Unit 3: Collecting Data

Introduction

We have looked at different types of data and ways of displaying data, and now we will start to look at collecting data and designing observational studies and experiments. This is an incredibly important topic, as flawed design or bad techniques of collecting data could potentially ruin years' worth of research. This has happened more often than you might think in real-world research, and it has cost huge amounts of money, resources, and even difficulty to the public-at-large receiving what they thought was good, scientific information.

There is a very old line, "There are lies, damned lies, and statistics" (implying that statistics are the worst form of lies). The reason this saying exists is that data from studies or experiments can be manipulated, either intentionally or unintentionally – it is our duty as statisticians and researchers to try to remove bias from our data, observations, and experiments as much as possible.

This is further confounded by the fact that variations that we observe may be random or they may have some kind of correlation or causation – but the conclusions are uncertain unless we design experiments and analyze data properly. Additionally, we need to ask good questions – sometimes we can find out what questions to ask by collecting data from observational studies, and we can use that to design good experiments. But it is essential that our data collection and experimentation be performed in such a way (using random selection of participants) so that we can generalize our results to a population that our sample represents.

3.1 Planning a Study, Random Sampling, and Data Collection

Objectives:

- Identify the type of study.
- Identify appropriate generalizations based on observational studies
- Determine methods of sampling and identify a sampling method.
- Explain why a particular sampling method is (or is not) appropriate for a given situation.

In using statistics, we have two main methods of approaching data. We can perform an *observational study* or an *experiment*.

- A study is an **observational study** if the investigator observes characteristics of a sample (ideally a random sample) from a population, but does not impose a treatment.
- An **experiment** differs from an observational study in that the investigator deliberately imposes a treatment (ideally treatments are randomly assigned to test subjects).
    - An investigator must identify at least one or more **explanatory variables, *x*,** also called **factors** or **treatment**, to manipulate and at least one **response variable, *y*.** A group is treated with some **level** of the explanatory variable, and the outcome on the response is measured.

Only *experiments* have the potential to provide *causation* rather than *correlation*!

When we perform an experiment, since we are assigning *treatments* (more on this later) to participants and (usually) testing against a *control*, we may be able to use the data we collect to establish a *causal relationship* (that is, a cause-and-effect relationship).

Causal Relationship:

- A **causal relationship** occurs when one variable has a *direct influence* on another variable. That is, in essence, one event triggers another event.
    - *Experiments* can establish causal relationships.
    - *Observational Studies* cannot establish causal relationships.

Given that observational studies cannot establish causality; why would we even do them?

There are actually a variety of answers to that question:

- First, observational studies can give us information about a correlation that can help us to realize that there may be a causal link to the correlation, and we can design an experiment (or experiments) to address the issue.
- Second, it may not be possible to perform an experiment ethically.
- Third, expense and time may prohibit certain types of experiments.

For example, suppose I am a statistician who wants to know whether having a pre-existing opioid addiction has an effect on surgery recovery times. To actually conduct an *experiment*, I would need to randomly select subjects, cause half of them to become opioid addicts, and then perform the same surgery on all of them and track the recovery time. There are obviously ethical concerns with this.

Example1: The Associated Press reported on an investigation that concluded that women who suffer severe morning sickness early in pregnancy are more likely to have a girl. This conclusion was reached by researchers in Sweden based on a "scientific study". Do you think that the "scientific study" referred to in the article was an experiment or an observational study?

This would likely be an observational study, as it is unlikely that they could control and apply the state of pregnancy to the women in the group.

Example 2: A study is to be designed to examine the life expectancy of tall people versus short people. Which is more appropriate, an observational study or an experiment?

Example 3: A study is to be designed to examine the GPAs of students who take marijuana regularly and those who don't. Which is more appropriate, an observational study or an experiment?

Example 4: To test the value of help sessions outside the classroom before a test, students could be divided into three groups, with one group receiving 4 hours of help sessions per week, a second group receiving 2 hours per week, and a third group receiving no help, and then all students take the same test. What are the explanatory and response variables and what are the levels?

Answer:

Explanatory Variable (x): Hours of help sessions

Response Variable (y): Grade on the test

Levels: 4 hours/week, 2 hours/week, 0 hours/week of help sessions.

When there is uncertainty with regard to which variable is causing an effect, we say the variables are **confounded**. It is easier to control **confounding variables** in an experiment rather than an observational study. Confounding variables are an important aspect of an experiment to control. They can be very hard to figure out at times, and there are many examples of "conventional scientific wisdom" being widely accepted but incorrect because of unknown confounding variable (saccharine causing cancer, the Alar scare and cancer, and eggs linked with cholesterol levels are all examples of this). We will address this much more in a later section.

For both observational studies and experiments, we want results that can potentially be applied to the general population (remember what we mean by population in statistics).

- **Generalizations:** We can only make generalizations about a population based on samples that are *randomly selected* or otherwise represent the population.

It would do well for us to recall certain statistics definitions at this point:

- **Population:** All items or subjects of interest.
- **Sample:** A subset of a population selected for study.

We must now consider how to take a good sample

<u>Methods of Sampling</u>

- **Simple Random Sample (SRS):** A simple random sample of size ***n*** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

  The definition of a simple random sample implies that every individual member of the population has an equal chance of being selected AND every group of size n is possible.

Example 5: A small private college has 4500 students enrolled. Assume that the university can provide a list of the students with the students numbered from 1 to 4500. Describe the procedure you will use to select a simple random sample of 20 students, and then identify (by number) which students from the list are included in your sample.

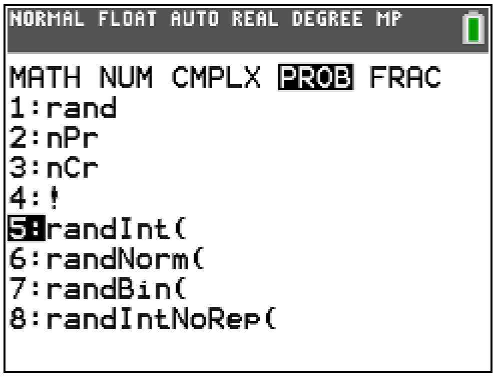
We have 3 methods to find a *simple random sample*:

1. Slips of paper – assign each member of the population a number on a piece of paper and thoroughly mix the slips in a hat (or some other container) and select without replacement slips until you have the desired number of elements in your sample.

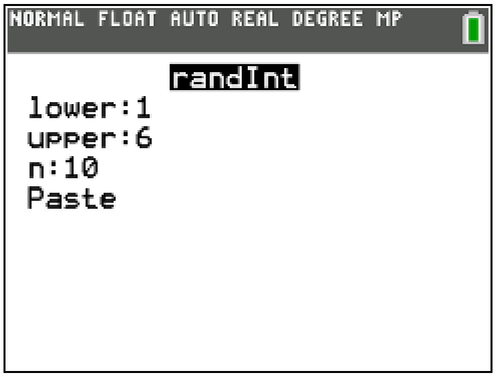
   - It is important to explain this exactly as stated above for the AP Test!

2. Calculator – assign each member of the population a number and use your calculator to generate random digits:

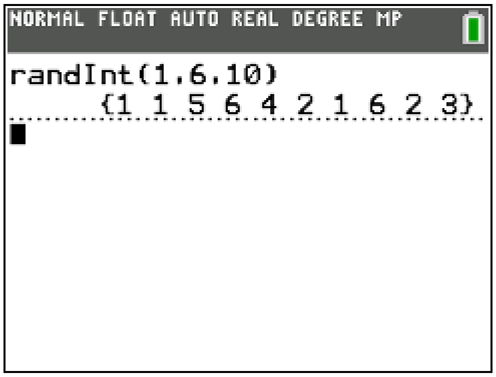
   - Press the **math** button, and move over to the PROB menu and select option 5.



Press the **enter** key and then the following menu shows up (in this example, I am generating 10 numbers with values from 1 to 6):



lower:1
upper:6
n:10

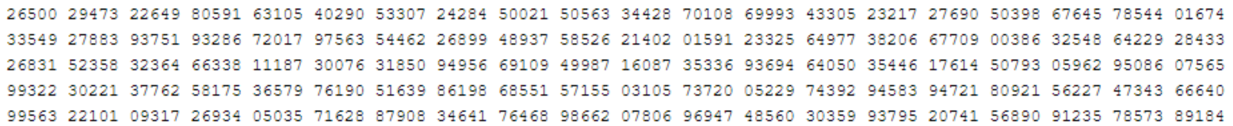


```
randInt(1,6,10)
     {1 1 5 6 4 2 1 6 2 3}
```

3. Random Digit Table: Assign a number to each member of the population (it is essential for this process to be valid, that if you have multi-digit numbers, all values have the same number of digits – for example, if you have 500 members of the population, you would assign each member of the population a number from 001 to 500). Then use a random digit table to find your sample.
    - All of the values having the same number of digits is essential for this to work.

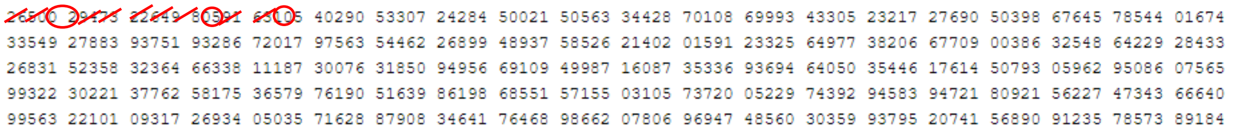
Example 6: Use the following random digit table to select 3 distinct members from the following list:

1: Alexander	2: Mason	3: Qi	4: Truong	5: Huang	6: O'Connor
7: Farad	8: Kowalski	9: Bueller	10: Kodos	11: Lenard	12: Murphy



```
26500 29473 22649 80591 63105 40290 53307 24284 50021 50563 34428 70108 69993 43305 23217 27690 50398 67645 78544 01674
33549 27883 93751 93286 72017 97563 54462 26899 48937 58526 21402 01591 23325 64977 38206 67709 00386 32548 64229 28433
26831 52358 32364 66338 11187 30076 31850 94956 69109 49987 16087 35336 93694 64050 35446 17614 50793 05962 95086 07565
99322 30221 37762 58175 36579 76190 51639 86198 68551 57155 03105 73720 05229 74392 94583 94721 80921 56227 47343 66640
99563 22101 09317 26934 05035 71628 87908 34641 76468 98662 07806 96947 48560 30359 93795 20741 56890 91235 78573 89184
```

First assign the numbers 01 to 12 for the individual members of the population above. Then pick two-digit numbers from right to left. If a number does not represent a member of your group, discard it and move to the next two-digit number. If a person is selected twice, disregard the second selection and keep going. Continue until you have gotten 5 individuals:



```
26500 29473 22649 80591 63105 40290 53307 24284 50021 50563 34428 70108 69993 43305 23217 27690 50398 67645 78544 01674
33549 27883 93751 93286 72017 97563 54462 26899 48937 58526 21402 01591 23325 64977 38206 67709 00386 32548 64229 28433
26831 52358 32364 66338 11187 30076 31850 94956 69109 49987 16087 35336 93694 64050 35446 17614 50793 05962 95086 07565
99322 30221 37762 58175 36579 76190 51639 86198 68551 57155 03105 73720 05229 74392 94583 94721 80921 56227 47343 66640
99563 22101 09317 26934 05035 71628 87908 34641 76468 98662 07806 96947 48560 30359 93795 20741 56890 91235 78573 89184
```

Looking at the table above, we find 02, 05, and 10. So the 3 people selected are Mason, Huang, and Kodos.

You might notice that if I were looking for a fourth candidate using this same method (with the initial question asking for 4 people, not 3), I would continue the same pattern:

Cross off the 64 which came next, and you would get 02 again. You would disregard this (as 02 was already selected) and continue the process, eliminating 90 53 30 72 42 84 50 02 (again) 15 05 (again) 63 34 42 87, and then we would finally get 01, which associates with Alexander.

Other Methods of Sampling

- **Stratified Sampling:** In stratified sampling, separate random samples are independently selected from each subgroup (called **strata**).
    - When the entire population can be divided into a set of non-overlapping, homogeneous subgroups, a method known as **stratified sampling** often proves easier to implement and more cost-effective than a *simple random sample*, and it may provide additional information as well.
    - If you wanted to poll the entire student body at SI you could break us up into Freshmen, Sophomores, Juniors, and Seniors…then take a *simple random sample* inside each of the **strata**.
        - Notice that the four classes are non-overlapping (you cannot be a freshman and a sophomore) and homogenous (everyone in the freshman class is, by definition, a freshman).

- **Cluster sampling** involves dividing the population of interest into non-overlapping heterogeneous subgroups, called clusters. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample.
    - **Clusters** differ from **strata** in that they are heterogeneous sub-groups.
    - Consider a 10-story college dorm building. We could get a list of each student living in the dorm, number them, get our random sample and track each kid down for an interview. It would be easier to pick a random floor in the dorm building and interview each student on that entire floor. The clusters are heterogeneous (the only similarity is what floor they are on).

- **Systematic sampling** is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement.
    - Let's say you want to TP your teachers' houses but you want to randomly choose which teachers to TP so that no teacher will take the attack personally. There are 120 faculty members at SI. You want to nab ten teachers in one awesome night. Break the teachers into 12 groups of ten, randomly select a number between 1 and 10…let's say you randomly select 7, then from each group of 12 TP the 7th faculty member on the list. :)

- **Convenience sampling** relies on using an easily available or convenient group to form a sample.
    - If you want to know how many hours the average students at SI studies, will asking just your friends give data that represents the entire student body?

Convenience sampling often leads to *bias* (more on this in the next section)

Example 7: If we took the 500 people attending a school in New York City, divided them by gender, and then took a random sample of the males and a random sampling of the females. This is an example of what type of sampling method?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Example 8: Say the target population in a study was church members in the United States. There is no list of all church members in the country. The researcher could, however, create a list of churches in the United States, choose a sample of churches, and then obtain lists of members from those churches to interview. This is an example of what type of sampling method?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Example 9: Determining the sample interval (represented by *k*), randomly selecting a number between 1 and *k*, and including each *k*th element in your sample are the steps for which form of sampling?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Answers: 7: (b), 8: (d), 9: (c)

Example 10: The financial aid officers of a university wish to estimate the average amount of money that students spend on textbooks each term. They are considering taking a stratified sample. For each of the following proposed stratification schemes, discuss whether you think it would be worthwhile to stratify the university students in this manner.

a) Strata corresponding to class standing (freshman, sophomore, junior, senior, graduate student)

b) Strata corresponding to field of study, using the following categories: engineering, architecture, business, other

c) Strata corresponding to the first letter of the last name : A –E, F – K, etc.

Summary:

- **Observational Study:** The investigator observes characteristics of a sample (ideally a random sample) from a population, but does not impose a treatment.

- **Experiment:** The investigator deliberately imposes a treatment (ideally treatments are randomly assigned to test subjects).
    - There is at least one **explanatory variable, *x*,** also called a **factor** or **treatment**, to manipulate and at least one **response variable, *y*.** A group is treated with some **level** of the explanatory variable, and the outcome on the response is measured.
- **Simple Random Sample (SRS):** A sample (of size *n*) that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.
- **Stratified Sampling:** In stratified sampling, separate random samples are independently selected from each subgroup (called **strata**).
- **Cluster sampling:** Divide the population of interest into non-overlapping heterogeneous subgroups, called clusters. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample.
- **Systematic sampling:** A procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement.
- **Convenience sampling:** Using an easily available or convenient group to form a sample.

Checkpoint 3.1

1. A researcher planning a survey of heads of households in a particular state has census lists for each of the 23 counties in that state. The procedure will be to obtain a random sample of heads of households from each of the counties rather than grouping all the census lists together and obtaining a sample from the entire group. Which of the following is a true statement about the resulting stratified sample?

   I. It is not a SRS.
   II. It is easier and less costly to obtain than a SRS.
   III. It gives comparative information that a SRS wouldn't give.

   (a) I only
   (b) I and II
   (c) I and III
   (d) I, II and III
   (e) None of the above gives the complete set of true responses.

2. Each of the 29 NBA teams has 12 players. A sample of 58 players is to be chosen as follows. Each team will be asked to place 12 cards with their players' names into a hat and randomly draw out two names. The two names from each team will be combined to make up the sample. Will this method result in a simple random sample of the 348 basketball players?

   (a) Yes, because each player has the same chance of being selected.
   (b) Yes, because each team is equally represented.
   (c) Yes, because this is an example of stratified sampling, which is a special case of simple random sampling.
   (d) No, because the teams are not chosen randomly.
   (e) No, because not each group of 58 players has the same chance of being selected (i.e. not every group is possible).

3. To survey the opinions of bleacher fans at Wrigley Field, a surveyor plans to select every one-hundredth fan entering the bleachers one afternoon. Will this result in a simple random sample of Cub fans who sit in the bleachers?

   (a) Yes, because each bleacher fan has the same chance of being selected.
   (b) Yes, but only if there is a single entrance to the bleachers.
   (c) Yes, because the 99 out of 100 bleacher fans who are not selected will form a control group.
   (d) Yes, because this is and example of systematic sampling, which is a special case of simple random sampling.
   (e) No, because not every sample of the intended size has an equal chance of being selected.

3.1 Homework

1) A study where researchers were investigating whether there was a relationship between the amount of sleep high school students get and the amount of caffeine that they consume. They randomly selected a group of 363 high school students and found that students who get less than eight hours of sleep on a school night were more likely to report falling asleep in class and consumed more caffeine on average than students who got more than eight hours of sleep.
   a. Is this an observational study or an experiment?
   b. Is it reasonable to conclude that getting less than eight hours of sleep on school nights causes teenagers to fall asleep during class and consume more caffeine on average? Explain.

2) A petition with 8000 signatures is submitted to a university's student council. The president of the student council would like to know the proportion of people who signed the petition but does not have enough time to check all of the names with the university registrar, so the student council president decides to take a simple random sample of 40 signatures. Describe how this might be done.

3) For each of the following situations, state the type of sampling procedure being used (simple random sampling, stratified random sampling, cluster sampling, systematic sampling, or convenience sampling).
   a. All Saint Ignatius freshman are enrolled in a one of 15 sections of a religious studies course. To select a sample of freshmen at SI, a researcher randomly selects 2 sections of this course from the 15 sections, and then uses all of the students in those two sections are included in the sample.
   b. A student doing a research project in his cognitive psychology class uses the 29 students in his class as a sample to represent the SI student body.
   c. To obtain a sample of students, faculty, and staff at SI, a researcher randomly selects 150 students from a list of students, 25 faculty members from a list of faculty, and 15 staff from a list of staff.
   d. To obtain a sample of the people attending the Bruce-Mahoney basketball game, a researcher selects the 15th person through the door and then every 20th person after that is included in the sample.
   e. To obtain a sample of seniors at SI, Dr. Quattrin, the school statistician, writes the name of each senior on identical slips of paper, places the slips in a box and mixes them thoroughly. He then selects 15 slips without replacement and uses all of the names on the selected slips of paper as his sample.

4) A newspaper reporter wanted to know how truthful people are on their social media platforms. A group of volunteers were recruited through a combination of print, online, and social media, and the resulting group was used as the sample. Their actual heights, weights, ages, and incomes were recorded and compared to what they portrayed on their social media accounts. The reporter used the resulting data to draw conclusions about how much people lied on their social media accounts.
   a) What concerns, if any, do you have about generalizing these results to all people on social media?
   b) What are some possible ways that you could address the concerns in part a)?

5) The article **"High Levels of Mercury Are Found in Californians" (*Los Angeles Times*, February 9, 2006)** describes a study in which hair samples of Californians were tested for mercury levels. The hair samples were obtained from more than 6000 people who voluntarily sent hair to researchers at Greenpeace and The Sierra Club. The researchers found that close to one-third of the individuals tested had mercury levels that exceeded the concentration that is thought to be safe. Is it reasonable to generalize this result to the larger population of adults in the United States? Explain why or why not?

6) In a certain college town, there was considerable debate among the public whether or not to hold a Mardi Gras parade. The parade had been popular with students and many residents, but some of the excessive celebrations of participants had led to complaints and a call to eliminate the parade. A local newspaper conducted both telephone and online surveys and the results were that for over 400 online surveys, the parade had a support rate of about 60%, while for the over 120 phone responses over 90% favored banning the parade. The editor was shocked by this disparity. What factors may have contributed to these different results to the surveys (identical questions were asked in both surveys)?

3.2 Potential Problems with Sampling

Objectives:

- Determine types of bias.
- Identify bias potential in surveys and data-collection.

A lot of people may have the question, “Why do we take samples of populations to conduct studies?”

There is actually a pretty simple answer to this – in almost every case, it isn’t actually possible (or wise) to test an entire population.

Just a quick note on the term population. Recall that it refers to all of the “things” about which we are asking our question. While in day-to-day life, “population” means people, that is not what we mean in statistics. It is important to remember this.

So how and why do we take samples?

Example 1: I want to know the voltage that can run through a computer chip a certain factory produces before that chip is destroyed. The company wants to know the overall ability to handle a voltage load for the chips it produces The factory has ten different production lines, with each line producing 150 identical chips per day (so the factory produces 1500 chips a day). Discuss which of the following you think is the best way to determine the voltage loads that the chips can hold.

A) Test 1 chip to failure and record the voltage. That would be the voltage it takes to destroy a chip.
B) Test all the chips that the factory produces and then average that data to find the average voltage which causes the chip to be destroyed.
C) Test all the chips from one line on one day until they fail and average those values to get the average voltage at which a chip will fail.
D) Over several days, take random samples of ships from each line and test them to failure, and average those values to get the average voltage at which a chip will fail.

What are potential problems that could occur with these methods?

The answer was D) – we want our sample to be representative of the population as much as possible. Obviously, we cannot test every chip, because then we would have none left to sell.

(Additionally, as we go into later chapters, we will find that it is better to give a *range of values* for the average than just getting a single number – much more on this when we talk about *confidence intervals* in Unit 7).

Example 2: Which of the following questions do you think would be best to include in a survey on a law banning plastic bags in supermarkets?

A) Given many experts believe that plastic waste is one of the biggest problems of our time, would you support a ban on plastic bags to help preserve our environment?
B) Given that plastic bags have little environmental impact when disposed of properly or recycled, would you oppose an over-reaching law that attempted to ban plastic bags for everyday use?
C) Would you support or oppose a law banning plastic bags in supermarkets?

Which of the above options do you think would get you closest to the actual opinion of the sample you surveyed?

Suppose you used question C) in a poll at an Earth day rally? Do you think you could generalize your results to the whole population? What if you used question C) in a poll asking people at a supermarket where most of the people leaving were using plastic bags? Do you think you could generalize your results to the whole population?

Depending on how we select our sample or ask questions, a lot can go wrong. Statisticians have a word for that - **bias**. Sometimes *bias* is intentional (when someone designs a survey or study to get a result that they desire), but more often, it is unintentional – either because of ignorance, accident, or simply lack of availability of better methods. Most of our discussion (for now) will focus on *bias* in sample selection, but we will address other forms of bias as well.

- **Bias:** A sample is **biased** if it systematically over-represents or under-represents a segment of our population of interest.
  - **The point of collecting data from a sample is to represent a population as closely as possible!**

There is no way to recover from a biased sample or a survey that asks biased questions!

A lot of people assume that sampling more people will help to overcome bias – it will not necessarily. If your questions are flawed, or your sample selection is biased, simply asking more people will not fix this.

Example 3: Suppose you want to know the general public's opinion on rap music. You construct a survey with unbiased questions. You go to a heavy metal concert and administer the survey to every 10th person as they enter the concert. Do you think you could generalize the results to the overall population? Would this survey be better if I double the number of people I asked?

> Hopefully, it should be obvious that you selected a group of people that was not necessarily representative of the whole population. If you wanted to know the opinion of heavy metal fans on rap, you may be able to generalize your survey to that population.

Types of Bias

- **Selection Bias:** The tendency for samples to differ from the corresponding population as a result of systematic exclusion of some part of the population.
  - **Voluntary Response Bias:** When a sample is comprised solely of volunteers (people who chose to participate), the sample will typically not be representative of the population.
    - Consider people listening to a radio program, where the host asks for phone calls with opinions about a controversial proposition. 80% of the people who call in are against the proposition – do you think these results could be generalized to the overall population?
  - **Undercoverage Bias:** A sampling scheme that fails to sample from some part of the population or that gives part of the population less representation than it has in the population.
    - Example 3 (about rap music at a heavy metal conference) is an example of undercoverage bias.
    - An example of undercoverage is the Literary Digest voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from undercoverage of low-income voters, who tended to be Democrats.

  - **Non-random sampling methods:** Samples chosen for convenience, using voluntary response, or which miss out on a segment of a population because of a design flaw introduce the potential for bias because they do not use chance to select the individuals in the sample.

- **Nonresponse Bias:** This occurs in a sample design when individuals selected from the sample fail to respond, cannot be contacted, or decline to participate.
  - When Dr. Quattrin sends out a survey to students via email, and there are very few students who respond, this is non-response bias.

- **Response Bias:** Response bias occurs in two main forms – when questions are misleading or with self-reported responses:
  - **Question Wording Bias:** When questions are misleading or confusing, or when there is a component of "social desirability" in the question (where the respondent feels social pressure to respond a certain way).
  - **Self-Reported Responses:** When people self-assess and report the results, there is often significant bias. For example, surveys will typically find that 90% of respondents rate themselves as "above average" drivers – since it is highly unlikely that 90% of a population is above average, these people are not assessing themselves properly, thus *self-reported response bias.*

It is very important to note that bias can be introduced into a observational study or experiment by *the way a sample is selected* or by *the way in which the data are collected.*

To reiterate an earlier point, one that is very important, remember that increasing the sample size, although possibly desirable for other reasons, DOES NOTHING TO REDUCE BIAS!

Example 4: Suppose a survey is constructed with the following question, "It is estimated that disposable diapers account for less than 2 percent of the trash in today's landfills. In contrast, beverage containers, third-class mail, and yard waste are estimated to account for about 21 percent of trash in landfills. Given this, in your opinion, would it be fair to tax or ban disposable diapers?" What kind of bias is introduced here?

Example 5: Until very recently (about 2016), political pollsters conducted surveys principally through live interviews from randomly selected home phone numbers (land lines).

a) Explain two ways this may have led to biased responses.

b) Recently, political polling has switched to almost exclusively online polling. Could there be bias issues with this method?

Answer:

For part a) there are several potential answers – there could be *non-response* bias because people may avoid the poll or hang up on the pollster. There could be *undercoverage bias* because people who do not have a land-line are not represented. There could be *social response bias* based on the fact that people may feel social pressure to respond a certain way to a person interviewing them.

For part b), you could introduce *undercoverage bias*, as people who do not have access to the internet may not be represented, and you could also have *voluntary response bias* depending on how the internet survey is given.

Example 6: According to the article "Effect of Preparation Methods on Total Fat Content, Moisture Content, and Sensory Characteristics of Breaded Chicken Nuggets and Beef Steak Fingers", sensory tests were conducted using 40 college volunteers at Texas Women's university. Give three reasons, apart from the relatively small sample size (although really it's not that small), why this sample may not be ideal as the basis for generalizing to the population of all college students.

Summary:

- **Bias:** A sample is **biased** if it systematically over-represents or under-represents a segment of our population of interest.
    - Good samples *represent the population.*
- **Voluntary Response Bias:** When a sample is comprised solely of volunteers.
- **Undercoverage Bias:** A sampling scheme that fails to sample from some part of the population or that gives part of the population less representation.
- **Non-random sampling methods:** Samples chosen for convenience or using voluntary response.
- **Nonresponse Bias:** This occurs in a sample design when individuals selected from the sample fail to respond, cannot be contacted, or decline to participate.
- **Response Bias:** Response bias occurs in two main forms – when *questions are misleading* or with *self-reported responses*.

Checkpoint 3.2

Multiple Choice Questions

1. Which of the following are true statements?

   I. If bias is present in a sampling procedure, it can be overcome by dramatically increasing the sample size.
   II. There is no such thing as a bad sample.
   III.Questions in a survey do not have the potential to be biased.

(a) I only (b) II only (c) III only (d) I and II only (e) None of the above

3.2 Homework

1) Three weeks before an election, a particular pollster is gathering data to see which candidates in a specific election are more or less likely to win. He polls a random selection of registered voters and asks them if they would prefer a Democrat or a Republican. A second pollster is trying to gather the same data on the same candidates, but uses a different methodology. She decides to poll a random sample of likely voters and asks if they would prefer one candidate or the other, identifying the candidates only by their names. Which method do you think would be a more accurate way of predicting the outcome of the election? Is their potential bias in either (or both) methods? Explain.

2) Describe the types of bias that might be present in each of the following examples.
    a. A survey is sent via email to all 350 SI seniors to determine senior's opinion about the school schedule. 37 students respond to the survey.
    b. A survey is sent via email to a random selection of SI students to determine student opinion about their own stress levels. They are asked to respond to the question, "What is the biggest cause of stress in your life?"
    c. The student council wants an idea of what would be good themes for Spirit Week, so they decide to write a survey. They administer the survey to all of the students that they know.
    d. A group of volunteers are testing a nutritional supplement. After taking the supplement, they are asked to rate how well they feel.
    e. A pollster is trying to determine which candidate is more popular among registered voters in the upcoming election. They call the home phone numbers of randomly selected registered voters, and if a voter does not have a home phone, a replacement is randomly selected.

3) An SI administrator is trying to determine the opinion of athletes at the school and designs a poll.
    a. He administers the poll to randomly selected basketball and football players and believes that he can generalize this result to the opinion of athletes. Is he correct? If he is not, what mistakes has he made, and what type(s) of bias are there in this poll?
    b. Which of the following do you think would yield better results? A simple random sample of SI athletes or a stratified sample where the strata are the different sports would yield better results? Explain.

3.3 Designing Experiments and Selecting an Experimental Design

Objectives:

- Identify possible confounding variables.
- Distinguish double-blind from single-blind experiments.
- Recognize the placebo effect.
- Use a control group in an experiment when appropriate.
- Use each of the 4 Key Concepts in Experimental Design.
- Design a completely randomized experiment.
- Design a block experiment or matched pairs experiment when appropriate.

We already mentioned (briefly) in section 3.1 the idea of *confounded variables*. When there is uncertainty with regard to which variable is causing an effect, we say the variables are **confounded**. It is easier to control **confounding variables** in an experiment rather than an observational study.

- **Confounding Variables:** Additional variables in an experiment or observational study which, if not controlled for can create uncertainty in the causal relationship(s) in the experiment or study.

For example, if I study marijuana use in teenagers, and after extensive surveying, I find that heavy marijuana use correlates with low GPAs (this is actually the finding after extensive study). However, I have many confounding variables that I have not controlled for (and likely cannot control for given it is an observational study) – for example, did the students struggle with school and turn to substance use to deal with lack of success, or do the students attracted to marijuana use have a more "counter-cultural" definition of success and do not define a high GPA as a valuable form of success (interestingly, there seems to be validity in all three of these statements based on the research, and it highlights just how difficult finding causal relationships can be).

While that example focused on an observational study, but this section will be looking much more at *experiments* and *experimental design*. Recall from section 3.1 our definition of an experiment:

- **Experiment:** An investigator deliberately imposes a treatment (ideally treatments are randomly assigned to test subjects) on test subjects, while controlling as many *confounding variables* as possible.
  - An investigator must identify at least one or more **explanatory variables, *x*,** also called **factors** or **treatment**, to manipulate and at least one **response variable, *y*.** A group is treated with some **level** of the explanatory variable, and the outcome on the response is measured.

Example 1: A study of human development showed two types of movies to groups of children. Crackers were available in a bowl, and the investigators compared the number of crackers eaten by children watching the different kinds of movies. One kind of movie was shown at 8 AM (right after the children had breakfast) and another at 11 AM (right before the children had lunch). It was found that during the movie shown at 11 AM, more crackers were eaten than during the movie shown at 8 AM. The investigators concluded that the different types of movies had an effect on appetite.

The results from this experiment cannot be trusted because

(a) boys and girls have different eating patterns.
(b) the investigators were biased. They knew beforehand what they hoped the study would show.
(c) the investigators should have used several bowls, with crackers randomly placed in each.
(d) the time the movie was shown and the type of movie are confounded.
(e) Children do not like movies.

The answer is (d). Because one movie is shown right after the children have eaten, and the other is shown after a long span of time after eating (right before the next meal), you have not controlled the variable of how hungry the children are while watching the movie, so you may get increased consumption of crackers based on their hunger rather than on the type of the movie they are watching.

It is important to control confounding variables in an experimental design, as you might end up reporting false conclusions based on your confounding variable rather than the cause-and-effect you thought you established. Lest you think that this risk is not actually that significant, here are 3 real world examples (and research is rife with examples – even the best researchers can make mistakes and/or miss things).

1) "The Marshmallow Effect" – In 1972, Stanford researchers tested 3 to 5 year olds on delayed gratification (making little children wait to eat a marshmallow). The children who had the longest delays in eating the marshmallows had the greatest success later in life (higher SAT scores, lower BMI, higher educational achievement, as well as other measures). Subsequent studies found that almost half of the effect found may have been due to the economic background rather than the delayed gratification.

    a. This was an experiment with subsequent studies to observe potential connections (the initial experiment occurred in 1972 with subsequent studies in 1988 and 1990 on the same children, as well as a brain imaging study in 2011).
    b. The likely issue with the study is that all of the subjects were chosen from the Bing Nursery School of Stanford University.

2) "Coffee Drinking Causes Pancreatic Cancer" – in 1981, a Harvard University research team studied the link between pancreatic cancer and coffee drinking. When they studied the results of their data collection, they found that there was a weak association of cigarette smoking with the pancreatic cancer, but a very strong association with coffee consumption. (Incidentally, they found no association between pancreatic cancer and alcohol consumption, pipe tobacco, or cigars).

   a. It turns out that many of the coffee drinkers in this study also smoked cigarettes, and once that was controlled for, the association vanished.
   b. In addition, this was an observational study and not an experiment, and in their excitement over the findings of their research, the researchers claimed causal links that were reported widely in newspapers (and in some cases exaggerated by authors who misunderstood the distinction between "likely contributes to an increased risk of" and "causes" cancer)

3) "Methylmercury in Fish Risk" – A 2008 study discovered that the toxic effects of methylmercury may be actually worse than previously thought. The phenomenon of mercury in certain kinds of fish is a well-documented phenomenon, and the effects of the mercury consumed by eating those fish is a will-studied phenomenon. In 2008, a research team discovered that methylmercury, in general, could be more hazardous to humans than previously believed, because of the confounding variable of the health benefit of a diet that had a high fish content.

   a. This is a case of *negative confounding* – where the confounding variable masks the effect rather than exaggerates the effect of the explanatory variable. In this case, the health benefits of the fish diet *masked the effect* of the methylmercury toxicity.
   b. Note that negative confounding is not specifically part of the AP Statistics curriculum, but I have included it because it is an interesting phenomenon to be aware of in statistical analyses.

Hopefully, you can see from the above examples that it is very possible for *anyone* to make these kind of mistakes. You may have noticed that I used teams from Stanford and Harvard in the examples – this is not to impugn their reputation, but to demonstrate that confounding variables (and experimental and/or observational design issues) can happen to even the most prominent members of a field.

In addition, these kinds of mistakes can have a huge impact on public policy and/or public behavior, and could cause a lot of problems for a lot of people based on relatively minor research mistakes.

Often, when we perform an experiment and apply a treatment, we need a group to whom the treatment is not applied. This is called a *control group.* The existence of a control group can lead to the *placebo effect*:

- If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment, called the **control group**.
- A **placebo** is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.
    - Subjects are subject to the **placebo effect**. Many patients respond favorable to *any* treatment, even a placebo, presumably because of trust in the doctor and expectations of a cure.
    - There is also something called the *no-cebo effect* – this occurs when an experimental subject has received the treatment but believes that they have not. They then respond unfavorably to the treatment, despite it being effective. Again, this is not part of the AP Statistics curriculum.
- A **single-blind experiment** is one in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.
- A **double-blind experiment** is one in which neither the subjects nor the individuals who measure the response know which treatment was received.

And in case you think the placebo effect is a purely psychological effect, there is evidence that physiological effects occur as well. Consider the following:

1) In research on the placebo effect and pain relief, higher levels of endorphins (a natural pain-killer, among other things, that is produced in your body normally) are found in participants experiencing the placebo effect. This means that they are actually experiencing pain relief, not just “believing” they are in less pain.

2) In another study about “bedside manner” of doctors, the experiment was conducted as follows: A doctor would come in and swab the arm of a person with a mild irritant, without talking to or interacting with the patient (this was the control). For the experimental group, the same irritant was introduced, but the doctors reassured them that they would get a small rash, but it would be minor and that they would recover quickly. The experimental group had less inflammation and shorter duration and severity of the rash

Example 2: Pismo Beach, California, has an annual clam festival that includes a clam chowder contest. Judges rate clam chowder from local restaurants, and the judging is done in such a way that the judges are not aware of which chowder is from which restaurant. One year, much to the dismay of the seafood restaurants on the waterfront, Denny's chowder was declared the winner (when asked what the ingredients were, the Denny's cook said he wasn't sure - he just had to add the right amount of nondairy creamer to the soup stock he got from the Denny's distribution center).

(a) Do you think that Denny's chowder would have won the contest if the judging had not been "blind"? Explain.

(b) Although this was not an experiment, use your answer in Part (a) to explain why experiments are often blinded this way.

The **design** of an experiment is the overall plan for conducting an experiment. A good design minimizes ambiguity in the interpretation of the results.

Four Key Concepts in Experimental Design:

- **Randomization:** Random assignment (of subjects to treatments or of treatments to trials) reduces bias by equalizing the effects of confounding variables.
    - o *Remember that random assignment – either* ***of subjects to treatments or of treatments to trials*** *– is a critical component of a good experiment.*
- **Blocking:** Using extraneous factors to create groups (blocks) that are similar. All experimental treatments are then tried in each block. Not required, but may improve your design. A **matched pairs design** is a type of blocking. *We block to reduce variability*. This is how we compare treatment groups.
- **Direct Control:** Holding extraneous factors constant so that their effects are not confounded with those of the experimental conditions.
- **Replication:** Ensuring that there is an adequate number of observations for each experimental condition.

A good experiment will have all four of these elements. On the AP Test, you will often be asked to design (in Free Response Questions) or evaluate the design (in Multiple Choice or Free Response Questions) of experiments. It is essential that you include these elements in your design and that you critique these elements in your analysis of an experiment.

Example 3: A study is made to determine whether studying Latin helps students achieve higher scores on the verbal section of the SAT exam. In comparing records of 200 students, half of whom have taken at least 1 year of Latin, it is noted that the average SAT verbal score is higher for those 100 students who have taken Latin than for those who have not. Based on this study, guidance counselors begin to recommend Latin for students who want to do well on the SAT exam. Which of the following statements are true?

I. While this study indicates a relation, it does not prove causation.
II. There could well be confounding variables responsible for the seeming relationship.
III. Self-selection here makes drawing the counselors' conclusion difficult.

(a) I and II (b) I and III (c) II and III (d) I, II, and III (e) None of the above

The correct answer would be (a). Since it is a study, it shows correlation, not causation. There could also be confounding variables (what if generally better students signed up for Latin – their good study habits preceding the Latin course could be responsible for the discrepancy). Students were selected from records, they were not self-selected for the survey, so III. is incorrect.

Example 4: Based on a survey conducted on the Dietsmart.com web site, investigators concluded that women who regularly watched Oprah were only one-seventh as likely to crave fattening foods as those who watched other daytime talk shows.

(a) Is it reasonable to conclude that watching Oprah causes a decrease in craving fattening foods?

(b) Is it reasonable to generalize the results this survey to all women in the United States? To all women who watch daytime talk shows? Why or why not?

It should be fairly obvious that the viewing of a TV show is not likely to cause a decrease in craving fattening foods, so there are likely other factors at play. If we truly believe that there is a link, then we should do follow-up studies and design an experiment to test this.

Example 5: (2003B Q4) There have been many studies recently concerning coffee drinking and cholesterol level. While it is known that several coffee-bean components can elevate blood cholesterol level, it is thought that a new type of paper coffee filter may reduce the presence of some of these components in coffee.

The effect of the new filter on cholesterol will be studied over a 10-week period using 300 nonsmokers who each drink 4 cups of caffeinated coffee per day. Each of these 300 participants will be assigned to one of two groups: the experimental group, who will only drink coffee that has been made with the new filter, or the control group, who will only drink coffee that has been made with the standard filter. Each participant's cholesterol level will be measured at the beginning and at the end of the study.

(a) Describe an appropriate method for assigning the subjects to the two groups so that each group will have an equal number of subjects.

(b) In this study, the researchers chose to include a group who only drank coffee that was made with the standard filter. Why is it important to include a control group in this study even though cholesterol levels will be measured at the beginning and at the end of the study?

(c) Why would the researchers choose to only use nonsmokers in the study?

There are several ways to answer part (a), here are some common answers:

> Assign each participant a number between 001 and 300. Then use a random number table or random number generator to select 150 of the 300 for the new filter group. The other 150 would be assigned to the standard filter group.
> OR
> For each subject, flip a coin. If it lands on "heads" assign the subject to the new filter group, otherwise assign them to the standard filter group. Continue this way until one or the other group has 150 subjects, then assign all remaining subjects to the other group.

For part (b), be sure to explain why you need a control group, specifically:

> Without a group to compare to, the cholesterol level could change overall, but we would not be able to assess whether the change was because of some other variable that changed in the 10-week period. For example, there may be a change in participants' diets as the time of year changes, and the diet might result in cholesterol changes that we might falsely attribute to the new coffee filter. Adding a control group would allow researchers to assess the mean change in cholesterol levels in both groups.

For part (c), we are clearly looking at a potential confounding variable that the researchers chose to exclude:

> If it is known that smoking impacts cholesterol levels, we should control for smoking by excluding a potential source of variability. This allows for more direct comparison between treatment and control groups, and it allows us more precise estimates of the effects of the treatments (though we could only be able to generalize our results to non-smokers as smokers were excluded).

Example 6: We are interested in determining how student performance on a calculus exam is affected by room temperature. There are four calculus classes, taught by two different teachers, at the school where we are running the experiment. Two classrooms are set at 65˚ and two classrooms are set at 75˚.

Design this experiment.

Your answer for example 6 should include each teacher having a class in each temperature environment (controlling for the teacher), students randomly assigned to the groups of different temperatures, and making sure that all 4 classes are identical in size. Note that this experiment is *blocked* by teacher.

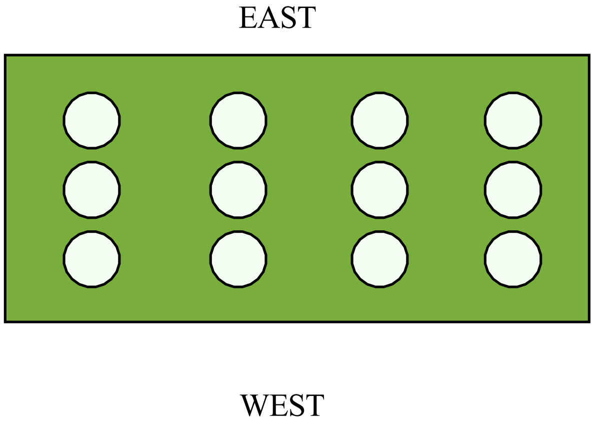
Sometimes there are confounding variables that we don't know or cannot control. One way to deal with this is *completely randomized design.*

- **Completely Randomized Design:** Treatments are assigned to experimental units completely at random.
  - *Random assignment* tends to balance out the effects of uncontrolled confounding variables so that we can do a better job of assessing whether responses can be attributed to treatments.

Example 7: An experiment to test the effectiveness of four different medications is being designed. Four hundred subjects have volunteered to participate. Describe a completely randomized experiment to randomly select subjects so that each subject will be assigned to one of the four treatment groups.

There are several ways you could randomly assign individuals to the treatments, but one is assigning each subject a number from 1 to 400. Write those numbers on identical slips of paper, put those papers into a container and mix thoroughly. Draw 100 slips of paper, without replacement, and the subjects corresponding to these numbers are assigned to the first treatment group. Repeat this process until you have all 4 treatments groups assigned. Apply each treatment and compare the results. (Note that this could easily be set up as a double blind experiment, where the researchers and the subjects do not know which treatment is applied to which group.

Example 8: A new type of fertilizer is being used that is meant to increase the mean overall weight of beans produced by bean plants over a six-week period. In order to test the efficacy of the fertilizer, the results need to be compared to bean plants grown with the previous fertilizer. The bean plants are to be grown in a garden as shown in the following diagram. However, it is believed that the direction of the sun will also have an effect on the way the bean plants grow.



Describe how an experiment may be constructed that compares the effectiveness of the new fertilizer with the old fertilizer and also accounts for the effect of the direction of the sun on the growth of the plant.

Example 9: (2002B Q3) A preliminary study conducted at a medical center in St. Louis has shown that treatment with small, low-intensity magnets reduces the self-reported level of pain in polio patients. During each session, a patient rested on an examining table in the doctor's office while the magnets, embedded in soft pads, were strapped to the body at the site of pain. Sessions continued for several weeks, after which pain reduction was measured.

A new study is being designed to investigate whether magnets also reduce pain in patients suffering from herniated disks in the lower back. One hundred male patients are available for the new study.

(a) Describe a completely randomized design for the new study. Discuss treatments used, methods of treatment assignment, and what variables would be measured.

(b) How could you modify the design above if, instead of 100 male patients, there were 50 male and 50 female patients available for the study? Why might you choose to do this?

Matched Pairs Design

A *matched pairs design* is a type of blocking. We will consider two types:

- **One Subject:** In this design one subject will receive both treatments. The order in which the subject receives the treatments is randomized.
- **Two Subjects:** In this design the two subjects are paired based on some common characteristic. One subject from the pair is randomly assigned a treatment, the other subject receives the other treatment.

Example 10: (2008B Q4) A researcher wants to conduct a study to test whether listening to soothing music for 20 minutes helps to reduce diastolic blood pressure in patients with high blood pressure, compared to simply sitting quietly in a noise-free environment for 20 minutes. One hundred patients with high blood pressure at a large medical clinic are available to participate in this study.

Propose a paired design for this study to compare these two treatments.

There are two ways we could answer example 10, with a *one subject matched pair* or a *two subject matched pair* design:

One Subject Matched Pairs: Each subject receives both treatments, with a suitable length of time in between treatments. Have the patient flip a coin to determine which treatment is administered first; heads = soothing music, tails = noise-free environment. Measure blood pressure, administer the treatment, and then measure blood pressure again. After a suitable amount of time, repeat the process with the other treatment. Calculate the changes in blood pressure and compare the results of the experiment.

Two Subject Matched Pairs: Measure the blood pressure of each patient and form 50 pairs of patients – the top two form a pair, then the next two highest form a pair, and so on until you get 50 pairs (this is so that we have similar individuals to compare). Have one patient flip a coin to determine which treatment they get; heads = soothing music, tails = noise-free environment, apply the other treatment to the other patient. Measure blood pressure of both in a pair, administer the respective treatments, and then measure blood pressure again. Calculate the change in blood pressure and compare results.

Summary:

- **Four Key Concepts in Experimental Design:**
    - **Randomization:** Random assignment (of subjects to treatments or of treatments to trials) reduces bias by equalizing the effects of confounding variables.
        - *Remember that random assignment – either **of subjects to treatments or of treatments to trials** – is a critical component of a good experiment.*
    - **Blocking:** Using extraneous factors to create groups (blocks) that are similar. All experimental treatments are then tried in each block. Not required, but may improve your design. A **matched pairs design** is a type of blocking. *We block to reduce variability*. This is how we compare treatment groups.
    - **Direct Control:** Holding extraneous factors constant so that their effects are not confounded with those of the experimental conditions.
    - **Replication:** Ensuring that there is an adequate number of observations for each experimental condition.
- **Confounding Variables:** Additional variables in an experiment or observational study which, if not controlled for can create uncertainty in the causal relationship(s) in the experiment or study.
- **Control group:** an experimental group that does not receive the treatment.
- **Placebo** is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.
    - **Placebo effect**: Response to a control treatment that mimics the response to the experimental treatment
- A **single-blind experiment** is one in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.
- A **double-blind experiment** is one in which neither the subjects nor the individuals who measure the response know which treatment was received.
- **Completely Randomized Design:** Treatments are assigned to experimental units completely at random.
- **One Subject Matched Pairs:** In this design one subject will receive both treatments. The order in which the subject receives the treatments is randomized.
- **Two Subjects Matched Pairs:** In this design the two subjects are paired based on some common characteristic. One subject from the pair is randomly assigned a treatment, the other subject receives the other treatment.

Checkpoint 3.3

Multiple Choice Questions

1. You are testing a new medication for relief of depression. You are going to give the new medication to subjects suffering from depression and see if their symptoms have lessened after a month. You have eight subjects available. Half of the subjects are to be given the new medication and the other half a placebo. The names of the eight subjects are given below.

1. Blumenthal	2. Costello	3. Duvall	4. Fan
5. House	6. Long	7. Pavlicova	8. Tang

Using the list of random digits 81507 27102 56027 55892 33063 41842 81868 71035 09001 43367 49497 starting at the beginning of this list and using single-digit labels, you assign the first four subjects selected to receive the new medication, while the remainder receive the placebo. The subjects assigned to the placebo are

(a) Blumenthal, Costello, Duvall, and Fan
(b) Blumenthal, House, Pavlicova, and Tang
(c) House, Long, Pavlicova, and Tang
(d) Costello, Duvall, Fan, and Long
(e) None of the above

2. A statistics class is made up of 20 female and 16 male students. A committee of 8 students needs to be selected. Each student is given a number from 1 to 36. A random table is used to repeatedly select two-digit numbers until eight different numbers in the range of 1 to 36 are generated, thus forming a committee of 8 students. After the committee was formed, it was discovered that all 8 of the students were male. One of the female students in the class complained that this could not be random since only male students were selected. Which of the following statements is true?

(a) A sample of size 8 is not large enough to produce random results.
(b) The method used did produce a random sample, even though only males were selected.
(c) It is so unlikely to have all 8 students be male that it is not a random sample.
(d) Since the results do not reflect the composition of the class, it is not representative; therefore, it is not random.
(e) A random number table cannot be used in this type of selection.

3. Which of the following is the best description of replication?

   (a) Asking subjects the same question in different ways
   (b) A technique of increasing the number of treatments used in an experiment
   (c) A technique of increasing the number of subjects in an experiment to help decrease the variation caused by chance
   (d) A tendency for subjects to be influenced by knowing what group they are in
   (e) A technique of distributing the subjects into random groups

This scenario applies to Questions 4 and 5: One hundred volunteers who suffer from severe depression are available for a study. Fifty are selected at random and are given a new drug that is thought to be particularly effective in treating severe depression. The other 50 are given an existing drug for treating severe depression. A psychiatrist evaluates the symptoms of all volunteers after four weeks in order to determine if there has been substantial improvement in the severity of the depression.

4. The study is an example of

   (a) a completely randomized design.
   (b) confounding. The effects of gender will be mixed up with the effects of the drugs.
   (c) a block design.
   (d) a matched-pairs design.
   (e) an observational study.

5. Referring to the study described above, suppose volunteers were first divided into men and women, and then half of the men were randomly assigned to the new drug and half of the women were assigned to the new drug. The remaining volunteers received the other drug. This would be an example of

   (a) a completely randomized design.
   (b) confounding. The effects of gender will be mixed up with the effects of the drugs.
   (c) a block design.
   (d) a matched-pairs design.
   (e) an observational study.

6. Which of the following are true about the design of matched-pairs experiments?

   I. Each subject might receive both treatments.
   II. Each pair of subjects receives identical treatment, and differences in their responses are noted.
   III. Matched-pair design is one form of blocking.

   (a) I only (b) II only (c) III only (d) I and III (e) II and III

7. Will a fluoride mouthwash used after brushing reduce cavities? Twenty sets of twins were used to investigate this question. One member of each set of twins used the mouthwash after each brushing; the other did not. After six months, the difference in the number of cavities of those using the mouthwash was compared with the number of cavities of those who did not use the mouthwash. This experiment uses

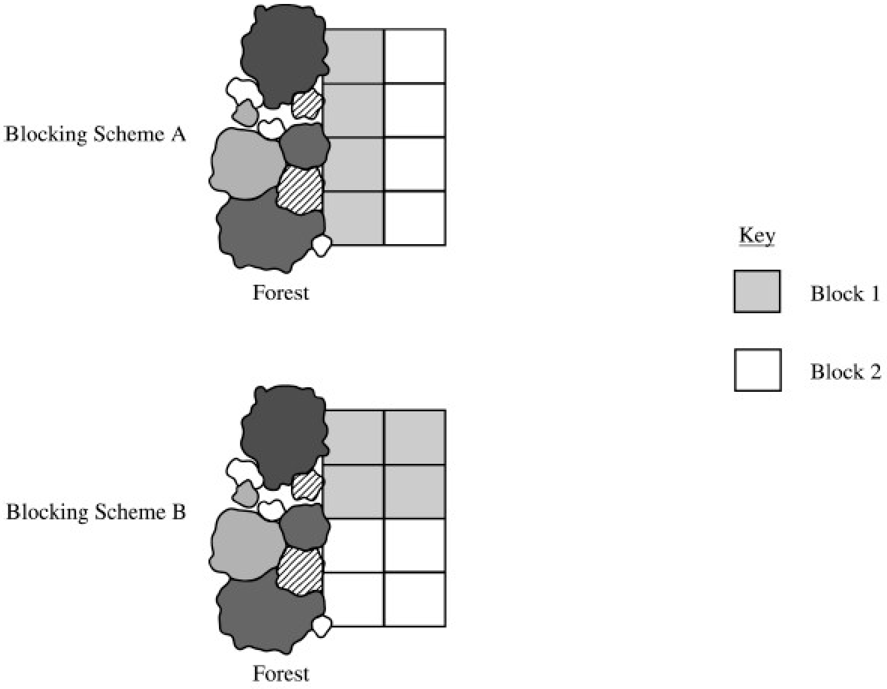
   (a) random placebos.
   (b) double-blinding.
   (c) double replication.
   (d) a matched-pairs design.
   (e) randomization of treatments.

Free Response Questions

1. Use a completely randomized design to construct an experiment that studies whether taking a garlic supplement in tablet form can reduce the occurrence of colds during the winter.

2. (2001 Q4) Students are designing an experiment to compare the productivity of two varieties of dwarf fruit trees. The site for experiment is a field that is bordered by a densely forested area on the west (left) side. The field has been divided into eight plots of approximately the same area. The students have decided that the test plots should be blocked. Four trees, two of each of the two varieties, will be assigned at random to the four plots within each block, with one tree planted in each plot.

The two blocking schemes shown below are under consideration. For each scheme, one block is indicated by the white region and the other block is indicated by the gray region in the figures.



a) Which of the blocking schemes, A or B, is better for this experiment? Explain your answer.

b) Even though the students have decided to block, they must randomly assign the varieties of trees to the plots within each block. What is the purpose of this randomization in the context of this experiment?

3. (2002 Q2) A manufacturer of boots plans to conduct an experiment to compare a new method of waterproofing to the current method. The appearance of the boots is not changed by either method. The company recruits 100 volunteers in Seattle, where it rains frequently, to wear the boots as they normally would for 6 months. At the end of the 6 months, the boots will be returned to the company to be evaluated for water damage.

   a) Describe a design for this experiment that uses the 100 volunteers. Include a few sentences on how it would be implemented.
   b) Could your design be double blind? Explain.

3.3 Homework

1) (2009 Q3) Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which the students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on the posttest minus score on the pretest).

   (a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

   (b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores for the two instructional programs? Provide a specific example to support your answer.

2) A medical study of heart surgery investigates the effect of a drug called a beta-blocker on the pulse rate of the patient during surgery. The pulse rate will be measured at a specific point during the operation. The investigators will use 20 patients facing heartsurgery as subjects. You have a list of these patients, numbered 1 to 20, in alphabeticalorder.

   (a) Describe a completely randomized experiment to test the effect of beta-blockers on pulse rate during surgery.

   (b) Use the section from the random digits table below to carry out the randomization required by your design and report the result.

```
96746 12149 37823 71868 18442 35119 62103 39244 96927 19931
36809 74192 77567 88741 48409 41903 43909 99477 25330 64359
40085 16925 85117 36071 15689 14227 06565 14374 13352 49367
81982 87209 36759 58984 68288 22913 18638 54303 00795 08727
```

3) Is the right hand of right-handed people generally stronger that the left? Paul Murky of Murky Research designs an experiment to test this question. He fastens an ordinary bathroom scale to a shelf five feet from the floor, with the end of the scale projectingout from the shelf. Subjects squeeze the scale between their thumb and their fingers on the top. A scale which reads in pounds will be used to measure hand strength. Youhave recruited 10 right-handed people to serve as subjects.

   (a) How would you conduct the experiment as a completely randomized design?

   (b) Are there potential flaws with this method?

   (c) Use the random digits below to do the randomization required by your design and report your results.

```
55588 99404 70708 41098 43563 56934 48394 51719
```

4) We wish to determine whether or not a new type of fertilizer is more effective than the type currently in use. Researchers have subdivided a 20-acre farm into twenty 1-acre plots. Wheat will be planted on the farm, and at the end of the growing season the number of bushels harvested will be measured. Describe a completely randomized design. What is the explanatory variable? What is the response variable? How many treatments are there? Are there any possible extraneous variables that would confoundthe results?

5) Suppose that the experiment described the previous problem (problem 4) has been redesigned in the following way: Ten 2-acre plots of land scattered throughout the county are randomly selected. Each plot is subdivided into two subplots, one of which is treated with the current fertilizer and the other of which is treated with the new fertilizer. Wheat is planted and the crop yields are measured. How is this experiment different from the previous problem? What advantages are there for this method? Which treatment is acting as the control group? What information, if any, can be gained by having a control group?

6) A local steel company wishes to test a new type of heat-resistant glove for workers who musthandle the molten steel. The company randomly selects 100 workers to test the gloves over afour-month period. Design an optimal experiment that will test whether the new gloves are more effective in resisting heat that the current gloves. Can your experiment be blinded? Explain your reasoning.

7) A researcher wants to determine whether performance in statistics classes can be influenced by the expectation of success. A statistics teacher will be teaching 15 sections of statistics over the next two years, 9 daytime classes and 6 evening classes. He wanted to tell the students in some sections that “females perform better in statistics than males”. In some othersections he wanted to say “males perform better in statistics than females”. Design an experiment that uses treatment and control groups, and blocks for the difference between dayand evening classes.

3.4 Inference and Experiments: What Can We Conclude?

Objectives:

- Interpret the results of a well-designed experiment.
- Explain the meaning of statistically significant results of an experiment.

When the results of an experiment show, through statistical methodologies, that the difference between a control and a treatment, or between two treatments is not due to random chance, then that difference is said to be **statistically significant**.

- **Statistically Significant Result:** When observed changes from control to experimental treatments are large enough that it is unlikely that they are due to random chance.
    - There are many tests for statistical significance that we will look at later in the course, but for now, we will look at this as a general concept.
    - *Not statistically significant* indicates that the difference or change is not large enough to establish that any change was causal.
    - *Statistically significant results* between experimental treatment groups are **evidence of a causal relationship** (that the explanatory variable caused the observed effect).
- **Generalizing to a Population**: If the experimental units (sample) used in an experiment *are representative of some larger group* (a population), then the results may be generalized to the larger group.
    - Remember, **random selection** of the experimental units will give a better chance that the units will be representative of the population.

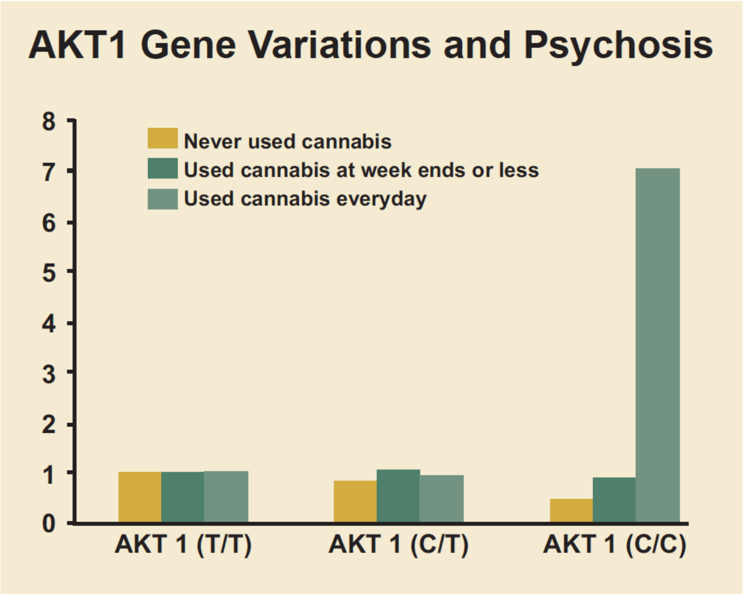
Example 1: A chemist is testing a new formulation of outdoor paint for a company. She obtains 50 identical blocks of wood and randomly assigns 25 blocks to be painted with the new formulation of the paint, and 25 blocks to be painted with the old formulation. All of the blocks are then subjected to 30 days of continuous exposure to identical weather conditions. After this exposure, the paint is examined for luster, brightness of color, shine, chipping, and other signs of weathering. Generally, the blocks with the new paint have better outcomes in the recording of the data, but after analysis, she determines that the results of the experiment were not statistically significant. Which of the following conclusions is correct?

a) The experiment is not valid because not enough pieces of wood were painted and tested.
b) There is not enough evidence to conclude that the new formulation weathers better than the old formulation over a one-month period.
c) The experiment is inconclusive because one month is not sufficient time for paint to become damaged from weather.
d) The new paint formulation will last longer than the old paint formulation on wooden surfaces.
e) Because the blocks were randomly assigned, there is evidence that the new formulation will last longer than the old formulation.

The correct answer is b), because it was determined that the results were not statistically significant, you cannot conclude that any difference that occurred was due to the new formulation of paint. Answer c) is tempting, but they told us that the results were not statistically significant, so b) is better. We could do another experiment over a longer period of time and see if we got statistically significant results in the next experiment if we suspected that one month was an insufficient amount of time to show any difference.

Hopefully, this image from a study on marijuana and the potential for marijuana-induced psychosis linked to a genetic component (*Di Forti et al. Biol Psychiatry. 2012*) can illustrate the magnitude of statistical significance (note that it is not always this obvious.

After a lot of research on the AKT1 gene variants (a gene that codes for proteins regulating dopamine production) and marijuana use, it was discovered that users with a particular variant of AKT1 were at a *statistically significantly* higher risk of developing psychosis. The following graph shows the *relative risks* of developing psychosis for people with this genetic marker:



It is interesting to note that carriers of the AKT1 (T/T) variant show virtually no variance in tendency to develop psychosis, while there is slight variation in the AKT1 (C/T) (which was not statistically significant), while the everyday users with AKT1 (T/T) were at 7 times greater risk – which was found to be statistically significant.

Important note about *relative risks*: sometimes these will be reported absent context to influence people's behavior, for example, if the risk of developing psychosis is 1 in 100,000 and jumps to 7 in 100,000, that is a seven-times increase in the rate, but it might not be statistically significant, because of the relatively small general risk (there are, as I have mentioned, a number of ways to test this).

As an example of the relative risks issue, it used to be common for anti-smoking advertisements to include that there was a 25% increased risk of developing lung cancer for people who were chronically exposed to second-hand smoke compared to those who were not. The study they cited had found a baseline rate (of non-smokers) for developing lung cancer to be 80 in 100,000. For the group who was exposed to second-hand smoke, their risk was 100 in 100,000. The authors of the study had found that this difference was not statistically significant. However, numerically it was a 25% increase so some groups began using it in anti-smoking campaigns. (Note that I am not including this to advocate for smoking or to try to indicate that smoke or second-hand smoke are not bad for you, I am simply illustrating how this particular data did not show the relationship that people were trying to say that it did).

Example 2: Researchers conduct a well-designed experiment to see how well a certain treatment mitigates symptoms of severe Covid-19 against a control group that is administered a placebo. Which of the following would allow us to conclude that the treatment causes mitigation of symptoms?

a) The data we collect could only be used to establish correlation, not causation.
b) The experiment cannot be used to determine the effectiveness of the new treatment because it is not being compared to an old treatment.
c) The difference between the responses to the new treatment and the placebo must be shown to be statistically significant to provide evidence that the new treatment causes mitigation of symptoms.
d) Any difference between the new treatment and the placebo is evidence that the treatment causes mitigation of symptoms.
e) Mitigation of symptoms would need to occur in all subjects who take the new treatment and in none of the subjects who take the placebo to provide evidence that the new treatment causes mitigation of symptoms.

The correct answer is c). Because this is an experiment that is well designed, we can potentially establish causation, so a) is incorrect. The reason b) is incorrect is that we are only establishing if this treatment works. If we wanted to know if this was more effective than other treatments, we would perform a different experiment. Differences must be statistically significant for the results to be evidence of causation, so d) is incorrect. Finally, the relief does not have to occur in everyone for it to be evidence, because individual responses can vary so widely, we look at the average mitigation of symptoms (this is why we perform experiments on relatively large samples – because people can respond to treatments very differently).

Example 3: People with agoraphobia (fear of open spaces) sometimes enroll in therapy sessions to help them overcome this fear. Typically, nine or ten therapy sessions are needed before improvement is noticed. A study was conducted to determine whether an anti-anxiety medication, gabapentin, used in combination with fewer therapy sessions, would help people with agoraphobia overcome this fear.

Each of 27 people who participated in the study received a pill before each of four therapy sessions. Seventeen of the 27 people were randomly assigned to receive a gabapentin pill, and the remaining 10 people received a placebo. After the four therapy sessions, none of the 27 people received additional pills or therapy. Three months after the administration of the pills and the four therapy sessions, each of the 27 people was evaluated to see if he or she had improvement.

a. Was this an experiment or an observational study? Provide evidence to support your assertion.

b. When the data were analyzed, the gabapentin group showed statistically significantly more improvement than the placebo group did. Based on this result, would the researchers be justified in concluding that the gabapentin pill and four therapy sessions are as beneficial as ten therapy sessions without the pill? Justify your answer.

c. A newspaper article that summarized the results of this study did not explain how it was determined which people received gabapentin and which received the placebo. Suppose the researchers allowed the therapists to choose which people received gabapentin and which received the placebo, and no randomization was used. Explain why such a method of assignment might lead to an incorrect conclusion.

Answers:

a. This was an experiment. Given that there was a treatment applied that was randomly assigned, and that it was compared to a control group, this was definitely an experiment.
b. No, they cannot. They can conclude that the gabapentin plus four therapy sessions is better than four therapy sessions with no medication. There is no basis for comparison to individuals with agoraphobia and ten therapy sessions, so no conclusions in that regard may be made.
c. If the therapists chose, they may (consciously or unconsciously) select patients who were more likely to be receptive to therapy for the control group. This could potentially skew the results and lead to an incorrect conclusion that the pill did not work (because the cases with less likelihood of improvement would be in the experimental group). Another possibility is that they assign more severe cases to one group and less severe to the other group. Our data would be confounded by receptiveness to therapy (in the first case) or severity of the illness (in the second case).

Example 4: The Physicians' Health Study, a large medical experiment involving 22,000 female medical doctors, attempted to determine whether aspirin could help prevent heart attacks. In this study, one group of about 11,000 physicians took an aspirin every day, while a control group took a placebo. After two years, it was determined that the medical doctors in the group that took aspirin had significantly fewer heart attacks than the medical doctors in the control group. Which of the following statements explains why it would **not** be appropriate to say that everyone should take an aspirin every day?

I. The study included only females, and the results may differ for males.
II. The study included only medical doctors, and the results may differ for other occupations.
III. Although aspirin may help prevent heart attacks, it may cause other negative health effects as well.

a) I only
b) II only
c) III only
d) I and II only
e) I, II, and III

The correct answer is e). This cannot be generalized to the whole population because we tested only on female physicians. In addition, before we advise the **everyone** do something, we should be aware of the potential negative effects of the treatment as well.

It is very important to pay attention to the wording in AP questions.

They give extensive clues in the wording of the questions, and they are looking for very precise, statistical responses. Remember the key ideas in terms of experimental design and statistical significance and look for how they are addressed within particular questions.

Summary:

- **Statistically Significant Result:** When observed changes from control to experimental treatments are large enough that it is unlikely that they are due to random chance.
    - *Not statistically significant* indicates that the difference or change is not large enough to establish that any change was causal.
    - *Statistically significant results* between experimental treatment groups are **evidence of a causal relationship** (that the explanatory variable caused the observed effect).
- **Generalizing to a Population**: If the experimental units (sample) used in an experiment *are representative of some larger group* (a population), then the results may be generalized to the larger group.
    - Remember, **random selection** of the experimental units will give a better chance that the units will be representative of the population.

Checkpoint 3.4

1. A researcher planning a survey of heads of households in a particular state has census lists for each of the 23 counties in that state. The procedure will be to obtain a random sample of heads of households from each of the counties rather than grouping all the census lists together and obtaining a sample from the entire group (using a stratified sample). After studying the resultant data, they find the mean income of the heads of households increases with the population of the county, and that the result is statistically significant. Which of the following statements are **FALSE**?

   I. There is a causal link between income and population.
   II. We cannot conclude anything as this is a survey and not an experiment.
   III. There is a correlation between population and income in the surveyed counties.

   (a) I only
   (b) I and II
   (c) I and III
   (d) I, II and III
   (e) None of the above

2. In order to assess the effects of exercise on reducing cholesterol, a researcher sampled 100 people. He assigned 50 people to exercise regularly and 50 people to not exercise regularly. They each reported to a clinic to have their cholesterol measured. The subjects were unaware of the purpose of the study, and the technician measuring the cholesterol was not aware of whether subjects exercised regularly or not. After analyzing the results of this double-blind study, the initial data shows a general decrease in cholesterol levels in the group that exercised, but further analysis showed that the results were not statistically significant. Which of the following is a correct conclusion?

   (a) The experiment was not well-designed and so the expected results did not happen.
   (b) There is not sufficient evidence to claim that exercise causes a reduction in cholesterol level.
   (c) Because the cholesterol level went down, we can conclude that the exercise caused the reduction.
   (d) This is an observational study, not an experiment, so no conclusions on a causal relationship can be made.
   (e) The lower cholesterol levels in the exercise group is likely the result of the placebo effect.

3. A study was done to compare the lung capacity of coal miners to the lung capacity of farm workers. The researcher studied 200 workers of each type. Other factors that might affect lung capacity are smoking habits and exercise habits. Then smoking habits of the two worker types are similar, but the coal miners generally exercise less than the farm workers, so the researcher blocks the groups by exercise habits. After collecting and analyzing the data, the researcher finds a statistically significant difference in lung capacity. Coal miners are found to have a statistically significantly lower lung capacity than farm workers.

   (a) We have sufficient information to conclude that working in a coal mine causes a decrease in lung capacity.
   (b) We have sufficient information to conclude that working in a coal mine is correlated with lower lung capacity compared to the general population.
   (c) We have sufficient information to conclude that working in a coal mine is correlated with lower lung capacity compared to working on a farm.
   (d) We cannot make any conclusions because this is an observational study and not an experiment.
   (e) We have sufficient information to conclude that working on a farm is correlated with lower lung capacity compared to working in a coal mine.

3.4 Homework

1) A chemist for a paint company conducted an experiment to investigate whether a new outdoor paint will last longer than the older paint. Fifty blocks made from the same wood were randomly assigned to be painted with either the new paint or the old paint. The blocks were placed into a weather-controlled room that simulated extreme weather conditions such as ice, temperature, wind, and sleet. After one month in the room, the blocks were removed, and each block was rated on texture, shine, brightness of color, and chipping. The results showed that the blocks painted with the new paint generally had higher ratings than the blocks painted with the old paint. However, an analysis of the results found that the difference in ratings was <u>not</u> statistically significant.
   a. Since we have noted that the results showed that the blocks with the new paint had generally higher ratings, can we conclude that the new paint is more weather resistant than the old paint? Explain why or why not.
   b. Suppose the experiment was repeated, but the blocks were placed in the room for three months rather than one month, and the new paint showed a lot less wear from ice and sleet, and that result was statistically significant. In addition, the new paint was generally better rated against wind and temperature, but the results were <u>not</u> statistically significant. What conclusions can you make from this result?

2) Researchers create a well-designed experiment to test the effectiveness of a weight loss product. They take a random sample of obese adults, half of whom receive the new weight loss product, and the other half of whom receive a placebo. After one month of treatment, it is determined that there is a statistically significant weight loss for the people who took the weight loss product.
   a. Can the results of this experiment be generalized to the entire population of the United States? Explain.
   b. The placebo group also lost weight. Can you conclude that the placebo caused this? Explain.

3) **Multiple Choice:** Researchers for a pharmaceutical company want to use a well-designed experiment to test the effectiveness of a new anti-viral versus a placebo in preventing severe flu symptoms. Which of the following will provide evidence that the new anti-viral prevents severe flu symptoms?
   (a) The experiment cannot be used to show the new anti-viral causes the prevention severe flu symptoms, only that it is correlated to the prevention of severe symptoms.
   (b) The experiment cannot be used to show the new anti-viral causes prevention of severe flu symptoms because the new anti-viral is not being compared to an older anti-viral.
   (c) Any difference between the responses to the new anti-viral and the placebo provides evidence that the new anti-viral is effective at causing the prevention of severe flu symptoms.
   (d) Prevention of severe symptoms would need to occur in all subjects who take the new anti-viral and in none of the subjects who take the placebo to provide evidence that the new anti-viral causes the prevention of severe symptoms.
   (e) The difference between the responses to the new anti-viral and the placebo must be shown to be statistically significant to provide evidence that the new anti-viral causes

the prevention of symptoms.

Unit 3 Practice Test