Censored Data and the Bootstrap

BRADLEY EFRON*

This article concerns setting standard errors and confidence intervals for the parameters of an unknown distribution when the data is subject to right censoring. The bootstrap, which is an elaboration of the jackknife, provides a general method for answering such questions. The validity of bootstrap methods is investigated using real data, computer simulations, and, in the final section, brief theoretical considerations.

KEY WORDS: Censored data; Bootstrap; Kaplan-Meier; Confidence interval; Jackknife.

1. INTRODUCTION

The *bootstrap* (Efron 1979) is a simple and straightforward method for calculating approximated biases, standard deviations, confidence intervals, and so forth, in almost any nonparametric estimation problem. Method is a key word here, since little is known about the bootstrap's theoretical basis, except that (a) it is closely related to the jackknife; (b) under reasonable conditions it gives asymptotically correct results; and (c) for some simple problems which can be analyzed completely, for example, ordinary linear regression, the bootstrap automatically produces standard solutions. The main line of argument in Efron (1979) is through a series of examples that show the bootstrap doing a reasonable job under a variety of situations.

Here we consider another example, namely, right-censored data, and use the bootstrap to answer several questions concerning the Kaplan-Meier (product limit) estimated survival curve (Kaplan and Meier 1958): (a) What is the standard error of the Kaplan-Meier curve? (b) What is the standard error of a location estimate, for example, a trimmed mean, based on the Kaplan-Meier curve? (c) What is a reasonable confidence interval for such an estimate?

Question (a) is perfectly standard, of course, but the bootstrap answer provides a new justification for Greenwood's formula, and suggests that the bootstrap approach may be reliable in more complicated censoring situations. (Turnbull and Mitchell 1978 have used the bootstrap for analyzing a quite complicated censored data problem, a serial sacrifice experiment, with results they describe as "appearing reasonable.") Question (b) is more challenging. Work by Miller (1974) and Reid (1979) propose jackknife solutions. For the specific data problem considered in Section 4, the bootstrap results are direct and graphic, suggesting that in this example the median is more variable than, say, the 10 percent trimmed mean.

Whether or not question (c) can be answered at all, even for uncensored data, is a matter of some speculation. Small sample nonparametric confidence intervals are well known for the median (Lehmann 1975, p. 182) but not for other estimators. Exceptions to this statement include Johnson's (1978) work on generalized student-t intervals for the mean and Hartigan's (1969) typical value theory, which applies to symmetric sampling distributions. Section 5 uses a bootstrap version of Hartigan's approach. Some justification for the resulting intervals is given in Sections 5 and 6, in the form of Monte Carlo results and brief theoretical considerations.

Channing House Data. Figure 1 shows the Kaplan-Meier estimated survival curve[1] for all 97 men who lived in Channing House, a Palo Alto retirement center, from its opening in 1964 to the data collection day, July 1, 1975. The curve is obtained from data appearing in Hyde (1976). Of the 97 lifetimes, 46 were observed exactly; that is, the men died while in Channing House. The remaining 51 observations were censored; five of the men moved elsewhere and 46 were still alive on July 1, 1975. (No distinction will be made between the two different types of censoring.) The curve has its median at 1,044 months (87 years), that is, it crosses .50 at 1,044, and appears roughly symmetric about that point, though there is too much uncertainty in the upper percentiles to rule out a slightly longer tail toward the left. We will examine bootstrap methods operating on this set of data in addition to some artificial Monte Carlo situations.

The Channing House situation has some features which suggest the use of methods more sophisticated than a simple survival curve. (a) The selection mechanism which brings men into Channing House is biased toward the more affluent and highly educated. We might attempt to adjust the curve using covariate information, such as income and education level, for better comparability with the general population. (Hyde 1980 uses the opposite tactic: he compares Figure 1 with the corresponding survival curve for male participants in the TIAA investment program, another group which is above average in income and education.) (b) Figure 1 assumes that the hazard rate for men in Channing House depends only on their cal-

* Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305. The author is grateful to Kent Bailey for several useful conversations, particularly concerning the method of constructing intervals for the median described at the beginning of Section 6.

[1] The Kaplan-Meier curve is described in Section 2.

June 1981, Volume 76, Number 374
Theory and Methods Section

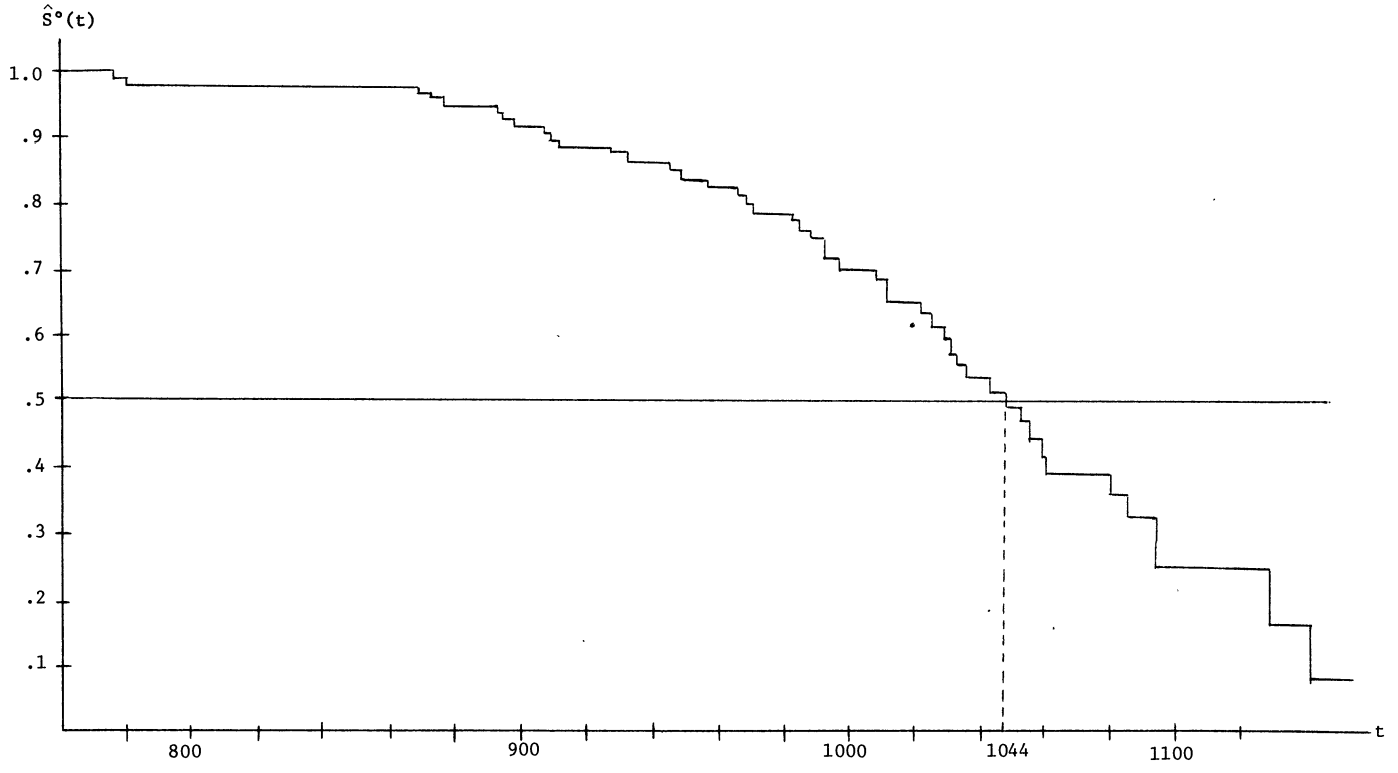

Figure 1. Kaplan-Meier Estimated Survival Curve for the Channing House Men; t = Age in Months. The Median Survival Age Is Estimated to Be 1,044 Months (87 years)

endar age. A proportional hazards model, as described in Kalbfleisch and Prentice (1980), could take account of other possible risk factors. (c) The men entered Channing House at different calendar ages. It is tacitly assumed that they would be alive at their recorded entry time if they had previously been living in Channing House rather than elsewhere. If the entry date data were available, and if we didn't trust this assumption, we could construct a survival curve taking account of left truncation as well as right censoring (see Turnbull 1976).

The bootstrap is unaffected by theoretical complications. The analysis which follows could be carried out just as easily starting from summary statistics more elaborate than the Kaplan-Meier curve. Since we do not do so, the reader should keep in mind the limitations of our analysis. For example, in saying that [1,029.5, 1,080.0] months is a 90 percent confidence interval for the median, the median referred to is that of an infinitely large data set collected and analyzed in the same way as our actual data set of 97 men.

2. THE BOOTSTRAP FOR CENSORED DATA

The bootstrap method for uncensored data is extremely simple, at least in theory. Suppose we observe $X_i = x_i$, $i = 1, 2, \ldots, n$, where the X_i are independent and identically distributed (iid) according to some unknown probability distribution F. The X_i may be real valued, two-dimensional, or take values in a more complicated space. A given parameter $\theta(F)$, perhaps the mean, median, correlation, and so forth, is to be estimated, and we agree to use the estimate $\hat{\theta} = \theta(\hat{F})$, where $\hat{F}$ is the empirical distribution function putting mass $1/n$ at each observed value x_i. We wish to assign some measure of accuracy to $\hat{\theta}$. (More general problems are also considered in Efron 1979.)

Let $\sigma(F)$ be some measure of accuracy that we would use if F were known, for example $\sigma(F) = \mathrm{SD}_F(\hat{\theta})$, the standard deviation of $\hat{\theta}$ when $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$. (The clumsier notation $\sigma_{n,\theta}(F)$ would be more explicit.) The bootstrap estimate of accuracy is simply $\hat{\sigma}_{\mathrm{BOOT}} = \sigma(\hat{F})$. In other words, $\hat{\sigma}_{\mathrm{BOOT}}$ is the measure of accuracy we would obtain if the true F equaled $\hat{F}$. Equivalently, $\hat{\sigma}_{\mathrm{BOOT}} = \sigma(\hat{F})$ is the nonparametric maximum likelihood estimate of $\sigma(F)$. Efron (1979) shows that the jackknife estimate of standard deviation is a linear approximation to $\hat{\sigma}_{\mathrm{BOOT}}$.

In order to calculate $\hat{\sigma}_{\mathrm{BOOT}}$ it is usually necessary to employ computer simulation methods. (a) A bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ is drawn from $\hat{F}$, in which each X_i^* independently takes value x_j with probability $1/n, j = 1, 2, \ldots, n$. In other words, $X_1^*, X_2^*, \ldots, X_n^*$ is an independent sample of size n drawn with replacement from the set of observations $\{x_1, x_2, \ldots, x_n\}$. (b) This gives a bootstrap empirical distribution function $\hat{F}^*$, the empirical distribution of the n values $X_1^*, X_2^*, \ldots, X_n^*$, and a corresponding bootstrap value $\hat{\theta}^* = \theta(\hat{F}^*)$. (c) Steps (a) and (b) are repeated, independently, a large number of times, say N, giving bootstrap values $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*N}$. (d) The value of $\hat{\sigma}_{\mathrm{BOOT}}$ is approxi-

mated, in the case where $\sigma(F)$ is the standard deviation, by the sample standard deviation of the $\hat{\theta}^*$ values,

$$\hat{\sigma}_{\text{BOOT}} = \sqrt{\frac{\sum_{j=1}^{N} (\hat{\theta}^{*j})^2 - \left(\sum_{j=1}^{N} \hat{\theta}^{*j}\right)^2 / N}{N-1}}. \quad (2.1)$$

Right-censored data is of the form $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$, where x_j is the jth observation, censored or not, and

$$d_j = \begin{cases} 1 & \text{if } x_j \text{ is uncensored} \\ 0 & \text{if } x_j \text{ is censored} \end{cases}. \quad (2.2)$$

For convenience we will assume $x_1 < x_2 < x_3 < \cdots < x_n$ in the calculations below to avoid notational difficulties and some minor technical problems arising from ties. The Channing House data begins (777, 1), (781, 1), (843, 0), (866, 0), (869, 1), . . . and ends with $(x_{97}, d_{97}) = (1{,}153, 0)$, a censored 96 year old. There are actually a few tied x values, but in the calculations which follow these have been broken by assigning the order given in Appendix I of Hyde (1976).

We have some estimated functional $\hat{\theta} = \theta(\text{data})$ based on $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$, for example the median $\hat{\theta} = 1{,}044$ for the Channing House data, and we wish to assign a measure of accuracy to it. We argue below that the appropriate bootstrap estimate $\hat{\sigma}_{\text{BOOT}}$ is the same as for the uncensored case, except that the individual data points are now the pairs (x_j, d_j). That is, (a) we draw a bootstrap sample $(X_1^*, D_1^*), (X_2^*, D_2^*), \ldots, (X_n^*, D_n^*)$ by independent sampling n times with replacement from $\hat{F}$, the distribution putting mass $1/n$ at each point (x_j, d_j); (b) letting data* represent this artificial data set, we calculate $\hat{\theta}^* = \theta(\text{data}^*)$; (c) we independently repeat steps (a) and (b) N times, obtaining $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*N}$; and (d) we calculate $\hat{\sigma}_{\text{BOOT}}$ as at (2.1).

This form of the bootstrap requires only that the observed pairs (x_i, d_i) are iid observations from a distribution F on $\mathcal{R}^1 \times \{0, 1\}$. We will consider statistics of the form $\hat{\theta} = \theta(\hat{S}^\circ)$, where $\hat{S}^\circ(t)$ is the Kaplan-Meier curve. It is possible to write $\hat{S}^\circ = \Phi(\hat{F})$, where Φ is a certain mapping, from distributions on $\mathcal{R}^1 \times \{0, 1\}$ to distributions on $\mathcal{R}^1$, described in Peterson (1977). Therefore $\hat{\theta} = (\Phi(\hat{F}))$ can be thought of as a statistic estimating the parameter $\theta = \theta(\Phi(F))$, and we are back in the situation described in the first paragraph of this section. The justification for $\hat{\sigma}_{\text{BOOT}}$ is then the same as before.

Suppose that in the Channing House situation a certain proportion of the men moved out of Channing House, perhaps to a hospital, a few months before death, and were then considered to be censored observations. (This was not actually the case.) The Kaplan-Meier curve would then be badly biased as an estimate of the true survival curve. The bootstrap estimate $\hat{\sigma}_{\text{BOOT}}$ is still a consistent estimate of $\sigma(F)$, the standard deviation of $\hat{\theta}$ under F, but $\sigma(F)$ itself may be meaningless. Valid use of the Kaplan-Meier curve requires an assumption that the censoring mechanism cannot look into the future as in the example just given. The random censorship model, described next, is one such assumption, perhaps the most commonly used. Random censorship constrains the class of possible distributions on $\mathcal{R}^1 \times \{0, 1\}$. Nevertheless, we will show that it leads to the same bootstrap method as previously described.

Consider the random censoring mechanism (Efron 1967, Gilbert 1962)

$$X_i = \min\{X_i^\circ, W_i\}, \quad (2.3)$$

where X_i° is the variable of interest (e.g., the age at death of the Channing House men) and W_i is some independent censoring variable. The observed quantity is the pair (X_i, D_i) with D_i equaling 1 or 0 as X_i equals X_i° or W_i, respectively. For example, the fourth Channing House data point $(x_4, d_4) = (866, 0)$ is equivalent to $x_4^\circ > 866$, $w_4 = 866$.

The Kaplan-Meier curve $\hat{S}^\circ(t)$ is a nearly unbiased estimate of the true survival curve for X°, say $S^\circ(t) \cong \text{prob}\{X^\circ > t\}$, and is given by the formula

$$\hat{S}^\circ(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1}\right)^{d_j}. \quad (2.4)$$

Here k_t is the value of k such that $t \in [x_k, x_{k+1})$; in other words, the largest observed value, censored or not, equal to or less than t.(If there is no censoring, then all $d_j = 1$, and $\hat{S}^\circ(t) = (n - k_t)/n$, the ordinary right-sided cumulative distribution function (cdf).) Kaplan and Meier (1958) show that $\hat{S}^\circ(t)$ is the nonparametric MLE for $S^\circ(t)$. If $d_n = 0$, as with the Channing House data, then $\hat{S}^\circ(x_n) > 0$; in this case we will make the arbitrary definition $\hat{S}^\circ(x_n+) = 0$, putting the missing probability mass just to the right of x_n.

An uncensored observation of X_i° corresponds to a censored observation of W_i, and vice versa, so the true survival curve for W, say $R(t) \equiv P\{W > t\}$, has nonparametric MLE

$$\hat{R}(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1}\right)^{1-d_j}. \quad (2.5)$$

Now (2.3) implies that the true survival curve for X, say $S(t) \equiv P\{X > t\}$, is the product $S(t) = S^\circ(t)R(t)$. Therefore the nonparametric MLE for $S(t)$ is

$$\hat{S}(t) = \hat{S}^\circ(t)\hat{R}(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1}\right) = \frac{n-k_t}{n}. \quad (2.6)$$

This represents the distribution putting mass $1/n$ at each observed x_j, censored or not. (Notice that (2.6) is not affected by the ambiguity in $\hat{S}^\circ$ if $d_n = 0$, or the corresponding ambiguity in $\hat{R}$ if $d_n = 1$.)

An obvious censored data version of the bootstrap procedure for randomly censored data is to independently obtain $X_i^{\circ *} \sim \hat{S}^\circ$ and $W_i^* \sim \hat{R}$, and define $X_i^* = \min(X_i^{\circ *}, W_i^*)$, D_i^* equaling 1 or 0 as X_i^* equals $X_i^{\circ *}$ or W_i^*,

respectively. Because of (2.6), X_i^* equals x_j with probability $1/n, j = 1, 2, \ldots, n$. Moreover, if $X_i^* = x_j$, then $D_i^* = d_j$, because (a) $\hat{S}^\circ$ puts mass only at those x_j having $d_j = 1$, (b) $\hat{R}$ puts mass only at those x_j having $d_j = 0$, and (c) we have assumed no ties. (Actually, all we need here is no ties between censored and uncensored values in the original data set; such ties, if they exist, are customarily assumed away by assigning censored events values just slightly larger than the values actually recorded.) This shows that the simple method of bootstrap sampling for censored data described earlier, sampling n times with replacement from $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$, is the same as the "obvious version" method given at the beginning of this paragraph.

Random censoring is a mathematically convenient assumption which can be completely unrealistic in some cases. For example, in one common situation the censoring times $w_1, w_2, \ldots, w_n$ have fixed values, all of which are known to the statistician, whether or not the x_i are censored.[2] An obvious bootstrap method in this case is to compute $\hat{S}^\circ$ as at (2.4), choose $X_1^{\circ *}, X_2^{\circ *}, \ldots, X_n^{\circ *} \overset{\text{iid}}{\sim} \hat{S}^\circ$, and define $X_i^* = \min\{X_i^{\circ *}, w_i\}$, with D_i^* equaling 1 or 0 as X_i^* equals $X_i^{\circ *}$ or w_i. This is *not* the same as sampling n times with replacement from $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$, but the Monte Carlo results in Table 2, at the end of Section 3, suggest that the numerical results may be very similar.

3. GREENWOOD'S FORMULA

We now use the bootstrap method to derive an estimate for the standard deviation of $\hat{S}^\circ(t)$ at a given value of t, and show how this estimate closely approximates the usual answer, "Greenwood's formula." Consider drawing a single bootstrap sample $(X_1^*, D_1^*), (X_2^*, D_2^*), \ldots, (X_n^*, D_n^*)$ and define

$$m_j^* = \#\text{ of times } (x_j, d_j) \tag{3.1}$$

appears in the bootstrap sample,
so $\mathbf{m}^* = (m_1^*, m_2^*, \ldots, m_n^*)$ is an n-category multinomial, n draws, probability $1/n$ for each category: $\mathbf{m}^* \sim \text{mult}(n, \mathbf{1}/n)$. For example, m_1^* might equal 2, in which case 2 of the n pairs (X_i^*, D_i^*) would equal (x_1, d_1), while $m_2^* = 0$ so (x_2, d_2) would not appear in that particular bootstrap sample. Also define

$$M_j^* \equiv \sum_{i=j}^{n} m_i^*, \quad j = 1, 2, \ldots, n, \tag{3.2}$$

so $M_1^* = n$, $M_2^* = n - m_1^*$, and so forth.

The Kaplan-Meier curve based on the bootstrap data $\{(x_1^*, d_1^*), (x_2^*, d_2^*), \ldots, (x_n^*, d_n^*)\}$ is

$$\hat{S}^{\circ *}(t) = \prod_{j=1}^{k_t} \left(1 - \frac{m_j^*}{M_j^*}\right)^{d_j}. \tag{3.3}$$

[2] This is the case for the Channing House data if we ignore the possibility of a resident leaving before death. Then w_i equals resident i's age on July 1, 1975.

(This is the time honored "life-table" estimate for a survival curve. When there are no ties between censored and uncensored observations, as is the case here, (3.3) can easily be derived from the Kaplan-Meier form (2.4).) The bootstrap estimate of standard deviation for $\hat{S}^\circ(t)$, t fixed, is

$$\hat{\sigma}_{\text{BOOT}} = \sqrt{\text{var}^*\, \hat{S}^{\circ *}(t)}, \tag{3.4}$$

where "var*" indicates the variance of (3.3) with the observed data $\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\}$ fixed and the vector $\mathbf{m}^*$ varying according to the multinomial distribution mult$(n, \mathbf{1}/n)$.

Greenwood's formula for the standard deviation of $\hat{S}^\circ(t)$ is

$$\hat{\sigma}_{\text{GREEN}} = \hat{S}^\circ(t) \sqrt{\sum_{j=1}^{k_t} \frac{d_j}{(n-j)(n-j+1)}}, \tag{3.5}$$

(see Kaplan and Meier 1958). Table 1 compares $\hat{\sigma}_{\text{GREEN}}$ with $\hat{\sigma}_{\text{BOOT}}$ for the Channing House data, at nine different values of t: $x_{10}, x_{20}, \ldots, x_{90}$. The values of $\hat{\sigma}_{\text{BOOT}}$ were derived by Monte Carlo simulation, as described in Section 2, with $N = 400$. The agreement between $\hat{\sigma}_{\text{GREEN}}$ and $\hat{\sigma}_{\text{BOOT}}$ is excellent.

It is easy to justify this agreement theoretically, and we do so at the end of this section, but first it is worth noting an important distinction. Greenwood's formula, as traditionally derived, is by no means trivial and requires a rather sophisticated analysis of censored data and the Kaplan-Meier curve (see Section 6 of Kaplan and Meier 1958). The bootstrap values $\hat{\sigma}_{\text{BOOT}}$ in Table 1 require a lot more computation than $\hat{\sigma}_{\text{GREEN}}$, but also a lot less analysis of the specific problem. In essence we have

Table 1. The Standard Deviation of the Kaplan-Meier Curve for the Channing House Data, Estimated by Greenwood's Formula and by the Bootstrap, N = 400 Bootstrap Replications. Estimates of Standard Deviation Based on Percentiles of the Bootstrap Distribution Are Also Given

$t =$	895 x_{10}	936 x_{20}	957 x_{30}	973 x_{40}	1001 x_{50}	1016 x_{60}	1033 x_{70}	1058 x_{80}	1098 x_{90}
$\hat{S}^\circ(t)$	.927	.862	.838	.786	.701	.653	.555	.442	.279
$\hat{\sigma}_{\text{GREEN}}$	.027	.036	.038	.044	.051	.055	.062	.067	.073
$\hat{\sigma}_{\text{BOOT}}$	.025	.036	.039	.044	.052	.057	.064	.069	.076
Estimates of $\hat{\sigma}$ based on the bootstrap percentiles									
$\frac{75\% - 25\%}{1.35}$	.024	.036	.040	.045	.054	.061	.066	.069	.074
$\frac{90\% - 10\%}{2.56}$	.023	.036	.039	.045	.052	.056	.066	.072	.074
Percentiles of the bootstrap distribution of $\hat{S}^{\circ *}(t)$									
10%	.895	.812	.790	.731	.636	.582	.471	.357	.183
25%	.910	.839	.813	.758	.668	.613	.509	.396	.233
50%	.927	.865	.841	.788	.700	.651	.548	.446	.284
75%	.942	.888	.867	.819	.741	.695	.598	.489	.332
90%	.955	.904	.890	.845	.768	.725	.641	.541	.373

Table 2. Ten Monte Carlo Trials of the Fixed Censoring Situation Described in the Text, N = 400 Bootstrap Replications per Trial. The Fixed Censoring Version of the Bootstrap Produces Values, $\hat{\sigma}_{FIXED}$, Close to $\hat{\sigma}_{BOOT}$ and $\hat{\sigma}_{GREEN}$

	$t = x_5$			$t = x_{10}$			$t = x_{15}$		
Trial	$\hat{\sigma}_{BOOT}$	$\hat{\sigma}_{FIXED}$	$\hat{\sigma}_{GREEN}$	$\hat{\sigma}_{BOOT}$	$\hat{\sigma}_{FIXED}$	$\hat{\sigma}_{GREEN}$	$\hat{\sigma}_{BOOT}$	$\hat{\sigma}_{FIXED}$	$\hat{\sigma}_{GREEN}$
1	.100	.092	.097	.114	.111	.111	.114	.111	.111
2	.090	.095	.097	.101	.114	.110	.135	.127	.129
3	.097	.104	.097	.114	.117	.112	.095	.101	.097
4	.094	.092	.097	.114	.110	.112	.112	.110	.110
5	.096	.097	.097	.096	.097	.097	.135	.154	.135
6	.096	.098	.097	.107	.107	.111	.128	.137	.123
7	.098	.097	.097	.112	.110	.111	.129	.125	.125
8	.098	.093	.097	.116	.104	.112	.114	.103	.111
9	.095	.096	.097	.113	.115	.112	.113	.113	.111
10	.093	.099	.097	.109	.111	.112	.104	.095	.102

made the computer derive Greenwood's formula for us. This point becomes more crucial in the following sections, where we discuss problems in which the equivalent of Greenwood's formula is not easily available.

Table 1 also gives the percentiles of the 400 bootstrap values of $\hat{S}^{\circ *}(t)$, for each value of t, and normal theory estimates of the standard deviation based on these percentiles. One might prefer such a definition of "$\hat{\sigma}_{BOOT}$" if outlying values of $\hat{S}^{\circ *}$ were of concern, but at least in this case the results are nearly the same.

The last paragraph of Section 2 discussed fixed censoring, and a different bootstrap method which seems more appropriate to this situation. Table 2 compares $\hat{\sigma}_{FIXED}$, the bootstrap estimate based on fixed censoring, with $\hat{\sigma}_{BOOT}$ and $\hat{\sigma}_{GREEN}$ for the following situation: $n = 20$; X_i° a standard exponential random variable ($S^{\circ}(t) = e^{-t}$); $w_1, w_2, \ldots, w_{10} = .693$; $w_{11}, w_{12}, \ldots, w_{20} = \infty$. Here half of the sample is potentially censored at the true median, and half is totally uncensored. Ten Monte Carlo trials[3] were run with $N = 400$ bootstrap replications per trial. We see close agreement between $\hat{\sigma}_{FIXED}$, $\hat{\sigma}_{BOOT}$, and $\hat{\sigma}_{GREEN}$. This same reassuring agreement was observed in other Monte Carlo trials, with the w_i values chosen differently.

Proof that $\hat{\sigma}_{GREEN} \approx \hat{\sigma}_{BOOT}$. From (3.3), $\log \hat{S}^{\circ *}(t) = \sum_{j=1}^{k_t} d_j \log(1 - m_j^*/M_j^*)$. We have $M_j^* \sim \text{bi}(n, n_j/n)$ and $m_j^* \mid M_j^* \sim \text{bi}(M_j^*, 1/n_j)$, where $n_j \equiv n - j + 1$. For n_j large, standard "delta theory" calculations give the approximation $\text{var}_*(\log \hat{S}^{\circ *}(t)) \doteq \sum_{j=1}^{k_t} d_j/n_j^2$, and so $\text{var}_* \hat{S}^{\circ *}(t) \doteq (S^{\circ}(t))^2 \sum_{j=1}^{k_t} d_j/n_j^2 \doteq \hat{\sigma}_{GREEN}^2$. (This proof is quite similar to the usual derivation of Greenwood's formula for grouped data.)

More careful calculations are possible, along the Martingale lines used by Kaplan and Meier (1958) in their derivation of Greenwood's formula, but in fact all derivations of (3.5) involve gross approximations near the right end of the data, where n_j gets small. Nevertheless, Table 1 shows excellent agreement between $\hat{\sigma}_{GREEN}$ and $\hat{\sigma}_{BOOT}$, even at x_{90} ($n_j = 8$).

4. LOCATION ESTIMATES

The median is often favored as a location estimate in censored data problems because, in addition to its usual advantage of easy interpretability, it least depends upon the right rail of the Kaplan-Meier curve $\hat{S}^{\circ}(t)$, which can be highly unstable if censoring is heavy. Using the bootstrap, we can estimate the bias and standard deviation of the sample median, or of any other location estimate. Table 3 gives estimated biases and standard deviations for seven such estimators, including the median, the mean, and various trimmed and Winsorized means. (As with the median, each estimate is defined as the value of the corresponding functional evaluated for the distribution $\hat{S}^{\circ}$.) The estimated standard deviation is $\hat{\sigma}_{BOOT}$, as given in (2.1). The bias estimate is the difference between the average of the bootstrap values and the observed value of the statistic, $(1/N) \sum_{j=1}^{N} (\theta^{*j}) - \hat{\theta}$, a quantity closely related to the jackknife estimate of bias (Efron 1979). The same $N = 1{,}600$ bootstrap replications were used for all seven estimators.

It is worth noting that the estimated bias and variance both decrease as the amount of trimming or Winsorizing decreases, the worst case being the median, the best being the mean. Normality and symmetry of the bootstrap distribution also tend to improve as we move toward the mean, as shown by the last four columns of Table 3. There is no extant theory for strictly interpreting Table 3, but it certainly does not demonstrate the superiority of the sample median as a point estimator for this particular data set. Although the point estimates for the median and mean are quite similar, 1039.9 and 1037.8, respectively, after bias correction (i.e., after subtraction of the bias estimate from the observed value), confidence intervals for the median will be about 33 percent larger than the corresponding intervals for the mean.

[3] Notice that there are two levels of Monte Carlo sampling involved: "trial" refers to a new drawing of $X_1^{\circ}, X_2^{\circ}, \ldots, X_{20}^{\circ}$, while "replication" refers to the bootstrap sampling, with the data $\{(x_1, d_1), \ldots, (x_{20}, d_{20})\}$ held fixed.

Table 3. Seven Location Estimates for the Channing House Data. Bias and Standard Deviation Estimates Are Based on N = 1,600 Bootstrap Replications. If the Bootstrap Distribution Were Perfectly Normal, $\hat{\sigma}_{BOOT}$ Based on (2.1) Would Exactly Equal the Standard Deviation Based on Percentiles. (Some smoothing has been done on the percentiles for the median. See Section 5.) (a) (75% − 25%)/1.350, (b) (90% − 10%)/2.564, (c) (95% − 5%)/3.290, (d) Asymmetry Ratio (95% − 50%)/(50% − 5%)

Statistic	Observed Value	Bootstrap Estimates		Percentiles of the Bootstrap Distribution							SD Based on Bootstrap Percentiles			
		Bias	$\hat{\sigma}_{BOOT}$	5%	10%	25%	50%	75%	90%	95%	a	b	c	d
1. Median	1,044.0	4.1	14.0	1,029.5	32.0	37.5	46.0	55.5	60.0	80.0	13.9	10.9	15.5	2.06
2. .25 trimmed (each tail)	1,047.4	1.8	12.5	1,028.1	32.8	39.5	47.6	56.4	65.3	70.0	12.5	12.7	12.7	1.15
3. .10 trimmed (each tail)	1,045.8	.6	11.4	1,026.7	30.4	37.6	45.3	52.7	60.4	64.2	11.2	11.7	11.4	1.02
4. .05 trimmed (each tail)	1,043.2	.4	10.8	1,025.1	28.5	35.2	42.5	49.8	56.7	61.0	10.8	11.0	10.9	1.06
5. .25 Winsorized (each tail)	1,044.5	4.5	12.5	1,028.6	32.5	38.8	47.4	56.9	64.2	68.8	13.4	12.4	12.2	1.14
6. .10 Winsorized (each tail)	1,041.4	1.4	10.7	1,025.0	28.3	34.7	41.6	49.2	55.8	59.9	10.8	10.7	10.6	1.10
7. Mean	1,038.2	.4	10.5	1,021.1	24.3	30.8	37.6	44.8	51.4	54.6	10.4	10.6	10.2	1.03

5. SMALL SAMPLE CONFIDENCE INTERVALS

We now discuss a simple method for constructing confidence intervals based on the bootstrap distribution. This method, which has nothing in particular to do with censored data, can be employed for any real-valued parameter θ, given the bootstrap distribution of $\hat{\theta}^*$, but will be illustrated in terms of the median. It is more ambitious than simply using $\hat{\theta} \pm c \cdot \hat{\sigma}_{BOOT}$, where c is a constant taken from the normal or t tables, and in particular can give quite asymmetric intervals in small samples. A brief theoretical rationale for this method is given in Section 6.

Suppose then that we have computed the cdf of the bootstrap distribution for some real-valued parameter θ, estimated by $\hat{\theta} = \theta(\hat{F})$, say

$$\mathrm{CDF}_*(t) \equiv \mathrm{prob}_*\{\hat{\theta}^* \le t\}. \tag{5.1}$$

For a given α, we can construct what we hope is a $1 - 2\alpha$ central confidence interval by using the appropriate percentiles of CDF_*, say

$$\hat{\theta}_{LOW} \equiv \mathrm{CDF}_*^{-1}(\alpha), \quad \hat{\theta}_{UP} \equiv \mathrm{CDF}_*^{-1}(1 - \alpha), \tag{5.2}$$

as the lower and upper points of the interval. Section 6 gives some theoretical arguments for believing that $[\hat{\theta}_{LOW}, \hat{\theta}_{UP}]$ is a reasonable candidate for a $1 - 2\alpha$ confidence interval. We will call this construction the Percentile Method. For example, looking at Table 3, the percentile method gives [1021.1, 1054.6] as a 90 percent central confidence interval for the mean, in the Channing House data.[4]

Bootstrap distributions are inherently discrete, but in most cases the probability atoms are small enough as to have no practical effect on the percentile method. An exception is the sample median. Figure 2 shows the bootstrap distribution of the sample median for the Channing House data, with the probability mass supported, as it must be, on the uncensored observed lifetimes. The percentiles for the median given in Table 3 are based on a smoothed version of the usual cdf: for t equal to an uncensored observed lifetime, $\mathrm{CDF}_*(t) = \mathrm{prob}_*\{\hat{\theta}^* < t\} + \frac{1}{2}\mathrm{prob}_*\{\hat{\theta}^* = t\}$. Linear interpolation is used to fill in the values of $\mathrm{CDF}_*(t)$ between the probability mass points, as shown in Figure 2.

Several Monte Carlo experiments were run to ascertain the actual coverage probabilities of the percentile method intervals for the median. In these experiments both X_i°

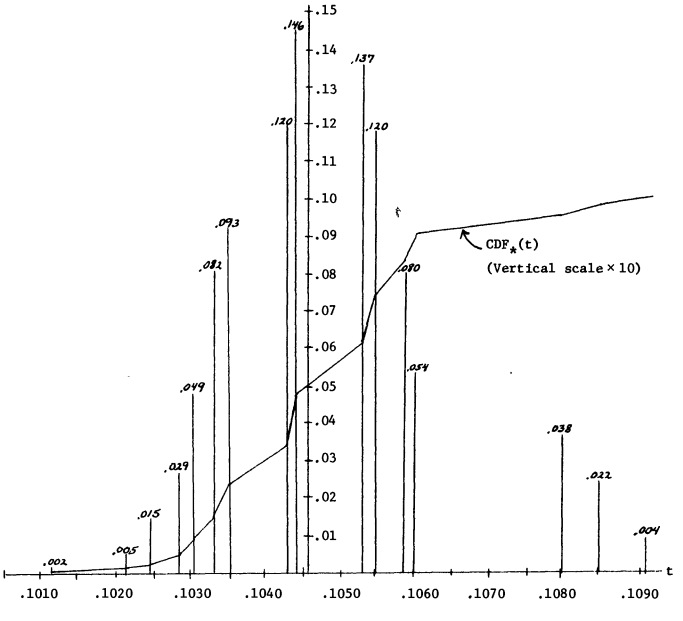

Figure 2. Bootstrap Distribution for the Sample Median, Channing House Data, Based on N = 1,600 Bootstrap Replications. The Distribution Is Supported on the Uncensored Observed Lifetimes. The Smoothed Version of CDF Described in the Text Is Also Shown*

[4] Most often we will be working with an approximation to CDF_*, based on some number N of Monte Carlo trials, $N = 1{,}600$ in this case. Empirically the author has found such large values of N necessary to numerically stabilize the extreme percentiles. For only estimating $\hat{\sigma}_{BOOT}$, $N = 100$ performs quite well in most examples.

Table 4. Eight Monte Carlo Experiments Comparing Nominal and Actual Probabilities for the Percentile Method. The Bootstrap Distribution of the Sample Median Is Used to Derive Confidence Intervals for the True Median. The Agreement Is Good. See Text for Details

Experiment Number	Sample Size n	Prob {Uncensored}	Actual Coverage Probabilities at Nominal (lower) level .90	.75	.50	.25	.10	# Bootreps per Trial, N	# Trials
1.	25	.50	.92	.76	.47	.22	.08	400	500 × 400
2.	25	.50	.93	.78	.49	.21	.08	200	200 × 100
3.	25	.50	.91	.73	.45	.20	.07	100	1,000 × 100
4.	21	.67	.91	.76	.48	.21	.08	1,000	1,000 × 200
5.	21	.67	.90	.75	.47	.21	.08	400	500 × 400
6.	21	.67	.91	.75	.49	.22	.08	200	200 × 50
7.	21	.67	.90	.74	.47	.21	.08	100	1,000 × 100
8.	21	.67	.90	.75	.48	.23	.09	50	400 × 50

and W_i had exponential distributions, prob$\{X_i^\circ > t\} = e^{-t}$, prob$\{W_i > t\} = e^{-t/c}$, $t \geq 0$. The probability of $X_i = \min\{X_i^\circ, W_i\}$ being uncensored (that is, $D_i = 1$) equals $c/(c + 1)$. Random samples $(X_i, D_i) = (x_i, d_i)$, $i = 1, 2, \ldots, n$ were drawn in this way, and the percentile method, based on the bootstrap distribution of the sample median, used to generate confidence intervals for the true median, exactly as described earlier. Table 4 shows generally good agreement between the nominal and actual coverage probabilities. For example, with $n = 25$, $c = 1$ (so prob$\{X_i$ uncensored$\} = .5$), and $\alpha = .10$ the first experiment had $\hat{\theta}_{\text{LOW}} < \theta$ 92 percent of the time, compared to a nominal 90 percent. (Technical note: It is equivalent to generate the X_i according to prob$\{X_i > t\} = \exp\{-[1 + (1/c)]t\}$, $t > 0$, and then $D_i = 1$ (or 0) with probability $c/(c + 1)$ (or $1/(c + 1)$), independently of $X_i = x_i$. For the first experiment, "500 × 400" number of trials refers to 500 realizations of $x_1, x_2, \ldots, x_{25}$, and for each of these 400 independent realizations of $d_1, d_2, \ldots, d_{25}$. Note that this does not constitute 20,000 independent replications.)

6. THEORETICAL JUSTIFICATION FOR THE PERCENTILE METHOD

First we consider the case of constructing a confidence interval for the median. Looking at Table 1, we can apply the percentile method to obtain a confidence interval for $S^\circ(t)$, t fixed. For $t = 1{,}016$ the percentile method gives the central 80 percentile interval $[\hat{S}_{\text{LOW}}^\circ(t), \hat{S}_{\text{UP}}^\circ(t)] = [.582, .725]$. For $t = 1{,}033$ the interval is $[.471, .641]$. A plausible method for constructing an 80 percentile confidence interval for the true median θ (defined as the minimum θ for which $S^\circ(\theta) = .5$) is to exclude any t for which $[\hat{S}_{\text{LOW}}^\circ(t), \hat{S}_{\text{UP}}^\circ(t)]$ does not include the value .5. Thus $t = 1016$ would be excluded and $t = 1{,}033$ would be included in the 80 percentile interval for θ.

This interval turns out to be the same as the 80 percentile interval $[\hat{\theta}_{\text{LOW}}, \hat{\theta}_{\text{UP}}]$, except for minor differences due to the smoothing of CDF$_*$ shown in Figure 2. The reason that the two methods agree is the equivalence of these two events,

$$\{S^{\circ *}(t) > \tfrac{1}{2}\} \Leftrightarrow \{\hat{\theta}^* > t\} \tag{6.1}$$

for every bootstrap sample. So we see that whether or not the percentile method is any good is general, it is at least consistent with itself in the case of the median.

This same argument can be used with any m estimator $\hat{\theta}$, defined as the value t for which

$$\sum_{i=1}^{n} \psi(x_i - t) = 0, \tag{6.2}$$

$\psi(z)$ monotone and antisymmetric about 0. The percentile interval for the true θ, which is defined by $\int \psi(x - \theta)\, dF(x) = 0$, can be constructed directly as at (5.2); or for each t we can construct the percentile interval for $M(t) = \int \psi(x - t)\, dF(x)$, based on the estimate $\hat{M}(t) = \sum \psi(x_i - t)/n$, and include in the interval for θ only those t for which $[\hat{M}_{\text{LOW}}(t), \hat{M}_{\text{UP}}(t)]$ includes 0. The two methods will agree, since $\{\hat{M}^*(t) > 0\} \Leftrightarrow \{\hat{\theta}^* > t\}$.

Brookmeyer and Crowley (1978) and Emerson (1979) have suggested constructing intervals for the median in the manner described at the beginning of this section, with one important difference: for each t the interval for $S^\circ(t)$ is derived in the standard way (based on the sampling distribution of $\hat{S}^\circ(t)$, using normal or binomial approximations) rather than by the bootstrap. Boos (1980) has made the same suggestion for m estimators. These methods require less computation than the percentile method, at the expense of greater reliance on approximation theory. In any case, it is nice to see that the two approaches agree in principle.

Efron (1980, Sec. 6) gives three general arguments supporting the percentile method for constructing confidence intervals for a real-valued parameter θ. The first is a Bayesian argument, which shows that a diffuse prior (actually a Dirichlet prior) on the class of all distributions F leads to posterior probability intervals for θ closely approximated by the percentile interval (5.2).[5]

The second argument involves monotone transformations $\phi = g(\theta)$ for which $\hat{\phi} = g(\hat{\theta})$ is pivotal, that is, $\hat{\phi}^* - \hat{\phi}$ has the same distribution under $\hat{F}$ as does $\hat{\phi} - \phi$

[5] In a recent unpublished paper, Rubin (1979), the bootstrap is criticized on these grounds, with the suggestion of instead doing the full Bayesian analysis beginning with a more sensible prior.

under F. If such a transformation exists, and if $\hat{\phi} - \phi$ is symmetrically distributed about 0, then the percentile interval (5.2) has the correct coverage probability. More strongly, it is the inverse mapping of the obvious interval for ϕ based on $\hat{\phi}$. The argument is reminiscent of transformations in parametric problems, for example, Fisher's transformation $\phi = .5 \log(1 + \rho)/(1 - \rho)$ for the correlation coefficient. It does not require knowledge of the transformation $g(\cdot)$, only its existence, but this itself may be hard to believe in nonparametric situations.

The third argument involves Hartigan's (1969) typical value theory. This is a method of constructing confidence intervals, quite similar to the percentile method, which yields exactly correct coverage probabilities for symmetric location problems. The two methods are shown to be asymptotically equivalent, with a particularly close connection in the case of the median.

None of these arguments is overwhelming, and in fact the percentile method sometimes performs poorly. One improvement, a bias correction depending on how far $P_*\{\hat{\theta}^* > \hat{\theta}\}$ deviates from .5, is suggested in Efron (1980). (These deviations were quite small for the estimates used on the Channing House data, Table 3, which is why the bias correction was not considered here.) On the basis of current evidence, both numerical and theoretical, the bias-corrected percentile method seems to merit further investigation as a general purpose tool for constructing nonparametric confidence intervals.

[*Received February 1980. Revised November 1980.*]

REFERENCES

BOOS, D. (1980), "A New Method for Constructing Approximate Confidence Intervals From *M*-Estimates," *Journal of the American Statistical Association,* 75, 142–145.

BROOKMEYER, R., and CROWLEY, J. (1978), "A Confidence Interval for the Median Survival Time," Technical Report No. 2, Wisconsin Clinical Cancer Center, University of Wisconsin.

EFRON, B. (1967), "The Two Sample Problem With Censored Data," *Proceedings of the Fifth Berkeley Symposium,* IV, 831–853.

——— (1979a), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics,* 7, 1–26.

——— (1979b), "Computers and the Theory of Statistics: Thinking the Unthinkable," *SIAM Review,* 21, 460–480.

EMERSON, J. (1979), "Nonparametric Confidence Intervals for Quantiles in the Presence of Partial Right Censoring," Technical Report 50Z, Sidney Farber Cancer Institute, Boston, Mass.

FORSYTHE, A., and HARTIGAN, J.A. (1970), "Efficiency and Confidence Intervals Generated by Repeated Subsample Calculations," *Biometrika,* 57, 629–640.

GILBERT, J.D. (1962), "Random Censorship," unpublished Ph.D. thesis, University of Chicago.

HARTIGAN, J.A. (1969), "Using Subsample Values as Typical Values," *Journal of the American Statistical Association,* 64, 1303–1317.

HYDE, J. (1980), "Testing Survival With Incomplete Observations," *Biostatistics Casebook,* New York: John Wiley.

JOHNSON, N.J. (1978), "Modified *t* Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association,* 73, 536–544.

KALBFLEISCH, J.D., and PRENTICE, R.L. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.

KAPLAN, E.L., and MEIER, P. (1958), "Nonparametric Estimation From Incomplete Samples," *Journal of the American Statistical Association,* 53, 457–481.

LEHMANN, E.L. (1975), *Nonparametrics: Statistical Methods Based on Ranks,* San Francisco: Holden-Day.

MARITZ, J.S. (1979), "A Note on Exact Robust Confidence Intervals for Location," *Biometrika,* 66, 163–166.

MILLER, R.G. (1974), "Jackknifing Censored Data," Technical Report No. 14, Department of Statistics, Stanford University.

PETERSON, A. (1977), "Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions," *Journal of the American Statistical Association,* 72, 854–858.

REID, N. (1979), "Influence Functions for Censored Data," Technical Report No. 46, Department of Statistics, Stanford University.

RUBIN, D.B. (1979), "A Bayesian Bootstrap," Unpublished report, Princeton, N.J.: Educational Testing Service.

TURNBULL, B. (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored, and Truncated Data," *Journal of the Royal Statistical Society,* Ser. B, 38, 290–295.

TURNBULL, B.W., and MITCHELL, T.J. (1978), "Exploratory Analysis of Disease Prevalence Data From Survival Sacrifice Experiments," *Biometrics,* 34, 555–570.