CENSORED DATA AND THE BOOTSTRAP

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 53
FEBRUARY 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 RO1 GM21215-05

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CENSORED DATA AND THE BOOTSTRAP

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 53
FEBRUARY 1980

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215-05

Also prepared under National Science Foundation Grant MPS 77-16974, and issued as Technical Report No. 143, Stanford University - Department of Statistics.

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
CENSORED DATA AND THE BOOTSTRAP

Bradley Efron

Abstract

This paper 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 methodology for answering such questions. The validity of bootstrap methods is investigated using real data, computer simulations, and, in the final section, theoretical considerations.
CENSORED DATA AND THE BOOTSTRAP

Bradley Efron

1. Introduction.

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

Here we consider another example, right-censored data, and use the bootstrap methodology to answer several questions concerning the Kaplan-Meier ("product limit") estimated survival curve, Kaplan and Meier (1958): (1) What is the standard error of the Kaplan-Meier curve? (2) What is the standard error of a location estimate, for example a trimmed mean, based on the Kaplan-Meier curve? (3) What is a reasonable confidence interval for such an estimate?

Question (1) 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 serial sacrifice experiment, a quite complicated censored data problem, with results they describe as "appearing reasonable"). Question (2) 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% trimmed mean.

Whether or not Question (3) 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, page 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 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).

---

† The Kaplan-Meier curve is described in Section 2.
Of the 97 lifetimes, 46 were observed exactly, that is the men died while in Channing House, while the remaining 51 were "censored", 5 of the men moving elsewhere and 46 still being alive on July 1, 1975. The curve has its "median" at 1044 months (87 years), i.e. it crosses .50 at 1044, and appears nearly 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 the bootstrap methodology operating on this set of data, in addition to some artificial Monte Carlo situations.

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 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 functional \( \theta(F) \), perhaps the mean, median, correlation, etc., 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 \( \sigma(F) = \text{SD}_F(\hat{\theta}) \), the standard deviation of \( \hat{\theta} \) when \( X_1, X_2, \ldots, X_n \overset{iid}{\sim} F \). (The clumsy notation \( \sigma_{n,\theta}(F) \) would be more explicit.) The bootstrap estimate of accuracy is simply \( \hat{\sigma}_{\text{BOOT}} = \sigma(\hat{F}) \).

In other words, \( \hat{\sigma}_{\text{BOOT}} \) is the measure of accuracy we would obtain if
the true $\hat{F}$ equaled the nonparametric MLE $\hat{F}$. Efron (1979) shows that
the jackknife estimate of standard deviation is a linear approximation
to $\hat{\sigma}_{\text{BOOT}}$.

In order to calculate $\hat{\sigma}_{\text{BOOT}}$, it is usually necessary to employ
computer simulation methods. (i) A "bootstrap sample" $X_1^*, X_2^*, \ldots, X_n^*$
is drawn from $\hat{F}$, in which each $X_j^*$ 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 $\{x_1, x_2, \ldots, x_n\}$. (ii) 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 $\hat{\theta}^* = \theta(\hat{F}^*)$. (iii)
Steps (i) and (ii) are repeated, independently, a large number of times,
say $N$, giving bootstrap values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_N^*$. (iv) The value of
$\hat{\sigma}_{\text{BOOT}}$ is approximated, in the case where $\sigma(F)$ is the standard devia-
tion, by the sample standard deviation of the $\hat{\theta}^*$ values,

$$\hat{\sigma}_{\text{BOOT}} = \sqrt{\frac{\sum_{j=1}^{N} (\hat{\theta}_j^*)^2 - \left( \frac{\sum_{j=1}^{N} \hat{\theta}_j^*}{N} \right)^2}{N-1}}.$$ (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 $j$th
ordered observation, censored or not, and

$$d_j = \begin{cases} 
1 & \text{if } x_j \text{ is uncensored} \\
0 & \text{if } x_j \text{ is censored}
\end{cases}.$$ (2.2)
We will assume \( x_1 < x_2 < x_3 < \ldots < x_n \), to avoid notational difficulties and some minor technical problems arising from ties. The Channing House data begins \((777, 1), (781, 1), (843, 0), (866, 0), (869, 1), \ldots \) and ends with \((x_{97}, d_{97}) = (1153, 0)\), an uncensored 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 \( \{(x_1, d_1^1), (x_1, d_1^2), \ldots, (x_n, d_n^1)\} \), for example the median \( \hat{\theta} = 1044 \) for the Channing House data, and we wish to assign a measure of accuracy to it. We argue below that the appropriate bootstrap estimate \( \hat{\theta}_{\text{BOOT}} \) is the same as for the uncensored case, except that the individual data points are now the pairs \((x_j, d_j^*)\). That is (i) 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 the set of pairs \( \{(x_1, d_1^1), (x_2, d_2^1), \ldots, (x_n, d_n^1)\} \); (ii) letting \( \text{data}^* \) represent this artificial data set, we calculate \( \hat{\theta}^* = \theta(\text{data}^*) \); (iii) we independently repeat steps (i) and (ii) \( N \) times, obtaining \( \hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*N} \); and (iv) we calculate \( \hat{\theta}_{\text{BOOT}} \) as at (2.1).

To see why this method is appropriate, consider the random censoring mechanism, Efron (1967), Gilbert (1962),

\[
X_i = \min\{X_i^*, W_i\} ,
\]  
(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 \(X^o_i\) or \(W_i\), respectively. For example, the fourth Channing House data point \((x_4^o, d_4) = (866, 0)\) is equivalent to \(x_4^o > 866, w_4 = 866\).

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

\[
\hat{S}^o(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1}\right)^{d_j}.
\] (2.4)

Here \(k_t\) is the value of \(k\) such that \(t \in [x_{k-1}, x_k]\); 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 \(\hat{S}^o(t) = (n-k_t)/n\), the ordinary right-sided c.d.f.) Kaplan and Meier (1958) show that \(\hat{S}^o(t)\) is the nonparametric MLE for \(S^o(t)\). If \(d_n = 0\), as with the Channing House data, then \(\hat{S}^o(x_n) > 0\); in this case we will make the arbitrary definition \(\hat{S}^o(x_n) = 0\), putting the missing probability mass just to the right of \(x_n\).

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

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

Now (2.3) implies that the true survival curve for \(X\), say \(S(t) \equiv \text{Prob}\{X > t\}\), is the product \(S(t) = S^o(t)R(t)\). Therefore the nonparametric MLE for \(S(t)\) is
\[ \hat{S}(t) = \hat{S}^o(t) \hat{R}(t) = \prod_{j=1}^{k_t} \frac{(n-j)}{(n-j+1)} \]
\[ = \frac{n-k_t}{n}. \] (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 \(\hat{S}^o\) if \(d_n=0\), or the corresponding ambiguity in \(\hat{R}\) if \(d_n=1\).)

The obvious censored data version of the bootstrap procedure is to independently obtain \(X^*_1 \sim \hat{S}^o\) and \(W^*_i \sim \hat{R}\), and define \(X^*_i = \min(X^*_i, W^*_i)\), \(D^*_i\) equaling 1 or 0 as \(X^*_i\) equals \(X^o_i\) 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) \(\hat{S}^o\) 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 or smaller 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 \(x_j\) are censored. The obvious bootstrap method in this case is to compute \(\hat{S}^o\) as at (2.4),
choose \( X_1^*, X_2^*, \ldots, X_n^* \) iid \( S^o \), and define \( X_i^* = \min\{X_i^o, w_i\} \), with 
\( D_i^* \) equaling 1 or 0 as \( X_i^* \) equals \( X_i^o \) or \( w_i \). This is not the same as sampling \( n \) times with replacement for \( \{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\} \), but the Monte Carlo results in Table 2, at the end of Section 3, suggest that the numerical results may be very similar.


We now use the bootstrap method to derive an estimate for the standard deviation of \( \hat{S}^o(t) \) 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

\[
m_j^* = \# of\ times\ (x_j^*, d_j)\ appears\ in\ the\ bootstrap\ sample, \quad (3.1)
\]

so \( m^* = (m_1^*, m_2^*, \ldots, m_n^*) \) is an \( n \)-category multinomial, \( n \) draws, probability \( 1/n \) for each category \( m_j^* \sim \text{Mult}(n, 1/n) \). For example, \( m_1^* \) might equal 2, in which case 2 of the \( n \) pairs \( (X_1^*, D_1^*) \) 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

\[
M_j^* = \sum_{i=j}^{n} m_i^*, \quad j=1, 2, \ldots, n, \quad (3.2)
\]

so \( M_1^* = n, M_2^* = n-m_1^* \), etc.

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

\[
\hat{S}^o(t) = \prod_{j=1}^{k_t} \left( 1 - \frac{m_j^*}{M_j^*} \right)^{d_j^*}. \quad (3.3)
\]
(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 $\hat{S}^o(t)$, $t$ fixed, is

$$\hat{\sigma}_{\text{BOOT}} = \sqrt{\text{Var}_* \hat{S}^o(t)}, \quad (3.4)$$

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

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

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

see Kaplan and Meier (1958). Table 1 compares $\hat{\sigma}_{\text{GREEN}}$ with $\hat{\sigma}_{\text{BOOT}}$ for the Channing House data, at 9 different values of $t$: $x_{10}', x_{20}', ..., x_{90}'$ 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.
<table>
<thead>
<tr>
<th>t</th>
<th>895</th>
<th>936</th>
<th>957</th>
<th>973</th>
<th>1001</th>
<th>1016</th>
<th>1033</th>
<th>1058</th>
<th>1098</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x_{10}</td>
<td>x_{20}</td>
<td>x_{30}</td>
<td>x_{40}</td>
<td>x_{50}</td>
<td>x_{60}</td>
<td>x_{70}</td>
<td>x_{80}</td>
<td>x_{90}</td>
</tr>
<tr>
<td>\hat{S}^o(t)</td>
<td>.927</td>
<td>.862</td>
<td>.838</td>
<td>.786</td>
<td>.701</td>
<td>.653</td>
<td>.555</td>
<td>.442</td>
<td>.279</td>
</tr>
<tr>
<td>\hat{GO}_{GREEN}</td>
<td>.027</td>
<td>.036</td>
<td>.038</td>
<td>.044</td>
<td>.051</td>
<td>.055</td>
<td>.062</td>
<td>.067</td>
<td>.073</td>
</tr>
<tr>
<td>\hat{GO}_{BOOT}</td>
<td>.025</td>
<td>.036</td>
<td>.039</td>
<td>.044</td>
<td>.052</td>
<td>.057</td>
<td>.064</td>
<td>.069</td>
<td>.076</td>
</tr>
</tbody>
</table>

Estimates of \( \hat{g} \) based on the bootstrap percentiles:

| 75\%–25\% | \frac{1.35}{1.35} | 0.024 | 0.036 | 0.040 | 0.045 | 0.054 | 0.061 | 0.066 | 0.069 | 0.074 |
| 90\%–10\% | \frac{2.56}{2.56} | 0.023 | 0.036 | 0.039 | 0.045 | 0.052 | 0.056 | 0.066 | 0.072 | 0.074 |

Percentiles of the bootstrap distribution of \( \hat{S}^o(t) \):

<table>
<thead>
<tr>
<th>Percentile</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.895</td>
<td>.910</td>
<td>.927</td>
<td>.942</td>
<td>.955</td>
</tr>
<tr>
<td></td>
<td>.812</td>
<td>.839</td>
<td>.865</td>
<td>.888</td>
<td>.904</td>
</tr>
<tr>
<td></td>
<td>.790</td>
<td>.813</td>
<td>.841</td>
<td>.867</td>
<td>.890</td>
</tr>
<tr>
<td></td>
<td>.731</td>
<td>.758</td>
<td>.788</td>
<td>.819</td>
<td>.845</td>
</tr>
<tr>
<td></td>
<td>.636</td>
<td>.668</td>
<td>.700</td>
<td>.741</td>
<td>.768</td>
</tr>
<tr>
<td></td>
<td>.582</td>
<td>.613</td>
<td>.651</td>
<td>.695</td>
<td>.725</td>
</tr>
<tr>
<td></td>
<td>.471</td>
<td>.509</td>
<td>.548</td>
<td>.598</td>
<td>.641</td>
</tr>
<tr>
<td></td>
<td>.357</td>
<td>.396</td>
<td>.446</td>
<td>.489</td>
<td>.541</td>
</tr>
<tr>
<td></td>
<td>.183</td>
<td>.233</td>
<td>.284</td>
<td>.332</td>
<td>.373</td>
</tr>
</tbody>
</table>

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.
than \( \hat{G}_{\text{GREEN}} \), but also a lot less analysis of the specific problem. In essence, we have 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 percentile of the 400 bootstrap values of \( \hat{S}^*(t) \), for each value of \( t \), and normal theory estimates of the standard deviation based on these percentiles. One might prefer such a definition of "\( \hat{G}_{\text{BOOT}} \)" if outlying values of \( \hat{S}^* \) 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 \( \hat{G}_{\text{FIXED}} \), the bootstrap estimate based on fixed censoring, with \( \hat{G}_{\text{BOOT}} \) and \( \hat{G}_{\text{GREEN}} \) for the following situation: \( n = 20; X^o_i \) a standard exponential random variable \((S^o(t) = e^{-t})\); \( w_1, w_2, \ldots, w_{10} = .693; w_4, w_{12}, \ldots, w_{20} = \infty \). Here half of the sample is potentially censored at the true median, and half is totally uncensored. Ten Monte Carlo trials were run, \( N = 400 \) bootstrap replications per trial. We see close agreement between \( \hat{G}_{\text{FIXED}}, \hat{G}_{\text{BOOT}}, \) and \( \hat{G}_{\text{GREEN}} \). This same reassuring agreement was observed in other Monte Carlo trials, with the \( w_i \) values chosen differently.

\[ \uparrow \] Notice that there are two levels of Monte Carlo sampling involved: "trial" refers to a new drawing of \( X_1^o, X_2^o, \ldots, X_{20}^o \), while "replication" refers to the bootstrap sampling, with the data \( \{(x_{1},d_1), \ldots, (x_{20},d_{20})\} \) held fixed.
<table>
<thead>
<tr>
<th>Trial</th>
<th>$\hat{\sigma}^{}_{\text{BOOT}}$</th>
<th>$\hat{\sigma}^{}_{\text{FIXED}}$</th>
<th>$\hat{\sigma}^{}_{\text{GREEN}}$</th>
<th>$\hat{\sigma}^{}_{\text{BOOT}}$</th>
<th>$\hat{\sigma}^{}_{\text{FIXED}}$</th>
<th>$\hat{\sigma}^{}_{\text{GREEN}}$</th>
<th>$\hat{\sigma}^{}_{\text{BOOT}}$</th>
<th>$\hat{\sigma}^{}_{\text{FIXED}}$</th>
<th>$\hat{\sigma}^{}_{\text{GREEN}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.100</td>
<td>.092</td>
<td>.097</td>
<td>.114</td>
<td>.111</td>
<td>.111</td>
<td>.114</td>
<td>.111</td>
<td>.111</td>
</tr>
<tr>
<td>2</td>
<td>.090</td>
<td>.095</td>
<td>.097</td>
<td>.101</td>
<td>.114</td>
<td>.110</td>
<td>.135</td>
<td>.127</td>
<td>.129</td>
</tr>
<tr>
<td>3</td>
<td>.097</td>
<td>.104</td>
<td>.097</td>
<td>.114</td>
<td>.117</td>
<td>.112</td>
<td>.095</td>
<td>.101</td>
<td>.097</td>
</tr>
<tr>
<td>4</td>
<td>.094</td>
<td>.092</td>
<td>.097</td>
<td>.114</td>
<td>.110</td>
<td>.112</td>
<td>.112</td>
<td>.110</td>
<td>.110</td>
</tr>
<tr>
<td>5</td>
<td>.096</td>
<td>.097</td>
<td>.097</td>
<td>.096</td>
<td>.097</td>
<td>.097</td>
<td>.135</td>
<td>.154</td>
<td>.135</td>
</tr>
<tr>
<td>6</td>
<td>.096</td>
<td>.098</td>
<td>.097</td>
<td>.107</td>
<td>.107</td>
<td>.111</td>
<td>.128</td>
<td>.137</td>
<td>.123</td>
</tr>
<tr>
<td>7</td>
<td>.098</td>
<td>.097</td>
<td>.097</td>
<td>.112</td>
<td>.110</td>
<td>.111</td>
<td>.129</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td>8</td>
<td>.098</td>
<td>.093</td>
<td>.097</td>
<td>.116</td>
<td>.104</td>
<td>.112</td>
<td>.114</td>
<td>.103</td>
<td>.111</td>
</tr>
<tr>
<td>9</td>
<td>.095</td>
<td>.096</td>
<td>.097</td>
<td>.113</td>
<td>.115</td>
<td>.112</td>
<td>.113</td>
<td>.113</td>
<td>.111</td>
</tr>
<tr>
<td>10</td>
<td>.093</td>
<td>.099</td>
<td>.097</td>
<td>.109</td>
<td>.111</td>
<td>.112</td>
<td>.104</td>
<td>.095</td>
<td>.102</td>
</tr>
</tbody>
</table>

Table 2. 10 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, $\hat{\sigma}^{}_{\text{FIXED}}$, close to $\hat{\sigma}^{}_{\text{BOOT}}$ and $\hat{\sigma}^{}_{\text{GREEN}}$. 

Proof that \( \hat{\sigma}_{\text{GREEN}} \approx \hat{\sigma}_{\text{BOOT}} \). From (3.3), \( \log \hat{S}^*(t) = \sum_{j=1}^{k} d_j \log(1-m_j^*/M_j^*) \).
We have \( M_j^* \sim \text{Bi}(n, n_j/n) \) and \( m_j^*/M_j^* \sim \text{Bi}(M_j^*, 1/n_j) \), where \( n_j = n-j+1 \).
For \( n_j \) large, standard "delta theory" calculations give the approximation \( \text{Var}_\delta(\log \hat{S}^*(t)) = \sum_{j=1}^{k} d_j / n_j^2 \), and so \( \text{Var}_\delta \hat{S}^*(t) = (\hat{S}^*(t))^2 \sum_{j=1}^{k} d_j / n_j^2 \approx \hat{\sigma}_{\text{GREEN}}^2 \).

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 \( \hat{\sigma}_{\text{GREEN}} \) and \( \hat{\sigma}_{\text{BOOT}} \), even at \( x_{90} \) (\( n_j = 8 \)).

4. Location Estimates.

The median is often favored as a location estimate in censored data problems, because it least depends upon the right tail of the Kaplan-Meier curve \( \hat{S}^*(t) \), 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 \( \hat{S}^* \).) The estimated standard deviation is \( \hat{\sigma}_{\text{BOOT}} \), 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, \( \frac{1}{N} \sum_{j=1}^{N} (\hat{\delta}^*_j) - \hat{\delta} \), a quantity closely related to the jackknife estimate of bias, Efron (1979). The same \( N = 1600 \) bootstrap replications were used for all seven estimators.
<table>
<thead>
<tr>
<th>Statistic</th>
<th>Observed Value</th>
<th>Bias</th>
<th>( \hat{\sigma} )_{BOOT}</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
<th>95%</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Median</td>
<td>1044.0</td>
<td>4.1</td>
<td>14.0</td>
<td>1029.5</td>
<td>32.0</td>
<td>37.5</td>
<td>46.0</td>
<td>55.5</td>
<td>60.0</td>
<td>80.0</td>
<td>13.9</td>
<td>10.9</td>
<td>15.5</td>
<td>2.06</td>
</tr>
<tr>
<td>2. .25 trimmed (each tail)</td>
<td>1047.4</td>
<td>1.8</td>
<td>12.5</td>
<td>1028.1</td>
<td>32.8</td>
<td>39.5</td>
<td>47.6</td>
<td>56.4</td>
<td>65.3</td>
<td>70.0</td>
<td>12.5</td>
<td>12.7</td>
<td>12.7</td>
<td>1.15</td>
</tr>
<tr>
<td>3. .10 trimmed (each tail)</td>
<td>1045.8</td>
<td>0.6</td>
<td>11.4</td>
<td>1026.7</td>
<td>30.4</td>
<td>37.6</td>
<td>45.3</td>
<td>52.7</td>
<td>60.4</td>
<td>64.2</td>
<td>11.2</td>
<td>11.7</td>
<td>11.4</td>
<td>1.02</td>
</tr>
<tr>
<td>4. .05 trimmed (each tail)</td>
<td>1043.2</td>
<td>0.4</td>
<td>10.8</td>
<td>1025.1</td>
<td>28.5</td>
<td>35.2</td>
<td>42.5</td>
<td>49.8</td>
<td>56.7</td>
<td>61.0</td>
<td>10.8</td>
<td>11.0</td>
<td>10.9</td>
<td>1.06</td>
</tr>
<tr>
<td>5. .25 Winsorized (each tail)</td>
<td>1044.5</td>
<td>4.5</td>
<td>12.5</td>
<td>1028.6</td>
<td>32.5</td>
<td>38.8</td>
<td>47.4</td>
<td>56.9</td>
<td>64.2</td>
<td>68.8</td>
<td>13.4</td>
<td>12.4</td>
<td>12.2</td>
<td>1.14</td>
</tr>
<tr>
<td>6. .10 Winsorized (each tail)</td>
<td>1041.4</td>
<td>1.4</td>
<td>10.7</td>
<td>1025.0</td>
<td>28.3</td>
<td>34.7</td>
<td>41.6</td>
<td>49.2</td>
<td>55.8</td>
<td>59.9</td>
<td>10.8</td>
<td>10.7</td>
<td>10.6</td>
<td>1.10</td>
</tr>
<tr>
<td>7. Mean</td>
<td>1038.2</td>
<td>0.4</td>
<td>10.5</td>
<td>1021.1</td>
<td>24.3</td>
<td>30.8</td>
<td>37.6</td>
<td>44.8</td>
<td>51.4</td>
<td>54.6</td>
<td>10.4</td>
<td>10.6</td>
<td>10.2</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 3. Seven location estimates for the Channing House Data. Bias and standard deviation estimates are based on \( N = 1600 \) 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%-0.05%)/3.290, (d) Asymmetry Ratio (95%-50%)/(50%-5%).
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 tends 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% larger than the corresponding intervals for the mean.

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 \( \theta \), 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. Some theoretical rationale for this method is given in Section 6.

Suppose then that we have computed the c.d.f. of the bootstrap distribution for some real-valued parameter \( \theta \), estimated by \( \hat{\theta} = \theta(\hat{F}) \), say
\[ \text{CDF}_\star(t) \equiv \text{Prob}_\star\{\hat{\theta}^* \leq t\}. \quad (5.1) \]

For a given \( \alpha \), we can construct what we hope is a 1-2\( \alpha \) central confidence interval by using the appropriate percentiles of \( \text{CDF}_\star \), say

\[ \hat{\theta}_\text{LOW} \equiv \text{CDF}_\star^{-1}(\alpha), \quad \hat{\theta}_\text{UP} \equiv \text{CDF}_\star^{-1}(1-\alpha), \quad (5.2) \]

as the lower and upper points of the interval. Section 6 gives some theoretical arguments for believing that \( [\hat{\theta}_\text{LOW}, \hat{\theta}_\text{UP}] \) 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\% central confidence interval for the mean, in the Channing House data.\(^\dagger\)

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 c.d.f.: for \( t \) equal to an uncensored observed lifetime, \( \text{CDF}_\star(t) = \text{Prob}_\star\{\hat{\theta}^* < t\} + \frac{1}{2} \text{Prob}_\star(\hat{\theta}^* = t) \). Linear interpolation is used to fill in the values

\(^\dagger\)Most often we will be working with an approximation to \( \text{CDF}_\star \), based on some number \( N \) of Monte Carlo trials, \( N = 1600 \) in this case. Empirically, the author has found such large values of \( N \) necessary to numerically stabilize the extreme percentiles. For only estimating \( \hat{\theta}_\text{BOOT} \), \( N = 100 \) performs quite well in most examples.
of $\text{CDF}_k(t)$ between the probability mass points, as shown in Figure 2. The Bayesian argument of Section 6 gives some support to "splitting the atoms" in this situation.

Figure 2. Bootstrap distribution for the sample median, Channing House data, based on $N = 1600$ bootstrap replications. The distribution is supported on the uncensored observed lifetimes. The smoothed version of $\text{CDF}_k$ described in the text is also shown.
Several Monte Carlo experiments were run to ascertain the actual coverage probabilities of the percentile method intervals for the median. In these experiments both $X_1^*$ and $W_1$ had exponential distributions, 

$\text{Prob}(X_1^* > t) = e^{-t}$, $\text{Prob}(W_1 > t) = e^{-t/c}$, $t \geq 0$. The probability of $X_1 = \min(X_1^*, W_1)$ being uncensored (that is, $D_1 = 1$) equals $\frac{c}{c+1}$. 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 $\text{Prob}(X_1 \text{ uncensored}) = .5$), and $\alpha=.10$ the first experiment had $\hat{\theta}_{\text{LOW}} < \theta$ 92% of the time, compared to a nominal 90%.

[Techical note: it is equivalent to generate the $X_1$ according to $\text{Prob}(X_1 > t) = \exp(-(1 + \frac{1}{c})t)$, $t > 0$, and then $D_1 = 1$ (or 0) with probability $c/(c+1)$ (or $1/(c+1)$), independently of $X_1 = x_1$. For the first experiment, "500 x 400" no. of trials refers to 500 realizations of $x_1, x_2, \ldots, x_{25}$, and for each of these 400 independent realizations of $d_1, d_2, \ldots, d_{25}$.

There is a nice connection\footnote{I am grateful to Kent Bailey for suggesting this connection to me. He points out that one can use Greenwood's formula and normality to approximate the percentile points in Table 1, and thereby construct an approximate interval for the median using the method of this paragraph, avoiding the bootstrap computations.} between Table 1 and the percentile method intervals for the median. For any fixed value of $t$ we could use the percentile method to construct a $1-2\alpha$ central confidence interval for $S^\circ(t)$. (For example, with $t = 1001 = x_{50}$, $\alpha = .10$, Table 1 gives the
<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>Sample Size n</th>
<th>Prob(Un-censored)</th>
<th>Actual coverage probabilities at nominal (lower) level</th>
<th># Bootreps per trial, N</th>
<th># Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>25</td>
<td>.50</td>
<td>.92, .76, .47, .22, .08</td>
<td>400</td>
<td>500 × 400</td>
</tr>
<tr>
<td>2.</td>
<td>25</td>
<td>.50</td>
<td>.93, .78, .49, .21, .08</td>
<td>200</td>
<td>200 × 100</td>
</tr>
<tr>
<td>3.</td>
<td>25</td>
<td>.50</td>
<td>.91, .73, .45, .20, .07</td>
<td>100</td>
<td>1000 × 100</td>
</tr>
<tr>
<td>4.</td>
<td>21</td>
<td>.67</td>
<td>.91, .76, .48, .21, .08</td>
<td>1000</td>
<td>1000 × 200</td>
</tr>
<tr>
<td>5.</td>
<td>21</td>
<td>.67</td>
<td>.90, .75, .47, .21, .08</td>
<td>400</td>
<td>500 × 400</td>
</tr>
<tr>
<td>6.</td>
<td>21</td>
<td>.67</td>
<td>.91, .75, .49, .22, .08</td>
<td>200</td>
<td>200 × 50</td>
</tr>
<tr>
<td>7.</td>
<td>21</td>
<td>.67</td>
<td>.90, .74, .47, .21, .08</td>
<td>100</td>
<td>1000 × 100</td>
</tr>
<tr>
<td>8.</td>
<td>21</td>
<td>.67</td>
<td>.90, .75, .48, .23, .09</td>
<td>50</td>
<td>400 × 50</td>
</tr>
</tbody>
</table>

Table 4. Eight Monte Carlo experiments comparing nominal and actual coverage 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.
the lower point of such an interval to be \( \hat{S}_{LOW}^*(t) = .768 \). But for any bootstrap realization \( \hat{S}_{LOW}^*(\cdot) \), we have \( \{\hat{S}_{LOW}^*(t) < .5\} \) iff \( \{\hat{\theta}^* < t\} \), where \( \hat{\theta}^* \) is the bootstrap median \( \hat{S}_{LOW}^*(\hat{\theta}^*) = .5 \). Therefore \( \hat{\theta}_{LOW} = \text{CDF}^{-1}(\alpha) \) is also that value of \( t \) for which \( \hat{S}_{LOW}^*(t) = .5 \); in other words, it is the smallest value of \( t \) for which the 1-2\( \alpha \) percentile method interval for \( S^*(t) \) includes the value .5.


We present three theoretical arguments supporting the percentile method for constructing confidence intervals, as described in Section 5, along with some criticism of this method. The theoretical arguments have nothing in particular to do with censored data. We can assume that the sample space \( \mathcal{X} \), in which the observations \( X_1, X_2, \ldots, X_n \) iid \( \sim F \) take their values, in anything at all, and that we are trying to construct a confidence interval for an arbitrary real-valued parameter \( \theta(F) \). (This includes the case of randomly censored data, for which the individual observations "\( X_i \)" are actually of the form \( (X_i, D_i) \), as explained in Section 2.)

I. A Bayesian Argument. For convenience, suppose that \( \mathcal{X} \) can be partitioned into \( L \) disjoint pieces, say \( \mathcal{X} = \bigcup_{\lambda=1}^{L} \mathcal{X}_\lambda \), and that we only observe into which \( \mathcal{X}_\lambda \) each \( X_i \) falls. In fact, this is no real restriction since \( L \) can be made arbitrarily large. If \( \mathcal{X} \) equals the real line, for instance, we might partition the interval \([-10^{10}, 10^{10}]\) into \( 2 \cdot 10^{20} \) subintervals of length \( 10^{-10} \). Then \( L = 2 \cdot 10^{20} + 2 \), counting the semi-infinite end intervals, and for most practical purposes the discretization will have no effect on our inferences.
The probability distribution $F$ on $X$ can be represented by a vector of probabilities $\tilde{f} = (f_1, f_2, \ldots, f_L)$, $f_L = \text{Prob}_F(X_1 \in \mathcal{X}_L)$. We take the prior distribution on $\tilde{f}$ to be a symmetric Dirichlet distribution, $\tilde{f} \sim \text{Di}(\alpha \tilde{f})$, i.e. the prior density of $\tilde{f}$ is taken proportional to $\prod f_{\mathcal{X}}^{\alpha - 1}$. Having observed $X_1, X_2, \ldots, X_n$, let $\hat{f}_L = \#\{X_1 \in \mathcal{X}_L\}/n$, $\hat{f} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_L)$. A posteriori we have $\tilde{f} | \hat{f} \sim \text{Di}(\alpha \tilde{f} + n\hat{f})$, and letting $\alpha \to 0$, to represent prior ignorance, we get the well known result

$$\tilde{f} | \hat{f} \sim \text{Di}(n\hat{f}).$$ \hspace{1cm} (6.1)

The bootstrap, applied to the discretized situation, selects $X_1^*, X_2^*, \ldots, X_n^*$ independently, $\text{Prob}_*\{X_1^* \in \mathcal{X}_L\} = \hat{f}_L$. Letting $\hat{f}_L^* = \#\{X_i^* \in \mathcal{X}_L\}/n$, $\hat{f}^* = (\hat{f}_1^*, \hat{f}_2^*, \ldots, \hat{f}_L^*)$,

$$\hat{f}^* | \hat{f} \sim \text{Mult}(n, \hat{f})/n.$$ \hspace{1cm} (6.2)

(See Remark G, Efron (1979).)

Distributions (6.1) and (6.2) are quite similar. Each is supported entirely on the nonzero components of $\hat{f}$. Given $\hat{f}$, each has conditional expectation $\hat{f}$. The conditional covariance matrices are also nearly equal, $\text{Cov}(\tilde{f} | \hat{f}) = \frac{\hat{f}\tilde{f}}{n+1}$, $\text{Cov}\hat{f}^* | \hat{f}) = \frac{\hat{f}\tilde{f}}{n}$, where $\frac{\hat{f}\tilde{f}}{n}$ has diagonal elements $\hat{f}_L^*(1-\hat{f}_L^*)$ and off-diagonal elements $-\hat{f}_L^*\hat{f}_m^*$. The point here is that the posterior distribution of $\theta(\tilde{f} | \hat{f})$ is likely to be well approximated by the bootstrap distribution $\theta(\hat{f}^* | \hat{f})$, if $\theta(\tilde{f})$ is any reasonably smooth function of $\tilde{f}$. If this is the case, the percentile method central confidence interval, level 1-2$\alpha$, will be a good approximation to the central Bayes interval of probability 1-2$\alpha$. 

22
The prior distribution \( Di(\alpha_1^{(x)}) \), \( \alpha \to 0 \), may seem unreasonable\(^{+}\), but for the specific case where \( X \) is the real line and \( \theta(F) \) is the median (and we let the discretization become infinitely fine) it gives a reasonable result: if \( x(z) \) is the \( z \)th ordered value of the observations, then
\[
\text{Prob}\{\theta(F) > x(z) \mid \hat{F}\} = \text{Prob}\{\text{Binomial}(n-1, \frac{1}{2}) \geq z\}.
\]
This is almost the same as the standard frequentist result for the median, \( \text{Prob}\{x(z) < \theta(F)\} = \text{Prob}\{\text{Binomial}(n, \frac{1}{2}) \geq z\} \), based on the pivotal quantity \( Z = \#\{X_i < \theta\} \).
The argument above then implies that the percentile method should give nearly the usual frequentist confidence intervals for the median. This is indeed the case, see Remark D, Efron (1979). (Notice that (6.1) gives a continuous distribution for \( \theta(F) \mid \hat{F} \), while (6.2) gives \( \theta(F) \mid \hat{F} \) a discrete distribution supported on the observed values \( x_1, x_2, \ldots, x_n \).

The continuity correction made in "Remark D" improves the agreement between \( \theta(F) \mid \hat{F} \) and \( \theta(F^*) \mid \hat{F} \) in an obvious way.)

II. Transformation to a Symmetric Pivotal Quantity. Suppose we try to use \( \hat{\theta} - \theta \) as a pivotal quantity; that is, we assume that its distribution is the same no matter what distribution \( F \) the \( X_i \) are sampled from. In particular, taking \( F = \hat{F} \), this implies \( \hat{\theta} - \theta \sim \hat{\theta} - \theta \). Then for any \( z_1, z_2 \), we have
\[
\text{Prob}_x(\theta \in [\hat{\theta} - z_1, \hat{\theta} + z_2]) = \text{Prob}_F(\theta \in [\theta - z_1, \theta + z_2])
\]
\[
= \text{Prob}_F(\theta \in [\hat{\theta} - z_2, \hat{\theta} + z_1]).
\]
Choosing \( z_1 = \hat{\theta} - \hat{\theta}_{LOW}, z_2 = \hat{\theta}_{UP} - \hat{\theta} \) gives central \( 1 - 2\alpha \) confidence interval \( \theta \in [2\hat{\theta}_{UP}, 2\hat{\theta}_{LOW}] \), in notation (5.2). This is the reflection of the percentile interval \( [\hat{\theta}_{LOW}, \hat{\theta}_{UP}] \) about the point \( \hat{\theta} \). (See Remark D, Efron (1979).)

\(^{+}\)In a recent unpublished paper, Rubin (1979), D. Rubin criticizes the bootstrap on just these grounds, suggesting that it would be better to do the full Bayesian analysis using a more sensible prior.
In many parametric problems it is not \( \hat{\theta} - \theta \) which is pivotal, but rather \( \hat{\phi} - \phi \), where \( \phi = g(\theta) \), \( \hat{\phi} = g(\hat{\theta}) \), \( g(\cdot) \) being some increasing function. [For example, if \( \theta \) is the Pearson correlation coefficient, then Fisher's transformation \( \phi = \tanh^{-1}(\theta) \) makes \( \hat{\phi} - \phi \) approximately \( N(\phi, (n-3)^{-1}) \) if \( F \) is bivariate normal. (Cramer, 1946, p. 399.).] If, for some \( g(\cdot) \), we are willing to assume \( \hat{\phi} - \phi \) pivotal, then the 1-2\( \alpha \) confidence interval is \( \phi \in [2\hat{\phi}_{UP}^*, 2\hat{\phi}_{LOW}^*] \) where \( \hat{\phi}_{LOW} \) is the \( \alpha \) percentile point of \( \hat{\phi}^* \)'s bootstrap distribution, and \( \hat{\phi}_{UP} \) the 1-\( \alpha \) percentile point. Using the fact that the bootstrap distribution of \( \hat{\phi}^* = g(\hat{\theta}^* \) is the obvious transformation of the bootstrap distribution of \( \hat{\theta}^* \), \( \text{Prob}_{*}\{ \hat{\phi}^* \leq g(t) \} = \text{Prob}_{*}\{ \hat{\theta}^* \leq t \} \), we have \( \hat{\phi}_{LOW} = g(\hat{\theta}_{LOW}) \), \( \hat{\phi}_{UP} = g(\hat{\theta}_{UP}) \).

Finally, suppose that we can choose \( g(\cdot) \) to make \( \hat{\phi} - \phi \) both pivotal and symmetrically distributed about zero. (This includes the case \( \hat{\phi} - \phi \sim N(0, \sigma^2) \), \( \sigma^2 \) not depending on \( \phi \).) Then \( 2\hat{\phi}_{UP} = \hat{\phi}_{LOW} \), \( 2\hat{\phi}_{LOW} = \hat{\phi}_{UP} \), since \( \hat{\phi}_{UP} - \hat{\phi} = \hat{\phi} - \hat{\phi}_{LOW} \), giving the central 1-2\( \alpha \) confidence interval \( \phi \in [\hat{\phi}_{LOW}, \hat{\phi}_{UP}] \). Transforming back to the \( \theta \)-scale gives the percentile method confidence interval \( \theta \in [\hat{\theta}_{LOW}, \hat{\theta}_{UP}] \).

The trouble with this argument is that there is no reason to believe that \( g(\cdot) \) exists. A necessary condition is that the bootstrap distribution of \( \hat{\theta}^* \) have its median at \( \hat{\theta} \). Otherwise \( \hat{\phi}^* \) cannot be symmetrically distributed about \( \hat{\phi} \), since \( \text{Prob}_{*}\{ \hat{\phi}^* < \hat{\phi} \} = \text{Prob}_{*}\{ \hat{\theta}^* < \hat{\theta} \} \) and \( \text{Prob}_{*}\{ \hat{\phi}^* > \hat{\phi} \} = \text{Prob}_{*}\{ \hat{\theta}^* > \hat{\theta} \} \). This necessary condition is satisfied when \( \hat{\theta} \) is the sample median and we are dealing with uncensored data, as in Section 3 of Efron (1979). Roughly speaking, it is satisfied for
the sample median of the Channing House data, as you can see by adding up
the bootstrap probabilities in Figure 2. If the necessary condition is not
satisfied, but we assume we can transform so that \( \hat{\phi}^* - \hat{\phi} \) has an approximately
normal bootstrap distribution not centered at zero, then the transformation
argument gives a different confidence interval for \( \theta \),

\[
\theta \in \left[ \operatorname{CDF}^{-1}(\phi(2z_0 - z_\alpha)), \quad \operatorname{CDF}^{-1}(\phi(2z_0 + z_\alpha)) \right],
\]

(6.3)

where \( \phi \) is the standard normal cumulative c.d.f., \( z_0 = \phi^{-1}(\operatorname{CDF}(\hat{\theta})) \), and
\( \phi(z_\alpha) = 1 - \alpha \). The centered case, \( z_0 = 0 \), in which \( \hat{\theta}^* \) is median unbiased
for \( \hat{\theta} \), gets us back to the percentile method interval. In other words,
(6.3) is a bias-corrected version of the percentile method.

Seemingly small deviations of \( z_0 \) from 0 can make the interval
(6.3) much different than the percentile method interval. As an example,
consider the 15 pairs of numbers discussed in Section 2 of Efron (1979b):
(576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07),
(661, 3.43), (651, 3.36), (605, 3.13), (653, 3.12), (5.75, 2.74), (545, 2.76),
(572, 2.88), (594, 2.96). The Pearson correlation coefficient is \( \hat{\rho} = .78; \)
N = 10000 bootstrap replications gave central 68\% confidence interval
\( [\operatorname{CDF}^{-1}(-1), \operatorname{CDF}^{-1}(1)] = [\hat{\rho} - .12, \hat{\rho} + .13] \) by the percentile method.
However \( \operatorname{CDF}(\hat{\theta}) = .433 \), so \( z_0 = -.17 \) and the central 68\% interval (6.3)
equals \( [\operatorname{CDF}^{-1}(-1.34), \operatorname{CDF}^{-1}(1.66)] = [\hat{\rho} - .17, \hat{\rho} + .10] \). This is much
closer to the normal theory interval \( [\hat{\rho} - .16, \hat{\rho} + .09] \) based on
inverting the bias-corrected approximation \( \tanh^{-1}(\hat{\rho}) \sim h(\tanh^{-1}(\rho) + \frac{\rho}{2(n+1)}), \frac{1}{n-3}). \)
III. Typical Value Theory. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a translation family on the real line, with density function $f(x-\theta)$, $f(\cdot)$ symmetric about 0. Let $\hat{\theta}$ be an $m$-estimator based on a monotone kernel, i.e. $\hat{\theta}$ is the solution of $\sum_{i=1}^{n} \psi(x_i - \theta) = 0$, where $\psi(\cdot)$ is anti-symmetric about 0, continuous, and strictly increasing. ($\psi(x) = x$ gives $\hat{\theta} = \bar{x}$.) Then Hartigan (1969), using suggestions of Tukey and Mallows, has developed an impressive theory of subsampling, including the "typical value theorem": for any nonempty subset $S$ of $\{1, 2, \ldots, n\}$, let $\hat{\theta}_S$ be the solution of $\sum_{i \in S} \psi(x_i - \theta) = 0$; also let $Y = \#\{S: \hat{\theta}_S < \theta\} + 1$; then $Y$ is a pivotal quantity,

$$\text{Prob}(Y = y) = \frac{1}{2^n}, \quad y = 1, 2, \ldots, 2^n. \quad (6.4)$$

(See also Forsythe and Hartigan (1970), and Maritz (1979).)

The typical value $1 - 2\alpha$ central confidence interval for $\theta$ is $[\hat{\theta}(a), \hat{\theta}(b)]$, where $\hat{\theta}(a)$ is the $a$th ordered value among $\{\hat{\theta}_S\}$ and $a/2^n = \alpha$, likewise $b/2^n = 1 - \alpha$. In practice, a random selection of $N$ of the $2^n - 1$ possible subsets will be used to construct the interval, just as with the bootstrap intervals. The bias correction (6.3) is also appropriate here.

The median is not included under the auspices of the typical value theorem, but a similar result can be developed. Let $Z = \#\{X_i < \theta\}$, so $Z$ is pivotal,

$$\text{Prob}(Z = z) = \binom{n}{z}/2^n, \quad z = 0, 1, 2, \ldots, n. \quad (6.5)$$

If $S$ is a subset of $\{1, 2, \ldots, n\}$ having an odd number of members, then the sample median $\hat{\theta}_S$ is uniquely defined as the middle value, but
if $S$ is even then $\hat{S}$ can be assigned any value between the two middle order statistics. The means that give the data $(x_1, x_2, \ldots, x_n)$, 

$Y = \#\{S: \hat{S} < \theta\} + 1$ can take on a range of values depending on how the even cases are defined. The connection with the typical value theorem is as follows:

**Theorem.** The event \{Z=z\} is equivalent to the event \{$\sum_{j=0}^{n-1} \binom{n}{j} < Y < \sum_{j=0}^{n} \binom{n}{j}$\}, where $Y$ can take on all integer values in the indicated interval, depending on how the $\hat{S}_S$ are assigned for $S$ even.

(The proof will not be presented here.) Adding up the probabilities, (6.4) shows that in a certain sense the typical value theorem reduces to the usual frequentist result for the median, based on (6.5). Symmetry is not required in the case of the sample median. This is not surprising since if $f(x-\theta)$ is not symmetric to begin with, a monotone transformation will make it symmetric, without changing the value(s) of $Z$ and $Y$.

We have seen that in the case of the sample median, uncensored data, the percentile method intervals are nearly the same as the usual frequentist intervals. The same is true for the typical value intervals, so the percentile and typical value methods give nearly the same results in this case. As a matter of fact, the two methods tend to give similar results in most situations.

To see why, consider choosing a random subsample of the observed data $(x_1, x_2, \ldots, x_n)$ by independently flipping a fair coin to determine whether or not each $x_i$ is included. This gives a random subsample $X_1$$^\star$, $X_2$$^\star$, $X_m$$^\star$, where $m$ is the number of elements selected. (We ignore the case $m=0$ which only has probability $2^{-n}$.) Using the finite sample
space notation of part I, define $\hat{f}_{**} = \#\{X_{**} \in \chi_{\lambda}\}/m$. Given $\hat{f}$, it is easy to show that $\hat{f}_{**} \mid \hat{f}$ is supported entirely on the nonzero components of $\hat{f}$, has conditional mean vector $\hat{\mu}_{\hat{f}}$, and conditional covariance matrix

$$\text{Cov}_{\hat{f}_{**}}(\hat{f}_{**} \mid \hat{f}) = \left( \frac{n+2}{n-1} + 0 \left( \frac{1}{n} \right) \right) \text{Cov}(\hat{f}_{**} \mid \hat{f}) .$$  \hspace{1cm} (6.6)

For smooth functions $\theta(\hat{f})$, we expect the typical value intervals, which are based on the distribution of $\theta(\hat{f}_{**}) \mid \hat{f}$, to be about $\sqrt{\frac{n+2}{n-1}}$ as wide as the percentile intervals, based on $\theta(\hat{f}_{**}) \mid \hat{f}$. An example is given in Table 5. Each of the 10 trials involves an i.i.d. sample $X_1, X_2, \ldots, X_{15}$ from a standard exponential distribution, $\text{Prob}\{X > x\} = e^{-x}$, $x > 0$. The goal is to obtain nonparametric confidence intervals for $\theta$, the expectation, using the bootstrap and subsample distributions of the sample mean. ($N = 1000$ replications for each method.)

Before analysis, each sample $(x_1, x_2, \ldots, x_{15})$ was translated to have $\bar{x} = 0$ and scaled to have $\Sigma(x_i - \bar{x})^2/14 = 1$. This stabilized the entries in Table 5, without affecting comparisons between the different methods. The percentile method and typical value intervals are seen to be quite similar. In this case $\sqrt{\frac{n+2}{n-1}} = 1.10$, and we see that this is just about the ratio of the widths. (For $1 - 2\alpha = .90$ the ratio is $(.46 + .44)/(.43 + .39) = 1.10$.) Both methods are disappointing in one major aspect: the true confidence limits, based on Pitman's method for the parametric location scale family $(1/\sigma)e^{-(x-\mu)/\sigma}$, $x > \mu$, are much more extended toward the right.

The bias correction (6.3) shifts all 10 percentile intervals in the right direction, as seen in columns d of Table 5. That method seems to
deserve further attention. (On the other hand, the typical value intervals were moved very little, and most often in the wrong direction, by application of (6.3) to the typical value distribution. In this example the typical value method appeared insensitive to asymmetry in the observed sample, which suggests that it should be used with caution when distributional asymmetry is a definite possibility.) The author is considering other nonparametric methods which better capture the effects of asymmetry, one promising candidate being Johnson's modified t-statistic, Johnson (1978).
| Sample | a   | b   | c   | d   | a   | b   | c   | d   | a   | b   | c   | d   | a   | b   | c   | d   | a   | b   | c   | d   | a   | b   | c   | d   | a   | b   | c   | d   | Skewness | μ₃ |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|----|
| 1      | -38 | -44 | -31 | -34 | -31 | -34 | -25 | -27 | .34 | .33 | .45 | .38 | .41 | .43 | .60 | .48 | 1.40 |
| 2      | -39 | -47 | -34 | -36 | -34 | -36 | -27 | -27 | .34 | .37 | .48 | .38 | .45 | .46 | .64 | .54 | 1.30 |
| 3      | -44 | -42 | -42 | -42 | -35 | -36 | -34 | -32 | .30 | .36 | .56 | .32 | .40 | .46 | .66 | .41 | 0.15 |
| 4      | -38 | -44 | -30 | -38 | -32 | -35 | -24 | -32 | .33 | .36 | .44 | .33 | .45 | .47 | .58 | .45 | 1.40 |
| 5      | -37 | -42 | -25 | -35 | -32 | -35 | -20 | -28 | .34 | .33 | .37 | .39 | .44 | .46 | .50 | .49 | 1.86 |
| 6      | -37 | -47 | -41 | -34 | -31 | -36 | -33 | -27 | .34 | .36 | .55 | .39 | .44 | .48 | .65 | .50 | 1.04 |
| 7      | -42 | -45 | -40 | -38 | -34 | -36 | -32 | -29 | .31 | .35 | .54 | .34 | .39 | .46 | .65 | .46 | 0.62 |
| 8      | -35 | -42 | -32 | -32 | -30 | -35 | -26 | -27 | .35 | .36 | .46 | .40 | .46 | .48 | .62 | .50 | 1.98 |
| 9      | -40 | -48 | -33 | -38 | -32 | -37 | -27 | -30 | .33 | .34 | .47 | .34 | .43 | .42 | .62 | .45 | 1.02 |
| 10     | -38 | -42 | -32 | -37 | -31 | -36 | -26 | -30 | .32 | .32 | .46 | .33 | .41 | .43 | .61 | .42 | 1.32 |
| AVE    | -39 | -44 | -34 | -36 | -32 | -36 | -27 | -29 | .33 | .35 | .48 | .36 | .43 | .46 | .61 | .47 |       |    |

Table 5. Nonparametric and parametric confidence intervals for the expectation, 10 standardized samples, n = 15, exponential distribution. (a) Percentile method (N = 1000), (b) Typical Value (N = 1000), (c) Pitman interval, exponential scale and location model, (d) Bias-corrected Percentile Interval (6.3).
REFERENCES


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


