

REPEATED TESTS ON SERIALY CORRELATED DATA



David Norman Geary

Lincoln College

A thesis submitted to the Faculty of Biological Sciences for the degree of Doctor of
Philosophy of the University of Oxford

Michaelmas term, 1986

Department of Biomathematics

5 South Parks Road

REPEATED TESTS ON SERIALY CORRELATED DATA

David Norman Geary, Lincoln College

D Phil thesis, Michaelmas term 1986

Abstract

Suppose two treatments are being compared in a clinical trial with a fixed number of subjects. Half the subjects receive treatment 1, while the other half receive treatment 2. Measurements of some relevant response variable are made on each subject at equally spaced time points. This thesis proposes a procedure for testing at each time point the null hypothesis H_0 of no treatment differences, while controlling the overall probability of type I error.

Chapter 1 describes alternative methods for estimating the unknown parameters of models generating short series of measurements. Maximum likelihood estimation, and modified variance-based estimation, are considered in 1.1 and 1.2 respectively. A 4-parameter model is described in 1.3. This allows for increasing or decreasing variances of measurements within subjects. In 1.4, maximum likelihood estimation is used in fitting the 4-parameter model to two data sets.

In Chapter 2, attention focuses on sequential testing, with data arising from the 4-parameter model. The possibility of using stopping rules based on posterior probabilities, rather than sequentially adjusted P values, is considered briefly in 2.1. Two methods for determining appropriate nominal significance levels for successive tests, when the model for data is assumed known, are described in 2.2. An algorithm for integrating quickly a special case of the multivariate normal density, in high dimensions, presented in 2.3, facilitates implementation of the two methods. In 2.4, test procedures based on the methods of 2.2 are described, in which model parameters are (re-) estimated at each time point, as data accumulate. Test statistics and z-values are adjusted for each test, according to the updated parameter estimates. Simulations show that these procedures give overall significance levels close to those required. One of the two procedures is preferred on the basis of power to detect deviations from H_0 .

To my parents
with love

Acknowledgements

I am especially grateful to Professor Peter Armitage for his steady supervision and guidance of this work.

Thanks to Keith Rennolls and The Forestry Commission for introduction and access to the "Interactive Minimizer", a computer program adapted for use in the maximum likelihood estimation of Chapters 1 and 2; to John Matthews for help with the proof of the conjecture in 1.1, and for often drawing my attention to relevant publications; to Mrs H V Worthington for the tooth decay data analysed in 1.4 and 2.4; to Patty Solomon for the blood pressure data analysed in 1.4; to Mahesh Parmar for loan of Schervish's multivariate normal integration algorithm, used in 2.2 to 2.4; to Dr F H C Marriott for help in understanding the unexpected results of 2.3; to Professor C R Heathcote for a useful discussion on the adaptation of functional least squares estimation, described in Appendix 3; to Hugh Possingham, John Matthews, Mahesh Parmar and Janet Spiller for constructive comments on reading the draft of the thesis; to Mrs Nancy Amery, Mrs Pat Searle and Mrs Allison Margetts, for assistance in administrative and computer matters; and to the staff of Daily Information, Oxford, for help in using their word-processing facilities.

I gratefully acknowledge receipt of a $2\frac{1}{4}$ -year Science Research Studentship from the Department of Education for Northern Ireland.

Contents

	page
Abstract	ii
Dedication	iii
Acknowledgements	iv
Contents	v
 Introductory remarks	 1
 CHAPTER 1 MODELS FOR SERIALLY CORRELATED DATA, AND ESTIMATION OF MODEL PARAMETERS	 2
1.1 Maximum likelihood estimation of the parameters of models used by Wilson, Hebel and Sherwin (1981), and Mansour, Nordheim and Rutledge (1985)	2
<i>The Wilson, Hebel and Sherwin model, and maximum likelihood estimation</i>	<i>2</i>
<i>The Mansour, Nordheim and Rutledge model, and maximum likelihood estimation</i>	<i>10</i>
<i>Proof of the Mansour, Nordheim and Rutledge conjecture</i>	<i>12</i>
1.2 Modified variance-based estimation	16
<i>Variance-based estimation</i>	<i>16</i>
<i>Modified variance-based estimator φ^{1*}</i>	<i>17</i>
<i>Modified variance-based estimators φ^{2*} and φ^{3*}</i>	<i>18</i>
1.3 Generalization of the models of Wilson, Hebel and Sherwin, and Mansour, Nordheim and Rutledge, by introduction of an extra parameter	25
<i>The $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model</i>	<i>25</i>
<i>Kenward's (1981) split plot autoregressive model, and its relationship to the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model</i>	<i>27</i>
<i>Simulation results using the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model</i>	<i>28</i>
1.4 Data analyses using the generalized model	34
<i>Analysis of dental caries data</i>	<i>34</i>
<i>Analysis of blood pressure data</i>	<i>38</i>
 CHAPTER 2 SEQUENTIAL TESTING WITH SERIALLY CORRELATED DATA	 42
2.1 Sequential stopping rules: classical or Bayesian?	42
2.2 Two methods for determining appropriate nominal significance levels, when the model for data is known a priori	47
<i>Group sequential testing with independent measurements</i>	<i>47</i>

	page
<i>Nominal significance levels for repeated measurements data: the integration method</i>	49
<i>Nominal significance levels for repeated measurements data: the transformation method</i>	53
2.3 Recursive integration of the multivariate normal density in a special case	59
<i>Recursive integration in a special case</i>	59
<i>Implementation of recursive integration</i>	65
2.4 Testing in practice, when the model for data is unknown a priori	68
<i>Three possible procedures</i>	68
<i>Simulation results</i>	72
<i>Sequential testing with dental caries data</i>	77
Concluding remarks	81
Appendix 1 Comparison between the Wilson, Hebel and Sherwin weighted and unweighted least squares estimators of the between-subject variance σ_a^2 (cf 1.1, p3)	83
Appendix 2 Results relating to the modified variance-based estimation of 1.2	86
(i) <i>Proof that, when $m = 3$, $-1 \leq \varphi_i^* \leq 3$, where φ_i^* is the WHS estimate of φ from the i^{th} subject (cf 1.2, p17)</i>	86
(ii) <i>Derivation of an approximate formula (as a function of φ) for the variance of the WHS estimator φ^*, when $m = 3$, and comparison of results from this formula with empirical variances obtained from simulations (cf 1.2, p19)</i>	87
(iii) <i>Modified variance-based estimator φ_{avad}^* (cf 1.2, p24)</i>	91
(iv) <i>Modified variance-based estimators φ_a^*, φ_b^* and φ_c^* (cf 1.2, p24)</i>	93
Appendix 3 The robust estimation by functional least squares of the autoregressive parameter in N short time series (cf 1.2, p24)	95
Appendix 4 The simulation program of 2.4 (cf p73)	98
References	103

Introductory remarks

Sequential analysis was developed during the Second World War as a technique for reducing the sample sizes required for quality control inspection plans. The book by Armitage (1960), "*Sequential Medical Trials*", helped popularize the idea of sequential testing in clinical trials. Implementation of a sequential design in a trial, rather than a fixed sample size design, allows the possibility of early stopping and so may be both more economical and more ethical, saving both money and lives.

At first, theory for sequential testing in a clinical trial was developed assuming that a test for significant treatment differences would be carried out each time a subject, or matched pair of subjects, completed the trial. Since the mid-1970's interest has focused more on testing after each *group* of subjects has finished the trial (see, for example, McPherson, 1974; Pocock, 1977). This is more convenient in practice.

In a "group sequential" trial, the maximum number of groups is determined at the outset. The relationship between the overall significance level and the "nominal significance levels" at successive tests has often been studied. Thus, for instance, to maintain an overall level of 0.05, in testing with constant nominal levels up to a maximum of 5 times, with normally distributed data, a nominal level of 0.0158 is required.

In group sequential testing, it is usually assumed that individual measurements are statistically independent. This may be reasonable when each subject gives rise to only one measurement. Often, however, the situation arises in which the number of subjects on each treatment is fixed, and at successive time points a measurement is made on each of these subjects. What are the appropriate nominal significance levels in this situation?

This question is addressed in a paper by Armitage, Stratton and Worthington (1985). They indicate the need to adjust nominal levels to allow for the ratio of between-subject to within-subject variance, and to allow for possible autocorrelation between successive measurements within subjects. The purpose of this thesis has been to develop further the work begun by Armitage et al.

CHAPTER 1 MODELS FOR SERIALLY CORRELATED DATA, AND

ESTIMATION OF MODEL PARAMETERS

1.1 Maximum likelihood estimation of the parameters of models used by Wilson, Hebel and Sherwin (1981), and Mansour, Nordheim and Rutledge (1985)

The Wilson, Hebel and Sherwin model, and maximum likelihood estimation

Throughout the thesis, the word *subject* is used of a *patient* or *individual* involved in a trial or experiment, and the word *measurement* is used of the random variable sometimes referred to as a subject *response* or *observation*.

Successive measurements on the same subject may depend on one another. One simple way of modelling this dependence is through a first-order autoregression. Wilson, Hebel and Sherwin (WHS, 1981) consider a model for m repeated measurements on each of N subjects, assuming within-subject errors for the i^{th} subject, $\{v_{it}, t=1, \dots, m\}$, follow a first-order autoregression,

$$v_{it} = \begin{cases} \epsilon_{i1} & , \quad t = 1 \\ \phi v_{i,t-1} + \epsilon_{it} & , \quad t = 2, \dots, m \end{cases} \quad (1)$$

Here, ϕ is the *autoregression coefficient*, ϵ_{i1} is normally distributed as $N(0, \sigma_e^2/(1-\phi^2))$, and $\{\epsilon_{it}, t=2, \dots, m\}$ are independent of ϵ_{i1} and are independent and identically distributed (iid) $N(0, \sigma_e^2)$. Subject means $\{\mu_i, i=1, \dots, N\}$ are assumed to be independent of $\{\epsilon_{it}\}$, and independent and normally distributed, $\mu_i \sim N(\mu, \sigma_a^2)$. In what follows, we shall refer to μ , σ_a^2 and σ_e^2 as the *overall mean*, the *between-subject variance* and the *within-subject variance* respectively.

Let $\underline{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$ be the measurement vector for the i^{th} subject from a sample of N subjects. Then, under the WHS model, $\underline{X}_i$ has a multivariate normal distribution,

$$\left. \begin{aligned} \underline{X}_i &\sim N(\underline{\mu} \underline{1}_m, \underline{\Sigma}) ; \text{ where } \underline{\Sigma} = \underline{u} \underline{u}' + \underline{V}, \quad \underline{u} = \sigma_a \underline{1}_m \\ \text{and } \underline{V} &= \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 & \dots & \varphi^{m-1} \\ \varphi & 1 & \varphi & \dots & \varphi^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{m-1} & \varphi^{m-2} & \varphi^{m-3} & \dots & 1 \end{bmatrix} \end{aligned} \right\} \quad (2)$$

Thus the WHS model has 4 parameters: μ , σ_a^2 , φ and σ_e^2 .

The log likelihood of data from the model is written down by WHS (p560; note that they use different notation for the parameters, and in their equation (23) for $|\underline{\Sigma}|$, the right hand side should be divided by $1-\beta^2$). To determine maximum likelihood estimates (mles) of the 4 parameters, WHS minimize the negative log likelihood with respect to the parameters. They find the 4-parameter minimization "*at least moderately unstable even with the use of good starting values*" (p560). They try estimating the overall mean μ and between-subject variance σ_a^2 by unweighted least squares,

$$\tilde{\mu}_i = \frac{1}{m} \sum_{j=1}^m X_{ij}, \quad \tilde{\mu} = \frac{1}{N} \sum_{i=1}^N \tilde{\mu}_i, \quad \tilde{\sigma}_a^2 = \frac{1}{N-1} \sum_{i=1}^N (\tilde{\mu}_i - \tilde{\mu})^2 - \frac{\sigma_e^{*2}}{m}, \quad (3)$$

where σ_e^{*2} is their "*variance-based*" estimator of the within-subject variance (see 1.2). (Note that WHS imply (p561) that their weighted least squares estimator of σ_a^2 (σ_μ^{*2} , of p560, (29)) is more efficient than the unweighted $\tilde{\sigma}_a^2$. This suggestion has been investigated empirically and found to be untrue. In Appendix 1, the weighted and unweighted estimators are compared, and empirical results are presented.) When $\tilde{\mu}$ and $\tilde{\sigma}_a^2$ are substituted for μ and σ_a^2 in the log likelihood, and the minimization is attempted on the remaining two parameters φ and σ_e^2 , the minimization process is reported as being "*quite stable when provided with good starting values*" (p561). Each of nine simulation studies performed by WHS using the 2-parameter minimization had 500 completed experiments (with $N = 100$ subjects per experiment, and $m = 3, 5$ or

10 repeated measurements per subject). However, experiments were aborted if reasonable convergence was not attained in the minimization process, and "*The number of aborted experiments in the various studies ranged up to several hundreds*" (p562).

The convergence problems encountered by WHS have been overcome using a minimization algorithm from the NAG library, E04KBF. This is a comprehensive quasi-Newton algorithm for finding a minimum of a function F of several variables, $\underline{\alpha}$ say, subject to fixed upper and lower bounds on the variables. The user must supply a starting point $\underline{\alpha}_0$, and first derivatives $\partial F/\partial \alpha_j$ at any point $\underline{\alpha}$.

In this application of E04KBF, $\underline{\alpha}$ was taken as $(\mu, \sigma_a^2, \varphi, \sigma_e^2)'$; μ was unbounded, σ_a^2 was bounded below by 0.0 and unbounded above, φ was bounded below by -0.99 and bounded above by 0.99 , and σ_e^2 was bounded below by 0.001 and unbounded above. The point $(\tilde{\mu}, \tilde{\sigma}_a^2, \varphi^*, \sigma_e^{*2})'$ (where $\tilde{\mu}$ and $\tilde{\sigma}_a^2$ are as defined in (3), and φ^* and σ_e^{*2} are the variance-based estimators discussed in 1.2) was used as the starting point for $\underline{\alpha}$, as in WHS.

Nine simulation studies, corresponding to those of WHS, were carried out. Each study consisted of 500 experiments. For each experiment, data were simulated using the NAG routines G05EAF and G05EZF. These routines generate a vector from a multivariate normal distribution with a given mean vector and covariance matrix. For each experiment, 100 vectors were generated, using the appropriate mean and covariance matrix (see (2)). Before generating any data, the NAG routine G05CCF was called to set the basic generator routine to a non-repeatable initial state. Parameter values were set at $\mu = 10.0$, $\sigma_a^2 = 1.0$ and $\sigma_e^2 = 0.6$; and the nine studies corresponded to all combinations of $\varphi = 0.2, 0.5, 0.8$ and $m = 3, 5, 10$.

Simulation results are presented in Table 1. Results for the mean μ are omitted from the table because the average, standard deviation and root mean squared error of the unweighted least squares estimator (the sample mean) of μ and the mle of μ were equal, to 2 decimal places, and very close to 10.0, in all 9 simulation studies.

Table 1 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of σ_a^2 -constrained maximum likelihood (ML_C) and unweighted least squares/variance-based (UL/VB) estimates of σ_a^2 , φ and σ_e^2*

Each simulation study consisted of 500 experiments, each conducted with $N=100$ subjects, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.

NBC = number of experiments for which the mle of the between-subject variance σ_a^2 was constrained at 0.0

/ indicates which of ML_C and UL/VB is better with respect to the RMSE criterion

Study	Par	ML_C			UL/VB		
		AV	SD	RMSE	AV	SD	RMSE
1 $m=3, \varphi=0.2$ NBC=2	σ_a^2	0.947	0.232	0.238	1.064	0.181	0.192/
	φ	0.208	0.161	0.161/	0.207	0.165	0.165
	σ_e^2	0.601	0.084	0.084/	0.600	0.084	0.084
2 $m=3, \varphi=0.5$ NBC=40	σ_a^2	0.893	0.403	0.417	1.295	0.215	0.365/
	φ	0.487	0.177	0.177/	0.489	0.192	0.192
	σ_e^2	0.593	0.081	0.081/	0.593	0.084	0.085
3 $m=3, \varphi=0.8$ NBC=172	σ_a^2	0.920	0.835	0.839/	2.215	0.343	1.262
	φ	0.751	0.140	0.148/	0.793	0.196	0.196
	σ_e^2	0.582	0.068	0.071/	0.595	0.082	0.082
4 $m=5, \varphi=0.2$ NBC=0	σ_a^2	0.981	0.170	0.171/	1.047	0.168	0.175
	φ	0.199	0.067	0.067/	0.201	0.099	0.099
	σ_e^2	0.598	0.045	0.045/	0.599	0.059	0.059
5 $m=5, \varphi=0.5$ NBC=0	σ_a^2	0.967	0.218	0.220/	1.228	0.200	0.303
	φ	0.496	0.080	0.080/	0.501	0.118	0.118
	σ_e^2	0.596	0.044	0.044/	0.599	0.058	0.058
6 $m=5, \varphi=0.8$ NBC=87	σ_a^2	0.906	0.621	0.628/	2.107	0.332	1.155
	φ	0.789	0.069	0.070/	0.797	0.128	0.128
	σ_e^2	0.597	0.044	0.044/	0.600	0.056	0.056
7 $m=10, \varphi=0.2$ NBC=0	σ_a^2	0.995	0.149	0.149/	1.036	0.150	0.154
	φ	0.200	0.039	0.039/	0.199	0.062	0.062
	σ_e^2	0.597	0.029	0.029/	0.597	0.038	0.038
8 $m=10, \varphi=0.5$ NBC=0	σ_a^2	0.977	0.174	0.176/	1.141	0.173	0.223
	φ	0.501	0.038	0.038/	0.505	0.070	0.070
	σ_e^2	0.602	0.030	0.031/	0.605	0.039	0.039
9 $m=10, \varphi=0.8$ NBC=4	σ_a^2	0.952	0.361	0.364/	1.854	0.272	0.896
	φ	0.798	0.037	0.037/	0.799	0.079	0.079
	σ_e^2	0.598	0.029	0.029/	0.599	0.038	0.038

In 7 out of 9 of the simulation studies of Table 1, the RMSE of the mle $\hat{\sigma}_a^2$ of the between-subject variance σ_a^2 was better (less) than that of the unweighted least squares estimator $\tilde{\sigma}_a^2$; exceptions were the studies with $m=3$, $\varphi=0.2$ and $m=3$, $\varphi=0.5$. The estimator $\hat{\sigma}_a^2$ is consistently negatively biased, with bias increasing as φ increases. The variance of $\hat{\sigma}_a^2$ also increases as φ increases, but decreases as m increases.

Mles $\hat{\varphi}$ and $\hat{\sigma}_e^2$ are better than variance-based estimators (vbes) φ^* and σ_e^{*2} in all 9 studies of Table 1: vbes are practically unbiased, but mles have smaller variance.

In the minimizations giving rise to Table 1, the values of the derivatives of the negative log likelihood with respect to the 4 parameters, at the solutions, were all less than 0.001 in absolute value (confirming the location of (minimum) turning points), except at constrained solutions. Absolute values of derivatives with respect to constrained parameters often exceeded 0.001 at solutions.

Only one constraint was operational for the minimizations giving rise to Table 1: namely, the lower bound of 0.0 on σ_a^2 . This bound was imposed because the log likelihood has a term involving

$$\log [1 + \sigma_a^2(1-\varphi)(m-m\varphi+2\varphi)/\sigma_e^2] . \quad (4)$$

However it is evident from (4) that the bound may be relaxed from 0.0 without necessarily causing the likelihood to become undefined. For those experiments in Table 1 which left $\hat{\sigma}_a^2$ constrained at 0.0, the minimization with E04KBF was resumed using a very large negative lower bound (-100000). At times the algorithm attempted to test parameter values erring into the region leading to a negative argument in (4); but at such times, the function value was left unchanged and control was returned to E04KBF, which proceeded to try different parameter values until eventually the minimum, or constrained minimum, was found. In these circumstances a different constraint came into play: the upper bound of 0.99 on φ . This was imposed because the log likelihood has a term involving $\log (1-\varphi^2)$. The effect on the RMSE's of $\hat{\sigma}_a^2$, $\hat{\varphi}$ and $\hat{\sigma}_e^2$ of allowing $\hat{\sigma}_a^2$ to be negative is illustrated by empirical results in Table 2, for 3 of the simulation studies of Table 1.

Table 2 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of $\hat{\sigma}_a^2$ -constrained mles (ML_C) and $\hat{\sigma}_a^2$ -unconstrained mles (ML) of σ_a^2 , φ and σ_e^2*

Each study consisted of 500 experiments, each conducted with $N=100$ subjects, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.
NBL = number of experiments for which the mle of the between-subject variance σ_a^2 was less than 0.0
NPC = number of experiments for which the autoregression constant φ was constrained at 0.99 ($\hat{\sigma}_a^2$ -unconstrained)
ANF = average number of function evaluations made by E04KBF in the minimization process ($\hat{\sigma}_a^2$ -unconstrained)
/ indicates which of ML_C and ML is better with respect to the RMSE criterion

Study	Par	ML _C			ML		
		AV	SD	RMSE	AV	SD	RMSE
2 m=3, $\varphi=0.5$ NBL=40 NPC=2, ANF=29	σ_a^2	0.893	0.403	0.417/	0.644	2.604	2.629
	φ	0.487	0.177	0.177/	0.492	0.187	0.187
	σ_e^2	0.593	0.081	0.081/	0.595	0.084	0.084
3 m=3, $\varphi=0.8$ NBL=172 NPC=77, ANF=82	σ_a^2	0.920	0.835	0.839/	-4.415	11.527	12.736
	φ	0.751	0.140	0.148/	0.778	0.168	0.169
	σ_e^2	0.582	0.068	0.071/	0.591	0.075	0.076
6 m=5, $\varphi=0.8$ NBL=87 NPC=5, ANF=36	σ_a^2	0.906	0.621	0.628/	0.276	3.467	3.541
	φ	0.789	0.069	0.070/	0.796	0.081	0.081
	σ_e^2	0.597	0.044	0.044/	0.599	0.046	0.046

Before discussing the results of Table 2, note that for simulations in which $\hat{\sigma}_a^2$ was constrained at 0.0, an alternative to resuming the minimization with a smaller lower bound on σ_a^2 was considered. This consisted in using the constrained mles as initial values in a routine to solve the 4 simultaneous non-linear equations obtained by setting the partial derivatives (with respect to the 4 parameters) of the negative log likelihood to zero. The NAG routine C05PCF was used to try and solve the 4 equations. It is a comprehensive routine to find a zero of a system of NF nonlinear functions in NF variables by a modification of the Powell hybrid method. For this application, NF=4. The user must provide the Jacobian; which in this case means supplying the second partial derivatives of the negative log likelihood $-L$,

$$\text{i.e. } -\partial^2 L/\partial \mu^2, \quad -\partial^2 L/\partial \mu \partial \sigma_a^2, \quad -\partial^2 L/\partial \mu \partial \varphi, \quad \text{etc.}$$

These derivatives were determined analytically and were checked using C05ZAF, a NAG routine designed to check user-supplied gradients of functions for consistency with the functions themselves. The derivatives were further checked using the NAG numerical differentiation algorithm D04AAF. Both checks confirmed the user-supplied derivatives to be correct (though the C05ZAF check had left room for some doubt, probably because of rounding error in the function and derivative evaluations). However, C05PCF did not always converge to the solution and convergence seemed highly dependent on the parameter starting values. This approach was eventually abandoned.

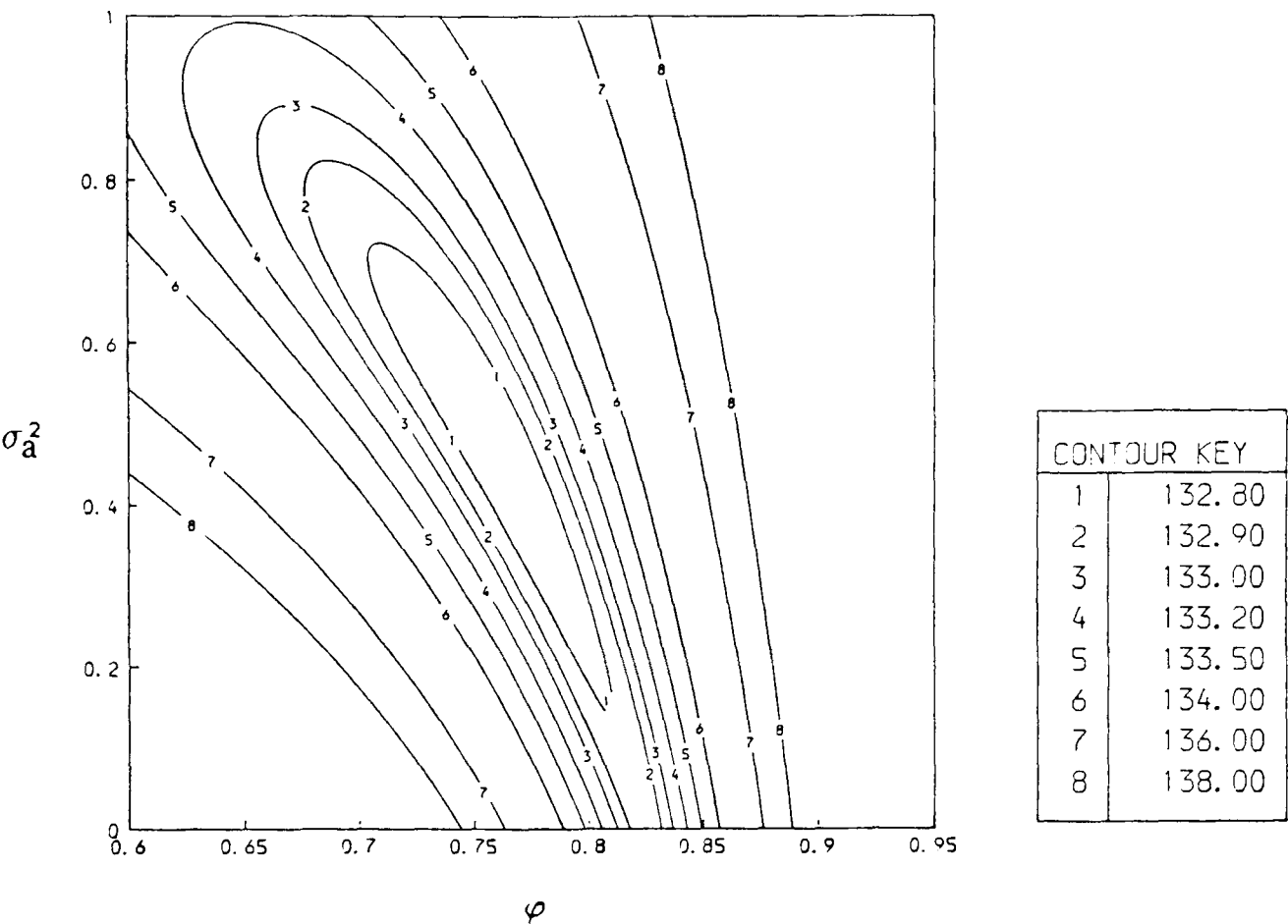
Table 2 indicates that, whereas mles of φ and σ_e^2 with $\hat{\sigma}_a^2$ unconstrained are slightly less biased than those with $\hat{\sigma}_a^2$ constrained to be non-negative, they have greater variance and worse RMSE. Imposing the constraint on σ_a^2 seems to introduce a small extra bias in the mles of φ and σ_e^2 , but this is more than compensated for by a corresponding decrease in variance.

In 77 of the experiments in study 3 (with $m=3$ and $\varphi=0.8$, and $\hat{\sigma}_a^2$ unconstrained) the upper constraint on φ was attained. For one of these, the corresponding mle of $\sigma_a^2 = 1.0$ was as low as -39.8 . This phenomenon, whereby large and negative mles of σ_a^2 were obtained with mles of φ approaching 1, was observed frequently for data generated using $m = 3$ and φ large (> 0.8 say). Examination of the variance matrix $\underline{\Sigma}$ in (2) from which data were generated, reveals that for φ near 1, "large" changes in σ_a^2 can be counterbalanced by "small" changes in φ , leaving the corresponding variance matrix almost unchanged, especially when m is small. This helps to explain the apparently erratic behaviour in study 3 (and, to a lesser extent, in studies 2 and 6) of the unconstrained $\hat{\sigma}_a^2$.

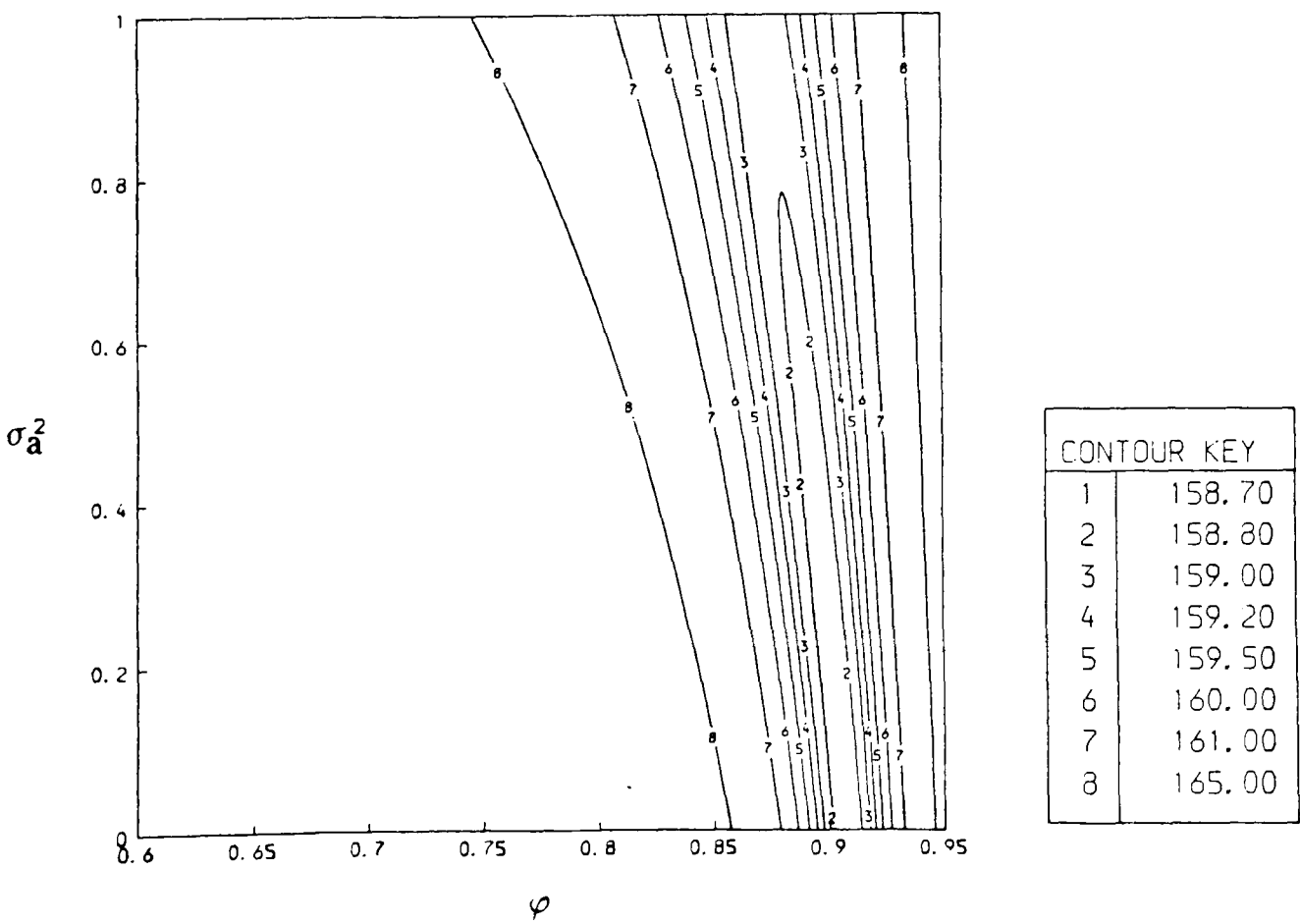
Contour plots were drawn to examine the behaviour of the likelihood as a function of σ_a^2 and φ for certain data sets. Two examples of these are given in Figure 1, (a) and (b): (a) shows a "well-behaved" likelihood surface, while (b) illustrates the sharp "ridge" which can sometimes occur in the likelihood.

Figure 1 Negative log likelihood (nll) as a function of σ_a^2 and φ (with μ and σ_e^2 fixed and equal to $\hat{\mu}$ and $\hat{\sigma}_e^2$ respectively), for two data sets with $N = 100$ subjects and $m = 3$ measurements per subject

(a) nll from data generated using $\sigma_a^2=1.0$ and $\varphi=0.7$; mles are $\hat{\sigma}_a^2=0.48$ and $\hat{\varphi}=0.76$; minimum of nll is 132.72



(b) nll from data generated using $\sigma_a^2=1.0$ and $\varphi=0.9$; mles are $\hat{\sigma}_a^2=-10.79$ and $\hat{\varphi}=0.98$; minimum of nll is 158.55



Note that the parameter XTOL of E04KBF specifies the accuracy in $\underline{\alpha}$ to which the solution is required. For the minimizations of Tables 1 and 2 this was set to 10^{-6} , ensuring that the elements of the solution vector $\underline{\alpha}_{\text{sol}}$ were usually accurate to about 6 decimal places. If XTOL is set too small, rounding error in forming the negative log likelihood makes attainment of the convergence conditions impossible. This happened with 155 of the 4500 experiments (in minimizations with σ_a^2 unbounded) – for each of these, the IFAIL parameter of E04KBF was equal to 3 after the minimization, indicating failure to meet the convergence conditions for a minimum. However, when XTOL was raised to 10^{-4} , further simulations gave IFAIL consistently equal to 0, indicating convergence (although with less accuracy in the solution vector).

The Mansour, Nordheim and Rutledge model, and maximum likelihood estimation

The Mansour, Nordheim and Rutledge (MNR, 1985) model for repeated measurements designs differs from the WHS model in two respects: the mean $\underline{\mu}$ of the m -vector $\underline{X}_i$ of measurements on the i^{th} subject is given by $\underline{\mu} = (\mu_1, \mu_2, \dots, \mu_m)'$ rather than by $\underline{\mu} = \mu(1, 1, \dots, 1)'$; and the variance of the first within-subject error v_1 (cf (1)) is σ_e^2 and not $\sigma_e^2/(1-\varphi^2)$. The more general mean allows the MNR model more flexibility, and also facilitates computation of the mle $\hat{\underline{\mu}}$ of $\underline{\mu}$, since, for $\underline{\mu} = (\mu_1, \mu_2, \dots, \mu_m)'$,

$$\hat{\mu}_j = \frac{1}{N} \sum_{i=1}^N x_{ij}$$

(see, for example, Mardia, Kent and Bibby (MKB, 1979), ps 103–105). A more complicated formula pertains for the mle of $\underline{\mu}$ in the WHS model (see MKB, ps 105, 106). The MNR assumption $\epsilon_1 \sim N(0, \sigma_e^2)$ gives rise to the "transient effect" whereby within-subject variances are increasing for a finite sequence m of measurements, but as m tends to infinity, stationarity in successive variances is attained, as in the WHS model. For the MNR model, the within-subject variance matrix $\underline{V}$ is given by

$$\underline{V} = \sigma_e^2 \begin{bmatrix} 1 & \varphi & \varphi^2 & \dots & \varphi^{m-1} \\ \varphi & 1+\varphi^2 & \varphi(1+\varphi^2) & \dots & \varphi^{m-2}(1+\varphi^2) \\ \varphi^2 & \varphi(1+\varphi^2) & 1+\varphi^2+\varphi^4 & \dots & \varphi^{m-3}(1+\varphi^2+\varphi^4) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{m-1} & \varphi^{m-2}(1+\varphi^2) & \varphi^{m-3}(1+\varphi^2+\varphi^4) & \dots & 1+\varphi^2+\dots+\varphi^{2m-2} \end{bmatrix}.$$

Comparison of this with the WHS $\underline{V}$ in (2) reveals that, whereas measurements in the WHS model have constant variance (are "stationary"), those in the MNR model have increasing variance (and so are "non-stationary"). The stationarity assumption of WHS seems reasonable for many applications (e.g. for repeated measurements of blood pressure); but the increasing variances of the MNR model may be preferable in some circumstances (e.g. for repeated measurements on growth).

In their paper, MNR show how to compute the mles of σ_a^2 , φ and σ_e^2 by writing down the derivatives of the log likelihood, with respect to the parameters, and solving numerically the three simultaneous non-linear equations obtained by setting the derivatives equal to zero. They present results of a Monte Carlo study examining the behaviour of the mles, using samples of size $N = 200$ subjects, and $m = 3$ repeated measurements per subject. They discuss biases of, and correlations between, the 3 mles.

A program was written to compute mles of the parameters of the MNR model, by minimizing the negative log likelihood with respect to the parameters, using E04KBF. Minimization was preferred to solving simultaneous non-linear equations, because problems were encountered in the equation-solving approach to finding mles for the WHS model, when the mle of σ_a^2 was negative (see note on ps 7,8). Simulation results analogous to those of MNR were obtained using the minimization program. MNR remark (p290) that biases in the mles $\hat{\sigma}_a^2$, $\hat{\varphi}$ and $\hat{\sigma}_e^2$ do not seem substantial, for their simulations with $N = 200$ and $m = 3$. However, it was found that for simulations with N smaller than 200, more appreciable biases arose in the mles.

Proof of the Mansour, Nordheim and Rutledge conjecture

Maximum likelihood estimates of the between-subject variance σ_a^2 are not always positive. In discussing this, MNR conjecture (p289) that when the mle $\hat{\sigma}_a^2$ is negative, and the estimate of σ_a^2 is constrained to be non-negative, the constrained mle of σ_a^2 is 0; and the σ_a^2 -constrained mles of φ and σ_e^2 are obtained by finding the mles of φ and σ_e^2 with $\sigma_a^2 = 0$ in the log likelihood. They report that "*Extensive numerical examination of likelihood surfaces has demonstrated the apparent correctness of this conjecture*". The conjecture may be proved formally as below.

Suppose an experiment involves a sample of N subjects, and let $\underline{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$ be the vector of measurements on the i th subject. For convenience of notation, let $x_{ij} = X_{ij} - \mu_j$, $i = 1, \dots, N$, $j = 1, \dots, m$. The log likelihood L ($\equiv L(\underline{X}; \mu, \sigma_a^2, \varphi, \sigma_e^2)$) for the model may be written (apart from a constant)

$$\begin{aligned}
 L = & -\frac{1}{2} N \log \left\{ (\sigma_e^2)^{m-1} \left[\sigma_e^2 + \sigma_a^2 \{1 + (m-1)(1-\varphi)^2\} \right] \right\} \\
 & - \frac{1}{2\sigma_e^2} \left\{ \sum_{i=1}^N (x_{i1}^2 + x_{im}^2) + (1+\varphi^2) \sum_{i=1}^N \sum_{j=2}^{m-1} x_{ij}^2 \right. \\
 & \quad \left. - 2\varphi \sum_{i=1}^N \sum_{j=2}^m x_{ij} x_{i,j-1} + \varphi^2 \sum_{i=1}^N x_{i1}^2 \right\} \\
 & + \frac{\sigma_a^2}{2\sigma_e^2 [\sigma_e^2 + \sigma_a^2 \{1 + (m-1)(1-\varphi)^2\}]} \left\{ (1+\varphi^2-\varphi)^2 \sum_{i=1}^N x_{i1}^2 \right. \\
 & \quad + 2(1+\varphi^2-\varphi)(1-\varphi)^2 \sum_{i=1}^N x_{i1} \sum_{j=2}^{m-1} x_{ij} \\
 & \quad + 2(1+\varphi^2-\varphi)(1-\varphi) \sum_{i=1}^N x_{i1} x_{im} + (1-\varphi)^4 \sum_{i=1}^N \left[\sum_{j=2}^{m-1} x_{ij} \right]^2 \\
 & \quad \left. + 2(1-\varphi)^3 \sum_{i=1}^N x_{im} \sum_{j=2}^{m-1} x_{ij} + (1-\varphi)^2 \sum_{i=1}^N x_{im}^2 \right\}, \tag{5}
 \end{aligned}$$

$$\text{i.e.} \quad L = -\frac{1}{2} N \log \left\{ (\sigma_e^2)^{m-1} [\sigma_e^2 + \sigma_a^2 (1 + (m-1)(1-\varphi)^2)] \right\} \\ - \frac{1}{2\sigma_e^2} \text{ (C1)} + \frac{\sigma_a^2}{2\sigma_e^2 [\sigma_e^2 + \sigma_a^2 (1 + (m-1)(1-\varphi)^2)]} \text{ (C2)} , \quad (6)$$

where C1 and C2 are functions only of μ , φ and the data $\underline{X}$.

Suppose $\hat{\sigma}_a^2 < 0$. We wish to determine the mles of σ_a^2 , φ and σ_e^2 with the constraint that the estimate of the variance σ_a^2 should be non-negative. Think of the log likelihood L as a function of σ_a^2 , φ and σ_e^2 , $L(\sigma_a^2, \varphi, \sigma_e^2)$. To prove the conjecture, we show that, for $\hat{\varphi}_c$ and $\hat{\sigma}_{ec}^2$ obtained by maximizing L with $\sigma_a^2 = 0$,

$$L(0, \hat{\varphi}_c, \hat{\sigma}_{ec}^2) > L(\sigma_a^2, \varphi, \sigma_e^2) \quad \text{for all } \sigma_a^2 > 0, \varphi, \sigma_e^2 > 0.$$

Suppose first that φ and σ_e^2 are fixed, and the likelihood is a function of σ_a^2 only.

Define

$$k = 1 + (m-1)(1-\varphi)^2$$

$$\text{and} \quad w = k\sigma_a^2/\sigma_e^2.$$

From (6), the log likelihood at w , $L(w)$, is given by

$$L(w) = -\frac{N}{2} \log \left\{ (\sigma_e^2)^m (1 + w) \right\} - \frac{1}{2\sigma_e^2} \text{ (C1)} + \frac{w}{2k\sigma_e^2 (1 + w)} \text{ (C2)}; \quad (7)$$

this is well defined for $w > -1$. From (7),

$$\frac{\partial L}{\partial w} = -\frac{N}{2(1+w)} + \frac{C2}{2k\sigma_e^2} \frac{1}{(1+w)^2};$$

$$\text{and} \quad \frac{\partial L}{\partial w} = 0 \implies 1 + w = \frac{C2}{Nk\sigma_e^2},$$

which means that there is *only one* turning point, at

$$w_0 = \frac{C2}{Nk\sigma_e^2} - 1. \quad (8)$$

Note that $C2 > 0$. (To see this, observe that C2 is obtained from

$$\sum_{i=1}^N \underline{x}_i' \underline{V}^{-1} \underline{u} (\underline{u}' \underline{V}^{-1} \underline{u} + 1)^{-1} \underline{u}' \underline{V}^{-1} \underline{x}_i = \frac{\sigma_a^2}{\sigma_e^2 [\sigma_e^2 + \sigma_a^2 \{1 + (m-1)(1-\varphi)^2\}]} C2,$$

where $\underline{V}$ is the covariance matrix of the within-subject measurements, and $\underline{u} = \sigma_a \underline{1}$.

It follows that

$$\sum_{i=1}^N \underline{x}_i' \underline{V}^{-1} \underline{u} \underline{u}' \underline{V}^{-1} \underline{x}_i = \frac{\sigma_a^2}{(\sigma_e^2)^2} C2, \quad \text{and} \quad C2 = (\sigma_e^2)^2 \sum_{i=1}^N t_i^2, \quad \text{where}$$

$t_i = \underline{x}_i' \underline{V}^{-1} \underline{1}$.) Thus, from (8), w_0 is certainly greater than -1, and a unique turning point always exists.

Observe also that

$$\frac{\partial^2 L}{\partial w^2} = \frac{N}{2(1+w)^2} - \frac{C2}{k\sigma_e^2(1+w)^3};$$

so, at the turning point w_0 ,

$$\frac{\partial^2 L}{\partial w^2} = \frac{N}{2 \left[\frac{C2}{Nk\sigma_e^2} \right]^2} - \frac{C2}{k\sigma_e^2 \left[\frac{C2}{Nk\sigma_e^2} \right]^3} = - \frac{N^3 k^2 \sigma_e^4}{2(C2)^2} < 0.$$

Therefore w_0 must be a *maximum* turning point, $w_{\max}$, say, and for all $w > w_{\max}$,

$$\frac{\partial L}{\partial w} < 0; \quad \text{but} \quad \frac{\partial L}{\partial \sigma_a^2} = \frac{k}{\sigma_e^2} \frac{\partial L}{\partial w}, \quad \text{so for all } \sigma_a^2 > \hat{\sigma}_a^2, \quad \frac{\partial L}{\partial \sigma_a^2} < 0.$$

Since $\hat{\sigma}_a^2$ is assumed to be less than 0, this shows that $L(0, \varphi, \sigma_e^2) > L(\sigma_a^2, \varphi, \sigma_e^2)$, for all $\sigma_a^2 > 0$, $\varphi, \sigma_e^2 > 0$. (The first mean value theorem can be used to show this more formally. Suppose $\sigma_a^2 > 0$. Then L (as a function of the between-subject variance) is continuous on $[0, \sigma_a^2]$ and differentiable on $(0, \sigma_a^2)$. Thus, by the mean value theorem, there exists $\tilde{\sigma}_a^2$ in $(0, \sigma_a^2)$ such that

$$\left. \frac{\partial L}{\partial \sigma_a^2} \right|_{\tilde{\sigma}_a^2} = \frac{L(\sigma_a^2) - L(0)}{\sigma_a^2 - 0}; \quad \text{but} \quad \left. \frac{\partial L}{\partial \sigma_a^2} \right|_{\tilde{\sigma}_a^2} < 0, \quad \text{so} \quad L(\sigma_a^2) < L(0).$$

So, finally,

$$\begin{aligned} L(0, \hat{\varphi}_c, \hat{\sigma}_{ec}^2) &\geq L(0, \varphi, \sigma_e^2) \quad \text{for all } \varphi, \sigma_e^2 > 0 \\ &> L(\sigma_a^2, \varphi, \sigma_e^2) \quad \text{for all } \sigma_a^2 > 0, \varphi, \sigma_e^2 > 0. \end{aligned}$$

Imposing the non-negativity constraint on σ_a^2 introduces extra bias into the estimates $\hat{\varphi}_c$ and $\hat{\sigma}_{ec}^2$; but simulation results show that, for the MNR model, as for the WHS model (see Table 2), the root mean squared errors (RMSEs) of σ_a^2 -constrained estimates of φ and σ_e^2 are smaller than the RMSEs of σ_a^2 -unconstrained estimates.

Alternative estimates to mles, for φ and σ_e^2 in the WHS model, based on WHS's variance-based estimation idea, are considered in 1.2.

1.2 Modified variance-based estimation

Variance-based estimation

Wilson, Hebel and Sherwin (WHS, 1985) show that, for their model (cf 1.1, ps 2,3),

$$E(S_n^2) = \frac{\sigma_e^2}{1-\varphi^2} \left\{ 1 - \frac{2}{n(n-1)} \sum_{k=1}^{n-1} (n-k) \varphi^k \right\},$$

where S_n^2 is the within-subject sample variance computed from any $n \leq m$ consecutive measurements (and m is the number of repeated measurements per subject). Hence

$$E(S_2^2) = \frac{\sigma_e^2}{1+\varphi} \quad (1)$$

and

$$E(S_3^2) = \frac{\sigma_e^2}{1+\varphi} \left\{ 1 + \frac{\varphi}{3} \right\}. \quad (2)$$

Solving these two equations for φ and σ_e^2 gives

$$\varphi = 3 \left\{ \frac{E(S_3^2)}{E(S_2^2)} - 1 \right\} \quad (3)$$

and

$$\sigma_e^2 = E(S_2^2) (1 + \varphi). \quad (4)$$

The WHS variance-based estimators of φ and σ_e^2 arise naturally from (3) and (4), and are given by

$$\varphi^* = 3 \left\{ \frac{S_3^2}{S_2^2} - 1 \right\} \quad (5)$$

and

$$\sigma_e^{*2} = S_2^2 (1 + \varphi^*), \quad (6)$$

where S_3^2 and S_2^2 are now redefined as the *averages* over all subjects of all possible variances which can be computed from 3 and 2 adjacent measurements respectively within a subject.

Modified variance-based estimator φ^{1*}

A single-subject version of φ^* in (5) can be simply defined for the i^{th} subject, by

$$\varphi_i^* = 3 \left\{ \frac{S_{3,i}^2}{S_{2,i}^2} - 1 \right\},$$

where $S_{3,i}^2$ and $S_{2,i}^2$ are the averages of all variances which can be computed from 3 and 2 adjacent measurements respectively from the i^{th} subject. (For $m=3$, $-1 \leq \varphi_i^* \leq 3$; this is shown in Appendix 2(i).) The average of these single-subject estimates of φ ,

$$\varphi_{av}^* = \frac{1}{N} \sum_{i=1}^N \varphi_i^*$$

was computed for a number of simulations with $m = 3$ and N (the number of subjects) = 100. (Other parameters were fixed at $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.) It was observed that, unlike φ^* , φ_{av}^* was biased as an estimator of φ ; but φ_{av}^* had smaller variance than φ^* . Also, the difference between φ_{av}^* and the true value of φ seemed closely related to the standard deviation (SD), skewness (SK) and kurtosis (KU) of the N individual estimates of φ , $\{\varphi_i^*, i=1, \dots, N\}$. This gave rise to the idea of a regression of $\varphi_{av}^* - \varphi$ on SD, SK and KU.

Data were simulated from 100 experiments (each with 100 subjects, and 3 repeated measurements per subject) for each of 10 values of φ : $\varphi = 0.0$ (0.1) 0.9. One thousand sets of $\varphi_{av}^* - \varphi$, SD, SK, KU and φ^* were thus obtained. Linear regression was performed, using the GENSTAT package, to predict $\varphi_{av}^* - \varphi$ from the other 4 variables, with the following results (t values corresponding to regression coefficients in brackets):

$$\varphi_{av}^* - \varphi = 0.433 + 0.460(\text{SD}) - 0.8133(\text{SK}) + 0.2205(\text{KU}) - 0.5837(\varphi^*). \quad (7)$$

(2.92) (3.22) (-14.19) (3.56) (-20.45)

The percentage variance accounted for was 30.4. Equation (7) suggests the new estimator φ^{1*} defined by

$$\varphi^{1*} = \varphi_{av}^* - 0.433 - 0.460(\text{SD}) + 0.8133(\text{SK}) - 0.2205(\text{KU}) + 0.5837(\varphi^*). \quad (8)$$

Table 1 (on the next page) presents some simulation results for this new estimator.

Table 1 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of φ^* , φ_{av}^* and φ^{1*} , for $\varphi = 0.2, 0.5, 0.8$*

Each simulation study consisted of 500 experiments, each with $N=100$ subjects, $m=3$ repeated measurements per subject, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.
/ indicates which estimator is best with respect to the RMSE criterion.

φ	Estimator	AV	SD	RMSE
0.2	φ^*	.193	.162	.162
	φ_{av}^*	.568	.134	.392
	φ^{1*}	.269	.127	.144/
0.5	φ^*	.510	.189	.189
	φ_{av}^*	.754	.138	.289
	φ^{1*}	.498	.139	.139/
0.8	φ^*	.804	.202	.202
	φ_{av}^*	.902	.143	.176
	φ^{1*}	.698	.140	.173/

For the three values of φ in Table 1, φ^{1*} is better than φ^* with respect to RMSE. However, although (8) essentially represents an attempt to adjust φ_{av}^* for bias, the estimator φ^{1*} is still biased towards 0.5. Biases of similar magnitude were obtained for φ^{1*} from simulations with $\sigma_a^2 = 0.1$ and $\sigma_a^2 = 10.0$ (instead of 1.0), with φ^{1*} consistently biased towards 0.5. Moreover, as φ decreases below 0.2 or increases above 0.8, the RMSE advantage of φ^{1*} over φ^* , apparent from Table 1, disappears.

Modified variance-based estimators φ^{2*} and φ^{3*}

The problem of adjusting φ_{av}^* by some appropriate linear function of SD, SK, KU and φ^* was exchanged for the equivalent problem of adjusting φ^* by some function of φ_{av}^* , SD, SK and KU.

It was observed from simulation results that the variance $v_{\varphi}(\varphi^*)$ of the estimator φ^* increases with φ , as indicated in Table 2; though φ^* remains essentially unbiased for all values of φ .

Table 2 Empirical variance ($\hat{v}_\varphi$) of WHS's φ^*

Each result in the table was obtained from a simulation with 500 experiments; and each experiment used $N=100$, $m=3$, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.

φ	$\hat{v}_\varphi$	φ	$\hat{v}_\varphi$	φ	$\hat{v}_\varphi$	φ	$\hat{v}_\varphi$
0.99	.0408	0.5	.0357	0.0	.0222	-0.5	.0083
0.9	.0400	0.4	.0342	-0.1	.0199	-0.6	.0056
0.8	.0396	0.3	.0289	-0.2	.0166	-0.7	.0031
0.7	.0396	0.2	.0292	-0.3	.0119	-0.8	.0014
0.6	.0357	0.1	.0259	-0.4	.0112	-0.9	.0004

A theoretical formula for the variance of φ^* was later derived; and the empirical variances of Table 2 are compared with the corresponding theoretical variances in Appendix 2(ii). The very marked differences in variances of φ^* between positive and negative φ , apparent from Table 2, are a consequence of the short series – only 3 measurements per subject – from which φ is being estimated.

In regressing to predict $\varphi^*-\varphi$, allowance should be made for the change in variance of φ^* over different values of φ , by using N_φ sets of $\varphi^*-\varphi$ and corresponding regressor variables in the regressions, where $N_\varphi \propto 1/\hat{v}_\varphi(\varphi^*)$. (In obtaining φ^{1*} of equation (8), N_φ was taken as constant: $N_\varphi = 100$, $\varphi = 0.0$ (0.1) 0.9.)

Significant regressions of $\varphi^*-\varphi$ on various regressor variables, for φ over the whole range $[-1.0, 1.0]$, were found to be unattainable. Moreover, in any given experiment, the true value of φ is unknown, so in general, regressions should be derived for φ^* (and not φ) in some interval. After some trial and error, the range of possible values of φ^* was divided into 4 intervals:

1 : $(-\infty, -0.7)$, 2 : $[-0.7, 0.0)$, 3 : $[0.0, 1.0)$ and 4 : $[1.0, \infty)$.

The idea was to produce a regression equation to predict $\varphi^*-\varphi$ for φ^* in each of the intervals 1 to 4. These equations would vary with the values of N and m , but would be constant over all values of μ , σ_a^2 and σ_e^2 . Equations were derived, and are presented below, for the case $N = 100$ and $m = 3$.

No significant regressions were obtained for φ^* in intervals 1 and 2. Table 3 indicates the number of experiments simulated, and the number of simulations used in deriving regression equations for φ^* , in intervals 3 and 4. In each of the regressions to predict $\varphi^*-\varphi$, possible regressor variables considered were φ_{av}^* , SD, SK, KU, $1/\varphi_{av}^*$, $(\varphi_{av}^*-\varphi^*)$ and $(\varphi_{av}^*-\varphi^*)^2$. Regressions were performed using GENSTAT, with stepwise forward regression.

Table 3 *Development of regressions to predict $\varphi^*-\varphi$*

Each simulation used $N=100$, $m=3$, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.
NEX = number of experiments simulated (roughly proportional to $1/\hat{v}_\varphi(\varphi^*)$)
NIN = number of experiments simulated for which φ^* was in the interval of interest

Interval 3, [0.0, 1.0)											
φ	-.3	-.2	-.1	.0	.1	.2	.3	.4	.5	.6	.7
NEX	400	287	239	214	189	163	163	139	133	133	120
NIN	1	25	69	105	143	148	155	138	131	131	110
φ	.8	.9	.99	Total							
NEX	120	119	117	2536							
NIN	100	82	60	1398							

Interval 4, [1.0, ∞)												
φ	.5	.55	.6	.65	.7	.75	.8	.85	.9	.95	.99	Total
NEX	250	250	250	238	225	225	225	224	223	221	219	2550
NIN	3	3	6	6	14	32	46	44	70	86	101	411

For φ^* in [0.0, 1.0), results from 1398 experiments (see Table 3) yielded the regression equation (t values in brackets)

$$\varphi^*-\varphi = 0.176+0.0955(\varphi_{av}^*)-0.2615(\varphi_{av}^*-\varphi^*)+0.407(\varphi_{av}^*-\varphi^*)^2-0.1395(SD), \quad (9)$$

(1.46)

(3.11)

(-3.51)

(3.04)

(-1.52)

with 1.8% variance accounted for.

For φ^* in [1.0, ∞), results from 411 experiments yielded

For φ non-negative and φ^* in $(-\infty, 1)$, results from 1031 experiments (see Table 4) yielded the regression equation (t values in brackets)

$$\varphi^* - \varphi = -0.0876 + 0.1928(\varphi_{av}^*) - 0.2086(\varphi_{av}^* - \varphi^*) , \quad (12)$$

(-3.86) (7.21) (-6.75)

with 14.3% variance accounted for. Equations (12) and (10) together suggest the further estimator φ^{3*} of φ , *appropriate when φ is known to be non-negative*,

$$\varphi^{3*} = \begin{cases} \varphi^* + 0.0876 - 0.1928(\varphi_{av}^*) + 0.2086(\varphi_{av}^* - \varphi^*) & , \text{ if } \varphi^* \in (-\infty, 1) \\ \varphi^* + 0.545 + 0.095(SK) + 0.7682(\varphi_{av}^* - \varphi^*) - 0.690(\varphi_{av}^*) & , \text{ if } \varphi^* \in [1, \infty) . \end{cases} \quad (13)$$

Simulations were carried out to test the performance of the two new estimators φ^{2*} and φ^{3*} , for $\varphi = 0.0$ (0.1) 0.9. A truncated variance-based estimator

$$\varphi^{t*} = \begin{cases} \varphi^* & , \text{ if } \varphi^* \in (-\infty, 1) \\ 1.0 & , \text{ if } \varphi^* \in [1, \infty) \end{cases}$$

was also tested, for comparison with φ^{2*} and φ^{3*} . Results are presented in Table 5.

Table 5 (on page 23) shows that both φ^{2*} and φ^{3*} improve on φ^* , with respect to the RMSE criterion, for almost the whole range of non-negative φ . The improvement attained by using φ^{3*} (assuming φ is non-negative) ranges between 18% and 22% for φ in the interval [0.3, 0.8]; while φ^{2*} improves on φ^* by between 11% and 19% for φ in the interval [0.6, 0.9]. However, improvements for values of φ in [0.0, 0.3] are not as large; and better RMSEs in φ^{2*} and φ^{3*} are usually accompanied by worse bias. In any case, RMSE results for φ^{2*} and φ^{3*} compare fairly well with the results obtained in 1.1 (see Table 1, p5) for the maximum likelihood estimator $\hat{\varphi}$ of φ .

Note that the regressions used in obtaining φ^{2*} and φ^{3*} are applicable only for $N=100$ subjects and $m=3$ measurements per subject. However, given $N=100$ and $m=3$, the values of μ , σ_a^2 and σ_e^2 do not much affect the accuracy of φ^{2*} and φ^{3*} .

(continued on page 24)

Table 5 *Simulation results: average (AV), standard deviation (SD), root mean squared error (RMSE) and improvement in RMSE as percentage of RMSE(φ^*), of φ^* , φ^{2*} , φ^{3*} and φ^{t*}*

Each simulation consisted of 500 experiments, each with $N=100$, $m=3$, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=0.6$.

φ	Estimator	AV	SD	RMSE	Improvement as % RMSE (φ^*)
0.0	φ^*	.011	.153	.154	0.6 -4.5
	φ^{2*}	.010	.153	.153	
	φ^{3*}	.104	.123	.161	
0.1	φ^*	.113	.161	.162	1.2 4.3
	φ^{2*}	.111	.159	.160	
	φ^{3*}	.185	.129	.155	
0.2	φ^*	.201	.168	.168	2.3 10.7
	φ^{2*}	.197	.163	.164	
	φ^{3*}	.255	.135	.146	
0.3	φ^*	.295	.182	.183	4.4 18.6
	φ^{2*}	.288	.175	.175	
	φ^{3*}	.331	.146	.149	
0.4	φ^*	.420	.183	.184	6.0 18.5
	φ^{2*}	.407	.173	.173	
	φ^{3*}	.431	.147	.150	
0.5	φ^*	.490	.184	.184	7.6 20.7
	φ^{2*}	.471	.168	.170	
	φ^{3*}	.486	.145	.146	
0.6	φ^*	.610	.193	.193	11.9 21.8 2.1
	φ^{2*}	.579	.169	.170	
	φ^{3*}	.581	.150	.151	
	φ^{t*}	.608	.189	.189	
0.7	φ^*	.706	.199	.199	16.1 22.1 5.5
	φ^{2*}	.658	.162	.167	
	φ^{3*}	.653	.148	.155	
	φ^{t*}	.700	.188	.188	
0.8	φ^*	.797	.204	.204	17.6 20.1 11.8
	φ^{2*}	.727	.151	.168	
	φ^{3*}	.717	.140	.163	
	φ^{t*}	.780	.178	.180	
0.9	φ^*	.895	.209	.209	18.7 16.7 23.0
	φ^{2*}	.786	.126	.170	
	φ^{3*}	.775	.122	.174	
	φ^{t*}	.854	.154	.161	

Further simulations were performed with $\mu = 10.0$, $\sigma_a^2 = 1.0$ (as before) and $\sigma_e^2 = 3.0$ (rather than 0.6). The results for φ^{2*} and φ^{3*} in relation to φ^* were comparable to those of Table 5.

It may be worthwhile to investigate and develop further these modified variance-based estimators. To use them for general N and m (and not just $N=100$, $m=3$) the following procedure is suggested.

Suppose, for example, we have data with $N = 57$ subjects and $m = 4$ measurements per subject. First, estimate μ , σ_a^2 and σ_e^2 (assuming the WHS model) from the data on 57 subjects, using unweighted least squares estimators $\tilde{\mu}$ and $\tilde{\sigma}_a^2$ (see 1.1, p3, (3)) for μ and σ_a^2 , and the variance-based estimator σ_e^{*2} (of (6)) for σ_e^2 . Then determine what prior knowledge (if any) is available on φ . Suppose φ is known to be non-negative. Then simulate data from N_φ experiments for each of $\varphi = 0.0$ (0.05) 0.95, where $N_\varphi \propto 1/\hat{v}_\varphi(\varphi^*)$ and the average of N_φ is about 80 say; using $\tilde{\mu}$, $\tilde{\sigma}_a^2$ and σ_e^{*2} as "true" values of the parameters, and using $N = 57$, $m = 4$. Finally, perform regression of $\varphi^* - \varphi$ on various regressor variables. This will result in an estimator of the form $\varphi^{S*} = \varphi^* + \text{terms from the regression}$, and this should improve on the unadjusted estimator φ^* . This whole procedure might be incorporated into a user-friendly computer program, in which the user inputs N , m , the range of values in which φ is known to lie and the specific data to be analysed. Output would consist of the actual estimate φ^{S*} of φ and the regression equation from which it was obtained. Such a program might use about 5 minutes of CPU time on the VAX 11/785.

Further attempts were made to improve on the WHS estimator φ^* , using various modifications, but these proved unsuccessful. They are described in Appendices 2(iii) and 2(iv).

A different procedure for estimating φ , based on the "*functional least squares*" estimation of Heathcote and Welsh (1983), is described in Appendix 3. This procedure may be appropriate for estimating φ when errors come from a variety of (non-normal) leptokurtic distributions.

1.3 Generalization of the models of Wilson, Hebel and Sherwin, and Mansour, Nordheim and Rutledge, by introduction of an extra parameter

The $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model

The models of Wilson, Hebel and Sherwin (WHS, 1981) and Mansour, Nordheim and Rutledge (MNR, 1985) differ in the variance associated with the first within-subject error for each subject: for WHS, this variance is $\sigma_e^2/(1-\varphi^2)$, larger than subsequent variances; while for MNR, it is σ_e^2 , the same as all subsequent variances. The consequences of these respective variance assumptions may be illustrated by reference to the simple case with $m = 3$ repeated measurements per subject. In this case: for WHS, data $\underline{X}_i$ for the i^{th} subject are normally distributed, $\underline{X}_i \sim N(\underline{\mu}_3, \underline{\Sigma}_W)$, where

$$\underline{\Sigma}_W = \sigma_a^2 \underline{1}_3 \underline{1}_3' + \underline{V}_W, \quad \underline{V}_W = \frac{\sigma_e^2}{1 - \varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 & \varphi \\ \varphi^2 & \varphi & 1 \end{bmatrix};$$

while for MNR, $\underline{X}_i \sim N(\underline{\mu}, \underline{\Sigma}_M)$, where

$$\underline{\Sigma}_M = \sigma_a^2 \underline{1}_3 \underline{1}_3' + \underline{V}_M, \quad \underline{V}_M = \sigma_e^2 \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1+\varphi^2 & \varphi(1+\varphi^2) \\ \varphi^2 & \varphi(1+\varphi^2) & 1+\varphi^2+\varphi^4 \end{bmatrix}.$$

Define a generalized covariance matrix by

$$\underline{\Sigma}_G = \sigma_a^2 \underline{1}_3 \underline{1}_3' + \underline{V}_G$$

$$\underline{V}_G = \frac{\sigma_e^2}{1 - K\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1-K\varphi^2+\varphi^2 & \varphi(1-K\varphi^2+\varphi^2) \\ \varphi^2 & \varphi(1-K\varphi^2+\varphi^2) & 1-K\varphi^2+\varphi^2-K\varphi^4+\varphi^4 \end{bmatrix},$$

where $K\varphi^2 < 1$.

Both $\underline{\Sigma}_W$ and $\underline{\Sigma}_M$ are special cases of $\underline{\Sigma}_G$: when $K = 1$, $\underline{V}_G = \underline{V}_W$, and when $K = 0$, $\underline{V}_G = \underline{V}_M$. So the model in which $\underline{X}_i \sim N(\underline{\mu}, \underline{\Sigma}_G)$ is a generalization of

both the WHS and the MNR models, arrived at by introduction of an extra parameter K . The extension of $\underline{V}_G$ for $m > 3$ is straightforward (see 2.3, p63).

It is interesting to inspect $\underline{V}_G$ at values of K outside $[0, 1]$. For example, if $K = 2$, then

$$\underline{V}_G = \frac{\sigma_e^2}{1 - 2\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 - \varphi^2 & \varphi(1 - \varphi^2) \\ \varphi^2 & \varphi(1 - \varphi^2) & 1 - \varphi^2 - \varphi^4 \end{bmatrix} \equiv \underline{V}_G^+ \text{ say;}$$

and if $K = -2$, then

$$\underline{V}_G = \frac{\sigma_e^2}{1 + 2\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 + 3\varphi^2 & \varphi(1 + 3\varphi^2) \\ \varphi^2 & \varphi(1 + 3\varphi^2) & 1 + 3\varphi^2 + 3\varphi^4 \end{bmatrix} \equiv \underline{V}_G^- \text{ say.}$$

It seems reasonable that either $\underline{V}_G^+$ or $\underline{V}_G^-$ may provide a better model for within-subject variance than either $\underline{V}_W$ or $\underline{V}_M$; $\underline{V}_G^+$ allows for decreasing variance in successive within-subject measurements, and $\underline{V}_G^-$ allows for variance to increase at a faster rate than $\underline{V}_M$. Only $\underline{V}_W$, with $K = 1$, represents a stationary process. In general, for $K > 1$, successive variances in $\underline{V}_G$ are monotone decreasing, while for $K < 1$, successive variances are monotone increasing. For fixed $K (\neq 1)$, the rate at which variances decrease or increase is governed by the value of $|\varphi|$.

In a data-modelling application of $\underline{\Sigma}_G$, we could proceed as follows. First estimate the 4 parameters σ_a^2 , φ , σ_e^2 and K by maximum likelihood. If the mle $\hat{K}$ of K is greater than 1, then the WHS (stationary) model provides a better fit to the data than the MNR (non-stationary) model; if $\hat{K}$ is negative, the MNR model is preferable; and if $0 < \hat{K} < 1$, examination of the likelihoods at $K = 1$ and $K = 0$, maximized with respect to the other 3 parameters, allows a criterion for choosing between the two models. Furthermore, if one model is preferred, say that with $K = 1$, it is still relevant to ask whether or not this model provides an "acceptable" fit to the data. Wilks' theorem may be used to test $H_0 : K=1$ against $H_1 : K \neq 1$. If

H_0 is rejected, use the model specified by the mles of σ_a^2 , φ , σ_e^2 and K , obtained in the 4-parameter minimization.

A 3-parameter model, with $K = 0$ or 1 , has the advantage of greater simplicity over the 4-parameter model, in which K is free to vary; although the 4-parameter model may be more accurate. Considerations of simplicity and accuracy, when weighed together, may indicate the best model choice in any given application.

We shall refer to the generalized model as the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model.

Kenward's (1981) split plot autoregressive model, and its relationship to the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model

Kenward (1981) provides a critical study of the more common approaches to the analysis of biological repeated measurements, emphasizing their application as practical statistical tools. In the first part of his three-part thesis, the split plot autoregressive (SPAR) model for errors is defined and discussed. It turns out that the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model is a special case of the SPAR model. Kenward gives an algorithm for finding mles of parameters in the SPAR model, and develops techniques for assessing the adequacy of data-fits using the model. These techniques are used in the analysis of some actual data sets.

In the SPAR model, the error for the t^{th} repeated measurement from a subject is written

$$e_t = g + v_t, \quad t = 1, \dots, m \quad (1)$$

where $E(g) = E(v_t) = 0$, $E(g^2) = \sigma_a^2$ (the between-subject variance) and

$$v_t = \begin{cases} \epsilon_1 & , \quad t = 1 \\ \varphi_1 v_1 + \epsilon_2 & , \quad t = 2 \\ \eta_{t-2} v_{t-2} + \varphi_{t-1} v_{t-1} + \epsilon_t & , \quad t = 3, \dots, m \end{cases} \quad (2)$$

where $E(v_t) = E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma_t^2$. (Note that (1) and (2) use different notation from Kenward (1981).) The model has $3m-2$ parameters: m σ 's, $m-1$ φ 's

$m-2$ η 's and σ_a^2 . The covariance matrix $\underline{\Sigma}$ for the errors $\underline{e}$ of (1) is given by

$$\underline{\Sigma} = \sigma_a^2 \underline{1}_m \underline{1}_m' + \underline{V},$$

where $\underline{V}$ is the covariance matrix of the within-subject errors $\underline{v}$.

Assuming a "first-order" SPAR model (with the η 's in (2) all zero), Kenward's covariance matrix $\underline{V}$ for the case $m = 3$ is given by

$$\underline{V} = \begin{bmatrix} \sigma_1^2 & \sigma_1^2 \varphi_1 & \sigma_1^2 \varphi_1 \varphi_2 \\ \sigma_1^2 \varphi_1 & \sigma_2^2 + \sigma_1^2 \varphi_1^2 & \varphi_2 (\sigma_2^2 + \sigma_1^2 \varphi_1^2) \\ \sigma_1^2 \varphi_1 \varphi_2 & \varphi_2 (\sigma_2^2 + \sigma_1^2 \varphi_1^2) & \sigma_3^2 + \sigma_2^2 \varphi_2^2 + \sigma_1^2 \varphi_1^2 \varphi_2^2 \end{bmatrix}. \quad (3)$$

It appears that the general first-order SPAR model is difficult to interpret or motivate via the matrix $\underline{V}$, even in the simple case $m = 3$. Its attraction and motivation are more readily understood from (1) and (2). However, if the φ 's are assumed to be all equal, $\underline{V}$ in (3) simplifies to

$$\underline{V} = \begin{bmatrix} \sigma_1^2 & \sigma_1^2 \varphi & \sigma_1^2 \varphi^2 \\ \sigma_1^2 \varphi & \sigma_2^2 + \sigma_1^2 \varphi^2 & \varphi (\sigma_2^2 + \sigma_1^2 \varphi^2) \\ \sigma_1^2 \varphi^2 & \varphi (\sigma_2^2 + \sigma_1^2 \varphi^2) & \sigma_3^2 + \sigma_2^2 \varphi^2 + \sigma_1^2 \varphi^4 \end{bmatrix}.$$

Furthermore, if $\sigma_1^2 = \sigma_e^2 / (1 - K\varphi^2)$ and $\sigma_2^2 = \sigma_3^2 = \sigma_e^2$, the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model is obtained, with $\underline{V} = \underline{V}_G$ as defined above. The WHS and MNR models are obtained by putting $K = 1$ and $K = 0$ respectively. Thus the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model and the WHS and MNR models are all just special cases of Kenward's SPAR model.

Simulation results using the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model

Let $\underline{X}_i = (X_{i1}, X_{i2}, \dots, X_{im})'$ be the vector of measurements on subject i , and suppose an experiment involves a sample of N subjects. Then the log likelihood L_G for the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model (apart from a constant) is given by

$$\begin{aligned}
L_G = & -\frac{1}{2} N \log \left\{ \frac{(\sigma_e^2)^{m-1}}{1 - K\varphi^2} \left[\sigma_e^2 + \sigma_a^2(1 + (m-1)(1-\varphi)^2 - K\varphi^2) \right] \right\} \\
& - \frac{1}{2\sigma_e^2} \left\{ \sum_{i=1}^N (x_{i1}^2 + x_{im}^2) + (1+\varphi^2) \sum_{i=1}^N \sum_{j=2}^{m-1} x_{ij}^2 \right. \\
& \quad \left. - 2\varphi \sum_{i=1}^N \sum_{j=2}^m x_{ij} x_{i,j-1} + \varphi^2(1-K) \sum_{i=1}^N x_{i1}^2 \right\} \\
& + \frac{\sigma_a^2}{2\sigma_e^2 [\sigma_e^2 + \sigma_a^2(1 + (m-1)(1-\varphi)^2 - K\varphi^2)]} \left\{ (1-K\varphi^2+\varphi^2-\varphi)^2 \sum_{i=1}^N x_{i1}^2 \right. \\
& + 2(1-K\varphi^2+\varphi^2-\varphi)(1-\varphi)^2 \sum_{i=1}^N x_{i1} \sum_{j=2}^{m-1} x_{ij} \\
& + 2(1-K\varphi^2+\varphi^2-\varphi)(1-\varphi) \sum_{i=1}^N x_{i1} x_{im} + (1-\varphi)^4 \sum_{i=1}^N \left[\sum_{j=2}^{m-1} x_{ij} \right]^2 \\
& \left. + 2(1-\varphi)^3 \sum_{i=1}^N x_{im} \sum_{j=2}^{m-1} x_{ij} + (1-\varphi)^2 \sum_{i=1}^N x_{im}^2 \right\} \quad (4)
\end{aligned}$$

(cf 1.1, p12, (5)), where $x_{ij} = X_{ij} - \mu_j$, $i=1, \dots, N$, $j=1, \dots, m$. Minimizing $-L_G$, with respect to the 4 parameters σ_a^2 , φ , σ_e^2 and K , gives mles of the parameters.

A series of simulations was performed using the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model. A $3 \times 8 \times 2$ factorial design was used: with m (the number of repeated measurements per subject) = 3, 5, 10; φ (the autoregression parameter) = -0.7 (0.2) 0.7; and $K = 0, 1$. For each simulation, data were simulated for $N = 100$ subjects, with $\underline{\mu} = \underline{1}_m$. The parameters σ_a^2 (between-subject variance) and σ_e^2 (within-subject variance) were fixed at 2 and 1 respectively. One thousand simulations were performed for each of the 48 combinations of m , φ and K .

For each simulated data set, 3 minimizations of $-L_G$ were performed (using the NAG subroutine E04KBF):

- (i) with respect to σ_a^2 , φ and σ_e^2 , assuming $K = 1$ (WHS),
- (ii) with respect to σ_a^2 , φ and σ_e^2 , assuming $K = 0$ (MNR)

and (iii) with respect to σ_a^2 , φ , σ_e^2 and K .

In each minimization the value of σ_a^2 was constrained to be non-negative. The average CPU time used on the VAX per simulation (over the 48000 simulations) was 0.93 seconds.

A comparison of the maximized log likelihoods corresponding to (i) and (ii) was made for each simulation. This served as a criterion for choosing between the models with $K = 1$ (WHS, stationary errors) and $K = 0$ (MNR, increasing error variances). Results from comparisons between the two maximized likelihoods are given in Table 1.

Table 1 *Number of times (out of 1000) the maximized log likelihood for the model with $K=1$ (WHS) was greater than that for the model with $K=0$ (MNR)*

data simulated from						
φ	K = 1			K = 0		
	m=3	m=5	m=10	m=3	m=5	m=10
-0.7	949	975	989	75	40	20
-0.5	700	782	817	254	177	155
-0.3	596	584	587	433	408	379
-0.1	503	481	516	509	481	448
0.1	565	503	526	523	489	470
0.3	566	556	585	480	374	399
0.5	688	694	755	354	298	216
0.7	795	872	907	192	149	93
Total	5362	5447	5682	2820	2416	2180

Table 1 displays some predictable trends: as m increases or $|\varphi|$ increases, the rate of detection of the correct model generally increases. It is difficult to distinguish between the two models when $|\varphi|$ is small, because in this case they are "nearly equivalent". When $\varphi=0$, the two models *are* equivalent, and the value of K in the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model is irrelevant. The asymmetry in Table 1 between results for positive and negative values of φ is related to the fact that (maximum likelihood) estimates of

model parameters are more precise when φ is negative than when φ is positive, for a given absolute value of φ .

Wilks' theorem (see 1.4, p35) was used to test the null hypothesis $K = 1$ (or $K = 0$) against the alternative K free to vary, by comparing the maximized log likelihoods corresponding to (i) and (iii) (or (ii) and (iii)). Tests were only performed for simulations in which the mle of σ_a^2 was non-zero, and therefore unconstrained, in all 3 minimizations; and in which all 3 minimizations were completed with E04KBF error code 0. A 5% significance level was used. Wilks' test results are presented in Table 2 (on page 32).

The proportions of rejections of the null hypothesis $K = 1$ (or $K = 0$) with data simulated from $K = 1$ (or from $K = 0$), are around 0.05 (see Table 2), and so agree well with the 5% significance level used in the tests. There is no obvious asymmetry between the number of rejections for positive and negative φ , when the true value of K (the value used in simulating the data) is assumed. However, when the wrong value of K is assumed, asymmetry is again quite clear, with the number of rejections being greater for negative than for positive φ .

Second derivatives of L_G in (4), $\partial^2 L_G / \partial (\sigma_a^2)^2$, $\partial^2 L_G / \partial \sigma_a^2 \partial \varphi$, etc, were evaluated analytically. These were used in the simulations as follows. For a minimization of type (iii), if the mle of σ_a^2 was non-zero, and the E04KBF error code was 0, then an estimate of the 4×4 matrix of second derivatives of the negative log likelihood with respect to the parameters was evaluated. This was used in the standard way to obtain an approximation to the covariance matrix of the mles. Average values of standard deviations and covariances of the mles were thus obtained for each of the 48 combinations of m , φ and K . The average standard deviation of $\hat{K}$ is presented in Table 3 (on page 33) for each of the 48 combinations.

Table 2 *Number of times, N_1 , Wilks' test rejected the null hypothesis $K = 1$, and N_0 , the null hypothesis $K = 0$, at the 5% level, with simulated data (NT = number of tests carried out)*

data simulated from									
φ	K = 1								
	m = 3			m = 5			m = 10		
	NT	N_1	N_0	NT	N_1	N_0	NT	N_1	N_0
-0.7	1000	43	852	993	44	964	997	43	989
-0.5	992	57	209	1000	47	372	998	58	454
-0.3	970	45	71	998	52	76	993	44	86
-0.1	855	40	44	952	34	31	977	47	47
0.1	865	45	40	951	51	50	968	47	53
0.3	957	45	57	994	45	65	987	43	83
0.5	936	38	138	996	59	213	999	48	297
0.7	760	38	168	947	62	589	990	48	798
Total	7335	351	1579	7831	394	2360	7909	378	2807
Proportion rejected		.048	.215		.050	.301		.048	.355

data simulated from									
φ	K = 0								
	m = 3			m = 5			m = 10		
	NT	N_1	N_0	NT	N_1	N_0	NT	N_1	N_0
-0.7	994	758	43	999	929	62	998	980	59
-0.5	994	237	56	999	322	48	997	424	57
-0.3	950	76	55	997	91	52	997	85	50
-0.1	857	40	32	935	50	46	975	54	55
0.1	851	43	36	957	41	41	972	57	58
0.3	952	52	40	995	74	49	992	72	44
0.5	972	135	42	994	182	47	998	278	54
0.7	875	328	21	994	512	50	999	747	53
Total	7445	1669	325	7870	2201	395	7928	2697	430
Proportion rejected		.224	.044		.280	.050		.340	.054

Table 3 *Average standard deviation of K in minimizations of type (iii)*
(number of simulations contributing to average in brackets)

data simulated from			
φ	K = 1		
	m = 3	m = 5	m = 10
-0.7	0.27 (1000)	0.21 (993)	0.18 (1000)
-0.5	1.08 (1000)	0.61 (1000)	0.50 (1000)
-0.3	96.10 (978)	2.58 (998)	1.74 (996)
-0.1	1139.16 (857)	441.54 (961)	70.81 (983)
0.1	804.67 (862)	487.68 (941)	110.30 (981)
0.3	222.93 (960)	3.99 (998)	2.09 (993)
0.5	96.01 (946)	1.05 (999)	0.67 (999)
0.7	3.57 (787)	0.41 (954)	0.27 (996)

data simulated from			
φ	K = 0		
	m = 3	m = 5	m = 10
-0.7	0.70 (1000)	0.38 (1000)	0.33 (1000)
-0.5	2.66 (999)	0.81 (1000)	0.66 (999)
-0.3	77.57 (966)	2.97 (998)	1.96 (998)
-0.1	569.64 (876)	370.62 (939)	89.46 (980)
0.1	871.59 (841)	391.67 (954)	91.23 (975)
0.3	168.35 (954)	4.31 (998)	2.29 (996)
0.5	32.86 (982)	1.42 (999)	0.91 (999)
0.7	12.87 (911)	1.64 (998)	0.68 (999)

It is not clear what biases may have been introduced in Table 3 by omission of results from simulations with $\hat{\sigma}_a^2$ constrained and from simulations with uncertain convergence (E04KBF error code non-zero). However, some general trends are clear. When, for instance, $m=3$ and $|\varphi|=0.1$, K is estimated with very large variance; but estimates of K improve as $|\varphi|$ increases and as m increases.

1.4 Data analyses using the generalized model

Analysis of dental caries data

A 3-year clinical trial on dentifrices was carried out in the north-west of England, and reported by Hodge et al (1980). At the outset, the trial involved 979 subjects, volunteers from 6 schools in the region, aged 11 to 12 years. Before the trial began, it had been well established that the use of properly formulated dentifrices containing sodium monofluorophosphate (smfp) reduced the incidence of dental caries in teenagers. This trial was designed to test whether the addition of sodium fluoride might enhance the effectiveness of smfp in preventing dental caries.

Subjects were stratified according to school and sex, and each subject was randomly assigned to one of 4 treatment groups. Treatments for groups 1 to 4 were:

- | | |
|---------------------|--|
| 1 test treatment I | 0.76% smfp plus 0.10% sodium fluoride, formulated in an alumina abrasive system |
| 2 test treatment II | 0.76% smfp plus 0.10% sodium fluoride, formulated in a dicalcium phosphate abrasive system |
| 3 positive control | 0.76% smfp, in an alumina abrasive system and |
| 4 negative control | non-fluoride, in an alumina abrasive system. |

Tests for significant differences between the test treatments and the controls were performed on the 3-year mean caries increments for subjects in the 4 groups. The two test treatments were judged significantly better than the negative control at the 1% level (yielding approximately a 25% improvement), and better than the positive control at the 5% level (yielding approximately a 16% improvement). It is of interest, and of some practical importance, to ask whether the duration of the trial was longer than "necessary" – could comparable strength of evidence (in terms of significance) have issued from say, a 2-year rather than a 3-year trial? This question is addressed in 2.4. Here, the intention is simply to illustrate application of the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model.

Complete data were obtained for 790 of the 979 subjects originally participating in the trial; and consisted of an integer score, the decayed, missing or filled teeth surface

(DMFS) score, obtained at the start of the trial and after 1, 2 and 3 years. Models were fitted to the DMFS *increments* from year to year, each subject thus providing a series of 3 data points.

Two possible parametrizations for the mean DMFS $\underline{\mu}$ were considered:

- (A) $\underline{\mu} = \mu(1,1,1)'$, 1 parameter
 and (B) $\underline{\mu} = (\mu_1, \mu_2, \mu_3)'$, 3 parameters.

Three models (cf 1.3),

- (i) $K = 1$ (WHS, stationary errors),
 (ii) $K = 0$ (MNR, increasing variances for errors)
 and (iii) K free to vary (the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model),

were fitted by maximum likelihood to data from each group, and to data from all the groups combined, assuming first (A) and then (B). The maximized log likelihoods were examined, and Wilks' theorem was applied, to provide a criterion for choosing between the fitted models.

Wilks' theorem states that, if Θ_0 and Θ_1 are subsets of the parameter space, with $\Theta_0 \subset \Theta_1$, and if Λ_n is defined by

$$\Lambda_n = \sup_{\theta \in \Theta_1} L_n(\theta) / \sup_{\theta \in \Theta_0} L_n(\theta) ,$$

where L_n is the likelihood based on n measurements, then

$$2 \log \Lambda_n \text{ is asymptotically (as } n \rightarrow \infty) \chi_v^2 ,$$

for all $\theta \in \Theta_0$, subject to certain regularity conditions, where $v = \dim \Theta_1 - \dim \Theta_0$.

In the formulations above, (A) is a special case of (B), and (i) and (ii) are special cases of (iii). Wilks' Theorem can thus be used to test (A) against (B), and (i) or (ii) against (iii). Table 1 (on page 36) gives maximized log likelihoods for the various models fitting the DMFS data. (Note that for each of the models in Table 1, the mle $\hat{\sigma}_a^2$ of the between-subject variance σ_a^2 was greater than 0. If $\hat{\sigma}_a^2$ had been constrained at 0, it would have been inappropriate to use Wilks' theorem. The same observations apply for the mles of σ_a^2 in the analysis of blood pressure data — see Table 3, p38.)

Table 1 *Maximized log likelihoods for DMFS data, assuming models (A)(i) to (B)(iii)*
('K f' means K free to vary)

Group	Model					
	(A) $\underline{\mu} = \mu(1,1,1)'$			(B) $\underline{\mu} = (\mu_1, \mu_2, \mu_3)'$		
	(i)K=1	(ii)K=0	(iii)K f	(i)K=1	(ii)K=0	(iii)K f
1	-854.71	-854.75	-853.98	-849.14	-849.21	-848.18
2	-857.51	-858.12	-855.09	-846.23	-847.14	-843.31
3	-879.11	-879.20	-878.93	-869.50	-869.65	-868.82
4	-923.16	-923.09	-921.40	-920.80	-920.75	-918.94
Combined	-3532.13	-3532.23	-3532.00	-3512.98	-3513.17	-3512.53

Consider testing (A) against (B),

i.e. $H_0 : \underline{\mu} = \mu(1,1,1)'$ against $H_1 : \underline{\mu} = (\mu_1, \mu_2, \mu_3)'$,

for treatment group 1 and model (iii). For this case,

$$\begin{aligned} 2 \log \Lambda_n &= 2(853.98 - 848.18) \quad (\text{see Table 1}) \\ &= 11.6 \end{aligned}$$

From Wilks' theorem, $2 \log \Lambda_n$ is approximately distributed as χ^2_2 . The upper 1% point of χ^2_2 is 9.21, so in this case we would reject H_0 . The corresponding statistic, $2 \log \Lambda_n$, is equal to 23.6, 20.2 and 4.9 for data from groups 2, 3 and 4 respectively; highly significant for groups 2 and 3, though not quite significant at the 5% level for group 4. Overall, it seems assumption (B), that $\underline{\mu} = (\mu_1, \mu_2, \mu_3)'$, is preferable to (A) for this DMFS data. The mles of $\underline{\mu}$ under (B) for the four groups are:

- 1 (2.32, 2.22, 1.59)' (mean from 196 subjects),
- 2 (2.32, 2.26, 1.35)' (mean from 201 subjects),
- 3 (2.60, 2.91, 1.82)' (mean from 193 subjects)
- and 4 (2.34, 2.92, 2.56)' (mean from 200 subjects).

For both test treatments, a decreasing average rate of tooth decay was realised over the 3 years; while for both control treatments, the rate increased from year 1 to year 2 before decreasing in the third year.

Now consider comparisons between models (i) and (ii), restricting attention to results obtained under assumption (B). The maximized log likelihoods are better (larger) for (i) than for (ii), for data from 3 of the 4 treatment groups; with group 4 (on the negative control treatment) again providing the exception. Thus the WHS model, implying stationary errors, seems generally better for this data than the MNR model.

Finally, it remains to test model (B)(i) against model (B)(iii),

$$H_0 : K=1 \qquad \text{against} \qquad H_1 : K \text{ free to vary.}$$

For data from treatment group 1,

$$\begin{aligned} 2 \log \Lambda_n &= 2(849.14 - 848.18) \quad (\text{see Table 1}) \\ &= 1.92 \end{aligned}$$

The upper 5% point of χ^2_1 is 3.84, so in this case, we would tend to accept $H_0 : K=1$. Corresponding statistics for data from groups 2, 3 and 4 are 5.84, 1.36 and 3.72 respectively. Although one of these is significant, and another almost significant, at the 5% level, it may be best to accept H_0 and consider the simpler model with $K = 1$ as generally the most appropriate. (For the significant result from treatment 2, the value of the mle $\hat{K}$ is 7.22 (with approximate standard deviation 5.21), suggesting decreasing error variances.)

The mles of σ_a^2 , φ and σ_e^2 for the preferred model, (B)(i), are presented in Table 2.

Table 2 *Mles of σ_a^2 , φ and σ_e^2 for DMFS data, assuming $\underline{\mu} = (\mu_1, \mu_2, \mu_3)'$ and $K = 1$ (approximate standard deviations in brackets)*

Group	$\hat{\sigma}_a^2$	$\hat{\varphi}$	$\hat{\sigma}_e^2$
1	1.417 (0.415)	-0.094 (0.088)	5.318 (0.502)
2	1.099 (0.295)	-0.239 (0.076)	4.808 (0.445)
3	2.051 (0.484)	-0.152 (0.091)	5.537 (0.563)
4	1.203 (0.522)	-0.054 (0.089)	6.809 (0.641)
Combined	1.469 (0.219)	-0.118 (0.044)	5.738 (0.274)

The negative values for $\hat{\varphi}$ in Table 2 were not unexpected. Armitage, Stratton and Worthington (1985) applied the WHS ($K=1$) model to DMFS data, from fewer subjects, and obtained $\hat{\varphi}$ equal to about -0.35 . (Estimates of standard deviation were not given.) They suggested that negative estimates of φ may be induced by independent measurement errors in the cumulative DMFS's. In their paper, $\hat{\sigma}_a^2$ was obtained as 0 for each of 2 groups, each having about half as many subjects as the groups considered here.

Analysis of blood pressure data

Blood pressure data were obtained from each of 25 hypertensive males in the UK (Solomon, 1985). Each subject was visited quarterly for a period of 4 years. At each visit, a subject's diastolic and systolic blood pressures were measured twice each. The physicians conducting treatment were attempting to control diastolic pressure to a level not exceeding 95 mm Hg. No such direct control was attempted for systolic pressure. For each subject-visit, one value for diastolic pressure was obtained by taking the average of the two diastolic measurements, and similarly for systolic pressure.

Models of type (A)(i) to (B)(iii), described above, were fitted to the diastolic data, and independently to the systolic data. (The data were first transformed from mm Hg to cm Hg, to avoid arithmetic overflow errors while running the minimizations used in model fitting.) Maximized log likelihoods under the various model assumptions are presented in Table 3.

Table 3 *Maximized log likelihoods for blood pressure data, assuming models (A)(i) to (B)(iii) ('K f' means K free to vary)*

pressure	Model					
	(A) $\underline{\mu} = \underline{\mu}_{1,6}$			(B) $\underline{\mu} = (\mu_1, \mu_2, \dots, \mu_{16})'$		
	(i)K=1	(ii)K=0	(iii)K f	(i)K=1	(ii)K=0	(iii)K f
diastolic	-93.60	-93.62	-93.54	-86.96	-86.99	-86.86
systolic	-354.19	-354.10	-353.41	-349.79	-349.69	-348.99

A test at the 5% significance level of

$$H_0 : \mu = \mu_{1,6} \quad \text{against} \quad H_1 : \mu = (\mu_1, \mu_2, \dots, \mu_{16})'$$

accepts H_0 for each of models (i), (ii) and (iii), for both diastolic and systolic data ($\chi^2_{15,0.05} = 25.00$). Under model (A), $\hat{\mu} = 9.2$ for diastolic data, and $\hat{\mu} = 14.9$ for systolic data.

The $K = 1$ (WHS) stationary error model is marginally better than the $K = 0$ (MNR) non-stationary model for diastolic data; and vice versa for systolic data. For model (A)(iii), the value of $\hat{K}$ is 6.14 (with standard deviation 15.31) for diastolic data, and -17.99 (with standard deviation 21.09) for systolic data. These values (rather weakly) suggest decreasing and increasing error variances respectively. Perhaps the decreasing variance implied for diastolic data arises because of the physicians' attempts to maintain diastolic pressure below 95 mm Hg. The mles of the parameters for the diastolic data (assuming the $K=1$, WHS model) and for the systolic data (assuming the $K=0$, MNR model) are given in Table 4.

Table 4 *Mles of σ_a^2 , φ and σ_e^2 for blood pressure data (approximate standard deviations in brackets)*

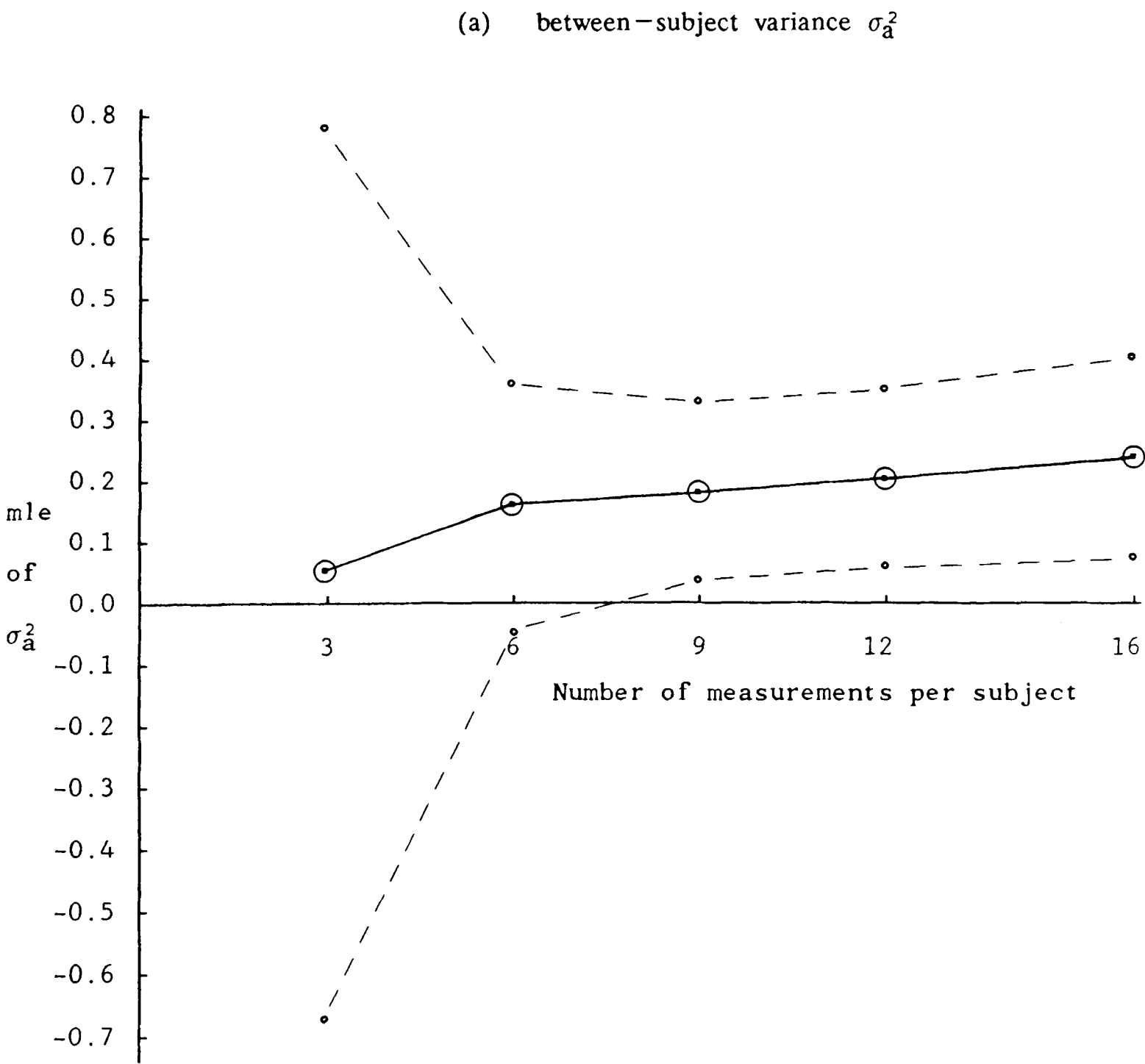
pressure	$\hat{\sigma}_a^2$	$\hat{\varphi}$	$\hat{\sigma}_e^2$
diastolic	0.243 (0.081)	0.140 (0.056)	0.521 (0.038)
systolic	1.836 (0.566)	0.162 (0.057)	1.846 (0.136)

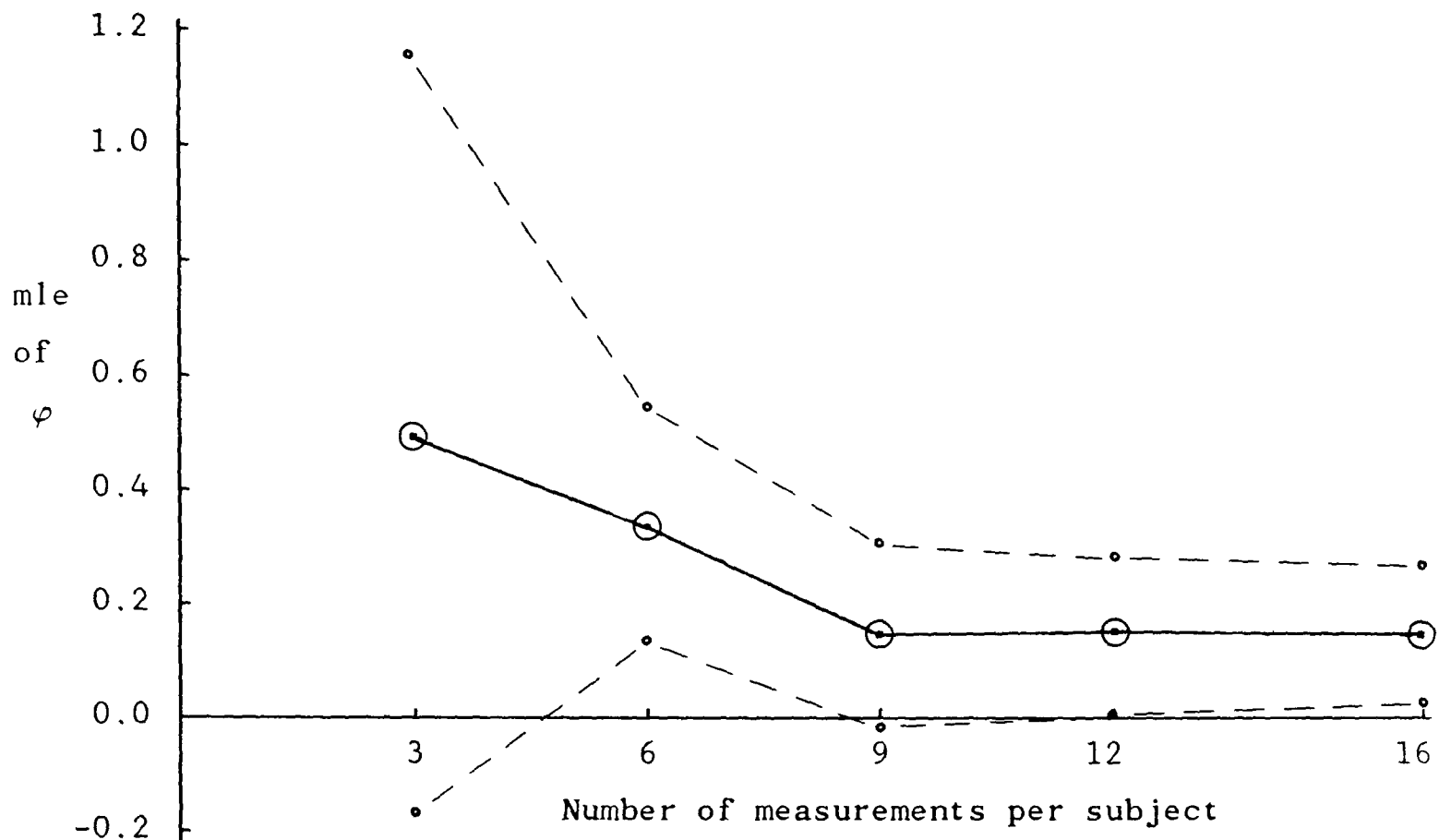
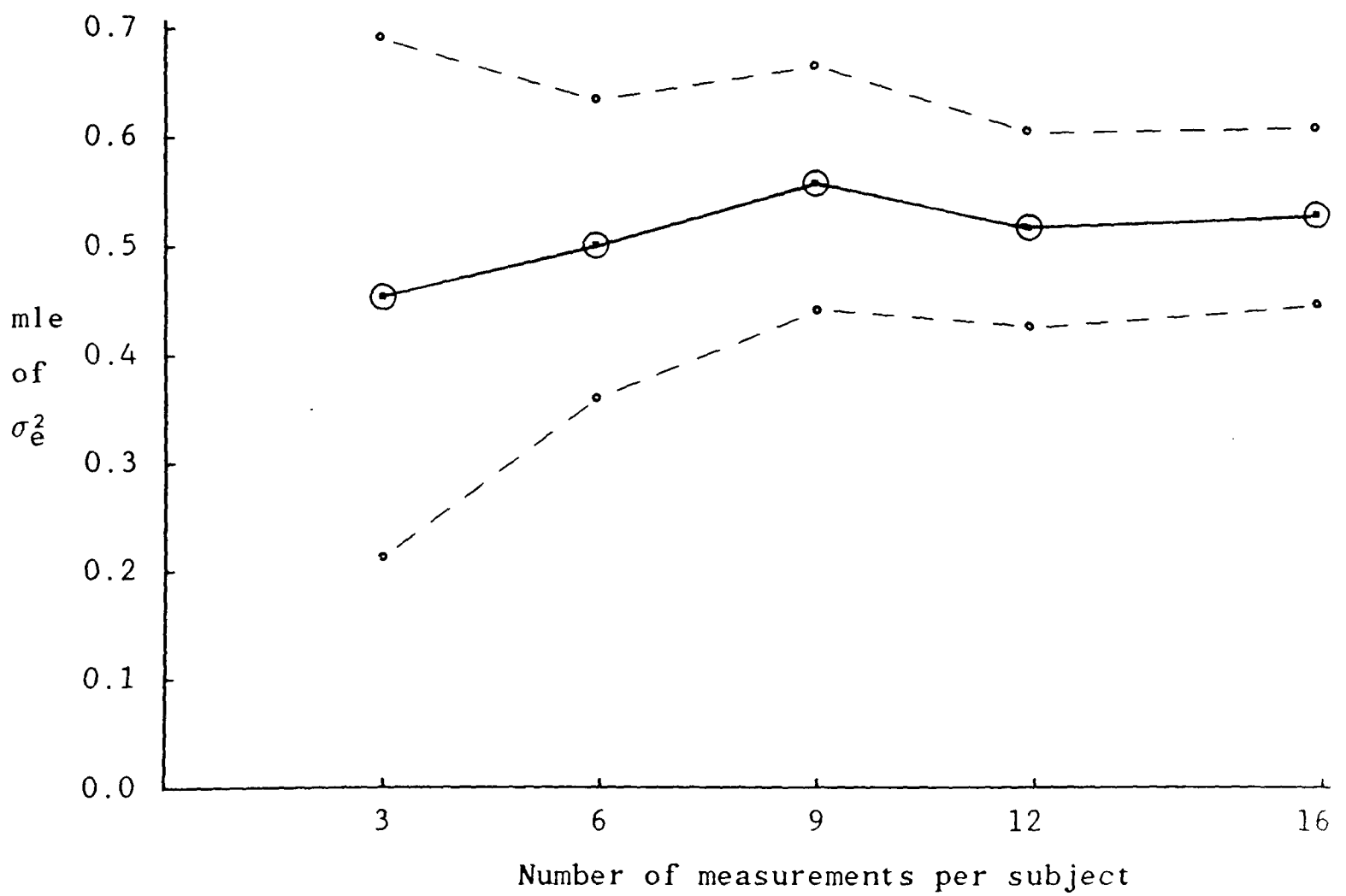
In the context of sequential testing (cf Chapter 2) it may be necessary or expedient to estimate model parameters successively as more data become available (see 2.4). With this in mind, and to see how precise the parameter estimates might be with comparatively small datasets (in this case only 25 subjects), model fitting was performed on truncated blood pressure data with 3, 6, 9, and 12 repeated measurements included successively. In each case, the $K = 1$ model was best for diastolic data, and $K = 0$ was best for systolic data, with the following exceptions: for 9 measurements on diastolic data, $K = 0$ seemed slightly better than $K = 1$; and for 6 and 9

measurements on systolic data, a test of $H_0 : K=0$ against $H_1 : K$ free to vary, would have rejected H_0 at the 5% level in favour of a negative K value.

Plots of mles $\hat{\sigma}_a^2$, $\hat{\varphi}$ and $\hat{\sigma}_e^2$ for diastolic data, against the number of repeated measurements included in the fitted models, are given in Figure 1, (a), (b) and (c). Analogous plots for systolic data displayed similar trends.

Figure 1 mles of σ_a^2 , φ and σ_e^2 for diastolic blood pressure data from 25 subjects, as a function of the number of repeated measurements per subject, assuming the $K = 1$ stationary error model (dotted lines represent approximate 95% confidence intervals for the mles)



(b) autoregression constant φ (c) within-subject variance σ_e^2 

CHAPTER 2 SEQUENTIAL TESTING WITH SERIALY CORRELATED DATA

2.1 Sequential stopping rules: classical or Bayesian?

- *This is no laughing matter, O'Casey, said Ulster; we're talking now about the affairs of God and not about affairs of men. I want to know if God is a catholic or a protestant: answer that, will ye?*
- *He's neither, said Sean laughingly; relationship with Him isn't sanctioned by the push-button of an opinion.* (O'Casey, 1952)

In a paper entitled "*Sequential stopping rules and sequentially adjusted P values: does one require the other?*", William D Dupont (1983a) calls into question the rationale behind using adjusted P values in sequential testing. He argues (p4) that sequentially adjusted P values "*can lead to contradictory conclusions when used as measures of inferential strength*". In developing this argument, he implies a basic dissatisfaction with classical or frequentist inference, and with its dependence on outcomes that were not observed. He inclines, rather, towards a system of inference based on the likelihood principle.

Dupont gives an example to show how the same trial results can yield a P value of 0.01 or 0.05, depending on the intentions of the experimenters at the outset of the trial. Uncomfortable with this, he draws attention to the likelihood ratio statistic, $R(\eta) \equiv L(\eta)/L(\hat{\eta})$, where η is the parameter indexing some hypotheses, and $\hat{\eta}$ is the value of η maximizing the likelihood L . If the null hypothesis $H_0 : \eta = \eta_0$ is tested, then $R(\eta_0) = 0.01$ means that the data are 100 times more likely to be observed under the alternative hypothesis $H_a : \eta = \hat{\eta}$ than under H_0 . In this case the data provide fairly strong evidence that H_0 is false, and this evidence is the same irrespective of the stopping rule used in the trial. Hence Dupont prefers inference based on the likelihood principle, to classical inference based on P values.

However, Dupont argues that P values are now widely used and understood in medical science. Moreover, for fixed-sample experiments with normally distributed data, the likelihood ratio statistic and the P value are equivalent in that, if two experiments yield P values P_1 and P_2 , and likelihood ratio statistics R_1 and R_2 , then

$$P_1 > P_2 \iff R_1 > R_2 .$$

For these reasons Dupont proposes, as a "*pragmatic compromise*", that fixed-sample P values are quoted at the end of a trial, irrespective of how frequently the data have been inspected. He claims that these fixed-sample P values provide a better measure of the strength of evidence justified by the data, than the sequentially adjusted values.

Three of the four discussants of Dupont's paper disagree with the proposed pragmatic compromise because it involves a misuse of the fixed-sample P value.

Greenhouse (1983) points out the failure of Dupont's paper to address the problem of providing the best criterion for early stopping; the likelihood principle is proposed only as a tool for inference once the trial is completed. In his rejoinder (1983b), Dupont replies that his original paper did not aim to address the stopping rule problem. He writes further (p28): *"The need for stopping rules that permit early termination yet insure that this outcome only occurs when the evidence is irrefutable has led to the development of the methods mentioned by Drs. Canner and Greenhouse. I have no quarrel with any of these methods as long as they are used solely as guides to decision making concerning the early termination of a trial."* The methods referred to here are those classical methods based on controlling type I and type II error probabilities.

McPherson (1982) reviews the rationale behind conventional stopping rules in group sequential clinical trials. He addresses the problem of how best to pre-specify the number of interim analyses to be undertaken in a trial. Four possible prior distributions are defined for the treatment effect. Eight designs, with fixed significance level α and power $1 - \beta$ ($\alpha = 0.05$, $1 - \beta = 0.90$) and with 1, 2, 4, 5, 8, 10, 20 and 40 interim analyses respectively, are compared on the basis of two indices:

- (i) the Average Sample Number (ASN), averaged with respect to the prior

distribution

and (ii) the ASN times the absolute treatment difference, averaged over the prior distribution.

These comparisons suggest that for a relatively disperse prior distribution, around 5–10 interim analyses are optimum, while for a more precise prior, between 2 and 4 analyses may be best, for power of 0.90. (For lesser power, fewer looks would be better, and vice-versa.)

Berry (1985) takes up the theme of interim analyses from a Bayesian perspective. He expresses an aversion to significance tests, to the extent of advocating the "*eradication of P-values and significance tests from statistics*" (p525).

Berry first describes the classical approach to interim analyses, then McPherson's (1982) "*quasi-classical*" approach. The Bayesian approach is presented in a framework analogous to that used by McPherson. Suppose the prior distribution of the treatment effect δ is $N(0, \tau^2)$, and the observed data $\{X_i, i=1, \dots, n\}$ are normally distributed, $N(\delta, 1)$. Then the posterior distribution of δ is $N(n\bar{X}/(n+\tau^{-2}), 1/(n+\tau^{-2}))$. Berry proposes testing the data after every N/g observations, where N is the maximum number of subjects (or subject pairs) and g is the maximum number of looks at the data. If at any stage, $P(\delta > 0 | X_1, X_2, \dots, X_n) > 0.95$ or $P(\delta < 0 | X_1, X_2, \dots, X_n) > 0.95$, then stop the trial. Otherwise, continue until the posterior standard deviation is small, say 0.05. Berry demonstrates by simulation that with this stopping rule, the ASN decreases indefinitely with the number of looks at the data. He contrasts this situation with McPherson's, where the minimum ASN sometimes occurs in the fixed-sample case, with one look.

With this stopping rule though, the problem of inflated type I error probabilities remains unresolved. Berry claims (p526) that "*the role of type I error is now played by the posterior distribution, or calculations made from this distribution*". This role play is convenient, as Berry points out, because of its effect on ASN. However, smaller ASN is only achieved at the expense of larger type I error. The true type I error varies with the number of repeated tests in much the same way for the Bayesian

scheme as for the usual P value scheme. This is demonstrated below.

Note that since the posterior distribution of δ is $N(n\bar{X}/(n+\tau^{-2}), 1/(n+\tau^{-2}))$, we decide $P(\delta>0|X_1,X_2,...,X_n) \geq 0.95$ and stop the trial if

$$\frac{\frac{n\bar{X}}{n+\tau^{-2}} - 0}{\left[\frac{1}{n+\tau^{-2}} \right]^{\frac{1}{2}}} \geq 1.96 ,$$

i.e. if $\sqrt{n} \bar{X} \geq \frac{1.96 (n+\tau^{-2})^{\frac{1}{2}}}{\sqrt{n}}$ (1)

(cf McPherson and Armitage, 1971, p20). Now under the null hypothesis $H_0 : \delta=0$, we have that

$$\bar{X} \sim N(0,1/n) \quad \text{and} \quad \sqrt{n} \bar{X} \sim N(0,1). \tag{2}$$

The z values of Table 1 are computed using (1) and (2). They are appropriate for the first test of $H_0 : \delta=0$, using the Bayesian procedure, on each of 6 possible designs, for the two τ^2 values used in Berry's paper.

Table 1 *z values for the first test in a Bayesian sequential testing scheme, with a maximum of N subjects and N/g equally spaced looks at the data, and prior distributions $N(0,\tau^2)$*

g	1	2	5	10	20	40
N=399, $\tau^2=1.0$	1.962	1.965	1.972	1.984	2.008	2.056
N=375, $\tau^2=0.04$	2.024	2.087	2.263	2.537	2.983	3.920

For $\tau^2 = 1.0$, with a maximum of 40 repeated tests on the data, the first 2-sided test is at a fixed sample P value of about 0.08. Subsequent tests have fixed sample P values of between 0.08 and 0.10. Hence the true 2-sided probability of type I error for 40 repeated tests with this stopping rule is something approaching 0.500 (see

Armitage et al, 1969, p239). (McPherson and Armitage (1971) observe that, when τ^2 is infinite, the Bayesian testing procedure is equivalent to the classical procedure.)

Berry points out (p525) that *"any costs and benefits (monetary and otherwise) that will accrue from the study can and should be explicitly considered at the design stage"*. He then immediately sets such considerations aside, as being tangential to the purpose of his paper. However, the costs associated with the potentially highly inflated type I error of the proposed Bayesian procedure may outweigh any benefits accruing from the procedure.

Moreover, Berry admits (p525) the legitimacy of the concern of classical statisticians over choice of an appropriate prior distribution for the treatment effect δ . The choice of such a prior distribution sometimes seems arbitrary, and then its influence on trial conclusions seems unjustified. Attempts to formalize choice of a suitable prior distribution, by measuring subjective prior clinical opinion, have recently been reported in, for example, Spiegelhalter and Freedman (1986). Nevertheless, the discussion between frequentist and Bayesian statisticians may continue to stick at this point. If there is convincing prior information about the nature of the distribution of a parameter, it seems prodigal to ignore it. Efforts to use such information may well be encouraged. But ultimately how convinced will the prior sceptics be?

Smith (1984), in an overview of Bayesian statistics, predicts (p256) *"that Bayes's theorem plus computer graphics will be the accepted form of most day-to-day statistical practice by the end of the century"*. Here, in sections 2.2 and 2.4, methodology is presented for sequential analysis, essentially assuming the classical standpoint, and using sequentially adjusted P values. The expectation is that both classical and Bayesian statistics will continue to fulfil useful functions for at least the next 14 years.

I want to know if Statistics is a classical catholic or a Bayesian protestant: onswer thot, wull ye?

2.2 Two methods for determining appropriate nominal significance levels, when the model for data is known a priori

Group sequential testing with independent measurements

Consider a clinical trial with two treatments, 1 and 2. In the usual group sequential testing situation (see, for example, Pocock, 1977), treatment responses are assumed to be independent. Subjects in the trial are divided into groups which enter the trial sequentially, and a test for significant treatment differences is conducted after each group of subjects has completed the trial. Suppose here that equal nominal significance levels α^* are used in the tests.

If there are $2n$ subjects in each group, then n are allocated to each treatment by a randomized permuted block design. Let μ_1 and μ_2 be the unknown mean measurements, including treatment effects, corresponding to treatments 1 and 2 respectively; and assume measurements are normally distributed with known equal variance σ^2 . Denote measurements by $\{X_{ijk}, i=1, \dots, n \text{ (subjects)}, j=1, \dots, m \text{ (groups)}, \text{ and } k=1, 2 \text{ (treatments)}\}$. Then S_g , the difference between the sums of measurements from subjects on treatments 1 and 2 in the first g groups, is given by

$$S_g = \sum_{i=1}^n \sum_{j=1}^g (X_{ij1} - X_{ij2}) ;$$

S_g is normal, with mean $ng(\mu_1 - \mu_2)$ and variance $2ng\sigma^2$.

Denote the sample equivalent of S_g by s_g . Then, under the null hypothesis $H_0 : \mu_1 = \mu_2$,

$$p_g \equiv P \left[|S_g| \geq |s_g| \right] = 2 \left[1 - \Phi \left[\frac{|s_g|}{(2ng\sigma^2)^{1/2}} \right] \right] .$$

If, for any g , p_g is less than the predetermined nominal significance level α^* (or equivalently, if $|s_g|/(2ng\sigma^2)^{1/2}$ is greater than the z -value z^* corresponding to α^*), then the trial is stopped, with the conclusion that there is a significant treatment

difference. Otherwise, the trial continues until m groups, of $2n$ subjects each, have participated. The level α^* is chosen so that the overall probability of falsely rejecting the null hypothesis, allowing for up to m tests, is α , say. It follows that if z^* is such that, when $\mu_1 = \mu_2$,

$$\int_{-z^*/\sqrt{v_m}}^{z^*/\sqrt{v_m}} \cdots \int_{-z^*/\sqrt{v_2}}^{z^*/\sqrt{v_2}} \int_{-z^*/\sqrt{v_1}}^{z^*/\sqrt{v_1}} f_{\underline{S}}(\underline{s}) d\underline{s} = 1 - \alpha, \quad (1)$$

where $v_i = \text{var}(S_i)$ and $\underline{S} = (S_1, S_2, \dots, S_m)'$, then

$$\alpha^* = 2(1 - \Phi(z^*)) . \quad (2)$$

Define

$$X_j = \sum_{i=1}^n (X_{ij_1} - X_{ij_2}) = X_{.j_1} - X_{.j_2},$$

the difference between sums of measurements on treatments 1 and 2 in the j^{th} group. For independent measurements, the covariance matrix of $\underline{X} = (X_1, X_2, \dots, X_m)'$ is given by

$$\underline{\Sigma}_X = 2n\sigma^2 \underline{I}_m, \quad (3)$$

while the covariance matrix of $\underline{S} = (S_1, S_2, \dots, S_m)'$ is

$$\underline{\Sigma}_S = 2n\sigma^2 \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & \dots & 2 \\ 1 & 2 & 3 & \dots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \dots & m \end{bmatrix}. \quad (4)$$

In this case the density of S_j may be defined recursively, enabling the joint density of $\underline{S}$ to be integrated (see Armitage, McPherson and Rowe, 1969). This allows the solution z^* to equation (1) to be determined numerically, and z^* may be used in (2) to obtain the desired nominal significance level α^* .

Nominal significance levels for repeated measurements data: the integration method

When measurements are not independent, $\underline{\Sigma}_X$ and $\underline{\Sigma}_S$ do not take the simple forms of (3) and (4). However, the method for determining the appropriate nominal level α^* for a given overall significance level α remains essentially unchanged: determine z^* satisfying (1) and then α^* follows immediately from (2). We shall refer to this as the *integration method*.

Consider now the repeated measurements trial, in which the total number of subjects is fixed, but each subject provides a small time series of up to m measurements. It is unlikely that such measurements will be independent. Models for data arising in this way are discussed in Chapter 1. Here, the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model of 1.3 will be assumed.

Suppose n subjects receive treatment 1 and another n receive treatment 2. The maximum possible number of measurements in this case is the same as in the group sequential situation above: this trial requires $2n$ subjects, each yielding up to m measurements, while the group sequential trial may require $2nm$ subjects, each giving rise to 1 measurement.

Now assume that measurements on treatments 1 and 2 come from the same $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model (except that they may have different means), where the parameters σ_a^2 , φ , σ_e^2 and K are known. Define X_j in this case as the difference between sums of measurements on treatments 1 and 2 at the j^{th} test (rather than in the j^{th} group). The variance matrix of $\underline{X}$ is now given by $\underline{\Sigma}_X = 2n\underline{\Sigma}$, where $\underline{\Sigma}$ has known form (see 2.3, p63); and the unknown assumed mean of $\underline{X}$ may be written $n(\mu_1 - \mu_2) \underline{1}_m \equiv \mu \underline{1}_m$, say. Now, if

$$\underline{A} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}_{m \times m},$$

then $\underline{S} = \underline{A} \underline{X}$ and $\underline{\Sigma}_S = \underline{A} \underline{\Sigma}_X \underline{A}'$. Hence

$$\underline{X} \sim N(\underline{\mu} \underline{1}_m, \underline{\Sigma}_X) \implies \underline{S} \sim N(\underline{\mu} \underline{A} \underline{1}_m, \underline{\Sigma}_S) .$$

Under the null hypothesis of no treatment differences, $\underline{\mu} = 0$, the pdf of $\underline{S}$ is the usual multivariate normal

$$f_{\underline{S}}(\underline{s}) = \frac{1}{(2\pi)^{m/2} |\underline{\Sigma}_S|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \underline{s}' \underline{\Sigma}_S^{-1} \underline{s} \right\} . \quad (5)$$

Theoretically, the integral in (1) can be evaluated for $f(\underline{s})$ in (5) using numerical integration algorithms. For instance, the NAG routine F01FCF is designed to evaluate a multi-dimensional integral (up to 15 dimensions) with constant and finite limits, to a specified relative accuracy, using an adaptive subdivision strategy. Alternatively, a subroutine MULNOR, for evaluation of multivariate normal probabilities up to dimension 7, is presented by Schervish (1984). Both these methods were programmed for the case $m=3$, and a comparison was made between them on the basis of time taken to evaluate

$$I = \int_{-1.96\sqrt{3}}^{1.96\sqrt{3}} \int_{-1.96\sqrt{2}}^{1.96\sqrt{2}} \int_{-1.96}^{1.96} f_{\underline{S}}(\underline{s}) d\underline{s} ,$$

with $\underline{S} = (X_1, X_1+X_2, X_1+X_2+X_3)'$ and $\underline{X} \sim N(\underline{0}, \underline{I}_3)$. Using accuracy parameter equal to 0.000005 for each method:

F01FCF gave I equal to 0.8927518, with 101541 integrand evaluations, taking 45 CPU seconds

and MULNOR gave I equal to 0.8927518, in less than 1 CPU second.

Thus MULNOR was much faster than F01FCF in evaluating I , and from this, and other 3-dimensional integrations, it seemed that MULNOR was generally preferable.

MULNOR was used in evaluating the appropriate nominal levels for data from differing $(\sigma_a^2, \phi, \sigma_e^2, K)$ models in the case $m = 3$. These levels are given in Table 1. They were obtained by solving for z^* the non-linear equation

$$\int_{-z^*/\sqrt{v_3}}^{z^*/\sqrt{v_3}} \int_{-z^*/\sqrt{v_2}}^{z^*/\sqrt{v_2}} \int_{-z^*/\sqrt{v_1}}^{z^*/\sqrt{v_1}} f_{\underline{S}}(\underline{s}) d\underline{s} = 0.95 , \quad (6)$$

where $v_j = \text{var}(S_j)$, using the NAG routine C05AZF. (C05AZF locates a simple zero of a continuous function on a given interval, by a combination of linear interpolation, linear extrapolation and bisection. It uses "reverse communication" for evaluating the function. For Table 1, the accuracy parameter of C05AZF was set at 0.000005, and the initial bounds for z^* were set at 1.96 and 2.5. The accuracy parameter for MULNOR was also set at 0.000005 (or 0.00002, for entries with $\varphi = -0.99$.) Solutions for z^* were found after about 7 or 8 triple integral evaluations. On obtaining the solution z^* to (6), the corresponding nominal level α^* was obtained from (2), using the NAG routine S15ABF (which returns the value of the cumulative normal distribution at a given point).

Table 1 *Nominal significance levels α^* for a maximum of 3 repeated tests on data from various $(\sigma_a^2, \varphi, \sigma_e^2, K)$ models, with overall significance level $\alpha = 0.05$ ($r = \sigma_a^2 / \sigma_e^2$)*

φ	K = 1				K = 0			
	r=0.0	r=0.5	r=1.0	r=2.0	r=0.0	r=0.5	r=1.0	r=2.0
-0.99	.0236	.0229	.0226	.0224	.0185	.0214	.0238	.0270
-0.75	.0197	.0220	.0238	.0262	.0184	.0222	.0247	.0280
-0.50	.0197	.0232	.0255	.0284	.0192	.0232	.0258	.0290
-0.25	.0206	.0244	.0268	.0298	.0205	.0244	.0268	.0300
0.00	.0221	.0257	.0280	.0310	.0221	.0257	.0280	.0310
0.25	.0240	.0272	.0293	.0320	.0238	.0271	.0292	.0321
0.50	.0267	.0292	.0309	.0332	.0255	.0285	.0305	.0332
0.75	.0313	.0326	.0337	.0353	.0273	.0299	.0316	.0341
0.99	.0451	.0451	.0451	.0451	.0289	.0309	.0324	.0345

The values in Table 1 reinforce the observations on pages 356–358 of Armitage, Stratton and Worthington (1985). High values of $r = \sigma_a^2 / \sigma_e^2$ tend to increase the correlations between S_1 , S_2 and S_3 , and so lead to values of α^* approaching α . The same is true of high positive values of the autoregression constant φ ; while negative values of φ tend to have the opposite effect, leading to smaller values of α^* . The

curious results for r small, whereby $\varphi = -0.99$ sometimes gives higher α^* values than $\varphi = -0.75$, are analogous to the correlation patterns for lowest values of φ presented in Table 3 (p358) of Armitage et al.

For data from the $K = 1$ (stationary error) model, correlations between S_1 , S_2 and S_3 tend to be smaller for negative values of φ (except when r is small) and larger for positive values of φ than the corresponding correlations for data from the $K = 0$ (non-stationary error) model. This leads to a range of α^* in Table 1 between $\varphi = -0.99$ and $\varphi = 0.99$ which is wider for $K = 1$ than for $K = 0$, for a given value of r .

Note that for $\varphi = 0$, the entries in Table 1 are the same for $K = 1$ as for $K = 0$, because in this case the two models are equivalent.

The model $(\sigma_a^2, \varphi, \sigma_e^2, K) = (0, 0, 1, 1)$ (or $(0, 0, 1, 0)$) corresponds to the usual independent measurements case, and the entry 0.0221 in Table 1 for this model is the usual value of α^* for group sequential testing with $m = 3$.

For the DMFS data analysed by Armitage et al and fitted by the model $\hat{\sigma}_a^2 = 0$, $\hat{\varphi} \cong -0.35$ and $\hat{\sigma}_e^2 \cong 12.8$ (with $K = 1$), the appropriate nominal significance level was computed as 0.0202. For the distinct DMFS data considered in 1.4, mles for the $K = 1$ model, with data from 4 groups combined, were $\hat{\sigma}_a^2 = 1.47$, $\hat{\varphi} = -0.12$ and $\hat{\sigma}_e^2 = 5.74$ (see Table 2, p37). The appropriate nominal level for such data was found to be 0.0235.

Suitable nominal levels for normal data from the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model, for a maximum of 3 repeated tests, may quickly be derived as indicated above, by numerical evaluation of a series of triple integrals. For 4 or more repeated tests, higher order numerical integration is required. Appropriate nominal levels were obtained, equal to .0182, for 4 repeated tests on $(\sigma_a^2, \varphi, \sigma_e^2, K) = (0, 0, 1, 1)$, corresponding to the usual independent measurements case; but the solution, involving seven 4-dimensional integrations, required just over 8 minutes of CPU time on the VAX 11/785 as compared with around 10 seconds for each of the 3-dimensional solutions of Table 1. Using accuracy parameter for MULNOR equal to 0.000005, not even one integral was evaluated in the 5-dimensional case, assuming the $(0, 0, 1, 1)$ model, in *one hour* of

CPU time. From this, and other attempts at 5-dimensional integrals, it seems that Schervish's routine uses excessive CPU time for the integrals of interest here, when there are 5 or more dimensions.

Thus two problems remain. One is the problem of integrating in high dimensions, if up to 5 or more repeated tests are to be performed. The other is that before commencement of a trial, it is not usually clear which $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model may best fit the data. Values of the 4 parameters may need to be estimated sequentially and some allowance then made for this estimation in the test procedure. A partial remedy for the integration problem is proposed in 2.3. Procedures for sequentially estimating parameters and testing for significance are discussed in 2.4.

Nominal significance levels for repeated measurements data: the transformation method

The *transformation method* circumvents the problem of integration in high dimensions. Essentially, the idea is to "transform the data to independence", assuming they come from a known $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model. For independent measurements, there is no problem about performing integrations of the kind in (1), even in high dimensions; and suitable nominal levels have been well documented by, for example, Pocock (1977).

Define $\underline{X}_j = (X_1, \dots, X_j)'$, where X_j is, as above, the difference between sums of measurements on treatments 1 and 2 at the j^{th} test. Then

$$\underline{X}_j \sim N(n(\mu_1 - \mu_2)\underline{1}_j, 2n\underline{\Sigma}_j) ,$$

where the $j \times j$ matrix $\underline{\Sigma}_j$ is the appropriate covariance matrix for a j -vector from the known $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model.

Define a sequence of transformation matrices, $\underline{T}_1, \underline{T}_2, \dots, \underline{T}_m$ (where $\underline{T}_j$ is $j \times j$) to satisfy

$$\underline{T}_j \underline{\Sigma}_j \underline{T}_j' = \underline{I}_j \quad (\text{the } j \times j \text{ identity matrix}) \quad j = 1, \dots, m$$

$$\text{and } \underline{T}_{j+1} = \begin{bmatrix} & & & & 0 \\ & & & & \vdots \\ & & \underline{T}_j & & 0 \\ a_{j+1,1} & a_{j+1,2} & \dots & a_{j+1,j+1} & \end{bmatrix} \quad j = 1, \dots, m-1 .$$

Such a sequence $\underline{T}_j$ exists for all symmetric positive definite matrices $\underline{\Sigma}_j$. (This is an immediate consequence of the Cholesky decomposition theorem.) Then the transformed data $\underline{X}_j^* \equiv \underline{T}_j \underline{X}_j$ are normally distributed, as $N(n(\mu_1 - \mu_2)\underline{T}_j \underline{1}_j, \underline{I}_j)$; and $\underline{X}_j^*$ is a vector of iid measurements. Note that

$$\mu_j \equiv n(\mu_1 - \mu_2) \underline{1}_j = \underline{0} \quad \leftrightarrow \quad \mu_j^* \equiv n(\mu_1 - \mu_2) \underline{T}_j \underline{1}_j = \underline{0} \quad \leftrightarrow \quad \mu_1 - \mu_2 = 0.$$

Thus to test the null hypothesis $\mu_1 = \mu_2$, we can either test $\mu_j = \underline{0}_j$, $j=1, \dots, m$ using the original data $\underline{X}_j$, or test $\mu_j^* = \underline{0}_j$, $j=1, \dots, m$ using the transformed data $\underline{X}_j^*$. Now for the iid transformed data, the usual group sequential procedure applies.

Hence if data are known a priori to come from a given $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model, transform the data at the j^{th} stage by $\underline{T}_j$, and test using the usual nominal levels.

Closed-form expressions for the elements of the matrices $\underline{T}_j$, in terms of the elements of the matrices $\underline{\Sigma}_j$ of interest, seem difficult to obtain. For instance, if the transformation matrix $\underline{T}_3$ is written

$$\underline{T}_3 = \begin{bmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{bmatrix},$$

then a , b , c , etc, may be determined by ensuring they satisfy certain equations. Specifically:

$$\text{for } \underline{X}_1^* = X_1^* = aX_1 = \underline{T}_1 \underline{X}_1, \text{ we require } \text{var}(X_1^*) = a^2 \text{var}(X_1) = 1 \quad \text{and}$$

solving this equation gives the desired value of a ;

$$\text{for } \underline{X}_2^* = (X_1^*, X_2^*)' = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \underline{T}_2 \underline{X}_2, \text{ we require}$$

$$\text{cov}(X_1^*, X_2^*) = 0 \quad \text{and} \quad \text{var}(X_2^*) = 1, \quad \text{and solving these two equations}$$

gives b and c ; and

$$\text{for } \underline{X}_3^* = (X_1^*, X_2^*, X_3^*)' = \begin{bmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \underline{T}_3 \underline{X}_3,$$

we require $\text{cov}(X_1^*, X_3^*) = 0$, $\text{cov}(X_2^*, X_3^*) = 0$ and $\text{var}(X_3^*) = 1$,

yielding 3 equations to be solved for d, e and f. If φ in the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model is zero, the appropriate covariance matrix for $\underline{X}_3$ is given by

$$\underline{\Sigma}_3 = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 \end{bmatrix}.$$

Then a, b, c and f are given by

$$\begin{aligned} a &= \frac{1}{(\sigma_a^2 + \sigma_e^2)^{\frac{1}{2}}}, & b &= \frac{-\sigma_a^2}{\{[(\sigma_a^2 + \sigma_e^2)^2 - (\sigma_a^2)^2](\sigma_a^2 + \sigma_e^2)\}^{\frac{1}{2}}}, \\ c &= \left\{ \frac{\sigma_a^2 + \sigma_e^2}{(\sigma_a^2 + \sigma_e^2)^2 - (\sigma_a^2)^2} \right\}^{\frac{1}{2}}, \\ f &= \frac{(\sigma_a^2 + \sigma_e^2)^{\frac{1}{2}} [2\sigma_a^2 \sigma_e^2 + (\sigma_e^2)^2]}{\{(\sigma_e^2)^3 [6(\sigma_a^2)^3 + 11(\sigma_a^2)^2 \sigma_e^2 + 6\sigma_a^2 (\sigma_e^2)^2 + (\sigma_e^2)^3]\}^{\frac{1}{2}}}, \end{aligned}$$

and (like f) d and e are complicated functions of σ_a^2 and σ_e^2 . For the more general covariance matrices $\underline{\Sigma}_j$ pertaining for dimensions greater than 3 and for φ non-zero, solutions for a, b, c, etc are correspondingly more complicated.

While it is difficult to write down algebraic expressions for the general elements of $\underline{T}_j$, it is nevertheless simple to compute these elements by numerical solution of a series of non-linear equations. The NAG subroutine C05NBF, for instance, can be used to find the solution of a system of j non-linear functions in j variables.

To give some idea of what the transformation matrices look like in particular cases, three examples of 5-dimensional transformation matrices corresponding to given $(\sigma_a^2, \varphi, \sigma_e^2, K)$ models, as computed using C05NBF, are presented below. The first example indicates the order of rounding errors arising when using the numerical algorithm for equation-solving.

(a) (0,0,1,1) – iid case

$$\begin{bmatrix} 1.00 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.69 \times 10^{-19} & 1.00 & 0.0 & 0.0 & 0.0 \\ 0.46 \times 10^{-26} & -0.24 \times 10^{-27} & 1.00 & 0.0 & 0.0 \\ 0.73 \times 10^{-26} & 0.16 \times 10^{-24} & 0.32 \times 10^{-27} & 1.00 & 0.0 \\ 0.73 \times 10^{-26} & -0.27 \times 10^{-34} & 0.00 & 0.32 \times 10^{-27} & 1.00 \end{bmatrix}$$

(b) (1,-0.5,1,1)

$$\begin{bmatrix} 0.66 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.10 & 0.66 & 0.0 & 0.0 & 0.0 \\ -0.45 & -0.05 & 0.80 & 0.0 & 0.0 \\ -0.31 & -0.46 & 0.12 & 0.86 & 0.0 \\ -0.24 & -0.35 & -0.35 & 0.21 & 0.89 \end{bmatrix}$$

(c) (1,0.5,1,1)

$$\begin{bmatrix} 0.66 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.67 & 0.94 & 0.0 & 0.0 & 0.0 \\ -0.12 & -0.59 & 0.94 & 0.0 & 0.0 \\ -0.11 & -0.05 & -0.58 & 0.95 & 0.0 \\ -0.10 & -0.05 & -0.05 & -0.57 & 0.95 \end{bmatrix}$$

To confirm the feasibility of the transformation method of testing when the model for data is known, and to observe the effect of using the method to perform repeated tests when the wrong model for data is assumed, simulations were performed. Data were simulated from each of 32 models – all combinations of

$$r = 0, 0.5, 1, 2 \quad (\text{where } r = \sigma_a^2/\sigma_e^2, \text{ with } \sigma_e^2 \text{ fixed at } 1),$$

$$\varphi = -0.7, -0.3, 0.3, 0.7$$

$$\text{and } K = 0, 1.$$

For each model,

- (i) 200000 simulations were performed with 1 subject per simulation and 3 repeated measurements per subject; and

(ii) for each simulation, the data were transformed assuming a given model, and tested for significance at the 5% level (with nominal significance levels equal to 0.0221, appropriate for group sequential testing with a maximum of 3 groups of subjects); this step was repeated assuming each of the 32 models in turn.

Empirical significance levels were obtained by computing the proportions of each set of 200000 simulations rejected under each of the model assumptions. Some of the results are presented in Table 2.

Table 2 Empirical overall significance levels obtained using the transformation method for a maximum of three repeated tests on data from various $(\sigma_a^2, \varphi, \sigma_e^2, K)$ models, with theoretical overall significance level $\alpha = 0.05$

model number		$(\sigma_a^2, \varphi, \sigma_e^2, K)$	model number		$(\sigma_a^2, \varphi, \sigma_e^2, K)$
	1	$(0, -0.7, 1, 1)$		5	$(2, -0.7, 1, 1)$
	2	$(0, -0.3, 1, 1)$		6	$(2, -0.3, 1, 1)$
	3	$(0, 0.3, 1, 1)$		7	$(2, 0.3, 1, 1)$
	4	$(0, 0.7, 1, 1)$		8	$(2, 0.7, 1, 1)$

		Assumed model, used in transforming data							
		1	2	3	4	5	6	7	8
True model, used in generating data	1	<u>.0508</u>	.0984	.0897	.0372	.0012	.0042	.0145	.0544
	2	.0645	<u>.0498</u>	.0249	.0042	.0000	.0002	.0007	.0047
	3	.2126	.1527	<u>.0499</u>	.0052	.0002	.0003	.0006	.0012
	4	.4461	.3802	.2112	<u>.0500</u>	.0134	.0138	.0123	.0065
	5	.5310	.5062	.3966	.2241	<u>.0494</u>	.0750	.1044	.1171
	6	.5300	.4808	.3368	.1453	.0411	<u>.0500</u>	.0535	.0422
	7	.5565	.5008	.3404	.1345	.0522	.0543	<u>.0495</u>	.0296
	8	.6161	.5648	.4082	.1872	.0942	.0927	.0824	<u>.0494</u>

Observe from Table 2 that when r is actually zero (as in models 1 to 4) but is erroneously assumed equal to 2 (as in models 5 to 8), the empirical significance level decreases. This happens because the actual variances of the generated data from the $r = 0$ models are smaller than the "estimated" variances under the $r = 2$ models;

and in these circumstances fewer rejections of the null hypothesis will occur. Similarly, when φ is over-estimated, the empirical significance level tends to increase (though not invariably). Table 2 only shows results for models with $K = 1$. Other results, not presented here, indicate that when K is overestimated (as 1, when data are generated using $K = 0$) the empirical significance level tends to decrease. These various trends reflect corresponding trends in the covariance matrices of the various models.

When the null hypothesis $H_0 : \mu_1 = \mu_2$ is true, the probability (assuming the correct model) that any test leads to a rejection of H_0 at the 5% level is 0.05. With 200000 tests, when H_0 is true, the distribution of the number of rejections NR , is binomial with mean 200000×0.05 and variance $200000 \times 0.05 \times 0.95$; $Bi(10000, 9500)$. Hence the standard deviation of NR is 97.5 and an approximate 95% percentile range for NR is $[9805, 10195]$. (The corresponding 95% range for the empirical significance level is $[0.0490, 0.0510]$.) One out of 32 values of NR , obtained when testing H_0 with the assumed model equal to the true model, lay outside the 95% range: NR was 10197 for data from $(\sigma_a^2, \varphi, \sigma_e^2, K) = (2, -0.7, 1, 0)$.

Both the integration method and the transformation method, as described above, concentrate on achieving a desired overall significance level when the null hypothesis H_0 is true. The relative power of the two methods has not been considered. With the integration method, deviations from H_0 might be expected to manifest themselves simply in increased absolute values of the test statistics S_j . However, it is less clear how deviations from H_0 may influence the test statistics S_j^* for the transformation method. Power comparisons, based on simulation results, are made between the two methods in 2.4.

2.3 Recursive integration of the multivariate normal density in a special case

Recursive integration in a special case

The problem of obtaining appropriate nominal significance levels for clinical trials with a fixed number of subjects and up to m repeated measurements per subject (with a known or assumed error structure), may be thought of as equivalent to the problem of evaluating an m -dimensional integral of an m -variate normal density (cf 2.2, p49).

The integral of interest in the context of repeated testing is of the form

$$I = \int_{-z_m/\sqrt{v_m}}^{z_m/\sqrt{v_m}} \cdots \int_{-z_2/\sqrt{v_2}}^{z_2/\sqrt{v_2}} \int_{-z_1/\sqrt{v_1}}^{z_1/\sqrt{v_1}} f_{\underline{S}}(\underline{s}) d\underline{s}, \quad (1)$$

where $\underline{S} = (S_1, S_2, \dots, S_m)'$ is the (multivariate normal) vector of cumulative sums of X 's, i.e. $S_1 = X_1$, $S_2 = X_1 + X_2$, ..., $S_m = X_1 + X_2 + \dots + X_m$, X_j is the j^{th} measurement on the subject (or the difference between sums of measurements on treatments 1 and 2 at the j^{th} test),

$$v_j = \text{var}(S_j)$$

and $z_1 = z_2 = \dots = z_m$ if the nominal significance levels are to be equal.

Schervish's (1984) algorithm MULNOR is written specifically to integrate a general multivariate normal density. However, with more than 4 dimensions and accuracy as good as 0.00001, computer time required by MULNOR is usually prohibitive (for 5 dimensions, usually more than 1 hour on the VAX (cf 2.2, ps 52,53)). This practical difficulty may be surmounted at the cost of imposing certain constraints on the correlation structure of the S 's.

In general,

$$\left. \begin{aligned} f(s_3, s_2, s_1) &= f(s_3 | s_2, s_1) f(s_2 | s_1) f(s_1) \\ f(s_m, s_{m-1}, \dots, s_2, s_1) &= f(s_m | s_{m-1}, \dots, s_2, s_1) \cdots f(s_2 | s_1) f(s_1) \end{aligned} \right\}. \quad (2)$$

Let $r_{ij,k}$ denote the partial correlation of S_i and S_j given S_k . If the constraints

$$r_{1j,j-1} = r_{2j,j-1} = \dots = r_{j-2,j,j-1} = 0, \quad j = 3, \dots, m \quad (3)$$

hold, then

$$f(s_j | s_{j-1}, \dots, s_1) = f(s_j | s_{j-1}), \quad j = 3, \dots, m$$

and (2) becomes

$$\left. \begin{aligned} f(s_3, s_2, s_1) &= f(s_3 | s_2) f(s_2 | s_1) f(s_1) \\ f(s_m, s_{m-1}, \dots, s_2, s_1) &= f(s_m | s_{m-1}) \dots f(s_2 | s_1) f(s_1) \end{aligned} \right\}. \quad (4)$$

In this case, the integral I of (1) may be expressed recursively (in a manner analogous to that used by Armitage, McPherson and Rowe, 1969, ps 237,238) as follows.

Suppose first that $m = 2$. Then

$$I = \int_{-z_2/\sqrt{v_2}}^{z_2/\sqrt{v_2}} f_{S_2}(s_2) ds_2, \quad (5)$$

where

$$\begin{aligned} f_{S_2}(s_2) &= \int_{-z_1/\sqrt{v_1}}^{z_1/\sqrt{v_1}} f_{S_2|S_1}(s_2 | s_1) f_{S_1}(s_1) ds_1 \\ &= \int_{-z_1/\sqrt{v_1}}^{z_1/\sqrt{v_1}} f_{X_2|S_1}(s_2 - s_1 | s_1) f_{S_1}(s_1) ds_1 \end{aligned} \quad (6)$$

is obtained for s_2 at a finite number of points on the interval $[-z_2/\sqrt{v_2}, z_2/\sqrt{v_2}]$. In (6), both $X_2|S_1$ and S_1 have normal distributions with known mean and variance. (Note that if $(X, Y)'$ is bivariate normal,

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N \left[\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right],$$

$$\text{then } X|Y=y \sim N \left[\mu_X + \left[\frac{\rho\sigma_X}{\sigma_Y} \right] (y - \mu_Y), \sigma_X^2(1 - \rho^2) \right].)$$

Integrals (5) and (6) are both evaluated numerically using integrand values at points spaced along the range of integration.

When $m = 3$, I may be written

$$I = \int_{-z_3/\sqrt{v_3}}^{z_3/\sqrt{v_3}} f_{S_3}(s_3) ds_3, \quad (7)$$

where

$$\begin{aligned}
f_{S_3}(s_3) &= \int_{-z_2/\sqrt{v_2}}^{z_2/\sqrt{v_2}} f_{S_3|S_2}(s_3|s_2) f_{S_2}(s_2) ds_2 \\
&= \int_{-z_2/\sqrt{v_2}}^{z_2/\sqrt{v_2}} f_{X_3|S_2}(s_3-s_2|s_2) f_{S_2}(s_2) ds_2
\end{aligned} \tag{8}$$

is obtained at a finite number of points in $[-z_3/\sqrt{v_3}, z_3/\sqrt{v_3}]$. In (8), the distribution of $X_3|S_2$ is normal, with known mean and variance, while $f(s_2)$ is determined in discrete form on $[-z_2/\sqrt{v_2}, z_2/\sqrt{v_2}]$ as above. (The consequence (4) of constraints (3) comes into play in (8); without it, (7) gives the wrong value for I.)

Integrals for m greater than 3 are similarly defined recursively. Thus, for any m ,

$$I = \int_{-z_m/\sqrt{v_m}}^{z_m/\sqrt{v_m}} f_{S_m}(s_m) ds_m ,$$

where

$$f_{S_j}(s_j) = \int_{-z_j/\sqrt{v_j}}^{z_j/\sqrt{v_j}} f_{X_j|S_{j-1}}(s_j-s_{j-1}|s_{j-1}) f_{S_{j-1}}(s_{j-1}) ds_{j-1}, \quad j=2, \dots, m.$$

The recursive method of integration depends on equation (4). This in turn holds if (and only if) (3) is true. To see the implication of this for the correlations between the X 's, consider the case with $m = 5$. In this case (3) requires that

$$r_{13.2} = 0 ,$$

$$r_{14.3} = r_{24.3} = 0$$

$$\text{and} \quad r_{15.4} = r_{25.4} = r_{35.4} = 0 .$$

(Recall that the r 's are partial correlations between the S 's, and $S_j = X_1 + X_2 + \dots + X_j$.)

Note that in general,

$$r_{ij.k} = (r_{ij} - r_{ik}r_{kj}) / \left\{ (1 - r_{kj}^2)(1 - r_{ik}^2) \right\}^{\frac{1}{2}} \quad i < k < j$$

(see e.g. Kendall and Stuart, 1973, Volume 2, p331); so

$$r_{ij.k} = 0 \iff r_{ij} = r_{ik}r_{kj} \quad i < k < j . \tag{9}$$

Let ρ_{ij} denote the correlation between X_i and X_j . Then, using (9) yields, after some algebra,

$$r_{13.2} = 0 \iff \rho_{13} = \rho_{23} \equiv \rho_3 \text{ say.}$$

Now, suppose the (lower triangle of the) correlation matrix for X_1 , X_2 and X_3 is given by

$$\underline{C}_3 = \begin{bmatrix} 1 & & \\ \rho_2 & 1 & \\ \rho_3 & \rho_3 & 1 \end{bmatrix};$$

$$\text{then } r_{14.3} = r_{24.3} = 0 \iff \rho_{14} = \rho_{24} \equiv \rho_4 \text{ say,}$$

$$\text{and } \rho_{34} = \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3} \right] \rho_4.$$

Further, if the correlation matrix $\underline{C}_4$ for X_1 , X_2 , X_3 and X_4 is given by

$$\underline{C}_4 = \begin{bmatrix} 1 & & & \\ \rho_2 & 1 & & \\ \rho_3 & \rho_3 & 1 & \\ \rho_4 & \rho_4 & \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3} \right] \rho_4 & 1 \end{bmatrix},$$

then $r_{15.4} = r_{25.4} = r_{35.4} = 0$ yields 3 equations in the 4 "unknowns" ρ_{15} , ρ_{25} , ρ_{35} and ρ_{45} :

$$\rho_{15} \left[3+\rho_2+3\rho_3+3\rho_4 + \left[\frac{2(1+2\rho_3)}{1+\rho_2+\rho_3} \right] \rho_4 \right] = (1+\rho_2+\rho_3+\rho_4)(\rho_{25}+\rho_{35}+\rho_{45}), \quad (10)$$

$$(\rho_{15}+\rho_{25}) \left[1+\rho_3+\rho_4 + \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3} \right] \rho_4 \right] = (1+\rho_2+\rho_3+\rho_4)(\rho_{35}+\rho_{45}) \quad (11)$$

and

$$(\rho_{15}+\rho_{25}+\rho_{35}) \left[1+2\rho_4 + \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3} \right] \rho_4 \right] = \left[3+2\rho_2+4\rho_3+2\rho_4 + \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3} \right] \rho_4 \right] \rho_{45}. \quad (12)$$

Subtracting (11) from (10) gives $\rho_{15} = \rho_{25} \equiv \rho_5$ say, but solutions for ρ_{35} and ρ_{45} are complicated functions of ρ_2 , ρ_3 , ρ_4 and ρ_5 .

In general, to move from an integral involving d dimensions to one involving $d+1$ dimensions using the recursive method, we must have

$$r_{1,d+1.d} = r_{2,d+1.d} = \dots = r_{d-1,d+1.d} = 0.$$

This yields $d-1$ equations in d "unknowns", $\rho_{1,d+1}$, $\dots$, $\rho_{d,d+1}$. On fixing one

of the unknowns, $\rho_{j,d+1} = \rho_{d+1}$, say, the others must be solutions (if they exist) to $d-1$ equations involving the known correlations $\rho_2, \rho_3, \dots, \rho_{d+1}$. These equations may be solved numerically, by using a standard computer subroutine for finding the zero of a function of $d-1$ non-linear functions in $d-1$ variables. However, despite being free to choose $\rho_{j,d+1}$ for some j , it seems unlikely that the correlation matrix for S's, corresponding to the solution in the ρ 's, will be of interest in many applications.

For the application in mind, repeated testing with a fixed number of subjects, the hypothesized $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model (of 1.3) supposes a covariance matrix for $\underline{X} = (X_1, X_2, \dots, X_m)'$ of the form

$$\underline{\Sigma} = \sigma_a^2 \underline{1} \underline{1}' + \left[\frac{\sigma_e^2}{1-K\varphi^2} \right] \underline{V}_c, \quad (13)$$

where

$$\underline{V}_c = \begin{bmatrix} 1 & \varphi & \varphi^2 & \dots & \varphi^{m-1} \\ \varphi & 1-K\varphi^2+\varphi^2 & \varphi(1-K\varphi^2+\varphi^2) & \dots & . \\ \varphi^2 & \varphi(1-K\varphi^2+\varphi^2) & 1-K\varphi^2+\varphi^2-K\varphi^4+\varphi^4 & \dots & . \\ \varphi^3 & \varphi^2(1-K\varphi^2+\varphi^2) & \varphi(1-K\varphi^2+\varphi^2-K\varphi^4+\varphi^4) & \dots & . \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi^{m-1} & \varphi^{m-2}(1-K\varphi^2+\varphi^2) & \varphi^{m-3}(1-K\varphi^2+\varphi^2-K\varphi^4+\varphi^4) & \dots & (1-K\varphi^2+\varphi^2-\dots \\ & & & & -K\varphi^{2m-2}+\varphi^{2m-2}) \end{bmatrix}.$$

For certain values of σ_a^2 , φ , σ_e^2 and K , the correlation matrix for $\underline{S}$ corresponding to $\underline{\Sigma}$ of (13), may be well approximated by a matrix satisfying the restrictions (3). However, the accuracy attainable in general by such an approximation might not be sufficient to warrant the effort involved in deriving and testing the approximation.

For the special case $\varphi = 0$ (when the value of K becomes irrelevant), Armitage, Stratton and Worthington (1985) show that

$$r_{ij} = r_{ik} r_{kj} \quad \text{for all } i \leq k \leq j,$$

where the r 's are correlations between the S's. Hence, from (9), $r_{ij,j-1} = 0$ for

all $i < j-1$, (3) is satisfied, and for this case, the recursive method of integration "works" exactly.

Consider an alternative means of establishing (3) for the case $\varphi = 0$. When $\varphi = 0$, the correlation matrix for X's may be written

$$\underline{C}_\rho = \begin{bmatrix} 1 & & & & \\ \rho & 1 & & & \\ \rho & \rho & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix},$$

(corresponding to the covariance matrix

$$\underline{\Sigma} = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & & & & \\ \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & & & \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma_e^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \dots & \sigma_a^2 + \sigma_e^2 \end{bmatrix};$$

$$\text{and } \text{cov}(S_i, S_j) = i + (j-1)i\rho \quad \text{and} \quad \text{var}(S_i) = i[1 + (i-1)\rho].$$

$$\text{Thus } r_{ij} = \frac{\text{cov}(S_i, S_j)}{(\text{var}(S_i) \text{var}(S_j))^{\frac{1}{2}}} = \frac{i + (j-1)i\rho}{(i[1 + (i-1)\rho] j[1 + (j-1)\rho])^{\frac{1}{2}}},$$

$$\text{and } r_{ij} = r_{i, j-1} r_{j-1, j} \quad (i < j-1)$$

$$\iff i + (j-1)i\rho = \frac{[i + (j-1)i\rho] [(j-1) + (j-1)^2\rho]}{(j-1)[1 + (j-2)\rho]}$$

$$\iff i + (j-2)i\rho + (j-1)i\rho + (j-1)(j-2)i\rho^2 = i + (j-2)i\rho + (j-1)i\rho + (j-1)(j-2)i\rho^2;$$

and again, it follows from (9) that (3) holds.

The fact that the recursive method of integration works exactly, when $\varphi = 0$, proves useful in 2.4.

Implementation of recursive integration

Recursive integration was programmed using the NAG subroutine D01GAF. (This routine integrates a function which is specified numerically at 4 or more points over the whole of its specified range, using 3rd-order finite difference formulae with error estimates. It uses a method due to Gill and Miller.)

In (a) to (c) below, evaluations are given of particular 4-dimensional multivariate normal integrals, computed using both the recursive method and the Schervish routine MULNOR. In (d), recursive results are compared with those of Armitage, McPherson and Rowe (1969), for the case of iid X's and integrals up to 100 dimensions.

In (a) to (d):

$\underline{C}$ = correlation matrix for X's,

$\underline{R}$ = corresponding correlation matrix for S's,

Rec() = integral as evaluated by the recursive method (with steplength H, or number of points in inner integral N, in brackets) and

Sch() = integral as evaluated by Schervish's (1984) MULNOR algorithm (with accuracy in brackets).

$$(a) \quad \int_{-1.5}^{1.5} \int_{-1.5}^{1.5} \int_{-1.5}^{1.5} \int_{-1.0}^{1.2} f_{\underline{S}}(\underline{s}) d\underline{s}$$

$$\underline{C} = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho & \rho & \rho & 1 \end{bmatrix} \quad \rho = 0.3 \quad \underline{R} = \begin{bmatrix} 1.00 & & & \\ 0.74 & 1.00 & & \\ 0.63 & 0.85 & 1.00 & \\ 0.57 & 0.77 & 0.90 & 1.00 \end{bmatrix}$$

$$\text{Rec} (H = 0.025) = 0.61960175$$

$$\text{Sch} (10^{-6}) = 0.61960173$$

$$(b) \quad \int_{-2}^1 \int_{-2}^2 \int_{-1}^1 \int_{-2}^2 f_{\underline{S}}(\underline{s}) d\underline{s}$$

$$\underline{C} = \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho & \rho & 1 & \\ \rho^* & \rho^* & \rho^* & 1 \end{bmatrix} \quad \begin{matrix} \rho = -0.2 \\ \rho^* = 0.4 \end{matrix} \quad \underline{R} = \begin{bmatrix} 1.00 & & & \\ 0.63 & 1.00 & & \\ 0.45 & 0.71 & 1.00 & \\ 0.44 & 0.69 & 0.98 & 1.00 \end{bmatrix}$$

Rec (H = 0.025) = 0.59173179

Sch (10⁻⁶) = 0.59173161

(c) $\int_{-1.96}^{1.96} \int_{-1.96}^{1.96} \int_{-1.96}^{1.96} \int_{-1.96}^{1.96} f_{\underline{S}}(\underline{s}) \, d\underline{s}$

$\underline{C} = \begin{bmatrix} 1 & & & \\ \rho_2 & 1 & & \\ \rho_3 & \rho_3 & 1 & \\ \rho_4 & \rho_4 & \left[\frac{1+2\rho_3}{1+\rho_2+\rho_3}\right]\rho_4 & 1 \end{bmatrix}$ $\rho_2 = 0.4$ $\rho_3 = 0.1$ $\rho_4 = -0.5$ $\underline{R} = \begin{bmatrix} 1.00 & & & \\ 0.84 & 1.00 & & \\ 0.73 & 0.87 & 1.00 & \\ 0.65 & 0.77 & 0.88 & 1.00 \end{bmatrix}$

Rec (N = 51) = 0.8874260

Sch (10⁻⁶) = 0.8874261

(d) $1 - \underbrace{\int_{-1.96}^{1.96} \dots \int_{-1.96}^{1.96}}_{m \leq 100} f_{\underline{S}}(\underline{s}) \, d\underline{s}$

$\underline{C} = \left[\begin{array}{cccc} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \dots & 1 \end{array} \right] \quad m \leq 100$

m	AMR (1969)	Rec (H=0.1)	Rec (H=0.025)
2	.08312	.08311	.08311
3	.10726	.10725	.10725
4	.12617	.12616	.12616
5	.14169	.14168	.14168
10	.19336	.19334	.19334
15	.22509	.22508	.22508
20	.24791	.24790	.24789
30	.28016	.28016	.28014
40	.30293	.30293	.30291
50	.32045	.32046	.32043
60	.33464	.33465	.33462
80	.35674	.35672	.35672
100	.37362	.37349	.37360

(H = 0.1 used 52 seconds CPU time on the VAX; H = 0.025 used 690 seconds)

If conditions (3) on the partial correlations of S's are eased to

$$r_{1j,j-1,j-2} = r_{2j,j-1,j-2} = \dots = r_{j-3,j,j-1,j-2} = 0, \quad j = 4, \dots, m,$$

a recursive approach to the integration of $f(s)$ may still be used, by at the d^{th} stage, retaining and using the function values obtained at the $(d-1)^{\text{th}}$ and $(d-2)^{\text{th}}$ stages. This would require more CPU time than the scheme above depending on (3); but would allow integration of $f(s)$ for $\underline{S}$ corresponding to X's correlated by the more general [than for (3)]

$$\underline{C}_m = \begin{bmatrix} 1 & & & & & & \\ \rho_{21} & 1 & & & & & \\ \rho_{31} & \rho_{32} & 1 & & & & \\ \rho_{41} & \rho_{42} & \tau_{43} & 1 & & & \\ \rho_{51} & \rho_{52} & \tau_{53} & \tau_{54} & 1 & & \\ \vdots & \vdots & \vdots & \vdots & & \ddots & \\ \rho_{m1} & \rho_{m2} & \tau_{m3} & \tau_{m4} & \tau_{m5} & \dots & 1 \end{bmatrix},$$

where ρ_{ij} , $i = 2, \dots, m$, $j = 1, 2$ (or any two correlations per row) may be chosen by the user, and the τ 's are fixed functions of the ρ 's above or adjacent to them.

2.4 Testing in practice, when the model for data is unknown a priori

Three possible procedures

It is hardly feasible to implement the integration method of 2.2 for obtaining nominal significance levels, for data with *up to 5 or more* repeated measurements, because of the difficulty in integrating the general multivariate normal in high dimensions. In the special case when φ in the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model is zero, the recursive integration algorithm of 2.3 resolves this integration difficulty. The transformation method of 2.2 avoids any integration problems by "transforming the data to independence" and using the standard group sequential nominal significance levels for iid data.

However, both the integration method and the transformation method, as presented in 2.2, assume data are generated from a *known* $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model, and this assumption is unlikely to be justified in practice. Instead, for any particular trial, parameters of the best-fitting $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model may need to be estimated sequentially, each time a new set of measurements is recorded. As new data are collected and used in estimation, parameter estimates will generally become more reliable. Thus in these circumstances, it seems advisable to exercise more caution than might otherwise (if the correct model were known a priori) seem necessary, before rejecting the null hypothesis on the basis of evidence from the early tests.

Instead, then, of using tests based on equal nominal significance levels, the integration method may be better implemented using the following three steps.

1. Suppose there are 2 treatments, and n subjects per treatment. Let the "rate of spending" the overall significance level α be predetermined. Thus if m is the maximum number of measurements per subject, choose $\alpha_1^*, \alpha_2^*, \dots, \alpha_m^*$, where α_j^* is the "absorption probability" for the j^{th} test, such that

$$\alpha_1^* + \alpha_2^* + \dots + \alpha_m^* = \alpha.$$

The α^* 's may be chosen as an increasing sequence, implying more stringent rejection criteria for the earlier tests.

2. Define $S_j = X_1 + X_2 + \dots + X_j$, where X_i is the difference between sums of the i^{th} measurements on treatments 1 and 2; and define $\underline{S}_j = (S_1, \dots, S_j)'$, $j=1, \dots, m$.

Observe the first measurement on each subject.

Use these measurements to estimate a common variance σ^2 , assuming the data are iid. (At this stage it is impossible to estimate between- and within-subject variances separately.)

Determine z_1^* from

$$\int_{-z_1^*/\hat{v}_{11}}^{z_1^*/\hat{v}_{11}} f_{S_1}(s_1) ds_1 = 1 - \alpha_1^* \quad (\text{cf 2.2, p48, (1)}),$$

where $\hat{v}_{11}$ is the estimated variance of S_1 , obtained using the estimate of σ^2 . (Note that this will be the z_1^* appropriate for iid data.)

Test $|s_1|/\sqrt{\hat{v}_{11}}$ against z_1^* . If $|s_1|/\sqrt{\hat{v}_{11}} > z_1^*$, stop the trial and reject the null hypothesis. Otherwise, set $j = 1$.

3. Set $j = j + 1$.

Observe the j^{th} measurement on each subject.

Use all the measurements made so far to estimate, by maximum likelihood, the between-subject variance σ_a^2 and the within-subject variance σ_e^2 (assuming $\varphi=0$).

Use the recursive integration algorithm of 2.3 to determine z_j^* satisfying

$$\int_{-z_j^*/\hat{v}_{jj}}^{z_j^*/\hat{v}_{jj}} \dots \int_{-z_1^*/\hat{v}_{j1}}^{z_1^*/\hat{v}_{j1}} f_{\underline{S}_j}(\underline{s}_j) d\underline{s}_j = 1 - (\alpha_1^* + \dots + \alpha_j^*)$$

where $\hat{v}_{ji}$ is the estimated variance of S_i using the new mles of σ_a^2 and σ_e^2 .

If $|s_j|/\sqrt{\hat{v}_{jj}} > z_j^*$, stop the trial and reject the null hypothesis. Otherwise, if $j < m$, repeat step 3.

Otherwise, stop the trial and do not reject the null hypothesis.

To apply the transformation method of testing, proceed with steps 1 and 2 as above, and replace 3 by 3_t as follows.

- 3_t . Set $j = j + 1$.

Observe the j^{th} measurement on each subject.

Use all the measurements accrued so far to estimate σ_a^2 and σ_e^2 (assuming $\varphi=0$) if $j = 2$, or σ_a^2 , φ , σ_e^2 and K if $j > 2$, by maximum likelihood.

Determine the transformation matrix corresponding to the estimated model (cf 2.2, ps 53–55), and transform the data using this matrix. Denote the $\underline{S}$ vector corresponding to the transformed data by $\underline{S}^*$. Use the recursive integration algorithm of 2.3 to determine z_j^* satisfying

$$\int_{-z_j^*/\sqrt{2nj}}^{z_j^*/\sqrt{2nj}} \cdots \int_{-z_1^*/\sqrt{2n}}^{z_1^*/\sqrt{2n}} f_{\underline{S}_j^*}(\underline{s}_j^*) d\underline{s}_j^* = 1 - (\alpha_1^* + \dots + \alpha_j^*) .$$

If $|s_j^*|/\sqrt{2nj} > z_j^*$, stop the trial and reject the null hypothesis. Otherwise, if $j < m$, repeat step 3_t.

Otherwise, stop the trial and do not reject the null hypothesis.

Note that in 3_t, $\text{var}(S_j^*) = 2nj$, because $\underline{S}^*$ is the vector of sums corresponding to iid X^* 's. Note too that, for the transformation procedure, the values z_1^* , z_2^* , ..., z_m^* do not depend on the observed data. (This is useful when performing many simulations with the same value of $\underline{\alpha}^*$.)

Another procedure for testing, in addition to the integration and transformation procedures described above, arises if the particular structure of the data is ignored and all measurements are assumed to be iid. This does not seem plausible in general, but the corresponding procedure is considered here for comparison with the first two procedures. For this iid procedure, the S 's are the same as in 3, but the $\hat{v}_{ji}$'s of 3 depend only on one estimated variance parameter, σ^2 say. The z_j^* 's for comparison with $|s_j|/\sqrt{\hat{v}_{jj}}$ are the same as in 3_t.

So three possible procedures are envisaged for sequential testing with repeated measurements data:

- (a) integration,
- (b) transformation
- and (c) iid.

It is impossible to estimate by maximum likelihood all 4 parameters of the $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model when only 2 measurements are available per subject. So, for the second test, the transformation procedure (b), like the integration procedure (a), uses only $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$, assuming $\varphi = 0$. In general, though, the two procedures are not equivalent, even at the second stage. To see this, observe that for the second stage the estimated variance for X's is

$$\hat{\Sigma} = 2n \begin{bmatrix} \hat{\sigma}_a^2 + \hat{\sigma}_e^2 & \hat{\sigma}_a^2 \\ \hat{\sigma}_a^2 & \hat{\sigma}_a^2 + \hat{\sigma}_e^2 \end{bmatrix}.$$

Thus $\text{var}(S_1) \equiv \hat{v}_{21} = 2n(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)$ and $\text{var}(S_2) \equiv \hat{v}_{22} = 2n(4\hat{\sigma}_a^2 + 2\hat{\sigma}_e^2)$.

For the *integration* procedure, z_2^* is obtained from

$$\int_{-z_2^*/\hat{v}_{22}}^{z_2^*/\hat{v}_{22}} \int_{-z_1^*/\hat{v}_{21}}^{z_1^*/\hat{v}_{21}} f_{\underline{S}_2}(\underline{s}_2) d\underline{s}_2 = 1 - (\alpha_1^* + \alpha_2^*).$$

Let this z_2^* be denoted by z_{2i}^* . Then $s_2/\hat{v}_{22} \equiv (x_1 + x_2)/\hat{v}_{22}$ is tested against z_{2i}^* .

For the *transformation* procedure, the appropriate transformation matrix for the second stage test is

$$\underline{T} = \frac{1}{(2n)^{\frac{1}{2}}} \begin{bmatrix} t_{11} & 0 \\ t_{21} & t_{22} \end{bmatrix},$$

$$\text{where } t_{11} = \frac{1}{(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)^{\frac{1}{2}}}, \quad t_{21} = \frac{-\hat{\sigma}_a^2}{\{[(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)^2 - (\hat{\sigma}_a^2)^2](\hat{\sigma}_a^2 + \hat{\sigma}_e^2)\}^{\frac{1}{2}}}$$

$$\text{and } t_{22} = \left\{ \frac{\hat{\sigma}_a^2 + \hat{\sigma}_e^2}{(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)^2 - (\hat{\sigma}_a^2)^2} \right\}^{\frac{1}{2}} \quad (\text{cf 2.2, p55}).$$

The transformation procedure tests $[(t_{11} + t_{21})x_1 + t_{22}x_2]/\sqrt{4n}$ against z_{2i}^* , the z value appropriate for X's which are iid.

A precise comparison between procedures (a) and (b) at the second stage cannot be made algebraically – z_{2i}^* is not a simple function of $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ – but simulations confirm that they are not equivalent. For example, for data with $n = 100$ pairs of subjects generated from the model (1,0.5,1,0), with $(\alpha_1^*, \alpha_2^*) = (0.010, 0.015)$, it was found that $\hat{\sigma}_a^2 = 0.96$, $\hat{\sigma}_e^2 = 0.34$, $x_1 = 16.84$, $x_2 = 42.54$ and $z_{2i}^* = 2.27$; so

$$\frac{x_1 + x_2}{[2n(4\hat{\sigma}_a^2 + 2\hat{\sigma}_e^2)]^{\frac{1}{2}}} = \frac{59.38}{30.07} = 1.97 < z_{2i}^* \equiv 2.27$$

and the integration procedure does not reject the null hypothesis $\mu = 0$; however,

$$\frac{[(t_{11} + t_{21})x_1 + t_{22}x_2]}{(4n)^{\frac{1}{2}}} = 2.69 > z_{2t}^* \equiv 2.36$$

and the transformation procedure clearly rejects the null hypothesis.

(Note that, when $\hat{\sigma}_a^2 = 0$, $z_{2i}^* = z_{2t}^*$ and procedures (a) and (b) are equivalent at the second stage.)

Simulation results

Simulations were carried out using data generated from each of six $(\sigma_a^2, \varphi, \sigma_e^2, K)$ models, with $m = 3, 5$ or 10 repeated measurements per subject. Significance tests were performed using procedures (a), (b) and (c). A theoretical significance level of 0.05 was used, with intermediate absorption probabilities

$$\underline{\alpha}_3^* = (.010, .015, .025)', \quad \underline{\alpha}_5^* = (.005, .0075, .010, .0125, .015)'$$

and

$$\underline{\alpha}_{10}^* = (.001, .002, .003, .004, .005, .005, .006, .007, .008, .009)'$$

The z values corresponding to these absorption probability vectors, for iid data, were determined as

$$\underline{z}_3^* = (2.576, 2.359, 2.094)', \quad \underline{z}_5^* = (2.808, 2.615, 2.447, 2.298, 2.165)'$$

and

$\underline{z}_{10}^* = (3.290, 3.058, 2.884, 2.740, 2.615, 2.548, 2.454, 2.362, 2.275, 2.192)'$,
respectively.

Empirical significance levels from test procedures (a), (b) and (c) are presented in Tables 1 and 2 (on pages 74 and 75, respectively). Simulations were performed separately for procedures (a) and (b), with (c) being used with both sets of simulations. (This situation arose by accident rather than by design. On first running the simulations, the version of (b) actually implemented was slightly different from the intended transformation procedure, described above. To rectify this, only the simulations for (b) were repeated – computer time required for implementation of (a) was between 30 and 50 times more than time required for (b). Results for (b) improved slightly when the intended procedure was used.) The simulation program used in producing the results is described in detail in Appendix 4.

Suppose, with a given test procedure, the probability of rejecting the null hypothesis $\underline{\mu} = \underline{0}$, given $\underline{\mu} = \underline{0}$, is equal to 0.05. Then, if NSIM simulations are performed, the number of rejections NR is binomially distributed, with mean $NSIM \times 0.05$ and variance $NSIM \times 0.05 \times 0.95$ (see also 2.2, p58). Hence intervals based on approximate 0.025 and 0.975 percentile points for the empirical significance levels of Table 1 are given by:

	[0.0438, 0.0562]	for NSIM = 5000,
	[0.0403, 0.0597]	for NSIM = 2000
and	[0.036, 0.064]	for NSIM = 1000.

For the integration procedure, 2 out of the 18 empirical significance levels of Table 1 fall outside their corresponding 0.95 percentile ranges (those for $m=3, (1, -0.5, 1, 1)$ and $m=10, (1, 0.5, 1, 0)$); while 4 out of 18 fall outside, for the transformation procedure (those for $m=3$, middle 4 models). For both procedures, the proportion of rejections at the first test tends to be higher than α_1^* , for data from all except the iid model (0,0,1,1). For instance, the numbers of rejections at the first test, with 5000

(continued on page 76)

Table 1 *Empirical significance levels from NSIM simulations, using procedures (a), (b) and (c), and theoretical overall significance level 0.05, with data generated using mean vector $\underline{\mu} = \underline{0}_m$*

Simulations were performed separately for the integration (a) and transformation (b) procedures. Each simulation had $n=100$ subject pairs, with m repeated measurements per subject pair. The intermediate absorption probabilities were $\underline{\alpha}_3^*$, $\underline{\alpha}_5^*$ or $\underline{\alpha}_{10}^*$ (see text). A $\checkmark$ in the table indicates which of procedures (a) and (b) gives a result closer to the desired 0.05 level.

	Model $(\sigma_a^2, \varphi, \sigma_e^2, K)$	Procedure for testing			
		(a) integ- ration	(c) iid	(b) transfor- mation	(c) iid
m = 3 $\underline{\alpha}_3^*$ NSIM=5000	(0, 0.0, 1, 1)	.0512 $\checkmark$	.0550	.0470	.0468
	(1, -0.5, 1, 1)	.0596	.1082	.0578 $\checkmark$	.1034
	(1, 0.0, 1, 1)	.0562 $\checkmark$	.1484	.0610	.1510
	(1, 0.5, 1, 1)	.0542 $\checkmark$	.1676	.0586	.1764
	(1, 0.5, 1, 0)	.0552 $\checkmark$	.1768	.0578	.1710
	(1, 0.5, 1, 2)	.0540	.1822	.0516 $\checkmark$	.1740
	Overall	.0551 $\checkmark$	.1397	.0556	.1371
m = 5 $\underline{\alpha}_5^*$ NSIM=2000	(0, 0.0, 1, 1)	.0405	.0465	.0485 $\checkmark$	.0505
	(1, -0.5, 1, 1)	.0500 $\checkmark$	.1805	.0545	.1640
	(1, 0.0, 1, 1)	.0570	.2355	.0565 $\checkmark$	.2070
	(1, 0.5, 1, 1)	.0585	.2490	.0565 $\checkmark$	.2550
	(1, 0.5, 1, 0)	.0565	.2590	.0495 $\checkmark$	.2505
	(1, 0.5, 1, 2)	.0545 $\checkmark$	.2540	.0550	.2665
	Overall	.0528 $\checkmark$	.2038	.0534	.1989
m = 10 $\underline{\alpha}_{10}^*$ NSIM=1000	(0, 0.0, 1, 1)	.038	.044	.055 $\checkmark$	.060
	(1, -0.5, 1, 1)	.043	.319	.050 $\checkmark$	.319
	(1, 0.0, 1, 1)	.049 $\checkmark$	.337	.055	.343
	(1, 0.5, 1, 1)	.054 $\checkmark$	.410	.062	.403
	(1, 0.5, 1, 0)	.065	.396	.054 $\checkmark$	.379
	(1, 0.5, 1, 2)	.042	.360	.058	.347
	Overall	.049 $\checkmark$	.311	.056	.309

Table 2 Empirical significance levels from NSIM simulations, using procedures (a), (b) and (c), and theoretical overall significance level $\alpha = 0.05$, with data generated using mean vector $\underline{\mu} = 0.2 \underline{1}_m$.

Simulations were performed separately for the integration (a) and transformation (b) procedures. Each simulation had $n=100$ subject pairs, with m repeated measurements per subject pair. The intermediate absorption probabilities were $\underline{\alpha}_3^*$, $\underline{\alpha}_5^*$ or $\underline{\alpha}_{10}^*$ (see text). A / in the table indicates which of procedures (a) and (b) gives the higher significance level, and so is better with respect to power.

	Model $(\sigma_a^2, \varphi, \sigma_e^2, K)$	Procedure for testing			
		(a) integ- ration	(c) iid	(b) transfor- mation	(c) iid
m = 3 $\underline{\alpha}_3^*$ NSIM=5000	(0, 0.0, 1, 1)	.9168	.9222	.9176/	.9202
	(1, -0.5, 1, 1)	.4360/	.5622	.4310	.5606
	(1, 0.0, 1, 1)	.4182/	.6152	.3876	.6088
	(1, 0.5, 1, 1)	.3244/	.5522	.2986	.5540
	(1, 0.5, 1, 0)	.3360/	.5548	.3242	.5844
	(1, 0.5, 1, 2)	.2814/	.5044	.2696	.5236
	Overall	.4521/	.6185	.4381	.6253
m = 5 $\underline{\alpha}_5^*$ NSIM=2000	(0, 0.0, 1, 1)	.9900	.9915	.9910/	.9925
	(1, -0.5, 1, 1)	.4590	.6840	.4620/	.7095
	(1, 0.0, 1, 1)	.4575/	.7255	.3925	.7170
	(1, 0.5, 1, 1)	.3625/	.6785	.3255	.6685
	(1, 0.5, 1, 0)	.3745/	.6830	.3075	.6630
	(1, 0.5, 1, 2)	.3165	.6235	.3165	.6460
	Overall	.4933/	.7310	.4658	.7328
m = 10 $\underline{\alpha}_{10}^*$ NSIM=1000	(0, 0.0, 1, 1)	1.000	1.000	1.000	1.000
	(1, -0.5, 1, 1)	.465/	.809	.445	.835
	(1, 0.0, 1, 1)	.502/	.854	.401	.848
	(1, 0.5, 1, 1)	.429/	.842	.351	.798
	(1, 0.5, 1, 0)	.407/	.798	.347	.820
	(1, 0.5, 1, 2)	.401/	.790	.384	.829
	Overall	.534/	.849	.488	.855

simulations using $m = 3$, for the 5 non-iid models were:

61, 60, 48, 53 and 63, for the procedure (a) simulations

and 61, 78, 56, 56 and 55, for the procedure (b) simulations;

compared with the expected value of 50 ($\alpha_1^* = 0.01$). In practice, special caution should be exercised before a trial is stopped and the null hypothesis rejected, after only one test.

Table 1 suggests that both procedures (a) and (b) are capable of yielding approximately (to within 1 or 2 per cent) a desired 5% significance level for data from a wide range of models. Procedure (c), assuming iid measurements, yields approximate 5% levels when data are in fact iid, but otherwise gives highly inflated levels, mainly because of its failure to take proper account of between-subject variance.

From Table 2, procedure (a) is generally better than (b) with respect to power, rejecting the null hypothesis more frequently than (b) when the null hypothesis is actually false (specifically, when $\mu = 0.2 \underline{1}_m$, rather than $\underline{0}_m$). This is to be expected, since procedure (a) is based on a sufficient statistic for μ and the test based on this statistic is uniformly most powerful of given size. By the same reasoning, it is anticipated that the superiority of (a) over (b) with respect to power may be maintained for higher powers than in Table 2, of the order of 0.8 say, rather than 0.4. (In Table 2, procedure (b) is "better" than (a) in 3 cases when power is relatively high, but this is thought to be a chance effect, perhaps explained in part by the fact that the significance tests were performed on different simulated data sets for (a) and (b).) On the basis of power, then, the integration procedure is preferred over the transformation procedure.

Each row of results in Table 1 required about 7, 8 and 14 hours of CPU time on the VAX, for $m = 3, 5$ and 10 respectively. (Testing one simulated dataset with all 3 procedures (a), (b) and (c) required, on average, 5, 15 and 50 CPU seconds for $m = 3, 5$ and 10 respectively.) Corresponding times for results in Table 2 were shorter, because of the increased number of rejections of the null hypothesis in Table 2.

(Note that in the simulations, z_j^* satisfying

$$\int_{-z_j^*/\hat{v}_{jj}}^{z_j^*/\hat{v}_{jj}} \cdots \int_{-z_1^*/\hat{v}_{j1}}^{z_1^*/\hat{v}_{j1}} f_{\underline{s}_j}(\underline{s}_j) d\underline{s}_j = 1 - (\alpha_1^* + \dots + \alpha_j^*)$$

were found to an accuracy of 0.001; while individual integrals

$$\int_{-z/\hat{v}_{jj}}^{z/\hat{v}_{jj}} \cdots \int_{-z_1^*/\hat{v}_{j1}}^{z_1^*/\hat{v}_{j1}} f_{\underline{s}_j}(\underline{s}_j) d\underline{s}_j ,$$

evaluated successively in determining z_j^* , were computed to an accuracy of 0.001 (with Schervish's algorithm, $j=2,3$), or with an inner steplength of $0.1/\hat{v}_{jj}$ (with the recursive algorithm, $j>3$). For testing in an individual trial, use of improved accuracies of 0.00001 say, and shorter steplengths of $0.05/\hat{v}_{jj}$ say, should be quite feasible.)

In conclusion, the integration procedure is recommended for repeated testing with a fixed number of subjects.

In practice, flexibility in choosing m , the maximum number of measurements per subject, and $\alpha_1^*, \alpha_2^*, \dots, \alpha_m^*$, the absorption probabilities, rests with the experimenter. As indicated above, it is inadvisable to choose equal α_j^* , since for "early" tests, little is known about the true underlying data structure. Thus an increasing sequence is recommended for $\alpha_1^*, \alpha_2^*, \dots, \alpha_m^*$.

Sequential testing with dental caries data

Recall the 3-year clinical trial with dentifrices, referred to in 1.4. Subjects in the trial were divided into 4 roughly equal groups, 2 of which were allocated "test" dentifrices containing a sodium fluoride additive and 2 of which were given non-sodium-fluoride "control" dentifrices. The trial conclusions (cf Hodge et al, 1980) were that the 2 test treatments were each significantly better than the 2 control treatments, in preventing incidence of dental caries in 11- to 15-year old subjects.

Could the trial have been stopped before the end of the 3-year period, with evidence of significant treatment differences? And how do results from the three possible procedures for sequential testing, described above, compare with one another

and with the published findings from the trial?

Sequential analyses using (a) integration, (b) transformation and (c) iid procedures were carried out on the DMFS data. Four treatments were involved, so 6 distinct pairs of treatments were available for comparison. The dilemma posed by "multiplicity" – in this case, several tests involving the same data – is ignored here. Parmar (1986), for example, discusses the multiplicity problem, and suggests possible ways of allowing for multiple testing.

It was assumed that data from each of the two DMFS groups in any given comparison arose from the same model. (This assumption was used in 2.2, and for the simulations described earlier in 2.4.) Results from sequential testing are presented in Table 3 (on page 79).

Table 3 shows that all three test procedures (a), (b) and (c) come to the same conclusions as Hodge et al with respect to rejection or acceptance of the null hypothesis H_0 , at the 5% level. The fairly close agreement between results from procedures (a) and (c) reflects the fact that the mles of σ_a^2 and σ_e^2 (with φ fixed at 0) for the caries data are consistently such that $\hat{\sigma}_a^2/\hat{\sigma}_e^2$ is small, around 0.15 say. Thus it is almost as though there is no between-subject variance and data are actually iid. However, since $\hat{\sigma}_a^2$ is never actually 0, the iid assumption does rather inflate the test statistic, and Table 3 illustrates how the integration procedure (a) consistently adjusts the (c) statistic downwards to allow for between-subject variance. (The z values for procedure (a) are also adjusted downwards from those for (c), but proportionally less than the statistics.)

Suppose the tests were performed at the 1% rather than the 5% level, with absorption probabilities 0.002, 0.003 and 0.005 say. Then the corresponding values for z_{bc} of Table 3 would be 3.090, 2.924 and 2.715 respectively. Thus, from Table 3, using any of the sequential procedures (a), (b) or (c), the two test treatments 1 and 2 would be judged significantly better than the negative control treatment 4, at the 1% level, as in Hodge et al; though no adjustment has been made here for multiple

(continued on page 80)

Table 3 *Results from sequential testing using procedures (a), (b) and (c) on DMFS data of 1.4*

An overall significance level of $\alpha=0.05$ was used, with intermediate absorption probabilities $\underline{\alpha}^* = (0.01, 0.015, 0.025)^*$; $\underline{z}_a$ and $\underline{z}_{bc}$ are z-values corresponding to $\underline{\alpha}^*$ for tests using procedure (a), and procedures (b) and (c), respectively; $\underline{s}_a$, $\underline{s}_b$ and $\underline{s}_c$ are test statistics from procedures (a), (b) and (c) respectively; (R) indicates rejection of the null hypothesis.

Group comparisons	Test	1	2	3
	$\underline{z}_{bc}$	2.576	2.359	2.094
1 vs 2	$\underline{z}_a$	2.576	2.352	2.069
	$\underline{s}_a$	0.026	0.121	0.390
	$\underline{s}_b$	"	0.124	0.480
	$\underline{s}_c$	"	0.125	0.428
1 vs 3	$\underline{z}_a$	2.576	2.347	2.057
	$\underline{s}_a$	0.985	2.287	2.166 (R)
	$\underline{s}_b$	"	2.346	2.156 (R)
	$\underline{s}_c$	"	2.415 (R)	—
1 vs 4	$\underline{z}_a$	2.576	2.347	2.064
	$\underline{s}_a$	0.088	1.790	3.119 (R)
	$\underline{s}_b$	"	1.893	3.393 (R)
	$\underline{s}_c$	"	1.885	3.531 (R)
2 vs 3	$\underline{z}_a$	2.576	2.349	2.065
	$\underline{s}_a$	0.955	2.224	2.661 (R)
	$\underline{s}_b$	"	2.272	2.780 (R)
	$\underline{s}_c$	"	2.333	2.970 (R)
2 vs 4	$\underline{z}_a$	2.576	2.349	2.073
	$\underline{s}_a$	0.061	1.712	3.673 (R)
	$\underline{s}_b$	"	1.798	3.998 (R)
	$\underline{s}_c$	"	1.790	4.019 (R)
3 vs 4	$\underline{z}_a$	2.576	2.345	2.064
	$\underline{s}_a$	0.921	0.564	0.880
	$\underline{s}_b$	"	0.513	1.183
	$\underline{s}_c$	"	0.599	0.998

testing, whereas Hodge et al apply Duncan's multiple range test.

Table 3 refers to sequential testing under the assumption that data from two treatment groups are generated from the same model. Further analyses were performed to investigate the effect of estimating variance-covariance matrices separately for data in each group. For procedure (a), this led to the use of a pooled estimate of the variance-covariance matrix and test results differed only slightly from those of Table 3. For the transformation procedure (b), however, different transformations for each group of subjects led to unreliable test statistics. Use of different transformations for each group may be inappropriate because of the possibility of a constant "period effect" – an overall increase or decrease in measurements between successive time points – which would normally be balanced between groups using the single transformation procedure, but may be confounded with treatment effect using two transformations. Assuming a single model for data seems preferable.

All three sequential procedures (a), (b) and (c) have been considered here for illustrative purposes. However, the previous recommendation, to use the integration procedure (a) in practice, remains unchanged.

Concluding remarks

The integration procedure for testing in clinical trials with a fixed number of subjects, as proposed in 2.4, assumes that subject measurements are normally distributed. This assumption may not always be justified, but it seems likely that the procedure, like many others based on normal theory, will be robust, in the sense that it will work quite well for non-normal measurements. Robustness, however, needs to be checked, either empirically or theoretically; so it should be of interest to investigate the properties of the procedure when measurements are from binary or exponential distributions, say.

The transformation procedure of 2.4, which gives fairly reliable significance levels when the null hypothesis is true, but does not have as much power to detect alternatives as the integration procedure, may yet prove useful in some circumstances. For instance, suppose the treatment effect is not constant in time. Specifically, suppose there is no treatment effect over the first time period, but there *is* an effect during the second time period. In this case, it may be preferable to test transformed measurements X_1^* and X_2^* separately, rather than testing X_1^* and then $X_1^* + X_2^*$. In general, if it is suspected that the treatment effect may appear suddenly sometime, then X_1^* , X_2^* , X_3^* , ... could be used for *independent* tests, controlling type I error according to Bonferroni's procedure.

Any improvements in numerical methods for evaluating multivariate normal integrals may facilitate improvements in the accuracy of the integration procedure. Number-theoretic methods for multivariate integration, such as those proposed by Hua and Yuan (1981), may be useful in this context. Both the integration and the transformation procedures may yet be refined by improvements in the methods for estimating parameters.

Providing the proposed procedure is considered acceptable, it may be useful to develop and make available an easy-to-use computer package for its implementation. In such a package, the user would specify at each stage of the trial

p the number of measurements per subject available so far

$z_1^*, \dots, z_{p-1}^*$ (for $p > 1$) the z -values corresponding to $\alpha_1^*, \dots, \alpha_{p-1}^*$, derived using the first $p-1$ measurements per subject and α_p^* the p^{th} absorption probability.

The user would further supply the data collected so far in the trial to a specified data file. The main output from the package at each stage would be the result of a significance test, though subsidiary information would be available. Some procedure may need to be adopted in the package

- (a) to transform the data by some scalar factor, or by some other mechanism,
- (b) to allow for missing data and
- (c) to alert the user to the presence of outliers in the data.

Appendix 1 Comparison between the Wilson, Hebel and Sherwin weighted and unweighted least squares estimators of the between-subject variance σ_a^2 (cf 1.1, p3)

Wilson, Hebel and Sherwin (WHS) suggest two possible ways (apart from maximum likelihood) of estimating the between-subject variance σ_a^2 for their model:

(i) **weighted least squares** (cf Rao, 1975, p549, 2.18)

$$\sigma_a^{*2} = \frac{1}{N-1} \sum_{i=1}^N (\mu_i^* - \mu^*)^2 = (\underline{1}' \underline{V}^{*-1} \underline{1})^{-1},$$

where $\mu_i^* = (\underline{1}' \underline{V}^{*-1} \underline{1})^{-1} \underline{1}' \underline{V}^{*-1} \underline{X}_i$, $\mu^* = \frac{1}{N} \sum_{i=1}^N \mu_i^*$

and $\underline{V}^*$ is the estimate (using variance-based estimates φ^* and σ_e^{*2} of φ and σ_e^2) of the covariance matrix of the within-subject errors;

and, "with somewhat less efficiency" (p561),

(ii) **unweighted least squares**

$$\tilde{\sigma}_a^2 = \frac{1}{N-1} \sum_{i=1}^N (\tilde{\mu}_i - \tilde{\mu})^2 = \frac{\sigma_e^{*2}}{m},$$

where $\tilde{\mu}_i = \frac{1}{m} \sum_{j=1}^m X_{ij}$ and $\tilde{\mu} = \frac{1}{N} \sum_{i=1}^N \tilde{\mu}_i$.

The unweighted least squares estimator $\tilde{\sigma}_a^2$ is used by WHS to provide a fixed estimate of σ_a^2 in their algorithm for finding "maximum likelihood" estimates of φ and σ_e^2 . However in Table 1 of 1.1 (p5), $\tilde{\sigma}_a^2$ is seen to be positively biased, especially for high values of φ and small values of m .

Simulations were performed (generating data with $m=3$ measurements per subject), to determine whether σ_a^{*2} actually improves on $\tilde{\sigma}_a^2$. The results, presented in Table 1, indicate that it does not. This is mainly because φ is estimated, and not known (as assumed in Rao's paper). Note that

$$(\underline{1}' \underline{V}^{*-1} \underline{1})^{-1} = \frac{\sigma_e^{*2}}{(1-\varphi^*)(m-m\varphi^*+2\varphi^*)}. \quad (1)$$

Table 1 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of σ_a^{*2} , $\tilde{\sigma}_a^2$ and $\sigma_a^{*2}(\varphi \text{ known})$*

Each simulation study consisted of 500 experiments, with $N=100$ subjects, $m=3$ measurements per subject, $\mu=10.0$ and $\sigma_e^2=0.6$.

Study	Estimator	AV	SD	RMSE
1 $\sigma_a^2=1.0$ $\varphi=0.2$	σ_a^{*2}	0.988	0.234	0.234
	$\tilde{\sigma}_a^2$	1.080	0.180	0.197
	$\sigma_a^{*2}(\varphi \text{ known})$	1.008	0.181	0.181
2 $\sigma_a^2=1.0$ $\varphi=0.5$	σ_a^{*2}	0.751	1.498	1.519
	$\tilde{\sigma}_a^2$	1.295	0.211	0.363
	$\sigma_a^{*2}(\varphi \text{ known})$	1.007	0.213	0.213
3 $\sigma_a^2=1.0$ $\varphi=0.8$	σ_a^{*2}	3.995	51.445	51.532
	$\tilde{\sigma}_a^2$	2.183	0.327	1.228
	$\sigma_a^{*2}(\varphi \text{ known})$	1.010	0.355	0.355
4 $\sigma_a^2=0.1$ $\varphi=0.2$	σ_a^{*2}	0.083	0.148	0.148
	$\tilde{\sigma}_a^2$	0.174	0.059	0.094
	$\sigma_a^{*2}(\varphi \text{ known})$	0.102	0.060	0.060
5 $\sigma_a^2=2.0$ $\varphi=0.2$	σ_a^{*2}	1.972	0.387	0.388
	$\tilde{\sigma}_a^2$	2.081	0.320	0.330
	$\sigma_a^{*2}(\varphi \text{ known})$	2.012	0.318	0.319
6 $\sigma_a^2=10.0$ $\varphi=0.2$	σ_a^{*2}	9.953	1.425	1.426
	$\tilde{\sigma}_a^2$	10.056	1.412	1.413
	$\sigma_a^{*2}(\varphi \text{ known})$	9.986	1.411	1.411
7 $\sigma_a^2=0.1$ $\varphi=0.8$	σ_a^{*2}	2.867	81.145	81.199
	$\tilde{\sigma}_a^2$	1.296	0.211	1.215
	$\sigma_a^{*2}(\varphi \text{ known})$	0.114	0.248	0.249
8 $\sigma_a^2=2.0$ $\varphi=0.8$	σ_a^{*2}	9.072	73.007	73.349
	$\tilde{\sigma}_a^2$	3.204	0.477	1.295
	$\sigma_a^{*2}(\varphi \text{ known})$	2.017	0.493	0.493
9 $\sigma_a^2=10.0$ $\varphi=0.8$	σ_a^{*2}	15.612	72.447	72.664
	$\tilde{\sigma}_a^2$	11.170	1.566	1.955
	$\sigma_a^{*2}(\varphi \text{ known})$	9.994	1.576	1.576

It is clear from (1) that, as φ^* approaches 1, $(\underline{1}' \underline{V}^{*-1} \underline{1})^{-1}$ becomes large (positive if $\varphi^* < 1$, negative if $\varphi^* > 1$). Hence σ_a^{*2} is very unstable for φ near 1, and this is reflected in Table 1. If however, φ is known instead of estimated, the weighted least squares estimator σ_a^{*2} is indeed preferable to the unweighted $\tilde{\sigma}_a^2$, and this is also apparent from Table 1.

Appendix 2 Results relating to the modified variance-based estimation of 1.2

(i) Proof that, when $m = 3$, $-1 < \varphi_i^* < 3$, where φ_i^* is the WHS estimate of φ from the i^{th} subject (cf 1.2, p17)

Let X_j be the j^{th} measurement on subject i , $j = 1, \dots, m$.

When $m = 3$, $S_{3,i}^2$ and $S_{2,i}^2$, the averages of all sample variances which can be computed from 3 and 2 adjacent measurements respectively on the i^{th} subject, are given by

$$\begin{aligned} S_{3,i}^2 &= \frac{1}{2} \left[X_1^2 + X_2^2 + X_3^2 - \frac{(X_1 + X_2 + X_3)^2}{3} \right] \\ &= \frac{1}{3} (X_1^2 + X_2^2 + X_3^2 - X_1 X_2 - X_1 X_3 - X_2 X_3) \end{aligned} \quad (1)$$

and

$$\begin{aligned} S_{2,i}^2 &= \frac{1}{2} \left[\left\{ X_1^2 + X_2^2 - \frac{(X_1 + X_2)^2}{2} \right\} + \left\{ X_2^2 + X_3^2 - \frac{(X_2 + X_3)^2}{2} \right\} \right] \\ &= \frac{1}{4} (X_1^2 + 2X_2^2 + X_3^2 - 2X_1 X_2 - 2X_2 X_3) . \end{aligned} \quad (2)$$

Now

$$\begin{aligned} -1 < \varphi_i^* < 3 &\iff -1 < 3 \left[\frac{S_{3,i}^2}{S_{2,i}^2} - 1 \right] < 3 \\ &\iff \frac{2}{3} < \frac{S_{3,i}^2}{S_{2,i}^2} < 2 . \end{aligned}$$

Consider first

$$\frac{2}{3} < \frac{S_{3,i}^2}{S_{2,i}^2} . \quad (3)$$

This holds iff $2 S_{2,i}^2 < 3 S_{3,i}^2$

$$\begin{aligned} \iff \frac{1}{2} (X_1^2 + 2X_2^2 + X_3^2 - 2X_1 X_2 - 2X_2 X_3) &< X_1^2 + X_2^2 + X_3^2 - X_1 X_2 - X_1 X_3 - X_2 X_3 \\ &\text{(substituting (1) and (2))} \end{aligned}$$

$$\iff \frac{1}{2} (X_1 - X_3)^2 > 0 .$$

It follows that (3) always holds, with equality iff $X_1 = X_3$; if $X_1 = X_3$, the estimate φ_1^* of φ is -1 .

Now consider

$$\frac{S_{3,i}^2}{S_{2,i}^2} < 2. \quad (4)$$

This holds iff

$$\begin{aligned} S_{3,i}^2 &< 2S_{2,i}^2 \\ \iff \frac{1}{3} (X_1^2 + X_2^2 + X_3^2 - X_1X_2 - X_1X_3 - X_2X_3) &< \frac{1}{2} (X_1^2 + 2X_2^2 + X_3^2 - 2X_1X_2 - 2X_2X_3) \\ \iff (X_1 - 2X_2 + X_3)^2 &> 0. \end{aligned}$$

Thus (4) always holds, with equality iff $X_1 + X_3 = 2X_2$; if $X_1 + X_3 = 2X_2$, the estimate φ_1^* of φ is 3 .

(Note that when $X_1 = X_2 = X_3$, $S_{3,i}^2 = S_{2,i}^2 = 0$, and the ratio $S_{3,i}^2/S_{2,i}^2$ is undefined.)

Inequalities (3) and (4) together imply that $-1 < \varphi_1^* < 3$.

(ii) Derivation of an approximate formula (as a function of φ) for the variance of the WHS estimator φ^* , when $m = 3$, and comparison of results from this formula with empirical variances obtained from simulations (cf 1.2, p19)

In deriving an approximation to the variance of the WHS estimator φ^* (see 1.2, p16, (5)), analytical formulae were obtained for $\text{cov}(S_{3,i}^2, S_{2,i}^2)$, $\text{var}(S_{3,i}^2)$ and $\text{var}(S_{2,i}^2)$, as follows (with $S_{3,i}^2$ and $S_{2,i}^2$ as defined in (1) and (2)).

First suppose $\underline{X}$ is a general m -variate normal vector, $\underline{X} \sim N(\underline{0}_m, \underline{\Sigma})$. Then

$$\underline{C} (\underline{X}' \underline{A} \underline{X}, \underline{X}' \underline{B} \underline{X}) = 2 \text{ trace } (\underline{A} \underline{\Sigma} \underline{B} \underline{\Sigma}), \quad (5)$$

where $\underline{C}$ is the covariance matrix, and $\underline{A}$ and $\underline{B}$ are any $m \times m$ symmetric matrices (see Mardia, Kent and Bibby, 1979, p95).

Now for subject i , with $m = 3$ measurements X_1 , X_2 and X_3 say, the quantities $S_{3,i}^2$ and $S_{2,i}^2$ (see (1) and (2)) may be written

$$S_{3,i}^2 = \frac{1}{6} (X_1 \ X_2 \ X_3) \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \equiv \underline{X}' \underline{A} \underline{X}, \quad (6)$$

where $\underline{A} \equiv \frac{1}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} ;$

and $S_{2,i}^2 = \frac{1}{4} (X_1 \ X_2 \ X_3) \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \equiv \underline{X}' \underline{B} \underline{X} , \quad (7)$

where $\underline{B} \equiv \frac{1}{4} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} .$

If $\underline{X} = (X_1, X_2, X_3)'$ is a vector from the WHS model, then, ignoring the variance between subjects, $\underline{X} \sim N(\underline{\mu}1, \underline{\Sigma})$, where

$$\underline{\Sigma} = \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 & \varphi \\ \varphi^2 & \varphi & 1 \end{bmatrix} \quad (\text{cf 1.1, p3, (2)}).$$

Assume without loss of generality that $\mu = 0$. Then (5), (6) and (7) together imply that $\text{cov}(S_{3,i}^2, S_{2,i}^2) = 2 \text{ trace } (\underline{A} \underline{\Sigma} \underline{B} \underline{\Sigma})$.

Now $\underline{A} \underline{\Sigma} = \frac{1}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 & \varphi \\ \varphi^2 & \varphi & 1 \end{bmatrix}$

$$= \frac{1}{6} \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} -\varphi^2-\varphi+2 & \varphi-1 & 2\varphi^2-\varphi-1 \\ -\varphi^2+2\varphi-1 & -2\varphi+2 & -\varphi^2+2\varphi-1 \\ 2\varphi^2-\varphi-1 & \varphi-1 & -\varphi^2-\varphi+2 \end{bmatrix} . \quad (8)$$

Also $\underline{B} \underline{\Sigma} = \frac{1}{4} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} 1 & \varphi & \varphi^2 \\ \varphi & 1 & \varphi \\ \varphi^2 & \varphi & 1 \end{bmatrix}$

$$= \frac{1}{4} \frac{\sigma_e^2}{1-\varphi^2} \begin{bmatrix} -\varphi+1 & \varphi-1 & \varphi^2-\varphi \\ -\varphi^2+2\varphi-1 & -2\varphi+2 & -\varphi^2+2\varphi-1 \\ \varphi^2-\varphi & \varphi-1 & -\varphi+1 \end{bmatrix} . \quad (9)$$

From (8) and (9),

$$\underline{A} \underline{\Sigma} \underline{B} \underline{\Sigma} = \frac{1}{24} \left\{ \frac{\sigma_e^2}{1-\varphi^2} \right\}^2 \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

$$\begin{aligned} \text{where } c_{11} &= (-\varphi^2 - \varphi + 2)(-\varphi + 1) + (\varphi - 1)(-\varphi^2 + 2\varphi - 1) + (2\varphi^2 - \varphi - 1)(\varphi^2 - \varphi) \\ &= 2\varphi^4 - 3\varphi^3 + 3\varphi^2 - 5\varphi + 3, \end{aligned}$$

$$\begin{aligned} c_{22} &= 2(-\varphi^2 + 2\varphi - 1)(\varphi - 1) + (-2\varphi + 2)^2 \\ &= -2\varphi^3 + 10\varphi^2 - 14\varphi + 6 \end{aligned}$$

$$\begin{aligned} \text{and } c_{33} &= (2\varphi^2 - \varphi - 1)(\varphi^2 - \varphi) + (\varphi - 1)(-\varphi^2 + 2\varphi - 1) + (-\varphi^2 - \varphi + 2)(-\varphi + 1) \\ &= c_{11}. \end{aligned}$$

Now

$$\begin{aligned} \text{trace } (\underline{A} \underline{\Sigma} \underline{B} \underline{\Sigma}) &= \frac{1}{24} \left\{ \frac{\sigma_e^2}{1-\varphi^2} \right\}^2 (c_{11} + c_{22} + c_{33}) \\ &= \frac{1}{6} \left\{ \frac{\sigma_e^2}{1-\varphi^2} \right\}^2 (\varphi^4 - 2\varphi^3 + 4\varphi^2 - 6\varphi + 3) \\ &= \frac{1}{6} \left\{ \frac{\sigma_e^2}{1+\varphi} \right\}^2 (\varphi^2 + 3) \end{aligned}$$

$$\begin{aligned} \text{and } \text{cov}(S_{3,i}^2, S_{2,i}^2) &= 2 \text{ trace } (\underline{A} \underline{\Sigma} \underline{B} \underline{\Sigma}) \\ &= \frac{1}{3} (\sigma_e^2)^2 \frac{\varphi^2 + 3}{(1+\varphi)^2}. \end{aligned} \quad (10)$$

By similar algebraic manipulations, using equations (5), (6) and (7), formulae for $\text{var}(S_{3,i}^2)$ and $\text{var}(S_{2,i}^2)$ were obtained as

$$\text{var}(S_{3,i}^2) = 2 \text{ trace } (\underline{A} \underline{\Sigma} \underline{A} \underline{\Sigma}) = \frac{1}{9} (\sigma_e^2)^2 \frac{5\varphi^2 + 6\varphi + 9}{(1+\varphi)^2} \quad (11)$$

$$\text{and } \text{var}(S_{2,i}^2) = 2 \text{ trace } (\underline{B} \underline{\Sigma} \underline{B} \underline{\Sigma}) = \frac{1}{4} (\sigma_e^2)^2 \frac{\varphi^2 - 2\varphi + 5}{(1+\varphi)^2}. \quad (12)$$

The WHS estimate φ^* of φ is given by

$$\varphi^* = 3 \left\{ \frac{S_3^2}{S_2^2} - 1 \right\} \quad (\text{see 1.2, p16, (5)}).$$

Hence (see e.g. Mood, Graybill and Boes, 1974, p181),

$$\text{var}_{\varphi}(\varphi^*) \cong 9 \left\{ \frac{E(S_3^2)}{E(S_2^2)} \right\}^2 \left\{ \frac{\text{var}(S_3^2)}{[E(S_3^2)]^2} - \frac{2\text{cov}(S_3^2, S_2^2)}{E(S_3^2)E(S_2^2)} + \frac{\text{var}(S_2^2)}{[E(S_2^2)]^2} \right\}. \quad (13)$$

$$\text{Now} \quad S_3^2 = \frac{1}{N} \sum_{i=1}^N S_{3,i}^2 \quad \text{and} \quad S_2^2 = \frac{1}{N} \sum_{i=1}^N S_{2,i}^2,$$

so, from equations (10) to (12),

$$\text{var}(S_3^2) = \frac{1}{N} \frac{1}{9} (\sigma_e^2)^2 \left\{ \frac{5\varphi^2 + 6\varphi + 9}{(1+\varphi)^2} \right\},$$

$$\text{cov}(S_3^2, S_2^2) = \frac{1}{N} \frac{1}{3} (\sigma_e^2)^2 \left\{ \frac{\varphi^2 + 3}{(1+\varphi)^2} \right\}$$

$$\text{and} \quad \text{var}(S_2^2) = \frac{1}{N} \frac{1}{4} (\sigma_e^2)^2 \left\{ \frac{\varphi^2 - 2\varphi + 5}{(1+\varphi)^2} \right\}.$$

$$\text{Also,} \quad E(S_3^2) = \frac{\sigma_e^2}{1+\varphi} \left\{ 1 + \frac{\varphi}{3} \right\} \quad \text{and} \quad E(S_2^2) = \frac{\sigma_e^2}{1+\varphi} \quad (\text{see 1.2}$$

p16, (1) and (2)).

Substituting into (13) gives

$$\begin{aligned} \text{var}_{\varphi}(\varphi^*) &\cong 9 \left\{ 1 + \frac{\varphi}{3} \right\}^2 \frac{1}{N} \left\{ \frac{5\varphi^2 + 6\varphi + 9}{9(1+\varphi/3)^2} - \frac{2(\varphi^2 + 3)}{3(1+\varphi/3)} + \frac{\varphi^2 - 2\varphi + 5}{4} \right\} \\ &= \frac{1}{4N} (\varphi^4 - 4\varphi^3 - 2\varphi^2 + 12\varphi + 9) \equiv \tilde{v}_{\varphi}(\varphi^*) \quad \text{say.} \end{aligned}$$

This approximation to the variance of φ^* is compared in Table 1 with empirical variances $\hat{v}_{\varphi}(\varphi^*)$ obtained from simulations (cf 1.2, p19, Table 2).

Table 1 *Theoretical variance ($\tilde{v}_\varphi$) and empirical variance ($\hat{v}_\varphi$, obtained from simulations with 500 experiments per simulation, with $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_\epsilon^2=0.6$) of WHS's φ^**

φ	$\tilde{v}_\varphi(\varphi^*)$	$\hat{v}_\varphi(\varphi^*)$	φ	$\tilde{v}_\varphi(\varphi^*)$	$\hat{v}_\varphi(\varphi^*)$
0.99	.0400	.0408	0.0	.0225	.0222
0.9	.0398	.0400	-0.1	.0195	.0199
0.8	.0392	.0396	-0.2	.0164	.0166
0.7	.0382	.0396	-0.3	.0133	.0119
0.6	.0369	.0357	-0.4	.0104	.0112
0.5	.0352	.0357	-0.5	.0077	.0083
0.4	.0331	.0342	-0.6	.0052	.0056
0.3	.0308	.0289	-0.7	.0031	.0031
0.2	.0282	.0292	-0.8	.0014	.0014
0.1	.0254	.0259	-0.9	.0004	.0004

(iii) Modified variance-based estimator φ_{avad}^* (cf 1.2, p24)

In general, for random variables X and Y, an approximation to the quotient of expected values $E(X)/E(Y)$ is given by

$$\frac{E(X)}{E(Y)} \cong E\left\{ \frac{X}{Y} \right\} + \frac{1}{E(Y)^2} \left\{ cov(X,Y) - \frac{E(X)}{E(Y)} var(Y) \right\} \tag{17}$$

(see e.g. Mood, Graybill and Boes, 1974, p181).

Recall (see 1.2, p16, (3)) that φ is given by

$$\varphi = 3 \left\{ \frac{E(S_3^2)}{E(S_2^2)} - 1 \right\} . \tag{18}$$

With the estimator

$$\varphi_{av}^* = \frac{1}{N} \sum_{i=1}^N \varphi_i^* = \frac{1}{N} \sum_{i=1}^N 3 \left\{ \frac{S_{3,i}^2}{S_{2,i}^2} - 1 \right\} = 3 \left\{ \frac{1}{N} \sum_{i=1}^N \frac{S_{3,i}^2}{S_{2,i}^2} - 1 \right\} , \tag{19}$$

the ratio $E(S_3^2)/E(S_2^2)$ of (18) is estimated by $(1/N)\sum(S_{3,i}^2/S_{2,i}^2)$. However

$$\frac{1}{N} \sum_{i=1}^N \frac{S_{3,i}^2}{S_{2,i}^2} \text{ may better estimate } E \left\{ \frac{S_3^2}{S_2^2} \right\}; \text{ and substituting}$$

$X = S_{3,i}^2$ and $Y = S_{2,i}^2$ into (17) suggests

$$\frac{E(S_3^2)}{E(S_2^2)} = \frac{E(S_{3,i}^2)}{E(S_{2,i}^2)}$$

$$\cong E \left\{ \frac{S_{3,i}^2}{S_{2,i}^2} \right\} + \frac{1}{E(S_{2,i}^2)} \left\{ \text{cov}(S_{3,i}^2, S_{2,i}^2) - \frac{E(S_{3,i}^2)}{E(S_{2,i}^2)} \text{var}(S_{2,i}^2) \right\}. \quad (20)$$

None of the quantities on the right-hand side of (20) is known, but each can be estimated by its sample equivalent. So, φ_{avad}^* , given by

$$\varphi_{\text{avad}}^* = 3 \left[\frac{1}{N} \sum_{i=1}^N \left[\frac{S_{3,i}^2}{S_{2,i}^2} + \frac{1}{S_{2,i}^2} \left\{ \text{c}\tilde{\text{ov}}(S_{3,i}^2, S_{2,i}^2) - \frac{S_{3,i}^2}{S_{2,i}^2} \text{v}\tilde{\text{ar}}(S_{2,i}^2) \right\} \right] \right], \quad (21)$$

where $\text{c}\tilde{\text{ov}}$ and $\text{v}\tilde{\text{ar}}$ are sample estimates, may improve on φ_{av}^* of (19). However, the simulation results of Table 2 show that φ_{avad}^* of (21) did not yield any improvement in RMSE.

Table 2 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of φ^* and φ_{avad}^**

Each simulation study consisted of 500 experiments, each with $N=100$ subjects, $m=3$ measurements per subject, $\mu=10.0$, $\sigma_a^2=1.0$ and $\sigma_e^2=1.1$.

Study	Estimator	AV	SD	RMSE
1 $\varphi=0.0$	φ^*	0.014	0.151	0.151
	φ_{avad}^*	-0.256	0.275	0.376
2 $\varphi=0.5$	φ^*	0.510	0.189	0.189
	φ_{avad}^*	0.318	0.385	0.426
3 $\varphi=0.9$	φ^*	0.888	0.200	0.201
	φ_{avad}^*	0.845	0.394	0.398

(iv) Modified variance-based estimators φ_a^* , φ_b^* and φ_c^* (cf 1.2, p24)

Expressions for $\text{cov}(S_{3,i}^2, S_{2,i}^2)$, $\text{var}(S_{3,i}^2)$ and $\text{var}(S_{2,i}^2)$ were derived for the case $m = 3$ (see equations (10) to (12)). Three additional ways of estimating φ were derived using these expressions.

- (a) Rearranging equation (20), and substituting expected values from equations (1) and (2) of 1.2, and from equations (10) and (12) gives

$$E \left[\frac{S_3^2}{S_2^2} \right] \cong 1 + \frac{\varphi}{3} - \left\{ \frac{\varphi^2+3}{3} - \left[1 + \frac{\varphi}{3} \right] \left[\frac{\varphi^2-2\varphi+5}{4} \right] \right\} \quad (22)$$

The expected value $E(S_{3,i}^2/S_{2,i}^2)$ can be estimated by

$$x = \frac{1}{N} \sum_{i=1}^N \frac{S_{3,i}^2}{S_{2,i}^2} ; \quad \text{and after rearranging, (22) gives}$$

$$\varphi^3 - 3\varphi^2 + 3\varphi + (15-12x) = 0 . \quad (23)$$

For $x \neq 4/3$, this cubic equation has one real root and two complex roots.

(When $x = 4/3$, it has 3 real roots – see e.g. Abramowitz and Stegun, p17.)

The equation can be solved numerically and the real root is taken as the estimator φ_a^* .

- (b) Let $\tilde{v}_3$ and $\tilde{v}_2$ be sample estimates of $\text{var}(S_{3,i}^2)$ and $\text{var}(S_{2,i}^2)$ obtained from N subjects. Then from (11) and (12),

$$\frac{\tilde{v}_3}{\tilde{v}_2} \cong \frac{4(5\varphi^2+6\varphi+9)}{9(\varphi^2-2\varphi+5)} .$$

$$\text{Hence } (9\tilde{v}_3 - 20\tilde{v}_2)\varphi^2 + (-18\tilde{v}_3 - 24\tilde{v}_2)\varphi + 45\tilde{v}_3 - 36\tilde{v}_2 = 0 . \quad (24)$$

Solving this quadratic for φ and choosing the root nearer φ^* , gives an estimate φ_b^* say, of φ .

- (c) Let $\tilde{c}_{23}$ be the sample estimate of $\text{cov}(S_{3,i}^2, S_{2,i}^2)$ obtained from N subjects. Then, from (10) and (11),

$$\frac{\tilde{v}_3}{\tilde{c}_{23}} \cong \frac{3(5\varphi^2+6\varphi+9)}{9(\varphi^2+3)} .$$

Hence $(3\tilde{v}_3-5\tilde{c}_{23})\varphi^2 - 6\tilde{c}_{23}\varphi + 9(\tilde{v}_3-\tilde{c}_{23}) = 0 .$ (25)

Solving this quadratic for φ and choosing the root nearer φ^* gives another estimate φ_c^* say, of φ .

Solutions to the cubic equation (23) and the quadratics (24) and (25) were obtained using the NAG subroutine C02AEF. (This finds all the roots of a real polynomial equation using the method of Grant and Hitchins.) Simulations were carried out to investigate the performance of φ_a^* , φ_b^* and φ_c^* . Table 3 summarizes the results, illustrating the failure of the 3 proposed estimators to improve on φ^* .

Table 3 *Simulation results: average (AV), standard deviation (SD) and root mean squared error (RMSE) of φ^* , φ_a^* , φ_b^* and φ_c^**

Each simulation consisted of 500 experiments, each conducted with $N=100$ subjects, $m=3$ measurements per subject, $\mu=0.0$, $\sigma_a^2=0.0$ and $\sigma_e^2=1.1$.

Study	Estimator	AV	SD	RMSE
1 $\varphi=0.0$	φ^*	0.016	0.148	0.149
	φ_a^*	-0.276	0.111	0.298
	φ_b^*	0.017	0.192	0.193
	φ_c^*	0.020	0.191	0.192
2 $\varphi=0.5$	φ^*	0.511	0.188	0.188
	φ_a^*	0.079	0.342	0.543
	φ_b^*	0.521	0.250	0.251
	φ_c^*	0.530	0.266	0.268
3 $\varphi=0.9$	φ^*	0.888	0.200	0.201
	φ_a^*	0.693	0.697	0.727
	φ_b^*	0.893	0.251	0.252
	φ_c^*	0.899	0.274	0.274

Appendix 3 The robust estimation by functional least squares of the autoregressive parameter in N short time series (cf 1.2, p24)

Heathcote and Welsh (1983) describe the estimation of a general autoregressive process by functional least squares. They define a family of estimators indexed by a real parameter, and the *functional least squares* (fls) estimator is chosen to have minimum variance with respect to this parameter. The chosen parameter value in some sense detects and adjusts for long-tailed error distributions. When data are normally distributed, this fls estimation is equivalent to ordinary least squares estimation. However, when data are generated from a first-order autoregressive model with a variety of (non-normal) leptokurtic error distributions, fls estimation is shown, by simulation, to be more robust than ordinary least squares (and generally more robust than M- and generalized M-) estimation.

What follows is a straightforward adaptation of fls estimation of a single autoregressive time series to estimation of the autoregressive constant φ in N short first order autoregressive time series. It is anticipated that this adaptation may provide a robust estimator of the N -series φ , with properties analogous to those of the one-series estimator; though this idea has yet to be substantiated by empirical or rigorous theoretical evidence.

Suppose $\underline{X}_j = (X_{j1}, X_{j2}, \dots, X_{jm})'$ is the measurement vector for the j^{th} subject from a sample of N subjects (cf 1.1, p2; the subscript j is used here instead of i , to avoid confusion with the imaginary number i , in what follows). Assume that measurements $\underline{X}_j$ satisfy

$$X_{jt} = \begin{cases} \mu + a_j + \epsilon_{j1} & , \quad t = 1 \\ \mu + \varphi X_{j,t-1} + a_j + \epsilon_{jt} & , \quad t = 2, \dots, m \end{cases} \quad j = 1, \dots, N . \quad (1)$$

In contrast to 1.1 ff (where normality is assumed), no assumption is made here about the distributions of the overall mean μ , the random subject effects $\{a_j\}$ and the within-subject errors $\{\epsilon_{jt}\}$. Consider the term $a_j + \epsilon_{jt}$ in (1) as "error", and

define the loss function

$$\begin{aligned}
 L_{N,t}(\varphi; s) &= -\frac{1}{s^2} \log \left| \frac{1}{N} \sum_{j=1}^N \exp \left\{ i s (X_{jt} - \mu - \varphi X_{j,t-1}) \right\} \right|^2 \\
 &= -\frac{1}{s^2} \log \left| \frac{1}{N} \sum_{j=1}^N \exp \left\{ i s (X_{jt} - \varphi X_{j,t-1}) \right\} \right|^2. \quad (2)
 \end{aligned}$$

Assume the free variable s is restricted to take values in a closed interval $\Upsilon = [0, s_1]$. Minimization of $L_{N,t}(\varphi; s)$ with respect to φ , for fixed s , leads to a family of estimators $\{\varphi_{N,t}^*(s), s \in \Upsilon\}$. The recommended estimator is that $\varphi_{N,t}^*(s)$ with estimated minimum asymptotic variance.

By analogy with Heathcote and Welsh, the *variance function* is defined by

$$\sigma^2(s) = \frac{u^2(s)\{1-u(2s)\} - 2u(s)v(s)v(2s) + v^2(s)\{1+u(2s)\}}{2s^2\{u^2(s)+v^2(s)\}^2} \quad (3)$$

$$\equiv \sigma^2\{u(s), v(s)\}, \quad (4)$$

where

$$u(s) = E[\cos\{s(X_{jt} - \varphi X_{j,t-1})\}] \quad (5)$$

and

$$v(s) = E[\sin\{s(X_{jt} - \varphi X_{j,t-1})\}]. \quad (6)$$

The estimated variance function $\sigma_{N,t}^2(s)$ is obtained by replacing $u(s)$ and $v(s)$ in equations (5) and (6) by their estimators

$$U_{N,t}\{\varphi_{N,t}^*(s), s\} = \frac{1}{N} \sum_{j=1}^N \cos \{s[X_{jt} - \varphi_{N,t}^*(s) X_{j,t-1}]\} \quad (7)$$

and

$$V_{N,t}\{\varphi_{N,t}^*(s), s\} = \frac{1}{N} \sum_{j=1}^N \sin \{s[X_{jt} - \varphi_{N,t}^*(s) X_{j,t-1}]\}. \quad (8)$$

The functions $u(2s)$ and $v(2s)$ in (3) are estimated by $U_{N,t}\{\varphi_{N,t}^*(s), 2s\}$ and $V_{N,t}\{\varphi_{N,t}^*(s), 2s\}$ respectively. The resulting estimator of $\sigma^2(s)$ is

$$\sigma_{N,t}^2(s) = \sigma^2(U_{N,t}\{\varphi_{N,t}^*(s), s\}, V_{N,t}\{\varphi_{N,t}^*(s), s\}). \quad (9)$$

This can be minimized with respect to s , to yield the value $s_{N,t}$.

The fls estimator $\varphi_{N,t}^*(s_{N,t}^*)$ is obtained by iterating in two stages:

(i) minimizing $\sigma_{N,t}^2(s)$ of (9) with respect to s

and (ii) minimizing $L_{N,t}(\varphi; s)$ of (2) with respect to φ ,

until convergence is achieved.

A distinct estimator $\varphi_{N,t}^*(s_{N,t}^*)$ exists for each fixed $t = 2, \dots, m$. The overall estimate of φ may be obtained from some simple function of $\{\varphi_{N,t}^*(s_{N,t}^*), t=2, \dots, m\}$, such as the mean.

Appendix 4 The simulation program of 2.4 (cf p73)

The Fortran program described below was used to produce the simulation results presented in 2.4. It generates data from a user-specified $(\sigma_a^2, \varphi, \sigma_e^2, K)$ model and "tests the data for significance" using the sequential procedures (a), (b) and (c) of 2.4.

The main program, TESTABC, is just 3 lines:

```
CALL SIMTEST
STOP
END .
```

The subroutine SIMTEST is compiled and loaded in an object code library, together with all subsequent subroutines.

In SIMTEST, the user is prompted for

NSIM	the number of simulations required
N	the number of subjects per simulation (data are envisaged here as coming from N subjects, rather than from N subject pairs)
M	the number of repeated measurements per subject (constrained to be 3, 5 or 10)
XMEAN	the vector of means, $(\mu_1, \mu_2, \dots, \mu_M)'$
SMUSQ	the between-subject variance, σ_a^2
BE	the autoregression parameter, φ
SWSQ	the within-subject variance, σ_e^2 and
K	the parameter which determines whether within-subject measurements are stationary; or have increasing or decreasing variances.

The user is asked to choose between repeatable and non-repeatable series of simulations; if a non-repeatable series is requested, a call is made to the NAG subroutine G05CCF. Also, the user is given the option of monitoring, for each simulated dataset, the minimization processes used in determining maximum likelihood estimates (mles).

SIMTEST then calls the subroutines INPUT and MINIMI, in sequence, NSIM times.

A call is made from INPUT to another subroutine SIMVAR, which has arguments depending on the value of M . In SIMVAR, the variance matrix of the M -dimensional normal vector $\underline{X}$, corresponding to the user-supplied model, is evaluated (on the first simulation only) and calls are made to subroutines SIMULATE and VARBASES. SIMULATE generates N vectors of length M , from the given mean vector and covariance matrix, using the NAG subroutines G05EAF and G05EZF for generating multivariate normal data. VARBASES computes estimates (unweighted least squares and variance-based, see 1.1 and 1.2) of the mean vector $\underline{\mu}$ and the parameters σ_a^2 , φ and σ_e^2 . These estimates are used as starting values for the minimization processes of the subroutine MINIMI. Sums and variances of measurements from the N subjects, based on the first $1, 2, \dots, M$ measurements successively, are evaluated in INPUT and retained for use in TRANTEST.

After setting some parameters, MINIMI calls the subroutine IACTMM (originally short for "Interactive Minimizer", cf Acknowledgements), where the actual minimizations are performed using the NAG routine E04KBF. The negative log likelihood and its derivatives are specified in the functions FN and FUNCT, and these are used by E04KBF in the actual minimizations. For the series of simulations described in 2.4, $2M-3$ minimizations of the negative log likelihood are performed for each dataset: $M-1$ with respect to 2 parameters σ_a^2 and σ_e^2 (involving $2, \dots, M$ measurements successively), assuming $\varphi = 0$; and $M-2$ with respect to 4 parameters σ_a^2 , φ , σ_e^2 and K (involving $3, \dots, M$ measurements successively). In these minimizations, σ_a^2 is constrained to be non-negative. If the user chose, in SIMTEST, to monitor the minimization processes, then the subroutines MONIT and MONWRI are used to output information on each minimization. The resulting mles of the parameters are recorded for each of the NSIM datasets.

Control returns to SIMTEST, where the subroutine TRANTEST is immediately called. In TRANTEST, the user is prompted to input the values $\alpha_1^*, \alpha_2^*, \dots, \alpha_M^*$, the absorption probabilities such that tests up to and including the j^{th} test "use up" significance $\alpha_1^* + \alpha_2^* + \dots + \alpha_j^*$; and such that $\alpha_1^* + \alpha_2^* + \dots + \alpha_M^*$ is equal to the required overall significance level α . The user is also asked to input $z_1^*, z_2^*, \dots, z_M^*$, the z

values corresponding to the α_j^* 's when the measurements are iid. (These are computed in a separate Fortran program ZF10.)

In TRANTEST, each of the NSIM datasets is tested for significance by the transformation procedure (procedure (b)), using the first-measurement variances, and the sums, evaluated in INPUT. Tests are performed using $P = 1, 2, \dots, M$ repeated measurements. If, for any P , a test is significant, no further tests are performed for the given dataset. For $P = 1$, measurements are assumed to be iid, and a test is performed in the usual way at level α_1^* . In this case, the "transformation matrix" is a constant, equal to $1/\hat{\text{var}}(X_1\text{'s})$. For $P = 2$, the mles of σ_a^2 and σ_e^2 are used to generate an estimated (2×2) covariance matrix $\underline{C}$ for the measurements. The transformation matrix

$$\underline{T} = \begin{bmatrix} t_{11} & 0 \\ t_{21} & t_{22} \end{bmatrix}$$

such that $\underline{T}' \underline{C} \underline{T} = \underline{I}_2$ is evaluated by solving two simultaneous non-linear equations, using the NAG routine C05NBF (cf 2.2, p55). The data are then "transformed" using $\underline{T}$. (At this stage, only the sums of the first, and the first and second, measurements over the N subjects, are needed.) The transformed data are tested using z_2^* , the user-supplied value appropriate for independent data. If there is a significant result, the routine records this and proceeds to the next dataset. Otherwise, the case $P = 3$ measurements per subject is considered. In this case, mles of the 4 parameters σ_a^2 , φ , σ_e^2 and K are used to generate the estimated (3×3) covariance matrix $\underline{C}$. The transformation test proceeds as before. If, after testing with $P = M$, the null hypothesis H_0 of "no treatment effect" is still not rejected, the routine proceeds to the next dataset.

Before leaving TRANTEST, the program tests H_0 for each of the NSIM datasets using procedure (c) of 2.4, assuming the measurements are iid. These tests use the sums and variances previously evaluated in INPUT, and the z_j^* values supplied by the user.

Results from TRANTEST are printed out.

Control returns to SIMTEST, which calls INTETEST, a subroutine which tests for significance using the integration procedure (procedure (a)).

For a given dataset, tests are performed after $P=1,2,\dots,M$ measurements per subject, or until a significant result is recorded. For $P = 1$, the procedure is the same as in TRANTEST: assume the measurements are iid and perform the test at level α_1^* . For $P = 2$, the appropriate z -value is determined by the function FINDZ. In FINDZ, the correlation $c_{1,2}$ between $S_1 = X_1$ and $S_2 = X_1 + X_2$ is estimated, by assuming the model with $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$, mles of σ_a^2 and σ_e^2 obtained from 2 measurements per subject. The NAG routine C05AZF is used to find the solution z_2 of the equation

$$\int_{-z_2}^{z_2} \int_{-z_1^*}^{z_1^*} f_{\underline{S}_2^*}(\underline{s}_2^*) d\underline{s}_2^* = 1 - (\alpha_1^* + \alpha_2^*), \quad (1)$$

where $\underline{S}_2^* = (S_1^*, S_2^*)'$ (the standardized version of $\underline{S}_2 = (S_1, S_2)'$) is distributed as

$$N \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & c_{1,2} \\ c_{1,2} & 1 \end{bmatrix} \right]$$

Initial bounds used by C05AZF, within which z_2 is assumed to lie, are 1.8 and 3.2. Intermediate values of the integral on the left hand side of (1) are obtained using the Schervish integration routine MULNOR. The absolute value of the sum S_2 (over the N subjects) of the first two measurements, divided by the theoretical standard deviation of S_2 (from the $(\hat{\sigma}_a^2, \hat{\sigma}_e^2)$ model) is compared with z_2 . If this value is larger than z_2 , H_0 is rejected and the program proceeds to consider the next dataset; otherwise the case $P = 3$ is considered. For $P = 3$ the theoretical correlations $c_{1,2}$, $c_{1,3}$ and $c_{2,3}$, between (S_1, S_2) , (S_1, S_3) and (S_2, S_3) respectively, are evaluated using $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ obtained from 3 measurements per subject. The solution z_3 to the equation

$$\int_{-z_3}^{z_3} \int_{-z_2}^{z_2} \int_{-z_1^*}^{z_1^*} f_{\underline{S}_3^*}(\underline{s}_3^*) d\underline{s}_3^* = 1 - (\alpha_1^* + \alpha_2^* + \alpha_3^*)$$

is then determined, using C05AZF and MULNOR as before. If the test involving S_3

and z_3 is non-significant, and if $M > 3$, INTETEST goes on to consider $P = 4$.

For $P = 4$ a different procedure comes into operation, and FINDZ calls the subroutine SETPM1. This evaluates

$$\int_{-z_3/\hat{v}_{43}}^{z_3/\hat{v}_{43}} \int_{-z_2/\hat{v}_{42}}^{z_2/\hat{v}_{42}} \int_{-z_1^*/\hat{v}_{41}}^{z_1^*/\hat{v}_{41}} f_{\underline{S}_4}(\underline{s}_4) d\underline{s}_4 \quad (2)$$

using the recursive method of integration of 2.3, and using the mles $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ based on 4 measurements per subject to determine $\{\hat{v}_{4j}\}$, the estimated variances of $\{S_j\}$. (The recursive integration uses the NAG routine D01GAF (cf 2.3, p65), and in this case takes advantage of the symmetry about 0 of the integral limits.) FINDZ reverts to C05AZF to compute the solution z_4 of the equation

$$\int_{-z_4/\hat{v}_{44}}^{z_4/\hat{v}_{44}} \cdots \int_{-z_1^*/\hat{v}_{41}}^{z_1^*/\hat{v}_{41}} f_{\underline{S}_4}(\underline{s}_4) d\underline{s}_4 = 1 - (\alpha_1^* + \alpha_2^* + \alpha_3^* + \alpha_4^*); \quad (3)$$

but now, instead of calling MULNOR for integral evaluations of the left hand side of (3), FINDZ calls the subroutine PRP. (Recall that MULNOR requires excessive (for these simulations) CPU time, for integrals of 4 dimensions or more.) In PRP, the recursive method uses numbers generated in SETPM1 to carry out the final (fourth) level of integration; PRP is called several times, until C05AZF converges to z_4 . The value z_4 is used for the fourth test of significance. This procedure, involving FINDZ, SETPM1 and PRP, is repeated up to $P = M$, or until H_0 is rejected. After all NSIM datasets have been tested in INTETEST, control returns to SIMTEST.

Finally, the user is given the option of computing and printing out some statistics on the mles used in the test procedures.

References

- Abramowitz, M and Stegun, I A (1964). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover Publications, New York.
- Armitage, P (1960, 1975). *Sequential Medical Trials*, first and second editions. Blackwell, Oxford.
- Armitage, P, McPherson, C K and Rowe, B C (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* 132, 235–244.
- Armitage, P, Stratton, I M and Worthington, H V (1985). Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up. *Biometrics* 41, 353–359.
- Berry, D A (1985). Interim analyses in clinical trials: classical vs. Bayesian approaches. *Statistics in Medicine* 4, 521–526.
- Dupont, W D (1983a). Sequential stopping rules and sequentially adjusted P values: does one require the other? *Controlled Clinical Trials* 4, 3–10.
- Dupont, W D (1983b). Rejoinder: statistical inference from trials with sequential stopping rules. *Controlled Clinical Trials* 4, 27–33.
- Greenhouse, S W (1983). Discussion on early stopping. *Controlled Clinical Trials* 4, 19–21.
- Heathcote, C R and Welsh, A H (1983). The robust estimation of autoregressive processes by functional least squares. *Journal of Applied Probability* 20, 737–753.
- Hodge, H C, Holloway, P J, Davies, T G H and Worthington, H V (1980). Caries prevention by dentifrices containing a combination of sodium monofluorophosphate and sodium fluoride. *British Dental Journal* 149, 201–204.
- Hua, L K and Yuan, W (1981). *Applications of Number Theory to Numerical Analysis*, revised edition. Springer-Verlag, Berlin.
- Kendall, M G and Stuart, A (1973). *The Advanced Theory of Statistics*, Volume 2, third edition. Griffin, London.
- Kenward, M G (1981). An investigation of certain methods for the analysis of repeated measurements. Ph D thesis, University of Reading.
- Mansour, H, Nordheim, E V and Rutledge, J J (1985). Maximum likelihood estimation of variance components in repeated measures designs assuming

- autoregressive errors. *Biometrics* 41, 287–294.
- Mardia, K V, Kent, J T and Bibby, J M (1979). *Multivariate Analysis*. Academic Press, London.
- McPherson, C K (1974). Statistics: the problem of examining accumulating data more than once. *New England Journal of Medicine* 290, 501–502.
- McPherson, C K (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine* 1, 25–36.
- McPherson, C K and Armitage, P (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society, Series A* 134, 15–25.
- Mood, A M, Graybill, F A and Boes, D C (1974). *Introduction to the theory of statistics*, third edition. McGraw–Hill, Singapore.
- O'Casey, S (1952). *Rose and Crown*. Macmillan, London.
- Parmar, M K B (1986). Multiplicity in clinical trials and other medical studies. D Phil thesis, University of Oxford.
- Pocock, S J (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191–199.
- Rao, C R (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* 31, 545–554.
- Schervish, M J (1984). Algorithm AS 195. Multivariate normal probabilities with error bound. *Applied Statistics* 33, 81–94.
- Smith, A F M (1984). Present position and potential developments: some personal views. Bayesian statistics. *Journal of the Royal Statistical Society, Series A* 147, 245–259.
- Solomon, P J (1985). Transformations for components of variance and covariance. *Biometrika* 72, 233–239.
- Spiegelhalter, D J and Freedman, L S (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine* 5, 1–13.
- Wilson, P D, Hebel, J R and Sherwin, R (1981). Screening and diagnosis when within-individual observations are Markov-dependent. *Biometrics* 37, 553–565.

