Understanding the Signal to Noise Ratio in Design of Experiments

Authored by: Aaron Ramert

31 August 2019

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

When considering test designs for stochastic systems, it is common to refer to the signal-to-noise ratio (SNR) and assess its effect on other design metrics. The signal is the difference in response values the tester desires to detect, and the noise is the natural variation within the (stochastic) system. For initial planning, it is often more productive to independently discuss and assess the two components that make up this ratio as the "difference to detect" and the "system noise." Each serves to address a specific need or property of the test and is derived by different methods. The methods and reasons to adjust each is also different. If this cannot be done due to a lack of data or understanding of the system, then combining these two measures into the SNR facilitates an understanding of the precision of the test. By following this method, the test planning process can be accomplished in a manner that best addresses the objectives.

Keywords: noise, variance, signal, stochastic, design evaluation, power

Introduction

An important property of a well-designed test is that it has high power to detect important effects on the response. When designing a test the power is only estimated with calculations involving other design properties. Some of the easiest values to manipulate when interacting with statistical software are the *signal* and the *noise*. The signal, or more specifically the significant difference-to-detect (δ) is the magnitude of the change in response the tester is seeking to detect when a factor changes levels. The *noise* (σ) is the natural variation that occurs within a stochastic system because of uncontrolled and often unknown changes to inputs (factors), variation to the system in operation, and inaccurate output measurements. With some small adjustments to δ, σ, or both, the power of the test can be easily manipulated. These metrics should not be adjusted in an ad hoc manner. Instead, they should both be thoroughly considered in test planning, long before any specific test designs are created and before any software is accessed, in order to determine the most accurate and relevant values for each. From that point on, these values should only be changed contingent upon a new understanding of the underlying system or the magnitude of the measured response.

Furthermore, it is also quite common to link these two metrics together and refer to them as the signal-to-noise ratio (SNR or δ/σ). This best practice explores the purpose of each metric and why it is often best to discuss them separately, especially in the test planning phase. Later, when comparing tests, SNR as a single metric gives testers information on how the effect of a factor under study is the measured in relation to the natural system variation.

Background

A Good Test Begins with Deliberate Test Planning

Prior to addressing SNR or any test metrics, the test team must define the test objectives. These objectives should be specific, unbiased, measurable, and of practical consequence (Montgomery and Coleman, 1993). With the objectives in mind, the test team can then determine the response(s) to measure and the factors to vary. Only then should the test team begin to design the test. This planning process is explained in more detail in the STAT COE "Guide to Developing an Effective STAT Test Strategy," referred to in this document as the Test Planning Guide (TPG). When designing a test, the test team will consider five interrelated test design metrics; confidence, power, sample size, difference to detect, and response noise. All of these are described in the following section and Table 1 provides a quick reference for their purpose and basic interactions. The information in Table 1 is based on the following null and two-sided alternative statistical hypotheses.

$$H_0: \mu_1 = \mu_2, \text{Changing factor levels has no influence on the response}$$
$$H_A: \mu_1 \neq \mu_2, \text{Changing factor levels does have an influence on the response}$$

If you are unfamiliar with statistical hypothesis testing, then it is recommended that you read, "Statistical Hypothesis Testing," a STAT COE best practice on the subject. In this test, the null hypothesis, which is only rejected in light of significant statistical evidence to the contrary, is H_0. The alternative hypothesis is H_A. When considering the five components in Table 1 it is also necessary to understand what α and β represent. A type I error in hypothesis testing is rejecting the null hypothesis when it is true (and H_0 should not be rejected). The probability of a type I error is symbolized by α. A type II error in hypothesis testing is failing to reject the null hypothesis when the null hypothesis is not true (and H_0 should be rejected). The probability of a type II error is symbolized by β. Test designers seek to minimize both α and β, however, "unfortunately, α and β work in opposition to one another" (Kensler, 2018).

Table 1: Five Test Components and their Effects

Component	Effect of Increase	Effect of Decrease
Number of runs (n)	More runs is better! When n increases the power of the test increases.	With fewer runs (and no other changes) the power of the test will decrease.
Difference-to-Detect (δ)	When δ increases (and all else equal) the power increases	When δ decreases (and all else equal) the power decreases
System Noise (σ)	If the system has a high level of noise, it is more difficult to determine which portion of change in the response is due to the factor change and which portion is from noise.	A lower level of noise in a system makes it easier to determine the change in response due to a change in factor levels. This results in higher power and could lead to a smaller test.
Confidence = $(1-\alpha)$ = $P(\text{fail to reject } H_0 \mid H_0 \text{ is true})$	In cases where changing factor levels does not influence the response (H_0 is true) there is a higher probability we correctly conclude there is no influence.	In cases where changing factor levels does not influence the response (H_0 is true) there is a higher probability that we **incorrectly** conclude that it does have influence.

Power = $(1-\beta)$ = $P(\text{reject } H_0 \mid H_0 \text{ is false})$	In cases where changing factor levels does have influence on the response (H_0 is false) there is a higher probability that we correctly conclude there is influence.	In cases where changing factor levels does have influence on the response (H_0 is false) there is a higher probability we **incorrectly** conclude there is no influence.

Five Components for Test Design Evaluation

Creating a designed test is an iterative process that seeks to find the best design for achieving the test objectives while balancing five metrics against each other. In most cases, the initial design produced will have to be altered and it is necessary to understand what characteristics each metric reveals about the test and how they affect each other.

Sample size is annotated as n. In nearly every application of statistics "more is better" when it comes to sample size. More runs in the test will not reduce the noise in the system under test (SUT), but they will provide a more accurate estimate of the noise. The result is a test with higher power. More runs also allow the testers to explore more of the design space and to evaluate more factors and interactions which affect the SUT.

Signal is what is measured in the test and is more formally referred to as the practical difference-to-detect. The difference-to-detect is denoted with δ and is set by the test team prior to designing the test at a level that minimizes testing required but fully supports the objectives. In general, the larger the difference to detect, the smaller the required test. This is because it is easier to detect large changes in the response compared to small changes.

Every stochastic system has some level of inherent variation, often referred to as noise. It is common practice to define the amount of variation in a system by the standard deviation (σ). The noise comes from small (and often imperceptible) changes in to inputs and to the SUT from uncontrollable conditions and from output measurement errors. In design of experiments (DOE), σ represents the inherent errors we cannot affect or attribute to a known source, such as variation of the response due to weather conditions across a flight path. In some cases, this is because the design factor cannot be controlled or measured with sufficient precision. In nearly all cases there are other input factors, both known and unknown to the testers, which change from run to run. These unintended and often uncontrollable changes to the input affect the SUT and create noise. The noise manifests itself as changes in the response variable with no apparent changes to the factors. As the TPG notes, randomization is essential to good testing and is the best way to guard against the uncontrollable factor effects biasing the results. A lower noise level makes it easier to design the test. In general, we expect less noise in the lab, where conditions are tightly controlled, and more noise in an operational environment.

Confidence is the probability of failing to reject the null hypothesis when it is true; which is in fact the correct decision. It is symbolized with $(1-\alpha)$, where α is the significance level and the probability of a *false positive* or Type I error, and is set by the test team prior to testing. This definition can also be stated as the probability we do not make a type I error and can be visualized in the upper left decision

quadrant of Table 2. In the context of DOE, confidence is the probability that you conclude a factor is not significant given that it truly is not. Confidence is set by the test team during test design and prior to test execution and typically ranges from 0.9 to 0.95, but can be as low as 0.8 and as high as 0.99. The goal is to set confidence as high as possible without creating an excessive penalty on the other design metrics. In some cases the upper limits of other metrics will limit the confidence a test can have.

Power is usually the most focused upon component when designing and comparing tests. According to *The Cambridge Dictionary of Statistics*, power gives, "a method of discriminating between competing tests of the same hypothesis. It is also the basis of a procedure for estimating the sample size needed to detect an effect of a particular magnitude." Power is symbolized with $(1-\beta)$, where β is the probability of a type II error. We can therefore think of power as the probability that we do not make a type II error. In a DOE context power is the probability that we correctly conclude a factor is significant. This outcome can be visualized in the lower right decision quadrant of Table 2. A common minimum value for power in DoD testing is 0.8.

Table 2: A comparison of the truth, the decision, and their relation to Power and Confidence

		Decision	
		Fail to reject H_0 Perception is, "Factor is not significant"	Reject H_0 Perception is, "Factor is significant"
Truth	H_0 is true "Factor is not significant"	Correct 1-α	Type I error α
	H_A is true "Factor is significant"	Type II error β	Correct 1-β

Proper SNR Starts in Test Planning

The Right Test Team

A key ingredient to a successful test is thorough planning by the test team. A successful test team needs to have a diversity of experience so the SUT, its desired impact, and all the test options can be accurately explored. The technical experts should understand the science and engineering principles behind the system and know the predicted behavior throughout the design space. System operators are expected to know how the final product is to be used and which characteristics are important for mission accomplishment. Test practitioners from the lab or range will know what equipment is available for testing, its sensitivity, and how it can be used to measure the responses of the SUT. They can also advise the test team if range or lab limitations present test constraints. The test team also requires program leadership so the team will have an understanding of the resources available and a conduit if more resources are requested. Finally, a STAT expert is required to properly frame each test and create a specific design.

Determine a Value for the Difference-to-Detect

The test team will have to examine the SUT and determine what magnitude of change in the response corresponds to a practical change in the performance of the system. A practical change is meaningful in a mission context and should elicit a response from the program. This specific amount of change could come from the requirements, but rarely does. The magnitude should be the smallest one that supports the test objective, but can still be measured. When considering the test objective, the question the test team must ask is, "What is the smallest change in response that we can record and conclude that a factor influences the SUT in a meaningful way?"

To answer this question the test team needs two pieces of information. First, they need to understand how precisely and accurately the responses can be measured. Test lab and range personnel will often be able to assist because they have experience with the equipment that will be used and understand its strengths and limitations. However, they can also fall into the trap of assuming that the way testing was done in a previous test is the only way it can be done. The test team may explore new ways of measuring the response to achieve a desired δ.

To answer the second part of the question, concluding a change in response demonstrates a cause and effect relationship with the factor, the test team needs to refer back to the test objective and determine what magnitude in response changes must be measured. This is not a trivial task and will nearly always involve discussions with program leadership, where decision makers can decide what a *meaningful* change is.

For example, suppose the military wants to increase the fuel economy of a logistics vehicle by 5% and a contractor proposes a fuel additive. If the vehicle can go 500 miles on a tank of fuel, it should now be able to go 525 miles. The factor, fuel, would have two levels; $x_1 = \text{with additive}$, and $x_2 = \text{without additive}$. The difference-to-detect is $\delta = 25$ miles. What if the vehicle were a small one that moved around a freight yard and had a maximum range of 10 miles? Now the 5% increase only equates to 10.5 miles and $\delta = 0.5$ miles.

Determine an Estimate of System Noise

Every stochastic system will exhibit some degree of variation in the results when all inputs remain the same. This further complicates the results from a test because the tester does not immediately know how much of the change in response was due to factor manipulation and how much was due to the natural system variation, or noise. In fact, during test planning the power of the test depends on the amount of noise. The following discussion describes specific methods for estimating noise. The assumption throughout is that the responses are normally distributed because that is also assumed for all power calculations.

The best source of information on σ is the root mean square error (RMSE) from a prior designed test or experiment. At the conclusion of a designed test, it is common to perform analysis of variance (ANOVA) to determine if the factor changes resulted in meaningful response changes. One of the metrics often displayed with the test results in design software is the root mean squared error (RMSE), which is the

best approximation of σ. The STAT COE best practice, "Understanding Analysis of Variance," is recommended for readers who desire more information on ANOVA (Natoli, 2018).

If results from a previous designed test are not available it can also be helpful to look at a similar system or test. This method requires an additional step that induces more imprecision. When the RMSE has been calculated the testers need subject matter expert (SME) input to make an adjusted RMSE, which should better estimate σ in the new SUT.

Observed data from the SUT can be used if there are no prior test results from a designed experiment. In an observed population a common measure of the amount of variation is sample standard deviation (s), which is calculated similarly to RMSE. Equation (1) details how to calculate s where the numerator is the cumulative squared deviation and the denominator is the degrees of freedom.

$$\sigma \approx s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{1}$$

In the absence of empirical data, testers are forced to rely on expert opinion. It is known that 95% of the data fall between $\pm 1.96\sigma$ of the mean in a normal distribution. The empirical rule generalizes this property and states that approximately 95% of the data will fall within two standard deviations of the mean of an approximately normal distribution (Wackerly, 2008). System experts, with experience testing similar systems and a sound understanding of the engineering within the SUT, can derive an expected minimum and maximum response value. Recall that this is the minimum and maximum response when all of the factors are held constant. Equation (2) shows how the range of data can then be divided by four to approximate one standard deviation. Range is simply $Y_{max} - Y_{min}$. This is due to the special properties of the range that provide an estimate for σ.

$$\sigma \approx \frac{\text{Range}}{4} \tag{2}$$

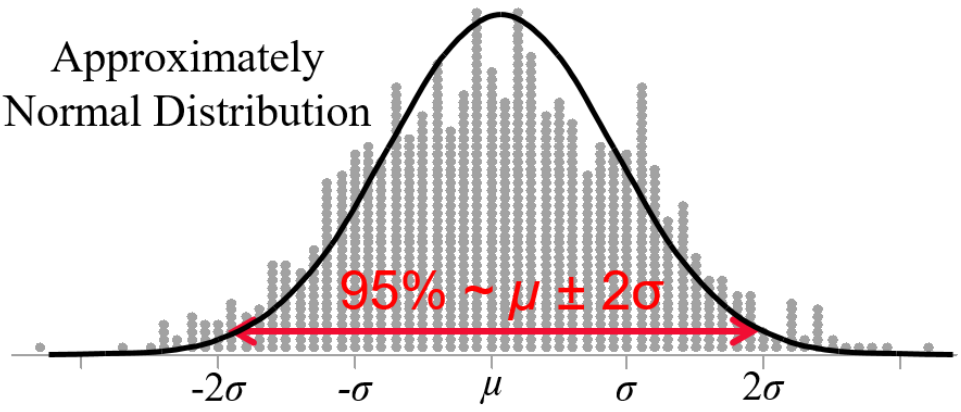

Figure 1: Illustration of the Empirical Rule

In his book, *Statistical Rules of Thumb*, Dr. Gerald van Belle presents Equation (3) as a method to bracket the standard deviation of a sample given only the range of data and number of samples. As noted earlier in this best practice, the sample standard deviation is a good estimator for σ. Even when Equation (3) produces some bounds for σ, someone needs to reduce it to a single point value for test planning.

$$\frac{\text{Range}}{\sqrt{2(n-1)}} \leq s \leq \frac{n}{n-1}\frac{\text{Range}}{2} \tag{3}$$

A study by the Institute for Defense Analysis found that the SNR for over 90% of observed effects in operational tests was less than 2 and 75% were less than 1.2 (Avery, 2014). Tests in wind tunnels and highly controlled laboratory environments can have a SNR of 6 or more (Landman et al., 2002). With this large discrepancy it is important that testers determine a reasonable value for their SUT.

The Effects of Signal and Noise Levels on Test Performance

The Effect of Noise on the Test Results

Noise in the SUT makes it more difficult to assess factor effects on the response from testing. The purpose of DOE is to assign causation of a measured change in the response to a measured change in a factor. Any imprecision in this measurement weakens the test and can only be overcome by either accepting the less precise response measure or by increasing the number of runs to overcome sampling variability. Figure 2 illustrates the effects of increasing the sample size (number of runs) while taking no action to decrease the noise. It is clear that as more results are recorded, the distribution of the sample mean becomes narrower and remains centered on the population mean; a better condition for testing. This is because of the central limit theorem (CLT) which states that independent and identically distributed random variables will tend to a normal distribution centered on the population mean and with the population variance (Montgomery, 2017). The CLT can be replaced with the simpler equation for its asymptotic properties, which preserves the relationship between the population variation, the observed variation, and sample size (Wackerly et al, 2008). Equation (4) defines this relationship within the CLT and shows that the observed variation ($\sigma_{\bar{y}}$) can be reduced by making the numerator smaller (less noise) or by making the denominator bigger (more runs).

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \tag{4}$$

Another method for reducing the noise is to introduce more control into the test. There are three basic areas to do this: input control, SUT operation, and response measurement. Precision is applied to the inputs by varying the factor levels in a systematic method and in the most accurate method possible. The maximum level of overall precision will depend not only on the method of adjustment to the SUT, but also the method of measuring the adjustment. This can only be done through close cooperation between the SMEs and the lab operators. Reducing the noise within the SUT is often not feasible.

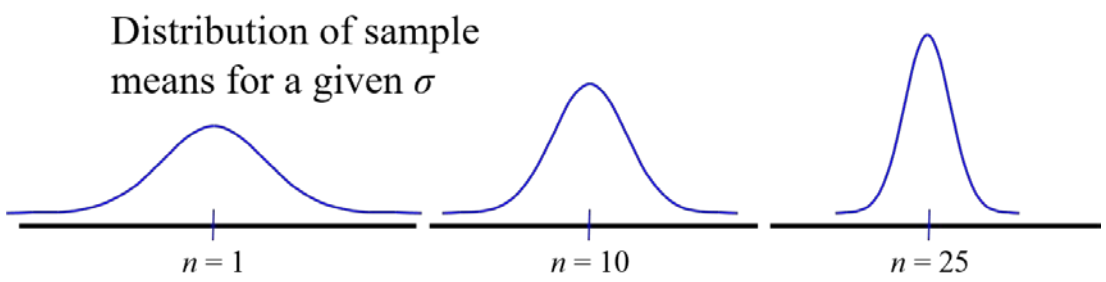

Figure 2: Distributions with the same standard deviation but increasing sample size

Much of the testing within the DoD is referred to a *black box* test where the tester can adjust inputs and record responses, but has no ability to adjust the SUT in any way. If the SUT can be adjusted, the SMEs should know the best way to do this. It is essential to document the changes to the test procedure and enforce adherence to it. The test team needs to keep their test objectives in mind when doing this. If the SUT is supposed to be production representative it may not be proper to give it a "tune up" for superior performance. Figure 3 looks similar to Figure 2, but the change that takes place is better precision, not increased sample size.

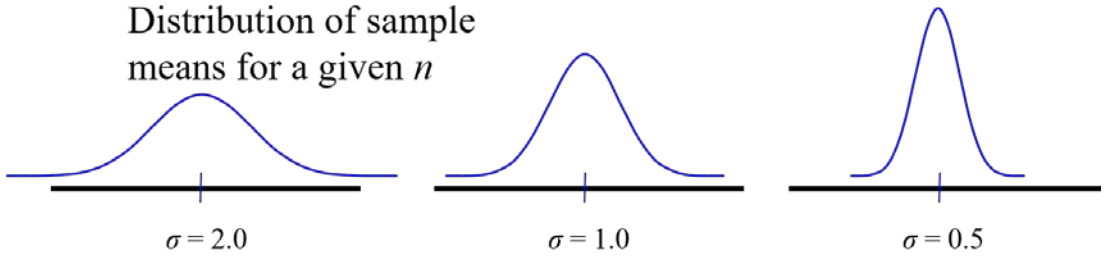

Figure 3: Distributions with the same sample size but decreasing standard deviation by better controlling test conditions

The Effect of Varying Difference-to-Detect on the Test Results

A small δ is more difficult to detect statistically, but yields more accurate results of the SUT. It is common for testers and program leadership to want more precision, which is understandable because it can lead to more knowledge about the SUT. This desired level of precision may not be practicable because δ is too small when compared to the noise. If one of the methods for reducing noise in the previous section does not work or the test cannot be executed as-is, then the δ will have to be altered. Figure 4 illustrates the interaction between δ and noise in a test where the factor is varied between -1 and 1, and the measured results, along with their distributions, are displayed along the *x*-axis. The worst case scenario is in the lower left corner. The small δ and high noise make it extremely difficult to determine if the response change was due to a factor change or the noise alone. The panel above this, in the upper left corner, has the same δ, but the noise is lower. This is better; it is also a difficult test because there is still overlap between the distributions. The best case scenario, with a large δ and low noise, is in the upper right corner. With almost no overlap it is clear that the change in response of that magnitude has to come from factor changes. The lower right corner is similar to the upper left, where

there is some separation between the mean responses, but not enough for the results to be clear without a very large sample size. It is also important to remember that at the start of the test the level of noise in the SUT is only an estimation.

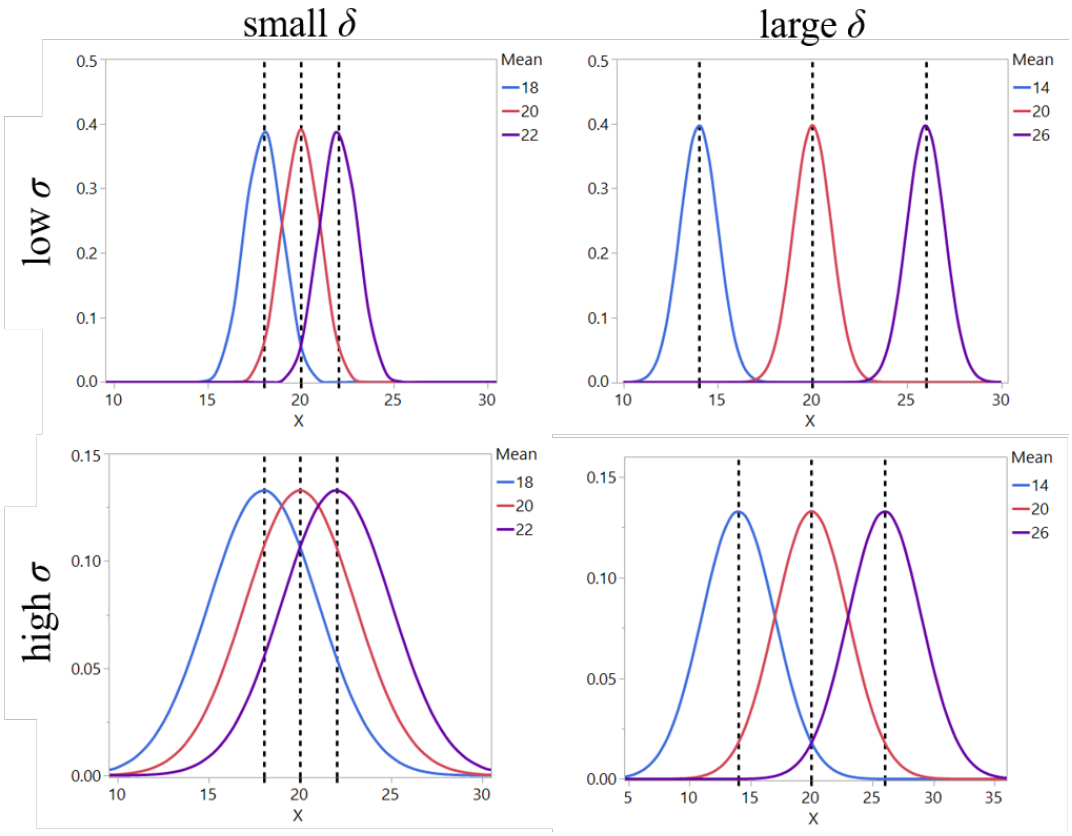

Figure 4. The interaction between δ and noise

How to Properly Adjust the SNR

When and How to Change the Difference to Detect

Changing δ is a last resort. When the STAT process is followed, δ is derived through careful examination of the information required to support an operationally meaningful objective. The test team should only change δ when forced to because the test is not executable. A good test plan can become impractical when the noise is higher than expected or budget adjustments no longer allow for the total runs necessary. Not all adjustments to the test plan will require δ to be adjusted, but some will. In a case where δ may be adjusted the test team needs to update the test design with the latest assumptions and reengage with leadership to determine if, in light of the assumption changes, a larger δ will be acceptable. In some cases the test team may present the arguments as, "Given our assumptions, the smallest δ that can be tested for is *Y*." Like other portions of the STAT process, this may also be an iterative process. When it is complete the result will be an inferior test, but one that is executable. It then becomes a program leadership decision to determine if the inferior test is an acceptable test.

When and How to Update the System Noise

The estimation of system noise should be updated whenever new information is gained. In the case of a sequential test strategy, which is recommended in the TPG, this is an expected step between test phases. The updated noise level should be entered into the planning software and the test reanalyzed in terms of estimated power. This may mean another round of adjusting other parameters. If the σ update is after a test, then the RMSE can be calculated and used in analysis to determine the actual power of the test. If the test team is following a sequential and progressive test strategy outlined in the TPG, then there will be multiple times throughout the test process to update noise levels based on the data produced.

Applying the SNR Concept to Binary Responses

All of the discussion in this best practice related to SNR has been predicated on having a stochastic SUT with a continuous response. However, there are methods for analyzing a binary response for the purpose of sizing the test. It is covered in detail in the best practice "Categorical Data in a Designed Experiment Part 2: Sizing with a Binary Response." In the TPG the STAT COE recommends that testers use continuous responses whenever possible because binary responses will require many more runs to get the same amount of information.

Conclusion

Although the term signal-to-noise ratio will continue to be used in DOE, it is helpful for disciplined test professionals to continue to refer to each portion separately in the test planning phase, where each term is a separate metric. This is easier when the difference-to-detect and the noise are understood for what they are and how they each affect the test planning and design process. They are not the only metrics to consider in test planning, but they each require deep understanding of the system, the test process, and the test objectives. Neither are easy to calculate, but they are critical to properly size a test to support the objectives.

References

Avery, Matt R. *Empirical Signal-to Noise Ratios from Operational Test Data.* Institute for Defense Analysis, https://www.ida.org/-/media/feature/publications/e/em/empirical-signal-to-noise-ratios-from-operational-test-data/d-5243.ashx, Aug. 2014

van Belle, Gerald. *Statistical Rules of Thumb*. 2nd ed., John Wiley & Sons Inc., 2008.

Coleman, David E., and Douglas C. Montgomery. "A Systematic Approach to Planning for a Designed Industrial Experiment." *Technometrics*, vol. 35, no. 1, 1993, pp. 1–27.

Everitt, B.S. *The Cambridge Dictionary of Statistics*. 2nd ed., Cambridge University Press, 2003

Kensler, Jennifer. "Statistical Hypothesis Testing" Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 27 Aug. 2018.

Landman, Drew et al. "Use of Designed Experiments in Wind Tunnel Testing of Performance Automobiles." *SAE Technical Paper Series* 111 (2002): 2339–2346.

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

Natoli, Cory. "Understanding Analysis of Variance" Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 24 Sep. 2018.

Ortiz, Francisco. "Categorical Data in a Designed Experiment Part 1: Avoiding Categorical Data" Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 23 Oct. 2018.

Ortiz, Francisco. "Categorical Data in a Designed Experiment Part 2: Sizing with a Binary Response" Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 19 Jul. 2014.

Wackerly, Dennis D., et al. Mathematical Statistics with Applications. 7th ed., Thomson Brooks/Cole, 2008.