The goals of this lesson are to introduce students to the logic of statistical inference. This lesson often comes very early in my course and we build in the formal process and terminology gradually over the next few lessons. In particular, I start with a lesson where the hypothesized probability is 0.50, but will illustrate a different scenario in this workshop. Such a lesson can be used at different points in the course to help students develop inferential thinking. So the lesson below is a mash up of a few different lessons I've used and is very adaptable to a different research context.

In my course I also tend to have half "regular classroom days" and half "lab days." The lab days are preceded by a pre-lab assignment. The regular classroom days are typically a guided activity introducing ideas that students will apply during the lab day. In the workshop, I'll present a lesson designed for a regular classroom day, but I would like you to complete the following pre-lab assignment.

Pre-Lab for Can Dogs Detect Covid (due before class)
Search online (e.g., here) for a news story about dogs ability to detect cancer from breath samples. Typically a dog would be given several "specimens," one of which corresponded to a patient who had tested positive for COVID. The dog would be trained to sniff the specimens and then sit down next to the COVID specimen. The dog's performance would be tested over several trials, with the location of the COVID specimen randomly determined each time. We will focus on one dog, Maika, a 3-year-old female Belgian Malinois whose specialty is search and rescue. Maika completed 57 trials where she would sniff four different sweat specimens, one of which was from a COVID positive person.
(a) Identify the observational unit in this study.
(b) Identify the variable of interest in this study.
(c) Is this a quantitative or categorical variable?
(d) Identify the parameter of interest in words.
(e) If Maika is not able to consistently detect the COVID sample but instead is guessing each time, what value would describe Maika's probability of identifying the correct specimen?

Currently I administer the pre-lab as multiple choice in Canvas. This allows me to take a quick look at student responses before I get to class to help identify any issues that need to be cleared up before students begin the lab activity. Students can also refer back to the correct answers if I ask similar questions in the lab as warm-up. The main goals are to have students engage in the study context so we can begin the class period by analyzing the data rather than reading the study background. Further along in the course I may expect them to proceed further into the study (e.g., run some of the descriptive statistics) before class so we can focus on the new material when we are together (but still maintaining the focus on the 6-step process). Students also learn that they are expected to complete the lab activity between class periods (often with a regular classroom day in between).

Day 4/5 – Making inferences/Statistical Significance
Can dogs detect COVID

Previously:

- *Probability* refers to the long-run proportion of times an outcome occurs in repeated trials under identical conditions
- Key terminology: obs unit, variable, statistic, sample, parameter, population/process
- Describing distribution of quantitative variable: shape, center, variability, unusual observations
    - With bell-shaped distributions, approximately 95% of observations fall within 2SD of mean

This is a variation of a STUB activity for use as a regular classroom activity. Students are provided with the word doc for recording their notes. Google doc version here

Example 1: Dogs have a keen sense of smell. They are used for search and rescue, explosive detection, sniffing out illegal drugs in luggage at airports, and locating game while hunting. Can they also tell whether someone has COVID-19 by sniffing a specimen of sweat from a person? We will be looking at a study that used several dogs to test this question. We will focus on one dog, Maika, a 3-year-old female Belgian Malinois whose specialty is search and rescue. Maika completed 57 trials where she would sniff four different sweat specimens, one of which was from a COVID positive person, and then sit in front of the specimen she determined to be the positive specimen. The researchers found that in 47 of the 57 trials Maika chose the COVID positive specimen.

(a) Identify the observational unit, variable (quantitative or categorical?), statistic, sample, parameter, and population.

Option: (b) Produce a bar graph representing the observed results. Summarize what you learn.

(c) Do these results convince you that dogs can detect COVID-19. Explain your reasoning.

(d) If Maika does not have some ability to identify the COVID specimen from the sweat of a person, what is another explanation for her getting so many correct?

There are two *possible* explanations for this tendency:

- In the long-run, Maika has a genuine tendency to correctly identify the COVID specimen
- In the long run, Maika does not have a genuine tendency and is simply guessing among the specimen each time (she got lucky)

(e) Suppose a friend thinks Maika just got lucky. How could you convince your friend otherwise? OR How likely do you think it is that Maika was able to identify so many correctly if she was simply guessing each time.

Our goal now is to decide whether the second explanation is *plausible*. To determine this, we need to know what pattern of results we might expect to see when Maika is merely guessing among the 4 specimens each time. We can *model* Maika's choices corresponding to the second explanation by using *simulation*.

(f) Suggest a method for modeling Maika's results assuming she is guessing each time.

To see what the results look like if Maika is guessing, let's first assume she was only give 5 trials rather than 57.

(g) Create a spinner (e.g., paper clip and pencil) that is _____% purple (success) and _____% blue (failure). Spin 5 times, how many times did it land in the purple section?

(h) Did everyone in class obtained the same result for their guessing dog? Why or why not?

In the One Proportion applet

- Specify the hypothesized probability of success _____
- Specify the sample size:_____
- Press **Draw Samples**

(i) Combine your simulation results for each repetition with your classmates on the axis below.

What are the observational units *in this graph*? What is the variable *in this graph*?

(j) Approximate the mean of your results. Does this value make sense? Explain.

(k) Based on these simulation results, would you consider it surprising for Maika to correctly identify the COVID specimen 4 times? Explain your reasoning.

(l) Now consider the actual study results – 47 out of 57 attempts. How do you think this will change the distribution from (i)? How do you think it will judge your judgment of whether Maika's result is surprising for a guessing dog?

(m) In the applet, change the Sample size (*n*) to **57** and the Number of Samples to **100**. Press **Draw Samples** until the distribution stops changing much, and you think you have a pretty good idea of the long-run pattern of these results. How many repetitions did you use? Sketch your graph below. How accurate were your predictions in (l)?

(n) Based on your graph in (m), do you think it's plausible that Maika was guessing each time and chose correctly 47 out of 57 times? How are you deciding?
OR
(n) Based on your answer to (m), which explanation appears more *plausible* to you?

- In the long-run, Maika has a genuine tendency to correctly identify the COVID specimen
- In the long run, Maika does not have a genuine tendency and is simply guessing among the specimens each time (she got lucky)

Explain how you are deciding as if to a friend not in a statistics class.

Exploration 1.2: Can dogs smell COVID?

LEARNING GOALS

- Applying the 6-step statistical investigation process to a research question about a single categorical variable
- Use the **One Proportion** applet to obtain the p-value after carrying out an appropriate simulation.
- Interpret the p-value.
- State a conclusion about the research question based on the p-value.

Dogs have a keen sense of smell. They are used for search and rescue, explosive detection, sniffing out illegal drugs in luggage at airports, and locating game while hunting. Can they also tell whether someone has COVID-19 by sniffing a specimen of sweat from a person? We will be looking at a study that used several dogs to test this question. We will focus on one dog, Maika, a 3-year-old female Belgian Malinois whose specialty is search and rescue. Maika completed 57 trials where she would sniff four different sweat specimens, one of which was from a COVID positive person, and then sit in front of the specimen she determined to be the positive specimen.

Names: >>

STEP 1: State the research question.

1. What is the research question that the researchers hoped to answer?

>>

STEP 2: Design a study and collect data.

2. Identify the observational unit in this study.

>>

3. Identify the variable. Is the variable quantitative of categorical?

>>

4. Describe the parameter of interest in this study (in words). (*Hint*: The parameter of interest is the long-run proportion of …?)

>>

5. One possibility here is that Maika can't smell COVID and is equally likely to choose any of the four scent specimens as the COVID positive specimen, essentially selecting one of the four specimens at random. In this case, what is the long-run proportion (i.e., probability) that Maika selects the COVID positive specimen in any particular attempt?

>>

6. Another possibility is that Maika can smell COVID and is more likely to select the COVID positive specimen than if she was randomly guessing. In this case, what can you say about the long-run proportion of times Maika selects the COVID positive specimen? (*Hint*: You are not to specify a particular value at this time, instead indicate a direction from a particular value.)

>>

Definitions
The ***null hypothesis*** typically represents the "by-random-chance-alone" explanation. The chance model (or "null model") is chosen to reflect this hypothesis.
The ***alternative hypothesis*** typically represents the "there is an effect" explanation that contradicts the null hypothesis. Researchers typically hope this hypothesis will be supported by the data they collect.

7. Your answers to #6 and #7 should be the null and alternative hypotheses for this study. Which is which?

>>

STEP 3: Explore the data.

The researchers found that in 47 of the 57 trials Maika chose the COVID positive specimen.

8. Calculate the value of the relevant statistic.

>>

STEP 4: Draw inferences.

10. Is the sample proportion of correct identifications in this study larger than the probability specified in the null hypothesis?

>>

11. Is it possible that this proportion could turn out to be this large even if the null hypothesis was true? (i.e., even if Maika couldn't smell COVID and was essentially selecting at random from the four specimens)?

>>

We will use simulation to investigate how surprising the observed sample result (47 of 57 correct COVID identifications) would be if in fact Maika could not smell COVID and so for each trial had a 0.25 probability of selecting the COVID specimen. (Note also that our null model assumes the same probability for each trial.)

Think About It

Can we use a single coin toss to represent the chance model specified by the null hypothesis? If not, can you suggest a different random device that we could use? What needs to be different about our simulation?

15. Explain why we cannot use a simple coin toss to simulate Maika's choices, as we did with a 50-50 chance of success.

>>

16. We could do the simulation using a set of four playing cards: one black and three red. Explain how the simulation would work in this case.

>>

17. Another option would be to use a spinner like the one shown here, like you would use when playing a child's board game. Explain how the simulation would work if you were using a spinner. In particular:
    a. What does each region represent? >>

    b. How many spins of the spinner will you need to do in order to simulate one repetition of the experiment when there is equal preference between the four specimens (null hypothesis is true)?

>>

18. We will now use the **One Proportion** applet to conduct this simulation analysis. Notice that the applet will show us what it would be like if we were simulating with spinners.
    a. First enter the **probability of heads/probability of success** value specified in the null hypothesis.
    b. Enter the appropriate **sample size** (number of Maika's trials in this study).
    c. Keep 1 for the number of samples, and press **Draw Samples**. Report the number of "successes" in this simulated sample.
    d. Now, select the radio button for "Proportion of successes." What value on the Proportion of successes graph is this simulated sample proportion of success close to? Use your answer to "c" to verify how this simulated value is calculated.
    e. Leaving the "Proportion of successes" radio button selected, click on **Draw Samples** four more times. Do you get the same results each time?
    f. Now enter 995 for the number of samples and click on **Draw Samples**, bringing the number of simulated samples to 1,000. Comment on the center, variability, and shape of the resulting distribution of sample proportions.

This distribution of simulated sample proportions is called the ***null distribution***, because it is created assuming the null hypothesis to be true.

>>

19. Recall that the observed value of the sample proportion of correctly identified COVID specimens in this study was = 47/57 ≈ 0.83. Looking at the null distribution you have simulated, is this a

very unlikely result when the null hypothesis is true? In other words, is this value far in the tail of the null distribution?

>>

In this case, the observed statistic is far out in the tail of the distribution and it is not hard to see that Maika's proportion of successful identifications is unlikely to happen by random chance. There will be studies when the observed statistic is not that far in the tail of the distribution, but also not near the middle of the distribution (e.g., what if Maika had been correct 35% of the time). To help make a judgement about strength of evidence in this case, we can count how many (and what proportion) of the simulated sample proportions are as extreme or more extreme than the observed value.

20. Use the applet to count how many (and what proportion) of the simulated sample proportions are as or more extreme than the observed value. Make sure that the ≥ inequality symbol is selected (to match the alternative hypothesis). Then enter **0.83** (the observed sample proportion of correct COVID positive identifications) in the box to the left of the **Count** button. Then click on the **Count** button. Record the number and proportion of simulated sample proportions that are as extreme or more extreme than the observed value.

>>

Definition

The ***p-value*** is the probability of obtaining a value of the statistic at least as extreme as the observed statistic when the null hypothesis is true. We can estimate the *p*-value by finding the proportion of the simulated statistics in the null distribution that are *at least as extreme* (in the direction of the alternative hypothesis) as the value of the statistic actually observed in the research study.

How do we *evaluate* this *p*-value as a judgment about strength of evidence provided by the sample data against the null hypothesis? One answer is: The smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis. But how small is small enough to regard as convincing? There is no definitive answer, but here are some guidelines:

Guidelines for evaluating the strength of evidence from p-values

0.10 < p-value	not much evidence against null hypothesis; null is plausible
0.05 < p-value ≤ 0.10	moderate evidence against the null hypothesis
0.01 < p-value ≤ 0.05	strong evidence against the null hypothesis
p-value ≤ 0.01	very strong evidence against the null hypothesis

The smaller the p-value, the stronger the evidence against the null hypothesis.

21. Is the approximate p-value from your simulation analysis (your answer to #20) small enough to provide convincing evidence against the null hypothesis that Maika was just guessing which of the four specimens was COVID positive? If so, how strong is this evidence? Explain.

>>

22. When computing p-values, “more extreme” is always measured in the direction of the alternative hypothesis. Use this fact to explain why you counted ≥ 0.83 in #20.

>>

STEP 5: Formulate conclusions.

23. Do you consider the observed sample result to be *statistically significant*? Recall that this means that the observed result is unlikely to have occurred by chance alone.

>>

24. How broadly are you willing to generalize your conclusions? Would you be willing to generalize your conclusions to all dogs? Explain your reasoning.

>>

STEP 6: Look back and ahead.

25. Suggest a new research question that you might investigate next, building on what you learned in this study.

>>

Example Pre-Lab (Multiple Choice)

In 2011, an article published by the medical journal *Gut—An International Journal of Gastroenterology and Hepatology* (**Sonoda et al** .) reported the results of a study conducted in Japan in which a dog was tested to see whether she could detect colorectal cancer. The dog used was an eight-year-old black Labrador named Marine. The study was designed so that the dog first smelled a bag that had been breathed into by a patient with colorectal cancer. This was the standard that the dog would use to judge the other bags. Marine then smelled the breath in five different bags from five different patients, only one of which contained breath from a different colorectal cancer patient; the other four bags contained breath from noncancer patients. The dog was then trained to sit next to the bag which she thought contained breath from a cancer patient (i.e., had the cancer scent). If she sat down next to the correct bag, she was rewarded with a tennis ball.
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3095480/)
Marine completed 33 attempts of this experimental procedure, with a different set of five patients each time: four noncancer patients and one cancer patient, and was correct 30 times.

1) Which of the following is the best definition of the *parameter* for this study?
- Marine's probability of picking the correct breath sample
- 30/33
  This is the statistic, not the parameter
- 33
  This is the sample size, not the parameter
- All possible trials of picking a breath sample
  This is the population, not a numerical summary of the population

2) Which of the following would be the most appropriate *null hypothesis?*
- Marine is simply guessing among the 5 bags
- Marine has a genuine tendency to identify the cancer breath
- Marine picks the correct breath sample half the time in the long run
- Marine picked the correct breath sample 30 of 33 times

3) Which of the following would be the most appropriate *alternative hypothesis?*
- Marine is simply guessing among the 5 bags
- Marine has a genuine tendency to identify the cancer breath
- Marine picks the correct breath sample half the time in the long run
- Marine picked the correct breath sample 30 of 33 times

Consider a simulation of the chance model (Marine is just guessing among the 5 breath bags)
4) How many attempts would we simulate in one repetition of the chance model?
  33 (with margin: 0)

5) Could we use coin flips to represent the chance model?
- Yes, because we have success (correct breath bag) and failure
- Yes, because the chance model has a long run probability of 0.50
- No, because if Marine is just guessing, the probability of identifying the correct bag differs from 0.50
- Yes, because we have repeat observations on the same dog

Example Post-Quiz

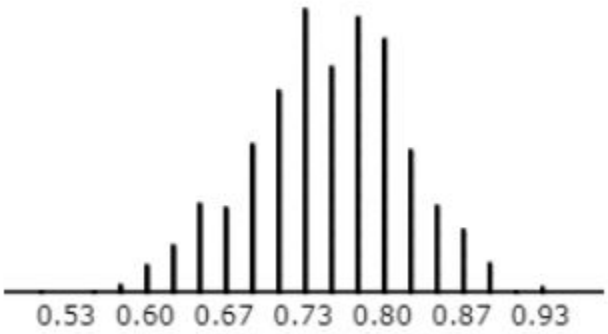
A reader wrote in to the “Ask Marilyn” column in *Parade* magazine to say that his grandfather told him that in 3/4 of all baseball games, the winning team scores more runs in one inning than the losing team scores in the entire game. (This phenomenon is known as a “big bang.”) Marilyn responded that this proportion seemed to be too high to be believable. To investigate the grandfather's claim, a statistics professor examined the 45 Major League baseball games played one weekend; he found 21 contained a big bang.

1) Calculate the sample proportion of games with a big bang
- 0.75
- 0.467
- 21
- 45
- 0.5

2) Is the number you found in Question 1 a parameter or a statistic?
- parameter
- statistic

Suppose we plan to test the following hypotheses:
Ho: the probability a Major League Baseball game includes a big bang is 0.74 (grandpa's right)
Ha: the probability a Major League Baseball game includes a big bang is less than 0.74 (Marilyn's right)

3) Below are 1000 “could have been” samples from a process assuming the grandfather’s claim is true.



This graph displays the distribution of what *variable*?

- Whether or not the game had a "big bang" inning
- The long-run probability an MLB game has a big bang
- Sets of 45 games
- Sample proportion with big bang

4) Based on the above simulated null distribution, would you consider the results of this study to be convincing evidence that the probability a Major League Baseball game has a big bang is less than 0.75?

- No, the distribution is centered at 0.75
- Yes, the p-value is going to be small
- No, the p-value is going to be small
- No, the p-value is going to be large
- Yes, the p-value is going to be large
- Yes, the distribution is centered at 0.75

5) Which of the following is the most appropriate interpretation of the p-value?

- The probability of a big bang
- The probability that the grandfather is right
- The probability of 21 or fewer games with a big bang if the long-run probability = 0.75
- The probability of 21 games with big bang if the long-run probability = 0.50
- The probability that Marilyn is correct