CHAPTER

19

Analysis of Variance and Experimental Design

Tom Stewart/Corbis/Glow Images, Inc.

USING CONTROLLED EXPERIMENTS IN BUSINESS

Can statistical principles and careful experimentation lead to improved products and lower costs? They certainly can, argues Rita Koselka in the article "The New Mantra: MVT" in *Forbes* magazine [Roselka (1996)]. Most products are a result of several controllable inputs. The question is, Which combination of input settings results in the best quality and lowest cost? This is where experimentation enters the picture. For years, companies experimented (if they experimented at all) by changing the level of one input at a time. This one-at-a-time method of experimentation is not only costly and time-consuming, but it often fails to identify the best *combination* of input settings. As we will discuss in Section 19-5, the input factors often interact, so that the best setting of one factor might depend on the settings of other factors, and one-at-a-time testing will probably not discover this fact. A better alternative is to test multiple factors *simultaneously*. This is called **multiple variable testing**, or **MVT**, and it is quickly becoming regarded as one of the most important statistical techniques for product improvement. If you have never heard of MVT, you probably will. It is a natural outgrowth of the quality control movement that has been so pervasive in the past two decades. As the article states, "It [MVT] doesn't just tell you how to raise the quality of your output. It tells you how to do that cost-effectively."

The potential improvements with MVT can occur in traditional manufacturing and service industries. The following are several examples discussed in the article.

- Several years ago a subsidiary of Raychem Corp. called Elo TouchSystems was losing $3 million annually making touch-sensitive computer screens for products like automatic teller machines. The problem was a bubbling between the screen and the coating, and it resulted in a disastrous

25% reject rate. Raychem had spent 18 months on quality improvement efforts, but they hadn't worked. Then the company hired statistical consultants who experimented with MVT. Their solution, which would never have surfaced with one-at-a-time testing, was to change three things at the same time: the type of polyester, the coversheet shaping process, and the adhesive. Within months, the reject rate had decreased to less than 1%, many fewer quality inspectors were required, and the company was breaking even. After the changes, it began making $15 million on $50 million in sales.

- For years, Boise Cascade had been experimenting, with limited success, with small variations in its pulping process at a Louisiana paper mill. After using MVT with eight variables, it came to the counterintuitive conclusion that the mill could maintain its paper quality while switching to a cheaper grade of wood. The result was paper of at least as high a quality level and a savings of $3 million per year.
- Saint Luke's Hospital in Kansas City was concerned about the misuse of warfarin, an anti-bloodclotting drug that can be fatal if used improperly. In 1992 the hospital worked with statistical consultants to experiment with ways to keep patients from misusing the drug. They tested seven variables to better educate patients and provide emergency access to nurses. They found that having a standardized instruction sheet and having the pharmacist discuss the drug with patients yielded a 68% improvement in patient understanding of how to use the drug appropriately.
- A shoe company selling sneakers in over 100 stores was considering a proposal to increase sales with a costly, high-tech display. Before doing so, it hired statistical consultants, who persuaded the company to experiment with a whole range of possible changes: in sales techniques, advertising, separation of shoes by color, and various discounts, as well as the displays. The findings were surprising. Although the new display had the potential to increase sales, its impact would not be nearly as great as a simple combination of using the old display case and arranging the shoes by color. Using this suggestion, the company did not spend money on new displays, and it was still able to increase sales by 33%.

MVT is not a new scientific method. Statisticians have studied and applied experimental designs for years, particularly in the physical sciences. However, it is relatively new to business. For the most part, the well-known strategic consultants and the big accounting firms have been strangers to MVT. Fortunately, this is changing. For example, experimental design is now being taught formally at Motorola University in Illinois. As Roselka concludes, "The design of experiments involves some cleverness. It may cost a lot of money to shut down a production line and rearrange it; it may take precious months for a billing department or a mail-order operation to see whether a novel way of doing things will pay off. MVT is the science of gleaning the most amount of information from the least amount of costly testing." ■

19-1 INTRODUCTION

One of the most frequent applications of statistics is the comparison of several populations on some characteristic. We discussed the simplest version of this comparison problem in Chapters 8 and 9 when we discussed the two-sample procedure for analyzing the difference between two population means. A natural extension is to *more* than two population means, which is the topic of this chapter. The resulting procedure is commonly called **analysis of variance**, or **ANOVA**.

There are two typical situations where ANOVA is used. The first is when there are several distinct populations. For example, consider recent graduates with BS degrees in one of three disciplines: Business, Engineering, and Computer Science. We might sample randomly from each of these populations to discover whether there are any significant differences among them with respect to mean starting salary. A second situation where ANOVA is used is in randomized experiments. In this case a *single* population is treated in one of several ways. For example, a pharmaceutical company might select a group of people who suffer from allergies and randomly assign each person to a different type of allergy medicine currently being developed. Then the question is whether any of the treatments differ from one another with respect to the mean amount of symptom relief.

These two examples illustrate two basic situations where ANOVA is used. The comparison of recent graduates is called an **observational study**. In this case we analyze the data that are already available to us, that is, the starting salaries of recent graduates from the three disciplines. Unfortunately we don't first get to choose which students should major in which disciplines. It might be nice to do so because it would help to rule out other possible causes besides discipline, such as unequal academic abilities, that might affect starting salaries. But we don't get to make these choices. The students themselves choose their disciplines, and all we can do is analyze the resulting data on starting salaries.

In an **observational study**, we analyze data already available to us. The disadvantage is that it is difficult or impossible to rule out factors over which we have no control for the effects we observe.

In contrast, the allergy example illustrates a **designed experiment**. The researchers in this example are interested in whether different allergy medicines cause different amounts of symptom relief. Therefore, they will select the subjects for the experiment so that the subjects receiving one allergy medicine are as much alike, in every characteristic that might matter—age, medical history, gender mix, and so on—as the subjects receiving any other allergy medicine. In this way, if there are any differences across groups with respect to symptom relief, the researchers will be able to attribute the differences to the types of medicine, not some extraneous factor.

In a **designed experiment**, we control for various factors such as age, gender, or socioeconomic status so that we can learn more precisely what is responsible for the effects we observe.

It should be clear from this discussion that designed experiments are generally preferable to observational studies. In a carefully designed experiment, where we can "control for" extraneous factors such as age or gender that are not of direct interest, we can be fairly sure that any differences across groups with respect to some measurement variable are due to the variables that we purposely manipulate. This ability to infer causal relationships is never possible with observational studies. For example, if recent Business graduates are found to make more, on average, than Computer Science graduates, we can never be sure whether this is a result of being a *Business* graduate rather than a *Computer Science* graduate or whether, say, it is due to the fact that the Business graduates in our study have more work experience than the Computer Science graduates. We didn't control for work experience, so we cannot rule out the possibility that it might have had an effect.

ANOVA has been used in many disciplines. In fact, it began in agricultural studies, where researchers wanted to learn, for example, which types of wheat produce the greatest yield per acre. Because the results from such an experiment can take many months to obtain, the agricultural researchers had to design their experiments very carefully, so that

they could obtain the most *information* from the resulting data. This idea of obtaining the most useful information from a limited amount of data continues to be crucial in ANOVA studies and has spawned a whole area of scientific research called **experimental design**. The essential goal of experimental design is to decide which observations to make, given a limited budget (in time and/or money), to maximize the chances of seeing differences across groups that actually exist. For example, the allergy researchers want to design their experiment so that if there really are differences across medicine types, the analysis will have a good chance of detecting them.

Experimental design is the science (and art) of setting up an experiment so that the most information can be obtained for the time and money involved.

We will concentrate on the most common and basic experimental designs in this chapter, leaving more complex designs to specialized books. However, because our audience is mostly *business* students, it is important to note that the use of designed experiments in business situations is probably less prevalent than in, say, medicine or agriculture. Business managers do not always have the luxury of being able to design a controlled experiment for obtaining data. Instead, they often have to rely on whatever data are available, that is, observational data. Nevertheless, as the introductory vignette to this chapter attests, there are many potentially profitable uses of experimental design in the business world, and many companies are beginning to use designed experiments for competitive advantage.

It is no coincidence that some of this terminology is the same as that used in regression analysis. We will see why later in this chapter, when we investigate the relationship between ANOVA and regression.

Before proceeding, there is some general terminology we should introduce. In all of our examples, there is a variable of primary interest that we wish to measure. It is called the **dependent variable** (or sometimes the **response** or **criterion variable**) and is the variable we measure to detect differences among groups. The groups themselves are determined by one or more **factors** (sometimes called **independent** or **explanatory variables**), each varied at several **treatment levels** (often shortened to **levels**). The number of factors determines the type of ANOVA. If there is a single factor, the procedure is called **one-way ANOVA**; if there are two factors, it is called **two-way ANOVA**; if there are three factors, it is called **three-way ANOVA**; and so on. The only types we will discuss in this book are the two most common types, one-way and two-way ANOVA. It is best to think of a factor as a categorical variable, with the possible categories being its levels. Finally, the "entities" measured at each treatment level (or combination of levels) are called **experimental units**. Some examples will help to clarify this terminology.

In **one-way ANOVA**, a single dependent variable is measured at various levels of a single factor. Each experimental unit is assigned to one of these levels. In **two-way ANOVA**, a single dependent variable is measured at various combinations of the levels of two factors. Each experimental unit is assigned to one of these combinations of levels.

Consider the observational study on graduates of Business, Engineering, and Computer Science. The dependent variable is starting salary, the experimental units are the individual graduates, and the *single* factor is the student's major discipline. This factor has three levels: Business, Engineering, and Computer Science, and each student is "assigned" to one of these levels. If we also wanted to see how gender affects starting salary, we could introduce a second factor, gender, at the two levels "male" and "female." Then each student would be "assigned" to one of the combinations of levels, such as a female in Business.

For the study on allergy medicines, the dependent variable is the amount of relief from allergy symptoms, the experimental units are the individual patients, and the single factor

is medicine type. Its levels are the various types of medicines used in the experiment. Each patient in the experiment receives exactly one of these types of medicine. We could also introduce a second factor here. For example, if we purposely wanted to see whether age has an effect on the dependent variable (or on which medicine type is most effective), we could introduce a second factor, age, with levels such as 5 to 20, 20 to 35, 35 to 50, and 50 or older. Then each patient would be at some combination of the two factors, such as a person 20 to 35 years old receiving the second type of medicine.

Although the experimental units in both of these two examples are people, this is not always the case. Suppose a company wants to see whether five different shelf layouts for its product lead to different levels of sales. The company could choose a sample of 50 supermarkets that sell its product, try the first layout in 10 of them, the second layout in another 10, and so on. Then the dependent variable is sales level, the single factor is shelf layout, varied at five levels, and the experimental units are the 50 supermarkets. Note that in this example, each of the experimental units, that is, each supermarket, is chosen (probably in some random way) to "receive" one of the five treatments and each treatment is applied to a separate subset of 10 supermarkets. When there are an equal number of experimental units assigned to each treatment level (or combination of levels, for a two-factor or multi-factor design), this is called a **balanced design**. Balanced designs are somewhat easier to analyze, and we prefer them whenever possible. In fact, the only two-factor design we will discuss in this book is a balanced design.

> In a **balanced design**, an equal number of experimental units is assigned to each combination of treatment levels.

19-2 ONE-WAY ANOVA

We begin our discussion with the simplest design to analyze, the one-factor design. As discussed in the introduction, there are basically two situations. First, the data could be observational data, in which case the levels of the single factor might best be considered as "subpopulations" of an overall population—graduates of Business, Engineering, and Computer Science, for example. Second, the data could be generated from a designed experiment, where a single population of experimental units, allergy patients, say, is treated by different types of allergy medicine. Fortunately, the data analysis is basically the same in either case. We normally ask two questions. First, are there any significant differences in the mean of the dependent variable across the different groups? If the answer to this question is "yes," then we typically ask the second question: Which of the groups differs significantly from which others, again with respect to the mean of the dependent variable?

19-2a The Equal-Means Test

We set up the first question as a hypothesis test. Let J be the number of levels of the single factor, and let μ_j be the mean of the dependent variable for level j. (As usual, this Greek letter is used as a "population" mean, the mean of the dependent variable if *all* experimental units received treatment level j.) The null hypothesis is that there are no differences in population means across treatment levels:

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_J$$

The alternative is then the opposite, namely, that at least one pair of μ's are not equal. If we can reject this null hypothesis at some typical level of significance (usually the 5% or 10% level), then we hunt further to see which means are different from which others. To do this, we typically calculate confidence intervals for differences between pairs of means and

see which of these confidence intervals do *not* include zero. For example, if the confidence interval for the difference $\mu_2 - \mu_4$ extends from 5.35 to 9.31, we would conclude that μ_2 and μ_4 are *not* equal (and that μ_2 is in fact larger than μ_4).

This is the general plan. Now we will see how to put it into action. First, we ask an obvious question: If ANOVA is basically a test of differences between means, why is it called analysis of *variance* and not analysis of *means?* The answer to this question is the key to the procedure. Consider the box plot in Figure 19.1. It corresponds to observations from four treatment levels with slightly different means and fairly large variances. (The large variances are indicated by the relatively wide boxes and long lines extending from them.) From these box plots, would you conclude that the population means differ across the four treatment levels? Would your answer change if the data were instead as in Figure 19.2?[1] We expect that it would.

Figure 19.1
Samples with Large Within Variation

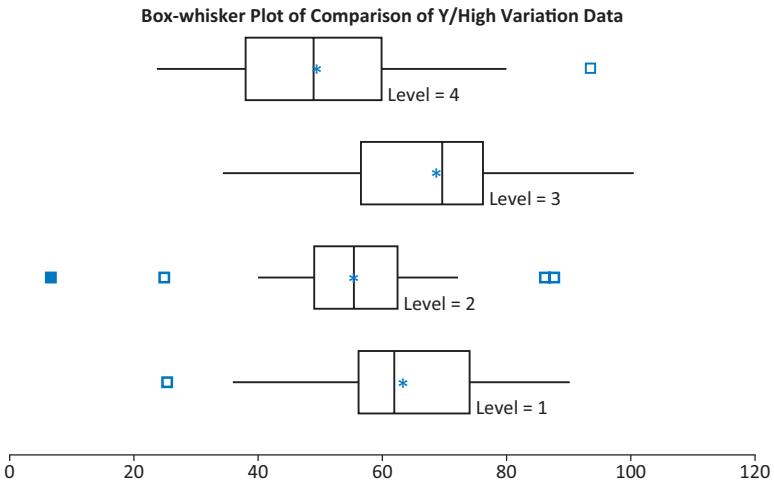
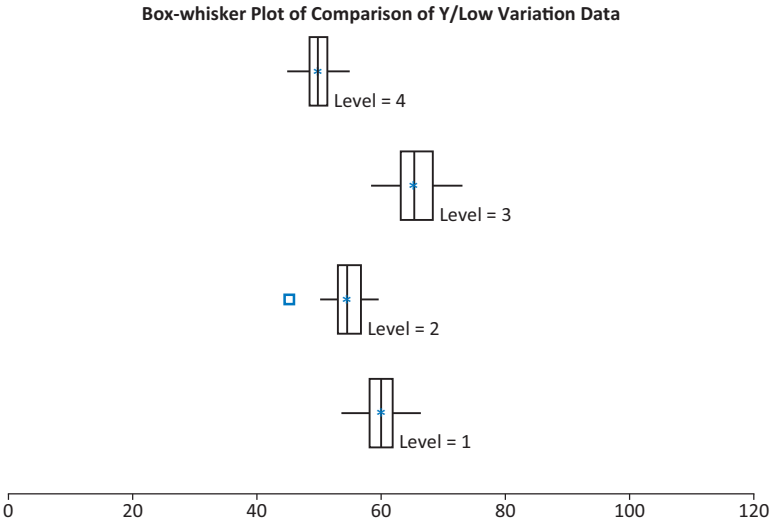


Figure 19.2
Samples with Small Within Variation



[1]Note that we keep the horizontal scale the same in both charts for a fair comparison.

The sample means in these two figures are virtually the same, but the variation *within* each treatment level in Figure 19.1 is quite large relative to the variation *between* the sample means. In contrast, there is very little variation within each treatment level in Figure 19.2. In the first case, the large "within" variation makes it difficult to infer whether there are really any differences across population means, whereas the small "within" variation in the second case makes it clearer that differences across population means probably exist.

This is the essence of the ANOVA procedure. We compare variation *within* the individual treatment levels to variation *between* the sample means. Only if the between variation is large relative to the within variation can we conclude with any assurance that there are differences across population means—and reject the equal-means hypothesis.

Remember that a robust *test is one in which the conclusions are approximately valid even when the assumptions behind it are violated to some extent.*

The test itself is based on two assumptions: (1) The population variances are all equal to some common variance σ^2, and (2) the populations are normally distributed. These are analogous to the assumptions we made for the two-sample procedures in Chapters 8 and 9. Although these assumptions are never satisfied exactly in any real application, you should keep them in mind and check for gross violations whenever possible. Fortunately, the test we present is fairly robust to violations of these assumptions, particularly when the sample sizes are large and roughly the same.

To run the test, let $\bar{Y}_j$, s_j^2, and n_j be the sample mean, sample variance, and sample size from treatment level j. Also, let n and $\bar{\bar{Y}}$ be the combined number of observations and the sample mean of all n observations. ($\bar{\bar{Y}}$ is called the **grand mean**.) Then a measure of the between variance is *MSB* (mean square between), given in Equation (19.1). Note that *MSB* is large if the individual sample means differ substantially from the grand mean $\bar{\bar{Y}}$, and this occurs only if they differ substantially from one another.

Measure of Between Variation

$$MSB = \frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{\bar{Y}})^2}{J - 1} \qquad \textbf{(19.1)}$$

A measure of the within variance is *MSW* (mean square within), given in Equation (19.2). This value is really just a weighted average of the individual sample variances, where the sample variance s_j^2 receives weight $(n_j - 1)/(n - J)$. In fact, *MSW* is the average of the sample variances if the sample sizes, the n_j's, are equal. In this sense, *MSW* is a pooled estimate of the (assumed) common variance σ^2, just as in the two-sample procedures from Chapters 8 and 9. Therefore, *MSW* is large if the individual sample variances are large. For example, *MSW* is much larger in Figure 19.1 than in Figure 19.2. However, *MSB* is about the same in both figures.

Measure of Within Variation

$$MSW = \frac{\sum_{j=1}^{J} (n_j - 1)s_j^2}{n - J} \qquad \textbf{(19.2)}$$

The numerators of Equations (19.1) and (19.2) are called **sums of squares** (often labeled *SSB* and *SSW*), and the denominators are called **degrees of freedom** (often labeled *dfB* and *dfW*). As you will see, they are always reported in ANOVA output. Finally, the ratio of the mean squares is the test statistic we use, the *F*-ratio in Equation (19.3). Under the null hypothesis of equal population means, this test statistic has an *F* distribution with *dfB* and *dfW* degrees of freedom. If the null hypothesis is *not* true, then we would expect *MSB* to be large relative to *MSW*, as in Figure 19.2. Therefore, the *p*-value for the test is found by finding the probability to the *right* of the *F*-ratio in the *F* distribution with *dfB* and *dfW* degrees of freedom.

***F*-ratio for ANOVA Test**

$$F\text{-ratio} = \frac{MSB}{MSW} \tag{19.3}$$

The elements of this test are usually presented in an **ANOVA table**, as you will see shortly. The "bottom line" in this table is the *p*-value for the *F*-ratio. If the *p*-value is sufficiently small, we can conclude that the population means are not all equal. Otherwise, we cannot reject the equal-means hypothesis.

19-2b Confidence Intervals for Differences between Means

If we cannot reject the equal-means hypothesis, then there is little incentive to examine differences between individual pairs of means. However, if we *can* reject the equal-means hypothesis, then it is customary to form confidence intervals for the differences between pairs of population means. This can lead to quite a few confidence intervals. For example, if there are $J = 5$ treatment levels, then there are 10 pairs of differences (the number of ways 2 means can be chosen from a total of 5 means). The confidence interval for any difference $\mu_i - \mu_j$ is of the form shown in Expression (19.4).

Confidence Interval for Difference Between Means

$$\bar{Y}_i - \bar{Y}_j \pm \text{multiplier} \times \sqrt{MSW(1/n_i + 1/n_j)} \tag{19.4}$$

As we will discuss in Section 19-4, there are several possibilities for the appropriate multiplier in this expression. Regardless of the multiplier, however, we are always looking for confidence intervals that do *not* include 0. If the confidence interval for $\mu_i - \mu_j$ is all positive, for example, then we can conclude with high confidence that these two means are not equal and that μ_i is indeed *larger* than μ_j. However, if the confidence interval for $\mu_i - \mu_j$ includes 0, that is, if it extends from a negative number to a positive number, we cannot conclude that these two means are different.

We have presented the formulas for one-way ANOVA because they lend some insight into the procedure. However, the StatTools One-Way ANOVA procedure takes care of all the calculations, as we illustrate in Example 19.1.

EXAMPLE 19.1 The Effect of Shelf Height on Cereal Sales

Does it matter which shelf a popular brand is placed on? It certainly might, because we tend to purchase items that are easiest to see. To test this, suppose that Midway is a large chain of supermarket stores with many stores in many locations. Midway selects 125 of these stores for an experiment. Specifically, it selects these particular 125 stores to be as alike as possible, so that store size, amount of customer traffic, types of customers, and other characteristics are as similar across stores as possible. Each store stocks cereal in a similar location in the store on five-shelf displays. In the experiment, 25 randomly selected stores place a particular popular brand of cereal—we'll call it Brand X—on the lowest shelf for a month. Another randomly selected 25 stores place Brand X on the next-to-lowest shelf, another 25 place it on the middle shelf, another 25 place it on the next-to-highest shelf, and the final 25 place it on the highest shelf. Then the number of boxes of Brand X sold is recorded at each of the stores for the last two weeks of the experiment. (The first two weeks allow customers to get used to the shelving arrangement.) The resulting data are in the file **Cereal Sales.xlsx**, as shown in Figure 19.3 (with some rows hidden). Does shelf height appear to make a difference in sales?

Figure 19.3
Data for Cereal Experiment

	A	B	C	D	E
1	Lowest	Next-to-lowest	Middle	Next-to-highest	Highest
2	340	347	444	456	358
3	376	428	281	471	427
4	378	219	378	484	325
5	371	431	425	448	428
24	389	345	284	564	461
25	417	329	349	395	375
26	250	374	346	546	399

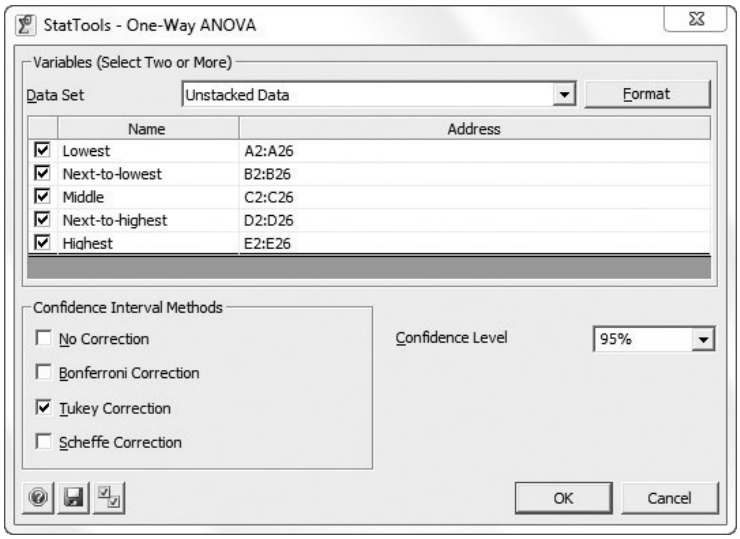
Objective To use one-way ANOVA to see whether shelf height makes any difference in mean sales of Brand X, and if so, to discover which shelf heights outperform the others.

Solution

First, the sample sizes are equal—this is a balanced design. This is not absolutely necessary in an experiment of this type, but since Midway is able to specify which stores use which shelving heights, it makes sense to use a balanced design. Second, this is a designed experiment, not an observational study. Midway deliberately chose the 125 stores in the experiment to be alike in as many ways as possible. This helps to ensure that any differences in sales across the five groups can be attributed to differences in shelf heights and not to other extraneous factors. Of course, it is virtually impossible to control for *all* other factors in an experiment such as this—the 125 stores are certainly not identical in *all* of their characteristics—but Midway has tried its best to keep them similar. Also, it has *randomly* assigned the stores to treatment levels (shelf heights), rather than arbitrarily assigning them. By using a random assignment, Midway avoids any possible bias it might have unconsciously introduced with a nonrandom assignment.

To analyze the data, select One-Way ANOVA from the StatTools Statistical Inference group, and fill in the resulting dialog box as shown in Figure 19.4. In particular, click the Format button and make sure the Unstacked option is checked (because there is a separate sale column for each of the shelf heights), and select all five variables for analysis. (Note that the Stacked option would be appropriate if there were two columns of length 125 each, one with the shelf height and the other with sales.) Finally, make sure only the Tukey option for confidence intervals is checked. (We will discuss these confidence interval options in Section 19-4.)

Figure 19.4
StatTools Dialog Box for Confidence Intervals



The effect of unequal variances is mitigated by having equal, or nearly equal, sample sizes for the treatment levels. This is another reason for using a balanced design.

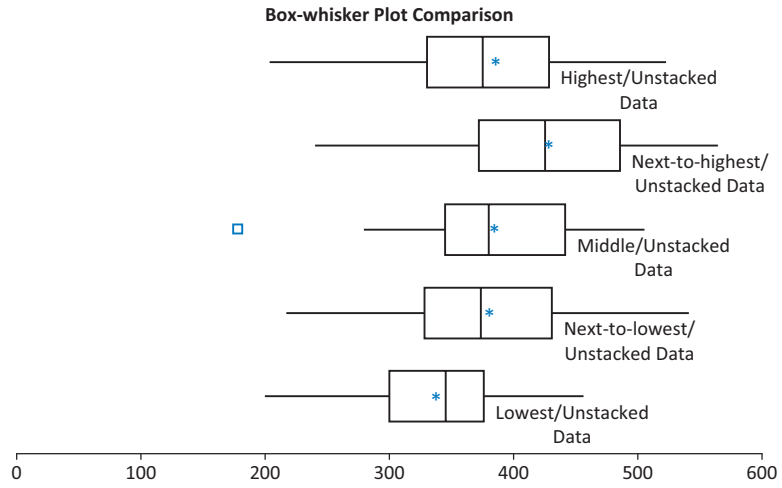
The one-way ANOVA output is shown in Figure 19.5. The summary statistics at the top indicate that the next-to-highest shelf height has the largest average sales, 426.3, almost 100 boxes larger than the lowest shelf height, which has the smallest average sales. This information is confirmed by the side-by-side box plots in Figure 19.6. (Although these box plots are not created as part of the ANOVA output, they are always a useful addition.) The sample standard deviations vary from about 61 to 85 over the five treatment levels. Although these tend to indicate unequal variances, the equal-variance assumption is almost never satisfied exactly in any study, and this much discrepancy in the standard deviations is nothing to worry about—it certainly does not invalidate the analysis.

It appears from the summary statistics and the box plots that mean sales differ for different shelf heights, but are the differences significant? The test of equal means answers

Figure 19.5 One-Way ANOVA Output

	A	B	C	D	E	F
7	***One-Way Anova for Selected Unstacked Variables***					
8	***ANOVA Summary***					
9	**Total Sample Size**	125				
10	**Grand Mean**	381.44				
11	**Pooled Std Dev**	75.63				
12	**Pooled Variance**	5719.19				
13	**Number of Samples**	5				
14	**Confidence Level**	95.00%				
15						
16						
17	***ANOVA Sample Stats***	**Lowest Data Set #1**	**Next-to-lowest Data Set #1**	**Middle Data Set #1**	**Next-to-highest Data Set #1**	**Highest Data Set #1**
18	**Sample Size**	25	25	25	25	25
19	**Sample Mean**	334.92	378.68	383.44	426.28	383.88
20	**Sample Std Dev**	61.04	84.08	75.63	85.05	69.62
21	**Sample Variance**	3726.24	7069.56	5719.17	7234.21	4846.78
22	**Pooling Weight**	0.2000	0.2000	0.2000	0.2000	0.2000
23						
24						
25	***One Way ANOVA Table***	**Sum of Squares**	**Degrees of Freedom**	**Mean Squares**	***F*-Ratio**	***p*-Value**
26	**Between Variation**	104807.68	4	26201.92	4.58	0.0018
27	**Within Variation**	686303.12	120	5719.19		
28	**Total Variation**	791110.80	124			
29						
30						
31	***Confidence Interval Tests***	**Difference of Means**	**Tukey Lower**	**Tukey Upper**		
32	**Lowest-Next-to-lowest**	–43.76	–103.0190249	15.49902487		
33	**Lowest-Middle**	–48.52	–107.7790249	10.73902487		
34	**Lowest-Next-to-highest**	–91.36	**–150.6190249**	**–32.10097513**		
35	**Lowest-Highest**	–48.96	–108.2190249	10.29902487		
36	**Next-to-lowest-Middle**	–4.76	–64.01902487	54.49902487		
37	**Next-to-lowest-Next-to-highest**	–47.60	–106.8590249	11.65902487		
38	**Next-to-lowest-Highest**	–5.20	–64.45902487	54.05902487		
39	**Middle-Next-to-highest**	–42.84	–102.0990249	16.41902487		
40	**Middle-Highest**	–0.44	–59.69902487	58.81902487		
41	**Next-to-highest-Highest**	42.40	–16.85902487	101.6590249		

Figure 19.6
Side-by-Side Box Plots of Sales



this question. It appears in rows 26–28 of the output. The values in this ANOVA table are based on Equations (19.1)–(19.3). (The only part we didn't discuss is the Total variation in row 28. It is based on the total variation of all observations around the grand mean in cell B10 and is used mainly as a check of the calculations. Note that *SSB* and *SSW* in cells B26 and B27 add up to the total sum of squares in cell B28. Similarly, the degrees of freedom add up in column C.) The *F*-ratio in cell E26 is 4.58, the ratio of the mean squares in cells D26 and D27. Its corresponding *p*-value is 0.0018, nearly zero. This leaves practically no doubt that the five population means are *not* all equal. Shelf height evidently does make a significant difference in sales.

The 95% **confidence intervals for ANOVA** in rows 32–41 indicate which shelf heights differ significantly from which others. Any difference whose confidence interval does *not* include 0 is boldfaced. In this example, there is only one such difference, the one between the next-to-highest height and the lowest height. Not surprisingly, these are the treatment levels with the largest and smallest average sales. None of the other differences are significant. For example, even though the difference between the next-to-highest and next-to-lowest heights is 47.6, the corresponding confidence interval extends from a negative number to a positive number. Therefore, we cannot declare this difference to be statistically significant.

The main conclusion from this example is that shelf height definitely appears to make a difference in mean sales, at least for the population of stores similar to the ones in the study. Customers tend to purchase fewer boxes of cereal when they are placed on the bottom shelf, and they tend to purchase more when they are placed on the next-to-highest shelf—presumably right around eye level. ■

19-2c Using a Logarithmic Transformation

Recall that the inferences based on the ANOVA procedure rely on two assumptions: equal variances across treatment levels and normally distributed data. Although these assumptions are never met *exactly* in any real study, you should check whether they are at least approximately valid. Often a look at side-by-side box plots, as in Figure 19.6, can indicate whether there are serious violations of these assumptions. For example, the box plots in this figure are reasonably symmetric and indicate reasonably similar variances, so that the ANOVA results should be valid. If the assumptions are seriously violated, however, you should not blindly report the ANOVA results. In some cases, a transformation of the data will help, as illustrated in Example 19.2.

EXAMPLE 19.2 PAYMENTS FOR ORDERS AT REBCO

Rebco is a manufacturing company that supplies parts to many other manufacturing companies, its customers. Rebco is concerned about the time it takes these customers to pay for their orders. The file **Rebco Payments.xlsx** contains data (a subset of which is shown in Figure 19.7) on the most recent payment from 91 of its customers. The customers are categorized as small, medium, and large. For each customer we see the number of days it took the customer to pay and the amount of the payment. Are there any differences in the mean time to pay across the three customer sizes? What about differences across the mean payment amounts?

Figure 19.7
Data for Rebco Example

	A	B	C	D	E
1	Customer	Customer Size	Days	Amount	Log(Amount)
2	1	Large	12	1352	7.209340257
3	2	Small	21	274	5.613128106
4	3	Small	20	267	5.587248658
5	4	Small	21	229	5.433722004
6	5	Large	14	1870	7.53369371
7	6	Small	29	246	5.505331536
90	89	Medium	22	372	5.918893854
91	90	Large	23	1045	6.951772164
92	91	Medium	15	671	6.508769137

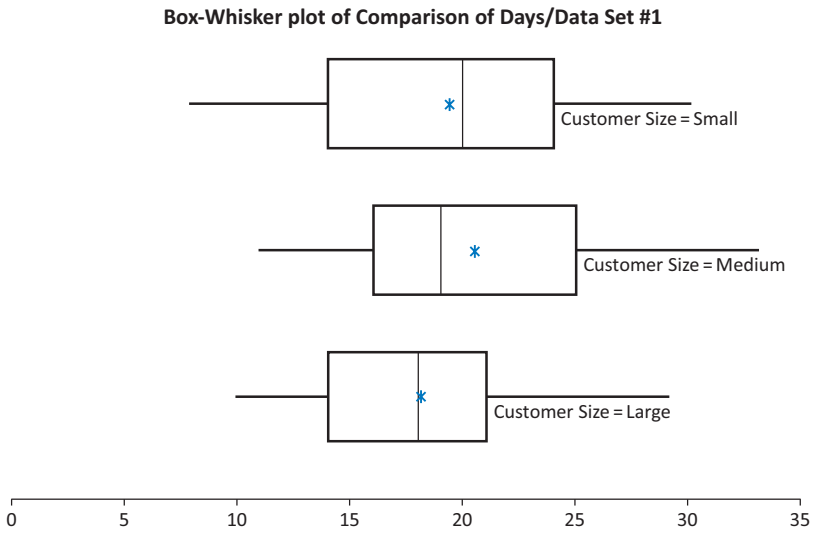
Objective To see how a logarithm transformation can be used to ensure the validity of the ANOVA assumptions, and to see how the resulting output should be interpreted.

Solution

This graph in Figure 19.8 indicates no violations of the ANOVA equal-variance and normality assumptions.

Unlike Example 19.1, this is a one-factor observational study, where the single factor is customer size at three levels: small, medium, and large. The experimental units are the bills for the orders, and there are two dependent variables, days until payment and payment amount, that will be examined. Focusing first on the days until payment, you can see from the side-by-side box plots in Figure 19.8 that whatever differences there are appear to be slight. Perhaps the large customers pay, on average, a bit more promptly, but it is difficult to see from the plots whether the apparent differences are significant. Therefore, we turn to the numerical results. The summary results and the ANOVA table in Figure 19.9 show

Figure 19.8
Box Plots for Days Until Payment



that the differences between the sample means are not even close to being statistically significant. The *p*-value for the test is only 0.318. Rebco cannot reject the null hypothesis that customers of all sizes take, on average, the same number of days to pay.

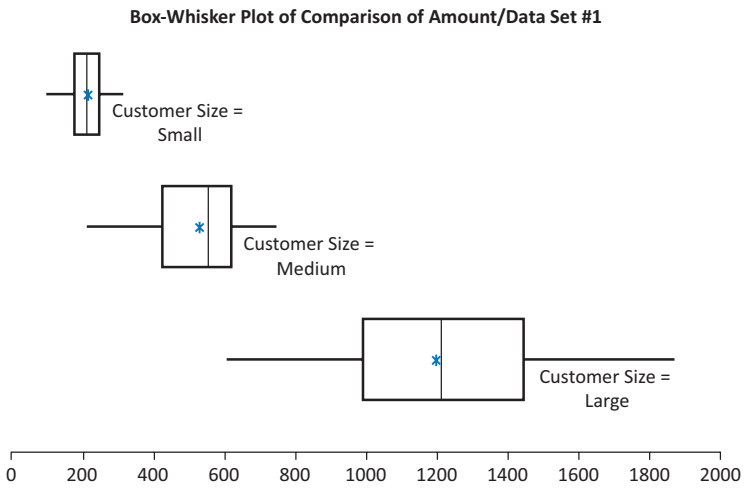
Figure 19.9
ANOVA Results for Days Until Payment

	A	B	C	D	E	F
7	***One-Way ANOVA for Days by Customer Size***					
8	***ANOVA Summary***					
9	**Total Sample Size**	91				
10	**Grand Mean**	19.571				
11	**Pooled Std Dev**	5.769				
12	**Pooled Variance**	33.285				
13	**Number of Samples**	3				
14	**Confidence Level**	95.00%				
15						
16		**Days (Large)**	**Days (Medium)**	**Days (Small)**		
17	***ANOVA Sample Stats***	**Data Set #1**	**Data Set #1**	**Data Set #1**		
18	**Sample Size**	20	39	32		
19	**Sample Mean**	18.100	20.487	19.375		
20	**Sample Std Dev**	4.887	5.707	6.318		
21	**Sample Variance**	23.884	32.572	39.919		
22	**Pooling Weight**	0.2159	0.4318	0.3523		
23						
24		**Sum of**	**Degrees of**	**Mean**		
25	***OneWay ANOVA Table***	**Squares**	**Freedom**	**Squares**	***F*-Ratio**	***p*-Value**
26	**Between Variation**	77.242	2	38.621	1.160	0.3181
27	**Within Variation**	2929.044	88	33.285		
28	**Total Variation**	3006.286	90			

The graph in Figure 19.10 indicates definite violations of the ANOVA equal-variance assumption.

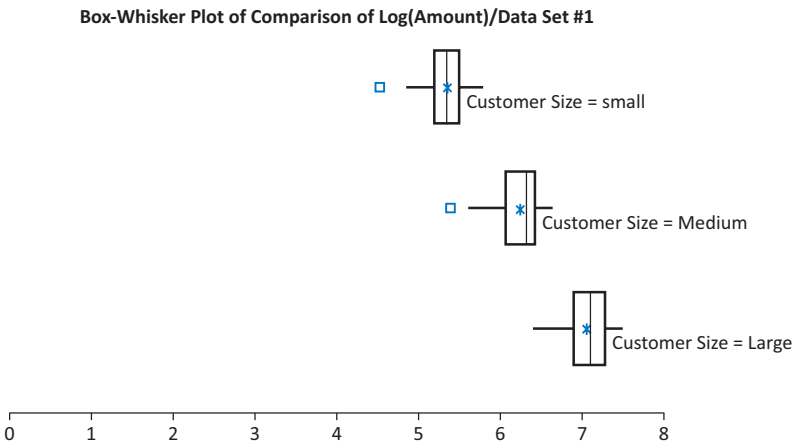
The analysis of the *amounts* these customers pay is quite different. This is immediately evident from the side-by-side box plots in Figure 19.10. Actually, two things are clear. First, there is little doubt that small customers tend to have lower bills than medium-size customers, who in turn tend to have lower bills than large customers. Second, however, you can see that the equal-variance assumption is grossly violated. There is very little variation in payment amounts from small customers and a large amount of variation from large customers. This situation should be remedied before running any formal ANOVA.

Figure 19.10
Box Plots for Payment Amounts



One common method for equalizing variances is to take logarithms of the dependent variable and then use the transformed variable as the new dependent variable. This log transformation tends to spread apart small values and compress together large values—exactly what is needed in this example. After taking the logarithms of the payment amounts, we obtain the box plots in Figure 19.11. The log transformation retains the ordering, so that logs of small amounts are still less than logs of large amounts, but the variances are now much closer to being equal. The resulting ANOVA on the log variable appears in Figure 19.12. The *p*-value in the ANOVA table is again the key for checking whether we can reject the equal-means hypothesis. The fact that it is virtually 0 indicates that the means of the log variables are *not* equal.

Figure 19.11
Box Plots of Log-Transformed Amounts



What does this say about the original variables? The bottom part of the output in Figure 19.12 answers this question, although we have to be very careful when interpreting the results. First, when we ran the StatTools One-Way ANOVA procedure on the log of the Amount variable, we requested the confidence intervals in rows 32–34.[2] However, each of these is a confidence interval for the difference between means of the *log-transformed* variables. Because they are in log units, these numbers have little practical meaning. The trick is to take their antilogarithms (with the EXP function), as shown in rows 37–39, and then interpret the antilogs correctly. It can be shown that the correct interpretation is that each antilog is a *ratio of medians* for the respective treatment levels. (If the populations are reasonably symmetric, the antilogs can also be interpreted as approximate ratios of *means*.) For example, our best guess is that the median amount paid by large customers is 2.253 times as large as the median amount paid by medium-sized customers, and we are 95% confident that this ratio is between 1.877 and 2.705. Because the populations are reasonably symmetric (see the box plots in Figure 19.10), this same statement applies, at least approximately, to the means.

The bottom line for Rebco is that its large customers have bills that are typically over twice as large as those for medium-sized customers, which in turn are typically over twice

[2]Again, we will discuss the type of confidence interval method shown here in Section 19-4.

Figure 19.12 ANOVA Results for Log-Transformed Amounts

	A	B	C	D	E	F
7	***One-Way ANOVA for Log(Amount) by Customer Size***					
8	***ANOVA Summary***					
9	**Total Sample Size**		91			
10	**Grand Mean**		6.0961			
11	**Pooled Std Dev**		0.2787			
12	**Pooled Variance**		0.0777			
13	**Number of Samples**		3			
14	**Confidence Level**		95.00%			
15						
16			**Log(Amount) (Large) Data Set #1**	**Log(Amount) (Medium) Data Set #1**	**Log(Amount) (Small) Data Set #1**	
17	***ANOVA Sample Stats***					
18	**Sample Size**		20	39	32	
19	**Sample Mean**		7.0474	6.2350	5.3323	
20	**Sample Std Dev**		0.3030	0.2863	0.2525	
21	**Sample Variance**		0.0918	0.0820	0.0637	
22	**Pooling Weight**		0.2159	0.4318	0.3523	
23						
24			**Sum of**	**Degrees of**	**Mean**	
25	***OneWay ANOVA Table***		**Squares**	**Freedom**	**Squares**	***F*-Ratio** ***p*-Value**
26	**Between Variation**		37.5191	2	18.7596	241.5401 <0.0001
27	**Within Variation**		6.8346	88	0.0777	
28	**Total Variation**		44.3538	90		
29						
30			**Difference**	**Tukey**		
31	***Confidence Interval Tests***		**of Means**	**Lower**	**Upper**	
32	**Log(Amount) (Large)-Log(Amount)(Medium)**		0.8123	**0.629549626**	**0.995099974**	
33	**Log(Amount) (Large)-Log(Amount)(Small)**		1.7151	**1.525627517**	**1.904488762**	
34	**Log(Amount) (Medium)-Log(Amount)(Small)**		0.9027	**0.744221983**	**1.061244697**	
35						
36	**Antilogs**					
37	**Amount(Large)-Amount(Medium)**		2.253	1.877	2.705	
38	**Amount(Large)-Amount(Small)**		5.557	4.598	6.716	
39	**Amount(Medium)-Amount(Small)**		2.466	2.105	2.890	

as large as those for small customers. Even though all customers currently tend to take about the same number of days to pay, there is a greater incentive to get the large customers to pay early—more money is at stake. ■

The "ratio of medians" interpretation discussed in this example is the correct interpretation in any comparison problem where a log transformation is used (probably to equalize variances) on the dependent variable. It applies not only to ANOVA studies such as Rebco's but to the two-sample *t* procedures discussed in Chapters 8 and 9.

PROBLEMS

Note: Student solutions for problems whose numbers appear within a colored box are available for purchase at www.cengagebrain.com.

Level A

1. An automobile manufacturer employs sales representatives who make calls on dealers. The manufacturer wishes to compare the effectiveness of four different call-frequency plans for the sales representatives. Thirty-two representatives are chosen at random from the sales force and randomly assigned to the four call plans (eight per plan). The representatives follow their plans for 6 months, and their sales for the 6-month study period are recorded. These data are listed in the file **P19_01.xlsx**.

a. Do the sample data support the hypothesis that at least one of the call plans helps produce a higher average level of sales? Perform an appropriate statistical test and report a *p*-value.

b. If the sample data indicate the existence of mean sales differences across the call plans, which plans produce significantly different average sales levels at the 95% level?

2. Consider a large chain of supermarkets that sell their own brand of potato chips in addition to many other name brands. Management would like to know whether the type of display used for the store brand has any effect on sales. There are four types of displays being considered, so management decides to choose 24 similar stores to serve as experimental units. A random six of these are instructed to use display type 1, another random six are instructed to use display type 2, a third random six are instructed to use display type 3, and the final six stores are instructed to use display type 4. For a period of one month, each store keeps track of the *fraction* of total potato chips sales that are of the store brand. The data for the 24 stores are listed in the file **P19_02.xlsx**. Note that one of the stores using display 3 is blank. This store did not follow instructions properly, so its observation is disregarded.

a. Why do you think each store keeps track of the fraction of total potato chips sales that are of the store brand? Why do they not simply record the total amount of sales of the store brand potato chips?

b. Do the data suggest different mean proportions of store brand sales at the 10% significance level? If so, construct 90% confidence intervals for the differences between all pairs of mean proportions to identify which of the display types are associated with higher fractions of sales.

3. National Airlines recently introduced a daily (i.e., early morning) nonstop flight between Houston and Chicago. The vice president of marketing for National Airlines decided to perform a statistical test to see whether National's average passenger load on this new flight was different from that of each of its two major competitors (which we will call competitor 1 and competitor 2). Ten early-morning flights were selected at random from each of the three airlines and the percent of *unfilled* seats on each flight was recorded. These data are listed in the file **P19_03.xlsx**.

a. Is there evidence that National's average passenger load on the new flight is different from that of its two competitors? Report a *p*-value and interpret the results of the statistical test.

b. Select an appropriate significance level and construct confidence intervals for all pairs of differences between means. Which of these differences, if any, are statistically significant at the selected significance level?

4. A hotel manager would like to know whether customers who pay by different methods have different-sized bills. She divides all customers into four categories: those who pay by check or cash, those who pay with a VISA or MasterCard, those who pay with an American Express card, and those who use some other type of credit card. She then collects data on daily bills, which are listed in the file **P19_04.xlsx**. Note that these bills contain the room charge, plus any other charges to the customer's account.

a. Test whether the different categories of customers have different-sized bills at the 10% significance level.

b. Based on 90% confidence intervals for all pairs of differences between means, which of these differences, if any, are significantly nonzero at the 10% significance level?

5. A company sells identical soap in four different packages at the same price. The sales of each package type for 12 months are listed in the file **P19_05.xlsx**. Is there any indication of differences in the mean sales of this brand of soap across the various package types? Perform an appropriate statistical test and report a *p*-value.

Level B

6. Do graduates of undergraduate business programs with different majors tend to earn disparate average starting salaries? Consider the data listed in the file **P19_06.xlsx**.

a. Is there any reason to doubt the equal-variance assumption made in the one-way ANOVA model in this particular case? Support your response to this question.

b. Assuming that the variances of the four underlying populations are indeed equal, can you reject at the 10% significance level that the mean starting salary

is the same for each of the given business majors? Explain why or why not.

c. Based on 90% confidence intervals for all pairs of differences between means, which of these differences, if any, are statistically significant at the 10% significance level?

7. A company that employs a large number of salespeople is interested in determining which of the following subsets of the sales staff sells, on average, the most: (1) those whose compensation consists of a fixed salary, (2) those whose compensation is based strictly on commission, and (3) those whose compensation is based on a smaller fixed portion *and* a reduced commission rate. Sales data (in dollars) from the previous quarter are collected for randomly selected salespeople who are compensated according to one of the three aforementioned schemes. The data are listed in the file **P19_07.xlsx**.

a. Is there any reason to doubt the equal-variance assumption made in the one-way ANOVA model in this particular case? Support your response to this question.

b. Can you reject at the 5% significance level that the mean sales are the same for each of the three groups of salespeople? Explain why or why not.

c. Based on 95% confidence intervals for all pairs of differences between means, which of these differences, if any, are statistically significant at the 5% significance level?

19-3 USING REGRESSION TO PERFORM ANOVA

The method we discussed in the previous section for performing ANOVA—calculating sums of squares by rather complex formulas and showing the results in an ANOVA table—is the traditional way of implementing ANOVA. Indeed, it is the method implemented in most statistical software packages, and it can be extended to many experimental designs besides one-way ANOVA. However, it is worth knowing that most of the same ANOVA results can be obtained by multiple regression analysis, as we will briefly discuss in this section. The advantage of using regression is that many people understand regression better than the formulas used in traditional ANOVA. The disadvantage is that some of the traditional ANOVA output, such as the confidence intervals for mean differences, can be obtained with regression only with some difficulty—they are not standard parts of the regression output. Therefore, regression is not a perfect substitute for traditional ANOVA, but it can supplement the analysis.

To perform ANOVA with regression, we run a regression with the same dependent variable as in ANOVA and use dummy variables for the treatment levels as the *only* explanatory variables. For example, if there is a single factor with 5 treatment levels, we create 4 dummy variables, one for each of the treatment levels except a designated "reference" level, and we run the regression with these 4 dummies as the only explanatory variables. In the resulting regression output, the ANOVA table will be *exactly* the same as the ANOVA table we obtain from traditional ANOVA, and the coefficients of the dummy variables will be estimates of the mean differences between the corresponding treatment levels and the reference level.

For example, if there are 5 treatment levels and level 5 is designated as the reference level, then the regression coefficients will estimate the mean differences $\mu_1 - \mu_5$, $\mu_2 - \mu_5$, $\mu_3 - \mu_5$, and $\mu_4 - \mu_5$. Therefore, the reported confidence intervals for these coefficients are really confidence intervals for these mean differences. However, we do not automatically obtain confidence intervals for other mean differences such as $\mu_2 - \mu_3$. Also, the confidence intervals we obtain are not of the "Tukey" type we obtained with ANOVA. (They are of the "no correction" type we will discuss in the next section.) On the plus side, however, the regression output provides an R^2 value, the percentage of the variation of the dependent variable explained by the various treatment levels of the single factor. This R^2 value is *not* part of the traditional ANOVA output.

To see how this works, we revisit Midway's cereal experiment from Example 19.1.

EXAMPLE 19.1 THE EFFECT OF SHELF HEIGHT ON CEREAL SALES (CONTINUED)

Recall that the Midway supermarket chain ran a study on 125 stores to see whether shelf height, set at five different levels, has any effect on sales of a popular brand of cereal. (See the file **Cereal Sales.xlsx**.) Does Midway get the same results as before if it analyzes the data with regression?

Objective To see how Midway can analyze its data with regression, using only dummy variables for the treatment levels.

Solution

Before we can run a regression, we must first reorganize the data. Recall that the original data in the file are in unstacked form—one sales column for each shelf height. For regression, the data must be in stacked form. This is easy to accomplish with StatTools. First, select Stack from the Data Utilities group in StatTools. In the resulting dialog box (not shown), check all five variables, and specify Shelf Height as the Category Name and Sales as the Value Name. This creates a new worksheet with two long variables called Shelf Height and Sales. Next, create a new StatTools data set for the stacked data, and then use StatTools to create dummies for the different shelf heights, based on the Shelf Height variable. The results for a few of the stores appear in Figure 19.13.

Figure 19.13 Stacked Variables and Dummy Variables

	A	B	C	D	E	F	G
1	Shelf Height	Sales	Highest	Lowest	Middle	Next-to-highest	Next-to-lowest
2	Lowest	340	0	1	0	0	0
3	Lowest	376	0	1	0	0	0
4	Lowest	378	0	1	0	0	0
5	Lowest	371	0	1	0	0	0
6	Lowest	395	0	1	0	0	0
124	Highest	461	1	0	0	0	0
125	Highest	375	1	0	0	0	0
126	Highest	399	1	0	0	0	0

We now run a multiple regression with the Sales variable as the dependent variable and the Shelf Height dummies as the explanatory variables. We used Lowest as the reference level, although any level could have been used. The regression output is shown in Figure 19.14.

The first thing to notice is that the ANOVA table from the regression output is *identical* to the ANOVA table from traditional ANOVA. (See Figure 19.5.) This will always be the case. You can infer, because of the extremely low *p*-value in this table, that the population regression coefficients are not all 0. However, because these regression coefficients are really mean differences between the various levels and the reference level, you can infer that these mean differences are not all 0. Specifically, at least one of the upper heights differs from the lowest height. The estimates of the mean differences, given in the range B20:B23, are the observed average differences in sales between upper heights and the lowest height. Also, the constant in cell B19 is the observed average sales for the lowest height.

If you compare the confidence intervals in the range F20:G23 of the regression output to the corresponding confidence intervals for the ANOVA output in Figure 19.5, you will see that they are somewhat different. For example, the confidence interval for $\mu_2 - \mu_1$

Figure 19.14 Regression Output for Cereal Example

	A	B	C	D	E	F	G
7	*Multiple Regression for Sales*						
8		Multiple		Adjusted	StErr of		
9	*Summary*	R	R-Square	R-Square	Estimate		
10		0.3640	0.1325	0.1036	75.62534408		
11							
12		Degrees of	Sum of	Mean of			
13	*ANOVA Table*	Freedom	Squares	Squares	*F*-Ratio	*p*-Value	
14	Explained	4	104807.68	26201.92	4.5814	0.0018	
15	Unexplained	120	686303.12	5719.192667			
16							
17			Standard			Confidence Interval 95%	
18	*Regression Table*	Coefficient	Error	*t*-Value	*p*-Value	Lower	Upper
19	Constant	334.92	15.12506882	22.1434	<0.0001	304.9734164	364.8665836
20	Highest	48.96	21.39007745	2.2889	0.0238	6.609135289	91.31086471
21	Middle	48.52	21.39007745	2.2683	0.0251	6.169135289	90.87086471
22	Next-to-highest	91.36	21.39007745	4.2711	<0.0001	49.00913529	133.7108647
23	Next-to-lowest	43.76	21.39007745	2.0458	0.0430	1.409135289	86.11086471

from Figure 19.14 extends from 1.41 to 86.11, whereas the similar confidence interval in Figure 19.5 extends from −15.50 to 103.02. (We had to reverse the signs to get the confidence interval for $\mu_2 - \mu_1$, not $\mu_1 - \mu_2$.) In particular, the confidence interval from regression, although centered around the same mean difference, is much narrower. In fact, it is entirely positive, leading us to conclude that this mean difference is significant. The ANOVA output led us to the opposite conclusion. The reason for this apparent discrepancy is the subject of the next section. It is basically because the Tukey intervals quoted in the ANOVA output are more "conservative" (wider) and typically lead to *fewer* significant differences.

One final comment about the regression output regards its R^2 value. We see that differences in the shelf height account for 13.25% of the variation in sales. This means that although shelf height has some effect on sales, there is a lot of "random" variation in sales across stores that cannot be accounted for by shelf height. ■

PROBLEMS

Level A

8. For the National Airlines data in Problem 3 (see the file **P19_03.xlsx**), perform a regression analysis using dummy variables for the airlines. Comment on the meaning of the regression output. How does it compare with the ANOVA output from Problem 3? Does it give you any extra insights?

9. For the soap data in Problem 5 (see the file **P19_05.xlsx**), perform a regression analysis using dummy variables for the different packages. Comment on the meaning of the regression output. How does it compare with the ANOVA output from Problem 8? Does it give you any extra insights?

Level B

10. In Problem 6, the salary data for graduates of the undergraduate business programs (see the file **P19_06.xlsx**) represent an unbalanced design—there are more students from some majors than others. Run a regression on starting salaries, using dummies for majors as the explanatory variables. Do you get the same results as with the ANOVA output from Problem 6? Specifically, consider the constant term and the regression coefficients in the output. Is the constant equal to the average starting salary for the reference major? (You can designate any of the majors as the reference major.) Are the regression coefficients equal to average differences between the various majors and the reference major?

19-4 THE MULTIPLE COMPARISON PROBLEM

In many statistical analyses, including ANOVA studies, we want to make statements about *multiple* unknown parameters. For example, in the cereal study (Example 19.1), we wanted to create confidence intervals for differences between each pair of means—10 confidence intervals in all.[3] Any time we make such a statement, there is a chance that we will be wrong; that is, there is a chance that the true population value will not be inside the confidence interval. If we create a 95% confidence interval, say, then the error probability is 0.05. In fact, as we explained in Chapter 8, the endpoints of the confidence interval are chosen so that the error probability will be 1 minus the confidence level chosen. What do we mean by "error probability," however, when we make several statements based on the same data? This is the issue addressed in this section.

We can use simulation to get an idea of the problem. In the file **Multiple Comparison .xlsx**, we simulate a data set very much like those encountered in one-way ANOVA. (See the range A10:H70 of Figure 19.15. Note that a lot of rows have been hidden.) The data in this range correspond to data from a one-factor design with 8 treatment levels and 60 observations per level—480 observations all together. However, we entered the same formula, **=NORMINV(RAND(),0,1)**, in each cell. This means that each treatment level is generating normal data with mean 0 and standard deviation 1. Therefore, we *know* that the equal-means hypothesis of ANOVA is true—all population means are equal to 0. Nevertheless, we calculate 95% confidence intervals for each of the 28 possible differences between means in rows 84–111, using the two-sample procedure described in Chapter 8. If a confidence interval does *not* include 0, we indicate it as a significant difference by putting "Yes" in column D. Then if *at least one* of these 28 confidence intervals is significant, we record "Yes" in cell D113. Finally, we replicate this procedure 500 times in columns J and K, each time recording the value in cell D113, and we report the percentage of replications with "Yes" in cell K5. As you can see, the reported percentage (in cell K5) is close to 50%.

What does this prove? Recall that all of the population means equal 0. This is how we simulated the random numbers in the first place. Therefore, if any one of the 28 confidence intervals in rows 84–111 turns out to be significant and we report it as such, we are making an error. That is, we are reporting that these two population means are not equal, when in fact they both equal 0. Because *none* of the 28 confidence intervals in rows 84–111 should be significant, we will have a perfect record only if we report "No" in cell D113. Of course, a perfect record cannot always be obtained, but by using a 95% confidence level, you might expect a perfect record in 95% of the replications. Unfortunately, the simulation shows that we are not even close to this. We get a perfect record only about half of the time! In statistical terms, if we run each confidence interval at the 95% level, the *overall* confidence level (of having *all* 28 statements correct) is much less than 95%. (Worse yet, we are never really sure how much less.) This is called the **multiple comparison problem**. It says that if we make a lot of statements, each at a given confidence level such as 95%, then the chance of making at least one wrong statement is much greater than 5%.

The question is how to get the overall confidence level equal to the desired value, such as 95%. Or in the simulation in Figure 19.15, how can we get the error rate in cell K5 to be approximately 5%? The answer is that we need to *correct* the individual confidence intervals, so that we do *not* calculate them exactly as described in Chapter 8. Several corrections have been proposed by statisticians, and StatTools includes three of the most popular correction methods in its one-way ANOVA procedure: the **Bonferroni**, **Tukey**, and **Scheffé methods**. (They can be chosen from the dialog box shown in Figure 19.16.) Although the

[3]Note that if a confidence interval for a difference such as $\mu_1 - \mu_2$ is reported, the confidence interval for its opposite, $\mu_2 - \mu_1$, is *not* reported. This is because the latter is simply the negative of the former.

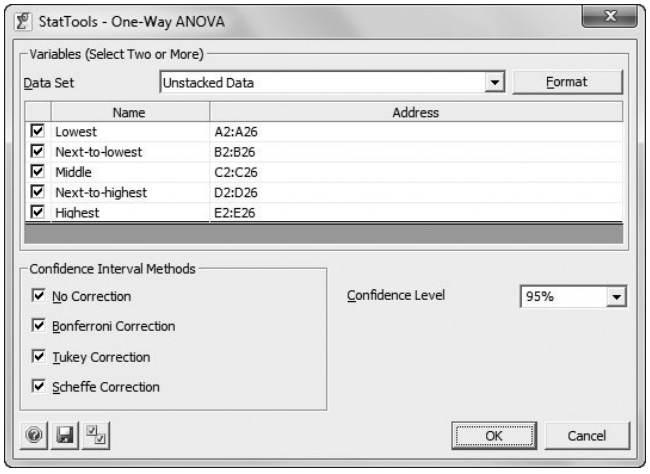
Figure 19.15 Simulation to Illustrate the Multiple Comparison Problem

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Simulation of multiple comparison problem**													
2														
3										**Percent of replications with any significant differences**				
4										(from data table below)				
5											48.2%			
6														
7														
8	**Simulation of data and confidence intervals**									**Data table for whether any differences are significant**				
9														
10	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8		Replication	Any Significant?			
11	0.217689683	1.81343727	0.7533667	0.978887152	−0.57045	0.983571	−0.75971	−1.38261			No			
12	−0.176004652	−0.0254418	0.59737356	1.039012856	2.309349	0.396659	0.637854	−0.33695		1	No			
13	−0.358308438	0.74692562	0.43801209	−0.02990212	0.300742	0.571373	1.083935	0.67425		2	No			
14	−1.238068644	−0.2595726	−0.4507764	0.757778728	−0.06452	−0.16077	0.482804	0.142446		3	Yes			
69	1.067266927	0.00504709	−0.2320311	−1.01662996	−0.92909	1.997633	0.028709	0.840802		58	No			
70	−2.719100304	−0.5515881	0.6977043	−1.00944356	−0.07783	0.232316	0.728025	−0.46184		59	No			
71										60	Yes			
72	Means									61	Yes			
73	−0.037980741	−0.1780704	−0.0541804	−0.09419457	0.023595	0.158603	0.12165	−0.04096		62	No			
74										63	Yes			
75	Stdevs									64	Yes			
76	0.980308296	1.01357647	0.97389301	0.913152569	0.818163	1.06218	0.897967	0.970762		65	Yes			
77										66	No			
78	Pooled stdev	0.95636008								67	Yes			
79	StErr of diff	0.17460666								68	Yes			
80	t-multiple	2.00171748								69	No			
81										70	Yes			
82	95% confidence intervals for differences between means									71	No			
83	Difference	Lower limit	Upper limit	Significant?						72	Yes			
84	1 minus 2	−0.2094235	0.48960289	No						73	No			
85	1 minus 3	−0.3333135	0.36571288	No						74	Yes			
86	1 minus 4	−0.2932994	0.40572704	No						75	No			
109	6 minus 7	−0.3125599	0.38646654	No						98	Yes			
110	6 minus 8	−0.1499552	0.54907127	No						99	Yes			
111	7 minus 8	−0.1869085	0.51211795	No						100	No			
112										101	Yes			
113		Any Significant?		No						102	No			
114										103	Yes			
115										104	Yes			
509										498	No			
510										499	No			
511										500	No			

In this simulation, samples from several normal populations, all with the *same* mean and standard deviation, will be simulated, and the usual 95% confidence interval is constructed for *each* mean difference. Then a data table is used to check the percentage of replications that have at least one confidence interval that does not include 0. It is *much* larger than 5%.

Figure 19.16

One-Way ANOVA Dialog Box with Confidence Intervals Options



details of these methods are beyond the scope of this book, they are all methods for coping with the multiple comparison problem. They differ only in the multiplier they use in the typical confidence interval formula for a difference between means:

$$\overline{Y}_i - \overline{Y}_j \pm \text{multiplier} \times \sqrt{MSW(1/n_i + 1/n_j)}$$

Recall that the multiplier used for the usual "no-correction" method from Chapter 8 is a *t*-value that, for a 95% confidence level, is approximately equal to 2. The correction methods all use multiples that are *larger* than this. The idea is that by using a larger multiplier, we get a wider confidence interval. This decreases the chance that the confidence interval will fail to include the true mean difference, which in turn decreases the chance that at least one of several such confidence intervals will fail to include its true mean difference. The larger the multiplier is, the more conservative the confidence intervals will be (where "conservative" means wider intervals). Scheffé's and Bonferroni's methods tend to be the most conservative, whereas Tukey's method strikes a balance between being too conservative and not conservative enough. It is the method favored by many researchers when the focus is on many confidence intervals for mean differences, as in Example 19.1.

To follow the simulation one step further, the multiplier used in the individual confidence intervals in rows 84–111 of Figure 19.15 is approximately equal to 2, as shown in cell B80. Using appropriate formulas (not presented here), it can be shown that the multipliers for the Tukey, Bonferroni, and Scheffé methods are 3.04, 3.27, and 3.77, respectively. Furthermore, if the Tukey multiplier is used in the simulation, the percentage in cell K5 becomes approximately 5%, exactly what we want.

To see how these correction methods might affect results, we report all four types of confidence intervals for the cereal data of Example 19.1 in Figure 19.17, with the results rearranged slightly to fit better on the printed page. (We reported only the Tukey intervals earlier.) You should note the following. First, the confidence intervals get wider as we move from no correction (from Chapter 8) to Tukey to Bonferroni to Scheffé. Second, all three correction methods report exactly the same significant differences. Specifically, they all report that the only significant difference is between the next-to-highest and lowest shelf heights. The three correction methods do not agree exactly in all data sets, but they usually produce similar results. In contrast, the no-correction method finds 7 of the 10 differences to be significant, a very different result. This is typical. Because this method does not correct for the number of confidence intervals being reported, it tends to find *too many* significant differences.

Figure 19.17 Confidence Intervals from Different Methods

	A	B	C	D	E	F	G	H	I	J
30		Difference	No Correction		Bonferroni		Tukey		Scheffe	
31	*Confidence Interval Tests*	of Means	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
32	Lowest-Next-to-lowest	–43.76	**–86.11086471**	**–1.409135289**	–104.9327306	17.41273061	–103.0190249	15.49902487	–110.6837587	23.16375867
33	Lowest-Middle	–48.52	**–90.87086471**	**–6.169135289**	–109.6927306	12.65273061	–107.7790249	10.73902487	–115.4437587	18.40375867
34	Lowest-Next-to-highest	–91.36	**–133.7108647**	**–49.00913529**	**–152.5327306**	**–30.18726939**	**–150.6190249**	**–32.10097513**	**–158.2837587**	**–24.43624133**
35	Lowest-Highest	–48.96	**–91.31086471**	**–6.609135289**	–110.1327306	12.21273061	–108.2190249	10.29902487	–115.8837587	17.96375867
36	Next-to-lowest-Middle	–4.76	–47.11086471	37.59086471	–65.93273061	56.41273061	–64.01902487	54.49902487	–71.68375867	62.16375867
37	Next-to-lowest-Next-to-highest	–47.60	**–89.95086471**	**–5.249135289**	–108.7727306	13.57273061	–106.8590249	11.65902487	–114.5237587	19.32375867
38	Next-to-lowest-Highest	–5.20	–47.55086471	37.15086471	–66.37273061	55.97273061	–64.45902487	54.05902487	–72.12375867	61.72375867
39	Middle-Next-to-highest	–42.84	**–85.19086471**	**–0.489135289**	–104.0127306	18.33273061	–102.0990249	16.41902487	–109.7637587	24.08375867
40	Middle-Highest	–0.44	–42.79086471	41.91086471	–61.61273061	60.73273061	–59.69902487	58.81902487	–67.36375867	66.48375867
41	Next-to-highest-Highest	42.40	**0.049135289**	**84.75086471**	–18.77273061	103.5727306	–16.85902487	101.6590249	–24.52375867	109.3237587

At this point, it is natural to ask why there are so many methods. The reason has to do with the purpose of the study. A researcher who initiates a study might have a particular interest in a few specific differences. For example, the analyst in Example 19.1 might be particularly interested in the differences between the lowest height and each of the other

four heights. The whole study is intended to study these specific differences. In this case, the differences of interest are called **planned comparisons**. On the other hand, the analyst might initiate the study just to see what differences there are. This analyst will examine all pairwise differences to see which are significant. Here we talk about **unplanned comparisons** because the analyst does not specify which differences to focus on *before* collecting the data.

In the case of planned comparisons, if there are only a few differences of interest, it is usually acceptable to report confidence intervals for these differences using the no-correction method. If there are more than a few planned comparisons (even trained statisticians do not always agree on the interpretation of "a few"), then it is better to report Bonferroni intervals. In the case of unplanned comparisons, the Tukey method is usually the preferred method. It keeps the overall confidence level close to the desired level (such as 95%) without making the intervals overly wide. More important, it keeps the entire study from becoming a "fishing expedition," where a few differences become significant just by the luck of the draw (as occurred in the simulation in Figure 19.15).

The Scheffé method can be used for planned or unplanned comparisons. It tends to produce the widest intervals because it is intended not only for differences between means, such as $\mu_2 - \mu_4$, but also for more general **contrasts**, where a contrast is the difference between weighted averages of means. For example, if the analyst in Example 19.1 is interested in how the lowest height compares to the *average* of the other four heights, then the difference $\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$ would be of interest. Although the analysis of general contrasts such as this is deferred until the next section, we note that Scheffé's method was developed specifically to deal with them. If we are interested only in simple differences like $\mu_2 - \mu_4$, then Tukey's method should be used instead.

PROBLEMS

Level A

11. Consider again the one-way ANOVA hypothesis test described in Problem 1. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 95%. Summarize your results. Recall that the relevant data are listed in the file **P19_01.xlsx**.

12. Consider again the one-way ANOVA hypothesis test described in Problem 2. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 90%. How do these results compare to the uncorrected 90% confidence intervals? Recall that the relevant data are listed in the file **P19_02.xlsx**.

13. Consider again the one-way ANOVA hypothesis test described in Problem 3. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 99%. Compare the widths of the confidence intervals generated with each of these methods with those of uncorrected 99% confidence intervals. Explain your findings. Recall that the relevant data are listed in the file **P19_03.xlsx**.

14. Consider again the one-way ANOVA hypothesis test described in Problem 4. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 90%. Summarize your results. Recall that the relevant data are listed in the file **P19_04.xlsx**.

15. Consider again the one-way ANOVA hypothesis test described in Problem 5. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 99%. How do these results compare to the uncorrected 99% confidence intervals? Recall that the relevant data are given in the file **P19_05.xlsx**.

Level B

16. Consider again the one-way ANOVA hypothesis test described in Problem 6. Suppose that we are interested in comparing the mean starting salary of accounting students with that of each of the other three majors (i.e., marketing, finance, and management). Recall that the relevant data are listed in the file **P19_06.xlsx**.

a. Which method for generating confidence intervals for mean differences is most appropriate in this situation? Explain your choice.

b. Apply the method identified in part **a** to estimate the mean differences of interest. Briefly interpret your findings.

17. Consider again the one-way ANOVA hypothesis test described in Problem 7. Suppose that we are interested in comparing the mean sales achieved by salespeople with varying compensation schemes. In other words, we are interested in making all possible comparisons at this point. Recall that the relevant data are listed in the file **P19_07.xlsx**.

a. Which method for generating confidence intervals for mean differences is most appropriate in this situation? Explain your choice.

b. Apply the method identified in part **a** to estimate the mean differences of interest. Briefly interpret your findings.

19-5 TWO-WAY ANOVA

The examples discussed so far in this chapter have been single-factor designs. There is a single factor, such as shelf height in Example 19.1 or customer size in Example 19.2, that we observe at several levels. The question then is whether the mean of a dependent variable is equal across all levels. In this section we allow two factors, each at several levels. As you will see, some of the ideas from one-way ANOVA carry over to two-way ANOVA. However, there are differences in the data setup, the analysis itself, and, perhaps most important, the types of questions we ask. Because an abstract discussion of two-way ANOVA can be difficult to follow, we immediately introduce Example 19.3.

EXAMPLE 19.3 DRIVING DISTANCES FOR GOLF BALL BRANDS

If you are a golfer, or even if you have ever seen golf ball commercials on television, you know that a number of golf ball manufacturers claim to have the "longest ball," that is, the ball that goes the farthest on drives. This example illustrates how these claims might be tested. We assume that there are five major brands, labeled A through E. A consumer testing service runs an experiment where 60 balls of each brand are driven under three temperature conditions. The first 20 are driven in cool weather (about 40 degrees), the next 20 are driven in mild weather (about 65 degrees), and the last 20 are driven in warm weather (about 90 degrees). The goal is to see whether some brands differ significantly, on average, from other brands and what effect temperature has on mean differences between brands. For example, it is possible that brand A is the longest ball in warm weather but some other brand is longest in cool temperatures.

Objective To use two-way ANOVA to analyze the effects of golf ball brands and temperature on driving distances.

Solution

This example represents a controlled experiment. The consumer testing service decides exactly how to run the experiment, namely, by assigning 20 randomly chosen balls of each brand to each of three temperature levels. In our general terminology, the experimental units are the individual golf balls and the dependent variable is the length (in yards) of each drive. There are two factors: brand and temperature. The brand factor has five treatment levels, A through E, and temperature has three levels, cool, mild, and warm. The design is balanced because the same number of balls, 20, is used at each of the $5 \times 3 = 15$ treatment level combinations. In fact, balanced designs are the only two-way designs we will discuss in this book. (The analysis of unbalanced designs is more complex and is best left to a more advanced book.) There is one further piece of terminology. We call this a **full factorial** two-way design because we test golf balls at *each* of the 15 possible treatment level combinations. If, for example, we decided not to test any brand A balls at a

temperature of 65 degrees, then the resulting experiment would be called an **incomplete** design. We will discuss incomplete designs briefly in the next section—and why they are sometimes used—but full factorial designs are preferred whenever possible.

In a **full factorial design**, we assign experimental units to *each* treatment level combination. In an **incomplete design**, we assign experimental units to some of the treatment level combinations but not to all of them.

How should the consumer testing service actually carry out the experiment? One possibility is to have 15 golfers, each of approximately the same skill level, hit 20 balls each. Golfer 1 could hit 20 brand A balls in cool weather, golfer 2 could hit 20 brand B balls in cool weather, and so on. You can probably see the downside of this design. Brand A might come out the longest ball just because the golfers assigned to brand A have good days. Therefore, if the consumer testing company decides to use human golfers, it should spread them evenly among brands and weather conditions. For example, it could employ 10 golfers to hit two balls of each brand in each of the weather conditions. Even here, however, the use of different golfers introduces an unwanted source of variation: the different abilities of the golfers (or how well they happen to be driving that day). Is the solution, then, to use a *single* golfer for all 300 balls? This has its own downside—namely, that the golfer might get tired in the process of hitting this many balls. Even if he hits the brands in random order, the fatigue factor could play a role in the results.

These are the types of things designers of experiments must consider. They must attempt to eliminate as many unwanted sources of variation as possible, so that any differences across the factor levels of interest can be attributed to these factors and not to extraneous factors. In this example, we suspect that the best option for the consumer testing service is to employ a "mechanical" golf ball driving machine to hit all 300 balls. This should reduce the inevitable random variation that would occur by using human golfers. Still, there will be some random variation. Even a mechanical device, hitting the same brand under the same weather conditions, will not hit every drive exactly the same length.

Data for StatTool's Two-Way ANOVA procedure cannot be in unstacked form. Also, a balanced design must be used.

Once the details of the experiment have been decided and the golf balls have been hit, we will have 300 observations (yardages) at various conditions. The usual way to enter the data in Excel®—and the *only* way the StatTools Two-Way ANOVA procedure will accept it—is in the stacked form shown in Figure 19.18. (See the file **Golf Ball.xlsx**.) There must be two categorical variables that represent the levels of the two factors (Brand and Temperature) and a measurement variable that represents the dependent variable (Yards). Although many rows are hidden in this figure, there are actually 300 rows of data, 20 for

Figure 19.18
Data for Golf Ball Example

	A	B	C
1	Brand	Temperature	Yards
2	A	Cool	214.3
3	A	Cool	208.0
4	A	Cool	208.8
5	A	Cool	216.7
6	A	Cool	212.1
7	A	Cool	219.2
8	A	Cool	220.6
9	A	Cool	229.1
10	A	Cool	204.0
11	A	Cool	215.3

each of the 15 combinations of Brand and Temperature. Again, this is a *balanced* design, which is what StatTools expects for its two-way ANOVA procedure. (StatTools will issue an error message if it finds an unbalanced design, that is, unequal numbers of observations at the various treatment level combinations.)

Now that we have the data, what can we learn from them? In fact, which questions should we ask? Here it helps to look at a table of sample means, such as in Figure 19.19. (This table is part of the output from the StatTools Two-Way ANOVA procedure. Alternatively, it can be obtained easily with an Excel pivot table, as we did here.[4]) Prompted by this table, here are some questions we might ask.

- Looking at column E, do any brands average significantly more yards than any others (where these averages are averages over all temperatures)?
- Looking at row 10, do average yardages differ significantly across temperatures (where these averages are across all brands)?
- Looking at the *columns* in the range B5:D9, do differences among averages of brands depend on temperature? For example, does one brand dominate in cool weather and another in warm weather?
- Looking at the *rows* in the range B5:D9, do differences among averages of temperatures depend on brand? For example, are some brands very sensitive to changes in temperature, while others are not?

Figure 19.19
Table of Sample Means in Golf Ball Example

	A	B	C	D	E
3	**Average of Yards**	**Temperature**			
4	**Brand**	**Cool**	**Mild**	**Warm**	**Grand Total**
5	A	218.8	236.5	258.4	237.9
6	B	224.1	245.1	258.3	242.5
7	C	228.0	242.7	263.0	244.6
8	D	215.0	237.6	256.1	236.2
9	E	224.8	255.7	270.9	250.5
10	**Grand Total**	**222.1**	**243.5**	**261.4**	**242.3**

It is useful to characterize the type of information these questions are seeking. Question 1 is asking about the **main effect** of the brand factor. If we ignore the temperature (by averaging over the various levels of it), do some brands tend to go farther than some others? This is obviously a key question for the study. Question 2 is also asking about a main effect, the main effect of the temperature factor. If we ignore the brand (by averaging over all brands), do balls tend to go farther in some temperatures than others? The answer to this question is obvious to golfers. They all know that balls compress better, and hence go farther, in warm temperatures than in cool temperatures. Therefore, this is not a key question for the study, although we would certainly expect the study to confirm what experience tells us.

Main effects indicate whether there are different means for treatment levels of one factor when averaged over the levels of the other factor.

Questions 3 and 4 are asking about **interactions** between the two factors. These interactions are often the most interesting results of a two-way study. Essentially, interactions (if there are any) provide information that could not be guessed by knowing the main

[4]Note that the default label Excel uses in cells L10 and P4 is Grand Total—and you cannot change them. However, they are really "grand averages."

effects alone. In this example, interactions are patterns of the averages in the range B5:D9 that could not be guessed by looking only at the "main effect" averages in column E and row 10. Specifically, the order of brands in column E, from largest to smallest average yardages, is E, C, B, A, D. If there were no interactions at all, this ordering would hold at each temperature. For these data, it is close. At the cool temperature, the ordering is C, E, B, A, D; for mild, it is E, B, C, D, A; for warm, it is E, C, A, B, D. Actually, having no interactions implies even more than the preservation of these rankings. It implies that the difference between any two brands' averages is the same at any of the three temperature levels. For example, the differences between brands E and D at the three temperatures are 224.8 − 215.0 = 9.8, 255.7 − 237.6 = 18.1, and 270.9 − 256.1 = 14.8. If there were no interactions at all, these three differences would be equal.

Interactions indicate patterns of differences in means that could not be guessed from the main effects alone. They exist when the effect of one factor on the dependent variable depends on the level of the other factor.

Remember that the interactions become stronger as the lines in either of these graphs become more nonparallel.

The concept of interaction is much easier to understand by looking at graphs. The graphs in Figures 19.20 and 19.21, which are both outputs from the StatTools Two-Way ANOVA procedure, represent two ways of looking at the pattern of averages for different combinations of brand and temperature—that is, the averages in the range B5:D9 of Figure 19.19. The first of these shows a line for each brand, where each point on the line corresponds to a different temperature. The second shows the same information with the roles of brand and temperature reversed. Neither graph is "better" than the other; they simply show the same data from different perspectives. The key to either is whether the lines are *parallel*. If they are, then there are no interactions—the effect of one factor on average yardage is the *same* regardless of the level of the other factor. The more nonparallel they are, however, the stronger the interactions are. The lines in either of these graphs are not exactly parallel, but they are nearly so. This implies that there is very little interaction between brand and temperature in these data.

Figure 19.20
One View of Interactions in Golf Ball Example

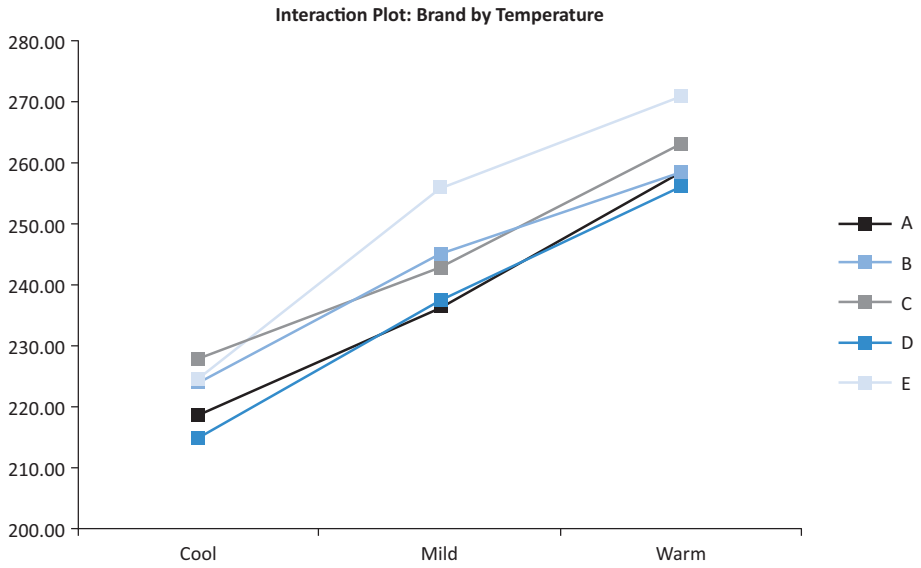
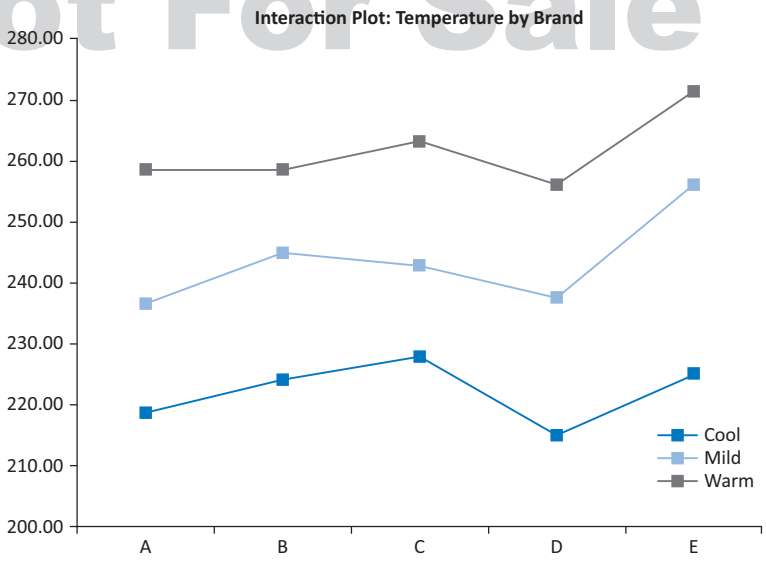


Figure 19.21
Another View of Interactions in Golf Ball Example



In general, interactions can be of several types. We show two contrasting types in Figures 19.22 and 19.23. (For simplification, these focus on two brands only. They are based on *different* data from those used in the **Golf Ball.xlsx** file.) In Figure 19.22, brand A dominates at all temperatures. However, there is an interaction because the difference between brands increases as temperature increases. In this situation the interaction effect is interesting, but the main effect of brand—brand A is better when averaged over all temperatures—is also interesting. The situation is quite different in Figure 19.23, where there is a "crossover." Brand A is somewhat better at cool temperatures, but brand B is better at mild and warm temperatures. In this case the interaction is the most interesting finding, and the main effect of brand is much less interesting. In simple terms, if you are a golfer, you would buy brand A in cool temperatures and brand B otherwise, and you wouldn't care very much which brand is better when averaged over *all* temperatures.

Figure 19.22
One Possible Pattern of Interactions

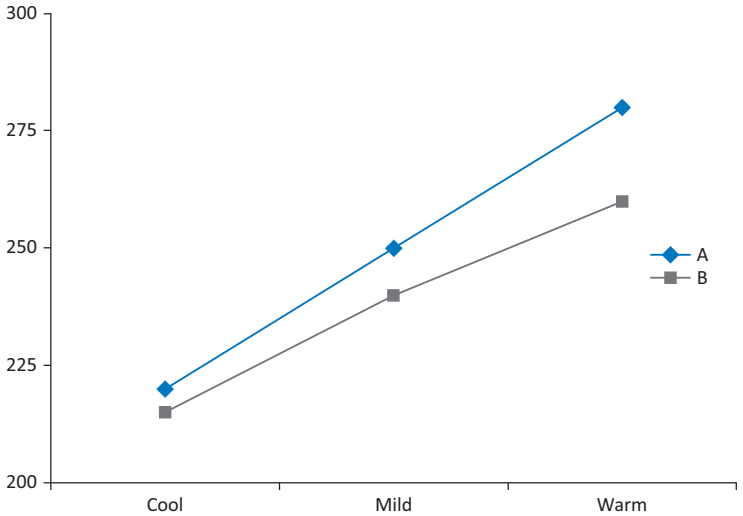
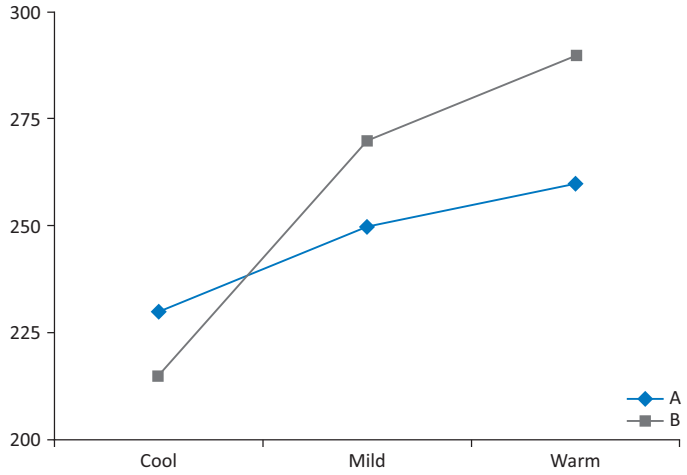


Figure 19.23
Another Possible Pattern of Interactions



Main effects are typically less important when interactions exist. Therefore, it is common to check for significant interactions first.

For these reasons, we check *first* for interactions in a two-way design. If there are significant interactions, then the main effects might not be as interesting. However, if there are no significant interactions, then main effects generally become more important.

Summing up what we have seen so far, main effects are differences in averages across the levels on one factor, where these averages are averages over *all* levels of the other factor. In a table of sample means, such as in Figure 19.19, we can check for main effects by looking at the averages in the "Grand Total" column and row. In contrast, the interactions are patterns of averages in the main body of the table and are best shown graphically, as in Figures 19.20 and 19.21. They indicate whether the effect of one factor depends on the level of the other factor.

The next question is whether the main effects and interactions we see in a table of sample means are statistically significant. As in one-way ANOVA, this is answered by an ANOVA table. However, instead of having just two sources of variation, within and between, as in one-way ANOVA, there are now four sources of variation: one for the main effect of each factor, one for interactions, and one for variation within treatment level combinations. For the golf ball data, two-way ANOVA separates the total variation across all 300 observations into four sources. First, there is variation due to different brands producing different average yardages. Second, there is variation due to different average yardages at different temperatures. Third, there is variation due to the interactions we saw in the interaction graphs. Finally, there is the same type of "within" variation as in one-way ANOVA. This is the variation that occurs because yardages for the 20 balls of the same brand hit at the same temperature are not all identical. (This within variation is usually called the "error" variation in statistical software packages.)

Two-way ANOVA collects this information about the different sources of variation, using fairly complex formulas, in an ANOVA table as shown in Figure 19.24. [This is the output from StatTools, by selecting Two-Way ANOVA from the Statistical Inference group in StatTools, selecting Brand and Temperature as the categorical (C1 and C2) variables and Yards as the measurement (Val) variable. The output includes tables of sample sizes, sample means, and sample standard deviations, as well as the ANOVA table.] The four sources of variation appear in rows 37–40. Rows 37 and 38 are for the main effects of brand and temperature, row 39 is for interactions, and row 40 is for the within (error) variation. Each source has a sum of squares and a degrees of freedom. Also, each has a mean square, the ratio of the sum of squares to the degrees of freedom. Finally, the first three sources have an F-ratio and an associated p-value, where each F-ratio is the ratio of the mean square in that row to the mean square error in cell D40.

Figure 19.24
StatTools Two-Way ANOVA Output for Golf Ball Data

	A	B	C	D	E	F
7	***Two-Way ANOVA for Yards by Brand and Temperature***					
8	***ANOVA Sample sizes***	**Cool**	**Mild**	**Warm**	**Totals**	
9	A	20	20	20	60	
10	B	20	20	20	60	
11	C	20	20	20	60	
12	D	20	20	20	60	
13	E	20	20	20	60	
14	Totals	100	100	100		
15	Balanced	TRUE				
16						
17						
18	***ANOVA Sample Means***	**Cool**	**Mild**	**Warm**	**Totals**	
19	A	218.82	236.45	258.44	237.90	
20	B	224.15	245.13	258.27	242.52	
21	C	228.00	242.72	263.04	244.58	
22	D	215.00	237.62	256.11	236.24	
23	E	224.79	255.75	270.94	250.49	
24	Totals	222.15	243.53	261.36		
25						
26						
27	***ANOVA Sample Std Dev***	**Cool**	**Mild**	**Warm**	**Totals**	
28	A	10.90	8.83	11.01	19.22	
29	B	11.70	9.80	8.93	17.36	
30	C	10.85	14.25	7.08	18.15	
31	D	13.64	10.18	12.13	20.69	
32	E	10.67	10.96	9.05	21.84	
33	Totals	12.28	12.78	10.98		
34						
35		**Sum of**	**Degrees of**	**Mean**	***F*-Ratio**	***p*-Value**
36	***TwoWay ANOVA Table***	**Squares**	**Freedom**	**Squares**		
37	Brand	7702.44	4	1925.61	16.47	<0.0001
38	Temperature	77086.00	2	38543.00	329.58	<0.0001
39	Interaction	1999.97	8	250.00	2.14	0.0325
40	Error	33329.13	285	116.94		
41	Total	120117.53	299			

We test whether main effects or interactions are statistically significant in the usual way—by examining *p*-values. Specifically, we claim statistical significance if the corresponding *p*-value is sufficiently small, less than 0.05, say. Looking first at the interactions, the *p*-value is about 0.03, which says that the lines in the interaction graphs are significantly nonparallel, at least at the 5% significance level. We might dispute whether this nonparallelism is *practically* significant, but there is statistical evidence that at least some interaction between brand and temperature exists. The two *p*-values for the main effects in cells F37 and F38 are practically 0, meaning that there *are* differences across brands and across temperatures. Of course, the main effect of temperature was a foregone conclusion—we already knew that balls do not go as far in cold temperatures—but the main effect of brand is more interesting. According to the evidence, some brands definitely go farther, on average, than some of the others. ■

ANOVA Tip: Unbalanced Designs

StatTools can handle only balanced *two-way designs. Unbalanced designs, where sample sizes are not equal across treatment level combinations, are mathematically more difficult to analyze. The best way to do this is with regression with dummy variables, similar to the method discussed in Section 19-3. You can find an excellent discussion of this issue in Chapters 23 and 24 of the textbook by Kutner et al. (2005).*

19-5a Confidence Intervals for Contrasts

If we find that main effects and/or interactions are significant, then we will probably want to check which factor levels, or factor level combinations, produce significantly larger means than others. Recall that the StatTools One-Way ANOVA procedure provides confidence intervals for differences between each pair of means. This same option is not provided in two-way ANOVA because there would typically be too much output to digest, and much of it would probably not be very useful. Given the purposes of any particular study, there are usually a few comparisons you would like to make, and this can be done fairly easily after the StatTools Two-Way ANOVA procedure has been run. We illustrate the methods in this section.

First, recall that a **contrast** is any difference between weighted averages of means. An example of a simple contrast is the difference between two means, such as $\mu_3 - \mu_1$. You would study this contrast if you were interested in whether μ_3 is different from μ_1. An example of a more complex contrast is $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4 + \mu_5)/3$. You would study this contrast if you were interested in whether the average of μ_1 and μ_2 is different from the average of μ_3, μ_4, and μ_5. Note that the coefficients of these contrasts sum to 0. For example, $(\frac{1}{2} + \frac{1}{2}) - (\frac{1}{3} + \frac{1}{3} + \frac{1}{3}) = 0$. All contrasts have this property. Obviously, many contrasts could be constructed. The ones you construct for any particular study depend entirely on what you are interested in. You might be interested in several simple contrasts or one or two more complex contrasts.

A **contrast** is any linear combination of means (sum of coefficients multiplied by means) such that the sum of the coefficients is 0. It is typically used to compare one weighted average of means to another.

The StatTools Two-Way ANOVA procedure finds MSW for this formula. However, you must calculate the other ingredients with Excel formulas.

Once StatTools has been used to run a two-way ANOVA, you can then form confidence intervals for any contrasts of interest. The general form of the confidence interval is given by Expression (19.5). Here, the point estimate of the contrast is formed by substituting sample means for the μ's in the contrast, *MSW* is the mean square error from the ANOVA table, n_j is the sample size corresponding to any particular mean in the contrast, c_j is the coefficient of the corresponding μ_j in the contrast, and the summation is over all terms in the contrast.

Confidence Interval for Contrast

$$\text{Point estimate of contrast} \pm \text{multiplier} \times \sqrt{MSW \sum_j c_j^2/n_j} \qquad \textbf{(19.5)}$$

As an example, if the contrast is a simple difference between means such as $\mu_1 - \mu_4$, then the point estimate is $\bar{Y}_1 - \bar{Y}_4$, and the c_j's are $c_1 = 1$ and $c_4 = -1$, so that $c_1^2 = 1$ and $c_4^2 = 1$. Therefore, the confidence interval becomes

$$\bar{Y}_1 - \bar{Y}_4 \pm \text{multiplier} \times \sqrt{MSW(1/n_1 + 1/n_4)}$$

which is the same as the formula given in Sections 19-2 and 19-4 for one-way ANOVA.

As in one-way ANOVA, the multiplier in the confidence interval can be chosen in several ways to handle the multiple comparison problem appropriately. We indicate typical possibilities in the following continuation of the golf ball example.

EXAMPLE 19.3 DRIVING DISTANCES FOR GOLF BALL BRANDS (CONTINUED)

One golf ball retail shop would like to test the claims that (1) brand C beats the average of the other four brands in cool weather and (2) brand E beats the average of the other four brands when it is not cool. Are these two claims supported by the data in the **Golf Ball.xlsx** file?

Objective To form and test contrasts for the golf ball data, and to interpret the results.

Solution

Let $\mu_{C,W}$ be the mean yardage for brand C balls hit in warm weather, and define similar means for the other brands and temperatures. Then the first claim concerns the contrast

$$\mu_{C,C} - \frac{\mu_{A,C} + \mu_{B,C} + \mu_{D,C} + \mu_{E,C}}{4}$$

and the second concerns the contrast

$$\frac{\mu_{E,M} + \mu_{E,W}}{2} - \frac{\frac{\mu_{A,M} + \mu_{A,W}}{2} + \frac{\mu_{B,M} + \mu_{B,W}}{2} + \frac{\mu_{C,M} + \mu_{C,W}}{2} + \frac{\mu_{D,M} + \mu_{D,W}}{2}}{4}$$
$$= \frac{\mu_{E,M} + \mu_{E,W}}{2} - \frac{\mu_{A,M} + \mu_{A,W} + \mu_{B,M} + \mu_{B,W} + \mu_{C,M} + \mu_{C,W} + \mu_{D,M} + \mu_{D,W}}{8}$$

(Note in this second contrast how we average over the mild and warm temperatures. This is because the second claim just specifies "not cool.") A good way to handle the calculations in Excel is illustrated in Figure 19.25. For either contrast, you first record the coefficients of the means in the contrast. These appear in the ranges H13:J17 and H26:J30. (Note that the sum of the values in each of these ranges is 0. This is required for contrasts, as we discussed previously.) Then the point estimate of a contrast is the SUMPRODUCT of the sample means and these coefficients. For example, you calculate the point estimate of the first contrast in cell H19 with the formula

=SUMPRODUCT(H13:J17,B19:D23)

The multiplier for these confidence intervals is always a thorny issue, but most statisticians agree that if only a small number of confidence intervals are being formed, as in this example, then you can use the usual *t*-value, where the degrees of freedom is the one corresponding to the error (within) variation. Therefore, the multiplier for each contrast is approximately 2, found in cell H18 with the formula

=TINV(1-H8,C40)

Remember that Excel's SUMSQ function sums the squares of the values in a given range.

Then because each sample size is 20 (the value in cell B9), you can find the lower and upper limits of the confidence intervals from Expression (19.5). For example, the confidence interval for the first contrast is found with the formulas

=H19-H18*SQRT(D40*SUMSQ(H13:J17)/B9)

and

=H19+H18*SQRT(D40*SUMSQ(H13:J17)/B9)

in cells H20 and H21.

Figure 19.25
Confidence Intervals for Contrasts in Golf Ball Example

	G	H	I	J	K
7	**Confidence intervals for contrasts**				
8	Confidence level	95%			
9					
10	Comparing brand C against average of others in cool weather				
11	Matrix of coefficients				
12			Cool	Mild	Warm
13		A	–0.25	0	0
14		B	–0.25	0	0
15		C	1	0	0
16		D	–0.25	0	0
17		E	–0.25	0	0
18	Multiplier		1.968		
19	Point estimate		7.31		
20	Lower limit		1.99		
21	Upper limit		12.63		
22					
23	Comparing brand E against average of others in non-cool weather				
24	Matrix of coefficients				
25			Cool	Mild	Warm
26		A	0	–0.125	–0.125
27		B	0	–0.125	–0.125
28		C	0	–0.125	–0.125
29		D	0	–0.125	–0.125
30		E	0	0.5	0.5
31	Multiplier		1.968		
32	Point estimate		13.62		
33	Lower limit		9.86		
34	Upper limit		17.38		

As you can see, both claims are supported. The confidence intervals for the two contrasts extend from 1.99 to 12.63 and from 9.86 to 17.38—all positive yardages. It looks like brand C beats the average of the competition by at least 1.99 yards in cool weather, and brand E beats the average of the competition by at least 9.86 yards in weather that is not cool. ■

If you want to examine a lot of contrasts, then you should use one of the other confidence interval methods discussed in Section 19-4, the two preferred methods being the Bonferroni and Scheffé methods. The only difference is in the multiplier used. For the Bonferroni method, suppose you want to form k confidence intervals. Then rather than using the t-value that has probability α in the tails, you should use the t-value that has probability α/k in the tails. For example, if you want to form $k = 2$ confidence intervals at the 95% confidence level, as above, then $\alpha = 0.05$ and $\alpha/k = 0.025$, so you put probability 0.025 in the tails rather than 0.05. The effect is that each of the two confidence intervals is constructed separately at the 97.5% level. The multiplier in the golf ball example (cell H18 in Figure 19.25) would use the formula

=TINV((1-H8)/2, C40)

which evaluates to 2.253. This larger multiplier would result in slightly wider confidence intervals.

The Scheffé method is the most "conservative" method in the sense that it generally produces the widest confidence intervals. However, the relevant multiplier for this method is rather complex and will not be discussed here.

19-5b Assumptions of Two-Way ANOVA

The assumptions for the two-way ANOVA procedure are basically the same as for one-way ANOVA. If we focus on any particular combination of factor levels, such as brand A golf balls hit in cool weather, then we assume that (1) the distribution of values (yardages) for this combination is normal, and (2) the variance of values at this combination is the same as at any other combination. It is always wise to check for at least gross violations of these assumptions, especially the equal-variance assumption. The StatTools output provides an informal check by providing a table of standard deviations for the factor level combinations. For the golf ball example, this table is shown in Figure 19.26. Obviously, these standard deviations are not all exactly equal, but we would never expect *exact* equality in any real study. Because these standard deviations are of similar magnitude, there is no reason to worry about the equal-variance assumption for these data. Besides, the equal-variance assumption is less important when the design is balanced, as this one is.

Figure 19.26
Checking the Equal-Variance Assumption

	A	B	C	D	E
27	*ANOVA Sample Std Dev*	Cool	Mild	Warm	Totals
28	A	10.90	8.83	11.01	19.22
29	B	11.70	9.80	8.93	17.36
30	C	10.85	14.25	7.08	18.15
31	D	13.64	10.18	12.13	20.69
32	E	10.67	10.96	9.05	21.84
33	Totals	12.28	12.78	10.98	

As we demonstrated in Example 19.2, however, the log transformation is often useful when variances are far from equal across factor level combinations. At least, this transformation is often worth trying. If it works—that is, if it tends to equalize the variances and maybe even make the data more normal—then two-way ANOVA can be carried out on the log-transformed data exactly as we demonstrated in Example 19.2.

PROBLEMS

Level A

18. Suppose a company that sells residential carpet cleaning equipment wants to judge the sales effectiveness of two factors: factor *A*, the type of sales presentation used by its salespeople, and factor *B*, the type of previous experience or training its salespeople have had in selling this type of equipment. Specifically, there are two types of presentations the company wishes to test. These two levels of factor *A* are the "hard-sell" approach, level 1, and the more relaxed "soft-sell" approach, level 2. The company also differentiates among four levels of past experience/training. These levels of factor *B* are labeled 1 through 4 and are defined as follows: (1) no past experience as a salesperson and no formal training in how to be a salesperson, (2) no past experience and some formal training, (3) some past experience and no formal training, and (4) some past experience and some formal training.

To see how presentation and experience/training affect sales, the company runs an experiment with 80 of its recently hired salespeople, 20 of whom fall into each of the four experience/training levels described above. Within each group of 20 salespeople at a given experience/training level, 10 are told to use a hard-sell approach and the other 10 are told to use a soft-sell approach. Of course, the hard-sellers and soft-sellers are instructed very carefully in the types of presentation they are supposed to use. During a 4-month period, the number of sales for each of the 80 salespeople is recorded. The data are listed in the file **P19_18.xlsx**. The company wishes to infer from these data whether the different presentations and experience/training backgrounds cause significant differences in sales.

a. Assess the main effect of the presentation approach factor upon sales.
b. Assess the main effect of the previous experience and training factor upon sales.
c. Do you find evidence of significant interactions between the two factors in this case? Explain.

19. A study is performed on a sample of residential homes to discover whether the size of the monthly heating bill depends on the type of heat or the type of home. In particular, three types of heat are examined: electric, natural gas, and oil. Also, all homes are classified into two types: those on a single level and those with at least two stories. In a single community, ten houses of each type, using each type of heat, are located and their heating bills for February of the past year are observed. These data are listed in the file **P19_19.xlsx**. Assume that the homes in this study are approximately equivalent in terms of overall square footage and level of insulation.
a. Do you find evidence of a significant main effect for the heat type factor? Explain.
b. Do you find evidence of a significant main effect for the home type factor? Explain.
c. Do you find evidence of significant interactions between the two factors? Explain.

20. An automobile dealer would like to know whether the amount of money spent on a new automobile depends on (1) the age of the buyer, and (2) whether the buyer is accompanied by his or her spouse. Data on 60 recent new vehicle purchases, including purchase prices, are listed in the file **P19_20.xlsx**. Test for any significant main effects and interactions at the 10% level, and briefly summarize your findings.

Level B

21. Consider again the two-way ANOVA hypothesis test described in Problem 18. Construct a 95% confidence interval for each possible pairwise difference between means. Interpret your results. Recall that the relevant data are listed in the file **P19_18.xlsx**.

22. Consider again the two-way ANOVA hypothesis test described in Problem 19. Recall that the relevant data are listed in the file **P19_19.xlsx**.
a. A natural gas supplier claims that homes which use gas heat generate an average February heating bill that is *less than* the average February heating bill of all other homes that use electricity or heating oil. Is this claim supported by the given data? Explain.
b. A heating oil supplier claims that homes that use heating oil generate an average February heating bill that is *less than* the average February heating bill of all other homes that use electricity or natural gas. Is this claim supported by the given data? Explain.

23. The file **P19_23.xlsx** lists data for a two-way ANOVA in which each of the two factors has two levels. Note that there are 25 observations for each of the four treatment combinations.
a. Are the assumptions of two-way ANOVA met for these data? If not, do what you can to correct any problem(s).
b. Test for any significant main effects and interactions at the 5% level. Briefly summarize your results.

19-6 MORE ABOUT EXPERIMENTAL DESIGN

The purpose of this chapter is to introduce key ideas and analysis techniques for the most common (and simple) single-factor and multi-factor models. In each of these models, we analyze how a dependent variable varies when one or more factors are varied at several levels. Although the same analysis can be used in observational studies or in designed experiments, we focus now on designed experiments. This is particularly important because many businesses are just now beginning to see the potential of designed experiments for reducing cost, increasing profit, and producing higher-quality items. (Examples of this were provided in the introductory vignette to this chapter.)

We can break up the topic of experimental design into two parts: (1) the actual design of the experiment, and (2) the analysis of the resulting data. In this section we will expand on these, mostly on the design but to some extent on the analysis, just to provide a sense of what is possible. Specifically, we will discuss some key issues in experimental design and some of the more popular designs. However, the discussion in this section is by no means complete. Many books have been written about experimental design and the required statistical analysis [see Berger and Maurer (2002), Schmidt and Launsby (1994), and DeVor et al. (1992), for example, for very readable accounts of the topic], so we can barely scratch the surface.

Experimental design, as opposed to the statistical methods for analyzing the resulting data, has to do with the selection of factors, the choice of the treatment levels, the way experimental units are assigned to the treatment level combinations, and the conditions under which the experiment is run. These decisions must be made *before* the experiment is performed, and they should be made very carefully. Experiments are typically costly and time-consuming, so the experiment should be designed (and performed) in a way that will provide the most useful information possible. Unfortunately, proper experimental design is by no means intuitive. We want the most for our money, but it is usually not clear how to achieve it. Therefore, a whole science of experimental design has developed through the years. We will summarize some of its most important results here.

19-6a Randomization

The purpose of most experiments is to see which of several factors have an effect on a dependent variable. The factors in question are chosen as those that are controllable and are most likely (among all possible factors) to have some effect. Often, however, there are "nuisance" factors that cannot be controlled, at least not directly. If nothing is done about these nuisance factors, they can possibly mask the effect of the "important" factors, so that we do not achieve the desired results. One important method for dealing with such nuisance factors is **randomization**, where we attempt to spread the levels of the nuisance factors randomly to the various levels of the experimental factors. We illustrate this extremely important idea in Example 19.4.

Randomization is the process of randomly assigning experimental units so that nuisance factors are spread uniformly across treatment levels.

EXAMPLE

19.4 Testing for Sharpness in Inkjet Printers

A computer magazine company regularly tests products from different manufacturers for differences in various aspects of quality. For its next issue, it would like to test sharpness of printed images across three popular brands of inkjet printers. It purchases one printer of each brand, prints several pages on each printer, and measures the sharpness of image on a 0–100 scale for each page. A subset of the data and the analysis appear in Figures 19.27 and 19.28. (See the file **Printers.xlsx**.) They indicate that printer A is best on average and C is worst. Why might these results be misleading?

Figure 19.27
Printer Data

	A	B
1	Printer Brand	Sharpness
2	A	92
3	A	85
4	A	88
5	A	92
6	A	85
28	C	85
29	C	76
30	C	86
31	C	87

Figure 19.28 Results from an Experiment Before Randomizing

	A	B	C	D	E	F
7	***One-Way ANOVA for Sharpness by Printer Brand***					
8	***ANOVA Summary***					
9	Total Sample Size	30				
10	Grand Mean	84.767				
11	Pooled Std Dev	3.285				
12	Pooled Variance	10.789				
13	Number of Samples	3				
14	Confidence Level	95.00%				
15						
16		**Sharpness (A)**	**Sharpness (B)**	**Sharpness (C)**		
17	***ANOVA Sample Stats***	**Data Set #1**	**Data Set #1**	**Data Set #1**		
18	Sample Size	10	10	10		
19	Sample Mean	88.800	83.700	81.800		
20	Sample Std Dev	3.011	2.669	4.022		
21	Sample Variance	9.067	7.122	16.178		
22	Pooling Weight	0.3333	0.3333	0.3333		
23						
24		**Sum of**	**Degrees of**	**Mean**		
25	***OneWay ANOVA Table***	**Squares**	**Freedom**	**Squares**	***F*-Ratio**	***p*-Value**
26	Between Variation	262.067	2	131.033	12.145	0.0002
27	Within Variation	291.300	27	10.789		
28	Total Variation	553.367	29			
29						
30		**Difference**	**Tukey**			
31	***Confidence Interval Tests***	**of Means**	**Lower**	**Upper**		
32	Sharpness (A)-Sharpness (B)	5.100	**1.45651367**	**8.74348633**		
33	Sharpness (A)-Sharpness (C)	7.000	**3.35651367**	**10.64348633**		
34	Sharpness (B)-Sharpness (C)	1.900	–1.74348633	5.54348633		

Objective To use randomization of paper types to see whether differences in sharpness are really due to different brands of printers.

Solution

This is a single-factor design, where the single factor, brand of printer, is varied at three levels. Suppose, however, that there is another factor, type of paper, that is not the primary focus of the study but might affect the sharpness of image. For the sake of discussion, suppose further that all type 1 paper is used in printer A, all type 2 paper is used in printer B, and all type 3 paper is used in printer C. Then it is very possible that the apparent effect of printer is really an effect of paper type. Specifically, it is possible that type 1 paper tends to produce the sharpest image, *regardless* of the printer used. We can't know this for sure, but it is certainly possible given our (flawed) experimental design. The solution is to randomize over paper type. For each sheet of paper to be printed by any printer, we randomly select a paper type. This will tend to even out the paper types across the printers. Then if the average sharpness of image from printer A is still higher than the averages from the other two brands, we will have more confidence that this is due to differences in printers, not types of paper. Note that it is *not* necessary to use equal numbers of sheets of each paper type in the experiment. For example, if paper type 1 is the most used paper type by

actual users, then we might use more of it in the experiment. The important point is that no printer is fed a much higher proportion of any paper type than any other printer.

We illustrate how this might be implemented with random numbers in Figure 19.29. Based on actual usage, suppose that approximately 50% of the paper used in the experiment is of type 1, 35% is of type 2, and 15% is of type 3. This information is entered in columns F and G. Then to randomize paper types across printers, we enter random numbers in column B with the RAND() function, enter the formula

=IF(B2<=$G$3,1,IF(B2<=$G$3+$G$4,2,3))

in cell C2, and copy down column C. Of course, Figure 19.29 shows only the experimental design. Now it is up to the company to run the experiment with the printers and paper types shown (one piece of paper per row), measure the sharpness levels, and perform the same statistical analysis as described in Section 19-2. That is, after we randomize and collect the data, the analysis is the usual one-way ANOVA. This time, however, because we have randomized over paper types, we can be more confident that any observed differences across printers are indeed due to the printers themselves and not differences in paper. ■

Figure 19.29
Experimental Design Using Randomization

	A	B	C	D	E	F	G	H
1	Printer Brand	Random Number	Paper Type	Sharpness		Distribution of paper type		
2	A	0.487642251	1			Type	Pct	
3	A	0.721544637	2			1	50%	
4	A	0.404243214	1			2	35%	
5	A	0.647971901	2			3	15%	
6	A	0.752277178	2					
7	A	0.036587599	1					
27	C	0.715408714	2					
28	C	0.598789252	2					
29	C	0.221325535	1					
30	C	0.974606651	3					
31	C	0.00353146	1					

19-6b Blocking

Randomization is one method for eliminating the effects of one or more nuisance factors. Another method is called **blocking**. Like randomization, blocking is extremely important and is used in many applications. Actually, you have already seen perhaps the simplest form of blocking in Chapters 8 and 9 in the discussion of the paired-sample procedure. In a study of differences between pretest and post-test performance scores, for example, each person is defined as a "block." The idea is that pretest and post-test tend to be correlated—some people do well on both, whereas some do poorly on both—so by using a paired-sample procedure, you "block out" the differences among people and are able to focus on the differences between the two tests.

There are many forms of blocking designs, but we will describe only the simplest: the **randomized block** design with a single experimental factor and a single blocking variable. Suppose there are T treatment levels of the single factor and B blocks. Then you use $T \times B$ experimental units and assign T of these to each block. If it is possible (or makes sense), you can also randomize the T experimental units in any block to the T treatment levels. We illustrate the typical setup in Example 19.5.

In a **randomized block design**, the experimental units are divided into several "similar" blocks. Then each experimental unit within a given block is randomly assigned a different treatment level.

EXAMPLE 19.5 The Effect of Soap Dispensers on Soap Sales

SoftSoap Company is introducing a new product into the market: liquid soap for washing hands. Four types of soap dispensers are being considered. SoftSoap has no idea which of these four dispensers will be perceived as the most attractive or easy to use, so it runs an experiment. It chooses eight supermarkets that have traditionally carried SoftSoap products, and it asks each supermarket to stock all four versions of its new product for a 2-week test period. It records the number of items purchased of each type at each store during this period. (See the file **Soap Sales.xlsx**.) How might we describe (and analyze) this experiment?

Objective To use a blocking design with store as the blocking variable to see whether type of dispenser makes a difference in sales of liquid soap.

Solution

As this example illustrates, the blocking variable—and even the decision whether to block—depends entirely on the specific problem.

At first glance, this might look exactly like a one-way design as described in Section 19-2. There is a single factor, dispenser type, varied at four levels, and there are eight observations at each level. For example, we obtain a count of sales for dispenser type 1 at each of eight stores. However, it is very possible that the dependent variable, number of sales, is correlated with store. That is, some stores might sell a lot of *each* dispenser type, whereas others might not sell many of any dispenser type. (For example, stores in areas where there are a lot of manual labor jobs might sell a lot more hand soap than stores in a university area.) Therefore, we treat each store as a block, so that the experimental design appears as in Figures 19.30 and 19.31. Each treatment level (dispenser type) is assigned exactly once to each block (store). As a practical matter, if each dispenser type is stocked on a different shelf in the store, randomization could also be used, where each store is instructed to randomize the order of dispenser types from top shelf to bottom shelf.

You can analyze these data essentially the same way you analyze a two-factor design, that is, with two-way ANOVA. There are two differences, one technical and one of inter-pretation. The technical difference is that because there is only one observation in each combination of treatment level and block, it is impossible to estimate an

Figure 19.30
Soap Sales Data

	A	B	C
1	Store	Dispenser	Sales
2	1	1	68
3	1	2	82
4	1	3	94
5	1	4	72
6	2	1	72
29	7	4	70
30	8	1	65
31	8	2	77
32	8	3	80
33	8	4	81

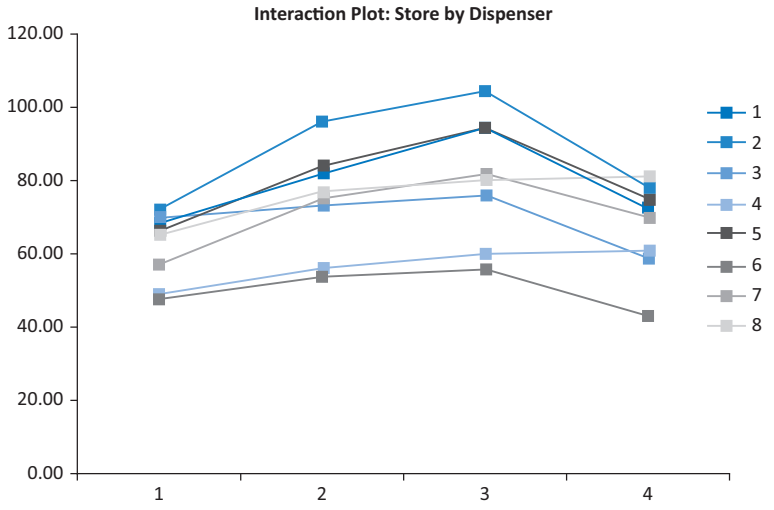
Figure 19.31 Randomized Block Design for Soap Example

	A	B	C	D	E	F
7	***Two-Way ANOVA for Sales by Store and Dispenser***					
8	***ANOVA Sample sizes***	**1**	**2**	**3**	**4**	**Totals**
9	**1**	1	1	1	1	4
10	**2**	1	1	1	1	4
11	**3**	1	1	1	1	4
12	**4**	1	1	1	1	4
13	**5**	1	1	1	1	4
14	**6**	1	1	1	1	4
15	**7**	1	1	1	1	4
16	**8**	1	1	1	1	4
17	**Totals**	8	8	8	8	
18	**Balanced**	TRUE				
19						
20						
21	***ANOVA Sample Means***	**1**	**2**	**3**	**4**	**Totals**
22	**1**	68.00	82.00	94.00	72.00	79.00
23	**2**	72.00	96.00	104.00	78.00	87.50
24	**3**	70.00	73.00	76.00	59.00	69.50
25	**4**	49.00	56.00	60.00	61.00	56.50
26	**5**	66.00	84.00	94.00	75.00	79.75
27	**6**	48.00	54.00	56.00	43.00	50.25
28	**7**	57.00	75.00	81.00	70.00	70.75
29	**8**	65.00	77.00	80.00	81.00	75.75
30	**Totals**	61.88	74.63	80.63	67.38	
31						
32						
33	***ANOVA Sample Std Dev***	**1**	**2**	**3**	**4**	**Totals**
34	**1**	0	0	0	0	11.60
35	**2**	0	0	0	0	15.00
36	**3**	0	0	0	0	7.42
37	**4**	0	0	0	0	5.45
38	**5**	0	0	0	0	12.01
39	**6**	0	0	0	0	5.91
40	**7**	0	0	0	0	10.21
41	**8**	0	0	0	0	7.37
42	**Totals**	9.37	14.04	16.72	12.48	
43						
44		**Sum of**	**Degrees of**	**Mean**	***F*-Ratio**	***p*-Value**
45	***TwoWay ANOVA Table***	**Squares**	**Freedom**	**Squares**		
46	**Store**	4313.50	7	616.21	17.75	<0.0001
47	**Dispenser**	1617.00	3	539.00	15.53	<0.0001
48	**Error**	729.00	21	34.71		
49	**Total**	6659.50	31			

interaction effect and a "within" error variance simultaneously. Therefore, we *assume* there are no important interaction effects between treatment levels and blocks, and we attribute all variation other than that from main effects to error variation. The output is

shown in Figure 19.31. To obtain this output, use the StatTools Two-Way ANOVA procedure, with Store and Dispenser as the two categorical variables. Because there is only one observation per store/dispenser combination, the ANOVA table has no Interaction row. Nevertheless, it still provides interaction charts, one of which appears in Figure 19.32, to check for the no-interaction assumption. If there were no interactions, the lines in this chart (one for each store) would be parallel. Although these lines are not *exactly* parallel, it appears that the effect of dispenser type is approximately the same at each store. Therefore, the no-interaction assumption appears justified.

Figure 19.32
Interaction Chart for Soap Example



There are two F-values and corresponding p-values in the ANOVA table in Figure 19.31. The one in row 47 is for the main effect of dispenser type, whereas the one in row 46 is for the main effect of store. The former is of more interest because this is the focus of the experiment. Its p-value is essentially 0, meaning that there *are* significant differences across dispenser types. In fact, judging by the sample means, the ranking of dispenser types in decreasing order is 3, 2, 4, 1, and there is a considerable gap between each of these. If SoftSoap had to market only one dispenser type, it would almost certainly select type 3. The p-value for the main effect of store is also essentially 0, which means that the stores differ significantly with respect to average sales. This is not as interesting a finding—in fact, we use a block design precisely because we suspect such an effect—but it does confirm that a block design is a good idea.

We can also confirm that blocking was useful by running a one-way ANOVA on the data, using Dispenser as the *single* factor and ignoring Store. The results appear in Figure 19.33. The differences across dispenser type are still significant at the 5% level (the p-value is still less than 0.05), but they are not as significant as when a blocking variable is used. By comparing the ANOVA tables in Figures 19.31 and 19.33, you can see that the error (within) sum of squares in the latter, 5042.5, is split into two parts in the former: the block sum of squares, 4313.5, and the error sum of squares, 729. By having a lower error sum of squares, you obtain a more powerful test for dispenser differences. The point is that when differences across stores are ignored, they tend to mask the differences across dispenser types.

Figure 19.33 Results for Soap Example Using One-Way ANOVA

	A	B	C	D	E	F
7	***One-Way Anova for Sales by Dispenser***					
8	***ANOVA Summary***					
9	**Total Sample Size**	32				
10	**Grand Mean**	71.13				
11	**Pooled Std Dev**	13.42				
12	**Pooled Variance**	180.09				
13	**Number of Samples**	4				
14	**Confidence Level**	95.00%				
15						
16		**Sales (1)**	**Sales (2)**	**Sales (3)**	**Sales (4)**	
17	***ANOVA Sample Stats***	**Data Set #1**	**Data Set #1**	**Data Set #1**	**Data Set #1**	
18	**Sample Size**	8	8	8	8	
19	**Sample Mean**	61.88	74.63	80.63	67.38	
20	**Sample Std Dev**	9.37	14.04	16.72	12.48	
21	**Sample Variance**	87.84	197.13	279.70	155.70	
22	**Pooling Weight**	0.2500	0.2500	0.2500	0.2500	
23						
24		**Sum of**	**Degrees of**	**Mean**		
25	***OneWay ANOVA Table***	**Squares**	**Freedom**	**Squares**	***F*-Ratio**	***p*-Value**
26	**Between Variation**	1617.00	3	539.00	2.99	0.0477
27	**Within Variation**	5042.50	28	180.09		
28	**Total Variation**	6659.50	31			

Blocking is one of the most powerful methods in experimental design. It allows you to "control" for a variable, such as store, that is not of primary interest but could introduce an unwanted source of variation. Experimental designers should always be on the lookout for possible blocking variables. They generally result in more powerful tests. ■

19-6c Incomplete Designs

Recall that the two-factor designs discussed in Section 19-5 are called *full factorial* designs. In a full factorial design you obtain one or more observations for *each* combination of treatment levels. For example, if there are two factors with 5 and 7 treatment levels, respectively, then you replicate the experiment at each of the $5 \times 7 = 35$ treatment level combinations. If there are three factors with 3, 5, and 7 treatment levels, respectively, then you replicate at each of the $3 \times 5 \times 7 = 105$ combinations. By running an experiment in this way, you can estimate all main effects and interactions. A full factorial design is the preferred way to run an experiment from a statistical point of view, but it can be very expensive, even infeasible, if there are more than a few factors.

In industrial settings, there are often a *large* number of input factors that can be varied to produce a product. (Think, for example, of the number of factors that might be varied in an attempt to produce a car door that doesn't rattle.) Each of these factors might have a main effect on some dependent variable of interest, and there might also be important interactions between input factors. The question is how to design an experiment so that you get as much useful information as possible and stay within budget (either time or money). To get an idea of the problem, suppose there are 12 input factors. Even if you use only two treatment levels ("low" and "high") for each factor, there are $2^{12} = 4096$ treatment level combinations in a full factorial design—probably many more than could be tested.

Because this is very common in real applications, statisticians have devised **incomplete**, or **fractional factorial**, designs that test only a fraction of the possible treatment level combinations. Obviously, something is lost by not gaining information on *all* of the possible combinations. Specifically, different effects are **confounded**, which means that they cannot be estimated independently. As an example, the main effect of factor D might be confounded with the three-way interaction effect of factors A, B, and C. In this case it is impossible to tell, because of the design, whether a particular set of observed differences is due to factor D or to the interaction of factors A, B, and C. You would probably conclude that the differences are due to factor D, simply because three-way interaction effects are typically not very important, but you cannot be absolutely sure.

This is a fairly difficult topic, and we will not be able to cover it in much detail. However, just to give you a taste of what is involved, we illustrate a "half-fractional" design with four factors, each at two levels, in Figure 19.34. (See the file **Fractional Design.xlsx**.) If this were a full factorial design, there would be $2^4 = 16$ combinations of treatment levels. The "half-fractional" design means that only half, or eight, of these are used. When using only two levels for each factor, it is customary to label the lower level with a -1 and the higher level with a $+1$. Therefore, each row in the figure represents one of eight combinations of the factor levels. For example, the first row uses the higher level of each factor. (Then when implementing the experiment, several experimental units would be assigned to each combination, so that there would be several observations per row.)

Figure 19.34
A Half-Fractional Design with Four Factors

	A	B	C	D
1	**Half-factorial design with 4 factors**			
2				
3	A	B	C	D
4	1	1	1	1
5	1	1	−1	−1
6	1	−1	1	−1
7	1	−1	−1	1
8	−1	1	1	−1
9	−1	1	−1	1
10	−1	−1	1	1
11	−1	−1	−1	−1

To see how the confounding works, it is useful to create new columns by multiplying the appropriate original A–D columns. For example, the AC column is the product, row by row, of the A and C columns. As in usual algebra, the result is $+1$ if there are an even number of -1's, and -1 if there are an odd number of -1's. The results appear in Figure 19.35. Note that there is now a column for each possible two-way and three-way interaction. If you compare these columns, you will notice that they come in pairs. For example, the A column has exactly the same pattern as the BCD column, the AB column has the same pattern as the CD column, and so on. When two columns are identical, we say that one is the **alias** of the other. The practical impact is that if two effects are aliases of one another, it is impossible to estimate their *separate* effects. Therefore, we try to design the experiment so that only one of these is likely to be important and the other is likely to be insignificant. In this particular design, each main effect (single letter) is aliased with a three-way interaction—A with BCD, B with ACD, and so on. If three-way interactions are unlikely to be important, then any significant findings can be attributed to main effects, not three-way interactions. But note that the two-way interactions are confounded with each other—AB with CD, AC with BD, and AD with BC. It will probably be difficult to unravel these.

Figure 19.35 Counfounding Effects in an Incomplete Design

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Half-factorial design with 4 factors**													
2					Two-way interactions						Three-way interactions			
3	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	−1	1	1
6	1	−1	1	−1	−1	1	−1	−1	1	−1	−1	1	−1	1
7	1	−1	−1	1	−1	−1	1	1	−1	−1	1	−1	−1	1
8	−1	1	1	−1	−1	−1	1	1	−1	−1	−1	1	1	−1
9	−1	1	−1	1	−1	1	−1	−1	1	−1	1	−1	1	−1
10	−1	−1	1	1	1	−1	−1	−1	−1	1	1	1	−1	−1
11	−1	−1	−1	−1	1	1	1	1	1	1	−1	−1	−1	−1

As we have indicated, there is a whole science devoted to creating incomplete designs such as the one in Figure 19.34, and to analyzing the resulting data. [We again refer to Schmidt and Launsby (1994) and DeVor et al. (1992) for introductory accounts of the topic.] The usual approach, especially when there are a large number of potentially important input factors, is to run a highly fractional experiment (a small fraction of all possible treatment level combinations) to "screen" for the relatively few factors that have important effects. Having found these, a more detailed experiment, perhaps even a full factorial experiment, can be run to investigate the few important factors more fully. As the introductory vignette to this chapter explains, the results are often very impressive. These experiments can lead to lower costs, higher sales, higher reliability, and higher customer satisfaction—in short, to better products.

PROBLEMS

Level A

24. Suppose that a producer of single-room air conditioners wishes to test four prototype air conditioning units. The dependent variable is the number of days an air conditioner will function properly before its motor needs major repair. In this case the producer is interested in only one factor, the type of air conditioner, at four different levels. However, the manufacturer suspects that the type of use might affect the time until major repair. Specifically, these air conditioning units are used in three environments: (1) in residential homes located in northern climates, where they are used only on an occasional basis during the summer months; (2) in residential homes located in moderate climates, where they are used frequently during the summer months and seldom during other seasons of the year; and (3) in residential homes in southern climates, where they are used frequently throughout the year except during the cooler winter months. The producer suspects that these different environments may tend to obscure real differences among the four types of air conditioners.

To conduct this experiment, the producer has allocated 20 air conditioners of each type. Provided that the air conditioner producer is interested primarily in how the type of unit affects the time until major repair, how can the company control for the type of environment? Assume that approximately 10% of all single-room air conditioners produced by this company are used in homes located in northern climates, 25% are used in homes located in moderate climates, and 65% are used in homes located in southern climates. Explain, in detail, how the producer should set up this experiment.

25. Consider again the one-way ANOVA hypothesis test described in Problem 6. How could blocking be employed to control for a factor that is not of primary interest yet could introduce an unwanted source of variation in this case?

26. Consider again the one-way ANOVA hypothesis test described in Problem 7. How could blocking be employed to control for a factor that is not of primary interest yet could introduce an unwanted source of variation in this case?

Level B

27. Following the example presented in Section 19-6c, illustrate a half-fractional design with *five* factors, each at two levels. Specifically, generate figures similar to Figures 19.34 and 19.35 to support your verbal explanation. Identify the aliases.

19-7 CONCLUSION

This chapter has focused on the design of experiments and the statistical analysis of the resulting data, called analysis of variance. This methodology has long played an important role in agriculture and many natural sciences, particularly the medical sciences. The business world is just beginning to realize the importance of designed experiments for designing and producing better products, and this trend will undoubtedly continue as more people receive training in the techniques of experimental design and analysis of variance. It is important to keep sight of the overall goal: to see whether variations in one or more factors have significant effects on a dependent variable of interest. The role of experimental design is to set up experiments in a way—using randomization, blocking, fractional factorial designs, or whatever—to get as much information from the resulting data as possible. Then the techniques of ANOVA indicate whether any main effects or interactions are significant. If there are significant effects, confidence intervals can be formed to measure the magnitudes of specific differences between means or other contrasts. The goal of good experimental design is to identify important factor effects when they exist.

Summary of Key Terms

Term	Explanation	Excel	Pages	Equation Number
Analysis of variance (ANOVA)	A collection of methods for testing for differences in means across subpopulations (or across a single population treated in different ways)		19-2	
Observational study	A study that uses readily available information		19-3	
Designed experiment	A study in which data are obtained under controlled experimental conditions		19-3	
Experimental design	The plan that determines how many observations to obtain at which combinations of experimental conditions		19-4	
Dependent variable	The variable that is measured in an ANOVA study		19-4	
Factors	The categorical variables that serve as the explanatory variables in an ANOVA study		19-4	
Treatment levels	The possible values of a factor		19-4	
Experimental units	The people, machines, or whatever, that are measured in an ANOVA study		19-4	
One-way ANOVA	An ANOVA study with a single factor	StatTools/Statistical Inference/ One-Way ANOVA	19-4	
Two-way ANOVA	An ANOVA study with two factors	StatTools/Statistical Inference/ Two-Way ANOVA	19-4	
Balanced design	An experimental design where the same number of experimental units is assigned to each treatment level combination		19-5	
ANOVA table	A table that includes the ingredients (sums of squares, degrees of freedom, mean squares, F-ratio, and p-value) for tests of equal means	StatTools/Statistical Inference/ One-Way (or Two-Way) ANOVA	19-8	19.1, 19.2, 19.3

(*continued*)

Summary of Key Terms (*Continued*)

Term	Explanation	Excel	Pages	Equation Number
Confidence intervals in ANOVA	Confidence intervals for differences between pairs of means (or contrasts)	StatTools/Statistical Inference/ One-Way ANOVA	19-8	19.4, 19.5
Multiple comparison problem	The problem that when many statements are made, each with a stated level of confidence, the probability that at least one will be wrong is much larger than anticipated		19-20	
Contrast	A weighted combination of means where the weights sum to 0; used to contrast one combination of means with another		19-23	
Bonferroni, Tukey, Scheffé methods	Methods that expand confidence interval lengths to correct for the multiple comparison problem	StatTools/Statistical Inference/ One-Way ANOVA	19-20	
Full factorial design	An experimental design in which observations are made at each combination of factor levels.		19-25	
Incomplete (or fractional) design	An experimental design in which observations are made only at a selected subset of the combinations of factor levels		19-25	
Main effects	Indications of differences across levels of one factor (when averaged over the levels of the other factor)	Statistical Inference/Two-Way ANOVA	19-26	
Interactions	Situation where the effect of one factor on a dependent variable depends on the level of another factor	Statistical Inference/Two-Way ANOVA	19-26	
Randomization	The random assignment of experimental units to various levels of factors		19-36	
Blocking	A technique of assigning experimental units to similar blocks of experimental units to decrease error variation	Statistical Inference/Two-Way ANOVA	19-38	
Confounding	The (unavoidable) confusion of some effects with others in an incomplete experimental design		19-43	

PROBLEMS

Conceptual Questions

C.1. ANOVA is always a test of the equality of *means*. So why is the method called analysis of *variance*?

C.2. In ANOVA terminology, there are "factors" and "treatment levels." Give at least two examples of possible factors and treatment levels to illustrate that you understand what these terms mean.

C.3. Explain what the key ratio, the *F* ratio, in a one-way ANOVA table is all about and why it is the basis for a test of equal means.

C.4. Explain why there are several choices (in StatTools and other packages) for the type of confidence intervals in one-way ANOVA. Specifically, what problem do these variations attempt to solve?

C.5. Explain exactly how regression can be used as an alternative to the "standard" one-way ANOVA methodology. Which *p*-values from the two methods are guaranteed to be the same?

C.6. In academic papers that are based on data analysis, you often see a correlation matrix from the observed data with asterisks indicating the "significant" correlations (those that are significantly different from 0). How does Section 19-4 on the multiple comparison problem relate to this situation? (By the way, academic researchers often offer no evidence that they are even aware of the multiple comparison problem. Maybe they are, but who knows?)

C.7. What is the difference between a full factorial design and an incomplete factorial design? If the former is "better," why would you ever use the latter?

C.8. A study is performed on a sample of residential homes to discover whether the size of the monthly heating bill depends on the type of heat or the type of home. In particular, three types of heat are examined: electric, natural gas, and oil. Also, all homes are classified into two types: those on a single level and those with at least two stories. What might an interaction effect look like in this situation? Intuitively, do you think there is any reason to expect an interaction effect?

C.9. Chapters 8 and 9 discussed paired comparisons as one possibility when analyzing the difference between two means. What does this have to do with blocking, as discussed in Section 19-6b?

C.10. Part of the title of this chapter is "experimental design." Why is the design of an experiment so important? What is the main disadvantage of an experiment that is not properly designed?

Level A

28. Although four similar-sized small-car models exhibit similar miles per gallon (mpg) sticker ratings, there is some skepticism as to whether their mean mpg values are really equal. To test this equal-means hypothesis, several cars of each model are driven for 10,000 miles under nearly identical driving conditions. The observed mpg values are listed in the file **P19_28.xlsx**. Use one-way ANOVA to help decide whether the different models have equal mean mpg values, and write a short report to summarize your findings.

29. A professional golf association wants to compare the mean distances traveled by four brands of golf balls when struck by the same driver. Specifically, a robotic golfer uses a driver to hit a random sample of 80 balls (i.e., 20 balls of each brand). Note that the 80 balls are hit in random order. The distance is recorded for each hit, and the results are listed in the file **P19_29.xlsx**.

a. Is there any indication of differences in the mean distances traveled by the four types of balls? Perform an appropriate statistical test and report a *p*-value.

b. Select an appropriate significance level and construct confidence intervals for all pairs of differences between means. Which of these differences, if any, are statistically significant at the selected significance level?

30. Boxes of a popular cereal brand are filled by five identical machines at a local production plant. Independent samples are randomly drawn from a large number of cereal boxes filled by each machine, and the number of ounces of cereal in each selected box is listed in the file **P19_30.xlsx**. Use one-way ANOVA to help decide whether the five machines are yielding essentially equivalent average fills (in ounces). Briefly summarize your findings.

31. Assume that we gather independent random samples from large batches of each of three different brands of lightbulbs. We then list the lifetime of each selected bulb in the file **P19_31.xlsx**.

a. Test whether the different brands of lightbulbs have equal average lifetimes at the 10% significance level.

b. Based on 90% confidence intervals for all pairs of differences between means, which of these differences, if any, are significantly nonzero at the 10% significance level?

32. Consider again the one-way ANOVA hypothesis test described in Problem 28. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 95%. How do these results compare to the uncorrected 95% confidence intervals?

33. Consider again the one-way ANOVA hypothesis test described in Problem 29. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 99%. Compare the widths of the confidence intervals generated with each of these methods with those of uncorrected 99% confidence intervals. Explain your findings.

34. Consider again the one-way ANOVA hypothesis test described in Problem 30. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 95%. Summarize your results.

35. Consider again the one-way ANOVA hypothesis test described in Problem 31. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 90%. Compare the widths of the confidence intervals generated with each of these methods with those of uncorrected 90% confidence intervals. Explain your findings.

36. A commuter airline wants to determine the combination of advertising medium (four levels) and

advertising agency (two levels) that would produce the largest increase in ticket sales per advertising dollar spent. Each of the two advertising agencies has prepared advertisements in formats required for distribution by each of the media (including television, radio, newspaper, and Web site). Forty small towns of roughly the same size have been selected for this experiment. Furthermore, groups of five of these small towns have been assigned to receive an advertisement prepared and distributed by each of the eight agency–medium combinations. The dollar increases in ticket sales per advertising dollar spent, based on a one-month period, are listed in the file **P19_36.xlsx**. Test for any significant main effects and interactions at the 5% level, and briefly summarize your results.

37. The file **P19_37.xlsx** lists the miles per gallon for each of three different octanes (Octane A, Octane B, and Octane C) of gasoline and three types of vehicles (light, medium, and heavy). Subsets of ten vehicles of each type have been randomly assigned to each octane level.

a. Do you find evidence of a significant main effect for the octane factor? Explain.

b. Do you find evidence of a significant main effect for the vehicle type factor? Explain.

c. Do you find evidence of significant interactions between the two factors? Explain.

38. In an effort to increase unit sales of particular products in the short run, many supermarkets reduce the price of these products and increase their display space. Consider three levels of each factor: for the price factor, (1) normal price, (2) moderately reduced price, and (3) heavily reduced price; and for the display factor, (1) normal display space, (2) moderately increased display space, and (3) heavily increased display space. Suppose that each of these nine treatment combinations was applied five times to a specific product at a particular supermarket. Each treatment application lasted 7 days, and the dependent variable was unit sales for the week. The data for this experiment are listed in the file **P19_38.xlsx**. Test for any significant main effects and interactions at the 1% level, and briefly summarize your results.

39. Consider again the one-way ANOVA hypothesis test described in Problem 29. Suppose now that the professional golf association wants to compare the mean distances traveled by four brands of golf balls using *human* golfers instead of a robotic golfer. Each human golfer who participates in the experiment will employ the same type of driver to hit a subset of the 80 balls.

a. Explain how a randomized experimental design could be used to perform this one-way ANOVA.

b. Explain how a randomized block design could be used to perform this one-way ANOVA.

Level B

40. A production manager believes that the time required to assemble a particular product depends on the type of training that workers on the line receive. Four different training programs have been administered to workers of roughly equal experience at the local plant during the past year. To test her hypothesis, the production manager gathers assembly time data for randomly selected subsets of workers who have participated in one of the four training programs. These times are listed in the file **P19_40.xlsx**. Use one-way ANOVA to help decide whether the different training programs yield equivalent average assembly times, and write a short report to summarize your findings.

41. Consider again the one-way ANOVA hypothesis test described in Problem 40. Address the multiple comparison problem by applying the Bonferroni, Tukey, and Scheffé methods to obtain an *overall* confidence level of approximately 95%. Briefly summarize your results.

CASE 19.1 Krentz Appraisal Services

Nancy Krentz, the owner and manager of a property appraisal service based in York, Pennsylvania, is concerned that her four appraisers (Allen, Felan, Maloy, and Nelson) are producing appraisals of comparable properties that are generally not equivalent. She wants to conduct an investigation to determine whether her concerns are valid. Nancy directs her administrative assistant, Katie Shaffer, to identify 40 similar properties in the York area for use in the study. Given the sample of comparable properties, Nancy then arbitrarily divides the 40 properties into four subsets of ten. Next, she randomly assigns each subset to one of the four appraisers for assessment. The appraisals of the given 40 properties are listed in the file **C19_01.xlsx**. Given Nancy's limited background in statistical analysis, she has asked for your expert assistance in evaluating the data that her assistant

has compiled. She recalls that at one point in her business studies she learned a systematic method, called analysis of variance, for comparing the averages of related groups of quantitative data. However, she cannot recall the assumptions that must be met to apply this methodology, nor the procedures for implementing the appropriate method and correctly interpreting the results. Nancy has prepared the following list of questions that she would like for you to help her answer:

1. What requirements must be met to apply analysis of variance? Is it appropriate to use analysis of variance in this case?
2. Assuming that it is appropriate to apply a form of analysis of variance here, how can she use the appropriate method to analyze the data?
3. Does the statistical analysis confirm her suspicion that there are individual differences among the four appraisers? If so, which of the four appraisers are typically generating evaluations that are larger or smaller than those of the others?
4. Has the statistical test been formulated in the best manner? In particular, was it appropriate for Nancy to divide the 40 selected properties into four subsets of ten and then assign each subset to one of the appraisers? If not, how could the design of the study be modified to discover the most useful information in evaluating the appraisal staff at Krentz? Be as specific as possible.
5. In light of the results of this data analysis, what steps, if any, should Nancy take to improve the situation in her organization?