Bootstrap Confidence Intervals for Ratios of Expectations

DENIS CHOQUET, PIERRE L'ECUYER, and CHRISTIAN LÉGER
Université de Montréal

We are concerned with computing a confidence interval for the ratio $E[Y]/E[X]$, where (X,Y) is a pair of random variables. This ratio estimation problem arises for instance in regenerative simulation. As an alternative to confidence intervals based on asymptotic normality, we study and compare different variants of the bootstrap, for one-sided and two-sided intervals. We point out situations where these techniques provide confidence intervals with coverage much closer to the nominal value than the classical methods.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]

General Terms: Algorithms

Additional Key Words and Phrases: Ratio estimation problem; regenerative simulation; confidence intervals; bootstrap;

This work has been supported by NSERC-Canada grants no. OGP0039996, ODGP0110050, and SMF0169893, and by FCAR-Québec grant no. 93ER1654. C. Léger wishes to acknowledge the support of the Centre de recherches mathématiques of Université de Montréal and of the MITACS Centre of Excellence. The authors thank two anonymous referees and the Area Editor James R. Wilson for their helpful comments.
Name: D. Choquet
Address: Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, H3C 3J7, Canada;
e-mail: `denisc@crt.umontreal.ca`
Name: P. L'Ecuyer
Address: Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, H3C 3J7, Canada;
e-mail: `lecuyer@iro.umontreal.ca`
Name: C. Léger
Address: Département de Mathématique et de Statistique, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, H3C 3J7, Canada;
e-mail: `leger@dms.umontreal.ca`

1. INTRODUCTION

Let (X, Y) be a pair of jointly distributed random variables, and assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independently and identically distributed (i.i.d.) copies of (X, Y). The aim is to construct a confidence interval for the ratio

$$\mu = \frac{E[Y]}{E[X]}, \tag{1}$$

based on this i.i.d. sample, assuming that $E[X] > 0$.

This problem arises in many different settings. For example, let $\{C(t),\ t \geq 0\}$ be a real-valued regenerative process with regeneration epochs $0 = \tau_0 < \tau_1 < \cdots$, which may represent a rate at which cost is accumulated. For $i \geq 1$, let

$$\begin{aligned} \tilde{Y}_i &= \int_{\tau_{i-1}}^{\tau_i} |C(t)| dt \\ Y_i &= \int_{\tau_{i-1}}^{\tau_i} C(s) ds \\ X_i &= \tau_i - \tau_{i-1}. \end{aligned}$$

If $E[\tilde{Y}_1 + X_1] < \infty$, then

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t C(s) ds \stackrel{\text{a.s.}}{=} \mu = E[Y_1]/E[X_1].$$

The steady-state mean of the process is thus expressed as a ratio of two expectations. For definitions and properties of regenerative processes, see, for example, Asmussen [1987], Meyn and Tweedie [1993], and Wolff [1989]. Other quantities of interest that can be expressed as a ratio of expectations include the infinite-horizon total expected discounted cost for a regenerative process, the expected hitting time of a specific set of states for a regenerative process, or an expectation conditional on a specific event (see Fox and Glynn [1989] and Glynn, L'Ecuyer, and Adès [1991] for details).

Let $\bar{X}$, S_X^2, $\bar{Y}$, S_Y^2, and S_{XY} denote the sample mean and variance of the X_j's, the sample mean and variance of the Y_j's, and the sample covariance between the X_j's and Y_j's, respectively. Define

$$\hat{\mu} = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \tag{2}$$

and put

$$Z_j = Y_j - \mu X_j, \tag{3}$$

which has zero expectation. The variance of Z_j can be estimated by

$$S_Z^2 = S_Y^2 - 2\hat{\mu} S_{XY} + \hat{\mu}^2 S_X^2. \tag{4}$$

Denote $\bar{Z} = \bar{Y} - \mu \bar{X}$ and define the studentized statistic (with mean 0 and variance 1)

$$T = \frac{\sqrt{n}\bar{Z}}{S_Z} = \frac{\sqrt{n}\bar{X}(\hat{\mu} - \mu)}{S_Z} = \frac{\hat{\mu} - \mu}{S_{\hat{\mu}}}, \tag{5}$$

where $S_{\hat{\mu}}^2 = S_Z^2/(n\bar{X}^2)$ is a variance estimator for $\hat{\mu}$.

If the Z_j's were *assumed* to be normally distributed, and the *true* value of μ was used instead of $\hat{\mu}$ in the estimator of the variance of Z in (4), then T would have a Student's t distribution with $n-1$ degrees of freedom. Under this assumption, one could construct a confidence interval for μ based on T^2 being less or equal to the appropriate quantile of an F distribution with 1 and $n-1$ degrees of freedom, by solving this quadratic inequality in μ. (Note that the denominator of T also involves μ.) This method is due to Fieller [1954] and has has been used in the context of regenerative simulation by Crane and Iglehart [1974] and Iglehart [1975]. A fact which is often overlooked in the discussion of Fieller's intervals is that the resulting confidence "interval" can be unbounded. Since the method consists of finding all values of μ such that a quadratic function of μ is less than a constant, the resulting solution is usually an interval, but it can also be the complement of a bounded interval, i.e., $(-\infty, a) \cup (b, \infty)$, or the interval $(-\infty, \infty)$. The latter cases happen with relatively high probability when $E[X]$ is close to 0 and n is small, see e.g., Hwang [1995]. We did not include Fieller's method in our experiments for the following reasons: (a) The bootstrap methods introduced in the next section can be adapted to more general situations than estimating the ratio of two expectations, e.g., they work for the derivative of a ratio of expectations (see Section 3); (b) in simulation studies reported by Iglehart [1975], Fieller's method gave results comparable to, and never significantly better, than the classical method which will be presented next; and (c) Fieller's method is exact only when the observations are normally distributed, an assumption rarely tenable in practice. We are especially interested in small sample sizes n and non-normal observations. No method can be fully reliable in this context, so our aim is not to propose a method with guaranteed good coverage probability, but simply to see if we can improve over the existing methods.

The *classical approach* consists of assuming that T has a standard normal distribution and constructing a confidence interval on μ in the usual way (see, e.g., Fishman [1973], Iglehart [1975], and Law and Kelton [1991] for details about its application in the context of regenerative simulation). Two drawbacks with this approach are: (i) $\hat{\mu}$ is a *biased* estimator of μ and (ii) the distribution of T is not normal (or Student's t) even if the Z_j's are normally distributed. These problems vanish asymptotically, as $n \to \infty$, because (a) both the bias and the variance of $\hat{\mu}$ are typically $O(1/n)$, so the bias is asymptotically negligible compared with the standard deviation, and (b) T is asymptotically normal due to the *delta method* version of the central-limit theorem. However, these are often important sources of error for small or moderate n since asymmetry in the distributions of X and Y and/or the covariance between them leads to asymmetry in the distribution of T.

Against (i), several bias-reduction methods have been proposed in the literature, such as the jackknife, the Beale and Tin estimators, etc. (see, e.g., Iglehart [1975], Law and Kelton [1991], and Shao and Tu [1995]). They typically reduce the bias from $O(1/n)$ to $O(1/n^2)$, but usually at the price of increasing the variance, so they do not necessarily improve the situation. Moreover, the bias-corrected estimators also suffer from having an asymmetric distribution which is the main source of the coverage error of confidence intervals based on these estimators.

Our aim in this paper is to examine how item (ii) can be addressed via the bootstrap approach [Efron 1979; Léger et al. 1992; Hall 1992; Efron and Tibshirani 1993]. The basic idea is to use the observations (X_i, Y_i) to estimate F, the joint distribution of X and Y, in order to simulate observations from that estimated distribution to estimate the distribution of T and therefore its quantiles which are used in constructing a confidence interval. The usual estimator of F is $\hat{F}_n$, the empirical distribution of (X_i, Y_i), $i = 1, \ldots, n$. Generating observations from $\hat{F}_n$ is equivalent to sampling n pairs of observations *with replacement* from the n pairs (X_i, Y_i).

Some of the uses of bootstrap methods in the context of computer simulation are surveyed by Cheng [1995]. An empirical study related to ours is that of Shiue, Xu, and Rea [1993]. These authors propose a heuristic way of constructing a bootstrap confidence interval for a ratio of expectations. Their method, based on a double bootstrap, is more complicated than the ones proposed here and its validity is not discussed. A general framework for constructing bootstrap confidence intervals for a smooth function of several expectations is given by Hall [1992], together with the relevant asymptotic theory. Our proofs of validity are based on this general theory. Since asymptotics only tell part of the story, empirical experiments with small and moderate sample sizes are also needed to assess how *practical* is the theory. We do perform such experiments in this paper.

We assume here that the goal is to provide one-sided or two-sided confidence intervals for a parameter, usually the ratio of two expectations, based on the available observations. The theory and our simulation results show that the bootstrap-t confidence intervals have better coverage properties than other competitors from the literature. This does not mean that the actual coverage probability is always close to the nominal one. Finding a method that performs well universally appears utopian. In particular, Hwang [1995] shows that if a confidence interval for μ, the ratio of expectations (1), is of finite length with probability 1, then the minimum coverage probability for a fixed n, over all possible values of $E[Y]$ and $E[X] \neq 0$, is 0. The problem occurs when $E[X]$ goes to 0. All the confidence intervals in this paper have finite length with probability 1 and could have arbitrarily small coverage probability for a given sample size. Fieller's method is exact when the observations are normally distributed, but the probability that the confidence interval is the complement of an interval or the whole real line increases as $E[X]$ goes to 0, and (as mentioned earlier), this method does not perform particularly well when the observations are non-normal.

We give an overview of the paper. In Section 2, we introduce different variants of the bootstrap in the context of the ratio estimation problem, explain how one-sided and two-sided confidence intervals are computed in each case, and give important theoretical properties of the bootstrap intervals. In particular, the coverage errors of the different one-sided and two-sided confidence intervals are studied. For many intervals, including the classical ones, the coverage error of the one-sided interval is an order of magnitude larger than for the two-sided interval. On the other hand, the coverage error of the bootstrap-t method is of the same order for one-sided and two-sided intervals, making it the preferred method. The improvement in one-sided intervals is of practical importance. In many cases, we are interested in upper bounds for the ratio, but not necessarily in lower bounds, such as when we want to make sure that the average cost is lower than a given constant. In these cases, one-sided intervals are more appropriate than two-sided intervals.

In certain situations, the numerator and denominator of the ratio are estimated by *independent* simulations. For example, if the cost process $C(\cdot)$ is influenced mostly by rare events, a sensible strategy may use some form of importance sampling to estimate the numerator $E[Y]$ and regular simulation to estimate (independently) $E[X]$ (see, e.g., Chang, Heidelberger, Juneja, and Shahabuddin [1994], L'Ecuyer and Champoux [1996], and our Example 3 for illustrations and justifications). We show how to adequately construct bootstrap confidence intervals in this independent case as well.

In Section 3, we explain how to construct bootstrap confidence intervals for the *derivative* of a ratio of expectations with respect to a real-valued parameter of the underlying probability measure of the random vector (X, Y). Derivative estimation can be useful for sensitivity analysis or for optimization (see, e.g., L'Ecuyer [1991], L'Ecuyer and Yin [1998], Rubinstein and Shapiro [1993] and other references therein). We consider a one-dimensional parameter for ease of explanation, but one can easily generalize our development to estimating the *gradient* with respect to parameter vectors. Glynn, L'Ecuyer, and Adès [1991] have proposed gradient estimators for this situation, including some estimators that incorporate bias-reduction methods. They suggested computing confidence intervals based on the asymptotic normality of the estimator. Again, some bootstrap methods will be shown to improve over the classical confidence intervals. As in Glynn, L'Ecuyer, and Adès [1991], we express the derivative as a function of four expectations, which enables us to apply the general theory of Hall [1992]. The development of Section 3 illustrates how the methodology generalizes to functionals other than a ratio of means.

Section 4 provides numerical illustrations and summarizes the results of our numerical investigations. Our numerical experiments are with a *single queue* where we estimate various steady-state means, such as the average sojourn time per customer, the proportion of customers whose waiting time in the queue exceeds some fixed threshold, and the loss ratio in a queue with finite waiting room (a model of the *leaky bucket* algorithm used in telecommunications). In the latter case, importance sampling is used for the numerator, and the variables X and Y are independent. We compute both one-sided and two-sided confidence intervals. A conclusion follows in Section 5. Additional numerical results can be found in Choquet [1997].

2. BOOTSTRAP FOR RATIO ESTIMATION

Let $J_n(x;F)$ be the exact distribution function of the statistic T defined in (5) computed from n i.i.d. pairs (X_i, Y_i) when the joint distribution of the pair (X, Y) is F. A two-sided confidence interval for μ, at level $1-\alpha_1-\alpha_2$, is given by

$$(\hat{\mu} - J_n^{-1}(1-\alpha_1;F)S_{\hat{\mu}},\ \hat{\mu} - J_n^{-1}(\alpha_2;F)S_{\hat{\mu}}). \tag{6}$$

This confidence interval is *exact* if the distribution function $J_n(\,\cdot\,;F)$ is continuous. Usually, $\alpha_1 = \alpha_2$ so that the interval is said to be equal-tailed. Right and left one-sided intervals at level $1-\alpha$ are

$$(-\infty,\ \hat{\mu} - J_n^{-1}(\alpha;F)S_{\hat{\mu}}) \tag{7}$$

and

$$(\hat{\mu} - J_n^{-1}(1-\alpha;F)S_{\hat{\mu}},\ \infty), \tag{8}$$

respectively. These intervals can of course be truncated appropriately if μ is known to lie in a specific area; e.g., if it is known to be non-negative.

The practical difficulty here is that, apart from a few exceptions, $J_n(\,\cdot\,;F)$, which depends on F, is unknown. If generating observations from F was cheap, then one could conceivably perform a simulation large enough to obtain an estimator of $J_n(\,\cdot\,;F)$ as precise as desired, by simulating several independent copies of T, leading to a confidence interval whose coverage error, the difference between the actual and the nominal coverage probability, could be arbitrarily small. But this would require much more than n observations from F and we want to assume here that only n pairs are available, so this solution is unacceptable. What we can do instead is to replace F by $\hat{F}_n$, the empirical distribution of the pairs (X_i, Y_i), $i = 1, \ldots, n$, in the formulas (6), (7), and (8). This is the *nonparametric bootstrap-t* solution (also known as the *percentile-t bootstrap* or the *studentized bootstrap*). For instance, the bootstrap-t equal-tailed $1 - 2\alpha$ two-sided confidence interval for μ is given by

$$(\hat{\mu} - J_n^{-1}(1-\alpha; \hat{F}_n) S_{\hat{\mu}},\ \hat{\mu} - J_n^{-1}(\alpha; \hat{F}_n) S_{\hat{\mu}}). \tag{9}$$

Here $J_n(x, \hat{F}_n)$ is the distribution function of the statistic T when it is computed from a sample of n pairs of i.i.d. observations (X^*, Y^*) from $\hat{F}_n$, i.e., it is the distribution function of

$$T^* = \frac{\hat{\mu}^* - \hat{\mu}}{S_{\hat{\mu}^*}}, \tag{10}$$

where $\hat{\mu}^*$ and $S^2_{\hat{\mu}^*}$ are the mean and variance (counterparts of $\hat{\mu}$ and $S^2_{\hat{\mu}}$) obtained from a random resample of size n drawn with replacement from the original sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Now, another difficulty is that $J_n(\cdot\,; \hat{F}_n)$ is generally unknown. However, it can easily and cheaply be approximated with as much precision as desired by simulating B samples of size n from $\hat{F}_n$ and using $\hat{J}_{n;B}(\cdot\,; \hat{F}_n)$, the empirical distribution function of the B copies of T^* computed on these samples, as an approximation. This simulation is fast since it only involves computing means, variances, and covariances on B samples obtained *with replacement* from the original sample. The cost of the bootstrap is thus independent of the cost of generating observations from F. Note that even if one could use $B = \infty$ samples, the resulting confidence intervals would still be approximate rather than exact because the observations used in computing it come from the estimator $\hat{F}_n$ rather than the true distribution F.

In practice, the bootstrap-t algorithm works as follows.

ALGORITHM 1.
From $(X_1, Y_1), \ldots, (X_n, Y_n)$, compute $\hat{\mu}$ and $S_{\hat{\mu}}$;
For $k = 1, \ldots, B$:
Resample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ with replacement from (X_1, Y_1), $\ldots$, (X_n, Y_n);

Compute $\hat{\mu}^*$ and $S_{\hat{\mu}^*}$ from $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$;

Let $P_k = (\hat{\mu}^* - \hat{\mu})/S_{\hat{\mu}^*}$;

Sort $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(B)}$

Return the bootstrap-t $1 - 2\alpha$ two-sided confidence interval

$$(\hat{\mu} - P_{(\lceil B(1-\alpha)\rceil)} S_{\hat{\mu}},\ \hat{\mu} - P_{(\lceil B\alpha\rceil)} S_{\hat{\mu}}), \tag{11}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x.

To construct bootstrap-t procedures, one needs an estimator of the variance of the statistic under study, here $\hat{\mu}$. The estimator $S_{\hat{\mu}}$ was obtained by the well-known *delta method* (see, e.g., Serfling [1980]). Other variance estimators, for instance the jackknife, could be used leading to different versions of bootstrap-t confidence intervals. Another example would be a bootstrap estimator of the variance, leading to a *double* bootstrap procedure, that is for each bootstrap sample, a further bootstrap sample of size B_2 would be used to compute bootstrap estimates $\hat{\mu}^{**}$ whose sample variance is the bootstrap estimate of the variance of $\hat{\mu}^*$. This bootstrap-t confidence interval, using a bootstrap estimate of variance, is the "crude bootstrap" method introduced by Shiue, Xu, and Rea [1993]. This requires much more computations than necessary since $S_{\hat{\mu}}^2$ is already a good variance estimator.

In some problems, it may be difficult to construct a variance estimator (or too costly to use a double bootstrap). A simple example is when the parameter of interest is the median. No simple variance estimator of the median exists. Moreover, the (ordinary) jackknife does not work for the median, unlike the bootstrap. In such difficult situations, it is useful to be able to construct confidence intervals without the need of a variance estimator. Consider $K_n(x, F)$, the distribution function of $\hat{\mu} - \mu$. An exact $1 - 2\alpha$ confidence interval for μ would be given by

$$(\hat{\mu} - K_n^{-1}(1-\alpha; F),\ \hat{\mu} - K_n^{-1}(\alpha; F)). \tag{12}$$

Again we estimate the quantiles $K_n^{-1}(\,\cdot\,; F)$ by the bootstrap quantiles $K_n^{-1}(\,\cdot\,; \hat{F}_n)$ of $\hat{\mu}^* - \hat{\mu}$ leading to the following bootstrap confidence interval

$$(\hat{\mu} - K_n^{-1}(1-\alpha; \hat{F}_n),\ \hat{\mu} - K_n^{-1}(\alpha; \hat{F}_n)). \tag{13}$$

This confidence interval goes under many different names such as the *hybrid method*

[Hall 1988; Shao and Tu 1995], the *basic bootstrap method* [Davison and Hinkley 1997], or the *percentile method* [Hall 1992]. This last name is confusing since for most authors it refers to a different method. We shall call the interval (13), the *basic bootstrap* method. In practice, an approximation of these bootstrap confidence intervals would be computed as follows. In Algorithm 1, for each bootstrap sample, compute $Q_k = \hat{\mu}^* - \hat{\mu},\ k = 1, \ldots, B$, sort them to obtain $Q_{(1)} \le Q_{(2)} \le \ldots \le Q_{(B)}$ so that the approximation to the basic bootstrap confidence interval (13) is given by

$$(\hat{\mu} - Q_{(\lceil B(1-\alpha)\rceil)},\ \hat{\mu} - Q_{(\lceil B(\alpha)\rceil)}). \tag{14}$$

The problem of the bias of $\hat{\mu}$ has been discussed before. It was mentioned that the jackknife can be used to estimate μ with less bias. The jackknife estimator of μ is

$$\hat{\mu}_J = n\hat{\mu} - \frac{(n-1)}{n} \sum_{i=1}^{n} \hat{\mu}_{-i}, \tag{15}$$

where $\hat{\mu}_{-i} = \sum_{j \neq i} Y_j / \sum_{j \neq i} X_j$. A jackknife estimator of the variance of $\hat{\mu}_J$ is

$$\hat{\sigma}_J^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\mu}_{-i} - \hat{\mu}_J)^2 \tag{16}$$

and a jackknife two-sided $1 - 2\alpha$ confidence interval for μ based on the point estimator $\hat{\mu}_J$ is

$$(\hat{\mu}_J - z_{1-\alpha}\hat{\sigma}_J,\ \hat{\mu}_J - z_\alpha \hat{\sigma}_J), \tag{17}$$

where z_α is the α^{th} quantile of the standard normal distribution. This symmetric confidence interval (with respect to $\hat{\mu}_J$) does not reflect the usual asymmetry of the distribution of $\hat{\mu}_J$. Instead, one could use a bootstrap-t confidence interval based on the bootstrap distribution of $(\hat{\mu}_J - \mu)/\hat{\sigma}_J$, i.e., the distribution of

$$T_J^* = \frac{\hat{\mu}_J^* - \hat{\mu}}{\hat{\sigma}_J^*}, \tag{18}$$

where $\hat{\mu}_J^*$ and $\hat{\sigma}_J^*$ are computed on each bootstrap sample. We call the resulting confidence interval the *bootstrap-t jackknife* interval. To compute this confidence interval, one would generate B bootstrap samples of size n with replacement from $\{(X_i, Y_i) : i = 1, \ldots, n\}$ as in Algorithm 1. For the kth set of bootstrap variates

$\{(X_i^*, Y_i^*) : i = 1, \ldots, n\}$, one computes

$$\hat{\mu}^*_{J,k} = n\hat{\mu}^*_k - \frac{(n-1)}{n}\sum_{i=1}^{n} \hat{\mu}^*_{-i,k}, \tag{19}$$

where

$$\hat{\mu}^*_k = \frac{\sum_{i=1}^n Y_i^*}{\sum_{i=1}^n X_i^*} \text{ and } \hat{\mu}^*_{-i,k} = \frac{\sum_{j\neq i} Y_j^*}{\sum_{j\neq i} X_j^*}, \quad i = 1, \ldots, n \tag{20}$$

in a way that parallels (15). Similarly, $(\hat{\sigma}^*_{J,k})^2$ is computed as in (16) from the corresponding quantities $\hat{\mu}^*_{J,k}$ and $\hat{\mu}^*_{-i,k}$, $i = 1, \ldots, n$. Using these quantities, one computes

$$R_k = \frac{\hat{\mu}^*_{J,k} - \hat{\mu}}{\hat{\sigma}^*_{J,k}}, \qquad k = 1, \ldots, B, \tag{21}$$

and the corresponding order statistics

$$R_{(1)} \leq R_{(2)} \leq \ldots \leq R_{(B)} \tag{22}$$

so that the bootstrap-t jackknife confidence interval is approximated by

$$(\hat{\mu}_J - R_{(\lceil B(1-\alpha)\rceil)}\hat{\sigma}_J,\ \hat{\mu}_J - R_{(\lceil B\alpha\rceil)}\hat{\sigma}_J). \tag{23}$$

The key in applying the bootstrap is the ability to estimate the model that generates the observations. So far, we have considered the simple model where all (pairs of) observations are i.i.d. from a distribution F so that the only unknown in this model is the distribution F. The bootstrap then consists of generating i.i.d. observations from $\hat{F}_n$, an estimate of F constructed from the original observations. To estimate the mean sojourn time in a queue, it would also be possible to fix n, simulate n customers, and take their mean sojourn time $\bar{C}$ as the estimate. In this case, the sojourn times C_j's of individual customers are not i.i.d. and the dependence structure between the different observations is complicated. There are bootstrap methods for such observations. Most of them are based on the idea of resampling blocks of consecutive observations. These methods are much more complicated, require more care in their application, and further research is still needed. The interested reader should consult Davison and Hinkley [1997] and references therein. Needless to say, the structure of a regenerative process, which provides i.i.d. observations, greatly simplifies the statistical analysis.

A practical example of a slightly different situation which can still easily be analysed with a bootstrap method is the case when separate (independent) simulations are being used to estimate $E(X)$ and $E(Y)$, as when importance sampling is used to estimate the numerator (see Example 3 in the next section). The statistical model for this situation is as follows. Let $X_1, \ldots, X_n$ be i.i.d. from F_X and $Y_1, \ldots, Y_n$ be i.i.d. from F_Y, independently from the X_i's. The estimator of μ remains $\hat{\mu}$, but the delta method variance estimator becomes

$$\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{n\bar{X}}(S^2_Y + \hat{\mu}^2 S^2_X). \tag{24}$$

Note that there is no term S_{XY} in $\hat{\sigma}^2_{\hat{\mu}}$ unlike in $S^2_{\hat{\mu}}$ since, by design, X and Y are independent. The bootstrap algorithm is modified as follows. Instead of resampling pairs (X_i^*, Y_i^*) with replacement from the n pairs (X_i, Y_i), the X_i^*'s are resampled with replacement from the X_i's, and the Y_j^*'s are independently resampled from the Y_j's. The rest of the algorithm remains the same except that $\hat{\sigma}_{\hat{\mu}}$ and $\hat{\sigma}^*_{\hat{\mu}}$ are used instead of $S_{\hat{\mu}}$ and $S^*_{\hat{\mu}}$.

We now study the asymptotic behavior of the bootstrap confidence intervals presented so far. The framework for the theoretical comparison is taken from Hall [1988, 1992]. Hall studies the *smooth function of means* model whereby the observations are d-dimensional i.i.d. vectors $V = (V_1, \ldots, V_d)$ with mean η and the parameter of interest is $\mu = g(\eta)$ where g is a smooth function. The estimator is $\hat{\mu} = g(\bar{V})$, where $\bar{V}$ is the mean of a sample of i.i.d. vectors V_i. It is assumed that the asymptotic variance of $\hat{\mu}$ is $n^{-1}\sigma^2 = n^{-1}h^2(\eta)$, for a fixed known function h. That the asymptotic variance is a function of the mean η is not much of a problem since the vector V can be augmented. For instance, the asymptotic variance of the univariate mean $\bar{Y}$ is $n^{-1}\sigma^2 = n^{-1}[E(Y^2) - (E(Y))^2]$ and so we just have to consider the vector $V = (Y, Y^2)$. Clearly, a ratio of means is part of this model.

Hall provides asymptotic series for the coverage error of different bootstrap confidence intervals by first studying Edgeworth expansions of the bootstrap distribution function of $(\hat{\mu}^* - \hat{\mu})/h(\bar{V})$ or of the studentized version $(\hat{\mu}^* - \hat{\mu})/h(\bar{V}^*)$, and then Cornish-Fisher expansions of their quantiles. We shall not be concerned with the technical regularity conditions here; for a rigorous treatment, see Sections 5.2 and 5.3 of Hall [1992]. The coverage error of a confidence interval is the difference between the actual coverage probability and the claimed value $1 - \alpha$.

PROPOSITION 1. *Consider the* $1-\alpha$ *one-sided and two-sided basic bootstrap confidence intervals from (13) given by* $\mathcal{I}_1(1-\alpha) = \{\mu : \mu \leq \hat{\mu} - K_n^{-1}(\alpha, \hat{F}_n)\}$ *and* $\mathcal{I}_2(1-\alpha) = \{\mu : \hat{\mu} - K_n^{-1}(1-\alpha/2, \hat{F}_n) \leq \mu \leq \hat{\mu} - K_n^{-1}(\alpha/2, \hat{F}_n)\}$ *and the* $1-\alpha$ *one-sided and two-sided bootstrap-t confidence intervals from (9) given by* $\mathcal{J}_1(1-\alpha) = \{\mu : \mu \leq \hat{\mu} - J_n^{-1}(\alpha, \hat{F}_n)S_{\hat{\mu}}\}$ *and* $\mathcal{J}_2(1-\alpha) = \{\mu : \hat{\mu} - J_n^{-1}(1-\alpha/2, \hat{F}_n)S_{\hat{\mu}} \leq \mu \leq \hat{\mu} - J_n^{-1}(\alpha/2, \hat{F}_n)S_{\hat{\mu}}\}$. *Under appropriate regularity conditions on the joint distribution of the pair* (X, Y),

$$\begin{aligned}
\Pr\{\mu \in \mathcal{I}_1(1-\alpha)\} - (1-\alpha) &= n^{-1/2}u_1(z_{1-\alpha})\phi(z_{1-\alpha}) + O(n^{-1}) \\
\Pr\{\mu \in \mathcal{I}_2(1-\alpha)\} - (1-\alpha) &= n^{-1}u_2(z_{1-\alpha})\phi(z_{1-\alpha}) + O(n^{-3/2}) \\
\Pr\{\mu \in \mathcal{J}_1(1-\alpha)\} - (1-\alpha) &= n^{-1}u_3(z_{1-\alpha})\phi(z_{1-\alpha}) + O(n^{-3/2}) \\
\Pr\{\mu \in \mathcal{J}_2(1-\alpha)\} - (1-\alpha) &= n^{-1}u_4(z_{1-\alpha})\phi(z_{1-\alpha}) + O(n^{-3/2}),
\end{aligned}$$

where u_1 *through* u_4 *are polynomials which depend on the moments of the distribution,* u_1 *is an even polynomial, and* $\phi(\cdot)$ *is the standard normal density.*

The Proposition describes the coverage error of the actual bootstrap confidence intervals, i.e., those based on $B = \infty$. In practice, the number of bootstrap samples used for a confidence interval, B, is about 1000. Hall [1986] has shown that the choice of B has very little impact on the *coverage*, but that a small value of B could lead to a larger confidence interval. Note that the length of the confidence interval is random due to the fact that $B < \infty$ and that the variance of the length increases when B decreases. He has also shown that if $\alpha = i/(B+1)$, the coverage probability using the quantiles from the simulated distribution of T^*, $\hat{J}_{n;B}(\,\cdot\,; \hat{F}_n)$, is identical to that using the exact distribution $J_n(\,\cdot\,; \hat{F}_n)$. Let us now look at the implications of the proposition.

First, note that the classical intervals have expansions similar to those of the basic bootstrap intervals. The result says that the coverage error of equal-tailed two-sided confidence intervals is of order n^{-1} for the classical, basic bootstrap, and bootstrap-t confidence intervals. More importantly, it says that the coverage error of one-sided confidence intervals is of order n^{-1} for the bootstrap-t compared to only $n^{-1/2}$ for the basic bootstrap and the classical intervals. So the improvement of the bootstrap-t confidence intervals over the classical interval is large, especially in one-sided intervals which, as we argued before, should be used more often in practice. The reason that the coverage error of equal-tailed two-sided confidence intervals based on the basic bootstrap or on the classical method is smaller than for one-sided intervals is that the errors of each side partially cancel. More precisely, the coverage probability of the two-sided intervals is the difference between the coverage probability of the $1-\alpha/2$ and $\alpha/2$ one-sided intervals. The $n^{-1/2}$ term in the coverage error of the two-sided interval is the difference between the corresponding terms of the one-sided coverage errors. But since both u_1 and ϕ are even functions, the leading term of the one-sided coverage error is an even function, so the difference between this function applied at $z_{1-\alpha}$ and at z_α is zero and the leading term of the two-sided interval is thus of order n^{-1}.

It is interesting to note that the basic bootstrap, which does not require an estimator of the variance of $\hat{\mu}$ has coverage errors of the same order as those of the classical intervals which do require an estimate of the variance of $\hat{\mu}$, but rely on the symmetric quantiles of the standard normal. Proposition 1 also applies to the modified bootstrap algorithm in the case of independent X and Y. The adequacy of these asymptotic developments in small samples will be checked in Section 4.

Other bootstrap confidence intervals have been introduced in the literature, such as the percentile method and improvements of it like the BC, the BCa, and the ABC. Interested readers should consult Efron and Tibshirani [1993], Hall [1992], and references therein. In particular, the behavior of the coverage error of the percentile and BC methods are like that of the basic bootstrap interval whereas that of the BCa and ABC methods follows that of the bootstrap-t interval.

3. BOOTSTRAP CONFIDENCE INTERVALS FOR THE DERIVATIVE OF A RATIO

In this section, we suppose that the probability space is parameterized by a real-valued parameter $\theta \in \Theta \subseteq \mathbb{R}$. Let P_θ and E_θ be the probability measure and the expectation operator, respectively, associated with the parameter value θ. For each $\theta \in \Theta$, the ratio of expectations can be expressed as

$$\mu(\theta) = \frac{u(\theta)}{\ell(\theta)} = \frac{E_\theta[Y]}{E_\theta[X]}. \tag{25}$$

The derivative with respect to θ, assuming that it exists, is

$$\mu'(\theta) = \frac{\ell(\theta)u'(\theta) - u(\theta)\ell'(\theta)}{\ell^2(\theta)} = \frac{u'(\theta) - \mu(\theta)\ell'(\theta)}{\ell(\theta)}. \tag{26}$$

Suppose one can obtain n i.i.d. replicates $V_i = (V_{1i}, V_{2i}, V_{3i}, V_{4i}) = (X_i, Y_i, X'_i, Y'_i)$, $i = 1, \ldots, n$, of a random vector $V = (X, Y, X', Y')$ whose components have finite second moments, and such that $E_\theta(X, Y, X', Y') = (\ell(\theta), u(\theta), \ell'(\theta), u'(\theta))$ and $E_\theta[X] > 0$ (here, X' and Y' are not necessarily the derivatives of X and Y). Two possible approaches for obtaining the derivative estimators X' and Y' are infinitesimal perturbation analysis and the score function method [Glasserman 1991; L'Ecuyer 1990; Rubinstein and Shapiro 1993]. A strongly consistent estimator for $\mu'(\theta)$ is then

$$\hat{\mu}' = \frac{\bar{Y}' - \hat{\mu}\bar{X}'}{\bar{X}} \tag{27}$$

where

$$(\bar{X}, \bar{Y}, \bar{X}', \bar{Y}') = \frac{1}{n}\sum_{i=1}^{n}(X_i, Y_i, X'_i, Y'_i).$$

The delta method yields the following estimator of the variance of $\hat{\mu}'$:

$$S^2_{\hat{\mu}'} = \frac{1}{n\bar{X}^2}\sum_{i=1}^{n}(\hat{W}_i - (\bar{X}'/\bar{X})\hat{Z}_i)^2,$$

where

$$\begin{aligned} \hat{Z}_i &= Y_i - \hat{\mu}X_i, \\ \hat{W}_i &= Y'_i - \hat{\mu}X'_i - \hat{\mu}'X_i. \end{aligned}$$

See Glynn, L'Ecuyer, and Adès [1991] for the details.

To apply the bootstrap, one resamples B batches of n vectors $V_i^* = (X_i^*, Y_i^*, X_i'^*, Y_i'^*)$, $i = 1, \ldots, n$ with replacement from $\{V_i = (X_i, Y_i, X'_i, Y'_i),\ i = 1, \ldots, n\}$, and computes $\hat{\mu}'^*$ and $S^2_{\hat{\mu}'^*}$ to obtain an approximation of the distribution of

$$T'^* = \frac{\hat{\mu}'^* - \hat{\mu}'}{S_{\hat{\mu}'^*}}$$

to construct bootstrap-t confidence intervals as in the previous section, or of the distribution of $\hat{\mu}'^* - \hat{\mu}'$ to construct basic bootstrap confidence intervals. The classical intervals are like the bootstrap-t intervals, except that the quantiles of the standard normal distribution are used instead of the bootstrap quantiles.

Since this problem also fits into the framework of smooth functions of means, the conclusion of Proposition 1 (with the obvious adaptation) holds here as well.

4. EXAMPLES AND NUMERICAL EXPERIMENTS

EXAMPLE 1. We consider an $M/M/1$ queue (single-server queue with i.i.d. exponential interarrival times and i.i.d. exponential service times) with arrival rate $\lambda = 1$ and service rate $1/\theta$, $0 < \theta < 1$. The load factor is θ. We are interested in the mean sojourn time in the system per customer, in steady-state, where the *sojourn time* of a customer is the sum of its waiting time and service time. For this simple model, the steady-state sojourn time $\mu = \mu(\theta)$ and its derivative are easily computed in closed form: One has $\mu(\theta) = \theta/(1-\theta)$ and $\mu'(\theta) = \theta/(1-\theta)^2$. Simulation is therefore not needed in this case. However, the availability of the exact solution makes this example a good testbed for comparing different methods of computing confidence intervals. It is a standard example, widely used for the numerical experimentation of simulation methodologies, and it is well-known that the classical confidence intervals for $\mu(\theta)$ do not have good coverage, especially when θ gets close to 1 [Iglehart 1975; Law and Kelton 1991]. Those for $\mu'(\theta)$ are even worse [Glynn et al. 1991]. These authors have also experimented with different variants of the jackknife on this $M/M/1$ example, and the improvements were *modest* at best. This simple example therefore provides a particularly difficult problem in terms of point and interval estimation, and it is interesting to see if the bootstrap can improve over the existing methods.

Let $C_0 = 0$ and C_j the sojourn time of the jth customer, for $j \geq 1$. These C_j obey the Lindley Equation

$$C_j = \max(0, C_{j-1} - \nu_j) + \theta\zeta_j,$$

where $\{\nu_1, \zeta_1, \nu_2, \zeta_2, \ldots\}$ are i.i.d. exponential random variables with mean 1. Here, $\theta\zeta_j$ represents the service time of the jth customer, whereas ν_j is the interarrival time between customers $j-1$ and j. Let $\tau_1 < \tau_2 < \cdots$ be the successive indices of the customers whose waiting time is zero: $\tau_0 = 0$ and $\tau_i = \min\{j > \tau_{i-1} : C_{j-1} - \nu_j \leq 0\}$. These τ_i are regeneration points for the discrete-time process

$\{C_j,\, j \geq 1\}$: They mark the beginning of the regenerative cycles. We thus define

$$X_i = \tau_{i+1} - \tau_i, \qquad Y_i = \sum_{j=\tau_i}^{\tau_{i+1}-1} C_j,$$

which are the number of customers and the total sojourn time, respectively, in the ith cycle. Our gradient estimators X_i' and Y_i' are obtained via the score function (or likelihood ratio) method, as explained by L'Ecuyer [1990] and Glynn, L'Ecuyer, and Adès [1991].

We computed 95% right one-sided, 95% left one-sided, and 90% two-sided confidence intervals with several different methods, based on n regenerative cycles for $n = 64, 128, 256, 512$, and for several values of θ. These three types of intervals have the form $[a, \infty)$, $[0, b]$, and $[a, b]$, respectively, where $a \geq 0$. All the intervals for a given n were computed from the *same* simulations. For $n = 128$, we added 64 new cycles to those already done with $n = 64$, and so on. There is thus strong dependence between the computed intervals and coverages. For the bootstrap intervals, $B = 1000$ bootstrap resamples from the empirical distribution were taken to estimate the distribution of the relevant statistic.

The experiment was repeated $N = 400$ times, thus obtaining 400 confidence intervals of each type, and the true coverage probability was estimated in each case. Tables 1 and 2 give the observed coverages of the confidence intervals for μ when the load factor is $\theta = 0.5$ and 0.9, respectively. Results for $\theta = 0.2$ and 0.8 can be found in Choquet [1997].

Consider the two-sided confidence intervals first. Even though the size of the coverage error for all intervals is theoretically of the same order $O(n^{-1})$, we see that the intervals based on the bootstrap-t constantly have better coverage than the other intervals, except for the easy case $\theta = 0.5$ and $n = 512$, where all intervals perform well. When θ increases towards 1, all confidence intervals become less reliable (the model and all estimators become unstable, a well-known fact), but the improvement in coverage of the bootstrap-t intervals over the classical ones also becomes more substantial. For example, for $\theta = 0.9$ and $n = 512$, the (estimated) coverage of the two-sided 90% bootstrap-t confidence interval is 85% whereas the classical and jackknife intervals cover with (estimated) probabilities of 70% and 73%, respectively. On the other hand, when θ decreases towards 0, the problem becomes easier and all confidence intervals perform well. Of course, the improved coverages of the bootstrap-t intervals is accompanied by wider intervals. The bootstrap-t jackknife intervals are also larger than the usual bootstrap-t for smaller values of n. Wider intervals might be viewed as a negative feature from a practical point of view, but one has to keep in mind that the reason they perform as well is that the bootstrap-t does a better job at estimating the asymmetric distribution of the statistics T or T_J (and their counterparts in the derivative case) as can be seen from the one-sided coverage. One has to go further away in the tails to have good coverage, thereby increasing the length of the interval. The length of competing intervals should only be an issue whenever the coverages are similar.

For one-sided intervals, as expected from Proposition 1, the bootstrap-t (as well as the bootstrap-t jackknife intervals) have much better coverage than the other intervals. The classical, jackknife, and basic bootstrap one-sided intervals tend to overcover for the left intervals and undercover for the right intervals. The two bootstrap-t right intervals also undercover, but not as much. As predicted from Proposition 1, the coverage errors of the one-sided intervals partially cancel in the two-sided intervals: Overcoverage on one-side cancels the undercoverage on the other side. The basic bootstrap intervals perform as well as the classical intervals without requiring any variance estimator.

Table 3 contains the observed coverages of the confidence intervals for $\mu'(\theta)$ at $\theta = 0.5$. Clearly, the estimators are much more noisy for $\mu'(\theta)$ than for $\mu(\theta)$, the intervals are much wider and their coverage is smaller. The situation improves when θ decreases and deteriorates when θ increases. Table 4 contains the results for $\theta = 0.9$. When $n \leq 512$, all intervals are very wide and have a bad coverage. Nevertheless, the coverage of the bootstrap-t methods is generally better than that of the other methods.

Example 2. Take the previous example but replace the sojourn time C_j by the indicator function

$$I_j = \begin{cases} 1 & \text{if } W_j > L, \\ 0 & \text{otherwise,} \end{cases}$$

where $W_j = C_j - \theta\zeta_j$ is the waiting time of customer j in the queue and $L > 0$

Table 1. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the mean sojourn time μ in an $M/M/1$ queue with load factor $\theta = 0.5$. The estimated coverage probability of the 95% left and right one-sided confidence intervals are given in the columns 95%l and 95%r, respectively. The estimated coverage probability of the 90% two-sided confidence interval is given in the column 90%c. The estimated mean length of the two-sided confidence intervals is given in the column identified Length.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	98	74	72	0.58	99	78	77	0.48
Jackknife	99	76	75	0.64	98	80	78	0.50
Basic boot	97	75	72	0.56	98	79	77	0.46
Boot-t	94	89	83	1.26	95	87	82	0.84
Boot-t Jack	98	86	84	1.16	97	85	82	0.74
	$n = 256$				$n = 512$			
Classical	98	85	83	0.38	98	89	87	0.28
Jackknife	98	86	84	0.40	98	90	88	0.28
Basic boot	98	85	83	0.38	98	89	87	0.28
Boot-t	95	91	86	0.56	93	93	86	0.34
Boot-t Jack	96	91	87	0.52	95	93	88	0.34

Table 2. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the mean sojourn time μ in an $M/M/1$ queue with load factor $\theta = 0.9$.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	95	46	41	5.66	98	55	53	5.90
Jackknife	99	57	56	9.86	99	62	61	8.32
Basic boot	89	53	42	6.28	94	59	53	6.04
Boot-t	90	82	72	13.04	94	83	77	14.78
Boot-t Jack	99	74	73	72.20	99	76	75	32.44
	$n = 256$				$n = 512$			
Classical	99	64	63	5.56	98	72	70	5.00
Jackknife	99	68	67	6.78	98	75	73	5.66
Basic boot	96	66	62	5.50	97	73	70	4.88
Boot-t	94	84	78	14.66	96	89	85	12.28
Boot-t Jack	99	79	78	18.08	97	85	82	11.02

Table 3. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the derivative of the mean sojourn time $\mu'(\theta)$ in an $M/M/1$ queue with load factor $\theta = 0.5$.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	98	44	42	4.22	100	52	52	4.40
Jackknife	99	51	50	5.58	100	57	57	5.16
Basic boot	94	48	42	3.90	98	55	53	4.62
Boot-t	90	66	56	10.92	92	79	71	13.32
Boot-t Jack	100	69	69	51.04	100	77	77	37.06
	$n = 256$				$n = 512$			
Classical	99	66	65	4.56	99	68	67	3.76
Jackknife	99	69	68	5.02	99	70	69	3.96
Basic boot	98	66	64	4.26	99	67	66	3.60
Boot-t	91	83	74	12.82	94	84	78	9.84
Boot-t Jack	99	81	80	19.10	99	80	79	11.32

Table 4. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the derivative of the mean sojourn time $\mu'(\theta)$ in an $M/M/1$ queue with load factor $\theta = 0.9$.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	95	13	8	38.44	95	22	17	47.50
Jackknife	96	24	20	115.22	97	31	28	137.70
Basic boot	92	18	10	49.16	91	25	16	65.32
Boot-t	92	29	21	178.52	92	40	32	200.04
Boot-t Jack	94	52	46	7361.26	96	52	48	5801.04
	$n = 256$				$n = 512$			
Classical	97	32	29	85.86	99	43	42	119.22
Jackknife	99	40	39	154.92	99	48	47	166.64
Basic boot	93	35	28	85.92	94	44	38	107.60
Boot-t	91	52	43	266.00	89	69	58	338.40
Boot-t Jack	100	58	58	3722.76	99	68	67	2468.88

Table 5. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the fraction of customers whose sojourn time exceeds $L = 2.303$ in an $M/M/1$ queue with load factor $\theta = 0.5$.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	99	57	56	0.086	99	68	67	0.086
Jackknife	99	58	57	0.094	99	70	69	0.092
Basic boot	98	54	52	0.084	99	66	65	0.086
Boot-t	95	81	76	0.294	95	96	91	0.276
Boot-t Jack	99	47	46	0.128	98	72	70	0.190
	$n = 256$				$n = 512$			
Classical	98	79	77	0.072	98	86	84	0.056
Jackknife	98	80	78	0.074	98	86	84	0.057
Basic boot	97	77	74	0.070	98	83	81	0.055
Boot-t	94	95	89	0.152	96	94	90	0.081
Boot-t Jack	96	90	86	0.138	96	94	90	0.080

Table 6. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the fraction of customers whose sojourn time exceeds $L = 26.013$ in an $M/M/1$ queue with load factor $\theta = 0.9$.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	96	20	16	0.036	98	31	29	0.059
Jackknife	99	21	20	0.065	99	33	32	0.083
Basic boot	92	20	12	0.035	95	30	25	0.055
Boot-t	93	23	16	0.061	94	38	32	0.115
Boot-t Jack	100	17	17	0.048	99	28	27	0.092
	$n = 256$				$n = 512$			
Classical	98	44	42	0.075	99	55	54	0.084
Jackknife	98	45	43	0.093	99	57	56	0.096
Basic boot	96	41	37	0.071	97	54	51	0.083
Boot-t	95	54	49	0.176	95	77	72	0.232
Boot-t Jack	99	39	38	0.153	99	64	63	0.313

is a fixed constant. Then, μ becomes the steady-state fraction of customers whose waiting time exceeds L. The exact value of μ here can be computed from the known steady-state distribution function F of the waiting time, that is $\mu = 1 - F(L) = \theta \exp(-L(1-\theta)/\theta)$ for $L > 0$.

We performed the same experiment as for the previous example, this time only with μ. Tables 5 and 6 report the (partial) results for $\theta = 0.5$ when the threshold L is 2.303 and $\theta = 0.9$ for a threshold of $L = 26.013$, respectively, corresponding to a probability of exceeding the threshold of 0.05 in both cases.

The results are qualitatively similar to those of Example 1. Once again, the bootstrap-t intervals are better than the classical, jackknife, and basic bootstrap intervals, especially for one-sided intervals. For $\theta = 0.5$, note that even though the two-sided intervals for the latter methods have good coverage for $n = 512$, their one-sided intervals still have relatively large coverage errors. When $\theta = 0.9$, large samples are necessary before the coverage error declines sufficiently, but the bootstrap-t nevertheless outperforms the other intervals. Since the probability that a customer's waiting time exceeds L is relatively small, there are cases where all Y_i's are 0 so that $\bar{Y}$ and $S_{\hat{\mu}}$ are also 0. In those cases, we define all intervals as the set $\{0\}$. There are also cases where only one Y_i differs from 0, and this affects the jackknife variance estimator more than the delta method variance estimator. This partly explains the poor performance of the bootstrap-t jackknife intervals in small samples.

Example 3. We now consider a TPBS/D/1/K queue, with deterministic service times equal to a constant d, a single server, and finite capacity K. Customers arriving when the system is full (one in service and $K - 1$ waiting) are lost. The arrivals follow a two-phase burst-silence (TPBS) process, where one customer arrives every τ units of time (τ is a constant) during a burst, then nobody arrives during the silence that follows, then another burst, and so on. The burst sizes are geometric random variables with mean $1/\alpha$, and the length of a silence is τ plus an exponential random variable with mean $1/\beta$. These random variables are mutually independent. This system regenerates whenever a customer arrives to an empty system. We want to estimate the steady-state fraction of time where the system is full. Again, this can be expressed as a ratio of two expectations: The expected length of time when the system is full in a cycle divided by the expected duration of a cycle.

This model is that of a source of emission of messages under the *leaky bucket* admission control policy, in a telecommunication network. The customers are packets of information and the packets are discarded *on purpose* by the network management when their number is too high, to make sure that no user sends significantly more traffic than allowed. See Nicola, Hagesteijn, and Kim [1994] for further details.

Our simulations are done using importance sampling, in exactly the same way as in Nicola, Hagesteijn, and Kim [1994]. The numerator of the ratio is estimated

Table 7. Coverage probabilities of one-sided and two-sided bootstrap confidence intervals for the fraction of time when a TPBS/D/1/K queue is full. The parameters are $\alpha = 0.1$, $\beta = 0.001$, $\tau = 4$, $d = 10$, and $K = 150$. The mean length of the two-sided confidence (column Length) is in units of 10^{-13}.

	$n = 64$				$n = 128$			
Method	95%l	95%r	90%c	Length	95%l	95%r	90%c	Length
Classical	98	91	89	4.06	96	94	90	2.86
Basic boot	99	89	88	4.16	98	92	90	2.88
Boot-t	98	91	90	4.10	96	94	90	2.86
	$n = 256$				$n = 512$			
Classical	96	94	90	2.02	97	94	91	1.42
Basic boot	98	92	90	2.02	97	93	90	1.42
Boot-t	95	94	89	2.02	96	94	90	1.42

by n regenerative cycles with a change of measure that increases β and decreases α in an appropriate way, whereas the denominator is estimated independently via n regenerative cycles without importance sampling. This strategy makes the estimation problem relatively easy even when the ratio μ is extremely small (in which case straightforward simulation is ineffective).

Again, we performed similar experiments as for the previous examples. Here, no analytic formula is available for the exact value of μ, so we estimated it via simulation with 5 million regenerative cycles, using importance sampling. These estimations can be considered as the "exact values" for our purposes. Here $\alpha = 0.1$, $\beta = 0.001$, $\tau = 4$, $d = 10$, and $K = 150$, leading to $\mu \approx 7.3 \times 10^{-13}$. Since the numerator and denominator are simulated independently, the jackknife and bootstrap-t jackknife intervals are not computed. Moreover, the bootstrap samples of X_i^*'s and Y_i^*'s are generated independently in the two bootstrap intervals. The results are in Table 7. We see that all intervals perform equally well in this case, even for small samples. The use of importance sampling, as well as the independence between the numerator and the denominator, leads to a more symmetric distribution for $\hat{\mu}$ than in the previous problems. The usual advantage of the bootstrap-t methods is not apparent here since the classical intervals are already very good for $n = 64$. Similar results were obtained by Choquet [1997] for the same model but with other arrival processes than the TPBS.

5. CONCLUSION

Simulation is often used to estimate unknown quantities. To supplement the point estimates, confidence intervals are computed to provide an idea of their uncertainty. Usually, classical intervals based on the asymptotic normality and an estimate of the asymptotic variance of the estimators are used. Alternatives include jackknife-based intervals. Although they are based on a less biased estimator, they also rely on asymptotic normality. In many cases, the distribution of the estimators is highly asymmetric, affecting the coverage of classical and jackknife intervals.

In this paper, we have shown how to adequately use the bootstrap method to construct confidence intervals for ratios of expectations and their derivatives. The bootstrap method can also be used for other more complicated functions of expectations. The basic bootstrap method provides a simple method of construction of confidence intervals with similar coverage properties to the classical and jackknife intervals, but without the necessity of computing a variance estimator for the estimator under study. While this is not an important advantage for the particular problems that we have studied, there are cases where it is difficult to come up with a good variance estimator. When a good variance estimator exists, the bootstrap-t generally provides superior confidence intervals than other available standard methods, especially for one-sided intervals. In practice, this is important since one may only be interested in a lower or upper bound for the performance measure of interest, such as for an expected cost, or the proportion of time that a queue is full. For such problems, one-sided intervals are more appropriate and the superiority of the bootstrap-t one-sided intervals makes it the tool of choice.

This does not mean that bootstrap-t confidence intervals always have good coverage properties. Their coverage error is typically smaller than that of the classical or jackknife intervals, but in some cases it is still quite important. In these situations, the estimators have a distribution which is particularly difficult to estimate with relatively small samples, and while the bootstrap-t intervals may not be satisfactory, they still outperform the other intervals in our study.

The bootstrap-t jackknife intervals illustrates how the bootstrap and the jackknife can be combined. In difficult problems with a relatively small sample size, the results are sometimes disappointing. The usual bootstrap-t should generally be the preferred method.

It would of course be highly desirable to have a practical and reliable rule of thumb for determining what sample size n suffices to ensure that the bootstrap-t method will deliver a valid confidence interval, i.e., with actual coverage close to the nominal coverage, for the problem at hand. Unfortunately, such reliable rules of thumb appear hard to obtain in this context, regardless of which method is used to compute the confidence intervals.

Bootstrap confidence intervals are more costly to compute than the classical intervals. However the cost is independent of the cost per run of the simulation itself and consists of resampling B samples of size n with replacement from the original observations and computing simple statistics, such as means, variances, and covariances on each of these samples. Usually, the number of bootstrap samples B is of the order of 1000. The relative cost of the bootstrap with respect to the cost of the simulation will be important if cycles are short and fast to simulate. But in this case, the cheap cost of the simulation will usually lead to relatively large sample sizes where the classical intervals do well. On the other hand, when the cost of simulating each cycle is large, one may have to considerably limit the sample size of the simulation and the small additional cost of computing bootstrap confidence intervals is well worth it in terms of improved coverage probabilities.

The bootstrap has been in existence for 20 years. Now well known in statistics, it would deserve to be used more often in the field of simulation.

REFERENCES

ASMUSSEN, S. 1987. *Applied Probability and Queues*. Wiley.

CHANG, C. S., HEIDELBERGER, P., JUNEJA, S., AND SHAHABUDDIN, P. 1994. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation 20*, 45–65.

CHENG, R. C. H. 1995. Bootstrap methods in computer simulation experiments. In *Proceedings of the 1995 Winter Simulation Conference* (1995), pp. 171–177. IEEE Press.

CHOQUET, D. 1997. Intervalles de confiance bootstrap pour la simulation de processus regénératifs. Master's thesis, Département d'IRO, Université de Montréal.

CRANE, M. A. AND IGLEHART, D. L. 1974. Simulating stable stochastic systems, I: General mutiserver queues. *Journal of the ACM 21*, 103–113.

DAVISON, A. C. AND HINKLEY, D. V. 1997. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

EFRON, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, 1–26.

EFRON, B. AND TIBSHIRANI, R. J. 1993. *An Introduction to the Bootstrap*. Chapman &

Hall, New York.

FIELLER, E. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B 16*, 175–185.

FISHMAN, G. S. 1973. *Concepts and Methods in Discrete Event Digital Simulation.* Wiley, New York.

FOX, B. L. AND GLYNN, P. W. 1989. Simulating discounted costs. *Management Science 35*, 11, 1297–1315.

GLASSERMAN, P. 1991. *Gradient Estimation Via Perturbation Analysis.* Kluwer Academic, Norwell, MA.

GLYNN, P. W., L'ECUYER, P., AND ADÈS, M. 1991. Gradient estimation for ratios. In B. L. NELSON, W. D. KELTON, AND G. M. CLARK Eds., *Proceedings of the 1991 Winter Simulation Conference* (Piscataway, NJ, 1991), pp. 986–993. IEEE Press.

HALL, P. 1986. On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics 14*, 1453–1462.

HALL, P. 1988. Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics 16*, 927–985.

HALL, P. 1992. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

HWANG, J. T. G. 1995. Fieller's problems and resampling techniques. *Statistica Sinica 5*, 161–171.

IGLEHART, D. L. 1975. Simulating stable stochastic systems V: Comparison of ratio estimators. *Naval Research Logistics Quarterly 22*, 553–565.

LAW, A. M. AND KELTON, W. D. 1991. *Simulation Modeling and Analysis* (Second ed.). McGraw-Hill, New York.

L'ECUYER, P. 1990. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science 36*, 11, 1364–1383.

L'ECUYER, P. 1991. An overview of derivative estimation. In B. L. NELSON, W. D. KELTON, AND G. M. CLARK Eds., *Proceedings of the 1991 Winter Simulation Conference* (Piscataway, NJ, 1991), pp. 207–217. IEEE Press.

L'ECUYER, P. AND CHAMPOUX, Y. 1996. Importance sampling for large ATM-type queueing networks. In *Proceedings of the 1996 Winter Simulation Conference* (1996), pp. 309–316. IEEE Press.

L'ECUYER, P. AND YIN, G. 1998. Budget-dependent convergence rate for stochastic approximation. *SIAM Journal on Optimization 8*, 1, 217–247.

LÉGER, C., POLITIS, D., AND ROMANO, J. 1992. Bootstrap technology and applications. *Technometrics 34*, 378–398.

MEYN, S. P. AND TWEEDIE, R. L. 1993. *Markov Chains and Stochastic Stability.* Springer-Verlag.

NICOLA, V. F., HAGESTEIJN, G. A., AND KIM, B. G. 1994. Fast simulation of the leaky bucket algorithm. In *Proceedings of the 1994 Winter Simulation Conference* (1994), pp. 266–273. IEEE Press.

RUBINSTEIN, R. Y. AND SHAPIRO, A. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* Wiley, New York.

SERFLING, R. J. 1980. *Approximation Theorems for Mathematical Statistics.* Wiley, New York.

SHAO, J. AND TU, D. 1995. *The Jackknife and Bootstrap.* Springer, New York.

SHIUE, W.-K., XU, C.-W., AND REA, C. B. 1993. Bootstrap confidence intervals for simulation outputs. *Journal of Statistical Computation and Simulation 45*, 249–255.

WOLFF, R. W. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, New York.