Design and Analysis of Experiments and Observational Studies

CHAPTER 21

Capital One

Not everyone graduates first in their class at a prestigious business school. But even doing that won't guarantee that the first company you start will become a Fortune 500 company within a decade. Richard Fairbank managed to do both. When he graduated from Stanford Business School in 1981, he wanted to start his own company, but, as he said in an interview with the *Stanford Business Magazine*, he had no experience, no money, and no business ideas. So he went to work for a consulting firm. Wanting to be on his own, he left in 1987 and landed a contract to study the operations of a large credit card bank in New York. It was then that he realized that the secret lay in data. He and his partner, Nigel Morris, asked themselves, "Why not use the mountains of data that credit cards produce to design cards with prices and terms to satisfy different customers?" But they had a hard time selling this idea to the large credit card issuers. At the time all cards carried the same interest rate—19.8% with a $20 annual fee, and almost half of the population didn't qualify for a card. And credit issuers were naturally resistant to new ideas.

Finally, Fairbank and Morris signed on with Signet, a regional bank that hoped to expand its modest credit card operation.

Using demographic and financial data about Signet's customers, they designed and tested combinations of card features that allowed them to offer credit to customers who previously didn't qualify. Signet's credit card business grew and, by 1994, was spun off as Capital One with a market capitalization of $1.1B. By 2000, Capital One was the ninth largest issuer of credit cards with $29.5B in cardholder balances.

Fairbank also introduced "scientific testing." Capital One designs experiments to gather data about customers. For example, customers who hear about a better deal than the one their current card offers may phone, threatening to switch to another bank unless they get a better deal. To help identify which potential card-hoppers were serious, Fairbank designed an experiment. When a card-hopper called, the customer service agent's computer randomly ordered one of three actions: match the claimed offer, split the difference in rates or fees, or just say no. In that way the company could gather data on who switched, who stayed, and how they behaved. Now when a potential card-hopper phones, the computer can give the operator a script specifying the terms to offer—or instruct the operator to bid the card-hopper a pleasant good-bye.

Fairbank attributes the phenomenal success of Capital One to their use of such experiments. According to Fairbank, "Anyone in the company can propose a test and, if the results are promising, Capital One will rush the new product or approach into use immediately." Why does this work for Capital One? Because, as Fairbank says, "We don't hesitate because our testing has already told us what will work."

In 2002, Capital One won the Wharton Infosys Business Transformation Award, which recognizes enterprises that have transformed their businesses by leveraging information technology.

21.1 Observational Studies

Fairbank started by analyzing the data that had already been collected by the credit card company. These data weren't from designed studies of customers. He simply *observed* the behavior of customers from the data that were already there. Such studies are called **observational studies.** Many companies collect data from customers with "frequent shopper" cards, which allow the companies to record each purchase. A company might study that data to identify associations between customer behavior and demographic information. For example, customers with pets might tend to spend more. The company can't conclude that owning a pet *causes* these customers to spend. People who have pets may also have higher incomes on

average or be more likely to own their own homes. Nevertheless, the company may decide to make special offers targeted at pet owners.

Observational studies are used widely in public health and marketing because they can reveal trends and relationships. Observational studies that study an outcome in the present by delving into historical records are called **retrospective studies.** When Fairbank looked at the accumulated experiences of Signet bank's credit card customers, he started with information about which customers earned the bank the most money and sought facts about these customers that could identify others like them, so he was performing a retrospective study. Retrospective studies can often generate testable hypotheses because they identify interesting relationships although they can't demonstrate a causal link.

When it is practical, a somewhat better approach is to observe individuals over time, recording the variables of interest and seeing how things turn out. For example, if we thought pet ownership might be a way to identify profitable customers, we might start by selecting a random sample of new customers and ask whether they have a pet. We could then track their performance and compare those who own pets to those who don't. Identifying subjects in advance and collecting data as events unfold would make this a **prospective study.** Prospective studies are often used in public health, where by following smokers or runners over a period of time we may find that one group or the other develops emphysema or arthritic knees (as you might expect), or dental cavities (which you might not anticipate).

Although an observational study may identify important variables related to the outcome we are interested in, there is no guarantee that it will find the right or the most important related variables. People who own pets may differ from the other customers in ways that we failed to observe. It may be this difference—whether we know what it is or not—rather than owning a pet in itself that leads pet owners to be more profitable customers. It's just not possible for observational studies, whether prospective or retrospective, to demonstrate a causal relationship. That's why we need experiments.

For Example — Observational studies

Amtrak launched its high-speed train, the Acela, in December 2000. Not only is it the only high-speed line in the United States, but it currently is the only Amtrak line to operate at a profit. The Acela line generates about one quarter of Amtrak's entire revenue.[1] The Acela is typically used by business professionals because of its fast travel times, high fares, business class seats, and free Wi-Fi. As a new member of the marketing department for the Acela, you want to boost young ridership of the Acela. You examine a sample of last year's customers for whom you have demographic information and find that only 5% of last year's riders were 21 years old or younger, but of those, 90% used the Internet while on board as opposed to 37% of riders older than 21 years.

Question: What kind of study is this? Can you conclude that Internet use is a factor in deciding to take the Acela?

Answer: This is a retrospective observational study. Although we can compare rates of Internet use between those older and younger than 21 years, we cannot come to any conclusions about why they chose to ride the Acela.

Just Checking

In early 2007, a larger-than-usual number of cats and dogs developed kidney failure; many died. Initially, researchers didn't know why, so they used an observational study to investigate.

1 Suppose that, as a researcher for a pet food manufacturer, you are called on to plan a study seeking the cause of this problem. Specify how you might proceed. Would your study be prospective or retrospective?

[1]Amtrak News Release ATK–09–074, October 2009.

21.2 Randomized, Comparative Experiments

Experiments are the only way to show cause-and-effect relationships convincingly, so they are a critical tool for understanding what products and ideas will work in the marketplace. An **experiment** is a study in which the experimenter *manipulates* attributes of what is being studied and observes the consequences. Usually, the attributes, called **factors**, are manipulated by being set to particular **levels** and then allocated or assigned to individuals. An experimenter identifies at least one factor to manipulate and at least one response variable to measure. Often the observed **response** is a quantitative measurement such as the amount of a product sold. However, responses can be categorical ("customer purchased"/ "customer didn't purchase"). The combination of factor levels assigned to a subject is called that subject's **treatment.**

The individuals on whom or which we experiment are known by a variety of terms. Humans who are experimented on are commonly called **subjects** or **participants**. Other individuals (rats, products, fiscal quarters, company divisions) are commonly referred to by the more generic term **experimental units.**

You've been the subject of marketing experiments. Every credit card offer you receive is actually a combination of various factors that specify your "treatment," the specific offer you get. For example, the factors might be *Annual Fee*, *Interest Rate*, and *Communication Channel* (e-mail, direct mail, phone, etc.). The particular treatment you receive might be a combination of *no Annual Fee* and a *moderate Interest Rate* with the offer being sent by *e-mail*. Other customers receive different treatments. The response might be categorical (do you accept the offer of that card?) or quantitative (how much do you spend with that card during the first three months you have it?).

Two key features distinguish an experiment from other types of investigations. First, the experimenter actively and deliberately manipulates the factors to specify the treatment. Second, the experiment assigns the subjects to those treatments at *random*. The importance of **random assignment** may not be immediately obvious. Experts, such as business executives and physicians, may think that they know how different subjects will respond to various treatments. In particular, marketing executives may want to send what they consider the best offer to the their best customers, but this makes fair comparisons of treatments impossible and invalidates the inference from the test. Without random assignment, we can't perform the hypothesis tests that allow us to conclude that differences among the treatments were responsible for any differences we observed in the responses. By using random assignment to ensure that the groups receiving different treatments are comparable, the experimenter can be sure that these differences are *due* to the differences in treatments. There are many stories of experts who were certain they knew the effect of a treatment and were proven wrong by a properly designed study. In business, it is important to get the facts rather than to just rely on what you may think you know from experience.

For Example — A marketing experiment

After finding out that most young riders of the Acela use the Internet while on board (see page 719), you decide to perform an experiment to see how to encourage more young people to take the Acela. After purchasing a mailing list of 16,000 college students, you decide to randomly send 1/4 a coupon worth 10% off their next Acela ride (*Coupon*), 1/4 a 5000 mile Amtrak mile bonus card (*Card*), and 1/4 a free Netflix download during their next Acela trip (*Movie*). The remaining 4000 students will receive no offer (*No Offer*). You plan to monitor the four groups to see which group travels most during the 12 months after sending the offer.

Question: What kind of study is this? What are the factors and levels? What are the subjects? What is the response variable?

Answer: This is an experiment because the factor (type of offer) has been manipulated. The four levels are *Coupon*, *Card*, *Movie*, and *No Offer*. The subjects are 16,000 college students. Each of four different offers will be distributed randomly to 1/4 of the college students. The response variable is *Miles Traveled* during the next 12 months on the Acela.

21.3 The Four Principles of Experimental Design

There are four **principles of experimental design.**

1. **Control.** We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. In a test of a new credit card, all alternative offers are sent to customers at the same time and in the same manner. Otherwise, if gas prices soar, the stock market drops, or interest rates spike dramatically during the study, those events could influence customers' responses, making it difficult to assess the effects of the treatments. So an experimenter tries to make any other variables that are not manipulated as alike as possible. Controlling extraneous sources of variation reduces the variability of the responses, making it easier to discern differences among the treatment groups.

   There is a second meaning of control in experiments. A bank testing the new creative idea of offering a card with special discounts on chocolate to attract more customers will want to compare its performance against one of their standard cards. Such a baseline measurement is called a control treatment, and the group that receives it is called the **control group.**

2. **Randomize.** In any true experiment, subjects are assigned treatments at random. Randomization allows us to equalize the effects of unknown or uncontrollable sources of variation. Although randomization can't eliminate the effects of these sources, it spreads them out across the treatment levels so that we can see past them. Randomization also makes it possible to use the powerful methods of inference to draw conclusions from your study. Randomization protects us even from effects we didn't know about. Perhaps women are more likely to respond to the chocolate benefit card. We don't need to test equal numbers of men and women—our mailing list may not have that information. But if we randomize, that tendency won't contaminate our results. There's an adage that says "Control what you can, and randomize the rest."

3. **Replicate.** Replication shows up in different ways in experiments. Because we need to estimate the variability of our measurements, we must make more than one observation at each level of each factor. Sometimes that just means making repeated observations. But, as we'll see later, some experiments combine two or more factors in ways that may permit a single observation for each *treatment*—that is, each combination of factor levels. When such an experiment is repeated in its entirety, it is said to be *replicated*. Repeated observations at each treatment are called **replicates**. If the number of replicates is the same for each treatment combination, we say that the experiment is **balanced**.

   A second kind of replication is to repeat the entire experiment for a different group of subjects, under different circumstances, or at a different time. Experiments do not require, and often can't obtain, representative random samples from an identified population. Experiments study the consequences of different levels of their factors. They rely on the random assignment of treatments to the subjects to generate the sampling distributions and to control for other possibly contaminating variables. When we detect a significant difference in response among treatment groups, we can conclude that it is due to the difference in treatments. However, we should take care in generalizing that result too broadly if we've only studied a specialized population. A special offer of accelerated checkout lanes for regular customers may attract more business in December, but it may not be effective in July. Replication in a variety of circumstances can increase our confidence that our results apply to other situations and populations.

4. **Blocking.** Sometimes we can identify a factor not under our control whose effect we don't care about, but which we suspect might have an effect either on

our response variable or on the ways in which the factors we are studying affect that response. Perhaps men and women will respond differently to our chocolate offer. Or maybe customers with young children at home behave differently than those without. Platinum card members may be tempted by a premium offer much more than standard card members. Factors like these can account for some of the variation in our observed responses because subjects at different levels respond differently. But we can't *assign* them at random to subjects. So we deal with them by grouping, or **blocking**, our subjects together and, in effect, analyzing the experiment separately for each block. Such factors are called **blocking factors**, and their levels are called **blocks.** Blocking in an experiment is like stratifying in a survey design. Blocking reduces variation by comparing subjects within these more homogenous groups. That makes it easier to discern any differences in response due to the factors of interest. In addition, we may want to study the effect of the blocking factor itself. Blocking is an important compromise between randomization and control. However, unlike the first three principles, blocking is not required in all experiments.

Just Checking

Following concerns over the contamination of its pet foods by melamine, which had been linked to kidney failure, a manufacturer now claims its products are safe. You are called on to design the study to demonstrate the safety of the new formulation.

2 Identify the treatment and response.

3 How would you implement control, randomization, and replication?

For Example Experimental design principles

Question: Explain how the four principles of experimental design are used in the Acela experiment described in the previous section (see page 720).

Answer:

Control: It is impossible to control other factors that may influence a person's decision to use the Acela. However, a control group—one that receives no offer—will be used to compare to the other three treatment levels.

Randomization: Although we can't control the other factors (besides *Offer*) that may influence a person's decision to use the Acela, by randomizing which students receive which offer, we hope that the influences of all those other factors will average out, enabling us to see the effect of the four treatments.

Replication: We will send each type of offer to 4000 students. We hope that the response is high enough that we will be able to see differences in *Miles Traveled* among the groups. This experiment is balanced since the number of subjects is the same for all four treatments.

Blocking: We have not blocked the experiment. Possible blocking factors might include demographic variables such as the region of the student's home or college, their sex, or their parent's income.

21.4 Experimental Designs

Completely Randomized Designs

When each of the possible treatments is assigned to at least one subject at random, the design is called a **completely randomized design**. This design is the simplest and easiest to analyze of all experimental designs. A diagram of the procedure can help in thinking about experiments. In this experiment, the subjects are assigned at random to the two treatments.

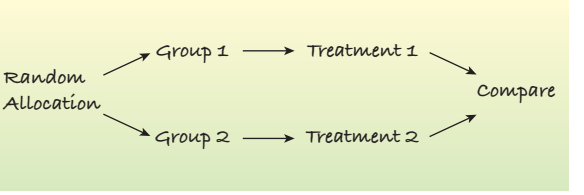


Figure 21.1 The simplest randomized design has two groups randomly assigned two different treatments.

Randomized Block Designs

When one of the factors is a blocking factor, complete randomization isn't possible. We can't randomly assign factors based on people's behavior, age, sex, and other attributes. But we may want to block by these factors in order to reduce variability and to understand their effect on the response. When we have a blocking factor, we randomize the subject to the treatments *within each block*. This is called a **randomized block design.** In the following experiment, a marketer wanted to know the effect of two types of offers in each of two segments: a high spending group and a low spending group. The marketer selected 12,000 customers *from each group* at random and then randomly assigned the three treatments to the 12,000 customers *in each group* so that 4000 customers in each segment received each of the three treatments. A display makes the process clearer.

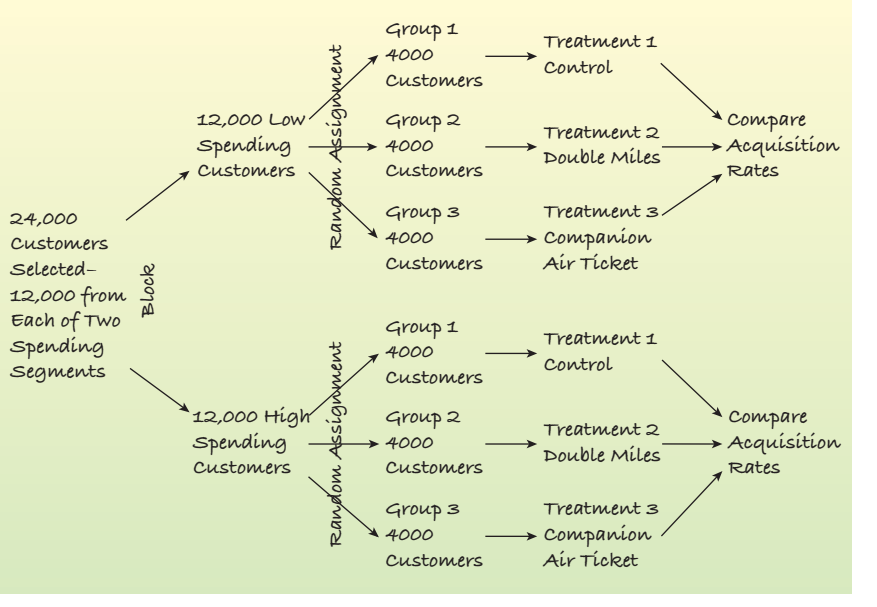


Figure 21.2 This example of a randomized block design shows that customers are randomized to treatments within each segment, or block.

Factorial Designs

An experiment with more than one manipulated factor is called a **factorial design.** A full factorial design contains treatments that represent all possible combinations of all levels of all factors. That can be a lot of treatments. With only three factors, one at 3 levels, one at 4, and one at 5, there would be $3 \times 4 \times 5 = 60$ different treatment combinations. So researchers typically limit the number of levels to just a few.

It may seem that the added complexity of multiple factors is not worth the trouble. In fact, just the opposite is true. First, if each factor accounts for some of the variation in responses, having the important factors in the experiment makes it *easier* to discern the effects of each. Testing multiple factors in a single experiment makes more efficient use of the available subjects. And testing factors together is the only way to see what happens at *combinations* of the levels.

An experiment to test the effectiveness of offering a $50 coupon for free gas may find that the coupon increases customer spending by 1%. Another experiment finds that lowering the interest rate increases spending by 2%. But unless some customers were offered *both* the $50 free gas coupon *and* the lower interest rate, the analyst can't learn whether offering both together would lead to still greater spending or less.

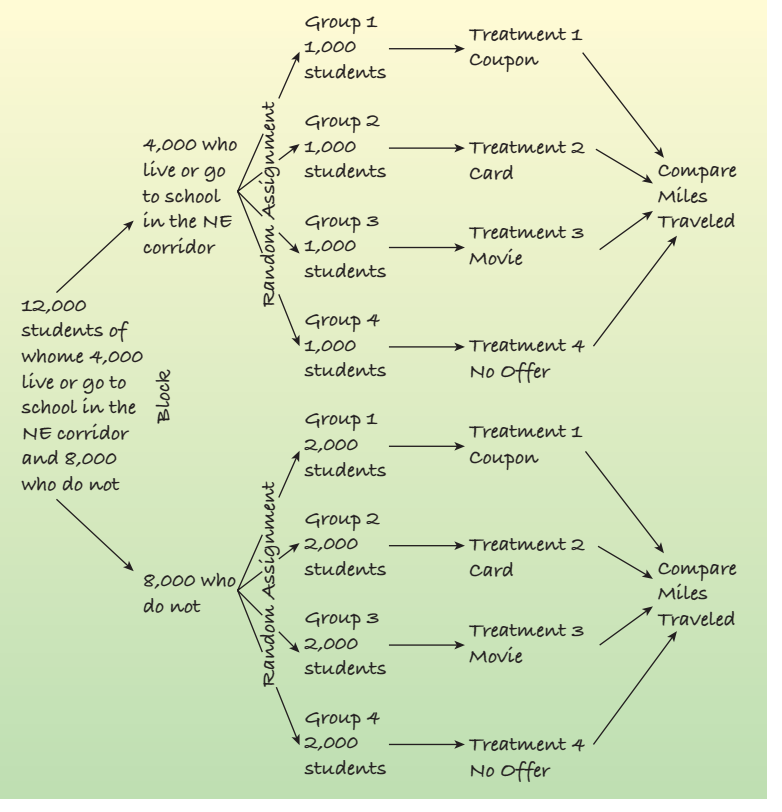
When the combination of two factors has a different effect than you would expect by adding the effects of the two factors together, that phenomenon is called an **interaction.** If the experiment does not contain both factors, it is impossible to see interactions. That can be a major omission because such effects can have the most important and surprising consequences in business.

For Example — Designing an experiment

Continuing the example from page 722, you are considering splitting up the students into two groups before mailing the offers: those who live or go to school in the Northeast corridor, where the Acela operates, and those who don't. Using home and school zip codes, you split the original 12,000 students into those groups and find that they split 4000 in the Northeast corridor and 8000 outside. You plan to randomize the treatments within those two groups and you'll monitor them to see if this factor, *NE corridor*, affects their *Miles Traveled* as well as the type of offer they receive.

Questions: What kind of design would this be? Diagram the experiment.

Answer: It is a randomized block experiment with *NE corridor* as the blocking factor.



Guided Example Designing a Direct Mail Experiment

At a major credit card bank, management has been pleased with the success of a recent campaign to cross-sell Silver card customers with the new SkyWest Gold card. But you, as a marketing analyst, think that the revenue of the card can be increased by adding three months of double miles on SkyWest to the offer, and you think the additional gain in charges will offset the cost of the double miles. You want to design a marketing experiment to find out what the difference will be in revenue if you offer the double miles. You've also been thinking about offering a new version of the miles called "use anywhere miles," which can be transferred to other airlines, so you want to test that version as well.

You also know that customers receive so many offers that they tend to disregard most of their direct mail. So, you'd like to see what happens if you send the offer in a shiny gold envelope with the SkyWest logo prominently displayed on the front. How can we design an experiment to see whether either of these factors has an effect on charges?

PLAN

State the problem.

We want to study two factors to see their effect on the revenue generated for a new credit card offer.

Response Specify the response variable.

Revenue is a percentage of the amount charged to the card by the cardholder. To measure the success, we will use the monthly charges of customers who receive the various offers. We will use the three months after the offer is sent out as the collection period and use the total amount charged per customer during this period as the response.

Factors Identify the factors you plan to test.

We will offer customers three levels of the factor **miles** for the SkyWest Gold card: regular (no additional) miles, double miles, and double "use anywhere miles." Customers will receive the offer in the standard envelope or the new SkyWest logo envelope (factor **envelope**).

Levels Specify the levels of the factors you will use.

Experimental Design Observe the principles of design:

Control any sources of variability you know of and can control.

We will send out all the offers to customers at the same time (in mid September) and evaluate the response as total charges in the period October through December.

Randomly assign experimental units to treatments to equalize the effects of unknown or uncontrollable sources of variation.

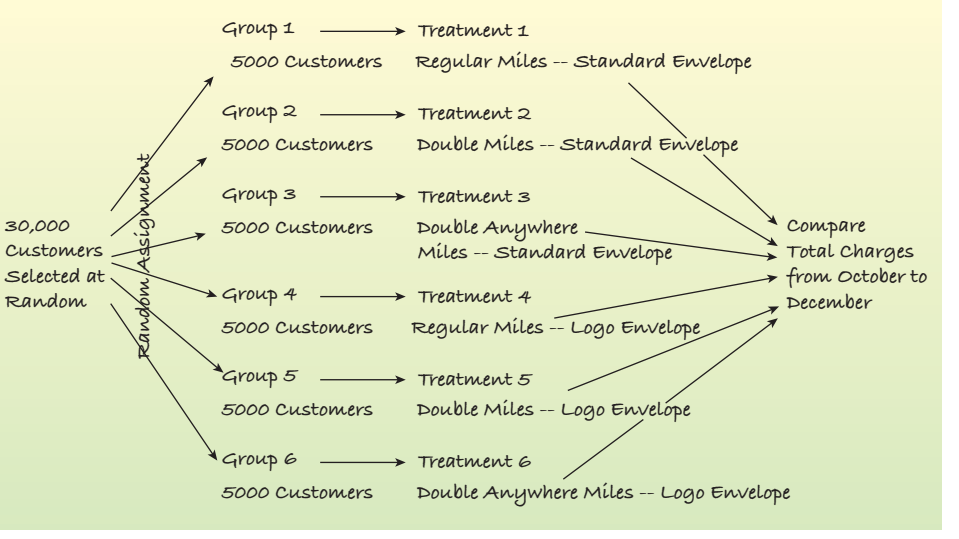
A total of 30,000 current Silver card customers will be randomly selected from our customer records to receive one of the six offers.

Replicate results by placing more than one customer (usually many) in each treatment group.

- ✓ Regular miles with standard envelope
- ✓ Double miles with standard envelope
- ✓ Double "anywhere miles" with standard envelope
- ✓ Regular miles with Logo envelope
- ✓ Double miles with Logo envelope
- ✓ Double "anywhere miles" with Logo envelope

(continued)

Make a Picture A diagram of your design can help you think about it.



Specify any other details about the experiment. You must give enough details so that another experimenter could exactly replicate your experiment.

It's generally better to include details that might seem irrelevant because they may turn out to make a difference.

Specify how to measure the response.

On January 15, we will examine the total card charges for each customer for the period October 1 through December 31.

DO

Once you collect the data, you'll need to display them (if appropriate) and compare the results for the treatment groups. (Methods of analysis for factorial designs will be covered later in the chapter.)

REPORT

To answer the initial question, we ask whether the differences we observe in the means (or proportions) of the groups are meaningful.

Because this is a randomized experiment, we can attribute significant differences to the treatments. To do this properly, we'll need methods from the analysis of factorial designs covered later in the chapter.

MEMO

Re: Test Mailing for Creative Offer and Envelope

The mailing for testing the Double Miles and Logo envelope ideas went out on September 17. On January 15, once we have total charges for everyone in the treatment groups, I would like to call the team back together to analyze the results to see:

- ✓ *Whether offering Double Miles is worth the cost of the miles*
- ✓ *Whether the "use anywhere miles" are worth the cost*
- ✓ *Whether the Logo envelope increased spending enough to justify the added expense*

21.5 Issues in Experimental Design

Blinding and Placebos

Blinding by Misleading

Social science experiments can sometimes blind subjects by disguising the purpose of a study. One of the authors participated as an undergraduate volunteer in one such (now infamous) psychology experiment. The subjects were told that the experiment was about 3-D spatial perception and were assigned to draw a model of a horse and were randomly assigned to a room alone or in a group. While they were busy drawing, a loud noise and then groaning were heard coming from the room next door. The *real* purpose of the experiment was to see whether being in a group affects how people reacted to the apparent disaster. The horse was only a pretext. The subjects were blind to the treatment because they were misled.

Humans are notoriously susceptible to errors in judgment—all of us. When we know what treatment is assigned, it's difficult not to let that knowledge influence our response or our assessment of the response, even when we try to be careful.

Suppose you were trying to sell your new brand of cola to be stocked in a school's vending machines. You might hope to convince the committee designated to make the choice that students prefer your less expensive cola, or at least that they can't taste the difference. You could set up an experiment to see which of the three competing brands students prefer (or whether they can tell the difference at all). But people have brand loyalties. If they know which brand they are tasting, it might influence their rating. To avoid this bias, it would be better to disguise the brands as much as possible. This strategy is called **blinding** the participants to the treatment. Even professional taste testers in food industry experiments are blinded to the treatment to reduce any prior feelings that might influence their judgment.

But it isn't just the subjects who should be blind. Experimenters themselves often subconsciously behave in ways that favor what they believe. It wouldn't be appropriate for you to run the study yourself if you have an interest in the outcome. People are so good at picking up subtle cues about treatments that the best (in fact, the only) defense against such biases in experiments on human subjects is to keep anyone who could affect the outcome or the measurement of the response from knowing which subjects have been assigned to which treatments. So, not only should your cola-tasting subjects be blinded, but also you, as the experimenter, shouldn't know which drink is which—at least until you're ready to analyze the results.

There are two main classes of individuals who can affect the outcome of the experiment:

- Those who could influence the results (the subjects, treatment administrators, or technicians)
- Those who evaluate the results (judges, experimenters, etc.)

When all the individuals in either one of these classes are blinded, an experiment is said to be **single-blind**. When everyone in both classes is blinded, we call the experiment **double-blind**. Double-blinding is the gold standard for any experiment involving both human subjects and human judgment about the response.

Placebos and Authority

The placebo effect is stronger when placebo treatments are administered with authority or by a figure who appears to be an authority. "Doctors" in white coats generate a stronger effect than salespeople in polyester suits. But the placebo effect is not reduced much, even when subjects know that the effect exists. People often suspect that they've gotten the placebo if nothing at all happens. So, recently, drug manufacturers have gone so far in making placebos realistic that they cause the same side effects as the drug being tested! Such "active placebos" usually induce a stronger placebo effect. When those side effects include loss of appetite or hair, the practice may raise ethical questions.

Often simply applying *any* treatment can induce an improvement. Every parent knows the medicinal value of a kiss to make a toddler's scrape or bump stop hurting. Some of the improvement seen with a treatment—even an effective treatment—can be due simply to the act of treating. To separate these two effects, we can sometimes use a control treatment that mimics the treatment itself. A "fake" treatment that looks just like the treatments being tested is called a **placebo**. Placebos are the best way to blind subjects so they don't know whether they have received the treatment or not. One common version of a placebo in drug testing is a "sugar pill." Especially when psychological attitude can affect the results, control group subjects treated with a placebo may show an improvement.

The fact is that subjects treated with a placebo sometimes improve. It's not unusual for 20% or more of subjects given a placebo to report reduction in pain, improved movement, or greater alertness or even to demonstrate improved health or performance. This **placebo effect** highlights both the importance of effective blinding and the importance of comparing treatments with a control. Placebo controls are so effective that you should use them as an essential tool for blinding whenever possible.

Just Checking

The pet food manufacturer we've been following hires you to perform the experiment to test whether their new formulation is safe and nutritious for cats and dogs.

4 How would you establish a control group?

5 Would you use blinding? How? (Can or should you use double-blinding?)

6 Both cats and dogs are to be tested. Should you block? Explain.

The best experiments are usually:

- Randomized
- Double-blind
- Comparative
- Placebo-controlled

Confounding and Lurking Variables

A credit card bank wanted to test the sensitivity of the market to two factors: the annual fee charged for a card and the annual percentage rate charged. The bank selected 100,000 people at random from a mailing list and sent out 50,000 offers with a low rate and no fee and 50,000 offers with a higher rate and a $50 annual fee. They discovered that people preferred the low-rate, no-fee card. No surprise. In fact, customers signed up for that card at over twice the rate as the other offer. But the question the bank really wanted to answer was: "How much of the change was due to the rate, and how much was due to the fee?" Unfortunately, there's simply no way to separate out the two effects with that experimental design.

If the bank had followed a factorial design in the two factors and sent out all four possible different treatments—low rate with no fee; low rate with $50 fee; high rate with no fee, and high rate with $50 fee—each to 25,000 people, it could have learned about both factors and could have also learned about the interaction between rate and fee. But we can't tease apart these two effects because the people who were offered the low rate were also offered the no-fee card. When the levels of one factor are associated with the levels of another factor, we say that the two factors are **confounded.**

Confounding can also arise in well-designed experiments. If some other variable not under the experimenter's control but associated with a factor has an effect on the response variable, it can be difficult to know which variable is really responsible for the effect. A shock to the economic or political situation that occurs during a marketing experiment can overwhelm the effects of the factors being tested. Randomization will usually take care of confounding by distributing uncontrolled factors over the treatments at random. But be sure to watch out for potential confounding effects even in a well-designed experiment.

Confounding may remind you of the problem of lurking variables that we discussed in Chapter 6. Confounding variables and lurking variables are alike in that they interfere with our ability to interpret our analyses simply. Each can mislead us, but they are not the same. A lurking variable is associated with two variables in such a way that it creates an apparent, possibly causal relationship between them. By contrast, confounding arises when a variable associated with a factor has an effect on the response variable, making it impossible to separate the effect of the factor from the effect of the confounder. Both confounding and lurking variables are outside influences that make it harder to understand the relationship we are modeling.

21.6 Analyzing a Design in One Factor—The One-Way Analysis of Variance

The most common experimental design used in business is the single factor experiment with two levels. Often these are known as champion/challenger designs because typically they're used to test a new idea (the challenger) against the current version (the champion). In this case, the customers offered the champion are the control group, and the customers offered the challenger (a special deal, a new offer, a new service, etc.) are the test group. As long as the customers are randomly assigned to the two groups, we already know how to analyze data from experiments like these. When the response is quantitative, we can test whether the means are equal with a two-sample t-test, and if the response is 0-1 (yes/no), we would test whether the two proportions are equal using a two proportion z-test.

But those methods can compare only two groups. What happens when we introduce a third level into our single factor experiment? Suppose an associate in a percussion music supply company, *Tom's Tom-Toms*, wants to test ways to increase the amount purchased from the catalog the company sends out every three months. He decides on three treatments: a coupon for free drum sticks with any purchase, a free practice pad, and a $50 discount on any purchase. The response will be the dollar amount of sales per customer. He decides to keep some customers as a control group by sending them the catalog without any special offer. The experiment is a single factor design with four levels: no coupon, coupon for free drum sticks, coupon for the practice pad, and $50 coupon. He assigns the same number of customers to each treatment randomly.

Now the hypothesis to test isn't quite the same as when we tested the difference in means between two independent groups. To test whether all k means are equal, the hypothesis becomes:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_A: \textit{at least one mean is different}$$

The test statistic compares the variance of the means to what we'd expect that variance to be based on the variance of the individual responses. Figure 21.3 illustrates the concept. The differences among the means are the same for the two sets of boxplots, but it's easier to see that they are different when the underlying variability is smaller.

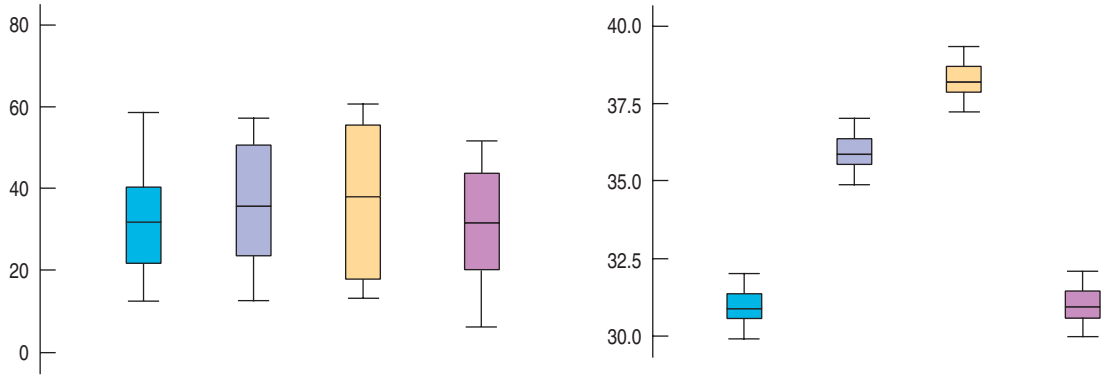


Figure 21.3 The means of the four groups in the left display are the same as the means of the four groups in the right display, but the differences are much easier to see in the display on the right because the variation within each group is less.

Why is it easier to see that the means[2] of the groups in the display on the right are different and much harder to see it in the one on the left? It is easier because we naturally compare the differences *between* the group means to the variation *within* each group. In the picture on the right, there is much less variation within each group so the differences among the group means are evident.

This is exactly what the test statistic does. It's the ratio of the variation among the group means to the variation within the groups. When the numerator is large enough, we can be confident that the differences among the group means are greater than we'd expect by chance, and reject the null hypothesis that they are equal. The test statistic is called the *F*-statistic in honor of Sir Ronald Fisher, who derived the sampling distribution for this statistic. The *F*-statistic showed up in multiple regression (Chapter 18) to test the null hypothesis that all slopes were zero. Here, it tests the null hypothesis that the means of all the groups are equal.

The ***F*-statistic** compares two quantities that measure variation, called *mean squares*. The numerator measures the variation *between* the groups (treatments) and is called the **Mean Square due to Treatments (MST)**. The denominator measures the variation *within* the groups, and is called the **Mean Square due to Error (MSE)**. The *F*-statistic is their ratio:

$$F_{k-1,\,N-k} = \frac{MST}{MSE}$$

We reject the null hypothesis that the means are equal if the *F*–statistic is too big. The critical value for deciding whether *F* is too big depends both on its degrees of freedom and the α-level you choose. Every ***F*-distribution** has two degrees of freedom, corresponding to the degrees of freedom for the mean square in the numerator and for the mean square (usually the MSE) in the denominator. Here, the MST has $k - 1$ degrees of freedom because there are k groups. The MSE has $N - k$ degrees of freedom where N is the total number of observations. Rather than compare the *F*-statistic to a specified critical value, we could find the P-value of this statistic and reject the null hypothesis when that value is small.

This analysis is called an **Analysis of Variance (ANOVA),** but the hypothesis is actually about *means*. The null hypothesis is that the means are all equal. The collection of statistics—the sums of squares, mean squares, *F*-statistic, and P-value—are usually presented in a table, called the **ANOVA table,** like this one:

Source	DF	Sum of Squares	Mean Square	*F*-Ratio	Prob > *F*
Treatment (Between)	$k - 1$	SST	MST	MST/MSE	P-value
Error (Within)	$N - k$	SSE	MSE		
Total	$N - 1$	SSTotal			

Table 21.1 An ANOVA table displays the treatment and error sums of squares, mean squares, *F*-ratio, and P-value.

- **How does the Analysis of Variance work?** When looking at side-by-side boxplots to see whether we think there are real differences between treatment means, we naturally compare the variation *between* the groups to the variation

[2]Of course the boxplots show medians at their centers, and we're trying to find differences among means. But for roughly symmetric distributions like these, the means and medians are very close.

within the groups. The variation between the groups indicates how large an effect the treatments have. The variation within the groups shows the underlying variability. To model those variations, the one-way ANOVA decomposes the data into several parts: the grand average, the treatment effects, and the residuals.

$$y_{ij} = \bar{\bar{y}} + (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i).$$

We can write this as we did for regression as

$$data = predicted + residual.$$

To estimate the variation *between* the groups we look at how much their means vary. The SST (sometimes called the *between* sum of squares) captures it like this:

$$SST = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2$$

where $\bar{y}_i$ is the mean of group i, n_i is the number of observations in group i and $\bar{\bar{y}}$ is the overall mean of all observations.

We compare the *SST* to how much variation there is *within* each group. The SSE captures that like this:

$$SSE = \sum_{i=1}^{k} (n_i - 1)s_i^2$$

where s_i^2 is the sample variance of group i.

To turn these estimates of variation into variances, we divide each sum of squares by its associated degrees of freedom:

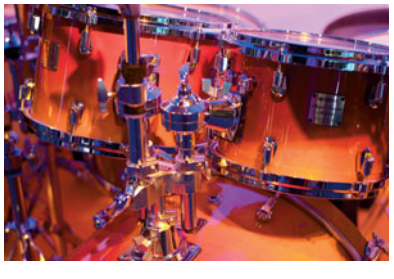
$$MST = \frac{SST}{k - 1}$$

$$MSE = \frac{SSE}{N - k}$$

Remarkably (and this is Fisher's real contribution), these two variances estimate the *same* variance when the null hypothesis is true. When it is false (and the group means differ), the MST gets larger.

The *F*-statistic tests the null hypothesis by taking the ratio of these two mean squares:

$F_{k-1,\,N-k} = \dfrac{MST}{MSE}$, and rejecting the hypothesis if the ratio is too large.



The critical value and P-value depend on the two degrees of freedom $k - 1$ and $N - k$.

Let's look at an example. For the summer catalog of the percussion supply company *Tom's Tom-Toms*, 4000 customers were selected at random to receive one of four offers[3]: *No Coupon*, *Free Sticks* with purchase, *Free Pad* with purchase, or $50 off next purchase. All the catalogs were sent out on March 15 and sales data for the month following the mailing were recorded.

[3]Realistically, companies often select equal (and relatively small) sizes for the treatment groups and consider all other customers as the control. To make the analysis easier, we'll assume that this experiment just considered 4000 "control" customers. Adding more controls wouldn't increase the power very much.

The first step is to plot the data. Here are boxplots of the spending of the four groups for the month after the mailing:

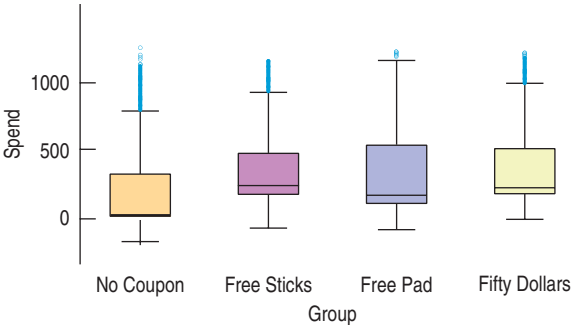


Figure 21.4 Boxplots of the spending of the four groups show that the coupons seem to have stimulated spending.

Here are summary statistics for the four groups:

SUMMARY			
Groups	*Median*	*Mean*	*SD*
No Coupon	$0.00	$216.68	$390.58
Free Sticks	$233.00	$385.87	$331.10
Free Pad	$157.50	$339.54	$364.17
Fifty Dollars	$232.00	$399.95	$337.07

The ANOVA table (from **Excel**) shows the components of the calculation of the *F*-test.

ANOVA					
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	20825966	3	6941988.66	54.6169	<0.0001
Within Groups	507905263	3996	127103.42		
Total	528731229	3999			

Table 21.2 The ANOVA table (from **Excel**) shows that the *F*-statistic has a very small P-value, so we can reject the null hypothesis that the means of the four treatments are equal.

The very small P-value is an indication that the differences we saw in the boxplots are not due to chance, so we reject the null hypothesis of equal means and conclude that the four means are not equal.

For Example — Analyzing a one-way design

You decide to implement the simple one factor completely randomized design sending out four offers (*Coupon*, *Card*, *Movie*, *or No Offer)* to 4000 students each (see page 720). A year later you collect the results and find the following table of means and standard deviations:

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Coupon	4,000	15.17	72.30	1.14	12.93	17.41
Card	4,000	11.53	62.62	0.99	9.59	13.47
Movie	4,000	13.29	66.51	1.05	11.22	15.35
No Offer	4,000	9.03	50.99	0.81	7.45	10.61

An ANOVA table shows:

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > *F*
Offer	3	81,922.26	27,307.42	6.75	0.0002
Error	15,996	64,669,900.04	4,042.88		
Total	15,999	64,751,822.29			

Question: What conclusions can you draw from these data?

Answer: From the ANOVA table, it looks like the null hypothesis that all the means are equal is strongly rejected. The average number of miles traveled seems to have increased about 2.5 miles for students receiving the *Card*, about 4.5 miles for students receiving the free *Movie*, and about 6 miles for those students receiving the *Coupon*.

21.7 Assumptions and Conditions for ANOVA

Whenever we compute P-values and make inferences about a hypothesis, we need to make assumptions and check conditions to see if the assumptions are reasonable. The ANOVA is no exception. Because it's an extension of the two-sample *t*-test, many of the same assumptions apply.

Independence Assumption

The groups must be independent of each other. No test can verify this assumption. You have to think about how the data were collected. The individual observations must be independent as well.

We check the **Randomization Condition**. Did the experimental design incorporate suitable randomization? We were told that the customers were assigned to each treatment group at random.

Equal Variance Assumption

ANOVA assumes that the true variances of the treatment groups are equal. We can check the corresponding **Similar Variance Condition** in various ways:

- Look at side-by-side boxplots of the groups to see whether they have roughly the same spread. It can be easier to compare spreads across groups when they have the same center, so consider making side-by-side boxplots of the residuals.

If the groups have differing spreads, it can make the pooled variance—the MSE—larger, reducing the *F*-statistic value and making it less likely that we can reject the null hypothesis. So the ANOVA will usually fail on the "safe side," rejecting H_0 less often than it should. Because of this, we usually require the spreads to be quite different from each other before we become concerned about the condition failing. If you've rejected the null hypothesis, this is especially true.

- Look at the original boxplots of the response values again. In general, do the spreads seem to change systematically with the centers? One common pattern is for the boxes with bigger centers to have bigger spreads. This kind of systematic trend in the variances is more of a problem than random differences in spread among the groups and should not be ignored. Fortunately, such systematic violations are often helped by re-expressing the data. If, in addition to spreads that grow with the centers, the boxplots are skewed with the longer tail stretching off to the high end, then the data are pleading for a re-expression. Try taking logs of the dependent variable for a start. You'll likely end up with a much cleaner analysis.
- Look at the residuals plotted against the predicted values. Often, larger predicted values lead to larger magnitude residuals. This is another sign that the condition is violated. If the residual plot shows more spread on one side or the other, it's usually a good idea to consider re-expressing the response variable. Such a systematic change in the spread is a more serious violation of the equal variance assumption than slight variations of the spreads across groups.

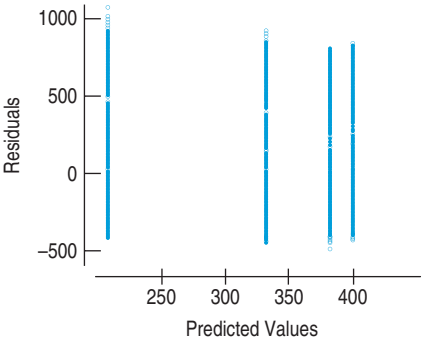


Figure 21.5 A plot of the residuals against the predicted values from the ANOVA shows no sign of unequal spread.

Normal Population Assumption

Like Student's *t*-tests, the *F*-test requires that the underlying errors follow a Normal model. As before when we faced this assumption, we'll check a corresponding **Nearly Normal Condition**.

Technically, we need to assume that the Normal model is reasonable for the populations underlying each treatment group. We can (and should) look at the side-by-side boxplots for indications of skewness. Certainly, if they are all (or mostly) skewed in the same direction, the Nearly Normal Condition fails (and re-expression is likely to help). However, in many business applications, sample sizes are quite large, and when that is true, the Central Limit Theorem implies that the sampling distribution of the means may be nearly Normal in spite of skewness. Fortunately, the *F*-test is conservative. That means that if you see a small P-value it's probably safe to reject the null hypothesis for large samples even when the data are nonnormal.

Check Normality with a histogram or a Normal probability plot of all the residuals together. Because we really care about the Normal model within each group, the Normal Population Assumption is violated if there are outliers in any of the groups. Check for outliers in the boxplots of the values for each treatment.

The Normal Probability plot for the *Tom's Tom-Toms* residuals holds a surprise.

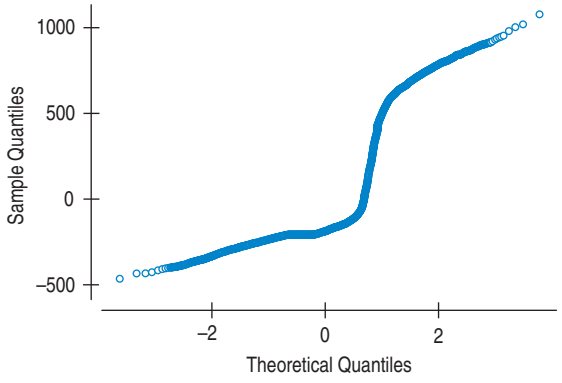


Figure 21.6 A normal probability plot shows that the residuals from the ANOVA of the *Tom's Tom-Toms* data are clearly not normal.

Investigating further with a histogram, we see the problem.

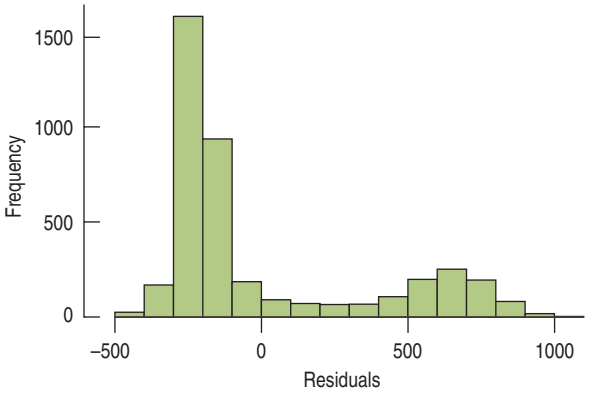


Figure 21.7 A histogram of the residuals reveals bimodality.

The histogram shows clear bimodality of the residuals. If we look back to histograms of the spending of each group, we can see that the boxplots failed to reveal the bimodal nature of the spending.

The manager of the company wasn't surprised to hear that the spending is bimodal. In fact, he said, "We typically have customers who either order a complete new drum set, or who buy accessories. And, of course, we have a large group of customers who choose not to purchase anything during a given quarter."

These data (and the residuals) clearly violate the Nearly Normal Condition. Does that mean that we can't say anything about the null hypothesis? No. Fortunately, the sample sizes are large, and there are no individual outliers that have undue influence on the means. With sample sizes this large, we can appeal to the Central Limit Theorem and still make inferences about the means. In particular, we are safe in rejecting the null hypothesis. When the Nearly Normal

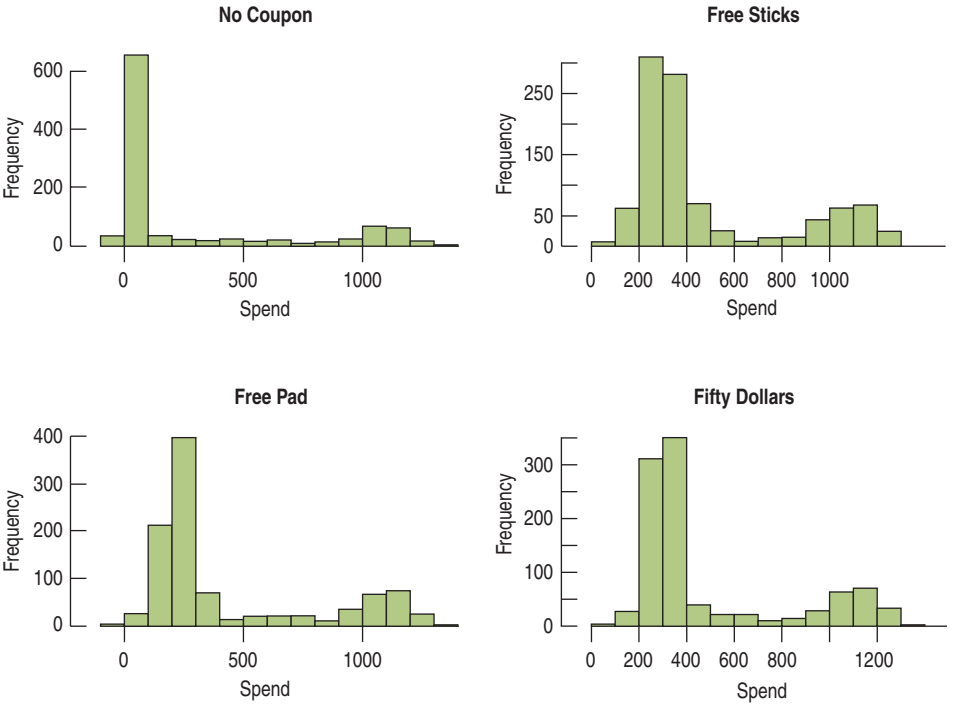


Figure 21.8 The spending appears to be bimodal for all the treatment groups. There is one mode near $1000 and another larger mode between $0 and $200 for each group.

Condition is not satisfied, the *F*-test will tend to fail on the safe side and be less likely to reject the null. Since we have a very small P-value, we can be fairly sure that the differences we saw were real. On the other hand, we should be very cautious when trying to make predictions about individuals rather than means.

For Example — Assumptions and conditions for ANOVA

Closer examination of the miles data from the Acela project (see page 720) shows that only about 5% of the students overall actually took the Acela, so the *Miles Traveled* are about 95% 0's and the other values are highly skewed to the right.

Question: Are the assumptions and conditions for ANOVA satisfied?

Answer: The responses are independent since the offer was randomized to the students on the mailing list. The distributions of *Miles Traveled* by *Offer* are highly right skewed. Most of the entries are zeros. This could present a problem, but because the sample size is so large (4000 per group), the inference is valid (a simulation shows that the averages of 4000 are Normally distributed). Although the distributions are right skewed, there are no extreme outliers that are influencing the group means. The variances in the four groups also appear to be similar. Thus, the assumptions and conditions appear to be met. (An alternative analysis might be to focus on the *Miles Traveled* only of those who actually took the Acela. The conclusion of the ANOVA would remain the same).

Just Checking

Your experiment to test the new pet food formulation has been completed. One hypothesis you have tested is whether the new formulation is different in nutritional value (measured by having veterinarians evaluate the test animals) from a standard food known to be safe and nutritious. The ANOVA has an F-statistic of 1.2, which (for the degrees of freedom in your experiment) has a P-value of 0.87. Now you need to make a report to the company.

7 Write a brief report. Can you conclude that the new formulation is safe and nutritious?

*21.8 Multiple Comparisons

Simply rejecting the null hypothesis is almost never the end of an Analysis of Variance. Knowing that the means are different leads to the question of which ones are different and by how much. Tom, the owner of *Tom's Tom-Toms*, would hardly be satisfied with a consultant's report that told him that the offers generated different amounts of spending, but failed to indicate which offers did better and by how much.

We'd like to know more, but the F-statistic doesn't offer that information. What can we do? If we can't reject the null hypothesis, there's no point in further testing. But if we can reject the null hypothesis, we can do more. In particular, we can test whether any pairs or combinations of group means differ. For example, we might want to compare treatments against a control or against the current standard treatment.

We could do t-tests for any pair of treatment means that we wish to compare. But each test would have some risk of a Type I error. As we do more and more tests, the risk that we'll make a Type I error grows. If we do enough tests, we're almost sure to reject one of the null hypotheses by mistake—and we'll never know which one.

Carlo Bonferroni (1892–1960) was a mathematician who taught in Florence. He wrote two papers in 1935 and 1936 setting forth the mathematics behind the method that bears his name.

There is a solution to this problem. In fact, there are several solutions. As a class, they are called methods for **multiple comparisons**. All multiple comparisons methods require that we first reject the overall null hypothesis with the ANOVA's F-test. Once we've rejected the overall null, we can think about comparing several—or even all—pairs of group means.

One such method is called the **Bonferroni method**. This method adjusts the tests and confidence intervals to allow for making many comparisons. The result is a wider margin of error (called the **minimum significant difference, or MSD**) found by replacing the critical t-value t^* with a slightly larger number. That makes the confidence intervals wider for each pairwise difference and the corresponding Type I error rates lower for each test, and it keeps the overall Type I error rate at or below α.

The Bonferroni method distributes the error rate equally among the confidence intervals. It divides the error rate among $\mathcal{J}$ confidence intervals, finding each interval at confidence level $1 - \frac{\alpha}{\mathcal{J}}$ instead of the original $1 - \alpha$. To signal this adjustment, we label the critical value t^{**} rather than t^*. For example, to make the

six confidence intervals comparing all possible pairs of offers at our overall α risk of 5%, instead of making six 95% confidence intervals, we'd use

$$1 - \frac{0.05}{6} = 1 - .0083 = .9917$$

instead of 0.95. So we'd use a critical t^{**} value of 2.64 instead of 1.96. The ME would then become:

$$ME = 2.642 \times 356.52\sqrt{\frac{1}{1000} + \frac{1}{1000}} = 42.12$$

This change doesn't affect our decision that each offer increases the mean sales compared to the *No Coupon* group, but it does adjust the comparison of average sales for the *Free Sticks* offer and the *Free Pad* offer. With a margin of error of \$42.12, the difference between average sales for those two offers is now $(385.87 - 339.54) \pm 42.12 = (\$4.21, \$88.45)$.

The confidence interval says that the *Free Sticks* offer generated between \$4.21 and \$88.45 more sales per customer on average than the *Free Pad* offer. In order to make a valid business decision, the company should now calculate their expected *profit* based on the confidence interval. Suppose they make 8% profit on sales. Then, multiplying the confidence interval

$$0.08 \times (\$4.21, \$88.45) = (\$0.34, \$7.08)$$

we find that the *Free Sticks* generate between \$0.34 and \$7.08 profit per customer on average. So, if the *Free Sticks* cost \$1.00 more than the pads, the confidence interval for profit would be:

$$(\$0.34 - \$1.00, \$7.08 - \$1.00) = (-\$0.66, \$6.08)$$

There is a possibility that the *Free Sticks* may actually be a less profitable offer. The company may decide to take the risk or to try another test with a larger sample size to get a more precise confidence interval.

Many statistics packages assume that you'd like to compare all pairs of means. Some will display the result of these comparisons in a table such as the one to the left. This table indicates that the top two are indistinguishable, that all are distinguishable from *No Coupon*, and that *Free Pad* is also distinguishable from the other three.

Fifty Dollars	\$399.95	A		
Free Sticks	\$385.87	A		
Free Pad	\$339.54		B	
No Coupon	\$216.68			C

Table 21.3 The output shows that the two top-performing offers are indistinguishable in terms of mean spending, but that the Free Pad is distinguishable from both those two and from *No Coupon*.

The subject of multiple comparisons is a rich one because of the many ways in which groups might be compared. Most statistics packages offer a choice of several methods. When your analysis calls for comparing all possible pairs, consider a multiple comparisons adjustment such as the Bonferroni method. If one of the groups is a control group, and you wish to compare all the other groups to it, there are specific methods (such as Dunnett's methods) available. Whenever you look at differences after rejecting the null hypothesis of equal means, you should consider using a multiple comparisons method that attempts to preserve the *overall* α risk.

For Example — Multiple comparisons

You perform a multiple comparison analysis using the Bonferroni correction on the Acela data (see page 733) and the output looks like:

Offer			Mean
Coupon	A		15.17
Movie	A		13.29
Card	A	B	11.53
No Offer		B	9.03

Question: What can you conclude?

Answer: From the original ANOVA we concluded that the means were not all equal. Now it appears that we can say that the mean *Miles Traveled is* greater for those receiving the *Coupon* or the *Movie* than those receiving *No Offer*, but we cannot distinguish the mean *Miles Traveled* between those receiving the *Card* or *No Offer*.

A further analysis only of those who actually took the Acela during the 12 months shows a slightly different story:

Coupon	A			300.37
Card	A	B		279.50
Movie		B		256.74
No Offer			C	206.40

Here we see that all offers can be distinguished from the *No Offer* group and that the *Coupon* group performed better than the group receiving the free *Movie*. Of those using the Acela, the *Coupon* resulted in nearly 100 more miles traveled on average during the year of those taking the Acela at least once.

21.9 ANOVA on Observational Data

So far we've applied ANOVA only to data from designed experiments. That application is appropriate for several reasons. The primary one is that randomized comparative experiments are specifically designed to compare the results for different treatments. The overall null hypothesis, and the subsequent tests on pairs of treatments in ANOVA, address such comparisons directly. In addition, the **Equal Variance Assumption** (which we need for all of the ANOVA analyses) is often plausible in a randomized experiment because when we randomly assign subjects to treatments, all the treatment groups start out with the same underlying variance of the experimental units.

Sometimes, though, we just can't perform an experiment. When ANOVA is used to test equality of group means from observational data, there's no *a priori* reason to think the group variances might be equal at all. Even if the null hypothesis of equal means were true, the groups might easily have different variances. But you can use ANOVA on observational data if the side-by-side boxplots of responses for each group show roughly equal spreads and symmetric, outlier-free distributions.

Balance

Recall that a design is called *balanced* if it has an equal number of observations for each treatment level.

Observational data tend to be messier than experimental data. They are much more likely to be unbalanced. If you aren't assigning subjects to treatment groups, it's harder to guarantee the same number of subjects in each group. And because you are not controlling conditions as you would in an experiment, things tend to be, well, less controlled. The only way we know to avoid the effects of possible lurking variables is with control and randomized assignment to treatment groups, and for observational data, we have neither.

ANOVA is often applied to observational data when an experiment would be impossible or unethical. (We can't randomly break some subjects' legs, but we *can* compare pain perception among those with broken legs, those with sprained ankles, and those with stubbed toes by collecting data on subjects who have already suffered those injuries.) In such data, subjects are already in groups, but not by random assignment.

Be careful; if you have not assigned subjects to treatments randomly, you can't draw *causal* conclusions even when the *F*-test is significant. You have no way to control for lurking variables or confounding, so you can't be sure whether any differences you see among groups are due to the grouping variable or to some other unobserved variable that may be related to the grouping variable.

Because observational studies often are intended to estimate parameters, there is a temptation to use pooled confidence intervals for the group means for this purpose. Although these confidence intervals are statistically correct, be sure to think carefully about the population that the inference is about. The relatively few subjects that you happen to have in a group may not be a simple random sample of any interesting population, so their "true" mean may have only limited meaning.

21.10 Analysis of Multifactor Designs

In our direct mail example, we looked at two factors: *Miles* and *Envelope*. *Miles* had three levels: *Regular Miles*, *Double Miles*, and *Double Anywhere Miles*. The factor *Envelope* had two levels: *Standard* and new *Logo*. The three levels of *Miles* and the two levels of *Envelope* resulted in six treatment groups. Because this was a completely randomized design, the 30,000 customers were selected at random, and 5000 were assigned at random to each treatment.

Three months after the offer was mailed out, the total charges on the card were recorded for each of the 30,000 cardholders in the experiment. Here are boxplots of the six treatment groups' responses, plotted against each factor.

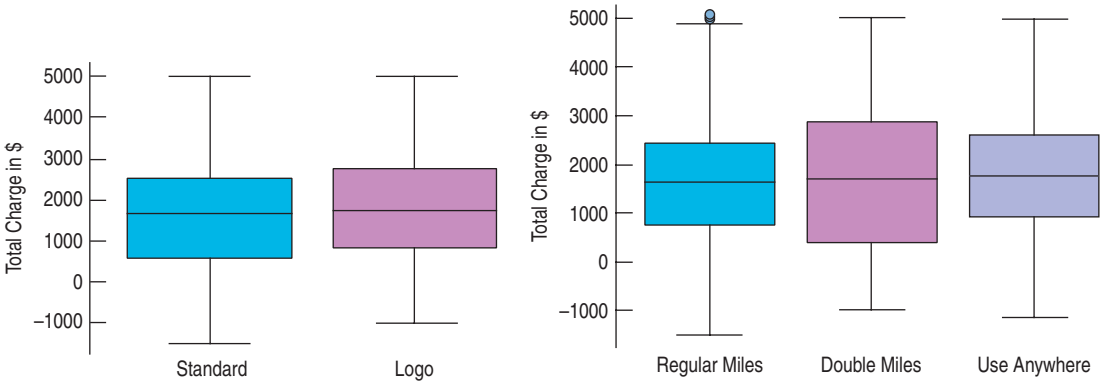


Figure 21.9 Boxplots of *Total Charge* by each factor. It is difficult to see the effects of the factors for two reasons. First, the other factor hasn't been accounted for, and second, the effects are small compared to the overall variation in charges.

If you look closely, you may be able to discern a very slight increase in the *Total Charges* for some levels of the factors, but it's very difficult to see. There are two reasons for this. First, the variation due to each factor gets in the way of seeing the effect of the other factor. For example, each customer in the boxplot for the *Logo Envelope* got one of three different offers. If those offers had an effect on spending, then that increased the variation within the *Logo* treatment group. Second, as is typical in a marketing experiment of this kind, the effects are very small compared to the variability in people's spending. That's why companies use such a large sample size.

The analysis of variance for two factors removes the effects of each factor from consideration of the other. It can also model whether the factors interact, increasing or decreasing the effect. In our example, it will separate out the effect of changing the levels of *Miles* and the effect of changing the levels of *Envelope*. It will also test whether the effect of the *Envelope* is the same for the three different *Miles* levels. If the effect is different, that's called an interaction effect between the two factors.

The details of the calculations for the two-way ANOVA with interaction are less important than understanding the summary, the model, and the assumptions and conditions under which it's appropriate to use the model. For a one-way ANOVA, we calculated three sums of squares (SS): the Total SS, the Treatment SS, and the Error SS. For this model, we'll calculate five: the Total SS, the SS due to Factor A, the SS due to Factor B, the SS due to the interaction, and the Error SS.

Let's suppose we have a levels of factor A, b levels of factor B, and r replicates at each treatment combination. In our case, $a = 2, b = 3, r = 5000$, and $a \times b \times r = N$ is 30,000. Then the ANOVA table will look like this.

Source	DF	Sum of Squares	Mean Square	*F*-Ratio	Prob > *F*
Factor A	$a - 1$	SSA	MSA	MSA/MSE	P-value
Factor B	$b - 1$	SSB	MSB	MSB/MSE	P-value
Interaction	$(a - 1) \times (b - 1)$	SSAB	MSAB	MSAB/MSE	P-value
Error	$ab(r - 1)$	SSE	MSE		
Total (Corrected)	$N - 1$	SSTotal			

Table 21.4 An ANOVA table for a replicated two-factor design with a row for each factor's sum of squares, interaction sum of square, error, and total.

There are now three null hypotheses—one that asserts that the means of the levels of factor A are equal, one that asserts that the means of the levels of factor B are all equal, and one that asserts that the effects of factor A are *constant* across the levels of factor B (or vice versa). Each P-value is used to test the corresponding hypothesis.

Here is the ANOVA table for the marketing experiment.

Source	DF	Sum of Squares	Mean Square	*F*-Ratio	Prob > *F*
Miles	2	201,150,000	100,575,000	66.20	<.0001
Envelope	1	203,090,000	203,090,000	133.68	<.0001
Miles × Envelope	2	1,505,200	752,600	0.50	0.61
Error	29,994	45,568,000,000	1,519,237		

Table 21.5 The ANOVA table for the marketing experiment. Both the effect of *Miles* and *Envelope* are highly significant, but the interaction term is not.

From the ANOVA table, we can see that both the *Miles* and the *Envelope* effects are highly significant, but that the interaction term is not. An **interaction plot,** a plot of means for each treatment group, is essential for sorting out what these P-values mean.

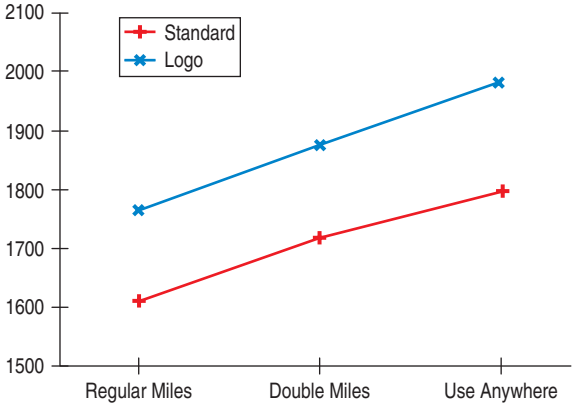


Figure 21.10 An interaction plot of the *Miles* and *Envelope* effects. The parallel lines show that the effects of the three *Miles* offers are roughly the same over the two different *Envelopes* and therefore that the interaction effect is small.

The interaction plot shows the mean *Charges* at all six treatment groups. The levels of one of the factors, in this case *Miles*, are shown on the *x*-axis, and the mean *Charges* of the groups for each *Envelope* level are shown at each *Miles* level. The means of each level of *Envelope* are connected for ease of understanding. Notice that the effect of *Double Miles* over *Regular Miles* is about the same for both the *Standard* and *Logo Envelopes*. And the same is true for the *Use Anywhere* miles. This indicates that the effect of *Miles* is constant for the two different *Envelopes*. The lines are parallel, which indicates that there is no interaction effect.

We reject the null hypothesis that the mean *Charges* at the three different levels of *Miles* are equal (with P-values < 0.0001), and also we reject that the mean *Charges* for *Standard* and *Logo* are the same (with P-value < 0.0001). We have no evidence, however, to suggest that there is an interaction between the factors.

After rejecting the null hypotheses, we can create a confidence interval for any particular treatment mean or perform a hypothesis test for the difference between any two means. If we want to do several tests or confidence intervals, we will need to use a multiple comparisons method that adjusts the size of the confidence interval or the level of the test to keep the *overall* Type I error rate at the level we desire.

When the interaction term is not significant, we can talk about the overall effect of either factor. Because the effect of *Envelope* is roughly the same for all three *Miles* offers (as we know by virtue of not rejecting the hypothesis that the interaction effect is zero), we can calculate and interpret an overall *Envelope* effect. The means of the two *Envelope* levels are:

Logo \$1871.75 *Standard* \$1707.19

and so the *Logo* envelope generated a difference in average charge of \$1871.75 − \$1707.19 = \$164.56. A confidence interval for this difference is (\$136.66, \$192.45), which the analysts can use to decide whether the added cost of the *Logo* envelope is worth the expense.

But when an interaction term *is* significant, we must be very careful not to talk about the effect of a factor, *on average*, because the effect of one factor *depends* on the level of the other factor. In that case, we always have to talk about the factor effect at a specific level of the other factor, as we'll see in the next example.

For Example Multifactor designs

Question: Suppose that you had run the randomized block design from Section 21.4 (see page 724). You would have had two levels of the (blocking) factor *NE Corridor* (*NE* or *not*) and the same four levels of *Offer* (*Coupon*, *Card*, *Movie*, and *No Offer*). The ANOVA shows a significant interaction effect between *NE Corridor* and *Offer*. Explain what that means. An analysis of the two groups (*NE* and *not*) separately shows that for the *NE* group, the P-value for testing the four offers is <0.0001 but for the *not* NE group, the P-value is 0.2354. Is this consistent with a significant interaction effect? What would you tell the marketing group?

Answer: A significant interaction effect implies that the effect of one factor is not the same for the levels of another. Thus, it is saying that the effect of the four offers is not the same for those living in the NE corridor as it is for those who do not. The separate analysis explains this further. For those who do not live in the NE, the offers do not significantly change the average number of miles they travel on the Acela. However, for those who live or go to school in the NE, the offers have an effect. This could impact where Amtrak decides to advertise the offers, or to whom they decide to send them.

Math Box

- **How does Two-Way Analysis of Variance work?** In Two-Way ANOVA, we have two factors. Each treatment consists of a level of each of the factors, so we write the individual responses as y_{ij}, to indicate the i^{th} level of the first factor and the j^{th} level of the second factor. The more general formulas are no more informative; just more complex. We will start with an unreplicated design with one observation in each treatment group (although in practice, we'd always recommend replication if possible). We'll call the factors A and B, each with a and b levels, respectively. Then the total number of observations is $n = a \times b$.

  For the first factor (factor A), the *Treatment Sum of Squares*, SSA, is the same as we calculated for one-way ANOVA:

$$SSA = \sum_{i=1}^{a} b(\bar{y}_i - \bar{\bar{y}})^2 = \sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_i - \bar{\bar{y}})^2$$

  where b is the number of levels of factor B, $\bar{y}_i$ is the mean of all subjects assigned level i of factor A (regardless of which level of factor B they were assigned), and $\bar{\bar{y}}$ is the *overall* mean of all observations. The mean square for treatment A (MSA) is

$$MSA = \frac{SSA}{a - 1}.$$

  The treatment sum of squares for the second factor (B) is computed in the same way, but of course the treatment means are now the means for each level of this second factor:

$$SSB = \sum_{j=1}^{b} a(\bar{y}_j - \bar{\bar{y}})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_j - \bar{\bar{y}})^2, \text{ and}$$

$$MSB = \frac{SSB}{b - 1}$$

  where a is the number of levels of factor A, and $\bar{\bar{y}}$, as before, is the overall mean of all observations. $\bar{y}_j$ is the mean of all subjects assigned the j^{th} level of factor B.

  The SSE can be found by subtraction:

$$SSE = SSTotal - (SSA + SSB)$$

  where

$$SSTotal = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{\bar{y}})^2.$$

  The mean square for error is $MSE = \dfrac{SSE}{N - (a + b - 1)}$, where $N = a \times b$.

  There are now two F-statistics, the ratio of each of the treatment mean squares to the MSE, which are associated with each null hypothesis.

  To test whether the means of all the levels of factor A are equal, we would find a P-value for $F_{a-1,\,N-(a+b-1)} = \dfrac{MSA}{MSE}$. For factor B, we would find a P-value for $F_{b-1,\,N-(a-b+1)} = \dfrac{MSB}{MSE}$.

  If the experiment is replicated (say r times), we can also estimate the interaction between the two factors, and test whether it is zero. The sums of squares for each factor have to be multiplied by r. Alternatively they can be written as a (triple) sum where the sum is now over factor A, factor B, and the replications:

$$SSA = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_i - \bar{\bar{y}})^2 \text{ and } SSB = \sum_{k=1}^{r}\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_j - \bar{\bar{y}})^2.$$

(continued)

We find the sum of squares for the interaction effect AB as:

$$SSAB = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j - \bar{\bar{y}})^2, \text{ and}$$

$$MSAB = \frac{SSAB}{(a-1)(b-1)}.$$

The SSE is the sum of the squared residuals:

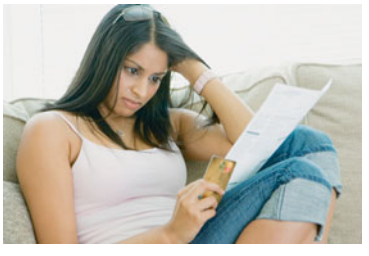
$$SSE = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{ij})^2 \text{ and } MSE = \frac{SSE}{ab(r-1)}.$$

There are now three *F*-statistics associated with the three hypotheses (factor A, factor B, and the interaction). They are the ratios of each of these mean squares with the MSE:

$$F_{a-1,\,ab(r-1)} = \frac{MSA}{MSE}, F_{b-1,\,ab(r-1)} = \frac{MSB}{MSE}, \text{ and } F_{(a-1)(b-1),\,ab(r-1)} = \frac{MSAB}{MSE}.$$

Note that $N = r \times a \times b$ is the total number of observations in the experiment.

Guided Example A Follow-up Experiment



After analyzing the data, the bank decided to go with the *Logo* envelope, but a marketing specialist thought that more *Miles* might increase spending even more. A new test was designed to test both the type of *Miles* and the amount. Again, total *Charge* in three months is the response.

PLAN

State the problem.

We want to study the two factors Miles and Amount to see their effect on the revenue generated for a new credit card offer.

Response Specify the response variable.

To measure the success, we will use the monthly charges of customers who receive the various offers. We will use the three months after the offer is sent out as the collection period and the total amount charged per customer during this period as the response.

Factors Identify the factors you plan to test.

Levels Specify the levels of the factors you will use.

We will offer each customer one of the two levels of the factor Miles for the SkyWest Gold card: SkyWest miles or Use Anywhere Miles. Customers are offered three levels of Miles: Regular Miles, Double Miles, and Triple Miles.

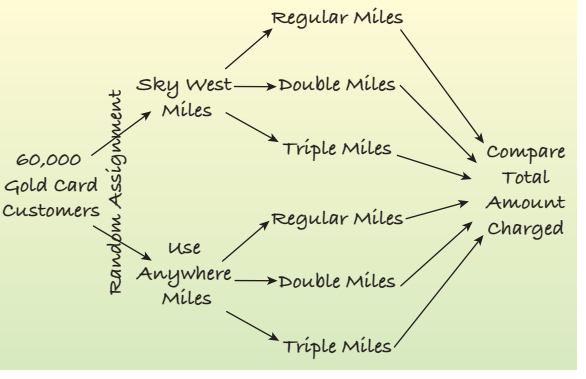
We will send out all the offers to customers at the same time (in mid March) and evaluate the response as total charges in the period April through June.

Experimental Design Specify the design.

A total of 60,000 current Gold card customers will be randomly selected from our customer records to receive one of the six offers.

- ✓ Regular SkyWest miles
- ✓ Double SkyWest miles
- ✓ Triple SkyWest miles
- ✓ Regular *Use Anywhere* Miles
- ✓ Double *Use Anywhere* Miles
- ✓ Triple *Use Anywhere* Miles

Make a Picture A diagram of your design can help you think about it. We could also draw this diagram like the one on page 726 with 6 treatment groups, but now we are thinking of the design as having two distinct factors that we wish to evaluate individually, so this form gives the right impression.



On June 15, we will examine the total card charges for each customer for the period April 1 through June 30.

Specify any other experimental details. You must give enough details so that another experimenter could exactly replicate your experiment.

We want to measure the effect of the two types of *Miles* and the three award Amounts.

It's generally better to include details that might seem irrelevant than to leave out matters that could turn out to make a difference.

Specify how to measure the response and your hypotheses.

The three null hypotheses are:

H_0: The mean charges for *Sky West Miles* and *Use Anywhere Miles* are the same (the means for *Miles* are equal).

H_0: The mean charges for *Regular Miles*, *Double Miles*, and *Triple Miles* are the same (the means for *Amount* are equal).

H_0: The effect of *Miles* is the same for all levels of *Amount* (and vice-versa) (no interaction effect).

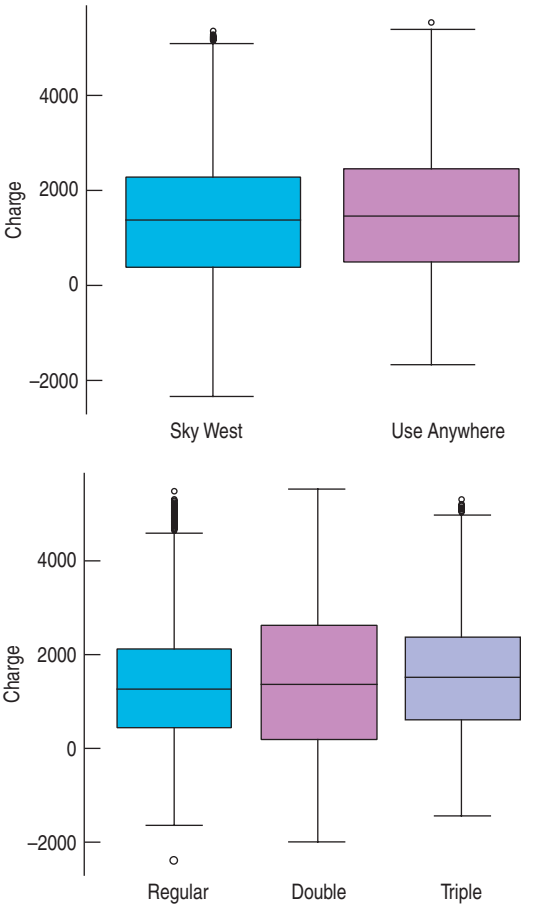
The alternative for the first hypothesis is that the mean *Charges* for the two levels of *Miles* are different.

The alternative for the second hypothesis is that at least one of the mean charges for the three levels of *Amount* is different.

The alternative for the third hypothesis is that there is an interaction effect.

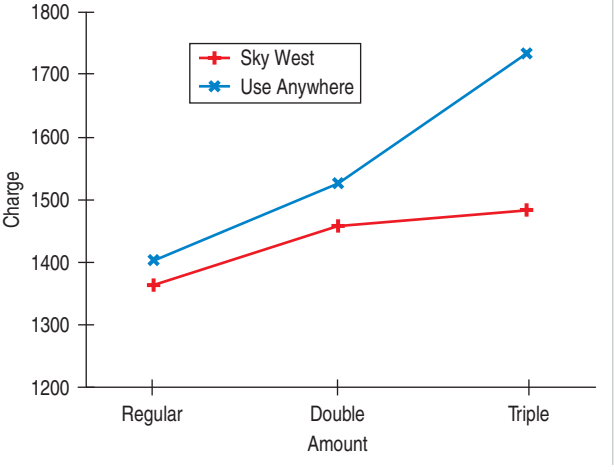
(continued)

Plot Examine the boxplots and interaction plots.



Boxplots by each factor show that there may be a slight increase in charges due to the *Use Anywhere* miles and the *Amount* of miles offered, but the differences are hard to see because of the intrinsic variation in *Charges*.

There are some outliers apparent in the boxplots, but none exerts a large influence on its group mean, so we will leave them in.



Assumptions and Conditions Think about the assumptions and check the conditions.

The interaction plot shows that offering *Triple* miles may have a much larger effect for *Use Anywhere* miles than for *Sky West* miles.

- ✓ **Independence Assumption, Randomization Condition.** The experiment was randomized to current cardholders.
- ✓ **Similar Variance Condition.** The boxplots show that the variances across all groups are similar. (We can recheck with a residual plot after fitting the ANOVA model.)
- ✓ **Outlier Condition.** There are some outliers, but none appear to be exerting undue influence on the group means.

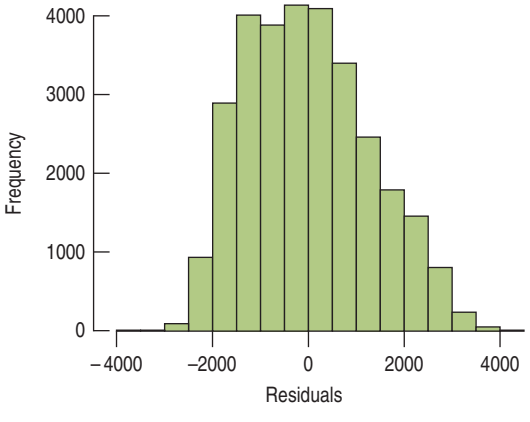
DO

Show the ANOVA table.

Source	Df	SS	MS	*F*-Ratio	P-Value
Miles	1	103,576,768	103,576,768	61.6216	<0.0001
Amount	2	253,958,660.1	126,979,330	75.5447	<0.0001
Miles × Amount	2	64,760,963.01	32,380,481.51	19.2643	<0.0001
Error	29,994	50,415,417,459	1,680,850		
Total	29,999	50,837,713,850			

Check the remaining conditions on the residuals.

- ✓ **Nearly Normal Condition.** A histogram of the residuals shows that they are reasonably unimodal and symmetric.



Under these conditions, it is appropriate to interpret the *F*-ratios and their *P*-values.

Discuss the results of the ANOVA table.

The *F*-ratios are all large, and the *P*-values are all very small, so we reject all three null hypotheses. Because the interaction effect is significant, we cannot talk about the overall effect of the amount of miles but must make the discussion specific to the type of Miles offered.

(continued)

Show a table of means, possibly with confidence intervals or tests from an appropriate Multiple Comparisons method.

Level					Mean
Use Anywhere, Triple	A				1732.21
Use Anywhere, Double		B			1526.93
Sky West, Triple		B			1484.34
Sky West, Double		B	C		1460.20
Use Anywhere, Regular			C	D	1401.89
Sky West, Regular				D	1363.94

REPORT

To answer the initial question, we ask whether the differences we observe in the means of the groups are meaningful.

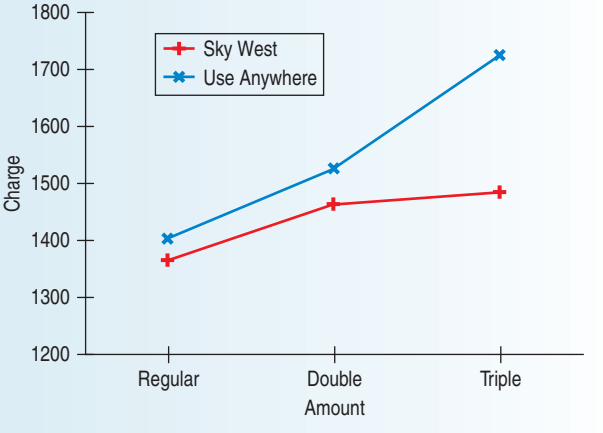
Because this is a randomized experiment, we can attribute significant differences to the treatments.

Be sure to make recommendations based on the context of your business decision.

MEMO

Re: Test Mailing for Creative Offer and Envelope

The mailing for testing the Triple Miles initiative went out in March, and results on charges from April through June were available in early July. We found that *Use Anywhere* miles performed better than the standard *Sky West* miles, but that the amount they increased charges depended on the amount offered.



As we can see, *Triple Miles* for the *Sky West* miles didn't increase *Charge* significantly and is probably not worth the added expense. However, *Triple Miles* for the *Use Anywhere* miles generated an average $205 more in average *Charge* (with a confidence interval from $131 to $279). Even at the low end of this interval, we feel that the added revenue of the Triple Miles justifies their cost.

In summary, we recommend offering Triple Miles for the Use Anywhere miles offers but would keep the Double miles offer for the Sky West miles.

What Can Go Wrong?

- **Don't give up just because you can't run an experiment.** Sometimes we can't run an experiment because we can't identify or control the factors. Sometimes it would simply be unethical to run the experiment. (Consider randomly assigning employees to two environments—one where workers were exposed to massive amounts of cigarette smoke and one a smoke-free environment—to see differences in health and productivity.) If we can't perform an experiment, often an observational study is a good choice.
- **Beware of confounding.** Use randomization whenever possible to ensure that the factors not in your experiment are not confounded with your treatment levels. Be alert to confounding that cannot be avoided, and report it along with your results.
- **Bad things can happen even to good experiments.** Protect yourself by recording additional information. An experiment in which the air-conditioning failed for two weeks, affecting the results, was saved by recording the temperature (although that was not originally one of the factors) and estimating the effect the higher temperature had on the response.[4] It's generally good practice to collect as much information as possible about your experimental units and the circumstances of the experiment. For example, in the direct mail experiment, it would be wise to record details of the general economy and any global events (such as a sharp downturn in the stock market) that might affect customer behavior.
- **Don't spend your entire budget on the first run.** Just as it's a good idea to pretest a survey, it's always wise to try a small pilot experiment before running the full-scale experiment. You may learn, for example, how to choose factor levels more effectively, about effects you forgot to control, and about unanticipated confounding.
- **Watch out for outliers.** One outlier in a group can change both the mean and the spread of that group. It will also inflate the Error Mean Square, which can influence the F-test. The good news is that ANOVA fails on the safe side by losing power when there are outliers. That is, you are less likely to reject the overall null hypothesis if you have (and leave) outliers in your data, so they are not likely to cause you to make a Type I error.
- **Watch out for changing variances.** The conclusions of the ANOVA depend crucially on the assumptions of independence and constant variance and (somewhat less seriously as the number of observations in each group increases) on Normality. If the conditions on the residuals are violated, it may be necessary to re-express the response variable to approximate these conditions more closely. ANOVA benefits so greatly from a judiciously chosen re-expression that the choice of a re-expression might be considered a standard part of the analysis.
- **Be wary of drawing conclusions about causality from observational studies.** ANOVA is often applied to data from randomized experiments for which causal conclusions are appropriate. If the data are not from a designed experiment, however, the Analysis of Variance provides no more evidence for causality than any other method we have studied. Don't get into the habit of assuming that ANOVA results have causal interpretations.

(continued)

[4]R. D. DeVeaux and M. Szelewski, "Optimizing Automatic Splitless Injection Parameters for Gas Chromatographic Environmental Analysis," *Journal of Chromatographic Science* 27, no. 9 (1989): 513–518.

- **Be wary of generalizing to situations other than the one at hand.** Think hard about how the data were generated to understand the breadth of conclusions you are entitled to draw.
- **Watch for multiple comparisons.** When rejecting the null hypothesis, you can conclude that the means are not *all* equal. But you can't start comparing every pair of treatments in your study with a *t*-test. You'll run the risk of inflating your Type I error rate. Use a multiple comparisons method when you want to test many pairs.
- **Be sure to fit an interaction term when it exists.** When the design is replicated, it is always a good idea to fit an interaction term. If it turns out not to be statistically significant, you can then fit a simpler two-factor main effects model instead.
- **When the interaction effect is significant, don't interpret the main effects.** Main effects can be very misleading in the presence of interaction terms. Look at this interaction plot:

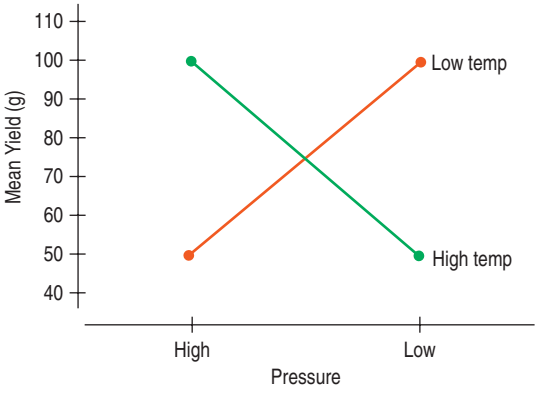


Figure 21.11 An interaction plot of *Yield by Temperature* and *Pressure*. The main effects are misleading. There is no (main) effect of *Pressure* because the average *Yield* at the two pressures is the same. That doesn't mean that *Pressure* has no effect on the *Yield*. In the presence of an interaction effect, be careful when interpreting the main effects.

The experiment was run at two temperatures and two pressure levels. High amounts of material were produced at high pressure with high temperature and at low pressure with low temperature. What's the effect of *Temperature*? Of *Pressure*? Both main effects are 0, but it would be silly (and wrong) to say that neither *Temperature* nor *Pressure* was important. The real story is in the interaction.

Ethics in Action

Professors at many state universities belong to a faculty union. The unionized faculty in one state's university system are preparing for contract negotiations. Cheryl McCrady, recently elected union president at one of the state university campuses, has long been concerned about the salary differential between male and female faculty. As union president, she now has access to faculty salary information, and she decides to run some analyses. After consulting with a few colleagues who regularly use statistics, she settles on using analysis of variance to determine if differences in salary can be attributed to gender accounting for faculty rank (assistant professor, associate professor, and full professor). She's not surprised by the results. While there is no significant interaction effect of gender and rank, she does find that both gender and rank are significant factors in explaining salary differences. Given that discrimination based on gender is a serious issue, she is wondering how she should proceed.

ETHICAL ISSUE *This is an observational study lacking the control of an experimental study. Confounding variables are likely to exist, but are not discussed. For instance, lower paid disciplines (e.g., Education) tend to have more female faculty than higher paid disciplines (e.g., Business). Related to Item A, ASA Guidelines. She should also check for outliers. Special cases, such as a star football coach or Nobel prize winner, may command unusually large salaries but not be relevant to the pay of ordinary faculty members.*

ETHICAL SOLUTION *Make all caveats explicit. This is a complex issue that should not be treated simply.*

What Have We Learned?

Learning Objectives

- Recognize observational studies.
  - A retrospective study looks at an outcome in the present and looks for facts in the past that relate to it.
  - A prospective study selects subjects and follows them as events unfold.
- Know the elements of a designed randomized experiment.
  - *Experimental units* (sometimes called *subjects* or *participants*) are assigned at random to *treatments*.
    - The experimenter manipulates *factors*, setting them to specified *levels* to establish the treatments.
  - A quantitative *response variable* is measured or observed for each experimental unit.
  - We can attribute differences in the response to the differences among the treatments.
- State and apply the Four Principles of Experimental Design.
  - *Control* sources of variation other than the factors being tested. Make the conditions as similar as possible for all treatment groups except for differences among the treatments.
  - *Randomize* the assignment of subjects to treatments. *Balance* the design by assigning the same number of subjects to each treatment.
  - *Replicate* the experiment on more than one subject.
  - *Block* the experiment by grouping together subjects who are similar in important ways that you cannot control.
- Work with *Blinding* and *Control groups*.
  - A *single-blind* study is one in which either all those who can affect the results or all those who evaluate the results are kept ignorant of which subjects receive which treatments.

- A *double-blind* study is one in which both those classes of actors are ignorant of the treatment assignment.
- A *control group* is assigned to a null treatment or to the best available alternative treatment.
  - Control subjects are often administered a *placebo* or null treatment that mimics the treatment being studied but is known to be inactive.

■ Understand how to use Analysis of Variance (ANOVA) to analyze designed experiments.

- ANOVA tables follow a standard layout; be acquainted with it.
- The *F*-statistic is the test statistic used in ANOVA. *F*-statistics test hypotheses about the equality of the means of two or more groups.

Terms

Analysis of Variance (ANOVA) An analysis method for testing equality of means across treatment groups.

ANOVA table The ANOVA table is convenient for showing the degrees of freedom, treatment mean square, error mean square, their ratio, *F*-statistic, and its P-value. There are usually other quantities of lesser interest included as well.

Blind, Blinding Any individual associated with an experiment who is not aware of how subjects have been allocated to treatment groups is said to be blinded.

Blocking, Blocking Factor When groups of experimental units are similar, it is often a good idea to gather them together into the same level of a factor. The factor is called a blocking factor and its levels are called blocks. By blocking we isolate the variability attributable to the differences between the blocks so that we can see the differences in the means due to the treatments more clearly.

Bonferroni method One of many methods for adjusting the margin of error to control the overall risk of making a Type I error when testing many pairwise differences between group means.

Confounded When a factor is associated with another factor in such a way that their effects cannot be separated, we say that these two factors are confounded.

Control When we limit the levels of a factor not explicitly part of the experiment design, we have controlled that factor. (By contrast, the factors we are testing are said to be *manipulated*.)

Control group The experimental units assigned to a baseline treatment level, typically either the default treatment, which is well understood, or a null, placebo treatment. Their responses provide a basis for comparison.

Designs

- **Randomized block design:** The randomization occurs only within blocks.
- **Completely randomized design:** All experimental units have an equal chance of receiving any treatment.
- **Factorial design:** Includes more than one factor in the same design and includes every combination of all the levels of each factor.

Double-blind, Single-blind There are two classes of individuals who can affect the outcome of an experiment:

those who could *influence the results* (subjects, treatment administrators, or technicians)

those who *evaluate the results* (judges, treating physicians, etc.)

When every individual in *either* of these classes is blinded, an experiment is said to be single-blind.
When everyone in *both* classes is blinded, we call the experiment double-blind.

Experiment An experiment *manipulates* factor levels to create treatments, *randomly assigns* subjects to these treatment levels, and then *compares* the responses of the subject groups across treatment levels.

Experimental units Individuals on whom an experiment is performed. Usually called subjects or participants when they are human.

F-distribution
: The *F*-distribution is the sampling distribution of the *F*-statistic when the null hypothesis that the treatment means are equal is true. The *F* is the ratio of two estimates of variance (mean squares), which are equal when the null hypothesis is true. It has two degrees of freedom parameters, corresponding to the degrees of freedom for the mean squares in the numerator and denominator respectively.

F-statistic
: The *F*-statistic for one-way ANOVA is the ratio MST/MSE. When the *F*-statistic is sufficiently large, we reject the null hypothesis that the group means are equal.

F-test
: The *F*-test tests the null hypothesis that all the group means are equal against the one-sided alternative that they are not all equal. We reject the hypothesis of equal means if the *F*-statistic exceeds the critical value from the *F*-distribution corresponding to the specified significance level and degrees of freedom.

Factor
: A variable whose levels are controlled by the experimenter. Experiments attempt to discover the effects that differences in factor levels may have on the responses of the experimental units.

Interaction
: When the effects of the levels of one factor change depending on the level of the other factor, the two factors are said to interact. When interaction terms are present, it is misleading to talk about the main effect of one factor because how large it is *depends* on the level of the other factor.

Interaction plot
: A plot that shows the means at each treatment combination, highlighting the factor effects and their behavior at all the combinations.

Level
: The specific values that the experimenter chooses for a factor are called the levels of the factor.

Mean Square
: A sum of squares divided by its associated degrees of freedom.

- **Mean Square due to Error (MSE)** The estimate of the error variance obtained by pooling the variance of each treatment group. The square root of the MSE is the estimate of the error standard deviation, s_p.
- **Mean Square due to Treatment (MST)** The estimate of the error variance under the null hypothesis that the treatment means are all equal. If the null hypothesis is not true, the expected value of the MST will be larger than the error variance.

Multiple comparisons
: If we reject the null hypothesis of equal means, we often then want to investigate further and compare pairs of treatment group means to see if the corresponding population means differ. If we want to test several such pairs, we must adjust for performing several tests to keep the overall risk of a Type I error from growing too large. Such adjustments are called methods for multiple comparisons.

Observational study
: A study based on data in which no manipulation of factors has been employed.

Placebo
: A treatment that mimics the treatment to be studied, designed so that all groups think they are receiving the same treatment. Many subjects respond to such a treatment (a response known as a *placebo effect*). Only by comparing with a placebo can we be sure that the observed effect of a treatment is not due simply to the placebo effect.

Placebo effect
: The tendency of many human subjects (often 20% or more of experiment subjects) to show a response even when administered a placebo.

Principles of experimental design
: - **Control** aspects of the experiment that we know may have an effect on the response, but that are not the factors being studied.
- **Randomize** subjects to treatments to even out effects that we cannot control.
- **Replicate** over as many subjects as possible. Results for a single subject are just anecdotes.
- **Block** to reduce the effects of identifiable attributes of the subjects that cannot be controlled.

Prospective study	An observational study in which subjects are followed to observe future outcomes. Because no treatments are deliberately applied, a prospective study is not an experiment. Nevertheless, prospective studies typically focus on estimating differences among groups that might appear as the groups are followed during the course of the study.
Random assignment	To be valid, an experiment must assign experimental units to treatment groups at random. This is called random assignment.
Response	A variable whose values are compared across different treatments. In a randomized experiment, large response differences can be attributed to the effect of differences in treatment level.
Retrospective study	An observational study in which subjects are selected and then their previous conditions or behaviors are determined. Because retrospective studies are not based on random samples, they usually focus on estimating differences between groups or associations between variables.
Subjects or Participants	When the experimental units are people, they are usually referred to as Subjects or Participants.
Treatment	The process, intervention, or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.

Technology Help: Analysis of Variance

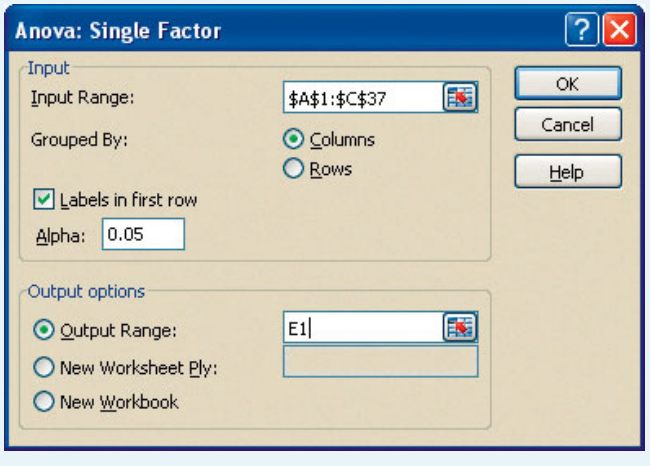
Most analyses of variance are performed with computers, and all statistics packages present the results in an ANOVA table much like the ones in the chapter. Technology also makes it easy to examine the side-by-side boxplots and check the residuals for violations of the assumptions and conditions. Statistics packages offer different choices among possible multiple comparisons methods. This is a specialized area. Get advice or read further if you need to choose a multiple comparisons method. There are two ways to organize data recorded for several groups. We can put all the response values in a single variable and use a second, "factor," variable to hold the group identities. This is sometimes called *stacked format.* The alternative is an unstacked format, placing the data for each group in its own column or variable. Then the variable identities become the group identifiers. Stacked format is necessary for experiments with more than one factor. Each factor's levels are named in a variable. Some packages can work with either format for simple one-factor designs, and some use one format for some things and the other for others. (Be careful, for example, when you make side-by-side boxplots; be sure to give the appropriate version of that command to correspond to the structure of your data.) Most packages offer to save residuals and predicted values and make them available for further tests of conditions. In some packages, you may have to request them specifically.

Some statistics packages have different commands for models with one factor and those with two or more factors. You must be alert to these differences when analyzing a two-factor ANOVA. It's not unusual to find ANOVA models in several different places in the same package. (Look for terms like "Linear Models.")

EXCEL XLSTAT

To compute a single-factor ANOVA:

- From the tools menu (or the Data Ribbon in Office 2007or 2010), select **Data Analysis**.
- Select **Anova Single Factor** from the list of analysis tools.
- Click the **OK** button.



- Enter the data range in the box provided.
- Check the **Labels in First Row** box, if applicable.

- Enter an alpha level for the *F*-test in the box provided.
- Click the **OK** button.

Comments

The data range should include two or more columns of data to compare. Unlike statistics packages, Excel expects each column of the data to represent a different level of the factor. However, it offers no way to label these levels. The columns need not have the same number of data values, but the selected cells must make up a rectangle large enough to hold the column with the most data values.

The Excel Data Analysis Add-in offers a two-way ANOVA "with and without replication." That command requires the data to be in a special format and cannot deal with unbalanced (i.e., unequal counts in treatment groups) data. See the Excel help files for specific instructions.

JMP

To compute a one-way ANOVA:

- From the **Analyze** menu, select **Fit Y by X**.
- Select variables: a quantitative Y, Response variable, and a categorical X, Factor variable.
- JMP opens the **Oneway** window.
- Click on the red triangle beside the heading, select **Display Options**, and choose **Boxplots**.
- From the same menu, choose the **Means/ANOVA t-test** command.
- JMP opens the one-way ANOVA output.

To compute a two-way ANOVA:

- From the **Analyze** menu, select **Fit Model**.
- Select variables and **Add** them to the **Construct Model Effects** box.
- To specify an interaction, select both factors and press the **Cross** button.
- Click **Run Model**.
- JMP opens a **Fit Least Squares** window.
- Click on the red triangle beside each effect to see the means plots for that factor. For the interaction term, this is the interaction plot.
- Consult JMP documentation for information about other features.

Comments

JMP expects data in "stacked" format with one continuous response and two nominal factor variables.

MINITAB

- Choose **ANOVA** from the Stat menu.
- Choose **One-way. . .** or **Two-way. . .** from the **ANOVA** submenu.
- In the dialog, assign a quantitative Y variable to the Response box and assign the categorical X factor(s) to the Factor box.
- In a two-way ANOVA, specify interactions.
- Check the **Store Residuals** check box.
- Click the **Graphs** button.
- In the ANOVA-Graphs dialog, select **Standardized residuals**, and check **Normal plot of residuals** and **Residuals versus fits**.
- Click the **OK** button to return to the ANOVA dialog.
- Click the **OK** button to compute the ANOVA.

Comments

If your data are in unstacked format, with separate columns for each treatment level, Minitab can compute a one-way ANOVA directly. Choose **One-way (unstacked)** from the **ANOVA** submenu. For two-way ANOVA, you must use the stacked format.

SPSS

To compute a one-way ANOVA:

- Choose **Compare Means** from the Analyze menu.
- Choose **One-way ANOVA** from the **Compare Means** submenu.
- In the One-Way ANOVA dialog, select the Y-variable and move it to the dependent target. Then move the X-variable to the independent target.
- Click the **OK** button.

To compute a two-way ANOVA:

- Choose **Analyze** > **General Linear Model** > **Univariate**.
- Assign the response variable to the **Dependent Variable** box.
- Assign the two factors to the **Fixed Factor(s)** box. This will fit the model with interactions by default.
- To omit interactions, click on **Model**. Select **Custom**. Highlight the factors. Select **Main Effects** under the **Build Terms** arrow and click the arrow.
- Click **Continue** and **OK** to compute the model.

Comments

SPSS expects data in stacked format. The **Contrasts** and **Post Hoc** buttons offer ways to test contrasts and perform multiple comparisons. See your SPSS manual for details.

Brief CASE

Design, carry out, and analyze your own multifactor experiment. The experiment doesn't have to involve human subjects. In fact, an experiment designed to find the best settings for microwave popcorn, the best paper airplane design, or the optimal weight and placement of coins on a toy car to make it travel farthest and fastest down an incline are all fine ideas. Be sure to define your response variable of interest before you start the experiment and detail how you'll perform the experiment, specifically including the elements you control, how you use randomization, and how many times you replicate the experiment. Analyze the results of your experiment and write up your analysis and conclusions including any recommendations for further testing.

Exercises

SECTION 21.1

1. For the following observational studies, indicate whether they are prospective or retrospective.

a) A company looked at a sample of returned registration cards to estimate the income level of households that purchased their product.

b) A retail outlet encouraged customers to join their "frequent buyers" program and studied whether those who joined were more likely to make use of discount coupons than those who were not members.

2. For the following observational studies, indicate whether they are prospective or retrospective studies.

a) An airline was concerned that new security measures might discourage air travelers. A year after the new security restrictions were put into place, the airlines compared the miles traveled by their frequent fliers before and after the change.

b) Does giving children a flu shot protect parents? Researchers questioned a random sample of families at the end of a flu season. They asked whether the children had been immunized, whether the parents had received flu shots, and who in the family had contracted the flu.

SECTION 21.2

3. For the following experiment, identify the experimental units, the treatments, the response, and the random assignment.

A commercial food lab compared recipes for chocolate chip cookies. They baked cookies with different kinds of chips (milk chocolate, dark chocolate, and semi-sweet). All other ingredients and amounts were the same. Ten trained tasters rated the cookies on a scale of 1 to 10. The cookies were presented to the tasters in a random order.

4. For the following experiment, identify the experimental units, the treatments, the response, and the random assignment.

An investment club decided to compare investment strategies. Starting with nine equal investment amounts, three invested in the "dogs of the Dow"—stocks in the Dow Industrial average that had been underperforming relative to the rest of the Dow average. The relative amounts to invest in each of the stocks were chosen randomly and differently for each fund. Three funds invested following the advice of a TV investment show host, again choosing the specific stocks and allocations randomly for the three funds. And three funds invested by throwing darts at a page from the *Wall Street Journal* that listed stocks on the NYSE, and invested in each of the stocks hit by a dart, throwing a different set of darts for each of the three funds. At the end of six months the funds were compared.

SECTION 21.3

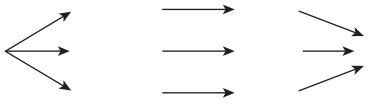
5. For the cookie recipe experiment of Exercise 3, identify how Control, Randomization, and Replication were used.

6. For the investment experiment of Exercise 4, identify how Control, Randomization, and Replication were used.

SECTION 21.4

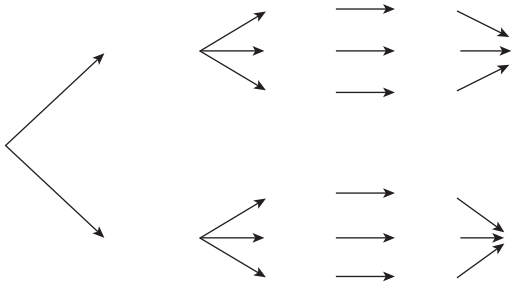
7. An Internet sale site randomly sent customers to one of three versions of its welcome page. It recorded how long each visitor stayed in the site.

Here is a diagram of that experiment. Fill in the parts of the experiment.



8. An Internet company was concerned that customers who came directly to their site (by typing their URL into a browser) might respond differently than those referred to the site from other sites (such as search engines). They decided to block according to how the customer arrived at their site.

Here is a diagram of that experiment. Fill in the parts.



SECTION 21.5

9. For the following experiment, indicate whether it was single-blind, double-blind, or not blinded at all. Explain your reasoning.

Makers of a new frozen entrée arranged for it to be served to randomly selected customers at a restaurant in place of the equivalent entrée ordinarily prepared in the kitchen. After their meal, the customers were asked about the quality of the food.

10. For the following experiment, indicate whether it was single-blind, double-blind, or not blinded at all. Explain your reasoning.

Does a "stop smoking" program work better if it costs more? Smokers responding to an advertisement offering to help them stop smoking were randomly offered a program costing $100 or the same program costing $250. The offer was made individually to each client by presenting a sealed envelope so the clerk providing the offer did not know the details of the offer. At the end of the program (a course and films along with diet and smoking cessation aids), clients were followed for six months to see if they had indeed quit smoking.

SECTION 21.6

11. In a completely randomized design, ten subjects were assigned to each of four treatments of a factor. Below is the partially completed ANOVA table.

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
Treatment (Between)		856.07			
Error (Within)					
Total		1177.97			

a) What are the degrees of freedom for treatment, error, and total?
b) What is SSE?
c) What is MST?
d) What is MSE?

12. Refer to Exercise 11.

a) State the null and alternative hypotheses.
b) Calculate the F-ratio.
c) What is the P-value?
d) State your conclusion at $\alpha = .05$.

SECTION 21.10

13. In the experiment described in Exercise 3, in fact the study also compared the use of butter or margarine in the recipes. The design was balanced, with each combination of chip type and oil type tested.

a) What were the factors and factor levels?
b) What were the treatments?
c) If an interaction was found to be significant, what would that mean?

14. The investment club described in Exercise 4 decided to repeat their experiment in a different way. Three members of the club took responsibility for one of each of the three investment "strategies," making the final choices and allocations of investment dollars. For this new experiment:

a) What were the subjects?
b) What were the factors and factor levels?
c) What were the treatments?

CHAPTER EXERCISES

15. **Laundry detergents.** A consumer group wants to test the efficacy of a new laundry detergent. They take 16 pieces of white cloth and stain each with the same amount of grease. They decide to try it using both hot and cold water settings and at both short and long washing times. Half of the 16 pieces will get the new detergent, and half will get a standard detergent. They'll compare the shirts by using an optical scanner to measure whiteness.

a) What are the factors they are testing?
b) Identify all the factor levels.
c) What is/are the response(s)?

16. Sales scripts. An outdoor products company wants to test a new website design where customers can get information about their favorite outdoor activity. They randomly send half of the customers coming to the website to the new design. They want to see whether the Web visitors spend more time at the site and whether they make a purchase.

a) What are the factors they are testing?
b) Identify all the factor levels.
c) What is/are the response(s)?

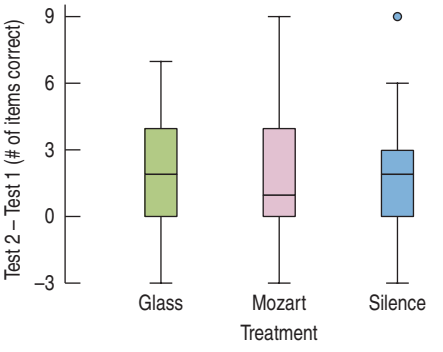
17. Laundry detergents, part 2. One member of the consumer group in Exercise 15 is concerned that the experiment will take too long and makes some suggestions to shorten it. Comment briefly on each idea.

a) Cut the runs to 8 by testing only the new detergent. Compare the results to results on the standard detergent published by the manufacturer.
b) Cut the runs to 8 by testing only in hot water.
c) Keep the number of runs at 16, but save time by running all the standard detergent runs first to avoid swapping detergents back and forth.

18. Swimsuits. A swimsuit manufacturer wants to test the speed of its newly designed $550 suit. They design an experiment by having 6 randomly selected Olympic swimmers swim as fast as they can with their old swimsuit first and then swim the same event again with the new, expensive swim suit. They'll use the difference in times as the response variable. Criticize the experiment and point out some of the problems with generalizing the results.

T **19. Mozart.** Will listening to a Mozart piano sonata make you smarter? In a 1995 study, Rauscher, Shaw, and Ky reported that when students were given a spatial reasoning section of a standard IQ test, those who listened to Mozart for 10 minutes improved their scores more than those who simply sat quietly.

a) These researchers said the differences were statistically significant. Explain what that means in this context.
b) Steele, Bass, and Crook tried to replicate the original study. The subjects were 125 college students who participated in the experiment for course credit. Subjects first took the test. Then they were assigned to one of three groups: listening to a Mozart piano sonata, listening to music by Philip Glass, and sitting for 10 minutes in silence. Three days after the treatments, they were retested. Draw a diagram displaying the design of this experiment.
c) The boxplots show the differences in score before and after treatment for the three groups. Did the Mozart group show improvement?
d) Do you think the results prove that listening to Mozart is beneficial? Explain.



20. More Mozart. An advertisement selling specially designed CDs of Mozart's music specifically because they will "strengthen your mind, heal your body, and unlock your creative spirit" claims that "in Japan, a brewery actually reported that their best sake is made when Mozart is played near the yeast." Suppose you wished to design an experiment to test whether this is true. Assume you have the full cooperation of the sake brewery. Specify how you would design the experiment. Indicate factors and response and how they would be measured, controlled, or randomized.

21. Cereal marketing. The makers of Frumpies, "the breakfast of rug rats," want to improve their marketing, so they consult you.

a) They first want to know what fraction of children, ages 10 to 13, like their celery-flavored cereal. What kind of study should they perform?
b) They are thinking of introducing a new flavor, maple-marshmallow Frumpies and want to know whether children will prefer the new flavor to the old one. Design a completely randomized experiment to investigate this question.
c) They suspect that children who regularly watch the Saturday morning cartoon show starring Frump, the flying teenage warrior rabbit who eats Frumpies in every episode, may respond differently to the new flavor. How would you take that into account in your design?

22. Wine marketing. A 2001 Danish study published in the *Archives of Internal Medicine* casts significant doubt on suggestions that adults who drink wine have higher levels of "good" cholesterol and fewer heart attacks. These researchers followed a group of individuals born at a Copenhagen hospital between 1959 and 1961 for 40 years. Their study found that in this group the adults who drank wine were richer and better educated than those who did not.

a) What kind of study was this?
b) It is generally true that people with high levels of education and high socioeconomic status are healthier than

others. How does this call into question the supposed health benefits of wine?

c) Can studies such as these prove causation (that wine helps prevent heart attacks, that drinking wine makes one richer, that being rich helps prevent heart attacks, etc.)? Explain.

23. SAT prep courses. Can special study courses actually help raise SAT scores? One organization says that the 30 students they tutored achieved an average gain of 60 points when they retook the test.

a) Explain why this does not necessarily prove that the special course caused the scores to go up.

b) Propose a design for an experiment that could test the effectiveness of the tutorial course.

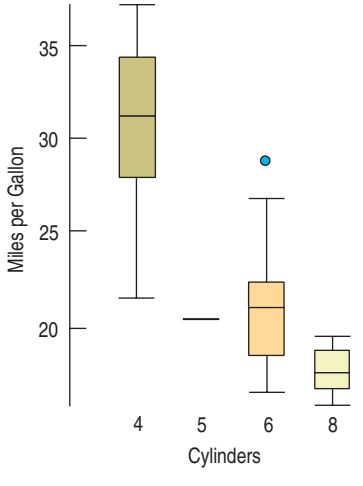
c) Suppose you suspect that the tutorial course might be more helpful for students whose initial scores were particularly low. How would this affect your proposed design?

24. Safety switch. An industrial machine requires an emergency shutoff switch that must be designed so that it can be easily operated with either hand. Design an experiment to find out whether workers will be able to deactivate the machine as quickly with their left hands as with their right hands. Be sure to explain the role of randomization in your design.

T **25. Cars (fuel efficiency).** These boxplots show the relationship between the number of cylinders in a car's engine and its fuel economy from a study conducted by a major car manufacturer.

a) What are the null and alternative hypotheses? Talk about cars and fuel efficiency, not symbols.

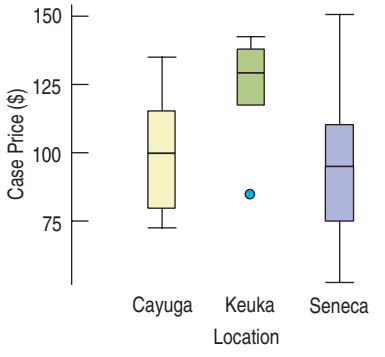
b) Do the conditions for an ANOVA seem to be met here? Why or why not?



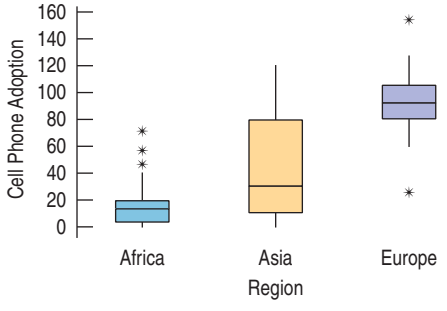
T **26. Wine production.** The boxplots display case prices (in dollars) of wine produced by wineries along three of the Finger Lakes in upstate New York.

a) What are the null and alternative hypotheses? Talk about prices and location, not symbols.

b) Do the conditions for an ANOVA seem to be met here? Why or why not?



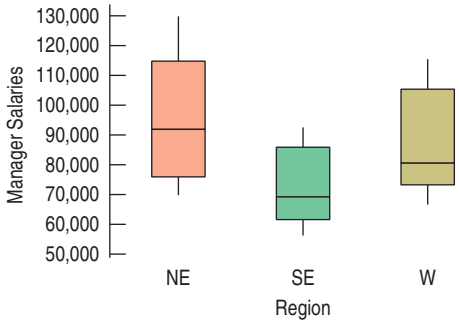
27. Cell phone adoption. Cell phone adoption rates are available for various countries in the United United Nations Database (unstats.un.org). Countries were randomly selected from three regions (Africa, Asia, and Europe), and cell phone adoption (per 100 inhabitants) rates retrieved. The boxplots display the data.



a) What are the null and alternative hypotheses (in words, not symbols)?

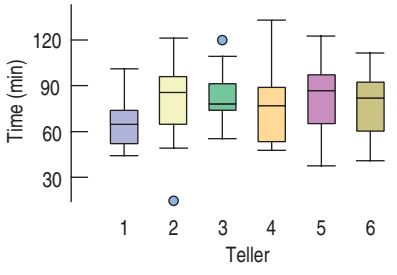
b) Are the conditions for ANOVA met? Why or why not?

T **28. Marketing managers' salaries.** A sample of eight states was selected randomly from each of three regions in the United States (Northeast, Southeast, and West). Mean annual salaries for marketing managers were retrieved from the U.S. Bureau of Labor Statistics (data.bls.gov/oes). The boxplots display the data.



a) What are the null and alternative hypotheses (in words, not symbols)?
b) Are the conditions for ANOVA met? Why or why not?

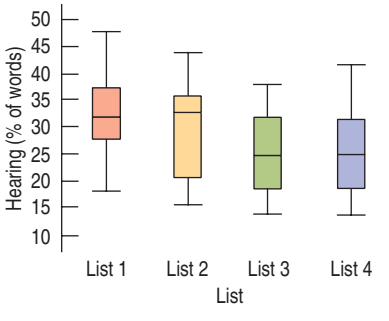
29. Bank tellers. A bank is studying the average time that it takes 6 of its tellers to serve a customer. Customers line up in the queue and are served by the next available teller. Here is a boxplot of the times it took to serve the last 140 customers.



Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Teller	5	3315.32	663.064	1.508	0.1914
Error	134	58919.1	439.695		
Total	139	62234.4			

a) What are the null and alternative hypotheses?
b) What do you conclude?
c) Would it be appropriate to run a multiple comparisons test (for example, a Bonferroni test) to see which tellers differ from each other? Explain.

T **30. Hearing.** Vendors of hearing aids test them by having patients listen to lists of words and repeat what they hear. The word lists are supposed to be equally difficult to hear accurately. But the challenge of hearing aids is perception when there is background noise. A researcher investigated four different word lists used in hearing assessment (Loven, 1981). She wanted to know whether the lists were equally difficult to understand in the presence of a noisy background. To find out, she tested 24 subjects with normal hearing and measured the number of words perceived correctly in the presence of background noise. Here are the boxplots of the four lists.



Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
List	3	920.4583	306.819	4.9192	0.0033
Error	92	5738.1667	62.371		
Total	95	6658.6250			

a) What are the null and alternative hypotheses?
b) What do you conclude?
c) Would it be appropriate to run a multiple comparisons test (for example, a Bonferroni test) to see which lists differ from each other in terms of mean percent correct? Explain.

31. E-security. A report released by the Pew Internet & American Life Project entitled *The Internet & Consumer Choice* focused on current online issues. Respondents were asked to indicate their level of agreement (1 = strongly agree to 4 = strongly disagree) with a variety of statements including "I don't like giving my credit card number or personal information online." A part of the data set was used to determine whether the type of community in which the individual resides (Urban, Suburban, or Rural) affected responses. Here are the results in the form of a partially completed Analysis of Variance table.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Community	2	6.615			
Error	183	96.998			
Total	185	103.613			

a) Is this an experimental or observational study? Explain.
b) Is this a prospective or retrospective study? Explain.
c) State the null and alternative hypothesis.
d) Calculate the *F*-statistic.
e) The P-value for this statistic turns out to be 0.002. State the conclusion. Can a causal link be established? Explain.

32. Internet usage. Internet usage rates are available for various countries in the United Nations Common Database (unstats. un.org). Countries were randomly selected from three regions (Africa, Asia, and Europe), and Internet usage (per 100 inhabitants) data from 2005 were retrieved. The data were analyzed to determine if Internet usage rates were the same across regions. The partially completed Analysis of Variance table is shown here.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Region	2	21607			
Error	93	20712			
Total	95	42319			

a) Is this an experimental or observational study? Explain.
b) Is this a prospective or retrospective study? Explain.
c) State the null and alternative hypotheses.
d) Calculate the *F*-statistic.
e) The P-value for this statistic turns out to be <0.001. State the conclusion. Can a causal link be established? Explain.

33. Colorful experiment? In a recent article published in *Quality Progress*, each student in a statistics class had a randomly assigned bag of peanut M&M's® and counted the number of each color (*Blue*, *Red*, *Orange*, *Green*, *Brown*, *Yellow*). The bags were all the same size (1.74 ounces). The investigators claimed to use a randomized block design, with *Bag* as the blocking factor. They counted the number of candies of each color in each bag. Their results are reproduced here (Lin, T., and Sanders, M.S., "A Sweet Way to Learn DOE," *Quality Progress*, Feb. 2006, p. 88).

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-ratio	P-value
Bag	13	4.726	0.364	0.10	1.000
Color	5	350.679	70.136	18.72	<0.001
Error	65	243.488	3.746		
Total	83	598.893			

a) Was this an observational or experimental study?
b) What was the treatment? What factors were manipulated?
c) What was the response variable?

34. Six Sigma training. A large financial institution is interested in training its college educated workforce in Six Sigma principles and methods. One part of the training involves basic statistical concepts and tools. Management is considering three approaches: *online*, *traditional classroom*, and *hybrid (a mix of both)*. Prior to launching the program throughout the entire organization, they decided to pilot test the three approaches. Because they believed that educational background may affect the results, they selected 3 employees from each of 10 different college major programs of study (*liberal arts*, *accounting*, *economics*, *management*, *marketing*, *finance*, *information systems*, *computer science*, *operations*, *other*), and randomly assigned each to one of the three approaches. At the end of training, each participant took an exam. The results are shown here.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-ratio	P-value
Major	9	2239.47	248.830	21.69	<0.001
Training	2	171.47	85.735	7.47	0.004
Error	18	206.53	11.474		
Total	29	2617.47			

a) Was this an observational study or an experiment?
b) What was the purpose of using Major as a blocking factor?
c) Given the results, was it necessary to use Major as a blocking factor? Explain.
d) State the conclusion from this analysis.

35. E-trust. Online retailers want customers to trust their websites and want to alleviate any concerns potential customers may have about privacy and security. In a study investigating the factors that affect e-trust, participants were randomly assigned to carry out online transactions on fictitious retailers' websites. The sites were configured in one of three ways: (1) *with a third-party assurance seal (e.g., BBBOnLine) displayed;* (2) *a self-proclaimed assurance displayed;* or (3) *no assurance*. In addition, participants made a transaction involving one of three products (*book*, *camera*, or *insurance*). These products represent varying degrees of risk. After completing the transaction, they rated how "trustworthy" the website was on a scale of 1 (not at all) to 10 (extremely trustworthy).

a) Is this an experiment or an observational study? Explain.
b) What is the response variable?
c) How many factors are involved?
d) How many treatments are involved?
e) State the hypotheses (in words, not symbols).

36. Injection molding. In order to improve the quality of molded parts, companies often test different levels of parameter settings in order to find the best combinations. Injection molding machines typically have many adjustable parameters. One company used three different mold temperatures (25, 35, and 45 degrees Celsius) and four different cooling times (10, 15, 20, and 25 minutes) to examine how they affect the tensile strength of the resulting molded parts. Five parts were randomly sampled and measured from each treatment combination.

a) Is this an experiment or an observational study? Explain.
b) What is the response variable?
c) What are the factors?
d) How many treatments are involved?
e) State the hypotheses (in words, not symbols).

37. Stock returns. Companies that are ISO 9000 certified have met standards that ensure they have a quality management system committed to continuous improvement. Going through the certification process generally involves a substantial investment that includes the hiring of external auditors. A group of such auditors, wishing to "prove" that ISO 9000 certification pays off, randomly selected a sample of small and large companies with and without ISO 9000 certification. Size was based on the number of employees. They computed the % change in closing stock price from August 2006 to August 2007. The two-way ANOVA results are presented here (data obtained from *Yahoo! Finance*).

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
ISO 9000	1	2654.4	2654.41	5.78	0.022
Size	1	0.2	0.18	0.004	0.984
Interaction	1	1505.5	1505.49	3.28	0.079
Error	36	16545.9	459.61		
Total	39	20705.9			

a) Is this an experiment or an observational study?
b) State the hypotheses.
c) Given the small P-value associated with the ISO 9000 factor and that the mean annual return for the companies

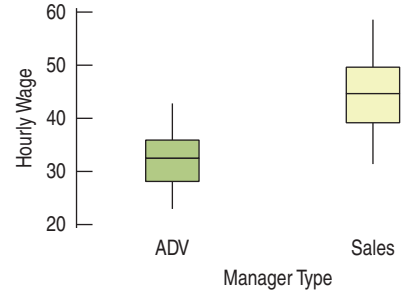
with ISO 9000 is 30.7% compared to 14.4% for those without, the auditors state that achieving ISO 9000 certification results in higher stock prices. Do you agree with their statement? Explain.

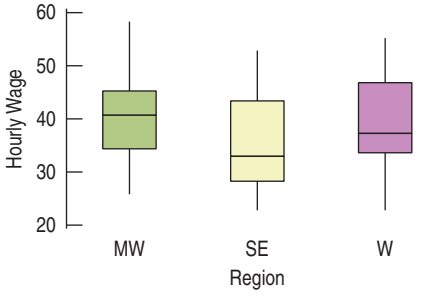
38. Company bonuses. After complaints about gender discrimination regarding bonus incentive pay, a large multinational firm collected data on bonuses awarded during the previous year (% of base pay). Human Resources (HR) randomly sampled male and female managers from three different levels: *senior*, *middle*, and *supervisory*. The two-way ANOVA results are presented here.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Gender	1	32.033	32.033	9.76	0.005
Level	2	466.200	233.100	70.99	0.000
Interaction	2	20.467	10.233	3.12	0.063
Error	24	78.800	3.283		
Total	29	597.500			

a) Is this an experiment or an observational study?
b) State the hypotheses.
c) Given the small P-value associated with the gender and that the mean annual bonus percent for females is 12.5% compared to 14.5% for males, HR concludes that gender discrimination exists. Do you agree? Explain.

39. Managers' hourly wages. What affects marketing managers' hourly wages? In order to find out, mean hourly wages were retrieved from the U.S. Bureau of Labor Statistics for two managerial occupations in marketing (*Sales managers*, *Advertising managers*) for a random sample of states from three regions (*Midwest*, *Southeast*, *West*) (www.bls.gov/data/#wages). Here are boxplots showing mean hourly wages for the two marketing occupations and the three regions as well as the results for a two-way ANOVA.

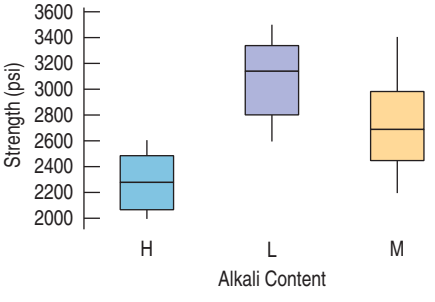


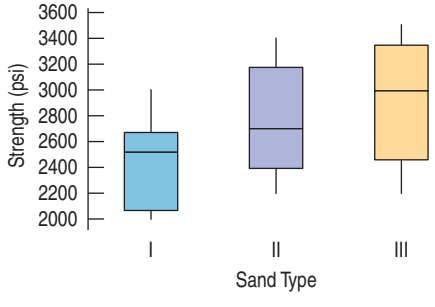


Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Manager Type	1	1325.93	1325.93	31.84	0.000
Region	2	153.55	76.78	1.84	0.176
Interaction	2	32.74	16.37	0.39	0.678
Error	30	1249.32	41.64		
Total	35	2761.55			

a) Is this an experiment or an observational study? Explain.
b) Are the conditions for two-way ANOVA met?
c) If so, perform the hypothesis tests and state your conclusions in terms of hourly wages, occupational type, and region.
d) Is it appropriate to interpret the main effects in this case? Explain.

40. Concrete testing. A company that specializes in developing concrete for construction strives to continually improve the properties of its materials. In order to increase the compressive strength of one of its new formulations, they varied the amount of alkali content (*low*, *medium*, *high*). Since the type of sand used may also affect the strength of concrete, they used three different types of sand (Types I, II, III). Four samples were randomly selected from each treatment combination to be tested. The boxplots show the test results on compressive strength (in psi) for the three levels of alkali content and three types of sand. Two-way ANOVA results are also given.





Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Alkali Content	2	4016600	2008300	46.38	0.000
Sand Type	2	1547817	773908	17.87	0.000
Interaction	4	177533	44383	1.02	0.412
Error	27	1169250	43306		
Total	35	6911200			

a) Is this an experiment or an observational study? Explain.
b) Are the conditions for two-way ANOVA met?
c) If so, perform the hypothesis tests and state your conclusions in terms of compressive strength, alkali content, and sand type.
d) Is it appropriate to interpret the main effects in this case? Explain.

41. Production problems. A manufacturing company that makes dental drills was experiencing problems with a specific part on the production line. Management suspected a machining problem that resulted in the length of the part to vary outside of target specification. Two factors were examined: the machine setting (at three levels) and the shift (morning, afternoon, and night). New hires were typically scheduled for night shift, and management believed that their relative inexperience may also be contributing to the variation. Three parts were randomly selected and measured from each treatment combination. The deviation from specified size was measured in microns. The data and two-way ANOVA results are shown.

Size Error	Machine Setting	Shift
1.1	1	Morning
3.6	2	Morning
3.3	3	Morning
2.1	1	Morning
0.9	2	Morning
2.6	3	Morning
0.6	1	Morning
2.3	2	Morning
3.2	3	Morning
2	1	Afternoon
2.4	2	Afternoon
5	3	Afternoon
1.8	1	Afternoon
4.3	2	Afternoon
3.2	3	Afternoon
2.5	1	Afternoon
5	2	Afternoon
2.3	3	Afternoon
3.8	1	Night
5.5	2	Night
5	3	Night
2.9	1	Night
6.7	2	Night
5.8	3	Night
2.8	1	Night
3	2	Night
5.3	3	Night

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
MachSet	2	17.1119	8.55593	7.3971	0.0045
Shift	2	24.9607	12.4804	10.790	0.0008
Interaction	4	1.4970	0.374259	0.32357	0.8585
Error	18	20.8200	1.15667		
Total	26	64.3896			

a) Is this an experiment or an observational study? Explain.
b) What is the response variable?
c) How many treatments are involved?
d) Based on the two-way ANOVA results, management concluded that shift has a significant impact on the length of the part and that consequently operator inexperience is the root cause of the part problems. Do you agree with this conclusion? Explain.

42. Process improvements. One way to improve a process is to eliminate non–value-added activities (e.g., extra movements) and wasted effort (e.g., looking for materials). A consultant was hired to improve the efficiency in a large shop floor operation. She tested three different workspace designs and two different storage/retrieval systems. She measured process flow time for three randomly selected operations through each of the combinations of workspace design and storage/retrieval systems. The data and two-way ANOVA results are shown here.

Workspace Design	Storage System	Flow Time (Days)
1	1	4.5
2	1	3.3
3	1	3.4
1	1	4.0
2	1	3.0
3	1	2.9
1	1	4.2
2	1	3.0
3	1	3.2
1	1	4.5
2	1	3.5
3	1	3.2
1	1	3.8
2	1	2.8
3	1	3.0
1	2	3.0
2	2	3.8
3	2	3.6
1	2	2.8
2	2	4.0
3	2	3.5
1	2	3.0
2	2	3.5
3	2	3.8
1	2	4.0
2	2	4.2
3	2	4.2
1	2	3.0
2	2	3.6
3	2	3.8

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Workspace					
Design	2	0.30867	0.15433	1.56	0.230
Storage					
System	1	0.07500	0.07500	0.76	0.392
Interaction	2	4.87800	2.43900	24.72	<0.001
Error	24	2.36800	0.09867		
Total	29	7.62967			

a) Is this an experiment or observational study? Explain.
b) What is the response variable?
c) How many treatments are involved?
d) Based on the two-way ANOVA results, management concludes that neither the workspace design nor the storage/retrieval system impacts process flow time (and that the consultant wasn't worth the money). Do you agree with this conclusion? Explain.

43. **Yogurt research.** An experiment to determine the effect of several methods of preparing cultures for use in commercial yogurt was conducted by a food science research group. Three batches of yogurt were prepared using each of three methods: traditional, ultrafiltration, and reverse osmosis. A trained expert then tasted each of the 9 samples, presented in random order, and judged them on a scale from 1 to 10. A partially complete Analysis of Variance table of the data follows.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Treatment	17.300			
Residual	0.460			
Total	17.769			

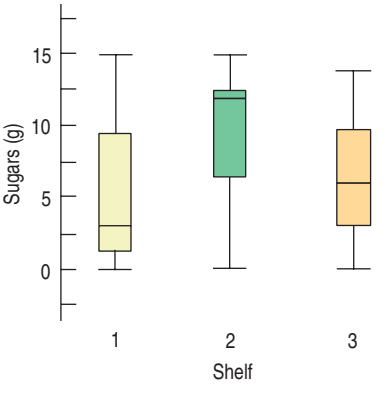
a) Calculate the mean square of the treatments and the mean square of the error.
b) Form the *F*-statistic by dividing the two mean squares.
c) The P-value of this *F*-statistic turns out to be 0.000017. What does this say about the null hypothesis of equal means?
d) What assumptions have you made in order to answer part c?
e) What would you like to see in order to justify the conclusions of the *F*-test?
f) What is the average size of the error standard deviation in the judge's assessment?

44. **Smokestack scrubbers.** Particulate matter is a serious form of air pollution often arising from industrial production. One way to reduce the pollution is to put a filter, or scrubber, at the end of the smokestack to trap the particulates. An experiment to determine which smokestack scrubber design is best was run by placing four scrubbers of different designs on an industrial stack in random order. Each scrubber was tested 5 times. For each run, the same material was produced, and the particulate emissions coming out of the scrubber were measured (in parts per billion). A partially complete Analysis of Variance table of the data is shown here.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Treatment	81.2			
Residual	30.8			
Total	112.0			

a) Calculate the mean square of the treatments and the mean square of the error.
b) Form the *F*-statistic by dividing the two mean squares.
c) The P-value of this *F*-statistic turns out to be 0.00000949. What does this say about the null hypothesis of equal means?
d) What assumptions have you made in order to answer part c?
e) What would you like to see in order to justify the conclusions of the *F*-test?
f) What is the average size of the error standard deviation in particulate emissions?

T **45.** **Cereal shelf placement.** Supermarkets often place similar types of cereal on the same supermarket shelf. The shelf placement for 77 cereals was recorded as their sugar content. Does sugar content vary by shelf? Here's a boxplot and an ANOVA table.



Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Shelf	2	248.4079	124.204	7.3345	0.0012
Error	74	1253.1246	16.934		
Total	76	1501.5325			

Level	*n*	Mean	StdDev
1	20	4.80000	4.57223
2	21	9.61905	4.12888
3	36	6.52778	3.83582

a) What kind of design or study is this?
b) What are the null and alternative hypotheses?
c) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of sugar content and shelf placement.)

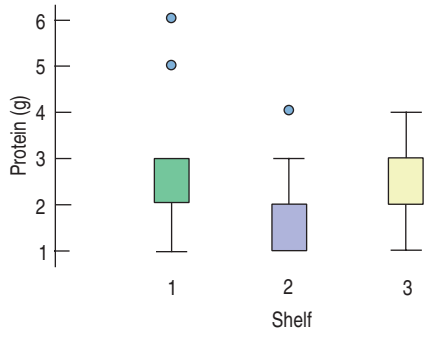
Dependent Variable: SUGARS (Exercise 45)

	(I) SHELF	(J) SHELF	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
Bonferroni						Lower Bound	Upper Bound
	1	2	−4.819	1.2857	0.001	−7.969	−1.670
		3	−1.728	1.1476	0.409	−4.539	1.084
	2	1	4.819	1.2857	0.001	1.670	7.969
		3	3.091	1.1299	0.023	0.323	5.859
	3	1	1.728	1.1476	0.409	−1.084	4.539
		2	−3.091	1.1299	0.023	−5.859	−0.323

d) Can we conclude that cereals on shelf 2 have a different mean sugar content than cereals on shelf 3? Can we conclude that cereals on shelf 2 have a different mean sugar content than cereals on shelf 1? What can we conclude?

e) To check for significant differences between the shelf means, we can use a Bonferroni test, whose results are shown here. For each pair of shelves, the difference is shown along with its standard error and significance level. What does it say about the questions in part d?

T 46. Cereal shelf placement, part 2. We also have data on the protein content on the 77 cereals in Exercise 45. Does protein content vary by shelf? Here's a boxplot and an ANOVA table.



Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Shelf	2	12.4258	6.2129	5.8445	0.0044
Error	74	78.6650	1.0630		
Total	76	91.0909			

Level	*n*	Mean	StdDev
1	20	2.65000	1.46089
2	21	1.90476	0.99523
3	36	2.86111	0.72320

a) What kind of design or study is this?

b) What are the null and alternative hypotheses?

c) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of protein content and shelf placement.)

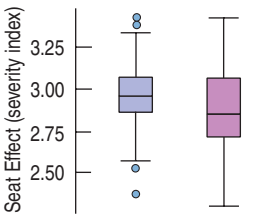
d) Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 3? Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 1? What can we conclude?

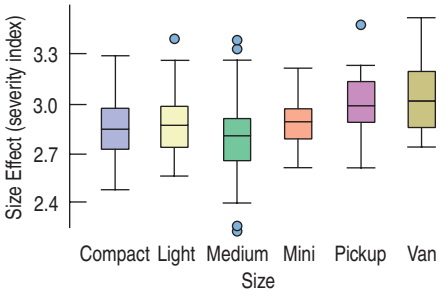
e) To check for significant differences between the shelf means, we can use a Bonferroni test, whose results are shown here. For each pair of shelves, the difference is shown along with its standard error and significance level. What does it say about the questions in part d?

Dependent Variable: PROTEIN
Bonferroni

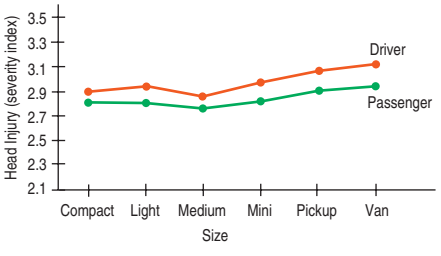
(I) SHELF	(J) SHELF	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	0.75	0.322	0.070	−0.04	1.53
	3	−0.21	0.288	1.000	−0.92	0.49
2	1	−0.75	0.322	0.070	−1.53	0.04
	3	−0.96	0.283	0.004	−1.65	−0.26
3	1	0.21	0.288	1.000	−0.49	0.92
	2	0.96	0.283	0.004	0.26	1.65

T 47. Automotive safety. The National Highway Transportation Safety Administration runs crash tests in which stock automobiles are crashed into a wall at 35 mph with dummies in both the passenger and the driver's seats. The THOR Alpha crash dummy is capable of recording 134 channels of data on the impact of the crash at various sites on the dummy. In this test, 335 cars are crashed. The response variable is a measure of head injury. Researchers want to know whether the seat the dummy is sitting in affects head injury severity, as well as whether the type of car affects severity. Here are boxplots for the 2 different *Seats* (*driver*, *passenger*) and the 6 different *Size* classifications (*compact*, *light*, *medium*, *mini*, *pickup*, *van*).

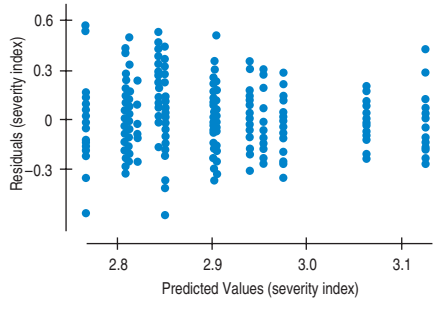




An interaction plot shows:



A scatterplot of residuals vs. predicted values shows:

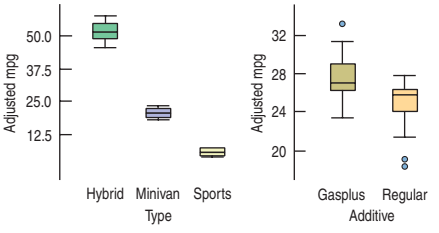


The ANOVA table follows:

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Seat	1	0.88713	0.88713	25.501	<0.0001
Size	5	1.49253	0.29851	8.581	<0.0001
Seat × Size	5	0.07224	0.01445	0.415	0.838
Error	282	9.8101	0.03479		
Total	293	12.3853			

a) State the null hypotheses about the main effects (in words, not symbols).
b) Are the conditions for two-way ANOVA met?
c) If so, perform the hypothesis tests and state your conclusion. Be sure to state it in terms of head injury severity, seats, and vehicle types.

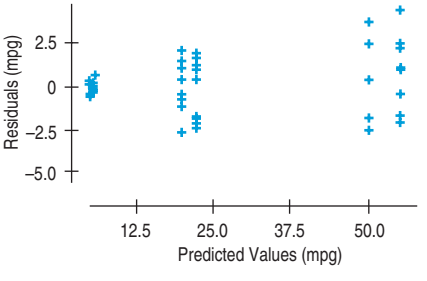
48. **Gas additives.** An experiment to test a new gasoline additive, *Gasplus*, was performed on three different cars: a sports car, a minivan, and a hybrid. Each car was tested with both *Gasplus* and regular gas on 10 different occasions, and their gas mileage was recorded. Here are the boxplots.



A two-way ANOVA with interaction model was run, and the following ANOVA table resulted.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Type	2	23175.4	11587.7	2712.2	<0.0001
Additive	1	92.1568	92.1568	21.57	<0.0001
Interaction	2	51.8976	25.9488	6.0736	0.0042
Error	54	230.711	4.27242		
Total	59	23550.2			

A plot of the residuals vs. predicted values showed:



What conclusions about the additive and car types do you draw? Do you see any potential problems with the analysis?

Just Checking Answers

1. Gather reports from veterinarians and pet hospitals. Look into the histories of sick animals. This would be a retrospective observational study.
2. Treatment: Feed the new food and a previously tested, known to be safe, food to pets.

   Response: Judge the health of the animals, possibly by having a veterinarian examine them before and after the feeding trials.
3. Control by choosing similar animals. Perhaps choose just one breed and test animals of the same age and health. Treat them otherwise the same in terms of exercise, attention, and so on.

   Randomize by assigning animals to treatments at random.

   Replicate by having more than one animal fed each formulation.
4. A control group could be fed a standard laboratory food, if we have one known to be safe. Otherwise we could prepare a special food in our test kitchens to be certain of its safety.
5. The veterinarian evaluating the animals should be blind to the treatments. For double-blinding, all technicians handling the animals should also be blinded. That would require making the control food look as much like the test food as possible.
6. Yes. Test dogs and cats separately.
7. No. We have failed to reject the hypothesis of a difference, but that's all we can conclude. There is insufficient evidence to discern any difference. But that should be sufficient for the company's purposes.

This page intentionally left blank