CK-12 Probability and Statistics - Advanced (Second Edition)

Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

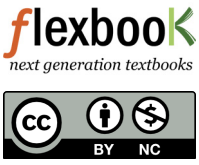
Chapter 6

Planning and Conducting an Experiment or Study

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Printed: December 17, 2014



AUTHORS
Ellen Lawsky
Larry Ottman
Raja Almukkahal
Brenda Meery
Danielle DeLancey

CHAPTER 6

Planning and Conducting an Experiment or Study

Chapter Outline

6.1 Surveys and Sampling

Learning Objectives

- Differentiate between a census and a survey or sample.
- Distinguish between sampling error and bias.
- Identify and name potential sources of bias from both real and hypothetical sampling situations.

Introduction

The New York Times/CBS News Poll is a well-known regular polling organization that releases results of polls taken to help clarify the opinions of Americans on pending elections, current leaders, or economic or foreign policy issues. In an article entitled "How the Poll Was Conducted" that explains some of the details of a recent poll, the following statements appear[1]:

"In theory, in 19 cases out of 20, overall results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking to interview all American adults."

"In addition to sampling error, the practical difficulties of conducting any survey of public opinion may introduce other sources of error into the poll. Variation in the wording and order of questions, for example, may lead to somewhat different results."

These statements illustrate two different potential problems with opinion polls, surveys, observational studies, and experiments. In this chapter, we will investigate these problems and more by looking at sampling in detail.

Census vs. Sample

A *sample* is a representative subset of a population. If a statistician or other researcher wants to know some information about a population, the only way to be truly sure is to conduct a census. In a *census*, every unit in the population being studied is measured or surveyed. In opinion polls, like the *New York Times* poll mentioned above, results are generalized from a sample. If we really wanted to know the true approval rating of the president, for example, we would have to ask every single American adult his or her opinion. There are some obvious reasons why a census is impractical in this case, and in most situations.

First, it would be extremely expensive for the polling organization. They would need an extremely large workforce to try and collect the opinions of every American adult. Also, it would take many workers and many hours to organize, interpret, and display this information. Even if it could be done in several months, by the time the results were published, it would be very probable that recent events had changed peoples' opinions and that the results would be obsolete.

In addition, a census has the potential to be destructive to the population being studied. For example, many manufacturing companies test their products for quality control. A padlock manufacturer might use a machine to see how much force it can apply to the lock before it breaks. If they did this with every lock, they would have

none left to sell! Likewise, it would not be a good idea for a biologist to find the number of fish in a lake by draining the lake and counting them all!

The U.S. Census is probably the largest and longest running census, since the Constitution mandates a complete counting of the population. The first U.S. Census was taken in 1790 and was done by U.S. Marshalls on horseback. Taken every 10 years, a Census was conducted in 2010, and in a report by the Government Accountability Office in 1994, was estimated to cost $11 billion. This cost has recently increased as computer problems have forced the forms to be completed by hand[3]. You can find a great deal of information about the U.S. Census, as well as data from past Censuses, on the Census Bureau's website: http://www.census.gov/ .

Due to all of the difficulties associated with a census, sampling is much more practical. However, it is important to understand that even the most carefully planned sample will be subject to random variation between the sample and the population. Recall that these differences due to chance are called *sampling error*. We can use the laws of probability to predict the level of accuracy in our sample. Opinion polls, like the *New York Times* poll mentioned in the introduction, tend to refer to this as *margin of error*. The second statement quoted from the *New York Times* article mentions another problem with sampling. That is, it is often difficult to obtain a sample that accurately reflects the total population. It is also possible to make mistakes in selecting the sample and collecting the information. These problems result in a non-representative sample, or one in which our conclusions differ from what they would have been if we had been able to conduct a census.

To help understand these ideas, consider the following theoretical example. A coin is considered fair if the probability, p, of the coin landing on heads is the same as the probability of it landing on tails ($p = 0.5$). The probability is defined as the proportion of heads obtained if the coin were flipped an infinite number of times. Since it is impractical, if not impossible, to flip a coin an infinite number of times, we might try looking at 10 samples, with each sample consisting of 10 flips of the coin. Theoretically, you would expect the coin to land on heads 50% of the time, but it is very possible that, due to chance alone, we would experience results that differ from this. These differences are due to sampling error. As we will investigate in detail in later chapters, we can decrease the sampling error by increasing the sample size (or the number of coin flips in this case). It is also possible that the results we obtain could differ from those expected if we were not careful about the way we flipped the coin or allowed it to land on different surfaces. This would be an example of a non-representative sample.

At the following website, you can see the results of a large number of coin flips: http://www.mathsonline.co.uk/nonmembers/resource/prob/coins.html . You can see the random variation among samples by asking for the site to flip 10 coins 10 times. Our results for that experiment produced the following numbers of heads: 3, 3, 4, 4, 4, 4, 5, 6, 6, 6. This seems quite strange, since the expected number is 5. How do your results compare?

Bias in Samples and Surveys

The term most frequently applied to a non-representative sample is *bias*. Bias has many potential sources. It is important when selecting a sample or designing a survey that a statistician make every effort to eliminate potential sources of bias. In this section, we will discuss some of the most common types of bias. While these concepts are universal, the terms used to define them here may be different than those used in other sources.

Sampling Bias

In general, sampling bias refers to the methods used in selecting the sample. The *sampling frame* is the term we use to refer to the group or listing from which the sample is to be chosen. If you wanted to study the population of students in your school, you could obtain a list of all the students from the office and choose students from the list. This list would be the sampling frame.

Incorrect Sampling Frame

If the list from which you choose your sample does not accurately reflect the characteristics of the population, this is called *incorrect sampling frame*. A sampling frame error occurs when some group from the population does not have the opportunity to be represented in the sample. For example, surveys are often done over the telephone. You could use the telephone book as a sampling frame by choosing numbers from the telephone book. However, in addition to the many other potential problems with telephone poles, some phone numbers are not listed in the telephone book. Also, if your population includes all adults, it is possible that you are leaving out important groups of that population. For example, many younger adults in particular tend to only use their cell phones or computer-based phone services and may not even have traditional phone service. Even if you picked phone numbers randomly, the sampling frame could be incorrect, because there are also people, especially those who may be economically disadvantaged, who have no phone. There is absolutely no chance for these individuals to be represented in your sample. A term often used to describe the problems when a group of the population is not represented in a survey is *undercoverage*. Undercoverage can result from all of the different sampling biases.

One of the most famous examples of sampling frame error occurred during the 1936 U.S. presidential election. The Literary Digest, a popular magazine at the time, conducted a poll and predicted that Alf Landon would win the election that, as it turned out, was won in a landslide by Franklin Delano Roosevelt. The magazine obtained a huge sample of ten million people, and from that pool, 2 million replied. With these numbers, you would typically expect very accurate results. However, the magazine used their subscription list as their sampling frame. During the depression, these individuals would have been only the wealthiest Americans, who tended to vote Republican, and left the majority of typical voters under-covered.

Convenience Sampling

Suppose your statistics teacher gave you an assignment to perform a survey of 20 individuals. You would most likely tend to ask your friends and family to participate, because it would be easy and quick. This is an example of *convenience sampling*, or convenience bias. While it is not always true, your friends are usually people who share common values, interests, and opinions. This could cause those opinions to be over-represented in relation to the true population. Also, have you ever been approached by someone conducting a survey on the street or in a mall? If such a person were just to ask the first 20 people they found, there is the potential that large groups representing various opinions would not be included, resulting in undercoverage.

Judgment Sampling

Judgment sampling occurs when an individual or organization that is usually considered an expert in the field being studied chooses the individuals or group of individuals to be used in the sample. Because it is based on a subjective choice, even by someone considered an expert, it is very susceptible to bias. In some sense, this is what those responsible for the Literary Digest poll did. They incorrectly chose groups they believed would represent the population. If a person wants to do a survey on middle-class Americans, how would this person decide who to include? It would be left to this person's own judgment to create the criteria for those considered middle-class. This individual's judgment might result in a different view of the middle class that might include wealthier individuals that others would not consider part of the population. Similar to judgment sampling, in *quota sampling*, an individual or

organization attempts to include the proper proportions of individuals of different subgroups in their sample. While it might sound like a good idea, it is subject to an individual's prejudice and is, therefore, prone to bias.

Size Bias

If one particular subgroup in a population is likely to be over-represented or under-represented due to its size, this is sometimes called *size bias*. If we chose a state at random from a map by closing our eyes and pointing to a particular place, larger states would have a greater chance of being chosen than smaller ones. As another example, suppose that we wanted to do a survey to find out the typical size of a student's math class at a school. The chances are greater that we would choose someone from a larger class for our survey. To understand this, say that you went to a very small school where there are only four math classes, with one class having 35 students, and the other three classes having only 8 students. If you simply choose students at random, it is more likely you will select students for your sample who will will say the typical size of a math class is 35, since there are more students in the larger class.

Here's one more example: a person driving on an interstate highway tends to say things like, "Wow, I was going the speed limit, and everyone was just flying by me." The conclusion this person is making about the population of all drivers on this highway is that most of them are traveling faster than the speed limit. This may indeed be true, but let's say that most people on the highway, along with our driver, really are abiding by the speed limit. In a sense, the driver is collecting a sample, and only those few who are close to our driver will be included in the sample. There will be a larger number of drivers going faster in our sample, so they will be over-represented. As you may already see, these definitions are not absolute, and often in a practical example, there are many types of overlapping bias that could be present and contribute to overcoverage or undercoverage. We could also cite incorrect sampling frame or convenience bias as potential problems in this example.

Response Bias

The term *response bias* refers to problems that result from the ways in which the survey or poll is actually presented to the individuals in the sample.

Voluntary Response Bias

Television and radio stations often ask viewers/listeners to call in with opinions about a particular issue they are covering. The websites for these and other organizations also usually include some sort of online poll question of the day. Reality television shows and fan balloting in professional sports to choose all-star players make use of these types of polls as well. All of these polls usually come with a disclaimer stating that, "This is not a scientific poll." While perhaps entertaining, these types of polls are very susceptible to *voluntary response bias*. The people who respond to these types of surveys tend to feel very strongly one way or another about the issue in question, and the results might not reflect the overall population. Those who still have an opinion, but may not feel quite so passionately about the issue, may not be motivated to respond to the poll. This is especially true for phone-in or mail-in surveys in which there is a cost to participate. The effort or cost required tends to weed out much of the population in favor of those who hold extremely polarized views. A news channel might show a report about a child killed in a drive-by shooting and then ask for people to call in and answer a question about tougher criminal sentencing laws. They would most likely receive responses from people who were very moved by the emotional nature of the story and wanted anything to be done to improve the situation. An even bigger problem is present in those types of polls in which there is no control over how many times an individual may respond.

Non-Response Bias

One of the biggest problems in polling is that most people just don't want to be bothered taking the time to respond to a poll of any kind. They hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly

past an interviewer on the street. We just don't know how much these individuals' beliefs and opinions reflect those of the general population, and, therefore, almost all surveys could be prone to *non-response bias*.

Questionnaire Bias

Questionnaire bias occurs when the way in which the question is asked influences the response given by the individual. It is possible to ask the same question in two different ways that would lead individuals with the same basic opinions to respond differently. Consider the following two questions about gun control.

"Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?"

"Do you believe that it is reasonable for the government to infringe on an individual's constitutional right to bear arms?"

A gun rights activist might feel very strongly that the government should never be in the position of limiting guns in any way and would answer no to both questions. Someone who is very strongly against gun ownership, on the other hand, would probably answer yes to both questions. However, individuals with a more tempered, middle position on the issue might believe in an individual's right to own a gun under some circumstances, while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the wording of a question to obtain a certain response to a poll question. Questionnaire bias is not necessarily always a deliberate action. If a question is poorly worded, confusing, or just plain hard to understand, it could lead to non-representative results. When you ask people to choose between two options, it is even possible that the order in which you list the choices may influence their response!

Incorrect Response Bias

A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individual intentionally responds to a survey with an untruthful answer, this is called *incorrect response bias*. This can occur when asking questions about extremely sensitive or personal issues. For example, a survey conducted about illegal drinking among teens might be prone to this type of bias. Even if guaranteed their responses are confidential, some teenagers may not want to admit to engaging in such behavior at all. Others may want to appear more rebellious than they really are, but in either case, we cannot be sure of the truthfulness of the responses.

Another example is related to the donation of blood. Because the dangers of donated blood being tainted with diseases carrying a negative social stereotype increased in the 1990's, the Red Cross has recently had to deal with incorrect response bias on a constant and especially urgent basis. Individuals who have engaged in behavior that puts them at risk for contracting AIDS or other diseases have the potential to pass these diseases on through donated blood[4]. Screening for at-risk behaviors involves asking many personal questions that some find awkward or insulting and may result in knowingly false answers. The Red Cross has gone to great lengths to devise a system with several opportunities for individuals giving blood to anonymously report the potential danger of their donation.

In using this example, we don't want to give the impression that the blood supply is unsafe. According to the Red Cross, "Like most medical procedures, blood transfusions have associated risk. In the more than fifteen years since March 1985, when the FDA first licensed a test to detect HIV antibodies in donated blood, the Centers for Disease Control and Prevention has reported only 41 cases of AIDS caused by transfusion of blood that tested negative for the AIDS virus. During this time, more than 216 million blood components were transfused in the United States. The tests to detect HIV were designed specifically to screen blood donors. These tests have been regularly upgraded since they were introduced. Although the tests to detect HIV and other blood-borne diseases are extremely accurate, they cannot detect the presence of the virus in the 'window period' of infection, the time before detectable antibodies or antigens are produced. That is why there is still a very slim chance of contracting HIV from blood that tests negative. Research continues to further reduce the very small risk." [4] Source:http://chapters.redcross.org/br/nypennregion/safety/mythsaid.htm

Reducing Bias

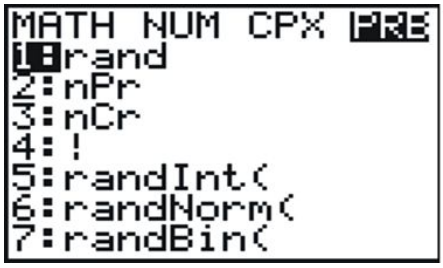
Randomization

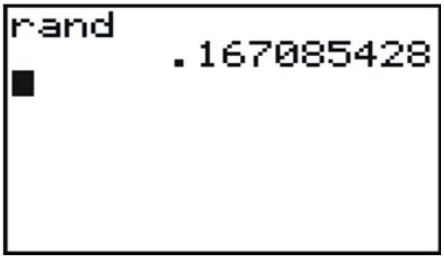
The best technique for reducing bias in sampling is *randomization*. When a *simple random sample* of size n (commonly referred to as an SRS) is taken from a population, all possible samples of size n in the population have an equal probability of being selected for the sample. For example, if your statistics teacher wants to choose a student at random for a special prize, he or she could simply place the names of all the students in the class in a hat, mix them up, and choose one. More scientifically, your teacher could assign each student in the class a number from 1 to 25 (assuming there are 25 students in the class) and then use a computer or calculator to generate a random number to choose one student. This would be a simple random sample of size 1.

A Note about Randomness

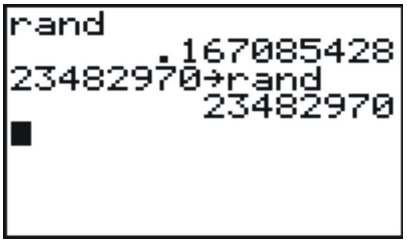
Technology Note: Generating Random Numbers on the TI-83/84 Calculator

Your graphing calculator has a random number generator. Press **[MATH]** and move over to the **PRB** menu, which stands for probability. (Note: Instead of pressing the right arrow three times, you can just use the left arrow once!) Choose '1:rand' for the random number generator and press **[ENTER]** twice to produce a random number between 0 and 1. Press **[ENTER]** a few more times to see more results.

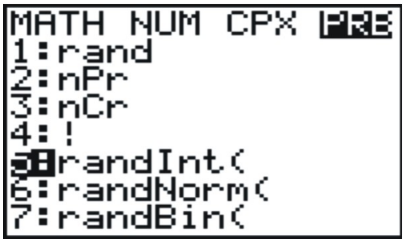




It is important that you understand that there is no such thing as true randomness, especially on a calculator or computer. When you choose the 'rand' function, the calculator has been programmed to return a ten digit decimal that, using a very complicated mathematical formula, simulates randomness. Each digit, in theory, is equally likely to occur in any of the individual decimal places. What this means in practice is that if you had the patience (and the time!) to generate a million of these on your calculator and keep track of the frequencies in a table, you would find there would be an approximately equal number of each digit. However, two brand-new calculators will give the exact same sequences of random numbers! This is because the function that simulates randomness has to start at some number, called a *seed value*. All the calculators are programmed from the factory (or when the memory is reset) to use a seed value of zero. If you want to be sure that your sequence of random digits is different from everyone else's, you need to seed your random number function using a number different from theirs. Type a unique sequence of digits on the home screen, press **[STO]**, enter the 'rand' function, and press **[ENTER]**. As long as the number you chose to seed the function is different from everyone else's, you will get different results.



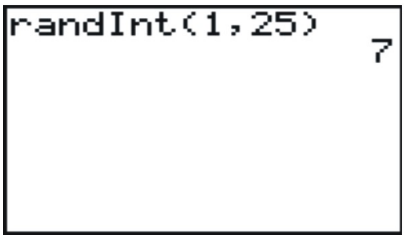
Now, back to our example. If we want to choose a student at random between 1 and 25, we need to generate a random integer between 1 and 25. To do this, press **[MATH][PRB]** and choose the 'randInt(' function.



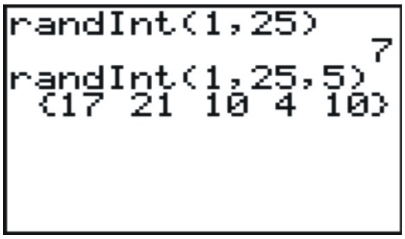
The syntax for this command is as follows:

'RandInt(starting value, ending value, number of random integers)'

The default for the last field is 1, so if you only need a single random digit, you can enter the following:

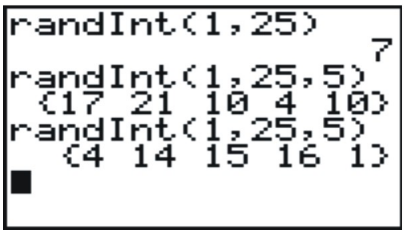


In this example, the student chosen would be student number 7. If we wanted to choose 5 students at random, we could enter the command shown below:



However, because the probability of any digit being chosen each time is independent from all other times, it is possible that the same student could get chosen twice, as student number 10 did in our example.

What we can do in this case is ignore any repeated digits. Since student number 10 has already been chosen, we will ignore the second 10. Press **[ENTER]** again to generate 5 new random numbers, and choose the first one that is not in your original set.



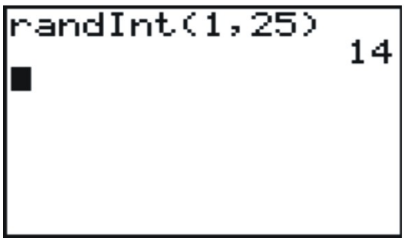
In this example, student number 4 has also already been chosen, so we would select student number 14 as our fifth student.

On the Web

http://tinyurl.com/395cue3 You choose the population size and the sample size and watch the random sample appear.

Systematic Sampling

There are other types of samples that are not simple random samples, and one of these is a systematic sample. In *systematic sampling*, after choosing a starting point at random, subjects are selected using a jump number. If you have ever chosen teams or groups in gym class by counting off by threes or fours, you were engaged in systematic sampling. The jump number is determined by dividing the population size by the desired sample size to insure that the sample combs through the entire population. If we had a list of everyone in your class of 25 students in alphabetical order, and we wanted to choose 5 of them, we would choose every 5^{th} student. Let's try choosing a starting point at random by generating a random number from 1 to 25 as shown below:



In this case, we would start with student number 14 and then select every 5^{th} student until we had 5 in all. When we came to the end of the list, we would continue the count at number 1. Thus, our chosen students would be: 14, 19, 24, 4, and 9. It is important to note that this is not a simple random sample, as not every possible sample of 5 students has an equal chance of being chosen. For example, it is impossible to have a sample consisting of students 5, 6, 7, 8, and 9.

Cluster Sampling

Cluster sampling is when a naturally occurring group is selected at random, and then either all of that group, or randomly selected individuals from that group, are used for the sample. If we select at random from out of that group, or cluster into smaller subgroups, this is referred to as *multi-stage sampling*. For example, to survey student opinions or study their performance, we could choose 5 schools at random from your state and then use an SRS (simple random sample) from each school. If we wanted a national survey of urban schools, we might first choose 5 major urban areas from around the country at random, and then select 5 schools at random from each of these cities. This would be both cluster and multi-stage sampling. Cluster sampling is often done by selecting a particular block or street at random from within a town or city. It is also used at large public gatherings or rallies. If officials take a picture of a small, representative area of the crowd and count the individuals in just that area, they can use that count to estimate the total crowd in attendance.

Stratified Sampling

In *stratified sampling*, the population is divided into groups, called strata (the singular term is 'stratum'), that have some meaningful relationship. Very often, groups in a population that are similar may respond differently to a survey. In order to help reflect the population, we stratify to insure that each opinion is represented in the sample. For example, we often stratify by gender or race in order to make sure that the often divergent views of these different groups are represented. In a survey of high school students, we might choose to stratify by school to be sure that the opinions of different communities are included. If each school has an approximately equal number of students, then we could simply choose to take an SRS of size 25 from each school. If the numbers in each stratum are different, then it would be more appropriate to choose a fixed sample (100 students, for example) from each school and take a number from each school proportionate to the total school size.

On the Web

http://tinyurl.com/2wnhmok This statistical applet demonstrates five basic probability sampling techniques for a population of size 1000 that comprises two sub-populations separated by a river.

Lesson Summary

If you collect information from every unit in a population, it is called a census. Because a census is so difficult to do, we instead take a representative subset of the population, called a sample, to try and make conclusions about the entire population. The downside to sampling is that we can never be completely sure that we have captured the truth about the entire population, due to random variation in our sample that is called sampling error. The list of the population from which the sample is chosen is called the sampling frame. Poor technique in surveying or choosing a sample can also lead to incorrect conclusions about the population that are generally referred to as bias. Selection bias refers to choosing a sample that results in a subgroup that is not representative of the population. Incorrect sampling frame occurs when the group from which you choose your sample does not include everyone in the population, or at least units that reflect the full diversity of the population. Incorrect sampling frame errors result in undercoverage. This is where a segment of the population containing an important characteristic did not have an opportunity to be chosen for the sample and will be marginalized, or even left out altogether.

Points to Consider

- How is the margin of error for a survey calculated?
- What are the effects of sample size on sampling error?

Review Questions

1. Brandy wanted to know which brand of soccer shoe high school soccer players prefer. She decided to ask the girls on her team which brand they liked.
    a. What is the population in this example?
    b. What are the units?
    c. If she asked all high school soccer players this question, what is the statistical term we would use to describe the situation?
    d. Which group(s) from the population is/are going to be under-represented?
    e. What type of bias best describes the error in her sample? Why?
    f. Brandy got a list of all the soccer players in the Colonial conference from her athletic director, Mr. Sprain. This list is called the what?
    g. If she grouped the list by boys and girls, and chose 40 boys at random and 40 girls at random, what type of sampling best describes her method?
2. Your doorbell rings, and you open the door to find a 6-foot-tall boa constrictor wearing a trench coat and holding a pen and a clip board. He says to you, “I am conducting a survey for a local clothing store. Do you own any boots, purses, or other items made from snake skin?” After recovering from the initial shock of a talking snake being at the door, you quickly and nervously answer, “Of course not,” as the wallet you bought on vacation last summer at Reptile World weighs heavily in your pocket. What type of bias best describes this ridiculous situation? Explain why.

In each of the next two examples, identify the type of sampling that is most evident and explain why you think it applies.

3. In order to estimate the population of moose in a wilderness area, a biologist familiar with that area selects a particular marsh area and spends the month of September, during mating season, cataloging sightings of moose. What two types of sampling are evident in this example?
4. The local sporting goods store has a promotion where every 1000^{th} customer gets a $10 gift card.

For questions 5-9, an amusement park wants to know if its new ride, The Pukeinator, is too scary. Explain the type(s) of bias most evident in each sampling technique and/or what sampling method is most evident. Be sure to justify your choice.

5. The first 30 riders on a particular day are asked their opinions of the ride.
6. The name of a color is selected at random, and only riders wearing that particular color are asked their opinion of the ride.
7. A flier is passed out inviting interested riders to complete a survey about the ride at 5 pm that evening.
8. Every 12^{th} teenager exiting the ride is asked in front of his friends: "You didn't think that ride was scary, did you?"
9. Five riders are selected at random during each hour of the day, from 9 AM until closing at 5 PM.
10. There are 35 students taking statistics in your school, and you want to choose 10 of them for a survey about their impressions of the course. Use your calculator to select a SRS of 10 students. (Seed your random number generator with the number 10 before starting.) Assuming the students are assigned numbers from 1 to 35, which students are chosen for the sample?

References

http://www.nytimes.com/2008/04/04/us/04pollbox.html

http://www.gao.gov/cgi-bin/getrpt?GAO-04-37

http://edition.cnn.com/2011/TECH/innovation/02/04/census.digital.technology/index.html

http://en.wikipedia.org/wiki/Literary_Digest

6.2 Experimental Design

Learning Objectives

- Identify the important characteristics of an experiment.
- Distinguish between confounding and lurking variables.
- Use a random number generator to randomly assign experimental units to treatment groups.
- Identify experimental situations in which blocking is necessary or appropriate and create a blocking scheme for such experiments.
- Identify experimental situations in which a matched pairs design is necessary or appropriate and explain how such a design could be implemented.
- Identify the reasons for and the advantages of blind experiments.
- Distinguish between correlation and causation.

Introduction

A recent study published by the Royal Society of Britain[1] concluded that there is a relationship between the nutritional habits of mothers around the time of conception and the gender of their children. The study found that women who ate more calories and had a higher intake of essential nutrients and vitamins were more likely to conceive sons. As we learned in the first chapter, this study provides useful evidence of an association between these two variables, but it is only an observational study. It is possible that there is another variable that is actually responsible for the gender differences observed. In order to be able to convincingly conclude that there is a cause and effect relationship between a mother's diet and the gender of her child, we must perform a controlled statistical experiment. This lesson will cover the basic elements of designing a proper statistical experiment.

Confounding and Lurking Variables

In an *observational study* such as the Royal Society's connecting gender and a mother's diet, it is possible that there is a third variable that was not observed that is causing a change in both the explanatory and response variables. A variable that is not included in a study but that may still have an effect on the other variables involved is called a *lurking variable*. Perhaps the existence of this variable is unknown or its effect is not suspected.

Example: It's possible that in the study presented above, the mother's exercise habits caused both her increased consumption of calories and her increased likelihood of having a male child.

A slightly different type of additional variable is called a confounding variable. *Confounding variables* are those that affect the response variable and are also related to the explanatory variable. The effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable. They are both observed, but it cannot be distinguished which one is actually causing the change in the response variable.

Example: The study described above also mentions that the habit of skipping breakfast could possibly depress glucose levels and lead to a decreased chance of sustaining a viable male embryo. In an observational study, it is impossible to determine if it is nutritional habits in general, or the act of skipping breakfast, that causes a change in gender birth rates. A well-designed statistical *experiment* has the potential to isolate the effects of these intertwined

variables, but there is still no guarantee that we will ever be able to determine if one of these variables, or some other factor, causes a change in gender birth rates.

Observational studies and the public's appetite for finding simplified cause-and-effect relationships between easily observable factors are especially prone to confounding. The phrase often used by statisticians is, "Correlation (association) does not imply causation." For example, another recent study published by the Norwegian Institute of Public Health[2] found that first-time mothers who had a Caesarian section were less likely to have a second child. While the trauma associated with the procedure may cause some women to be more reluctant to have a second child, there is no medical consequence of a Caesarian section that directly causes a woman to be less able to have a child. The 600,000 first-time births over a 30-year time span that were examined are so diverse and unique that there could be a number of underlying causes that might be contributing to this result.

Experiments: Treatments, Randomization, and Replication

There are three elements that are essential to any statistical experiment that can earn the title of a randomized clinical trial. The first is that a *treatment* must be imposed on the subjects of the experiment. In the example of the British study on gender, we would have to prescribe different diets to different women who were attempting to become pregnant, rather than simply observing or having them record the details of their diets during this time, as was done for the study. The next element is that the treatments imposed must be *randomly assigned*. Random assignment helps to eliminate other confounding variables. Just as randomization helps to create a representative sample in a survey, if we randomly assign treatments to the subjects, we can increase the likelihood that the treatment groups are equally representative of the population. The other essential element of an experiment is *replication*. The conditions of a well-designed experiment will be able to be replicated by other researchers so that the results can be independently confirmed.

To design an experiment similar to the British study, we would need to use valid sampling techniques to select a representative sample of women who were attempting to conceive. (This might be difficult to accomplish!) The women might then be randomly assigned to one of three groups in which their diets would be strictly controlled. The first group would be required to skip breakfast, the second group would be put on a high-calorie, nutrition-rich diet, and the third group would be put on a low-calorie, low-nutrition diet. This brings up some ethical concerns. An experiment that imposes a treatment which could cause direct harm to the subjects is morally objectionable, and should be avoided. Since skipping breakfast could actually harm the development of the child, it should not be part of an experiment.

It would be important to closely monitor the women for successful conception to be sure that once a viable embryo is established, the mother returns to a properly nutritious pre-natal diet. The gender of the child would eventually be determined, and the results between the three groups would be compared for differences.

Control

Let's say that your statistics teacher read somewhere that classical music has a positive effect on learning. To impose a treatment in this scenario, she decides to have students listen to an MP3 player very softly playing Mozart string quartets while they sleep for a week prior to administering a unit test. To help minimize the possibility that some other unknown factor might influence student performance on the test, she randomly assigns the class into two groups of students. One group will listen to the music, and the other group will not. When the treatment of interest is actually withheld from one of the treatment groups, it is usually referred to as the *control group*. By randomly assigning subjects to these two groups, we can help improve the chances that each group is representative of the class as a whole.

Placebos and Blind Experiments

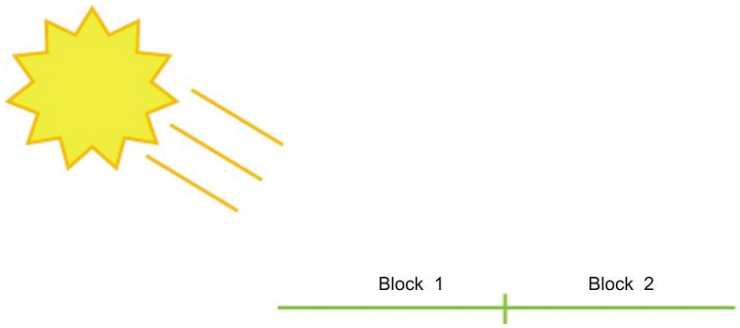
In medical studies, the treatment group usually receives some experimental medication or treatment that has the potential to offer a new cure or improvement for some medical condition. This would mean that the control group would not receive the treatment or medication. Many studies and experiments have shown that the expectations of participants can influence the outcomes. This is especially true in clinical medication studies in which participants who believe they are receiving a potentially promising new treatment tend to improve. To help minimize these expectations, researchers usually will not tell participants in a medical study if they are receiving a new treatment. In order to help isolate the effects of personal expectations, the control group is typically given a *placebo*. The placebo group would think they are receiving the new medication, but they would, in fact, be given medication with no active ingredient in it. Because neither group would know if they are receiving the treatment or the placebo, any change that might result from the expectation of treatment (this is called the *placebo effect*) should theoretically occur equally in both groups, provided they are randomly assigned. When the subjects in an experiment do not know which treatment they are receiving, it is called a *blind experiment*.

Example: If you wanted to do an experiment to see if people preferred a brand-name bottled water to a generic brand, you would most likely need to conceal the identity of the type of water. A participant might expect the brand-name water to taste better than a generic brand, which would alter the results. Also, sometimes the expectations or prejudices of the researchers conducting the study could affect their ability to objectively report the results, or could cause them to unknowingly give clues to the subjects that would affect the results. To avoid this problem, it is possible to design the experiment so that the researcher also does not know which individuals have been given the treatment or placebo. This is called a *double-blind experiment*. Because drug trials are often conducted or funded by companies that have a financial interest in the success of the drug, in an effort to avoid any appearance of influencing the results, double-blind experiments are considered the gold standard of medical research.

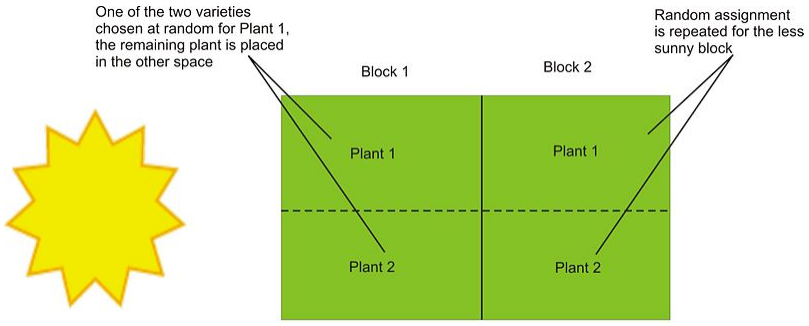
Blocking

Blocking in an experiment serves a purpose similar to that of stratification in a survey. For example, if we believe men and women might have different opinions about an issue, we must be sure those opinions are properly represented in the sample. The terminology comes from agriculture. In testing different yields for different varieties of crops, researchers would need to plant crops in large fields, or blocks, that could contain variations in conditions, such as soil quality, sunlight exposure, and drainage. It is even possible that a crop's position within a block could affect its yield. Similarly, if there is a sub-group in the population that might respond differently to an imposed treatment, our results could be confounded. Let's say we want to study the effects of listening to classical music on student success in statistics class. It is possible that boys and girls respond differently to the treatment, so if we were to design an experiment to investigate the effect of listening to classical music, we want to be sure that boys and girls were assigned equally to the treatment (listening to classical music) and the control group (not listening to classical music). This procedure would be referred to as blocking on gender. In this manner, any differences that may occur in boys and girls would occur equally under both conditions, and we would be more likely to be able to conclude that differences in student performance were due to the imposed treatment. In blocking, you should attempt to create blocks that are homogenous (the same) for the trait on which you are blocking.

Example: In your garden, you would like to know which of two varieties of tomato plants will have the best yield. There is room in your garden to plant four plants, two of each variety. Because the sun is coming predominately from one direction, it is possible that plants closer to the sun would perform better and shade the other plants. Therefore, it would be a good idea to block on sun exposure by creating two blocks, one sunny and one not.



You would randomly assign one plant from each variety to each block. Then, within each block, you would randomly assign each variety to one of the two positions.



This type of design is called *randomized block design*.

Matched Pairs Design

A *matched pairs design* is a type of randomized block design in which there are two treatments to apply.

Example: Suppose you were interested in the effectiveness of two different types of running shoes. You might search for volunteers among regular runners using the database of registered participants in a local distance run. After personal interviews, a sample of 50 runners who run a similar distance and pace (average speed) on roadways on a regular basis could be chosen. Suppose that because you feel that the weight of the runners will directly affect the life of the shoe, you decided to block on weight. In a matched pairs design, you could list the weights of all 50 runners in order and then create 25 matched pairs by grouping the weights two at a time. One runner would be randomly assigned shoe A, and the other would be given shoe B. After a sufficient length of time, the amount of wear on the shoes could be compared.

In the previous example, there may be some potential confounding influences. Factors such as running style, foot shape, height, or gender may also cause shoes to wear out too quickly or more slowly. It would be more effective to compare the wear of each shoe on each runner. This is a special type of matched pairs design in which each experimental unit becomes its own matched pair. Because the matched pair is in fact two different observations of the same subject, it is called a *repeated measures design*. Each runner would use shoe A and shoe B for equal periods of time, and then the wear of the shoes for each individual would be compared. Randomization could still be important, though. Let's say that we have each runner use each shoe type for a period of 3 months. It is possible that the weather during those three months could influence the amount of wear on the shoe. To minimize this, we could randomly assign half the subjects shoe A, with the other half receiving shoe B, and then switch after the first 3 months.

Lesson Summary

The important elements of a statistical experiment are randomness, imposed treatments, and replication. The use of these elements is the only effective method for establishing meaningful cause-and-effect relationships. An experiment attempts to isolate, or control, other potential variables that may contribute to changes in the response variable. If these other variables are known quantities but are difficult, or impossible, to distinguish from the other explanatory variables, they are called confounding variables. If there is an additional explanatory variable affecting the response variable that was not considered in an experiment, it is called a lurking variable. A treatment is the term used to refer to a condition imposed on the subjects in an experiment. An experiment will have at least two treatments. When trying to test the effectiveness of a particular treatment, it is often effective to withhold applying that treatment to a group of randomly chosen subjects. This is called a control group. If the subjects are aware of the conditions of their treatment, they may have preconceived expectations that could affect the outcome. Especially in medical experiments, the psychological effect of believing you are receiving a potentially effective treatment can lead to different results. This phenomenon is called the placebo effect. When the participants in a clinical trial are led to believe they are receiving the new treatment, when, in fact, they are not, they receive what is called a placebo. If the participants are not aware of the treatment they are receiving, it is called a blind experiment, and when neither the participant nor the researcher is aware of which subjects are receiving the treatment and which subjects are receiving a placebo, it is called a double-blind experiment.

Blocking is a technique used to control the potential confounding of variables. It is similar to the idea of stratification in sampling. In a randomized block design, the researcher creates blocks of subjects that exhibit similar traits that might cause different responses to the treatment and then randomly assigns the different treatments within each block. A matched pairs design is a special type of design where there are two treatments. The researcher creates blocks of size 2 on some similar characteristic and then randomly assigns one subject from each pair to each treatment. Repeated measures designs are a special matched pairs experiment in which each subject becomes its own matched pair by applying both treatments to the subject and then comparing the results.

Points to Consider

- What are some other ways that researchers design more complicated experiments?
- When one treatment seems to result in a notable difference, how do we know if that difference is statistically significant?
- How can the selection of samples for an experiment affect the validity of the conclusions?

Review Questions

1. As part of an effort to study the effect of intelligence on survival mechanisms, scientists recently compared a group of fruit flies intentionally bred for intelligence to the same species of ordinary flies. When released together in an environment with high competition for food, the percentage of ordinary flies that survived was significantly higher than the percentage of intelligent flies that survived.
    a. Identify the population of interest and the treatments.
    b. Based on the information given in this problem, is this an observational study or an experiment?
    c. Based on the information given in this problem, can you conclude definitively that intelligence decreases survival among animals?
2. In order to find out which brand of cola students in your school prefer, you set up an experiment where each person will taste two brands of cola, and you will record their preference.

a. How would you characterize the design of this study?
b. If you poured each student a small cup from the original bottles, what threat might that pose to your results? Explain what you would do to avoid this problem, and identify the statistical term for your solution.
c. Let's say that one of the two colas leaves a bitter after-taste. What threat might this pose to your results? Explain how you could use randomness to solve this problem.

3. You would like to know if the color of the ink used for a difficult math test affects the stress level of the test taker. The response variable you will use to measure stress is pulse rate. Half the students will be given a test with black ink, and the other half will be given the same test with red ink. Students will be told that this test will have a major impact on their grades in the class. At a point during the test, you will ask the students to stop for a moment and measure their pulse rates. In preparation for this experiment, you measure the at-rest pulse rates of all the students in your class.

Here are those pulse rates in beats per minute:

TABLE 6.1:

Student Number	At Rest Pulse Rate
1	46
2	72
3	64
4	66
5	82
6	44
7	56
8	76
9	60
10	62
11	54
12	76

a. Using a matched pairs design, identify the students (by number) that you would place in each pair.
b. Seed the random number generator on your calculator using 623.

Use your calculator to randomly assign each student to a treatment. Explain how you made your assignments.
c. Identify any potential lurking variables in this experiment.
d. Explain how you could redesign this experiment as a repeated measures design?

4. A recent British study was attempting to show that a high-fat diet was effective in treating epilepsy in children. According to the *New York Times*, this involved, "...145 children ages 2 to 16 who had never tried the diet, who were having at least seven seizures a week and who had failed to respond to at least two anticonvulsant drugs."

a. What is the population in this example?
b. One group began the diet immediately; another group waited three months to start it. In the first group, 38% of the children experienced a 50% reduction in seizure rates, and in the second group, only 6

percent saw a similar reduction prior to beginning the diet. What information would you need to be able to conclude that this was a valid experiment?

c. Identify the treatment and control groups in this experiment.
d. What conclusion could you make from the reported results of this experiment?

5. Researchers want to know how chemically fertilized and treated grass compares to grass grown using only organic fertilizer. Also, they believe that the height at which the grass is cut will affect the growth of the lawn. To test this, grass will be cut at three different heights: 1 inch, 2 inches, and 4 inches. A lawn area of existing healthy grass will be divided up into plots for the experiment. Assume that the soil, sun, and drainage for the test areas are uniform. Explain how you would implement a randomized block design to test the different effects of fertilizer and grass height. Draw a diagram that shows the plots and the assigned treatments.

Further reading:

http://www.nytimes.com/2008/05/06/health/research/06epil.html?ref=health

Part One: Multiple Choice

1. A researcher performs an experiment to see if mice can learn their way through a maze better when given a high-protein diet and vitamin supplements. She carefully designs and implements a study with the random assignment of the mice into treatment groups and observes that the mice on the special diet and supplements have significantly lower maze times than those on normal diets. She obtains a second group of mice and performs the experiment again. This is most appropriately called:

   a. Matched pairs design
   b. Repeated measures
   c. Replication
   d. Randomized block design
   e. Double blind experiment

2. Which of the following terms does not apply to experimental design?

   a. Randomization
   b. Stratification
   c. Blocking
   d. Cause and effect relationships
   e. Placebo

3. An exit pollster is given training on how to spot the different types of voters who would typically represent a good cross-section of opinions and political preferences for the population of all voters. This type of sampling is called:

   a. Cluster sampling
   b. Stratified sampling
   c. Judgment sampling
   d. Systematic sampling
   e. Quota sampling

Use the following scenario to answer questions 4 and 5. A school performs the following procedure to gain information about the effectiveness of an agenda book in improving student performance. In September, 100 students are selected at random from the school's roster. The interviewer then asks the selected students if they intend to use their agenda books regularly to keep track of their assignments. Once the interviewer has 10 students who will use their book and 10 students who will not, the rest of the students are dismissed. Next, the selected students' current

averages are recorded. At the end of the year, the grades for each group are compared, and overall, the agenda-book group has higher grades than the non-agenda group. The school concludes that using an agenda book increases student performance.

4. Which of the following is true about this situation?
    a. The response variable is using an agenda book.
    b. The explanatory variable is grades.
    c. This is an experiment, because the participants were chosen randomly.
    d. The school should have stratified by gender.
    e. This is an observational study, because no treatment is imposed.
5. Which of the following is not true about this situation?
    a. The school cannot conclude a cause-and-effect relationship, because there is most likely a lurking variable that is responsible for the differences in grades.
    b. This is not an example of a matched pairs design.
    c. The school can safely conclude that the grade improvement is due to the use of an agenda book.
    d. Blocking on previous grade performance would help isolate the effects of potential confounding variables.
    e. Incorrect response bias could affect the selection of the sample.

Part Two: Open-Ended Questions

1. During the 2004 presidential election, early exit polling indicated that Democratic candidate John Kerry was doing better than expected in some eastern states against incumbent George W. Bush, causing some to even predict that he might win the overall election. These results proved to be incorrect. Again, in the 2008 New Hampshire Democratic primary, pre-election polling showed Senator Barack Obama winning the primary. It was, in fact, Senator Hillary Clinton who comfortably won the contest. These problems with exit polling lead to many reactions, ranging from misunderstanding the science of polling, to mistrust of all statistical data, to vast conspiracy theories. The Daily Show from Comedy Central did a parody of problems with polling. Watch the clip online at the following link. Please note that while "bleeped out," there is language in this clip that some may consider inappropriate or offensive. http://www.thedailyshow.com/video/index.jhtml?videoId=156231&title=team-daily-polls What type of bias is the primary focus of this non-scientific, yet humorous, look at polling?
2. Environmental Sex Determination is a scientific phenomenon observed in many reptiles in which air temperature when eggs are growing tends to affect the proportion of eggs that develop into male or female animals. This has implications for attempts to breed endangered species, as an increased number of females can lead to higher birth rates when attempting to repopulate certain areas. Researchers in the Galapagos wanted to see if the Galapagos Giant Tortoise eggs were also prone to this effect. The original study incubated eggs at three different temperatures: $25.50°C$, $29.50°C$, and $33.50°C$. Let's say you had 9 female tortoises, and there was no reason to believe that there was a significant difference in eggs from these tortoises.
    a. Explain how you would use a randomized design to assign the treatments and carry out the experiment.
    b. If the nine tortoises were composed of three tortoises each of three different species, how would you design the experiment differently if you thought that there might be variations in response to the treatments?
3. A researcher who wants to test a new acne medication obtains a group of volunteers who are teenagers taking the same acne medication to participate in a study comparing the new medication with the standard prescription. There are 12 participants in the study. Data on their gender, age, and the severity of their condition are given in the following table:

TABLE 6.2:

Subject Number	Gender	Age	Severity
1	M	14	Mild
2	M	18	Severe
3	M	16	Moderate
4	F	16	Severe
5	F	13	Severe
6	M	17	Moderate
7	F	15	Mild
8	M	14	Severe
9	F	13	Moderate
10	F	17	Moderate
11	F	18	Mild
12	M	15	Mild

(a) Identify the treatments, and explain how the researcher could use blinding to improve the study.

(b) Explain how you would use a completely randomized design to assign the subjects to treatment groups.

(c) The researcher believes that gender and age are not significant factors, but is concerned that the original severity of the condition may have an effect on the response to the new medication. Explain how you would assign treatment groups while blocking for severity.

(d) If the researcher chose to ignore pre-existing condition and decided that both gender and age could be important factors, he or she might use a matched pairs design. Identify which subjects you would place in each of the 6 matched pairs, and provide a justification of how you made your choice.

(e) Why would you avoid a repeated measures design for this study?

Keywords

Bias

Blind experiment

Blocking

Census

Cluster sampling

Confounding variables

Control group

Convenience sampling

Double blind experiment

Experiment

Incorrect response bias

Incorrect sampling frame

Judgement sampling

Lurking variable

Margin of error

Matched pairs design

Multi-stage sampling

Non-response bias

Observational study

Placebo

Placebo effect

Questionnaire bias

Quota sampling

Random sample

Randomization

Randomized block design

Randomly assigned

Repeated measures design

Replication

Response bias

Sample

Sampling error

Sampling frame

Seed value

Simple random sample

Size bias

Stratified sampling

Systematic sampling

Treatment

Undercoverage

Voluntary response bias