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*
{npereira, emt, bbatista, lpinho}@dei.isep.ipp.pt; ianb@cs.york.ac.uk

Abstract

Modern real-time systems, with a more flexible and adaptive nature, demand approaches for timeliness evaluation based on probabilistic measures of meeting deadlines. In this context, simulation can emerge as an adequate solution to understand and analyze the timing behaviour of actual systems. However, 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 from the tail of a probability distribution (near worst-case probabilities), instead of deriving confidence on mean values. We approach this subject by considering the random nature of simulation output data. 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 is the basis for a discussion on the applicability of such approaches to derive confidence on the tail of distributions, where the worst-case is expected to be.

1. Motivation

Timeliness analysis of real-time systems is dominated by the notion of absolute temporal guarantees. In such frameworks, computational and communication loads are presumed to be bounded and known, and the worst-case (at least believed to be) conditions are assumed.

However, guaranteed approaches pose problems when applied to many complex systems, including modern distributed control systems. Worst case analysis-based formulations tend to be overwhelmed with simplifications that often lead to results which, when the results are compared to real system behaviour, are either pessimistic (where the worst case scenario considered for analysis cannot occur) or inadequate (e.g. where the probability of the bad case scenarios is low and we wish to trade performance for acceptable levels of deadline failures).

To deal with the more flexible and adaptive nature of such systems, we consider approaching the timeliness evaluation problem in a different way: instead of using a guaranteed approach, we explore the use of 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. The recent advent of fast and inexpensive computational power makes accurate modelling of the system feasible; simulation may be used to observe behaviour that is almost identical to the real system.

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 hides many traps. Special care must be taken under the penalty of obtaining results with lack of credibility [2]. In this paper, we focus our attention on appropriate *analysis* of simulation output data; that is, a valid model of the system is assumed to be already constructed and appropriate source(s) of randomness have been applied.

It is important to note that, for real-time systems, we are particularly interested in near-worst case values, which are infrequently observed (i.e. in the tail of the distribution). This is unlike the main body of statistical analysis which considers mainly mean values. In order to be able to reason about probabilistic measures of meeting deadlines, we need to be able to understand the results of simulation for maximum values and the significance of rare events.

In the following, we discuss well known approaches for estimating *distributions* from simulation output, and the confidence which can be applied to its *mean* values. This will serve the basis to discuss the applicability of these approaches to derive understanding of the tail of the distribution for consideration of *worst case* values.

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. 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 only 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 queue, with a random service time. The waiting time of 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.

One 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.

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 mean delay observed, the number of messages dropped, or the maximum response time.

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, for example, the mean delay over a communication medium. Consider that we would like to observe this delay during a defined period, because the system is shutdown or restarted after that period (e.g., a system that is disconnected at the end of a working day). This is called a *terminating simulation*.

One run of the simulation will produce one estimate for the mean message delay. The value of just one sample of a random process has little significance by itself. However, executing multiple runs of the simulation (*simulation replicas*) will provide a set of values, characterized by some distribution. It will also be IID, as we have seen. The mean of the samples is a natural estimator of the (unknown) true mean message delay.

But, how reliable is this estimate? If we make another set of simulation replicas, the result would, almost certainly, be different. Indeed, an estimate without an indication of its precision is of little value.

To derive a useful conclusion, one would have to know something about the distribution of the sample. The *central limit theorem* is a basic, but very useful and important concept which basically states that the sum and the average of many random values present a distribution close to normal distribution. Typically, a normal approximation is sufficient if about 30 or more values are used [3]. Then, well-known methods can be used to draw confidence intervals from normal distributions. However, the standard procedures for inference are intended for situations where the standard deviation for the entire population is known. As there is not usually knowledge of the entire population, it is necessary 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 true characteristics. The validity of this estimation, and of its confidence interval, is well known for the mean value of a distribution. The applicability to the tail of the distribution (the worst- or near to worst-case) will be then discussed in Section 3.

Suppose that $X_1, X_2, \cdots X_n$ 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 ($\bar{X}(n)$), is an unbiased (point) estimator of μ, and is:

$$\bar{X}(n) = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

That is, the expected value of $\bar{X}(n)$ is μ: $E_i[\bar{X}(n)] = \mu$. If we perform a great number of independent experiments, each resulting in a $\bar{X}(n)$, their average will be μ. While $\bar{X}(n)$ is the estimator of μ, in a similar way, the sample variance ($S^2(n)$) is an unbiased estimator of σ^2:

$$S^2(n) = \frac{\sum_{i=1}^{n} [X_i - \bar{X}(n)]^2}{n-1} \tag{2}$$

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:

$$\bar{X}(n) \pm t_{n-1,1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}} \tag{3}$$

The estimate ($\bar{X}(n)$, 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. 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, we note that the systems of interest will not always have a terminating event, and we are interested in analysing the system's behaviour over a long time. 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*.

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 initialisation 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. These methods have practical impairments, thus more formal methods are commonly used.

The replication approach may be used in the context of non-terminating simulations. To extract the steady-state means from each simulation replica is simply a case of deleting the warm-up data (samples 1 to l) from each replica of length m samples. Equation (4) generates IID variables which can be used to generate steady-state mean, variance and confidence values in the same way as for the terminating simulations. In the context of non-terminating simulations, this method is usually called *replication/deletion*.

$$X_i = \frac{\sum_{j=l+1}^{m} Y_{i,j}}{m-l} \quad \forall i \in \{1 \cdots n\} \tag{4}$$

In order to know when the steady-state is reached, techniques based on the assumption that the variance of the samples is substantially lower in the steady-state than in the

warm-up period, are used to *detect* when a steady-state is reached.

Another method for achieving independent samples from non-terminating simulations is to perform one long simulation run and obtain independent observations from subsets of the data. The method of *batch means* [4], is similar to the replication/deletion, except that one 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 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 significant problem is choosing the batch size m, or equivalently, the number of batches k. A number of guidelines extracted from research literature, and a general recommended strategy may be found in [4].

Other methods based on one long simulation are encountered [5], such as the *autoregressive* method or *spectral analysis*, which 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. 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. We note that [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.

It can be argued [5] that no procedure that fixes the run length before the simulation begins will always produce a satisfactory confidence interval. A *sequential* procedure where the simulation's end is determined by a relative statistical error that is verified in consecutive checkpoints is a more viable approach. Sequential procedures are recognised as a practical approach allowing control on the error of the final results of stochastic simulations [2].

Sequential procedures are not without problems. They are more complex, requiring computing the estimates at several points of the simulation to check if the stopping rule has been satisfied. The approach may not be easily applicable when multiple measures of performance are needed. Finally, because of random nature of simulation, the relative stopping rule can be accidentally satisfied, resulting in premature termination of the simulation, and wrong results.

Currently, sequential procedures are not well supported by existing software packages. A simulation package supporting sequential procedures is Akaroa2, designed at 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].

3. What about Worst-Case Assessment?

All the previous methods seek to obtain a mean value for the output point estimator. We noted that other measures about worst-case values are required for analysis of real-time systems. For example: the probability that a queue length is greater than k messages. In this section, we briefly discuss how to extract measures of performance such as proportions, probabilities and 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 delay 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.

If Y is the original steady-state random variable obtained from simulation and B is a set of real numbers smaller than x then this is special case of estimating the mean, by letting the random variable Z be:

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

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

However, a performance measure that does not fit in the same reasoning is a quantile. For instance, if the variable represents the delay then 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. For the steady-state estimation of a 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 requires 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 that do not require storage have been proposed. In [9], a number of such approaches are presented and evaluated.

3.2. 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 the other extreme values distributions (Gumbel, Fréchet and Weibull), *t*-student and Paretto distributions.

An important property of heavy-tailed distributions is that they are (essentially) invariant under maximisation, tending to a (Fréchet) distribution. This may suggest a generalisation for extreme values, similar to the central limit theorem for means.

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 proposals for modelling and analysing heavy-tail distributions for estimation of rare event probabilities with computable tractable techniques [11, 12].

3.3. 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 this 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, 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.

3.4. Rare event simulation

It is possible to view the occurrences or near worst-case scenarios as *rare event* (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. 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 analysing 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 modelling the middle of distributions, but there are numerous problems when trying to modelling 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 Comm.*, vol. 40, pp.132-139, 2002.

[3] David S. Moore and George McCabe. From probability to Inference, an 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(c)": 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. Quantile Estimation 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)*, Warsaw, pp. 168-174, 1999.

[10] S. Edgar and A. Burns. Statistical Analysis of WCET for Scheduling. In *Proc. of 22nd IEEE Real-Time Systems Symposium. (RTSS'01)*, London, UK, pp. 215-224, 2001.

[11] D. Starobinski and M. Sidi. Modeling and Analysis of Heavy-Tailed Distributions via Classical Telegraffi c Methods. *QUESTA*, vol. 36, pp. 243-267, 2000.

[12] S. Asmussen, D. P. Kroese, and R. Rubinstein. Heavy Tails, Importance Sampling and Cross-Entropy. Uni. of Aarhus August 2003. `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 Comm*, vol.36, pp.36-41, 1998.