A Few What-Ifs on Using Statistical Analysis of Stochastic Simulation Runs to Extract Timeliness Properties

Nuno Pereira[1], Eduardo Tovar[1], Berta Batista[1], Luis Miguel Pinho[1], Ian Broster[2]

[1] Polytechnic Institute of Porto, Porto, Portugal

[2] University of York, York, UK

Abstract

Analytical models to provide timeliness guarantees tend to either be based on simplifications often leading to very pessimistic assumptions or consider more accurate techniques but at the cost of adding rather complex, difficult to handle, abstractions. Moreover, the more flexible and adaptive nature of current systems demand approaches for the timeliness evaluation problem based on probabilistic measures of meeting deadlines.

It is in this context that simulation can emerge as an adequate solution to understand and analyze the timing behaviour of actual systems. However, special care must be taken with the obtained outputs under the penalty of obtaining results with lack of credibility. Particularly important, is to consider that we are more interested in values which are at the tail of a probability distribution (near worst-case probabilities), instead of deriving confidence on mean values.

We approach this subject considering the randomness nature of simulation output data. In this, we will start by discussing well known approaches for estimating distributions out of simulation output, and the confidence which can be applied to its mean values. This will serve the basis to discuss on the applicability of these approaches to derive confidence on the tail of the distributions, where the worst-case is expected to be.

1.Motivation

Timeliness analysis of real-time systems has been commonly exploited in a framework dominated by the notion of absolute temporal guarantees. In those systems, computational and communication loads are presumed to be bounded and known, and the worst-case (at least believed to be) conditions are assumed.

However, it is now accepted that those guaranteed approaches may pose serious problems when applied to some targeted systems, such as those related to distributed real-time applications. In fact, for many distributed systems, analytical-based worst-case formulations tend to be overwhelmed with simplifications that often lead to pessimistic assumptions, and therefore to very pessimistic worst-case results. There are, however, a number of techniques

that can be applied in order to reduce the pessimism level. These include considering precedence relationships, event phasing and inheritance of time characteristics, (or combinations of these) into the analytical formulations. However, and good things come also often with a caveat, improving tightness of the schedulability analysis may be achieved at the cost of adding rather complex abstractions, which unfortunately makes it further difficult to handle and reason those analytical abstractions.

To add on top of this, the more flexible and adaptive nature of those “more complex” distributed systems creates the eagerness to approach the timeliness evaluation problem in a different way: instead of using a guaranteed approach, why not tackling the problem by trying to find a probabilistic measure of meeting deadlines?

It is in this context that simulation can emerge as an adequate solution to tackle the problem of engineering complex distributed systems. On the other hand, the relatively recent advent of fast and inexpensive computational power allows the approach of trying to model the system as faithfully as possible, and then use simulation to obtain accurate characteristics.

Can simulation be used to extract useful timeliness results about the modelled systems?

A simulation is the imitation of a real-world process or system over time [1]. It is based on the construction of a simulation model that will allow the expression and investigation of a wide variety of "what-if" questions, and in that way be used to obtain some temporal inferences about the real world system.

Although a powerful tool, simulation may hide some traps, so special care must be taken under the penalty of obtaining results with lack of credibility [2]. Particularly important, is to consider that, within real-time systems, we are more interested in values which are at the tail of the distribution (near worst-case probabilities), instead of deriving mean values. In order to be able to reason about probabilistic measures of meeting deadlines, we need to be able to understand the result of simulation for the case of rare events, as we expect worst-case to be.

The first necessary step of any performance evaluation studies based on stochastic simulation is to use a valid simulation model. The next step is to ensure that valid simulation experiments take place. Two main issues have to be addressed for assuring the validity of the stochastic experiment [2]:

1) application of appropriate elementary source(s) of randomness;
2) appropriate analysis of simulation output data.

In this work, we focus our attention into the second point; that is, a valid model of the system is assumed to be already constructed. For the analysis of the simulation output data, we will

start by discussing the randomness nature of this output, and what are the appropriate approaches to analyse the results. In the presented reasoning, we will start by discussing well known approaches for estimating distributions out of simulation output, and the confidence which can be applied to its mean values. This will serve the basis to discuss on the applicability of these approaches to derive confidence on the tail of the distribution, where the worst-case is expected to be.

2.Simulation Output Data

By their nature, stochastic simulation models will produce random outputs. Thus, simulation has to be regarded as a computer-based statistical experiment, and, to have any meaning, appropriate statistical techniques must be employed to analyse the simulation experiments.

Moreover, the data resulting from a simulation cannot be directly analysed using traditional statistical methods, since most of these apply to Independent and Identically Distributed (IID) data. This is an important topic of concern for the remainder of this text.

Let us consider a simple example of a waiting queue, with a random service time. The waiting time experienced by the first user will always be zero. On the other hand, the waiting time of the second user will depend on the departure of the first one, and so on. If we are interested in studying the waiting time in the queue, it is easy to observe that the distribution of these times is neither identically distributed nor independent.

A method commonly used to overcome this problem is to make observations from the results of multiple, and independent, simulation runs (or simulation replicas). Typically this is performed by making multiple simulation runs with the same initial conditions and parameters, but yet different seeds for the random numbers used to drive the simulation through time. In this way, it is possible to obtain independent and identically distributed variables. Hence, it is possible to make estimates of variables of interest, such as the average delay observed in each simulation, the number of messages dropped in each simulation, or the maximum response time observed in each simulation, just to roll a few examples.

Next, we will briefly survey the typical mathematical formulations for obtaining an estimator for a mean value, as well as its respective confidence interval. Another aspect of concern is how to get confidence intervals with some specified precision. These are crucial pieces of basic statistical reasoning used in the majority of the approaches for simulation output data analysis.

2.1 Statistical Ground for the Analysis of Simulations Output Data

Suppose we would like to obtain an estimate for the mean of an output variable. By the way of example, let us say it is the mean message delay in queue to access a communication medium. For a matter of simplicity, consider that we would like to observe this delay during a defined period, because the system is shutdown or restarted after that period (imagine the case of a system that is disconnected at the end of a working day) – a terminating simulation.

One run of the simulation will produce one estimate for the mean message delay. Noticeably, the value of just one sample of a random process has no significance by itself. However, executing multiple runs of the simulation will provide a set of mean delay values, characterized by some distribution. Moreover, it will be IID, as we have seen.

The sample mean (remember that, in this example, our samples are the set of mean delays observed for each simulation replica) is a natural estimator of the (unknown) true mean message delay.

But, how reliable is this estimate?

If we would make another set of simulation replicas, the result would, most likely, be different. Indeed, an estimate without an indication of its precision is of little value.

However, to come up with this type of conclusions, one would have to know something about the distribution of the sample. There is a basic, but very useful and important concept in statistics, called the central limit theorem. This theorem basically states that the sum and the average of many random values present a distribution close to normal. Typically, a normal approximation is sufficiently good if about 30 or more values are used in the sum (or average) [3]. Then, well-known methods can be used to draw confidence intervals from normal distributions.

There is, however, an important aspect to point out. The standard procedures for inference are developed for situations where the standard deviation for the entire population is known. As usually there is not the knowledge of the entire population, there is the need to also estimate the standard deviation from the available data, in which case, the statistic will not have a normal distribution, but a *t*-distribution.

For the sake of completeness, let us now lay down some basic statistics, applied to the estimation of the model true characteristics. The validity of this estimation, and of its confidence interval, is well known for the mean value of a distribution. But will it be applicable to the tail of the distribution (the worst- or near to worst-case)? This will be then discussed in Section 3.

Suppose that $X_1, X_2, \ldots X_n$ are IID random variables with a mean μ (in our example, the mean message delay in queue to access a communication medium) and a variance σ^2. Our primary objective is to estimate μ. The sample mean ($\overline{X}(n)$), is an unbiased (point) estimator of μ, and is defined by:

$$\overline{X}(n) = \frac{\sum_{i=1}^{n} X_i}{n} \quad \textbf{(1)}$$

That is, the expected value of $\overline{X}(n)$ is μ: $E[\overline{X}(n)] = \mu$. If we perform a very large number of independent experiments, each resulting in a $\overline{X}(n)$, their average will be μ.

While $\overline{X}(n)$ is the estimator of μ, in a similar way, the sample variance ($S^2(n)$) is an unbiased estimator of σ^2:

$$S^2(n) = \frac{\sum_{i=1}^{n} \left[X_i - \overline{X}(n) \right]^2}{n-1} \quad \textbf{(2)}$$

As we have discussed, it is important to have an assessment of the estimation precision. The usual way to do this is to construct a confidence interval. An approximate
$100(1 - \alpha)\%$ confidence interval for μ is given by:

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2} \sqrt{\frac{S^2(n)}{n}} \quad \textbf{(3)}$$

The estimate ($\overline{X}(n)$, in our case) represents the guess for the value of interest. The margin of error (terms after the $\pm$ sign) gives a measure on how accurate the estimation is, based on the variability of the estimation.

The confidence level reflects the amount of confidence that, in the long run, this approach will be able to approximate the true value of interest. As we increase the confidence level, the confidence interval gets wider. It can be shown that to cut the length of the confidence interval in half, four times more samples are required.

2.2 Non-terminating Simulations

So far, we have been concerned with a finite set of samples extracted from a terminating simulation. Nevertheless, non-terminating simulations are an important class that must be target of our attention. Indeed, often our systems of interest will not have a terminating event, and we will probably be interested in analysing the system's behaviour in the long run.

There are several subtypes of non-terminating simulations. We will consider a subtype where the outputs of the simulation model tend to stabilize; that is, the system reaches a steady state. A measure of performance for such simulation is said to be a steady-state parameter. We will then focus our attention to analysing non-terminating, steady-state parameters and thus assuming that, as the amount of data becomes large, distributions will converge to a common distribution in the steady-state.

The analysis of steady-state parameters raises a very important problem, which is how to choose the simulation data that actually represents the steady-state. Mostly due to the choice of starting conditions, the initial output data of the simulation is usually not very representative of the steady-state behaviour. This period, affected by the initialization bias, is usually referred to as the *warm-up* period. Using data from this period for the estimation of system's steady-state parameters may yield deceptive results.

To circumvent the warm-up period problem, one may simply resort to very long runs, such that the data from the initial phase has a negligible impact, or to start the simulation in a state supposed to be close to the steady-state. Effectively, these methods have some serious practical impairments, thus somewhat more elaborate methods are commonly used. These methods typically ignore data from the warm-up period, utilizing some techniques, based on the assumption that the variance of the samples is substantially lower in the steady-state than in the warm-up period, to detect when it ends.

The replication approach described earlier may still be used in the context of non-terminating simulations. All what is necessary is to define how to extract the steady-state means from each simulation replica (X_i used to compute the sample mean in (1)). Suppose that we make n simulation replicas, each of length m, where m is much larger than l (the length of deleted data used to eliminate the impact from initial conditions). As a rule of thumb, m-l should be at least 10 times the size of l. In the context of non-terminating simulations, this method is commonly called *replication/deletion.*

Let X_i be IID variables given by the mean in each simulation replica i, from the set of values collected between l and m (Y_{ij}):

$$X_i = \frac{\sum_{j=l+1}^{m} Y_{ij}}{m-l} \quad for \;\; i = 1,2,...n \tag{4}$$

Similarly to the terminating case, $\overline{X}(n)$ (1) is an approximately unbiased point estimation for the steady-state mean μ, and a confidence interval may be obtained with (3).

An informal description of the method may be as follows:

1. Define the size of the initial phase *l* from test simulation runs.
2. Perform *n* independent simulation replicas of length *m* (with *m* much larger than *l*).
3. For each simulation replica *i*, compute the mean of all observations after the initial phase *l*.
4. Apply usual point estimate and confidence intervals on the IID means obtained (given by equations (1), (2) and (3)).

As referred previously in Section 2.1, the confidence interval depends on the variance of X_i, which will be unknown when the first *n* simulation replicas are performed. If we make a fixed number of replications, the resulting confidence interval may be too wide for our particular purpose. However, also as pointed out in Section 2.1, we can decrease the length of the confidence interval by a factor of 2, by performing 4 times as many replications.

There are other methods that apply some variations. Instead of achieving independence trough multiple simulation runs, one can perform one long simulation run and try to obtain independent observations from subsets of data. The method of *batch means* [4], similarly to the replication/deletion, attempts to obtain independent observations, but, in this case, the single simulation run is divided into *batches*, where a batch takes the role of a single replica. It can be shown that, for a sufficiently large number of batches, the mean of the several batches will be approximately IID normal.

One of the most relevant advantages of this method is that it only has to go through one warm-up phase, on other hand, a major problem is on choosing the batch size *m*, or equivalently, the number of batches *k*. A number of guidelines from research literature, and a general recommended strategy may be found in [4].

In the group of methods based on one long simulation, other methods may be encountered [5]. These methods, such as the *autoregressive method* or *spectral analysis*, try to use estimates of the autocorrelation structure of the underlying stochastic process to obtain an estimate of the variance of the sample and then to construct a confidence interval. For sake of simplicity, we refer the reader to the literature [4, 5] for further information on other methods.

All the procedures described to this point are, usually, classified as fixed-sample procedures, where the sample sizes taken (the whole simulation, in the case of replication/deletion or the batch, in batch means) are of a fixed size. Generally, some conclusions may be established for all of these fixed-sample procedures [5]:

- if the total sample size is chosen too small, the actual coverage may be lower than the desired;
- the appropriate choice of the total sample size is extremely model dependent and impossible to choose arbitrarily.

Evidently, no procedure that fixates the run length before the simulation begins will always produce a satisfactory confidence interval. A sequential scenario where the simulation's end is determined by a relative statistical error that is verified in consecutive checkpoints is a more interesting approach. Sequential methods are commonly based on the same methods for non-terminating simulations as batch means or spectral analysis, in conjunction with absolute or relative-error stopping rules. These procedures are more complex, requiring computing the estimates at several points of the simulation to check if the stopping rule has been satisfied, which can be computationally very expensive. Additionally, these procedures may not be easily applicable when multiple measures of performance are needed, and, because of random nature of simulation, the relative stopping rule can be accidentally satisfied, resulting in premature termination of the simulation, and on wrong estimation results.

Another typical problem with sequential procedures is that they are not very popular among existent software packages. A simulation package supporting sequential procedures is Akaroa2, designed ate the University of Canterbury, New Zealand [2, 6]. One interesting feature is that it is possible to integrate Akaroa2 with other open simulation packages, such is the case of OMNeT++ [7].

Besides the problems described, sequential procedures are recognised as the a practical approach allowing control on the error of the final results of stochastic simulations [2].

3.What about Worst-Case Assessment?

All the previous methods seek to obtain a mean value for the output point estimator. What about other kind of measures? Consider that we would like to estimate the probability of a value belonging to an interval, for example, imagine the case of investigating the probability that a queue length is greater than k messages. Another different performance measure is a quantile. Quantiles describe the level of performance that can be delivered with a given probability p.

Next, we will briefly discuss how to extract such measures of performance like proportions, probabilities or quantiles. Then, we will address some ideas about extracting worst-case measurements from stochastic simulations.

3.1 Probabilities and Quantiles

Suppose we need to estimate the steady-state probability (p) of the mean message delays in queue to access a communication medium being less than a value x. The variable under analysis may be represented by 1 if the queue delay exceeds the value x, and 0 otherwise.

Making $p = P(Y \in B)$, where B is a set of real numbers smaller than x, and Y is the original steady-state random variable, we are just in the presence of a special case of estimating the mean, by letting the random variable Z be defined by:

$$Z = \begin{cases} 1, & if \quad Y \in B \\ 0, & otherwise \end{cases} \qquad \textbf{(5)}$$

It can be shown that estimating p is equivalent to estimating the steady-state mean for the expected value of Z ($E(Z)$).

A performance measure that does not fit in the same reasoning is a quantile. For instance, if the variable represents the delay in queue that a client experiences, the 0.90-quantile is the value x such that 90% of all messages experienced a delay shorter than x.

Estimating quantiles is both conceptually and computationally (in terms of number of observations required) a more difficult problem than estimating the steady-state mean. Additionally, most of the procedures for estimating these performance measures are based on order statistics and require storage and sorting of the observations. Nevertheless, the general reasoning is similar to the one for obtaining the interval estimator for a steady-state mean.

An example taken from [8] points-out one major problem with quantile estimation describes that, for the steady-state estimation of 0.99 quantile of waiting times, an estimate with relative a precision of 10% required about 500.000 observations, and a 0.999 quantile needed a samples size of approximately 2.300.000. Because quantile estimation require storage and sorting of observed values, obtaining small quantile estimations, with a good accuracy is often impractical. However, this is a problem under investigation, and several techniques already exist that do not require storage and sorting have already been attained and implemented. In [9], a number of such approaches are presented and evaluated.

3.2 Worst-case

As we have seen throughout this paper, simulation is usually a very good tool from evaluation of average behaviour.

But, what about worst-case behaviour?

3.2.1 Extreme value theory

Goodness-of-fit tests may be used to evaluate the likeness between the sample data distribution and a theoretical distribution. If it is possible to obtain a good approximation from the theoretical distribution, then it is usually feasible to obtain good estimates of the output variables. However, for the purpose of drawing worst-case estimates from these distributions we are considering the tails of the probabilistic distributions and it is known that these are the areas where less accuracy exists. Considering that we capture a sufficient number of values close to the worst-case value during the simulation runs, we will probably end up with a heavy-tailed distribution. A distribution function or random variable is said to be heavy-tailed it presents a high coefficient of variance. For example, in [10], the authors found that the distribution of execution times was better represented by a Gumbel distribution (a heavy tail-distribution). Other examples of heavy tail distributions include all extreme values distributions (Gumbel, Fréchet and Weibull), t-student or Paretto distributions.

An important property related to heavy tail distributions is that they are (essentially) invariant under maximisation (extreme value theory): This means that, if X(1), X(2), … are independent and identically distributed with common distribution function F that is heavy-tailed and M(n) = max(X(1),…,X(n)), then $M(n)/\sqrt[\alpha]{n} \to G$, as $n \to \infty$, in which G is the Fréchet distribution. This may suggest a generalisation for extreme values, similar to the central limit theorem for means, when *n* is large enough.

Heavy-tail distributions have been object of several (recent) studies in the fields of load balancing (CPU, network), job scheduling (Web servers) and complex system studies. Particularly, there are some proposals for modelling and analysing heavy-tail distributions for estimation of rare event probabilities with computable tractable techniques [11, 12].

3.2.2 Average maximums

Derived from the previously referred methods for simulation output analysis, an intuitive approach, for trying to obtain an estimator for the worst-case value of the output variable is to pick the maximum value in the set of data from each simulation replica, instead of calculating a mean value. The problem of such approach is the assumption of a normally distributed variable, needed for the applicability of the previously mentioned methods for estimating means. A possible solution could be to group the values obtained in batches and apply the assumption of a normally distributed average over the means of each batch, in a similar way to the batch means procedure. Doing this could result in an additional statistical error introduced by this second grouping. Additionally, the results obtained in this way, would not be exactly worst-case values, but average maximums, which can be a rather different thing and, to

achieve the conditions of the central limit theorem, much more data would be necessary, most likely making this an impractical approach.

3.2.3 Rare event simulation

It is possible to view the event of a worst-case as a rare event, and the average system behaviour tends to be far apart from the worst-case. Obtaining precise estimates of such rare event probabilities using classical simulation can require prohibitively long run lengths.

A popular technique applied for the simulation of rare events is called *importance sampling*. Basically, importance sampling comprises of two different approaches. One, that attempts to modify the probability dynamics, in such a way that rare events will occur more frequently. An alternative important sampling technique is trajectory splitting, based on the assumption that there exist some well identifiable intermediate system states that occur much more often than the rare events of interest. The idea is to detect these intermediate states during simulation execution and split the simulation execution into several independent sub-trajectories, simulated from that state. Naturally, to obtain the final estimator, the results must be adjusted accordingly to the modification introduced. See [13] and references within for further information about importance sampling techniques.

Importance sampling may indeed obtain a significant reduction in the amount of observations required to obtain the same estimator precision as would be obtained in a simulation that does not use importance sampling, however, this requires a considerable amount of problem-specific knowledge from the simulation designer and how the modified distributions introduced will affect the distribution of the target events of interest. Reducing the simulation length, while simultaneously retaining the ease and flexibility of simulation is an important issue, receiving increasing attention from researchers. But, will the application of all these techniques still make simulation an appealing tool, compared to analytical approaches?

4.Conclusion

This paper has promoted the idea that simulation is a useful tool for analyzing and understanding complex systems. As the complexity of systems increases (perhaps to the point where analysis techniques will fail to be useful), simulation or a combination of simulation with other techniques may be essential. We note that simulation can be very good at modeling the middle of distributions, but there are numerous problems when trying to modeling the tails of distributions. Yet it is the tails which are the most relevant part of the distribution from the perspective of providing predictions of future correct behaviour.

We must recognize that both simulation and analysis approaches have weaknesses. This paper, therefore, poses the following research questions. What are the essential roles of simulation? How can simulation be used, in a statistically valid way? How can simulation be combined with other analysis approaches to produce accurate analysis of systems?

References

[1] George S. Fishman, Concepts and Methods in Discrete Event Digital Simulation. New York: John Wiley, 1973.

[2] K. Pawlikowski, H. D. J. Jeong, and J. S. R. Lee, "On credibility of simulation studies of telecommunication networks," IEEE Communications Magazine, vol. 40, pp. 132-139, 2002.

[3] David S. Moore and George McCabe, "From probability to Inference",in Introduction to the practice of statistics, 3rd ed: W.H. Freeman and Company, 1999, pp. 373-431.

[4] Jerry Banks, John S. II Carson, Barry L. Nelson, and David M. Nicol, Discret-Event System Simulation. Upper Saddle River: Prentice Hall, 2001.

[5] Averill M. Law and W. David Kelton, Simulation modeling and analysis, 3rd ed. New York: McGraw-Hill, 2000.

[6] Simulation Research Group, "Akaroa2©": Department of Computer Science, University of Canterbury, 2003. Web Site: http://www.cosc.canterbury.ac.nz/research/RG/net_sim/simulation_group/akaroa/about.chtml.

[7] A. Varga, "OMNeT++ Discrete Event Simulation System", v2.3, 2004. Web Site: http://www.omnetpp.org/.

[8] P. Heidelberger and P. A. W. Lewis, "QuantileEstimation in Dependent Sequences," Operations Research, vol. 31, pp. 185-209, 1984.

[9] J.-S. R. Lee, D. McNicle, and K. Pawlikowski, "Quantile Estimations in Sequential Steady-State Simulation", in proceedings of the European Simulation Multiconference (ESM'99), International Society for Computer Simulation, Warsaw, pp. 168-174, 1999.

[10] A. Burns and E. Stewart, "Predicting Computation Time for Advanced Processor Architectures", in proceedings of the 12th Euromicro Conference on Real-Time Systems (ECRTS'00), Stockholm, Sweden, pp. 89-96, 2000.

[11] D. Starobinski and M. Sidi, "Modeling and Analysis of Heavy-Tailed Distributions via Classical Telegraffic Methods," Queueing Systems (QUESTA), vol. 36, pp. 243-267, 2000.

[12] S. Asmussen, D. P. Kroese, and R. Rubinstein, "Heavy Tails, Importance Sampling and Cross-Entropy", University of Aarhus August 2003. Available online at http://mefast.sta.unipi.gr/iwap2004/Abstracts/FinalAbstracts/IWAP2004_Rubinstein.pdf.

[13] J.K. Townsend, Z. Haraszti, J.A. Freebersyser, and M. Devetsikiotis, "Simulation of rare events in communications networks," IEEE Communications Magazine, vol. 36, pp. 36-41, 1998.