Minitab

LESSON: SAMPLING
This lesson includes an overview of the subject, instructor notes, and example exercises using Minitab.

Sampling

Lesson Overview

A ***population*** is the entire set, actual or conceptual, of subjects in a survey or study for which the question of interest is asked. It is the entire set of subjects about which the statistician wants to draw conclusions. A ***sample*** is a subset of subjects from a population for which observations are actually made.

Sampling bias occurs when a chosen sample is not representative of the larger population. Examples of sampling bias are provided in this lesson.

This lesson provides a foundation for determining how samples of data should be generated in order for their results to be valid.

Two methods of creating valid samples are discussed: ***simple random sampling*** and ***stratified sampling***. Sampling using Minitab will also be demonstrated.

Prerequisites

None. In Minitab, samples will be drawn on a single column of data.

Learning Targets

This lesson teaches students:

- The difference between a population and a sample. Numerical summaries of population data are called parameters and numerical summaries of sample data are called statistics.
- About sampling bias, including selection bias and nonresponsive bias.
- The definition and creation of a simple random sample and a stratified sample.
- When to sample with or without replacement.

Time Required

It will take the instructor 30 minutes in class to introduce sampling. The exercises on the activity sheet will take an additional 20 minutes and can be used as homework or quiz problems.

Materials Required

- Minitab 19 or Minitab Express

Assessment

The activity sheet contains exercises for students to assess their understanding of the learning targets for this lesson.

Possible Extensions

The instructor may want to do this lesson before the ***Describing Data Numerically*** and ***Describing Data Graphically*** lessons since they refer to sample data drawn from a population.

References

Case Study I: The 1936 *Literary Digest* poll:
http://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html

Barron's *How to Prepare for the AP Statistics*, 3rd Edition, by Marty Sternstein.

Instructor Notes with Examples

Sampling Definitions

Definition: A ***population*** is the entire set, actual or conceptual, of subjects in a survey or study for which the question of interest is asked. It is the entire set of subjects about which the statistician wants to draw conclusions.

The population ***must*** be clearly defined – including the time frame – for a sample to be drawn from it. Sampling errors often occur because the population of interest is not clearly defined at the onset of a study.

Definition: A ***parameter*** is a **numerical measure** that characterizes an entire **population**.

Definition: A ***census*** is a study that attempts to acquire data **on every subject** in a **population**.

Definition: A ***sample*** is a subset of subjects from a population for which observations are actually made.

Definition: A ***statistic*** is a **numerical measure** that characterizes a **sample**. Each sample will have a unique set of statistics.

Definition: A ***survey*** is a study that collects data **on every subject** in a **sample**.

Summary of Populations versus Samples

Population	Sample
Entire set of subjects	Subset of subjects
Described using parameters	Described using statistics
Specified using Greek letters: μ, σ	Not specified using Greek letters: $\bar{x}$, s
Data collected with a census	Data collected with a survey
Related to the field of probability	Related to the field of statistics

Fundamental Idea behind Sampling

If data is to be used to make either practical or statistical decisions about a population, then ***how*** the data is collected is critical. For sample data to provide reliable information about a population, it must be representative of that population!

Some statisticians like to compare sampling from a population to tasting a spoonful of soup from a large pot. Suppose you are having a large party and you make a huge pot of chili using many different ingredients and seasonings. Some of your guests don't like fiery foods, so you need to check the spiciness of the chili.

Some things to consider:

- How would you sample the chili? You stir the pot, reach in with a spoon, take out a little bit, and taste it. You draw your conclusion about the whole pot of chili based on the taste of the spoonful.
- If your spoonful is taken in a "fair" manner, then you get a good idea of how spicy the chili is without tasting all of it.

- As with tasting chili, you need to make sure your population is properly "stirred" so that the sample represents the entire population. Not doing so can lead you to make false conclusions about the larger population.
- If your spoonful is taken fairly, then using a larger spoon will not provide additional information about the chili's spiciness. In the same manner, taking a larger sample from a population does not necessarily provide a more representative sample or additional information.

In statistics, we use samples to help us draw conclusions about the population. If the samples are not representative of the population from which they are drawn, then conclusions based on that sample are not necessarily valid. Below are examples of two real cases of poor sampling.

Example 1

1936 Presidential Election. Reference: "Case Study I: The 1936 *Literary Digest* poll"

The U.S. presidential election of 1936 was between Alfred Landon, the Republican governor of Kansas, and the incumbent President, Franklin D. Roosevelt.

At that time, there was a popular general interest magazine called *Literary Digest* that polled American citizens before the 1936 election to make predictions about the outcome.

- A **survey**, asking participants to fill out a mock ballot, was mailed to 10 million people whose names came from every telephone directory in the United Sates, lists of magazine subscribers, rosters of clubs and associations, and other sources. Approximately 2.4 million people mailed back their ballots.
- Based on the returned ballots, the *Literary Digest* predicted that Landon would get 57% of the vote versus Roosevelt's 43%. These values are **statistics** since they were computed with sample data.
- Along with the *Literary Digest's* prediction, George Gallup predicted a Roosevelt victory based on a much smaller sample of about 50,000 people.

Who won the presidential election in 1936? *Literary Digest's* prediction was incorrect – Roosevelt won with 62% of the vote.

What went wrong with the *Literary Digest's* prediction?

- Reason #1: **Selection Bias**
  - Names on the mailing list for the mock ballots were taken from telephone directories, club membership lists, lists of magazine subscribers, etc.
  - In 1936 toward the end of the Great Depression, having a telephone was a luxury, and only middle-class and upper-class citizens had them.
  - Also, since there were nearly 9 million people unemployed at that time, the names of these citizens were not on club membership or magazine subscription lists.
  - In other words, the *Literary Digest's* prediction was based on a sample that was not representative of the larger population! Statisticians call this **selection bias.**
- Reason #2: **Nonresponse Bias**
  - Out of the 10 million people who were mailed a mock ballet, only 2.4 million people replied.
  - People who tend to respond to surveys and those who do not are different types of people. Often people who respond to surveys are those who feel strongly about an issue or choice.
  - When a response rate is low (e.g. 24%), statisticians call this **nonresponse bias.** Nonresponsive people are excluded from the sample, and those peoples' views may differ from the responders' views.

Important conclusions from this real-world example:

- A large, poorly-chosen sample is worse than a small, well-chosen sample. Gallup made the correct prediction using a much smaller sample that represented the true population of interest.
- When choosing a sample, **sampling bias** in the form of selection or nonresponse bias should be avoided. However, it can be difficult to avoid nonresponse bias in surveys since there are people who will fill out surveys and those who will not, even when financial prizes or discounts are involved.

Types of Bias

Here are more formal definitions of bias:

Definition: A sampling method has **sampling bias** if all subjects in the population are not equally likely to be included in a sample.

Definition: Selection bias is a type of sampling bias that occurs when objects are selected from the population in a non-random fashion. With selection bias, the exclusion of certain objects from possible samples affects statistical results based on those samples.

Definition: Nonresponse bias is a type of sampling bias that occurs because of the absence of certain objects or subjects from a sample. For example, some subjects don't respond to surveys because they refuse, cannot be contacted, or have a lack of interest in the survey content.

Bias occurs when:

- Subjects self-select into a sample. The resulting sample is called a **voluntary sample**.
- Subjects in a sample are chosen for simplicity. The resulting sample is called a **convenience sample**.
- Subjects are arbitrarily chosen, not randomly selected. The resulting sample is called a **haphazard sample**.

Example 2

Ann Landers Advice Column

In the 1970's, famed newspaper advice columnist Ann Landers asked her readers "If you had it to do over again, would you have children?" She received nearly 10,000 responses with 70% saying "NO!" A few weeks later, her column was headlined: "70% OF PARENTS SAY KIDS NOT WORTH IT."

- The results were biased because of the **nonresponse** of people who were happy with their children and did not respond to the survey. The respondents to the survey did not represent all parents.
- A **voluntary sample** was created since the majority of respondents were parents who felt strongly enough about not having kids to take the time to write to Ann Landers. Their letters showed that many of them were angry at their children.
- After Ann Lander's survey, a similar survey was conducted using more sound random selection techniques, and this survey found that 91% of parents *would* have children again.

4-C The Grand Rapids Press, Monday, O

Ann Landers

Hairdo obstructs theatergoer's view

Dear Ann Landers: My husband and I live in the Middle West. We go to New York once a year to visit the art galleries and Broadway shows. Something extraordinary happened on our last trip. What would you have done?

We had a pair of choice (and expensive) seats for a terrific musical — fifth row center. Three minutes before curtain time a couple arrived and sat down directly in front of us. The woman had a two-story hairdo which stood mysteriously erect — about nine inches above her empty skull. She blocked my line of vision completely.

I was angry and frustrated. If she had been wearing a hat I'd have known what to say. But you can't tell a woman to take down her hair. This may never happen to me again in a hundred years, but if it does — what should I do? Or, to put it differently — what would YOU have done?

— Anne Noyd

Dear Anne: I would have tapped the woman on the shoulder and, in a voice dripping with molasses, said, "Your hair-do is simply beautiful but it's interfering with my ability to see. Would you mind moving your head about five inches to the left?"

Dear Ann Landers: This is no big deal but if my husband sees it in print it would help me a lot. Question: When a couple is out in public, who is supposed to carry the baby?

According to my husband it is proper for the man

Sampling Methods

There are many different ways to sample from a population to help ensure that the sample is not biased and is representative of the population. Two of these sampling methods are described below.

Definition: A ***simple random sample* (*SRS*)** is a sample of size n such that every collection of size n is equally likely to be sampled. This is the most commonly used method for random sampling.

One way to obtain a simple random sample is to use a **lottery method**:

- Each of the N objects in the population is assigned a unique number less than or equal to N. For example, if there are 100 objects in a population, then any number from 1 to 100 will be assigned to each object where no two objects have the same number.
- The numbers are placed in a bowl and thoroughly mixed.
- A blind-folded researcher selects n numbers.
- Population members having the selected numbers are included in the sample.

Definition: A ***stratified sample*** is a sampling method in which the population is first divided into groups, or strata, based on some characteristic prevalent in the population. Within each group, a simple random sample is taken.

Advantages of using a stratified sample instead of a simple random sample:

- Stratifying can help to ensure that the sample reflects the diversity of the population. Under-represented groups can still be represented in the sample.
- Stratifying can help to avoid samples that do not represent the population, even when randomization is used (e.g. a sample of only sophomores from an entire university population).
- We can obtain information about the sample subgroups, as well as the entire sample.

Example 3

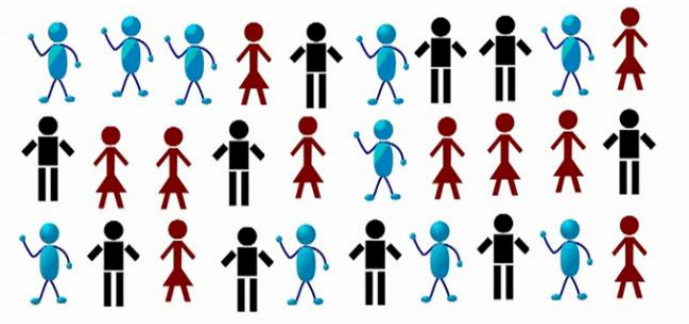
Suppose we are interested in computing the mean GPA at a certain university. We decide that an appropriate sample size with respect to the size of our population is $n = 100$. To estimate the mean GPA, we could use a simple random sample to select 100 students and average their GPA's. Since freshmen GPA's tend to be lower than senior GPA's, we want to make sure both classes are represented in our sample, so we decide to use a stratified sample.

According to the university's registrar, the student body consists of 35% freshmen, 25% sophomores, 20% juniors, and 20% seniors. We will sample from each stratum proportional to its size. Specifically, we will take simple random samples of 35 freshmen, 25 sophomores, 20 juniors, and 20 seniors. We will then average the GPA's of these students to estimate the GPA for the entire university.

Instead of class, we could subgroup the population by academic major. We need to make sure each student is assigned one major. When stratifying into subgroups, the subgroups must be mutually exclusive. If they aren't, then some subjects will have a higher probability of being chosen since they reside in more than one subgroup.

Example 4

Choose a random sample of size $n = 9$ using simple random sampling. Then create a sample of size $n = 9$ using stratified sampling. The population size is $N = 30$ and the various colors represent different subgroups.



Simple Random Sampling

Assign each person a unique number from 1 to 30. In Minitab, create a column of integers from 1 to 30. Then randomly sample 9 numbers from that column, without replacement. Select the 9 subjects according to the numbers that Minitab has chosen.

To create a column of integers, 1 through 30, in a blank Minitab worksheet:

Minitab 19 (Mac and PC)

1 Choose **Calc > Make Patterned Data > Simple Set of Numbers**.
2 In **Store patterned data in**, type *C1*.
3 In **From first value**, type *1*. In **To last value**, type *30*. **In steps of**, type 1.

4 In **Number of times to list each value**, type *1*. In **Number of times to list the sequence**, type *1*.
5 Click **OK**.

	C1	C2
1	1	19
2	2	9
3	3	22
4	4	21
5	5	5
6	6	3
7	7	30
8	8	16
9	9	2
10	10	

Minitab Express

1 Open the generate patterned numeric data dialog box.
   - Mac: **Data > Generate Patterned Data > Numeric**
   - PC: **DATA > Patterned Data > Numeric**
2 From **Form of data**, select **Equally spaced numbers**.
3 In **First number**, enter *1*.
4 In **Last number**, enter *30*.
5 In **Number of times to list each value**, enter *1*.
6 In **Number of times to list the whole sequence**, enter *1*.
7 Click **OK**.

To randomly draw a sample of size $n = 9$ from column C1:

Minitab 19

1 Choose **Calc > Random Data > Sample From Columns**.
2 In **Number of rows to sample**, enter *9*. In **From columns**, enter *C1*.
3 In **Store samples in**, enter *C2*, then click **OK**.

Minitab 19 Mac

1 Choose **Calc > Sample From Columns**.
2 In **Number of rows to sample**, enter *9*. In **From columns**, enter *C1*.
3 In **Store samples in**, enter *C2*, then click **OK**.

Minitab Express

1 Open the sample from columns dialog box.
   - Mac: **Data > Sample from Columns**
   - PC: **DATA > Sample from Columns**
2 In **Take a sample from the following columns**, enter *C1*.
3 In **Number of rows in each sample**, enter 9.
4 From **Method**, select **Sample without replacement**.
5 Click **OK**.

Stratified Sampling

Separate the population by color into 3 strata: blue, black, and red. Each strata has 10 members, so the strata are equally weighted. Assign values 1 to 10 to the team members in each strata. Then use simple random sampling for each stratum with sample size $n = 4$.

Using steps similar to those above, create a column of integers 1 through 10, using a blank Minitab worksheet. Then randomly draw a sample of size $n = 4$ from column C1 and repeat this step two more times, once for each color.

↓	C1	C2	C3	C4
		Blue	Black	Red
1	1	6	3	5
2	2	2	1	7
3	3	10	7	6
4	4	4	5	4
5	5			
6	6			
7	7			
8	8			
9	9			
10	10			

Sampling with versus without Replacement

When **sampling with replacement**, an object in a population can be selected *more than* once. When **sampling without replacement**, an object in a population can be selected *only* once.

For example, suppose we use the lottery method to select a simple random sample:

- After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl.
- If we put the number back in the bowl, it may be selected more than once. In this case, we are sampling **with** replacement.
- If we put the number aside, it can be selected only one time. In this case, we are sampling **without** replacement.

Independent events in statistics and probability are easier to analyze and result in simpler formulas. Two events or objects are **independent** of each other when the selection of one does

not influence the selection of another. This is where sampling with replacement becomes important.

Why Sample with Replacement?

Sampling with replacement results in independent events that are unaffected by previous outcomes.

When selecting a relatively small sample from a large population, obtaining a sample of independent subjects occurs whether we sample with replacement or without replacement.

- For example, if we sample from a bathtub full of M&M's, we don't need to sample with replacement because drawing one red M&M doesn't influence the next color to be drawn. We can say that the M&M's are **independent** of each other, which means the selection of one M&M doesn't influence the selection of another.

When the population is small, obtaining a sample such that subjects are independent is difficult.

- For example, if we sample from a small bag of M&M's, removing one red can influence the next color to be drawn. In this case, we would sample with replacement.

Example 5

A very tired professor has 10 keys on her key ring, and one of them is the key to her locked office door. Should she sample the keys with or without replacement?

- If she randomly tries the keys one by one, but does not eliminate the ones she tries, then she is sampling with replacement. In this case, the long-run average number of tries to unlock her door is 10.
- If she randomly tries the keys one by one, eliminating the ones that do not work, then she is sampling without replacement. In this case, the long-run average number of tries to unlock her door is 5.5.

In this case, sampling without replacement makes sense.