Chapter 1: Introduction to Statistical Thinking and Data

In the previous notes, we encountered basic examples that required us to think statistically in order to investigate a question of interest. Before we move on to slightly more complex examples, we will discuss some basic definitions that will be used throughout the semester.

DEFINITIONS

- **Statistics:**

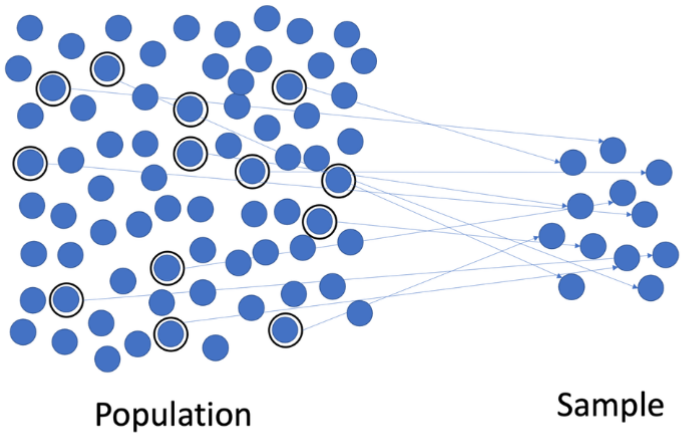
- **Categorical (or qualitative) data:** Measurements that are classified into one of a group of categories.
- **Numerical (or quantitative) data:** Measurements that are recorded on a naturally occurring numerical scale.

Most of what we'll be doing in this course centers on trying to understand a set of information. This set of information is from a...

- **Population:** The complete collection of ALL elements that are of interest for a given problem.

The population is often so big that obtaining all information about its elements is either difficult or impossible. So, we work with a more manageable set of data that we obtain from a...

- **Sample:** A subcollection of elements drawn from a population. The number of elements drawn is called the **Sample Size**.
- **Observation:** The collection of measurements from a particular unit in a sample.
- **Variable:** Any measurable characteristic of an observation. Variables can be classified as either *categorical* or *numerical*.



Revisit Example 0.2: Can our class speak Martian?

1. Identify the following in the context of this example:

- Population of interest:

- Sample:

- Variable of interest:

- Data type:

2. Suppose we were interested in whether all Cal Poly students could speak Martian. What would change from your answers above?

3. What might be concerning about our sample if we wanted to know if all Cal Poly students could speak Martian?

DEFINITIONS

Representative Sample: individuals in the sample accurately reflect the characteristics of the population from which it is drawn.

Convenience Sample: individuals who are easily accessible are more likely to be included in the sample.

In general, we always seek to **randomly** select a sample from a population.

See https://openintro-ims.netlify.app/data-design#sec-samp-methods for more information on sampling methods.

When working with statistical research questions, the information is usually stored in a data set so it can be shared, visualized, or analyzed.

DEFINITIONS

Tidy Data is a standard way of mapping the meaning of a data set to its structure. In tidy data,

- each *variable* forms a column
- each *observation* forms a row
- each *cell* is a *single measurement*.

For the observed data our class collected in Example 0.1, the tidy data set of 35 students might look like the following:

Student	Outcome
1	Correct
2	Correct
3	Correct
4	Correct
5	Correct
6	Correct
7	Correct
8	Incorrect
9	Correct

10	Incorrect
11	Incorrect
12	Incorrect
13	Correct
14	Correct
15	Incorrect
16	Incorrect
17	Incorrect
18	Correct
19	Incorrect
20	Correct
21	Correct
22	Correct
23	Incorrect
24	Correct
25	Incorrect
26	Correct
27	Correct
28	Correct
29	Correct
30	Correct
31	Correct
32	Correct
33	Correct
34	Correct
35	Correct

There might be extra variables collected on each observation and included in the data set than just our variable of interest. For example, we might have collected the `age`, `major`, `dorm`, and `height` of the student.

Student	Outcome	Age (years)	Major	Dorm	Height
1	Correct	21.3	Microbiology	Shasta	
2	Correct	22.9	Biochemistry	Sequoia	
3	Correct	21.5	Animal Science	Sierra Madre	
4	Correct	20.8	Animal Science	Poly Canyon Village	
5	Correct	20.5	Microbiology	Fremont	
6	Correct	21.3	Food Science	Fremont	
7	Correct	22.4	Biological Sciences	yak it ut u	
8	Incorrect	21.0	Environmental Management & Protection	Sierra Madre	
9	Correct	18.7	Animal Science	Tenaya	
10	Incorrect	20.9	Psychology	Fremont	
11	Incorrect	20.1	Biological Sciences	Tenaya	
12	Incorrect	19.0	Psychology	Poly Canyon Village	
13	Correct	19.2	Biological Sciences	Fremont	
14	Correct	20.5	Nutrition	Fremont	
15	Incorrect	22.3	Nutrition	Cerro Vista	
16	Incorrect	20.2	Statistics	Poly Canyon Village	
17	Incorrect	18.8	Nutrition	yak it ut u	
18	Correct	18.9	Nutrition	yak it ut u	
19	Incorrect	19.8	Marine Sciences	Poly Canyon Village	
20	Correct	21.3	Microbiology	Poly Canyon Village	
21	Correct	21.3	Dairy Science	Cerro Vista	
22	Correct	19.8	Dairy Science	Shasta	
23	Incorrect	21.3	Nutrition	Sequoia	
24	Correct	19.1	Statistics	Cerro Vista	
25	Incorrect	18.4	Psychology	Shasta	
26	Correct	20.9	Nutrition	Lassen	
27	Correct	21.3	Kinesiology	Trinity	
28	Correct	22.9	Psychology	Lassen	
29	Correct	22.9	Biochemistry	Fremont	
30	Correct	20.2	Environmental Management & Protection	Trinity	
31	Correct	21.4	Psychology	Poly Canyon Village	
32	Correct	21.6	Biochemistry	Yosemite	
33	Correct	19.7	Microbiology	Sierra Madre	
34	Correct	19.6	Statistics	Sequoia	

35	Correct	18.1	Psychology	Sequoia

4. Identify each of the variables contained in the data set above and determine whether the variable contains measurements of categorical or numeric data types.

Example 1.1: Helper vs. Hinderer?

- Helper Triangle: https://www.youtube.com/watch?v=j4n_Qh4Gg9Q
- Hinderer Square: https://www.youtube.com/watch?v=ExcxDMEHlHY
- 10-month old choice: https://www.youtube.com/watch?v=NsWICFLt_-g

In a study reported in a November 2007 issue of Nature, researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction (Hamlin, Wynn, and Bloom 2007). In one component of the study, sixteen 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character ("helper") and one where the climber was pushed back down the hill by another character ("hinderer"). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The color and shape and order (left/right) of the toys were varied and balanced out among the 16 infants (Holcomb et al. 2010).

? Research Question: Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?

1. Why was it important for the researchers to balance out the color, shape, and order of the toys across the study? For example, how would the study results have been affected if the researchers always made the helper toy a blue circle and the hinderer a yellow triangle?

DEFINITION

Confounding Variable: characteristics other than the variable of interest (e.g., helper/hinderer) that may be related to the outcome (e.g., choice).

2. Identify the following in the context of this example:

- Population of interest:

- Sample:

- Variable of interest:

- Data type:

- How would you store this information in tidy data format? Think about what your rows and columns represent.

4. Recall that this study involves 16 infants. If the population of all 10-month-old infants has no real preference for one toy over the other, how many infants do you expect to choose the helper toy? Explain.

5. Suppose that 10 out of 16 infants choose the helper toy (62.5%). Since this value is higher than 50%, a researcher argues that these data show that the majority of all 10-month-old infants would choose the helper toy. What is wrong with their reasoning?

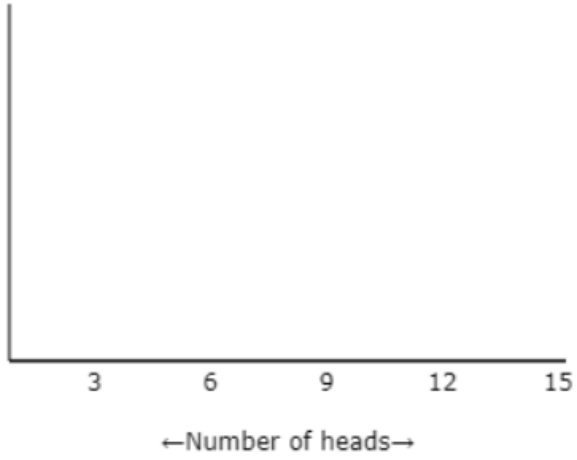
Once again, the key question is how to determine whether the study's result is surprising under the assumption that there is no real preference for one toy over the other in the population of all 10-month-old infants. To answer this, we will simulate the process of 16 infants simply choosing a toy at random, over and over again. Each time we simulate the process, we'll keep track of how many infants out of the 16 chose the helper toy (note that you could also keep track of the number that chose the hinderer toy). Once we've repeated this process a large

number of times, we'll have a pretty good sense for what outcomes would be very surprising, somewhat surprising, or not so surprising if the population of all 10-month-old infants has no real preference.

Carry out the simulation via the `Online Simulation Applets > One Proportion Inference`. Note that you should consider the following questions when designing your simulation study:

- What are the two possible outcomes on each of the trials?
- What is the probability that the helper is selected under the assumption that the population of all 10-month-old infants has no real preference for either toy? Change your `Probability of heads` accordingly.
- How many infants were used in this study? Keep this value in mind when setting the `Number of tosses` value.

Carry out the simulation study 100 times overall, keeping track of the number of infants that choose the helper toy in each of the simulated experiments. Sketch in your results below:



6. What does each dot on this plot represent?

7. Suppose that in the actual study 10 out of 16 infants chose the helper toy. Would this convince you that the majority of the population of all 10-month-old infants had a preference for the helper toy? Why or why not?

8. The actual study results are as follows: 14 out of 16 infants chose the helper toy. Mark this on the axis above the results of your simulations study. Based on this statistical investigation, what should the researchers conclude? Recall that their research question was stated as follows: Do 10-month-old infants tend to *prefer* the helper toy over the hinderer toy?

Example 1.2: Are Women Passed Over for Managerial Training?

> ⚠ Warning
>
> *It is important to acknowledge that the data collected in this study inherently assumes all employees identify as either woman or man. However, we recognize that this depiction does not mirror the diverse realities of all individuals.*

This example involves possible discrimination against women employees. Suppose a large supermarket chain occasionally selects employees to receive management training. A group of women employees has claimed that they are less likely than men employees of similar qualifications to be chosen for this training.

The large employee pool that can be tapped for management training is 60% women and 40% men; however, since the management program began, 9 of the 20 employees chosen for management training were women (only 45%). Do the women employees have a valid statistical argument that they are being discriminated against?

? Research Question: *Is there statistical evidence for sex discrimination against women?*

1. Identify the following in the context of this example:

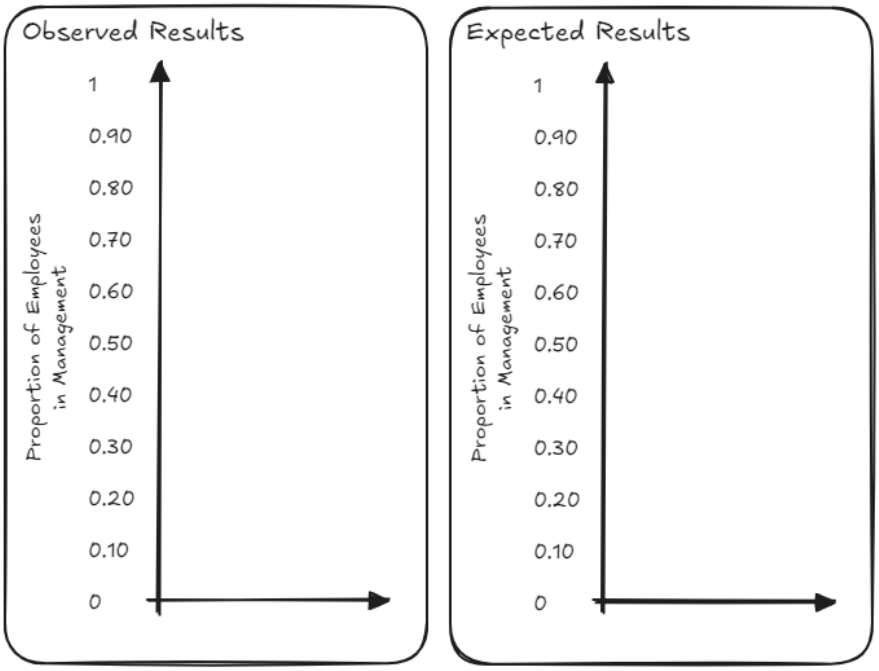
- Population of interest:

- Sample:

- Variable of interest:

- Data type:

2. Sketch out **stacked barplots** to compare the observed results since the management program began and the expected results based on the large employee pool. Think about what your two possible outcomes are. What do you notice?



3. If the selection process was unbiased, how many of the 20 employees selected for management do you expect to be women? Explain.

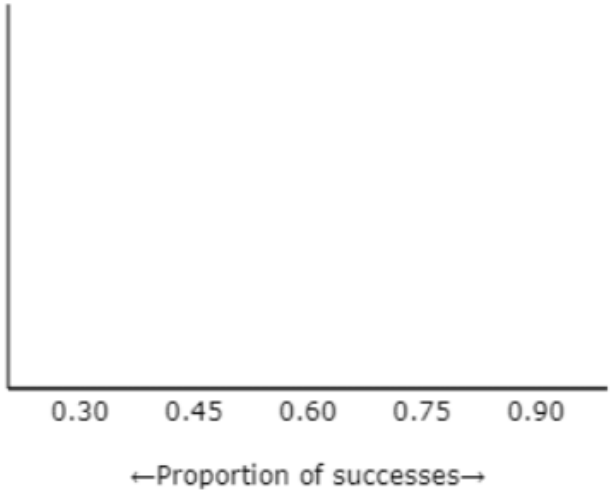
Once again, the key question is how to determine whether the result is surprising under the assumption that the selection process is unbiased. To answer this, we will simulate the process of an unbiased selection process, over and over again. Each time we simulate the process, we'll keep track of how many of the 20 employees selected for management were women. Once

we've repeated this process a large number of times, we'll have a pretty good sense for what outcomes would be very surprising, somewhat surprising, or not so surprising if there was no discrimination in the selection process.

Carry out the applet simulation. Note that you should consider the following questions when designing your simulation study:

- What are the two possible outcomes on each of the trials?
- What is the probability that a woman is selected for a managerial position under the assumption that there is no sex discrimination in the selection process? Change your `Probability of heads` accordingly.
- How many subjects were there in this study? Keep this value in mind when setting the `Number of tosses` value.

Carry out the simulation study 1000 times overall, keeping track of the probability of employees chosen for management that were women on each of the simulated experiments. Sketch in your results below:



4. What does each dot on this plot represent?

5. Recall that since the management program began, only 9 of the 20 employees (45%) chosen for management training were women Does this outcome convince you that the selection process is biased against women? Why or why not?

6. Can we use the results from the employee program from this large supermarket chain to say anything about employee programs from competing supermarkets?

Example 1.3: Font Preferences

Researchers carried out a marketing field study in order to study preferences of potential consumers in the U.S. They used silver cardboard boxes to contain chocolate truffles in a forced choice task. All of the box tops were decorated in the same way, and a white label was attached to each bearing the name *"Indulgence"* in either Signet font or Salem font. The text on each label was approximately equal-sized. For each of the 40 subjects in the study, one box labeled with the Signet font and another box labeled with the Salem font were placed on a tray, and the subject was simply asked to choose a truffle from one of the two boxes that were on the tray in front of them. The researchers randomized the order in which the fonts were presented to each participant.

Font Style	
Signet	Salem
Indulgence	Indulgence

½ of the people were presented a tray like this	
The other ½ were presented a tray like this	

The researchers aren't sure which font is more appropriate for the label and simply want to know whether the majority of all consumers will choose the truffles with one font more than the other. In the sample of 40 subjects, 30 chose to take a truffle from the box that had Signet font.

? Research Question: Do the majority of consumers have a preference for one font over the other?

1. Identify the following in the context of this example:

- Population of interest:

- Sample:

- Variable of interest:

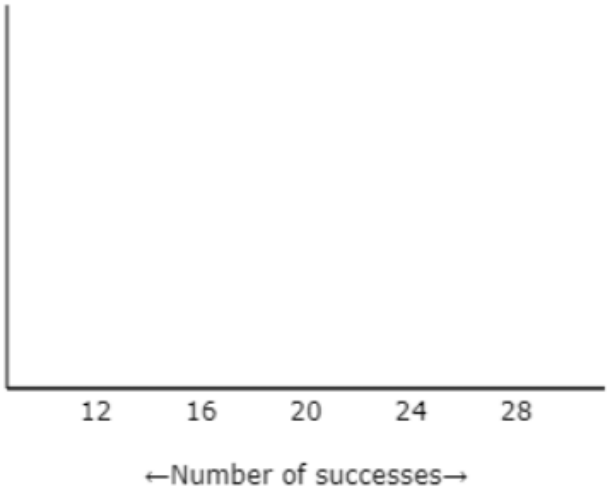
- Data type:

2. If there was no preference in the population, how many of the 40 consumers do you expect to choose Signet font? Explain.

To gain an understanding of what outcomes we expect to see if there is no real preference in the population of all consumers, we will simulate this experiment under the condition that there is no preference for one font over the other. Carry out the Applet simulation. Note that you should consider the following questions when designing your simulation study:

- What are the two possible outcomes on each of the trials?
- What is the probability that the Signet font is selected under the assumption that there is no preference in the population? Change your `Probability of heads` accordingly.
- How many subjects were there in this study? Keep this value in mind when setting the `Number of tosses` value.

Carry out the simulation study 1000 times overall, keeping track of the number that choose Signet on each of the simulated experiments. Sketch in your results below:



3. What does each dot on this plot represent?

4. In the actual study, 30 of the 40 selected the Signet font. Does this outcome convince you that there is a preference for one font over the other? Why or why not?

5. Why was it important for the researchers to present out the order in which the fonts were presented across the study? For example, how would the study results have been affected if the researchers always presented the Signet font on the left?

References

Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. 2007. "Social Evaluation by Preverbal Infants." *Nature* 450 (7169): 557–59. https://doi.org/10.1038/nature06288.

Holcomb, John, Beth Chance, Allan Rossman, Emily Tietjen, and George Cobb. 2010. "Introducing Concepts of Statistical Inference via Randomization Tests." *Data and Context in Statistics Education: Towards an Evidence-Based Society (ICOTS8), Voorburg, The Netherlands*.