Activity 1: Martian Alphabet

Your Name: _____

September 20, 2022

Learning outcomes

- Describe the statistical investigation process.
- Identify observational units, variables, and variable types in a statistical study.

Terminology review

Statistics is the study of how best to collect, analyze, and draw conclusions from data. Today in class you will be introduced to the following terms:

- Observational units or cases
- Variables: categorical or quantitative
- Proportions
- Graphs: frequency bar plot and relative frequency bar plot
- Distribution

Can you read "Martian"?

How well can humans distinguish one "Martian" letter from another? In today's activity, we'll find out. When shown the two Martian letters, Kiki and Bouba, write down whether you think Bouba is on the left or on the right.

1. Were you correct or incorrect in identifying Bouba?

Steps of the statistical investigation process

Step 1: The first step of any statistical investigation is to *ask a research question.* In this study the research question is: Can we as a class read Martian? (We will refine this later on!).

Step 2: To answer any research question, we must *design a study and collect data.* For our question, the study consists of each student being presented with two Martian letters and asking which was Bouba. Your responses will become our observed data that we will explore.

Observational units or **cases** are the subjects data are collected on. In a spreadsheet of the data set, each *row* will represent a single observational unit.

2. What are the observational units in this study?

3. How many students are in class today? This is the **sample size**.

A **variable** is information collected or measured on each observational unit or case. Each *column* in a data set will represent a different variable. Today we are only measuring one variable on each observational unit.

4. **Identify the variable we are collecting on each observational unit in this study, i.e., what are we measuring on each student?** *Hint*: Your answer to question 1 is the outcome for the variable measured on one observational unit.

We will look at two types of variables: **quantitative** and **categorical**.

Quantitative variables are numerical measurements that can be discrete (whole, non-negative numbers) or continuous (any value within an interval). The number of pets one owns would be a discrete variable as you can not have a partial pet. GPA would be a continuous variable ranging from 0 to 4.0.

The outcome of a categorical variable is a group or category such as eye color, state of residency, or whether or not a student lives on campus. Categorical variables with a natural ordering are considered *ordinal* variables (e.g., A-F letter grades or days of the week). Categorical variables without a natural ordering are considered *nominal* variables (hair color, gender identity, favorite color). Although these are important differences, in this course we will treat all categorical variables as nominal variables.

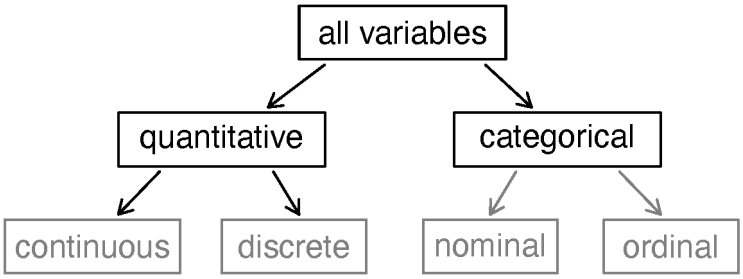


Figure 1: Types of variables.

5. Is the variable identified in question 4 categorical or quantitative?

Step 3: Once we have collected data, the next step is to *summarize and visualize the data*.

6. How many people in your class were correct in identifying Bouba? Using the class size from question 3, calculate the proportion of students who correctly identified Bouba.

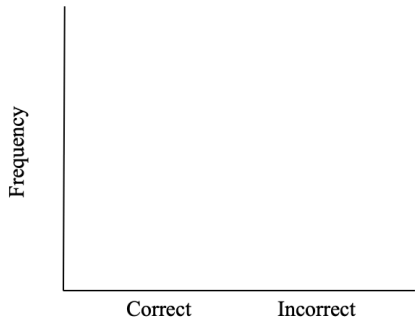
$$\text{proportion} = \frac{\text{number of students who correctly identified Bouba}}{\text{total number of students}}$$

The proportion in question 6 is called a **summary statistic**—a single value that summarizes the data set. It is important to note that a variable is different than a summary statistic. A *variable* is measured on a *single observational unit* while a summary *statistic* is calculated from a *group* of observational units.

For example, the variable "whether or not a student lives on campus" can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who live on campus, the summary statistic. Look back and make sure you wrote the variable in question 4 as a variable, NOT a summary statistic.

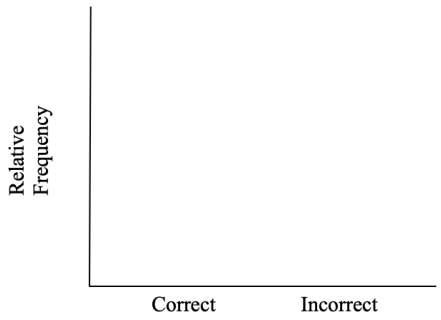
Looking at the data set and the summary statistic is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels or outcomes, correct or incorrect, we will create two bars—one for each level.

7. Plot the observed class data using a frequency bar plot. Be sure to add a scale to the y-axis.



We can also visualize the data as a proportion in a **relative frequency bar plot**. Relative frequency is the proportion calculated for each level of the categorical variable.

8. Plot the observed class data using a relative frequency bar plot. Be sure to add a scale to the y-axis.



Step 4: Now that we've visualized and summarized what we observed in our sample, the next step is to use statistical methods to draw inferences from our results. To answer the question of whether our class can read Martian, we need to know what *could* have happened in our class just by random chance. We can then compare what happened in our class to these "random chance" statistics to decide if we honestly believe our statistic is unlikely to have occurred if we all were just guessing.

To do this we need to simulate what statistics could have happened if we all were just guessing. With lots of these simulated statistics, we can understand the *variability* we might expect to see between different "randomly guessing" classes. We can then compare our class's observed data to the distribution of simulated statistics to get an idea of how often the class's result would occur if all of use were merely guessing. This comparison allows us to determine if the data provides evidence that we as a class can in fact read Martian.

9. If humans really don't know Martian and are just guessing which is Bouba, what are the chances of getting it right?

10. How could we use a coin to simulate each student "just guessing" which Martian letter is Bouba?

11. How could we use coins to simulate the entire class "just guessing" which Martian letter is Bouba?

12. How many people in your class would you expect to choose Bouba correctly just by chance? Explain your reasoning.

13. Each student will flip a coin one time to simulate your "guess" under the assumption that we can't read Martian. Let Heads = correct, Tails = incorrect.

- What was the result of your one simulation?

- What was the result from your class's simulation? What proportion of students "guessed" correctly in the simulation?

14. If students really don't know Martian and are just guessing which is Bouba, which seems more unusual: (1) the result from your class's **simulation** or (2) the **observed proportion** of students in your class that were correct (this is your summary statistic from question 6)? Explain your reasoning.

While your observed class data is likely different from the simulated "just-guessing" class, comparing our class data to a single simulation does not provide enough information to decide whether we believe our class can read Martian! The differences between these two statistics seen could just be due to the randomness of that set of coin flips! Let's simulate another class.

15. Each student should flip their coin again. What was the result from your class's second simulation? What proportion of students "guessed" correctly in the second simulation? Create a plot to compare the two simulated results with the observed class result.

We still only have a couple of simulations to compare our class data to. It would be much better to be able to see how our class compared to *hundreds* or *thousands* of "just-guessing" classes. Since we don't want to flip coins all class period, your instructor will use a computer simulation to get 1000 "random guessing" statistics.

16. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 statistics.

    Probability of correct guess: _____

    Sample size: _____

    Number of repetitions: _____

Sketch the distribution displayed by your instructor here. Label each axis appropriately.

17. What does one dot on the plot above represent in context of the problem?

18. Is your class particularly good or bad at Martian? Use the plot in question 17 to explain your answer.

19. Is it *possible* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

20. Is it *likely* that we could see our class results just by chance if everyone was just guessing? Explain your reasoning.

Step 5: The next step in the statistical investigation process is to *communicate the results and answer the research question.*

21. Does this activity provide strong evidence that students were not just guessing at random? If so, what do you think is going on here? Can we as a class read Martian?

If you are interested, this activity was inspired by a TED talk given by Vilayanur Ramachandran in 2007. The synesthesia part begins at roughly 17:30 minutes: https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.

Take-home messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses" when asked to identify Bouba) to a distribution of many simulated results under an assumption like "blind guessing."
2. Blind guessing between two outcomes will be correct only about half the time. We can simulate data using a computer program to fit the assumption of blind guessing.
3. Unusual observed results will make us doubt the assumptions used to create the simulated distribution. A large number of correct "guesses" is evidence that a person was not just blindly guessing.

Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered, and to write down the names and contact information of your teammates.