

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

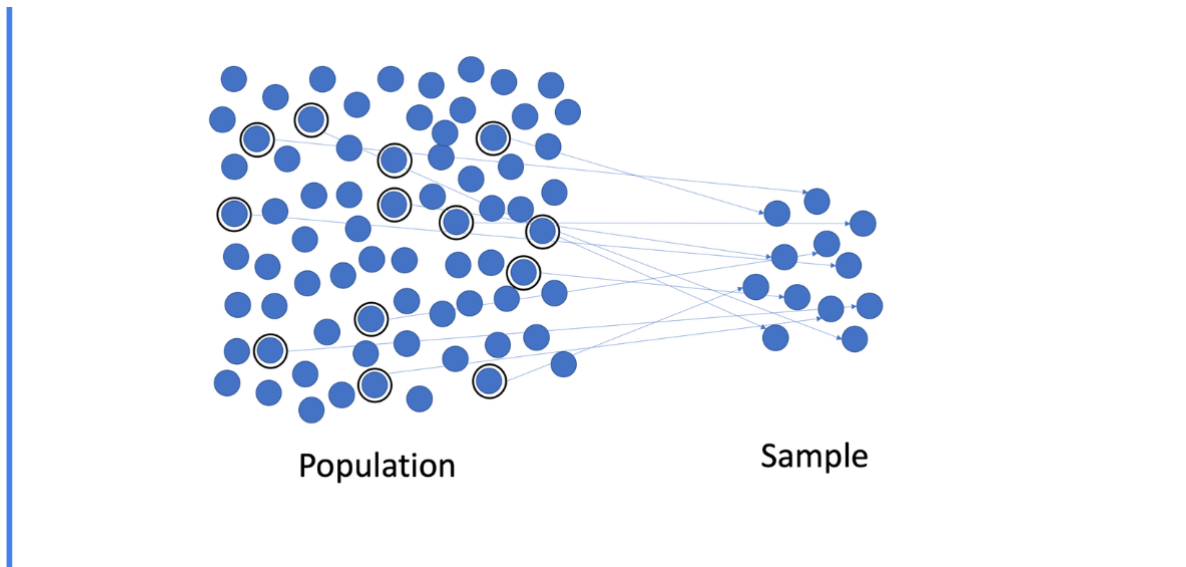
- **Statistics:**
- **Categorical (or qualitative) data:** Measurements that are classified into one of a group of categories.
- **Numerical (or quantitative) data:** Measurements that are recorded on a naturally occurring numerical scale.

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.



Example 1.1 Alleged Hearing Loss

Consider the following case study centered on potential insurance fraud regarding deafness. This case study was presented in Binder (1992), “A Forced-Choice Technique to Evaluate Deafness in the Hysterical or Malingering Patient.” The following is an excerpt from the article:

The patient was a 27-year-old male with a history of multiple hospitalizations for idiopathic convulsive disorder, functional disabilities, accidents, and personality problems. His hospital records indicated that he was manipulative, exaggerated his symptoms to his advantage, and that he was a generally disruptive patient. He made repeated attempts to obtain compensation for his disabilities. During his present hospitalization he complained of bilateral hearing loss, left-sided weakness, left-sided numbness, intermittent speech difficulty, and memory deficit. There were few consistent or objective findings for these complaints. All of his symptoms disappeared quickly with the exception of the alleged hearing loss.

To assess his alleged hearing loss, testing was conducted through earphones with the subject seated in a sound-treated audiology testing chamber. Visual stimuli utilized during the investigation were produced by a red and a blue light bulb, which were mounted behind a one-way mirror so that the subject could see the bulbs only when they were illuminated by the examiner. The subject was presented several trials on each of which the red and then the blue light were turned on consecutively for 2 seconds each. On each trial, a 1,000-Hz tone was randomly paired with the illumination of either the blue or red light bulb, and the subject was instructed to indicate with which light bulb the tone was paired. Because the researchers

were implementing a “forced-choice” technique, the subject was forced to answer each time with either “red” or “blue.”

The subject was correct on 7 out of 20 trials when they were asked to identify whether the tone played with either the red or the blue lightbulb.

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

- Population of interest:

- Sample:

- Variable of interest:









































- Data type:

Let’s carry out a simulation study to determine whether this patient who was suspected of malingering had obtained too few correct answers. Recall the results of the simulation study indicate what outcomes we expect from a _____ subject:

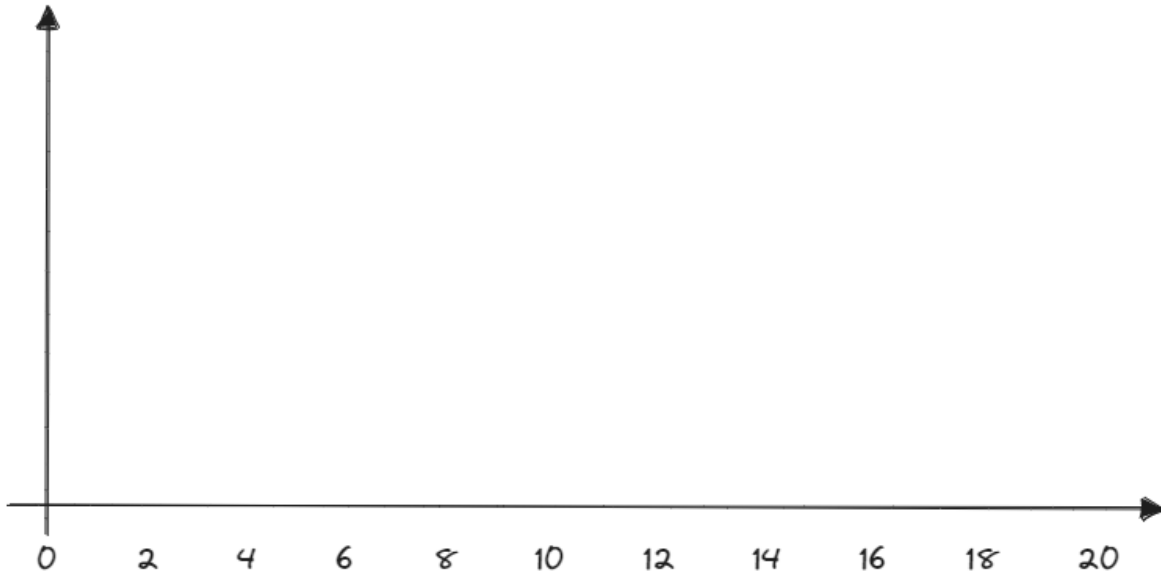
An applet has been constructed so that you can conduct your own repeated trials of this hearing experiment: http://course1.winona.edu/cmalone/afc_hearing/.

Recall that the goal is to mimic the outcomes of a deaf person. Therefore, when conducting this experiment, you should mute the speakers on your computer.

2. Conduct 20 repeated trials of the hearing experiment. Record the number of correct results below.

Trial	Choice	Correct?		Trial	Choice	Correct?
1	  Red Blue	<input type="checkbox"/>		11	  Red Blue	<input type="checkbox"/>
2	  Red Blue	<input type="checkbox"/>		12	  Red Blue	<input type="checkbox"/>
3	  Red Blue	<input type="checkbox"/>		13	  Red Blue	<input type="checkbox"/>
4	  Red Blue	<input type="checkbox"/>		14	  Red Blue	<input type="checkbox"/>
5	  Red Blue	<input type="checkbox"/>		15	  Red Blue	<input type="checkbox"/>
6	  Red Blue	<input type="checkbox"/>		16	  Red Blue	<input type="checkbox"/>
7	  Red Blue	<input type="checkbox"/>		17	  Red Blue	<input type="checkbox"/>
8	  Red Blue	<input type="checkbox"/>		18	  Red Blue	<input type="checkbox"/>
9	  Red Blue	<input type="checkbox"/>		19	  Red Blue	<input type="checkbox"/>
10	  Red Blue	<input type="checkbox"/>		20	  Red Blue	<input type="checkbox"/>
Total Number of Correct Results: _____						

3. Collect the simulation outcomes from everybody in the class. Place a dot for each outcome on the following number line. Make sure to add labels to the graph



4. Based on the results of this simulation study, do you believe the patient's outcome of 7 correct out of 20 was consistent with guessing, or do these results indicate that the subject may have been answering incorrectly on purpose in order to mislead the researchers into believing they were deaf?
5. In reality, the subject was correct on 36 out of 100 trials when they were asked to identify whether the tone played with either the red or the blue lightbulb. How does this change the process we just conducted above?

Tidy Data

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 subject's observed data in Example 1.1, the tidy data set of 20 trials might look like the following:

Trial	Outcome
1	Incorrect
2	Correct
3	Correct
4	Incorrect
5	Incorrect
6	Correct
7	Incorrect
8	Incorrect
9	Incorrect
10	Incorrect
11	Incorrect
12	Correct
13	Incorrect
14	Incorrect
15	Incorrect
16	Correct
17	Incorrect
18	Incorrect
19	Correct
20	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 number of seconds it took for the subject to answer.

Trial	Outcome	Time to Response (s)
1	Incorrect	7.9
2	Correct	8.7
3	Correct	4.0
4	Incorrect	7.1
5	Incorrect	2.7
6	Correct	6.2
7	Incorrect	0.9
8	Incorrect	5.6
9	Incorrect	8.2
10	Incorrect	1.0
11	Incorrect	5.1
12	Correct	5.1
13	Incorrect	2.7
14	Incorrect	0.3
15	Incorrect	2.6
16	Correct	0.1
17	Incorrect	0.6
18	Incorrect	2.5
19	Correct	1.5
20	Correct	9.0

Example 1.2: Helper vs. Hinderer?

- Helper Triangle: https://www.youtube.com/watch?v=j4n_Qh4Gg9Q
- Hinderer Square: <https://www.youtube.com/watch?v=ExcxDMEHIIHY>
- 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?

i 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.

-
- 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.
 - 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?

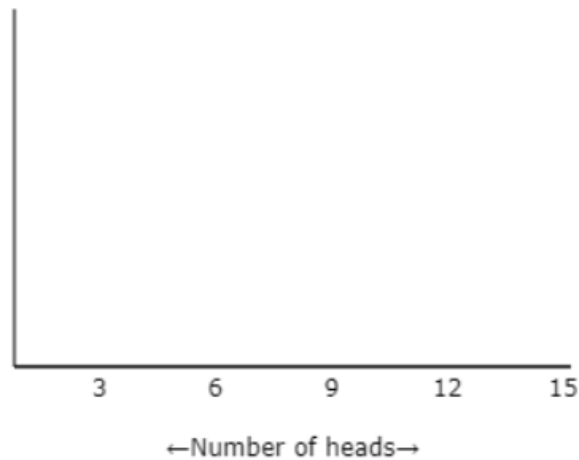
Once again, the key question is how to determine whether the study's result is surprising under the assumption that there is no real preference for one toy over the other in the population of all 10-month-old infants. To answer this, we will simulate the process of 16 infants simply choosing a toy at 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 1.3: 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?

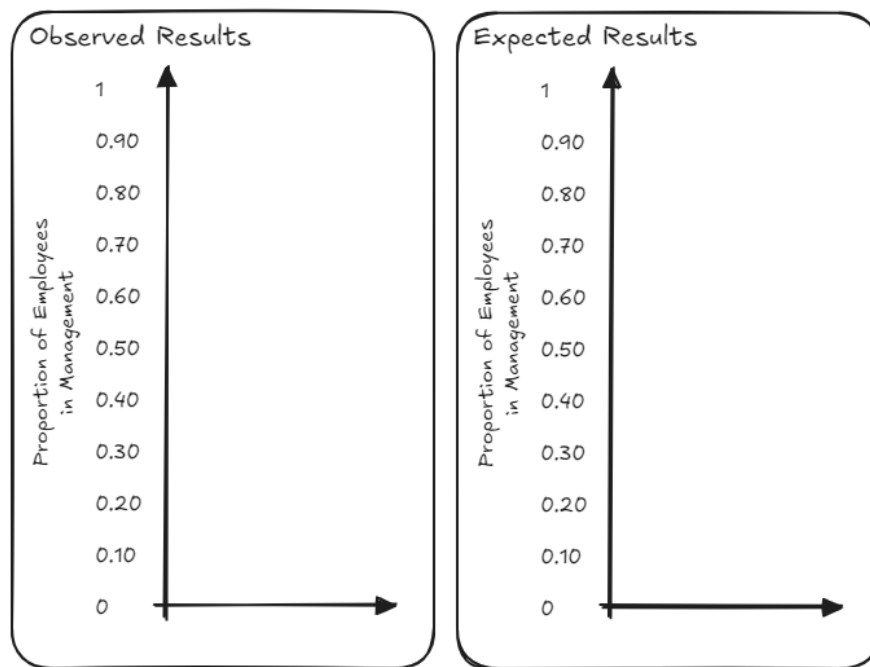
? Research Question: *Is there statistical evidence for sex discrimination against women?*

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.

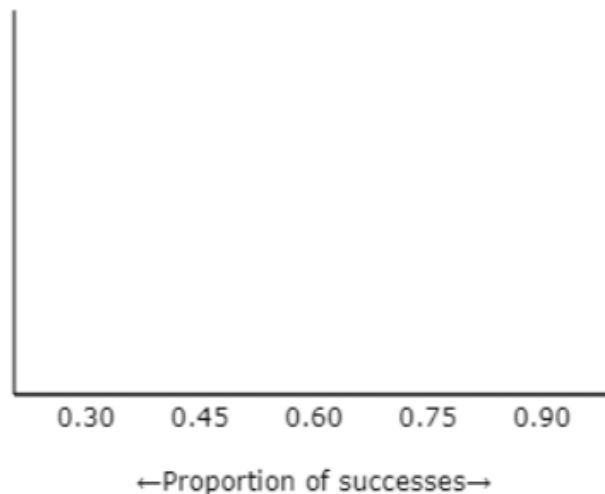
Once again, the key question is how to determine whether the result is surprising under the assumption that the selection process is unbiased. To answer this, we will simulate the process of an unbiased selection process, over and over again. Each time we simulate the process, we'll keep track of how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

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.

Example 1.4: 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.

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:

Font Style	
Signet	Salem
<i>Indulgence</i>	<i>Indulgence</i>

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

- Sample:

- Variable of interest:

- Data type:

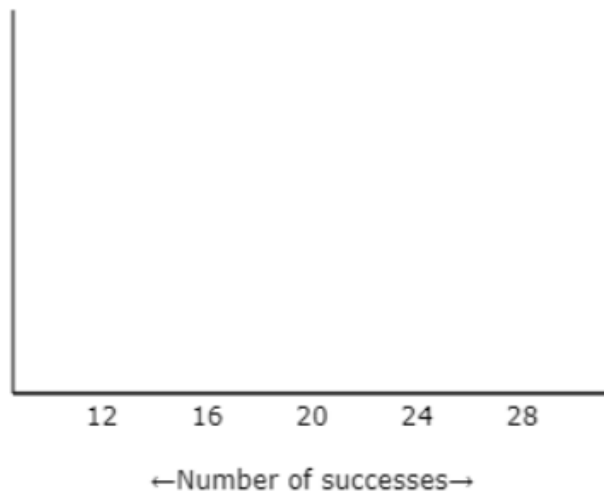
2. If there was no preference in the population, how many of the 40 consumers do you expect to choose Signet font? Explain.

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

- Binder, L. 1992. “Malingering Detected by Forced Choice Testing of Memory and Tactile Sensation: A Case Report.” *Archives of Clinical Neuropsychology* 7 (2): 155–63. [https://doi.org/10.1016/0887-6177\(92\)90009-c](https://doi.org/10.1016/0887-6177(92)90009-c).
- 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.