Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

SINGLE SIMULATION CONFIDENCE INTERVALS

Christoph Roser
Masaru Nakano

Software Science Laboratory
Toyota Central Research and Development Laboratories
Nagakute, Aichi, 480-1192, JAPAN
croser@robotics.tytlabs.co.jp
nakano@robotics.tytlabs.co.jp

ABSTRACT

This paper presents a method to determine the confidence intervals of many simulation performance measures based on a single simulation. The confidence intervals of independent variables can be calculated directly. The confidence interval of performance functions of means can be calculated using the delta method, as for example for utilizations, frequencies, and throughputs. This allows the measurement of the accuracy of the utilizations, frequencies and throughputs using only a single simulation by using the variation of the underlying means of the performance function. The presented method is highly accurate, fast, and easy to apply. In addition, the method can also be used for short simulations or rare event applications, where methods based on batch means fail. Furthermore, this method can easily be implemented into existing simulation software.

1 INTRODUCTION

Discrete event simulation is a powerful tool to predict the behaviour of complex systems. However, the accuracy of the results depends on many factors, as for example the level of modelling detail of the simulation, the number and magnitude of random effects, and the simulation length. This paper concentrates on the difference between the average results of a finite simulation and the true average results of a theoretical infinite simulation.

While for simple queuing systems the true results can be predicted theoretically, this is very difficult for more complex simulations, and practically every simulation has some small errors between the simulation results and the true average results. Confidence intervals are used to measure the accuracy of the possible true mean values of randomly data. The standard equations for the calculation of confidence intervals, however, require independent and identically distributed data (i.i.d). While simulation data is frequently not independent, there are many performance measures in simulation that are indeed independent. This paper shows, that the results of simulations are frequently independent and, if so, the standard deviation and the confidence interval can be calculated using standard equations.

Additionally, many performance measures in discrete event simulation are a function of one or more means. For example, the throughput is the inverse of the mean time between completions of two parts, or the utilization is the mean work time divided by the mean time between the completions of two parts. While this allows the calculation of the mean throughput or utilization, it gives only one mean value, which is insufficient to calculate a variance, which in turn is necessary to calculate a confidence interval.

This paper applies the delta method to determine the variance of the function of the means of one or more variable (Henderson 2000), (Oliveira, Santana, and Lopes 1997). This method is a valuable alternative to other currently existing methods to estimate the variance and confidence intervals of simulations and other non-independent data, as for example batching (Alexopoulos, and Seila 2000), (Law, and Kelton 1991), (Banks 1998).

2 INDEPENDENCE OF DATA

A single simulation can produce large amounts of data. Some of this data is heavily dependent, i.e. the value at one simulation step depends on the value of the previous simulation step. A prime example are queues, where a long waiting time for one part is most likely followed by a long waiting time of the next part. Currently, dependent data cannot be analysed directly but requires the construction of independent batch means in order to estimate the variance and the confidence intervals. However, other simulation data is quite independent. For example, the time to process one part at one machine is in most simulations described by an independent random variable, hence the resulting data is also independent. The variance of independent data can easily be analysed using the standard equation as shown in equation (1), where σ is the standard deviation if a data set x with n elements and a mean of $\bar{x}$.

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n_i}(x_i - \bar{x})^2}{n-1}} \tag{1}$$

(Neumann 1941; Neumann 1942) developed a measure to determine if sequential data is dependent or independent, known as the von Neumann ratio or the ratio of the mean squared successive difference to the variation η (RMSSDV). Equation (2) shows the calculation of the RMSSDV η based on a set of data x, where the mean squared difference between successive data is divided by the variance of the data.

$$\eta = \frac{n}{(n-1)} \cdot \frac{\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2}{\sum_{i=1}^{n_i}(x_i - \bar{x})^2} \tag{2}$$

If the data x is independent and normally distributed, then the RMSSDV will also be normally distributed with a mean of two and a standard deviation of $4(n\text{-}2)/(n^2\text{-}1)$. Unfortunately, in simulations the data is usually not normally distributed. However, in this case independent data still has a mean value of two. Thus this method can be used to determine if the collected data is independent (i.e. with a mean value at or near two) or not (i.e. the mean differs from two). Variants of equation (2) can be found in (Kleijnen 1987) or (Steiger, and Wilson 1999). Two examples will be used to show the frequent occurrence of independent variables in simulation systems. The first example is a simple one machine queuing system. The second example is a complex simulation involving seven machines and two different part types.

2.1 Queuing system example

The queuing system consists of only one machine, which is at any given time in either of two possible states, idle and working. In the simulation, the idle time and the working time are randomly distributed. As expected, queue related performance measures as for example the waiting time or the queue length were heavily dependent. However, performance measures related to the machine performance were very independent. Table 1 shows an example of a one machine queuing system with a utilization of 80%. The RMSSDV has been determined for both the duration and the time between the beginning of a duration for the working and idle times. The working times are very independent, as is the time between the beginnings of the idle periods. Only the idle periods itself are slightly dependent, but fortunately, this measure is rarely needed in practice.

Table 1: RMSSDV of one machine queuing system

Measure	Number of Events	RMSSDV	
		Duration	Time between Occurrences
Working	79972	2.01	1.96
Idle	16036	2.57	1.98

2.2 Complex production system example

The presented method was also verified using a complex simulation example, consisting of seven machines in a complex setting and a mixture of two different products. The simulation was performed using the GAROPS simulation software as shown in Figure 1 (Kubota, Sato, and Nakano 1999), (Nakano et al. 1994).

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

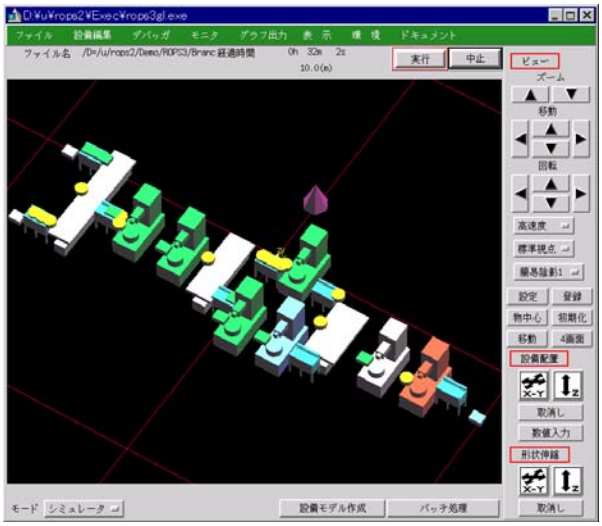

Figure 1: GAROPS simulation example

The independence of the resulting simulation data was measured using the von Neumann ratio as shown in equation (2). As expected, measures related to the queue performance were heavily dependent. However, despite the complex interactions of the system, most machine performance measures were independent. In fact, out of 46 measured parameters as for example the working times or the time between failures, all but four were approximately independent with a RMSSDV between 1.7 and 2.2 as shown in Table 2. This allows the calculation of a valid standard deviation and a confidence interval for these values as described in more detail below.

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

Table 2: RMSSDV of complex simulation

Measure	Number of Events	RMSSDV	
		Duration	Time between Occurrences
M1 Working	49049	2.0	2.0
M1 Blocked	49050	2.0	2.0
M2 Working	49049	2.0	2.0
M2 Blocked	14261	1.8	2.0
M2 Repair	1043	2.0	2.0
M3 Working	16349	2.0	1.8
M3 Idle	6151	1.7	1.8
M3 Blocked	319	6.1	2.1
M3 Repair	1196	2.1	2.0
M4 Working	16349	2.0	1.8
M4 Idle	16061	1.8	1.8
M4 Blocked	8	Insufficient Data	
M4 Repair	494	4.6	2.1
M5 Working	3037	1.7	1.7
M5 Idle	50	2332.1	2.1
M5 Blocked	1721	3.7	2.1
M5 Repair	1291	2.1	2.0
M6 Working	49046	2.0	1.9
M6 Idle	11934	1.8	2.0
M6 Blocked	48205	2.0	1.9
M6 Repair	893	1.9	2.1
M7 Working	49046	2.0	1.9
M7 Idle	12755	1.7	1.7
M7 Repair	1172	1.9	2.0

3 DELTA METHOD CONFIDENCE INTERVALS

The variance and the confidence interval of simulation data can be easily constructed if numerous independent data values are available as shown in equation (1). Unfortunately, other performance measures can be measured only once per simulation, as for example the throughput q (i.e. the number of parts produced divided by the total time) or the utilization u (i.e. the total working time divided by the total time). Subsequently, no variance can be determined and no confidence interval can be constructed.

However, it is possible to express the throughput q or the utilization u as a function of random variables that can be measured repeatedly in the simulation. For example the throughput q is the inverse of the mean time between parts, and the utilization u is the mean working time divided by the mean time between parts as shown in equation (3).

$$q = \frac{Number\,of\ Parts}{TotalTime} = \frac{1}{E(Time\,between Parts)}$$
$$u = \frac{Total\,Working\,Time}{TotalTime} = \frac{E(Working\,Time)}{E(Time\,between Parts)} \quad (3)$$

This has the benefit that the variation of the working time and the variation of the time between parts can be used to calculate the variation of the throughput q and the utilization u using the delta method (Rinne 1997). The delta method determines the variance of a general performance function of one or more mean values based on the gradient of the performance function using a Taylor series

expansion. Assume there is a general performance measure z as a function f of the mean value of one variables μ_x as shown in equation (4).

$$z = f(\mu_x) \tag{4}$$

Yet, if the mean values μ_x are applied to the function f, only one performance measure z is generated. The variation of the performance measure y and subsequently the confidence interval is yet unknown. While it is possible to enter the individual values x_i into the equation f, the resulting mean and variation of the performance measure y would be incorrect for all non-linear functions, i.e. the function of the mean would differ from the expected value of the function of the individual data values. Only if the function f is linear will the function of the means and the mean of the function be equal (Papoulis 1991).

This leads naturally to the idea to replace the function f by its tangent f^* at the mean value μ_x. Using this tangent f^*, it is possible to determine the standard σ_z of the function f of the mean μ_x based on the standard deviation σ_x. Figure 2 visualizes the throughput example for a tangential line f^* replacing the function f.

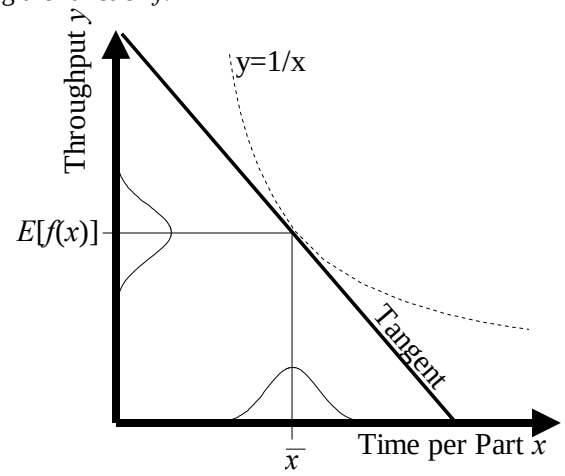

Figure 2: Function replaced by Tangent at the Mean Value

The delta method calculates the deviation σ_z of the function $f_z=(\mu_x, \mu_y)$ of one or more mean values μ_x and μ_y. Equation (5) shows the use of the Delta method for a function of two variables with respect to the covariance between the two variables x and y (Papoulis 1991). This approach is used for example by (Freedman 2001; Moore, and Sa 1999). The equation can be simplified if needed for example if there is no covariance or if only one variable is used.

$$\sigma_z^2 = \left(\left[\frac{df}{dx} \right]_{\substack{x=\mu_x \\ y=\mu_y}} \cdot \sigma \right)^2 + \left(\left[\frac{df}{dy} \right]_{\substack{x=\mu_x \\ y=\mu_y}} \cdot \sigma_y \right)^2 + 2 \cdot \left[\frac{df}{dx} \right]_{\substack{x=\mu_x \\ y=\mu_y}} \cdot \left[\frac{df}{dy} \right]_{\substack{x=\mu_x \\ y=\mu_y}} \cdot Cov[x,y] \tag{5}$$

The unbiased estimate of the covariance $Cov[x,y]$ between two paired variables x and y can be measured as shown in Equation (6).

$$Cov[x,y] = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n-1} \tag{6}$$

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

The resulting standard deviation σ_z can now be used to calculate a confidence interval as shown in Equation (7) using the student t distribution from William Gosset (Student 1908). The confidence interval CI_Z can be calculated from a set of data of size n with a variation σ_z for a confidence level of 1-α/2.

$$CI_z = t_{n-1,\alpha/2} \cdot \frac{\sigma_z}{\sqrt{n}} \tag{7}$$

The validation of standard deviations of functions of means is difficult, as no exact reference method exists to which the results could be compared to. Therefore, this paper validates the confidence interval of the performance functions of the mean. Per definition, out of an infinite number of confidence intervals, the percentage of the confidence intervals containing the true value equals the confidence level (Devore 1995). Subsequently, testing the validity of a confidence interval method requires comparing a large number of confidence intervals with the true value. The percentage of the confidence intervals containing the true value has to be approximately equal to the confidence level 1-α/2 of the confidence intervals. This validation approach is applied to the two examples described above.

3.1 Queuing system example

The queuing system consists of only one machine, as described above. Simple queuing systems have the advantages that the true value of the system performances are known based on the random distributions of the system parameters. To test the method, the validity of the confidence intervals of these queuing systems were tested by calculating 1000 confidence intervals for each setting and comparing the confidence intervals with the true value.

The method has been thoroughly verified for a wide range of possible settings. One varied factor was the confidence level, were the method was tested for frequently used confidence levels of 90%, 95% and 99%, but also for less common confidence levels of 30%, 50% and 75%. The random distributions of the idle time and the working time included exponential distributions, two additional different types of lognormal distributions, and a Weibull distribution. The system was tested for different machine utilizations of 10%, 30%, 50%, 70%, and 95%. Various sample sizes tested the dependence of the algorithm on the sample size. In addition, the system was tested in different identical, but scaled versions, where the scaling factors were 10, 100, and 1000. Finally, the relation between the idle time and the working time was set to be either uncorrelated (idle time independent of working time), positively correlated (long working time causes long idle time and vice versa), or negatively correlated (long working time causes short idle time and vice versa). The true values of the simulation can easily be determined for uncorrelated data using queuing theory. For correlated data, the true value was approximated using very large samples.

All of the above possible system settings have been tested, including a large number of combinations. Altogether about 600 different experiments have been tested, each including about 1000 confidence intervals, creating a total of 600,000 simulations of queuing systems performed for validation purposes. The overall results are shown in Table 3 for the frequencies confidence interval and in Table 4 for the percentage confidence interval. The shown actual coverage is the percentage of confidence intervals containing the true value, where each row of the tables is the result of 100,000 confidence intervals. As can be seen, the actual coverage is always very close to the desired coverage, indicating a very good prediction of the presented method.

Table 3: Accuracy of the frequency confidence interval

Desired Coverage	Actual Coverage	Coverage Deviation	Too Small	Too Large
99%	98.4%	0.9%	1.0%	0.7%
95%	94.1%	1.4%	3.3%	2.6%
90%	88.6%	2.1%	6.4%	5.0%
75%	73.3%	2.7%	14.5%	12.2%
50%	48.9%	1.9%	27.5%	23.6%
30%	29.2%	1.4%	37.5%	33.3%

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

Table 4: Accuracy of the percentage confidence interval

Desired Coverage	Actual Coverage	Coverage Deviation	Too Small	Too Large
99%	98.3%	0.8%	1.0%	0.7%
95%	93.9%	1.5%	3.2%	3.0%
90%	88.7%	1.8%	5.8%	5.5%
75%	73.5%	2.6%	13.2%	13.3%
50%	48.2%	2.1%	25.6%	26.2%
30%	29.1%	1.8%	34.8%	36.0%

The tables also shows the percentage of the simulations with the true value above the confidence interval is approximately equal to the number of cases with the true value below the confidence interval. This indicates a symmetric behaviour and a good fit of the confidence interval. Subsequently, the presented method produces valid and highly accurate confidence intervals for the queuing system.

3.2 Complex manufacturing system example

The presented method was also verified using the complex simulation example as described above. The total simulation time of almost two years was split into 101 subsets with a simulation time of 6 days each. For each subset, the frequencies and the percentages of all machines working, idle, blocked or repaired were measured and the 95% confidence intervals calculated if the underlying data was independent. In total, for all machines, 6219 individual confidence intervals were calculated for frequencies and also for percentages.

These confidence intervals were then compared to the overall averages, which are very close to the unknown true value. Ideally, for confidence intervals with a confidence level of 95%, 95% of the confidence intervals contain the true value, i.e. the desired coverage is 95%. The closer the actual coverage is to the desired coverage, the more accurate is the confidence interval method. Table 5 shows an overview of the coverage results of the complex simulation.

Table 5: Coverage of the simulation example

Performance Measure	Desired Coverage	Actual Coverage	Too Small	Too Large
Frequency	95%	94.44%	2.86%	2.70%
Percentage	95%	92.86%	4.33%	2.81%

Out of the 6219 confidence intervals of frequencies with a desired coverage of 95%, the actual coverage was 94.44%. The instances where the long-term average was outside of the confidence interval were also symmetrically distributed with 2.8% under prediction and 2.7% over prediction. This indicates a very good overall fit.

Out of the 6219 confidence intervals of percentages with a desired coverage of 95%, the actual coverage was 92.86%. The instances where the long-term average was outside of the confidence interval contained 4.3% under prediction and 2.8% over prediction. Overall, the actual coverage is almost identical with the desired coverage. Furthermore, the actual coverage is also nicely centred, with the number of over and under predictions being almost equal. Subsequently, the confidence interval method calculation based on the delta method performs very well in the actual complex simulation.

Figure 3 shows the comparison of the confidence interval coverage of the delta method and the batching method. To verify the batching method, the confidence intervals of the frequencies and percentages have been obtained from 100 simulations, using a fixed number of 30 batches with independent batch means. A total of 2180 confidence intervals for both the frequencies and percentages have been evaluated, of which only 498 and 1503 confidence intervals contained the true mean value. Therefore the batch means method had a coverage of only 22.8% and 68.8% for the frequencies and throughputs respectively, missing the desired coverage of 95 by a wide margin. For the demonstrated example, the batching method is clearly inferior to the delta method for independent data.

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

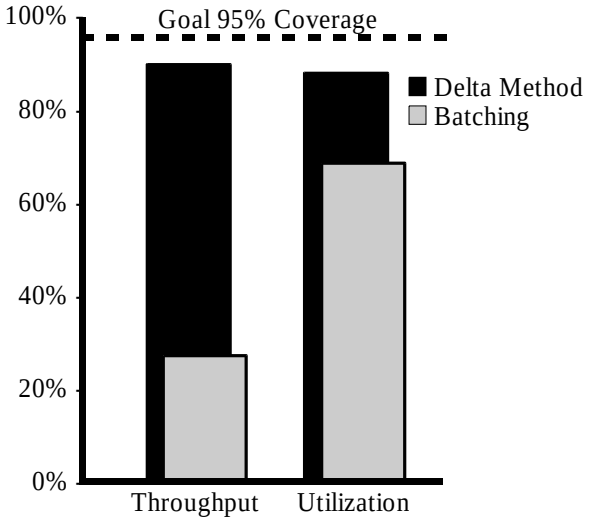

Figure 3: Batching vs. Delta method accuracy

4 CONCLUSION

This paper described how to measure the independence using the von Neumann RMSSDV, and how to calculate confidence intervals of functions of mean values using the delta method. Both, the frequent availability of independent data in simulations and the accuracy of the resulting confidence intervals have been demonstrated using different examples. The delta method is clearly superior to the batching method if independent data is available. More complex and imprecise methods as for example batching have to be used only if the resulting data is not independent. For independent data however, the confidence interval can be calculated directly. This also allows the calculation of related confidence intervals as for example the utilization or the throughput. The method provides extremely accurate results for independent and identically distributed data, as it was tested for a large number of different queuing system conditions. Additional tests with complex simulations also created highly accurate results for independent data. In addition to the accuracy, the method has a large number of benefits.

Compared to batching, the delta method is very fast to calculate the confidence interval, as it is not necessary to calculate different batch sizes and perform complex statistical tests. Furthermore, the method works also with small sets of data. While for batching, each batch has to satisfy certain statistical requirements and subsequently has a minimum size; the presented method requires only one set of data, allowing the calculation of a confidence interval at a much earlier stage during the simulation. This is extremely useful for example to analyse rare events, where even a long simulation does not have many events of interest, and subsequently batch means methods cannot be applied.

In summary, the method provides a preferable alternative for approximately independent data to calculate the function of one or more means, as for example frequencies or percentages. Therefore batching methods should only be used if the data required for the performance measure is not independent.

REFERENCES

Alexopoulos, Christos, and Seila, Andrew F. 2000. Output Analysis for Simulations. In *Winter Simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 101-108, Orlando, Florida, USA.

Banks, Jerry 1998. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice.* John Wiley & Sons.

Devore, Jay L. 1995. *Probability and Statistics for Engineering and the Sciences.* Belmont: Duxbury Press, Wadsworth Publishing.

Freedman, Laurence S. 2001. Confidence intervals and statistical power of the 'Validation' ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, 96(1), 143-153.

Henderson, Shane G. 2000. Mathematics for Simulation. In *Winter Simulation Conference*, P. A. Fishwick, K. Kang, J. A. Joines, and R. R. Barton, 137-146, Orlando, Florida, USA.

Kleijnen, Jack P. C. 1987. *Statistical Tools for Simulation Practitioners.* New York and Basel: Marcel Dekker.

Roser, Christoph, and Masaru Nakano. “Single Simulation Confidence Intervals.” In Proceedings of the Japan-USA Symposium on Flexible Automation, 289–94. Hiroshima, Japan, 2002.

Kubota, Fumiko, Sato, Shuichi, and Nakano, Masaru 1999. Enterprise Modeling and Simulation Platform Integrated Manufacturing System Design and Supply Chain. In *IEEE Conference on Systems, Man, and Cybernetics*, 511-515, Tokyo, Japan.

Law, Averill M., and Kelton, David W. 1991. *Simulation Modeling & Analysis.* McGraw Hill.

Moore, Linda J., and Sa, Ping 1999. Comparisons with the best in response surface methodology. *Statistics & Probability Letters*, 44(2), 189-194.

Nakano, Masaru, Sugiura, Norio, Tanaka, Minoru, and Kuno, Toshitaka 1994. ROPSII: Agent Oriented Manufacturing Simulator on the basis of Robot Simulator. In *Japan-USA Symposium on Flexible Automation*, 201-208, Kobe, Japan.

Neumann, John von 1941. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Annals of Mathematical Statistic*, 12, 367-395.

Neumann, John von 1942. A Further Remark Concerning the Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Annals of Mathematical Statistic*, 13, 86-88.

Oliveira, Nelson F., de, Santana, Vilma S., and Lopes, Antonio Alberto 1997. Ratio of proportions and the use of the delta method for confidence interval estimation in logistic regression. *Revista de Saude Publica*, 31(1), 90-99.

Papoulis, Athanasios 1991. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill.

Rinne, Horst 1997. *Taschenbuch der Statistik.* Frankfurt am Main: Verlag Harri Deutsch.

Steiger, Natalie Miller, and Wilson, James R. 1999. Improved Batching for Confidence Interval Construction in Steady State Simulation. In *Winter Simulation Conference*, P. A. Farrington, D. T. Nembhard, D. T. Sturrock, and G. W. Evans, 442-451, Phoenix, AZ, USA.

Student 1908. The probable error of a mean. *Biometrika*, 6, 1-25.