Procedures for Comparing Samples with Multiple Endpoints

Peter C. O'Brien
Section of Medical Research Statistics, Mayo Clinic,
Rochester, Minnesota 55905, U.S.A.

SUMMARY

Five procedures are considered for the comparison of two or more multivariate samples. These procedures include a newly proposed nonparametric rank-sum test and a generalized least squares test. Also considered are the following tests: ordinary least squares, Hotelling's $T^2$, and a Bonferroni per-experiment error-rate approach. Applications are envisaged in which each variable represents a qualitatively different measure of response to treatment. The null hypothesis of no treatment difference is tested with power directed towards alternatives in which at least one treatment is uniformly better than the others. In all simulations the nonparametric procedure provided relatively good power and accurate control over the size of the test, and is recommended for general use. Alternatively, the generalized least squares procedure may also be useful with normally distributed data in moderate or large samples. A convenient expression for this procedure is obtained and its asymptotic relative efficiency with respect to the ordinary least squares test is evaluated.

1. Introduction

Clinical trials are often conducted for the purpose of evaluating the relative efficacy of two or more modes of therapy. When the therapies are administered to independent groups and efficacy is measured on the basis of a single response variable, appropriate parametric and nonparametric one-way analysis-of-variance (ANOVA) procedures and their properties are well-known. Often, however, efficacy is measured by more than one variable. Although univariate methods for assessing each characteristic individually are useful in this setting, there is often the additional need for a single, overall, objective probability statement that addresses the question of whether or not the experimental therapy is efficacious. This need is particularly acute when the medical question is controversial and the sample size is small.

Although a large literature exists on the comparison of multivariate samples, the procedures tend to be infrequently used in practice. As pointed out by Meier (1975, p. 523), the standard analysis for the comparison of two multivariate samples, which is based on Hotelling's $T^2$ statistic, addresses somewhat the wrong question and consequently has very poor power for the alternatives of primary interest. Specifically, the $T^2$ procedure addresses the question 'Are one or more of the treatments different?', making no distinction between variables that change favorably and variables that change unfavorably.

A second approach would be to assign per-experiment error rates to each of the univariate tests by using Bonferroni's inequality (i.e. by multiplying each univariate $P$-value by the number of variables studied). The desired overall probability statement is then given by the minimum of the per-experiment univariate $P$-values. Although this approach may be useful in some settings, it may lack power for alternatives in which most or all measures of efficacy are improved. This will be of particular concern when the number, $K$, of endpoints studied is large relative to sample size. In fact, statistical significance will not be possible in some instances. This problem will be exacerbated when the measures of efficacy are highly

Key words: Multivariate; Analysis of variance; Nonparametric; Least squares; Maximum likelihood; Clinical trials.
correlated, as is often the case. To illustrate, with perfect correlation a test with nominal size $\alpha$ will have true size $\alpha/K$.

We have therefore considered three additional procedures. These are described in §2, with some theoretical results in §3 and additional comparisons via Monte-Carlo simulations in §4. The nonparametric procedure is illustrated in §5 with data from a clinical trial of an experimental treatment for diabetes.

2. Notation and Statement of the Procedures

Let $Y_{ijk}$ represent the $k$th variable for the $j$th subject in Group $i$ ($k = 1, \ldots, K$, $j = 1, \ldots, n_i$, $i = 1, \ldots, I$), and define

$$E(Y_{ijk}) = \mu_{ijk},$$

$$\text{cov}(Y_{ijk}, Y_{i'j'k'}) = \begin{cases} 
\sigma_{kk'} & \text{if } i, j = i', j', \\
0 & \text{otherwise.}
\end{cases}$$

In matrix notation the vectors $Y_{ij}$ ($j = 1, \ldots, n_i$, $i = 1, \ldots, I$) are independently distributed with mean $\mu_i$ and covariance matrix $\Sigma$. The null hypothesis is $H_0: \mu_1 = \cdots = \mu_K$. To simplify the notation, we shall assume that $Y_{ijk}$ is defined so that large values are better than small values for each $k = 1, \ldots, K$. Thus, alternatives for which $\mu_{ijk} > \mu_{i'k}$ for $k = 1, \ldots, K$ are of primary interest.

The proposed nonparametric procedure is a rank-sum-type test. Let $R_{ijk}$ represent the rank of $Y_{ijk}$ among all values of variable $k$ in the pooled set of $I$ samples, and define $S_{ij}$ as the sum of the ranks assigned to the $j$th person in Sample $i$ ($i = 1, \ldots, I$, $j = 1, \ldots, n_i$, $k = 1, \ldots, K$). Perform a one-way analysis of variance on the $S_{ij}$ values. Equivalently, for comparing two samples, one may use a two-sample $t$ test; an alternative is the two-sample rank-sum test.

We next consider two parametric approaches. In both, it is necessary first to express the data in common units via a transformation to relative deviates, subtracting the overall variable mean from each observation and dividing by the pooled within-group sample standard deviation. To simplify the notation, we assume this step has been completed in obtaining $Y_{ijk}$.

The usual approach for testing among-group main effects in repeated-measures designs is to compute the mean of the observations on each subject, then perform a one-way analysis of variance. As pointed out by Anderson (1958), this test may be derived via ordinary least squares (OLS) or maximum likelihood (ML) methods. The corresponding OLS–ML procedure in the present setting is to perform the analogous computations on the standardized data.

Since the generalized least squares (GLS) procedure provides best linear unbiased estimates and corresponding optimal tests, it is of interest to consider the analogous procedure in which the unknown covariance matrix is replaced by its sample estimate. The use of the GLS method has been studied in the context of repeated-measures designs by Grizzle and Allen (1969). We propose the following procedure. Compute

$$F = \sum_i n_i |J' \hat{\Sigma}^{-1}(\bar{Y}_{i.} - \bar{Y}_{..})|^2 / [(I - 1)J' \hat{\Sigma}^{-1}J],$$

where

$$J' = (1, \ldots, 1),$$

$$\bar{Y}_{i.} = \sum_j Y_{ij} / n_i,$$

$$\bar{Y}_{..} = \sum_{ij} Y_{ij} / \sum n_i.$$
and

$$\hat{\Sigma}_{\nu} = \sum_{ij} (Y_{ij\nu} - \bar{Y}_{i\nu})(Y_{ij\nu} - \bar{Y}_{i\nu})/\Sigma_i(n_i - 1).$$

Reject \(H_0\) if \(F\) exceeds the \((1 - \alpha) \times 100\) percentile of the standard \(F\) distribution with \(I - 1\) and \(\Sigma(n_i - K)\) degrees freedom (df) in the numerator and denominator, respectively.

Notice that the procedure consists essentially of weighting the \(K\) variables according to the row totals of \(\hat{\Sigma}^{-1}\). An alternative computational procedure for obtaining \(F\) is to define \(S^*_\eta = J' \hat{\Sigma}^{-1}Y_{ij}\) and compute the \(F\) statistic associated with a one-way ANOVA on \(\{S^*_\eta\}\). For comparing two samples, one may use a two-sample \(t\) test with \(\Sigma(n_i - K)\) df.

3. Theoretical Considerations

Since \(\text{cov}(R_{ij\beta}, R_{i'j', \beta}) \to 0\) as \(\Sigma n_i \to \infty\) for \(ij \neq i'j',\) the \(\{S_{ij}\}\) are uncorrelated asymptotically. Thus the central limit theorem is sufficient to ensure that the nonparametric procedure maintains the size of the test in large samples. The error rates in small samples are evaluated on the basis of Monte Carlo simulations in §4.

To derive the parametric procedures, consider the model

$$\mu_i = \mu + \beta_i J, \quad i = 1, \ldots, I,$$

where \(\{\beta_i\}\) are scalars such that \(\Sigma \beta_i = 0\). The GLS estimator for \(\beta_i\) is given by

$$\hat{\beta}_i = J' \Sigma^{-1}(\check{Y}_i - \bar{Y})/(J' \Sigma^{-1} J)$$

with

$$\text{var}(\hat{\beta}_i) = (1/n_i + 2/In_i + 1/I\check{n})/(J' \Sigma^{-1} J),$$

where \(\check{Y} = \Sigma \check{Y}_i/I\) and \(\check{n}\) is the harmonic mean sample size (see Appendix). The GLS test criterion is defined by \(F\) with \(\hat{\Sigma}\) replaced by \(\Sigma\).

Similarly, the OLS estimator is

$$\hat{\beta}_i = J' (\check{Y}_i - \bar{Y})/K$$

with

$$\text{var}(\hat{\beta}_i) = [(1/n_i + 2/In_i + 1/I\check{n})J' \Sigma J]/K^2.$$

It follows that the efficiency of the OLS procedure relative to the GLS procedure is

$$\text{RE} = \frac{\text{var}(\hat{\beta}_i) / \text{var}(\hat{\beta}_i)}{K^2/[(J' \Sigma J)(J' \Sigma^{-1} J)]}.$$ 

From the extended Cauchy–Schwartz inequality (Johnson and Wichern, 1982, p. 66), \(\text{RE} \leq 1\), with equality holding if and only if the row totals of \(\Sigma\) are constant. In particular, it is sufficient but not necessary that the variables be equicorrelated with common variance.

If the underlying distributions are multivariate normal, the OLS test statistic follows a standard \(F\) distribution. For the GLS procedure, it is natural to consider an \(F\) distribution, since the test statistic depends on the data only through mean vectors that have a multivariate normal distribution and an independently distributed sample covariance matrix with a Wishart distribution. It was observed in simulations that an \(F\) distribution with \(I - 1\) and \(\Sigma(n_i - K)\) df provided an accurate fit in a wide variety of parametric configurations, even for quite small sample sizes. That this approximation holds exactly
asymptotically follows as an immediate consequence of the following lemma, stated without proof (see Billingsley, 1968, Theorems 4.4 and 5.1).

Lemma. Let $A_n$ and $B_n$ be sequences of random variables, such that $A_n$ converges in distribution to $A$, and $B_n$ converges in probability to $b$, a constant, and let $g$ be a continuous function. Then $g(A_n, B_n)$ converges in distribution to $g(A, b)$.

That the same result holds for any underlying distribution with finite second moments follows from the central limit theorem for multivariate distributions (Anderson, 1958, p. 74). The adequacy of the approximation in small samples is considered in §4.

As a consequence of the preceding discussion, we obtain the following generalization of the well-known result for the combination of independent estimators that have unequal variances.

Theorem. Let $X' = (X_1, \ldots, X_k)$ represent $K$ unbiased estimators of a scalar parameter $\mu$ with covariance matrix $\Sigma$. Then the best linear unbiased estimate of $\mu$ is $(J'\Sigma^{-1}X)/(J'\Sigma^{-1}J)$, with variance $1/(J'\Sigma^{-1}J)$.

4. Simulations

In order to evaluate the operating characteristics of the procedures with small to moderate sample sizes, we simulated 1000 experiments for a variety of parametric configurations. Although each configuration involved a different set of 1000 experiments, the various procedures were applied to the same data in each instance. Two samples of size $n = 5, 20, 50$ were compared in each experiment, with size and power evaluated on the basis of one-sided tests, larger values being anticipated for the first sample. In all simulations, components of the vectors $\{Y_{ij}\}$ were generated by means of the formula $Y_{ijk} = c_kX_{ij0} + (1 - c_k^2)^{1/2}X_{ijk}$, $k = 1, \ldots, K$. The $X$ variables were generated independently from a variety of distributions (normal, exponential and Cauchy). The variable $X_{ij0}$ may be interpreted as representing variability among subjects, and $\{X_{ijk}\}$ the variability within subjects. Various correlation structures were evaluated by varying the values of $c_k$.

In evaluating the ability of the procedures to control the size of the test, our concern centered on the newly proposed nonparametric test and the GLS test. The OLS test was also evaluated for comparative purposes. When the underlying distributions were multivariate normal, results from OLS tests were used to improve estimates for the other two procedures (by subtracting the difference between the observed and expected rejection rates obtained with OLS procedure). In generating multivariate normal distributions, variables were all NID(0, 1). Since the procedures are all invariant under change in scale of the individual variables, only different correlation structures were considered: $c_k = 1/\sqrt{2}$ for equal correlation [asymptotic relative efficiency (ARE) = 1], and $c_1 = c_2 = c_3 = c_4 = \sqrt{.9}$ and $c_5 = 0$ for unequal correlation (ARE = .76). As indicated in the top half of Table 1, both of these procedures provide accurate control over the size of the test, even in quite small samples. Additional simulations not reported here for $n = 5$ and $k = 3, 4$ showed equally accurate control over the size.

In order to evaluate the robustness of the procedures, two non-normal distributions were simulated. The first was highly skewed: each of the $X$ variables was independent, having a common exponential distribution with unit variance. To evaluate the effect of outliers, the variables $X_1, \ldots, X_k$ were NID(0, 1), but $X_0$ was taken to have a standard Cauchy distribution. The variables were equally correlated in both cases. As would be anticipated, the presence of outliers causes the normal-theory procedures to become quite conservative, but has little effect on the probability of Type-I error for the nonparametric procedure (bottom half, Table 1).
To evaluate relative power, we repeated the simulations, adding .5 to each observation in the first sample. In addition, a staggered configuration was used for the equicorrelated multivariate normal distribution, in which \( .5k/K \) was added to \( Y_{ijk} \), \( k = 1, \ldots, K \). We also include the results from Hotelling’s \( T^2 \) test and from application of Bonferroni’s inequality to multiple \( t \) tests.

The results in Table 2 suggest the following conclusions:

(i) \( T^2 \) tests are very insensitive to the alternatives considered here and should not be used in the present context.

(ii) The nonparametric test is relatively efficient in all the parametric configurations studied.

(iii) The GLS procedure appears to be nearly as efficient as the ordinary least squares OLS approach when correlations are equal and somewhat more efficient otherwise.

(iv) The Bonferroni approach is generally less powerful for the main-effect types of alternatives considered here, particularly when the number of variables is large relative to sample size. When applied to multiple \( t \) tests, power is also adversely affected by departures from normality (more so than for the GLS and OLS procedures). On the other hand, one would expect the Bonferroni approach to be more powerful than the rank-sum, GLS or OLS procedures when only one endpoint is affected by treatment. The staggered configuration appears to be a midway point between these two alternatives.

5. Example

This research was motivated by a randomized trial comparing two therapies, experimental and conventional, for the treatment of diabetes. The objective of the study was to determine whether the experimental therapy resulted in better nerve function, as measured by

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Correlation</th>
<th>( n )</th>
<th>( K )</th>
<th>( \alpha = .05 )</th>
<th>( \alpha = .01 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rank-sum*</td>
<td>GLS*</td>
</tr>
<tr>
<td>Normal</td>
<td>Equal</td>
<td>5</td>
<td>2</td>
<td>.045</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2</td>
<td>.050</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>5</td>
<td>.051</td>
<td>.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>5</td>
<td>.048</td>
<td>.058</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>50</td>
<td>.041</td>
<td>—</td>
</tr>
<tr>
<td>Normal</td>
<td>Unequal</td>
<td>20</td>
<td>5</td>
<td>.047</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>5</td>
<td>.053</td>
<td>.051</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>50</td>
<td>.050</td>
<td>—</td>
</tr>
<tr>
<td>Exponential</td>
<td>Equal</td>
<td>5</td>
<td>2</td>
<td>.043</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2</td>
<td>.062</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>5</td>
<td>.056</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>5</td>
<td>.035</td>
<td>.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>50</td>
<td>.052</td>
<td>—</td>
</tr>
<tr>
<td>Cauchy</td>
<td>Equal</td>
<td>5</td>
<td>2</td>
<td>.047</td>
<td>.021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>2</td>
<td>.063</td>
<td>.032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>5</td>
<td>.051</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>5</td>
<td>.052</td>
<td>.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>50</td>
<td>.053</td>
<td>—</td>
</tr>
</tbody>
</table>

* Estimates for normal distributions were refined using variance-reduction techniques based on OLS results.
<table>
<thead>
<tr>
<th>Correlation</th>
<th>( n )</th>
<th>( K )</th>
<th>( \alpha = .05 )</th>
<th>( \alpha = .01 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rank-sum</td>
<td>GLS</td>
</tr>
<tr>
<td>Normal, equal correlation</td>
<td>20</td>
<td>5</td>
<td>.639</td>
<td>.669</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>.914</td>
<td>.922</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
<td>.275</td>
<td>—</td>
</tr>
<tr>
<td>Normal (staggered means), equal correlation</td>
<td>20</td>
<td>5</td>
<td>.319</td>
<td>.347</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>.620</td>
<td>.643</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
<td>.138</td>
<td>—</td>
</tr>
<tr>
<td>Normal, unequal correlation</td>
<td>20</td>
<td>5</td>
<td>.625</td>
<td>.745</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>.920</td>
<td>.967</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
<td>.177</td>
<td>—</td>
</tr>
<tr>
<td>Skewed, equal correlation</td>
<td>20</td>
<td>5</td>
<td>.765</td>
<td>.661</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>.990</td>
<td>.927</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
<td>.377</td>
<td>—</td>
</tr>
<tr>
<td>Outliers, equal correlation</td>
<td>20</td>
<td>5</td>
<td>.230</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5</td>
<td>.419</td>
<td>.081</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>50</td>
<td>.106</td>
<td>—</td>
</tr>
</tbody>
</table>

\((c = .5)\)
Comparisons with Multiple Endpoints

Table 3
Nonparametric analysis of diabetes trial

<table>
<thead>
<tr>
<th>Variables included</th>
<th>Number of variables</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>34</td>
<td>.033</td>
</tr>
<tr>
<td>Speed of conduction</td>
<td>21</td>
<td>.044</td>
</tr>
<tr>
<td>Sensory conduction</td>
<td>13</td>
<td>.045</td>
</tr>
<tr>
<td>Lower extremity function</td>
<td>13</td>
<td>.156</td>
</tr>
<tr>
<td>Distal function</td>
<td>8</td>
<td>.164</td>
</tr>
<tr>
<td>Proximal function</td>
<td>6</td>
<td>.001</td>
</tr>
<tr>
<td>Elevated from normal* initially</td>
<td>14</td>
<td>.156</td>
</tr>
<tr>
<td>Most reliably measured</td>
<td>8</td>
<td>.029</td>
</tr>
</tbody>
</table>

* Overall mean $> \mu + .15$ for normals.

34 electromyographic (EMG) variables, than the standard therapy. Six subjects were randomized to standard therapy, five to experimental therapy. At eight months, the change in EMG measurements from baseline was determined for each subject.

Despite the small sample size, the objective of the study was viewed primarily as confirmatory rather than exploratory: to test the hypothesis of no improvement in nerve function rather than to identify hypotheses (or subgroupings of EMG measurements) for further study. Furthermore, the medical question at issue was controversial, requiring an overall quantitative and objective probability statement.

Nonparametric analysis was used throughout because of severe departures from normality. The results of 34 separate rank-sum tests (one-sided), as summarized below, suggest that an improvement had occurred with the experimental therapy.

Interval: 0-.1  .1-.2  .2-.3  .3-.4  .4-.5  .5-.6  .6-.7  .7-.8  .8-.9  .9-1.0
Percent of 34 P-values in interval: 38 12 9 15 6 2 9 3 0 3

There were six EMG variables for which the difference between groups was statistically significant ($P < .05$). The two smallest P-values were $P = .002$ and $P = .015$. The treated group did better in 28 of 34 variables, with $P < .50$ as a criterion; this indicated the type of main effect hoped for. Although these results appear to support the hypothesis of a beneficial effect associated with the experimental therapy, the appropriate method for quantifying this impression was unclear. For example, a $T^2$ test is obviously not possible. Application of a multiple-comparison-type per-experiment error rate is also meaningless here owing to small sample size relative to the large number of variables. Specifically, the smallest P-value of .002 is also the smallest possible with these sample sizes, representing complete separation of the groups. Yet the corresponding per-experiment error rate obtained via the Bonferroni inequality ($P = .068$) is not significant at the .05 level.

These considerations led to our formulation of the nonparametric procedure. When applied to all the data, it yielded a P-value of .033. As indicated in Table 3, the investigators then formed seven overlapping subgroups (without viewing the data) in an attempt to understand the nature of the effect. In accord with the biological mechanisms believed to operate, it was found that the effect on nerve function was most apparent proximally.

6. Discussion

Reports in the medical literature often compare two or more treatments in terms of several qualitatively different endpoints. Analyses, though often quite extensive in terms of descrip-
tion and hypothesis testing, typically fail to provide a single overall probability statement as to whether or not a difference among treatments exists. Typically the reader is left with the caveat that, since multiple tests have been performed, corresponding probability statements should be interpreted with caution. In this context, our purpose has been to augment existing practice by providing procedures for making such overall probability statements.

Specifically, we sought a single overall test of the null hypothesis of no treatment effect, which would be sensitive to departures in which some improvement was demonstrated consistently among the various endpoints. Of particular concern is the situation in which sample sizes are small relative to the number of endpoints studied so that, while separate tests on each variable individually may fail to demonstrate statistical significance, the combined evidence from all measures may be convincing. As illustrated by the numerical simulations, conventional \( T^2 \) tests are insensitive in this respect; per-experiment error rates may also be insensitive, though less so. As illustrated by the diabetes data, neither approach may be feasible in some situations.

We recommend the nonparametric test for general use, particularly if the variables are not normally distributed or the sample size is small. There is little loss of efficiency when variables are normally distributed, and gains in power may be considerable otherwise. Alternatively, the GLS procedure is remarkably robust and achieves optimality in the normal-theory setting by utilizing information contained in the correlation matrix. Both procedures have been shown to provide accurate control over the probability of Type-I error, with the approximations becoming exact asymptotically.

We emphasize that the proposed procedures are suggested as a supplement (not an alternative) to univariate methods. They may not be helpful in certain situations. Clearly, the methods are not designed to detect treatment effects that are expected to occur in only relatively few measurements. In this case, Bonferroni-type per-experiment error rates would be more appropriate. Alternatively, it may be advantageous in some situations to apply the proposed methods to a subset of variables for which a treatment effect is judged most likely. In this case, it would be important that the subset be identified on an \textit{a priori} basis that will be recognized by other investigators. Unfortunately, it is our experience that this is often not possible. The proposed methodology may also be inappropriate in some exploratory settings in which the direction of change (improvement or worsening) cannot be anticipated in advance and may not be consistent among the variables under study. In this case, a multiple-comparison approach or \( T^2 \) analysis may be more useful.

Finally, we evaluated the performance of the nonparametric GLS and OLS procedures in the repeated-measures setting, i.e. the same characteristic is measured under differing conditions. For excellent discussions, see Koch (1969) and Grizzle and Allen (1969). The step involving transformation to standardized variables was omitted, although a modified GLS test with this step reinserted was also evaluated. A full discussion of our findings is available in an unpublished report by P. C. O'Brien (Technical Report No. 25, Section of Medical Research Statistics, Mayo Clinic, 1983). To summarize, all procedures except the GLS provided accurate control over the size of the test. For example, if variables follow a multivariate normal distribution with grossly unequal variances, the estimated probability of Type-I error with \( \alpha = .05 \), \( n = 50 \) and \( k = 5 \) is \(.058 \) (rank-sum), \(.067 \) (GLS) and \(.052 \) (modified GLS). However, the gain in power with the GLS procedure is dramatic, with observed rejection rates of \(.219 \) (rank-sum), \(.547 \) (GLS), \(.328 \) (modified GLS) and \(.147 \) (OLS) when \(.5 \) was added to each observation in Group 1. Our conclusions in the repeated-measures setting were essentially the same as we report here except that, when variances are grossly dissimilar, the GLS procedure may considerably enhance power at the expense of a slight increase in the size of the test.
ACKNOWLEDGEMENTS

I would like to thank Dr Alan Zinsmeister for his helpful comments, Dr F. J. Service for the use of his data from the diabetes trial, and Mr Jon Kosanke for programming of the Monte Carlo simulations.

REFERENCES


Received August 1983; revised April 1984

APPENDIX

We indicate the derivation of \( \hat{\beta}_i \) for the case \( I = 2 \). Let \( Y' = (Y_{i1}, \ldots, Y_{in}, Y_{12}, \ldots, Y_{2m}) \) so the model may be represented as \( Y = X\theta + \epsilon \), where \( X \) is the design matrix, \( \theta' = (\beta_1, \mu_2, \ldots, \mu_k) \) after reparameterizing with \( \mu_1 = 0 \) to obtain a full-rank model, and \( \epsilon \) is the error whose covariance matrix will be denoted by \( V \). The GLS estimator for \( \theta \) is given by \( \hat{\theta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \), where

\[
(X'V^{-1}X) = \begin{bmatrix}
(n_1 + n_2)J'\Sigma^{-1}J & J'\Sigma^{-1}(n_1 - n_2) \\
\Sigma^{-1}J(n_1 - n_2) & (n_1 + n_2)\Sigma^{-1}
\end{bmatrix},
\]

\[
(X'V^{-1}X)^{-1} = (4n_1n_2J'\Sigma^{-1}J)^{-1} 	imes 
\begin{bmatrix}
(n_1 + n_2) & -(n_1 - n_2)J' \\
-(n_1 - n_2)J & \frac{JJ'}{n_1 + n_2} (n_1 - n_2)^2 + \frac{\Sigma}{n_1 + n_2} (4n_1n_2J'\Sigma^{-1}J)
\end{bmatrix}
\]

\[
X'V^{-1}Y = \begin{bmatrix}
J'\Sigma^{-1}(n_1\bar{Y}_1 - n_2\bar{Y}_2) \\
\Sigma^{-1}(n_1\bar{Y}_{12} + n_2\bar{Y}_{22})
\end{bmatrix}
\]

The desired expression for \( \hat{\beta}_i \) is obtained from the indicated matrix multiplication.