THE CONSULTANT’S FORUM

Robust Analysis of Within-Unit Variances in Repeated Measurement Experiments

A. R. Padmanabhan,1 V. M. Chinchilli,2 and Gutti Jogesh Babu3

1Department of Mathematics, Monash University, Clayton, Victoria 3168, Australia
2Center for Biostatistics and Epidemiology, College of Medicine, Pennsylvania State University, Hershey, Pennsylvania 17033, U.S.A.
3Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

SUMMARY

The objective of some experiments is to compare the within-unit variances of two or more treatments, products, or techniques. In this situation, a repeated measurement design involving a random effects model, with possibly heterogeneous variances, is appropriate. Under the assumption that the random errors have a normal or a multivariate $t$-distribution, this design was analyzed in Chinchilli, Esinhart, and Miller (1995, *Biometrics* 51, 215–216). However, the resulting methodology is quite vulnerable to skewness and outliers. We propose two distribution-free procedures that are quite robust for balanced designs when the number of repeated measurements is the same for all units and for all treatments. We then show how these procedures are modified to handle unbalanced situations. We illustrate the methodology with an example from a trial comparing serum cholesterol measurements from a routine laboratory analyzer with those of a standardized method.

1. Introduction

The comparison of the variances of two treatments, products, or techniques is an important objective for some experiments. Usually it is more critical to compare the within-unit variances rather than between-unit variances, and therefore investigators need to design repeated measurement experiments in these situations. (Strictly speaking, we should say within-unit variances of the treatments. However, by an abuse of the language, we shall simply say within-unit variances.) Two examples are the comparison of the within-unit variances of (1) a new laboratory analytical technique to a standard technique and (2) a generic pharmaceutical formulation to a standard pharmaceutical formulation.

This design was analyzed in Chinchilli, Esinhart, and Miller (1995) under each of two assumptions:

**Assumption 1:** The random errors have a normal distribution.

**Assumption 2:** The random errors have a multivariate $t$-distribution.

However, the corresponding methodologies are quite vulnerable to violations of these assumptions. The tests corresponding to Assumption 1 may become quite liberal, while those corresponding to Assumption 2 may become quite conservative, resulting in significant reduction in power.

*Key words:* Bootstrap; Confidence limits; Intra-subject variances; Sum of squared ranks test; Variance ratios; Wilcoxon test.
In Section 2, dealing with balanced designs, we introduce distribution-free procedures based on the Wilcoxon statistic $h_W$, the sum of squared ranks statistic $h_{SQ}$, and an asymptotically distribution-free procedure based on the bootstrap.

In Section 3, we briefly indicate how the $h_W$ and $h_{SQ}$ procedures are modified for unbalanced designs. In Section 4, we present results of Monte Carlo studies involving normal and nonnormal random errors, which show that the best all-round performers are $h_{SQ}$ and $h_W$, in that order. In Section 5, we apply the $h_{SQ}$ and $h_W$ procedures to data from a trial comparing a routine laboratory analyzer to a standardized method.

2. Distribution-Free Procedures for a Balanced Design

For later reference, we begin by recalling a result from the two-sample scale problem. Let $t_1 = (t_{11}, \ldots, t_{1n_1})$ and $t_2 = (t_{21}, \ldots, t_{2n_2})$ be samples, respectively, from distribution functions $F_1$ and $F_2$ such that $F_1(u) = F(u/r_1)$, $F_2(u) = F(u/r_2)$, where $F$ has support $(0, \infty)$ (which ensures that the random variables are positive valued) and $F$ is right skewed. Then, for inference regarding $r_2/r_1$, the ratio of scale parameters, the test based on the Wilcoxon statistic $h_W$ performs well, while the test based on the sum of squared ranks statistic $h_{SQ}$ performs even better (cf., Duran and Mielke, 1968, Table 1).

In what follows, we shall use the terms units and subjects interchangeably. We consider an experimental design in which the response is measured on multiple occasions under each treatment or experimental condition for every experimental unit. Our model is

$$Y_{ijk} = \mu_{ij} + U_{ij} + E_{ijk},$$

(1)

$i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, and $k = 1, 2, \ldots, m$, where $Y_{ijk}$ is the $k$th measurement under the $j$th treatment for the $i$th unit, $\mu_{ij}$ is a location parameter under the $j$th treatment for the $i$th unit, and $U_{ij}$ is the random effect. For fixed $i$ and $j$, let $E_{ij}$ denote the error vector $(E_{ij1}, \ldots, E_{ijm})$, $E_{ij} = (1/m) \Sigma E_{ijk}$, and $v_{ij}^2 = [1/(m - 1)] \Sigma (E_{ijk} - E_{ij})^2$, the summation being taken over $k$ in these cases. We make the following assumptions:

(A1) $U$, the matrix of random effects, is distributed according to some unspecified np-variate distribution function.

(A2) (i) The vectors $E_{ij}$ ($1 \leq i \leq n$, $1 \leq j \leq p$) are independent.

(ii) For fixed $j$, $E_{ij}$’s have an unknown common multivariate distribution (so that $v_{ij}^2$, $1 \leq i \leq m$ can be regarded as a sample from a common univariate distribution, say $G_j$), and

(iii) The $G_j$’s differ at most by a scale factor; that is, for some unknown distribution $G$ and unknown numbers $\sigma_1, \sigma_2, \ldots, \sigma_p$, we have $G_1(x) = G(x/\sigma_1^2), \ldots, G_p(x) = G(x/\sigma_p^2)$. Moreover, the $p \times p$ diagonal matrix $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is finite and positive definite.

The variance components $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$ in $\Sigma$ represent the within-unit variances for the $p$ treatments, and we are interested in conducting inference with respect to $\Sigma$.

If the investigator is also interested in statistical inference with respect to the mean vector $\mu$, then we could assume a linear or nonlinear structure of the form $\mu_{ij} = f(X_{ij}^T \beta)$, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, where $X_{ij}$ is an $r$-vector of explanatory variables, $\beta$ is an $r$-vector of parameters, and $f$ is a specified function. However, for the purposes of this article, we concentrate on $\Sigma$, so we do not specify any functional form for $\mu$ and regard it as a nuisance parameter.

For balanced designs, the above model is the same as that considered in Chinchilli et al. (1995), except that their assumption (A2) regarding the distributions of the $E_{ij}$’s is much stronger. They assume that for fixed $j$, $E_{ij}$’s have a common multivariate normal or multivariate central $t$-distribution, satisfying some additional requirements.

Like Chinchilli et al. (1995), we also make no assumptions regarding the existence or nonexistence of relationships between $U_{ij}$’s and $E_{ijk}$’s because, as explained in Remark 2.3, such assumptions will severely restrict the applicability of our methodology.

Remark 2.1. If the object is to conduct inference for the mean vector $\mu$ (of the responses) and the associated variance–covariance matrix, then we have to do the following additional work (cf., Huggins, 1993a,b; Jensen, 1982; Rao, Sutradhar, and Kim, 1993; Richardson and Welsh, 1995; Wu, Holt, and Holmes, 1988).

(1) Make assumptions such as symmetry or normality regarding the distributions of $U_{ij}$ and $E_{ijk}$.

(2) Model correlations between pairs of $U_{ij}$’s and also between $U_{ij}$ and $E_{ijk}$.
However, we are not doing all these because we are considering an entirely different problem. As a result, for our model,

(a) $U_{ij}$ and $E_{ijk}$ could have different and possibly highly skewed distributions, and
(b) the correlations between pairs of $U_{ij}$'s and between $U_{ij}$ and $E_{ijk}$ could be much more complicated than those in the papers of the above authors.

We begin with the case of two treatments ($p = 2$). Now the parameter of interest is $\Delta = \sigma^2 / \sigma_1^2$. For $1 \leq i \leq n$ and $1 \leq j \leq 2$, write $\bar{Y}_{ij} = (1/m) \sum Y_{ijk}, s_{ij}^2 = [1/(m - 1)] \sum (Y_{ijk} - Y_{ij})^2$.

For fixed $i (= 1, 2, \ldots, n)$, note that $\bar{Y}_{i1k} = \mu_{i1} + U_{i1} + E_{i1k}$ and $\bar{Y}_{i2} = \mu_{i1} + U_{i1} + \bar{E}_{i1}$. Hence,

$$\begin{align*}
(\bar{Y}_{i1k} - \bar{Y}_{i1}) &= (\mu_{i1} + U_{i1} + E_{i1k}) - (\mu_{i1} + U_{i1} + \bar{E}_{i1}) \\
&= (E_{i1k} - \bar{E}_{i1}).
\end{align*}$$

Therefore, $\Sigma (\bar{Y}_{i1k} - \bar{Y}_{i1})^2 = \Sigma (E_{i1k} - \bar{E}_{i1})^2$, or

$$s_{i1}^2 = \frac{1}{m - 1} \sum (Y_{i1k} - \bar{Y}_{i1})^2 = \frac{1}{m - 1} \sum (E_{i1k} - \bar{E}_{i1})^2 = v_{i1}^2.$$ 

Thus, $s_{i1}^2 = v_{i1}^2$ and similarly $s_{i2}^2 = v_{i2}^2$, $i = 1, 2, \ldots, n$. Although $v_{i1}^2$ and $v_{i2}^2$ are unobservable, $s_{i1}^2$ and $s_{i2}^2$ are observable and so we can work with $s_{i1}^2$ and $s_{i2}^2$, $i = 1, 2, \ldots, n$. Note that $(Y_{111}, \ldots, Y_{11m})$ and $(Y_{121}, \ldots, Y_{12m})$ are, in general, dependent. This is because the random effects $U_{11}$ (corresponding to $Y_{111}$) and $U_{12}$ (corresponding to $Y_{121}$) could be paired and hence strongly correlated. However, $s_{11}^2$ and $s_{12}^2$ are free of the random effects and depend only on the error sets $(E_{111}, E_{112}, \ldots, E_{11m})$ and $(E_{121}, E_{122}, \ldots, E_{12m})$ and hence are independent. Similarly for unit $i$, $s_{i1}^2$ and $s_{i2}^2$ are independent. In fact, $s_{11}^2 = (s_{111}^2, \ldots, s_{1m}^2)$ and $s_{22}^2 = (s_{21}^2, \ldots, s_{2m}^2)$ are two independent samples with positive valued observations.

Therefore, the Wilcoxon ($h_W$) and the sum of squared ranks ($h_{SQ}$) tests can be applied to $(s_{11}^2, s_{12}^2)$, although they cannot be applied to the original responses $Y_{ij}$ because of dependence and the possible presence of negative values. Clearly, $v_{i1}^2 = (v_{111}^2, \ldots, v_{1n1}^2)$ and $v_{i2}^2 = (v_{21}^2, \ldots, v_{2n2}^2)$ come from distributions that (i) have support $(0, \infty)$, (ii) are right skewed, and (iii) differ only by a scale factor. Since $s_{i1}^2 = v_{i1}^2$, (i)–(iii) also hold for $s_{11}^2 = (s_{111}^2, \ldots, s_{1n1}^2)$ and $s_{22}^2 = (s_{21}^2, \ldots, s_{2n2}^2)$. Therefore (in view of the results of paragraph 1, Section 2), for inference regarding $\Delta$, the $h_W$ and $h_{SQ}$ tests (based on $s_{11}^2$ and $s_{22}^2$) are appropriate. We shall describe the estimators and confidence intervals based on these tests later in this section.

Remark 2.2. Note that $h_W$ and $h_{SQ}$ are based on the vectors of sample variances $(s_{11}^2, s_{12}^2)$, which are free of the $U_{ij}$'s. Consequently, the $U_{ij}$'s play no further role in the analysis. This, however, is not serious because, although the $U_{ij}$'s play a useful role in making the model realistic, they contain no information about $\Delta = \sigma^2 / \sigma_1^2$. On the other hand, all the information about $\Delta$ is contained in $(s_{11}^2, s_{12}^2)$. Therefore, elimination of $U_{ij}$'s from the analysis entails little loss of efficiency. Further details are given in Remark 2.3.

The bootstrap. For $j = 1, 2$, let

$$T_j = \sum_i s_{ij}^2, \quad \bar{T}_j = T_j / n, \quad n_j = \frac{n \sum (s_{ij}^2 - \bar{T}_j)^2}{T_j^2},$$

and

$$V = \frac{\sqrt{n} (\bar{T}_1 - \Delta)}{\sqrt{2n_j}}.$$ 

Let $s_{11}^* = (s_{111}^2, \ldots, s_{1n1}^2)$ and $s_{12}^* = (s_{112}^2, \ldots, s_{1n2}^2)$ be bootstrap samples from $s_{11}^2 = (s_{111}^2, \ldots, s_{1n1}^2)$ and $s_{22}^2 = (s_{21}^2, \ldots, s_{2n2}^2)$, respectively. Write

$$V_1^* = \frac{\sqrt{n} (\bar{T}_1^* - \Delta)}{\sqrt{2n_j^*}},$$

where $T_j^*$ and $n_j^*$ are the bootstrap counterparts of $T_j$ and $n_j$. This process is repeated 499 more times to yield $V_1^*, V_2^*, \ldots, V_{500}^*$. Let the corresponding ordered values be $V_{(1)}^* \leq V_{(2)}^* \leq \cdots \leq V_{(500)}^*$. 

Then $V_{(13)}$ and $V_{(488)}$ are estimates of the 2.5% and 97.5% quantiles of $V$, respectively. Using $V_{(13)}$ and $V_{(488)}$, tests of $\Delta$ can be performed and a 95% confidence interval centered at $T_2/T_1$ (now the estimator of $\Delta$) can be constructed.

For completeness, we now briefly describe the procedures None, Normal, and Multivariate $t$ used in Chinchilli et al. (1995). None is a distribution-free procedure based on the order statistics of the ratios $s_{2k}^2/s_{1k}^2$, $1 \leq k \leq m$.

For fixed $i$ and $j$, consider the vector $E_{ij} = (E_{ij1}, \ldots, E_{ijm})$. Normal and Multivariate $t$ are special cases of the partial likelihood analysis given below, when $E_{ij}$ has, respectively, a multivariate normal and multivariate central $t$-distribution, satisfying some additional requirements (cf., Chinchilli et al., 1995, pp. 208, 210).

The partial likelihood analysis consists of (i) expressing the full log-likelihood $L$ (of the combined responses $Y_{ijk}$) as a sum of a conditional log-likelihood $L_{[12]}$, which involves the $U_{ij}$’s and a marginal log-likelihood, say $L_{[2]}$, that is free of the $U_{ij}$’s and (ii) working only with $L_{[2]}$. Chinchilli et al. (1995, p. 207) point out that the resulting loss in efficiency is only slight and, moreover, goes to zero as the sample size $m$ increases. However, even the partial likelihood analysis is rather complicated, as it involves the Nelder–Mead algorithm (Chinchilli et al., 1995, p. 210).

Remark 2.3. Any analysis involving the $U_{ij}$’s should be based on the full log-likelihood $L$ and hence are even more complicated, although its performance is not going to be much better, as pointed out in the preceding paragraph. As a result, its performance for nonnormal distributions will be quite unsatisfactory (since this is true of the partial likelihood procedures Normal and Multivariate $t$). Therefore, in order to achieve gains by including $U_{ij}$ in the analysis, we must impose stringent requirements regarding the distributions and correlations of $U_{ij}$ and $E_{ij}$. However, such gains will be more than nullified from the greatly diminished applicability of the resulting methodology.

When there are more than two treatments ($p > 2$), the $p$-sample versions of the $h_W$ test (which is the Kruskal–Wallis test) and the $h_{SQ}$ test (cf., Hajek and Sidak, 1967, pp. 104–105) are appropriate. It may be possible to develop a $p$-sample version of the bootstrap, but we are not attempting it because, even in the case of two treatments, its performance fell short of expectations (cf., Section 4). The procedure None does not seem to extend easily to the $p$-sample situation. However, the Multivariate $t$ and Normal procedures easily extend to the case of $p$-treatments and also to unbalanced designs (Chinchilli et al., 1995).

Suppose now there are only two treatments. We describe the estimators and confidence intervals based on $h_W$ and $h_{SQ}$. Recall that $s_1^2 = (s_1^{21}, \ldots, s_n^{21})$ and $s_2^2 = (s_1^{22}, \ldots, s_n^{22})$. For $i = 1, 2, \ldots, n$, let $R_i$ be the rank of $s_i^2$ in the pooled sample $(s_1^2, s_2^2)$. Then $h_W = R_1 + \cdots + R_n$, $U = h_W - n(n+1)/2$ is the corresponding Mann–Whitney statistic and $h_{SQ} = R_1^2 + \cdots + R_n^2$. As $h_W$ and $U$ are equivalent, they yield the same estimator and confidence intervals. Let $U_{0.25}$ and $U_{0.975}$ be the 2.5% and 97.5% quantiles of the (null) distribution of $U$. Let $h_{SQ,.025}$ and $h_{SQ,.975}$ have similar meanings in relation to $h_{SQ}$.

Consider the ratios $\{s_{i2}^2/s_{i1}^2; i = 1, 2, \ldots, n\}$. Clearly, there are $n \times n = n^2 = M$ such ratios (although they need not all be distinct). Denote these ratios, rearranged in ascending order by $D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(M)}$. A slight modification of the arguments in the location model shows that the 95% confidence interval corresponding to $U$ (or $h_W$) is $[D_{(U_{0.25})}, D_{(U_{0.975})}]$. The estimator is precisely the Hodges–Lehmann type scale estimator given by the median of the ratios $\{D_{(i)}, i = 1, 2, \ldots, M\}$ (cf., Lehmann, 1975, pp. 82–83.)

Next, let $[\Delta_r, \Delta_s]$ be the confidence interval based on $h_{SQ}$. This is not known in a closed form and has to be computed iteratively (cf., Bauer, 1972). However, even for practical sample sizes, the number of iterations could become somewhat large. Therefore, we suggest the following short cut, which utilizes the fact that $[\Delta_r, \Delta_s]$ is not far away from the Wilcoxon confidence interval and, therefore, starting with the latter, the former can be obtained in a few iterations. For any positive number $p$, set $s_r^2 = s_{r2}^2/p, \ldots, s_s^2 = s_{s2}^2/p$. Start with the upper endpoint of the Wilcoxon interval, namely $D_{(U_{0.975})}$. For some positive integer $k$, write

$$a_1 = D_{(U_{0.975} - k)}, \ldots, a_k = D_{(U_{0.975} - 1 - k)}, \quad a_{k+1} = D_{(U_{0.975})}, \ldots, a_{2k} = D_{(U_{0.975} - 1 + k)}.$$

(Normally, for small to moderate samples, it is enough to take $k = 4$ or $k = 5$.)

Denote the midpoints of $(a_1, a_2), (a_2, a_3), \ldots, (a_{2k}, a_{2k+1})$ by $e_1, e_2, \ldots, e_{2k}$, respectively. Among these $e_i$’s, there will be an $e$, say $e_\ell$, with

$$h_{SQ}(s_1^2, s_2^2/e_\ell) \geq h_{SQ,.025} \quad \text{and} \quad h_{SQ}(s_1^2, s_2^2/e_{\ell+1}) \leq h_{SQ,.975}.$$

Then $\Delta_\ell = a_{\ell+1} - a_\ell$.
Write \( b_1 = D_{(U,025-k)} \), \( b_2 = D_{(U,025+1-k)} \), \ldots, \( b_{2k+1} = D_{(U,025+k)} \). Let \( f_1, f_2, \ldots, f_{2k} \) be the midpoints of \((b_1, b_2), (b_2, b_3), \ldots, (b_{2k}, b_{2k+1})\), respectively. There is an \( f \)-value, say \( f_q \), such that \( h_{SQ}(s^2_1, s^2_2, f_q) \geq h_{SQ},.975 \) and \( h_{SQ}(s^2_1, s^2_2, f_q + 1) \leq h_{SQ},.975 \). Then \( b_{q+1} \) is the required endpoint; that is, \( \Delta_r = b_{q+1} \). Hence, the confidence interval is \([\Delta_r, \Delta_s] = [b_{q+1}, a_{\ell+1}]\).

For estimation, recall that the estimator corresponding to \( h_{SQ} \), say \( \Delta_W \), is simply the Hodges–Lehmann type scale estimator, that is, the median of the \( D \)'s. The estimator corresponding to \( h_{SQ} \), say \( \Delta_{SQ} \), is exactly or very nearly the midpoint of \((b_{q+1}, a_{\ell+1})\). More precisely, \( \Delta_{SQ} \) is characterized as follows: Let \( \mu_{SQ} \) be the (null) expectation of \( h_{SQ} \). Then \( h_{SQ}(s^2_1, s^2_2, \rho) \) is \( \geq \mu_{SQ} \) or \( \leq \mu_{SQ} \) according as \( \rho < \Delta_{SQ} \) or \( \rho > \Delta_{SQ} \). Once again, starting with \( \Delta_W \) as an initial estimate, \( \Delta_{SQ} \) can be computed in a few iterations. The details are omitted, being similar to those of confidence intervals.

3. Unbalanced Designs

We initially extend the methodology to unbalanced designs satisfying Assumption 3 and then show that even this assumption can be dropped.

Assumption 3: For each treatment, \( m_1 \) measurements are made on each of \( n_1 \) subjects (these \( n_1 \) subjects need not be the same for all treatments), \( m_2 \) measurements are made on each of \( n_2 \) subjects (these \( n_2 \) subjects need not be the same for all treatments), \ldots, and finally \( m_\ell \) measurements are made on each of \( n_\ell \) subjects (once again, these \( n_\ell \) subjects need not be the same for all treatments), where \( n_1 + n_2 + \cdots + n_\ell = n \).

When Assumption 3 is satisfied, we can divide the design into \( \ell \) groups, say \( G_1, G_2, \ldots, G_\ell \) such that \( m_i \) repeated measurements are made in \( G_i \) (\( i = 1, 2, \ldots, \ell \)). As a result, each \( G_i \) represents a balanced design and the methodology of Section 2 applies to \( G_i \). Let \( h_i \) be the statistic based on \( G_i \) and \( h = h_1 + h_2 + \cdots + h_\ell \). The \( h_i \)'s are independent and asymptotically normal if \( p = 2 \) and asymptotically Chi-square if \( p > 2 \). Hence, the same is true of \( h \) due to the independence of the \( h_i \)'s. Therefore, tests for \( \Delta \) can be based on \( h \). The \( h \)-test will be reasonably good if each \( n_i \geq 15 \) and quite good if each \( n_i \geq 20 \). This is because simulation studies show that the asymptotic theory enforcing the good properties of \( h_W \) and \( h_{SQ} \) is reasonably effective or quite effective according as the minimum sample size is 15 or 20.

When \( p = 2 \), confidence intervals for \( \Delta \) are found as in Section 2. The only difference is that now we have to consider \( E_1 \cup E_2 \cup \cdots \cup E_\ell \), where \( E_k \) is the set of ratios \( \{s^2_{i',k}/s^2_{i,k}, 1 \leq i, i' \leq n_\ell \} \) for the \( k \)th design \( G_k \) (\( k = 1, 2, \ldots, \ell \)).

Finally, consider the most general unbalanced design, where even Assumption 3 is not satisfied; that is, \( m_1 \) measurements are made on \( n_1 \) subjects for some treatments but not for others, etc. By discarding some measurements, if necessary, we can ensure Assumption 3 is satisfied and apply the corresponding methodology.

4. Monte Carlo Studies

For some \( a > 0 \), let \( N(0,a) \) denote the normal distribution with mean zero and standard deviation \( a \). The following situations were simulated by using random samples based on IMSL subroutines.

Fifty subjects were given two treatments, three measurements being made for each subject and for each treatment, and the error distributions were

(i) \( N(0,6) \),
(ii) the contaminated normal \( .9N(0,6) + .1N(0,18) \),
(iii) the standard exponential, and
(iv) the standard lognormal.

There were 2000 replications. The procedures considered were \( h_W, h_{SQ} \), None, Bootstrap, Normal (which assumed the error distributions were always normal), and Multivariate t (which assumed the error distributions always had a multivariate t-distribution). For each procedure, the following problems were studied. The total number of rejections divided by 2000 gave the empirical levels. Subsequently, the second treatment values were multiplied by 1.2 and the above process was repeated to yield the empirical powers. Finally, the median confidence interval lengths were also computed. The results are displayed in Table 1.

Remark 4.1. As explained in Section 2, \( h_W, h_{SQ} \), Bootstrap, and None were based on the vectors of sample variances, which were free of the random effects \( U_{ij} \). The Normal and Multivariate t procedures were based on the marginal likelihood and thus were also free of \( U_{ij} \) (cf., Section 2). Therefore, there was no need to simulate the \( U_{ij} \)'s.
### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Mult. t</th>
<th>None</th>
<th>Bootstrap</th>
<th>$h_W$</th>
<th>$h_{SQ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normal Distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.0500</td>
<td>0.0345</td>
<td>0.0645</td>
<td>0.0510</td>
<td>0.0485</td>
<td>0.0470</td>
</tr>
<tr>
<td>Power</td>
<td>0.4295</td>
<td>0.3680</td>
<td>0.2885</td>
<td>0.4035</td>
<td>0.3440</td>
<td>0.3845</td>
</tr>
<tr>
<td>Med CI width</td>
<td>0.8120</td>
<td>0.9045</td>
<td>0.7615</td>
<td>0.8286</td>
<td>0.9505</td>
<td>0.9440</td>
</tr>
<tr>
<td><strong>Lognormal Distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.5740</td>
<td>0.0150</td>
<td>0.0670</td>
<td>0.0555</td>
<td>0.0510</td>
<td>0.0520</td>
</tr>
<tr>
<td>Power</td>
<td>0.6185</td>
<td>0.0125</td>
<td>0.1330</td>
<td>0.0870</td>
<td>0.1445</td>
<td>0.1435</td>
</tr>
<tr>
<td>Med CI width</td>
<td>0.8029</td>
<td>9.4074</td>
<td>2.0174</td>
<td>3.2518</td>
<td>1.7456</td>
<td>1.7274</td>
</tr>
<tr>
<td><strong>Contaminated Normal Distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.2350</td>
<td>0.0515</td>
<td>0.0550</td>
<td>0.0515</td>
<td>0.0470</td>
<td>0.0485</td>
</tr>
<tr>
<td>Power</td>
<td>0.4545</td>
<td>0.1445</td>
<td>0.2270</td>
<td>0.2205</td>
<td>0.2655</td>
<td>0.2730</td>
</tr>
<tr>
<td>Med CI width</td>
<td>0.8244</td>
<td>3.8542</td>
<td>1.2812</td>
<td>1.4864</td>
<td>1.1187</td>
<td>1.1106</td>
</tr>
<tr>
<td><strong>Exponential Distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.2445</td>
<td>0.0365</td>
<td>0.0575</td>
<td>0.0545</td>
<td>0.0510</td>
<td>0.0535</td>
</tr>
<tr>
<td>Power</td>
<td>0.4830</td>
<td>0.0575</td>
<td>0.1680</td>
<td>0.1875</td>
<td>0.2045</td>
<td>0.2140</td>
</tr>
<tr>
<td>Med CI width</td>
<td>0.8051</td>
<td>8.9854</td>
<td>1.5738</td>
<td>1.4981</td>
<td>1.3867</td>
<td>1.3739</td>
</tr>
</tbody>
</table>

For the three nonnormal distributions, the Normal procedure had excessively high empirical levels, while the Multivariate $t$ had very low empirical levels, resulting in drastically reduced powers. None had relatively low powers for the normal and lognormal distributions (especially taking into account its somewhat inflated empirical levels). Bootstrap gave a somewhat unsatisfactory performance for the lognormal distribution, especially in terms of median confidence interval width. Thus, only the $h_W$ and $h_{SQ}$ procedures had robustness of level as well as robustness of efficiency. While these two performed similarly for the three nonnormal distributions, the $h_{SQ}$ procedure was well ahead for the normal distribution, making it the best all-round performer, which is consistent with theory (cf., Section 2).

### 5. An Illustration

Consider the data given in Section 5 of Chinchilli et al. (1995). These are based on the trial conducted at Virginia Commonwealth University on 100 subjects in which serum cholesterol was measured 10 times on each subject with a routine laboratory analyzer (Ektachem 700, the test procedure, say $T$) and 10 times on each subject with a Center for Disease Control standardized method (Cobas Bio, the reference procedure). For convenience, we use the notation $T$ (test) and $R$ (reference) to represent the two assays so that $\Delta = \sigma_T^2/\sigma_R^2$. The 95% confidence bounds for $\Delta$ based on $h_W$ and $h_{SQ}$ are respectively 2.376 and 2.210. Note that both these bounds are sharper than the bound 2.5893 given by None (Chinchilli et al., 1995, Table 2, p. 216), with the sharpest bound given by $h_{SQ}$. As the Normal and Multivariate $t$ procedures are not robust, we have not considered the corresponding bounds given in Chinchilli et al. (1995). Both $h_W$ and $h_{SQ}$ yield the same estimator, namely 1.857. Finally, the 95% confidence intervals based on $h_W$ and $h_{SQ}$ are, respectively, [1.401, 2.482] and [1.404, 2.414]. Because both intervals exclude 1, the null hypothesis is rejected, suggesting that the variance of the test method is significantly higher than that of the reference method. This is consistent with the findings of Chinchilli et al. (1995).

**Remark 5.1.** Note that the above example refers not to the original measurements, which were highly dependent, but only to the vectors of the corresponding sample variances, which were independent and positive valued, as shown in Section 2. (The original data, being classified, could not be submitted for publication.) However, this did not entail any loss of efficiency, as our procedures were based only on sample variances.

### Acknowledgements

We are deeply grateful to the associate editor and the referees for valuable suggestions and comments. Research of G.J.B. is supported in part by NSA grant MDA904-97-1-0023 and NSF grant DMS-9626189.
Résumé

Les objectifs de certains essais sont de comparer les variances entre les unités de deux ou plusieurs traitements, produits ou techniques. Dans ce cas un essai en "mesures répétées" avec un modèle à effet aléatoire et éventuellement des variances hétérogènes est approprié. Avec l'hypothèse que les erreurs aléatoires ont une distribution normale ou 1 multidimensionnelle l'essai a été analysé par Chinchilli, Eshinart, et Miller (1995, Biometrics 51, 215–216). Cependant cette analyse reste très sensible au biais et aux données aberrantes. Nous proposons deux méthodes non-paramétriques robustes dans les essais équilibrés lorsque le nombre de mesures répétées est le même pour toutes les unités et pour tous les traitements. Nous montrons ensuite comment adapter ces procédures afin de traiter les essais non équilibrés. Nous illustrons le principe avec un exemple d'essai de comparaison de mesures de cholestérol entre une méthode habituelle de laboratoire d'analyse et une méthode standardisée.

Références


Received December 1996; accepted May 1997.