

Introduction to Data Science

Unit 3

Introduction to Data Science Daily Overview: Unit 3

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Testing, Testing... 1, 2, 3... (7 days)	1	Lesson 1: Anecdotes vs. Data		Reading articles critically, data	7
	2	Lesson 2: What is an Experiment?		Experiments, causation	10
	3	Lesson 3: Let's Try an Experiment!		Random assignments, confounding factors	13
	4	Lesson 4: Predictions, Predictions		Visualizations, predictions	15
	5	Lesson 5: Time Perception Experiment		Elements of an experiment	17
	6	Lab 3A: The results are in!		Analyzing experiment data	19
	7	Practicum: Music to my Ears		Design an experiment	20
Would You Look at That? (4 days)	8	Lesson 6: Observational Studies		Observational study	23
	9	Lesson 7: Observational Studies vs. Experiments		Observational study, experiment	25
	10	Lesson 8: Monsters that Hide in Observational Studies		Observational study, confounding factors	27
	11	Lab 3B: Confound it all!		Confounding factors	31
Are You Asking Me? (9 days)	12	Lesson 9: Survey Says...		Survey	35
	13	Lesson 10: We're So Random		Data collection, random samples	38
	14	Lesson 11: The Gettysburg Address		Sampling bias	42
	15	Lab 3C: Random Sampling		Random sampling	47
	16	Lesson 12: Bias in Survey Sampling		Bias, sampling methods	49
	16	Lesson 13: The Confidence Game		Confidence intervals	52
	17	Lesson 14: How Confident Are You?		Confidence intervals, margin of error	55
	18	Lab 3D: Are You Sure about That?		Bootstrapping	57
What's the Trigger? (5 days)	19	Practicum: Let's Build a Survey!		Non-biased survey design	60
	20	Lesson 15 Ready, Sense, Go!		Sensors, data collection	63
	21	Lesson 16: Does it have a Trigger?		Survey questions, sensor questions	66
	22	Lesson 17: Creating Our Own Participatory Sensing Campaign		Participatory sensing campaign creation	69
	23	Lesson 18: Evaluating Our Own Participatory Sensing Campaign		Statistical questions, evaluate campaign	72
Webpages (6 days)	24 [^]	Lesson 19: Implementing Our Own Participatory Sensing Campaign	Class Campaign—data	Mock-implement campaign, campaign creation, data collection	74
	29	Lesson 20: Online Data-ing	Class Campaign—data	Data on the internet	78
	30	Lab 3E: Scraping web data	Class Campaign—data	Scraping data from the internet	82
	31	Lab 3F: Maps	Class Campaign—data	Making maps with data from the internet	84
	32	Lesson 21: Learning to Love XML	Class Campaign—data	Data storage, XML	86
	33 ⁺	Lesson 22: Changing Orientation	Class Campaign—data	Converting XML files	88
End of Unit Project (5 days)	34	Practicum: What Does Our Campaign Data Say?	Class Campaign	Statistical questions, visualizations, numerical summaries	90
	35-40	End of Unit Project: TB or Not TB	Class Campaign	Simulation using experiment data	91

[^]=Data collection window begins.

⁺=Data collection window ends.

IDS Unit 3: Essential Concepts

Lesson 1: Anecdotes vs. Data

Data beat anecdotes. In science, we need to closely examine the quality of evidence in order to make sound conclusions. Anecdotes can contain personal bias, might be carefully selected to represent a particular point of view, and, in general, may be completely different from the general trend.

Lesson 2: What is an Experiment?

Science is often concerned with the question "What causes things to happen?" To answer this, controlled experiments are required. Controlled experiments have several key features: (1) there is a treatment variable and a response variable, and we wish to see if the treatment causes a change that we can measure with the response variable; (2) There is a comparison/control group; (3) Subjects are assigned randomly to treatment or control (randomized assignment); (4) Subjects are not aware of which group they are in (a 'blind'). This may require the use of a placebo for those in the control group; and (5) those who measure the response variable do not know which group the subjects were in (if both 4 and 5 are satisfied, this is a 'double blind' experiment).

Lesson 3: Let's Try an Experiment!

Randomized assignment is required to determine cause-and-effect.

Lesson 4: Predictions, Predictions

Designing an experiment requires making many decisions, including what to measure and how to measure it.

Lesson 5: Time Perception Experiment

Designing and carrying out an experiment helps us answer specific statistical questions of interest.

Lesson 6: Observational Studies

Observational studies are those for which there is no intervention applied by researchers.

Lesson 7: Observational Studies vs. Experiments

Experiments are not always possible because of various factors such as ethics, cost limitations, and feasibility.

Lesson 8: Monsters that Hide in Observational Studies

Confounding factors/variables make it difficult to determine a cause-and-effect relation between two variables.

Lesson 9: Survey Says...

Surveys ask simple, straightforward questions in order to collect data that can be used to answer statistical questions. Writing such questions can be hard (but fun)!

Lesson 10: We're So Random

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson 11: The Gettysburg Address

Statistics vary from sample to sample. If the typical value across many samples is equal to the population parameter, the statistic is 'unbiased.' Bias means that we tend to “miss the mark.” If we don't do random sampling, we can get biased estimates.

Lesson 12: Bias in Survey Sampling

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to ‘center’ on the population parameter value.

Lesson 13: The Confidence Game

We can estimate population parameters. This means that we can give an estimate “plus or minus” some amount that we are confident contains the true value (the population parameter).

Lesson 14: How Confident Are You?

We can estimate population parameters. This means that we can give an estimate “plus or minus” some amount that we are confident contains the true value (the population parameter).

Lesson 15 Ready, Sense, Go!

Sensors are another data collection method. Unlike what we have seen so far, sensors do not involve humans (much). They collect data according to an algorithm.

Lesson 16: Does it have a Trigger?

A key feature that distinguishes the way sensors collect data from more traditional approaches is that sensors collect data when a 'trigger' event occurs. In Participatory Sensing, this event is something we humans agree upon beforehand. Every time that trigger happens, we collect data.

Lesson 17: Creating Our Own Participatory Sensing Campaign

Creating a Participatory Sensing Campaign requires that survey questions must be completed whenever they are “triggered”. Research questions provide an overall direction in Participatory Sensing Campaign.

Lesson 18: Evaluating Our Own Participatory Sensing Campaign

Statistical questions guide a Participatory Sensing Campaign so that we can learn about a community or ourselves. These Campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson 19: Implementing Our Own Participatory Sensing Campaign

Practicing data collection prior to implementation allows optimization of a Participatory Sensing Campaign.

Lesson 20: Online Data-ing

We stretch students' conception of data, to help them see that many web pages present information that can be turned into data.

Lesson 21: Learning to Love XML

XML is a programming language that we use with our campaigns. We create basic XML "tags" in the code, which help us store data in a format we understand.

Lesson 22: Changing Orientation

Converting XML to spreadsheet format helps us better understand and view our data.

Testing, Testing...1, 2, 3...

Instructional Days: 7

Enduring Understandings

An experiment is a data collection method in which the effects of different treatments on an outcome of interest are measured. In an experiment, a treatment is applied to subjects and then observations about the effect of the treatment are made. To isolate the effects from unexplained variation, randomization (or chance) assignment to treatments is applied.

Engagement

Students will view Hans Rosling's video *How Not to Be Ignorant About the World* and will participate in his interactive quiz in order to learn how anecdotes and personal experience can influence what we know and, alternatively, how data provides basis for evidence. The video can be found at:

https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world

Learning Objectives

Statistical/Mathematical:

S-IC 1: Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Focus Standards for Mathematical Practice for All of Unit 3:

SMP-1: Make sense of problems and persevere in solving them.

SMP-4: Model with mathematics.

SMP-8: Look for and express regularity in repeated reasoning.

Data Science:

Understand that differences between the measured outcomes of the treatment and control groups in an experiment can be tested. Understand the roles of randomization and of random sampling in statistical inference.

Applied Computational Thinking:

- Test for differences between experimental groups.
- Create graphical representations to compare data between experimental groups.
- Write code to randomly assign subjects to treatment groups

Real-World Connections:

Experiments are used to ensure safety and efficacy of medicines, reliability of electronics and structural materials and find patterns in human behavior.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

1. Students will gather data generated through a simple experiment.

Data File:

1. Students' *Time Perception* experiment data.

Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

Lesson 1: Anecdotes vs. Data

Objective:

Students will learn the difference between anecdotes and data. They will begin to read articles critically to discern whether the evidence presented is based on anecdotes or data.

Materials:

1. Hans Rosling's video *How Not to Be Ignorant About the World* found at https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world
2. Article: *Miracle at the KK Café* (also available in the LMR folder) <https://archives.sfweekly.com/sanfrancisco/miracle-at-the-kk-cafe/Content?oid=2144741>
3. Article: *Can Trophy Hunting Actually Help Conservation?* (also available in the LMR folder) <https://lastwordwildlife.com/2014/01/21/can-trophy-hunting-actually-help-conservation/>

Vocabulary:

anecdote, data

Essential Concepts: Data beat anecdotes. In science, we need to closely examine the quality of evidence in order to make sound conclusions. Anecdotes can contain personal bias, might be carefully selected to represent a particular point of view, and, in general, may be completely different from the general trend.

Lesson:



1. Prepare your video player to show the first 5 minutes and 23 seconds of Hans Rosling's video *How Not to Be Ignorant About the World* found at: https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world
2. Ask students to play along as they watch the video. Each time Hans Rosling asks the audience to choose an answer to each of the three questions, pause the video for about 5 seconds and ask students to write down what they think is the answer to each question.



3. After viewing the video, engage students in *T-I-P-S* (see strategies) with the questions below:
 - a. Why did the chimps at the zoo score better than the people? **Answer: Anyone can select the correct answer just by chance.**
 - b. On the second question, Hans Rosling says that "everyone is aware that there are countries and there are areas where girls have great difficulties and they are stopped when they go to school." How could this information influence the answer choice? **Answer: Personal knowledge and experiences can influence what we think we know.**
 - c. Why do you think only a few people know the correct answer to these three questions? **Answer: People do not know enough about the data that can help them answer these questions.**

4. Display the following statements to students:

- "My skin glows more...I feel pretty confident." Melissa for Proactiv®
- "Within four months, I'd lost a grand total of 63 pounds* and was down to my goal weight." Marianne G. for Nutrisystem®
- "The customer service is obnoxious. The employees are patronizing, smug, and intractable." Seymour773 for Bank of America®



5. Discuss each statement with students by asking the following questions:
 - a. Is _____ a good product? **For example, is Proactiv® a good skin product?, is Nutrisystem® a good diet program?, is Bank of America® a good bank?**

- b. Do you think this person's experience is "typical?" Why? *Maybe it is typical but maybe not. Their own experience might be very different.*
 - c. Do you think the company chose this person? How do you know? *Each company may have chosen the first 2 statements because they were a success. In the case of the Seymour773, a competing company may have chosen his experience to make them appear better.*
 - d. What about all the other people? How many were successes, how many failures? *We don't know for sure.*
 - e. How could we answer such questions? *Collect data!*
6. Inform students that the statements are called testimonials and they are examples of **anecdotes**. Anecdotes are stories that someone tells about his/her own experience or the experience of someone he/she knows. Anecdotes are good for some things like witness statements in a police report but are not useful for reaching conclusions about groups of people because the assumptions they are based on are not always true. Their claims are easily debunked. Many anecdotes do not equal data.

Note to teacher about witness statements: Lots of evidence suggests that witness testimony needs to be examined very closely. "As perhaps the single most effective method of proving the elements of a crime, eyewitness testimony has been vital to the trial process for centuries. However, the reliability of eyewitness testimony has recently come into question with the work of organizations such as The Innocence Project, which works to exonerate the wrongfully convicted. This thesis examines previous experiments concerning eyewitness testimony as well as court cases in which eyewitnesses provided vital evidence in order to determine the reliability of eyewitness testimony as well as to determine mitigating or exacerbating factors contributing to a lack of reliability." Information gathered from digitalcommons.liberty.edu

- 7. On the other hand, **data** are a series of observations, measurements, or facts. Data are information and tell a story.
- 8. Quickly survey students about whether the video they watched at the beginning of class is based on anecdotes or data. Then inform students that they will analyze two articles to find out if their claims are based on anecdotes or data.
- 9. Students will read one of two articles, *Miracle at the KK Cafe* or *Can Trophy Hunting Actually Help Conservation?* to analyze whether the claims each makes are based on anecdotes or data. The articles can be found at the following links or in the LMR folder:

Miracle at KK Café

<https://archives.sfweekly.com/sanfrancisco/miracle-at-the-kk-cafe/Content?oid=2144741>

Can Trophy Hunting Actually Help Conservation?

<https://lastwordwildlife.com/2014/01/21/can-trophy-hunting-actually-help-conservation/>

- 10. Ask students to number themselves off as 1 or 2. Students whose number is 1 will read *Miracle at KK Café* and those whose number is 2 will read *Can Trophy Hunting Actually Help Conservation?*
- 11. Ask students to find a partner with the same number.
- 12. Before reading, ask students what they think their article will be about. Have students pair-share their thoughts.
- 13. During reading, students will take turns reading each paragraph out loud to each other. The student listening will verbally summarize what his/her partner just read.
- 14. After reading, student pairs will answer the following questions in their DS Journals:
 - a. What was the article about?
 - b. What claim(s) was/were this article making? Cite examples from the article.
 - c. Was this article based on anecdotes or data? Cite examples from the article.
 - d. How believable are the claims?



15. After answering the questions, students will find a partner with a different number. Each student will report to their new partner the following information about the article he/she read:

- a. The name and publisher of the article.
- b. His/her response to the four questions in the DS journal.

16. Quickly survey students about which article was based more on anecdotes and which one was based more on data. Ask a couple of students to explain their choices and give examples.

Miracle at KK Café makes claims that are anecdotal. Students may cite a customer's claim as an example of an anecdote. Can Trophy Hunting Actually Help Conservation? uses data to make their claims. Students may refer to a statistic used in the article as an example.



17. Class discussion: ***Data Beat Anecdotes!*** Ask students to come up with reasons why this statement is true. Have them come up with situations where you **have** to have an anecdote. For example, if asked what it's like to walk on the moon, only a few people would be able to tell us.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will do a *Last Word Review* for the words DATA and ANECDOTE.

Last Word Review: Write the word vertically. Students come up with a word or phrase for each letter of the word. Each letter of the word should summarize something about what the students learned about the topic.

Lesson 2: What is an Experiment?

Objective:

Students will learn about the elements of an experiment and the meaning of "causation". Students will learn to distinguish claims of causation from claims of association.

Materials:

1. Video: MythBusters' *Is Yawning Contagious?*
<https://www.discovery.com/tv-shows/mythbusters/videos/is-yawning-contagious>

Note: If video is not found using link, please use a search engine (e.g., Google Video) and type "MythBusters Is Yawning Contagious" to find it. The clip is a little over 5 minutes in length.

Vocabulary:

experiment, subjects, treatment, treatment group, control group, random assignment, representative sample, outcome, research question, confounding factors

Essential Concepts: Science is often concerned with the question "What causes things to happen?" To answer this, controlled experiments are required. Controlled experiments have several key features: (1) there is a treatment variable and a response variable, and we wish to see if the treatment causes a change that we can measure with the response variable; (2) There is a comparison/control group; (3) Subjects are assigned randomly to treatment or control (randomized assignment); (4) Subjects are not aware of which group they are in (a 'blind'). This may require the use of a placebo for those in the control group; and (5) those who measure the response variable do not know which group the subjects were in (if both 4 and 5 are satisfied, this is a 'double blind' experiment).

Lesson:

1. Display the following headlines to students:
 - a. Stop Global Warming: Become a Pirate
 - b. Lack of sleep may shrink your brain
 - c. Early language skills reduce preschool tantrums
 - d. Dogs walked by men are more aggressive



2. Discuss each headline by asking the following questions:
 - a. What is the headline implying with its wording? *1a is implying that you can stop global warming by becoming a pirate, 1b is implying that it's possible to shrink your brain if you aren't getting enough sleep, 1c is implying that having early language skills will decrease preschool tantrums, 1d is implying that dogs are more aggressive when they've been walked by men.*
 - b. Is it implying causation or association? *Discuss definitions of causation and association. Causation means there is a cause and effect relationship between variables. For example, heat causes water to boil; whereas association or correlation means that high values of one variable tend to be associated with high values of the other (or high values tend to be with low values). However, this is not necessarily cause-and-effect at play. For example, blanket sales in Canada are associated with brush fires in Australia - not because Canadian blankets cause the fires, but because Canadian winters cause blanket sales, and Canadian winters are Australian summers, which cause fires. 1a, 1c and 1d are implying causation and 1b is implying association.*
 - c. How can you tell the difference between causation and correlation? What words stand out in these headlines? *Answers will vary but some terms for causation include: cause, increase/ decrease, benefits, impacts, effect/ affect, etc.; and for correlation*

include: get, have, linked, more/ less, tied, connected, etc. In 1a, “become” stands out; in 1b, “may” stands out; in 1c, “reduce” stands out; in 1d, “are” stands out.

- d. Change each causal version of a headline into a non-causal version and vice versa. *Answers will vary but an example for 1a is to instead say Global Warming linked to increase of pirates.*

3. Introduce the MythBusters video clip by answering the following questions, in teams, for their headline “Is Yawning Contagious?”

- What is the headline implying with its wording? *That yawning may cause other people to yawn.*
- Is it implying causation or correlation? How do you know? *Causation because “contagious” yawns means that you are yawning because someone else has yawned.*
- How can we determine if this is true? *Split the class into groups and have each team come up with a way to determine if this is true. Each group should assume that they get to examine 50 people.*



4. Show the MythBusters video clip called *Is Yawning Contagious?* The clip can be found at:

<https://www.discovery.com/tv-shows/mythbusters/videos/is-yawning-contagious>

5. Focus students on the following guiding questions and ask them to take notes as they watch the video clip:

- How did the MythBusters design the investigation?
- What steps did they take?
- How is this different than your team’s headline responses?

6. After viewing the clip, inform students that the MythBusters have just conducted an **experiment**, which is one method of data collection.

7. We begin with a brief introduction into “what is an experiment” but the definition will be developed over the next several lessons.

8. Guide students to identify the elements of an experiment by referring back to the video clip:

- Research Question**—the question to be answered by the experiment (*Is Yawning Contagious?*)
- Subjects** – people or objects that are participating in the experiment (*the 50 adults*)
- Treatment** – the procedure that is assigned to a group of subjects (*Kari yawned before subject entered the room*)
- Treatment group** – the group of subjects that receive the treatment (*two out of every three subjects who were placed into rooms – yawn from Kari*)
- Control group** – the group that does not receive a treatment (*one out of every three subjects who were placed into rooms – no yawn from Kari*)
- Random assignment** – subjects are randomly assigned to either the treatment or control group (*two out of every three subjects received the treatment*) **Note:** In this experiment, random assignment was not used (or if it was, we were not told so.)
- Outcome** – the variable that the treatment is meant to influence. (*whether or not a person yawned*)
- Statistic**—A method for comparing the outcomes of the control and treatment groups is needed. *In this case, the MythBusters used the difference between the percent of subjects that yawned in the treatment group was 4% higher than the control group.*



Note: In this experiment, and in those found in the IDS curriculum, we use a treatment and a control group. However, a control group is not a *necessary* element of an experiment. Sometimes it is more appropriate to have two treatment groups with no control group (e.g., medical professionals testing different doses of drugs). The effect that is being studied will dictate whether to feature a control group or not.

9. Display the following questions on the board or projector. Using *T-I-P-S*, ask students to discuss them.

- a. Why did the MythBusters follow all of these steps to design their experiment? *In order to determine if watching someone yawn can cause you to yawn.*
 - b. We don't know how MythBusters chose who would be in the treatment group and who would be in the control group. Suppose that the people who showed up first, early in the morning, were assigned to the treatment group, and the last few people, later in the day, ended up in the control group. Would you believe in the conclusions? *No, because the two groups were different. The first group might have been sleepier, and so more likely to yawn anyways. Explain that this --another explanation for the cause-and-effect--is caused a confounding variable.*
 - c. Explain that in order to make the two groups as similar as possible, experimenters usually assign subjects randomly. How might we randomly assign about half of the subjects to the treatment and half to the control? *We might flip a coin, and those who get Heads go to Treatment.*
 - d. Why would random assignment improve the MythBusters study? *Because then the two groups would be more similar. So we wouldn't have a confounding variable to worry about.*
10. **Emphasize that without random assignment, we cannot determine causation because we are not comparing two similar groups.**

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 3: Let's Try an Experiment!

Objective:

Students will explore the importance of randomized assignment in experiments. They will understand that without random assignment, there might be confounding variables and will be able to suggest possible confounding variables.

Materials:

1. Measuring Tape

Essential Concepts: Randomized assignment is required to determine cause and effect.

Lesson:

1. Inform students that they will be exploring the question “Why do we need randomized assignment?” by conducting an experiment. Tell students that you have a treatment that can make people taller. Explain that the class will be divided into two groups, one group will get the treatment, and one group will not. The group that does not receive the treatment will be the control group. After the treatment, they will measure the groups to see which is taller. Now divide the class into two groups by placing the boys in the treatment group and the girls in the control group.

Remember that in an experiment we typically have a treatment group and a control group. In the MythBusters experiment, they compared number of yawns after treatment, and not any measurements before treatment, because they were comparing the treatment group to the control group (the control group is specifically here because it is a comparable untreated group - this allows us to not need “before” measurements). Therefore, in this case, we will run the experiment and *then* compare average height of the treatment group to the control group.

2. Tell them that after the treatment group takes the treatment, your statistic to compare groups will be to measure the heights. If the treatment group is taller, then the treatment must have worked. There are two possible outcomes to dividing the class this way:
 - a. The students will protest (as they should) and you can start a discussion as to why this is not a good way to divide the class.
 - b. OR the students don't protest and you continue with the experiment. The treatment should be something silly, like waving a ruler in front of the person's face or by asking them to chant “grow, grow, grow!” three times. After treatment, measure the heights of each group and ask them if they think this is good evidence (**do not say “proves”**) that the treatment is effective.
3. Regardless of the outcome, students should recognize that by putting the boys in one group, the outcome was pre-determined, since boys tend to be taller than girls to begin with. This is an example of a **confounding factor**. Confounding factors are variables that provide an alternative explanation of the effect of the treatment on the outcome variable.
4. Ask students: “How should students be put into groups?”
5. Discuss various other methods of grouping students. Someone will probably say to split the groups into equal numbers of boys and girls. At this suggestion, divide the class into two groups by placing the tallest boys and tallest girls in the treatment group, and the shorter boys and shorter girls in the control group. *Students should be able to recognize that you shouldn't use any characteristics to decide the groups.*
6. Continue discussion of other ways to decide the groups. Use the following questions as a guide:
 - a. What about flipping a coin?
 - b. What will the gender balance look like? *Each group should have about the same balance as the class, though not exactly.*



- c. Why is it important that the groups be similar? *Because otherwise, something else might be the cause of the response changing.*
7. Inform students that today the class will begin to design their own experiment using what they have learned over the last few lessons. The question they will investigate is:

How does our perception of time change when exposed to a stimulus?

8. They will be trying to determine the length of one minute without the use of time-aids. In their experiment, they will subject some students to a stimulus and others to no stimulus. They will then analyze the data to determine if subjecting students to a stimulus affects the perception of how long a minute of time lasts.
9. In their DS journals, ask students to answer the following questions about the elements of their experiment:
- a. What is the research question we're interested in addressing?
 - b. Who are the subjects that will be participating in the experiment?
 - c. How should we randomly assign the subjects into treatment and control groups? (See step 12 for an RStudio method that the teacher can use)
 - d. What is the outcome variable that we will be measuring? What unit of measurement should we use?

Note: Students will decide on a treatment to apply to each group on the following day.



10. As a class, discuss the responses to the questions above (step #9, a-d) and come to a consensus for each question's answer.
11. Inform the class that they will be using the answers they have agreed upon as the final design of the class's experiment.
12. At the end of the class, the students should be assigned to the treatment or control groups using the randomization method they chose as a class in step #9c.

Note: One method to determine group assignment would be to use the class roster and the `sample()` function in RStudio. The students have a number that corresponds to their placement on the roster (i.e. student 1's last name most likely starts with an A, and then we move alphabetically through the roster). You can then use RStudio to randomly select which half of the numbers/students will be assigned to the treatment group.

```
> sample(1:30, size = 15, replace = FALSE)
```

13. Students will conduct the experiment in the next lesson.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 4: Predictions, Predictions

Objective:

Students will continue to read articles critically. They will anticipate visualizations about the data that will be collected from the class experiment and make predictions about the outcome.

Materials:

1. Article: PsyBlog's *10 Ways Our Minds Warp Time* found at: <http://www.spring.org.uk/2011/06/10-ways-our-minds-warp-time.php>
2. *Experiment Predictions* handout (LMR_3.1_Experiment Predictions)

Vocabulary:

theory

Essential Concepts: Designing an experiment requires making many decisions, including what to measure and how to measure it.

Lesson:



1. Students will read the article *10 Ways Our Minds Warp Time* found at: <http://w.spring.org.uk/2011/06/10-ways-our-minds-warp-time.php>.
2. They will read the article critically to answer the following questions (displayed or written on the board):
 - a. Who was observed and what were the variables measured? *People and their perceptions of time.*
 - b. What statistical questions were the researchers trying to answer? *How is time perception affected by different stimuli?*
 - c. Who collected the data? *Researchers such as cave expert Michel Siffre collected data.*
 - d. How were the data collected? *Data were collected through various experiments/studies (13 were cited).*
 - e. What claim(s) did the article make? *There were 10 claims made regarding time perception.*
 - f. What are some statistics that the article used to make the claim(s)? *Answers may vary. Article has several percentage statistics.*
3. In their teams, ask students to share their responses from reading the *10 Ways Our Minds Warp Time* article and agree on the responses as a team.
4. Do a quick *Whip Around* of the responses (see step #2 for responses).
5. Remind students that they designed a class experiment during the previous lesson but did not select an actual treatment. As a class, decide on a treatment to use for the experiment. Students can use the methods found in the article for inspiration, or come up with something novel on their own.

Note: Stimuli examples include music (genres determined by the class), lights off, physical activity (e.g., holding arms out), relaxation/meditation techniques, heads down, eyes closed, etc. Ensure that the experiment can be completed in **one** 50-60 minute class period. Treatments requiring excessive preparation time (e.g., running a mile) are less than ideal.
6. Before they conduct the experiment, students will test their theories by making predictions about the data and the outcomes. A **theory** is an idea used to explain a situation.
7. Display the class experiment's research question:

How does our perception of time change when exposed to a stimulus?

8. Take a poll of the students who believe that there will be differences in the estimate of the length of a minute between the treatment and control groups. The remaining students, then, do not believe that there will be differences.
9. Then, ask those students who believe there are differences, how small or large they think the difference will be.
10. Distribute the [Experiment Predictions handout \(LMR 3.1\)](#) and, in pairs, have students discuss and complete the answers for the handout.



Note: What will the distribution of time perceptions look like? *The distributions will likely have more points that are closer to 60 seconds, but will also have values that are shorter and longer than 60 seconds. Appropriate plots to use will include histograms, dotplots or boxplots.*

Name: _____ Date: _____

Experiment Predictions

Instructions:

Answer the following before conducting your time perception experiment. Remember, the variable that we're measuring is the number of seconds that actually elapse until each person believes one minute has passed.

1. In the boxes below, draw a plot of what you predict the distribution of each group's data will look like. Be sure to add numbers and labels.

Treatment
Control

2. Based on your prediction, write down how the *treatment* group's distribution will compare to the *control* group's in terms of its *center*, *shape* and *spread*.
3. What do these differences in *center*, *shape* and *spread* tell us about how people in the *treatment* group perceive time?

LMR_3.1

11. Using *Anonymous Author*, select student work to share with the whole class.
12. Give student teams time (about 2 minutes) to discuss each product that is shared/presented.
13. Teams will offer their thoughts using a modified *Two Cents* strategy where, instead of two cents, each team will receive one cent (or a token) and, in order to turn it in, the team will have to make comments or ask questions about the student work that is being shared. Call on teams until you have collected every cent. This ensures that all teams contribute to the discussion.
14. Inform students that they will conduct the experiment in which they will estimate the length of time of one minute during the next lesson.



Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 5: Time Perception Experiment

Objective:

Students will engage in a collectively designed experiment.

Materials:

2. RStudio's **stopwatch()** function
3. IDS UCLA App or Browser-Based survey-taking tool

Essential Concepts: Designing and carrying out an experiment helps us answer specific statistical questions of interest.

Lesson:

1. Begin the lesson by eliciting the elements of an experiment from students (they may refer back to their DS journals for their responses from Lesson 2).
2. Inform students that they will be using RStudio to get a precise measurement of their estimate. Ask for a student volunteer.
3. Demonstrate the stopwatch function using RStudio by typing in the following code:

```
> stopwatch()
```

4. Then, ask the student volunteer to stand in front of your computer and get ready to estimate the length of time of one minute without looking at a clock. Once he/she thinks a minute has passed, ask him/her to press the enter/return key on the keyboard to see the result of the estimate.
5. Inform students that you have just demonstrated how they will measure their one-minute estimates.
6. Begin conducting the experiment by reviewing the research question:
How does our perception of time change when exposed to a stimulus?
7. Refer back to the experiment design.
8. Review the specific treatment that the subjects in the treatment group will receive. If necessary, demonstrate to the treatment group how to do the experiment. For example, if standing with open arms is the stimulus, the estimate begins when the student starts the **stopwatch()** function and engages in the stimulus, and ends when the subject presses enter/return in RStudio to stop the timer.
9. For the control group, the students can simply sit at their desks with their eyes closed. Each student will run the **stopwatch()** function and stop the timer when they believe a minute has elapsed.
10. Conduct the experiment in its entirety. Use team roles effectively to ensure the experiment is done correctly.
11. Have each student use a computer and the **stopwatch()** function to record her/his estimate of one minute. Ensure each student records her/his estimate in the DS journal.
12. When the experiment is completed, have students enter their data in the *Time Perception* survey found in the Survey Taking Tool at <https://portal.idsucla.org> or by using the IDS UCLA App in their iOS or Android devices.
13. Inform students that they will be analyzing the results from the experiment *in Lab 3.1: The results are in!*

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3A: The results are in!

Complete Lab 3A prior to Practicum.

Lab 3A - The results are in!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Conducting experiments

- Previously in class, you conducted an experiment to gauge how a stimulus affected people's perception of time.
 - Some people were given a treatment, others were not.
- In this lab, we'll use the data cycle to analyze the *research question*:
Does the stimulus your class chose change people's perception of time?

Coming up with questions

- **Write down two statistical questions that will help you answer the *research question* from the previous slide.**
- Then, *export, upload, import* your experiment data into RStudio.
 - If you're having trouble coming up with good statistical questions, try loading the data and looking at the variables.
 - Ask yourself, *How would I use these variables to answer the research question?*

Analyzing our data

- Create appropriate plots to answer your statistical questions.
 - **Are there any outliers or unusual observations that require some cleaning before you can interpret your plots?**
- Calculate appropriate numerical summaries to answer your statistical questions.
- Interpret your plots and summaries.
 - **Write down a few sentences with your interpretations.**

Wrapping it up

- Is it possible your initial results occurred by chance alone?
 - **Use repeated shuffling to determine how likely the typical difference between the two groups occurred by chance alone.**
 - **Create a plot and use it to justify your answer.**
- What do you conclude about the *research question*?
 - **Write a report using the plots and analysis you conducted to answer the *research question*.**
 - Be sure to describe how you conducted your experiment.

Practicum: Music to my Ears

Objective: Students will design a simple experiment.

Materials:

1. Practicum: *Music to my ears* (LMR_U3_Practicum_Music to My Ears)



Note to Teacher: Before assigning the practicum to your students, engage the class in a discussion about experiments. Use the following questions as a guide to assess student understanding.

1. When is random assignment used? Why is it important? *Random assignment is used when you wish to determine whether a treatment causes changes in an outcome variable. It's important because it creates a "balance" of the groups so that the only way the groups differ, on average, is that one gets the treatment and one does not. Thus, if there is a change in the outcome variable, only the treatment could have caused it.*
2. Below are some headlines, determine if they are causal or not. If not causal, re-write so that it is. If causal, state why it's causal.
 - Straight A's in high school may mean better health later in life. *not causal, re-writing answers will vary*
 - Murder rates affect IQ test scores: Study. *causal, explanations will vary*
 - Microbe linked to Alzheimer's Disease. *not causal, re-writing answers will vary*
 - Luckiest people "born in summer" *causal, explanations will vary*
3. Why is a control group important? *The control group is important because it allows us to measure the effects of the treatment group with an untreated comparable group. Without the control group, we don't know what would have happened if we had done nothing. [Think of a new vaccine for the flu. If there is no control group, and we see the treatment group improving, we will never know if they would have improved anyways, without the vaccine.]*

Practicum Music to my Ears

In class, you designed and conducted the *Time Perception* experiment to find out if a person's perception of time changed when exposed to a stimulus. This experiment was designed so that it used random assignment, which is the process of using a chance device (e.g., dice, RStudio, etc.) to determine the placement of subjects into the treatment and control groups. By randomizing, you are removing other possible explanations for why the results happened the way they did.

Now we are asking you to design an experiment to determine whether doing math homework with music playing in the background affects student's test scores. Work with your team to design this experiment.

Submit a paper that clearly lays out your team's design plan. Be sure to include:

1. Descriptions of each element of the experiment by answering the following questions:
 - a. What is the research question we are interested in addressing?
 - b. Who are the subjects that would be participating in the experiment? How should we select them?
 - c. What could be possible treatments? What treatment do you choose and why? What will the control group do in your study?
 - d. Describe how to randomize the subjects into the treatment and control groups.
 - e. What is the outcome variable that we are measuring? Is it categorical or numerical? What other variables will you measure for each subject?
2. An analysis plan:
 - a. What statistical questions will you ask to address your research question?
 - b. What analyses (graphical and numerical) will you use to answer these questions?
 - c. An explanation of how you will determine whether the treatment affects test scores.

Would You Look at That?

Instructional Days: 4

Enduring Understandings

An observational study is a data collection method in which subjects are observed and outcomes are recorded. Unlike experiments, it may not be possible to assign subjects to treatment and control groups in observational studies, which impedes our ability to control for confounding factors. This means that researchers must rely on existing control and treatment groups to observe the outcomes. Observational studies can show associations in the data, but cause and effect relationships can only be concluded with experiments.

Engagement

Students will participate in the *Observational Studies Activity* described in Lesson 5. They will record information that can be obtained through pictures. The data will then be analyzed to see if there are any variables related to the number of friends a person has on social media.

Learning Objectives

Statistical/Mathematical:

S-IC 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that data from observational studies can help us find associations among variables. Explain why some variables that are not related in reality might look as though they are due to the presence of confounding factors.

Applied Computational Thinking using RStudio:

- Download data from the Internet that was collected via an observational study.
- Clean data set by adding variable names.
- Create scatterplots of two variables and determine possible relationships between them, as well as identify potential confounding variables.

Real-World Connections:

Economists, psychologists, and biologists conduct observational studies to study human behavior. For example, observational studies are used in epidemiology to study outbreaks of illnesses and people's behavioral patterns.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

1. Students will record information about a set of high school students by observing characteristics given in a picture.

Data File:

1. *Lung Capacity of Children* data set found at
<https://jse.amstat.org/v13n2/datasets.kahn.html>

NOTE: The raw data set can be found at
<https://jse.amstat.org/datasets/fev.dat.txt>

Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

Lesson 6: Observational Studies

Objective:

Students will learn that an observational study is a data collection method in which subjects are observed and outcomes are recorded. They will learn how to collect this type of data and make informal inferences about the results.

Materials:

1. *Stick Figures Cutouts* (LMR_1.2_ Stick Figures) from Unit 1, Lesson 1
Note: Advanced preparation required (see step 1 below).
2. *Turning Observations into Data* handout (LMR_3.2_ Observations_to_Data)

Vocabulary:

observational study

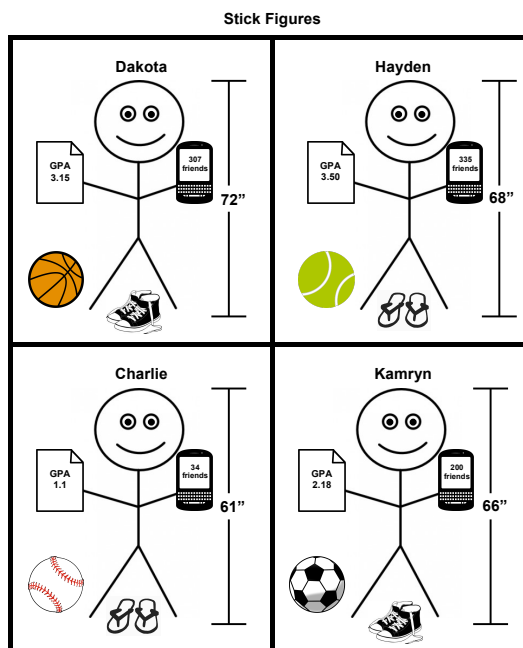
Essential Concepts: Observational studies are those for which there is no intervention applied by researchers.

Lesson:

1. From Unit 1, Lesson 1, redistribute one full set of 8 cards from the *Stick Figures* handout (LMR_1.2) to each student team.

Advanced preparation required: Print the *Stick Figures* handout (LMR_1.2). The handout can then be cut into the 8 cards. You will need enough sets of the cards for each student team to share a full set. For example, if there are 5 student teams in a class, then 5 copies of the file will need to be printed so that each team gets all 8 cards.

2. Have students recall that they used these cards in Unit 1, Lesson 1. When they used them in Lesson 1, the data was collected, recorded, and organized, but without particular structure to it.



LMR_1.2

- Then, distribute one copy per student of the *Turning Observations into Data* handout (LMR_3.2).

Name: _____ Date: _____

Turning Observations into Data

Part 1

Name the 8 variables that can be recorded from the picture on your card. What variable names could you use to represent these? Record your answers in the correct column below.

Numerical Variables	Categorical Variables

Part 2

Using the variable names you chose in Part 1, create a data table and use the first row to record the information about the person on your card.

Compare your card's values to your team members' and add their data to your table. If there are extra cards left over, record those observations as well.

LMR_3.2

- Every student from the team will then select one of the cards from the team's pile of 8, and should begin working through the *Turning Observations into Data* handout individually.
- As the students finish each part of the handout, they should compare their responses with their student teams.
- Go over the names of the variables in Part 1 by doing a quick Whip Around by teams. Then, select a couple of teams to share the information on the first row and one of the columns.
- Part 3 of the handout asks the students to consider the following research question:

What determines the number of friends a person has on social media?



- Once the students have completed the handout, discuss the variable that they thought was best associated with the number of friends on social media. *They should have seen that a person's GPA was related to the number of friends. More specifically, the higher a person's GPA, the more friends he/she had.*
- Ask a few students to share out their responses to the very last question: "Can you think of another variable (not necessarily given in the pictures) that might impact both the number of friends AND the variable you selected? Give an example and explain how it might impact each of the variables." *Answers will vary, but one example could be: a person's self-esteem level (if he/she is confident in school, his/her grades might be higher; higher confidence could also be a reason for a person having more friends).*



- Remind students that in the previous section, they learned about the elements of an experiment. In teams, ask students to discuss how collecting this data is similar or different from experiments. Then have a whole class discussion about this comparison, guiding *students to realize that there were no assignments to groups and no treatment was applied. The subjects (i.e. the people displayed on the cards) were simply observed, and then information about them was recorded.*
- Inform students that an **observational study** is a data collection method in which subjects are observed and outcomes are recorded. No treatment is applied to the subjects. Instead, researchers are simply watching something happen and have absolutely no control over it.
- In lesson 7, students will learn more about the differences between experiments and observational studies and what conclusions they can make about each.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 7: Observational Studies vs. Experiments

Objective:

Students will learn how observational studies differ from experiments, and will classify different research scenarios based on which method would be most appropriate. They will also learn about the roles of ethics, cost limitations, and feasibility when deciding between the two data collection methods.

Materials:

1. *What Should We Do?* handout (LMR_3.3_ObsStudies vs Experiments)

Vocabulary:

ethics, cost limitations, feasibility

Essential Concepts: Experiments are not always possible because of various factors such as ethics, cost limitations, and feasibility.

Lesson:

1. Remind students that in observational studies, we can never randomly assign subjects to treatment and control groups. Conversely, in experiments, we always need to have random assignment into these groups.
2. Pose the following question to students: Why can't we just always do experiments? Have students discuss this question in their student teams and write down a few responses in their DS journals.
3. Inform students that a researcher wants to perform studies to answer the research questions below. In teams, have students come up with reasons for why an experiment would not be possible for each scenario.
 - a. Does smoking cause lung cancer? *Unethical. You cannot make people smoke cigarettes and then see if they have lung cancer later in life.*
 - b. Does drinking water from Mars keep you healthier than drinking water from Earth? *Cost. It would be incredibly expensive to design a space shuttle that can successfully transport people to Mars and have them live there for an extended period of time and most researchers would not have the funding to do this.*
 - c. Do people with higher IQ scores score better on the SAT than people with lower IQ scores? *Not feasible/not possible. You cannot randomly assign IQ scores to people because it is a measurement based on aptitude.*
4. Select three teams and assign a scenario above to each team. Ask each team to report out on their assigned scenario. As teams share, be sure to discuss the following issues regarding why we cannot always do experiments:
 - a. **Ethics:** Sometimes, experiments cannot be performed because it would be unethical to give certain treatments to subjects. For example, we could not inject an HIV infection into participants because the long-term effects might lead to death.
 - b. **Cost Limitations:** Sometimes, experiments would be very costly and much too expensive to perform. Some possibilities could be with technology.
 - c. **Feasibility,** impossible to randomize: In certain cases, you cannot perform an experiment because it is impossible to randomly assign people to particular groups. For example, you cannot assign a gender to a person.
5. Distribute *What Should We Do?* handout (LMR_3.3). In teams, students will identify whether the research question could best be answered via an experiment or an observational study.



- Once all student teams have completed the handout, assign one research question to each team to report out. As each response is shared, conduct a whole-class discussion to compare which data collection method was most appropriate for each research question. Ensure everyone understands the reasons each method was chosen before moving on to the next scenario.

Note: Page 2 of the handout is an answer key for teacher reference only!

Name: _____ Date: _____

What Should We Do?
Deciding between Experiments and Observational Studies

For each research question posted in column one below, decide which type of data collection method would be best, or most appropriate. Then circle the appropriate method in column two. Explain why you chose this method in column three. Be sure to include why the other method would not be appropriate.

Research Question	Best Data Collection Method	Why this method?
Do people who live alongside freeways suffer from asthma more than people who do not live near freeways?	Observational Study OR Experiment	
How does being alone or with others affect a person's stress or chill ratings?	Observational Study OR Experiment	
Are males or females more frequently late to class?	Observational Study OR Experiment	
What are the effects of nuclear radiation 48 hours after exposure?	Observational Study OR Experiment	
Are males who play violent video games more prone to engage in violent actions than males who play E-rated video games?	Observational Study OR Experiment	
Who speaks up more when you cut the line, adults or teenagers?	Observational Study OR Experiment	
What types of rockets and fuel mixtures gets us closest to achieving the speed of light?	Observational Study OR Experiment	

LMR_3.3



- Next, student teams will generate three research questions on their own. They need to identify the best data collection method for answering their question and should provide an explanation. At least one of the three research questions should use an observational study for data collection.
- Using a share-out strategy, have the reporter of each team share one of their investigation questions with the rest of the class.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 8: Monsters that Hide in Observational Studies

Objective:

Students will learn about confounding factors that may impact the results of an observational study, which is why causation can never be concluded with observational studies, only associations between variables.

Materials:

1. Computers
2. *Spurious Correlations* website (tylervigen.com)

Vocabulary:

cause, confounding factors, associated

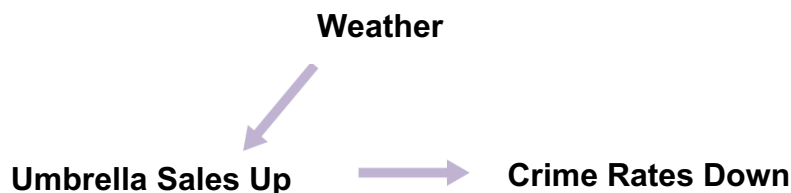
Essential Concepts: Confounding factors/variables make it difficult to determine a cause-and-effect relation between two variables.

Lesson:

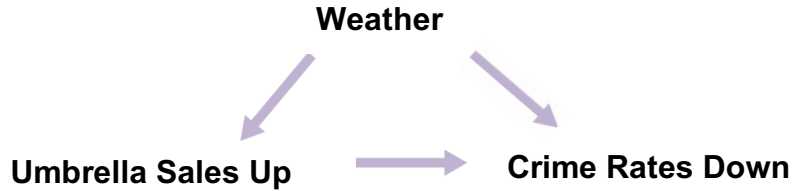
1. Ask students to recall that they looked at the relationship between a student's GPA and the number of friends that person has on social media during lesson 6. It seemed that students with higher GPAs had more friends than students with lower GPAs. But did this mean that the **cause** of a person's GPA is the amount of friends they have? NO!
2. They also identified other variables that could have contributed to the relationship, these outside variables are called **confounding factors**. Confounding factors are variables that are related to both the explanatory variable and the response variable in an observational study.
3. Propose the following statement to students: "Research suggests that a rise in umbrella sales leads to decreased crime rates."
4. Allow the students to work in teams to think about possible confounding factors. They should choose a variable that is related to umbrella sales, and that might lead to decreased crime rates. After they've come up with a few possibilities, use the following diagram progression to further explain the impact of confounding factors.
 - a. Step 1: Draw an arrow showing that "a rise in umbrella sales leads to decreased crime rates" since that is what researchers have stated.

Umbrella Sales Up → **Crime Rates Down**

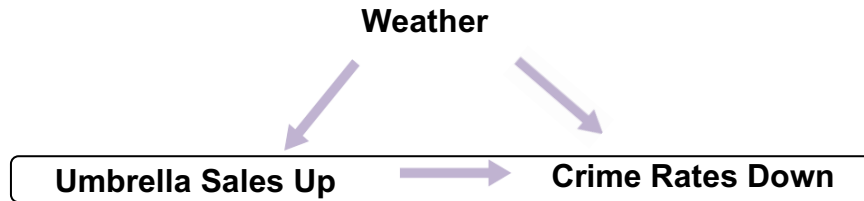
- b. Step 2: Include the variable that might be related to people buying more umbrellas (i.e., the confounding factor). For example, when the weather is rainy, people buy more umbrellas.



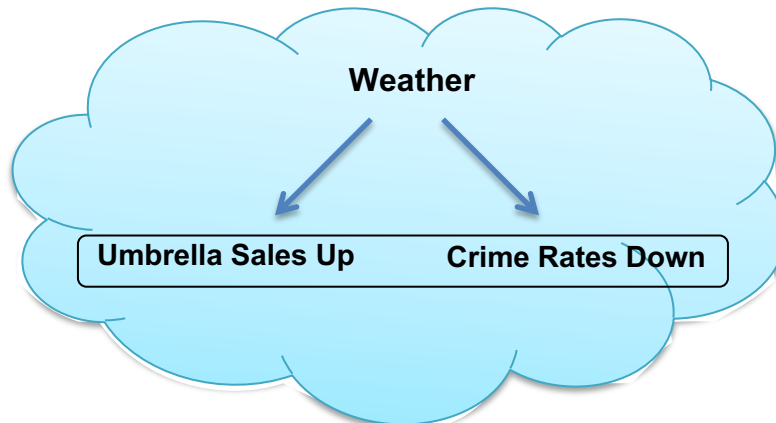
- c. Step 3: Draw an additional arrow from “Weather” to “Crime Rates Down” because it is well known that when the weather is bad, people are less likely to be outside committing crimes.



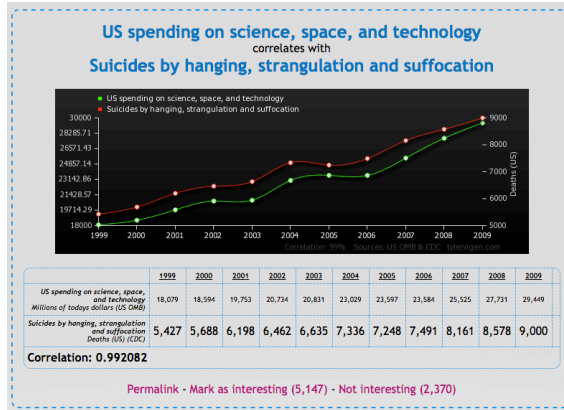
- d. Step 4: Remind students that the original claim was that “a rise in umbrella sales leads to decreased crime rates.” However, we’ve now shown that maybe buying umbrellas is not the only thing that could be contributing to a decrease in crime, which makes us question the link between the two variables.



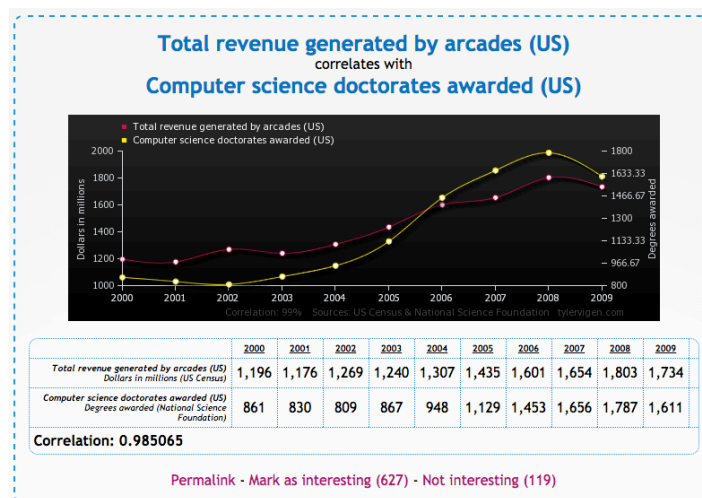
- e. Step 5: Therefore, we have found a confounding factor with the variable “crime rates.” This means we can erase the original “link” between a rise in umbrella sales and decreased crime rates since there are outside variables interfering. We can’t say buying umbrellas *causes* decreased crime rates, but we can say that a rise in umbrella sales are **associated** with decreased crime rates.



5. Once the students grasp what confounding factors are and how to identify them, introduce them to the website *Spurious Correlations* by Tyler Vigen. This site shows many explanatory and response variables that are randomly associated with each other. Spurious Correlations can be found at: <http://www.tylervigen.com/spurious-correlations>.



6. For the example given above, we see that as the US spends more money on science, space, and technology, more people are dying by way of suicide. Clearly, it does not make sense that if the US keeps spending money on science, then more people are going to commit suicide. It simply happened by chance (or a bizarre chain of confounding factors) that the two variables are related to each other.
7. Allow the students to explore the website on their own (Note: there are multiple pages of graphs, so they are not restricted to simply the homepage). They should choose a graph that interests them and answer the following questions in their DS journals:
 - a. What are the two variables shown in your graph?
 - b. Is there a positive association or a negative association between the variables?
 - c. Write an interpretation of this plot in the context of the data.
 - d. Write the data points in a "spreadsheet format" in a form that RStudio could read. Each row should represent a point on the graph, and each column one of the two variables.
 - e. By hand, make a scatterplot of the association. Describe whether the association seems strong or weak or moderate to you.
 - f. Do you think that the explanatory variable *causes* the response variable? Explain.
 - g. If you answered 'no' to f, then draw a diagram like in #4 with possible confounding factors. **Note:** this can be difficult, depending on the graph chosen. Some factors to consider: weather, economy, fashion trends.
8. Example answers to Step 7 are given below:

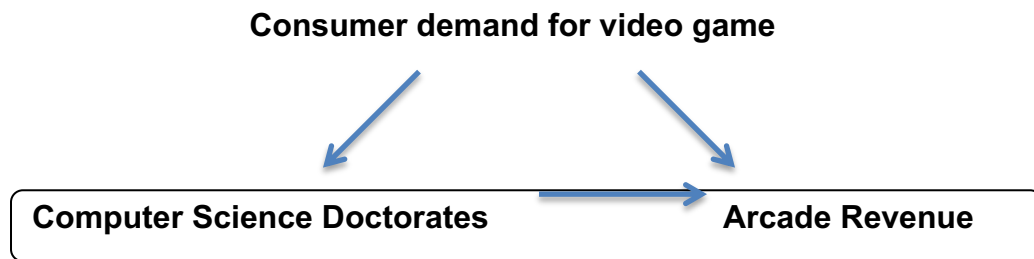


- What are the two variables shown in your graph? *Total revenue generated by arcades in the US and the number of computer science doctorates awarded.*
- Is there a positive association or a negative association between the variables? *There is a direct relationship because the lines have the same shape (they follow the same pattern).*
- Write an interpretation of this plot in the context of the data. *It seems that as more doctorates are awarded to computer scientists, arcades are generating more revenue.*

Arcade Revenue	CS doctorates
1196	861
1176	830

etc.

- Answers will vary.
- Can you conclude that the one variable causes the other? *No. Although the two variables are associated with one another, we do not have evidence to say that more doctorate awards cause arcades to make more money because the data do not come from a controlled experiment.*
- Draw a diagram like the one we did together earlier (in step 4 of lesson) with possible confounding factors. *Student's diagram should look like the one below:*



- Once all students have selected a graph and have answered the above questions, have them share their responses with a partner. They should explain why they thought their particular graph was interesting, how the two variables are related (directly or inversely), and whether or not there is a causal link between the variables.
- At the end of this lesson, students should understand that causation can only be concluded when an experiment is performed, but associations can be concluded for observational studies.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3B: Confound it all!

Complete Lab 3B prior to Lesson 9.

Lab 3B - Confound it all!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Finding data in new places

- Since your first forays into doing data science, you've used data from two-sources:
 - Built-in datasets from RStudio.
 - Campaign data from IDS Campaign Manager.
- Data can be found in many other places though, especially online.
- In this lab, we'll read an *observational study* dataset from a website.
 - We'll use this data to then explore what factors are associated with a person's lung capacity.

Our new data

- You can find the data online here:
 - (Right-click and select *Open in New Window*)
<https://raw.githubusercontent.com/IDSUCLA/dataset/main/fev.csv>
- Variables that were measured include:
 - Age in years.
 - Lung capacity, measured in liters.
 - The youth's heights, in inches
 - Genders; "1" for males, "0" for females.
 - Whether the participant was a smoker, "1", or non-smoker "0".

Importing our data

- Rather than *export*-ing the data and then *upload*-ing and *importing*-ing it, we'll pull the data straight from the webpage into R.
- Click on the *Import Dataset* button under the *Environment* tab.
 - Then click on the *From CSV* option.
 - Type or copy/paste the URL into the box and then hit *Update*.
- Before importing, change the following *Import Options*:
 - Name: `lungs`
 - *Uncheck* the *First Row as Data* box
 - Change *Delimiter* to *Whitespace*

About the data

- The data come from the *Forced Expiratory Volume (FEV)* study that took place in the late 1970's.
 - The observations come from a sample of 654 youths, aged 3 to 19, in/around East Boston.
 - Researchers were interested in answering the *research question*:
What is the effect of childhood smoking on lung health?

Cleaning your data

- Now that we've got the data loaded, we need to clean it to get it ready for use (*Look at lab 1F for help*). Specifically:
 - We want to name the variables: "age", "lung_cap", "height", "gender", "smoker", in that order.
 - Change the type of variable for gender and smoker from *numeric* to *character*.
- After changing the variable types for gender and smoker:
 - For gender, use `recode` to change "1" to "Male" and "0" to "Female".
 - For smoker, use `recode` to change "1" to "Yes" and "0" to "No".

Analyzing our data

- Our lungs data is from an observational study.
- **Write down a reason the researchers couldn't use an experiment to test the effects of smoking on children's lungs.**
- Observational studies are often helpful for analyzing how variables are related:
- **Do you think that a person's age affects their lung capacity? Make a sketch of what you think a scatterplot of the two variables would look like and explain.**
- Use the lungs data to create an `xypplot` of age and lung_cap.
 - **Interpret the plot and describe why the relationship between the two variables makes sense.**

Smoking and lung capacity

- Make a plot that can be used to answer the statistical question:
Do people who smoke tend to have lower lung capacity than those who do not smoke?
- **Use your plot to answer the question.**
 - **Were you surprised by the answer? Why?**
 - **Can you suggest a possible confounding factor that might be affecting the result?**

Let's compare

- Create three subsets of the data:
 - One that includes *only* 13 year olds ...
 - One that includes *only* 15 year olds ...
 - and one that includes *only* 17 year olds.
- Make a plot that compares the lung capacity of smokers and non-smokers for each subset.
- **How does the relationship between smoking and lung capacity change as we increase the age from 13 to 15 to 17?**

Sum it up!

- **Does smoking affect lung capacity? If so, how?**
 - Support your answers with appropriate plots.
 - Explain why you included the variables you used in your plots.

Are You Asking Me?

Instructional Days: 9

Enduring Understandings

A survey is a data collection method that is administered to a sample. The sample is fraction of the target population. The sample must be representative of the population and random sampling is used to ensure an equal chance of being selected. A census is a special survey that collects data from the entire population. Sampling error and bias cause problems in analysis made from survey data.

Engagement

Students will view a video clip from the game show *Family Feud* to begin to think about survey components. The video can be found at:

<https://www.youtube.com/watch?v=-3Nk9t7-rCs>

Learning Objectives

Statistical/Mathematical:

S-IC 1: Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that bias and sampling error should be minimized when conducting surveys. The wording of questions, as well as who is asked to participate in a survey, can lead to bias. Learn that sampling error can be minimized when larger random samples are collected from a population.

Applied Computational Thinking Using RStudio:

- Create random samples of different sizes to make estimates about a population.
- Create informal confidence intervals based on sample medians.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will write questions that emphasize differences in data science concepts and skills.

Data File or Data Collection Method

Data File:

1. American Time-Use Survey (ATUS) data

Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

Lesson 9: Survey Says...

Objective:

Students will learn that a survey is another data collection method. They will learn what a survey is, what types of questions are used in a survey, and how a survey is conducted.

Materials:

1. Video: *Family Feud*'s "Shocking Fast Money" found at:
<https://www.youtube.com/watch?v=-3Nk9t7-rCs> (good quality, but sad ending)
OR
Video: *Family Feud* video clip titled "Family Feud – Comeback of the Century" found at:
<https://www.youtube.com/watch?v=ofQkOfeg60g> (bad quality, but happy ending)
2. *Designing a Survey* handout (LMR_3.4_Designing a Survey)

Vocabulary:

survey, self-reported, open-ended questions, closed-ended questions

Essential Concepts: Surveys ask simple, straightforward questions in order to collect data that can be used to answer statistical questions. Writing such questions can be hard (but fun)!

Lesson:



1. Introduce one of the videos listed above by informing students that they will be watching a clip from the television game show *Family Feud*. This segment of the show is called *Fast Money*, where the winning family plays for an additional \$20,000. Two family members are chosen to play and must reach a combined score of 200 points to win the money. The goal is to guess the most common responses to five questions. For example, if the question "What animal is a common pet?" were asked, each family member might answer with "dog" or "cat" since these are popular household pets. The first person accumulates as many points as possible during the 20-second first round. The second person is given 25 seconds to earn points with different answers.
2. As students watch the video, have them answer the following questions in their DS journals:
 - a. How many people were surveyed? **100**
 - b. Who was represented in the survey? **Single men**
 - c. How many survey questions were asked? **5**
 - d. When the host says "survey said" and we see the response, what does it mean? **It means that X number of people out of the 100 gave that response to the survey question.**
3. *Family Feud* uses surveys as its main data collection tool. In their DS journals, students should write down what they know about surveys individually.
4. Then, with a partner, students will share what each one of them knows about surveys.
5. Select a couple of students to either share their response or their partner's response with the whole class.
6. Inform students that a **survey** is a data collection method where the data are **self-reported**, meaning that participants answer questions themselves. Surveys are composed of:
 - a. Questionnaires or a series of questions
 - b. A representative sample of the population of interest
 - c. Carefully worded questions
7. Surveys rely on questions. There are two types of questions that can be asked in a survey: **open-ended questions** and **closed-ended questions**. Open-ended questions offer a free-response/text approach, whereas closed-ended questions give a fixed set of choices.



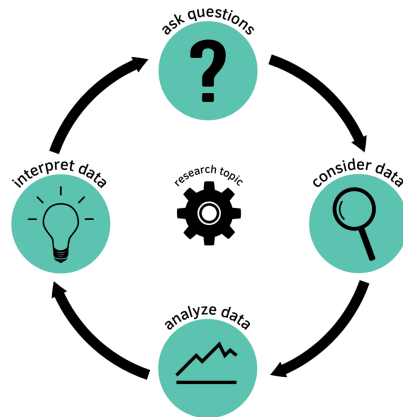


8. Display the following list to the class. With a partner, have the students categorize the following types of questions as either open-ended or closed-ended:

- (a.) Multiple choice (*closed*)
- (b.) Write a paragraph (*open*)
- (c.) Yes/No (*closed*)
- (d.) Comments (*open*)
- (e.) Essays (*open*)
- (f.) On a scale from 1 to 5 (*closed*)
- (g.) Choose from a list (*closed*)
- (h.) Write a sentence (*open*)
- (i.) Check a box (*closed*)

9. Do a quick *Whip Around* to share the categorization for each type of question. Be sure that students make corrections to the list if any items were miscategorized.

10. Quickly review the Data Cycle.



11. To give students an introduction to conducting surveys, they will first go through a practice scenario as a class to try to answer the following research question:

What are ‘families’ in the United States?

12. Distribute the *Designing a Survey* handout (LMR_3.4) and let students fill in the boxes for “Research Topic” (*Families*) and “Research Question” (*What are ‘families’ in the United States?*).

Name: _____ Date: _____

Designing a Survey

Research Topic

↓

Research Question

Statistical Question #1

Statistical Question #2

Survey Question #1

Survey Question #4

Survey Question #2

Survey Question #5

Survey Question #3

Survey Question #6



13. Inform students that the left side of the handout will be completed as a class, and then student teams will work together to complete the right side.

14. Using the Data Cycle as a guide, students should brainstorm a statistical question that is related to the research question. One might be: *What is the typical family size in the United States?*

Note: This requires a definition of “family,” which can have a variety of meanings to different people. Different definitions will likely guide the discussion of possible survey questions in the following step.

15. Next, students need to determine 3 survey questions to help answer the statistical question. The goal in creating survey questions is to make sure they (1) are unambiguous, and (2) address the statistical question. Some examples are listed below (which come from different definitions of “family”):

Note: Survey questions MUST match the statistical question.

- (a.) How many siblings do you have?
- (b.) How many people live with you?

16. It may help to actually collect data once the first survey question has been created. For example: “How many siblings do you have?” – each student would give a response and the values could be recorded in a dotplot (if desired). If the question is too vague (do we include half-siblings, step-siblings, etc.?), students can revise the question.

17. Once the class has agreed upon 3 survey questions for the first statistical question, allow students to join their student teams for the remainder of the activity.

18. Each team should come up with a statistical question that might answer the research question, then determine 3 survey questions that match their statistical question. Have the students create both open- and closed-ended questions in the handout. Each survey question should be a different type (see Step 8).

19. Have student teams share out their statistical questions and related survey questions with the rest of the class.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

For homework, students should choose one of their team’s survey questions and rewrite it 3 ways, using 3 different question types (see Step 8). Example rewrites for the statistical question “How many siblings do you have?” are given below for reference.

(a.) *Multiple choice:*

How many siblings do you have? Select one option.

- (a) 0 siblings
- (b) 1 sibling
- (c) 2 siblings
- (d) 3 siblings
- (e) more than 3 siblings

(b.) *Write a paragraph:*

In your own words, describe your siblings.

(c.) *Yes/No:*

Do you have any siblings?

Lesson 10: We're So Random

Objective:

Students will learn how to collect random samples from a population in order to estimate a parameter.

Materials:

1. *Populations & Samples* handout (LMR_3.5_Populations and Samples)
2. RStudio
3. Projector for RStudio functions
4. Dotplot titled "Percent of Students Who Have Met Friends Online"
5. *Parameters & Statistics* handout (LMR_3.6_Parameters and Statistics)

Vocabulary:

population, sample, representative, random sample, parameter, statistic

Essential Concepts: Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson:

1. Introduce today's lesson by displaying the following statement made by the Pew Research Center in their August 2015 report titled *Teens, Technology & Friendships*:

"For today's teens, friendships can start digitally: 57% of teens have met new friends online. The margin of error is plus or minus 3.7 percentage points. Social media and online gameplay are the most common digital venues for meeting friends."

Note: The data for this report were collected via interviews of 1,060 teenagers between the ages of 13 and 17.

2. Discuss the results of the Pew poll with the following prompting questions:

- a. The report says that 57% of teens have met new friends online. Since the report was based on a sample of 1,060 teens, how many of the teens reported that they have met friends online? $0.57(1060) = 604.2$. *This means that approximately 604 teens in the sample have met friends online.*
- b. Do you think 57% of students in our class have met friends online? Why or why not? *Answers will vary by class. The discussion should include points about how similar and different samples can be. The sample of students in the Pew poll may not represent the students in our class.*
- c. What percent of students in our class have met friends online while a teenager? *Answers will vary by class. Calculate the percentage by dividing the number of students who have met friends online by the total number of students who came to class today.*
- d. What if [absent person] were in class today? Would that change our percentage? What effect would it have on the percentage if [absent person] answered "yes?" What effect would it have if [absent person] answered "no?" *Answers will vary by class. If a student who was absent for today's lesson had actually come to class, we would have a different sample of students. It would change the percentage because our sample size now includes 1 more person. If the person answered "yes," the values in the numerator and denominator of the percentage would change. If the person answered "no," the value in the denominator would change. (Students should calculate these values.)*
- e. If we were able to interview every single teenager in the United States, would exactly 57% of them say they have met friends online? *Probably not because there are many more teenagers in the US than the 1,060 they interviewed. It would be unlikely for a group of 1,060 teenagers to exactly represent all teenagers in the entire country.*

- f. Why do you think the Pew Research Center only interviewed 1,060 teenagers, and not all teenagers in the US? *It would be impossible to talk to all teenagers in the US in a short period of time, or even a fairly long period of time.*



3. Introduce students to the terms **population** and **sample**. Explain that a population consists of all of the people we want to learn something about. A sample consists of people (or objects) that are selected *from* the population. In pairs, ask students to discuss and record answers to the following two questions:

- a. What was the population of interest to the researchers for the Pew poll above? *All teenagers in the US right now.*
- b. Based on your answer in (a), what characteristics should people have in order to be included in a sample for this poll? *People would need to be in the US and be between the ages of 13 and 17. People could be from many states, but we would not want to sample only people from California, or only people from Los Angeles. It would be impossible to survey every single person in the US; this is why we create a random representative sample of the population.*

Note: Steer the discussion so that students see that a sample has to be “like” or “similar to” or “representative of” the population.

- 4. Select pairs to share their responses to the questions and let students revise their responses.
- 5. Distribute the *Populations & Samples* handout (LMR_3.5), which contains survey results from other Pew Research Center reports. Give students time to determine the population and sample for each scenario, and then have them verify their results with a partner.

Name: _____ Date: _____

Populations & Samples

Instructions:
For each Pew Research Center survey data listed below, identify the population of interest, the sample that was taken, and the sample size.

Example 1:

Social media users among all adults
Among all American adults ages 18+, the % who use the following social media sites

Facebook	68
LinkedIn	23
Pinterest	22
Instagram	21
Twitter	19

Source: Pew Research Center's Internet Project, September 2014. Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=2,000 adults in the U.S. ages 18+.

PEW RESEARCH CENTER

<http://www.pewinternet.org/2015/01/09/social-media-update-2014/>

Population: _____

Sample: _____

Sample size: _____

Example 2:

Why Get Married?
Percent saying each is a very important reason to marry, by marital status

Love	MARRIED: 93%	UNMARRIED: 84%
Making a lifelong commitment	87%	74%
Companionship	81%	63%
Having children	59%	44%
Financial stability	31%	30%

Adults of married and unmarried separately, n=1,306 for married and 1,885 for unmarried.

Pew Research Center

<http://www.pewsocialtrends.org/2013/02/13/love-and-marriage/>

Population: _____

Sample: _____

Sample size: _____

LMR_3.5

6. State that we want to know the percentage of students in our class that have made friends online, but we don't want to ask every single student. Instead, we would like to ask only 10 students and then make some guesses about our class from those 10. Ask:

- a. What is the population of interest? *The students in our classroom.*
- b. How can we select 10 students to be part of our sample? *Answers will vary by class. There may be a variety of suggestions; here are some examples of what may be given: (1) put every student's name in a hat and pick out 10; (2) select the 10 students sitting closest to the teacher's desk; (3) have 10 students volunteer to be in the sample.*



7. Inform students that, in general, we want samples to “look like” the population. One way to get a representative sample like this is to take a **random sample**. Ask the students:

- a. Would the selection techniques we came up with in Step 6 result in random samples of our class. *Answers will vary by class. Using the examples from Step 6: (1) putting each student's name in a hat and then picking out 10 would be a random sample as long as each piece of paper is the same size; (2) selecting the 10 students sitting closest to the teacher's desk would not be a random sample because those students might not represent the whole class; (3) if 10 students volunteer, we would not have a random sample because those students selected themselves to be part of the group and may not represent everyone in the class.*
8. Next, assign each student in the class a number by having them count off from 1 to N (N being the total number of students in the class). Show students that we can use RStudio to create random samples with the following function:

```
> sample(1:N, size = 10, replace = FALSE)
```

Note: we use `replace = FALSE` because we only want each student to be selected once.

9. Using the results given in the output of RStudio, ask the students whose numbers were chosen to stand. Inform them that they are “in” the sample. Then, determine what percent of the sample (these 10 students) have made friends online. How does this percentage compare to the overall class percentage we found in Step 2(c)? *Answers will vary by class.*
10. Create a dotplot on the board titled “Percent of Students Who Have Met Friends Online.” Record the sample percentage from Step 9 on this dotplot.
11. Have the students return to their seats so that we can select a new sample. Before we do this, ask:
 - a. What do you predict the percentage to be for the next sample of 10 students? *Answers will vary by class. They might say the expected percentage will be close to the class's overall percentage.*
12. Using RStudio, create a new sample, calculate the percentage of those 10 students who have met friends online, and record the value in the dotplot.
13. Repeat Step 11 for a few more rounds (at least 5 samples should be taken). Be sure that the students give a prediction before finding each new sample.
14. Display the following questions. Refer to the dotplot of sample percentages. Ask students to discuss the questions in teams:
 - a. What do you notice about the sample percentages? What is the “typical” value? *Answers will vary by class. The typical value should be close to the class's overall percent calculated in Step 2(c) since it is the population percent.*
 - b. What is the smallest value? What is the biggest value? *Answers will vary by class. There might be a lot of variability in the dotplot based on the selected samples. Most sample percentages will be within 30% of the population value, so that really gives a wide variety of possible sample values.*
 - c. If we took a larger sample – maybe of size 15 or 20 – would there be more or less variability in the dotplot? *There will be less variability because adding more people gets us closer to the population size. Be sure to point out that if the sample were exactly the same as the population, then there would be no variability in the plot.*
15. Select teams at random to share their responses to the questions above with the whole class. The rest of the teams should be in full agreement with the responses before moving on to the next question.
16. Explain that the population percentage (the percentage of all students in the class who met friends online) we have been using is called a **parameter**. A parameter is any number that summarizes a population. So, our class has been the population, and the percentage of students that have met friends online is the parameter.
17. Similarly, the term **statistics** is used for numbers that summarize a sample. Ask students what sample statistics they have seen today? *Each value we included in the dotplot is a statistic.*

18. Be sure to point out that there can be multiple values for a sample statistic (i.e. “We had 5 sample percentages in our dotplot.”), but there is always only one parameter value.
19. Using these new definitions, ask students to consider the original Pew poll data, which found that 57% of teens have met friends online. Ask the students:
 - a. Is 57% a parameter or statistic? *This is a statistic because it is based on a sample. Remind them that the population was ALL US teenagers and the sample included 1,060 teens.*
 - b. What is the population parameter then? *We don't know! We would have to interview every teenager in the US to determine the parameter, and that is not possible.*
20. Conclude the lesson by telling the students: although we cannot determine the actual population parameter for the percent of teens that have made friends online, we can estimate it using random samples.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



Students should complete the *Parameters & Statistics* handout (LMR_3.6) for homework.

Note: Page 2 of the handout is an answer key for teacher reference only!

Name: _____ Date: _____

Parameters & Statistics

Instructions:
For each study below, identify the population, sample, parameter of interest, and any statistics.

1. A poll is a type of survey that is used to make statistical inference, a conclusion about a population based on a sample. In 2013, Gallup, a polling agency, surveyed 2,048 adults to find out Americans' main source of news. 55% of adults responded that they get their news from television.

Population: _____	Parameter: _____
_____	_____
Sample: _____	Statistic: _____
_____	_____
2. In 2009, Time Magazine conducted an Internet poll of affluent adults (people whose income is \$150,000 per year or more). A total of 603 affluent adults over the age of 18 were interviewed. They found that 95% of affluent Americans made online purchases in the last year.

Population: _____	Parameter: _____
_____	_____
Sample: _____	Statistic: _____
_____	_____
3. In a 2013 article published by The Guardian, an English newspaper, a survey found that 62% of 16-24 year-olds prefer print books over digital books. In this survey, 1,420 young adults aged 16-24 were interviewed.

Population: _____	Parameter: _____
_____	_____
Sample: _____	Statistic: _____
_____	_____
4. The Centers for Disease Control (CDC) collected data from 20,015 Americans between 2007 and 2010. The CDC wanted to know wanted to know the typical height of women over age 20. 5,971 women age 20 and over were part of the study. They found that the average height in centimeters is 63.8.

Population: _____	Parameter: _____
_____	_____
Sample: _____	Statistic: _____
_____	_____

LMR_3.6

Lesson 11: The Gettysburg Address

Objective:

Students will learn the definition of sampling bias and will learn that random samples reduce bias when estimating a population parameter. They will gain practice collecting a random sample from a small population and estimating the population parameter.

Materials:

1. *Gettysburg Address* handout (LMR_3.7_Gettysburg_Address)
2. *Sampling the Gettysburg Address* handout (LMR_3.8_Sampling the Gettysburg Address)
3. Dotplot titled “Mean Word Length, Self-Selected Sample, Size = 10” – on board or poster paper
4. *Gettysburg Address – Word Length Histogram* file (LMR_3.9_Gettysburg Histogram)
5. *Gettysburg Word Lengths* handout (LMR_3.10_Gettysburg_Words)
6. RStudio
7. Projector for RStudio functions
8. Dotplot titled “Mean Word Length, Random Sample, Size = 10” – on board or poster paper

Note: This dotplot will be used again during Lesson 13, so the results need to be kept in some way (this can be either on poster paper or by simply taking a photo).

Vocabulary:

sampling bias

Essential Concepts: Statistics vary from sample to sample. If the typical value across many samples is equal to the population parameter, the statistic is 'unbiased.' Bias means that we tend to “miss the mark.” If we don't do random sampling, we can get biased estimates.

Lesson:

1. Introduce the lesson by describing the Gettysburg Address:
 - a. President Lincoln delivered the Gettysburg Address in November 1863.
 - b. It is one of the most famous speeches in the United States.
 - c. In it, Lincoln invoked the principles of human equality contained in the U.S. Constitution and Declaration of Independence.
2. Read the Gettysburg Address aloud to the class OR have students read it aloud. The text of the speech can be found in the *Gettysburg Address* handout (LMR_3.7).

Name: _____ Date: _____

The Gettysburg Address
By President Abraham Lincoln

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

LMR_3.7

3. Today we will use the Gettysburg Address to learn about different sampling techniques.
4. Inform students that the Gettysburg Address contains 272 words. We can consider these 272 words to be our population because it includes all words in the entire speech. From the population, we can sample 10 words that we think represent the speech. It is ok for this step to be vague – students can come up with their own concept for what they think “representative” means in this case.

- Distribute the *Sampling the Gettysburg Address* handout (LMR_3.8), which includes the actual speech, as well as 2 sampling activities. For this part of the lesson, we will only be looking at Sampling Activity 1 on page 1 of the handout.

Note: This activity was originally created by Allan Rossman and Beth Chance, and has been modified for the IDS curriculum.

Name: _____ Date: _____

Sampling the Gettysburg Address

The Gettysburg Address
By President Abraham Lincoln

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Sampling Activity 1

- Circle 10 words that you think might be representative of all words in the speech.
- Record your self-selected words and their corresponding word lengths in the table.

Word #	Word	Word Length	Word #	Word	Word Length
1			6		
2			7		
3			8		
4			9		
5			10		

- Summarize your word lengths data in a dotplot.

- Calculate the mean word length of your sample.

LMR_3.8

- Inform students that they will get 30 seconds to select 10 words that they think are representative of all words in the speech.

Note: It is important that students work fast so they are forced to choose based on first impressions and don't have time to reflect. Also, this activity tends to not work well if students are informed of the punch line (that random samples are unbiased) before they begin.

- At this point, explain to students that we are actually interested in answering a specific question:

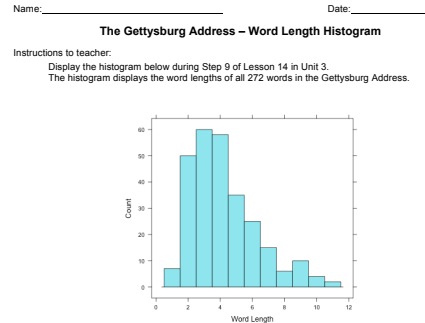
What is the typical word length in the Gettysburg Address?

- Next, students should record each circled word, as well as the number of letters each word has (this is the word length) in the table on the handout. Then, they should summarize the data in a dotplot and calculate the mean word length of the sample.
- On the board, create a class dotplot (may also be done on poster paper) titled "Mean Word Length, Self-Selected Sample, Size = 10." Once all students have completed the first page of the *Gettysburg Address* handout (LMR_3.8), ask each student to record the mean word length of his or her sample on the class's dotplot.
- When all students have recorded their sample statistics in the dotplot, conduct a class discussion based on the questions listed below.

Note: You might need to do a reality check. Students will often make mistakes when adding the word lengths and when dividing. Be suspicious (and double-check) extreme values.

- What does each point on the plot represent? *Each point represents one student's estimate of the mean length of all of words in the Gettysburg address.*
- What is the typical value represented in the dotplot? *Answers will vary by class. You should indicate the approximate location of the mean of the distribution (the balancing point, on the dotplot. Remind students that when we ask for the 'typical' value we mean the value in the center of the distribution.*

- c. How much variability is there in the distribution? *Answers will vary by class. One reasonable approach is for students to give the range (the difference between the largest and smallest values).*
 - d. What is the shape of the distribution? *Answers will vary by class. Often, the shape is right-skewed, but it might not be for you. Note that outliers here will often be arithmetic errors, but not always.*
11. Next, display the histogram from the *Gettysburg Address – Word Length Histogram* file (LMR_3.9), which shows the distribution of word lengths for the entire population of words in the Gettysburg Address.



LMR_3.9



12. Remind students that the population is the 272 words from the speech, and inform them that the mean word length of the population, or the population parameter, is 4.22. Using *Think-Pair-Share*, ask:
- a. How does the typical value of our class's sample means compare to the actual population mean of 4.22? *Almost always, the class's typical mean will be higher (sometimes much higher) than 4.22. Some students will be close to 4.22. But point out that we are talking about the "trend" or typical outcome. The typical outcome is usually too high.*

13. Explain that self-selected samples often produce biased results. **Sampling bias** is a description of the process, or the sampling plan, that is used to collect data. If the resulting samples tend to produce results that are influenced in one particular direction, we say that the sampling plan is biased.

Note: Bias is NOT the same as prejudice. Bias is a tendency to lean towards a certain belief or viewpoint, and is mostly unconscious. Prejudice is a very conscious phenomenon though, where a person actively makes a decision to dislike something based on unfounded facts.



14. Refer back to the dotplot of sample means and point out how it is biased. Ask:
- a. Why was our original sampling procedure biased? *When we look for 'representative' words, and do so quickly, our eyes are drawn by the larger, more unusual words, and we tend to overlook small words such as "in," "a," "we," etc.*
15. Go back to the *Gettysburg Address* handout (LMR_3.8), and direct students to page 2 for Sampling Activity 2. Inform students that they will now do a sampling procedure that results in a better representation of the population of words in the speech.
16. Explain that a random sample tends to produce unbiased sample results.
17. Before students begin the activity, demonstrate how to generate 10 random numbers from a possible 272 using RStudio.

```
> sample((1:272), size = 10, replace = FALSE)
```

18. Each student should generate his or her own set of 10 random numbers. Once students have created their random numbers, distribute the *Gettysburg Address Word Lengths* handout (LMR_3.10).

Name: _____ Date: _____

The Gettysburg Address - Word Lengths

Number	Word	Length	Number	Word	Length	Number	Word	Length
001	Four	4	046	nation	6	091	might	5
002	score	5	047	so	2	092	live,	4
003	and	3	048	conceived	9	093	It	2
004	seven	5	049	and	3	094	is	2
005	years	5	050	so	2	095	altogether	10
006	ago	3	051	dedicated,	9	096	fitting	7
007	our	3	052	can	3	097	and	3
008	fathers	7	053	long	4	098	proper	6
009	brought	7	054	endure.	6	099	that	4
010	forth	5	055	We	2	100	we	2
011	on	2	056	are	3	101	should	6
012	this	4	057	met	3	102	do	2
013	continent,	9	058	on	2	103	this.	4
014	a	1	059	a	1	104	But,	3
015	new	3	060	great	5	105	in	2
016	nation,	6	061	battle-	6	106	a	1
017	conceived	9	062	field	5	107	larger	6
018	in	2	063	of	2	108	sense,	5
019	liberty,	7	064	that	4	109	we	2
020	and	3	065	war.	3	110	can	3
021	dedicated	9	066	We	2	111	not	3
022	to	2	067	have	4	112	dedicate --	8
023	the	3	068	come	4	113	we	2
024	proposition	11	069	to	2	114	can	3
025	that	4	070	dedicate	8	115	not	3
026	all	3	071	a	1	116	consecrate --	10
027	men	3	072	portion	7	117	we	2
028	are	3	073	of	2	118	can	3
029	created	7	074	that	4	119	not	3
030	equal.	5	075	field,	5	120	hallow --	6
031	Now	3	076	as	2	121	this	4
032	we	2	077	a	1	122	ground.	6
033	are	3	078	final	5	123	The	3
034	engaged	7	079	resting	7	124	brave	5
035	in	2	080	place	5	125	men,	3
036	a	1	081	for	3	126	living	6
037	great	5	082	those	5	127	and	3
038	civil	5	083	who	3	128	dead,	4
039	war,	3	084	here	4	129	who	3
040	testing	7	085	gave	4	130	struggled	9
041	whether	7	086	their	5	131	here,	4
042	that	4	087	lives	5	132	have	4
043	nation,	6	088	that	4	133	consecrated	11
044	or	2	089	that	4	134	it,	2
045	any	3	090	nation	6	135	Far	3

LMR_3.10

19. Explain that the table contains the word number, word, and length of each word in the Gettysburg Address. Demonstrate how to find a word that corresponds to one of the random numbers generated by RStudio, and explain that this word is now part of our random sample.
20. Then, each student will complete the handout by creating a dotplot and calculating the mean of their random sample.
21. On the board, near the first dotplot, create another class dotplot (may also be done on poster paper) titled "Mean Word Length, Random Sample, Size = 10." Once all students have completed the second page of the *Gettysburg Address* handout (LMR_3.8), ask each student to record the mean word length of his or her random sample on the class's dotplot.

Note: As in Step 9, be sure to check arithmetic for outliers!

22. When all students have recorded their sample statistics in the dotplot, conduct a class discussion based on the following questions:



- What does each point in the dotplot represent? *Each dot represents one student's estimate of the mean word length. But this time, the estimates are based on a random sample of 10 words.*
- What do you notice about the typical value in this distribution? *Answers will vary by class. They should notice that the means of the random samples are fairly close to the population mean of 4.22. (Again, you might have to discard or correct outliers.)*
- What shape does this distribution have? What does that tell us? *Typically, the distribution of sample means for random samples will be symmetric and unimodal.*
- What does this distribution tell us about the benefits of random samples? *We can reduce bias by collecting random samples instead of self-selected samples.*
- Why do we need sampling? Why can't we just get the information from the actual population? *It is usually impossible to include every person or object from a population. Even for the population of size 272 words in the Gettysburg Address, it would take a long time to calculate the word lengths of every single word.*

23. Conclusions and takeaways:

- a. It turns out that there are approximately $5.17 \cdot 10^{17}$ different possible samples of 10 words from the Gettysburg Address.
- b. If we could determine the mean for each of these samples and produce a dotplot for all of these means, then the center of the dotplot would lie exactly at 4.22.
- c. The resulting distribution of the means from all possible samples is called the sampling distribution for the sample mean (for samples of size 10 from this population).
- d. The above dotplot is an approximation to the actual sampling distribution.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students should write a reflection about why random sampling is better at reducing bias than other sampling procedures.

Lab 3C: Random Sampling

Complete Lab 3C prior to Lesson 12.

Lab 3C - Random Sampling

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Learning by sampling

- In many circumstances, there's simply no feasible way to gather data about everyone in a *population*.
 - For example, the Department of Water & Power (DWP) wants to determine how much water people in Los Angeles use to take a shower. They've created a survey to pass out to collect this information.
 - **Write down two reasons why getting everyone in Los Angeles to fill out the survey would be difficult. Also, write a sentence why the DWP might consider using a sample of households instead.**
- In this lab, we'll learn how *sampling methods* affect how *representative* a sample is of a *population*.

Loading a population

- In previous labs, we used the cdc data as a sample for young people in the United States.
 - In this lab, we'll consider these survey respondents to be our population.
- Load the cdc data into R and fill in the blanks to take a *convenience* sample of the first 50 people in the data:

```
s1 <- slice(____, 1:____)
```

- **Why do you think we call this method a *convenience* sample?**

Comparing your convenience sample

- A convenience sample is a sample from a population where we collect data on subjects because they're easy-to-find.
- Using your convenience sample, create a bargraph for the number of people in each grade.
 - **Do you think the distribution of grade for your sample would look similar when compared to the whole cdc data?**
 - **Which groups of people do you think are over or under represented in your convenience sample? Why?**
- Create a bargraph for grade using the cdc data.
 - **Compare the distributions of the cdc data and your convenience sample and write down how they differ.**

Using randomness

- Fill in the blanks below to create a sample by randomly selecting 50 people in the cdc data, without replacement. Call this new sample s2:

```
__ <- sample(____, size = __, replace = __)
```

- **Write a sentence that explains why you think the distribution of grade for this *random sample* will look more or less similar to the distribution from the whole cdc data.**
 - Create a bargraph for grade based on this *random sample* to check your prediction.

Increasing sample size

- Create bargraphs for grade based on each of the following sample sizes: 10, 100, 1,000, 10,000.
 - Compare each distribution to that of the population.
- **How do the distributions change as the size of the sample increases? Why do you think this occurs?**
- tally() the proportion of grades for your *convenience* sample and all your *random* samples.
 - **Which set of proportions looks most similar to the proportions of the population?**

Lessons learned

- The mean, or proportion, from a *random* sample might not always be closer to that of the true population when compared to a *convenience* sample.
- However, as sample sizes get larger:
 - *Random* samples will tend to be better estimates for the population.
 - With *convenience* samples, this might not be the case.
- **Write down a reason why estimates based on *convenience* samples might not improve even as sample size increases.**

Lesson 12: Bias in Survey Sampling

Objective:

Students will learn about bias in relation to survey sampling. More specifically, they will learn what types of sampling methods could result in a biased sample, who might be over/under-represented in the sample, and how to select a better sample.

Materials:

1. *Identifying Biased Samples* handout (LMR_3.11_Identifying Biased Samples)
2. Poster paper

Vocabulary:

survey sample, over-represented, under-represented, random sampling

Essential Concepts: Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to ‘center’ on the population parameter value.

Lesson:

1. Remind students that they learned about biased samples during the last few lessons. Today, they will continue with this topic and discuss how people are selected to be in a sample.
2. The people who are asked to participate in a survey are known as the **survey sample**. Ideally, the people who are included in the survey sample are a representative group of the target population, or the population we would like to make inferences about.
3. Propose the following scenario to the class: “An elementary school is going to start serving ice cream in the cafeteria every Friday during lunch, and needs to know the favorite flavor of its students.”
4. In pairs, ask students to come up with two examples of samples that might be biased. For instance, one biased sample might include only the four 3rd grade classes at the school. For each biased sample, the students should answer the following questions in their DS journals:
 - a. Who is the target population? *All students at the elementary school.*
 - b. Who is included in your biased sample? *Only 3rd grade students. These students are **overrepresented** in the sample.*
 - c. Who is not included in your biased sample? *All other students in the school (Kindergartners, 1st, 2nd, 4th, and 5th graders). These students are **underrepresented** in the sample.*
 - d. Is your sample representative of the target population? *No! We’re only including 3rd graders, and they may not have the same preferences as other students.*
5. Once pairs have come up with their biased samples and answered the questions in Step 4, they should share out with their student teams and answer the following questions:
 - a. How is your biased sample different from the samples created by other pairs in your team? *Answers will vary by class. An example might be that one pair sampled only 3rd graders and the other pair sampled only girls.*
 - b. Which do you think is more representative of the target population? Why? *Answers will vary by class. Using the example above, we could argue that either one is more representative. We could maybe say that, since 3rd graders include both boys and girls, we have a more representative sample than if we just sampled girls. Or, we could say that since girls come from all grade levels, they’re more representative of the entire school than just 3rd graders.*

6. After the teams have discussed their samples, they should select one pair's biased sample to share with the rest of the class. Record each team's biased sample on a sheet of poster paper with the following layout:

Biased Sample	Who is overrepresented?	Who is underrepresented?
<i>All 3rd grade students</i>	<i>3rd grade students</i>	<i>All other students (kindergartners, 1st, 2nd, 4th, and 5th graders)</i>

7. Distribute the *Identifying Biased Samples* handout (LMR_3.11). In this activity, students will explain why a particular sampling method might result in a biased sample – a sample that is not representative of the population of interest.

Note: It is NOT enough for students to say that the “sample is not random.” They need to explain how the sample is biased.

Note: Page 2 of the handout provides sample answers for teacher reference ONLY. Do NOT distribute page 2 to students.

Name: _____ Date: _____

Identifying Biased Samples

Instructions:
For each example given below, explain why the resulting sample might be biased.

- A researcher sends out 500 questionnaires about pollution in Los Angeles to local residents by mail. She receives 340 responses.
Population of interest: _____
Why might the sample be biased? Explain. _____

- A researcher is interested in learning the typical number people per household who own cell phones. He conducts a survey by randomly calling phones that have land-lines.
Population of interest: _____
Why might the sample be biased? Explain. _____

- A researcher has concluded that dolphins are nice animals by surveying people who were assisted by one in a shark attack.
Population of interest: _____
Why might the sample be biased? Explain. _____

- A radio station host wants to know what proportion of her listeners enjoy the “Daily Dilemma” segment. She asks listeners to call into the station and respond.
Population of interest: _____
Why might the sample be biased? Explain. _____

- A researcher wants to know how many students at UCLA own pets. He stands outside the student health center and asks students before they enter the building.
Population of interest: _____
Why might the sample be biased? Explain. _____

LMR_3.11



8. Each student should complete the handout independently. Afterwards, conduct a whole-class discussion to compare and contrast different students' explanations of how the samples might be biased. For each example given in the handout, discuss who is most likely over-represented and who is most likely under-represented in the sample.
9. Ask the students:
- Now that we have learned about sampling biases, how can we eliminate this type of bias? *Answers will vary by class.*

Note: Allow students to collaborate and come up with a few ideas on their own of how to eliminate sampling bias. If desired, ideas can be written on the board for discussion and comparison.

10. Conclude with the actual answer: **random sampling.**

- a. If we randomly sample people from our population of interest, we can reduce the bias of any sample statistics obtained from the survey responses.
- b. If we have a biased sample, we can only give descriptions about that particular sample; we CANNOT generalize to the population of interest.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



Students will complete the *Survey Sampling* handout (LMR_3.12) for homework.

Name: _____ Date: _____

Survey Sampling

Instructions:

Read the survey sampling scenario below. Then, read the questions that follow. Re-read the scenario focused on the questions. Finally, write your response to each question.

Scenario

A researcher was asked to design an investigation based on the following research question:

Do American adult males spend more of their prime time hours online or watching regular television?
(Prime time is defined as weekday evenings from 7pm to 9pm.)

The researcher decided to conduct a survey sample. He determined that he would choose 1,000 American males over the age of 18 through a landline telephone survey. The researcher would randomly select the telephone numbers of adult males that live on the east coast of the United States.

Answer the following questions:

1. What is the target population and what is the sample?
2. Describe what 'bias' means. If a bias exists in the scenario you just read, explain why the sampling plan might be biased.
3. If a sampling plan is biased, what can the researcher do to reduce the bias?
4. A study was conducted at a particular high school about whether girls that attend that school prefer Red Vines[®] or Twizzlers[®]. Identify the following as either a parameter or a statistic by circling the correct term:
 - a. A random sample of 100 girls was selected. The results were that 40% preferred Red Vines[®]. The number "40%" is:
Parameter or Statistic
 - b. All of the girls at the high school were asked about their preference. The results were that 45% preferred Red Vines[®]. The number "45%" is
Parameter or Statistic

LMR_3.12

Lesson 13: The Confidence Game

Objective:

Students will learn about informal confidence intervals for making estimates about population parameters based on statistics from random samples.

Materials:

1. *The Confidence Game* handout (LMR_3.13_Confidence Game)
2. Dotplot titled “Number Correct” displayed on the board (or on poster paper)

Vocabulary:

inferences, interval, confidence interval

Essential Concepts: There is uncertainty when we estimate population parameters. Because of this, it is better to give a range of plausible values, rather than a single value.

Lesson:

1. Remind students that they have been learning about why sampling allows us to make **inferences** about a population. Some methods of sampling produce biased sample statistics, which does not allow us to generalize the results from a sample to the population of interest. To obtain unbiased statistics, random sampling methods need to be used.
2. Conduct a class brainstorm about what it means to “estimate” something. Have students come up with possible synonyms for the word “estimate.” *Some example synonyms include: guess, approximation, projection, opinion, impression, etc.*
3. Inform students that, in statistics, to provide an estimate means that we can give a range of values that we are confident include the population parameter value.
4. In today’s lesson, explain that the students will be playing a game, called *The Confidence Game*, in which they will be asked a series of questions that each have one numerical answer. However, instead of guessing what the exact answer is, the students will create a range of possible values that they think might include the real answer. They should be 90% confident that the true value is within their interval.
5. Introduce *The Confidence Game* to students by first going through an example using the question:

How tall is the Empire State Building, in feet (including the spire at the very top)?

- a. Ask the students to write down an **interval**, or range, of values that they think contains the true height of the building.
- b. Have a few students share their intervals with the class and discuss any obvious similarities or differences between them.

For example: If Student A gives an interval from 500 to 2000 feet and Student B gives an interval from 1100 to 1400 feet, one discussion could stem from asking Student A why he or she isn’t as sure of the answer as Student B is (since Student B gave a narrower interval). Then see if Student A wants to change his or her interval.

- c. After the discussion, tell the students that the actual height of the Empire State Building is 1,454 feet tall. Take a poll to see how many students’ intervals contained this value. We will learn what it means to have the true value in our intervals after we play the game.
6. Now, we can actually play the game! Distribute *The Confidence Game* handout (LMR_3.13) and explain the rules. Students will have about 5 minutes to complete the handout, which gives them approximately 30 seconds per question.

Note: The rules are printed at the beginning of the handout. They are included here for your convenience.

- a. Each question must be answered WITH AN INTERVAL.
- b. You should choose your interval so that you are “90% confident” (whatever that means to you).
- c. You CANNOT use any reference tools (i.e. no cell phones or computers to find answers).
- d. A question is “correct” if the true answer is inside your interval.
- e. The winner is determined by who got the most questions correct. In the case of a tie, the winner is chosen by whose intervals were narrower.

Name: _____ Date: _____

The Confidence Game

Rules of the game:

- a. Each question must be answered WITH AN INTERVAL.
- b. You should choose your interval so that you are “90% confident” (whatever that means to you).
- c. You CANNOT use any reference tools (i.e. no cell phones or computers to find answers).
- d. A question is “correct” if the true answer is inside your interval.
- e. The winner is determined by who got the most questions correct. In the case of a tie, the winner is chosen by whose intervals were narrower.

- 1) In what year did Mickey Mouse make his film debut?
- 2) What is the lowest temperature (in degrees Fahrenheit) ever recorded in California?
- 3) During the year 2014, how many television series were aired?
- 4) How far away, in miles, is the Earth from the Moon?
- 5) What is the greatest number of children officially recorded that were all born to one mother?
- 6) In what year did Orville and Wilbur Wright, more commonly known as the Wright brothers, make the first-ever powered flight in their self-built plane?
- 7) As of June 2015, how many songs by music artist Rihanna have reached the Number 1 spot on Billboard’s “Dance Club Hits” chart?
- 8) How many years have actors Will Smith and Jada Pinkett Smith been married?
- 9) How many hours will it take to complete a cross-country road trip from Los Angeles to New York City, according to Google Maps?
- 10) How many baseball fans can attend game at Dodger Stadium during any given day?

LMR_3.13

7. Once each student has completed *The Confidence Game* handout (LMR_3.13), have students choose partners and exchange handouts so that they can grade each other. Remind them that a question is marked as “correct” if the actual value (see answers in Step 8) falls within the interval.
8. Display the answers for each of the 10 questions from the handout found below:
 - 1) In what year did Mickey Mouse make his film debut? **1928**
 - 2) What is the lowest temperature (in degrees Fahrenheit) ever recorded in California? **-45 degrees Fahrenheit**
 - 3) During the year 2014, how many television series were aired? **1,715 TV shows**
 - 4) How far away, in miles, is Earth from the moon? **238,900 miles**
 - 5) What is the greatest number of children officially recorded that were all born to one mother? **69 children**
 - 6) In what year did Orville and Wilbur Wright, more commonly known as the Wright brothers, make the first-ever powered flight in a plane? **1903**
 - 7) As of June 2015, how many of Rihanna’s songs have reached the Number 1 spot on *Billboard’s* “Dance Club Hits” chart? **23 songs**
 - 8) How many years have actors Will Smith and Jada Pinkett-Smith been married? **18 years**
 - 9) How many hours will it take to complete a cross-country road trip from Los Angeles to New York City according to Google Maps? **41 hours (2,789.5 miles)**
 - 10) How many baseball fans can attend game at Dodger Stadium during any given day? **56,000 fans**

9. Each student should write the total number of “correct” responses at the top of his or her partner’s handout, and then return it.



10. Engage the students in a discussion about how well they did at estimating the true values with their intervals. The following questions can be used to steer the discussion:

- a. Remember that we were aiming to be 90% confident for each question. Based on this, how many of the 10 questions should we each have gotten correct? *If we are 90% confident, then we would expect 90% of the 10 intervals to include the true value, which is 9 intervals.*
- b. Did anyone in the class get exactly 9 correct? Did anyone get all 10 correct? *Answers will vary by class. However, it is very unlikely that many students will have gotten 9 or 10 correct responses on this first round.*



11. Create a dotplot on the board (or on poster paper) titled “Number Correct” and have each student record his or her value. Then, ask:

- a. How many students got 9 correct? In other words, how many students were actually 90% confident of their intervals? *Answers will vary by class.*
- b. What is the typical number of correct responses for our class? Does it seem too high or too low? Explain. *Answers will vary by class. Most likely, the typical number of correct responses will be fairly low (maybe even 4 or less).*
- c. Why is our typical score so much lower than 9? *We tend to be more confident than we should be, so we create narrower intervals.*
- d. It looks like, even though we thought we were 90% confident, most of us (or all of us) did not succeed 90% of the time. How could we increase our level of confidence? *We could use wider intervals.*

12. Recall from Step 3 that, in statistics, to estimate something means that we can give a range of values that we are confident include the population parameter value. This range of values, like the ones the students created during *The Confidence Game* activity, is known as a **confidence interval**.

13. Students will continue to learn about confidence intervals during the next lesson.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



In your own words, write a description of what a confidence interval is and why it is used in statistics.

Lesson 14: How Confident Are You?

Objective:

Students will learn about informal confidence intervals and estimates for the margin of error.

Materials:

1. Dotplot titled “Mean Word Length, Random Sample, Size = 10” – from Lesson 11

Vocabulary:

margin of error, bootstrapping

Essential Concepts: The margin of error expresses our uncertainty in an estimate. The estimate, plus or minus the margin of error, gives us an interval in which we are very confident the true value lies.

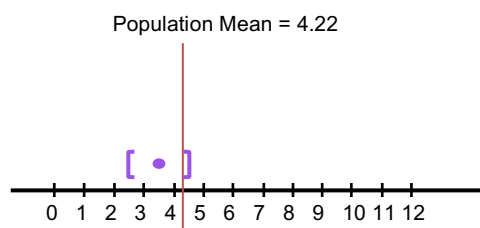
Lesson:

1. In this lesson, students will learn about confidence intervals in more detail.
2. Display the dotplot the class created during Lesson 11 (The Gettysburg Address) titled “Mean Word Length, Random Sample, Size = 10.”
3. Have students recall that each dot represents one student’s calculation of the mean word length of a sample of 10 randomly selected words from the Gettysburg Address.
4. Also remind them that the population parameter, which is the mean word length of all words in the speech, was 4.22. There should already be a vertical line on the dotplot to indicate this value, but if it is not present, please add it during this step. Ask:
 - a. What vocabulary word was used to describe each of the sample means we each created during Lesson 11? *The sample statistic. Every dot on the graph represents one sample statistic, more specifically each dot corresponds to a different sample mean.*
 - b. How many of us got exactly the right value? *Probably none.*
 - c. Thinking back on The Confidence Game we played yesterday, what approach could we do so that 90% of us would be correct? *We could give an interval.*
5. Show students that they can give an interval in the form:

Your sample statistic plus or minus AMOUNT

6. Ask them to calculate what their AMOUNT must be so that their interval includes the parameter value. Ask them to write this as an interval.
7. Choose one student to illustrate what is to be done. Ask them for their AMOUNT. On the dotplot, find their value, and use bars to go out plus and minus the AMOUNT. Confirm that it includes the parameter value.

For example: The purple dot represents a sample mean of 3.5. The AMOUNT we have chosen for this particular case is 0.8, so the lower bracket is 0.8 below the sample mean, and the upper bracket is 0.8 above the sample mean. Notice that the population parameter is included within the brackets.



8. Now, convert this to an interval of the form [lowest value, highest value] by subtracting the amount from the sample statistic to get the lowest value, and adding to get the highest.
9. Inform students that this AMOUNT is called the **margin of error**. Explain that the students all now have different margins of error because in this unusual 'game' they know the population value. But in real life we do not, and so we have to choose one single margin of error that will work 90% of the time.
10. Ask the students what margin of error they should use so that 90% of the estimates will have a 'successful' interval. You might want to tell them how many estimates that is for your class. A ballpark figure for the margin of error is 1.3.
11. Explain: If we were to start all over, we could imagine picking one of these sample statistics at random.
 - a. What's the probability that the sample statistic plus or minus the margin of error would include the parameter value? **90%**.
12. Because of this, we call these 'confidence intervals.' When we report an interval, for example 2.7 to 4.3, we say "We are 90% confident that the population parameter value is between 2.7 and 4.3." This is another way of saying "We don't know what the exact true value is, but we're confident it is somewhere in this interval."
13. Remind students of the Pew Poll they discussed during Lesson 10. For reference, the Pew Research Center made the following statement in their August 2015 report titled *Teens, Technology & Friendships: Pew Poll*

"For today's teens, friendships can start digitally: 57% of teens have met new friends online. The margin of error is plus or minus 3.7 percentage points. Social media and online gameplay are the most common digital venues for meeting friends."

Note: The data for this report were collected via interviews of 1,060 teenagers between the ages of 13 and 17.
14. Now that students have learned about the margin of error, have them write an *Exit Slip* about what the margin of error means in context of the Pew Poll.
15. Conclusions and takeaways:
 - a. Estimates that are based on random samples vary.
 - b. We can measure this variation.
 - c. The margin of error can tell us how much estimates vary.
 - d. We can use the estimate from our random sample, along with the margin of error, to give us a range of plausible values for the population parameter. This is called a confidence interval.
16. If time allows, introduce students to the idea of **bootstrapping**, which is where we take random samples of really large samples. For example, if we were looking at Twitter data, it would be almost impossible to compile every single tweet that exists in the population. Instead, we might be able to access 500,000 tweets, which is a very large sample. From this sample, we could create smaller random samples of size 100 and make inferences about the overall population of tweets from these samples. This will be discussed further in Lab 3D.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3D: Are You Sure about That?

Complete Lab 3D prior to the Let's Build a Survey! Practicum.

Lab 3D - Are you sure about that?

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Confidence and intervals

- Throughout the year, we've seen that:
 - Means are used for describing the typical value in a sample or population, but we usually don't know what they are, because we can't see the entire population.
 - Means of samples can be used to *estimate* means of populations.
 - By including a margin of error with our estimate, we create an interval that increases our confidence that we've located the correct value of the population mean.
- Today, we'll learn how we can calculate *margins of error* by using a method called the *bootstrap*.
 - Which comes from the phrase, *Picking yourself up by your own bootstraps*.

In this lab

- Load the built-in `atus` (*American Time Use Survey*) data set, which is a survey of how a sample of Americans spent their day.
 - **The United States has an estimated population of 327,350,075. How many people were surveyed for this particular data set?**
- The statistical question we wish to investigate is:

What is the mean age of people older than 15 living in the United States?
- **Why is it important that the ATUS is a random sample?**
- **Use our `atus` data to calculate an estimate for the average age of people older than 15 living in the U.S.**

One bootstrap

- A *bootstrapped* sample is when we take a random `sample()` of our original data (`atus`) *WITH* replacement.
 - The size of the sample should be the same size as the original data.
- We can create a single *bootstrapped* sample for the mean in 3 steps:
 1. Sample the number of the rows to use in our *bootstrap*.
 2. `slice` those rows from our original data into our *bootstrap* data.
 3. Calculate the mean of our *bootstrapped* data.

Our first bootstrap

- Fill in the blanks to `sample` the row numbers we'll use in our *bootstrapped* sample.
 - Be sure to re-read what a *bootstrapped* sample is from the previous slide to help you fill in the blanks.
 - Use `set.seed(123)` before taking the sample.

```
bs_rows <- ____ (1:____, size = ____, replace = ____)
```

- We can use the `slice` function to create a new data set that includes each row from our sample

```
bs_atus <- slice(atus, bs_rows)
```

Take a look

- Look at the values of `bs_rows` and `bs_atus`.
 - **Write a paragraph that explains to someone that's not familiar with R how you created `bs_rows` and `bs_atus`. Be sure to include an explanation of what the *values* of `bs_rows` mean and how those values are used to create `bs_atus`. Also, be sure to explain what each argument of each function does.**

One strap, two strap

- Calculate the mean of the age variable in your bootstrapped data, then use a different value of `set.seed()` to create your own, personal *bootstrapped* sample. Then calculate its mean.
 - Compare this second *bootstrapped* sample with three other classmates and write a sentence about how similar or different the *bootstrapped* sample means were.

Many bootstraps

- To use *bootstrapped* samples to create *confidence intervals*, we need to create many *bootstrapped* samples.
 - Normally, the more *bootstrapped* samples we use, the better the *confidence interval*.
 - In this lab, we'll `do()` 500 *bootstrapped* samples.
- To make `do()`-ing 500 *bootstraps* easier, we'll code our 3-step bootstrap method into a function.
 - Open a new R script (File -> New File -> R Script) to write your function into.

Bootstrap function

- Fill in the blank space below with the 3-steps needed to create a *bootstrapped* sample mean for our `atus` data.
 - Each step should be written on its own line between the curly braces.

```
bs_func <- function() {  
  
  }  
}
```

- Highlight and *Run* the code you write.

Visualizing our bootstraps

- Once your function is created, fill in the blanks to create 500 *bootstrapped* sample means:

```
bs_means <- do(____) * bs_func()
```

- **Create a histogram for your bootstrapped samples and describe the *center*, *shape* and *spread* of its distribution.**
 - These bootstrapped estimates no longer estimate the average age of people in the U.S.
 - Instead, they estimate how much the estimate of the average age of people in the U.S. varies.
- In the next slide, we'll look at how we can use these bootstrapped means to create *90% confidence intervals*.

Bootstrapped confidence intervals

- To create a 90% confidence interval, we need to decide between which two *ages* the middle 90% of our bootstrapped estimates are contained.
- **Using your histogram, fill in the statement below:**
The lowest 5% of our estimates are below _____ years and the highest 5% of our estimates are above _____ years.
- Use the `quantile()` function to check your estimates.
- **Based on your bootstrapped estimates, between which two ages are we 90% confident the actual mean age of people living in the U.S. is contained?**

On your own

- Using your *bootstrapped* sample means, create a 95% confidence interval for the mean age of people living in the U.S.
 - **Why is the 95% confidence interval wider than the 90% interval?**
 - **Write down how you would explain what a 95% confidence interval means to someone not taking *Introduction to Data Science*.**

Practicum: Let's Build a Survey!

Objective: Students will design a non-biased survey.

Materials:

1. Practicum: *Let's Build a Survey!* (LMR_U3_Practicum_Build a Survey)

**Practicum
Let's Build a Survey!**

Based on what you have learned in Lessons 9 through 14, you will now design a survey. You and your team members must do all of the following:

1. Select a topic from the list below:
 - a. Social Media
 - b. Entertainment
 - c. Sports
 - d. The Environment
 - e. Health
 - f. Education
 - g. Other topic of interest
2. Create a research question about your topic of interest.
3. Create a statistical question that is related to the research question.
4. Identify the population of interest.
5. Describe how you will select your sample from the population so that you'll be able to make generalizations about your population of interest.
6. Identify the number of people who will be in your sample.
7. Create five survey questions that will try to answer your statistical question and describe how you have made sure that they are non-leading questions.
8. Identify a statistic that can be used to summarize the responses from this survey. Can you identify a parameter?
9. Submit a typed paper that details the survey you just designed.

What's the Trigger?

Instructional Days: 5

Enduring Understandings

Sensors are data collection devices that collect data either continuously or whenever they are triggered. A sensor is a converter that measures a physical quantity and converts it into a signal, which can be read by an observer or by an instrument. Participatory Sensing is a specific data collection method that uses sensor technology. This method emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. Triggers play an important role in the Participatory Sensing data collection process. The response to the triggers may or may not be the same each time.

Engagement

Students will view and discuss a video clip called *Play Like Nadal With a Smart Tennis Racket* to begin to think about the sensors as data collection devices found ubiquitously in today's world. The video can be found at: <https://youtu.be/lcBnzddQECc>

Learning Objectives

Statistical/Mathematical:

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that sensors provide a continuous stream of data. Participatory Sensing provides real-time data from a user who is willingly providing the data. What differentiates a sensor as a data gathering method is the use of a trigger that signals a data collection session.

Applied Computational Thinking:

- Create a Participatory Sensing campaign using a campaign Authoring Tool.

Real-World Connections:

Sensors are found everywhere in today's world. They can provide data about environmental conditions as well as personal habits. More and more, sensors are being used for personal tracking, especially in the medical field, to inform people about what they do.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use write questions that use emphasize differences in data science concepts and skills.

Data File or Data Collection Method

Data Collection Method:

1. Students will gather data generated through a class-generated Participatory Sensing campaign.

Data File:

1. Students' Participatory Sensing campaign data

Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

Lesson 15: Ready, Sense, Go!

Objective:

Students will learn what sensors are and how they are used to collect data.

Materials:

1. Video: *Play Like Nadal With a Smart Tennis Racket*
<https://youtu.be/lcBnzddQECC>
2. Computers (see Step 5)
3. Poster paper
4. Flags in 3 different colors
Advanced preparation required (see Step 10 below)

Vocabulary:

sensor, trigger, algorithm

Essential Concepts: Sensors are another data collection method. Unlike what we have seen so far, sensors do not involve humans (much). They collect data according to an algorithm.

Lesson:

1. *Entrance Ticket:* What are some of the data collection methods we have learned about so far in this unit? *We have learned about experiments, observational studies, surveys, and getting data from a URL (in Lab 3B).*
2. Inform students that, in this lesson, they will be introduced to another data collection method known as sensors.
3. With a partner, ask students to discuss what they think a sensor is. Ask each pair to write down their ideas.
4. Show the *Play Like Nadal With a Smart Tennis Racket* video found at: <https://youtu.be/lcBnzddQECC>. As students watch the video, they should think about other sensors they may have come across, particularly ones used with smartphones. After watching the video, ask students to add to their definition of a sensor.
5. Now, inform students that they will work in teams to compile a list of data-collecting sensors. They may use computers to conduct online research for this part of the lesson. Challenge each team to generate the longest list in the class.
6. After students have had time to research and create their lists, ask students in each team to number off one through four (or five, depending on team sizes).
7. Share out in rounds. First, ask students in each team whose number is one to share one sensor from their list. On the poster paper, create a class list of sensors as shared by the students. Repeat with the rest of the numbers.
8. Score keeping: Each person gets five seconds to respond. You may hold up your hand with the palm facing the students and count down. The rules for teams are as follows:
 - a. add a sensor to the list, get 1 point
 - b. repeat an answer, lose 1 point
 - c. do not answer in five seconds, lose a turn
 - d. do not have an answer to contribute, may pass

Note: You may reward the winning team with extra credit points, if desired.

9. Next, students will engage in an activity to see sensors in action.
10. Create 3 groups of students:

- a. Group 1 – Triggers (3 students)
Provide each Trigger a different colored flag (for example: **Pink**, **Purple**, **Green**). The teacher will call out a color, at random, and the Trigger assigned to that color will raise his or her flag. Each flag corresponds to a research question of interest.

Pink – Who is in our class?

Purple – What is on our classroom walls?

Green – What do we like to do after school?

- b. Group 2 – Sensors (2 students)
Each Sensor should be assigned to one Trigger, or colored flag (**Pink** = Sensor A, **Purple** = Sensor B, **Green** = no sensor assigned to it). When the Sensors see their assigned Trigger, they send a signal to the Collector (see below) telling him or her to collect data from another student in the class. The Sensors are basically go-betweens for the Triggers and the Collector.

- c. Group 3 – Collector (1 student)
One student is the Collector of all the data. The Collector is in charge of asking survey questions related to the research question of the original Trigger. Survey questions are provided here:

Survey questions related to the **Pink** trigger:

- (1) How did you get to school today (bus, car, walking, etc.)?
- (2) What size shoe do you wear?
- (3) What is your favorite pizza topping?

Survey questions related to the **Purple** trigger:

- (1) What is your favorite wall decoration?
- (2) What type of poster is it (motivational, reference, class work, etc.)?
- (3) What color is most prevalent in the poster?

Survey questions related to the **Green** trigger:

- (1) Do you have a sports team practice or club meeting today?
- (2) Are you hanging out with friends today after school?
- (3) Will you be working today after school?

Each time a sensor is active, the Collector must ask a new student in the class the appropriate survey questions.

11. Explain the activity to your students. Then, call out a flag color at random. Repeat several times. Make sure you call out the flag that has no assignment at least once so that students see that no action took place. Reflect on the activity with the following discussion questions:

- a. What data were missed? Why? *Data about what our class likes to do after school. They were missed because there was no Sensor connected to the Green trigger, so the Collector never knew to collect this type of data.*
- b. Grocery stores keep track of customer data when purchases are made with a loyalty card. What is the trigger in this case? What data are being collected? *The trigger is checking out at a grocery store. There are lots of data that are collected, including: items bought, cost of items, number of items on sale, etc.*

12. After they engage in the sensor activity, ask students to revisit their definition of a sensor (see Step 3). Have them revise their definition based on the following concepts:

- a. A **sensor** is a converter that measures a physical quantity and converts it into a signal, which can be read by an observer or by an instrument.
- b. A sensor collects data continuously, or whenever a **trigger** is activated. A trigger is a something that responds to an event so that an action can occur.

- c. Sensors collect data according to an **algorithm**. An algorithm is a process or set of rules that are followed (just like the rules followed during the activity).
- d. Sensors may also collect data automatically, without anyone's knowledge or input. Examples include GPS location, time, and date.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



Now that students learned what sensors are, ask them what data they would they like to see collected on a sensor that they couldn't collect in an experiment or survey. They must explain why it is difficult to collect that data in an experiment or survey, and how a sensor would make it easier to collect that data.

Lesson 16: Does It Have a Trigger?

Objective:

Students will learn to identify and categorize survey questions versus sensor questions, and will practice writing sensor questions.

Materials:

1. Poster paper (one per student team)
2. Sticky notes
3. *Sensor or Survey?* handout (LMR_3.14_Sensor or Survey)

Vocabulary:

Participatory Sensing

Essential Concepts: A key feature that distinguishes the way sensors collect data from more traditional approaches is that sensors collect data when a 'trigger' event occurs. In Participatory Sensing, this event is something we humans agree upon beforehand. Every time that trigger happens, we collect data.

Lesson:

1. Refer back to the list of sensors the class created during the previous lesson. Distribute a piece of poster paper to each student team, and have them create the following table:

Sensor	How is it triggered?
1.	1.
2.	2.
3.	3.

2. Assign each student team 3 sensors from the class's list.
3. Then, each team should complete the table using their knowledge of triggers discussed during the previous lesson. Remind students that when a trigger occurs, a sensor reacts to it and sends a signal to a data collector.
4. Conduct a *Gallery Walk* of the posters. Each team will get to write one reaction or question about what they see on each poster.
5. After the Gallery Walk, ask each team to return to their posters. If the posters include questions, have teams take turns responding to the questions.
6. *Quickwrite:* In their DS Journals, ask student to respond to the following questions. They will have two minutes to write as much as they can:
 - a. When you learned about survey questions, what were the two categories of questions you learned about? *Answer: Open-ended and Closed-ended are the categories.*
 - b. What are some examples of these types of questions? *Open-ended: write a paragraph, comments, essays, write a sentence, single answer. Closed-ended: multiple or single choice, yes/no, scales (e.g. 1-5), choose from a list, check a box.*
7. In teams, ask students to share their responses using the *Give One/Get One* strategy. You may use a timer to keep track of time.
8. Remind students that one of the most important things they learned about sensors is that there is a trigger that reminds either a device or a person to answer a question or to collect data.
9. For this class, students have already had experience with using sensors as a data collection tool – all the **Participatory Sensing** campaigns.



10. Explain that survey questions are asked in Participatory Sensing campaigns. There is no difference in the type of questions that are asked when collecting data via surveys and when collecting data via PS campaigns.
11. When deciding whether to use a survey or a PS campaign for data collection, we have to look at the research question of interest. Some questions are better answered with survey data, while others with PS campaigns. Research questions that include variation across time or across locations are good candidates for PS. Some questions might be answered by both.

For example:

Consider the research question: How does my sense of safety and security change as I go about my daily routine? This question would best be answered via a PS campaign because students could collect data in real time about their sense of security. A possible trigger could be "whenever you change locations" or "once at the start of every hour" or perhaps whenever a random alarm goes off.

Consider the research question: What proportion of high school students are superstitious? This question could be done with a survey based on a random sample from the population of all high school students.



12. Distribute the *Sensor or Survey?* (LMR_3.14) handout. In teams, students will determine whether a sensor or survey is better for a given research scenario.

Name: _____ Date: _____

Sensor or Survey?

Instructions:
For each scenario in the table below, identify which data collection method is more appropriate – a sensor (Participatory Sensing campaign) or a survey. Include your reasoning in the appropriate column.

Research Scenario	Sensor or Survey	Why did you choose this method?
You want to know the percentage of students in the school district who complete all of their homework each night.		
You want to know how many times per day, and at what times, students in the district play sports.		
You want to know what foods your class eats most often.		
You want to know your class' favorite food.		
A department store wants to know popular trends.		
You want to know what time of day students in the district wake up on school mornings.		
You are interested in determining patterns in your heart rate before, during, and after exercise sessions.		

LMR_3.14

13. Once the teams have completed the handout, assign each team one research scenario from the *Sensor or Survey* activity.



14. Conduct a *Whip Around* and have each team share their responses with the class. Allow students time to revise any incorrect responses.
15. Summarize the lesson by highlighting that PS campaigns and surveys use similar questions. However, depending on the research topic of interest, the decision to use one or the other relies on whether or not a trigger is involved.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



Suppose we wish to know more about whether people behave superstitiously. Write two research scenarios, using the following questions as a guide:

- a. How would you collect data to address this using PS? Include the trigger event you would use, and the data you would like to collect when the trigger happens.
- b. How would you collect data to collect this using a survey based on a random sample of people in California?
- c. Describe the differences between these two approaches. What can you learn in one approach that you can't in the other?

Lesson 17: Creating Our Own Participatory Sensing Campaign

Objective: Students will be guided through the creation of a new Participatory Sensing campaign and survey on a topic of interest chosen by the class.

Materials:

1. *Food Habits Campaign Questions* handout (LMR_3.15_Food Habits Qs)
2. *Campaign Creation Brainstorm* handout (LMR_3.16_Campaign Creation)

Essential Concepts: Creating a Participatory Sensing Campaign requires that survey questions must be completed whenever they are “triggered”. Research questions provide an overall direction in Participatory Sensing Campaign.

Lesson:



1. Review homework questions by asking a couple of students to share their responses. The rest of students will engage in *Agree/Disagree* as the questions are shared.
2. Display the following definition of Participatory Sensing that some computer scientists have agreed to and ask students to read and record this definition in their DS journals:

At its heart, Participatory Sensing is data collection and interpretation. Participatory Sensing emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. It can range from private personal observations to the combination of data from hundreds, or even thousands, of individuals that reveals patterns across an entire city. Most important, Participatory Sensing begins and ends with people, both as individuals and members of communities. The type of information collected, how it is organized, and how it is ultimately used, may be determined in a traditional manner by a centrally organized body, or in a deliberative manner by the collection of participants themselves. The latter case, in particular, emphasizes the novelty of Participatory Sensing as an approach and underscores the importance of using widely available and familiar technology. [Source: "Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world."]

3. Activate prior knowledge: Based on this definition, ask students to recall the Participatory Sensing campaigns in which they have engaged thus far. **Answer: Food Habits, Time Use, Stress/Chill.**

Note: *Personality Color* and *Time Perception* were surveys, not Participatory Sensing campaigns because they were only completed once. Their data was not collected over time.

4. Inform students that they will be creating a new, whole class Participatory Sensing campaign, but before they do that, they will analyze the *Food Habits Campaign Questions* handout (LMR_3.15).

Name: _____ Date: _____

Food Habits Campaign Questions

Prompt	Variable	Data Type
What's the name of your snack?	name	text
Is your snack salty or sweet?	salty_sweet	categorical
About how many servings do you actually eat?	servings_eat	numerical
How many calories per serving?	calories	numerical
How many grams of total fat per serving?	total_fat	numerical
How many milligrams of sodium per serving?	sodium	numerical
How many grams of sugar per serving?	sugar	numerical
How healthy do you think this snack is?	healthy_level	categorical 5 - Very healthy 4 - healthy 3 - Unhealthy 2 - Unhealthy 1 - Super Unhealthy
In one word, describe why you are eating this snack.	why	text
How much does this snack cost?	cost	numerical
How many ingredients are in your snack?	ingredients	numerical
Take a picture?	snack_image	photo
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date
AUTOMATIC	user	user id

In teams, analyze the Food Habits Campaign questions by responding to and recording your team's answer to the following questions.

- a. How many questions does the campaign have and what do you notice about the questions?
- b. When do these questions need to be answered?
- c. Who collects the data for the campaign?



5. In teams, allow students two minutes to discuss the following as they analyze the Food Habits Campaign questions:
 - a. How many questions does the campaign have and what do they notice about the questions? *Answers will vary. Students may notice that they are survey type of questions and may identify the type of questions such as open-ended, single-choice, etc.*
 - b. When do these questions need to be answered? *Each time they eat a snack.*
 - c. Who collects the data for this campaign? *The participants collect their own data.*
6. Ask a few teams to share their insights about the discussion. In the share-out, guide students to see that the questions are in fact survey questions. Although survey questions are answered once, when we collect data every time a 'trigger' event occurs, then we are engaging in Participatory Sensing.
7. Ensure that team roles have defined duties to keep teams on task for the rest of this lesson. Creating this class campaign will follow a process in which consensus (or a majority rule) will be reached in each step of the campaign development within each team. Inform students that they will be creating a Participatory Sensing campaign on a topic of their interest using LMR_3.16.

Name: _____ Date: _____

Campaign Creation

Instructions:
In teams, work together to fill in the information in this handout. You will be deciding, as a class, what information will be used in your class campaign during each round.

Round 1: Topic
This is a hobby, area of interest, or place or process that you want to know more about.

Team Ideas of Topics:

Class Decided Topic:

Round 2: Research Question
This is the main question you want to answer about the topic and will be the focus of the Campaign.

NOTE: You should NOT be able to simply search the Internet to find the answer to this question; data collection is required.

Team Research Questions:

Class Decided Research Question:

LMR 3.16

8. **Round 1:** First, teams will discuss their hobbies, areas of interest, or places or processes they want to know more about. Prompt students to think about whether they want to learn about "where they live, work, or play." All students within the group must agree on a hobby or area of interest to be their topic of interest to create a campaign for. *An example of a hobby is practicing cello. An area of interest might be 'the environment.' A place of interest might be "our school" or "my church" or "Disneyland."*
9. Once teams have decided on a topic for their group, have teams share out their topic of interest. As a class, decide on one topic that will be used for creating a new Campaign.
10. **Round 2:** Now have teams consider what research questions you might ask about this topic of interest. *An example of a research question for practicing cello is "How can I improve my playing?" or "How can I practice more effectively?"*
11. Once teams have decided on a research question for their group, have teams share out their research question. As a class, decide on one research question that will be used for creating a new Campaign.
12. **Round 3:** Next, they will examine what kind of data needs to be collected in order to answer this research question. They should discuss possible triggers that will determine when data should be collected. Allow teams to engage in a discussion about when is the best time to trigger the data collection/completion of the survey. For example: every day at 8am; whenever they practice the cello; whenever they see an advertisement; etc. They should record it in their DS Journals. *An*

example of a trigger for practicing cello is whenever you play the cello. In this case, it could be any time of day or even multiple times of the day.

13. Once teams have decided on a trigger for their group, have teams share out their possible trigger. As a class, decide on one trigger that will be used for creating a new Campaign.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework



Using the class topic, research question, trigger event, and discussion of the data they plan to collect. Classify our class campaign under the appropriate category with your justification:

(A) Individual; (B) Groups of people; (C) Community

Note to teacher: To determine which category a campaign should be placed under, consider the question "Who or what will we learn about?" If the answer is "only one person", then place in the Individual category. The cello campaign is an example of this. If we might learn about lots of people, put it in the Groups of People category. The Food Habits, Stress/Chill, and Time Use campaigns fit into this category (they learn about the students in the class). A campaign that wanted to know where all of the churches in the neighborhood were located, or wanted to try to keep people from littering, or wasting water, these should go into the "community" category.

Lesson 18: Evaluating Our Own Participatory Sensing Campaign

Objective: Students will create statistical questions and evaluate their Participatory Sensing Campaign.

Materials:

1. *Campaign Creation* handout (LMR_3.16_Campaign Creation) from previous lesson
2. Class Campaign Information from Lesson 16

Essential Concepts: Statistical questions guide a Participatory Sensing Campaign so that we can learn about a community or ourselves. These Campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson:

1. Review homework by giving students about five minutes to share their classifications in their teams. They will decide as a team which classification is the most fitting.
2. Once the five minutes have passed, have a class discussion of classifications and their justifications. Explain to the class that the campaign must be carried out by the whole class so if it has been classified in the Individual category, it must be revised. Also discuss whether the campaign is feasible. (For example, is the trigger so rare that no one will collect data? Are the questions too intrusive?).
3. Inform students that one of the promises of PS is its potential for helping people bring about social and civic change. Ask teams to consider the following questions and report back:

- a. Does our campaign try to do this?
- b. Could it be changed or modified to do this?

Note: Feasible campaigns fall under the groups of people or community categories. If a campaign is in the individual category, it should be modified to fall under the other categories before moving to round 4.

4. Display the campaign information students generated (and selected as a class) the previous day or revised today: Topic, Research question, Trigger, and Type of Data needed.
5. Now they will continue the rounds using the Campaign Creation handout LMR 3.16 from the previous lesson.
6. **Round 4:** Now that the class has decided on a trigger and the type of data needed, they will create survey questions to ask when the trigger is set. The questions should consider all of the possible data they might collect at this trigger event. It's ok if the list is long; the goal is to be creative and think of lots of different ideas.

Examples of survey questions for practicing cello are:

“How long did you practice?”

“What did you play?”

“How would you rate your practice session: 1 to 5?”

“Any thoughts or comments about your practice?”

7. Once teams have created 4 survey questions for their group, have teams share out their survey questions. As a class, decide on no more than 10 survey questions that will be used for creating a new Campaign.
 - a. Then, evaluate each survey question. For each question they should consider:
 - i. What type of data will this question collect? (Numerical, discrete numerical, text, categories, photos, location).
 - ii. How does this question help address the research question?

- iii. Does the question need to be reworded? (Is it clear what is being asked for? Do they know how to answer it?)
 - b. If the survey questions need to be rewritten, assign teams to rewrite survey questions. Then, as a class, decide on the changes.
 - c. Once finalized, write the survey question that goes along with that data variable, being cognizant of question bias.
- 8. Round 5: In teams, now generate two to three statistical questions that they might answer with these data. Make sure your statistical questions are interesting and relevant to the class topic of interest. They may keep a record in their DS Journals. Remind students that they will also have data about the date, time, and place of data collection.

Examples of statistical questions that can be answered for practicing cello are:

“How frequently do I practice?”

“When I practice more frequently, do I rate my sessions higher?”

“Are higher-rated sessions associated with time of day?”

- 9. Once teams have generated their statistical questions, have them share out with the class. Confirm that the questions are statistical and that they can be answered with the data the students propose to collect. As a class, decide on no more than 3 statistical questions to guide your campaign.



- 10. Now that they have all the pieces of the campaign, evaluate whether it's a reasonable and ethically sound campaign. Engage the class in a whole group discussion on the following questions:
 - a. Are answers to your survey questions likely to vary when the trigger occurs? (If not, you'll get bored entering the same data again and again)
 - b. Can the entire class carry out the campaign?
 - c. Do triggers occur so rarely that you'll have very little data? Do they occur so often that you'll get frustrated entering too much data?
 - d. Ethics: Would sharing these data with strangers or friends be embarrassing or undermine someone's privacy?
 - e. Can you change your trigger or survey questions to improve your evaluation?
 - f. Will you be able to gather enough relevant data from your survey questions to be able to answer your statistical questions?
- 11. Students have collaboratively created their first Participatory Sensing campaign. Inform them that you will be demonstrating one tool used to create the campaigns that they see on their smart devices or the computer. Students should take notes in their DS journals, as they will be using the tool later.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 19: Implementing Our Own Participatory Sensing Campaign


Objective: Students will mock-implement, create their Participatory Sensing campaign, survey on their topic of interest, then begin data collection.

Materials:

1. *Campaign Creation* handout (LMR_3.16_Campaign Creation) from previous lesson
2. Campaign Authoring Tool (<https://portal.idsucla.org>)

Essential Concepts: Practicing data collection prior to implementation allows optimization of a Participatory Sensing Campaign.

Lesson:

- 
1. Display the class generated campaign information for the class to clearly see.
 2. In teams, have students mock-implement the campaign they have created. They can do this by asking each other the survey questions to make sure they make sense/ will generate relevant data to their research question and statistical questions. They can use the evaluative questions from Lesson 17 step #10.
 3. If there are suggestions for improvement, have teams propose them to the class and make final changes to the campaign.
 4. Inform students that you will now demonstrate the tool used to create the campaigns that is displayed on their mobile devices or computers.
 5. Login to the **IDS Home Page** found at <https://portal.idsucla.org>. Click on the **Campaigns tab** on the navigation bar at the top of the page. Then, follow the steps in the tool:
 - a. **Campaign Info Window:**
 - i. **Campaign Name:** Give your campaign a name. A name related to the topic is recommended.
 - ii. **Select your class/period.**
 - iii. **Description:** Provide a one-sentence description of your campaign.
 - iv. **Data Sharing:** Select Disabled in order to monitor for improper responses.
 - v. **Campaign Status:** Select Running.
 - vi. **Click the **+Add Survey** button.**
 - b. **Survey Window:**
 - i. **Title:** Give the survey a title (again, it may or may not be the same as the campaign name). Users see the title and the all the prompts that follow.
 - ii. **ID:** Give the survey a name (it may or may not be the same as the campaign name). Users do not see the survey ID.
 - iii. **Description:** Provide a short description of the survey for display.
 - iv. **Submit Text:** Provide a brief message to be displayed after survey submission.
 - v. **Anytime:** Select the checkbox if you want the survey to be available at anytime.
 - vi. **Click the **+Add Prompt** button and select the prompt type for your first survey question. Note:** You should only select from the following choices: Single choice, number, photo, and text. Multiple-choice does not mean select one choice; it means select many choices. It is not recommended that multiple-choice be used at this point.
 - c. **Prompt Information:**
 - i. **Click the new prompt bar.**
 - ii. **Prompt ID:** This will be your first variable. A short one-word name or short two-word name separated by an underscore is recommended.

- iii. **Prompt Label:** This is the variable name that will be displayed (it may be the same as the prompt ID without the underscore, if used).
 - iv. **Question Text:** Type the survey question about which you want to collect data.
 - v. **Additional Prompt Information:** Depending on the prompt type, you will be asked to enter additional information. For example, if your prompt is Text, you will be asked a minimum and a maximum value for the number of characters the participant can enter.
 - vi. **Skippable:** Select the checkbox if you would like the prompt to be skipped. It is recommended that photo prompts be skippable, since some users will submit their responses via a browser.
- d. **Repeat step c for the remaining survey questions by clicking the **+Add Prompt** button.**
 - e. **XML Code:** As you create the campaign, the code that creates it will be displayed. You may select the checkbox titled **Enable Syntax Highlighting** so that students can keep track of where the information you are adding is embedded in the code. Inform students that they will be learning about XML syntax in the next several lessons.
 - f. **Click the **Submit Campaign** button on the top, right hand side of the page once all prompts have been added.** This action will send the campaign to the server for users to see.
6. Once all prompts have been created, students may use their smart devices or login to the IDS Home Page to view the new campaign. Remember to **Refresh Campaigns**.
 7. Students should go through the entire participatory sensing survey to see how their questions are displayed. They do not have to upload the survey.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 5 days, students will collect data using their newly created Participatory Sensing campaign.

Webpages

Instructional Days: 6

Enduring Understandings

Data takes on a variety of forms online and requires a different style of representation.

Engagement

Students will view a video clip about a data farm, specifically, Google's Street View Data Center to begin thinking about data formats and accessing data online. The video can be found at:

<https://www.engadget.com/2012-10-17-google-inside-data-centers.html>

Learning Objectives

Statistical/Mathematical:

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

DS: Use different techniques to access data from the web and understand why different data representations are useful for different software platforms.

Applied Computational Thinking using RStudio:

- Read data from xml and html table and convert to R data frames
- Use latitude and longitude coordinates of mountain data and overlay it on a map

Real-World Connections:

Data from the web has been used to predict outbreaks of the flu and is a source of extremely rich data sets.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will engage in discussions regarding internet research as it applies to data science.

Data File or Data Collection Method

Data Collection Method:

1. Students will scrape data from online HTML and XML sources.

Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

Lesson 20: Online Data-ing

Objective:

Students will discover that data exists on the Internet in a variety of areas, formats, and for a variety of purposes.

Materials:

1. *Video: Explore a Google Data Center with Street View* found at: <https://www.engadget.com/2012-10-17-google-inside-data-centers.html>
2. *Wikipedia – Video Games* handout (LMR_3.17_Wikipedia – Video Games)
3. *Wikipedia – Video Games – CSV Format* handout (LMR_3.18_Video Games – CSV)
4. *Online Data-ing* handout (LMR_3.19_Online Data-ing)

Vocabulary:

data farm, tags, HTML

Essential Concepts: We stretch students' conception of data, to help them see that many web pages present information that can be turned into data.

Lesson:

1. By a show of hands, ask students if they have ever heard of the term **data farm**. If any of them have, ask him or her to share what they know about it.
2. Inform students that a data farm is a physical space where high capacity servers are placed to store large amounts of data.
3. Introduce the video titled *Explore a Google data center with Street View* found at <https://www.engadget.com/2012-10-17-google-inside-data-centers.html> by explaining that the data center they are about to see is one of these large data farms used to store vast amounts of data.
4. After students watch the video, have a class discussion using the following questions:
 - a. We have been talking about data for a few months now. How would you respond if someone asked you, “What are data?” *Answers will vary by class.*
 - b. What are some ways that we have stored data? *Data frames in R, Excel spreadsheets, .csv files.*
5. Explain that one of the main ways data are distributed is through the Internet. Storing and sharing data on the Internet requires a different format than what we have seen. For example, Wikipedia has a page dedicated to the top video games.
6. Distribute the *Wikipedia – Video Games* handout (LMR_3.17), and have students explain the information that the data table provides.



Name: _____ Date: _____

Wikipedia – Video Games

Background:
The Wikipedia website contains many informative webpages. One such page is dedicated to the top video games of all time, which can be found at https://en.wikipedia.org/wiki/List_of_video_games_considered_the_best.

This screenshot from the website shows the first five rows of the "List of Best Games" data table.

Game	Original release year	Genre	Number of lists	Platform of original release	Lists / References
The Legend of Zelda: Ocarina of Time	1998	Action-Adventure	42	Nintendo 64	CNET ^[1] CVG000 ^[2] Esp000 ^[3] Esp001 ^[4] Esp004 ^[5] Esp013 ^[6] Esp104 ^[7] ECG001 ^[8] ECG004 ^[9] Esp0400 ^[10] Esp0414 ^[11] Fam0400 ^[12] Da ^[13] GameRanking ^[14] GameRabbit ^[15] GPR204 ^[16] GPR205 ^[17] GPR206 ^[18] GPR207 ^[19] GPR208 ^[20] GPR209 ^[21] GPR210 ^[22] GPR211 ^[23] GPR212 ^[24] GPR213 ^[25] GPR214 ^[26] GPR215 ^[27] GPR216 ^[28] GPR217 ^[29] GPR218 ^[30] GPR219 ^[31] GPR220 ^[32] GPR221 ^[33] GPR222 ^[34] GPR223 ^[35] GPR224 ^[36] GPR225 ^[37] GPR226 ^[38] GPR227 ^[39] GPR228 ^[40] GPR229 ^[41] GPR230 ^[42] GPR231 ^[43] GPR232 ^[44] GPR233 ^[45] GPR234 ^[46] GPR235 ^[47] GPR236 ^[48] GPR237 ^[49] GPR238 ^[50] GPR239 ^[51] GPR240 ^[52] GPR241 ^[53] GPR242 ^[54] GPR243 ^[55] GPR244 ^[56] GPR245 ^[57] GPR246 ^[58] GPR247 ^[59] GPR248 ^[60] GPR249 ^[61] GPR250 ^[62] GPR251 ^[63] GPR252 ^[64] GPR253 ^[65] GPR254 ^[66] GPR255 ^[67] GPR256 ^[68] GPR257 ^[69] GPR258 ^[70] GPR259 ^[71] GPR260 ^[72] GPR261 ^[73] GPR262 ^[74] GPR263 ^[75] GPR264 ^[76] GPR265 ^[77] GPR266 ^[78] GPR267 ^[79] GPR268 ^[80] GPR269 ^[81] GPR270 ^[82] GPR271 ^[83] GPR272 ^[84] GPR273 ^[85] GPR274 ^[86] GPR275 ^[87] GPR276 ^[88] GPR277 ^[89] GPR278 ^[90] GPR279 ^[91] GPR280 ^[92] GPR281 ^[93] GPR282 ^[94] GPR283 ^[95] GPR284 ^[96] GPR285 ^[97] GPR286 ^[98] GPR287 ^[99] GPR288 ^[100] GPR289 ^[101] GPR290 ^[102] GPR291 ^[103] GPR292 ^[104] GPR293 ^[105] GPR294 ^[106] GPR295 ^[107] GPR296 ^[108] GPR297 ^[109] GPR298 ^[110] GPR299 ^[111] GPR300 ^[112] GPR301 ^[113] GPR302 ^[114] GPR303 ^[115] GPR304 ^[116] GPR305 ^[117] GPR306 ^[118] GPR307 ^[119] GPR308 ^[120] GPR309 ^[121] GPR310 ^[122] GPR311 ^[123] GPR312 ^[124] GPR313 ^[125] GPR314 ^[126] GPR315 ^[127] GPR316 ^[128] GPR317 ^[129] GPR318 ^[130] GPR319 ^[131] GPR320 ^[132] GPR321 ^[133] GPR322 ^[134] GPR323 ^[135] GPR324 ^[136] GPR325 ^[137] GPR326 ^[138] GPR327 ^[139] GPR328 ^[140] GPR329 ^[141] GPR330 ^[142] GPR331 ^[143] GPR332 ^[144] GPR333 ^[145] GPR334 ^[146] GPR335 ^[147] GPR336 ^[148] GPR337 ^[149] GPR338 ^[150] GPR339 ^[151] GPR340 ^[152] GPR341 ^[153] GPR342 ^[154] GPR343 ^[155] GPR344 ^[156] GPR345 ^[157] GPR346 ^[158] GPR347 ^[159] GPR348 ^[160] GPR349 ^[161] GPR350 ^[162] GPR351 ^[163] GPR352 ^[164] GPR353 ^[165] GPR354 ^[166] GPR355 ^[167] GPR356 ^[168] GPR357 ^[169] GPR358 ^[170] GPR359 ^[171] GPR360 ^[172] GPR361 ^[173] GPR362 ^[174] GPR363 ^[175] GPR364 ^[176] GPR365 ^[177] GPR366 ^[178] GPR367 ^[179] GPR368 ^[180] GPR369 ^[181] GPR370 ^[182] GPR371 ^[183] GPR372 ^[184] GPR373 ^[185] GPR374 ^[186] GPR375 ^[187] GPR376 ^[188] GPR377 ^[189] GPR378 ^[190] GPR379 ^[191] GPR380 ^[192] GPR381 ^[193] GPR382 ^[194] GPR383 ^[195] GPR384 ^[196] GPR385 ^[197] GPR386 ^[198] GPR387 ^[199] GPR388 ^[200] GPR389 ^[201] GPR390 ^[202] GPR391 ^[203] GPR392 ^[204] GPR393 ^[205] GPR394 ^[206] GPR395 ^[207] GPR396 ^[208] GPR397 ^[209] GPR398 ^[210] GPR399 ^[211] GPR400 ^[212] GPR401 ^[213] GPR402 ^[214] GPR403 ^[215] GPR404 ^[216] GPR405 ^[217] GPR406 ^[218] GPR407 ^[219] GPR408 ^[220] GPR409 ^[221] GPR410 ^[222] GPR411 ^[223] GPR412 ^[224] GPR413 ^[225] GPR414 ^[226] GPR415 ^[227] GPR416 ^[228] GPR417 ^[229] GPR418 ^[230] GPR419 ^[231] GPR420 ^[232] GPR421 ^[233] GPR422 ^[234] GPR423 ^[235] GPR424 ^[236] GPR425 ^[237] GPR426 ^[238] GPR427 ^[239] GPR428 ^[240] GPR429 ^[241] GPR430 ^[242] GPR431 ^[243] GPR432 ^[244] GPR433 ^[245] GPR434 ^[246] GPR435 ^[247] GPR436 ^[248] GPR437 ^[249] GPR438 ^[250] GPR439 ^[251] GPR440 ^[252] GPR441 ^[253] GPR442 ^[254] GPR443 ^[255] GPR444 ^[256] GPR445 ^[257] GPR446 ^[258] GPR447 ^[259] GPR448 ^[260] GPR449 ^[261] GPR450 ^[262] GPR451 ^[263] GPR452 ^[264] GPR453 ^[265] GPR454 ^[266] GPR455 ^[267] GPR456 ^[268] GPR457 ^[269] GPR458 ^[270] GPR459 ^[271] GPR460 ^[272] GPR461 ^[273] GPR462 ^[274] GPR463 ^[275] GPR464 ^[276] GPR465 ^[277] GPR466 ^[278] GPR467 ^[279] GPR468 ^[280] GPR469 ^[281] GPR470 ^[282] GPR471 ^[283] GPR472 ^[284] GPR473 ^[285] GPR474 ^[286] GPR475 ^[287] GPR476 ^[288] GPR477 ^[289] GPR478 ^[290] GPR479 ^[291] GPR480 ^[292] GPR481 ^[293] GPR482 ^[294] GPR483 ^[295] GPR484 ^[296] GPR485 ^[297] GPR486 ^[298] GPR487 ^[299] GPR488 ^[300] GPR489 ^[301] GPR490 ^[302] GPR491 ^[303] GPR492 ^[304] GPR493 ^[305] GPR494 ^[306] GPR495 ^[307] GPR496 ^[308] GPR497 ^[309] GPR498 ^[310] GPR499 ^[311] GPR500 ^[312] GPR501 ^[313] GPR502 ^[314] GPR503 ^[315] GPR504 ^[316] GPR505 ^[317] GPR506 ^[318] GPR507 ^[319] GPR508 ^[320] GPR509 ^[321] GPR510 ^[322] GPR511 ^[323] GPR512 ^[324] GPR513 ^[325] GPR514 ^[326] GPR515 ^[327] GPR516 ^[328] GPR517 ^[329] GPR518 ^[330] GPR519 ^[331] GPR520 ^[332] GPR521 ^[333] GPR522 ^[334] GPR523 ^[335] GPR524 ^[336] GPR525 ^[337] GPR526 ^[338] GPR527 ^[339] GPR528 ^[340] GPR529 ^[341] GPR530 ^[342] GPR531 ^[343] GPR532 ^[344] GPR533 ^[345] GPR534 ^[346] GPR535 ^[347] GPR536 ^[348] GPR537 ^[349] GPR538 ^[350] GPR539 ^[351] GPR540 ^[352] GPR541 ^[353] GPR542 ^[354] GPR543 ^[355] GPR544 ^[356] GPR545 ^[357] GPR546 ^[358] GPR547 ^[359] GPR548 ^[360] GPR549 ^[361] GPR550 ^[362] GPR551 ^[363] GPR552 ^[364] GPR553 ^[365] GPR554 ^[366] GPR555 ^[367] GPR556 ^[368] GPR557 ^[369] GPR558 ^[370] GPR559 ^[371] GPR560 ^[372] GPR561 ^[373] GPR562 ^[374] GPR563 ^[375] GPR564 ^[376] GPR565 ^[377] GPR566 ^[378] GPR567 ^[379] GPR568 ^[380] GPR569 ^[381] GPR570 ^[382] GPR571 ^[383] GPR572 ^[384] GPR573 ^[385] GPR574 ^[386] GPR575 ^[387] GPR576 ^[388] GPR577 ^[389] GPR578 ^[390] GPR579 ^[391] GPR580 ^[392] GPR581 ^[393] GPR582 ^[394] GPR583 ^[395] GPR584 ^[396] GPR585 ^[397] GPR586 ^[398] GPR587 ^[399] GPR588 ^[400] GPR589 ^[401] GPR590 ^[402] GPR591 ^[403] GPR592 ^[404] GPR593 ^[405] GPR594 ^[406] GPR595 ^[407] GPR596 ^[408] GPR597 ^[409] GPR598 ^[410] GPR599 ^[411] GPR600 ^[412] GPR601 ^[413] GPR602 ^[414] GPR603 ^[415] GPR604 ^[416] GPR605 ^[417] GPR606 ^[418] GPR607 ^[419] GPR608 ^[420] GPR609 ^[421] GPR610 ^[422] GPR611 ^[423] GPR612 ^[424] GPR613 ^[425] GPR614 ^[426] GPR615 ^[427] GPR616 ^[428] GPR617 ^[429] GPR618 ^[430] GPR619 ^[431] GPR620 ^[432] GPR621 ^[433] GPR622 ^[434] GPR623 ^[435] GPR624 ^[436] GPR625 ^[437] GPR626 ^[438] GPR627 ^[439] GPR628 ^[440] GPR629 ^[441] GPR630 ^[442] GPR631 ^[443] GPR632 ^[444] GPR633 ^[445] GPR634 ^[446] GPR635 ^[447] GPR636 ^[448] GPR637 ^[449] GPR638 ^[450] GPR639 ^[451] GPR640 ^[452] GPR641 ^[453] GPR642 ^[454] GPR643 ^[455] GPR644 ^[456] GPR645 ^[457] GPR646 ^[458] GPR647 ^[459] GPR648 ^[460] GPR649 ^[461] GPR650 ^[462] GPR651 ^[463] GPR652 ^[464] GPR653 ^[465] GPR654 ^[466] GPR655 ^[467] GPR656 ^[468] GPR657 ^[469] GPR658 ^[470] GPR659 ^[471] GPR660 ^[472] GPR661 ^[473] GPR662 ^[474] GPR663 ^[475] GPR664 ^[476] GPR665 ^[477] GPR666 ^[478] GPR667 ^[479] GPR668 ^[480] GPR669 ^[481] GPR670 ^[482] GPR671 ^[483] GPR672 ^[484] GPR673 ^[485] GPR674 ^[486] GPR675 ^[487] GPR676 ^[488] GPR677 ^[489] GPR678 ^[490] GPR679 ^[491] GPR680 ^[492] GPR681 ^[493] GPR682 ^[494] GPR683 ^[495] GPR684 ^[496] GPR685 ^[497] GPR686 ^[498] GPR687 ^[499] GPR688 ^[500] GPR689 ^[501] GPR690 ^[502] GPR691 ^[503] GPR692 ^[504] GPR693 ^[505] GPR694 ^[506] GPR695 ^[507] GPR696 ^[508] GPR697 ^[509] GPR698 ^[510] GPR699 ^[511] GPR700 ^[512] GPR701 ^[513] GPR702 ^[514] GPR703 ^[515] GPR704 ^[516] GPR705 ^[517] GPR706 ^[518] GPR707 ^[519] GPR708 ^[520] GPR709 ^[521] GPR710 ^[522] GPR711 ^[523] GPR712 ^[524] GPR713 ^[525] GPR714 ^[526] GPR715 ^[527] GPR716 ^[528] GPR717 ^[529] GPR718 ^[530] GPR719 ^[531] GPR720 ^[532] GPR721 ^[533] GPR722 ^[534] GPR723 ^[535] GPR724 ^[536] GPR725 ^[537] GPR726 ^[538] GPR727 ^[539] GPR728 ^[540] GPR729 ^[541] GPR730 ^[542] GPR731 ^[543] GPR732 ^[544] GPR733 ^[545] GPR734 ^[546] GPR735 ^[547] GPR736 ^[548] GPR737 ^[549] GPR738 ^[550] GPR739 ^[551] GPR740 ^[552] GPR741 ^[553] GPR742 ^[554] GPR743 ^[555] GPR744 ^[556] GPR745 ^[557] GPR746 ^[558] GPR747 ^[559] GPR748 ^[560] GPR749 ^[561] GPR750 ^[562] GPR751 ^[563] GPR752 ^[564] GPR753 ^[565] GPR754 ^[566] GPR755 ^[567] GPR756 ^[568] GPR757 ^[569] GPR758 ^[570] GPR759 ^[571] GPR760 ^[572] GPR761 ^[573] GPR762 ^[574] GPR763 ^[575] GPR764 ^[576] GPR765 ^[577] GPR766 ^[578] GPR767 ^[579] GPR768 ^[580] GPR769 ^[581] GPR770 ^[582] GPR771 ^[583] GPR772 ^[584] GPR773 ^[585] GPR774 ^[586] GPR775 ^[587] GPR776 ^[588] GPR777 ^[589] GPR778 ^[590] GPR779 ^[591] GPR780 ^[592] GPR781 ^[593] GPR782 ^[594] GPR783 ^[595] GPR784 ^[596] GPR785 ^[597] GPR786 ^[598] GPR787 ^[599] GPR788 ^[600] GPR789 ^[601] GPR790 ^[602] GPR791 ^[603] GPR792 ^[604] GPR793 ^[605] GPR794 ^[606] GPR795 ^[607] GPR796 ^[608] GPR797 ^[609] GPR798 ^[610] GPR799 ^[611] GPR800 ^[612] GPR801 ^[613] GPR802 ^[614] GPR803 ^[615] GPR804 ^[616] GPR805 ^[617] GPR806 ^[618] GPR807 ^[619] GPR808 ^[620] GPR809 ^[621] GPR810 ^[622] GPR811 ^[623] GPR812 ^[624] GPR813 ^[625] GPR814 ^[626] GPR815 ^[627] GPR816 ^[628] GPR817 ^[629] GPR818 ^[630] GPR819 ^[631] GPR820 ^[632] GPR821 ^[633] GPR822 ^[634] GPR823 ^[635] GPR824 ^[636] GPR825 ^[637] GPR826 ^[638] GPR827 ^[639] GPR828 ^[640] GPR829 ^[641] GPR830 ^[642] GPR831 ^[643] GPR832 ^[644] GPR833 ^[645] GPR834 ^[646] GPR835 ^[647] GPR836 ^[648] GPR837 ^[649] GPR838 ^[650] GPR839 ^[651] GPR840 ^[652] GPR841 ^[653] GPR842 ^[654] GPR843 ^[655] GPR844 ^[656] GPR845 ^[657] GPR846 ^[658] GPR847 ^[659] GPR848 ^[660] GPR849 ^[661] GPR850 ^[662] GPR851 ^[663] GPR852 ^[664] GPR853 ^[665] GPR854 ^[666] GPR855 ^[667] GPR856 ^[668] GPR857 ^[669] GPR858 ^[670] GPR859 ^[671] GPR860 ^[672] GPR861 ^[673] GPR862 ^[674] GPR863 ^[675] GPR864 ^[676] GPR865 ^[677] GPR866 ^[678] GPR867 ^[679] GPR868 ^[680] GPR869 ^[681] GPR870 ^[682] GPR871 ^[683] GPR872 ^[684] GPR873 ^[685] GPR874 ^[686] GPR875 ^[687] GPR876 ^[688] GPR877 ^[689] GPR878 ^[690] GPR879 ^[691] GPR880 ^[692] GPR881 ^[693] GPR882 ^[694] GPR883 ^[695] GPR884 ^[696] GPR885 ^[697] GPR886 ^[698] GPR887 ^[699] GPR888 ^[700] GPR889 ^[701] GPR890 ^[702] GPR891 ^[703] GPR892 ^[704] GPR893 ^[705] GPR894 ^[706] GPR895 ^[707] GPR896 ^[708] GPR897 ^[709] GPR898 ^[710] GPR899 ^[711] GPR900 ^[712] GPR901 ^[713] GPR902 ^[714] GPR903 ^[715] GPR904 ^[716] GPR905 ^[717] GPR906 ^[718] GPR907 ^[719] GPR908 ^[720] GPR909 ^[721] GPR910 ^[722] GPR911 ^[723] GPR912 ^[724] GPR913 ^[725] GPR914 ^[726] GPR915 ^[727] GPR916 ^[728] GPR917 ^[729] GPR918 ^[730] GPR919 ^[731] GPR920 ^[732] GPR921 ^[733] GPR922 ^[734] GPR923 ^[735] GPR924 ^[736] GPR925 ^[737] GPR926 ^[738] GPR927 ^[739] GPR928 ^[740] GPR929 ^[741] GPR930 ^[742] GPR931 ^[743] GPR932 ^[744] GPR933 ^[745] GPR934 ^[746] GPR935 ^[747] GPR936 ^[748] GPR937 ^[749] GPR938 ^[750] GPR939 ^[751] GPR940 ^[752] GPR941 ^[753] GPR942 ^[754] GPR943 ^[755] GPR944 ^[756] GPR945 ^[757] GPR946 ^[758] GPR947 ^[759] GPR948 ^[760] GPR949 ^[761] GPR950 ^[762] GPR951 ^[763] GPR952 ^[764] GPR953 ^[765] GPR954 ^[766] GPR955 ^[767] GPR956 ^[768] GPR957 ^[769] GPR958 ^[770] GPR959 ^[771] GPR960 ^[772] GPR961 ^[773] GPR962 ^[774] GPR963 ^[775] GPR964 ^[776] GPR965 ^[777] GPR966 ^[778] GPR967 ^[779] GPR968 ^[780] GPR969 ^[781] GPR970 ^[782] GPR971 ^[783] GPR972 ^[784] GPR973 ^[785] GPR974 ^[786] GPR975 ^[787] GPR976 ^[788] GPR977 ^[789] GPR978 ^[790] GPR979 ^[791] GPR980 ^[792] GPR981 ^[793] GPR982 ^[794] GPR983 ^[795] GPR984 ^[796] GPR985 ^[797] GPR986 ^[798] GPR987 ^[799] GPR988 ^[800] GPR989 ^[801] GPR990 ^[802] GPR991 ^[803] GPR992 ^[804] GPR993 ^[805] GPR994 ^[806] GPR995 ^[807] GPR996 ^[808] GPR997 ^[809] GPR998 ^[810] GPR999 ^[811] GPR1000 ^[812] GPR1001 ^[813] GPR1002 ^[814] GPR1003 ^[815] GPR1004 ^[816] GPR1005 ^[817] GPR1006 ^[818] GPR1007 ^[819] GPR1008 ^[820] GPR1009 ^[821] GPR1010 ^[822] GPR1011 ^[823] GPR1012 ^[824] GPR1013 ^[825] GPR1014 ^[826] GPR1015 ^[827] GPR1016 ^[828] GPR1017 ^[829] GPR1018 ^[830] GPR1019 ^[831] GPR1020 ^[832] GPR1021 ^[833] GPR1022 ^[834] GPR1023 ^[835] GPR1024 ^[836] GPR1025 ^[837] GPR1026 ^[838] GPR1027 ^[839] GPR1028 ^[840] GPR1029 ^[841] GPR1030 ^[842] GPR1031 ^[843] GPR1032 ^[844] GPR1033 ^[845] GPR1034 ^[846] GPR1035 ^[847] GPR1036 ^[848] GPR1037 ^[849] GPR1038 ^[850] GPR1039 ^[851] GPR1040 ^[852] GPR1041 ^[853] GPR1042 ^[854] GPR1043 ^[855] GPR1044 ^[856] GPR1045 ^[857] GPR1046 ^[858] GPR1047 ^[859] GPR1048 ^[860] GPR1049 ^[861] GPR1050 ^[862] GPR1051 ^[863] G

Name: _____ Date: _____

Wikipedia – Video Games – CSV Format

Background:
The data below represent the same 5 video games from the Wikipedia webpage, but in CSV format. CSV stands for "comma separated values."

```
Game,Original release year,Genre,Number of lists,Platform of original release,Lists/References
The Legend of Zelda: Ocarina of Time,1998,Action-Adventure,42,Nintendo 64,"CNET,[1]
CVG2000,[2] Edge2000,[3] Edge2007,[4] Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM2001,[8]
EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13] GameRankings,[14]
GamingBolt,[15] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GI1999,[20] GI2005,[21]
GI2009,[22] GameSpot2006,[23] GameSpot2014,[24] GuinnessConsole,[25] IGN10s,[26] IGN2003,[27]
IGN2005,[28] IGN2006Readers,[29] IGN2008Readers,[30] IGN2007,[31] IGN2008Readers,[32]
Maniac,[33] Metacritic,[34] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38]
Stuff2008,[39] Stuff2014,[40] UnkGamer,[41] Yahoo[42]"
Chrono Trigger,1995,Role-Playing Game,39,Super Nintendo Entertainment
System,"ActionButton,[43] Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45]
EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13]
GameRankings,[14] GamesRadar,[46] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19]
GameSpot2006,[23] GamingBolt,[15] GI1998,[47] GI1999,[20] GI2001,[21] GI2009,[22]
GuinnessConsole,[25] IGN10s,[26] IGN2003,[27] IGN2005,[28] IGN2006Readers,[29]
IGN2006Readers,[30] IGN2008Readers,[32] IGN10s,[26] Maniac,[33] NowGamer,[35] PM2014,[36]
Slant2014,[38] Time,[48] UnkGamer[41]"
Final Fantasy VII,1997,Role-Playing Game,39,PlayStation,"CVG2000,[2] CVG2007PC,[49]
Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45] EGM1997Dev,[50]
EGM1997Readers,[51] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12]
G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GameSpot2006,[23] GamingBolt,[15]
GI1998,[47] GI2001,[21] GI2009,[22] GuinnessConsole,[25] IGN2005,[28] IGN2006Readers,[29]
IGN2006Readers,[30] IGN2007,[31] IGN2008Readers,[32] NowGamer,[35] PM2014,[36] RetroGamer,[37]
Slant2014,[38] Stuff2008,[39] Stuff2014,[40] Time,[48] UnkGamer,[41] Yahoo[42]"
Super Mario 64,1996,Platformer,38,Nintendo 64,"CVG2000,[2] Edge2000,[3] Edge2007,[4]
Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM1997,[45] EGM1997Dev,[50] EGM1997Readers,[51]
EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] G4,[13] GF2004,[16] GF2005,[17]
GF2009,[18] GF2014,[19] GamingBolt,[15] GI2001,[21] GI2009,[22] GamesRadar,[46]
GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2006Readers,[29] IGN2006Readers,[30]
Maniac,[33] NextGen1996,[32] NowGamer,[35] GameRankings,[14] IGN2007,[31] IGN2008Readers,[32]
PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Time,[48] Yahoo[42]"
Street Fighter II,1991,Fighting,37,Arcade,"CVG2000,[2] Edge2000,[3] EGM1992,[53] EGM1997,[45]
EGM1997Arcade,[54] EGM1997Dev,[50] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11]
Famitsu2006,[12] FHM,[55] G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19]
GameSpot2006,[23] GameSpyArcade,[56] GamingBolt,[15] GI2001,[21] GI2009,[22]
GuinnessArcade,[57] GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2007,[31]
NextGen1996,[52] NextGen1999,[58] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38]
Stuff2008,[39] Stuff2014,[40] Time,[48] Yahoo[42]"
```

LMR_3.18

- 10. Inform students that a file with the CSV format is easily readable by R. Then ask:
 - a. Where are the variable names stored? *The variable names are stored in the first row*
 - b. How are values of the variables separated? *The values are separated by commas.*
 - c. If we were interested in using the online data, how would we obtain it? *This is a challenging problem – one which students may not know how to answer at this point. The objective is for them to struggle with how they would obtain data and recognize that it is not always as simple as “export, upload, import.”*



- 11. Split the class into their student teams and distribute the *Online Data-ing* handout (LMR_3.19). Assign each team a different website (each page of the handout lists a different site) and have them use this site to complete the questions in the handout.

Name: _____ Date: _____

Online Data-ing

Instructions:
Visit the assigned website and, with your team, answer each question below. If you do not see data at the top of the page, explore the website a bit to find some.

Assigned Website: Yelp – www.yelp.com

1. Describe the data on this website.

2. What variables are present?

3. What type of values do the variables have (i.e. words, numbers, dates, places, categories, etc.)?

4. Where do the values of the data come from?

5. How often are they updated? Can you tell?

6. Who collected the data (regular people, professionals, scientists, etc.)?

7. Who is the target audience for the data? In other words, who would most likely use this data, and why?

8. Can you think of ways you might get the data from the website into RStudio for analysis? If so, explain how. If not, why?

LMR_3.19



12. Have each student team share their findings with one other team. They should have their website displayed while discussing their results.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next 2 Days

For the next 4 days, students will collect data using their newly created Participatory Sensing campaign.

Lab 3E: Scraping Web Data

Lab 3F: Maps

Complete Labs 3E and 3F prior to Lesson 21.

Lab 3E - Scraping web data

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

The web as a data source

- The internet contains huge amounts of information.
 - Using computers to gather this information in an automated fashion is referred to as *scraping web data*.
 - Scraping data from the web can be difficult because each website displays & stores data differently.
- In this lab, we'll learn how to scrape data in two steps:
 - Step 1: Gather information from the web.
 - Step 2: Clean it up and turn it into a usable data frame for Lab 3F.

Our first web scraper

- Copy and paste the link below into a web browser to view the website of data we'd like to *scrape* and analyze.
<https://labs.idsucla.org/extras/webdata/mountains.html>
- **Briefly describe what the data on the website is about.**
 - **Then write down 3 questions you'd be interested in answering by analyzing this data.**

HTML

- HTML is the code that's used to render every website you've ever visited.
- The following slide shows the HTML code used to create the first two rows of the web data.
 - **How is the data table in HTML different than the data tables we're used to seeing in R, for example, when we use the `View()` function?**
 - **What do you think the *tags* `<TABLE>`, `<TR>`, `<TH>`, `<TD>` mean? How does HTML use these *tags* to display the table?**

```
<TABLE>
  <TR>
    <TH>peak</TH>
    <TH>range</TH>
    <TH>state</TH>
    <TH>long</TH>
    <TH>lat</TH>
    <TH>elev_ft</TH>
    <TH>elev_m</TH>
    <TH>prominence_ft</TH>
    <TH>prominence_m</TH>
    <TH>rank</TH>
  </TR>
  <TR>
    <TD>Denali (Mount McKinley)</TD>
    <TD>Alaska Range</TD>
    <TD>Alaska</TD>
    <TD>-151.0063</TD>
    <TD>63.0690</TD>
```

```
<TD>20236</TD>
<TD>6168</TD>
<TD>20174</TD>
<TD>6149</TD>
<TD>1</TD>
</TR>
</TABLE>
```

Get to scraping!

- Use your browser to go back to the website with the data we're interested in scraping.
- Find the URL address for the site and assign it the name `data_url` in R.
 - Then fill in the blanks below to have R scrape *every* web table available on the site:

```
tables <- readHTMLTable(____)
```

Find our data

- Since `readHTMLTable()` scrapes *every* table that is on a particular web URL, we need to find out which table has the data we're interested in.
 - For example, `wikipedia.org` often has articles with 3 or more tables.
 - This means we need to check all 3 tables to find the one we're interested in.
- Use the `length()` function to find out how many tables of data were scraped in our set of tables.

Saving tables

- Now that we know how many tables we've scraped, we can go back and scrape individual tables by adding the `which` argument to the `readHTMLTable()` function.
 - Use `readHTMLTable()` to re-scrape the data from the web but this time use the `which` argument to scrape just the individual table.
 - The `which` argument should be the integer denoting which table you want scraped.
 - Assign the scraped data the name `mtns`

Check, save and use!

- After scraping the data, the only thing left to do is to save it and use it.
- Fill in the blanks to save the data and give it a file name

```
save(____, file = "____.Rda")
```

- **What is the mean and standard deviation of `elev_ft`?**
- **Which state has the most mountains in our data?**

Lab 3F - Maps

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Informative and Fun!

- Maps are some of the most interesting plots to make because the info represents:
 - Where we live.
 - Where we go.
 - Places that interest us.
- Maps are also helpful to display geographic information.
 - John Snow (the physician, not the character from *Game of Thrones*...) once famously used [a map to discover how cholera was transmitted](#).
- In this lab, we'll use R to create an interactive map of the mtns data we scraped in Lab 3E.

Getting ready to map

- The map we'll be creating will end up in RStudio's *Viewer* pane.
 - Which means you'll need to alternate between building the map and loading the lab.
- You'll find it very helpful, for this lab, to write all of the commands, including the `load_lab(23)` command, as an R script.
 - This way you can edit the code that builds the map and quickly reload the lab.

Load your data!

- In Lab 3E you created a dataset. Load it into RStudio now by filling in the blank with the file name of the data.

```
load("____.Rda")
```

- Didn't finish the lab or save the data file? Ask a friend to share it!

Build a Basic Map

- Let's start by building a basic map!
- Use the `leaflet()` function and the mtns data to create the leaf that we can use for mapping.

```
mtns_leaf <- leaflet(____)
```

- Then, insert `mtns_leaf` into the `addTiles()` function and assign the output the name `mtns_map`
- Run `mtns_map` in the console to look at your basic map with no data displayed.
 - Be sure to try clicking on the map to pan and zoom.

Including our data

- Now we can add markers for the locations of the mountains using the `addMarkers()` function.
 - Fill in the blanks below with the basic map we've created and the values for latitude and longitude.

```
addMarkers(map = _____, lng = ~____, lat = ~____)
```

- Supply the peak variable, in a similar way as we supplied the `lat` and `long` variables, to the `popup` argument and include it in the code above.

- Click on a marker within California and write down the name of the mountain you clicked on.

Colorize

- Our current map looks pretty good, but what if we wanted to add some colors to our plot?
- Fill in the blanks below to create a new variable that assigns a color to each mountain based on the state its located.

```
mtns <- mutate(____, state_colors = colorize(____))
```

- Now that we've added a new variable, we need to re-build `mtns_leaf` and `mtns_map` to use it.
 - Create `mtns_leaf` and `mtns_map` as you did before.
 - Then change `addMarkers` to `addCircleMarkers` and keep all of the arguments the same.

Showing off our colors

- To add the colors to our plot, use the `addCircleMarkers` like before but this time include `color = ~state_colors` as an argument.
- It's hard to know just what the different colors mean so let's add a legend.
 - First, assign the map with the circle markers as `mtns_map`.
 - Then, fill in the blanks below to place a legend in the top-right hand corner.

```
addLegend(____, colors = ~unique(____), labels = ~unique(____))
```

Lesson 21: Learning to Love XML

Objective:

Students will understand the need for data to be stored in different ways - specifically, why it makes sense for web data to be formatted as XML.

Materials:

1. *Online Data-ing* handout (LMR_3.19_Online Data-ing)
Note: This should have been completed during the previous class.
2. Mountain Peak XML data found at:
<https://labs.idsucla.org/extras/webdata/mountains.html>
Note: Open with Google Chrome or Firefox browsers, NOT with Safari.
3. Projector
4. *Mountains – HTML vs. XML* handout (LMR_3.20_Mountins – HTML vs. XML)

Vocabulary:

XML

Essential Concepts: XML is a programming language that we use with our campaigns. We create basic XML "tags" in the code, which help us store data in a format we understand.

Lesson:

1. Allow time for student teams to present their findings from the *Online Data-ing* handout (LMR_3.19) if there was not sufficient time during the previous lesson.
2. Remind students that in the previous lesson they learned about a variety of ways that data can be presented online.
3. They've been working with comma separated (CSV) files and R data frames. Last time and in the lab, they worked with HTML tables. Today they are going to learn how HTML can be displayed as an XML table.
4. **XML**, or Extensible Mark up Language, is a popular format for storing data on the Internet. It is useful because it creates readable web pages, and also because it allows programmers to easily update values in the data table if those values change.
5. In pairs, ask students to brainstorm ways in which data that is found online is different than the way we see data in RStudio. Then, create a class brainstorm from the student pair responses.
6. After the brainstorm, emphasize the following:
 - a. RStudio's default way to work with data is as large data frames (tables) where rows represent observations and columns represent variables.
 - b. Data that is viewed online often has a different structure.
 - c. Data structures found on the web might be displayed in tables, such as those on Wikipedia, or streams, such as Twitter, and might even include data spread across multiple sections of a web page, such as Yelp.

Show students, on a projector, the Mountain Peak XML data found at
<https://labs.idsucla.org/extras/webdata/mountains.html>

Ask students to look at the data and determine if they have seen it before. Hint: They have! It was the data they scraped during Lab 3E.

7. Once students figure out that the XML is just the same data as the website they scraped during Lab 3E, distribute the *Mountains – HTML vs. XML* handout (LMR_3.20), which displays both HTML and XML versions of the data.

Note: The handout only includes the first 3 mountains.

Name: _____

Date: _____

Mountains – HTML vs. XML

Background:

The Mountain Peak data are displayed below in two different formats – the first is HTML, and the second is XML.

peak	range	state	long	lat	elev_ft	elev_m	prominence_ft	prominence_m	rank
Mount McKinley (Denali)	Alaska Range	Alaska	-151.0063	63.0689	20236	6168	20174	6149	1
Mount Saint Elias	Saint Elias Mountains	Alaska	-140.9264	60.2931	18009	5489	11250	3429	2
Mount Foraker	Alaska Range	Alaska	-151.3998	62.9604	17400	5304	7250	2210	3

```
<mountainpeaks>
  <data>
    <mountain>
      <countin>
        <peak>Mount McKinley (Denali)</peak>
        <range>Alaska Range</range>
        <state>Alaska</state>
        <long>-151.0063</long>
        <lat>63.0689</lat>
        <elev_ft>20236</elev_ft>
        <elev_m>6168</elev_m>
        <prominence_ft>20174</prominence_ft>
        <prominence_m>6149</prominence_m>
        <rank>1</rank>
      </mountain>
    <mountain>
      <peak>Mount Saint Elias</peak>
      <range>Saint Elias Mountains</range>
      <state>Alaska</state>
      <long>-140.9264</long>
      <lat>60.2931</lat>
      <elev_ft>18009</elev_ft>
      <elev_m>5489</elev_m>
      <prominence_ft>11250</prominence_ft>
      <prominence_m>3429</prominence_m>
      <rank>2</rank>
    </mountain>
    <mountain>
      <peak>Mount Foraker</peak>
      <range>Alaska Range</range>
      <state>Alaska</state>
      <long>-151.3998</long>
      <lat>62.9604</lat>
      <elev_ft>17400</elev_ft>
      <elev_m>5304</elev_m>
      <prominence_ft>7250</prominence_ft>
      <prominence_m>2210</prominence_m>
      <rank>3</rank>
    </mountain>
  </data>
</mountainpeaks>
```

LMR_3.20

8. Ask student pairs to answer the following:
 - a. Why are certain XML tags indented in the XML version of the data? *The indentations tell us how to structure the HTML table. For example, all the mountains are contained in the <data> section, but are further tagged by each particular mountain within the <mountain> and </mountain> tags. All information stored between those two tags will be displayed as one row of the HTML table.*
 - b. What are the role of tags (ex. <state>) and end tags (ex. </state>) in the XML code? *Tags tell us when a certain type of data begins, and end tags tell us when the data should end. In other words, it tells us where to find the specific values of a variable (ex. Alaska would be the value of the “state” variable since it is between the <state> and </state> tags.*
 - c. Where are the variable names? *The variable names can be found between each <mountain> and </mountain> tags. Specifically, the first variable is “peak” and the last variable is “rank.”*
 - d. Where are the observations? *The observations are located within each of the variable tags. For example, the observation “Mount McKinley (Denali)” is found between the <peak> and </peak> tags.*
9. Assign student pairs one of the above questions to share out with the class. Student pairs that did not receive an assignment must participate using the *Agree/Disagree* strategy.
10. As a class, discuss the answers to the questions above.
11. XML formats make it easier to display data on the web in a pleasant matter and make it easier for programmers to find and alter data if the values change or if, for example, they wish to add a new row to a table.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 3 days, students will collect data using the class’s newly created Participatory Sensing campaign (see Lessons 16-18).



For homework, students should reflect about how XML and HTML data are displayed. They should discuss when each format is appropriate.

Lesson 22: Changing Format

Objective:

Students will learn how to convert XML files to the more familiar data table format and vice versa.

Materials:

1. *There and Back Again: From XML to Data Tables* handout (LMR_3.21_From XML to Data Tables)
2. *There and Back Again: From Data Tables to XML* handout (LMR_3.22_From Data Tables to XML)

Essential Concepts: Converting XML to spreadsheet format helps us better understand and view our data.

Lesson:

1. Take a few minutes to compare the structure of XML code to HTML data tables (refer to Step 7 from Lesson 21).
2. Inform students that in today's lesson, they will learn how to translate information from XML code into a data table.
3. Distribute the *There and Back Again: From XML to Data Tables* handout (LMR_3.21) to students.

Name: _____ Date: _____

**There and Back Again:
From XML to Data Tables**

Instructions:
Translate the XML data into an R data table, and answer the questions on page 2.

```
<volunteers>
  <data>
    <volunteer>
      <name>Salazar</name>
      <organization>No Kill LA (NOLA)</organization>
      <time1>1</time>
    </volunteer>
    <volunteer>
      <name>Hayden</name>
      <organization>Vosemite Foundation</organization>
      <time1>3</time>
    </volunteer>
    <volunteer>
      <name>Charlie</name>
      <organization>Vosemite Foundation</organization>
      <time1>2</time>
    </volunteer>
    <volunteer>
      <name>Emerson</name>
      <organization>City of Hope</organization>
      <time1>1</time>
    </volunteer>
    <volunteer>
      <name>Jessie</name>
      <organization>Roundwood Warrior Project</organization>
      <time1>2</time>
    </volunteer>
    <volunteer>
      <name>Sawyer</name>
      <organization>City of Hope</organization>
      <time1>2</time>
    </volunteer>
    <volunteer>
      <name>Karyn</name>
      <organization>No Kill LA (NOLA)</organization>
      <time1>1</time>
    </volunteer>
    <volunteer>
      <name>London</name>
      <organization>LA Regional Food Bank</organization>
      <time1>4</time>
    </volunteer>
  </data>
</volunteers>
```

Name of Data: _____

LMR_3.21

4. Inform the students that XML code is provided on page 1 of the handout, and their goal is to transfer all the information to the empty data table.
5. As a guide, ask a volunteer to find and name one of the variables in the XML code and then have all the students write the name of the variable in the first column of the top row in the data table.
6. Next, ask another student to find the first value of the variable named in Step 5. This value should be placed in the correct column and row of the data table.



- Provide time for students to complete the handout individually.
- Using the Anonymous Author strategy, share a couple of the completed data tables. Ask teams to discuss how they are alike and how they are different.

Note: Most tables will probably be the same, but could vary slightly based on which columns each variable name was placed in, and in what order the observations were listed in the rows. Ultimately, the information contained in the data tables is the same.

- Then, conduct a whole class discussion regarding student responses to the questions on page 2 of the handout.
- Distribute the *There and Back Again: From Data Tables to XML* (LMR_3.22) to student teams and allow them time to complete it.

Name: _____ Date: _____

**There and Back Again:
From Data Tables to XML**

Instructions:
Translate the data table into an XML data file using appropriate tags and end tags.

Name of Data: **Yosemite Hiking Trails**

trail_name	park_area	miles
North Rim	Yosemite Valley	27.4
South Rim	Yosemite Valley	21.6
Glen Aulin	Tuolumne Meadows	10.6
10 Lakes Basin	Tioga Road	12.4
Clouds Rest	Tuolumne Meadows	14.0

```

<hikingTrails>
<data>
<trail>
< > </ >
< > </ >
< > </ >
</trail>
< > </ >
< > </ >
< > </ >
</ >
< > </ >
< > </ >
< > </ >
</ >
< > </ >
< > </ >
< > </ >
</ >
< > </ >
< > </ >
< > </ >
</ >
</data>
</hikingTrails>

```

LMR_3.22

- Once teams have finished, teams will guide you to write the correct XML code.
- Using a *Whip Around*, teams will tell you the first line of the XML code you need to write. Teams waiting their turn will check if the team is guiding you correctly. If not, they need to stop you and propose their line of code. You may not continue writing the lines of code until all teams are in agreement.

Class Scribes:



One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students will continue to collect data using the class's Participatory Sensing campaign (see Lessons 17-19). They will analyze the data the next day during the practicum.

Practicum: What Does Our Campaign Data Say?

Objective:

Students will answer the statistical question they generated at the beginning of the Participatory Sensing campaign creation lesson. They will use RStudio to make graphical representations or numerical summaries of their data to answer their question.

Materials:

1. *Our Own Campaign* (LMR_U3_Practicum_Our Own Campaign)

Practicum Our Own Campaign

At the start of the Participatory Sensing campaign creation in lesson 16, the class developed a research question about your class's topic of interest.

It is now time to analyze and interpret your class campaign data. You will use the data from your class-created campaign only. Based on the analysis, you can also wonder about what other data would be necessary to better answer your question, if any.

Based on the class's campaign data collected:

1. Refer back to the statistical questions your class generated in lessons 16-18 that address the research question.
2. Choose one of these statistical questions and determine which variables will answer this question.
3. Analyze the data to answer the question you've chosen. Your analysis should include graphs and numerical summaries. You should:
 - a. Provide the plot and numerical summary.
 - b. Describe what the plot shows.
 - c. Explain why you chose to make that particular plot.
 - d. Explain how the plot and numerical summary answers your statistical question.
 - e. Include the code you used in RStudio to make your plot.
4. After analyzing your data, determine if additional data would better answer your statistical question. If so, propose what that data would be. Different variables? Different data collection approach? Same variables, but more people? Same variables and people but more time?
5. Now, choose two more statistical questions that address the research question.
6. Analyze and interpret the data to answer these questions.
7. Sometimes, when analyzing data, we think of new statistical questions to ask, or we realize that the data need to be cleaned before we can answer. Explain whether this is the case with any of your statistical questions.
8. Write a one-page report and present it to another member of the class who is not in your team.

End of Unit Project and Oral Presentation: TB or Not TB?

Objective:

Students will apply what they have learned in the unit.

Materials:

1. Computers
2. *IDS Unit 3 – Project and Oral Presentation* (LMR_U3_End of Unit Project)

IDS Unit 3 – End of Unit Project TB or Not TB

Experiments in the medical field that involve new treatments (new medications) are called clinical trials. You have received a data set that shows the results from Sir Austin Bradford Hill's first randomized study in 1948 examining the effects of the antibiotic Streptomycin on 107 tuberculosis patients. You and a partner will use this data set to find out if Streptomycin is an effective treatment for tuberculosis.

A short article about tuberculosis facts can be found at:



<http://www.cdc.gov/tb/publications/factsheets/general/tb.htm>

Since this is an experiment, answer the following questions below. You may need to research the answer to some of the questions.

- a. What is the research question?
- b. Who are the subjects that participated in the experiment?
- c. What is the treatment?
- d. Who is in the treatment group?
- e. Who is in the control group?
- f. How were the subjects assigned to each group?
- g. What population is this experiment representative of?
- h. What is the variable that we will be measuring?
- i. What is the outcome of this experiment?

To answer your research question, you and a partner will compare the outcome of the data with the outcomes given by a chance model (in which Streptomycin has no effect on TB).

1. First, scrape the data. Refer to the web scraping lab if you need to recall how to scrape data. To access Sir Hill's data, go to: <https://labs.idsucla.org/extras/webdata/tb.html>
2. Second, determine the percentages of subjects in the study that died and the percentages of the subjects that recovered for each group.
3. Third, assuming that the treatment had no effect, use the data to:
 - a. Calculate the percentage of people with tuberculosis we would expect to die.
 - b. Use the *expected* percentage for (a), above, to calculate the number of people we expect to die from the treatment group.
 - c. Compare the percentage from (b) to the percentage from the treatment group *actually* died.
4. Then, if we assume that the outcome does not depend on the treatment, design and complete an appropriate simulation in RStudio using a chance model to replicate Sir Hill's study:
 - a. Shuffle the treatment and control labels 300 times; each time, calculate the percentage of treatment patients who "died". Plot the distribution of the 300 percentages. Refer to the simulation labs if you need to recall how to create a simulation.
 - b. Use the results from the chance model (shuffling) to determine whether (i.) or (ii.) below is the most reasonable explanation for the actual data in Sir Hill's study and state why:

- i. Streptomycin is a much better treatment for tuberculosis than bed rest. So, the outcome depends on the treatment.
- ii. The actual difference between treatments is due to chance; Streptomycin may not be effective on tuberculosis. So, it is possible that treatment and outcome are independent.

5. Can we say that Streptomycin **causes** the recovery of tuberculosis patients? Explain your answer.

Create a 4-5 slide, 5-minute presentation that shows your results. Be sure to include a detailed explanation of how you and your partner decided to conduct your simulation. Each person must participate in the presentation. In addition to the presentation, submit a 2-4 page, double-spaced summary of your analysis.