Credibility of the Final Results from Quantitative Stochastic Simulation

G. Ewing, D. McNickle and K. Pawlikowski
Department of Computer Science
University of Canterbury

TR-COSC 05/95, May 1995

The contents of this work reflect the views of the authors who are responsible for the facts and accuracy of the data presented. Responsibility for the application of the material to specific cases, however, lies with any user of the report and no responsibility in such cases will be attributed to the author or to the University of Canterbury.
This technical report contains a research paper, development report or tutorial article which has been submitted for publication in a journal or for consideration by the comissioning organization. We ask you to respect the current and future owner of the copyright by keeping copying of this article to the essential minimum. Any requests for further copies should be sent to the author.

The credibility of estimated confidence intervals for mean values produced by quantitative stochastic simulation is considered. Basic rules of proper experimental studies of quality of such interval estimators is formulated and applied in evaluation of two methods of confidence-interval construction: Batch Means and Spectral Analysis, in the case of simulations run on single and multiple processors.

Credibility of the Final Results from Quantitative Stochastic Simulation

G. Ewing, D. McNickle and K. Pawlikowski*

University of Canterbury
Christchurch, New Zealand

The credibility of estimated confidence intervals for mean values produced by quantitative stochastic simulation is considered. Basic rules of proper experimental studies of quality of such interval estimators is formulated and applied in evaluation of two methods of confidence-interval construction: Batch Means and Spectral Analysis, in the case of simulations run on single and multiple processors.

1. INTRODUCTION

There are many aspects that have to be taken into account in stochastic discrete-event simulation to produce credible final results. They include the necessity for verification of the simulation model (does a given simulation model perform as intended ?) and its validation (is a given simulation model an acceptable model of the real-world system under study ?), selection of statistically correct generator(s) of pseudo-random numbers and, finally, statistically correct analysis of output data collected during simulation. In this paper we address the last of these problems, in the context of *mean value analysis in sequential steady-state stochastic simulation*, ie. simulation conducted for studying systems' behaviour over a long period of time. Sequential analysis of simulation output is generally accepted as the only efficient way for securing representativeness of samples of collected observations (see for example [1]), by stopping the simulation experiment when the relative precision of estimates (defined as the relative width of confidence intervals at an assumed confidence level) reaches the required level. The main analytical problems of sequential estimation of the width of steady-state confidence intervals are discussed eg. in [2]. They are caused by strong correlations between events in typical simulated processes, as well as by initial non-stationary periods.

At least a dozen methods have been proposed for analysing confidence intervals of correlated time-series of observations collected during simulation experiments. A survey of such methods until 1990 can be found in [2]. Newer proposals can be found eg. in [3, 4]. So far only a few implementations of these methods in an automated sequential simulation framework have been reported (see for example [3, 5, 6]) and incorporated in some simulation packages. The problem is that no satisfactorily exhaustive comparative studies of these methods have been reported yet, and it is difficult to find a good method for a specific range of applications. All methods involve different approximations and their quality should be assessed by analysing properties of the final confidence intervals they generate. A good method should produce narrow and stable confidence intervals, which should of course be valid, ie. they should contain the true value of the estimated performance measure. Theoretical studies of various estimators of confidence intervals, reported before 1990, are surveyed in [2]. Newer results can be found for example in [7]. Most of them relate to simulation experiments run on single processors, and very little is known about quality of the methods that could be used in fast concurrent sequential simulation based on Multiple Replications in Parallel

* email address: krys@cosc.canterbury.ac.nz

(MRIP), where multiple processors cooperate in production of data for the global output samples [5].

The theoretical studies of confidence intervals reveal general conditions which have to be satisfied to secure validity of the final confidence intervals, but correctness of any practical implementation of a specific method has to be additionally tested experimentally. In this paper we formulate a new methodology of such experimental studies of the methods used in sequential stochastic simulation for determining the final precision of results, and present the results of our comparative studies of two selected methods: the classical method of (non-overlapping) Batch Means, and SA/HW (the method of Spectral Analysis in its version proposed by Heidelberger and Welch [8]), both in sequential simulations on single processors [2] and in sequential simulations on multiple processors in MRIP scenario. Further directions of research in this area are indicated in the Conclusions.

2. EXPERIMENTAL ANALYSIS OF COVERAGE

In any performance evaluation studies of dynamic systems by means of stochastic discrete-event simulation the final estimates should be determined together with their statistical errors, which are usually measured by the half-width of the final confidence intervals. Restricting our attention to estimators of means, let us assume that we estimate theoretical mean $\mu = EX$ by

$$\bar{X}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

where $x_1, x_2, \ldots, x_n$ are observations collected during simulation. Then, one should also determine

$$P\left(\bar{X}(n) - \Delta \leq \mu \leq \bar{X}(n) + \Delta\right) = 1-\alpha \qquad (2)$$

ie. the confidence interval (c.i.) of μ, at a given confidence level $1-\alpha$, $0<\alpha<1$. Δ is *the half-width of the c.i.*, defined as $\Delta = t_{\kappa,1-\alpha/2}\,\hat{\sigma}[\bar{X}(n)]$, where $\hat{\sigma}^2[\bar{X}(n)]$ is an estimator of the variance of $\bar{X}(n)$, with κ degrees of freedom and $t_{\kappa,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of Student t-distribution.

Problems associated with estimating $\sigma^2[\bar{X}(n)]$ in steady-state simulations are discussed eg. in [2]. Various estimators of this variance have been proposed. This in sequel created the need for an assessment of quality of these estimators and associated with them specific methods of running simulation and pre-processing simulation output data.

Let us note that in an ideal case the final c.i. would contain μ with the probability $1-\alpha$, or equivalently, if an experiment is repeated many times, one would expect to have μ contained in about $(1-\alpha)100\%$ of final confidence intervals. *Coverage of confidence intervals* is defined as the frequency with which the final confidence intervals $(\bar{X}(n)-\Delta \leq \mu \leq \bar{X}(n)+\Delta)$ contain the true value μ. While some interesting results have been achieved in theoretical studies of coverage (see eg. [7, 9-11]), experimental analysis of coverage is still required for assessing the quality of practical implementations of methods used for determining confidence intervals is steady-state simulation. Of course, such analysis is limited to analytically tractable systems, since the value of μ has to be known.

As for any other point estimate, the coverage can be determined together with its c.i. :

$$\left(c - z_{1-\alpha/2}\sqrt{\frac{c(1-c)}{n_c}},\ c + z_{1-\alpha/2}\sqrt{\frac{c(1-c)}{n_c}}\right) \qquad (3)$$

where c is the coverage, $z_{1-\alpha/2}$ is the (1-α/2) quantile of the standard normal distribution and n_c is the (suitably large) number of replicated experiments in the coverage analysis.

An estimator of $\hat{\sigma}^2[\bar{X}(n)]$ used for determining the c.i. of μ is considered as valid, ie. producing valid 100(1-α)% confidence intervals of μ, if the upper bound of the confidence interval of the coverage c in Eq.(3) equals at least (1-α); see [12]. Results of experimental coverage analysis were reported in many publications, although majority of these results was related to simulations on single processors, and very little is known about coverage of estimators that could be used in parallel simulation executed in MRIP scenario [5]. It is strange, but while sequential simulation is generally recognised as the only way of producing results with the required precision since "*... no procedure in which the run length is fixed before the simulation begins can be relied upon to produce a c.i. that covers the true steady-state mean with the desired probability level*" [1, 13], even the original advocates of sequential simulation have applied non-sequential (fixed-sample size) approach in their simulation studies of coverage. In addition, most of reported results on coverage were based on 50-200 replications (see for example [10, 12-18]), which obviously put in question the statistical representativeness of such experimental data. In all these cases, the estimates of coverage were based only a few (!) confidence intervals# which did not cover μ !

It is also generally known that sequential steady-state simulation can produce wrong estimates if the stopping criterion is only temporary satisfied, resulting in too short simulation length. Thus, this cause of *bad confidence intervals* , ie. confidence intervals which do not cover μ, should be also eliminated in coverage analysis, to avoid obscuring the statistical properties of interval estimators by directly unrelated effects. Recognising this fact, we have decided that the following rules should be applied in coverage analysis of sequential interval estimators to produce credible results:

R1. Coverage should be analysed sequentially, following the ordinary rules of sequential simulation, ie analysis of coverage should be stopped when the relative precision (the relative half-width of c.i.) of estimated coverage falls below an assumed level.
R2. An estimate of coverage has to be calculated from a representative sample of data, ie. the analysis can start only after a minimum number of bad confidence intervals has been recorded.
R3. Results from too short simulation runs should be not taken into account.

These rules have been applied in our comparative studies of various methods proposed for controlling the length of sequential steady-state simulation, initial results of which are reported in the next section.

3. NUMERICAL RESULTS

As mentioned, all previously reported studies of experimental coverage analysis conducted for assessing the quality of sequential steady-state interval estimators in stochastic simulation used a fixed-sample size approach [1, 19]. In this section we show initial results of our analysis of coverage conducted in sequential way, following the three rules formulated in Section 2, and compare them with the results one would obtained applying a non-sequential approach.

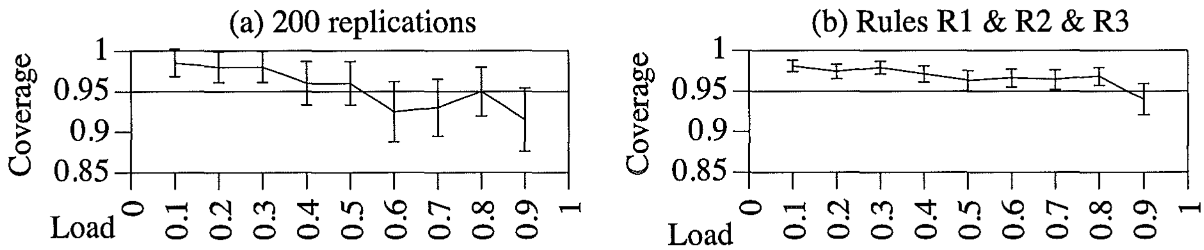
The results presented here are limited to simulations of M/M/1/∞ queuing system only, and two sequential methods of steady-state analysis of the mean values and their confidence intervals: the method of non-overlapping Batch Means (BM), and SA/HW (the method of Spectral Analysis in its version proposed by Heidelberger and Welch [8]). Our implementations of these methods on single processors followed exactly procedures specified in [2], including the procedure described there for detecting the length of initial transient period. In case of parallel

An exemption is [7] where the reported results were averaged over 1000 replications.

simulations in MRIP scenario, BM was used independently by each simulation engine. Thus, the global analyser dealt with a composition of subsequences of (almost independent) batch means, but means submitted by different simulation engines could be calculated over different batch sizes. The parallel version of SA/HW is described in [5].

All reported results were obtained stopping simulations when the final steady-state results gave a relative precision of at least 0.05, at the 0.95 confidence level. All series of replicated simulations were executed using strictly non-overlapping sequences of pseudo-random numbers generated by a multiplicative congruential generator with multiplier 7^5=16807 and modulus 2^{31}-1, which is used for example in such simulation languages as SIMSCRIPT II.5 and GPSS/H [1]. Simulations run for obtaining data for comparative studies of different methods or strategies were initiated using identical pseudo-random number.

Following our three rules of sequential analysis of coverage, the analysis started when N_{min}=30 bad confidence intervals were recorded (rule R2). At this point, the mean and standard deviation of simulation run lengths were calculated, and data obtained from simulations shorter than one standard deviation from the mean run length were discarded (rule R3). Next, if the number of bad confidence intervals was not smaller than N_{min}, the coverage was estimated sequentially, with calculations being repeated after each suffiently long replication.



(c) 200 replications

Load	Good c.i.	Bad c.i.
0.1	197	3
0.2	196	4
0.3	196	4
0.4	192	8
0.5	192	8
0.6	185	15
0.7	186	14
0.8	190	10
0.9	183	17

(d) Rules R1 & R2

Load	Good c.i.	Bad c.i.
0.1	1590	30
0.2	1146	30
0.3	1385	30
0.4	1005	30
0.5	616	30
0.6	515	30
0.7	473	30
0.8	523	30
0.9	444	42

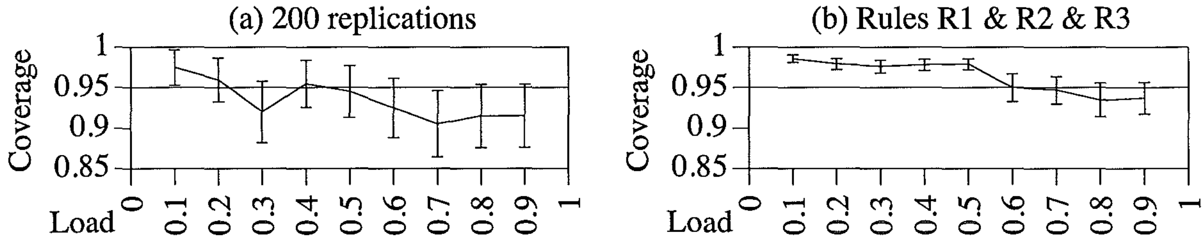
(e) Rules R1 & R2 & R3

Load	Good c.i.	Bad c.i.	Rejected
0.1	1590	30	0
0.2	1146	30	0
0.3	1385	30	0
0.4	1005	30	0
0.5	772	30	152
0.6	865	30	143
0.7	793	30	146
0.8	895	30	164
0.9	470	30	98

Figure 1. Coverage analysis of BM, one processor

The results obtained for simulations on single processors are shown in Fig.1 and 2, with case (a) showing results of traditional fixed-sample size coverage analysis (over 200 replications), case (b) showing results obtained following our methodology of sequential analysis of coverage, ie. applying rule R1&R2&R3. The three tables in each figure show the number of good and bad confidence intervals used to estimate coverage in: the non-sequential approach; the sequential approach with adopted rule R1&R2 only (case (c) and(d), respectively); as well as the number of good and bad confidence intervals, and the number of too short simulation runs recorded in sequential coverage analysis based on rule R1&R2&R3.

It appears that the traditional approach cannot produce reliable estimates of coverage, although in these specific examples it underestimated the quality of results produced by BM and SP/HW. Comparing case (d) and (e) one can see that rule R3 becomes more important in more heavier loaded systems, when correlations between observations increase. In the range of traffic load considered, $0.1 \le \rho \le 0.9$, both methods remain valid.



(c) 200 replications

Load	Good c.i.	Bad c.i.
0.1	195	5
0.2	192	8
0.3	184	16
0.4	191	9
0.5	189	11
0.6	185	15
0.7	181	19
0.8	183	17
0.9	183	17

(d) Rules R1 & R2

Load	Good c.i.	Bad c.i.
0.1	885	30
0.2	593	30
0.3	395	32
0.4	536	30
0.5	427	30
0.6	416	36
0.7	416	36
0.8	453	44
0.9	435	40

(e) Rules R1 & R2 & R3

Load	Good c.i.	Bad c.i.	Rejected
0.1	2133	30	163
0.2	1376	30	104
0.3	1199	30	71
0.4	1352	30	83
0.5	1394	30	78
0.6	573	30	55
0.7	538	30	66
0.8	432	30	66
0.9	449	30	69

Figure 2. Coverage analysis of SA/HW, one processor

Fig.3 depicts the results obtained from sequential coverage analysis of BM and SP/HW on P=2 and 4 processors. As these results show, the quality of BM decreases with the level of parallelization and the method becomes even invalid for heavier loaded systems. On the other hand, the quality of SP/HW remains practically unchanged. A word of caution: our preliminary results should not be used for drawing general conclusions about the quality of BM and SP/HW when $\rho > 0.9$ or $P > 4$, as well as about their behaviour in simulations of arbitrary systems, until more exhaustive studies are conducted.

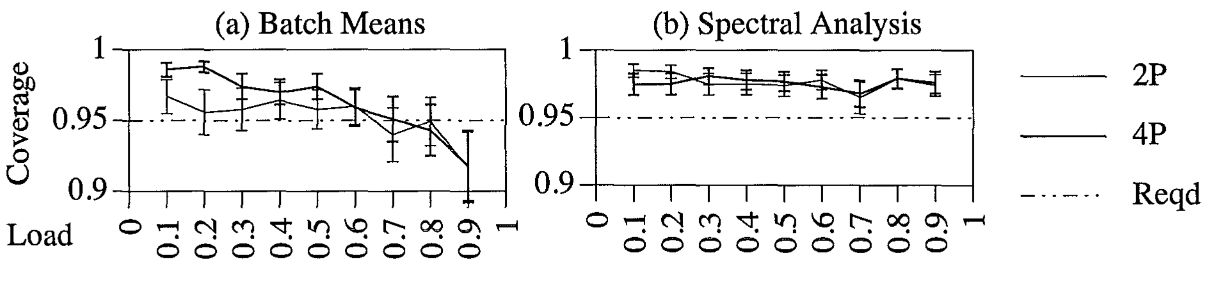


Figure 3. Coverage of BM and SA/HW in MRIP, P = 2 and 4, rules R1 & R2 & R3

4. CONCLUSIONS

We have formulated basic rules that should be followed in proper experimental analysis of coverage of different steady-state interval estimators. Our main argument is that such analysis should be done sequentially, The numerical results of our preliminary coverage analysis of the method of Batch Means and Spectral Analysis have been also presented and compared with those obtained by traditional, non-sequential approach. As advocated in [19], to draw more general conclusions about performance of interval estimators used in various methods of sequential steady-state simulation one needs to consider a number of different simulation models, since the results obtained for one system (in this paper: $M/M/1\infty$) are not sufficient.

REFERENCES

1. Law, A.M., and W.D.Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, NY, 1992.
2. Pawlikowski, K. "Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions". *ACM Computing Surveys*, no.2, 1990, 123-170
3. Fox, B.L., D.Goldsman and J.Swain. "Spaced Batch Means".*Operations Res. Letters*, vol.10, 1991, 255-263
4. Howard, R.B. et al. “Confidence Intervals for Univariate Discrete-Event Simulation Output Using the Kalman Filter”. Proc. 1992 Winter Simulation Conf., IEEE Press, 1992, 586-593
5. Pawlikowski, K., V. Yau and D. McNickle. "Distributed Stochastic Discrete-Event Simulation in Parallel Time Streams". Proc. 1994 Winter Simulation Conf., IEEE Press, 1994, 723-730
6. Yau, V., and K.Pawlikowski. "AKAROA: a Package for Automatic Generation and Process Control of Parallel Stochastic Simulation". Proc. 16th Australian Computer Science Conf., *Australian Computer Science Comms.,* 1993, 71-82
7. Kang, K., and D.Goldsman. "The Correlation Between Mean and Variance Estimators in Computer Simulation". *Trans. of IIE,* March 1990, 15-23
8. Heidelberger, P., and P.D. Welch. "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulation". *Comms. of ACM*, vol.25, 1981, 233-245
9. Glynn, P.W. "Coverage Error for Confidence Intervals Arising in Simulation Output Analysis". Proc. 1982 Winter Simulation Conf., IEEE, N.Y., 1982, 369-375.
10. Schriber, T.J. and R.W. Andrews. "A Conceptual Framework for Research in the Analysis of Simulation Output." *Comm. of the ACM*, 1981, 218-232
11. Schruben, L.W. "A Coverage Function for Interval Estimators of Simulation Response." *Management Sci.*, 1980, 18-27.
12. Sauer, C.H., and S.S.Lavenberg. "Confidence Intervals for Queueing Simulations of Computer Systems". *ACM Performance Evaluation Review*, vol.8, no.1-2, 46-55
13. Law, A.M., and W.D.Kelton. “Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures”. Management Sci., vol.28, no.5, 1982, 550-562
14. Adam, N.R. "Achieving a Confidence Interval for Parameters Estimated by Simulation". *Management Sci.*, 1983, 856-866.
15. Heidelberger, P., and P.D. Welch. "Simulation Run Length Control in the Presence of an Initial Transient". *Operations Res..*, 1983, 1109-1144
16. Kelton, W.D., and A.M.Law. "An Analytical Evaluation of Alternative Strategies in Steady-State Simulation". *Operations Res.*, 1984, 169-184.
17. Lavenberg, S.S., and C.H.Sauer. “Sequential Stopping Rules for the Regenerative Method of Simulation”. IBM J. Research Development, vol.21, 1977, 667-678
18. Law, A.M., and J.S.Carson. "A Sequential Procedure for Determining the Length of a Steady-State Simulation". *Operations Res.,* 1979, 1011-1025
19. Law, A.M. "Statistical Analysis of Simulation Output Data". *Operations Res.*, 1983, 983-1029