Resampling Data:
Using Bootstraps

S. Sawyer — Washington University — March 12, 2005

1. Introduction. Suppose that we want to estimate a parameter θ that depends on a random quantity sample $X = (X_1, X_2, \ldots, X_n)$ in a complicated way. For example, θ might be the sample variance of X or the log sample variance. If the X_i are vector valued, θ could be the Pearson correlation coefficient.

Assume that we have an estimator

$$\widehat{\theta}(X_1, X_2, \ldots, X_n) = \phi(X_1, X_2, \ldots, X_n) \tag{1.1}$$

of θ but do not know the probability distribution of $\phi(X)$ given θ. This means that we cannot estimate the error involved in estimating θ by $\phi(X)$. In particular, we cannot tell if we can conclude $\theta \neq 0$ from an observed $\phi(X) \neq 0$, no matter how large.

Can we get a confidence interval for θ depending only on the observed $X_1, X_2, \ldots, X_n$, or test $H_0 : \theta = \theta_0$ just using the data $X_1, X_2, \ldots, X_n$?

The *bootstrap* and *jackknife* are two methods for answering these questions without any prior assumptions about the distribution of $\widehat{\theta}(X)$. See the accompanying Jackknife Handout for details about the jackknife. See Efron (1982, 1987) and Efron and Tibshirani (1993) in the reference section below for a more detailed discussion of bootstrap methods.

2. The Basic Bootstrap Recipe: Given a sample $X = (X_1, X_2, \ldots, X_n)$ of size n, a *bootstrap resample* of X is a sample

$$X^* = (X_1^*, X_2^*, \ldots, X_n^*) \tag{2.1}$$

where each value X_j^* in (2.1) is a random sample from $X = (X_1, X_2, \ldots, X_n)$ *with replacement*. That is, given distinct values $X_1, X_2, \ldots, X_n$,

$$\Pr(X_j^* = X_i) = 1/n, \quad 1 \leq i \leq n \tag{2.2}$$

with independent choices of X_j^* for $1 \leq j \leq n$. In particular, repeated values $X_{j_1}^* = X_{j_2}^* = X_i$ are allowed. Since the sample size of X^* is n, repeated values of X_j^* means that other values in X must be left out.

We repeat this process B times to form B independent resamples $(X^*)^{(1)}, (X^*)^{(2)}, \ldots, (X^*)^{(B)}$. The *bootstrap resampled values* for the estimator (1.1) are

$$\widehat{\theta}_k^* = \widehat{\theta}((X^*)^{(k)}) \qquad 1 \leq k \leq B \tag{2.3}$$

The values $\widehat{\theta}_k^*$ in (2.3) will be treated as an independent random sample with mean θ. Typical values of B are $B = 50$, $B = 1000$, or $B = 10{,}000$. In particular, the number of resampled values $\widehat{\theta}_k^*$ that we can use is not limited by the sample size n.

3. A Motivation for the Bootstrap: Let $F(z) = P(X_i \le z)$ be the distribution function of the sample $X_1, X_2, \ldots, X_n$. One can argue that, given $X_1, X_2, \ldots, X_n$, our best estimate of $F(z)$ is the sample distribution function $\widehat{F}_s(z)$, where $\widehat{F}_s(z)$ is the distribution function that puts mass $1/n$ at each point $z = X_i$ (with appropriate changes for multiplicity if X_i has tied values). By (2.2), each bootstrap resample X^* is a random sample of size n from $\widehat{F}_s(z)$.

If the estimator $\widehat{\theta}(X)$ in (1.1) does not depend on the order of the values X_i, then $\widehat{\theta}(X)$ is a function of $\widehat{F}_s(z)$, and the process of obtaining $\widehat{\theta}(X)$ from X can then be pictured as

$$\begin{array}{ccccc} (\theta,\, F) & \longrightarrow & \widehat{F}_s & \longrightarrow & \widehat{\theta}(X) = \widehat{\theta}(\widehat{F}_s) \\ \text{(background} & & \text{data} & & \\ \text{population)} & & X_1, \ldots, X_n & & \end{array} \tag{3.1}$$

In general, P-values for $H_0 : \theta = \theta_0$ and confidence intervals for θ depend on knowing what $X_1, \ldots, X_n$ might have looked like for samples for various values of θ. Using the idea that our best estimate of $F(z)$ is $\widehat{F}_s(z)$, and that bootstrap resamples X^* are samples from $\widehat{F}_s(z)$, we can modify (3.1) to

$$\begin{array}{ccccc} (\widehat{\theta, F_s}) & \longrightarrow & \widehat{F}_s^* & \longrightarrow & \widehat{\theta}(X^*) = \widehat{\theta}(\widehat{F}_s^*) \\ \text{(estimated} & & \text{new data} & & \\ \text{population)} & & X_1^*, \ldots, X_n^* & & \end{array} \tag{3.2}$$

We then make inferences about the distribution of $\widehat{\theta}(X)$ given θ by studying the distribution of $\widehat{\theta}(X^*)$ given X. We can do this theoretically, or, usually with less effort, we can simulate the distribution of $\widehat{\theta}(X^*)$ by computing $\widehat{\theta}_k^* = \widehat{\theta}((X^*)^{(k)})$ for B simulated resampled data sets $(X^*)^{(k)}$.

4. Bootstrap Tests, Estimates, and CIs: For any value of B, we can use the resampled values $\widehat{\theta}_k^*$ in the same way as jackknife pseudovalues (see the Jackknife Handout). Specifically, we can estimate a mean and variance for θ by

$$\widehat{\theta}^*_{\text{mean}}(X) = \frac{1}{B}\sum_{k=1}^{B} \widehat{\theta}_k^*, \qquad V_\theta(X) = \frac{1}{B-1}\sum_{k=1}^{B} \left(\widehat{\theta}_k^* - \widehat{\theta}^*_{\text{mean}}(X)\right)^2 \tag{4.1}$$

and a corresponding 95% confidence interval for θ

$$\left(\widehat{\theta}^*_{\rm mean}(X) - 1.960\sqrt{\frac{1}{B}V_\theta(X)}\,,\quad \widehat{\theta}^*_{\rm mean}(X) + 1.960\sqrt{\frac{1}{B}V_\theta(X)}\right) \tag{4.2}$$

However, it is usually not much more work to generate $B \geq 1000$ bootstrap resampled values than for a smaller value of B, and in that case we can usually do better than (4.1)–(4.2).

Assume that the distribution of $\widehat{\theta}(X)$ given θ is symmetric about θ for each θ. Then the distribution of $\widehat{\theta}(X^*)$ will also be symmetric about θ, and a good estimator of θ will be

$$\widehat{\theta}^*_{\rm med}(X) = \text{median}\{\,\widehat{\theta}^*_k : 1 \leq k \leq B\} \tag{4.3}$$

Similarly, a 95% confidence interval for θ can be constructed as the upper and lower 2.5% quantiles of the sampled values $\widehat{\theta}^*_k$. Specifically, the *bootstrap percentile* 95% confidence interval for θ is

$$\big(\widehat{\theta}^*_{(U)},\, \widehat{\theta}^*_{(B+1-U)}\big) \tag{4.4}$$

In (4.4), $\widehat{\theta}^*_{(k)}$ are the values $\widehat{\theta}^*_k$ sorted in increasing order, so that

$$\widehat{\theta}^*_{(1)} \;\leq\; \widehat{\theta}^*_{(2)} \;\leq\; \ldots \;\leq\; \widehat{\theta}^*_{(B)}$$

and $U = 0.025B$, with U rounded down to the nearest integer unless this would result in $U = 0$ (in which case $U = 1$). In particular, if $B = 1000$, $U = 25$ and $B + 1 - U = 951$. In general, the lower value $\widehat{\theta}^*_{(U)}$ in (4.4) is the $U^{\rm th}$ value from the bottom of the sorted array $\widehat{\theta}^*_{(k)}$ and $\widehat{\theta}^*_{(B+1-U)}$ is the $U^{\rm th}$ value from the top.

One advantage of (4.4) over (4.2) is that the endpoints of the confidence interval (4.4) are realizable values of $\theta = \widehat{\theta}(X)$. For example, assume that θ is a sample proportion, so that we know $0 \leq \theta \leq 1$ in advance, and that $0 \leq \widehat{\theta}(X) \leq 1$ for any sample X. The endpoints of (4.2) can be well below zero or well above 1. However, each value $\widehat{\theta}^*_k = \widehat{\theta}((X^*)^{(k)})$ is a value of $\widehat{\theta}(X)$ for some sample X, so that the endpoints of (4.4) are within the range $0 \leq \theta \leq 1$. (Of course, one could always round up the lower endpoint of (4.2) to zero and round down the upper endpoint to one if needed, with equal precision or imprecision, but with (4.4) this is not necessary.)

5. Bootstrap Samples are Invariant. Recall that, in some cases, the jackknife seemed to work better if we applied the method to $\log\big(\widehat{\theta}(X)\big)$ instead of $\widehat{\theta}(X)$ and then exponentiated the resulting estimates and confidence

interval. However, by (2.3),

$$\widehat{\log\theta}_k^* = \log\bigl(\widehat{\theta}_k^*\bigr)$$

and more generally

$$\widehat{g(\theta)}_k^* = g\bigl(\widehat{\theta}_k^*\bigr) \tag{5.1}$$

for any monotonic function $g(\theta)$ of θ, where $\widehat{g(\theta)}_k^*$ means the k^{th} bootstrap sampled value for $g(\widehat{\theta}(X))$. In other words, replacing θ by any monotonic function of θ replaces the bootstrap resamped values for θ by the same monotonic function of those values. This implies that generating the median estimate (4.3) and confidence interval (4.4) for $\log(\theta)$ and then exponentiating the results leads to exactly the same results as if we had worked with θ in the first place.

This suggests a second advantage of the bootstrap percentile confidence interval (4.4) over the jackknife-like confidence interval (4.2). The proper scaling of θ is not clear in many instances. If θ is the size of a dust particle, it may not be clear whether $\theta = x$ (the width), $\theta = x^2$ (surface area), or $\theta = x^3$ (volume or weight) is more important. One may be tempted to log-transform the values in (4.2) to eliminate outliers, or not to log-transform to avoid making the distribution of the values $\widehat{\theta}_k^*$ too even. The invariance (5.1) means that we obtain the same confidence intervals and median estimate in (4.3)–(4.4) for any monotonic transformation of θ, so that we do not need to worry. (Of course, this invariance does not apply to the basic data X_i itself, as opposed to the parameter θ, unless the form of $\widehat{\theta}(X)$ is suitable.)

Recall that the bootstrap percentile estimators (4.3) and (4.4) were motivated by the assumption that the distribution of $\widehat{\theta}(X)$, as a function of θ, was always symmetric about θ. The basic invariance property (5.1) means that this assumption can be weakened to assume that there exists a strictly-increasing monotonic function $g(\theta)$ such that $g\bigl(\widehat{\theta}(X)\bigr)$, as a function of θ, is always symmetrically distributed about $g(\theta)$. This is potentially a larger class of estimators.

6. How Many Values are Left Out of a Bootstrap Resample? Given a sample $X_1, X_2, \ldots, X_n$ of size n, the probability that a particular value X_i is left out of a resample $X^* = (X_1^*, \ldots, X_n^*)$ is

$$\Pr(X_j^* \neq X_i, 1 \le j \le n) = \left(1 - \frac{1}{n}\right)^n$$

by (2.2), where we assume that the X_i are distinct for simplicity. This means that the *expected proportion* of values X_i that are not represented in

a particular resample X^* is also $\big(1-(1/n)\big)^n$. If n is large, this proportion is approximately $e^{-1} \approx 0.37$. This means that approximately 37% of the values X_i are left out of any particular resample X^*, although (of course) these values may appear in other bootstrap resamples.

Recall that one of the disadvantages of the jackknife was that the pseudovalues of the delete-one jackknife were too close together for large n, so that a delete-k jackknife with a correspondingly smaller sample size had to be used instead. Since bootstrap resamples normally drop about 37% of the original sample values for all values of n, bootstrap resample recipes should be valid for all sample sizes.

7. The Parametric Bootstrap: In the recipe (3.1), we approximated the unknown distribution (θ, F) by the sample distribution function $\widehat{F}_s(z)$. Suppose, instead, that we assume that we know the distribution of X_i

$$F(z) = F(\theta, z)$$

within a parameter θ. In this case, we could use $F(\widehat{\theta}, z)$ instead of $\widehat{F}_s(z)$ in (3.1). The recipe (3.2) would then mean that the resamples X_j^* would be taken from $F(\widehat{\theta}, z)$ instead of from $\widehat{F}_s(z)$.

For example, assume that $F(\theta, z)$ is the double exponential distribution with density $(1/2)e^{-|x-\theta|}$. The maximum likelihood estimator of θ for this family is

$$\widehat{\theta}(X_1, X_2, \ldots, X_n) \;=\; \widehat{\theta}_m \;=\; \text{median}\{X_i : 1 \le i \le n\}$$

If we use $F(\widehat{\theta}, z)$ instead of $\widehat{F}_s(z)$ in (3.2), then the resamples $X^* = (X_1^*, \ldots, X_n^*)$ are independent random variables with density

$$f(x) = (1/2)e^{-|x-\widehat{\theta}_m|}$$

instead of samples with replacement from the n values $X_1, X_2, \ldots, X_n$. Since this procedure uses more information about the distribution of X_i, this should give a more reasonable set of resampled values $\widehat{\theta}_k^* = \widehat{\theta}((X^*)^{(k)})$.

Using $F(\widehat{\theta}, z)$ in (3.2) in this way instead of $\widehat{F}_s(z)$ is called the *parametric bootstrap*. In contrast, using $\widehat{F}_s(z)$ is the *nonparametric bootstrap*. These ideas are connected by the fact that, within the class of all possible distribution functions $F(z)$, it can be argued that the maximum likelihood estimator $\widehat{F}^*_{s,\text{MLE}}(X_1, \ldots, X_n)(z)$ of $F(z)$ is the sample distribution function $\widehat{F}_s(z)$.

8. Biased Bootstrap Estimates: Suppose $\widehat{\theta}^*_{\text{med}}(X) < \widehat{\theta}(X)$, or equivalently

$$\widehat{\theta}^*_{\text{med}}(X) \;=\; \text{median}\{\,\widehat{\theta}((X^*)^{(k)}) : 1 \le k \le B\} \;<\; \widehat{\theta}(X) \tag{8.1}$$

In this case, one could argue that estimating $\theta = \widehat{\theta}^*_{\rm med}(X)$ is not sensible, since (8.1) implies that the resampling process tends to yield smaller values of θ. If the process of sampling X from nature is like the process of resampling $(X^*)^{(k)}$ from X (which is implicit in the bootstrap recipe (3.1)–(3.2)), then (8.1) should imply $\theta > \widehat{\theta}(X)$, not $\theta = \widehat{\theta}^*_{\rm med}(X) < \widehat{\theta}(X)$.

One way of quantifying this concern is the notion of median bias. Define "bias" as the proportion of resampled values $\widehat{\theta}^*_k$ that are less than or equal to $\widehat{\theta}(X)$:

$$\text{bias} \;=\; \frac{\#\{\, k \le B :\; \widehat{\theta}((X^*)^{(k)}) \;\le\; \widehat{\theta}(X)\,\}}{B} \tag{8.2}$$

where $\#A$ means the number of elements in the set A. This is an approximation of $P(\widehat{\theta}(X^*) \le \widehat{\theta}(X) \mid X)$, where bias $= 1/2$ means that there is no median bias about $\widehat{\theta}(X)$ in the resampled values $(X^*)^{(k)}$.

A reasonable rule of thumb is not to be concerned if $0.40 \le$ bias ≤ 0.60 or $0.35 \le$ bias ≤ 0.65. Otherwise, one should consider finding a better-behaved estimator $\widehat{\theta}(X)$ of θ.

LITERATURE CITED

1. Efron, B. (1982) The jackknife, the bootstrap, and other resampling plans. CBMS Regional Conference Series in Applied Mathematics Vol. 38, Soc. Ind. Appl. Math..

2. Efron, B. (1987) Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82,** 171–200 (with discussion).

3. Efron, B., and R. J. Tibshirani (1993) An introduction to the bootstrap. Chapman and Hall, New York.