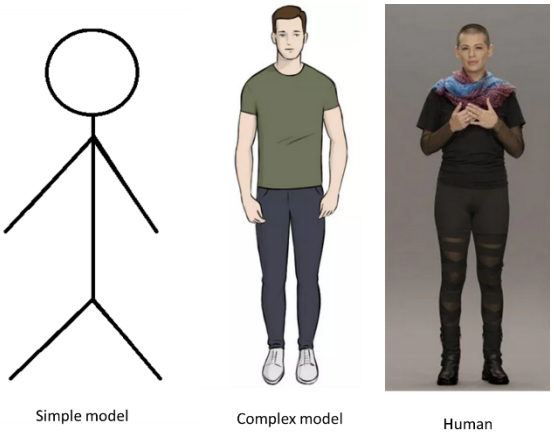
Exploration 1.2: Can dogs smell COVID?

Measuring the Strength of Evidence

LEARNING GOALS

- Develop a simple statistical model using verbal description
- Use appropriate statistical terminology for describing the model.
- State the null and alternative hypotheses in words and in terms using mathematical symbols.
- Develop a simple statistical model based on a null hypothesis
- Use the **One Proportion** applet to simulate the model and calculate a p-value for observed data
- Interpret the p-value as a probability of null hypothesis
- Evaluate validity of the model and identify criteria for rejecting the null hypothesis
- State a conclusion about the null and alternative hypothesis based on the p-value.

Life is complicated, but we often do not need to know all the details and can boil down the process to its most essential elements. This process in science usually referred as developing *a model*. Those simplifications are based on a set of rules, called *assumptions*. For example, a stick figure is a model of a human body, where we assume that the head is circular, and arms and legs are linear. It is very rude simplification, but you all will recognize what it represents. Adding more assumptions makes the model more complex. A common aphorism states: "All models are wrong, but some are useful." That is, no matter how sophisticated, all models are approximations of something real. While they are not the "real thing" (and are thus wrong), models are useful when they allow us to make predictions about real life that we can use.



Simple model | Complex model | Human

In Statistics, we refer to a model as a set of mathematical rules (assumptions) which allows us to generate random observations. For example, we may develop a model of flipping a coin by assuming that both outcomes (heads or tails) are equally likely. We then proceed of generating a set of random numbers, say zeros and ones, where zero represent a head and one represent a tail. Such a model, often referred as a chance model, is often used to simulate the process with equally likely outcome. However, not all situations call for a coin-flipping model. The coin-flipping model would work well to simulate guessing on a true-false test with a 50% chance of getting a question correct, but what about a multiple-choice test where you are guessing between four possible answers with a 25% chance of getting a question correct? Computers can easily generate a set of four random numbers (say from zero to three), so we can calculate how many of of each we tend to get by random chance alone.

We use chance models to will apply the 6-step statistical investigation method in the more general setting of an arbitrary probability of "success." We will also learn helpful terminology that is commonly used in statistical investigations to describe the process of drawing conclusions from data. We will work toward formalizing the procedure of a *test of significance* and give some guidelines to help determine when we have strong evidence that our chance model is not correct. We will also introduce some symbols, for convenience, but the big picture of assessing evidence against a claim is the same.

We will look at a study that tests a dog's sense of smell. 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? Before class you have read the Jendrny et al, 2020 article which describes a sample protocol for COVID-19 sample presentation to a dog. We will be looking at a study that used several dVarogs 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 would sit in front of the specimen she determined to be the positive specimen. You can watch a video of a trial on the article website.

STEP 1: State the research question.

1. What is the research question that the researchers hoped to answer? We may not expect Maika to be correct 100% of the time - how often would she need to be right to convince you see wasn't simply guessing and getting lucky? Consult with your teammates and formulate concise research question which the Maika trial can address. Write your question in a box below.

The research question is: can Maika consistently detect covid in sweat samples?

STEP 2: Design a study and collect data.

2. Identify the variable(s) of interest. Is the variable(s) quantitative of categorical? Is it a **binary** variable?

In a simple model, there would be one variable: if the dog correctly identifies if the sample is positive or negative for Covid. This is a binary categorical variable. Its outcome can be 'yes' or 'no.'

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

If Maika is randomly choosing a sample out of the 4 presented, there will be a 1/4th, or 25%, chance of choosing the correct (positive sample.)

4. 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 proportion of times Maika selects the COVID positive specimen? (*Hint*: You are not to specify a particular value at this time, instead describe a trend in the variable which you would observe)

If Maika can smell COVID, she will select the COVID positive specimen more than 25% of the time.

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.

5. Some of your previous answers contains (at least in part) null and alternative hypotheses for this study. Which is which? In a space below please state, in your own words, the null and alternative hypotheses for Maika's specimen presentation.

Null Hypothesis: When presented with four samples, Maika will randomly identify the COVID-positive sample.

Alternative Hypothesis: When presented with four samples, Maika will identify the COVID-positive sample non-randomly.

STEP 3: Explore the data.

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

Definition
Descriptive statistics are brief numerical metric that summarize a given data set, which can be either a representation of the entire population or a sample of a population.

6. What would be a relevant statistic for this trial (in words)? Calculate the value of the relevant statistic. Provide one-sentence interpretation of the calculated value of the relevant statistics.

The relevant statistic for this experiment would be the proportion of trials where Maika correctly selected the covid-positive sweat specimen.

Use of Symbols

We can use mathematical symbols to represent quantities and simplify our writing. In class notes and in the course textbook we will emphasize written explanations but will also show you mathematical symbols which you are free to use as a short-hand once you are comfortable with the material. The distinction between parameter and statistic is so important that we always use different symbols to refer to them.

When dealing with the probability that Maika would choose the COVID positive specimen, we use the Greek letter π (pronounced "pie"). But when working with a statistic that is the proportion of "successes" in a sample, such as the proportion of trials where Maika *did* choose the COVID positive sample, we use the symbol $\hat{p}$ (pronounced "p-hat"). Finally, we use the symbol n to represent the sample size.

7. What is the value of $\hat{p}$ in this study? What is the value of n in this study?

p hat = 47, n = 57

8. Hypotheses are always conjectures about the unknown parameter. You can also use H_0 and H_a as short-hand notation for the null and alternative hypotheses, respectively. A colon, ":", is used to represent the word "is." Restate the null and alternative hypotheses using π to represent the unknown probability that Maika will choose the positive specimen.

H0: $\pi = 0.25$
Ha: $\pi =/= 0.25$

STEP 4: Draw inferences.

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

The sample proportion of correct identifications in the study is greater than the probability from the null hypothesis. If there are no interfering factors, it is extremely unlikely, that this result occurred while the null hypothesis is true.

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.)

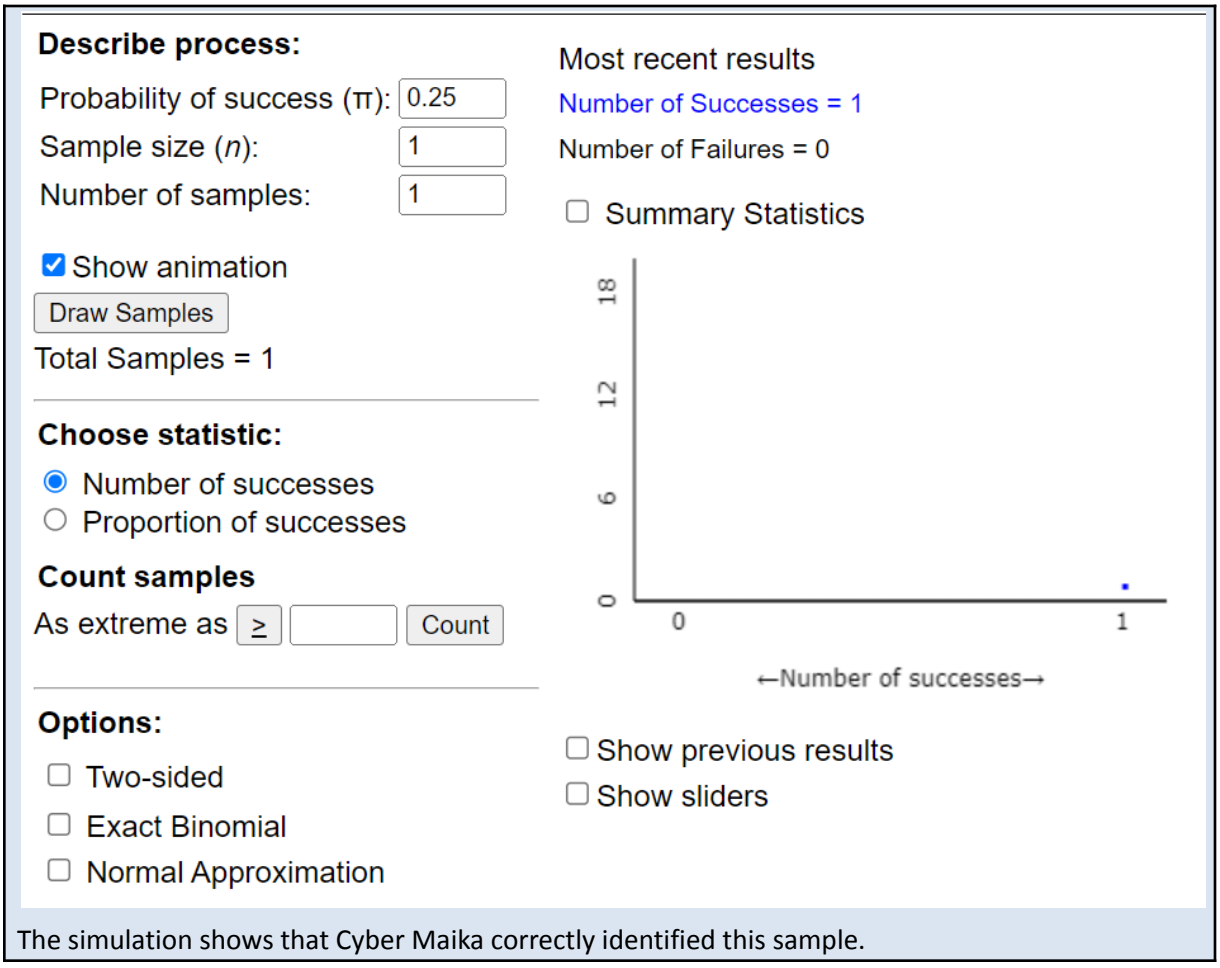
Think About It

We will develop a model of Maika (let's call it Cyber-Maika) using our null hypothesis as the assumption and then simulate the Cyber-Maika model many-many times to see if she can have success proportion of COVID sample detection as high as Maika.

We will simulate Cyber-Maika using one simple assumption provided by our null hypothesis:
When offered 4 specimens, Cyber-Maika select one of the specimens completely randomly
In other words, we will simulate many typical values of the sample proportions for Cyber-Mika who we KNOW is simply guessing each time (because that's the model we create).

10. We will now use the **One Proportion** applet to conduct this simulation. You already familiarized yourself with the basic functionality of this applet before the class so now you can consult with your teammates and simulate the following scenario:

    a. Single Trial. Cyber-Maika is offered 4 specimens and chooses one completely randomly
        1. Run this simulation in app. Give Cyber-Maika model one attempt with 25% chance of success. Insert the screenshot of your simulation in a box below. Did Cyber-Maika correctly identify the COVID specimen?



The simulation shows that Cyber Maika correctly identified this sample.

11. Now give Cyber-Maika 57 attempts at choosing correctly out of 4 specimen.
    a. Run the appropriate simulation. Insert the screenshot of the simulation in the box below. How many times did Cyber-Maika choose correctly. Are you surprised with this number of successes? Why/why not?

Describe process:

Probability of success (π): 0.25

Sample size (*n*): 57

Number of samples: 1

☑ Show animation

Draw Samples

Total Samples = 1

Choose statistic:

◉ Number of successes

○ Proportion of successes

Count samples

As extreme as ≥ Count

Options:

☐ Two-sided

☐ Exact Binomial

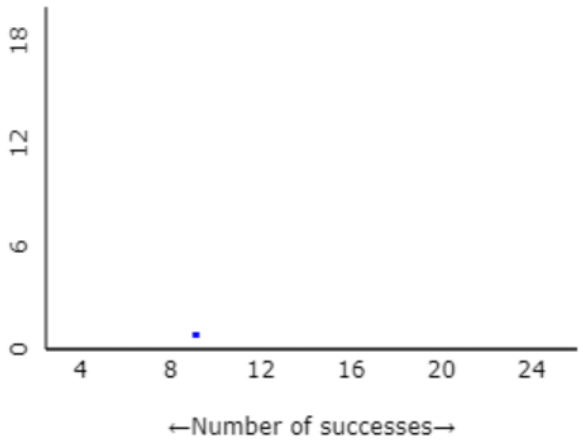
☐ Normal Approximation

Most recent results

Number of Successes = 9

Number of Failures = 48

☐ Summary Statistics



☐ Show previous results

☐ Show sliders

Out of 57 attempts, Cyber Maika correctly identified 9 samples. P(Success)*n=14.25. Since Maika correctly identified 9 samples, and 9<14.25, Maika was not as successful as predicted using this model. This is not surprising, since successful identification occurs merely by chance and our sample size is not very large.

12. Now we will run our model 10 times. That is, we will simulate 10 dogs like Cyber-Maika. Our dogs are tested one at a time. Since we simulate our dogs by generating random numbers, every dog will have slightly different number of successes in their 58 trials (although all of them selecting completely randomly).
    a. Run your simulation 10 times. Each repetition will be plotted as a separate dot on the plot on the right. Insert a screenshot of your simulation in the box below. Comment on your worst and the best trial. How expected/unexpected are those results under the chance model?

Describe process:

Probability of success (π): 0.25

Sample size (*n*): 57

Number of samples: 10

☑ Show animation

Draw Samples

Total Samples = 10

Choose statistic:

◉ Number of successes

○ Proportion of successes

Count samples

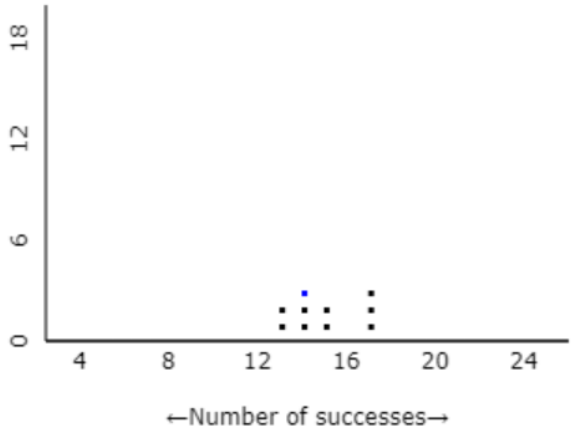
As extreme as ≥ Count

Most recent results

Number of Successes = 14

Number of Failures = 43

☐ Summary Statistics



Under the chance model(null hypothesis) these results make sense, as all of the simulated dogs were within 3 successes of 14, the predicted value. The best trials were those with 17 successes and the worst trials were those with 13 successes, both far worse than how Maika actually performed.

When you repeat your simulation many times, you can generate a frequency distribution of almost all potential outcomes. This distribution of simulated sample proportions is called the ***null distribution***, because it is created assuming the null hypothesis to be true.

Simulate 1000 dogs like Cyber-Maika. In the box below insert two graphs of null distributions for :

- Number of Successes
- Proportion of Successes

What is the most common number of success? What is the most common proportion of successes?

Number of successes:

Describe process:

Probability of success (π): 0.25

Sample size (*n*): 57

Number of samples: 1000

☑ Show animation

Draw Samples

Total Samples = 1000

Choose statistic:

◉ Number of successes

○ Proportion of successes

Count samples

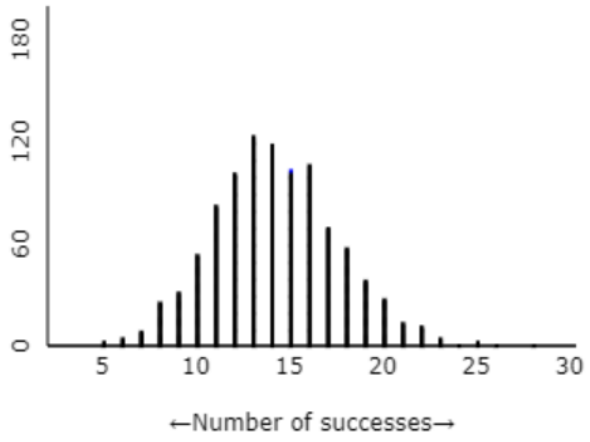
As extreme as ≥ Count

Most recent results

Number of Successes = 15

Number of Failures = 42

☐ Summary Statistics



The most common number of successes is 13, closely followed by 14.

Proportion of successes:

Describe process:

Probability of success (π): 0.25

Sample size (*n*): 57

Number of samples: 1000

☑ Show animation

Draw Samples

Total Samples = 1000

Choose statistic:

○ Number of successes

◉ Proportion of successes

Count samples

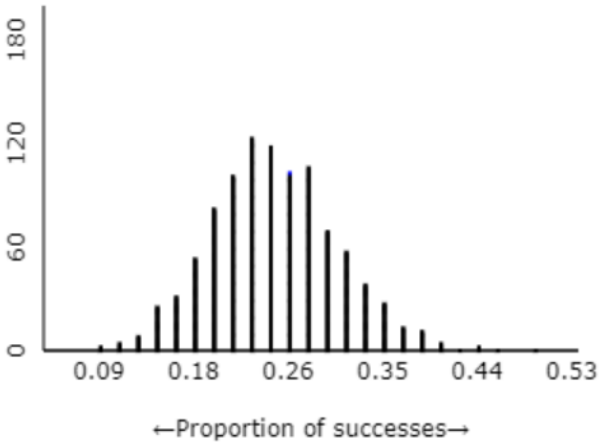
As extreme as ≥ Count

Most recent results

Number of Successes = 15

Number of Failures = 42

☐ Summary Statistics



The most common proportion of successes is about 0.23

13. Recall that the observed value of the sample proportion of correctly identified COVID specimen in this study was $\hat{p}$ = 47/57 ≈ 0.82. 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?

1. Use "**Count samples**" option of the applet to calculate how many samples in your distribution larger than 0.82

Out of the 1000 simulated trials, there are 0 samples above a 47 count (above a 0.82 proportion).

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.

14. Is the approximate p-value from your simulation analysis (your answer to previous question) 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.

The p value drawn from the simulation is 0/1000. With a p value less than or equal to 0.01, there is very strong evidence against the null hypothesis.

STEP 5: Formulate conclusions.

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

The result is statistically significant because there is virtually no chance it could have arisen randomly. Based on the study design, the result is conclusive, but only for the dogs in the study, which were specially trained detection dogs. It is very unlikely that these results can be replicated for all dogs.

STEP 6: Look back and ahead.

16. Suggest a new research question that you might investigate next, building on what you learned in this study. What kind of experiment would you design to test this question? List the variable(s) of the study and relevant statistics you will need to calculate.

A new research question could ask whether the age of detection dogs impacts their ability to accurately identify COVID positive sweat samples. This could be tested with a similar experiment with two or three sets of detection dogs of different age groups. The relevant variables in the study would be dog age and number of correct identifications. Again, the statistic calculated would be the proportion of correct identifications, and this value could be averaged and compared among the different age groups.