Martian Alphabet

Learning Outcomes

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

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 bumba, write down whether you think bumba is on the left or the right.

Steps of Statistical Investigation

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!). To answer any research question, we must design a study and collect data. (This will normally be provided for you in class.) For our question, the study consists of each student being presented with two Martian letters and asking which was bumba. Your responses will become our observed data that we will explore. To answer the research question we will simulate what *could* have happened in our class given random chance, repeat that many times to understand the expected variability between different "randomly guessing" classes, then comparing our class's observed data to the simulation. This gives us an estimate of how often (or the probability of) our class's result would occur if we were all merely guessing, allowing us to determine if we as a class can in fact read Martian.

Let's explore the data. **Observational units** or **cases** are the subjects data is collected on. In a data set the rows will represent a single observational unit.

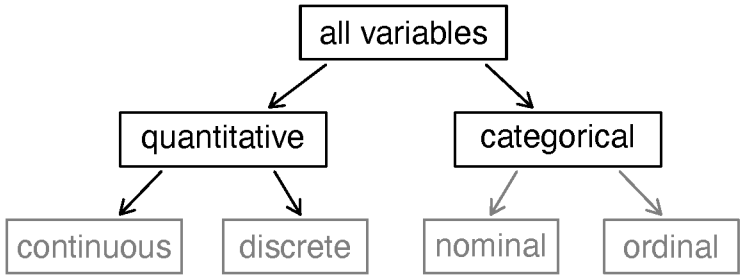
1. What are the observational units in this study?

2. 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. We will look at two types of variables: **quantitative** and **categorical**. Each column in a data set will represent a different variable.

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

Categorical variables are data that are in groups or categories such as eye color, state of residency, or whether or not a student is considered in-state. Categorical variables with a natural ordering are considered ordinal variables while those without a natural ordering are considered a nominal variable. All variables will be treated as nominal for analysis.



3. Identify the variable we are collecting on each observational unit in this study. What are we measuring on each student?

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 is considered in-state** can be measured on each individual student. In a class of 50 students we can calculate the proportion of students who are considered in-state, the summary statistic. Make sure you wrote the variable in question 3 as a variable **NOT** a summary statistic.

4. Is the variable identified in question 3 categorical or quantitative?

5. Were you correct or incorrect in identifying bumba?

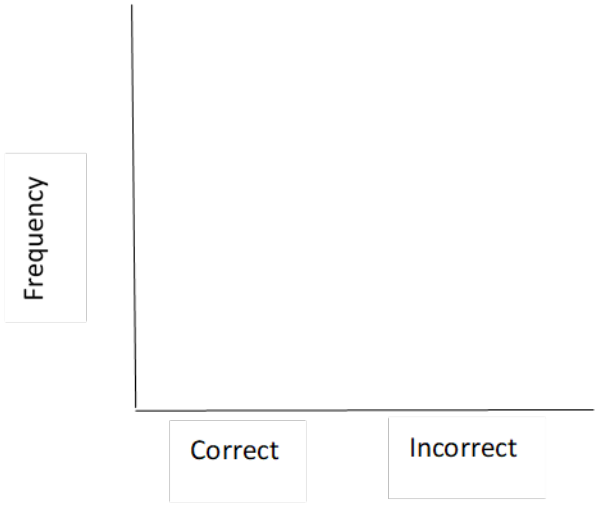
We will now collect the data from the entire class.

6. How many people in your class were correct in identifying bumba? Using the class size from question 2, calculate the proportion of students who correctly identified bumba.

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

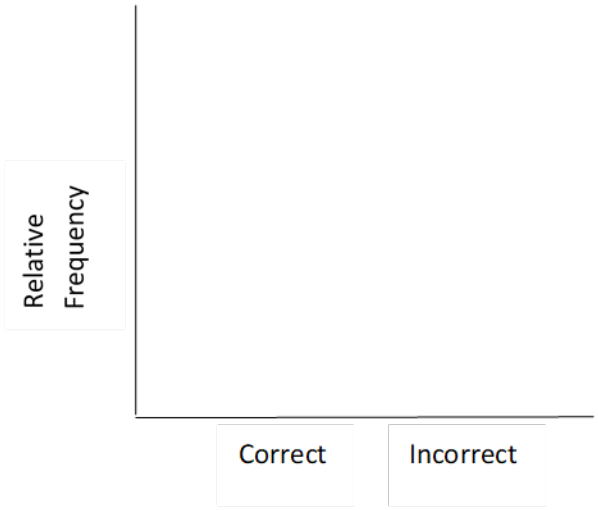
Looking at the data set and the summary statistics is only one way to display the data. We will also want to create a visualization or picture of the data. A **frequency bar plot** is used to display categorical data as a count or frequency. Since our variable has two levels, correct or incorrect, we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

7. Plot the observed class data using a frequency bar plot.



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.



9. The next step is to analyze the data. If humans really don't know Martian and are just guessing which is bumba, what are the chances of getting it right?

   How could we use a coin to simulate each student "just guessing" which martian letter is bumba?

How could we use coins to simulate the entire class "just guessing" which martian letter is bumba?

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

10. Each of you will flip a coin one time to simulate your "guess". Let Heads = correct, Tails = incorrect. What was the result of your simulation?

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

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

12. While your observed class data is likely far different from the simulated "just-guessing" class, comparing our class data to a single simulation does not seem to give enough information. The differences seen could just be due to that set of coin flips! Let's simulate another class. Each student should flip your 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.

13. We still unfortunately 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 web app to get several 1000 trials. Fill in the following blanks to describe how we would create a simulation of random guessing with 1000 trials.

    Probability of correct guesses: _____

    Sample size: _____

    Number of repetitions: _____

14. Sketch the distribution displayed by your instructor here, being sure to label each axis appropriately.

15. Is your class particularly good or bad at Martian? How can you use the plot in question 14 to tell?

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

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

18. 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?

Take Home Messages

1. In this course we will learn how to evaluate a claim by comparing observed results (classes' "guesses") 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 create data (via computer simulation) 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.