MEDIAN CONFIDENCE INTERVALS - GROUPING DATA INTO BATCHES AND COMPARISON WITH OTHER TECHNIQUES

Johann Christoph Strelen
Rheinische Friedrich–Wilhelms–Universität Bonn
Römerstr. 164, 53117 Bonn, Germany
E-mail: strelen@cs.uni-bonn.de

KEYWORDS

Confidence Intervals, Discrete Simulation, Median

ABSTRACT

Confidence intervals around the median of estimators are proposed as a substitute for confidence intervals around the expectation. This is adequate since for many estimators the median and the expectation are close together, or even coincide, particularly if the sample size is large. Median 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. Grouping data into batches which is known from the batch means method is also useful for median confidence intervals.

INTRODUCTION

Stochastic simulation profits from fast computers, many more and much longer simulation runs than a decade ago can be carried out in reasonable time, estimation can rely on many data. This increases the trustworthiness in the statistical simulation results, and in many situations, the distributions of estimators can be expected to be close to a normal distribution, in which the expectation and the median coincide. This observation motivates us ([8]) to consider confidence intervals around the median as a substitute for confidence intervals (CI) around the expectation.

These median confidence intervals (MCI) are easier to obtain than usual confidence intervals, and furthermore they even apply in situations where usual confidence intervals cannot be used, in general, namely for functions of two or more estimators, or when the variance of the estimator does not exist.

Median confidence intervals are obtained by means of a small number w of replications, typically w =4, 5, or 6. They have attractive features and some minor disadvantages compared to classical confidence intervals:

1. The variance of estimators is not needed. Here is a main difficulty when confidence intervals are constructed because “simulation output data are always correlated” (Law and Kelton [5]). 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 [4], Bratley, Fox, and Schrage [3], Banks [2], or Law and Kelton [5]. This is omitted for median confidence intervals.

2. Even the variance of an estimator may not exist, for example in the case of some heavy-tailed distributions ([7]). Nevertheless the median confidence intervals can be constructed, whereas classical confidence intervals cannot.

3. 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 ([6]). For example, in a queueing system, the mean waiting time in the queue, $E[W]$, and the arrival rate λ are estimated. The mean number of jobs in the queue is $\lambda E[W]$ (Little's

law). Hence this number can be estimated indirectly with a product of estimates for which a median confidence interval can be given.

4. If in a simulation a median confidence interval is too wide it can be narrowed. A smaller median confidence interval can be obtained when the w replications are augmented, the runs are continued from the last state. Sometimes this is called a “sequential procedure”.

5. For some simple examples, e.g. the sum of independent normally distributed random numbers, we found median confidence intervals which are slightly wider than classical confidence intervals, but not too much ([8]).

6. In more realistic examples which involve dependent simulation output with unknown distribution, the width of median confidence intervals and classical confidence intervals did not differ significantly. More important, the median confidence intervals seem to be more accurate, i.e. the coverages are closer to the predefined confidence level. That means, in repeated simulations of the same model, the proportion of median confidence intervals which contain the real value of the estimated parameter is nearer to the theoretically expected value, the confidence level.

7. It must be remarked that not each level of confidence is possible with the new technique, only values like $1-0.5^j,\ j=1,2,\ldots$, or similar – this becomes clear in the theorem in the next section.

Some independent replications (simulations) are performed for median confidence intervals, say w. If steady state measures are to be estimated, in each replication the statistical equilibrium must be reached before data can be collected. The median confidence interval technique shares this drawback with the replication/deletion approach.

But this can be omitted: A single simulation run is produced instead with only one transient phase and w consecutive batches each of which is taken as a substitute of a distinct replication. We call this approach “batch median confidence intervals (BMCI)”. The idea to consider consecutive batches of output data as independent is similarly applied in the well-known batch means method.

MEDIAN CONFIDENCE INTERVALS

Let $X_{1,1},...,X_{1,n}$ be random variables with the distribution function $F_{X,\theta}(x)$ where $\theta,\ \theta\in\Theta$, is a parameter, for example the mean or the variance, and Θ a set of possible parameters. Let $T(X_{1,1},...,X_{1,n})$ denote an estimator for this parameter 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

$$T^{\min}=\min_{1\le i\le w} T_i$$

and

$$T^{\max}=\max_{1\le i\le w} T_i.$$

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 $X_{1,1},...,X_{1,n}$.

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

Theorem 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.

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 speak of “median confidence intervals”. This is the case for all estimators with symmetric distributions, for example if the estimator is unbiased and normally distributed. 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 $1-0.5^{w-1},\ w=2,3,\ldots$.

If the median is merely close to the expectation, the median confidence interval is only approximate. The skewness of the distribution is the reason for the error of the MCI, more precisely for the error of the conficence level, namely 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.

Confidence intervals for estimates in simulation are usually approximate since some assumptions for the applied method are not satisfied, for example

- the estimator has not the assumed distribution
- variables which are assumed to be independent are dependent.

For the constructionn of the confidence interval it is often assumed that the estimator has a normal or Student distribution and that it is based on an independent sample or some independent replications of the simulation.

For median confidence intervals the assumptions are weaker, only independency of the w replications and symmetry of the estimators distribution are required.

This last assumption needs not to be satisfied if the confidence level is determined with (2). Under these circumstances we speak of "min-max confidence intervals" (MMCI). An MMCI 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's distribution function at θ, the unknown parameter which is to be estimated. We do not know how to calculate this F in general. It is known for some toy simulations, in [8] we consider min-max confidence intervals for the estimation of the variance of normally distributed random variables. Or it can be estimated in a very long and expensive simulation, in the section "numerical experience" we present an example in 5.

BATCH MEDIAN CONFIDENCE INTERVALS

Batch median confidence intervals serve the purpose to estimate steady state measures. To this end, a single simulation run is performed. It begins with a transient phase which continues until the steady state is nearly reached, and w phases follow, the batches. They are considered to be independent as with the batch means method. With each batch, the desired parameter is estimated, and the minimum and the maximum of these estimates are considered for the median confidence interval (1).

NUMERICAL EXPERIENCE

Many simulation studies were accomplished in which median confidence intervals and classical confidence intervals are compared. 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 (batch) median confidence intervals.

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 classical methods, particularly with the batch means method. For both of these techniques only one transient phase at the beginning of the single run must be simulated before data are considered for steady-state simulations.

In each study, many independent experiments were performed. In each experiment we noticed if the true value of an estimated parameter (which was known here) was contained in the (median) confidence interval or not. So we estimated the "coverage" C, i.e. the fraction of (median) confidence intervals which contained the true value. 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.

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 [5, 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 percent 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.

Law and Kelton performed 400 independent simulation experiments with all indicated sample sizes n and each method (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 $CL - C$ where $CL = 0.9$.

As the authors remark, this model is known to be statistically difficult. This means that the assumptions for the methods are not strictly satisfied, hence the confidence intervals are not very accurate, the coverage differs a good deal from the theoretical 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 percent confidence level CL. We counted the proportion of 400 similar simulations for each run length which contained the true value and thus obtained the coverages C and the errors $CL - C$.

In all cases, the coverages of the batch median confidence intervals were nearer to the theoretical value of 93.75 percent than all coverages of the Law and Kelton study to 90 percent, 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 the following table. 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, the observed coverage is 69% since the theoretical confidence level was chosen to be 90%. Or for batch median confidence intervals (BMCI) and sample size $n = 1280$ the error is 0.060. This means, the observed coverage is 87.75% since the theoretical confidence level was chosen to be 93.75%.

$n(m)$	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

Errors $CL - C$

We had the impression that the statistical relevance from 400 independent experiments is not sure. Therefore we compared the batch means method and batch median confidence intervals in a study with 25600 independent experiments and 90 percent confidence intervals for the errors.

Here again the new technique is more accurate, especially for short sample sizes, as can be seen in the following table.

$n(m)$	B	BMCI
320(64)	0.195 ± 0.005	0.153 ± 0.005
640 (128)	0.133 ± 0.005	0.104 ± 0.005
1280 (256)	0.091 ± 0.004	0.071 ± 0.004
2560 (512)	0.063 ± 0.004	0.045 ± 0.004

Errors $CL - C$

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 order to obtain statistically significant comparisons.

In the short simulations, the obtained median confidence intervals are more accurate than the classical confidence intervals from the replication/deletion approach, the observed coverages were all closer to the confidence level. In the long simulations for light and medium load, no statistically significant differences were observed: The coverages were all close to the confidence level, 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 which contain the true value with the confidence level probability.

ρ	Run	RD	MCI
0.25	Short	0.023 ± 0.003	0.017 ± 0.003
	Long	0.003 ± 0.003	0.002 ± 0.003
0.5	Short	0.032 ± 0.004	0.022 ± 0.003
	Long	0.000 ± 0.003	0.004 ± 0.003
0.8	Short	0.056 ± 0.004	0.043 ± 0.004
	Long	0.012 ± 0.003	0.004 ± 0.003

Errors $CL - C$

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

ρ	What	Run	RD, Jackknife	MCI
0.25	$\hat{\lambda}^{(\mathrm{r})}$	Short	0.105 ± 0.005	0.002 ± 0.003
		Long	0.077 ± 0.004	0.002 ± 0.003
	$\hat{W}^{(\mathrm{r})}$	Short	0.091 ± 0.004	0.015 ± 0.003
		Long	0.076 ± 0.004	0.001 ± 0.003
0.5	$\hat{\lambda}^{(\mathrm{r})}$	Short	0.125 ± 0.005	0.008 ± 0.003
		Long	0.076 ± 0.004	-0.001 ± 0.003
	$\hat{W}^{(\mathrm{r})}$	Short	0.100 ± 0.005	0.019 ± 0.003
		Long	0.073 ± 0.004	0.003 ± 0.003
0.8	$\hat{\lambda}^{(\mathrm{r})}$	Short	0.172 ± 0.005	0.018 ± 0.003
		Long	0.083 ± 0.004	0.002 ± 0.003
	$\hat{W}^{(\mathrm{r})}$	Short	0.121 ± 0.005	0.037 ± 0.003
		Long	0.085 ± 0.004	0.003 ± 0.003

Errors $CL - C$

4. Law and Kelton present in their book [5, p. 535] another study where they compare the different confidence level techniques except the standardized time series technique for the model of a time-shared computer system. Although the model is more complicated than the M/M/1 queue, it is statistically friendly: The confidence intervals are much more accurate compared to the M/M/1 model. We do not cite the results because they rely only on 200 independent replications each.

Instead we took the model for a comparative study between the batch means method and batch median confidence intervals.

The model which is due to Adiri and Avi-Itzhak [1] consists of a single central processing unit (CPU) and K terminals. The think times at the terminals are independent exponential random variables with mean 25 seconds. After the think time, the jobs are sent to the CPU and demand there a service time, exponentially distributed with mean 0.8 second. In the CPU, there is a queue, and the jobs are served in a round-robin manner. The CPU allocates to each job at the head of the queue a service quantum of length q =0.1 second. If the (remaining) service time of a job, s seconds, is no more than q, the CPU spends s seconds, plus a fixed overhead of τ=0.015 second, processing the job, which then returns to its terminal, where another think time begins. However, if $s > q$, the CPU spends $q + \tau$ seconds processing the job, which then joins the end of the queue, and its remaining service time is decremented by q seconds. For K=35 terminals, the steady-state mean response time is known to be 8.25 seconds.

In our study we considered sample sizes $n = 320$, 640, 1280, 2560. For batch means the number of batches was 5, hence the batch sizes were $m = 64$, 128, 256, 512. Batch median confidence intervals were constructed with $w = 5$ batches, each with sample size m. This implies a 93.75 percent confidence level CL which we also adopted for the batch means confidence intervals.

The coverages C and the errors $CL - C$ were estimated with 50,200 independent simulations. In the table, errors are given with 90 percent confidence intervals. The figures indicate that for smaller sample sizes the new method is again more accurate than the batch means method. For longer simulations, $m = 256$ and 512, there is no statistically significant difference.

$n(m)$	B	BMCI
320(64)	0.060 ± 0.003	0.050 ± 0.003
640 (128)	0.030 ± 0.003	0.023 ± 0.002
1280 (256)	0.013 ± 0.002	0.012 ± 0.002
2560 (512)	0.004 ± 0.002	0.007 ± 0.002

Errors $CL - C$

5. Now we present an example where the estimator has a very skewed and nonnormal 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 [5, 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.

In simulations we constructed median confidence intervals with w =5 replications. This implies a 93.75 percent confidence level CL which we also adopted for classical confidence intervals. Each replication consisted in m = 1, 2, 3, 4, or 5 outcomes, hence the total sample sizes were $n = wm$ = 5, 10, 20, 40.

The coverages C and the errors $CL - C$ were estimated with 8000 independent simulations. In the table the errors are given for classical confidence intervals (CI) and median confidence intervals (MCI). The 90 percent confidence intervals for the errors are $\pm\epsilon$ with $\epsilon < 0.008$.

Again the MCIs are clearly more accurate than the CIs.

$n(m)$	5 (1)	10 (2)	20 (4)	40 (8)
CI	0.191	0.143	0.105	0.069
MCI	0.147	0.079	0.049	0.032

Errors $CL - C$

Now we present the dream of min-max confidence intervals (MMCI): If the modeller were able to obtain in an efficient way $F_\theta(\theta)$, the value of the estimator's distribution function at θ which is the value of the unknown parameter to be estimated, the real confidence level could be determined according to (2).

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 $\hat{F}$ we calculated the confidence level $\hat{C}L$ according to (2). In the table it can be seen that these estimated confidence levels are very close to the observed coverages C, even in this pathological example.

$n(m)$	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

Coverages and Estimated Confidence Levels

This is a brute force appoach which we cannot recommend due to the needed high effort - until now min-max confidence intervals remain a dream except in some special cases.

6. In our last 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. This implies a 93.75 percent confidence level CL which we also adopted for classical confidence intervals for the median. For the expectation these classical confidence intervals do not exist due to the non-existing variance.

Each replication consisted in $m = 1000$ observations, hence the total sample size was $n = wm = 5000$.

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, $CL - C = 0.016 \pm 0.014$, for $a = 1.5$ quite bad, $CL - C = 0.107 \pm 0.020$, 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 the table the errors are given for classical confidence intervals (CI) and median confidence intervals (MCI) for this order statistic (with 90 percent confidence intervals for the errors).

a	CI	MCI
2	0.082 ± 0.018	0.000 ± 0.013
1.5	0.079 ± 0.018	0.000 ± 0.013
1.1	0.086 ± 0.019	0.005 ± 0.013

Errors $CL - C$

Again the MCIs are clearly more accurate than the CIs.

CONCLUSION

All example results, and the inherent advantages of the median confidence interval technique which are quoted at the beginning of the paper indicate that this new technique is accurate, easy to apply, efficient, and generally applicable.

References

[1] I. Adiri and B. Avi-Itzhak. A time-sharing queue with a finite number of customers. *Journal ACM*, 16:315–323, 1969.

[2] J. Banks. *Handbook of simulation.* Wiley and Sons, New York, 1998.

[3] P. Bratley, B.L. Fox, and L.E. Schrage. *A guide to simulation.* Springer, New York, second edition, 1987.

[4] G. S. Fishman. *Principles of Discrete Event Simulation.* John Wiley and Sons, New York, 1978.

[5] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis.* McGraw-Hill, New York, third edition, 2000.

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

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

[8] J. Ch. Strelen. Median Confidence Intervals. In E.J.H. Kerkhoffs and M. Snorek, editors, *Modelling and Simulation 2001 - Proceedings of the ESM 2001*, pages 771–775. Society for Computer Simulation, 2001.

Johann Christoph Strelen was born in Wiesbaden, Germany, in 1941. He received the Dipl.-Math. and Dr. rer. nat. degrees in mathematics and the Habilitation degree in Computer Science from the Technische Hochschule Darmstadt, Germany, in 1968, 1973, and 1981, respectively. There he was affiliated to the Computing Center (1968-1973), and assistant at the Computer Science Department (1974-1982). In 1973 he was a post doctoral fellow at the IBM Scientific Center, Grenoble, France. Since 1982 he has been a Professor of Computer Science at the Rheinische Friedrich–Wilhelms–Universität Bonn, Germany. His reseach interests include performance evaluation, distributed systems, and simulation. Dr. Strelen is a member of the Gesellschaft für Informatik (GI).