

# Monte Carlo approximation of bootstrap variances

James G. Booth and Somnath Sarkar\*

October, 1996

## Abstract

It is widely believed that the number of resamples required for bootstrap variance estimation is relatively small. An argument due to Efron (1987), based on the unconditional coefficient of variation of the Monte Carlo approximation, suggest that as few as 25 resamples will give reasonable results. In this paper we argue that the number of resamples should in fact be determined by the conditional coefficient of variation, involving only resampling variability. Our analysis indicates that approximately 800 resamples are required in order that conclusions of statistical analyses utilizing bootstrap variance estimates remain unaffected by Monte Carlo error. Our approach can be generalized to the multivariate setting and a simple formula is given for determining a lower bound on the number of resamples required to approximate an  $m$ -dimensional bootstrap variance-covariance matrix.

## 1 Introduction

A conscientious scientist returns from lunch excited about his new data that he had analyzed that morning, using a state-of-the-art *bootstrap* method,

---

\*James G. Booth is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.. Somnath Sarkar is Senior Statistician, Lilly Research Laboratories, Eli-Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285. This work was completed while the first author was Visiting Professor, Department of Statistics, Colorado State University.

revealing a scientifically important result. He decides to double check his analysis, but finds to his disappointment that the result is no longer significant. A further run of the analysis produces yet another P-value. Perplexed, the scientist decides to call a statistician who informs him that the fluctuation in P-values is due to Monte Carlo error. At this point the scientist decides to use another, more reliable, statistical procedure.

The reason the scientist did not get consistent answers when he repeatedly analyzed the same data is that the number of resamples required for approximating a bootstrap variance was underestimated in his statistical package. In this paper we show that this number 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  $Y_1, \dots, Y_n$  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 non-parametric 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  $\sigma^2(F) = E\{(\hat{\theta} - E(\hat{\theta}))^2\}$ . The bootstrap estimate of  $\sigma^2$  is therefore

$$\hat{\sigma}^2 := \sigma^2(\hat{F}) = E^* \{(\hat{\theta}^* - E^*(\hat{\theta}^*))^2\}, \quad (1)$$

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

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

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

Efron (1987) addresses the issue of the number of resamples for approximating bootstrap variances. Efron's argument 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 paper 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 suggested by Efron.

## 2 A conditional criterion for choosing $B$

We contend that resampling methods will only be trusted by practitioners 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 only happen if the relative error of  $\hat{\sigma}_B^2$  *due to resampling* is small; i.e.

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

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

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

for some small probability  $\alpha$ .

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

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

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 given in

$\delta$	Probable range of P-values
.1	(.040,.063)
.2	(.032,.080)
.5	(.016,.166)

Table 1: Probable ranges of approximate P-values due to Monte Carlo error when the true P-value is .05

Table 1 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 contrast, if the relative error of  $\hat{\sigma}_B^2$  can reach 50% then *the conclusions of the analysis are 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 determine a rough formula for  $B$  which may be applied in a wide range of problems. Thus a high degree of accuracy in our asymptotic approximations is not necessary.) Under mild regularity conditions, the resampling distribution of  $\hat{\theta}^*$  will also be approximately normal. Since  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$  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 chisquared 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 chisquared, we obtain the approximation

$$\begin{aligned}
P^* \left( 1 - \delta < \frac{\hat{\sigma}_B^2}{\hat{\sigma}^2} < 1 + \delta \right) &\approx P \left( B(1 - \delta) < \chi_B^2 < B(1 + \delta) \right) \\
&\approx P \left( B(1 - \delta) < B + \sqrt{2B}Z < B(1 + \delta) \right) \\
&= 1 - 2\Phi \left( -\sqrt{\frac{B}{2}}\delta \right). \tag{6}
\end{aligned}$$

Combining equations (4) and (6) reveals the formula

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

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

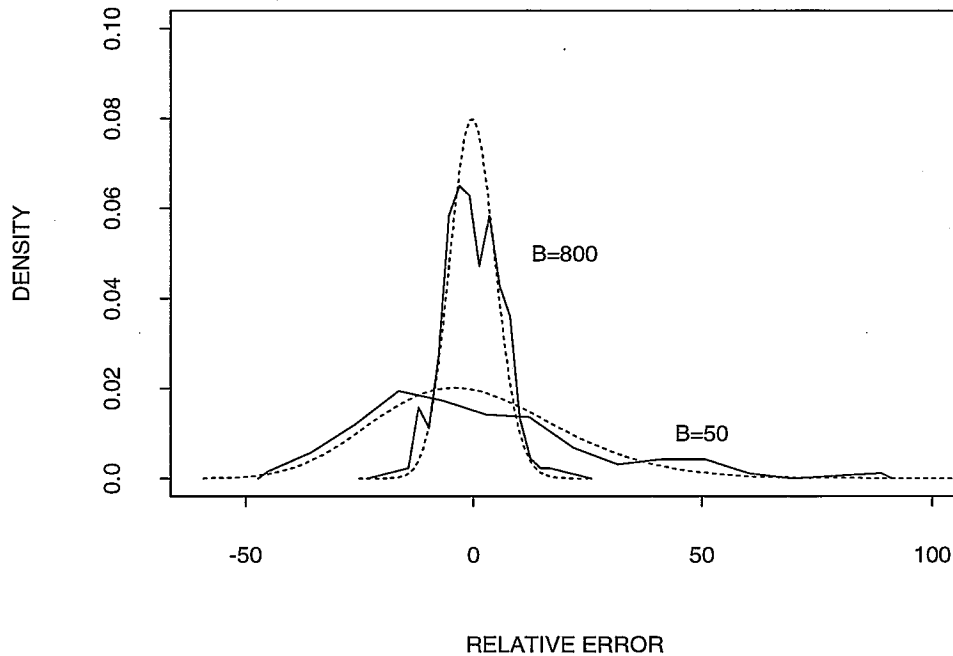


Figure 1: Histograms of the percentage relative errors of 200 Monte Carlo approximations along with the corresponding chisquared density approximations.

In order to illustrate the accuracy of the chisquared approximation used in (6), consider the Law School data from Efron (1982). This dataset consists of  $n = 15$  bivariate values with a sample correlation coefficient  $\hat{\theta} = .776$ . Histograms of 200 values of the percent relative error  $P_B = 100(\hat{\sigma}_B^2/\hat{\sigma}^2 - 1)$  are displayed in Figure 1 for  $B = 50$  and  $B = 800$ . Note that the percent relative error is related to  $X_B = (B - 1)\hat{\sigma}_B^2/\hat{\sigma}^2$  through the location-scale transformation,  $P_B = 100(X_B - B + 1)/(B - 1)$ . The transformed chisquared densities corresponding to two histograms are also displayed on the plot. In both cases the chisquared approximation appears to quite good. For example, in our simulation, 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 chisquared approximation. The value of  $\hat{\sigma}^2$  was taken to be the average of the 200 values in each simulation. In both cases  $\hat{\sigma}^2 = .0179$  ( $\hat{\sigma} = .134$ ).

## 4 Comparison with Efron's approach

The chisquared approximation,  $B\hat{\sigma}_B^2/\hat{\sigma}^2 \sim \chi_B^2$  implies that

$$CV^*(\hat{\sigma}_B^2) \approx \sqrt{\frac{2}{B}}. \quad (8)$$

Thus the 10% relative error criterion discussed in Sections 2 and 3 is approximately equivalent to  $CV^*(\hat{\sigma}_B^2) \leq \sqrt{2/800} = .05$ . In contrast, Efron's (1987) criterion for determining the value of  $B$  is the unconditional coefficient of variation,

$$CV(\hat{\sigma}_B^2) \approx \sqrt{CV(\hat{\sigma}^2)^2 + \frac{2}{B}}. \quad (9)$$

In practice the sampling variability of  $\hat{\sigma}^2$ , measured by  $CV(\hat{\sigma}^2)$ , will generally swamp the resampling error in (8). Efron concludes that "For values of  $CV(\hat{\sigma}_B^2) \geq .1$ , typical in practice, *there is little improvement past  $B = 100$* . In fact,  $B$  as small as 25 gives reasonable results." Essentially the same argument and conclusion is reproduced in Efron and Tibshirani (1993, Section 6.4). The unconditional argument is founded on an assumption that Monte Carlo error can be ignored if it is small relative to sampling variability. In contrast, our conditional argument is based on a belief that Monte Carlo error should not be allowed to affect the conclusions of a statistical analysis.

## 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,  $\theta = \theta(F)$  of  $F$ . The bootstrap estimate of  $\Sigma = \text{Cov}(\hat{\theta})$  is

$$\hat{\Sigma} = \text{Cov}^*(\hat{\theta}^*) = E^* \{ (\hat{\theta}^* - E^*(\hat{\theta}^*)) (\hat{\theta}^* - E^*(\hat{\theta}^*))^t \}. \quad (10)$$

The Monte Carlo approximation of  $\hat{\Sigma}$ , generalizing the univariate approximation in (2), is

$$\hat{\Sigma}_B = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - E^*(\hat{\theta}^*)) (\hat{\theta}_b^* - E^*(\hat{\theta}^*))^t. \quad (11)$$

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

$$\begin{aligned} 1 - \alpha &= P \left( 1 - \delta < \inf_{\mathbf{a} \neq 0} \frac{\mathbf{a}^t \hat{\Sigma}_B \mathbf{a}}{\mathbf{a}^t \hat{\Sigma} \mathbf{a}} \leq \sup_{\mathbf{a} \neq 0} \frac{\mathbf{a}^t \hat{\Sigma}_B \mathbf{a}}{\mathbf{a}^t \hat{\Sigma} \mathbf{a}} < 1 + \delta \right) \\ &= P(1 - \delta < l_{(1)} \leq l_{(m)} < 1 + \delta), \end{aligned} \quad (12)$$

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

Now, assuming that the resampling distribution of  $\hat{\theta}^*$  is  $m$ -variate normal with variance-covariance matrix  $\hat{\Sigma}$  implies that  $(B-1)\hat{\Sigma}^{-1}\hat{\Sigma}_B$  has an  $m$ -dimensional Wishart distribution with an identity scale matrix. It follows that

$$P(1 - \delta < l_{(1)} \leq l_{(m)} < 1 + \delta) \leq P \left\{ B(1 - \delta) < X_{(1)} \leq X_{(m)} < B(1 + \delta) \right\}, \quad (13)$$

where  $X_1, \dots, X_m$  is a random sample of  $\chi_{B-1}^2$  variates (Muirhead, 1982, Theorem 9.7.5). Applying the normal approximation to the chisquared distribution on the right side of (13) we obtain the approximate upper bound

$$P(1 - \delta < l_{(1)} \leq l_{(m)} < 1 + \delta) \leq \left\{ 1 - 2\Phi \left( -\sqrt{\frac{B}{2}} \delta \right) \right\}^m. \quad (14)$$

Combining equation (12) with (14) and the Taylor series approximation  $(1 - (1 - \alpha)^{1/m})/2 \approx \alpha/(2m)$  reveals the generalized formula

$$B \geq \frac{2 \left| \Phi^{-1} \left( \frac{\alpha}{2m} \right) \right|^2}{\delta^2}. \quad (15)$$

Note that (15) only provides a lower bound for  $B$  in the multivariate setting. Thus, for example, the formula implies that *at least*  $2(2.81/.1)^2 = 1580$  (or approximately 1600) resamples are required in order to achieve simultaneous accuracy of less than 10% relative error with probability .95 when  $m = 10$ .

## References

- Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.
- Efron, B. (1987), "Better bootstrap confidence intervals" (with discussion), *J. Amer. Statist. Assoc.* **82**, 171-185.
- Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*: Chapman and Hall.
- Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory*: Wiley.