Unbiasing the Bootstrap—Bootknife Sampling vs. Smoothing

Tim C. Hesterberg, Research Department, Insightful Corp.

This work was supported by NSF Phase I SBIR Award No. DMI-9861360.

Abstract

Bootstrap standard errors are generally biased downward, which is a primary reason that traditional bootstrap confidence intervals have coverage probability which is too low. For the sample mean the downward bias is a factor of $\frac{n-1}{n}$ (for the squared standard error); the same bias holds approximately for asymptotically-linear statistics. In the case of stratified or two-sample bootstrapping, the bias depends on the individual strata or sample sizes, not the combined total; hence the bias is substantial in situations with many small strata.

We discuss two remedies. One, specifically for bootstrapping the sample mean, is to do a smoothed bootstrap, sampling from a kernel density estimate of the underlying distribution, with kernel width chosen to correct the bias. The second, more generally-applicable, is "bootknife sampling," in which bootstrap samples are drawn from jackknife samples. These provide more accurate inferences than ordinary bootstrap sampling – better confidence interval coverage and less-biased or unbiased standard errors. These methods are implemented in downloadable S-PLUS software.

Key Words: bootstrap, jackknife, kernel smoothing, smoothing.

1 Introduction

We begin with a short introduction to the bootstrap; for a more complete introduction to the bootstrap see [4] or [1]. The original data are $\mathcal{X} = (X_1, X_2, \ldots, X_n)$, a sample from an unknown distribution (which may be multivariate). Let $\theta = \theta(F)$ be a real-valued functional parameter of the distribution, such as its mean or slope of a regression line, and $\hat{\theta} = \theta(\hat{F})$ the value estimated from the data. The sampling distribution of $\hat{\theta}$

$$G(a) = P_F(\hat{\theta} \leq a) \tag{1}$$

is used for statistical inference. In other examples the sampling distribution of a test statistic such as $T = (\hat{\theta} - \theta)/s$ is required where s is also computed from the data; in that case define G as the distribution of T.

In simple problems the sampling distribution can be approximated using methods such as the central limit theorem and the substitution of sample moments such as $\overline{x}$ and s into formulas obtained by probability theory. This may not be sufficiently accurate or even possible in many real, complex situations.

The bootstrap principle is to estimate some aspect of G, such as its standard deviation, by replacing F with an estimate $\hat{F}$. In the usual nonparametric bootstrap, $\hat{F}$ is the empirical distribution $\hat{F}_n$. Then sampling from $\hat{F}$ corresponds to generating samples of size n with replacement from the original data.

When it is known that the underlying distribution is continuous, it may be desirable to sample from a continuous distribution rather than from the empirical distribution [2]. The most common way to do this is by kernel smoothing, in which $\hat{F}$ is created by convolving $\hat{F}_n$ with a kernel distribution. This is implemented by sampling with replacement from $\hat{F}_n$, then independently adding one random observation from the kernel distribution to each of the n observations. A common choice for the kernel distribution is a normal distribution (or multivariate normal distribution), which should be fine in most applications; see e.g. [8] and the references therein. But there does not seem to be a standard choice for the kernel parameter, e.g. the standard deviation (or covariance matrix) for the (multivariate) normal distribution. In Section 2 we propose a standard value for the smoothing parameter, based on making variances approximately correct.

In Section 3 we propose a new bootstrap sampling method, "bootknife sampling," which also corrects the downward bias. The new method is also applicable to discrete data, and is compatible with bootstrap tilting.

2 Choosing the Bootstrap Smoothing Parameter to Unbias Standard Errors

Smoothing has two effects. The first, and the original motivation for smoothing, was to sample from continuous rather than discrete bootstrap distributions in applications where the underlying distribution is known to be discrete. However, [5] shows that bootstrap distributions are practically continuous under fairly general conditions, so this effect is relatively unimportant. More important is that smoothing produces estimates with larger variance. We propose to choose the smoothing parameter to produce unbiased variance estimates.

The usual sample standard deviation is

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2}$$

The choice of $n-1$ for a divisor instead of n makes the estimate unbiased for σ^2, the variance of the underlying distribution.

The usual bootstrap corresponds to using a divisor of n. In particular, if the statistic being bootstrapped is $\bar{X}$, the sample mean of univariate data, then the bootstrap estimate of $\mathrm{Var}(\bar{X})$ is

$$\mathrm{Var}(\bar{X}^*) = \hat{\sigma}^2/n, \tag{3}$$

where

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{4}$$

This differs from the usual unbiased estimate s^2/n by the factor $(n-1)/n$.

Note that (3) can be calculated exactly; the usual Monte Carlo implementation of the bootstrap is needed only in more complicated situations.

Now suppose that smoothed bootstrapping is used, that the ith observation in a bootstrap sample is $x_{i^*} + z_i$, where x_{i^*} is randomly chosen with equal probabilities from the original sample and z_i is an independent random variable with mean 0 and standard deviation h. The resulting variance of the sampling distribution

$$\mathrm{Var}(\bar{X}^*) = \hat{\sigma}^2/n + h^2/n. \tag{5}$$

We suggest using

$$h = s/\sqrt{n} \tag{6}$$

which makes (5) unbiased. We have made this the default choice for the smoothed bootstrap in S+RESAMPLE; see the summary, Section 4, for download information.

In multivariate problems, let S^2 denote the usual unbiased sample variance-covariance matrix,

$$S^2_{jk} = (n-1)^{-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k). \quad (7)$$

We suggest smoothing by letting z_i have mean zero and variance-covariance matrix S^2/n which makes the bootstrap estimate for the variance-covariance matrix of the sampling distribution of $\bar{X}$ unbiased.

In nonlinear problems these smoothing choices will in general not result in unbiased estimates, though they will typically have smaller bias than with simple bootstrap sampling.

This smoothing will also tend to improve the coverage accuracy of bootstrap confidence intervals. Bootstrap confidence intervals tend to be anti-conservative (see simulation results collected in [8]), in large part because bootstrap variances tend to be too small. Figure 1 demonstrates this—both smoothing and bootknife sampling (described below) give higher coverage probability, closer to the nominal value, though still short of the nominal value.

3 Bootknife Sampling

Kernel smoothing is not always possible, for example with discrete data. Even where it is possible it may be undesirable. For example, it can lead to "impossible data," data which violates known constraints (e.g. that the data must be positive). In high-dimensional problems it is difficult to smooth in a way that respects the (unknown) structure that is usually present. If the data are skewed, or arise from a regression problem with heteroskedastic errors, then simple kernel smoothing will lead to some biases.

Furthermore, kernel smoothing creates new data points which were not in the original sample. This is incompatible with some bootstrap procedures, such as the fast importance-sampling implementation for bootstrap tilting inferences [3, 6].

In this section we propose "bootknife sampling," which generates bootstrap samples solely from the original data. The name comes from a combination of jackknife and bootstrap, and provides a short description of the procedure. To generate a single bootstrap sample, we first create a jackknife sample by omitting one observation, then draw a bootstrap sample of size n with replacement from the remaining $n-1$ observations. (The original name "jackboot sampling" [7] was even more descriptive, in that it reflected the order of the operations, but the name was too Nazi.)

Using bootknife sampling, the bootstrap estimate of the variance of a sample mean is unbiased, in both univariate and multivariate problems. This result is obtained by conditioning on the omitted observation, say o^*:

$$\begin{aligned} \mathrm{Var}(\bar{X}^*) &= \mathrm{E}(\mathrm{Var}(\bar{X}^*|o^*)) + \mathrm{Var}(\mathrm{E}(\bar{X}^*|o^*)) \\ &= \mathrm{E}(s^2)/n = \mathrm{Var}(\bar{X}) \end{aligned} \quad (8)$$

We omit the details of algebraic simplification after the conditioning step.

The omissions can be random, but better results are obtained using stratification. If B bootstrap samples are to be generated let $k = \lfloor B/n \rfloor$, and omit each of the observations deterministically in k of the bootstrap samples. For the remaining bootstrap samples, generate a sample of size $B - nk$ without replacement from the numbers $1 \ldots n$, and omit the corresponding observations. Thus each original data point is omitted either k or $k+1$ times.

Figures 2 and 3 show coverage properties of three bootstrap intervals, using simple bootstrap and bootknife sampling. For the most part the intervals obtained using bootknife sampling (dotted lines) have higher coverage, particularly for smaller values of n (this is not universally true because of randomness in sampling). And these higher coverage levels

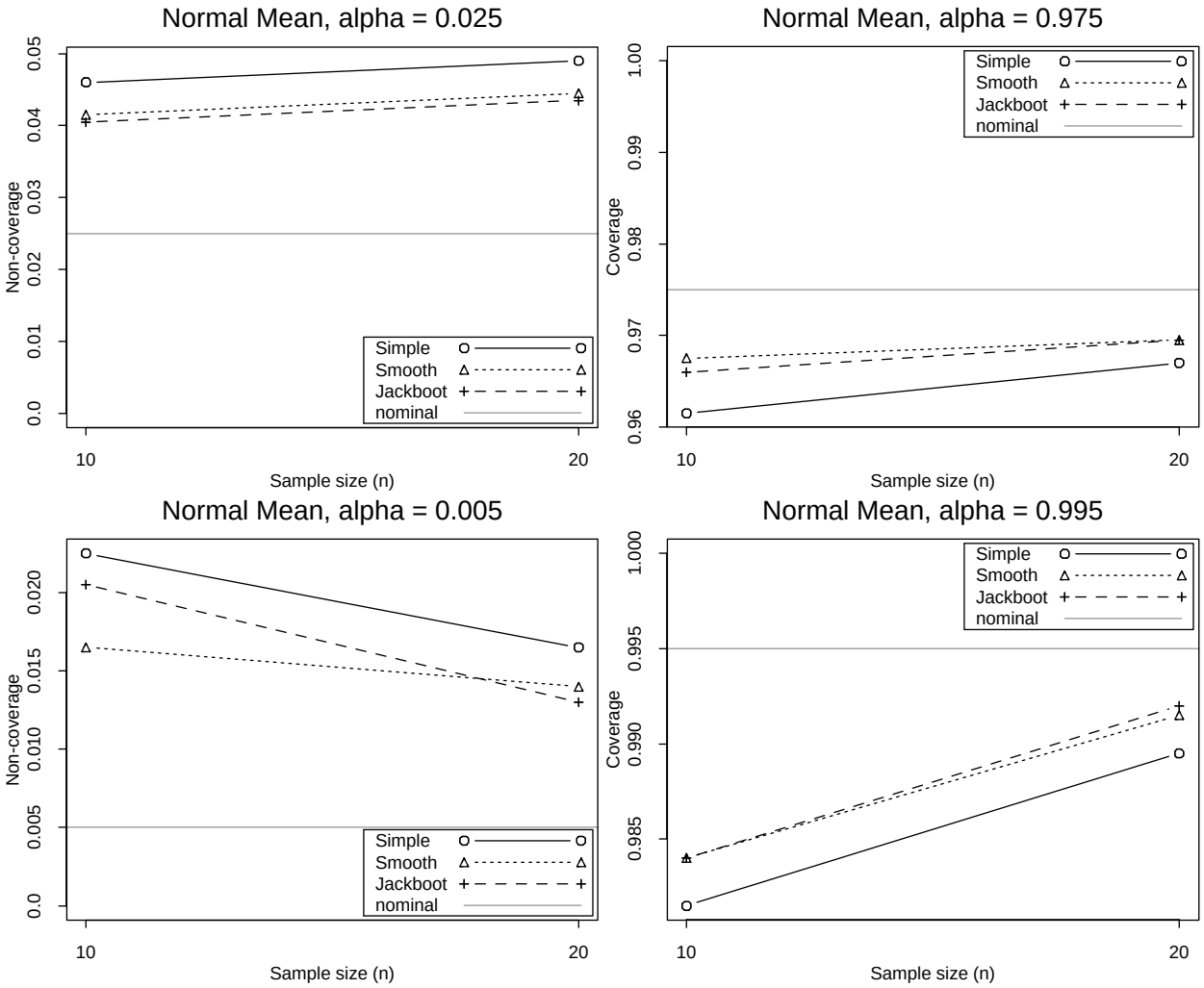

Figure 1: Coverage Accuracy. One-sided bootstrap-percentile confidence intervals for the sample mean of normal data. Solid lines are for simple bootstrap sampling, dotted lines for bootknife sampling. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated, and two sets of bootstrap samples were generated—one of size $B = 1999$ using simple bootstrap sampling, and one of size $B = 2000$ using stratified bootknife sampling. Bootstrap means for the smooth bootstrap were computed by adding a random normal variate with mean 0 and variance s^2/n^2 to the simple bootstrap mean. The standard errors are approximately $(.025 * .975/2000) = .0035$ for intervals with nominal coverage of 0.025 or 0.975, and about 0.0016 for nominal coverages of 0.005 or 0.995.

are closer to nominal, except for lower endpoints for the mean of exponential data, where the bootstrap percentile interval tended to over-cover.

Stratification does not affect the unbiasedness property in (8)—the result should be interpreted as holding for a bootstrap sample chosen randomly from the B such samples, so that o^* is still randomly chosen from the numbers $1, \ldots, n$.

In some situations it is appropriate to omit multiple observations. For example, when bootstrapping a linear regression problem by resampling residuals, if there are p coefficients including the intercept, then unbiased estimates of residual variance are obtained by omitting p randomly-chosen observations when generating a bootstrap sample (under the usual linear regression assumptions).

4 Summary

Both smoothing with the standard choice of smoothing parameter, and bootknife sampling, provide unbiased estimates for the variance of a sample mean. More generally, they provide bootstrap distributions with larger and approximately-correct variance, and larger and more accurate coverage probabilities.

Both procedures adapt readily to stratified sampling and two-sample applications.

These methods are available in S+RESAMPLE, available for free download from `www.insightful.com/downloads/libraries`.

References

[1] A. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, 1997.

[2] B. Efron. Bootstrap methods: another look at the jackknife. (with discussion). *Annals of Statistics*, 7:1–26, 1979.

[3] B. Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139 – 172, 1981.

[4] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[5] P. Hall. On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, 14(4):1453–1462, 1986.

[6] Tim C. Hesterberg. The bootstrap and empirical likelihood. In *Proceedings of the Statistical Computing Section*, pages 34–36. American Statistical Association, 1997.

[7] Tim C. Hesterberg. Smoothed bootstrap and jackboot sampling. Research Department 87, MathSoft, Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109, 1999.

[8] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995.

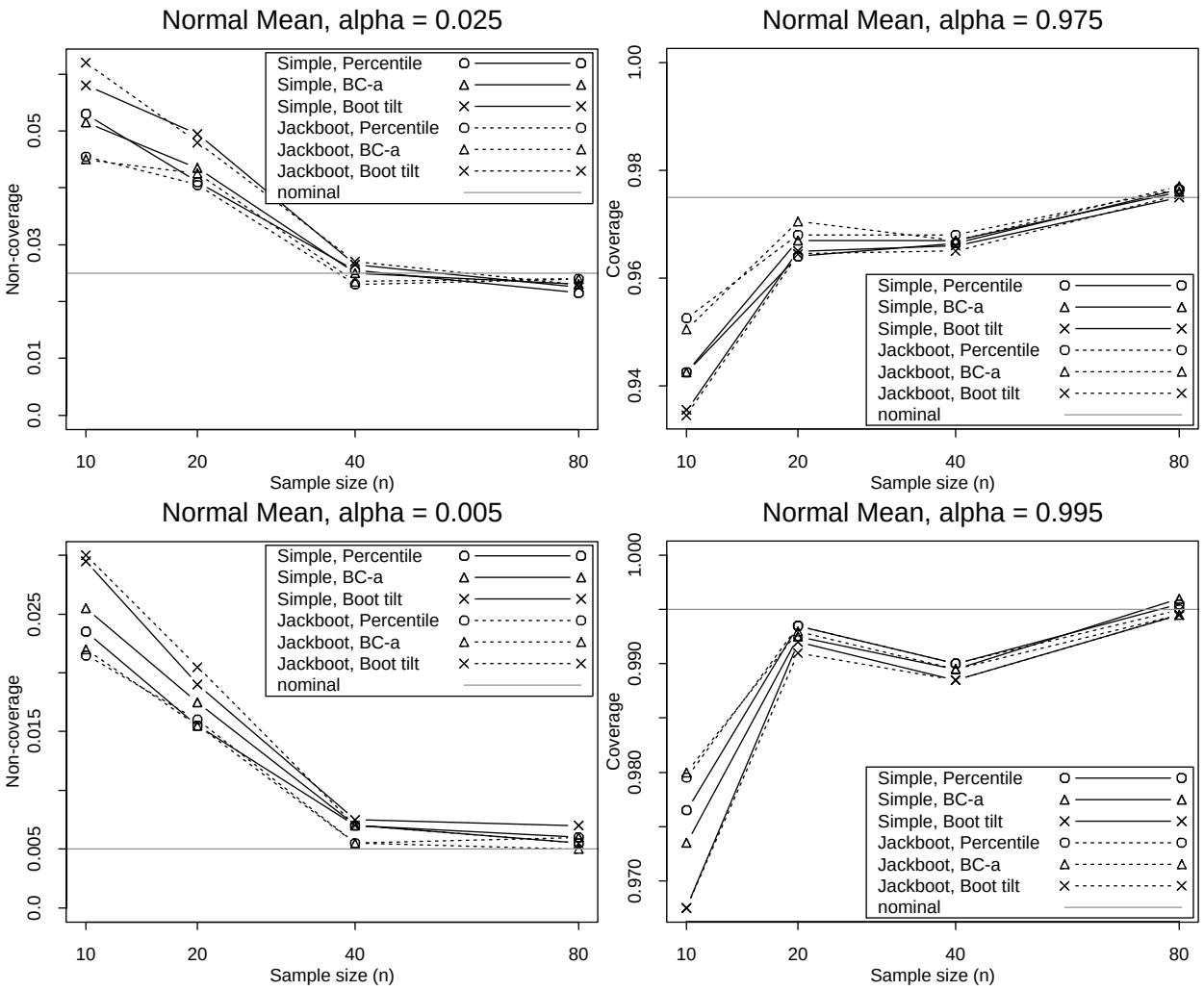

Figure 2: Coverage Accuracy. One-sided confidence intervals for the sample mean of normal data. Solid lines are for simple bootstrap sampling, dotted lines for bootknife sampling. Results are from 2000 bootstrap experiments; in each experiment a random data set was generated, bootstrap samples were generated using simple random bootstrap sampling or (stratified) bootknife sampling, and one of each kind of bootstrap confidence interval was generated. using $B = 200$ bootstrap samples for the bootstrap tilting intervals and $B = 1999$ bootstrap samples for the bootstrap percentile and BC-a intervals. The standard errors are approximately $(.025 * .975/2000) = .0035$ for intervals with nominal coverage of 0.025 or 0.975, and about 0.0016 for nominal coverages of 0.005 or 0.995.

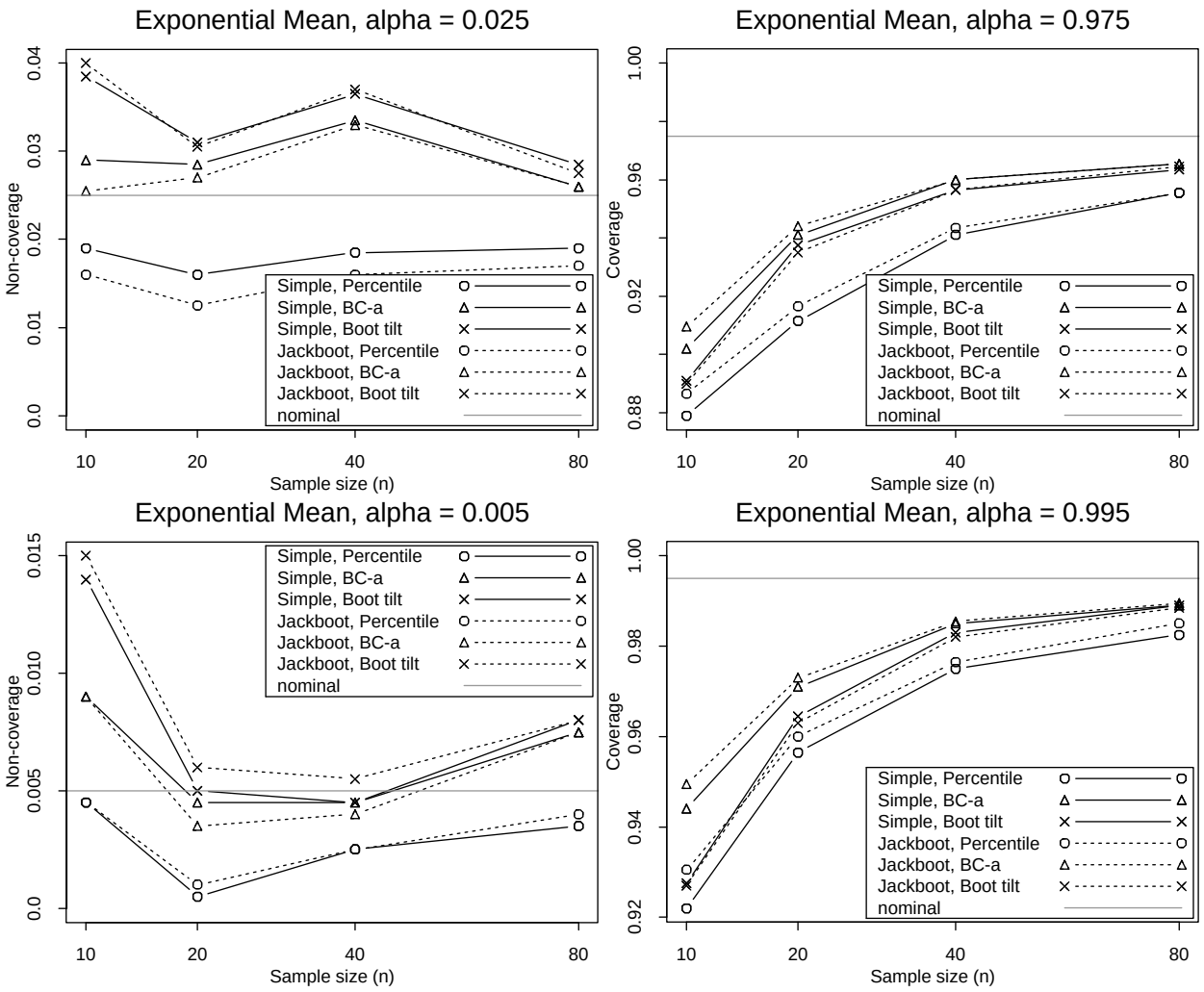

Figure 3: Coverage Accuracy. One-sided confidence intervals for the sample mean of exponential data. Other details are the same as for Figure 2