The Annals of Statistics
1981, Vol. 9, No. 1, 130–134

THE BAYESIAN BOOTSTRAP

BY DONALD B. RUBIN

Educational Testing Service

The Bayesian bootstrap is the Bayesian analogue of the bootstrap. Instead of simulating the sampling distribution of a statistic estimating a parameter, the Bayesian bootstrap simulates the posterior distribution of the parameter; operationally and inferentially the methods are quite similar. Because both methods of drawing inferences are based on somewhat peculiar model assumptions and the resulting inferences are generally sensitive to these assumptions, neither method should be applied without some consideration of the reasonableness of these model assumptions. In this sense, neither method is a true bootstrap procedure yielding inferences unaided by external assumptions.

1. Introduction. Efron (1979) has discussed the use of a technique called the bootstrap to generate sampling distributions of statistics and thereby to draw inferences about parameters. The bootstrap can be viewed as a generalization of the jackknife (cf., Miller, 1974). The Bayesian bootstrap (BB) is a natural Bayesian analogue of the bootstrap.

Section 2 defines the simple versions of the bootstrap and the BB, and shows that operationally they are very similar. Section 3 illustrates the BB in two simple examples from Efron (1979). Section 4 proves that the BB is simulating the posterior distribution of a parameter under particular model specifications and shows that the BB and bootstrap are quite similar inferentially. Section 5 points out that the BB model specifications can be unreasonable and lead to inappropriate inferences; thus the bootstrap, because of its inferential similarity to the BB, can also lead to inappropriate inferences.

Serious data analyses require serious consideration of the effects of model assumptions. Although Efron avoids making any general claims for the bootstrap, its name and Efron's examples may suggest to some that it is indeed a technique for "pulling ourselves up by our bootstraps" in a data analysis, that is, for obtaining inferences insensitive to model assumptions. Neither the bootstrap nor the Bayesian bootstrap presents a general panacea for avoiding sensitivity to model assumptions.

2. The bootstrap and the Bayesian bootstrap. Suppose we have a sample of size n, say $x_1, \cdots, x_n$, which is viewed as n i.i.d. realizations of a random variable X. A statistic $\hat{\phi}$ is chosen to estimate a parameter ϕ of the distribution of X; x_i, X, ϕ, and $\hat{\phi}$ may be vectors. The bootstrap distribution of $\hat{\phi}$ is generated by taking repeated bootstrap replications from $x_1, \cdots, x_n$. One bootstrap replication from $x_1, \cdots, x_n$ is a simple random sample of size n from $x_1, \cdots, x_n$ *with replacement*, and one bootstrap replication of the statistic $\hat{\phi}$ is the value of $\hat{\phi}$ calculated on the bootstrap replicated sample. The bootstrapped distribution of $\hat{\phi}$ is generated by considering all possible bootstrap replications of $\hat{\phi}$.

For example, consider the sample mean to be the statistic $\hat{\phi}$, and let f_i be the proportion of times x_i is drawn in a bootstrap replication, $f_i = 0, 1/n, \cdots, n/n$. Then in each bootstrap replication we treat the f_i as if they were the proportions of sampled values equal to x_i and so calculate the bootstrap sample mean as $\sum_1^n f_i x_i$; the distribution of $\sum_1^n f_i x_i$ over all bootstrap replications (i.e., generated by repeated draws of the f_i) is the bootstrap distribution of the sample mean.

Note that this method of generating the sampling distribution of a statistic essentially assumes that the sample cdf is the population cdf; that is, each bootstrap replication is drawn independently from the sample cdf. More complicated estimates of the population cdf may be

Received May 1979; revised September 1979.
AMS 1970 *subject classifications.* Primary 62A15; secondary 62F15, 62G05.
Key words and phrases. Model-free inference, Dirichlet, jackknife.

drawn from, and Efron considers some. Unless otherwise stated, we will use "bootstrap" to describe the simple bootstrap that draws from the sample cdf.

The Bayesian bootstrap is analogous to the bootstrap. Each BB replication generates a posterior probability for each x_i where values of X that are not observed have zero posterior probability, just as they have zero probability under the sample cdf. The posterior probability for each of the n x_i is centered at $1/n$ but has variability. Specifically, one BB replication is generated by drawing $(n-1)$ uniform (0, 1) random variates $u_1, \cdots, u_{n-1}$, ordering them, and calculating the gaps $g_i = u_{(i)} - u_{(i-1)}$, $i = 1, \cdots, n-1$ where $u_{(0)} = 0$ and $u_{(n)} = 1$. Then $g = (g_1, \cdots, g_n)$ is the vector of probabilities to attach to the data values $x_1, \cdots, x_n$ in that BB replication. Considering all BB replications gives the BB distribution of the distribution of X and thus of any parameter of this distribution.

For example, with ϕ = mean of X, in each BB replication we calculate the mean of X as if g_i were the probability that $X = x_i$; that is, we calculate $\sum_1^n g_i x_i$. The distribution of the values of $\sum_1^n g_i x_i$ over all BB replications (i.e., generated by repeated draws of the g_i) is the BB distribution of the mean of X.

Operationally the bootstrap and BB differ only in how the probabilities attached to each x_i are drawn, and the methods of drawing the probabilities are really quite similar. It is simple to show that (a) $E(f_i) = E(g_i) = 1/n$; (b) $V(f_i) = V(g_i)(n+1)/n = (n-1)/n^3$; and (c) $C(f_i, f_j) = C(g_i, g_j) = -1/(n-1)$ where $E(\ )$, $V(\ )$, and $C(\ )$ refer to the expectation, variance, and correlation over the respective replications.[1] Consequently, if the form of the estimator, $\hat{\phi}$, is chosen so that it mimics ϕ, i.e., $\hat{\phi} = \phi$ when all $f_i = g_i$, the BB distribution of ϕ will be similar to the bootstrap distribution of $\hat{\phi}$. However, the interpretations of the resulting distributions will be different because the BB simulates the posterior distribution of the parameter ϕ, whereas the bootstrap simulates the estimated sampling distribution of a statistic $\hat{\phi}$ estimating ϕ. Thus, the BB has an inherent advantage over the bootstrap with respect to the resulting inferences about parameters: the BB generates likelihood statements about parameters rather than frequency statements about statistics under assumed values for parameters. Efron notes this shortcoming of the bootstrap in Remark F. We will use the operational similarity of the BB and the bootstrap and the inferential directness of the BB to criticize the BB and the bootstrap as general inferential tools.

3. Examples; dichotomous X and bivariate X. Suppose X is dichotomous (0 or 1) where the parameter of interest, ϕ, is the probability that X is 1. Let n_1 be the number of $x_i = 1$ and $n - n_1$ be the number of $x_i = 0$. A BB replication first obtains n probabilities from the n gaps generated by drawing $n-1$ uniform (0, 1) random numbers, and then assigns these probabilities to the n observed values of X. Thus a BB replication assigns n_1 of the probabilities to the $x_i = 1$ observations (call the sum P_1) and the remaining $n - n_1$ probabilities to the $x_i = 0$ observations; hence, the BB replication value of ϕ is P_1. Another BB replication is obtained by drawing $n-1$ new uniform random numbers and calculating a new value of P_1; continued replications generate the BB distribution of ϕ.

In this simple example, we can calculate the BB distribution of ϕ analytically: since the $n-1$ variables in each replication are i.i.d. $U(0, 1)$, the gaps follow the $n-1$ variate Dirichlet $(1, \cdots, 1)$ distribution (see Wilks, 1962, page 238). Consequently, P_1 which equals the sum of n_1 gaps, is distributed as beta$(n_1, n - n_1)$ and the BB distribution of ϕ is beta$(n_1, n - n_1)$. Note that beta$(n_1, n - n_1)$ is the Bayesian posterior distribution of ϕ from this sample with prior distribution of ϕ proportional to $[\phi(1-\phi)]^{-1}$. The BB has simply simulated the posterior distribution of ϕ under the model x_i i.i.d.

$$\begin{aligned} P(X = k \mid \phi) &= \phi && \text{if } k = 1 && \text{and} \quad P(\phi) \propto [\phi(1-\phi)]^{-1} && \text{if } 0 \le \phi \le 1 \\ &= 1 - \phi && \text{if } k = 0 && \propto 0 && \text{otherwise.} \\ &= 0 && \text{otherwise,} \end{aligned}$$

[1] In a private communication regarding an earlier draft of this paper, B. Efron pointed out these relations as a possible justification for the bootstrap procedure; the remarks of Section 5 offer a more guarded interpretation.

As Efron shows, the bootstrap distribution of $\hat{\phi}$ = "proportion of sampled x_i equal to 1" is $1/n \times$ binomial$(n, n_1/n)$, so that the mean of both ϕ and $\hat{\phi}$ is n_1/n, the variance of ϕ is $n_1(n - n_1)/[n^2(n + 1)]$ and the variance of $\hat{\phi}$ is $n_1(n - n_1)/n^3$.

Now consider a more complicated example, specifically, the correlational example presented in Efron (1979), where $x_i = (y_i, z_i)$ is bivariate and $n = 12$. Each BB replication is as follows: draw 11 $U(0, 1)$ random variables, create the 12 gaps $g_1, \cdots, g_{12}$, and calculate the correlation between Y and Z assuming $g_1, \cdots, g_{12}$ are the probabilities of $x_1, \cdots, x_{12}$; that is, calculate

$$\phi = \frac{\sum_1^{12} g_i y_i z_i - (\sum_1^{12} g_i y_i)(\sum_1^{12} g_i z_i)}{([\sum_1^{12} g_i y_i^2 - (\sum_1^{12} g_i y_i)^2][\sum_1^{12} g_i z_i^2 - (\sum_1^{12} g_i z_i)^2]^{1/2}}. \tag{3.1}$$

The result of 1000 BB replications of $\phi - \hat{\phi}^*$, where $\hat{\phi}^*$ is the observed sample correlation, is displayed in Figure 1. Notice the general similarity of our Figure 1 to Figure 1 in Efron (1979)

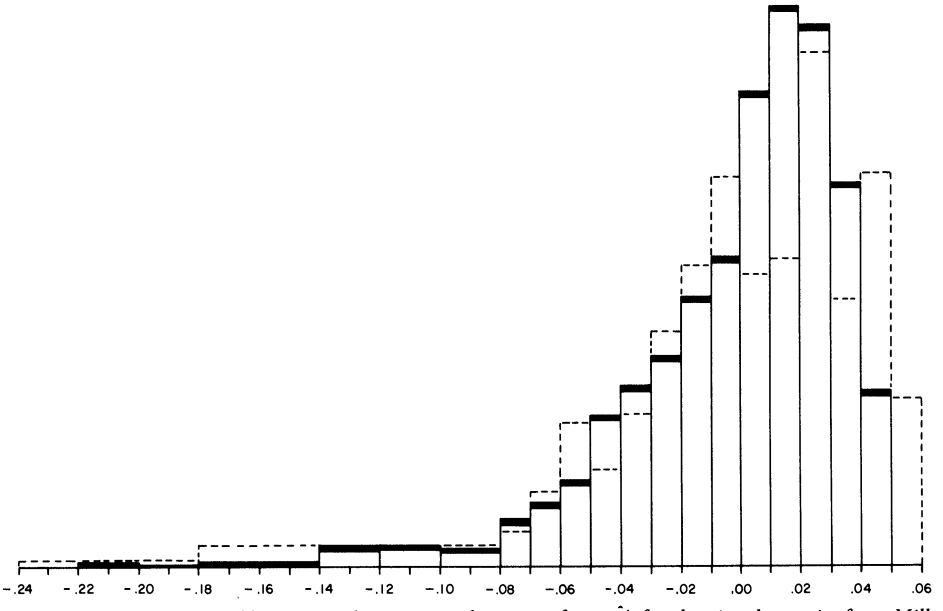

FIG. 1. *Histogram of* 1000 *Bayesian bootstrap replications of* $\phi - \hat{\phi}^*$ *for the nine data pairs from Miller* (1974); ϕ = *population correlation,* $\hat{\phi}^*$ = *observed sample correlation* = .945. *Dashed lines represent histogram of 1000 bootstrap replications of the sampling distribution of* $\hat{\phi} - \hat{\phi}^*$ *from Efron* (1979); $\hat{\phi}$ = *sample correlation.*

giving the bootstrap distribution of $\hat{\phi} - \hat{\phi}^*$, but also notice our Figure 1's smoother character due to its smoother choices of g_i (i.e., the bootstrap correlation is (3.1) with f_i in place of g_i where the possible values of f_i are 0, 1/12, $\cdots$, 1). The BB has simulated the posterior distribution of ϕ, the correlation between Y and Z, under a rather odd model that leads to the posterior probability for X spread equally among the observed values of X. The fact that Figure 1 is simulating the posterior distribution of the correlation under a specific model is proved in the next section.

4. Theory. Let $d = (d_1, \cdots, d_K)$ be the vector of all possible distinct values of X, and let $\theta = (\theta_1, \cdots, \theta_K)$ be the associated vector of probabilities

$$P(X = d_k \mid \theta) = \theta_k, \qquad \sum \theta_k = 1. \tag{4.1}$$

Let $x_1, \cdots, x_n$ be an i.i.d. sample from (4.1) and let n_k = the number of x_i equal to d_k. If the prior distribution of θ is proportional to

$$\prod_{k=1}^K \theta_k^{l_k} \qquad (0 \text{ if } \sum \theta_k \neq 1), \tag{4.2}$$

then the posterior distribution of θ is the $K-1$ variate Dirichlet distribution $D(n_1 + l_1 + 1, \cdots, n_K + l_K + 1)$ which is proportional to

$$\prod_{k=1}^{K} \theta_k^{(n_k + l_k)} \qquad (0 \text{ if } x_i \neq d_k \text{ for some } i, k \text{ or if } \sum \theta_k \neq 1). \tag{4.3}$$

This posterior distribution can be simulated using $m-1$ independent uniform random numbers, where $m = n + K + \sum_1^K l_k$. Let $u_1, \cdots, u_{m-1}$ be i.i.d. $U(0, 1)$, and let $g_1, \cdots, g_m$ be the m gaps generated by the ordererd u_i. Partition the $g_1, \cdots, g_m$ into K collections, the kth having $n_k + l_k + 1$ elements, and let P_k be the sum of the g_i in the kth collection, $k = 1, \cdots, K$. Then $(P_1, \cdots, P_K)$ follows the $K-1$ variate $D(n_1 + l_1 + 1, \cdots, n_K + l_K + 1)$ distribution (Wilks, 1963, page 238). Consequently, the BB which assigns one gap to each x_i is simulating the posterior distribution of θ and thus of a parameter $\phi = \Phi(\theta, d)$ under the improper prior distribution proportional to $\prod_{k=1}^{K} \theta_k^{-1}$.[2]

Let $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_K)$ be the statistic giving the proportion of values equal to each d_K in a sample of size n, and let $\hat{\theta}^* = (\hat{\theta}_1^*, \cdots, \hat{\theta}_K^*)$ be the observed value of $\hat{\theta}$, $\hat{\theta}_k^* = n_k/n$. The BB posterior distribution of $(\theta - \hat{\theta}^*)$ is quite similar to the distribution of $(\theta - \hat{\theta})$ assuming $\theta = \hat{\theta}^*$ (i.e., the bootstrap distribution of $(\theta - \hat{\theta})$). Specifically, the means of both distributions are 0, the variances of the kth components of the distributions are $\hat{\theta}_k^*(1 - \hat{\theta}_k^*)/(n + 1)$ and $\hat{\theta}_k^*(1 - \hat{\theta}_k^*)/n$ respectively, and the correlation between the kth and k'th components is $-\{(\hat{\theta}_k^*\hat{\theta}_{k'}^*)/[(1 - \hat{\theta}_k^*)(1 - \hat{\theta}_{k'}^*)]\}^{1/2}$ for both distributions.[3] Hence, in practice the BB inference about $\phi = \Phi(\theta, d)$ will be similar to the bootstrap inference about ϕ based on the statistic $\hat{\phi} = \Phi(\hat{\theta}, d)$. Because of this similarity, criticisms of BB inferences are criticisms of bootstrap inferences.

5. Discussion of model specifications. The model specification (4.1) is no real restriction because all data as observed are discrete. However, the choice of all $l_k = -1$ in the prior distribution for the simple BB is very questionable, and even the more general prior specification for the distribution of θ in (4.2) is questionable. These doubts about the appropriateness of the prior distribution of θ address the important issue of sensitivity of inferences to model specifications and the attempts of the bootstrap and BB to avoid this issue.

First, is it reasonable to use a model specification that effectively assumes all possible distinct values of X have been observed? Both the BB and the bootstrap operate under this assumption. In some cases inferences may be insensitive to this assumption but not always. For an extreme example, consider the probability that $X > C$ where C is larger than the largest observed X (or the probability that $A > X > B$ where A and B are between two order statistics). The simple BB and bootstrap estimate such probabilities as 0 with zero variability, which is clearly inappropriate if X can take many more than n values.

If the bootstrap and the BB can be so obviously inappropriate, what guidance do we have for when they are appropriate? Even commonly estimated parameters like the mean and variance are sensitive to model specifications. More explicitly, we expect the probability that X is greater or equal to $x_{(n)}$ to be about $(n + 1)^{-1}$; inferences about population moments will be different if we assume all this probability is concentrated at $x_{(n)}$ than if we assume the probability is spread uniformly from $x_{(n)}$ to some very large number. Because the observed

---

[2] Simulations corresponding to other prior distributions with integer l_k can also be performed; for example, with a uniform prior distribution on θ, (i.e., all $l_k = 0$) generate $n + K - 1$ uniform random variables, form $n + K$ gaps, add the first $(n_1 + 1)$ gaps together to yield the simulated value of θ_1, add the second $(n_2 + 1)$ gaps together to yield the simulated value of θ_2, and so on. However, when using a proper prior distribution, all a priori possible values of X must be specified because they have positive posterior probability.

[3] Furthermore, the predictive distribution of $(\theta - \hat{\theta})$ found by averaging the distribution of $(\theta - \hat{\theta})$ given θ over the BB posterior distribution of θ (instead of conditioning on $\theta = \hat{\theta}^*$ as with the bootstrap) has the same first two moments as the BB posterior distribution of $(\theta - \hat{\theta}^*)$. This predictive distribution for $(\theta - \hat{\theta})$ is the basis for frequency inferences for θ using the statistic $\hat{\theta}$ and the full Bayesian specification.

data cannot distinguish between these two assumptions, inferences about moments will be sensitive to the model specification of tail probabilities.

Second, even assuming all distinct values of X have been observed, is it reasonable to assume a priori independent parameters, constrained only to sum to 1, for these values? If two values of X are "close" isn't it often realistic to assume that the associated probabilities of their occurrence should be similar? Shouldn't the parameters be smoothed in some way? Both the BB and the bootstrap effectively make the assumption that these parameters are unrelated. Efron discusses more complicated versions of the bootstrap that in effect smooth the probabilities attached to values of X. The smoothed probabilities are implicitly based on some kind of model more complicated than the K distinct parameter model of the simple bootstrap. Of course there exist Bayesian analogues to the bootstrap under such models (e.g., the normal distribution, mixture models), and these models are in fact the more standard tools of Bayesian inference (cf., Box and Tiao, 1973). The trick of simulating posterior distributions of parameters by simply using gaps may no longer work, but other methods for summarizing posterior distributions are certainly available. In any case, inferences for parameters will in general change as different methods are used to smooth the probabilities attached to values of X, and neither the bootstrap nor the Bayesian bootstrap can avoid this sensitivity of inference to model assumptions.

Serious data analyses should always include serious consideration of model constraints; although knowledge of the context of a data set may make the incorporation of reasonable model constraints obvious, and although the bootstrap and the BB may be useful in many particular contexts, there are no general data analytic panaceas that allow us to pull ourselves up by our bootstraps.

Acknowledgments. I wish to thank H. Braun, B. Efron and a referee for helpful comments on an earlier draft of this paper.

REFERENCES

BOX, G. E. P. and TIAO, G. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1–26.

MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1–15.

WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.

EDUCATIONAL TESTING SERVICE
PRINCETON, N.J. 08541