REPEATED MEASUREMENT: SENSITIVE TESTS FOR EXPERIMENTS WITH FEW ANIMALS[1]

J. L. Gill[2]

Michigan State University[3]
East Lansing, 48824

ABSTRACT

Animal scientists who conduct experiments involving repeated measurements of animals often are frustrated by low statistical power of tests for comparisons of treatment means. In many cases, low power of the traditional tests simply is the consequence of low replication (few animals per treatment) that was forced by cost or complexity of experimental technique. A method is given for comparing treatments in a way that permits sensitive tests (via partition of the treatment × period interaction), often when the number of animals per treatment is not more than five or six. Modifications for the procedure are given for the case of heterogeneous variances and covariances. An example from mammary physiology is used to illustrate the procedure and to compare it with standard methods of analysis.
(Key Words: Statistical Methods, Experimental Designs, Treatment, Variance-Covariance Matrix.)

Introduction

Few animals may be available (or few used, because of complex technique) in experiments with non-random repeated measurement (e.g., n animals in each of t treatment groups, each measured in p periods). In such cases, the use of (i) summary statistics for each animal to eliminate the time factor (e.g., area under a response curve), or (ii) ordinary univariate split-plot tests of the treatment means (Gill and Hafs, 1971) or (iii) multivariate analysis (Cole and Grizzle, 1966) is inadvisable, because comparisons of treatments are not sufficiently sensitive for any of those procedures. The problem with (ii) is that main effects of treatments (i.e., unconditiional or "marginal" means, averaged across periods) must be tested by the mean square for animals within treatments, which is inflated by positive correlations among repeated observations. Even conditional tests (i.e., comparisons of treatments within periods), as well as tests of means of summary statistics (i), cannot be very sensitive because, with low replication, the standard errors of mean differences (SED) are not much smaller than the ordinary (error) standard deviation among animals treated alike, without the influence of correlations induced by repeated measurement. For example, if n = 4 animals/group, then the SED is 70% as large as the ordinary error standard deviation. In (iii), severe restriction of numbers of animals leaves few (or no) degrees of freedom (df) for error, either reducing statistical power drastically or preventing multivariate analysis entirely.

In this paper, a method is given for partitioning the treatment × period interaction (of the univariate split-plot analysis) to permit sensitive comparisons of treatments. The procedure does not test for unconditional differences of treatment means (nor for the ordinary conditional differences, i.e., within periods), but for differences in treatment trends in response between specified periods. For example, one might ask, "Does response to treatment 1 change more than response to treatment 2 during the time between measurements at periods 3 and 4?. Logically, the answers to such questions usually are equally informative about the effects of treatments, as are the results of traditional comparisons of means, when both have sufficient statistical support, and the comparisons of trends have a better chance to achieve statistical significance when treatment

[1] Journal article number 11752 of the Michigan Agr. Exp. Sta.

[2] Dept. of Anim. Sci.

[3] The author is grateful for support, during the preparation of this paper, from the Dept. of Anim. Sci. and the Research Commission of the Agricultural Univ. of Wageningen, Netherlands. Helpful comments were provided by W. J. Koops and H.A.M. van der Steen. Also, discussions with D.F. Cox were stimulating and useful.

Received September 16, 1985.
Accepted March 24, 1986.

differences are not trivial. The procedure achieves its power via positive correlations among repeated observations, in much the same way that intelligent pairing or grouping of animals reduces error in randomized block experiments (Gill, 1978a,b). Modifications for the procedure are given for the case of a variance-covariance matrix that is not homogeneous. An example from mammary physiology is given to illustrate the procedure and compare it with alternative methods ordinarily used.

Procedure

The most common type of experiment with repeated measurement of animals has simple split-plot structure (Gill and Hafs, 1971). Therefore, that model is used as a vehicle for describing the method. Let the linear model for response variable Y (e.g., a serum hormone concentration) be

$$Y_{ijk} = M + T_i + A_{(i)j} + P_k + (TP)_{ik} + E_{ijk}, \quad (i = 1 \cdots t,\ j = 1 \cdots n \text{ per } i,\ k = 1 \cdots p), \quad (1)$$

where the terms are overall mean (M), treatment effect (T), effect of animal within treatment (A), period effect (P), treatment × period interaction (TP), and residual (E). It is assumed that $\Sigma_i T = 0$, $\Sigma_k P = 0$, $\Sigma_i (TP) = 0$, and $\Sigma_k (TP) = 0$. The animal term (A) represents error associated with treatment differences, and variability among animals ordinarily is substantial. A common problem in trying to overcome variation among animals is that, because of cost or complexity of experimental technique, replication (n) is rather small (i.e., few animals are in each treatment group).

Structure of the split-plot analysis of variance is shown in table 1 [see Gill and Hafs (1971) or Gill (1978b) for calculation of sums of squares and mean squares]. Expectations of mean squares are given for the case where the correlation (ρ) between any two periods is assumed to be constant. Modifications are given, in succeeding sections, for cases where the assumption clearly is not valid. The symbol σ^2 stands for intra-treatment animal variation when $\rho = 0$ (i.e., as for error variation without repeated measurement). Estimates of ρ and σ^2, i.e., $\hat{\rho}$ and $\hat{\sigma}^2$, can be obtained by setting the two error mean squares, $MS_{E(1)}$ and $MS_{E(2)}$, equal to their respective expectations and solving for the estimates of the two parameters. That leads to the results

$$\hat{\sigma}^2 = [MS_{E(1)} + (p-1)MS_{E(2)}]/p \text{ and} \quad (2)$$

$$\hat{\rho} = [MS_{E(1)} - MS_{E(2)}]/p\hat{\sigma}^2. \quad (3)$$

If one naively ignored the split-plot structure, i.e., ignored the term $A_{(i)j}$ in equation (1), then the ordinary two-way analysis of variance would have residual mean square equal to $\hat{\sigma}^2$. That is equivalent to pooling sums of squares and df for $A_{(i)j}$ and E_{ijk} in equation (1). Alternatively, if one calculated p separate one-way analyses of variance, one for each period (which would make tests of treatments conditional on time), and averaged the resulting mean squares for error, that would equal $\hat{\sigma}^2$, also. So, however one might obtain it, $\hat{\sigma}^2$ estimates variation of animals treated alike,

TABLE 1. STRUCTURE OF SPLIT-PLOT ANALYSIS OF VARIANCE FOR THE MOST COMMON TYPE OF EXPERIMENT WITH REPEATED MEASUREMENT[a]

Source of variation (model)	df	MS	Expectations of mean squares[b]
T = treatments (T_i)	$(t-1)$	MS_T	$\sigma^2[1 + (p-1)\rho] + np\Sigma T_i^2/(t-1)$
E(1) = animals/T ($A_{(i)j}$)	$t(n-1)$	$MS_{E(1)}$	$\sigma^2[1 + (p-1)\rho]$
P = periods (P_k)	$(p-1)$	MS_P	$\sigma^2(1-\rho) + tn\Sigma P_k^2/(p-1)$
TP = interaction $(TP)_{ik}$	$(t-1)(p-1)$	MS_{TP}	$\sigma^2(1-\rho) + n\Sigma\Sigma(TP)_{ik}^2/(t-1)(p-1)$
E(2) = residual error (E_{ijk})	$t(n-1)(p-1)$	$MS_{E(2)}$	$\sigma^2(1-\rho)$

[a] See Gill (1978b) for details.

[b] ρ = Correlation between any two periods, within treatments (assumed constant over time); σ^2 = variation among animals, within treatments, when $\rho = 0$ (or without repeated measurement); $(AP)_{(i)jk}$, animal × period interaction, is ignored; if it exists, then σ^2_{AP} is included in expectations of mean squares for P, TP and E(2), but tests of P and TP still are valid.

unaffected by the inter-period correlations induced by repeated measurement. In no case is $\hat{\sigma}^2$ an appropriate measure of error for comparing the unconditional effects of treatments or the effects of periods, nor for testing interaction of treatments and periods. It is appropriate, however, in conditional testing of treatments (i.e., within periods) and for calculating standard errors of period means (but not for standard errors of differences between period means).

Relative Efficiency of Repeated Measurement. As the expectations of mean squares show (table 1), the consequences of inducing positive inter-period correlation by repeated measurement are reduction of $MS_{E(2)}$ and an increase of $MS_{E(1)}$, relative to $\hat{\sigma}^2$. Therefore, statistical procedures that can make legitimate use of $MS_{E(2)}$ as error are likely to be substantially more efficient than those that must use $MS_{E(1)}$ or $\hat{\sigma}^2$.

First, note that the ratio of expectations of the two types of error mean squares is

$$E[MS_{E(1)}]/E[MS_{E(2)}] = \sigma^2[1 + (p-1)\rho]/\sigma^2(1-\rho) = 1 + p[\rho/(1-\rho)], \quad (4)$$

which is a function only of the inter-period correlation (ρ) and the number of periods (p). Ratios of expectations for several combinations of values of ρ and p are shown in table 2. Residual error, $MS_{E(2)}$, is expected to be not more than one-third of the magnitude of animal error, $MS_{E(1)}$, if the correlation is positive and moderately strong ($\rho > .5$), or if the number of periods is substantial when correlation is not strong (say, $p > 10$, unless ρ is near 0).

Second, note that the expected value of $MS_{E(2)}$ is $\sigma^2(1-\rho)$. Therefore, if correlation is positive and moderately strong ($\rho > .5$), then residual error is expected to be not more than one-half of the magnitude of σ^2, the ordinary error among animals in experiments without repeated measurement.

So, it is clear that repeated measurement ordinarily should be efficient, in the sense of reducing residual error (in cases with negative correlation, of course, that will not occur). If that efficiency is to be utilized in comparisons of treatments, however, it must be done by partitioning the interaction of treatments and periods, because the expectations of mean squares indicate that $MS_{E(2)}$ is relevant as error only for assessing differences among periods and for testing interaction of treatments and periods. A significant interaction can be evaluated by making conditional comparisons of treatment means (within periods), or conditional comparisons of period means (within treatments), or both.

Conditional Comparisons of Treatments. In experiments with repeated measurement of few animals, typical results of univariate split-plot analysis are significant differences among periods and significant interaction of treatments and periods, but weak evidence for differences among treatments. If interaction exists, then some conditional differences of treatments must exist, i.e., at least one treatment must differ from the others, at least in one period, because of different trends, over time, in responses to different treatments. However, the usual comparisons of treatments within periods may not provide statistical significance at the level reflected in the test of interaction, because interaction is tested with $MS_{E(2)}$, whereas conditional differences of treatments must be tested with $\hat{\sigma}^2$.

Consider the hypothesis, H: $\Delta_k = 0$, where Δ_k is a specified contrast among treatment means within period k. Two or more con-

TABLE 2. RATIOS OF EXPECTED VALUES OF $MS_{E(1)}$ AND $MS_{E(2)}$ FOR COMBINATIONS OF INTER-PERIOD CORRELATION (ρ) AND NUMBER OF PERIODS (p)

ρ	p = 2	3	4	5	10	20	30	50
.1	1.22	1.33	1.44	1.56	2.11	3.22	4.33	6.56
.2	1.50	1.75	2.00	2.25	3.50	6.00	8.50	13.50
.3	1.85	2.29	2.71	3.14	5.29	9.57	13.86	22.43
.5	3.00	4.00	5.00	6.00	11.00	21.00	31.00	51.00
.7	5.66	8.00	10.33	12.67	24.33	47.67	71.00	117.67
.9	19.00	28.00	37.00	46.00	91.00	181.00	271.00	451.00

trasts may be planned and tested for a given period. Let the sample contrast be

$$\bar{q}_k = \sum_{i=1}^{t} c_i \bar{y}_{i \cdot k}, \text{ with } \Sigma c_i = 0, \quad (5)$$

where the $\bar{y}_{i \cdot k}$ are treatment means in period k. The simplest contrast involves only two means, with corresponding c_i of 1 and −1, and all other $c_i = 0$. The variance of $\bar{q}_k$ does not involve $MS_{E(1)}$ alone (as often supposed), but $\hat{\theta}^2$, which is a function of both $MS_{E(1)}$ and $MS_{E(2)}$, shown in equation (2). The variance of the contrast is

$$V(\bar{q}_k) = (\sum_{i=1}^{t} c_i^2)(\hat{\theta}^2/n), \quad (6)$$

which is just $(2\hat{\theta}^2/n)$ for the simple comparison of two treatments in peri... k.

The test statistic for hypothesis H is

$$t = \bar{q}_k / \sqrt{V(\bar{q}_k)}, \quad (7)$$

but one must consider appropriate ways to evaluate it for different types of contrasts.

Consider the five situations listed in table 3, in order of characteristics leading to decreasing sensitivity of tests (Miller, 1981). To select an appropriate method from those five, one can ask questions in the following order:

1) Are the proposed contrasts orthogonal (Gill, 1978a)? If the anwer is "yes", evaluate (7) as Student's t (i); if not, continue.
2) Are the necessary contrasts only those that compare each experimental group with the control group? If the anwer is "yes", evaluate (7) as Dunnett's t (ii); if not, continue.
3) Are the proposed (nonorthogonal) contrasts relatively few in number and based on objectives (instead of results)? If the answer is "yes", evaluate (7) as the Bonferroni t (iii); if not, continue.
4) Are the treatments not structured (by concentration, or otherwise), so that all pair-wise comparisons can provide unambiguous scientific inferences (i.e., inferences that establish uniqueness of causality with small probability of error)? If the answer is "yes", evaluate (7) as Tukey's t (iv); if not, then the proposed contrasts must be either quite numerous or based on results (instead of objectives), and one should evaluate (7) as Scheffé's t (v).

Experience suggests that the procedure most likely to fit a small set of planned contrasts is (iii), the Bonferroni t, because the most relevant questions about treatments often do not permit orthogonal contrasts and the extra statistical power that could be gained via Student's t. Ordinarily, one should try to avoid methods (iv) and (v), i.e., those of Tukey and Scheffé, to avoid the low power inherent in them. For each of the procedures, the approximate number of df determined by the method of Satterthwaite (1946), is

$$\hat{\upsilon} \cong (\hat{\theta}^2)^2 / \{[(MS_{E(1)})^2 + (p-1)(MS_{E(2)})^2] / tp^2(n-1)\}, \quad (8)$$

where $\hat{\theta}^2$ is from (2).

In some experiments, one may encounter heterogeneous variance among the treatment groups for the data of a given period. That may be suggested by significantly different p x p variance-covariance matrices for different treatments (Gill and Hafs, 1971; Gill, 1978b), or one may test, more specifically, for heterogeneous variance within a given period by using one of the well-known procedures of Bartlett or Hartley (Gill, 1978a), or a procedure by Levene (1960) that is more robust against nonnormality. If heterogeneous variance is evident, and no transformation of scale has been applied to correct the situation, then one can reduce the error df to compensate for the problem in judging significance of test statistics, i.e., to assess more accurately the magnitude of type I error ("P-value"). Grimes and Federer (1979) have shown that adaptations of a procedure attributed to Welch are versatile and reasonably accurate. First, the test statistic in (7) is altered by changing the variance from the expression in (6) to

$$V(\bar{q}_k) = (\sum_{i=1}^{t} c_i^2 S_{ik}^2)/n, \quad (9)$$

where S_{ik}^2 is the sample variance among observations taken in period k for n animals in treatment group i. Then, the critical values in table 3 are altered by replacing the approximate df of equation (8) by reduced approximate df for a given period k,

TABLE 3. CRITICAL VALUES OF TESTS FOR METHODS OF MAKING CONDITIONAL COMPARISONS OF TREATMENTS (WITHIN PERIODS), IN ORDER OF DECREASING SENSITIVITY

Types of treatment comparisons	Test	Critical values[a]	Tables (Gill, 1978c)
(i) Orthogonal[b] (including polynomials)	Student t	$t_{\alpha/2,\hat{\upsilon}}$	A.4
(ii) Each treatment vs control	Dunnett t	$t_{D,\alpha,m,\hat{\upsilon}}$	A.9
(iii) Non-orthogonal[b] (planned and few)	Bonferroni t	$t_{B,\alpha/2,m,\hat{\upsilon}}$	A.10
(iv) All possible pairs of treatments[c]	Tukey t	$(q_{\alpha,t,\hat{\upsilon}})/\sqrt{2}$	(A.8)[d]
(v) Data-based (or many planned)[e]	Scheffé t	$[(t-1)F_{\alpha,t-1,\hat{\upsilon}}]^{.5}$	(A.5)[f]

[a] α = Probability of type I error; $\hat{\upsilon}$ = approximate df for $\hat{\sigma}^2$, from equation (8); m = Number of comparisons within a period; t = Number of treatments in the experiment.

[b] For determination of orthogonality or nonorthogonality of contrasts, see Gill (1978a).

[c] Avoid if treatments are structured (by concentration or otherwise).

[d] Table of Studentized Range does not provide division by $\sqrt{2}$.

[e] Use only in exceptional cases, with many nonorthogonal planned contrasts, or without enough prior knowledge about structure or expected outcome to plan contrasts in advance.

[f] Table of F-values does not provide multiplication by $(t-1)$, nor the square root operation.

RP 9447

$$\hat{v}_k \cong \left(\sum_{i=1}^{t} c_i^2 S_{ik}^2\right)^2 / \left[\sum_{i=1}^{t} (c_i^2 S_{ik}^2)^2/(n-1)\right], \quad (10)$$

where each S_{ik}^2 has (n−1) df.

The accuracy of (10), with respect to valid assessment of the strength of evidence against a proposed hypothesis in the face of heterogeneous variance, may be somewhat hampered when n is as small as three or four (ordinarily, however, experiments with so few animals should be repeated, anyway). In the unlikely event that the variance-covariance matrices differ significantly from treatment to treatment, but each matrix is relatively homogeneous with respect to time, then one could improve the accuracy a little by pooling the variances within each matrix ($S_i^2 = \sum_{k=1}^{p} S_{ik}^2/p$) to replace the S_{ik}^2 in (9) and (10). Pooling, however, would still provide only n − 1 df [not p(n − 1)df] for each variance, because only n animals are involved in one treatment group. Possibly, the pooling would produce somewhat more stable estimates of the variances associated with the treatments, and it would cause the reduced df to be the same for testing the same contrast of treatments in each period.

Consider cases with variance-covariance structure that is homogeneous across treatments, but heterogeneous across time (Gill and Hafs, 1971; Gill 1978b). Then in (6) S_k^2 replaces $\hat{\sigma}^2$, where S_k^2 is from the p × p matrix of variances and covariances pooled over treatments, and has t(n − 1) df, which replace (8). In that case (which is quite common in practice), tests of the $\bar{q}_k$ are likely to be somewhat more sensitive than they would be if the variance-covariance structure were heterogeneous across treatments as well as time.

Conditional Comparisons of Periods. In some experiments, one may have no interest in testing hypotheses about the effects of time, but in others one may wish to examine specific hypotheses about the time trend for a given treatment. Such tests should be relatively sensitive, because the smaller mean square error, $MS_{E(2)}$, is involved. Consider the hypothesis, H: $\Delta_i = 0$, where Δ_i is a specified contrast among period means within treatment group i. Two or more such contrasts may be planned and tested for a given treatment. Orthogonal polynomial contrasts of means may be especially useful (and sensitive) if the mean response trend for a given treatment is smooth enough to permit adequate flexibility of fit to the polynomials.

Let the sample contrast be

$$\bar{q}_i = \sum_{k=1}^{p} c_k \bar{y}_{i\cdot k} \text{ with } \Sigma c_k = 0, \quad (11)$$

where the $\bar{y}_{i\cdot k}$ are period means for treatment group i. The variance of $\bar{q}_i$ is

$$V(\bar{q}_i) = \left(\sum_{k=1}^{p} c_k^2\right)(MS_{E(2)}/n), \quad (12)$$

or just $2MS_{E(2)}/n$ for a simple comparison of two periods for treatment group i. Note that (12) is not the same as (6), the variance used for conditional tests of treatments; (12) is likely to be considerably smaller for a given type of contrast (e.g., the comparison of just two means), because $MS_{E(2)}$ is likely to be much smaller than $\hat{\sigma}^2$.

The test statistic for hypothesis H is

$$t = \bar{q}_i / \sqrt{V(\bar{q}_i)}, \quad (13)$$

and the critical values for various evaluative procedures are the same as in table 3, except that the df for error, $\hat{v}$, are replaced by the df for $MS_{E(2)}$, t(n − 1)(p − 1), in all of the procedures. Also, the Scheffé critical value (procedure v) requires that (p − 1) replace (t − 1), in the multiplier of (and in the df for the numerator of) the tabulated F-value. Procedure (iv), for comparing all possible pairs of period means for a given treatment, should not be used at all, because time always gives structure to a set of means, and one should utilize that structure to formulate a limited set of specific contrasts that may be tested with a more sensitive procedure. Also, the Scheffé procedure (v) should be avoided, if possible, when the number of periods is large, because of extremely low statistical power that occurs in such cases. Use of it may be unavoidable, however, if prior information about expected magnitudes of response over time is scanty or lacking entirely.

If the p × p matrix of sample variances and covariances for p periods (for treatment group i) is found to be heterogeneous (Gill and Hafs, 1971; Gill, 1978b), and no transformation of scale has been made to eliminate the problem, then some alterations may be made to compen-

sate (approximately) for that. First, the test statistic in (13) is altered by changing the variance from the expression in (12) to

$$V(\bar{q}_i) = (\sum_{k=1}^{p} c_k^2 S_{ik}^2 + 2 \sum\sum_{k<k'} c_k c_{k'} S_{ikk'})/n, \quad (14)$$

where S_{ik}^2 is the sample variance among observations taken at period k, for n animals in treatment group i, and $S_{ikk'}$ is the sample covariance between observations taken in periods k and k′ (k<k′) for the same animals. For a simple comparison of only two periods (e.g., periods 3 and 4), (14) becomes

$$V(\bar{q}_i) = (S_{i3}^2 + S_{i4}^2 - 2S_{i34})/n. \quad (15)$$

The critical values of table 3 must be altered by replacing the approximate df for $\hat{\sigma}^2$ (i.e., $\hat{v}$) by the number of degrees of freedom for animals of a given treatment group, n − 1.

However, if the variance-covariance matrices for the t treatment groups are not significantly different, but heterogeneity over time exists (a combination of circumstances frequently encountered), then sensitivity can be improved by using, in (14), the variances and covariances of the pooled matrix

$$(\text{i.e., } S_k^2 = \sum_{i=1}^{t} S_{ik}^2/t \text{ and } S_{kk'} = \sum_{i=1}^{t} S_{ikk'}/t, k < k').$$

In that case, the number of df per contrast is t(n−1).

Comparisons of Treatment Trends Between Specified Periods. As pointed out in the section on efficiency, comparisons of treatments can utilize the reduced error created by repeated measurement only via partition of the interaction of treatments and periods. That is, unconditional tests of treatments require $MS_{E(1)}$ as error, and conditional comparisons within periods require $\hat{\sigma}^2$, but the interaction may be tested validly with $MS_{E(2)}$. A simple and natural way to examine specific aspects of interaction is to create contrasts involving differences of the same desired treatment contrasts from one period to another (not necessarily adjacent).

Consider the hypothesis, H: $(\Delta_k - \Delta_{k'}) = 0$, where Δ_k is a specified contrast among treatment means within period k, and $\Delta_{k'}$ is the same contrast in period k′ (k ≠ k′). Let the sample contrast be

$$\bar{q}_{kk'} = (\bar{q}_k - \bar{q}_{k'}) = \sum_{i=1}^{t} c_i(\bar{y}_{i\cdot k} - \bar{y}_{i\cdot k'}), \quad \text{with } \Sigma c_i = 0 \text{ and } k \neq k', \quad (16)$$

where the $\bar{y}_{i\cdot k}$ and the $\bar{y}_{i\cdot k'}$ are treatment means in the two relevant periods. The simplest contrast involves only two treatments. For example, suppose that one wants to know if the difference between treatments 1 and 2 has changed in magnitude between periods 3 and 4. In that case, (16) becomes

$$\begin{aligned}\bar{q}_{34} = (\bar{q}_3 - \bar{q}_4) &= [(\bar{y}_{1\cdot 3} - \bar{y}_{1\cdot 4}) \\ &\quad - (\bar{y}_{2\cdot 3} - \bar{y}_{2\cdot 4})] \\ &= [(\bar{y}_{1\cdot 3} - \bar{y}_{2\cdot 3}) \\ &\quad - (\bar{y}_{1\cdot 4} - \bar{y}_{2\cdot 4})],\end{aligned}$$

the last form reflecting directly the verbal question of change in treatment difference. The variance of $\bar{q}_{kk'}$ is

$$(\sum_{i=1}^{t} c_i^2)(2MS_{E(2)}/n), \quad (17)$$

or for the simple case of two treatments, just $4MS_{E(2)}/n$.

The test statistic for hypothesis H is

$$t = \bar{q}_{kk'}/\sqrt{V(\bar{q}_{kk'})}, \quad (18)$$

and the critical values for various evaluative methods are the same as in table 3, except that df for $\hat{\sigma}^2$ (i.e., $\hat{v}$) are replaced by df for $MS_{E(2)}$, t(n − 1)(p − 1). Procedures (iv) and (v), i.e., for comparisons of all possible pairs of treatments and data-based contrasts, should be avoided completely.

If the variance-covariance structure is heterogeneous with respect to treatments and time (Gill and Hafs, 1971; Gill, 1978b), and no transformation of scale has been made to eliminate the problem, then some alterations may be made to compensate (approximately) for that. First, the test statistic in (18) is altered by changing the variance calculation from the expression in (17) to

$$V(\bar{q}_{kk'}) = \sum_{i=1}^{t} c_i^2 (S_{ik}^2 + S_{ik'}^2 - 2S_{ikk'})/n, \quad k \neq k', \quad (19)$$

where S_{ik}^2 and $S_{ik'}^2$ are sample variances among observations in periods k and k′ for n animals in

treatment group i, and $S_{ikk'}$ is the sample covariance between periods k and k′ for the same animals. For the simplest case, illustrated by $\bar{q}_{34}$ (previously), the variance becomes

$$V(\bar{q}_{34}) = [(S_{13}^2 + S_{14}^2 - 2S_{134}) + (S_{23}^2 + S_{24}^2 - 2S_{234})]/n.$$

The critical values in table 3 must be altered by replacing the df for $\hat{\sigma}^2$ (i.e., $\hat{v}$) by the approximation:

$$\hat{v}_{kk'} \cong [(n-1)/2]\{[(\sum_{i=1}^{t} c_i^2 S_{ik}^2)^2 / \sum_{i=1}^{t} (c_i^2 S_{ik}^2)^2] + [(\sum_{i=1}^{t} c_i^2 S_{ik'}^2)^2 / \sum_{i=1}^{t} (c_i^2 S_{ik'}^2)^2]\}, k \neq k'. \quad (20)$$

For the simplest case, illustrated by $\bar{q}_{34}$, the approximate df become

$$\hat{v}_{34} \cong [(n-1)/2]\{[(S_{13}^2 + S_{23}^2)^2 / \{(S_{13}^2)^2 + (S_{23}^2)^2\}] + [(S_{14}^2 + S_{24}^2)^2 / \{(S_{14}^2)^2 + (S_{24}^2)^2\}]\}.$$

The accuracy of (20) may be somewhat threatened when n is as small as three or four animals per treatment group. However, if the variance-covariance matrices for the different treatment groups are not significantly different, but the pooled matrix is heterogeneous with respect to time (a common joint result), then accuracy can be improved by using, in (19) and (20), the variances and covariances of the pooled matrix (i.e., S_k^2 and $S_{kk'}$, $k < k'$). In that case, the df from (20) can be replaced by $t(n-1)$. Possibly, (20) may be somewhat less accurate (relatively) than (10), which is used in conditional tests of treatments within periods when variance is heterogeneous across treatment groups, because (20) ignores the sample covariances and involves an average of two approximations for the degrees of freedom.

Perhaps, the source of sensitivity of the tests for partitioned interaction (i.e., for specific aspects of treatment trends for two or more treatments) can be seen most clearly by examining the variance, $V(\bar{q}_{34})$, for the simple illustration of the change in the difference between two treatments from period 3 to period 4 when the variance-covariance structure is heterogeneous. In that case, the two covariance terms, $-2S_{134}$ and $-2S_{234}$, can reduce the variance of the contrast substantially if the correlations between observations in periods 3 and 4, in the two relevant treatment groups, are at least modestly strong and positive. For simplicity, suppose that the correlations are $\rho_{134} = \rho_{234} = .5$, and that $\sigma_{13}^2 = \sigma_{23}^2$ and $\sigma_{14}^2 = \sigma_{24}^2$. Then, $V(\bar{q}_{34})$ would be expected to have only 50% of the magnitude it would have in uncorrelated data.

Example

Generality might be served best by giving an example with three or more treatment groups. However, simplicity and clarity (as well as a certain sense of continuity) can be served better by using an example with two treatments, from mammary physiology (Paape and Tucker, 1969), that has been analyzed and discussed in some statistically-oriented publications (Gill and Hafs, 1971; Gill, 1978b). The original data have been presented in all three of the referenced publications, so it is necessary now only to repeat the summary statistics.

The experiment involved measurement of litter weight gains for n = 6 pregnant lactating rats (treatment group 1) and n = 6 nonpregnant lactating rats (treatment group 2). Litter size was adjusted to maintain one suckling pup per mammary gland, and litters were replaced every 4 d to maintain intense suckling stimulus. The p = 4 periods of measurement were days of lactation 8 to 12, 12 to 16, 16 to 20, and 20 to 24. The eight treatment × period means and the univariate split-plot analysis of variance are shown in table 4. The results are typical for experiments with few animals: strong evidence for period differences and for treatment × period interaction, but weak evidence for the main effect of treatments (i.e., for mean difference between pregnant and nonpregnant rats).

It has been shown (Gill and Hafs, 1971) that these data have sample variance-covariance matrices that are not significantly different for pregnant and nonpregnant animals, but that the pooled matrix is not uniform, i.e., the variances and covariances are not stable with respect to time. That condition could exaggerate the significance levels claimed for periods and interaction, but in this example, evidence for

TABLE 4. MEANS AND ANALYSIS OF VARIANCE OF LITTER WEIGHT GAINS (G) FOR PREGNANT AND NONPREGNANT LACTATING RATS

Item	Period 1	2	3	4
Means				
Pregnant (n = 6)	9.92	10.82	8.87	3.30
Nonpregnant (n = 6)	10.93	11.13	10.28	8.08

Source of variation	df	MS	F-ratio	Significance level
T = treatments (pregnancy status)	1	42.56	2.54	P>.10
E(1) = animals/T	10	16.75		
P = periods	3	68.38	113.59	P<.001
TP = interaction	3	11.87	19.72	P<.001
E(2) = residual error	30	.60		

effects of P and TP is so strong that those inferences are not doubted.

Efficiency. Equations (2) and (3) can be used to estimate the animal variation that would occur without repeated measurement (i.e., with zero correlation) and the average correlation between two periods:

$$\hat{\theta}^2 = [16.75 + 3(.60)]/4 = 4.64 \text{ and}$$

$$\hat{\rho} = (16.75 - .60)/4(4.64) = .87.$$

From (4), the estimated efficiency of repeated measurement in reducing residual error, relative to animal error with correlation, is

$$\{1 + 4[.87/(1 - .87)]\}(100\%) = 2{,}777\%.$$

The estimated efficiency, relative to animal error without correlation, is

$$[1/(1 - .87)](100\%) = 769\%.$$

So, for this example, comparisons that can legitimately use $MS_{E(2)}$ as error should be substantially more sensitive than those that require $MS_{E(1)}$ or $\hat{\theta}^2$. That is, comparisons of specific aspects of treatment trends (via partitioned interaction) and conditional (intra-treatment) comparisons of periods should provide much stronger inferences than the traditional unconditional or conditional (intra-period) comparisons of treatment means.

For the moment, ignore the problem of heterogeneous variance-covariance structure over time, and examine (in table 5) the standard errors of various means and differences of means that would be appropriate otherwise. The standard errors involving $\hat{\theta}^2$ or $MS_{E(2)}$ lack general validity because of the problem of heterogeneity (see the second footnote for table 5). Note that both unconditional and conditional treatment differences have standard errors approximately twice as large as the standard error for a simple difference in treatment trends. One peculiarity is that the standard error of a single period mean involves $\hat{\theta}^2$, whereas the standard errors of unconditional and conditional differences of periods involve $MS_{E(2)}$.

Conditional Comparisons. For comparative purposes, it is useful to analyze specific differences in the traditional way (i.e., to make conditional tests of treatments and periods), as well as in the context of partitioned interaction (i.e., trend differences). First, consider conditional comparisons of treatments. If the variance-covariance structure were homogeneous across both treatments and time, one would use equations (5) through (8). However, in this example the pooled variance-covariance matrix is heterogeneous over time. Therefore, (6) must be altered by replacing $\hat{\theta}^2$ with S_k^2, and the df are $t(n - 1) = 10$, instead of the approximation given by (8). Because the experiment involves only two treatments, the question of how to evaluate multiple test statistics (within a given period) does not arise. Each $\bar{q}_k$ is just the difference between the means for pregnant and nonpregnant rats in period k (table 4), and the relevant standard errors of differences are 1.082, 1.062, 1.122 and 1.623 for the four periods (based on S_k^2 from table 6), instead

TABLE 5. STANDARD ERRORS OF ESTIMATES OF MEANS AND DIFFERENCES OF MEANS (SEM[a] AND SED) FOR THE EXAMPLE OF TABLE 4 (t = 2 TREATMENTS, n = 6 ANIMALS PER TREATMENT, p = 4 PERIODS), ASSUMING HOMOGENEOUS[b] VARIANCES AND COVARIANCES

Item	Standard error
Estimates requiring $MS_{E(1)} = 16.75$	
Treatment mean ($\bar{y}_{i\cdot\cdot}$)	$SEM = [16.75/(6)(4)]^{.5} = .84$
Unconditional treatment difference ($\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$)	$SED = [2(16.75)/(6)(4)]^{.5} = 1.18$
Estimates requiring $\hat{\sigma}^2 = 4.64$	
Conditional (intra-period) treatment difference ($\bar{y}_{1\cdot k} - \bar{y}_{2\cdot k}$)	$SED = [2(4.64)/6]^{.5} = 1.24$
Period mean ($\bar{y}_{\cdot\cdot k}$)	$SEM = [4.64/(6)(2)]^{.5} = .62$
Estimates requiring $MS_{E(2)} = .60$	
Unconditional period difference ($\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdot\cdot k'}$), $k \neq k'$	$SED = [2(.60)/(6)(2)]^{.5} = .32$
Conditional (intra-treatment) period difference ($\bar{y}_{i\cdot k} - \bar{y}_{i\cdot k'}$), $k \neq k'$	$SED = [2(.60)/6]^{.5} = .45$
Treatment trend difference between two periods $[(\bar{y}_{1\cdot k} - \bar{y}_{2\cdot k}) - (\bar{y}_{1\cdot k'} - \bar{y}_{2\cdot k'})]$, $k \neq k'$	$SED = [4(.60)/6]^{.5} = .63$

[a] Generally, it is not advisable to report a standard error (SEM) for the treatment × period combination means $[(\hat{\sigma}^2/n)^{.5} = (4.64/6)^{.5} = .88$ in this example], because too many readers would try to use it in a conditional; comparative context, which requires two grossly different answers for SED, depending on whether one is considering intra-period comparisons of treatments (SED = 1.24, here) or intra-treatment comparisons of periods (SED = .45, here).

[b] For heterogeneous variance-covariance over time, with variances (S_k^2) and covariance ($S_{kk'}$) from a pooled matrix (table 6), replace $\hat{\sigma}^2 = 4.64$ with S_k^2, and replace $(2)(MS_{E(2)}) = (2)(.60)$ with $(S_k^2 + S_{k'}^2 - 2S_{kk'})$, $k < k'$.

of SED = 1.24 from table 5. Therefrom, t-statistics for the four respective periods are .933, .292, 1.257 and 2.945. The appropriate df for the tests are $t(n - 1) = 10$. Only in the fourth period is the evidence for treatment difference (i.e., pregnant rats vs non-pregnant rats) judged to be more than trivial, that test providing $P < .05$ ($t_{.025,10} \cong 2.228$). If one were to choose a Bonferroni critical value, to compensate for dependencies among the four tests (arising from inter-period correlations), then the inferences would be essentially the same, but weaker for the fourth period ($t_{B,.025,4,10} = 3.038$).

For conditional comparisons of periods, suppose that the following contrasts were selected after examining the data (a practice that is not desirable, but often necessary because of prior ignorance when period effects are in question):

1) Pregnant; peak at period 2?: 2 vs (1, 3), $(C_k) = (-1,2,-1,0)$;

2) Pregnant; decline from 3 to 4?: 3 vs 4, $(C_k) = (0,0,1,-1)$;

3) Nonpregnant; decline to 3?: (1, 2) vs 3, $(C_k) = (1,1,-2,0)$;

4) Nonpregnant; decline from 3 to 4?: 3 vs 4, $(C_k) = (0,0,1,-1)$.

From equation (11), the corresponding $\bar{q}_i$ statistics are 2.85, 5.57, 1.50, and 2.20. For example, $\bar{q}_1 = [-9.92 + 2(10.82) - 8.87] = 2.85$, where period 2 is given twice the weight of period 1 or 3.

The pooled variance-covariance matrix is not uniform over time, so the most valid estimates of variances, $V(\bar{q}_i)$, will come from (14), instead of (12). The pooled matrix is shown in table 6, along with calculations for the four necessary variances. There will be $2(n - 1) = 10$ df for each contrast, instead of $t(n - 1)(p - 1) = 30$, because of heterogeneity of the variance-covariance matrix over time. It should be noted that if the matrices for the two treatment groups were significantly different, then variances and covariances for the first two comparisons would come from the matrix for pregnant rats, and those for the last two comparisons would come from the matrix for nonpregnant rats.

TABLE 6. VARIANCE-COVARIANCE MATRIX FOR FOUR PERIODS (POOLED ACROSS TREATMENTS IN THE EXAMPLE OF TABLE 4) WITH CALCULATIONS OF CONTRAST VARIANCES, $V(\bar{q}_i)$, FOR FOUR DATA-BASED COMPARISONS OF PERIODS WITHIN TREATMENTS

Item				
Pooled matrix =	3.512	3.420	3.479	4.418
	3.420	3.368	3.397	4.346
	3.479	3.397	3.778	5.170
	4.418	4.346	5.170	7.901

Pregnant rats – period 2 vs 1 and 3 and period 3 vs 4

$$V(\bar{q}_1) = [S_1^2 + 4S_2^2 + S_3^2 + 2(-2S_{12} + S_{13} - 2S_{23})]/n$$
$$= [3.512 + 4(3.368) + 3.778 - 4(3.420) + 2(3.479) - 4(3.397)]/6 = .0753$$
$$V(\bar{q}_2) = (S_3^2 + S_4^2 - 2S_{34})/n = [3.778 + 7.901 - 2(5.170)]/6 = .2232$$

Nonpregnant rats – periods 1 and 2 vs 3 and period 3 vs 4

$$V(\bar{q}_3) = [S_1^2 + S_2^2 + 4S_3^2 + 2(S_{12} - 2S_{13} - 2S_{23})]/n$$
$$= [3.512 + 3.368 + 4(3.778) + 2(3.420) - 4(3.479) - 4(3.397]/6 = .2213$$
$$V(\bar{q}_4) = (S_3^2 + S_4^2 - 2S_{34})/n = V(\bar{q}_2) = .2232$$

The four test statistics, from (13), are 10.386, 11.790, 3.189 and 4.657. For data-based contrasts among p periods, the Scheffé critical value for a t-statistic (table 3) is $[(p-1)F_{\alpha,p-1,t(n-1)}]^{.5}$ or $(3F_{\alpha,3,10})^{.5}$. Then, for $\alpha = .05$ and .01, the critical values are 3.34 and 4.43, respectively. Three of the test statistics are significant at $P < .01$, the other being not quite significant at $P = .05$. So, one can conclude with confidence that, for pregnant rats, litter gains peak at period 2 and decline from period 3 to 4. Also, for nonpregnant rats, litter gains possibly decline from the first two periods to the third, but clearly decline from period 3 to period 4.

Comparisons of Treatment Trends Between Two Specified Periods. Suppose that the primary interest, at the beginning, was in sequential differences in responses to treatments. That is, pregnant and nonpregnant rats should be compared for changes in litter weight gains from period 1 to 2, from 2 to 3 and from 3 to 4 (In general, one need not restrict comparisons to adjacent periods). The $\bar{q}_{kk'}$ of (16) can be written as

$$\bar{q}_{12} = (\bar{y}_{2\cdot 1} - \bar{y}_{1\cdot 1}) - (\bar{y}_{2\cdot 2} - \bar{y}_{1\cdot 2})$$
$$= (10.93 - 9.92) - (11.13 - 10.82)$$
$$= +.70,$$

$$\bar{q}_{23} = (\bar{y}_{2\cdot 2} - \bar{y}_{1\cdot 2}) - (\bar{y}_{2\cdot 3} - \bar{y}_{1\cdot 3})$$
$$= (11.13 - 10.82) - (10.28 - 8.87)$$
$$= -1.10 \text{ and}$$

$$\bar{q}_{34} = (\bar{y}_{2\cdot 3} - \bar{y}_{1\cdot 3}) - (\bar{y}_{2\cdot 4} - \bar{y}_{1\cdot 4})$$
$$= (10.28 - 8.87) - (8.08 - 3.30)$$
$$= -3.37.$$

Because the pooled variance-covariance matrix is heterogeneous over time, one cannot use $MS_{E(2)}$ in (17). One must use (19) to find the $V(\bar{q}_{kk'})$. Because of pooling of the variances and covariances over treatments df will be $t(n-1) = 10$, instead of being determined from (20). Therefore, (19) simplifies to the following expressions (with values from table 6):

$$V(\bar{q}_{12}) = 2(S_1^2 + S_2^2 - 2S_{12})/n = [2(3.512) + 2(3.368) - 4(3.420)]/6$$
$$= .0133,$$

$$V(\bar{q}_{23}) = 2(S_2^2 + S_3^2 - 2S_{23})/n = [2(3.368) + 2(3.778) - 4(3.397)]/6$$
$$= .1173 \text{ and}$$

$$V(\bar{q}_{34}) = 2(S_3^2 + S_4^2 - 2S_{34})/n = [2(3.778) + 2(7.901) - 4(5.170)]/6$$
$$= .4463.$$

The three comparisons are not orthogonal. Therefore, the t-statistics (6.071, −3.216, −5.045) may be evaluated, conservatively, by the Bonferroni critical values, $\pm t_{B,\alpha/2,3,10}$. For $\alpha = .05$ and .01, the respective values are ± 2.870 and ± 3.827. Therefore, evidence is clear that the treatment trends in litter weight gains for pregnant and nonpregnant lactating rats are not the same between any two adjacent periods.

The clarity of those results is in sharp distinction from those of the conditional (intra-period) comparisons of treatments, which showed only a possible treatment difference in period 4. Thus, the relative sensitivity of tests for treatment trends is demonstrated for experiments with few animals. Those tests are more powerful than the traditional tests of treatments within periods even in large experiments, but there the relative advantage in sensitivity may be somewhat less important for clarity of inference.

Literature Cited

Cole, J.W.L., and J. E. Grizzle. 1966. Application of multivariate analysis of variance to repeated measurements experiments. Biometrics 22:810

Gill, J. L. 1978a. Design and Analysis of Experiments in the Animal and Medical Sciences, Vol. 1. Iowa State Univ. Press, Ames.

Gill, J. L. 1978b. Design and Analysis of Experiments in the Animal and Medical Sciences, Vol. 2. Iowa State Univ. Press, Ames.

Gill, J. L. 1978c. Design and Analysis of Experiments in the Animal and Medical Sciences, Vol. 3. Iowa State Univ. Press, Ames.

Gill, J. L., and H. D. Hafs. 1971. Analysis of repeated measurements of animals. J. Anim. Sci. 33:331.

Grimes, B. A., and W. T. Federer. 1979. Cochran-like and Welch-like approximate solutions to the problem of comparison of means from two or more populations with unequal variances. Amer. Statist. Assoc., Proc. Soc. Statist. Sec. p 628.

Levene, H. 1960. Robust tests for equality of variances. Contributions to Probability and Statistics, pp 278–292 In: I. Olkin (Ed.) Stanford Univ. Press, Palo Alto, CA.

Miller, R. G., Jr. 1981. Simultaneous Statistical Inference (2nd Ed.). Springer-Verlag, New York, NY.

Paape, M. J., and H. A. Tucker. 1969. Mammary nucleic acid, hydroxyproline, and hexosamine of pregnant rats during lactation and post-lactational involution. J. Dairy Sci. 52:380.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biometrics 2:110.