CONFIDENCE INTERVALS FROM A SINGLE SIMULATION USING THE DELTA METHOD[1]

Christoph Roser

Masaru Nakano

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

Keywords:

Measurement and Control

Flexible Manufacturing System (FMS)

Confidence Intervals

Prediction Accuracy

Word count: 3557 words

[1] This paper is based on the paper "Single Simulation Confidence Intervals using the Delta Method" by Christoph Roser and Masaru Nakano, presented at the "International Symposium on Scheduling" in Hamamatsu on June 4-6, receiving the "Best Paper" award! The paper has been edited and expanded for the special issue of the "International Journal of the Japan Society of Mechanical Engineering".

ABSTRACT

This paper describes a method for the calculation of confidence intervals of simulation throughputs and utilizations. The method is based on the delta method and uses only a single simulation, where the variation of the underlying means is used to determine the variation of the performance function by using the first derivative of the performance function. While the method has some limitations, it can be used frequently in practice. In addition, the method can also be used for short simulations or rare event applications, where methods based on batch means fail. This method can easily be implemented into existing simulation software.

INTRODUCTION

Scheduling is the process to arrange a number of tasks in a sequence. A frequent goal is to reduce the overall time for the performance of all tasks, or to ensure the completion of some tasks before a deadline. The time needed for the tasks have to be known to ensure the timely completion of the tasks. Unfortunately, these times are rarely static but vary depending on outside influences and random events. Often, simulation is used to estimate these times, and confidence intervals are used to determine the accuracy of these estimates. The underlying equations to calculate a confidence interval are well known [1].

However, there are some complications to calculate confidence intervals in discrete event simulation. One complication is, that independent and identically distributed (i.i.d.) data is required, but simulation data is frequently neither independent (e.g. waiting times) nor identically distributed (e.g. warming up period) [2], [3]. 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. This further complicates the calculation of confidence intervals

This paper describes a method to determine the variance of the function of the means of one or more variable using the mean and variance of the variables and the gradient of the function at the mean value. This paper is based on a paper presented at the International Symposium on Scheduling 2002 in Hamamatsu, Japan [4].

References [5], [6], and [7] give good overviews of the currently available methods for confidence interval calculation in simulations, with the most popular method being the batch means method. There are a number of different batch means and related methods developed. References [8], [9], [10], [11] give an overview of different batching methods, like overlapping batch means, non overlapping batch means, or fixed number of batches methods. There are a number of problems associated with the batch mean method. First, it is difficult to decide on the number of batches. Secondly, large data sets are needed to achieve valid confidence intervals. Third, different batching methods differ widely in their results. Finally, it is computationally intensive to calculate the confidence intervals, and therefore the confidence intervals are usually not calculated continuously during simulations.

This paper addresses the calculation of confidence intervals of functions of mean values. The presented method avoids most of the above problems for i.i.d. data. Example applications are given for throughputs and utilizations. The method is then validated experimentally using a complex simulation.

INDEPENDENCE OF DATA

While queue performance measures in a manufacturing system are usually heavily dependent, machine performance measures are surprisingly often independent or near independent. These independent responses can be seen as an individual machine with an independent working time distribution, failure time distribution, etc, giving independent and identically distributed data. Subsequently, this data can be used for subsequent statistical analyses as for example the calculation of the standard deviation or the confidence intervals as shown in equation (1).

$$\bar{x}_i = n_i^{-1}\sum_{j=1}^{n_i} x_{i,j} \qquad \sigma_{x_i} = \sqrt{\frac{\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2}{n_i - 1}} \tag{1}$$

The independence can be tested using the von Neumann ratio η of the mean squared successive difference to the variation (RMSSDV) [12, 13]. Equation (2) shows the calculation of the RMSSDV η based on a set of data x of size n, where the mean squared difference between successive data is divided by the variance of the data. The variable $x_{i,j}$ denotes the j^{th} element of the data set x_i. Variants of equation (2) can be found in [3] or [14].

$$\eta_i = \frac{n_i}{n_i - 1}\cdot\frac{\sum_{j=1}^{n_i-1}\left(x_{i,j+1} - x_{i,j}\right)^2}{\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2} \tag{2}$$

If the data x is independent the RMSSDV η has a value of two. Thus this method can be used to determine if the collected data is approximately independent (i.e. with a mean value at or near two) or not (i.e. the mean differs from two). The independence of the data of the selected example will be shown in more detail below.

DERIVED PERFORMANCE MEASURES

The second problem in measuring the manufacturing system performance measures is the handling of derived performance measures, i.e. performance measures which are based on a function of the mean value of another performance measure. Common examples are frequencies or throughputs, which are the inverse of the mean time between the respective events (TBE), for example the throughput is the inverse of the mean time between the completion of the parts. The delta method will be demonstrated in detail for frequencies, before other performance measures as for example percentages or general functions are explained.

Frequencies and Throughputs

Frequencies are a measurement of the number of occurrences in a given time. Throughputs are a type of frequencies, measuring the number of parts produced in a given period of time. Other frequencies are for example failure rates, i.e. the number of failures in a given period of time. These frequencies are defined by the number of occurrences of an event in a given period of time. This can also be described as the inverse of the mean TBE. Subsequently, the frequency can be defined as the inverse of the average TBE $\bar{x}_1$ as shown in equation (3).

$$y = f(\bar{x}_1) = \frac{1}{\bar{x}_1} \tag{3}$$

Problem Statement

Equation (3) allows the calculation of a frequency for one set of data. However, since equation (3) generates only one value y from a set of data x, it is not possible to calculate a valid standard deviation or confidence interval. However, if the data for the TBE is i.i.d, a standard deviation, and a confidence interval of the time between events can be calculated. Yet, this variation cannot be simply transformed using equation (3), as this would not only create an incorrect variance, but also an incorrect mean frequency as shown in Figure 1. In general, equation (4) holds true for all nonlinear functions. Only for linear functions is the mean of the functions of the data values equal to the function of the means [15].

$$f(\bar{x}_1, \bar{x}_2 \ldots) \neq E[f(x_1, x_2 \ldots)] \tag{4}$$

(Insert Figure 1 about here)

Conventional Method Batching

Currently, in order to evaluate the variance of the derived performance measure, batching is used. Generally speaking, batching replaces the distribution of the data with a distribution of the batch means. This distribution of the batch means has a much smaller variance than the original data, and the batch means are usually sought to be i.i.d. Due to this smaller variance, the underlying performance function can be seen as approximately linear over the range of the batch means. Therefore, the variance of the functions of the batch means represents approximately the variance of the mean function, or for the case of the frequencies, the variance of the inverse batch means represent approximately the variance of the mean frequency. This is also visualized in Figure 2.

(Insert Figure 2 about here)

Of course, this is only an approximation, whose accuracy depends on the range of the batch means and the curvature of the performance function over this range, and significant errors are possible. Again, only for linear functions will the function variance be estimated correctly.

The Delta Method for Frequencies

This leads to the natural conclusion, to replace the function with a tangent at the mean value in order to predict the variance of the function. This approach is called the “delta method”, also occasionally known as moment matching method. The delta method replaces the function *f* by its tangent f^* at the mean values $\bar{x}_i$. Using this tangent f^*, it is possible to determine the standard deviation σ_f of the function *f* of the mean $\bar{x}_1$ based on the deviation of the variables σ_{x1}. The functional evaluation of the standard deviation of the frequency in equation (3) based on the mean $\bar{x}_1$ and the standard deviation σ_{x1} of the TBE is shown in equation (5). Figure 3 visualizes the throughput example for a tangential line f^* replacing the function *f*.

$$\sigma_y = \sqrt{\left(\frac{-1}{\bar{x}_1^2} \cdot \sigma_{x_1}\right)^2} \tag{5}$$

(Insert Figure 3 about here)

The resulting standard deviation σ_y of the function value *y* can then be used to calculate desired measures of accuracy, as for example a confidence interval as shown in equation (6), where *t* is the student-t distribution and α is the confidence level [16].

$$CI_y = t_{n_1 - 1, \alpha/2} \cdot \frac{\sigma_y}{\sqrt{n_1}} \tag{6}$$

Of course, it is also possible to use the tangent to translate other measures of variation from the TBE to the frequency. Figure 4 for example shows how the confidence interval of the TBE is translated directly into a confidence interval of the frequency. This may be useful for cases where the TBE is non-normal distributed, and more complex calculations are needed to describe the behavior of the TBE.

(Insert Figure 4 about here)

It is even possible to combine the delta method with the batch means method, for example in cases where the underlying TBE data is not i.i.d. In this case, batch means may be used to represent the variance of the underlying TBE data while generating independent batch means. A standard deviation and a confidence interval may then be calculated based on the batch means of the TBE. This standard deviation and confidence interval may then be translated into a standard deviation and a confidence interval of the frequency using the delta method. This approach combines the accuracy of the delta method with the ability to handle dependent data using batch means. If the batch means itself are non-normal distributed, it is also possible to translate the batch means using the tangent function to determine a set of representative data of the frequency.

The Delta Method for Percentages

Another common performance measure in discrete event simulation are percentages of times, as for example the percentage of time a machine is working, or the percentage of time a machine is under repair. In general, a percentage can be calculated by dividing the total time a machine is in a certain state by the total simulation time. This can also be represented as the mean duration a machine is in a certain state $\bar{x}_2$ divided by the mean duration between the beginnings of a certain state $\bar{x}_1$. For example, the percentage repair is the mean time to repair divided by the mean time between the beginnings of repairs. The function of a percentage based on two mean values is shown in equation (7).

$$y = f(\bar{x}_1, \bar{x}_2) = \frac{\bar{x}_1}{\bar{x}_2} \qquad \mathit{for}\ \bar{x}_2 \geq \bar{x}_1 \geq 0,\ \bar{x}_2 > 0 \qquad (7)$$

Replacing the function equation (7) with a tangential plane at the mean values, the standard deviation of the percentage can be determined as shown in equation (8). This equation (8) also includes the effect of the covariance between the two variables x_1 and x_2 as shown in equation (9). Note that equation (9) requires the size n of the data sets x_1 and x_2 to be equal. If the two variables are independent of each other, the covariance is zero and the term can be dropped. A subsequent confidence interval can be calculated as shown in equation (6).

$$\sigma_y^2 = \left(\frac{\bar{x}_2}{\bar{x}_1^2} \cdot \sigma_{x_1} \right)^2 + \left(\frac{1}{\bar{x}_1} \cdot \sigma_{x_2} \right)^2 + \\ 2 \cdot \frac{\bar{x}_2}{\bar{x}_1^2} \cdot \frac{1}{\bar{x}_1} \cdot Cov[x_1, x_2] \tag{8}$$

$$Cov[x_1, x_2] = \frac{\sum_{j=1}^{n_1} x_{1,j} \cdot x_{2,j} - \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1,j} \cdot \sum_{j=1}^{n_1} x_{2,j}}{n_1 - 1} \tag{9}$$

The General Delta Method

Assume there is a general performance measure y as a function f of the mean values one or more variables $\bar{x}_i$. Common examples are a throughput as an inverse of the time between parts, or utilization as the working time divided by the time between parts. The mean values $\bar{x}_i$ are calculated based on a set of n_i data values $x_{i,j}$, where the mean $\bar{x}_i$ and the standard deviation σ_{xi} is calculated using the well-known equations as shown in (1).

Using the standard deviation and the covariance of the variables $\bar{x}_i$, the standard deviation of the function value y can be determined using the delta method as shown in equation (10) [17]. Equation (10) includes the effect of the correlation between two paired variables, where $cov[x_1, x_2]$ is the unbiased estimate of the covariance as shown in equation (9) [15].

$$\sigma_y^2 = \left(\left[\frac{df}{dx_1} \right]_{\substack{x_1=\bar{x}_1 \\ x_2=\bar{x}_2}} \cdot \sigma_{x_1} \right)^2 + \left(\left[\frac{df}{dx_2} \right]_{\substack{x_1=\bar{x}_1 \\ x_2=\bar{x}_2}} \cdot \sigma_{x_2} \right)^2 + \tag{10}$$

$$2 \cdot \left[\frac{df}{dx_1} \right]_{\substack{x_1=\bar{x}_1 \\ x_2=\bar{x}_2}} \cdot \left[\frac{df}{dx_2} \right]_{\substack{x_1=\bar{x}_1 \\ x_2=\bar{x}_2}} \cdot Cov[x_1, x_2]$$

COMPLEX MANUFACTURING SYSTEM

The presented method was 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 5 [18], [19].

(Insert Figure 5 about here)

The total simulation time was almost two years of simulation. After removing the warming up period, this data was then split into 101 subsets with a simulation time of 6 days each. In order to calculate valid confidence intervals, the data has to be independent. Therefore, the RMSSDV has been calculated for the data using equation (2) to determine if the data is independent. While simulations are notorious for dependent data, the actual machine performance data was surprisingly often independent or near independent.

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 69 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 1. Table 1 shows not only the RMSSDV of the durations of the events (working, repair, blocked and idle), but also the time between the start of the events and the time between the end of the event and the beginning of the next event. This allows the calculation of a valid standard deviation and a confidence interval for these values as described in more detail below. This allows the calculation of a valid standard deviation and a confidence interval for these values as described above.

For each of the 101 subsets, the frequencies and the percentages of all machines working, idle, blocked or repaired were measured and the 95% confidence intervals calculated. These confidence intervals were then compared to the overall average, 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%. However, in the real case, the percentage of the confidence intervals containing the true value may differ from the ideal case, i.e. the actual coverage differs from the desired coverage. The closer the actual coverage is to the desired coverage, the more accurate is the confidence interval method. Table 2 shows an overview of the coverage results of the complex simulation.

(Insert Table 2 about here)

Out of the 6219 frequency confidence intervals 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 percentage confidence intervals 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. While the fit is not as good as for the frequencies, the coverage is still very close to the desired coverage. Overall, the actual coverage is almost identical with the desired coverage. Furthermore, the actual coverage is also nicely centered, with the number of over and under predictions being almost equal.

The presented method has been compared to the batching method, where the confidence interval is based on the batch means. 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 coverage of only 22.8% and 68.8% for the frequencies and throughputs respectively, missing the desired coverage of 95 by a wide margin and is clearly inferior to the delta method for independent data. Figure 6 shows the results of the delta method compared to the batch means method.

(Insert Figure 6 about here)

CONCLUSION

In conclusion, the method provides very accurate results for near independent and identically distributed data. While simulation data is known to be dependent, the machine performance data was actually found to be frequently independent, allowing the calculation of the confidence intervals using the delta method.

Compared to batching, it is very fast to calculate the confidence interval, as it is not necessary to calculate different batch sizes and perform complex statistical tests. Moreover, if additional data becomes available, this data can easily be integrated into the previous calculation, and the confidence interval can be updated. This allows a sequential adding of data while updating the confidence interval.

Furthermore, the method works also with small sets of data. This is extremely useful for example to analyze rare events, where even a long simulation does not have many occurrences of the rare event, and subsequently batch means methods cannot be applied.

In summary, the method provides a preferable alternative to calculate the confidence intervals for approximately independent data.

REFERENCES

(1) Devore, J.L., *Probability and Statistics for Engineering and the Sciences*. 4th edition ed. 1995, Belmont: Duxbury Press, Wadsworth Publishing.

(2) Rinne, H., *Taschenbuch der Statistik*. 2 ed. 1997, Frankfurt am Main: Verlag Harri Deutsch.

(3) Kleijnen, J.P.C., *Statistical Tools for Simulation Practitioners*, ed. D.B. Owen. 1987, New York and Basel: Marcel Dekker. 448.

(4) Roser, C. and M. Nakano. *Single Simulation Confidence Intervals using the Delta Method*. in *International Symposium on Scheduling*. 2002. Hamamatsu, Japan.

(5) Alexopoulos, C. and A.F. Seila. *Output Analysis for Simulations*. in *Winter Simulation Conference*. 2000. Orlando, Florida, USA.

(6) Law, A.M. and D.W. Kelton, *Simulation Modeling & Analysis*. 2 ed. 1991: McGraw Hill.

(7) Banks, J., *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. 1998: John Wiley & Sons.

(8) Seila, A.F. *Advanced output analysis for simulation*. in *Winter Simulation Conference*. 1992. Arlington, VA USA.

(9) Goldsman, D. *Simulation output analysis*. in *Winter Simulation Conference*. 1992. Arlington, VA USA.

(10) Schmeiser, B.W. and W.T. Song. *Batching methods in simulation output analysis: what we know and what we don't*. in *Winter Simulation Conference*. 1996. San Diego, CA USA.

(11) Pawlikowski, K., *Steady-state simulation of queueing processes: survey of problems and solutions.* ACM Computing Surveys, 1990. **22**(2): p. 123-170.

(12) Neumann, J. v., *Distribution of the Ratio of the Mean Square Successive Difference to the Variance.* Annals of Mathematical Statistic, 1941. **12**: p. 367-395.

(13) Neumann, J. v., *A Further Remark Concerning the Distribution of the Ratio of the Mean Square Successive Difference to the Variance.* Annals of Mathematical Statistic, 1942. **13**: p. 86-88.

(14) Steiger, N.M. and J.R. Wilson. *Improved Batching for Confidence Interval Construction in Steady State Simulation*. in *Winter Simulation Conference*. 1999. Phoenix, AZ, USA.

(15) Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. 3rd edition ed. 1991: McGraw-Hill.

(16) Student, *The probable error of a mean.* Biometrika, 1908. **6**: p. 1-25.

(17) Rao, C.R., *Linear statistical inference and its applications*. 2nd Edition ed. 2001: Wiley Press. 656.

(18) Kubota, F., S. Sato and M. Nakano. *Enterprise Modeling and Simulation Platform Integrated Manufacturing System Design and Supply Chain*. in *IEEE Conference on Systems, Man, and Cybernetics*. 1999. Tokyo, Japan.

(19) Nakano, M., N. Sugiura, M. Tanaka and T. Kuno. *ROPSII: Agent Oriented Manufacturing Simulator on the basis of Robot Simulator*. in *Japan-USA Symposium on Flexible Automation*. 1994. Kobe, Japan.

TABLES

Table 1: RMSSDV of Complex Simulation

Measure	Occurrences	RMSSDV		
		Duration	Time between Occurrences	Interval without Occurrence
M1 Working	49049	2.0	2.0	2.0
M1 Blocked	49050	2.0	2.0	2.0
M2 Working	49049	2.0	2.0	2.0
M2 Blocked	14261	1.8	2.0	2.0
M2 Repair	1043	2.0	2.0	2.0
M3 Working	16349	2.0	1.8	1.8
M3 Idle	6151	1.7	1.8	1.8
M3 Blocked	319	6.1	2.1	2.1
M3 Repair	1196	2.1	2.0	2.0
M4 Working	16349	2.0	1.8	1.8
M4 Idle	16061	1.8	1.8	2.0
M4 Blocked	8	Insufficient Data		
M4 Repair	494	4.6	2.1	2.1
M5 Working	3037	1.7	1.7	1.9
M5 Idle	50	2332.1	2.1	2.2
M5 Blocked	1721	3.7	2.1	2.1
M5 Repair	1291	2.1	2.0	2.0
M6 Working	49046	2.0	1.9	1.9
M6 Idle	11934	1.8	2.0	2.0
M6 Blocked	48205	2.0	1.9	2.0
M6 Repair	893	1.9	2.1	2.1
M7 Working	49046	2.0	1.9	1.9
M7 Idle	12755	1.7	1.7	1.7
M7 Repair	1172	1.9	2.0	2.0

Table 2: Simulation Example Coverage

Performance Measure	Desired Coverage	Actual Coverage	Too Small	Too Large
Frequency	95%	94.4%	2.9%	2.7%
Percentage	95%	92.9%	4.3%	2.8%

FIGURES

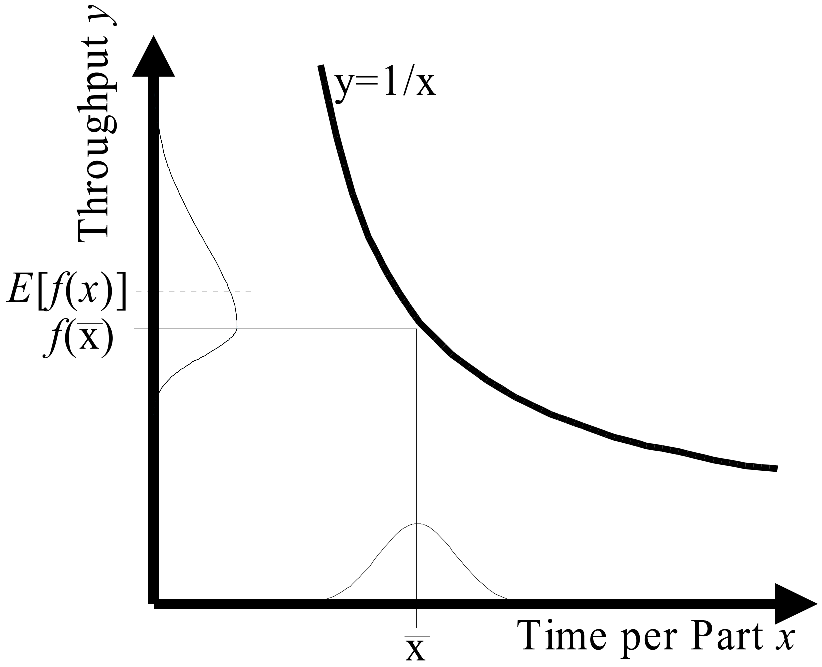

Figure 1: Inverse of a Set of Data

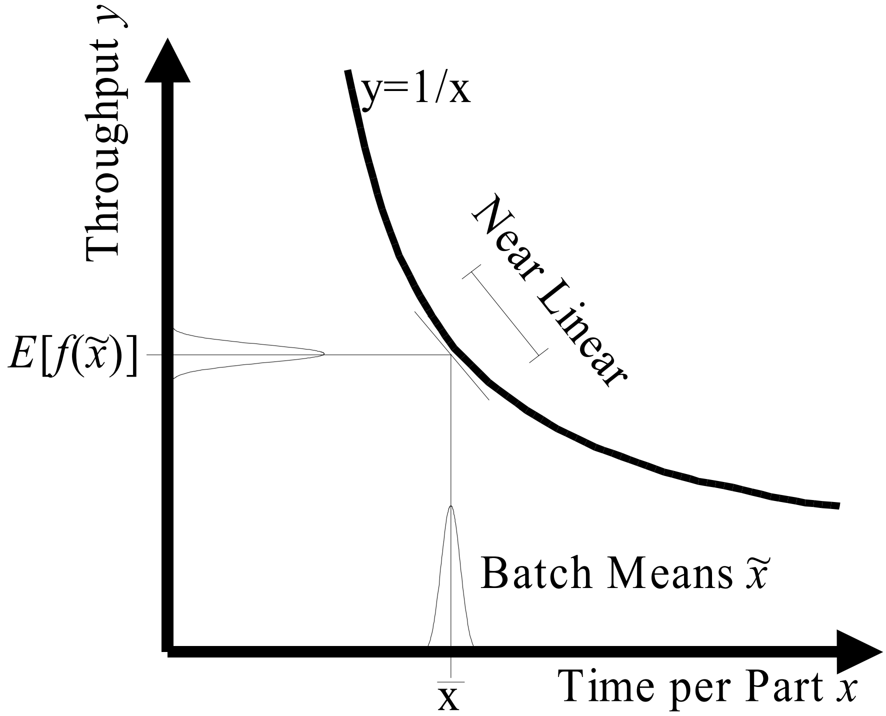

Figure 2: Batch Means

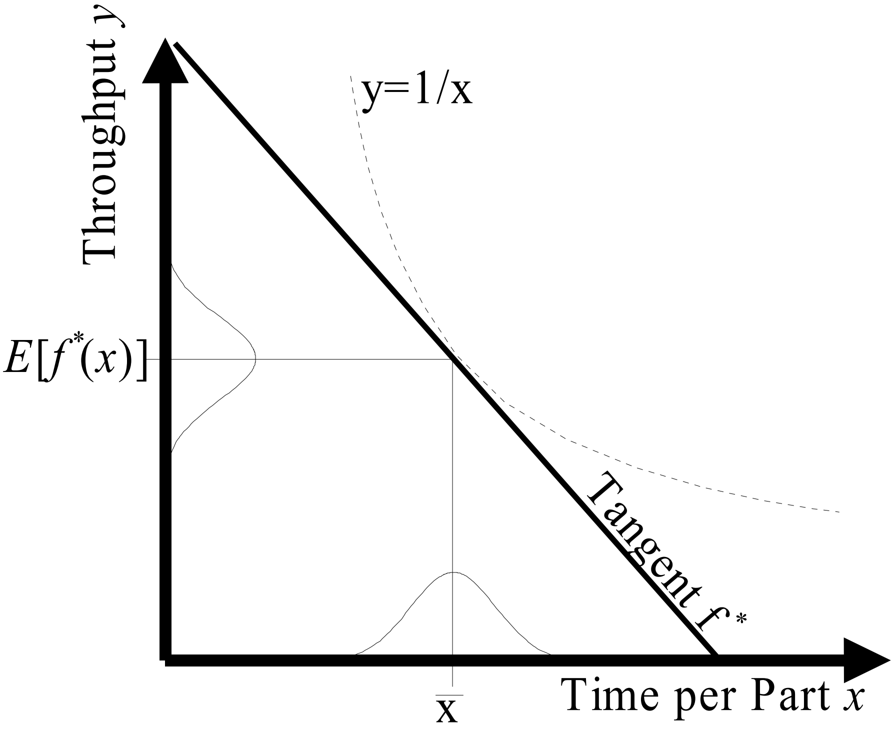

Figure 3: Tangent at the Mean Value

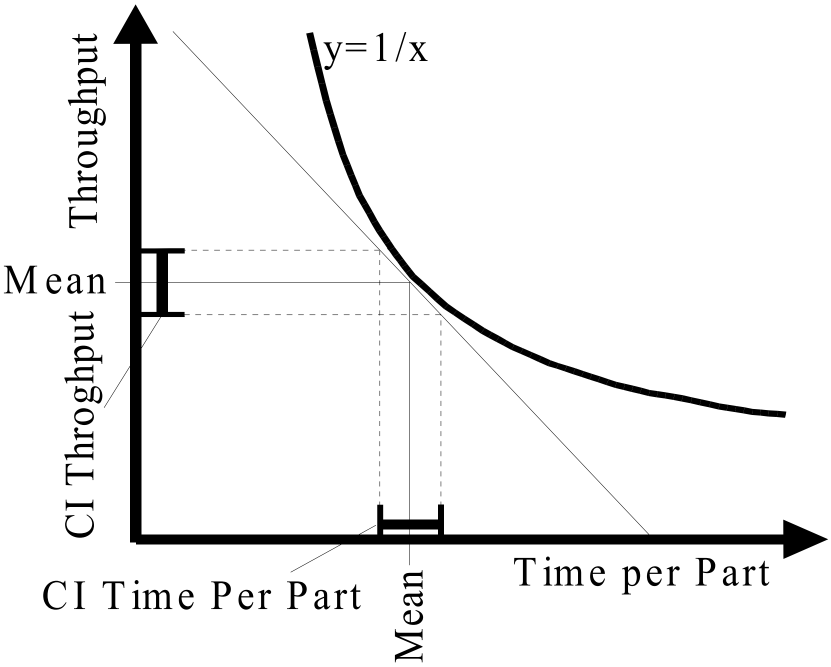

Figure 4: Translation of Confidence Interval

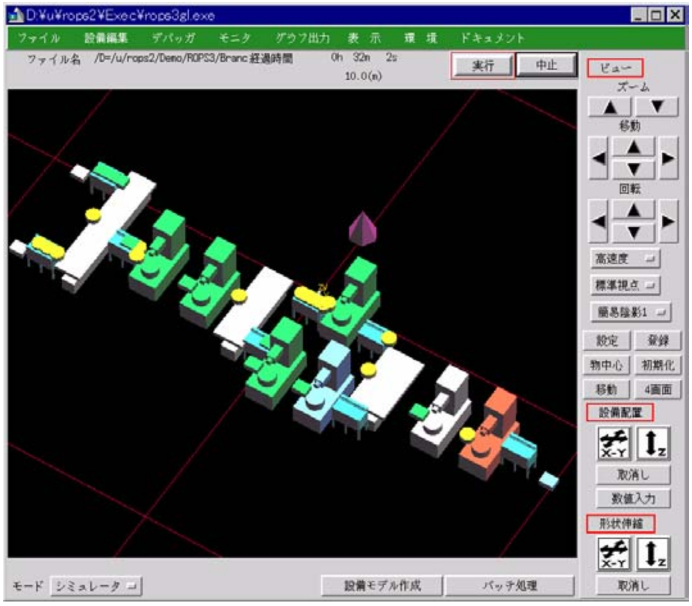

Figure 5: GAROPS Simulation Example

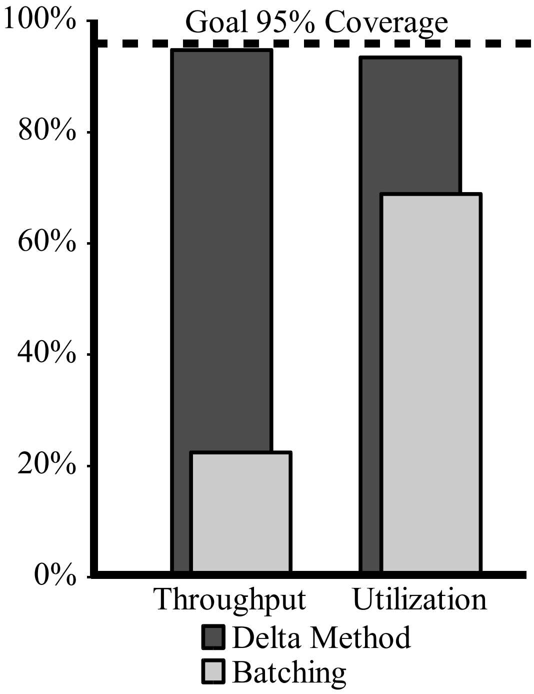

Figure 6: Complex Example Coverage Results