SLAC–PUB–4665
STAN. U. TECH RPT–287
June 1988
(M)

BOOTSTRAP CONFIDENCE INTERVALS IN A COMPLEX SITUATION: A SEQUENTIAL PAIRED CLINICAL TRIAL

SALLY C. MORTON

Department of Statistics
and
Stanford Linear Accelerator Center
Stanford University, Stanford, California 94309

ABSTRACT

This paper considers the problem of determining a confidence interval for the difference between two treatments in a simplified sequential paired clinical trial, which is analogous to setting an interval for the drift of a random walk subject to a parabolic stopping boundary. Three bootstrap methods of construction are applied: Efron's accelerated bias-corrected, the DiCiccio-Romano, and the bootstrap-t. The results are compared with a theoretical approximate interval due to Siegmund. Difficulties inherent in the use of these bootstrap methods in a complex situation are illustrated. The DiCiccio-Romano method is shown to be the easiest to apply and to work well.

Key Words: accelerated bias-corrected method; bootstrap; percentile method; random walk; sequential test.

---

*Work supported in part by the Department of Energy, contract DE–AC03–76SF00515 and by the National Science Foundation.

(Presented at the Joint Statistical Meetings, New Orleans, LA. August 22-25, 1988)

1. Introduction

The bootstrap (Efron 1982) is an ideal tool to construct confidence intervals in complex situations where explicit analytical solutions cannot be found. Several methods have been proposed and theoretical comparisons of these methods are available. The purpose of this article is to consider a challenging problem of practical importance in which a theoretical approximate interval has been determined analytically and to compare this interval with various bootstrap method results. Hopefully, this illustration will provide guidelines for choosing among the methods in similar situations. The problem to be examined is the construction of an interval for the drift of a random walk, which arises in a particular paired sequential clinical trial test.

Section 2 reviews the bootstrap intervals to be calculated. Sections 3 and 4 describe the example and the theoretical interval. Sections 5 and 6 consist of a description of the bootstrap procedure and a discussion of the results.

2. Bootstrap Intervals

Five bootstrap intervals will be constructed in this paper and each will be described briefly in this section. For a more thorough discussion, the reader is referred to the surveys by DiCiccio and Romano (1987), Hall (1988), and Tibshirani (1984), and to the particular references given for each interval.

The clinical trial example to be considered later is a one-parameter problem in which the parametric bootstrap sampling procedure may be applied and the bootstrap notation will be explained in this context. We would like to construct a confidence interval for θ which indexes a scalar-parameter family of distributions F_θ. Our interval is based on the observed value $\hat{\theta}$, generally the maximum-likelihood estimate of θ. The sampling distribution of $\hat{\theta}$ under θ is $G_\theta(s) \equiv P_\theta\{\hat{\theta} \leq s\}$. The bootstrap distribution is the distribution of the estimator at $\hat{\theta}$, $G_{\hat{\theta}}(s) \equiv P_{\hat{\theta}}\{\hat{\theta}^* \leq s\}$, where $\hat{\theta}^*$ is obtained by sampling from $F_{\hat{\theta}}$. In

practice, we resample B times from $F_{\hat{\theta}}$ to produce $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ and approximate $G_{\hat{\theta}}$ by the empirical cumulative distribution function

$$\hat{G}_{\hat{\theta}}(s) = \#\{\hat{\theta}_b^* \leq s\}/B \quad . \tag{2.1}$$

We will seek to construct a central $1-2\alpha$ interval $[\theta_{L,\alpha}, \theta_{U,\alpha}]$ whose endpoints have the property that

$$\begin{aligned} G_{\theta_{U,\alpha}}(\hat{\theta}) &= \alpha \text{ , and} \\ G_{\theta_{L,\alpha}}(\hat{\theta}) &= 1-\alpha \quad . \end{aligned} \tag{2.2}$$

Any upper (lower) endpoint that satisfies (2.2) will be called 'exact'. Otherwise, as in the terminology of Efron (1987), a proposed upper endpoint θ_U will be called 'i^{th}-order correct' if

$$| \theta_U - \theta_{U,\alpha} | = O_p(n^{-(i+1)/2}) \quad , \tag{2.3}$$

and similarly for a proposed lower endpoint. The two-sided coverage error for an interval whose endpoints satisfy (2.3) is typically $O(n^{-i/2})$.

2.1 The Percentile, BC, and BC_a Intervals

The popular standard interval $[\hat{\theta} - \hat{\sigma}z^{(1-\alpha)}, \hat{\theta} - \hat{\sigma}z^{(\alpha)}]$ where $z^{(\alpha)}$ is the α^{th} percentile point of the $N(0,1)$ distribution, relies on the assumption that the statistic is normal with a constant variance. Efron's (1987) BC_a interval is based on the more general assumption that for some monotone transformation g, bias constant z_0, and acceleration constant a, the following is true:

$$g(\hat{\theta}) - g(\theta) \sim N(-z_0(1+ag(\theta)), (1+ag(\theta))^2) \quad . \tag{2.4}$$

That is, there exists a transformation which normalizes the statistic and linearizes, not necessarily stabilizes, the variance. The resulting interval $[\theta_L, \theta_U]$ is

$$[G_{\hat{\theta}}^{-1}(\Phi(z[\alpha])), G_{\hat{\theta}}^{-1}(\Phi(z[1-\alpha]))] \quad , \tag{2.5}$$

where Φ is the $N(0,1)$ c.d.f., and

$z[\alpha] = z_0 + (z_0 + az^{(\alpha)})/(1 - a(z_0 + z^{(\alpha)}))$.

The distribution $G_{\hat{\theta}}$ is approximated using (2.1).

If the transformation g stabilizes the variance, resulting in an acceleration constant of zero, the interval (2.5) reduces to Efron's earlier BC interval. If the bias constant is also zero, it reduces to the percentile interval $[G_{\hat{\theta}}^{-1}(\alpha), G_{\hat{\theta}}^{-1}(1-\alpha)]$. Beginning with the percentile interval, the BC and BC_a methods successively adjust the chosen percentiles, the former taking into account bias and the latter additionally considering variance.

The appeal of this formulation is that the transformation g need not be known in order to form the BC_a interval. However, the two constants must still be determined. The bias correction can be estimated easily from the approximate bootstrap distribution by

$$\hat{z}_0 = \Phi^{-1}(\hat{G}_{\hat{\theta}}(\hat{\theta})) \quad .$$

The acceleration constant is more difficult to calculate. Efron gives one estimate which entails the skewness of the score function:

$$\hat{a} = 1/6 \ \mathrm{SKEW}(\dot{l}_\theta(\hat{\theta}))\ |_{\theta=\hat{\theta}} \quad , \tag{2.6}$$

where $\dot{l}_\theta = \partial/\partial\theta \ log f_\theta(\hat{\theta})$,

$f_\theta(\hat{\theta})$ is the density of $\hat{\theta}$ for a particular value θ, and

SKEW is the skewness.

In addition, he proves that if $\hat{\theta}$ is the maximum-likelihood estimate, a is approximately equal to z_0. DiCiccio and Romano (1987a) discuss other estimates based on the moments of $\hat{\theta}$, which apply in the non-maximum-likelihood case and when nuisance parameters are present.

If the transformation assumption (2.4) is true, then the BC_a interval (2.5) is an exact interval. Even if this requirement is not strictly met, the interval is second-order correct as defined in (2.3). However, the BC interval and percentile intervals are only first-order correct as is the ordinary standard interval.

DiCiccio and Romano (1987b) voice a criticism of the BC_a method which is shared by Hall (1988). The objection is that the acceleration constant must be derived theoretically, which means this particular bootstrap interval is not purely automatic, the most appealing and fundamental feature of bootstrap methods. This problem will surface when the clinical trial example is examined.

2.2 The DiCiccio-Romano Interval

As Schenker (1987) points out, an exact upper endpoint $\theta_{U,\alpha}$ exists, namely that value of θ such that $\hat{\theta} = G_{\theta}^{-1}(\alpha)$ and analagously for the lower endpoint. However, simulation of G_θ for numerous values of θ in order to conduct a search, would be prohibitive. The DiCiccio-Romano method (1987b) bypasses this obstruction by prudently choosing the values of θ at which the bootstrap distribution is simulated and then constructing an interval based on a theoretical relationship between these distributions. In so doing, they not only deal with the acceleration constant calculation problem, as their method does not require this parameter, but also they achieve exactness when

$$G_{\hat{\theta}}(\theta) \text{ is a pivot} \quad , \tag{2.7}$$

which is a weaker condition than (2.4). Their iterative method for the upper endpoint θ_U is:

Let $\theta_0 =$ any value of θ,

in particular it could be the percentile point $G_{\hat{\theta}}(1-\alpha)$,

then $\theta_i' = G_{\theta_i}^{-1}(\alpha)$, and

$$\theta_{i+1} = G_{\hat{\theta}}^{-1}\{G_{\theta_i'}(\theta_i)\} \quad \text{for } i = 1, 2, \ldots . \tag{2.8}$$

The i$^{\text{th}}$ step DiCiccio-Romano endpoint is defined as θ_i. The lower endpoint endpoint is found by replacing α by $1-\alpha$ in the above. In practice, (2.1) is used at each step to estimate the bootstrap distribution.

If the exactness condition (2.7) holds, θ_1 will be an exact upper (lower) endpoint. Otherwise, the method can be iterated as outlined in (2.8) until the endpoint is deemed satisfactory as discussed in Section 6. The number of steps required for either endpoint may differ. Notably, the amount of computation required is linear in the number of iterations. In addition, DiCiccio and Romano (1987b) show that each iteration reduces the error by $O_p(n^{-1/2})$.

2.3 The Bootstrap-t Interval

A third approach is the bootstrap-t method (Efron 1981, Hall 1988), which results from studentizing in analogy with the location-scale problem. In this case, a stable estimate $\hat{\sigma}$ of the standard deviation of $\hat{\theta}$ must be known. The sampling distribution which we would like to estimate is $H_\theta(s) \equiv P_\theta\{(\hat{\theta}-\theta)/\hat{\sigma} \leq s\}$. The bootstrap method provides the bootstrap distribution $H_{\hat{\theta}}(s) \equiv P_{\hat{\theta}}\{(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^* \leq s\}$ where $\hat{\sigma}^*$ is the estimated standard deviation of $\hat{\theta}^*$. The resulting interval is

$$[\hat{\theta} - \hat{\sigma}H_{\hat{\theta}}^{-1}(1-\alpha), \hat{\theta} - \hat{\sigma}H_{\hat{\theta}}^{-1}(\alpha)] \quad . \tag{2.9}$$

In practice, a sample of size B is taken and $H_{\hat{\theta}}$ is is approximated by

$$\hat{H}_{\hat{\theta}}(s) = \#\{(\hat{\theta}_b^* - \hat{\theta})/\hat{\sigma}^* \leq s\}/B \quad .$$

Hall (1988) shows that the bootstrap-t interval (2.9) is second-order correct and provides an interesting theoretical framework for comparing it with the BC$_a$

interval. He advises the former based on its philosophical appeal, even though both have the same error. Basically, if an interval for a mean when the variance is unknown were desired, the bootstrap-t method is analagous to the correct approach: the $(1-\alpha)^{\text{th}}$ and α^{th} percentile points from the appropriate t-table are used for the lower and upper endpoints respectively.

3. The Clinical Trial Example

We consider a simplified medical experiment in which two treatments are to be compared as described by Siegmund (1985). Paired patients enter the trial sequentially, perhaps matched on some external factors such as age . The difference between the treatment 1 and treatment 2 responses for the i^{th} pair is denoted x_i, $i = 1, \ldots$. The x_i's are assumed to be i.i.d. normal with mean μ and known variance which, without loss of generality, is set equal to one. For expositional purposes, we first consider a test of the mean and then construct the analogous confidence interval. The null hypothesis is $H_0 : \mu = 0$ versus the alternative $H_a : \mu \neq 0$. The trial should be terminated as soon as enough evidence is gathered to adequately favor one treatment over the other. In the fixed sample size setting, the null hypothesis would be rejected if

$$S_n \equiv \mid \sum_{i=1}^{n} x_i \mid > b\sqrt{n} \quad ,$$

where b is chosen to achieve the desired significance level of the test.

One intuitive sequential stopping rule is to first observe a minimum number of observations (m_0) before even considering stopping the trial, and to define the time T as

$$T \equiv \inf\{n : n \geq m_0, \mid S_n \mid > b\sqrt{n}\} \quad .$$

Since we do not want to go on sampling indefinitely, a maximum of m patient pairs are allowed to enter the experiment. Combining these elements, the trial

is stopped at $\tau \equiv \min(T, m)$. The null hypothesis is rejected if and only if the cumulative sum S_τ is less than or equal to some level where the level and the other test parameters m_0, b, and m are chosen to achieve the desired test significance.

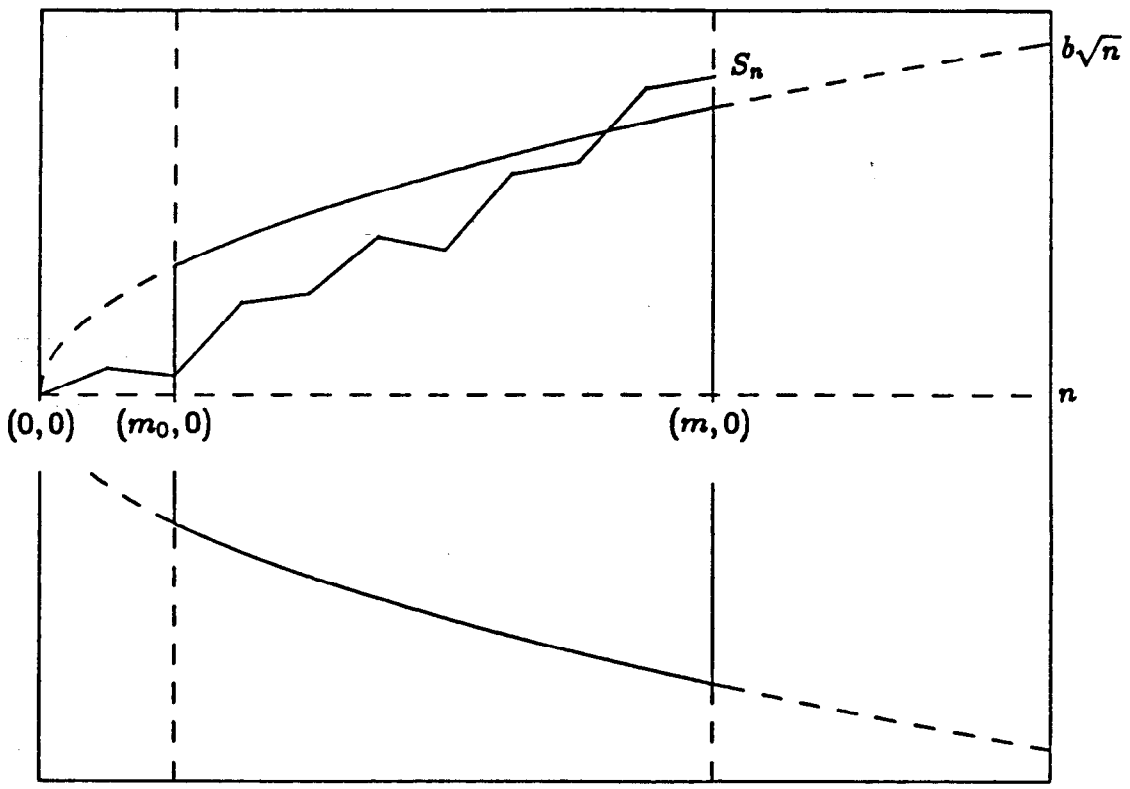

Figure 1

The stopping boundary is a parabola lying horizontally with a vertical boundary at m_0 after which the process is first observed and one at m at which the process is cut-off. *Figure 1* shows a typical process S_n which hits the parabolic part of the boundary.

4. The Theoretical Interval

Siegmund (1978, 1985) develops a theoretical confidence interval by noting that points on the stopping boundary are ordered, with small values corresponding to large μ values, which favor rejection of the null hypothesis. The upper limit $\mu_{U,\alpha}$ for a $1-2\alpha$ interval is

$$\mu_{U,\alpha} = \sup\{\mu : P_\mu\{\text{stopping point is equal to or smaller than the observed stopping point}\} \geq \alpha\} \quad . \tag{4.1}$$

A similar expression is available for $\mu_{L,\alpha}$. Siegmund has worked out analytical approximations for the probabilities required in (4.1) and he solves for the endpoints by doing a numerical search for the sup (inf) values. Because the sample points are ordered simply by their stopping time and any excess over the boundary $(S_T - b\sqrt{T})$ is ignored, some information is inherently lost. The approximations are also dependent on the normality assumption.

5. The Bootstrap Procedure

Two trial situations will be used to compare the theoretical and bootstrap intervals (Siegmund 1978, 1985). All intervals will be two-sided with a confidence level of 0.90. Situations I and II have five and four hypothetical outcomes respectively:

Situation I	**Situation II**
$m_0 = 1, m = 148, b = 3.45$	$m_0 = 16, m = 144, b = 3.12$
$T = 19$ and $S_T > 0$	$T = 39$ and $S_T > 0$
$T = 32$ and $S_T > 0$	$T = 61$ and $S_T > 0$
$T = 68$ and $S_T > 0$	$T > m$ and $S_m = 36.0$
$T > m$ and $S_m = 40.0$	$T > m$ and $S_m = 28.8$
$T > m$ and $S_m = 30.0$	

The boundary parameters for each Situation are given in the first line of the

table. In Situation I, the trial is stopped as soon as possible and a maximum of 148 patient pairs will be allowed to enter the trial. In Situation II, sixteen patient pairs will be allowed to enter the experiment before the results are examined and a maximum of 144 pairs will be allowed.

Consider the first and fourth Situation I outcomes. The first outcome is that '$T = 19$ and $S_T > 0$', which means that the process hit the upper parabolic boundary at time $T = 19$. In clinical trial terms, this means that after 19 patient pairs entered the experiment, the sum of the differences between treatment 1 and treatment 2 levels was greater than $3.45\sqrt{19}$. Unfortunately, and perhaps somewhat unrealistically, we do not know the actual statistic S_T value. We only know that it was as least as large as the boundary that point.

The fourth outcome is that '$T > m$ and $S_m = 40$', which means that the observed process did not hit the parabolic boundary by the time the maximum allowable number of patient pairs had entered the trial. Rather, the process hit the vertical boundary at a height of $S_m = 40$.

5.1 Estimators of the Drift

In order to conduct the resampling, an estimator of μ must be chosen. The likelihood function for μ is

$$l(\tau, S_\tau; \mu) = exp(\mu S_\tau - \mu^2 \tau/2)$$

regardless of the stopping rule. The maximum-likelihood estimator is

$$\hat{\mu}_{mle} \equiv S_\tau/\tau \quad , \tag{5.1}$$

which is just the usual sample mean in the fixed sample size case. This will be one of the estimators used in the simulations. However, as Woodroofe (1982) notes, though the likelihood function is not dependent on the stopping rule, the

distribution of the sample mean is. Siegmund (1978) proves that the bias is of order $1/b$. He suggests an estimator of the form

$$\hat{\mu} \equiv \begin{cases} (S_T/T)/(1+2/b^2) & \text{if } T \leq m, \\ S_m/m & \text{if } T > m \quad . \end{cases}$$

The second estimator which will be used in the simulations is a continuous version of the above:

$$\hat{\mu}_{ub} \equiv (S_\tau/\tau)/(1+2/b^2) \quad . \tag{5.2}$$

As pointed out earlier, if the process hits before the parabolic boundary, the outcomes do not include the specific stopping height S_T. Realistically, the investigator would keep a record of the patient pair differences. A reasonable estimate is $b\sqrt{T}$, the actual height of the boundary at the stopping time, though the true value is at least as large as this height. This approximation will be used in (5.1) and (5.2) to estimate μ for Situation I's first three outcomes and Situation II's first two outcomes.

5.2 The Bootstrap Sample

For each outcome, an estimate of the drift μ is calculated and B bootstrap samples $\hat{\mu}_b^*$ are drawn from $F_{\hat{\mu}}$ in the following manner:

1. Let $S^*_{m_0-1} = \sum_{i=1}^{m_0-1} x_i$ where each $x_i \sim N(\hat{\mu}, 1)$.

2. For $k = m_0$ to m

begin

2.1 Let $S_k^* = S_{k-1}^* + x_k$ where $x_k \sim N(\hat{\mu}, 1)$.

2.2 If $S_k^* \geq b\sqrt{k}$, then the bootstrap sample point is (S_k^*, k). **GOTO 4.**

end

3. The bootstrap sample is (S_m^*, m). **GOTO 4.**

4. Calculate $\hat{\mu}_b^*$.

In other words, we simulate a random walk and observe it after a suitable wait. We have no way of knowing if the process may have crossed the parabolic boundary and then recrossed it before it is looked at initially (before m_0).

The bootstrap samples do differ slightly from the original outcomes in that if the bootstrap process crossed the parabolic boundary (Step **2.2**), we know the exact height at the stopping time, S_T. In most bootstrap situations, the statistician mirrors the original observation in the resampling procedure. However, if the excess over the boundary is ignored, the resulting bootstrap distribution $G_{\hat{\mu}}$ is discrete as the only stopping heights which can be observed are $b\sqrt{m_0}$, $b\sqrt{m_0+1}, \ldots$. Initially, this approach was tried in the simulation work but it required smoothing the resulting bootstrap distribution. Better results were obtained by an easier procedure if the bootstrap samples consisted of (S_τ^*, τ). Siegmund's approach, as described in Section 4, avoids this choice as his approximate probabilities are in terms of the stopping time only.

5.3 Estimator of the Standard Deviation

In order to calculate the bootstrap-t interval, an estimate of the standard deviation of $\hat{\mu}$ must be chosen. If the number of patient pairs was fixed at n, $\hat{\sigma}$ would be $1/\sqrt{n}$. In the sequential analysis situation, the standard interval is

$$[\hat{\mu}_{mle} - z^{(1-\alpha)}/\sqrt{\tau}, \hat{\mu}_{mle} - z^{(\alpha)}/\sqrt{\tau}] \quad . \tag{5.3}$$

For lack of a better solution, we will use $1/\sqrt{\tau}$ for an estimate of the standard deviation for both $\hat{\mu}_{mle}$ and $\hat{\mu}_{ub}$. The problem of finding a stable $\hat{\sigma}$ is a difficulty inherent in the bootstrap-t method.

5.4 The Acceleration Constant

For the simulations using $\hat{\mu}_{mle}$, the acceleration constant a will be approximated by $\hat{z}_0$ (Efron 1987). A satisfactory analytical answer could not be found in the $\hat{\mu}_{ub}$ case, illustrating a problem with the BC_a method as discussed in Section 2.1. Disregarding the vertical boundaries at m_0 and m, an estimate of the expected value of the stopping time for a parabolic boundary is given by Siegmund (1985). Combining this result with the score function estimate (2.6), Wald's identity and Bartlett's formula, produced the estimate

$$\hat{a} \equiv -(b^2 - 1)^{-1/2} \quad , \tag{5.4}$$

which equals -0.303 and -0.338 in Situations 1 and 2 respectively. However, this estimate did not work well in practice. In outcomes when the process hit the vertical boundary at m, such as the fourth and fifth in Situation I, (5.4) was especially poor. This result is intuitive as these situations are close to the fixed sample size case in which no transformation g (2.4) is necessary and both the bias and acceleration constants are zero. Therefore, all BC_a intervals will be constructed with the acceleration constant set equal to the bias constant, regardless of which estimator is used.

6. Discussion

For both Situations, intervals were calculated using both the maximum-likelihood and unbiased estimators for bootstrap sample sizes of B=30,000 and B=5,000. For each of the eight combinations, the Siegmund (Section 4), standard (5.3), DiCiccio-Romano 1st through 4th step (2.8), percentile, BC, and BC_a (2.5), and bootstrap-t (2.9) intervals were constructed.

A reasonable method for evaluating and comparing the confidence intervals is to use Efron's correctness criterion (2.2). For each interval, $[\mu_L, \mu_U]$, the tail probabilities $1 - G_{\mu_L}(\hat{\mu})$ and $G_{\mu_U}(\hat{\mu})$ are estimated using (2.1) with B equal to

the bootstrap sample size used in the interval construction. As stated in DiCiccio and Romano (1987b), the standard error in these approximate tail probabilities can be estimated by

$$((\alpha(1-\alpha))/B)^{1/2} \quad .$$

For α=0.05, this estimated standard error is 0.0013 for B=30,000 and 0.0031 for B=5,000.

The right-to-left ratio

$$(\mu_U - \hat{\mu})/(\hat{\mu} - \mu_L)$$

is also reported, indicating the skewness of the interval.

Appendix A contains the Situation 2, B=30,000 intervals for both the unbiased (5.4) and maximum-likelihood (5.2) estimators in Sections A.1 and A.2 respectively. Situation 2 was chosen because its large value of m_0 presents greater difficulties for the methods. Each entry consists of the name of the method, the interval, the upper-to-lower ratio (R/L Ratio), and the upper and lower tail probabilities (P_U and P_L).

6.1 The Unbiased Estimator Results

The Siegmund method does well in all cases. The DiCiccio-Romano procedure also performs well except perhaps for the third outcome which is the most difficult because the observed process stopped close to the vertex between the parabolic boundary and the vertical boundary at m. Though four steps are given for this method, we could stop iterating an endpoint as soon as the respective tail probability is close enough to 0.05, where closeness could be defined using the standard error (6.1). For example, one step might be acceptable for the lower endpoint and two for the upper for the first outcome. The bootstrap-t method does less well than the Siegmund and DiCiccio-Romano procedures.

The percentile, BC and BC_a intervals are all skewed to the upper as indicated by large upper-to-lower ratios and consequently have large lower tail probabilities and small upper tail probabilities.

The standard interval is skewed in the opposite direction. Understandably, this method does better at the lower endpoint than at the upper since the lower corresponds to the fixed sample size situation in which the standard interval is exact.

6.2 The Maximum-likelihood Estimator Results

If the maximum-likelihood estimator is used instead, the Siegmund interval is the same since it does not incorporate $\hat{\mu}$. The tail probabilities change as the different estimate is used to generate the relevant bootstrap distributions. The bootstrap-t intervals also stay roughly the same.

Perhaps because setting the acceleration constant equal to the bias constant is now a better approximation, the BCa interval does better in the maximum-likelihood situation than in the unbiased one.

6.3 Further Comments

The results for both Situations and for both estimators are summarized in A.3. If the unbiased estimator is used, the DiCiccio-Romano method constructs endpoints whose tail probabilities are within [0.035,0.065] for eight out of nine outcomes. Siegmund's intervals based on the maximum-likelihood estimator satisfy this criterion for seven outcomes.

A bootstrap sample size of 30,000 helps to ensure that the observed differences between the intervals are results of the methods themselves rather than simulation error. However, in practice the computer cost is prohibitive and suggestions on how to choose B (Efron 1987), and importance sampling ideas should be considered. The Situation 2 unbiased estimator intervals using B=5,000 are given in A.4. In general, the intervals exhibit the same behavior as discussed in

Section 6.1. The DiCiccio-Romano and BC_a upper endpoints are the least stable as the bootstrap distributions have long upper tails.

6.4 Conclusions

This exercise has demonstrated problems encountered in the application of bootstrap methods to a complicated yet practical example. The DiCiccio-Romano procedure has proved especially promising. This new method retains the automatic nature of the bootstrap, not requiring the type of analytical work which turns out to be difficult and even impossible in other approaches. More complicated bootstrap procedures involving prepivoting (Beran 1987) or double bootstrapping (Hall 1986) also might work well in this situation and should be considered in the future. We hope that when faced with a similar problem, the reader will find this illustration a useful one.

Acknowledgements: The author gratefully acknowledges the help of Professors Thomas DiCiccio, Bradley Efron, Jerome Friedman, Joseph Romano and David Siegmund.

References

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457-468.

DiCiccio, T. and Romano, J. (1987a). A review of bootstrap confidence intervals. Tech. Rep. 279, Dept. of Statistics, Stanford University.

DiCiccio, T. and Romano, J. (1987b). Accurate bootstrap confidence limits. Tech. Rep. 281, Dept. of Statistics, Stanford University.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Can. J. Statist.* **9**, 139-172.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* CMBS 38, SIAM-NSF.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82**, 171-185.

Hall, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14**, 1431-1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. To appear in *Ann. Statist.*

Schenker, N. (1987). Comment. *J. Amer. Statist. Assoc.* **82**, 192-194.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika* **65**, 341-349.

Siegmund, D. (1985). *Sequential analysis, tests and confidence intervals.* Springer–Verlag.

Tibshirani, R. (1984). Bootstrap confidence intervals. Tech. Rep. 216, Dept. of Statistics, Stanford University.

Woodroofe, M. (1982) *Nonlinear renewal theory in sequential analysis.* Society for Industrial and Applied Mathematics.

Appendix A

A.1 Situation 2, B=30,000 using the unbiased estimator

Outcome 1: $T = 39$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.160, 0.740]	1.279	0.043	0.036
Standard	[0.151, 0.678]	1.000	0.037	0.078
DiCiccio-Romano 1	[0.155, 0.696]	1.084	0.039	0.062
DiCiccio-Romano 2	[0.182, 0.701]	1.230	0.053	0.056
DiCiccio-Romano 3	[0.169, 0.702]	1.173	0.045	0.057
DiCiccio-Romano 4	[0.180, 0.707]	1.244	0.052	0.055
Percentile	[0.241, 0.694]	1.611	0.103	0.066
BC [$\hat{z}_0$=0.202]	[0.266, 0.771]	2.394	0.129	0.023
BC_a	[0.290, 0.990]	4.611	0.164	0.000
Bootstrap-t	[0.233, 0.718]	1.671	0.096	0.050

Outcome 2: $T = 61$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.130, 0.590]	1.284	0.049	0.037
Standard	[0.121, 0.542]	1.000	0.045	0.076
DiCiccio-Romano 1	[0.100, 0.587]	1.102	0.029	0.039
DiCiccio-Romano 2	[0.157, 0.573]	1.388	0.074	0.049
DiCiccio-Romano 3	[0.114, 0.564]	1.074	0.039	0.054
DiCiccio-Romano 4	[0.145, 0.564]	1.250	0.058	0.055
Percentile	[0.164, 0.637]	1.823	0.078	0.015
BC [$\hat{z}_0$=0.176]	[0.193, 0.693]	2.613	0.115	0.004
BC_a	[0.227, 0.863]	5.097	0.167	0.000
Bootstrap-t	[0.166, 0.589]	1.564	0.082	0.037

A.1 Situation 2, B=30,000 using the unbiased estimator, cont.

Outcome 3: $T > m$ and $S_m = 36.0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.080, 0.380]	1.355	0.054	0.045
Standard	[0.070, 0.344]	1.000	0.045	0.097
DiCiccio-Romano 1	[0.085, 0.371]	1.341	0.057	0.052
DiCiccio-Romano 2	[0.061, 0.376]	1.157	0.035	0.048
DiCiccio-Romano 3	[0.087, 0.368]	1.337	0.060	0.056
DiCiccio-Romano 4	[0.060, 0.382]	1.182	0.034	0.040
Percentile	[0.057, 0.475]	1.781	0.033	0.002
BC [$\hat{z}_0$=0.260]	[0.095, 0.634]	3.780	0.070	0.000
BC_a	[0.121, 0.934]	8.380	0.114	0.000
Bootstrap-t	[0.082, 0.358]	1.200	0.056	0.073

Outcome 4: $T > m$ and $S_m = 28.8$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.030, 0.330]	1.207	0.032	0.055
Standard	[0.029, 0.303]	1.000	0.028	0.101
DiCiccio-Romano 1	[0.048, 0.296]	1.104	0.046	0.115
DiCiccio-Romano 2	[0.050, 0.357]	1.649	0.048	0.027
DiCiccio-Romano 3	[0.051, 0.320]	1.337	0.050	0.068
DiCiccio-Romano 4	[0.051, 0.340]	1.513	0.048	0.041
Percentile	[0.024, 0.396]	1.618	0.024	0.009
BC [$\hat{z}_0$=0.334]	[0.071, 0.636]	4.935	0.072	0.000
BC_a	[0.098, 0.441]	4.080	0.128	0.002
Bootstrap-t	[0.037, 0.308]	1.106	0.035	0.092

A.2 Situation 2, B=30,000 using the maximum-likelihood estimator

Outcome 1: $T = 39.0$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.160, 0.740]	0.708	0.043	0.036
Standard	[0.236, 0.763]	1.000	0.100	0.024
DiCiccio-Romano 1	[0.131, 0.798]	0.808	0.029	0.014
DiCiccio-Romano 2	[0.234, 0.750]	0.940	0.098	0.033
DiCiccio-Romano 3	[0.131, 0.725]	0.611	0.029	0.044
DiCiccio-Romano 4	[0.258, 0.716]	0.896	0.121	0.049
Percentile	[0.351, 0.913]	2.779	0.272	0.002
BC [$\hat{z}_0$=-0.329]	[0.304, 0.780]	1.432	0.186	0.020
BC_a	[0.177, 0.684]	0.571	0.050	0.074
Bootstrap-t	[0.235, 0.716]	0.816	0.100	0.050

Outcome 2: $T = 61.0$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.130, 0.590]	0.707	0.049	0.037
Standard	[0.189, 0.610]	1.000	0.110	0.026
DiCiccio-Romano 1	[0.050, 0.654]	0.729	0.014	0.011
DiCiccio-Romano 2	[0.273, 0.583]	1.451	0.260	0.043
DiCiccio-Romano 3	[0.050, 0.579]	0.514	0.013	0.043
DiCiccio-Romano 4	[0.272, 0.577]	1.387	0.266	0.046
Percentile	[0.279, 0.821]	3.515	0.281	0.000
BC [$\hat{z}_0$=-0.325]	[0.211, 0.666]	1.412	0.140	0.008
BC_a	[0.077, 0.561]	0.502	0.022	0.060
Bootstrap-t	[0.179, 0.574]	0.788	0.094	0.049

A.2 Situation 2, B=30,000 using the m.l.e., cont.

Outcome 3: $T > m$ and $S_m = 36.0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.080, 0.380]	0.765	0.054	0.045
Standard	[0.113, 0.387]	1.000	0.098	0.037
DiCiccio-Romano 1	[0.012, 0.376]	0.531	0.011	0.047
DiCiccio-Romano 2	[0.106, 0.372]	0.852	0.088	0.051
DiCiccio-Romano 3	[0.024, 0.372]	0.539	0.015	0.054
DiCiccio-Romano 4	[0.106, 0.374]	0.862	0.086	0.048
Percentile	[0.113, 0.651]	2.920	0.099	0.000
BC [$\hat{z}_0$=-0.222]	[0.076, 0.501]	1.443	0.048	0.001
BC_a	[-0.034, 0.414]	0.578	0.004	0.017
Bootstrap-t	[0.084, 0.387]	0.827	0.057	0.036

Outcome 4: $T > m$ and $S_m = 28.8$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.030, 0.330]	0.765	0.032	0.055
Standard	[0.063, 0.337]	1.000	0.063	0.046
DiCiccio-Romano 1	[0.046, 0.298]	0.636	0.042	0.109
DiCiccio-Romano 2	[0.057, 0.368]	1.174	0.054	0.021
DiCiccio-Romano 3	[0.053, 0.307]	0.725	0.050	0.092
DiCiccio-Romano 4	[0.054, 0.357]	1.071	0.053	0.028
Percentile	[0.063, 0.554]	2.579	0.062	0.000
BC [$\hat{z}_0$=-0.051]	[0.054, 0.517]	2.174	0.051	0.000
BC_a	[0.041, 0.472]	1.707	0.036	0.000
Bootstrap-t	[0.029, 0.337]	0.803	0.029	0.047

A.3 Summary

The interval names are abbreviated: Si (Siegmund), St (Standard), DR (DiCiccio-Romano), Pe (Percentile)), BC, BC_a, and Bt (Bootstrap-t).

In the summary below, '$\star$' denotes a tail probability in in [0.035,0.065], '+' in (0.065,0.080], '++' above 0.080, '−' in [0.020,0.035), '−−' below 0.020. The DiCiccio-Romano intervals consist of the best endpoints chosen independently from the 4 steps where a '−' is considered better than a '+', a '+' better than a '−−', and a '−−' better than '++', as over-coverage is considered better than under-coverage.

Situation 1, B=30,000 using the maximum-likelihood estimator

Interval	Si	St	DR	Pe	BC	BC_a	Bt
Outcome 1	($\star$,−)	(++,−)	($\star$,$\star$)	(++,−−)	(++,−−)	(+,$\star$)	($\star$,$\star$)
Outcome 2	($\star$,−)	(++,−)	($\star$,$\star$)	(++,−−)	(++,−−)	(+,$\star$)	($\star$,$\star$)
Outcome 3	($\star$,$\star$)	(++,−)	(−−,$\star$)	(++,−−)	(++,−−)	(−−,$\star$)	($\star$,$\star$)
Outcome 4	($\star$,$\star$)	(++,$\star$)	(−,$\star$)	(++,−−)	($\star$,−−)	(−−,−−)	($\star$,$\star$)
Outcome 5	($\star$,$\star$)	($\star$,$\star$)	($\star$,−)	($\star$,−−)	($\star$,−−)	($\star$,−−)	(−,$\star$)

Situation 1, B=30,000 using the unbiased estimator

Interval	Si	St	DR	Pe	BC	BC_a	Bt
Outcome 1	($\star$,−)	($\star$,+)	($\star$,$\star$)	(++,−−)	(++,−−)	(++,−−)	($\star$,$\star$)
Outcome 2	($\star$,−)	($\star$,+)	($\star$,$\star$)	(++,−−)	(++,−−)	(++,−−)	($\star$,$\star$)
Outcome 3	($\star$,$\star$)	($\star$,+)	($\star$,$\star$)	(+,−−)	(++,−−)	(++,−−)	($\star$,$\star$)
Outcome 4	($\star$,$\star$)	($\star$,++)	($\star$,−)	(−,−−)	(+,−−)	(++,−−)	($\star$,+)
Outcome 5	($\star$,$\star$)	(−,++)	($\star$,$\star$)	(−,−)	(+,−−)	(++,−−)	(−,++)

Situation 2, B=30,000 using the maximum-likelihood estimator

Interval	Si	St	DR	Pe	BC	BC_a	Bt
Outcome 1	($\star$,$\star$)	(++,−)	(−,$\star$)	(++,−−)	(++,−)	($\star$,+)	(++,$\star$)
Outcome 2	($\star$,$\star$)	(++,−)	(−−,$\star$)	(++,−−)	(++,−−)	(−,$\star$)	(++,$\star$)
Outcome 3	($\star$,$\star$)	(++,$\star$)	(−−,$\star$)	(++,−−)	($\star$,−−)	(−−,−−)	($\star$,$\star$)
Outcome 4	(−,$\star$)	($\star$,$\star$)	($\star$,−)	(+,−−)	($\star$,−−)	($\star$,−−)	(−,$\star$)

Situation 2, B=30,000 using the unbiased estimator

Interval	Si	St	DR	Pe	BC	BC_a	Bt
Outcome 1	($\star$,$\star$)	($\star$,+)	($\star$,$\star$)	(++,+)	(++,−)	(++,−−)	(++,$\star$)
Outcome 2	($\star$,$\star$)	($\star$,+)	($\star$,$\star$)	(+,−−)	(++,−−)	(++,−−)	(++,$\star$)
Outcome 3	($\star$,$\star$)	($\star$,++)	($\star$,$\star$)	(+,−−)	(+,−−)	(++,−−)	($\star$,+)
Outcome 4	(−,$\star$)	(−,++)	($\star$,$\star$)	(−,−−)	(+,−−)	(++,−−)	($\star$,++)

A.4 Situation 2, B=5,000 using the unbiased estimator

Outcome 1: $T = 39.0$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.160, 0.740]	1.279	0.042	0.038
Standard	[0.151, 0.678]	1.000	0.040	0.078
DiCiccio-Romano 1	[0.127, 0.687]	0.951	0.023	0.071
DiCiccio-Romano 2	[0.204, 0.690]	1.306	0.072	0.068
DiCiccio-Romano 3	[0.165, 0.690]	1.104	0.043	0.068
DiCiccio-Romano 4	[0.182, 0.687]	1.173	0.057	0.066
Percentile	[0.241, 0.687]	1.574	0.101	0.070
BC [$\hat{z}_0$=0.208]	[0.265, 0.774]	2.410	0.142	0.025
BC_a	[0.290, 0.959]	4.356	0.163	0.001
Bootstrap-t	[0.239, 0.718]	1.726	0.101	0.045

Outcome 2: $T = 61.0$ and $S_T > 0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.130, 0.590]	1.284	0.045	0.038
Standard	[0.121, 0.542]	1.000	0.039	0.082
DiCiccio-Romano 1	[0.127, 0.687]	0.951	0.023	0.071
DiCiccio-Romano 2	[0.204, 0.690]	1.306	0.072	0.068
DiCiccio-Romano 3	[0.165, 0.690]	1.104	0.043	0.068
DiCiccio-Romano 4	[0.182, 0.687]	1.173	0.057	0.066
Percentile	[0.167, 0.646]	1.917	0.083	0.014
BC [$\hat{z}_0$=0.180]	[0.197, 0.695]	2.713	0.121	0.004
BC_a	[0.229, 0.884]	5.391	0.171	0.000
Bootstrap-t	[0.164, 0.584]	1.509	0.080	0.041

A.4 Situation 2, B=5,000 using the unbiased estimator, cont.

Outcome 3: $T > m$ and $S_m = 36.0$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.080, 0.380]	1.355	0.051	0.044
Standard	[0.070, 0.344]	1.000	0.043	0.091
DiCiccio-Romano 1	[0.089, 0.370]	1.378	0.065	0.052
DiCiccio-Romano 2	[0.056, 0.372]	1.089	0.034	0.053
DiCiccio-Romano 3	[0.092, 0.367]	1.387	0.065	0.058
DiCiccio-Romano 4	[0.053, 0.379]	1.110	0.032	0.045
Percentile	[0.059, 0.466]	1.739	0.033	0.003
BC [$\hat{z}_0$=0.255]	[0.094, 0.635]	3.756	0.062	0.000
BC_a	[0.120, 0.651]	5.050	0.111	0.000
Bootstrap-t	[0.083, 0.356]	1.198	0.051	0.074

Outcome 4: $T > m$ and $S_m = 28.8$				
	Interval	R/L Ratio	P_L	P_U
Siegmund	[0.030, 0.330]	1.207	0.034	0.061
Standard	[0.029, 0.303]	1.000	0.034	0.106
DiCiccio-Romano 1	[0.048, 0.303]	1.160	0.045	0.103
DiCiccio-Romano 2	[0.048, 0.352]	1.579	0.046	0.032
DiCiccio-Romano 3	[0.047, 0.327]	1.357	0.045	0.057
DiCiccio-Romano 4	[0.049, 0.329]	1.396	0.045	0.057
Percentile	[0.024, 0.413]	1.747	0.026	0.003
BC [$\hat{z}_0$=0.319]	[0.070, 0.624]	4.769	0.072	0.000
BC_a	[0.096, 0.425]	3.726	0.129	0.003
Bootstrap-t	[0.033, 0.307]	1.068	0.032	0.092