Categorical Data in a Designed Experiment Part 1: Avoiding Categorical Data

Authored by: Francisco Ortiz, PhD

Revised 27 August 2018

Revised 23 October 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary 1
Introduction 2
Background 3
Data Types 3
Confidence and Power 4
About Responses 5
Generic Missile Targeting System Example 5
About Factors 8
Generic Missile Targeting System Example Revisited 8
Conclusion 11
References 12

Revision 1, 27 Aug 2018: Formatting and minor typographical/grammatical edits.

Executive Summary

The STAT COE has observed that many TEMPs and test plans use categorical data types to describe factors and responses. In the case of responses, categorical data types contain a relatively poor amount of information in comparison to continuous data types (38% to 60% less in some cases). This reduced information results in an increased difficulty for tests to detect significant changes in the presence of noise. In other words, these tests will have a poor signal-to-noise ratio and in turn will require more replication/runs to compensate for the lack of information. Using categorical data types to describe factors may also have an effect on the overall size of the test and the quality of the analysis. For one, categorical data types do not allow for inference between levels. Also, the number of center points needed to test for curvature (or nonlinearity) in a design experiment increases rapidly as a function of the number of categorical factors and their levels. The following best practice is an argument against the use of categorical data types in a designed experiment if at all possible. This best practice is the first part of a series that will show how to deal with categorical data types in a designed experiment.

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

Introduction

The data type chosen to represent test design factors (inputs) and responses (outputs) in an experiment can have a major affect on the resources needed to conduct an experiment and the quality of its respective analysis. The STAT COE has observed through a review of program TEMPs and test plans that categorical data types are often used when describing factor levels and responses. Perhaps this is due to requirements and test objectives that are defined too generally and are non-specific. It could also be that planners find it easier to conceptualize and plan experiments using generic settings and measures. Perhaps there is difficultly in measuring the inputs and outputs in great detail. For example, live fire testing could be destructive and thus impossible to precisely measure impact point. Thus, it is easier to use a count of hit and misses instead of measured missed distance from the target. Another example would be that we are testing the system during different times of the day and list the factor settings as “day” and “night” instead of using some measure of illumination. Whatever the reason, any perceived savings in planning time to define factors and responses using numeric data types instead of categorical data types is paid for in wasted resources and complications during analysis.

The following best practice is broken up into three parts. Part 1 will provide justification for using numeric measures (continuous, discrete) over categorical (nominal, ordinal, binary) measures when possible; Part 2 is a tutorial on how to properly size a designed experiment when a binary response must be used; and Part 3 will illustrate how to conduct the analysis and interpret results using logistic regression.

The following discussion will begin with a brief overview of data types, power, and confidence, followed by separate discussions on the benefits of using numeric measures (specifically continuous measures) for responses and factors respectively. For these discussions a generic missile targeting system example will be used to compare different approaches. This paper won't go into great details on the statistics and mathematics behind some of the calculations but it will reference relevant papers and presentations if the reader wishes to research the subject further.

Background

Data Types

Data can be classified as either numerical (quantitative) or categorical (qualitative).

Figure 1 depicts some commonly used data types and their relationships to each other. Table 1 provides a more detailed description of each type as well as some examples.

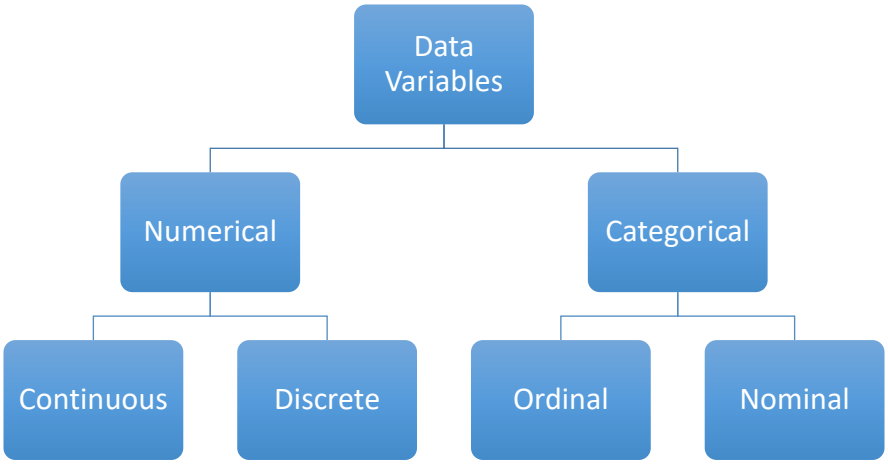


Figure 1: Commonly used data types

Table 1: Summary of commonly used data types

Data Type	Definition	Example	Information Content
Continuous	Data take a value based on a measurement at any point along a continuum. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows.	Height, time, age, temperature, time	Most Information

Data Type	Definition	Example	Information Content
Discrete	Data take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value.	Number of registered cars, number of business locations, and number of children in a family, all of which are measured as whole units (i.e. 1, 2, 3 cars).	More Information
Ordinal	Data take on values that can be logically ordered or ranked. The categories associated with ordinal variables can be ranked higher or lower than another but do not necessarily establish a numeric difference between the each category.	Academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra large) and attitudes (i.e. strongly agree, agree, disagree, strongly disagree). Rank order of preferences on a Scale of 1-5, Order in Races, Letter Grades	Less Information
Nominal	Data take on values that are not able to be organized in a logical sequence. Binary data (0-1, on/off, hit/miss) is an example of a nominal data.	Binary data (Pass/Fail, Hit/Miss, Detect/Non-Detect) gender, business type, eye color, religion and brand.	Least information

In depth studies have been conducted to quantify the amount of information loss when using categorical (specifically binary) responses. Cohen (1983) shows that using a binary response results in the reduction of statistical power equivalent to discarding 38%-60% of the data. Hamada (2002) shows that confidence intervals for binary responses can be multiple times larger than if continuous responses were used, particularly when the conformance probability is high. Of the four data types listed, continuous variables are considered to be the most information rich, meaning it has greater precision, detail, and potential for inference.

Confidence and Power

In order to illustrate the benefits of using continuous over categorical variables we need to first discuss the concepts of confidence and power. Confidence is a measure of how accurate and reliable your statistical judgments will be. In statistical terms, confidence is the probability of not committing a Type I error, or making a false positive decision. For example consider a scenario where we are comparing two systems to see if there is a difference in average performance, a confidence of 95% means that there is a 5% chance (denoted as area α in Figure 2) that we will say there is a difference between these two means (μ_0 and μ_1 in Figure 2) when in fact there is no difference. Note that confidence is always set before beginning the test.

Power is a measure of how likely your statistical test will be to detect changes (δ) of a given size. In statistical terms, Power is the probability of not committing a Type II error, or making a false negative decision. Again consider the scenario where we are comparing two systems. A power of 80% means that there is a 20% chance (denoted as area β in Figure 2) that we will say there is no difference between these two means when in fact there is a difference.

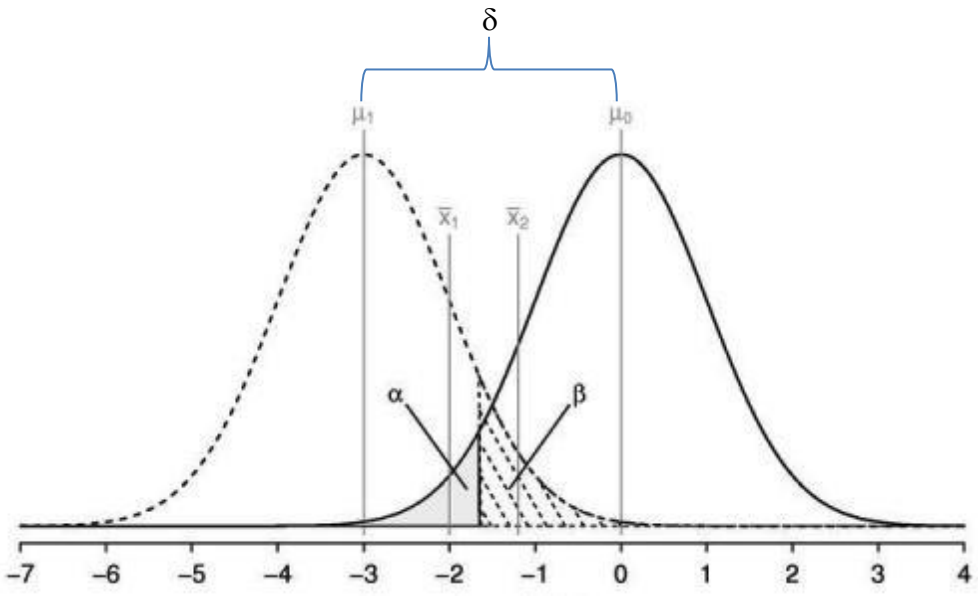


Figure 2: Confidence and power demonstration

Power is dependent on the number of trial runs collected as well as the difference between means (δ) that the practitioner wishes to detect.

About Responses

Generic Missile Targeting System Example

To further elaborate on power and confidence let's consider a simple test in which we are comparing the ability of two different systems to accurately assess the coordinates of a target within a tolerance radius of 10 feet. The data collected from both systems would follow a bivariate normal distribution (see Figure 3).

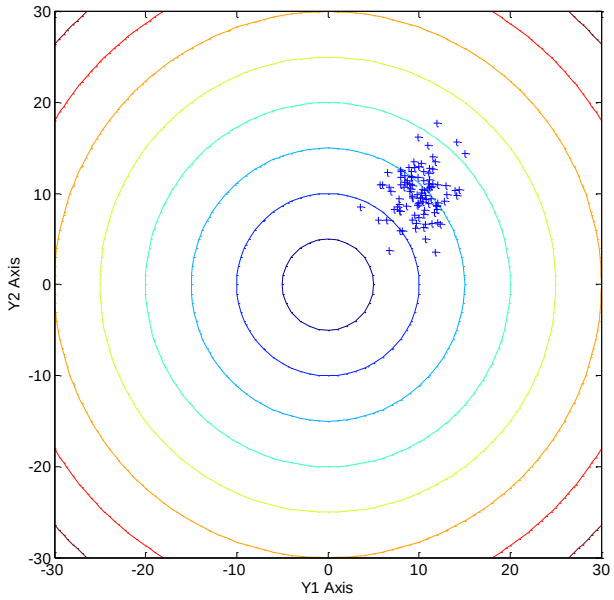
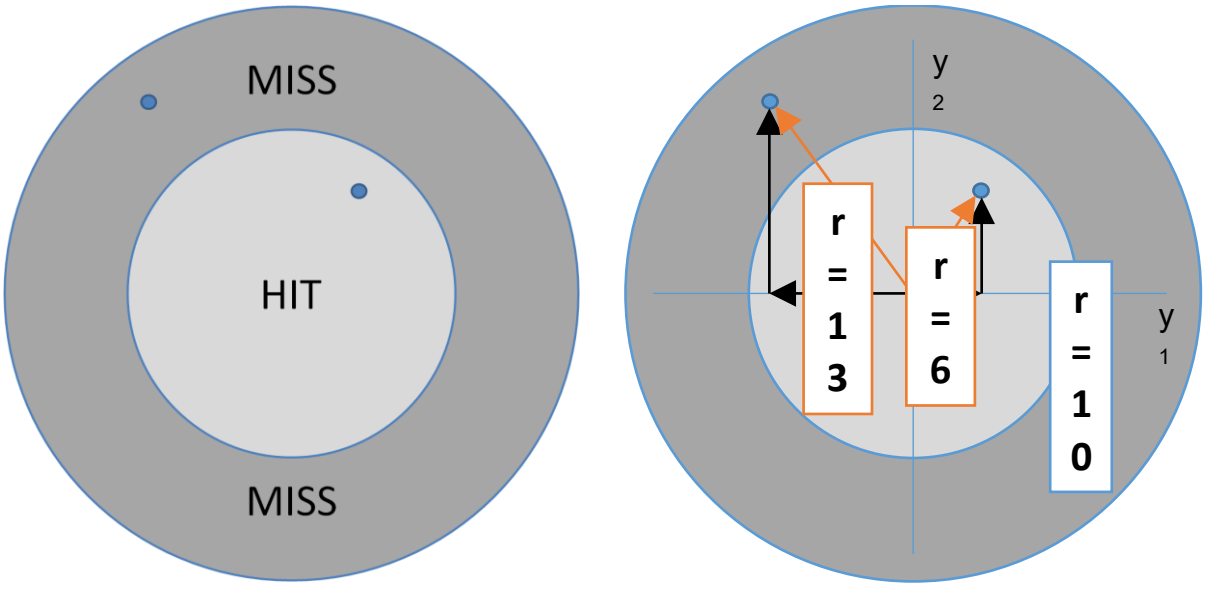


Figure 3: Distribution of error distance example

One way we could 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 4a). This is a type of nominal response, specifically a binary response.



(a) **(b)**

Figure 4: Measuring using a) binary (hit/miss) and b) continuous (distance from center) response

Another way to evaluate performance is to measure the error distance between the system generated coordinates and the actual coordinates of the target (see Figure 4b). This is a continuous response and requires two measures of the error distance along y1 and y2 to calculate the resultant vector r, where:

$$r = \sqrt{y_1^{\,2} + y_2^{\,2}} \tag{1}$$

Note that the continuous response can easily be converted to a binary (hit/miss) response by simply stating that any measured distance greater than 10 feet from the center is a miss. However, it is not possible to convert the binary response back into the continuous distance measure if that data was not originally collected. That information is lost when recording the system performance on the binary scale.

For the sake of simplicity let's just examine the continuous response in one dimension (y_1), see Figure 5. Let's assume that for System A, the system is normally distributed and the average distance from center is 0 feet and it has standard deviation of 10 feet. This means that system will hit within 10 feet of the target center approximately 68% of the time. System B is also normally distributed, with an average distance from center of 10 feet and it also has a standard deviation of 10 feet. Hence, System B will hit within 10 feet of the target center 48% of the time.

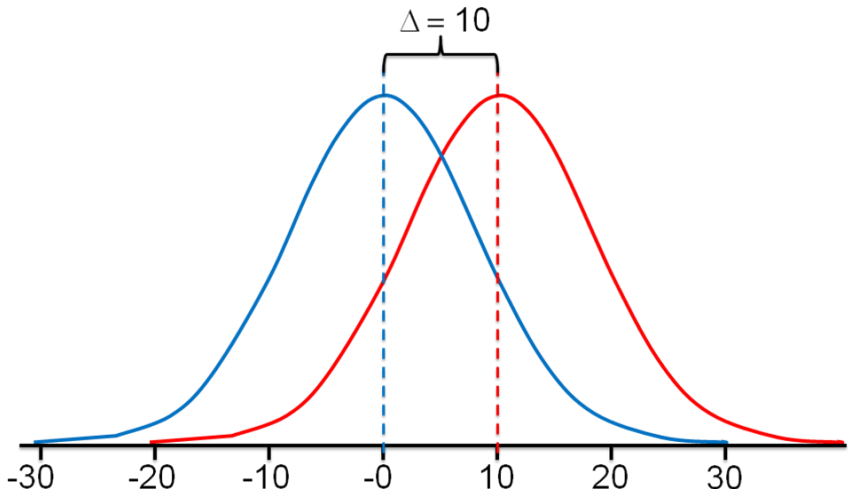
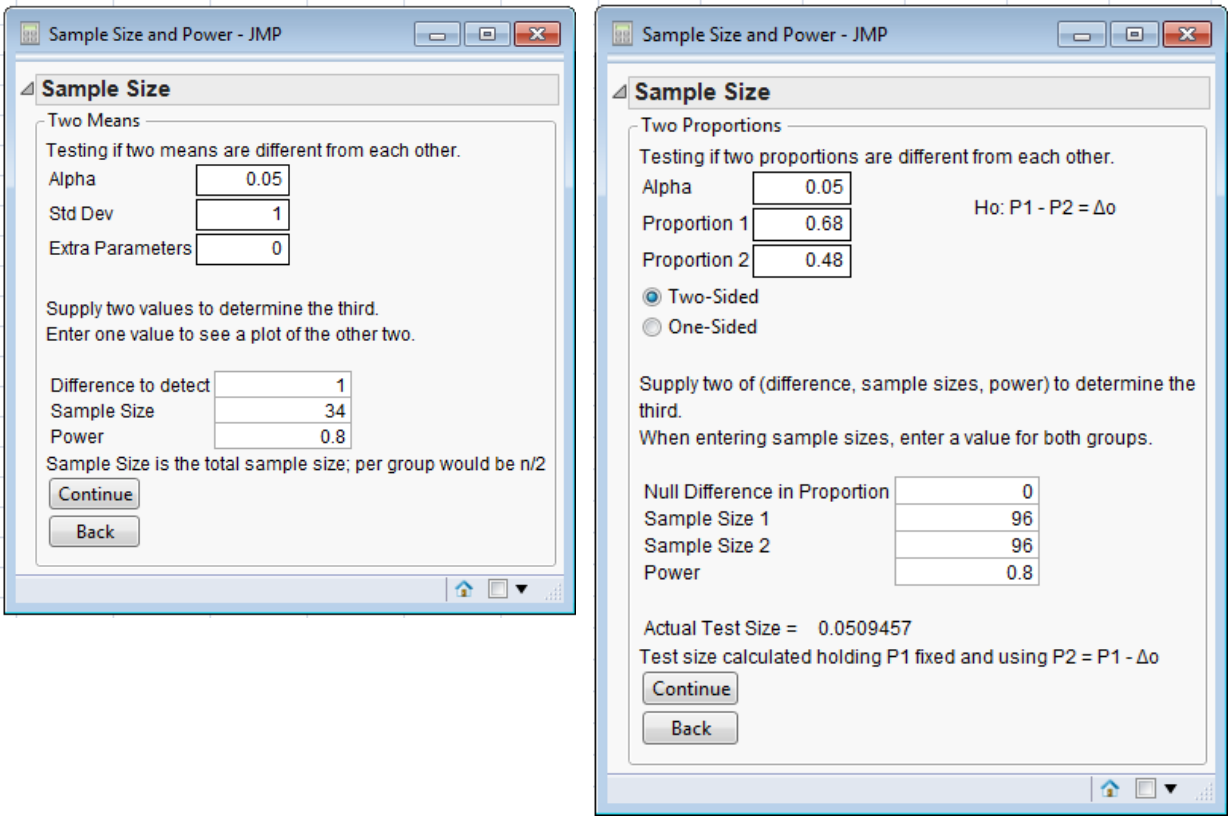


Figure 5: Comparison of system A versus B.

The goal again is to design a comparative experiment capable of detecting this level of change (with a 95% confidence and a power of 80%). If a continuous response is used, then we are testing whether the two means differ from each other, with a delta of 10 (the signal/difference in distance to detect) and the standard deviation equal to 10 (noise in the response). These values can be expressed as a signal-to-noise ratio, in this case 10/10 or 1. Using any sample size calculator readily available in most statistical software or online (see Figure 6(a) for example JMP input) we can determine that the number of samples needed from both systems is 17 (34 total samples).



(a) **(b)**

Figure 6: JMP input and output for a) two means and b) two proportions

If a binary response is used, then we are testing if two proportions are different from each other (0.68 and 0.48). Again using a sample size calculator (see Figure 6(b) for JMP input) we can determine that the number of samples needed from both systems would need to be 96 (192 total samples). This is almost six times the number of samples required if a continuous response had been used instead.

About Factors

Generic Missile Targeting System Example Revisited

Returning back to our Missile Targeting System example, let's suppose the purpose of testing is to determine what factors affect the error distance between the system generated target coordinates and the actual surveyed coordinates of the target. Figure 7 is a fishbone diagram generated from a brainstorming session with system subject matter experts.

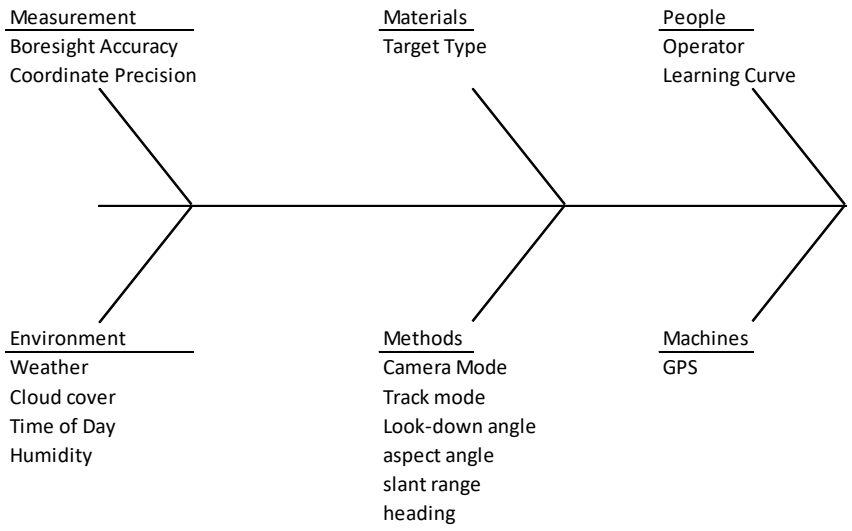


Figure 7: Fishbone diagram for missile targeting system

Level settings for many of these factors could be described in either numeric or categorical terms. For example, the test designer could set levels for slant range to be 6800 ft or 30000 ft or could simply use ordinal (unit-less) terms like "near" or "far". If the designer is generating a 2-level factorial design, using a numeric or categorical data types for this factor won't really affect the size of the overall design. For example let's say we wish to build a design that will vary Look Down Angle, Slant Range, and Aspect Angle, each at two levels. We will come up with a simple 2^3 design (8 runs) whether or not factors are categorical or numerical (see table 2(a) and (b)).

Table 2: 2^3 design for missile targeting system using (a) binary factors versus (b) continuous factors

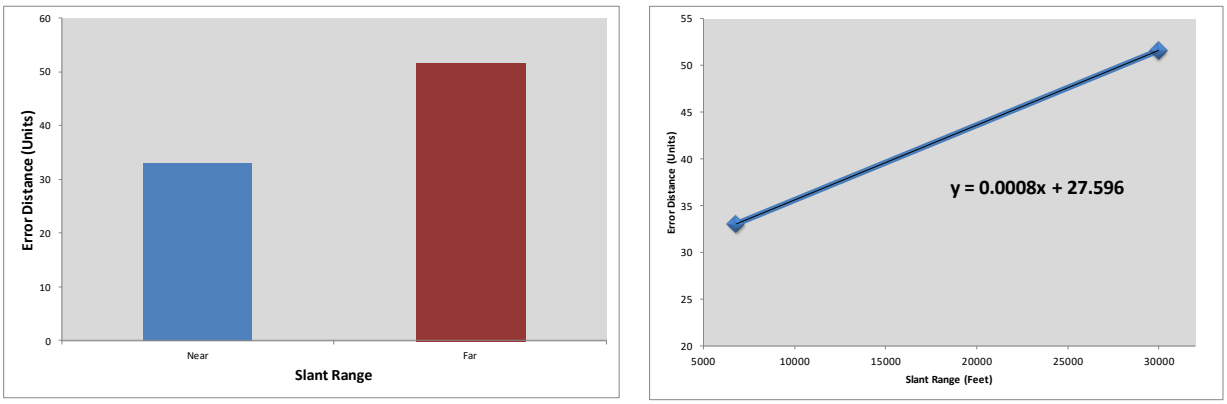
Run	Factor A: Look Down Angle	Factor B: Slant Range	Factor C: Aspect Angle
1	14	Near	0
2	68	Near	0
3	14	Far	0
4	68	Far	0
5	14	Near	180
6	68	Near	180
7	14	Far	180
8	68	Far	180

(a)

Run	Factor A: Look Down Angle	Factor B: Slant Range	Factor C: Aspect Angle
1	14	6800	0
2	68	6800	0
3	14	30000	0
4	68	30000	0
5	14	6800	180
6	68	6800	180
7	14	30000	180
8	68	30000	180

(b)

However, since there is information lost it will obviously affect the analysis and the final empirical models generated from the data. If a categorical data type is used one cannot infer about anything happening between level settings, essentially you have a different model for each factor level (see Figure 8(a)). However, if a numerical data type is used (actual or coded units for example) you can build a simple regression model that can be used to interpolate what's happening between levels (see Figure 8(b)).



(a) **(b)**

Figure 8: Interpretation of response with (a) binary factors versus (b) continuous factors

Another shortfall of using categorical factor settings comes if we decide we want to test for curvature (or nonlinearity) in the design space. This leads to the subject of pseudo center points, where because there is no actual single center value for categorical factors, center points for the numerical factors are duplicated at each level of each categorical factor. In a 2 level design this doubles the number of center points for EACH categorical factor in the design. So in the case of our 2^3 factorial with 2 numeric and 1 categorical factor 5 center points (the recommended amount) to test for curvature and estimate pure error would require a total of 10 runs as opposed to 5 runs (see tables 3(a) and (b)).

Table 3: 2^3 design for missile targeting system with center points, using (a) binary factors versus (b) continuous factors

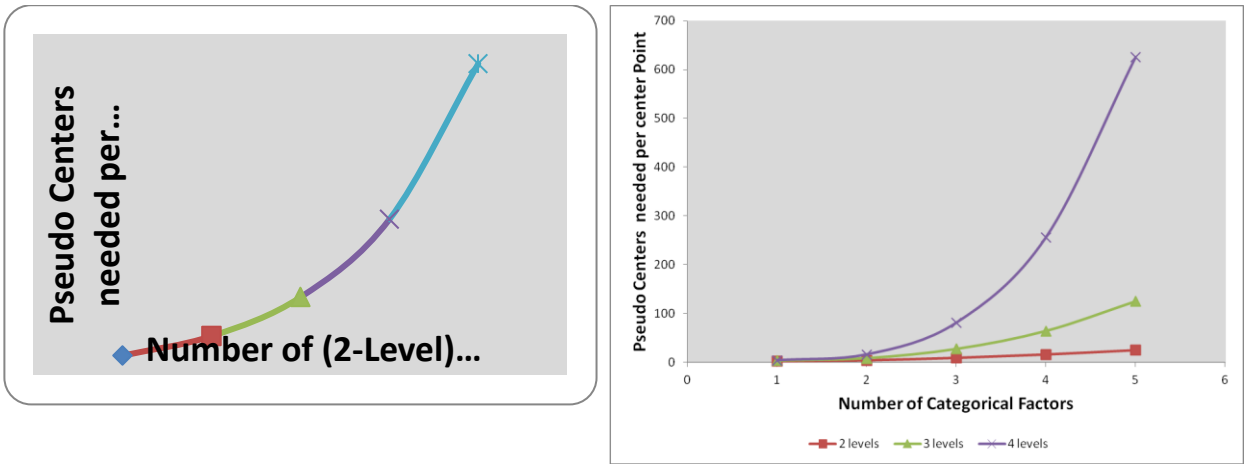
Run	Factor A: Look Down Angle	Factor B: Slant Range	Factor C: Aspect Angle
1	14	Near	0
2	68	Near	0
3	14	Far	0
4	68	Far	0
5	14	Near	180
6	68	Near	180
7	14	Far	180
8	68	Far	180
9	41	Near	90
10	41	Far	90
11	41	Near	90
12	41	Far	90
13	41	Near	90
14	41	Far	90
15	41	Near	90
16	41	Far	90
17	41	Near	90
18	41	Far	90

(a)

Run	Factor A: Look Down Angle	Factor B: Slant Range	Factor C: Aspect Angle
1	14	6800	0
2	68	6800	0
3	14	30000	0
4	68	30000	0
5	14	6800	180
6	68	6800	180
7	14	30000	180
8	68	30000	180
9	41	18400	90
10	41	18400	90
11	41	18400	90
12	41	18400	90
13	41	18400	90

(b)

The number of pseudo centers increase rapidly as both the number of categorical factors and the number of levels for each increase, see Figure 9(a) and (b). Note no more information is gained despite the larger design.



(a) **(b)**

Figure 9: Number of pseudo centers needed as the number of categorical factors increases

Conclusion

Using a continuous data type for a response measure over a binary (pass/fail) metric maximizes test effectiveness. Tests using a continuous measure require a much smaller sample size and produce tighter confidence intervals than tests using just a binary response. In essence, using a binary response is equivalent to discarding 38% to 60% of the test runs Cohen (1983). Also, if specification limits are provided, a continuous response can easily be converted to a binary (hit/miss) response however it is not possible to convert the binary response to the continuous distance measure after testing completes. That information is lost when viewing the system performance on the binary scale.

Using a continuous data type for describing factor levels also has advantages over using a categorical data type. The type of data type used for factor settings affects the analysis and the quality of the final empirical models generated from the data. If a categorical data type is used, one cannot infer anything about the response between level settings. Also, more center points will be needed if the user wishes to test for curvature (nonlinearity) in the design space.

Unfortunately, there will be cases where numeric data types cannot be used, in part 2 we will introduce a tutorial and calculator to help properly size a designed experiment when a binary response is used. In part 3, we will illustrate how to conduct the analysis and interpret the results using logistic regression.

References

Cohen, Jacob. “The Cost of Dichotomization.” *Applied Psychological Measurement,* vol. 7, no. 3, 1983, pp. 249-253., doi:10.1177/014662168300700301.

Hamada, Michael. “The Advantages of Continuous Measurements over Pass/Fail Data.” *Quality Engineering,* vol. 15, no. 2, 2002. pp. 253-558.