1. INTRODUCTION

A conscientious scientist returns from lunch excited about a statistical analysis of some new data carried out that morning revealing a scientifically important result. The scientist decides to double check the analysis which uses a state-of-the-art *bootstrap* method, but is disappointed to find that the result is no longer significant. A further run of the analysis produces yet another P-value. Perplexed, the scientist calls a statistician who explains that the fluctuation in P-values is due to Monte Carlo error inherent to the implementation of the bootstrap method. At this point the scientist decides to use another, more reliable, statistical procedure.

In this article, the fluctuation in P-values observed by the scientist when repeatedly analyzing the same data is explained in terms of the number of resamples used to approximate a bootstrap variance. We argue that the number of resamples required for this purpose is far larger than is generally thought. We begin with a brief description of bootstrap variance estimation and its corresponding Monte Carlo approximation.

Suppose that $Y_1, \ldots, Y_n$ is a random sample from an unknown distribution $F$ and let $\hat{F}$ denote an estimate of $F$ or fitted distribution. For example, $F$ might be known to be a member of a parametric family, $\{F_\lambda : \lambda \in \Lambda\}$, in which case $\hat{F} = F_\hat{\lambda}$, where $\hat{\lambda}$ is an estimate of the parameter $\lambda$ based on the sample. The bootstrap estimate of an arbitrary functional, $T(F)$, is then given by $T(\hat{F})$. Most of the literature on the bootstrap concerns the nonparametric situation, where $F$ is completely unknown and $\hat{F}$ is the empirical distribution function, $\hat{F}(y) = \#\{i : Y_i \leq y\}/n$.

Now, suppose that $\hat{\theta}$ is an estimator of a scalar characteristic $\theta = \theta(F)$ based on the sample. Then the sampling variance of $\hat{\theta}$ is given by $\sigma^2(\hat{\theta}) = E\{(\hat{\theta} - E(\hat{\theta}))^2\}$. The bootstrap estimate of $\sigma^2$ is therefore

$$\hat{\sigma}^2 = \sigma^2(\hat{\theta}) = \sigma^2\left(\hat{\theta} = E^*\{\hat{\theta} - E^*(\hat{\theta})\}^2\right),$$

where $E^*$ denotes expectation with respect to $\hat{F}$ (i.e., conditional on the observed sample) and $\hat{\theta}^*$ is the version of $\hat{\theta}$ computed using a sample drawn from $\hat{F}$. Samples drawn from the fitted distribution are referred to as resamples in the bootstrap literature. In particular, if $\hat{F}$ is the empirical distribution, then resamples are obtained by sampling with replacement from the original sample.

In most problems the bootstrap variance formula in (1) is analytically intractable. In such cases a Monte Carlo approximation to $\hat{\sigma}^2$ is obtained as

$$\hat{\sigma}_B^2 = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_b)^2,$$

where $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ denote versions of $\hat{\theta}$ computed using $B$ independent resamples drawn from $\hat{F}$ and $\hat{\theta}_b^* = B^{-1} \sum \hat{\theta}_b$.

In this article we address the issue of the number of resamples required for approximating bootstrap variances. One argument, due to Efron (1987), is based on the unconditional coefficient of variation of $\hat{\sigma}_B^2$ which involves both sampling and resampling variability. His analysis suggests that only a small number of resamples are required to give reasonable answers, a result that has been widely quoted in the statistical literature ever since. In the remainder of this article we argue that only resampling variability should be considered when choosing $B$. The adoption of this principle leads to the conclusion that much larger values of $B$ are
required in practice than is suggested by the unconditional argument.

2. A CONDITIONAL CRITERION FOR CHOOSING $B$

We contend that resampling methods will be trusted by practitioners only if the Monte Carlo error incurred has little or no impact on their conclusions. In the context of Monte Carlo approximation of a bootstrap variance this will happen only if the relative error of $\hat{\sigma}_B^2$ (hereafter referred to as relative error) due to resampling is small; that is,

$$1 - \delta < \frac{\hat{\sigma}_B^2}{\sigma^2} < 1 + \delta$$  \hspace{1cm} (3)

for a small positive $\delta$. In practice, since $\hat{\sigma}_B^2$ is a random variable, we can only ask that (3) hold with high probability. Thus, we aim to choose $B$ such that

$$1 - \alpha = P^*(1 - \delta < \frac{\hat{\sigma}_B^2}{\sigma^2} < 1 + \delta)$$  \hspace{1cm} (4)

for some small probability $\alpha$, where $P^*$ denotes probability computed under $F$ (i.e., conditional on the observed sample).

To put this in perspective, consider the effect of Monte Carlo error on the P-value of a hypothesis test of $\theta = 0$ versus $\theta \neq 0$ based on the approximate standard normal pivot $Z = \hat{\theta}/\hat{\sigma}$. If $\hat{\sigma}$ is replaced by $\hat{\sigma}_B$ satisfying (3), then the range of possible values of the resulting test statistic is

$$Z\sqrt{1 - \delta} < Z_B = \frac{\hat{\theta}}{\hat{\sigma}_B} < Z\sqrt{1 + \delta}.$$  \hspace{1cm} (5)

Suppose that $Z = 1.96$ so that the true P-value is .05. The probable ranges of approximate P-values obtained due to Monte Carlo error are displayed in Figure 1 for $\delta$ between 0 and .75. (The actual values are given for $\delta = .1, .2$ and .5.) Thus, a 10% relative error requirement for $\hat{\sigma}_B^2$ ensures that the conclusion of the statistical test is relatively unaffected by Monte Carlo error in the sense that the P-value remains in the borderline significant range, .04 to .06. In contrast, if the relative error of $\hat{\sigma}_B^2$ can reach 50% then the conclusion of the analysis is essentially determined by the seed of a random number generator. We will show in the next section that approximately 800 resamples are required for the relative error to be less than 10% with probability .95.

3. A SIMPLE FORMULA FOR $B$

Suppose that the sampling distribution of $\hat{\theta}$ is approximately normal. (We emphasize the word approximately here because our objective is to determine a rough formula for $B$ which may be applied in a wide range of problems. If the exact sampling distribution of $\theta$ is known, or a higher order approximation is available, it may be possible to refine the arguments presented in the following.) Under mild regularity conditions, the resampling distribution of $\hat{\theta}^*$ will also be approximately normal. Since $\hat{\theta}_1^*, \ldots, \hat{\theta}_{n_B}^*$ is a sample from a distribution with variance $\hat{\sigma}^2$, it follows that $(B - 1)\hat{\sigma}_{B}^2/\hat{\sigma}^2$ approximately has a chi-squared distribution with $(B - 1)$ degrees of freedom. Moreover, since $B$ is large, we can ignore the difference between it and $B - 1$. Thus, using the normal approximation to the chi-squared, we obtain the approximation

$$P^*\left(1 - \delta < \frac{\hat{\sigma}_B^2}{\sigma^2} < 1 + \delta\right) \approx P\left(B(1 - \delta) < \chi_B^2 < B(1 + \delta)\right) \approx P\left(B(1 - \delta) < B\right) + \sqrt{2BZ} < B(1 + \delta) \approx 1 - 2\Phi\left(-\sqrt{\frac{B}{2\delta}}\right).$$  \hspace{1cm} (6)

Combining equations (4) and (5) reveals the formula

$$B \approx \frac{2}{\delta^2} \Phi^{-1}\left(\frac{\alpha}{2}\right)^2.$$  \hspace{1cm} (7)

It follows, for example, that approximately 800 resamples ($B = 2(1.96^2)/.1 = 768$) are required to achieve a relative error less than 10% with probability .95.

To illustrate the accuracy of the chi-squared/normal approximation used in (5), consider the law school data from

![Figure 1. Probable Ranges of Approximate P-Values Due to Monte Carlo Error When the True P-Value is .05](image)

![Figure 2. Law School Data. Histograms of the percentage relative errors of 200 Monte Carlo approximations to the bootstrap variance estimate for the sample correlation along with the corresponding chi-squared density approximations.](image)
Efron (1982). This data set consists of \( n = 15 \) bivariate values with a sample correlation coefficient \( \hat{\theta} = .776 \). (We note that this is a situation in which the normal sampling distribution assumption is questionable.) Histograms of 200 values of the percent relative error \( P_B = 100(\hat{\sigma}_B^2 / \hat{\sigma}^2 - 1) \) are displayed in Figure 2 for \( B = 50 \) and \( B = 800 \). Note that the percent relative error is related to \( X_B = B \hat{\sigma}_B^2 / \hat{\sigma}^2 \) through the location-scale transformation, \( P_B = 100(X_B / B - 1) \). The transformed normal densities corresponding to the two histograms are also displayed on the plot. In both cases the normal approximation appears to be quite good. For example, in our simulation, the relative error was less than 10\% on 62 occasions when \( B = 50 \) and 185 occasions when \( B = 800 \). Both of these numbers are in remarkable agreement with the proportions predicted by the normal approximation. The value of \( \hat{\sigma}^2 \) was taken to be the average of the 200 values in each simulation. In both cases \( \hat{\sigma}^2 = .0179 \) \( (\hat{\theta} = .134) \).

A second example involves the scores on five exams for \( n = 88 \) students taken from Mardia, Kent, and Bibby (1979). Let \( \hat{\theta} \) denote the percentage of variation explained by the first principle component. Efron and Tibshirani approximate the bootstrap variance of \( \hat{\theta} \) using 200 resamples. Figure 3 shows a histogram of percent relative error for 200 Monte Carlo approximations. Once again the chi-square/normal approximation used in (5) is remarkable accurate. Clearly, a relative error of more than 10\% is quite likely using \( B = 200 \) resamples.

4. COEFFICIENT OF VARIATION

The approximation, \( \hat{\theta}^2 / \hat{\sigma}^2 \sim B + \sqrt{2BZ} \), implies that the conditional coefficient of variation of \( \hat{\sigma}_B^2 \) is

\[
\text{CV}^*(\hat{\sigma}_B^2) = \sqrt{\text{var}(\hat{\sigma}_B^2) / \hat{\sigma}_B^2} \approx \sqrt{\frac{2}{B}}.
\]

Also, application of the delta method reveals

\[
\text{CV}^*(\hat{\sigma}_B) \approx \frac{1}{2} \text{CV}^*(\hat{\sigma}_B^2) \approx \sqrt{\frac{1}{2B}}.
\]

Thus, a relative error bound of \( \delta \) with 95\% confidence is equivalent to the bound, \( \text{CV}^*(\hat{\sigma}_B) \leq \frac{1}{2} \delta \), on the conditional coefficient of variation for the bootstrap standard error approximation. In particular, our recommendation of \( \delta = .1 \) is equivalent to \( \text{CV}^*(\hat{\sigma}_B) \leq .025 \).

Our conditional argument is based on a belief that Monte Carlo error should not be allowed to affect the conclusions of a statistical analysis. In contrast, Efron’s (1987, sec. 9) criterion for determining the value of \( B \) is the unconditional coefficient of variation,

\[
\text{CV}(\hat{\sigma}_B) \approx \sqrt{\text{CV}(\hat{\sigma})^2 + \frac{1}{2B}}.
\]

In practice the sampling variability of \( \hat{\theta} \), measured by \( \text{CV}(\hat{\theta}) \), will generally swamp the resampling component in (10). Using this argument, it follows that “For values of \( \text{CV}(\hat{\theta}) \geq .1 \), typical in practice, there is little improvement past \( B = 100 \). In fact, \( B \) as small as 25 gives reasonable results” (Efron, 1987; see also Efron and Tibshirani 1993, sec. 6.4, for essentially the same argument). The unconditional argument appears to be founded on an assumption that Monte Carlo error can be ignored if it is small relative to sampling variability. Note, however, that \( B = 100 \) translates into a 28\% relative error bound (2\( \sqrt{2}/100 = .28 \)) and with \( B = 25 \) the bound is 56\%! Thus, quite different answers are possible for these values of \( B \) depending upon the seed of the random number generator.

5. A MULTIVARIATE EXTENSION

The formula for the number resamples required for approximating a single bootstrap variance given in (7) can be extended to the \( m \)-dimensional setting using some standard results from multivariate normal theory. Let \( \hat{\theta} \) be an estimate of an \( m \)-dimensional parameter, \( \hat{\theta} = \Theta(F) \) of \( \Theta \). The bootstrap estimate of \( \Sigma = \text{cov}(\hat{\theta}) \) is

\[
\hat{\Sigma} = \text{cov}*(\hat{\theta}^*) = E^* \left\{ (\hat{\theta}^* - E^*(\hat{\theta}^*)) (\hat{\theta}^* - E^*(\hat{\theta}^*))^t \right\}.
\]

The Monte Carlo approximation of \( \hat{\Sigma} \), generalizing the univariate approximation in (1), is

\[
\Sigma_B = \frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\theta}_b^* - E^*(\hat{\theta}^*)) (\hat{\theta}_b^* - E^*(\hat{\theta}^*))^t.
\]

Notice that the resampling variance of \( a^t \hat{\theta}^* \) is equal to \( a^t \Sigma a \), for any \( m \)-vector \( a \) and that a Monte Carlo approximation of this variance is given by \( a^t \Sigma_B a \). Following the development in the univariate case we attempt to choose \( B \) such that the ratio \( a^t \Sigma_B a / a^t \Sigma a \) is close to one for all non-zero \( a \) with high probability. More formally, we require \( B \) such that

\[
1 - \alpha = P \left( 1 - \delta < \inf_{a \neq 0} \frac{a^t \Sigma_B a}{a^t \Sigma a} \leq \sup_{a \neq 0} \frac{a^t \Sigma_B a}{a^t \Sigma a} < 1 + \delta \right) = P(1 - \delta < l_{(1)} < l_{(m)} < 1 + \delta),
\]

where \( l_{(j)} \) denotes the \( j \)th smallest eigenvalue of \( \Sigma^{-1} \Sigma_B \).

Now, assuming that the resampling distribution of \( \hat{\theta}^* \) is \( m \)-variate normal with variance-covariance matrix \( \Sigma \) im-
plies that \((B - 1)\Sigma^{-1}\tilde{\Sigma}_B\) has an \(m\)-dimensional Wishart distribution with an identity scale matrix. It follows that
\[
P(1 - \delta < l(1) \leq l(m) < 1 + \delta) \leq P\{B(1 - \delta) < X(1) \leq X(m) < B(1 + \delta)\},
\]
where \(X_1, \ldots, X_m\) is a random sample of \(\chi^2_{B-1}\) variates (Muirhead 1982, theorem 9.7.5). Applying the normal approximation to the chi-squared distribution on the right side of (14) we obtain the approximate upper bound
\[
P(1 - \delta < l(1) \leq l(m) < 1 + \delta) \leq \left(1 - 2\Phi\left(-\sqrt{\frac{B}{2}\delta}\right)\right)^m.
\]
(15)

Combining equation (13) with (14) and the Taylor series approximation \((1 - (1 - \alpha)^{1/m})/2 \approx \alpha/(2m)\) reveals the generalized resample size formula
\[
B \geq \frac{2\Phi^{-1}(\frac{\alpha}{2m})^2}{\delta^2}.
\]
(16)

Note that (16) provides only a lower bound for \(B\) in the multivariate setting. Thus, for example, the formula implies that at least \(2(2.24/.1)^2 \approx 1,000\) resamples are required to achieve simultaneous accuracy of less than 10% relative error with probability .95 when \(m = 2\). Similarly, (14) only provides an upper bound on the confidence level. For example, with \(m = 2\) an upper bound on the probability of simultaneously achieving less than 10% relative error with \(B = 200\) resamples is \((1 - 2\Phi(-1))^2 = .47\). In fact, these formulas may significantly underestimate \(B\) and overestimate the confidence level. To illustrate, consider the open and closed book data discussed in Section 3. Suppose that bootstrap variance estimates are required for the percentage of variability explained by each of the first two principle components. In the 200 bootstrap samples, each of size \(B = 200\), used in the previous section, the inequality \(.9 < l(1) \leq l(2) < 1.1\) was satisfied only 53 times, significantly fewer than the 94 (200 \times .47) predicted by (14).

6. BOOTSTRAP CONFIDENCE INTERVALS

We conclude by noting that our conditional analysis is closely related to the usual argument for the endpoints of bootstrap confidence intervals. To see this, consider the naive confidence interval, \(\hat{\theta} \pm z_\alpha \hat{\sigma}\), and its Monte Carlo approximation, \(\hat{\theta} \pm \hat{z}_\alpha \hat{\sigma}_B\), based on \(B\) resamples. Efron (1987, sec. 9) argues that the number of resamples should be based on the conditional coefficient of variation of the halfwidth. Since \(CV^*(z_\alpha \hat{\sigma}_B) = CV^*(\hat{\sigma}_B)\), the conditional resample size formulas for bootstrap standard error approximation and naive confidence intervals agree exactly.

More sophisticated bootstrap confidence intervals involve the approximation of quantiles. For example, let \((\bar{\theta}_{B[\alpha]}, \bar{\theta}_{B[1 - \alpha]})\) denote a Monte Carlo approximation of a \((1 - 2\alpha)\)-level percentile method, bootstrap confidence interval for \(\theta\). In this case, the coefficient of variation of the halfwidth is approximately given by
\[
CV^*(\bar{\theta}_{B[\alpha]} - \hat{\theta}) \approx \frac{1}{B^{1/2}z_\alpha} \left\{\frac{\alpha(1 - \alpha)}{\phi(z_\alpha)^2}\right\}^{1/2}
\]
(Efron 1987, sec. 9). When \(B = 800\) and \(\alpha = .025\) equation (17) reveals a coefficient of variation of .048. Thus, roughly four times this number are required to achieve the target coefficient of variation of .025.

The analysis in the previous paragraph indicates that accurate approximation of the endpoints of bootstrap confidence intervals generally requires more resamples than the approximation of bootstrap variances or standard errors. Qualitatively, this finding is similar to Efron’s (1987). However, our findings are based on a level playing field using a criterion based on conditional (i.e., resampling) variability in both instances. In contrast, Efron’s conclusions are based on the use of an unconditional resample size criterion for standard errors and a conditional one for confidence intervals.

[Received October 1996. Revised January 1998.]

REFERENCES

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