Biometrics 59, 237–247
June 2003

# Inferences on the Common Mean of Several Normal Populations Based on the Generalized Variable Method

**K. Krishnamoorthy* and Yong Lu**

Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana 70504, U.S.A.
**email:* krishna@louisiana.edu

Summary. This article presents procedures for hypothesis testing and interval estimation of the common mean of several normal populations. The methods are based on the concepts of generalized p-value and generalized confidence limit. The merits of the proposed methods are evaluated numerically and compared with those of the existing methods. Numerical studies show that the new procedures are accurate and perform better than the existing methods when the sample sizes are moderate and the number of populations is four or less. If the number of populations is five or more, then the generalized variable method performs much better than the existing methods regardless of the sample sizes. The generalized variable method and other existing methods are illustrated using two examples.

Key words: Combined estimator; Combined test; Fisher's test; Generalized confidence interval; Generalized p-value; Power.

## 1. Introduction

The statistical analysis that combines or integrates the results of several independent studies is known as *meta-analysis,* and it is commonly used in clinical trials and social and behavioral sciences. If independent samples are collected from different normal populations with a common mean but possibly with different variances, then the problem of interest is to combine the summary statistics of the samples to estimate or test the common mean. This problem arises in situations where different instruments, different methods, or different laboratories are used to measure like substances or products to assess their average quality. The article by Meier (1953) presents an example in which four different methods are used to estimate the mean percentage of albumin in the plasma protein of normal people (see Example 1 in Section 7). Eberhardt, Reeve, and Spiegelman (1989) illustrated an example where different analytical methods are used to estimate a chemical substance in nonfat milk powder (see Example 2 of Section 7).

To formulate the present problem, consider $k$ different normal populations with a common mean $\mu$ but possibly different variances $\sigma_1^2, \ldots, \sigma_k^2$. Let $\bar{X}_i$ and $S_i^2$ denote, respectively, the mean and variance of a random sample of $n_i$ observations from the $i$th normal population, $i = 1, \ldots, k$. An extensive number of papers have been written on the point estimation of $\mu$, and a landmark result is due to Graybill and Deal (1959), who first showed for the two-sample case that the combined estimator

$$\hat{\mu}_{GD} = \frac{\frac{n_1}{S_1^2}\bar{X}_1 + \frac{n_2}{S_2^2}\bar{X}_2}{\frac{n_1}{S_1^2} + \frac{n_2}{S_2^2}} \tag{1}$$

has a smaller variance than either of the sample mean, provided $n_1 \geq 11$ and $n_2 \geq 11$. Since then, many authors improved and extended this result to the case of more than two populations. For a good exposition of the work in point estimation, we refer to Bhattacharya (1984) and the references therein.

Several authors proposed approximate confidence intervals which are centered at $\hat{\mu}_{GD}$. Meier (1953) proposed a normal-based approximate confidence interval centered at $\hat{\mu}_{GD}$. Using Welch's (1947) method, Maric and Graybill (1979) and Pagurova and Gurskii (1979) derived approximate confidence intervals which are also centered at $\hat{\mu}_{GD}$. Another approximate confidence interval based on the likelihood ratio test is proposed in Hinkley (1979). The article by Fairweather (1972) seems to be the first work that provides an exact confidence interval for $\mu$. This exact confidence interval is based on a linear combination of individual $t$-test statistics. Following the idea of Fairweather (1972), Jordan and Krishnamoorthy (1996) developed an exact confidence interval centered at $\hat{\mu}_{GD}$. This interval is obtained by inverting a linear combination of individual $F$-test statistics. The results are generalized to a multivariate normal case in Jordan and Krishnamoorthy (1995a). Jordan and Krishnamoorthy (1996) compared these two exact confidence intervals along with two other exact confidence intervals. Their numerical studies showed that Fairweather's intervals and Jordan and Krishnamoorthy's intervals are in general shorter than the other two intervals. For smaller variance ratios, Fairweather's intervals are narrower than Jordan and Krishnamoorthy's intervals; otherwise the latter intervals are shorter than the former intervals. Yu, Sun, and Sinha (1999) considered the four confidence intervals in Jordan and Krishnamoorthy (1996) and some other intervals that are obtained by inverting the tests based on different combinations of the p-values from individual tests (for example, see Fisher's test in Section 6 of this paper). Based on their numerical studies, they recommended the intervals due to Fairweather (1972), Jordan and Krishnamoorthy (1996) and the one based on Fisher's test in Section 6. It should be noted that all the methods (except Fairweather's) considered in Yu et al. (1999) do not always produce nonempty confidence intervals. They required satisfying some conditions in order to yield nonempty intervals.

Regarding hypothesis testing about the common mean $\mu$, Cohen and Sackrowitz (1984) considered several tests, including a normal approximate test based on $\hat{\mu}_{GD}$ in (1). Their power comparison studies showed that there is no clear-cut winner among the tests considered. In particular, the variance ratio must be known to choose the best test. In the context of the balanced incomplete block design, Cohen and Sackrowitz (1989) proposed a test that combines the individual tests by weighting with their sample variances. Following their idea, Zhou and Mathew (1993) developed two combined tests for the present problem, and compared them with Fisher's test. Their power comparison studies, for the case of $k = 2$, showed that one of the new tests is better than Fisher's test; however, knowledge about the magnitude of the variance ratio must be known to choose the better of the two new tests. Even though some of these combined tests are easy to use and better than Fisher's test and other approximate tests, it is difficult to invert them to obtain confidence intervals for the common mean.

We see from the above literature review that there is no single procedure that performs better than the other methods for both hypothesis testing and interval estimation. The main purpose of this article is to develop a single *generalized pivot variable* that is simple to use for both hypothesis testing and interval estimation of $\mu$. Toward this, we want to develop inferential procedures based on the generalized p-value approach. The concept of generalized p-value was introduced by Tsui and Weerahandi (1989) for hypothesis testing. Weerahandi (1993) extended the idea for constructing confidence intervals. The book by Weerahandi (1995b) gives a detailed discussion along with numerous examples. The concepts of generalized p-values and generalized confidence intervals have turned out to be very satisfactory for obtaining tests and confidence intervals for many complex problems; see Zhou and Mathew (1994), Weerahandi (1995a), and Weerahandi and Berger (1999). For these reasons, we want to construct a pivot variable using the idea of a generalized p-value approach, so that it can be used for both hypothesis testing and interval estimation of the common mean.

This article is organized as follows. In the following section, we briefly explain the concept of the generalized p-value and generalized limits in a typical setup, and then construct a generalized variable for the present problems. This generalized variable is obtained by combining the generalized variables of the means based on individual samples, using the reciprocals of the generalized variables of the variances as weights. In Section 3, we outline the methods of constructing confidence intervals and hypothesis testing about the common mean $\mu$ using the generalized variable given in Section 2. For this purpose, a computational algorithm is provided. In Section 4, we describe the methods of computing powers of the test and coverage probabilities of the confidence interval based on the generalized variable. A computational algorithm for computing powers of the generalized test and coverage probabilities of the generalized confidence interval is also given. In Section 5, the generalized confidence interval is compared to confidence intervals (i) due to Fairweather (1972) and (ii) due to Jordan and Krishnamoorthy (1996). Of the two intervals (i) and (ii), there is no clear-cut winner. Comparison among these three intervals show that the expected lengths of the generalized intervals are close to the minimums of the expected lengths of (i) and (ii). If the variances are drastically different, then the generalized interval has the shortest expected length. In Section 6, we present some power studies. Powers of the generalized test are compared with those of two combined tests in Zhou and Mathew (1993), an approximate test due to Mathew, Sinha, and Zhou (1993), and Fisher's test. When $k = 2$ and the sample sizes are moderate (at least nine), the powers of the generalized test are very close to those of the test that has the maximum power among the other three tests and in some situations, the generalized test has the largest power. For $k \geq 3$, the sample size requirement is somewhat relaxed. For example, when $k = 3$, the generalized test seems to be the most powerful test of all the tests considered when the sample sizes are five or more. For moderate $k$, our power comparison studies show that the generalized test is the best when sample sizes are four or more. In Section 7, the generalized variable methods and other methods considered in this article are illustrated using two examples. Some concluding remarks are given in Section 8.

## 2. Generalized Variable for the Common Mean

### 2.1 *The Generalized P-Value and Confidence Limits*

We shall first explain the method of constructing a generalized pivot variable and the definition of the generalized p-value in a general setup. Let $X$ be a random variable (that could be a vector) whose distribution depends on the parameters $(\theta, \delta)$, where $\theta$ is a scalar parameter of interest, and $\delta$ is a nuisance parameter. Suppose we are interested in testing the hypotheses

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_a : \theta > \theta_0, \tag{2}$$

where $\theta_0$ is a specified quantity. Let $x$ denote the observed value of $X$. In other words, $x$ is known after the data have been collected. A generalized pivot variable, to be denoted by $T_1(X;\, x,\, \theta,\, \delta)$, is a function of $X$, $x$, $\theta$, and $\delta$, and it satisfies the following conditions:

(i) For a fixed $x$, the distribution of $T_1(X; x, \theta, \delta)$ is free of the nuisance parameter $\delta$.

(ii) The value of $T_1(X; x, \theta, \delta)$ at $X = x$ is free of any unknown parameters.

(iii) For fixed $x$ and $\delta$, the distribution of $T_1(X; x, \theta, \delta)$ is either stochastically increasing or stochastically decreasing in $\theta$. That is, $P\{T_1(X; x, \theta, \delta) \geq a\}$ is an increasing function of $\theta$, or is a decreasing function of $\theta$, for every $a$. (3)

Let $t_1 = T_1(x; x, \theta, \delta)$, the value of $T_1(X; x, \theta, \delta)$ at $X = x$. If $T_1(X; x, \theta, \delta)$ is stochastically increasing in $\theta$, the generalized p-value for testing the hypotheses in (2) is given by

$$\sup_{H_0} P\{T_1(X; x, \theta, \delta) \geq t_1\} = P\{T_1(X; x, \theta_0, \delta) \geq t_1\},$$

and if $T_1(X; x, \theta, \delta)$ is stochastically decreasing in $\theta$, the generalized p-value for testing the hypotheses in (2) is given by

$$\sup_{H_0} P\{T_1(X; x, \theta, \delta) \leq t_1\} = P\{T_1(X; x, \theta_0, \delta) \leq t_1\}.$$

Note that the computation of the generalized p-value is possible in view of the conditions (i) and (ii) in (3), i.e., the distribution of $T_1(X; x, \theta, \delta)$ is free of the nuisance parameter $\delta$ and $t_1 = T_1(x; x, \theta, \delta)$ is free of any unknown parameters.

A generalized confidence interval for $\theta$ is computed using the percentiles of a generalized pivot variable, say $T_2(X; x, \theta, \delta)$, satisfying the following conditions:

(i) For a fixed $x$, the distribution of $T_2(X; x, \theta, \delta)$ is free of all unknown parameters.
(ii) The value of $T_2(X; x, \theta, \delta)$ at $X = x$ is $\theta$, the parameter of interest. (4)

Appropriate quantiles of $T_2$ form a $(1-\alpha)$ confidence limit for $\theta$. For example, if $T_2(x; p)$ is the $p$th quantile of $T_2(X; x, \theta, \delta)$, then $(T_2(x; \alpha/2), T_2(x; 1-\alpha/2))$ is a 95% confidence interval for $\theta$.

For further details on the concepts of generalized p-values and generalized confidence intervals, along with numerous examples, we refer the reader to the book by Weerahandi (1995b).

### 2.2 *The Generalized Variable for Testing and Interval Estimation of the Common Mean*

Let $\bar{x}_i$ and $s_i^2$ denote respectively the observed values of $\bar{X}_i$ and $S_i^2$, $i = 1, \ldots, k$. Further, let $v_i^2 = (n_i - 1)s_i^2$ and $V_i^2 = (n_i - 1)S_i^2$, $i = 1, \ldots, k$. The generalized pivot variable for estimating $\mu$ based on the $i$th sample is given by

$$T_i = \bar{x}_i - \left(\frac{\bar{X}_i - \mu}{\sigma_i/\sqrt{n_i}}\right)\frac{\sigma_i}{V_i}\frac{v_i}{\sqrt{n_i}} = \bar{x}_i - \frac{Z_i}{U_i}\frac{v_i}{\sqrt{n_i}}, \quad i = 1, \ldots, k, \tag{5}$$

where $U_i^2 = V_i^2/\sigma_i^2$ and

$$Z_i = (\bar{X} - \mu_i)/(\sigma_i/\sqrt{n_i}),$$

$i = 1, \ldots, k$. Notice that $U_i^2$'s and $Z_i$'s are independent random variables with $U_i^2 \sim \chi^2_{n_i - 1}$ and $Z_i \sim N(0, 1)$, $i = 1, \ldots, k$. Furthermore, note that

$$T_i = \bar{x}_i - t_i s_i/\sqrt{n_i},$$

where

$$t_i = Z_i\sqrt{n_i - 1}/U_i$$

follows a student's $t$-distribution with df $= n_i - 1$, $i = 1, \ldots, k$. For a given $\bar{x}_i$ and $s_i$, let $T_{i,p}$ denote the the $p$th quantile of $T_i$, $i = 1, \ldots, k$. Then, the $1-\alpha$ confidence interval based on $T_i$ is $(T_{i,\alpha/2}, T_{i,1-\alpha/2})$ and it coincides with the usual $t$-interval based on the $i$th sample alone, $i = 1, \ldots, k$.

The generalized pivot variable for estimating $\sigma_i^2$ based on the $i$th sample is given by

$$R_i = \frac{\sigma_i^2}{V_i^2}v_i^2 = \frac{v_i^2}{Q_i^2}, \quad i = 1, \ldots, k, \tag{6}$$

where $Q_i^2 = V_i^2/\sigma_i^2$ are independent $\chi^2_{n_i-1}$ random variables, $i = 1, \ldots, k$. For a given $v_i^2$, let $R_{i,p}$ denote the $p$th quantile of $R_i$, $i = 1, \ldots, k$. Then $(R_{i,\alpha/2}, R_{i,1-\alpha/2})$ is an exact $(1-\alpha)$ confidence interval for $\sigma_i^2$ (based on the $i$th sample alone) and it coincides with the usual confidence interval based on the chi-squared distribution.

The generalized variable that we propose is a weighted average of the generalized variables $T_i$'s of $\mu$ based on individual samples. The weights are inversely proportional to the generalized variables $R_i$'s for the variances and are directly proportional to the sample sizes. Let $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_k)$ and $\mathbf{V} = (V_1, \ldots, V_k)$, and let $\bar{\mathbf{x}}$ and $\mathbf{v}$ denote respectively the observed values of $\bar{\mathbf{X}}$ and $\mathbf{V}$. Then the generalized variable can be expressed as

$$T(\bar{\mathbf{X}}, \mathbf{V}; \bar{\mathbf{x}}, \mathbf{v}) = \frac{\sum_{i=1}^{k} \frac{n_i Q_i^2}{v_i^2}\left(\bar{x}_i - \frac{Z_i}{U_i}\frac{v_i}{\sqrt{n_i}}\right)}{\sum_{j=1}^{k} \frac{n_j Q_j^2}{v_j^2}} = \sum_{i=1}^{k} W_i T_i, \tag{7}$$

where $W_i = (n_i Q_i^2/v_i^2)/(\sum_{j=1}^{k} n_j Q_j^2/v_j^2)$, $i = 1, \ldots, k$, and $T_i$ is given in (5).

To construct confidence limits based on $T$, we need to verify that $T$ in (7) satisfies the two conditions in (4). The value of $T_i$ in (5) at $(\bar{X}_i, V_i) = (\bar{x}_i, v_i)$ is $\mu$, $i = 1, \ldots, k$. It follows from (6) that $W_i = (n_i/\sigma_i^2)/(\sum_{j=1}^{k} n_j/\sigma_j^2)$ when $V_i = v_i$, $i = 1, \ldots, k$. Therefore, $T = \mu$ at $(\bar{\mathbf{X}}, \mathbf{V}) = (\bar{\mathbf{x}}, \mathbf{v})$. It is also clear from (5) and (6) that, for a given $(\bar{\mathbf{x}}, \mathbf{v})$, the distribution of $T$ is independent of any unknown parameters. Therefore, $T$ is a bona fide generalized variable, and its quantiles can be used to construct confidence limits for $\mu$.

For hypothesis testing about $\mu$, a generalized variable, say $T^*$, can be defined as $T^* = T - \mu$, where $T$ is defined in (7). We see that, for a given $(\bar{\mathbf{x}}, \mathbf{v})$ and $t$, $P(T > t) = P(T^* > t - \mu)$, and so $T$ is stochastically decreasing in $\mu$.

Remark 2.1 In the definitions of $T_i$ in (5) and $R_i$ in (6), we used different chi-squared random variables $U_i^2$ and $Q_i^2$, even though both are related to the same sample sums of squares $\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$. We used different chi-squared variables because our preliminary numerical studies showed that the generalized variable based on the same chi-squared variable (that is, the one with $R_i = v_i^2/U_i^2$) produced confidence limits that are too liberal. In some situations, it yielded confidence limits with coverage probabilities less than 0.88 when the nominal level is 0.95. As will be seen later, the confidence limits based on $T$ in (7) are either slightly conservative or almost exact.

## 3. Hypothesis Testing and Interval Estimation

For a given $(\bar{\mathbf{x}}, \mathbf{v})$, the distribution of $T$ in (7) is independent of any unknown parameters and hence the Monte Carlo method can be used to test the interval estimation of $\mu$. Because $T$ is stochastically decreasing with respect to $\mu$, the generalized p-value for testing

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_a : \mu > \mu_0 \tag{8}$$

is given by

$$P(T < \mu_0 \,|\, \bar{\mathbf{x}}, \mathbf{v}). \tag{9}$$

The generalized test rejects $H_0$ in (8) whenever the p-value in (9) is less than $\alpha$. If one is interested in testing

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0, \tag{10}$$

then the generalized p-value is given by

$$2 \min\{P(T < \mu_0 \,|\, \bar{\mathbf{x}}, \mathbf{v}), P(T > \mu_0 \,|\, \bar{\mathbf{x}}, \mathbf{v})\}. \tag{11}$$

The null hypothesis in (10) will be rejected whenever the p-value in (11) is less than $\alpha$.

The $(1-\alpha)$ confidence interval is given by

$$(T_{\alpha/2}, T_{1-\alpha/2}), \tag{12}$$

where $T_p$ denotes the $p$th quantile of $T$. One-sided confidence intervals are similarly constructed.

The following algorithm is useful in constructing a test and confidence interval for $\mu$.

### Algorithm 1

For a given $(n_1, \ldots, n_k)$, $(\bar{x}_1, \ldots, \bar{x}_k)$, and $(v_1, \ldots, v_k)$:

For $j = 1, m$:
  Generate $t_{n_1-1}, \ldots, t_{n_k-1}$.
  Generate $Q_l^2 \sim \chi^2_{n_l-1}$, $l = 1, \ldots, k$.
  Compute $W_1, \ldots, W_k$.
  Compute $T_j = \sum_{l=1}^k W_l(\bar{x}_l - t_{n_l-1} \frac{s_l}{\sqrt{n_l}})$.
(end $j$ loop)

Compute the $100(\alpha/2)$th percentile $T_{\alpha/2}$ and the $100(1 - \alpha/2)$th percentile $T_{1-\alpha/2}$ of $T_1, \ldots, T_m$. Then, $(T_{\alpha/2}, T_{1-\alpha/2})$ is a $1 - \alpha$ confidence interval for $\mu$. The generalized p-value for testing (8) is estimated by the proportion of $T_i$'s which are less than $\mu_0$.

## 4. Computation of Coverage Probabilities and Powers

Even though, for a given $(\bar{\mathbf{x}}, \mathbf{v})$, the distribution of the $T$ in (7) does not depend on any parameters, the coverage probabilities of the generalized limits depend on the sampling distributions of $\bar{\mathbf{X}}$ and $\mathbf{V}$, which depend on the parameters $\sigma_i^2$'s. Hence the properties of the generalized limits and generalized test are to be evaluated numerically. We computed the coverage probabilities, expected lengths of the generalized confidence limits, and powers using the Monte Carlo method given in the following algorithm.

### Algorithm 2

For a given $(n_1, \ldots, n_k)$, $(\sigma_1^2, \ldots, \sigma_k^2)$, and $\mu$:

For $i = 1, n$:
  Generate $Z_1, \ldots, Z_k$ from $N(0, 1)$.
  Generate $\chi^2_{n_1-1}, \ldots, \chi^2_{n_k-1}$.
  Set $\bar{x}_j = \mu + Z_j\sigma_j/\sqrt{n_j}$, $j = 1, \ldots, k$. $v_j^2 = {\sigma_j}^2\chi^2_{n_j-1}$, $j = 1, \ldots, k$.
  Use Algorithm 1 to construct a $(1-\alpha)$ confidence interval $E_i$. If $E_i$ contains $\mu$, set $\delta_i = 1$; otherwise, set $\delta_i = 0$.
  Use Algorithm 1 to compute the generalized p-value. If this p-value is less than $\alpha$, set $\eta_i = 1$; otherwise, set $\eta_i = 0$.
(end $i$ loop)

The proportion $\frac{1}{n}\sum_{i=1}^n \delta_i$ is the estimated coverage probability of the generalized confidence interval. The proportion $\frac{1}{n}\sum_{i=1}^n \eta_i$ is the estimated power of the generalized test. The expected length of the generalized limits is estimated by the average length of $E_i$'s.

## 5. Comparison between Confidence Intervals

We shall now present the other confidence intervals that are to be compared in our numerical studies.

5.0.1 *The interval estimate due to Fairweather* (1972). Let $t_i$ denote the student's $t$ variable with df $= n_i - 1$, and $u_i = \{\mathrm{Var}(t_i)\}^{-1}/[\sum_{j=1}^k \{\mathrm{Var}(t_j)\}^{-1}]$, $i = 1, \ldots, k$. Further, let $b$ denote the $(1-\alpha)$th quantile of $|\sum_{j=1}^k u_i t_i|$. Then

$$\frac{\sum_{i=1}^k u_i \bar{x}_i \sqrt{n_i}/s_i}{\sum_{j=1}^k u_j \sqrt{n_j}/s_j} \pm \frac{b}{\sum_{j=1}^k u_j \sqrt{n_j}/s_j} \tag{13}$$

is an exact $(1 - \alpha)$ confidence interval for $\mu$. For the case of $k = 2$, the critical point $b$ can be obtained using numerical integration involving the probability density function of a $t$-random variable. For $k \geq 2$, Fairweather (1972) suggested a moment approximation.

5.0.2 *The confidence interval due to Jordan and Krishnamoorthy* (1996). Let $F_i$ denote the $F$ random variable with numerator df $= 1$ and the denominator df $= n_i - 1$, $i = 1, \ldots, k$, and let $a$ denote the $(1 - \alpha)$th quantile of $\sum_{i=1}^k w_i F_i$, where $w_i = \{\mathrm{Var}(F_i)\}^{-1}/[\sum_{j=1}^k \{\mathrm{Var}(F_j)\}^{-1}]$, $i = 1, \ldots, k$. Then

$$\sum_{i=1}^k p_i \bar{x}_i \pm \frac{a}{\sum_{j=1}^k \frac{w_j n_j}{s_j^2}} - \left[\left(\sum_{j=1}^k p_j \bar{x}_j^2 - \left(\sum_{j=1}^k p_j \bar{x}_j\right)^2\right)\right]^{1/2} \tag{14}$$

is an exact $(1 - \alpha)$ confidence interval for the common mean $\mu$. Again, for the case of $k = 2$, exact values of $a$ can be obtained using numerical integration involving the probability density function of an $F$ random variable; for other cases, Jordan and Krishnamoorthy (1996) proposed a moment approximation following the idea of Fairweather (1972).

**Table 1(a)**
*Simulated expected widths of 95% confidence intervals*

| | $n_1 = n_2 = 11$ | | | $n_1 = n_2 = 16$ | | |
|---|---|---|---|---|---|---|
| $(\sigma_1^2,\, \sigma_2^2)$ | 1 | 2 | 3 | 1 | 2 | 3 |
| (5,5) | 2.04 | 2.30 | 2.14(0.95) | 1.60 | 1.85 | 1.70(0.95) |
| (5,10) | 2.36 | 2.68 | 2.47(0.96) | 1.88 | 2.14 | 1.97(0.96) |
| (5,15) | 2.57 | 2.87 | 2.63(0.96) | 2.04 | 2.28 | 2.08(0.95) |
| (5,20) | 2.73 | 2.95 | 2.73(0.96) | 2.14 | 2.36 | 2.14(0.95) |
| (5,25) | 2.80 | 3.03 | 2.77(0.95) | 2.23 | 2.41 | 2.18(0.95) |
| (5,30) | 2.85 | 2.98 | 2.82(0.95) | 2.29 | 2.45 | 2.22(0.95) |
| (5,40) | 3.00 | 3.10 | 2.86(0.95) | 2.38 | 2.50 | 2.25(0.95) |
| (5,50) | 3.12 | 3.14 | 2.87(0.95) | 2.45 | 2.52 | 2.27(0.95) |
| (5,100) | 3.36 | 3.25 | 2.93(0.95) | 2.64 | 2.59 | 2.32(0.95) |
| (5,250) | 3.60 | 3.33 | 2.95(0.95) | 2.85 | 2.63 | 2.36(0.95) |
| (5,500) | 3.74 | 3.34 | 2.93(0.95) | 2.96 | 2.64 | 2.34(0.95) |
| (5,1000) | 3.86 | 3.36 | 2.92(0.95) | 3.04 | 2.65 | 2.35(0.95) |

**Table 1(b)**
*Simulated expected widths of 95% confidence intervals*

| | $n_1 = 31,\, n_2 = 11$ | | | $n_1 = 11,\, n_2 = 31$ | | |
|---|---|---|---|---|---|---|
| $(\sigma_1^2,\, \sigma_2^2)$ | 1[a] | 2[b] | 3[c] | 1[a] | 2[b] | 3[c] |
| (5,5) | 1.42 | 1.55 | 1.46(0.95) | 1.62 | 1.83 | 1.46(0.95) |
| (5,10) | 1.58 | 1.63 | 1.55(0.95) | 1.97 | 2.30 | 1.86(0.95) |
| (5,15) | 1.66 | 1.66 | 1.59(0.95) | 2.18 | 2.56 | 2.09(0.95) |
| (5,20) | 1.72 | 1.68 | 1.61(0.95) | 2.33 | 2.72 | 2.24(0.95) |
| (5,25) | 1.76 | 1.69 | 1.61(0.95) | 2.84 | 2.85 | 2.34(0.95) |
| (5,30) | 1.79 | 1.69 | 1.61(0.95) | 2.53 | 2.95 | 2.43(0.95) |
| (5,40) | 1.83 | 1.70 | 1.62(0.95) | 2.67 | 3.07 | 2.54(0.95) |
| (5,50) | 1.87 | 1.71 | 1.62(0.95) | 2.78 | 3.16 | 2.62(0.95) |
| (5,100) | 1.95 | 1.71 | 1.63(0.95) | 3.08 | 3.36 | 2.77(0.95) |
| (5,250) | 2.03 | 1.72 | 1.63(0.95) | 3.41 | 3.49 | 2.88(0.95) |
| (5,500) | 2.08 | 1.73 | 1.64(0.96) | 3.61 | 3.56 | 2.92(0.95) |
| (5,1000) | 2.11 | 1.73 | 1.64(0.95) | 3.77 | 3.57 | 2.92(0.95) |

[a]1—the interval (13) due to Fairweather (1972).
[b]2—the interval (14) due to Jordan and Krishnamoorthy (1996).
[c]3—the generalized interval (12)

In simulation studies, we used Algorithm 2 with $n = 2500$ runs and Algorithm 1 with $m = 5000$ runs to estimate the coverage probabilities of the confidence intervals. Because the sampling distributions of $\bar{x}_i$'s are location invariant, without loss of generality we can let $\mu = 0$ in our numerical studies. The estimated coverage probabilities of the generalized confidence interval along with their expected lengths are given in Tables 1(a) and 1(b) for various sample sizes and parameter configurations considered in Jordan and Krishnamoorthy (1996). We also presented expected lengths of the exact confidence intervals in (13) and (14). It is clear from these table values that the generalized variable method produced limits that are either very close to the shorter of the intervals (13) and (14) or are the shortest among the three intervals.

## 6. Power Comparison with Other Combined Tests

The combined tests given in Zhou and Mathew (1993) are obtained by combining the p-values of the usual tests based on the individual samples. As already pointed out, these tests are exact; however, it is difficult to invert them to find confidence limits for $\mu$. We shall now briefly describe the combined tests. Note that, for hypotheses in (8), the p-value of the usual $t$-test based on the $i$th sample is given by $P(t^2_{n_i-1} > t^2_{0i})$, where $t_m$ denotes the student's $t$ variable with df $= m$, and $t^2_{0i}$ is an observed value of $n_i(\bar{X}_i - \mu_0)^2/S_i^2$, $i = 1, \ldots, k$. Let $P_i = -\ln P(t^2_{n_i-1} > t^2_{0i})$, $i = 1, \ldots, k$. The weight of the $i$th test is given by

$$a_i = \frac{n_i(n_i+1)\left\{n_i(\bar{x}_i - \mu_0)^2 + v_i^2\right\}^{-1}}{\sum_{j=1}^{k}(n_j+1)\left\{n_j(\bar{x}_j - \mu_0)^2 + v_j^2\right\}^{-1}}, \quad i = 1, \ldots, k. \tag{15}$$

Note that the weights given below the display (3.1) of Zhou and Mathew (1993) are only for equal sample sizes. For unequal sample sizes, the correct weights are the $a_i$'s given in (15); for more details, see Jordan and Krishnamoorthy (1995b). Zhou and Mathew (1993) present an exact expression for computing the p-value of their combined test. Their test rejects the null hypothesis in (8) whenever the p-value

$$\sum_{i=1}^{k} \frac{a_i^{k-1} e^{-P_0/a_i}}{\prod_{j=1^k; j\neq i} (a_i - a_j)} \le \alpha(1+\eta), \tag{16}$$

where

$$P = \sum_{i=1}^{k} a_i P_i, \quad \eta = \frac{\sum_{i<j} \operatorname{sign}\{(\bar{x}_i - \mu_0)(\bar{x}_j - \mu_0)\}}{k(k-1)/2}, \tag{17}$$

and $P_0$ is an observed value of $P$. The derivation of the p-value in the left-hand side of (16) utilizes the fact that the test statistics $t_i^2$'s and the weights are statistically independent. Therefore, the test based on (16) is applicable only for two-sided alternative hypothesis (see Zhou and Mathew (1993) for more details).

Fisher's test is based on the fact that $2\sum_{i=1}^k P_i \sim \chi^2_{2k}$, and it rejects the null hypothesis in (8) whenever

$$P\left(\chi^2_{2k} > c\right) < \alpha, \tag{18}$$

where $c$ is an observed value of $2\sum_{i=1}^k P_i$.

Finally, we consider the approximate test due to Mathew et al. (1993). Let $B_i = n_i(\bar{x}_i - \mu_0)^2/\{n_i(\bar{x}_i - \mu_0)^2 + v_i^2\}$, $i = 1,\, 2$. The test statistic is given by

$$T^* = \frac{n_1 B_1 + n_2 B_2 \hat{\theta}^2 + 2\hat{\theta}\eta\sqrt{n_1 n_2 B_1 B_2}}{(\sqrt{n_1} + \hat{\theta}\sqrt{n_2})^2},$$

where $\hat{\theta}^2 = (n_2^2/n_1^2)\{n_1(\bar{x}_1 - \mu_0)^2 + v_1^2\}/\{n_2(\bar{x}_2 - \mu_0)^2 + v_2^2\}$. It has been shown that $T^*$ follows a beta$(a,\, b)$ distribution approximately. The expressions for $a$ and $b$ are given by $a = E_1(E_1 - E_2)/(E_2 - E_1^2)$ and $b = (1 - E_1)a/E_1$, where

$$E_1 = E(T^* \mid \hat{\theta}) = \frac{1 + \hat{\theta}^2}{(\sqrt{n_1} + \hat{\theta}\sqrt{n_2})^2}$$

and

$$E_2 = E(T^{*2} \mid \hat{\theta}) = \frac{\frac{3}{2}\left(\frac{n_1}{n_1/2+1} + \frac{\hat{\theta}^4 n_2}{n_2/2+1}\right) + 6\hat{\theta}^2}{(\sqrt{n_1} + \hat{\theta}\sqrt{n_2})^4}.$$

For an observed value $T_0^*$ of $T^*$, this test rejects the null hypothesis in (10) whenever

$$P\left(X > T_0^*\right) < \alpha, \tag{19}$$

where $X$ is a beta($a$, $b$) random variable. This test is available only for $k = 2$, and is applicable only when the alternative hypothesis is two-sided.

The generalized test based on $T$ in (7) and the other combined tests given above are scale invariant, and hence the powers of these tests depend on the variances via the ratio $\sigma_2^2/\sigma_1^2$ when $k = 2$. Therefore, for the case $k = 2$, it suffices to study the power properties for various values of $\sigma_2/\sigma_1$. The powers of the generalized test are computed using Algorithm 2 with $n = 2500$ runs and Algorithm 1 with $m = 5000$ runs. The powers of the combined tests are computed using Monte Carlo method with 100,000 runs. The powers of

(a) the test in (16) with $\eta$ in (17),
(b) the test in (16) with $\eta = 0$,
(c) Fisher's test (18),
(d) the approximate test in (19), and
(e) the generalized test in (11)

are given in Tables 4(a)–4(d) for various values of $\mu$. Test (d) is included only for the case of $k = 2$, and other tests are compared when $k$ = 2, 3, 4, and 7. The power comparison studies indicate the followings:

(i) Among the tests (a), (b), (c), and (d), no test is uniformly better than the others when $k = 2$. Test (d) seems to be asymptotically less efficient than other tests; when $n_1 = n_2 = 4$, $n_1 = n_2 = 9$, and $\sigma_2/\sigma_1$ is small, it performs better than all other tests (see Table 4(a)); when $n_1 = n_2 = 12$ it is inferior to all other tests, while when

**Table 4(a)**
*Simulated powers of the tests when $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ and $\alpha = 0.05$*

| | | $n_1 = n_2 = 4$ | | | | | | | $n_1 = n_2 = 9$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mu$ | | | | | | | $\mu$ | | | | | |
| $\sigma_2/\sigma_1$ | Tests[a] | 0 | 0.4 | 0.8 | 1.2 | 1.6 | 2.0 | 2.4 | 0 | .2 | .4 | .6 | .8 | 1 |
| 1 | (a) | | .14 | .39 | .70 | .90 | .98 | 1.0 | | .12 | .31 | .60 | .84 | .95 |
| | (b) | | .09 | .25 | .53 | .79 | .94 | .99 | | .09 | .22 | .47 | .73 | .90 |
| | (c) | | .11 | .31 | .61 | .86 | .97 | 1.0 | | .10 | .25 | .53 | .79 | .94 |
| | (d) | .05 | .15 | .45 | .78 | .95 | .99 | 1.0 | .05 | .12 | .34 | .63 | .86 | .96 |
| | (e) | .03 | .10 | .30 | .59 | .84 | .95 | .99 | .05 | .10 | .30 | .58 | .84 | .96 |
| 2 | (a) | | .10 | .26 | .50 | .71 | .86 | .94 | | .09 | .22 | .42 | .64 | .82 |
| | (b) | | .08 | .19 | .36 | .56 | .74 | .86 | | .08 | .18 | .35 | .57 | .76 |
| | (c) | | .08 | .19 | .38 | .60 | .79 | .91 | | .08 | .17 | .33 | .55 | .75 |
| | (d) | .05 | .11 | .28 | .52 | .73 | .86 | .93 | .05 | .09 | .22 | .43 | .65 | .82 |
| | (e) | .04 | .08 | .19 | .39 | .60 | .78 | .90 | .05 | .08 | .20 | .41 | .64 | .82 |
| 3 | (a) | | .10 | .24 | .44 | .64 | .79 | .87 | | .09 | .19 | .37 | .56 | .73 |
| | (b) | | .09 | .19 | .36 | .54 | .70 | .81 | | .08 | .18 | .35 | .56 | .75 |
| | (c) | | .08 | .17 | .31 | .49 | .67 | .81 | | .07 | .15 | .29 | .48 | .67 |
| | (d) | .05 | .09 | .23 | .43 | .61 | .74 | .82 | .05 | .09 | .20 | .38 | .59 | .76 |
| | (e) | .04 | .07 | .17 | .35 | .53 | .71 | .84 | .05 | .09 | .20 | .36 | .57 | .78 |
| 4 | (a) | | .09 | .22 | .41 | .60 | .74 | .83 | | .08 | .18 | .34 | .52 | .67 |
| | (b) | | .09 | .20 | .37 | .55 | .71 | .82 | | .08 | .19 | .36 | .56 | .75 |
| | (c) | | .08 | .16 | .29 | .45 | .61 | .74 | | .07 | .15 | .27 | .45 | .63 |
| | (d) | .05 | .09 | .21 | .38 | .55 | .66 | .74 | .05 | .08 | .20 | .37 | .57 | .74 |
| | (e) | .05 | .08 | .17 | .33 | .51 | .69 | .81 | .05 | .09 | .19 | .36 | .54 | .74 |
| 5 | (a) | | .09 | .21 | .39 | .57 | .70 | .79 | | .08 | .18 | .32 | .49 | .62 |
| | (b) | | .09 | .20 | .37 | .55 | .71 | .82 | | .08 | .19 | .35 | .56 | .75 |
| | (c) | | .08 | .15 | .27 | .42 | .58 | .72 | | .07 | .14 | .27 | .44 | .62 |
| | (d) | .05 | .09 | .20 | .37 | .52 | .63 | .70 | .05 | .08 | .19 | .36 | .56 | .74 |
| | (e) | .04 | .07 | .16 | .32 | .50 | .68 | .80 | .05 | .08 | .18 | .33 | .55 | .72 |
| 6 | (a) | | .09 | .21 | .37 | .54 | .67 | .75 | | .08 | .17 | .31 | .47 | .60 |
| | (b) | | .09 | .20 | .38 | .57 | .73 | .85 | | .08 | .18 | .35 | .56 | .75 |
| | (c) | | .08 | .15 | .27 | .39 | .54 | .66 | | .07 | .14 | .26 | .43 | .61 |
| | (d) | .05 | .08 | .20 | .35 | .50 | .61 | .68 | .05 | .09 | .19 | .36 | .56 | .74 |
| | (e) | .05 | .08 | .17 | .33 | .50 | .66 | .79 | .05 | .08 | .17 | .33 | .57 | .74 |
| 10 | (a) | | .09 | .19 | .34 | .49 | .59 | .66 | | .08 | .16 | .28 | .42 | .53 |
| | (b) | | .09 | .21 | .38 | .58 | .75 | .87 | | .08 | .18 | .35 | .56 | .75 |
| | (c) | | .08 | .15 | .26 | .39 | .54 | .66 | | .07 | .14 | .26 | .42 | .59 |
| | (d) | .05 | .08 | .18 | .34 | .49 | .59 | .66 | .05 | .08 | .19 | .36 | .56 | .75 |
| | (e) | .04 | .08 | .18 | .33 | .51 | .67 | .79 | .05 | .07 | .17 | .38 | .54 | .70 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (d) the approximate test in (19); (e) the generalized test in (11).

**Table 4(b)**
*Simulated powers of the tests when* $H_0 : \mu = 0$ vs $H_a : \mu \neq 0$ *and* $\alpha = 0.05$

| | | $n_1 = n_2 = 12$ | | | | | | $n_1 = 15,\ n_2 = 4$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mu$ | | | | | | $\mu$ | | | | | |
| $\sigma_2/\sigma_1$ | Tests[a] | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| 1 | (a) | | .39 | .71 | .91 | .99 | 1.0 | | .11 | .28 | .54 | .77 | .90 |
| | (b) | | .28 | .58 | .84 | .97 | 1.0 | | .08 | .21 | .43 | .67 | .84 |
| | (c) | | .32 | .64 | .88 | .98 | 1.0 | | .10 | .27 | .55 | .81 | .95 |
| | (d) | .05 | .15 | .44 | .77 | .95 | .99 | .05 | .12 | .35 | .65 | .87 | .96 |
| | (e) | .04 | .35 | .69 | .91 | .99 | 1.0 | .04 | .10 | .26 | .53 | .83 | .94 |
| 2 | (a) | | .27 | .51 | .74 | .89 | .96 | | .10 | .27 | .49 | .68 | .80 |
| | (b) | | .23 | .45 | .69 | .87 | .96 | | .10 | .27 | .51 | .74 | .88 |
| | (c) | | .21 | .42 | .66 | .85 | .96 | | .09 | .23 | .47 | .73 | .90 |
| | (d) | .05 | .11 | .30 | .56 | .80 | .93 | .05 | .11 | .31 | .58 | .81 | .98 |
| | (e) | .04 | .24 | .49 | .74 | .90 | .98 | .04 | .10 | .28 | .52 | .74 | .89 |
| 3 | (a) | | .23 | .44 | .67 | .85 | .95 | | .10 | .26 | .46 | .62 | .72 |
| | (b) | | .18 | .36 | .59 | .79 | .92 | | .11 | .29 | .56 | .79 | .92 |
| | (c) | | .22 | .41 | .59 | .74 | .95 | | .09 | .23 | .46 | .71 | .89 |
| | (d) | .05 | .10 | .27 | .51 | .74 | .89 | .05 | .11 | .30 | .57 | .81 | .93 |
| | (e) | .05 | .22 | .46 | .71 | .87 | .95 | .05 | .11 | .28 | .57 | .79 | .92 |
| 4 | (a) | | .22 | .41 | .59 | .74 | .95 | | .10 | .25 | .43 | .59 | .67 |
| | (b) | | .22 | .44 | .67 | .85 | .95 | | .11 | .30 | .57 | .81 | .94 |
| | (c) | | .18 | .35 | .56 | .75 | .89 | | .09 | .23 | .45 | .70 | .88 |
| | (d) | .05 | .10 | .26 | .49 | .72 | .88 | .05 | .11 | .30 | .58 | .81 | .94 |
| | (e) | .05 | .23 | .43 | .69 | .85 | .94 | .05 | .11 | .29 | .56 | .79 | .93 |
| 5 | (a) | | .21 | .38 | .56 | .69 | .77 | | .10 | .24 | .42 | .56 | .64 |
| | (b) | | .23 | .44 | .67 | .85 | .95 | | .11 | .30 | .58 | .81 | .94 |
| | (c) | | .17 | .34 | .54 | .74 | .88 | | .09 | .22 | .45 | .70 | .88 |
| | (d) | .05 | .10 | .25 | .49 | .72 | .88 | .05 | .11 | .30 | .58 | .82 | .94 |
| | (e) | .05 | .22 | .43 | .66 | .85 | .95 | .05 | .11 | .29 | .57 | .82 | .93 |
| 6 | (a) | | .20 | .37 | .53 | .66 | .73 | | .10 | .24 | .41 | .54 | .62 |
| | (b) | | .23 | .44 | .67 | .85 | .95 | | .11 | .30 | .57 | .82 | .95 |
| | (c) | | .17 | .33 | .53 | .73 | .87 | | .09 | .22 | .44 | .70 | .88 |
| | (d) | .05 | .10 | .25 | .48 | .71 | .88 | .05 | .11 | .30 | .58 | .82 | .95 |
| | (e) | .05 | .22 | .44 | .66 | .85 | .95 | .05 | .11 | .29 | .56 | .82 | .95 |
| 10 | (a) | | .19 | .34 | .48 | .58 | .64 | | .10 | .23 | .39 | .51 | .57 |
| | (b) | | .23 | .44 | .67 | .85 | .95 | | .11 | .30 | .58 | .82 | .95 |
| | (c) | | .17 | .32 | .52 | .72 | .86 | | .09 | .22 | .45 | .69 | .87 |
| | (d) | .05 | .10 | .25 | .48 | .71 | .88 | .05 | .11 | .30 | .58 | .82 | .94 |
| | (e) | .05 | .22 | .43 | .67 | .84 | .95 | .05 | .12 | .29 | .58 | .81 | .95 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (d) the approximate test in (19); (e) the generalized test in (11).

$n_1 = 15$, $n_2 = 4$, it performs as good as other tests. For moderate $k$, Fisher's test seems to be better than the tests (a) and (b) (see Tables 4(d) and 4(e)). The test (a) is better than (b) when $\sigma_2/\sigma_1$ is not too different from 1; otherwise (b) is better than (a).

(ii) The generalized test (e) is slightly conservative when the differences among the variances are not large and/or sample sizes are very small.

(iii) For smaller $k$, conditions on the sample sizes are necessary for the generalized test (e) to dominate the other tests; for $k = 2$, a sufficient condition is that both $n_1$ and $n_2$ are at least nine. This condition does not seem to be necessary; see Table 4(b). For $k = 3$, the sample sizes must be at least seven; for $k \geq 4$, they must be at least six. If the sample sizes satisfy these conditions and/or $k$ is moderate, then the powers of the generalized test are close to those of the one that has the highest power among the other three tests ((a), (b), and (c)). In some situations, the powers of the generalized test are much higher than those of the other tests (see Tables 4(d), 4(e), and 5).

## 7. Illustrative Examples

We shall now illustrate the generalized variable methods using two examples given in Meier (1953) and Eberhardt et al. (1989). These examples are also used by Jordan and Krishnamoorthy (1996) for constructing 95% confidence intervals. Applicability of the present problem to these examples and other details are discussed in the article just cited.

**Table 4(c)**
*Simulated powers of the tests when $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ and $\alpha = 0.05$*

| | | $n_1 = n_2 = n_3 = 7$ | | | | | | | $n_1 = n_2 = n_3 = n_4 = 6$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mu$ | | | | | | | $\mu$ | | | | | |
| $(\sigma_1, \sigma_2, \sigma_3)$ | Tests[a] | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | $(\sigma_1, \sigma_2, \sigma_3, \sigma_4)$ | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| (1,1,1) | (a) | | .10 | .26 | .52 | .78 | .93 | (1,1,1,1) | | .09 | .22 | .47 | .74 | .91 |
| | (b) | | .08 | .18 | .40 | .66 | .87 | | | .07 | .16 | .36 | .62 | .84 |
| | (c) | | .09 | .24 | .50 | .78 | .94 | | | .10 | .23 | .50 | .79 | .95 |
| | (e) | .05 | .10 | .30 | .60 | .85 | .96 | | .05 | .11 | .29 | .62 | .89 | .97 |
| (3,1,1) | (a) | | .08 | .19 | .38 | .62 | .81 | (3,2,1,1) | | .07 | .16 | .31 | .52 | .72 |
| | (b) | | .08 | .16 | .34 | .57 | .78 | | | .07 | .14 | .27 | .47 | .67 |
| | (c) | | .08 | .16 | .33 | .55 | .76 | | | .07 | .13 | .24 | .41 | .60 |
| | (e) | .05 | .07 | .23 | .49 | .67 | .84 | | .05 | .07 | .14 | .36 | .56 | .76 |
| (5,3,1) | (a) | | .07 | .15 | .26 | .42 | .59 | (9,6,3,1) | | .07 | .13 | .22 | .35 | .50 |
| | (b) | | .07 | .15 | .27 | .43 | .60 | | | .07 | .13 | .22 | .36 | .51 |
| | (c) | | .06 | .10 | .16 | .25 | .36 | | | .06 | .08 | .12 | .18 | .25 |
| | (e) | .05 | .07 | .16 | .25 | .40 | .56 | | .04 | .07 | .14 | .22 | .34 | .46 |
| (8,4,1) | (a) | | .07 | .14 | .26 | .41 | .58 | (2,2,1,1) | | .07 | .16 | .32 | .53 | .73 |
| | (b) | | .07 | .15 | .27 | .43 | .60 | | | .07 | .14 | .27 | .46 | .67 |
| | (c) | | .06 | .10 | .16 | .25 | .35 | | | .07 | .13 | .25 | .42 | .62 |
| | (e) | .05 | .06 | .15 | .27 | .43 | .58 | | .05 | .07 | .18 | .34 | .54 | .76 |
| (10,5,1) | (a) | | .07 | .14 | .26 | .41 | .58 | (16,9,6,1) | | .07 | .13 | .22 | .35 | .50 |
| | (b) | | .07 | .15 | .27 | .43 | .60 | | | .07 | .13 | .23 | .36 | .51 |
| | (c) | | .06 | .10 | .16 | .24 | .35 | | | .06 | .08 | .12 | .17 | .24 |
| | (e) | .05 | .08 | .15 | .26 | .44 | .55 | | .05 | .08 | .12 | .22 | .35 | .51 |
| (1,.8,.5) | (a) | | .48 | .95 | .99 | 1.0 | 1.0 | (3,1,.8,5) | | .37 | .87 | .98 | .99 | 1.0 |
| | (b) | | .42 | .93 | .99 | 1.0 | 1.0 | | | .35 | .85 | .97 | .99 | 1.0 |
| | (c) | | .33 | .88 | 1.0 | 1.0 | 1.0 | | | .23 | .71 | .97 | 1.0 | 1.0 |
| | (e) | | .41 | .95 | 1.0 | 1.0 | 1.0 | | .05 | .37 | .86 | .99 | 1.0 | 1.0 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (e) the generalized test in (11).

To compute the 95% generalized confidence limits, we used Algorithm 1 with $m = 100{,}000$ runs.

*Example 1 (Meier, 1953).* Four experiments are used to estimate the mean percentage of albumin $\mu$ in the plasma protein of normal human subjects. The summary statistics, along with sample sizes, are given in Table 2. The confidence intervals are given following Table 2. All three methods yielded confidence intervals with centers very close to each other. The interval (13) due to Fairweather (1972) is the shortest among the three intervals. The generalized interval (12) is close to the interval (13), and shorter than the interval (14) due to Jordan and Krishnamoorthy (1996). For the sake of illustration, we also computed the p-values for testing $H_0: \mu = 59.5$ vs. $H_a: \mu \neq 59.5$. Noting that $\eta$ in (17) for this example is 1, we see that the test (16) and the generalized test reject $H_0$ at the level of 0.05. Fisher's test in (18) does not reject the null hypothesis.

*95% confidence intervals.*

Fairweather's (1972) Interval (13): $61.04 \pm 1.15$

Jordan and Krishnamoorthy's (1996) Interval (14): $61.00 \pm 1.44$

Generalized limits in (12): $61.01 \pm 1.22$

*P-Values for testing.* $H_0: \mu = 59.5$ vs. $H_a: \mu \neq 59.5$

Zhou and Mathew (1993) Test (16): 0.088; $\eta$ in (17) is 1

Fisher's Test in (18): 0.055

Generalized p-value in (11): 0.016

*Example 2.* Here, we are interested in estimating the mean selenium content in nonfat milk powder using four different analytical methods. The methods and summary statistics are given in Table 3. The interval estimates are given below. We

**Table 2**
*Percentage of albumin in plasma protein*

| Experiment | $n_i$ | $\bar{x}$ | $s^2$ |
|---|---|---|---|
| A | 12 | 62.3 | 12.986 |
| B | 15 | 60.3 | 7.840 |
| C | 7 | 59.5 | 33.433 |
| D | 16 | 61.5 | 18.513 |

**Table 3**
*Selenium in nonfat milk powder*

| Methods | $n_i$ | $\bar{x}$ | $s^2$ |
|---|---|---|---|
| Atomic absorption spectrometry | 8 | 105.0 | 85.711 |
| Neutron activation instrumental | 12 | 109.75 | 20.748 |
| Radiochemical | 14 | 109.5 | 2.729 |
| Isotope dilution mass spectrometry | 8 | 113.25 | 33.640 |

**Table 4(d)**
*Simulated powers of the tests when $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ and $\alpha = 0.05$*

| $k = 7$; $n_i = 4$, $i = 1, \ldots, 7$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\mu$ | | | | | |
| $(\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7)$ | Test[a] | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| (1,1,1,1,1,1,1) | (a) | | .07 | .14 | .29 | .52 | .75 |
| | (b) | | .06 | .10 | .20 | .39 | .62 |
| | (c) | | .08 | .18 | .40 | .69 | .90 |
| | (e) | .05 | .08 | .23 | .49 | .76 | .94 |
| (1,.9,.9,.9,.9,.9,.9) | (a) | | .07 | .16 | .35 | .61 | .83 |
| | (b) | | .06 | .12 | .25 | .48 | .72 |
| | (c) | | .08 | .21 | .25 | .48 | .72 |
| | (e) | .05 | .09 | .28 | .57 | .84 | .97 |
| (1,1,1,1,.9,.9,.9) | (a) | | .07 | .15 | .32 | .57 | .79 |
| | (b) | | .06 | .11 | .23 | .43 | .67 |
| | (c) | | .08 | .19 | .43 | .74 | .93 |
| | (e) | .05 | .08 | .26 | .53 | .81 | .96 |
| (1,.9,.9,.9,.8,.8,.8) | (a) | | .08 | .18 | .40 | .67 | .87 |
| | (b) | | .07 | .13 | .29 | .54 | .78 |
| | (c) | | .09 | .23 | .54 | .84 | .97 |
| | (e) | .04 | .08 | .31 | .63 | .89 | .97 |
| (1,.9,.9,.9,.6,.6,.6) | (a) | | .09 | .25 | .56 | .83 | .95 |
| | (b) | | .08 | .19 | .44 | .73 | .90 |
| | (c) | | .10 | .32 | .70 | .94 | .98 |
| | (e) | .04 | .12 | .41 | .78 | .95 | .99 |
| (1,.8,.8,.5,.5,.4,.4) | (a) | | .14 | .50 | .86 | .97 | .99 |
| | (b) | | .11 | .39 | .78 | .94 | .98 |
| | (c) | | .16 | .59 | .94 | 1.0 | 1.0 |
| | (e) | .05 | .19 | .69 | .97 | 1.0 | 1.0 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (e) the generalized test in (11).

again see that the centers of the intervals are very close to each other. The generalized confidence interval (12) is the shortest among the three intervals. The p-values for testing $H_0: \mu = 110.5$ vs. $H_a: \mu \neq 110.5$ are given below. At the level of significance 0.05, we see that only the Zhou and Mathew (1993) test (16) rejects $H_0$.

*95% confidence intervals.*

Fairweather's (1972) Interval (13): $109.7 \pm 1.11$

Jordan and Krishnamoorthy's (1996) Interval (14): $109.6 \pm 1.08$

Generalized limits in (12): $109.6 \pm 0.93$

*P-Values for testing.* $H_0: \mu = 110.5$ vs. $H_a: \mu \neq 110.5$

Zhou and Mathew's (1993) Test (16): 0.042; $\eta$ in (17) is 0

Fisher's Test (18): 0.071

Generalized p-value in (11): 0.064

## 8. Concluding Remarks

In this article, we have shown yet another problem where the generalized variable approach yielded efficient inferential procedures. Unlike other methods, it yielded the generalized pivot variable $T$ in (7) which is simple to use for both hypothesis testing and for constructing confidence intervals for the common mean. Even though this approach is computationally involved, it is as easy as other procedures once it is programmed using Algorithm 1. In order to get consistent results irrespective of the seed used for random number generators, we recommend a Monte Carlo simulation of 100,000 runs.

So far, the majority of the articles in this area have considered comparison studies only for the case of $k = 2$ because powers or coverage probabilities are computed numerically. Another reason could be many authors assumed that the results that hold for smaller $k$ will also hold for large $k$. However, we showed that the results that hold for smaller $k$ (say $\leq 6$) may not hold for moderately large $k$. In particular, we showed that Fisher's test, although it does not appear to be superior to other tests for smaller $k$, seems to be better than the tests (a) and (b) given in Zhou and Mathew (1993) for moderate values of $k$ (see Tables 4(d) and 4(e)). We also did power comparison studies for large $k$ for a few values of $(\sigma_1^2, \ldots, \sigma_k^2)$ and $n_i = 4$, $i = 1, \ldots, k$, in an arbitrary manner (these values are not reported here). For instance, when $k = 20$, we randomly picked 20 numbers from the interval [1,50] for population variances, and 20 integers from [3,15] for sample sizes, and computed the powers of all the tests considered. The

**Table 4(e)**

*Simulated powers of the tests when* $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ *and* $\alpha = 0.05$

$k = 7$; $n_i = 4$, $i = 1, \ldots, 7$

| | | $\mu$ | | | | | |
|---|---|---|---|---|---|---|---|
| $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_6^2, \sigma_7^2)$ | Test[a] | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| (5,3,3,3,2,2,1) | (a) | | .06 | .10 | .18 | .31 | .48 |
| | (b) | | .06 | .08 | .14 | .23 | .37 |
| | (c) | | .06 | .11 | .21 | .38 | .58 |
| | (e) | .05 | .05 | .17 | .28 | .50 | .69 |
| (10,1,1,1,1,2,2) | (a) | | .06 | .12 | .23 | .41 | .62 |
| | (b) | | .06 | .09 | .17 | .31 | .50 |
| | (c) | | .07 | .13 | .28 | .50 | .74 |
| | (e) | .04 | .08 | .17 | .38 | .61 | .83 |
| (10,9,4,1,1,1,1) | (a) | | .06 | .11 | .22 | .38 | .59 |
| | (b) | | .06 | .09 | .17 | .30 | .48 |
| | (c) | | .07 | .12 | .24 | .42 | .64 |
| | (e) | .04 | .06 | .18 | .31 | .52 | .77 |
| (10,10,10,3,3,3,3) | (a) | | .06 | .07 | .10 | .14 | .21 |
| | (b) | | .05 | .06 | .08 | .11 | .16 |
| | (c) | | .06 | .07 | .11 | .16 | .25 |
| | (e) | .04 | .06 | .08 | .14 | .23 | .30 |
| (10,9,8,7,6,5,4) | (a) | | .05 | .06 | .08 | .10 | .14 |
| | (b) | | .05 | .06 | .07 | .08 | .11 |
| | (c) | | .05 | .06 | .09 | .12 | .17 |
| | (e) | .04 | .05 | .08 | .10 | .17 | .23 |
| (10,10,10,8,8,8,8) | (a) | | .05 | .06 | .07 | .09 | .11 |
| | (b) | | .05 | .05 | .06 | .07 | .09 |
| | (c) | | .05 | .06 | .07 | .10 | .13 |
| | (e) | | .05 | .06 | .08 | .13 | .17 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (e) the generalized test in (11).

overall pattern that we observed is similar to the one reported in Table 5. For moderately large $k$, the generalized method seems to be the most efficient among all the methods considered. This may be the most common situation in clinical trials where $k$ (the number of groups of patients) is large and $n_i$ (the number of patients in the $i$th group) is small. Thus, we conclude that the generalized variable methods are very efficient, and readily applicable for practical use.

## Acknowledgements

The authors would like to thank a referee and an associate editor for providing valuable comments and suggestions.

## Résumé

Cet article présente des procédures pour effectuer un test d'hypothèse et une estimation par intervalle de la moyenne commune de plusieurs populations normales. Les méthodes

**Table 5**

*Simulated powers of the tests when* $H_0: \mu = 0$ vs $H_a: \mu \neq 0$ *and* $\alpha = 0.05$

$k = 20$, $n_i = 4$, $i = 1, \ldots, 20$

| | | $\mu$ | | | | | | | $\mu$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $(\sigma_1^2, \ldots, \sigma_{20}^2)$ | Tests[a] | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | $(\sigma_1^2, \ldots, \sigma_{20}^2)$ | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
| (1,1,1,1,1, | (a) | .05 | .06 | .11 | .22 | .41 | .64 | (10,9,8,7,6, | .05 | .06 | .08 | .13 | .21 | .34 |
| 2,2,2,2,2, | (b) | .05 | .06 | .09 | .17 | .32 | .53 | 9,9,9,8,8, | .05 | .05 | .07 | .11 | .17 | .27 |
| 3,3,3,3,3, | (c) | .05 | .07 | .17 | .38 | .69 | .92 | 3,3,3,3,3, | .05 | .06 | .10 | .18 | .33 | .53 |
| 4,4,4,4,4) | (e) | .05 | .11 | .34 | .62 | .89 | .97 | 1,2,3,4,5) | .05 | .09 | .16 | .36 | .59 | .80 |
| (30,20,10,12,11, | (a) | .05 | .06 | .09 | .15 | .24 | .38 | (1,.9,.9,.8,.8, | .05 | .09 | .30 | .71 | .93 | .98 |
| 20,25,24,12,1, | (b) | .05 | .06 | .08 | .13 | .21 | .33 | .7,.7,.7,.6,.6, | .05 | .08 | .23 | .60 | .88 | .97 |
| 3,6,5,4,3, | (c) | .05 | .06 | .09 | .15 | .26 | .41 | .6,.6,.5,.5,.5, | .05 | .14 | .60 | .97 | 1 | 1 |
| 30,31,35,1,2) | (e) | .04 | .08 | .17 | .32 | .50 | .73 | .5,.4,.4,.4,.4) | .05 | .30 | .81 | .98 | 1 | 1 |

[a](a) the test in (16) with $\eta$ in (17); (b) the test in (16) with $\eta = 0$; (c) Fisher's test (18); (e) the generalized test in (11).

sont basées sur le concept du p généralisé et de l'intervalle de confiance généralisé. L'intérêt des méthodes proposées est quantifié et comparé à celui de méthodes existantes. Les comparaisons numériques montrent que les nouvelles procédures sont correctes et ont des performances supérieures à celles des méthodes existantes quand les échantillons sont de taille modérée et que le nombre de populations est quatre ou moins. Si le nombre de populations est cinq ou plus, cette nouvelle méthode est bien supérieure aux méthodes existantes, quelles que soient les tailles des échantillons. Nous illustrons cette nouvelle méthode et les méthodes existantes avec des données de deux exemples.

## References

Bhattacharya, C. G. (1984). Two inequalities with an application. *Annals of the Institute of Statistical Mathematics* **36,** 129–134.

Cohen, A. and Sackrowitz, H. B. (1984). Testing hypotheses about the common mean of normal distributions. *Journal of Statistical Planning and Inference* **9,** 207–227.

Cohen, A. and Sackrowitz, H. B. (1989). Exact tests that recover interblock information in balanced incomplete blocks designs. *Journal of the American Statistical Association* **84,** 556–560.

Eberhardt, K. R., Reeve, C. P., and Spiegelman, C. H. (1989). A minimax approach to combine means, with practical examples. *Chemometrics and Intelligent Laboratory Systems* **5,** 129–148.

Fairweather, W. R. (1972). A method of obtaining an exact confidence interval for the common mean of several normal populations. *Applied Statistics* **21,** 229–233.

Graybill, F. A. and Deal, R. B. (1959). Combining unbiased estimators. *Biometrics* **15,** 543–550.

Hinkley, D. V. (1979). A note on the weighted means problems. *Scandinavians Journal of Statistics* **6,** 37–40.

Jordan, S. J. and Krishnamoorthy, K. (1995a). Confidence regions for the common mean vector of several multivariate normal populations. *The Canadian Journal of Statistics* **23,** 283–297.

Jordan, S. J. and Krishnamoorthy, K. (1995b). On combining independent tests in linear models. *Statistics and Probability Letters* **23,** 117–122.

Jordan, S. J. and Krishnamoorthy, K. (1996). Exact confidence intervals for the common mean of several normal populations. *Biometrics* **52,** 77–86.

Maric, N. and Graybill, F. A. (1979). Small samples confidence intervals on common mean of two normal distributions with unequal variances. *Communications in Statistics—Theory and Methods* **A8,** 1255–1269.

Mathew, T., Sinha, B. K., and Zhou, L. (1993). Some statistical procedures for combining independent tests. *Journal of the American Statistical Association* **88,** 912–919.

Meier, P. (1953). Variance of a weighted mean. *Biometrics* **9,** 59–73.

Pagurova, V. I. and Gurskii, V. V. (1979). A confidence interval for the common mean of several normal distributions. *Theory of Probability and Its Applications* **88,** 912–919.

Tsui, K. W. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association* **84,** 602–607.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association* **88,** 899–905.

Weerahandi, S. (1995a). ANOVA under unequal error variances. *Biometrics* **51,** 589–599.

Weerahandi, S. (1995b). *Exact Statistical Methods for Data Analysis.* New York: Springer-Verlag.

Weerahandi, S. and Berger, V. W. (1999). Exact inference for growth curves with intraclass correlation structure. *Biometrics* **55,** 921–924.

Welch, B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika* **34,** 28–35.

Yu, P. L. H., Sun, Y., and Sinha, B. K. (1999). On exact confidence intervals for the common mean of several normal populations. *Journal of Statistical Planning and Inference* **81,** 263–277.

Zhou, L. and Mathew, T. (1993). Combining independent tests in linear models. *Journal of the American Statistical Association* **88,** 650–655.

Zhou, L. and Mathew, T. (1994). Some tests for variance components using generalized p-values. *Technometrics* **36,** 394–402.

*Received August* 2002. *Revised November* 2002.
*Accepted December* 2002.