Interval Estimators of Proportions in Coverage Analysis of Simulation Output Results

Jong-Suk R. Lee, Don McNickle[†] and Krzysztof Pawlikowski
Department of Computer Science and [†]Department of Management
University of Canterbury, New Zealand
E-mail: {{ruth, krys}@cosc, [†]dcm@mang}.canterbury.ac.nz

Abstract

A normal distribution is commonly used for defining confidence interval estimators for proportions even though alternative, more accurate and efficient estimators have been proposed in the past. This is probably because the normal approximation has been easier to use in practice than other estimators. However, current computing technology can now deal with such estimators quite efficiently. One application of proportions is in the coverage analysis of simulation output results. Coverage analysis is often used to measure the robustness of methods of simulation output data analysis. In this paper we study three interval estimators of proportions to find out a more accurate estimator for the applications in coverage analysis. The estimators (based on the normal distribution, the *arcsin* transformation and the F distribution) were compared with exact values, which are calculated by the binomial probability density function. The numerical results show that estimators based on the F distribution should be used in practice.

Keywords: proportions, interval estimators, confidence intervals, coverage analysis

1 Introduction

Statistical analysis of output data of a stochastic steady-state simulation is made difficult by the degree of serial correlation often present in the output. Methods such as batch means, regenerative cycles, and spectral analysis are used to overcome this. An important measure of the robustness of any output analysis method is the coverage of the final confidence intervals (CIs), defined as the proportion of CIs which contain the true value; see Figure 1. Any good method of analysis should produce narrow and stable CIs, and the probability of such an interval containing the true value of the estimated performance measure should be close to the assumed confidence level.

Some interesting results have been achieved in theoretical studies in terms of coverage error for CIs arising in simulation output data analysis (see [4]). A coverage function (which is defined for all confidence levels between zero and one) to measure the robustness of CIs has been proposed [16], and coverage properties of CIs based on the average Bayesian *posterior* probability have been studied [14]. Nevertheless, experimental analysis of coverage is still required to assess the quality of practical implementations of methods used for determining the final CIs, especially in the context of a stochastic steady-state simulation.

The conventional interval estimator of proportions based on the normal approximation has been widely used in experimental coverage analysis (see for example [9], [12], [16]). Alternative, possibly more efficient and accurate interval estimators of proportions are also known [5]. However, they were too slow for practical use until the early 1990's. Recently, one of these estimators (based on the *arcsin* transformation) has been used for the analysis of proportions in a sequential steady-state simulation ([13]), but a comparative study of the properties of these estimators was not undertaken.

For example, Figure 2 shows the typical results of a coverage experiment, where the method of non-overlapping batch means has been used to analyse the mean response time of an $M/M/1/\infty$ queueing model in a sequential steady-state simulation. One can see that the actual coverage of the CIs drops away from the assumed confidence level (95%) as the traffic intensity increases. Choosing the optimal batch size for reducing or eliminating the autocorrelation is not easy, especially in the case of very strong autocorrelations existing between data (observations) collected during the simulation ([2]). In general, the autocorrelations in queueing systems with unlimited buffer capacity rapidly increase as traffic load approaches 1.0, so in such cases one needs to collect a huge number of observations to obtain the final simulation results with small statistical errors. As discussed in [2], in the method of batch means, larger batch sizes are required to obtain less correlated batch means if observations are more correlated. The major reason for poor coverage in heavier traffic intensities may be that, even with a sophisticated sequential algorithm for batch size selection in the batch means method, such as the one described in [11], correlations between batch means in heavily loaded traffic may not be completely eliminated.

There are a number of facts related to the analysis of coverage. Firstly, it is naturally limited to analytically tractable systems only, since the theoretical value of the parameter of interest has to be known. Because of that, it has even been claimed that there is no justification for experimental coverage analysis, since there is no theoretical basis for extrapolating results found for simple, analytically tractable systems to more complex systems, which are the subjects of practical simulation

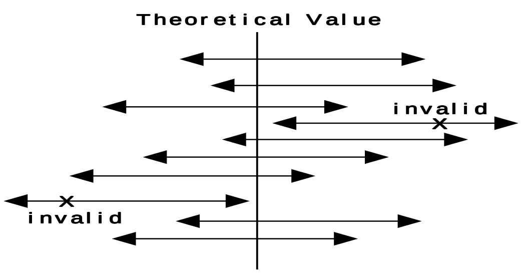

Figure 1: Valid and invalid CIs in coverage analysis. In this example, the coverage is 80% since 8 out of 10 CIs contain the theoretical value

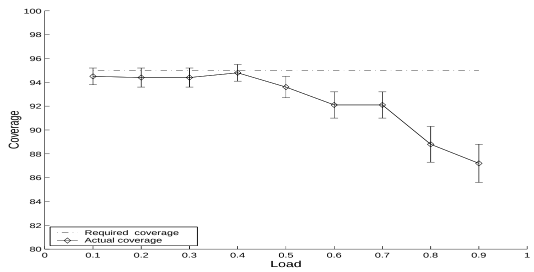

Figure 2: Coverage analysis of the method of non-overlapping batch means (Steady-state simulation of $M/M/1/\infty$ queueing system)

studies [3]. On the other hand, no theory of coverage for finite sample sizes exists, and in this situation, experimental coverage analysis of analytically tractable systems remains the only method available for testing the validity of methods proposed for simulation output analysis. Certainly nobody is ready to accept an analysis method showing very poor quality in experimental studies of coverage.

Secondly, coverage analysis requires the execution of multiple, independent replications of simulations. Very large numbers of replications are often needed to determine coverage with a satisfactory precision. Traditionally, coverage analysis was performed with a fixed number of replications, mostly ranging from 50 - 500 ([7], [8], [9], and [15]). However, newer results of coverage analysis reported ([10] and [12]), clearly show the existence of a high initial instability of coverage in the region of 50 - 500 replications for three different methods of mean value analysis: non-overlapping batch means, spectral analysis, and regenerative cycles; also see Figure 3. To avoid taking the final result from this region, coverage analysis has to be done over a sufficiently large sample of data or, as advised in [12], sequential analysis of coverage is recommended. In any case, one needs to accurately estimate the proportion of CIs covering the theoretical value of interest.

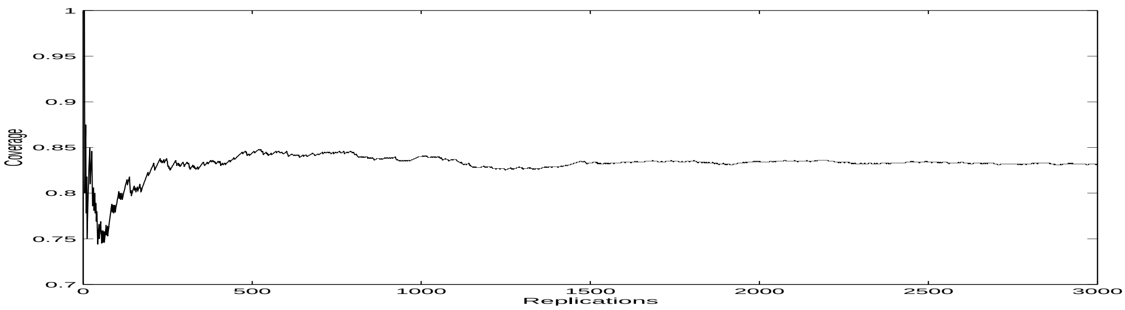

Figure 3: Convergence of coverage of the sequential method of non-overlapping batch means ($M/M/1/\infty$ queueing system at the load $\rho = 0.9$, confidence level of 0.95)

2 Interval Estimators for Proportions

Binomial experiments consist of repeated trials, each with two possible outcomes, which may be labelled *success* or *failure*. The point estimator of the proportion p in a binomial experiment is simply given by the statistic

$$\hat{p} = \frac{count\ of\ successes\ in\ sample}{size\ of\ sample} = \frac{X}{n}. \quad (1)$$

If a binomial experiment can result in a success with probability p and a failure with probability $(1-p)$, then the probability distribution of the binomial random variable X, the number of successes in n independent experiments, is

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \ \ x = 0, 1, \ldots, n. \quad (2)$$

The accuracy with which it estimates an unknown proportion p can be assessed by the width of its CI at a given confidence level, i.e, by the probability

$$Pr(\hat{p} - \Delta_1 \le p \le \hat{p} + \Delta_2) = 1 - \alpha$$

where $\hat{p}$ is the estimate of the proportion p, Δ_1 and Δ_2 are the offset for the lower and the upper limit of the CI of p, and $(1-\alpha)$ is the confidence level, $0 < \alpha < 1$. Ideally, this would mean that if the simulation experiment is repeated many times, the resulting CIs would contain the parameter p in $100(1-\alpha)\%$ of cases.

To determine Δ_1 and Δ_2 we need the exact distribution of $\hat{p}$, or at least to know $Var(\hat{p})$. Calculating exact confidence limit values of p is possible only using Equation (2). However, expanding and inverting the polynomials of order n becomes impractical, even using a computer algebra system, as n increases. The time complexity of the polynomials of order n is $O(p^n)$ [6]. Therefore, approximation methods for a binomial distribution have been suggested. Three interval estimators of the proportion p, based on the normal distribution, the *arcsin* transformation, and the F distribution are described in the following subsections.

2.1 Interval Estimator Based on the Normal Distribution

Finding a CI for the binomial parameter p, $0 \leq p \leq 1$, using an interval estimator based on the normal distribution, is based on the approximation of the binomial distribution of p by areas under the normal distribution, with mean $\hat{p}$ and variance $\hat{p}(1-\hat{p})/n$ [5].

For large n, the random variable

$$Z = \frac{p - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

is approximately standard normal; see for example [17]. Thus, an approximate CI for the proportion p is

$$Pr(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha,$$

where $\hat{p}$ is the sample proportion, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, and n is the sample size. Note, that this is a symmetric CI.

The accuracy of the normal approximation improves as the sample size n increases. However, it is most accurate when p is close to $1/2$, and becomes quite inaccurate when p is near 0 or 1. Much of this inaccuracy is then due to the skewed nature of the binomial distribution. This is exactly the situation in simulation coverage analysis, where typically p is between 0.9 and 0.99. Thus, we need an interval estimator for coverage analysis which can produce an asymmetric CI.

2.2 Interval Estimator Based on *Arcsin* Transformation

An asymmetric CI for proportions based on the *arcsin* transformation was originally proposed by Fisher (see [5]). Having defined a new random variable $\hat{Y} = 2\, arcsin\sqrt{\hat{p}}$, we can construct an approximate 100(1-α)% CI for a proportion, $(\hat{p}_l, \hat{p}_u)$, where $\hat{p}_l = sin(l/2)^2$ and $\hat{p}_u = sin(u/2)^2$. Here

$$l = arcsin\sqrt{\hat{p} - 1/(2n)} - z_{1-\alpha/2}/\sqrt{n} \quad \text{and} \quad u = arcsin\sqrt{\hat{p} + 1/(2n)} + z_{1-\alpha/2}/\sqrt{n},$$

where $\hat{p}$ is the sample proportion, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, and n is the sample size ([5] and [13]). This is based on the fact that if the sample proportion is approximately normally distributed with mean

p and variance $p(1-p)/n$, then $\hat{Y} = 2\,arcsin\sqrt{\hat{p}}$ is also approximately normally distributed with mean $\mu_{\hat{p}} = 2\,arcsin\sqrt{p}$ and variance $\sigma^2_{\hat{p}} = 1/n$; see [5] for more details.

2.3 Interval Estimator Based on the F Distribution

CIs for proportions can also be formulated from the relationship between the F distribution and the binomial distribution with the incomplete and complete beta functions. The ratio of two successive terms in a binomial distribution $b(x;n,p)$ of Equation (2) is

$$\frac{b(x+1;n,p)}{b(x;n,p)} = (\frac{n-x}{x+1})(\frac{p}{1-p}), \quad x = 0, 1, \ldots, n-1,$$

where x is the observed number of successes in the sample; see Equation (1). Using the transformations shown, for example, in [1] and [5], the quantiles of the binomial distribution can be obtained from those of the F distribution, as

$$Pr\{F(r_1,r_2) < (\frac{n-n\hat{p}}{n\hat{p}+1})(\frac{p}{1-p})\} = Pr\{\frac{(n\hat{p}+1)F(r_1,r_2)}{(n-n\hat{p})+(n\hat{p}+1)F(r_1,r_2)} < p\},$$

where $F(r_1,r_2)$ is a random variable with the F distribution of $r_1 = 2*(n\hat{p}+1)$ and $r_2 = 2*(n-n\hat{p})$ degrees of freedom. Thus, an 100(1-α)% CI for a proportion is given by $(\hat{p}_l, \hat{p}_u)$, where

$$\hat{p}_u = \frac{(n\hat{p}+1)f_{1-\alpha/2}(r_1,r_2)}{(n-n\hat{p})+(n\hat{p}+1)f_{1-\alpha/2}(r_1,r_2)} \quad \text{and} \quad \hat{p}_l = \frac{n\hat{p}}{n\hat{p}+(n-n\hat{p}+1)f_{1-\alpha/2}(r_3,r_4)}.$$

Here, n is the sample size, and $f_{1-\alpha/2}(r_1,r_2)$ is the $(1-\alpha/2)$ quantile of the F distribution with (r_1,r_2) degrees of freedom, where $r_1 = 2*(n\hat{p}+1)$ and $r_2 = 2*(n-n\hat{p})$, while $f_{1-\alpha/2}(r_3,r_4)$ is the $(1-\alpha/2)$ quantile of the F distribution with (r_3,r_4) degrees of freedom, where $r_3 = 2*(n-n\hat{p}+1)$ and $r_4 = 2*n\hat{p}$ [5].

3 Comparisons of Three Interval Estimators with Exact Values

To more closely investigate the three interval estimators, CIs of proportions based on the normal, the *arcsin*, and the F distribution[1] at the given confidence level (1 - α = 0.99) and sample size[2] $n = 20$ are depicted in Figure 4. The upper limits of CIs of proportions from 0.025 to 0.975 using each interval estimator and the 'exact' upper limits of the CIs of proportions, which are calculated by the binomial probability density function, are in Table 1. The relative accuracy of the upper confidence limits of the three interval estimators when compared to the 'exact' values of the upper confidence limit is also presented in parentheses. For higher proportions, the interval estimator based on the F distribution produces the closest values to the exact values, while the interval estimator based on the normal distribution produces values exceeding the upper limit of 1.0.

[1]The numerical values for the F distribution were obtained from a carefully validated implementation of the method proposed in [1].

[2]We have chosen a small sample size of twenty for visibility and to obtain the exact values of proportions from the binomial distribution (as discussed before, it is impossible to calculate them at a large sample size). Similar results obtained from larger sample sizes will be presented later.

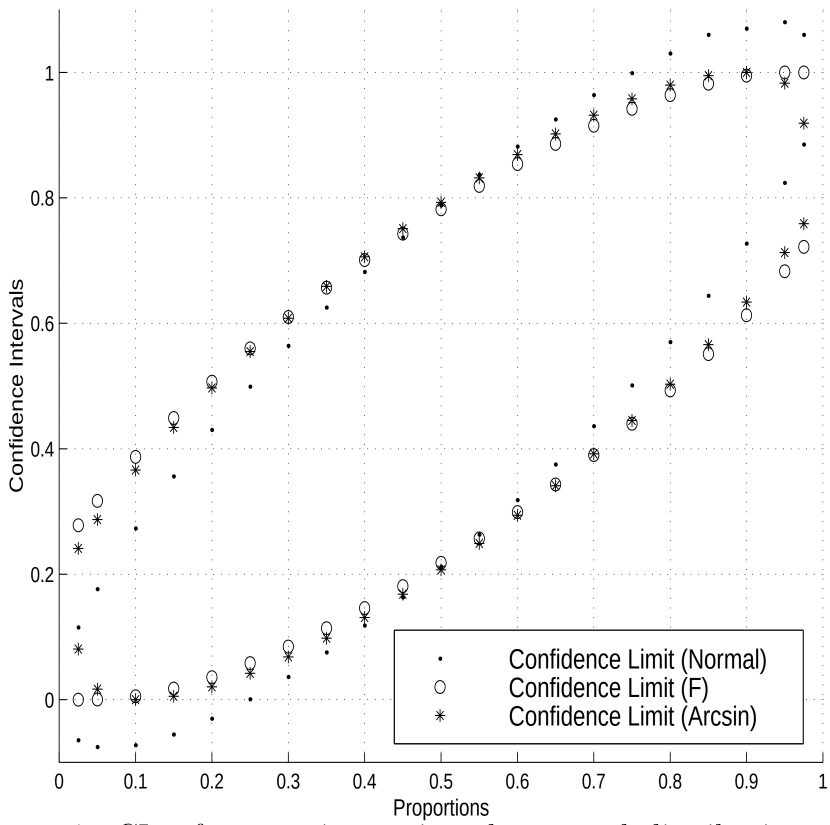

Figure 4: CIs of proportions using the normal distribution, the *arcsin* transformation, and the F distribution ($\alpha = 0.01$ & $n = 20$)

To determine whether the interval estimator based on the normal distribution produces invalid CIs when the sample size is increased, we tested larger sample sizes n (ranging between 10,000 and 1,000,000) and an α of 0.05 - 0.001. The results are presented in Table 2. The invalid CI regions (CI < 0.0 or CI > 1.0) have shrunk as the sample sizes n increased, but even taking the very large sample size of one million, the invalid CI regions of proportions still exist.

Figure 4 and Table 2 confirm the theoretical claims that interval estimators of proportions based on the *arcsin* transformation and the F distribution never exceed the practical lower and upper limits of the CIs. However, it can be seen that the lower and upper limits of the interval estimator of proportions based on the normal distribution can exceed the lower limit of 0.0 and the upper limit of 1.0, making it inappropriate for coverage analysis of sequential interval estimators. The results also show that the interval estimator based on the F distribution produces more accurate CIs than the other interval estimators.

4 Conclusions

While some interesting results have been achieved in theoretical studies of coverage analysis of interval estimators used in simulation output data analysis, experimental analysis of coverage based on accurate estimators of proportions is still required to assess the quality of the practical implementations of the methods used to determine CIs in stochastic simulation. In this paper we have studied three interval estimators of proportions to find out a more accurate estimator for the applications

Table 1: Upper limits of CIs of proportions and their relative inaccuracy ($\alpha = 0.01$ & $n = 20$)

Proportions	*Exact Values*	*Normal Distribution*	*Arcsin Transformation*	*F Distribution*
0.05	0.289	0.176 (- 39.1 %)	0.287 (- 0.7 %)	0.317 (+ 9.7 %)
0.1	0.358	0.273 (- 23.7 %)	0.366 (+ 2.2 %)	0.387 (+ 8.1 %)
0.15	0.421	0.356 (- 15.4 %)	0.434 (+ 3.1 %)	0.449 (+ 6.7 %)
0.2	0.478	0.430 (- 10.0 %)	0.497 (+ 4.0 %)	0.507 (+ 6.1 %)
0.25	0.532	0.499 (- 6.2 %)	0.555 (+ 4.3 %)	0.560 (+ 5.3 %)
0.3	0.583	0.564 (- 3.3 %)	0.608 (+ 4.3 %)	0.610 (+ 4.6 %)
0.35	0.631	0.625 (- 1.0 %)	0.659 (+ 4.4 %)	0.657 (+ 4.1 %)
0.4	0.677	0.682 (+ 0.7 %)	0.706 (+ 4.3 %)	0.701 (+ 3.5 %)
0.45	0.720	0.737 (+ 2.4 %)	0.751 (+ 4.3 %)	0.743 (+ 3.2 %)
0.5	0.761	0.788 (+ 3.5 %)	0.793 (+ 4.2 %)	0.782 (+ 2.8 %)
0.55	0.800	0.837 (+ 4.6 %)	0.832 (+ 4.0 %)	0.819 (+ 2.4 %)
0.6	0.837	0.882 (+ 5.4 %)	0.869 (+ 3.8 %)	0.854 (+ 2.0 %)
0.65	0.871	0.925 (+ 6.2 %)	0.902 (+ 3.6 %)	0.886 (+ 1.7 %)
0.7	0.902	0.964 (+ 6.9 %)	0.932 (+ 3.3 %)	0.915 (+ 1.4 %)
0.75	0.931	0.999 (+ 7.3 %)	0.958 (+ 2.9 %)	0.942 (+ 1.2 %)
0.8	0.956	1.03 (+ 7.7 %)	0.980 (+ 2.5 %)	0.964 (+ 0.8 %)
0.85	0.977	1.06 (+ 8.5 %)	0.995 (+ 1.8 %)	0.982 (+ 0.5 %)
0.9	0.992	1.07 (+ 7.9 %)	1.0 (+ 0.8 %)	0.995 (+ 0.3 %)
0.95	0.999	1.08 (+ 8.1 %)	0.983 (- 1.6 %)	1.0 (+ 0.1 %)

Table 2: Invalid CIs of proportions using the normal distribution

	Sample Size = 10000		*Sample Size = 100000*		*Sample Size = 1000000*	
	Proportions	*Upper CIs*	*Proportions*	*Upper CIs*	*Proportions*	*Upper CIs*
α = *0.05*	0.99997	1.000004	0.99997	1.000004	0.999997	1.0000004
	0.99998	1.0000077	0.99998	1.0000077	0.999998	1.0000008
	0.99999	1.0000096	0.99999	1.0000096	0.999999	1.000001
α = *0.01*	0.9994	1.0000309	0.99994	1.0000031	0.999994	1.0000003
	0.9995	1.0000759	0.99995	1.0000076	0.999995	1.0000008
	0.9996	1.0001151	0.99996	1.0000115	0.999996	1.0000012
	0.9997	1.0001461	0.99997	1.0000146	0.999997	1.0000015
	0.9998	1.0001643	0.99998	1.0000164	0.999998	1.0000016
	0.9999	1.0001576	0.99999	1.0000158	0.999999	1.0000016
α = *0.001*	0.999	1.0000401	0.9999	1.0000041	0.99999	1.0000004
	0.9991	1.0000868	0.99991	1.0000087	0.999991	1.0000009
	0.9992	1.0001304	0.99992	1.0000131	0.999992	1.0000013
	0.9993	1.0001703	0.99993	1.0000171	0.999993	1.0000017
	0.9994	1.0002058	0.99994	1.0000206	0.999994	1.0000021
	0.9995	1.0002357	0.99995	1.0000236	0.999995	1.0000024
	0.9996	1.000258	0.99996	1.0000258	0.999996	1.0000026
	0.9997	1.0002699	0.99997	1.000027	0.999997	1.0000027
	0.9998	1.0002653	0.99998	1.0000265	0.999998	1.0000027
	0.9999	1.0002291	0.99999	1.0000229	0.999999	1.0000023

in coverage analysis. These estimators (based on the normal distribution, the *arcsin* transformation and the F distribution) were compared with exact values, which are calculated by the binomial probability density function.

CI estimators for proportions using the (symmetric) normal approximation have been commonly used for coverage analysis of simulation output even though alternative estimators of (asymmetric) CIs for proportions have been proposed in the past. This is probably because the normal approximation had been easier to use in practice than the other estimators. However, current computing technology can now deal with alternative estimators. Even CIs for coverage analysis based on the F distribution can be calculated easily by a standard computer. Therefore, being concerned about their validity, we would point to the estimator based on the F distribution as a more accurate and appropriate one in coverage studies, especially if a higher value of confidence level is assumed.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Government Printing Office. New York: Dover Publications, Inc., 1964.

[2] C. Chien. Batch Size Selection for the Batch Means Method. In *Proceedings of the 1994 Winter Simulation Conference*, pages 345–352, Lake Buena Vista, Florida, 1994. J. D. Tew, S. Manivannan, D. A. Sadowski and A. F. Seila (eds.).

[3] B. Fox. Estimation and Simulation. *Management Science*, 24:860–861, 1978.

[4] P. W. Glynn. Coverage Error for Confidence Intervals Arising in Simulation Output Analysis. In *Proceedings of the 1982 Winter Simulation Conference*, pages 369–375. Highland, Chao, and Madrigal (eds.), 1982.

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

[6] E. Horowitz and S. Sahni. *Fundamentals of Data Structures in PASCAL*. Computer Science Press, 1984.

[7] S. S. Lavenberg and C. H. Sauer. Sequential Stopping Rules for the Regenerative Method of Simulation. *IBM Journal of Research Development*, pages 545–558, 1977.

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

[9] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd Ed., 1991.

[10] J. R. Lee, K. Pawlikowski, and D. McNickle. Do Not Trust Too Short Sequential Simulation. In *Proceedings of the 1999 Summer Computer Simulation Conference*, pages 97–102, Chicago, Illinois, 1999.

[11] K. Pawlikowski. Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions. *ACM Computing Surveys*, 22(2):122–170, 1990.

[12] K. Pawlikowski, D. C. McNickle, and G. Ewing. Coverage of Confidence Intervals in Sequential Steady-State Simulation. *Simulation Practice and Theory*, 6:255–267, 1998.

[13] K. E. E. Raatikainen. Simulation-Based Estimation of Proportions. *Management Science*, 41(7):1202–1223, 1995.

[14] D. B. Rubin and N. Schenker. Efficiently Simulating the Coverage Properties of Interval Estimates. *Applied Statistics*, 35(2):159–167, 1986.

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

[16] L. W. Schruben. A Coverage Function for Interval Estimators of Simulation Response. *Management Science*, 26:18–27, 1980.

[17] R. E. Walpole and R. H. Myers. *Probability and Statistics for Engineers and Scientists*. Macmillan Publishing Co., Inc., New York, 2nd Ed., 1978.