THIRD EDITION

DOE Simplified

THIRD EDITION

DOE Simplified

Practical Tools for Effective Experimentation

Mark J. Anderson • Patrick J. Whitcomb

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150427

International Standard Book Number-13: 978-1-4987-3090-7 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

Without deviation from the norm, progress is not possible.

Frank Zappa

Design of experiments (DOE) is a planned approach for determining cause and effect relationships. It can be applied to any process with measurable inputs and outputs. DOE was developed originally for agricultural purposes, but during World War II and thereafter it became a tool for quality improvement, along with statistical process control (SPC). Until 1980, DOE was mainly used in the process industries (i.e., chemical, food, pharmaceutical) perhaps because of the ease with which engineers could manipulate factors, such as time, temperature, pressure, and flow rate. Then, stimulated by the tremendous success of Japanese electronics and automobiles, SPC and DOE underwent a renaissance. The advent of personal computers further catalyzed the use of these numerically intense methods.

This book is intended primarily for engineers, scientists, quality professionals, Lean Six Sigma practitioners, market researchers, and others who seek breakthroughs in product quality and process efficiency via systematic experimentation. Those of you who are industrial statisticians won't see anything new, but you may pick up ideas on translating the concepts for nonstatisticians. Our goal is to keep DOE simple and make it fun.

By necessity, the examples in this book are generic. We believe that, without much of a stretch, you can extrapolate the basic methods to your particular application. Several dozens of case studies, covering a broad cross section of applications, are cited in the Recommended Readings at the end of the book. We are certain you will find one to which you can relate.

DOE Simplified: Practical Tools for Effective Experimentation evolved from over 50 years of combined experience in providing training and computational tools to industrial experimenters. Thanks to the constructive feedback

of our clients, the authors have made many improvements in presenting DOE since our partnership began in the mid-1970s. We have worked hard to ensure the tools are as easy to use as possible for nonstatisticians, without compromising the integrity of the underlying principles. Our background in process development engineering helps us stay focused on the practical aspects. We have gained great benefits from formal training in statistics plus invaluable contributions from professionals in this field.

What's New in This Edition

A major new revision of the software that accompanies this book (via download from the Internet) sets the stage for introducing experiment designs where the randomization of one or more hard-to-change factors can be restricted. These are called *split plots*—terminology that stems from the field of agriculture, where experiments of this nature go back to the origins of DOE nearly a century ago. Because they make factors such as temperature in an oven so much easier to handle, split-plot designs will be very tempting to many experimenters. However, as we will explain, a price must be paid in the form of losses in statistical power; that is, increasing the likelihood of missing important effects. After studying the new chapter on split plots, you will know the trade-offs for choosing these designs over ones that are completely randomized.

This edition adds a number of other developments in design and analysis of experiments, but, other than the new material on split plots, remains largely intact. The reviews for *DOE Simplified* continue coming in strongly positive, so we do not want to tamper too much with our system. Perhaps the biggest change with this third edition is it being set up in a format amenable to digital publishing. Now web-connected experimenters around the globe can read *DOE Simplified*.

Another resource for those connected to the Internet is the "Launch Pad"—a series of voiced-over PowerPoint® lectures that cover the first several chapters of the book for those who do better with audiovisual presentation. The goal of the Launch Pad is to provide enough momentum to propel readers through the remainder of the *DOE Simplified* text. The reader can contact the authors for more information about the Launch Pad.

After publication of the first edition of this book, the authors wrote a companion volume called *RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments* (Productivity Press, 2004).

It completes the statistical toolset for achieving the peak of performance via empirical modeling. If *DOE Simplified* leaves you wanting more, we recommend you read *RSM Simplified* next.

We are indebted to the many contributors to development of DOE methods, especially George Box and Douglas Montgomery. We also greatly appreciate the statistical oversight provided by our advisors, University of Minnesota statistics professors Kinley Larntz and Gary Oehlert.

Mark J. Anderson
(mark@statease.com)

Patrick J. Whitcomb
(pat@statease.com)

Introduction

> There are many paths to enlightenment. Be sure to take one with a heart.
>
> **Lao Tzu**

This book provides the practical tools needed for performing more effective experimentation. It examines the nuts and bolts of design of experiments (DOE) as simply as possible, primarily by example. We assume that our typical reader has little or no background in statistics. For this reason, we have kept formulas to a minimum, while using figures, charts, graphs, and checklists liberally. New terms are denoted by quotation marks and also are included in a glossary for ready reference. As a spoonful of sugar to make the medicine go down, we have sprinkled the text with (mostly) relevant text boxes. Please enjoy (or forgive) the puns, irreverent humor, and implausible anecdotes.

Furthermore, we assume that readers ultimately will rely upon software to set up experimental designs and do statistical analyses. Many general statistical packages now offer DOE on mainframe or personal computers. Other software has been developed specifically for experimenters. For your convenience, one such program accompanies this book. You will find instructions for downloading the software (and viewing its tutorials) at the back of the book. However, you must decide for yourself how to perform the computations for your own DOE.

Chapter 1 presents the basic statistics that form the foundation for effective DOE. Readers already familiar with this material can save time by skipping ahead to Chapter 2 or Chapter 3. Others will benefit by a careful reading of Chapter 1, which begins with the most basic level of DOE: comparing two things, or two levels of one factor. You will need this knowledge to properly analyze more complex DOEs.

Chapter 2 introduces more powerful tools for statistical analysis. You will learn how to develop experiments comparing many categories, such as various suppliers of a raw material. After completing this section, you will be equipped with tools that have broad application to data analysis.

Chapters 3 through 5 explain how to use the primary tool for DOE: two-level factorials. These designs are excellent for screening many factors to identify the vital few. They often reveal interactions that would never be found through one-factor-at-a-time methods. Furthermore, two-level factorials are incredibly efficient, producing maximum information with a minimum of runs. Most important, these designs often produce breakthrough improvements in product quality and process efficiency.

Chapter 6 introduces more complex tools for two-level factorials. Before you plow ahead, be sure to do some of the simpler factorials described in prior chapters. Practice makes perfect.

Chapter 7 goes back to the roots of DOE, which originated in agriculture. This chapter provides more general factorial tools, which can accommodate any number of levels or categories. Although these designs are more flexible, they lack the simplicity of focusing on just two levels of every factor.

At this point, the book begins to push the limits of what can be expected from a DOE beginner. Chapters 8 and 9 definitely go beyond the boundary of elementary tools. They offer a peek over the fence at more advanced tools for optimization of processes and mixtures. Because these final chapters exceed the scope of the working knowledge meant to be provided, "DOE Simplified," we did not include practice problems. However, advanced textbooks, such as the companion volume to this book—*RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments* (Productivity Press, 2004)—are readily available to those of you who want to expand your DOE horizons.

Chapter 11 brings readers back to the basics with keys to doing good DOE. It also provides a process for planning experiment designs that takes statistical power into account.

Chapter 12 details split plots, which, as explained in the Preface, provide a workaround for factors that experimenters find difficult to change in random fashion. However, the relief from randomization comes at the cost of power. Consider the trade-offs carefully.

The flowchart in Figure I.1 provides a chapter-by-chapter "map." At the end of Chapters 1 through 7, you will find at least one practice problem. We strongly recommend that readers work through these problems (answers to which are posted on the Internet; see About the Software at

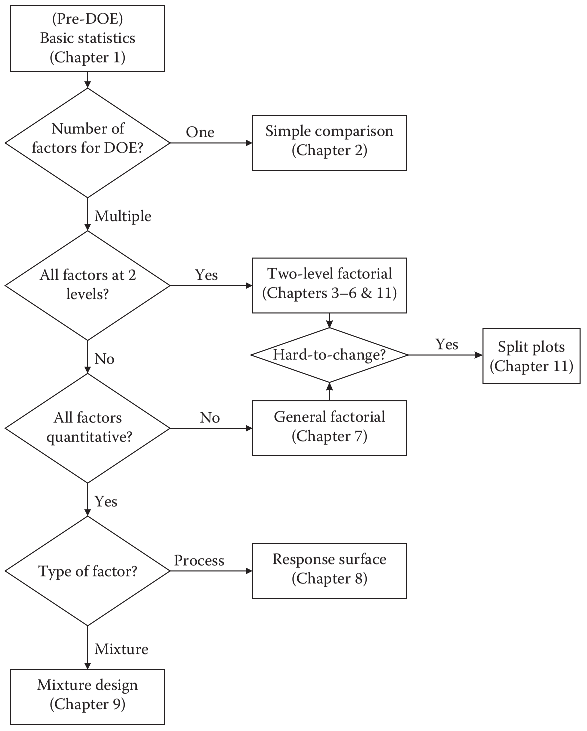

(Pre-DOE)
Basic statistics
(Chapter 1)
Number of factors for DOE?
One
Simple comparison (Chapter 2)
Multiple
All factors at 2 levels?
Yes
Two-level factorial (Chapters 3–6 & 11)
Hard-to-change?
Yes
Split plots (Chapter 11)
No
All factors quantitative?
No
General factorial (Chapter 7)
Yes
Type of factor?
Process
Response surface (Chapter 8)
Mixture
Mixture design (Chapter 9)

Figure I.1 Flowchart guide to *DOE Simplified*.

the back of the book for details). As with any new tool, the more you know about it, the more effectively you will use it.

Our hope is that this book inspires you to master DOE. We believe that reading this book, doing the exercises, and following up immediately with your own DOE will give you a starting point, a working knowledge of simple comparative and factorial designs. To foster this "DOE it yourself" attitude, we detail several practice experiments in Chapter 12. No answers are provided because we do not want to bias your results, but you may contact us for data from our experiments.

Chapter 1

Basic Statistics for DOE

One thing seems certain—that nothing certain exists.

Pliny the Elder, Roman scholar (CE 23–79)

Statistics means never having to say you're certain.

Slogan on shirt sold by the American Statistical Association (ASA)

Most technical professionals express a mixture of fear, frustration, and annoyance when confronted with statistics. It's hard even to pronounce the word, and many people, particularly after enduring the typical college lecture on the subject, prefer to call it "sadistics." Statistics, however, are not evil. They are really very useful, especially for design of experiments (DOE). In this chapter, we present basic statistics in a way that highlights the advantages of using them.

Statistics provide a way to extract information from data. They appear everywhere, not only in scientific papers and talks, but in everyday news on medical advances, weather, and sports. The more you know about statistics the better, because they can be easily misused and deliberately abused.

Imagine a technical colleague calling to give you a report on an experiment. It wouldn't make sense for your colleague to read off every single measurement; instead, you would expect a summary of the overall result. An obvious question would be how things came out on average. Then you would probably ask about the quantity and variability of the results so you could develop some degree of confidence in the data. Assuming that the

experiment has a purpose, you must ultimately decide whether to accept or reject the findings. Statistics are very helpful in cases like this; not only as a tool for summarizing, but also for calculating the risks of your decision.

GO DIRECTLY TO JAIL

When making a decision about an experimental outcome, minimize two types of errors:

1. Type I: Saying something happened when it really didn't (a false alarm). This is often referred to as the alpha (α) risk. For example, a fire alarm in your kitchen goes off whenever you make toast.
2. Type II: Not discovering that something really happened (failure to alarm). This is often referred to as the beta (β) risk. For example, after many false alarms from the kitchen fire detector, you remove the battery. Then a piece of bread gets stuck in the toaster and starts a fire.

The following chart shows how you can go wrong, but it also allows for the possibility that you may be correct.

Decision-Making Outcomes		*What You Say Based on Experiment:*	
		Yes	*No*
The Truth:	*Yes*	Correct	Type 2 Error
	No	Type 1 Error	Correct

The following story illustrates a Type I error. Just hope it doesn't happen to you!

A sleepy driver pulled over to the side of the highway for a nap. A patrolman stopped and searched the vehicle. He found a powdery substance, which was thought to be an illegal drug, so he arrested the driver. The driver protested that this was a terrible mistake; that the bag contained the ashes from his cremated grandmother. Initial screening tests gave a positive outcome for a specific drug. The driver spent a month in jail before subsequent tests confirmed that the substance really was ashes and not a drug. To make matters worse, most of grandmother's ashes were consumed by the testing. The driver filed a lawsuit seeking unspecified damages. (Excerpted from a copyrighted story in 1998 by the *San Antonio Express-News.*)

The “X” Factors

Let’s assume you are responsible for some sort of system, such as:

- Computer simulation
- Analytical instrument
- Manufacturing process
- Component in an assembled product
- Any kind of manufactured “thing” or processed “stuff”

In addition, the system could be something people-related, such as a billing process or how a company markets its products via the layout of an Internet web page or point-of-purchase display. To keep the example generic, consider the system as a black box, which will be affected by various controllable factors (Figure 1.1). These are the inputs. They can be numerical (e.g., temperature) or categorical (e.g., raw material supplier). In any case, we will use the letter “X” to represent the input variables.

Presumably, you can measure the outputs or responses in at least a semiquantitative way. To compute statistics, you must at least establish a numerical rating, even if it’s just a 1 to 5 scale. We will use the letter “Y” as a symbol for the responses.

Unfortunately, you will always encounter variables, such as ambient temperature and humidity, which cannot be readily controlled or, in some cases, even identified. These uncontrolled variables are labeled “Z.” They can be a major cause for variability in the responses. Other sources of variability are deviations around the set points of the controllable factors, sampling and measurement error. Furthermore, the system itself may be composed of parts that also exhibit variability.

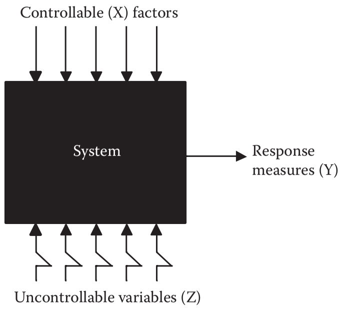


Figure 1.1 System variables.

Table 1.1 How DOE differs from SPC

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

How can you deal with all this variability? Begin by simply gathering data from the system. Then make a run chart (a plot of data versus time) so you can see how much the system performance wanders. 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—you must apply DOE. Table 1.1 shows how the tools of SPC and DOE differ.

TALK TO YOUR PROCESS (AND IT WILL TALK BACK TO YOU)

Bill Hunter, one of the co-authors of a recommended book on DOE called *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed. (Wiley-Interscience, 2005), said that doing experiments is like talking to your process. You ask questions by making changes in inputs, and then listen to the response. SPC offers tools to filter out the noise caused by variability, but it is a passive approach, used only for listening. DOE depends completely on you to ask the right questions. Asking wrong questions is sometimes called a Type III error (refer to the earlier text on Type I and II errors). Therefore, subject matter knowledge is an essential prerequisite for successful application of DOE.

> *“When I took math class, I had no problem with the questions, it was the answers I couldn’t give.”*
>
> **Rodney Dangerfield**

Does Normal Distribution Ring Your Bell?

When you chart data from a system, it often exhibits a bell-shaped pattern called a *normal distribution*. However, not all distributions will be

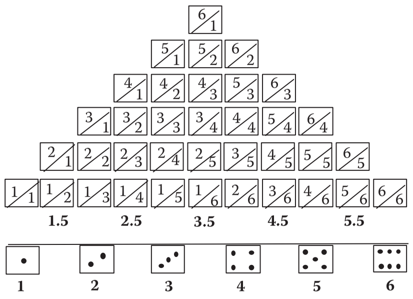


Figure 1.2 Rolling one die (bottom row) versus a pair of dice (pyramid at top).

normal. For example, if you repeatedly roll a six-sided die, the frequency of getting 1 through 6 will be approximately equal (see bottom row of Figure 1.2). This is called a *uniform* distribution. However, if you roll a pair of dice, 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 1s (snake eyes). On the other hand, there are three ways you can roll an average of 2: (1, 3), (2, 2), or (3, 1). The combinations of two dice are represented by the pyramid at the top of Figure 1.2 (above the line). Average values of 1.5, 2.5, and so on now become possible. An average outcome of 3.5 is most probable from a pair of dice.

Notice how the distribution becomes more bell-shaped (normal) as you go from one die to two dice. If you roll more than two dice repeatedly, the distribution becomes even more bell-shaped and much narrower.

DON'T SLICE, JUST DICE

Rather than fight a war over a disputed island, King Olaf of Norway arranged to roll dice with his Swedish rival. The opponent rolled a double 6. "You can't win," said he. Being a stubborn Norwegian, Olaf went ahead anyway—in the hope of a tie. One die turned up 6; the other split in two for a total of 7 (because the opposing sides always total 7). So Norway got the island with a lucky—and seemingly impossible—score of 13. This outcome is called an *outlier,* which comes from a special cause. It's not part of the normal distribution. Was it a scam? (From Ivar Ekeland. 1996. *The Broken Dice and Other Mathematical Tales of Chance.* Chicago: University of Chicago Press.)

For example, let's say you put five dice in a cup. Consider how unlikely it would be to get 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 you collect, the more normal the distribution of averages and the closer you get to the average outcome (for the dice the average is 3.5). The normal distribution is "normal" because all systems are subjected to many uncontrolled variables. As in the case of rolling dice, it is very unlikely that these variables will conspire to push the response in one direction or the other. They will tend to cancel each other out and leave the system at a stable level (the mean) with some amount of consistent variability.

THE ONLY THEOREM IN THIS ENTIRE BOOK

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 purposes of SPC or DOE, we recommend that you base your averages 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, which is a function of the increasing sample size for the subgroup. The more data you collect the better.

Descriptive Statistics: Mean and Lean

To illustrate how to calculate descriptive statistics, let's assume your "process" is rolling a pair of dice. The output is the total number of dots that land face up on the dice. Figure 1.3 shows a frequency diagram for 50 rolls.

Notice the bell-shaped (normal) distribution. The most frequently occurring value is 7. A very simplistic approach is to hang your hat on this outpost, called the *mode*, as an indicator of the location of the distribution. A much more effective statistic for measuring location, however, is the *mean*, which most people refer to as the *average*. (We will use these two terms interchangeably.)

EDUCATION NEEDED ON MEAN

A survey of educational departments resulted in all 50 states claiming their children to be above average in test scores for the United States. This is a common fallacy that might be called the "Lake Wobegon Effect" after the

Result	*Tally*	*Number (n)*	*Product*
12	X	1	12
11	X	1	11
10	XXXXX	5	50
9	XXXX	4	36
8	XXXXXXXX	8	64
7	XXXXXXXXXXX	11	77
6	XXXXXXX	7	42
5	XXXXXX	6	30
4	XXXX	4	16
3	XX	2	6
2	X	1	2
Sum		50	346

Figure 1.3 Frequency distribution for 50 rolls of the dice. (Data from SPC Simplified.)

mythical town in Minnesota, where, according to author Garrison Keillor, "… all women are strong, all the men good-looking, and all the children above average."

In a related case, a company president had all his employees tested and then wanted to fire the half that were below average. Believe it or not.

The formula for the mean of a response (Y) is shown below, where "n" is the sample size and "i" is the individual response:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

The mean, or "Y-bar," is calculated by adding up the data and dividing by the number of "observations." For the dice:

$$\bar{Y} = \frac{346}{50} = 6.92$$

This is easy to do on a scientific calculator. (Tip: If you don't have a calculator handy, look for an app on your smartphone, tablet, or computer. Many of these require changing the view to an advanced scientific mode to enable doing squares, roots, and other functions needed for statistical calculations.)

STATISTICALLY (BUT NOT POLITICALLY) CORRECT QUOTES

"Even the most stupid of men, by some instinct of nature,
is convinced that the more observations [n] have been made,
the less danger there is of wandering from one's goal."

Jacob Bernoulli, 1654–1705

"The ns justify the means."

(Slogan on shirt seen at an American Statistical Association meeting)

Means don't tell the whole story. For example, when teaching computer-intensive classes, the authors often encounter variability in room temperature. Typically, it is frigid in the morning but steamy by the afternoon, due to warmth from the student bodies and heat vented off the computers and projector. Attendees are never satisfied that on average the temperature throughout the class day is about right.

The most obvious and simplest measure of variability is the "range," which is the difference between the lowest and highest response. However, this is a wasteful statistic because only two values are considered. A more efficient statistic that includes all data is "variance" (see formula below).

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}$$

Variance (s^2) equals the sum of the squared deviations from the mean, divided by one less than the number of individuals. For the dice:

$$s^2 = \frac{233.68}{(50-1)} = 4.77$$

The numerator can be computed on a calculator or a spreadsheet program. It starts in this case with Y_1 of 12, the first response value at the top row of Figure 1.3. Subtracting the mean (Y-bar) of 6.92 from this value nets 5.08, which when squared, produces a result of 25.81. Keep going in this manner on the 49 other responses from the dice-rolling process. These squared differences should then add up to 233.68 as shown above. The denominator (n – 1) is called the *degrees of freedom* (df). Consider this to be the amount of information available for the estimate of variability after calculating the mean. For example, the degrees of freedom to estimate variability from one observation would be zero. In other words, it is impossible to estimate variation. However, for each observation after the first, you get one degree of freedom to estimate variance. For example, from three observations, you get two degrees of freedom.

A FUNCTION BY ANY OTHER NAME WILL NOT BE THE SAME

When using statistical calculators of spreadsheet software, be careful to select the appropriate function. For example, for the denominator of the variance, you want the sum of squares (SS) corrected for the mean. Microsoft Office Excel® offers SUMSQ worksheet function, but this does not correct for the mean. The proper function is DEVSQ, which does correct for the mean.

Variance is the primary statistic used to measure variability, or dispersion, of the distribution. However, to get units back to their original (not squared) metric, it's common to report the "standard deviation(s)." This is just the square root of variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}}$$

For the dice:

$$s = \sqrt{4.77} = 2.18$$

SAMPLING: A MEATY SUBJECT

The symbol "s" is used when we compute standard deviation from sample data. Some statistical calculators offer this function under the label of "σ_{n-1}," where the Greek letter sigma refers to the true value of standard deviations for the "population." Sampling is a necessity in process improvement, because the population would be all the possible outcomes, which are unknown and unknowable. Even if you could test the entire population, it would be inefficient to do so. This became a huge political issue in the year 2000 when U.S. census takers wanted to use sampling rather than counting every individual in the entire population. The statistical treatment for populations is simpler than for sample data, but because it's not relevant for industrial experimenters, we won't spend time on population statistics. You can get this information from any book on basic statistics.

"You do not have to eat the whole ox to know the meat is tough!"

Shakespeare

Confidence Intervals Help You Manage Expectations

In the dice example, the sample mean was calculated from 50 rolls. What happens if the pair of dice is rolled another 50 times? What if you repeat this experiment many times? You are encouraged to try this, but you can probably guess the outcome—the means will vary somewhat around the true value of 7 (assuming perfect dice). This variation, measured by taking a standard deviation of the means, is called the *standard error* (SE) of the mean.

It would be a terrible waste of time to repeat an experiment many times to get an estimate of the standard error. To save time, statisticians use an approximation based on part two of the central limit theorem (see above boxed text):

$$SE \equiv s_{\bar{Y}} \cong \sqrt{\frac{s^2}{n}}$$

For the dice:

$$SE \cong \sqrt{\frac{4.77}{50}} = 0.31$$

The standard error gets smaller as the number sampled (n) gets larger. You get what you pay for; the more you sample, the more precisely you can estimate the "true" outcome.

Very few people understand standard error, but almost everybody recognizes a related statistic called the confidence interval, particularly in election years. Any political poll that calls itself "scientific" reveals its so-called margin of error (MOE). For example, a pollster may say that the leading candidate got a 60% approval rating, plus or minus 3%. In most cases, when such a margin or interval is reported, it's based on a confidence level of 95%. Intervals constructed in this way will contain the true value of the calculated statistic 95% of the time.

A POLITICIAN WHO GETS IT: THE SECRET WEAPON

A clever twist by U.S. Senator Amy Klobuchar on Teddy Roosevelt's famous advice for being an effective leader:

> *"Speak softly, but carry a big statistic."* (From CBS radio interview, 11/28/2012)

The formula for the confidence interval (CI) of a mean is

$$CI = \bar{Y} \pm t \,\square\, SE$$

where "t" is a factor that depends on the confidence desired and the degrees of freedom generated for the estimate of error. It represents the number of standard deviations from the estimated mean. The t-statistic can be looked up in the table provided in Appendix 1.1 at the back of the book.

BELIEVE IT OR NOT: A UNIVERSITY STUDENT WHO'S AN EXPERT ON BEER

An Oxford graduate in chemistry and mathematics who worked at Guinness brewery in Dublin, Ireland, characterized the t-distribution in 1908. His name was W. S. Gosset, but out of modesty, or perhaps due to restrictions from Guinness, he published under the pen name "Student." Prior to Student's t-distribution, statisticians had directed their attention to large samples of data, often with thousands of individuals. Their aims were primarily academic. Guinness couldn't afford to do pure research

and did only limited field trials on natural raw materials, such as barley. In the process, Gossett/Student discovered: "As we decrease the number of experiments, the value of the standard deviation found from the sample of experiments becomes itself subject to increasing error."

He developed practical tools for dealing with variability in soil and growing conditions so decisions could be made with some degree of confidence. Gossett/Student then described the distribution (t) of small samples: "If the sample is very small, the 'tails' of its distribution take longer to approach the horizontal and its 'head' is below the head of the normal distribution; as the number in the sample increases, the tails move 'inwards' toward the tails of the normal distribution, and the 'head' of the curve moves toward the 'head' of the normal distribution."

(From Gavin Kennedy, 1983. *Invitation to statistics*. Oxford, U.K.: Martin Robinson & Co.)

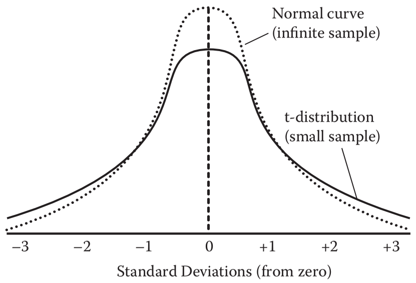


Figure 1.4 Normal curve (dotted) versus t-distribution (solid).

As you can see in Figure 1.4, the t-distribution looks like a normal curve with "heavy" tails (fatter, longer). However, for practical purposes, when you collect a large amount of data, say n > 30, the t-distribution becomes normal. Figure 1.5 shows the percent of individual outcomes as a function of standard deviation for the normal curve (or for the t-distribution based on a very large sample size).

Figure 1.5 demonstrates that the majority of individuals (specifically 68%) from a given population will fall within one standard deviation of the mean (the zero point). By expanding the range of standard deviation to plus or minus two, we can include 95% of the population. Any individuals outside of this wider range certainly could be considered to be special. Therefore,

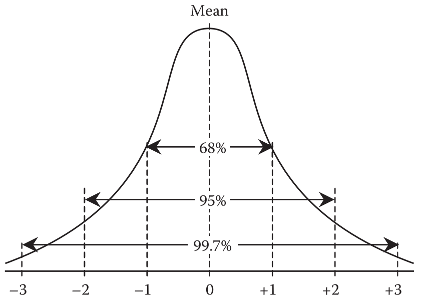


Figure 1.5 Percentage of individuals falling within specified limits of standard deviation for normal distribution (or t-distribution based on large sample size).

for large samples, it makes sense to set the critical value of t at 2 when constructing a confidence interval. In the example of the dice:

$$CI = 6.92 \pm 2(0.31) = 6.92 \pm 0.62$$

Intervals constructed in this manner will include the true population mean 95% of the time. The calculated range from 6.30 (6.92 minus 0.62) to 7.54 (6.92 plus 0.62)—commonly called a 95% confidence interval—does encompass the theoretical outcome of 7 (thank goodness).

Confidence intervals are an extremely useful tool for experimenters, because they allow an assessment of uncertainty. Using statistics allows you to quantify probabilities for error and play a game of calculated risks.

HOW NOT TO INSPIRE CONFIDENCE

A young engineer, obviously not well-schooled in statistics, made this statement about a problem encountered while testing a new piece of equipment: “This almost always hardly ever happens.” What do you suppose is the probability of “this” happening again?

Graphical Tests Provide Quick Check for Normality

The basic statistical tools discussed in this section are very powerful. However, they all depend on the assumption of normality. As anyone who does technical work knows, it always pays to check assumptions before making a final report. So, before going any farther, it will be worthwhile to learn how to check your data with a special type of graph called a *normal plot*. This graph is very handy because normally distributed data will fall nicely in line when plotted properly. You will see the normal plot—and its cousin, the “half-normal” plot—throughout this book.

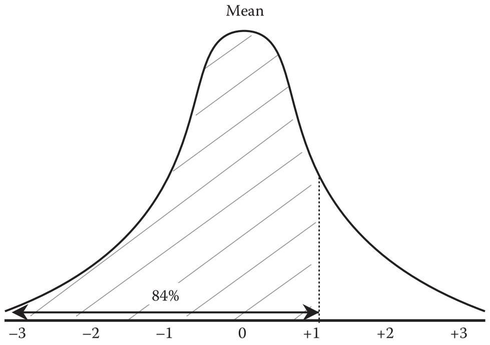


Figure 1.6 Cumulative probability (area under curve) at +1 standard deviation.

The normal plot requires a properly scaled template called *probability paper*. Long ago, you would get this template from a commercial vendor of graph paper; today, most people generate normal plots from statistical software. However, it’s not that hard to make it yourself. The x-axis is easy. It’s made just like any other graph with regularly spaced tic marks. The y-axis, on the other hand, is quite different, because it’s scaled by “cumulative probability.” This is a measure of the percentage of individuals expected to fall below a given level, benchmarked in terms of standard deviation. For example, as shown by the shaded area in Figure 1.6, the cumulative probability at the benchmark of 1 standard deviation is approximately 84%.

The cumulative probability can be determined by measuring the proportion of area under the normal curve that falls below any given level. This has been enumerated in great detail by statisticians. Table 1.2 provides the highlights.

We are now ready to tackle the mysterious y-axis for probability paper. Figure 1.7 shows how this can be done by putting standard deviations on a linear scale at the right side and recording the associated cumulative probability on the left side. We added a lot more detail to make the plot more usable. As you can see, the cumulative probability axis is very nonlinear.

Table 1.2 Cumulative probability versus number of standard deviations (from the mean)

Standard Deviations	*Cumulative Probability*
–2.0	2.3%
–1.0	15.9%
0.0	50.0%
1.0	84.1%
2.0	97.7%

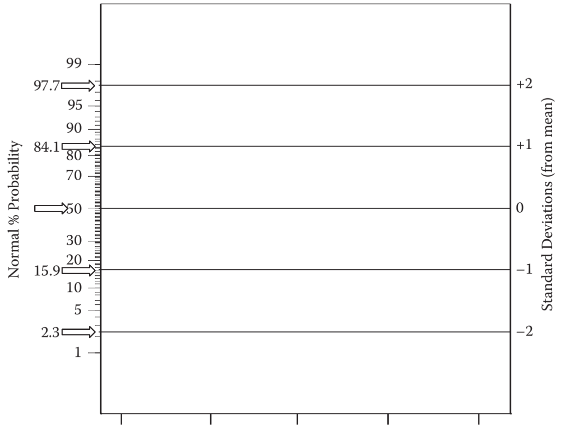


Figure 1.7 Probability paper with standard deviation scale at the right.

Now that we have done all of this work, let's go for the payoff: checking data for normality. The following 10 weights (in pounds) come from a random sample of men at the 25th reunion of an all-boys high school class: 199, 188, 194, 206, 235, 219, 200, 234, 171, 172. Various statistics could be computed on these weights (and snide comments made about their magnitude), but all we want to do is check the data for normality. A few things must be done before plotting the data on the probability paper:

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

Table 1.3 Values to plot on probability paper

Point	*Weight*	*Cumulative Probability*
1	171	5%
2	172	15%
3	188	25%
4	194	35%
5	199	45%
6	200	55%
7	206	65%
8	219	75%
9	234	85%
10	235	95%

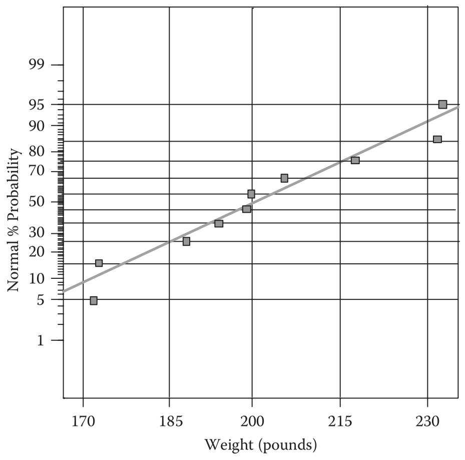


Figure 1.8 Normal plot of weights.

In this case, the sample size (n) is 10, so each probability segment will be 10% (100/10). The lowest weight will be plotted at 5%, which is the midpoint of the first segment. Table 1.3 shows this combination and all the remaining ones.

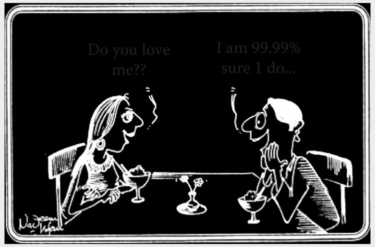
Now all we need to do is plot the weights on the x-axis of the probability paper and the cumulative probabilities on the y-axis (Figure 1.8).

The interpretation of the normal plot is somewhat subjective. Look for gross deviations from linear, such a big "S" shape. A simple way to check for linearity is to apply the "pencil test." If you can cover all the points with a pencil, the data are normal. It's okay if only a portion of some points touch the pencil. Don't be too critical. By the way, to discourage overly precise engineers from being too picky about these plots, statisticians suggest a "statistics test"; but in this case perhaps we shouldn't emphasize being fat. According to these criteria, the weight data passes the pencil test: fat or the standard diameter. Therefore, the alumni apparently conform to the normal bell-shaped curve weight-wise (and for the heavier men, shape-wise).

This concludes our coverage of basic statistical tools that form the foundation for DOE. You are now ready to move on to DOE procedures for making simple comparisons, such as which material works best, or whether you will benefit by changing suppliers.

DO NOT TRY THIS AT HOME

Please, for your own good, do not get carried away with statistical thinking to the extent of the fellow in this cartoon,[*] who took to heart the advice from the ASA as quoted at the outset of this chapter; that is, never saying one is certain.



[*] Used by permission from the cartoonist: Nadeem Irfan Bukhari.

Practice Problems

To practice using the statistical techniques you learned in Chapter 1, work through the following problems.

Problem 1.1

You are asked to characterize the performance of a motor-shaft supplier. The data shown below is a measure of the endplay:

61, 61, 57, 56, 60, 52, 62, 59, 62, 67, 55, 56, 52, 60, 59, 59, 60, 59, 49, 42, 55, 67, 53, 66, 60.

Determine the mean, standard error, and approximate 95% confidence interval for the endplay using the guidelines given in this book. (Suggestion: Use the software provided with the book. First do the simple sample feature tour included with the program's tutorials. It's keyed to the data above. See the end of the book for software instructions and details on the tutorial files.)

Problem 1.2

The alumni from the case described at the end of the chapter reminisced about their weights at the time of graduation. One of them contacted the school nurse and dug up the old records. These are shown below:

153, 147, 148, 161, 190, 167, 155, 178, 130, 139.

Simple inspection of these results versus those reported earlier will reveal the obvious impact of aging, so we need not bother doing any statistics (enough said). Your job is to construct a normal plot of these 10 data points. (Suggestion: To save time, reuse the normal plot shown earlier in Figure 1.8. Just subtract 40 from each tic mark. Then plot the new points with a red pen or pencil so you can see them clearly.) Do you see anything grossly abnormal about the data?

THE "BODY" IMPOLITIC

Former wrestler, Governor Jesse Ventura, of the author's home state of Minnesota, described an unfavorable political poll as "skewered." The proper term for a nonnormal distribution that leans to the left or right is "skewed." However, Governor Ventura, who tries to stay in the center, may be on to something with his terminology.

> *"Keep it simple and stupid."*
>
> **Jesse Ventura**

Chapter 2

Simple Comparative Experiments

Many of the most useful designs are extremely simple.

Sir Ronald Fisher

We now look at a method for making simple comparisons of two or more "treatments," such as varying brands of toothpaste, and testing their efficacy for preventing cavities. Called the *F-test*, it is named after Sir Ronald Fisher, a geneticist who developed the technique for application to agricultural experiments. The F-test compares the variance among the treatment means versus the variance of individuals within the specific treatments. High values of F indicate that one (or more) of the means differs from another. This can be very valuable information when, for example, you must select from several suppliers or materials or levels of a process factor. The F-test is a vital tool for any kind of design of experiments (DOE), not just simple comparisons, so it's important to understand as much as one can about it.

PUT ON YOUR KNEE-LENGTH BOOTS

In his landmark paper on DOE entitled "The Differential Effect of Manures on Potatoes," Fisher analyzes the impact of varying types of animal waste on yield of spuds. The agricultural heritage explains some of the farm jargon you see in writings about DOE: blocks, plots, treatments, environmental factors, etc. It's a lot to wade through for nonstatisticians, but worth the effort.

The F-Test Simplified

Without getting into all the details, the following formula for F can be derived from part two of the central limit theorem:

$$F \equiv \frac{ns_{\bar{Y}}^2}{s_{pooled}^2}$$

This formula assumes that all samples are of equal size n. You might think of F as a ratio of signal (differences caused by the treatments) versus noise. The F-ratio increases as the treatment differences become larger. It becomes more sensitive to a given treatment difference as the sample size (n) increases. Thus, if something really does differ, you will eventually find it if you collect more data. On the other hand, the F-ratio decreases as variation (s_{pooled}^2) increases. This noise is your enemy. Before you even begin experimentation, do what you can to dampen system variability via statistical process control (SPC), quality assurance on your response measurement, and control of environmental factors.

If the treatments have no effect, then the F-ratio will be near a value of 1. As the F-ratio increases, it becomes less and less likely that this could occur by chance. With the use of statistical tables, such as those provided in Appendix 1, you can quantify this probability ("p"). The "p-value" becomes a good "bottom-line" indicator of significance. When the F-ratio gets so high that the p-value falls below 0.05, then you can say with 95% confidence that one or more of the treatments is having an effect on the measured response. This still leaves a 5% "risk" that noise is the culprit. (Technically, this is a Type I error—detailed in the "Go Directly to Jail" at the beginning of Chapter 1.) If 5% risk is too much, you can set a standard of 1%, thus ensuring 99% confidence. Conversely, you may want to live dangerously by taking a 10% risk (90% confidence). It's your choice. Our choice for examples shown in this book will be 95% confidence for all tests and intervals.

ANOVA: NEITHER A CAR, NOR AN EXPLODING STAR

The F-test uses variance as its underlying statistic. Therefore, statisticians call the overall procedure *analysis of variance*, or ANOVA. When this is applied to simple comparisons, it's called a *one-way* ANOVA. With the advent of built-in spreadsheet functions and dedicated statistical software, ANOVA can be accomplished very easily. Nonstatisticians, however, may still find it somewhat scary.

(P.S.: The allusion to a car pays homage to the 1960s vintage Chevrolet Nova, which legend has that it did not sell well in Latin America because in Spanish "no va" means "it doesn't go.")

A Dicey Situation: Making Sure They Are Fair

Let's look at another example: further experimentation with dice. One of the authors purchased a game called Stack® developed by Jeffrey L. Strunk. The game came with 56 six-sided dice in 4 colors (14 dice per each).

Let's assume that it's advantageous to get dice that tend to roll high. The first question is whether the change in colors causes a significant effect on the average outcome. Table 2.1 shows the results obtained by dumping the dice on a desk. The bag was first shaken to thoroughly mix the dice, thus assuring a "random" toss.

Notice that the data forms a histogram when viewed sideways. You can see the relative uniform shape (nonnormal) of the distribution for individual dice. Furthermore, you will observe a great deal of overlap between colors and tosses. Viewing the table, it is hard to see any differences, but why take a chance—perform an ANOVA.

Table 2.1 Frequency distribution for 56 rolls of the dice

Result	*White (1)*	*Blue (2)*	*Green (3)*	*Purple (4)*
6	6/6	6/6	6/6	6
5	5	5	5	5
4	4	4/4	4/4	4
3	3/3/3/3/3	3/3/3/3	3/3/3/3	3/3/3/3/3
2	2/2/2	2/2/2/2	2/2/2/2	2/2/2/2/2
1	1/1	1	1	1
Mean ($\bar{Y}$)	3.14	3.29	3.29	2.93
Var. (s^2)	2.59	2.37	2.37	1.76

Let's get started on the calculation of F. The simplest part is determining the sample size (n). The answer is 14—the number of dice of each color. Next, the variance (s^2) of the means must be calculated. This can be done on any statistical calculator or standard spreadsheet program, or by chugging through the following equation:

DON'T GET PARANOID ABOUT PATTERNS IN DATA

There appears to be an inordinate number of 3s and 2s in Table 2.1, but this was just a chance occurrence. (Dubious readers are encouraged to look at Problem 2.2 at the end of this chapter that shows results from a second toss.) Coincidences occur more often than most people think. Although we recommend you always look over your raw data, it's best to be skeptical about seemingly unusual patterns. Wait until you apply statistical tools, such as ANOVA and diagnostic plots, before tampering with the results, and even then be very sure you can attribute a special cause for those results. (For example, the die being split in the Chapter "Don't Slice, Just Dice.")

> *"You can see a lot just by looking."*
>
> **Yogi Berra**
>
> *"Unexpected events happen unexpectedly often."*
>
> **Freeman Dyson**

$$S_{\bar{Y}}^2 = \frac{(3.14-3.1625)^2+(3.29-3.1625)^2+(3.29-3.1625)^2+(2.93-3.1625)^2}{4-1}$$

$$= 0.029$$

where the value 3.1625 is the grand mean of all the responses, or equivalently, the average of the four group means (3.14, 3.29, 3.29, 2.93). (The extra digits on the grand mean are carried along for the sake of accuracy.) Finally, the last element in the F equation, the variance of the individuals pooled (S^2_{pooled}), is computed by simply taking the mean of the variance within each type.

$$S_{pooled}^2 = \frac{2.59+2.37+2.37+1.76}{4} = 2.28$$

Therefore, for toss one of the colored dice:

$$F = \frac{n * S_{\bar{Y}}^2}{s_{pooled}^2} = \frac{14 * 0.029}{2.28} = 0.18$$

The value of 14 for n (sample size) represents the number of dice of each color. The resulting F-ratio falls below 1, which implies that the variation between mean results by color is not excessive when compared with the variation within each color. We provide F-tables in Appendix 1, but it's not even worth looking at them at this point because the F is below a value of 1. (Don't worry, we will get back to the tables in part two of this example.) It's correct to assume that the results are not significant. Therefore, it appears safe to go ahead and play the game.

NEVER SAY NEVER, BUT IT'S OKAY TO STATE A DEFINITE MAYBE

Statisticians think of experimentation as a "hypothesis test." Their approach is to assume that there is no effect. This is the "null" hypothesis. It must be disproved beyond a specified risk, such as 5%. This can be likened to the jury system of law where the standard is innocent until proven guilty beyond a shadow of a doubt. A subtlety of this statistical approach is that you can never prove the null—no effect. It would require an infinite sample size to do so. For example, in the dice case, a very subtle doctoring could be obscured by natural variability over many cycles of testing, but in the long run it could provide enough of an advantage to make a good living for the crooked gambler.

> *"Absence of evidence is not evidence of absence."*
>
> **Carl Sagan**

Figure 2.1 shows graphs of the data with "least significant difference" (LSD) bars, a form of confidence interval, superimposed. These are set at 95% confidence. Don't be alarmed that results fall outside the interval because it's related to the average outcome, not the individual tosses. The more data you collect, the narrower the interval becomes.

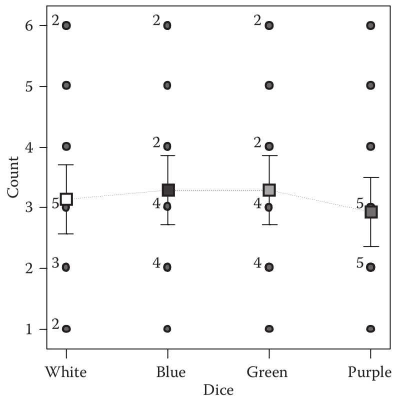


Figure 2.1 Effects plot for dice of various colors (for first toss only).

The number next to some points indicates repeated results; for example, two 6s, five 3s, three 2s, and two 1s turned up for the white dice on the first toss. The middle square on the bars represents the mean value for the dice subgroup. The bars are set to provide 95% confidence. They obviously overlap, thus confirming the lack of significance in the outcome of the colored dice experiment. However, in this case, you already know this from the insignificant F-test.

The formula for the LSD, with equal sample sizes (n), is

$$LSD = t_{critical} \square s_{pooled}\sqrt{2/n}$$

This formula is derived from the equation for a confidence interval shown in Chapter 1. As discussed earlier, "t" is a factor that depends on the confidence desired and the degrees of freedom generated for the estimate of standard deviation, pooled from all the subgroups.

BE CAREFUL WHEN USING LSD

Using LSD without the protection of a significant F-test could be a real "downer" because of the risk of getting false positive differences. For example, if you were to compare 10 things two at a time, you would get 45 combinations. Chances are very good that by arbitrarily focusing only on the highest and lowest outcomes you would get a significant test even from a random set of data. So, just say “no” to pair-wise testing before conducting the F-test. (There are a number of special procedures for multiple pair-wise testing in the reference by George Box et al. in the Recommended Readings section at the back of this book.)

If you are tripping badly on statistics at this point, consider mellowing out with “Lucy in the Sky with Diamonds” by the Beatles, evidently inspired by this LSD procedure.

A sample calculation for LSD is provided in part two of this example, which illustrates what happens when you get a significant F-test. Read on for a thrilling conclusion to this case study!

Catching Cheaters with a Simple Comparative Experiment

To add some spice to this whole discussion, let’s assume that a nefarious gamer tinkers with the dice in an attempt to gain an advantage. Our villain can do this in any number of ways: by doctoring the surfaces, rounding the edges, or “loading” the dice with tiny weights. The actual cheating method will not be revealed yet, just the results seen in Table 2.2. The data from just this one toss were sufficient to catch the cheaters.

THE ONE THING CERTAIN ABOUT GAMBLING IS THAT SOMEONE WILL TRY TO CHEAT

Archeologists discovered several pairs of sandstone dice in an ancient Egyptian tomb. These dice, now residing in a Chicago museum, are weighted to favor twos and fives. It seems that loaded dice are as old as the pyramids. The statisticians have a term that relates to unfairly favoring certain events over others—it’s called *bias*. It’s safe to say that the objective is to eliminate bias if at all possible.

(From “The Roll of the Dice,” *Technology Magazine,* March 1998.)

> *“If the dice roll long enough, the man who knows what numbers are favored is going to finish with a fatter bankroll than when he started.”*
>
> **Gambling authority John Scarne**

Table 2.2 Frequency distribution for 56 rolls of the doctored (?) dice

Result	*White (1)*	*Blue (2)*	*Green (3)*	*Purple (4)*
6	6	6/6/6	6/6/6/6/6/6	6
5	5	5/5	5/5/5/5/5	5/5
4	4	4/4/4	4	4/4/4
3	3/3	3/3/3/3	3	3
2	2/2/2/2/2	2		2/2
1	1/1/1/1	1	1	1/1/1/1/1
Mean ($\bar{Y}$)	2.50	3.93	4.93	2.86
Var. (s^2)	2.42	2.38	2.07	3.21

The equation expressed in terms of statistical functions is

$$F = \frac{ns_{\bar{Y}}^2}{s_{pooled}^2} = \frac{14 * Var(2.50, 3.93, 4.93, 2.86)}{Mean(2.42, 2.38, 2.07, 3.21)} = \frac{14 * 1.21}{2.52} = 6.71$$

A reminder: The value of 14 for n (sample size) represents the number of dice of each color.

F-tables for various levels of risk are provided in Appendix 1. You can use these to determine whether your results are significant. However, you must first determine how much information, or degrees of freedom (df), went into the calculations for the variances. Calculation of the df is not complicated, but it can be tedious. See boxed text below if you want the details for this case, but the appropriate df are 3 (4 minus 1) and 52 (4 times [14 minus 1]) for the variances between treatments (numerator of F) and within treatments (denominator), respectively. If you go to column 3 and row 40 and row 60 (there is no row for 52) for the 5% table in Appendix 1, you will see F-values of 2.839 and 2.758, respectively. To be conservative, let’s use the value of 2.839. Because this critical value is exceeded by the actual F of 6.71, you can be more than 95% confident of significance for the test. Just out of curiosity, look up the critical F for 1% risk. You should find a value of about 4.2, which is still less than the actual F, thus indicating significance at greater than 99% confidence. It really looks bad for the cheaters!

DEGREES OF FREEDOM: WHO NEEDS 'EM?

(WARNING: STATISTICAL DETAILS FOLLOW)

In the two dice experiments, the variance for the treatments (numerator of F) is based on four numbers (means for each color of dice), so just subtract 1 to get the df of 3. The pooled variance within each treatment (denominator of F) is accumulated from 4 groups of 14 results (56 total), but 4 df must be subtracted for calculating the means, leaving 52 df for estimating error.

With the significant F-test as protection, we now use LSD as a comparative tool:

$$LSD = t_{critical} \Box s_{pooled}\sqrt{2/n}$$

where the pooled standard deviation (s_{pooled}) is calculated by taking the square root of the average variance:

$$s_{pooled} = \sqrt{(2.42+2.38+2.07+3.21)/4} = \sqrt{2.52} = 1.59$$

Recall that this estimate of error is based on 52 degrees of freedom (df). As a rough rule of thumb, when you have more than 30 df, which we easily exceed, the t-distribution becomes approximately normal. In Figure 1.5 (Chapter 1), you saw that about 95% of a population falls within two standard deviations. Thus, a reasonable approximation for the critical t at 95% confidence will be 2. Plugging in the subgroup size (n) of 14, we get:

$$LSD \cong 2 * 1.59\sqrt{2/14} = 1.20$$

The LSD of 1.2 is exceeded by the 1.43 difference between white and blue (3.93 – 2.50). Therefore, this comparison is statistically significant. You can see this on Figure 2.2: The LSD bars for the white and blue groups of dice do not overlap. The graph reveals that the blue and green dice are rolling much higher than the white and purple dice. Further tosses presumably would strengthen this finding, but were deemed unnecessary.

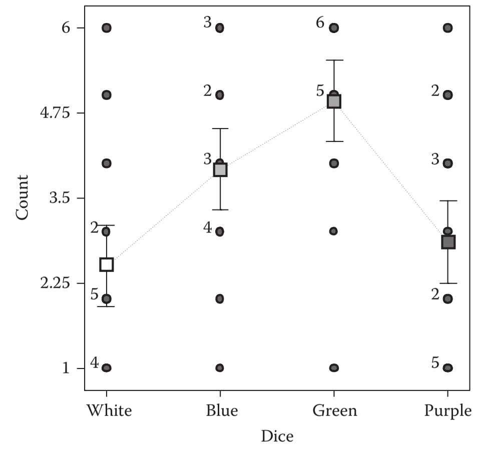


Figure 2.2 Effects plot for unfair dice game.

In this case, two cheaters (playing the blue and green dice) conspired against the other two (white and purple). While one partner in crime distracted the victims, the other turned most of the white and purple dice down one dot and most of the blue and green dice up one dot. To allay suspicion, they did not change a die that was the last of its kind, for example, the last 3. This all had to be done in a hurry, so mistakes were made which contributed to the overall variability in response. Despite this, the scheme proved effective, perhaps too much so, because all but young children would see through it fairly quickly. Professional cheaters would be much more subtle. Nevertheless, given enough data, statistical analysis would reveal the effect.

Blocking Out Known Sources of Variation

Known sources of variation caused by changes in personnel, materials, or machinery can be a real nuisance. Fortunately for us, statisticians, such as Fisher, developed techniques to "block" out these nuisance variables. In a landmark field trial on barley (used for making beer) in the authors' home state of Minnesota, agronomists grew 10 varieties of the crop at 6 sites in the early 1930s. Within each block of land, they planted barley in random locations. The yields varied considerably due to variations in soil and climate, but the ranking of barley types remained remarkably consistent.

This is called a *randomized block* experiment. A good rule for DOE is to block what you can and randomize what you cannot. There are many ways to incorporate blocking in a DOE, but we will illustrate only the more common and simple approaches.

Blocking is an especially effective tool for experiments that involve people. Each person will behave differently, but in many cases a consistent pattern will emerge. George Box et al. (see Recommended Readings) describe a clever approach for blocking out variation caused by the natural differences between boys. The objective is to measure wear of two alternative raw materials for shoe soles. The problem is that boys vary tremendously in terms of energy, from "hyperactives" at one end of the normal curve to TV-viewing "couch potatoes" at the other extreme. The resulting variation in shoe wear caused by boy-to-boy differences was blocked by giving each boy one shoe of each material, applied on a random basis to the left or right shoe. A clear difference in materials emerged despite an overwhelming variation from one boy to the next.

The following example illustrates the procedure for blocking. It involves the same dice used in previous studies, but with different rules. A group of preschool children were asked to pick out and stack three dice with 1, 2, 3, 4, 5, or 6 dots (any color). Because young children develop at differing rates, the time needed to accomplish this task varied widely. However, by repeating the entire exercise for each child, the differences between children were blocked out. The final results, illustrated in Table 2.3, were somewhat surprising.

LEARNING FROM SIMPLE COMPARATIVE EXPERIMENTATION

The famous child psychologist, Jean Piaget, suggested that thinking is a flexible, trial-and-error process. He observed that children, at their earliest stage of development, experiment with objects and learn from experience. For example, they often find it difficult at first to grasp the concept of volume. To confirm this observation, preschoolers in a class taught by Mark's wife Karen were shown a tall, narrow container filled with water. They were asked to guess what would happen when the teacher poured the water into a broad and shallow bowl of equal volume. Here are the more interesting hypotheses:

> *"It will be a volcano—it will start on fire."*
> *"It won't fit. It will explode all over."*
> *"It will be juice."*
> *"It will turn pink."*
> *"It will bubble."*
> *"It will turn red. It got full! It's magic. The teacher's magic!"*

Only a few of the children guessed correctly that water would not overflow. Presumably many of the others learned from their mistakes, but for some of the preschoolers it took a few repetitions and further cognitive development for the concept of volume to completely sink in.

> *"When consequences and data fail to agree, the discrepancy can lead, by a process called induction, to modification of the hypothesis."*
>
> **Box, Hunter, and Hunter**

Table 2.3 shows the data for four children participating in the dice experiment. You can assume that each result is actually an average from several trials, because individual times would normally vary more than those shown. The last row shows how each child's time differs from the overall average of 36.9 seconds. For example, the K1 child completed the tasks 18.47 seconds faster on average than the group as a whole. These differences are substantial relative to the differences due to the various dot patterns. For example, child K3 evidently required some coaxing to do the task, because the times far exceed those of the other individuals.

The F-test with no correction for blocks is shown below:

$$F = \frac{ns_{\bar{Y}}^2}{s_{pooled}^2} = \frac{4 * Var(20.6, 25.3, 46.5, 33.1, 56.1, 39.6)}{367.8} = \frac{4 * 176.9}{367.8} = 1.92$$

This F-value is based on 5 df for the between means comparison (numerator) and 18 df for the within means comparison. The critical F for 5% risk is 2.773, so the actual F fails to reject the null hypothesis at the 95% confidence level. In other words, the results are not significant. Figure 2.3

Table 2.3 Times (in seconds) for stacking three dice with same number of dots

Child (block)	*K1*	*K2*	*K3*	*K4*	*Within-Dot Mean*	*Within-Dot Variance*
1 dot	7.2	13.2	39.9	22.2	20.6	203.3
2 dots	10.0	21.6	45.3	24.1	25.3	216.8
3 dots	25.6	36.2	79.7	44.6	46.5	548.3
4 dots	15.2	30.0	54.5	32.9	33.1	262.5
5 dots	33.0	48.1	90.8	52.7	56.1	604.7
6 dots	19.5	32.0	65.0	42.0	39.6	370.9
Mean	18.4	30.2	62.5	36.4	36.9	367.8
Difference	–18.47	–6.71	25.64	–0.46	0.0	

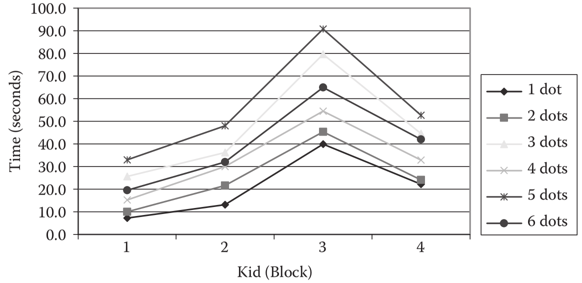


Figure 2.3 Scatter plot of times to stack dice.

reveals a pattern, but it's obscured statistically by the variability between children. How can we properly recognize the pattern statistically?

The answer is to remove the variability between children by subtracting the difference shown on the last line of Table 2.3 from the raw data in the associated columns. This is how you can block out a known source of variation. The results are given in Table 2.4.

The means for each dot pattern (1 through 6) remain the same, but notice the huge reduction in overall within-dot variance after removal of blocks. It drops more than 10-fold from 367.8 down to 19.8. Obviously this will have a very beneficial impact on the F-ratio. The calculation is

$$F = \frac{4 * 176.9}{19.8 * 18 / 15} = 29.8$$

Table 2.4 Times (in seconds) for stacking three dice (corrected for blocks)

Kid (block)	*K1*	*K2*	*K3*	*K4*	*Within-Dot Mean*	*Within-Dot Variance*
1 dot	25.7	19.9	14.3	22.7	20.6	23.6
2 dots	28.5	28.3	19.7	24.6	25.3	17.0
3 dots	44.1	42.9	54.0	45.1	46.5	25.7
4 dots	33.7	36.7	28.8	33.4	33.1	10.5
5 dots	51.5	54.8	65.2	53.2	56.1	38.0
6 dots	38.0	38.7	39.4	42.5	39.6	3.9
Mean	36.9	36.9	36.9	36.9	36.9	19.8
Kid Diff.	0.00	0.00	0.00	0.00		

Notice that the numerator remains unchanged from the unblocked case, because the treatment means are unaffected. The degrees of freedom for the numerator also remain the same as before at 5 df. However, the degrees of freedom for the denominator of F must be reduced because we took the block means into account. Specifically, the calculation of the 4 block means causes a loss of 3 df (n minus 1 with n = 4) for calculating error, so only 15 df remain. This loss of information for estimating error must be accounted for when calculating F (notice the 18/15 correction to the within-dot variance in the denominator) and when looking up the critical value for F. On the 5% table, you will find a value of 2.901 under column 5 and row 15. The actual F of 29.8 far exceeds this critical F, so the outcome is statistically significant. In this case, blocking proved to be the key to success.

Figure 2.4 shows the times as a function of the number of dots. It does not show the blocks, but they are accounted for in calculating the LSD bars. Aided by this tool of statistics, you can see the unexpected outcome: The children found it most difficult to pick out the 5-spot dice. They were very quick to identify the 1-spot dice, significantly so in relation to the 5 (also the 3, 4, and 6). The 3-spot seemed to cause more trouble than the 4-spot dice.

Many more children would need to be tested to confirm these findings. The point of this exercise is simply to show the advantage of blocking out known sources of variation. Blocking can be applied to any kind of DOE, not just simple comparisons such as those illustrated in this chapter. It often reveals results that would otherwise be obscured.

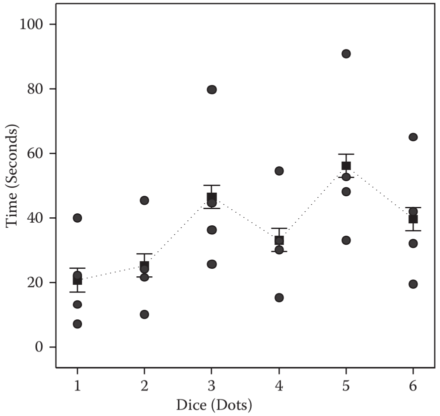


Figure 2.4 Effects plot of times to stack dice (corrected for blocks).

Practice Problems

Problem 2.1

Three bowlers must compete for the last position on the company team. They each bowl six games (see data in Table 2.5). The best bowler will fill the opening, with the runner-up as the alternate. The worst bowler is out.

Assume that you are the captain. You know better than to simply pick the bowler with the highest average score and drop the person who scores lowest. Maybe it's a fluke that Mark scored highest and Pat's score is low.

Table 2.5 Bowling scores

Game	*Pat*	*Mark*	*Shari*
1	160	165	166
2	150	180	158
3	140	170	145
4	167	185	161
5	157	195	151
6	148	175	156
Mean	153.7	178.3	156.2

Are the scores significantly different, given the variability in individual scores? Do an analysis of variance. If it produces significant results, the choice may be clear; otherwise, the bowlers must return to the alley for more games. (Suggestion: Use the software provided with the book. First do the one-factor tutorial that comes with the program. It's keyed to the data in Table 2.5. See the end of the book for software instructions and details on the tutorial files.)

Problem 2.2

To double-check the dice comparison detailed in Table 2.1, the players tossed the dice a second time. The data is shown below in Table 2.6.

Do an analysis of variance for the second toss according to the procedures shown earlier. Do you have any reason to dispute the earlier conclusion that color of dice does not significantly affect the outcome (so long as no one cheats)? (Suggestion: Follow up by using the software provided with the book. Set up a one-factor design similar to that shown in the tutorial that comes with the program. After doing the ANOVA, generate an effects plot with the LSD bars.)

Table 2.6 Frequency distribution for 56 rolls of the dice, second toss

Result	*White (1)*	*Blue (2)*	*Green (3)*	*Purple (4)*
6	6/6	6/6/6	6/6/6	6/6/6
5	5/5/5/5	5	5/5/5	5/5/5
4	4/4/4	4/4	4/4	4/4/4
3	3/3/3	3/3/3/3	3/3	3
2	2/2	2/2/2	2/2	2
1		1	1/1	1/1/1
Mean ($\bar{Y}$)	___	___	___	___
Var. (s^2)	___	___	___	___

AGE-OLD ADVICE: IF YOU FEEL IT'S REAL, DO IT AGAIN

As evidenced by the following quote passed along from *Games, Gods and Gambling: A History of Probability and Statistical Ideas* by Florence

Nightingale David (Dover Publications, 1998, p. 24), the value of replication for confirming things has been known for some time:

> *"Suppose you made a hundred casts and the [same] throw appeared a hundred times; could you call that accidental?"*
>
> **Quintus Tullius Cicero (102 – 43 BCE)**

Table 2.7 Data from four incoming shipments

Lot	*Data*
A	61,61,57,56,60,52,62,59,62,67,55,56,52,60,59,59,60,59,49,42,55,67,53,66,60
E	56,56,61,67,58,63,56,60,55,46,62,65,63,59,60,60,59,60,65,65,62,51,62,52,58
I	62,62,72,63,51,65,62,59,62,63,68,64,67,60,59,59,61,58,65,64,70,63,68,62,61
M	70,70,50,68,71,65,70,73,70,69,64,68,65,72,73,75,72,75,64,69,60,68,66,69,72

Problem 2.3

In problem 1.1 (Chapter 1) you analyzed performance of a supplier. Let's assume that you get three more incoming shipments. The delivery person observes that something looks different than before. You decide to investigate using analysis of variance. The data are collected in Table 2.7.

Do you see a significant difference between lots? If so, which ones stand out, either higher or lower than the others? (Suggestion: Use the software provided with the book. Set up a one-factor design similar to that shown in the tutorial that comes with the program. After doing the ANOVA, generate an effects plot with the LSD bars.)

Problem 2.4

A research group in the military decided to test a new fabric for making uniforms. However, the researchers realized that if they made two pairs of pants with differing material, only one pair could be worn at a time. Even if they restricted the testing to just one subject, variations could occur due to changes in that person's daily activities. They knew that the problem would get much worse as the testing expanded to more subjects, because variability between people would be severe. The experimenters overcame this problem by making special pants with the current fabric for one leg and the new fabric for the other leg. Conceivably, one leg might consistently get more wear (probably the right one), so they alternated which material got sewn on which leg.

Table 2.8 Wear ratings for differing fabric for military pants

Block (Subject)	*Old Fabric*	*New Fabric*
1	8.1	9.0
2	5.1	5.8
3	6.8	7.2
4	8.5	8.8
5	6.7	7.6
6	4.4	4.3
7	6.4	6.9
8	6.7	7.3
9	5.8	6.1

The researchers randomly assigned these special pants to nine subjects. (No problem finding "volunteers" in the military.) The results are shown in Table 2.8. The response is a subjective rating of wear on a scale of 1 to 10, the higher the better, assessed by a panel of inspectors and then averaged.

No two subjects participated in exactly the same activities; however, both materials for any one person endured the same activities. By blocking the experiments by subject, the person-to-person variability due to differing activities (wear conditions) can be set aside during the analysis. Your mission is to determine whether the fabric wears differently, and, if so, which one lasts longest. (Suggestion: Set this up as a one-factor, blocked design. The ANOVA, if done properly, will remove the block variance before calculating the F-test on the effect of the fabric. If the test is significant, make an effects plot.)

Chapter 3

Two-Level Factorial Design

If you do not expect the unexpected, you will not find it.

Heraclitus

If you have already mastered the basics discussed in Chapters 1 and 2, you are now equipped with very powerful tools to analyze experimental data. Thus far, we have restricted discussion to simple, comparative one-factor designs. We now introduce *factorial design*—a tool that allows you to experiment on many factors simultaneously. The chapter is arranged by an increasing level of statistical detail. The latter portion becomes more mathematical, but the added effort required to study these details will pay off in increased understanding of the statistical framework and more confidence when using this powerful tool.

The simplest factorial design involves two factors, each at two levels. The top part of Figure 3.1 shows the layout of this two-by-two design, which forms the square "X-space" on the left. The equivalent one-factor-at-a-time (OFAT) experiment is shown at the upper right.

The points for the factorial designs are labeled in a "standard order," starting with all low levels and ending with all high levels. For example, runs 2 and 4 represent factor A at the high level. The average response from these runs can be contrasted with those from runs 1 and 3 (where factor A is at the low level) to determine the effect of A. Similarly, the top runs (3 and 4) can be contrasted with the bottom runs (1 and 2) for an estimate of the effect of B.

Later, we will get into the mathematics of estimating effects, but the point to be made now is that a factorial design provides contrasts of averages,

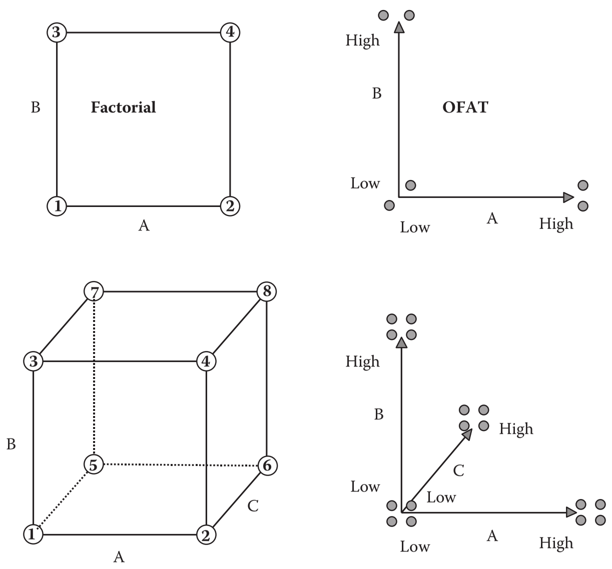


Figure 3.1 Two-level factorial versus one-factor-at-a-time (OFAT).

thus providing statistical power to the effect estimates. The 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 6 runs versus only 4 for the two-level design.

HAY!

Perhaps the earliest known multifactor experiment of this kind was laid out in 1898 by agronomists from Newcastle University in the United Kingdom, who varied the amounts of nitrogen (N), potash (K), and phosphate (K) via a two-level factorial design on Palace Leas meadow. In this field (last ploughed during the Napoleonic Wars), they applied the eight combinations of N, P, and K into individual plots. The results led to an impressive increase in crop yields. (Source: Shirley Coleman, et. al. 1987. The effect of weather and nutrition on the yield of hay from Palace Leas meadow hay plots, at Cockle Park Experimental Farm, over the period from 1897 to 1980. *Grass and Forage Science* 42: 353–358.)

The advantage of factorial design becomes more pronounced as you add more factors. For example, with three factors, the factorial design requires only 8 runs (in the form of a cube) versus 16 for an OFAT experiment with equivalent power. In both designs (shown at the bottom of Figure 3.1), the effect estimates are based on averages of 4 runs each: right-to-left, top-to-bottom, and back-to-front for factors A, B, and C, respectively. The relative efficiency of the 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 offers two additional advantages over OFAT:

- Wider inductive basis, i.e., it covers a broader area or volume of X-space from which to draw inferences about your process.
- It reveals "interactions" of factors. This often proves to be the key to understanding a process, as you will see in the following case study.

Two-Level Factorial Design: As Simple as Making Microwave Popcorn

We will illustrate the basic principles of two-level factorial design via an example.

What could be simpler than making microwave popcorn? Unfortunately, as everyone who has ever made popcorn knows, it's nearly impossible to get every kernel of corn to pop. Often a considerable number of inedible "bullets" (unpopped kernels) remain at the bottom of the bag. What causes this loss of popcorn yield? Think this over the next time you stand in front of the microwave waiting for the popping to stop and jot down a list of all the possible factors affecting yield. You should easily identify 5 or even 10 variables on your own; many more if you gather several colleagues or household members to "brainstorm."

In our example, only three factors were studied: brand of popcorn, time of cooking, and microwave power setting (Table 3.1). The first factor, brand, is clearly "categorical"—either one type or the other. The second factor, time, is obviously "numerical," because it can be adjusted to any level. The third factor, power, could be set to any percent of the total available, so it's also numerical. If you try this experiment at home, be very careful to do some range finding on the high level for time (see related boxed text below). Notice that we have introduced the symbols of minus (–) and plus (+) to designate low and high levels, respectively. This makes perfect sense for numerical factors, provided you do the obvious and make the lesser value correspond to the low level. The symbols for categorical factor levels are completely arbitrary, although perhaps it helps in this case to assign minus as "cheap" and plus as "costly."

Table 3.1 Test-factors for making microwave popcorn

Factor	*Name*	*Units*	*Low Level (–)*	*High Level (+)*
A	Brand	Cost	Cheap	Costly
B	Time	Minutes	4	6
C	Power	Percent	75	100

BE AGGRESSIVE IN SETTING FACTOR LEVELS, BUT DON'T BURN THE POPCORN

One of the most difficult decisions for design of experiments (DOE), aside from which factors to choose, is what levels to set them. A general rule is to set levels as far apart as possible so you will more likely see an effect, but not exceed the operating boundaries. For example, test pilots try to push their aircraft to the limits, a process often called pushing the envelope. The trick is not to break the envelope, because the outcome may be "crash and burn." In the actual experiment on popcorn (upon which the text example is loosely based), the experiment designer (Mark) set the upper level of time too high. In the randomized test plan, several other combinations were run successfully before a combination of high time and high power caused the popcorn to erupt like a miniature volcano, emitting a lava-hot plasma of butter, steam, and smoke. Alerted by the kitchen smoke alarm, the family gathered to observe the smoldering microwave oven. Mark was heartened to hear the children telling his wife not to worry because "in science, you learn from your mistakes." Her reaction was not as positive, but a new microwave restored harmony to the household. Lessen learned—always conduct highly controlled pretrials on extreme combinations of factors!

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 lower the weight, the better.

The results from running all combinations of the chosen factors, each at two levels, are shown in Table 3.2. Taste ranged 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 even get a taste. Obviously, this particular 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.

ALWAYS RANDOMIZE YOUR RUN ORDER

You must randomize the order of your experimental runs to satisfy the statistical requirement of independence of observations. Randomization acts as insurance against the effects of lurking time-related variables, such as the warm-up effect on a microwave oven. For example, let's say you forget to randomize and first run all low levels of a factor and then all high levels of a given factor that actually creates no effect on response. Meanwhile, an uncontrolled variable causes the response to gradually increase. In this case, you will mistakenly attribute the happenstance effect to the nonrandomized factor. By randomizing the order of experimentation, you greatly reduce the chances of such a mistake. Select your run numbers from a table of random numbers or mark them on slips of paper and simply pull them blindly from a container. Statistical software can also be used to generate random run orders.

> *"Designing an experiment is like gambling with the devil: only a random strategy can defeat all his betting systems."*
>
> **Sir Ronald Fisher**

The first column in Table 3.2 lists the standard order, which can be cross-referenced to the labels on the three-factor cube in Figure 3.1. We also placed the mathematical symbols of minus and plus, called "coded factor levels," next to the "actual" levels at their lows and highs, respectively. Before proceeding with the analysis, it will be very helpful to re-sort the test matrix on the basis of standard order, and list only the coded factor levels. We also want to dispense with the names of the factors and responses, which just get in the way of the calculations, and show only their mathematical symbols. You can see the results in Table 3.3.

Table 3.2 Results from microwave popcorn experiment

Standard	*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

Table 3.3 Test matrix in standard order with coded levels

Standard	*Run*	*A*	*B*	*C*	*Y_1*	*Y_2*
1	8	–	–	–	74	3.1
2	1	+	–	–	75	3.5
3	2	–	+	–	71	1.6
4	4	+	+	–	80	1.2
5	3	–	–	+	81	0.7
6	5	+	–	+	77	0.7
7	7	–	+	+	42	0.5
8	6	+	+	+	32	0.3
Effect $\mathbf{Y_1}$		–1.0	–20.5	–17.0	66.5	
Effect $\mathbf{Y_2}$		–0.05	–1.1	–1.8		1.45

The column labeled "Standard" and the columns for A, B, and C form a template that can be used for any three factors that you want to test at two levels. The standard layout starts with all minus (low) levels of the factors and ends with all plus (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 look them up in statistical handbooks.

ORTHOGONAL ARRAYS: WHEN YOU HAVE LIMITED RESOURCES, IT PAYS TO PLAN AHEAD

The standard two-level factorial layout shown in Table 3.3 is one example of a carefully balanced “orthogonal array.” Technically, this means that there is no correlation among the factors. You can see this most easily by looking at column C. When C is at the minus level, factors A and B contain an equal number of pluses and minuses, thus, their effect cancels. The same result occurs when C is at the plus level. Therefore, the effect of C is not influenced by factors A or B. The same can be said for the effects of A and B and all the interactions as well. The authors have limited this discussion of orthogonal arrays to those that are commonly called the *standard arrays* for two-level full and fractional factorials. However, you may come across other varieties of orthogonal arrays, such as Taguchi and Plackett-Burman. Note, however, that any orthogonal test array is much preferred to unplanned experimentation (an oxymoron). Happenstance data are likely to be highly correlated (nonorthogonal), which makes it much more difficult to sort out the factors that really affect your response. (For an in-depth explanation of the dangers in dealing with nonorthogonal matrices, see Chapter 2, Lessons to Learn from Happenstance Regression, in *RSM Simplified*.)

Let’s begin the analysis by investigating the “main effects” on the first response (Y_1)—taste. It helps to view the results in the cubical factor space. We will focus on factor A (brand) first (Figure 3.2).

The right side of the cube contains all the runs where A is at the plus level (high); on the left side, the factor is held at the minus level (low).

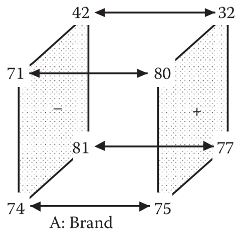


Figure 3.2 Cube plot of taste ratings with focus on brand (Factor A).

Now, simply average the highs and the lows to determine the difference or contrast. This is the effect of Factor A. Mathematically, the calculation of an effect is expressed as follows:

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

where the ns refer to the number of data points you collected at each level. The Ys refer to the associated responses. You can pick these off the plot or from the matrix itself. For A, the effect is

$$Effect = \frac{75 + 80 + 77 + 32}{4} - \frac{74 + 71 + 81 + 42}{4} = 66 - 67 = -1$$

In comparison to the overall spread of results, it looks like A (brand) has very little effect on taste. Continue the analysis by contrasting the averages from top-to-bottom and back-to-front to get the effects of B and C, respectively. Go ahead and do the calculations if you like. The results are –20.5 for B and –17 for C. The impact, or “effect,” of factors B (power) and C (time) are much larger than that of A (the brand of popcorn).

Before you jump to conclusions, however, consider the effects caused by interactions of factors. The full-factorial design allows estimation of all three two-factor interactions (AB, AC, and BC) as well as of the three-factor interaction (ABC). Including the main effects (caused by A, B, and C), this brings the total to seven effects; the most you can estimate from the 8-run factorial design, because one degree of freedom is used to estimate the overall mean. Table 3.4 lists all seven effects. The main effects calculated earlier are listed in the A, B, and C columns.

The pattern of pluses and minuses for interaction effects is calculated by multiplying the parent terms. For example, the AB column is the product of columns A and B, so for the first standard row, the combination of –A times –B produces +AB. Remember that numbers with like signs, when multiplied, produce a plus; whereas multiplying numbers with unlike signs produces a minus. The entire array exhibits a very desirable property of balance called *orthogonality* (see related sidebar in the above boxed text).

Now, it’s just a matter of computing the effects using the general formula shown previously. The results are shown on the bottom line of Table 3.4.

Table 3.4 Complete matrix, including interactions, with effects calculated

Standard	*Main Effects*			*Interaction Effects*				*Response*
	A	*B*	*C*	*AB*	*AC*	*BC*	*ABC*	Y_1
1	–	–	–	+	+	+	–	74
2	+	–	–	–	–	+	+	75
3	–	+	–	–	+	–	+	71
4	+	+	–	+	–	–	–	80
5	–	–	+	+	–	–	+	81
6	+	–	+	–	+	–	–	77
7	–	+	+	–	–	+	–	42
8	+	+	+	+	+	+	+	32
Effect	–1.0	–20.5	–17.0	0.5	–6.0	–21.5	–3.5	66.5

Notice that the interaction effect of BC is even greater on an absolute scale than its parents B and C. In other words, the combination of time (B) and power (C) produces a big (negative) impact on taste. With that as a clue, look more closely at the response data (Y_1). Notice the big drop-off in taste when both B and C are at their high levels. We will investigate this further after sorting out everything else.

On an absolute value scale, the other interaction effects range from near 0 (for AB) to as high as 6 (for AC). Could these just be chance occurrences due to normal variations in the popcorn, the tasting, the environment, and the like? To answer this question, let's go back to a tool discussed at the end of Chapter 1: the normal plot. Then we can see whether some or all of the effects vary normally. Ideally, we will discover one or more effects at a significant distance from the remainder. Otherwise we have wasted a lot of experimental effort chasing noise from the system.

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 paper called the *half-normal*, which is literally based on the positive half of the full normal curve. (Imagine cutting out the bell-shaped curve and folding it in half at the mean.) As before, the vertical (y) 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.

Table 3.5 Values to plot on half-normal probability paper

Point	*Effect*	*Absolute Value of Effect*	*Cumulative Probability*
1	AB	|0.5|	7.14%
2	A	|–1.0|	21.43%
3	ABC	|–3.5|	35.71%
4	AC	|–6.0|	50.00%
5	C	|–17.0|	64.29%
6	B	|–20.5|	78.57%
7	BC	|–21.5|	92.86%

Remember that, before plotting this data on the probability paper, you must:

1. Sort the datapoints (in this case seven effects) in ascending order.
2. Divide the 0 to 100% cumulative probability scale into (seven) equal segments.
3. Plot the data at the midpoint of each probability segment.

In this case, each probability segment will be approximately 14.28% (100/7). The lowest weight will be plotted at 7.14%, which is the midpoint of the first segment. Table 3.5 shows this combination and all the remaining ones.

Now all we need to do is plot the absolute values of the effect on the x-axis versus the cumulative probabilities on the specially scaled y-axis on half-normal paper (Figure 3.3).

Figure 3.4 shows the completed half-normal plot for the effects on taste of popcorn. This particular graph has some features you will not usually see:

- A half-normal curve for reference
- A "dot-plot" on the x-axis representing the actual effects projected down to the x-axis number line

Notice that the biggest three effects fall well out on the tail of the normal curve (to the right). These three effects (C, B, and BC) are most likely significant in a statistical sense. We wanted to draw attention to these big effects, so we labeled them. Observe the large gap before you get to the next lowest

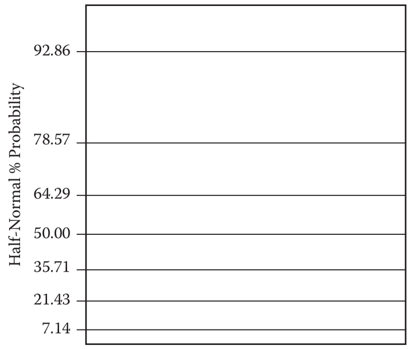


Figure 3.3 Blank half-normal paper (set up for plotting seven effects).

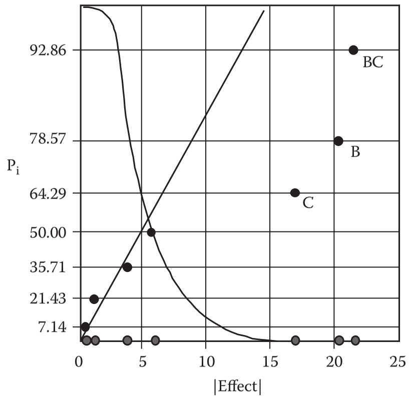


Figure 3.4 Half-normal plot of effects for taste (curve and dot-plot added for reference).

effect. From this point on, the effects (AC, ABC, A, and AB—from biggest to smallest, respectively) fall in line, which represents the normal scatter. We deliberately left these unlabeled to downplay their importance. These four trivial effects (nearest 0) will be used as an estimate of error for the upcoming analysis of variance (ANOVA).

The pattern you see in Figure 3.4 is very typical: The majority of points fall 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, if anything, is significant.

THE VITAL FEW VERSUS THE TRIVIAL MANY

A rule of thumb, called sparsity of effects, says that, in most systems, only 20% of the main effects (MEs) and two-factor interactions (2 fi) will be significant. The other ME and 2 fis, as well as any three-factor interactions (3 fi) or greater will vary only to the extent of normal error. (Remember that the effects are based on averages, so their variance will be reduced by a factor of n.) This rule of thumb is very similar to that developed a century ago by economist Vilfredo Pareto, who found that 80% of the wealth was held by 20% of the people. Dr. Joseph Juran, a preeminent figure in the twentieth-century quality movement, applied this 80/20 rule to management: 80% of the trouble comes from 20% of the problems. He advised focusing effort on these “vital few” problems and ignoring the “trivial many.”

Table 3.6 Effects calculated for second response (bullets)

Standard	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*	Y_2
1	–	–	–	+	+	+	–	3.1
2	+	–	–	–	–	+	+	3.5
3	–	+	–	–	+	–	+	1.6
4	+	+	–	+	–	–	–	1.2
5	–	–	+	+	–	–	+	0.7
6	+	–	+	–	+	–	–	0.7
7	–	+	+	–	–	+	–	0.5
8	+	+	+	+	+	+	+	0.3
Effect	–0.05	–1.1	–1.8	–0.25	–0.05	0.80	0.15	1.45

Let’s apply this same procedure to the second response for microwave popcorn: the weight of the bullets. In the last row of Table 3.6, the seven effects are calculated using the formula shown earlier:

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

Table 3.7 shows the effects ranked from low to high in absolute value, with the corresponding probabilities.

Table 3.7 Values to plot on half-normal plot for bullets

Point	*Effect*	*Absolute Value of Effect*	*Cumulative Probability*
1	A	|−0.05|	7.14%
2	AC	|−0.05|	21.43%
3	ABC	|0.15|	35.71%
4	AB	|−0.25|	50.00%
5	BC	|0.80|	64.29%
6	B	|−1.10|	78.57%
7	C	|−1.80|	92.86%

Notice that the probability values are exactly the same as for the previous table on taste. In fact, these values apply to any three-factor, two-level design, if you successfully perform all 8 runs and gather the response data.

Figure 3.5 shows the resulting plot (computer generated) for bullets, with all effects labeled so you can see how it's constructed. For example, the smallest effects, A and AC, which each have an absolute value of 0.05, are plotted at 7.1 and 21.4% probability. (When effects are equal, the order is arbitrary.) Next comes effect ABC at 35.7%, and so on.

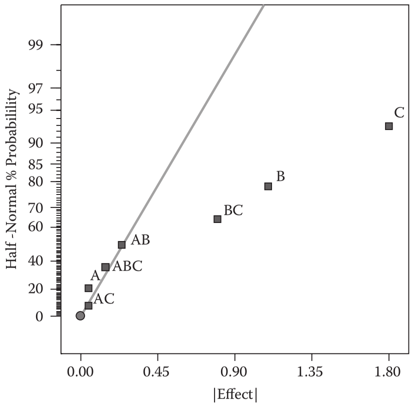


Figure 3.5 Half-normal plot of effects for bullets (all effects labeled).

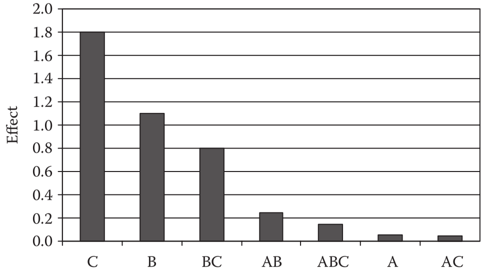


Figure 3.6 Pareto chart of effects for bullets (all effects labeled).

Notice how four of the effects (AB, ABC, A, and AC) fall in a line near zero. These effects evidently vary only due to normal causes—presumably as a result of experimental error (noise in the system), so they are probably insignificant. You will almost always find three-factor interactions, such as ABC, in this normal population of trivial effects. Interactions of four or more factors are even more likely to fall into this near-zero group of effects.

The effects of B, C, and BC are very big relative to the other effects. They obviously do not fall on the line. In a statistical sense, each of these three standout effects should be considered significant populations in their own right. In other words, we need to focus on factors B and C and how they interact (as BC) to affect the response of bullets.

Figure 3.6 offers a simpler view of the relative effects via an ordered bar graph called a Pareto chart, which serves as a graphic representation of the principle (also called the 80/20 rule) discussed above. This becomes manifest by the vital few bars at the left towering over the trivial many on the right. (See the Appendix to this chapter for a more sophisticated form of the Pareto chart that provides statistical benchmarks for assessing statistical significance of the effects.)

A word of caution: To protect against spurious outcomes, it is absolutely vital that you verify the conclusions drawn from the half-normal plots and Pareto charts by doing an analysis of variance (ANOVA) and the associated diagnostics of "residual error." As you will see later in this chapter, the statistics in this case pass the tests with flying colors. Please take our word on it for now. We will eventually show you how to generate and interpret all the statistical details, but it will be more interesting to jump ahead now to the effect plot.

How to Plot and Interpret Interactions

Interactions occur when the effect of one factor depends on the level of the other. They cannot be detected by traditional one-factor-at-a-time (OFAT) experimentation, so don't be surprised if you uncover previously undetected interactions when you run a two-level design. Very often, the result will be a breakthrough improvement in your system.

The microwave popcorn study nicely illustrates how to display and interpret an interaction. In this case, both of the measured responses are greatly impacted by the interaction of time and power, so it is helpful to focus on these two factors (B and C, respectively). Table 3.8 shows the results for the two responses: taste and bullets. These are actually averages of data from Table 3.3, which we have cross-referenced by standard order. For example, the first two experiments in Table 3.3 have both time and power at their low (minus) levels. The associated taste ratings are 74 and 75, which produces an average outcome of 74.5, as shown in Table 3.4.

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. However, when power is high (plus), the taste goes very bad—from 79 to 37. This is much clearer when graphed (Figure 3.7).

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). However, 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. Table 3.8 shows the average result to be 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.

Table 3.8 Data for interaction plot of microwave time versus power

Standard	*Time (B)*	*Power (C)*	*Taste (Y_1 Avg)*	*Bullets (Y_2 Avg)*
1,2	–	–	74.5	3.3
3,4	+	–	75.5	1.4
5,6	–	+	79.0	0.7
7,8	+	+	37.0	0.4

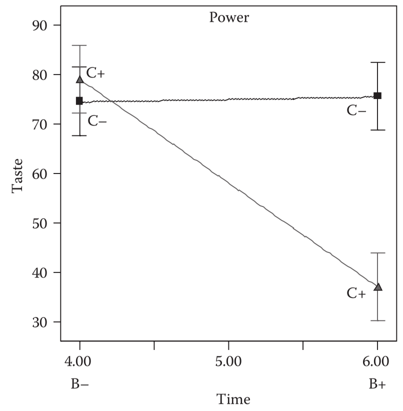


Figure 3.7 Interaction of time (B) versus power (C) on popcorn taste.

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.

TASTE IS IN THE MOUTH OF THE BEHOLDER

Before being rescaled, the popcorn taste was rated on a scale of 1 (worst) to 10 (best) by a panel of Minnesotans. The benchmarks they used reflect a conservative, Scandinavian heritage:

10: Just like lutefisk*
9: Not bad for you
8: Tastes like Mom's
7: Not so bad
6: Could be better
5: Could be worse
4: My spouse made something like this once
3: I like it, but …
2: It's different
1: Complete silence

* Dried cod soaked in lye that is rumored to be good for removing wallpaper—just throw the fish up by the ceiling and let it slide all the way down for broadest application.

If you are not from Minnesota, we advise that you use an alternative scale used by many sensory evaluators, which goes from 1 to 9, with 9 being the best. All nine numbers are laid out in line. The evaluator circles the number that reflects his or her rating of a particular attribute. To avoid confusion about orientation of the scale, we advise that you place sad (☹), neutral (😐), and happy (☺) faces at the 1, 5, and 9 positions on the number line, respectively. This is called a "hedonic" scale. Rating scales like this can provide valuable information on subjective responses, particularly when you apply the averaging power of a well-planned experiment.

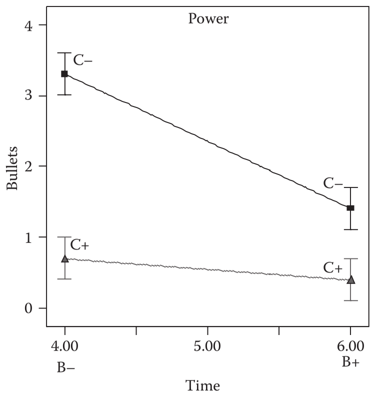


Figure 3.8 Interaction of time (B) versus power (C) on popcorn bullets.

Figure 3.8 shows how time and power interact to affect the other response, the bullets. The pattern differs from that for taste, but it again exhibits the nonparallel lines that are characteristic of a powerful two-factor interaction.

The effect of time on the weight of bullets depends on the level of power, represented by the two lines on the graph. On the lower line, notice the overlap in the least significant difference (LSD) bars at left versus right. This indicates that at high power (C+), there's not much, if any, effect. However, the story differs for the top line on the graph where power is set at its low level (C–). Here the LSD bars do not overlap, indicating that the effect of time is significant. Getting back to that bottom line, it's now obvious that when using the bullets as a gauge, it is best to make microwave popcorn at the highest power setting. However, recall that high time and high power resulted in a near-disastrous loss of taste. Therefore, for "multiresponse optimization" of the microwave popcorn, the best settings are high power at low time. The brand, factor A, does not appear to significantly affect either response, so choose the one that's cheapest.

Protect Yourself with Analysis of Variance (ANOVA)

Now that we have had our fun and jumped to conclusions on how to make microwave popcorn, it's time to do our statistical homework by performing the analysis of variance (ANOVA). Fortunately, when factorials are restricted to two levels, the procedure becomes relatively simple. We have already done the hard work by computing all the effects. To do the ANOVA, we must compute the sums of squares (SS), which are related to the effects as follows:

$$SS = \frac{N}{4}\left(Effect^2\right)$$

where N is the number of runs. This formula works only for balanced two-level factorials.

HOW SS SIMPLIFIES: A QUICK EXPLANATION

(Try following along with this text only if you really care where the sum of squares formula came from.) The difference from either the plus or minus results to the overall experimental response average (also known as the somewhat pompous *grand mean*) is equal to one-half of the effect. (See Figure 3.10 for illustration.) Half of the runs are plus and half are minus. Thus:

$$SS = SS_{plus} + SS_{minus} = N/2(Effect/2)^2 + N/2(Effect/2)^2 = 2\ (N/2(Effect/2)^2$$

$$= N/4(Effect)^2$$

So now you know.

The three largest effects (B, C, and BC) are the vital few that stood out on the half-normal plot. Their sum of squares are shown in the italicized rows in Table 3.9 below. For example, the calculation for sum of squares for effect B is

$$SS_B = \frac{8}{4}\left(-20.5^2\right) = 840.5$$

You can check the calculations for the sum of squares associated with effects C and BC. The outstanding effects will be incorporated in the "model" for predicting the taste response. (We provide more details on the model later.)

When added together, the resulting sum of squares provides the beginnings of the actual ANOVA. Here's the calculation for the taste response:

$$SS_{Model} = SS_B + SS_C + SS_{BC} = 840.5 + 578 + 924.5 = 2343$$

The smaller effects, which fell on the near-zero line, will be pooled together and used as an estimate of error called residual. The calculation for this taste response is

$$\begin{aligned} SS_{\mathrm{Re}sidual} &= SS_A + SS_{AB} + SS_{AC} + SS_{ABC} \\ &= \frac{8}{4}\left(-1^2\right) + \frac{8}{4}\left(0.5^2\right) + \frac{8}{4}\left(-6^2\right) + \frac{8}{4}\left(-3.5^2\right) \\ &= 2 + 0.5 + 72 + 24.5 = 99 \end{aligned}$$

The sum of squares for model and residual are shown in the first column of data in the ANOVA, shown in Table 3.9. The next column lists

Table 3.9 ANOVA for taste

Source	*Sum of Squares (SS)*	*Df*	*Mean Square (MS)*	*F Value*	*Prob > F*
Model	2343.0	3	781.0	31.5	<0.01
B	*840.5*	*1*	*840.5*	*34.0*	*<0.01*
C	*578.0*	*1*	*578.0*	*23.3*	*<0.01*
BC	*924.5*	*1*	*924.5*	*37.3*	*<0.01*
Residual	99.0	4	24.8		
Cor Total	2442.0	7			

the degrees of freedom (df) associated with the sum of squares (derived from the effects). Each effect is based on two averages, high versus low, so it contributes 1 degree of freedom (df) for the sum of squares. Thus, you will see 3 df for the three effects in the model pool and 4 df for the four effects in the residual pool. This is another simplification made possible by restricting the factorial to two levels. The next column in the ANOVA is the mean square: the sum of squares divided by the degrees of freedom (SS/df). The ratio of mean squares ($MS_{\mathrm{Model}}/MS_{Residual}$) forms the F value of 31.5 (= 781.0/24.8).

The F value for the model must be compared to the reference distribution for F with the same degrees of freedom. In this case, you have 3 df for the numerator (top) and 4 df for the denominator (bottom). The critical F-values can be obtained from table(s) in Appendix 1 at the back of the book by going to the appropriate column (in this case the third) and row (the fourth). Check these against the values shown in Figure 3.9.

If the actual F value exceeds the critical value at an acceptable risk value, you should reject the null hypothesis. In this case, the actual F of 31.5 is bracketed by the critical values for 0.1% and 1% risk. We can say that the probability of getting an F as high as that observed, due to chance alone, is less than 1%. In other words, we are more than 99% confident that taste is significantly affected by one or more of the effects chosen for the model. That's good.

However, we are not done yet, because it's possible to accidentally carry an insignificant effect along for the ride on the model F. For that reason, always check each individual effect for significance. The F-tests for each effect are based on 1 df for the respective numerators and df of the residual for the denominator (in this case, 4). The critical F at 1% for these dfs (1 and 4) is 21.2. Check the appropriate table in Appendix 1 to verify this. The actual F-values for all three individual effects exceed the critical F, so we can say they are all significant, which supports the assumptions made after viewing the half-normal plot.

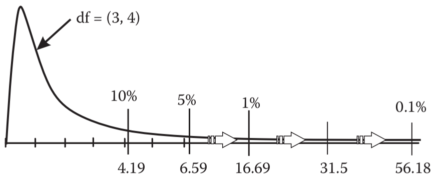


Figure 3.9 The F-distribution with various critical values noted.

Table 3.10 ANOVA for bullets

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	10.18	3	3.39	75.41	0.0006
B	*2.42*	*1*	*2.42*	*53.78*	*0.0018*
C	*6.48*	*1*	*6.48*	*144.00*	*0.0003*
BC	*1.28*	*1*	*1.28*	*28.44*	*0.0060*
Residual	0.18	4	0.045		
Cor Total	10.36	7			

We haven't talked about the last line of the ANOVA, labeled "Cor Total." This is the total sum of squares corrected for the mean. It represents the total system variation using the average response as a baseline. The degrees of freedom are also summed, so you can be sure nothing is overlooked. In this case, we started with eight data points, but 1 df is lost to calculate the overall mean, leaving 7 df for the ANOVA.

The ANOVA for the second response, bullets, can be constructed in a similar fashion. The one shown in Table 3.10 is from a computer program that calculates the probability (p) value to several decimals (reported as "Prob > F"). The p-values are traditionally reported on a scale from 0 to 1. In this book, p-values less than 0.05 are considered significant, providing at least 95% confidence for all results. None of the p-values exceed 0.05 (or even 0.01), so we can say that the overall model for bullets is significant, as are the individual effects.

WHY POPCORN POPS (OR DOESN'T)

Before designing any experiment, it pays to gather knowledge about the subject matter so you can more intelligently choose factors. This studious approach also provides perspective for assessing the apparent effects that emerge from the statistical analysis. For example, here is some subject matter background that may "a-maize" you. Corn used for popping comes from a special strain called pericarp, characterized by an outer covering that is stronger and more airtight than that of other corn varieties. Like all corn, pericarp contains moisture, which when heated, becomes superheated steam. At some point, the pressure causes an explosive rupture in the coating of the popcorn. The white ball of well-popped corn is made up of mostly protein and starch granules that expand to 30 times in volume when popped. Unpopped kernels may be caused by seemingly slight scratches in the coating that allow the heated moisture to gradually escape, rather than build up. Too little moisture in the kernels may also cause problems with popping. On the other hand, excessive moisture results in tough, rather than crunchy, popcorn.

Let's recap the steps taken so far for analyzing two-level factorial designs:

1. Calculate effects: Average of highs (pluses) versus average of lows (minuses)
2. Sort absolute value of effects in ascending order
3. Lay out probability values P_i as shown by examples
4. Plot effects on half-normal probability paper
5. Fit line through near-zero points ("residual")
6. Label significant effects off the line ("model")
7. Calculate each effect's sums of squares (SS) using formula
8. Compute SS_{Model}: Add SS for points far from line
9. Compute $SS_{Residuals}$: Add SS for points on line
10. Construct ANOVA table
11. Using tables, estimate the p-values for calculated F-values; if <0.05, proceed
12. Plot main effect(s) and interaction(s); interpret results

We have now completed most of the statistical homework needed to support the conclusions made earlier. However, one final step must be taken for absolute protection against spurious results: Check the assumptions underlying the ANOVA.

Modeling Your Responses with Predictive Equations

This is a good place to provide details on the model tested in the ANOVA. The model is a mathematical equation used to predict a given response. To keep it simple, let's begin the discussion by looking at only one factor. The linear model is

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

where Y with the circumflex or "hat" (^) is the predicted response, β_0 (beta zero) is the intercept, and β_1 (beta one) is the model coefficient for the

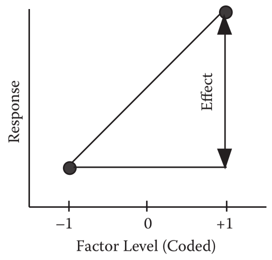


Figure 3.10 Graph of response versus coded factor level.

input factor (X_1). For statistical purposes, it helps to keep factors in coded form: –1 for low and +1 for high. As shown in Figure 3.10, changing the factor from low to high causes the measured effect on response.

The model coefficient β_1 represents the slope of the line, which is the "rise" in response (the effect) divided by the corresponding "run" in factor level (2 coded units). Therefore, the β_1 coefficient in the coded model is one-half the value of the effect (effect/2).

As more factors are added, the number of terms in the model increases. The factorial model for two factors, each at two levels, is

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$$

The fitted model for the popcorn taste with the factors of time (B) and power (C) in coded form is

$$\text{Taste} = 66.5 - 10.25\text{ B} - 8.50\text{ C} - 10.75\text{ BC}$$

The value for the intercept (β_0) of 66.5 represents the average of all actual responses. The coefficients can be directly compared to assess the relative impact of factors. In this case, for example, we can see that factor B (coefficient –10.25) causes a bigger effect than factor C (coefficient –8.50).

The one drawback to the coded model is that you must convert actual factor levels to coded levels before plugging in the input values. Using standard statistical regression, we produced an alternative predictive model that expresses the factors of time and power in their original units of measure:

$$\text{Taste} = -199 + 65\text{ Time} + 3.62\text{ Power} - 0.86\text{ Time} * \text{Power}$$

Use this uncoded model to generate predicted values, but don't try to interpret the coefficients. The intercept loses meaning when you go to the uncoded model because it's dependent on units of measure. For example, a –199 result for taste makes no sense. Similarly, in the uncoded model, you can no longer compare the coefficient of one term with another, such as time versus power.

We advise that you work only with the coded model. This is shown below for the second response:

$$\text{Bullets} = 1.45 - 0.55\text{ B} - 0.90\text{ C} + 0.40\text{ BC}$$

A good way to check your models is to enter factor levels from your design and generate the predicted response. When you compare the predicted value with the actual (observed) value, you will normally see a discrepancy. This is called the "residual."

Diagnosing Residuals to Validate Statistical Assumptions

For statistical purposes, it's assumed that residuals are normally distributed and independent with constant variance. Two plots are recommended for checking the statistical assumptions:

- Normal plot of residuals
- Residuals versus predicted level

Let's look at these plots for the taste response from the popcorn experiment. Table 3.11 provides the raw data.

The column of predicted (Pred) values for taste is determined by plugging the coded factor levels into the coded model. For example, the predicted taste for standard order 1 is

$$\text{Taste} = 66.5 - 10.25(-1) - 8.50\ (-1) - 10.75\ (+1) = 74.5$$

The residuals (Resid), calculated from the difference of actual versus predicted response, can be plotted on normal probability paper. The procedure for creating a *full*-normal plot is the same as that shown earlier for the *half*-normal plot, but you don't need to take the absolute value of the data. Just be sure you have the correct variety of graph paper. In this case,

Table 3.11 Residuals for taste data

Standard	*B*	*C*	*BC*	*Taste Actual*	*Taste Pred*	*Resid*
1	–1	–1	+1	74	74.5	–0.5
2	–1	–1	+1	75	74.5	0.5
3	+1	–1	–1	71	75.5	–4.5
4	+1	–1	–1	80	75.5	4.5
5	–1	+1	–1	81	79.0	2
6	–1	+1	–1	77	79.0	–2
7	+1	+1	+1	42	37.0	5
8	+1	+1	+1	32	37.0	–5

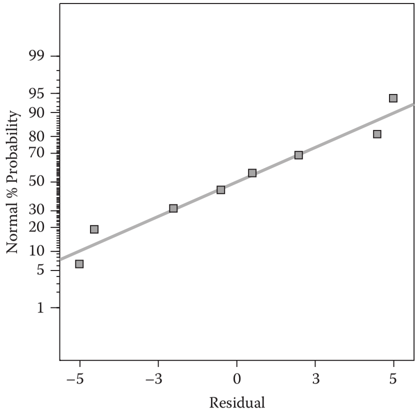


Figure 3.11 Normal plot of residuals for popcorn taste.

we have eight points so, to be evenly dispersed across the range of 100 percent, the cumulative probabilities P_i should be set at these eight intervals: 6.25, 18.75, 31.25, 43.75, 56.25, 68.75, 81.25, and 93.75%. The resulting plot is shown on Figure 3.11.

If the residuals are normally distributed, they will all fall in a line on this special paper. In this case, the deviations from linear are very minor, so it supports the assumption of normality. Watch for clearly nonlinear patterns, such as an "S" shape. Then consider doing a response transformation, a topic that will be discussed in the next section.

THE PENCIL TEST

Recall from Chapter 1 that a simple, but effective, way to check for linearity is to place a pencil on the normal probability graph. If the pencil covers all the points, consider it in line. A large marker pen would solve all your problems!

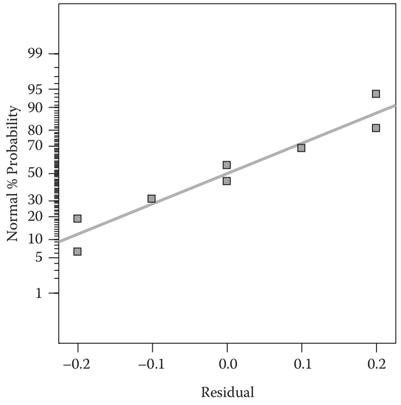


Figure 3.12 Normal plot of residuals for popcorn bullets.

Figure 3.12 shows the normal plot of residuals for the second response (bullets). Give it the pencil test. You will find that residuals for the bullets exhibit no major deviations from the normal line.

The other recommended plot for diagnostics is the residuals versus predicted response. Using the data from Table 3.11, we constructed the plot shown in Figure 3.13.

Ideally, the vertical spread of data will be approximately the same from left to right. Watch for a megaphone (<) pattern, where the residuals increase with the predicted level. In a design as small as that used for the popcorn experiment, only 8 runs, it's hard to detect patterns. However, it's safe to say that there is no definite increase in residuals with predicted level, which supports the underlying statistical assumption of constant variance. In the next chapter, we will show you what to do if the residuals are not normal and exhibit nonconstant variance.

Figure 3.14 shows the residual versus predicted plot for the second response (bullets). You don't need to apply the rule of thumb because there is no obvious increase in residuals as the predicted value increases.

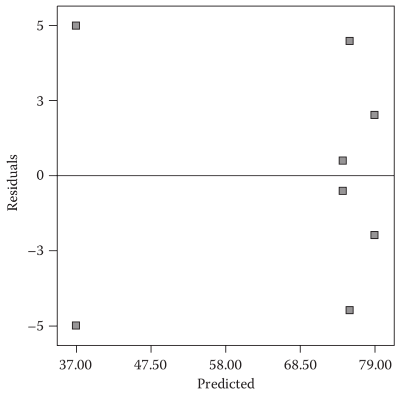


Figure 3.13 Residuals versus predicted taste.

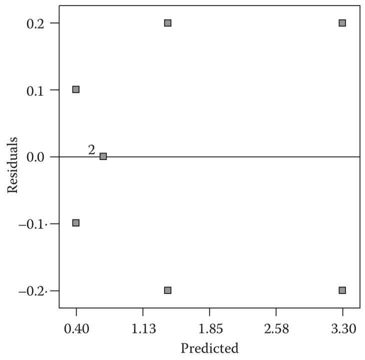


Figure 3.14 Residuals versus predicted bullets.

A LITERAL RULE OF THUMB

ANOVA and other statistical analyses are relatively robust to deviations from normality and constancy of variance. Therefore, you should not overreact to slight nonlinearity on the normal plot of residuals, or to vague patterns on the residuals versus predicted plot. As a rule of thumb, if you think you see a pattern, but it disappears when you cover one point with your thumb, then don't worry. However, if you have a touch screen computer, do not get carried away with this procedure.

P.S.: Despite rumors to the contrary, the term *rule of thumb* probably came from use of the thumb as a crude measure of length. However, it may refer to the traditional practice of brewmasters who check temperature of beer by dipping their thumbs in the batch. This latter explanation is most consistent with the origins of the t-test and other statistical innovations.

Practice Problems

Problem 3.1

Montgomery describes a two-level design on a high-pressure vessel for making waferboard glue (see Recommended Readings: *Design and Analysis of Experiments* (2012), example 6.2). A full-factorial experiment is carried out in the pilot plant to study four factors thought to influence the filtration rate of the product. Table 3.12 shows actual high and low levels for each of the factors.

At each combination of these machine settings, the experimenters recorded the filtration rate. The goal is to maximize the filtration rate and also try to find conditions that would allow a reduction in the concentration of formaldehyde, factor C.

The response data are tabulated in standard order, with factor levels coded, in Table 3.13.

Do an analysis of the data to see if any effects are significant. Recommend operating conditions that maximize rate with a minimum of formaldehyde. (Suggestion: Use the software provided with the book. First

Table 3.12 Factors and levels for two-level factorial design on a reactor

Factor	*Name*	*Units*	*Low Level (–)*	*High Level (+)*
A	Temperature	Deg C	24	35
B	Pressure	Psig	10	15
C	Concentration	Percent	2	4
D	Stir Rate	RPM	15	30

Table 3.13 Design layout and response data for reactor study

Standard	*A*	*B*	*C*	*D*	*Filtration Rate (gallons per hour)*
1	–	–	–	–	45.0
2	+	–	–	–	71.0
3	–	+	–	–	48.0
4	+	+	–	–	65.0
5	–	–	+	–	68.0
6	+	–	+	–	60.0
7	–	+	+	–	80.0
8	+	+	+	–	65.0
9	–	–	–	+	43.0
10	+	–	–	+	100.0
11	–	+	–	+	45.0
12	+	+	–	+	104.0
13	–	–	+	+	75.0
14	+	–	+	+	86.0
15	–	+	+	+	70.0
16	+	+	+	+	96.0

do the two-level factorial tutorial that comes with the program. It's keyed to the data in Table 3.13. See About the Software for software installation instructions and details on the associated tutorials.)

Problem 3.2

Modern cars are built with such precision that they become hermetically sealed when locked. As a result, the interior becomes unbearably hot in cars parked outdoors on warm, sunny days. A variety of window covers can be purchased to alleviate the heat. The materials vary, but generally the covers present either a white or shiny, metallic surface that reflects solar radiation. In some cases, they can be flipped to one side or the other. The white variety of cover usually displays some sort of printed pattern, such as a smiling sun or the logo of a local sports team. These patterns look good,

but they may detract from the heat-shielding effect. A two-level factorial design was conducted to quantify the effects of several potential variables: cover (shiny versus white), direction of the parked car (east versus west), and location (near the office in an open lot versus far away from the office under a shade tree).

FOR THOSE CONSUMERS WHO MAY NOT BE FIRING ON ALL CYLINDERS

Operating instructions seen on accordion-style, front-window shade for automobiles: "Remove before driving."

The resulting 8-run, two-level DOE was performed during a period of stable weather in Minneapolis during early September. Anticipating possible variations, the experimenter (Mark) recorded temperature, cloudiness, wind speed, and other ambient conditions. Outside temperatures ranged from 66 to 76°F under generally clear skies. Randomization of the run order provided insurance against the minor variations in weather. The response shown in Table 3.14 is the difference in temperature from inside to outside as measured by a digital thermometer.

Even in this relatively mild late-summer season, the inside temperatures of the automobile often exceeded 100°F. It's not hard to imagine how hot it could get under extreme midsummer weather. Analyze this data to see what, if any, factors prove to be significant. Make a recommendation on how to shade the car and how and where to park it. (Suggestion: Use the

Table 3.14 Results from car-shade experiment

Std	*A: Cover*	*B: Orientation*	*C: Location*	*Temp Increase (°F)*
1	White	East	Close/Open	42.1
2	Shiny	East	Close/Open	20.8
3	White	West	Close/Open	54.3
4	Shiny	West	Close/Open	23.2
5	White	East	Far/Shaded	17.4
6	Shiny	East	Far/Shaded	10.4
7	White	West	Far/Shaded	11.7
8	Shiny	West	Far/Shaded	16.0

software provided with the book. Set up a factorial design, similar to the one you did for the tutorial that comes with the program, for three factors in eight runs. Sort the design by standard order to match the table above and enter the data. Then do the analysis as outlined in the tutorial.)

Appendix: How to Make a More Useful Pareto Chart

The Pareto chart is useful for showing the relative size of effects, especially to nonstatisticians. However, if a two-level factorial design gets botched, for example, due to a breakdown during one particular run, it becomes unbalanced and nonorthogonal, thus causing the effects to exhibit differing standard errors. In such cases, the absolute magnitude of an effect may not reflect its statistical significance. To make the Pareto chart more robust to experimental mishaps, we recommend it be plotted with the t-values of the effects. Furthermore, in this dimensionless statistical scale, it becomes appropriate to superimpose benchmarks for significance as detailed in this appendix, which follows up on the less-sophisticated Pareto chart laid out earlier for the effects on popcorn bullets (unpopped kernels).

The t-value is computed by simply dividing the numerical effect by its associated standard error, which is easy to calculate for a balanced, orthogonal experiment like that done on microwave popcorn. Below is the formula in this special case:

$$\text{t-value}_i = \frac{|Effect_i|}{\sqrt{MS_{Residual}\left(\frac{1}{n_+}+\frac{1}{n_-}\right)}}$$

where n represents the number of responses from each of the two levels tested and MS is the mean square for the residuals computed by the ANOVA. For the largest effect on the weight of popcorn bullets—C (seen at the bottom of Table 3.7)—the t-value is calculated as follows:

$$\text{t-value}_C = \frac{|-1.8|}{\sqrt{0.045\left(\frac{1}{4}+\frac{1}{4}\right)}} = \frac{|-1.8|}{\sqrt{0.045\left(\frac{2}{4}\right)}} = \frac{|-1.8|}{\sqrt{0.0225}} = \frac{1.8}{0.15} = 12$$

Similar calculations can be performed to obtain t-values for the other six effects, which can then be bar-charted in descending order. However, before doing this, it will be very helpful to look up the two-tailed t-value for 0.05 probability (or some other p-value that you establish depending on your threshold for risk) from the table in Appendix 1.1. The degrees of freedom (df) can be read off the ANOVA for the residuals. In this case, the df are 4 (from Table 3.10), so the critical t-value is 2.776. A more conservative t-value, named after its inventor (Bonferroni), takes the number of estimated effects into account by dividing it into the desired probability for the risk value alpha (α). For the popcorn experiment, seven effects are estimated, so the Bonferroni-corrected p-value becomes:

$$t_{\left(\alpha_{2-tail}=0.05\right)\big/_{k=7},\,df=4} = t_{(0.007,4)} \cong 5.1$$

This value can be calculated precisely by statistical software, but approximated values from a table (such as the one in Appendix 1) may suffice for seeing which effects, if any, stand out from those caused by chance variation.

The Pareto chart in terms of t-values with the two threshold limits is shown in Figure 3.15.

All three labeled effects (C, B, and BC) exceed even the more conservative Bonferroni limit, thus providing a high level of confidence; greater than 95%.

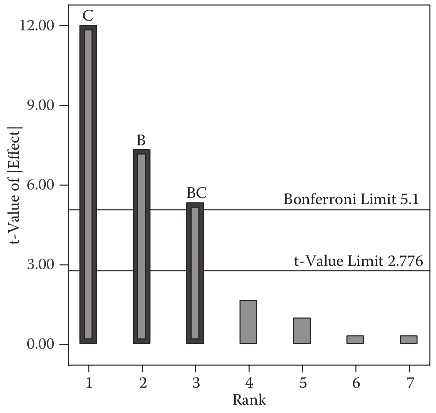


Figure 3.15 Pareto chart for bullets rescaled by t-values with limits superimposed.

ANOTHER ITALIAN MATHEMATICIAN: BONFERRONI

Carlo Emilio Bonferroni (1892–1960) joined his predecessor Pareto in the pantheon of Italian mathematicians that include such luminaries as Fibonacci. Bonferroni's famous Correction states that if an experimenter tests n independent hypotheses, then the statistical significance level should be n times smaller than usual. For example, when testing two hypotheses, instead of assigning a p-value of 0.05, one should cut this value in half, to a much stricter level of 0.025. The Bonferroni Correction safeguards against multiple tests of statistical significance on the same data.

For example, when the authors worked for General Mills, the head of quality control in Minneapolis would mix up a barrel of flour and split it into 10 samples, carefully sealed in air-tight containers, for each of the mills to test in triplicate for moisture. A naïve analyst would have simply taken the highest moisture value and compared it to the lowest one. But given that there are 45 possible pair-wise comparisons (10*9/2), this selection (high versus low) is likely to produce a result that tests significant at the 0.05 level (1 out of 20). Refer to the 6/6/10 StatsMadeEasy blog "Bonferroni of Bergamo" (www.statsmadeeasy.net/2010/06/bonferroni-of-bergamo/) for the results of a simulation that graphically demonstrates the bias, if uncorrected, introduced by such a multiple pair-wise comparison.

> *"I do not feel obliged to believe that the same God who has endowed us with sense, reason, and intellect has intended us to forgo their use."*
>
> **Galileo Galilei**

Chapter 4

Dealing with Nonnormality via Response Transformations

No experiments are useless.

Thomas Edison

At the end of Chapter 3, we showed you how to check for normality—a fundamental assumption of the statistical analysis for design of experiments (DOE). In this chapter, we discuss how to deal with nonnormality (and nonconstant variance) via transformation of the response data. The most common transformation, the logarithm, is illustrated with a case study. This is the most complex DOE shown thus far: a two-level design on four factors, requiring 16 runs for all the combinations. After detailing this DOE, we will use fractional designs to squeeze more factors into the same number of runs.

Skating on Thin Ice

The data in this chapter comes from an exercise called *tabletop hockey*. It works well as an in-class experiment for workshops on DOE. The objective of the hockey experiment is to learn how to shoot a puck for distance with a flexible, 15-centimeter (cm) ruler. The puck is made of two or more quarters (25-cent coins) stuck together with a gum adhesive. A simple wooden block acts as a fixture for the ruler. The response is the distance the puck slides over a smooth tabletop.

WENT TO A FIGHT AND A HOCKEY GAME BROKE OUT

True hockey fans, particularly at the college level, appreciate the planning and organization of a well-coached team. The rink-long breakaways and resounding checks may arouse the crowd, but good passing and discipline win out in the end. Similarly, a hit-or-miss approach to experimentation may achieve goals in spectacular fashion, but the odds over the long haul favor the well-planned factorial approach.

> *"Ice hockey is a form of disorderly conduct in which the score is kept."*
>
> **Doug Larson**

Table 4.1 Test factors for tabletop hockey

Factor	*Name*	*Units*	*Low Level (–)*	*High Level (+)*
A	Puck weight	Quarters	4	6
B	Stick length	Centimeters	7.5	15
C	Wind-up	Centimeters	5	10
D	Puck place	Percent	0	100

After some brainstorming, a team of workshop students decided to study four factors at the levels shown in Table 4.1.

Figure 4.1 is a template for the tabletop hockey exercise. The circles show the various locations for the factor named *stick length*. The marks below the line of the ruler relate to the factor called *windup*. The last factor, named *puck place*, is not shown on the figure. The low level for puck place is 0%, which means it is kept at the original position in front of the ruler and slapped. The high level for puck place is 100%, which puts it against the ruler so it can be flung forward, much like a wrist shot in the real game of hockey.

Table 4.2 lists the runs in standard order with factors in coded form. The actual experiment was performed in random order. The response (Y) covers a broad range, from 3 to 190 centimeters; more than a 60-fold change. When data spans an order of magnitude (10-fold) or more, model fitting is often simplified by applying a logarithm to the response. Therefore, we added a column for the log base 10 of the response. This is called a *transformation*.

Ultimately, we will show that the logarithmic scale works best, but for comparison sake, let's first analyze the data without a transformation.

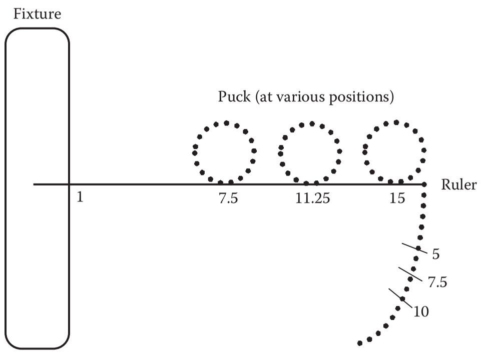


Figure 4.1 Template for tabletop hockey (dimensions in centimeters).

Table 4.2 Data for tabletop hockey (response Y is distance in centimeters)

Standard	*A*	*B*	*C*	*D*	*Y*	*$Log_{10}Y$*
1	–	–	–	–	38.2	1.582
2	+	–	–	–	23.3	1.367
3	–	+	–	–	3.0	0.477
4	+	+	–	–	7.6	0.881
5	–	–	+	–	110.0	2.041
6	+	–	+	–	90.6	1.957
7	–	+	+	–	20.6	1.314
8	+	+	+	–	18.9	1.276
9	–	–	–	+	36.6	1.563
10	+	–	–	+	38.0	1.580
11	–	+	–	+	47.4	1.658
12	+	+	–	+	44.9	1.652
13	–	–	+	+	190.0	2.279
14	+	–	+	+	116.8	2.067
15	–	+	+	+	137.5	2.138
16	+	+	+	+	84.5	1.927

Table 4.3 Table of effects

Term	*Effect (Y)*	*Effect (Log_{10} Y)*
A	–19.8375	–0.045
B	–34.8875	–0.39
C	66.2375	0.53
D	47.9375	0.50
AB	6.6875	0.078
AC	–16.9875	–0.091
AD	–11.9875	–0.062
BC	–26.5875	–0.035
BD	18.1125	0.36
CD	24.2375	–0.043
ABC	2.7875	–0.066
ACD	–14.2875	–0.088
ABD	–2.6125	–0.013
BCD	0.9625	–0.081
ABCD	3.2375	0.076

(In Table 4.2 and those that follow, ignore the column(s) labeled "Log_{10}" for now.) From the 16 unique combinations that result from the 2^4 factorial (2 * 2 * 2 * 2 = 16), 15 effects can be estimated, as shown in Table 4.3. A computer does the calculations much faster and more accurately than most people, with the possible exception of statisticians. The tricky part for nonstatisticians is understanding the terminology and interpreting the outputs, but the reader who has gone over all the nuts and bolts of two-level factorials in Chapter 3 has nothing to fear.

Once again you see data transformed by the logarithm. We will be getting back to the "logged" data very soon, but for now let's proceed with the analysis of the data as it was originally collected, before being transformed mathematically. The next step is to view the untransformed effects (Figure 4.2).

The students working on the experiment could not see any obvious division of effects on this plot, so they simply focused on the largest effect, C. The analysis of variance (ANOVA) (not shown) did produce a significant probability for the resulting model (intercept plus main effect C), but the residual plots looked odd (Figure 4.3a,b).

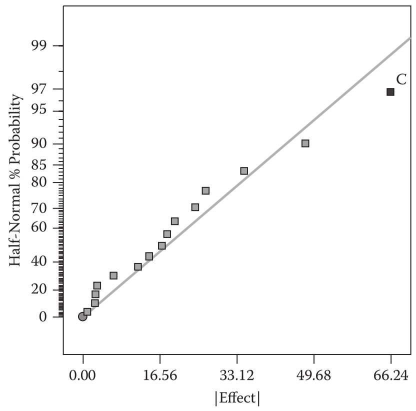


Figure 4.2 Half-normal plot of effects for tabletop hockey.

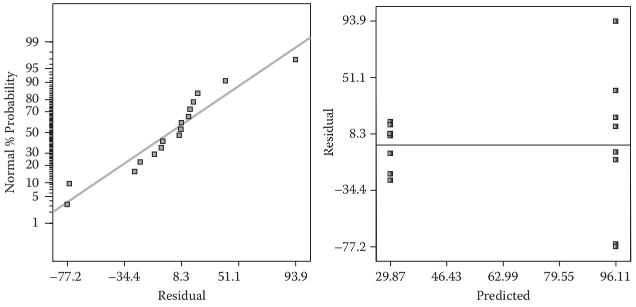


Figure 4.3a,b Normal plot of residuals (left) and residuals versus predicted (right).

The residuals did not line up on the normal plot and they clearly increased with the predicted level, forming the unwanted megaphone pattern. The students knew enough from prior training to recognize that these patterns indicate problems in the statistics. The analysis required a different approach.

Log Transformation Saves the Data

The tabletop hockey data demonstrates a very common characteristic—as the response increases, the variance does too. This is a case of constant

percent error. For example, a system may exhibit 10% error, so at a level of 10 the standard deviation is 1; however, at a level of 100, the standard deviation becomes 10, and so on. Transforming such a response with the logarithm will stabilize the variance.

WHEN RESIDUALS MISBEHAVE, HIT THEM WITH A LOG

Logarithms are used as a scaling function for many measurements, including decibels of sound, the Richter scale for earthquakes, the pH rating of acidity, and astronomical units for stellar brightness. We encourage you to try rescaling your response to log when residual diagnostic plots show abnormalities, but this advice comes with a cautionary note. Don't expect much of an impact if the range of response is threefold or less. In this case, the response transformation may create more trouble than it's worth. Also, remember that you cannot take the log of a negative number. Overcome this obstacle by adding a constant to the responses so that all become positive.

Luckily, we anticipated the need for a log after seeing the wide range of response. Using the transformed effects from Table 4.3, the workshop participants generated the half-normal plot shown in Figure 4.4.

The transformation was amazing, revealing a subset of relatively large effects, including interaction BD. Moreover, as one would expect from

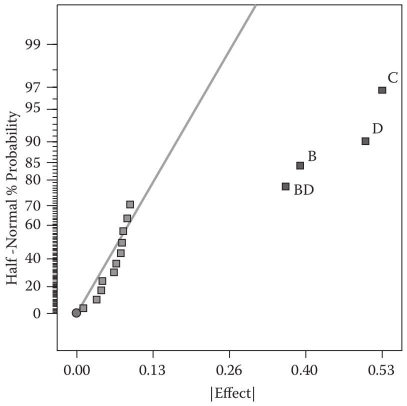


Figure 4.4 Half-normal plot of transformed effects (Log_{10}).

Table 4.4 ANOVA for transformed response

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	3.23	4	0.81	46.19	< 0.0001
B	*0.60*	*1*	*0.60*	*34.24*	*0.0001*
C	*1.11*	*1*	*1.11*	*63.66*	*< 0.0001*
D	*0.99*	*1*	*0.99*	*56.76*	*< 0.0001*
BD	*0.53*	*1*	*0.53*	*30.11*	*0.0002*
Residual	0.19	11	0.017		
Cor Total	3.43	15			

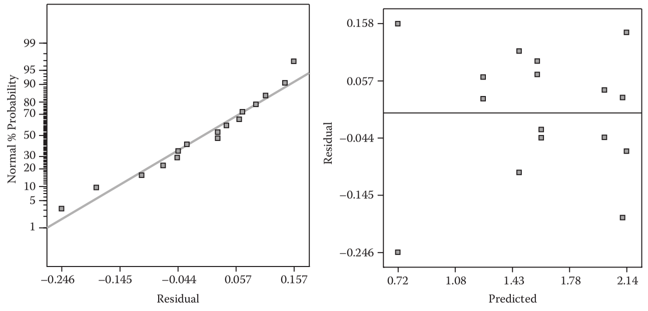


Figure 4.5a,b Residual plots after log transformation.

seeing such a dramatic half-normal plot, the subsequent analysis of variance indicated that all four chosen effects (B, C, D, and BD) were highly significant (Table 4.4).

The residuals from the transformed model now looked much better (Figure 4.5a,b).

In the final analysis, the students found that the distance of the shot is most affected by factor C: the windup of the ruler (Figure 4.6).

As hockey aficionados might expect, larger windups produced longer shots. The weight of the puck (Factor A) had no effect, at least over the range tested. On the other hand, the significant interaction between B and D (Figure 4.7) yielded results that may be surprising.

The effect of stick length (B) depends on puck placement (D). The D+ (100% setting) line is flat, with overlapping least significant difference (LSD) bars at either end. This indicates that stick length from 7.5 to 15 cm makes no difference when you fling the puck at the D+ level. However, when the puck is left at the original ruler line (0% setting) and slapped at this D– level, the longest shot comes with the shorter stick (B–). This is counterintuitive for most students who participate in this exercise, and thus it provides a good lesson on why one ought to conduct experiments rather than relying only on instinct.

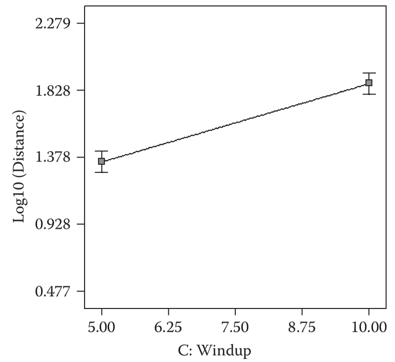


Figure 4.6 Main effect of factor C (windup of the ruler).

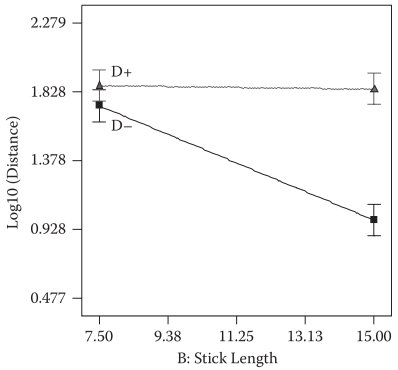


Figure 4.7 Interaction BD (stick length and puck place).

To see the combined effect of the three significant factors, view the cube plots in Figure 4.8. The left cube is in transformed units to be consistent

BOOM-BOOM: THE MASTER AT SLAPPING A PUCK

Bernie Geoffrion, a hockey Hall of Famer who passed away in 2006, invented the slap shot. The sound of the puck coming off his stick and almost instantaneously smashing into the boards soon earned him the nickname "Boom-Boom." Unfortunately, although the nickname reflects the power of the slap shot, Bernie's shot was hard, but not very accurate and usually missed the goal net. But, as hockey great Wayne Gretzky pointed out, it is better to try and fail than not to try at all:

> *"You miss 100% of the shots you never take."*

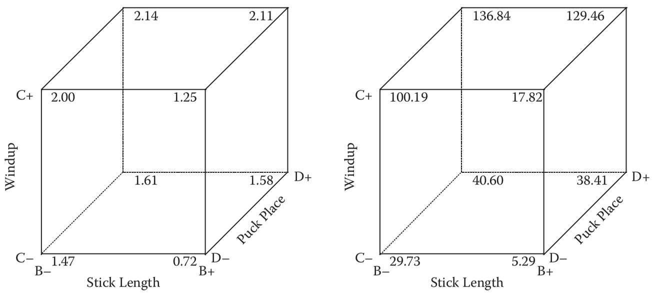


Figure 4.8a,b Cube plot of predicted response in log (left) versus original (median) units (right).

with the previous effect graphs done in the as-analyzed (log) metric. But, what you really want to know is the distance in original units. This can be generated by taking the antilog of the predicted responses. The right cube in Figure 4.8 presents the results after taking this reverse transformation.

The best results, well over 100 centimeters, can be seen at the upper, back side with long windup (C+) and the puck placed against the stick (D+). At these settings of C and D, the stick length makes little difference, but by choosing the shorter level (B–) you lessen the impact of potential variations in the placement of the puck.

Now that we are back in original scale, let's revisit the interaction plot, illustrated in Figure 4.9. It was produced with factor C set at its optimal level: high (+).

MEAN PREDICTIONS BIASED WHEN TRANSFORMING BACK FROM LOG SCALE

You sharp-eyed readers may be wondering why the captions for the graphs in original scale are noted to be at the median. It turns out that the process of transforming back from log scale creates a bias in the mean predictions, the results being somewhat underestimated (but correct as median values). For example, the extreme predicted median values of 5.29 and 136.84 shift upward a bit via a bias correction to 5.53 and 143.19 at their means. Good software will either note (as done here) that predictions are at the median and/or apply the necessary correction.

(Stop here if you prefer not wracking your brain.)

The reason for the bias relates to some simple math using logs. Consider a set of data ranging from 10 to 100. Now apply the logarithm in base 10 to produce data that are normally distributed (bell-shaped) with a transformed range of 1 to 2. Take the middle value of 1.5 and antilog it. The result is 31.6; not 55 as you might have thought from the original data. The end result is a skewed distribution (think of a hill with a skier sliding down to the right) where the median falls to the left of the mean. This is what creates the bias. If you are great at mathematics and wish to learn more about this problem, search the Internet on "retransformation bias" and look for detailing of "homoskedastic error" and "smearing estimators." (We warned you.)

> *"Taking the logarithm of a set of numbers squashes the right tail of the distribution."*
>
> **Andy Field, Discovering Statistics, Exploring Data: The Beast of Bias, StatisticsHell.com**

After untransforming the responses to put them back in the original scale of distance in centimeters, two things become noticeable:

- The curvilinear shape that is most pronounced at the low level of "puck place" (D–).
- How small the LSD bar gets at the short distance resulting from longest stick length (15 cm) at D–.

It makes sense that variation will be much reduced at this extremely low median response level of 5.29 centimeters versus what you would expect at (median) 136.84 centimeters, the predicted extremes in original scale (you can

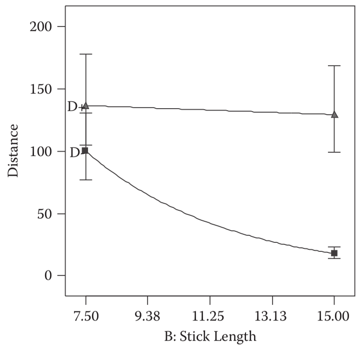


Figure 4.9 Interaction BD in original (median) units with C set high.

read these values off the right-hand cube in Figure 4.8). Now you really see the reason for applying the log transformation to stabilize variances at all levels of response so they can be properly pooled for statistical purposes.

Choosing the Right Transformation

The abnormal residual plots shown on Figure 4.3 exhibit a relatively common "power law" relationship between the standard deviation and the mean response. Statistically, this situation is symbolized as follows:

$$\sigma_y \propto \quad {}^{\alpha}$$

where the Greek letter sigma is the true standard deviation (of response y), which is proportional to the true mean (mu) to some power (alpha). Table 4.5 shows just a few of the possibilities for this power law, along with the appropriate transformations.

Ideally, there is no power law relationship (α = 0), so no transformation is needed. In some cases, such as counts of imperfections, the standard deviation increases with the mean to the 0.5 power. The direct power relation (α = 1) implies that the error is a constant percent of the response. This is a very common problem, which is remedied by a log transformation.
In some cases, particularly when the response is a rate (e.g., liters per second), the standard deviation increases with the square of the mean (power of 2). Then the inverse transformation is indicated (e.g., seconds per liter).

Table 4.5 Variance-stabilizing transformations

Power (α)	*Transformation*	*Comment*
0	None	Normal
0.5	Square root	Counts
1	Logarithm	Constant percent error
2	Inverse	Rate data

THE DEATHLY COUNT

Counts of traffic accidents and deaths follow the Poisson distribution, where the standard deviation is a function of the mean. A practical application of this was demonstrated by Ladisclaus Bortkiewicz, who kept track of Prussian cavalry soldiers killed by their horses between 1875 and 1894 (*The Law of Small Numbers,* 1898). Pity the poor soldiers who became statistics.

Transformations, such as those shown in Table 4.5, may stabilize the residual variance, satisfying the assumptions for ANOVA. They may be supported by scientific knowledge of the underlying relationship between factor(s) and response. For example, material scientists studying spider silk discovered a remarkable exponential relationship between extension (x) and the external force (f): $f = e^{x/k}$ where k is a constant (see H. Zhou, Polymers and Filaments: The Elasticity of Silk, *Max Planck Institute Biannual Report 2003–2004*, p. 122). In this and similar cases, a log transformation is the obvious remedy for linear modeling.

If you are uncertain whether or not a transformation will help, try one. However, if you don't see a definite improvement in the ANOVA (F-test) and residuals, you may find that the transformation actually complicates matters. If you do use a transformation, remember to reverse the process by applying the inverse function, such as the antilog for a logged response. Otherwise you might get some grief about your goofy predictions!

A PLOT THAT ADVISES WHEN A RESPONSE WOULD BEST BE TRANSFORMED

In *RSM Simplified* (Productivity Press, 2004), we detail the Box–Cox plot for transformations (see the appendix in Chapter 5). The plot, which can be generated by the software provided with this book and other programs like it, pinpoints the power (α) for the response transformation that minimizes residuals. It is very handy!

Practice Problem

Problem 4.1

This problem demonstrates that design of experiments can be applied to any system, even one that does not involve manufacturing. It addresses a question debated by producers of goods and services aimed at a technical audience: Would this data-driven personality type react favorably to fancy four-color printing on a direct mail piece? Conventional wisdom says the answer will be yes, but, on second thought, this “yes” may not be absolute. While it may apply to nontechnical consumers, a cheaper, two-color printing might work as well (or better) for technical types.

A ROSE BY ANY OTHER NAME

In its March 11, 1996 issue, *Forbes* magazine introduced the concept of design of experiments to the business world. The title of the article, The New Mantra MVT, coined a new acronym (MVT), which stands for multivariable testing. Admittedly, the reference to multivariable testing conveys a major benefit to this style of experimentation. However, the case studies presented in the article show that MVT is simply design of experiments applied to business and marketing problems. Nevertheless, whether you call it DOE or MVT, it is a powerful methodology that can be applied to any system for which one can manipulate inputs and measure outputs.

> *“If you test factors one at a time, there’s a very low probability that you’re going to hit the right one before everybody gets sick of it and quits.”*
>
> ***Forbes***

In addition to considering the color factor, market researchers looked at two postcard sizes (small versus big) and two types of paper stock (thin versus thick). The eight resulting postcard designs (2^3) were sent to eight equal segments of the company’s client list, chosen at random. To garner more response, the researcher offered a free technical report to anyone who faxed back the reply side of the postcard. The postcards incorporated the standard two-level code to facilitate measurement. For example, the first and last combinations in standard order were coded:

- – – – (= two-color, small card on thin stock)
- + + + (= four-color, big card on thick stock)

Table 4.6 shows the number of requests generated by each postcard configuration.

The cost of printing the cards is also shown for reference. This is a “deterministic” response because it depends only on the factor levels. The four-color, large, thick postcard was the most expensive combination.

Table 4.6 Results from postcard experiment

Standard	*A: Color*	*B: Size*	*C: Thickness*	*Requests (Count)*	*Printing Cost (Cents/Card)*
1	Two	Small	Thin	152	6
2	Four	Small	Thin	57	10
3	Two	Large	Thin	258	8
4	Four	Large	Thin	31	12
5	Two	Small	Thick	250	8
6	Four	Small	Thick	131	12
7	Two	Large	Thick	398	10
8	Four	Large	Thick	96	14

OBTAIN ENOUGH RESPONSES TO GENERATE STATISTICAL SIGNIFICANCE

To get good, reliable results from tests on direct-mail pieces, you must generate a significant response from every configuration. Market researchers advise a minimum of 20 responses per row (or “cell”). For simple comparison, where only two varieties are tested on a split mailing list, the following rule of thumb is often applied: If the difference between the test results is two times greater than the square root of the total, it is a significant difference. Because responses may continue to trickle in for many weeks, market researchers often extrapolate the early returns in order to generate preliminary findings. For example, they might double the response received after two weeks’ time.

Analyze this data. Given that this section of the book focuses on the use of transformations, consider trying one. (Hint: The response is a count.) Determine the combination that maximizes response. You might be surprised by the results.

(Suggestion: Use the software provided with the book. Set up a factorial design, similar to the one you did for the tutorial that comes with the program, for three factors in 8 runs with two responses. Sort the design by standard order to match Table 4.6, enter the data, and do the analysis as outlined in the tutorial. Then go back and reanalyze after first choosing the square root as a response transformation. Compare the model and resulting residual plots before and after doing the transformation.)

Chapter 5

Fractional Factorials

Believe nothing and be on your guard against everything.

Latin Proverb

The full-factorial approach to experimentation covers all combinations of factors, providing valuable information on interactions. However, the number of experimental runs increases rapidly, even if you test the factors at only two levels each. Fortunately, by cutting back to a “fractional factorial,” you may find it possible to study more factors within a given experimental budget. Table 5.1 shows how few runs we suggest you pare back to for four to seven factors (see Appendix 2 for details on layout.). These particular designs are especially good for “screening” the main effects of many factors in search of the vital few or more in-depth “characterizing” for possible two-factor interactions. For reference, we show the number of main effects and two-factor interactions (2 fi). A minimal experiment for screening requires at least two times the number of runs as the number of factors, e.g., 8 runs for 4 factors. An efficient fraction for characterization contains at least this many runs plus one more for estimating the overall average, but no more. The five-factor, 16-run half fraction is an ideal example—a “minimum run” design for resolving main effects and 2 fis. Unless runs come free, consider choosing a fractional factorial for four or more factors.

We hope this whets your appetite, because we will soon get into the nitty-gritty of fractional factorial design. You will learn that there is a price to pay in the form of “aliasing” effects and, due to the reduction in runs, an inevitable power loss. The more you know about this the better, because conducting fractional factorials is like playing with fire. It’s a powerful tool that can burn you if not handled carefully.

Table 5.1 Number of runs in selected two-level fractional factorial designs for screening or characterization

Factors (Main Effect)	*Two-Factor Interactions*	*Full Factorial Runs*	*Screening (Runs)*	*Characterization (Runs)*
4	6	16	8	12
5	10	32	12	16
6	15	64	14	22
7	21	128	16	30

THE “SCO” STRATEGY OF EXPERIMENTATION

A rule-of-thumb for process development is not to put more than 25% of your budget into the first experiment, thus allowing the chance to adapt as you work through the project (or abandon it altogether). Furthermore, a sound strategy of experimentation is to proceed in three stages:

1. Screening the vital few factors (typically 20%) from the trivial many (80%)
2. Characterizing main effects and interactions
3. Optimizing (typically via response surface methods)

For a great overview of this SCO path for successful design of experiments (DOE), refer to Ronald D. Snee. 2010. Implementing Quality by Design. *Pharmaceutical Processing Magazine*, March 23.

Of course, at the very end, one must not overlook one last step: Confirming your results via a statistically valid, that is, adequately powered, experiment design around the optimal settings or by simply repeatedly running at this ideal point.

Example of Fractional Factorial at Its Finest

Let’s begin with an example of a very safe fractional factorial: five factors at two levels each, done in 16 runs (a half fraction). This data comes from an

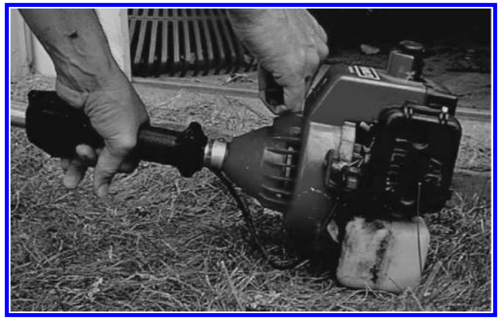


Figure 5.1 The weedwacker engine.

actual experiment on an uncooperative grass trimmer, commonly known as a weedwacker. A one-cylinder engine that runs on a mixture of gasoline and oil powers the weedwacker (Figure 5.1). Perhaps you have become frustrated trying to start small engines such as this one.

The weedwacker engine features a manual starter (pull cord), and its performance is largely a function of three controls: primer pump, choke, and gas. Before starting the engine, you must first give the primer several pumps, choke the air intake, and pull the starter cord several times. To start the engine, the choke must be open. Table 5.2 lists the tested factors and levels.

It would require 32 runs to perform all the combinations of these five factors. This full, two-level factorial would reveal 31 effects: 5 main effects, 10 two-factor interactions, 10 three-factor interactions, 5 four-factor interactions, and 1 five-factor interaction. Interactions of three or more factors are highly unlikely in an engine or any other system. Moreover, the principle of sparsity of effects states that only 20% of the main effects and two-factor interactions are likely to be significant in any particular system. If this is true, then only about three effects will be significant, which leaves 28 effects for estimation of error; far more than necessary. Therefore, a full factorial on five factors (or more) will waste much of its effects on unneeded estimates of error.

Table 5.2 Factors for weedwacker experiment

Factor	*Name*	*Low Level (–)*	*High Level (+)*
A	Prime pumps	3	5
B	Pulls at full choke	3	5
C	Gas during full choke pulls	0	100%
D	Final choke setting	0	50%
E	Gas for start	0	100%

Table 5.3 Design layout for weedwacker experiment

Std	*A*	*B*	*C*	*D*	*E*	*Pulls*
1	–	–	–	–	+	1
2	+	–	–	–	–	4
3	–	+	–	–	–	4
4	+	+	–	–	+	2
5	–	–	+	–	–	8
6	+	–	+	–	+	2
7	–	+	+	–	+	3
8	+	+	+	–	–	5
9	–	–	–	+	–	3
10	+	–	–	+	+	1
11	–	+	–	+	+	3
12	+	+	–	+	–	4
13	–	–	+	+	+	3
14	+	–	+	+	–	4
15	–	+	+	+	–	6
16	+	+	+	+	+	5

A properly constructed half fraction (16 runs) estimates the overall mean, the 5 main effects, and the 10 two-factor interactions. The design for the weedwacker, constructed by standard methods, is shown in Table 5.3. Experiments are listed in standard (Std) order, not randomized run order. Factors are shown in coded format. The response (Y) is pulls needed to start the engine.

You can use this as a template for your own experiment on five factors. Notice that the first four factor columns form a full factorial with 16 runs. The fifth factor column (E) is the product of the first four columns (ABCD). Check it out. This and many other fractional designs can be generated via statistical software or from tables in referenced textbooks on DOE. We will give you more clues on their construction later on in this section.

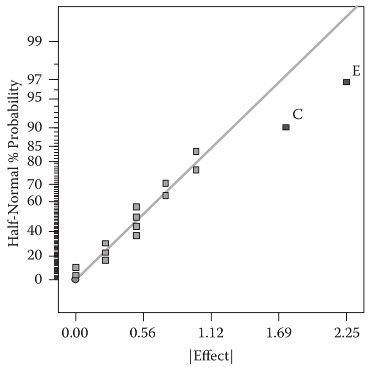


Figure 5.2 Half-normal plot of effects for weedwacker.

The half-normal plot of effects (Figure 5.2) reveals two large main effects: factors E and C. The Pareto plot in Figure 5.3 supports the hypothesis that these two effects are the vital few versus the others being the trivial many. Do not forget, however, that E was constructed from ABCD and that these two effects are, therefore, aliased. The effect of C is also aliased with a four-factor interaction: ABDE. (Check this out by multiplying the appropriate columns and comparing signs.) Four-factor interactions are not plausible, so we will ignore these aliases. This assumption follows the generally accepted practice of statisticians. However, you must keep in mind that you rarely get something for nothing; fractional factorials save on runs but they produce aliases, which can be troublesome. (See the next section, (Potential Confusion Caused by Aliasing in Lower Resolution Factorials), for more on this subject.)

The F-test on the resulting model was significant at a 99.9% confidence level. Residual analysis showed no abnormalities. Figure 5.4 shows the four combinations of the two significant factors. The best combination (least number of pulls) occurs at low C and high E at the upper left of the square plot.

None of the other factors created a statistical significant effect, so for ease of operation they were set at their most advantageous level. The ideal procedure for minimizing pulls is:

- Prepare engine with three primer pumps and three pulls at 100% choke with 0% gas
- Start at 0% choke at 100% gas

In the confirmation trial, the engine started immediately on the first pull.

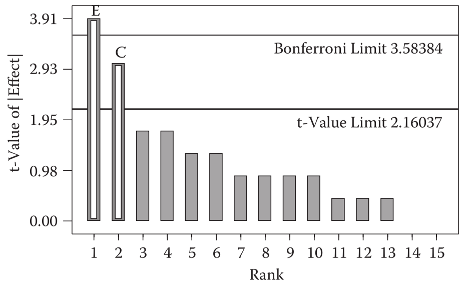


Figure 5.3 Pareto chart of effects for weedwacker.

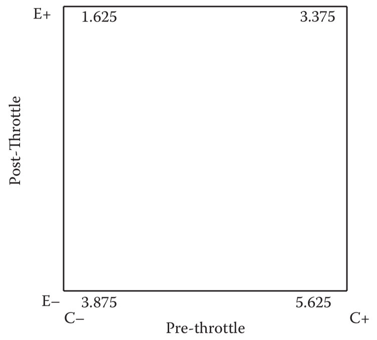


Figure 5.4 Square plot of significant factors for weedwacker.

POWER TO CUT THE GRASS

Fractional factorial designs, such as those shown in this book, exhibit a property called *projection*. This means that the subset of significant factors becomes equivalent to a full-factorial design. For example, in the weedwacker experiment, the design projects to a two-by-two factorial replicated four times. Each point on the square plot in Figure 5.4 is based on an average of four results. This provides greater power to see small effects hidden in the underlying variation, much like an engine-powered weedwacker cuts through tall grass to reveal lost golf balls, etc.

DEATH BY A THOUSAND CUTS?

Think carefully before changing a factor that creates no significant effect. The most conservative approach, one that should be your default, is to leave these factors be, that is, remain at their mid-level set point. Otherwise, you may find that after several iterations a factor ratchets far enough away from a satisfactory setting to cause a noticeable deterioration in process performance. Perhaps a better analogy than the painful method of execution alluded to in this section's title may be the boiling frog story. Supposedly, experiments showed that by increasing water temperature very slowly, a hapless amphibian will never notice itself getting cooked. Al Gore used this unproven story in the 2006 movie *An Inconvenient Truth* to describe humankind's oblivious attitude about global warming. The moral of this story for practitioners of DOE is not to be too cavalier about assuming that insignificant factors do not matter. Perversely, the more variation a process exhibits, the less likely a given experiment will reveal significant effects and the greater the danger of then changing settings to a more convenient level—that is the inconvenient truth.

Potential Confusion Caused by Aliasing in Lower Resolution Factorials

You may be wondering whether the application of fractions to factorial design appears to be too good to be true, and, as noted above, this is a valid point to consider. The price you pay when you cut down the number of runs is the aliasing of effects. This is measured by the "resolution" of the fractional factorial. The half-fractional design on the five weedwacker factors is a resolution V (the resolution is represented by Roman numerals), which indicates aliasing of at least one main effect with one or more four-factor interaction(s), and/or at least one two-factor interaction(s) with one or more three-factor interaction(s). Think of these as relationships of 1 with 4 and 2 with 3, both of which total to 5. As detailed in the sidebar on A Handy Way To Put Your Finger On The Concept Of Resolution, you can figure this out with your fingers!

As you might expect, as resolution decreases, alias problems increase. Let's go to a worst-case scenario for two-level factorials—a resolution III design, which we will demonstrate by revisiting the popcorn case from Chapter 3.

Table 5.4 Popcorn experiment with half of runs (shaded) removed

Std	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*	*Taste*
1	–	–	–	+	+	+	–	74
2	+	–	–	–	–	+	+	75
3	–	+	–	–	+	–	+	71
4	+	+	–	+	–	–	–	80
5	–	–	+	+	–	–	+	81
6	+	–	+	–	+	–	–	77
7	–	+	+	–	–	+	–	42
8	+	+	+	+	+	+	+	32
Effect	–1.0	–20.5	–17.0	0.5	–6.0	–21.5	–3.5	Full
Effect	–22.5	–26.5	–16.5	–16.5	–26.5	–22.5	64.75	Fraction

The unshaded rows in Table 5.4 represent a half-fraction of the original data. It was created by eliminating the minus ABC runs (shaded in the table).

Can you see any problems with the resulting pattern of pluses and minuses in the clear rows?

At first glance, the design looks good, with a nicely balanced pattern for the main effects of A, B, and C. However, upon closer inspection, notice that each effect column has an identical twin in terms of the pattern of pluses and minuses (shown below in parentheses) that you see from top to bottom in the clear rows:

A = BC (+, –, –, +)
B = AC (–, +, –, +)
C = AB (–, –, +, +)

These equalities are called *confounding* relationships, or aliases. The most troublesome of these aliases involves the interaction BC. Recall that the full factorial revealed this to be a significant effect. However, as shown above, the half fraction attributes the BC effect to A, which is completely misleading. The effect labeled "A" is actually the combination of A and BC, which can be expressed mathematically as: A = A + BC. (To check this equation, observe that the effects of A and BC from the full factorial are –1.0 and –21.5, respectively, which sum to the value of –22.5 for A in the half fraction.)

Finally, notice that the last effect column, ABC, displays only plus symbols in the clear rows of Table 5.4. Because there is no contrast in the signs, this effect can no longer be estimated. The calculated value of 64.75, labeled as an effect, is actually an estimate of the overall mean or intercept.

You can now see the downside to creating a fractional design: aliasing of effects. In this case, we cut a full factorial to a half fraction by eliminating the negative ABC rows. The resulting loss of information about the ABC effect is not critical, because three-factor interactions rarely occur. However, you should be very concerned about the aliasing of main effects with two-factor interactions. Statisticians consider such a design to be resolution III, the lowest possible for standard fractional two-level factorials. We recommend that you avoid resolution III designs if at all possible, because if your system contains any interactions, the true cause will be disguised by the aliasing. However, because shortcuts must be taken from time to time under extreme deadlines, we will provide more details on resolution III designs and show how to "de-alias" them in the next chapter.

DESPERATE MEASURES

DOE guru George Box, who passed away in 2013, said that resolution III designs are like kicking the television to make it work. You may succeed, but you won't know which component dropped into place or whatever else actually caused the improvement. The authors concur with this assessment and with Scottish scholar Andrew Lang's colorful exhortation: "[Don't] use statistics as a drunken man uses lamp posts, for support rather than illumination."

So far we have shown examples of resolution V (good) and resolution III (bad). Let's fill the gap with a resolution IV design, which represents a reasonable compromise between minimal runs and maximum information on the main effects. Table 5.5 shows a complete matrix of effects for four factors. The rows where ABCD is minus are shaded. The remaining rows form the half fraction. Look over the resulting columns very carefully, ignoring the gray cells. Can you see the aliases? (Hint: Start in the middle, at AD–BC and work outward.)

In this case, every main effect is aliased with a three-factor interaction, and all the two-factor interactions are aliased with each other, so this design is resolution IV.

Table 5.5 Four-factor design matrix with half fraction unshaded

Std	*A*	*B*	*C*	*D*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	*ABC*	*ABD*	*ACD*	*BCD*	*ABCD*
1	–	–	–	–	+	+	+	+	+	+	–	–	–	–	+
2	+	–	–	–	–	–	–	+	+	+	+	+	+	–	–
3	–	+	–	–	–	+	+	–	–	+	+	+	–	+	–
4	+	+	–	–	+	–	–	–	–	+	–	–	+	+	+
5	–	–	+	–	+	–	+	–	+	–	+	–	+	+	–
6	+	–	+	–	–	+	–	–	+	–	–	+	–	+	+
7	–	+	+	–	–	–	+	+	–	–	–	+	+	–	+
8	+	+	+	–	+	+	–	+	–	–	+	–	–	–	–
9	–	–	–	+	+	+	–	+	–	–	–	+	+	+	–
10	+	–	–	+	–	–	+	+	–	–	+	–	–	+	+
11	–	+	–	+	–	+	–	–	+	–	+	–	+	–	+
12	+	+	–	+	+	–	+	–	+	–	–	+	–	–	–
13	–	–	+	+	+	–	–	–	–	+	+	+	–	–	+
14	+	–	+	+	–	+	+	–	–	+	–	–	+	–	–
15	–	+	+	+	–	–	–	+	+	+	–	–	–	+	–
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

A HANDY WAY TO PUT YOUR FINGER ON THE CONCEPT OF RESOLUTION

To determine the alias structure of a given resolution, count it with your fingers. For a resolution III design hold up three fingers. In this design, at least one main effect (represented by the first finger or thumb) is aliased with at least 1 two-factor interaction (represented by the other two fingers). It's very simple: 1 plus 2 equals 3. If you choose a resolution IV design, hold up four fingers. In this design, at least one main effect (represented by the first finger) is aliased with at least 1 three-factor interaction (represented by the other three fingers). 1 plus 3 equals 4. Another option presents itself in this case: 2 plus 2 also equals 4. This represents the presence of at least one alias between a pair of two-factor interactions. For a resolution V design, you need a hand full of fingers, because main effects are aliased only with four-factor interactions (thumb plus four fingers), and two-factor interactions are aliased only with three-factor interactions (2 plus 3 equals 5).

Assuming that three-factor interactions are unlikely, you can rely on main effect estimates from resolution IV designs, but don't jump to any conclusions on any significant two-factor interactions. For example, if AB shows significance, it might really be due to CD, which changes in exactly the same way: +, +, −, −, −, −, +, + from top to bottom, clear rows only. We will show you how to escape from this trap in the next chapter of the book.

BLOCKING TWO-LEVEL FACTORIALS

What if you had only enough time in a day to do half the runs in a full four-factor design? You could run the half fraction shown in Table 5.5, but perhaps you don't want to take a chance on the resulting aliasing of two-factor interactions. In this case, you could run the first fraction on day 1 and the second fraction on day 2. Within each day, you should randomize the run order. Each day must be considered as a block of time. The block difference can be removed in the analysis, but only at a cost—the estimate of ABCD will be sacrificed. This may be a small price to pay in return for gaining resolution of the two-factor (and three-factor) interactions. Deploying similar approaches to those used for generating fractional designs, statisticians have worked out optimal schemes to block two-level factorials into two, four, or more groups. The objective is to sacrifice a minimal number of lower-order effects. As discussed earlier in this book, the tool of blocking is very powerful for removing known sources of variation, such as time or material or operators.

Not all resolution IV designs are the same. Some have only a few two-factor interactions aliased with each other. For example, the seven factors in 32-run design (1/4 fraction) shown on Table 5.6 are listed as a resolution IV, but only six of the two-factor interactions (those involving D, E, F, and G) are actually aliased (DE = FG, DF = EG, and DG = EF). The other 15 two-factor interactions are aliased only with three-factor or higher interactions. Therefore, you would be wise to assign factors that are least likely to interact with the labels D, E, F, or G, and those factors most likely to interact with A, B, and C.

Table 5.6 Resolutions (Res) for standard two-level designs, some replicated (Rep) with reasonable number of runs

Factors	**2**	**3**	**4**	**5**	**6**	**7**
4 Runs	Full	½ Rep Res III	—	—	—	—
8	2 Rep	Full	½ Rep Res IV	¼ Rep Res III	⅛ Rep Res III	1/16 Rep Res III
16	4 Rep	2 Rep	Full	½ Rep Res V	¼ Rep Res IV	⅛ Rep Res IV
32	8 Rep	4 Rep	2 Rep	Full	½ Rep Res VI	¼ Rep Res IV
64	16 Rep	8 Rep	4 Rep	2 Rep	Full	½ Rep Res VII

On the other hand, you might be tempted to cut back to only 16 runs for the seven-factor design, a 1/8 fraction. After all, as shown in Table 5.6, it too is resolution IV. However, as you can see from the alias structure shown in Table 5.7, no matter how you do the labeling, even one active interaction may cause confusion. For example, let's say that only C and D interact in a particular system. This interaction (CD) will be mislabeled as AG due to the aliasing. It should be apparent by now that you must

Table 5.7 Alias structure for 7 factors in 16 runs (2-factor interactions only)

Labeled as	*Actually*
AB	AB + CE + FG
AC	AC + BE + DG
AD	AD + CG + EF
AE	AE + BC + DF
AF	AF + BG + DE
AG	AG + BF + CD
BD	BD + CF + EG

investigate the alias structure for anything less than resolution V designs. You will find the details on construction of optimal fractions and their alias structures from referenced textbooks by Box et al. and Montgomery, found in Recommended Readings. Good DOE software also will set up the appropriate fraction and supply information on the specific aliases of your chosen design.

Plackett–Burman Designs

The standard two-level designs, which we recommend, provide the choice of 4, 8, 16, 32, or more runs, but only for powers of 2. In 1946, Robin Plackett and J. P. Burman invented alternative two-level designs that are multiples of 4. The 12-, 20-, 24-, and 28-run Plackett–Burman designs are of particular interest because they fill gaps in the standard designs. Unfortunately, these particular Plackett–Burman designs also have very messy alias structures. For example, the 11 factor in the 12-run choice, which is very popular, causes each main effect to be partially aliased with 45 two-factor interactions. In theory, you can get away with this if absolutely no interactions are present, but this is a very dangerous assumption in our opinion.

Of course, you could cut down the number of factors in a given Plackett–Burman, but the alias structure may not be optimal. Assuming that your software provides such information, be sure to look at the alias structure before doing your experiment.

HUSH-HUSH MILITARY SECRETS

Thomas Barker in his book, *Quality by Experimental Design,* 3rd ed. (Chapman and Hall/CRC, 2005), reports that he heard from ASQ's (American Society for Quality) past-president Richard Freund that the motivation for development of Plackett-Burman designs was control of stack-up tolerance for mechanical systems in wartime materials (bomb fuses, for example?), but this remained classified as a military secret when Plackett and Burman published their work in 1946.

Overall, because of the unexpected aliasing that occurs with many Plackett–Burman designs, we recommend that you avoid them in favor of the standard two-level designs. However, if you must use Plackett–Burman,

consider following up with a second set of runs, but not an exact replicate of the initial block. In the next chapter, we show how to do a "foldover" that doubles the number of runs in a way that increases the resolution of highly aliased Plackett–Burmans and standard fractional factorials.

Irregular Fractions Provide a Clearer View

At this point, you have seen several options for the standard two-level design, which apply fractions that are negative powers of 2 (1/2, 1/4, etc.). However, it is possible to use other "irregular" fractions and still maintain a relatively high resolution. A prime example of this is the three-quarter (3/4) replication (rep) for four factors (the first design in Table 5.1). It can be created by identifying the standard quarter fraction, and then selecting two more quarter fractions. This design contains only 12 runs, yet it estimates all main effects and two-factor interactions aliased only by three-factor or higher interactions, thus making it a viable alternative to the 16-run full factorial. (Details on this and other designs of the like can be found in Peter John's *Statistical Design and Analysis of Experiments* (see Recommended Readings).) In Table 5.8 is an example of this irregular fraction with four factors in 12 runs.

In the 1990s, the authors frequently presented computer-intensive workshops on DOE in corporate training rooms equipped with RGB (red-green-blue) projectors. Students often found it difficult to read the projected statistical outputs. To improve readability, co-author Mark investigated four factors at two levels as shown in Table 5.8. The assignment of minus versus plus levels for the categorical variables (B, C, D) is completely arbitrary.

During one of our workshops, students worked together on the exercise. Mark (the instructor) displayed a column of numbers in random order. One student transcribed the data from top to bottom while the other timed it. The actual experiment was performed by many students. By treating each

Table 5.8 Test factors for RGB projector study

Factor	*Name*	*Units*	*Low Level (–)*	*High Level (+)*
A	Font size	Point	10	18
B	Font style	(Categorical)	Arial	Times
C	Background	(Categorical)	Black	White
D	Lighting	(Categorical)	Off	On

Table 5.9 Irregular fraction design on RGB projection

Std	*A: Font Size (point)*	*B: Font Style*	*C: Background*	*D: Lighting*	*E: Readability (seconds)*
1	10 (–)	Arial (–)	Black (–)	Off (–)	52
2	18 (+)	Times (+)	Black (–)	Off (–)	39
3	10 (–)	Arial (–)	White (+)	Off (–)	42
4	18 (+)	Arial (–)	White (+)	Off (–)	27
5	10 (–)	Times (+)	White (+)	Off (–)	37
6	18 (+)	Times (+)	White (+)	Off (–)	31
7	10 (–)	Arial (–)	Black (–)	On (+)	57
8	18 (+)	Arial (–)	Black (–)	On (+)	28
9	10 (–)	Times (+)	Black (–)	On (+)	52
10	18 (+)	Times (+)	Black (–)	On (+)	30
11	18 (+)	Arial (–)	White (+)	On (+)	19
12	10 (–)	Times (+)	White (+)	On (+)	47

student as an individual block, Mark eliminated variability caused by differing distance from the screen and reading ability. However, to keep it simple, the results are shown as one block of data. Table 5.9 provides the complete design with the resulting readability measured in seconds. Shorter times are desired.

The four main effects (A, B, C, D) can be calculated in the usual way by contrasting the average of the plus values with the average of the minus values for the associated response (symbolized by Y). For example, the main effect of factor A is

$$Effect_A = \frac{\sum Y_{A+}}{n_{A+}} - \frac{\sum Y_{A-}}{n_{A-}} = \frac{39+27+31+28+30+19}{6} - \frac{52+42+37+57+52+47}{6}$$

$$= 29.00 - 47.83 = -18.83$$

In other words, with a font size of 18 points, students read the projected display 18.83 seconds faster (on average) than with a font size of 10 points. In this case, the bigger the better. The effect of A (font size), in absolute value, stands out on the half-normal plot (Figure 5.5).

The main effect of C (background) and the interaction AD (font size times lighting) also fall off the normally distributed line of near-zero effects.

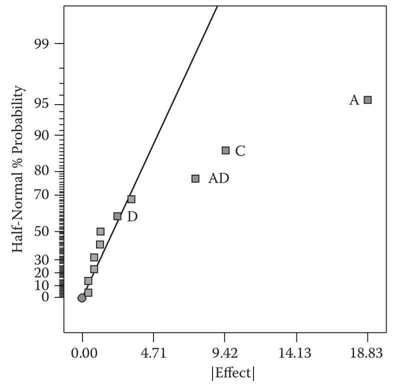


Figure 5.5 Half-normal plot of effects (readability in seconds) from projector DOE.

As usual, we did not label any of the smaller, presumably insignificant effects, with one exception: D. For statistical reasons, which will be discussed later, this main effect must be chosen to support the choice of the AD interaction.

The sharper-eyed (and particular) readers among you may be wondering why 11 points are displayed on Figure 5.5. Ten of the points show estimates of the effects of interest: four main effects plus 6 two-factor interactions. The eleventh point is a three-factor interaction, which provides an estimate of error. Remember that we started with 12 runs. We used one bit of information (degree of freedom) to estimate the overall mean. That leaves us with 11 bits of information, of which 10 get used for designed-for effects, with one left over to use as an estimate of error. There is no reason to throw away information if it can be used to serve some purpose.

WARNING: IRREGULAR FRACTIONS MAY PRODUCE IRREGULARITIES IN EFFECT ESTIMATES

Irregular fractions have somewhat peculiar alias structures. For example, when evaluated for fitting a two-factor interaction model, they exhibit good properties: main effects aliased with three-factor interactions, etc. However, if you fit only the main effects, they become partially aliased with one or more two-factor interactions. Appendix 2.1 provides the gory details for the four-factor in a 12-run design. Mathematically, this is deemed a "nonorthogonal" matrix (see boxed text in Chapter 3 titled Orthogonal Arrays for background). Do not be surprised when analyzing irregular designs like this if your software warns that it is not orthogonal. It should provide a mechanism to recalculate effects after you initially select terms for your model.

Table 5.10 Analysis of variance for RGB readability data

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	1501.58	4	375.40	60.64	< 0.0001
A	*1064.08*	*1*	*1064.08*	*171.89*	*< 0.0001*
C	*266.67*	*1*	*266.67*	*43.08*	*0.0003*
D	*16.67*	*1*	*16.67*	*2.69*	*0.1448*
AD	*168.75*	*1*	*168.75*	*27.26*	*0.0012*
Residual	43.33	7	6.19		
Cor Total	1544.92	11			

Table 5.10 shows the analysis of variance for the RGB data.

You may still be curious as to why we include factor D in the model because it is not significant in the analysis of variance (ANOVA). The factor was chosen to support the significant AD interaction, thus maintaining model "hierarchy" (see Preserving Family Unity below).

PRESERVING FAMILY UNITY

Statisticians advise that you maintain hierarchy in your regression models. The idea of hierarchy can be likened to the traditional structure of a family, with parents and children. In this analogy, a two-factor interaction, such as AD, is considered a child. To maintain family hierarchy you must include both parents (A and D). Similarly, if you select a two-factor interaction for a predictive model, be sure to also select both of the main effects. Otherwise, under certain circumstances, some statistics may not get computed correctly. (For details, see J. L. Peixoto. 1990. A Property of Well-Formulated Polynomial Regression Models. *The American Statistician* February, 44 (1).) An even better reason to abide by this rule is to avoid misleading your audience by saying a factor is not significant when it really does make a difference, albeit only in conjunction with one or more other factors. The RGB experiment is a case in point.

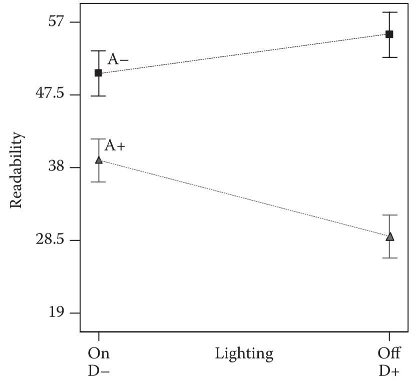


Figure 5.6 Interaction of factors A and D.

Although factor D is not significant on its own, it does have an effect on readability, but only in conjunction with factor A. This becomes very clear when you look at the interaction graph in Figure 5.6 (produced with other factors averaged).

If you split the differences from left (D–) to right (D+), you get a nearly flat line, indicating that D has no effect, thus explaining why it falls on the near-zero effect line shown in the half-normal plot. However, to say that D has no effect makes no sense. Factor D does affect the response, but it depends on the level of Factor A (font size). When A is low (–), increasing D increases the response. However, when A is high (+), increasing D *decreases* the response. In either case, factor D does cause an effect. Therefore, it should not be excluded.

From a practical perspective, the upper line on the AD interaction tells us that students find it more difficult to read the RGB screen when the font size is small (A–) with the lights turned on (D+). The lower line shows that when font size is large enough, it doesn't matter if you turn on the lights; in fact, the readability results actually improved. This was a very desirable outcome, because it allowed students to read their notes while viewing the RGB screen.

One other effect stood out on the half-normal plot: the main effect of C (background). As you can see in Figure 5.7, the readability improves (lower the time, the better) at the high level of this factor, which represents the background set to white.

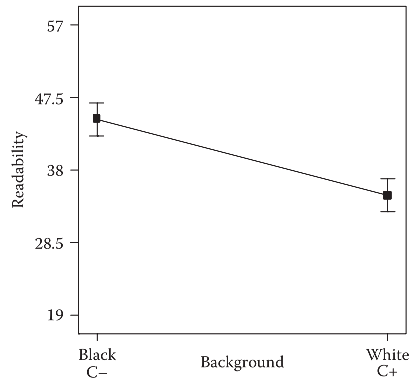


Figure 5.7 Main effect plot for factor C (background).

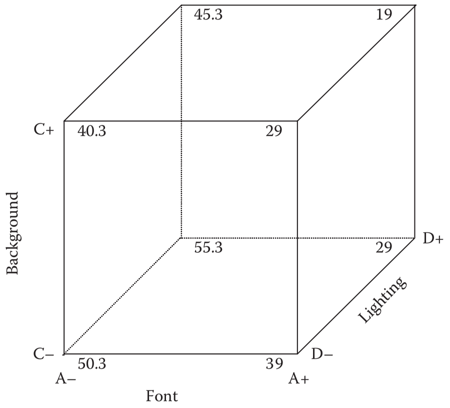


Figure 5.8 Cube plot of RGB (red-green-blue) readability as a function of factors A, C, and D.

Finally, we put all three of the significant factors together in the form of the cube plot shown in Figure 5.8. It shows the best result (19 seconds) at the right, back, and upper corner where all factors are set at their high levels.

Note that B (font style) was not significant, so either style can be chosen. One author likes Arial and the other Times New Roman, but neither is likely to affect readability. Thankfully, projection technology has improved far beyond RGB to a point where its brightness, contrast, and resolution make it far easier to read all the numbers we put up on the screen.

Practice Problem

Problem 5.1

An injection molder wanted to gain better control of part shrinkage. The experimenter set aside two parallel production lines for a study of seven factors (Table 5.11).

All possible combinations of these factors require 128 runs, but only 32 of these were actually done: a 1/4 fraction. This DOE is symbolized mathematically as 2^{7-2}. It is one of the recommended fractional factorial designs for screening (see Table 5.1). From the 32 runs, you can get information on all the main effects and nearly all two-factor interactions. Table 5.12 shows the design matrix in terms of coded factor levels, and the results for shrinkage.

Notice that the experiment is divided into two blocks of 16 runs in a standard way that preserved the greatest possible information on main effects and interactions. The experimenters then ran the DOE on the two parallel lines, greatly reducing the time needed to generate the data, as well as providing information on machine-to-machine variation.

The blocking does cause further loss of information in the form of additional aliasing (revealed by statistical DOE software):

[Block 1] = Block 1 + CDG + CEF + ABDE + ABFG
[Block 2] = Block 2 – CDG – CEF – ABDE – ABFG

Table 5.11 Factors for molding case

	Factor	*Units*	*Low (–)*	*High (+)*
A.	Mold temperature	degrees F	130	180
B.	Cycle time	seconds	25	30
C.	Booster pressure	psig	1500	1800
D.	Moisture	percent	0.05	0.15
E.	Screw speed	inches/sec	1.5	4.0
F.	Holding pressure	psig	1200	1500
G.	Gate size	inches (10^{-3})	30	50

Table 5.12 Design matrix for molding case study

Std	*Run*	*Line*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Shrinkage (%)*
1	19	2	–1	–1	–1	–1	–1	+1	+1	0.833
2	4	1	+1	–1	–1	–1	–1	–1	–1	0.784
3	10	1	–1	+1	–1	–1	–1	–1	–1	0.966
4	25	2	+1	+1	–1	–1	–1	+1	+1	0.898
5	28	2	–1	–1	+1	–1	–1	–1	–1	0.916
6	11	1	+1	–1	+1	–1	–1	+1	+1	1.130
7	3	1	–1	+1	+1	–1	–1	+1	+1	0.760
8	29	2	+1	+1	+1	–1	–1	–1	–1	0.730
9	13	1	–1	–1	–1	+1	–1	–1	+1	0.838
10	17	2	+1	–1	–1	+1	–1	+1	–1	0.669
11	27	2	–1	+1	–1	+1	–1	+1	–1	1.060
12	14	1	+1	+1	–1	+1	–1	–1	+1	0.956
13	8	1	–1	–1	+1	+1	–1	+1	–1	1.780
14	22	2	+1	–1	+1	+1	–1	–1	+1	1.660
15	18	2	–1	+1	+1	+1	–1	–1	+1	1.080
16	12	1	+1	+1	+1	+1	–1	+1	–1	1.230
17	16	1	–1	–1	–1	–1	+1	+1	–1	0.922
18	24	2	+1	–1	–1	–1	+1	–1	+1	0.815
19	20	2	–1	+1	–1	–1	+1	–1	+1	1.100
20	9	1	+1	+1	–1	–1	+1	+1	–1	0.858
21	2	1	–1	–1	+1	–1	+1	–1	+1	1.170
22	30	2	+1	–1	+1	–1	+1	+1	–1	1.040
23	26	2	–1	+1	+1	–1	+1	+1	–1	0.780
24	15	1	+1	+1	+1	–1	+1	–1	+1	1.020
25	21	2	–1	–1	–1	+1	+1	–1	–1	0.939
26	1	1	+1	–1	–1	+1	+1	+1	+1	0.909
27	5	1	–1	+1	–1	+1	+1	+1	+1	1.060
28	31	2	+1	+1	–1	+1	+1	–1	–1	0.916
29	23	2	–1	–1	+1	+1	+1	+1	+1	1.680
30	6	1	+1	–1	+1	+1	+1	–1	–1	1.440
31	7	1	–1	+1	+1	+1	+1	–1	–1	1.330
32	32	2	+1	+1	+1	+1	+1	+1	+1	1.210

Continued

Table 5.12 (*Continued*) Design matrix for molding case study

The loss of the three-factor and higher interactions didn't cause much worry. The experimenters also carefully reviewed the alias structure (see Appendix 2.4) before assigning labels. By labeling the most likely interactors—booster pressure and moisture—as C and D, they avoided deliberate aliasing of this potential effect with other two-factor interactions.

Analyze the data. Look for conditions that minimize and/or stabilize shrinkage. Remember to check the significant factors against the alias structure.

(Suggestion: Use the software provided with the book. Look for a data file called "5-P1 Molding" and open it. Otherwise, create the design by choosing a factorial for 7 factors in 32 runs with 2 blocks and enter the data (from Table 5.12) in standard order. Then do the analysis as outlined in the factorial tutorial that comes with the program.)

Chapter 6

Getting the Most from Minimal-Run Designs

The best carpenters make the fewest chips.

German Proverb

In the previous chapter, we demonstrated how to shave runs from a two-level factorial design by performing only a fraction of all possible combinations. In this chapter, we explore minimal designs with one fewer factor than the number of runs, e.g., seven factors in 8 runs. Statisticians consider such designs to be “saturated” with factors. These resolution III designs confound main effects with two-factor interactions—a major weakness. However, they may be the best option when time and other resources are limited. If you are lucky, nothing will be significant and any questions about aliasing become moot. However, if the results exhibit significance, you must make a big leap of faith to assume that the reported effects are correct. To be safe, do further experimentation (known as *design augmentation*) to de-alias the main effects and/or two-factor interactions. The most popular method of design augmentation is the foldover. We will illustrate this method with a case study.

RUGGEDNESS TESTING

Before transferring any system, such as a product, process, or test method, find out what the recipients will do to it. For example, a small U.S. medical device manufacturer redesigned the electronics in its 220-volt unit for European customers. It worked without fail in all but one country, where every single unit burned out. The country where this occurred was relatively undeveloped, and its power supply varied more than that of any other European country the device manufacturer supplied. One of the components in the new design could not tolerate such wide variation in voltage. After this fiasco, the engineers developed a standard fractional factorial design to test new designs against this variation and half a dozen or so other variables that could affect the units. This led to redesigning components and the overall system to make it more robust.

For a cookbook approach, see ASTM (American Society for Testing and Materials) Standard E1169 “Standard Practice for Conducting Ruggedness Tests,” available for download from American National Standards Institute (ANSI) eStandards Store at http://webstore.ansi.org.

Minimal-Resolution Design: The Dancing Raisin Experiment

The dancing raisin experiment provides a vivid demonstration of the power of interactions. It normally involves just two factors:

1. Liquid: Tap water versus carbonated
2. Solid: A peanut versus a raisin

Only one out of the four possible combinations produces an effect. Peanuts will generally float in water, and raisins usually sink. Peanuts are even more likely to float in carbonated liquid. However, when you drop in a handful of raisins, the results can be delightful. Most will drop to the bottom. There the raisins become coated with tiny bubbles, which lift some of them to the surface. At the surface, the bubbles pop, and some raisins drop to the bottom again. The up-and-down process can continue for some time, creating a “dancing” effect. Conduct this experiment with your family and friends and follow up by getting their ideas on the cause of this interaction of factors. Consider also why some raisins fail to dance, which may be due to their freshness, the specific brand of carbonated liquid, and so forth. For an alternative, try using a popcorn kernel rather than a raisin. These and other factors listed in Table 6.1 became the subject of a two-level factorial design. Note that the aging of the objects, factor G, was accelerated by baking them under a 100-watt lightbulb for 15 minutes.

Table 6.1 Factors for initial DOE on dancing objects

Factor	*Name*	*Low Level (–)*	*High Level (+)*
A	Material of container	Plastic	Glass
B	Size of container	Small	Large
C	Liquid	Club soda	Lemon lime
D	Temperature	Room	Ice cold
E	Cap on container	No	Yes
F	Type of object	Popcorn	Raisin
G	Age of object	Fresh	Stale

The full two-level factorial for seven factors requires 128 runs. This can be computed as 2 * 2 * 2 * 2 * 2 * 2 * 2 = 128 or, more conveniently, by exponential notation: 2^7. We chose the 1/16th fraction (2^{-4} in scientific notation) that requires only 8 runs (= $2^7 * 2^{-4} = 2^{7-4}$ in exponential notation). This is a minimal design with resolution III. At each set of conditions, we rated the dancing performance of 10 objects on a scale of 1 to 5; the higher, the more delightful. Table 6.2 shows the results. The actual order was randomized within one block (Blk) of runs.

The half-normal plot of effects is shown in Figure 6.1.

Three effects stood out: cap (E), age of object (G), and size of container (B). The analysis of variance (ANOVA) on the resulting model revealed highly significant statistics: An F-value of 27.78 with an associated probability of 0.0039, which falls far below the maximum threshold of 0.05.

Table 6.2 Results from first dancing raisin experiment

Std	*Blk*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*Dancing Rating*
1	1	–	–	–	+	+	+	–	1.5
2	1	+	–	–	–	–	+	+	2.0
3	1	–	+	–	–	+	–	+	1.0
4	1	+	+	–	+	–	–	–	4.0
5	1	–	–	+	+	–	–	+	1.5
6	1	+	–	+	–	+	–	–	1.0
7	1	–	+	+	–	–	+	–	5.0
8	1	+	+	+	+	+	+	+	1.0

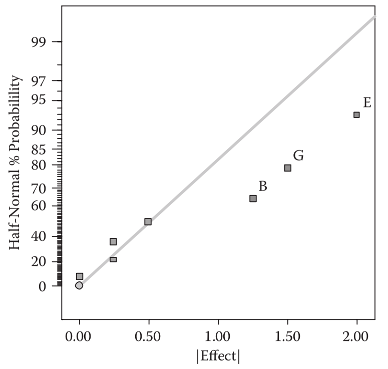


Figure 6.1 Half-normal plot of effects.

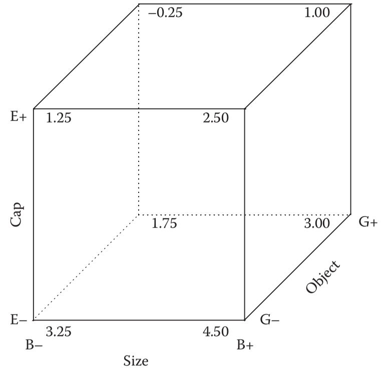


Figure 6.2 Cube plot of predicted responses.

(Reminder: A probability value of 0.05 indicates a 5% risk of a false positive, i.e., saying something happened because of a specific effect when it actually occurred by chance. An outcome like this is commonly reported to be significant at the 95% confidence level.) The cube plot on Figure 6.2 shows the predicted responses for the three listed factors.

The worst rating (lowest) occurs at the upper-left, back corner of the cube; essentially no reaction at all. It reflects negative impacts of stale objects (G+) and capped liquid (E+), both of which make sense. However, the effect of container size (B) does not make much sense. Could this be

Table 6.3 Alias structure for 2^{7-4} design (7 factors in 8 runs)

Labeled as	*Actually*
A	A + BD + CE + FG
B	B + AD + CF + EG
C	C + AE + BF + DG
D	D + AB + CG + EF
E	E + AC + BG + DF
F	F + AG + BC + DE
G	G + AF + BE + CD

an alias for the real culprit, perhaps an interaction? Take a look at the alias structure for this resolution III design, shown in Table 6.3. Each main effect in this experiment is actually aliased with 15 other effects, but we have simplified the table to list only two-factor interactions.

Can you pick out the likely suspect from the lineup for B? Although the number of possibilities may seem overwhelming, they can be narrowed by assuming that the effects form a family.

A VERY SCARY THOUGHT

Could a positive effect be cancelled by an "antieffect"?

If you use a resolution III design, be prepared for the possibility that a positive main effect may be wiped out by an aliased interaction of the same (but negative) magnitude. The opposite could happen as well. Therefore, if nothing significant emerges from a resolution III design, you cannot always be certain that there are no active effects. Two or more big effects may have cancelled each other out.

The obvious alternative to B (size) is the interaction EG. However, this is only one of several feasible "hierarchical" models (ones that maintain family unity):

- E, G, and EG (disguised as B)
- B, E, and BE (disguised as G)
- B, G, and BG (disguised as E)

Figure 6.3 illustrates these three alternative models.

Notice that each figure predicts the same maximum outcome. However, the actual cause remains murky. The EG interaction seems far more plausible than the alternatives, but further experimentation is needed to verify this.

Complete Foldover of Resolution III Design

By adding a second block of runs with signs reversed on all factors, you can break the aliases between main effects and two-factor interactions. This procedure is called a *complete foldover*. It works on any resolution III factorial. It's especially popular with Plackett–Burman designs, such as the 11 factors in 12-run choice. Table 6.4, which shows the second block of experiments with all signs reversed on the control factors, illustrates how the foldover method works on the dancing raisin experiment.

Table 6.4 Second block of runs after complete foldover (selected interactions included)

Std	*Blk*	*A*	*B*	*E*	*C*	*D*	*F*	*G*	*AD*	*BE*	*BG*	*EG*	*Dancing Rating*
1	1	–	–	+	–	+	+	–	–	–	+	–	1.5
2	1	+	–	–	–	–	+	+	–	+	–	–	2.0
3	1	–	+	+	–	–	–	+	+	+	+	+	1.0
4	1	+	+	–	–	+	–	–	+	–	–	+	4.0
5	1	–	–	–	+	+	–	+	–	+	–	–	1.5
6	1	+	–	+	+	–	–	–	–	–	+	–	1.0
7	1	–	+	–	+	–	+	–	+	–	–	+	5.0
8	1	+	+	+	+	+	+	+	+	+	+	+	1.0
9	2	+	+	–	+	–	–	+	–	–	+	–	1.2
10	2	–	+	+	+	+	–	–	–	+	–	–	0.9
11	2	+	–	–	+	+	+	–	+	+	+	+	4.6
12	2	–	–	+	+	–	+	+	+	–	–	+	1.4
13	2	+	+	+	–	–	+	–	–	+	–	–	0.6
14	2	–	+	–	–	+	+	+	–	–	+	–	1.3
15	2	+	–	+	–	+	–	+	+	–	–	+	1.2
16	2	–	–	–	–	–	–	–	+	+	+	+	4.5

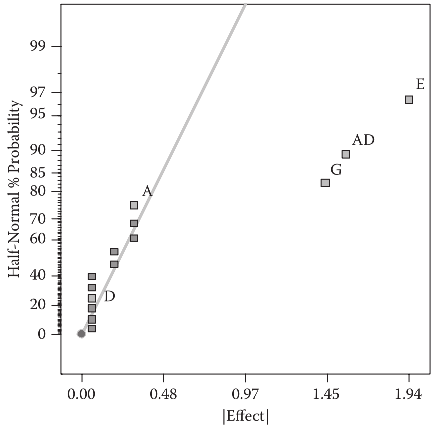


Figure 6.4 Half-normal plot of effects after foldover.

Notice that the signs of the two-factor interactions do not change from block 1 to block 2. For example, in block 1 the signs of columns B and EG are identical, but in block 2 they differ, thus the combined design no longer aliases B with EG. If B really is the active effect, it should come out on the plot of effects for the combined design.

As you can see in Figure 6.4, factor B has disappeared and AD has taken its place. But, hold on. What happened to family unity? Neither of AD's parents (A or D) appear in this chosen model.

The problem is that complete foldover of a resolution III design does not break the aliasing of two-factor interactions, so AD remains aliased with EG as well as CF. The listing of the effect as AD—the interaction of container material with beverage temperature—is arbitrary, by alphabetical order. Figure 6.5 shows the AD interaction with all other factors set arbitrarily at their low levels (specified in Table 6.1). It makes no sense physically for the effect of material (A) to depend on temperature of the beverage (room temperature D– versus ice-cold D+).

Discounting the CF interaction (liquid type versus object type) is not easy, but this new evidence clearly shows that the interaction between E and G is the most plausible, particularly since we now know that these two factors are present as main effects. Figure 6.6 shows the EG interaction.

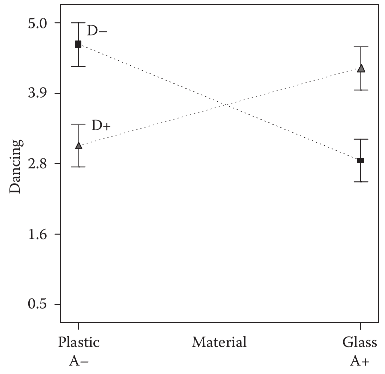


Figure 6.5 Interaction plot for AD.

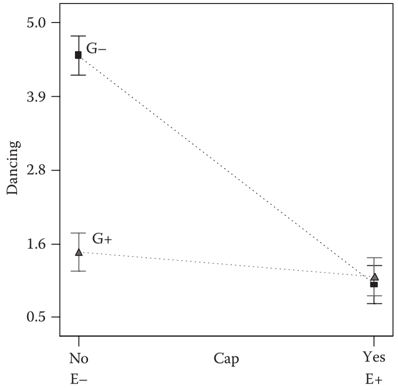


Figure 6.6 Plot of interaction EG.

It appears that the effect of cap (E) depends on the age of the object (G). When the object is stale (the G+ line at the bottom of Figure 6.6), twisting on the bottle cap (going from E– at left to E+ at right) makes little difference. However, when the object is fresh (the G– line at the top), the bottle cap quenches the dancing reaction. More experiments are required to confirm this interaction. One obvious way to do this is to conduct a full factorial on E and G alone. Other ideas on de-aliasing are presented after the next sidebar.

AN ALIAS BY ANY OTHER NAME IS NOT NECESSARILY THE SAME

You might be surprised that aliased interactions, such as AD and EG, do not look alike. Their coefficients are identical, but the plots differ because they combine the interaction with their parent terms. To check this, construct the interaction plot for CF using the methods described in Chapter 3. This interaction also is aliased with AD. By comparing the graphs of aliased interactions (such as AD versus EG versus CF), you can hazard an educated guess about which interaction merits further investigation.

Single-Factor Foldover

Another simple way to de-alias a resolution III design is the "single-factor foldover." Like a complete foldover, this requires a second block of runs, but in this variation of the general method, you change signs on only one factor. This factor and all its two-factor interactions become clear of any other main effects or interactions. To see how this works, go back to the original resolution III design on the dancing raisins. It makes sense to focus on the biggest effect: E (refer to Figure 6.1). The end result of the foldover on factor E (only) is a design with 16 runs in two blocks of eight. The resulting alias structure (main effects and two-factor interactions only) is shown in Table 6.5.

Table 6.5 Alias structure for 2^{7-4} design after foldover on factor E

Labeled as	*Actually*
A	A + BD + FG
B	B + AD + CF
C	C + BF + DG
D	D + AB + CG
E	E
F	F + AG + BC
G	G + AF + CD
AC	AC + BG + DF
AE, BE, CE, DE, EF, EG	AE, BE, CE, DE, EF, EG

The combined design remains at resolution III because, with the exception of E, all main effects remained aliased with 2 two-factor interactions. Factor E is a resolution V, because the main effect is clear of troublesome aliases (anything less than a four-factor interaction), and the two-factor interactions are aliased only with three-factor or higher-order interactions. In the case of the dancing raisins, the single-factor foldover would have revealed that B was actually EG, not AD. On the other hand, factor G remains aliased with 2 two-factor interactions. You can't win either way.

FOLDOVER DOESN'T ALWAYS ADD A NEW WRINKLE

The complete foldover of resolution IV designs may do nothing more than replicate the design so that it remains resolution IV. This would happen if you folded over the 16 runs in Table 6.4. By folding only certain columns of a resolution IV design, you might succeed in de-aliasing some of the two-factor interactions. Other than trying different combinations of columns to fold over, the only sure way to eliminate aliases is the single-factor foldover, which works on resolution IV the same as it would on resolution III designs: The main effect and all the two-factor interactions of the factor you choose will be cleared. If you are really adventurous (or possess design of experiments (DOE) software offering this design augmentation tool), try cutting the single-factor foldover in half to create a "semifoldover." For details, see How to Save Runs, Yet Reveal Breakthrough Interactions, by Doing Only a Semifoldover on Medium-Resolution Screening Designs, a paper presented by Mark Anderson and Patrick Whitcomb at the 55th Annual Quality Congress of the American Society of Quality in Milwaukee, 2001 (online at www.statease.com).

Choose a High-Resolution Design to Reduce Aliasing Problems

The best way to reduce aliasing problems is to run a higher resolution design in the first place by selecting fewer factors and/or a bigger design. For example, in the dancing raisin experiment, we would have prevented much confusion by testing the 7 factors in 32 runs (2^{7-2}). This option was discussed in the previous chapter on fractional factorials. It is a resolution IV design, but all 7 main effects and 15 of the 21 two-factor interactions are clear of other two-factor interactions. The remaining 6 two-factor interactions are shown in Table 6.6.

Table 6.6 Problem aliases for 2^{7-2} design (7 factors in 32 runs)

Labeled	*Actually*
DE	DE + FG
DF	DF + EG
DG	DG + EF

The trick is to label the likely interactors anything but D, E, F, and G. For example, knowing now that capping and age interact in the dancing raisin experiment, we would not label these factors E and G. If only we knew then what we know now.

Another option, one that requires no premonition, is to use a minimum-run resolution V (MR5) design (described in the appendix of this chapter). With just 30 runs, *all* main effects and two-factor interactions can be estimated.

> *"As for the Research Department, the Board feels you should try to find whatever you're looking for the first time you search for it."*
>
> **Caption on cartoon in *American Scientist* 95 (March–April 2007): 120.**

Practice Problems

Problem 6.1

(Warning: The following story contains liberal doses of fantasy.) One of the authors became envious of the skating ability of his co-author. The "wannabe" skater secretly got together with a local manufacturer of in-line skates and borrowed the latest and greatest experimental gear. Bewildered by all the options, he decided to try various combinations at the local domed stadium, which opened its concourse to skaters when not in use for sporting events. The special skates would have to be returned fairly soon, so the wannabe skater set up a quick-and-dirty fractional factorial. If anything proved to be statistically significant, the result would be a faster time around the track

Table 6.7 First experiment on in-line skates

Std	A: Pad	B: Bearing	C: Gloves	D: Helmet	E: Wheels	F: Covers	G: Neon	Time (sec)
1	Out	Old	On	Front	Soft	Off	Off	195
2	In	Old	On	Back	Hard	Off	On	192
3	Out	New	On	Back	Soft	On	On	200
4	In	New	On	Front	Hard	On	Off	165
5	Out	Old	Off	Front	Hard	On	On	190
6	In	Old	Off	Back	Soft	On	Off	195
7	Out	New	Off	Back	Hard	Off	Off	166
8	In	New	Off	Front	Soft	Off	On	201

(and, of course, some ego gratification). The experimenter knew that aliasing of effects would obscure a true picture of what really enhanced speed, but he anticipated that this could be figured out later through follow-up designs using the foldover method. Table 6.7 shows the initial design, a 2^{7-4} resolution III fractional factorial, and the resulting times around the track.

Here is more background on the factors and levels to help you interpret the outcome:

A. Pad goes inside skate to elevate the heel: Out (–), In (+)
B. Bearing constructed either from old material (–) or new high-tech alloy (+)
C. Gloves made specially for in-line skating to protect wrists: On (–), Off (+)
D. Helmet fits with logo to back (–) or front (+): Can't tell which is correct, so try both and ignore laughs when wrong
E. Wheels can be made of hard (–) or soft (+) polymer
F. Covers go on wheels to make them look faster: On (–), Off (+)
G. Neon lighting (from generator on skates) for night-time use: Off (–), On (+)

Analyze the data to see if any of these factors appear to be significant. Do the results make sense? Could the real answer be disguised by an alias? (Suggestion: Refer to Table 6.3. Use the software provided with the book. Create a two-level factorial design for 7 factors in 8 runs and sort the resulting layout by standard order, then enter the time data from Table 6.7. Do the analysis as outlined in the factorial tutorial that comes with the program.)

Table 6.8 Follow-up design (foldover) to initial DOE on in-line skates

Std	A: Pad	B: Bearing	C: Gloves	D: Helmet	E: Wheels	F: Covers	G: Neon	Time (sec)
9	In	New	Off	Back	Hard	On	On	175
10	Out	New	Off	Front	Soft	On	Off	211
11	In	Old	Off	Front	Hard	Off	Off	202
12	Out	Old	Off	Back	Soft	Off	On	205
13	In	New	On	Back	Soft	Off	Off	212
14	Out	New	On	Front	Hard	Off	On	175
15	In	Old	On	Front	Soft	On	On	204
16	Out	Old	On	Back	Hard	On	Off	201

Problem 6.2

This is a continuation of the skating saga from Problem 6.1. Rolling right along, the experimenter decided to do a complete foldover on the initial design. Table 6.8 shows the factor levels and results for this follow-up design.

Add these data to the design from Problem 6.1 and analyze it as a second block of data. Do any of the significant model terms turn out to be interactions rather than main effects? Remember that the foldover upgrades the resolution III design to resolution IV, but interactions may still be aliased with other interactions. If interactions do appear, do they make more sense than the aliased alternatives? (Suggestion: Use the software provided with the book. Look for a data file called "6-P2 Skate2," open it, then do the analysis. View the aliased interactions and try substitutions. Graphically compare the alternatives in a similar manner to that outlined for the dancing raisin case.)

Appendix: Minimum-Run Designs for Screening

In the presence of two-factor interactions, only designs of resolution IV (or higher) can ensure accurate screening. If you are limited by time, materials, or other experimental resources, in most cases (other than eight factors), modern "minimum run" resolution IV (MR4) designs offer savings in runs over the equivalent standard two-level fractional factorial design (2^{k-p}). A nine-factor MR4 design, for example, requires only 18 runs, far fewer than

Table 6.9 MR4 design done to screen a fermentation process

#	*A: Temp. °C*	*B: pH*	*C: O² %*	*D: Load Density*	*E: Complex g/min*	*F: Inducer g/min*	*Yield Density*
1	39	6.8	40	6	4.8	4.8	70.39
2	35	6.8	40	20	4.8	7.2	101.40
3	39	6.8	40	20	3.2	4.8	86.88
4	39	6.8	5	20	4.8	7.2	79.71
5	39	6.8	40	6	3.2	7.2	61.68
6	35	6.8	5	6	3.2	4.8	50.2
7	35	7.4	5	20	3.2	7.2	97.23
8	39	7.4	5	6	3.2	4.8	73.33
9	35	7.4	5	6	4.8	7.2	67.51
10	35	7.4	40	6	3.2	4.8	71.71
11	35	7.4	5	20	4.8	4.8	111.20
12	39	7.4	40	20	4.8	7.2	74.92

the 32 runs needed for a standard 2^{k-p}. In general, the MR4 designs require only two times the number of factors (2k). Table 6.9 details a minimum-run screening design on six factors that a biotechnologist ran on a fermentation process (from: Mark Anderson and Patrick Whitcomb, October 2006, *Chemical Processing*. Online at www.chemicalprocessing.com/articles/2006/166/).

DEFINITIVE SCREENING DESIGNS

Definitive screening designs (DSD) are recently invented near-minimal run (2K+1) with three (not just two) levels of each factor (Jones, B. and C. Nachtsheim. 2011. A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects. *Journal of Quality Technology*, January). DSDs produce clean estimates of main effects. They also generate squared terms, but these are badly aliased with two-factor interactions and, thus, must be approached with caution. For screening continuous numeric factors that can be readily and precisely controlled, running three levels of each, as required by DSD, might be of interest due to it covering the ranges better than simply testing the extremes via more conventional two-level factorials. Normally, when screening commences, it falls on to a planar region, but if you suspect the region will be curvy, then give DSDs strong consideration.

What you do as a result of running such a screening design depends on the statistical analysis of the effects:

Scenario 1: Nothing significant. Look for other factors that affect your response(s).
Scenario 2: Only main effects significant. Change these factors to their best levels.
Scenario 3: Two-factor interaction(s) significant. De-alias by performing a semifoldover.

By following this strategy you will increase your odds of uncovering breakthrough main effects and interactions at a relatively minimal cost in experimental runs.

JUST IN CASE RUNS DO NOT GO AS PLANNED

By choosing minimum-run resolution IV designs, you go to the brink of falling back to an experiment that aliases two-factor interactions with main effects: The loss of one result will push you over the edge. How sure are you that things will not break down somewhere along the way as the experiment progresses? How likely is it that a measurement will be lost? To be on the safe side, choose an MR4 with 2 runs added—an option presented in the software that accompanies this book. Using this option, the six-factor design laid out in Table 6.9 would be tested in 14 runs. Safer yet, and slightly more powerful, would be the 16-run standard resolution IV fraction noted in Table 5.6. When in doubt, build it stout.

PS: For more details on minimum-run resolution IV designs, see the 2004 talk by Anderson and Whitcomb on "Screening Process Factors in the Presence of Interactions," online at www.statease.com.

The development of minimum-run, two-level factorial designs has gone to the next level with resolution V "characterization" designs such as the MR5 (minimum-run resolution five) shown in Table 6.10 for six factors in only 22 runs. The following boxed text provides some background and a reference for more details on the new class of MR5 designs.

Table 6.10 MR5 layout for six factors

Row	*A*	*B*	*C*	*D*	*E*	*F*
1	–	–	+	+	–	+
2	+	+	–	–	–	–
3	–	–	–	–	+	–
4	+	+	+	+	–	–
5	–	+	–	+	–	+
6	+	+	+	–	–	+
7	+	–	–	+	–	+
8	–	+	–	+	+	–
9	+	–	+	–	–	–
10	+	–	+	+	+	+
11	+	–	–	+	+	–
12	+	+	–	+	+	+
13	–	+	–	–	+	+
14	+	+	+	–	–	–
15	–	–	+	+	+	–
16	+	–	–	–	+	+
17	–	+	+	–	–	–
18	–	–	–	–	–	+
19	–	–	–	+	–	–
20	–	+	+	+	+	+
21	+	+	+	–	+	–
22	–	–	+	–	+	+

CUSTOM-MADE OPTIMAL DESIGNS MADE POSSIBLE WITH AID OF COMPUTERS

Most statistical software offering capabilities for design of experiments includes computer-generated test matrices that are custom built to the model you specify. For example, if you want to produce a resolution V

design for 7 factors, specify a model with all 7 main effects, the 21 possible two-factor interactions, and the intercept (overall average of the response) for a total of 29 terms. The computer then performs a search via the algorithm programmed into the DOE software, which typically uses D-optimal criterion. (See Optimality from A to Z (or Maybe Just D) (Chapter 7) of *RSM Simplified* (Productivity Press, 2004).) The MR5 templates noted above were developed in a similar manner, but balanced off to create equal numbers of low and high levels, i.e., "equireplicated." (See Oehlert and Whitcomb's 2002 talk on Small, Efficient, Equireplicated Resolution V Fractions of 2^k Designs, online at www.statease.com.) With this in place, the 7-factor design lays out 30 runs—15 each at the low and high levels. The straight D-optimal design will present some factors with 16 low and 14 high and others with 14 low and 16 high, which is figuratively and literally quite odd.

Chapter 7

General Multilevel Categoric Factorials

> Invention is discernment, choice. … The useful combinations are precisely the most beautiful.
>
> **Henri Poincaré**

In Chapter 2, we showed how to perform simple comparative experiments on individual factors with any number of levels. In the chapters that followed, we shifted to experiments with multiple factors, but restricted to just two levels. You might find it helpful at this point to revisit the flowchart in the Introduction, which directs you to the various types of experiments covered in this book. Your decision on which designs will prove most useful depends on:

- Number of factors
- Number of levels
- Nature of factors: numerical or categorical, process or mixture.

Quite often, you may be confronted with a number of categorical alternatives, such as three suppliers (A versus B versus C), as well as other categorical factors that could interact with the first. A good example of this is what happens when brewing several varieties of coffee with a number of flavoring additives—a second categorical factor. A simple solution to this

problem is to run all combinations of all factors. Unfortunately, this “general factorial” design is not very popular and with good reason:

1. The number of combinations becomes excessive after only a few factors. It may be possible to perform fractional designs, but except for special cases documented in the literature, you must resort to computer-aided selection via matrix-based methods. This goes well beyond the scope of *DOE Simplified*, but see the boxed text on optimal design at the end of Chapter 6 for some clues.
2. Each design is unique to the given situation, with many possible levels for the specified varying numbers of factors. Therefore, you may not find a template or example that you can follow for design and analysis.
3. Due to the already large number of combinations, general factorials for three or more factors are often done without replication. Because no pure error is available, you must assume that highest-order interactions are insignificant and build your analysis of variance (ANOVA) from this basis.
4. Predictive models for categorical factors must be coded in a nonintuitive manner.

To avoid these complications, stick with simple comparisons and two-level factorials. In the coffee example cited above, for example, you can begin by screening several varieties of coffee via a one-factor, simple-comparative taste test. After eliminating all but the top two varieties, combine these with your two favorite flavor additives, with additional factors at two levels, such as brewing temperature low versus high, in a standard factorial. It is important to note, however, this simplistic approach to experiment design may overlook critical interactions that will be revealed in a general factorial design. For example, perhaps one of the coffee varieties eliminated early on may have worked really well if paired with a particular flavor additive. Therefore, rather than taking the simpler route, we will detail the more comprehensive general factorial choice in the following case study.

Putting a Spring in Your Step: A General Factorial Design on Spring Toys

A coiled spring, made to specific dimensions (see boxed text below), will gracefully “walk” down an incline. The most obvious factor affecting speed is the degree of incline, which must be between 20 and 40 degrees. If the incline is too shallow, the coil will not move. Too steep an incline causes the coil to tumble or roll out of control. Friction also plays a role. During our coiled-spring experiment, after observing slippage on bare wood, we added a high-friction rubber mat for the walking surface. Many other variables are associated with the construction of the coil, such as the spring constant, mass, diameter, and height.

ONE PERSON WHO KNEW IT WAS A SLINKY

Richard James was a naval engineer. During World War II, he worked on springs for keeping sensitive instruments steady at sea. One day, he accidentally knocked an experimental spring off a table onto a pile of books. The spring tumbled each step of the way in a delightful walking motion. After seeing the reaction of neighborhood kids to his new toy, James decided to develop it commercially. The first Slinky hit the Philadelphia market in 1945. It became an instant success. The Slinky, now made by Poof-Slinky, Inc., in Pennsylvania, remains popular not only with children, but also with physics teachers who use the toy to illustrate wave properties and energy states. The original Slinky measures 87 feet when stretched (a bit of trivia you can use to impress your friends).

What walks down stairs, alone or in pairs, and makes a
slinkity sound?
A spring, a spring, a marvelous thing, Everyone knows it's Slinky …
It's Slinky, it's Slinky, for fun it's a wonderful toy
It's Slinky, it's Slinky, it's fun for a girl and a boy

Advertising jingle that most Baby Boomers will never forget.

Several varieties of coiled springs are made by Poof-Slinky, including the Slinky and the smaller Slinky Junior. They come in traditional metal or styrene-butadiene plastic. The first two "bent" springs (resting in curved formation) shown in Figure 7.1 (starting from bottom left and then going on up the picture) are a metal Slinky in traditional size and a plastic Slinky Junior. We also looked at a larger plastic Giant Slinky (shown in the hands of tester in Figure 7.1), but it walked too slowly and frequently stopped.

Two additional generic plastic spring toys (the coiled ones in the picture) were tested. The smaller of these walked too fast for valid comparison to the Slinky brands. That left three coiled spring toys to test at two inclines in a full factorial experiment. We replicated each of the six combinations (3 * 2) in a completely randomized test plan. The 12 results are sorted by standard order in Table 7.1.

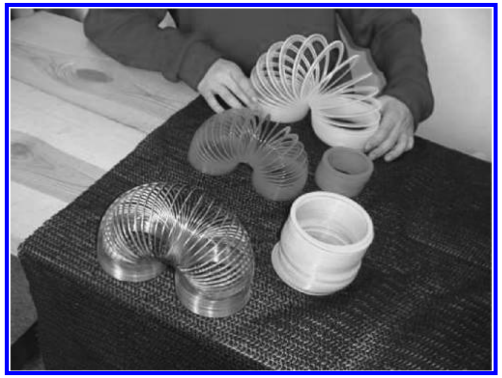


Figure 7.1 Various spring toys on high-friction rubber mat.

Table 7.1 General factorial design (replicated) on spring toys

Std	*A: Spring toy*	*B: Incline*	*Time (seconds)*
1, 2	Metal Slinky	Shallow	5.57, 5.75
3, 4	Slinky Junior	Shallow	5.08, 5.36
5, 6	Generic plastic	Shallow	3.03, 3.34
7, 8	Metal Slinky	Steep	4.67, 4.95
9, 10	Slinky Junior	Steep	4.23, 4.98
11, 12	Generic plastic	Steep	3.58, 4.50

The response is time in seconds for the springs to walk a four-foot inclined plank. The two results per cell represent the replicated runs. As noted earlier, these runs were performed at random, not one after the other, as shown. Therefore, the difference in time reflects the variations due to setting of the board, placement of the coil, how the operator made it move, and so forth. The ANOVA is seen in Table 7.2. Procedures for calculating the sums of squares are provided in referenced textbooks and encoded in many statistical software packages.

Table 7.2 ANOVA for general factorial on spring toys

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	7.73	5	1.55	10.96	0.0056
A	*5.90*	*2*	*2.95*	*20.90*	*0.0020*
B	*0.12*	*1*	*0.12*	*0.88*	*0.3848*
AB	*1.71*	*2*	*0.85*	*6.05*	*0.0365*
Residual	0.85	6	0.14		
Cor Total	8.58	11			

Look first at the probability value (Prob > F) for the model. Recall that we consider values less than 0.05 to be significant. (Reminder: A probability value of 0.05 indicates a 5% risk of a false positive, i.e., saying something happened because it was influenced by a factor when it actually occurred due to chance. An outcome like this is commonly reported to be significant at the 95% confidence level.) In this case, the probability value of 0.0056 for the model easily passes the test for significance, falling well below the 0.05 threshold. Because we replicated all six combinations, we obtained six degrees of freedom (df) for estimation of error. This is shown in the line labeled *Residual*. The three effects can be individually tested against the residual. The result is an unexpectedly large interaction (AB). This is shown graphically in Figure 7.2.

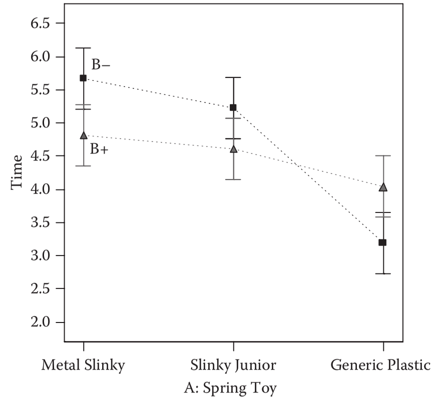


Figure 7.2 Interaction plot for walking times at shallow (B–) versus steep (B+) incline.

The two Slinky springs behaved as expected by walking significantly faster down the steeper incline (B+). However, just the opposite occurred with the generic plastic spring. It was observed taking unusually long, but low steps down the shallow slope, much like the way the comedic actor Groucho Marx slouched around in the old movies. While it didn't look very elegant, it was fast. On the steep slope, the generic spring shifted into the more pleasing, but slower and shorter high-step, like the Slinky toys.

THE ELEPHANT "GROUCHO RUN"

A study published in the April 3 issue of Nature *solves a longstanding mystery about elephant speeds by clocking the animals at 15 miles per hour. … Their footfall pattern remains the same as that in walking, and never do all four feet leave the ground at the same time. … Biomechanists have dubbed this gait "Groucho running." …*

For their experiments … a Stanford postdoctoral research fellow in the Department of Mechanical Engineering … marked [elephants'] joints with large dots of water-soluble, nontoxic paint. They videotaped 188 trials of 42 Asian elephants walking and running through a 100-foot course and measured their speed with photosensors and video analysis.

(From Dawn Levy. 2003. Stanford Report, April 3. Online at http://news-service.stanford.edu/news/2003/april9/elephants-49.html (Published in the April 3, 2003 issue of *Nature*, p. 493—"Biomechanics: Are fast-moving elephants really running?")

How to Analyze Unreplicated General Factorials

The example shown above involved only six combinations for the full factorial, so doing a complete replicate of the design was feasible. When you add a third factor, the cost of replication goes up twofold or more, depending on how many levels you must run. In this case, it's common practice to assume that interactions of three or more factors are not significant. In other words, whatever effects you see for these higher order interactions are more than likely due to normal process variability. Therefore, these effects can be used for estimating error. Other insignificant interactions also may be pooled as error in much the same way as with two-level factorials discussed earlier in this book. A new example based on the same toys as before illustrates this concept in a way that keeps things simple and fun.

Before conducting the earlier case study on the springs, we rejected two abnormal toys:

- A larger one: The "original plastic Slinky"
- A smaller one: A generic plastic variety

In this new example, we will go back and test these two springs, along with two from the earlier test, excluding only the generic plastic toy. The factor of incline will be retained at two levels. A third factor will be added, which is whether an adult or a child makes the springs walk. The resulting 16 combinations (4 * 2 * 2) are shown in Table 7.3, along with the resulting times.

The impact of the three main effects and their interactions can be assessed by analysis of variance. In general factorials like this, it helps to view a preliminary breakdown of sum of squares (a measure of variance) before doing the actual ANOVA. In many cases you will see that some or all of the two-factor interactions contribute very little to the overall variance. These trivial effects then can be used as estimates of error, in addition to the three-factor or higher interactions already destined for error. Table 7.4 provides the sum of squares, degrees of freedom, and the associated mean squares for the time data from the experiment on spring toys.

Table 7.3 Second experiment on spring toys

Std	*A: Spring Toy*	*B: Incline*	*C: Operator*	*Time Seconds*
1	Metal Slinky	Shallow	Child	5.57
2	Slinky Junior	Shallow	Child	5.08
3	Plastic Slinky	Shallow	Child	6.37
4	Small Generic	Shallow	Child	3.03
5	Metal Slinky	Steep	Child	4.67
6	Slinky Junior	Steep	Child	4.23
7	Plastic Slinky	Steep	Child	4.70
8	Small Generic	Steep	Child	3.28
9	Metal Slinky	Shallow	Adult	6.51
10	Slinky Junior	Shallow	Adult	5.21
11	Plastic Slinky	Shallow	Adult	6.25
12	Small Generic	Shallow	Adult	3.47
13	Metal Slinky	Steep	Adult	4.88
14	Slinky Junior	Steep	Adult	3.39
15	Plastic Slinky	Steep	Adult	6.72
16	Small Generic	Steep	Adult	2.80

Table 7.4 Breakdown of variance for second experiment on spring toys

Term	*Sum of Squares*	*DF*	*Mean Square*
A	18.68	3	6.23
B	2.91	1	2.91
C	0.33	1	0.33
AB	0.88	3	0.29
AC	1.03	3	0.34
BC	0.014	1	0.014
ABC	1.71	3	0.57
Total	25.55	15	

The main effect from factor A ranks first on the basis of its mean square. The main effect of B comes next. The other main effect, from factor C, contributes relatively little to the variance. Continuing down the list, you see fairly low mean squares from each of the 3 two-factor interactions (AB, AC, BC). Finally, you get to the three-factor interaction effect ABC, which (as stated earlier) will be used as an estimate of error. Because none of the other interactions caused any larger mean square, these also will be thrown into the error pool (labeled "Residual" in the ANOVA). The rankings in terms of mean square (the last column in Table 7.4) provide support for this decision. Although it appears to be trivial, the C term will be kept in the model because it is a main effect. The resulting main-effects-only (A, B, C) model is the subject of the ANOVA shown in Table 7.5.

The model passes the significance test with a probability of less than 0.001 (>99.9% confidence). As expected, A and B are significant, as indicated by their probability values being less than 0.05. By this same criterion, we conclude that the effect of factor C (child operator versus adult operator) is not significant. The probability is high (0.3626) that it occurred by chance. In other words, it makes no difference whether the adult or the child walked the spring toy. The effect of A—the type of spring toy—can be seen in

Table 7.5 ANOVA for selected model (main effects only)

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	21.92	5	4.38	12.06	0.0006
A	*18.68*	*3*	*6.23*	*17.14*	*0.0003*
B	*2.91*	*1*	*2.91*	*8.00*	*0.0179*
C	*0.33*	*1*	*0.33*	*0.91*	*0.3626*
Residual	3.63	10	0.36		
Cor Total	25.55	15			

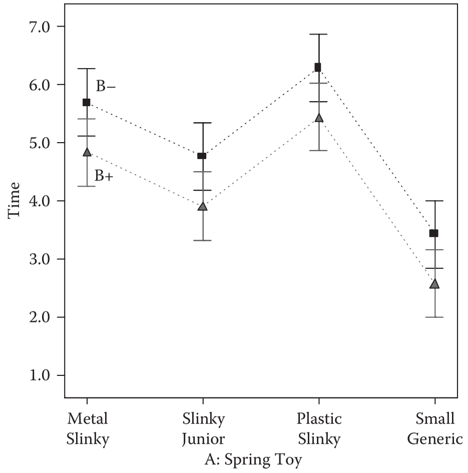


Figure 7.3 Plot for walking times at shallow (B–) versus steep (B+) incline.

Figure 7.3, the interaction plot for AB. (Normally we would not show an interaction plot when the model contains only main effects, but it will be helpful in this case for purposes of comparison.)

Compare Figure 7.3 with Figure 7.2. Notice that in this second design the lines are parallel, which indicates the absence of an interaction. The effects of A and B evidently do not depend on each other. You can see from Figure 7.3 that the shallower incline (B–) consistently produced longer predicted times. The main-effect plot in Figure 7.4 also shows the effect of B, but without the superfluous detail on the type of spring.

QUANTIFYING THE BEHAVIOR OF SPRING TOYS

A more scientific approach to understanding the walking behavior of the spring toys is to quantify the variables. To do this, you would need a spring-making machine. Then you could make coils, of varying height and diameter, from plastics with varying physical properties. The general factorial indicates that walking speeds increase as the spring diameter decreases, but the exact relationship remains obscure. Similarly, the impact of incline is known only in a general way. By quantifying the degree of incline, a mathematical model can be constructed. We will provide an overview of ways to do this in Chapter 8 on "response surface methods" (RSM).

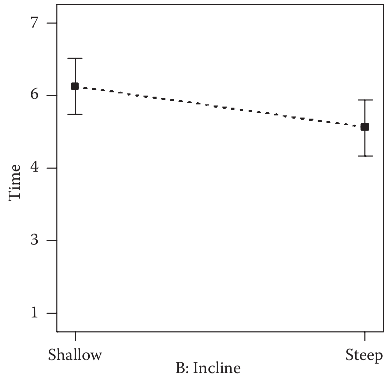


Figure 7.4 Main effect plot for incline.

As expected, the spring toys walked faster down the steeper slope.

This example created a template for general factorial design. Setup is straightforward: Lay out all combinations by crossing the factors. With the aid of computers, you can create a fractional design that fits only your desired model terms, but this goes beyond the scope of the current book. Check your software to see if it supports this "optimal design" feature. If you don't include too many factors or go overboard on the number of levels, you might prefer to do the full factorial. Replication is advised for the two-factor design, but it may not be realistic for larger factorials. In that case, you must designate higher-order interactions as error and hope for the best.

If at all possible, try to redesign your experiment to fit a standard two-level factorial. This is a much simpler approach. Another tactic is to quantify the factors and do a response surface method design (see above boxed text).

Practice Problems

Problem 7.1

Douglas Montgomery describes a general factorial design on battery life (see Recommended Readings, *Design and Analysis of Experiments*, 2012, ex. 5.3.1). Three materials are evaluated at three levels of temperature. Each experimental combination is replicated four times in a completely randomized design. The responses from the resulting 36 runs can be seen in Table 7.6.

Which, if any, material significantly improves life? How are results affected by temperature? Make a recommendation for constructing the battery. (Suggestion: Use the software provided with the book. First do the tutorial on general factorials that comes with the program. It's keyed to the data in Table 7.6.)

Problem 7.2

One of the authors (Mark) lives about 20 miles due east of his workplace in Minneapolis. He can choose among three routes to get to work: southern loop via freeway, directly into town via stop-lighted highway, or northern loop on the freeway. The work hours are flexible so Mark can come in early or late. (His colleagues frown on him coming in late and leaving early, but they don't object if he comes in early and leaves late.) In any case, the time of Mark's commute changes from day to day, depending on weather conditions and traffic problems. After mulling over these variables, he decided that the only sure method for determining the best route and time would be to conduct a full factorial replicated at least twice. Table 7.7 shows the results from this general factorial design. The run order was random.

Table 7.6 General factorial on battery (response is life in hours)

Material Type	*Temperature (Deg F)*					
	15		*70*		*125*	
A1	130	155	34	40	20	70
	74	180	80	75	82	58
A2	150	188	136	122	25	70
	159	126	106	115	58	45
A3	138	110	174	120	96	104
	168	160	150	139	82	60

Table 7.7 Commute times for different routes at varying schedules

Std	*A: Route*	*B: Depart*	*Commute minutes*
1, 2	South	Early	27.4, 29.1
3, 4	Central	Early	29.0, 28.7
5, 6	North	Early	28.5, 27.4
7, 8	South	Late	33.6, 32.9
9, 10	Central	Late	30.7, 29.1
11, 12	North	Late	29.8, 30.9

Analyze this data. Does it matter which route the driver takes? Is it better to leave early or late? Does the route depend on the time of departure? (Suggestion: Use the software provided with the book. Look for a data file called "7-P2 Drive," open it, then do the analysis. View the interaction graph. It will help you answer the questions above.)

Appendix: Half-Normal Plot for General Factorial Designs

Stat-Ease statisticians developed an adaptation of the half-normal plot—so very handy for selecting effects for two-level factorial designs—for similar use in the analysis of experimental data from general factorial designs. This is especially useful for unreplicated experiments, such as the second one done on spring toys. Notice, for example, how the main effect of factor A (and to some extent B) stands out in Figure 7.5.

The other (unlabeled) five effects, C, AB, AC, BC, and ABC, line up normally near the zero-effect level and, in all likelihood, they vary because

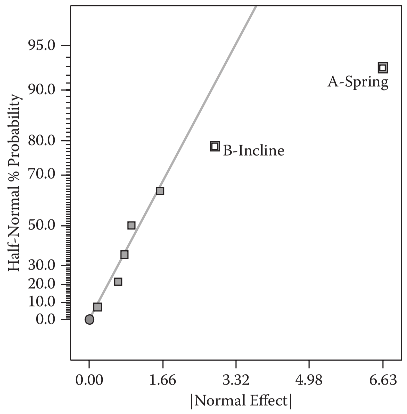


Figure 7.5 Half-normal plot of effects from second spring toy experiment.

of experimental error (noise). This graphical approach provides a clear signal that A and B ought to be tested via the analysis of variance (ANOVA) for statistical significance. In this case, having access to this half-normal plot (not yet invented when the initial baseline experiment was performed) would have saved the trouble of testing the main effect of factor C for significance (Table 7.5). For details, see the proceeding for "Graphical Selection of Effects in General Factorials" by Patrick J. Whitcomb and Gary W. Oehlert. Paper submitted for the 2007 Fall Technical Conference of the American Society for Quality (ASQ) and American Statistical Association (ASA), October.

Chapter 8

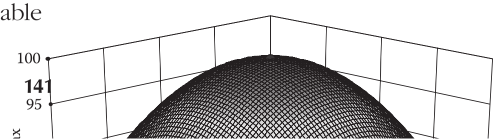
Response Surface Methods for Optimization

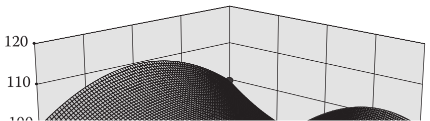
> The first step comes in a burst. Then "bugs," as such little faults and difficulties are called, show themselves. Months of intense study and labor are required before commercial success.
>
> **Thomas Edison (1878)**

This chapter provides a broad overview of more advanced techniques for optimization called *response surface methods* or RSM. (For a much more detailed description of this technique, read *RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments* (Productivity Press, 2004).) RSM should be applied only after completing the initial phases of experimentation:

1. Fractional two-level designs that screen the vital few from the trivial many factors.
2. Full factorials that study the vital few factors in depth and define the region of interest.

The goal of RSM is to generate a map of response, either in the form of contours or as a 3-D rendering. These maps are much like those used by a geologist for topography, but instead of displaying elevation they show your response, such as process yield. Figure 8.1 shows examples of response surfaces. The surface on the left exhibits a "simple maximum," a very desirable





outcome because it reveals the peak of response. The surface on the right, called a “saddle,” is much more complex. It exhibits two maximums. You also may encounter other types of surfaces, such as simple minimums or rising ridges.

A two-level design cannot fit the surfaces shown in Figure 8.1, but it can detect the presence of curvature with the addition of “center points.” We will show how to add center points; then, if curvature is significant, how to augment your design into an RSM.

Center Points Detect Curvature in Confetti

Center points are created by setting all factors at their midpoint. In coded form, center points fall at the all-zero level. An experiment on confetti easily illustrates this concept.

The objective is to cut strips of paper that drop slowly through the air. If the dimensions are optimal, the confetti spins and moves at random angles that please the eye. Table 8.1 shows the factors and levels to be tested. Note the addition of center points (coded as 0). This is a safety measure that plugs the gap between low (–) and high (+) levels.

For convenience of construction, the confetti specified above is larger and wider than the commercially available variety. The actual design is shown in Table 8.2. We replicated the center point four times to provide more power for the analysis. These points, along with all the others, were performed in random order. The center points act as a barometer of the variability in the system.

Table 8.1 Two-level factorial with center points for confetti

Factor	*Name*	*Units*	*Low Level (–)*	*Center (0)*	*High Level (+)*
A	Width	Inches	1	2	3
B	Height	Inches	3	4	5

TRUE REPLICATION VERSUS REPEAT MEASUREMENT

It often helps to repeat response measurements many times for each run. For example, in the confetti experiment, each strand was dropped 10 times and then averaged. However, this cannot be considered a true replicate because some operations are not repeated, such as the cutting of the paper. In the same vein, when replicating center points, you must repeat all the steps. For example, it would have been easy just to reuse the 2 × 4-inch confetti, but we actually recut to this dimension four times. Therefore, we obtained an accurate estimate of the “pure error” of the confetti production process.

Table 8.2 Design layout and results for confetti experiment

Std	*A: Width (inches)*	*B: Length (inches)*	*Time (seconds)*
1	1.00	3.00	2.5
2	3.00	3.00	1.9
3	1.00	5.00	2.8
4	3.00	5.00	2.0
5	2.00	4.00	2.8
6	2.00	4.00	2.7
7	2.00	4.00	2.6
8	2.00	4.00	2.7

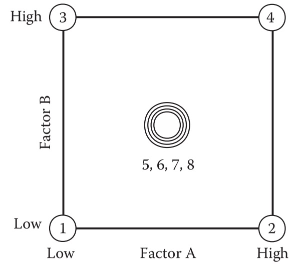


Figure 8.2 Two-level factorial design with center point(s).

Figure 8.2 shows where the design points are located.

The response is flight time in seconds from a height of five feet. The half-normal plot of effects for this data is shown in Figure 8.3.

Factor A, the width, stands out as a very large effect. On the other end of the effect scale (nearest zero), notice the three triangular symbols. These come from the replicated center points, which contribute three degrees of freedom for estimation of “pure error.” In line with the pure error, you will

ADDING A CENTER POINT DOES NOT CREATE A FULL THREE-LEVEL DESIGN

The two-level design with center point(s) (shown in Figure 8.2) requires all factors to be set at their midlevels around which you run only the combinations of the extreme lows and highs. It differs from a full three-level factorial, which would require nine distinct combinations, including points at the midpoints of the edges. The two-level factorial with center point(s) will reveal curvature in your system, but it does not provide the complete picture that would be obtained by doing a full three-level factorial.

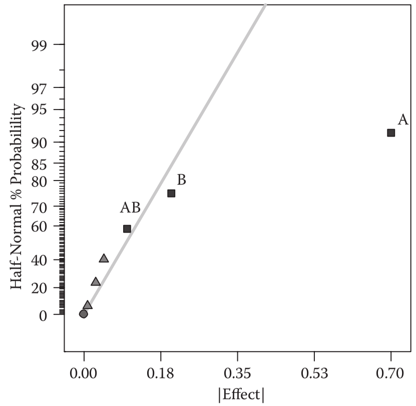


Figure 8.3 Half-normal plot of effects for confetti experiment.

find the main effect of B (length) and the interaction AB. These two relatively trivial effects, chosen here only to identify them, will be thrown into the residual pool under a new label, “lack of fit,” to differentiate these estimates of error from the “pure error.” (Details on lack of fit can be found below in the boxed text on this topic.) The pure error is included in the residual subtotal in the analysis of variance (ANOVA), shown in Table 8.3, which also exhibits a new row labeled “Curvature.”

Apply the usual 0.05 rule to assess the significance of the curvature. In this case, the probability value of 0.005 for curvature falls below the acceptable threshold of 0.05, so it cannot be ignored. That’s bad. It means

Table 8.3 ANOVA for confetti experiment (effects B and AB used for lack-of-fit test)

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Model	0.49	1	0.49	35.00	0.0020
A	*0.49*	*1*	*0.49*	*35.00*	*0.0020*
Curvature	0.32	1	0.32	22.86	0.0050
Residual	0.070	5	0.014		
Lack of Fit	*0.050*	*2*	*0.025*	*3.75*	*0.1527*
Pure Error	*0.020*	*3*	*0.0067*		
Cor Total	0.88	7			

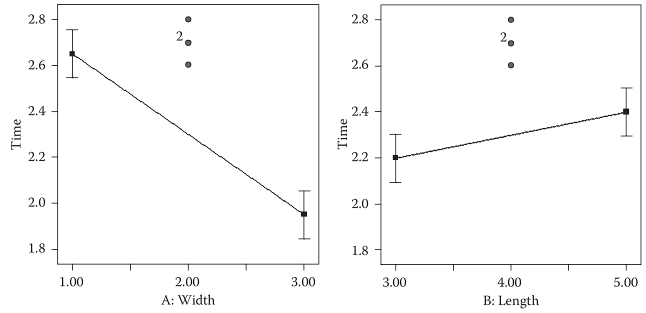


Figure 8.4a,b Effect plots confetti flight time versus factors A and B (not significant).

that the results at the center point were unexpectedly high or low relative to the factorial points around it. In this case, as illustrated by Figure 8.4 effect plots of the response versus factors A and B (insignificant at this stage), the center points fall much higher than one would expect from the outer ones.

The relationships obviously are not linear. Notice that the center point responses stay the same in both plots. (Disregard the slope shown for the effect of factor B, which as one can see from the overlapping LSD bars at either end, is not significant.) Because all factors are run at their respective midlevels, we cannot say whether the observed curvature occurs in the A or the B direction, or in both. Statisticians express this confusion as an alias relationship: Curvature = $A^2 + B^2$. It will take more experimentation to pin this down. The next step is to augment the existing design by response surface methods (RSM).

LACK OF FIT MAY BE FASHIONABLE, BUT IT IS NOT DESIRABLE FOR EXPERIMENTERS

You may have noticed a new line in the ANOVA called *lack of fit*. This tests whether the model adequately describes the actual response surface. It becomes possible only when you include replicates in your design. The lack-of-fit test compares the error from excess design points (beyond what is needed for the model) with the pure error from the replicates. As a rule of thumb, a probability value of 0.05 or less for the F-value indicates a significant lack of fit—an undesirable result.

Augmenting to a Central Composite Design (CCD)

The remedy for dealing with significant curvature in two-level factorial design is to add more points. By locating the new points along the axes of the factor space, you can create a central composite design (CCD). If constructed properly, the CCD provides a solid foundation for generating a response surface map. Figure 8.5 shows the two-factor and three-factor CCDs.

For maximum efficiency, the "axial" (or "star") points should be located a specific distance *outside* the original factor range. The ideal location can be found in textbooks or provided by software, but it will be very close to the square root of the number of factors. For example, for the two-factor design used to characterize confetti, the best place to add points is 1.4 coded units from the center. The augmented design is shown in Table 8.4. The new points are designated as block 2. The additional center points provide a link between the blocks and add more power to the estimation of second-order effects needed to characterize curvature.

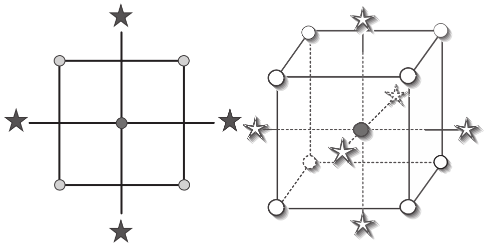


Figure 8.5a,b Central composite designs for two and three factors, respectively.

Table 8.4 Central composite design for confetti

Std	*Block*	*Type*	*A: Width (Inches)*	*B: Length (Inches)*	*Time (Seconds)*
1	1	Factorial	1.00	3.00	2.5
2	1	Factorial	3.00	3.00	1.9
3	1	Factorial	1.00	5.00	2.8
4	1	Factorial	3.00	5.00	2.0
5	1	Center	2.00	4.00	2.8
6	1	Center	2.00	4.00	2.7
7	1	Center	2.00	4.00	2.6
8	1	Center	2.00	4.00	2.7
9	2	Axial	0.60	4.00	2.5
10	2	Axial	3.40	4.00	1.8
11	2	Axial	2.00	2.60	2.6
12	2	Axial	2.00	5.40	3.0
13	2	Center	2.00	4.00	2.5
14	2	Center	2.00	4.00	2.6
15	2	Center	2.00	4.00	2.6
16	2	Center	2.00	4.00	2.9

The CCD contains five levels of each factor: low axial, low factorial, center, high factorial, and high axial. With this many levels, it generates enough information to fit a second-order polynomial called a *quadratic*. Standard statistical software can do the actual fitting of the model. The quadratic model for confetti flight time is

$$\text{Time} = 2.68 - 0.30\ \text{A} + 0.12\ \text{B} - 0.050\ \text{AB} - 0.31\ \text{A}^2 + 0.020\ \text{B}^2$$

This model is expressed in terms of the coded factor levels shown in Table 8.1. The coding eliminates problems caused by varying units of measure, such as inches versus centimeters, which can create problems when comparing coefficients. In this case, the A-squared (A^2) term has the largest coefficient, which indicates curvature along this dimension. The ANOVA, shown in Table 8.5, indicates a high degree of significance for this term and the model as a whole. Notice that the AB and B^2 terms are insignificant,

Table 8.5 ANOVA for CCD on confetti

Source	*Sum of Squares*	*DF*	*Mean Square*	*F Value*	*Prob > F*
Block	0.016	1	0.016		
Model	1.60	5	0.32	15.84	0.0003
A	*0.72*	*1*	*0.72*	*35.47*	*0.0002*
B	*0.12*	*1*	*0.12*	*5.77*	*0.0397*
A^2	*0.75*	*1*	*0.75*	*37.37*	*0.0002*
B^2	*0.003*	*1*	*0.003*	*0.15*	*0.7031*
AB	*0.010*	*1*	*0.010*	*0.50*	*0.4991*
Residual	0.18	9	0.020		
Lack of Fit	*0.071*	*3*	*0.024*	*1.30*	*0.3578*
Pure Error	*0.11*	*6*	*0.018*		
Cor Total	1.79	15			

but we let them be because there is no appreciable benefit to eliminating them from the model; the response surface will not be affected one way or the other (with or without these two terms).

Lack of fit is not significant (because the probability value of 0.3578 exceeds the threshold value of 0.05) and diagnosis of residuals showed no abnormality. Therefore, the model is statistically solid. The resulting contour graph is shown in Figure 8.6.

Each contour represents a combination of input factors that produces a constant response, as shown by the respective labels. The actual runs are shown as dots. (The number 8 by the center point indicates the number of replicates at this set of conditions. In other words, at eight random intervals throughout the experiment, we reproduced confetti with the midpoint dimensions of 2 × 4 inches.) Normally, we would restrict the axes to the factorial range to avoid extrapolation beyond the experimental space, but here we wanted to show the entire design. Notice the darkened areas outside of the actual design space, especially in the corners. These represent areas where predictions will be unreliable due to lack of information. Figure 8.7 shows a 3-D response surface with the ranges reduced to their proper levels. It bisects the center points nicely (those above the surface in black and those below displayed in gray).

The maximum flight time within this factorial range occurs at a width of 1.44 inches and length of 5 inches. Longer confetti might fly even longer, but this cannot be determined without further experimentation.

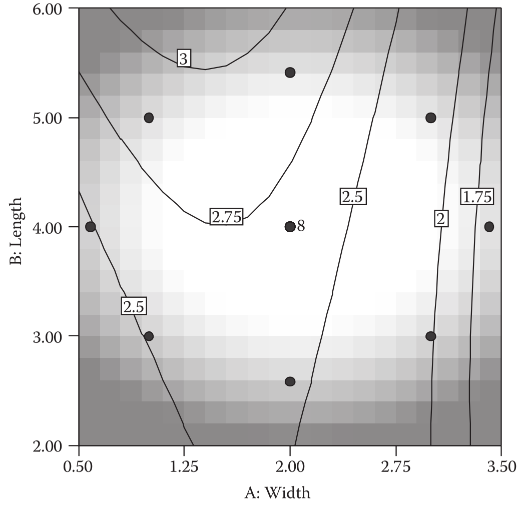


Figure 8.6 Contour graph for confetti flight time.

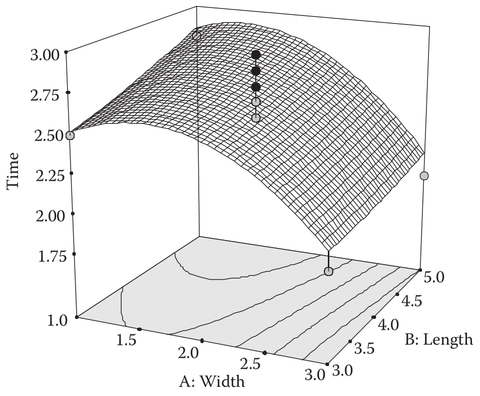


Figure 8.7 Response surface for confetti flight time.

WHERE THERE'S SMOKE, THE PROBABILITY IS HIGH THERE'S A FIRE

A scientist, engineer, and statistician watched their research lab burn down as a result of combustion of confetti in a paper shredder. They speculated as to the cause.

"It's an exothermic reaction," said the scientist.

"That's obvious," replied the engineer. "What's really important is the lack of heat transfer due to inadequate ventilation."

The two technical professionals then turned to their statistical colleague, who said, "I have no idea what caused the fire, but I can advise that you do some replication. Let's burn down another lab and see what happens."

Finding Your Sweet Spot for Multiple Responses

Leaders of projects for process improvement, such as Six Sigma Black Belts, quickly learn that they cannot break the "iron triangle" shown in Figure 8.8. The triangle depicts the unavoidable tradeoffs that come with any attempt to make the highest-quality product on schedule at minimal cost.

When pressured by never-ending demands of management, point to this triangle and ask, "Which two of these three things do you want—cheaper, better or faster?" While this response may not be diplomatic, it is very realistic. It may be possible to produce a top-quality product within schedule, but only at high cost. Going for a lowest-cost product in the fastest possible time will invariably cause a decline in quality. And, if quality and cost are tightly guarded, the production timeline will almost certainly be stretched. In other words, you cannot achieve the ideal level at all three objectives simultaneously.

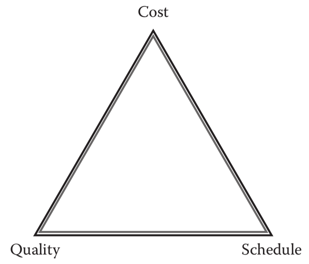


Figure 8.8 The iron triangle of tradeoffs in process improvement.

Fortunately, a tool called *desirability*—when coupled with optimization algorithms—can achieve the best compromise when dealing with multiple demands. The application of desirability (or some other single objective function, such as overall cost) is essential for response surface methods, but is of lesser value for the simpler two-level factorial designs that are the focus of this book. Nevertheless, you may find it helpful for confirming what you ferret out by viewing plots of main effects and any two-factor interactions.

For example, in the popcorn case, it becomes evident that the best compromise for great taste with a minimal number of unpopped kernels (bullets) will be achieved by running the microwave at high power for the shorter time. Figure 8.9 illustrates this most desirable outcome with dots located along the number lines that ramp up (8.9a) for taste (goal: maximize) and down for bullets (goal: minimize).

You can infer from these figures that the desirability scale is very simple. It goes from zero (d = 0) at the least to one at the most (d = 1).

The predicted taste rating of 79 highlighted in Figure 8.9a exceeds the minimum threshold of 65, but it falls short of perfection: a rating of 100. (The smaller numbers (32 – 81) benchmark the experimental range for the taste response.) Similarly, the weight of bullets falls below the maximum threshold

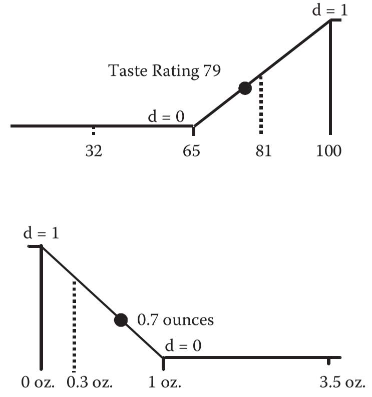


Figure 8.9 (a) Desirability ramp for taste; (b) desirability ramp for bullets.

of 1 ounce, which is desirable, but it comes up short of perfection—none at all (0 ounces). The overall desirability, symbolized "D," is computed as

$$d_1 = \frac{79 - 65}{100 - 65} = \frac{14}{35} = 0.40$$

$$d_2 = \frac{1 - 0.7}{1 - 0} = \frac{0.3}{1} = 0.30$$

$$D = (0.40 * 0.30)^{\frac{1}{2}} = \sqrt{0.12} = 0.35$$

where the lower case ds represent the individual responses for taste (d_1), a measure of quality, and bullets (d_2), an indicator for the yield of the microwave popcorn process. This is a multiplicative or "geometric" mean rather than the usual additive one. Thus, if any individual response falls out of specification, the overall desirability becomes zero.

GOLDILOCKS AND THE TWO BEARS: A DESIRABILITY FABLE

Think back to the old story of Goldilocks and the three bears and imagine that the Mama bear has passed away, leaving only Papa Bear and Baby Bear. Everything in the home now comes in two flavors, two sizes, etc.

Now grown a bit from before, Goldilocks comes along and sits in each of the two chairs remaining. One is too small and the other is too large, but on average she finds comfort.

You can see where this is going. It makes no sense to be off equally bad at both ends of a specification and state that everything is all right on an arithmetic average. This really is a fable: Goldilocks needs the missing mother's chair that would sit just right.

In real life, there seems to be an iron-clad rule that tradeoffs must be made. Desirability analysis may find a suitable compromise that keeps things sweet with your process management and product customers, but this will work only if you design a good experiment that produces a statistically valid model within a region that encompasses a desirable combination of the responses of interest. That takes a lot of subject matter knowledge, good design of experiments (DOE), and perhaps a bit of luck.

Chapter 9

Mixture Design

> The good things of life are not to be had singly, but come to us with a mixture.
>
> **Charles Lamb**

The cliché image of experimentation is a crazed person in a lab coat pouring fluorescent liquids into bubbling beakers. Ironically, the standard approaches for design of experiments (DOE) don't work very well on experiments that involve mixtures. The following illustration shows why.

For example, what would happen if you performed a factorial design on combinations of lemonade and apple juice? Table 9.1 shows the experimental layout with two levels of each factor—either one or two cups.

Notice that standard orders 1 and 4 call for mixtures with the same ratio of lemonade to apple juice. The total amount varies, but will have no effect on responses, such as taste, color, or viscosity. Therefore, it makes no sense to do the complete design. When responses depend only on proportions and not the amount of ingredients, factorial designs don't work very well.

Another approach to this problem is to take the amount variable out of the experiment and work on a percentage basis. Table 9.2 shows the layout for a second attempt at the juice experiment, with each "component" at two levels: 0 or 100%.

This design, which asks for impossible totals, does not work any better than the first design. It illustrates a second characteristic of mixtures: The total is "constrained" because ingredients must add up to 100%.

Table 9.1 Factorial design on mixture of fruit juices

Std Order	*A: Lemonade (cups)*	*B: Apple juice (cups)*	*Ratio*
1	1	1	1/1
2	2	1	2/1
3	1	2	1/2
4	2	2	1/1

SHOTGUN APPROACH TO MIXTURE DESIGN

Chemists are famous for creating mysterious concoctions seemingly by magic. A typical example is Hoppe's Nitro Powder Solvent Number 9, invented by Frank August Hoppe. While fighting in the Spanish–American War, Captain Hoppe found it extremely difficult to ream corrosion from his gun barrel. The problem was aggravated by the antagonistic effects of mixing old black powder with new smokeless powder. After several years of experimenting in his shed, Hoppe came up with a mixture of nine chemicals that worked very effectively. A century or so later, his cleaning solution is still sold. The composition remains a trade secret.

Table 9.2 Alternative factorial design on fruit juices in terms of percentage

Std Order	*A: Lemonade (%)*	*B: Apple juice (%)*	*Total (%)*
1	0	0	0
2	100	0	100
3	0	100	100
4	100	100	200

Two-Component Mixture Design: Good as Gold

Several thousand years ago, a jewelry maker discovered that adding copper to gold reduces the melt point of the resulting mixture. This led to a breakthrough in goldsmithing, because small decorations could be soldered to a main element of pure gold with a copper-rich alloy. The copper blended in with no noticeable loss of luster in the finished piece.

Table 9.3 lays out a simple mixture experiment aimed at quantifying this metallurgical phenomenon.

Table 9.3 A mixture experiment on copper and gold

Blend	*A: Gold (wt%)*	*B: Copper (wt%)*	*Melt Point (deg C)*
Pure	100	0	1039, 1047
Binary	50	50	918, 922
Pure	0	100	1073, 1074

This is a fully replicated design on the pure metals and their binary (50/50) blend. For a proper replication, each blend must be reformulated, not just retested for melt point. If the blends are simply retested for melt point without reformulation, the only variation is due to testing, not the entire process, and error will be underestimated. It is also critical to randomize the run order. Don't do the same formulation twice in a row, because you will more than likely get results that reflect less than the normal process variation.

Notice the depression in melt point in the blend. This is desirable and, therefore, an example of "synergistic" behavior. It can be modeled with the following equation:

$$\text{Melt point} = 1043.0\ A + 1073.5\ B - 553.2\ AB$$

where components A and B are expressed in proportional scale (0 to 1). This second-order "mixture model," developed by Henri Scheffé, is easy to interpret. The coefficients for the main effects are the responses for the purest "blends" for A and B. The second-order term AB looks like a two-factor interaction, but in the mixture model, it's referred to as a *quadratic*, or sometimes *nonlinear*, blending term. The negative coefficient on AB indicates that a combination of the two components produces a response that is less than what you would expect from linear blending. This unexpected curvature becomes more obvious in the response surface graph shown in Figure 9.1.

GOOD OR BAD: WATCH OUT FOR NONLINEAR BLENDING

There are two ways that components can unexpectedly combine: positively (synergism) or negatively (antagonism). With synergism, you get a better response than what you would expect to get from simply adding the effects of each ingredient alone. In other words, adding one to one gives you more than two. For example, if you combined two different types of firecrackers and got an explosion like an atomic bomb, you would be the beneficiary (victim?) of synergism. On the other hand, with antagonism, you get a poorer response than what you would expect from the combination of ingredients. As an analogy, one of the authors (Mark) observed antagonism at an early age between his two similarly aged sons, and two similarly aged daughters. These are classic cases of sibling rivalry or, in the parlance of mixture design, negatively nonlinear blending. A child may act very positively on his or her own, but in the presence of a sibling, they compete in a negative way for parental attention. Although antagonism is the norm, siblings can act synergistically, such as when they happily play together. Wouldn't that be nice?

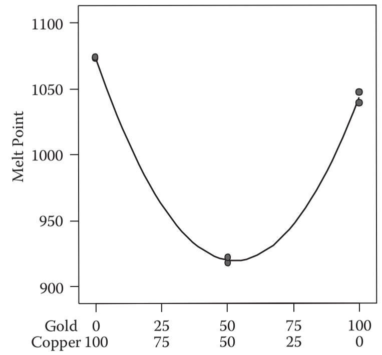


Figure 9.1 Response surface for melt point of copper–gold blend.

This is a highly simplified view of the actual behavior of copper and gold mixtures. To pin it down further, one would need to add a number of check blends to fill in the blanks between these compositions. The predicted value for equal amounts of gold and copper is

$$\text{Melt point} = 1043.0\ (0.5) + 1073.5\ (0.5) - 553.2\ (0.5 * 0.5)$$
$$= (1043.0 + 1073.5)/2 - 553.2/4 = 1058.25 - 138.3 = 919.95$$

The equation predicts a deflection of 138.3°C from the melt point of 1058.25 that one would expect from linear blending; simply the average of the melt points for the two metals. Statistical analysis of the overall model and the interaction itself reveals a significant fit.

WORTH ITS WEIGHT IN GOLD?

An ancient king suspected that his goldsmith had mixed some silver into a supposedly pure gold crown. He asked the famous mathematician Archimedes to investigate. Archimedes performed the following experiment:

1. Create a bar of pure gold with the same weight as the crown.
2. Put the gold in a bath tub. Measure the volume of water spilled.
3. Do the same with the crown.
4. Compare the volumes.

Archimedes knew that silver would be less dense than gold. Therefore, upon finding that the volume of the crown exceeded the volume of an equal weight of gold, he knew that the crown contained silver. According to legend, the naked Archimedes then ran from his bath into the street shouting, “Eureka!” (Greek for “I have found it.”)

The principles of mixture design can be put to work in this case. Gold and silver have densities of 10.2 and 5.5 troy ounces per cubic inch, respectively. Assume that no density interactions exist between silver and gold. We then can apply a linear mixture model to predict weight (in ounces) of one cubic inch (enough to make a crown?) as a function of proportional volume for gold (A) versus silver (B).

Weight = 10.2 A + 5.5 B

Notice that the coefficients of the model are simply the densities of the pure metals. The weight of a blend of half gold and half silver is calculated as follows:

Weight = 10.2 (0.5) + 5.5 (0.5) = 7.85 troy ounces per cubic inch

Three-Component Design: Teeny Beany Experiment

Formulators usually experiment on more than two components. A good example of this is a mixture design on three flavors of small jelly candies called *teeny beanies*. Table 9.4 shows the blends and resulting taste ratings. The rating system, similar to the one used for the popcorn experiments in Chapter 3, goes from 1 (worst) to 10 (best). Each blend was rated by a panel, but in a different random order for each taster. The ratings were then averaged. Several of the blends were replicated to provide estimates of pure error.

Table 9.4 Teeny beany mixture design (three-component)

Blend	*A: Apple %*	*B: Cinnamon %*	*C: Lemon %*	*Taste Rating*
Pure	100.00	0.00	0.00	5.1, 5.2
Pure	0.00	100.00	0.00	6.5, 7.0
Pure	0.00	0.00	100.00	4.0, 4.5
Binary	50.00	50.00	0.00	6.9
Binary	50.00	0.00	50.00	2.8
Binary	0.00	50.00	50.00	3.5
Centroid	33.33	33.33	33.33	4.2, 4.3

This design is called a *simplex centroid* because it includes a blend with equal proportions of all the components. The centroid falls at the center of the mixture space, which forms a simplex—a geometric term for a figure with one more vertex than the number of dimensions. For three components, the simplex is an equilateral triangle. The addition of a fourth component creates a tetrahedron, which looks like a three-sided pyramid. Because only two-dimensional space can be represented on a piece of paper or computer screen, response data for mixtures of three or more components are displayed on triangular graphs such as the one shown in Figure 9.2.

Notice that the grid lines increase in value as you move from any of the three sides toward the opposing vertex. The markings on this graph are in terms of percentage, so each vertex represents 100% of the labeled component. The unique feature of this trilinear graph paper is that only two components need to be specified. At this point, the third component is fixed. For example, the combination of X_1 at 33.3% (1/3rd of the way up from the bottom) plus X_2 at 33.3% (1/3rd of the distance from the right side to lower left vertex) is all that is required to locate the centroid (shown on the graph). You then can read the value of the remaining component (X_3), which must be 33.3% to bring the total to approximately 100%.

Figure 9.3 shows a contour map fitted to the taste responses for the various blends of teeny beanies.

Notice the curves in the contours. This behavior is modeled by the following second-order Scheffé polynomial:

Taste = 5.14 A + 6.74 B + 4.24 C + 4.12 AB − 7.28 AC − 7.68 BC

Analysis of variance (not shown) indicates that this model is highly significant. Component B, cinnamon, exhibits the highest coefficient for the main

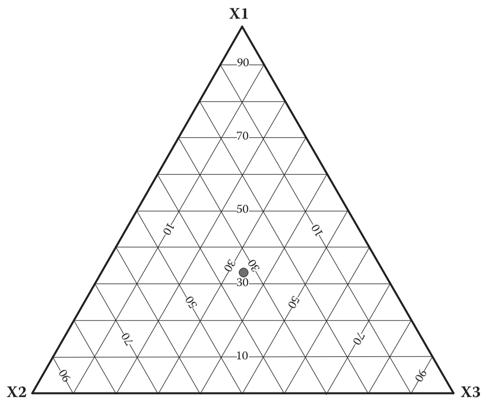


Figure 9.2 Trilinear graph paper for mixtures with point at centroid.

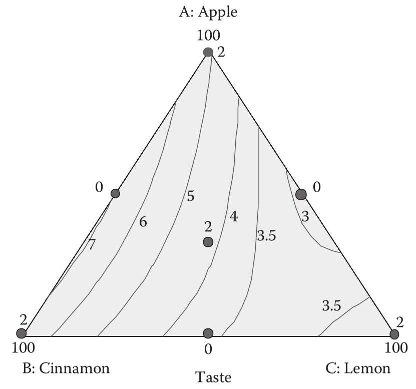


Figure 9.3 Contour graphs for teeny beanies.

effects. Thus, one can conclude that the best pure teeny beanie is cinnamon. Component C, lemon, was least preferred. Looking at the coefficients for the second-order terms, you can see by the positive coefficient that only the AB combination (apple–cinnamon) was rated favorably. The tasters gave low ratings to the apple–lemon (AC) and lemon–cinnamon (BC) combinations. These synergisms and antagonisms are manifested by upward and downward curves along the edges of the 3-D response surface shown in Figure 9.4.

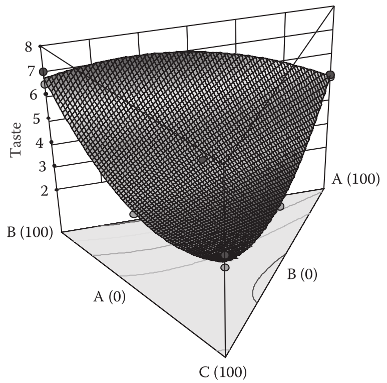


Figure 9.4 Response surface for taste of teeny beanies.

This concludes our overview of mixture design, but we have only scratched the surface of this DOE tool geared for chemists of food, pharmaceuticals, coatings, cosmetics, metals, plastics, etc. For more details, see “A Primer on Mixture Design: What’s In It for Formulators?” online at www.statease.com/formulator and study the referenced texts by Cornell or Smith.

Chapter 10

Back to the Basics: The Keys to Good DOE

> When running an experiment, the safest assumption is that unless extraordinary precautions are taken, it will be run incorrectly.
>
> **George Box, J. Stuart Hunter, and William Hunter (see Recommended Readings)**

In the last few chapters, we provided a high-level glimpse at the most sophisticated tools for design of experiments. Now we come back to Earth with the ground-level basics of design of experiments (DOE), in particular the planning stage. Barring luck, any experiment that is misguided from the start will almost certainly fail. In such cases, you can try to salvage some information with help from a statistician. However, in this role, statisticians (as Sir Ronald Fisher observed) are more like pathologists than physicians. They cannot bring an experiment back to life, they can only determine why it died.

LOST IN TRANSLATION

The Six Sigma quality movement brings statistical tools, such as DOE, to nontechnical professionals.

A group of salespeople, for example, completed a web-based training module assigned from the first edition of this book. At a follow-up session (face-to-face), one of the authors (Mark) used an in-class experiment to test what the group had learned. These aspiring Six Sigma experts decided to do whatever they felt like run-by-run, rather than be governed by a standard test matrix. For example, one of the factors was set at its low level for the first run and high for all the others. None of the factors were balanced between their two levels, and the matrix as a whole was utterly haphazard. Needless to say, this poorly designed experiment produced nothing much of value, but it provided a great learning experience on what not to do. Aided by the DOE software provided for the class, the group then laid out a standard, two-level factorial design for their second try. This experiment proved successful.

"That's not an experiment you have there, that's an experience."

Sir Ronald A. Fisher

A Four-Step Process for Designing a Good Experiment

In our workshops on DOE, we drill our students to rigorously follow a four-step planning process:

1. Define the objective in terms of measurable responses.
2. Calculate the "signal-to-noise ratio" (needed for calculating "power" later on):
   a. Determine the difference (symbolized by delta: Δ_y) that, at a minimum, is important to detect for each response.
   b. Estimate experimental error for *each* response in terms of its overall standard deviation (σ).
   c. Divide the signal by the noise for the ratio Δ/σ.
3. Select the input factors to study and set their levels—the farther apart the better for generating a difference in response Δ_y.

FISHING FOR FACTORS VIA A CAUSE-AND-EFFECT DIAGRAM

It does no good to experiment on the wrong factors, so it is very important not to overlook any variables that may affect your process. We recommend that at this third step in the experiment design process you gather a group of subject matter experts, as many as a dozen. Designate a leader who will be responsible for maintaining a rapid flow of ideas. Another individual should record all the ideas as they are presented.

To provide some structure for brainstorming, make use of the fishbone diagram as shown in Figure 10.1. Label the five big fish bones any way you like, but quality professionals typically categorize them by five major causes: Material, Method, Machine, People, and Environment (spine).

The first thing your group must do is agree on the objective. Write this down at the head of the fishbone diagram. For example, let's say that you and your friends form a team for a fishing contest aimed at catching the highest total weight for the limit of bass per day. That becomes your objective in measurable terms (Step 1 of the experiment design process). Next, start collecting ideas and recording them, such as the type of fishing rod and lure under Material and whether you will troll or cast under Method, and so forth.

Encourage participants not to be overly negative about others' ideas. By remaining open-minded, they can branch off from what may at first seem to be somewhat absurd. Be inventive. Encourage participation by asking everyone to note variables on sticky notes that then can be posted on the fishbone diagram and moved around as desired.

Before too long, you will get a big bone pile with lots of branching into dozens of potential variables—far more than you need for a single experiment. At this point, for the sake of efficiency, consider paring the group down to three key people. Why three? Because then it will be two against one if disagreements arise; no chance of stalemates on the tough decisions. They must critically evaluate the collection of variables. Those variables not chosen as factors for the experiment should be held fixed if possible, or perhaps blocked. Record any other variable that can be monitored, e.g., ambient temperature.

A fun, but practical, idea for prioritizing variables is to give each evaluator 100 units of imaginary currency to invest in his/her favorite variables. Tally up the totals and pick the top ones for your experiment. This gives everyone a say and helps to weed out variables that are not especially useful.

> *"It is easier to tone down a wild idea than to think up a new one."*
>
> **Alex Osborne**

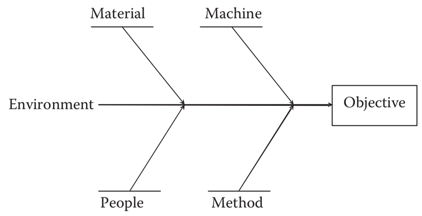


Figure 10.1 Fishbone diagram for a brainstorm on factors.

4. Select a design, then:
   a. Evaluate aliases, if any (refer to Chapter 5).
   b. Determine statistical power—the probability that the experiment will reveal important effects, assuming they occur, which depends on factors chosen in Step 3 and their levels.
   c. Examine the design layout to ensure all the factor combinations can be run at their specified levels safely and that they will produce meaningful results. Suggestion: Consider doing some pre-experiment range-finding.

The hardest part of all this is getting a handle on the power, the ability of your experiment design to detect an effect of a specified size, generally stated relative to the standard deviation of the process (in other words, the signal-to-noise ratio). Power is expressed as a probability. It depends on the risk level, symbolized by α (alpha), you establish for saying an effect is real when it may be caused by chance. The generally acceptable risk is 0.05 or 5%. However, what really matters is that power is directly related to the number of experimental runs. For example, Table 10.1

Table 10.1 Power for designs on five factors for various signal-to-noise ratios

Runs = >	*8*	*12*	*16*	*32*
Resolution = >	*III*	*Min-Run IV*	*V*	*Full*
Signal/Noise Ratio (Δ_y/σ)				
0.5	7.3%	9.0%	14.8%	27.5%
1	13.8%	21.3%	44.0%	77.7%
2	35.7%	61.9%	94.9%	99.9%
4	80.0%	99.1%	99.9%	>99.9%

shows power calculated at a 5% type I risk level to detect main effects for two-level designs on five factors. (To compute these power statistics, we used the software that accompanies this book. For the details on the calculations, see *Sizing Fixed Effects for Computing Power in Experimental Designs* by Gary W. Oehlert and Pat Whitcomb online at www.statease.com.)

Notice how power increases as more runs are performed and the signal-to-noise ratio goes up. Ideally, your design will achieve at least 80% power, but it is possible to overdo this. For example, consider this scenario:

- You want to screen only for the five main effects.
- Runs are very expensive.
- The signal-to-noise ratio is 4 (not unheard of at early experimental stages on pilot systems where engineers can be bold in setting factor levels).

In this case, a minimum-run resolution IV design of 12 runs would be plenty powerful (99.1%). Perhaps the 16-run option might be chosen to achieve the higher resolution V that would uncover two-factor interactions, but it would be very wasteful to do the full 32-run factorial design. Even the 8-run design offers a minimally acceptable level of power (80%), but it achieves only resolution III, which is insufficient for screening purposes because main effects become aliased with two-factor interactions.

You would be wise to run through this four-step process on the highest priority response first. It may take a few iterations before an experiment can be reconciled, i.e., a reasonable number of runs will probably detect an important difference, provided this effect is produced by the chosen factors at the changes in levels specified by the design.

EXPERIMENT ON HOTEL ROOM LAYOUT

In the early 1980s, one of the authors was offered $20 to view eight hotel rooms laid out by a factorial design with varying layouts. At this time, hotels were not well differentiated for varying types of traveler, such as a business person versus a vacationing family. The experiment included rooms with well-lit work spaces and ergonomic desk chairs. Some had separate sitting areas and others had big closets. Bathroom layouts ranged from simplistic to luxurious.

The hotel that devised this experiment brought in sufficient numbers of testers to develop powerful insights on what business travelers in particular really wanted, and, over the course of the next few years, they developed a very successful new chain that specifically catered to the needs of these travelers: a no-nonsense layout for a night or two, suitable for the typical road warrior. Replication is the key in such an experiment; if done often enough, a significant result will eventually emerge. At the time this experiment was performed, there was no end to the line of business travelers willing to pocket an extra $20 for a half-hour survey on room preferences. Nowadays the price would be much higher.

A Case Study Showing Application of the Four-Step Design Process

To demonstrate the utility of DOE for improving any system and put our four-step process to the test, let's look at a sales process involving a call center for customer service. Our case study stems from Paul Selden's *Crash Course on DOE for Sales & Marketing* (The Paul Selden Companies, Inc., 2005$^{©}$). With Dr. Selden's permission, we made certain adjustments in the material to suit educational purposes.

The case study details a situation that many companies must confront: responding to service calls. This is a great opportunity to improve the company bottom line. When a potential customer calls in, the business wants to make an immediate sale. In a variation on this theme, the proper response is to alleviate or prevent something that has a negative impact: A caller wants to cancel an ongoing service, such as an Internet connection, and the goal becomes one of retention.

In the case provided by Selden, a manufacturer has mailed an offer to extend the warranty for the company's product at an attractive, but very profitable, price. Customers who respond to such mailings by calling generally intend to extend their warranties. However, not every call results in a sale. During the calls, many customers decide to decline the offer, expressing uncertainty about what repairs are covered, which product the notice was in reference to, the likelihood of repairs being needed, emergency services covered, how much is actually covered, and so forth.

The problem is compounded because each caller and each warranty is different. The customer service reps (CSRs) in this busy call center can be likened to air traffic controllers. They must quickly assess the nature of the inquiry and channel it through the path leading to the highest probability of a sale.

After hanging up, these CSRs must be immediately ready for a new call coming in from the automated system and must quickly adjust to the needs of the new caller and the specific warranty about which he or she is calling.

Clearly, this is a very challenging job. The steps taken by the sales engineers tasked with setting up an experiment aimed at improving performance of this call center are presented below:

1. Define the objective in terms of measurable responses. *At present, the CSRs average $200 in sales per hour. Management desires an increase of 10% in this rate, which would more than pay for the relatively inexpensive changes in the protocol.*
2. Calculate the "signal-to-noise ratio" (needed for calculating "power" later on):
   a. Determine the difference (symbolized by delta: Δ) that is important to detect for each response. *The Δ_y is $20/hour (10% of $200).*
   b. Estimate experimental error for *each* response in terms of its overall standard deviation (σ). *Historical records reveal a sigma (σ) of $10/hour.*
   c. Divide the signal by the noise for the ratio Δ_y/σ. *In this case it is 2 ($20/$10).*
3. Select the input factors to study and set their levels. The farther apart the better for generating a difference in response Δ_y. *The sales engineers, after a lot of brainstorming with everyone concerned, came up with the five factors and levels shown in Table 10.2. The low levels represent the current operations, which are managed by people coming up from the ranks of call center specialists, who rarely have a college degree. Perhaps a new graduate with a business degree can do better. This will be tested as the high level of this factor (C) of manager's experience. The other factors should be self-explanatory.*

Table 10.2 Factors for call-center experiment

Factor	*Name*	*Low Level (–)*	*High Level (+)*
A	Training	Same	More
B	Script	No	Yes
C	Manager's experience	CSR	College degree
D	Meetings	Weekly	Daily
E	Monetary incentives	No	Yes

4. Select a design. *Test the five factors in 16 runs done on 16 teams of 10 people each who are selected at random out of the pool of 1,000 CSRs at the call center.* Then:
   a. Evaluate aliases, if any (refer to Chapter 5). *This is a resolution V design, for which aliasing need not be a concern.*
   b. Determine statistical power—the probability that the experiment will reveal important effects, assuming they occur, which depends on the chosen factors and their levels (previous step). *The power for a signal-to-noise ratio of 2 for this 16-run design is 94.9% at a type I risk level of 0.05 (see Table 10.1).*
   c. Examine the design layout to ensure that all the factor combinations can be run at their specified levels safely and that they will produce meaningful results. Suggestion: Consider doing some pre-experimental range-finding. *In this case, it would be prudent for the sales engineers to try handling some calls with their new script. Other than that, there will probably not be much to worry about; nothing that may create a big upset in the call-center operations.*

PARTING ADVICE: ESSENTIAL ELEMENTS FOR EXPERIMENT DESIGN

Replication: Everyone knows that the more data one collects the more precisely something can be estimated; this is the power of averaging. To achieve an exact fit for handiwork at home, for example, you would want to measure twice and cut once. By the same token, two-level factorial designs, such as the one laid out in Table 10.2, build averaging into the test matrix, in this case, by balancing 8 runs at the low level versus the same number high. This dampens down overall process variation, including that stemming from measurements. However, if the measurement is the primary source of variation, you can simply repeat the test several times and enter the average measurement into your response column.

Randomization: Do not run the experiments according to the order laid out in your template, which may list a given factor at low levels first and then at high levels. This leaves you vulnerable to lurking variables that change over time, such as people getting tired, aging raw materials, or machine warm-up. Randomization provides insurance against the biasing of your effects that might otherwise occur.

As a further precaution, record any of these lurking variables that can be measured, such as the ambient temperature and humidity. Then, if you think they caused a significant effect relative to the experimental factors, consult with an expert about "regressing out the covariates." (If you get a blank look after saying this, find another expert.)

Blocking: If you cannot complete all experimental runs with everything else held constant, consider breaking them into two or more smaller blocks according to optional plans that can be found in DOE textbooks like this or laid out by software developed for this purpose. Typical variables to block out are day-to-day, machine-to-machine, or person-to-person differences, none of which can be held constant indefinitely. After all runs are completed, simple arithmetic techniques remove the effect of blocks from the analysis so it can more clearly estimate the effects of the experimental factors. Your statistical software will handle this.

Confirmation: If you are smart enough to pick some important factors that generate significant effects, your experiment will lead you to a new setup that promises better results. However, until this can be independently confirmed, it must be considered only a hypothesis.

> *"Theory guides, experiment decides."*
>
> **I. M. Kolthoff**

Unfortunately, things can change quickly and dramatically, and this sometimes makes it impossible to replicate and reproduce outcomes precisely. When this situation arises, all you can do then is rethink what is affecting your process. Even in this unhappy event, you will more than likely learn far more from a well-designed multifactor experiment than by the simplistic one-factor-at-a-time (OFAT) approach.

Appendix: Details on Power

In the case study illustrating our four-step process for DOE, everything worked out satisfactorily in the end, including the requirement for power being at least 80% within a reasonable number of experimental runs. In real life, things do not always go so well, particularly for power, when laying out a proposed design. What can you do if your probability of detecting the desired effect

(signal Δ) comes up short, for example, less than 50%? We have some suggestions for short-circuiting an alarming lack of power in the design of your experiment. They are listed in order of ease and sensibility with the simplest one first.

WHY IT IS SO VITAL TO SIZE YOUR DESIGN TO PROVIDE ADEQUATE POWER

What's "important" to any experimenter is detecting a signal if it's there to be found. It matters not if a result is "significant" but so small that it is of no importance—the ambivalent result illustrated in the upper-right quadrant of Figure 10.2. This can happen when robotic equipment, for example, automated analyzers with multiple 96 well plates, make it easy to do far more runs than needed for adequate power, thus generating significant results for miniscule effects. However, that's not nearly as bad as underpowered experiment designs that provide too few runs for seeing important effects. Then, you may unknowingly fall into the abyss of the lower-left quadrant where, if only you would have worked a bit harder, success could have been achieved. The really sad aspect of this predicament is you will not ever know what you missed.

P.S.: One last possibility is illustrated by the upper-left quadrant of Figure 10.2—an unimportant effect that is not statistically significant. By properly sizing your design to adequate power (80% or higher), you can say that you have "been there and done that," that is, having not discovered what you hoped to find, this combination of factors can be eliminated from future consideration. Often, success is achieved by this systematic process of elimination.

> *"A bigger effect produced in a study with a big margin of error is more impressive than a smaller effect that was measured more precisely."*
>
> **Stephen Ziliak**
>
> *"Making a Stat Less Significant," online at the Wall Street Journal website (online.wsj.com) by The Numbers Guy.*

Managing Expectations for What the Experiment Might Reveal

This is the simplest way to deal with a power problem: Expect less from your experiment. The initial tendency is to say any difference Δ (delta)

		Significant	
		No	Yes
Important	No	☺	😐
	Yes	☹	☺

Figure 10.2 The difference between significance and importance.

is important, which leads to unrealistic expectations and/or overly large designs (excessive number of runs). As a more practical alternative for purposes of calculating power, choose a delta large enough that you will take action if it is detected, not a value that is too small to provide economic payback. For example, if you start out thinking it would be nice to detect a 1-unit difference in a key response, but the probability of this happening (given the number of runs in your design) is only 50%, reset your sights to 2 units. Bigger effects are easier to detect and increase power.

Increase the Range of Your Factors

At the early stages of an experimental program, where you are only interested in screening the vital few factors from the trivial many, it pays to be bold in setting levels as wide apart as you dare. This greatly improves the chances of producing a signal that significantly exceeds the noise in the system (σ). Be careful, though. Like a test pilot flying a new aircraft, you cannot push your process beyond its limits. We urge you to do some pre-experimental range-finding first. For example, if you try reproducing the example for making microwave popcorn, detailed in Chapter 3 as an initial example of two-level factorial design, watch what happens as you increase time and be ready to press the off button at the first sight of smoke (this establishes the high end of your range for this factor). Wider ranges generate bigger effects, which are easier to detect and thus increase your statistical power. Note, however, that curvature may become more pronounced as the factor levels are pushed apart, so consider adding center points as discussed in Chapter 8.

Decrease the Noise (σ) in Your System

The first thing you must do after determining which factors will be varied in your experiment is to determine how to hold all other factors constant

or block them out. Next, if you think your measurements may contribute a major component of variance, consider investing in more precise test equipment. Another option is to repeat all measurements several times per run and enter the average result. This will add to the cost of your experiment, but may be far cheaper than replicating runs. For example, if you are baking cakes and rating the taste, it is a lot easier and a lot cheaper to line up 10 people to take bites from one cake than to bake 10 cakes the same way with the identical recipe (i.e., completely replicating the runs). Decreasing noise increases the ratio of Δ/σ, thus increasing power.

Accept Greater Risk of Type I Error (α)

One option for increasing power that will test your statistical mettle is to increase your critical alpha (α) value. This is typically set at 0.05, so you would increase it to 0.10, for example. The result will be a reduction in your risk of missing real effects (a type II error), thus, it increases power. On the other hand, it also increases the risk of declaring an effect is significant when it's not—a type I error. (For more detail on error types, refer back to the "Go Directly to Jail" boxed text at the beginning of Chapter 1.) We recommend this remedy for low power only if your experiment is aimed at screening process factors and you plan to follow up with further studies. At this phase of process improvement, you may be more concerned about missing a real effect (type II error) than selecting one that turns out to be false (type I error).

Select a Better and/or Bigger Design

If you plan to run a full-factorial design, then consider one of the previously listed suggestions for increasing your power. The only other choice is to replicate the design. If, on the other hand, you are contemplating a fraction, consider a larger one, such as a one-half rather than one-quarter. Larger fractions not only increase power, they reduce aliasing as well as filling more of the design space by providing additional combinations of factors. If you initially selected a design, such as the 12-run Plackett–Burman, then trade up to a standard fraction (such as one with 16 runs). One thing you should never do is replicate a fractional factorial to increase the power of your experiment. It is better to do a complete foldover, for example, than to rerun a resolution III design. This approach increases both power and resolution.

DO YOU FIND YOURSELF FORGETTING STATISTICAL DETAILS?

> *"A retentive memory may be a good thing, but the ability to forget is the true token of greatness."*
>
> **Elbert Hubbard**

Do not feel bad if at this stage of the book you cannot remember what you studied in the beginning chapters. This is only natural, although we may not entirely agree with Hubbard that it is a sign of greatness. The rate of forgetting can be predicted by the "Forgetting Curve" discovered by psychologist Hermann Ebbinghaus in the late nineteenth century. Without getting into the details (they are very forgettable), suffice it to say that memory decays exponentially and, thus, it can be characterized in half-life similar to radioactivity.

Ebbinghaus also found that by repeated review and practice, details can be remembered for a much longer period of time. To make the most of what you have learned in this book, go back and review everything in it at once. Then, as quickly as possible, get out and apply your new knowledge on an actual experiment.

Chapter 11

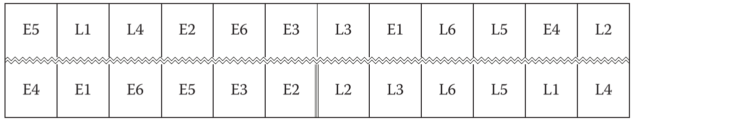
Split-Plot Designs to Accommodate Hard-to-Change Factors

> Oftentimes, complete randomization of all test parameters is extremely inefficient, or even totally impractical.
>
> **Alex Sewell**
> *53rd Test Management Group, 28th Test and Evaluation Squadron, Eglin Air Force Base*

Researchers often set up experiments with the best intentions of running them in random order, but they find that a given factor, such as temperature, cannot be easily changed, or, if so, only at too much expense in terms of time or materials. In this case, the best test plan might very well be the *split-plot* design. A split plot accommodates both hard-to-change (HTC) factors, e.g., the cavities in a molding process, and those that are easy to change (ETC); in this case, the pressure applied to the part being formed.

How Split Plots Naturally Emerged for Agricultural Field Tests

Split-plot designs originated in the field of agriculture where experimenters applied one treatment to a large area of land, called a *whole plot*, and other treatments to smaller areas of land within the whole plot, called *subplots*. For example, when D. R. Cox wrote about the “Split Unit Principle” in his classic *Planning of Experiments* (John Wiley & Sons, 1958), he detailed an experiment on six varieties of sugar beets (number 1 through 6) that are sown either early (E) or late (L). Figure 11.1 shows the alternatives of a completely randomized design versus one that is split into two subplots.



E5	L1	L4	E2	E6	E3	L3	E1	L6	L5	E4	L2
E4	E1	E6	E5	E3	E2	L2	L3	L6	L5	L1	L4

Figure 11.1 A completely randomized experiment (top row) versus one that is divided into split plots (bottom row).

MUST WE RANDOMIZE OUR EXPERIMENT?*

DOE guru George Box addressed this frequently asked question in his 1989 report: “Must We Randomize Our Experiment” (#47, Center for Quality and Productivity Improvement, University of Wisconsin/Madison), advising that experimenters:

- Always randomize in those cases where it creates little inconvenience.
- When an experiment becomes impossible, being subjected to randomization, and you can safely assume your process is stable, that is, any chance variations will be small compared to factor effects, then run it as you can in nonrandom order. However, if due to process variation, the results would be “useless and misleading” without randomization, abandon it and first work on stabilizing the process.
- Consider a split-plot design.

> *“Designing an experiment is like gambling with the devil: only a random strategy can defeat all his betting systems.”*
>
> **Sir Ronald A. Fisher**

* Excerpted from StatsMadeEasy blog 12/31/13 www.statsmadeeasy.net/2013/12/must-we-randomize-our-experiment/

Block 1		Block 2		…	Block 6	
V3 N4	V3 N3	V3 N3	V3 N4	…	V1 N4	V1 N1
V3 N2	V3 N1	V3 N1	V3 N2	…	V1 N3	V1 N2
V1 N1	V1 N2	V2 N1	V2 N3	…	V2 N4	V2 N3
V1 N4	V1 N3	V2 N4	V2 N2	…	V2 N1	V2 N2
V2 N1	V2 N2	V1 N2	V1 N3	…	V3 N1	V3 N2
V2 N3	V2 N4	V1 N4	V1 N1	…	V3 N3	V3 N4

Figure 11.2 A two-factor split plot replicated and blocked.

The split-plot layout made it far easier to sow the seeds because of the continuity in grouping. This is sweet for the sugar beet farmer (pun intended). However, as with anything that seems too easy, there is a catch to the more convenient test plan at the bottom of Figure 11.1: It confounds the time of sowing with the blocks of land, leaving no statistical power for assessing this factor (early versus late). The only way around this is to replicate the whole plot, that is, repeat this entire experiment on a number of fields, as is provided in another agricultural split plot illustrated by Figure 11.2. It lays out a classic field study on three varieties (V) of oats treated with four levels of nitrogen (N) fertilizer described by Sir Fisher's protégé (DOE pioneer Frank Yates. 1935. Complex experiments. *Journal of the Royal Statistical Society Supplement* 2: 181–247). The varieties were conveniently grouped into randomized whole plots within six blocks of land (we only show three blocks). It was then fairly easy to apply the different fertilizers chosen at random.

Applying a Split Plot to Save Time Making Paper Helicopters

Building paper helicopters as an in-class design of experiments (DOE) project is well established for a hands-on learning experience (see, for example, George's Column: Teaching Engineers Experimental Design with a Paper Helicopter, by George Box. 1992. *Quality Engineering* 4 (3): 453–459).

REPORT FROM THE FIELD

Paul Nelson (Ph.D., C.Stat.) in a private correspondence to the authors passed along this story of his exposure to split-plot experimentation from where these first sprung up:

The Classical (agricultural) Experiments conducted at Rothamsted Agricultural Research Station, Harpenden, United Kingdom (www.rothamsted.ac.uk), are the longest running agricultural experiments. Of the nine started between 1843 and 1856, only one has been abandoned. Unfortunately, we had to wait until the 1920s before the introduction of factorial treatment structures, so the designs are simple and follow the one-factor-at-a-time (OFAT) dictum. What is fascinating about these experiments is the clearly visible lines between strips, plots, and subplots of land due to the treatments (species, fertilizers, nitrogen sources, etc.) applied. The Rothamsted Classical Experiments are described in detail and beautifully illustrated in a document online at www.rothamsted.ac.uk/sites/default/files/LongTermExperiments.pdf.

The first split-plot experiment I came across was as an undergraduate student. My soon-to-be PhD supervisor, Rosemary Bailey (later professor of Statistics and head of the School of Mathematical Sciences at Queen Mary, University of London), who at the time was still working at Rothamsted, was the lecturer. The experiment was the comparison of three different varieties of rye grass in combination with four quantities of nitrogen fertilizer. The two fields of land (replicate blocks) were each divided into three strips (whole plots) and each strip divided into four subplots per strip. The rye grass varieties were sown on the strips by tractor, whilst the fertilizers could be sown by hand on the subplots. The response was percent dry-matter harvested from each plot.

Figure 11.3 pictures an example—the top flyer from the trials detailed in this case study.

Inspired by news of a supreme paper (Conqueror CX22) made into an airplane that broke the Guinness World Record™ for greatest distance flown (detailed by Sean Hutchinson in The Perfect Paper Airplane, *Mental Floss*, January 14, 2014), engineers at Stat-Ease (Mark and colleagues) designed a split-plot experiment on paper helicopters. They put CX22 to the test against a standard copy paper. As laid out below, five other factors were added to round out the flight testing via a half-fraction, high-resolution (Res VI) two-level design with 32 runs (= 2^{6-1}):

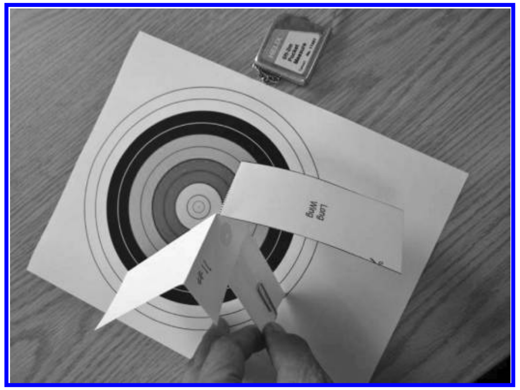


Figure 11.3 Paper helicopter.

a. Paper: 24# Navigator Premium (standard) versus 26.6# Conqueror CX22 (supreme)
b. Wing (technically a rotor in aviation terms) Length: Short versus Long
c. Body Length: Short versus Long
d. Body Width: Narrow versus Wide
E. Clip: Off versus On
F. Drop: Bottom versus Top

Being attributes of construction, the first four factors (identified with lower-case letters) are hard to change, i.e., HTC. The other factors come into play in the flights (operation). They are easy to change (ETC).

To develop the longest flying, most-accurate helicopter, the experimenters measured these two responses (Y):

1. Average time in seconds for three drops from ceiling height
2. Average deviation in centimeters from target

The averaging dampened out drop-to-drop variation, which, due to human factors in the release, air drafts, and so forth, can be considerable.

The experimenters enlisted a worker at Stat-Ease with the patience and skills needed for constructing the paper helicopters. Grouping the HTCs number of flying machines into whole plots, as shown in Table 11.1, reduced the manufacturing time by one half.

Table 11.1 Partial listing (first three groups and the last) of 32-run split-plot test plan

Group	*Run*	*a: Paper*	*b: Wing*	*c: Body Length*	*d: Body Width*	*E: Clip*	*F: Drop*
1	1	Nav Ultra	Long	Short	Narrow	Off	Top
1	2	Nav Ultra	Long	Short	Narrow	On	Bottom
2	3	CX22	Short	Short	Narrow	Off	Top
2	4	CX22	Short	Short	Narrow	On	Bottom
3	5	Nav Ultra	Long	Long	Narrow	Off	Bottom
3	6	Nav Ultra	Long	Long	Narrow	On	Top
…	…	…	…	…	…	…	…
16	31	CX22	Short	Long	Wide	Off	Top
16	32	CX22	Short	Long	Wide	On	Bottom

After the 16 helicopters were built (numbered by group with a marker), the engineers put each one (in run order) to the test with combinations of the two ETC factors as noted in the table.

Trade-Off of Power for Convenience When Restricting Randomization

As shown in Table 11.2, of the two responses, the distance from target (Y2) produced the lowest signal-to-noise ratio. It was based on a 5-cm minimal deviation of importance relative to a 2-cm standard deviation measured from prior repeatability studies.

Assuming a whole-plot to split-plot variance ratio of 2 (see boxed text below: Heads-Up about Statistical Analysis of Data from Split Plots, for background), a 32-run, two-level factorial design generates the power shown in Table 11.3 at 2.5 signal-to-noise for targeting.

Table 11.2 Signal-to-noise for the two paper helicopter responses

Response	*Signal*	*Noise*	*Signal/Noise*
Time avg	0.5 sec	0.15 sec	3.33
Target avg	5.0 cm	2.00 cm	2.50

Table 11.3 Power for main effects for design being done as split plot versus fully randomized

Design	*Hard (a–d)*	*Easy (E, F)*
Split plot	82.1%	99.9%
Randomized	97.5%	97.5%

What's really important to see here is that, by grouping the HTC factors, the experimenters lost power versus a completely randomized design. (It mattered little in this case, but it must be noted that the other factors (the ETCs) gained a bit more power, going from 97.5% to 99.8%.) However, the convenience and cost savings (the CX22 stock being extremely expensive) of only building half the paper helicopters—16 out of the 32 required in the fully randomized design of experiments—outweighed the loss in power, which in any case remained above the generally acceptable level of 80%. Thus, the split-plot test plan laid out in Table 11.1 got the thumbs up from the flight engineers.

Much more could be said about split plots and the results of this experiment in particular (for what it's worth, CX22 did indeed rule supreme). The main point here is how power was assessed to "right size" the test plan while accommodating the need to reduce the builds.

One More Split Plot Example: A Heavy-Duty Industrial One

George Box in a Quality Quandaries column on Split Plot Experiments (*Quality Engineering*, 8(3), 515–520, 1996) detailed a clever experiment that discovered a highly corrosion-resistant coating for steel bars. Four different coatings were tested (easy to do) at three different furnace temperatures (hard to change), each of which was run twice to provide power. See Box's design (a split plot) in Table 11.4 (results for corrosion resistance shown in parentheses). Note the bars being placed at random by position.

The HTC factor (temperature) creates so much noise in this process that in a randomized design it would overwhelm the effect of coating. The application of a split plot overcomes this variability by grouping the heats, in essence, filtering out the temperature differences. The effects graph in Figure 11.4 tells the story.

Table 11.4 Split plot to increase corrosion resistance of steel bars

Group	*Heats (Deg C) (Whole plots)*	*Positions (Subplots)*			
1	360	C2 (73)	C3 (83)	C1 (67)	C4 (89)
2	370	C1 (65)	C3 (87)	C4 (86)	C2 (91)
3	380	C3 (147)	C1 (155)	C2 (127)	C4 (212)
4	380	C4 (153)	C3 (90)	C2 (100)	C1 (108)
5	370	C4 (150)	C1 (140)	C3 (121)	C2 (142)
6	360	C1 (33)	C4 (54)	C2 (8)	C3 (46)

STRAYING FROM RANDOM ORDER FOR THE SAKE OF RUNNABILITY

Observe in this experiment design layout (Table 11.4) how Box made it even easier, in addition to grouping by heats, by increasing the furnace temperature run-by-run and then decreasing it gradually. This had to be done out of necessity due to the difficulties of heating and cooling a large mass of metal. The saving grace is that, although shortcuts like this undermine the resulting statistics when they do not account for the restrictions in randomization, the effect estimates remain true. Thus, the final results can still be assessed on the basis of subject matter knowledge as to whether they indicate important findings. Nevertheless, if at all possible, it will always be better to randomize levels in the whole plots and, furthermore, reset them when they have the same value, *e.g.,* between Groups 3 and 4 in this design.

> *"All industrial experiments are split plot experiments."*
>
> **Cuthbert Daniel**

The combination of high temperature with coating C4, located at the back corner, towers above all others. This finding, the result of the two-factor interaction aB between temperature and coating, achieved significance at the $p < 0.05$ threshold. The main effect of coating (B) also came out significant. If this experiment had been run completely

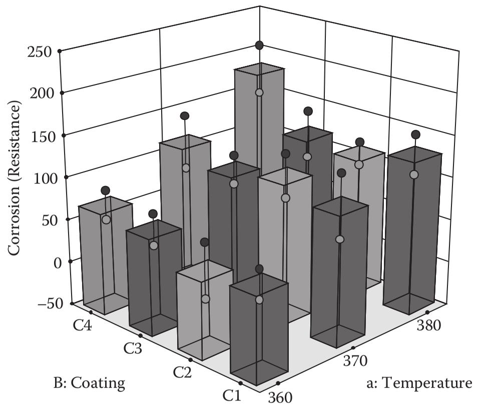


Figure 11.4 3-D bar chart of temperature (a) versus coating (B) effects.

randomized, p-values for the coating effect and the coating-temperature interaction would have come out to approximately 0.4 and 0.85 p-values, respectively, i.e., nowhere near to being statistically significant. Box concludes by suggesting the metallurgists try even higher heats with the C4 coating while simultaneously working at better controlling furnace temperature. Furthermore, he urges the experimenters to work at understanding better the physiochemical mechanisms causing corrosion of the steel. This really was the genius of George Box—his matchmaking of empirical modeling tools with subject matter expertise.

Our brief discussion of split plots ends on this high note.

HEADS-UP ABOUT STATISTICAL ANALYSIS OF DATA FROM SPLIT PLOTS

Split plots essentially combine two experiment designs into one. They produce both split-plot and whole-plot random errors. For example, as pointed out by Jones and Nachtsheim in their October 2009 primer on Split-Plot Designs: What, Why, and How (*Journal of Quality Technology* 41 (4)), the corrosion-resistance design introduces whole-plot error with each furnace reset due to variation by operators dialing in the temperature, inaccurate calibration, changes in ambient conditions, and so forth. Split-plot errors arise from bad measurements, variation in the distribution of heat within the furnace, differences in the thickness of the steel-bar coatings, and so forth.

This split error-structure creates complications in computing proper p-values for the effects, particularly when departing from a full-factorial balanced and replicated experiment, such as the corrosion-resistance case. If you really must go this route, be prepared for your DOE software applying specialized statistical tools that differ from standard analysis of variance (ANOVA). When that time comes, take full advantage of the help provided with the program and do some Internet searching on the terms. However, try not to get distracted by all the jargon-laden mumbo jumbo that may appear in the computer output; just search out the estimated p-values and thank goodness for statisticians and programmers who sort this all out.

> *"If the Lord Almighty had consulted me before embarking on Creation, I should have recommended something simpler."*
>
> **Alphonso the Wise (1221–1289)**

Chapter 12

Practice Experiments

Practice Experiment #1: Breaking Paper Clips

It's easy to find various brands, sizes, and types (e.g., virgin or recycled, rough-edged or smooth) of paper clips. These can be tested for strength in a simple comparative experiment, such as those illustrated in Chapter 2. (You also might find this to be a good way to relieve stress.) The procedure shown below specifies big versus little paper clips. It's intended for a group of people, with each individual doing the test. The person-to-person variation should be treated as blocks and removed from the analysis. The procedure is as follows:

1. Randomly select one big and one regular-sized paper clip. Flip a coin to randomly choose the first clip to break: heads = big, tails = standard size.
2. Gently pull each clip apart with the big loop on the right. Use the drawing in Figure 12.1 as a template. The angle affects performance, so be precise.
3. As pictured in Figure 12.1, move the smaller loop of the clip to the edge of your desk. The bigger loop should now project beyond the edge.
4. Hold the small loop down firmly with your left thumb. Grasp the big loop between right thumb and forefinger. Bend the big loop straight up and back. Continue bending the big loop back and forth until it breaks. Record the count for each clip. (Each back and forth movement counts as two bends.)

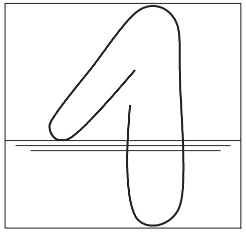


Figure 12.1 Paper clip positioned at edge of desk.

Tabulate the results from the entire group. Do a statistical analysis of the data in a manner similar to that outlined in Chapter 2, Problem 4 (wear tests on old versus new fabric). Is one paper clip significantly stronger than the other? (Suggestion: Use the software provided with the book. Set up a one-factor design similar to that shown in the tutorial that comes with the program. Be sure to identify this as a blocked design. Treat each person as a block. The ANOVA will then remove the person-to-person variance before doing the F-test on the effect of changing the type of paper clip. If the test is significant, make an effects plot.)

Practice Experiment #2: Hand–Eye Coordination

This is a simple test of hand–eye coordination that nicely illustrates an interaction of two factors:

1. Hand: Left (L) versus right (R)
2. Diameter of a target: Small versus large

Obviously, results will vary from one person to another depending on which hand is dominant. The procedure specifies four combinations in a two-by-two factorial (Figure 12.2).

We recommend you replicate the design at least twice (8 runs total) to add power for the statistical analysis. The procedure is as follows:

1. Draw two pairs of circles, centers equally distant, but with diameters that differ by a factor of two or more. (Use the template shown below in Figure 12.3.)

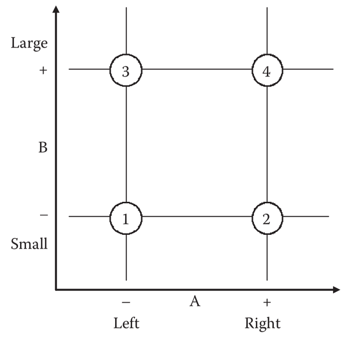


Figure 12.2 Design layout for hand–eye coordination test.

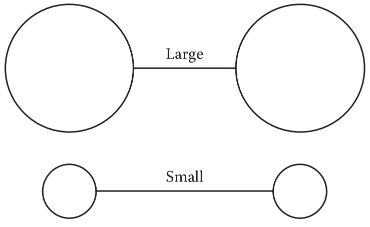


Figure 12.3 Template for hand–eye coordination test.

2. Randomly select either your left or right hand and one set of circles.
3. Alternating between circles in a set, mark as many dots as you can in 10 seconds. Your score will consist of the number of paired, in-target cycles that you complete in 10 seconds. A cycle means a dot in both circles, back and forth once. We recommend that you count the cycles as you are performing the experiment, because with so many dots in a confined space, it's very hard to tell them apart.
4. When the time is up, subtract one from your count for each dot outside either circle. Record your corrected count as the score.
5. Repeat the above steps to complete all four combinations at least twice each, for a total of at least 8 runs.

Analyze the results in the same manner as that outlined in Chapter 3. Do an analysis of variance (ANOVA) to see if anything is significant. Watch for an interaction of factors. (Suggestion: Use the software provided with the book. Set up a factorial design, similar to the one you did for the tutorial that comes with the program, for two factors in 4 runs, with two replicates. Be sure to enter the data properly so the inputs match up with the outputs. Then do the analysis as outlined in the tutorial.)

Other Fun Ideas for Practice Experiments

Ball in Funnel

This experiment is loosely based on W. Edwards Deming's funnel experiment. Time how long it takes a ball to spin through a funnel set at various heights. The ball can be fed through a tube. Vary the inclination and entry angle. Consider using different types of balls. Fasten the funnel so it is somewhat loose. You might then find that the effect of ball size depends on whether or not you hold the funnel—an interaction. Many more factors can be studied. (For details, see Bert Gunter's Through a Funnel Slowly with Ball Bearing and Insight to Teach Experimental Design. *The American Statistician* 47, Nov. 1993.)

Flight of the Balsa Buzzard

This is a fun DOE that anyone can do. Depending on how complex you want the design to be, purchase 10 to 20 balsa airplanes at your local hobby shop. Statistician Roger Longbotham, who contributed this idea to the authors, suggests testing five factors: vertical stabilizer frontward or backward, horizontal stabilizer frontward or backward, wing position to the front or back, pilot in or out, and nose weight as is or increased. If you do test these five factors, try a half-fraction of a two-level factorial. For each configuration, make two flights. Input the mean distance and range as separate responses. Caution: You may discover that certain factors cause increased variation in the flight path.

Paper Airplanes

This experiment makes school teachers cringe, but students won't need much help if you let them use their imagination to identify factors. Some ideas from grad students at North Carolina Tech include using multiple sheets, altering the design, changing the width and length, or increasing launch height and/or angle. Desired responses are length and accuracy.

(For details, see Sanjiv Sarin's Teaching Taguchi's Approach to Parameter Design, *Quality Progress*, May 1997.)

Impact Craters

Drop ball bearings (or marbles) of varying size into shallow containers filled with fine sand or granular sugar. Measure the diameter of the resulting crater. Try different drop heights and any other factors you come up with. Be prepared for some powerful interactions. If you have children do this experiment, put some little dinosaurs in the sand. Count how many become extinct. (For details, see Bert Gunter's *Linking High School Math and Science Through Statistical Design of Experiments*, Macomb Intermediate School District, Clinton Township, Michigan, 1995, p. 2.1. Also search on MISD "Design of Experiments" crater on the Internet for details on this experiment, including video.)

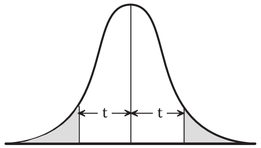
ASTRONOMICAL BOWLERS DENIED THEIR LOFTY DREAMS

To simulate the impact of meteorites, members of the Salt Lake Astronomical Society wanted to drop bowling balls from very high altitudes onto the salt flats of Utah. However, workers in the target area from the U.S. Bureau of Land Management objected to the experiment.

(From Chuck Shepherd, *News of the Weird,* March 6, 2003. Online at www.newsoftheweird.com/)

Appendix 1

A1.1 Two-Tailed *t*-Table



Probability points of the t-distribution with degrees of freedom (df)

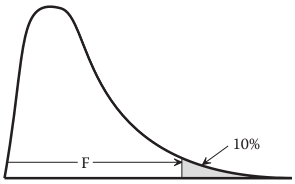
df	*Two-Tail Area Probability*					
	0.2	*0.1*	*0.05*	*0.01*	*0.005*	*0.001*
1	3.078	6.314	12.706	63.657	127.321	636.619
2	1.886	2.920	4.303	9.925	14.089	31.599
3	1.638	2.353	3.182	5.841	7.453	12.924
4	1.533	2.132	2.776	4.604	5.598	8.610
5	1.476	2.015	2.571	4.032	4.773	6.869
6	1.440	1.943	2.447	3.707	4.317	5.959
7	1.415	1.895	2.365	3.499	4.029	5.408
8	1.397	1.860	2.306	3.355	3.833	5.041
9	1.383	1.833	2.262	3.250	3.690	4.781
10	1.372	1.812	2.228	3.169	3.581	4.587
11	1.363	1.796	2.201	3.106	3.497	4.437

Continued

Probability points of the t-distribution with degrees of freedom (df) (*Continued*)

df	*Two-Tail Area Probability*					
	0.2	*0.1*	*0.05*	*0.01*	*0.005*	*0.001*
12	1.356	1.782	2.179	3.055	3.428	4.318
13	1.350	1.771	2.160	3.012	3.372	4.221
14	1.345	1.761	2.145	2.977	3.326	4.140
15	1.341	1.753	2.131	2.947	3.286	4.073
16	1.337	1.746	2.120	2.921	3.252	4.015
17	1.333	1.740	2.110	2.898	3.222	3.965
18	1.330	1.734	2.101	2.878	3.197	3.922
19	1.328	1.729	2.093	2.861	3.174	3.883
20	1.325	1.725	2.086	2.845	3.153	3.850
21	1.323	1.721	2.080	2.831	3.135	3.819
22	1.321	1.717	2.074	2.819	3.119	3.792
23	1.319	1.714	2.069	2.807	3.104	3.768
24	1.318	1.711	2.064	2.797	3.091	3.745
25	1.316	1.708	2.060	2.787	3.078	3.725
26	1.315	1.706	2.056	2.779	3.067	3.707
27	1.314	1.703	2.052	2.771	3.057	3.690
28	1.313	1.701	2.048	2.763	3.047	3.674
29	1.311	1.699	2.045	2.756	3.038	3.659
30	1.310	1.697	2.042	2.750	3.030	3.646
40	1.303	1.684	2.021	2.704	2.971	3.551
60	1.296	1.671	2.000	2.660	2.915	3.460
120	1.289	1.658	1.980	2.617	2.860	3.373
1000	1.282	1.646	1.962	2.581	2.813	3.300
10000	1.282	1.645	1.960	2.576	2.808	3.291

A1.2 *F*-Table for 10%



Percentage points of the F-distribution: Upper 10% points

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
1	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.858	60.195	61.220	61.740
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.425	9.441
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.200	5.184
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.870	3.844
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.238	3.207
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.871	2.836
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.632	2.595
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.464	2.425
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.340	2.298
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.244	2.201
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.167	2.123

Continued

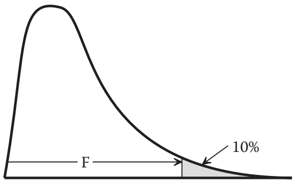
Percentage points of the F-distribution: Upper 10% points (*Continued*)

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.105	2.060
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.053	2.007
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.010	1.962
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059	1.972	1.924
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	1.940	1.891
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001	1.912	1.862
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977	1.887	1.837
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956	1.865	1.814
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937	1.845	1.794
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920	1.827	1.776
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904	1.811	1.759

23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890	1.796	1.744
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877	1.783	1.730
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866	1.771	1.718
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855	1.760	1.706
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845	1.749	1.695
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836	1.740	1.685
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827	1.731	1.676
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819	1.722	1.667
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763	1.662	1.605
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.603	1.543
120	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652	1.545	1.482
100K	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599	1.487	1.421

K: Multiply this value by 1,000.

A1.3 *F*-Table for 5%



Percentage points of the F-distribution: Upper 5% points

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.429	19.446
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.703	8.660
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.858	5.803
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.619	4.558
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	3.938	3.874
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.511	3.445
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.218	3.150
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.006	2.936
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.845	2.774
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.719	2.646

12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.617	2.544
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.533	2.459
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.463	2.388
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.403	2.328
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.352	2.276
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.308	2.230
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.269	2.191
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.234	2.155
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.203	2.124
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.176	2.096
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.151	2.071
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.128	2.048
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.108	2.027
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.089	2.007
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.072	1.990
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.056	1.974
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.041	1.959
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.027	1.945
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.015	1.932

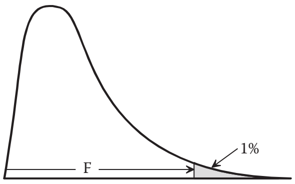
Continued

Percentage points of the F-distribution: Upper 5% points (*Continued*)

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	1.924	1.839
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.836	1.748
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910	1.750	1.659
100K	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.939	1.880	1.831	1.666	1.571

K: Multiply this value by 1,000.

A1.4 *F*-Table for 1%



Percentage points of the F-distribution: Upper 1% points

df_{num} df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
1	4052.2	4999.5	5403.3	5624.6	5763.6	5859.0	5928.3	5981.1	6022.5	6055.8	6157.3	6208.7
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.433	99.449
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	26.872	26.690
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.198	14.020
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.722	9.553
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.559	7.396
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.314	6.155
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.515	5.359
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	4.962	4.808
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.558	4.405
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.251	4.099

Continued

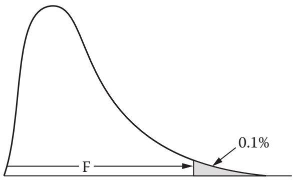
Percentage points of the F-distribution: Upper 1% points (*Continued*)

df_{num} df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.010	3.858
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.815	3.665
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.656	3.505
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.522	3.372
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.409	3.259
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.312	3.162
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.227	3.077
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.153	3.003
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.088	2.938
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.030	2.880
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	2.978	2.827

23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	2.931	2.781
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	2.889	2.738
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.850	2.699
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.815	2.664
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.783	2.632
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.753	2.602
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.726	2.574
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.700	2.549
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.522	2.369
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.352	2.198
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472	2.192	2.035
100K	6.635	4.605	3.782	3.319	3.017	2.802	2.640	2.511	2.408	2.321	2.039	1.878

K: Multiply this value by 1,000.

A1.5 *F*-Table for 0.1%



Percentage points of the F-distribution: Upper 0.1% points

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
1	405.2K	500.0K	540.4K	562.5K3	576.4K	585.9K	592.9K	598.1K	602.3K92	605.6K	615.8K	620.9K
2	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.43	999.45
3	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	127.37	126.42
4	74.137	61.246	56.177	53.436	51.712	50.525	49.658	48.996	48.475	48.053	46.761	46.100
5	47.181	37.122	33.202	31.085	29.752	28.834	28.163	27.649	27.244	26.917	25.911	25.395
6	35.507	27.000	23.703	21.924	20.803	20.030	19.463	19.030	18.688	18.411	17.559	17.120
7	29.245	21.689	18.772	17.198	16.206	15.521	15.019	14.634	14.330	14.083	13.324	12.932
8	25.415	18.494	15.829	14.392	13.485	12.858	12.398	12.046	11.767	11.540	10.841	10.480
9	22.857	16.387	13.902	12.560	11.714	11.128	10.698	10.368	10.107	9.894	9.238	8.898
10	21.040	14.905	12.553	11.283	10.481	9.926	9.517	9.204	8.956	8.754	8.129	7.804
11	19.687	13.812	11.561	10.346	9.578	9.047	8.655	8.355	8.116	7.922	7.321	7.008
12	18.643	12.974	10.804	9.633	8.892	8.379	8.001	7.710	7.480	7.292	6.709	6.405

13	17.815	12.313	10.209	9.073	8.354	7.856	7.489	7.206	6.982	6.799	6.231	5.934
14	17.143	11.779	9.729	8.622	7.922	7.436	7.077	6.802	6.583	6.404	5.848	5.557
15	16.587	11.339	9.335	8.253	7.567	7.092	6.741	6.471	6.256	6.081	5.535	5.248
16	16.120	10.971	9.006	7.944	7.272	6.805	6.460	6.195	5.984	5.812	5.274	4.992
17	15.722	10.658	8.727	7.683	7.022	6.562	6.223	5.962	5.754	5.584	5.054	4.775
18	15.379	10.390	8.487	7.459	6.808	6.355	6.021	5.763	5.558	5.390	4.866	4.590
19	15.081	10.157	8.280	7.265	6.622	6.175	5.845	5.590	5.388	5.222	4.704	4.430
20	14.819	9.953	8.098	7.096	6.461	6.019	5.692	5.440	5.239	5.075	4.562	4.290
21	14.587	9.772	7.938	6.947	6.318	5.881	5.557	5.308	5.109	4.946	4.437	4.167
22	14.380	9.612	7.796	6.814	6.191	5.758	5.438	5.190	4.993	4.832	4.326	4.058
23	14.195	9.469	7.669	6.696	6.078	5.649	5.331	5.085	4.890	4.730	4.227	3.961
24	14.028	9.339	7.554	6.589	5.977	5.550	5.235	4.991	4.797	4.638	4.139	3.873
25	13.877	9.223	7.451	6.493	5.885	5.462	5.148	4.906	4.713	4.555	4.059	3.794
26	13.739	9.116	7.357	6.406	5.802	5.381	5.070	4.829	4.637	4.480	3.986	3.723
27	13.613	9.019	7.272	6.326	5.726	5.308	4.998	4.759	4.568	4.412	3.920	3.658
28	13.498	8.931	7.193	6.253	5.656	5.241	4.933	4.695	4.505	4.349	3.859	3.598
29	13.391	8.849	7.121	6.186	5.593	5.179	4.873	4.636	4.447	4.292	3.804	3.543
30	13.293	8.773	7.054	6.125	5.534	5.122	4.817	4.581	4.393	4.239	3.753	3.493
40	12.609	8.251	6.595	5.698	5.128	4.731	4.436	4.207	4.024	3.874	3.400	3.145

Continued

Percentage points of the F-distribution: Upper 0.1% points (*Continued*)

df_{num} / df_{den}	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*15*	*20*
60	11.973	7.768	6.171	5.307	4.757	4.372	4.086	3.865	3.687	3.541	3.078	2.827
120	11.380	7.321	5.781	4.947	4.416	4.044	3.767	3.552	3.379	3.237	2.783	2.534
100K	10.828	6.908	5.422	4.617	4.103	3.743	3.475	3.266	3.098	2.959	2.513	2.266

K: Multiply entries by 1,000 in first row of F-values, and last value for df.

Appendix 2

A2.1 Four-Factor Screening and Characterization Designs

Screening Main Effects in 8 Runs

This is an 8-run standard fraction (1/2 replicate) for four factors. The design allows you to estimate all main effects aliased only by three-factor or higher-order interactions.

Screening Design Layout

Std	*A*	*B*	*C*	*D*
1	–	–	–	–
2	+	–	–	+
3	–	+	–	+
4	+	+	–	–
5	–	–	+	+
6	+	–	+	–
7	–	+	+	–
8	+	+	+	+

Alias Structure

[Intercept] = Intercept
[A] = A

[B] = B
[C] = C
[D] = D
[AB] = AB + CD
[AC] = AC + BD
[AD] = AD + BC

Characterizing Interactions with 12 Runs

This is a 12-run irregular* fraction (3/4 replicate) for four factors. The design allows you to estimate all main effects and two-factor interactions (2fi) aliased only by three-factor or higher-order interactions. However, if effects are calculated hierarchically starting with main effects, these will be partially aliased with one or more interactions. In this case, be sure to review the probabilities in the ANOVA for the 2fi model. Exclude any main effects that are not significant.

Characterization Design Layout

Std	*A*	*B*	*C*	*D*
1	–	–	–	–
2	+	+	–	–
3	–	–	+	–
4	+	–	+	–
5	–	+	+	–
6	+	+	+	–
7	–	–	–	+
8	+	–	–	+
9	–	+	–	+
10	+	+	–	+
11	+	–	+	+
12	–	+	+	+

* The number of runs is not a power of 2 (4, 8, or 16) as in a standard two-level factorial for four factors.

Alias Structure for Factorial Two-Factor Interaction Model

[Intercept] = Intercept – 0.5 * ABC – 0.5 * ABD
[A] = A – ACD
[B] = B – BCD
[C] = C – ABCD
[D] = D – ABCD
[AB] = AB – ABCD
[AC] = AC – BCD
[AD] = AD – BCD
[BC] = BC – ACD
[BD] = BD – ACD
[CD] = CD – ABD
[ABC] = ABC – ABD

Alias Structure for Factorial Main Effect Model

[Intercept] = Intercept – 0.333 * CD – 0.333 * ABC – 0.333 * ABD
[A] = A – 0.333 * BC – 0.333 * BD – 0.333 * ACD
[B] = B – 0.333 * AC – 0.333 * AD – 0.333 * BCD
[C] = C – 0.5 * AB
[D] = D – 0.5 * AB

A2.2 Five-Factor Screening and Characterization Designs

Screening Main Effects in 12 Runs

This is a minimum-run resolution IV design with 2 runs added in reserve to cover for any that cannot be performed or turn out to be statistically discrepant. The design allows you to estimate all main effects aliased only by three-factor or higher-order interactions.

MINIMUM-RUN DESIGNS

Minimum-run, two-level factorial designs for screening and characterization were invented in 2002 by Patrick Whitcomb and Gary Oehlert (“Small, Efficient, Equireplicated Resolution V Fractions of 2^K Designs and Their Application to Central Composite Designs.” Paper presented at the Proceedings of 46th Fall Technical Conference of the American Society of Quality, Chemical and Process Industries Division and Statistic Division, 2002; American Statistical Association, Section on Physical and Engineering Sciences).

Minimum-run resolution IV (“MR4”) screening designs estimate all main effects, clear of two-factor or higher interactions, in a minimum of experimental runs equal to two times the number of factors. Minimum-run resolution V (“MR5”) factorial designs estimate all main effects and two-factor interactions.

Screening Design Layout

Std	*A*	*B*	*C*	*D*
1	+	+	+	–
2	–	+	–	+
3	+	–	–	–
4	+	–	+	+
5	+	+	–	+
6	–	–	+	–
7	–	–	–	+
8	+	–	+	–
9	–	+	+	+
10	–	+	–	–
11	–	–	+	–
12	+	+	–	+

Alias Structure

All main effects are aliased only with four-factor or higher order interactions, which aren’t worth noting. Be very wary of any interactions that look significant. Due to aliasing issues, these must be verified via a follow-up characterization design.

[Intercept] = Intercept + BD – CD + DE
[A] = A
[B] = B
[C] = C
[D] = D
[E] = E
[AB] = AB – BD – CE
[AC] = AC – BE – CD
[AD] = AD + BD – CD + DE
[AE] = AE + 2 * BD – BE – 2 * CD + CE + DE
[BC] = BC + 2 * BD – BE – 2 * CD + CE + 2 * DE

Characterizing Interactions with 16 Runs

This is a 16-run standard fraction (1/2 replicate) for five factors. The design allows you to estimate all main effects and two-factor interactions aliased only by three-factor or higher-order interactions.

Design Layout

Std	*A*	*B*	*C*	*D*	*E*
1	–	–	–	–	+
2	+	–	–	–	–
3	–	+	–	–	–
4	+	+	–	–	+
5	–	–	+	–	–
6	+	–	+	–	+
7	–	+	+	–	+
8	+	+	+	–	–
9	–	–	–	+	–
10	+	–	–	+	+
11	–	+	–	+	+

Continued

Std	*A*	*B*	*C*	*D*	*E*
12	+	+	–	+	–
13	–	–	+	+	+
14	+	–	+	+	–
15	–	+	+	+	–
16	+	+	+	+	+

Alias Structure for Factorial Two-Factor Interaction (2FI) Model

[Intercept] = Intercept + ABCDE
[A] = A + ABCD
[B] = B + BCDE
[C] = C + ABDE
[D] = D + ABCE
[E] = E + ABCD
[AB] = AB + CDE
[AC] = AC + BDE
[AD] = AD + BCE
[AE] = AE + BCD
[BC] = BC + ADE
[BD] = BD + ACE
[BE] = BE + ACD
[CD] = CD + ABE
[CE] = CE + ABD
[DE] = DE + ABC

A2.3 Six-Factor Screening and Characterization Designs

Screening Main Effects in 14 Runs

This is a minimum-run resolution IV design (see boxed text in Section A2.2 above for details on these MR4 templates) with 2 runs added in reserve to cover for any that cannot be performed or turn out to be statistically discrepant.

Screening Design Layout

Std	*A*	*B*	*C*	*D*	*E*	*F*
1	–	+	–	+	+	–
2	–	–	–	+	–	–
3	–	+	+	+	–	+
4	–	–	+	–	+	–
5	–	–	–	–	–	+
6	+	–	+	+	+	+
7	–	+	–	–	+	+
8	+	–	+	–	–	+
9	+	+	+	–	+	+
10	+	–	–	–	+	–
11	+	+	–	+	–	+
12	+	+	+	+	+	–
13	–	+	–	–	–	–
14	+	–	+	+	–	–

Alias Structure

All main effects are aliased only with four-factor or higher order interactions, which are not worth noting. Be very wary of any interactions that look significant. Due to aliasing issues, these must be verified via a follow-up characterization design.

[Intercept] = Intercept + BF – DE
[A] = A
[B] = B
[C] = C
[D] = D
[E] = E
[F] = F

[AB] = AB + 0.5 * BC + 0.5 * BE + 0.5 * BF – 0.5 * CD + 0.5 * CE – 0.5 * CF – 0.5 * DE – 0.5 * DF
[AC] = AC + BE – BF + DE – DF
[AD] = AD – 0.5 * BC – 0.5 * BE – 0.5 * BF + 0.5 * CD + 0.5 * CE – 0.5 * CF + 0.5 * DE + 0.5 * DF
[AE] = AE + 0.5 * BC – 0.5 * BE – 0.5 * BF + 0.5 * CD + 0.5 * CE + 0.5 * CF + 0.5 * DE + 0.5 * DF
[AF] = AF – 0.5 * BC – 0.5 * BE – 0.5 * BF – 0.5 * CD + 0.5 * CE + 0.5 * CF + 0.5 * DE + 0.5 * DF
[BD] = BD – BF + DE – EF

Characterizing Interactions with 22 Runs

This is a minimum-run resolution V design (see boxed text in Section 2.2 above for details on these MR5 templates). The design allows you to estimate all main effects and two-factor interactions aliased only by three-factor or higher-order interactions.

Design Layout

Std	*A*	*B*	*C*	*D*	*E*	*F*
1	–	–	+	+	–	+
2	+	+	–	–	–	–
3	–	–	–	–	+	–
4	+	+	+	+	–	–
5	–	+	–	+	–	+
6	+	+	+	–	–	+
7	+	–	–	+	–	+
8	–	+	–	+	+	–
9	+	–	+	–	–	–
10	+	–	+	+	+	+
11	+	–	–	+	+	–
12	+	+	–	+	+	+

Continued

Std	*A*	*B*	*C*	*D*	*E*	*F*
13	–	+	–	–	+	+
14	+	+	+	–	–	–
15	–	–	+	+	+	–
16	+	–	–	–	+	+
17	–	+	+	–	–	–
18	–	–	–	–	–	+
19	–	–	–	+	–	–
20	–	+	+	+	+	+
21	+	+	+	–	+	–
22	–	–	+	–	+	+

Alias Structure for Factorial Two-Factor Interaction (2FI) Model

Each main effect and two-factor interaction is partially aliased only with three-factor interactions. Only the first main effect ([A]) and two-factor interaction ([AB]) on the entire list is shown below. The other aliasing is very similar, so this provides an idea of what to expect.

[Intercept] = Intercept
[A] = A + 0.333 * BCD + 0.333 * BCE + 0.333 * BCF – 0.333 * BDE – 0.333 * BDF – 0.333 * BEF – 0.333 * CDE – 0.333 * CDF – 0.333 * CEF + 0.333 * DEF …
[AB] = AB – 0.333 * ACD – 0.333 * ACE – 0.333 * ACF + 0.333 * ADE + 0.333 * ADF + 0.333 * AEF – 0.333 * BCD – 0.333 * BCE – 0.333 * BCF + 0.333 * BDE + 0.333 * BDF + 0.333 * BEF – 0.333 * CDE – 0.333 * CDF – 0.333 * CEF + 0.333 * DEF …

A2.4 Seven-Factor Screening and Characterization Designs

Screening Main Effects in 16 Runs

This is a 16-run standard fraction (1/8 replicate) for seven factors. Main effects are clear of two-factor interactions. Two-factor interactions are completely aliased with each other.

Screening Design Layout

Std	*A*	*B*	*C*	*D*	*E*	*F*	*G*
1	–	–	–	–	–	–	–
2	+	–	–	–	+	–	+
3	–	+	–	–	+	+	–
4	+	+	–	–	–	+	+
5	–	–	+	–	+	+	+
6	+	–	+	–	–	+	–
7	–	+	+	–	–	–	+
8	+	+	+	–	+	–	–
9	–	–	–	+	–	+	+
10	+	–	–	+	+	+	–
11	–	+	–	+	+	–	+
12	+	+	–	+	–	–	–
13	–	–	+	+	+	–	–
14	+	–	+	+	–	–	+
15	–	+	+	+	–	+	–
16	+	+	+	+	+	+	+

Alias Structure

[Intercept] = Intercept
[A] = A + BCE + BFG + CDG + DEF
[B] = B + ACE + AFG + CDF + DEG
[C] = C + ABE + ADG + BDF + EFG
[D] = D + ACG + AEF + BCF + BEG
[E] = E + ABC + ADF + BDG + CFG
[F] = F + ABG + ADE + BCD + CEG
[G] = G + ABF + ACD + BDE + CEF
[AB] = AB + CE + FG
[AC] = AC + BE + DG
[AD] = AD + CG + EF
[AE] = AE + BC + DF
[AF] = AF + BG + DE

[AG] = AG + BF + CD
[BD] = BD + CF + EG

Characterizing Interactions with 30 Runs

This is a minimum-run resolution V design (see boxed text in section A2.2 above for details on these MR5 templates). The design allows you to estimate all main effects and two-factor interactions aliased only by three-factor or higher-order interactions.

Design Layout

Std	*A*	*B*	*C*	*D*	*E*	*F*	*G*
1	–	–	–	+	–	–	+
2	–	+	–	–	+	–	–
3	–	+	+	+	–	–	+
4	–	–	+	–	–	–	+
5	+	+	+	+	+	+	+
6	–	–	–	–	+	+	–
7	+	+	–	–	–	+	–
8	–	+	–	+	+	+	–
9	+	–	+	–	+	–	–
10	+	+	–	+	–	+	+
11	–	–	–	+	–	+	–
12	–	–	–	+	+	+	+
13	–	–	+	+	+	–	–
14	+	–	–	+	+	–	–
15	+	+	–	+	–	–	–
16	–	+	+	–	–	–	–
17	+	+	–	–	+	+	+
18	+	+	+	–	–	+	+
19	–	+	+	–	+	+	–
20	+	+	+	+	+	–	–
21	+	–	–	–	+	–	+
22	+	–	–	–	–	+	+
23	+	–	+	+	+	+	–
24	–	–	+	–	–	+	–
25	+	–	+	+	+	–	–
26	–	+	+	–	–	–	+
27	+	+	–	–	–	–	+
28	–	–	+	+	–	+	+
29	+	–	+	+	+	–	+
30	–	+	–	–	–	+	+

Continued

Alias Structure for Factorial Two-Factor Interaction (2FI) Model

Each main effect and two-factor interaction is partially aliased only with three-factor interactions. Only the first main effect ([A]) and two-factor interaction ([AB]) on the entire list is shown below. The other aliasing is very similar so this provides an idea of what to expect.

[Intercept] = Intercept – 0.0667 * ABC + 0.2 * ABD – 0.333 * ABE + 0.333
* ABF + 0.0667 * ABG + 0.333 * ACD + 0.333 * ACE + 0.2
* ACF – 0.0667 * ACG + 0.0667 * ADE – 0.0667 * ADF – 0.333
* ADG – 0.0667 * AEF + 0.2 * AEG + 0.333 * AFG + 0.133
* BCE + 0.667 * BCG + 0.133 * BEF + 0.133 * BFG + 0.133
* CDE + 0.133 * CDG + 0.667 * DEF + 0.133 * DFG
[A] = A – 0.467 * ABC – 0.6 * ABD – 0.333 * ABE + 0.333 * ABF + 0.467
* ABG + 0.333 * ACD + 0.333 * ACE – 0.6 * ACF – 0.467 * ACG
+ 0.467 * ADE – 0.467 * ADF – 0.333 * ADG – 0.467 * AEF – 0.6
* AEG + 0.333 * AFG – 0.0667 * BCE – 0.333 * BCG – 0.0667 * BEF
– 0.0667 * BFG – 0.0667 * CDE – 0.0667 * CDG – 0.333 * DEF
– 0.0667 * DFG …
[AB] = AB + 0.25 * BCD + 0.417 * BCE + 0.0833 * BCF – 0.417 * BCG + 0.25
* BDE– 0.25 * BDF – 0.25 * BDG – 0.417 * BEF + 0.0833 * BEG
+ 0.417 * BFG + 0.583 * CDE – 0.0833 * CDF – 0.583 * CDG – 0.417
* CEF – 0.417 * CEG + 0.417 * CFG – 0.583 * DEF – 0.0833 * DEG
+ 0.583 * DFG + 0.417 * EFG …

Glossary

Statistical Symbols

df	degrees of freedom
k	number of factors in design
i	individual datum
n	number of observations in sample
p	fraction of design (example 2^{k-p}) or probability value (ex. Prob > F)
PI	Prediction Interval
r	sample correlation coefficient
R^2	index of determination
s	sample standard deviation
s^2	sample variance
t	t-value
X	independent variable
Y	observed response value
Z	uncontrolled variable
*	multiplication symbol
—	(bar) average (e.g., $\bar{Y}$)
^	(hat) predicted (e.g., $\hat{Y}$)
α	(alpha) Type I error rate
β	(beta) coefficient or Type II error rate
Δ	(delta) difference (e.g., ΔY)
σ	(sigma) population (true) standard deviation
Σ	(capital sigma) mathematical operator to take the sum of a number series
µ	(mu) population (true) mean

Terms

Actual value: The observed value of the response from the experiment. The physical levels of the factors in their units of measure (as opposed to their coded levels, such as –1 or +1).

Adjusted R-squared: R-squared adjusted for the number of terms in the model relative to the number of points in the design. An estimate of the fraction of overall variation in the data accounted for by the model.

Alias: Other term(s) that is (are) correlated with a given coefficient. The resulting predictive model is then said to be *aliased*. (Also called *confounding*.)

Analysis of variance (ANOVA): A statistical method, based on the F-test, that assesses the significance of experimental results. It involves subdividing the total variation of a set of data into component parts.

Antagonism: An undesirable interaction of two factors where the combination produces a response that is not as good as what would be expected from either one alone. The same concept can be applied to higher-order interactions.

Average: *See* **Mean**.

Axial points: Design points that fall on the spatial coordinate axes emanating from the overall center point (or centroid in mixture space), often used as a label for star points in a central composite design.

Balanced design: Designs in which low and high levels of any factor or interaction occur in equal numbers.

Bias: A systematic error in estimation of a population parameter.

Block: A group of trials based on a common factor. Blocking is advantageous when there is a known factor that may influence the experimental result, but the effect is not of interest. For example, if all experiments cannot be conducted in one day or within one batch of raw material, the experimental points can be divided in such a way that the blocked effect is eliminated before computation of the model. Removal of the block effect reduces the noise in the experiment and improves the sensitivity to effects.

Case statistics: Diagnostic statistics calculated for each case, that is, each design point in the design after the model has been selected.

Categoric variable: Factors whose levels fall into discrete classes, such as metal versus plastic material. (Also called *class* or *qualitative* variable.)

Cell: The blank field to be filled with a response resulting from a given set of input factor levels.

Center point: An experiment with all numerical factor levels set at their midpoint value.

Central composite design (CCD): A design for response surface methods (RSM) that is composed of a core two-level factorial plus axial points and center points.

Central limit theorem: In its simplest form, this theorem states that the distribution of averages approximates normal as the sample size (n) increases. Furthermore, the variance of the averages is reduced by a factor of n when compared with the variance of individual data.

Centroid: The center point of mixture space within the specified constraints.

Class variable: *See* **Categoric variable**.

Coded factor level: *See* **Coding**.

Coding: A way to center and normalize factors, e.g., by converting low and high factor levels to –1 and +1, respectively.

Coefficient: *See* **Model coefficient**

Coefficient of variation (C. V.): Also known as the *relative standard deviation*, the coefficient of variation is a measure of residual variation of the data relative to the size of the mean. It is the standard deviation (root mean square error from ANOVA) divided by the dependent mean, expressed as a percent.

Component: An ingredient of a mixture.

Confidence interval (CI): A data based interval constructed by a method that covers the true population value a stated percentage (typically 95%) of the time in repeated samplings.

Confounding: *See* **Alias**.

Constraint: Limit in respect to component ranges for a mixture experiment.

Continuous variable: *See* **Numeric variable**.

Contour plot: A topographical map drawn from a mathematical model, usually in conjunction with response surface methods (RSM) for experimental design. Each contour represents a continuous response fixed at some value.

Cook's distance: A measure of how much the regression would change if that particular run were omitted from the analysis. Relatively large values are associated with cases with high leverage and large externally studentized residuals. Cases with large values relative to other cases may cause undue influence on the fitting and should be investigated. They could be caused by recording errors, an incorrect model, or a design point far from the remaining cases.

Corrected total: The total sum of squares (SS) corrected for the mean (calculated by taking the sum of the squared distances of each individual response value from its overall average).

Count data: Data based on discrete occurrences rather than from a continuous scale.

Crash and burn: Exceed the operating boundaries (envelope) of a process.

Cumulative probability: The proportion of individuals in a population that the fall below a specified value.

Curvature: A measure of the offset at the center point of actual versus predicted values from a factorial model. If significant, consider going to a quadratic model, which can be fitted to data from a response surface design.

Degree of equation: The highest order of terms in a model. For example, in an equation of degree two, you will find terms with two factors multiplied together as well as squared terms.

Degrees of freedom (df): The number of independent pieces of information available to estimate a parameter.

Dependent mean: The mean of the response over all the design points.

Design matrix: An array of values presented in rows and columns. The columns usually represent design factors. The values in the rows represent settings for each factor in the individual runs of the design.

Design parameters: The number of levels, factors, replicates, and blocks within the design.

Design of experiment space: An imaginary area bounded by the extremes of the tested factors. (It is also called the *experimental region.*)

Deterministic: An outcome that does not vary (i.e., it is always the same) for a given set of input factors.

Diagnostics: Statistics and plots, often involving model residuals, which assess the assumptions underlying a statistical analysis.

Distribution: A spatial array of data values.

D-optimal: A criterion for choosing design points that minimizes the volume of the joint confidence intervals for the model coefficients, thereby making them most precise.

Dot plot: A method for recording a response by simply putting points on a number line.

Effect: The change in average response when a factor, or interaction of factors, goes from its low level to its high level.

Envelope: The operating boundaries of a process.

Error term: The term in the model that represents random error. The data residuals are used to estimate the nature of the error term. The usual assumption is that the error term is normally and randomly distributed about zero, with a standard deviation of sigma.

Experiment: A series of test runs for the purpose of discovery.

Experimental region: *See* **Design of experiment space**.

Externally studentized residual: *Also see* **Studentized residual**. This statistic tests whether a run is consistent with other runs, assuming the chosen model holds. Model coefficients are calculated based on all design points except one. A prediction of the response at this point is then produced. The externally studentized residual measures the number of standard deviations difference between this new predicted value (lacking the point in question) and the actual response. As a rule of thumb, an externally studentized residual greater than 3.5 indicates that the point should be examined as a possible outlier. (For a more exact rule, apply the Bonferroni correction (α/n) and use the two-tailed t-statistic with residual df for limits.) Note: This statistic becomes undefined for points with leverages of one.

Factor: The independent variable to be manipulated in an experiment.

Factorial design: A series of runs in which combinations of factor levels are included.

F-distribution: A probability distribution used in analysis of variance. The F-distribution is dependent on the degrees of freedom (df) for the mean square in the numerator and the df of the mean square in the denominator of the F-ratio.

Foldover: A method for augmenting low-resolution, two-level factorial designs that requires adding runs with opposite signs to the existing block of factors.

Fractional factorial: An experimental design including only a subset of all possible combinations of factor levels, causing some of the effects to be aliased.

F-test: *See* **F-value**.

Full factorial: An experimental design including all possible combinations of factors at their designated levels.

F-value: The F-distribution is a probability distribution used to compare variances by examining their ratio. If they are equal, the F-value is 1. The F-value in the ANOVA table is the ratio of model mean square (MS) to the appropriate error mean square. The larger their ratio, the larger the F-value and the more likely that the variance contributed by the model is significantly larger than random error. (Also called the *F-test*.)

General factorial: A type of full factorial that includes some categoric factors at more than two levels, also know as "multilevel categoric."

Half-normal: The normal distribution folded over to the right of the zero point by taking the absolute value of all data. Usually refers to a plot of effects developed by statistician Cuthbert Daniel.

Heredity: *See* **Hierarchy**.

Hierarchy: (It is referred to as *heredity*.) The ancestral lineage of effects flowing from main effects (parents) down through successive generations of higher order interactions (children). For statistical reasons, models containing subsets of all possible effects should preserve hierarchy. Although the response may be predicted without maintaining hierarchy when using the coded variables, predictions will not be the same in the actual factor levels unless hierarchy is preserved. Without hierarchy, the model will be scale-dependent.

Homogeneous: Consistent units such as lot-to-lot or operator-to-operator.

Hypothesis (H): A mathematical proposition set forth as an explanation of a scientific phenomena.

Hypothesis test: A statistical method to assess consistency of data with a stated hypothesis.

Identity column (I): (Alternatively: Intercept.) A column of all pluses in the design matrix used to calculate the overall average.

Independence: A desirable statistical property where knowing the outcome of one event tells nothing about what will happen from another event.

Individuals: Discrete subjects or data from the population.

Interaction: The combined change in two factors that produces an effect different than that of the sum of effects from the two factors. Interactions occur when the effect one factor has depends on the level of another factor.

Intercept: The constant in the regression equation. It represents the average response in a factorial model created from coded units.

Internally studentized residual: The residual divided by the estimated standard deviation of that residual.

Irregular fraction: A two-level fractional factorial design that contains a total number of runs that is not a power of two. For example, a 12-run fraction of the 16-run full-factorial design on four factors. This is a 3/4 irregular fraction.

Lack of fit (LOF): A test that compares the deviation of actual points from the fitted surface, relative to pure error. If a model has a significant lack of fit, it should be investigated before being used for prediction.

Lake Wobegon Effect: A phenomenon that causes all parents to believe their children are above the mean. It is named after the mythical town in Minnesota, where, according to author Garrison Keillor, all women are strong, men are good looking, and children are above average.

Least significant difference (LSD): A numerical value used as a benchmark for comparing treatment means. When the LSD is exceeded, the means are considered to be significantly different.

Least squares: *See* **Regression analysis**.

Level: The setting of a factor.

Leverage: The potential for a design point to influence its fitted value. Leverages near 1 should be avoided. If leverage is 1, then the model is forced to go through the point. Replicating such points reduces their leverage.

Linear model: A polynomial model containing only linear or main effect terms.

LSD bars: Plotted intervals around the means on effect graphs with lengths set at one-half the least significant difference. Bars that do not overlap indicate significant pair-wise differences between specific treatments.

Lurking variable: An unobserved factor (one not in the design) causing a change in response. A classic example is the study relating to the number of people and the number of storks in Oldenburg, which led to the spurious conclusion that storks cause babies.

Main effect: The change in response caused by changing a single factor.

Mean: The sum of all data divided by the number of data—a measure of location. (It also is called *average*.)

Mean square: A sum of squares divided by its degrees of freedom (SS/df). It is analogous to a variance.

Median: The middle value.

Mixed factorial: *See* **General factorial**.

Mixture model: *See* **Scheffé polynomial**.

Mode: The value that occurs most frequently.

Model: An equation, typically a polynomial, that is fit to the data.

Model coefficient: The coefficient of a factor in the regression model. (It is also called *parameter* or *term*.)

Multicollinearity: The problem of correlation of one variable with others, which arises when the predictor variables are highly interrelated (i.e., some predictors are nearly linear combinations of others). Highly collinear models tend to have unstable regression coefficient estimates.

Multilevel categoric: *See* **General factorial**.

Multiple response optimization: Method(s) for simultaneously finding the combination of factors giving the most desirable outcome for more than one response.

Nonlinear blending: A second-order effect in a mixture model that captures synergism or antagonism between components. This differs from a simpler factor interaction by the way it characterizes curvature in the response surface for predicted behavior of the mixture as a function of its ingredients.

Normal distribution: A frequency distribution for variable data, represented by a bell-shaped curve symmetrical about the mean with a dispersion specified by its standard deviation.

Normal probability plot: A graph with a y-axis that is scaled by cumulative probability (Z) so normal data plots as a straight line.

Null: Zero difference.

Numeric variable: A quantitative factor that can be varied on a continuous scale, such as temperature.

Observation: A record of factors and associated responses for a particular experimental run (trial).

OFAT: One-factor-at-a-time method of experimentation (as opposed to factorial design).

Order: A measure of complexity of a polynomial model. For example, first-order models contain only linear terms. Second-order models contain linear terms plus two-factor interaction terms and/or squared terms. The higher the order, the more complex shapes the polynomial model can approximate.

Orthogonal arrays: Test matrices exhibiting the property of orthogonality.

Orthogonality: A property of a design matrix that exhibits no correlation among its factors, thus allowing them to be estimated independently.

Outlier: A design point where the response does not fit the model.

Outlier t-test: *See* **Externally studentized residual**.

Parameter: *See* **Model coefficient**.

Pencil test: A quick and dirty method for determining whether a series of points fall on a line.

Plackett–Burman design: A class of saturated orthogonal (for main effects) fractional two-level factorial designs where the number of runs is a multiple of four, rather than 2^k. These designs are resolution III.

Poisson: A distribution characterizing discrete counts, such as the number of blemishes per unit area of a material surface.

Polynomials: Mathematical expressions, composed of powers of predictors with various orders, used to approximate a true relationship.

Population: A finite or infinite collection of all possible individuals who share a defined characteristic, e.g., all parts made by a specific process.

Power: The probability that a test will reveal an effect of stated size.

Power law: A relationship where one variable (e.g., standard deviation) is proportional to another variable (such as the true mean) raised to a power.

Predicted R-squared: Measures the amount of variation in new data explained by the model. It makes use of the predicted residual sum of squares (PRESS) as shown in the following equation: Predicted R-squared = $1 - SS_{PRESS}/(SS_{TOTAL} - SS_{BLOCKS})$.

Predicted residual sum of squares (PRESS): A measure, the smaller the better, of how well the model fits each point in the design. The model is repeatedly refitted to all the design points except the one being predicted. The difference between the predicted value and actual value at each point is then squared and summed over all points to create the PRESS.

Predicted value: The value of the response predicted by the mathematical model.

Prob > F (Probability of a larger F-value): The p-value for a test conducted using an F-statistic. If the F-ratio lies near the upper tail of the F-distribution, the probability of a larger F is small and the variance ratio is judged to be significant. The F-distribution is dependent on the degrees of freedom (df) for the mean square in the numerator and the df of the mean square in the denominator of the F-ratio.

Prob > t (Probability of a larger t-value): The p-value for a test conducted using a t-statistic. Small values of this probability indicate significance and rejection of the null hypothesis.

Probability paper: Graph paper with specially scaled y-axis for cumulative probability. The purpose of the normal probability paper is to display normally distributed data as a straight line. It is used for diagnostic purposes to validate the statistical assumption of normality.

Process: Any unit operation, or series of unit operations, with measurable inputs and outputs (responses).

Pure error: Experimental error, or pure error, is the normal variation in the response, which appears when an experiment is repeated. Repeated experiments rarely produce exactly the same results. Pure error is the minimum variation expected in a series of experiments. It can be estimated by replicating points in the design. The more replicated points, the better will be the estimate of the pure error.

p-value: Probability value, usually relating to the risk of falsely rejecting a given hypothesis.

Quadratic: A second order polynomial.

Qualitative: *See* **Categoric variable**.

Quantitative: *See* **Numeric variable**.

Randomization: Mixing up planned events so each event has an equal chance of occurring in each position, particularly important to ensure that lurking variables do not bias the outcome. Randomization of the order in which experiments are run is essential to satisfy the statistical requirement of independence of observations.

Range: The difference between the largest and smallest value—a measure of dispersion.

Regression analysis: A method by which data are fitted to a mathematical model. (It is also called the method of *least squares*.)

Replicate: An experimental run performed again from start to finish (not just resampled and/or remeasured). Replication provides an estimate of pure error in the design.

Residual (or "Residual error"): The difference (sometimes referred to as *error*) between the observed (actual) response and the value predicted by the model for a particular design point.

Response: A measurable product or process characteristic thought to be affected by experimental factors.

Response surface methods (RSM): A statistical technique for modeling responses via polynomial equations. The model becomes the basis for 2-D contour maps and 3-D surface plots for purposes of optimization.

Risk: The probability of making an error in judgment (i.e., falsely rejecting the null hypothesis). (*Also see* **Significance level**.)

Root mean square error: The square root of the residual mean square error. It estimates the standard deviation associated with experimental error.

R-squared: The coefficient of determination. It estimates the fraction (a number between zero and one) of the overall variation in the data accounted for by the model. This statistic indicates the degree of relationship of the response variable to the combined linear predictor variables. Because this raw R-squared statistic is biased, use the adjusted R-squared instead.

Rule of thumb: A crude method for determining whether a group of points exhibit a nonrandom pattern: If, after covering any point(s) with your thumb(s), the pattern disappears, there is no pattern.

Run: A specified setup of process factors that produces measured response(s) for experimental purposes. (It is also called a *trial.*)

Run order: Run order is the randomized order for experiments. Run numbers should start at one and include as many numbers as there are experiments. Runs must be continuous within each block.

Sample: A subset of individuals from a population, usually selected for the purpose of drawing conclusions about specific properties of the entire population.

Saturated: An experimental design with the minimum number of runs required to estimate all effects.

Scheffé polynomial: A form of mathematical predictive model designed specifically for mixtures. These models are derived from standard polynomials, of varying degrees, by accounting for the mixture constraint that all components sum to the whole. (It is also called a *mixture model.*)

Screening: Sifting through a number of variables to find the vital few. Resolution IV two-level fractional factorial designs are often chosen for this purpose.

Significance level: The level of risk, usually 0.05, established for rejection of the null hypothesis.

Simplex: A geometric figure with one more vertex than the number of dimensions. For example, the two-dimensional simplex is an equilateral triangle. This shape defines the space for three mixture components that each can vary from 0 to 100%.

Simplex centroid: A mixture design comprised of the purest blends, binary combinations, etc., up to and including a centroid blend of all components.

Sparsity of effects: A rule-of-thumb that about 20% of main effects and two-factor interactions will be active in any given system. The remainder of main effects, two-factor interactions, and all higher-order effects, are near zero, with a variation based on underlying error.

Split plot: An experiment design that conveniently groups hard-to-change (HTC) factors, which are set up in randomized order, within which the

easy-to-change (ETC) factors vary according to a random plan. The groups are called *whole plots* and the splits are referred to as *subplots.*

Standard deviation: A measure of variation in the original units of measure, computed by taking the square root of the variance.

Standard error: The standard deviation usually associated with a parameter estimate rather than individuals.

Standard error of a parameter: The estimated standard deviation of a parameter or coefficient estimate.

Standard order: A conventional ordering of the array of low and high factor levels versus runs in a two-level factorial design.

Star points: Axial points in a central composite design.

Statistic: A quantity calculated from a sample to make an estimate of a population parameter.

Studentized: A value divided by its associated standard error. The resulting quantity is a Z-score (number of standard deviations) useful for purposes of comparison.

Stuff: Processed material such as food, pharmaceutical, or chemical (as opposed to “thing”).

Subplot: Experimental units that the whole plot is split into.

Sum of squares (SS): The sum of the squared distances of the actual values from an estimate.

Synergism: A desirable interaction of two factors where the combination produces a response that is better than what would be expected from either one alone. The same concept can be applied to higher-order interactions.

Tetrahedron: A three-dimensional geometric figure with four vertices. It is a simplex. The tetrahedron looks like a pyramid, but has only three sides, not four.

Thing: Manufactured hard goods, such as electronics, cars, medical devices (as opposed to “stuff”).

Transformation: A mathematical conversion of response values (for example, logY).

Treatment: A procedure applied to an experimental unit. One usually designs experiments to estimate the effect of the procedures (treatments) on the responses of interest.

Trivial many: Nonactive effects. The sparsity of effects principal predicts that all interactions of third order or higher will fall into this category, as well as 80% of all main effects and two-factor interactions.

True: Related to the population rather than just the sample.

Trial: *See* **Experiment**.

t-value: A value associated with the t-distribution that measures the number of standard deviations separating the parameter estimate from zero.

Type 1 error: Saying something happened when it really didn't (a false alarm).

Type 2 error: Not discovering that something really happened (failure to alarm).

Uniform distribution: A frequency distribution where the expected value is a constant, exhibiting a rectangular shape.

Variable: A factor or response that assumes assigned or measured values.

Variance: A measure of variability computed by summing the squared deviations of individual data from their mean, then dividing this quantity by the degrees of freedom.

Vertex: A point representing an extreme combination of input variables subject to constraints. Normally used in conjunction with mixture components.

Vital few: The active effects. (*Also see* **Sparsity of effects**.)

Whole plot: Largest experimental unit (group) in a split-plot design.

X-space: *See* **Design of experiment space**.

Y-bar ($\bar{Y}$): A response mean.

Recommended Readings

Textbooks

Amsden, R., and H. Butler. 1998. *SPC simplified: Practical steps to quality,* 2nd ed. New York: Productivity Press.

Anderson, M. and P. Whitcomb. 2009. *A primer on mixture design: What's in it for formulators?* Minneapolis: Stat-Ease, Inc. Online at: www.statease.com/pubs/MIXprimer.pdf

Anderson, M., and P. Whitcomb. 2004. *RSM simplified: Optimizing processes using response surface methods for design of experiments*. New York: Productivity Press.

Box, G., J. S. Hunter, and W. Hunter. 2005. *Statistics for experimenters,* 2nd ed. New York: John Wiley & Sons.

Cornell, J. 2002. *Experiments with mixtures,* 3rd ed. New York: John Wiley & Sons.

Gonick, L., and W. Smith. 1994. *The cartoon guide to statistics*. New York: HarperCollins.

John, P. 1969. *Statistical design and analysis of experiments*. New York: Macmillan.

Montgomery, D. 2012. *Design and analysis of experiments,* 8th ed. New York: John Wiley & Sons.

Montgomery, D., R. Myers, and C. Anderson-Cook. 2009. *Response surface methodology,* 3d ed. New York: John Wiley & Sons.

Phillips, J. 1999. *How to think about statistics,* 6th ed. New York: Owl Books.

Smith, W. 2005. *Experimental design for formulation*. Philadelphia: ASA-SIAM Series on Statistics and Applied Probability.

Case Study Articles

For a wealth of case studies detailing application of DOE over a broad spectrum of industries, see www.statease.com/publications/case-studies.html.

This site is a particularly good one to check if you want to see whether this powerful statistical tool has proved useful in your field. The authors maintain an extensive file of articles on applications of DOE. We welcome queries about or requests for materials in our database and invite you to contact us via stathelp@statease.com.

Index

About the Authors

Mark J. Anderson, PE, CQE, MBA, is a principal and general manager of Stat-Ease, Inc. (Minneapolis, Minnesota). A chemical engineer by profession, he also has a diverse array of experience in process development (earning a patent), quality assurance, marketing, purchasing, and general management. Prior to joining Stat-Ease, he spearheaded an award-winning quality improvement program (which generated millions of dollars in profit for an international manufacturer) and served as general manager for a medical device manufacturer. His other achievements include an extensive portfolio of published articles on design of experiments. Anderson co-authored (with Whitcomb) *RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments* (Productivity Press, 2004).

Patrick J. Whitcomb, PE, MS, is the founding principal and president of Stat-Ease, Inc. Before starting his own business, he worked as a chemical engineer, quality assurance manager, and plant manager. Whitcomb developed Design-Ease® software, an easy-to-use program for design of two-level and general factorial experiments, and Design-Expert® software, an advanced user's program for response surface, mixture, and combined designs. He has provided consulting and training on the application of design of experiments (DOE) and other statistical methods for decades. In 2013, the Minnesota Federation of Engineering, Science, and Technology Societies (MFESTS) awarded Whitcomb the Charles W. Britzius Distinguished Engineer Award for his lifetime achievements.

This is the third edition of Anderson and Whitcomb's book.

About the Software

To make DOE easy, this book is augmented with fully functional time-limited version of a commercially available computer program from Stat-Ease, Inc.—Design-Expert® software. Download this Windows-based package, as well as companion files in Adobe's portable document format (PDF) that provide tutorials on the one-factor, factorial, general multilevel categoric factorial and other, more advanced, designs, from www.statease.com/simplified.html. There you will also find files of data for most of the exercises in the book: The datasets are named and can be easily cross-referenced with corresponding material in the book. Some data is in Microsoft Excel spreadsheet format ("xls*").

You are encouraged to reproduce the results shown in the book and to explore further. The Stat-Ease software offers far more detail in statistical outputs and many more graphics than can be included in this book. You will find a great deal of information on program features and statistical background in the on-line hypertext help system built into the software.

BEFORE YOU START USING
THE SOFTWARE CHECK FOR UPDATES!

Before getting started with the software, check www.statease.com/simplified.html for update patches. Add this path to the Favorites folder in your Internet web browser. You can also download the data for case studies and problems discussed throughout the book from this website. Also, from this web page link to answers posted for all the practice problems.

Technical support for the software can be obtained by contacting:

Stat-Ease, Inc.
2021 East Hennepin Ave, Suite 480
Minneapolis, MN 55413
Telephone: 612-378-9449
Fax: 612-378-2152
E-mail: support@statease.com
Website: www.statease.com