0.0	0.0	0.0	0.0	0.0	0.2	0.4	0.9	1.4	14.0	75.0

The deviance is 23.93 with 41 degrees of freedom and Pearson's Chi-square is 49.72

Table 3(a)

Results from Dutch data, fitting the basic model

Parameters

p	0.75	0.18	0.03	0.03	0.01
se	0.02	0.01	0.01	0.01	0.01

a	2.566
se	0.396

doubling time (months)	11.093
se	0.551

Expected reported incidences from the exponential model

	1	2	3	4	5	6	7	8	9	10	11
5	2.5	3.7	5.2	7.4	8.6						
6	0.0	0.0	0.2	0.2	2.1	12.6					
7	0.0	0.0	0.0	0.2	0.3	3.0	18.3				
8	0.0	0.0	0.0	0.0	0.3	0.4	4.4	26.6			
9	0.0	0.0	0.0	0.0	0.0	0.5	0.6	6.4	38.7		
10	0.0	0.0	0.0	0.0	0.0	0.1	0.7	0.9	9.2	56.3	
11	0.0	0.0	0.0	0.0	0.0	0.1	0.3	1.0	1.4	13.5	81.9

The deviance is 33.22 with 50 degrees of freedom and Pearson's Chi-square is 47.29

Table 3(b)

Results from Dutch data, fitting the exponential model

Parameters

p	0.26	0.34	0.25	0.04	0.11
se	0.03	0.03	0.03	0.02	0.03

a	4.374
se	0.737

doubling time (months)	11.453
se	0.754

Expected reported incidences from the exponential model

	1	2	3	4	5	6	7	8	9	10	11
9	4.2	6.1	8.6	12.3	17.3	23.8	32.7	32.9	20.4		
10	0.0	0.1	0.2	0.4	0.7	1.5	1.6	14.1	26.9	29.3	
11	0.0	0.1	0.2	0.4	0.7	1.5	4.4	2.3	20.3	38.7	42.1

The deviance is 41.63 with 24 degrees of freedom and Pearson's Chi-square is 43.21

Table 4

Results from United Kingdom data, fitting the exponential model

Parameters

p	0.54	0.27	0.06	0.05	0.07
se	0.00	0.04	0.02	0.01	0.02

a	1.347
se	0.253

doubling time (months)	6.936
se	0.260

Expected reported incidences from the exponential model

	1	2	3	4	5	6	7	8	9	10	11
5	1.3	2.3	3.9	6.6	8.1						
6	0.0	0.0	0.2	0.5	4.0	14.7					
7	0.0	0.0	0.0	0.4	0.9	7.4	26.7				
8	0.0	0.0	0.0	0.2	0.8	1.6	13.4	48.7			
9	0.0	0.0	0.0	0.2	0.4	1.4	2.9	24.4	88.7		
10	0.0	0.0	0.0	0.2	0.4	1.0	2.5	5.3	44.5	161.5	
11	0.0	0.0	0.0	0.2	0.4	1.0	3.7	4.5	9.6	81.1	294.3

The deviance is 108.46 with 50 degrees of freedom and Pearson's Chi-square is 162.39

Table 5

Results from Italian data, fitting the exponential model

We present here the results of the computations of the model with one extra parameter. Note that the degrees of freedom for the deviance now equal exactly the numbers of observations minus the numbers of parameters. This is not the case for the model where no observations are allowed after some delay time T; the numbers of cells of Table 2 that are 'structural zeros' must then be subtracted from the degrees of freedom. This must be done by 'hand': the OPTIMIZE directive naturally does not recognise this. The approximate confidence limits for the expectation of the number of cases are constructed with the delta-method; however, as we think it more appropriate to assume Normality on the log-scale, we constructed the limits on that scale, and transformed back. This prevented awkward things such as negative limits, and produced of course skew intervals. The limits are given for the expectations only, and not for new observations, which we think is less appropriate since what we observe are reported incidences. It is possible to construct approximate confidence limits for future reported incidence, which we hope to do in the near future. The exponential models are displayed in Figures 1(a) to (c).

One could argue that the model, complicated as it is, is still over-simplifying the real-world problem by considering the reporting as some independent Poisson process, and disregarding dependencies between the expected incidences. The model could be made somewhat more realistic, and still manageable, by assuming e.g. a gamma distribution for N_j. This is not entirely without foundation, as one could, for example, regard the expected incidence as a result from some mechanistic stochastic process, described by a stochastic logistic equation ([10]). It can be proved that the stationary distribution is a gamma distribution.

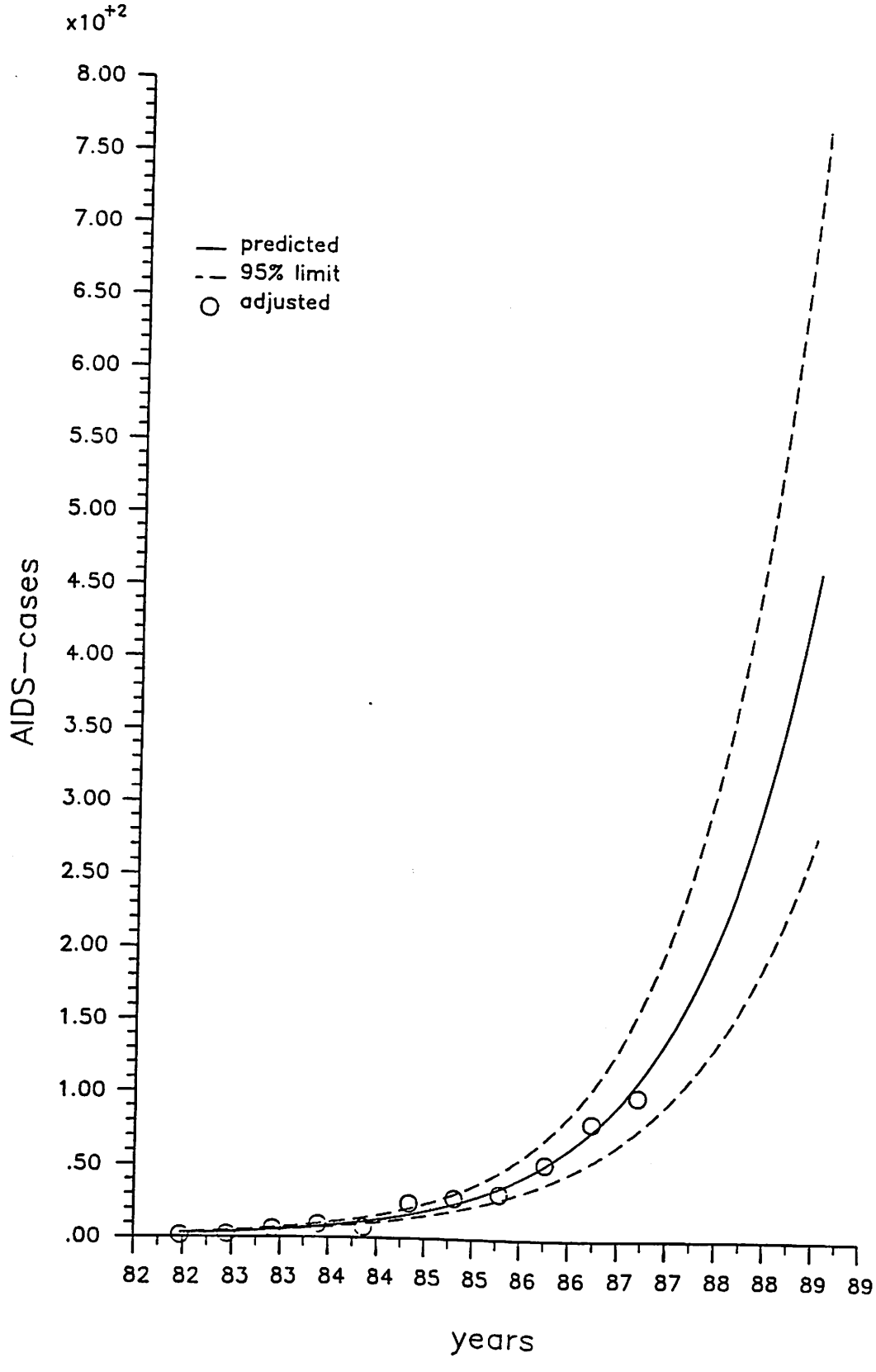

Figure 1(a)

Dutch data new cases reported by 30 June 1987
Exponential mode a 2.57 t 11.09 last 11 data

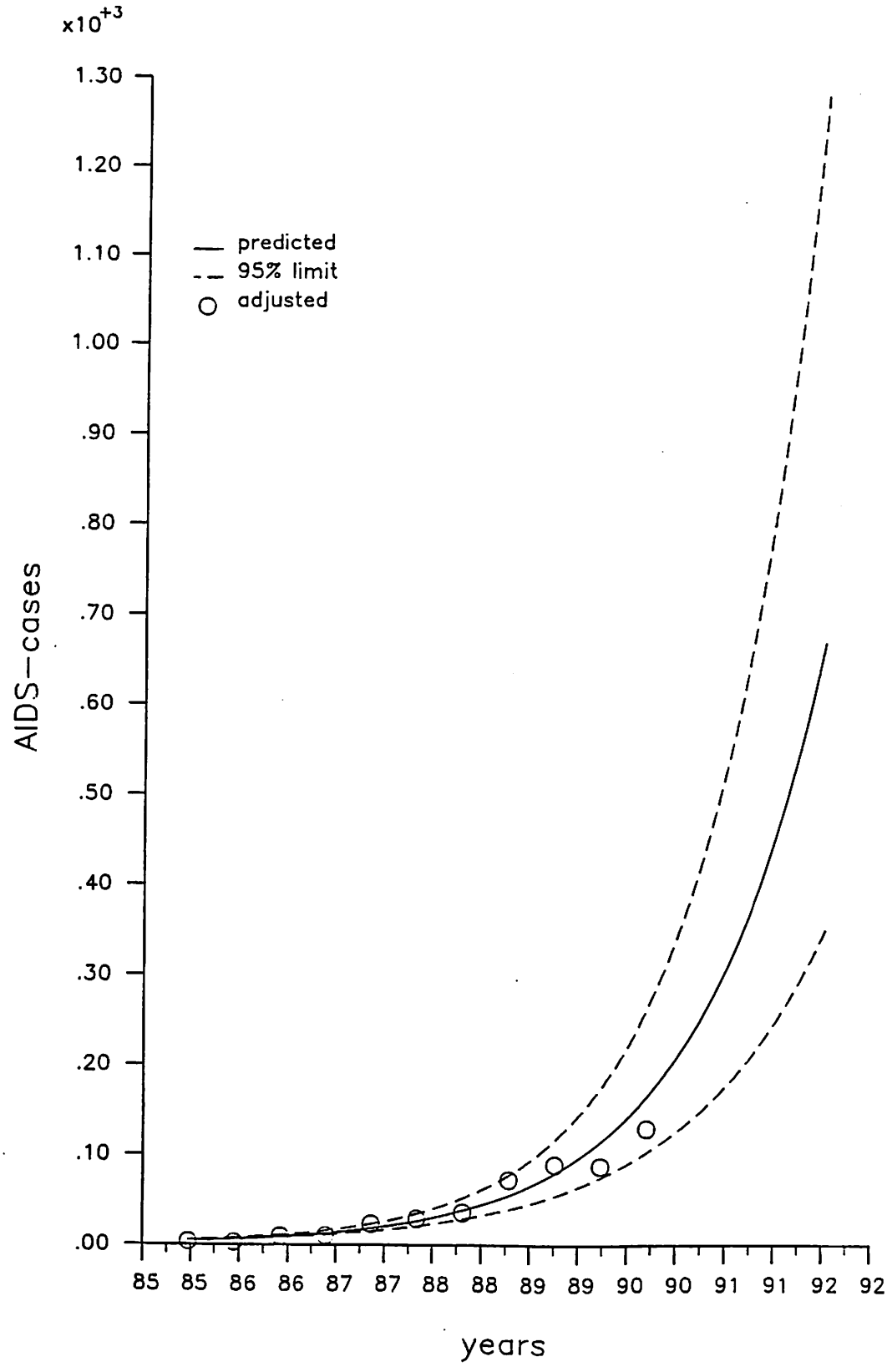

Figure 1(b)

United Kingdom new cases from 1985a onwards u
Exponential mode a 4.36 t 11.45 last 11 data

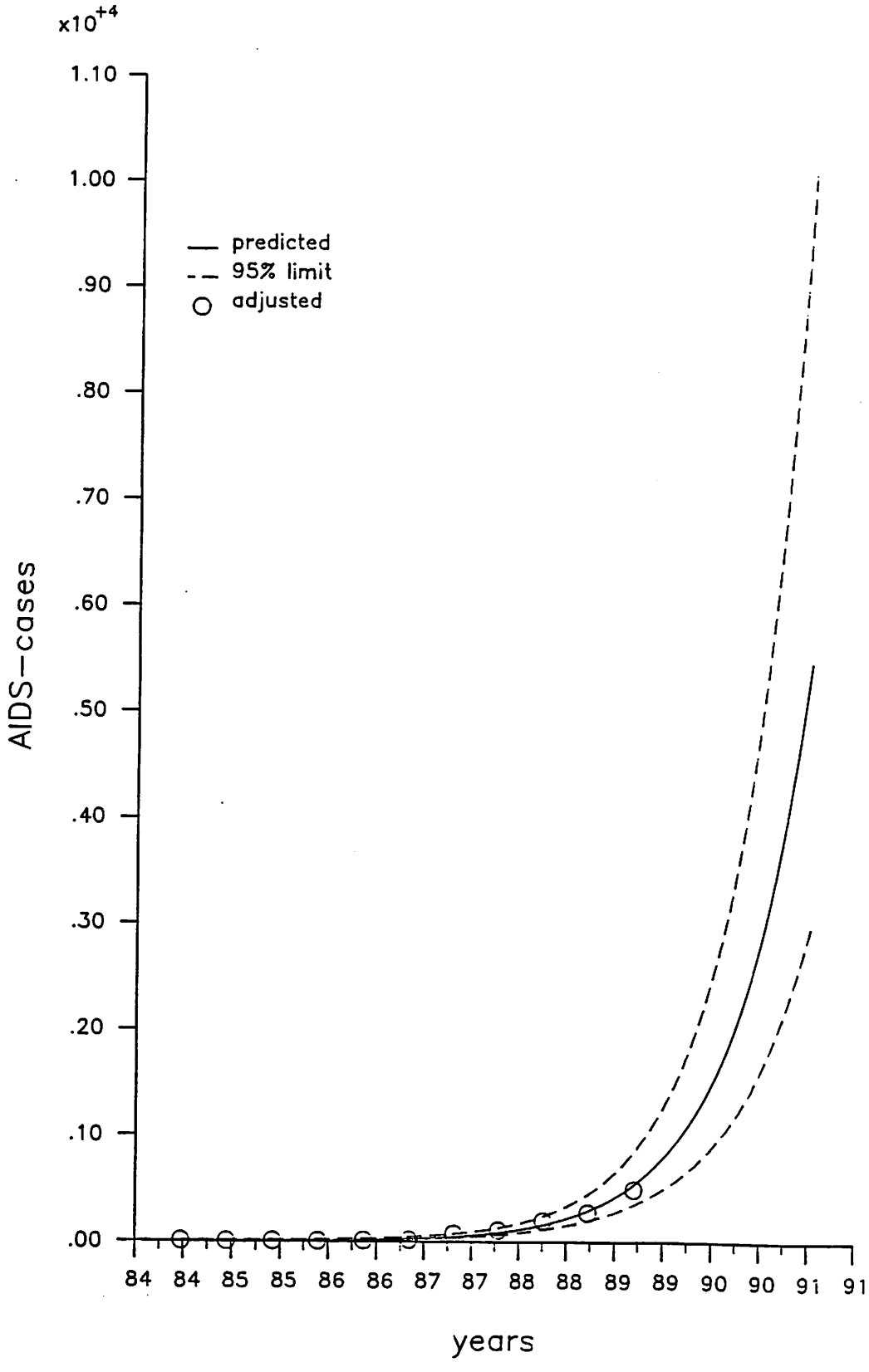

Figure 1(c)

Italy new cases from 1984a onwards until 1987a
Exponential mode a 1.35 t 6.94 last 11 data

9. Conclusions

We think that 'adjusted incidence' is useful for describing an epidemic where a considerable delay occurs between reporting and diagnosis. For short-term predictions a model, provided that it exists, can be fitted directly to the data, without making use of the adjusted cases.

Genstat is an extremely efficient and flexible tool for fitting such a baroque model as we propose here. The effort in programming such a model in Genstat is much less than in achieving the same result in, for example, Fortran, which as a matter of fact we did recently on a Personal Computer (just before we received Genstat 4.03E). This applies to the output and to the ease of changing the model as well. In fact, one can easily imagine that such a model would not be constructed if a language such as Genstat did not exist. Our plans for the future are: addressing the problem of confidence limits more thoroughly, refining the procedure by allowing for over-reporting, splitting the data in order to study sub-groups, and eventually replacing the function used for short-term prediction by mechanistic models for long-term prediction. The latter seems rather difficult in Genstat 4.04; but Genstat 5, with the new PASS directive, might be more suitable for that.

10. Acknowledgements

The discussion during the Genstat conference in Pavia (which not only took place in the conference room but also on the Piazza della Vittoria until the late hours) was interesting and stimulating. Special thanks go to David Brown for his ideas on making the model more robust, which in fact are used in the results presented here.

11. References

[1] Alvey, N.G. *et. al.*
Genstat. A General Statistical Program.
Numerical Algorithms Group, Oxford, 1983.

[2] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W.
Discrete Multivariate Analysis: Theory and Practice.
The MIT press, Cambridge Massachusetts, 1975.

[3] Brookmeyer, R. and Gail, M.H.
Minimum Size of the Acquired Immunodeficiency Syndrome (AIDS) Epidemic in the United States.
The Lancet, pp. 1320-1322, 1986.

[4] Cox, D.R. and Hinkley, D.V.
Theoretical Statistics.
Chapman and Hall, London, 1974.

[5] Downs, A.M., Ancelle, R., Brunet, J.B., Jager, J.C., Heisterkamp, S.H., Druten, J.A.M. van, de Boo, Th. and Ruitenberg, E.J.
Trends in the Incidence of AIDS in Europe: Statistical Analysis of Surveillance Data (1981-1986).
Proceedings of the EC Workshop on Statistical Analysis and Mathematical Modelling of AIDS, Bilthoven, December 15-17th.
(in press).

[6] Downs, A.M., Ancelle, R.A., Jager, J.C. and Bruner, J.S.
AIDS in Europe: Current Trends and Short-term Predictions Estimated from Surveillance Data, January 1981-June 1986.
AIDS, **1**, pp. 53-57, 1987.

[7] Druten, J.A.M. van, de Boo, Th., Jager, J.C., Heisterkamp, S.H., Coutinho, R.A. and Ruitenberg, E.J.
AIDS Prediction and Intervention.
The Lancet, **12**, pp. 852-853, 1986.

[8] Efron, B. and Thisted, R.
Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?
Biometrika, **63**, 3, pp. 435-447, 1976.

[9] Heisterkamp, S.H., Jager, J.C., Downs, A.M., Druten, J.A.M. van, de Boo, Th. and Ruitenberg, E.J.
Statistical Estimation of AIDS Incidence from Surveillance Data and the Link with Modelling of Trends (1987).
Proceedings of the EC Workshop on Statistical Analysis and Mathematical Modelling of AIDS, Bilthoven, December 15-17th.
(in press).

[10] Karlin, S. and Taylor, H.M.
A Second Course in Stochastic Processes.
Academic Press, New York, 1981.

[11] May, R.M. and Anderson, R.M.
Transmission Dynamics of HIV Infection.
Nature, **326**, pp. 137-142, March 1987.

[12] McCullagh, P. and Nelder, J.A.
Generalised Linear Models.
Chapman and Hall, London, 1983.

[13] McEvoy, M. and Tillet, H.E.
Some Problems in the Prediction of Future Numbers of Cases of the Acquired Immunodeficiency Syndrome in the UK.
The Lancet, II: pp. 541-542, 1985.

[14] Morgan, W.M. and Curran, J.W.
Acquired Immunodeficiency Syndrome: Current and Future Trends.
Public Health Reports, **101**, pp. 459-465, 1986.

[15] Nelder, J.A. and Wedderburn, R.W.M.
Generalised Linear Models.
J. Roy. Stat. Soc., A, **135**, pp. 370-384, 1972.

A Procedure for Robust Pairwise Comparisons Between Means

D Brown
R W Marrs
D E Walters

Statistics and Computing Group
AFRC Institute of Animal Physiology
& Genetics Research
Babraham Hall
Cambridge
United Kingdom *CB2 4AT*

1. Introduction

In some animal experimentation within the area of basic science rather than applied research or development, simple experimental designs in which the unit is an individual animal occur frequently. Even though the set of treatments applied, one to each of these animals, might have a factorial structure, many biologically significant questions hinge on pairwise comparisons of the treatments, and sometimes on contrasts between the mean response of these treatments.

A typical example concerns the amount of hormone secreted by the pituitary gland of experimental rats under electrical stimulation. The gland is surgically removed from the animal and impaled on a stimulating electrode, and the ratio of amount of hormone released during a control stimulation and the amount when the gland is kept in a medium containing the substances of interest is the response variable. The treatments are application of Yohimbine (Y) and of Propanolol (P) in a 2×2 factorial arrangement: 0, Y, P, YP. The questions asked were:

(1) whether there is a non-specific effect of Y, i.e. an effect of Y in the absence of P;

(2) whether there is an effect of Y in the presence of P; and

(3) whether Y and P interact.

The data from this experiment, itself one of a much longer series, are given in Table 1.

Control	Y	P	YP
		0.8 0.9 0.9 0.9	
1.0 1.1 1.1		1.1	1.1
1.3 1.3		1.2 1.3 1.3 1.3	1.2 1.2 1.3 1.3
1.5	1.4	1.4	
1.7 1.7			1.7
1.9	1.8		
2.1	2.0		
	2.3		
	2.6		
	3.7		
	4.2		

Table 1

Data for an example from neurophysiology.

Though these experiments are planned as balanced experiments in which the replication of each treatment is equal, circumstances often make this extremely difficult to achieve. For example, for some treatment combinations it might be very difficult to ensure that the gland responds to stimulation at all. Necessarily the replication of the experiments is kept to a minimal level both because of ethical considerations, and of cost. Common features of the responses are non-Normality and inequality of variance which cannot be removed by a simple transformation. Thus there are four aspects which together render almost valueless the usual parametric pairwise comparison tests and estimation procedures: unequal and low replication, non-Normality and inequality of variance.

2. Requirements for a Genstat 5 Procedure

The advent of Genstat 5 provided us with the opportunity to write a procedure for analysing data such as these quickly and conveniently. Our requirements were that the procedure should be as easy to use as a oneway ANOVA directive, and it should provide a reasonably complete analysis in a compact printed and graphical form. The writing of reports and presentation of results for awkward experimental results such as these can be very time-consuming; equally difficult sometimes is the job of the scientist (or the statistician returning to the work months, or even years, later) in appreciating and comparing the various approaches. A standardised and compact presentation would be a great help.

3. Analysis Performed

The procedure eventually written incorporated the following analyses:

(1) for each pair of treatments, the following significance test results:

  (a) the usual t-test using the pooled variance from the anova;

  (b) the approximate version of the Welch test, [5], using only the variances of the pair of treatments in question, which is highly robust to inequality of variance; Best and Rayner [1] concluded that this simple test is approximately equal in power to a range of tests for the Behrens-Fisher problem;

  (c) an observation randomisation test using the difference of the treatment means as the test statistic (an exact version [2]), and one involving sampling the randomisation distribution [4]); and

  (d) a rank randomisation test, or Mann-Whitney U-test, which uses the difference of the medians as the test statistic;

(2) for each pair of treatments, the difference between the means, its standard error using the Welch approach, and its effective degrees of freedom;

(3) for each treatment, an individual mean and median and 95% confidence intervals on both, standard deviation and standard error, and inter-quartile range;

(4) for each specified contrast between treatment means a robust test and interval estimate; this was achieved – providing only a partial solution to the robustness problem – using the modification of the Welch analysis by Satterthwaite [3] in which the estimated variance of the contrast is assumed to be distributed as a scaled chi-squared variable.

4. A Graphical Representation of the Results

A further need was a graphical representation of as much of this information as possible. The printed and graphical output is given for the above example in Table 2 and Figure 1(a) respectively. The annotated printed output is self-explanatory, but the graph might require some justification and explanation. The axes of the box at the lower and lefthand edges give the scale of measurement; these are not repeated elsewhere to prevent the graphs becoming cluttered. At the top of the figure is a schematic plot of the standard deviation of each treatment, the length of

each arm being proportional to the standard deviation within that treatment. Below this are plotted the treatment means with attached lines representing 95% confidence intervals on the means. Immediately above the box two character treatment labels – the labels given to the levels of the treatment factor – are given, together with the treatment replication. The left hand edge of these labels is positioned at the treatment mean except where two means are so close together that their labels would overlap. If necessary, these labels are automatically spread out by the procedure, in order of increasing mean, just sufficiently to prevent overlap. This is repeated at the right of the figure for the medians, 95% confidence intervals on the medians, and the interquartile ranges centred on the median (this last plot is of some help in detecting assymmetry in the distribution, and more robustly comparing the treatment variances).

```
!   TABLE OF MEANS, STANDARD DEVIATIONS & STANDARD ERRORS.                  !
!    NV  Number of values for each treatment level                          !
!    MN  Treatment means                                                    !
!    SDR Standard Deviation                                                 !
!    SER Standard Error (Mean)                                              !
!    L95CI   Lower 95% Confidence Limit                                     !
!    U95CI   Upper 95% Confidence Limit                                     !
!     All the above are calculated for each treatment level separately      !
!                                                                           !
!    SEP Standard Error of Mean calculated from Residual M.S. in ANOVA      !
!    RATIO   Ratio of SER to SEP                                            !
!              This gives an indication of whether the variance in that     !
!              particular treatment level is above or below average.        !
!            ---------------------------------------                        !
!                                                                           !
!    TABLE OF MEDIANS & QUARTILES                                           !
!    MD      Median                                                         !
!    IQR     Interquartile range                                            !
!    LCI, UCI Lower and upper conservative 95% CI s on median - exact       !
!         confidence given by CONF                                          !
!    LQ, UQ  Lower & upper quartiles                                        !
!-------------------------------------------------------------------------!

 RESDF          RESV
    29       0.28092


 Name   NV      MN     SDR   L95CI   U95CI     SER     SEP   RATIO
   _P   10   1.139   0.216   0.984   1.294   0.068   0.168   0.408
   YP    6   1.325   0.207   1.108   1.542   0.085   0.216   0.391
   __   10   1.498   0.368   1.235   1.761   0.116   0.168   0.694
   Y_    7   2.571   1.024   1.624   3.519   0.387   0.200   1.932


 Name   NV      MD     IQR     LCI     UCI    CONF      LQ      UQ
   _P   10   1.195   0.390   0.890   1.340  97.860   0.920    1.31
   YP    6   1.255   0.180   1.130   1.710  96.880   1.210    1.39
   __   10   1.455   0.570   1.160   1.860  97.860   1.180    1.75
   Y_    7   2.300   1.900   1.400   4.200  98.440   1.800    3.70


 ------------------------------------------------------------------------
!   DIFFERENCES OF treatment MEANS, STANDARD ERRORS OF DIFFERENCES,        !
!                       T-TESTS & SIGNIFICANCES                            !
!   In the square matrices below,the lower right triangle gives            !
!   Standard Errors, T-Values, and levels of significance using the        !
!   Residual Mean Square from the ANOVA above as the Error Mean Square     !
!   in each comparison.                                                    !
!   The upper left  triangle of the matrix gives Standard Errors etc using !
!   only the within treatment variances of the 2 factor levels in question.!
!-------------------------------------------------------------------------!
```

Table 2

(continued on next page)

```
     DFM
 x   _P   YP   __   Y_
 y
Y_    7    6    8 1000
__   15   13 1000   29
YP   13 1000   29   29
_P 1000   29   29   29

            MDIFF
 x            _P            YP            __            Y_
 y
Y_       -1.4324       -1.2464       -1.0734        0.0000
__       -0.3590       -0.1730        0.0000        1.0734
YP       -0.1860        0.0000        0.1730        1.2464
_P        0.0000        0.1860        0.3590        1.4324

            SEDIFF
 x            _P            YP            __            Y_
 y
Y_        0.3931        0.3963        0.4042             *
__        0.1349        0.1438             *        0.2612
YP        0.1088             *        0.2737        0.2949
_P             *        0.2737        0.2370        0.2612

            T
 x        _P        YP        __        Y_
 y
Y_    -3.64     -3.15     -2.66         *
__    -2.66     -1.20         *      4.11
YP    -1.71         *      0.63      4.23
_P        *      0.68      1.51      5.48

     SIG
 x   _P   YP   __   Y_
 y
Y_  1.1  2.0  3.3    *
__  1.9 25.1    *  0.1
YP 11.2    * 53.3  0.1
_P    * 50.3 14.1  0.1

 ---------------------------------------------
! Non-Parametric Tests                        !
! Above diagonal - obsn randomization         !
!  below         - rank randomization         !
 ---------------------------------------------

          NPARMAT
 x   _P   YP   __   Y_
 y
Y_  0.5  1.1  0.8    *
__  2.1 32.5    *  1.2
YP 12.0    * 44.8  0.7
_P    * 28.0  4.6  0.5

------ Contrasts ------

   __       Y_       _P       YP CONTRST      SEC      DFC       TC     SIGC
-1.00     1.00     1.00    -1.00 -0.8874   0.4186    8.130    2.120    6.732
```

Table 2

The printer output from the procedure applied to the data in Table 1. The initial analysis of variance which is usually output has been suppressed.

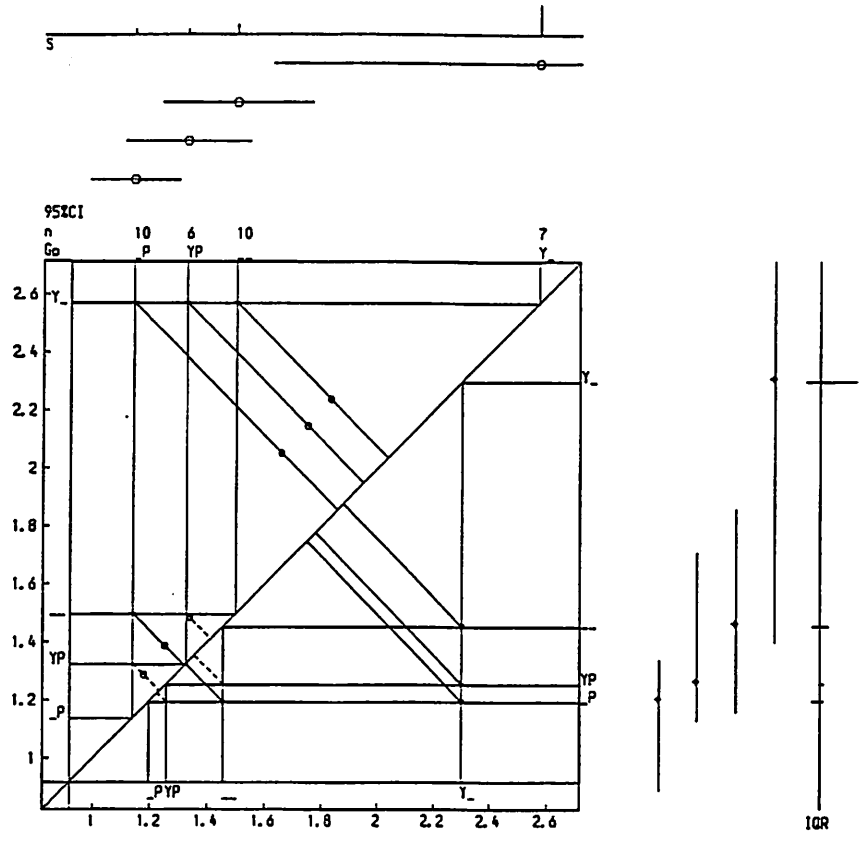

Figure 1(a)

Unlogged

The main part of the figure is split into two diagonally: the top left represents the treatment comparisons based on means and Welch tests; the lower right on medians and rank randomisation test. Consider just the top left. Each point in this half represents a pair of means; which pair is given by the labels on the top and left axes. Faint lines are drawn from these labels through the treatment mean points to aid identification of the points. When the plot is done in colour, these are usually drawn in a light colour which can easily be disregarded by the eye when necessary. It can easily be seen that the size of the treatment differences corresponding to a given point is equal to the distance from the point to the diagonal in a direction parallel to the y-axis. Lines representing all these differences could be drawn on the figure but the problem then would be that the lines for a number of treatment differences would necessarily lie on top of each other. To avoid this, the lines are rotated anticlockwise through 45 degrees, thus producing 'treatment-difference-lines' which join each 'treatment-pair-point' to the diagonal, in a direction perpendicular to the diagonal. The length of these lines is thus reduced by a factor of $\sqrt{2}$. In this form the lines representing treatment differences would only occasionally overlap. The base of each perpendicular is positioned along the diagonal a distance from a zero origin equal to the average of the means in question divided by $\sqrt{2}$. All that has been done is to rotate the axes through 45 degrees. The new coordinates relative to the diagonal axes, if we assume for the moment that the origins of our main axes are at zero, are then

$$x' = \frac{x + y}{\sqrt{2}} \qquad y' = \frac{y - x}{\sqrt{2}}$$

In order to facilitate an understanding of the treatment differences, the lines corresponding to non-significant differences according to the Welch test are plotted as dashed lines, and the 'significant' differences are plotted as solid lines, colour coded according to the level of significance, necessarily discretely, with black denoting 5%, red denoting 1% and blue denoting

0.1%. Additionally the symbol representing the treatment mean pair is plotted differently depending on the level of significance: no symbol implying non-significant, a star for a result significant at 5%, a circle at 1%, and a cross and circle at 0.1%. This would enable the diagram to be photocopied without losing information.

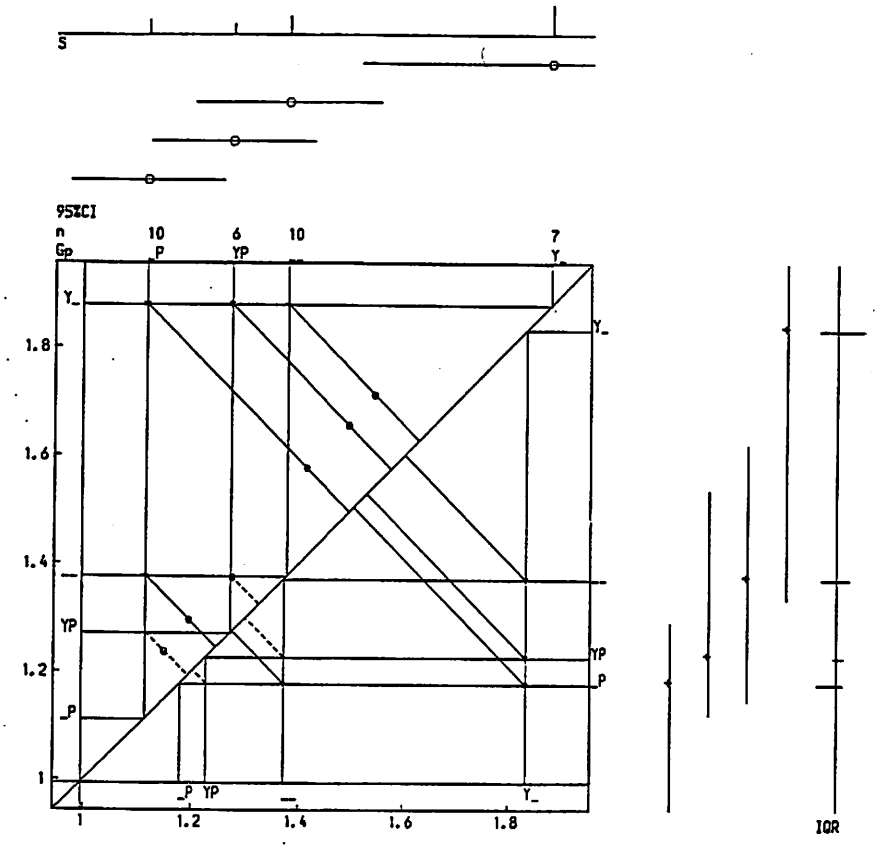

Figure 1(b)

$Log(Y) + 1$

It should be stressed that the object of the representation of the levels of significance in the figure is to aid interpretation: a preliminary filter enabling us to give less weight to differences for which there is not strong evidence in the present data. The intention is not to facilitate a screening of all pairwise differences; in pilot experimentation with totally unstructured treatments, a multiple comparison procedure could perhaps be used, although we do not favour such an approach. Our aim is to present a picture of the relationship between the means which can be compared with expectations established *a priori*, and – if those expectations are not fulfilled – to aid the process of interpretation and formation of new hypotheses which could then be assessed by further experimentation.

Significance tests by themselves are of little value: it is important when assessing the results of the tests to consider the standard errors of the differences of the means and these are plotted, appropriately scaled by $1/\sqrt{2}$, up from the diagonal, along each 'treatment-difference-line' with the symbol e. The common baseline of the diagonal enables the viewer to see directly which comparisons have the greatest uncertainty attached to them.

The lower right hand half of the box gives the same results but based on medians and on rank-randomisation tests. Standard errors are not given. Thus the rank based comparisons of the medians can be contrasted with the parametric comparisons of the means. The medians and means can also be directly compared by considering their relative positions on the right and left, or bottom and top margins of the box, respectively.

Figure 1(b) is the graphical output resulting from a re-analysis after a log transformation of the data, which was suggested by the standard deviation plot at the top of Figure 1(a). Log transformation removes most of the heteroscedasticity, but the pairwise comparisons of the treatments have not changed in essence as a result of the transformation. This illustrates the value of a graphical summarisation for easy comparison of the results of a transformation.

5. The Structure of the Procedure

The procedure is written in Genstat 5 with the exception of the code for the randomisation tests. This illustrates the great value of being able to move outside Genstat for those calculations which can be performed more efficiently in other languages or for which code in another language already exists. Fortran 77 programs for these analyses had been written by us and used for a number of years and were thus available for incorporation into the procedure using the PASS directive (which unfortunately might only be available on certain computers and certain operating systems). The same code is used for the rank and the observation randomisation tests, thus obviating the need to tabulate or calculate the percentage points of the Mann-Whitney U-statistic.

The Genstat code alone is about 600 lines long. We originally envisaged a modular approach to the procedure in which the main procedure called further procedures but avoided this for much of the development as early in the project we discovered that there could be considerable overheads both in time and space requirements when using procedures (see Section 6), and we wished to minimize these. In the end, with more understanding of the efficiency of Genstat, we returned to the modular form of a procedure calling a number of sub-procedures.

Enabling the procedure to deal with restricted variates in parameter list and still preserving the labelling of the variates and the factors caused us some initial headaches. The solution we found was to copy the original unrestricted versions of the variates and factors into temporary variates and factors, and to replace all restricted out individuals by missing values in the original variates, and then to proceed being careful to make proper allowance for missing values, which in some cases necessitated further copying into structures of reduced size. At the end of the procedure the original variates and factors are copied from their temporary accommodation back to their original homes.

It seemed that it would be convenient to re-order the levels of the treatment factor in any tables produced so that they are in order of increasing mean. This greatly eases interpretation of matrices of significance levels etc. This was done initially and necessitated a further copying of the variates and factors.

The randomisation tests were performed by two separate Fortran 77 programs, one determining an exact significance level by summing the probability in the tail of the randomisation distribution, the other obtaining an estimate by sampling the randomisation distribution (written by R W Marrs and D E Walters respectively). The exact randomisation program produced estimates very quickly when sample sizes were small (typically under a tenth of a second on a VAX 750 under VMS for two samples of size 6 and 6). The procedure determines automatically whether to use the 'sampling' or the 'exact' subprogram, on the basis of the two sample sizes. The precision required in the estimated significance level when using the sampling routine was dependent on the true significance level; there is little point in accurately determining a significance level of 25%, for example, but if the value is nearer 5% or 1% then conventionally greater accuracy is required. A two stage sampling scheme ensuring appropriate accuracy for our needs was incorporated into the procedure.

6. Computing Resources

The procedure is expensive to use in resources of both space and CPU time – at least in the form we implemented it on the VAX 750. We tried to improve the efficiency by eliminating

unnecessary code, and by considering various forms of performing the analysis such as running the code in the form of a procedure and a program. We used exactly the same code in various different forms:

(1) as a single very long procedure with no sub-procedure calls;

(2) as a procedure with sections of the code put into sub-procedures and called from the main procedure, which is probably the most convenient and reliable way of writing procedures; checking and compiling of the code of these sub-procedures can be done as they are written;

(3) substituted in the main program like a macro, still calling the sub-procedures;

(4) substituted in the main program with all sub-procedures inserted directly into the code of the main program. The results of this exercise are given in Table 3. It should be noted that all the comparisons were performed on a VAX 750 operating under VMS, using Release 1.2 of Genstat 5 as mounted on the AFRC's network of computers. The results might be quite different for a different computer or operating system.

Form of Program/Procedure			
(1) Procedure without sub-procedures	(2) Procedure calling sub-procedures	(3) Program calling sub-procedures	(4) Program without sub-procedures
Data set 1: four treatments (n=10,10,7,6)			
9.0 (243%)	7.7 (208%)	5.5 (149%)	3.7
Data set 2: six treatments (n=13,12,10,5,4,4)			
10.0 (263%)	8.3 (218%)	5.6 (147%)	3.8
Data set 3: eight treatments (n=13,10,10,7,6,6,5,5)			
11.7 (249%)	9.6 (204%)	7.0 (149%)	4.7

Table 3

CPU Time (in minutes) required for running the ONEWAY procedure on three different datasets on a VAX 750 operating under VMS. The bracketed figures give the CPU time as a percentage of the time required for the most efficient version (4).

It is clear that there are substantial overheads in actually using a procedure as opposed to a macro. This applies to a large procedure; i.e. one which uses a few large structures or many small ones, and performs a large number of manipulations (hundreds of lines of Genstat) of these structures. The CPU time of our 600 line procedure was 150% higher than the equivalent straight through program, but this was reduced by 50% by breaking the large procedure into modules (sub-procedures) of about 100 lines and calling these from the procedure. On the other hand calling these subprocedures from the main program increased the CPU time of the main program version by about 50%. It is more difficult to generalise about the storage requirements, but it does appear that using a procedure can increase the store required, but in general by much smaller amounts than the CPU time, typically in our case by about 20%.

The part of the program dealing with the graphical output was rather long, and involved many structures partly because many separate lines needed plotting. It would be useful if a plot directive were available in Genstat which connected all points – in a specified or increasing order of x or y – which had the same level of a factor. Such a facility might have reduced the space requirements of the procedure appreciably.

The CPU time required by the Fortran part of the procedure has been virtually negligible for most of the cases we have considered, a matter of fractions of a second for the exact randomisation test on small samples, or a few seconds for the sampling version with 1000 samples generated, compared to minutes for the rest of the procedure. When we attempted to code the same calculations in Genstat, using the randomise directive, we found that one randomisation test using the sampling method to determine the significance level for two treatments was slower by a factor of about 100-1000. This illustrates the considerable value of being able to call Fortran code for those calculations for which it is more efficient.

7. Conclusions

We feel that the extra facilities which came with Genstat 5 are of great value to statisticians and other users, most particularly the facilities for extension of the language: the use of procedures, and the ability to call code written in other languages. It is pity that there are considerable overheads in the use of procedures in some cases, for example, the CPU time was doubled in the most efficient version of the procedure we could produce; a pity too that the method – even the possibility – of invoking code in other languages might vary from one make of computer to another. More tuning of the Genstat method for calling procedures might reduce the first problem. Greater standardisation on the part of computer and operating system manufacturers about how different processes can interact might solve the second.

A further important change is the introduction of plotter graphics. The integration of a powerful calculating facility and flexible graphics is very valuable: to statisticians, when investigating a complex problem, detailed graphs can facilitate the appreciation of a large volume of numerical results; to statisticians and their clients, the presentation of data and fitted models helps understanding and gives an immediate opportunity for verification that the analysis presented is at least plausible.

Our procedure provides a simple example of what can be done with the new facilities. This procedure makes a fairly comprehensive analysis quickly available, so that different approaches, e.g. different transformations of the response variable, the use of different contrasts between the means, can be easily and efficiently compared.

8. References

[1] Best, D. J. and Rayner, J. C. W.
Welch's Approximate Solution for the Behrens-Fisher Problem.
Technometrics, **29**, pp. 205-210, 1987.

[2] Fisher, R.A.
The Design of Experiments, (pp. 44-49).
Oliver and Boyd, Edinburgh, (Seventh Edition) 1960.

[3] Satterthwaite, F.E.
An Approximate Distribution of Estimates of Variance Components.
Biometrics, **2**, pp. 110-114, 1946.

[4] Walters, D.E.
Sampling the Randomisation Distribution.
The Statistician, **30**, pp. 289-295, 1981.

[5] Welch, B.L.
The Significance of the Difference between Two Means when the Population Variances are Unequal.
Biometrika, **29**, pp. 350-362, 1937.

Using Genstat to Fit Regression Models to Ordinal Data

H Jansen
Institute of Horticultural Plant Breeding
P.O. Box 16
6700 AA Wageningen
The Netherlands

1. Introduction

Regression models for ordered categorical or ordinal data (McCullagh [1]), have proved useful in many practical applications. This paper is concerned with computational methods for fitting such models to data. Basic properties of McCullagh's models are presented in Section 2.

Thompson and Baker [2] mentioned that regression models for ordinal data can be embedded into the framework of generalised linear models by using composite link functions. Consequently, regression models for ordinal data can be fitted by iterative weighted least-squares. In Section 3 the method of Thomson and Baker is considered in detail.

Genstat provides excellent facilities for iterative weighted least-squares. Thompson and Baker's method can be implemented in a Genstat procedure using these facilities. In Section 4, implementation of the method in a Genstat procedure is discussed. Section 5 contains an example of the use of the procedure.

2. Regression Models for Ordinal Data

Suppose Y is a non-observable continuous random variable with mean η and scale parameter σ. Typical distributions for Y are the Normal disribution and the Logistic distribution, of which the distribution functions are given by

$$F(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(y-\eta)/\sigma} e^{-z^2/2} dz$$

and

$$F(y) = \frac{1}{1 + e^{-(y-\eta)/\sigma}}$$

respectively. In practice, the aim is to compare different treatments with respect to their values of η. The scale parameter σ is set to unity, i.e. σ is the unit of measurement on the y-scale.

The model assumes that the y-axis can be divided in k non-overlapping intervals by $(k-1)$ cut-points $\theta_1, \theta_2 \ldots \theta_{k-1}$. Furthermore, $\theta_0 = -\infty$ and $\theta_k = \infty$. See Figure 1. Since the random variable Y cannot be observed, the cut-points $\theta_1, \theta_2 \ldots \theta_{k-1}$ are unknown parameters.

A data set involving ordinal data can be presented by an $m \times k$ matrix $\underline{N}$, of which the i^{th} row refers to treatment i $(= 1,2 \ldots m)$ and the j^{th} column to category j $(= 1,2 \ldots k)$. It is assumed that the rows of $\underline{N}$ are independent and follow a multinomial distribution with parameters n_i and $\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik}$. The probabilities π_{ij} are given by

$$\pi_{ij} = F(\theta_j - \eta_i) - F(\theta_{j-1} - \eta_1).$$

In practice the parameter vector $\underline{\eta} = (\eta_1, \eta_2 \ldots \eta_m)$ takes the form of a linear model $\underline{\eta} = \underline{X}\,\underline{\beta}$, where $\underline{X}$ is the $m \times p$ design matrix for treatments and $\underline{\beta}$ is the vector of unknown parameters. The main problem is to obtain the maximum likelihood estimate of the parameter vector $\underline{\beta}$.

The log-likelihood function l can be written as

$$l = c + \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij} \, ln(\pi_{ij})$$

where c is a function not involving unknown parameters. The likelihood equations for $\underline{\beta}$ read

$$\frac{\partial l}{\partial \beta_q} = -\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{n_{ij}}{\pi_{ij}} v_{ij} x_{iq} = 0, \qquad (q = 1,2,...p)$$

where $v_{ij} = f[\theta_j - \mu_i] - f[\theta_{j-1} - \mu_i]$ and f is the first derivative of F. It should be noted that maximum likelihood estimation involves both $\underline{\beta}$ and $\underline{\theta}$. In order to obtain estimates one restriction on the parameters has to be imposed, e.g. $\theta_1 = 0$.

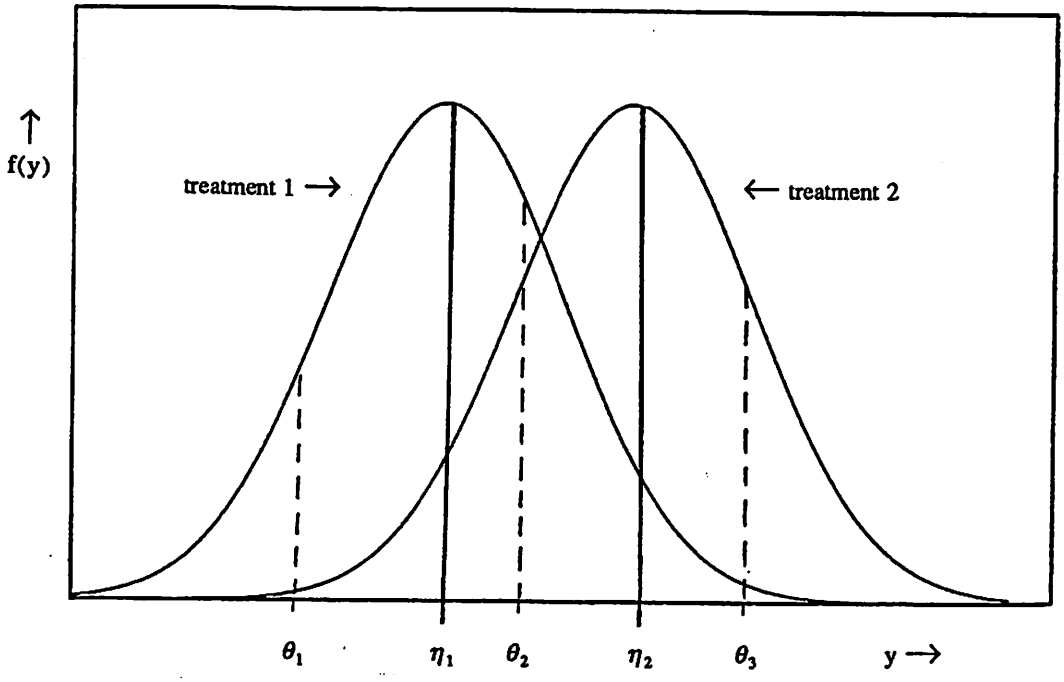

Figure 1
Graphical Presentation of a Regression Model for Ordinal Data

3. Thompson and Baker's Method

The Newton-Raphson method for solving the likelihood equations required first and second derivatives of the log-likelihood. Fisher's scoring technique uses expected second derivatives. The general form of step s (= 1,2,...) of the iteration process is given by

$$\underline{\alpha}_s = \underline{\alpha}_{s-1} + \underline{A}_{s-1}^{-1} \underline{b}_{s-1}$$

where $\underline{\alpha}$ is the vector of parameters, $\underline{b}$ is the vector of first derivatives and $\underline{A}$ is the matrix of expected second derivatives of the log-likelihood. A direct advantage of Fisher's scoring technique is that at convergence it provides directly an estimate of the covariance matrix of the vector of parameters. The form of the vector $\underline{b}$ and the matrix $\underline{A}$ will now be considered in detail.

The probabilites π_{ij} can be written as

$$\pi_{ij} = F(\tau_{i1})$$
$$\pi_{ij} = F(\tau_{ij}) - F(\tau_{i,j-1}) \qquad (j = 2...k-1)$$
$$\pi_{ik} = F(\tau_{ik})$$

where

$$\tau_{ij} = \theta_j - \underline{x}_1^T\underline{\beta} \qquad (j = 1,2...k-1)$$
$$\tau_{ik} = \theta_k + \underline{x}_1^T\underline{\beta}$$

In matrix notation the above results are

$$\underline{\mu}_i = N_i\underline{\pi}_i = \underline{C}_i\underline{F}(\underline{\eta}_i)$$
$$\underline{\tau}_i = \underline{Z}_i\underline{\alpha}$$
$$\underline{\alpha}^T = (\underline{\theta}^T,\underline{\beta}^T)$$

For example, for $k = 5$ the matrices $\underline{C}_i$ and $\underline{Z}_i$ are given by

$$\underline{C}_i = N_i\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \underline{Z}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & \underline{x}_i^T \\ 0 & 1 & 0 & 0 & \underline{x}_i^T \\ 0 & 0 & 1 & 0 & \underline{x}_i^T \\ 0 & 0 & 0 & 1 & \underline{x}_i^T \\ 0 & 0 & 0 & -1 & \underline{x}_i^T \end{bmatrix}$$

Furthermore, let $\underline{H}_i$ and $\underline{W}_i$ be diagonal matrices with j^{th} diagonal elements given by

$$h_{ij} = \frac{\partial F(\tau_{ij})}{\partial \tau_{ij}} \quad \text{and} \quad w_{ij} = \frac{1}{\mu_{ij}},$$

respectively. The first and expected second derivatives of the log-likelihood can be written as

$$\underline{b} = \sum_{i=1}^{m} \underline{Z}_i^T\ \underline{H}_i\ \underline{C}_i^T\ \underline{W}_i\ (\underline{n}_i - \underline{\mu}_i)$$

$$\underline{A} = \sum_{i=1}^{m} \underline{Z}_i^T\ \underline{H}_i\ \underline{C}_i^T\ \underline{W}_i\ \underline{C}_i\ \underline{H}_i\ \underline{Z}_i.$$

By writing $\underline{Z}^T = [\underline{Z}_1^T \ ... \ \underline{Z}_m^T]$, $\underline{H} = \text{diag}[\underline{H}_1 \ ... \ \underline{H}_m]$, $\underline{C} = \text{diag}[\underline{C}_1 \ ... \ \underline{C}_m]$, $\underline{W} = \text{diag}[\underline{W}_1 \ ... \ \underline{W}_m]$, $\underline{n}^T = [\underline{n}_1^T \ ... \ \underline{n}_m^T]$, $\underline{\mu}^T = [\underline{\mu}_1^T \ ... \ \underline{\mu}_m^T]$, $\underline{\tau}^T = [\underline{\tau}_1^T \ ... \ \underline{\tau}_m^T]$ the iteration process becomes

$$\underline{\alpha}_s = \underline{\alpha}_{s-1} + (\underline{Z}^T\underline{H}\ \underline{C}^T\underline{W}\ \underline{C}\ \underline{H}\ \underline{Z})^{-1}\ [\underline{Z}^T\ \underline{H}\ \underline{C}^T\underline{W}\ (\underline{n}-\underline{\mu})]$$
$$= (\underline{Z}^T\underline{H}\ \underline{C}^T\underline{W}\ \underline{C}\ \underline{H}\ \underline{Z})^{-1}\ \underline{Z}^T\underline{H}\ \underline{C}^T\underline{W}\ \underline{C}\ \underline{H}\ [\underline{Z}\ \underline{\alpha}_{s-1} + \underline{H}^{-1}\underline{C}^{-1}(\underline{\eta}-\underline{\mu})].$$

Since $\underline{C}$ is not a diagonal matrix (it is block-diagonal), solving the above equation requires a generalised least-squares program. However, it can be modified to fit in a weighted least-squares framework by writing

$$\underline{\alpha}_s = (\underline{Z}^T\underline{H}\ \underline{C}^T\underline{W}\ \underline{C}\ \underline{H}\ \underline{Z})^{-1}\ \underline{Z}^T\underline{H}\ \underline{C}^T\underline{W}\ [\underline{C}\ \underline{H}\ \underline{Z}\ \underline{\alpha}_{s-1} + (\underline{n}-\underline{\mu})]$$

Thus, a step of the iteration process may be carried out by a weighted least-squares regression with

- working dependent variate: $\underline{C}\ \underline{H}\ \underline{\eta} + (\underline{n}-\underline{\mu})$
- weights: $\underline{W}$
- regression variates: $\underline{C}\ \underline{H}\ \underline{X}$

Starting values can be obtained easily for the iteration process described above. An initial value for $\underline{\mu}_i$ is obtained by taking $\underline{\mu}_i^0 = \underline{n}_i$. It then follows that

$$\underline{\tau}_i^0 = \underline{F}^{-1}[\underline{C}_i^{-1}\underline{n}_i]$$

In order to avoid the complication of zeros in the data matrix, n_{ij} has to be replaced by $n_i(n_{ij}+\frac{1}{2})/(n_i+m/2)$. It can be shown easily that for $m = 5$:

$$\underline{C}_i^{-1} = n_i^{-1}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

4. Procedure ORDINAL

The computations described above have been implemented in a Genstat procedure, called ORDINAL. An example of the use of ORDINAL is given in Section 6. A listing of a preliminary version of the procedure is available on request from the author. The final version of the procedure will be submitted to the Genstat Procedure Library.

The procedure ORDINAL uses the Genstat regression facilities to carry out the computations described above. It appeared that it is able to cope with data sets of considerable size, and that if the model fits the data reasonably well, convergence is attained in a few iteration steps. Initial values are calculated from the data by the method described in Section 4.

As the procedure uses variates of length $m \times k$ (m = number of rows of the data matrix; k = number of response categories), use of the procedure requires a fair amount of computer space and time.

A disadvantage of the procedure is that linear models for treatment effects have to be supplied as variates; no use can be made of the Genstat model formulae.

5. Example

The example concerns a small part of a larger experiment concerning vascular wilt disease (caused by ***Fasarium oxysporum ssp Dianthi***) in carnation (***Dianthus cariophyllus L.***). In this example two isolates of fasarium have been applied on four genotypes of carnation. In the experiment there were about 35 plants with each combination of isolate and genotype. At the end of the experiment plants were assigned to one of four categories:

category 1: plant not affected,
category 2: plant showed discolouration of the vessels,
category 3: plant showed discolouration of the vessels, and also wilting symptoms,
category 4: plant has died.

The data are shown in Table 1:

Isolate	Genotype	y[1]	y[2]	y[3]	y[4]	x[1]	x[2]	x[3]	x[4]	x[5]	x[6]	x[7]	x[8]
1	1	1	3	12	19	1	0	0	0	0	0	0	0
1	2	0	6	23	6	1	0	1	0	0	0	0	0
1	3	0	13	20	2	1	0	0	1	0	0	0	0
1	4	20	12	0	0	1	0	0	0	1	0	0	0
2	1	1	12	18	4	1	1	0	0	0	0	0	0
2	2	1	16	17	1	1	1	1	0	0	1	0	0
2	3	16	19	0	0	1	1	0	1	0	0	1	0
2	4	27	8	0	0	1	1	0	0	1	0	0	1

Table 1
Data from the fasarium experiment and explanatory variates representing main effects and interactions

Also shown in Table 1 are the explanatory variates, representing the main effects of isolates and genotypes and their interaction. Unfortunately, the procedure does not allow the user to specify main effects and interactions using Genstat model formulae. A Genstat program for fitting a regression model for ordinal data to the data given in Table 1 could be of a form shown below:

```
JOB "Fasarium experiment"
UNITS [8]
FACTOR [LEVELS=2; VALUES=4(1,2)] isolate
FACTOR [LEVELS=4; VALUES=(1....4)2] genotype
READ y[1...4]
 data
VARIATE [VALUES=8(1)] x[1]
CALCULATE x[2]=isolate.EQ.2
&         x[3...5]=genotype.EQ.2...4
ORDINAL NCATEGORIES=4; Y=y; X=x
"default options: PRINT=SUMMARY,ESTIMATES; LINK=LOGIT"
STOP
```

The default output form the procedure ORDINAL after fitting the main effects of isolates and genotypes is given below:

```
*** Regression analysis of ordinal data ***
    Link function: logit

*** Summary of analysis ***
    Dispersion parameter is 1

    Residual diviance:      33.84
    Degrees of freedom:        17

*** Estimates of regression coefficients ***

                          Estimate       s.e.
    Theta_ 2                 2.970     0.3461  } estimates of
    Theta_ 3                 5.817     0.4792  } cut-points (θ1=0)
     Beta_ 1                 5.743     0.5489  }
     Beta_ 2                -2.079     0.3192  } estimates of
     Beta_ 3                -0.986     0.3824  } regression coefficients
     Beta_ 4                -2.929     0.4376  } corresponding with
     Beta_ 5                -5.786     0.5885  } x[1...5]
```

An analysis of deviance table is easily constructed from the output of the procedure.

6. References

[1] McCullagh, P.
Regression Models For Ordinal Data (with discussion).
J. R. Statist. Soc. Ser. B, 42, pp. 109-142, 1980.

[2] Thompson, R. and Baker, R.J.
Composite Link Functions in Generalised Linear Models.
Appl. Statist., 30, pp. 125-131, 1981.

Genstat Analyses for Complex Balanced Designs with Non-interacting Factors

D A Preece
Institute of Horticultural Research
East Malling
Maidstone
Kent
United Kingdom *ME19 6BJ*

1. Introduction

Some Genstat users have reported puzzlement and misunderstanding in connection with non-orthogonality amongst treatment factors in ANOVA. Indeed such non-orthogonality can provide pitfalls for the unwary. It therefore seems worthwhile to illustrate the points at issue by examining analyses for some designs of a general type used in orchard experimentation at East Malling. Genstat 5 notation is used throughout. The reproduced computer-output is as obtained from the version of Genstat 5 that was in use at East Malling in the summer of 1987.

In orchard experimentation, a set of trees must often be used in successive years to test successive sets of treatments. So designs are needed where a new set of treatments is superimposed on a previous set. Often, residual effects of the previous treatments must be allowed for in designing the superimposition, even though they may be assumed not to interact with the effects of the new treatments. If each of the two sets of treatments is non-factorial, the design will then have one factor for the new set, one for the old, and possibly one or more blocking factors from the original design. As the residual effects may be of little or no interest in themselves, the old treatments may sometimes be most conveniently viewed as the levels of a blocking factor, not a treatment factor. But whichever choice is made, the design is a so-called *main-effect plan*, i.e. a multi-factor design for factors assumed not to interact. Such designs, although important in orchard experimentation, have non-agricultural uses too, usually with all factors present from the outset.

One of the simplest such multi-factor designs is a Graeco-Latin square, with the rows and columns used for blocking factors, the Roman letters used for an initial set of treatments, and the Greek letters used for a superimposed set. However, the complete orthogonality of the Graeco-Latin square is usually unattainable in practice. Instead, the experimenter may have to use a non-orthogonal design that may or may not be balanced in the sense recognized by Genstat.

2. Three-factor Designs Where Two Mutually Orthogonal Factors are Each Balanced With Respect to a Third

This article is concerned with balanced multi-factor designs that can be analysed by Genstat's ANOVA without the use of pseudo-factors. However, recognising such a design is not as simple as the novice might suppose. Consider, for example, the following block design (unrandomized) for two non-interacting sets of treatments (say T1 and T2, denoted by upper- and lower-case letters respectively) in 10 blocks each of size 3:

			Block						
1	2	3	4	5	6	7	8	9	10
Af	Bf	Cf	Df	Ef	Aa	Bb	Cc	Dd	Ee
Bd	Ce	Da	Eb	Ac	Cd	De	Ea	Ab	Bc
Ec	Ad	Be	Ca	Db	Dc	Ed	Ae	Ba	Cb

(1)

Inspection shows that this design has the following properties:

(1) the allocation of the upper-case letters to blocks constitutes a balanced incomplete block design;

(2) the allocation of the lower-case letters to blocks constitutes a balanced incomplete block design;

(3) the two sets of letters are mutually orthogonal, as each letter of each set is paired the same number of times (once) with each letter of the other set.

So, if we are interested in one set of treatments and ignore the other, we have a balanced analysis. This may lead us to suppose that the analysis will remain balanced after introduction of the other set, this being itself orthogonal or balanced with respect to each of the other factors. So, with the block factor and treatment factors appropriately defined, we may try specifying the analysis as follows:

```
BLOCKS Block
TREATMENTS T1+T2
ANOVA
```

However, this ANOVA statement will rightly reject the design as unbalanced (Fault Code AN 1). The reason is that, in the blocks containing any one particular treatment from either set, the treatments of the other set are unequally replicated. (For example, treatment e occurs in blocks 2, 3, 7, 8, 10, and these blocks have A and D twice each, B and C four times each, and E three times.) Examination of the normal equations quickly shows that such unequal replication unbalances the design.

Consider now, however, this alternative (again unrandomized):

			Block						
1	2	3	4	5	6	7	8	9	10
Af	Bf	Cf	Df	Ef	Aa	Bb	Cc	Dd	Ee
Cd	De	Ea	Ab	Bc	Bd	Ce	Da	Eb	Ac
Dc	Ed	Ae	Ba	Cb	Ec	Ad	Be	Ca	Db

(2)

Not only does this second design have properties 1, 2 and 3 as before, it also satisfies the following:

4) In the blocks containing any particular letter from either set, the letters of the other set are equally replicated.

So ANOVA with block and treatment factors as before will accept this as a balanced design. Indeed, the four properties together have the further valuable consequence that the efficiency factor (e.f.) for each set of treatments (5/6 = 0.833 for the upper-case treatments and 4/5 = 0.800 for the lower-case) is the same as if the other set were absent. Another way of saying this is that, because of property 4, the blocks do not induce any non-orthogonality between the two mutually orthogonal sets of treatments. In the terminology of Eccleston and Russell [3,4], the two sets have *adjusted orthogonality* as well as *pairwise orthogonality*, when adjustment is made for blocks.

2.1. Example 1

The balanced design (2) was quoted because of the fairly systematic way in which it can be generated (from blocks 1 and 6). A combinatorially different design with the same statistical properties was given, with data, by Potthoff [6, p. 201]. In the notation of the present article they were as follows, where the variate values are square roots of counts; as Potthoff worked with the square roots rounded to two decimal places, they are used and presented in the same form here:

	Block								
1	2	3	4	5	6	7	8	9	10
Aa	Ba	Ac	Ad	Be	Ab	Cc	Ae	Bc	Af
8.49	7.00	7.68	7.28	6.16	10.05	7.00	3.16	9.90	8.83
Bb	Cd	Ca	Cf	Df	De	Db	Bf	Ce	Bd
7.07	7.07	7.87	7.35	8.19	6.78	5.92	3.46	8.54	9.59
Dc	Eb	Ee	Da	Ea	Ed	Ef	Cb	Dd	Ec
9.43	7.42	5.82	7.21	7.55	6.78	6.93	3.61	9.06	10.00

(3)

Potthoff's paper should be consulted for details of the experiment. For present purposes we need no more than the following details of the factors:

Blocks 1, 2,...,10 = Locations 1, 2,...,10
Treatments A, B, C, D, E = Days: Mon, Tues, Wed, Thurs, Fri
Treatments a, b, c, d, e, f = Times, a.m.: 8:00, 8:10, 8:20, 8:30, 8:40, 8:50

With Genstat factors defined in the obvious way, and the variate identified as `Sqrtnmbr`, we can specify an analysis as follows:

```
BLOCKS Location
TREATMENTS Day + Time
ANOVA Sqrtnmbr
```

This produces the following analysis of variance and tables of means:

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

Variate: Sqrtnmbr

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

Location stratum
Day                             4    20.1277     5.0319
Time                            5    55.4349    11.0870

Location.*Units* stratum
Day                             4     1.6614     0.4153     0.63
Time                            5     8.3513     1.6703     2.54
Residual                       11     7.2266     0.6570

Total                          29    92.8019


***** Information summary *****

Model term                    e.f.  non-orthogonal terms

Location stratum
  Day                        0.167
  Time                       0.200

Location.*Units* stratum
  Day                        0.833  Location
  Time                       0.800  Location
```

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

Variate: Sqrtnmbr

Grand mean  7.37

     Day       Mon      Tues       Wed     Thurs       Fri
               7.77      7.04      7.51      7.38      7.17

    Time      8:00      8:10      8:20      8:30      8:40      8:50
              7.60      7.54      8.20      7.08      6.27      7.54

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

Table                    Day      Time
rep.                       6         5
s.e.d.                 0.513     0.573
```

Because the effects of the factors `Day` and `Time` remain mutually orthogonal after allowance for differences between Locations, no difficulty of interpretation arises over the sums of squares in the second stratum of the analysis: the sum of squares for `Day` is that for 'Days after fitting Locations' (sometimes called 'Days eliminating Locations'), and that for `Time` is for 'Times after fitting Locations'. The means in the one-way tables for the factors `Day` and `Time` are *adjusted* means (adjusted for Locations, but without recovery of inter-Location information) and agree with the Day and Time effects calculated by Potthoff.

For this example, it may be questioned whether `Location` should truly have been a block factor rather than a treatment factor. Indeed the following specification is plausible:

```
BLOCKS
TREATMENTS Location + Day + Time
ANOVA Sqrtnmbr
```

This produces the following:

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

Variate: Sqrtnmbr

Source of variation          d.f.       s.s.       m.s.     v.r.
Location                        9    75.5626     8.3958    12.78
Day                             4     1.6614     0.4153     0.63
Time                            5     8.3513     1.6703     2.54
Residual                       11     7.2266     0.6570
Total                          29    92.8019


***** Information summary *****

Model term                    e.f.  non-orthogonal terms
  Day                        0.833  Location
  Time                       0.800  Location
```

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

Variate: Sqrtnmbr

Grand mean  7.37

 Location        1        2        3        4        5        6        7

              8.33     7.16     7.13     7.28     7.30     7.87     6.62

                 8        9       10
              3.41     9.17     9.47

     Day      Mon     Tues      Wed    Thurs      Fri
             7.77     7.04     7.51     7.38     7.17

    Time     8:00     8:10     8:20     8:30     8:40     8:50
             7.60     7.54     8.20     7.08     6.27     7.54


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

Table              Location         Day        Time
rep.                      3           6           5
s.e.d.                0.662       0.513       0.573
```

Here the sums of squares for `Day` and `Time` are as before, as are the means for these two factors. However, we must now note that the analysis of variance reports *sequential fitting* of the terms. (If the analysis of variance table is to be published without risk of ambiguity, the symbol + should perhaps be inserted at the start of each of the three treatment lines, as in accumulated analyses of variance from the regression part of Genstat.) Thus the `Location` sum of squares is for 'Locations before fitting Days and Times' (sometimes called 'Locations ignoring Days and Times'). Likewise, the one-way table of means for the factor `Location` consists of *unadjusted* means (i.e. unadjusted for Days and Times); great care must be taken to remember this. The key to the interpretation of the output here is the Information Summary, with its identification of the non-orthogonality.

Having reached this point, the user may well try writing

```
BLOCKS
TREATMENTS Day + Time + Location
ANOVA Sqrtnmbr
```

in the hope of thereby obtaining adjusted `Location` means and also a sum of squares for `Location` after fitting `Day` and `Time`. However, this fails (Fault Code `AN 1`) and prints the message

```
Model term Location (non-orthogonal to term Day) is unbalanced.
```

The point here is that *balance is not, in general, a symmetric relationship.* Although `Day` (or `Time`) is balanced with respect to `Location`, `Location` is not balanced with respect to `Day` (or `Time`). Consequently, the differences between pairs of adjusted `Location` means do not all have the same standard error. To proceed, the user should use Genstat's regression facilities:

```
MODEL Sqrtnmbr
FIT Day, Time, Location
PREDICT [PRINT=description,prediction,se; VCOVAR=VCovLoc] Location
```

This produces:

```
***** Regression Analysis *****

      Y variate: Sqrtnmbr
   Fitted terms: Constant, Day, Time, Location


*** Accumulated analysis of variance ***

Change              d.f.         s.s.         m.s.         v.r.
+ Day                  4       2.6869       0.6717         1.02
+ Time                 5      22.8525       4.5705         6.96
+ Location             9      60.0359       6.6707        10.15
Residual              11       7.2266       0.6570

Total                 29      92.8019       3.2001
```

and

```
*** Predictions from regression model ***

The predictions have been standardized by averaging fitted values
over the levels of some factors with the stated weighting policy:

         Factor   Weighting policy   Status of weights
           Time   Marginal weights   Constant over levels of other factors
            Day   Marginal weights   Constant over levels of other factors

      Table contains predictions followed by standard errors

      Y variate: Sqrtnmbr


     Location
            1       7.898       0.514
            2       7.263       0.514
            3       7.033       0.514
            4       7.064       0.514
            5       7.712       0.514
            6       8.212       0.514
            7       6.251       0.514
            8       3.596       0.514
            9       9.421       0.514
           10       9.287       0.514
```

These predictions are the adjusted `Location` means (adjusted for `Day` and `Time`) as obtained by Potthoff; standard errors of differences between these predictions can be calculated from the variance-covariance matrix `VCovLoc` obtained by use of the `VCOVAR` option of `PREDICT`. (All the covariances are very small. To two decimal places, every s.e.d. is 0.73.) In the accumulated analysis of variance, the sums of squares are for 'Days before fitting Locations', 'Times before fitting Locations', and 'Locations after fitting Days and Times'. Using the last of these, we can now complete Potthoff's *non-sequential* analysis of variance table [6, p. 204]:

Source of variation	d.f.	s.s.	m.s.
Days after fitting Locations	4	1.661	0.415
Times after fitting Locations	5	8.351	1.670
Locations after fitting Days and Times	9	60.036	6.671
Residual	11	7.227	0.657

Here, of course, the four sums of squares do not add up to the total sum of squares.

3. Block Designs with $2v$ Blocks and Two or More Sets of v Treatments, Each Set Being Balanced with Respect to Blocks and to the Other(s)

Akin to design (2) are the following two further three-factor designs, chosen from many similar examples given by Preece [8]:

	Block								
1	2	3	4	5	6	7	8	9	10
Aa	Bb	Cc	Dd	Ee	Aa	Bb	Cc	Dd	Ee
Cb	Dc	Ed	Ae	Ba	Bd	Ce	Da	Eb	Ac
De	Ea	Ab	Bc	Cd	Ec	Ad	Be	Ca	Db

(4)

and

	Block								
1	2	3	4	5	6	7	8	9	10
Aa	Bb	Cc	Dd	Ee	Aa	Bb	Cc	Dd	Ee
Cd	De	Ea	Ab	Bc	Be	Ca	Db	Ec	Ad
Dc	Ed	Ae	Ba	Cb	Eb	Ac	Bd	Ce	Da

(5)

Each of these two designs has the same number of treatments (five) in each set, but satisfies properties 1 and 2 as before. Also, with A and a (or B and b, etc.) described as 'corresponding' letters from the two sets, designs (4) and (5) satisfy modified properties 3 and 4 as follows:

3b) Each set of letters is balanced with respect to the other, as each letter of each set is paired equally often (once) with each non-corresponding letter of the other set.

4b) In the blocks containing any particular letter from either set, the non-corresponding letters of the other set are equally replicated, namely four times in (4) and three times in (5).

As a result of satisfying properties 1, 2, 3b and 4b, each of designs (4) and (5) is recognized as balanced by an analysis of variance with a single block factor and the two treatment factors. However, the difference between (4) and (5) that was noted in property 4b has the consequence that the two designs differ in their efficiency factors.

Each set of letters in (4) and (5) is arranged in a balanced incomplete block design whose e.f. is 5/6 = 0.833. In (4), the e.f. for either set drops to 20/27 = 0.741 when the effects of the other set are fitted as well as block effects. In (5), however, the e.f. for either set remains 5/6 = 0.833. In other words, design (5) has the striking property that the two sets of treatment effects become mutually orthogonal when each is adjusted for blocks. In the terminology of Eccleston and Russell [3,4], the two sets have 'adjusted orthogonality' but not 'pairwise orthogonality'.

Let us denote the two sets of treatments in (4) as T1 and T2, respectively, and those in (5) as T1 and T3. Then an analysis of (4) can be specified by:

```
BLOCKS Block
TREATMENTS T1 + T2
```

The Information summary then obtained from ANOVA will be:

```
***** Information summary *****

Model term                      e.f.  non-orthogonal terms

Block stratum
  T1                           0.167
  T2                           0.093  T1

Block.*Units* stratum
  T1                           0.833  Block
  T2                           0.741  Block  T1
```

The value of 0.833 given for T1 in the bottom stratum is the e.f. for T1 before fitting T2; to confirm that the e.f. for T1 after fitting T2 is 0.741, the analysis can be repeated with the treatment terms fitted in the opposite order. Once again, too, the tables of means obtainable from ANOVA arise from sequential fitting of the treatment terms. Thus, with T1 fitted before T2, the T1 means are unadjusted for T2.

An analysis of (5) can be specified by:

```
BLOCKS Block
TREATMENTS T1 + T3
```

The treatment terms here need be fitted in one order only, as each is orthogonal to the other after adjustment for blocks. In the Block stratum, the information on T3 is aliased with that on T1, and the only treatment information that Genstat records for that stratum is for T1. The user must deduce for him- or herself where the 'missing' information on T3 is hidden, and why the Block stratum of the analysis of variance seems to have no d.f. for T3.

To obtain greater insight into designs (4) and (5), we can view them as parts of the following five-factor design for four sets of treatments T1, T2, T3 and T4:

				Block					
1	2	3	4	5	6	7	8	9	10
Aa*Aa*	Bb*Bb*	Cc*Cc*	Dd*Dd*	Ee*Ee*	Aa*Aa*	Bb*Bb*	Cc*Cc*	Dd*Dd*	Ee*Ee*
Cb*De*	Dc*Ea*	Ed*Ab*	Ae*Bc*	Ba*Cd*	Bd*Ec*	Ce*Ad*	Da*Be*	Eb*Ca*	Ac*Db*
DeCb	Ea*Dc*	Ab*Ed*	Bc*Ae*	Cd*Ba*	Ec*Bd*	Ad*Ce*	Be*Da*	Ca*Eb*	Db*Ac*

(6)

If *any* two treatment-sets are omitted from (6), the resultant design is equivalent to either (4) or (5).

If three of the treatment-sets of (6) are retained, or indeed all four, the design remains balanced for all its treatment-sets. As all designs with just three of the sets are equivalent, we can restrict attention to the block design with T1, T2 and T3. Although (as we have already seen) the effects of T1 and T3 become mutually orthogonal after fitting blocks, they are not mutually orthogonal after fitting T2 as well as blocks. Accordingly, if adjusted means are needed for all three sets of treatments, each set of means being adjusted for all other factors, at least three TREATMENTS specifications must be used, each having a different treatment-factor last. Analyses so specified show that the efficiency factors for each treatment-set after fitting all other factors are these:

T1 and T3	T2
35/48 = 0.729	35/54 = 0.648

So far as I am aware, the literature of balanced designs does not indicate that different e.f.s can apply to different sets of equally replicated treatments when each set has the same number of treatments; this phenomenon can however readily arise when there are three sets of treatments.

For amusement and completeness, we may as well record that if all factors in (6) are retained, the e.f. for each treatment-set, after fitting all other factors, is 25/42 = 0.595. Intuition might well have suggested a worse value.

4. Block Designs with *v* Blocks and Two or More Sets of *v* Treatments, Each Set Being Balanced with Respect to Blocks and to the Other(s)

Hoblyn, Pearce and Freeman [5] gave versions of the following design for seven blocks each of four units:

	Block					
1	2	3	4	5	6	7
Ag	Be	Cb	Dc	Ec	Fb	Ge
Dd	Eb	Fa	Gb	Ad	Bf	Cf
Ff	Gg	Af	Bg	Ce	De	Eg
Ga	Aa	Bc	Cd	Da	Ed	Fc

(7)

In this three-factor design, each of the two sets of treatments is arranged in a balanced incomplete block design, each set of treatments is 'balanced' with respect to the other, and the blocks are 'balanced' with respect to each set of treatments. (Here, as in similar contexts below, 'balanced' connotes the same sort of balance as that of a balanced incomplete block design.) The easiest way of seeing that, for example, blocks and T1 are balanced with respect to T2 is to rewrite the design like this:

a	b	c	d	e	f	g
1G	2E	3B	4C	5C	6B	7E
2A	3C	4D	5A	6D	7C	1A
3F	4G	5E	6E	7G	1F	2G
5D	6F	7F	1D	2B	3A	4B

(8)

Here, the entry 1G in the column headed 'a' indicates that block 1 of the original design had G paired with treatment a; and so on. The sets of upper-case letters and of digits in the columns of (8) each constitute the blocks of a balanced incomplete block design, which confirms that T1 and the blocks of (7) are balanced with respect to T2.

With each factor of (7) being balanced with respect to any other, the design might well be expected to have balance overall. But, like design (1), it does not. However, the nature of the unbalance is not apparent from cursory inspection.

For a design that is balanced overall and in other respects similar to (7), sets of letters must be taken from the following:

	Block					
1	2	3	4	5	6	7
Aa*AaAa*	Bb*BbBb*	Cc*CcCc*	Dd*DdDd*	Ee*EeEe*	Ff*FfFf*	Gg*GgGg*
Bc*DeFg*	Cd*EfGa*	De*FgAb*	Ef*GaBc*	Fg*AbCd*	Ga*BcDe*	Ab*CdEf*
Ce*GbDf*	Df*AcEg*	Eg*BdFa*	Fa*CeGb*	Gb*DfAc*	Ac*EgBd*	Bd*FaCe*
Eb*FcGd*	Fc*GdAe*	Gd*AeBf*	Ae*BfCg*	Bf*CgDa*	Cg*DaEb*	Da*EbFc*

(9)

As shown by Dall'Aglio [2] and Preece [7], efficiency depends once again on which two sets of letters are chosen. Discussion of this and related matters is perhaps easiest if we regard the first factor of the design as a treatment-set T0 rather than as a blocking factor. Then we have seven possible seven-level factors T0, T1, ... , T6.

Two distinct three-factor designs are obtainable from (9). Firstly there is the design with factors T0, T1 and T2; as can be checked by ANOVA, its e.f. for each factor after fitting the other two is 3/4 = 0.750. Secondly we can use the factors T0, T1 and T3; then the e.f. for each factor after fitting the others is 5/8 = 0.625.

One further factor from (9) can be squeezed into a design if we are prepared to drop to 3 d.f. for error. Indeed, the distinct four-factor designs obtainable from (9) are again two in number. (These are obtainable as the complements of the two three-factor designs, but can alternatively be given as follows.) Firstly there is the design with factors T0, T1, T2 and T3; its efficiency factors for each treatment-set after fitting the others are these:

T0 and T3	T1 and T2
7/24 = 0.292	7/20 = 0.350

Secondly there is the design with T0, T1, T2 and T4, with the following efficiency factors:

T0	T1, T2 and T4
7/10 = 0.700	7/12 = 0.583

The reader is invited to obtain these values too by ANOVA.

4.1. Example 2

Potthoff [6, p. 205] gave data for the former of the above two designs for three factors. His levels of T0 were 'individuals' (people), those of T1 were 'shapes', and those of T2 were 'colours'. The original data were the individuals' estimates of the areas of some geometric figures of the designated shapes and colours. The variate analysed was the logarithm (to the base 10) of the ratio of estimated area to actual area, the logarithms being rounded to three decimal places:

Individual						
1	2	3	4	5	6	7
Aa	Bb	Cc	Dd	Ee	Ff	Gg
–0.113	–0.083	–0.182	0.076	–0.057	0.046	–0.297
Bc	Cd	De	Ef	Fg	Ga	Ab
–0.153	–0.246	–0.100	–0.137	–0.109	0.305	–0.414
Ce	Df	Eg	Fa	Gb	Ac	Bd
–0.085	–0.255	–0.191	0.074	–0.290	0.268	–0.384
Eb	Fc	Gd	Ae	Bf	Cg	Da
–0.045	–0.255	–0.063	0.118	–0.129	0.170	–0.360

(10)

Readers may like to use three sequential ANOVAs to obtain Potthoff's non-sequential analysis of variance table as follows:

Source of variation	d.f.	s.s.	m.s.
Individuals after fitting Shapes and Colours	6	0.68296	0.11383
Shapes after fitting Individuals and Colours	6	0.02450	0.00408
Colours after fitting Individuals and Shapes	6	0.06593	0.01099
Residual	9	0.07533	0.00837

5. Superimposed Youden squares, etc.

Arrangements such as (4), (5), (6) and (10) above can also be used as row-and-column designs, i.e. their rows (as printed) can be used as the blocks of a further blocking system. As all other factors are orthogonal to the rows, no new complication is introduced into the analysis by the rows; the d.f. for rows are merely removed from the Residual d.f. to a Row stratum of their own.

If (9) is regarded as a row-and-column array, then each set of letters there is disposed in a so-called 'Youden square' (a design which is *not* square, where each treatment occurs exactly once in each row, and where the columns are the blocks of a balanced incomplete block design with as many blocks as treatments). A row-and-column design formed with different sets of letters from (9) can thus be described as a balanced superimposition of Youden squares one upon another. Other such superimpositions can be obtained by deleting one row from a complete set of mutually orthogonal $v \times v$ Latin squares; an example with $v = 5$ is as follows:

Bc*De*	Cd*Ea*	De*Ab*	Ea*Bc*	Ab*Cd*
Ce*Bd*	Da*Ce*	Eb*Da*	Ac*Eb*	Bd*Ac*
Db*Ec*	Ec*Ad*	Ad*Be*	Be*Ca*	Ca*Db*
Ed*Cb*	Ae*Dc*	Ba*Ed*	Cb*Ae*	Dc*Ba*

(11)

In general for such $v \times (v-1)$ designs, the e.f. for any factor other than the Row factor, after fitting i other such factors, is

$$\frac{v(v-1-i)}{(v-1)(v-i)}.$$

In particular, the e.f. for the treatments of a simple $v \times (v-1)$ Youden square (no superimpositions) is obtained by taking $i = 1$ (for the Column factor).

6. Balanced Four-factor Designs with Two Independent Non-orthogonalities

Youden squares feature also in some row-and-column designs such as the following, due to Clarke [1]:

Aa	Bb	Cc	Dd	Ea
Bd	Ac	Eb	Ca	Db
Cb	Da	Ad	Ec	Bc
Dc	Ed	Ba	Ab	Cd

(12)

Let the two sets of treatments be T1 and T2 respectively. As the treatments of T1 are disposed in a Youden square, T1 and the columns are balanced with respect to one another. Additionally, T2 and the rows are balanced with respect to one another. Otherwise, the factors are orthogonal to one another. The design can thus be said to have 'two independent non-orthogonalities' (one involving T1 and columns, the other involving T2 and rows). Thus, to obtain fully adjusted means for both T1 and T2, only one analysis sequence is needed, the order of fitting the two sets of treatments being immaterial; the following specification will suffice:

```
BLOCKS      Row * Column
TREATMENTS T1 + T2
```

7. Implications for Genstat Users and for Genstat Itself

Before writing this paper, I had not explored the use of ANOVA for the designs that I have been discussing, was uncertain of its suitability for them, and – despite much experience of ANOVA – was still hazy about its handling of non-orthogonality amongst treatment factors. Now that the computing is complete, I find that ANOVA is an effective tool for providing the required analyses, and that the designs in question are very useful for illustrating how ANOVA handles balanced non-orthogonalities. There are, however, implications for Genstat users and for Genstat itself. (Almost identical considerations apply to the analysis of balanced change-over designs for the estimation of direct and residual effects.)

The implications all arise from the sequential fitting of the treatment terms in ANOVA. Except where there is adjusted orthogonality between these terms, recognition of the sequential fitting is crucial to correct interpretation of (a) the analysis of variance table, (b) the information summary, and (c) the tables of means. Of these three components of the output (as obtained from the version of Genstat 5 that I was using in the summer of 1987), only the information summary provides an implicit warning of problems of interpretation. This warning may well pass unrecognized by a user who is not well trained in statistics, and it does not indicate how the sums of squares and tables of means are calculated; the unsophisticated user may well be unaware of how a sum of squares depends on the sequence of fitting. So further annotation of the ANOVA output is needed to provide reasonable precautions against misinterpretation. (The Genstat 5 Manual indicates with great clarity that the fitting is sequential, but output is often used independently of the Manual.)

When I considered the analogous problem with non-orthogonal multiple-regression analyses [9, p. 37], I suggested printing messages such as

```
EACH SUM OF SQUARES CALCULATED
AFTER FITTING ALL TERMS ABOVE IT
```

with all analysis of variance tables to which they apply. Any modified such wording for analyses with more than one stratum would however be cumbersome, and the best expedient might be merely to print the symbol + at the start of each treatment line in the analysis, as in accumulated analyses of variance from the regression part of Genstat. This special printing could be triggered easily whenever the calculations for the information summary detect non-orthogonality between any treatment terms. An additional desirable feature would be a printed indication of any within-stratum aliasing such as occurs in the Block stratum of the analysis of design (5) above.

The correspondence of the printed efficiency factors and tabulated means to steps of a sequential process could perhaps best be signalled by the amplified headings:

```
***** Information summary (model terms fitted sequentially) *****
```

and

```
***** Tables of means (model terms fitted sequentially) *****
```

These precautions apart, a user who regularly analyses designs of the types in question might well be best served by a Genstat Procedure whose output is assembled from ANOVA having the required different sequences for the treatment terms. The Procedure could then print (a) an analysis of variance table where each sum of squares for a treatment term was calculated after fitting all other treatment terms, (b) an information summary where each efficiency factor is for the corresponding term when fitted last, and (c) tables of means that are all adjusted for all other treatment factors.

5. References

[1] Clarke, G.M.
A Second Set of Treatments in a Youden Square Design.
Biometrics, **19**, pp. 98-104, 1963.

[2] Dall'Aglio, G.
Blocs Incomplets Equilibres Orthogonaux.
Colloques Internat. C.N.R.S., **110**, pp. 195-214, 1963.

[3] Eccleston, J. and Russell, K.
Connectedness and Orthogonality in Multi-factor Designs.
Biometrika, **62**, pp. 341-345, 1975.

[4] Eccleston, J. and Russell, K.
Adjusted Orthogonality in Nonorthogonal Designs.
Biometrika, **64**, pp. 339-345, 1977.

[5] Hoblyn, T.N., Pearce, S.C. and Freeman, G.H.
Some Considerations in the Design of Successive Experiments in Fruit Plantations.
Biometrics, **10**, pp. 503-515, 1954.

[6] Potthoff, R.F.
Three-factor Additive Designs more General than the Latin Square.
Technometrics, **4**, pp. 187-208, 1962.

[7] Preece, D.A.
Some Row and Column Designs for Two Sets of Treatments.
Biometrics, **22**, pp. 1-25, 1966.

[8] Preece, D.A.
Some Balanced Incomplete Block Designs for Two Sets of Treatments.
Biometrika, **53**, pp. 497-506, 1966.

[9] Preece, D.A.
Genstat Analysis of Variance and the Distant Client.
Genstat Newsletter, **13**, pp. 35-38, 1984.

An Index for Genstat Newsletters 1-21

S J Welham
Rothamsted Experimental Station
Harpenden
Hertfordshire
United Kingdom *AL5 2JQ*

Introduction

The following index includes all articles (excepting editorials) running from the first issue of the Genstat Newsletter in December 1975 up to the present. It is written in three parts:

- the first section indexes all articles by author,
- the second categorises articles by subject, and
- the third lists all macros and procedures discussed in the Newsletter.

Specifically, the first section orders articles alphabetically by authors, and chronologically within authors. Any piece produced by joint authors appears under each author. There are also two columns of numbers associated with the articles: the first column numbers articles from 1 to 188, and thus identifies articles for use with the other parts of the index. In order that each piece has a unique number, those articles with more than one author are numbered under the main author (as credited in the Newsletter). The second column indicates the situation of the article in the series: [x.y] labels a piece as Newsletter x, Article y.

The second section lists the identifying numbers of articles under certain subject areas, and so indicates where articles in specific areas may be found.

The third section lists alphabetically all macros and procedures mentioned by name in any article, together with the identifying number(s) of the piece(s). It should be noted that not all of these articles include a listing of the procedure (although many do), and some give only a general description.

1. Index by Author

11 [12.04] Appleby, P.N.
Savage's Log-Rank Test.

12 [18.06] Arnold, G.M.
A Generalised Procrustes Macro for Sensory Analysis.

13 [9.08] Astier, R.
Post-Graduate Use of Genstat.

14 [10.09] Atkinson, J.
Genstat and Prime CPL.

15 [16.02] Atkinson, J.
A SIR/Genstat Interface.

16 [5.03] Baines, C.
Hints on the Use of Genstat – Orthogonal Polynomials.

17 [19.04] Baird, D.B.
A Genstat 5 Procedure for a First Difference Analysis.

18 [3.07] Banfield, C.
New Facilities for Genstat 3.09 – Multivariate Analysis

19 [4.02] Banfield, C.
New Library Macros – Multivariate Analysis of Variance

20 [4.02] Banfield, C.
New Library Macros – Starting Classification.

21 [5.02] Banfield, C.
New Library Macros.

22 [8.11] Barnard, J.
Duncan's Multiple Range Test.

23 [10.07] Barnard, J. and Norelli, J.
A Regression Model for Genotypical Stability.

24 [12.05] Barnard, J.
Canonical Analysis of a Response Surface.

[6.06] Bassill, L., Digby, P.G.N. and McLaren, N.
New Library Macros.

[3.01] Bell, G. and Miller, B.G.
Random Sampling of a Data Matrix.

25 [9.06] van den Bol, M.E.
The Use of Genstat at IWIS-TNO.

26 [20.03] van den Bol, M.E.
The Analysis of a Mixed Model.

27 [13.03] Bouvier, A.
Interfacing Genstat and a Database Management System.

28 [15.09] Bouvier, A.
Some Uses of the OWN Directive: Interfaces Between Genstat and Other Packages and Interruption of Genstat Sessions.

29 [8.12] ter Braak, C.J.F.
Dummy Covariates in Genstat.

30 [21.03] Brown, D., Marrs, R.W. and Walters, D.E.
A Genstat 5 Procedure for Robust Pairwise Comparisons in a One-way Design.

31 [8.08] Bryan-Jones, J.
Simply Reading Data.

32 [9.03] Bryan-Jones, J.
Genstat Macro Library

33 [9.10] Bryan-Jones, J.
A Conversational Approach to Using Genstat.

34 [11.03] Bryan-Jones, J.
The Genstat Macro Library.

35 [17.08] Bryan-Jones, J.
Case Study – a Fortran Influence.

36 [17.11] Bryan-Jones, J.
The Genstat Macro Library.

37 [12.06] Cole, T.J.
Indirect Calorimetry Analysed Using Genstat.

38 [17.10] Coursol, J.
Genstat 4.03E.

39 [13.04] Derobert, E.
Utilisation de Genstat pour les Traitment de Statistique de Donnees Fromageres.

40 [9.12] Dickson, J. and Hunter, E.A.
Displaying Residuals.

41 [6.06] Digby, P.G.N., Bassill, L. and McLaren, N.
New Library Macros.

42 [7.03] Digby, P.G.N.
Principal Coordinate Analysis of Grouped Data.

43 [7.05] Digby, P.G.N.
Avoiding VA 18.

44 [7.09] Digby, P.G.N.
Line Spacing with PRINT.

45 [8.03] Digby, P.G.N.
Reply to letter – Problems with PCP in Genstat.

[8.15] Digby, P.G.N. and Harding, S.A.
The INDSCAL Macro.

46 [10.08] Digby, P.G.N.
Non-Hierarchical Classification – CLASSIFY (and CLASSF).

47 [11.09] Digby, P.G.N.
Processing the Results of Multivariate Analysis.

48 [11.10] Digby, P.G.N.
Plotting Variables in Pairs.

49 [12.07] Digby, P.G.N. and Harding, S.A.
Rationalising the Macros for Multivariate Analysis.

50 [14.06] Digby, P.G.N.
Dendograms and Ziggurats.

51 [14.07] Digby, P.G.N.
Drawing Pretty Dendograms.

[15.07] Digby, P.G.N. and Lane, P.W.
Linking Fortran Subprograms into Genstat.

[19.03] Digby, P.G.N., Ainsley, A.E., Harding, S.A., Lane, P.W., Payne, R.W. and Simpson, H.R.
Conversion from Genstat 4 to Genstat 5.

[20.06] Digby, P.G.N. and Payne, R.W.
Genstat 5 Procedure Library: Instructions for Authors.

[11.04] Dixon, T.J. and Payne, R.W.
The Analysis of Experiments with Repeated Measures.

52 [11.05] Dixon, T.J. and Payne, R.W.
Macro REPMEAS.

53 [14.04] Dixon, T.J.
Genstat – An Easier Way to Learn.

[21.02] Downs, A.M., Heisterkamp, S.H., Jager, J.C. and van Druten, J.A.M.
Use of Genstat in the Estimation of Expected Numbers of AIDS Cases Adjusted for Reporting Delays.

[9.12] Hunter, E.A. and Dickson, J.
Displaying Residuals.

[8.10] Hurley, M. and Tunnicliffe Wilson, G.
Spectral Analysis.

[20.07] Iles, R.M.J., Lane, P.W. and Nelder, J.A.
Accessing the NAG Fortran Library from Within Genstat and Other Ways of Extending Genstat.

[21.02] Jager, J.C., Heisterkamp, S.H., van Druten, J.A.M. and Downs, A.M.
Use of Genstat in the Estimation of Expected Numbers of AIDS Cases Adjusted for Reporting Delays.

71 [16.07] Jagger, G., Whinney, K.A. and Fincham, J.
The Testing of Anti-Dandruff Shampoos – An Application of Genstat.

72 [21.04] Jansen, J.
Experiences with the Use of Genstat Regression Facilities when Fitting Extensions of Generalised Linear Models.

73 [9.04] John, J.A.
The Second Genstat Conference, Wageningen 1981 October 7-9.

74 [13.11] Keen, B.
Modelling in Genstat.

75 [16.06] Kempson, R.E.
A Genstat Program for General Block Designs.

76 [7.04] Keuls, M., Koops, W.J. and Thissen, J.T.N.M.
Hidden Connectedness and Analysis of Variance.

77 [8.13] Kirby, E.J.M.
Why I Like Genstat.

[7.04] Koops, W.J., Keuls, M. and Thissen, J.T.N.M.
Hidden Connectedness and Analysis of Variance.

78 [2.08] Lane, P.W.
New Regression Facilities.

79 [3.07] Lane, P.W.
New Facilities for Genstat 3.09 – Regression.

80 [4.02] Lane, P.W.
New Library Macros – Generalised Linear Models.

81 [4.02] Lane, P.W.
New Library Macros – Probit Analysis.

82 [4.02] Lane, P.W.
New Library Macros – Censored Data.

83 [5.01] Lane, P.W.
Some Changes in Release 4.02 – Changes to Regression Facilities.

84 [5.03] Lane, P.W.
Hints on the Use of Genstat – Using the FOR Directive.

85 [8.14] Lane, P.W.
New Genstat Books and Foreign Language Manuals.

86 [9.07] Lane, P.W.
Making Predictions from a Regression Model.

87 [9.11] Lane, P.W.
Absorption – A Method for Fitting Models with Many Parameters.

88 [9.13] Lane, P.W.
How to Print Structure Identifiers Without their Values.

89 [10.03] Lane, P.W., Payne, R.W. and Simpson, H.R.
Genstat 4.04.

90 [10.04] Lane, P.W. and Todd, A.D.
Backing Store Changes in Genstat 4.04.

111 [8.05] Nelder, J.A.
The Manipulation of Multi-way Tables in Genstat.

112 [14.05] Nelder, J.A.
Genstat: Origins and Prospects.

[20.07] Nelder, J.A. , Lane, P.W. and Iles, R.M.J.
Accessing the NAG Fortran Library from Within Genstat and Other Ways of Extending Genstat.

[10.07] Norelli, J. and Barnard, J.
A Regression Model for Genotypical Stability.

113 [17.03] Nunn, P.A.
Fitting and Assessing a Non-Linear Response Curve with Genstat.

114 [1.06] Payne, R.W.
Partial Aliasing and Confounding.

115 [3.07] Payne, R.W.
New Facilities for Genstat 3.09 – Analysis of Designed Experiments.

116 [4.02] Payne, R.W.
New Library Macros – Aliased Model Terms in ANOVA.

[5.03] Payne, R.W. and Alvey, N.
Hints on the Use of Genstat – Analysis of Covariance for Each Level of a Factor.

[10.03] Payne, R.W., Lane, P.W. and Simpson, H.R.
Genstat 4.04.

117 [11.04] Payne, R.W. and Dixon, T.J.
The Analysis of Experiments with Repeated Measures.

[11.05] Payne, R.W. and Dixon, T.J.
Macro REPMEAS.

118 [12.09] Payne, R.W.
Plans for Genstat Mark 5.

[19.03] Payne, R.W., Ainsley, A.E., Digby, P.G.N., Harding, S.A., Lane, P.W. and Simpson, H.R.
Conversion from Genstat 4 to Genstat 5.

119 [20.06] Payne, R.W. and Digby, P.G.N.
Genstat 5 Procedure Library: Instructions for Authors.

120 [6.02] Pearlman, J.
Addition of Time Series Facility.

121 [1.01] Phelps, K.
'Between and Within Groups' Analysis of Variance.

122 [1.02] Phelps, K.
The Beauties of JOIN.

123 [12.11] Philipe, O.
Essai de Modelisation des Relations Rendement-Peuplement Epis de Bles d'Hiver.

[9.05] Phillips, P. and Thompson, R.
Combination of Information from Different Strata.

124 [2.01] Pilcher, C.
Genstat at East Malling Research Station.

125 [13.08] Pistone, G. and Repetto, I.
Teaching Genstat to Undergraduate Students in Applied Mathematics in the University of Genoa, Italy.

[17.09] Pons, O. and de Turctheim-Lesquoy, E.
Modelling the Feeding Pattern of Rabbits with Cox's Regression Model.

126 [2.03] Potter, F.
Setting Values Missing.

127 [17.05] Poultney, R.F.A. and Riley, J.
A Genstat Macro for the Bivariate Analysis of Intercropping Data.

128 [13.10] Preece, D.A.
Genstat Analysis of Variance and the Distant Client.

129 [17.06] Preece, D.A.
The Use of Pseudo-Factors when Treatments were Superimposed in an Orchard Experiment.

130 [19.05] Preece, D.A.
The Use of Pseudo-Factors for a Balanced 6x6 Row-and-Column Design for Nine Treatments.

131 [21.05] Preece, D.A.
Genstat Analyses for Complex Balanced Designs with Non-interacting Factors.

[13.08] Repetto, I. and Pistone, G.
Teaching Genstat to Undergraduate Students in Applied Mathematics in the University of Genoa, Italy.

132 [7.02] Richardson, M.G. and Simpson, H.R.
Status Report (March 1981).

133 [8.04] Richardson, M.G. and Simpson, H.R.
Status Report (September 1981).

134 [9.02] Richardson, M.G. and Simpson, H.R.
Status Report (March 1982).

135 [10.02] Richardson, M.G. and Simpson, H.R.
Genstat 4.03 Status Report.

136 [11.02] Richardson, M.G. and Simpson, H.R.
Status Report (March 1983).

137 [14.02] Richardson, M.G. and Simpson, H.R.
Status Report (October 1984).

138 [17.07] Richardson, M.G.
Survey of Genstat Users – Preliminary Report.

139 [18.03] Richardson, M.G.
Survey of Genstat Users.

140 [20.05] Ridout, M.S.
Prediction on the Scale of the Linear Predictor for Generalised Linear Models.

141 [15.04] Riley, J.
Genstat by Post.

142 [16.05] Riley, J.
A Genstat Analysis for Intercropping Stability.

[17.05] Riley, J. and Poultney, R.F.A.
A Genstat Macro for the Bivariate Analysis of Intercropping Data.

143 [18.07] Rogers, C.A.
A Genstat Macro for Partial Least Squares Analysis with Cross-Validation Assessment of Model Dimensionality.

144 [4.01] Ross, G.J.S.
Optimisation in Genstat.

145 [11.08] Ross, G.J.S.
Fitting General Models with OPTIMISE.

146 [15.02] Shea, B.L.
Letter – The Role of a Subroutine Library for Statisticians.

147 [15.10] Sherington, J. and Gilson, D.
Efficient Performance of Genstat on a VAX.

148 [3.06] Shih, T.M. and Whittaker, J.
A Genstat Program for the Iterative Scaling Algorithm.

149 [1.04] Simpson, H.R.
The Secondary Output Channel in Genstat.
150 [2.04] Simpson, H.R.
Termination Codes.
151 [2.05] Simpson, H.R.
Text Indicator.
152 [2.06] Simpson, H.R.
Restricting the Unrestrictable.
153 [3.02] Simpson, H.R.
Directory Full!
154 [3.07] Simpson, H.R.
New Facilities for Genstat 3.09 – Syntax and Program Structure.
155 [5.03] Simpson, H.R.
Hints on the use of Genstat – Restricting the Unrestrictable – again.
156 [6.11] Simpson, H.R.
Inconsistency – or Things Ain't What They Seem to Be.
[7.02] Simpson, H.R. and Richardson, M.G.
Status Report (March 1981).
157 [7.07] Simpson, H.R.
When READ fails...
[8.04] Simpson, H.R. and Richardson, M.G.
Status Report (September 1981).
158 [8.07] Simpson, H.R.
Comment – Points of General Interest to Macro Writers.
159 [8.16] Simpson, H.R.
Genstat Error Notice S1020.
[9.02] Simpson, H.R. and Richardson, M.G.
Status Report (March 1982).
160 [9.14] Simpson, H.R.
Genstat Error Notice S2010.
[10.02] Simpson, H.R. and Richardson, M.G.
Genstat 4.03 Status Report.
[10.03] Simpson, H.R., Lane, P.W. and Payne, R.W.
Genstat 4.04.
161 [10.10] Simpson, H.R.
Genstat Error Notice S3010.
[11.02] Simpson, H.R. and Richardson, M.G.
Status Report (March 1983).
162 [11.11] Simpson, H.R.
Genstat Error Notice S4010.
[14.02] Simpson, H.R. and Richardson, M.G.
Status Report (October 1984).
163 [14.09] Simpson, H.R.
Manual and Notice Board Amendments.
164 [16.09] Simpson, H.R.
Macro Library, Manual and Notice Board Amendments.
[19.03] Simpson, H.R., Ainsley, A.E., Digby, P.G.N., Harding, S.A., Lane, P.W. and Payne, R.W.
Conversion from Genstat 4 to Genstat 5.
165 [8.02] Sparrow, P.E. and Franks, C.R.
Letter – Problems with PCP in Genstat.
166 [11.07] Stevens, G.V.
Box-Cox Transformation.

167 [12.03] Stevens, G.V.
Reply to letter [12.02].
168 [12.12] Talbot, H.
How Steep is the Genstat Learning Curve?
169 [20.02] Talbot, H.
Changing from Genstat 4 to Genstat 5.
[3.07] Tett, P. and Alvey, N.
New Facilities for Genstat 3.09 – Basic Data Operations.
[7.04] Thissen, J.T.N.M., Keuls, M. and Koops, W.J.
Hidden Connectedness and Analysis of Variance.
170 [16.03] Thompson, C.J.
Genstat in New Zealand.
171 [9.05] Thompson, R. and Phillips, P.
Combination of Information from Different Strata.
172 [10.06] Thompson, R.
Diallel Crosses, Partially Balanced Incomplete Block Designs with Triangular Association Schemes and Rectangular Lattices.
173 [12.13] Thompson, R.
Multiple copies, the Genstat Analysis of Variance and Neighbour Analyses.
174 [13.12] Thomsen, L.K. and Thyregod, P.
Poisson Models for the Analysis of Road Traffic Accidents.
[16.08] Thomsen, L.K. and Engel, U.
An Enquiry into the Relation of Accident Numbers to Traffic Flow and Vehicle Speeds.
[13.12] Thyregod, P. and Thomsen, L.K.
Poisson Models for the Analysis of Road Traffic Accidents.
175 [5.01] Todd, A.D.
Some Changes in Release 4.02 – Small Userfiles.
176 [6.10] Todd, A.D.
Simple Transfer of Data Structures to Disc and Magnetic Tape.
177 [6.12] Todd, A.D.
The Saving of Punching and Programming Time in the Analysis of Experiments.
178 [6.14] Todd, A.D.
Suggested Changes to Backing Store.
[10.04] Todd, A.D. and Lane, P.W.
Backing Store Changes in Genstat 4.04.
179 [17.12] Trinder, K.I.
Genstat and Workstations.
180 [8.10] Tunnicliffe Wilson, G. and Hurley, M.
Spectral Analysis.
181 [18.05] Tunnicliffe Wilson, G.
The Finite Fourier Transform.
182 [17.09] de Turckheim-Lesquoy, E. and Pons, O.
Modelling the Feeding Pattern of Rabbits with Cox's Regression Model.,
183 [15.08] Underhill, L.G.
An Implementation of the Genstat OWN Directive.
[21.03] Walters, D.E., Brown, D. and Marrs, R.W.
A Genstat 5 Procedure for Robust Pairwise Comparisons in a One-way Design.
184 [6.03] Watson, S.
A New Directive for Drawing Histograms.
185 [6.04] Watson, S.
Smooth Lines in GRAPH.

2. Index by Subject

Information on Genstat Releases:

1, 6, 8, 10, 18, 38, 78, 79, 83, 89, 90, 91, 98, 115, 120, 132, 133, 134, 135, 136, 137, 154, 159, 160, 161, 162, 163, 164, 175, 179, 184, 185

Genstat 5:

1, 17, 30, 72, 92, 94, 118, 119, 130, 131, 140, 169

3. Index of Macros and Procedures