Using Dynamic Applets to Develop Statistical Thinking

I. Teens and Cyberbullying *Inference for random sampling*

From April 14 to May 4, 2022, the Pew Research Center surveyed a random sample of 1,316 U.S. teenagers about their use of digital devices, social media, and online platforms. One question asked of all survey respondents was:

> Thinking about your experiences online or on your cellphone, which of the following, if any, has ever happened to you personally?
> a. Been called offensive names
> b. Been threatened with physical harm
> c. Had someone spread false rumors about you
> d. Had someone share explicit images of you without your consent
> e. Had someone send you explicit images you did not ask for
> f. Had someone, other than a parent, constantly ask where you were, who you were with or what you were doing

Among the 1,316 teens who responded to the survey, 46% report having experienced at least one of these cyberbullying behaviors. (*Source:* Pew Research Center, December 2022, "Teens and Cyberbullying 2022".)

Let p = the proportion of all U.S. teens who would say they have experienced cyberbullying behavior.

- What is our "best guess" for the value of the population proportion p based on the sample data?

- Do you believe that the population proportion p is <u>exactly equal</u> to this value? Why or why not?

- Do you believe that the population proportion p is <u>close</u> to 0.46? Why or why not?

- BIG QUESTION: What values of p are <u>plausible</u> based on the sample result?

Is $p = 0.50$ a plausible value of the population proportion? (Don't answer yet!)

1. Point your browser to www.stapplet.com. Then launch the *1 Categorical Variable, Single Group* applet.
2. Type "Cyberbullying experience" for the variable name and choose to input the data as Counts in Categories.
3. Enter "Yes" and "No" as the category names, and input the corresponding frequencies.
4. Click the Begin analysis button. A bar graph of the data should be displayed, along with summary statistics. Change the graph to display relative frequency.
5. Scroll down to the Perform Inference section. Select the inference procedure "Simulate sample proportion" and choose "Yes" as the Category to indicate a success. (*Note:* If you only see the option to "Simulate sample percent", click on the link to "Adjust color, rounding, and percent/proportion preferences", choose the option to Display proportions as decimals, and click Done at the bottom of the window.)
6. Choose the option to Simulate the distribution of the sample proportion for samples of the original size assuming that the true proportion is equal to a hypothesized value. Then type 0.50 as the hypothesized proportion.
7. Input 1 as the Number of samples to add, then click the Add samples button. What just happened?

Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)

8. Click the Add samples button 9 more times, so that you have a total of 10 simulated sample proportions. Make a guess about what the distribution of sample proportions will look like if you keep adding more samples.

   Shape: Center (Mean): Variability (SD):

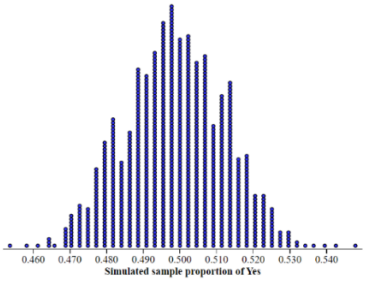
9. Change the Number of samples to add to 990 and click the Add samples button. You should now have a distribution with 1000 simulated sample proportions. Describe the simulated sampling distribution of $\hat{p}$.

   Shape: Center (Mean): Variability (SD):

10. What percentage of dots on the graph have values ≤ 0.46 *or* ≥ 0.54? _____

    In other words, IF the true proportion of all U.S. teens who would say they have experienced cyberbullying behavior is $p = 0.50$, the estimated probability of obtaining a sample result at least as surprising (in either direction) as $\hat{p} = 0.46$ purely by chance due to random sampling is _____.

11. Based on the estimated probability in Question 10, is $p = 0.50$ a plausible value for the population proportion? Explain your answer.



Note: You have just carried out an informal significance test!

- *What values of p are plausible (believable) based on the sample result of 46%?* (Don't answer yet!)
  Option 1: Use simulation to classify other potential values of p as believable or not.
  Option 2: Answer a different question! If p is "close" to the sample result of 46%, what is the maximum distance we expect the sample proportion $\hat{p}$ to be from the population parameter (i.e., the *margin of error*)?

12. Choose the option to Simulate the distribution of the sample proportion for samples of the original size assuming that the true proportion is equal to the observed value. (This process is known as *bootstrapping*.)
13. Input 1 as the Number of samples to add, then click the Add samples button. What just happened?

14. Change the Number of samples to add to 999 and click the Add samples button. You should now have a distribution with 1000 simulated sample proportions. Describe the simulated sampling distribution of $\hat{p}$.

    Shape: Center (Mean): Variability (SD):

15. Use the results of your simulation to answer the following two questions.
    (a) What's the maximum distance we expect the sample result to be from the population parameter? That is, what is the estimated *margin of error* for this survey? (*Hint:* In a normal distribution, about 95% of values are within ____ standard deviations of the mean.)

    (b) What values of p = the proportion of all U.S. teens who would say they have experienced cyberbullying behavior are believable based on the sample result of 0.46?

Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)

Is talking on a cell phone while driving more distracting than talking to a passenger? David Strayer and his colleagues at the University of Utah designed an experiment to help answer this question. They used 48 undergraduate students as subjects. The researchers randomly assigned half of the subjects to drive in a simulator while talking on a cell phone, and the other half to drive in the simulator while talking to a passenger. One response variable was whether the driver stopped at a rest area that was specified by researchers before the simulation started. The table below shows the results:

		Distraction	
		Passenger	Cell phone
Stopped at rest area?	Yes	21	12
	No	3	12

1. Calculate the observed difference (passenger – cell phone) in the proportion of drivers who stopped at the rest area in the two groups.

2. Go to www.rossmanchance.com/applets/index2021.html and launch the Statistical Inference applet called *Two-way Tables* (look under Randomization test for categorical response).
3. Click the button to Enter table 2×2. Input the labels for the two columns (Passenger and Cell phone) and the labels for the two rows (Yes and No), as well as the numbers of successes and failures in the two groups. Then click on Use Table.
4. A segmented bar chart of the data from the study should appear, along with the observed difference in the proportion of successes for the two groups.
5. Do the data give *some* evidence that talking on a cell phone while driving is more distracting than talking to a passenger? Justify your answer.

BIG QUESTION: Is it plausible (believable) that there's really no difference in the effect of talking to a passenger versus talking on a cell phone on driving distraction, and that random chance alone produced the observed difference between these two groups? Or are the results of this study statistically significant? To find out, let's see what would happen if we randomly reassign the 48 people in this experiment to the two groups many times, *assuming the type of distraction received doesn't affect whether a driver stops at the rest area*.

Simulation of the random assignment

- If there is no treatment effect, then the values of the response variable (whether the driver stopped at the rest area) will be the same as in the original study for all subjects, irrespective of the random assignment.
- Randomly reassign the subjects to the treatments: 24 to talk with a passenger and 24 to talk on a cell phone.
- Calculate the difference (Passenger – Cell phone) in the proportion of drivers who stop at the rest area for the two groups.
- Repeat the simulation process many times to build a simulated distribution of $\hat{p}_{passenger} - \hat{p}_{cellphone}$.

6. On the right side of the applet screen, click on Show Shuffle Options. Be sure that Cards is chosen under Select display. Keep the Statistic as Difference in proportions.
7. Click the Shuffle button. What just happened?

Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)

8. Click the Shuffle Responses button 9 more times, so that you have a total of 10 simulated differences in sample proportions. Make a guess about what the distribution of shuffled differences in proportions will look like if you keep doing more shuffles.

   Shape: Center (Mean): Variability (SD):

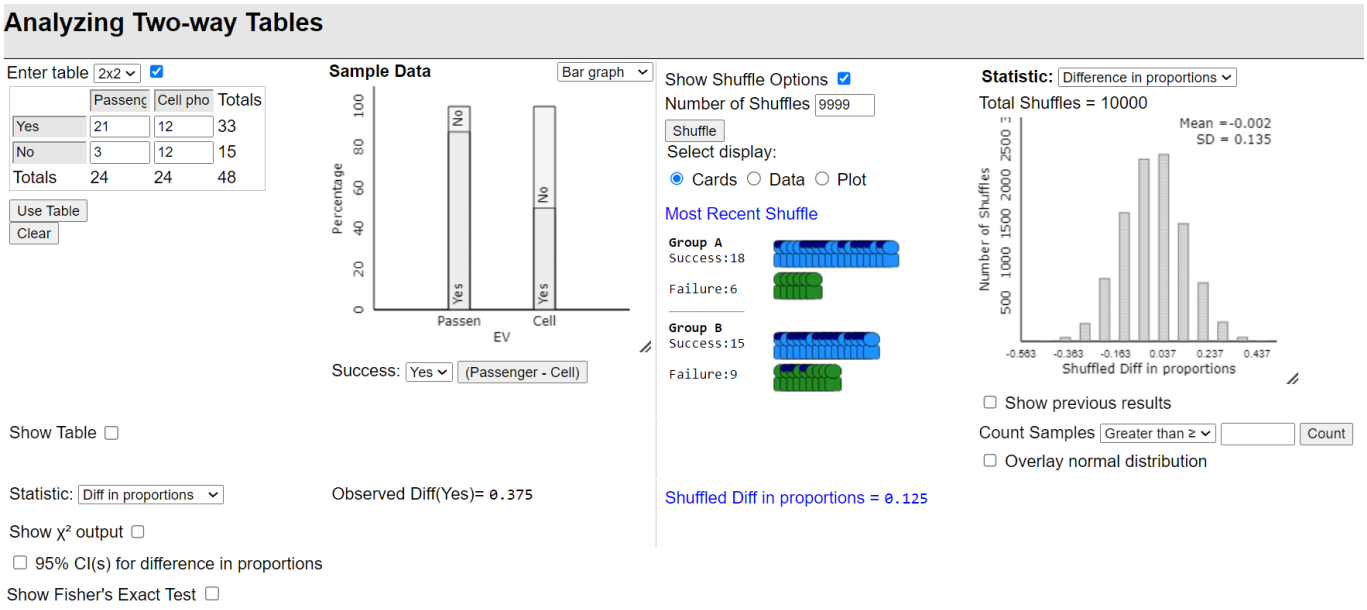
9. Change the Number of Shuffles to add to 9990 and click the Shuffle Responses button. You should now have a distribution with 10,000 simulated differences in sample proportions. Describe the simulated sampling distribution of $\hat{p}_{passenger} - \hat{p}_{cellphone}$.

   Shape: Center (Mean): Variability (SD):

10. What percent of dots on the graph have values greater than or equal to 0.375? _____

    In other words, IF the type of distraction received doesn't affect whether a driver like the ones in this study stops at the rest area, the estimated probability of obtaining a difference in sample proportions at least as surprising as $\hat{p}_{passenger} - \hat{p}_{cellphone} = 0.375$ purely by chance due to the random assignment is _____.

11. Based on the estimated probability in Question 10, do these data provide convincing evidence that talking on a cell phone while driving is more distracting than talking to a passenger for people like the ones in this study? Explain your answer.



Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)

Meadowfoam seed oil is used in making various skin care products. Researchers interested in maximizing the productivity of meadowfoam plants designed an experiment to investigate the effect of different light intensities on plant growth. The researchers planted 120 meadowfoam seedlings in individual pots, randomly assigned 10 pots to each of 12 trays, and put all the trays into a controlled enclosure. Two trays were then randomly assigned to each light intensity level (micromoles per square meter per second): 150, 300, 450, 600, 750, and 900. The number of flowers produced by each plant was recorded. Here are data on the average number of flowers per plant on each tray and the corresponding light intensity level.

Light intensity	150	150	300	300	450	450	600	600	750	750	900	900
Average number of flowers	62.3	77.4	55.3	54.2	49.6	61.9	39.4	45.7	31.3	44.9	36.8	41.9

(*Source:* M. Seddigh and G.D. Jolliff, "Light Intensity Effects on Meadowfoam Growth and Flowering," *Crop Science* 34 (1994): 497–503.)

1. Go to www.rossmanchance.com/applets/index2021.html. Choose Data Analysis→Least Squares Regression.
2. Open the Meadowfoam Experiment data file.
3. Copy and paste the data into the Enter Data panel.
4. Click the Use Data button. A scatterplot of the data should appear.
5. Do the data give *some* evidence of a linear relationship between light intensity and average number of flowers produced for meadowfoam plants? Explain your answer.

6. Click the button to Show Regression Line. Identify and interpret the slope of the sample regression line.

BIG QUESTION: Do these data provide *convincing* evidence of a linear relationship between light intensity and average number of flowers for meadowfoam plants like the ones in this experiment? Or is it believable that there is not actually a linear relationship between the two variables, and that the observed result ($b = -0.041$) occurred purely by the chance involved in the random assignment?

To find out, let's simulate re-doing the random assignment under the assumption that there is no relationship between average number of flowers and light intensity. This amounts to assuming that each tray would have the same average number of flowers regardless of light intensity, and randomly reassigning the trays to the various light intensities, recording the slope b of the resulting sample regression line each time.

7. On the right side of the applet screen, click the box to Show Shuffle Options. Be sure that Plot is chosen under Select display. In the Choose statistic drop-down menu at the top right, select Slope.
8. Click the Shuffle Y-values button. What just happened?

Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)

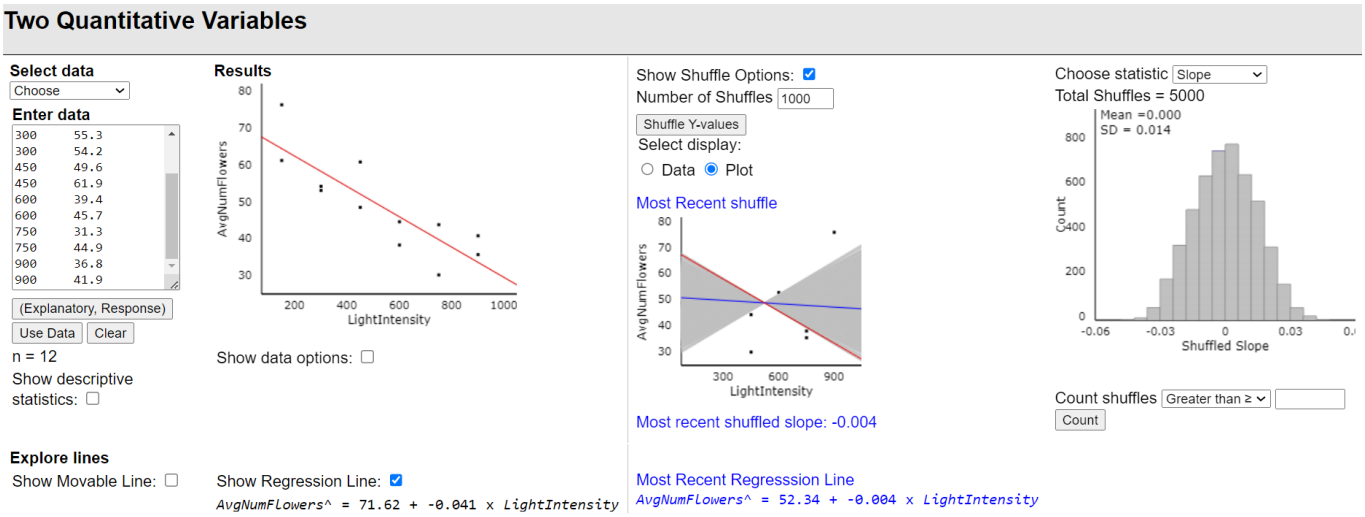
9. Change the Number of Shuffles to 1000 and click Shuffle several times. What do you notice about the simulated distribution of the sample slopes?

   Shape: Center: Variability:

10. What percent of dots on the graph have values at least as extreme (in either direction) as −0.0411? _____ (Use the Count shuffles Beyond feature under the graph to help answer this question.)

11. In other words, IF there is actually no linear relationship between light intensity and average number of flowers produced in the population of meadowfoam plants like the ones in this study, the estimated probability of obtaining a sample regression line with a slope at least as surprising (in either direction) as −0.0411 purely by chance due to random assignment is _____.

12. Based on the estimated probability in Question 11, do these data provide convincing evidence of a linear relationship between light intensity and average number of flowers produced in the population of meadowfoam plants like the ones in this study? Explain your answer.



Daren Starnes (StarnesStats@gmail.com) and Josh Tabor (TaborStats@gmail.com)