STAT COE-Report-04-2015

Using Statistical Intervals to Assess System Performance

Authored by:

Francisco Ortiz, PhD

Lenny Truett, PhD

17 April 2015

Revised 12 October 2018

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

Background 3

- Missile Warning System (MWS) Case Study 3
- Requirement/Problem Statement 5

Statistical Intervals 6

- Confidence Intervals 6
  - Definition 6
  - Questions 6
  - Case Study 6
  - Things to note 7
- Prediction intervals 9
  - Definition 9
  - Questions 9
  - Case Study 9
  - Things to note 10
- Tolerance Intervals 10
  - Definition 10
  - Questions 11
  - Case Study 11
  - Things to note 12

Conclusion 12

References 13

Revision 1, 12 Oct 2018: Formatting and minor typographical/grammatical edits.

Executive Summary

Statistical intervals should be reported when assessing a system's performance within its operational space. This best practice demonstrates how to use different types of statistical intervals in conjunction with design of experiments (DOE) and regression analysis to best tackle the principal questions behind testing. Using such an approach adds greater rigor in the assessment of a system, extracts more information from limited resources, and avoids the much criticized folly of reporting a single average across all test conditions.

Keywords: Confidence, Prediction, Tolerance, Intervals, Regression, Analysis, Design of Experiments

Introduction

The Director, Operational Test and Evaluation (DOT&E) FY 2012 Annual Report criticized the practice/folly of reporting a single average of a system's performance across all test conditions (Gilmore, 2012). The same report advocated the use of advanced statistical methods in conjunction with test designs developed using design of experiments (DOE). Statistical methods such as regression analysis and statistical intervals combined with DOE allow programs to assess a system's performance with greater rigor, while extracting more information from limited resources. The Scientific Test and Analysis Techniques Center of Excellence (STAT COE) has observed a lot of confusion in the test and evaluation (T&E) community regarding the interpretation and application of some commonly used statistical intervals. In Department of Defense (DoD) testing, we often make an assessment about a system's performance based on limited sample data. Due to this limitation, there is always some level of uncertainty in the system performance estimates. A way to quantify the uncertainty of the estimate is by constructing a statistical interval. In this best practice, we provide clarification on how to use three commonly calculated intervals in DoD testing: confidence intervals, prediction intervals, and tolerance intervals. For each interval, we provide a layman's definition as well as demonstrate its use on a Missile Warning System case study in which a designed experiment and regression analysis are employed. This best practice will not go into details regarding the mathematics and formulation of each statistical interval. The formulation of the intervals varies based on test methodology used and parameter of interest. Most statistical software will do these calculations by default, so there is no need to go into details; however, it is important to understand the underlying assumptions behind each statistical interval. There are many good sources available if you are interested in learning more about the mathematical details of these intervals (see, for example, Montgomery [2017] or Anderson-Cook [2009]).

Background

Missile Warning System (MWS) Case Study

To illustrate how to use these statistical intervals, we'll use a generic example of a designed experiment applied to assess a Missile Warning System (MWS) as shown in Figure 1. An MWS works in conjunction with a counter measure (CM) tracker in order to defeat guided seeker threats to aircrafts. The MWS acts as a cueing system by detecting, declaring, and eventually handing off a potential threat to the CM tracker. The ultimate goal of the analysis is to assess various performance measures and help make a determination on the suitability of the MWS. One such performance measure is "time to handoff," which has a threshold requirement to be under 500ms. Note that **all data presented in this best practice are notional and used for demonstrative purposes only.**

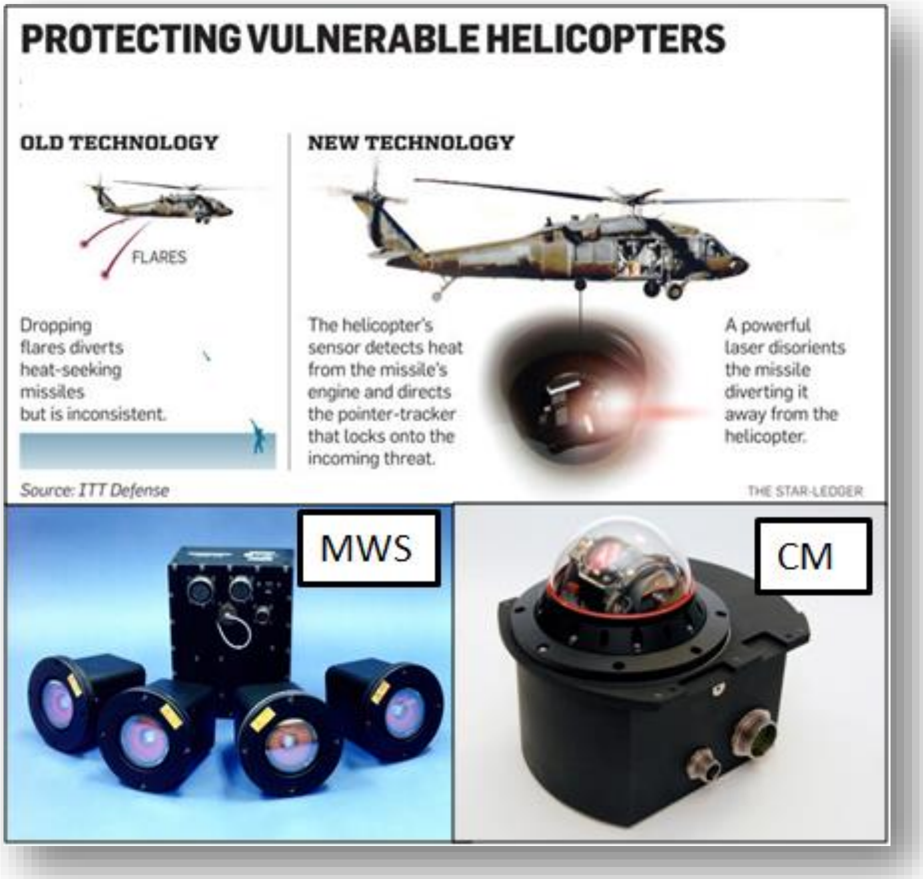

PROTECTING VULNERABLE HELICOPTERS
OLD TECHNOLOGY
FLARES
Dropping flares diverts heat-seeking missiles but is inconsistent.
NEW TECHNOLOGY
The helicopter's sensor detects heat from the missile's engine and directs the pointer-tracker that locks onto the incoming threat.
A powerful laser disorients the missile diverting it away from the helicopter.
Source: ITT Defense
THE STAR-LEDGER
MWS
CM

Figure 1: Missile Warning System application (Source: ITT Defense)

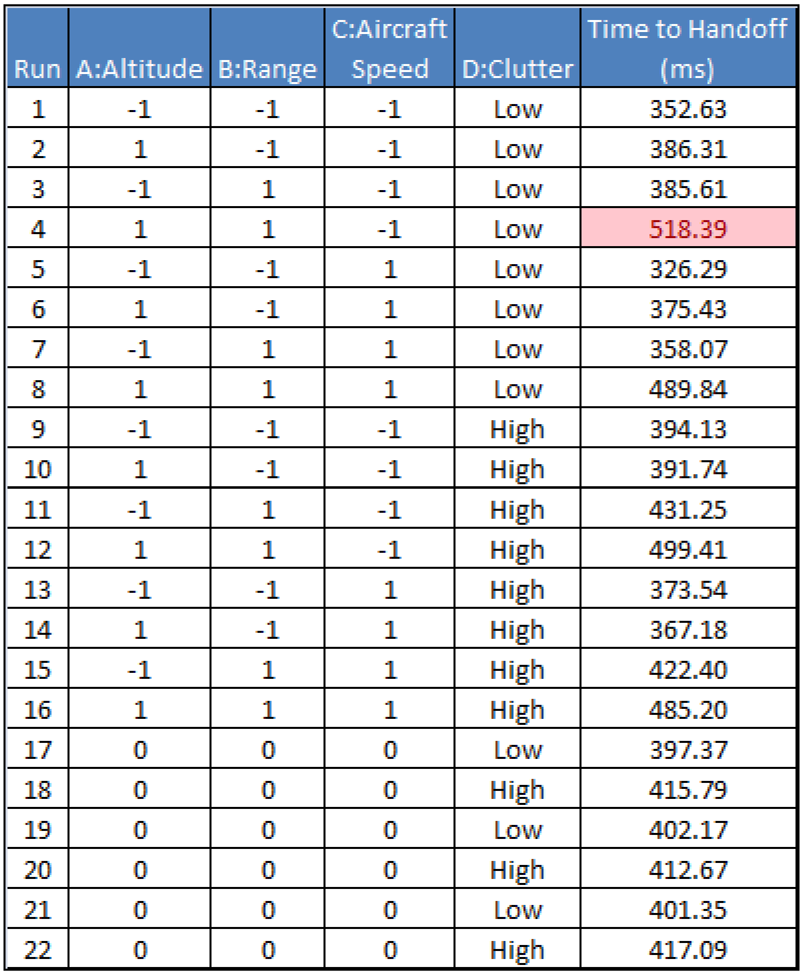
MWS handoff capabilities and timelines vary according to threat type, engagement slant range, atmospheric conditions, clutter level, and platform flight profile.

For simplicity, the designed experiment will only consider one threat type and will vary the following factors at a high and low level (+1, -1 in coded units, respectively):

- Altitude
- Range
- Aircraft Speed
- Clutter

The following 2^4 design (with 6 center points) shown in Table 1 was created and executed for the MWS. The performance measure of interest (i.e., response) is time to handoff and is shown in the last column in Table 1.

Table 1: 2^4 design for MWS test



Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)
1	-1	-1	-1	Low	352.63
2	1	-1	-1	Low	386.31
3	-1	1	-1	Low	385.61
4	1	1	-1	Low	518.39
5	-1	-1	1	Low	326.29
6	1	-1	1	Low	375.43
7	-1	1	1	Low	358.07
8	1	1	1	Low	489.84
9	-1	-1	-1	High	394.13
10	1	-1	-1	High	391.74
11	-1	1	-1	High	431.25
12	1	1	-1	High	499.41
13	-1	-1	1	High	373.54
14	1	-1	1	High	367.18
15	-1	1	1	High	422.40
16	1	1	1	High	485.20
17	0	0	0	Low	397.37
18	0	0	0	High	415.79
19	0	0	0	Low	402.17
20	0	0	0	High	412.67
21	0	0	0	Low	401.35
22	0	0	0	High	417.09

Requirement/Problem Statement

The MWS program wishes to demonstrate that the time to handoff will not exceed 500 ms throughout the operational region as defined by the factors and levels. A more statistically precise statement would be that the program wants to show, with 95% confidence, that the probability of success is at least 99%. That is,

$$\Pr(\mu_{time} < 500\text{ms}) \geq 0.99$$

at any point within the design space. Note that μ_{Time} represents the population mean time to handoff. Run number 4 (highlighted in red in Table 1) already demonstrates that the MWS can exceed 500ms under certain conditions.

The following regression model was created based on the collected data and can be used to predict time to handoff performance (in milliseconds) across the design space:

$$\text{Time to handoff} = 409.27 + 29.35A + 38.93B - 10.09C + 9.86D + 20.09AB - 14.07AD$$

The regression model is represented graphically in Figure 2 with aircraft speed and clutter held constant. This surface plot clearly shows the relationships between the response (time to handoff) and two of its input factors (altitude and range). You see that as range and altitude increase, the MWS takes longer to handoff. The model allows for interpolation within the design space, thus allowing for prediction of untested scenarios.

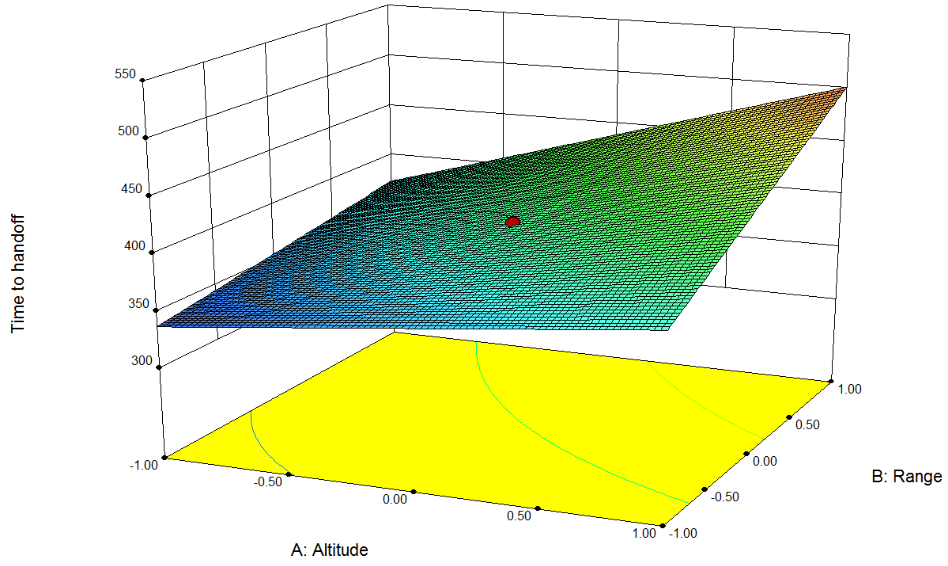


Figure 2: 3-D graphical representation of the regression model developed using design of experiments for MWS example

Statistical Intervals

For this MWS case study, the parameter of interest is the mean time to handoff. The basic form for a statistical interval for the mean is as follows:

$$\bar{y} \pm c_{(\text{level, n})} \cdot s$$

where

- $\bar{y}$ is the sample mean
- s is the standard error
- n is the sample size
- $c_{(\text{level, n})}$ is a critical value that changes depending on the interval type and a specified confidence level.

Confidence Intervals

Definition

A confidence interval (CI) is an estimated range of values constructed using a sample drawn from a population so that, if we repeated the sampling method and CI estimation an infinite number of times, such intervals would contain the true parameter value the 100(level)% of the time. In layman's term, a confidence interval is a calculated range of values based on sampled data where the true population parameter (e.g., mean) likely resides.

Questions

Some sample questions that may require the calculation of a confidence interval for the mean:

- What is the average performance of my system at a specific condition?
- Is the average performance of the system below/above the specification limits?

Case Study

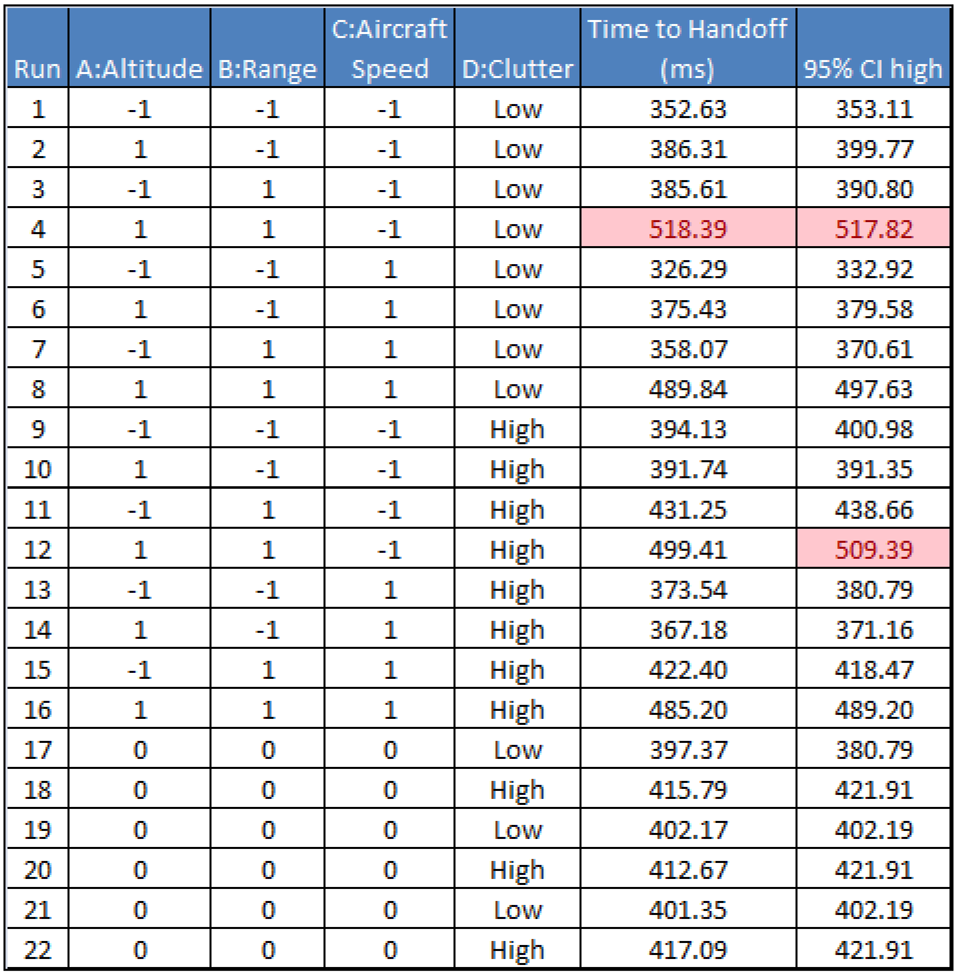
Confidence intervals are used in hypothesis testing and statistical inference. For our MWS example, let's say the null and alternate hypothesis are as follows:

$$H_0: \mu_{time} \geq 500$$
$$H_1: \mu_{time} < 500$$

In this case, we are assuming that the system is bad ($\mu_{time} \geq 500$) and want to find evidence that the system is good ($\mu_{time} < 500$). The first step in constructing the interval is to set the confidence level $(1 - \alpha)$, where α is the acceptable risk level for making the wrong conclusion that the system is good when it is actually bad (i.e., rejecting the null hypothesis when the null hypothesis is actually true). This degree of certainty must be specified up front and prior to testing. Based on the MWS problem statement, the confidence level is set to 95% ($\alpha = 0.05$). In other words, there is a 5% probability that we will say the system is good when in fact it is bad purely by chance.

Table 2 shows the results of the data for the MWS test along with the 95% upper confidence bound for each test condition. We can see that the 95% upper confidence bound for runs 4 and 12 exceeds 500ms. This suggests that there is evidence to not reject the null hypothesis (i.e. the true mean of the population could be over 500ms when altitude and range are both at the high level and aircraft speed is at the low level).

Table 2: Calculated upper confidence intervals for MWS designed experiment.



Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high
1	-1	-1	-1	Low	352.63	353.11
2	1	-1	-1	Low	386.31	399.77
3	-1	1	-1	Low	385.61	390.80
4	1	1	-1	Low	518.39	517.82
5	-1	-1	1	Low	326.29	332.92
6	1	-1	1	Low	375.43	379.58
7	-1	1	1	Low	358.07	370.61
8	1	1	1	Low	489.84	497.63
9	-1	-1	-1	High	394.13	400.98
10	1	-1	-1	High	391.74	391.35
11	-1	1	-1	High	431.25	438.66
12	1	1	-1	High	499.41	509.39
13	-1	-1	1	High	373.54	380.79
14	1	-1	1	High	367.18	371.16
15	-1	1	1	High	422.40	418.47
16	1	1	1	High	485.20	489.20
17	0	0	0	Low	397.37	380.79
18	0	0	0	High	415.79	421.91
19	0	0	0	Low	402.17	402.19
20	0	0	0	High	412.67	421.91
21	0	0	0	Low	401.35	402.19
22	0	0	0	High	417.09	421.91

Things to note

Important things to note and remember about CIs on the mean:

- *It does not tell you the true population mean.*

The CI tells you about the likely location of the true population mean.

- *It does not tell you the <u>probability</u> that the true mean will be between the estimated confidence limits.*

This is perhaps the most common misunderstanding regarding CIs. The interval describes the uncertainty associated with the sampling method, not the parameter. For example, let's assume that under a particular scenario (run) the true population mean of MWS handoff time is 400ms. Figure 3 shows 95% confidence intervals for 100 samples (20 observations per sample) of the MWS handoff time (vertical lines). The black horizontal line represents the true population mean handoff time. You can see that 96 of 100 (approx. 95%) samples yield confidence intervals that cover the true population mean of 400ms. The four red lines indicate samples where the estimated CI did not cover the true population mean.

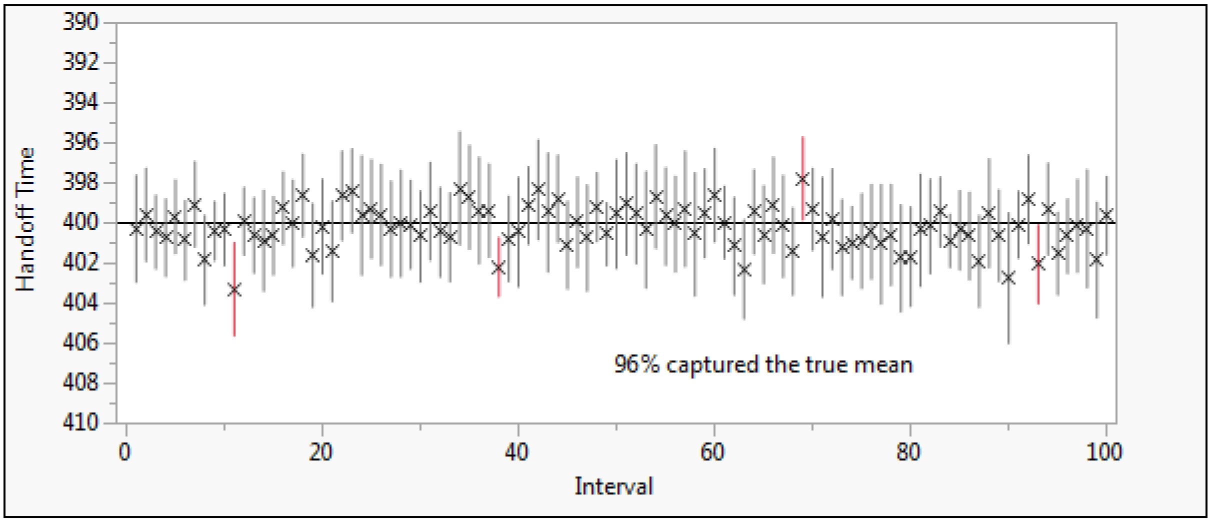


Figure 3: 95% confidence intervals for 100 samples $(n = 20)$ from the population of MWS handoff times

We cannot ever know whether the interval we calculate is one of the intervals that contains the true value of the parameter or one of the intervals that does not.

- *CIs do not allow you to predict future sample points from the population.*

Confidence intervals take into account the variation in the estimation (sampling error) but not in the response (standard deviation). In the next section, we cover prediction intervals which do encompass the variation in the estimation and in the response, thus allowing us to predict future sample points.

- *CIs do not tell you that a certain percentage of the population is between your limits.*

Again, because the confidence interval does not encompass the variation in the response, we cannot determine if 90%, 95%, etc. of the population will fall below a threshold specification (e.g. 500ms). Later in this paper, we introduce tolerance intervals which will allow us to do just that.

- *The more data in your sample, the smaller your confidence interval is for the stated parameters.*

As you increase the sample size, the sampling error decreases. If we were to sample the entire population, the sampling error would be zero and we would know the true mean of the system under test. Thus taking larger samples gives us a sampling error closer to zero, which narrows the confidence interval calculated.

Prediction intervals

Definition

A prediction interval (PI) is an estimated range of values in which future observations will fall at a specified level given what has already been observed. In layman's term, a PI gives you a range of values you can expect the response at a future tested or untested scenario. PIs are often used in regression analysis, where the intent could be to create an empirical model that will interpolate within the design space and estimate untested scenarios (i.e., settings of factors).

Questions

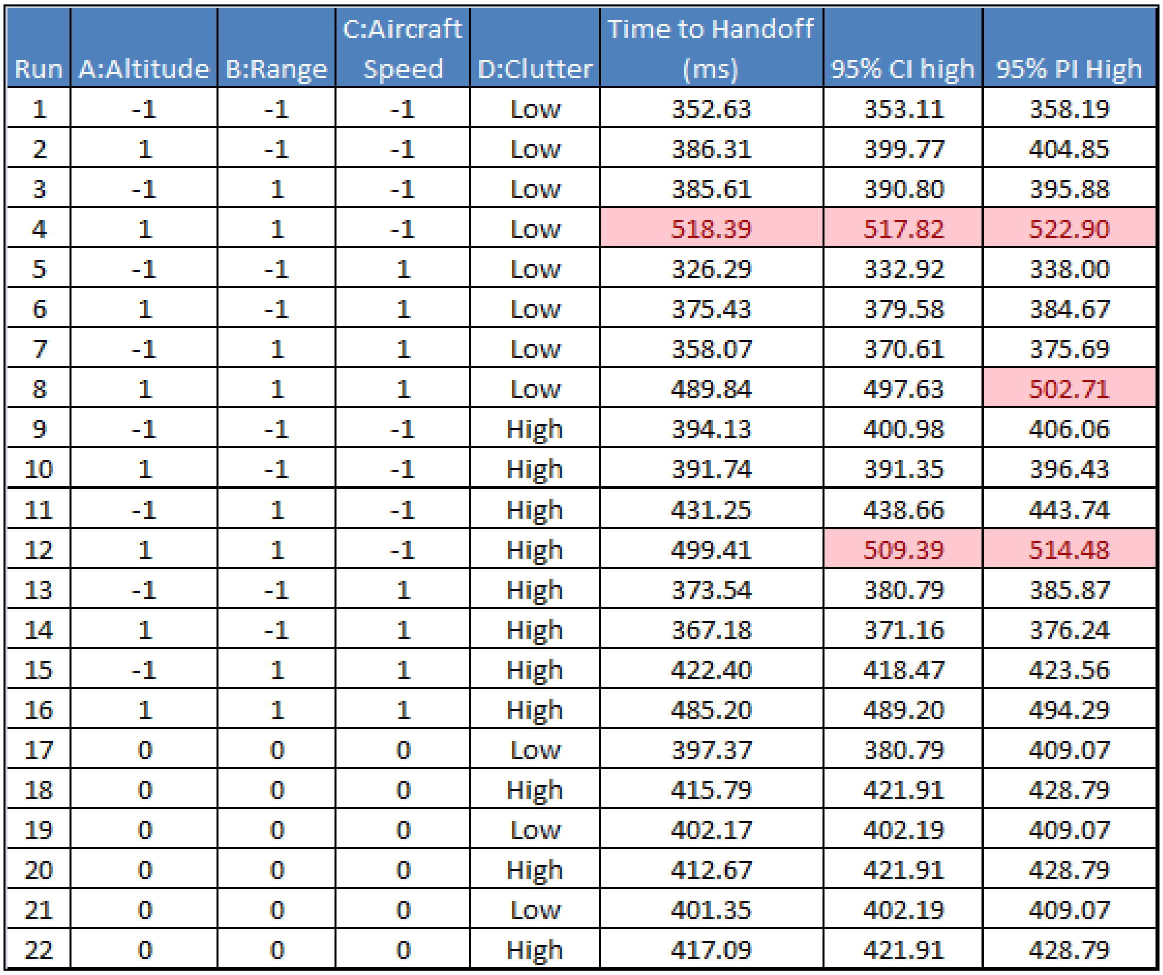
A sample question that may require the calculation of a prediction interval is:

- What is the expected (predicted) performance of my system at a specific condition?

Case Study

PIs encompass both the variation in the estimation and in the response. Therefore, PIs tend to be wider than confidence intervals. Table 3 shows the results for the MWS case study with a new column of the 95% prediction intervals. You can see runs 4, 8, and 12 all have a value greater than 500ms (highlighted in red in Table 3). These results suggest that while the true mean could be under 500ms at run 8, the response variation can lead us to see values that exceed 500ms.

Table 3: Calculated upper prediction intervals for MWS designed experiment.



Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high	95% PI High
1	-1	-1	-1	Low	352.63	353.11	358.19
2	1	-1	-1	Low	386.31	399.77	404.85
3	-1	1	-1	Low	385.61	390.80	395.88
4	1	1	-1	Low	518.39	517.82	522.90
5	-1	-1	1	Low	326.29	332.92	338.00
6	1	-1	1	Low	375.43	379.58	384.67
7	-1	1	1	Low	358.07	370.61	375.69
8	1	1	1	Low	489.84	497.63	502.71
9	-1	-1	-1	High	394.13	400.98	406.06
10	1	-1	-1	High	391.74	391.35	396.43
11	-1	1	-1	High	431.25	438.66	443.74
12	1	1	-1	High	499.41	509.39	514.48
13	-1	-1	1	High	373.54	380.79	385.87
14	1	-1	1	High	367.18	371.16	376.24
15	-1	1	1	High	422.40	418.47	423.56
16	1	1	1	High	485.20	489.20	494.29
17	0	0	0	Low	397.37	380.79	409.07
18	0	0	0	High	415.79	421.91	428.79
19	0	0	0	Low	402.17	402.19	409.07
20	0	0	0	High	412.67	421.91	428.79
21	0	0	0	Low	401.35	402.19	409.07
22	0	0	0	High	417.09	421.91	428.79

Things to note

- *Prediction intervals assume normality*

If the data collected does not follow the normal distribution, the interval reported is not appropriate. Diagnostic plots and tests for normality should be conducted to ensure this assumption is not violated. If there is a violation of normality, a transformation of the response could be employed such that the transformed response is normal. However, interpretation of results can be difficult (since it is on a transformed scaled and not in the real-world scale) and the PIs can be inflated (Perry, 2015).

Tolerance Intervals

Definition

A tolerance interval is a statistical interval within which, with some confidence level, a specified proportion of the population falls. In layman's terms, a tolerance interval will give you a range of values where X% (specified by the user) of the population should fall. Tolerance intervals are not as well-known compared to prediction and confidence intervals and have been underutilized in DoD testing (Rucker, 2014).

Questions

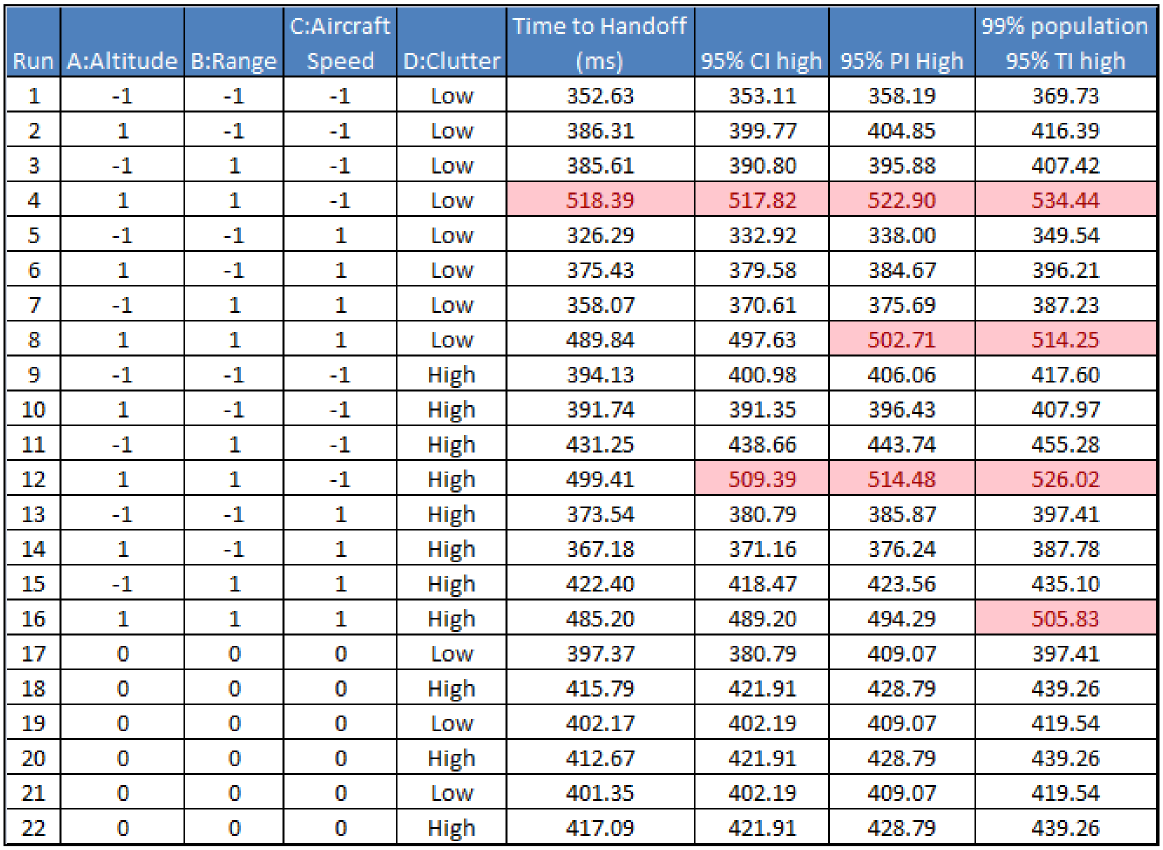
A sample question that may require the calculation of a tolerance interval:

- Will 99% of my observations fall under the threshold specification at least 95% of the time?

Case Study

A column for the 95% confidence/99% tolerance intervals for MWS case study data has been added in Table 4. You can see that now run 16 has a value greater than 500ms.

Table 4: Calculated upper tolerance intervals for MWS designed experiment.



Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high	95% PI High	99% population 95% TI high
1	-1	-1	-1	Low	352.63	353.11	358.19	369.73
2	1	-1	-1	Low	386.31	399.77	404.85	416.39
3	-1	1	-1	Low	385.61	390.80	395.88	407.42
4	1	1	-1	Low	518.39	517.82	522.90	534.44
5	-1	-1	1	Low	326.29	332.92	338.00	349.54
6	1	-1	1	Low	375.43	379.58	384.67	396.21
7	-1	1	1	Low	358.07	370.61	375.69	387.23
8	1	1	1	Low	489.84	497.63	502.71	514.25
9	-1	-1	-1	High	394.13	400.98	406.06	417.60
10	1	-1	-1	High	391.74	391.35	396.43	407.97
11	-1	1	-1	High	431.25	438.66	443.74	455.28
12	1	1	-1	High	499.41	509.39	514.48	526.02
13	-1	-1	1	High	373.54	380.79	385.87	397.41
14	1	-1	1	High	367.18	371.16	376.24	387.78
15	-1	1	1	High	422.40	418.47	423.56	435.10
16	1	1	1	High	485.20	489.20	494.29	505.83
17	0	0	0	Low	397.37	380.79	409.07	397.41
18	0	0	0	High	415.79	421.91	428.79	439.26
19	0	0	0	Low	402.17	402.19	409.07	419.54
20	0	0	0	High	412.67	421.91	428.79	439.26
21	0	0	0	Low	401.35	402.19	409.07	419.54
22	0	0	0	High	417.09	421.91	428.79	439.26

We can assume that runs 4, 8, 12, and 16 fail to meet our requirement that the probability of success (P_S) is at least 99%. Note that neither the CI nor the PI calculations were able to address this requirement directly. The TI is the only interval that tells us what scenario will result in failures more than 1% of the time. However, the TI does not provide an estimate for P_S. In order to get an estimate for P_S, the inverse of the TI needs to be found as shown in Table 5.

Table 5: Calculated upper bound for P_s for MWS designed experiment.

			C:Aircraft		Time to Handoff			99% population	
Run	A:Altitude	B:Range	Speed	D:Clutter	(ms)	95% CI high	95% PI High	95% TI high	% Below Spec
1	-1	-1	-1	Low	352.63	353.11	358.19	369.73	>99%
2	1	-1	-1	Low	386.31	399.77	404.85	416.39	>99%
3	-1	1	-1	Low	385.61	390.80	395.88	407.42	>99%
4	1	1	-1	Low	518.39	517.82	522.90	534.44	<0.1%
5	-1	-1	1	Low	326.29	332.92	338.00	349.54	>99%
6	1	-1	1	Low	375.43	379.58	384.67	396.21	>99%
7	-1	1	1	Low	358.07	370.61	375.69	387.23	>99%
8	1	1	1	Low	489.84	497.63	502.71	514.25	78.9%
9	-1	-1	-1	High	394.13	400.98	406.06	417.60	>99%
10	1	-1	-1	High	391.74	391.35	396.43	407.97	>99%
11	-1	1	-1	High	431.25	438.66	443.74	455.28	>99%
12	1	1	-1	High	499.41	509.39	514.48	526.02	11.1%
13	-1	-1	1	High	373.54	380.79	385.87	397.41	>99%
14	1	-1	1	High	367.18	371.16	376.24	387.78	>99%
15	-1	1	1	High	422.40	418.47	423.56	435.10	>99%
16	1	1	1	High	485.20	489.20	494.29	505.83	97.1%
17	0	0	0	Low	397.37	380.79	409.07	397.41	>99%
18	0	0	0	High	415.79	421.91	428.79	439.26	>99%
19	0	0	0	Low	402.17	402.19	409.07	419.54	>99%
20	0	0	0	High	412.67	421.91	428.79	439.26	>99%
21	0	0	0	Low	401.35	402.19	409.07	419.54	>99%
22	0	0	0	High	417.09	421.91	428.79	439.26	>99%

A column for the upper bound of P_S for MWS case study data has been added in table 5. You can see that for runs 4, 8, and 12 we fail to meet the spec by a large margin and run 16 fails as well although by a smaller margin

Things to note

- More sensitive to normality assumption violation.

Like the PI, the TI also requires that the data be normal distributed. However, if diagnostic plots and a test for normality indicates this assumption has been violated, a transformation of the response is not recommended. Rather, a distribution-free (nonparametric) calculation of a tolerance interval should be employed (see Natrella [1963] for details).

Conclusion

This best practice has demonstrated how to use three statistical intervals in conjunction with design of experiments and regression analysis to address the underlying questions behind testing. The combination of these tools allows programs to assess a system's performance with greater rigor than the general practice of reporting a single average of a system's performance across all test conditions. Design of experiments helps define the operational space and helps determines which scenarios (settings of the input factors) should be run that would best aid the analysis. Regression analysis allows us to build an empirical model that informs us which input factor or combination of input factors influences performance and by how much. The empirical model created with regression analysis can be used to predict performance for future untested scenarios.

Statistical intervals help quantify the level of uncertainty in our system performance estimates. The appropriate statistical interval to use is dependent on the question that is being asked. Tolerance intervals are perhaps the best suited for many DoD applications but are currently underutilized in the T&E community. The end results from testing and analysis must aid senior leaders (the decision makers). The combined use of these tools provides a rigorous examination of a system's performance to achieve just that.

References

Anderson-Cook, Christine M. "Interval Training: Answering the right question with the right interval." *Quality Progress* vol. 42, no. 10, 2009, pp. 58-59.

De Gryze, Steven, Ivan Langhans, and Martina Vandebroek. "Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression." *Chemometrics and Intelligent Laboratory Systems* vol. 87, no. 2, 2007, pp. 147-154.

Hahn, G. J., & Meeker, W. Q. *Statistical intervals: a guide for practitioners*. John Wiley & Sons, Inc., 2011.

Gilmore, J.M. "FY 2012 Annual Report." Director, Operational Test and Evaluation, 2012.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017.

Natrella, Mary Gibbons. Experimental Statistics, National Bureau of Standards Handbook 91, US Department of Commerce, 1963.

Perry, M.B. and Walker, M.L. "A prediction interval estimator for the original response when using box-cox transformations." *Journal of Quality Technology*, vol. 47, no. 3, 2015, pp. 278-297.

Rucker, "Improving statistical rigor in defense test and evaluation: use of tolerance intervals in designed experiments." *Defense Acquisition Research Journal*, vol. 21, no. 4, 2014.