Design Institute for Six Sigma

The Power to Know.

A SAS White Paper

Table of Contents

An *experiment* is a process or study that results in the collection of data. The results of experiments are not known in advance. Usually, statistical experiments are conducted in situations in which researchers can manipulate the conditions of the experiment and can control the factors that are irrelevant to the research objectives. For example, a rental car company compares the tread wear of four brands of tires, while also controlling for the type of car, speed, road surface, weather, and driver.

Experimental design is the process of planning a study to meet specified objectives. Planning an experiment properly is very important in order to ensure that the right type of data and a sufficient sample size and power are available to answer the research questions of interest as clearly and efficiently as possible.

Six Sigma is a philosophy that teaches methodologies and techniques that provide the framework to define a business strategy that focuses on eliminating mistakes, waste, and re-work. Six Sigma establishes a measurable status for achieving a strategic problem-solving methodology in order to increase customer satisfaction and dramatically enhance financial performance. For more information about the Design Institute for Six Sigma at SAS, see http://support.sas.com/training/us/css/.

Most Six Sigma training programs include some information about experimental design. However, the amount of training in these programs can vary from nothing about experimental design to one-week of instruction about this subject. The purpose of this paper is to summarize the basic concepts of traditional experimental design that would apply to a Six Sigma project. These basic concepts also apply to a general experimental setting. In addition, this paper shows how to apply some of these basic concepts by using examples of common experimental design and analysis.

This paper is written for people who have a basic understanding of experimental design.

This section discusses the basic concepts of experimental design, data collection, and data analysis. The following steps summarize the many decisions that need to be made at each stage of the planning process for the experiment. These steps are not independent, and it might be necessary to revise some earlier decisions that were made. A brief explanation of each step, which will help clarify the decisions that should be made during each stage, is given in the section that follows this list.

Perform the following steps when designing an experiment:

1. Define the problem and the questions to be addressed.
2. Define the population of interest.
3. Determine the need for sampling.
4. Define the experimental design.

Before data collection begins, specific questions that the researcher plans to examine must be clearly identified. In addition, a researcher should identify the sources of variability in the experimental conditions. One of the main goals of a designed experiment is to partition the effects of the sources of variability into distinct components in order to examine specific questions of interest. The objective of designed experiments is to improve the precision of the results in order to examine the research hypotheses.

A *population* is a collective whole of people, animals, plants, or other items that researchers collect data from. Before collecting any data, it is important that researchers clearly define the population, including a description of the members. The designed experiment should designate the population for which the problem will be examined. The entire population for which the researcher wants to draw conclusions will be the focus of the experiment.

A *sample* is one of many possible sub-sets of units that are selected from the population of interest. In many data collection studies, the population of interest is assumed to be much larger in size than the sample so, potentially, there are a very large (usually considered infinite) number of possible samples. The results from a sample are then used to draw valid inferences about the population.

A *random sample* is a sub-set of units that are selected randomly from a population. A random sample represents the general population or the conditions that are selected for the experiment because the population of interest is too large to study in its entirety. Using techniques such as random selection after stratification or blocking is often preferred.

An often-asked question about sampling is: How large should the sample be? Determining the *sample size* requires some knowledge of the observed or expected variance among sample members in addition to how large a difference among treatments you want to be able to detect.

Another way to describe this aspect of the design stage is to conduct a prospective *power analysis,* which is a brief statement about the capability of an analysis to detect a practical difference. A power analysis is essential so that the data collection plan will work to enhance the statistical tests primarily by reducing residual variation, which is one of the key components of a power analysis study.

A clear definition of the details of the experiment makes the desired statistical analyses possible, and almost always improves the usefulness of the results. The overall data collection and analysis plan considers how the experimental factors, both controlled and uncontrolled, fit together into a model that will meet the specific objectives of the experiment and satisfy the practical constraints of time and money.

The data collection and analysis plan provides the maximum amount of information that is relevant to a problem by using the available resources most efficiently. Understanding how the relevant variables fit into the design structure indicates whether the appropriate data will be collected in a way that permits an objective analysis that leads to valid inferences with respect to the stated problem. The desired result is to produce a layout of the design along with an explanation of its structure and the necessary statistical analyses.

The *data collection protocol* documents the details of the experiment such as the data definition, the structure of the design, the method of data collection, and the type of analyses to be applied to the data.

Defining the experimental design consists of the following steps:

1. Identify the experimental unit.

2. Identify the types of variables.

3. Define the treatment structure.

4. Define the design structure.

In our experience in the design and implementation of previous studies, often, a number of extenuating circumstances arise that require last minute adjustments to the data collection plan. Therefore, contingency plans should be available to keep the structure of the design intact in order to meet the stated objectives.

The following discussion provides further insight into the decisions that need to be made in each of the steps for defining the experimental design.

The first step in detailing the data collection protocol is to define the experimental unit. An *experimental or sampling unit* is the person or object that will be studied by the researcher. This is the smallest unit of analysis in the experiment from which data will be collected. For example, depending on the objectives, experimental or sampling units can be individual persons, students in a classroom, the classroom itself, an animal or a litter of animals, a plot of land, patients from a doctor's office, and so on.

A data collection plan considers how four important variables: *background*, *constant, uncontrollable*, and *primary*, fit into the study. Inconclusive results are likely to result if any of these classifications are not adequately defined. It is important to consider all the relevant variables (even those variables that might, at first, appear to be unnecessary) before the final data collection plan is approved in order to maximize confidence in the final results.

Background variables can be identified and measured yet cannot be controlled; they will influence the outcome of an experiment. Background variables will be treated as covariates in the model rather than primary variables. *Primary variables* are the variables of interest to the researcher. When background variables are used in an analysis, better estimates of the primary variables should result because the sources of variation that are supplied by the covariates have been removed. Occasionally, primary variables must be treated as covariates in order to keep the size of the experiment to a manageable level. Detailed measurements of all relevant variables should be made, preferably at the time the actual measurements are collected.

Constant variables can be controlled or measured but, for some reason, will be held constant over the duration of the study. This action increases the validity of the results by reducing extraneous sources of variation from entering the data. For this data collection plan, some of the variables that will be held constant include:

- the use of standard operating procedures
- the use of one operator for each measuring device
- all measurements taken at specific times and locations

The standard operating procedure of a measuring device should be used in the configuration and manner in which the developer and the technical representative consider most appropriate. Operator error might also add to the variability of the results. To reduce this source of variation, one operator is often used for each measuring device, or specific training is given with the intent of achieving uniform results.

Uncontrollable (Hard-to-Change) variables are those variables that are known to exist, but conditions prevent them from being manipulated, or it is very difficult (due to cost or physical constraints) to measure them. The experimental error is due to the influential effects of uncontrollable variables, which will result in less precise evaluations of the effects of the primary and background variables. The design of the experiment should eliminate or control these types of variables as much as possible in order to increase confidence in the final results.

Primary variables are independent variables that are possible sources of variation in the response. These variables comprise the treatment and design structures and are referred to as *factors*. Primary variables are referred to as *factors* in the rest of this paper.

The treatment structure consists of factors that the researcher wants to study and about which the researcher will make inferences. The primary factors are controlled by the researcher and are expected to show the effects of greatest interest on the response variable(s). For this reason, they are called *primary factors*.

The levels of greatest interest should be clearly defined for each primary factor. The levels of the primary factors represent the range of the inference space relative to this study. The levels of the primary factors can represent the entire range of possibilities or a random sub-set. It is also important to recognize and define when combinations of levels of two or more treatment factors are illogical or unlikely to exist.

Factorial designs vary several factors simultaneously within a single experiment, with or without replication. One-way, two-way, three-way, 2n, 3n, D-optimal, central composite, and two-way with some controls are examples of treatment structures that are used to define how data are collected. The treatment structure relates to the objectives of the experiment and the type of data that's available.

Drawing a design template is a good way to view the structure of the design factors. Understanding the layout of the design through the visual representation of its primary factors will greatly help you later to construct an appropriate statistical model.

Fixed effects treatment factors are usually considered to be "fixed" in the sense that all levels of interest are included in the study because they are selected by some non-random process, they consist of the whole population of possible levels, or other levels were not feasible to consider as part of the study. The fixed effects represent the levels of a set of precise hypotheses of interest in the research. A fixed factor can have only a small number of inherent levels; for example, the only relevant levels for gender are male and female. A factor should also be considered fixed when only certain values of it are of interest, even though other levels might exist (types of evaluation tests given to students). Treatment factors can also be considered "fixed" as opposed to "random" because they are the only levels about which you would want to make inferences.

Because of resource limitations or missing data, all combinations of treatment factors might not be present. This is known as the *missing* or *empty cell problem*.

Certain designs, known as *fractional factorials,* are designed to enable you to study a large number of factors with a relatively small number of observations. Analyses of such data assume that specific interactions among factors are negligible.

If a treatment condition appears more than one time, it is defined to be replicated. *True replication* refers to responses that are treated in the same way. Misconceptions about the number of replications have often occurred in experiments where sub-samples or repeated observations on a unit have been mistaken as additional experimental units.

Notice that replication does not refer to the initial similarity among experimental units; the important issue is not the similarity of the experimental units but how many will be needed per treatment, given their current differences and the differences to be expected as the experiment proceeds.

Replication is essential for estimating experimental error. The type of replication that's possible for a data collection plan represents how the error terms should be estimated. Two or more measurements should be taken from each experimental unit at each combination of conditions. In addition, it is desirable to have measurements taken at a later period in order to test for repeatability over time. The first method of replication gives an estimate of pure error, that is, the ability of the experimental units to provide similar results under identical experimental conditions. The second method provides an estimate of how closely the devices can reproduce measurements over time.

Most experimental designs require experimental units to be allocated to treatments either randomly or randomly with constraints, as in blocked designs (Montgomery 1997).

Blocks are groups of experimental units that are formed to be as homogeneous as possible with respect to the block characteristics. The term *block* comes from the agricultural heritage of experimental design where a large block of land was selected for the various treatments, that had uniform soil, drainage, sunlight, and other important physical characteristics. Homogeneous clusters improve the comparison of treatments by randomly allocating levels of the treatments within each block.

The design structure consists of those factors that define the blocking of the experimental units into clusters. The types of commonly used design structures are described next.

Completely Randomized Design. Subjects are assigned to treatments completely at random. For example, in an education study, students from several classrooms are randomly assigned to one of four treatment groups (three new types of a test and the standard). The total number of students in 4 classrooms is 96. Randomly assign 1/4 of them, or 24 students, to each of the 4 types of tests.

Note: Numbers that are evenly divisible by 4 are used here; equal sample size in every cell, although desirable, is not absolutely necessary.

Test Method			
Standard	New Test 1	New Test 2	New Test 3
24 students	24 students	24 students	24 students

Randomized Complete Block Design. Subjects are divided into *b* blocks (see description of blocks above) according to demographic characteristics. Subjects in each block are then randomly assigned to treatments so that all treatment levels appear in each block. For example, in the education study that involved several classrooms, the classrooms might differ due to different teaching methods. Students are randomly assigned to one of the four types of tests within each classroom. There might be significant variability between the subjects in each classroom, each of which contains 24 students. Randomly assign 6 students to each of the three types of test s and the standard. The classroom is now the 'block'. The primary interest is in the main effect of the test.

Class-room	Test Method			
	Standard	Test 1	Test 2	Test 3
1	6 students	6 students	6 students	6 students
2	6 students	6 students	6 students	6 students
3	6 students	6 students	6 students	6 students
4	6 student	6 students	6 students	6 students

The improvement of this design over a completely randomized design enables you to make comparisons among treatments after removing the effects of a confounding variable, in this case, different classrooms.

It is important to follow the data collection protocol exactly as it is written when the data are collected. Prior to collecting the data, it is important to double check that all the instruments are valid, reliable, and calibrated. After that is confirmed, take time to explain the data collection procedures to the person who will be doing the actual data collection. It might seem counter-intuitive to technicians and machine operators to execute a randomized design. They might re-organize the data collection scheme in an effort to be more efficient, without realizing the impact that it might have on the experiment.

The basis for the analysis of the data from designed experiments is discussed in this section. There are many thousands of experimental designs. Each design can be analyzed by using a specific analysis of variance (ANOVA) that is designed for that experimental design. One of the jobs of a statistician is to recognize the various experimental designs, and to help clients create the design and analyze the experiments by using appropriate methods and software. The examples at the end of this paper will generate the experimental design and respective analysis by using JMP software.

An *effect* is a change in the response due to a change in a factor level. There are different types of effects. One objective of an experiment is to determine if there are significant differences in the responses across levels of a treatment (a fixed effect) or any interaction between the treatment levels. If this is always the case, the analysis is usually easily manageable, given that the anomalies in the data are minimal (outliers, missing data, homogeneous variances, unbalanced sample sizes, and so on).

A *random effect* exists when the levels that are chosen represent a random selection from a much larger population of equally usable levels. This is often thought of as a sample of interchangeable individuals or conditions. The chosen levels represent arbitrary realizations from a much larger set of other equally acceptable levels.

Elements of the design structure (for example, the blocks) are usually treated as random effects. Blocks are a sub-set of the larger set of blocks over which inference is to be made. It is helpful to assume that there is no interaction among elements of the design structure and elements of the treatment structure if blocks are considered a fixed effect. If blocks are treated as random effects, you can determine interaction among elements of treatment structure and design structure.

The number of potential human subjects that are available is often very large compared to the actual number of subjects that are used. Subjects who are chosen are likely to be just as reasonable to collect data from as potential subjects who were not chosen, and inferences for how individual subjects respond is usually not of primary importance, whereas a measure of the variation in responses is important to know.

One additional consideration that is essential in the evaluation of the treatment and design structure with two or more treatment/design factors is to differentiate whether the levels of the factors are either *crossed* or *nested* with each other.

Two factors that are crossed with one another means that all levels of the first factor appear in combination with all levels of the second factor, which produces all possible combinations. For example, in an education program, male and female students receive the same educational tests, thus, gender is crossed with test.

One factor that is nested in a second factor implies a hierarchy. This means that a given level of the nested factor appears in one level of the nesting factor. For example, in a study of educational programs, teachers are usually nested within schools because, usually, teachers teach only at one school.

Data from experiments are analyzed using linear regression and analysis of variance. The standard assumptions for data analysis that apply to linear regression and the analysis of variance are now summarized. To the degree that they are not satisfied implies that the results are merely numbers on an output listing. However, when the assumptions are satisfied, the estimates from sample data inform us of the structure of relationships in the real world. The following list briefly describes the assumptions behind linear regression:

I. No model specification error

- The response Y is the dependent variable.
- The independent variables, x1,..,xp, influence Y.
- The form of the relationship between Y and (x1,..,xp) is linear (not nonlinear) in the parameters.

II. No measurement error

- The dependent variable(s) are interval or ratio data (not ordinal or nominal).
- The independent variables are measured accurately.

III. No collinearity (a high correlation between any two independent variables is not allowed).

IV. The error term, residuals, are well-behaved when the following conditions hold:

- a zero mean
- homoscedasticity
- no autocorrelation (usually of most concern with time series or spatial data)
- no 'large' correlations between any of the independent variables
- normal distribution

When any of these assumptions are not met, then statistical significance tests lose their meaning. What if some of them are not met? In practice, assumptions from I, II, and III are usually reasonably satisfied. It is the behavior of the errors in IV that can cause the most severe problems with estimation and hypothesis testing.

The majority of these assumptions also apply to the analysis of variance because ANOVA is nothing more than a special case of regression.

Although a few of the assumptions can be relaxed (such as measurement error, which usually does not apply with categorical data), you should not categorize continuous data because the process of doing this creates a source of measurement error in the model that is not necessary just use regression. Additional assumptions or desirable conditions for data analysis with analysis of variance include:

- Homogeneous variances across levels of data categories.
- Elements of the design structure are random effects (blocks).

- Nearly equal sample sizes with outliers absent and missing data can be considered missing at random (MAR).
- No interaction exists among elements of the design structure and elements of the treatment structure.

The *inference space* of experimental results defines the range of conditions for which the subjects form a representative sample, and are assumed to be randomly selected from this well-defined population. The first question to ask before any experimental design is administered is, "What population do I want to apply the final results to?" By strict definition, the results of the experiment are only applicable to the population that the experimental units were selected from in the beginning. This means that the relevant range of the independent variables that are covered by the experiment is limited by the types of experimental units in the sample. Relaxing this assumption requires considerable caution because it assumes that the model remains valid outside the inference space, which in many cases might not be true.

Factors that are defined as fixed are limited to those factors that are specified in the experimental design. Therefore, extrapolation to other values, if any, is not justified. Inferences from a study should be confined to the type and range of independent variables that are included in the design.

Don't allow generalizations that are made for a group to extend far beyond the legitimate scope of the research; that is, the sampling method that's used is not consistent with the larger group to which inferences are made (for example, generalizing to high school students when, in fact, only 6th graders were studied).

JMP software provides the capabilities to design experiments and perform the statistical analyses. JMP provides capabilities to create screening, mixed-level, response surface, mixture, Taguchi, and custom designs. This section of the paper demonstrates the two designs that are discussed in the design structure portion of this paper and the respective analyses. These designs are among the most commonly used designs. The first design is a completely randomized design that begins with a power analysis. The second design is a randomized complete block design. The examples come from Montgomery (1997).

A company uses drill tips to determine the hardness of the metal plates it produces. There are different types of drill tips that the company can use, and the company wants to determine whether the different types of tips result in similar hardness measurements.

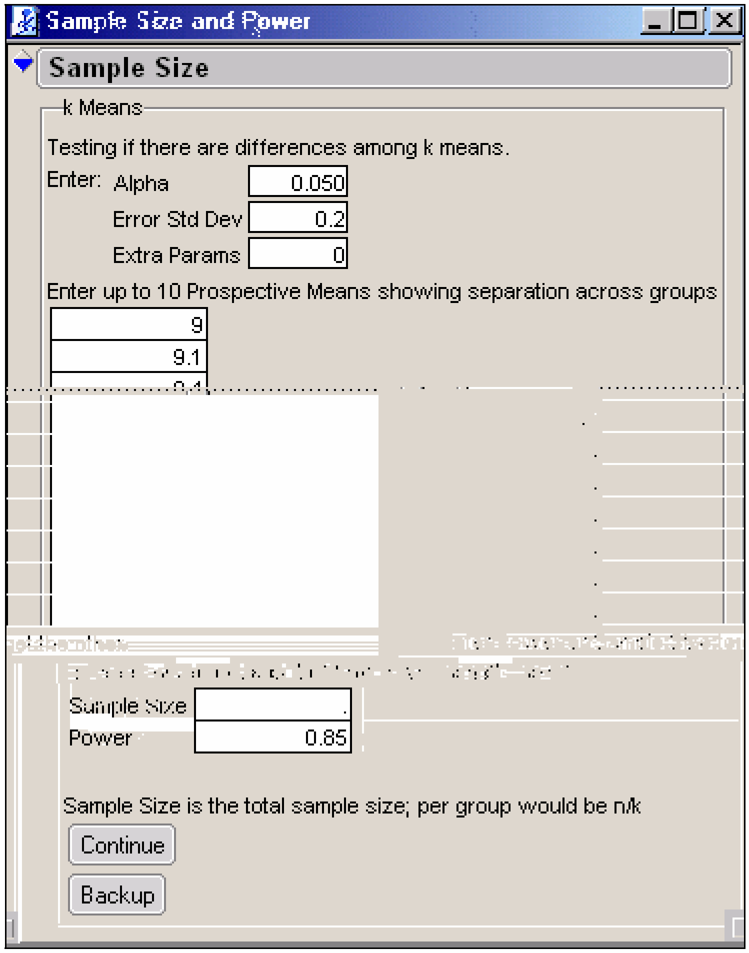
There are four types of drill tips which are made of different alloys. To test the hardness of the metal sheet, a drill press is used to press the drill tip into the metal sheet, causing an indentation. The resulting depression is measured, and this measurement translates into the hardness of the metal sheet. The purpose of the experiment is to determine whether these measurements differ because of the type of drill tip used.

In this experiment, the metal sheet is divided into four sections (quadrants). One measurement is taken from each quadrant. In this experiment, the experimental unit is the quadrant of the metal sheet.

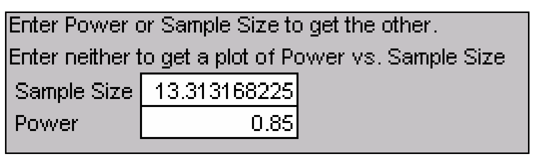
This experiment has one factor tip type, with four levels. The experimental unit is the quadrant of the metal sheet, with one tip type assigned to each quadrant. As the researcher, you have decided to use metal sheets from the same lot and assume that the sheets and the different quadrants of the sheets are homogeneous. Based on this, the most appropriate design is the completely randomized design. In a completely randomized design, treatments are randomly assigned to the experimental units.

In preparation for your experiment, you have consulted with industry experts and reviewed previous experiments on drill tips. Based on this initial research, you expect that the average hardness measurements will be between 9 and 9.6. The best estimate for the Error Std Dev is the Root Mean Square Error that you get from previous experiments or knowledge. In this case, you expect the Error Std Dev to be 0.2. Based on business needs, you have decided on an alpha equal to 0.05. You also expect tip types 1 and 2 to have means closer to each other and lower than the other two tip types. The actual means that you decide to use are 9, 9.1, 9.4, and 9.6. Finally, you would like to have power of at least 85%. Based on this information, you must determine the necessary sample size for your experiment.

1. Open JMP from your desktop by selecting **Start** ⇨ **Programs** ⇨ **JMP** ⇨ **JMP 5.1**. When JMP opens you will see menus across the top.
2. Click the **DOE** menu and select **Sample Size and Power**.
3. In the **Sample Size and Power** window, click **k Means**.
4. In the **Error Std Dev** field, type **0.2**.
5. In the **Prospective Means** fields, type **9**, **9.1**, **9.4**, and **9.6**.
6. In the **Power** field, type **0.85**.



7. Click **Continue**.

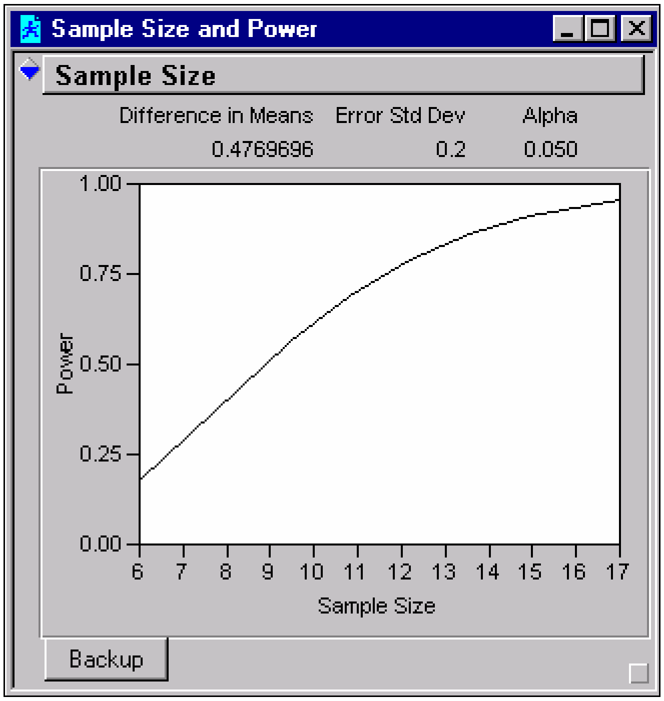


The necessary sample size is just over 13 observations. This is the total sample size. Because you have 4 groups and the power analysis presumes equal sample sizes in all groups, this would suggest a total sample size of 16, or 4 observations per treatment.

If you are interested in seeing how a change in power might affect the sample size requirements, you can leave both the **Sample Size** and **Power** fields blank and obtain a plot of power versus sample size.

8. Clear the **Sample Size** and **Power** fields by highlighting the entries and pressing **Delete** on the keyboard.

9. Click **Continue**.



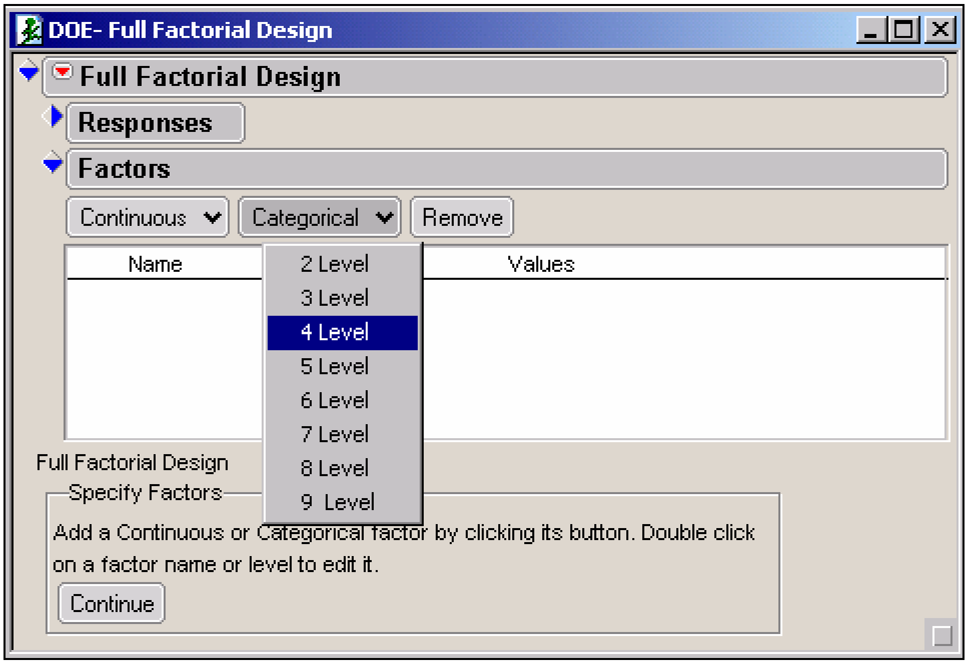
As you can see from the resulting plot, as power increases, the required sample size increases and vice versa. You can use the crosshairs tool to explore the plot further. Using the crosshairs tool, you determine that a sample size of 16 should result in power close to 94%.

10. Close the **Sample Size and Power** window.

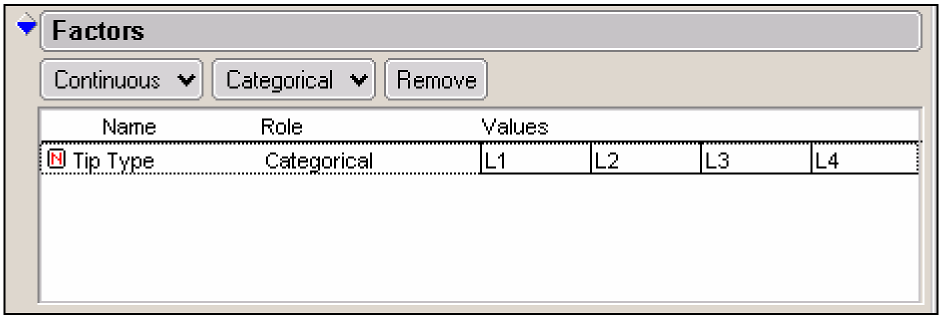
After the preliminary work has been done (including the determination of the sample size) and the appropriate design is selected, you can generate the design including the randomization.

1. Click the **DOE** menu and select **Full Factorial Design**.

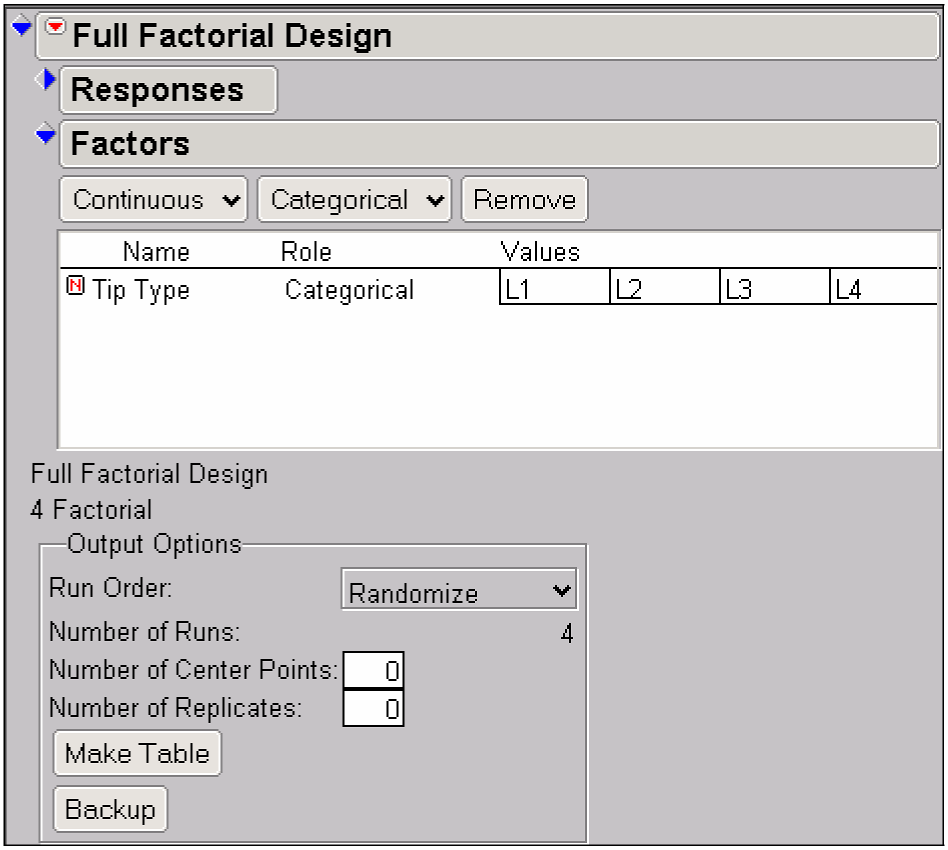
2. Under **Factors**, click **Categorical** ⇨ **4 Level**.



3. The default name for the factor is **X1**. To change this, double-click **X1** and change the name to **Tip Type**. This is the only factor for this experiment.



4. Click **Continue**.

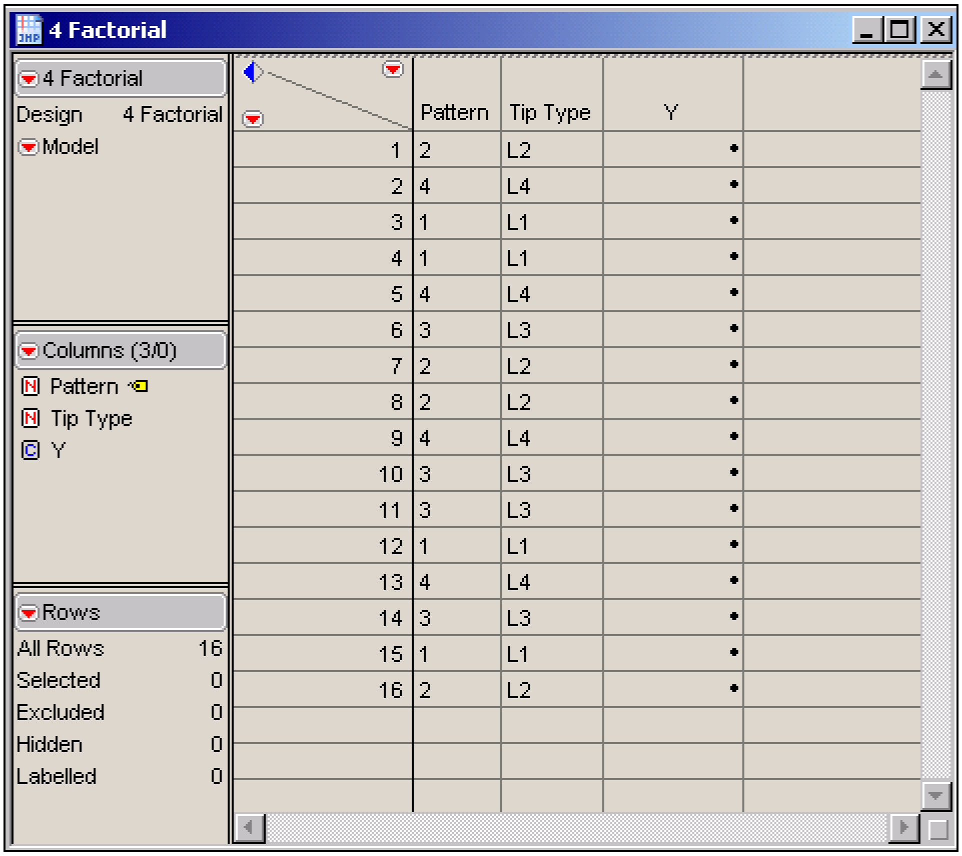


Examine the **Output Options**. A *run* is the collection of one value of the response variable, so it is one observation. The default is that the run order will be randomized. This essentially randomizes the assignment of the treatments to the experimental units, presuming that you have the experimental units in a set order before generating the table.

Also, notice that the default number of runs is 4. Therefore, you would have one observation per treatment. Recall that it was determined that this experiment would consist of 16 observations or runs. In order to obtain the desired number of runs, change the number of replicates to 3. This results in the original 4 observations plus 3 more sets of 4 observations, for a total of 16 runs.

5. Click in the **Number of Replicates** field and change the value to **3**.

6. Click **Make Table**.



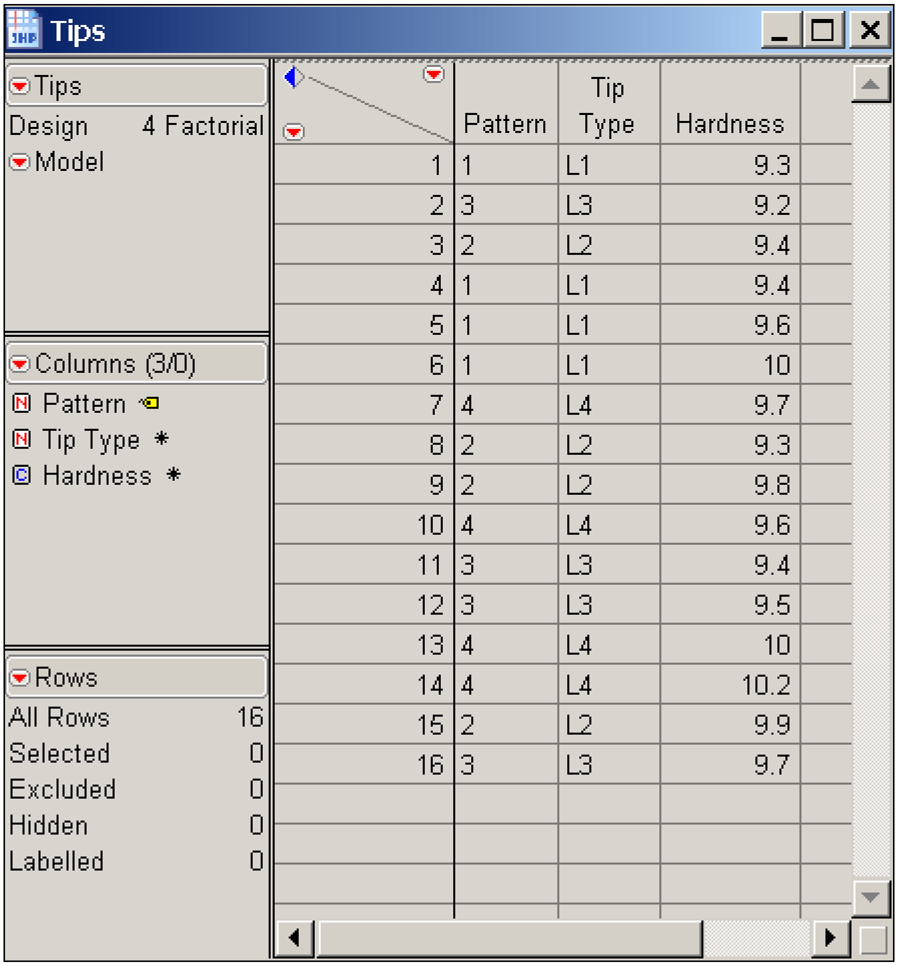
	Pattern	Tip Type	Y
1	2	L2	•
2	4	L4	•
3	1	L1	•
4	1	L1	•
5	4	L4	•
6	3	L3	•
7	2	L2	•
8	2	L2	•
9	4	L4	•
10	3	L3	•
11	3	L3	•
12	1	L1	•
13	4	L4	•
14	3	L3	•
15	1	L1	•
16	2	L2	•

The resulting table has three columns. The first column is **Pattern** that, for more than one factor analysis, would contain the combination of factors or treatments that are used for a specific observation. For this experiment, the second column is the factor **Tip Type**. Lastly, a column labeled **Y** is included to allow you to enter the response data as you collect it.

After the design has been generated, the data can be collected and analyzed.

7. Close the **4 Factorial** window.

The experiment was conducted and the data table is shown next.



	Pattern	Tip Type	Hardness
1	1	L1	9.3
2	3	L3	9.2
3	2	L2	9.4
4	1	L1	9.4
5	1	L1	9.6
6	1	L1	10
7	4	L4	9.7
8	2	L2	9.3
9	2	L2	9.8
10	4	L4	9.6
11	3	L3	9.4
12	3	L3	9.5
13	4	L4	10
14	4	L4	10.2
15	2	L2	9.9
16	3	L3	9.7

The data set is the original design with the response values added and the response column **Y** is re-named **Hardness**.

Conduct an analysis of variance to determine whether there are differences between the types of drill tips.

1. Click the **Analyze** menu and select **Fit Y by X**.

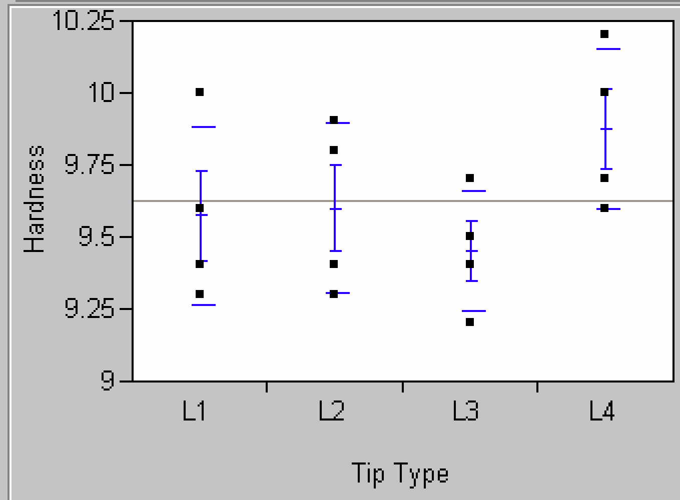
2. Click **Hardness** ⇨ **Y**, **Response**.

3. Click **Tip Type** ⇨ **X, Factor**.

4. Click **OK**.

5. Click the red triangle at the left of **Oneway Analysis of Hardness By Tip Type** and select **Means and Std Deviations**.

Oneway Analysis of Hardness By Tip Type



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
L1	4	9.57500	0.309570	0.15478	9.0824	10.068
L2	4	9.60000	0.294392	0.14720	9.1316	10.068
L3	4	9.45000	0.208167	0.10408	9.1188	9.781
L4	4	9.87500	0.275379	0.13769	9.4368	10.313

The means for the different types of tips are fairly close to each other. In fact, they are closer than anticipated prior to the experiment. The standard deviations are also fairly close to each other, but they are larger than anticipated prior to the study. The confidence intervals for the mean overlap each other, which is an indication that there are no significant differences between the tip types.

6. To generate the ANOVA, click the red triangle next to **Oneway Analysis of Hardness By Tip Type** and select **Means/Anova/t Test**.

Oneway Anova

Summary of Fit

Rsquare	0.29845
Adj Rsquare	0.123062
Root Mean Square Error	0.274621
Mean of Response	9.625
Observations (or Sum Wgts)	16

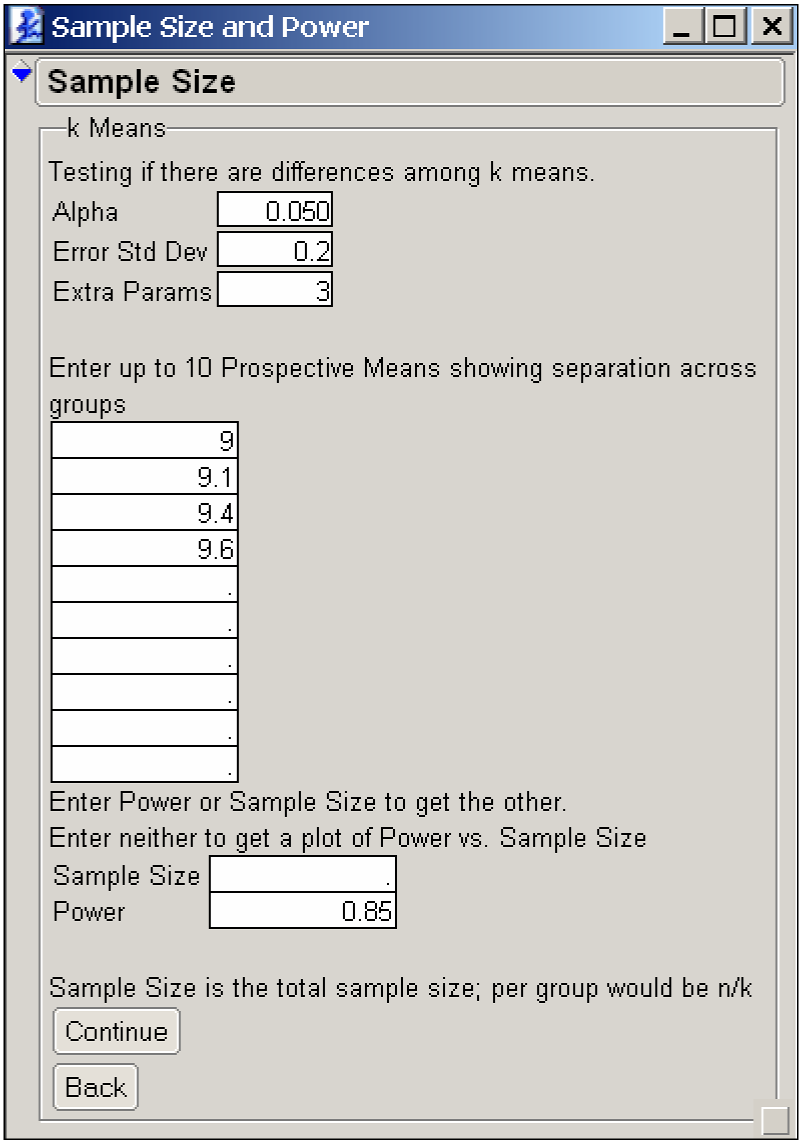
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Tip Type	3	0.3850000	0.128333	1.7017	0.2196
Error	12	0.9050000	0.075417		
C. Total	15	1.2900000			

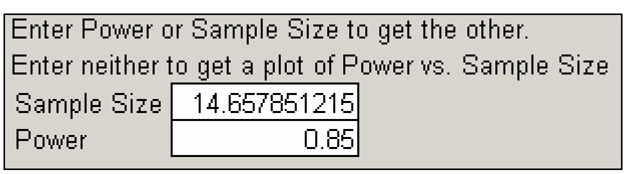
The *p*-value for the analysis of variance is 0.2196. This is larger than the stated alpha of 0.05.

average hardness measurements will be between 9 and 9.6. You also expect the Root Mean Square Error to be approximately 0.2. This value is used for the **Error Std Dev** field in the **Sample Size and Power** window. Based on business needs, you have decided on an alpha equal to 0.05. You also expect tip types 1 and 2 to have means closer to each other and lower than the other two tip types. The actual means that you decide to use are 9, 9.1, 9.4, and 9.6. You would like to have power of at least 85%. Finally, there is a blocking factor. Because there are 4 observations per metal sheet and the experiment has 4 treatments, there will be 4 blocks, each with 4 runs. To use the blocking information in the power analysis, use the **Extra Params** field. For a randomized complete block design, set **Extra Params** equal to the number of blocks minus 1. In this case, there are 4 blocks; 4 blocks minus 1= 3. Based on this information, you must determine the necessary sample size for your experiment.

1. In JMP, click the **DOE** menu and select **Sample Size and Power**.
2. In the **Sample Size and Power** window, click **Sample Size**.
3. In the **Error Std Dev** field, type **0.2**.
4. In the **Extra Params** field, type **3**.
5. In the **Prospective Means** fields, type **9**, **9.1**, **9.4**, and **9.6**.
6. In the **Power** field, type **0.85**.



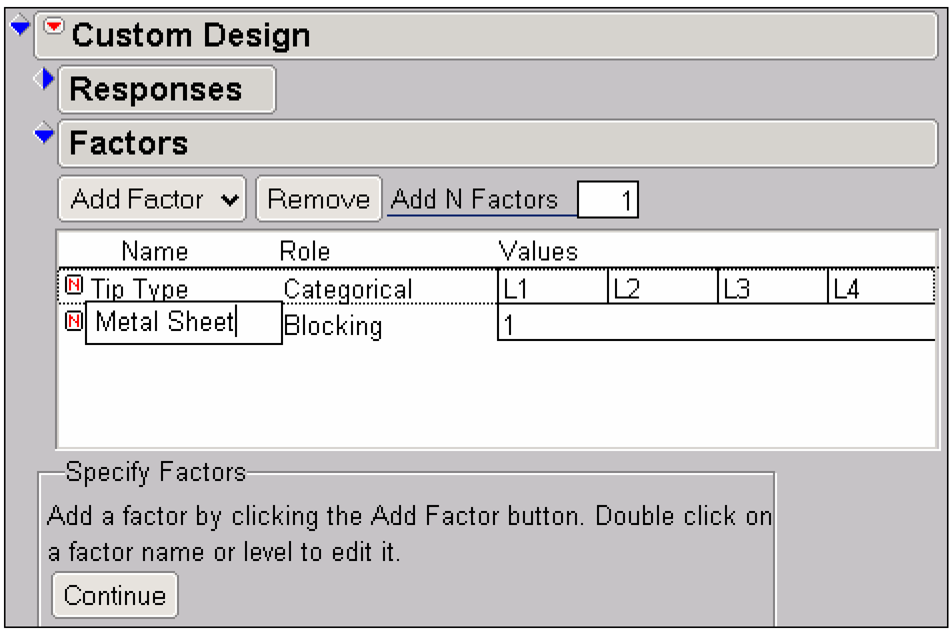
7. Click **Continue**.



The necessary sample size is just over 14 observations. This is the total sample size. Because you have 4 groups and 4 blocks and the power analysis presumes equal sample sizes in all groups, this would suggest a total sample size of 16, or 4 observations per treatment and block combination.

You want to determine whether different types of drill tips result in different hardness readings. As the subject matter expert, you have determined that there will be variability due to the metal sheets, therefore, you have decided to use the metal sheet as the blocking factor. Use the **Custom Design** platform because it enables you to specify the role of the second factor as blocking and to randomize the runs within the blocks. Generate a design for this experiment:

1. In JMP, click the **DOE** menu and select **Custom Design**.
2. In the **DOE** window under **Factors**, click **Add Factor** ⇨ **Categorical** ⇨ **4 Level**. This adds the factor **X1** to the design.
3. To change the name of the factor, double-click **X1** and type **Tip Type**.
4. Click **Add Factor** ⇨ **Blocking** ⇨ **4 runs per block**.
5. To change the name of the blocking factor, double-click **X2** and type **Metal Sheet**.



6. Click **Continue**.
7. Under **Design Generation**, notice that the default design will have 16 runs, which is acceptable.
8. Click **Make Design**.

Under **Output Options** in the now expanded **Custom Design** window, notice that the default for the run order is to randomize within blocks.

9. Under **Output Options**, click **Make Table**.

Custom Design
Design Custom Design
Model

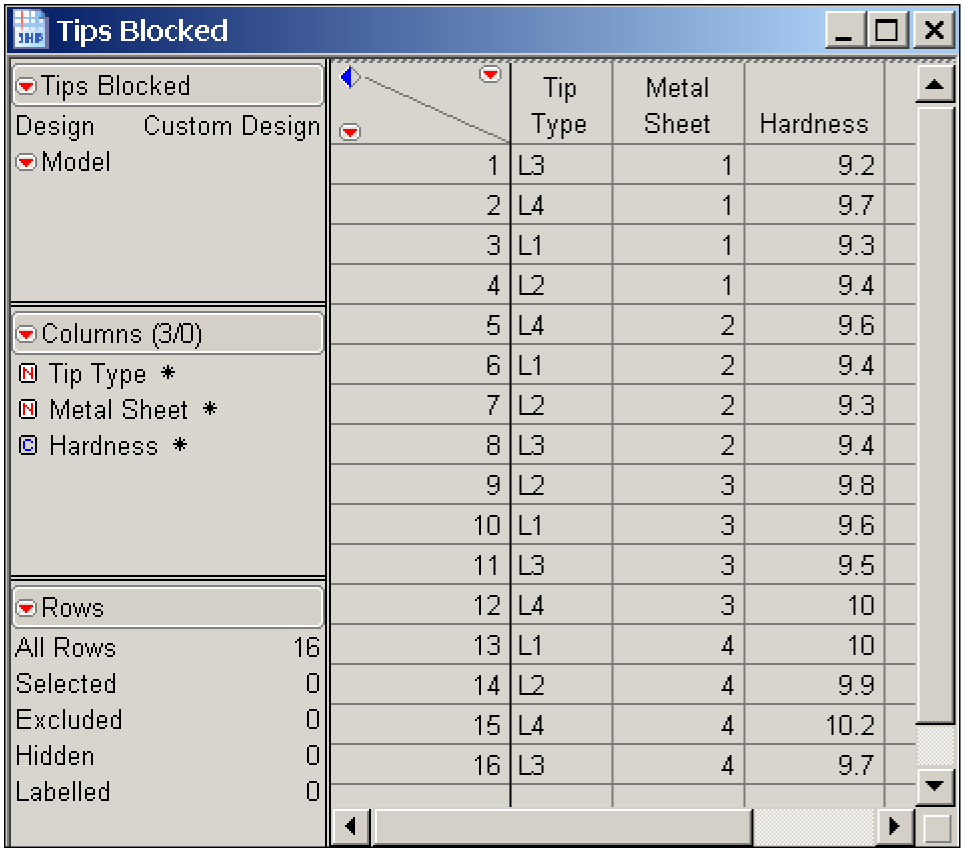
Columns (3/0)
Tip Type
Metal Sheet
Y

Rows
All Rows 16
Selected 0
Excluded 0
Hidden 0
Labelled 0

	Tip Type	Metal Sheet	Y
1	L3	1	•
2	L4	1	•
3	L1	1	•
4	L2	1	•
5	L4	2	•
6	L1	2	•
7	L2	2	•
8	L3	2	•
9	L2	3	•
10	L1	3	•
11	L3	3	•
12	L4	3	•
13	L1	4	•
14	L2	4	•
15	L4	4	•
16	L3	4	•

This design includes 4 blocks and, as expected, each treatment appears once in each block. The order of the treatments within each block has been randomized.

The experiment has been conducted and the data table is shown in the next window.



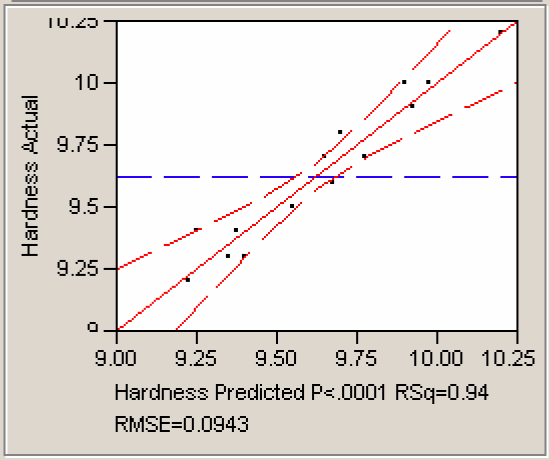
	Tip Type	Metal Sheet	Hardness
1	L3	1	9.2
2	L4	1	9.7
3	L1	1	9.3
4	L2	1	9.4
5	L4	2	9.6
6	L1	2	9.4
7	L2	2	9.3
8	L3	2	9.4
9	L2	3	9.8
10	L1	3	9.6
11	L3	3	9.5
12	L4	3	10
13	L1	4	10
14	L2	4	9.9
15	L4	4	10.2
16	L3	4	9.7

This data set contains the original design with the response values added and the response column re-named as **Hardness**.

The analysis for a randomized complete block design is the same as for a completely randomized design, except that the blocking factor is included as an independent variable in the model.

1. In JMP, click the **Analyze** menu and select **Fit Model**.
2. In the **Fit Model** window, the variables are already assigned to the appropriate roles because the original data table was created under the DOE platform.
3. Click **Run Model** in the **Fit Model** window.

Actual by Predicted Plot



Summary of Fit

RSquare	0.937984
RSquare Adj	0.896641
Root Mean Square Error	0.094281
Mean of Response	9.625
Observations (or Sum Wgts)	16

In the **Actual by Predicted Plot**, the distance from a point to the

- horizontal dashed line represents the residual to the baseline model.
- diagonal solid line represents the residual to the fitted model.

Because the confidence curves cross the horizontal dashed line, this implies that the model is useful in explaining the variation in the data.

The **Summary of Fit** report shows that both the R^2 and adjusted R^2 are high. The average hardness measurement is 9.625 for these 16 observations.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	1.2100000	0.201667	22.6875
Error	9	0.0800000	0.008889	Prob > F
C. Total	15	1.2900000		<.0001

The **Analysis of Variance** report confirms the conclusion that's shown in the plot. The *p*-value is less than 0.0001. Presuming an alpha of 0.05, the *p*-value is less than alpha. Therefore, at least one of the factors **Tip Type** or **Metal Sheet** in the model is explaining a significant amount of the variability in the response **Hardness**.

Notice that the degrees of freedom (DF) for the model is 6. There are four levels of the factor **Tip Type**, so there are 3 degrees of freedom. In addition, there are 4 blocks (metal sheets), so there are 3 degrees of freedom. Because you presume there is no interaction between the factor of interest and the blocking factor, the degrees of freedom for the model is the sum of the degrees of freedom for the two factors.

Examine the **Effect Tests** table to determine whether **Tip Type** is significant.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Tip Type	3	3	0.38500000	14.4375	0.0009
Metal Sheet	3	3	0.82500000	30.9375	<.0001

The *p*-value for **Tip Type** is 0.0009. Therefore, there are significant differences between the average hardness measurements from the different types of drill tips. The F Ratio for **Metal Sheet** is larger than 1. This indicates that including the blocking factor in the model was beneficial to the analysis.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.625	0.02357	408.35	<.0001
Tip Type[L1]	-0.05	0.040825	-1.22	0.2518
Tip Type[L2]	-0.025	0.040825	-0.61	0.5554
Tip Type[L3]	-0.175	0.040825	-4.29	0.0020
Metal Sheet[1]	-0.225	0.040825	-5.51	0.0004
Metal Sheet[2]	-0.2	0.040825	-4.90	0.0008
Metal Sheet[3]	0.1	0.040825	2.45	0.0368

The **Parameter Estimates** table shows the estimate for each term, its standard error, and a *t* Ratio and *p*-value to test whether the true parameter is 0.

The intercept term is the overall mean for all observations. The parameter estimates for **Tip Type** are the difference between the mean for that level of **Tip Type** and the overall mean. Therefore, the *t* Ratios are testing whether the mean for that type of drill tip is significantly different from the average for all drill tips. Notice that only the first three levels of **Tip Type** are listed in the table. The estimate for the last level is the additive inverse of the sum of the estimates for the first three levels. No *t* Ratio is provided for the last level because the estimate is not independent. After the overall mean and the estimates for the first three levels are determined, the last level can take on only one value.

The estimates for any factor in an ANOVA model are computed with this method. Each estimate is the difference between the overall mean and that specific level of the factor.

Scroll to the right to examine the least squares means for the different types of drill tips.

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
L1	9.5750000	0.04714045	9.57500
L2	9.6000000	0.04714045	9.60000
L3	9.4500000	0.04714045	9.45000
L4	9.8750000	0.04714045	9.87500

The **Least Squares Means Table** for **Tip Type** shows the mean and least squares mean for each factor level. In this case, the least squares means and the means are the same because there is an equal number of observations for each type of drill tip in every block. The standard error is also included in the table.

Because **Tip Type** is a significant effect, there are significant differences between the average hardness measurements from the different types of drill tips. To determine which levels are different from each other, you need to compare all pairs of means.

Recall that there are two types of multiple comparisons: those that control only the comparisonwise error rate and those that control the experimentwise error rate. Presume that you want to compare all pairs of means and control the experimentwise error rate. Tukey's HSD is the appropriate multiple comparison.

4. Click the red triangle at the left of **Tip Type** and select **LSMeans DifferencesTukey HSD**.

Tip Type

LSMeans Differences Tukey HSD

LSMean[j]

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	L1	L2	L3	L4
L1		0	-0.025	0.125	-0.3
		0	0.06667	0.06667	0.06667
		0	-0.2331	-0.0831	-0.5081
		0	0.18312	0.33312	-0.0919
L2		0.025	0	0.15	-0.275
		0.06667	0	0.06667	0.06667
		-0.1831	0	-0.0581	-0.4831
		0.23312	0	0.35812	-0.0669
L3		-0.125	-0.15	0	-0.425
		0.06667	0.06667	0	0.06667
		-0.3331	-0.3581	0	-0.6331
		0.08312	0.05812	0	-0.2169
L4		0.3	0.275	0.425	0
		0.06667	0.06667	0.06667	0
		0.09188	0.06688	0.21688	0
		0.50812	0.48312	0.63312	0

Level			Least Sq Mean
L4	A		9.8750000
L2		B	9.6000000
L1		B	9.5750000
L3		B	9.4500000

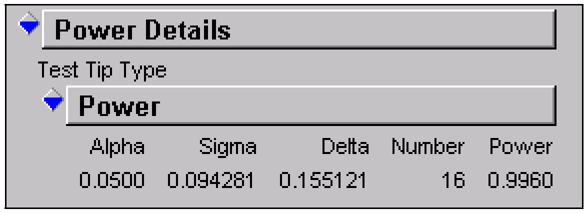
Levels not connected by same letter are significantly different

The table shows the difference between the two sample means, the standard error of the difference, and the 95% confidence interval for the difference. Those differences that are found to be statistically significant at the 5% level of significance are displayed in red. The table at the bottom can also be used. This experiment determines that tip type 4 is significantly different from all other tip types. The hardness measurements from tip type 4 are higher than those for all other types of drill tips.

5. Click the red triangle at the left of **Tip Type** and select **Power Analysis**.

6. In the **Power Details** dialog box, click the box next to **Solve for Power**.

7. Click **Done**.


Power Details
Test Tip Type
Power

Alpha	Sigma	Delta	Number	Power
0.0500	0.094281	0.155121	16	0.9960

The retrospective power for detecting true differences among the various types of drill tips in this experiment is very high.

Most Six Sigma training programs include some information about designed experiments, but the amount of training in these programs varies from nothing about designed experiments to one week of training about the subject. The purpose of this paper is to summarize the basic concepts of traditional experimental design that would apply to a Six Sigma project, and would also apply to a general experimental setting.

There are many thousands of experimental designs. Each design can be analyzed by using a specific ANOVA that is designed for that experimental design. One of the jobs of a Master Black Belt, a Black Belt, or a statistician is to recognize the various experimental designs and help clients create a design and analyze the experiments by using appropriate methods and software.

Montgomery, Douglas C. (1997). *Design and Analysis of Experiments* (4th edition). New York: John Wiley & Sons.

World Headquarters
and SAS Americas
SAS Campus Drive
Cary, NC 27513 USA
Tel: (919) 677 8000
Fax: (919) 677 4444
U.S. & Canada sales:
(800) 727 0025

SAS International
PO Box 10 53 40
Neuenheimer Landstr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160
Fax: (49) 6221 474850

www.sas.com

328767US.805