Introduction to Bootstrap Methods in Statistics

by

Dimitris N. Politis
Purdue University

Technical Report #93-49

Department of Statistics
Purdue University

September 1993

Introduction to bootstrap methods in statistics

Dimitris N. Politis

Department of Statistics
Purdue University
W. Lafayette, IN 47907

Abstract

A tutorial introduction to the recently developed resampling and subsampling methods for statistical inference, i.e., the bootstrap, the jackknife, and their variations, is presented.

1. Resampling and the bootstrap

The goal of this paper is to provide a readable, self-contained introduction to the bootstrap and jackknife methodology for statistical inference; in particular, the focus is on the derivation of confidence intervals in general situations.

1.1 The general non-parametric set-up. Suppose that $\mathbf{X} = (X_1, \ldots, X_N)$ is an independent, identically distributed (i.i.d.) sample from a population with distribution F. In other words, $F(x) = Prob(X_i \leq x)$, for $i = 1, \ldots, N$, where x is any real number. The sample is studied in order to estimate a certain parameter $\theta(F)$ associated with the distribution F. A statistic $T = T(\mathbf{X})$ might be used to estimate $\theta(F)$ from the data. However, a measure of the statistical accuracy of the point estimator $T(\mathbf{X})$ is also desired. For example, the bias and the variance of the estimator T are of interest, and are defined as follows:

$$Bias_F(T) = E_F T(\mathbf{X}) - \theta(F) \tag{1}$$

$$Var_F(T) = E_F T^2(\mathbf{X}) - [E_F T(\mathbf{X})]^2 \tag{2}$$

where E_F denotes expectation under the F distribution.

To fix ideas, consider that $\theta(F)$ is a location parameter, say the mean or median of F, and $T(\mathbf{X})$ is the corresponding sample statistic (sample mean, sample median, etc.). In many practical situations the Central Limit Theorem can be invoked to assert that the estimator $T(\mathbf{X})$ is approximately distributed as a Gaussian random variable. This will typically be true for most 'good' estimators, provided the sample size N is large enough, in which case the estimator is said to be *asymptotically* normal, and an approximate interval estimate, i.e., a confidence interval, for $\theta(F)$ can be formed, in addition to the point estimate $T(\mathbf{X})$.

1.2 Confidence intervals based on asymptotic normality. If the bias $Bias_F(T)$ is negligible (compared to the square root of the variance $Var_F(T)$), a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ will be of the usual form

$$[T(\mathbf{X}) - z\sqrt{Var_F(T)}, T(\mathbf{X}) + z\sqrt{Var_F(T)}], \tag{3}$$

where $z = z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution. If $Bias_F(T)$ is not negligible, the confidence interval must be adjusted appropriately; generally, a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ will be given by

$$[T(\mathbf{X}) - Bias_F(T) - z\sqrt{Var_F(T)}, T(\mathbf{X}) - Bias_F(T) + z\sqrt{Var_F(T)}] \tag{4}$$

Note that the aforementioned confidence interval is based on the fact that the asymptotic normal distribution of $T(\mathbf{X}) - \theta(F)$ does not depend on the unknown parameter $\theta(F)$; in other words, $T(\mathbf{X}) - \theta(F)$ is an approximate *pivot*. However, to formulate this confidence interval one needs to know $Bias_F(T)$ and $Var_F(T)$.

Estimates of $Bias_F(T)$ and $Var_F(T)$ might be available in the statistical literature for different problems. For example, if $T(\mathbf{X}) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the sample mean, and $\theta(F) = E_F X_1$ is the population mean, then it is well known that $Bias_F(T) = 0$, and $Var_F(T) =$

$\frac{1}{N}Var_F(X_1)$, where $Var_F(X_1)$ can be estimated by the sample variance $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$. If $T(\mathbf{X})$ is the sample median and $\theta(F)$ is the population median, estimates of $Bias_F(T)$ and $Var_F(T)$ can still be calculated (cf. Lehmann (1983)), but are substantially more complicated.

The bootstrap (and the closely related jackknife (cf. Efron (1979, 1982)) could alternatively be used to easily obtain estimates of $Bias_F(T)$ and $Var_F(T)$ for a wide variety of statistics $T(\mathbf{X})$. However, before going into that, let us look at this problem from a different angle.

1.3 The usefulness of Monte Carlo randomization. Suppose, for the sake of argument, that the population and its distribution F were in fact known. Then, $Bias_F(T)$ and $Var_F(T)$ could be calculated exactly by analytical methods, or approximately by Monte Carlo simulation, in case the analytical computation is difficult.

The idea behind Monte Carlo simulation is the following. Since the population is considered known, we can draw any number of i.i.d. samples from it. Suppose that we draw B samples, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(B)}$, where each sample consists of N i.i.d. observations from the population F. If B is large enough, the strong law of large numbers can be invoked to claim that

$$E_F g(T(\mathbf{X})) \simeq \frac{1}{B}\sum_{i=1}^{B} g(T(\mathbf{X}^{(i)})) \tag{5}$$

where $g(\cdot)$ is some function, e.g. $g(x) = x$ or $g(x) = x^2$. Then we would have

$$Bias_F(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{(i)}) - \theta(F) \tag{6}$$

$$Var_F(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{(i)})]^2 \tag{7}$$

But if the population is considered known, we could also directly evaluate the sampling distribution of the approximate pivot $T(\mathbf{X}) - \theta(F)$, without reference to the asymptotic (for large N) normal distribution. Define $P_F(A)$ to be the probability of event A occurring, under the assumption that the population has distribution F, and let

$$Dist_{T,F,\theta}(x) = P_F(T(\mathbf{X}) - \theta(F) \leq x). \tag{8}$$

Knowledge of $Dist_{T,F,\theta}(x)$, for all real x, would immediately yield a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ in the form

$$[T(\mathbf{X}) - q(1-\alpha/2), T(\mathbf{X}) - q(\alpha/2)], \tag{9}$$

where $q(\alpha/2)$ and $q(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T,F,\theta}(x)$ distribution respectively. The above confidence interval is equal-tailed, meaning that the probability the interval's left end-point is bigger than $\theta(F)$ is equal to the probability the interval's right end-point is smaller than $\theta(F)$. Other constructions (e.g., symmetric, shortest length, etc.) for confidence intervals are also available (cf. Hall (1988, 1992)) and possess some interesting theoretical properties; nevertheless, the confidence intervals that are most often used in practice are equal-tailed (cf. Efron and Tibshirani (1993)).

Again, although F is considered known, the analytical evaluation of $Dist_{T,F,\theta}(x)$ may be difficult, and we might resort to Monte Carlo. Observe that $Dist_{T,F,\theta}(x)$ is just a shifted (centered) version of

$$Dist_{T,F}(x) = P_F(T(\mathbf{X}) \leq x) \tag{10}$$

so that $Dist_{T,F,\theta}(x) = Dist_{T,F}(x + \theta(F))$. If we define the indicator function of event A by the formula

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{else} \end{cases}$$

then, using equation (5) with $g(T(\mathbf{X})) = \mathbf{1}(T(\mathbf{X}) \leq x)$, and the fact that $E_F\mathbf{1}(A) = P_F(A)$, we have

$$Dist_{T,F}(x) \simeq \frac{1}{B}\sum_{i=1}^{B}\mathbf{1}(T(\mathbf{X}^{(i)}) \leq x) = \frac{1}{B}(\#T(\mathbf{X}^{(i)}) \leq x) \tag{11}$$

i.e., the theoretical probability should be approximately equal to the observed sample proportion if B is large.[1] From equation (11), the quantiles of $Dist_{T,F}(x)$ and of $Dist_{T,F,\theta}(x)$) can be approximately calculated, and the confidence interval (9) constructed.

1.4 The bootstrap principle. To summarize, *if* the population and its distribution F *were* known, then we would be able to calculate (analytically or by Monte Carlo simulations)

[1] Note that $(\#T(\mathbf{X}^{(i)}) \leq x)$ reads: number of the $T(\mathbf{X}^{(i)})$'s among $T(\mathbf{X}^{(1)}), \ldots, T(\mathbf{X}^{(B)})$ that are observed to be less or equal to x.

$Bias_F(T)$, $Var_F(T)$, and $Dist_{T,F,\theta}(x)$. However, in the practical problem the population and its distribution F are *not* known. The bootstrap method now is an outcome of the following simple idea: *since you do not have the whole population, do the best with what you* do *have, which is the observed sample* $\mathbf{X} = (X_1, \ldots, X_N)$.

In other words, the bootstrap method (cf. Efron (1979)) amounts to treating your observed sample as *if* it *exactly* represented the whole population. In this fashion, the Monte Carlo procedure in which B i.i.d. samples were drawn from the population is modified to read:

- Draw B i.i.d. samples $\mathbf{X}^{*(1)}, \ldots, \mathbf{X}^{*(B)}$ (each of size N) from the sample population consisting of the observations $\{X_1, \ldots, X_N\}$. In the bootstrap terminology, these B i.i.d. samples are called *resamples.* Of course, drawing an i.i.d. sample from a finite population such as $\{X_1, \ldots, X_N\}$, amounts to sampling with replacement from the set $\{X_1, \ldots, X_N\}$.

Note that, as the whole population has distribution F, the sample population has distribution $\hat{F}$, the so-called *empirical* distribution, which is defined as

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(X_i \le x) = \frac{1}{N}(\# X_i \le x). \tag{12}$$

To elaborate, in order to form the ith resample $\mathbf{X}^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, we sample with replacement from the set $\{X_1, \ldots, X_N\}$, or, using a different terminology, we take an i.i.d. sample of size N from a population with distribution $\hat{F}$.

1.5 The bootstrap as a 'plug-in' method. This last observation suggests a different perspective for the implementation of the bootstrap as a simple '*plug-in*' method. Namely, if at a certain formula the unknown distribution F appears, you just substitute $\hat{F}$ in place of F to get its bootstrap approximation. For example, the bootstrap approximations to $Bias_F(T)$ and $Var_F(T)$ are given (cf. equations (1), (2)) by

$$Bias^*(T) = Bias_{\hat{F}}(T) = E_{\hat{F}}T(\mathbf{X}) - \theta(\hat{F}) \tag{13}$$

$$Var^*(T) = Var_{\hat{F}}(T) = E_{\hat{F}}T^2(\mathbf{X}) - [E_{\hat{F}}T(\mathbf{X})]^2. \tag{14}$$

It should be noted that $\theta(\hat{F})$ is just the sample statistic corresponding to the population parameter $\theta(F)$. In most cases, the statistic $T(\mathbf{X})$ is chosen to be just $\theta(\hat{F})$. For example, if $\theta(F)$ is the population median, then we might want to use the sample median to estimate it, i.e. $T(\mathbf{X}) = \theta(\hat{F})$. We will henceforth assume that $\theta(\hat{F}) \equiv T(\mathbf{X})$ for simplicity; in a different situation the 'plug-in' principle can be appropriately modified.

1.6 A parametric set-up. This 'plug-in' viewpoint permits one to see how the bootstrap would work in a parametric problem as well. For example, if the distribution $F(x) = F_\theta(x)$ is known up to some parameter θ, then the parametric bootstrap method would be to approximate quantities such as $Bias_F(T)$, $Var_F(T)$, and $Dist_{T,F}(x)$ by $Bias_{F_{\hat{\theta}}}(T)$, $Var_{F_{\hat{\theta}}}(T)$, and $Dist_{T,F_{\hat{\theta}}}(x)$ respectively, where $\hat{\theta} = T(\mathbf{X})$ is the estimated (from our sample) value of the parameter θ. All the Monte Carlo approximations remain valid, except that in the parametric set-up, to form the ith resample $\mathbf{X}^{*(i)} = (X_1^{*(i)}, \ldots, X_N^{*(i)})$, we take an i.i.d. sample from a population with distribution $F_{\hat{\theta}}$. Note that in parametric problems, the theory of Maximum Likelihood estimation and Fisher information are traditionally used to get point and interval estimates of the unknown parameter θ (cf. Miller (1986)); however, the bootstrap will tend to give more accurate estimates in general (cf. Hall (1992), Efron and Tibshirani (1993)). Having said that, let us return and focus our attention on the general non-parametric problem, that is, the problem where F is completely unknown, since here the bootstrap is more urgently needed.

1.7 Construction of bootstrap confidence intervals. As was mentioned before, to calculate $Bias_F(T)$ and $Var_F(T)$ we might have to resort to Monte Carlo simulation even if the distribution F were known; see equations (6), (7). Thus to calculate $Bias_{\hat{F}}(T)$ and $Var_{\hat{F}}(T)$ we might use the following Monte Carlo approximations:

$$Bias^*(T) = Bias_{\hat{F}}(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{*(i)}) - \theta(\hat{F}) \tag{15}$$

$$Var^*(T) = Var_{\hat{F}}(T) \simeq \frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{*(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{*(i)})]^2 \tag{16}$$

The above mentioned bootstrap approximations to $Bias_F(T)$ and $Var_F(T)$ can be used to yield a confidence interval for $\theta(F)$ based on the normal approximation of equation (4).

Alternatively, we can by-pass the normal approximation and set confidence intervals for $\theta(F)$ based on the exact distribution of the pivotal quantity $T(\mathbf{X}) - \theta(F)$ given in equation (8).

Of course, this exact distribution is not known, but a bootstrap approximation is available. More specifically, the bootstrap approximation to $Dist_{T,F,\theta}(x)$ is given by

$$Dist^*_{T,\theta}(x) = Dist_{T,\hat{F},\theta}(x) = P_{\hat{F}}(T(\mathbf{X}) - \theta(\hat{F}) \leq x) \tag{17}$$

and an equal-tailed $(1-\alpha)100\%$ *bootstrap* confidence interval for $\theta(F)$ would be

$$[T(\mathbf{X}) - q^*(1-\alpha/2), T(\mathbf{X}) - q^*(\alpha/2)], \tag{18}$$

where $q^*(\alpha/2)$ and $q^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{T,\hat{F},\theta}(x)$ distribution respectively. It should be noted at this point that this just one of many possible ways to construct a bootstrap confidence interval; see Efron and Tibshirani (1993, chapter 22) for a thorough discussion, and DiCiccio and Romano (1988), Hall (1988, 1992) for a comparison of bootstrap confidence intervals. Note that in the terminology of Hall (1988), equation (18) represents a confidence interval based on the 'hybrid' method, whereas in Hall (1992) equation (18) is the 'other percentile method' confidence interval; to avoid the confusion we will refer to equation (18) as the *pivotal* method for bootstrap confidence intervals.

The bootstrap distributions $Dist_{T,\hat{F},\theta}(x)$ and $Dist_{T,\hat{F}}(x) = P_{\hat{F}}(T(\mathbf{X}) \leq x)$ –and therefore their quantiles as well– can be easily evaluated by Monte Carlo as follows:

$$Dist_{T,\hat{F}}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} 1(T(\mathbf{X}^{*(i)}) \leq x) = \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x) \tag{19}$$

and

$$Dist^*_{T,\theta}(x) = Dist_{T,\hat{F},\theta}(x) = Dist_{T,\hat{F}}(x + \theta(\hat{F})) \simeq \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x + \theta(\hat{F})) \tag{20}$$

It should be noted that since the resampling procedure implicit in equation (20) is done with the sample $X_1, \ldots, X_N$ being *fixed* and playing the role of a population with distribution $\hat{F}$, the sample statistic $\theta(\hat{F})$ is just a fixed number, calculated once and for all from the original sample $X_1, \ldots, X_N$. In the bootstrap literature, the terminology is that the resampling is done *conditionally* on the data $X_1, \ldots, X_N$.

1.8 Higher order accuracy of the bootstrap and studentization. The reason for the success and popularity of the bootstrap methodology is twofold: (a) it provides answers (confidence intervals, standard error estimates, etc.) in complicated situations, and (b) it provides *more accurate* answers in standard settings, more accurate as compared to the ubiquitous normal approximation. So far we have discussed only part (a) above; we will know focus on (b).

Suppose that we have at our disposal a consistent[2] estimator of the variance $Var_F(T)$; let us call this estimator $\widehat{Var}_F(T)$. To fix ideas, consider the simplest case where the statistic $T(\mathbf{X})$ of interest is the sample mean $\bar{X}$. In this case, there is available a simple consistent estimator of $Var_F(T)$, namely $\widehat{Var}_F(T) = s^2/N$, where $s^2 = (N-1)^{-1}\sum_{k=1}^{N}(X_k - \bar{X})^2$ is the sample variance. Dividing the statistic $T(\mathbf{X})$ by its estimated standard deviation $\sqrt{\widehat{Var}_F(T)}$ is usually referred to as 'studentization', since –*if* the data were Gaussian– this would result in Student's t-distribution. Consider then the sampling distribution of the 'studentized' statistic, i.e.,

$$Dist_{student}(x) = P_F(\frac{T(\mathbf{X}) - \theta(F)}{\sqrt{\widehat{Var}_F(T)}} \le x). \tag{21}$$

Knowledge of $Dist_{student}(x)$ for all real x would yield a $(1-\alpha)100\%$ confidence interval for $\theta(F)$ in the form

$$[T(\mathbf{X}) - u(1-\alpha/2)\sqrt{\widehat{Var}_F(T)}, T(\mathbf{X}) - u(\alpha/2)\sqrt{\widehat{Var}_F(T)}], \tag{22}$$

where $u(\alpha/2)$ and $u(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist_{student}(x)$ distribution respectively.

Note however that for general statistics, or even for the sample mean if we are not willing to assume that data are Gaussian, the distribution $Dist_{student}(x)$ and its quantiles are unknown; nevertheless, it can be estimated by the bootstrap, similarly to what was discussed in the previous sections. In particular, the bootstrap distribution $Dist^*_{student}(x)$ that can be used to approximate $Dist_{student}(x)$ is given by

$$Dist^*_{student}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(\#T(\mathbf{X}^{*(i)}) \le x\sqrt{\widehat{Var}_F{}^{*(i)}(T)} + \theta(\hat{F}))$$

[2] Loosely speaking, an estimator is consistent if it is accurate for large samples, i.e., *asymptotically* correct.

$$= \frac{1}{B}(\#T(\mathbf{X}^{*(i)}) \leq x\sqrt{\widehat{Var_F}^{*(i)}(T)} + \theta(\hat{F})), \tag{23}$$

where $\widehat{Var_F}^{*(i)}(T)$ is the estimate of the variance of the statistic $T(\mathbf{X})$ *as computed from the* $\mathbf{X}^{*(i)}$ *resample.* For example, in the sample mean case, $\widehat{Var_F}^{*(i)}(T) = (N-1)^{-1}\sum_{k=1}^{N}(X_k^{*(i)} - \bar{X}^{*(i)})^2$, where $\bar{X}^{*(i)} = N^{-1}\sum_{k=1}^{N} X_k^{*(i)}$.

Note that, if a variance estimate is not readily available, $\widehat{Var_F}(T)$ itself could be a bootstrap estimate constructed as in section 1.3; in that case, $\widehat{Var_F}^{*(i)}(T)$ is the bootstrap variance estimate computed from the $\mathbf{X}^{*(i)}$ resample! In other words, we have an *iterated* or *nested* bootstrap –a bootstrap simulation for each of the original bootstrap resamples; cf. Hall (1992) or Efron and Tibshirani (1993) for more details.

In any case, an equal-tailed $(1-\alpha)100\%$ bootstrap confidence interval for $\theta(F)$

$$[T(\mathbf{X}) - u^*(1-\alpha/2)\sqrt{\widehat{Var_F}(T)}, T(\mathbf{X}) - u^*(\alpha/2)\sqrt{\widehat{Var_F}(T)}], \tag{24}$$

where $u^*(\alpha/2)$ and $u^*(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap $Dist^*_{student}(x)$ distribution respectively; the confidence interval of equation (24) is called a bootstrap-t or a percentile-t interval due to the 'studentization'.

As it turns out (cf. Singh (1981)), the confidence interval of equation (24) is *more* accurate than *either* the pivotal bootstrap interval of equation (18), *or* the normal confidence interval of equation (3); this is what is meant by 'higher order accuracy of the bootstrap', or that the bootstrap 'captures the skewness' of the underlying distribution. This higher order accuracy comes at a price, since the iterated bootstrap is much more computer intensive than the simple bootstrap; however, in the sample mean case the extra computational burden is minuscule, because a variance estimate can be computed without Monte Carlo simulation.

1.9 Trasformations and variance stabilization. The reader should also refer to the textbook by Efron and Tibshirani (1993) for a different construction of higher order accurate bootstrap confidence intervals, the BC_a intervals, that are based on the idea of 'bias correction'. It is quite interesting to note that the BC_a intervals have the additional desirable property of being 'transformation invariant', a property not shared by the bootstrap confidence intervals of equations (18) and (24), nor by the normal approximation interval of equation (3).

To explain the property of 'transformation invariance', consider a (strictly) monotone function $g(\cdot)$, and its inverse $g^{-1}(\cdot)$. Since $T = T(\mathbf{X})$ is considered to be a good estimator of $\theta = \theta(F)$, then it follows that $g(T)$ is a good estimator of $g(\theta)$. Suppose $[l, u]$ is an equal-tailed $(1-\alpha)100\%$ approximate confidence interval for $\theta(F)$ constructed using any of the available methods, i.e., normal theory of equation (3), pivotal bootstrap of equation (18), bootstrap-t of equation (24), or bootstrap BC_a.

Observe that $g(T)$ is just a statistic based on our sample, and it can be 'bootstrapped' as well. In other words, the sampling distribution of $g(T)$ can be estimated, and an equal-tailed $(1-\alpha)100\%$ confidence interval for $g(\theta)$ can be formed, by the same method used to obtain the interval for $\theta(F)$; say this interval is $[g_l, g_u]$. It then follows that $[g^{-1}(g_l), g^{-1}(g_u)]$ is an approximate $(1-\alpha)100\%$ confidence interval for $\theta(F)$, and this new confidence interval should be compared to the interval $[l, u]$ found directly. If the two intervals for $\theta(F)$ are identical, then the property of 'transformation invariance' holds; if not, it makes sense to ask "which of the two intervals is better?", in which case one is led to search for an 'optimal' transformation $g(\cdot)$ to use in connection with the construction of confidence intervals.

In some isolated cases, e.g., Fisher's hyperbolic tangent transformation for the correlation coefficient (cf. Efron and Tibshirani (1993, p. 54 and p. 163)), a transformation is available in the literature that approximately 'normalizes' and 'variance stabilizes' the estimator $T(\mathbf{X})$; in other words, the estimator $g(T)$ has a distribution that is closer to being Gaussian than the distribution of $T(\mathbf{X})$, and the variance of $g(T)$ does not depend on the parameter $\theta(F)$, at least not significantly. As a consequence, such a transformation is 'optimal' to use in connection with the construction of confidence intervals based on the normal approximation of equation (3).

In most cases however, it may not be possible to simultaneously normalize and variance stabilize the estimator $T(\mathbf{X})$ by a single transformation. As it turns out, the 'optimal' transformation associated with constructing bootstrap-t confidence intervals should primarily achieve variance stabilization. Now if $Var_F(T)$ were known as a function of $\theta(F)$, then an approximate variance stabilizing transformation $g(\cdot)$ could be found by the δ-method (cf. Miller (1986), Efron and Tibshirani (1993)). The problem of course is that $Var_F(T)$, $\theta(F)$, as well as the functional relationship between the two are generally unknown!

Nonetheless, an approximate 'optimal' transformation for variance stabilization can be computed using an iterated bootstrap –much like the iterated bootstrap described in the previous section on studentization– to calculate estimates of $Var_F(T)$ from each resample; details can be found in Efron and Tibshirani (1993, p. 163). It should be noted that if an iterated bootstrap is carried out to calculate the variance stabilizing transformation, then there is no need to do another iterated bootstrap to get the bootstrap-t confidence interval. In other words, there is no need for the studentization any more since the variance can be considered constant, and a bootstrap confidence interval for $g(\theta)$ based on the pivotal method of equation (18) would be obtained and then inverted (using g^{-1}) to give a good bootstrap confidence interval for $\theta(F)$.

2. Subsampling and the jackknife

While one reason for the success of the bootstrap is its widespread applicability, there are certainly situations where the bootstrap is *not* applicable; for example, in the case where the statistic $T(\mathbf{X})$ is linear, i.e., of the sample mean type, the validity of the bootstrap crucially hinges on whether the statistic is asymptotically normal or not. As a matter of fact, a huge statistical literature on the bootstrap has accumulated since Efron's (1979) pioneering paper, with main focus to show the applicability of the bootstrap in many different settings; see our section 3 for some bibliographical comments.

At another level, recall that performing the bootstrap in practice requires sampling *with* replacement from the observations $X_1, \ldots, X_N$, to get a resample of size N. The *exact* computation of the bootstrap distribution would actually involve taking into account *all* the possible resamples, weighted by the corresponding multinomial probabilities; however, the number of possible resamples is $\frac{(2N-1)!}{N!(N-1)!}$ which is impractically large. Doing the Monte Carlo random bootstrap sampling gets around this problem, but there is also another way of lowering the computational complexity: the jackknife and subsampling.

2.1 The jackknife idea. Consider sampling *without* replacement from the observations $X_1, \ldots, X_N$, to get a resample (now called a *subsample*) of size b, where of course $b < N$. If $b = N - 1$, this is exactly the original jackknife of Quenouille and Tukey (cf. Efron (1979, 1982) and Efron and Tibshirani (1993) for details), and there are only N possible different subsamples. Since these subsamples are all equally probable under the sampling without replacement scheme, formulas much like (15), (16), (19), and (20) can be constructed to estimate bias, variance, and distribution of the statistic $T(\mathbf{X})$; these will be given in a more general form in what follows.

In general, one can take an arbitrary b, not necessarily equal to $N-1$, yielding the so-called delete-d jackknife, where $d = N - b$; the number of possible subsamples now rises to $\frac{N!}{b!(N-b)!}$ and again a Monte Carlo method can be employed to randomly chose a smaller number, say B, among these subsamples to be included in the jackknife. As long as b is large enough (but of smaller order of magnitude than N) the subsampling distribution estimates are asymptotically correct.

In some sense, subsampling can be thought to be even more intuitive than the bootstrap, because the subsamples are actually samples (of smaller size) from the *true* distribution F, whereas the bootstrap resamples are samples from an estimator of F. As can be shown (cf. Politis and Romano (1992), distribution estimates based on subsampling are valid in a wider range of situations than their resampling (i.e., bootstrap) analogs, even in cases where the statistic $T(\mathbf{X})$ is not asymptotically normal; however, they do not possess the property of higher order accuracy, and this is essentially due to the fact that the subsampling size is b and not N.

This difference between the subsample size and the original sample size has an additional consequence, namely that a re-scaling is in order in computing the subsampling distribution estimator. Suppose that the variance of $T(\mathbf{X})$ is approximately proportional to c^2/τ_N^2, for large N, where c is some constant; in regular cases, $\tau_N^2 = N$. It follows that the variance of T calculated from a sample of size b is approximately proportional to c^2/τ_b^2; here the need for a re-scaling becomes apparent. The subsampling procedure can finally be summarized as follows:

- Choose B subsamples $\mathbf{X}^{\star(1)}, \ldots, \mathbf{X}^{\star(B)}$ among all the possible subsamples of size b of the sample population $\{X_1, \ldots, X_N\}$. Suppose the ith subsample is $\mathbf{X}^{\star(i)} = (X_1^{\star(i)}, \ldots, X_b^{\star(i)})$; the final step now is to evaluate the statistic T over each of the chosen subsamples, creating the pseudo-replications $T(\mathbf{X}^{\star(1)}), \ldots, T(\mathbf{X}^{\star(B)})$.

2.2 Confidence intervals based on subsampling. The subsampling estimates of $Bias_F(T)$, $Var_F(T)$, $Dist_{T,F}(x)$, and $Dist_{T,F,\theta}(x)$ are $Bias^{\star}(T)$, $Var^{\star}(T)$, $Dist^{\star}_{T,F}(x)$, and $Dist^{\star}_{T,F,\theta}(x)$ respectively which are presented below; note that if $B = \frac{N!}{b!(N-b)!}$ and Monte Carlo randomization is not used, i.e., *all* possible subsamples are taken into account, the approximation signs ($\simeq$) below can be replaced by equality signs.

$$Bias^{\star}(T) \simeq \frac{\tau_b}{\tau_N}\left(\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{\star(i)}) - T(\mathbf{X})\right) \tag{25}$$

$$Var^{\star}(T) \simeq \frac{\tau_b^2}{\tau_N^2}\left(\frac{1}{B}\sum_{i=1}^{B} T^2(\mathbf{X}^{\star(i)}) - [\frac{1}{B}\sum_{i=1}^{B} T(\mathbf{X}^{\star(i)})]^2\right) \tag{26}$$

$$Dist^{\star}_{T,F}(x) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(T(\mathbf{X}^{\star(i)}) \le x\frac{\tau_N}{\tau_b}) = \frac{1}{B}(\#T(\mathbf{X}^{\star(i)}) \le x\frac{\tau_N}{\tau_b}) \tag{27}$$

and

$$Dist^{\star}_{T,F,\theta}(x) \simeq \frac{1}{B}(\#T(\mathbf{X}^{\star(i)}) \le x\frac{\tau_N}{\tau_b} + T(\mathbf{X})). \tag{28}$$

Similarly to the interval (18), an equal-tailed $(1-\alpha)100\%$ confidence interval for $\theta(F)$ based on subsampling would be

$$[T(\mathbf{X}) - q^{\star}(1-\alpha/2), T(\mathbf{X}) - q^{\star}(\alpha/2)], \tag{29}$$

where $q^{\star}(\alpha/2)$ and $q^{\star}(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the $Dist^{\star}_{T,F,\theta}(x)$ distribution respectively.

As a final remark, it is worth noting that if the B subsamples that are used to construct the subsampling estimates are chosen (without Monte Carlo randomization) to be the $N-b+1$ subsamples characterized by the property that each contains b *consecutive* observations from the original sample $X_1, \ldots, X_N$, then the subsampling estimates given above are valid *even if the sample exhibits serial correlation*; see Politis and Romano (1992) for more details regarding subsampling stationary time series.

3. Some bibliographical comments

At this moment, there are four published books on the bootstrap: the original monograph of Efron (1982), the textbook by Hall (1992) that contains a lot of material concerning the higher order accuracy of the bootstrap and the effects of 'studentization', the collection of research papers in LePage and Billard (1992), and the new textbook by Efron and Tibshirani (1993). There are also two collections of lecture notes: Beran and Ducharme (1991) provide theoretical expositions of the concept of 'prepivoting', a method related to 'studentization', and of bootstrap balanced confidence intervals and prediction regions, and Mammen (1992) focuses mainly on the bootstrap for linear models.

Several review articles are now available in the literature: Efron and Gong (1983) and Efron and Tibshirani (1986) have a more applied flavor, whereas DiCiccio and Romano (1988) give a theoretical treatment. Swanepoel (1990), and Léger, Politis, and Romano (1992) review more recent developments and provide discussion on more advanced applications of the bootstrap methodology; both papers also contain an extensive list of references. Léger *et al.* (1992) and Bose and Politis (1993) provide reviews of the bootstrap for dependent samples. Finally, the reference for most of our section on subsampling is Politis and Romano (1992) that also contains a good number of examples where the bootstrap does *not* work.

Acknowledgement

The present report was written as a part of a tutorial set of notes for students in the Purdue University course STAT526 'Advanced Statistical Methodology'. Many thanks are due to Professor Mary Ellen Bock of the Department of Statistics, Purdue University, for her critical reading and helpful comments.

References

[1] Beran, R. and Ducharme, G.R. (1991), *Asymptotic Theory for Bootstrap Methods in Statistics*, Les Publications CRM, Montreal.

[2] Bose, A. and Politis, D.N. (1993), A review of the bootstrap for dependent samples, Technical Report No. 93-4, Department of Statistics, Purdue University.

[3] DiCiccio, T., and Romano, J. (1988), A review of bootstrap confidence intervals (with discussion), *J. Roy. Statist. Soc., Ser. B*, vol. 50, 338-370.

[4] Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *Ann. Statist.*, 7, 1-26.

[5] Efron, B. (1982),*The Jackknife, the Bootstrap, and other Resampling Plans*, SIAM NSF-CBMS, Monograph 38.

[6] Efron, B., and Gong, G. (1983), A leisurely look at the Bootstrap, the Jackknife, and Cross-Validation, *Amer. Statistician*, vol. 37, No. 1, pp. 36-48.

[7] Efron, B. and Tibshirani, R.J. (1986), Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Statist. Sci.* **1**, 54-77.

[8] Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[9] Hall, P. (1988), Theoretical Comparison of Bootstrap Confidence Intervals, *Ann. Statist.*, 16, 927-953.

[10] Hall, P.(1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag.

[11] Léger, C., Politis, D.N., and Romano, J.P. (1992), Bootstrap Technology and Applications, *Technometrics*, vol. 34, pp. 378-399 .

[12] Lehmann, E.L. (1983), *Theory of point estimation*, John Wiley.

[13] LePage, R. and Billard, L. (eds.) (1992), *Exploring the Limits of Bootstrap*, John Wiley.

[14] Mammen, E. (1992), *When does bootstrap work? asymptotic results and simulations*, Lecture notes in Statistics # 77, Springer, New York.

[15] Miller, R. (1986), *Beyond ANOVA: Basics of Applied Statistics*, John Wiley.

[16] Politis, D.N., and Romano, J.P. (1992), Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, Technical Report No. 92-36, Department of Statistics, Purdue University; also submitted to *Ann.Statist.*.

[17] Singh, K.(1981), On the asymptotic accuracy of Efron's bootstrap, *Ann.Statist.*, 9, 1187-1195.

[18] Swanepoel, J.W.H. (1990), A review of bootstrap methods, *South African Statist. J.*, vol. 24, pp. 1-34.