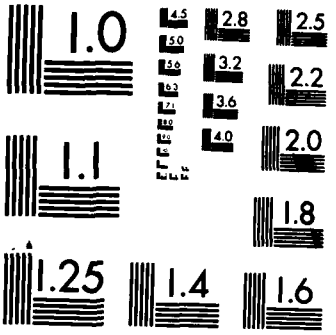
AD-A160 043 HOW MANY BOOTSTRAPS?(U) STANFORD UNIV CA DEPT OF STATISTICS R TIBSHIRANI 22 AUG 85 TR-362 1/1
N00014-76-C-0475

UNCLASSIFIED F/G 12/1 NL



HOW MANY BOOTSTRAPS?

BY

ROBERT TIBSHIRANI

TECHNICAL REPORT NO. 362
AUGUST 22, 1985

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

How Many Bootstraps?

Robert Tibshirani

1. Introduction.

The bootstrap (Efron 1979) is a non-parametric method for assessing statistical accuracy. Consider for example estimation of the variance of a statistic T under a true distribution F, i.e. $VAR_F T$. The bootstrap works by replacing the unknown distribution F by the empirical distribution function $\hat{F}$ and the bootstrap estimate of $VAR_F T$ is defined as $VAR_{\hat{F}} T$ Unless T is very simple, this can't be computed analytically, and hence must be approximated by a monte carlo simulation. To do this, we sample N times with replacement from the original data (N is the sample size), then evaluate the statistic of interest for this "bootstrap sample". This process is repeated B times, where B is typically 100 to 1000. The monte carlo estimate of $VAR_{\hat{F}} T$ is the sample variance of the B bootstrap values of T. (Note that sampling with replacement from the data is equivalent to sampling from $\hat{F}$).

If we could take $B = \infty$, our monte carlo estimate would exactly equal $VAR_{\hat{F}} T$. Because of computational costs, we might not be able to take B much larger than 1000; for complicated statistics, 100 might be our limit. In this paper we study the question of how many bootstraps to take. We provide an adaptive method, based on the bootstrap samples already taken, for assessing how many more bootstrap samples (if any) we need. Our experience shows that for estimating variance or bias, B=100 to 300 is usually adequate. An estimate of a percentile, however, may require B = 1000.

The formulae used here are closely related to those derived in

Efron (1984), Section 6 and are based on standard results that can be found in many books (e.g. Kendall and Stuart (1958), chapter 10).

2. The Bootstrap Method and a Statement of the Problem.

Suppose that we have a sample $\mathbf{X}=(X_1,X_2,\ldots X_N)$, with X_i assumed to be i.i.d from a distribution F. (X_i may be real or vector valued). Our statistic of interest is some symmetric function $T(X_1,X_2,\ldots X_N)$. We require an estimate of a functional $Q(T,F,\mathbf{X})$. Denoting the empirical distribution function by $\hat{F}$, the bootstrap estimate is defined as $Q(T,\hat{F},\mathbf{X})$. Usually, we can't compute this analytically so we estimate it through a monte carlo simulation. This is done by a) writing $Q(T,\hat{F},\mathbf{X})$ in terms of quantities of the form $E_{\hat{F}}R$, then b) estimating each quantity by a monte carlo estimate of expectation. In the case of $Q(T,F,\mathbf{X}) = VAR_F T$, for example, we write $VAR_{\hat{F}}T=E_{\hat{F}}T^2 - (E_{\hat{F}}T)^2$. We draw B bootstrap samples (that is, samples of size N drawn with replacement from $X_1, X_2,\ldots X_N$.) and compute the bootstrap values $T_1^*, T_2^*,\ldots T_B^*$. Our monte carlo estimate of $VAR_F T$ is then $\Sigma T_i^{*2}/B - [\Sigma T_i^*/B]^2$.

If $Q(T,F,\mathbf{X}) = Prob_F(T > c)$, we write $Prob_{\hat{F}}(T > c) = E_{\hat{F}}I(T > c)$ and our monte carlo estimate is $\#\{T_i^* > c\}/B$.

Now let $Q_B(T,\hat{F},\mathbf{X})$ be an approximation to $Q(T,\hat{F},\mathbf{X})$ based on B bootstrap samples. As $B \to \infty$ we have, for sufficiently well-behaved Q, $Q_B(T,\hat{F},\mathbf{X}) \to Q(T,\hat{F},\mathbf{X})$. The question we address here is: how big should B be so that $Q_B(T,\hat{F},\mathbf{X})$ is (on the average) sufficiently close to $Q(T,\hat{F},\mathbf{X})$?

The approach we will take is the following. For a specific Q, we choose a measure of Q_B's accuracy in estimating Q . (This measure will be conditional on the observed data). Then we take a small number of bootstrap samples (say 50 or 100), and estimate the accuracy of $Q_B(T,\hat{F},\mathbf{X})$. If $Q_B(T,\hat{F},\mathbf{X})$ is not accurate enough, we take more bootstrap samples until the estimated accuracy is sufficiently high.

In the following sections we illustrate this for three specific Q's: standard error, percentile, and bias. In the final section we discuss a

number of points including Efron's (1984) approach to this problem.

3. Number of bootstraps for standard error estimation.

Here we take $Q=\sigma\equiv(Var_F(T))^{1/2}$. Let $\mathbf{X}_1^*$, $\mathbf{X}_2^*$... $\mathbf{X}_B^*$ denote bootstrap samples generated from $\hat{F}$. Then $Q_B=\sigma_B$, the sample standard deviation of the $\mathbf{X}_i$'s. A reasonable measure of the accuracy of σ_B is its coefficient of variation conditional on $\mathbf{X}$. Standard calculations show

$$CV(\sigma_B \mid \mathbf{X}) = [(\delta+2)/4B]^{1/2} + O(1/B) \qquad (1)$$

where δ is the kurtosis of the bootstrap distribution of T. Table 1 shows $CV(\sigma_B \mid \mathbf{X})$ for $\delta=0$ and $\delta=3$ (the kurtosis of the t distribution on 6 degrees of freedom).

Table 1.

$CV(\sigma_B \mid \mathbf{X})$ as a function of B and δ

B	10	20	50	100	200	500	1000
$\delta=0$	.22	.14	.10	.07	.05	.03	.02
$\delta=3$	.35	.25	.16	.11	.08	.05	.04

We propose estimation of $CV(\sigma_B \mid \mathbf{X})$ from the bootstrap distribution as a guideline for the choice of B. The suggested procedure is to take say 50 bootstrap samples, estimate $CV(\sigma_B \mid \mathbf{X})$ from the bootstrap distribution, take more samples etc. until the estimated coefficient of variation is small enough. Note that estimation of $CV(\sigma_B \mid \mathbf{X})$ requires an estimate of δ. For this we use the sample kurtosis of the bootstrap values. For

non-robust T, it would be preferable to use a more robust measure.

What is a "small" coefficient of variation? One way to determine this is to examine the effect of an error in ϵ_B on the coverage of a confidence interval of the form $[T(\mathbf{X})+\epsilon_B z^{(\alpha)},\ T(\mathbf{X})+\epsilon_B z^{(1-\alpha)}]$ where $z^{(\alpha)}$ is the 100α-th percentage point of the standard normal distribution. Let θ be the parameter of F that T estimates and let $g(\epsilon_B)=P(\theta \in [T(\mathbf{X})+\epsilon_B z^{(\alpha)}, T(\mathbf{X})+\epsilon_B z^{(1-\alpha)}])$, the coverage of the standard interval. A taylor series argument gives for the (conditional) standard deviation of $g(\epsilon_B)$

$$SD(g(\epsilon_B) \mid \mathbf{X}) \approx 2\,\varphi(z^{(1-\alpha)})z^{(1-\alpha)}\ CV(\epsilon_B \mid \mathbf{X}) \qquad (2)$$

This relationship is illustrated in Table 2.

Table 2.

$SD(g(\epsilon_B) \mid \mathbf{X})$ and $CV(\epsilon_B \mid \mathbf{X})$
for α=.025 and .05

	CV	
SD	α=.025	α=.05
.01	.04	.03
.02	.09	.06
.05	.22	.15

Thus if we aim for a 90% confidence interval and we're willing to allow a standard error of 2%, we require $CV(\epsilon_B \mid \mathbf{X})$ to be about .06. For a 95% interval, we require $CV(\epsilon_B \mid \mathbf{X})$ to be .09. In the examples that follow, we will use .06 as our target, although this is of course up to the statistician to choose in any particular problem.

Example 1. The correlation coefficient.

The data for this example come from the SAS Basics manual (1982) page 510. They consist of 50 measurements of chest and abdomen skinfold thickness. The statistic we chose was the sample correlation coefficient which had a value of .620 for the original data. Table 3 shows the results of successively increasing the number of bootstraps.

Table 3
Results for the correlation coefficient
applied to the skinfold data

B	$\hat{\delta}$	σ_B	CV
50	-.11	.090	.065
100	-.44	.097	.062
200	.13	.097	.052
500	-.20	.100	.030
1000	.25	.106	.024

We see that with 200 bootstraps the CV is below .06, and even 50 may be adequate.

Example 2. Cox's model.

In this example we bootstrapped Cox's partial likelihood estimate for the proportional hazards model. The data consisted of 200 measurements on mice taken from Kalbfleisch and Prentice (1981) pg 233. The outcome was survival in days, the covariate was % antibody level. The

bootstrapping was performed by treating the response, covariate and censoring indicator for each mouse as the sampling unit. The partial likelihood estimate for the original data was -.015. The bootstrap results are shown in Table 4.

Table 4.
Results for Cox's estimator
applied to mouse leukemia data.

B	δ	ϵ_B	CV
50	3.02	.010	.16
100	2.04	.009	.10
200	1.75	.008	.07
300	2.00	.008	.06
500	1.86	.008	.04
1000	3.00	.009	.04

About 300 bootstraps are necessary to get CV down to about .06, despite the fact that the value of ϵ_B changes very little as B increases.

4. Number of bootstraps for percentile points.

For the problem of estimating percentile points, the functional $Q(T,F,\mathbf{X})$ is $G^{-1}(\alpha)$ where G is the distribution of T under F. The bootstrap estimate of $G^{-1}(\alpha)$ is $\hat{G}^{-1}(\alpha)$ where $\hat{G}$ is the bootstrap distribution of $T(\mathbf{X}^*)$, that is the distribution of $T(\mathbf{X}^*)$ under $\hat{F}$. Letting $\hat{G}_B$ be the empirical distribution function of $T_1^*,\ldots T_B^*$, the monte carlo approximation to $\hat{G}^{-1}(\alpha)$ is $\hat{G}_B^{-1}(\alpha)$. (If $G^{-1}(\alpha)$ is not uniquely defined, we will assume some reasonable definition like $\inf\{t: G(t)> \alpha\}$ and similarly for $\hat{G}^{-1}(\alpha)$ and $\hat{G}_B^{-1}(\alpha)$). Standard calculations give

$$CV[\hat{G}_B^{-1}(\alpha)] = \frac{[\alpha(1-\alpha)]^{\frac{1}{2}}}{[B^{\frac{1}{2}}\hat{G}^{-1}(\alpha)\,\hat{g}(\hat{G}^{-1}(\alpha))]} + O(1/B) \qquad (3)$$

In the above $\hat{g}(.)$ is $d\hat{G}(t)/dt$. Now if $\hat{G}$ is normal, then $\hat{g}(\hat{G}^{-1}(\alpha)) = |z^{(\alpha)}/\hat{G}^{-1}(\alpha)\ (\psi(z^{(\alpha)})$ and hence $CV[\hat{G}_B^{-1}(\alpha)] = [\alpha(1-\alpha)]^{\frac{1}{2}}/[B^{\frac{1}{2}}|z^{(\alpha)}|\psi(\ z^{(\alpha)})]$. For this case, Table 5 shows a tabulation of $CV[\hat{G}_B^{-1}(\alpha)]$.

Table 5

$CV\ [\hat{G}_B^{-1}(\alpha)]$ for $\hat{G}$ normal

	α	.75	.90	.95	.975
B					
50		.29	.19	.18	.19
100		.20	.13	.13	.14
200		.14	.09	.09	.10
500		.09	.06	.06	.06
1000		.06	.04	.04	.04

As we did in the previous section, we must address the question: What is a "small" coefficient of variation for this problem? Let $\hat{\alpha}=P_{\hat{F}}(T_i^* < \hat{G}_B^{-1}(\alpha))$. We can measure the size of $CV\ [\hat{G}_B^{-1}(\alpha)]$ by assessing its effect on $SD(\hat{\alpha})$. If G is approximately normal, it is easy to show that

$$SD(\hat{\alpha}) \approx z^{(\alpha)}|\psi(z^{(\alpha)})\ CV\ [\hat{G}_B^{-1}(\alpha)] \qquad (4)$$

Table 6 tabulates this for various values of α and $SD(\hat{\alpha})$.

Table 6.

$CV[\hat{G}_B^{-1}(\alpha)]$ as a function of $\hat{\alpha}$ and $SD(\hat{\alpha})$

$SD(\hat{\alpha})$ \ α	.75	.90	.95	.975
.005	.02	.02	.03	.04
.01	.05	.04	.06	.09
.02	.09	.09	.12	.17
.05	.23	.22	.29	.44

Thus for example if we want to estimate the 90th percentile with a standard error of .01 in the coverage, $CV[\hat{G}_B^{-1}(\alpha)]$ has to be less than or equal to .04. Assuming again that G is normal, Table 5 then tells us that B should be at least 1000.

Instead of assuming normality, we can take the approach discussed in section 2: estimate $\hat{G}^{-1}(\alpha)$ and $\hat{g}(\hat{G}^{-1}(\alpha))$ from the bootstrap distribution and thus get an estimate of $CV[\hat{G}_B^{-1}(\alpha)]$. We tried this in the next two examples, using a kernel density estimate of the form $\Sigma(1/h\epsilon_B)\psi((T_i^*-y)/h\epsilon_B)$. A value of .5 was used for the window parameter h; fortunately the results changed very little when h was varied.

Example 3. Percentile points for the setup of example 1.

We applied the bootstrap to estimate a percentile of the bootstrap distribution for the correlation in the skinfold data estimate. Table 7 shows the results for α=.10 and .05.

Table 7.

Percentile results for skinfold data

		$\alpha=.025$		$\alpha=.10$	
		$CV[\hat{G}_B^{-1}(\alpha)]$	$\hat{G}_B^{-1}(\alpha)$	$CV[\hat{G}_B^{-1}(\alpha)]$	$\hat{G}_B^{-1}(\alpha)$
	100	.014	.752	.013	.711
B	200	.011	.761	.011	.725
	500	.006	.772	.006	.732
	1000	.005	.786	.005	.743

We see that 100 bootstraps is adequate both for $\alpha=.10$ and $\alpha=.025$.

Example 4. Percentile points for Example 2 (Cox model).

The results for the Cox model applied to the mouse leukemia data are shown in Table 8.

Table 8.

Percentile results for Cox model

		$\alpha=.025$		$\alpha=.10$	
		$CV[\hat{G}_B^{-1}(\alpha)]$	$\hat{G}_B^{-1}(\alpha)$	$CV[\hat{G}_B^{-1}(\alpha)]$	$\hat{G}_B^{-1}(\alpha)$
	100	.250	-.0046	.095	-.010
B	200	.151	-.0047	.090	-.008
	500	.098	-.0047	.054	-.008
	1000	.062	-.0044	.040	-.007

In order to get CV down to .04, 1000 bootstraps are needed for $\alpha=.10$; more than 1000 bootstraps are required for $\alpha=.025$. In the second case, $\hat{G}^{-1}(\alpha)$ isn't changing much but the estimates still have large variability.

5. Bias estimation.

The mean bias of the estimator T is defined as $E_F T-\theta$, and is estimated by the bootstrap quantity $E_{\hat{F}} T(X_1^*, X_2^* \ldots X_N^*)-T(X_1, X_2, \ldots X_N)$. The monte carlo approximation to the bootstrap estimate is $\Sigma T_i^*/B - T(X_1, X_2, \ldots X_N)$. Assuming $E_{\hat{F}} T(X_1^*, X_2^* \ldots X_N^*) = T(X_1, X_2, \ldots X_N)$, this has coefficent of variation $\epsilon / B^{\frac{1}{2}} |T(X_1, X_2, \ldots X_N)|$. Using ϵ_B as an estimate of ϵ, Table 9 shows the CV as B increases for the setup of Example 1.

Table 9.
CV for estimate of bias
(correlation coefficient, example 1)

B	CV
50	.027
100	.016
200	.011
500	.007
1000	.005

There seems to be no natural way to decide what a "small" CV for bias is; one might arbitrarily decide that .05 is small. In that case, as few as 50 bootstraps is satisfactory in this example. For the Cox model example, about 200 bootstraps are necessary to get the CV down below .05.

Median bias is defined as $G^{-1}(\frac{1}{2})-\theta$ and is estimated by $\hat{G}^{-1}(\frac{1}{2})-T(X_1, X_2 \ldots X_N)$. The monte carlo approximation $\hat{G}_B^{-1}(\frac{1}{2})-T(X_1, X_2 \ldots X_N)$

has coefficient of variation that can be estimated by the method described in section 3.

A related problem is the estimation of $\alpha=P(T > t)$, where α is near $\frac{1}{2}$. The bootstrap estimate is $\hat{\alpha} = P_{\hat{F}}(T_i^* > t)$ and the monte carlo approximation is $\hat{\alpha}_B=\#(T_i^* > t)/B$. The standard deviation of $\hat{\alpha}_B$ is $[\hat{\alpha}(1-\hat{\alpha})/B]^{\frac{1}{2}}$, which equals $1/2B^{\frac{1}{2}}$ when $\hat{\alpha}=\frac{1}{2}$. In order for the standard deviation to get down to .02, B must be 625; a standard deviation of .01 requires B=2500. Efron (1984) notes this fact in the context of estimation of the bias-corrected percentile interval and suggests a better method of approximation.

6. Discussion.

We have presented here an adaptive method for determining the number of bootstraps necessary to achieve a pre-specified accuracy. These techniques should be used with some caution. In the first two problems considered, estimates of kurtosis and the density of the bootstrap distribution were required. The estimation of both these quantities is quite delicate and shouldn't be attempted for bootstrap sample sizes less than 50 and 100, respectively. It goes without saying that if you can feasibly take 1000 or more bootstraps, then take them! One reason is that often attention is not focussed on a single aspect like variance; for example the general shape of the bootstrap distribution may be of interest. For this it is hard to quantify how large B should be and the bigger the better. Some further remarks:

Remark A. *An unconditional view of the problem.* Efron (1984) computes the unconditional coefficient of variation of ϵ_B and $\hat{G}^{-1}(\alpha)$ as a way of determining how large B should be. He takes the point of view that B is large enough if the conditional variation in ϵ_B is small compared to its unconditional variation, that is, if increasing B doesn't decrease the unconditional CV substantially. For standard error estimation he assumes

that $E(\delta)=0$ and shows that B as small as 25 or 50 can be adequate. For percentile estimation, his results, based on normal approximations, indicate that B=1000 is necessary. In this paper, we have taken a slightly differerent view--- that even if the sampling variability of the estimate is large, one might still want an accurate measure of its standard error. Also the adaptive approach allows one to take less or more bootstraps depending on the variability of the bootstrap values observed.

Remark B. *Variance reduction techniques*. The use of monte carlo variance reduction techniques can greatly reduce the number of bootstraps necessary to achieve a given accuracy. Therneau (1983) looked at a number of methods, most notably control variates, for variance and bias estimation. Johns (1984 personal communication) has had success with importance sampling for percentile estimation.

Remark C. *Relationship between sample size and number of bootstraps*. If $T = \bar{X}$, it is straightforward to show that the unconditional CV of ϵ_B is of the form $[a/N + (b/NB) + c/B]^{\frac{1}{2}}$, where a, b and c are constants depending on F (the distribution of X_i). Hence as N increases, $CV(\epsilon_B)$ goes down at the rate $1/N^{\frac{1}{2}}$. This makes senses intuitively: as the sample size increases, the kurtosis of the bootstrap distribution of $\bar{X}^*$ decreases. By linearizing a non-linear statistic, one could presumably show that this is approximately true in general. This doesn't mean, however, that less computations are necessary since each evaluation of the statistic will typically be at least O(N).

Remark D. *The parametric and smooth bootstraps*. The techniques discussed in this paper did not make any special use of the fact that F was estimated by the empirical distribution function $\hat{F}$. Hence they are still appropriate if some other estimate of F is used, for example a parametric estimate ("the parametric bootstrap") or a semi-parametric estimate ("the smooth bootstrap").

Acknowledgement: this research was supported by a Ontario Ministry of Health fellowship.

REFERENCES

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of statistics 7, 1-26.

Efron, B. (1984). Better bootstrap confidence intervals. Stanford University technical report LCS 14.

Kalbfleisch, J. and Prentice, R. (1980). The statistical analysis of failure time data. Wiley, New York.

Kendall, M. and Stuart, A. (1958). The advanced theory of statistics. Griffen, London.

SAS User's Guide: Basics manual (1982). SAS Institute Incorporated, Cary, North Caroline.

Therneau, T. (1983). Variance reduction techniques for the bootstrap. Stanford University technical report 200.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 362	2. GOVT ACCESSION NO. AD-A160043	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) How Many Bootstraps?		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Robert Tibshirani		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 411SP		12. REPORT DATE August 22, 1985
		13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bootstrap, monte carlo approximation

This document

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In approximating bootstrap quantities by monte carlo simulation, one must decide how many bootstrap samples to generate. We propose an adaptive sequential method that estimates the accuracy based on the current bootstrap samples. Bootstrap sampling is continued until the estimated accuracy is high enough. In the examples given, 100 to 300 bootstraps are sufficient for standard error and bias estimation, while 1000 bootstraps may be necessary for estimating a percentile.

Additional keywords: tables (data).

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

11-85

DTIC