

INSTITUTE FOR DEFENSE ANALYSES

Power Analysis Tutorial for Experimental Design Software

Laura J. Freeman, *Project Leader*
Thomas H. Johnson
James R. Simpson

November 2014

Approved for public release;
distribution is unlimited.

IDA Document D-5205
Log: H 14-000639

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-0001, Project BD-9-2299(90), Test Science Applications, for the office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments

The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Ms. Denise Edwards, Dr. Colin Anderson, Dr. Steve Movit, Dr. Dan Pechkis, and Dr. George Khoury of the Operational Evaluation Division, and Dr. Dennis DeRiggi of the System Evaluation Division.

Copyright Notice

© 2014 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-5205

Power Analysis Tutorial for Experimental Design Software

Laura J. Freeman, *Project Leader*
Thomas H. Johnson
James R. Simpson

Executive Summary

A. Purpose and Overview

The Department of Defense (DoD) Test and Evaluation (T&E) community is increasing its employment of Design of Experiments (DOE), a rigorous methodology for planning and evaluating test designs. An essential capability that DOE provides is the ability to quantitatively and qualitatively assess the adequacy of a test design. Assessing the adequacy of the test design involves evaluating:

- the goals of the test
- the response variables (or measures)
- the range of possible test conditions (factors and levels)
- the amount of testing, in terms of the number of test points and where they are placed across the test region.

The last consideration is addressed quantitatively by the calculation of statistical power. Since power is one of the primary quantitative metrics used to determine test adequacy, it is important that we understand what it is and generally how it is computed.

This guide provides both a general explanation of the power analysis and specific guidance to successfully interface with two software packages, JMP and Design Expert (DX). A detailed discussion of how to interact with the software is necessary because the software packages make different assumptions that can result in different, or even misleading estimates of statistical power. The guide provides recommendations for inputs for statistical power calculations between the different software packages for both continuous and binary response variables.

In a designed experiment, statistical power is the probability that we conclude that a factor matters (or, more generally, that a model term matters), given that it truly does matter. Power analysis for a designed experiment involves setting or estimating several parameters including:

- the number of factors (and number of levels for each factor)
- the proposed statistical model
- the number of test points dedicated to estimating error
- the acceptable levels of statistical error
- the desired detectable change in the response (δ)

- the magnitude of the system noise variability (σ), and
- the statistical model anticipated.

While all of the above affect statistical power, the two most important assumptions are the estimates of δ and σ. The ratio of these two quantities is often referred to as the signal-to-noise ratio (SNR).

B. Software Package Approaches to Power Analysis

This guide focuses on Design Expert and JMP products because of the robustness of their experimental design packages. Many other good software programs exist to construct experimental designs, but both DX and JMP provide all of the analysis capabilities that the Director, Operational Test and Evaluation (DOT&E) has requested in evaluating test designs. DOT&E provides the guidance for all DoD operational testing. For a detailed description of how DOT&E reviews test designs please see the July 23, 2013 DOT&E memorandum, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation."

Unfortunately, the various software packages and their versions use different terminology and default methodologies in the calculation of statistical power. These defaults can lead to different power calculations between organizations that might be using different versions of software, and sometimes to misleading results. The primary difference between packages lies in the definition of detectable difference in the SNR.

C. Guidebook Overview

This guidebook provides an overview of power calculations and detailed instructions for calculating power across a variety of software packages. The first chapter introduces statistical power for designed experiments, highlighting key points and essential assumptions that can lead to different power estimates. Additionally, the chapter provides an overarching framework for calculating statistical power. The first chapter concludes by introducing the software packages and the notation used by each package.

Chapter 2 of the guidebook outlines power calculations for two-level factors. The power calculations discussed in this section apply to continuous factors, two-level categorical factors, and interaction effects for both continuous and two-level categorical factors. Estimating power is straightforward for two-level factors across all software packages.

Chapter 3 focuses on estimating power for multiple-level categorical factors. Here significant differences in power exist depending on the assumptions. The section strongly recommends that users do not accept JMP 11 default coefficients without first understanding the implications.

The final chapter provides default values for conducting power calculations in each of the software packages that can be used when program-specific information is not available.

The appendices of the guidebook outline approaches for calculating power for binary response variables and provide mathematical details behind the power calculations.

D. Conclusions and Recommendations

Testers should always try to base the SNR on the specifics of the test that is being planned. The detectable difference (or signal) should be based on what differences are operationally significant using input from operators and other subject matter experts. The noise estimate should be based on past test data collected in similar conditions whenever possible. Pilot tests provide excellent estimates of the noise in many cases.

However, when reasonable estimates of the SNR ratio are not available, we can provide some guiding principles based on past operational test experience. The final chapter of this user guide recommends using a SNR (δ/σ) between 1.5 and 2.0 and a 95 percent confidence level. Larger values (up to 2.0) should be used only in highly controlled test environments. Smaller values (less than 1.0) drive extremely large tests and have not resulted in operationally significant results in previous tests. These SNR values only apply to continuous response variables. Recommendations for specifying the SNR for binary responses are provided in Appendix B.

Additionally, we can control for differences in the software packages by using the scaled values for the SNR. Table 1, below, provides recommended default values.

Table 1. Recommended Inputs for Signal-to-Noise Ratio in Software Packages

Software	2 Level Factors/ Continuous Factors/ Interactions for 2 Level Factors	Multiple Level Categorical Factors and their Interactions	Quadratic Terms
Design Expert 8, 9	δ/σ*	δ/σ	$\delta/2\sigma$
JMP 9	$\delta/2\sigma$	$\delta/2\sigma$**	$\delta/2\sigma$
JMP 10	δ/σ	δ/σ	δ/σ
JMP 11	Under advanced options use "apply delta for power" of δ/σ	Under advanced options use "apply delta for power" of δ/σ Adjust all but two coefficients to zero (conservative method described in Chapter 4)	Under advanced options use "apply delta for power" of δ/σ

*If using the generic signal-to-noise ratios suggested in the previous section this value would be between 1.5 and 2.0.

**Dividing the signal-to-noise ratio by 2 only provides an exact power calculation to match the other packages for two-level factors. JMP 9 only provides power calculations for coefficients and is not comparable to the other packages. However, using this value typically provides reasonable test sizes, despite the limitations in the power calculations.

Contents

1. Introduction – Power Analysis Concepts 1-1
- A. Motivation 1-1
  - 1. Motivating Example 1-1
- B. Guide Overview and Intended Use 1-2
- C. Power for a Designed Experiment 1-3
  - 1. Overview 1-3
  - 2. Essential Elements of Power Calculations 1-5
  - 3. Statistical Model 1-6
  - 4. Factor Effect Power versus Coefficient Power 1-7
  - 5. Error Degrees of Freedom 1-8
  - 6. Power Analysis Process 1-9
- D. Response Types - Continuous versus Binary Responses 1-9
- E. Power Analysis Process Flow 1-11
- F. Software Packages for Computing Statistical Power 1-13
- G. Summary of General Power Concepts 1-14

2. Power for Two-Level Designs 2-1
- A. Two-level Design Generation and Design Choices 2-1
- B. Two-Level Design Generation and Power in Design Expert 2-2
  - 1. Design Expert Test Design Generation 2-2
  - 2. Design Expert Power Calculations 2-3
- C. Two-Level Design Generation and Power in JMP 2-4
  - 1. JMP Test Design Generation 2-4
  - 2. JMP 9 and JMP 10 Power Calculations 2-5
  - 3. JMP 11 Power Calculations 2-6
- D. Two-level Design Power Overall Comparison 2-9
- E. Summary of Power for Two-Level Designs 2-9

3. Power for Designs with Multi-level Categorical Factors 3-1
- A. Introduction to Categorical Factors 3-1
  - 1. Design Efficiency – Achieved by Trimming Factors or Levels? 3-2
  - 2. Coding Categorical Factors and Factor Parameters 3-4
- B. Power Analysis with Multi-level Categorical Factors 3-6
  - 1. Options for Conducting the Power Analysis 3-6
- C. Power Analysis using Design Expert 3-8
  - 1. Generating a Designed Experiment in Design Expert 3-8
  - 2. Calculating Power Once a Design is Generated 3-10
- D. Power Analysis using JMP 11 3-13

1. Introduction 3-13
2. Specifying Anticipated Responses 3-14
3. Specifying Anticipated Coefficients 3-15
4. Specifying Power using Advanced Options 3-19
5. Default Anticipated Coefficient Power 3-20
6. Configuring the JMP 11 Coefficients for Most Conservative Power 3-21
7. JMP 11 Power Reporting 3-27
8. Alternative Power Specification (JMP Semi-Conservative) 3-32

E. Power Comparison across Packages 3-33
F. Power Analysis Practice Tips 3-36
G. Summary of Power for Multi-Level Categorical Factors 3-39

4. Conclusions and Recommendations 4-1

A. Extension to Additional Analysis Model 4-1
B. Summary of Results 4-1
C. Power Analysis Software Recommendations 4-2
1. General Recommendations for Risk Specification 4-2
2. General Recommendations for Signal-to-Noise Ratio Estimation 4-2
3. General Recommendations for Software Inputs 4-4
D. Overall Recommendations 4-5

References R-1

Appendix A – Acronyms A-1
Appendix B – Binary Response Power B-1
Appendix C – JMP 11 Power Calculation Details C-1
Appendix D – Design Expert Power Calculation Details D-1
Appendix E – JMP Monte Carlo Simulation Script E-1

1. Introduction – Power Analysis Concepts

A. Motivation

The Department of Defense (DoD) Test and Evaluation (T&E) community is increasing employing Design of Experiments (DOE) as a methodology for planning and evaluating test designs. As the adoption of DOE (or experimental design) increases, it is essential that we consider both quantitative and qualitative aspects of test adequacy. Three major aspects of test planning collectively answer the adequacy question. The first and most important aspect is whether we are attempting to solve the right problem. Do we have our objectives stated correctly and completely? The second aspect, which only careful, team-based planning can provide is whether all the relevant performance measures are listed, and whether the associated test design(s) span the range of possible test conditions (factors and levels). The third aspect, and the reason for this guide, is whether we are planning to test too little, just the right amount, or too much relative to the insight we need for the stated objectives. This third consideration is addressed by the calculation and assessment of statistical power.

Power is one of the primary metrics used to determine test adequacy, so it is important that: (1) we understand what it is and generally how it is computed, and (2) how to interact with selected software to obtain accurate power values for a given design strategy. This guide provides both a general explanation of the power analysis strategy and specific guidance to successfully interface with two software packages, JMP and Design Expert. The guide addresses three versions of JMP and two versions of Design Expert, and provides recommendations for inputs for statistical power calculations across these different software packages for both continuous and binary response variables.

1. Motivating Example

Figure 1-1 shows the power for multiple designed experiments each with two factors, but with varying numbers of levels from two to six. The power is provided for different versions of JMP software and Design Expert. Notice for two-level factors the power is consistent across all packages, but for categorical factors with multiple levels the power changes, often dramatically, between software packages. This power guide will explain those changes and provide recommendations.

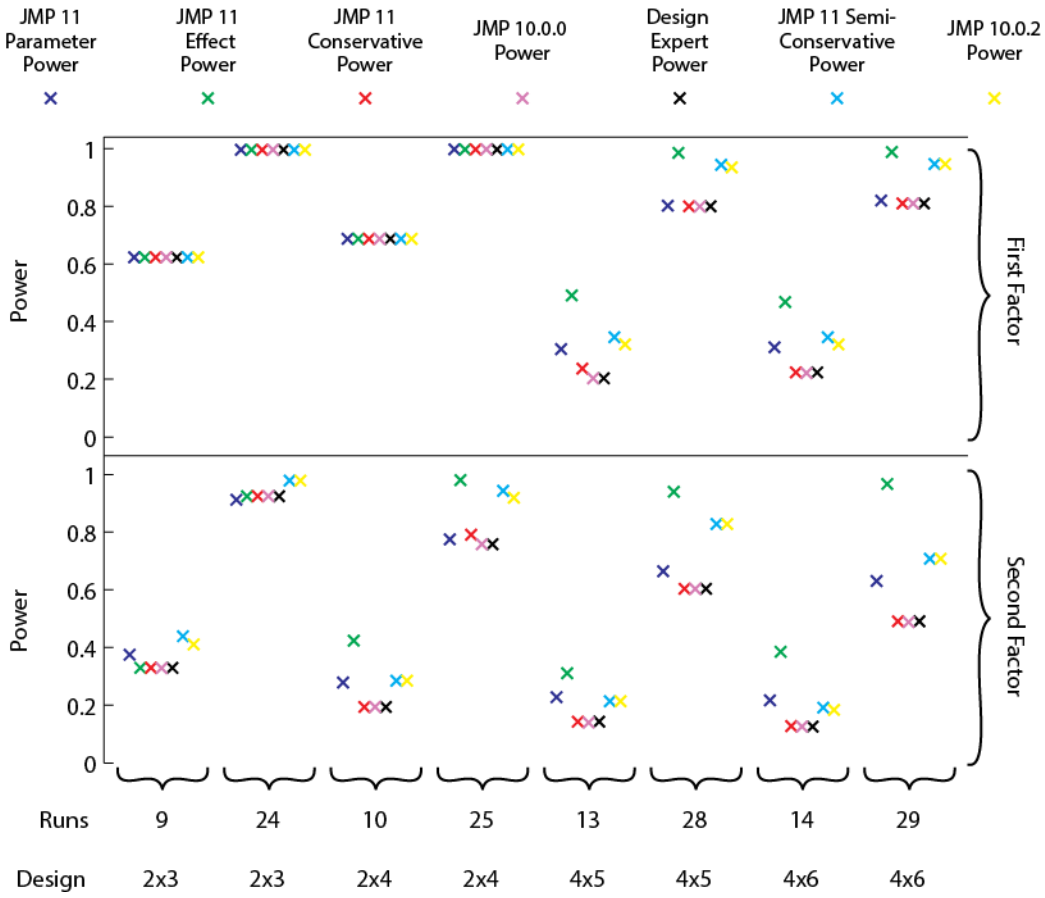


Figure 1-1 Power analysis comparison across software platforms using a signal-to-noise ratio of 2 and all methods discussed in this guide.

B. Guide Overview and Intended Use

This guide provides both a general explanation of the power analysis procedure and specific guidance to successfully interface with two software packages, JMP and Design Expert (DX). A detailed discussion of how to interact with the software is necessary because the software packages make different assumptions that can result in different and/or misleading estimates of statistical power. The guide provides recommendations for inputs for statistical power calculations across the difference software packages for both continuous and binary response variables.

This guide is intended for analysts who need to use statistical software packages to calculate power. While the appendices provide the detailed mathematics behind the power calculations, the main body of the guide is intended to walk users through the software packages without sidetracking to cover the mathematical details. A user should be able to use the guide to calculate and interpret power from any of the packages discussed.

The remainder of this chapter provides an overview of statistical power for designed experiments, highlighting key points, and essential assumptions that can lead to different power estimates. We discuss the selection of response variables in the context of statistical power. That discussion is followed by descriptions of all the relevant parameters involved in power analysis, along with an overarching framework for calculating statistical power. This framework is intended to guide the user through the key choices they must make when calculating statistical power. Finally, we introduce the software packages and the notation used by each package.

Chapter 2 of the guidebook details power calculations and considerations for designs with two-level factors. The power calculations discussed in this section apply to continuous factors for main effects (first order model), two-level categorical factors, and interaction effects for both continuous and two-level categorical factors. Estimating power is straightforward for two-level designs with both factor types and is consistent across all software versions. The guidebook walks through the process in both JMP and DX software packages.

Chapter 3 focuses on estimating power for categorical factors with more than two levels. Here the software packages make different assumptions of the questions of interest, which is important for the user to understand. The differences are discussed and summarized.

The final chapter of the guidebook highlights important recommendations, expands the recommendations of the guidebook to additional statistical models not covered by the guide, and provides recommendations for specific inputs to each of the packages.

C. Power for a Designed Experiment

1. Overview

One of the primary practices in ensuring test design adequacy is to sufficiently mitigate risk associated with the probabilities of drawing incorrect conclusions post-test. The method known as statistical power analysis is used mostly to determine the number of runs (also called design points or test events) needed in order to control the two types of error probabilities (α and β) in testing. The two types of errors manifest themselves in a number of ways in statistics. In DOE they are associated with probabilities of either incorrectly concluding that a factor matters in affecting system performance when it truly does not (α), or concluding that a factor is not influential when it really does affect system performance (β). Power and the corresponding errors for a designed experiment are defined below:

α = Probability (the test conclusion is that a factor matters, given the factor has no effect)
β = Probability (the test conclusion is that a factor has no effect, given the factor matters)
Power = **1-β** = Probability (the test conclusion is that a factor matters, given the factor matters)

Consider the simple example of an air-to-air missile operating both in low and high clutter environments. The primary response variable for this simple example is the miss distance (MD) and it is a function of just a single factor, the level of clutter in the environment. In this simple experiment, the null hypothesis (H_0) is that clutter has no effect on the missile miss distance. The alternative hypothesis (H_1) is that clutter does have an effect on the missile miss distance. Figure 1-2 illustrates the α and β probabilities under the null and alternative hypotheses. The standard process for calculating power is to set an acceptable α error (step a – b), then to compute the β error (step c – d). Statistical power is defined as the complement (1- β) of the β error. In this example, power is the probability that we conclude clutter does impact the missile miss distance, when it truly does have an effect. Note that the α and β errors are depicted as areas (probabilities) under a probability distribution. In this example the reference distribution shapes drawn are notional. It is often the case that miss distance response variables are not symmetric. While non-symmetric distributions are not discussed in this guide, it is typically reasonable to treat non-symmetric distributions the same as symmetric distributions for power calculations because the shape of the distribution has less of an effect on power than many of the other assumptions.

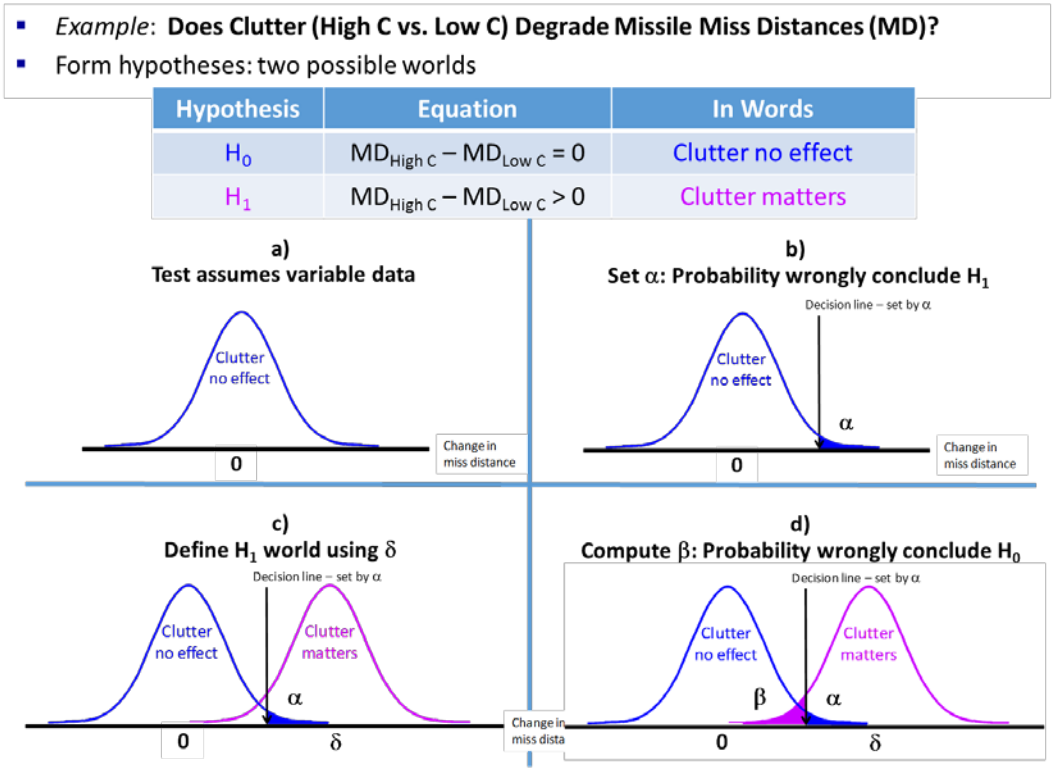


Figure 1-2. Air-to-Air Missile Example: sequence for determining statistical power

Once α is set, the objective is to size the test such that a high statistical power is achieved. The relationship between power and sample size is one of marginally decreasing returns. Power grows rapidly initially but as the number of trials continues to increase, power improvement slows. Assuming a stable system under test and little chance of missing data, extra trials to obtain power values above 95 percent are usually not necessary.

2. Essential Elements of Power Calculations

Most tests are more complex than the notional one-factor air-to-air missile experiment described above. Power analysis for a designed experiment involves setting or estimating parameters for:

- the number of factors (and number of levels for each factor)
- the number of test points dedicated to estimating experimental error (due to system noise)
- an acceptable α risk
- the desired detectable change in the response (δ)
- and the magnitude of the system noise variability σ.

Figure 1-3 outlines all of the elements necessary for calculating power from a designed experiment. The numbers of factors and levels tend to arise from the planning process, although those parameters can also influence the final power or test size value. **The most useful endeavors in power analysis investigations are in obtaining accurate estimates for the detectable difference (δ) and the system noise (σ).** The ratio δ/σ is also called the signal-to-noise ratio (SNR). Subject matter experts for the system under consideration are the best sources of information in determining the detectable difference, while pilot studies or historical data of similar systems under like conditions usually lead to sufficient noise estimates. In cases where the response is suspected to be non-symmetric and historical data are available, the estimates of δ and σ can take into account the shape of the distribution.

The test planning process and subject matter expert are essential in producing defensible power estimates. Test team input into power calculations include: the responses and number of factors which often come from the test team's process decomposition, estimates of σ from historical or pilot data, α is based on acceptable risk, and δ is uncovered in discussion with system or technical experts.

Parameter	Description	How Obtained	Relevance in Planning
k: **factors**	Number of factors in the experiment	Determined in process decomposition	Key finding from process decomposition
df_{error}: **model error**	Amount of data reserved for estimating system noise	Desired model order (e.g. interaction, quadratic)	Estimate of complexity of input-output relation
α: **alpha**	Probability (declaring factor matters when it doesn't)	Set by test team	Fix and leave alone
δ: **delta**	Size of response change expert wants to detect	Experts and management determine	Some ability to vary
σ: **sigma**	System noise – run-to-run variability or repeatability	Historical data; pilot tests; expert judgment	System driven but can be improved by planning
1-β: **power**	Probability of declaring a factor matters when it does	Lower bound set by test team	Primary goal is to set *N* to achieve high power
N: **test size**	Number of test events including replication	Usually computed based on all other parameters	Direct, should modify to satisfy power

Figure 1-3. Parameters included in a power analysis, along with a description of each and ways to provide estimates. The gear diagram shows the sequence for setting or estimating each parameter.

3. Statistical Model

In addition to the items outlined in Figure 1-3 it is important to consider the resulting statistical analysis that will be conducted as a result of the test. A simple math characterization of a system is:

$$y = f(x) + \varepsilon$$

where y is the response, the x's are factors, so ε represents differences between the math model and the observed outcome (due to noise). The parameter σ is the standard deviation of that noise. It is important to note is that σ is estimated with the effect of factors removed. Hence the data used to estimate σ should be data obtained from similar operating and environmental conditions (factor settings). Using data collected under dissimilar conditions will excessively inflate the noise estimate such that the estimate would represent variability in excess of noise variability.

Prior to constructing a design, the test team must consider alternative anticipated statistical models and determine which polynomial form (e.g., first order plus interaction) best aligns with the test objectives (e.g., screen, vs. characterize vs. optimize), and which types of model terms might be significant for that system under study. One common polynomial form is the main effects (or first order) model which is given by:

$$y_i = \beta_0 + \sum_{j=1}^{k} (\beta_j x_j) + \varepsilon_i$$

where k is the number of factors, the x_j are specific settings for each factor j, β_0 is the overall intercept, and β_j are coefficients reflecting the change in the response per unit change in x for each factor. This first order model allows for shifts in the overall mean as a function of the factors. For continuous factors the shift in the mean is a linear function. For categorical factors shifts in the mean apply to each level of the categorical factors. Therefore, for categorical factors with more than two levels, more than one value of x is needed to account for the appropriate mean shifts.

A first order plus interaction model is by far the more prevalent model form used as the general model, because it captures the preponderance of significant effects occurring in real world systems, assuming the objectives are screening or characterization. This model provides more flexibility in the analysis of the test outcomes. A first order plus interaction model is:

$$y_i = \beta_0 + \sum_{j=1}^{k} (\beta_j x_j) + \sum_{j=1}^{k} \sum_{l=j+1}^{k} (\beta_{jl} x_j x_l) + \varepsilon_i$$

This model adds commonly occurring two-way interaction terms for all factors, to the first order model. Higher order models can be generated by adding quadratic polynomial terms, which is a popular model form for optimization objectives.

Connected with the statistical model an important assumption for multi-level categorical factors is the relationship of the coefficients to the overall factor effect. This assumption is at the root of all differences in the software packages. Therefore, by understanding the differences, we can understand the power calculations that software provides.

4. Factor Effect Power versus Coefficient Power

We previously defined δ as the desired detectable change in the response or detectable difference. Essentially the focus is on the factor effect. Another approach to power analysis instead considers assessing the change in the model regression coefficients. However, when we perform hypothesis testing on the coefficient of the factor, we are formally testing if the model coefficient β is significantly different from zero. Therefore, a translation must be made between the difference that we seek to detect in the response and the difference we seek to detect in the coefficient. For a two-level factor, this translation is straightforward, in that the coefficient detectable difference is half of the response detectable difference.

For multiple-level categorical factors this calculation is not straightforward because we must define exactly which of the levels we expect to produce the change in the response through a customized contrast. The situation is further complicated in that there are many ways to code the contrasts for a given factor and design, affecting the final power outcomes. These contrasts are discussed in more detail in Chapter 3.

However, for now it is important to note that there are different types of power calculations. Factor effect power is based on the hypothesis that we are looking for any difference in outcomes for any level of the factor. So if at least one level of the factor has a significant effect the factor effect hypothesis test will capture this. On the other hand, coefficient power tests the statistical significance of each coefficient. For two-level factors there is only a single coefficient per factor, so effect and coefficient power calculations are equivalent. For multiple-level categorical factors the two are clearly different because multiple coefficients make up the overall factor significance.

5. Error Degrees of Freedom

To conduct statistical significance testing in building a statistical model, it is essential to plan a test to collect sufficient data to estimate the error term, which can be assessed by the error degrees-of-freedom. Degrees-of-freedom is a concept formally tied to the rank (number of independent rows or columns) of the model matrix in quadratic form. Practically, the concept of degrees-of-freedom refers to the number of independent elements (one per experimental run) available to estimate model parameters or error. The total degrees-of-freedom value for an experiment is usually equal to the number of runs minus one, to account for the grand mean or model intercept. This total is then partitioned into degrees-of-freedom required to estimate the model terms, plus those degrees-of-freedom needed for error. An insufficient number of degrees-of-freedom dedicated to estimating error can drastically alter power downward. Figure 1-4 shows power for a 5-factor 2^{5-1} half fraction factorial design, and assumes the general model contains the full first order plus interaction set of terms, for a total of 16 model degrees-of-freedom. The plot shows that power can change from 0 percent (saturated model with 0 error degrees-of-freedom) to over 90 percent in as few as five additional runs due to sensitivities in the number of available error degrees-of-freedom.

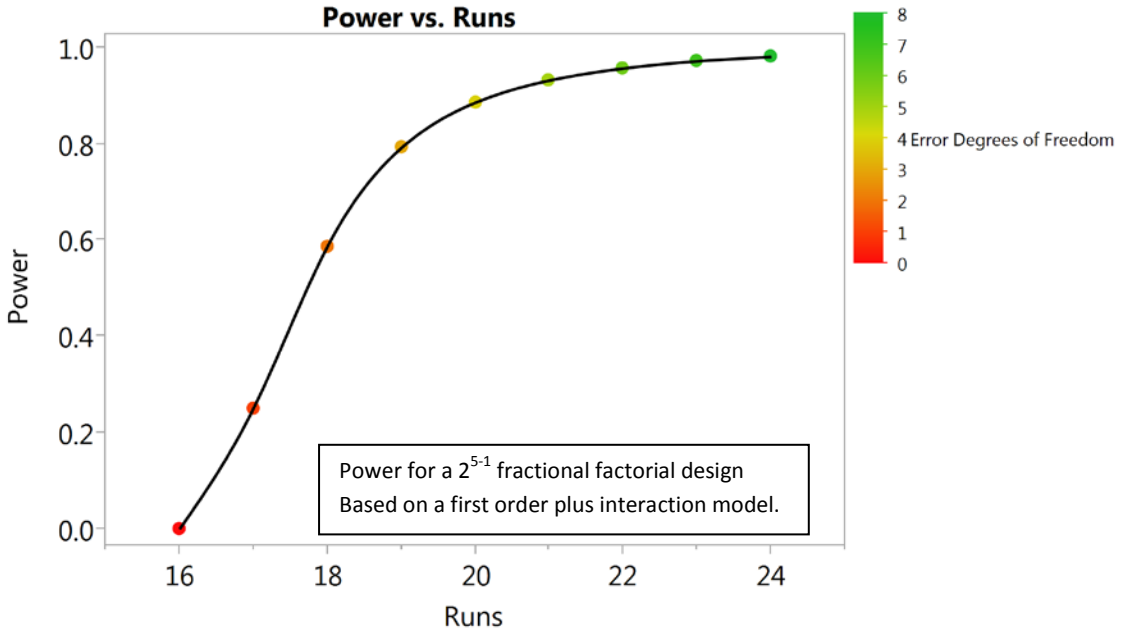


Figure 1-4. Power as a function of error degrees-of-freedom. Power plotted for a 2^{5-1} fractional factorial design with SNR=2 and an assumed two-factor interaction model.

6. Power Analysis Process

The standard power analysis process involves manipulating the sample size (number of test points) until an acceptably high statistical power is achieved. It is during this process that the analyst utilizes statistical software to iterate on the design test size until the desired power is obtained. Recall, that power is reported per factor and per response, so even for a single designed experiment, a number of power values should be tuned and reported. It is also important to note that numerous parameters must be estimated in order to compute power, so a low precision estimate (integer percent values) is probably more than adequate. It cannot be overstated, the key to right-sizing a test is the hard work put in by the test team to obtain appropriate and accurate estimates of δ (from discussions with system experts) and σ (from actual data).

D. Response Types - Continuous versus Binary Responses

Before we detail the steps for conducting a power analysis with various statistical software packages, it is very important to recognize the nature of response variables, and the impact response variable types have on the richness of the information acquired, and hence on statistical power. If the test is planned using a binary response (pass/fail, detect/no-detect, hit/miss, lock/break-lock, yes/no, etc.) as the primary or only measure of performance, then the estimated power value will be markedly lower than the same design using a continuous response variable. Figure 1-5 shows power curves for a design

with four two-level factors, comparing continuous versus binary response variables. A large increase in sample size is required for the binary response to achieve the same power. Appendix B discusses useful approximations of the SNR for binary responses, which allows for using either DX or JMP directly.

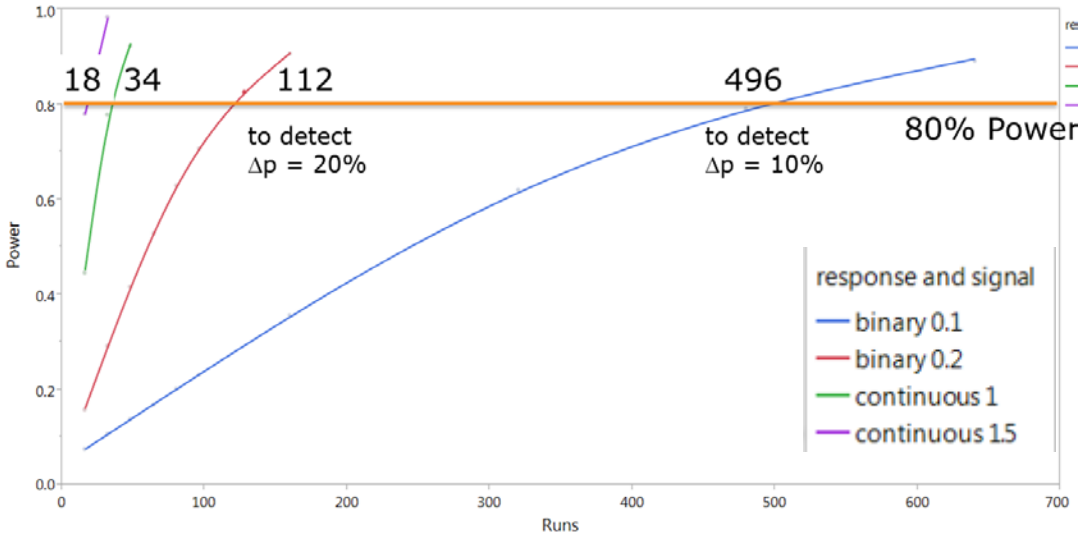


Figure 1-5. Sample size for binary versus continuous responses for a 2^4 full factorial design to achieve 80 percent statistical power

As a general rule more information is obtained from the outcome of a test when the measures of effectiveness or suitability are continuous random variable (assuming measurement error is minimal and the measure appropriately reflects system performance associated with test objectives). The best type of response for data analysis, empirical modeling and ultimately improved knowledge of system behavior is one that varies continuously over a wide range of values, and that is largely affected by system factor changes, and less by noise. Continuous responses not only allow for statistical model predictions across the range of performance, but require substantially fewer test runs than a design based on a binary response. Take for example the Joint Precision Airdrop System (JPADS), with steerable parachutes and an onboard computer to direct loads to a designated point of impact in a drop zone. Suppose the requirement for JPADS is that the loads land within, say 200 yards of the designated point of impact. Two response options are a binary pass/fail (inside/outside the 200-yard ring) and miss distance (in yards) from the target point. The miss distance response is vastly superior in information content per test condition and should be the primary response, but there is no reason not to collect, model and report both measures.

It is often beneficial and useful to collect binary response random variables alongside their continuous counterparts, for the same objective. However, whenever possible the test size and power analysis should be based on the continuous response variable. Two of the more compelling reasons for collecting binary responses are: (1) for

each test condition, they directly answer some specification or system requirement, and (2) they are easily interpreted and reported. A third reason for assessing binary responses is that they can provide additional insight relative to the story told by the continuous measure, especially if regarding a target end state. Miss distance is a popular continuous response, but it does not always perfectly correlate with its binary counterpart (e.g., target destruction). So adding binary responses of battle damage assessment kill/no-kill is informative. Level of destruction is a good example of an intermediate type of response (ordinal) that provides less information than a continuous response, but more than a binary one. It is important to recognize though that as the response variable types progress from continuous towards binary, less information is retrieved per test run, and so by comparison more and more data are needed to accurately answer the same questions.

E. Power Analysis Process Flow

There are several actions to take and decisions to make during the course of a power analysis, which also involves the need to iterate on the design in order to achieve the desired power values. All of the steps and considerations along the way are addressed in this guide. Figure 1-6 provides a process flow diagram that may be helpful to visualize the entire process. Some of the steps have been described already, but others are covered in more detail in the sections to follow.

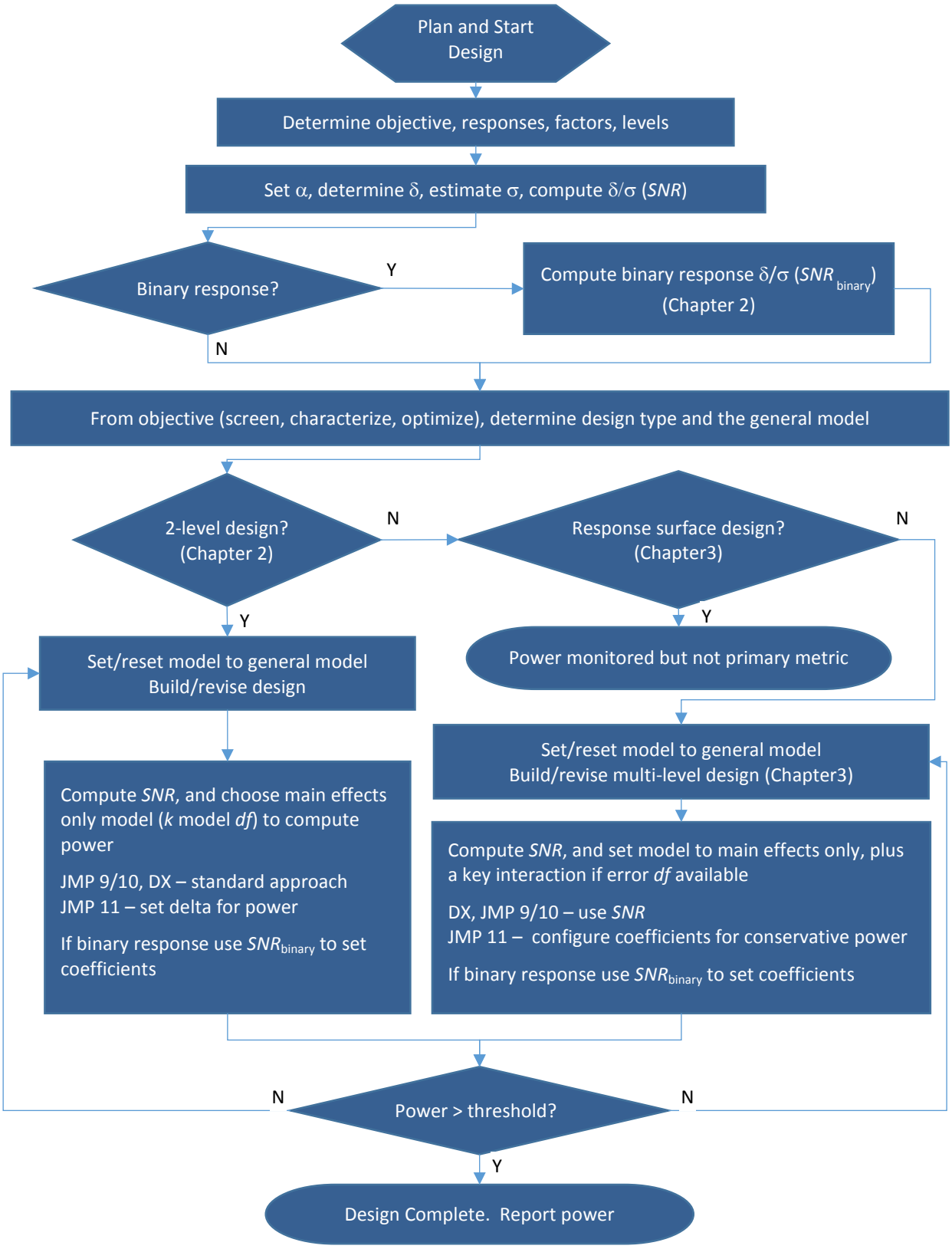


Figure 1-6. Power analysis process flow diagram

This iterative process is integral to building an effective and efficient design. Needed up front are specifics about the design to include response types, factors, levels, and the design objective (e.g., screen, characterize). From a power analysis perspective, the user must also supply α, δ, and σ. If a binary response is primary, there is an additional step to take prior to building the design, to determine the binary SNR value. Based on the design objective, the user proceeds to build the design and interacts with the power values to increase or decrease the design size until desired power values are obtained.

F. Software Packages for Computing Statistical Power

Several very capable DOE software packages are available. The quantity and diversity of statistical power analysis programs within these packages has increased. This guide provides recommendations to IDA and DoD analysts on general power analysis guidelines, cautions on pitfalls to avoid, and suggestions for working with specific software versions. This guide focuses on Design Expert and JMP products due to the robustness of their experimental design and analysis capabilities. Several other high quality DOE software programs exist to construct and analyze experimental designs. However, both Design Expert and JMP provide all of the analysis capabilities that DOT&E has requested in evaluating test designs. For a detailed description of how DOT&E reviews test designs please see the July 23, 2013 DOT&E memorandum, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation."

Probably the most inconsistent aspect of software enabled power analysis across software platforms is the notation and terminology used to describe the desired detectable response change and the system noise. Table 1-1 shows the various terms used to capture essentially the same information.

JMP Statistical Software (JMP v.11, 2014) has undergone the most dramatic changes in its design of experiments power analysis platform, particularly in JMP version 11. The primary purpose of the new interface is to provide the user flexibility to shape the nature of the design based on user prior knowledge or information, particularly for categorical factors with more than two levels. Computing power for factors with more than two categorical levels is not trivial and requires assumptions regarding the expected differences among factor levels. This guide provides some general power analysis insights along with example-based guidelines for successfully accomplishing power assessments using JMP and Design Expert.

Table 1-1. Terminology in Software to Request or Report Delta (δ) and Sigma (σ) Estimates for Power Analysis

Software	Delta	Sigma	Delta/Sigma
Design Expert 8, 9	**Delta**, Difference to detect in the response, "Signal", refers to the change in the response.	**Sigma**, Est. Std. Dev., "Noise"	**Delta/Sigma**, Signal/Noise Ratio*
JMP 9	Implied, as **Signal** but cannot enter directly, refers to a change in the coefficient as opposed to the response	Implied, as **Noise** but can't enter directly	**Signal to Noise Ratio** (for coefficient)
JMP 10	Implied, as **Signal** but cannot enter directly, refers to a change in the response.	Implied, as **Noise** but cannot enter directly	**Signal to Noise Ratio** (for response)
JMP 11	Indirectly either using Anticipated Responses or Anticipated Coefficients, or directly using **delta** under Advanced Options)	**Anticipated Root mean square error (RMSE)**	If using Advanced Options, and Power Analysis interface, then **delta/RMSE,** assuming RMSE = 1; delta refers to a change in the coefficient

* Note: In Design Evaluation, several default delta/sigma ratios (0.5, 1.0, 2.0) are shown as e.g., 2 Std. Dev.

G. Summary of General Power Concepts

- Two test risks should be considered in test planning, α and β. Both are probabilities associated with incorrect conclusions based on a pair of complementary hypotheses conjectured prior to test.
- Of the two risks, the α risk is set prior to designing the test. The β risk is usually computed then iterated on by changing the test size until β is sufficiently small.
- Power is a probability (1- β) and is the complement of the β risk associated with test.
- Because of the way we address the two risks, power becomes the final risk typically managed in design construction.
- Power depends on the total test or sample size, the α risk, and also the size of the change in response the test team seeks to discover, the noise standard deviation, the number of test factors, and the anticipated model.
- There are many different assumptions in the anticipated model that can affect the power calculations. Different software packages use different assumptions so it is important for the user to understand these assumptions.

- Higher power values are desired, and while designs can be under-powered, right-sized or over-powered, but because of resource restrictions we are usually striving to right-size a test that would otherwise be under-powered.
- Continuous responses are vastly more informative than categorical responses, especially binary categorical responses, so work hard to identify continuous responses as primary measures of interest.
- Power is only one of many design metrics, but one of the more important indicators of test design adequacy.

2. Power for Two-Level Designs

This section provides a user's guide for calculating power for two-level design. These designs provide a good introduction to power calculations for each of the software packages because differences between the packages are relatively minor and easily explained due to the simplicity of the designs. The next chapter discusses power calculations for multiple level categorical factors, which are significantly more complex in terms of their power calculations.

A. Two-level Design Generation and Design Choices

The most common two-level designs are 2^k full factorial and 2^{k-p} fractional factorial regular designs. Non-regular two-level optimal designs are also addressed by this section. Typically in a two-level design the goal of the test is screen for important factors affecting the test outcomes. Two-level designs are the most efficient and powerful method for screening for important factors, whether they are numeric (continuous or discrete) or categorical.

Two-level designs support first order models (main effects) and interaction effect. The degree of the interaction effect supported by the design depends on the design type. Full factorial designs support all possible interactions. Regular 2^{k-p} fractional factorial designs are often categorized by the resolution of the design:

- **Resolution III** designs support modeling of main effect. However, some main effects may be indistinguishable from some two-factor interactions.
- **Resolution IV** designs support modeling main effect and some two-way interactions. However, while all main effects can be estimated independently, some two-factor interactions will be indistinguishable from other two-factor interactions.
- **Resolution V** designs support modeling main effects and two factor interactions. All main effects and two-factor interactions can be estimated independently from each other. Some two-factor interactions may be indistinguishable from some three-factor interactions.

Resolution II designs are not viable alternatives because main effects are perfectly aliased (confounded) with other main effects. Designs with resolution > V are also available. Typically a Resolution V design is considered a safe and robust test design strategy.

Optimal fraction designs are another class of two-level designs that allow for the specification of a completely customized model. Depending on the sample size selected and the exact model selected optimal fraction designs may result in partially correlated model coefficients. This is in contrast to regular fractional factorial designs which have either zero correlation between model terms or complete confounding (100 percent correlation) between model terms. Optimal fraction designs can be useful for generating test designs for sample sizes that are not equal to a power of 2 (i.e., 8, 16, 32, 64, etc.).

There is a trade-off in design capabilities between selecting anticipated model terms (main effects and interactions) to be perfectly correlated (aliased) as they are in a regular 2^{k-p} fractional factorial and partially correlated (correlations >0 and <1) as they are in an optimal fraction design. Partially correlated terms allow for the possibility that the estimated effects in question are sufficiently uncorrelated such that reasonable estimates can be obtained. If you choose designs where main effects are partially correlated with interactions or two-factor interactions are partially correlated with each other, it is highly recommended that when analyzing the data, use a model building variable selection technique such as stepwise regression to find the likely important effects. Adding runs to partially aliased designs is recommended although it can be difficult to determine the exact runs which offer the most value in uncovering the correct model.

By contrast, designs with perfectly correlated model effects (regular fractional factorials) do not have the model estimation issues and possible variable identification issues due to partial correlation, but they have their own challenges. Sequential testing is key to success in these designs, as it should be with partially aliased designs. To uncover the correct model in these designs, the first step is to analyze the data from the initial design and identify the important effect sets (alias chains) via typical hypothesis tests and least squares fitting. The next step is to attempt to determine the most logical contributors to each significant effect set using the principles of effect sparsity and model heredity. For example, a main effect aliased with two or more four-factor interactions yields a simple solution that with near certainty the main effect is driving the large effect observed. Two-factor interactions aliased with other two-factor interactions can often be resolved via model heredity, which is a long-standing empirical finding that significant interactions carry along one, or more likely both of their main effects as significant too. The important third step is to identify the unresolved effect sets to decouple via a small set of additional tests, and execute those runs and analyze the combined two sets of data.

B. Two-Level Design Generation and Power in Design Expert

1. Design Expert Test Design Generation

The process for generating two-level designs in Design Expert is relatively straight-forward for regular ($N=2^r$, where $r=2, 3, 4, \ldots$) designs, while non-regular or optimal

options are also available. Two-level full factorial and regular fractional factorial designs are available under the *Two-Level Factorial Design* section. Figure 3-1 below shows a screen shot for generating regular two-level designs. Notice in the figure that Design Expert provides the resolution of the design with color coding, identifying resolution V designs and higher as optimal.

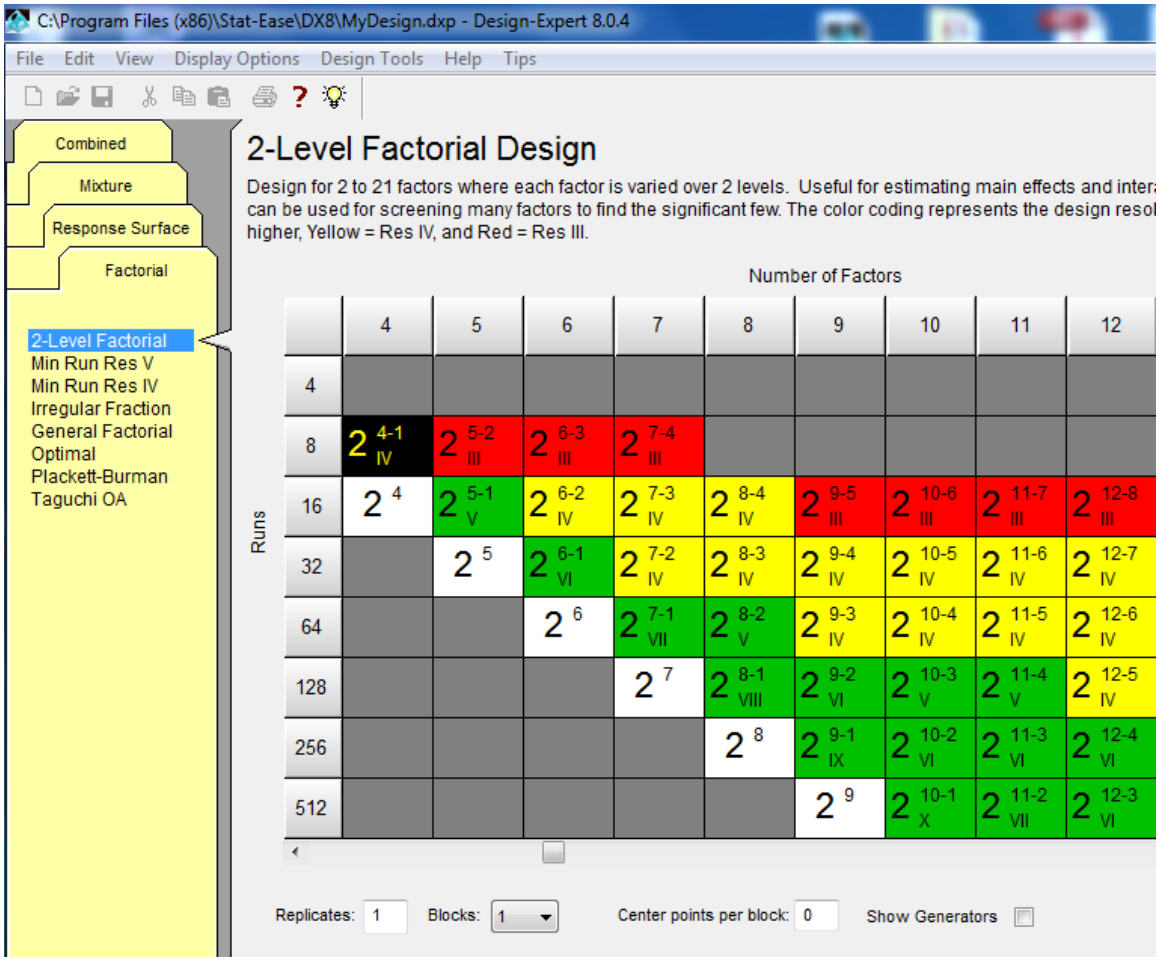


Figure 2-1. Design Expert Regular Two-Level Design Interface

Options are available to replicate the design to improve power and add center points. Center points are useful for continuous variables for checking for deviations from the linearity assumption implicit in the two-level designs.

The *Optional* Design option under the factorial tab in Design Expert provides the ability to generate non-regular fractional factorial designs.

2. Design Expert Power Calculations

Design Expert provides power calculations in two locations. The first is provided before the test design is finalized. The Design Expert power wizard provides a tool where the user can directly input estimates for δ and σ. Design Expert defaults to using a

main effects model for estimating power, which on the surface appears inconsistent with our standard recommendation to build designs to fit the main effects plus interaction model. The rationale supporting only main effects for power analysis is, due to effect sparsity, not all the main effects and interactions are typically significant. Only a subset of the main effects plus interactions will be significant, and the total number of main effects is a reasonable value for the purposes of power analysis. The final statistical model is expected to contain some combination of main effects and interactions, but all that is needed for power analysis is an adequate estimate of the number of model degrees of freedom. If more than *k* model terms are anticipated in the final model, this default can be changed by selecting the "Edit model for power…" button. Again, the most important consideration here is to select an appropriate *number* of anticipated effects, and it doesn't matter which effects are checked, and in fact we really don't know prior to test. For 2-level designs selecting main effects or interactions is immaterial as all model terms, whether main effects or interactions, require only 1 degree of freedom for the model.

The second location that Design Expert provides power is after the design is finalized in the *Design→Evaluation→Results* output. Here power is provided for three signal-to-noise ratios (SNRs). The defaults are 0.5σ, 1.0σ, and 2.0σ, denoted. The default SNR can be changed under *Design→Evaluation→Results→Options*. The default α level for Design Expert is 0.05, which is recommended unless extenuating circumstances justify an alternative value. The default value of 0.05 can be changed under *Edit→Preferences→Math.*

C. Two-Level Design Generation and Power in JMP

1. JMP Test Design Generation

In JMP, two-level designs can be generated using Custom Design, Screening Design, or Full-Factorial Design. The Custom Design procedure provides the largest flexibility in constructing designs and the ability to evaluate the design inside the design generation application. To generate a two-level design in the Custom Design tool:

1. Input the factors, factor types, and levels
2. Choose the appropriate model, usually main effects plus 2-factor interactions (select Interactions, 2nd).
3. Choose the desired number of runs. If practical, choose a regular design with $N=2^k$ runs, where *k* is integer. Remember that the single replicate full or fractional design is only the base design, and additional runs are needed for replicates, center points, augmentation to decouple interactions, augmentation for second order terms, or to increase power, and finally additional runs for validation.
4. If not choosing a regular two-level design, be sure to select as many model terms as possible (make them Necessary) based on available model degrees of freedom, then include the remaining desired model terms (If Possible).

5. In assessing the constructed design be sure to perform not only a power analysis, but check for perfect aliasing (factor effects perfectly correlated), estimation efficiency, and color map correlations.

Please note that, although it is possible to build classical 2^{k-p} designs in JMP *DOE→Custom Design*, we suggest you select *DOE→Screening Design*.

The following tips apply to users of JMP regarding initially specifying the desired anticipated model vs. trimming the model to a realistic size for assessing power.

- When building designs, note the default model when constructing a custom design is main effects only. Be sure to specify the fully desired anticipated model (usually at least including all two-factor interactions) prior to building the design.
- If the desired number of runs is less than the number of model term degrees of freedom (less than the *Minimum* in the JMP interface), go back to the model and change the setting on some or all of the included two-factor interaction terms to, *If Possible* (requires left mouse click under *Estimability* heading).
- Once the design is built, check the aliasing or correlations using the alias matrix and the color map on correlations
- To perform power analysis after the design is built (after *Make Design* and *Make Table*), from the data table:

1. Choose DOE→*Evaluate Design.*
2. Place the Response(s) in Y, Response and the factors in X, Factor.
3. Under Model, select a model with approximately the right number of model terms. A main effects model is often a good choice.
4. Page down to the Power tab and check the model terms to ensure the changes you made are reflected in the power section. Read the power values from either the coefficient power or the factor effect power – should be the same for 2-level designs.

2. JMP 9 and JMP 10 Power Calculations

Power analysis is found under the *Design Evaluation* section in all versions of JMP. This section appears in the custom design tool after the design is generated. In JMP 9, power calculations are found under the *Relative Variance of Coefficients* tab. As the title of the tab suggests the power calculations are for the change in the coefficient instead of the change in the response variable. JMP 9 provides a box to change the significance level and enter the SNR above the power calculations. For two-level designs the SNR is based on the anticipated change in the regression coefficient, and not the effect (or change in the response). Accordingly, in JMP 9 be sure to divide your delta/sigma ratio by 2 before entering that value as the SNR.

JMP 10 power calculations follow a similar process to JMP 9, except there is an inherent difference in the use and interpretation of SNR. Both versions refer to the

numerator of the SNR as signal, but whereas JMP 9 is expecting the user to supply the anticipated change in the regression coefficient, JMP 10 is expecting the signal to refer to the change in the response, or the effect (twice the regression coefficient for a two-level design). The rationale for the JMP 10 interpretation of the signal to be the anticipated change in the response (as opposed to the regression coefficient), is that the change in the response seems to be more natural for the system expert (usually an engineer) to determine. During test planning, the engineer might be asked directly, "what size change in the response (use the response units) do you want to detect during testing?"

Figure 3-2 shows power estimates from JMP 9 and 10 to illustrate the difference between interpretations of SNRs. Power is shown for five different factorial designs at three different SNRs. The SNRs are user inputs to the JMP interface. As seen in the graph, the power for JMP 9 at a SNR of 0.5 is equivalent to the power for JMP 10 at a SNR of 1.0. A similar case is evident for SNRs of 1.0 and 2.0. JMP 9's interpretation of SNR is two times larger than JMP 10's, which greatly inflates JMP 9's power estimates compared to JMP 10. The general guidance is to use JMP 10 given the choice, and if using JMP 9, then divide your anticipated SNR by 2 before inputting the value into JMP 9.

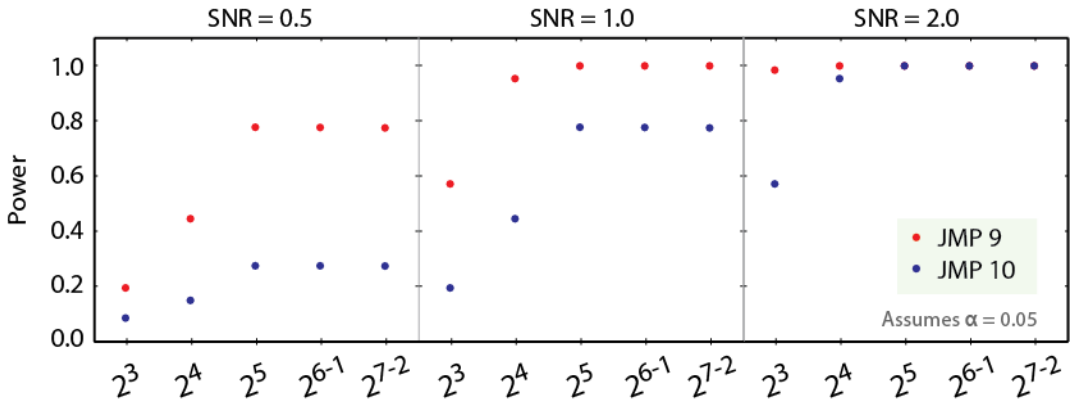


Figure 2-2. Comparison of JMP 9 and JMP 10 Power Calculations

3. JMP 11 Power Calculations

The JMP 11 power analysis interface is different from all previous JMP versions. The new interface provides multiple options and flexibility in entering the information necessary for power analysis. This increased flexibility makes matching the power calculations from any other package possible, but also increases the chances for error.

For two-level factors, the default JMP 11 and JMP 10 power computations agree. This is because the default anticipated coefficients used result in a δ/σ ratio (or *SNR*) of two. Differences do exist for categorical factors with 3 or more levels, and are discussed extensively later in discussions on designs involving categorical factors more than two levels.

The JMP 11 interface provides three primary methods for entering the information for power analysis:

- Entering the anticipated responses
- Entering the anticipated coefficients
- Entering a general SNR (similar to JMP 9 and 10).

a. Power Using Anticipated Responses

JMP 11 provides the option to directly enter the anticipated response for each run of the test design. While this option provides useful educational information by illustrating the translation between anticipated responses and anticipated coefficients, in practice it is nearly impossible to know the anticipated response for each test condition.

b. Power Using Anticipated Coefficients

The anticipated coefficients are all defaulted to a value of one, which implies a two-unit change in the response as the factor changes from the low level (-1 in coded) to the high level (+1 coded). If a different SNR is desired, just set the anticipated coefficients to *SNR*/2. This method is most useful if prior knowledge is available such that the anticipated change in the response per factor varies and this is known prior to test. For example, suppose that a four-factor design is being considered and it is anticipated that two factors (say A and B) will cause twice the change in response as the other two factors (C and D). One could set the anticipated coefficients for A and B to 1, while setting C and D to 0.5. The resulting power analysis will show lower power values for C and D, everything else the same.

c. Power Using *SNR* in Advanced Options

JMP 11 provides the traditional method of entering a delta and sigma, but is not immediately obvious. The developers of JMP 11 were sensitive to a potential need for a JMP 11 power analysis input option similar to the JMP 10/9 interfaces (and similar to Design Expert). They added the ability to independently specify the anticipated signal (or δ), and the noise (σ) estimate (anticipated RMSE). There are two options discussed below for entering the SNR, but Option 1 is simpler and recommended.

Option 1: Enter Delta/Sigma for *delta* – The easiest way to input the necessary information is to form your estimated δ/σ ahead of time as a ratio, similar to the form JMP 10 requests. That ratio will then be entered into JMP as a delta, assuming $\sigma = 1$. To enter the δ/σ (which is the same as δ for $\sigma=1$), go to the red triangle at the top of the dialog box to the left of Custom Design. Left mouse click on the triangle, choose *Advanced Options, Set Delta for Power*. It asks to enter the delta for power and notes

that the anticipated coefficients will be half of this value. Enter your estimated δ/σ value here. Note that the default sigma, referred to as *Anticipated RMSE* = 1.

Example: Planning has determined the test will have 6 factors, each with two levels. Initially, it is decided to construct a fractional factorial design with 16 runs, which allows for estimation of a main effects and some two-factor interactions, aliased in pairs. Choose *Make Design*. The Custom Design dialog appears with the power analysis interface along with other features that permit a detailed assessment of the proposed design (Figure 3-3). Assume the test planning has uncovered the following:

1. System experts determine the difference to detect, delta = $\hat{\delta}$ = 25.0
2. Historical data provides noise estimate, sigma = $\hat{\sigma}$ = 13.5.
3. Then, compute $\hat{\delta}/\hat{\sigma}$ = 25.0/13.5 = 1.85/1 = 1.85

Using Option 1, take the scaled $\hat{\delta}/\hat{\sigma}$ and enter 1.85 for delta, as it assumes $\hat{\sigma}$= RMSE = 1.

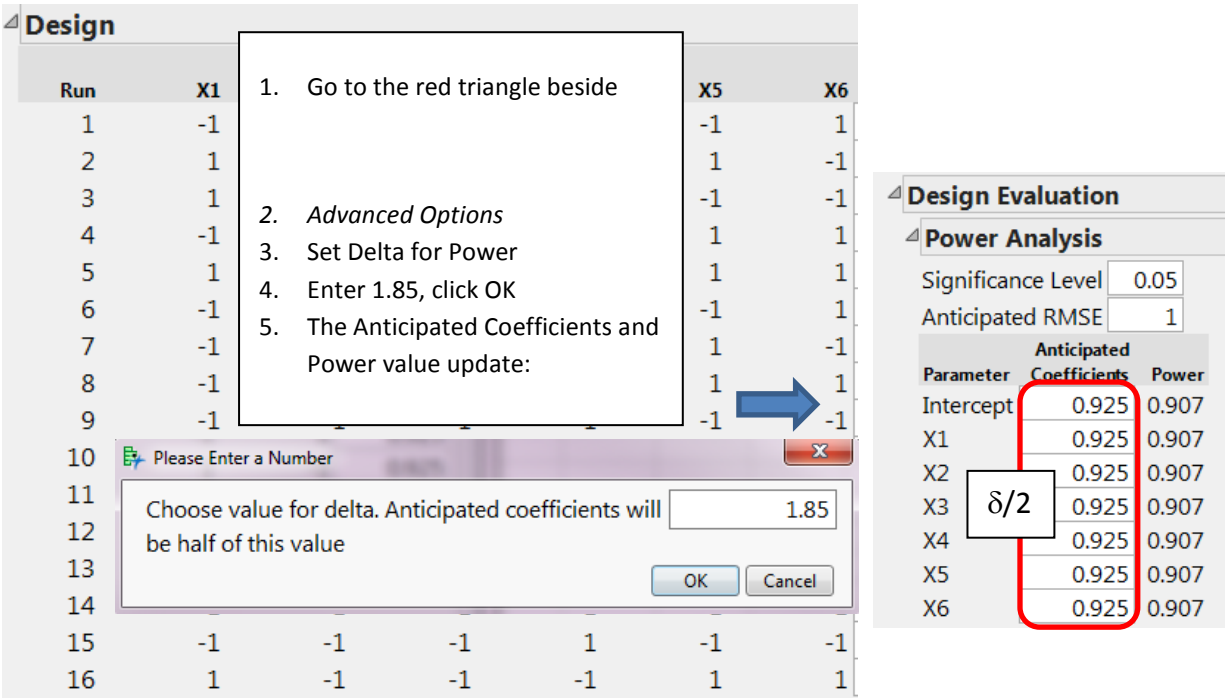


Figure 2-3. Assessing power in JMP 11 by specifying the estimated δ/σ ratio directly

Option 2: Provide the actual delta value in the *Set Delta for Power* option under *Advanced Options*, and then input sigma in the Anticipated RMSE input cell. The challenge with this approach is that the anticipated coefficients are no longer a function of delta, so they lose interpretation, and because the delta input cell is not visible, it makes it hard to see what you have set for delta and sigma. With option one, the anticipated coefficients provide some feedback regarding δ/σ or *SNR*, because of the relationship that with RMSE = 1, the Anticipated Coefficients = SNR/2.

For JMP 10 users, remember that the SNR is easily modified in the power analysis input section. Power values are reported as Effect Power, which is now different in JMP 11. This differences between JMP 10 and 11 Effect Power does not manifest itself in two-level designs, so suggestions for interacting with the JMP 11 interface are discussed in Chapter 4 regarding categorical factors with more than 2-levels.

D. Two-level Design Power Overall Comparison

For the same design, model and SNR, JMP 10 and 11, as well as DX report consistent power values. Design Expert and JMP 10 are setup quite similarly for two-level designs. The user inputs are nearly identical and the SNRs both assume the signal is the change in the response. While the JMP 11 interface is different, it still provides the same results. Figure 3-4 shows that all three software packages provide identical power estimates for two-level factorial designs.

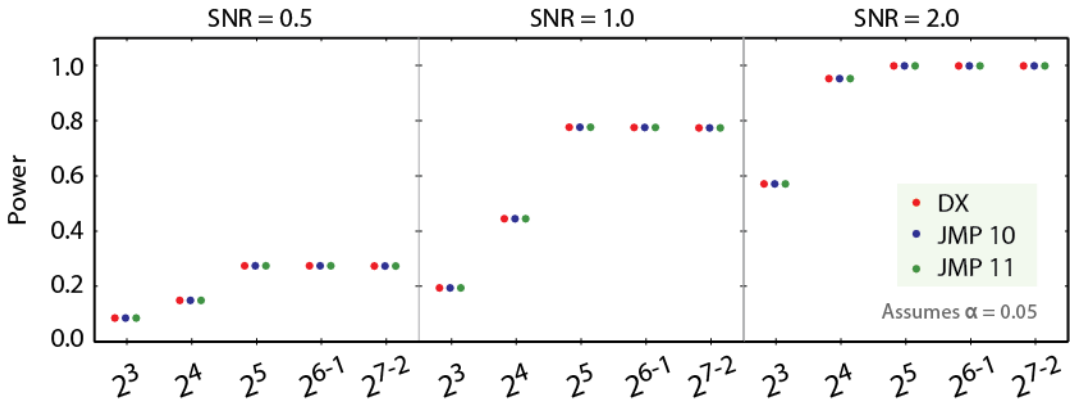


Figure 2-4. Two-level design comparison of DX, JMP 10, and JMP 11 power calculations

E. Summary of Power for Two-Level Designs

- Among the many benefits of two-level designs are efficiencies in the number of tests required, and the benefit that all model terms, regardless of complexity (main effects and interactions), consume only 1 degree of freedom each.
- For orthogonal (all design factor columns are pairwise linearly independent), balanced two-level designs, without knowledge that certain factors are expected to have larger effects, all model terms will have the same power.
- Power is adversely affected by partial aliasing in a design.
- Classic 2^{k-p} fractional factorials have model effects completely aliased, so power is for alias chains. Resolution IV and V designs have attractive coupling of model terms such that effect sparsity and model heredity often point to the correct model term interpretation. These designs can be built in JMP and Design Expert.

- The JMP 11 power analysis interface differs substantially from previous versions. Using the standard SNR to provide δ and σ is possible only through the red triangle in the *Advanced Options*.
- JMP 9 differs from JMP 10, JMP 11 and Design Expert. In JMP 9 the SNR signal represents the change in the regression coefficients, which is half the change in the response due to an effect used in all the other software variants. The resulting interpretation is $SNR_{JMP9} = ½\ SNR_{JMP10,\ JMP11,\ DX}$. If using JMP 9, multiply the computed SNR (based on δ and σ) by ½ before performing power analysis.
- For two-level designs, the power calculations are identical for JMP 10, JMP 11, and Design Expert.

3. Power for Designs with Multi-level Categorical Factors

A. Introduction to Categorical Factors

Many real world designs contain one or more categorical factors. That said, it is always recommended that the test team carefully study each categorical factor to determine whether it best thought of as numeric factor, either continuous or discrete or a discrete factor. Numeric factors provide higher power than multiple level categorical factors and also support more robust trend analysis. Discrete factors are appropriate when the groups are truly nominal. Consider a test designed to assess tactics against ground targets using various ordinance options. The objective is to determine a delivery mode and weapon that can be successful against a variety of targets on the ground, stationary or moving. Miss distance is a reasonable response variable. Table 3-1 shows two possible listings of the factors for this test.

Table 3-1. Possible Factors and Levels for a Test Characterized both as Categorical Levels and Numeric Continuous Levels

Factor	*Categorical Levels*	*Numeric Factor*	*Levels*
Weapon	GBU-10, GBU -16, GBU-12	Weapon Weight	500, 1000, 2000
Delivery	Loft, Level, Dive	Release Angle	+10, 0, -30 deg
Location	Eglin, Nellis	Visibility	5, 9 nm
Target Type	Car, Tractor Trailer	Target Size	60, 568 sq ft
Target Motion	Stationary, Moving	Target Speed	0, 30 mph
Time of Day	Day, Dusk, Night	Ambient Light	100, 500, 800 lumens
Range	Edge of Launch Acceptability Region (LAR), Center of LAR	Range	5, 10 nm

Hence, the first task when working with categorical factors is to convert as many as possible to numeric factors. Numeric factors provide the capability to make predictions of performance at levels not explicitly tested. One example of a more challenging conversion for the above problem would be a composite of weapon types such as GBU-12 (Paveway II), GBU-22 (Paveway III), AGM-65 Maverick and GBU-38 JDAM. Enough differences in guidance, control, propulsion exist such that categorical levels are most likely the better choice.

1. Design Efficiency – Achieved by Trimming Factors or Levels?

Introducing multi-level categorical factors to a test can complicate the design and analysis procedures from an experimental design perspective. The first challenge is that the analysis is limited only to those combinations prescribed as the categorical levels. For example, suppose that due to resource limitations, only the GBU-22 of the Paveway III class weapon was tested. Suppose further that the Paveway III showed the most promise, but because of its near miss performance, a heavier variant of the Paveway III might well have performed better. If weapon weight (within class) was another variable, the statistical model might have revealed the weapon weight most effective against the full classes of target types and target speeds.

The second aspect of concern with specifying categorical factors has to do with the number of factor levels. As the number of levels of a factor increases, statistical power typically decreases substantially for a fixed sample size. Table 3-2 shows the number of observations per level for factors with two, four, five, and ten levels. The number of observations per level (or pseudo-replication) is the main source of statistical power. Therefore, a test with 40 test points might have high power if all of the factors considered have two levels, but low power if one factor has two levels and the other factors have 3 or more.

Table 3-2. Distribution of Observations per Factor Level as the Number of Levels Increases

Factor	Levels	Obs per level (N=20)	Obs per level (N=40)	Obs per level (N=60)
A	2	10	20	30
B	4	5	10	15
C	5	4	8	12
D	10	2	4	6

From the table it is clear that increasing the number of levels decreases the number of observations per level in a linear fashion. It is this reduction in observations per level, coupled with the assumptions made about how many factor levels contribute to the factor effect that often leads to low power for the many-level factors in a test design.

The following graphs depict the relative effect on power due to the number of factors versus the number of levels of a factor. Figure 3-1 shows the power for each two level factors in a 16-run test. Note the relatively mild power reductions as the number of factors increases for two level factors. By contrast, Figure 3-2 shows dramatic power reductions as the number of levels (q) increases for a single factor. These graphs illustrate the benefit of converting multiple level categorical factors to continuous factors for statistical power.

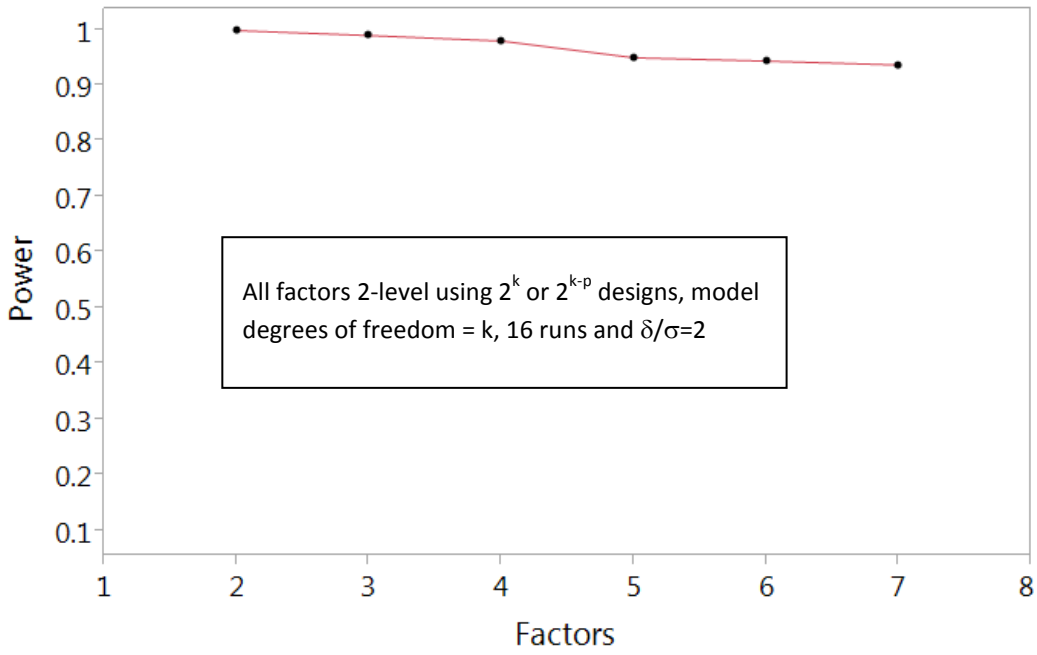


Figure 3-1. Statistical power of two-level full (*k*=2, 3, 4) and fractional (*k*=5, 6, 7), 16-run designs assuming the number of significant model terms = *k*

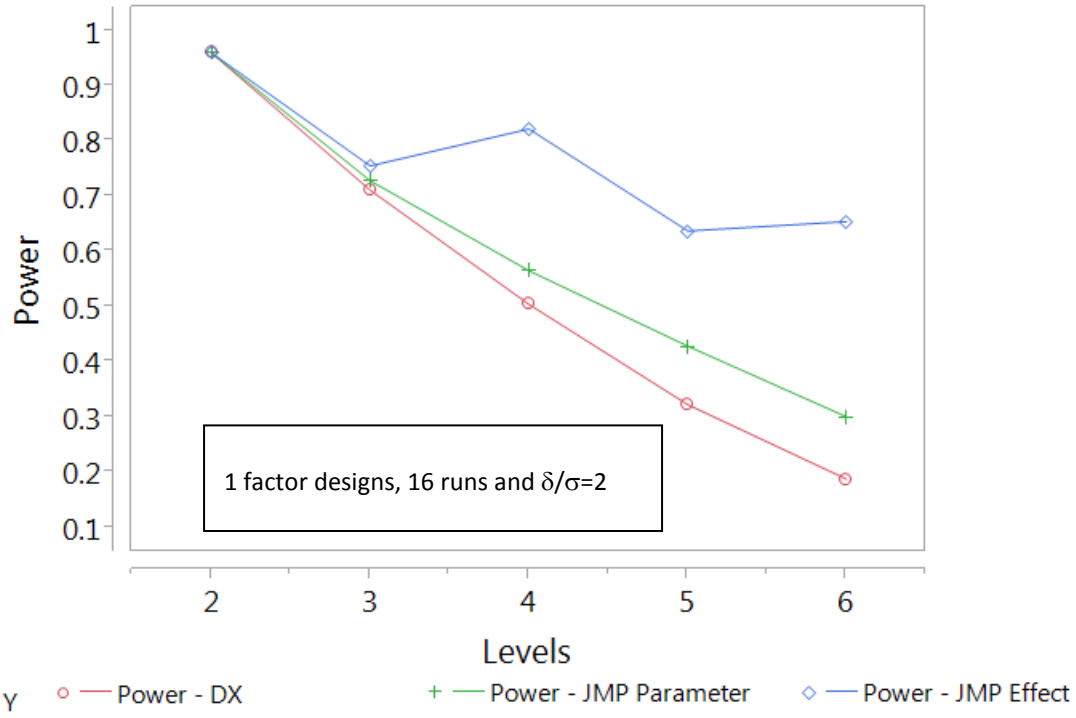


Figure 3-2. Statistical power for one-factor designs with multiple levels. Designs are all 16-runs, with replicates. JMP Power calculations are for version 11

Figure 3-2 also shows an interesting trend for JMP 11 power calculations. Previous versions of JMP match either the Design Expert power or the Coefficent power. However, the default coding for JMP 11 categorical factors results in this saw tooth power curve. This result is not intuitive until we understand the JMP 11 default coding structure and the graph suggests that users should not accept the default anticipated coefficient values given in JMP 11.

2. Coding Categorical Factors and Factor Parameters

For factors with more than two levels, analysts have the advantage of flexibility in choosing a way to parameterize the model that accounts for, and ultimately explains the factor effects. One of the most common approaches is to use a contrast scheme that permits a direct (all levels but one) and indirect (the last level) comparison between that level average and the overall average. Table 3-3 shows this coding strategy for a 4-level categorical factor. Most software packages (including JMP and Design Expert) use this choice of contrasts, sometimes called simple or effects coding.

Table 3-3. Nominal Factor Contrast Coefficients for a 4-level Factor using Simple Coding

Factor Level	*A[1]*	*A[2]*	*A[3]*
L1	+1	0	0
L2	0	+1	0
L3	0	0	+1
L4	–1	–1	–1

The scheme in Table 3-3 allows parameters A[1], A[2], and A[3] to estimate the differences from factor levels L1, L2 and L3 to the grand mean respectively, while the expression –(A[1]+A[2]+A[3]) is the L4 difference from the grand mean. As interactions are added to the model, direct interpretation of the main effects becomes more difficult, so it is recommended that graphical interpretation be used instead.

It appears by inspection of any A[i] column that the comparison for that variable is between the i^{th} level and the last level L(q). However, because all the contrasts have a value = –1 in that last cell, the comparison is ultimately with the grand mean. Consider the single 4-level factor example above, and a statistical model of the form,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

where,

y_i is the i^{th} observed response

$\beta_0, \beta_1, \beta_2 \; and \; \beta_3$ are parameters for the intercept, the first indicator variable parameter (such as A[1] above), the second, and the third factor parameter, respectively.

x_{ji}, represents the values for indicator variable j, for observation i

ε_i are the assumed independent and normally distributed error terms

$$x_{1i} = \begin{cases} 1 & if\ level = 1 \\ 0 & if\ level = 2 \\ 0 & if\ level = 3' \\ -1 & if\ level = 4 \end{cases} \quad x_{2i} = \begin{cases} 0 & if\ level = 1 \\ 1 & if\ level = 2 \\ 0 & if\ level = 3' \\ -1 & if\ level = 4 \end{cases} \quad x_{3i} = \begin{cases} 0 & if\ level = 1 \\ 0 & if\ level = 2 \\ 1 & if\ level = 3 \\ -1 & if\ level = 4 \end{cases}$$

The estimates for the means of each of the four levels of the factor, relative to the model parameters, as well as the grand mean μ, are

$$\mu_1 = \beta_0 + \beta_1$$
$$\mu_2 = \beta_0 + \beta_2$$
$$\mu_3 = \beta_0 + \beta_3$$
$$\mu_4 = \beta_0 - \beta_1 - \beta_2 - \beta_3$$
$$\beta_0 = \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} = \mu$$

Another common coding scheme is to use indicator variables for all but the last level of the categorical factor. However, the effect coding scheme shown above has the advantage that the grand mean is preserved in the intercept parameter β_0. From above, we see the remaining parameters are just the difference between the factor level average and the grand mean.

$$\beta_1 = \mu_1 - \mu$$
$$\beta_2 = \mu_2 - \mu$$
$$\beta_3 = \mu_3 - \mu$$

Not only is it instructive to see the result of what is referred to as *simple* or *effects coding* on the estimates of the means for each factor level, but the above model formulation also reveals that the values of the model parameters (sometimes called coefficients) directly contribute to the factor effects. Based on data from a test, these parameters are estimated and evaluated for statistical significance. The parameter values can be hypothesized prior to test as part of a power analysis. JMP does this in JMP 11, by providing default *anticipated coefficients* for each model parameter. During the following discussion, we will be consistent with JMP terminology in using the term "anticipated coefficients" when referring to the parameter estimates of indicator variables ahead of test as part of a power analysis. We will refer to the individual power values associated with each individual indicator variable as *coefficient power*. By contrast, *factor effect power* refers to the probability of correctly detecting a true effect from the

combined contribution of all the indicator variables for a factor or interaction. Note, as shown in Figure 3-1, the factor effect power and coefficient power are equal for any two-level factor or continuous factor.

The next sections are intended to assist a user in performing a power analysis with Design Expert and JMP 11. The input information and dialog boxes for JMP are different from those in Design Expert, so this guide provides a step-by-step guide for both Design Expert and JMP. More detail is provided for JMP 11, as it is substantially distinct from the other versions of JMP. The interfaces for JMP 10 and JMP 9 are nearly identical and relatively simple. The only substantive difference is the software's expectation for the SNR input. We start with considerations for power when initially constructing designs.

B. Power Analysis with Multi-level Categorical Factors

In building the design, the user must input the factors, factor types, levels, intended model and number of runs in building the design. The most important aspect (but easily bypassed) is the anticipated model. The anticipated model contains all the model terms the team is interested in estimating based on the design. For example, in screening it is often necessary to be able to estimate main effects and low order interactions. Recall from the two-level design discussion that JMP Custom Design defaults to a main effects only model, while Design Expert optimal design defaults to a main effect plus two-factor interaction model. A main effects only model is only appropriate if there is strong evidence interactions are not important in a system. In nearly all initial design situations, the model appropriate for the test is either one with main effect plus interactions, or perhaps a model of a full second order polynomial.

Recall too that after a design is built, remember to set the model degrees of freedom equal to anticipated number of significant effects – usually equal to the number of main effects. It is always a good idea to include two-factor interactions of interest iteratively, making sure sufficient error degrees of freedom are included. One possible approach would be to first perform the power analysis using only the main effects in the model, and record those power estimates. Then substitute main effects for two-factor interactions of interest. Repeat this process until all the two-factor interactions of interest have power estimates.

1. Options for Conducting the Power Analysis

This power analysis discussion focuses on the iterative process of designing a test with sufficient power to meet test objectives. Therefore, the guide assumes that the test planning process is complete and all that is left is to construct a test run matrix. More specifically, the guide assumes that:

- Continuous responses have been identified as the primary responses of interest (or an approximate SNR has been determined for binary responses).
- The number of factors and levels is set, and constraints or restrictions on randomization incorporated
- The design has been drafted
- Due diligence has resulted in reliable estimates for δ and σ.

This guide focuses on the iterative process of generating the test design size to achieve adequate power. The iterative process involves determining power values for a given design, then modify the design until the power is acceptable. Design modification can be in terms of the point distribution or locations, or by increasing the size of the test. Higher power values are desirable, more specifically power values greater than 90 percent should be the goal. Rarely do tests planned grow in size during execution and often the test size shrinks. Accepting power values lower than 90 during test planning often results in lower power probabilities when missing data are taken into account. Figure 3-3 shows a notional power curve for a one-sample hypothesis test. All power curves generally have this shape. As you can see from the curve, after a certain level (90 to 95 percent) the power gains are not worthwhile for the increased sample size. Additionally, below 80 percent power drops off dramatically. Testers should seek power values just above the knee in the curve. This avoids over testing while protecting against the possibility of data loss in testing.

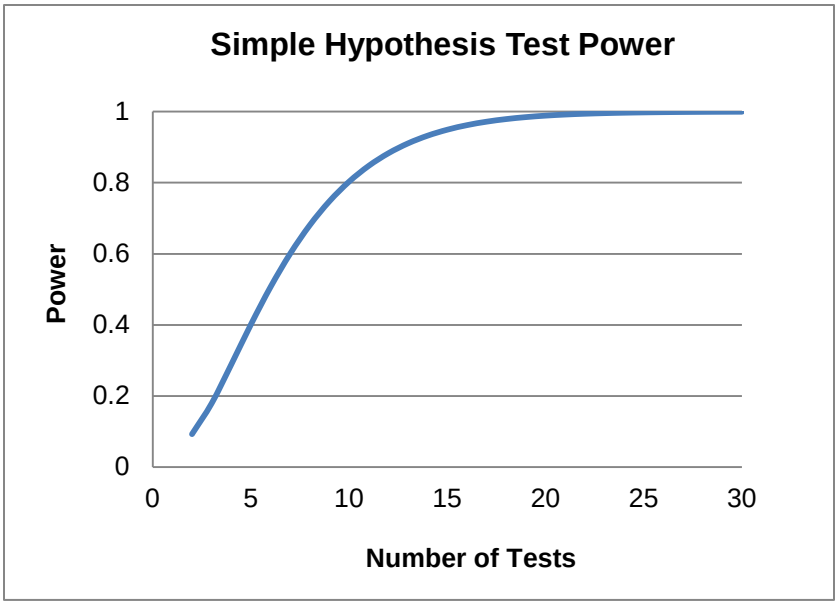


Figure 3-3. Power Analysis Curve for a One-Sample Hypothesis Test

If the final test size is too large for cost, schedule or resource limitation reasons, there are several options to consider. Test design adaptations can include changing more than just the size of the design. Perhaps the low power values for certain factor effects

are due to having too many factor levels, or perhaps too few factor level changes (for hard-to-change factors). Other considerations can involve selecting a different design type, or including fewer terms (if justified) in the anticipated model. Iterate until you have achieved the best possible design power given your limitations. In the end, it is recommended always to identify ways to reduce system noise (σ) prior to testing, which can greatly aid in improving power and in ultimately ensuring test success.

C. Power Analysis using Design Expert

This section forms a user's guide for calculating power in Design Expert. The instructions specified herein are intended for analysts with limited experience with Design Expert. After reading this guide, the analyst will be able to perform rudimentary power calculations on basic experiments that are typically encountered in operational testing.

1. Generating a Designed Experiment in Design Expert

Whether power is needed for a design created from scratch or for an existing design, the first step in calculating power in Design Expert is to generate the design. Begin by selecting "new design" under the file menu. The two-level factorial design page will appear. The yellows tabs on the left side of the screen contain four types of designs. The design types are distinguished by the regression models they support and their placement of points within the design space.

Common designs encountered in operational testing include general factorial designs, two-level factorial designs, and optimal designs. A general factorial design is a full factorial that uses categorical factors, which can have two or more levels. A two-level factorial design is a full or fractionated factorial that uses continuous factors. General factorial and two-level factorial designs create sample sizes that are multiples of powers of two. Optimal designs, on the other hand, support sample sizes with an arbitrary number of runs. To create an optimal design, the user specifies the factors, sample size, and model. Then, an optimization algorithm finds the factors settings that provide the most favorable statistical properties for the design. Optimal designs support continuous and categorical factors.

a. Generating a New Design

To generate a new design, continue by selecting a general factorial experiment from the yellow factorial tab. Design Expert prompts the user to input the number of factors in the design and the number of levels for each factor. In most cases, the number of factors in a design is inconsequential for power (except when the design is saturated or unbalanced), but power does decrease as the number of levels within a factor increases.

For this reason, consideration should be given to minimizing the number of levels within a factor, when possible.

An example of a case in operational testing where the levels could be minimized involves the "light" factor (also known as the "time of day" factor). The analyst is interested in characterizing the performance of the system under test during different segments of the day and decides to partition the categorical factor into day, dawn, dusk, and night. A design with a four-level categorical factor requires approximately 34 runs to have 90 percent power (with $\alpha = 0.05$ at an SNR of two). In contrast, if a three-level categorical factor were used, 22 runs would achieve 90 percent power, while if a two-level factor were used, only 13 runs would provide 90 percent power. By collapsing the four levels into two (day, night), the test size could be 61 percent shorter. This tester should carefully weigh the importance of resolving four levels of a factor versus the consequence on test length.

After inputting a four-level, a three-level, and a two-level factor, press continue. A nearly completely gray screen will appear that provides options to adjust the number of replicates or blocking scheme. The number of replicates (here replicates are runs at each test condition of the full factorial, so incrementing replicates from 1 to 2 doubles the design size) will affect power by increasing the sample size, while the blocking scheme will have little effect on power. Press continue again to generate the experiment and arrive at the *MyDesign* page (Figure 3-4).

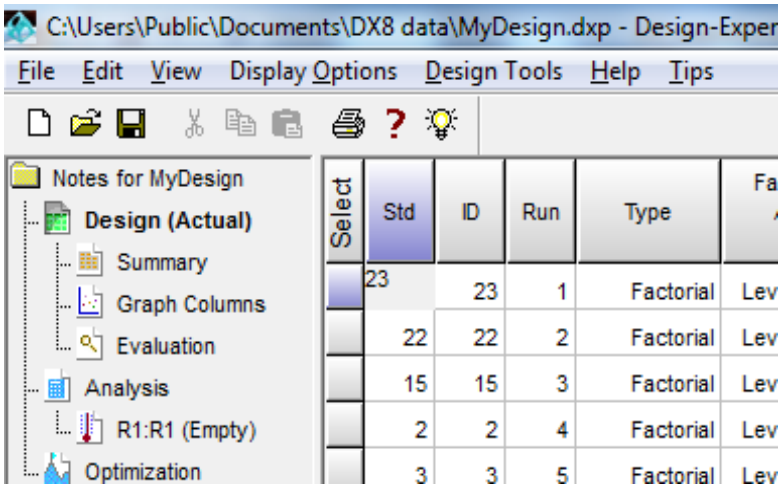


Figure 3-4. The MyDesign page in Design Expert

b. Importing an Existing Design

Importing an existing design is just as easy as creating a design from scratch. To import a design from Excel that has all categorical factors, first, build a general factorial experiment with the corresponding quantity of factors and levels. When inputting the details of each factor, it is critical the name of each level matches the name from the

excel spreadsheet. After properly inputting the number of factors and levels, and the names for each level, click continue in the bottom right to generate the design and to proceed to the *MyDesign* page. The newly generated design looks similar to the existing design from excel, but the settings of each factor and the sample sizes are different. To fix this, adjust the design within Design Expert so that it has same number of runs as the existing design. This is done by right clicking a row in the design and selecting *Insert Row* or *Delete Rows(s)* (Figure 3-5). Once the sample sizes match, copy and paste the design from Excel into Design Expert. If all of the cells in the design are populated then that is indication that the import was successful. If any cells are blank, the import failed and is probably caused by a spelling mistake between the names of the levels of the existing and generated design.

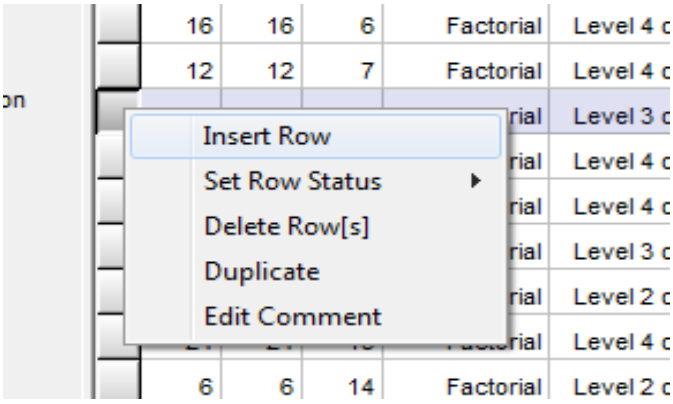


Figure 3-5. Inserting a row into a design

The process for importing a design with continuous variables is quite similar. Choose to build a "Historical Data" response surface design. Select the number of continuous factors in the design and set the low and high level of each factor. Enter the number of runs (rows) in the design and hit continue. Then, copy and paste the data from the existing design. The "Historical Data" approach can be used to build designs with continuous factors, both categorical and continuous factors, but not categorical factors only.

Importing a design from an existing file is also possible via a drop down menu from the opening page. The option, *File > Import from File*, allows users to import a design from a text file. In our experience, Design Expert can be a bit awkward when it comes to reading-in and interpreting the text file. The time it takes to relearn the proper formatting for the text file is probably not worth the hassle. For this reason, we suggest the copy and paste approach.

2. Calculating Power Once a Design is Generated

Returning to the *MyDesign* page, notice the hierarchy-tree under *Notes for MyDesign*. In the upper left corner of the screen, there are three primary branches: *Design*

(Actual), Analysis, and *Optimization*. The *Design (Actual)* branch allows the analyst to evaluate the goodness of the design by providing calculations for power, and other metrics of design quality. The *Analysis*, and *Optimization* branches do not involve power and are not covered in this guide. Under the *Design (Actual)* branch, click on *evaluation* to proceed to calculate power.

After clicking on *evaluation*, the *Model* page (under the "Model" tab) appears. Power of a designed experiment is influenced by α, signal/noise ratio, *N*, model form, and the placement of points within the design (design structure). The sample size and design structure are predetermined at this point, but the Model page provides the option to change the signal/noise ratio and the model form.

An important option on the *Model* page is the *Order*. *Order* is the model form the user intends to fit to the data, once that data are collected. Two common model forms used in operational testing are a main effects model and a two-factor interaction model (2FI). In the *Model* page, Design Expert has chosen the two-factor interaction model by default, which is recommended if the number of runs are affordable. If the user chooses a number of runs less than what is required for a two-factor interaction model, take care to assess the degree of aliasing, potential factor effect correlations, and especially the degree of model saturation, resulting in dangerously few error degrees of freedom. Recommended design alternatives include 2^{k-p} fractional factorials and optimal designs.

Another important option on the *Model* page can be found by selecting the *Options* button. The options menu provides the ability to change the signal/noise ratio (Figure 3-6). By default, Design Expert calculates power for signal/noise ratios of 0.5,1, and 2. As the name suggests, the signal/noise ratio is a ratio of the signal δ, to the noise. Design Expert uses a default α value of 0.05, which is recommended unless a higher level of risk can be fully justified. In those extenuating circumstances, α can be modified under the *Edit* drop down at the top, choose *Preferences*… and then click on the *Math* tab. Within the *Math* page there is an option to change the significance threshold for power.

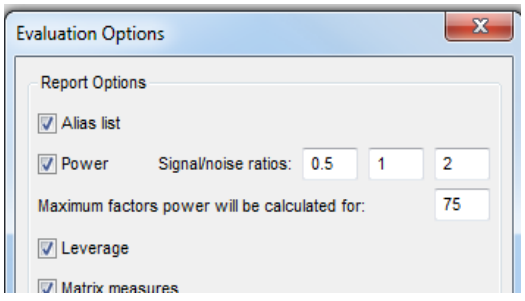


Figure 3-6. Signal-to-noise options

Now that the important options have been configured, click on the *Results* tab to proceed to calculate power (Figure 3-7). On the *Results* page, power is shown for each term in the model for three different signal/noise ratios. Different variations in the

presentation of the power table are possible depending on the type of factors in the design (continuous, categorical) or the model form.

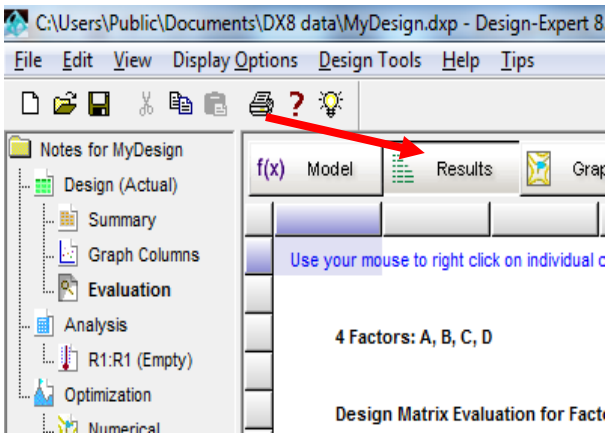


Figure 3-7. Results tab

A power table for a design with a main effects model that has two-level categorical factors or continuous factors is easy to interpret; one power value is presented for each term in the model. Figure 3-8 shows the power output for a design that includes two-level categorical factors with a main effects model. Figure 3-9 shows identical output for a design that has continuous factors with a main effects model. Power for two-level categorical factors and continuous factors are simpler to interpret than multi-level categorical factors because the signal, in the SNR, is easier to define.

				Power at 20 % alpha level to detect sig		
Term	StdErr**	VIF	Ri-Squared	0.5 Std. Dev.	1 Std. Dev.	2 Std. Dev.
A	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %
B	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %
C	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %

lasis Std. Dev. = 1.0

Figure 3-8. Power output for a design with two-level categorical factors.

				Power at 20 % alpha level to detect sig		
Term	StdErr**	VIF	Ri-Squared	0.5 Std. Dev.	1 Std. Dev.	2 Std. Dev.
A	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %
B	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %
C	0.35	1.00	0.0000	28.6 %	49.8 %	89.0 %

lasis Std. Dev. = 1.0

r Categorical Terms, The minimum Power for each group of terms is reported.

Figure 3-9. Power output for a design with continuous factors

The signal becomes more difficult to define as the number of levels within a factor increases. For a two-level categorical factor, the signal is simply the change in the response as the factor changes from one level to the other. For a multi-level categorical factor, there are numerous possible contributors to "signal." The signal could be defined as the change in the response as one level changes to one of the other levels, or possibly the change in the average response between a group of levels compared to a group of other levels. As a greater number of levels for a factor are introduced, the definition of the signal can become more complicated because there are several ways to configure the factor level coefficients to generate an effect.

Design Expert has formulated a definition of the signal for multi-level categorical factors. Details on this formulation can be found in Appendix D. To avoid reporting power for the numerous possible signals for a single categorical factor, Design Expert takes a conservative approach, stating, "For Categorical Terms, The minimum power for each group of terms is reported." The minimum power is obtained by iteratively searching for and finding the pair of factor levels which yield the lowest power for that factor in the given design.

D. Power Analysis using JMP 11

1. Introduction

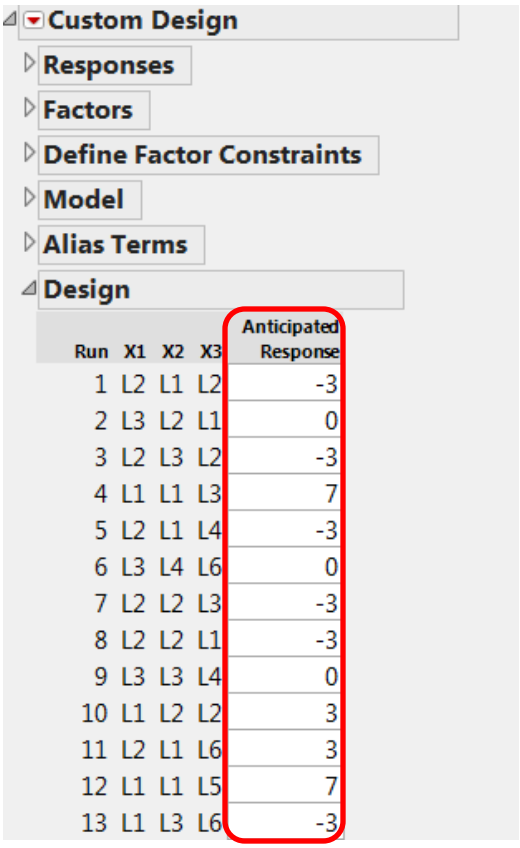
JMP 11 provides the user with an increased flexibility to specify the anticipated signal (on the response), compared to previous versions of JMP that simply requested a single value for the SNR. All of the following discussion assumes the use of JMP's Custom Design. The added flexibility in JMP 11 can be daunting, even for those familiar with JMP, for a couple of reasons. The first challenge is that the user is faced with inputting values for anticipated responses or anticipated coefficients as an alternative to the single-value SNR. Default values are set in place and provide a starting point, but as we will see, the default coefficients should nearly always be modified to a more conservative configuration. If anticipated responses or anticipated coefficients are reasonably known (past test data of essentially the same system), then the user can modify the defaults. For those who are unfamiliar with anticipated responses or anticipated coefficients, the following sections (and appendices) provide an explanation on how they are involved in the power calculations, as well as guidance for their use. The second challenge is that the *Signal to Noise Ratio* cell in JMP 10 is absent on the JMP 11 *Design Evaluation, Power Analysis* page. Recall that the user could use the option to *Set Delta for Power* in *Advanced Options* under the red triangle, but as we will see that is not necessarily the best choice if the desire is to match multi-level categorical factor power with other software platforms. Each of the three methods (anticipated responses, anticipated coefficients, and setting delta for power) will be addressed.

2. Specifying Anticipated Responses

The first option available in the JMP 11 power interface is to work with anticipated responses for each of the design runs. Although this alternative is arguably the best source of information to directly compute power, it is rarely practical. Most system experts, even with a tremendous amount of historical data, would be at best, wildly guessing on the vast majority of response values for the typically encountered multiple factor designs. Obviously, poor estimates for the responses would lead to misleading power estimates. As is pointed out by the developers of JMP, this response-value interface does have pedagogical utility, so that students can see real-time the connections between altering response values and changes to the corresponding factor power estimates. There is also a benefit in seeing hypothetical response values for a pre-determined set of anticipated coefficients, as will be discussed in the next section.

Other than obtaining feedback in planning by performing some sort of sensitivity analysis on potential misspecification in the coefficient estimates (or just the SNR), using the response values as the primary determinant of effect magnitudes in a power analysis would most likely lead to off-the-mark estimates of factor SNRs and hence power estimates.

When you create your initial design, and click *Make Design*, a dialog box appears that first lists the experimental runs along with anticipated responses for each run. The anticipated response values are all filled in, which can be a little confusing. You have the option of keeping the defaults or supplying your own values (Figure 3-10). The response values shown essentially originate from the default anticipated coefficients (Figure 3-11). Initially bypassing the default anticipated responses makes sense, to study the values of the anticipated coefficients in the next dialogue, to better understand the relationship between the anticipated coefficients and anticipated responses. This exercise will enable you to make a more informed decision as to whether to change any of the anticipated responses.



Run	X1	X2	X3	Anticipated Response
1	L2	L1	L2	-3
2	L3	L2	L1	0
3	L2	L3	L2	-3
4	L1	L1	L3	7
5	L2	L1	L4	-3
6	L3	L4	L6	0
7	L2	L2	L3	-3
8	L2	L2	L1	-3
9	L3	L3	L4	0
10	L1	L2	L2	3
11	L2	L1	L6	3
12	L1	L1	L5	7
13	L1	L3	L6	-3

Figure 3-10. Power Analysis Anticipated Response Dialog

One other aspect of the default anticipated response values can make it difficult for the user to associate with those values. Most of the values are near zero and many are negative, which is out of the norm for most analysts accustomed to strictly positive responses and can make it harder to recognize effect sizes.

3. Specifying Anticipated Coefficients

The next set of values displayed in the power analysis section in JMP 11 is the anticipated coefficients. These estimates are another method of obtaining information from the users in order to discern the anticipated effects due to changes in the factor settings. The option to modify anticipated coefficients serves as a direct input to the software in providing the information necessary for power. We expect this method to be easier than the anticipated responses for the user to supply the necessary information. In general, the software provides the option to specify the magnitude of the change in the response as each of the factor settings change during test (Figure 3-11). For a two-level factor, the levels are coded -1 and +1. The difference between these two levels is two-units, a coefficient of one means that there is a two-unit change from level 1 to level 2 of the factor.

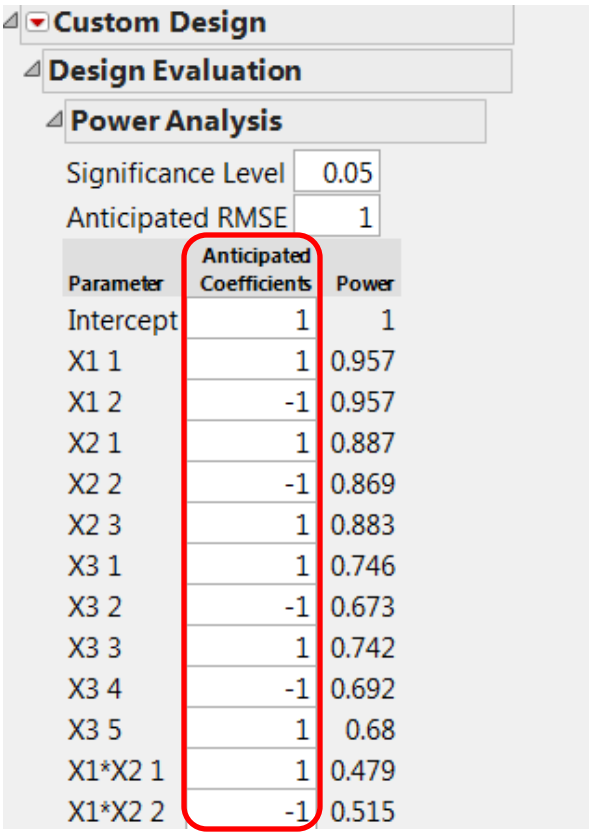


Figure 3-11. Power Analysis Anticipated Coefficients Input Dialog

The anticipated coefficients for factors with more than two levels require more deliberation. Recall (Section 4.A.2) that there are the number of levels (q) minus one or $q - 1$ coefficients for each factor, so for a three-level factors, two coefficients must be estimated. Implied in this parameterization (which is standard in most statistical software), is that there is a coefficient associated with the final level of a factor as well. It turns out that the coefficient associated with the last level (if used directly in the model would cause a problem mathematically) can simply be calculated by knowing the $q - 1$ other coefficients. This coefficient associated with the last level is just the negative sum of the $q - 1$ coefficients, such that all q coefficients sum to zero. The coefficients are derived from the response values, so they can be used to tune each factor (and two-factor interaction) for power according to anticipated influences on the response.

As a default, JMP 11 sets all anticipated coefficients associated with a factor to an absolute value of 1. For factors with more than two levels, the anticipated coefficients take on a value of +1 for the first coefficient, then alternate in sign to return -1 for the second anticipated coefficient, then +1 for the third coefficient, then -1 for the fourth, and so on. The coefficients across all q levels, including the last level (without a parameter), sum to zero. So for factors with an odd number of levels, the value of the coefficient for the last level is zero, implying that level is not actively contributing to the response effect for a power calculation. For factors with an even q, that last coefficient (Level q) is -1,

again such that $\sum c_i = 0$. One way of understanding JMP's default coefficient values is as follows: With an odd number of factor levels, all but one level differ from each other. For factors with an even number of levels, all factor level means differ from one another. This choice of values for anticipated coefficients will affect the factor effect power for a given factor in such a way that it matters whether the factor has even or odd number of levels. Consider the one-factor case with increasing q and fixed N=16 back in Figure 3-2. Note the saw-tooth trend in factor effect power reflecting the inconsistent percentage of active effects with even q vs. odd q caused by the default anticipated coefficients.

Example: Two-factor 3 x 4 replicated full factorial

Consider an example design involving two factors, one with three levels and the other with four levels. For the purposes of this example and in order to demonstrate increasing complexities in power estimation, we will look first at a balanced replicated full factorial design, then consider an unbalanced fractional factorial design.

For the replicated full factorial, this example has two replicates of a 3 x 4 design, resulting in 24 runs (Figure 3-12). Because this design is a balanced replicated full factorial, the software packages should all generate the same design. The design is balanced and orthogonal, and regardless of the underlying anticipated general model, there are 12 degrees of freedom for pure error from replication to be used in the power calculation. It is standard practice to consider an anticipated model containing only the main effects, a reasonable approximation for the model degrees of freedom needed for the final predictive model.

Design

Run	X1	X2	Anticipated Response
1	L2	L1	1
2	L3	L2	0
3	L1	L3	3
4	L3	L4	0
5	L1	L1	3
6	L1	L2	1
7	L1	L3	3
8	L2	L4	-1
9	L1	L1	3
10	L3	L2	0
11	L2	L3	1
12	L2	L4	-1
13	L2	L1	1
14	L1	L2	1
15	L2	L3	1
16	L1	L4	1
17	L3	L1	2
18	L2	L2	-1
19	L3	L3	2
20	L3	L4	0
21	L3	L1	2
22	L2	L2	-1
23	L3	L3	2
24	L1	L4	1

Apply Changes to Anticipated Responses

Figure 3-12. Design Runs for a 3 x 4 two-factor factorial with two replicates

Factor effect power is typically calculated for each factor or interaction and estimates the probability that a factor effect is declared significant when the factor effect truly is significant. In this example the 24-run design supports a main effects model, containing two effects: the three-level factor, and four-level factor. The coefficients in the user interface are the coefficient values assumed under the alternative hypothesis, while the coefficients under the null are equal to zero (not shown in the interface).

Factor effect power depends on the area under a reference F distribution associated with the alternative hypothesis. The F-distribution is the appropriate reference distribution for a statistical test to determine effect significance as a ratio of mean squares, which are variances. Factor effect power is calculated for the three-level factor by first computing a critical F value, which is the F value at the 100(1-α) percent cumulative percentile of the F distribution under the null hypothesis. The F distribution, shown in Figure 3-13, under the null hypothesis is a function of the numerator and denominator degrees of freedom. The numerator degrees of freedom for a factor's effect power is equal to q-1 (q-1=2 for three-level factor). The denominator degrees of freedom is equal to the number of runs in the design (24) minus the number of parameters (p) in the model (p=6, including the intercept). Then, a non-centrality parameter is calculated

using the coefficients under the alternative hypothesis. Power (1-β) is determined from a non-central F distribution using the non-centrality parameter, numerator degrees of freedom, denominator degrees of freedom, and critical F value. For complete details on JMP 11's factor effect power calculation, see Appendix C.

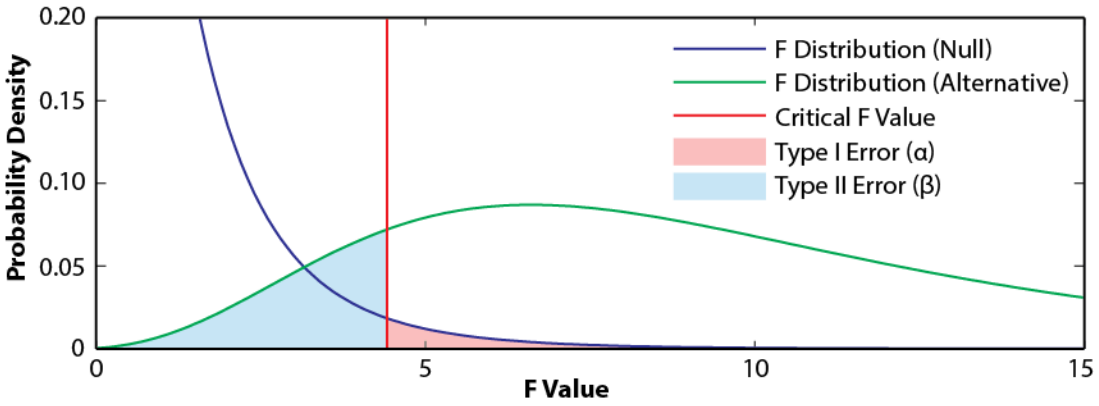


Figure 3-13. Distribution of F statistic Under the Null and Alternative Hypotheses

Coefficient power is typically calculated for each parameter in the model. The three-level and four-level factors are made up of two and three parameters, respectively, so for this example power would be calculated for five parameters. The coefficient power calculation is quite similar to the factor effect power calculation. The numerator degrees of freedom for coefficient power are always equal to one. Similar to factor effect power, the denominator degrees of freedom is equal to the number of runs in the design minus the number of parameters in the model. Power is determined from a non-central F distribution using the non-centrality parameter, numerator degrees of freedom, denominator degrees of freedom, and critical F value. For complete details on JMP 11's coefficient power calculation, see Appendix C.

4. Specifying Power using Advanced Options

The *Advanced Option* approach for entering a delta value for power described in Chapter 3 is also available for designs with multi-level categorical factors. However, because of the way the SNR is distributed to the anticipated coefficients, this approach should only be used without modification if the user desires that all the factor indicator variables be active (nonzero and different anticipated coefficients). The direct applications of the SNR approach will lead to overly optimistic power estimates in JMP 11 for multi-level categorical factors. Multi-level categorical factor power values in JMP 11 using the default anticipated coefficients, even if manipulated by this advanced option, essentially give effect sizes larger than that specified by the SNR, because all the factor parameter coefficients are set to SNR/2. As a result the corresponding effect size grows as the number of factor levels increases, inflating the power estimates. Further explanation is provided starting in Section 6 below.

5. Default Anticipated Coefficient Power

Consider two-factor replicated factorial example shown in Figure 3-12. Once the two-replicate full factorial design has been specified in JMP 11 (factors, levels, model, and *N*), choose *Make Design*. The design will be constructed, displayed, and the *Power Analysis* interface will appear. Take note of the default coefficients for this example (Figure 3-14). An intercept coefficient is followed by the three-level parameters, X1 1 and X1 2. The alternating +1, -1 coefficients are displayed for both of the factors. It may be helpful to think of coefficients for the q^{th} level of each factor. For the three-level odd factor X1, the 3rd level (X1 3, not permitted in the model because it would result in too many parameters) is 0, while the four-level even factor X2 forth level (X2 4) is -1. Consideration of these last levels is useful when contemplating the calculation of the factor effect power values.

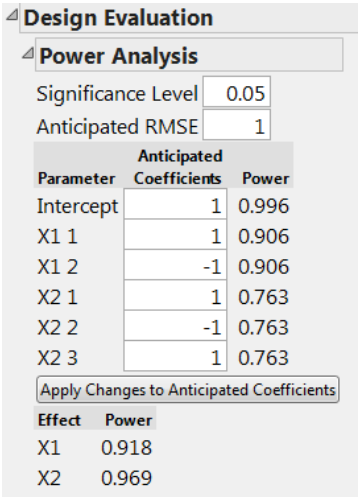


Figure 3-14. Default Coefficients and Power for 3 x4 24-run Design

Take a look at the reported power values using the default coefficients. First note that the coefficient power estimates differ from the factor effect power, for the same factors. Secondly, the coefficient power values within a factor are all the same, which is due to the balanced, orthogonal aspects of the design. The coefficient power values reflect only the contribution of one indicator variable to the response, while the factor effect power reflects the cumulative contribution of all a factor's indicator variables. Naturally it makes sense that the factor effect power would be larger than the coefficient power, especially since factor effect power assumes each of the *q*-1 parameters contributes to changing the response average. In nearly all instances, placing nonzero values in each of the parameter cells for a factor will result in factor effect powers greater than the coefficient power values, and the relative difference between coefficient power and factor effect power increases as *q* increases.

Next note the power for the four-level factor (X2) versus the three level (X1) power. The four-level factor has greater factor effect power than the three-level factor, despite

the fact that each parameter has lower power. The default coefficients in JMP 11 result in the even *q* factor with four levels having four levels active (nonzero), as opposed to the odd *q*, three-level factor having only two active levels. The other point is that the factor effect power for the three-level factor is the same as the factor effect power from JMP 10 and DX (what we call most conservative power), because the default coefficients have only two active levels which is the same approach taken in the most conservative approaches to estimating power. The next section describes this most conservative approach to computing factor effect power obtained by zeroing out some of the JMP 11 anticipated coefficients.

6. Configuring the JMP 11 Coefficients for Most Conservative Power

If the user is interested in modifying the anticipated coefficients and desires to configure a multi-level factor for a conservative estimate of power, there is a general approach which will match the values obtained in JMP 10.0.0 (whereas the JMP 10.0.2 approach is presented in Section 4.D.8) and Design Expert. This approach to be used with categorical factors with more than two levels mimics the approach used with two-level factors by assuming that the mean response of one level differs from the another level, while all the other levels do not contribute to changing the response mean. To implement such a condition, just set the values of the anticipated coefficients of all but one or two factor's anticipated coefficients to zero. In our two-factor replicated design example, X1 is a three-level and X2 a four-level factor, and a 24-run replicated full factorial design is built. In this case all the parameters (factor levels) have the same number of observations, and the coefficient power probabilities within a factor using the default anticipated coefficients (Figure 3-14) are all identical.

A coefficient solution for most conservative power is simply to set one of the factor anticipated coefficients to one, while changing all the remaining coefficients (for that factor) to zero, which is displayed in Figure 3-15. Note that setting one coefficient to one results in the invisible last parameter level Xi[*q*] = -1, so this assignment of coefficients assumes δ/σ = SNR = 2. In general, set the nonzero coefficient(s) equal to SNR/2. Be sure to *Apply Changes to the Anticipated Coefficients*. After setting the nonzero coefficient(s) to the desired nonzero SNR/2, *the coefficient power should be ignored*; the factor effect power value should reflect that minimum power estimate for that multilevel factor. In Figure 3-15, note that the conservative power is 0.918 for X1, and 0.744 for X2, which better agrees with the notion that increasing *q* decreases power.

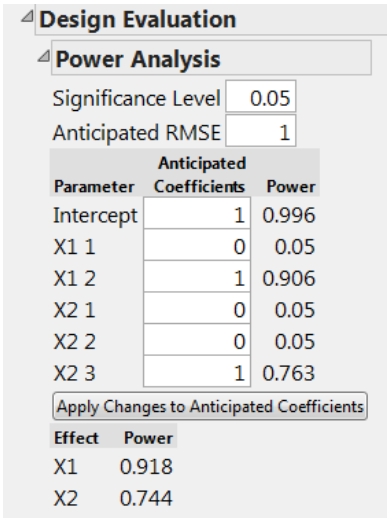


Figure 3-15. Coefficients for Most Conservative Power for 3 x4 24-run Design

As you can see in Figure 3-16, this most conservative power estimate for the 3 x 4, two-replicate example, agrees with Design Expert and JMP 10.

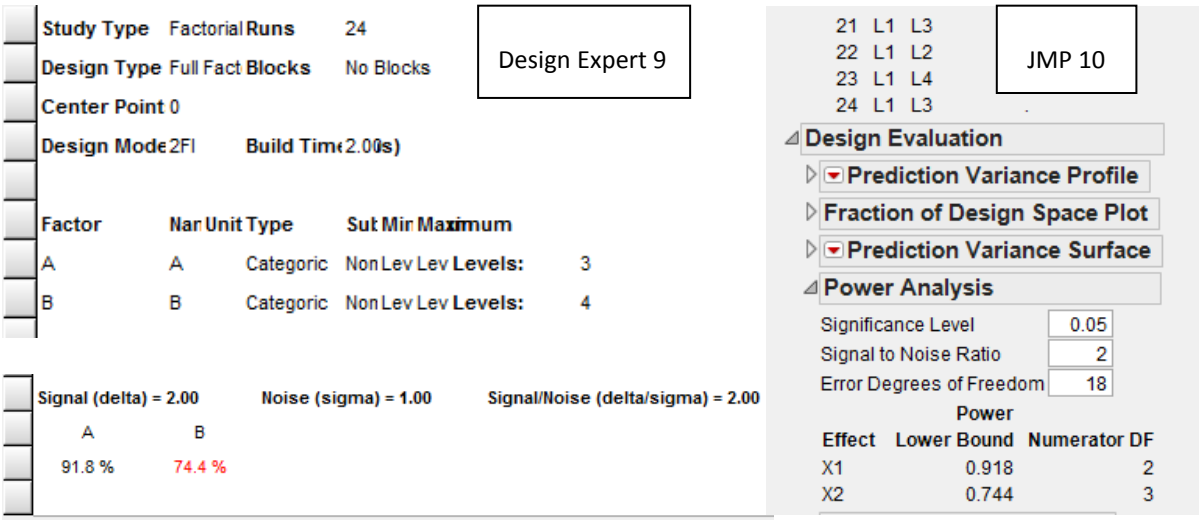


Figure 3-16. Design Expert 9 and JMP 10 output showing power calculation for 3 x 4 design with two replicates and 24 runs. Power matches JMP 11 conservative power.

For the case of unbalanced, non-orthogonal fractional designs, it is recommended that the user replace the value given the coefficient with the smallest coefficient power with the desired δ/σ. In some cases there is a tie for minimum default coefficient power, so in that case form a + δ/σ, - δ/σ contrast with the two smallest power parameters. The goal is to locate the two parameters which would result in the lowest (most conservative) power for that factor. The parameters with the lowest coefficient power under the default combination of anticipated coefficients are most likely those that will result in the factor having the lowest power when only a pair of coefficients are nonzero. The parameters

displaying the lower power with the default coefficients typically have the least observations per level for that factor. An example of this relationship is given in the next section.

Example: Most Conservative Power for 3 x 4 x 5 Fractional Factorial Design

Consider an example design involving three factors, one with three levels, one with four levels and the last with five levels, and 38 runs. This design could potentially estimate all the main effects plus two-factor interactions assuming the points are appropriately placed, as the main effect plus interaction model requires 36 degrees of freedom. Two additional runs are added for lack of fit, such that $N = 38$. This 38-run design will be unbalanced for all factors, as N is not a multiple of three, four or five. In JMP or DX, the design is fairly easy to build, one just needs to specify the general model as including all the main effects and two-factor interactions, and specify the number of runs. Figure 3-17 shows the JMP 11 steps before and after building the design along with the screen captures.

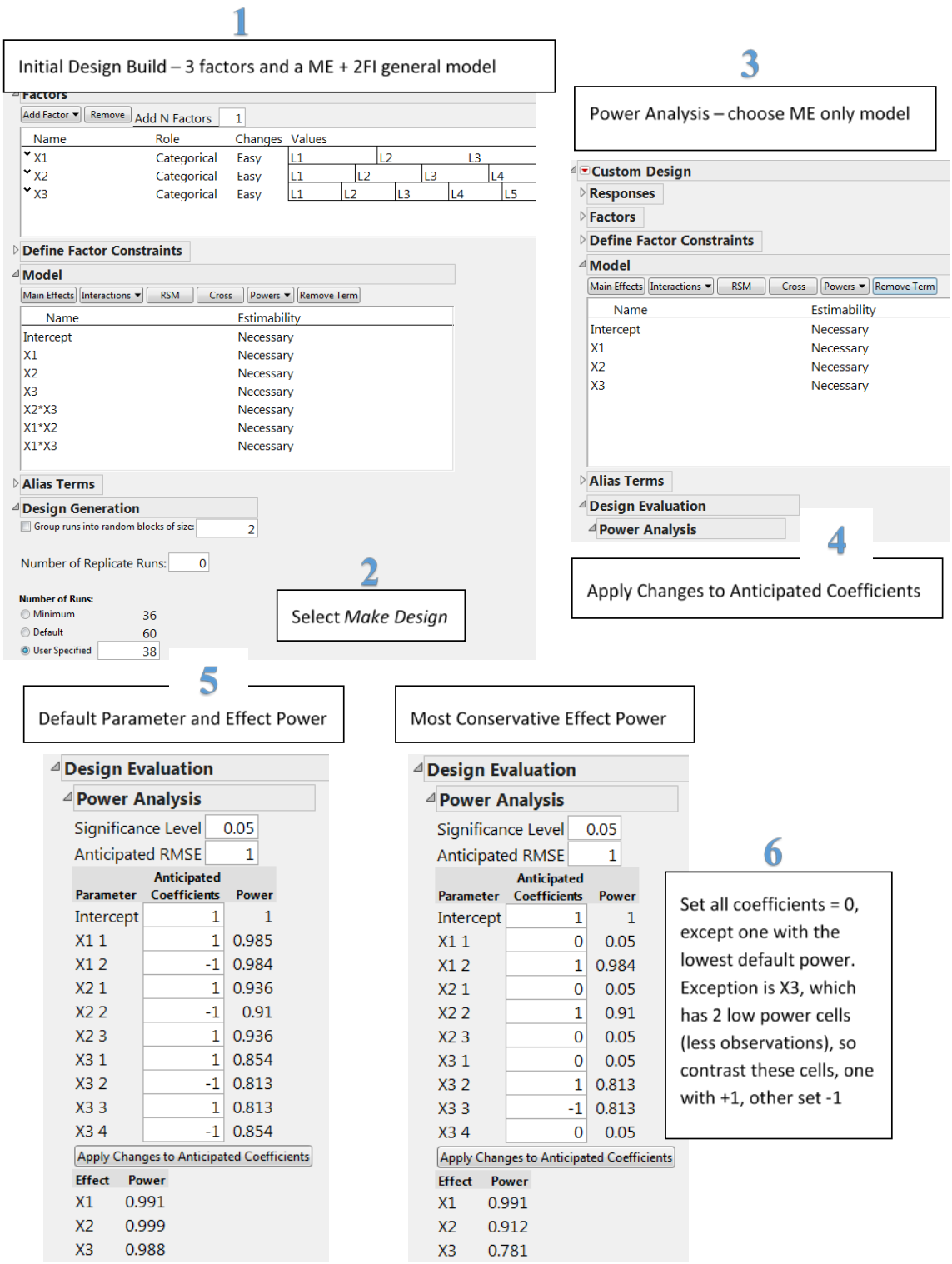


Figure 3-17. Power analysis example using the most conservative power approach in JMP 11. Most conservative power shown as Factor Effect Power bottom right.

Note that there is a purposeful change to the model form after building the design and before performing the power analysis. The model used in building the design includes the full polynomial order determined in planning (e.g., main effects plus two-factor interactions), while the model selected for power analysis contains a reduced model, having approximately the number of terms anticipated to be significant. The importance of this step is emphasized in the upcoming practice tips, and is due to not having sufficient degrees of freedom for error (only the two from lack of fit), so power will be drastically underestimated. If the model is not changed to only main effects prior to power analysis, the power values for this design are all less than 40 percent, including the default effects power estimates!

The default anticipated coefficient power values vary over a range within each factor, due to the unbalanced nature of the design. The first step in obtaining the most conservative power estimates is to make note of the parameters with the lowest power for each factor, to later set those anticipated coefficients nonzero, while setting all other coefficients for that factor equal to zero. For this example, see that the coefficient power values are all greater than 80 percent with power values decreasing as q per factor increases. Effect Power is also reported, and those values are all impressive, with the even numbered four-level factor having the highest power.

The next step for most conservative power is to modify the anticipated coefficients such that each factor has mostly zero entries, and the nonzero cell(s) are set to a value of SNR/2. Choose the nonzero entry to be the parameter having the lowest coefficient power for that factor. In this example, we select X1 2, X2 2, and for X3 we select a contrast based on the two parameters X3 2, and X3 3. The reason for choosing the contrast between Level 2 and Level 3 for X3, is that to obtain the most conservative power estimate requires selection of the two levels with the fewest observations. Because those two levels have the same default coefficient power of 0.813, it is likely those two levels have the fewest observations. So instead of contrasting a chosen level with the last level, as is done for X1 and X2, we make the contrast in X3 between two displayed levels. In general, the most conservative power should be based on a comparison of the two levels with the fewest observations. Some trial and error might be used to find the combinations giving the lowest power. The general procedure for determining the most conservative power in JMP 11, or to find the equivalent power in JMP 10 is provided in the process flow diagram (Figure 3-18).

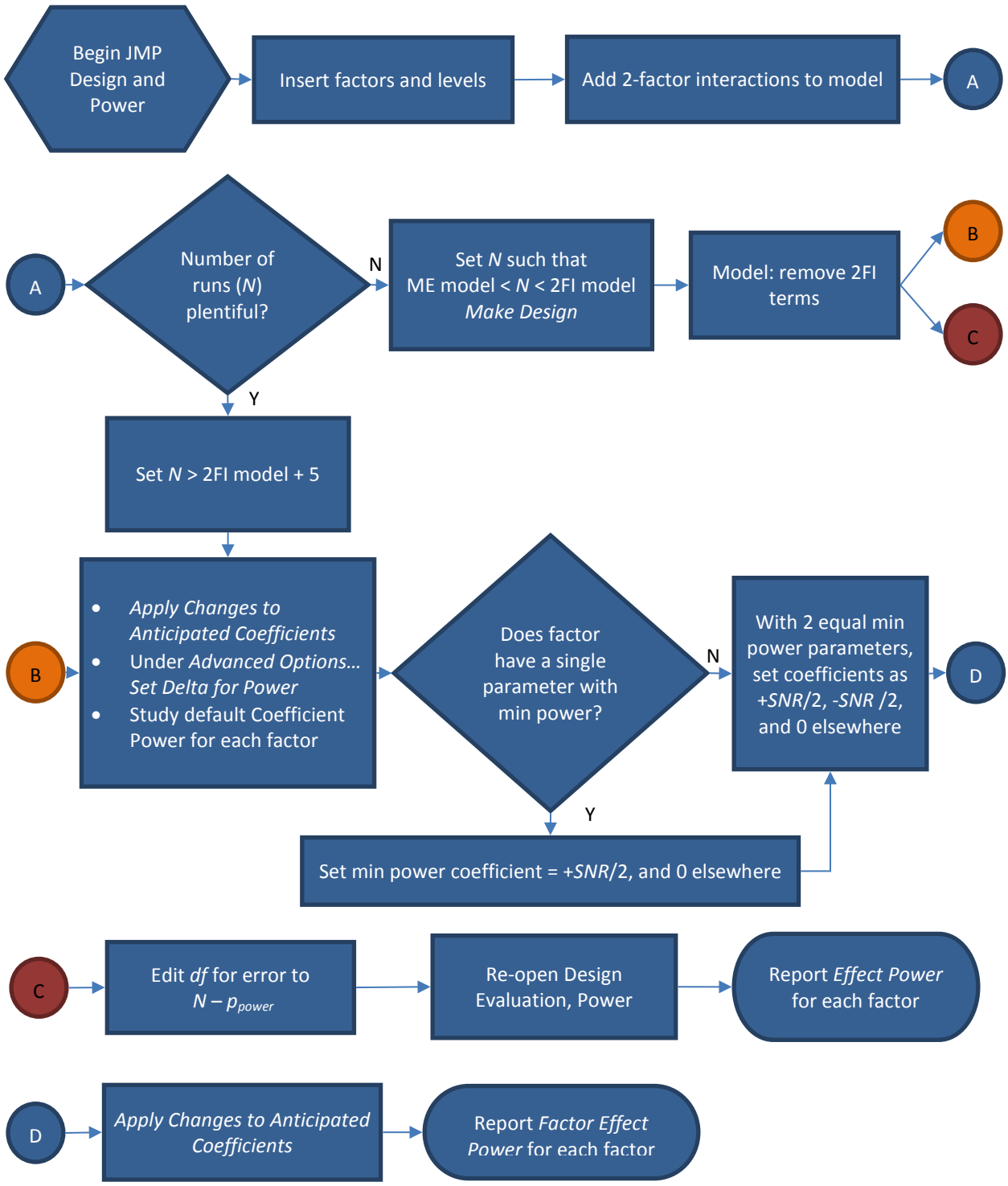


Figure 3-18. Process flow diagram for using JMP 11 or JMP 10 to estimate equivalent power

In this example, the most conservative power approach generates power estimates lower than both the default effect and default coefficient power, for the same design and SNR. The most conservative power also gives lower power values for factors with more

levels, which we will see later, agrees with the other software platforms. The next section provides a survey comparison of this JMP 11 most conservative power approach with standard reported power from JMP 10 and DX.

7. JMP 11 Power Reporting

a. Power for Coefficient Estimates

The most significant change to JMP 11 power reporting for designs involving multi-level categorical factors, is that it provides two assessments of power for such a factor type. The first variant of power estimate is referred to in JMP 11 as power for the parameter using anticipated coefficients. There are q-1 parameters, hence q-1 coefficient power values reported per factor, and depending on the design built, these estimates within a factor may not agree. For balanced, orthogonal designs the coefficient power estimates within a factor should agree for each factor parameter. If an optimal design is generated, and especially if the number of runs is not a multiple of q for a factor, there must be an imbalance in terms of the number of observations per factor level. Accordingly, some levels (and associated parameters) will receive correspondingly lower or higher power values. Figure 3-19 shows coefficient power for an unbalance design. JMP 11 reports those power values for each model parameter, essentially a level of detail further than factor effect power. You can assume the last level's (q^{th} level) power will be similar to the other coefficient power values and to be conservative, you could just choose the minimum coefficient power to report for that factor. Typically though, just a single power value for a multi-level factor (or interaction) is sufficient.

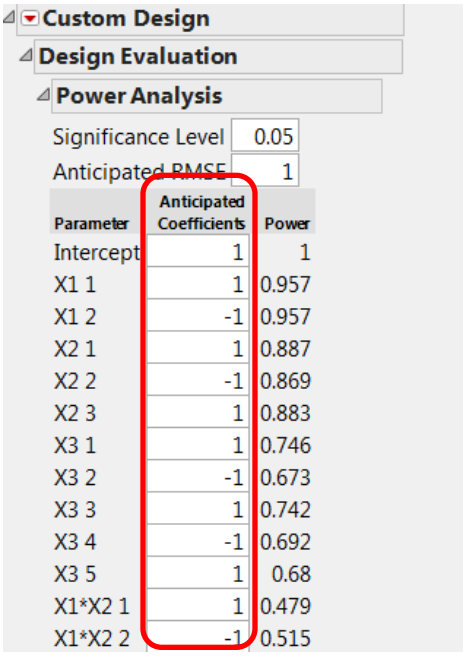


Figure 3-19. Power Analysis Results for Coefficient Power

The power calculation for the parameter estimate is computed by treating each factor parameter independently and computing a probability that the single parameter with anticipated coefficient magnitude shown can be detected given it is truly significant. An anticipated coefficient of magnitude one is equivalent to a SNR ratio of two for that parameter. The power calculation ultimately uses these anticipated coefficients to compute the non-centrality parameter for the alternate hypothesis, which is then used to find the power probability. More details regarding the calculation of the non-centrality parameter for the coefficient power is given in the appendix.

In many situations, the coefficient power values reported within a factor, with equivalent magnitude anticipated coefficients for a proposed design typically vary over a range of less than 10 percent, relative to each other. As we will see in a subsequent section, JMP 11 coefficient power estimates are more closely aligned with JMP 10 and DX factor effect power estimates than they are with JMP 11 factor effect power. The JMP 11 factor effect power, discuss following this section, tends to provide more favorable power estimates than most other software power estimates, assuming the exact same design and SNR ratio.

Coefficient power formulations are detailed in Appendix C, but an example calculation is provided here. Consider a one-factor design, where the factor has four levels and the design contains four replicates of each of the four levels, for a total of 16 runs. To determine the power of the second parameter of the four-level factor, say X1 2 in JMP, the non-centrality parameter λ_2 is calculated as

$$\lambda_2 = (\boldsymbol{Q}_3\boldsymbol{b})^T(\boldsymbol{Q}_3(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Q}_3^T)^{-1}\boldsymbol{Q}_3\boldsymbol{b} = 5.33 \text{ , where}$$

$$\boldsymbol{Q} = [0 \quad 0 \quad 1 \quad 0]\,,\ \boldsymbol{b} = [1 \quad 1 \quad -1 \quad 1]^{\boldsymbol{T}},$$

and $\boldsymbol{X}$ is the model matrix with 16 rows and four columns. The first column of $\boldsymbol{X}$ contains 1's and corresponds to the intercept, while columns two through four makeup the simple coding for the q-1 factor parameters. $\boldsymbol{Q}_j$ is a one-dimensional row vector of length p, the number of model parameters, and contains all zeroes except for the jth parameter, which is set equal to one. The index j begins with j=1 for the model intercept, so j=3 here corresponds the second factor parameter. The critical F value is calculated as $\acute{F}_2 = F^{-1}\{1-\alpha, 1, n-p\} = F^{-1}\{1-0.05, 1, 16-4\} = 4.75$. Power is then computed as $P_2 = 1-\tilde{F}\{\acute{F}, 1, n-p, \lambda_2\} = 1-\tilde{F}\{4.75, 1, 16-4,\ 5.33\} = 0.56$. For further details on these calculations, see Appendix C.

b. Power for Effects

In addition to the coefficient power, JMP 11 also reports a factor effect power probability for each factor and interaction effect. JMP 11 factor effect power provides the user flexibility to specify the contribution to the factor or interaction effect by specifying values for each of its parameters. This capability can be advantageous for those with knowledge or even reliable expert judgment about the potential causal effect of factor level subgroups on the response. Suppose, for example, that there is a five-level factor to be varied in an upcoming test. We will assume a one-factor design with three replicates, giving 15 runs. All the planning has been completed including the other factors, the design, the anticipated model, as well as reliable estimates of δ and σ. For the five-level factor, it is presumed that levels 1, 2, and 4 will most likely act in a similar fashion, and will all positively affect the response by about the same amount. Levels 3 and 5 are expected to move the response in the opposite or negative direction, again by about the same amount each. For the purposes of this example, and to keep the math fairly simple, assume that the estimated $\delta/\sigma = 2.5$. This scenario with all the levels active (some positive, some negative) actually corresponds to a factor effect much larger than the $\delta/\sigma = 2.5$ specified, reporting power values often much larger than the conservative power value. Table 3-4 shows the coefficient estimates that would give the desired effect contribution from all the factor levels.

Table 3-4. Example Factor Coefficients for a 5-level Factor

Level	JMP Parameter	Anticipated Coefficient
1	X1 1	1
2	X1 2	1
3	X1 3	-1.5
4	X1 4	1
5		-1.5*

*anticipated coefficient obtained such that all coefficients sum to zero, but this coefficient is not shown in the software and not a model parameter

The approach taken in generating the anticipated coefficients (c_i) is to assign a +1 to factor levels 1, 2, and 4 because they are all positive effects of the same magnitude. Then, assign negative coefficients to levels 3 and 5, such that they are equal in magnitude and adhere to the requirement of a contrast, which is $\sum_{i=1}^{q} c_i = 0$. The resulting coefficients now have a range of ($\max(c_i) - \min(c_i)$) = +1 – (–1.5) = 2.5 which is one way to view the detectable effect magnitude (δ). Again, what makes this factor effect power different from other software package multi-level categorical factor power computations, is that more than the minimum number of levels are active (nonzero coefficient estimates). So in this case, the effective δ is more than just the gap between the smallest and largest coefficients, because up to q levels are contributing to changes in the response.

To illustrate the calculation for factor effect power using the instructions supplied in Appendix C, consider the one-factor design above with 15 runs. The non-centrality parameter λ is calculated as

$$\lambda = (\boldsymbol{Lb})^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}\boldsymbol{Lb} = 22.5,$$

where

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{b} = [1 \quad 1 \quad 1 \quad -1.5 \quad 1]^{\boldsymbol{T}},$$

and $\boldsymbol{X}$ has 15 rows and five columns. The first column of $\boldsymbol{X}$ contains 1's and corresponds to the intercept, while columns two through five makeup the contrast coding for each of the five levels. The critical F value is calculated as $\acute{F} = F^{-1}\{1-\alpha, q, n-p\} = F^{-1}\{1-0.05, 4, 15-5\} = 3.48$. Power is then computed as $P = 1 - \tilde{F}\{\acute{F}, q, n-p, \lambda\} = 1 - \tilde{F}\{3.48, 4, 15-5, 22.5\} = 0.87$. For complete details on this calculation procedure see Appendix C.

c. JMP 11 Default Factor Effect Power and Comparison with Coefficient Power

The default coefficients for *q*-level categorical factors consists of all active coefficients of magnitude = 1, alternating sign starting with +1. In the case of odd *q*, the final coefficient is set = 0 (again to satisfy the contrast constraint $\sum_{i=1}^{q} c_i = 0$). These nonzero coefficients all contribute to δ, such that the majority of the time for factors with more than three levels the factor effect power greatly exceeds the most conservative power.

The two variants of a two-factor design (3 x 4) used in the previous sections seem to point to problem with outright accepting JMP 11's default coefficient estimates. Figure 3-20 shows a simple investigation for a one factor test design into the influence the number of runs (*N*) has on parameter and factor effect power for a three-level categorical factor. Power for δ/σ=2 is plotted for not only JMP 11 coefficient power and factor effect power, but also for Design Expert (DX) 9. As we noted earlier for the three-level design, DX agrees perfectly with JMP 11 factor effect power, because JMP 11 defaults to the conservative coding for a three level factor. Interestingly, for the three-level design, the coefficient power is larger than the factor effect power for small *N*, while the reverse is true as *N* increases, albeit minor differences.

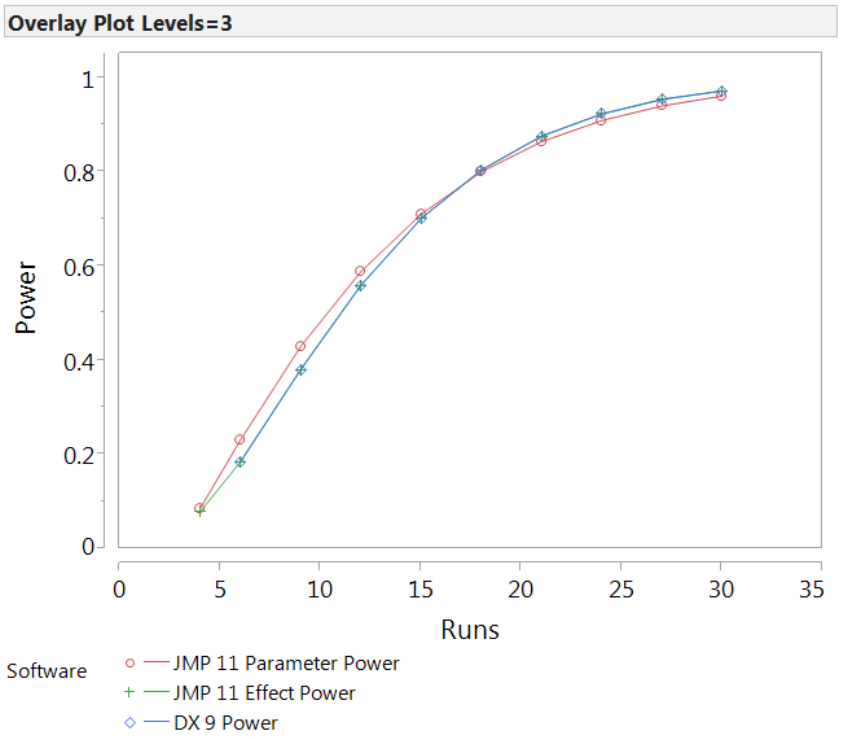


Figure 3-20. Power estimates for a 1-factor categorical three-level design, comparing three sources: JMP 11 Coefficient Power, JMP 11 Factor Effect Power, and Design Expert (DX) 9 power

The situation for the eight-level design is quite different. Figure 3-21 shows a similar comparison for an eight-level design. None of the three default power sources agree, creating possible confusion for a user or an organization. Design Expert is the more conservative, JMP 11 coefficient power gives 10 to 15 percent higher power than DX, while JMP 11 factor effect power differs substantially, reporting 20 to 45 percent higher power than the JMP 11 coefficient power. Notice that when JMP 11 anticipated coefficients are modified to the conservative approach the power calculations match the Design Expert values exactly.

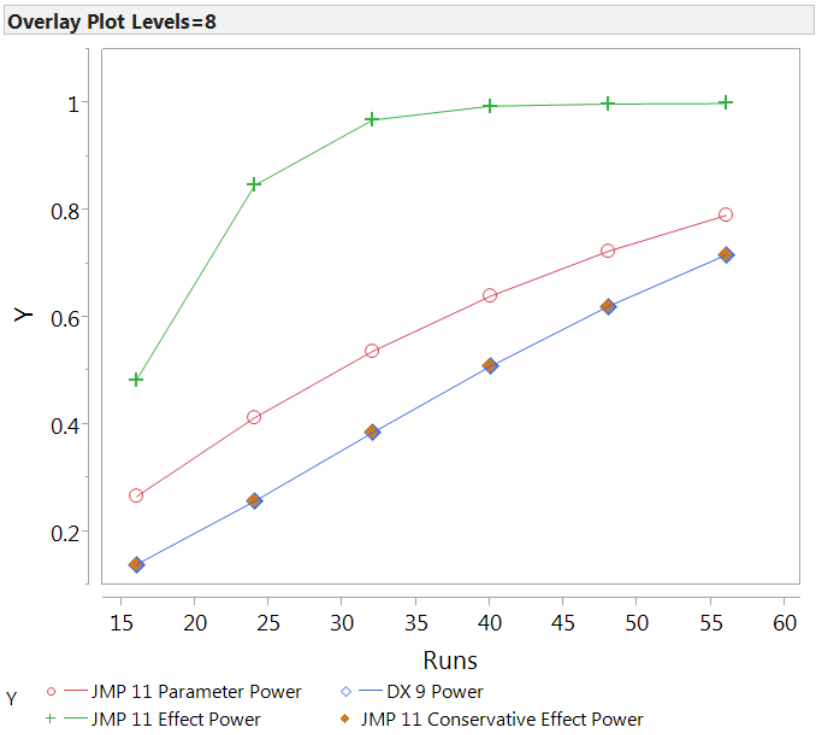


Figure 3-21. Power estimates for a 1-factor categorical 8-level design, comparing: JMP 11 Coefficient Power, JMP 11 Factor Effect Power, Design Expert (DX) power, and JMP 11 conservative power.

The bottom line recommendation from this guide is to **use the most conservative power algorithm to configure the anticipated coefficients in JMP 11** in order to achieve power estimates consistent across software packages. Additionally, this power ensures that the test will be adequate if at least two levels of the categorical factor are active. The default specifications require that all levels of the factor are active to achieve the specified power.

8. Alternative Power Specification (JMP Semi-Conservative)

As an alternative to most conservative power, consider a factor effect in which all the levels are active (non-zero) but the combined contribution by each level is less than those proposed by the JMP 11 defaults. Perhaps some prior knowledge exists that, for a given factor, all but one of the levels have similar contribution to the factor effect in their sign and magnitude, and are opposite in sign with the remaining level of that factor. This alternative, called the semi-conservative approach (used in JMP 10.0.2), is executed by setting one level of the coefficient to:

$$c_1 = \frac{-(\delta/\sigma) * (q-1)}{q}$$

where q is the number of levels for that particular factor. The remainders of the coefficients are set to:

$$c_i = \frac{(\delta/\sigma)}{q}$$

Notice that this formulation results in the sum of the coefficients being equal to zero. Table 3-5 below provides some examples of coefficients in terms of the SNR for multi-level categorical factors using this semi-conservative approach.

Table 3-5. Semi-Conservative Coefficients for Multi-Level Categorical Factors

	3 Level Factor	**4 Level Factor**	**5 Level Factor**
1st Coefficient	$-0.667(\delta/\sigma)$	$-0.75(\delta/\sigma)$	$-0.8(\delta/\sigma)$
Remaining Coefficients	$0.333(\delta/\sigma)$	$0.25(\delta/\sigma)$	$0.2(\delta/\sigma)$

These coefficients are termed semi-conservative because they will result in higher power calculations than the conservative approach. This is because the absolute value of the large coefficient is larger than the conservative approach for a fixed SNR.

E. Power Comparison across Packages

In general, the power analysis approach that both JMP 10 (we omit JMP 9 due to the SNR definition difference) and DX 9/8 software programs take for power analysis is to determine the most conservative power configuration (for more than two factors) of the factor level contributions by iteratively computing power for all pairs of factor levels, until the lowest power combination is found.

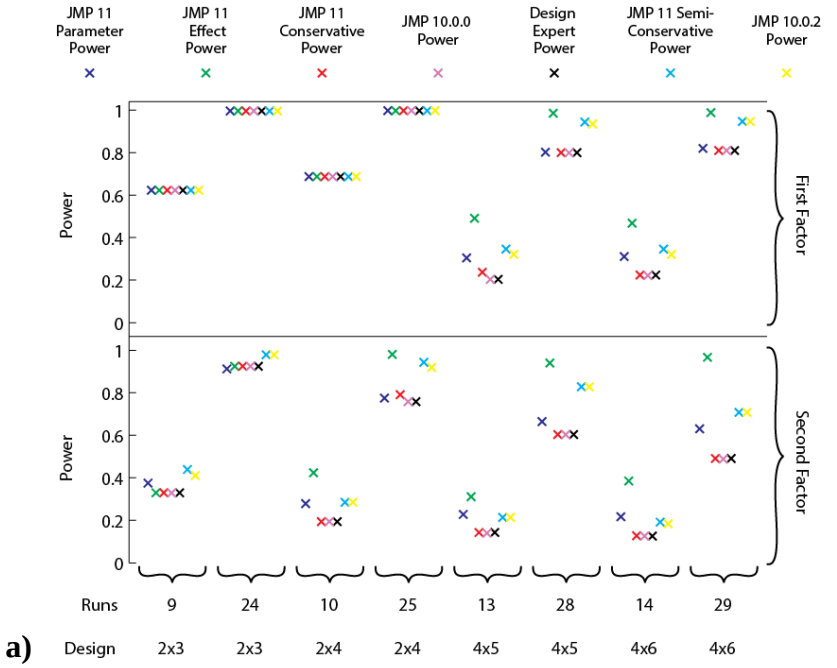
Several design types which involve varying the number of factors, number of levels per factor, and total number of runs are considered for a power analysis comparison. Software versions JMP 10, JMP 11 and DX 9 are compared. Power is reported as factor effect power. The designs considered differ in terms of the number of factors (k = 2, 3), number of levels per factor (q=2, 3, 4, 5, 6, or 8), number of runs ($N=p+5$, $p+20$, where $p=k+1$)). For example a 3-factor design, may have Factor A with q_A = 4, and Factor B with q_B = 6, and Factor C with q_C = 8, which is written as 4 x 6 x 8. Table 3-6 provide a summary of the 16 designs considered to compare across the software packages. Each of the 8 design types consisted of a smaller and moderate number of runs, for 16 total designs. Smaller run designs have 5 degrees of freedom error beyond a main effects' model, and larger designs have 20 degrees of freedom error beyond a main effects' model.

Table 3-6. Designs for Power Comparison

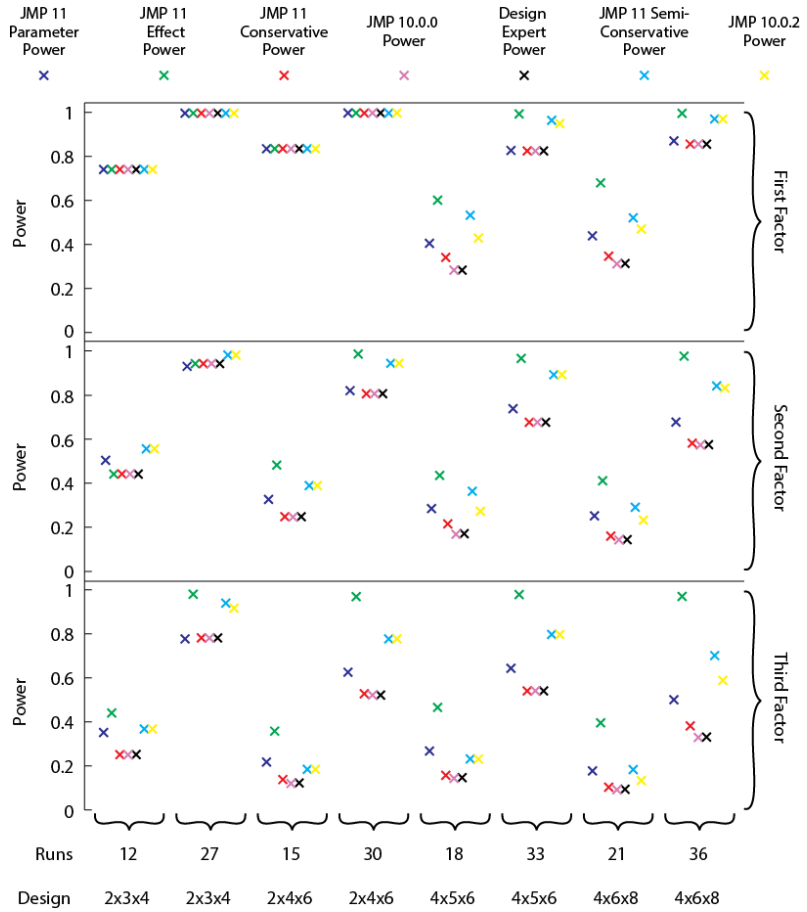
2-Factor Designs	Runs (2 designs)	3-Factor Designs	Runs (2 designs)
2 x 3	9, 24	2 x 3 x 4	12, 27
2 x 4	10, 25	2 x 4 x 6	15, 30
4 x 5	13, 28	4 x 5 x 6	18, 33
4 x 6	14, 29	4 x 6 x 8	21, 36

The designs in Table 3-6 provide a diverse set of likely scenarios for testing in a restricted resource environment, especially when multi-level categorical variables dominate a system description. The designs are not unusual in run number too, ranging from 9 to 36 runs. We would expect some of the designs to have reasonable power, while others would have moderate to low power. The set of 16 designs were used to compare all the power estimation approaches discussed in this guide. Included in the comparison are JMP 9, 10 (listed as 10.0.0), 11 default, 11 parameter, 11 conservative, and Design Expert. Also included are the power estimates from JMP 10.0.2 and power obtained by manipulating anticipated coefficients in JMP 11 to mimic JMP 10.0.2 power (called semi-conservative power). Power for JMP 10.0.2 is detailed in Appendix C. For the two-level factors, the power is the same, but the disparities increase as the design considered increases in number of factors and in the number of factor levels (Figures 3-22 a. and b).

Notice in Figure 3-22 that JMP 11 conservative power compares well with JMP 10.0.0 and Design Expert, while JMP 10 semi-conservative power compares well with JMP 10.0.2. In fact, differences in the power reported are due to slight differences in the optimal designs generated by the two packages. JMP 11 coefficient power can be used to understand the JMP 9.0 power calculations in these scenarios.



a)



b)

Figure 3-22. Power analysis comparison across software platforms using SNR=2 and all methods discussed in this guide for 16 2- and three-factor mixed level designs. Part a) contains 2-factor designs while part b) shows three-factor designs. Smaller run designs have 5 df error over a main effects model, and larger designs have 20 df error.

F. Power Analysis Practice Tips

- Recall that power is only one of many design metrics. Develop a table of metrics not just for assessing one design, but several design alternatives. Other notable design metrics include the anticipated model form, error degrees of freedom, correlation measures such as the variance inflation ratio, the number of replicate runs, the number of factor levels, the range of power values, and prediction variance metrics such as the median and 90 percentile standard error of the mean across the design space (using fraction of the design space). For more information on these other useful metrics see, DOT&E Memo, July 23, 2013 memorandum, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation."

- Make sure you know what is assumed if you are to accept all the default settings in the power analysis. Each software and version has built-in defaults for the model (number of terms, hence degrees of freedom for the model and degrees of freedom for error), and either the values for δ and σ, or the δ/σ ratio. It is always a good idea to check the defaults and even better to modify them to suit your problem.

- Often we build very efficient designs, particularly for screening. If few (or no) design runs are available to estimate error, power is greatly compromised (or un-estimable if degrees of freedom for error = 0). Be careful to consider whether the design built is near saturated in runs *(N)* relative to the anticipated general model. A design is saturated if $N = p$ (the anticipated general model parameters). For design construction and power analysis, there are two model forms to construct. The first model is used in building the design points and contains all the model terms you want to be able to estimate. This initial model, used to generate the design, is often called the general model (e.g., main effects plus two-factor interaction). The second model is used directly in power analysis only and the objective in creating this model is that it contains about the right number of model terms (translated to degrees of freedom) that will be in the final model. The effect sparsity principle combined with decades of DOE experience has suggested that this second model for power analysis contain approximately the number of degrees of freedom in a model with only the main effects. The anticipated general model has p_{gen}degrees of freedom, while the model for power has p power degrees of freedom, where $p_{\text{gen}} > p_{\text{power}}$. For a two-level design with a first order (main effects) plus two-factor interaction anticipated model, $p_{\text{gen}} = 1 + k + \frac{k(k-1)}{2}$. When computing power for a design where $N \cong p_{\text{gen}}$, choose a model for power analysis to include only main effects, p_{power}. The idea is that the number of main effects is roughly the number of model terms (main effects or interactions) we expect to be significant, so this approach is sound. So, when

$N \cong p_{\text{gen}}$, power is only reported for main effects. Power for specific interactions can be estimated by adding one interaction effect to the model at a time. With multi-level categorical factors, it is a best to check the power of interactions as well, so you can iteratively exchange main effects for interactions, keeping enough (more than two) degrees of freedom for error.

- In Design Expert, when constructing a design, you are asked for delta and sigma. The default model for power analysis is main effects only, but it can be changed as well, under Options (Figure 3-23).

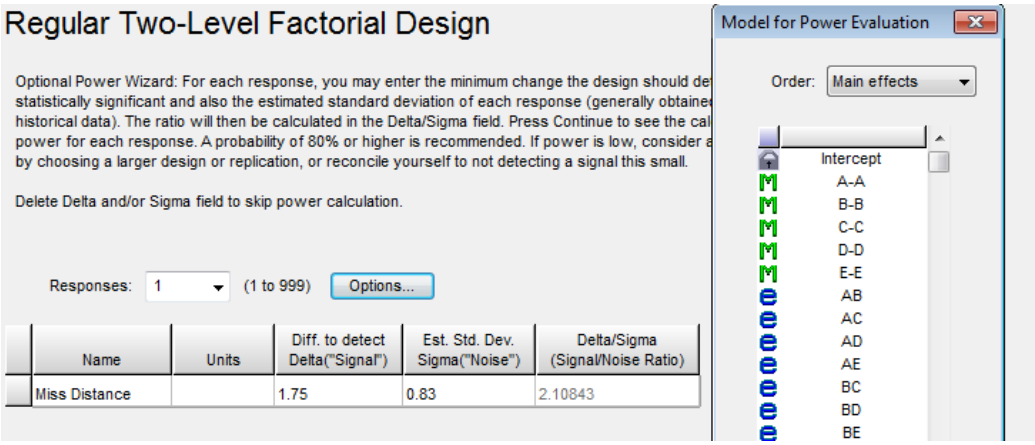


Figure 3-23. Design Expert Power Wizard defaults to a model with main effects only, but can be modified

- For the near saturated ($N \cong p_{\text{gen}}$) case in JMP, first make sure to **specify the full anticipated model** prior to building the design. For example, suppose you have five two-level factors and are considering 16 runs for screening and you need to build a design to assess power. Understand the default model in JMP is main effects only, so you'll need to add the two-factor interactions (Figure 3-24).

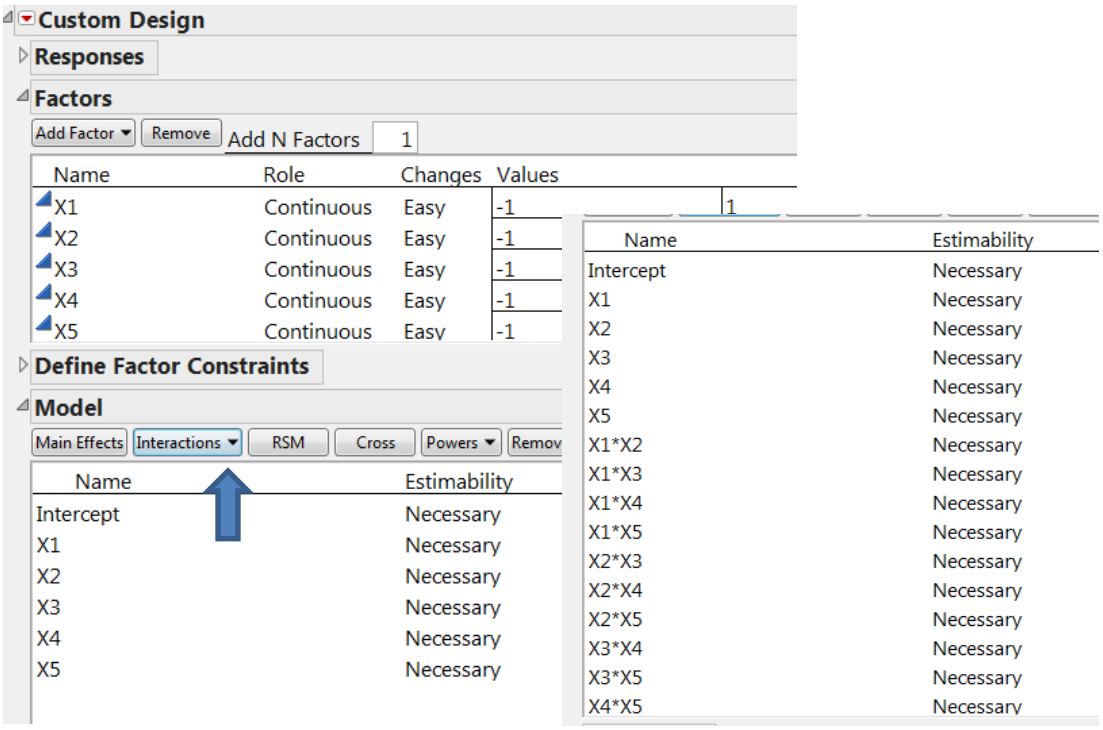


Figure 3-24. JMP model default is main effects, so for a screening design, also select *Interactions, 2nd*

- Then build the design, and prior to performing the power analysis, ensure sufficient error degrees of freedom. In the above example of a two-level screening design, choosing the $p_{\text{power}} = k$ is reasonable. Remove the interactions from the model after the design is built, then go to the power analysis section and *Apply Changes to the Anticipated Coefficients*. Only the main effects should remain in the power analysis. If the model is not reduced from $p_{\text{gen}}$to p_{power} the power values will be very low because the error degrees of freedom are near zero.
- If sufficient error degrees of freedom exist for the anticipated model (i.e., more than five error degrees of freedom), you can just accept the same model used to build the design, and check the power for all the model anticipated effects of interest.
- If the design objective involves optimization (as opposed to characterization or screening), then a response surface design is most likely needed. For most response surface investigations the sample size is adequate and statistical power is not important relative to the prediction variance metrics such as plots of standard error of prediction, the fraction of design space plot, and various optimality efficiencies. Please note though, that second order screening designs are now available, which are small run designs, so be careful not to neglect power in response surface designs where $N < \frac{(k+1)(k+2)}{2}$, that is the

number of runs (N) is less than what would be required for the k-factor full quadratic model.

G. Summary of Power for Multi-Level Categorical Factors

- Statistical power for categorical factors with more than two levels requires an additional decision or assumption be made regarding the nature of the factor effect.
- Because each of the factor levels can be thought to stand on their own, a common modeling approach used is indicator variables.
- For a factor of this type, one must decide how many levels are active, assuming that the effect is real.
- Standard approaches historically (and currently in JMP 9/10 and DX) for active levels is to assume the most conservative scenario with only a pair of levels different by δ.
- Conservative power is reported by default in JMP 9/10 and DX, whereas JMP 11 allows the user to specify the factor level effects.
- JMP 11 power analysis is purposefully adapted to provide the user flexibility in tailoring factor effect power for categorical factors with more than two levels.
- JMP 11 default anticipated coefficients make all factor levels active (with coefficient SNR/2), except the last level for factors with odd numbered levels.
- JMP 11 also provides coefficient power, which gives the power for that level's indicator variable. JMP coefficient power values more closely align with effect most conservative power, but enough differences exist (Figure 3-22) not to use it that way.
- JMP 11 provides an option for the user to input delta under Advanced Options. While this option is useful for two-level designs, when using it for multi-level categorical factors, JMP inserts a $SNR/2$ value for every parameter, equivalent to the JMP 11 default approach. As such the coefficient and factor effect power values will generally be overly optimistic relative to other software, including JMP 10.
- JMP 11 anticipated coefficients can be structured fairly easily for most conservative factor effect power.
- Due to the differences between most conservative factor effect power, coefficient power and default factor effect power, it is important that the user understand the assumptions and interpretation of these different estimates when using software to estimate power involving more than two level factors.

- **It is highly recommended,** to ensure test adequacy and for consistent reporting across software platforms**, that users of JMP 11 configure the anticipated coefficients for most conservative power.**

4.Conclusions and Recommendations

A. Extension to Additional Analysis Model

This guide provided a detailed discussion on power calculations for continuous factors, two-level categorical factors, and multi-level categorical factors. The models discussed included main effects and second order interactions. The concepts behind the power calculations can be simply extended to higher order models involving higher order interactions, quadratic terms, and other polynomial terms.

Higher order interactions follow the rules of the factors involved in the interaction. Therefore, if all of the factors involved in the interaction have two-levels the simple recommendations for the two-level factors apply. If one or more of the factors involved in the interaction has more than two levels than the multi-level categorical response rules apply.

Quadratic terms provide another interesting contrast between the software packages. Design Expert actually doubles the coefficient in the power calculation for the quadratic effect. DX does this because they only consider quadratic effects over half of the design space (0 to 1 in coded values) so essentially double the size of the coefficient to get the same change in the response. JMP on the other hand (except JMP 10.0.0), because they focus power on the coefficients, do not double the coefficient. We recommend that when quadratic terms are of interest that users divide the Design Expert SNR by two.

B. Summary of Results

While, the models and mathematics behind statistical power calculations are always the same, different software packages have implemented the default assumptions differently. JMP 11 provides the most flexibility and the ability to reproduce power calculations from DX and other versions of JMP. However, the defaults values for the coefficients for JMP 11 can lead to very misleading power calculations especially for multi-level categorical variables if the user is not aware of the implicit assumptions. Design Expert provides a consistent methodology for calculating power. However, the power for quadratic terms may be overly optimistic. However, both of these packages provide defensible and useful power estimates when the underlying assumptions are clearly understood.

C. Power Analysis Software Recommendations

Below we provide general recommendations for inputs to the software packages. These general recommendations should only be considered when no other information is available to help inform these assumptions. While they are based on previous experiences employing DOE in operational tests, they should be modified for the specific of each individual test/system. They are by no means a substitute for engaging subject matter experts to determine the detectable difference and using past test data to inform the estimation of the noise.

1. General Recommendations for Risk Specification

The first decision that one must make in a power analysis is the acceptable level of α risk. Typical values used in test and evaluation range from 0.01 to 0.20. In the lack of any other information in selecting the risk of falsely detecting a factor as significant users should use $\alpha = 0.05$. It is important to note that selecting an α level for determining factor significance of 0.05 does not imply that all model predictions have to be made at a 95 percent confidence level. In fact, it is reasonable to select significant model factors using an α cutoff of 0.05 and then making prediction statements at a different level of confidence. Previous experience across a variety of programs has shown that using a 0.20 α cutoff for testing for significant model factors can result in an over characterization (fitting a more complex model than necessary and declaring insignificant factors significant) of system performance.

2. General Recommendations for Signal-to-Noise Ratio Estimation

The estimation of the signal-to-noise ratio is the most important aspect of power calculations. Ideally, the signal (or detectable difference in the response) should be based on the operational impact of changes in performance. The noise should be estimated based on previous test data under similar test conditions. However, it is sometimes infeasible to obtain defensible estimates of both the detectable difference and noise. In these cases it is reasonable to consider default values for the signal-to-noise ratios. Figure 4-1 shows the trade space in the power analyses for multiple test sizes and values of the signal-to-noise ratio.

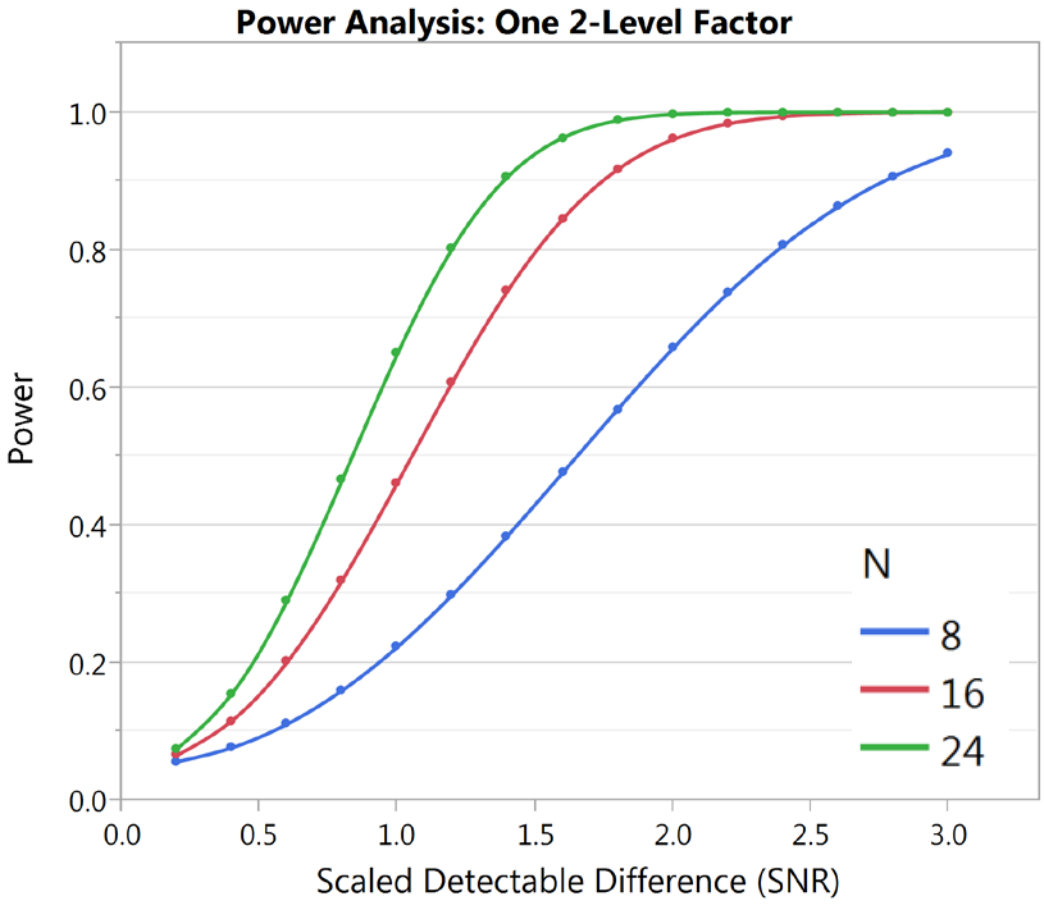


Figure 4-1 Power analysis sensitivity for a single factor with 2 levels

Figure 4-1 captures the trade space between the SNR and the test size that applies to all test designs. Power analysis is most useful for signal-to-noise ratios between 0.5 and 2.0. For continuous response variables, signal-to-noise ratios below 0.5 require extremely large test sizes to detect, and tend not to result in operationally or practically meaningful differences. On the other end, signal-to-noise ratios larger than two tend to be very obvious and visible in even extremely small tests. If we also consider the diminishing return in statistical power as sample size increases a "sweet spot" emerges for conducting power calculations.

In the lack of any information about what detectable difference and noise estimates to use in planning tests, test teams should use values between 1.5 and 2.0 at a significance level of 0.05. These generic ratios are in terms of the response outcomes so if software employs tests on coefficients the signal-to-noise ratios should be adjusted accordingly (see the next general recommendation). These inputs will provide reasonable test sizes for operational tests, and have been successful in previous operational tests. There is still a large difference in the power results for a signal-to-noise ratio of 1.5 and 2.0. Tests that employ a high level of control over the operational environment and therefore expect lower unexplained variability should lean towards 2.0. Highly operationally realistic

tests, where sources of unexplained variability are expected to be high should lean towards 1.5 or even smaller. These general recommendations only apply to continuous metrics; binary metric approaches are discussed in more detail in Chapter 2. If a significance level of 0.20 is selected instead of the recommended 0.05, then users should also adjust the SNR accordingly. In these cases, the reasonable SNR drops to between 1.0 and 1.5, with the same considerations. In terms of power, one should ideally aim for above 90 percent, and 95 percent is equivalent to a β risk of 0.05 to match the α risk recommendation. In every case, it is important to balance the risks associated with power and confidence. In some cases higher power may be appropriate when it is related to the key reasons for conducting the test.

3. General Recommendations for Software Inputs

Table 4-1 below summarizes the general recommendation for entering the signal-to-noise ratio for different types of factors for each of the software packages provided in this guide. In Table 4-1, and throughout this guide, δ refers to the change expected in the response variable as a function of changing the factor, σ refers to the model corrected estimate of the standard deviation (which in a classical Regression/ANOVA context is the root mean square error).

Table 4-1. Recommended Inputs for Signal-to-Noise Ratio in Software Packages

Software	***2 Level Factors/ Continuous Factors/ Interactions for 2 Level Factors***	***Multiple Level Categorical Factors and their Interactions***	***Quadratic Terms***
Design Expert 8, 9	δ/σ*	δ/σ	$\delta/2\sigma$
JMP 9	$\delta/2\sigma$	$\delta/2\sigma$**	$\delta/2\sigma$
JMP 10	δ/σ	δ/σ	δ/σ
JMP 11	Under advanced options use "apply delta for power" of δ/σ	Under advanced options use "apply delta for power" of δ/σ Adjust all but two coefficients to zero (conservative method described in Chapter 3)	Under advanced options use "apply delta for power" of δ/σ

*If using the generic signal-to-noise ratios suggested in the previous section this value would be between 1.5 and 2.0.

**Dividing the signal-to-noise ratio by 2 only provides an exact power calculation to match the other packages for two-level factors. JMP 9 only provides power calculations for coefficients and is not comparable to the other packages. However, using this value typically provides reasonable test sizes, despite the limitations in the power calculations.

D. Overall Recommendations

This guide provides a plethora of recommendations on the calculation of statistical power and its implementation in statistical software. The most important of these suggestions are:

- Understand the interpretation of statistical power and that test designs with more than one factor should involve multiple estimates of statistical power.
- Understand how software calculates power.
- Use continuous response variables whenever possible, when continuous response variables are not available use the SNR approximations discussed in Chapter 2.
- In the lack of good estimates of SNR use values between 1.5 and 1.5. Divide these values by two for JMP 9.0 and for quadratic effects for Design Expert.
- Use a default α risk of 0.05. Remember that you can always change the confidence level for reporting results.
- Aim for power of 90 percent or higher.

But above all else, use all of the information available including advice from subject matter experts, previous test data, operational relevance, etc. to make these decisions and discard the recommendations above when they are not consistent with the goal of the test at hand!

References

1. Barker, Clay. *Power and Sample Size Calculation in JMP*. White Paper, Cary, NC: JMP, 2011.
2. Bisgaard, Soren and Fuller, Howard. "Sample Size Estimates for 2^{k-p} Designs with Binary Responses," *Journal of Quality Technology*, 1995.
3. DOT&E Memo, July 23, 2013 memorandum, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation."
4. Freeman, Laura, Johnson, Thomas, Anderson, Colin. "Power Analysis Method for Test and Evaluation." *IDA Paper P-4887*. 2012.
5. Gotwalt, Chris. "JMP Script for Computing Binary Power using the Logit Transformation," JMP, 2012.
6. Gotwalt, Chris. "Webex on Computing Power in JMP," JMP, 2014.
7. Lenth, Russ. Piface Java Applet version 1.76, University of Iowa, 2011.
8. Lenth, Russ. "Some Practical Guidelines for Effective Sample Size Determination." *The American Statistician*, 2001: 187-193.
9. Montgomery, Douglas. *Design and Analysis of Experiments.* Hoboken, New Jersey: John Wiley and Sons, 2008.
10. Montgomery, Douglas. *Applied Statistics and Probability for Engineers.* 5thEdition. Hoboken, New Jersey: John Wiley and Sons, 2011.
11. Myers, Raymond, Montgomery, Douglas, and Anderson-Cook, Christine. *Response Surface Methodology: Process and Product Optimization using Designed Experiments.* Third Edition. New York: John Wiley and Sons, 2009.
12. Muthen, Linda, and Bengt Muthen. "How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power." *Structural Equation Modeling* (Structural Equation Modeling), 2002: 599-620.
13. Oehlert, Gary W. and Whitcomb, Pat. "Sizing Fixed Effects for Computing Power in Experimental Designs." *Quality and Reliability Engineering International*, 2001: 291-306.
14. Ortiz, Francisco. "Dealing with Categorical Data 3 Part: Binary Power Best Practices," *Scientific Test and Analysis Techniques for Test and Evaluation Center of Excellence Technical Document*, 2014.

15. SAS Institute Inc., *JMP Statistical Software v9.0*, Cary, NC, 20XX
16. SAS Institute Inc., *JMP Statistical Software v10.0*, Cary, NC, 20XX
17. SAS Institute Inc., *JMP Statistical Software v11.0*, Cary, NC, 2013.
18. SAS Institute Inc., *JMP 10 Design of Experiments Guide*, Cary, NC: SAS Institute Inc., 2012.
19. Simpson, James, Listak, Charles, and Hutto, Gregory. "Guidelines for Planning and Evidence for Assessing a Well Designed Experiment," *Quality Engineering*, Vol 25, No. 4, 2013.
20. Stat-ease Inc., *Design Expert v8.0*, Minneapolis, MN, 2007.
21. Stat-ease Inc., *Design Expert v9.0*, Minneapolis, MN, 2013.
22. Stat-ease Inc., *Handbook for Experimenters v 09.01.03*, Minneapolis, MN: Stat-ease, 2014.
23. Whitcomb, Pat, and Anderson, Mark. *Excel Sample Size Calculator for Binary Responses*, Stat-ease Inc., 2000.

Appendix A
Acronyms and Glossary

k	number of factors
q	number of factor levels
c	value of an anticipated coefficient
df	degrees of freedom
p	number of model parameters
N	total number of design runs, or total sample size
λ	reference distribution non-centrality parameter
SNR	signal-to-noise ratio, and equivalent to δ/σ for this guide
Model parameter	also called a regression coefficient indicates the change in a response when a factor variable changes from 0 to 1
Anticipated coefficient	values set as model parameter estimates to compute power
Effect	average change in the response due to changing factor levels

Appendix B
Binary Response Power

It is often beneficial to collect both continuous responses and binary response variables. In these situations analysts should make the clear distinction between the continuous and binary responses so that calculations can be conducted using different methods for the continuous responses versus the binary responses.

Design Expert and JMP both currently assume responses are continuous in power computations (although Design Expert v9.1 is expected to allow the user the capability of performing binary response power analysis using the methods described in this appendix). Design Expert currently support through documentation and help files the use of the Bisgaard-Fuller (1997) method for binary response power described below, and JMP plans to address this problem in a future version.

There are many other response types between binary and continuous. However, binary and continuous tend to be the most common in test and evaluation. Other types of responses include multi-level (>2) categorical (multinomial), ordinal, and integer. Power calculations for multi-level, ordinal, and integer responses are beyond the scope of this guide. However, they can be done using Monte Carlo simulation modified to the appropriate distributions described in the next section, or using a signal to noise approximation as described in section 2.C of the guide, again appropriately modified.

A. Monte Carlo Methods for Calculating Binary Statistical Power

Monte Carlo simulation can be used to calculate power for a binary response. Monte Carlo simulations are a set of computational algorithms in which data are generated from a population with given parameter values. They have been widely used in computer simulations of physical systems. However, they also provide a general approach for statistical power calculations. Muthen and Muthen (2002) provide a good tutorial for calculating power via a Monte Carlo simulation.

The general steps in using a Monte Carlo simulation to calculate binary response power for a designed experiment are:

1. Select a candidate test design.
2. Generate a sample of test data for the proposed design.
   a. Data generation involves specifying the proportion of expected successful outcomes for each condition of the test and generating a potential outcome

using the binomial distribution. The proportion can be specified either through setting selecting a statistical model and the corresponding model coefficients or through directly assigning and outcome probability to the specific set of conditions. It is important that the data generation reflect the detectable difference of interest.

3. Perform the analysis of the simulated test data (typically a logistic regression).
4. Determine if each of the factors in the test design (or higher order model terms) are significant.
5. Repeat steps 2 through 4 several times (typically 10,000 is a good number). Note the data generated will change, because of the use of the binomial distribution.
6. Calculate power for each factor in the test. Power for each factor or model term is simply the number of times the factor/model term was significant divided by the number of simulations (e.g., 10,000). Recall, we know it is a correct rejection because we generated data under the detectable differences of interest.
7. Bonus Step: It is always good to verify that the size of the test (α error) is reasonably close to the predetermined specified level. This is especially important in cases where the proposed sample size is small and the assumed distribution is the binomial. To do this, repeat steps 2 through 4, except this time generate data with a common proportion of success across all conditions. The α error (Type I) will be the total number of times the null hypothesis was incorrectly rejected divided by the total number of simulation iterations (e.g., 10,000).

Monte Carlo simulations can be difficult to implement the first time. JMP has provided a JMP script that can be used in JMP software to conduct the Monte Carlo simulation. The JMP script is provided in Appendix D.

B. Bisgaard-Fuller Arcsin Square Root Method – For Two-Level Designs

The first method for binary response power is based on an approach outlined in Bisgaard and Fuller (1995). The approach deals with the binomial data by using a variance stabilizing transformation on the response, the observed proportion of success($\hat{p}$). The purpose of a variance stabilizing transformation is to satisfy the statistical model assumption that the model errors have constant variance across the range of predicted values. The transformation uses the arcsine square root transformation. This new transformed response would be the response used in the analysis.

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

This transformation provides a way of estimating the number of replicates needed for a 2^{k-f} factorial design. The authors' formulation of the signal of interest (the change in the response we wish to detect) on the transformed scale is:

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

where $\bar{p}$ is the expected proportion across the design space, and Δ is the signal (as a proportion) to be detected. The individual point number of replicates (n) per test condition is then calculated by the following formula:

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

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the critical values of the normal distribution based on the specified power and confidence, N is the total number of unique test conditions (2^{k-f}) and δ is defined in equation (2). Bisgaard and Fuller (1995) provide several tables of test sizes for reference.

C. Signal to Noise Approximation Methods – Any Design

Signal to noise approximation methods are extremely useful for approximating power using standard software packages. The following three approaches provide similar outcomes and can be applied to any design type, especially designs with multi-level categorical factors. To use the approximation methods, one first generates an estimate of the binary response δ/σ, and then inserts that δ/σ into standard statistical software to calculate power. Ortiz (2014) describes and compares the various approaches from the literature. He has also developed an interactive Excel spreadsheet, which calculates either the power/sample size, or an equivalent δ/σ ratio which can be input directly into either Design Expert or JMP during design construction and assessment.

1. Arcsine Formulation

The first approximation of the signal to noise ratio introduced in Ortiz (2014) is the arcsine method. The arcsine formulation, where δ (in the transformed scale) is the same as in shown above, and 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:

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

where the number of replicates (or tests involving 0/1 outcomes per test condition), $n = 1$ since we wish to determine the ratio prior to replication. In the case of power for the binomial response, the ratio is first determined, then replication is used to achieve the desired power.

2. Logit Transformation Formulation

This second approximation of the signal to noise ratio uses the logit transformation, which stems from the traditional solution used when applying logistic regression to fit a model where the dependent variable is a proportion. The transformation takes the log of the odds

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

where p is the (binomial) proportion of successes, $1-p$ is the proportion of non-successes, $\frac{p}{1-p}$ is the odds of the event. δ in the transformed scale is defined below.

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

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

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

where $n = 1$ again, since we wish to determine the ratio prior to replication.

It is important to note for this method that the SNR is computed as SNR = $\delta_2 * \sigma_2$, which stems from knowing that the information matrix for logistic regression is.

$$\mathbf{X'WX}$$

where $\mathbf{X}$ is the design matrix, and the $\mathbf{W}$ is made up of the $p_i(1\text{-}p_i)$ – the variance of the observations. We assume that there is some p that is close enough that

$$\mathbf{(X'X)p(1-p)}$$

is a 'reasonable' approximation to $\mathbf{X'WX}$. The variance of the model parameters is then approximated by $\mathbf{(X'X)^{-1}/p(1-p)}$. Since $\mathbf{p(1-p)}$ is in the denominator of the variance it becomes part of the numerator of the SNR.

3. Normal Approximation Formulation

The final approximation of the signal to noise ratio is based on the Normal approximation of the binomial. This is the simplest of the formulations presented in this paper. In this case δ is defined as:

$$\delta_3 = \left| p_1 - p_2 \right|$$

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

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

where $n = 1$ here since we wish to determine the ratio before replication.

4. General Recommendations for Binary Response Power Calculations

The signal to noise approximation provides an easy to implement solution for approximating power for binary responses to designed experiments. This is typically adequate for providing approximate power calculations. Table 2-1 shows the results from an example design used to compare the three formulations when varying p. The power estimates, as seen by the signal to noise ratios, are quite similar across methods for a fixed p. Note that the normal approximation method consistently produces the most conservative estimate of the signal to noise ratio. On the other hand, the assumed value of p can result in very different approximate signal to noise ratios. As Table 2-1 shows in most cases signal to noise ratios between 0.2 and 0.4 are useful for planning tests for binary responses. Table 2-1 only shows signal to noise ratios for proportions greater than 50 percent because proportions less than 50 percent have symmetric signal to noise ratios.

Table B-1. Comparison of Approximate Signal to Noise (SNR) Ratios

p	*Δ*	*SNR (arcsin)*	*SNR (logit)*	*SNR (normal)*
0.9	0.10	0.34	0.36	0.33
0.85	0.10	0.28	0.29	0.28
0.8	0.10	0.25	0.25	0.25
0.75	0.10	0.23	0.23	0.23
0.7	0.10	0.22	0.22	0.22
0.65	0.10	0.21	0.21	0.21
0.6	0.10	0.20	0.20	0.20
0.55	0.10	0.20	0.20	0.20
0.5	0.10	0.20	0.20	0.20
0.9	0.20	0.93	N/A	0.67
0.85	0.20	0.60	0.66	0.56
0.8	0.20	0.52	0.54	0.50
0.75	0.20	0.47	0.48	0.46
0.7	0.20	0.44	0.45	0.44
0.65	0.20	0.42	0.43	0.42
0.6	0.20	0.41	0.42	0.41
0.55	0.20	0.40	0.41	0.40
0.50	0.20	0.40	0.41	0.40

The Bisgaard-Fuller method for two-level designs, and the three signal-to-noise formulations for any design, can easily be implemented in a spreadsheet program. Ortiz (2014) has developed spreadsheet and user guides in addition to the technical report to implement all of these calculations.

Appendix C
JMP 11 Power Calculation Details

This appendix shows how JMP 11 calculates effect, and parameter power, and also shows how to calculate most conservative power to reproduce Design Expert's and JMP 10's calculations in JMP 11.

A. Effect Power

Realizing that more than one parameter is required for multi-level categorical factors, JMP provides power estimates for the multiple individual parameters for a factor or interaction, as well as the combined contribution of all the factor or interaction parameters, called effect power. The effect power calculations shown below assume a linear model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the design matrix of size $N\ x\ p$, N is the number of runs, p is the number of parameters in the model, $\boldsymbol{y}$ is the response vector of size $N\ x\ 1$, $\boldsymbol{b}$ is the coefficient vector of size $p\ x\ 1$, and $\boldsymbol{\varepsilon}$ is an error term that is uncorrelated and normally distributed with a mean of zero and variance σ^2. For this tutorial, all factors in the model are assumed to be categorical. The coefficient vector ($\boldsymbol{b}$) contains one or more coefficients for each factor or interaction effect. For a model with k effects, the coefficient vector can be written in terms of subsets of coefficients, i.e. $\boldsymbol{b} = [1 \quad \boldsymbol{b}_1 \quad \boldsymbol{b}_2 \quad \cdots \quad \boldsymbol{b}_k]$, where the ith subset of coefficients $(i = 1,2,3, \ldots, k)$ corresponds to the ith factor or interaction effect in the model. In other words, $\boldsymbol{b}_i$ is the subset of $\boldsymbol{b}$ that belongs to the ith effect in the model. For factor or interaction effect power, JMP 11 tests the hypothesis $\boldsymbol{b}_i = \boldsymbol{0}$ versus the alternative $\boldsymbol{b}_i \neq \boldsymbol{0}$ for each effect in the model. Power for the ith effect (P_i) is calculated as

$$P_i = 1 - \tilde{F}\{\acute{F}_i, g_i, N - p, \lambda_i\}$$

where $\tilde{F}$ is the non-central cumulative F distribution, F^{-1} is the inverse F central cumulative distribution, and the critical F value for the ith effect is calculated as $\acute{F}_i = F^{-1}\{1 - \alpha, g_i, n - p\}$. g_i is equal to one less than the number of levels $(q - 1)$ in the factor corresponding to the ith effect, and δ_i is the non-centrality parameter that is calculated for the ith effect and is equal to

$$\delta_i = (\boldsymbol{L}_i\boldsymbol{b})^T(\boldsymbol{L}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_i^T)^{-1}\boldsymbol{L}_i\boldsymbol{b}$$

In the calculation for λ_i, $\boldsymbol{L}_i$ (sometimes called the "hypothesis matrix") is used to isolate the subset of coefficients that are under test and is of size $g_i \; x \; p$. $\boldsymbol{L}_i$ takes the form $\boldsymbol{L}_i = [\boldsymbol{A} \quad \boldsymbol{B} \quad \boldsymbol{C}]$, where $\boldsymbol{A}$ is a matrix of zeroes of size $g_i \; x \; E$, where E is the number of parameters in the model preceding the ith effect (including the intercept). $\boldsymbol{B}$ is the identity matrix of size $g_i \; x \; g_i$. $\boldsymbol{C}$ is a matrix of zeroes of size $g_i \; x \; F$, where F is the number of parameters in the model following the ith effect.

Example:

Consider a full-factorial design with a 3-level categorical factor and a 4-level categorical factor, with one replicate and N = 12 that supports a main effects model. For this example, by default JMP 11 assumes

$$\boldsymbol{b} = [1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1]^T = [1 \quad \boldsymbol{b}_1 \quad \boldsymbol{b}_2]^T.$$

Then,

$$\boldsymbol{L}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{L}_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix},$$

$\lambda_1 = (\boldsymbol{L}_1\boldsymbol{b})^T(\boldsymbol{L}_1(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_1^T)^{-1}\boldsymbol{L}_1\boldsymbol{b} = 8.0$, and

$$\lambda_2 = (\boldsymbol{L}_2\boldsymbol{b})^T(\boldsymbol{L}_2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_2^T)^{-1}\boldsymbol{L}_2\boldsymbol{b} = 12.0 \ .$$

Proceeding with the calculation,

$\acute{F}_1 = F^{-1}\{1-\alpha, g_1, N-p\} = F^{-1}\{1-.05,2,12-6\} = 5.14$, and

$$\acute{F}_2 = F^{-1}\{1-\alpha, g_2, N-p\} = F^{-1}\{1-.05,3,12-6\} = 4.76 \ .$$

Power is calculated as

$P_1 = 1-\tilde{F}\{\acute{F}_1, g_1, N-p, \lambda_1\} = 1-\tilde{F}\{5.14, 2, 12-6, 8.0\} = 0.49$, and

$$P_2 = 1-\tilde{F}\{\acute{F}_2, g_2, N-p, \lambda_2\} = 1-\tilde{F}\{4.76, 3, 12-6, 12.0\} = 0.54.$$

1. Semi-Conservative Power (JMP 10.0.2)

Semi-conservative power is calculated similarly to effect power and provides power estimates that are very close to the values provided by JMP 10.0.2. Borrowing from the

previous example, here is an illustration of how to calculate semi-conservative power in JMP 11.

Consider a full-factorial design with a 3-level categorical factor and a 4-level categorical factor that supports a main effects model. For this example, the default anticipated coefficients supplied by JMP 11 are

$\boldsymbol{b} = [1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1]^T = [1 \quad \boldsymbol{b}_1 \quad \boldsymbol{b}_2]^T$, where

$\boldsymbol{b}_1 = [1 \quad -1]^T$ and $\boldsymbol{b}_2 = [1 \quad -1 \quad 1]^T$.

For semi-conservative power, we change the anticipated coefficients to

$\boldsymbol{b}_1 = [SNR/q_1 \quad SNR/q_1]^T$ and $\boldsymbol{b}_2 = [SNR/q_2 \quad SNR/q_2 \quad SNR/q_2]^T$,

where *SNR* is the signal to noise ratio, q_1 is the number of levels in the first factor, and q_2 is the number of levels in the second factor. *SNR* in this calculation is the *SNR* inputted into JMP 10.0.2. In this example we'll assume the signal to noise is equal to one, so *SNR* = 1.0. We then have

$\boldsymbol{b}_1 = [0.33 \quad 0.33]^T$ and $\boldsymbol{b}_2 = [0.25 \quad 0.25 \quad 0.25]^T$.

Then,

$$\boldsymbol{L}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} ,$$

$$\boldsymbol{L}_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ,$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix},$$

$\lambda_1 = (\boldsymbol{L}_1\boldsymbol{b})^T(\boldsymbol{L}_1(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_1^T)^{-1}\boldsymbol{L}_1\boldsymbol{b} = 2.67$, and

$$\lambda_2 = (\boldsymbol{L}_2\boldsymbol{b})^T(\boldsymbol{L}_2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_2^T)^{-1}\boldsymbol{L}_2\boldsymbol{b} = 2.25 \ .$$

Proceeding with the calculation,

$\acute{F}_1 = F^{-1}\{1-\alpha, g_1, N-p\} = F^{-1}\{1-.05,2,12-6\} = 5.14$, and

$$\acute{F}_2 = F^{-1}\{1-\alpha, g_2, N-p\} = F^{-1}\{1-.05,3,12-6\} = 4.76 \ .$$

Power is calculated as

$P_1 = 1 - \tilde{F}\{\acute{F}_1, g_1, N-p, \lambda_1\} = 1 - \tilde{F}\{5.14, 2, 12-6, 2.67\} = 0.19$, and

$$P_2 = 1 - \tilde{F}\{\acute{F}_2, g_2, N-p, \lambda_2\} = 1 - \tilde{F}\{4.76, 3, 12-6, 2.25\} = 0.13.$$

2. Conservative Power (Design Expert and JMP 10.0.0)

Conservative power is calculated similarly to effect power and provides power estimates that are very close to the values provided by Design Expert and JMP 10.0.0.

Borrowing from the previous example, here is an illustration of how to calculate conservative power in JMP 11.

Consider a full-factorial design with a 3-level categorical factor and a 4-level categorical factor that supports a main effects model. For this example, the default anticipated coefficients are

$\boldsymbol{b} = [1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1]^T = [1 \quad \boldsymbol{b}_1 \quad \boldsymbol{b}_2]^T$, where

$\boldsymbol{b}_1 = [1 \quad -1]^T$ and $\boldsymbol{b}_2 = [1 \quad -1 \quad 1]^T$.

For conservative power, we change the anticipated coefficients to

$\boldsymbol{b}_1 = [0 \quad SNR/2]^T$ and $\boldsymbol{b}_2 = [0 \quad 0 \quad SNR/2]^T$,

where *SNR* is the signal to noise ratio as defined by Design Expert of JMP 10. The nonzero coefficients of magnitude *SNR*/2 are selected as those with the lowest parameter power when JMP 11 default anticipated coefficients are targeted. Most analysts who use this conservative power calculation will be doing so to replicate the power provided in Design Expert and JMP 10.0.0. *SNR* in this calculation is the *SNR* inputted into Design Expert or JMP 10.0.0. In this example we'll assume that we'd like to reproduce the power estimates provided by Design Expert or JMP 10.0.0 where the signal to noise ratio in those packages is equal to one, so *SNR* = 1.0. We then have

$\boldsymbol{b}_1 = [0 \quad 0.5]^T$ and $\boldsymbol{b}_2 = [0 \quad 0 \quad 0.5]^T$.

Then,

$$\boldsymbol{L}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$\boldsymbol{L}_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix},$$

$\lambda_1 = (\boldsymbol{L}_1\boldsymbol{b})^T(\boldsymbol{L}_1(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_1^T)^{-1}\boldsymbol{L}_1\boldsymbol{b} = 2.0$, and

$$\lambda_2 = (\boldsymbol{L}_2\boldsymbol{b})^T(\boldsymbol{L}_2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_2^T)^{-1}\boldsymbol{L}_2\boldsymbol{b} = 1.5 \ .$$

Proceeding with the calculation,

$\acute{F}_1 = F^{-1}\{1-\alpha, g_1, N-p\} = F^{-1}\{1-.05,2,12-6\} = 5.14$, and

$$\acute{F}_2 = F^{-1}\{1-\alpha, g_2, N-p\} = F^{-1}\{1-.05,3,12-6\} = 4.76 \ .$$

Power is calculated as

$P_1 = 1-\tilde{F}\{\acute{F}_1, g_1, N-p, \lambda_1\} = 1-\tilde{F}\{5.14, 2, 12-6, 2.0\} = 0.15$, and

$$P_2 = 1-\tilde{F}\{\acute{F}_2, g_2, N-p, \lambda_2\} = 1-\tilde{F}\{4.76, 3, 12-6, 1.5\} = 0.10.$$

3. Parameter Power

The parameter power calculations below assume a linear model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the design matrix of size $N\ x\ p$, n is the number of runs, p is the number of parameters in the model, $\boldsymbol{y}$ is the response vector of size $N\ x\ 1$, $\boldsymbol{b}$ is the coefficient vector of size $p\ x\ 1$, and $\boldsymbol{\varepsilon}$ is an error term that is uncorrelated and normally distributed with a mean of zero and variance $\boldsymbol{\sigma^2}$ .Recalling from before, effect power deals with hypotheses on all the coefficients within an effect, i.e. the null hypothesis $\boldsymbol{b}_i = \boldsymbol{0}$ versus the alternative $\boldsymbol{b}_i \neq \boldsymbol{0}$. Parameter power takes a slightly different approach by testing the hypothesis $b_j = 0$ versus the alternative $b_j \neq 0$ for the jth parameter within model. Effect power tests a set of coefficients, whereas parameter power test a single parameter at a time. Power for the j^{th} parameter is calculated as

$$P_j = 1 - \tilde{F}\{\acute{F}, 1, n-p, \lambda_j\} \quad ,$$

where the critical F value is calculated as $\acute{F} = F^{-1}\{1-\alpha, 1, n-p\}$. The non-centrality parameter is then

$$\lambda_j = \left(\boldsymbol{Q}_j\boldsymbol{b}\right)^T\left(\boldsymbol{Q}_j(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Q}_j^T\right)^{-1}\boldsymbol{Q}_j\boldsymbol{b} \quad ,$$

where $\boldsymbol{Q}_j$ is a one-dimensional row vector of length equal to the column vector $\boldsymbol{b}$ and contains all zeroes except for the jth parameter, which is set equal to one.

Example:

Consider, again, a full-factorial design with a 3-level categorical factor and a 4-level categorical factor that supports a main effects model. The design matrix is

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} ,$$

and for this example, by default, JMP 11 provides

$$\boldsymbol{b} = [1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1]^T \ .$$

The first column of the design matrix corresponds to the intercept. The second and third columns correspond to the three-level factor, while the last three columns correspond to the four-level factor. To calculate power for the third parameter in the model (which belongs to the second level of the three-level categorical factor), we have

$$\boldsymbol{Q}_3 = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0].$$

$$\delta_3 = (\boldsymbol{Q}_3\boldsymbol{b})^T(\boldsymbol{Q}_3(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{Q}_3^T)^{-1}\boldsymbol{Q}_3\boldsymbol{b} = 6.0.$$

Proceeding with the calculation,

$$\acute{F} = F^{-1}\{1-\alpha, 1, N-p\} = F^{-1}\{1-.05,1,12-6\} = 5.99 \ .$$

Power is calculated as

$$P_3 = 1 - \tilde{F}\left\{ \acute{F}, 1, N - p, \lambda_3 \right\} = 1 - \tilde{F}\{ 5.99, 1, 12 - 6, 6.0 \} = 0.54 \, .$$

Appendix D
Design Expert Power Calculation Details

The Design Expert power calculations below assume a linear model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the design matrix of size $N\ x\ p$, n is the number of runs, p is the number of parameters in the model, $\boldsymbol{y}$ is the response vector of size $N\ x\ 1$, $\boldsymbol{b}$ is the coefficient vector of size $p\ x\ 1$, and $\boldsymbol{\varepsilon}$ is an error term that is uncorrelated and normally distributed with a mean of zero and variance σ^2. It is assumed that all factors are categorical.

In practice, we wish to determine the power to observe whether a model term, such as a main effect or interaction, is significant. Suppose there are δp coefficients associated with that model term. The coefficients vector can be split into two vectors, one, $\delta\boldsymbol{\beta}$, containing the coefficients to be tested and the other, $\boldsymbol{\beta_0}$, containing the remainder of the coefficients. The design matrix can be similarly partitioned into corresponding matrices so the model can be written as:

$$y = (\boldsymbol{X_0} \quad \delta\boldsymbol{X})\begin{pmatrix}\boldsymbol{\beta_0}\\ \delta\boldsymbol{\beta}\end{pmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X_0}\boldsymbol{\beta_0} + \delta\boldsymbol{X}\delta\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The model consisting of only $\boldsymbol{X_0}$ and $\boldsymbol{\beta_0}$ is the null or restricted model. If the term under test is insignificant, then $\delta\boldsymbol{\beta} = \boldsymbol{0}$. Hence, the null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \quad \delta\boldsymbol{\beta} = \boldsymbol{0} \\ H_1 &: \quad \delta\boldsymbol{\beta} \neq \boldsymbol{0}. \end{aligned}$$

The test statistic is constructed from the residual sum of squares under the full and restricted models and is defined as:

$$f = \frac{N-p}{\delta p}\frac{\boldsymbol{y}^T(\boldsymbol{H}_0 - \boldsymbol{H})\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{H_0} = \boldsymbol{X_0}(\boldsymbol{X_0}^T\boldsymbol{X_0})^{-1}\boldsymbol{X_0}^T$ (these are the hat matrices). The test statistic is F-distributed under both hypotheses. Specifically,

$$f \sim \begin{cases} F(\delta p, N-p) & : \ H_0 \ is \ true \\ F(\delta p, N-p;\ \lambda) & : \ H_1 \ is \ true \end{cases}$$

where $F(\nu_1, \nu_2)$ is the F-distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom. $F(\nu_1, \nu_2;\ \lambda)$ is the singly non-central F-distribution with non-centrality parameter λ, and

$$\lambda = \left(\delta \boldsymbol{X} \frac{\delta \boldsymbol{\beta}}{\sigma}\right)^T (\boldsymbol{I} - \boldsymbol{H}_0) \left(\delta \boldsymbol{X} \frac{\delta \boldsymbol{\beta}}{\sigma}\right).$$

Given the distribution of the test statistic under the null and alternative hypotheses, power is calculated in the usual way. The null hypothesis would be rejected with significance α if

$$f > f_U = P^{-1}[F(\delta p, N-p)](1-\alpha).$$

where $P_{-1}[*]$ is the inverse cumulative distribution function of $*$. Thus the power of the test is:

$$\mathcal{P} = P^{-1}[F(\delta p, N-p;\ \lambda)](f_U).$$

Model terms for continuous factors are associated with a single coefficient. Hence, $\delta p = 1$ and $\delta \boldsymbol{\beta}$ is a one-dimensional vector: $\delta \boldsymbol{\beta} = (\delta \beta)$ and $\delta \boldsymbol{X}$ is an $N \times 1$ matrix. The unit change induced by $\delta \boldsymbol{\beta}$ is generally defined as the largest observable change in the response. If factor levels are scaled to the interval $[-1, 1]$, then the range of a linear term of the form βx is 2β. The effect size is the same for a two-factor interaction term of the form βxy. For a quadratic term of the form βx^2, the size of the effect is only β since x^2 varies from 0 to 1 on $[-1, 1]$. In general, if the model term contains any odd powers, the effect size is twice the coefficient; otherwise it is equal to the coefficient. If we seek to determine the power to observe an effect of size $\Delta\sigma$ (Δ is the signal-to-noise ratio), then we should set

$$\delta\beta = \begin{cases} \dfrac{\Delta\sigma}{2} & : \quad term \ includes \ an \ odd \ power \\ \Delta\sigma & : \quad term \ includes \ no \ odd \ powers \end{cases}$$

Design Expert and JMP 10 follow this philosophy; however, JMP 9 chooses to subscribe to another. Instead of sizing the effect based on the change in the response,

JMP 9 sizes the effect based on the change in the coefficient. Therefore, $\delta\beta{=}\Delta\sigma$ regardless of the type of model term. Table 1 summarizes the effect size for common model terms.

Table D-1. $\delta\boldsymbol{\beta}$ Values Used by JMP and Design Expert

Package	Main Effects	Two Factor Interactions
JMP 9	$\delta\beta = \Delta\sigma$	$\delta\beta = \Delta\sigma$
JMP 10	$\delta\beta = \Delta\sigma/2$	$\delta\beta = \Delta\sigma/2$
Design Expert	$\delta\beta = \Delta\sigma/2$	$\delta\beta = \Delta\sigma/2$

Categorical factors complicate the calculation of power since categorical model terms are usually described by more than one coefficient. The definition of the effect size for a categorical term and the principle for determining $\delta\boldsymbol{\beta}$ used by Design Expert is described by Oehlert and Whitcomb (2001) and will be repeated here.

Suppose we have a model with two categorical factors, A and B, that accounts for main effects and the interaction. If A has N levels and B has M, then a general linear model can be written for the predicted response using indicator variables:

$$\hat{y} = \mu + \sum_{i=1}^{N} A_i a_i + \sum_{i=1}^{M} B_i b_i + \sum_{i=1}^{N}\sum_{j=1}^{M} AB_{ij} a_i b_j \ .$$

In this example, the lowercase variables are indicator variables, which are equal to one when the treatment is in the associated level and zero otherwise. μ represents the overall mean. A_i and B_i represent the main effects and AB_{ij} represents the interaction.

The coefficients must sum to zero over any index since each coefficient represents the departure from the overall mean at the associated combination of levels. This constraint means that an q-level main effect can be sufficiently described by $q{-}1$ coefficients. These sufficient coefficients represent the effect of contrasts between the factor levels. Typically, the contrasts are defined so that the $N{-}1$ contrast coefficients are equal to the first $q{-}1$ level coefficients. This definition leaves the contrast coefficients with a straightforward interpretation. Following this convention, the level coefficients of a three-level main effect, Z, can be described by two contrast coefficients, β_1^Z and β_2^Z, through the following relationship:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1^Z \\ \beta_2^Z \end{pmatrix} \ ,$$

and in general

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} = \boldsymbol{C}_q \begin{pmatrix} \beta_1^Z \\ \vdots \\ \beta_{q-1}^Z \end{pmatrix} .$$

The columns of $\boldsymbol{C}_q$ must sum to zero to enforce the constraint on the level coefficients. The coding system in $\boldsymbol{C}_q$ is often referred to as a simple coding. Other coding systems for categorical factors exist such as the forwards difference coding or backwards difference coding, but simple coding is most common and is employed by default by JMP 9, 10, and Design Expert.

While the level coefficients are useful for describing the behavior of the model, the contrast coefficients are used in the regression. We can rewrite the model using the level coefficients:

$$\hat{y} = \mu + \sum_{i=1}^{q-1} \beta_i^A \tilde{a}_i + \sum_{i=1}^{M-1} \beta_i^B \tilde{b}_i + \sum_{i=1}^{q-1} \sum_{j=1}^{M-1} \beta_{ij}^{AB} \tilde{a}_i \tilde{b}_j \ .$$

Doing so introduces new indicator variables that are related to the previous indicator variables through the contrast matrix. In general,

$$\begin{pmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_{q-1} \end{pmatrix} = \boldsymbol{C}_q{}^T \begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} .$$

We introduced the distinction between the level coefficients and the contrast coefficients because in Design Expert the effect size is defined in terms of the differences between levels, but the regression and power calculation are carried out on the contrast coefficients. Design Expert defines the size of the effect, ϵ, induced by the main effect of a factor to be the largest absolute difference between any two levels, i.e.

$$\epsilon = \max_{i,i'} A_i - A_{i'}$$

The effect size due to a two-factor interaction is the largest quartet difference between interaction terms:

$$\epsilon = \frac{1}{2}\max_{i,i',j,j'} AB_{ij} - AB_{i'j} - AB_{ij'} + AB_{i'j'}$$

The effect size due to a three-factor interaction is the largest octet difference, and the effect size due to higher-order interactions continues the pattern.

The non-centrality parameter dictates the separation between the distribution of the test statistic under the null and alternative hypotheses – the lower the non-centrality parameter, the lower the power. Design Expert appeals to the principle of conservatism and searches for the $\delta\boldsymbol{\beta}$ that produces the desired effect size as just described and minimizes the non-centrality parameter (hence, power). Design Expert searches for the $\delta\boldsymbol{\beta}$ that minimizes the non-centrality parameter λ, and If we represent each trial solution as the column of a matrix $\delta\boldsymbol{\beta}$, then

$$\lambda = \min \text{diag}\left(\left(\delta\boldsymbol{X}\frac{\delta\boldsymbol{\beta}}{\sigma}\right)^T (\boldsymbol{I} - \boldsymbol{H}_0)\left(\delta\boldsymbol{X}\frac{\delta\boldsymbol{\beta}}{\sigma}\right)\right).$$

Design Expert Power Calculation Example

Consider a design that has one factor with three levels. The design is fully replicated resulting in six total runs. The regression model for this design is a linear model of the form $\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the design matrix of size $n\ x\ p$, n is the number of runs, p is the number of parameters in the model, $\boldsymbol{y}$ is the response vector of size $n\ x\ 1$, $\boldsymbol{b}$ is the coefficient vector of size $p\ x\ 1$, and $\boldsymbol{\varepsilon}$ is an error term that is uncorrelated and normally distributed with a mean of zero and variance σ^2. The design matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix} .$$

The first column of $\boldsymbol{X}$ corresponds to the model intercept, while the second and third columns correspond to the indicator variable settings. The design matrix is partitioned into the null design matrix and the augmentation under the alternative model:

$$\boldsymbol{X}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad , \quad \delta\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} .$$

The hat matrix under the null model is

$$\boldsymbol{H}_0 = \boldsymbol{X}_0\left(\boldsymbol{X}_0{}^{\boldsymbol{T}}\boldsymbol{X}_0\right)^{-1}\boldsymbol{X}_0{}^{\boldsymbol{T}} = \frac{1}{6}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and

$$\delta\boldsymbol{X}\frac{\delta\boldsymbol{\beta}}{\sigma} = SNR\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & -1/2 \end{bmatrix} = \frac{SNR}{2}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} .$$

Then the non-centrality parameter is

$$\lambda = \min \text{diag}\left(\left(\delta \boldsymbol{X}\frac{\delta \boldsymbol{\beta}}{\sigma}\right)^{T}(\boldsymbol{I}-\boldsymbol{H}_0)\left(\delta \boldsymbol{X}\frac{\delta \boldsymbol{\beta}}{\sigma}\right)\right) = SNR^2 \min \text{diag}\left(\frac{1}{2}\begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}\right)$$
$$= SNR^2 .$$

The design includes 6 runs ($N = 6$), the model includes three parameters ($p = 3$), and two parameters are being tested ($\delta p = 2$). Hence, the critical F-value is:

$$f_U = P^{-1}[F(\delta p, N-p)](1-\alpha) = P^{-1}[F(2,3)](1-0.2) = 2.886 \ ,$$

and power for an effect with a signal-to-noise ratio of 1 is:

$$\mathcal{P} = P^{-1}[F(\delta p, N-p;\ \lambda)](f_U) = P^{-1}[F(2,3;1)](2.886) = 0.294.$$

Appendix E
JMP Monte Carlo Simulation Script

```
dt = Current Data Table();
nsim = 100;
p = J( nsim, 4, 0 );
For( sim = 1, sim <= nsim, sim++,
     Column( 5 ) << Eval Formula;

     glm = Fit Model(
          Y( :Y ),
          Effects( :X1, :X2, :X3 ),
          Personality( Generalized Linear Model ),
          GLM Distribution( Binomial ),
          Link Function( Logit ),
          Overdispersion Tests and Intervals( 0 ),
          Name( "Firth Bias-adjusted Estimates" )(1),
          Run
     );
     rpt = Report( glm );
     pValues = Report( glm )[Outline Box( "Parameter
Estimates" )][Number Col Box( 4 )
     ] << get as matrix;
     p[sim, 0] = pValues`;
     rpt << Close Window;
);

power=j(1,4,0);
for (i=1,i<=4,i++,
     ps = Sort Ascending(p[0,i]);
     power[1,i] = min(loc(ps>0.05))/nsim;
);
as table(power);
Column(1)<<set name("Intercept Power");
Column(2)<<set name("X1 Power");
Column(3)<<set name("X2 Power");
Column(4)<<set name("X3 Power");
```

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-11-2014	Final	— — —

4. TITLE AND SUBTITLE

Power Analysis Tutorial for Experimental Design Software

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER
— — —

5c. PROGRAM ELEMENT NUMBER
— — —

6. AUTHOR(S)

Freeman, Laura J.; Johnson, Thomas H.; Simpson, James R.

5d. PROJECT NUMBER
— — —

5e. TASK NUMBER
BD-9-229990

5f. WORK UNIT NUMBER
— — —

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER
D-5205
H14-000639

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Director, Operational Test and Evaluation
The Pentagon
1700 Defense
Washington, D.C. 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
— — —

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited. Office of the Director, Operational Test and Evaluation, 12 December 2014.

13. SUPPLEMENTARY NOTES
Project Leader: Freeman, Laura J.

14. ABSTRACT

Statistical power calculations for designed experiments are essential to right-sizing tests during planning. Under-sized tests will fail to uncover true contributors affecting system effectiveness and suitability, while over-sized tests are wasteful. Although the concepts of statistical power are reasonably well understood, the mechanics of computations are not necessarily well publicized. The statistical software packages are not necessarily consistent in requesting user information, nor are they clear or consistent in the assumptions made for the necessary power information not requested. Most likely the least understood concept and the one software companies fail to agree upon is the method for sizing effects for categorical factors with more than two levels. These effects are critical ingredients to the power equation. This document reviews basic statistical power concepts as they relate to the design of experiments, discuss the differences between and the proper steps for continuous response variable power versus binary response power, describe the power formulation intricacies for designs involving multi-level categorical factors, and finally to compare software platform interfaces and power computation differences. The intent is to make you aware of the differences in power estimates across software packages, but even more importantly to equip you to confidently and successfully estimate power for your testing.

15. SUBJECT TERMS
JMP Software, Design Expert Software, Statistical Power, Experimental Design, Hypothesis Testing, Test Size

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			— — —
Unclassified	Unclassified	Unclassified	Unlimited	116	**19b. TELEPHONE NUMBER** *(include area code)* — — —