Nonparametric Bootstrap Confidence Intervals for Variance Components Applied to Interlaboratory Comparisons

Brent D. BURCH

Exact confidence intervals for variance components in linear mixed models rely heavily on normal distribution assumptions. If the random effects in the model are not normally distributed, then the true coverage probabilities of these conventional intervals may be erratic. In this paper we examine the performance of nonparametric bootstrap confidence intervals based on restricted maximum likelihood (REML) estimators. Asymptotic theory suggests that these intervals will achieve the nominal coverage value as the sample size increases. Incorporating a small-sample adjustment term in the bootstrap confidence interval construction process improves the performance of these intervals for small to intermediate sample sizes. Simulation studies suggest that the bootstrap standard method (with a transformation) and the bootstrap bias-corrected and accelerated (BC_a) method produce confidence intervals that have good coverage probabilities under a variety of distribution assumptions. For an interlaboratory comparison of mercury concentration in oyster tissue, a balanced one-way random effects model is used to quantify the proportion of the variation in mercury concentration that can be attributed to the laboratories. In this application the exact confidence interval using normal distribution theory produces misleading results and inferences based on nonparametric bootstrap procedures are more appropriate.

Key Words: Bootstrap BC_a method; Bootstrap standard method; One-way random effects model; Small-sample adjustment.

1. INTRODUCTION

Quantifying the influence that factors have on a response variable has been a continuing subject of study in statistical modeling. In many applications the factors themselves are random and thus contribute a certain amount of uncertainty or dispersion to the model. If one measures uncertainty using variance components, then inferences pertaining to a ratio

Brent D. Burch (✉) is Professor, Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ 86011-5717, USA (E-mail: *brent.burch@nau.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 17, Number 2, Pages 228–245
DOI: 10.1007/s13253-012-0087-9

of variance components can shed light on the relative importance of a factor or source of dispersion. Specifically, if σ_1^2 and σ_2^2 denote the variance components associated with two sources of dispersion, then the ratio σ_1^2/σ_2^2 can be used to compare the impact that the two sources of dispersion have on the response. In the context of one-way random effects models, if the ratio is small, then the variability in the responses is due largely to the variation within classes (or laboratories). Alternatively, if the ratio is large, then the variability in the responses is due for the most part to the variation amongst the classes (or laboratories).

Calculating point estimates and/or constructing confidence intervals for a ratio of variance components can help an investigator reach meaningful conclusions. In linear mixed models, the restricted maximum likelihood (REML) method is commonly used to compute point estimates of variance components. Unlike maximum likelihood estimators, Searle, Casella, and McCulloch (1992) and West, Welch, and Galecki (2007) indicate that REML estimators take into account the degrees of freedom that are lost from estimating the fixed effects in the model. Harville (2004) suggests that REML has become the preferred method to estimate variance components. Furthermore, Jiang (1996, 2005) described the large-sample properties of REML estimators so one can construct approximate or large-sample confidence intervals for variance components. Of course, the asymptotic standard error of the point estimator must be estimated to actually implement this approach. In some applications, exact confidence intervals for ratios of components are available under normal distribution assumptions. See Harville and Fenech (1985) and Burch and Iyer (1997) for some examples.

The main focus of this paper is to understand how bootstrap resampling procedures can be employed to construct confidence intervals for ratios of variance components in balanced one-way random effects models. Specifically, let $\theta = \sigma_1^2/\sigma_2^2$ and $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$. Then consider the performance of nonparametric bootstrap confidence intervals for θ or ρ under normal and nonnormal distribution assumptions. Ukoumunne et al. (2003) presented results of bootstrap confidence intervals for ρ for a specific nonnormal case using small values of ρ. Ren et al. (2010) examined bootstrap confidence intervals for individual variance components in hierarchical data. Field and Welsh (2007) showed that bootstrap variances in the balanced one-way random effects model are asymptotically correct under certain conditions.

The nonparametric bootstrap confidence intervals considered in this paper are commonly referred to as the standard method, the percentile method, and the bias-corrected and accelerated (BC_a) method. Efron and Tibshirani (1986) provide an expanded list of bootstrap confidence interval methods. According to DiCiccio and Efron (1996), many bootstrap confidence interval procedures, including the BC_a method developed by Efron (1987), can be thought of as automatic algorithms that incorporate improvements over conventional confidence intervals. Chernick (2008) contends that while bootstrapping is not a panacea, it does allow scientists to discard restrictive modeling and distributional assumptions.

Section 2 of this paper describes an application involving measurements of mercury concentration in oyster tissue. An overview of the one-way random effects model is

given in Section 3 which also features the large-sample properties of the REML estimator and the normal-distribution-based confidence interval for θ. Section 4 presents the nonparametric bootstrap confidence intervals for θ or ρ and includes the necessary details to implement these procedures. Section 5 provides confidence interval simulation results for a variety of scenarios and distributions. In Section 6 we revisit the mercury in oyster tissue application and provide guidance as to which confidence interval procedures are appropriate. A discussion and summary comments are given in Section 7.

2. MERCURY IN OYSTER TISSUE

The National Oceanic and Atmospheric Administration (NOAA) National Status and Trends Program conducted ongoing comparisons of laboratory measurements of trace metals in marine sediments and biological tissues. In the ninth round of these interlaboratory comparisons, Willie and Berman (1995) reported that a total of 53 laboratories received materials to be analyzed. The materials included freeze dried marine sediment collected from Nova Scotia, freeze dried mussel collected in Boston Harbor, and certified reference materials (CRMs) known as NRC sediment BCSS-1 and NIST oyster tissue SRM 1566a. The data considered in this paper come from the National Institute of Standards and Technology (NIST) Standard Reference Material (SRM) 1566a.

Altogether, 13 NIST SRM 1566a elements including arsenic, mercury, and lead were under investigation. Each laboratory was requested to perform five replicate measurements for each element. Of the 53 laboratories who received materials, 44 laboratories submitted data. Furthermore, not all the laboratories provided results for all of the elements. In this paper we consider the NIST SRM 1566a mercury concentration measurements. Samples of the oyster material originated from a commercial source that had been ground and freeze dried. The oyster tissue material was then processed at NIST and bottled. In the end, 26 laboratories provided five replicate measurements each of NIST SRM 1566a mercury concentration (mg/kg) in the oyster tissue samples. Figure 1 displays side-by-side dotplots of the mercury measurements for the 26 laboratories. If, for a particular laboratory, five distinct vertical dots are not visible in the figure, then some of the individual measurements are equal to the same value. Ideally, measurements from different laboratories should be consistent because they are samples from the same source, for which the certified value of mercury concentration is 0.0654 mg/kg with a 95 % confidence interval of 0.0654 mg/kg $\pm$ 0.0067 mg/kg.

Compared to shark, swordfish, king mackerel and tilefish, oysters have relatively low levels of mercury. However, mercury exposure over long periods of time can affect the nervous system and can permanently damage the brain. In particular, high levels of methylmercury in the bloodstream of unborn babies and young children may harm the developing nervous system. See *http://www.epa.gov/mercury/* for more details.

Rather than focus on safety concerns, the original purpose of the aforementioned annual exercises was to assess the capabilities of laboratories to analyze the sediments and tissues

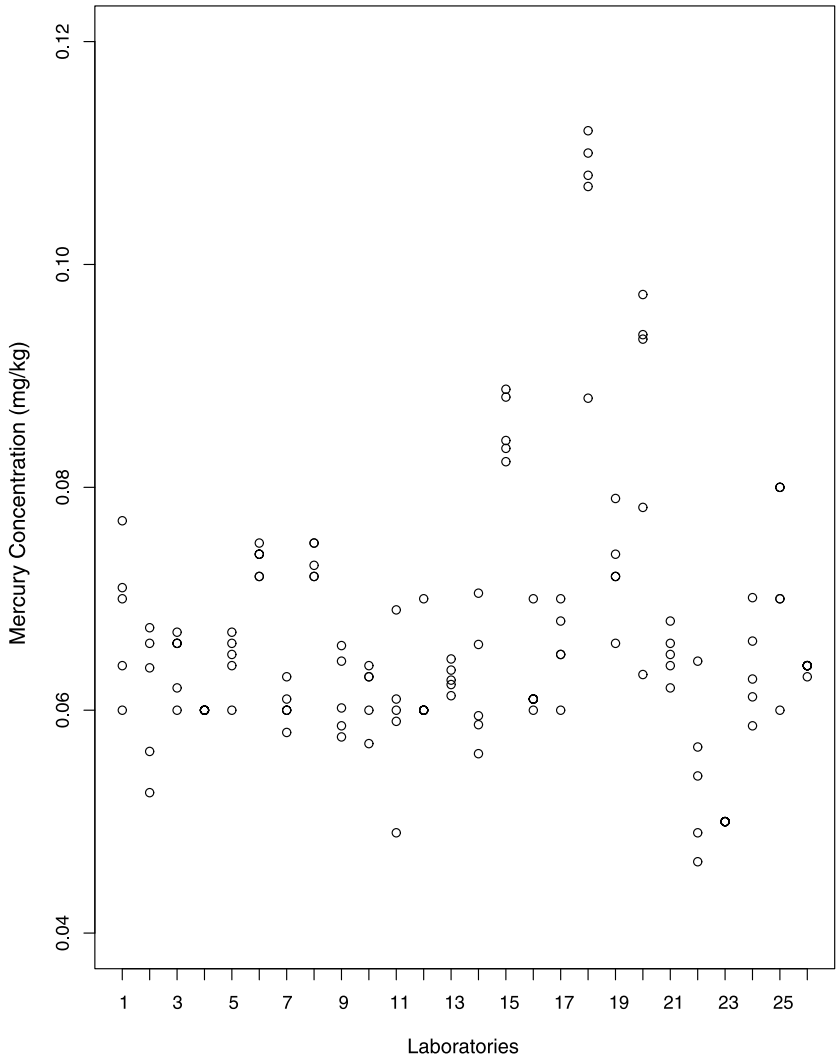

Figure 1. Dotplots of the mercury concentration measurements (mg/kg) for each of the 26 laboratories.

for trace metals. Table 1 displays the analysis of variance table for the mercury concentration measurements. If σ_1^2 denotes the variance between laboratories and σ_2^2 denotes the variance within laboratories, then $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ is the proportion of variation in mercury concentration that can be attributed to the laboratories. A balanced one-way random effects model is used for this data and confidence intervals for θ or ρ are of primary concern. If the resulting confidence interval for θ or ρ is comprised of values that are close to zero, then the variability in mercury concentration amongst the laboratories is negligible. Based on Figure 1, however, it appears that there is a substantial amount of interlaboratory variation.

Table 1. ANOVA table for the mercury concentration measurements in oyster tissue.

Source of variation	df	Sum of squares	Mean square
Between Labs	25	0.015381	0.000615
Within Labs	104	0.002969	0.000029
Total	129		

3. THE ONE-WAY RANDOM EFFECTS MODEL

The balanced one-way random effects model can be expressed as

$$Y_{ij} = \mu + A_i + e_{ij},$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $N = nm$ is the total sample size. In this hierarchical model, where there are two stages of sampling, the n classes of factor A represent a random sample from a population of classes. Secondly, for each class of A, a random sample of size m is selected and thus Y_{ij} is the jth observation associated with the ith class of A. In the mercury in oyster tissue application, one may consider the 26 laboratories in the study as a random selection of laboratories from a larger collection of laboratories. For a given laboratory, five mercury measurements are obtained and thus a total of 130 observations constitute the sample. The e_{ij}'s are referred to as random errors and may be interpreted as the deviations of the observations within the classes (or laboratories). In this paper the focus of attention rests on the variance components implicit in the model. Furthermore, μ, the overall mean of the Y_{ij}'s, is treated as a nuisance parameter.

It is assumed that the A_i's are independent and identically distributed with $E(A_i) = 0$, $\text{var}(A_i) = \sigma_1^2$, and (excess) kurtosis given by $\kappa_1 = E({A_i}^4)/\sigma_1^4 - 3$. Similarly, the e_{ij}'s are independent and identically distributed with $E(e_{ij}) = 0$, $\text{var}(e_{ij}) = \sigma_2^2$, and (excess) kurtosis given by $\kappa_2 = E({e_{ij}}^4)/\sigma_2^4 - 3$. In addition, suppose that the A_i's and e_{ij}'s are mutually independent. Note that the underlying distributions of the A_i's and the e_{ij}'s are not specified. Observations within the same class are correlated since $\text{cov}(Y_{ij}, Y_{ij'}) = \sigma_1^2$ for $j \neq j'$ and observations from different classes are uncorrelated. It follows that $E(Y_{ij}) = \mu$ and $\text{var}(Y_{ij}) = \sigma_1^2 + \sigma_2^2$ so that the two sources of variation, A_i and e_{ij}, combine in a linear manner when describing the variance of the responses.

The value of the ratio of variance components, $\theta = \sigma_1^2/\sigma_2^2$, serves as a measure of the relative importance of the two sources of variation. For instance, if the ratio is small, then the variability in responses is due largely to the dispersion within classes. Alternatively, if the ratio is large, then the variability in responses is due mainly to the dispersion amongst the classes. By definition, $0 \leq \theta < \infty$. The parameter $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ and θ are one-to-one functions and $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ is often referred to as the intraclass correlation coefficient since it is the correlation between two observations within the same class. The parameter ρ is also the proportion of the total variance in the Y_{ij}'s due to the random factor A since $\rho = \text{var}(A_i)/\text{var}(Y_{ij})$. Note that $0 \leq \rho < 1$ where a value close to zero indicates that the overall variance in the responses can be attributed to the small variance of A_i

(relative to the variance of e_{ij}) and a value close to one is consistent with a large variance of A_i (relative to the variance of e_{ij}).

In this paper, point estimators of variance components are obtained using the restricted maximum likelihood (REML) method. Furthermore, bootstrap confidence intervals for θ or ρ can be constructed using the large-sample properties of the REML estimators. Jiang (1996, 2005) provided a description of the large-sample properties of the REML estimator of θ and Burch (2011) considered REML-based confidence intervals for θ or ρ in one-way random effects models. Under normal theory, the restricted log-likelihood function with respect to θ and σ_2^2 can be written as

$$L_R\left(\theta, \sigma_2^2\right) \propto -\left[(nm-1)\ln\sigma_2^2 + (n-1)\ln(1+m\theta)\right] - \frac{1}{\sigma_2^2}\left[\text{SSE} + \frac{\text{SSA}}{1+m\theta}\right],$$

where the sum of squares within classes is $\text{SSE} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \overline{Y}_{i.}\right)^2$ and the sum of squares between classes is $\text{SSA} = \sum_{i=1}^{n} m\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$. $\overline{Y}_{i.}$ and $\overline{Y}_{..}$ denote the ith class sample mean and overall sample mean, respectively. The corresponding mean squares are $\text{MSE} = \text{SSE}/(n(m-1))$ and $\text{MSA} = \text{SSA}/(n-1)$. Differentiating L_R and maximizing over the parameter space yields the REML estimators

$$\widehat{\theta} = \frac{1}{m}\left(\frac{\text{MSA}}{\text{MSE}} - 1\right),$$

$$\widehat{\sigma_2^2} = \text{MSE},$$

if $\text{MSA} > \text{MSE}$. Otherwise, $\widehat{\theta} = 0$ and $\widehat{\sigma_2^2} = (\text{SSA} + \text{SSE})/(nm-1)$. One can verify that $P(\widehat{\theta} = 0)$ decreases as n increases. The corresponding REML estimator of ρ is $\widehat{\theta}/(\widehat{\theta}+1)$.

Assuming that the regularity conditions hold, Jiang (1996) proved that REML estimators are asymptotic normally distributed even in nonnormal applications. Westfall (1987) and Burch and Harris (2001) demonstrated that the regularity conditions for balanced one-way random effects models correspond to the number of classes, n, approaching infinity while the class size, m, remains finite. Jiang (1996, 2005) derived the asymptotic covariance matrix of REML estimators which leads to

$$\widehat{\theta} \overset{\text{asymp}}{\sim} N\left(\theta, \text{var}(\theta)\right) \tag{3.1}$$

where

$$\begin{aligned}\text{var}(\widehat{\theta}) &= \frac{(1+m\theta)^2}{m^2\sigma_2^4}\left(\text{var}(\text{MSE}) - \frac{2\text{cov}(\text{MSA}, \text{MSE})}{1+m\theta} + \frac{\text{var}(\text{MSA})}{(1+m\theta)^2}\right)\\ &= \frac{(1+m\theta)^2}{m^2}\left(\frac{\text{var}(\text{SSE})}{(E(\text{SSE}))^2} - \frac{2\text{cov}(\text{SSE}, \text{SSA})}{E(\text{SSE})E(\text{SSA})} + \frac{\text{var}(\text{SSA})}{(E(\text{SSA}))^2}\right).\end{aligned}$$

In practice, the distribution of $\widehat{\theta}$ is positively skewed so directly using the normal approximation in (3.1) may not produce acceptable results. The commonly used logarithmic transformation is applied to the REML estimator of θ using Slutzsky's theorem, and it follows that

$$\ln(1+m\widehat{\theta}) \overset{\text{asymp}}{\sim} N\left(\ln(1+m\theta), \frac{\text{var}(\text{SSE})}{(E(\text{SSE}))^2} - \frac{2\text{cov}(\text{SSE}, \text{SSA})}{E(\text{SSE})E(\text{SSA})} + \frac{\text{var}(\text{SSA})}{(E(\text{SSA}))^2}\right).$$

In general, it is difficult to estimate $\text{var}(\ln(1+m\widehat{\theta}))$. For example, see Burch (2011) where estimates of kurtosis were used. A simplification, however, does occur under normal distribution theory, where $\text{var}(\text{SSE}) = 2n(m-1)\sigma_2^4$, $\text{var}(\text{SSA}) = 2(n-1)\sigma_2^4(1+m\theta)^2$, and $\text{cov}(\text{SSE}, \text{SSA}) = 0$ giving $\text{var}(\ln(1+m\widehat{\theta})) = 2m/(n(m-1))$ which is free of θ. It follows that a large-sample confidence interval for θ is readily available for this simple case.

An exact equal-tailed $100(1-\alpha)$ % confidence interval for θ comes from an F-distributed pivotal quantity so that

$$P\left(\frac{1}{m}\left[\frac{\text{MSA}}{\text{MSE}}F_{\alpha/2} - 1\right] \le \theta \le \frac{1}{m}\left[\frac{\text{MSA}}{\text{MSE}}F_{1-\alpha/2} - 1\right]\right) = 1-\alpha, \tag{3.2}$$

where $F_{\alpha/2}$ and $F_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $F(n(m-1), n-1)$ distribution, respectively. The performance of this confidence interval procedure, in terms of actual coverage probability, can be very poor for nonnormal cases. See Burch (2011) for details.

4. NONPARAMETRIC BOOTSTRAP CONFIDENCE INTERVALS FOR θ OR ρ

In this section we consider some nonparametric bootstrap confidence intervals using the standard, percentile, and BC_a methods. The resampling with replacement procedure employed in this paper recognizes that observations within classes may be correlated whereas observations from different classes are independent. By notation, let the $N \times 1$ vector $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$ represent the original sample where the $m \times 1$ subvector $\mathbf{Y}_i$ corresponds to the observations within the ith class, $i = 1, \ldots, n$. In this two-stage sampling process the vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ are independent and identically distributed. A bootstrap sample of the original data is denoted by $\mathbf{Y}^* = (\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_n^*)$, where $(\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_n^*)$ is a simple random sample with replacement from $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$. In other words, if the ith class is randomly selected, then the entire set of m measurements is included in forming $\mathbf{Y}_i^*$. This is the same sampling scheme suggested by Davison and Hinkley (1997). Field and Welsh (2007) call this cluster bootstrap sampling.

Following the approach established by the REML estimators discussed in the previous section, the bootstrap estimators of θ and σ_2^2 are

$$\widehat{\theta}^* = \frac{1}{m}\left(\frac{\text{MSA}^*}{\text{MSE}^*} - 1\right),$$

$$\widehat{\sigma_2^2}^* = \text{MSE}^*,$$

if $\text{MSA}^* > \text{MSE}^*$. Otherwise, $\widehat{\theta}^* = 0$ and $\widehat{\sigma_2^2}^* = (\text{SSA}^* + \text{SSE}^*)/(nm-1)$. Note that SSA^* and SSE^* are constructed from the bootstrap sample $\mathbf{Y}^* = (\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_n^*)$.

The standard method considered here estimates $\text{var}(\ln(1+m\widehat{\theta}))$ using cluster bootstrap sampling. In other words, find a bootstrap estimator of the standard error of $\ln(1+m\widehat{\theta})$ in order to construct an approximate confidence interval for θ. This procedure makes use of the standard bootstrap approach to construct a confidence interval accompanied by a transformation from $\ln(1+m\theta)$ to θ. Let $E^*(\text{SSE}^*)$, $E^*(\text{SSA}^*)$, $\text{var}^*(\text{SSE}^*)$, $\text{var}^*(\text{SSA}^*)$, and

$\operatorname{cov}^*(\mathrm{SSA}^*, \mathrm{SSE}^*)$ denote the cluster bootstrap expectations, variances, and covariances associated with the bootstrap sums of squares SSE^* and SSA^*. Using the results of Field and Welsh (2007),

$$E^*\big(\mathrm{SSE}^*\big) = \mathrm{SSE},$$

$$E^*\big(\mathrm{SSA}^*\big) = \frac{n-1}{n}\mathrm{SSA},$$

$$\operatorname{var}^*\big(\mathrm{SSE}^*\big) = \sum_{i=1}^{n}\left[\sum_{j=1}^{m}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2\right]^2 - \frac{1}{n}\mathrm{SSE}^2,$$

$$\operatorname{var}^*\big(\mathrm{SSA}^*\big) = \frac{(n-1)^2}{n^2}\sum_{i=1}^{n}\left[m\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2\right]^2 - \frac{(n-1)(n-3)}{n^3}\mathrm{SSA}^2$$

and

$$\operatorname{cov}^*\big(\mathrm{SSA}^*, \mathrm{SSE}^*\big) = \frac{n-1}{n}\sum_{i=1}^{n}\left[m\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2\sum_{j=1}^{m}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2\right] - \frac{n-1}{n^2}(\mathrm{SSA})(\mathrm{SSE}).$$

Furthermore, one can show that

$$E\big(\operatorname{var}^*\big(\mathrm{SSE}^*\big)\big) = \frac{n-1}{n}\operatorname{var}(\mathrm{SSE}),$$

$$E\big(\operatorname{var}^*\big(\mathrm{SSA}^*\big)\big) = \left(\frac{n-1}{n}\right)^3\operatorname{var}(\mathrm{SSA}) + \frac{2(n-1)}{n^3}E\big(\mathrm{SSA}^2\big)$$

and

$$E\big(\operatorname{cov}^*\big(\mathrm{SSA}^*, \mathrm{SSE}^*\big)\big) = \left(\frac{n-1}{n}\right)^2\operatorname{cov}(\mathrm{SSA}, \mathrm{SSE}).$$

As described by Field and Welsh (2007), the bootstrap estimators of means, variances and covariances are asymptotically correct. Using their definition, a_n is asymptotic to b_n means that $a_n/b_n \to 1$ as $n \to \infty$. We define the bootstrap estimator of $\operatorname{var}(\ln(1+m\widehat{\theta}))$ to be

$$\operatorname{var}^*\big(\ln\big(1+m\widehat{\theta}^*\big)\big) = \frac{\operatorname{var}^*(\mathrm{SSE}^*)}{\left(E^*(\mathrm{SSE}^*)\right)^2} - \frac{2\operatorname{cov}^*(\mathrm{SSE}^*, \mathrm{SSA}^*)}{E^*(\mathrm{SSE}^*)E^*(\mathrm{SSA}^*)} + \frac{\operatorname{var}^*(\mathrm{SSA}^*)}{\left(E^*(\mathrm{SSA}^*)\right)^2}.$$

The cluster bootstrap sampling procedure produces a nonparametric estimator of $\operatorname{var}(\ln(1+m\widehat{\theta}))$. Also see Ukoumunne et al. (2003) for a similar approach. Note that since $E^*(\mathrm{SSE}^*)$, $E^*(\mathrm{SSA}^*)$, $\operatorname{var}^*(\mathrm{SSE}^*)$, $\operatorname{var}^*(\mathrm{SSA}^*)$, and $\operatorname{cov}^*(\mathrm{SSA}^*, \mathrm{SSE}^*)$ are asymptotically correct, it follows that $\operatorname{var}^*(\ln(1+m\widehat{\theta}^*))$ is an asymptotically correct estimator. This result does not depend on distribution assumptions and is valid as $n \to \infty$ with m fixed. In addition, the bootstrap estimator of $\operatorname{var}(\ln(1+m\widehat{\theta}))$ is available in closed form so the investigator does not need to rely on Monte Carlo approximations. Using this approach, an equal-tailed $100(1-\alpha)$ % bootstrap standard confidence interval for θ is

$$\left(\frac{1}{m}\left[(1+m\widehat{\theta})e^{z_{\alpha/2}\sqrt{\operatorname{var}^*(\ln(1+m\widehat{\theta}^*))}} - 1\right], \frac{1}{m}\left[(1+m\widehat{\theta})e^{z_{1-\alpha/2}\sqrt{\operatorname{var}^*(\ln(1+m\widehat{\theta}^*))}} - 1\right]\right), \tag{4.1}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal distribution, respectively. The corresponding confidence interval for ρ is readily available. That is, if (L, U) represents a confidence interval for θ, then a confidence interval for ρ is given by $(L/(L+1), U/(U+1))$.

It is well known that bootstrap confidence interval procedures are valid if the bootstrap principle holds. In other words, if the bootstrap distribution of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ converges to the sampling distribution of $\sqrt{n}(\widehat{\theta} - \theta)$ under large-sample theory (where m is fixed and $n \to \infty$), then bootstrap confidence intervals are asymptotically correct. See Field, Pang, and Welsh (2010) and Field and Welsh (2007). In particular, Theorem 2 of Field, Pang, and Welsh (2010) is satisfied for cluster bootstrap sampling and we conclude that $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) \approx \sqrt{n}(\widehat{\theta} - \theta)$ in distribution when n is large.

Efron and Tibshirani (1986) indicate that the percentile method works well when there exists a monotonic function of θ, say $g(\theta)$, such that $g(\widehat{\theta}) \overset{\text{asymp}}{\sim} N(g(\theta), \tau^2)$ where $\tau^2 = \text{var}(g(\widehat{\theta}))$ does not depend on θ. Applying the bootstrap principle to the percentile framework yields $g(\widehat{\theta}^*) \overset{\text{approx}}{\sim} N(g(\widehat{\theta}), \tau^2)$. Let $G(x) = P^*(\widehat{\theta}^* \le x)$ be the bootstrap cumulative distribution function of $\widehat{\theta}^*$. The equal-tailed $100(1-\alpha)$ % bootstrap percentile confidence interval for θ is simply

$$\left(G^{-1}(\alpha/2), G^{-1}(1-\alpha/2)\right), \tag{4.2}$$

where $G^{-1}(\alpha/2)$ and $G^{-1}(1-\alpha/2)$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution of $\widehat{\theta}^*$, respectively.

The quality of the intervals derived from the bootstrap percentile method depend in part on the magnitude of n since large-sample theory has been employed. In practice G is unknown, so one typically uses a Monte Carlo simulation to approximate the distribution of $\widehat{\theta}^*$. Let $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$ denote values of $\widehat{\theta}^*$ from B bootstrap replications. The empirical cumulative distribution function of $\widehat{\theta}^*$, denoted by $\widehat{G}$, is used in place of G to compute the endpoints of the bootstrap percentile confidence interval for θ. Furthermore, a confident interval for ρ is readily available since the percentile method is transformation-respecting.

Efron and Tibshirani (1986) suggest that the bias-corrected and accelerated (BC_a) method works well when there exists a monotonic function $g(\theta)$ such that $g(\widehat{\theta}) \overset{\text{asymp}}{\sim} N(g(\theta) - z_0\tau, \tau^2)$, where z_0 is a bias-correction parameter and $\tau^2 = \text{var}(g(\widehat{\theta}))$ depends on θ in the form of $\tau = 1 + ag(\theta)$. The acceleration parameter a measures the change in the standard error of $g(\widehat{\theta})$ as $g(\theta)$ changes. Confidence intervals for θ can be built from the bootstrap distribution of $\widehat{\theta}^*$ provided estimates of the bias-correction and acceleration parameters are available. See Efron (1987) for the development of the BC_a method which is in response to Schenker's (1985) criticism of the precursor bias-corrected (BC) method.

Applying the bootstrap principle to the BC_a framework yields $g(\widehat{\theta}^*) \overset{\text{approx}}{\sim} N(g(\widehat{\theta}) - z_0(1 + ag(\widehat{\theta})), (1 + ag(\widehat{\theta}))^2)$. In this manner a nonzero bias and a standard error which is a linear function of g are incorporated in the procedure. Efron's (1987) bootstrap BC_a confidence interval, which is also discussed by Shao and Tu (1995) as well as other authors, is given by

$$\left(G^{-1}\left(\Phi\left\{\frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})} + z_0\right\}\right), G^{-1}\left(\Phi\left\{\frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})} + z_0\right\}\right)\right), \tag{4.3}$$

where Φ is the cumulative distribution function of the standard normal distribution. As previously stated, the empirical cumulative distribution function, $\widehat{G}$, is used in place of G so that the endpoints of the bootstrap BC_a confidence interval for θ can be computed. Note that the appropriate quantiles of the bootstrap distribution of $\widehat{\theta}^*$ now depend on z_0 and a.

While the BC_a method accounts for specific bias and nonconstant variance features of an estimator, z_0 and a must be estimated in order to implement the procedure. It is known that $z_0 = \Phi^{-1}(G(\widehat{\theta}))$ and thus $\widehat{z}_0 = \Phi^{-1}(\widehat{G}(\widehat{\theta}))$. Furthermore, a commonly used nonparametric estimator of a is

$$\widehat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} (\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^3}{[\sum_{i=1}^{n} (\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)})^2]^{3/2}},$$

where in our case $\widehat{\theta}_{(i)}$ is the REML estimator of θ computed using $(\mathbf{Y}_1, \ldots, \mathbf{Y}_{i-1}, \mathbf{Y}_{i+1}, \ldots, \mathbf{Y}_n)$ and $\widehat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \widehat{\theta}_{(i)}/n$. Both z_0 and a are order of magnitude $O(n^{-1/2})$ in the sample size n. See Efron (1987) and DiCiccio and Efron (1996) for details. A confident interval for ρ is readily available since the BC_a method is invariant under reparameterization.

5. SIMULATION STUDY: CONFIDENCE INTERVALS FOR ρ

The confidence intervals for θ using the pivotal quantity and the three nonparametric bootstrap methods, namely, Equations (3.2), (4.1), (4.2), and (4.3), are converted to confidence intervals for ρ and the simulated coverage probabilities are computed for a variety of scenarios. The values of ρ in the simulations include 0.10, 0.25, 0.50, 0.75, and 0.90. Since asymptotic results hold as n increases for m fixed, let $n = 10$, 25, 50 and 100, and consider $m = 5$. The overall sample size ($N = nm$) in this study ranges from 50 to 500. Without loss of generality, let $\mu = 100$. Furthermore, we begin by assuming that $A_i \overset{\text{iid}}{\sim} N(0, \sigma_1^2)$, $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_2^2)$, where A_i and e_{ij} are mutually independent.

For the simulations in this paper, nonnegativity restrictions on $\widehat{\theta}$, $\widehat{\theta}^*$, and $\widehat{\theta}_{(i)}$ are not imposed so there is the possibility that a simulated confidence interval may have a negative lower bound. This is more likely to occur in a finite sample scenario when n is small and θ is close to zero. The main reason why nonnegative restrictions are not considered in this section is that it is more appropriate to compare the expected lengths of the different confidence interval methods without interval truncation. This approach was also taken by Ukoumunne et al. (2003). A similar tactic was used by Burch (2007) when assessing the performance of generalized confidence intervals. In practice, of course, the investigator can simply report the truncated interval to avoid negative endpoints.

For each combination of ρ, number of classes (n), and class size $m = 5$, we generated 10,000 Monte Carlo random samples and computed simulated coverage probabilities using the pivotal quantity (PQ) and the three nonparametric bootstrap methods. For the bootstrap percentile and bootstrap BC_a methods, the number of bootstrap replications is $B = 2000$. The simulated coverage probabilities for the nominal 95 % confidence intervals for ρ are displayed in Table 2.

Table 2. Simulated coverage probabilities of 95 % nominal confidence intervals for ρ. The results assume $m = 5$ and an underlying normal distribution. The probabilities in the last three columns include a small-sample adjustment.

			Nonparametric bootstrap			Nonparametric bootstrap with small-sample adjustment		
ρ	n	PQ	Standard	Percentile	BC$_a$	Standard	Percentile	BC$_a$
0.10	10	0.949	0.874	0.833	0.862	0.938	0.888	0.919
	25	0.951	0.914	0.898	0.913	0.942	0.925	0.942
	50	0.952	0.932	0.927	0.933	0.947	0.939	0.948
	100	0.948	0.937	0.933	0.939	0.944	0.941	0.945
0.25	10	0.951	0.873	0.833	0.865	0.939	0.892	0.919
	25	0.951	0.915	0.899	0.914	0.942	0.926	0.942
	50	0.954	0.931	0.926	0.934	0.947	0.939	0.948
	100	0.950	0.938	0.934	0.939	0.946	0.941	0.945
0.50	10	0.950	0.868	0.832	0.858	0.942	0.888	0.919
	25	0.952	0.915	0.897	0.918	0.943	0.925	0.944
	50	0.953	0.933	0.924	0.934	0.946	0.938	0.947
	100	0.948	0.938	0.934	0.939	0.946	0.941	0.945
0.75	10	0.951	0.873	0.831	0.863	0.942	0.894	0.920
	25	0.952	0.916	0.897	0.919	0.944	0.923	0.946
	50	0.952	0.932	0.925	0.932	0.945	0.939	0.946
	100	0.947	0.937	0.933	0.938	0.944	0.940	0.945
0.90	10	0.951	0.873	0.836	0.865	0.942	0.898	0.922
	25	0.952	0.917	0.898	0.919	0.946	0.927	0.947
	50	0.953	0.932	0.924	0.932	0.945	0.939	0.946
	100	0.947	0.937	0.931	0.936	0.945	0.939	0.944

When n is large, the pivotal quantity and the three nonparametric bootstrap methods yield similar results in terms of coverage probabilities. In general, the bootstrap standard and bootstrap BC$_a$ methods outperform the bootstrap percentile method for the cases considered. When n is small, however, all of the nonparametric bootstrap methods fall short in terms of coverage probabilities. This is to be expected since the bootstrap methods do not use the fact that A_i and e_{ij} are normally distributed. This prompts one to consider a small-sample adjustment to the bootstrap intervals in an effort to match the results obtained by using the F-distributed pivotal quantity. The simulated coverage probabilities for the modified nonparametric bootstrap methods are included in Table 2 and we now discuss the small-sample adjustment in detail.

For the bootstrap standard procedure, it can be shown that a simple and yet useful approach is to unbiasedly estimate the individual terms $E^*(\text{SSE}^*)$, $E^*(\text{SSA}^*)$, $\text{var}^*(\text{SSE}^*)$, $\text{var}^*(\text{SSA}^*)$, and $\text{cov}^*(\text{SSA}^*, \text{SSE}^*)$ in $\text{var}^*(\ln(1 + m\widehat{\theta}^*))$. The result, ignoring terms of $O(1/n^2)$, is the estimator of $\text{var}(\ln(1+m\widehat{\theta}))$ given by $n/(n-1)\text{var}^*(\ln(1+m\widehat{\theta}^*))$. That is, the original bootstrap estimator of $\text{var}(\ln(1+m\widehat{\theta}))$ is multiplied by $n/(n-1)$. This is similar to multiplying the bootstrap estimator of $\text{var}(\overline{Y})$ by $n/(n-1)$ in the simple case where $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$. While Efron (1982) states that there is no real advantage to using an adjustment in the simple case, our simulation results suggest that the bootstrap standard in-

tervals are too narrow since $\text{var}^*(\ln(1+m\widehat{\theta}^*))$ underestimates $\text{var}(\ln(1+m\widehat{\theta}))$. Of course, any term of the form $(n+c)/(n-1)$, where c is an integer, approaches one as $n \to \infty$ and thus can serve as a small-sample adjustment to the estimator of $\text{var}(\ln(1+m\widehat{\theta}))$. In this paper, the value of c selected is tuned to give adequate coverage results for $n = 25$ as this is very similar to the sample size in the mercury in oyster tissue application. Empirical results indicate that $(n+5)/(n-1)$ serves as a reasonable small-sample adjustment term so that

$$\widehat{\text{var}}\big(\ln(1+m\widehat{\theta})\big) = \frac{n+5}{n-1}\text{var}^*\big(\ln\big(1+m\widehat{\theta}^*\big)\big),$$

and

$$\left(\frac{1}{m}\big[(1+m\widehat{\theta})e^{z_{\alpha/2}\sqrt{\widehat{\text{var}}(\ln(1+m\widehat{\theta}))}}-1\big], \frac{1}{m}\big[(1+m\widehat{\theta})e^{z_{1-\alpha/2}\sqrt{\widehat{\text{var}}(\ln(1+m\widehat{\theta}))}}-1\big]\right) \quad (5.1)$$

is used as the bootstrap standard confidence interval adjusted for finite sample sizes.

The small-sample adjustment term used here is based solely on simulations, gives reasonable confidence interval coverage results for the finite-sample cases considered, and produces confidence intervals that are asymptotically correct. On a related subject, Polansky (2000) presents a theoretically motivated adjustment to the standard error of an estimator to build bootstrap-t confidence intervals. One can show that this adjustment is equivalent to multiplying the bootstrap variance of the estimator by $(n+c)/(n-1)$, where c itself depends on n. While this theory-based small-sample adjustment is useful for bootstrap-t intervals involving the mean or variance, Polansky (2000) notes that its effectiveness is less impressive when considering confidence intervals for the correlation coefficient.

The small-sample adjustment to the variance is also applied to the percentile and BC_a methods. For the bootstrap percentile method, consider $\text{var}^*(g(\widehat{\theta}^*)) = ((n-1)/(n+5))\tau^2$. Then the bootstrap percentile confidence interval adjusted for a finite sample size is

$$\left(\widehat{G}^{-1}\left(\Phi\left\{\sqrt{\frac{n+5}{n-1}}z_{\alpha/2}\right\}\right), \widehat{G}^{-1}\left(\Phi\left\{\sqrt{\frac{n+5}{n-1}}z_{1-\alpha/2}\right\}\right)\right). \quad (5.2)$$

Similarly, for the bootstrap BC_a method, consider $\text{var}^*(g(\widehat{\theta}^*)) = ((n-1)/(n+5))(1+ag(\widehat{\theta}))^2$. It follows that the corresponding bootstrap BC_a interval adjusted for a finite sample size is

$$\left(\widehat{G}^{-1}\left(\Phi\left\{\sqrt{\frac{n+5}{n-1}}\left(\frac{z_0+z_{\alpha/2}}{1-a(z_0+z_{\alpha/2})}+z_0\right)\right\}\right),\right.$$
$$\left.\widehat{G}^{-1}\left(\Phi\left\{\sqrt{\frac{n+5}{n-1}}\left(\frac{z_0+z_{1-\alpha/2}}{1-a(z_0+z_{1-\alpha/2})}+z_0\right)\right\}\right)\right). \quad (5.3)$$

The simulated coverage probabilities given in Table 2 indicate that incorporating a small-sample adjustment to the nonparametric bootstrap methods can help improve performance. The modification assists the three bootstrap methods in a similar manner so that the bootstrap standard and bootstrap BC_a methods still outperform the bootstrap percentile method for the cases considered. Furthermore, when $n \geq 25$, the bootstrap standard

and bootstrap BC_a methods offer results that are similar to those of the F-distributed pivotal quantity. Including a small-sample adjustment term enables the bootstrap standard and bootstrap BC_a methods to more successfully approximate the normal theory results for small to intermediate sample sizes. In this paper the bootstrap percentile method will no longer be considered as a viable candidate since its performance is overshadowed by the performances of the bootstrap standard and bootstrap BC_a methods.

While not explicitly displayed in the terms of var(SSE), var(SSA), and cov(SSE, SSA), the asymptotic variance of the REML estimator of θ depends in part on the kurtosis of A_i and e_{ij}, which is denoted by κ_1 and κ_2, respectively. If A_i and e_{ij} are normally distributed, then $\kappa_1 = \kappa_2 = 0$. If κ_1 and κ_2 are positive, then the coverage probabilities of confidence intervals using the pivotal quantity method may fall short of the nominal coverage value of 0.95. In contrast, if κ_1 and κ_2 are negative, then the coverage probabilities of confidence intervals using the pivotal quantity method may exceed the nominal coverage value of 0.95. See Burch (2011) for details.

For examples of nonzero kurtosis values, the random variable $X \sim \text{Gamma}(2, 1)$ has $\kappa = 3$. For more extreme kurtosis values, $X \sim \text{t}(5)$ has $\kappa = 6$ and $X \sim \text{lognormal}(0, 0.395)$ has $\kappa = 12$. Alternatively, if $X \sim \text{Beta}(1/3, 2/3)$, then $\kappa = -1$. We will consider these asymmetric and symmetric distributions having various κ-values in the following manner. Suppose that A_i and e_{ij} are built from the same underlying distribution. Furthermore, if the underlying distribution does not have the appropriate mean and/or variance for a particular simulation, then the distribution is simply located and scaled so that it does. For instance, if one considers the random variable $X_i \sim \text{Gamma}(2, 1)$ for factor A_i, then let $A_i = (\sigma_1/\sqrt{2})(X_i - 2)$, where $i = 1, \ldots, n$. In this manner $E(A_i) = 0$ and $\text{var}(A_i) = \sigma_1^2$.

Simulated coverage probabilities using the pivotal quantity, standard, and BC_a methods are computed assuming the underlying distribution is either Beta(1/3, 2/3), N(0, 1), Gamma(2, 1), t(5), or lognormal(0, 0.395). Consider $\rho = 0.1$, 0.5, and 0.9 with the number of classes (n) equal to 10, 25, 50, and 100. Assume a fixed class size of $m = 5$. The different combinations of distributions, ρ-values, and number of classes (n) gives 60 scenarios in which to evaluate the performances of the three confidence interval procedures currently under study. The coverage probabilities of the 95 % confidence intervals for ρ are displayed in Figure 2. The top panel in Figure 2 summarizes the results for $\rho = 0.1$. The nominal coverage probability of 0.95 is depicted by the horizontal dashed line. The middle and bottom panels of Figure 2 show the simulated coverage probabilities for $\rho = 0.5$ and $\rho = 0.9$, respectively.

Figure 2 suggests that the pivotal quantity, standard, and BC_a methods yield somewhat similar results when $\rho = 0.1$ for the cases considered. Note that there is an overall trend of decreasing coverage probabilities as the kurtosis of the underlying distribution increases. Also, when $n = 10$, the performance of the BC_a method often trails that of the other two methods. The middle and bottom panels of Figure 2, where $\rho = 0.5$ and 0.9, respectively, display cases where the pivotal quantity and the nonparametric bootstrap procedures produce markedly different results. The pivotal quantity confidence intervals exhibit overcoverage when the kurtosis is less than zero and severe undercoverage when the kurtosis is greater than zero. Furthermore, for cases where the kurtosis is greater than zero, the pivotal quantity's underperformance becomes more pronounced as the sample size increases.

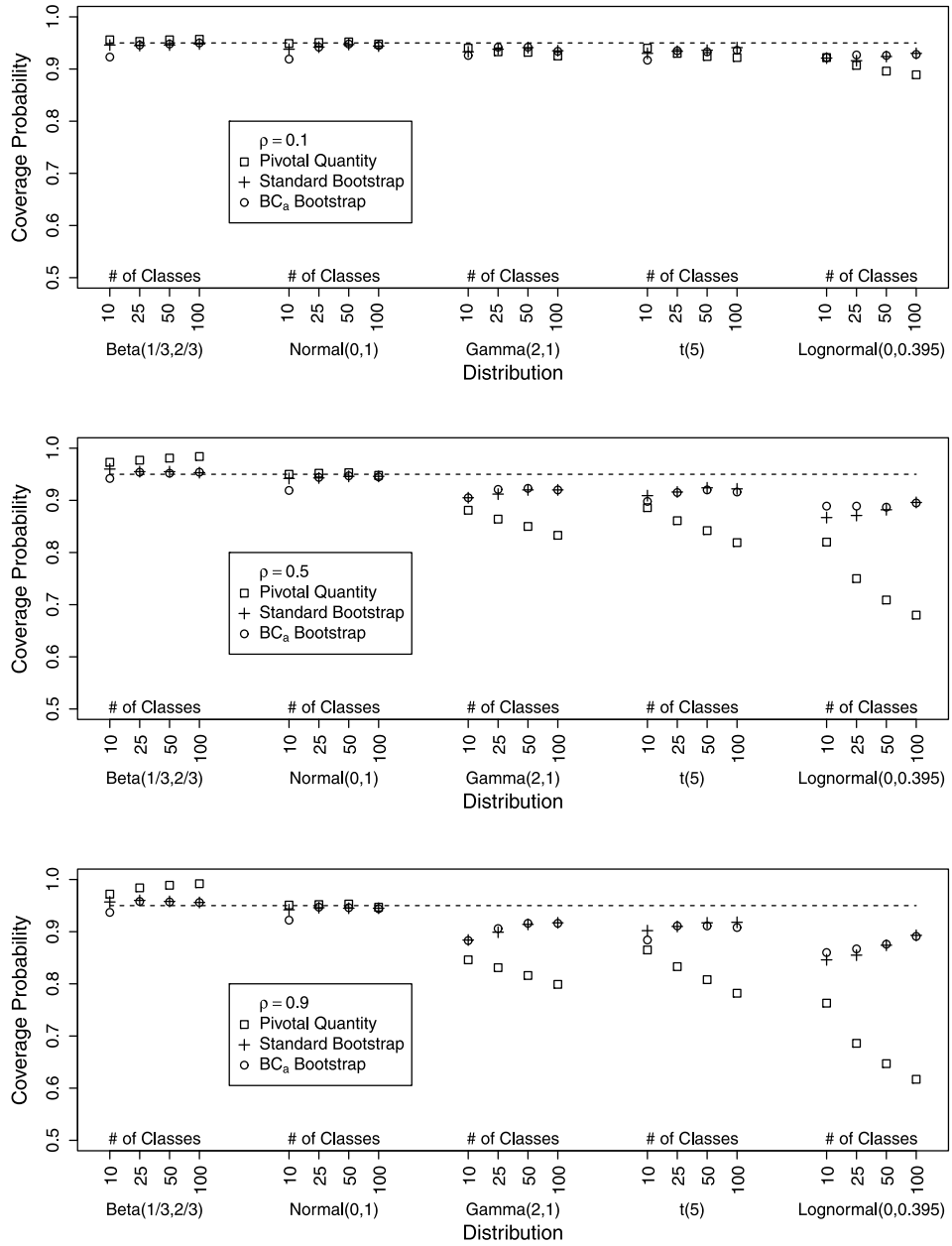

Figure 2. Simulated coverage probabilities of 95 % nominal confidence intervals for $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ when the true value of ρ is equal to 0.1, 0.5, and 0.9. Results are for the pivotal quantity, standard bootstrap, and BC$_a$ bootstrap procedures using number of classes equal to 10, 25, 50, and 100 for a variety of distributions. The nonparametric bootstrap procedures include a small-sample adjustment.

These seemingly errant results are simply due to the fact that pivotal quantity itself is not based on large-sample theory. As was the case for $\rho = 0.1$, the lower two panels in Figures 2 show a trend of decreasing coverage probabilities as the kurtosis of the underlying

distribution increases. Also note that the standard and BC_a methods yield similar coverage probabilities for $n \geq 25$.

It is clear from the bottom panel of Figure 2 that the pivotal quantity method has difficulty maintaining the nominal coverage probability of 0.95 for cases where both ρ and the kurtosis are large. For example, when $\rho = 0.9$, $n = 25$, $m = 5$, and using the t(5) distribution, the simulated coverage probability for the pivotal quantity method is 0.833. In contrast, the simulated coverage probabilities for the bootstrap standard and bootstrap BC_a methods are 0.910 and 0.911, respectively. The lognormal$(0, 0.395)$ distribution yields even more disparate simulated coverage probabilities when comparing the pivotal quantity and nonparametric bootstrap methods.

6. REVISITING THE MERCURY IN OYSTER TISSUE EXAMPLE

In the mercury in oyster tissue application, $\widehat{\theta} = 4.11$ and $\widehat{\rho} = 0.80$. The 95 % confidence interval estimates for ρ using the pivotal quantity, standard, and BC_a procedures are $(0.69, 0.89)$, $(0.60, 0.92)$, and $(0.58, 0.91)$, respectively. The pivotal quantity interval for ρ comes from the appropriate transformation of the endpoints of the interval in (3.2). The standard interval for ρ is related to the standard interval for θ, which is given in (5.1). Similarly, the BC_a interval for ρ comes from the appropriate transformation of the endpoints of the interval in (5.3). Note that the length of the pivotal quantity interval is relatively narrow since it is less than two thirds of the length of the nonparametric bootstrap intervals.

Recall that the pivotal quantity confidence interval for ρ assumes that A_i and e_{ij} are normally distributed. While one cannot directly ascertain whether or not the normal distribution assumptions are valid in this application, we computed $\widehat{A}_i = \overline{Y}_{i.} - \overline{Y}_{..}$, $i = 1, \ldots, 26$ and $\widehat{e}_{ij} = Y_{ij} - \overline{Y}_{i.}$, $i = 1, \ldots, 26$, $j = 1, \ldots, 5$, and considered their distributions in lieu of the distributions of A_i and e_{ij}. Both distributions are asymmetric with estimated kurtosis values of $\widehat{\kappa_1} = 4.86$ and $\widehat{\kappa_2} = 3.98$. The estimator of kurtosis used here for κ_1 and κ_2 is known as G_2. See Joanes and Gill (1998) for details. Furthermore, Shapiro–Wilk tests for normality indicate that the distributions of $\widehat{A}_i$ and $\widehat{e}_{ij}$ are not normal. All indications suggest that it is very unlikely that the true confidence level of the pivotal quantity procedure is 95 %. That is, the underlying kurtosis values are substantially greater than zero so that the actual coverage probabilities of confidence interval procedures that rely on normal theory will fall short of the nominal level.

In this example, where there is evidence that ρ is large and the kurtosis values are greater than zero, the standard and BC_a confidence interval coverage probabilities are much more likely to be closer to the stated coverage probability. In summary, an investigator can be confident in concluding that likely values of ρ range from about 0.6 to 0.9. In other words, roughly 60 to 90 % of the variation in the mercury concentration is due to the variability amongst the laboratories. Interlaboratory variation of this magnitude indicates that the measurements are not consistent across laboratories.

7. DISCUSSION

Nonparametric bootstrap methods, augmented with a small-sample adjustment, are viable options when constructing confidence intervals for θ or ρ using a variety of underlying distributions. In particular, the BC_a method exhibited reasonable simulated coverage probability results when $n \geq 25$. It is the most automated interval procedure considered in this paper in that the investigator is not required to specify a particular transformation to construct the interval. In contrast, the bootstrap percentile method is not the best choice for constructing confidence intervals for θ or ρ since the simulations suggest that $\tau^2 = \text{var}(g(\widehat{\theta}))$ is not free of the parameter θ. Even with the small-sample adjustment, the bootstrap percentile intervals did not produce acceptable coverage probabilities for small values of n. The bootstrap standard method, using a logarithmic transformation and a small-sample adjustment, performed well for the most of the cases considered. Interestingly, simulation results indicate that this method can also outperform the BC_a method when $n = 10$.

In a mildly unbalanced one-way random effects design, the modified bootstrap intervals maintain their simulated coverage probabilities. Consider a design with $n = 25$ where twenty of the classes have $m = 5$ and five classes have $m = 4$. Let $\rho = 0.5$ and suppose the underlying distribution is t(5). The bootstrap BC_a coverage value of 0.915 for the balanced design (where all twenty-five classes have $m = 5$) is now 0.918 for the unbalanced design. If the underlying distribution is lognormal$(0, 0.395)$, then the bootstrap BC_a coverage value of 0.889 for the balanced design becomes 0.885 for the unbalanced design. The usefulness of the small-sample adjustment appears to carry over to mildly unbalanced designs.

An investigator, however, should be cautious when utilizing bootstrap confidence intervals in small sample size applications when the underlying kurtosis is extremely large. Schenker (1985), Good and Chernick (1993), Hall and Padmanabhan (1997), Chernick and LaBudde (2010, 2011), Sun, Chernick, and LaBudde (2011), and Good (2011) comment on the underperformance of bootstrap intervals, in terms of coverage probability, when the parameter under study is the variance or a ratio of variances. In particular, a simulation study by Chernick and LaBudde (2010) using $X \sim$ lognormal$(0, 1)$ where $\kappa = 110.94$ shows that bootstrap interval coverage probabilities for variances can fall far below nominal coverage values.

In our study, consider the scenario where $\rho = 0.5$, $n = 25$, and $m = 5$. The bootstrap standard, percentile, and BC_a nominal 95 % confidence intervals for ρ using an underlying lognormal$(0, 1)$ distribution have simulated coverage probabilities of 0.762, 0.834, and 0.825, respectively. The poor performance of the standard method arises from the fact that estimating $\text{var}(\ln(1 + m\widehat{\theta}))$ for this case is challenging. The degraded performance of the BC_a method is in part due to the difficulty of accurately estimating a and z_0 in small-sample problems involving the lognormal$(0, 1)$ distribution. While the performance of the percentile method now rivals that of the BC_a method and the standard method's performance lags behind, none of the modified bootstrap procedures considered here produce satisfactory results.

The pivotal quantity method, which is based on the normal distribution as opposed to large-sample theory, performed erratically in terms of coverage probability under nonnormal conditions. The poor performance is magnified in those scenarios where ρ is intermediate to large in value and the underlying kurtosis is large. If the underlying distribution is truly not normal, then using normal-based pivotal quantities to make inferences on functions of variance components can lead to erroneous conclusions which cannot be corrected by increasing the size of the sample. However, if it is known that ρ is small (see the top panel of Figure 2), in many scenarios the pivotal quantity method may result in coverage probabilities that are similar to those of the other confidence interval procedures.

The bootstrap confidence interval procedures considered in this paper depend on the REML estimator of the parameter under study. Under regularity conditions, the bootstrap principle holds and the large-sample properties of the REML estimator are well known. Combining these two statistical methods, in conjunction with a small-sample adjustment, enables the investigator to pursue methods to construct nonparametric confidence intervals for ratios of variance components in a number of finite sample applications.

ACKNOWLEDGEMENTS

The author thanks the two anonymous reviewers for their valuable suggestions which substantially improved the quality of the article.

[Received November 2011. Accepted March 2012. Published Online April 2012.]

REFERENCES

Burch, B. D. (2007), "Generalized Confidence Intervals for Proportions of Total Variance in Mixed Linear Models," *Journal of Statistical Planning and Inference*, 137, 2394–2404.

——— (2011), "Assessing the Performance of Normal-Based and REML-Based Confidence Intervals for the Intraclass Correlation Coefficient," *Computational Statistics & Data Analysis*, 55, 1018–1028.

Burch, B. D., and Harris, I. R. (2001), "Closed-Form Approximations to the REML Estimator of a Variance Ratio (or Heritability) in a Mixed Linear Model," *Biometrics*, 57, 1148–1156.

Burch, B. D., and Iyer, H. K. (1997), "Exact Confidence Intervals for a Variance Ratio (or Heritability) in a Mixed Linear Model," *Biometrics*, 53, 1318–1333.

Chernick, M. R. (2008), *Bootstrap Methods: A Guide for Practitioners and Researchers*, Hoboken, NJ: Wiley.

Chernick, M. R., and LaBudde, R. A. (2010), "Revisiting Qualms about Bootstrap Confidence Intervals," *American Journal of Mathematical and Management Sciences*, 29, 437–456.

——— (2011), *Bootstrap Methods With Applications to R*, Hoboken, NJ: Wiley.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, Cambridge, UK: Cambridge University Press.

DiCiccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–228.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: SIAM.

——— (1987), "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 82, 171–185.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54–77.

Field, C. A., and Welsh, A. H. (2007), "Bootstrapping Clustered Data," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 69, 369–390.

Field, C. A., Pang, Z., and Welsh, A. H. (2010), "Bootstrapping Robust Estimates for Clustered Data," *Journal of the American Statistical Association*, 105, 1606–1616.

Good, P. I. (2011), *A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling*, Boca Raton, FL: Chapman & Hall/CRC Press.

Good, P., and Chernick, M. (1993), "Testing the Equality of Variances of Two Populations," Unpublished manuscript.

Hall, P., and Padmanabhan, A. R. (1997), "Adaptive Inference for the Two Sample Problem," *Technometrics*, 39, 412–422.

Harville, D. A. (2004), "Making REML Computationally Feasible for Large Data Sets: Use of the Gibbs Sampler," *Journal of Statistical Computation and Simulation*, 74, 135–153.

Harville, D. A., and Fenech, A. P. (1985), "Confidence Intervals for a Variance Ratio, or Heritability, in an Unbalanced Mixed Linear Model," *Biometrics*, 41, 137–152.

Jiang, J. (1996), "REML Estimation: Asymptotic Behavior and Related Topics," *Annals of Statistics*, 24, 255–286.

——— (2005), "Partially Observed Information and Inference About Non-Gaussian Mixed Linear Models," *Annals of Statistics*, 33, 2695–2731.

Joanes, D. N., and Gill, C. A. (1998), "Comparing Measures of Sample Skewness and Kurtosis," *The Statistician*, 47, 183–189.

Polansky, A. M. (2000), "Stabilizing Bootstrap-t Confidence Intervals for Small Samples," *Canadian Journal of Statistics*, 28, 501–516.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., and Lai, S. (2010), "Nonparametric bootstrapping for hierarchical data," *Journal of Applied Statistics*, 37, 1487–1498.

Schenker, N. (1985), "Qualms about Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 80, 360–361.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer.

Sun, J., Chernick, M. R., and LaBudde, R. A. (2011), "A Bootstrap Test for Comparing Two Variances: Simulation of Size and Power in Small Samples," *Journal of Biopharmaceutical Statistics*, 21, 1079–1093.

Ukoumunne, O., Davison, A. C., Gulliford, M. C., and Chinn, S. (2003), "Non-parametric Bootstrap Confidence Intervals for the Intraclass Correlation Coefficient," *Statistics in Medicine*, 22, 3805–3821.

West, B. T., Welch, K. B., and Galecki, A. T. (2007), *Linear Mixed Models: A Practical Guide Using Statistical Software*, Boca Raton, FL: Chapman and Hall/CRC Press.

Westfall, P. H. (1987), "A Comparison of Variance Component Estimates for Arbitrary Underlying Distributions," *Journal of the American Statistical Association*, 82, 866–874.

Willie, S., and Berman, S. (1995), "NOAA National Status and Trends Program Ninth Round Intercomparison Exercise Results for Trace Metals in Marine Sediments and Biological Tissues," NOAA Technical Memorandum ORCA 93, pp. 52 plus appendices.