BOOTSTRAP CONFIDENCE INTERVALS: GOOD OR BAD?

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 116
MARCH 1987

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 RO1 GM21215-12

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
BOOTSTRAP CONFIDENCE INTERVALS: GOOD OR BAD?

BY

BRADLEY EFron

TECHNICAL REPORT NO. 116
MARCH 1987

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215-12

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
BOOTSTRAP CONFIDENCE INTERVALS: GOOD OR BAD?

B. Efron

Abstract

The bootstrap is a nonparametric technique for estimating standard errors and approximate confidence intervals. Rasmussen has used a simulation experiment to suggest that bootstrap confidence intervals perform very poorly in the estimation of a correlation coefficient. Part of Rasmussen's simulation is repeated. A careful look at the results shows the bootstrap intervals performing quite well. Some remarks are made concerning the virtues and defects of bootstrap intervals in general.
Bootstrap Confidence Intervals: Good or Bad?

B. Efron

The bootstrap is a nonparametric technique for assessing standard errors and setting approximate confidence intervals, see Efron and Tibshirani (1986), Efron and Diaconis (1983), or Efron (1982). Rasmussen (1987) suggests that bootstrap confidence intervals perform disasterously in the case of estimating a correlation coefficient. In fact the bootstrap intervals perform rather well in this situation, as will be shown here by repeating a small portion of Rasmussen's simulation experiment.

Suppose then that \( x_1, x_2, \ldots, x_{30} \) is a random sample from a bivariate normal population \( N_2(0,1) \), with mean vector 0 and covariance matrix the identity. For this distribution the true Pearson correlation coefficient \( \rho \) equals 0. On the basis of \( x_1, x_2, \ldots, x_{30} \), we desire to set a confidence interval for \( \phi = .5 \cdot \log[(1+\rho)/(1-\rho)] \), the "z transform" of \( \rho \). The corresponding statistic \( \hat{\phi} = .5 \cdot \log[(1+\hat{\rho})/(1-\hat{\rho})] \), where \( \hat{\rho} \) is the sample correlation coefficient of \( x_1, x_2, \ldots, x_{30} \), is approximately normally distributed around \( \phi \), with constant standard deviation \( 1/\sqrt{27} = .192 \),

\[
\hat{\phi} \sim N(\phi, .192^2).
\]

This approximation is quite good for \( \phi \) near 0. (It simplifies discussion to work in terms of \( \phi \) and \( \hat{\phi} \), but all of our conclusions hold equally well in terms of \( \rho \) and \( \hat{\rho} \), see Remark 3.)

One hundred independent random samples \( x_1, x_2, \ldots, x_{30} \) were drawn from a \( N_2(0,1) \) distribution, and for each sample two confidence intervals were constructed: the normal theory parametric 90% confidence interval \( \hat{\phi} \pm .317 = (1.645 \cdot .192) \);
and the "percentile method" 90% bootstrap confidence interval based on 2000 bootstrap replications per sample, i.e. the interval with endpoints at the 100th and 1900th order statistics of the 2000 bootstraps. Figure 1 displays both intervals for the first 10 of the 100 samples.

The most important thing to notice is how similar the two sets of intervals are. For most practical purposes, except perhaps for strict hypothesis testing, the two sets of intervals would be effectively equivalent. This must be counted as a success for the bootstrap method which in a totally automatic and nonparametric way is closely approximating a quite specific parametric result: namely that in sampling 30 times from a bivariate normal distribution, the z-transform $\hat{\phi}$ is distributed approximately $N(\phi, 0.192^2)$.

Figure 1. Confidence intervals for $\phi$, confidence level 90%, for the first ten samples in the simulation experiment. Solid lines are the nonparametric percentile method (bootstrap) intervals; dashed lines are the normal theory parametric intervals. The central hashmark indicates $\hat{\phi}$. 

- .7 -.6 -.5 -.4 -.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7
Table 1 gives a more precise description of how the bootstrap intervals performed in our sampling experiment. Column F shows that the ratio

$$R_{len} = \frac{\text{length of bootstrap interval}}{\text{length of parametric interval}}$$

(parametric length = .633)

had mean .986 and standard deviation .171 for the 100 samples. In other words the bootstrap intervals averaged almost the right length, but with a coefficient of variation (standard deviation/mean) of 17%.

Column G gives the ratio of the nonparametric to the parametric estimate of standard deviation for $\hat{\phi}$,

$$R_{sd} = \frac{\text{bootstrap estimate of standard dev.}}{\text{parametric estimate of standard dev.}}$$

(parametric sd = .192)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>-0.002</td>
<td>-0.316</td>
<td>0.309</td>
<td>-0.319</td>
<td>0.315</td>
<td>0.986</td>
<td>0.988</td>
<td>0.882</td>
<td>0.190</td>
</tr>
<tr>
<td>SD</td>
<td>0.201</td>
<td>0.214</td>
<td>0.214</td>
<td>0.201</td>
<td>0.201</td>
<td>0.171</td>
<td>0.170</td>
<td>0.057</td>
<td>0.033</td>
</tr>
<tr>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.392</td>
<td>-0.698</td>
<td>-0.142</td>
<td>-0.709</td>
<td>-0.075</td>
<td>0.878</td>
<td>0.883</td>
<td>0.847</td>
<td>0.170</td>
</tr>
<tr>
<td>2</td>
<td>0.186</td>
<td>-0.102</td>
<td>0.477</td>
<td>-0.131</td>
<td>0.503</td>
<td>0.914</td>
<td>0.935</td>
<td>0.867</td>
<td>0.180</td>
</tr>
<tr>
<td>3</td>
<td>-0.189</td>
<td>-0.457</td>
<td>0.054</td>
<td>-0.506</td>
<td>0.128</td>
<td>0.807</td>
<td>0.811</td>
<td>0.815</td>
<td>0.156</td>
</tr>
<tr>
<td>4</td>
<td>0.167</td>
<td>-0.184</td>
<td>0.487</td>
<td>-0.150</td>
<td>0.484</td>
<td>1.060</td>
<td>1.070</td>
<td>0.918</td>
<td>0.206</td>
</tr>
<tr>
<td>5</td>
<td>-0.193</td>
<td>-0.484</td>
<td>0.121</td>
<td>-0.510</td>
<td>0.124</td>
<td>0.956</td>
<td>0.941</td>
<td>0.883</td>
<td>0.181</td>
</tr>
<tr>
<td>6</td>
<td>-0.058</td>
<td>-0.402</td>
<td>0.226</td>
<td>-0.375</td>
<td>0.259</td>
<td>0.992</td>
<td>0.998</td>
<td>0.893</td>
<td>0.192</td>
</tr>
<tr>
<td>7</td>
<td>0.033</td>
<td>-0.318</td>
<td>0.381</td>
<td>-0.284</td>
<td>0.350</td>
<td>1.104</td>
<td>1.102</td>
<td>0.931</td>
<td>0.212</td>
</tr>
<tr>
<td>8</td>
<td>0.252</td>
<td>-0.038</td>
<td>0.552</td>
<td>-0.065</td>
<td>0.569</td>
<td>0.932</td>
<td>0.920</td>
<td>0.875</td>
<td>0.177</td>
</tr>
<tr>
<td>9</td>
<td>0.033</td>
<td>-0.265</td>
<td>0.317</td>
<td>-0.284</td>
<td>0.350</td>
<td>0.919</td>
<td>0.894</td>
<td>0.869</td>
<td>0.172</td>
</tr>
<tr>
<td>10</td>
<td>0.062</td>
<td>-0.269</td>
<td>0.399</td>
<td>-0.255</td>
<td>0.379</td>
<td>1.055</td>
<td>1.070</td>
<td>0.917</td>
<td>0.206</td>
</tr>
</tbody>
</table>

Table 1. Results for the first ten samples in the simulation experiment; also means and standard deviations for all one hundred samples. Key: A = $\hat{\phi}$; B and C = endpoints of bootstrap interval; D and E = endpoints of normal theory interval $\hat{\phi} \pm .317$; F = (C-B)/(E-F), the ratio of lengths of the two intervals; G = ratio of bootstrap standard deviation estimate to normal theory estimate $1/\sqrt{27}$; H = actual coverage probability of bootstrap interval; I = bootstrap standard deviation estimate for $\hat{\phi}$.
Comparing entries in F and G, we see that almost all of the variability in the bootstrap interval lengths is explained by the corresponding variability in the standard deviation estimates. In fact the correlation between $R_{\text{len}}$ and $R_{\text{sd}}$ is .997. The 17% coefficient of variation for the bootstrap interval lengths comes directly from the 17% coefficient of variation for the bootstrap estimate of standard deviation. There is good evidence that no fully nonparametric estimate of standard deviation can do much better in this situation, see Efron (1981). By implication no nonparametric confidence intervals can better match the parametric intervals $\hat{\phi} \pm .317$. (But see Remark 8.)

Table 2 shows the Achieved Significance Level (ASL) of the true value $\phi = 0$ for all 100 samples, as computed almost exactly by the parametric method [where $\text{ASL} = 1 - \Phi(\sqrt{27} \hat{\phi})$, $\Phi(\cdot)$ being the standard normal cumulative] or approximately by the bootstrap method [where ASL equals the proportion of the 2000 bootstrap replications lying below 0.] If the bootstrap method were exact, then an expected 15 of the 100 samples would have bootstrap ASL between .10 and .25, etc. Table 2 shows that in our sampling experiment the bootstrap method yielded quite reasonable inferences concerning the true value $\phi = 0$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>or</td>
<td>or</td>
<td>or</td>
<td>or</td>
<td>or</td>
<td></td>
</tr>
<tr>
<td>[1,.975]</td>
<td>[.975,.950]</td>
<td>[.95,.90]</td>
<td>[.90,.75]</td>
<td>[.75,.50]</td>
<td></td>
</tr>
<tr>
<td>Expected:</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Bootstrap:</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>21</td>
<td>51</td>
</tr>
<tr>
<td>Parametric:</td>
<td>6</td>
<td>5</td>
<td>13</td>
<td>24</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 2. Achieved Significance Level of the true value $\phi = 0$ according to the bootstrap and parametric confidence methods, all 100 samples. For example, in 14 of the 100 samples, 0 lay between either the 5th and 10th percentiles or the 90th and 95th percentiles of the 2000 bootstrap values. This compares with an expected value of 10 such samples if the bootstrap confidence method were exact.
The first two columns of Table 2 show that 14 of the 100 bootstrap intervals had either \( \text{ASL} < 0.05 \) or \( \text{ASL} > 0.95 \), or equivalently had two-sided 90% bootstrap confidence interval not containing the true value \( \phi = 0 \). Likewise, 0 did not lie in 11 of the 100 parametric 90% intervals. In the latter case we know that the expected number of such intervals is almost exactly 10 for the parametric model. Therefore \( \frac{14}{11} = 12.7\% \) is a reasonable estimate for the noncoverage probability of the bootstrap intervals. See Remark 8.

In the 14 samples where the bootstrap interval didn't cover 0, the parametric \( \text{ASL} \sim \Phi(\sqrt{27\hat{\phi}}) \) equalled \( 0.01, 0.03, 0.08, 0.017, 0.040, 0.050, 0.063, 0.063, 0.934, 0.956, 0.959, 0.962, 0.976, 0.979 \). In other words, if the bootstrap interval "rejected 0" then the parametric interval did likewise, or at least came close to doing so. We see again that the bootstrap and parametric methods provide nearly equivalent inferences.

Some Remarks

(1) Figure 2 shows a random sample of size 30 from a nonnormal bivariate distribution (which won't be revealed here, in order to put the reader in the usual situation of the data analyst). The sample Pearson correlation coefficient is

![Figure 2](image_url)

Figure 2. A random sample \( x_1, x_2, \ldots, x_{30} \) from a non-normal bivariate distribution. The sample correlation coefficient is \( \hat{\rho} = 0.820 \). What is a reasonable confidence interval for \( \rho \)?
\( \hat{\rho} = .820. \) Two thousand bootstrap replications gave estimated standard deviation \( \hat{\sigma}_{\text{BOOT}} = .0436, \) and approximate 90\% bootstrap confidence interval \( [.723,.875]. \) The usual normal theory standard deviation estimate is \( \hat{\sigma}_{\text{NORM}} = .0631, \) and the usual 90\% confidence interval is \( [.682,.899]. \) In this case the normal theory standard deviation and confidence interval are about 45\% larger than their bootstrap counterparts, but of course there is no particular reason to believe the normal theory. Situations like this bring home the virtues of a nonparametric approach. A considerable amount of recent theoretical work supports the bootstrap method, see Efron (1987), Efron and Tibshirani (1986), Hail (1987), although nonparametric confidence intervals are still not completely understood.

(2) Both the bootstrap and normal theory intervals in Remark 1 include a simple bias correction, as discussed in Efron (1982) and Efron and Tibshirani (1986). This is important in most applications, though not in Rasmussen's situation where the bias of \( \hat{\phi} \) for estimating \( \phi \) is near 0.

(3) When sampling from a bivariate normal distribution we know that the z-transform normalizes the very non-normal distribution of \( \hat{\rho}. \) The z-transform is strongly nonlinear, which is why the normal-theory confidence intervals for \( \rho \) can be so asymmetric about the point estimate \( \hat{\rho}. \) For instance the interval \( [.682,.899] \) in Remark 1 is highly skewed to the left of \( \hat{\rho} = .820, (.820-.682)/(.899-.820) = 1.75. \)

The bootstrap percentile methods, including various corrections, are transformation invariant. This means that the methods can be applied on the original scale, the \( \hat{\rho} \) scale in Remark 1, with the assurance that if any normalizing transformation exists then that transformation will in effect be automatically incorporated into the computation of the bootstrap confidence intervals. See Section 2 of Efron (1987). In the example of Remark 1, the bootstrap interval \( [.723,.875] \) is skewed to the left of \( \hat{\rho} \) in almost the same ratio as the normal-theory interval, \( (.820-.723)/(.875-.820) = 1.76. \) This indicates that something like the z-transform
might be appropriate for our non-normal population. Fortunately we do not have to settle this question in order to calculate bootstrap intervals.

(4) As few as 10 or 20 bootstrap replications can give quite useful bootstrap estimates of standard deviation, and more than 200 replications are hardly ever needed for this purpose. Nonparametric confidence intervals demand far more computation. It is not excessive to use 2000 replications, as in this paper, though we might have stopped at 1000. Section 9 of Efron (1987) presents the theory behind these statements. [Current research aimed at decreasing these numbers appears in Davidson, Hinkley and Schechtman (1986) and Ogbonmwan and Wynn (1987).] Rasmussen's choice of 500 replications per sample is dangerously low, especially for the extreme confidence limits (99%) he investigates.

(5) In asking if bootstrap confidence intervals are good or bad, an important question is "compared to what?". The standard method, which is currently the method of choice in most situations where bootstrap confidence intervals might be seriously considered, assigns

$$\hat{\theta} \pm 1.645 \hat{\sigma}$$

as an approximate 90% confidence interval for $\theta$, where $\hat{\theta}$ is a point estimate of $\theta$, and $\hat{\sigma}$ is an approximate estimated standard deviation for $\hat{\theta}$, often calculated by Taylor series methods.

The standard method is extremely useful in statistical practice since it can be applied easily to a wide variety of problems. It is far from perfect however, and usually will not perform well in the kind of tests Rasmussen proposes for the bootstrap intervals. In our main simulation experiment, described in Table 1, the standard intervals for $\hat{\phi}$, using the usual delta method estimate of standard deviation, (p. 43, Efron (1982)), were shorter than the true parametric interval
\( \hat{\phi} \pm .317 \) in 76 out of the 100 samples (compared to 58 out of 100 for the bootstrap intervals). The standard interval lengths averaged .914 of the true interval length, compared to .986 for the bootstrap intervals. All 100 standard intervals were shorter than their bootstrap counterparts.

A good way to think of bootstrap intervals is as a cautious improvement over the standard intervals, using large amounts of computation to overcome certain deficiencies of the standard method, for example its lack of transformation invariance. The bootstrap is not intended to be a substitute for precise parametric results, but rather a way to reasonably proceed when such results are unavailable.

(6) In fact it is easy to find situations where the nonparametric bootstrap percentile-type intervals do not perform very well. Table 10.3 of Efron (1982) describes one such situation, and Schenker (1985) describes another. (Schenker's example is considered further in Efron (1987).) In both of these situations the percentile intervals are an improvement, but too of a cautious improvement, over the standard intervals. Less cautious methods, in particular the "bootstrap t" can do better in these cases, at the expense of erratic performance in other situations. Hall (1986), and Abramovitch and Singh (1985) are recent references on the bootstrap t theory.

(7) The simulation experiment of Table 1 was run at confidence level .90, rather than .95 and .99 as in Rasmussen's experiment. In the author's experience nonparametric bootstrap intervals perform better when not pushed too far toward extreme coverage probabilities.

The bootstrap method can be applied parametrically rather than nonparametrically, see Figure 8 of Efron and Tibshirani (1986). This is a useful approach when one has a believable parametric family but a complicated statistic that is difficult to analyze theoretically (for example a trimmed correlation coefficient rather than Pearson's \( \hat{\phi} \)). The parametric bootstrap, if the parametric assumptions are
justified, gives good confidence intervals even for high coverage probabilities, see Efron (1987).

(8) The smoothed bootstrap is a compromise between the fully nonparametric bootstrap used in Table 1, and the fully parametric bootstrap just mentioned. Instead of drawing bootstrap data points from the empirical distribution of the data as the nonparametric bootstrap does, points are drawn from a smoothed empirical distribution. The 100 samples in Table 1 were reanalyzed using the uniform smoothed bootstrap as in lines 4 and 5 of Table 1, Efron (1981), 2000 bootstrap replications per sample. The smoothing window for each sample had covariance matrix $1/2$ that of the data. Now $R_{\text{len}}$, the ratio of smoothed bootstrap to parametric interval length, had mean .988 and standard deviation .109, so its coefficient of variation was reduced from 17% to 11%.

(9) In situation (1) the confidence interval $\phi \in [\hat{\phi}-a,\hat{\phi}+b]$ has true coverage probability

$$\Phi(\frac{a}{.192}) - \Phi(-\frac{b}{.192}).$$

Column H of Table 1 gives the true coverage probability for each of the 100 bootstrap intervals. These averaged .882 rather than the intended value .900. If coverage probability is independent of $\hat{\phi}$, which appeared to be nearly true in our sampling experiment, then .882 is a superior estimate of the true coverage probability of the bootstrap method. This calculation estimates the non-coverage probability as 11.8%, compared to the value 12.7% derived by simply counting non-coverages of $\phi = 0$.

The smoothed bootstrap intervals of Remark 8 had average coverage .890. Moreover the standard deviation of the coverages was .038, compared to .057 for the ordinary bootstrap intervals. Smoothing markedly improves the bootstrap intervals in this case.
Summary

There is now a considerable body of theory supporting bootstrap confidence intervals. This support mainly concerns asymptotic correctness at a high order of precision. Simulation studies of small-sample performance are an important adjunct to this theory.

However it is easy to disagree about the outcome of a simulation study. The main points made here differ from Rasmussen's results more in emphasis than on fact:

- At 90% coverage level, the bootstrap intervals closely match the parametric intervals on a sample by sample basis.

- "Closely match" means more than just coverage probability; it also takes into account how similar were the inferences from the two methods.

- The bootstrap intervals worked close to the optimum possible for a fully nonparametric method. However semiparametric methods such as the smoothed bootstrap can do considerably better. The parametric bootstrap does better still, assuming the parametric assumptions are justified.

- Bootstrap methods are intended to supplement rather than replace parametric analysis, particularly when parametric methods can't be used because of modeling uncertainties or theoretical intractability.

- The various bootstrap percentile methods can be viewed as improvements on the standard intervals $\hat{\theta} \pm z_{0.05} \hat{\sigma}$ ($z_{0.05} = 1.645$). In particular the bootstrap intervals automatically incorporate transformations to normality, such as the z-transform for correlation coefficients.

- The standard method is currently the most-used technique for setting approximate confidence intervals. It should always be included as a point of comparison in simulation studies of the bootstrap, or of other nonparametric techniques.
References


Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. SIAM CBMS Monograph No. 38.


