NPS-TE-17-011

ACQUISITION RESEARCH PROGRAM SPONSORED REPORT SERIES

Allocating a Data Collection Budget to Support Test and Evaluation Plans

14 November 2016

Dashi I. Singham

Graduate School of Business & Public Policy

Naval Postgraduate School

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Abstract

Acquisition decisions are often made using experimental and simulation data. It can be unclear how much to budget to collect such data. This research delivers guidelines for determining the amount of data needed to evaluate uncertain performance of a system using different sampling techniques. These methods can be applied to help determine the budget for performing test and evaluation on new systems under uncertain conditions. We implement these sampling techniques to develop a tool that can be used to analyze different experimental conditions to estimate the approximate testing budget.

THIS PAGE LEFT INTENTIONALLY BLANK

NPS-TE-17-011

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Allocating a Data Collection Budget to Support Test and Evaluation Plans

14 November 2016

Dashi I. Singham

Graduate School of Business & Public Policy

Naval Postgraduate School

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

THIS PAGE LEFT INTENTIONALLY BLANK

Table of Contents

THIS PAGE LEFT INTENTIONALLY BLANK

LIST OF FIGURES

THIS PAGE LEFT INTENTIONALLY BLANK

LIST OF TABLES

THIS PAGE LEFT INTENTIONALLY BLANK

Confidence Intervals for Data Collection

Introduction

Data collection, either through physical experimentation or modeling and simulation, is often a critical part of any acquisitions decision. In all situations, the decision-maker must decide how much data to collect. A number of factors influence this decision, including the cost of data collection, the expected variation in the data, the level of risk associated with the project, and the desired precision in the results needed to make a decision. If the decision on how much data to collect is not made carefully, the allocated budget could fall short of what is required to evaluate the potential acquisition. Alternatively, if too much effort is invested in data collection beyond what would be required to make a decision, then time and money have been wasted.

This research studies methods for determining the data collection effort (also known as *the sample size effort*) to evaluate the trade-offs in the input factors and simplify the process needed to make this decision. In particular, we focus on a Test and Evaluation (T&E) setting to motivate the use of sampling methods. We describe the factors that are needed to determine the sample size effort, and create an Excel tool where the decision-maker can enter their parameter values to estimate their proposed sampling effort and the associated risk in the output. This tool can help make the decision for how much data collection is needed. Additionally, a sequential sampling procedure can be implemented so that as samples are collected, we can estimate how many more samples might be needed to obtain a narrow confidence interval.

We attempt to enable better understanding of statistical methods required to obtain valid output results through sampling to support an acquisitions decision, and these methods could be directly implemented in a T&E environment. Sequential sampling rules are a part of many T&E studies, and we describe the validity of such rules in the face of recent theoretical research in this area.

We state our two research objectives:

- Define explicitly how the sample size effort should be determined within the T&E environment and how it can be integrated into existing frameworks.
- Develop the foundations to support sequential sample decisions and deliver the results to enable the T&E analyst to evaluate their risk and determine their data collection budget using a spreadsheet tool.

There are two components pertaining to the objectives above. The first component is to define how the sample size decision fits within a T&E framework. Modeling and simulation can be an important part of augmenting T&E analysis when there are limits on physical experimentation, though it should not replace operational testing (U.S. Marine Corps, 2013). The Simulation, Test and Evaluation Process (STEP) is a way of integrating simulation with the test process (U.S. Department of Defense, 2005). Understanding Integrated Testing and Evaluation is critical to efficient implementation of modeling and simulation results (U.S. Marine Corps, 2010). In some cases, simulation is available to estimate performance in the Developmental T&E (DT&E) phase, and collecting performance estimates using computer models is cheap. In Operational T&E (OT&E), experimentation is usually much more costly, and testing under operationally realistic conditions is critical to the final evaluation of the system.

We first provide background on the different types of sampling rules. There are fixed sampling rules that determine the number of samples to be collected ahead of time. Sequential sampling can be used when we are unsure of the number of samples needed in order to obtain some measure of accuracy on the mean performance of the system. We check our observations as data arrives and determine whether we need to continue sampling. We describe each of the parameters and inputs needed to determine fixed and sequential sampling rules, and analyze the trade-offs. Next, we present how these rules can be used in a decision-based context, bridging the gap toward applications in T&E. We describe how these concepts can be used in T&E specific contexts. Finally, we describe the spreadsheet to implement the sampling rules discussed.

Fixed and Sequential Sampling

Sequential sampling is useful both when samples are expensive and when they are cheap. If collecting samples is cheap, sequential sampling can be used to determine when to stop collecting samples and aggregate results. In a simulation context where generating replications on the computer is relatively cheap (except possibly in terms of time), sequential sampling rules help determine an appropriate time to stop.

In T&E, collecting samples is usually extremely expensive, and there may be a limited budget to spend on testing. In this case, sequential sampling is critical because stopping as early as possible could lead to potential cost savings. It may be unclear ahead of time how many test samples are needed, and it is better not to spend more than necessary on tests. Any saved resources from first-stage tests in DT&E could be used later in OT&E.

There are two main ways of evaluating mean system performance. The first is through taking the average of performance estimates and calculating a confidence interval for the mean. The second is to calculate the probability of success, which is the mean of a binary success/failure response. We first focus on confidence intervals and later describe methods for estimating the probability of success of the system.

We define the following parameters:

- n is the number of samples collected.
- $\bar{x}$ is the sample mean of n samples collected.
- μ is the true mean performance of the system, which is unknown.
- σ is the square root of the variance (standard deviation) of the data. This is usually unknown.
- s is the estimate of the standard deviation of the data.
- η is the confidence coefficient desired in the result, which is usually 90%, 95%, or 99%.
- α is $1-\eta$, and is the Type I error associated with the test.
- $t_{\alpha,n-1}$ is the t-value associated with the t-distribution with $n-1$ degrees of freedom with tail probability α/2.
- z_{α} is the z-value associated with the standard normal distribution with tail probability α/2.
- δ is the desired precision (half-width) of the resulting confidence interval.

A confidence interval under the assumption of normality and known variance in the collected data takes the following form:

$$\left[\bar{x} \pm z_{\alpha,n-1}\frac{\sigma}{\sqrt{n}}\right].$$

Because the variance is usually unknown, we suggest updating the estimate of the variance after each data point is collected using the variance estimate formula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

If the data is normally distributed and the variance is estimated, then the confidence interval takes the following form:

$$\left[\bar{x} \pm t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right].$$

If the assumption of normality in the data is met, repeated collections of this interval from sampled data will result in *(1-α)*100% of the intervals correctly including the true mean of the data *μ*. The goal is usually to estimate mean performance, though in T&E other metrics may be of interest, like the variance, quantiles, or the minimum/maximum performance values.

Fixed Sampling Rules

If we have an estimate of the standard deviation *s*, and have some desired upper bound on the precision in our confidence interval *δ*, then we can choose the smallest sample size such that

$$n \geq \left(\frac{t_{\alpha,n-1} \cdot s}{\delta}\right)^2$$

and this sample size will yield a confidence interval for *μ* that has a half-width smaller than *δ*. This is often the best first step to estimating the sample size needed and can help determine whether the budget will allow for the given level of precision. Thus, a high variance and a small precision level will require a large number of samples. Decreasing the Type I error will also increase the sample size needed.

There exists much research to quantify and optimize the sampling rule decision in a simulation framework, where each simulation replication is relatively inexpensive (Singham & Schruben, 2012; Singham, 2014). In a T&E framework, we need to consider that collecting performance sample estimates, either through physical experimentation or simulation modeling, has a real cost that must be considered. This cost results in a trade-off between the statistical precision in the results that can be obtained, and the budget needed to execute the experiment.

As an example, consider the following situation where the quality of an experiment is judged by the length of the half-width of a confidence interval that can be generated, where the confidence interval with the corresponding half-width is defined according to the usual formula as

$$\left[\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

The sample mean of the observations is the center-point, the standard deviation is σ, *n* is the sample size, and *z* is the usual quantile of the normal distribution. If a higher half-width is worse and corresponds to a higher cost in terms of uncertainty, and we have a proportional cost *c* for each sample that needs to be collected relative to the performance measure, we can create a rescaled objective function for total cost:

$$cost = z_\alpha \frac{\sigma}{\sqrt{n}} + cn.$$

Taking the derivative with respect to *n* and setting it equal to zero yields an optimal value *n** that minimizes the cost objective function:

$$optimal\ sample\ size = n^* = \left(\frac{z_\alpha \sigma}{2c}\right)^{\frac{2}{3}}.$$

This tells us the optimal number of samples to help us make a decision. We can use this formula under assumptions of normality in the data and a known variance *σ*. The difficulty is in estimating the cost *c* associated with each sample. In T&E, the cost can be quite high, and higher costs mean a smaller optimal sample size.

Sequential Sampling Rules

Sequential sampling rules allow for the tester to stop when a given criteria has been met. The criteria usually involve some precision around an estimate of a metric to ensure that we have a reasonable value for the mean performance of the system. The most common example is to stop when we can generate a confidence interval for the mean that is narrow enough to meet some precision requirement. Let n^* be the random sample size associated with meeting a stopping rule. The stopping rule to end sampling when we have collected enough samples is written as

$$n^* = \text{argmin}_n \, t_{\alpha,n-1} \frac{s}{\sqrt{n}} \leq \delta,$$

which means we choose the smallest value of *n* for which the half-width of the confidence interval collected with *n* samples is smaller than the desired precision δ. The parameter *δ* is the precision desired from the output confidence interval. Testers will require some level of precision in their mean estimate, and *δ* represents an absolute level of precision. For example, if the confidence interval for the mean time to failure should be smaller than 10 hours, and δ is the desired half-width of the confidence interval, then δ is 5. Another option is relative precision, where the precision is expressed in units of the sample mean as in

$$n^* = \text{argmin}_n \, t_{\alpha,n-1} \frac{s}{\sqrt{n}} \leq \delta \bar{x}.$$

An example is that the confidence interval for the mean length of parts produced by a machine should be within 2% of the average length of the part. This setting is called relative precision and is useful when it is not clear what the scale of the mean is ahead of time, but we know the potential error should be small relative to the mean.

If instead of *δ*, we have a cost of sampling that determines the number of samples, we still need to estimate the variance to employ a sequential sampling rule. If the variance of the data is not readily accessible (and usually it won't be for any previously untested system), then sequential estimation can be employed where we

estimate the variance until a stopping condition is met and we have enough samples. For example, we can use the following rule:

$$sample\ and\ increase\ n\ until\ \hat{\sigma}_n \leq \frac{2cn^{\frac{3}{2}}}{z_\alpha}.$$

This rule increases the sample size and estimates the sample variance until it is small relative to the cost of an additional sample. The rule is found by using the fixed sampling rule for n^* and solving for σ to determine what the sample variance needs to be relative to the cost and the number of samples.

THIS PAGE LEFT INTENTIONALLY BLANK

Parameters of Sampling Rules

This section discusses the inputs to sampling procedures that determine the quality of the results. Four of the inputs are chosen, and one is outside the control of the user. We list them here and then discuss each one in detail.

- The desired precision level δ representing the absolute upper-bound on the half-width, or the relative precision coefficient.
- The cost c for each test.
- The desired confidence coefficient $1-\alpha$.
- The starting sample size of the procedure. To construct a confidence interval, at least two samples are needed. In some cases, the sequential stopping rule will stop after two samples. An initial sample size may already be budgeted by the tester, or may be part of some minimum requirements.
- Finally, the variance of the underlying system is an input and greatly affects the quality of the sequential stopping rule. Unfortunately this is uncontrollable and usually unknown.

The effect of each of these factors is discussed in the following section.

Precision and Cost

The parameters δ and cost c play similar roles in sequential sampling rules. They are chosen/known quantities that control the number of samples needed. The parameter δ represents the desired precision in the result and is chosen by the tester to determine when enough accuracy in the mean result has been obtained. A small value of δ means that a large sample size will be required in order to obtain a narrow confidence interval, as discussed in the fixed sampling section. A small value of δ also means better confidence interval coverage results, as discussed later in this paper. However, a larger choice of δ could save money by allowing stopping to happen earlier.

A high cost c means that it is optimal to stop earlier, while a lower cost would allow for more replications. In a T&E setting, it may be more reasonable to use cost as the stopping criterion than δ, because the cost is a real factor that needs to be

managed, whereas the precision of a confidence interval has less of a real impact on whether the system is implemented, and thus δ is often chosen on an ad hoc basis.

Variance

The variance of the underlying data plays a major role in the quality and results of sequential sampling. It is often unknown and must be estimated. If the variance is large, then many samples may be needed to obtain a narrow enough precision around the mean. In these cases, the mean may not even be a good measure of performance because the variance of the system is so high that it cannot be assumed that the system will perform near mean performance when it is actually deployed. It may make sense to invest more time in reducing the variation of the system before further testing continues. If the variance is small, a small sample size may be enough to obtain a good estimate of the mean and stop sampling.

There are a few potential issues to consider when estimating the variance. It could take many samples just to obtain a good estimate of the variance. Also, if two samples happen to be close together at the start of the experiment, the variance could appear lower than it actually is, incorrectly implying that a good estimate of the mean has been reached. The sequential stopping rule could stop prematurely leading to a confidence interval that appears narrow with a good estimate of the mean, when in fact the variance is much larger and the interval should be centered around an entirely different mean.

Key point: Underestimating the variance of the data will lead to too few samples, resulting in a poor confidence interval!

Figure 1 shows an example of the distribution of the stopping time of a sequential rule when the data is normally distributed with mean zero and variance 1, and an absolute precision stopping rule is used with tolerance $\delta=0.3$ and desired confidence $1-\alpha=90\%$.

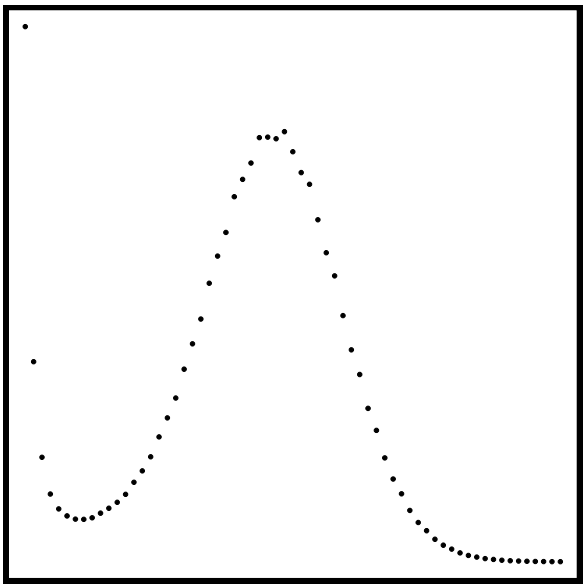

Figure 1: The x-axis is the value of the stopping time n, and the y-axis is the probability distribution for the stopping time. Figure from Singham (2014).*

As we can see, the mode of the distribution is actually at sample *n=2*. This means that there is a relatively high probability that a sequential confidence interval procedure will stop after two samples with a belief that the desired precision has been obtained, when in fact it was just by chance that the first two samples are close to each other, not necessarily close to the sample mean. If the first few samples happen to be close to each other, the estimate of the variance will be much lower than the true variance. This means that it is important for testers to be aware of this potential underestimation of variance (meaning underestimation of risk), and not simply stop after a few samples if the results happen to be close together.

The Initial Sample Size

To resolve the problems with early stopping discussed in the previous section, we discuss the choice of an initial sample size. This choice of an initial sample size is often made without much thought, with the assumption being that only the information collected at the end of the experiment with all the samples matters. However, choosing the initial sample size to be large enough to avoid a poor confidence interval is critical to preventing early stopping. Singham (2014) recommends 30 sample sizes, which is usually large enough to avoid problems with

sampling output, but this work was conducted in a simulation context where replications are cheap. In T&E, the budget may be limited to a few tests. The key is not to stop earlier than the budget allows. For example, if the budget allows for five replications of a test, it is critical not to stop after two samples if they appear similar. The probability of inadvertently meeting a stopping rule is high after two or three samples, but drops off quickly. Thus, sticking with five samples reduces the risk of receiving an unexpectedly bad outcome. Of course, five samples may still be too low to get a precise answer, but it is better to have an uncertain outcome of an experiment that represents the real risk in the system, than to have an apparent "certain" output that is actually incorrect.

Assuming there is some flexibility in the sampling budget, rather than fixing a sample size, a sequential sampling rule can be employed. The difficulty is in knowing when a sequential stopping rule is likely to stop. This means it is uncertain how much to budget for experimenting. The point of this research is to show how we can attempt to get an idea for how many samples will be required by a sequential rule to better predict the budget needed. The sample size from a sequential experiment is a random variable, but with updated information as data is collected, we may be able to have an idea of how good it might be.

Confidence Coefficients and Sequential Sampling Bias

What many analysts don't realize is that sequential stopping rules cause a bias in the confidence interval coverage. This means that testers may think they are getting a 90% confidence interval, but the true confidence associated with the confidence interval is actually 80%. We call this decrease in the confidence coefficient the "loss in coverage" or the "sequential sampling bias." It can be hard to quantify the real impact of what is meant by a 90% confidence interval, and what it means to only unknowingly obtain 80% confidence. But the general idea is that we have more risk and uncertainty with an 80% confidence interval than with a 90% interval. Thus, the real performance of our system may be very different from our estimate, more than we think we have quantified, and we have underestimated the risk.

Singham and Schruben (2012) derived a method for explaining and calculating the loss in coverage. This method is outside the scope of this report, but the general idea is that because a sequential stopping rule relies on an uncertain variance estimate, the stopping rule could stop too early. The theoretical solution to this problem (Chow & Robbins, 1965) is to let δ approach zero, requiring the precision to shrink to zero. This results in the desired confidence coefficient $\eta=1-\alpha$ being obtained as the sample size increases to infinity. Of course, in any practical setting increasing the sample size to infinity is unrealistic, and in T&E we often are only allowed small sample sizes so we cannot even approximate the results to the limiting case. Thus, our goal is to acknowledge and quantify the loss in coverage associated with small sample sizes to obtain a better idea of the risk involved. We evaluate the trade-offs associated with different inputs and determine the best way to manage the risk and costs to obtain the best analysis from limiting experiments.

Output Measures

We briefly highlight the four outputs that are collected as the sequential sampling rule procedure is running.

- The sample mean varies in trajectory as the samples are collected. This is often the main measure of performance of the system. With a large enough set of samples in a sequence, it should approach the true mean μ. The confidence interval around the sample mean narrows over time.
- The sample variance of the mean is updated as samples are collected, and this helps determine the precision in the mean estimate and the size of the half-width of the confidence interval.
- The sample size n^* is the random stopping time of the sequential procedures and is unknown ahead of time. It is a random variable as a result of the sequential sampling procedure.
- The true confidence interval coverage of a procedure is uncertain. While $1-\alpha$ is the intended coverage, the sequential confidence interval induces a bias as discussed earlier that could reduce the coverage.

These values are all uncertain at the start of the experiment. In a fixed sampling experiment, the sample size is known, as is the confidence interval coverage, but the mean, variance, and half-width of the confidence interval are unknown. While the sequential rule allows for potential efficiency in sampling (by

stopping as soon as possible) and a fixed half-width, there is additional uncertainty in the confidence interval coverage and the sample size. It is important for the T&E practitioner to choose which outputs are most important to help decide whether to use fixed or sequential sampling.

We summarize the inputs and the outputs of sampling rules in Table 1.

Table 1: Inputs and Outputs of Sampling Experiments

INPUT PARAMETERS	OUTPUTS
Variance σ	Sample variance
Risk probabilities η, α	Actual confidence interval coverage
Precision to make a decision δ, cost c	Sample mean
Initial sample size n	Optimal number of samples, n^*

Table 2 summarizes the effect of increasing each of the input factors on the output metrics. We see there is a trade-off between the two metrics of the sample size n^* and the loss in confidence interval coverage. Trying to decrease the sample sizes leads to a higher loss in confidence interval coverage. The analyst needs to determine whether it is better to have fewer samples with lower cost and a larger risk in the output, or invest in more samples at a higher cost in order to obtain better confidence intervals.

Table 2: Effect of Changing Inputs on Outputs

Parameter (increasing)	Effect on n^*	Effect on coverage loss
Confidence coefficient η	Increasing	Decreasing
Precision δ	Decreasing	Increasing
Cost of data collection c	Decreasing	Increasing
Variance σ	Increasing	(Depends on parameters)

The effect of these parameters will be made apparent when trying different sampling rules in the Excel spreadsheet. The goal of the spreadsheet is for the analyst to be able to estimate the effect of different parameter choices before committing to an experimental decision or a testing budget.

Decision-Based Sampling Rules

This section links sampling rules with the T&E decision process so that experiments can be designed to directly support the decision of whether to adopt a system.

Estimating a Mean Value Relative to a Standard

We define the following terms:

- *d** is the critical performance value (performance specification). This is the value, or standard, at which the system must perform in order to be successful, or deemed operationally suitable.
- *s* is the standard deviation calculated or estimated from the data,
- *n* is the number of samples taken,
- *α* is the Type I error, which is the probability that we would incorrectly reject a null hypothesis that the mean performance of the data is equal to or better than *d**.
- *1-β* is the power of the test, which is the probability of correctly rejecting the null hypothesis when it is false, and *β* is the Type II error.
- $|\mu_1 - \mu_0|$ is the absolute difference between the desired objective (μ_1), and the minimum threshold for performance needed (μ_0), and hence is called the military significant difference.

As we collect samples, we calculate a t-statistic as

$$t_s = \frac{\bar{x} - d^*}{s/\sqrt{n}}.$$

The confidence coefficient associated with this t-statistic is the area under the density curve of the t-distribution with *n-1* degrees of freedom between

$$(-|t_s|, |t_s|).$$

In order to determine the samples needed to achieve a certain confidence coefficient *1-α*, we choose the *t*-value associated with the desired confidence coefficient and *n*=∞*,* and call this t_α. This is the same as the *z*-statistic for the normal distribution. Then solving for the number of samples yields

$$n = \left(\frac{t_\alpha \cdot s}{\bar{x} - d^*}\right)^2.$$

Given an updated value of *n*, we can now calculate t_α and repeat the calculation iteratively until we converge on a value of *n*. The problem is this assumes that a value of *s* is available, when in reality this must be estimated. Thus, our spreadsheet updates the value of *n* given updates to *s* as samples are collected. This equation is similar to the original fixed sampling equation based on *δ*, but in this case *δ* is defined as the difference between the sample mean and the desired standard performance. This is because we need a confidence interval narrow enough to distinguish the observed performance as being different from the standard *d**. Then we can determine if the system performance is better than, or worse than, the standard.

Wald's Sequential Probability Ratio Test

We can apply a particular sequential test when we are trying to estimate the fraction of defective items or failures in a system. The mean performance in this case is a proportion and is the average of success=1 and failure=0 data values. The parameter *α* is the producer risk that we accidently reject a good system, while *β* is the consumer risk that we accidentally accept a bad system. Wald's sequential probability ratio test (SPRT) is a sequential sampling rule that helps decide between a null hypothesis for some parameter and an alternative hypothesis. We focus on the case where this parameter is the probability of success. The value of p_1 is the null hypothesis for the probability of success, and p_2 is the alternative hypothesis for the probability of success. Rejecting the null implies we use the alternative, while failure to reject implies we use the null.

Generally, Wald's SPRT relies on the fact that boundaries for choosing the null or alternative can be chosen as:

$$a = \log\frac{\beta}{1-\alpha}, \; b = \log\frac{1-\beta}{\alpha},$$

where we stop and choose the null whenever some value *X* falls below *a*, and choose the alternative if it goes above *b*. The value *X* is the log-likelihood ratio between the null being true and the alternative being true. In the case of estimating a probability from successes and failures, the likelihood function given a particular

hypothesis p_i for the probability of success is proportional to $p_i^S(1-p_i)^F$ where *S* is the number of successes and F is the number of failures. It is straightforward to calculate the log-likelihood ratio between the two hypotheses as samples (successes or failures) are collected and determine when the boundaries have been crossed.

THIS PAGE LEFT INTENTIONALLY BLANK

Our intention is to guide the T&E practitioner in analyzing the trade-offs discussed above and in making the decision on how much to budget for testing. Sequential testing can help update the testing budget as needed, but in some cases the entire budget must be decided completely ahead of time. While complex numerical integration is often needed to determine the actual performance of sequential rules, here we use approximations to allow the calculations to be completed in Excel. The main idea is that the user should be aware of the loss in coverage that could occur when a sequential rule does not require many samples. We have two separate cases, one when data is available to calibrate the tests and one when it is not available.

First, we need an estimate of the variance of the data from the user. This estimate is very important as it drives the sampling requirements, and underestimating the variance will lead to too few samples and overestimating confidence in the results. This leads to an underestimation of risk associated with the system, and so we encourage the tester to overestimate risk to avoid an interval that is too narrow.

Next, we establish the precision needed in the result. This asks the tester to determine the accuracy needed in the mean result, meaning the absolute or relative amount that the mean estimate can be wrong and the system will still be operationally suitable. This is a difficult value to quantify because we are operating in terms of averages, and it can be hard for the user to estimate what tolerance to differences in the mean value exists. In some cases, there may be a significant cost that drives the sample size, and we will allow the user to put in a cost if they wish to include this in their sample size determination.

The confidence coefficient also needs to be selected. Usually this is chosen on an ad hoc basis as 90%, 95% or 99%. A larger choice will lead to more samples being required. The choice of confidence coefficient can be difficult to quantify, and

thus is largely left to the analyst, though we encourage higher confidence coefficients to be safe and offset the effect of a potential loss in coverage. The actual coverage delivered by the procedure may be less than what is desired if a sequential rule is used, so choosing a higher confidence coefficient allows for some leeway if there is a loss in coverage.

The user will employ either a sequential test or a fixed sampling test, depending on whether an estimate of the variance is available. In the fixed sampling test, they put in their parameters ahead of time and the spreadsheet reports the sample size they should use. In the sequential sampling test, a preliminary estimate of the number of samples is given as samples are collected. Then, a stopping rule is employed to determine when data collection can stop.

Providing exact statistical results on the quality of a sampling rule is not possible without knowing the true distribution of the data, and in any testing environment, it is highly unlikely that the distribution is known or that there is even data available to estimate the distribution. Thus, our goal is not to provide exact results, but rather conservative approximations that reveal the trade-offs and potential risks associated with sampling. We emphasize that the overall goal of the test should not be to obtain an exact estimate of performance, as that would not be possible with a finite sample size. The goal is to obtain a confidence interval for likely performance, and to understand the risks associated with those confidence interval results.

Additionally, the tool will help the user be aware of the trade-offs between the different inputs. By changing the precision and confidence coefficient, they can see the changes in expected performance and sample size. If the tester has a limited budget, he or she can see what performance they can obtain within the budget and in theory can lobby for a higher budget if the tests prove insufficient to evaluate the system. In a constrained resource environment, some trade-offs must be made, and we hope the tester becomes aware of the trade-offs and can make an informed decision.

Test and Evaluation

This section describes how sampling rules can be specifically applied to T&E. The performance measures to be collected could relate to the requirements, or to critical operational issues (COIs). Focusing on the COIs helps provide direction for the test program and can allocate effort toward data requirements effectively. It is important to define specific test objectives as well, which organize measure of effectiveness and measures of performance under a particular COI.

Criteria characteristics for determining evaluation criteria and objectives often are defined on absolute or relative (to the baseline) terms and may have statistical confidence levels as part of the wording. These confidence coefficients are often applied to reliability measurements, like the probability of success of a task. In order to estimate things like maintainability, we need a reasonable number of maintenance events to occur, which could require a large number of tests. Qualitative measurements, like interoperability, may not be able to be used in sequential rules because they cannot be numerically quantified using a mean. The sampling plan should be established when the Test and Evaluation Master Plan (TEMP) is drafted prior to Milestone A decisions and updated again in the TEMP prior to Milestone B, to incorporate all these issues. Early Operational Assessments can be used to help calibrate the parameters of a sampling rule, either by helping with variance estimation or having a rough idea of what the sample mean will be. These assessments can sometimes rely on Modeling and Simulation (M&S), and if a simulation model is available, it can be used with its own sequential sampling rule to generate samples and obtain a variance estimate.

The precision/tolerance of a sampling rule can be decided based on the capabilities gap and requirements that are decided at the start of a test process. Knowledge of the requirements can help determine the level of confidence needed, or the precision in the result that is needed to meet the capabilities gap. For a particular performance specification, the threshold is the minimum acceptable performance, and the objective is the desired value that is better than the threshold.

These values can be used to decide the needed precision in the result in order to evaluate the system as acceptable for use.

Additionally, it is important to know the decision point of performance, which is the performance value that is needed for the system to be labeled operationally suitable. This target value can be used to help determine the required precision in the output, and the confidence level represents the risk associated with making an incorrect decision about the operational suitability. Choosing the confidence level in T&E depends on the critical issues of mission accomplishment, the expense of more tests, the cost of fixing and redoing the experiment if the runs fail, the safety of tests, and contract incentives.

Developmental Test and Evaluation (DTE) can be used to make early operational assessments (EOA) and collect measures of performance estimates, which might be appropriate for using confidence analysis. Often factors are changed one at a time to estimate different effects. Operational Test and Evaluation (OTE) is often dedicated to looking at overall effectiveness, operational suitability, and answering the question “does it work,” which gives a binary response. The measure of effectiveness could be a probability of detection, or reliability probability, which is an average of a binary response. It can take many samples to estimate these probabilities. Understanding the nature of the T&E process is critical to choosing the appropriate inputs for sampling rules.

Excel Tool to Implement Sampling Rules

This section describes the Excel tool that implements the sampling rules. This tool is meant to be a way of incorporating the sampling ideas discussed in this report and allowing for a simple and easy application of the methods to obtain a quick estimate of the testing effort required. The first worksheet "Parameters" is where all the user inputs are entered. The key inputs are listed and described in Table 3:

Table 3: Input Parameters for Sampling Rule Analysis

Performance Threshold (minimum acceptable level)	23
Performance Objective (desired level)	24
Confidence interval tolerance	0.1
Alpha (Type I error)	0.1
Beta (Type II error)	0.2
Proportion or Value (PROP or VAL)	PROP
Variance/Proportion Estimate (number or EST)	EST
Stopping Type (ABS or REL)	ABS

All yellow cells need to contain inputs from the user. Across the whole workbook, the user should enter or change values only in the yellow cells; everything else will calculate automatically. On the Parameters sheet, many yellow cells have comments included in the Excel file to help explain the inputs. The first two cells on the Parameters page are the Performance Threshold and the Performance Objective. The Performance Threshold is the minimum performance needed to determine whether the system is a success. The Performance Objective is the desired performance, which is usually higher than the minimum threshold. If there are not clear performance thresholds or objectives, then the Confidence interval tolerance cell (next) can be used to enter a desired value of δ. The first three cells will be used to determine the desired precision in the result needed to make a decision.

The next two cells are the values of α and β, the desired Type I and Type II errors of the experiment. The value of α is important for the fixed and sequential sampling rules, while both α and β are used in Wald's SPRT. The next cell asks for the value PROP or VAL. This is asking whether the desired metric is a proportion between 0 and 1, or just any value that is a real number. This is because our implementation of Wald's SPRT is based on estimating a proportion. The cell for the Variance/Proportion Estimate depends on whether PROP or VAL is entered in the previous box. If PROP is used, then a nominal estimate of the proportion should be entered. If VAL is used, then the user can enter an estimate of the variance if available, or else EST can be entered, meaning the variance will be estimated using a sequential sampling procedure. The last cell asks for a stopping rule type, where ABS stands for absolute precision and REL stands for relative precision. The definitions of these rules are included in the Section "Confidence Intervals for Data Collection." The error boxes will highlight if an inappropriate value is entered in any of these boxes.

The Parameters page then includes some outputs that depend on the values entered in the yellow cells (see Table 4).

Table 4: Supplementary Parameters Calculated in the Excel Tool

Worksheet to use:	FixedSampling
Difference Threshold and Objective	1
Precision (delta)	0.1
Confidence coefficient	0.9
Power	0.8

The user should not change these values; hence they are not highlighted in yellow. The first cell in blue directs the user to a worksheet depending on the values entered above. There are three worksheets, FixedSampling, SequentialSampling, and WaldsSPRTest. The user can then go to these worksheets to enter their data, if

needed and obtain the results. Values used in calculations in these pages will link to values in the Parameters page, so it is crucial that the Parameters page is kept updated at all times. The Difference Threshold and Objective cell takes the difference between the threshold and objective, which can be used as a proxy for the tolerance. The final value for the Precision (δ) is the smaller of the confidence interval tolerance and the difference between the threshold and objective values. A smaller precision value is better for increasing the sample size and obtaining a more accurate confidence interval. Finally, the confidence coefficient is calculated as $\eta=1-\alpha$ and the power is $1-\beta$.

Next we describe each of the worksheets and how they should be used. The first worksheet, FixedSampling, can be used in the rare case where the variance of the system is known ahead of time and entered in the Variance/Proportion Estimate yellow cell. This is unlikely to happen in practice, but of course the user can put in different values of the variance and see what the sample size should be. The point of this workbook is to allow for experimentation and see the effect of different parameters. The worksheet FixedSampling just has one output based on the values entered in the Parameters page, and that is the estimate of the number of samples needed to conduct a fixed sampling experiment to obtain a confidence interval with level $1-\alpha$, and a precision level δ given the variance in the system.

The SequentialSampling worksheet is more involved. It still draws most of its inputs from the Parameters page. The main information to be entered by the user is the values of the samples as they are collected. This information should be entered in each row in the yellow cells as collected. Then, the columns to the right of the yellow cells will populate based on the values of the samples. The first column to the right highlights “YES” or “NO” depending on whether the stopping rule is met. The user should continue sampling until the stopping rule is met and “YES” appears. If the stopping rule has been met, then a confidence interval for the mean appears to the right. For all samples, an estimate of the total number of samples needed is also provided. This means that as the user is collecting samples, he or she can have an estimate of the number of samples needed to meet the stopping rule. This will help

reallocate the testing budget and give information on how long the procedure may need to continue. Finally, a flag is given if the expected number of samples is smaller than 50, meaning there could be a potential loss in coverage in the confidence interval. Sequential stopping rules that predict stopping with fewer than 50 samples could be prone to bias in the results, and thus the user is aware that the sequential rule may have worse coverage than the value of $\eta=1-\alpha$ predicted. The user should assume some additional risk in the system in this case.

Finally, the WaldsSPRT worksheet is for entering samples when the goal is to estimate a probability. As in the sequential sampling page, there are yellow cells where the results of the tests can be entered. In this case, a value of 1 is entered if the test was a success, and 0 if it was a failure. Wald's test is a sequential test and asks for additional samples until it can make the correct determination between two hypotheses (one null and one alternate) with Type I error α and Type II error β. The yellow cells take the sample values, and the column to the right highlights the result if the test can determine whether to accept or reject the null, and the resulting hypothesis to use. The user needs to enter additional information for the values of the null and alternate hypotheses at the top of the page in the yellow cells (see Table 5).

Table 5: Inputs for WaldsSPRT Worksheet

Estimate	0.5
p1	0.3
p2	0.35

The cell Estimate links to the parameters page where the user may have entered in a nominal value of what they believe the proportion of success is. This value could be used as the null hypothesis, though we leave it open to be changed in the cell "p1." Then "p2" can be used to enter an alternate hypothesis for the proportion of success. The worksheet asks the user to continue to collect samples until Wald's test returns a result in favor of either the null or alternate hypothesis.

Summary

This research provides a link between sequential sampling theory and a major issue in T&E: choosing the number of tests to conduct. The goal is to enable test analysts to better understand the number of samples needed to make a determination about the quality of their system. In some cases, we can predict ahead of time the approximate number of samples needed. In most cases, it is easier to collect samples sequentially and stop when some conditions have been met. The main idea is that there are trade-offs between the parameters of the confidence coefficient, precision, variance, and sample number. This research aims to make these trade-offs easy to understand, and provide a tool for testing different options to estimate the number of samples needed.

In some cases, the tester will be limited by their testing budget. The spreadsheet allows the tester to try different parameter values to see what options allow them to work within their budget. The spreadsheet also updates the estimate of the number of samples needed as results are collected so that the budget can be adjusted. The intention of this project is to remove some of the guesswork in making the sampling decision, as this decision must be made for every experiment. With better knowledge about the quality of results we are likely to receive, we can balance the cost of testing with the precision in the results carefully without leaving these factors to chance.

THIS PAGE LEFT INTENTIONALLY BLANK

List of References

Chow, Y.S., & Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics, 36*(2), 457–462.

Singham, D. I. (2014). Selecting stopping rules for confidence interval procedures. *ACM Transactions on Modeling and Computer Simulation*, *24*(3), Article 18.

Singham, D. I., & Schruben, L. W. (2012). Finite-sample performance of absolute precision stopping rules. *INFORMS Journal on Computing*, *24*(4), 624–635.

U.S. Department of Defense. (2005). *Test and evaluation management guide*. Fort Belvoir, VA: Defense Acquisition University Press.

U.S. Marine Corps. (2010, May). *Integrated test and evaluation handbook*. Quantico, VA: Marine Corps Systems Command.

U.S. Marine Corps. (2013, February). *MCOTEA operational test and evaluation manual* (3rd ed.). Quantico, VA: Author.

Wald, A. (1973). *Sequential analysis.* North Chelmsford, MA: Courier Corporation.

THIS PAGE LEFT INTENTIONALLY BLANK

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net