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.

ℹ DEFINITIONS

- **Statistics:** The study of how to collect, analyze, and draw conclusions from data.

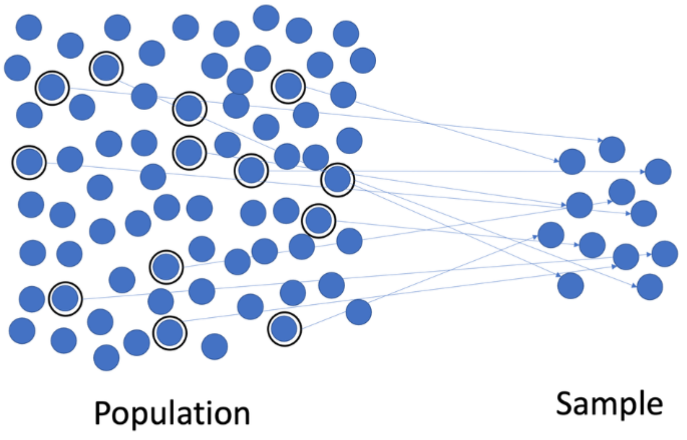
- **Categorical (or qualitative) data:** Measurements that are classified into one of a group of categories. correct/incorrect, hair color, etc.
- **Numerical (or quantitative) data:** Measurements that are recorded on a naturally occurring numerical scale. height, speed, etc.

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. Not necessarily people (subjects)

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*.



Notice it is hard to collect ALL elements in the population, thus we take a subset of elements (n = 10) and measure variables (e.g., hair color, height) on each observational unit in the sample.

Revisit Example 0.2: Can our class speak Martian?

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

- Population of interest: All students enrolled in Stat 218 section xx at Cal Poly

- Sample: n = 36 students in class that day

- Variable of interest: Outcome

- Data type: Categorical (correct/incorrect)

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

The population is now all students enrolled at Cal Poly, not just our class.

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

Students in our Stat 218 class might not be *representative* of all students at Cal Poly (e.g., maybe Life Science majors have something that makes them speak martian).

ℹ 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.

ℹ 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

Student	Outcome
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` of the student.

Student	Outcome	Age (years)	Major	Dorm	Height (in)
1	Correct	21.3	Microbiology	Fremont	66.7
2	Correct	22.9	Biochemistry	Cerro Vista	85.9
3	Correct	21.5	Animal Science	PCV	67.2
4	Correct	20.8	Animal Science	yak it ut u	61.1
5	Correct	20.5	Microbiology	yak it ut u	66.8
6	Correct	21.3	Biological Sciences	PCV	77.4
7	Correct	22.4	Food Science	PCV	60.0
8	Incorrect	21.0	Animal Science	Cerro Vista	67.5
9	Correct	18.7	Biological Sciences	Shasta	65.6
10	Incorrect	20.9	Biological Sciences	Sequoia	60.9
11	Incorrect	20.1	Nutrition	Cerro Vista	68.4
12	Incorrect	19.0	Nutrition	Shasta	64.4
13	Correct	19.2	Nutrition	Lassen	60.5
14	Correct	20.5	Nutrition	Trinity	72.2
15	Incorrect	22.3	Marine Sciences	Lassen	58.3
16	Incorrect	20.2	Microbiology	Fremont	68.5
17	Incorrect	18.8	Dairy Science	Trinity	73.9
18	Correct	18.9	Dairy Science	PCV	80.5
19	Incorrect	19.8	Nutrition	Yosemite	81.2
20	Correct	21.3	Nutrition	Sierra Madre	64.6
21	Correct	21.3	Kinesiology	Sequoia	76.5
22	Correct	19.8	Biochemistry	Sequoia	61.0
23	Incorrect	21.3	Food Science	Cerro Vista	53.9
24	Correct	19.1	Biochemistry	Yosemite	71.3
25	Incorrect	18.4	Microbiology	Sierra Madre	75.4
26	Correct	20.9	Food Science	Sequoia	78.1
27	Correct	21.3	Microbiology	Yosemite	56.2
28	Correct	22.9	Animal Science	Muir	83.6
29	Correct	22.9	Nutrition	Sequoia	69.5
30	Correct	20.2	Nutrition	Lassen	64.8
31	Correct	21.4	Animal Science	Sierra Madre	55.4
32	Correct	21.6	Biochemistry	Fremont	65.7
33	Correct	19.7	Nutrition	Fremont	60.1
34	Correct	19.6	Biochemistry	Tenaya	66.9
35	Correct	18.1	Nutrition	yak it ut u	77.2

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.

Student (technically this ID's our observational unit, categorical), Outcome (this is our response of interest, categorical), Age (numeric), Major (categorical), Dorm (categorical), Height (numeric)

Note that the number of correct is not a variable, this counts up and summarizes the sample. Similarly, the average age is not a legit variable as this summarizes the variable age in our sample.

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?

Eliminate biaws in the results. We want to narrow the conlcusion to say the helper/hinderer rather than an an infants' preference for say red or the right side.

> 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: ALL 10-month old infants

- Sample: n = 16 10-month olds in the study (note: the observational unit is one 10-month old)

- Variable of interest: Choice

- Data type: Categorical – (helper/hinderer)

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

Infant	Choice
1	Hinderer
2	Helper
3	Hinderer
4	Helper
5	Helper
6	Helper
7	Hinderer
8	Hinderer
9	Hinderer
10	Hinderer
11	Hinderer
12	Hinderer
13	Helper
14	Hinderer
15	Helper
16	Hinderer

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.

Expect 8 out of 16 (50% chance of selecting helper if no preference)

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?

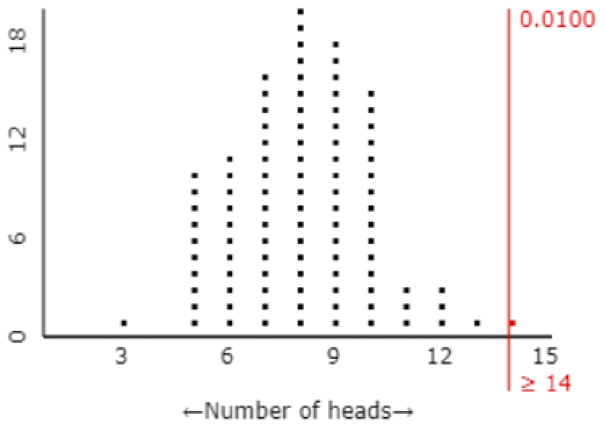
Although 62.5% is higher than 50%, the researchers should conduct a simulation study to determine what results could have happened if there was no preference.

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? Helper/Hinderer
- 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. 0.5
- How many infants were used in this study? Keep this value in mind when setting the `Number of tosses` value. n = 16

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:



Notice, most simulated sets had 8 out of 16 choose the helper.

6. What does each dot on this plot represent?

One set of simulated 16 10-month olds under the assumption there is no preference. (Note: there are 100 of these sets!)

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?

Since 10 out of 16 is consistent with the results from our simulation study, we are not convinced that a majority of the population of all 10-month old infants have a preference for the helper toy.

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?

Since 14 out of 16 infants is in the tail end of our simulated distribution, we are convinced that a majority of the population of all 10-month old infants have a preference for helper 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:

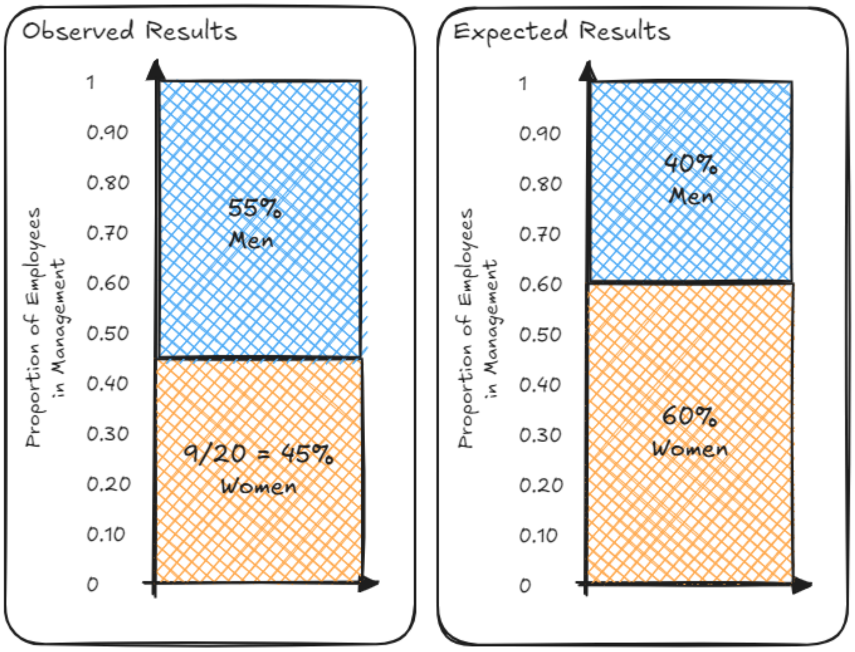
- Population of interest: ALL employees who will be in the management program at this supermarket chain (i.e., think of this program going on for years and years).

- Sample: n = 20 employees selected for training

- Variable of interest: sex of employee

- Data type: categorical (man/woman)

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.

0.60*20 = expect 12 employees selected for management training to be women. Note that employees for the program are selected from the larger employee pool where 60% are women and 40% are men. If the process is fair, there should be 60% women in the program and 40% men to match the random chance of being selected from the large pool.

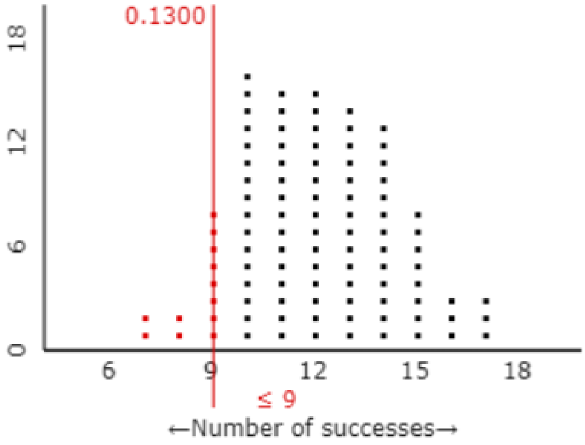
Once again, the key question is how to determine whether the result is surprising under the assumption that the selection process is unbiased. To answer this, we will simulate the process of an unbiased selection process, over and over again. Each time we simulate the process, we'll keep track of how 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? man/woman

- 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. 0.6
- How many subjects were there in this study? Keep this value in mind when setting the `Number of tosses` value. n = 20

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:



Note the simulated distribution is centered at about 60%.

4. What does each dot on this plot represent?

One set of 20 employees simulated under the assumption the process is unbiased

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?

I am not convinced the selection process is biased toward women since 45% (or less) regularly occurred in our simulation study. The observed result of 0.45 is consistent with what we would see if the process was fair.

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

No, our sample of n = 20 employees only came from the one supermarket and is not representative of other 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**. Observed results

? **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: ALL potential consumers in the U.S.

- Sample: n = 40 subjects in the study

- Variable of interest: Font

- Data type: Categorical (Signet / Salem)

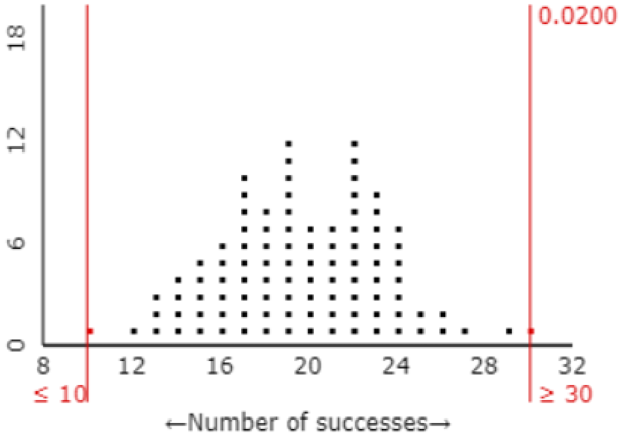
2. If there was no preference in the population, how many of the 40 consumers do you expect to choose Signet font? Explain.

Expect 20 out of 40 (50% chance of randomly choosing)

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? Signet / Salem
- 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. 0.5
- How many subjects were there in this study? Keep this value in mind when setting the `Number of tosses` value. n = 40

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?

One set of n = 40 consumers simulated under the assumption there is no preference for a font.

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?

I am convinced there is a preference for one font over the other since the observed results of 30 (or more) or 10 (or less) are not consistent with the results from the simulation study.

Note that 75% observed Signet but 25% observed signet would be just as unusual.

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.*