Coverage of confidence intervals in sequential steady-state simulation*

Krzysztof Pawlikowski, Donald C. McNickle♠ and Gregory Ewing
Department of Computer Science and Department of Management♠
University of Canterbury
Christchurch, New Zealand

Abstract

Stochastic discrete-event simulation has become one of the most-used tools for performance evaluation in science and engineering. But no innovation can replace the responsibility of simulators for obtaining credible results from their simulation experiments. In this paper we address the problem of the statistical correctness of simulation output data analysis, in the context of sequential steady-state stochastic simulation, conducted for studying long run behaviour of stable systems. Such simulations are stopped as soon as the relative precision of estimates, defined as the relative half-width of confidence intervals at a specified confidence level, reaches the required level. We formulate basic rules for the proper experimental analysis of the coverage of steady-state interval estimators. Our main argument is that such an analysis should be done sequentially. The numerical results of our coverage analysis of the method of Non-overlapping Batch Means and Spectral Analysis are presented, and compared with those obtained by the traditional, non-sequential approach. Two scenarios for stochastic simulation are considered: traditional sequential simulation on a single processor, and fast concurrent simulation based on Multiple Replications in Parallel (MRIP), with multiple processors cooperating in the production of output data.

Keywords: stochastic simulation, simulation output analysis, confidence intervals, observed coverage

---

* An abbreviated version of this paper was presented as [4] at the 1995 EUROSIM Congress in Vienna.

1. Introduction

There are many aspects that have to be taken into account in stochastic discrete-event simulation to produce credible results. They include the necessity for verification of the simulation model (does a given simulation model perform as intended?) and for 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 the simulation. In this paper we address the last problem, in the context of *estimating means in sequential steady-state stochastic simulation*, i.e., simulation conducted for studying mean system behaviour over a long period of time. Sequential analysis of simulation output is generally accepted as the only efficient way for ensuring representativeness of samples of collected observations; see for example [11, 18, 19]. In this scenario, a simulation experiment is stopped as soon as the relative precision of estimates, defined as the relative width of confidence intervals at a specified confidence level, reaches the required size. The main analytical problems of sequential estimation of the width of steady-state confidence intervals are discussed in for example [21]. They are caused by strong correlations between events in typical simulated processes, as well as by the presence of 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 [21]. Newer proposals can be found in for example [3, 7, 12]. So far only a few implementations of these methods in an automated sequential simulation framework have been reported (see for example [3, 11, 22, 24, 29]). The problem is that no reliable 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 experimentally assessed by studying the properties of the final confidence intervals they generate. A good method should produce narrow and stable confidence intervals, which should of course yield a probability of such an interval containing the true value of the estimated performance measure that does not differ from the assumed confidence level. Theoretical studies of various interval estimators up to 1990 are surveyed in [21]. Newer results can be found, for example, in [6] and [13]. Most of them relate to simulation experiments run on single processors, and very little is known about the quality of methods that could be used in fast concurrent sequential simulation based on Multiple Replications in Parallel (MRIP), where multiple processors cooperate in the production of data for the global output samples [22 - 24].

The theoretical studies of confidence intervals can reveal general conditions which have to be satisfied to secure correct coverage, but correctness of any practical implementation of a specific method also has to be tested experimentally. Unfortunately, no appropriate

methodology of experimental coverage analysis had been proposed, and this prompted us to formulate such a methodology in Section 2 of this paper. We apply this methodology to compare the quality of two selected methods of (automated) output data analysis in sequential steady-state simulation: the classical method of (non-overlapping) Batch Means, and SA/HW (the method of Spectral Analysis in its version proposed by Heidelberger and Welch [10]), both in the case of sequential simulations on single processors and in sequential simulations on multiple processors in the MRIP scenario. The main results of these analyses are presented in Section 3. Further directions of research, and related practical problems, are indicated in the Conclusions.

2. Experimental analysis of coverage

In any performance evaluation study 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 the theoretical mean $\mu_x = EX$ by

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

where $x_1, x_2, ..., x_n$ are observations collected during simulation. Then, one should also determine the confidence interval (c.i.) for μ_x, at a given confidence level $1-\alpha$, $0 < \alpha < 1$

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

where Δ is *the half-width of the c.i.*, typically estimated by $\hat{\Delta} = t_{\kappa,1-\alpha/2}\, \hat{\sigma}\,[\bar{X}(n)]$ 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 the Student t-distribution.

Problems associated with estimating $\sigma^2[\bar{X}(n)]$ in steady-state simulations are discussed for example in [21]. Various estimators of this variance have been proposed, which has created the need for an assessment of the quality of these estimators and of specific methods of running the simulation and pre-processing simulation output data.

In an ideal case the final c.i. would contain μ_x with the probability $1-\alpha$, or equivalently, if an experiment were repeated many times, one would expect to have μ_x in about $(1-\alpha)100\%$ of the final confidence intervals. *Coverage of confidence intervals* , c, is defined as the relative frequency with which the final confidence interval $\left(\bar{X}(n) - \hat{\Delta}\ ,\ \bar{X}(n) + \hat{\Delta}\right)$ contains the true value μ_x. While some interesting results have been achieved in theoretical studies of coverage (see for example [5, 11, 13, 26, 27]), 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_x has to be known.

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

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

where c is the coverage, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and n_c is the (suitably large) number of replicated coverage experiments. This is based on the fact that, while the number of confidence intervals containing the true value μ_x has a binomial distribution with mean $n_c\mu_c$, $(c-\mu_c)\sqrt{c(1-c)}$ tends to the standard normal distribution as $n_c \to \infty$; see for example [20].

An estimator of $\hat{\sigma}^2[\overline{X}(n)]$ used for determining the c.i. of μ_x is considered as valid, ie. producing valid $100(1-\alpha)\%$ confidence intervals of μ_x, if the upper bound of the confidence interval of the coverage c in Eq.(3) equals at least $(1-\alpha)$; see [25]. Results of experimental coverage analysis have been reported in many publications but, unfortunately, the statistical validity of many of these results can be questioned. The coverage was often not analysed on the basis of a large number of replications. We have found only four studies [8, 9, 12, 13] where at least a thousand replications were used. Unfortunately in many reported cases as few as 50-200 replications were used (see for example [1, 11, 14-17, 25, 26, 28]) which obviously puts in question their statistical representativeness. Inevitably, in these cases, the estimates of coverage were based on only a few confidence intervals which did not cover μ_x !

Additionally, while sequential simulation is generally regarded 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*" ([7, 17]), even the original advocates of sequential simulation have applied non-sequential (fixed-sample size) approaches in their own simulation studies of coverage. Certainly, if one accepts the arguments for a sequential approach as the only practical way, then meta-simulation experiments, such as those for coverage analysis, *should also be run sequentially* !

Sequential coverage analysis does raise the problem that some of the simulation experiments may stop after an abnormally short time, because, by chance, the stopping criterion has been temporarily satisfied. While of course this occurs in actual simulation experiments here it has the effect of introducing considerable “noise” into our estimates of coverage, and making them difficult to compare with results from fixed sample size studies. Also we believe that, in practice, a careful analyst might well eliminate such obviously flawed

runs. Thus we decided to eliminate such runs to avoid obscuring the statistical properties of interval estimators.

Taking these facts into account, we adopt the following rules in coverage analysis of sequential steady-state interval estimators :

R1. Coverage should be analysed sequentially, ie. analysis of coverage should be stopped when the relative precision (the relative half-width of c.i.) of the estimated coverage satisfies a specified level.

R2. 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.

R3. Results from simulation runs that are clearly too short should not be taken into account.

Details of our implementation of these rules of sequential coverage analysis for studying quality of the final steady- state interval estimators of mean values are discussed in the next section.

3. Numerical results

In this Section we consider two sequential methods of steady-state analysis of means 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 [10]). Our implementations of these methods on single processors followed exactly the procedures specified in [21], including the procedure described there for detecting the length of the initial transient period. For parallel simulations using the MRIP scenario, BM was used independently by each simulation engine ([22, 29]). Thus, the global analyser dealt with a composition of subsequences of (almost independent) batch means, but the mean values submitted by different simulation engines could be calculated over different batch sizes. The parallel version of SA/HW is described in [22].

All the numerical results presented in this Section were obtained using the M/M/1/∞ queuing system as the reference simulation model. Simulations of this queueing system were stopped as soon as the steady-state results reached a relative precision of at least 0.05 at the 0.95 confidence level, where relative precision is defined as the ratio of the current half-width of the confidence interval of mean to the current value of the estimated mean. 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. This generator is used in simulation languages such as SIMAN

and SLAM II [19]. Simulations runs for comparing different methods or strategies were performed using identical pseudo-random numbers.

In a practical implementation of the rules R1-R3 of sequential coverage analysis we have to decide on (i) the minimum number of bad confidence intervals, N_{Bmin}, which have to be recorded before the sequential analysis of coverage can start, and (ii) the minimum length of simulation runs for producing valid steady-state estimates.

To determine N_{Bmin} we looked at the convergence of coverage to its limiting value as a function of the number of replications. Figures 1 and 2 show the coverage convergence curves obtained for the BM method, having run multiple independent replications of the M/M/1 queueing system loaded at $\rho = 0.7$, for simulations on P = 1 processor, and for P = 2 and 4 processors, respectively. (Note that the curves are drawn to different scales.) The curve in Figure 1a shows the type of convergence of coverage for P = 1 processor when none of the rules R1-R3 is applied. The next three curves (Figures 1b, c and d) were obtained assuming N_{bmin} = 30, 100, and 200. Additionally, to implement R3, when N_{Bmin} "bad" confidence intervals had been recorded, the average length of a simulation run was calculated, and all simulation runs shorter by more than one standard deviation than the average simulation run length were discarded. At the points where these operations were executed, the coverage has improved, as indicated by jumps in the convergence curves in Fig. 1b, c and d. Thus, filtering out simulation runs that are too short removes significant bias in the results.

Next, if the number of bad confidence intervals was not smaller than N_{Bmin}, the coverage was estimated sequentially, taking into account only sufficiently long replications. Otherwise, more "bad" confidence intervals would need to be recorded first. Sequential analysis of coverage was stopped when the relative precision of coverage dropped to within its required level (in our case, the threshold was 5% at 0.95 confidence level).

Comparing the locations of the points at which N_{bmin} = 30, 100, and 200 "bad" confidence intervals were recorded (Figures 1b, c, d) with the curve of Figure 1a, one can clearly see that at the point corresponding to N_{bmin} = 30 the coverage curve has not yet settled down, and that about N_{bmin} = 100 is necessary for reasonable convergence. (On this basis, the preliminary results of sequential coverage analysis for N_{bmin} = 30, published in [4], cannot be regarded as reliable.) This is consistent with other results for BM, obtained from simulations of the M/M/1 queueing system at different load levels. Similar effects can be also observed when running simulations concurrently on multiple processors; see Figures 2a and b, showing coverage convergence curves for P = 2 and 4 processors, respectively. Similar conclusions could also be drawn when studying the results we obtained for the method of Non-overlapping Batch Means proposed by Heidelberger and Welch. Thus, in our studies we used N_{bmin} = 200. The convergence curves in Figures 1b, c and d, as well as in Figure 2, end at the stopping points of the reported studies of coverage.

The results of our coverage analysis of the method of BM and SA/HW, when they are applied in simulations on single processors, are shown in Figures 3 and 4, and Table 1 and 2. The results following the traditional fixed-sample size approach (coverage analysis over 200 replications) are shown in Figures 3a and 4a, and Tables 1a and 2a, while Figures 3b and 4b, and Tables 1b and 2b, show results following our methodology of sequential analysis of coverage, ie. applying the rules R1, R2 and R3. The second and third columns in the tables show the total number of confidence intervals and the number of "bad" confidence intervals used in the analysis of coverage at a given load level. The fourth column in Tables 1b and 2b shows the number of replications whose results were discarded because of insufficient simulation run lengths, i.e., because their lengths were below one standard deviation from the mean simulation length (of the M/M/1 queueing system at a given load level).

It appears that the traditional approach cannot produce reliable estimates of coverage. In these specific examples it underestimated the quality of the results produced by BM and SP/HW.

Figures 5 and 6 depict the results obtained from sequential coverage analysis of BM and SA/HW when the simulations were run concurrently on P = 2 and 4 processors. As these results show, the quality of both BM and SA/HW improves as the number of simulation engines (processors) increases. This is a reasonable consequence of using the MRIP scenario, since by increasing the number of independent simulation engines one introduces more independent subsequences of output data (one subsequence per simulation engine), and the quality of the pooled estimators used in BM and SA/HW improves.

The question of whether these methods remain valid for heavier loaded systems, i.e. for $\rho > 0.9$, as well as for larger number of processors, remains open. Experimental studies of heavily loaded systems require very long simulation runs. On the other hand, the theoretical properties of SA/HW suggest the existence of an upper level of P, above which this method of analysis of simulation output data can become invalid.

4. Conclusions

We have formulated basic rules that should be followed in proper experimental analysis of the coverage of different steady-state confidence interval estimators. Our main argument is that such a meta-analysis should be done sequentially. Coverage results for the methods of Batch Means and Spectral Analysis have been presented and compared with those obtained by the traditional, non-sequential approach. As advocated in [17] and [26], to draw more general conclusions about performance of interval estimators used in various methods of steady-state simulation, we need to consider a number of different simulation models, since the results obtained for one system (in this paper: M/M/1∞) are not sufficient. Additionally, it

is unlikely that a single method of sequential analysis of simulation output data could be universally valid. To encourage cooperation and to intensify research in this area, representative, but analytically tractable, systems in various application areas of simulation (computer systems, telecommunication networks, industrial processes, etc.) should be selected and used as standards in experimental analysis of coverage. Unfortunately, even though such a demand was formulated for the first time in 1980 [26], no progress in reaching such an agreement can be reported.

Acknowledgements

The authors thank an anonymous referee for valuable comments which improved the style of this paper.

References

[1] N. R. Adam, Achieving a Confidence Interval for Parameters Estimated by Simulation, *Management Sci.* **29** (1983) 856-866.

[2] J. M. Charnes and E. I. Chen, Vector-Autoregressive Inference for Equally Spaced Time-Averaged Multiple Queue Length Processes, in Proc. 1994 Winter Simulation Conf. (IEEE, N.Y., 1994) 312-315

[3] B. L. Fox, D. Goldsman and J. Swain, Spaced Batch Means, *Operations Res. Letters* **10** (1991) 255-263

[4] G. Ewing, D. McNickle and K. Pawlikowski, Credibility of the Final Results from Quantitative Stochastic Simulation, in Proc. 1995 EUROSIM Conference (Elsevier Science Publ., 1995) 189-194

[5] P. W. Glynn, Coverage Error for Confidence Intervals Arising in Simulation Output Analysis, in Proc. 1982 Winter Simulation Conf. (IEEE, N.Y., 1982) 369-375.

[6] P. W. Glynn and W. Ward, The Asymptotic Validity of Sequential Stopping Rules for Stochastic Simulations, *Annals of Applied Probability* **2** (1) (1992) 180-198

[7] D. Goldsman and K. Kang, Cramer-von Mises Variance Estimators for Simulations, in Proc. 1991 Winter Simulation Conf. (IEEE Press, 1991) 916-920

[8] D. Goldsman, K. Kang and R. G. Sargent, Large and Small Sample Comparisons of Various Variance Estimators, in Proc. 1986 Winter Simulation Conf. (IEEE Press, 1991) 278-284

[9] D. Goldsman and L. Schruben, New Confidence Interval Estimators Using Standardised Time Series, *Management Sci*., **36** (3) (1990) 393-397

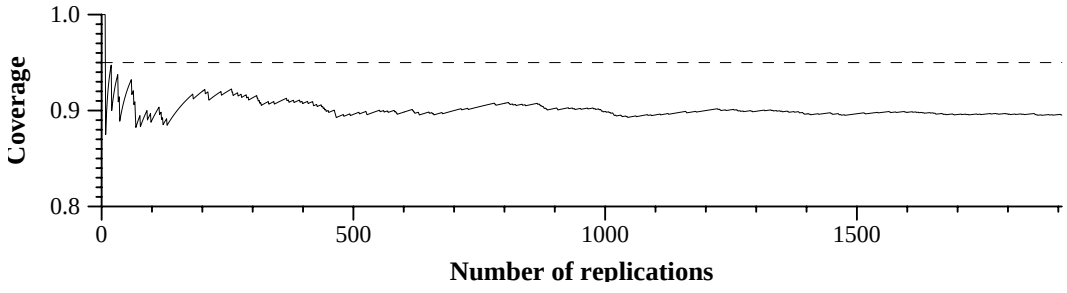
[10] P. Heidelberger and P. D. Welch, A Spectral Method for Confidence Interval Generation and Run Length Control in Simulation, *Comms. of ACM* **25** (1981) 233-245

[11] P. Heidelberger and P. D. Welch, Simulation Run Length Control in the Presence of an Initial Transient, *Operations Res*.. **31** (1983) 1109-1144

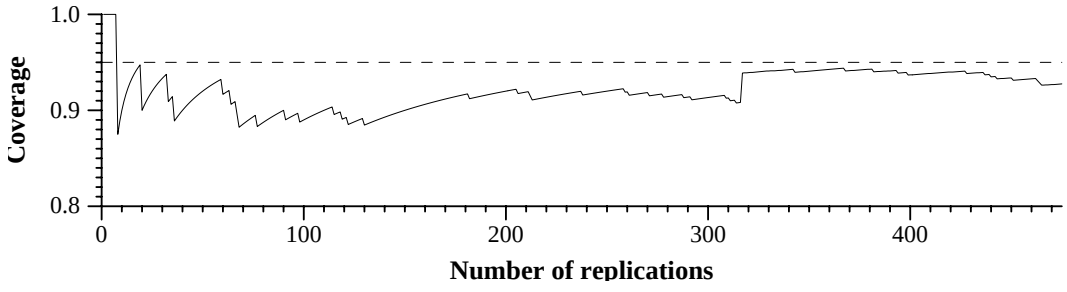
[12] R. B. Howard, M. A. Gallagher, K. W. Bauer and P. S. Maybeck, Confidence Intervals for Univariate Discrete-Event Simulation Output Using the Kalman Filter, in Proc. 1992 Winter Simulation Conf. (IEEE Press, 1992) 586-593

[13] K. Kang and D. Goldsman, The Correlation Between Mean and Variance Estimators in Computer Simulation. *Trans. of IIE* **22** (1) (1990) 15-23

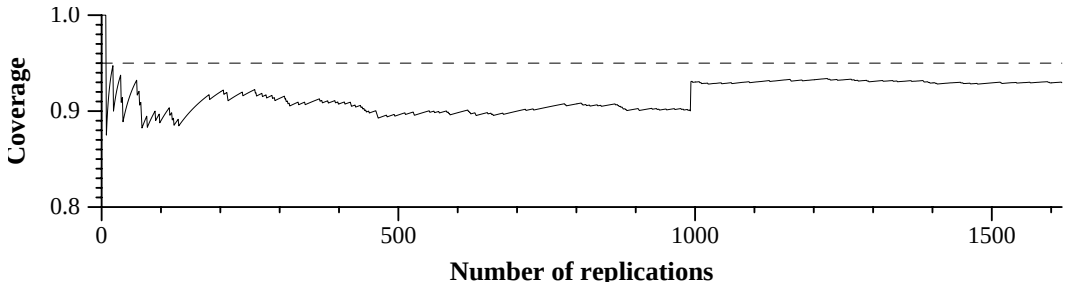
[14] W. D. Kelton and A. M. Law, An Analytical Evaluation of Alternative Strategies in Steady-State Simulation, *Operations Res*.. **32** (1) (1984) 169-184.
[15] S. S. Lavenberg and C. H. Sauer, Sequential Stopping Rules for the Regenerative Method of Simulation, *IBM J. Research and Development* **21** (1977) 667-678
[16] A. M. Law and J. S. Carson, A Sequential Procedure for Determining the Length of a Steady-State Simulation, *Operations Res.* **27** (1979) 1011-1025
[17] A. M. Law and W. D. Kelton, Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures, *Management Sci.* **28** (5) (1982) 550-562
[18] A. M. Law, Statistical Analysis of Simulation Output Data, *Operations Res*. **31** (6) (1983) 983-1029
[19] A. M. Law and W.D.Kelton, *Simulation Modeling and Analysis*. (McGraw-Hill, NY, 1992)
[20] F. Mosteller, R. E. K. Rourke and G.B.Thomas, *Probability with Statistical Applications*. (Addison Wesley, Reading, 1970)
[21] K. Pawlikowski,. Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions, *ACM Computing Surveys* **22** (2) (1990) 123-170
[22] K. Pawlikowski, V. Yau and D. McNickle, Distributed Stochastic Discrete-Event Simulation in Parallel Time Streams, in Proc. 1994 Winter Simulation Conf. (IEEE Press, 1994) 723-730
[23] K. E. E. Raatikainen, Rune Length Control Using Parallel Spectral Method, in Proc. 1992 Winter Simulation Conf. (IEEE Press, 1992) 594-602
[24] V. J. Rego and V. S. Sunderam, Experiments with Concurrent Stochastic Simulation: the EcliPSe Paradigm". *J. Parallel and Distributed Computing* **14** (1992) 66-84
[25] C. H. Sauer and S. S. Lavenberg, Confidence Intervals for Queueing Simulations of Computer Systems, *ACM Performance Evaluation Review* **8** (1-2) 46-55
[26] T. J. Schriber and R. W. Andrews, A Conceptual Framework for Research in the Analysis of Simulation Output, *Comm. of the ACM* **24** (4) (1981) 218-232
[27] L. W. Schruben, A Coverage Function for Interval Estimators of Simulation Response, *Management Sci*.. **26** (1980) 18-27.
[28] A. F. Seila, A Batching Approach to Quantiles Estimation in Regenerative Simulations, *Management Sci*., **28** (5) (1982) 573-581
[29] V. Yau and K. Pawlikowski, AKAROA: a Package for Automatic Generation and Process Control of Parallel Stochastic Simulation, in Proc. 16th Australian Computer Science Conf., *Australian Computer Science Comms.* **15** (1) (1993) 71-82



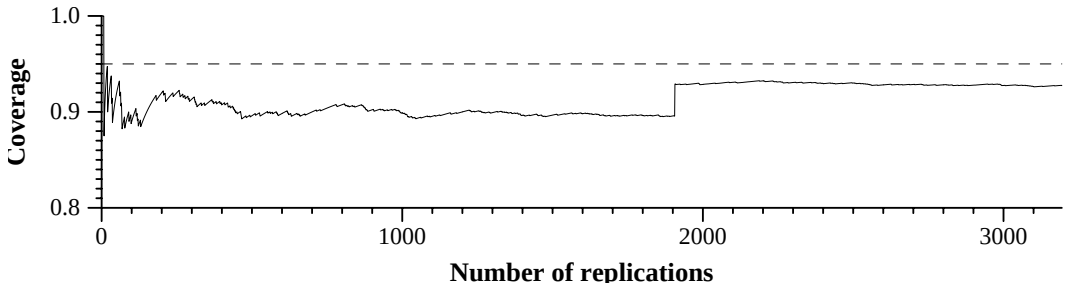
(a)



(b)

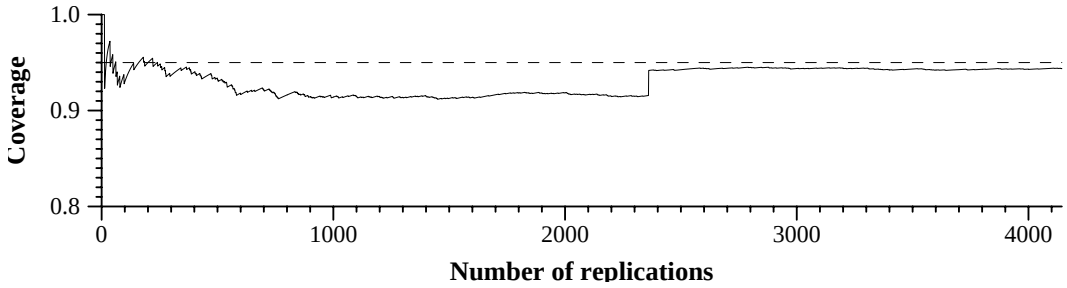


(c)

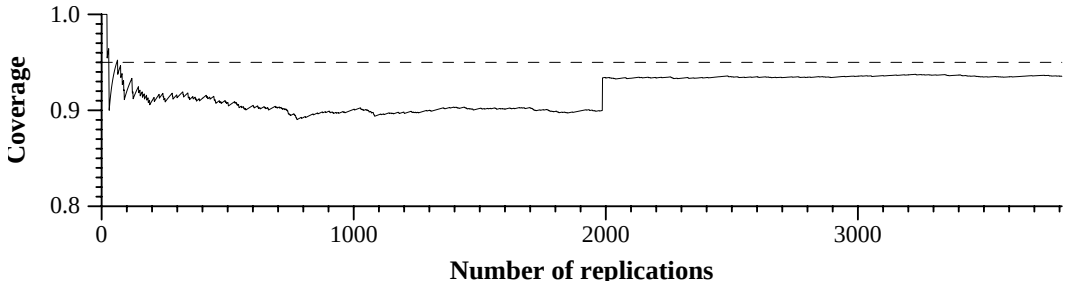


(d)

Figure 1. Coverage as a function of the sample size for BM in steady-state simulation of an M/M/1/∞ queueing system for ρ = 0.7, P = 1 processor. (a) No filtering of output data, (b) N_{Bmin} = 30, (c) N_{Bmin} = 100, (d) N_{Bmin} = 200.



(a)



(b)

Figure 2. Coverage as a function of the sample size for BM in steady-state simulation of an M/M/1/∞ queueing system for ρ = 0.7, N_{Bmin} = 200. (a) P = 1 processor, (b) P = 2 processors.

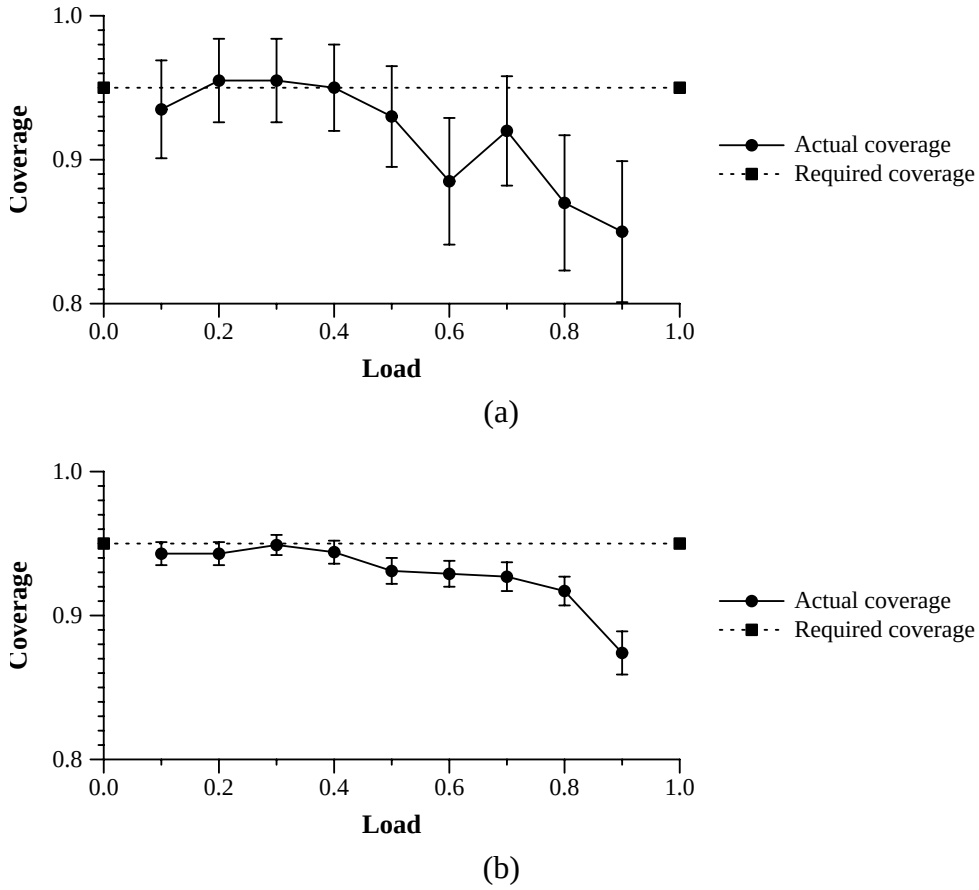


Figure 3. Coverage analysis of BM for P = 1 processor.
(a) Fixed sample size of 200 replications;
(b) sequential analysis for N_{Bmin} = 200.

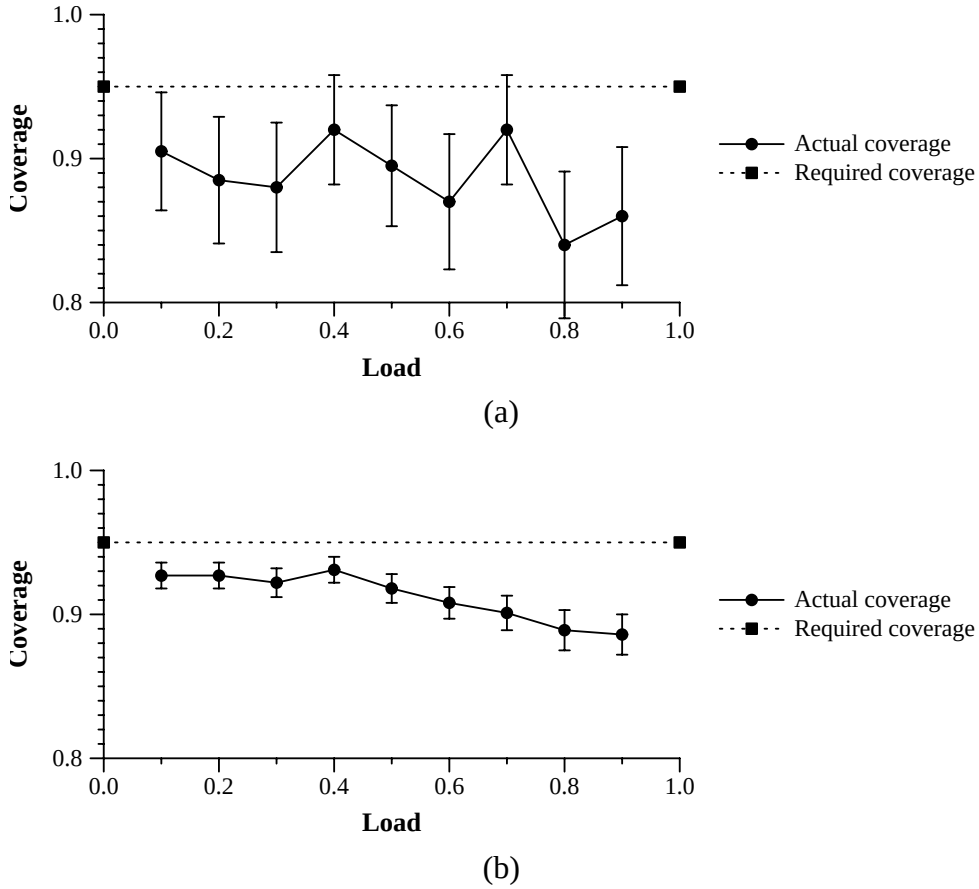


Figure 4. Coverage analysis of SA/HW for $P = 1$ processor.
(a) Fixed sample size of 200 replications;
(b) sequential analysis for $N_{Bmin} = 200$.

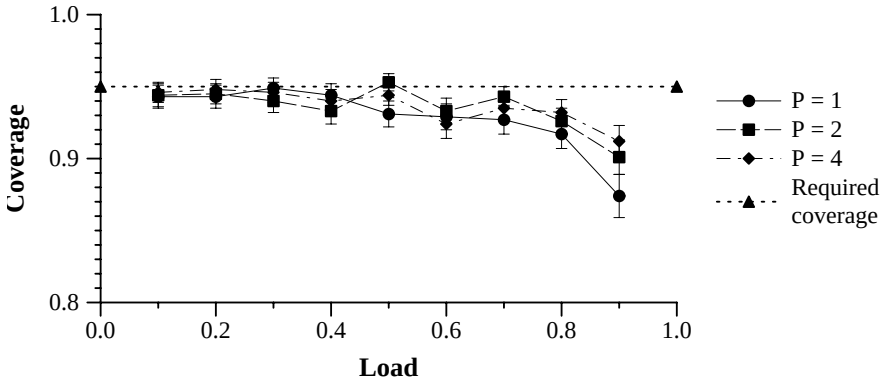


Figure 5. Coverage of BM when simulation is executed on P = 1, 2 and 4 processors.

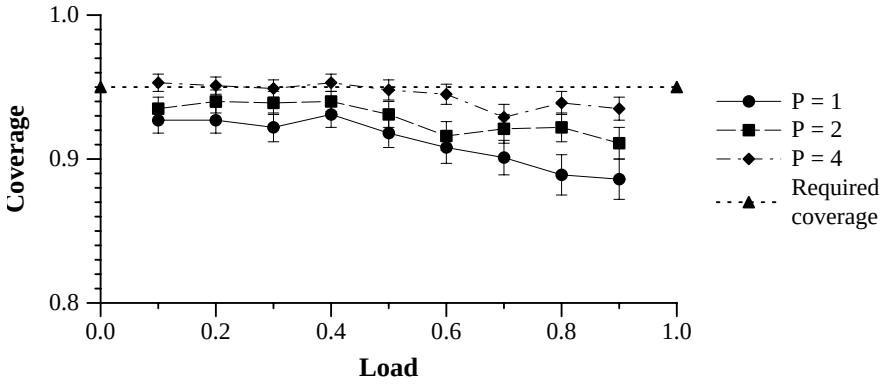


Figure 6. Coverage of SA/HW when simulation is executed on P = 1, 2 and 4 processors.