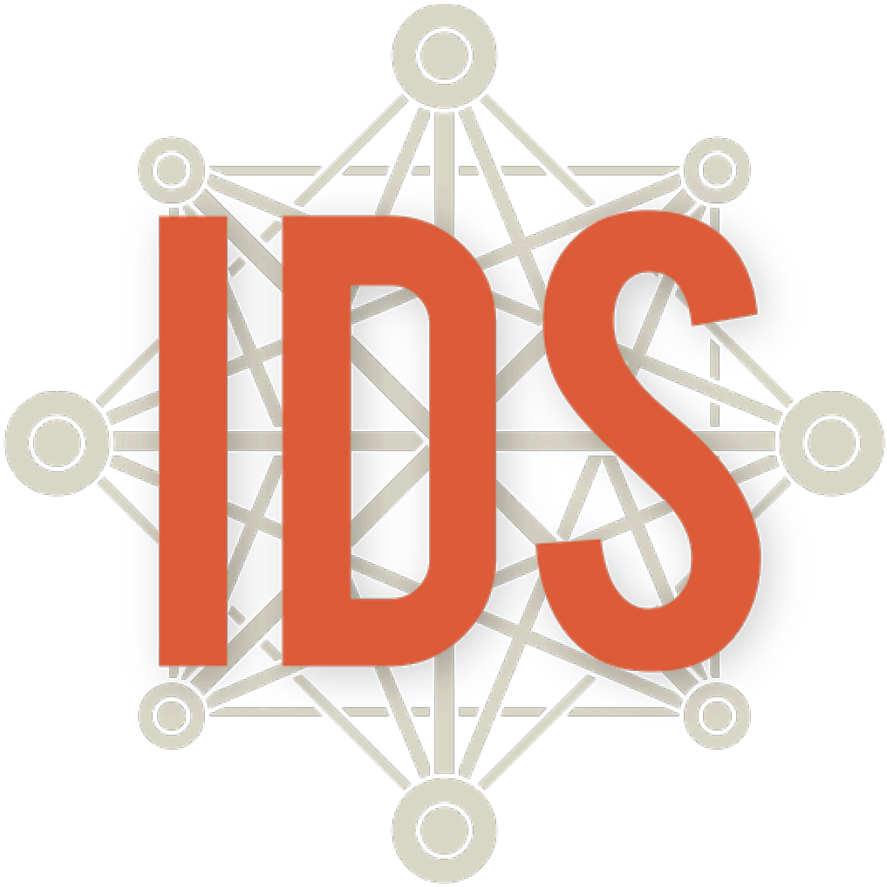


Introduction to Data Science

Unit 2

Introduction to Data Science
Daily Overview: Unit 2

Theme	Day	Lessons and Labs	Campaign	Topics	Page
What Is Your True Color? (10 days)	1	Lesson 1: What Is Your True Color?	Personality Color - data	Subsets, relative frequency	7
	2	Lesson 2: What Does *Mean* Mean?	Personality Color	Measures of center – mean	10
	3	Lesson 3: Median In the Middle	Personality Color	Measures of center – median	14
	4	Lesson 4: How Far Is It from Typical?	Personality Color	Measures of spread – MAD	18
	5	Lab 2A: All About Distributions	Personality Color	Measures of center & spread – mean, median, MAD	22
	6	Lesson 5: Human Boxplots		Boxplots, IQR	24
	7	Lesson 6: Face Off		Comparing distributions	27
	8	Lesson 7: Plot Match		Comparing distributions	29
	9	Lab 2B: Oh, the Summaries…	Personality Color	Boxplots, IQR, numerical summaries, custom functions	32
	10	Practicum: The Summaries	Food Habits or Time Use	Statistical questions, comparing distributions	35
How Likely Is It? (7 days)	11	Lesson 8: How Likely Is It?		Probability, simulations	39
	12	Lesson 9: Bias Detective		Simulations to detect bias	43
	13	Lesson 10: Marbles, Marbles		Probability, with replacement	47
	14	Lab 2C: Which Song Plays Next?		Probability of simple events, do loops, set.seed()	49
	15	Lesson 11: This AND/OR That		Compound probabilities	52
	16	Lab 2D: Queue It Up!		Probability with & without replacement, sample()	56
	17	Practicum: Win, Win, Win		Probability estimation through repeated simulations	59
Are You Stressing or Chilling? (8 Days)	18^	Lesson 12: Don't Take My Stress Away	Stress/Chill – data	Introduction to campaign	62
	19	Lesson 13: The Horror Movie Shuffle	Stress/Chill – data	Chance differences – cat var	66
	20	Lab 2E: The Horror Movie Shuffle	Stress/Chill – data	Inference for categorical variable, do loops, shuffle()	70
	21	Lesson 14: The Titanic Shuffle	Stress/Chill – data	Chance differences – num var	73
	22	Lab 2F: The Titanic Shuffle	Stress/Chill – data	Inference for numerical variable, do loops, shuffle()	77
	23+	Lesson 15: Tangible Data Merging	Stress/Chill – data	Merging data sets	79
	24	Lab 2G: Getting It Together	Stress/Chill & Personality Color	Merging data sets, stacking vs. joining	81
	25	Practicum: What Stresses Us?	Stress/Chill & Personality Color	Answering statistical questions of merged data	83
What's Normal? (5 Days)	26	Lesson 16: What Is Normal?		Introduction to normal curve	86
	27	Lesson 17: Normal Measure of Spread		Measures of spread - SD	90
	28	Lesson 18: What's Your Z-Score?		z-scores, shuffling	93
	29	Lab 2H: Eyeballing Normal		Normal curves overlaid on distributions & simulated data	98
	30	Lab 2I: R's Normal Distribution Alphabet		Normal probability, rnorm(), pnorm(), quantiles, qnorm()	100
Unit 2 Project (5 Days)	31-35	End of Unit Project and Oral Presentations: Asking and Answering Statistical Questions of Our Own Data	Stress/Chill, Personality Color, Food Habits, or Time Use	Synthesis of above	102

^=Data collection window begins.
+=Data collection window ends.

IDS Unit 2: Essential Concepts

Lesson 1: What Is Your True Color?

Students will understand that the 'typical' value is a value that can represent the entire group, even though we know that not all members of the group share the same value.

Lesson 2: What Does Mean Mean?

The center of a distribution is the 'typical' value. One way of measuring the center is with the mean, which finds the balancing point of the distribution. The mean gives us the typical value, but does not tell the whole story. We need a way to measure the variability to understand how observations might differ from the typical value.

Lesson 3: Median In the Middle

Another measure of center is the median, which can also be used to represent the typical value of a distribution. The median is preferred for skewed distributions or when there are outliers, because it better matches what we think of as 'typical.'

Lesson 4: How Far Is It from Typical?

MAD measures the variability in a sample of data - the larger the value, the greater the variability. More precisely, the MAD is the typical distance of observations from the mean. There are other measures of spread as well, notably the standard deviation and the interquartile range (IQR).

Lesson 5: Human Boxplots

A common statistical question is “How does this group compare to that group?” This is a hard question to answer when the groups have lots of variability. One approach is to compare the centers, spreads, and shapes of the distributions. Boxplots are a useful way of comparing distributions from different groups when all of the distributions are unimodal (one hump).

Lesson 6: Face Off

Writing (and saying) precise comparisons between groups in which variability is present based on the (a) center, (b) spread, (c) shape, and (d) unusual outcomes help to make statements in context of the data. Actual comparison statements should use terms such as "less than," "about the same as," etc.

Lesson 7: Plot Match

Boxplots are an alternative visualization of histograms or dot plots. They capture most, but not all, of the features we can see in a dotplot or histogram.

Lesson 8: How Likely Is It?

Probability is an area about which we humans have poor intuition. Probability measures a long-run proportion: 50% chance means the event happens 50% of the time *if you repeated it forever*. When we don't repeat forever, we see variability.

Lesson 9: Bias Detective

In the short-term, actual outcomes of chance experiments vary from what is 'ideal.' An ideal die has equally likely outcomes. But that does not mean we will see exactly the same number of one-dots, two-dots, etc.

Lesson 10: Marbles, Marbles…

There are two ways of sampling data that model real-life sampling situations: with and without replacement. Larger samples tend to be closer to the "true" probability.

Lesson 11: This AND/OR That

What does "A or B" mean versus "A and B" mean? These are compound events and two-way tables can be used to calculate probabilities for them.

Lesson 12: Don't Take My Stress Away!

Generating statistical questions is the first step in a Participatory Sensing campaign. Research and observations help create applicable campaign questions.

Lesson 13: The Horror Movie Shuffle

We can "shuffle" data based on categorical variables. The statistic we use is the difference in proportions. The distribution we form by shuffling represents what happens if chance were the only factor at play. If the actual observed difference in proportions is near the center of this shuffling distribution, then we would conclude that chance is a good explanation for the difference. But if it is extreme (in the tails or off the charts), then we should conclude that chance is NOT to blame. Sometimes, the apparent difference between groups is caused by chance.

Lesson 14: The Titanic Shuffle

We can also "shuffle" data based on numerical variables. The statistic we use is the difference in means. The distribution we form by this form of shuffling still represents what happens if chance were the only factor at play. When differences are small, we suspect that they might be due to chance. When differences are big, we suspect they might be 'real.'

Lesson 15: Tangible Data Merging

We can enhance the context of a statistical problem by merging related data sets together. To merge data, each data set must have a "unique identifier" that tells us how to match up the lines of the data.

Lesson 16: What Is Normal?

The Normal curve, also called the Gaussian distribution and the "bell curve," is a model that describes many real-life distributions and is usually called the Normal Model.

Lesson 17: A Normal Measure of Spread

The standard deviation is another measure of spread. This is commonly used by statisticians because of its role in common models and distributions, such as the Normal Model.

Lesson 18: Shuffling with Normal

Z-scores allow us a way to measure how extreme a value is, regardless of the units of measurement. Usually, z-scores will range between -3 and +3, and so values that are at or more extreme than -3 or +3 standard deviations are considered extremely rare.

What is Your True Color?

Instructional Days: 10

Enduring Understandings

Statistics enable us to make sense of large amounts of data. Numerical summaries capture important elements of a distribution. Measures of center, also known as measures of central tendency, show the tendency of quantitative data to gather around a central value. Measures of spread, also known as measures of variability, show how much the quantitative data is spread out. Measurements of the propensity for the data to cluster on a central location and the range of variability within the data can provide insightful indicators about the data.

Engagement

Students will complete the *True Colors Personality Test* to discover the qualities and characteristics of their personality styles. Students will use the results from the personality color test to learn about subsetting data and finding measures of center and spread. The data from their personality test will be collected in a survey using the IDS UCLA App or via web browser at https://portal.idsucla.org/

Learning Objective

Statistical/Mathematical:

S-ID 2: Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID 3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-IC 6: Evaluate reports based on data.

Focus Standards for Mathematical Practice for All of Unit 2:

SMP-4: Model with mathematics.

SMP-5: Use appropriate tools strategically.

Data Science:

Understand the information that numerical summaries provide about the data. Understand that a boxplot is a graphical representation of a numerical summary.

Applied Computational Thinking Using RStudio:

- Calculate numerical summaries (mean, median, Sum of Absolute Deviations (SAD), and Mean of Absolute Deviations (MAD)).
- Create graphical representations to compare two or more data sets, including boxplots.

Real-World Connections:

We must be able to synthesize vast amounts of data into coherent, comprehensible measures. Today's media is continuously publishing articles that include statistical references. Critical consumerism requires that we understand the information provided in summaries of data.

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.

Data File or Data Collection Method

Data Collection Method:

1. **True Colors Personality Test**: Students will complete the *Personality Color* survey that will collect their data about their personality styles.

Data Files:

1. Students' *Personality Color* survey data *(colors)*

Legend for Activity Icons

 Video clip

 Discussion

 Articles/Reading

 Assessments

 Class Scribes

Lesson 1: What is Your True Color?

Objective:

Students will collect data that might tell them about their personality type, and will understand how to subset their data.

Materials:

1. *True Colors Personality Test* (LMR_2.1_True Colors Personality Test)
2. Posted signs for each Personality Color: Blue, Gold, Green, and Orange
   Advanced preparation required (see step 5 below)
3. Poster paper
4. Markers
5. Data collection devices

Vocabulary:

subsets

Essential Concepts: Students will understand that the 'typical' value is a value that can represent the entire group, even though we know that not all members of the group share the same value.

Lesson:

1. Ask students to consider the following questions (they do not need to record any responses):
   a. How well do you know yourself?
   b. How well do you know your classmates?
2. There are things students know and don't know about themselves. The *True Colors Personality Test* (LMR_2.1) claims to identify personality types (Later, students can gather more evidence to test these claims). Students will use these data to explore fundamental statistical concepts.

Name:_____ Date:_____

Discovering Our Personality through TRUE COLORS

(Adapted from Head Start of Greater Dallas – http://hsgd.org)

Instructions: Compare all 4 boxes in each row. Do NOT analyze each word; just get a general sense of each box. **Score each of the 4 boxes in each row from most to least as it describes you:**

4 = most, 3 = a lot, 2 = somewhat, 1 = least.

Row 1	A Active Variety Sports Opportunities Spontaneous Flexible Score	B Organized Planned Neat Parental Traditional Responsible Score	C Warm Helpful Friends Authentic Harmonious Compassionate Score	D Learning Science Quiet Versatile Inventive Competent Score
Row 2	E Curious Ideas Questions Conceptual Knowledge Problem Solver Score	F Caring People Oriented Feelings Unique Empathetic Communicative Score	G Orderly On-time Honest Stable Sensible Dependable Score	H Action Challenges Competitive Impetuous Impactful Score
Row 3	I Helpful Trustworthy Dependable Loyal Conservative Organized Score	J Kind Understanding Giving Devoted Warm Poetic Score	K Playful Quick Adventurous Confrontive Open Minded Independent Score	L Independent Exploring Competent Theoretical Why Questions Ingenious Score
Row 4	M Follow Rules Useful Save Money Concerned Procedural Cooperative Score	N Active Free Winning Daring Impulsive Risk Taker Score	O Sharing Getting Along Feelings Tender Inspirational Dramatic Score	P Thinking Solving Problems Perfectionistic Determined Complex Composed Score
Row 5	Q Puzzles Seeking Info Making Sense Philosophical Principled Rational Score	R Social Causes Easy Going Happy Endings Approachable Affectionate Sympathetic Score	S Exciting Lively Hands On Courageous Skillful On Stage Score	T Pride Tradition Do Things Right Orderly Conventional Careful Score

Total Orange Score	Total Gold Score	Total Blue Score	Total Green Score
A,H,K,N,S	B,G,I,M,T	C,F,J,O,R	D,E,L,P,Q

LMR_2.1

3. Distribute the first 2 pages of the *True Colors Personality Test* (LMR_2.1). DO NOT include the final page, which contains the descriptions of each color. Instruct students on how to complete the test, and allow time for them to complete it (see page 2 in handout).

   Note: When adding scores for each color at the bottom of the test, make sure that students have **NOT** added straight down each column.

4. Students should have a score for each of the 4 colors. Ask students to record each color and its respective score in their IDS journal. Inform them that the color with the *highest* score describes their personality. We can refer to this as their predominant color. They should record their predominant personality color in their IDS journal as well. Tell students that you will show them what each color means at the end of the lesson.

5. Post a sign for each personality color on different walls of the classroom. For example, Blue on the north wall, Gold on the east wall, etc.

6. Ask students to gather by the wall corresponding to their predominant personality color. The students should record answers to the following questions in their IDS journals.

   a. How many students are in your color group?
   b. How many students are in each of the other color groups?
   c. What is the predominant personality color in your class?

7. Then ask students to determine some common characteristics of the people in their group. Questions to help steer the discussions are included below. Each team should come up with a consensus to describe their team and will share their descriptions with the whole class. The goal is to get the students to think about "what is typical?" within their groups.

   a. What are your likes and dislikes?
   b. What things do you have in common?
   c. What are your favorite activities?
   d. What's your favorite color?
   e. Do you prefer mornings or nights?
   f. What's your favorite type of music?

8. As the groups are presenting, record some dominant characteristics on the board for each color. The students will be able to compare their perceived traits with the actual descriptions from the activity at the end of the class.

9. Next ask students in each color group to gather into two **subsets**: introvert and extrovert. Inform them that subsetting is another way to organize collected data. Create a two-way frequency table like the one below on the board to record the results.

		Color				
		Orange	Gold	Blue	Green	**Total**
Introvert/ Extrovert	Introvert					
	Extrovert					
	Total					

10. Distribute poster paper and markers to each team.

11. Inform the students that they will be creating visuals for this data by comparing subsets. The Orange and Gold groups should create visualizations that subset the color variable by introverts and extroverts, and the Blue and Green groups should create visualizations that subset introverts and extroverts by color.

12. If students are confused or stuck, have them recall the topic of two-way tables and relative frequencies from Unit 1 (Lessons 16 & 17). The Orange and Gold groups will be looking at the columns and comparing introverts/extroverts, while the Blue and Green groups will be looking at the rows and comparing colors.

13. Once all groups have completed their visuals, the Orange and Gold teams should choose one of their 2 posters to display to the class. The Blue and Green groups should do the same and select one of their visuals.

14. Display both visuals on the board and discuss their similarities and differences. Ask students to analyze and interpret the visualizations by discussing the following questions for each of the visualizations:
    a. What type of plot is this and how many variables are present? ***Answers will vary by class.***
    b. What information about this subset can I gather from this visualization? ***Answers will vary by class.***
    c. What do I see the most/least of? ***Answers will vary by class.***
    d. What is the typical personality color for this subset? Or, what is the typical group (introverts/extroverts) for this subset? ***Answers will vary by class.***
15. Ask students to summarize their impressions of the class's personality color data by writing this summary in their IDS journals.
16. Distribute the description of each personality color to students (page 3 of LMR_2.1). Remind them that the highest score is considered their predominant color and the second highest score is considered their secondary color. If there is a tie for their predominant or secondary colors, ask students to choose the color that describes their personality better.
17. Compare the given descriptions on the handout to the characteristics listed on the board for each group during step 7. Do the descriptions match what the students originally thought? How accurate are the descriptions? If time allows, ask a couple of students to share their comparisons.
18. Students will now record their data by completing the *Personality Color* campaign on the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org.

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

If not finished in class, students should complete the *Personality Color* survey either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org/

Lesson 2: What Does Mean Mean?

Objective:

Students will learn that values that gather around the center of a distribution show the typical value. This value is also referred to as the mean, or average.

Materials:

1. *Pennies on a Ruler* handout (LMR_2.2_Pennies on a Ruler)
2. Markers (1 for each table)
3. Rulers (1 for each table)
4. Pennies (6 for each table group)
5. Tape

Digital Option:

IDS Balancing Point app
Balancing Point handout (LMR_2.2b)

6. Exported, printed, and reproduced class's *Personality Color* survey data
   Advanced preparation required: The teacher must share students' data on the IDS Home page (https://portal.idsucla.org) before it can be exported and printed. Students will keep for use in subsequent lessons.
7. *Mr. Jones Mile Run Times* handout (LMR_2.3_Mr. Jones Run Times)

Vocabulary:

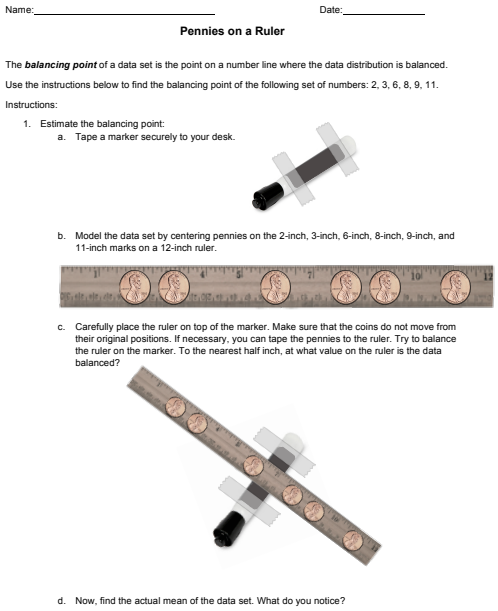
measures of central tendency (or center), typical, measures of variability (or spread), mean, average, balancing point

Essential Concepts: The center of a distribution is the 'typical' value. One way of measuring the center is with the mean, which finds the balancing point of the distribution. The mean gives us the typical value, but does not tell the whole story. We need a way to measure the variability to understand how observations might differ from the typical value.

Lesson:

1. In student pairs, ask students to discuss what they think the following terms mean:
   a. Measures of central tendency. ***A value that shows the tendency of quantitative data to gather around a central, or typical, value. Also known as measures of center. Students will learn about two such measures: the mean and the median.***
   b. Measures of variability. ***Values that show how much the quantitative data varies. Also known as measures of spread. Note: This is not taught during this lesson, but will be addressed as part of Lesson 4.***
2. Ask a pair to share what they think these two terms mean. Pairs who are listening must decide whether they agree or disagree with the pair that shared. Lead a discussion based on their statements of agreement or disagreement.
3. Communicate to the class that they will be learning more about these measures and what they tell us about data as we progress through this unit.
4. By a show of hands, ask students how many are familiar with finding the **mean**, or **average**.
5. Select a student to share his/her process for finding the mean. ***Possible answer: Add up all of the numbers. Then divide by how many numbers there are.***
6. Another way to find the mean is to find the **balancing point** of a distribution. They will learn about the balancing point via the activity in Steps 7 & 8.

7. Distribute the *Pennies on a Ruler* handout (LMR_2.2) along with a marker, ruler, tape, and 6 pennies to each table group. If you prefer to not print the document, you can project it on the board instead.



Name:_____ Date:_____

Pennies on a Ruler

The ***balancing point*** of a data set is the point on a number line where the data distribution is balanced.

Use the instructions below to find the balancing point of the following set of numbers: 2, 3, 6, 8, 9, 11.

Instructions:

1. Estimate the balancing point:
   a. Tape a marker securely to your desk.

   b. Model the data set by centering pennies on the 2-inch, 3-inch, 6-inch, 8-inch, 9-inch, and 11-inch marks on a 12-inch ruler.

   c. Carefully place the ruler on top of the marker. Make sure that the coins do not move from their original positions. If necessary, you can tape the pennies to the ruler. Try to balance the ruler on the marker. To the nearest half inch, at what value on the ruler is the data balanced?

   d. Now, find the actual mean of the data set. What do you notice?

LMR_2.2

8. Guide the students through the handout and have them share their findings throughout the activity. Be sure to emphasize the idea that the mean of a distribution can be identified by finding its balancing point.
9. Next, distribute the class's *Personality Color* survey data to the students.
10. Have student pairs find the variable ***Blue*** (whether or not that was their predominant color) in the class's printed data.
11. As a class, make a dot plot on the board to show the distribution of ***Blue*** values. Each student should come to the board and draw a dot to indicate where their value is in the distribution. Ask the students:
    a. What do you think the typical ***Blue*** score is? ***Answers will vary by class. They should be driven to an answer in the center of the distribution.***
    b. Are the data roughly symmetric? Where is the balancing point of this distribution? ***Answers will vary by class. Once a value is chosen, indicate the location on the dot plots.***
12. As a class, compute the mean ***Blue*** score for the entire class on the board and compare this value to the class's prediction of the balancing point. Students may not remember exactly how to compute the mean, so you can remind them of the general algorithm or refer them back to their responses from Step 5 above.
13. Show the students the formula for calculating the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

14. Now that they have calculated the mean for the ***Blue*** score, ask them to identify each symbol in the formula with a step in their algorithm for finding the mean, and discuss the meaning of the

symbols in the formula as a class. x_i *represents each individual data point and* n *represents the total number of observations.*

15. Indicate the location of the calculated mean on the dot plots by drawing a vertical line at the value on the x-axis. Ask student pairs to engage in a conversation about how close the mean value is to their predicted balancing point and why their prediction was made that way. Select a pair to share their discussion with the whole class.
16. Using the *Personality Color* survey data from Step 9, ask student pairs to compute the mean score for each of the other three personality colors.
17. Inform the students that, during the next lesson, they will learn about another method that can be used for measuring the center of a distribution.
18. Now, you can inform the class about an even easier method of calculating the mean – using RStudio! Explain that the command RStudio uses to calculate the mean incorporates the algorithm of summing up all the data and dividing by the total number of observations. Students will be able to use this command for quick calculations now.

    Note: If you have already *"Exported, Downloaded, Imported"* the class's *Personality Color* campaign data, you can simply use the exact command below to calculate the mean ***Blue*** score:

    ```
    > mean(~blue, data = colors)
    ```

    In general, the function can be denoted as follows:

    ```
    > mean(~variable, data = datafile)
    ```

    So, for our specific example, **`blue`** is the **`variable`** we want to find the mean value of, and **`colors`** is the **`datafile`**.

19. Have the students *Think-Pair-Share* to discuss how the mean value of a group of data could be used to easily describe complicated things. For example, instead of giving someone the entire class's ***Blue*** scores, we could just tell him/her the mean score and he/she would have a general idea about 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

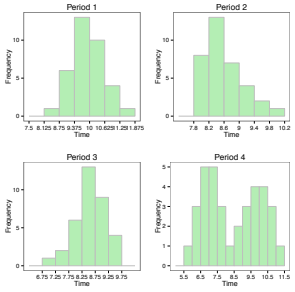
Students should complete the *Mr. Jones Mile Rule Times* handout (LMR_2.3) for homework. They can practice finding the mean of distributions by determining a balancing point for the data. Answers to the handout are below. **Note:** The mean values in part (3) do NOT need to be exact.

Name:_____ Date:_____

Mr. Jones
Mile Run Times

Background:

Mr. Jones is a physical education teacher at a high school. Every year, his students must run a distance of one mile in a specified amount of time. At the beginning of the year, he gives the students a practice one-mile run and records their times. He hopes to compare these run times to others throughout the year to see if there is any improvement in the students' running paces.



Answer the following questions:

1. What kind of plots did Mr. Jones create for his classes? _____
2. Where does each distribution balance? Find and label the balancing point on each plot above.
3. Based on the balancing points you found, what would you say is the mean mile run time for each of Mr. Jones' classes?
   i. Period 1: _____
   ii. Period 2: _____
   iii. Period 3: _____
   iv. Period 4: _____

LMR_2.3

1. What kind of plots did Mr. Jones create for his classes? ***Histograms.***
2. Where does each distribution balance? Find and label the balancing point of each distribution. ***The balancing point for all of these distributions is at the mean.***
3. Based on the balancing points you found, what would you say the mean mile run time is for each class?
   i. Period 1: ***9.91***
   ii. Period 2: ***8.48***
   iii. Period 3: ***8.45***
   iv. Period 4: ***8.17***

Lesson 3: Median in the Middle

Objective:

Students will learn that the median is another way to measure the center, or typical-ness, of a distribution, and will understand how medians compare and contrast with the mean.

Materials:

1. Sticky notes (one per student)
   Advanced preparation required (see step 6 below)
2. Poster paper
3. Graphics from *Medians – Dotplots or Histograms?* (LMR_2.4_Medians – Dotplots or Histograms)
4. *Where is the Middle?* handout (LMR_2.5_Where is the Middle)
5. Exported, printed, and reproduced class's *Personality Color* survey data

Vocabulary:

median

Essential Concepts: Another measure of center is the median, which can also be used to represent the typical value of a distribution. The median is preferred for skewed distributions or when there are outliers because it better matches what we think of as 'typical.'

Lesson:

1. Remind students that, during the previous lesson, they learned about the mean as the balancing point of a distribution and as a measure of center. In statistics, there are a few values that can be considered as measures of center – the mean is one, and another is the **median**. The median is the middle value in a group of ordered observations.
2. As a simple example, write or display the following group of numbers on the board:

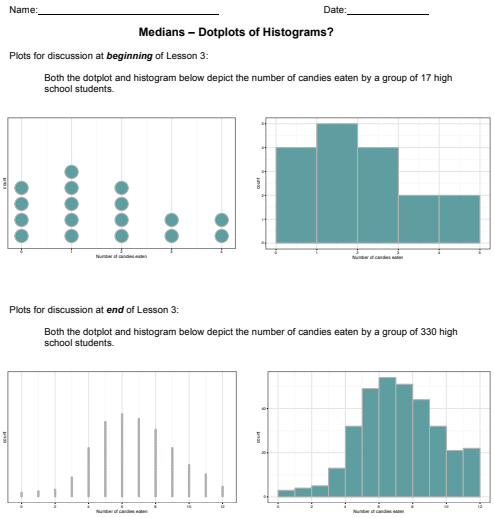
   8, 2, 6, 3, 7, 4, 9, 5, 5
3. Since there are 9 numbers in the list above, we should use the 5th number as the median because it is directly in the middle and there are 4 numbers above it, and 4 numbers below it.
4. However, students should realize that they cannot simply pick the middle number of the list as it is currently written (this would give a median value of 7). Instead, they must first arrange the numbers in numerical order (from lowest to highest).

   2, 3, 4, 5, 5, 6, 7, 8, 9
5. Now they can identify that the true median value of this list of numbers is 5.
6. Next, randomly distribute one sticky note to each student.

   Advanced preparation required: There should be one card for every student in the class. All of the cards, except one, need to have the value 0 written on them. One card should have the value 1,000,000 written on it.
7. Place poster paper on the board and have the students create a dot plot by placing their sticky notes at the corresponding values on the axis. Then, ask and record answers to the following questions:
   a. What is the typical value of these data? ***0 – all sticky notes but one have a value of 0.***
   b. Using the formula we learned in class, calculate the mean, or average, value of this distribution. ***Answer will vary by class/class size. Example: for a class with 28 students enrolled, there would be 27 values of 0 and 1 value of 1,000,000. Therefore, the mean value would be (0*27 + 1,000,000)/28 ≈ 35,714.3.***
   c. Does the mean you calculated match your understanding of "typical?" Why is the mean not capturing our notion of "typical?" ***The 1,000,000 value is heavily skewing the***

calculation of the mean. It is pulling the mean to a higher value than what we consider to be typical for these data.

8. Since we introduced the idea of the median as a measure of center at the beginning of class, have the students find the median value of the data on their sticky notes. If time permits, have them place the sticky notes in a line across the board in order (from least to greatest) and have them find the middle number. The median value will be 0.
9. Ask students why there is such a large difference between the mean and median values even though they are both measures of center? Is there a specific reason why the mean is larger than the median for this particular set of data? ***In this case, there was an outlier value that skewed the distribution and forced the balancing point to move to the right.***
10. Display the first 2 plots in the *Medians – Dotplots or Histograms?* file (LMR_2.4). They are labeled as plots for discussion for the *beginning* of class. Both the dotplot and histogram depict the number of candies eaten by a group of 17 high school students.


Name:_____ Date:_____

Medians – Dotplots of Histograms?

Plots for discussion at ***beginning*** of Lesson 3:

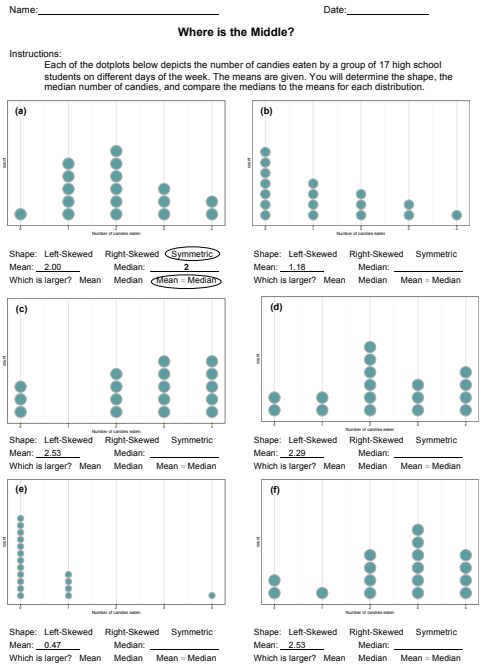
Both the dotplot and histogram below depict the number of candies eaten by a group of 17 high school students.

Plots for discussion at ***end*** of Lesson 3:

Both the dotplot and histogram below depict the number of candies eaten by a group of 330 high school students.

LMR_2.4

11. For the first 2 plots, ask students:
    a. Which plot makes it easier to find the median number of candies eaten – the dot plot or the histogram? Why? ***The dot plot is easier because we can simply find the middle dot and record the value. It is harder on the histogram, because we would have to add up amount in each bar to find the middle person.***
    b. What is the median value? ***The median number of candies eaten is 1 candy.***
12. Inform the students that they will practice finding medians of distributions using the *Where is the Middle?* handout (LMR_2.5). They will be determining medians when distributions have different shapes (e.g., symmetric, left-skewed, right-skewed).
13. Distribute the *Where is the Middle?* handout (LMR_2.5). Students should complete the handout individually first, then compare answers with their team members. Once each team has agreed upon their answers, discuss the handout as a class.



Name:_____ Date:_____

Where is the Middle?

Instructions:
Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given. You will determine the shape, the median number of candies, and compare the medians to the means for each distribution.

(a)

Shape: Left-Skewed Right-Skewed (Symmetric)
Mean: 2.00 Median: 2
Which is larger? Mean Median (Mean = Median)

(b)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 1.18 Median: _____
Which is larger? Mean Median Mean = Median

(c)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.53 Median: _____
Which is larger? Mean Median Mean = Median

(d)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.29 Median: _____
Which is larger? Mean Median Mean = Median

(e)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 0.47 Median: _____
Which is larger? Mean Median Mean = Median

(f)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.53 Median: _____
Which is larger? Mean Median Mean = Median

LMR_2.5

14. Ask the following questions to elicit a team discussion about the relationship between means and medians:

    a. What did you notice about the relationship between the mean and median values for the symmetric distributions? ***The mean and median values in the symmetric distributions - plots (a) and (d) - are fairly similar. For plot (a), the mean and median are exactly equal. For plot (d), the mean is actually larger than the median, but not by much (2.29 > 2).***
    b. What did you notice about the relationship between the mean and median values for the left-skewed distributions? ***The mean value was smaller than the median value in both of the left-skewed distributions - plots (c) and (f). Both plots had the same values for the mean (2.53) and the median (3.00) - clearly, the mean is much smaller than the median (2.53 < 3).***
    c. What did you notice about the relationship between the mean and median values for the right-skewed distributions? ***The mean value was larger than the median value in both of the right-skewed distributions - plots (b) and (e). For plot (b), the mean was only slightly higher than the median (1.18 > 1). For plot (e), the mean was a decent amount higher than the median (0.47 > 0).***

15. Steer the discussion towards the relationship between the shape of a distribution and its corresponding mean and median values.

    a. Is there a pattern that emerges between the mean and median values for differently shaped distributions? ***Yes! It seems that symmetric distributions will produce similar mean and median values, left-skewed distributions will produce smaller means and higher medians, and right-skewed distributions will produce higher means and smaller medians.***
    b. For each of the plots in the *Where is the Middle?* handout (LMR_2.5), which value better matches your idea of "typical" for that specific distribution? ***For plot (a), both the mean and median agree and appear to be the balancing point of the distribution – both match what we think is typical. For plot (b), the median seems to be more typical, but the values are very close. For plot (c), the median appears to be a more typical value. For plot (d), both the mean and median appear to be capturing our idea of***

typical. For plot (e), the median is a better match to typical. For plot (f), the median is also a better match.

16. Steer the discussion so that students recognize that the better measures of center for skewed distributions are typically medians, and the better measures for center for symmetric distributions are typically means.
17. Display the last 2 plots in the *Medians – Dotplots or Histograms* file (LMR_2.4). They are labeled as plots for discussion for the *end* of class. Both the dot plot and histogram depict the number of candies eaten by a group of 330 high school students.
18. For the last 2 plots, ask students:

    a. Which plot makes it easier to find the median number of candies eaten – the dot plot or the histogram? Why? ***The histogram is easier because we can estimate based on the distribution's shape. There are too many dots in the dot plot to find the exact middle person.***
    b. What is the median value? ***The median number of candies eaten is 7 candies.***

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 calculate the median values for each of their personality color scores. They should compare the median values to the mean values (calculated in Lesson 2) and make a decision about the possible shape of the distribution if we were to create a dot plot of the scores.

Lesson 4: How Far is it from Typical?

Objective:

Students will understand that the mean of the absolute deviations (MAD) is a way to assess the degree of variation in the data from the mean and adjusts for differences in the number of points in the data set (*n*). The MAD measures the total distance between all the data values from the mean and divides it by the number of observations in the data set.

Materials:

1. Masking tape (or painter's tape) – approximately 4-5 feet long – one for each student team
2. *How Far Apart?* handout (LMR_2.6_How Far Apart) – will be used again in Lesson 17
3. Exported, printed, and reproduced class's *Personality Color* survey data

Vocabulary:

measures of variability (spread), deviation, mean of absolute deviations (MAD)

Essential Concepts: MAD measures the variability in a sample of data - the larger the value, the greater the variability. More precisely, the MAD is the typical distance of observations from the mean. There are other measures of spread as well, notably the standard deviation and the interquartile range (IQR).

Lesson:

1. Remind students that they learned about 2 different measures of center during the previous 2 lessons: the mean and the median. Have the students recall when it is appropriate to use each value based on the shape of the distribution.
    a. Mean – use with symmetric distributions.
    b. Median – use with skewed distributions or when there are outliers.
2. Inform the students that, during today's lesson, they will learn about **measures of variability** – also known as measures of **spread**. These values show us how much the quantitative data varies from the center of a distribution. Similar to measures of center, we will use two different measures of spread: (1) the mean of absolute deviations (MAD), and (2) the interquartile range (IQR).

    Note: IQR will be discussed in detail during Lesson 5.

3. Introduce the term **deviation**. Using *Think, Pair, Share*, ask students what they think this word means and how it could relate to variability. ***A deviation is the act of departing from an established course or accepted standard. Common synonyms include departure, detour, difference, digression, divergence, fluctuation, inconsistency, modification, shift, etc.***
4. On the classroom floor next to each student team, place a 4-5 foot long piece of masking tape (or painter's tape). Then, propose the following scenario:

    Your team has been invited to guest star at the circus! You have been asked to perform as part of the tightrope act – a routine that requires tremendous focus and balance to walk across a tightly pulled rope that is suspended high in the air. In order to practice your balancing skills, the circus has provided your team with a line of tape that will represent the tightrope.

5. Have the students consider the piece of tape (aka the rope) to be the "typical" path they must take to finish the circus act. Since they do not want to fall from the suspended tightrope while performing at the actual circus, they will need to practice walking directly on the middle of the line at all times. If they *deviate* from the line, they will no longer be walking the "typical" path, and will likely fall.
6. Each team should select one student to be their starting performer.

7. In teams of 4, one student is the performer, two are measuring the distance of the deviation (one on each side of the tape), and one is the recorder.
8. Place a ruler perpendicular to the "rope" and measure the distance, in centimeters, from the path to the center of the back of their heel as the student walks and attempts to balance across the "rope."
9. The performer will walk the tightrope by looking straight up to the sky – first they look to place a foot on the line, then walk naturally while looking up to the sky, and repeating one step at a time for 4 steps, measuring after each step. Any time the performer missteps, this is considered a variation from the typical value. *You can have students take turns so everyone gets a chance to balance, walk, and to measure, depending on time in your class.*
10. Now that the students have an idea about what it means to deviate from something they consider "typical," they can start looking at distributions to see how data points vary from their typical value.
11. Inform students that they were observing deviations from typical while calculating actual differences between the rope and the performer's steps. When data are quantified with numbers, we can then calculate how far away each value is from the center.
12. One such calculation that is popular among data scientists is the mean of absolute deviations (MAD). Ask students to consider the components of the MAD in math terms, and brainstorm what the MAD value might represent.

    mean – an average

    absolute – in mathematics, we talk about absolute value, the positive difference between 2 numerical values

    deviation – as discussed earlier in the lesson, deviation represents how much things vary

13. Using the 3 components in Step 12, explain that the MAD measures the absolute distance of each data point from the mean, and then finds the average of all those distances.
14. Display the formula for the MAD distribution for the whole class to see.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

15. Discuss what each symbol in the formula means and how we use it to perform the calculation. x_i ***represents each individual data point,*** $\bar{x}$ ***represents the mean value, and*** n ***represents the total number of observations. The*** $\sum$ ***symbol represents the summation – this tells us to add up all the absolute distances from each point to the mean.***
16. To practice using this formula with actual data, students will calculate and compare the MAD values for 2 distributions.
17. Distribute the *How Far Apart?* handout (LMR_2.6), which contains 2 of the dot plots - plots (a) and (c) from the *Where is the Middle?* handout (LMR_2.5) used in Lesson 3. As before, the dot plots depict the number of candies eaten by a group of 17 high school students on different days of the week. The means are also given.

Name:_____ Date:_____

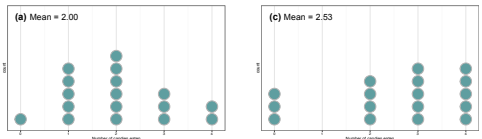
How Far Apart?

Instructions:

Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given.

Note: the plots are labeled (a) and (c) to correspond with the plots on the *Where is the Middle?* handout (LMR_2.5).

Answer questions (i) – (iii) below.



Shape: Left-Skewed Right-Skewed Symmetric

Shape: Left-Skewed Right-Skewed Symmetric

i. Determine the shape of each distribution by circling the corresponding option below the dotplot.

ii. Without doing any calculations, just by looking at the distributions, which one do you think will have a larger MAD value? Why?

iii. Calculate the MAD for each distribution by using the formula. Space has been provided to show your work on the following page.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

LMR_2.6

The calculations for each plot are shown below for the teacher's reference.

MAD for plot (a)

$$\begin{aligned} MAD &= \frac{1\,|0-2| + 5\,|1-2| + 6\,|2-2| + 3\,|3-2| + 2\,|4-2|}{17} \\ &= \frac{1(2) + 5(1) + 6(0) + 3(1) + 2(2)}{17} \\ &= \frac{2+5+0+3+4}{17} \\ &= \frac{14}{17} \\ &\approx 0.8235 \end{aligned}$$

MAD for plot (c)

$$MAD = \frac{3|0-2.53| + 0|1-2.53| + 4|2-2.53| + 5|3-2.53| + 5|4-2.53|}{17}$$

$$= \frac{3(2.53) + 0(1.53) + 4(0.53) + 5(0.47) + 5(1.47)}{17}$$

$$= \frac{7.59 + 0 + 2.12 + 2.35 + 7.35}{17}$$

$$= \frac{19.41}{17}$$

$$\approx 1.1418$$

18. Students may work in pairs to complete the handout. After all student pairs have come to an agreement on their answers, pose the following questions to the class as a whole:

    a. Which MAD value did you think would be larger based only on the look/shape of the distributions? Why? ***Since plot (c) is skewed to the left, it probably has a larger MAD because more points will be further away from the mean than in plot (a).***
    b. Which MAD value was actually larger when you calculated it? ***The MAD value for plot (c) was larger (1.1418 > 0.8253).***
    c. Did your prediction match the actual calculated values, or were you surprised by the results? ***Yes. The distribution with the wider spread (more variability) had the larger MAD value.***

19. To continue exploring with the class's Personality Color survey data, student teams should calculate the MAD value for their ***Blue*** scores. Does the MAD value seem reasonable based on the dot plot they created during Lesson 2?

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 calculate the MAD values for each of the other 3 personality color scores and compare the values of the 4 color scores.

LAB 2A: All About Distributions

Complete Lab 2A prior to Lesson 5.

Lab 2A - All About Distributions

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

In the beginning...

- Most of the labs thus far have covered how to visualize, summarize, and manipulate data.
  - We used visualizations to explore how your class spends their time.
  - We also learned how to clean data to prepare it for analyzing.
- Starting with this lab, we'll learn to use R to answer statistical questions that can be answered by calculating the mean, median and MAD.

How to talk about data

- When we make plots of our data, we usually want to know:
- Where is the *bulk* of the data?
- Where is the data more *sparse*, or *thin*?
- What values are *typical*?
- How much does the data *vary*?
- To answer these questions, we want to look at the *distribution* of our data.
  - We describe *distributions* by talking about where the *center* of the data are, how *spread* out the data are, and what sort of *shape* the data has.

Let's begin!

- *Export*, *upload* and *import* your class' *Personality Color* data.
  - Name your data `colors` when you load it.
- Before analyzing a new data set, it's often helpful to get familiar with it. So:
  - **Write down the `names` of the 4 variables that contain the point-totals, or *scores*, for each personality color.**
  - **Write down the names of the variables that tell us an observation's *introvert/extrovert designation* and whether they participated in playing *sports*.**
  - **How many variables are in the data set?**
  - **How many observations are in the data set?**

Estimating centers

- Create a `dotPlot` of the scores for your *predominant color*.
  - Pro-tip: If the `dotPlot` comes out looking wonky, include the `nint` and `cex` options.
- Based on your `dotPlot`:
  - **Which values came up the most frequently? About how many people in your class had a score similar to yours?**
  - **What, would you say, was a *typical* score for a person in your class for your predominant color? How does your own score for this color compare?**

Means and medians

- *Means* and *medians* are usually good ways to describe the *typical* value of our data.
- Fill in the blank to calculate the `mean` value of your predominant color score:

```
mean(~____, data = colors)
```

- **Use a similar line of code to calculate the `median` value of *your* predominant color.**
  - **Are the `mean` and `median` roughly the same? If not, use the `dotPlot` you made in the last slide to describe why.**

Estimating Spread

- Now that we know how to describe our data's *typical* value we might also like to describe how closely the rest of the data are to this *typical* value.
  - We often refer to this as the **variability** of the data.
  - Variability is seen in a `histogram` or `dotPlot` as the horizontal *spread*.
- Re-create a `dotPlot` of the scores for your predominant color and then run the code below filling in the blank with the name of your predominant color.

```
add_line(vline = mean(~____, data = colors))
```

- **Look at the spread of the scores from the mean score then complete the sentence below:**

  Data points in my plot will usually fall within ____ units of the center.

Mean Absolute Deviation

- The **mean absolute deviation** finds how far away, on average, the data are from the mean.
  - We often write *mean absolute deviation* as *MAD*.
- Calculate the MAD of your *predominant color* by filling in the blanks:

```
MAD(~_____, data = colors)
```

- **How close was your estimate of the spread for your predominant color (from the previous slide) to the actual value?**

Comparing introverts/extroverts

- Do introverts and extroverts differ in their typical scores for your predominant color?
  - Answer this investigative question using a dotPlot and numerical summaries.
- Make a `dotPlot` of your predominant color again; but this time, facet the plot by the introvert/extrovert variable. Include the layout option to stack the plots as well as the `nint` and `cex` options.
- **Describe the shape of the distribution of scores for the extroverts. Do the same for the introverts.**
- Using similar syntax to how you facet plots, calculate either the `mean` or `median` to describe the center of your predominant color for introverts and extroverts.
- Do introverts and extroverts differ in their typical scores for your predominant color?
- Based on the MAD, which group (introverts or extroverts) has more variability for your predominant color's scores?

On your own

- Do introverts and extroverts in your class differ in their color scores?
  - **Perform an analysis that produces *numerical summaries* and *graphs*.**
  - **Then, write a few sentences that address this statistical question and considers the *shape*, *center* and *spread* of the distributions of the graphs you create.**

Lesson 5: Human Boxplots

Objective:

Students will learn how and when to use boxplots to compare groups of data. They will learn how to compute and interpret another measure of spread: the IQR.

Materials:

1. Poster paper, 3-4 feet long
   Advanced preparation required (see Step 9 below)
2. Tape
3. Poster paper
4. Markers
5. *Ages of Oscar Winners* handout (LMR_2.7_Oscar Ages)

Vocabulary:

boxplot, quartiles, first quartile (Q_1), third quartile (Q_3), quantiles, minimum, maximum, five-number summary, range, interquartile range (IQR)

Essential Concepts: A common statistical question is "How does this group compare to that group?" This is a hard question to answer when the groups have lots of variability. One approach is to compare the centers, spreads, and shapes of the distributions. Boxplots are a useful way of comparing distributions from different groups when all of the distributions are unimodal (one hump).

Lesson:

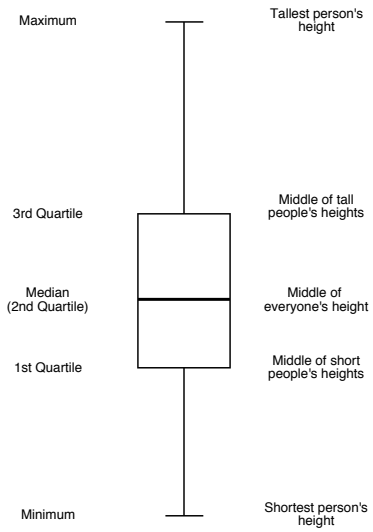
1. Remind students that we have been using the following numerical and graphical summaries to look at data:
   a. Measures of center – mean, median
   b. Measures of spread – MAD
   c. Graphing – dotPlots, histograms
2. Explain that all of these tools help us describe data to someone who may not actually be viewing it. Today, we will explore another way to summarize and describe data to others with the use of another type of statistical plot that involves breaking data up into distinct pieces: a **boxplot**.
3. For the next activity, students will need to carry their IDS journals and a pen with them.
4. Instruct students to stand up and move their chairs away from the longest wall in the classroom. Ask them to line up against the wall (in no particular order).

   Note: If there isn't enough room for everyone to line up together inside the classroom, you may do this activity outside along a building wall.
5. Say, "I want to know which person represents the typical height of students in our class. Can I tell by looking at the line as it currently stands? How would I be able to tell?" Students should discuss with a partner.
6. Ask students to share their discussions. Call on students to contribute to what has been shared if needed. Guide students to see that organizing the data (in other words, themselves) can give you a visual for their heights. Then tell them to line up in height order from shortest to tallest along the wall.
7. Once students are arranged (and this may take a little time—allow students to develop their own algorithm for finding the ordering), ask them how they might be able to describe their distribution of heights. ***Possible answers include: mean, median, MAD.***
8. Ask them to split themselves into two groups, one half that is taller and one half that is shorter, and have them decide which student represents the class's median height.

9. Have the median student stand next to the wall directly in front of the poster paper.

   Advanced preparation required: Before class begins, tape a piece of poster paper, approximately 3-4 feet long, vertically to a wall in the classroom. The students will be creating a plot using lines drawn at certain students' heights.

10. Draw a horizontal line on the poster paper to mark the location of the median by having the actual student stand in front of the poster paper so you can mark his/her exact height. Be sure to label this point as the median and include the student's actual height, in inches.
11. Next, ask the two halves to split again, so there are now four groups of students.
12. The breaks between each group are called **quartiles** because they break the data into four groups (*quartile* comes from the Latin word *quartus*, which is also the root of the Spanish word *cuatro*). The lower break represents the **first quartile** (because 25% of the class is shorter than this student's height), and the upper break represents the **third quartile** (because 75% of the class is shorter than this height). Another term that can be used in place of percentiles is **quantiles** because they represent the *quantity* of data that is lower than that value.
13. Using the student who represents the first quartile, draw another horizontal line on the poster paper marking his/her height. The student should stand in the same spot as the student who represented the median so that the line for this student is drawn underneath the median line. Be sure to label this point as the first quartile (or $\mathbf{Q_1}$) and include the student's actual height, in inches.
14. Using the student who represents the third quartile, draw another horizontal line on the poster paper marking his/her height. The student should stand in the same spot as the student who represented the median so that the line for this student is drawn above the median line. Be sure to label this point as the third quartile (or $\mathbf{Q_3}$) and include the student's actual height, in inches.
15. Finally, ask the tallest and shortest student to stand in front of the poster paper and draw horizontal lines at their heights. The shortest person represents the **minimum** height of the students in the class, and the tallest person represents the **maximum** height. Be sure to label the points as the minimum and maximum, and include the students' actual heights, in inches.
16. When you finish, you should have five lines, which represent the **five-number summary**: minimum, first quartile, median, third quartile, and maximum. Draw a box using the first and third quartiles as the edges of the box. The median line will be contained within the box. Extend a line from the first quartile down to the minimum and extend a line from the third quartile up to the maximum. Your class's boxplot should look similar to the following:



17. Students should now be facing the newly created boxplot. Allow students time to sketch the boxplot in their IDS journals, with the appropriate labels.

18. Ask students:
    a. What is the difference between the largest and smallest heights? Is there a large difference between the tallest and shortest person? ***Students should calculate maximum – minimum.*** Inform students that this difference is known as the **range** of the data set.
    b. What is the difference between the quartiles Q_1 and Q_3? What percent of our class falls within these two values? ***Students should calculate $Q_3 - Q_1$. 50% of the class falls between these two height values.*** Inform students that this difference is known as the **interquartile range (or IQR)**.

19. Remind students that they learned about one measure of spread (the MAD) during the previous lesson, and tell them that we now have another measure of spread – the IQR. Pose the following questions to the students:
    a. What does it mean when the IQR is small? ***The middle 50% of heights are close to each other.***
    b. What does it mean when the IQR is large? ***The middle 50% of heights are more spread out.***

20. Finally, subset the class into introverts and extroverts. Ask each group of students (the introverts and extroverts) to create a boxplot of their group's heights on a piece of poster paper using the techniques they just learned as a class.

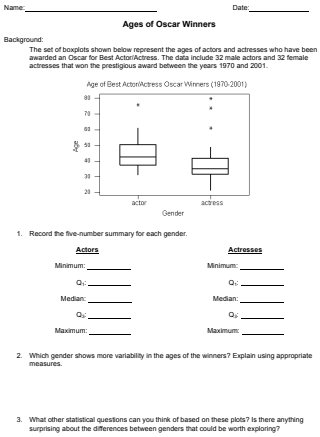
21. Ask each group to share their boxplot with the class. Lead a discussion about the similarities and differences between the plots, and be sure to include how they compare to the overall combined boxplot of heights they created earlier. In the discussion, have the students calculate the IQR for both plots and make a comparison by asking: What does the IQR tell us about each group? ***Answers will vary by 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

Students should complete the *Ages of Oscar Winners* handout (LMR_2.7) for homework using their newly acquired knowledge of boxplots.



Name:_____ Date:_____

Ages of Oscar Winners

Background:
The set of boxplots shown below represent the ages of actors and actresses who have been awarded an Oscar for Best Actor/Actress. The data include 32 male actors and 32 female actresses that won the prestigious award between the years 1970 and 2001.

Age of Best Actor/Actress Oscar Winners (1970-2001)

1. Record the five-number summary for each gender.

Actors		Actresses	
Minimum:	_____	Minimum:	_____
Q_1:	_____	Q_1:	_____
Median:	_____	Median:	_____
Q_3:	_____	Q_3:	_____
Maximum:	_____	Maximum:	_____

2. Which gender shows more variability in the ages of the winners? Explain using appropriate measures.

3. What other statistical questions can you think of based on these plots? Is there anything surprising about the differences between genders that could be worth exploring?

LMR_2.7

Lesson 6: Face Off

Objective:

Students will informally compare two or more distributions using their knowledge of shape, center, and spread to answer statistical questions. They will learn how to find the difference between two means and two medians using a histogram or dotplot.

Materials:

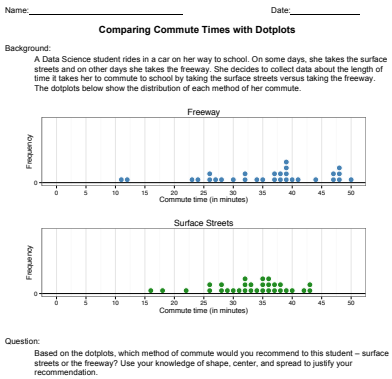
1. *Comparing Commute Times with Dotplots* handout (LMR_2.8_Commute Times – Dotplots)
2. *Comparing Exam Scores with Histograms* handout (LMR_2.9_Exam Scores – Histograms)
3. Timer
4. *Comparing Fuel Efficiency with Boxplots* handout (LMR_2.10_Fuel Efficiency – Boxplots)

Vocabulary:
rebuttal

Essential Concepts: Writing (and saying) precise comparisons between groups in which variability is present based on the (a) center, (b) spread, (c) shape, and (d) unusual outcomes help to make statements in context of the data. Actual comparison statements should use terms such as "less than," "about the same as," etc.

Lesson:

1. Poll students about the method of transportation they use for their daily school commute. How many of them walk, ride in a car, take the bus, ride a bike, etc.? Record their responses on the board. Ask them to estimate the typical amount of time it takes for them to get to school, in minutes.
2. Inform students that they have learned important features of distributions that will allow them to make decisions when working with data. More specifically, they will be able to use their knowledge of measures of center and measures of spread to compare 2 distributions in order to make a decision.
3. In teams, have students complete the *Comparing Commute Times with Dotplots* handout (LMR_2.8). Allow students time to read the "Background" portion of the handout, and then discuss what statistical question(s) the student in the scenario is trying to answer.



Name:_____ Date:_____

Comparing Commute Times with Dotplots

Background:

A Data Science student rides in a car on her way to school. On some days, she takes the surface streets and on other days she takes the freeway. She decides to collect data about the length of time it takes her to commute to school by taking the surface streets versus taking the freeway. The dotplots below show the distribution of each method of her commute.

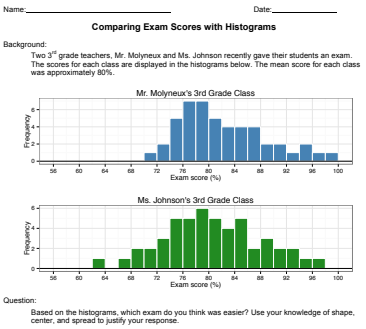
Question:

Based on the dotplots, which method of commute would you recommend to this student – surface streets or the freeway? Use your knowledge of shape, center, and spread to justify your recommendation.

LMR_2.8

4. Once teams decide on their recommendation, engage half of the class in an *Active Debate*. Half of the students will stand in a debate line and the other half will "fishbowl" the debate. Roles will reverse later in the lesson (see Step14).
5. Of those students standing on the debate line, half will argue the reasons why they recommend street travel and the other half will argue the reasons why they recommend freeway travel.

6. On the debate line, each student will stand face to face with a student who has the opposite recommendation. In other words, a student who recommends street travel will stand facing a student who recommends freeway travel.
7. Using a timer, allow one minute for students who recommend freeway travel to argue their point to the person they are facing. Then, repeat for students who recommend street travel. Students should not interrupt or respond; they should only listen to the other side.
8. Next, give debaters two minutes to prepare a **rebuttal** of the other person's argument. For example, if one student claimed that freeway travel is better, the other student may ask where the evidence is in the data or show that the data does not support the claim.
9. Allow each debater two minutes to present his/her rebuttal.
10. Finally, ask debaters if any of them changed their recommendations after engaging in the debate.
11. In teams, have students complete the *Comparing Exam Scores with Histograms* handout (LMR_2.9). Allow students time to read the "Background" portion of the handout, and then discuss what statistical question(s) the student in the scenario is trying to answer.


Name:_____ Date:_____

Comparing Exam Scores with Histograms

Background:

Two 3rd grade teachers, Mr. Molyneux and Ms. Johnson recently gave their students an exam. The scores for each class are displayed in the histograms below. The mean score for each class was approximately 80%.

Question:

Based on the histograms, which exam do you think was easier? Use your knowledge of shape, center, and spread to justify your response.

LMR_2.9

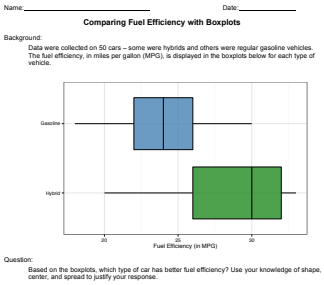
12. Repeat debate process (Steps 4 - 10) with the other half of the class.
13. Summarize the lesson by conducting a class discussion about what to look for when comparing distributions. Students should be precise when estimating values of means, medians, MAD, and IQR. They should also be able to comment on when it is most appropriate to use each measure of center and spread. ***If a distribution is symmetric, it is best to use the mean as a measure of center and the MAD as a measure of spread. If a distribution is skewed, or has outliers, it is best to use the median as a measure of center and the IQR as a measure of spread.***

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

Similar to the activities they did during class today, for homework, students should complete the *Comparing Fuel Efficiency with Boxplots* handout (LMR_2.10).


Name:_____ Date:_____

Comparing Fuel Efficiency with Boxplots

Background:

Data were collected on 50 cars – some were hybrids and others were regular gasoline vehicles. The fuel efficiency, in miles per gallon (MPG), is displayed in the boxplots below for each type of vehicle.

Question:

Based on the boxplots, which type of car has better fuel efficiency? Use your knowledge of shape, center, and spread to justify your response.

LMR_2.10

Lesson 7: Plot Match

Objective:

Students will learn how to create a boxplot from an already-established dotplot.

Materials:

1. *From Dotplots to Boxplots* handout (LMR_2.11_Dotplots to Boxplots)
2. Sets of plots from *Plot Match* file (LMR_2.12_Plot Match) – one for each team
   Advanced preparation required (see Step 7 below)

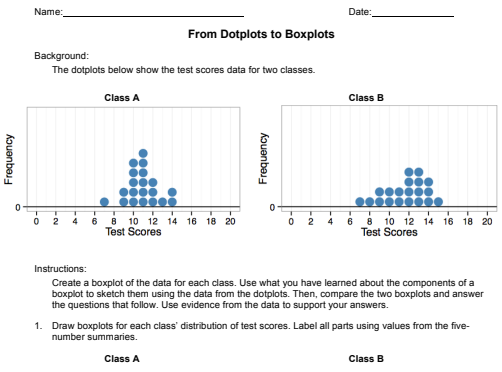
Vocabulary:

representation

Essential Concepts: Boxplots are an alternative visualization of histograms or dotplots. They capture most, but not all, of the features we can see in a dotplot or histogram.

Lesson:

1. Ask students to complete an *Entrance Slip* by recalling the components of the five-number summary that make up a boxplot. ***Five-number summary: minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3), maximum.***
2. Randomly select students to share the components and briefly discuss what each means in a boxplot. If students are missing a component, ask them to add the component to their list.
3. Remind students that during Lesson 5, they created a boxplot from students' heights.
4. Explain that a boxplot is one **representation** of the distribution of a variable in a data set. They have worked with other representations of distributions. Ask students:
   a. What other representations of distributions have we seen? ***Answers may include: dotPlots, bar charts, scatterplots, histograms, and tables.***
5. Distribute the *From Dotplots to Boxplots* handout (LMR_2.11). In teams, students will sketch boxplots from dotplots. They will need to determine the five-number summaries of each plot, and should clearly label each value on their boxplots.



Name:_____ Date:_____

From Dotplots to Boxplots

Background:

The dotplots below show the test scores data for two classes.

Class A

Class B

Instructions:

Create a boxplot of the data for each class. Use what you have learned about the components of a boxplot to sketch them using the data from the dotplots. Then, compare the two boxplots and answer the questions that follow. Use evidence from the data to support your answers.

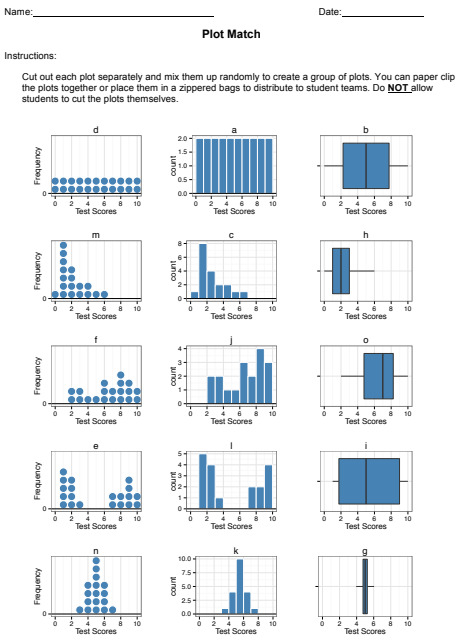
1. Draw boxplots for each class' distribution of test scores. Label all parts using values from the five-number summaries.

Class A **Class B**

LMR_2.11

6. Students should answer the 3 questions included in the handout. They can discuss their answers in pairs, and then have a class share out of the responses.
7. Once the discussion wraps up, inform the students that they will now attempt to find plots that represent the same data but are plotted differently.
8. Distribute one set of plots, from the *Plot Match* file (LMR_2.12), to each student team.

Advanced preparation required: Each student team will receive a set of plots containing all 15 plots from the *Plot Match* file (LMR_2.12). Copies will need to be cut and sorted prior to class time. To keep the plots together, you can either paper clip them or place them in zippered bags. **Note:** Do not distribute the handout for students to cut out the plots!


Name:_____ Date:_____

Plot Match

Instructions:

Cut out each plot separately and mix them up randomly to create a group of plots. You can paper clip the plots together or place them in a zippered bags to distribute to student teams. Do **NOT** allow students to cut the plots themselves.

LMR_2.12

9. Inform students that they are now going to gather in their teams and practice matching different representations of distributions. Each group will receive 15 plots (5 dotPlots, 5 histograms, and 5 boxplots). Their task is to determine which dotplots, histograms, and boxplots represent the same data.
10. Once each group has decided upon their 5 groupings, engage the students in a class share out until all students agree. Then, have the students record their responses to the following statements and/or questions in their IDS journals:
    a. What types of data are best for using a histogram? ***Histograms are useful for almost any type of data. They can easily show the shape of a distribution (including skewness and multiple peaks). They are usually best with larger data sets.***
    b. What types of data are best for using a dotplot? ***Dotplots can also easily show the shape of a distribution. They are preferred over histograms when there is a relatively small amount of data.***
    c. What types of data are best for using a boxplot? ***Boxplots are useful when the distribution has one mode (one peak). They are also useful to describe data that are heavily skewed or that contain outliers.***
    d. Describe some characteristics of data that become hidden when a boxplot is used instead of a dotplot or histogram. ***Dotplots and histograms can show the number of modes in a distribution, but a boxplot cannot. If a distribution is bimodal, we will not be able to tell in a boxplot. In general, we lose the ability to talk about the overall shape of the distribution.***
11. Display the uncut version of the *Plot Match* file (LMR_2.12) so that students see the letters that correspond to each set of representations.

    Solution key:
    Set 1: Plots (d), (a), (b)
    Set 2: Plots (m), don't, (h)
    Set 3: Plots (f), (j), (o)

Set 4: Plots (e), (l), (i)
Set 5: Plots (n), (k), (g)

12. Have a few students share out their responses. For homework, students will record some pros and cons of using different types of graphical representations to display the same data.

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 reflect on today's class discussion and record their ideas of some pros and cons of using different types of graphical representations to display the same data.

LAB 2B: Oh the Summaries…

Complete Lab 2B prior to the Practicum.

Lab 2B - Oh the Summaries ...

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

Just the beginning

- Means, medians, and MAD are just a few examples of *numerical summaries*.
- In this lab, we will learn how to calculate and interpret additional summaries of distributions such as: minimums, maximums, ranges, quartiles and IQRs.
  - We'll also learn how to write our first custom function!
- Start by loading your *Personality Color* data again and name it `colors`.

Extreme values

- Besides looking at *typical* values, sometimes we want to see *extreme* values, like the smallest and largest values.
  - To find these values, we can use the `min`, `max` or `range` functions. These functions use a similar syntax as the `mean` function.
- **Find the `min` value and `max` value for your predominant color.**
- **Apply the `range` function to your predominant color and describe the output.**
  - The *range* of a variable is the difference between available smallest and largest value.
  - Notice, however, that our `range` function calculates the maximum and minimum values for a variable, but not the difference between them.
  - Later in this lab you will create a custom `Range` function that will calculate the difference.

Quartiles (Q1 & Q3)

- The *median* of our data is the value that splits our data in half.
  - Half of our data is smaller than the *median*, half is larger.
- *Q1* and *Q3* are similar.
  - 25% of our data is smaller than *Q1*, 75% are larger.
- Fill in the blanks to compute the value of *Q1* for your predominant color.

```
quantile(~____, data = ____, p = 0.25)
```

- **Use a similar line of code to calculate *Q3*, which is the value that's larger than 75% of our data.**

The Inter-Quartile-Range (IQR)

- Make a `dotPlot` of your *predominant* color's scores.
- Visually (Don't worry about being super-precise):
  - Cut the distribution into quarters so the *number* of *data points* is equal for each piece. (Each piece should contain 25% of the data.)
    - Hint: You might consider using the `add_line(vline = )` to add vertical lines to the quarter marks.
  - **Write down the numbers that split the data up into these 4 pieces.**
  - **How long is the interval of the middle two pieces?**
  - This length is the *IQR*.

Calculating the IQR

- The IQR is another way to describe *spread*.
  - It describes how *wide* or *narrow* the middle 50% of our data are.
- Just like we used the `min` and `max` to compute the `range`, we can also use the *1st* and *3rd* quartiles to compute the *IQR*.
- **Use the values of *Q1* and *Q3* you calculated previously and find the *IQR* by hand.**
  - **Then use the iqr() function to calculate it for you.**
- **Which personality color score has the widest spread according to the *IQR*? Which is narrowest?**

Boxplots

- By using the medians, quartiles, and min/max, we can construct a new single variable plot called the *box and whisker* plot, often shortened to just *boxplot*.
- **By showing someone a `dotPlot`, how would you teach them to make a *boxplot*? Write out your explanation in a series of steps for the person to use.**
  - **Use the steps you write to create a sketch of a *boxplot* for your predominant color's scores in your journal.**
  - **Then use the `bwplot` function to create a *boxplot* using `R`.**

Our favorite summaries

- In the past two labs, we've learned how to calculate numerous *numerical summaries*.
  - Computing lots of different summaries can be tedious.
- Fill in the blanks below to compute some of our *favorite* summaries for your predominant color all at once.

```
favstats(~____, data=colors)
```

Calculating a range value

- We saw in the previous slide that the `range` function calculates the maximum and minimum values for a variable, but not the difference between them.
- We could calculate this difference in two steps:
  - Step 1: Use the `range` function to `assign` the max and min values of a variable the name `values`. This will store the output from the `range` function in the *environment* pane.

    ```
    values <- range(____, data = ____)
    ```
  - Step 2: Use the `diff` function to calculate the difference of `values`. The input for the `diff` function needs to be a vector containing two numeric values.

    ```
    diff(values)
    ```
- **Use these two steps to calculate the *range* of your predominant color.**

Introducing custom functions

- Calculating the *range* of many variables can be tedious if we have to keep performing the same two steps over and over.
  - We can combine these two steps into one by writing our own custom `function`.
- Custom functions can be used to combine a task that would normally take many steps to compute and simplify them into one.

- The next slide shows an example of how we can create a custom function called `mm_diff` to calculate the absolute difference between the `mean` and `median` value of a `variable` in our data.

Example function

```
mm_diff <- function(variable, data) {
mean_val <- mean(variable, data = data)
med_val <- median(variable, data = data)
abs(mean_val - med_val)
}
```

- The function takes two *generic* arguments: `variable` and `data`
- It then follows the steps between the curly braces { }
  - Each of the *generic* arguments is used inside the `mean` and `median` functions.
- Copy and paste the code above into an *R script* and *run* it.
- The `mm_diff` function will appear in your Environment pane.

Using `mm_diff()`

- After running the code used to create the function, we can use it just like we would any other numerical summary.
  - In the *console*, fill in the blanks below to calculate the absolute difference between the `mean` and `median` values of your predominant color:

```
____(~____, data = ____)
```

- **Which of the four colors has the largest absolute difference between the `mean` and `median` values?**
  - **By examining a `dotPlot` for this personality color, make an argument why either the `mean` or `median` would be the better description of the *center* of the data.**

Our first function

- Using the previous example as a guide, create a function called Range (*Note the capital 'R'*) that calculates the *range* of a variable by filling in the blanks below:

```
____ <- function (____, ____) {
values <- range(____, data = ____)
diff(___)
}
```

- **Use the `Range` function to find the personality color with the largest difference between the `max` and `min` values.**

On your own

- **Create a function called myIQR that uses the `quantile` function to compute the middle 30% of the data.**

Practicum: The Summaries

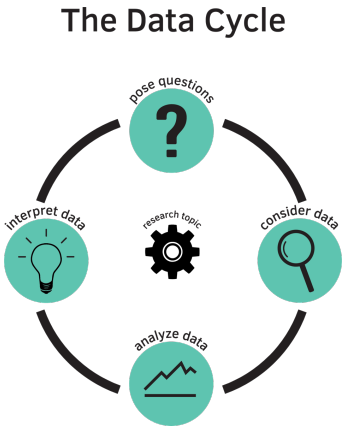
Objective:

Students will engage in their first statistical investigation using the Data Cycle. They will pose a statistical question based on a data set from a Participatory Sensing campaign, analyze, and interpret the data.

Materials:

1. *The Summaries Practicum* (LMR_U2_Practicum_The Summaries)

Note to Teacher: Before assigning the practicum to your students, engage the class in a discussion about the sample statistical questions below. Guide the discussion so that students identify not only the groups being compared in each question, but also what is being compared about the groups. Remind them of the Data Cycle from Unit 1.



Practicum
The Summaries

Using the *Food Habits* campaign data or *Personality Color* survey data, develop a new statistical question that compares two or more groups. Some sample statistical questions (about other data sets) are below:

- Which gender shows a bigger range in age, male or female Oscar winners?
  Grouping variable: gender (male, female)
  Variable: ages
- Do children, teenagers, or adults spend more money on candy?
  Grouping variable: age group (child, teenager, adult)
  Variable: the amount of money spent on candy
- How does the median height of teenage males compare to that of females?
  Grouping variable: gender (male, female)
  Variable: height
- How do the average temperatures of Los Angeles, Las Vegas, and San Francisco compare?
  Grouping variable: city (Los Angeles, Las Vegas, San Francisco)
  Variable: daily maximum temperature

Remember, a statistical question is one that anticipates variability in the question and then addresses the variability in the answer:

Based on the data you chose (*Food Habits* or *Personality Color)*, you need to:

1. Write down your question and think about ways you could answer it using RStudio.
2. Describe the data you are using to answer your question and explain why it is appropriate.
3. Analyze the data to provide evidence that supports the answer to your question. Include plot(s) and numerical summaries (mean, median, MAD, IQR, etc.) related to your plots.

4. Interpret the data to answer your statistical question. You should:
    a. Provide the plot(s) and numerical summaries related to your plot(s).
    b. Describe what the plot shows.
    c. Explain why you chose to make that particular plot(s) and the related numerical summaries.
    d. Explain how the plot and numerical summary answer your statistical question.
5. 5. Write and submit a one-page report.

Note: You may use the scoring guide in Unit 1 to give you an idea of how to score the Practicum.

How Likely Is It?

Instructional Days: 7

Enduring Understandings

Probability measures the long run frequency of occurrence for chance outcomes. Probabilities can be approximated by designing and conducting simulations, and also via mathematical calculation.

Engagement

Students will watch a scene from the movie *Rosencrantz and Guildenstern are Dead* and discuss the likelihood of getting "heads" when tossing a coin 78 times in a row. The scene can be found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc

Learning Objectives

Statistical/Mathematical:

S-CP 2: Understand that two events A and B are independent if the probability of A and B occurring together is the product of their probabilities, and use this characterization to determine if they are independent.

S-CP 9: (+) Use permutations to perform [informal] inference.
*This standard will be addressed in the context of data science.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand how algorithms are used to design simulations.

Applied Computational Thinking using RStudio:

- Design and conduct simulations in RStudio.
- Compare actual data to simulated data using RStudio.
- Re-randomization of permuted data.
- Use estimated probabilities from samples to determine theoretical probabilities

Real-World Connections:

Learn to use simulations to determine expectations of events.

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.

Data File or Data Collection Method

Simulated data in RStudio.

Legend for Activity Icons

 Video clip

 Discussion

 Articles/Reading

 Assessments

 Class Scribes

Lesson 8: How Likely Is It?

Objective:

Students will understand the basic rules of probability. They will learn that previous outcomes do not give information about future outcomes if the events are independent.

Materials:

1. Video: "Heads" from the movie *Rosencrantz and Guildenstern are Dead* found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc
2. Projector for RStudio functions

Vocabulary:

probability, simulation, model, sample proportion, chance, independence

Essential Concepts: Probability is an area that we humans have poor intuition about. Probability measures a long-run proportion: 50% chance means the event happens 50% of the time *if you repeated it forever*. When we don't repeat forever, we see variability.

Lesson:

1. Ask students to consider possible synonyms to the word **chance**. If someone says, "That just happened by chance," what does that mean? ***Synonyms: possibility, prospect, expectation, unintentional, unplanned. The actual definition of chance is "a possibility of something happening."***
2. Then, ask them which game – chess or the board game, "Sorry" – is more based on chance? Why?

Note: any game can be chosen. ***"Sorry" is more based on chance because many outcomes are determined by dice rolls. In chess, there are certain strategies and movements that can be planned, so it is more a game of skill. With Sorry, the players roll a die (number cube), so the numbers they roll have an impact on how well they do in the game.***

3. Next ask students if they can think of situations where chance is the only force at play. ***Possible responses: card games, slot machines, the lottery, coin flipping, and rock-paper-scissors.***
4. Play the "Heads" video from the movie *Rosencrantz and Guildenstern are Dead* found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc.
5. In their IDS journals, ask students to write down their initial reactions to the clip by responding to the following questions:
    a. Is it *possible* to get 78 heads in a row when tossing a coin? ***Yes, it is possible to get 78 heads in a row since one coin toss does not determine the next coin toss.***
    b. Do you think it is *likely* to get 78 heads in a row? ***No, although it is possible to get 78 heads in a row.***
    c. How many times should we get heads when tossing a coin? ***1 out of 2 times or 50% of the time.***
    d. On average, how many times out of the 78 tosses should the characters have gotten heads? ***Roughly about 39 times.***
6. Ask students to discuss their findings with their team members and come to an agreement on their responses. Afterwards, conduct a *Whip Around* and ask each team to share its findings. Are there any differences between the teams? Any similarities?
7. As teams share their responses, students should add to or revise their individual findings in their IDS journals.

8. Explain to students that, from the concept of chance, we can start learning about **probability**. Chance is simply the possibility that something will happen, and probability is a measurement for how often something happens in the "long run." Students may have ideas about how to calculate probabilities based on prior classes or knowledge, but inform them that IDS will be taking a different approach by using simulations (see next step).

9. Since we don't want to actually flip a coin 78 times like the actors did in the video, we can have RStudio simulate them for us. A **simulation** is a way of creating random events that are close to real-life situations without actually doing them. It is a kind of **model**, which is a way of representing real world situations so that predictions can be made.

10. Explain to students that R has a function that does coin flipping for us, and that it assumes an equal probability of heads and tails. Using a projector to display your computer screen to the whole class, demonstrate how to do one simulation of a coin flip in RStudio. Use the following function:

11. `rflip(1)`Explain that the value of 1 in the argument part of the function tells R to flip the coin 1 time. If we want to flip the coin 10 times, we could simply change the function to **`rflip(10)`**.

12. Run the function again using 10 as the number of times to flip the coin. Ask students:
    a. How many heads ("H"s) were there? ***Answers will vary for each sample.***
    b. How many Tails ("T"s) were there? ***Answers will vary for each sample.***
    c. In the output, what does **`Flipping 10 coins [Prob(Heads) = 0.5]`** mean? ***this is RStudio telling us that we are tossing the coin 10 times and that the* probability *of getting heads should be 0.5 (it is flipping a fair coin).***
    d. In the output, what does **`Number of Heads: 3 [Proportion Heads: 0.3]`** mean? **Note:** This is example of an output. Your sample may have a different value for the number of heads that appeared, and thus a different value for the proportion of heads. ***this is RStudio telling us that in our sample, we got heads 3 out of the 10 times we flipped the coin. The* sample proportion *is automatically calculated for us by dividing the number of heads by the total number of tosses (in this case, 3/10 = 0.3).***

13. To relate back to the video at the beginning of class, repeat the simulation once more, but use 78 as the number of coin flips **`rflip(78)`**. Ask students:
    a. How many heads ("H"s) were there? Since we know to expect about 39 heads if the coin is fair, does the value seem reasonable? ***Answers will vary for each sample. Most likely, you will see values near 39.***
    b. How many Tails ("T"s) were there? ***Answers will vary for each sample.***
    c. What proportion of the coin flips were heads? ***Answers will vary for each sample.***

14. Using the **`rflip(78)`** command, run the simulation 3–5 more times and have students record the values for the number of heads and the proportion of heads.

    As an example, we ran the function 3 times and saw the following values:

    Sample 1 – amount of heads: 45
    proportion of heads: 0.577

    Sample 2 – amount of heads: 33
    proportion of heads: 0.423

    Sample 3 – amount of heads: 42
    proportion of heads: 0.538

15. Have students answer the questions listed below. The important thing to note is that the values can and (almost always) WILL change each time you run the simulation to create a new sample.
    a. How do the proportions of heads in the samples compare to each other? ***Answers will vary.***
    b. How do the proportions of heads compare to the true probability of heads (1/2 or 50%)? ***Answers will vary, but students should notice that most of the probabilities are close to 50%.***

c. Why is there a 50% chance of getting heads during each coin flip? ***Since there are two sides to a coin, both should be equally likely to come up. So there is a 1 out of 2 chance of getting heads and 1 out of 2 chance of getting tails.***

16. Ask students to engage in a discussion with their group about the statement below, then have a few group reporters share out.
    a. If a coin was flipped 78 times, I would claim that the coin is unfair if I got less than **#** heads or more than **#** heads.
17. Inform students that you are going to perform 500 simulations. Each simulation represents a coin being flipped 78 times. For each simulation, the computer will record the number of heads in the 78 flips. A histogram will be created that represents the number of heads in each of the simulations. The histogram is a model that will display what typically happens when a fair coin is flipped 78 times.
18. Copy and paste the code below in an RScript and run each line of code, one at a time, for the students:

```
set.seed(11) #reproducibility
flips <- do(500)*rflip(78)
View(flips) # 4 variables
histogram(~heads, data = flips)
favstats(~heads, data = flips)
```

19. Engage the students in a discussion about the histogram:
    a. What is the distribution telling us? ***When flipping a fair coin 78 times, what typically happened was that it landed on heads between 36 and 40 times (36/78 = 0.46 to 40/78 = 0.51). It was not uncommon for the coin to land on heads 31-35 (0.40-0.45) times or 41-45 (0.52-0.58) times. Even landing on heads between 46-50 (0.59-0.64) times was not too uncommon. What was very uncommon, however, was landing on heads less than 30 times (less than 38%) or more than 51 times (more than 65%).***
    b. Were the group's cut-offs (item #16) similar to what the chance model displayed? ***Answers will vary. Some groups' intervals might be very wide and others very narrow.***
    c. Use the chance model (histogram) which displays what typically happens when a fair coin is flipped 78 times, make a call for the scenarios below – fair or unfair?
        i. You flip a coin 78 times and get 37 heads. ***Fair. 37 was very common based on the histogram.***
        ii. You flip a coin 78 times and get 46 heads. ***Fair. 46 was less common but not too uncommon.***
        iii. You flip a coin 78 times and get 20 heads. ***Unfair. In the 500 simulations, not once did we see a FAIR coin land on heads 20 times.***

20. Next, pose the following question:
    a. If you get a heads on the first toss of a coin, will you definitely get a heads on the next toss? Will you definitely get a tails on the next toss? ***No. One coin toss should not affect another coin toss. Each time you flip the coin, the chances of getting heads versus tails remains the same.***
21. Introduce the concept of **independence**. Explain that, when tossing a fair coin, there is no relationship between each toss. The second toss does NOT depend on the first toss; therefore, the coin tosses are independent of each other.

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 create a Tweet (they do not have to post it online). Using 280 characters or fewer, write a Tweet about the meaning of probability.

Lesson 9: Bias Detective

Objective:

Students will learn how to use simulations to detect biased probability.

Materials:

1. Poster paper – with 6 columns labeled 1, 2, 3, 4, 5, 6
2. 2 dice (number cubes)
   Note: You can use regular hard dice, or soft foam dice (can be found at dollar stores)
3. Projector for RStudio functions

Vocabulary:

bias

Essential Concepts: In the short-term, actual outcomes of chance experiments vary from what is 'ideal.' An ideal die has equally likely outcomes. But that does not mean we will see exactly the same number of one-dots, two-dots, etc.

Lesson:

1. In pairs, ask students to quickly share their Tweets from the previous lesson's homework. Collect the Tweets and select a few to share with the class. Of the Tweets shared, ask students which one is closest to the definition: probability measures how often something happens in the "long run."
2. Remind students that, during the previous lesson, they were introduced to simulations. The progression following this path: chance ➔ probability ➔ simulations. The motivation for using simulations is that we can use the calculated sample proportions to estimate probabilities of real-life events.
3. During today's lesson, we will be continuing to learn about probability and simulations to determine if an event is not fair (one example: a coin is weighted and lands on heads more often than tails).
4. Ask students what they know about dice (number cubes). If they have never heard of them, show one to the class and explain how it works. ***A die (number cube) is a 6-sided cube. Each face of the cube is labeled with dots to represent a number between 1 and 6. For example, if the face has 3 dots, then it represents the number 3. The cube itself is weighted so that there is an equal probability of rolling each of the 6 numbers.***
5. Have a discussion about what the students would expect the probability of rolling the number 1 should be if a die (number cube) were tossed into the air and allowed to fall back to the ground (or table). ***Since there are 6 numbers on the die, each number should be equally likely to occur, so the probability of rolling a 1 is 1/6.***
6. Display a piece of poster paper on the board with columns labeled 1, 2, 3, 4, 5, 6. Explain that each column represents the numbers on the die (number cube). We will be using poster to tally the results of actual dice (number cube) rolls.
7. Select two students to be dice (number cube) rollers and give each student one die. As noted in the *Materials* section above, you can have the students use either regular hard dice, or softer foam ones (can be found at dollar stores).
8. Tell the class that each student roller will be rolling the dice 6 times (so there will be a total of 12 rolls for our sample). Ask:
   a. If they are rolling the dice 6 times, how often do you think Student 1 will roll a 3? Would you expect it to be the same for Student 2? ***Out of 6 rolls, we would expect to see each of the numbers one time, so we will most likely see about one 3 for Student 1.***

b. Would you expect the Student 2 to roll a 3 just as often? Why? *Yes, we should expect the same thing from Student 2 because we have independent events. There are actually two ways that independence plays a part here: (1) each student is independent from the other and has no effect on what the other will roll, (2) the 6 die rolls for Student 1 are all independent of each other because each face of the cube has an equal chance of happening on any given roll. So, if Student 1 gets a 3 during his/her first roll, that doesn't give us any information about what he/she will get on the second roll.*

c. Since we will have 12 rolls (and therefore 12 samples), how many tally marks should we expect in each column on the chart? *We would expect to see 2 tally marks in each column (each number will probably be rolled twice).*

9. Have each student roller toss his/her die one time and share the outcomes with the rest of the class. As they do this, place a tally mark in the corresponding column on the chart. Repeat this process 5 more times so that each student has a total of 6 rolls.

10. As a class, observe the results in the chart and discuss the following:

a. Do the data from these 12 rolls match what we expected (see responses from Step 8)? Is this surprising? *Answers will vary by class. Some values may have shown up more than we expected (example: the number 3 was rolled 3 times), and others may not have been rolled at all (example: the number 5 was never an outcome). We only have a small sample of data, so it's not surprising for our results to vary from the expected outcomes.*

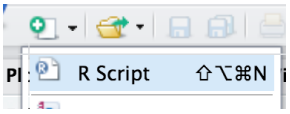
b. If the data do not match our expectations, does this mean the dice are unfair in some way? *Even if they don't match our expectations, this does not mean the dice are unfair – we simply don't have enough data yet to know. We would need to roll the dice more.*

c. If we wanted to purposely create an unfair die, what are some ways we could achieve that? *Answers will vary by class. Some examples include: (1) We could add tape to one face of the die to give that side more weight. This would increase the chances of the number that is directly opposite of it appearing because the die will land on the heavier side more (and therefore the side facing up will be the number opposite). (2) We could chip the edge of one corner of the die. This would throw off the original balance and favor certain sides.*

11. Similar to the previous day's lesson with coin flipping, we can also simulate dice rolls in RStudio. The function required is called **`roll_die()`**. The arguments for this function are a bit different than the **`rflip()`** function from yesterday. We cannot simply put **`roll_die(1)`** for the computer to roll a die one time. Instead, the function was built with 2 possible dice to choose from – die A and die B.

12. Inform the students that one of the dice in the function is fair and the other has been created with **bias**. Bias is the act of favoring one outcome over another. They will attempt to determine which dice is the biased one by doing multiple simulations.

    Note to Teacher: Many simulations require multiple functions, or code, to perform. This is where RScripts are helpful. An RScript can be used to test code, write notes, and let us easily execute multiple lines of code at one time. This would be a good place to introduce students to RScripts.



13. Using a projector to display your computer screen to the whole class, demonstrate how to open an RScript. Type the following function on your script and click Run. Run simply pastes the function onto the console.

14. `roll_die("A", times = 1)`The output will show one number that represents what value on the die the computer rolled. Go back to your script and modify the function to roll die A 12 Times.

```
roll_die("A", times = 12)
```

15. Compare the results of these 12 simulated rolls to the results of the 12 actual rolls completed by the two students during Step 9. If there is space available on the tally chart, you can add the computer results to it for an easy comparison.

16. Ask students how we could record data from these simulations if we wanted to roll the die 100 times. Would they want to hand count the number of times each value occurred? Is there a function in RStudio that will count them for us? ***It would be difficult to count every individual value in the output on the screen. However, we can use the*** `tally()` ***function to find out how many times each die value appeared.***

17. To make using the **`tally()`** function easier, we should assign a name to each simulation so we don't have to type the entire function multiple times. We can also have it calculate the proportions for each value. Add the functions below to your RScript and run them one at a time.

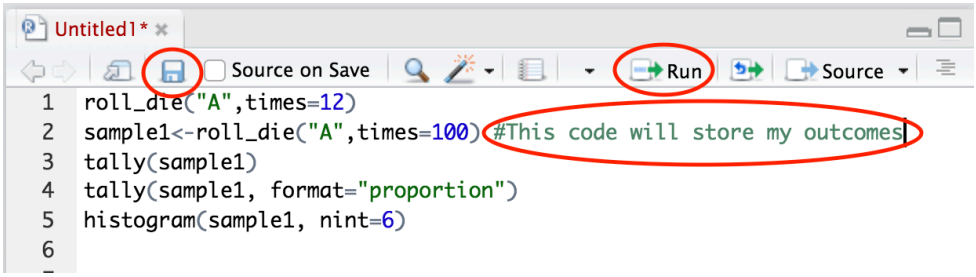
```
sample1 <- roll_die("A", times = 100)
tally(sample1)
tally(sample1, format = "proportion")
```

18. Remind the students that if the die is fair, then each side of the die should appear roughly the same amount of times. Therefore, the proportions should be fairly similar to each other and to the true probability of 1/6.

19. Add the function below to your script, but before running it, ask the students what they think a histogram of the simulated data might look like and then run the command on your screen. Note: Be sure to include the argument **`nint = 6`** so that the resulting histogram has six bars. ***If the die is fair, each of the bars in the histogram should be roughly the same height.***

```
histogram(sample1, nint=6)
```

Note to teacher: Show students how to save an RScript. Inform students that they can take notes on their RScript by including a hashtag (also known as a pound sign or #) at the beginning of the note. Data scientists refer to these types of notes as "code comments" or simply "comments". See image below.



The Script will be stored in the files tab. To run each function individually, place your cursor on the line and hit the Run button. To run multiple lines of code at once, highlight them and hit Run.

20. Allow the students to access their school computers now to start creating their own simulations in an RScript using die B. Students can pair up, if needed. Have them begin by asking RStudio to roll the die 100 times. They should note their output from both the **`tally()`** function and the histogram. They can then compare the results to those from Step 14. Are they similar? Can they determine which die is unfair yet? ***Answers will vary by class. The results will be similar, but not exact. With the sample sizes of each simulation being fairly small, we cannot see a clear difference between the two dice yet.***

21. Let the students explore by changing the number of times RStudio rolls the dice. Remind them that the goal is to determine which of the two dice is biased. The sample sizes need to be very large in order for them to see a clear difference between the 2 histograms. The pattern becomes more visible when **`times = 2000`**.

Note: The maximum value for `times` within the `roll_die()` command is 500. Simulations can be combined using the concatenate function `c()`. For example, suppose s1 represents 500 rolls

of die A and s2 is a second sample of 500 rolls for die A. To combine these two samples the following can be used `more_rolls <- c(s1, s2)`

22. When students have had enough time to make a decision regarding which dice is biased and how, engage the class in a discussion to verify that everyone agrees. ***Die B is biased; Die A is fair. Die B favors the number 3.***
23. Then, steer the conversation towards why the sample size affected the results. ***The sample size needed to be large because the difference between the probabilities of the die rolls was very small. In order to detect small differences, we must have larger sample sizes.***

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 consider a four-sided die and imagine rolling it 20 times. They should sketch a histogram of (a) the ideal, expected outcome, (b) an outcome that they think is "realistic," and (c) an outcome they might see if the die were biased to produce more 4's.

Lesson 10: Marbles, Marbles…

Objective:

Students will understand that random events vary, so that the percentage predicted by a probability isn't exact, but varies. Students practice converting percentages to proportions.

Materials:

1. For each student team: 50 marbles – 20 of one color, and 30 of another color
   Note: Marbles can be substituted for other materials, such as small blocks, as long as they are the same size.

Vocabulary:

proportion, percentage, event, with replacement, without replacement

Essential Concepts: There are two ways of sampling data that model real-life sampling situations: with and without replacement. Larger samples tend to be closer to the "true" probability.

Lesson:

1. Remind students that, during the previous two lessons, they learned how to estimate probabilities for a population with the help of simulations to create sample data. Both lessons had nice, prepackaged functions already available in RStudio, which made the simulations fairly quick and easy – in Lesson 8, the **`rflip()`** function was used to simulate flipping a coin; and in Lesson 9, the **`roll_die()`** function was used to simulate rolling one of two dice.
2. But what if we don't have a nice function to perform a simulation for us? Can we create our own method? Yes! We will actually learn to create a simulation from scratch during Lab 2C.
3. Ask students:

   a. If you have a bag of 50 marbles, where 20 of them are blue and 30 of them are red, what is the probability of drawing a red marble? ***30/50 or 60%.***
4. Select a student to answer the question. Ask the class if they agree or disagree. If they agree, ask them to raise their hand. If there are students who disagree, lead a class discussion until a consensus is reached.
5. Ask students to share their strategies on how to convert the **proportion** into a **percentage**. As strategies are being shared, students should take notes in their IDS journals. Review how to turn fractions into percentages, if necessary.
6. Ask students:

   a. What if we actually drew out one marble, recorded its color, then replaced it back in the bag, and did this 10 times? ***Answers will vary by class.***
   b. Would the percentage of red marbles in this sample necessarily be exactly the same as the probability? Identify that each time a marble is drawn, we are creating an **event**. ***Answers will vary by class.***
7. Distribute the bags of marbles to each team. Ask each team to:

   a. Shake the bag of marbles.
   b. Draw one marble out of the bag.
   c. Record the marble's color in their IDS journal.
   d. Replace the marble back into the bag. Inform them that this is called sampling **with replacement**. Ask them to consider what "with replacement" means and discuss with the class. ***"With replacement" means that after you select a marble from the bag, you have to put it back into the bag (replace it) before you select another marble.***

   They should draw 10 marbles from the bag and record the observed colors.

8. Do a *Whip Around* to find out how many times each team drew a red marble out of their 10 draws. Have them calculate the corresponding sample proportions. For example, if one team drew 7 red marbles out of their 10 draws, their sample proportion is 7/10 = 0.70 (which is the sample as a sample percentage of 70%).
9. Ask students why the proportions are perhaps different from each other and from the "true" probability of drawing a red marble?
10. Have the student teams continue drawing marbles, one at a time and with replacement, until they have 50 events recorded. Discuss the following questions:

    a. How many times did they draw a red marble out of these 50? ***Answers will vary by class.***
    b. What's the corresponding sample proportion? Is it closer to the true probability than when you only drew 10 marbles? ***Answers will vary by class. But, they should notice that, as the sample size increases, the sample proportion gets closer to the true population proportion.***

11. Engage students in a discussion about how the sample size affects the sample proportion. They should see that as they draw more marbles, their sample probability gets closer and closer to the true probability. If we were to continue drawing marbles forever, in the long run, our sample proportion should equal our true probability.
12. Have students consider what it might mean to sample **without replacement**. How would they do that with their bag of marbles? ***"Without replacement" means that after you select a marble from the bag, you never put it back into the bag (do not replace it). Instead, you simply select another marble from the bag immediately. Students should recognize that, by not replacing the marble each time, the probabilities will change. This means each draw from the marble bag is NOT independent from another draw because removing one marble impacts the next event.***
13. *Exit Slip*: Based on this lesson, ask students to describe a sample, an event, and replacement.

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 2C: Which Song Plays Next?

Complete Lab 2C prior to Lesson 11.

Lab 2C – Which Song Plays Next?

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

A new direction

- For the past two labs, we've looked at ways that we can summarize data with numbers.
  - Specifically, you learned how to describe the *center*, *shape* and *spread* of variables in our data.
- In this lab, we're going to *estimate the probability* that a rap song will be chosen from a playlist with both rap and rock songs, if the choice is made at random.
  - The playlist we'll work with has 100 songs: 39 are rap and 61 are rock.

Estimate what ... ?

- To *estimate the probability*, we're going to imagine that we select a song at random, write down its genre (*rock* or *rap*), put the song back in the playlist, and repeat 499 more times for a total of 500 times.
- The statistical question we want to address is: *On average, what proportion of our selections will be rap?*
- **Why do we *put a song back* each time we make a selection?**
- **What would happen in our little experiment if we did not do this?**

Calculating probabilities

- Remember that a *probability* is the long-run proportion of time an event occurs.
  - Many probabilities can be answered exactly with just a little math.
  - The probability we draw a single rap song from our playlist of 39 rap and 61 rock songs is `39/100`, `0.39` or `39%`.
- Probabilities can also be answered exactly if we were willing to randomly select a song from the playlist, write down its *genre*, place the song back in the list, and repeatedly do this *forever*.
  - Literally, *forever…*
  - But we don't have that much time. So we're only going to do it 500 times which will give us an *estimate of the probability*.

Estimating probabilities

- You might ask, *Why are we estimating the probability if we know the answer is 39%?*
  - Sometimes, probabilities are too hard to calculate with simple division as we did above. In which case, we can often program a computer to run an experiment to estimate the probability.
  - We refer to these programs as *simulations*.
- The techniques you learn in this lab could be applied to very simple probability calculations or very hard and complex calculations.
  - In both cases, your *estimated* probability would be very close to the *actual* probability.

Getting ready

- Simulations are meant to mimic what happens in real-life using randomness and computers.
  - Before we can start simulating picking songs from a playlist, we need to simulate that playlist in R.

- To simulate our 39 `rap` songs, we'll use the repeat (`Rep`) function.

```
rap <- rep("rap", times = 39)
```

- Look in the `Environment` pane for the vector containing your rap songs.
- Use a similar line of code to simulate the rock songs in our playlist of 100.

Put the songs in the playlist

- Now that we've got some different songs, we need to combine them together.
  - To do this, we can use the combine function `c()` in R.
- Fill in the blanks to combine your different songs:

```
songs <- __(rap, ____)
```

- And with that, our playlist of songs should be ready to go.
  - Type `songs` into the console and hit enter to see your individual *songs*.

Pick a song, any song

- Data scientists call the act of choosing things randomly from a set, *sampling*.
  - We can randomly choose a song from our playlist by using:

```
sample(songs, size = 1, replace = TRUE)
```

- Run this code 10 times and compute the proportion of `"rap"` songs you drew from the 10.
  - Vocabulary Check: A *proportion* is a fraction of the whole.
    - For example, if 2 rap songs were drawn from the 10, the *proportion* would be 2/10
    - It is more common to express a *proportion* as a decimal, in this case, 0.20
    - It is even more common to express a *proportion* as a percentage, 20%
- **Once everyone in your class has computed their proportions, calculate the *range* of proportions (the largest proportion minus the smallest proportion) for your class and write it down.**

Now `do()` it some more

- Instead of running the same line of code multiple times ourselves we can use R to `do()` multiple repetitions for us.
  - Fill in the blanks below to do the `sample` code from the previous slide *50* times run:

```
do(___) * sample(___, ___ = ___, ___ = ___)
```

- Recall that we need to store our results to be able to perform analysis.
- *Assign* the 50 selected songs the name `draws` and then `View` your file.
- What is the variable name?
  - `R` defaulted to naming the variable based on the function used. You may use the data cleaning skills you learned in lab 6 to `rename` the variable if you wish.
- Fill in the blank below to tally how often each genre was selected:

```
tally(~___, data = draws)
```

- **Compute the proportion of `"rap"` songs for your 50 draws and find out if the *range* for your class' proportions is bigger or smaller than when we drew 10 songs.**

Proportions vs. Probability

- To review, so far in this lab we've:
  - Simulated a "playlist" of songs.
  - Repeatedly simulated drawing a song from the playlist, noting its genre and placing it back in the playlist.
  - Computed the proportion of the draws that were `"rap"`.
- These proportions are all *estimates* of the theoretical probability of choosing a rap song from a playlist.
  - As we increase the number of draws, the *range* of proportions should shrink.

When using simulations to estimate probabilities, using a large number of repeats is better because the estimates have less variability and so we can 'e confident we're closer to the actual value.

Non-rando' Randomness

- We've seen that random simulations can produce many different outcomes.
  - Some estimated probabilities in your class were smaller/larger relative to others.
- There are instances where you might like the same random events to occur for everyone.
  - We can do this by using `set.seed()`.
- For example, the output of this code will always be the same:

```
set.seed(123)
sample(songs, size = 1, replace = TRUE)

## [1] "rap"
```

Playing with seeds

- With a partner, choose a number to include in `set.seed` then redo the simulation of 50 songs.
  - Both partners should run `set.seed(___)` just before simulating the 50 draws.
  - The blank in `set.seed(___)` should be the same number for both partners.
  - Verify that both partners compute the same proportion of `"rap"` songs.
- Redo the 50 simulations one last time but have each partner choose a different number for `set.seed(___)`.
  - **Are the proportions still the same? If so, can you find two different values for `set.seed` that give different answers?**

On your own

- Suppose there are 1,200 students at your school. 400 of them went to the movies last Friday, 600 went to the park and the rest read at home.

If we select a student at random, what is the probability that this student is one of the one's who went to the movies last Friday?

- **Answer this by estimating the probability that a randomly chosen student went to the movies using 500 simulations.**
  - **Write down both the estimated probability and the code you used to compute your estimate. You might find it helpful to write your answer in an R Script *(File -> New File -> R Script)***
  - **Include `set.seed(123)` in your code before you do 500 repeated samples.**

Lesson 11: This AND/OR That

Objective:

Students will understand how AND/OR probabilities are defined and will be able to use frequency tables to compute these probabilities.

Materials:

1. *Compound Probabilities* handout (LMR_2.13_Compound Probabilities)
2. Blue sticky notes
3. Gold sticky notes
4. Four signs on the board labeled: *Pickles, Mayonnaise, Both, None* (in that exact order, and equally spaced across the length of the board)

Vocabulary:

compound probabilities

Essential Concepts: What does "A or B" mean versus "A and B" mean? These are compound events and two-way tables can be used to calculate probabilities for them.

Lesson:

1. Remind the students that they have been learning about estimating probabilities of single events based on sample proportions. Inform them that, today, they will learn how to calculate proportions when multiple events happen.
2. Review the basic idea of computing probabilities; in other words, the number of outcomes we are interested in divided by the total number of outcomes possible.
3. Pose the questions below to the class.

   Note: They do not need to come up with specific answers; this is a time for them to make suggestions.
   a. How would we compute the probability of two outcomes occurring at the same time? For example, what is the probability that a randomly chosen student likes both pickles AND mayonnaise?
   b. How would we compute the probability of either of two outcomes occurring? For example, what is the probability that a randomly chosen student likes either pickles OR mayonnaise?
4. For both questions, steer the students towards using the definition from Step 2. That is, we want the students to realize that they can count the number of people that qualify for the given circumstance and divide by the total number of people to calculate a probability.
5. In order to define AND/OR probabilities, students will participate in an activity where they are grouped by their food preferences.
6. Divide the board into 4 groups and write the words “Pickles,” “Both,” “Mayonnaise,” and “None,” in that order, from left to right.
7. Ask for 10 volunteers to stand by the word that represents their preferences. That is, if they only like pickles, they should stand by the word “Pickles.” If they like both pickles and mayonnaise, they should stand by the word “Both.”

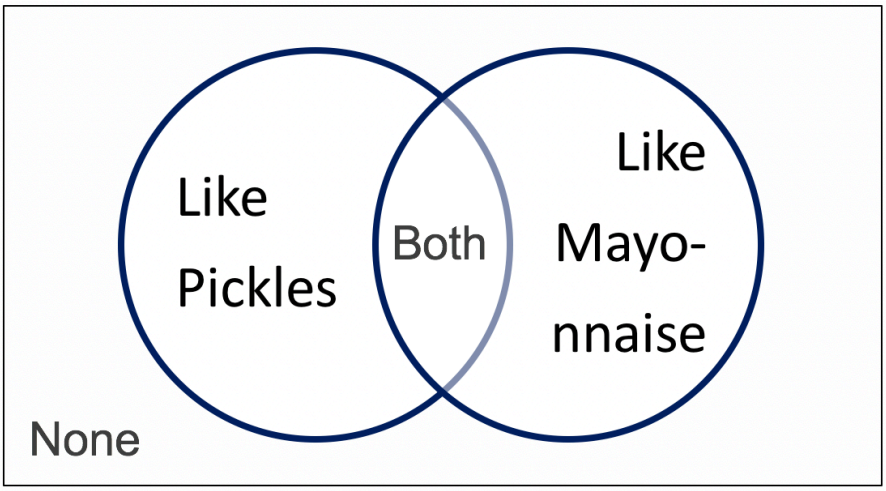
   Note: If all 4 groups do not have at least one student in them, select a few more students to stand at the board.
8. Ask the remaining students (those still seated) to count the total number of people standing by the board and have a student volunteer share the answer with the class. ***Answers will vary by class.***

9. Next, create a 2-way frequency table like the one below to organize the values of student preferences as follows:
   - Counts for students who like both go in the Yes/Like Mayonnaise and Yes/Like Pickles box.
   - Counts for students who like none go in the No/Like Mayonnaise and No/Like Pickles box
   - Counts for students who like only mayonnaise go in the Yes/Like Mayonnaise and No/Like Pickles box.
   - Counts for students who like only pickles go in the No/Like Mayonnaise and Yes/Like Pickles box.

   Note: A Venn diagram like the one below may be used as well, depending on student understanding and at teacher discretion.

Like Mayonnaise

Like Pickles	Yes	No	TOTAL
Yes			
No			
TOTAL			



10. Next, ask the students sitting down the following questions:

    a. How many students like both pickles AND mayonnaise? ***Answers will vary by class.***
    b. What is the probability that a randomly selected student at the board likes both pickles AND mayonnaise? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Both" (number given in Step 9(a)) by the number of students at the board (number given in Step 8).***

$$\frac{\text{\# students under "Both"}}{\text{\# students standing at the board}}$$

11. Now, ask a student from the audience:

    a. How many students like pickles? ***Answers will vary by class.***

       Note: Avoid phrasing the question with "Students that like ONLY pickles." Students need to see that students who like "Both" items also belong to the groups liking the individual items.

If students mistakenly report the number of students who like ONLY pickles, ask the people at the board to raise their hands if they like pickles and then ask the mistaken student to recount.

b. What is the probability that a randomly selected student at the board likes pickles? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Pickles" and "Both" by the total number of students at the board.***

$$\frac{\text{(\# students under "Pickles") + (\# Students under "Both")}}{\text{\# students standing at the board}}$$

12. Finally, ask one more student from the audience:

a. How many students like pickles OR mayonnaise? ***Answers will vary by class.***

Note: Avoid phrasing the question with "Students that like ONLY pickles OR ONLY mayonnaise."

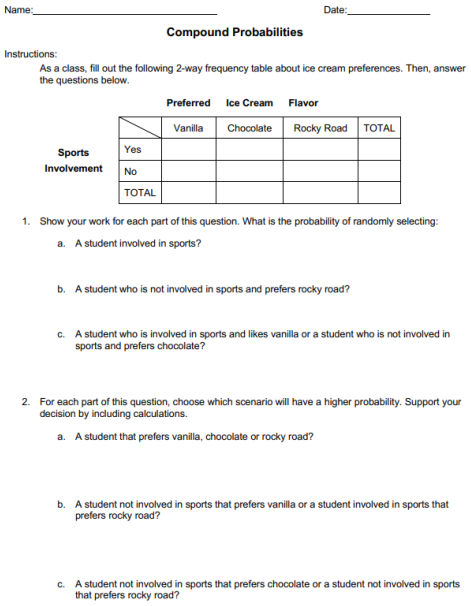
If students mistakenly report the number of students who like ONLY pickles plus the students who like ONLY mayonnaise, ask the people at the board to raise their hands if they like either pickles or mayonnaise (All students at the board should raise their hand except for the students who like "None") and then ask the mistaken student to recount.

b. What is the probability that a randomly selected student at the board likes pickles OR mayonnaise? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Pickles," "Mayonnaise," and "Both" by the total number of students at the board.***

$$\frac{\text{(\# students under "Pickles") + (\# Students under "Mayonnaise") + (\# Students under "Both")}}{\text{\# students standing at the board}}$$

13. Informs students that AND/OR probabilities are called **compound probabilities.** In teams, have students record their own definitions of AND/OR probabilities based on the activity they just completed. ***A compound probability is the probability of some combination of events occurring.***

14. Distribute the *Compound Probabilities* handout (LMR_2.13).



Name:_____ Date:_____

Compound Probabilities

Instructions:

As a class, fill out the following 2-way frequency table about ice cream preferences. Then, answer the questions below.

Preferred Ice Cream Flavor

Sports Involvement	Vanilla	Chocolate	Rocky Road	TOTAL
Yes				
No				
TOTAL				

1. Show your work for each part of this question. What is the probability of randomly selecting:
   a. A student involved in sports?
   b. A student who is not involved in sports and prefers rocky road?
   c. A student who is involved in sports and likes vanilla or a student who is not involved in sports and prefers chocolate?
2. For each part of this question, choose which scenario will have a higher probability. Support your decision by including calculations.
   a. A student that prefers vanilla, chocolate or rocky road?
   b. A student not involved in sports that prefers vanilla or a student involved in sports that prefers rocky road?
   c. A student not involved in sports that prefers chocolate or a student not involved in sports that prefers rocky road?

LMR_2.13

15. Pass out a blue sticky note to each student who plays a sport and a gold sticky note to each student who does not play a sport.
16. Draw the table from the worksheet on the board (make it large and legible).
17. Have each student who plays a sport hold up their sticky note. Count them and record the number of students who play a sport in the appropriate row of the TOTAL column in the table.
18. Have each student who does not play a sport hold up their sticky note. Count them and record the number of students who do not play a sport in the appropriate row of the TOTAL column in the table.
19. Ask each student which of the following ice cream flavors they most prefer (each student must choose exactly one option): Vanilla, Chocolate or Rocky Road.
    a. Have the students write their ice cream preference on their sticky note.
    b. Fill out the remainder of the table by asking each group of students, those who play a sport and those who do not, to hold up their preference.
    c. Make sure the totals for preferred ice cream flavors and sports involvement add up to the same number.
20. Instruct the students to work in pairs to answer the questions on the *Compound Probabilities* handout (LMR_2.13).

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

If not completed in class, students should finish the *Compound Probabilities* handout (LMR_2.13).

LAB 2D: Queue It Up!

Complete Lab 2D prior to the Practicum.

Lab 2D - Queue it up!

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

Where we left off

- In the last lab, we looked at how we can use computer simulations to compute estimates of simple probabilities.
  - Like the probability of drawing a song genre from a playlist.
- We also saw that performing *more* simulations:
  - Took *longer* to finish.
  - Had estimates that *varied less*.
- In this lab, we'll extend our simulation methods to cover situations that are more complex.
  - We'll learn how to estimate their probabilities.
  - We also look at the roll of sampling *with* or *without replacement*.

Back to songs

- In R, simulate a *playlist of songs* containing 30 `"rap"` songs, 23 `"country"` songs and 47 `"rock"` songs.
  - *Assign* the combined playlist the name `songs`.
- Simulate choosing a single song 50 times. Then use your simulated draws to estimate the probability of choosing a *rap* song.
  - The actual (theoretical) probability of choosing a *rap song* in this case is `0.30`.
  - **Write a sentence comparing your estimated probability to the actual probability.**

With or Without?

- So far, you've selected songs *with replacement*.
  - We called it that, because each time you made a selection, you started with the same playlist. That is, you chose a song, wrote down its data, and then placed it back 'n the list.
- It's also possible to select *without replacement* by setting the `replace` option in the `sample` function to `FALSE`.
- Take a sample of `size` 100 from our playlist of songs *without replacement*. Assign this sample the name `without`.
  - **Run `tally(~without)` and describe the output. Does something similar happen if you sample *with replacement*?**
    - Notice that the tilde `~` was not needed with the `tally` function. This is because `without` was not a variable within a data frame but rather a vector which acts like a lone variable.
  - **What happens if `size = 101` and `replace = FALSE`?**

Sample with? Or without?

- Imagine the following two scenarios.
  1. You have a coin with two sides: *Heads* and *Tails*. You're not sure if the coin is fair and so you want to estimate the probability of getting a *Head*.
  2. A child reaches into a candy jar with 10 *strawberry*, 50 *chocolate* and 25 *watermelon* candies. The child is able to grab three candies with their hand and you're interested in probability that all three candies will be chocolate.

- **Which of these scenarios would you sample *with replacement* and which would you sample *without replacement*? Why?**
  - **Write down the line of code you would run to `sample` from the candy jar. Assume the simulated jar is named `candies`.**

Simulations at work

- In reality, songs from a playlist are chosen without replacement.
  - This way, you won't hear the same song several times in a row.
- Let's write a more realistic simulation and estimate the probability that if we select two songs at random, without replacement, that both are rap songs.
  - Use the `do` function to perform 10 simulated `samples` of `size` 2, with replacement and *assign* the simulations the name `draws` and then `View` your file. Use `set.seed(1)`.

What are the variable names? What happened in the first simulation? Did any of your 10 simulations contain two rap songs?

Simulations and probability

- To estimate the probability from our simulations, we need to find the proportion of times that the event we're interested in occurs in the simulations.
- In other words, we need to count the number of times the desired events occurred, divided by the number of attempts we made (the number of simulations).
- The next slides will show you two ways to do this.

Counting similar outcomes

- One way we can estimate the probability of drawing two songs of the *same* genre is to use the following trick to count the number of *rap* songs in each of the 10 simulations:

```
mutate(draws, nrap = rowSums(draws=="rap"))
```

- **Let's break down the code above by running each part of the code one piece at a time. As you run each line of code below describe the output**

```
draws == "rap"
rowSums(draws == "rap")
mutate(draws, nrap = rowSums(draws=="rap"))
```

- Remember to assign a name to your mutated data set.

Counting other outcomes

- Another method we can use to estimate the probability of complex events is to use the following 2-step procedure:
  1. Subset the rows of the simulations that match our desired outcomes.
  2. Count the number of rows in the subset and divide by the number of simulations.
- The result that you obtain is an estimate of the probability that a specific combination of events 'occurred.
- We'll see an example of this method on the next slide.

Step 1: Creating a subset

- Fill in the blanks below to:

1. Create a subset of our simulations when both draws were `"rap"` songs.
2. Count the number of rows in this subset
3. And divide by the total number of repeated simulations.

```
draws_sub <- filter(draws, ___ == "rap",  ___ == "rap")

nrow(___) / ___
```

Estimating probabilities

- Answer the following questions by performing 500 simulations of sampling 2 songs from a playlist of 30 rap, 23 country and 47 rock songs:
- **Calculate estimated probabilities for the following situations:**
  1. You draw two `"rap"` songs.
  2. You draw a `"rap"` song in the first draw and a `"country"` song in the 2nd.
- **Create a histogram that displays the number of times a `"rap"` song occurred in each simulation. That is, how often were zero rap songs drawn? A single rap song? Two rap songs?**

On Your Own

- Using what you've learned in the previous two labs, answer the following question by performing two computer simulations with 500 repetitions a piece:

If we draw 5 songs from a playlist of 30 rap, 23 country and 47 rock songs, how does the estimated probability of all 5 songs being rap songs change if we draw the songs with or without replacement?

- For each simulation:
  - **Create a histogram for the number of *rap* songs that occurred for each of the 500 repetitions.**
- **Describe how the distribution of the number of *rap* songs changes depending on if we use replacement or not.**

Practicum: Win Win Win

Objective:
Students will create and combine simulations to assess probabilities.

Materials:
1. *Win Win Win Practicum* (LMR_U2_Practicum_Win Win Win)

Practicum
Win Win Win

The California lottery has a game called the *Daily 3*.

- It consists of 3 numbers between 0 - 9 that are drawn daily.
- The numbers are drawn *with replacement*
- Winners are usually awarded a couple hundred dollars.
- To win the maximum amount of money, players must correctly choose the numbers that are drawn, in order.

Based on what you learned in *Lab 2C* and *Lab 2D* (*Which Song Plays Next* and *Queue it Up!)* and using the rules of the *Daily 3*, you need to:

1. Write down the code to correctly simulate the *Daily 3* once.
2. Use your code to simulate the *Daily 3* 500 times.
3. Compute the estimated probability of getting the first 2 numbers of the *Daily 3* correct.
4. Should the estimated probability of correctly guessing the last 2 numbers of the *Daily 3* be less than, the same as, or more than guessing the first 2 numbers? Why?
5. In teams of 4:
    a. Each team member chooses 3 numbers for the *Daily 3*.
    b. Each team member simulates the *Daily 3* game 500 times.
    c. Within your group, combine the team simulation estimates to estimate the probability of winning the *Daily 3*.
6. Write and submit a one-page report. Your report should include the code.

What Are the Chances That You Are Stressing or Chilling?

Instructional Days: 8

Enduring Understandings

Permutations of data provide a model that shows us how the world behaves if chance is the only reason for differences between groups or for associations between variables. If our actual observation is a rare permutation, this suggests that chance is not a good explanation for the difference or association. On the other hand, if the actual observation is a common permutation, this suggests that chance may be a valid explanation. Differences between permuted data and actual data suggest that the chance model can be rejected and there is a dependent relationship between two variables.

Engagement

Students will read the Huffington Post article titled *Don't Take My Stress Away* to set the stage for the Stress/Chill Campaign. High school students who expected, and wanted, to feel stressed out by school wrote this article. The article is found at:

http://www.huffingtonpostdon't/jack-cahn/dont-take-my-stress-away_b_2090203.html

Learning Objectives

Statistical/Mathematical:

S-IC 2: Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that a chance model serves as an indicator of whether or not associations in the actual data are due to chance (understand why a plot might appear to have a trend, but may actually be the result of randomness). Understand that simulations provide a way of comparing expected chance outcomes to real outcomes in order to determine if a model and actual data appear consistent. Learn about merging data sets by understanding the structure of both data sets and the logic of the way they will be combined.

Applied Computational Thinking using RStudio:

- Permutations of data, determining if actual data is similar to permuted data
- Merge multiple data sets together based on a common variable
- Create permutations using a merged data set

Real-World Connections:

In media, citizens read about results and scientific studies in which treatments are applied. In real life, one can ask the question: Does this happen by chance? Understanding chance helps us interpret media reports of scientific and medical findings.

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. **Stress/Chill Participatory Sensing Campaign**: Students will monitor how they feel at different times of the day – whether they are "stressing" or "chilling." Along with how they feel, they will make observations regarding other factors, such as being alone or with others, what they are doing at that moment, and why they are doing that activity.

Data Files:

1. Students' *Personality Color* survey data (`colors`)
2. Students' *Stress/Chill* campaign data
3. Titanic data set (`titanic.rda`)
4. Horror Movie data set (`slasher.rda`)

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 12: Don't Take My Stress Away!

Objective:

Students will read the Huffington Post article titled *Don't Take My Stress Away* to spark their interest about how they spend their time, and will continue to read reports critically to look for claims that may or may not be based on data.

Materials:

1. Article: *Huffington Post's Don't Take My Stress Away* found at: http://www.huffingtondon't.com/jack-cahn/dont-take-my-stress-away_b_2090203.html
2. Data collection devices

Essential Concepts: Generating statistical questions is the first step in a Participatory Sensing campaign. Research and observations help create applicable campaign questions.

Lesson:

1. Become familiar with the *Stress/Chill Campaign Guidelines* (shown at the end of this lesson), particularly the questions, to help guide students during the campaign (see Campaign Guidelines in Teacher Resources).
2. Ask students the following questions and conduct a brief share out of their responses.
   a. Do you know anyone who seems to be always stressed or anyone who seems to be always chilled? ***Answers will vary by class.***
   b. What are some observations you have made that make that person extremely stressed or chilled? ***Answers will vary by class.***
3. Inform students that they will be learning about some high school students who view stress as a part of life in the Huffington Post article titled *Don't Take My Stress Away.*
4. Provide students the link to the article and allow time for them to read it: http://www.huffindon'tpost.com/jack-cahn/dont-take-my-stress-away_b_2090203.html
5. As they read the article, students should note whether they agree or disagree with the authors and should write down their comments and/or reactions to the article in their IDS journals.
6. Ask student pairs to share if they agree or disagree with the authors of the article and why. Conduct a *Share Out* of student responses.
7. Inform students that for this unit, we will be investigating how stressed or chilled they are at certain times of the day.
8. Students will collect data using the *Stress/Chill* Participatory Sensing campaign. They will add the *Stress/Chill* campaign to their list of available campaigns either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org
9. Ask students to complete their first survey.
10. After students have completed their first survey, use a random number generator to generate two random times a day for the next 6 days (RStudio example given below). It is recommended that you create 6 sets of random numbers so that students are polled at different times each day.

    Example for RStudio (assuming students are awake between the hours of 7:00 am and 11:00 pm):

    ```
    > sample(7:23, size = 2, replace = FALSE)
    ```

    Note: If a time falls within the school day, it is up to the discretion of the teacher to use this time or not.

11. Based on the times generated, ask students to set reminders on the IDS UCLA App for the next 6 days. Students without a mobile device may set reminders using a method available to them.
12. Focus students' attention on the *Stress/Chill* survey questions (you may display the questions on the Campaign Guidelines document). Ask students to generate three statistical questions that could be answered using the *Stress/Chill* data.
13. Then, ask them to write down in their IDS journals some predictions about what they think they will see after they collect some data.

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 6 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

Campaign Guidelines – Stress/Chill

1. **The Issue:**

   People report being more and more stressed everyday. This trend is extending beyond adults, it is also reported by children and teenagers. The amount of work for which people are responsible has been increasing. To understand what makes us feel stressed, some important questions to ask are:

   a) What factors affect my stress/chill level?
   b) Do different personality types have different things that make them happy/sad?
   c) Do you like to be alone or with people?
   d) Is your stress/chill level a function of the environment in which you are?

2. **Objectives:**

   Upon completing this campaign, students will have compared groups and gained an understanding of variability within and between groups. They will have learned how to conduct and use permutations to model variability, perform informal inference, and how to do simulations to make predictions.

3. **Survey Questions:**

   Use a random number generator to generate two random times a day for the next 6 days, including a weekend if possible. If a time falls within the school day, it is up to the discretion of the teacher to use this time or not.

Prompt	Variable	Data Type
Take a photo of what you are doing right now.	photo	photo
How stressed are you feeling right now (3 is very stressed, 0 is not stressed at all)?	level	integer
How many people are you with (not counting yourself) up to 107,282?	howmanypeople	integer
Who are you with? -alone -friends -family -friends and family -classmates -teacher -teacher and classmates -strangers	who	categorical
Where are you? -school -work -home -public place -others' house -commuting	where	categorical
Why are you here (in one word)?	why	text
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date

When? Surveys are taken two to three times per day at pre-determined randomly selected times.

How Long? About two weeks. Ideally, two of these days include a weekend.

4. Motivation:

Students must understand that they need to keep collecting data. Use the Plot App to look at the data after the first day and have a discussion.

Ask: Why were most people stressed? Guide students along the way.

Ask students to predict the following: What is your stress/chill level in the evening versus morning? Does it change everyday? How about during the weekend? What is the difference between groups?

Data collection: After the first day, use the Campaign Monitoring tool to see who has collected the most data.

5. Technical Analysis:

Students will use RStudio.

6. Guiding Questions:

a) Have students generate predictions and check up on their predictions.
b) What's the typical stress/chill level of the class across the campaign?
c) What's my typical stress/chill level and how does it compare to whole class?
d) Do the stress/chill levels vary by weekday or weekend or the type of people you are with?
e) Under which conditions is my stress/chill level affected?
f) Encourage students to generate their own questions.

7. Report:

Students will complete the Stress/Chill Practicum. They will analyze their stress/chill data using data analysis skills and RStudio skills learned in the unit.

Lesson 13: The Horror Movie Shuffle

Objective:

Students will understand that, just by chance, we will see differences between two groups. They will understand that these differences are usually small. Specifically, they will learn that we can determine if outcomes are due to chance for categorical variables by calculating differences in the proportions between two groups.

Materials:

1. 3" x 5" cards (1 per student)

Vocabulary:

chance, simulations, randomness, shuffle

Essential Concepts: We can "shuffle" data based on categorical variables. The statistic we use is the difference in proportions. The distribution we form by shuffling represents what happens if chance were the only factor at play. If the actual observed difference in proportions is near the center of this shuffling distribution, then we would conclude that chance is a good explanation for the difference. But if it is extreme (in the tails or off the charts), then we should conclude that chance is NOT to blame. Sometimes, the apparent difference between groups is caused by chance.

Lesson:

1. **Data Collection Monitoring:** Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org. Click on **Campaign Monitor** and sign in.
2. Inform students that you will be monitoring their data collection. Ask:
    a. Who has collected the most data so far? *See User List and sort by Total*.
    b. How many active users are there? How many inactive users are there? *Click on pie chart*.
    c. How many responses were submitted yesterday and today? *See Total Responses*.
    d. How many responses have been shared? How many remain private? *Click on pie chart.*
    e. Using TPS, ask students to think about what they can do to increase their data collection.
3. Conduct a discussion about the data that has been collected.
4. Have students recall what they have learned about **chance** (see Lesson 8). ***Synonyms: possibility, prospect, expectation, unintentional, unplanned. The actual definition of chance is "a possibility of something happening."***
5. To expand on the flow chart from Lesson 9 (chance ➔ probability ➔ simulations), explain that we can use **simulations** to show that sometimes, when we think two groups are different, the difference is really just because of chance, or **randomness**, and does not mean anything. This brings us back to "chance" in the flow chart.
6. Remind students that a simulation is a model for creating random outcomes. Randomness means that something just happens without a specific order.
7. In pairs, ask students to name situations where two groups could be compared, and then have the students record these situations in their IDS journals. Some examples include:

- ***Men earn more money than women for some work.***
- ***Basketball players are faster runners than baseball players.***
- ***Los Angeles students are smarter than _____.***
- ***UCLA football players are better athletes than USC football players.***
- ***You and a friend flipped a coin 10 times, and you got more "heads."***

8. Then, ask students to write next to each situation whether they think the differences are either real or due to chance because sometimes differences between two groups are real, but sometimes they might just be due to chance, and they will be learning ways to tell the difference.

9. Explain to the class that we are interested in finding out who will survive by the end of a horror movie. Ask the students:
    a. Do you think men and women have an equal likelihood of surviving by the end of a horror movie? ***Answers will vary by class.***
10. Have a few students share out their opinions along with their reasoning.
11. Inform the students that they will be pretending to be actors from horror movies during today's lesson.
12. Explain that data from horror movies (sometimes called slasher films) were collected of 485 characters from 50 films. For each character, 2 variables were recorded: Gender and Survival. The values for Gender were "Male" and "Female." The values for Survival were "Dies" and "Survives."

	Gender	
Survival	Female	Male
Dies	172	228
Survives	50	35
Total	222	263

Notice that there were more male characters than female characters and that most characters in slasher films do not survive.

13. From this data, the proportion of survivors was calculated for each gender. In other words, for all female characters, the number of female survivors was divided by the total number of females. Similarly, for all male characters, the number of male survivors was divided by the total number of males.

$$\frac{\#\text{ (“Female” \& “Survives”)}}{\#\text{ (“Female”)}} \quad \text{or} \quad \frac{\#\text{ (“Male” \& “Survives”)}}{\#\text{ (“Male”)}}$$

14. The percent of females who survived by the end of a horror movie was about **23%**, and the percent of males who survived by the end of a horror movie was about **13%**. Ask the students:
    a. Is this what you expected? (Refer back to the discussion from Step 9.) ***Answers will vary by class. If students thought males would survive more often, then these results would be unexpected. If students thought females would survive more often, then these results would be exactly what they expected. If students thought there was an equal likelihood of survival, these results would also be surprising.***
    b. What is the difference in the proportions of survival rates between genders? What does this mean in the context of surviving a horror movie? ***The difference is 23% - 13% = 10%. This means that 10% more women characters survived than men.***
    c. Is this difference "big" or "small"? How can they define what is "big" and what is "small?" ***Answers will vary by class. Upon first glance, it may seem like 10% is a big difference, but we do not know for sure.***
15. Explain that they will participate in an activity to determine if the 10% difference seen in the actual data set is big or small. This will help them determine if there really is a difference in survival rates for males versus females, or if the 10% difference was just due to chance.
16. Split the class into two groups, 46% of them on one side of the room and the other 54% on the other side of the room. Tell the smaller group they have been assigned to play female characters

in the horror movie (regardless of gender) and tell the larger group that they have been assigned to play male characters in the horror movie (regardless of their gender). Once those groups have been created, have the class calculate the number of students in each group that would have survived a horror film using the actual proportions given in Step 14.

For example: For a class of 30 students:

- ***46% of 30 (0.46x30) ≈ 14 students representing female characters.***
- ***Of those 14 female characters, 23%, or 3 (0.23x14 ≈ 3), are survivors.***
- ***The remaining 16 students (30 – 14 = 16) represent male characters.***
- ***Of those 16 male characters, 13%, or 2 (0.13x16 ≈ 2), are survivors.***

17. Each group should then decide which students will be survivors. Using the 3" x 5" cards, students should write either "dies" or "survives" on their card.

    For example (continued from above):

    Three of the females are survivors; so 3 female characters from the group should write "survives" on their cards. The rest of the group should write "dies" on their cards.

    Two of the male characters are survivors; so 2 males from the group should write "survives" on their cards. The rest of the group should write "dies" on their card.

18. Explain to students that *IF* there really is no difference between genders in horror films, then the characters who survived would only have done so by chance. In other words, males and females would have an equal likelihood of surviving. Have students discuss the following questions:

    a. How many total people in our class are survivors? What is the total proportion of survivors? ***Answers will vary by class. Using the example above, there would be a total of 5 survivors from the class of 30 students. The proportion of survivors would be 5/30 = 0.17 = 17%.***
    b. How many of the survivors would we expect to be male? How many would we expect to be female? ***Answers will vary by class. Using the example above, we would expect to see 17% of males and 17% of females survive since that was the overall proportion of survivors. So, we would expect 0.17*16 ≈ 3 male survivors, and 0.17*14 ≈ 2 female survivors.***

19. Collect all of the 3" x 5" cards from the students and explain that you are going to **shuffle** the cards and redistribute them so that their genders have no influence on whether or not they survive the horror movie.

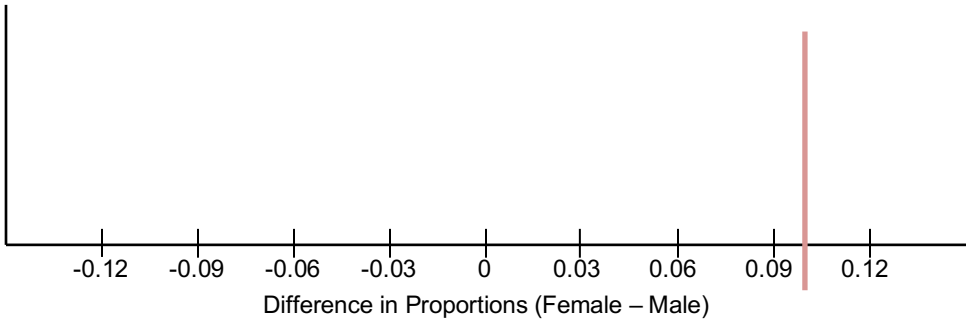
20. Visibly shuffle the survives/dies cards to create a random shuffle. Once the cards have been well-shuffled, pass them back out to the students face down. After all the cards are given out, each group should identify the number of people that are survivors and calculate the corresponding proportion of the survivors.

21. On the board, create a table to display the proportions of survivors for each gender, and include a column for the difference (female survivors – male survivors). Fill in the table with the values the students found in Step 20. **Note:** The first row has been filled in with the example data from above BEFORE the shuffles have taken place. Exact numbers were not used so that the proportions would match the actual horror movie data set.

# of Female Survivors	# of Male Survivors	Proportion of Female Survivors	Proportion of Male Survivors	Difference in Proportions (Female – Male)
3.22	2.08	3.22/14 = 0.23	2.08/16 = 0.13	0.23 – 0.13 = 0.10

22. Note that values in the "Difference in Proportions" column can be positive or negative because sometimes more women will survive, and other times more men will survive.

23. Draw a dotplot on the board labeled "Difference in Proportions." Include a vertical line at 10% to represent the actual difference in gender survival rates in real horror movies (see example below).



24. Using the information from Steps 20 and 21, place a dot at the corresponding value for the shuffled data's difference in proportions. Ask the students:
    a. How does this difference compare to the actual data set's difference of 10%? ***Answers will vary by class. Most likely, the difference in proportions will be much smaller than 10%. In fact, the difference in proportions will be centered around 0.***
25. Repeat Steps 19 – 24 a few more times (depending on how much class time you have available).
26. Ask the students to record their responses to the following questions:
    a. What was the biggest difference we saw from our shuffles? What was the smallest? ***Answers will vary by class.***
    b. What do you think this dotplot would look like if we shuffled our survival cards 1000 times? ***The dotplot would look roughly symmetric and centered around 0, meaning that if there were no relationship between a character's gender and whether or not they survive, the difference in proportions would typically be 0.***
27. Have a discussion about how the actual difference in gender survival (10%) is rarely seen when we assign "survives" or "dies" just by chance (aka when shuffling). What does this mean in terms of who will die in actual horror movies? ***Since we never (or rarely) saw a 10% difference in the proportions of female survivors versus male survivors, it seems that horror movies actually favor female survivors.***
28. Ask the students:
    a. If you were going to be cast in a horror movie, would you want to be a male character or a female character? ***You would want to be a female character because they are more likely to survive by the end of the film.***
29. Inform the students that they will learn how to shuffle in RStudio in order to determine if an event is real or simply due to chance.

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 the next 5 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

LAB 2E: The Horror Movie Shuffle

Complete Lab 2E prior to Lesson 14.

Lab 2E - The Horror Movie Shuffle

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

Playing with permutations

- *Slasher* films are notoriously gory and are said to contain recurring biases.
  - One such bias, is that women in slasher films are more likely to survive than men.
- This lab will focus on the statistical question: *Are women in slasher films more likely to survive until the end of the film than men?*
- To answer this question, we'll learn how to use permuted data to gauge how likely an event occurs by chance.
- To begin, use the `data` function to load the `slasher` data file.
  - The data contains information about 485 characters from a random sample of 50 *slasher* horror films.

Initial thoughts...

- To familiarize yourself with the data, answer the following:
  - **How many variables and observations are contained in the data and what are the possible values of the variables?**
  - **Which gender had more survivors? Write down a few sentences as to how you came to your conclusion. Be sure to look at both the *counts* and *percentages* of survivors in each group before deciding.**
  - **Calculate the difference between the percentage of females who survived and the percentage of males who survived. Is the difference large enough to conclude that women tend to survive more often than men?**

Tally whoa ... !

- Something you might have noticed is that these two lines of code aren't equivalent:

```
tally(gender ~ survival, data = slasher)

tally(survival ~ gender, data = slasher)
```

- One of these lines takes the group of *survivors* and tells us how many of them were `Male` or `Female`.
- The other takes the group of *females / males* and tells us how many of them `Dies` or `Survives`.
- **The last question on the previous slide can be answered using the 2nd line of code. Why?**
  - Pro-tip: Include the option `format = "percent"` to obtain a two-way table with percentages.

```
tally(survival ~ gender, format = "percent", data=slasher, margin = TRUE)
```

```
##          gender
## survival    Female       Male
##  Dies       77.47748     86.69202
##  Survives   22.52252     13.30798
##  Total      100.00000    100.00000
```

Examining differences

- When we're comparing the difference between two quantities, such as survival rates of slasher films, it can be difficult to decide how *different* two values need to be before we can conclude that the difference didn't just happen by chance.
  - To help us decide when a difference is not due to chance, we'll use repeated random shuffling.
- By using repeated random shuffling, we'll estimate how often our *actual* difference occurs by *chance*.

Do the shuffle!

- When we shuffle data, we use our original data set as a starting point.
  - Run the following and write down the resulting table on a piece of paper.

```
tally(survival ~ gender, data = slasher)
```

- Now run the following to randomly reassign each `survival` status to each observation. Compare the resulting table to the one you wrote down.

```
tally(shuffle(survival) ~ gender,
  data = slasher)
```

Let's compare ...

- **How many people, in total, survived the slasher film before shuffling? How many people survived after shuffling?**
- **How has shuffling our data changed the percentage of women who survived compared to men who survived?**
  - **Is the difference in proportions from your shuffled data larger or smaller than the difference from the original data? Interpret what this means.**
- **Explain why shuffling our data one time is not enough to decide if the difference seen in our *actual* data occurs by chance or not.**

Detecting differences

- To help us decide if the difference in percentages in our *actual* data occurs by chance or not, we can use the `do()` function to shuffle our data many times and see how often our *actual* difference occurred by chance.
- Run the following lines of code:

```
set.seed(7)
shuffled_outcomes <- do(10) * tally(shuffle(survival) ~ gender,
  format = "percent", data = slasher)
View(shuffled_outcomes)
```

- **In how many simulations did a higher percentage of males survive than females?**
- **What is the largest difference in percentages of survival between males and females?**
- **What patterns are emerging from these simulations?**
- Ten simulations is not enough. Use `do`, `tally` and `shuffle` functions to `shuffle` the `survival` variable and `tally` the percentage of women who survived 500 times. `Assign` your 500 shuffles the name `shuffled_survivors`. Use `set.seed(1)`

Now what?

- The next step to find out how often our *actual* difference occurs by chance is to compare it to the differences in our shuffled data.
- To compute the differences for each shuffle we can use the `mutate` function.
  - Fill in the blanks to add the difference between `Survives.Female` and `Survives.Male` to our `shuffled_survivors` data.

```
shuffled_survivors <- mutate(shuffled_survivors,
                             diff = ____ - ____)
```

Time to decide

- Create a `histogram` of the `differences` in our `shuffled_survivors` data. Based on your plot, answer the following
  - **What was the typical difference in percentages between men and women survivors?**
- Include a vertical line in your histogram of the actual difference by running the code below:

```
add_line(vline = 22.52252 - 13.30798)
```

- **Does the actual difference occur very often by chance alone?**
- **Does gender play a role in whether or not a character will survive in a horror film? Explain your reasoning.**
- **If you wanted to survive in a horror film, would you want to play a female character or a male character?**

Summary

- By shuffling the `survival` label, we made it so that the proportion of males and females who survived the slasher film was random.
  - The males and females survived by chance alone.
- If surviving the film occurred purely by chance, then most of the time the difference in survival proportions was close to zero.
  - Notice how most values in the histogram occur close to zero.
- When we look to see how often our actual difference occurs in our shuffled data, if the actual difference doesn't occur very often then perhaps there is something more going on than just chance alone ...

On your own

- Carry out another 500 simulations but this time shuffle the `gender` variable instead of the `survival` variable.
  - Include the code `set.seed(1)` before your 500 simulations to make your answer reproducible.
- **Does shuffling the `gender` variable instead of the `survival` variable change your answer to the question: *Does gender play a role in whether or not a character will survive in a horror film?***
  - **Why or why not?**

Lesson 14: The Titanic Shuffle

Objective:

Students will continue to understand that, just by chance, we will see differences between two groups. They will understand that these differences are usually small.

Materials:

1. *LMR_Titanic Strips*
   Advanced preparation required (see Step 8 of lesson)
2. Poster paper
3. Markers

Essential Concepts: We can also "shuffle" data based on numerical variables. The statistic we use is the difference in medians. The distribution we form by this form of shuffling still represents what happens if chance were the only factor at play. When differences are small, we suspect that they might be due to chance. When differences are big, we suspect they might be 'real.'

Lesson:

1. Remind students that they previously learned how to determine if a difference is due to chance by shuffling based on categorical variables (gender and survival).
2. Display the dotplot created during Lesson 13 of the difference in proportions between female and survivors of horror movies. Remind the students that, "by chance," the differences were typically zero. Most of the time, they were pretty small. Sometimes they were bigger, but that was rare and this tells us that if we see "small" differences, we might think they are due to chance. But if we see "big" differences, they are not.
3. Lead a short discussion about what students think small and big differences mean. Make sure they answer in units (which are percentage points for the horror movie data). So, for example, a "big" difference might be 5 percentage points (but don't let them just say "5").
4. Inform students that, during today's lesson, they will learn how to determine if there is a difference between groups when a numerical variable is involved.
5. In particular, they will assume the roles of passengers in the *Titanic* for today's lesson. In case some students may not know about the *Titanic*, ask a volunteer to share what he/she knows.
6. Explain that, at its time, the *Titanic* was the largest cruise ship ever built and was declared to be unsinkable. However, on its first voyage, it sank and was one of the worst maritime disasters in history. About 40% of passengers survived; however, your chances of survival depended very much on your age, gender, and wealth.
7. Inform the students that we are going to look at whether the amount of money a passenger paid for his/her cabin (the fare price) had anything to do with whether or not he/she survived.
8. Each student will need a strip from the *LMR_Titanic Strips* file—see below for instructions.

   Advanced preparation required:

   The Titanic Strips LMR contains data from 40 actual passengers on the titanic. Each strip represents the data from one passenger: the left hand side shows the fare paid and right hand side contains the survival information of that passenger after the collision. Cut the LMR into strips such that the fare price is attached to the survivor status for each of the 40 observations.

Titanic Strips

$7.75	Survivor
$26.00	Survivor
$56.93	Survivor
$7.75	Survivor
$80.00	Survivor
$26.55	Survivor
$35.50	Survivor

LMR_Titanic Strips

9. 40 strips were created for large classes. If your class has less than 40 students, assign the students to two groups such that roughly 40% of them are in the survivor group (15/40 = 37.5% ≈ 40%), and the rest are in the victim group. If your class is small (smaller than 10), then put the students in two equal sized groups. The split does not have to be exactly 40%.
10. Inform the smaller group that they are the survivors and distribute a survivor strip with its corresponding fare to each student. Set aside any leftover strips. Tell them that the price on the strip represents the amount of money paid for their ticket to board the *Titanic*. Notify them that $20 in 1912 is worth about $500 today.
11. Divulge to the larger group that they, unfortunately, are the victims and distribute a victim strip with its corresponding fare to each student. Set aside any leftover strips.
12. Ask each group to create a dotplot of their fare prices on a poster. Lead a quick discussion comparing the two dotplots visually. Then, ask each group to calculate the mean fare for their group.
13. As a class, find the difference between the mean fares for the two groups.

    median of “Survivor” fares – median of “Victim” fares

    For example:

    If all 15 survivor cards and all 25 victim cards are used, the difference is medians would be:

    $26.00 – $13.00 = $13.00

14. Explain that one of the controversies of the *Titanic* disaster was that some people felt that the rich people were given better access to the lifeboats than were the poor, so rich people were more likely to survive. Note that the data represented on the fare cards are only a subset of the actual *Titanic* data, which had over 800 passengers. However, the data were randomly selected from the real data and are considered representative of the 800 passengers.
15. In pairs, ask students to discuss the following:
    a. Based on the data from our dotplots, do you think rich people were more likely to survive? In other words, did passengers who paid more for their tickets have a better chance of survival? ***Yes, there is evidence that rich passengers survived more often than poorer passengers. The median difference between the fare prices of the survivors and the victims is $13.00 (see Step 13). Most survivors had higher fare prices than the victims, so the distribution of survivor fares is shifted to the right and is more right-skewed.***
16. Share out a couple of responses with the whole class.
17. Have students tear their strip such that they separate the fare from the outcome (survivor or victim). Collect only the outcomes and randomly shuffle them. Students will keep their fare.

Distribute the shuffled outcome strips face down to the students. Once everyone has a new outcome strip, ask students to turn their outcome strip over and re-group based on their new survival status.

18. Ask the students:

    a. Why do we shuffle the survivor/victim strips and not the fare strips? ***We want to know if the price someone paid for his/her ticket affects whether or not he/she survived. So, when we shuffle, we assume that fare price has nothing to do with survival, so the prices should be irrelevant.***

    b. What do you think the median fare difference of our shuffled groups will be? ***The median fare difference of the shuffled groups should be close to 0, meaning that there should be NO difference in fare price for the survivors and the victims. Everyone would have the same chances of surviving, regardless of their ticket price.***

19. Have each group calculate the median fare price for their new groups. Then, ask:

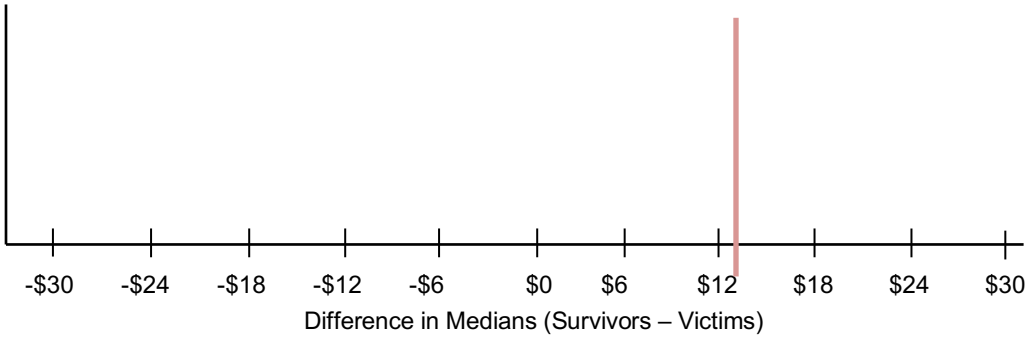
    a. Do you think this difference, of ___ dollars, is real or due to chance? ***Answers will vary by class. Since the data were shuffled, any difference should be due to chance.***

20. On the board, create a table to display the median fare prices for each group, and include a column for the difference (median "Survivor" fare – median "Victim" fare). Fill in the table with the values the students found in Step 13. **Note:** The first row has been filled in with the example data from above BEFORE the shuffles have taken place.

Median Fare Price of Survivors	Median Fare Price of Victims	Difference in Medians (Survivors - Victims)
$26.00	$13.00	$26.00 - $13.00 = $13.00

21. Note that values in the "Difference in Medians" column can be positive or negative because sometimes the survivors will pay more for their tickets, and other times the victims will pay more for their tickets.

22. Draw a dotplot on the board labeled "Difference in Medians." Include a vertical line at $13.00 (or whatever value was calculated in Step 13 by the class) to represent the actual difference in the median fare prices between the survivors and the victims (see example below).



23. Using the information from Steps 19 and 20, place a dot at the corresponding value for the shuffled data's difference in medians. Ask the students:

    a. How does this difference compare to the actual difference of $13.00 (from Step 13)? ***Answers will vary by class. Most likely, the difference in medians will be much smaller than $13.00. In fact, the difference in medians will be centered around 0.***

24. Remind students that small differences might be due to chance and big differences typically mean that there is a "real" difference between groups. In this case, a big difference might mean that the rich passengers were more likely to survive. And a small difference might mean that survival was just a matter of plain luck.
25. Repeat Steps 17 – 23 a few more times (depending on how much class time you have available).
26. In pairs, ask students to discuss whether they think the real difference in median fare prices they calculated in Step 13 ($13.00 if all cards were used) is small or large. ***Answers will vary by class. Guide students to look at the MAD value of the distribution of differences in median fares.***
27. Explain that one way that we can decide what is "large" or "small" is by creating cut-off values that we think are too far away from the center of the distributions of differences. In general, we can assign a rule that states that any difference in mean fare prices that is greater than 2 MAD values above or below the mean is considered unusual. This means that any value in the outer edges of the plot would indicate that a passenger's ticket price impacted his/her chances of survival.
28. Inform students that they will use RStudio to shuffle the actual *Titanic* data of all 800 passengers during the next class and can decide if the difference in survival rates of rich passengers and poor passengers was real, or just due to chance.

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 the next 3 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

LAB 2F: The Titanic Shuffle

Complete Lab 2F prior to Lesson 15.

Lab 2F - The Titanic Shuffle

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

Previously ...

- In the previous lab, we learned that by using a `do`-loop and the `shuffle` function, we could simulate randomly shuffling our data many times.
  - This helps us determine how likely it is that a difference between groups is due to chance.
- For this lab, will extend these ideas to *numerical* variables by using random shuffling and numerical summaries.
- The question we will investigate in this lab is:

  Is there any evidence to suggest that wealthier passengers on the Titanic were more likely to survive than poorer passengers?

- We will consider wealthier passengers to be those that paid a higher `fare` for their ticket.

The Titanic

- The Titanic was a ship that sank en route to the U.S.A. from England after hitting an Iceberg in 1912.
  - At the time, it was claimed that the Titanic was *unsinkable* ... it wasn't ... because it did.
- Use the `data` function to load the `titanic` passenger and survival data.
- Create a boxplot of the `fares` paid by passengers and facet the plot based on whether the passenger survived or not.
  - **Based on the plot, do you believe richer passengers were more likely to survive? Explain why and describe how certain you are of being correct.**

The search begins!

- Start your analysis by calculating how much more the *typical* survivor paid than the *typical* non-survivor in our data.
  - Based on the distributions of fares paid, which numerical summary that describes the *typical* value might be preferred?
- **What was the *typical* fare paid by survivors? Non-survivors? How much more did the typical survivor pay?**

Do the shuffle!

- Use the `do` and the `shuffle` functions to shuffle the passenger's survival status 500 times.
  - Use the previous lab if you need some help on how to do this.
  - For each shuffle, compute each group's `median` fare paid.
  - `Assign` your shuffled data the name `shuffled_survival`.
- After shuffling your data, use the `mutate` function to create a variable called `diff` which is the median fare of survivors minus the median fare of non-survivors. (Assign your mutated data the name `shuffled_survival` again).

Put your simulations to use

- **By using your shuffled data, answer the research question we posed at the beginning of the lab.**

Is there any evidence to suggest that wealthier passengers on the Titanic were more likely to survive than poorer passengers?

- **Write up your answer as a statistical analysis. Create a plot and explain how the plot supports your conclusion. Be sure to also explain why shuffling your data is important.**

Comparing Mean Fares

- What about if instead of calculating the median fare price for each group after a shuffle, we calculated the mean fare price and took the difference (mean_survivor – mean_victim).
- **If we did this 500 times, what do you predict the distribution of differences will look like?**
- Use the `do` and the `shuffle` functions to shuffle the passenger's survival status 500 times.
  - For each shuffle, compute each group's mean fare paid.
  - After shuffling your data, use the `mutate` function to create a variable called `diff` which is the mean fare of survivors minus the mean fare of non-survivors.
- **What does the shuffled data reveal? Does the answer to the research question below change when using the mean fares instead of the median fares?**

Is there evidence to suggest that those who survived paid a higher fare than those who died?

Lesson 15: Tangible Data Merging

Objective:

Students will learn how to merge two data sets and ask statistical questions about the merged data.

Materials:

1. *Tangible Data Merging* file (LMR_2.14_Tangible Data Merging)
   Advanced preparation required *(*see Step 4 of lesson*)*
2. Copy paper in two colors
   Advanced preparation required (see Step 4 of lesson)

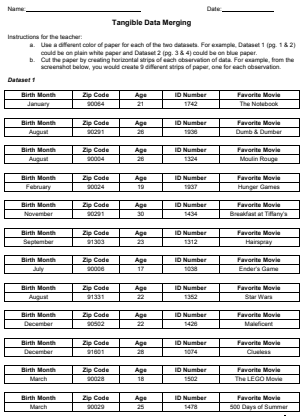
Vocabulary:

merge

Essential Concepts: We can enhance the context of a statistical problem by merging related data sets together. To merge data, each data set must have a "unique identifier" that tells us how to match up the lines of the data.

Lesson:

1. Inform students that they are going to examine the research question "Does the personality color test really work?" To answer this, we're going to examine whether the different color groups actually differ on particular beliefs or attitudes, or if these differences might just be due to chance. In particular, we are going to use the *Stress/Chill* data to see if there is evidence that the "colors" actually differ.
2. Show students the variables in each of these data sets. Give students time to brainstorm statistical questions of interest with their teams and record their questions in their IDS journals. Encourage them to think of two- and three-variable questions.
3. Conduct a share out of some of the questions students came up with. Examples include: ***(1) Do people whose predominant color is Gold tend to stress more than people whose predominant color is Blue? (2) Is there a difference between the sorts of things that stress out the different personality colors?***
4. In order to answer the above questions, we will need to merge our class's 2 data sets together (*Personality Color* and *Stress/Chill*). In order to do this, we will be practicing how to merge data sets today.
5. Print out the material from the *Tangible Data Merging* file (LMR_2.14). Use a different color of paper for each of the two data sets. For example, Data set 1 could be on plain white paper and Data set 2 could be on blue paper. Cut the paper by creating horizontal strips of each observation of data. For example, from the screenshot below of the first page of Data Set 1, you would create 12 different strips of paper, one for each observation.



Name:_____ Date:_____

Tangible Data Merging

Instructions for the teacher:

a. Use a different color of paper for each of the two datasets. For example, Dataset 1 (pg. 1 & 2) could be on plain white paper and Dataset 2 (pg. 3 & 4) could be on blue paper.
b. Cut the paper by creating horizontal strips of each observation of data. For example, from the screenshot below, you would create 9 different strips of paper, one for each observation.

Dataset 1

Birth Month	Zip Code	Age	ID Number	Favorite Movie
January	90064	21	1742	The Notebook
Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	90291	26	1936	Dumb & Dumber
Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	90004	26	1324	Moulin Rouge
Birth Month	Zip Code	Age	ID Number	Favorite Movie
February	90024	19	1937	Hunger Games
Birth Month	Zip Code	Age	ID Number	Favorite Movie
November	90291	30	1434	Breakfast at Tiffany's
Birth Month	Zip Code	Age	ID Number	Favorite Movie
September	91303	23	1312	Hairspray
Birth Month	Zip Code	Age	ID Number	Favorite Movie
July	90066	17	1538	Ender's Game
Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	91331	22	1352	Star Wars
Birth Month	Zip Code	Age	ID Number	Favorite Movie
December	90502	22	1426	Maleficent
Birth Month	Zip Code	Age	ID Number	Favorite Movie
December	91801	28	1074	Clueless
Birth Month	Zip Code	Age	ID Number	Favorite Movie
March	90028	18	1502	The LEGO Movie
Birth Month	Zip Code	Age	ID Number	Favorite Movie
March	90029	25	1478	500 Days of Summer

LMR_2.14

6. Hand each student in the class a strip of paper. Ask them to try to find someone with the other data set (i.e., a person with a different colored strip of paper) that they can "match up," or **merge**, with.
7. For example, a student with the first row of data listed below from Data set 1 might want to match up with the second row of data listed below from Data set 2 because a person who is 21 has probably graduated high school.

Birth Month	Zip Code	Age	ID Number	Favorite Movie
January	90064	21	1742	The Notebook

Zip Code	ID Number	Birth Month	Siblings	Education
91331	1352	August	2	High School

8. However, they should notice that they cannot just make guesses about a person's characteristics in order to match up the data. They should realize that only 3 of the variables are the same in both data sets: *Birth Month*, *Zip Code,* and *ID Number.*
    a. Since multiple people have the same *Birth Month*, discuss why this may not be the best variable to merge with. ***Multiple people are born in January, so we would have no way of differentiating between those people.***
    b. The same is true for the *Zip Codes* variable. Although there are less repeats with *Zip Codes*, we still see some overlap between observations.
    c. So, the only *UNIQUE* identifier in both data sets is *ID Number*. So the students should end up in pairs at the end of the exercise – a student from Data set 1 is matched with the student from Data set 2 that has the same *ID Number*.
9. Have the students write about the experience of tangible data merging in their IDS journals and ask:
    a. Why is it important to have at least one unique identifier for both data sets? ***It is the only way to know which information belongs to which person. We want to make sure we do not match up observations (in this case, people) incorrectly because that will compromise any analysis we do later.***
10. Inform students that they will learn to merge data sets using RStudio during the next lab.

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 collect data for one more day for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org/

LAB 2G: Getting it Together

Complete Lab 2G prior to the Practicum.

Lab 2G - Getting It Together

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

Putting data together

- In the labs so far, we've only ever looked at individual data files.
- But often times, we gain additional insights by including additional information from a separate data set.
- In this lab, we will learn how to merge information from our *personality color* data with our *stress/chill* data.
- *Export, upload, import* your *Personality Color* data set and name it `colors`.
- Then, *export, upload, import* your *Stress/Chill* data set and name it `stress`.

Looking at Stress/Chill

- We would like to analyze the research question:

  How do people's personality colors and/or sports participation affect their stress levels?

- We already have data about *personality color* and a separate data set about *stress*.
  - What we don't have is a single data set with information from both ... yet.
- We'll start then by strategizing how to merge our data together.

Deciding how to merge

- Before we merge data, we need to decide *how* we plan to merge it:
- We can *stack* our data sets, that is, take one data set's rows and add them to the bottom of the other data set.
- We can also *join* our data sets horizontally. This is where we take one data set's columns and add them to the end of the other data set's columns based on matching an *ID* variable.
  - The *ID* variable will have entries that we use to *match* observations in both data sets.
- **To answer the statistical question of interest, would it make more sense to *stack* or *join* our `colors` and `stress` data?**

Finding variables in common:

- Look at the `names` of the variables in each data set.
  - To merge different data sets together, we need to find variables they have in common.
- **Which variables do the data sets have in common?**
- **Which variable would make sense to merge the data sets together with? Why not the others?**

Caution required

- Whether *stacking* or *joining*, we need to be careful when we merge data:
- When *stacking* data, we need to be absolutely certain that the variables we're stacking represent the exact same measurements.
  - We wouldn't want to stack `height` in meters and `height` in inches, for instance (without converting one to the other).
- When *joining* data, we need to make sure that the *id* variable in our primary data set matches to *one and only one* observation in the joining data.

  - Otherwise, R won't know which observation to match to.

Getting ready

- Our goal is to add the variables from the `colors` data onto the `stress` data.
- Start by ensuring that every `user.id` in the `colors` data is unique.
  - If there's a duplicate, have your teacher remove the duplicate from the IDS *Response Manager* and then re-*export*, *upload*, *import* your `colors` data.
- **After we add the data from *colors* to *stress*, how many rows should our merged data have? Write this number down.**

Putting them together

- We can use the `merge` function to *join* our data sets together using the variables that appear in both sets.
- **Fill in the blanks below to join the information from the `colors` data onto the `stress`.**

```
merge(____, ____, by = "____")
```

- `Assign` this `merged` data set the name `stress_colors`.
  - Make sure your data has the same number of observations that you wrote down on the previous slide.

Saving your data:

- `View` your merged data and make sure nothing appears to be blatantly wrong with it.
- **Why didn't we stack the rows of data instead?**
- **What happens if you swap the order of the data sets in the `merge` function?**
- Fill in the blank below to `save` our `stress_colors` data for later use.

```
save(stress_colors, file = "stress_colors.rda")
```

- Be sure to look in the *Files* tab to make sure your data was saved.

Moving on

- In the next lab, we'll begin analyzing our merged data. In the meantime:
- **Make a few plots using variables from the `stress` data and *facet* or *group* the plots based on variables from the `colors` data.**
  - **Write down the most interesting discovery you make by just exploring your data. Write out how you found your discovery and interpret what it means for the people in your class.**
- **With our *colors* data, we could answer questions about the *typical* color scores in your class. Why can we no longer answer this question in our `stress_colors` data?**

Practicum: What Stresses Us?

Objective:

Students will use RStudio to make graphical representations or numerical summaries of their *Stress/Chill* and *Personality Color* data to answer research questions.

Materials:

1. *What Stresses Us? Practicum* (LMR_U2_Practicum_What Stresses Us)

Practicum
What Stresses Us?

We made a data set that combined our *Stress/Chill* data with our *Personality Color* data. You will use this data to answer the following research questions:

- Do color personalities really predict a person's personality?
- Do people with different personality colors tend to have different stress levels?

Based on the merged data, you need to:

1. Write a one-page report to address these research questions. Use the Data Cycle. Your analysis should include both numerical methods (means, medians, etc.) and graphical methods (plots). The research questions are fairly broad, and you should first think of simpler statistical questions you could ask that would address these research questions.
2. In your report, be sure to:
    a. Provide the plot(s) and numerical summary (or summaries).
    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.
3. Present your report to another member of the class who is not in your team.
    a. Make sure to include any relevant plots or numerical summaries that you use.

    Note: You may use the scoring guide in Unit 1 to give you an idea of how to score the Practicum.

What's Normal?

Instructional Days: 5

Enduring Understandings

Students learn that the Normal curve can be used as a model that describes many real phenomena. Drawing plots of the Normal curve over histograms helps data scientists determine if the distribution represented by the histogram is close to Normal. The Normal curve suggests that one is more likely to obtain values that are close to typical (average), which are found in the center of the curve, and less likely to obtain values that are extreme and farther away from typical.

Engagement

Students will learn about the Normal curve by watching the first 35 seconds the New York Times Video "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html.

Learning Objectives

Statistical/Mathematical:

S-ID 4: Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Understand that there are data sets for which such a procedure is not appropriate. Use calculators and RStudio to estimate areas under the normal curve.

S-IC 6: Evaluate reports based on data.

Data Science:

Learn to eyeball Normal distributions and overlay a Normal curve on a histogram; learn to simulate draws from a Normal distribution, and the impact of sample size; learn that estimating probabilities with a model leads to stable estimates; and estimate probabilities by finding the area under the Normal curve using RStudio.

Applied Computational Thinking Using RStudio:

- Use software to find the area under a Normal curve
- Use software to compare sample distributions (with histograms, for example) with the Normal distribution and make a decision as to whether the distribution appears Normally distributed.
- Draw random samples from a Normal distribution using software.

Real-World Connections:

The Normal curve is used to make inferences about a population. The model makes it possible to estimate the probability of occurrence of any value of a Normally distributed variable. For example, heights are Normally distributed. Using a Normal curve, we can find the probability of that a person would be a height of 6' 2".

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.

Data File or Data Collection Method

Data Files:

1. CDC data (`cdc`)
2. Titanic data (`titanic`)

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 16: What is Normal?

Objective:

Students will learn what a Normal distribution is and learn how to identify a Normal distribution.

Materials:

1. Video: *New York Times'* "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html

   Note: Show only the first 41 seconds of the video.
2. Graphics from the *Normal Plots* file (LMR_2.15_Normal Plots)
3. Projector to display plots
4. 3" x 5" cards (1 per student)

Vocabulary:

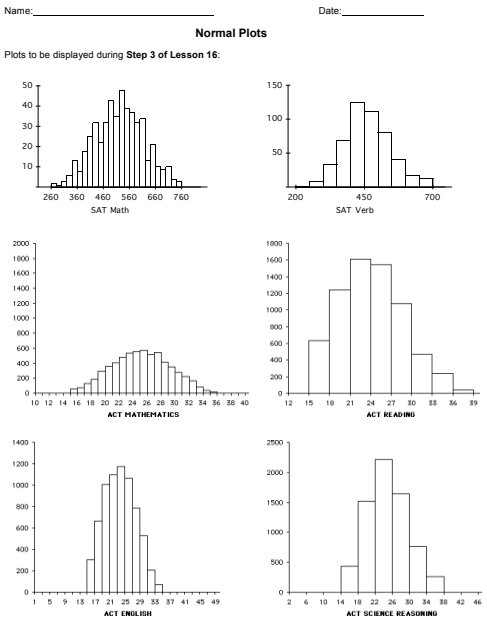
bell-shaped, normal curve, normal distribution

Essential Concepts: The Normal curve, also called the Gaussian distribution and the "bell curve," is a model that describes many real-life distributions and is usually called the Normal Model.

Lesson:

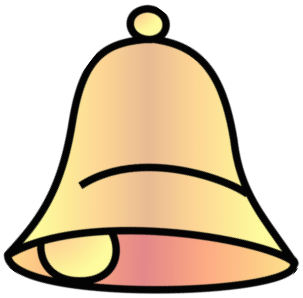
1. Remind students that in Unit 1, Lesson 11 (*What Shape Are You In?*), they sorted histograms into groups based on their shapes.
2. The *Normal Plots* file (LMR_2.15) contains some of the unimodal **bell-shaped** distributions from the original handout of that lesson (*Sorting Histograms* handout (LMR_1.10)).

   Note: You do not need the original handout from Unit 1 – all relevant plots have been compiled in the *Normal Plots* file (LMR_2.15) for accessibility. Six plots are included: SAT Math, SAT Verb, ACT Mathematics, ACT Reading, ACT English, and ACT Science Reasoning.



LMR_2.15

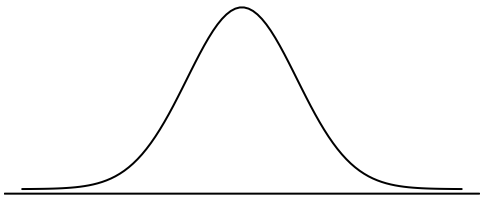
3. Display the group of bell-shaped distributions from page 1 of the *Normal Plots* file (LMR_2.15) to the class and ask the students:
    a. What characteristic does this particular group share? ***All of these plots are unimodal (one mode/peak) and symmetric.***
    b. Inform students that these types of distributions are often referred to as bell-shaped. Why might this term be used? ***The histograms look very similar to the shape of a bell.***
4. To show the similarities between the shape of a bell and the shape of these distributions, a clip-art image (shown here) has been included in the *Normal Plots* file (LMR_2.15) on page 2.



5. Explain that this shape occurs often in real-life. It occurs so often that it's been given its own name: the **normal curve**, or **normal distribution**. Can the students think of distributions where they have seen Normal curves in previous labs?
6. To give some more background on the normal distribution, play the New York Times video titled "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html

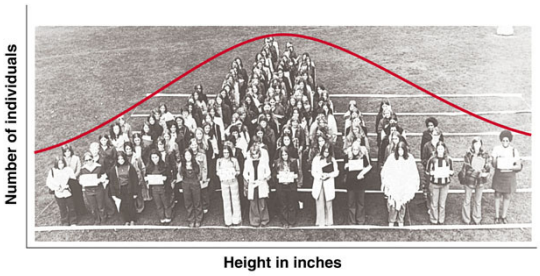
    Note: Show only the first 41 seconds of the video.
7. Discuss that the normal curve has a very precise mathematical definition, which is pretty complex. But the result is a curve that looks like the one in the "Bunnies" video. In general, the curve looks like the plot shown below.

    Note: You can either draw the diagram below on the board or display it via a projector – the image can be found on page 2 of the *Normal Plots* file (LMR_2.15).



8. Explain that normal distributions are good for describing some populations of people. For example, people's heights are often considered to be normally distributed. Display the famous Frank Anscombe photograph (shown below) via a projector. The graphic can be found on page 2 of the *Normal Plots* file (LMR_2.15). Inform the students that this photo was taken of a group of randomly selected college women who stood in height order.

Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6



9. Next, lead a discussion about why the normal curve is a good fit to the histogram in the above picture. Notice that more people are near the center of the distribution, and fewer are in the outer edges, or tails. Engage the class in a conversation using the following probing questions:
    a. Notice that there is a peak in the center of the distribution. What height do you think is at the center? ***The average height of American women is approximately 5'5" (5 feet, 5 inches) tall. We might therefore expect the average height of this group to be close to 5'5" as well.***
    b. Why are more people in the center, and less people in the edges, or tails, of the distribution? ***The center represents the mean height. Most women will fall somewhere close to the mean and may be a few inches shorter or taller than it. However, less people are likely to be MUCH shorter or MUCH taller than the mean. For example, we would not expect to see many women who are 4'10" tall nor would we expect to see many women who are 6'0" tall.***

10. Explain that the normal curve is a good description of a distribution when it makes sense that there is a single 'typical' value with random deviations above and below that value. Ask students:
    a. Why does this make sense with heights but not with incomes? ***With heights, we expect most people to be near the average, with some deviations above and below the mean (people who are taller or shorter than the mean height). However, with incomes, the distribution will have more deviations that are above the typical value, since there is no upper bound for a maximum income (ex. Bill Gates, Warren Buffett, etc.***
    b. Are there more real-life examples, other than height, that students think might follow a normal distribution? ***Answers will vary by class. Some examples include: (1) scores on standardized math and reading tests (like the SAT and ACT), (2) IQ scores, and (3) body temperatures.***
    c. Does it matter that the curve drawn on the photograph does not match exactly to the women's heights? ***No. We often refer to the curve as the "normal model" because the curve is a just a model of the true population distribution. So, even though the red curve is not exactly the same as the women's heights, it is a close enough approximation of the shape of their heights.***

    Note to teacher: The main role the normal distribution has historically played has been in modeling errors (most measurements will be close to the actual value while larger errors occur less often) and sample means.

11. Inform the students that, during the next few lessons, they will be learning more about the normal distribution. In particular, they will learn about a new measure of spread used to describe a normal distribution, how to calculate probabilities from this distribution, and how to randomly sample from this distribution.

12. *Cheat Card*: Distribute an index card to students and ask them to create a cheat card that will help them remember information about the normal curve.

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 their cheat cards if they were not able to finish in class.

Lesson 17: A Normal Measure of Spread

Objective:
Students will learn that standard deviation is another way to measure variability.

Materials:
1. *How Far Apart?* handout (LMR_2.6_How Far Apart) – completed during Lesson 4
2. *How Far Apart? (with standard deviation – SD)* handout (LMR_2.16_How Far Apart SD)
3. Projector to display visuals using RStudio
4. RScript with the functions in this lesson

Vocabulary:
standard deviation (SD)

Essential Concepts: The standard deviation is another measure of spread. This is commonly used by statisticians because of its role in common models and distributions, such as the Normal Model.

Lesson:
1. In their IDS journals, ask students to create a two-column table and label the left-column as *Measures of Center (Central Tendency)* and the right column as *Measures of Spread (Dispersion).*
2. In pairs, ask students to recall methods they have learned so far for measuring center and measuring spread in distributions.

   Measures of Center: ***mean (average or typical value), median***
   Measures of Spread: ***mean absolute deviation (MAD), interquartile range (IQR)***
3. Share out a pair's explanation and ask the rest of the pairs to agree or disagree. If there is disagreement, hold a class discussion until the lists are correct.
4. Point out that a measure of center or a measure of spread depicts one value for a distribution. Ask student pairs to discuss the following question:
   a. What does the value of each measure tell us about the data in the distribution? ***Possible answer: A measure of center tells us the value that is typical, or in the center. A measure of spread tells us how variable, or how spread apart, the data are.***
5. Next, ask students to add the term **standard deviation (SD)** to their *Measures of Spread* column.
6. Inform students that the standard deviation of a distribution is another way to measure spread, or variability. The standard deviation is similar to the mean absolute deviation (MAD).
7. Ask students to recall the formula for calculating the MAD:

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

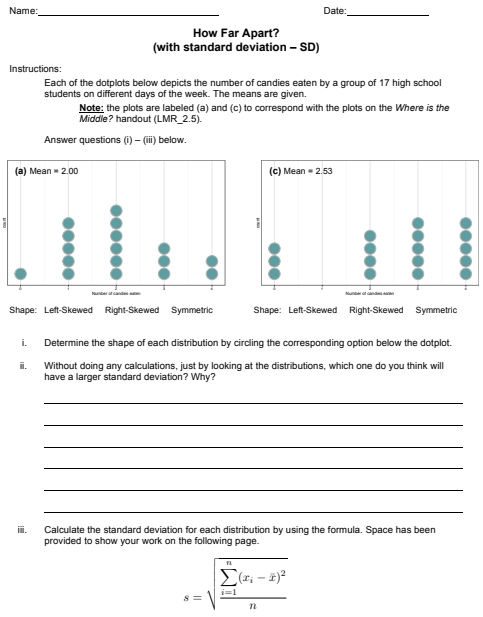
8. While the MAD measures the absolute distance of each data point from the mean, the standard deviation squares the distances of each data point from the mean. Both methods result in positive measurements because distance is always positive.
9. Ask students to recall that they calculated MAD values in the *How Far Apart?* handout (LMR_2.6) during Lesson 4 of this unit.

10. Show and discuss the formula for calculating the standard deviation of a data set:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Note to teacher: There are different formulas for the standard deviation. We are presenting the simpler one, which divides by *n*. In AP Statistics (or college introductory statistics), students will learn that if they are using a sample of data to estimate the standard deviation for the population, then dividing by $n-1$ is a better estimator than dividing by *n*. But this technically requires a lot of scaffolding and leads to little understanding, and so we will stick with the simpler version. (In some books, this is called the "population value of the standard deviation" and the $n-1$ version is called the "sample estimate of the standard deviation.")

11. Guide the class to complete the *How Far Apart? (with standard deviation -- SD)* handout (LMR_2.16) to calculate standard deviations of the dotplots using the formula listed above.



Name:_____ Date:_____

How Far Apart?
(with standard deviation – SD)

Instructions:

Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given.

Note: the plots are labeled (a) and (c) to correspond with the plots on the *Where is the Middle?* handout (LMR_2.5).

Answer questions (i) – (iii) below.

(a) Mean = 2.00

(c) Mean = 2.53

Shape: Left-Skewed Right-Skewed Symmetric

Shape: Left-Skewed Right-Skewed Symmetric

i. Determine the shape of each distribution by circling the corresponding option below the dotplot.

ii. Without doing any calculations, just by looking at the distributions, which one do you think will have a larger standard deviation? Why?

iii. Calculate the standard deviation for each distribution by using the formula. Space has been provided to show your work on the following page.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

LMR_2.16

Answers: Plot (a) – SD = 1.0847 candies; Plot (c) – SD = 1.3770 candies

12. As a whole group, ask students to compare and contrast the standard deviations with the MAD values for the two plots in the handout.

Similarities between SD and MAD:

- ***Measure the same idea: variability, or spread***
- ***Are based on looking at the "deviations" from the mean: the difference between an observation and the mean***
- ***Uses the "typical" deviation***

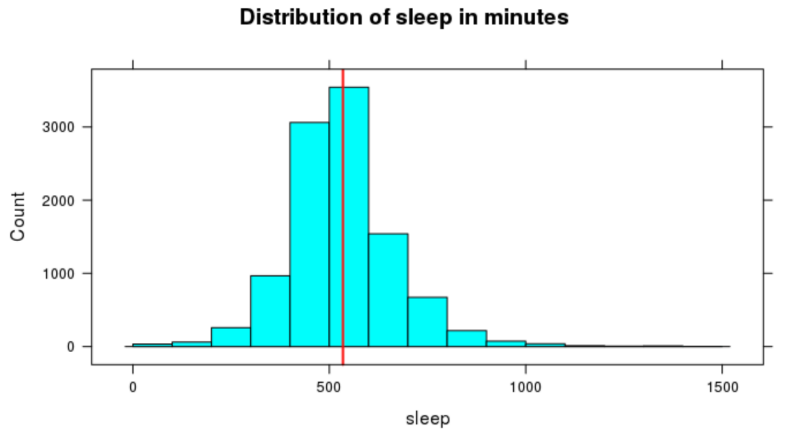
Differences between SD and MAD:

- ***The MAD uses the absolute value and finds the average of the absolute deviations from the mean***
- ***The SD uses the square of each deviation from the mean, and finds the average of the squares***
- ***The SD takes the square root of the average of the squares***

13. Ask students why they think the SD takes the square root of the average of the squares.
Possible response: Taking the square root of the average of the squares returns the measurements to their original units instead of square units.

14. To reinforce students' conceptual understanding of standard deviation, student teams will estimate the standard deviation for a few numerical distributions and explain the reasoning for their estimate. Load and view the atus data, then run the following functions one by one:

```
> histogram(~sleep, data=atus, breaks=seq(0,1500,by=100),
                main = "Distribution of sleep in minutes")
> sleep_mean<-mean(~sleep, data=atus)
> add_line(vline=sleep_mean)
```



15. Zoom in on the visual and give student teams a few minutes to discuss what they estimate the standard deviation to be. Have the reporter from each team report their estimate using the following sentence frame:

"The time spent sleeping (in minutes) typically varies from the mean by _____minutes."

16. Reveal the actual standard deviation by running the function:

```
> sd(~sleep, data=atus)
```

17. Choose a reporter from a student team that had a good approximation to explain their reasoning.

18. Repeat this process with a few more numerical variables. Functions are provided below.

Household size

```
> histogram(~household_size, data=atus, nint=13)
> household_mean<-mean(~household_size, data=atus)
> add_line(vline=household_mean)
```

"Household sizes typically vary from the mean by _____people."

```
> sd(~socializing, data=atus)
```

Socializing

```
> histogram(~socializing, data=atus, breaks=seq(0,2000,by=100))
> social_mean<-mean(~socializing, data=atus)
> add_line(vline=social_mean)
```

"The time spent socializing (in minutes) typically varies from the mean by _____minutes."

```
> sd(~socializing, data=atus)
```

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 18: What's Your Z-Score?

Objective:

Students will understand that a z-score can be used to measure how far away - or how many standard deviations - an observation is away from the mean. Typically z-scores will range between -3 and +3. For simulations involving shuffling, if we compute a z-score that lies far away from the mean, then we might conclude that the outcome was not due to chance. If we see a z-score that lies close to the mean, then we might conclude it was by chance.

Materials:

1. Projector to display RStudio function
2. *RScript with all of the functions in this lesson*
3. *A ruler with centimeter marks on it*

Vocabulary:

z-score, standardized score, Empirical Rule

Essential Concepts: z-scores offer us a way to measure how extreme a value is, regardless of the units of measurement. Typically z-scores will range between -3 and +3, so values that are at or are more extreme than -3 or +3 standard deviations are considered extremely rare.

Lesson:

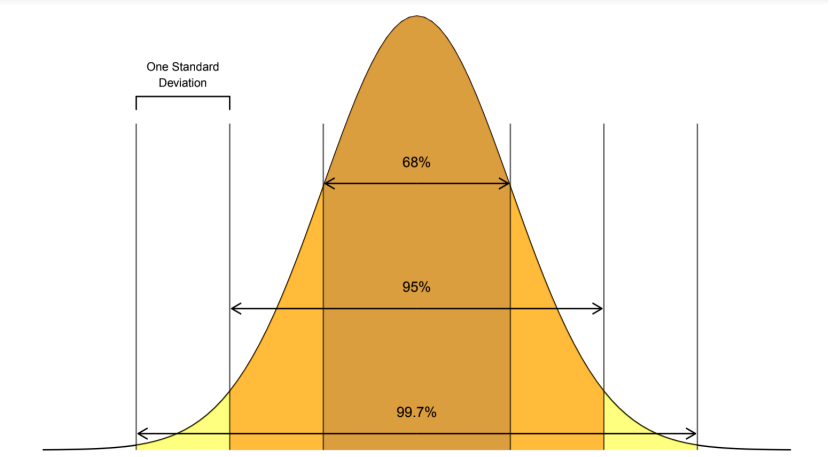
1. Ask students to recall what they remember about normal distributions.

   Answer: Normal distributions are unimodal and symmetric, and are often referred to as bell-shaped. Some real-life examples of variables that produce normal distributions are people's heights, scores on standardized tests, and body temperatures.

2. Display the following statement to students: "All normal distributions are bell-shaped, but not all bell-shaped distributions are normal." Then inform students that normal distributions have special properties.

3. Display the image below and introduce the **Empirical Rule**, which states:

   - Approximately 68% of the observations in a normal distribution fall within one standard deviation of the mean
   - Approximately 95% of the observations in a normal distribution fall within 2 standard deviations of the mean
   - Approximately 99.7% of the observations in a normal distribution fall within 3 standard deviations of the mean



4. Open RStudio and project for students to see. Read in the babies data set by following these steps:

    - On your **Environment Pane** go to Import Dataset
    - Choose **From Text (readr)**…
    - Paste the following in the **File/URL** box: http://people.hsc.edu/faculty-staff/blins/classes/spring17/math222/data/babies.csv
    - Click on **Update** and then **Import**

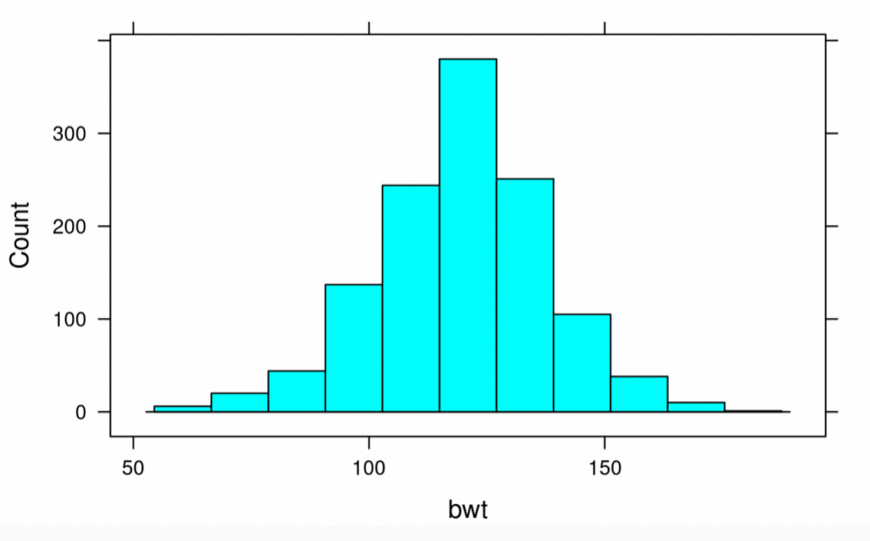
    Scroll through the spreadsheet so that students can see the variables. Ask student teams to predict which of the variables in the "babies" dataset they think might be normally distributed. Choose a couple of teams to share out.

    Description of variables:

    - bwt – birth weight (in ounces)
    - gestation – length of pregnancy (in days)
    - parity – 1 if baby was first born, 0 otherwise
    - age – mother's age (in years)
    - height – mother's height (in inches)
    - weight – mother's weight (in lbs.)
    - smoke – 1 if the mother is a smoker, 0 otherwise

5. Create histograms using the variables shared by student teams. There are a few variables that look normally distributed, such as birth mother's heights. We will investigate the babies' birth weights.

    ```
    histogram(~bwt, data = babies)
    ```



6. Ask students:
    a. Does the distribution of baby birth weights look approximately normal? Explain. ***Answer: The distribution of baby birth weights is unimodal, roughly symmetric, and somewhat bell-shaped, so it might be approximately normal.***
    b. What do you approximate the mean weight of the distribution to be? How about the standard deviation? ***Answers will vary. Use this as a check for understanding of standard deviation as well as estimating the mean using the balancing point concept. See next step for calculating the actual mean weight and standard deviation.***

7. Use RStudio to calculate the actual mean and standard deviation.

```
mean_bwt <- mean(~bwt, data = babies)

sd_bwt <- sd(~bwt, data = babies)
```

mean_bwt	119.576860841424
sd_bwt	18.2364518669726

8. Have students draw a number line with seven equally spaced intervals and label it “Baby birth weight in ounces.” Make sure students leave about 5 centimeters of space above the number line to draw a normal curve. Have students label the middle tick mark with the mean baby weight (round to the nearest tenth of an ounce=119.6 ounces). Then ask students:
    a. What weight is one standard deviation above the mean? ***Answer: A baby whose weight is 137.8 ounces is one standard deviation above the mean baby weight.***
    b. What weight is one standard deviation below the mean? ***Answer: A baby whose weight is 101.4 ounces is one standard deviation below the mean baby weight.***

    Have students label their number line with these values.

9. Have students continue filling their number line with the corresponding weights that are two and three standard deviations from the mean. ***Answer: A baby who weighs 156 ounces is two standard deviations above the mean weight, and a baby who weights 174.2 ounces is three standard deviations above the mean weight. A baby who weighs 83.2 ounces is two standard deviations below the mean weight, and a baby who weighs 65 ounces is three standard deviations below the mean weight.***

10. Ask students: If the distribution of baby weights is approximately normal, what percentage of babies weigh between 101.4 and 137.8 ounces? ***Answer: If the distribution of baby weights is approximately normal, about 68% of babies should be between 101.4 ounces and 137.8 ounces.***

11. Use RStudio to confirm if indeed the distribution of baby weights is approximately normal.

```
> one_sd_bwt <- filter(babies, bwt > 101.4, bwt < 137.8)
```

    Answer: In this sample of 1236 observations, there are 861 babies whose weights are one standard deviation from the mean, so 861/1236 = 0.697. This means that around 69.7% of the weights of babies in this sample fall within one standard deviation from the mean baby weight. This is close to 68%, so it seems that the distribution of baby weights is approximately normally distributed.

    Note: If you continue this process for this sample, you will find that the distribution of baby weights is normally distributed as defined by the Empirical Rule. In this sample, 1171/1236 = 94.7% of the baby weights fall within two standard deviations of the mean, and 1229/1236 = 99.4% of the baby weights fall within three standard deviations of the mean.
    .

12. Now that it has been verified that a normal distribution is an appropriate model for this distribution, have students draw a normal curve above the number line. Suggested method to obtain a decent normal curve:
    - Step 1: Draw a dot 4 centimeters above the mean height
    - Step 2: Draw dots 2.4 cm above the heights that are 1 standard deviation from the mean
    - Step 3: Draw dots 0.36 cm above the heights that are 2 standard deviations from the mean

- Step 4: Draw dots right above the number line for the heights that are 3 standard deviations from the mean
- Step 5: Connect the dots with a smooth curve

13. Tell students that we are using this normal curve as a model to represent the distribution of all baby weights. This will allow us to make comparisons, draw conclusions, and make predictions about baby weights. Let's see:
    a. What percentage of babies weigh less than 119.6 ounces? Explain. ***Answer: About 50% of babies weigh less than 119.6 ounces. Since normal distributions are symmetric, the mean and the median are about the same. Since the median divides a distribution into equal halves, in this case so does the mean.***
    b. What percentage of babies weigh between 119.6 and 137.8 ounces? ***Answer: About 34% of babies weigh between 119.6 and 137.8 ounces. According to the Empirical rule, 68% of the observations fall within one standard deviation of the mean, and since normal distributions are symmetric, the area under the curve from the mean to one standard deviation is half of 68% or 34%.***
    c. What percentage of babies weigh more than 137.8 ounces? ***Answer: About 16% of babies weigh more than 137.8 ounces. From part a and b above, we know that 50%+34%=84% of babies weigh less than 137.8 ounces, so 100%-84%=16% weigh more than 137.8 ounces.***

14. Explain that statisticians use something called a **z-score** to compare values. A z-score tells us how many standard deviations away from the mean an observation is. Another name for z-score is a **standardized score**.
15. Introduce the formula for calculating a z-score and discuss what each symbol in the formula means.

$$z = \frac{x - \bar{x}}{s}$$

16. Explain that z-scores answer the question: "How typical is x?" If x is the same as the typical value (the mean), then z = 0. If x is one standard deviation away from the mean, then z = -1 or +1. Remind students from the normal curve that as you move farther from the center (from the mean), there are fewer observations. Therefore, a large z-score is considered an unusual value.
17. Have students calculate the z-score for a baby that weighs 100 ounces:

    z = (100 – 119.6) / 18.2 = -1.08

    Ask the class:
    a. What does a negative z-score mean? ***A negative z-score means the x value is below the mean. This means that the weight is below average.***
    b. What does a positive z-score mean? ***A positive z-score means the x value is above the mean. This means that the weight is above average.***
    c. What is the most negative z-score you think we will find? What is the most positive z-score? ***Typically, values in a normal distribution rarely fall outside two or three standard deviations from the mean. So, if our data is purely by chance, we probably won't see any values that are less than -3 or greater than +3.***
18. Ask students: "Where does a baby that weighs 100 ounces fall within the distribution of baby weights?" Have students find 100 ounces on the x-axis of the normal curve and draw a vertical line from the x-axis until it intersects the normal curve. Have them shade the area under the curve to the left of the vertical line.

19. Tell students that the shaded area represents a percentile in the distribution. A percentile is the exact value in which the desired proportion of observations lie below the specific value in a distribution. Use RStudio to calculate the percentile.

```
pnorm(100, mean = 119.6, sd = 18.2) = 0.140
```

20. Doctors report percentiles to describe a child's development compared to other children their age. For a baby that weighs 100 ounces, a doctor would report the following: "The baby is at the 14th percentile in weight." This means that the baby weighs more than 14% of all babies.

    Note: A z-score can also be used to calculate a percentile, but since a z-score is a standardized score, the mean of the distribution would be zero and the standard deviation would be one.

```
pnorm(-1.08, mean = 0, sd = 1) = 0.140
```

21. Inform the class that they will be using RStudio during the next few days to practice using normal models.

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 2 Days

LAB 2H: Eyeballing Normal

LAB 2I: R's Normal Distribution Alphabet

Complete Labs 2H and 2I prior to the End of Unit Design Project.

Lab 2H - Eyeballing Normal

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

What's normal?

- The *normal distribution* is a curve we often see in real data.
  - We see it in people's blood pressures and in measurement errors.
- When data appears to be *normally distributed*, we can use the *normal model* to:
- Simulate *normally distributed* data.
- Easily compute probabilities.
- In this lab, we'll look at some previous data sets to see if we can find data that are roughly normally distributed.

The normal distribution

- The normal distribution is *symmetric about the mean*:
  - The `mean` is found in the very center of the distribution.
  - And the curve looks the same to the left of the mean as it does on the right.
- Use the following to draw a normal distribution:

```
plotDist('norm', mean = 0, sd = 1)
```

The mean and sd of it

- To draw a normal curve, we need to know exactly 2 things:
  - The `mean` and `sd`.
- The `sd`, or *standard deviation*, is a measure of spread that's similar to the `MAD`.
- **Which part of the normal curve changes when the value of the `mean` changes?**
- **Which part of the normal curve changes when the value of the `sd` changes?**
- *Hint*: Try changing the `mean` and `sd` values in the `plotDist` function.

Finding normal distributions

- Load the `cdc` data and use the `histogram` function to answer the following:
- **Think about the `height` and `weight` variables. Based on what you know about these variables, which of the variables do you think have distributions that will look like the normal distribution?**
  - **Make histograms of these variables. Which ones look like the normal distribution?**
  - *Hint*: To help answer this question, try including the option `fit = "normal"` in the histogram function. You might also try faceting by `gender`.

Using normal models

- Data scientists like using normal models because it often resembles real data.
  - *But not EVERYTHING is normally distributed.*
- As a data scientist in training, you must decide when a normal model seems appropriate.
  - No model is ever perfect 100% of the time.
  - If you choose a model, you should be able to justify why you chose it.

On your own

- **For each of the following, determine which, if any, appear to be normally distributed. Explain your reasoning:**
- Hint: Refer to Lab 2E and 2F
  - **The difference in `percentages` between male and female survivors in a slasher film for 500 random shuffles.**
  - **The difference in `median` fares between survivors and non-survivors on the Titanic for 500 random shuffles.**
  - **The difference in `mean` fares between survivors and non-survivors on the Titanic for 500 random shuffles.**

Lab 2I - R's Normal Distribution Alphabet

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

Where we're headed

- In the last lab, you were able to overlay a normal curve on histograms of data to help you decide if the data's distribution is close to a normal distribution.
  - We also saw that calculating the `mean` of random shuffles also produces differences that are normally distributed.
- In this lab, we'll learn how to use some other R functions to:
  - Simulate random draws from a normal distribution.
  - Calculate probabilities with normal distributions.

Get set up

- Start by loading the `titanic` data and calculate the `mean` `age` of people in the data but `shuffle` their `survival` status 500 times.
  - `Assign` this data the name `shfls`.
- After creating `shfls`, use `mutate` to add a new variable to the data set. This new variable should have the name `diff` and should be the `mean` `age` of those who survived minus those who died.
- Finally, calculate the `mean` and `sd` of the `diff` variable.
  - `Assign` these values the name `diff_mean` and `diff_sd`.

Is it normal?

- Before we proceed, we need to verify that our `diff` variable looks approximately normally distributed.
  - **Is the distribution close to normal? Explain how you determined this. Describe the center and spread of the distribution.**
  - **Compute the mean difference in the age of the *actual* survivors and the actual non-survivors.**

Using the normal model

- Since the distribution of our `diff` variable appears normally distributed, we can use a normal model to estimate the probability of seeing differences that are more extreme than our actual data.
  - **Draw a sketch of a normal curve. Label the mean age difference, based on your shuffles, and the actual age difference of survivors minus non-survivors from the actual data. Then shade in the areas, under the normal curve, that are *smaller* than the actual difference.**
- Fill in the blanks to calculate the probability of an even smaller difference occurring than our actual difference using a normal model.

```
pnorm(____, mean = diff_mean, sd = ____)
```

Extreme probabilities

- The probability you calculated in the previous slide is an estimate for how often we expect to see a difference smaller than the actual one we observed, by chance alone.

- If you wanted to instead calculate the probability that the difference would be larger than the one observed, we could run (fill in the blanks):

```
1 - pnorm(____, mean = diff_mean, sd = ____)
```

Simulating normal draws

- We can simulate random draws from a normal distribution with the `rnorm` function.
  - Fill in the blanks in the following two lines of code to simulate 100 heights of randomly chosen men. Assume the `mean` height is 67 inches and the `standard deviation` is 3 inches.
  - Plot your simulated heights with a `histogram`.

```
draws <- rnorm(____, mean = ____, sd = ____)

histogram(draws, fit = ____)
```

P's and Q's

- We've seen that we can use `pnorm` to calculate *probabilities* based on a specified *quantity*.
  - Hence, why we call it "P" norm.
- Now we'll see how to do the opposite. That is, calculate the *quantity* for a specific *probability*.
  - Hence why we'll call this a "Q" norm.
- How tall can you be and still be in the shortest 25% of heights if the mean height is 67 inches with a standard deviation of 3 inches?

```
qnorm(____, mean = ____, sd = ____)
```

On your own

Conduct one of the statistical investigations below:

- Using the `titanic` data, answer the following statistical question:
  - **Were women on the Titanic typically younger than men?**
  - **Use a histogram, 500 random shuffles and a normal model to answer the question in the bullet above.**
- Using the `cdc` data
  - **Using 500 random shuffles and a normal model, how much taller would the typical male have to be than the typical female in order for the difference to be in the upper 1% by chance alone?**
  - **How can we use this value to justify the claim that the average Male in our data is taller than the average Female?**

End of Unit Design Project and Oral Presentation: Asking and Answering Statistical Questions of Our Own Data

Objective:

Students will apply their learning of the first and second units of the curriculum by completing an end of unit design project.

Materials:

1. *IDS Unit 2 – Design Project and Oral Presentation* (LMR_U2_Design Project)

End of Unit 2 Design Project and Oral Presentation: Asking and Answering Statistical Questions of Our Own Data

Available data sets:

1. Food Habits
2. Time Use
3. Stress/Chill
4. Personality Color

Your mission is to ask and answer a statistical question using at least one data set above.

1. Your question must include a comparison of two distinct groups.
2. Your analysis should address whether any observed differences are real or could be simply due to chance.
3. You should use at least two of the following methods to answer your question with appropriate explanations:
    - ❑ Merge data
    - ❑ Create simulations
    - ❑ Calculate probabilities based on simulations
    - ❑ Use a Normal model
    - ❑ Shuffle/permute data

You will have 5 days to complete this project with your assigned partner. You need to:

- ❑ Prepare an oral presentation (both partners need to participate) that includes:
    - A 4-slide, 5-minute presentation
    - An explanation of why you think your statistical question is interesting.
    - An interpretation of supporting plots and summaries that answer your question.
    - A reasoning of whether you think the outcome might be due to chance.

- ❑ Submit a 2 -4 page typed, double-spaced summary of your analysis.

<u>Project Assignment Sequence</u>:

- ❑ Day 1: Decide on a statistical question with assigned partner; get approval from teacher
- ❑ Day 2: Working day for analysis – create plots and numerical summaries
- ❑ Day 3: Working day for analysis – create presentation (4 slides maximum)
- ❑ Day 4: Presentations
- ❑ Day 5: Presentations