Designing a Test for Prediction

Authored by: Kyle Kolsti, PhD

Kaitlyn Jones

2 March 2021

The HS CoBP core mission involves leading a consortium of government, industry, and academic experts to assess future homeland threats to inform strategic plans across nine DHS T&E capability areas.

Homeland Security Community of Best Practices
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

The fundamental questions of test design are how many test points to include and which test conditions should be chosen, given the constraints like time and budget. These questions are best addressed by clearly identifying the objective(s) of the test and the amount of risk the program can tolerate of making an incorrect decision based on the data.

This best practice provides a process to design a test where the precision of predictions of the mean response is the primary consideration. The method described here details the process of determining the maximum allowable width for a confidence interval based upon the program's risk tolerance. The intent is to allow testers to create and compare multiple test designs which will likely lead to acceptable confidence interval widths over sufficient areas of the design space.

Keywords: relative prediction variance, margin of error, fraction of design space, JMP

Introduction

"How many test runs do we need to do, and what should they be?" These fundamental questions of test design cannot be answered without knowing why the test is being conducted. Knowing this fact, a proper response to these questions would be more questions: "What do you need to know, and how well do you need to know it?" In other words, the test objectives and the risk tolerance of the program must be clearly stated.

Crafting effective test objectives is therefore the first step in this process (Truett, 2018). For this discussion we may think of two categories of test objectives: to explain or to predict, which is precisely the title of the (Shmueli, 2012) paper. As Shmueli defines them,

> *"…I define explaining as causal explanation and explanatory modeling as the use of statistical models for testing causal explanations." … "In summary, explanatory modeling refers here to the application of statistical models to data for testing causal hypotheses about theoretical constructs."*
>
> *"I define predictive modeling as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations." … "Predictions include point or interval predictions, prediction regions, predictive distributions, or rankings of new observations."*

It is important for the test team to understand which of these purposes is most applicable to the about-to-be-designed test. Sometimes the two are complementary, but they also may lead to substantially different designs. Designing a test in practice is usually iterative; as each candidate design is produced, a suite of metrics is generated. These metrics may then be compared among candidate designs so the "best" design can be chosen (Harman, 2014). The test purpose will point to which metrics are most pertinent.

Perhaps the most used metric is power. The general definition of power is "the probability of rejecting the null hypothesis when it is false" (Kensler & Cortes, 2014). It is intended to be of use only prior to collecting data, not as a post-test analysis tool (Johnson D. H., 1999). Power calculation requires the specification of a signal magnitude, δ, also called the effect size – a meaningful change in the response. It also requires an estimate of the standard deviation of the random aspects of the response not accounted for in the model, σ. Together these form the signal to noise ratio, according to which a test design will have a certain power (Ramert, 2019). In the context of Design of Experiments, power is most often applied to the parameters of the model: Is a factor significant or not? Did the drug increase the survival of the patients? Does the sensor work just as well during the day as at night? This application of statistical power has been called "effect power," where "Effect power is the probability of concluding that an effect impacts the response variable when it is truly active." (Johnson, Freeman, Simpson, & Anderson, 2018) The power calculations in JMP use a similar definition: "Power is the probability of detecting an active effect of a given size." (SAS Institute Inc., 2018).

These applications of the power metric are relevant, if not primary, when the goal is to explain. Going back to Shmueli's definition, the purpose is to test the theoretical construct of the model itself (whether terms are significant) and/or to test the hypotheses of causation (the improved survival rate may be attributed to the new drug) (Shmueli, 2012). But what if the users are mainly concerned about future performance, especially when it is already known which factors affect it?

Consider a requirement like "The effective range must be at least 3 statute miles" within a well-defined operational space. Physics and experience with similar technologies provide several meaningful factors. The test results not only will be used to evaluate the system against the requirement but will also feed a system-wide stochastic model. This scenario illustrates a purpose of prediction: based on the test results, what effective range should we expect after fielding and how precisely can we estimate it?

This paper proposes a process for designing a test based on the anticipated margin of error, M. The margin of error is the half-width of the confidence interval about the estimated mean, $\hat{y}$; so after testing, the results can be reported as the upper/lower confidence limit (LCL/UCL) form, "95% CI = [$\hat{y} - M$, $\hat{y} + M$]." The goal of the test design process would be to select the smallest suitable design that provides an acceptably small M in a sufficient portion of the operational space. Building on the previous example of effective range, the analysts could be tasked to provide a design with margins of error less than ± 0.3 miles.

This Best Practice will provide a step-by-step process for identifying the optimal test design for linear regression analysis based on margin of error. The margin of error will be calculated using another related metric, the relative prediction variance, which can be calculated mathematically or using statistical software like JMP. This Best Practice will demonstrate both ways to obtain it. Readers who are not comfortable with confidence interval calculation or interpretation are recommended to refer to (Kensler & Cortes, 2014). Likewise, readers not comfortable with Design of Experiments terminology like factors, factor interactions, and residuals may wish to refer to (Burke, et al., 2019) for additional background.

Method

Development of the Approach

This section will provide the conceptual and mathematical development of the proposed approach. See the next section for the abbreviated checklist.

At any point x_i in the factor space, the confidence interval following a test is

$$CI_i = \hat{y}_i \pm M_i \tag{1}$$

where $\hat{y}$ is the estimate of the mean and M is the margin of error. The estimate of the mean across the space will be generated using a linear data model in the form $y_i = \sum \beta_j x_{ij} + \varepsilon_i$, $i = 1..n, j = 1..K$, where the error term ε_i is normally distributed with zero mean and standard deviation σ. For a test that provided n observations and a data model that contains K parameters β, the margin of error may be calculated using the formula

$$M_i = t_{1-\alpha/2,\nu}\hat{\sigma}\sqrt{r_i} \tag{2}$$

where the number of degrees of freedom is $\nu = n - K$. The value $t_{1-\alpha/2,\nu}$ is the appropriate "Student's t" score corresponding to the confidence, where $\alpha = (100 - \text{confidence})/100$, and the degrees of freedom ν. The estimated standard deviation about the prediction at all data points is $\hat{\sigma}$; it is the sample-based estimate of the true standard deviation, σ.

The final variable in the equation, r, is the relative prediction variance. The relative prediction variance is the ratio of the variance of the standard error of the mean to the variance of the random part of the response. In other words, r tells us how precise our estimate of the mean is, given the spread of the data points. A lower value of r indicates a more precise estimate of $\hat{y}$ for a given sample $\hat{\sigma}$ and therefore a smaller margin of error. Note that r can vary across the factor space. Statistical software like JMP calculates relative prediction variance along with other metrics like power. Without access to such software, r can be calculated using matrix formulas provided in Appendix A.

Unfortunately, during the planning stage, we do not have any data for calculating $\hat{\sigma}$. It must be estimated. The estimate σ_{est} can be obtained by any justifiable means, including relevant historical data or simulation. Another method is to ask subject matter experts (SMEs) what range they would expect 95% of the data points to fall within, then divide that range by four (taking advantage of the fact that in a normal distribution, 95% of the area is within 2 standard deviations of the mean (Ramert, 2019)).

Recognize that $\hat{\sigma}$ is a random variable – even though the true standard deviation σ remains constant, each time the test is repeated the calculated $\hat{\sigma}$ will be different. It can be shown that after many tests, if

the guess σ_{est} were perfect (equal to σ), then $\hat{\sigma}$ would exceed σ_{est} approximately 45% of the time just by random chance. That means the estimated margin of error would also be larger than expected 45% of the time. If the goal is to limit M, a "factor of safety" must be added to σ_{est}. To add this cushion the design value σ_{des} will be used instead of σ_{est}. The design value may be obtained using the formula

$$\sigma_{des} = \sigma_{est}\sqrt{\frac{\chi^2_{1-\gamma,\nu}}{\nu}} \tag{3}$$

Define the tolerance as the percent chance of seeing a standard deviation less than the estimate (assuming the estimate is correct). Correspondingly, the probability of seeing a standard deviation greater than the estimate is γ, where $\gamma = (100 - \text{tolerance})/100$. Table 1 provides computed values for the ratio $\sigma_{des}/\sigma_{est}$ using this formula.

Table 1: Computed values for $\boldsymbol{\sigma_{des}/\sigma_{est} = \sqrt{\chi^2_{1-\gamma,\nu}/\nu}}$

Deg. of Freedom ν	Tolerance (γ)		
	80%	**90%**	**95%**
5	1.21	1.36	1.49
10	1.16	1.26	1.35
20	1.12	1.19	1.25
30	1.10	1.16	1.21
50	1.08	1.12	1.16
100	1.06	1.09	1.12
1000	1.02	1.03	1.04

The goal of the test is to see margins of error less than M_{max}, the maximum acceptable M. Equation 4 may be used wherever the margin of error is a concern to see if the anticipated precision is adequate.

$$M_i = t_{1-\alpha/2,\nu}\left(\sigma_{est}\sqrt{\frac{\chi^2_{1-\gamma,\nu}}{\nu}}\right)\sqrt{r_i} \tag{4}$$

Since statistics software often directly provides the relative prediction variance, it may be more convenient to use some algebra to find the formula for r_{max}, the maximum acceptable value for r. Note that r_{max} is the same everywhere in the factor space, while r may vary.

$$r_{max} = \left(\frac{\nu}{\chi^2_{1-\gamma,\nu}}\right)\left(\frac{M_{max}}{t_{1-\alpha/2,\nu}\sigma_{est}}\right)^2 \tag{5}$$

For a given test design and estimated data variability, the relative prediction variance r can therefore be compared to r_{max} to evaluate the test design.

Metrics for Design Comparison and Optimization

We introduce a useful metric for optimizing the design: $FDS_{r_{max}}$, defined as the Fraction of the Design Space ("FDS") within which $r \leq r_{max}$. FDS can be presented as a plot of relative prediction ratio as a function of the proportion of the factor space, a plot which JMP provides in its design evaluations (Anderson-Cook, Zahran, & Myers, 2003). $FDS_{r_{max}}$ is identified as the FDS value where the relative prediction ratio equals r_{max}. The metric is broadly applicable because the same plot is produced for complex multi-dimensional problems. Some special accommodations may have to be made for categorical variables. For example, if the problem involves two 2-level categorical factors, the factor space consists only of 4 discrete points (the corners of the square) – there is no interior within which to make predictions. The variance profile and FDS plots as constructed by a software product may interpolate using straight lines or step functions. Instead, the analyst could check for $r \leq r_{max}$ at each of the four points and report $FDS_{r_{max}}$ as 0, 0.25, 0.5, 0.75, or 1.0 accordingly.

The goal of the optimization process presented here is to find the smallest suitable test where $FDS_{r_{max}} = 1$. If this criterion is met, the margin of error will likely be within the desired limits throughout the factor space. (Of course, this statement and similar ones throughout this Best Practice carry the caveat that the estimate of the true standard deviation is correct or conservative.)

After Test is Executed

After the test is executed and the data are in hand, the estimate of the standard deviation may be updated using the standard deviation of the residuals,

$$\hat{\sigma} = \frac{SSE}{\nu} = \frac{1}{n-K}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{6}$$

Therefore, the confidence interval at any point r_k is

$$\hat{y}_k \pm M_k = \hat{y}_k \pm t_{1-\alpha/2,\nu}\hat{\sigma}\sqrt{r_k} \tag{7}$$

A final note about confidence intervals: it is important to understand how to interpret them and to avoid technically incorrect claims. Here is a succinct list of cautions and advice:

- If the confidence interval does not contain the requirement value, it may be claimed that "The evidence indicates that the true mean is above/below the requirement." (Equivalent to rejection of the null hypothesis that the true mean equals the requirement).

- If the confidence interval contains the requirement value, the test is inconclusive, meaning "The evidence is not sufficient to claim that the true mean is above/below the requirement." (Equivalent to a failure to reject the null hypothesis).
- The statement "There is a 95% chance the confidence interval contains the true mean" is only correct *before testing*. After testing, the resulting confidence interval either contains the true mean or it does not – it is unknown which is the case. Therefore, *after testing,* it is correct to state "The interval contains the true mean with 95% confidence."

Abbreviated Steps

Table 2 provides a succinct summary of the steps in the proposed process. Refer to the previous section or the following examples for elaboration and explanation.

Table 2: Abbreviated steps for designing a test for prediction

Phase	Step
Pre-planning decisions	1. Choose the maximum desired margin of error, M_{max} 2. Choose m prediction points x_p where M is to be managed 3. Choose the confidence level (typically 80%, 90%, or 95%) 4. Choose the tolerance (typically 80%, 90%, or 95%) 5. Choose a model, $y_i = \sum \beta_j x_{ji} + \varepsilon_i$ with K parameters 6. Estimate the standard deviation of ε, σ_{est}
Iterate to optimal design	7. Create a test design 8. Evaluate the test design at the points selected in Step 2 Option 1: Compare M (Eq. 4) to M_{max} Option 2: Compare r to r_{max} (Eq. 5); useful if M is not readily available but r is provided by statistical software 9. Repeat Steps 7-8 to find smallest suitable design
Test Execution	10. Execute the test
Analysis and Evaluation	11. Calculate CI at desired points using Eqs. 6 and 7 12. If evaluating the system against a requirement, compare the confidence intervals to the requirement

Examples

Example 1: No Factors (Single-Sample)

This first example is a simple one to demonstrate the process. Testing will occur at only one test condition, so there are no factors to consider. The test team needs to determine how many runs are necessary to achieve the desired precision in the estimate of the mean. The system also has a requirement that the mean must meet or exceed 12. Table 3 shows the pre-planning decisions made after a series of meetings.

Table 3: Single-sample example – Decisions made to inform test design

Phase	Step
Pre-planning decisions	1. $M_{max} = 1$ is the desired precision. 2. There is only one possible x_p … the lone test condition. 3. 95% confidence, so $\alpha = 0.05$ 4. 80%, tolerance, so $\gamma = 0.2$ 5. The model has only the intercept: $y_i = \beta_0 + \varepsilon_i$, so $K = 1$ 6. Experts estimate the data spread to be $\sigma_{est} = 1.5$

It can be shown that for this scenario with no factors, the relative prediction ratio is a simple function of sample size: $r = 1/n$ (see Appendix A for the derivation). Table 4 simulates the table an analyst may make in finding the optimal solution. The second-to-last column contains the relative prediction variance. It is useful here as there is only one point for calculation – for more complex problems, this column would have to be replaced by multiple columns or removed altogether. Also, because there is testing at only one point, $FDS_{r_{max}}$ will either be zero or one. In this exercise, the analyst started with guesses of $n = 10$ and $n = 20$ and correctly recognized that these tests bound the solution. The analyst then continued exploring the space to find that $n = 14$ is the smallest test where $M \leq M_{max}$ and therefore $FDS_{r_{max}} = 1$.

Table 4: Single-sample example – Iterations to arrive at optimum test design

n	ν	$t_{1-\alpha/2,\nu}$	$\sqrt{\chi^2_{1-\gamma,\nu}/\nu}$	r_{max}	r	M
10	9	2.262	1.166	0.064	0.100	1.25
13	12	2.179	1.148	0.071	0.077	1.04
14	**13**	**2.160**	**1.143**	**0.073**	**0.071**	**0.99**
15	14	2.145	1.139	0.075	0.067	0.95
20	19	2.093	1.122	0.081	0.050	0.79

Although power was not the driver of this test design, it is useful for insight to see the power of the selected test. Figure 1 depicts the power of the test to detect a range of δ's (signal) given $\sigma_{est} = 1.5$ (noise) as calculated using JMP. The results were confirmed by running simulations using the proposed method's confidence interval evaluation procedure, Eq. 6. Recall that the desired margin of error was $M_{max} = 1$. If the true mean is ± 1 from the hypothesized mean, there is a 63.6% chance this test will detect it. If the purpose of the test were different – for example, that 80% of many confidence intervals do not overlap the requirement if the true mean differs by ± 1 – the optimum test size would be $n = 20$. The caution here is that the oft-used guide of 80% power to detect a meaningful difference may not be achieved through this proposed process if M_{max} is thought of as that meaningful difference.

After the test is complete, the sample standard deviation $\hat{\sigma}$ may be calculated using Eq. 5. Because $r = 1/n$ for this example, Eq. 6 will take on the familiar classical confidence interval formula,

$$\hat{y} \pm t_{1-\frac{\alpha}{2},n-1}\frac{\hat{\sigma}}{\sqrt{n}} \tag{8}$$

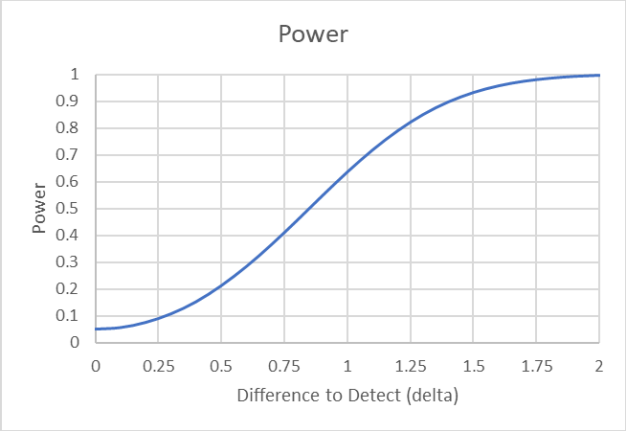


Figure 1: Power of the test design $n = 14$ with $\sigma = 1.5$ and 95% confidence

Example 2: One Factor

Suppose system response is acknowledged by consensus to be affected by a single continuous factor, x. Steps 1-9 have been completed: $M_{max} = 1$, x_p= 101 uniformly spaced points between -1 and 1 (the team would like to control M throughout the operational space), 90% confidence, 90% tolerance, and $\sigma_{est} = 0.6$. The experts say the response in the region to be tested is a mildly nonlinear function of x, so a quadratic model is chosen to capture any curvature. In this case, $K = 3$ since there are three βs.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{8}$$

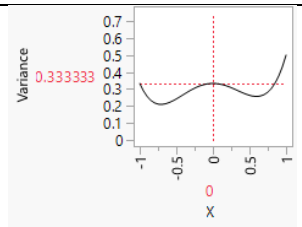
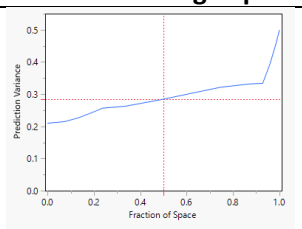
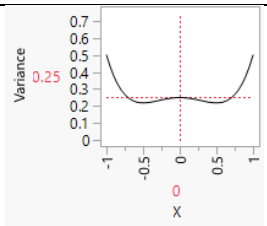
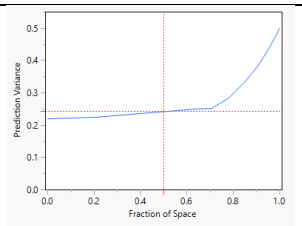
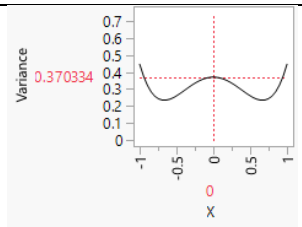
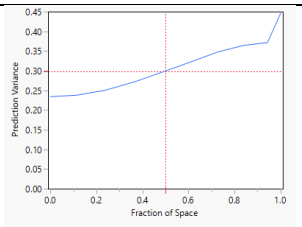
The program thinks funding is enough for up to eight runs, so the analyst begins there by creating three designs. Eq. 4 provides the maximum allowable relative prediction variance, $r_{max} = 0.370$.

> **Design 1**: "3-3-2". The analyst used JMP to create a D Optimal design. D optimality increases the effect power of a design, which is not the goal of this test. As background, the different types of optimality discussed in this best practice, D, I, and G, are well-defined and address distinctively different aspects of a test design (Goos & Jones, 2011) (Hardin & Sloane, 1993). This criterion tends to place more points at the edges of the factor space. This unbalanced design has three points at $x = -1$, three points at $x = 0$, and two points at $x = 1$ – hence the name "3-3-2"

Design 2: "2-4-2". The analyst used JMP to create an I optimal design. I optimality minimizes the integral of the relative prediction variance over the factor space (Goos & Jones, 2012). This approach is obviously relevant to the goal of this test. It tends to place more points in the interior to bring r down over a wider region of the factor space.

Design 3: "2-1-2-1-2". Noting a tradeoff between r at the center and at the ends of the factor space, the analyst utilized a design that had two points at each end and two points at the center, leaving two points $\pm\eta$ to be determined by setting $r_{x=0} = r_{max}$. The solution is $\eta = 0.62175$. This design has a lower maximum value of r in the factor space, making it the most G optimal of the three designs.

Which of these three designs is the "best" one? Diagnostic charts and metrics from JMP are shown in Figure 2. Figure 3 reproduces and overlays the Variance Profiles and adds equivalent plots for Margin of Error; as Eq. 2 shows, M is proportional to $\sqrt{r}$. Figure 3 was generated using the formulas in Appendix A. The Variance Profile plots depict the value of r across the space. These plots give the approximate shape of the margin of error across the space; The Fraction of Design Space plots depict the cumulative distribution of r across the space – it is from these plots that the values for $FDS_{r_{max}}$ were extracted.

Design	Variance Profile	Fraction of Design Space	Design Diagnostics
#1: 3-3-2			**Design Diagnostics** D Optimal Design D Efficiency 52.0021 G Efficiency 75 A Efficiency 34.61538 Average Variance of Prediction 0.288889 Design Creation Time (seconds) 11
#2: 2-4-2			**Design Diagnostics** I Optimal Design D Efficiency 50 G Efficiency 75 A Efficiency 37.5 Average Variance of Prediction 0.266667 Design Creation Time (seconds) 11
#3: 2-1-2-1-2			**Design Diagnostics** D Efficiency 47.65391 G Efficiency 83.96179 A Efficiency 29.5505 Average Variance of Prediction 0.303877 Design Creation Time (seconds) 0

Design

#1: 3-3-2

Run	X
1	0
2	-1
3	1
4	-1
5	1
6	0
7	-1
8	0

#2: 2-4-2

Run	X
1	0
2	0
3	1
4	0
5	-1
6	-1
7	0
8	1

#3: 2-1-2-1-2

Run	X
1	-1
2	-1
3	-0.62175
4	0
5	0
6	0.62175
7	1
8	1

Figure 2: JMP design diagnostic products for the three designs of a single-factor test

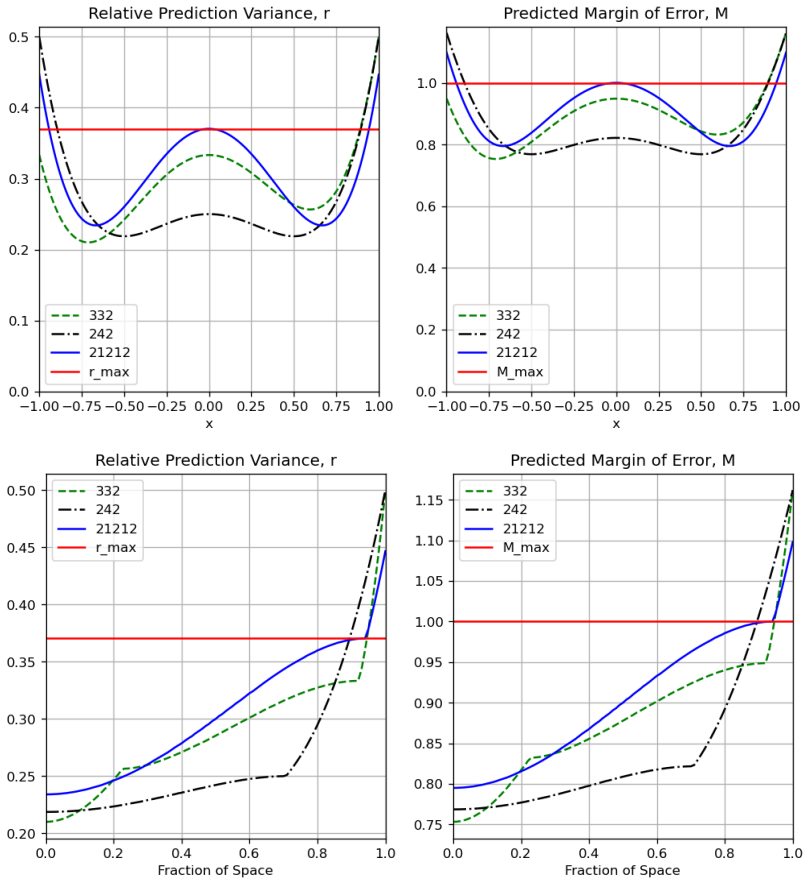


Figure 3: Relative Prediction Variance and Margin of Error for the three designs of a single-factor test

Table 5 illustrates a table which can be constructed for comparing the designs, following the format of (Harman, 2014). The "best" values of each metric are bolded. The I optimal Design #2 produces the lowest average r, as expected, but not by much. Interestingly the D optimal Design #1 has the highest $FDS_{r_{max}}$ and would be preferred based on that metric alone. More importantly, the analyst should be concerned that none of these designs meets the goal of $FDS_{r_{max}} = 1$. In other words, the resources allocated by the program are not likely to be adequate to provide the needed prediction precision throughout the factor space. The analyst should report this potential risk as well as the degree and the nature of the risk.

Table 5: One-factor example design metrics

Metric	**#1**	**#2**	**#3**
Avg. Variance of Prediction	0.289	**0.267**	0.304
$FDS_{r_{max}}$	**0.946**	0.894	0.943
r at center of space ($x = 0$)	0.33	**0.250**	0.370
Max value of r in the space	0.5	0.5	**0.447**
FDS plot, area of r > r_{max}	0.0065	0.013	**0.0041**
FDS plot, area of M > M_{max}	0.0083	0.016	**0.0054**

The Variance Profile plots are helpful in revealing the shortcomings and benefits of each design. For example, if the "heart of the envelope" ($x = 0$) is where the system will be most of its operational life, the I optimal Design #2 would be preferred. The D optimal Design #1 is unbalanced, so the margins of error will be larger at one end of the space then the other. Whether this behavior is adverse would depend again on the operational impacts across the space; the final design chosen may not be I-, D-, or G-optimal if another design provides a more appropriate solution for the problem at hand. (It should also be noted that an unbalanced design also introduces aliasing between factors and interactions, but for this test that may not matter.) The method proposed here is prescriptive to provide an actionable starting point – as always, however, judgment, user input, and understanding on the purpose of the test should inform the decisions.

Example 3: Multiple Factors (Design of Experiments)

This last example will illustrate the applicability of the proposed test design process for more complicated problems. Suppose system response is acknowledged by consensus to be affected by a mix of continuous and categorical factors. In this instance, we consider a system affected by three continuous factors, x_1, x_2, and x_3; and a single 3-level categorical factor called "CAT", as depicted in Figure 4.

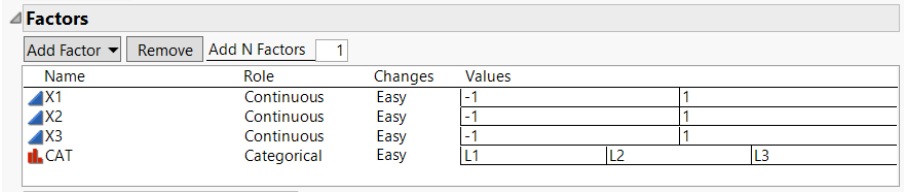


Figure 4: Factors affecting designed experiment

The test team follows the proposed process, making the following choices for the design, as shown in Table 6. We emphasize that Step 6, estimating the standard deviation, is a critical step which requires much consideration and research.

Table 6: Multi-factor example – Decisions made to inform test design

Planning Step choices made
1. $M_{max} = 1.0$
2. x_p = every point in the factor space
3. Confidence = 90%
4. Tolerance = 90%
5. (See Figure 5 below)
6. $\sigma_{est} = 0.4$

The experts design a response surface model to measure all model parameters. In this scenario, $K = 14$ since we wish to capture all main effects, two-factor interactions, and quadratic effects of the continuous factors, listed in Figure 5.

Model

Main Effects | Interactions | RSM | Cross | Powers | Remove Term

Name	Estimability
Intercept	Necessary
X1	Necessary
X2	Necessary
X3	Necessary
CAT	Necessary
X1*X2	Necessary
X1*X3	Necessary
X1*CAT	Necessary
X2*X3	Necessary
X2*CAT	Necessary
X3*CAT	Necessary
X1*X1	Necessary
X2*X2	Necessary
X3*X3	Necessary

Figure 5: List of main factors, two factor interactions, and quadratic terms to be estimated in the model

For Step 7, the analyst used JMP to create an I optimal design with 24 runs, as shown in Figure 6. Figures 7 through 9 show the JMP outputs, which will be explained as we evaluate this design.

Design

Run	X1	X2	X3	CAT
1	0	0	0	L3
2	-1	0	-1	L2
3	1	-1	1	L1
4	1	1	0	L2
5	0	-1	1	L2
6	0	1	1	L1
7	-1	0	1	L1
8	-1	-1	-1	L3
9	0	1	-1	L2
10	-1	-1	0	L2
11	-1	1	-1	L1
12	0	0	0	L1
13	1	1	-1	L3
14	1	1	1	L3
15	1	0.52	-1	L1
16	1	-1	-1	L2
17	1	-1	0	L3
18	0	0	0	L3
19	-1	1	0	L3
20	-1	1	1	L2
21	0	-1	-1	L1
22	-1	-1	0	L1
23	-1	-1	1	L3
24	1	0	1	L2

Figure 6: Sample I-optimal design created in JMP, with randomized run order

For Step 8, Option 2 was used to leverage the Prediction Variance Profile plots produced by JMP. Using Eq. 4, $r_{max} = 0.862$.

The first diagnostic step is to refer to the Design Diagnostics shown in Figure 7. The Average Variance of Prediction is 0.476, below the allowable maximum of 0.862. While that is not in and of itself a discriminator, it is a good sign when the average r is less than the maximum allowable r.

Design Diagnostics

I Optimal Design

D Efficiency	55.71919
G Efficiency	59.02563
A Efficiency	40.12408
Average Variance of Prediction	0.476424
Design Creation Time (seconds)	11

Figure 7: JMP Design Diagnostics metrics

The Prediction Variance Profile plots in Figure 8 are useful for inspecting r at various points in the factor space. To do so, click on the X axis of any of the factor plots in JMP and read the value of r to the left (shown in red). By using the "Maximize Variance" item from the red triangle dropdown on the Prediction Variance Profile, the analyst can have JMP snap to the factor combination with the highest relative prediction variance. Figure 8 shows that in this case the maximum value of r in the factor space is 1.4185 (well above r_{max}), and it occurs at $x_1 = x_2 = x_3 = 1$ and $CAT = L1$. Clearly there is a region in the design space where the confidence interval will likely exceed M_{max}, which should be concerning. Further investigation is warranted to see how large this region is and whether there are ramifications to the adequacy of this design.

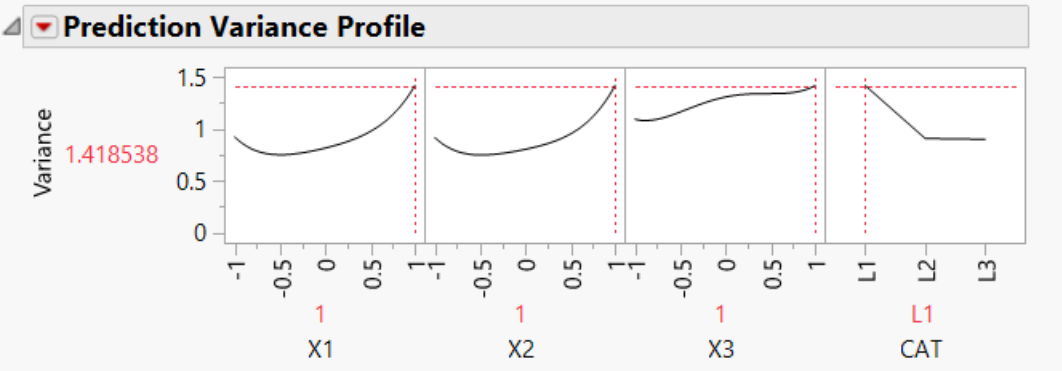


Figure 8: JMP Prediction Variance Profile plots

It is also important to notice that for this design, the level of the categorical factor affects the width of the confidence interval. Specifically, at $x_1 = x_2 = x_3 = 1$, the confidence interval for $CAT = L1$ will be about 50% larger than those for $CAT = L2$ and $CAT = L3$. This difference may be problematic if CAT

represents three systems under test that are in competition with each other, or where CAT represents three target types and L1 is the most operationally important.

Figure 8 shows the FDS plot. A dashed green line has been added to the figure at $r_{max} = 0.862$ for reference. Despite the greater complexity of this example, the plot is interpreted the same way as shown in Example 2. Over 95% of the design space has a relative prediction variance less than r_{max}. Thus, there is only a small region where the confidence intervals may be expected to exceed the desired width, and we know where this region is from Figure 8. This information may be used to determine whether the risk of using this design (i.e., the likelihood that the confidence interval in that 5% of the envelope may exceed the desired width) is acceptable.

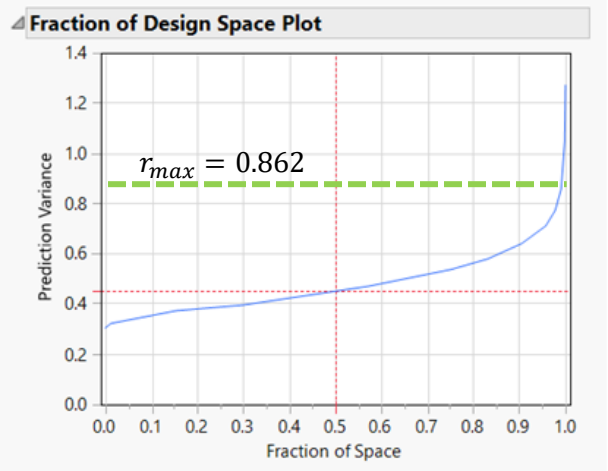


Figure 8: JMP Fraction of the Design Space (FDS) Plot

In addition to the standard prediction variance profiler available in JMP, the analyst can also use the Prediction Variance Surface to see where the maximum prediction variance occurs between any two given factors, as shown in Figure 9. The sliding scales allow the analyst to pinpoint the precise levels of each factor that maximize the prediction variance, showing the combinations of factors and levels which we can expect to create the largest prediction variance.

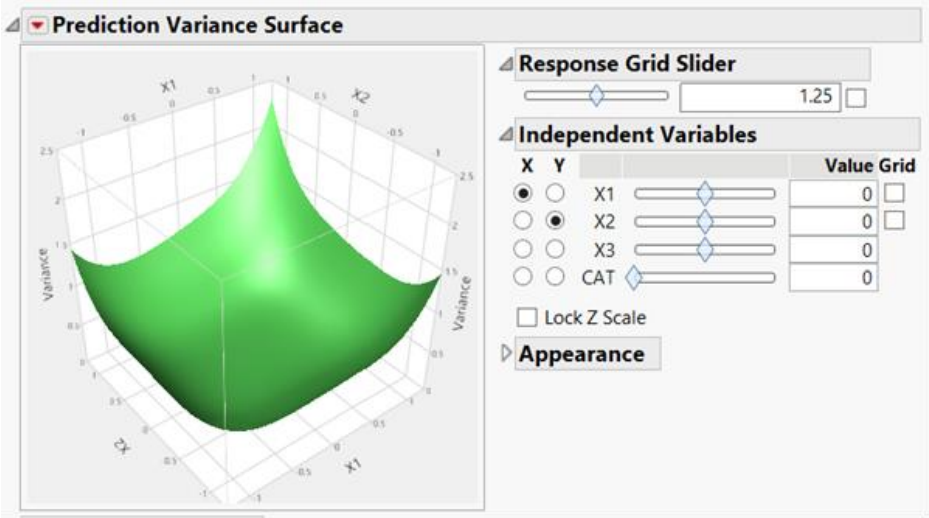


Figure 9: JMP Prediction Variance Surface plot

Conclusion

The method of using the relative prediction variance or margin of error of a confidence interval about the mean of the response to select a test design presents testers with an opportunity to design a test based on a given level of confidence that the system operates within a prescribed performance envelope. The procedure outlined in this best practice allows for an intuitive exploration and comparison of test designs while ensuring that the required precision will likely be achieved during testing.

One parting word of caution regarding this technique is that it applies specifically to confidence intervals, despite the fact we used the terminology “tolerance” in calculating the factor of safety. Before applying this procedure, test teams are highly encouraged to reconsider whether confidence intervals are appropriate for the given situation rather than tolerance intervals (Splinter, Sigler, Harman, & Kolsti, 2020).

Works Cited

Anderson-Cook, C. M., Zahran, A., & Myers, R. H. (2003). Fraction of Design Space to Assess Prediction Capabilities of Response Surface Designs. *Journal of Quality Technology, 35*(4), 377-386.

Burke, S., Divis, E., Guldin, S., Harman, M., Kolsti, K., McBride, A., . . . Welker, T. (2019). *Guide to Developing an Effective STAT Test Strategy V7.0.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

Goos, P., & Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach.* United Kingdom: John Wiley and Sons, Inc.

Goos, P., & Jones, B. (2012). *I-optimal versus D-optimal split-plot response surface designs.* Antwerp: University of Antwerp.

Hardin, R. H., & Sloane, N. J. (1993). *A New Approach to the Construction of Optimal Design.* Murray Hill: AT&T Bell Laboratories.

Harman, M. (2014, August 19). *Test Design Comparison and Selection.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE). Retrieved from Scientific Test and Analysis Techniques Center of Excellence.

Johnson, D. H. (1999). The Insignificance of Statistical Significance Testing. *The Journal of Wildlife Management, 63*(3), 763-772.

Johnson, T. H., Freeman, L., Simpson, J., & Anderson, C. (2018). Power approximations for generalized linear models using the signal-to-noise transformation method. *Quality Engineering*, 511-524.

Kensler, J., & Cortes, L. A. (2014, December 24). *Interpreting Confidence Intervals.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE). Retrieved from Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

Kiefer, J., & Wolfowitz, J. (1959). Optimum Designs in Regression Problems. *Annals of Mathematical Statistics, 30*(2), 271-294.

Ramert, A. (2019, August 31). *Understanding the Signal to Noise Ratio in Design of Experiments.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE). Retrieved from Scientific Test and Analysis Techniques Center of Excellence.

SAS Institute Inc. (2018). *JMP 14 Design of Experiments Guide.* Cary: SAS Institute Inc.

Shmueli, G. (2012). To Explain or to Predict? *Statistical Science, 25*(3), 289-310. doi:10.1214/10-STS330

Splinter, K., Sigler, G., Harman, M., & Kolsti, K. (2020). *Tolerance Intervals Demystified.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

Truett, L. (2018). *Developing Effective Test Objectives.* Dayton: Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

Appendix A: Mathematical Perspective

This appendix will provide a summary of the matrix formulas which can be used in lieu of statistical software.

TEST DESIGN PHASE

First, choose a model with factors, factor interactions, and higher-degree terms as desired. For example, for a quadratic model of two factors and their interaction, the model equation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \varepsilon_i \quad \text{(A.1)}$$

There are six coefficients in this example, so $K = 6$. For the matrix algebra to follow, the formula can be recast using row and column vectors:

$$y_i = \xi_i \beta + \varepsilon_i \quad \text{(A.2)}$$

where

$$\xi_i = [1 \quad x_{i1} \quad x_{i2} \quad x_{i1} x_{i2} \quad x_{i1}^2 \quad x_{i2}^2]$$

$$\beta = [\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_{12} \quad \beta_{11} \quad \beta_{22}]^T$$

To make one matrix equation with all n data points, the $n \times K$ design matrix X consists of n row vectors ξ_1,…, ξ_n stacked vertically.

$$X = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}$$

Note that D optimality is obtained by minimizing the determinant of $(X^T X)^{-1}$ (also known as the information matrix in JMP documentation) and A-optimality is obtained by minimizing the trace of $(X^T X)^{-1}$. The model in matrix form for n test points is then

$$y = X\beta + \varepsilon \quad \text{(A.3)}$$

At this stage in the planning, the relative prediction variance can be calculated at any point r_i (Goos & Jones, 2011).

$$r_i = \xi_i (X^T X)^{-1} {\xi_i}^T \quad \text{(A.4)}$$

I optimality is obtained by minimizing the integral of r over the entire factor space. G optimality ("G" is for "global", (Kiefer & Wolfowitz, 1959)) is obtained by minimizing the maximum value of r over the entire factor space.

AFTER COLLECTING THE DATA

With the data in hand, the fitted predictions of y are called $\hat{y}$ and the prediction formula is

$$\hat{y} = X\beta \tag{A.5}$$

Some algebra produces the formula for the best-fit parameter values. Note the explicit presence of the information matrix, $(X^T X)^{-1}$.

$$\beta = (X^T X)^{-1} X^T \hat{y} \tag{A.6}$$

To arrive at a solution without having to take the inverse of a matrix, Equation A.6 is the canonical linear system $A\beta = b$ where $A = X^T X$ and $= X^T \hat{y}$. The overall standard deviation of the response is estimated from the standard deviation of the residuals,

$$\hat{\sigma} = \frac{SSE}{\nu - 1} = \frac{1}{n - K - 1} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{A.7}$$

where $\nu = n - K$. Therefore, the confidence interval at any point r_k is

$$\hat{y}_k \pm M_k = \hat{y}_k \pm t_{1-\alpha/2,\nu} \hat{\sigma} \sqrt{r_k} \tag{A.8}$$

SINGLE-SAMPLE CASE

When all data are taken at one test condition (a single-sample case), as was shown in Example 1 of this Best Practice, there is no factor and hence there are no values for x. The model is

$$y_i = \beta_0 + \epsilon_i$$

where $\beta = [\beta_0]$ and $\xi = [1]$. With $K = 1$, the design matrix for a sample of n observations is the $n \times 1$ matrix X,

$$X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The prediction occurs at a single point so $X_p = 1$. From Eq. A.4,

$$r = X_p (X^T X)^{-1} {X_p}^T$$

$$r = 1 \cdot \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \cdot 1$$

$$r = 1 \cdot (1 + 1 + \cdots + 1)^{-1} \cdot 1$$

$$r = \frac{1}{n}$$

Since $\sqrt{r} = 1/\sqrt{n}$, this result can be substituted into Eq. 2 of this Best Practice to show that it produces the familiar classical confidence interval formula.