2017-03

Decision-Based Metrics for Test and Evaluation Experiments

Singham, Dashi

Monterey, California. Naval Postgraduate School

SYM-AM-17-082

Proceedings
of the
Fourteenth Annual Acquisition Research Symposium

Thursday Sessions
Volume II

Acquisition Research:
Creating Synergy for Informed Change

April 26–27, 2017

Published March 31, 2017

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

Acquisition Research Program
Graduate School of Business & Public Policy
Naval Postgraduate School

Decision-Based Metrics for Test and Evaluation Experiments

Dashi Singham—is a Research Assistant Professor of Operations Research at the Naval Postgraduate School, where she researches, teaches, and advises student theses. Dr. Singham's primary areas of focus include simulation modeling, simulation analysis, and applied statistics, with most of her work on developing new methods and metrics for analyzing simulation output. Her areas of application include energy and intelligence systems. She received her PhD in Industrial Engineering and Operations Research from the University of California Berkeley in 2010. [dsingham@nps.edu]

Abstract

We develop a new decision-based metric for determining sample sizes in Test and Evaluation experiments. Traditional confidence intervals for the mean can be used, and we present sequential confidence interval procedures as a way to derive efficient intervals. We discuss decision rules for analyzing the observed output and how to choose confidence interval methods for calibrating these decision rules. The metric presented can help determine if a fast decision on the quality of the system can be made or if many more tests are needed to ensure an accurate estimate of performance relative to a desired standard.

Introduction

Test and Evaluation (T&E) experiments are often conducted with the intent of answering a question about the feasibility of a new system. This system may have properties that are unknown, so rigorous testing is required to ensure the safety and performance of the system before it is adopted. We will use the term "system" to include any object under scrutiny via testing, be it a weapon, computer program, or piece of equipment. This work mainly applies to Developmental T&E where different performance metrics are analyzed individually, though it could also apply to Operational T&E where many varying factors are jointly tested.

In this paper, we will assume that there is a quantifiable non-binary metric for evaluating system performance so that averages and confidence intervals can be easily constructed. The intent of this work is to show how to better use quantifiable metrics to answer research questions or make a decision about the quality of the system in Developmental T&E. We will outline basic metrics for quantifying uncertainty in system output and show how these metrics can be mapped to a decision rule.

The main goal of data analysis is often to estimate the performance of a system using experimental data, sometimes using the sample mean or a confidence interval for the mean as the metric for evaluating the quality of the system. While these metrics are useful, they would be even more useful if they could be mapped directly to a decision. For example,

- If the system mean performance is greater than some value D, then we should adopt the system.
- If the system mean performance is greater than some value D with probability x, then we should adopt the system.
- f a 95% confidence interval for mean performance has a lower bound greater than D, then we should adopt the system.

In the examples above, D is the decision threshold that is used to determine whether or not to implement the system. It is important to decide beforehand the metrics for success and determine what D should be to ensure that the system is selected only if it will satisfy its

intended purpose. Waiting to choose *D* until after the system has been tested can lead to bias based on initial test results, and these initial results can be misleading if the system has a high variance.

In this paper, we will describe how confidence interval procedures can be used to design statistical test rules that link the data analysis to the decision threshold. By making a simple change to standard confidence interval procedures, new rules can be developed that incorporate the decision threshold D. These new rules will enable a better determination of whether the system should be implemented and can potentially save testing costs.

Confidence Intervals

A simple way to evaluate the effectiveness of the system is by taking the average of the test results. Let $\bar{x}_n$ be the sample mean of *n* test replications. This average can be compared to the status quo, to the averages of competing systems, or to a decision threshold. However, looking at the average alone does not account for variability and uncertainty in system behavior. Assuming that the system will always perform near the mean when it is implemented could significantly underestimate risk.

Confidence intervals provide a method for assessing the uncertainty in mean results. Let s_n be the estimate of the standard deviation of the data based on *n* samples. Define s_n^2as the variance estimate based on *n* samples, calculated as

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

The value of s_n^2 estimates the real variance of the system σ^2, which is usually unknown. Let $t_{\alpha,n-1}$ be the *t*-value associated with the *t*-distribution with *n-1* degrees of freedom and tail probability *α/2*. Furthermore, let *η* be the confidence coefficient desired in the resulting confidence interval. This coefficient is usually 90%, 95%, or 99% and *α* is 1-*η*. The Type I error associated with the test is often denoted using *α*. If the data is normally distributed and the variance is estimated, then the confidence interval for mean system performance using *n* samples takes the following form:

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

This confidence interval can be compared to the desired system performance *D*, or to confidence intervals for other systems, as will be discussed later. The center point of the confidence interval giving the estimate of mean performance can be compared to *D*, as well as the width of the confidence interval. Formally, we can define the half-width of the confidence interval as

$$\text{half-width} = \left[t_{\alpha,n-1}\frac{s_n}{\sqrt{n}}\right]$$

where narrower half-widths imply less uncertainty in the mean performance of the system. If the assumption of normality in the data is met, repeated collections of confidence intervals from new experiments will result in (1-α)x100% of the intervals including the true mean of the data *μ*, and ideally this value will be around 90%, 95%, or 99%, depending on the choice of *η*.

Confidence intervals help determine the quality of a mean estimate. A narrow confidence interval (small half-width) implies less variability around the estimated system mean and is desirable, while a wide confidence interval makes it more difficult to predict the behavior of the system. When fixing the sample size used for testing ahead of time, there is no control over the half-width.

Let δ be the desired precision in the resulting confidence interval, which is the maximum half-width that is acceptable to the T&E analyst. Smaller values of δ are desirable because narrower confidence intervals provide more precise information on mean performance of the system. Suppose 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

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

and this sample size will ideally (though not necessarily) yield a confidence interval for *μ* that has a half-width smaller than δ. This method can be used to estimate a sample size ahead of time that would be needed to produce a small confidence interval. Oftentimes, budgetary constraints are the driving force behind the choice of *n*. A quick comparison between the budgeted number of samples with the ideal choice of *n* using Equation 1 can help determine ahead of time whether the experiment will yield enough precision to get an adequate idea of the true performance.

We note that many methods exist for choosing the sample size for T&E experiments, and guidelines incorporating sampling for different settings are presented in the *Test and Evaluation Management Guide* (2005), the *2010 Integrated Test and Evaluation Handbook* (United States Marine Corps [USMC], 2010), and the *Operational Test and Evaluation Manua*l (USMC, 2013). This work aims to deliver specific sequential sampling techniques that can be used in conjunction with these guidelines to better inform the sample size decision so that appropriate budgetary effects can be considered.

Sequential Sampling

Sequential sampling rules can be an improvement over fixed sample-size testing because they allow for adjustments to the sample size conditional on system performance as it is observed. Thus, after each test is conducted, the cumulative results are aggregated and an estimated confidence interval is computed. The decision to continue testing depends on the confidence interval produced from past samples. Sequential testing avoids the issue associated with Equation 1 where knowledge of the sample variance is required.

For example, if after 30 test runs of a system the confidence interval for the mean is very narrow, it may be unlikely that more runs will produce any additional information or variety in the results. In this case, testing could stop to avoid wasting money on future samples. However, if the confidence interval is quite wide, more tests should be conducted to better understand the uncertainty in the system. Additional tests will narrow the confidence interval to give more precision in the results and will also help better assess the risk in the system. Sequential rules check the confidence interval after each sample, and determine whether testing should continue. In this section, we provide the mathematical notation for understanding sequential confidence interval procedures.

The benefits of sequential sampling can be immediate when applied to a T&E setting. It is cost effective to stop as early as possible and may be wasteful to continue to test after prior tests have established the performance of the system. However, there is a potential for statistical bias associated with sequential sampling, as will be discussed at the end of this section. In this paper, we will not address this statistical bias directly for brevity, but we acknowledge that when sequential sampling stops with only a few test results, there is a high potential for bias in the results. This bias decreases as the number of samples increases.

Another benefit of sequential sampling is that the tester may have no idea ahead of time how many samples are needed to generate a narrow confidence interval for mean performance. If the underlying variance of the system is known, or can be estimated, then a formula such as Equation 1 can be used. But for new systems, the variance is usually not known and must be estimated as data is collected. Thus, it is difficult to know ahead of time how many tests are needed. Sequential sampling removes the need to make this decision and allows the sample size to be variable and adjust to the conditions of the data.

Sequential sampling rules allow for the tester to stop when some specified criterion is reached. This criterion is often a statistical property of the data collected up to that point. The main example we will use is to stop sampling when a confidence interval with a half-width smaller than some precision value can be generated from the data. Instead of a fixed sample size *n*, let n^*be the number of samples collected as the result of a sequential stopping procedure. This value is random, in that it will vary depending on the output values of the test. The values of n^* can be represented using

$$n^* = \underset{n}{\operatorname{argmin}}\, t_{\alpha,n-1}\frac{s_n}{\sqrt{n}} \leq \delta \tag{2}$$

where n^* is the smallest value of *n* (the first time the criterion is observed) where the half-width of the confidence interval collected with *n* samples is smaller than the desired precision δ. This value of δ is similar to the one used in Equation 1 and represents the allowable uncertainty in the sample mean estimate (the confidence interval). Recall that s_n is calculated as samples are collected, and this will make n^* random and depend on the particular values of the samples observed up to that point. Equation 2 is called an absolute precision rule, because the desired precision in the confidence interval is fixed ahead of time. Another type of rule is relative precision, where the precision can depend on the mean of the data. An example of a relative precision rule is:

$$n^* = \underset{n}{\operatorname{argmin}}\, t_{\alpha,n-1}\frac{s_n}{\sqrt{n}} \leq \delta\bar{x}_n$$

where the required precision of the confidence interval will be smaller for data that have smaller values (when $\bar{x}_n$ is small). Relative precision is useful when the tester does not have any information on what the mean of the data will be, but wants the error in the mean estimate to be within some percentage of the overall performance (for example the half-width should be within 5% of the estimated mean performance value). Note that for both absolute and relative precision rules, the variance must be estimated with each additional sample.

As an example, some specifications for small-arms tests involve absolute precision rules while others involve relative precision. The report TOP 3-2-045 outlines the maximum permissible error of measurement for small arms tests (United States Army Developmental Test Command, 2007). Some metrics require absolute precision in the results (thermograph reading measurement error must be within 0.6 degrees Celsius) while others require relative precision (the viscometer error should be within 0.5% of the full-scale reading). These error values provided are presumed to be two standard deviations over the data, and these values can be used as is or modified to be used as the input δ in a sequential procedure.

Decision-Based Performance

Simulation experiments are often used to help make decisions on whether to implement or modify a system. Because computer models allow for systems that have not yet been constructed to be tested, we can experiment with lower costs than building a

physical model. Test and evaluation plans can include simulation system tests, as well as physical tests of real systems.

A common question is "How should we determine what metrics to use in collecting experimental output?" If the system exhibits variable and uncertain behavior, we usually seek to estimate some measure of performance, μ. Confidence intervals are used to measure variability of an estimate. The risk of the confidence interval estimate is measured using its confidence coefficient (η), and the precision is measured using the interval half-width (δ). Estimates of the mean are collected using a sampling rule, and a confidence interval is constructed to help make a decision.

However, the experimental parameters used are often independent of system performance. We ask for the same risk and precision regardless of the data output and even if the output gives mixed results. We would expect a user to want more strict requirements on the precision when the system performance is close to the boundary between deciding to implement or not. Or, a risk-averse individual may want more confidence in a result suggesting that the system be implemented, and may be quicker to decline to implement a system that is unlikely to be better than the status quo.

For example, it may be critical that a system has performance greater than some threshold *D* (recall the examples from the Introduction). If the first set of experiments shows conclusively that $\bar{x}_n > D$ so that it is highly likely that $\mu > D$, then it is not necessary to obtain a narrower confidence interval. However, if $\bar{x}_n$ is close to *D*, the original value of δ might be too wide to differentiate if μ is actually better than *D*. In this case, a smaller value of δ should be used to drive up the number of samples needed. The type of output confidence interval should depend on the potential effect on the final decision to be made. More precise intervals with higher confidence coefficients should be required when the results of system experiments are close to the boundary between implementation or not. Less strict confidence intervals are needed if the system is performing exceptionally well or poorly, in which case the implementation decision is clear.

While we do not know what the true performance of the system is (hence requiring a T&E study), we do know what the decision would be if the true performance were known. We can choose confidence interval parameters based on the type of risk and precision we wish to have for different levels of performance. The confidence coefficient and precision are usually chosen before starting a simulation experiment and are static in that they do not change based on the resulting observations collected. We propose changing the precision parameter depending on the values of the observations collected as the procedure is running. This means we could obtain high-precision results for systems that are close to the decision point *D*, while stopping earlier with less precise results if it becomes clear early in the experiment that the system should not be implemented.

As a way of measuring the effectiveness of sequential confidence interval procedures, confidence interval coverage is often used, where coverage is the proportion of intervals generated by the procedure that cover the true mean μ. Nominal coverage is important for establishing validity of a procedure. However, here we consider the possibility that while a confidence interval may not cover the true system performance mean μ, it still may cover values that would lead to the same decision. There is usually some asymmetry in the type of error the tester will accept. For example, overestimating cost may be better than underestimating cost. But, if the procedure overestimates cost so much that an otherwise profitable system is no longer implemented, then the procedure has failed in two ways: in estimating the true cost and in failing to lead to the correct decision.

Let μ be the unknown true mean performance of the system. The threshold point *D* determines a binary decision for whether or not to implement the system. Suppose higher values correspond to better performance. If $\mu > D$, then we might choose to implement the system, and if $\mu > D$, we might decline to implement the system. The confidence interval can help determine the decision by comparing the values it covers to *D.* For example, if an interval lies completely above *D*, the decision would be to implement the system, while if the interval contained *D* and values below it, then the decision may be to delay or decline implementing the system.

In addition to an interval covering μ, we want to estimate the probability that the procedure results in an interval that leads to a correct decision being made. Consider the following four possibilities for an interval in Figure 1. The interval can either cover (include) μ or not, and it can either lead to the correct decision or not depending on its location relative to *D*.

Cover & Correct	Cover & Incorrect
Fail to cover & Correct	Fail to cover & Incorrect

Figure 1. Four Possible Confidence Interval Situations

Figure 2 illustrates the four situations presented in Figure 1. We assume that the mean performance of the system, μ, is greater than the decision threshold *D*. Thus the "correct" result of the experiment is that the system should be implemented. The top-left plot shows a confidence interval using parentheses that covers the true mean μ, and also lies on the right side of the decision threshold, thus making the correct decision. The top-right figure shows a different confidence interval that also covers the true mean μ. However, it fails to correctly predict that performance is greater than the decision threshold, because the confidence interval includes values on the left and right of *D*. The bottom-left confidence interval fails to cover the true mean μ. However, it is so far to the right that it still correctly estimates performance as greater than *D*. The bottom-right confidence interval not only fails to include *μ*, but it lies on the wrong side of *D*, so it will incorrectly predict that system performance is worse than *D*.

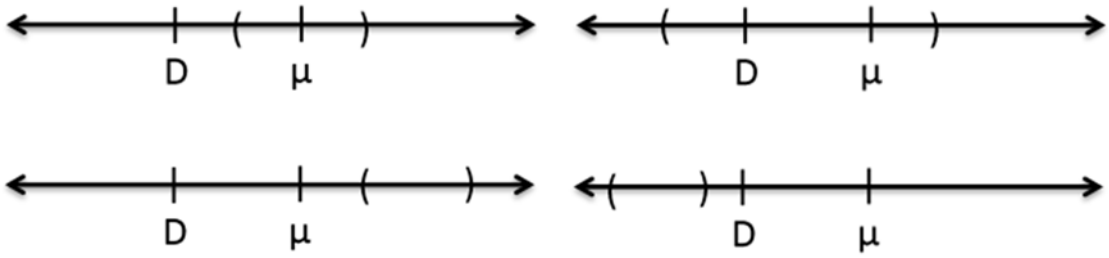

Figure 2. Visual Representation of Confidence Interval Situations

The goal of most confidence interval procedures is to provide adequate coverage of μ so that the procedure produces an interval that includes μ with probability 1-*α*. However, in the decision context, the correctness of the decision is potentially even more important. Both coverage and correctness are likely correlated, but correctness usually has more of an immediate impact than the effects of confidence interval coverage, which are only realized in the long term.

Implementation

The main result of this analysis is that we should choose δ to be small enough to distinguish μ from *D* in a sequential procedure. A confidence interval that is larger than | μ - *D*| may include μ, but may not be able to distinguish system performance from δ, as seen in the top-right plot of Figure 2. The catch is that we do not know μ at the start of the experiment. However, as samples are collected, $\bar{x}_n$can be used to estimate μ and will be updated with each sample. This value of $\bar{x}_n$ will be the center of each confidence interval. Thus, the sequential stopping rule can be changed to:

$$n^* = \underset{n}{\operatorname{argmin}}\, t_{\alpha,n-1}\frac{S_n}{\sqrt{n}} \leq |\bar{x}_n - D|$$

so that the stopping criterion is such that the experiment will not end until an interval that is small enough to distinguish $\bar{x}_n$ from the decision threshold *D* can be formed.

Making this adjustment would increase the efficiency in standard sequential stopping rules (the absolute and relative precision rules defined above) in a few ways. Sequential stopping rules can be "efficient" because they allow the user to stop as early as possible without wasting effort once a narrow confidence interval has been achieved. However, stopping depends on the choice of δ, which could be arbitrary. What we propose is choosing δ=$|\bar{x}_n - D|$ in an absolute precision rule so that the threshold for the half-width updates and adjusts based on how far away $\bar{x}_n$ is relative to *D*. This way it will be impossible to end with a confidence interval that looks like the top-right plot of Figure 2, because the half-width of the confidence interval will always be smaller than the distance between $\bar{x}_n$ and *D*, so it will never include *D*.

If it turns out the sample mean is close to *D*, then $|\bar{x}_n$-*D*| will be small. This will force the number of samples to increase in order to decrease the confidence interval half-width enough to distinguish whether the system performance is better or worse than D. If the sample mean is far away from *D*, then $|\bar{x}_n$-*D*| will be large so it will be easy to meet the stopping criterion after a few samples. Effort will not be wasted when it is clear that μ is on one side or the other of *D*.

The end result is that with this simple change, we can better allocate effort to test systems with a clear idea of the decision threshold. Our decision-making criterion informs the sequential test, and this means we only need to exert the minimum test effort to make a decision. The choice of *D* is very important and should not be made lightly. If $\bar{x}_n$is close to *D*, even if the confidence interval can distinguish system performance, there may still be high levels of risk that require more tests before making a decision about the system.

Of course, standard caveats associated with confidence interval coverage still apply. If too few samples are taken in a sequential procedure, confidence interval coverage can be poor, so the actual confidence could be much lower than the nominal 90% or 95% expected. This is a problem that can be addressed (e.g., in increasing the sample size or changing the expectation in confidence). Chow and Robbins (1965) is the classic reference showing that this bias in coverage decreases to zero as the sample size increases to infinity. However, large sample sizes are often not available in a T&E setting. The other option is to adjust expectations. For example, the tester can run the procedure trying for a 95% confidence interval, while acknowledging that in reality only a 90% interval will be achieved. For more details on calculating and preventing this bias in confidence interval procedures, see Singham and Schruben (2012) and Singham (2014).

Confidence intervals and sampling rules play a major role in determining whether systems meet specified performance thresholds using T&E experiments. For example, in

evaluating the performance of body armor in terms of resistance to penetration and deformation, confidence intervals are calculated, and the lower and upper confidence limits are compared to the requirements (Office of the Secretary of Defense, 2010). Specific methods, such as the Clopper-Pearson method, are suggested as a way to calculate confidence interval for probabilities when the output of the experiment is a binary measure of success/failure. We note that other sequential rules may exist for evaluating binary outputs or comparing two hypotheses (Wald, 1973).

While sequential testing may be useful in establishing sampling rules that have the desired precision, adding the decision component *D* would be an easy way of ensuring that the output confidence interval is not only precise but also useful for making the final decision.

Conclusion

Confidence intervals are a useful tool for evaluating T&E data. Sequential confidence interval procedures are a type of sequential testing that determines the sample size by computing a confidence interval after each sample is collected. These procedures potentially allow for a more efficient way of choosing the sample size than fixing it ahead of time. This paper proposes a new type of sequential confidence procedure using a decision threshold that determines whether or not a new system should be implemented based on observed samples. This new metric can potentially be used to either save testing costs or encourage more sampling when system performance is close to the decision threshold. Future work will test the statistical properties of this new metric and simulate the application of decision-based sequential testing using data from past experiments.

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.

Office of the Secretary of Defense. (2010). *Standardization of hard body armor testing* [Memorandum]. Washington, DC: Author.

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

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

Test and evaluation management guide. (2005, January). Fort Belvoir, VA: Defense Acquisition University Press.

United States Army Developmental Test Command. (2007). *Small arms—Hand and shoulder weapons and machineguns* (TOP 3-2-045). Aberdeen Proving Ground, MD: Author.

United States Marine Corps (USMC). (2010, May). *2010 integrated test and evaluation handbook*.

United States Marine Corps (USMC). (2013, February). *Operational test and evaluation manua*l (3rd ed.).

Wald, A. (1973). *Sequential analysis*. Courier Corporation.

Acquisition Research Program
Graduate School of Business & Public Policy
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net