Success With DOE

Understanding and controlling variables results in usefulness of statistics.

Mark J. Anderson

When most people hear the word "statistics," their reaction is usually a mixture of fear, frustration, and annoyance, especially after enduring the typical college lecture on the subject. However, statistics prove useful, especially for design of experiments (DOE).

Statistics are a tool for extracting information from data. Imagine a technical colleague giving a report on an experiment. It wouldn't make sense for him to detail every measurement. Instead, he delivers a summary of the overall results. One question would be, "What's the average result?" Subsequent questions could focus on the quantity and variability of the results to build confidence in the data. Assuming that the experiment has a purpose, it's possible to accept or reject the findings. Statistics become not only a tool for summarization, but for calculating risks.

QUALITY TECH TIPS

- DOE targets systematic improvements in manufacturing, rather than eliminating specific causes of process problems.
- Normal distribution charts depend on enough data being collected to deliver the best possible average outcome.
- Experimenting with many factors and variables simultaneously is more efficient than experimenting with one factor at a time.
- Plotting experimental results puts extreme outcomes in perspective.

The 'x' factors

With responsibility for some manufacturing systems, one should be aware of the various factors that affect the statistics. Some factors are random, while others are controllable. Controllable factors are inputs. They can be numerical, such as temperature, or categorical, such as raw material suppliers. In charting an experiment, an engineer can use the letter "x" to represent the sum of all input variables.

Measuring the outputs, or responses, must be done in some quantitative manner such as a 1 to 5 scale. Engineers can use the letter "y" to represent the sum of the responses.

There are variables, such as ambient temperature and humidity, which can't be easily controlled or, in some cases, even identified. These uncontrolled variables, labeled "z," can cause variability in responses. Other sources of variability are deviations around the set points of the controllable factors, plus sampling and measurement error.

Furthermore, the system itself may be made up of parts that exhibit variability. How can all this variability be dealt with? By gathering system data, a run chart can be made to trace the wandering responses. Statistical process control (SPC) offers more sophisticated tools for assessing the natural variability of a system. However, to make systematic improvements, rather than just eliminating special causes, DOE must be applied.

Doing DOE is "talking" to a manufacturing process. "Questions" are asked by making changes in inputs, then "listening" to the response. SPC filters out the noise caused by variability, but it is a passive approach. DOE depends on asking the right questions. Therefore, subject matter knowledge is an essential prerequisite for successful application of DOE.

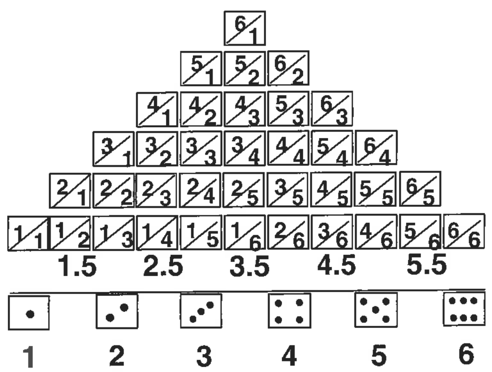
	SPC	DOE
Who ?	Operator	Engineer
How ?	Hands-off (monitor)	Hands-on (change)
Result.	Control	Breakthrough
Cause for Variability.	Special (upset)	Common (systemic)

Design of experiments (DOE) and statistical process control (SPC) are both very dependent on analysis of numbers, but that's where the similarity ends. DOE is used as a systematic tool and SPC is a tool for analyzing specific operations.

Normal distribution

When charting data from a system, there is often a bell-shaped pattern, called normal distribution, that emerges. However, not all distributions will be normal.

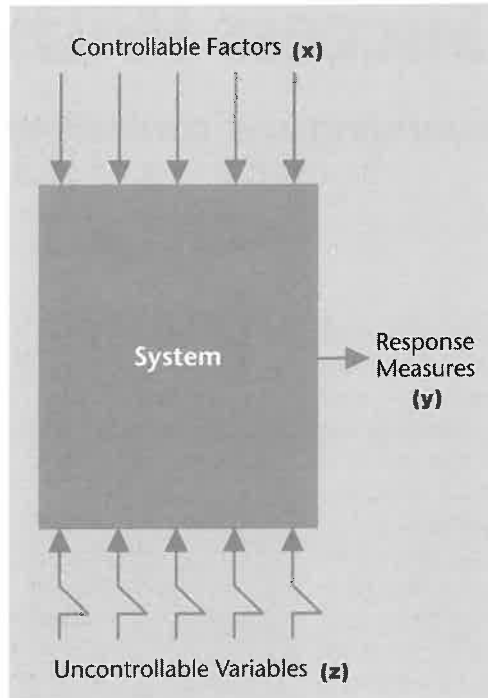
For example, if a six-sided die is repeatedly rolled, the frequency of getting 1 through 6 will be almost equal. This is called uniform distribution. If a pair of dice is rolled, the chances of them averaging to the extreme values of 1 or 6 are greatly reduced. The only way to hit an average of 1 from two dice is to roll two ones. On the other hand, there are three ways an average of 2 can be had: 1 and 3, 2 and 2, or 3 and 1. Average values of 1.5, 2.5, and so on, become possible. An average outcome of 3.5 is the most probable from a pair of dice.



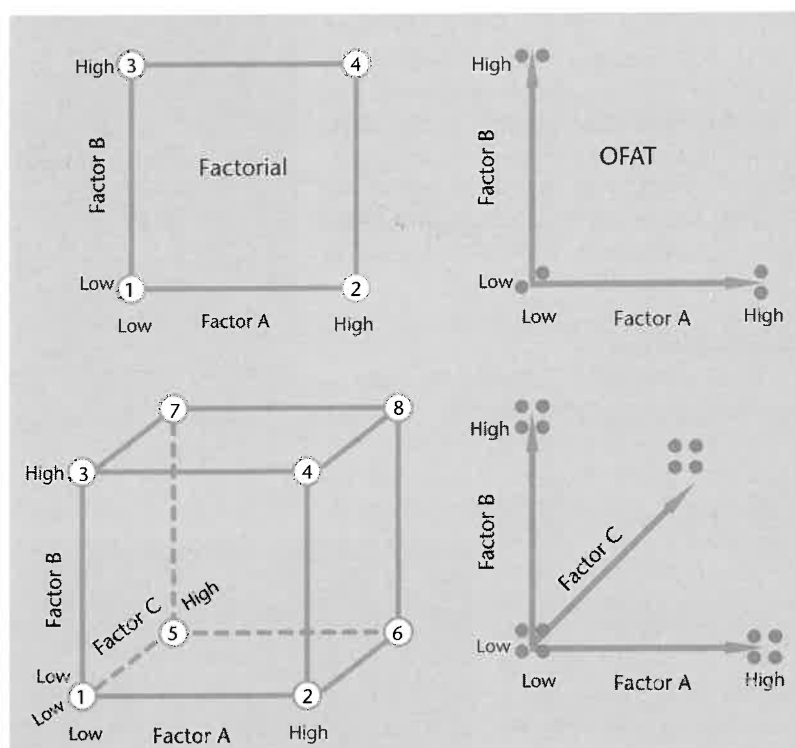
Rolling one die (bottom row) versus a pair of dice (pyramid at top).

The shape of the distribution becomes more bell-shaped as one die is increased to two. If more than two dice are repeatedly rolled, the distribution becomes even more bell-shaped and much narrower. If five dice are put in a cup, it is increasingly unlikely it would result in the extreme averages of 1 or 6—all five dice would have to come up 1 or 6 respectively.

The dice play illustrates the power of averaging—the more data collected, the more normal the distribution of averages, and the closer one gets to the average outcome. The normal distribution is "normal" because all systems are subjected to many uncontrolled variables. As in the case of rolling dice, it's unlikely that the variables will push the response in one direction or the other. Instead, they will tend to cancel each other and leave the system at a stable level, the mean, with some amount of consistent variability.



Not all variables are uncontrollable, some can be controlled and represent regular input to a system. The responses obtained from both types of variables must be measured in some quantifiable manner.



Factorial design lets an engineer experiment with multiple factors simultaneously vs. one-factor-at-a-time (OFAT) method. The result is a more efficient way to obtain results using fewer experimental runs. The square and cube on the left demonstrate how as factors are added to this two-level factorial design method, statistical power is gained quicker than the OFAT methods illustrated on the right.

Standard Order	Run Order	A: Brand	B: Time (minutes)	C: Power (percent)	Y_1: Taste (rating)	Y_2: "bullets" (ounces)
2	1	Costly (+)	4 (-)	75 (-)	75	3.5
3	2	Cheap (-)	6 (+)	75 (-)	71	1.6
5	3	Cheap (-)	4 (-)	100 (+)	81	0.7
4	4	Costly (+)	6 (+)	75 (-)	80	1.2
6	5	Costly (+)	4 (-)	100 (+)	77	0.7
8	6	Costly (+)	6 (+)	100 (+)	32	0.3
7	7	Cheap (-)	6 (+)	100 (+)	42	0.5
1	8	Cheap (-)	4 (-)	75 (-)	74	3.1

In building the results of factorial design, an engineer can get a quick glance at the outcome of his experiments. Using coded factor levels, illustrated with the "+" and "-," next to actual levels, the engineer gets an idea of highs and lows in the results.

Regardless of the shape of the original distribution of "individuals," the taking of averages results in a normal distribution. This comes from the central limit theorem. As shown in the dice example, the theorem works imperfectly with a subgroup of two. For SPC or DOE purposes, averages should be based on subgroups of four or more. A second aspect of the central limit theorem predicts the narrowing of the distribution as seen in the dice example. This is a function of the increasing sample size for the subgroup—the more data you collect, the better.

When making a decision about an experimental outcome, there are two types of errors minimized:

■ *Type I.* When something seems to have happened when it really didn't, it is called a false alarm.

■ *Type II.* Not discovering that something really happened, a failure to alarm, when it really did occur.

Playing with variables

The more factors that are tested, the less the chance is for error. One way to more accurately test is to use factorial design, which allows manufacturers to experiment on many factors simultaneously. The simplest factorial design involves two factors, each at two levels. Factorial design provides contrasts of averages, thus providing statistical power to the effect estimates. The one-factor-at-a-time (OFAT) experimenter must replicate runs to provide equivalent power. The end result for a two-factor study is that to get the same precision for effect estimation, OFAT requires six runs vs. only four for a two-level design.

The advantage of factorial design becomes more pronounced as more factors are added. For example, with three factors, the factorial design requires only eight runs vs. 16 for an OFAT experiment with equivalent power. The relative efficiency of factorial design is now twice that of OFAT for equivalent power. The relative efficiency of factorials continues to increase with every added factor.

Factorial design has two additional advantages from OFAT:

■ Wider inductive basis—it covers a broader area from which to draw inferences about a manufacturing process.

■ It reveals interactions of factors. This often proves to be the key to understanding a process.

The basic principles of two-level factorial design can be explained in the example of making microwave popcorn.

It's nearly impossible to get every kernel of corn to pop. Often, there's a considerable number of inedible kernels, or "bullets," at the bottom of the bag. What causes this loss of popcorn yield? In this example, only three factors were studied: brand of popcorn, time of cooking, and microwave power setting. Brand is a categorical factor—one type of popcorn or another. Time is a numerical factor because it can be adjusted to any level. Power could be set to any percent of the total available, so it's also a numerical factor. If this experiment is attempted, do some range finding on the high level for time. Using the minus (-) and plus (+) symbols to designate low and high levels, makes sense for numerical factors, provided the lesser value is made the low level. The symbols for categorical factor levels are arbitrary.

Two responses were considered for the experiment on microwave popcorn: taste and bullets. Taste was determined by a panel of testers who rated the popcorn on a scale of 1, worst, to 10, best. The ratings were averaged and multiplied by 10. This is a linear transformation that eliminates a decimal point to make data entry and analysis easier. It does not affect the relative results. The second response, bullets, was measured by weighing the unpopped kernels—the less the weight, the better.

The results from doing all combinations of the chosen factors, each at two levels, were taste ranging from a 32 to 81 rating, and bullets from 0.7 to 3.5 ounces. The latter result came from a bag with virtually no popped corn, barely enough to get a taste. Such a setup is one to avoid. The run order was randomized to offset any lurking variables, such as machine warm-up and degradation of taste buds.

Building and analyzing

The first item to list in the results of this popcorn experiment is the standard order. The mathematical symbols of minus and plus, the "coded factor levels," are next to the actual levels at their lows and highs, respectively. To take advantage of established methods for analysis, it's helpful to resort the test matrix on the basis of standard order and list only the coded factor levels. Dispense with the names of the factors and responses, which get in the way of the calculations, and show only their mathematical symbols.

A column labeled "Standard," plus the columns for brand, time, and power, form a template that can be used for any three factors tested at two levels. The standard layout starts with all low levels of the factors and ends with all high levels. The first factor changes sign every other row, the second factor every second row, the third every fourth row, and so on, based on powers of 2. You can extrapolate the pattern to any number of factors, or research in statistical handbooks.

The analysis can begin by investigating the main effects on the first response, taste. It helps to view test results in a cubical factor space.

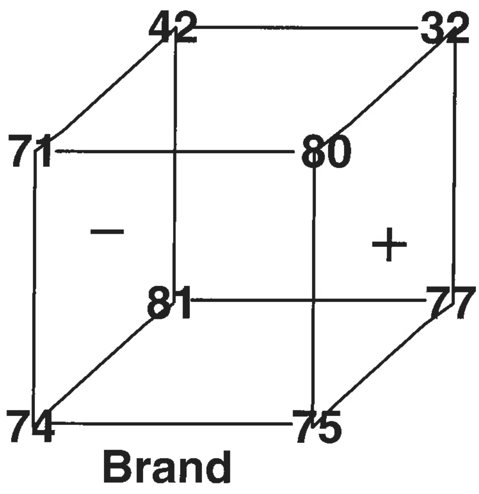
The right side of the cube contains all the runs where brand is at the high level, vs. the left side where the factor is held at the low level. Average the highs and the lows, and determine the difference or contrast. This is the effect of the brand factor. Mathematically, the calculation of an effect is expressed as:

$$\text{Effect} = \frac{\sum Y_+}{n_+} - \frac{\sum Y_-}{n_-}$$

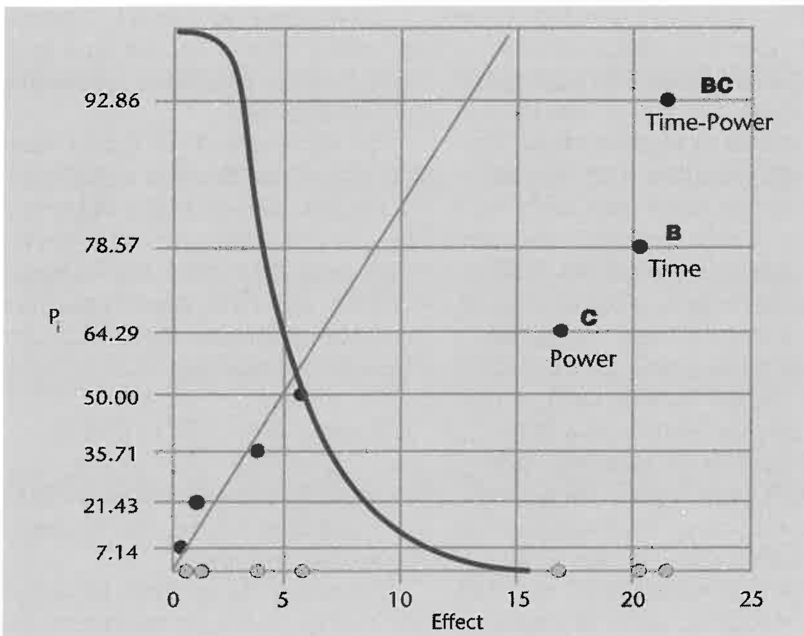
The n's refer to the number of data points collected at each level. The Y's refer to the associated responses.

Continue the analysis by contrasting the averages from top to bottom and back to front to get the effects of time and power, respectively.

Before jumping to conclusions, it's important to consider the effects caused by interactions of factors. The full-factorial design allows estimate of all three two-factor interactions: brand-time, brand-power,



Cube plot of taste ratings with focus on brand (Factor A).



Some experiments will deliver extreme results in the combination of factors. To get an idea of whether seeming extremes are the norm, it's best to plot all results on a half-normal distribution chart, which is based on the positive half of a full-normal curve.

and time-power; and the three-factor interaction of brand-time-power. Including the main effects, caused by brand, time, and power, there are seven effects, the most that can be estimated from the eight-run factorial design, because 1 degree of freedom is used to estimate the overall mean.

Once the effects have been listed, it becomes a matter of computing the effects using the general calculation of an effect. On an absolute value scale, some of the other interaction effects are extremely low or extremely high. Could such extremes be chance occurrences caused by normal variations in the popcorn, the tasting, the environment, and the like? The answer to this question is found in plotting the normal distribution.

Plotting results

Before plotting the effects, it helps to convert them to absolute values, a more sensitive scale for detection of significant outcomes. The absolute value scale is accommodated via a variety of normal papers called the half-normal, which is based on the positive half of the full normal curve. The vertical axis of the half-normal plot displays the cumulative probability of getting a result at or below any given level. However, the probability scale for the half-normal is adjusted to account for using the absolute value of the effects.

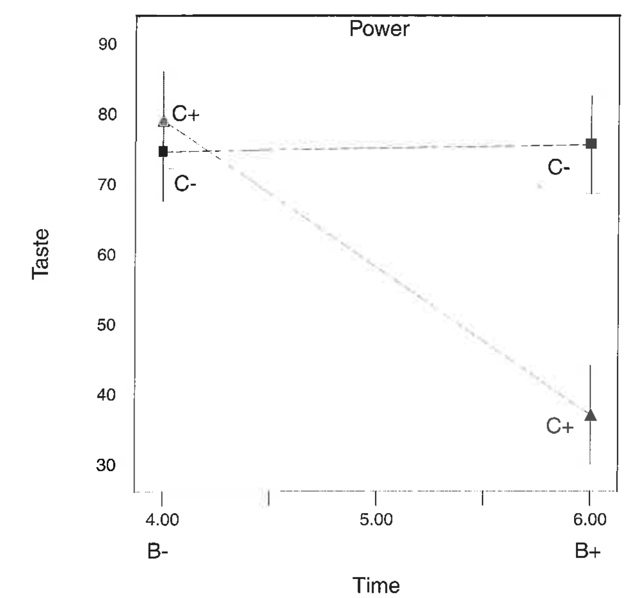
Before plotting data on the probability paper:

- Sort the data points in ascending order.
- Divide the 0 to 100% cumulative probability scale into equal segments.
- Plot the data at the midpoint of each probability segment.

After that is complete, plot the absolute values of the effect on the x-axis vs. the cumulative probabilities on the specially scaled y-axis on half-normal paper. Half-normal plots can be generated using statistical software.

The pattern that can result typically has the majority of points in a line emanating from the origin, followed by a gap, and then one or more points fall off to the right of the line. The half-normal plot of effects makes it very easy to see at a glance what's significant. This same procedure can be applied to the second response for microwave popcorn, the weight of the bullets.

To protect against spurious outcomes, it is important to verify the conclusions drawn from the half-normal plots by doing an analysis of variance and the associated diagnostics of residual error.



Interaction of time (B) versus power (C) on popcorn taste.

Interpretation

Notice that the effect of time depends on the level of power. For example, when power is low (minus), the change in taste is small--from 74.5 to 75.5. But when power is high (plus), the taste goes very bad--from 79 to 37. This is much clearer when graphed (see the "Interactions of time versus power" graph). Two lines appear on the plot, bracketed by least significant difference (LSD) bars at either end. The lines are far from parallel, indicating quite different effects of changing the cooking time. When power is low (C-), the line is flat, which indicates that the system is unaffected by time (B). But when power goes high (C+), the line angles steeply downward, indicating a strong negative effect due to the increased time. The combination of high time and high power is bad for taste. The average result is only 37 on the 100-point rating scale. The reason is simple: the popcorn burns. The solution to this problem is also simple: turn off the microwave sooner. Notice that when the time is set at its low level (B-), the taste remains high regardless of the power setting (C). The LSD bars overlap at this end of the interaction graph, which implies that there is no significant difference in taste.

Interactions occur when the effect of one factor depends on the level of the other. They cannot be detected by OFAT experimentation so don't be surprised if previously undetected interactions are uncovered when running a two-level design. Q

From "DOE Simplified: Practical Tools for Effective Experimentation," *Mark J. Anderson and Patrick J. Whitcomb. Copyright 2000, Productivity Inc., P.O. Box 13390, Portland, OR 97213. Contact at (800) 394-6868 or at www.productivity.inc. Reprinted with permission.*

Mark J. Anderson is principal of Stat-Ease Inc., Minneapolis. He may be contacted at (612) 378-9449 or mark@statease.com.

Reprinted from **QUALITY** April 2000, with additional information included from author's book.