Section 2: Data Collection, Sampling, and Experimental Design

The following maps the videos in this section to the Texas Essential Knowledge and Skills for Mathematics TAC §111.47(c).

2.01 Sample Surveys

- Statistics (1)(A)
- Statistics (2)(B)
- Statistics (2)(C)
- Statistics (2)(E)

2.02 Sources of bias in sampling and surveys

- Statistics (2)(A)
- Statistics (2)(C)

2.03 Sampling Methods – Part 1

- Statistics (1)(A)
- Statistics (2)(A)

2.04 Sampling Methods – Part 2

- Statistics (1)(A)
- Statistics (2)(A)
- Statistics (2)(G)

2.05 Experiments vs Observational Studies

- Statistics (2)(B)

2.06 Three Principles of Experimental Design

- Statistics (1)(B)
- Statistics (2)(C)
- Statistics (2)(E)
- Statistics (2)(F)

2.07 Lurking and Confounding Variables

- Statistics (2)(G)

2.08 Confidence Intervals in the Real World

- Statistics (1)(B)
- Statistics (2)(C)
- Statistics (2)(E)
- Statistics (2)(F)
- Statistics (2)(G)

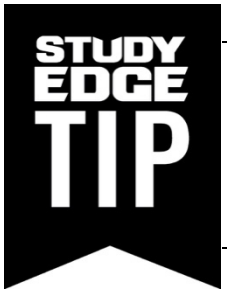
Note: Unless stated otherwise, any sample data is fictitious and used solely for the purpose of instruction.

2.01

Sample Surveys

Sample survey – Designed to gather information about a small group from a population of interest

- A sample of subjects is taken from the _____ and asked questions.
- The sample _____ matters, not the population size. The _____ our sample, the more precise our estimates will be.
- To reduce bias in surveys, it is best to have _____ samples.
- _____ samples are made up of people or subjects that are easy to obtain.



The design of a survey can have a major impact on results. A poorly designed and implemented study may be meaningless or misleading.

1. Suppose Dr. Malcolm is researching the side effects of a drug intended to relieve pain. She surveys 15 of her patients and finds that 40% of them experienced fatigue or muscle ache. Two other doctors researching the same drug found that from a random sample of 760 patients across the country, 30% experienced fatigue or muscle ache. Which study is more credible? Justify your answer.

When designing a survey, the following questions should be considered:

- What are the goals?
- What is the targeted population?
- What questions will be asked?
- How will the participants be selected?
- How will the data be analyzed and presented?

Other questions may need to be considered as well, depending on the circumstances.

2. Below are three hypothetical survey questions and statements. For each, identify any flaw(s) and suggest a way to improve it.

   i. "How satisfied are you with your pay, health benefits, and work load?"

   ii. "With the increase in speeding tickets on I-10 these past few months, do you think the speed limit should be increased since it seems everyone is driving fast anyway?"

   iii. "What is your marital status?"

2.02

Sources of Bias in Sampling and Surveys

Suppose that you work for a research agency, and your job includes surveying voters and predicting who is going to win the upcoming election. You call 1,000 voters, ask which candidate they intend to vote for, record and tabulate the answers, and make some predictions. After you collect the data, you discover that almost every survey respondent was a low-income voter and a registered Democrat. Is there a problem with your data?

In survey sampling, _____ is defined as the systematic favoring of certain outcomes.

A good sample is _____, meaning that each person or item in the sample is equally likely to be selected from the population.

Bias – The results obtained from a sample do not accurately represent the population.

Sources of Bias

- ***Undercoverage*** – A portion of the population is excluded or underrepresented.

  Example: The population is high school students, but our sample contains only freshmen.

- ***Nonresponse bias*** – People do not respond to the study.

  Example: One hundred people are selected from the phonebook to be surveyed, but 40 of them hang up before responding.

- ***Response bias*** – The survey design influences the responses.

  Examples: "Do you really think we should have band and chorus in school when we have so many core subjects to learn?" or "Do you only care about looks?"

Sampling Techniques That Create Bias

- ***Voluntary response sample*** – The respondents _____ choose to submit their responses.
- ***Convenience sample*** – The sample is selected based on ease and cost to the interviewer.

1. Suppose you take a survey about school attendance. The woman conducting the survey reminds you of your aunt, so you respond "No" when she asks if you have ever been late to class, even though you have. Which of the following types of bias is/are present in your response: undercoverage, nonresponse bias, or response bias?

2. Two polls show significantly different results. One asks, "Do you think school hours should be decreased?" and the other asks, "Do you think school hours should be decreased considering there is no time for outside play?" Which of the following types of bias is/are present in this scenario: undercoverage, nonresponse bias, or response bias?

3. Suppose that a politician wants to know how the residents of his district will react to a bill that raises the full retirement age to 70. He runs the following ad during *Sunday Night Football*:

   "Let us know what you think! Would you be in favor of raising the retirement age to 70, or would you rather keep the retirement age at 67, when many seniors are still productive and healthy enough to stay working? Give us a call at 1-800-555-1111, and give us your opinion!"

   Identify the sources of bias present in this ad, and justify your answer.

2.03

Sampling Methods – Part 1

Suppose you are examining a study and you find that the results are biased because of factors such as undercoverage and convenience sampling. You are very interested in the topic and would like to replicate the study to produce results that are unbiased and represent the population.

In general, _____ _____ use the laws of probability in the selection process and minimizes bias.

Random Sampling Techniques

- ***Simple random sample (SRS)*** – Each subject in the population has an equal chance of being selected. The use of random number generators is useful in this process.
- ***Stratified random sampling*** – The population is divided into homogeneous groups, and SRSs are taken from each group.

  Example: Dividing a high school into freshmen, sophomores, juniors, and seniors, and taking a simple random sample from each group
- ***Systematic random sampling*** – A random number generator is used to determine values for k, and then every k^{th} observation is sampled.
- ***Cluster random sampling*** – The population is divided into heterogeneous groups, and SRSs are taken from each group.

  Example: Dividing a high school based on class period, randomly selecting several classrooms, and sampling everyone in those classrooms
- ***Multistage random sampling*** – This approach involves a combination of the random sampling methods listed above.

1. Suppose Ms. Abernathy is conducting research on teachers' attitudes toward homeschooling. She is particularly interested in describing the attitudes of teachers from rural, small urban, and large urban school districts. Which sampling procedure should Ms. Abernathy use to ensure her sample is representative of these types of school districts?

2. Suppose you are a quality control engineer for a bottling company in Laredo, Texas, that focuses on soft drinks. Your job is to plan and oversee the various steps that are involved in processing and manufacturing each soft drink product to make sure the products maintain the highest standards of quality. Every time you want to test a product, you use random digits to randomly select five cans of soda from each batch of 50 cans. What sampling method are you using? Give an example of the random selection of cans.

3. You are conducting a poll in the mall and want to randomly select five people to interview from the next 100 people that walk past you. If you use the following random digits, what are the first two numbers that will be added to your sample?

41241 17562 70184 05752 81565 92499

4. You are conducting a poll in the mall and want to randomly select 10 people to interview from the next 200 people that walk past you. If you use the following random digits, what are the first two numbers that will be added to your sample?

41241 17562 70184 05752 81565 92499

2.04

Sampling Methods – Part 2

1. Which type of random or nonrandom sampling technique is being used in each scenario below?

   i. You are a quality control engineer for an orange juice company. You use the following random digits to randomly select five cartons of orange juice from a batch of 50 cartons of orange juice.

   74378 83000 36123 41232 54238 81261

   ii. Percy Weasley is conducting an investigation. He randomly samples from all students at Hogwarts until he has 25 Slytherins, 20 Gryffindors, 20 Ravenclaws, and 15 Hufflepuffs.

   iii. A polling company in New York wants to determine whether registered voters in the state recently voted "Yes" on a new amendment. They randomly select 200 registered voters from every county in the state and ask each person whether he or she voted "Yes" or "No."

2. The owner of a business park wants to survey the offices in the complex to estimate her tenants' overall satisfaction. The business park has two buildings, one older and one newer. Each building has three floors of standard office space (20 offices per floor) and one top floor of large offices, which are double the size (10 large offices per floor). Half of the offices in each building face the two main roads, while the other half face a lake and several trees. Currently, 75% of the offices are occupied.

   i. How many offices are there total?

   ii. Explain how to select a simple random sample of 20 offices.

   iii. Explain how to select a stratified random sample of 20 offices.

   iv. Explain why selecting two random floors would not be a good way to obtain a cluster sample.

2.05

Experiments and Observational Studies

Observational study – Researchers do not try to change anything; they just observe.

- ***Retrospective study*** – Researchers select subjects and determine their previous conditions or behaviors.
- ***Prospective study*** – Researchers follow subjects to observe future outcomes.

  Example: Suppose you are studying the question "Do musicians have better grades?" You can take a high school and look up records (retrospective), or you can take a high school and track the next decade (prospective).

Experiment – Researchers manipulate factor levels to create treatments and compare responses.

- ***Experimental units*** – Subjects or participants
- ***Response*** – What we want to measure
- ***Factor*** – Categorical variable with at least two levels
  - ***Level*** – The specific values of a factor that the experimenter chooses
  - ***Treatment*** – All possible values of the explanatory variable

1. Determine whether each of the following research projects is an experiment or an observational study. If it is an observational study, identify it as retrospective or prospective.

   i. A researcher wants to know which of three sororities has the highest GPA. She randomly samples 10 women from each sorority and records their GPAs.

   ii. A researcher wants to compare the effects of three headache medications. She randomly assigns 20 patients suffering from headaches to receive one of four pills, including a placebo, and measures the time until each patient no longer feels the headache.

   iii. A researcher wants to compare three brands of tires. He randomly assigns a tire brand to 24 cars, ensures that the cars are driven under similar conditions, and records how long each car's tires last.

   iv. A researcher wants to compare the salaries of four college majors. He randomly samples 20 recent graduates from each major and records their salaries over the next 5 years.

2. Is happiness affected by diet and exercise? A group of people is involved in a study to test this. They are told whether to exercise a low, moderate, or high amount each week and whether to maintain a healthy or unhealthy diet. In this experiment, identify each of the following:

   i. Experimental units

   ii. Factor(s)

   iii. Level(s)

   iv. Treatment(s)

   v. Response variable

2.06

Three Principles of Experimental Design

Three fundamental principles that help to make a good experiment:

- ***Control*** – The treatment the control group receives is a placebo.
    - If the subjects don't know which pill they get, the subjects are ***blind***.
    - If the researcher doesn't know either, the study is ***double-blind***.
- ***Randomization*** – Subjects are chosen at random.
    - A ***completely randomized*** is a design in which the experimental units are randomly assigned to the treatment groups.
    - A ***block design*** may be used to place subjects into similar groups
    - A ***matched pair design*** is a special case of a block design, used when the experiment has only two treatments.
- ***Replication*** – Several experimental units are assigned to each treatment. The number of replications is typically the ratio of the number of experimental units to the number of treatments.

1. Several methods of control exist: placebo group, blinding, and double-blinding. Write the correct method of control for each of the following situations.

    i. The treatment group is given a fake pill because people often get better simply because they think they are taking something that will make them better.

    ii. Neither the researcher(s) nor the subjects know who is getting what treatment. This helps to ensure all treatment groups are treated as equally as possible.

    iii. The subjects do not know which treatment they are getting. This helps to ensure all treatment groups are treated as equally as possible.

2. Determine whether each of the following scenarios is a completely randomized design, a randomized block design, or a matched pairs design. Justify your answer.

   i. To test for differences between three brands of gasoline, 20 compact cars, 20 midsized cars, 20 hybrid cars, and 20 luxury cars will be randomly assigned one of the three gasoline brands. All the cars will be driven under similar conditions. Then their miles per gallon (mpg) will be calculated and compared.

   ii. A farmer wishes to study the effect of two different pesticides on his strawberries. He will randomly assign side-by-side plots of field to receive either pesticide or no pesticide (control group). Then he will measure and compare the yield of strawberries within a pre-determined period of time for each plot.

   iii. To compare the abilities of male and female firefighters, 10 male and 10 female firefighters, matched according to their weights, will be asked to run up five flights of stairs while carrying the same fire hose. The differences in times will be determined and compared.

2.07

Lurking and Confounding Variables

Earlier we discussed variables that influence the results of an experiment. Two types of variables can affect the results of a study: _____ variables and _____ variables.

Lurking variable – A variable that is not incorporated into the design of a research study

Confounding variable – A variable that is incorporated into the design of a research study

These two types of variables have similar effects. Each drives the behavior of two other variables observed in a study, creating an apparent association between those two variables. However, when two variables are confounded, they are intertwined in such a way that figuring out which of them (or perhaps which combination) is affecting another variable is a huge challenge.

1. For the following scenarios, state whether the issue is with lurking variables or confounding variables.

    i. A legislator proposes a reduction of funding for fire departments in his state because a study concluded that the more firefighters involved in a fighting a fire, the worse damage the fire causes.

    ii. A student conducts research on the difference in body mass index (BMI) between economics students and education students. He was inspired by a previous study that concluded that economics students have higher BMIs than education students. The student's advisor suggests that he include biological sex in his study. The student finds that there is no difference between economics and education students' BMIs, challenging previous findings.

2.08

Generalization of Results and Conclusions

What results can be drawn from observational studies, experiments, and surveys?

_____ refers to the extent to which the findings from a study can be generalized, or applied, to a larger population. It requires _____ selection.

1. Which of the following describes when it is appropriate to generalize study results to a larger population?

   A. In experimental studies, when participants are conveniently selected from a large population
   B. In experimental studies, when participants are randomly selected from a large population
   C. In observational studies, when participants are selected from a pool of interested people

2. A study evaluates the efficacy of a specific treatment in elderly Caucasian men who have coronary heart disease.

   i. Discuss when can we generalize or apply the results of this study to a larger population.

   ii. Suppose that a health center in your city wants to apply the results of this study to develop outreach programs and services for Hispanic women and African-American men. Discuss the generalized issues that may arise from this action.

3. Suppose a popular radio station conducted a telephone survey from 11:00 a.m. to 1:00 p.m. to evaluate the popularity of the station among the local population. What is the problem with the generalizability of the results and conclusions of this study?