Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

THE ACCURACY OF A NEW CONFIDENCE INTERVAL METHOD

Johann Christoph Strelen

Rheinische Friedrich–Wilhelms–Universität Bonn
Römerstr. 164, 53117 Bonn, GERMANY

ABSTRACT

Confidence intervals for the median of estimators or other quantiles were proposed as a substitute for usual confidence intervals in terminating and steady-state simulation. This is adequate since for many estimators the median and the expectation are close together or coincide, particularly if the sample size is large. Grouping data into batches is useful for median confidence intervals. The novel confidence intervals are easy to obtain, the variance of the estimator is not used. They are well suited for correlated simulation output data, apply to functions of estimators, and in simulation they seem to be particularly accurate, namely they follow the confidence level better than other confidence intervals. This paper states their accuracy which is the difference between the nominal confidence level and the actual coverage. The accuracy is evaluated with analytical models and simulation. For the estimation of quantiles by order statistics, the new confidence intervals are exact.

1 INTRODUCTION

Modeling and simulation has been one of the modern key technologies with increasing importance for some decades. Discrete-event stochastic simulation is a very important subarea for evaluating performance and reliability. It is used in many application fields.

Due to the stochastic nature of the results, careful statistic analysis must be done for the correct interpretation of calculated values. If this is omitted, there is a significant probability of making erroneous inferences about the system under study.

In our view, many modellers and tool designers are not aware of the importance of correct statistical inference. There are many special-purpose simulation packages available which provide comfortable means for programming simulation models on a high level of abstraction, resulting in a significant decrease of programming time and in a reduction in overall project costs. As a matter of fact, many of them do not provide sufficient support for obtaining correct statistical analysis of results.

However, we are convinced that this problem is partly based on a misunderstanding that can be removed, namely the opinion that it is very difficult to achieve improvements. Clearly, statistics is a hard discipline, but it is quite easy to understand which are the crucial aspects with simulation output data, and it is easy to apply some simple techniques which are much more correct and reliable than "performing one simulation run and taking the obtained figures as the true solution".

An important feature for correct statistical analysis of simulation results are confidence intervals. Only if they are easy to apply they will be used, and only if they are statistically correct the results are credible.

In simulation, confidence intervals tend to be inaccurate, since assumptions concerning the sample and the estimator are not fulfilled literally, more precisely, the assumed confidence level is not the probability that the real parameter value lies within the calculated confidence interval. Quite often, this probability is smaller than the confidence level, the confidence interval is too optimistic. Law and Kelton (2000) performed an empirical study in order to see how (in)accurate confidence intervals may be. They found substantial errors.

We proposed (Strelen 2001 and 2002) an alternative technique for confidence intervals which we call *median confidence intervals*, and *min-max confidence intervals* which are more general: Median confidence intervals (MCI) are a special case of min-max confidence intervals (MMCI), but they seem to be particularly useful. Min-max confidence intervals are suitable for the estimation of quantiles, and they have some potential for further development.

Both can be obtained with the replication/deletion approach or with batch means. The technique is very easy to apply in steady-state or terminating (finite horizon) simulation; hopefully this motivates tool designers providing them in their tools and users applying them. Median confidence intervals can be used where other methods cannot, e.g. when the variance of an estimator does not exist, and they seem to be more accurate than usual confidence intervals.

This paper deals with the accuracy. Theorem 2 states that min-max confidence intervals are exact when quantiles are estimated with order statistics. Median confidence intervals are compared with classical confidence intervals by analytical models and by simulation studies.

In section 3 the term accuracy is specified more precisely. For the evaluation of accuracy, we present analytical models in section 4. In section 5 we present empirical studies for the accuracy; more can be found in (Strelen 2002). All results indicate a better accuracy of median confidence intervals compared to other methods. But first we explain in section 2 what min-max and median confidence intervals are.

2 MIN-MAX AND MEDIAN CONFIDENCE INTERVALS

Min-max confidence intervals (and median confidence intervals which are a special case) are obtained by means of a small number of replications, typically 4, 5, or 6. They have attractive features and some minor disadvantages compared to classical confidence intervals.

The variance of estimators is not needed for them, but this is usually a main difficulty when confidence intervals are constructed because “simulation output data are always correlated” (Law and Kelton 2000). Special procedures must be applied for this variance, the replication/deletion approach, batch means, the regenerative method, autoregressive processes, the spectral estimation method, or the standardized time series method, all of which are not free from obstacles, see Fishman (1978), Bratley, Fox, and Schrage (1987), Banks (1998), or Law and Kelton (2000). This difficulty is omitted for min-max confidence intervals.

Even the variance of an estimator may not exist, for example in the case of some heavy-tailed distributions (Sigman 1999). Nevertheless, min-max confidence intervals can be constructed, whereas classical confidence intervals cannot.

It is easy to obtain median confidence intervals for functions of two or more estimators whereas it is difficult to get confidence intervals with other known methods, in general, except for jackknife intervals (Miller 1974); an example is given in section 5.

In realistic models which involve dependent simulation output with unknown distribution, median confidence intervals seem to be more accurate, i.e. the coverages are closer to the predefined confidence level.

Some independent estimations of a measure of interest are used for a min-max or median confidence interval, say w, e.g. 5 or 6. The confidence level depends on w, only values like $1 - 0.5^w$, $w = 1, 2, \ldots$, or similar are possible – this becomes clear later in theorem 1. In terminating simulation, such an estimation can be obtained from a single terminating simulation run, e.g. a mean, or it can be a mean taken from some terminating simulation runs.

In steady state simulation, each independent estimation can be obtained from an independent simulation run, as for the replication/deletion approach. In each replication the statistical equilibrium must be reached before data can be collected. The new technique shares this drawback with the replication/deletion approach.

This can be omitted: A single simulation run is produced with only one transient phase and w consecutive batches which are considered to be nearly independent as with the batch means method. Each batch is taken as a substitute of a distinct replication. We call this *batch median confidence intervals*.

Now we explain the new technique in detail. The random variables of the sample $X_{1,1}, ..., X_{1,n}$ may have the distribution function $F_{X,\theta}(x)$ when they stem from a steady-state simulation run of length n, or from n independent terminating simulation runs. θ, $\theta \in \Theta$, is a parameter, for example the mean or a quantile, and Θ a set of possible parameters. Or the sample stems form a terminating simulation and has a common distribution with the parameter θ.

Let $T(X_{1,1}, ..., X_{1,n})$ denote an estimator for the parameter θ with the distribution function $F_\theta(x)$, $\theta \in \Theta$.

We consider a novel kind of confidence interval

$$[T^{\min},\ T^{\max}) \tag{1}$$

where

$$T^{\min} = \min_{1 \le i \le w} T_i \quad \text{and} \quad T^{\max} = \max_{1 \le i \le w} T_i.$$

Here, the $T_i = T(X_{i,1}, ..., X_{i,n})$, $i = 1, \ldots, w$, are estimators for w independent replications $X_{i,1}, ..., X_{i,n}$ of the sample $X_{1,1}, ..., X_{1,n}$. We call (1) a “min-max confidence interval”.

For $F = F_\theta(\theta)$, the value of the estimator distribution function at θ, the following theorem holds.

Theorem 1 *The interval (1) is a confidence interval for the parameter θ with the confidence level $1 - F^w - (1 - F)^w$, i.e.*

$$P\{T^{min} \le \theta < T^{max}\} = 1 - F^w - (1 - F)^w \tag{2}$$

holds.

The **proof** is very simple. The probability that T_i is less than or equal to θ is $P\{T_i \le \theta\} = F_\theta(\theta) = F$, the probability that $T^{\max}$ is less than or equal to θ is $P\{T^{\max} \le \theta\} = P\{\text{all } T_i \le \theta\} = F^w$ due to the independency. Similarly, $P\{T_i > \theta\} = 1 - F_\theta(\theta) = 1 - F$, $P\{T^{\min} > \theta\} = P\{\text{all } T_i > \theta\} = (1 - F)^w$. Hence $P\{T^{\min} > \theta \text{ or } T^{\max} \le \theta\} = F^w + (1 - F)^w$, and (2) follows. □

Remarks

1. The distribution function $F_\theta(x)$ of the estimator may not be known, only the value $F_\theta(\theta)$ is needed.

2. The variance of the estimator is not needed, the question whether the random variables $X_{i,1}, \ldots, X_{i,n}$ are independent does not arise.

3. The confidence level cannot be chosen arbitrarily, only the values $1 - F^w - (1-F)^w,\ w = 2, 3, \ldots$ are allowed.

Now we consider the most important special case where $F_\theta(\theta) = 1/2$, i.e. the unknown parameter is the median of the estimator. Therefore we use the term “median confidence intervals”. This is the case for all unbiased estimators with symmetric distributions, for example the normal distribution. Then,

$$P\{T^{\min} \le \theta < T^{\max}\} = 1 - 0.5^{w-1} \tag{3}$$

holds, the confidence level can be one of the values $1 - 0.5^{w-1},\ w = 2, 3, \ldots$.

Symmetry of the estimator distribution, the absence of skewness, is a sufficient condition for median conficence intervals to be exact. It is not necessary, there are unsymmetric distributions for which the mean and the median coincide, hence median conficence intervals would be exact for them.

If the median is merely close to the expectation of an unbiased estimator $F \neq 0.5$ but $F \approx 0.5$ holds, and the median confidence interval is only approximate. The error of the conficence level is the difference between (2) and (3). This happens quite often, due to the central limit theorem, when the summed random variables are not normally distributed, but n, the number of summands, is large. Then the distribution function of the estimator is approximately a normal distribution, hence approximately symmetric, and the median is near to the expectation.

A min-max confidence interval is exact if the w replications are independent, even the estimator may be biased. This sounds very interesting, but the serious problem is the value $F = F_\theta(\theta)$, the value of the estimator distribution function at θ, the unknown parameter which is to be estimated. We do not know how to calculate this F in general.

But there is an interesting application where F can be calculated: Order statistics as estimates for quantiles. Consider samples $X_1, \ldots, X_n$ and the according ordered sequence $X_{(1)}, \ldots, X_{(n)},\ X_{(i)} \le X_{(j)}$ if $i < j$, where the X_i are IID with the strictly increasing distribution function $F(x)$. The q-quantile $\theta = x_q,\ q \in (0,1),\ F(x_q) = q$, is estimated by $X_{(r)},\ r \in \{1, 2, \ldots, n\}$. Let $F_\theta(x)$ denote the distribution function of the estimator, namely $X_{(r)}$. Here, $F = F_\theta(x)$ is known:

Theorem 2 *If the q-quantile x_q is estimated by $X_{(r)}$, the min-max confidence interval (1) has precisely the confidence level (2) with*

$$F = \sum_{i=r}^{n} \binom{n}{i} q^i (1-q)^{n-i}. \tag{4}$$

Proof For any $k,\ 0 < k < n$, the probability $P\{X_{(k)} \le x < X_{(k+1)}\}$ equals the probability that k of the random variables X_i of the sample are less or equal x, hence $P\{X_{(k)} \le x < X_{(k+1)}\} = \binom{n}{k} F^k(x)[1 - F(x)]^{n-k},\ k = 1, \ldots, n-1$, and $P\{X_{(n)} \le x\} = F^n(x)$ hold. Using this we conclude $F_\theta(x) = P\{X_{(r)} \le x\} = P\{X_{(r)} \le x < X_{(r+1)}\} + P\{X_{(r+1)} \le x < X_{(r+2)}\} + \ldots + P\{X_{(n)} \le x\} = \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i}$. With $F(x_q) = q$, (4) follows. □

Remarks

1. Here the value $F = F_\theta(x_q)$ is independent of the actual distribution function of the sample elements X_i.

2. Theorem 2 is not useful for the simulation of the extremes, $q = 0$ or $q = 1$. Here we get the confidence level 0.

3. Usually, $r \approx qn$ is chosen.

Corollary 1 *If the sample size n is odd, $r = \lceil n/2 \rceil$ and $q = 0.5$, i.e. the median is estimated, $F = 0.5$ holds.*

Proof Here $F = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} = 0.5^n \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i}$. Due to $\binom{n}{i} = \binom{n}{n-i}$, we have also $F = 0.5^n \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{n-i} = 0.5^n \sum_{j=0}^{n-\lceil n/2 \rceil} \binom{n}{j} = 0.5^n \sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{j}$. We add these two equations: $2F = 0.5^n \sum_{j=0}^{n} \binom{n}{j} = 0.5^n 2^n = 1$. $F = 0.5$ follows. □

The critical value F is also known for some toy simulations. In (Strelen 2001) we consider min-max confidence intervals for the estimation of the variance of normally distributed random variables. Moreover, it can be estimated in a very long and expensive simulation; in the section 6 we present a brute force example.

3 ACCURACY OF CONFIDENCE INTERVALS

Consider a sample of realizations of random variables with a specific distribution. From this sample, a parameter of the distribution is estimated. In the statistical theory, confidence intervals for such an estimation are provided for any given confidence level CL. When some assumptions concerning the sample and the estimator are fulfilled, CL is the probability that the real parameter value is contained in the confidence interval.

In simulation, it is very common to estimate means and to assume normality and independence of the random variables. Usually both is not true literally, only more or less, depending on the length of the simulation runs. Due to this, the confidence intervals are inaccurate, more precisely, the assumed confidence level is not the probability that the real parameter value lies within the calculated confidence interval. Quite often, the actual probability C is smaller than the confidence level CL, hence the confidence level is too optimistic. Sometimes the actual probability C that the real parameter value is covered by a confidence interval is called *coverage*.

The empirical comparative study of Law and Kelton (2000), p. 535, exhibits how (in)accurate confidence intervals may be. They found remarkable errors, e.g. $C = 69\%$ instead of $CL = 90\%$.

For such an empirical study, a model with a known parameter is considered, e.g. the expectation of the customers' delays in the queue of an M/M/1 queueing system. This parameter is estimated by simulation. Many simulations are performed, sometimes the known value is within the calculated confidence interval, sometimes it is not. The frequency of the confidence interval covering the real value estimates the coverage.

In many empirical studies we compared the accuracy of median confidence intervals with classical confidence intervals, see (Strelen 2002); in the section after the following one we will present some figures. But first, in the next section, we present analytical models for the coverage of classical confidence intervals (with Student's distribution) and for the coverage of median confidence intervals.

4 ANALYTICAL MODELS FOR THE COVERAGE

In an ideal world, we would give theorems which state the accuracy of the new confidence intervals, but at present this seems impossible. Instead, we evaluated the accuracy with examples and simulation; in section 5 we present results. Some scientists prefer analytical models, and this is why we built models which allow to calculate coverages.

Each model consists of two parts, one part defines the distribution of an estimator in a specific simulation model. With this distribution, the other part defines the distribution of a confidence interval for the estimated parameter, namely two random variables for the boundary points of the confidence interval. We consider classical confidence intervals with the Student distribution, and median confidence intervals.

With the distributions of the boundary points, the probability that an estimation lies within the confidence interval is calculated – that is the expectation of the coverage.

We consider simulations with the replication/deletion (R/D) approach: In w independent simulation runs, the replications, a parameter θ is estimated. To this end, in the replications, means $\bar{X}_i,\ i = 1, \dots, w$, are sampled.

The distribution of these means depends on the specific model, the parameter under consideration, and on the mode of simulation, terminating or steady state. We consider the delays in an M/M/1/N queue with N buffer places, in steady state. For the moment, we assume that we know the distribution of the means. Later we point out how to obtain it.

A Student confidence interval for the parameter θ can be obtained as follows: The great mean $\bar{\bar{X}} = (\bar{X}_1 + ... + \bar{X}_w)/w$ estimates θ, and with the empirical variance $S^2 = \sum_{i=1}^{w}(\bar{X}_i - \bar{\bar{X}})^2/(w-1)$ one obtains the confidence interval $\bar{\bar{X}} \pm B^{\text{half}}$ to the confidence level $1-\alpha$ where the half width is $B^{\text{half}} = t_{w-1,1-\alpha/2}\sqrt{S^2/w}$; $t_{w-1,1-\alpha/2}$ is the upper critical value of Student's t-distribution with $w-1$ degrees of freedom.

If this distribution of the means $\bar{X}_i$ is given as density $f(x)$, the coverage is

$$C = \int_{\mathbb{R}^w} \mathcal{I}\{|\bar{\bar{X}} - \theta| \le B^{\text{half}}\} f(x_1)...f(x_w)dx_1...dx_w$$

where $\bar{\bar{X}}$ and B^{half} depend on $x_1, ..., x_w$ instead of $\bar{X}_1, ..., \bar{X}_w$, and $\mathcal{I}$ denotes the indicator function. This formula looks simple but the w-fold integral can not be evaluated neither in closed form nor numerically.

Therefore we use a common distribution for the random variables $S = \bar{X}_1 + ... + \bar{X}_w$ and $Q = \bar{X}_1^2 + ... + \bar{X}_w^2$ instead. Using this common distribution one gets "random confidence intervals", i.e. two random variables which are the lower und the upper boundary point as follows: $\bar{\bar{X}} = S/w$, $S^2 = (Q - w\bar{\bar{X}}^2)/(w-1)$, B^{half} as above, and the boundary points are $\bar{\bar{X}} \pm B^{\text{half}}$. The coverage C is the probability that the parameter θ is in this random confidence interval.

For a discrete distribution of the means $\bar{X}_i,\ i = 1, \dots, w$, $P\{X_i = x\} = f_x,\ x \in \mathbb{Z}$, the common distribution $P\{S = s, Q = q\} = f_{s,q}^{(w)},\ s, q \in \mathbb{Z}$, can be efficiently calculated, as will be seen soon. With this distribution, the coverage is

$$C = \sum_{(s,q)\in\mathbb{Z}^2} \mathcal{I}\{\bar{\bar{X}} - B^{\text{half}} \le \theta \le \bar{\bar{X}} + B^{\text{half}}\} f_{s,q}^{(w)}. \qquad (5)$$

The common distributions $f_{s,q}^{(w)}$ are calculated recursively for $w = 2, 3, ...$ according to

$$f_{s,q}^{(2)} = \sum_{\substack{(i,j)\in\mathbb{Z}^2 \\ i+j=s, i^2+j^2=q}} f_i f_j, \quad s, q \in \mathbb{Z}, \qquad (6)$$

and for $w > 2$

$$f_{s,q}^{(w)} = \sum_{\substack{(i,t,r)\in\mathbb{Z}^3 \\ i+t=s, i^2+r=q}} f_i f_{t,r}^{(w-1)}, \quad s, q \in \mathbb{Z}. \qquad (7)$$

The reader may note the similarity of these formulas with convolution. Many $f_{s,q}^{(w)}$ are zero, therefore we store these probabilities in a hash table.

Now we determine the distribution of the estimator, namely the means $\bar{X}_i$ of the delays in an M/M/1/N queue with N buffer places, in steady state. We give a Discrete Time Markov Chain (DTMC) for the desired probabilities which is finite state, inhomogeneous, and has absorbing states. The probabilities in the absorbing states provide the probabilities of the estimator. The DTMC is embedded, the observation times are the beginnings of services. As in steady state simulation, the delays of the customers in the transient phase are not taken into consideration; we call them *ignored customers*.

After a while when the steady state is (nearly) reached, the first of *n observed* customers whose delays are sampled arrives. He sees I^{init} ignored customers in front of himself which are served before him, $I^{\text{init}} \in \{0, \ldots, N-1\}$, with probability $P\{I^{\text{init}} = i\} = p_i/(1-p_N)$. Here p_i is the probability that i customers are in an M/M/1/N queue as given in any book on queueing theory, e.g. (Allen 1990).

The states of the DTMC are given by some random variables: I, the number of ignored customers in the system, Z observed customers are present whose delays are sampled, $0 \leq Z+I \leq N$, K observed customers arrived already which were delayed W service times up to now, $0 \leq K \leq n$.

We consider two phases of the DTMC. In the first phase, ignored customers are present and are served, and observed customers arrive and wait. When the last ignored customer leaves the system, the second phase begins where observed customers are served, more customers arrive, wait, and are served. After n services, they all are served, the DTMC is in an absorbing state.

In an M/M/1/N queue, while a customer is served, $A^{(S)}$ new customers arrive with probability

$$\begin{aligned} P\{A^{(S)} = a\} &= \int_0^\infty \exp(-\lambda t)\frac{(\lambda t)^a}{a!}\mu \exp(-\mu t)dt \\ &= \left(\frac{\alpha}{1+\alpha}\right)^a \frac{1}{1+\alpha},\ a = 0, 1, \ldots \end{aligned}$$

($\alpha = \lambda/\mu$ where λ is the arrival rate and μ the service rate), but some of them are lost when the buffer is full.

In our model, we restrict the number of arivals, within a service only $A \leq A^{\max}$ customers arrive instead of $A^{(S)}$:

1. $A \leq a^{\max}$ where $P\{A^{(S)} > a^{\max}\}$ is very small,
2. $A \leq n-K$, we are interested only in n delays,
3. $A \leq N-Z-I$ restricts the number of customers in the buffer.

If A is smaller than this maximum we take the probabilities $P\{A=a\} = P\{A^{(S)} = a\}$, otherwise the arrival probability is $1 - P\{A = A^{\max} - 1\} - \ldots - P\{A = 0\}$.

In the first (transient) phase ignored customers are served. Within the first service of this phase, at least one customer arrives, hence we take the arrival probabilities conditioned on this fact, $P\{A = a|A > 0\}$ instead of $P\{A = a\}$ which is valid afterwards.

The state transitions in this first phase are as follows.

- $I \longrightarrow I-1$, one ignored customer less,
- $Z \longrightarrow Z+a$, a observed customers more in system,
- $K \longrightarrow K+a$, a observed customers more arrived,
- $W \longrightarrow W+Z+a$, all present and all new observed customers wait one service time,

the transition probability is $P\{A = a\},\ a = 0, \ldots, A^{\max}$.

Initially, $I = I^{\text{init}},\ Z = 0,\ K = 0,\ W = 0$ holds with probability $P\{I^{\text{init}} = i\},\ i = 0, \ldots, N-1$.

The transitions in the second phase where the n observed customers are served occur with the transition probability $P\{A = a\},\ a = 0, \ldots, A^{\max}$. They are as follows ($I$ is always 0). If $Z > 1$ or $A > 0$:

- $Z \longrightarrow Z-1+a$, one customer finished, a new customers,
- $K \longrightarrow K+a$, a customers more arrived,
- $W \longrightarrow W+Z+a$, all present and all new customers wait one service time.

If $Z = 1$ and $A = 0$ and $K < n$ (after the service, the system is empty; later a customer arrives):

- $Z \longrightarrow 1$, the new customer,
- $K \longrightarrow K+1$, one customer more,
- W remains unchanged.

If $Z = 1$ and $K = n$, no arrivals occur and the last service follows. The absorbing states are reached with transition probability 1:

- $Z \longrightarrow 0$, the last customer leaves the system,
- $K = n$ remains unchanged, no more arrivals,
- W remains unchanged.

The absorbing probabilities $P\{W = j,\ Z = 0,\ K = n\}$ equal the probabilities $P\{W = j\},\ j = 0, 1, \ldots$. $W = j$ means: The sum of all delays consists in j service times. With these probabilities, the distribution function of $n\bar{X}_i$ which is in principle the estimator, is

$$\sum_{j=0}^{\infty} P\{W = j\} F_{E(j,j/\mu)}(x)$$

where $F_{E(j,j/\mu)}$ denotes the Erlang-j distribution function with expectation j/μ.

But we need a discrete distribution for the means $\bar{X}_i$. Therefore we approximate the exponential distribution with a geometric distribution. Let X have an exponential distribution with distribution function $F_E(x) = 1 - \exp(-\lambda t)$

and expectation $E[X] = 1/\lambda$, and let G have a geometric distribution with the probabilities $P\{G = x\} = p(1-p)^x$, $x = 0, 1, \ldots$, and expectation $E[G] = (1-p)/p$. For $x = 0, 1, \ldots$, $F_E(x+1) - F_E(x) = (1 - e^{-\lambda})(e^{-\lambda})^x$. This motivates the approximation $X \approx G$ where G is geometrically distributed with parameter $p = 1 - e^{-\lambda}$. For the expectations $0 < E[X] - E[G] < 0.5$ holds, hence the relative error $|(E[X] - E[G])/E[X]|$ is smaller than 0.5λ, the approximation is better for a small λ.

Sums $X^{(j)}$ of j iid. exponentially with parameter λ distributed random variables have an Erlang-j distribution with expectation j/λ, sums $G^{(j)}$ of j iid. geometrically with parameter p distributed random variables have a negative binomial distribution with the probabilities

$$P\{G^{(j)} = x\} = \binom{j+x-1}{x} p^j (1-p)^x,\ x = 0, 1, \ldots,$$

and expectation $j(1-p)/p$. This motivates the approximation

$$X^{(j)} \approx G^{(j)} + \delta_j, \quad \delta_j = \left\lfloor j\left(\frac{1}{\lambda} - \frac{1-p}{p}\right) + 0.5 \right\rfloor$$

which accounts for the error $E[X] - E[G]$.

Together with the probabilities for W we obtain the approximate discrete probabilities for $n\bar{X}_i$:

$$P\{n\bar{X}_i = x\} = \sum_{j=0, x \geq \delta_j}^{\infty} P\{W = j\} P\{G^{(j)} = x - \delta_j\}, \quad (8)$$

$x = 0, 1, \ldots$.

Due to the finite DTMC for the probabilities $P\{W = j\}$, these are zero for large j. For large $x > x^{\max}$ when $P\{n\bar{X}_i = x\}$ becomes very small, we set these probabilities to zero and normalize as follows: $P\{n\bar{X}_i = x^{\max}\} = 1 - P\{n\bar{X}_i = x^{\max} - 1\} - \ldots - P\{n\bar{X}_i = 0\}$.

Now we change the time skale, one old unit equals n new units. Than the probabilities in (8) are the desired distribution for the means $\bar{X}_i$, $P\{\bar{X}_i = x\} = f_x$.

This distribution is used for the calculation of coverages for Student confidence intevals and median confidence intevals: We apply them in (6) and in (7), get the common distribution for S and Q, and calculate the coverage C^{Student} with (5). This is the desired figure for confidence intervals with the Student distribution.

For median confidence intervals, the coverage C^{MCI} equals the confidence level (2) of a min-max confidence interval according to theorem 1, $C^{\text{MCI}} = 1 - [F(\theta)]^w - [1 - F(\theta)]^w$ where $F(y)$ denotes the distribution function of the estimators $\bar{X}_i$, $F(y) = \sum_{x \leq y} f_x$.

Some figures for the errors $CL - C^{\text{Student}}$ and $CL - C^{\text{MCI}}$ are given in Table 1, the theoretical confidence level CL is 93.75% (w=5):

Table 1: Errors $CL - C$

ρ	N	n	$CL - C^{\text{Student}}$	$CL - C^{\text{MCI}}$
0.1	10	20	0.056	0.043
0.3	20	5	0.103	0.076
0.3	10	10	0.065	0.047
0.3	10	20	0.040	0.027
0.5	10	10	0.080	0.051
0.5	5	40	0.016	0.014
0.8	10	10	0.040	0.020
0.9	20	20	0.035	0.017

Obviously, in all examples the median confidence interval is more accurate - we did not find counter-examples.

5 NUMERICAL EXPERIENCE

Many simulation studies were accomplished which compared median confidence intervals and classical confidence intervals. We summarize some results.

In these studies, simulation experiments were done with different models. In each experiment confidence intervals were calculated with well-known methods and with the new technique. Especially the replication/deletion method is compared with median confidence intervals. For both techniques some independent replications of the simulation must be done. In steady-state simulations each replication begins with a transient phase. Batch median confidence intervals are compared with the batch means method. Both techniques need only one transient phase.

We chose the confidence level 93.75% for all confidence intervals, hence $w = 5$ replications or batches.

In each study, many independent experiments were performed. In each experiment we noticed if the true value of an estimated parameter (which is known here) was contained in the (median) confidence interval or not. So we estimated the coverage C. This coverage should be near to the theoretical confidence level CL if the confidence intervals are accurate. The error $CL - C$ measures the accuracy of the confidence interval technique, the smaller the better. These errors serve the purpose to compare the accuracy of different techniques. Compared confidence intervals are calculated with equal total sample sizes but for the regeneration method this is possible only approximately.

The overall result of these studies is as follows: All confidence intervals are accurate or too small. When they are too small, the new technique is more accurate with slightly wider confidence intervals.

1. An M/M/1 queueing system is considered. The arrival rate is 1.0 and the service rate 1.25, hence the system is heavily loaded with utilization 0.8. Law and Kelton (2000),

p. 535, performed a comparative study in order to see how accurate the confidence intervals are. They applied different well-known methods for confidence intervals: Batch means (B), autoregressive method (A), spectrum analysis(SA), regenerative method (R) (classical (C) and jackknife (J)), and standardized time series (STS). 90% confidence intervals were constructed for the steady-state mean delay which is known to be 3.2.

For each of the methods and for different simulation run lengths, confidence intervals are considered. The total sample sizes are $n = 320$, 640, 1280, 2560. For batch means and standardized time series, the number of batches is 5, hence the batch sizes are $m = 64$, 128, 256, 512 (10 and 20 batches were also tried, but with worse results). These batch sizes m are also the numbers of regeneration cycles because the mean length of these cycles here is 5.

For each sample size and each method, Law and Kelton estimated the coverages C and the errors $CL - C$ for the confidence level $CL = 0.9$ with 400 independent simulation experiments (they took the results for the standardized time series method from another source). They counted how often the known value of the mean delay was inside the confidence interval and thus got the coverages C and the errors $CL - C$.

In this statistically difficult model, the coverage differs a good deal from the nominal confidence level, at least for small sample sizes. The longer the run was, the more accurate were the confidence intervals, as one would expect.

We conducted an according simulation study with the same model and the same run lengths. Batch median confidence intervals were constructed with $w = 5$ batches in each simulation. This implies a 93.75% confidence level CL. The transient phase is eliminated in our M/M/1 examples as follows: Initially, there are q customers in the system with probability $(1-\rho)\rho^q$ where ρ is the utilization. Since these are the steady-state probabilities, there is no transient phase – this is a result of queueing theory, see e.g. Allen (1990).

Unfortunately, we did not repeat the experiments of Law and Kelton. They used the confidence level of 90% which is impossible for the new technique. Hence, 90% confidence intervals and 93.75% median confidence intervals are compared; this is not entirely correct. But the reader may note that we compared the difference between the assumed confidence level CL and the coverage C which is the error, not the coverage itself. This remark only applies to the figures in Table 2; all other confidence intervals are calculated for the level 93.75%.

In all cases, the coverages of the batch median confidence intervals were nearer to the theoretical value of 93.75% than all coverages of the Law and Kelton study to 90%, the BMCI errors were smaller. That means, in the considered examples, the new technique is more accurate than all the other methods.

An overview of the errors of the Law-and-Kelton study and our study is given in Table 2. Here the entries are the errors $CL - C$. For example, for the batch means method (B) and sample size $n = 320$ the error is 0.210. This means that the observed coverage is 69% = 90% (nominal confidence level) - 21%. Or for batch median confidence intervals (BMCI) and sample size $n = 1280$ the error is 0.060, hence the observed coverage is 87.75% = 93.75% - 6%.

Table 2: Errors $CL - C$

$n(m)$	320(64)	640 (128)	1280 (256)	2560 (512)
B	0.210	0.177	0.120	0.102
STS	0.380	0.272	0.170	0.102
SA	0.187	0.140	0.117	0.067
A	0.212	0.177	0.147	0.145
RC	0.340	0.217	0.195	0.155
RJ	0.230	0.172	0.152	0.137
BMCI	0.127	0.102	0.060	0.045

We had the impression that the statistical relevance from 400 independent experiments is insufficient. Therefore we made more experiments in all our studies and calculated 90% confidence intervals for the errors.

In Table 3 we compare the batch means method and batch median confidence intervals with 25600 independent experiments. In this table and others in section 5, the entries are the errors $CL - C$, together with the average of the half lengths of the (median) confidence intervals, and the empirical variance of the half lengths. The confidence intervals of the errors are $CL - C \pm \epsilon$, $\epsilon < 0.005$.

Table 3: Errors $CL - C$, avg. Half Length, var. Half Length

$n(m)$	B	BMCI
320(64)	0.189 2.4 3.3	0.153 2.5 3.4
640 (128)	0.135 2.2 2.7	0.107 2.3 2.9
1280 (256)	0.094 1.8 1.7	0.072 1.9 1.8
2560 (512)	0.059 1.4 0.8	0.046 1.5 0.9

The bach median confidence intervals are more accurate and only slightly wider than classical confidence intervals.

2. In another study, median confidence intervals and classical confidence intervals which were achieved with the replication/deletion approach were compared. The M/M/1 queue was simulated with light, medium, and heavy load ρ (arrival rate 1.0, service rate 4.0, 2.0, 1.25). For each load, short simulations with run lengths 150, 200, and 500 delays, respectively, and long simulations with 2,400, 3,200, and 8,000 delays, respectively, were performed. We did 25,600 independent simulation experiments for each case.

In the short simulations, the obtained median confidence intervals are more accurate than the classical confidence intervals from the replication/deletion approach. In the long simulations for light and medium load, no statistically significant differences were observed: Both techniques yielded accurate confidence intervals. We conjecture that here the estimator is nearly normally distributed, and for normally distributed estimators, both techniques provide exact confidence intervals. In Table 4, the confidence intervals of the errors are $CL - C \pm \epsilon$, $\epsilon < 0.004$.

Table 4: Errors $CL-C$, avg. Half Length, var. Half Length

ρ	Run	RD			MCI		
0.25	Short	0.022	0.03	≈ 0	0.017	0.03	≈ 0
	Long	0.001	0.01	≈ 0	0.000	0.01	≈ 0
0.5	Short	0.032	0.19	0.01	0.024	0.20	0.01
	Long	0.003	0.05	≈ 0	0.004	0.05	≈ 0
0.8	Short	0.058	1.44	0.85	0.043	1.53	0.94
	Long	0.005	0.42	0.03	0.005	0.45	0.04

3. In study 2. we compared also median confidence intervals and jackknife intervals for ratios of estimators. In particular, we estimated the expected throughput, $\hat{\lambda}^{(r)}$, as the ratio of the mean number of jobs in the waiting room, $\hat{Q}$, and the mean delay, $\hat{W}$, $\hat{\lambda}^{(r)} = \hat{Q}/\hat{W}$ (Little's law), and we estimated the mean delay $\hat{W}^{(r)}$ by $\hat{W}^{(r)} = \hat{Q}/\hat{\lambda}$. For these ratios, $\hat{Q}$, $\hat{W}$, and the throughput $\hat{\lambda}$ were estimated directly.

For the ratios, we calculated median confidence intervals and jackknife intervals. In all examples, the median confidence intervals are much more accurate than the jackknife intervals. In Table 5, the confidence intervals of the errors are $CL - C \pm \epsilon$, $\epsilon < 0.005$.

Table 5: Errors $CL-C$, avg. Half Length, var. Half Length

ρ	What	Run	RD, Jackknife			MCI		
0.25	$\hat{\lambda}^{(r)}$	Short	0.105	0.07	≈ 0	0.002	0.10	≈ 0
		Long	0.074	0.02	≈ 0	-0.002	0.02	≈ 0
	$\hat{W}^{(r)}$	Short	0.093	0.02	≈ 0	0.015	0.04	≈ 0
		Long	0.075	0.01	≈ 0	0.000	0.01	≈ 0
0.5	$\hat{\lambda}^{(r)}$	Short	0.126	0.06	≈ 0	0.005	0.09	≈ 0
		Long	0.078	0.01	≈ 0	0.000	0.02	≈ 0
	$\hat{W}^{(r)}$	Short	0.098	0.14	0.01	0.019	0.20	0.01
		Long	0.080	0.04	≈ 0	0.003	0.05	≈ 0
0.8	$\hat{\lambda}^{(r)}$	Short	0.178	0.04	≈ 0	0.017	0.05	≈ 0
		Long	0.079	0.01	≈ 0	-0.001	0.01	≈ 0
	$\hat{W}^{(r)}$	Short	0.123	1.07	0.50	0.036	1.55	0.96
		Long	0.079	0.31	0.02	0.005	0.45	0.04

4. In the next example we consider the heavy-tailed Pareto distribution with the distribution function $F(x) = 1 - (b/x)^a$, $0 < a <= 2$, $0 < b <= x$ where a is a shape parameter and b a skale parameter. The expectation is $ab/(a-1)$ if $a > 1$, the median $2^{1/a}b$, and the variance does not exist. This distribution is very skewed.

In simulations we constructed median confidence intervals with w =5 replications for the expectation and the median and classical confidence intervals only for the median; for the expectation they do not exist due to the non-existing variance.

Each replication consisted in $m = 999$ independent observations, hence the total sample size was $n = wm = 4995$. The coverages C and the errors $CL-C$ were estimated with 1000 independent simulations.

For shape parameter $a = 2$ and scale parameter $b = 1$, the accuracy of the median confidence intervals for the expectation of the Pareto distribution was quite good, $CL - C = 0.003 \pm 0.013$, for $a = 1.5$ quite bad, $CL - C = 0.086 \pm 0.019$, for $a = 1.1$ inacceptably bad, even for much bigger sample sizes. These results confirm the general observation that the mean may converge poorly towards the expectation for heavy-tailed distributions.

Here one should resort to alternative estimators; we chose the suitable order statistic for the median and found very accurate estimates and very accurate median confidence intervals.

In Table 6 the errors are given for classical confidence intervals (CI) and median confidence intervals (MCI) for this order statistic. Their confidence intervals are $\pm\epsilon$ with $\epsilon < 0.019$.

Table 6: Errors $CL-C$, avg. Half Length, var. Half Length

a	CI			MCI		
2	0.007	0.02	≈ 0	0.007	0.03	≈ 0
1.5	0.011	0.04	≈ 0	0.002	0.04	≈ 0
1.1	-0.005	0.06	≈ 0	0.001	0.06	≈ 0

Clearly both kinds of confidence intervals are accurate. Due to theorem 2, one would expect the median confidence intervals to be exact. The observed figures seem to confirm this.

5. Now we present an example where the estimator has a very skewed and non-normal distribution. Hence, confidence intervals and median confidence intervals are quite inaccurate, but again the latter are better.

The considered reliability model from Law and Kelton (2000), p. 508, consists of three components and will function as long as component 1 works and either component 2 or 3 works. If G is the time to failure of the whole system and G_i is the time to failure of component i, $i = 1, 2, 3$, then $G = \min\{G_1, \max\{G_2, G_3\}\}$. It is further assumed that the random variables G_i are independent and that each G_i has a Weibull distribution $F(x) = 1 - \exp(-x/b)^a$, $x > 0$, with shape parameter $a = 0.5$ and scale parameter $b = 1$. This particular Weibull distribution is extremely skewed and nonnormal.

We constructed median confidence intervals with $w = 5$ replications each of which consisted of $m = 1$, 2, 4, or 8 outcomes, and the mean of them is the estimator, hence the total sample sizes were $n = wm = 5$, 10, 20, 40.

The sample of n realisations of the time to failure G is independent. Hence we calculated a classical confidence interval with the Student quantile $t_{n-1,1-\alpha/2}$, as indicated in the beginning of section 4.

The coverages C and the errors $CL-C$ were estimated with 8000 independent simulations. In Table 7, the errors are given. Their 90% confidence intervals are $\pm\epsilon$ with $\epsilon < 0.008$. Again the MCIs are clearly more accurate than the CIs.

Table 7: Errors $CL-C$, avg. Half Length, var. Half Length

$n(m)$	CI			MCI		
5 (1)	0.191	1.15	1.31	0.147	1.17	1.33
10 (2)	0.143	0.77	0.37	0.079	0.95	0.59
20 (4)	0.105	0.55	0.11	0.049	0.73	0.23
40 (8)	0.069	0.40	0.03	0.032	0.55	0.09

6 FUTURE RESEARCH

In section 2 we mentioned that min-max confidence intervals would be exact, but the modeller would have to be able to obtain $F_\theta(\theta)$, the value of the estimator distribution function at θ which is the value of the unknown parameter to be estimated, in an efficient way. With this value, the real confidence level according to (2) could be determined very accurately.

We illustrate this with example 5: In very long and expensive simulations we estimated first the empirical distribution of G and then the distribution of the estimator $\hat{F}_\theta(x)$ which is essentially the m-fold convolution of this empirical distribution. Using $\hat{\theta}$, the estimation of the unknown parameter θ, we obtained $\hat{F} = \hat{F}_\theta(\hat{\theta})$. With this estimated $\hat{F}$ we calculated the confidence level $\hat{C}L$ according to (2). Table 8 shows that these estimated confidence levels are very close to the observed coverages C, even in this pathological example.

Table 8: Coverages and Estimated Confidence Levels

$n(m)$	C	$\hat{C}L$
5 (1)	0.791±0.002	0.791
10 (2)	0.852±0.002	0.848
20 (4)	0.886±0.002	0.884
40 (8)	0.909±0.002	0.907

This is a brute force appoach which we cannot recommend due to the high effort - min-max confidence intervals would require efficient estimation of $F_\theta(\theta)$.

REFERENCES

Allen, A. O. 1990. *Probability, statistics and queueing theory*. Academic Press.

Banks, J. 1998. *Handbook of simulation*. New York: Wiley and Sons.

Bratley, P., B. Fox, and L. Schrage. 1987. *A guide to simulation*. second ed. New York: Springer.

Fishman, G. S. 1978. *Principles of discrete event simulation*. New York: John Wiley and Sons.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*. third ed. New York: McGraw-Hill.

Miller, R. 1974. The jackknife - a review. *Biometrika* 61:1–15.

Sigman, K. 1999. A primer on heavy-tailed distributions. *Queueing Systems* 33:261–275.

Strelen, J. C. 2001. Median confidence intervals. In *Modelling and Simulation 2001 - Proceedings of the ESM 2001*, ed. E. Kerkhoffs and M. Snorek, 771–775. Society for Computer Simulation.

Strelen, J. C. 2002. Median confidence intervals - grouping data into batches and comparison with other techniques. In *Proceedings of the Business and Industry Symposium - ASTC 2002*, ed. M. Ades and L. Deschaine, 169–175. San Diego: Society for Computer Simulation.

AUTHOR BIOGRAPHY

JOHANN CHRISTOPH STRELEN received the Dipl.-Math. and Dr. rer. nat. degrees in mathematics and the Habilitation degree in Computer Science from the Technische Hochschule Darmstadt, Germany. He is Professor of Computer Science at the university of Bonn, Germany. His research interests include performance evaluation, distributed systems, and simulation. His e-mail address is <strelen@bonn.edu> and his web address is <www.informatik.uni-bonn.de/∼strelen>.