Dealing with Categorical Data Types in a Designed Experiment
Part II: Sizing a Designed Experiment When Using a Binary Response
Best Practice

Authored by: Francisco Ortiz, PhD STAT T&E COE

The goal of the STAT T&E 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 T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Introduction

From Part I of this best practice series we learned that the data type used to represent a response can affect the size of the experiment and the quality of its analysis. Categorical data types such as binary (pass/fail) measures contain a relatively poor amount of information in comparison to continuous data types. This reduction in information increases the number of samples needed to detect significant changes of a response in the presence of noise. The use of categorical data types for responses should be avoided; however, there will be circumstances in which a pass/fail measure is the only practical way to characterize a systems performance.

Three methods to estimate the samples size needed for a designed experiment using binary responses will be presented in this paper, the arcsine transformation approach, the signal-to-noise method and the inverse binomial sampling scheme method. The circumstances in which each method can be applied will be described in the paper. The methods will be demonstrated using a Target Location Error (TLE) example originally presented in part I of this series. All methods presented are available with the Binary Response Calculator on at the STAT COE website.

Keywords: binary responses, categorical factors, sample size, test and evaluation, design of experiments, confidence, power

Background

Target Location Error (TLE) Example Revisited

Let's revisit the missile targeting system example from part I of this best practice series. In this example we are comparing the ability of a missile targeting system to accurately assess the coordinates of a target within a tolerance radius of 10 feet. An example of the error distribution (using notional data) is shown in figure 1.

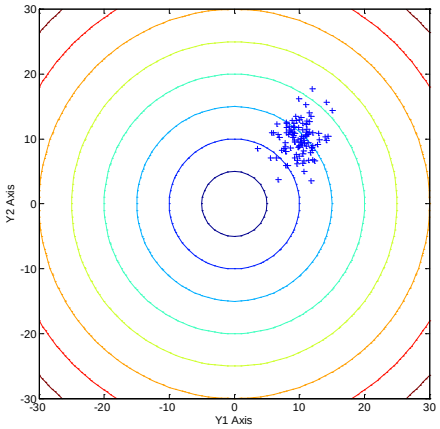


Figure 1: Distribution of error distance example (notional data).

Let's assume the only way to measure the systems is to categorize each attempt as either a "Pass" or "Fail" based on whether it falls within a 10 feet radius of the target (see Figure 2). This is a type of nominal response, specifically a binary response.

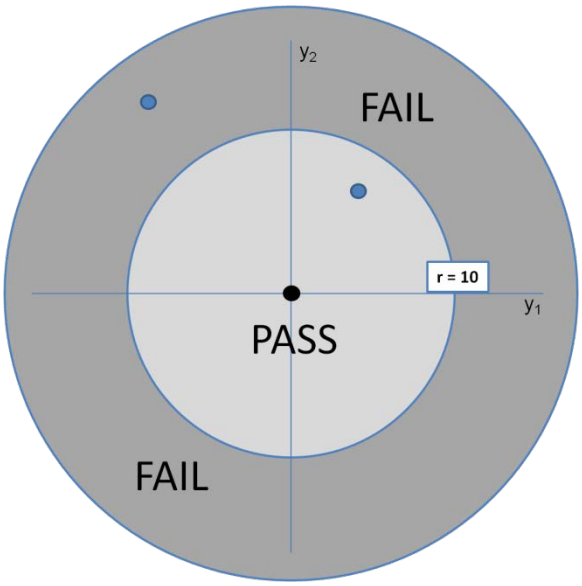


Figure 2: Measuring using binary (Pass/Fail) response.

The purpose for this test is to evaluate the performance of the system in various simulated engagements and also to determine what factors affect the probability to successfully hit the target. The objective and threshold probability of success (P_s) for the system is 90% and 80% respectively, under all expected conditions.

The test engineers have decided to vary four different factors, Altitude , Range, Aircraft Speed, and AOA A 2^4 factorial design has been chosen to setup the simulated engagements, see the coded matrix below (-1 and 1 are the low and high level values of a factor respectively). This design will allow the practitioner the ability to test what main effects and interactions terms significantly affect the response, the observed proportion of success $(\hat{p})$.

Table 1: 2^4 factorial design for TLE example.

Runs	X_1: Altitude	X_2: Range	X_3: Aircraft Speed	X_4: AOA
1	-1	-1	-1	-1
2	1	-1	-1	-1
3	-1	1	-1	-1
4	-1	-1	1	-1
5	-1	-1	-1	1

Runs	X_1: Altitude	X_2: Range	X_3: Aircraft Speed	X_4: AOA
6	1	1	-1	-1
7	1	-1	1	-1
8	1	-1	-1	1
9	-1	1	1	-1
10	-1	1	-1	1
11	-1	-1	1	1
12	1	1	1	-1
13	1	1	-1	1
14	1	-1	1	1
15	-1	1	1	1
16	1	1	1	1

The question that must now be answered is how many replicates of each design point must be run in order to achieve an appropriate level of power and confidence. For a review of the concepts of confidence and power please see the related section in Part I. Let's assume for this example that we will go with the DoD standard of 80% confidence and 80% power for a test ($\alpha = \beta = 0.2$) in this paper we will introduce a calculator/app that will aid practitioners in answering this question but before doing so let's first discuss the distribution the data comes from, the binomial distribution.

The Binomial Distribution

The binomial distribution is a discrete probability distribution of the number of successes in a series of n independent Bernoulli trials (pass/fail experiments), each trial yields success with probability p. The probability mass function is defined as:

$$P(Y = m) = \binom{n}{m} p^m (1-p)^{n-m} \tag{1}$$

For $m = 1,2, \ldots n$. The cumulative distribution function can be expressed as

$$P(Y \leq m) = \sum_{j=0}^{m} P\left(Y = j\right) \tag{2}$$

If a large enough sample size n is used the binomial distribution begins to look like the normal distribution and its parameters can be approximated with the following formulas.

Mean:

$$\mu = np \tag{3}$$

Standard Deviation:

$$\sigma = \sqrt{np(1-p)} \tag{4}$$

A rule of thumb commonly used to ensure that the distribution can be approximated by the normal distribution is the "rule of five":

$$\begin{gathered} np \geq 5 \\ \text{and} \\ n(1-p) \geq 5 \end{gathered} \tag{5}$$

The farther p is from 0.5 the larger n needs to be in order for this approximation to work. So for various p's the number of reps (n)needed are as follows:

Table 2: Number of reps (n)needed versus p based on the "Rule of Five".

p	n
0.1	50
0.2	25
0.3	17
0.4	13
0.5	10
0.6	13
0.7	17
0.8	25
0.9	50

The following graphs provide a visual representation of how the binomial distribution behaves with varying sample sizes n while keeping p at 0.1. You see that around $n = 50$ the shape of the histogram begins to look like the normal distribution curve.

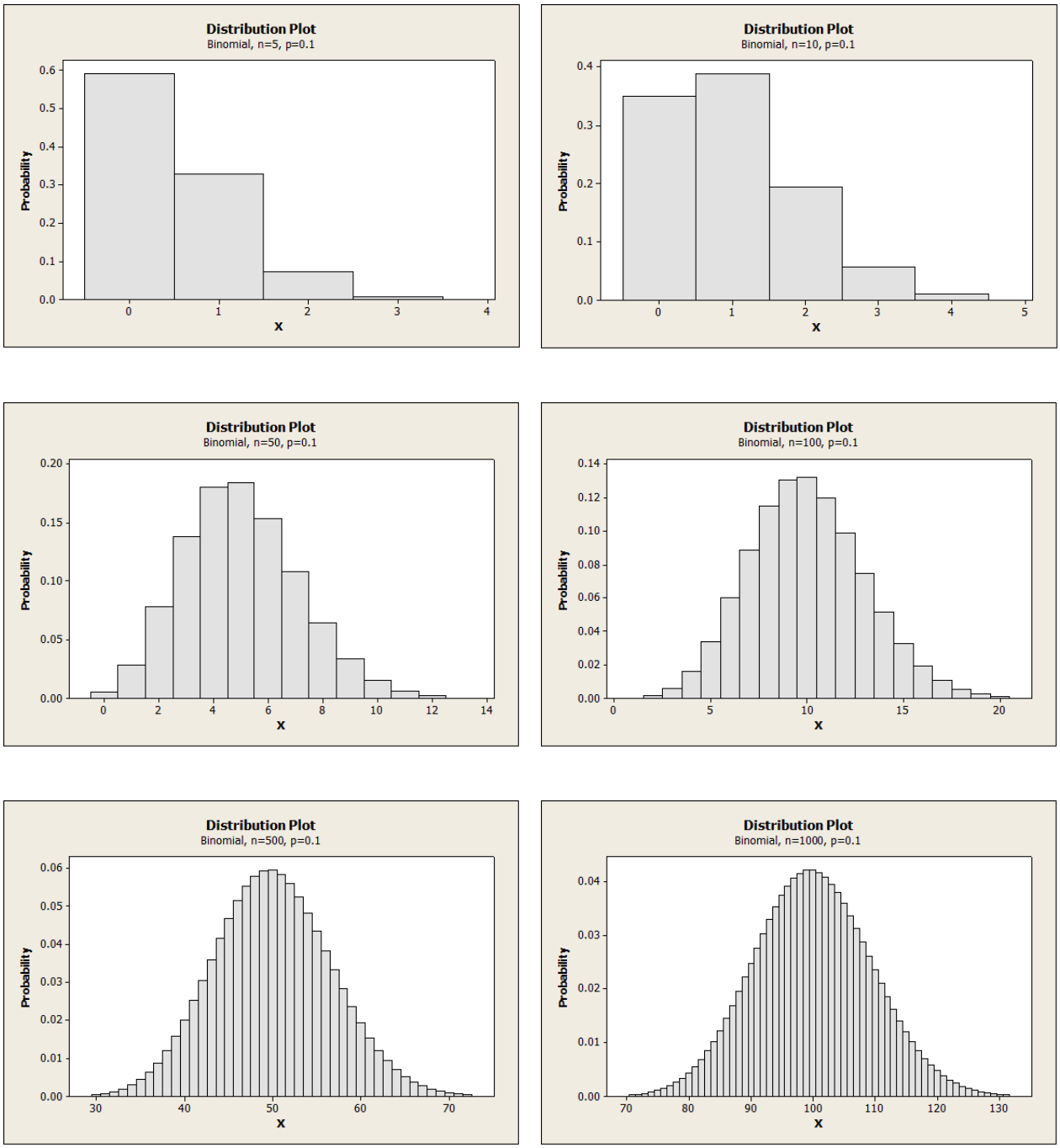
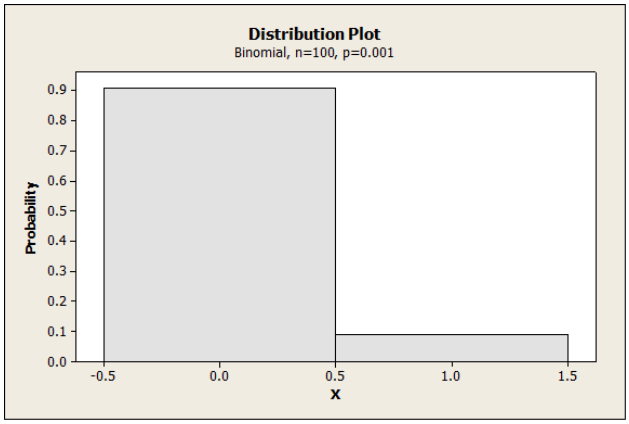
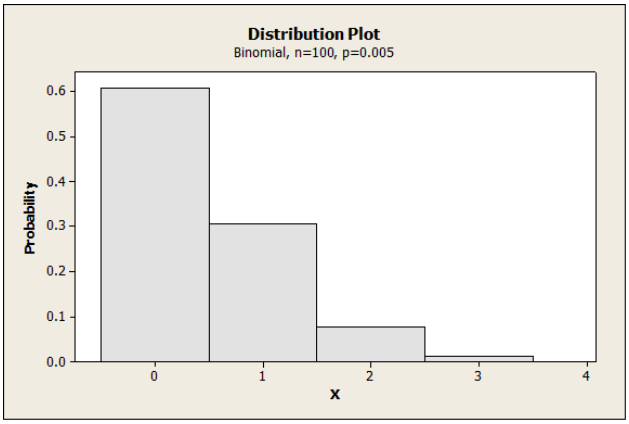
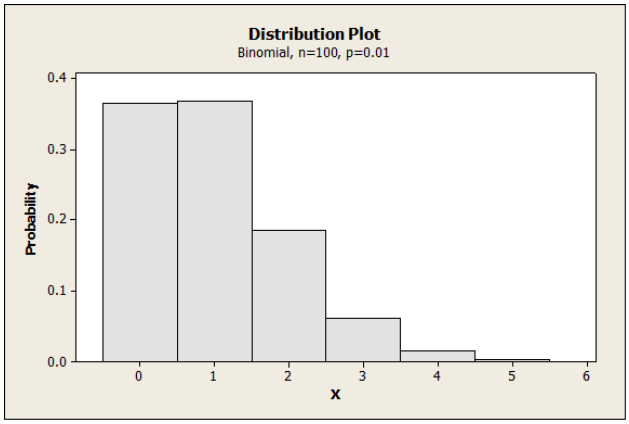


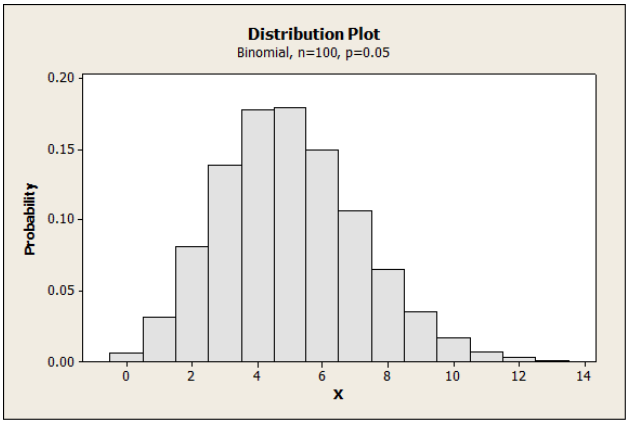
Figure 3: Distribution plots for $p = 0.1$ for varying sample sizes, n.

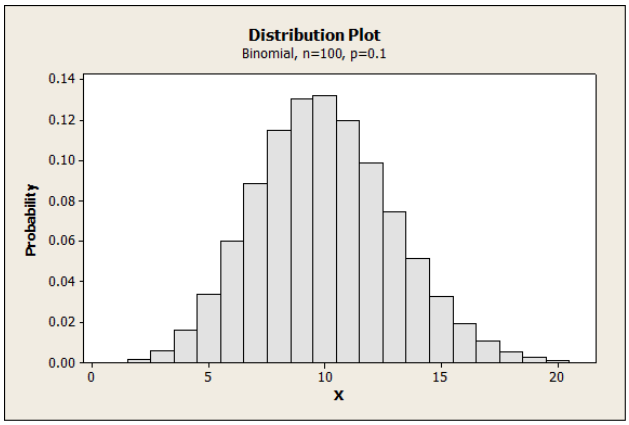
The following graphs provide a visual representation of how the binomial distribution behaves with varying proportions p and a constant sample size $n = 100$. You can see that the closer you are to the min and max values of 0 and 1 the distribution begins to looks less normal. Therefore caution should be taken when dealing with (P_s) greater than 95% or less than 1%..











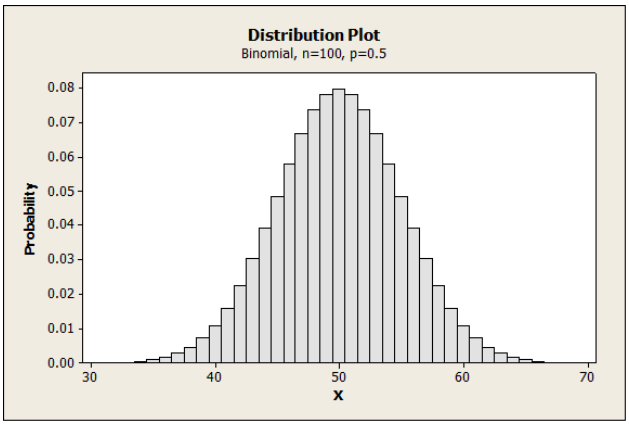


Figure 4: Distribution plots for $n = 100$ for varying proportions p.

Method

Method 1: Arcsine Transformation Approach

Note that the formulation for standard deviation in equation (4) is a function p, which is the very response we are monitoring and wish to change by varying factor levels. Due to this, the assumption of constant variance is violated. An approach to deal with this problem is to perform a variance stabilizing transformation on the observed response $\hat{p}$. The most commonly used transformation when dealing with binomial data is the arcsine square root transformation (see equation 6). This new transformed response would be the response used in the analysis.

$$\hat{p}_1^* = \arcsin\sqrt{\hat{p}} \tag{6}$$

Bisgaard and Fuller (1995) use this transformation to derive the number of replicates needed when using a 2^{k-f} factorial design with binary responses. Their formulation for the signal of interest (the change in the response we wish to detect) on the transformed scale is,

$$\delta = \arcsin\left(\sqrt{\bar{p} + \frac{\Delta}{2}}\right) - \arcsin\left(\sqrt{\bar{p} - \frac{\Delta}{2}}\right) \tag{7}$$

Where $\bar{p}$ is the expected proportion across the design space, and Δ is the signal in the original scale/units.

The individual point sample size is then calculated by the following formula

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{N\delta^2} \tag{8}$$

Where $z_{1-\alpha/2}\, z_{1-\alpha/2}$ and $z_{1-\beta}$ are the critical z values based on the specified power and confidence, N is the total number of design points and δ is defined in equation (7).

To demonstrate further, let's use the TLE example introduced earlier with the Binary Response Calculator available on at the STAT COE website. The following is a screen shot of the calculator.

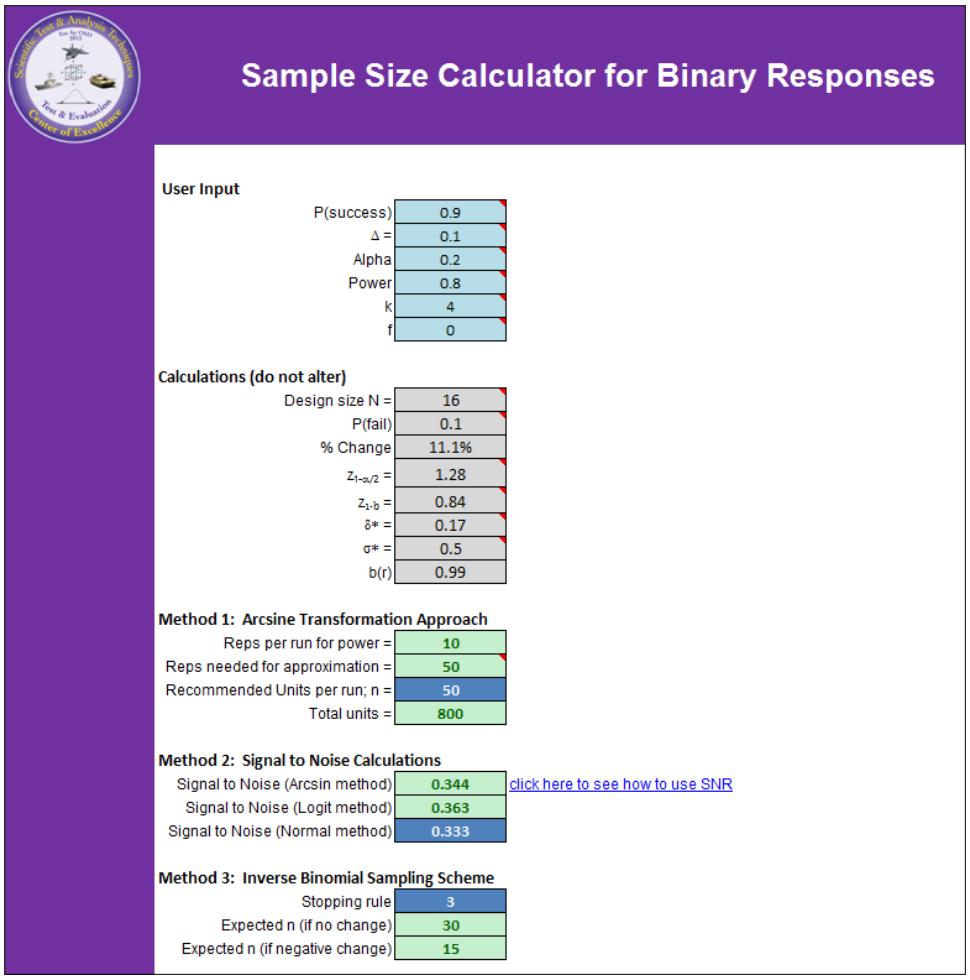


Figure 5: Sample Size Calculator for Binary Responses screenshot.

The calculator consists of 5 sections.

The User Input section, this is where the basic information about the test needs to be specified.

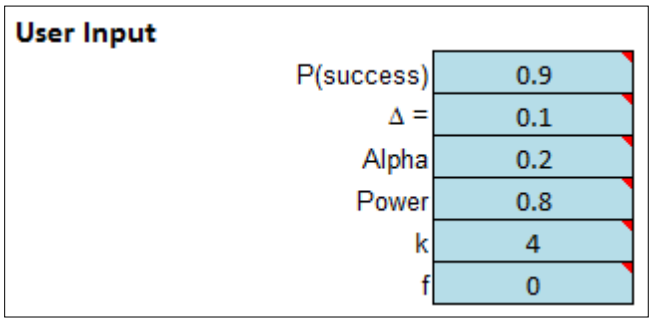


Figure 6: User Input section.

The following information must be specified in this section:

- P(success): The expected probability of success across the design space
- Δ: The signal of interest (the change in the response we wish to detect)
- Alpha: Allowable Type I error. Confidence is 1-Alpha.
- Power: The probability of detecting Δ. 1-Power is the type II error
- k: The number of factors in the design (4 in this case).
- f: The level we wish to fractionate the factorial (0 in this case since this is a full factorial)

Remember, for our example the objective and threshold P_s for the system are 90% and 80% respectively. Therefore, $P(success)$ is set to 0.9 and Δ is equal to 0.1. We're going with the DOD standard of 80% confidence and 80% power for a test $(\alpha = \beta = 0.2)$. We are using a 2^4 full factorial design so $k = 4$ and $f = 0$.

The Calculations section simply displays some values of interest and the inputs used in equation (8).

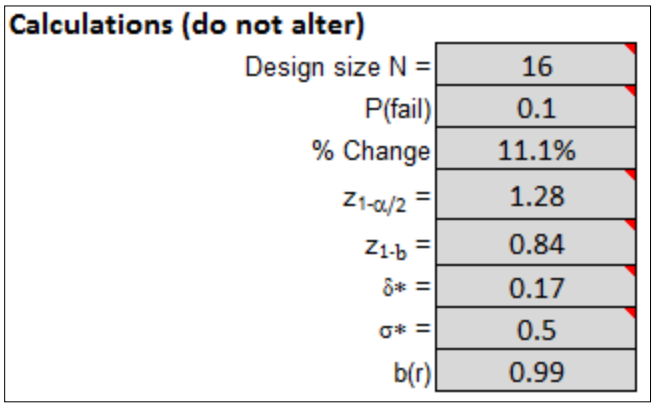


Figure 7: Calculations section.

The Method 1 section displays the results from applying the approach defined by Bisgaard and Fuller (1995).

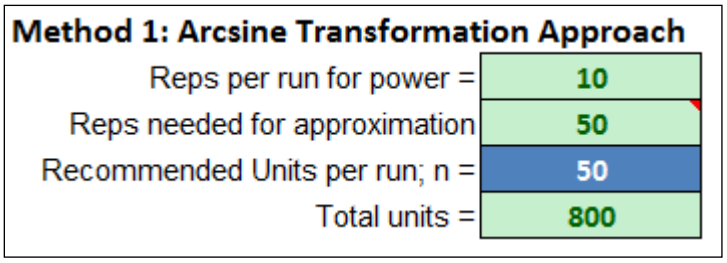


Figure 8: Method 1- Arcsine Transformation Method.

The following describes the output:

- Reps per run for power: Based on equation (8)

- Reps needed for approximation: Based on the "Rule of 5"
- Recommended Units per run: Takes the maximum value between the reps needed for power and the reps needed for the approximation
- Total units: The total number of runs X the recommended number of reps.

For the TLE example, the calculator is recommending 50 reps for each of the $2^4 = 16$ design points, resulting in a total of 800 runs.

The information provided in the sections Method 2: Signal to Noise Calculations and Method 3: Inverse Binomial Sampling Scheme will be discussed in the following sections.

Method 2: Signal to Noise Calculations

Method 1, the Arcsine Transformation Approach, only works if a 2^{k-f} designs is used. If another type of design is used a better approach would be to used the signal-to-noise ratio (SNR) method. The signal to noise ratio is simply the ratio between the measured change in the response we wish to detect (δ, the signal of interest) and the estimated standard deviation of the system (noise), see the formula below.

$$SNR = \frac{\delta}{\sigma} \tag{9}$$

Three methods of calculating the SNR are presented in the calculator.

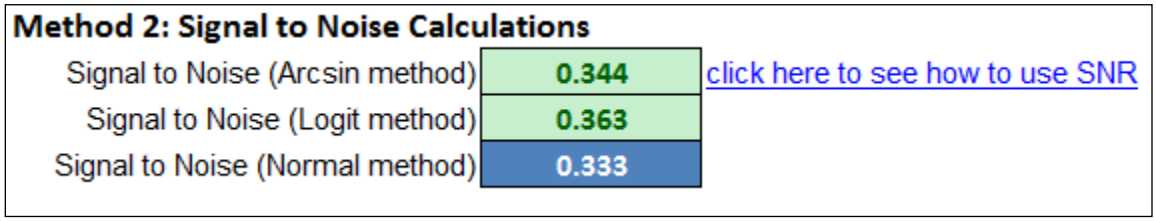


Figure 9: Method 2- Signal to Noise Calculations.

Arcsine Formulation

This method uses the same formulations for delta and sigma that were derived in the Bisgaard and Fuller (1995) paper. Delta (in the transformed scale) is the same as in equation (7) repeated here for convenience.

$$\delta_1 = \arcsin\left(\sqrt{\bar{p} + \frac{\Delta}{2}}\right) - \arcsin\left(\sqrt{\bar{p} - \frac{\Delta}{2}}\right)$$

and the standard deviation for the arcsine transformation is as follows

$$\sigma_1 = \frac{1}{\sqrt{4n}} = \frac{1}{2} \tag{10}$$

where $n = 1$ here since we wish to determine what the SNR is before replication.

Logit Formulation

This approaches uses the Logit transformation which is the traditional solution used when applying logistic regression to fit a model where the dependant variable is a proportion. The transformation takes the log of the odds

$$\hat{p}_2^* = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) \tag{11}$$

Where p is the probability of an event occurring, $1-p$ is the probability of an event not occurring and $\frac{p}{1-p}$ is the odds of the event. Delta in the transformed scale is defined below.

$$\delta_2 = \left|\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right)\right| \tag{12}$$

where $p_1 = \overline{p} + \frac{\Delta}{2}$ and $p_2 = \overline{p} - \frac{\Delta}{2}$. The standard deviation is defined as follows,

$$\sigma_2 = \sqrt{n\overline{p}(1-\overline{p})} = \sqrt{\overline{p}(1-\overline{p})} \tag{13}$$

where $n = 1$ here since we wish to determine what the SNR is before replication.

Normal Approximation Formulation

The final SNR formulation is based on the Normal Approximation of the binomial. This is the simplest of the formulations presented in this paper. Delta is defined as

$$\delta_3 = \left|p_1 - p_2\right| \tag{14}$$

where $p_1 = \overline{p} + \frac{\Delta}{2}$ and $p_2 = \overline{p} - \frac{\Delta}{2}$. The standard deviation is defined the same as the logit formulation

$$\sigma_3 = \sqrt{n\overline{p}(1-\overline{p})} = \sqrt{\overline{p}(1-\overline{p})}$$

where $n = 1$ here since we wish to determine what the SNR is before replication.

Figure 9 are the results from calculator using the TLE Example. Note that all three methods provide similar results. Table 3 shows the SNR results of the three methods when varying p. The Normal Approximation method consistently produces the most conservative estimate of the SNR.

Table 3: Comparison of SNR calculation methods.

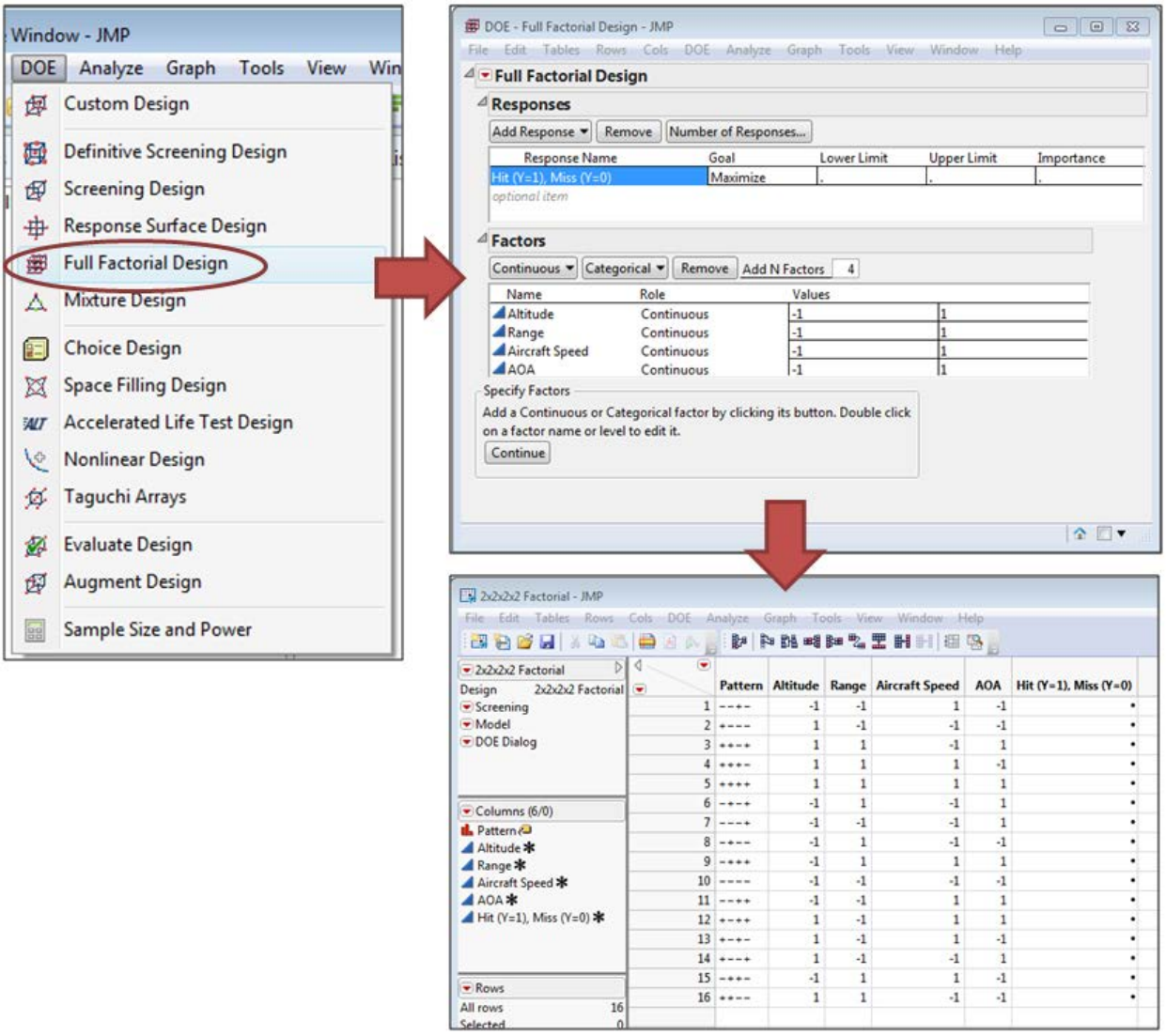
p	Δ	SNR (arcsin)	SNR (logit)	SNR (normal)
0.9	0.100	0.3444	0.3630	0.3333
0.85	0.100	0.2838	0.2896	0.2801
0.8	0.100	0.2518	0.2544	0.2500
0.75	0.100	0.2320	0.2334	0.2309
0.7	0.100	0.2189	0.2198	0.2182
0.65	0.100	0.2102	0.2107	0.2097
0.6	0.100	0.2045	0.2050	0.2041
0.55	0.100	0.2014	0.2017	0.2010
0.5	0.100	0.2003	0.2007	0.2000
0.45	0.100	0.2014	0.2017	0.2010
0.4	0.100	0.2045	0.2050	0.2041
0.35	0.100	0.2102	0.2107	0.2097
0.3	0.100	0.2189	0.2198	0.2182
0.25	0.100	0.2320	0.2334	0.2309
0.2	0.100	0.2518	0.2544	0.2500
0.15	0.100	0.2838	0.2896	0.2801
0.1	0.100	0.3444	0.3630	0.3333

Using the signal-to-noise ratio (JMP 10 Demo)

Most DOE software will allow you to input the SNR in order to calculate the power of the test. In this section we will demonstrate how to use the SNR with JMP 10.

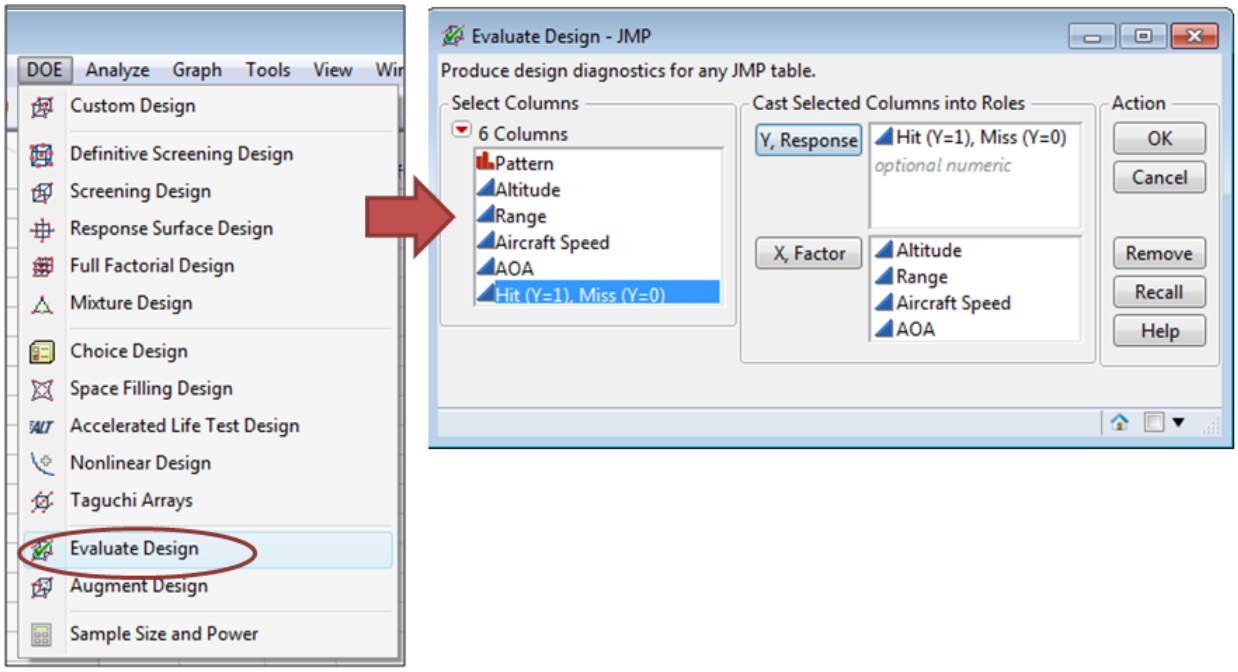
Step 1: Create your design in JMP

- Create a 2^4 factorial design to match our TLE example. The columns X1, X2, X3, X4 columns represent our factors Altitude, Range, Aircraft Speed, AOA respectively. The Y column is our pass/fail response.

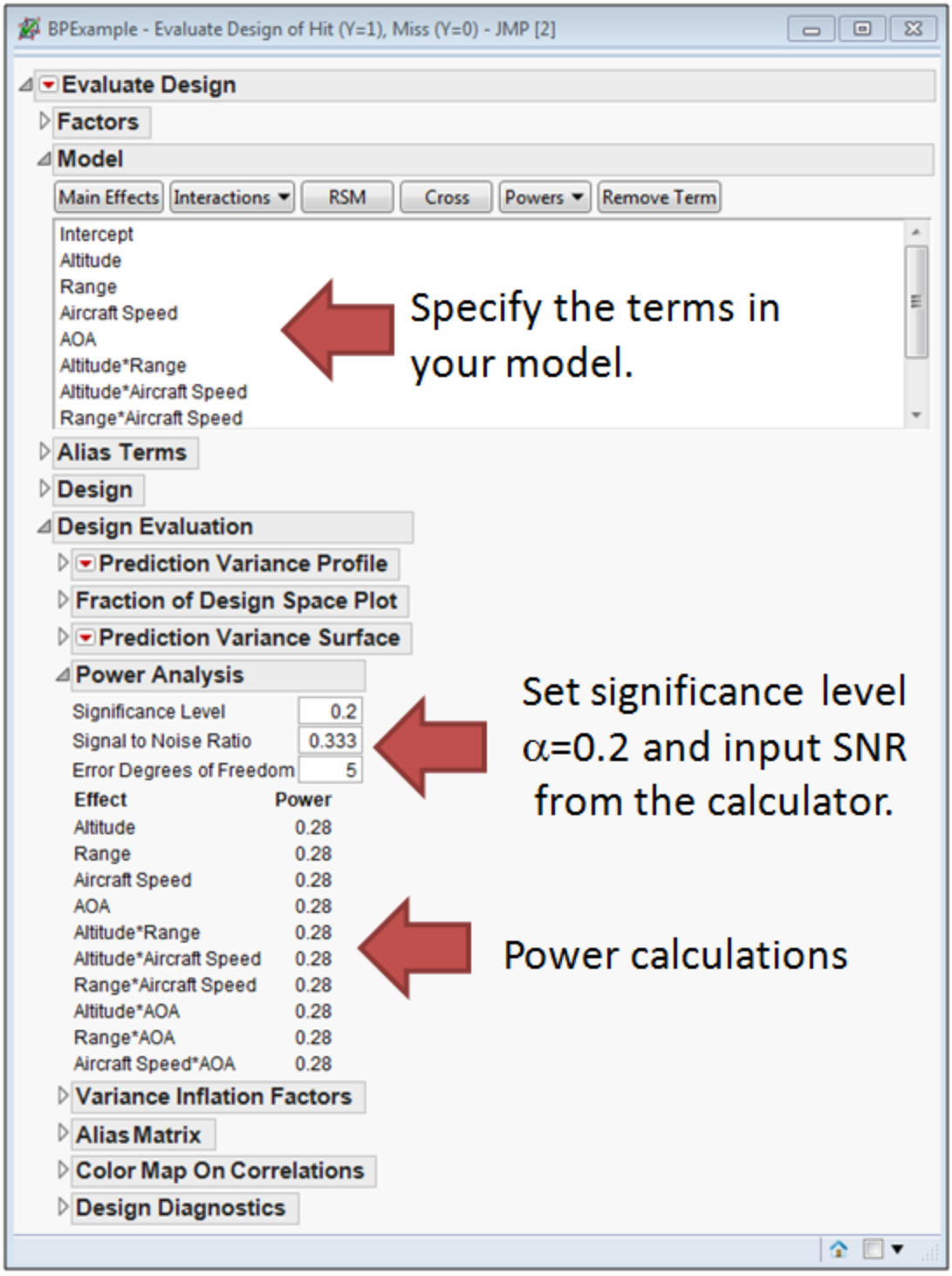


Step 2: Evaluate design

- Once the design is created, select DOE > Evaluate Design.



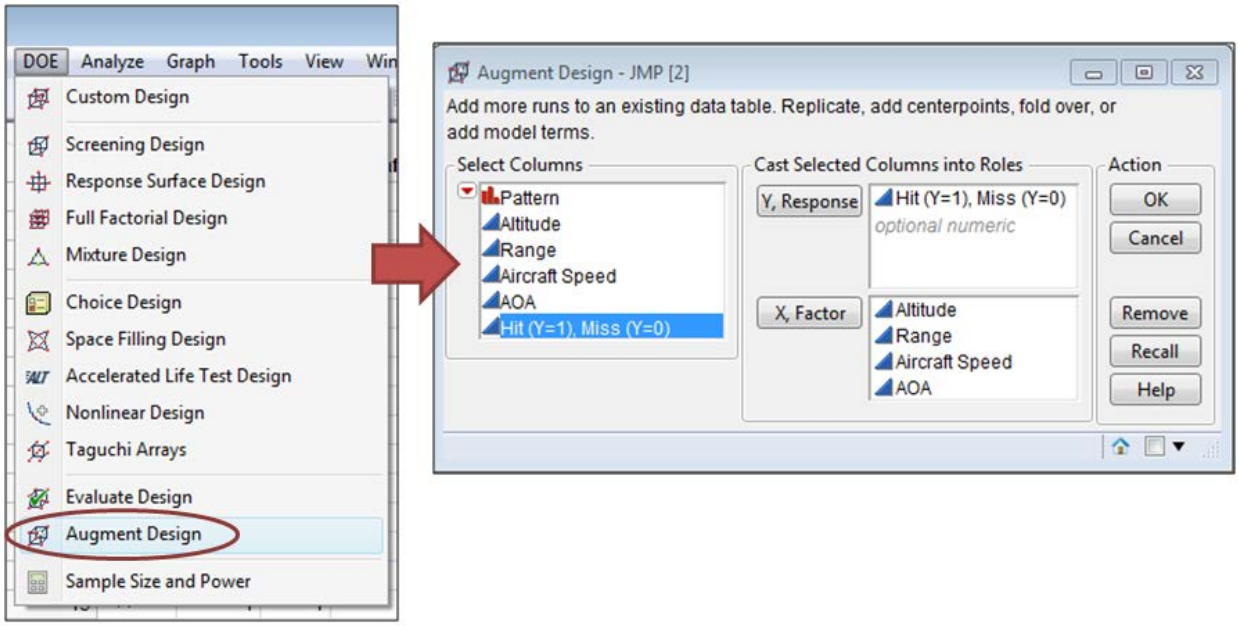
- Specify the response and factor columns and then click OK. The Evaluate Design dialog box will appear.
- In the Evaluate Design dialog box:
  - Specify the terms in your model. For this example we are interested in main effects and two factor interactions.
  - Set significance level $\alpha = 0.2$.
  - Input SNR from the calculator.



- Note: Power is 28% for all terms if we only run each setting once.

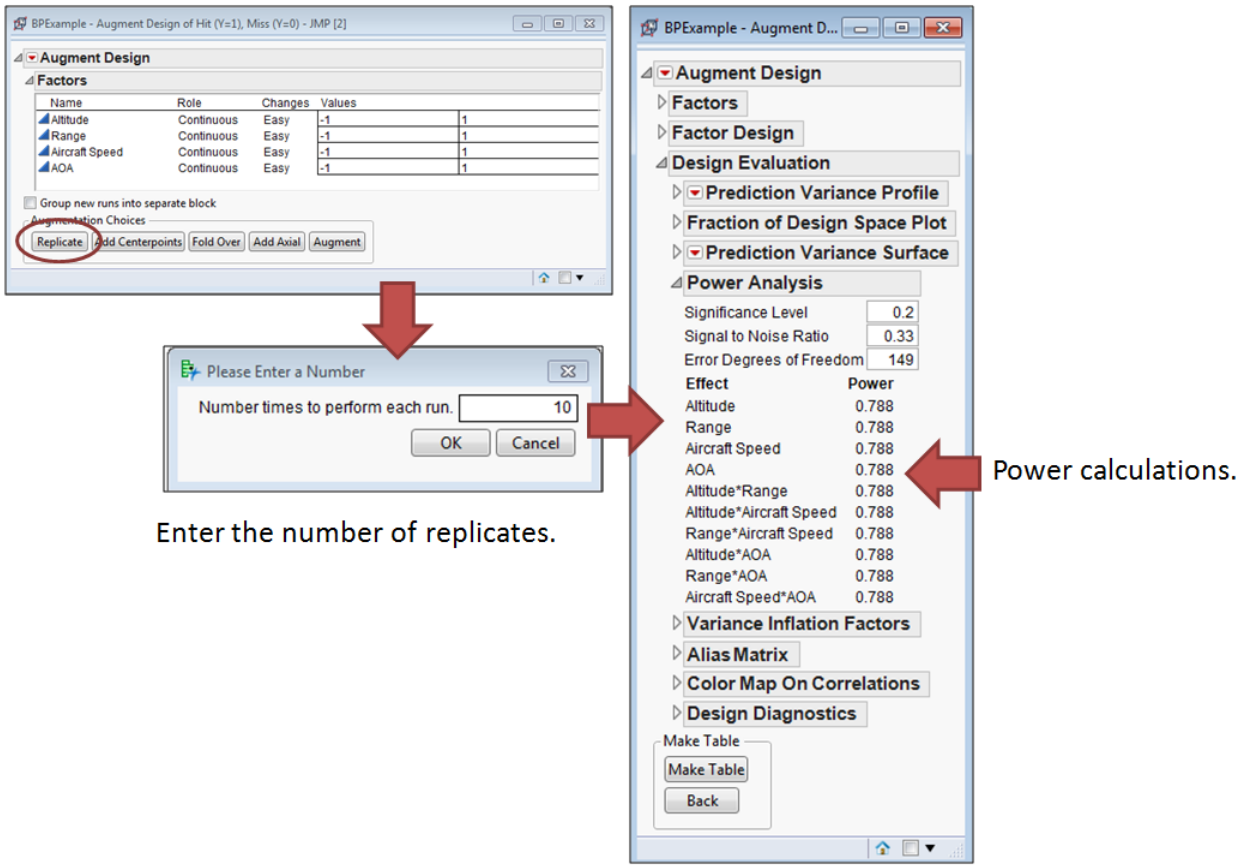
Step 3: Augment design

- Augment the design to add replicates and bring power up to appropriate level. For this example 80%.
- Select DOE > Augment Design.



- Specify the response and factor columns and then click OK. The Augment Design dialog box will appear.

Click the "Replicate" button.



- Make sure all factors to be replicated are listed.
- Click the "replicate" button.
- Enter the number of times to replicate each design point.
- Check the Power Analysis section of the resulting design. Confirm that calculations are above or close to predetermined power objectives. In this case, 80%.
- If not, click the back button and increase the number of replicates until you achieve your power objective.

Method 3: Inverse Binomial Sampling Scheme

The final method provided by the calculator is the Inverse Binomial Sampling Scheme proposed in Bisgaard and Gertsbakh (2000). This method can be used with a 2^{k-f} design where the purpose is to reduce the rate of defectives. Instead of determining a fixed sample size for each design run this approach suggests sampling until a fixed number of defects r, are observed. The derivation for the stopping rule will not be covered in this paper, for more details please refer to Bisgaard and Gertsbakh (2000). The number of defects *r* observed is based on the number of factorial trials in a 2^{k-f} design, the change in probability to detect Δ, and the fixed levels of α and β. The total number of reps until r

defects occurs is used as the response. This approach could significantly reduce the number of total runs needed if the system does not meet the P_s requirement.

Method 3: Inverse Binomial Sampling Scheme	
Stopping rule	3
Expected n (if no change)	30
Expected n (if negative change)	15

The following describes the calculators output:

- Stopping rule: The number of defects to observe for each design run.
- Expected n (if no change): Based on the estimated P_s, this is the expected number of reps needed to observe the stopping rule.
- Expected n (if negative change): If a negative change of Δ has occurred this is the expected number of reps needed to observe the stopping rule.

Note that an unequal number of reps for each design run is likely therefore a general linear model (GLM) or weighted least squares (WLS) approach is recommended for the analysis. A tutorial on how to analyze the data will be presented in part 3 of this best practice series.

Conclusion

Three methods to estimate the samples size needed for a designed experiment using binary responses were presented in this paper. The arcsine transformation approach can be used if a 2^{k-f} design is employed. The signal-to-noise method can be used for any design but requires iterative exploration of the number of replicates needed using statistical software. A JMP 10 tutorial on how to do this was provided. The Inverse Binomial Sampling Scheme method can also be used if a 2^{k-f} design is employed. This method could be a potential resource saving approach for system with a high expected probability of success (P_s) and if the goal is to simply demonstrate that the system meets that objective P_s. All methods presented are available for use on the Binary Response Calculator available on at the STAT COE website.

There are opportunities for future work on this subject. A Monte Carlo approach should be considered in order to produce more accurate power calculations that are robust to the experimental design used. Also future research should explore the use of OC curves and sequential probability ratio testing in order to truncate and quickly stop testing if it is abundantly clear that the system is passing or failing the requirements.

References

Bisgaard, S. and Fuller, H., "Sample Size Estimates for 2^{k-p} Designs with Binary Responses," *Journal of Quality Technology,* 1995.

Bisgaard, S. and Gertsbakh, I. "2^{k-p} Experiments With Binary Responses: Inverse Binomial Sampling," Journal of Quality Technology, 2000.

Gotwalt, C., "JMP Script for Computing Binary Power using the Logit Transformation," JMP, 2012.

Lenth, R., Piface Java Applet version 1.76, University of Iowa, 2011.

Whitcomb, P. and Anderson, M., Excel Sample Size Calculator for Binary Responses, Stat-ease, 2000.