Experimental Coverage Analysis of Interval Estimators for Sequential Stochastic Simulation

Jong-Suk Ruth Lee Krzysztof Pawlikowski
Department of Computer Science
Donald C. McNickle
Department of Management
University of Canterbury
Christchurch, New Zealand

Abstract

Stochastic discrete-event simulation has become one of the most widely used tools for performance evaluation of complex stochastic dynamic systems in many areas of science and engineering. We address the problem of the statistical correctness of final simulation results in the context of sequential steady-state simulation, conducted for studying long run behaviour of stable dynamic systems. Due to various approximations, the final confidence intervals can cover estimated theoretical values at much below the frequency suggested by the assumed confidence levels. On the basis of exhaustive experimental analysis, we formulate basic rules for the proper experimental analysis of the coverage of sequential steady-state interval estimators. The numerical results of our coverage analysis for the method of non-overlapping batch means, spectral analysis and the regenerative method are presented.

1 Introduction

Stochastic discrete-event simulation has become one of the most widely used tools for performance evaluation of complex stochastic dynamic systems in many areas of science and engineering. One problem faced by simulators is to obtain credible final results. Attempts at solving this problem in the case of steady-state simulation, when simulation output data are typically strongly correlated, have led to different interval estimators, used in different methods of simulation output data analysis. In this paper, we address the problem of the statistical correctness of the final simulation results in the context of sequential steady-state simulation, conducted for studying long run mean values of performance measures of stable dynamic systems.

Sequential statistical analysis of output data in stochastic simulation, used for controlling the length of simulation, is regarded as the only practical way of securing an appropriate level of credibility of the final simulation results (Law and Kelton [9], and Pawlikowski [13]). In this approach, simulation progresses from one checkpoint to the next one, until prespecified accuracies of the point estimators of interest

are obtained. Probably the most commonly used stopping criterion for sequential steady-state simulation is the *relative precision* of results, defined as the ratio of the current half-width of the confidence interval (at a given confidence level) to the current point estimate of a given estimated performance measure (Pawlikowski [12]). An experiment is stopped at the checkpoint at which the required relative precision of the final results is reached.

One of the main quality criteria used for assessing methods of simulation output data analysis in stochastic simulation is the coverage of the final confidence intervals they produce, defined as the proportion of confidence intervals which contain the true value of an estimated performance measure. Any good method of analysis of simulation output data should produce narrow and stable confidence intervals, and the relative frequency with which such intervals contain the true value of the estimated performance measure should not differ too much from the assumed theoretical confidence level.

As recently argued in Pawlikowski et al. [14], coverage analysis should be also conducted sequentially, to secure statistically accurate results. The rules of sequential coverage analysis have been proposed in Pawlikowski et al. [14]. In this paper, a revision of those rules is presented. This is an enhanced version of sequential coverage analysis, based on the F distribution, which leads to more efficient interval estimators of proportions, see Lee et al. [10]. In this paper coverage analysis is also studied experimentally, in the context of sequential steady-state simulation, assuming three different methods of mean value analysis: non-overlapping Batch Means (BM), Spectral Analysis (SA) in its version proposed by Heidelberger and Welch [7], and regenerative method (RM) (Crane and Iglehart [1], Crane and Lemoine [2], and Shedler [17]). The theoretical bases of these three methods of simulation output data analysis, and sequential implementations of two first methods, are given for example in Pawlikowski [12]. Sequential implementation of RM is described in Lee [8].

In steady-state simulation output data analysis, such as SA and BM, one has to discard data collected during the initial transient period. Determination of the length of the initial transient period is often nontrivial and likely to require sophisticated statistical techniques (Pawlikowski [12]). RM, also known as regenerative simulation, avoids this problem. During the regenerative simulation, simulated processes are regarded as regenerative stochastic processes. Following RM, one groups data collected during different regenerative cycles (RCs) in batches (of independent and identically distributed data), and the final precision of results depends on the number of RCs recorded.

Numerical results of coverage analysis of the three sequential methods applied for estimating steady-state means were obtained in our quest for the most robust method of sequential analysis of simulation output data, to be implemented in Akaroa2 (Ewing et al. [4]), a fully automated controller of distributed stochastic simulation on multiple networked processors, in the Multiple Replications In Parallel (MRIP) scenario (Pawlikowski et al. [15]). The results of coverage analysis of sequential methods of estimation of steady-state quantiles were reported in Lee et al. [11].

Rules of experimental coverage analysis are formulated in Section 2, and the numerical results are presented in Section 3.

2 Experimental Analysis of Coverage

Following the most basic principles of scientific experimentation, the final results from performance evaluation studies of stochastic dynamic systems by means of discrete-event simulation should be always determined together with their statistical errors. These errors are usually measured by the half-width of the final confidence intervals. But the methods proposed for estimating the confidence intervals of different performance measures (such as mean values, variances, probabilities, quantiles etc.) are based on different simplifying assumptions, which can cause the experimental confidence level (or coverage) of the final confidence intervals to differ significantly from the assumed (theoretical) confidence level. The aim of experimental coverage analysis is to find the best method(s) (in the sense of coverage) for sequential analysis of simulation output data.

As justified in Pawlikowski et al. [14], only sequential coverage analysis can lead to credible final conclusions regarding the quality of any method of simulation output analysis. Experimental results presented there clearly show high initial instability of coverage. To avoid this region, coverage analysis has to be done over a sufficiently large sample of data (in this case: after sequential simulation is repeated sufficiently many times).

Following this approach, one needs to estimate the proportion of confidence intervals covering the theoretical value of interest. This of course means that experimental coverage analysis is restricted to analytically tractable systems.

Let us consider independently repeated simulation experiment, and let p be the probability that a final confidence interval (obtained from a replication) covers the theoretical value. Then, this probability p can be estimated by a proportion $\hat{p}$

$$\hat{p} = \frac{count\ \ of\ \ successes\ \ in\ \ sample}{size\ \ of\ \ sample} = \frac{s}{n},$$

where a "success" means that a confidence interval covers the theoretical value, and n is the number of replications executed.

The accuracy with which $\hat{p}$ estimates p can be assessed by the probability

$$P(\hat{p} - \Delta_1 \leq p \leq \hat{p} + \Delta_2) = 1 - \alpha$$

where Δ_1 and Δ_2 are the offset for the lower and the upper limit of the confidence interval of p, and $(1 - \alpha)$ is the confidence level, $0 < \alpha < 1$. Note that, if the simulation results are obtained at an assumed confidence level of $(1 - \alpha_0)$ and the simulation experiment were repeated many times, the final confidence intervals would be expected to contain the exact theoretical value in $p = 100(1 - \alpha_0)\%$ of cases. Unfortunately, methods used for determining confidence intervals are based on various approximations, which can lower the coverage (experimental confidence level) of the method much below the assumed confidence level. As mentioned, the robustness of methods of sequential estimation is usually discussed in the context of

coverage. Sauer [16] proposed that the method used for determining the confidence interval of the point estimate at a given confidence level $(1 - \alpha_0)$ is considered as valid, if the upper limit of the confidence interval of p equals at least $(1 - \alpha_0)$.

As discussed in Lee et al. [10], the interval estimator of coverage should be based on the F distribution, to ensure that the sequential analysis of coverage produces realistic estimates. A $100(1 - \alpha)\%$ lower limit of the confidence interval for the proportion p is

$$\hat{p}_l = \hat{p} - \Delta_1 = \frac{n\hat{p}}{n\hat{p} + (n - n\hat{p} + 1)f_{1-\alpha/2}(r_1, r_2)}$$

and a $100(1 - \alpha)\%$ upper limit of the confidence interval for the proportion p is

$$\hat{p}_u = \hat{p} + \Delta_2 = \frac{(n\hat{p} + 1)f_{1-\alpha/2}(r_3, r_4)}{(n - n\hat{p}) + (n\hat{p} + 1)f_{1-\alpha/2}(r_3, r_4)}$$

where $f_{1-\alpha/2}(r_1, r_2)$ and $f_{1-\alpha/2}(r_3, r_4)$ are the $(1-\alpha/2)$ quantile of the F distribution with (r_1, r_2) and (r_3, r_4) degrees of freedom, where $r_1 = 2 * (n - n\hat{p} + 1)$, $r_2 = 2 * n\hat{p}$, $r_3 = 2 * (n\hat{p} + 1)$, and $r_4 = 2 * (n - n\hat{p})$ (Hald [6]).

Any sequential simulation experiments may stop after too few simulation runs have been collected, if, by chance, the stopping criteria has been temporarily satisfied. As shown in Pawlikowski et al. [14], this happens in real simulation experiments from time to time and can make estimates of coverage unreliable.

On the basis of exhaustive experimental analysis, results of which are presented in Section 3, we reformulate basic rules for the proper experimental analysis of the coverage of sequential steady-state interval estimators as follows:

- **Rule 1 :** Coverage should be analysed sequentially, i.e. analysis of coverage should be stopped when the *absolute precision* of the estimated coverage satisfies a specified level which is sufficiently small.
- **Rule 2 :** An estimate of coverage has to be calculated from a representative sample of data, so the coverage analysis can start only after a minimum number of 'bad' confidence intervals have been recorded.
- **Rule 3 :** Results from simulation runs that are clearly too short should not be taken into account.
- **Rule 4 :** An interval estimator which is based on the F distribution of coverage should be used to ensure that the sequential analysis of coverage produces realistic estimates.

Details of our experimental results of these revised rules of sequential coverage analysis for studying quality of the final steady-sate interval estimators of mean values are discussed in Section 3.

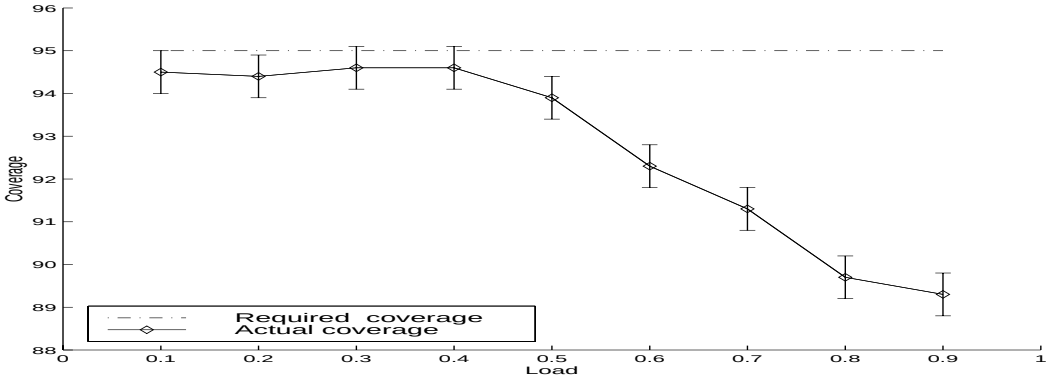

Figure 1: Sequential coverage analysis for sequential BM using F distribution ($M/M/1/\infty$)

3 Numerical Results

We consider three sequential methods of steady-state analysis of means: the method of non-overlapping BM, SA in its version proposed by Heidelberger and Welch [7], and RM (Crane and Iglehart [1], Crane and Lemoine [2], and Shedler [17]). Our implementations of the first two methods follow exactly the procedures specified in Pawlikowski [12] and the last method is in Lee [8].

All results of our sequential coverage analysis were obtained assuming the required precision of the final result was 1% or better, at a confidence level of 0.95 using an $M/M/1/\infty$ queueing system as a reference simulation model. Additionally, the interval estimator of coverage was based on the F distribution to ensure that the sequential analysis of coverage produces more realistic estimates (Lee et al. [10]).

As justified in Pawlikowski et al. [14], one can clearly see that the sequential coverage analysis, with filtering of too short simulation runs and with a requirement of a minimum of bad confidence intervals, produces more reliable results. Therefore, in a practical implementation of **Rules 1 - 3**, we assume that representativeness of data for coverage analysis requires that a minimum of 200 bad confidence intervals have to be recorded before sequential analysis of coverage can commence and then the results from all simulation runs shorter than a threshold (one standard deviation below the mean of the run lengths) are discarded. Removing the statistical 'noise' introduced by too short and unrepresentative simulation runs improves the conclusions we can make about the quality of a given method of simulation output analysis.

The results of our sequential coverage analysis of the method of BM, RM, and SA are presented in Figures 1, 2, and 3. These numerical results are very similar to the results reported in Pawlikowski et al. [14]. The difference is that the final half-width of the confidence intervals at different traffic levels are the same. (Note that the figures are drawn to different scales.)

Ideally, the confidence interval of coverage for a method of simulation output data analysis should contain the confidence level assumed for the final results (Sauer [16]).

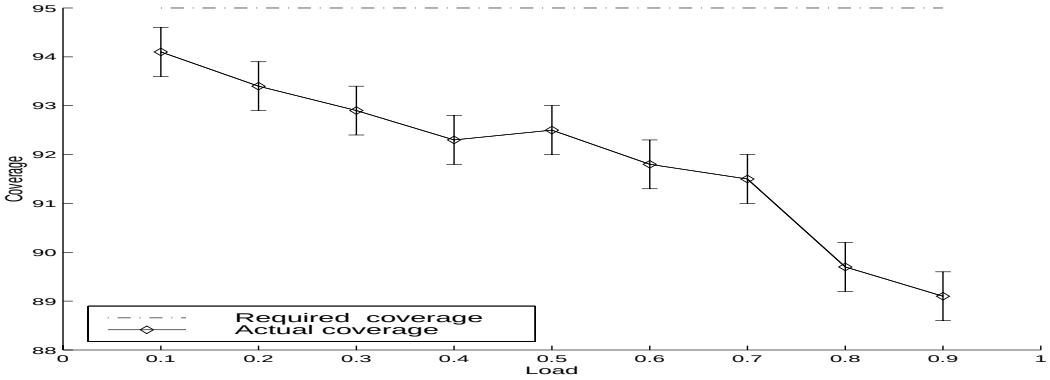

Figure 2: Sequential coverage analysis for sequential RM using F distribution ($M/M/1/\infty$)

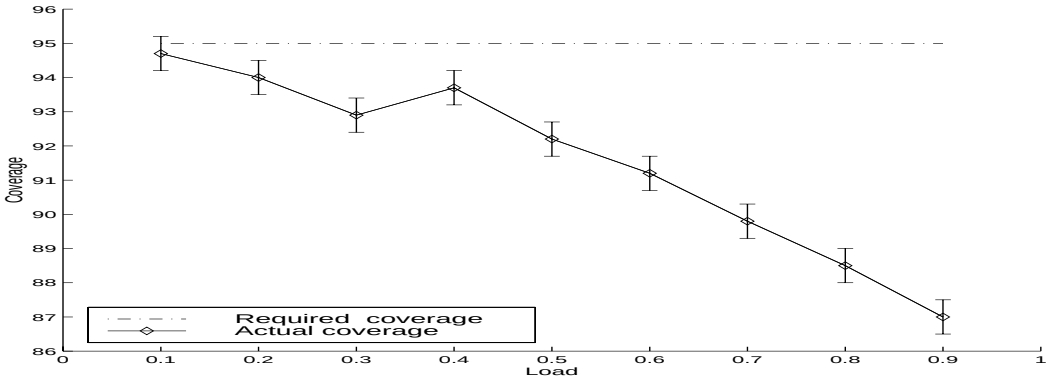

Figure 3: Sequential coverage analysis for sequential SA using F distribution ($M/M/1/\infty$)

In practice, this criterion is hardly met by any method, so, making this requirement weaker, we accept the method for practical applications if the confidence interval of its coverage is sufficiently close to the confidence level assumed.

The final coverage of three sequential methods is still far from the required level, especially in highly correlated systems. These could be caused by the fact that an insufficient number of observations is collected in each simulation run, as shown in Figure 4. This is clear from the comparison of the theoretical and experimental run lengths of sequential simulation, under BM, RM, and SA, obtained from 10,000 independent replications of steady-state simulations of the $M/M/1/\infty$, $M/D/1/\infty$, and $M/H_2/1/\infty$ queueing systems, estimating the mean response time, with at least 0.1 as the upper level of the acceptable relative error of the final results, at a confidence level of 0.95. Figures 4 - 6 show the average numbers of observations collected in experiments and these required theoretically (see Daley [3] and Fishman [5]).

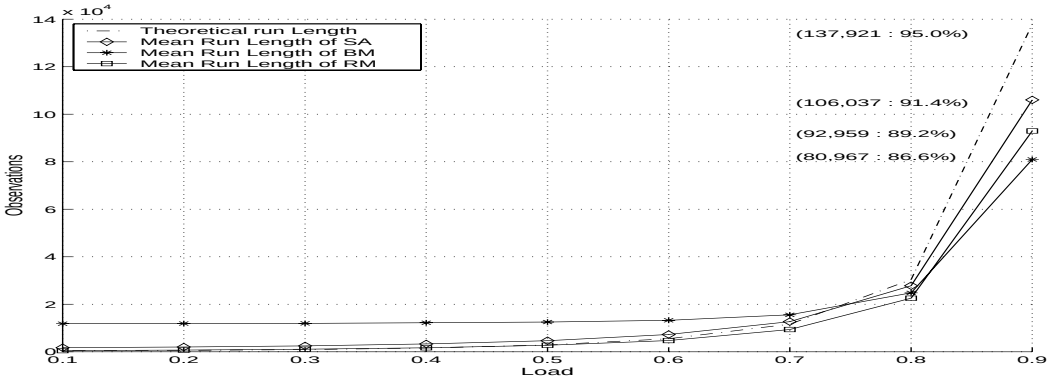

Figure 4: Mean run length of 10,000 independent simulation runs of estimating mean response time at $M/M/1/\infty$ queueing system

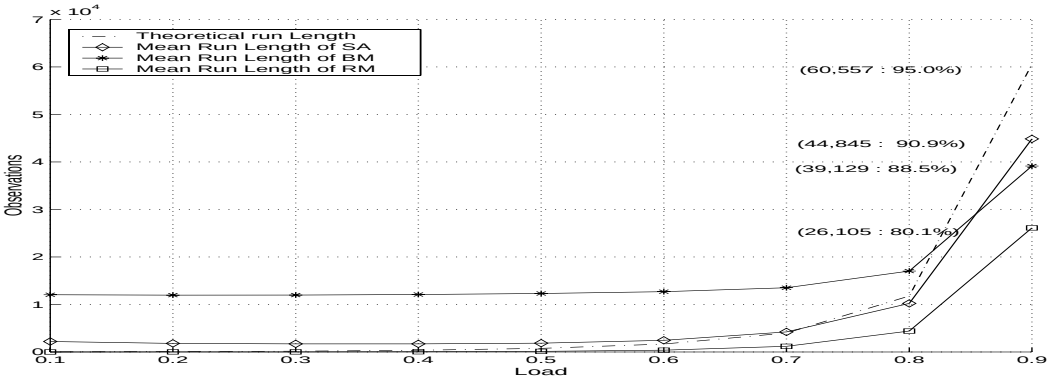

Figure 5: Mean run length of 10,000 independent simulation runs of estimating mean response time at $M/D/1/\infty$ queueing system

When we compare the numbers of observations needed theoretically with those used in experiments, they are nearly the same at the light traffic intensities, except the sequential BM. This is because the finally accepted batch size in sequential BM is unnecessarily large. As a consequence, all three methods produce acceptable coverage at light traffic intensities. For heavier traffic intensities, all three methods collect smaller numbers of observations than theoretically required.

4 Conclusions

In this paper, we address the problem of the statistical correctness of the final simulation results in the context of sequential steady-state simulation, conducted for studying long run behaviour of stable dynamic systems. Such simulations are often stopped when the relative half-width of confidence intervals of point estimates becomes sufficiently small. The problem is that, due to various approximations,

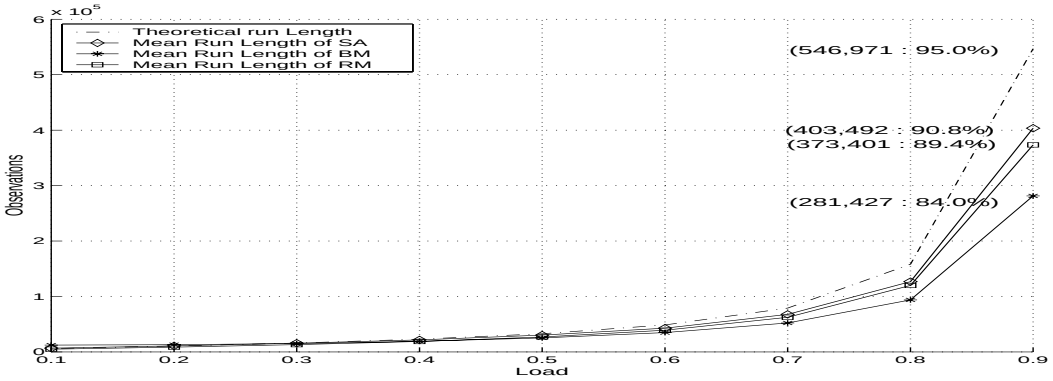

Figure 6: Mean run length of 10,000 independent simulation runs of estimating mean response time at $M/H_2/1/\infty$ queueing system

the final confidence intervals can cover estimated theoretical values at much lower frequency than that suggested by the assumed confidence level. On the basis of exhaustive experimental analysis, we reformulate basic rules for the proper experimental analysis of the coverage of sequential steady-state interval estimators. The numerical results of our coverage analysis for the method of non-overlapping batch means, spectral analysis and regenerative method are presented. We also compare the theoretical and empirical run lengths of sequential steady-state simulation, to indicate the cause of poor coverage of results from simulations of highly correlated processes.

References

[1] Crane, M. A. and Iglehart, D. L. (1975), Simulating Stable Stochastic Systems:III. Regenerative Processes and Discrete Event Simulations, *Operations Research*, **23(1)**, 33-45.

[2] Crane, M. A. and Lemoine, A. J. (1977), *An Introduction to the Regenerative Method for Simulation Analysis*, Springer Verlag.

[3] Daley, D. J. (1968), The Serial Correlation Coefficients of Waiting Times in a Stationary Single Server Queue, *Journal of the Aust. Math. Soc.*, **8**, 683-699.

[4] Ewing, G., Pawlikowski, K., and McNickle, D. (1999), Akaroa 2: Exploiting Network Computing by Distributed Stochastic Simulation, in *Proceedings of 13th European Simulation Multiconference*, Warsaw, Poland, June 1999.

[5] Fishman, G. S. (1978), Grouping Observations in Digital Simulation, in *Management Science*, **24(5)**, 510-521.

[6] Hald, A. (1952), *Statistical Theory with Engineering Applications*, John Wiley and Sons, Inc.

[7] Heidelberger, P. and Welch, P. D. (1981), A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations, *Communications of the ACM*, **25**, 233-245.

[8] Lee, J.-S. R. (1999), Coverage Analysis of Sequential Regenerative Simulation, *Technical Report TR-COSC 02/99*, Department of Computer Science, University of Canterbury, Christchurch, New Zealand.

[9] Law, A. M. and Kelton, W. D. (1982), Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures, *Management Science*, **28(5)**, 550-562.

[10] Lee, J. R., McNickle, D., and Pawlikowski, K. (1999a), Confidence Interval Estimators for Coverage Analysis in Sequential Steady-State Simulation, in *Proceedings of the 22nd Australasian Computer Science Conference*, Auckland, New Zealand, Jan., 87-98.

[11] Lee, J. R., McNickle, D., and Pawlikowski, K. (1999b), Quantile Estimation in Sequential Steady-State Simulation, in *Proceedings of 13th European Simulation Multiconference*, Warsaw, Poland, June 1999.

[12] Pawlikowski, K. (1990), Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions, *ACM Computing Surveys*, **2**, 123-170.

[13] Pawlikowski, K. (1999), Simulation Studies of Telecommunication Networks and Their Credibility, in *Proceedings of 13th European Simulation Multiconference*, Warsaw, Poland, June 1999.

[14] Pawlikowski, K., McNickle, D., and Ewing, G. (1999), Coverage of Confidence Intervals in Sequential Steady-State Simulation, *Simulation Practice and Theory*, **6(2)**, 255-267 and **7**, 105.

[15] Pawlikowski, K., Yau, V., and McNickle, D. C. (1994), Distributed and Stochastic Discrete-event Simulation in Parallel Time Streams, In *Proceedings of the 1994 Winter Simulation Conference*, Orlando, Florida, Dec., 723-730.

[16] Sauer, C. H. (1979), Confidence Intervals for Queueing Simulations of Computer Systems, *ACM Performance Evaluation Review*, **8(1-2)**, 46-55.

[17] Shedler, G. S. (1993), *Regenerative Stochastic Simulation*, Academic Press, Inc.