Availability Confidence Intervals from Bootstrap Sampling

Authored by:

Kyle Kolsti, PhD

Sean Tomlin

14 Jan 2020

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary 2
Introduction 2
Role in the STAT Process 2
Definitions 2
Uncertainty 4
Motivating Scenario 4
Proposed Method 5
Inputs 5
Monte Carlo Simulation Procedure 6
MTBF 6
MTTR 7
MLDT 8
Availability Distribution 8
Example 8
Conclusion 13
References 13

Executive Summary

A method is proposed to calculate confidence intervals for operational availability. The method is based on Monte Carlo simulation and utilizes bootstrap sampling of failure and repair times. The mean time to repair (MTTR) is estimated using a biased predictor to account for the typical right skew of repair time distributions. The mean logistics delay time (MLDT) is assumed to be normally distributed where the mean and standard deviation of the mean are estimated from logistics studies. The method was designed to be robust to a variety of scenarios, meaning the user does not have to identify or assume any underlying population distributions. An example is used to demonstrate application of the method.

Keywords: availability, reliability, bootstrap

Introduction

Role in the STAT Process

Scientific Test and Analysis Techniques (STAT) are "deliberate, methodical processes and procedures that seek to relate requirements to analysis in order to inform better decision-making" (Guide to Developing an Effective STAT Test Strategy V7.0). The STAT Center of Excellence (STAT COE) utilizes a construct called the STAT Process to provide a structure for applying STAT to test and evaluation programs. The STAT Process is an iterative process that begins with requirements and ends at program decisions. In between there are four phases: Plan, Design, Execute, and Analyze.

The method introduced in this paper will be most relevant to the Analyze phase. An availability requirement would have driven appropriate test planning, design, and execution to obtain sufficient reliability data for analysis. This method is then applied to the test data to inform decisions. This method could potentially be used to aid in the Design phase, perhaps through the use of simulation; however, that application is outside the scope of this document.

Definitions

Operational availability is a metric that can be useful when applied along with a suitable reliability metric. Some definitions are provided to lay the groundwork for the rest of the paper. First, from the DOD Reliability, Availability, Maintainability, and Cost Rationale Report Manual (DOD RAM-C):

> **Reliability**: "*Reliability measures the probability that the system will perform without failure over a specified interval under specified conditions…Considerations of reliability must support both availability metrics [i.e., materiel and operational as described below]. Reliability may be expressed initially as a desired failure-free interval that can be converted to a failure frequency for use as a requirement*."

Reliability can be stated in terms of a probability (of functioning without failure for a given length of time) or as a time (the average failure-free interval). Both perspectives are equivalent and can be

calculated from each other. Many acquisition programs and this Best Practice use the failure-free interval formulation where the metric is mean time between failures (MTBF).

Reliability testing is necessary to provide the data for calculation of availability because availability is a derived metric for which there is no direct test. There are two types of availability: materiel (A_m) and operational (A_o). This paper will focus entirely on operational availability, but for completeness both are defined below from the DOD RAM-C Manual (2009). A key distinction is that operational availability uses "active" time. In other words, the A_o metric counts only for the time a unit is actively in service, not when it's in storage, used as a spare, in transport, or the like.

- **Materiel Availability**. "*Materiel Availability is a measure of the percentage of the total inventory of a system operationally capable (ready for tasking) of performing an assigned mission at a given time, based on materiel condition. This measure can be expressed mathematically as number of operational end items [divided by the] total population*."

- **Operational Availability**. "*Operational Availability indicates the percentage of time that a system or group of systems within a unit are operationally capable of performing an assigned mission and can be expressed as (uptime/(uptime + downtime))*."

While uptime is represented by MTBF, the downtime is typically broken into two components: mean time to repair (MTTR) and mean logistics delay time (MLDT). The metrics MTBF, MTTR, and MLDT are the average times over the useful lifetime of the system. When a system takes a long time to repair, its availability will suffer. Times to repair are random according to some distribution over the system's lifetime, but the availability depends on the average time to repair the system which is called MTTR.

MLDT is the measure of how long it takes the maintainers to acquire the resources needed to perform the repair. A common example is when something breaks and the maintainers are ready to perform the repair, but the parts, tools, people, or other necessary resources are not on hand. The time spent waiting to begin repairs is represented by MLDT. Since MLDT is dependent on logistics organizations and supply processes which might not be considered part of the system under test (SUT), different programs may choose to include it or not in the availability calculations. The STAT COE recommends A_o be reported with several possible values for MLDT to inform decision makers about the intrinsic system behavior and the less-than-ideal availability in the deployed environment.

Assuming the MTBF, MTTR, and MLDT estimates accurately predict the respective mean times over the entire life of the system, the point prediction for operational availability is

$$A_o = \frac{MTBF}{MTBF + MTTR + MLDT} \tag{1}$$

An important concept is that the availability metric may be misleading when used in isolation. For example, consider two systems with $A_o = 90\%$. Over its lifetime the first system can be expected to go

about 9 days before breaking and it takes one day on average to fix it – the reliability is $MTBF = 9$ days and the $MTTR = 1$ day. The second system has $MTBF = 9$ months and $MTTR = 1$ month. Those two profiles are drastically different from the user's point of view, even though the operationally availabilities are identical. The operationally availability metric by itself provides no information about the probability of completing a mission of a certain time duration. Therefore, it is generally recommended when drafting requirements that if operational availability is stipulated, it should be accompanied by a reliability requirement such as MTBF and a logistics requirement such as MLDT as appropriate.

Uncertainty

The test team should strive to report not only the A_o point estimate, but also the associated uncertainty. A confidence interval about A_o is one way to do this. Recall the proper interpretation of a confidence interval: given a method to compute a confidence interval that claims the confidence to be 90%, if the test were re-accomplished many times and the method were used to calculate the new confidence interval each time, the true availability value would lie within approximately 90% of the confidence intervals. This statement is the basis for evaluating the proposed method later in this paper.

To capture uncertainty in availability, the uncertainties for the constituent terms MTBF, MTTR, and MLDT must be estimated. Notable methods in the literature for calculating the confidence interval for A_o assume certain distributions as shown in Table 1. The distributions are independent – times between failure are not affected by times to repair and vice versa. None of these methods include MLDT.

Table 1: Existing methods in the literature and their assumed population distributions

Method	Times between Failure	Times to Repair
Keesee (1965)	Exponential	Exponential
Masters and Lewis (1987)	Gamma	Lognormal
Masters et al (1991)	Weibull	Lognormal
Ananda (2003)	Exponential	Lognormal
Ananda and Gamage (2004)	Weibull, Gamma, Lognormal	Lognormal

Numerical evaluation has shown that these methods generally perform well when their distributional assumptions are met. However, when the distributional assumptions are not met, the results suffer to some degree – the intervals are too wide or not wide enough on average, causing the actual confidence of a method (also called its coverage) to deviate significantly from its stated confidence. Unfortunately, in a real-world test program without extensive prior information, it's unlikely there will be sufficient reliability data to distinguish the population's true distribution. If it's not feasible to confirm the assumptions were met, the test team may apply an unsuitable method and report misleading results.

Motivating Scenario

Consider the acquisition of a system for which little to no reliability information exists. The users have an urgent need for the system's promised capabilities, so they have coordinated a compressed acquisition

timeline with the program and decision makers. The test and evaluation (T&E) portion of the timeline is based on verifying performance requirements rather than reliability test planning (i.e., system effectiveness rather than suitability). Despite the short test duration, the decision makers want predictions of reliability and availability – times to failure and times to repair will be recorded. Logistic delay times will be predicted by logisticians in a study because the logistical system has not been stood up yet. The logisticians can only provide a point estimate, but the test team is able to ascertain through communication with the logisticians how certain they are of their prediction. Ultimately the decision makers are willing to accept a high degree of risk of accepting a system that does not meet RAM requirements; however, they do want some idea of how much risk they would be accepting.

In response to this type of scenario, the confidence interval method proposed in this paper was developed to provide a robust procedure for quantifying uncertainty that is suitable at very small sample sizes. Robust in this context means the method must be suitable across common combinations of distributions without the analysts having to decide or determine which one is truly the case.

Proposed Method

Inputs

The method requires information about failures, repairs, and logistics in order to provide confidence limits on A_o. The data required by this method are shown in Table 2. The failure and repair times may be observed from any number of units in operation. As mentioned in the motivating scenario, it is expected that the test team will only receive a point prediction for MLDT which we will label as $\bar{d}$. One way to incorporate uncertainty is to communicate with the logisticians to translate their certainty into a standard deviation of the mean, defined here as $\sigma_{\bar{d}}$. For example, if the logisticians were to agree there's a 95% chance the true MLDT is in the range $[\bar{d}_{low}, \bar{d}_{high}]$, assuming a normal distribution where about 95% of the population is within two standard deviations results in an estimate of $\sigma_{\bar{d}} \cong (\bar{d}_{high} - \bar{d}_{low})/4$.

Table 2: Data required

Constituent	Source	Data
Times between failure	Test data	n failure times, x_i
Times to Repair	Test data and/or Maintenance Study	m repair times, y_j
Logistics Delay Times	Logistics Analysis	Mean and standard deviation of MLDT, $\bar{d}$ and $\sigma_{\bar{d}}$ respectively

The analyst must also decide on the type of confidence interval to report (upper 1-sided, lower 1-sided, or 2-sided; typically the upper 1-sided is of highest concern for accepting a system) and the desired level of confidence, typically 80%, 90%, or 95%. This decision should be agreed upon by the stakeholders based on the risk profile and information needed to make decisions.

Monte Carlo Simulation Procedure

The method proposed here is based on Monte Carlo simulation. One simulation consists of a large number of runs, N. For each run, the algorithm generates a random value for MTBF, MTTR, and MLDT based on the information provided by the user (described below). From these values, the availability for that run is calculated according to Equation 1. At the conclusion of the simulation, there will be a set containing N values of availability which can be used to make inferences, including the expected value (the arithmetic mean) and confidence intervals.

The following paragraphs describe the procedures for generating MTBF, MTTR, and MLDT that are used to calculate availability in a given run.

MTBF

The method uses one of two procedures for calculating MTBF, the choice of which depends on how many failures were observed during testing.

If there are five or more times between failure, MTBF is generated using the bootstrap sampling procedure with replacement (Efron 1979). The reason bootstrap is selected is to avoid assuming a distribution for the failure time population. Bootstrap sampling, in summary, is described as follows: Given a sample of times between failure from the test, $x = \{x_1, x_2, \dots, x_n\}$, one of the values x_i is randomly selected and called x_1^*. This random selection is performed n times until a bootstrap sample has been created: $x^* = \{x_1^*, x_2^*, \dots, x_n^*\}$. The bootstrap sample x^* consists entirely of times between failure that are present in the original test data x. Some values of x may have been selected numerous times; others may have not been selected at all. After the bootstrap sample has been created, the MTBF for the simulation run is calculated to be the average of the bootstrap sample.

$$MTBF = \bar{x}^* = \frac{1}{n}\sum x_i^*$$

For $1 \leq n \leq 4$, the population of times between failure is assumed to be exponential because the bootstrap procedure doesn't have enough possible values from which to choose in order to emulate the population. Given an exponential population of times between failure, the following ratio follows a Chi-Square distribution. The value T is the total test time, which is the sum of the observed times between failure.

$$\frac{2T}{MTBF} \sim \chi^2(2n)$$

Therefore, after sampling a random number a from the distribution $\chi^2(2n)$, the value of MTBF used for this run is

$$MTBF = \frac{2T}{a}$$

MTTR

An MTTR value is generated for each Monte Carlo run using bootstrap with replacement as described for MTBF. To estimate the MTTR from a sample of m times to repair, $y^* = \{y_1^*, y_2^*, \dots, y_m^*\}$, it is simple to use the average as was done earlier for MTBF,

$$\bar{y}^* = \frac{1}{m}\sum y_i^*$$

Unfortunately the arithmetic average tends to underestimate the mean of populations that are skewed to the right, as times to repair often are. The right tail of the times to repair distribution has a significant impact on the lower confidence limit of availability. Therefore, for this method a biased estimator for the mean was chosen that is based on the lognormal distribution (O'Hagan, 2003). A biased estimate was not employed for times between failure because the primary concern for accepting a system lies with the lower confidence bounds (Ananda 2003). To estimate the population MTTR from the sample of m times to repair y^*, first transform the times to repair using the natural logarithm,

$$Y_i = \ln(y_i^*)$$

Then the mean and standard deviations of the logarithmic times to repair respectively are

$$\bar{Y} = \frac{1}{m}\sum Y_i$$

$$s^2 = \frac{1}{m-1}\sum (Y_i - \bar{Y})^2$$

The variables $\bar{Y}$ and s are the estimates of the location and shape parameters of the population's lognormal distribution, μ and σ. Finally, MTTR is estimated using the formula

$$MTTR \equiv \tilde{y} = \exp\left(\bar{Y} + \frac{s^2}{2m}\right)\left(1 - \frac{s^2}{m-1}\right)^{-(m-1)/2}$$

Note that imaginary numbers result if $s^2 > m - 1$ and m is even. As desired, the biased estimate $\tilde{y}$ tends to be higher than the average, $\bar{y}$. This behavior is more pronounced with small sample sizes and higher standard deviations as indicated by higher values for σ. The two estimators tend to provide nearly equal estimates of the mean for data sets with low standard deviations ($s \leq 0.2$), where the data indicate the population is nearly normal.

The difference between the two estimators $\tilde{y}$ and $\bar{y}$ can be unrealistically large for some samples, such as those with one extremely high observation. This kind of sample occurs more often with smaller sample sizes and more highly skewed populations. Therefore, for the code used in this paper, a limiter was put in place that limited s^2 to a maximum value of 2.25. This limit corresponds to an estimated shape parameter of $s = 1.5$, which avoids the numerical difficulties, but is sufficiently large for the range of distributions to cover most plausible situations.

MLDT

The MLDT is treated differently than MTBF and MTTR. Logistics information is likely to be provided from a study rather than through test observations. It is also likely that only a point estimate of MLDT will be provided and the test team will have to estimate the spread of the data if uncertainty is to be accounted for. In this case, the user of this method can provide the current state of knowledge of MLDT by entering the expected value, $\bar{d}$, and the standard deviation of the expected value, $\sigma_{\bar{d}}$. Given this information, for each simulation run, a value for MLDT may be sampled from a selected distribution. If MLDT were assumed to be normally distributed, the values would be randomly sampled from a normal distribution with mean $\bar{d}$ and standard deviation $\sigma_{\bar{d}}$.

Availability Distribution

Finally, the availability for each run is calculated using the randomly generated constituent values,

$$A_o = \frac{MTBF}{MTBF + MTTR + MLDT}$$

The resulting set of N values for A_o may be used to make inferences about the system's availability. Statistics such as the mean, median, percentiles, and so forth may be used as with any other sample. In this method, the percentiles are used directly to determine the confidence interval. For example, the 80% upper confidence interval is defined by the lower confidence limit, which is the 20th percentile of the set of A_o values.

Example

Application of the proposed method is demonstrated using an example from a journal article. This method was coded in the programming language R to receive the necessary input data and produce plots with accompanying numerical results. This code was written to report only the lower confidence limit for a one-sided upper confidence interval. For access to the web-hosted tool based on this code, navigate to the tool section of the STAT COE web site www.afit.edu/STAT or contact the STAT COE for assistance.

The example comes from the journal article "Confidence Intervals for Steady State Availability of a System with Exponential Operating Time and Lognormal Repair Time" in which the author proposed a method for determining the lower confidence limit for availability assuming exponential times between failure and lognormal times to repair (Ananda, 2003). The times between failure were randomly generated from an exponential distribution with $\theta = 100$ where the parameter θ is the mean failure time, or MTBF. The times to repair were randomly generated from a lognormal distribution with parameters $\mu = 1.0$ and $\sigma = 1.0$ (the mean and standard deviation of the natural logarithms of the times to repair, respectively). From these distributions the true A_o of the system can be calculated as 95.7%. The sample size was $n = m = 10$ with the data shown below:

- Times between failure: 75.69, 46.50, 393.30, 476.17, 15.76, 340.92, 21.20, 14.06, 33.24, 2.83
- Times to repair: 3.69, 1.22, 0.43, 3.14, 4.59, 2.96, 12.11, 3.06, 1.45, 2.20

From this data, the journal article's method produced an LCL of 93.39%. To demonstrate the method in this Best Practice, the same data set was entered into the R code-based tool as shown in Figure 1.

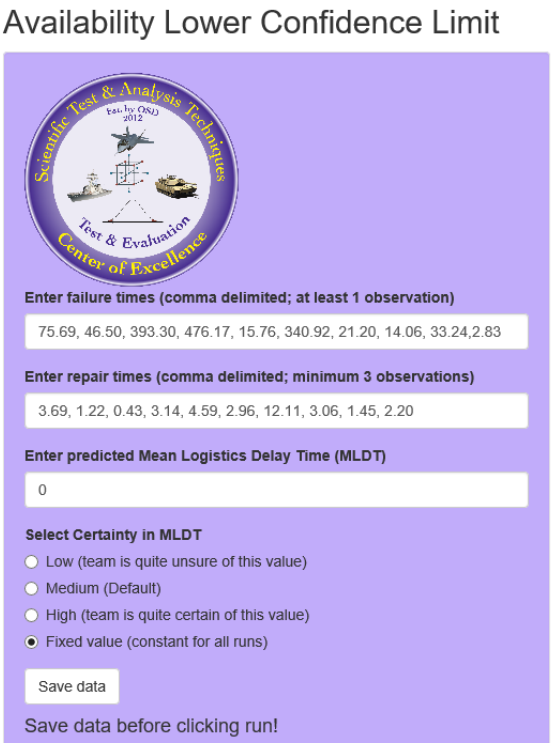

Figure 1: Tool input section.

Since the journal article's method does not include MLDT, the value for MLDT was set to zero. After clicking the "Save data" button, the tool echoed the data back to the user for quality verification as shown in Figure 2.

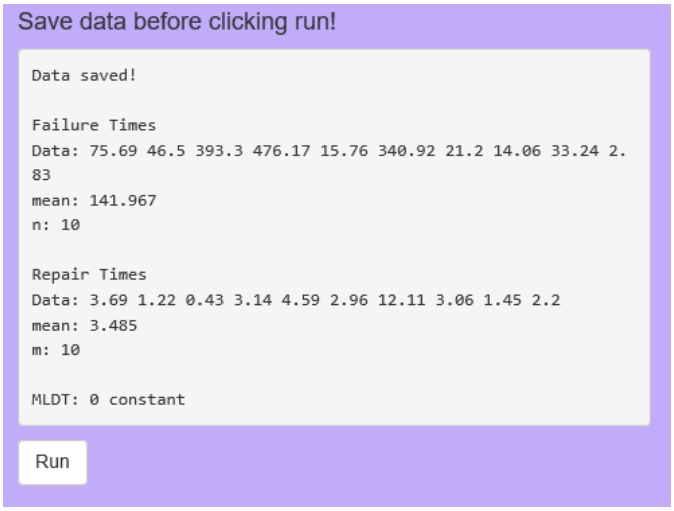

Figure 2: Tool data input confirmation section.

After clicking the "Run" button and waiting a few seconds for the tool to perform the large number of simulation runs (10,000), the results appear on the "Analysis" tab as shown in Figures 3 through 8. Figure 3 depicts how the tool documents the input data and the algorithm's settings. It also displays the availability point prediction of 96.8%, which is 1.1% higher than the true value. In contrast, consider calculating availability from the raw data only. According to the tool, the mean times predicted from the samples are simply the sample averages, $MTBF = 141.967$ and $MTTR = 3.485$. Plugging these values into Equation 1 results in an availability estimate of 97.6%, which is 1.9% higher than the true value. For this particular example, the proposed method provides an estimate for availability which is closer to the true value.

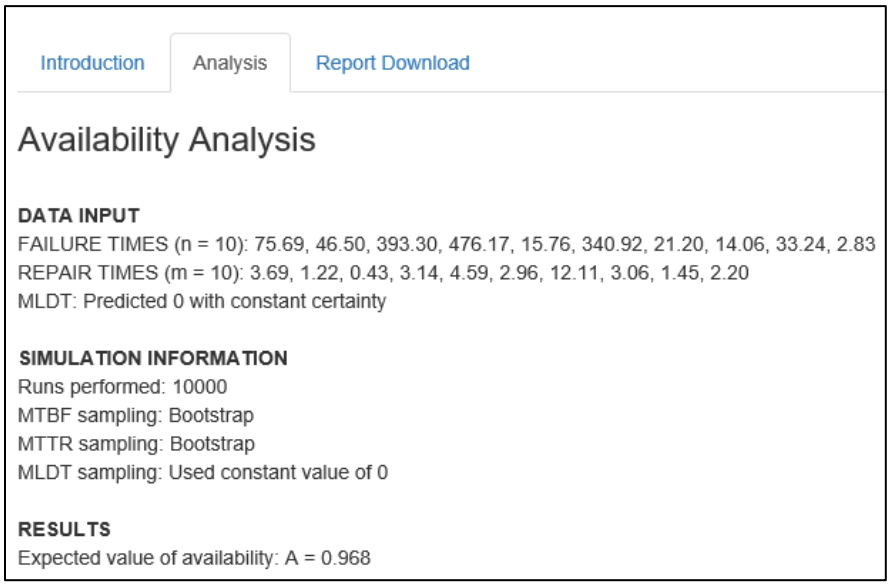

Figure 3: Tool analysis tab – simulation statistics and numerical results.

Figure 4 depicts the confidence level results. A table shows the LCL for a variety of confidence levels. Each LCL is simply the appropriate percentile of the 10,000 availability values generated by the method's simulation. For example, the 95% LCL is the 5^{th} percentile value of the 10,000 availability values. From this table the 95% LCL was 93.0%, which is within 0.4% of the journal article method's LCL of 93.39%. This comparison demonstrates the method proposed here provided a valid result for this example's data. Note that because of the random process used in the simulation, the tool may produce a slightly different result from run to run. If desired, the analyst can run the tool multiple times to obtain a mean and standard deviation for the LCL. Numerical evaluations performed by the STAT COE indicate this method remains valid for a variety of underlying distributions, and is therefore robust without the user having to make assumptions about them. The evaluation results are outside the scope of this Best Practice – for more information please contact the STAT COE.

Below the table of LCLs there is a slider that permits the user to select any confidence level from 80% to 99%. The slider value is used in the statements and in the availability plots of Figure 5 and 6. The statements were designed to provide the numerical results in technically accurate terms that could be copied and pasted into a report.

Figure 4: Tool analysis tab –Lower Confidence Limit table and statements.

Figures 5 through 8 show the plots generated by the tool to enhance the user's understanding of the results. The cumulative distribution function (CDF) in Figure 5 enables the analyst to graphically determine any percentile value of interest. Figure 6 depicts the probability distribution function (PDF) of availability. The vertical red line in Figures 5 and 6 is the LCL as set by user using the slider.

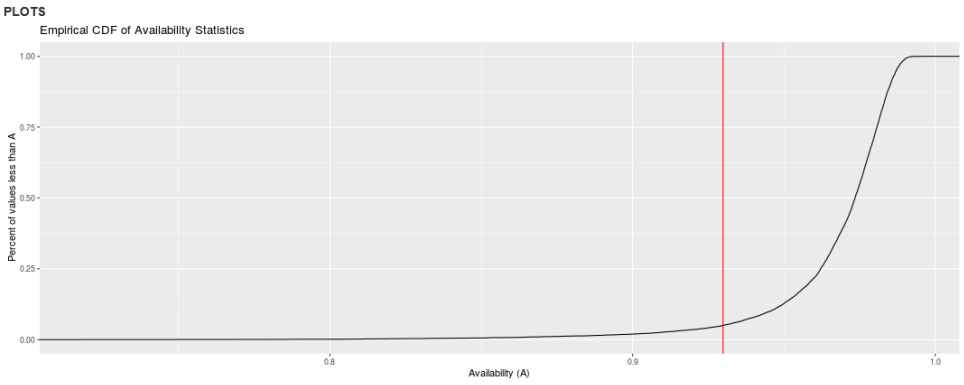

Figure 5: Tool analysis tab –Availability cumulative distribution function (CDF).

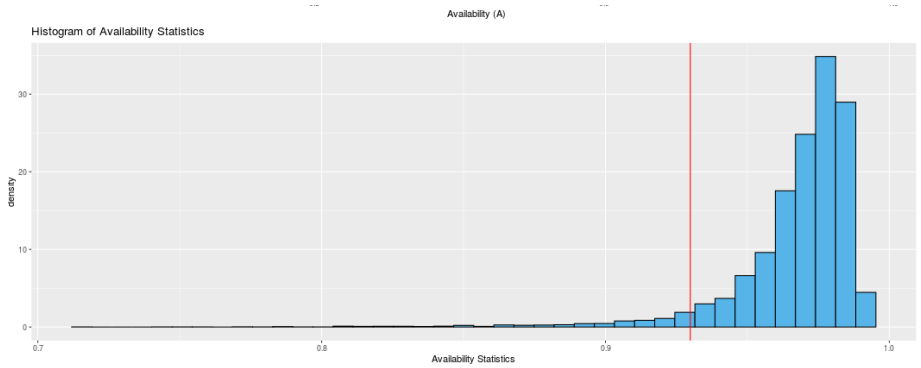

Figure 6: Tool analysis tab –Availability histogram.

Finally, Figures 7 and 8 show the histograms of the 10,000 simulation MTBFs and MTTRs to provide additional insight. The multiple peaks in the histograms are an artifact of the bootstrap sampling procedure as there is a finite combination of observations that can appear in any given sample. It should be emphasized that the plotted distributions are of the mean values, not the times between failure and times to repair themselves. Therefore as the sample sizes become very large, the Central Limit Theorem says the MTBF and MTTR distributions will each approach a normal distribution, even in this example where the failure and times to repair come from exponential and lognormal populations.

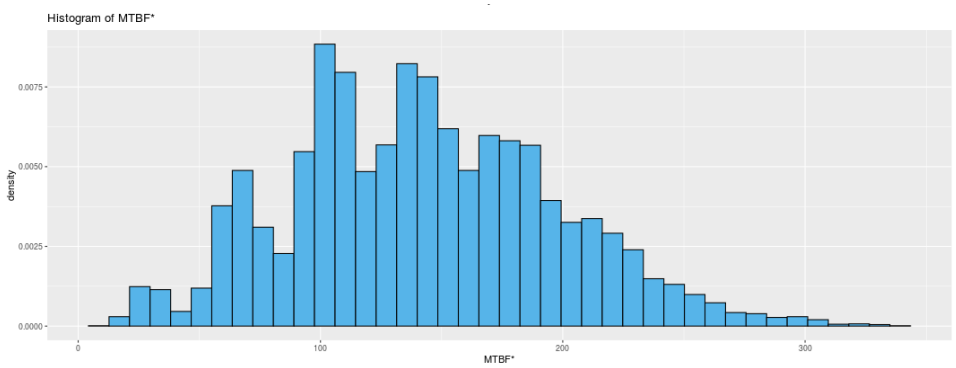

Figure 7: Tool analysis tab –MTBF histogram.

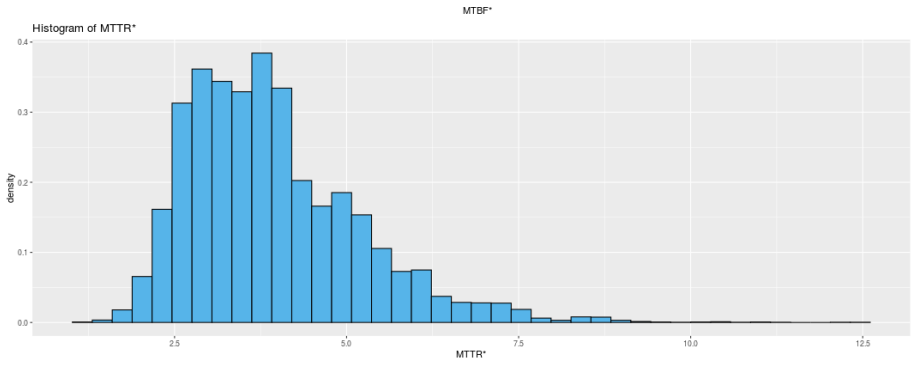

Figure 8: Tool analysis tab –MTTR histogram.

Conclusion

This Best Practice proposed a method for calculating confidence intervals for operational availability. An example using a STAT COE-developed tool demonstrated the validity of the results by comparison to a method in the literature. The method is robust to a variety of underlying population distributions of interest, meaning it produces suitable results without the user having to declare which distributions to assume. Other known methods for calculating availability confidence intervals can provide inaccurate results if the underlying distributional assumptions are violated. With small sample sizes it is difficult to validate the distributional assumptions, so the analyst may not know the risks involved in applying a certain method.

After a test is concluded, this method produces an empirical distribution of availability results from the simulations. The distribution can be used to answer questions like "What is the probability that the availability is less than the threshold?" for a decision maker who is balancing the risks.

References

Ananda, M.M.A., "Confidence Intervals for Steady State Availability of a System with Exponential Operating Time and Lognormal Repair Time," *Applied Mathematics and Computations*, Vol. 137, Issues 2-3, 2003, pp. 499-509., doi:10.1016/S0096-3003(02)00155-8.

Ananda, M.M.A. and Gamage, J., "On Steady State Availability of a System with Lognormal Repair Time," *Applied Mathematics and Computations*, Vol. 150, 2004, pp. 409-416., doi:10.1016/S0096-3003(03)00281-9.

Burke, Sarah et al. "Guide to Developing an Effective STAT Test Strategy V7.0," Scientific Test and Analysis Techniques Center of Excellence (STAT COE), Dec 2019.

Department of Defense Reliability, Availability, Maintainability, and Cost Rationale Report Manual. 2009. Washington, DC: Office of the Secretary of Defense.

Efron, B., "Bootstrap methods: Another look at the jackknife", The Annals of Statistics, Vol. 7, No. 1, 1979, pp. 1–26., doi:10.1214/aos/1176344552.

Keesee, W. R., "A Method of Determining a Confidence Interval for Availability," U.S. Naval Missile Center Misc. Publication No. NMC-MP-65-8, 1965., doi:10.21236/ad0617716.

Masters, B. N. and Lewis, T. O., "A Note on the Confidence Interval for the Availability Ratio," Microelectronics Reliability, Vol. 27, No. 3, 1987, pp. 487-492., doi:10.1016/0026-2714(87)90467-7.

Masters, B. N., Lewis, T. O., and Kolarik, W.J., "A Confidence Interval for the Availability Ratio for Systems with Weibull Operating Time and Lognormal Repair Time," Microelectronics Reliability, Vol. 32, Issues 1-2, 1992, pp. 89-99., doi:10.1016/0026-2714(92)90089-4.

O'Hagan and Stevens, "Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality?", *Health Economics,* Vol 12, Issue 1, 2003, pp. 33-49., doi:10.1002/hec.699