01-introduction-to-statistical-thinking

Wednesday, September 20, 2023 9:10 PM

01-introduct...

Chapter 1: Introduction to Statistical Thinking and Data

In last week's 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

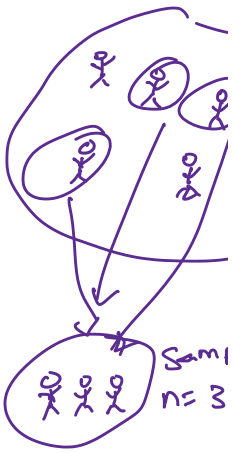
- **Statistics:** the study of collecting, analyzing, and drawing conclusions from data.
- **Categorical (or qualitative) data:** Measurements that are classified into one of a group of categories.
- **Numerical (or quantitative) data:** Measurements that are recorded on a naturally occurring numerical 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.



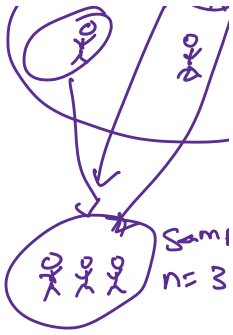
- **Numerical (or quantitative) data:** Measurements that are recorded on a naturally occurring numerical scale.

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.

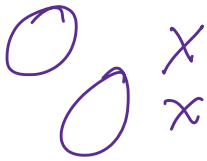


ject seated in a sound-treated audiology testing chamber. Visual stimuli utilized during the investigation were produced by a red and a blue light bulb, which were mounted behind a one-way mirror so that the subject could see the bulbs only when they were illuminated by the examiner. The subject was 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:

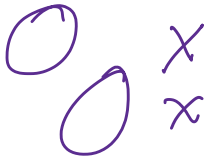
- Population of interest: ALL trials of red/blue lightbulb given to the patient
- Sample: The trials the patient participated in; n=20
- Variable of interest: outcome (correct/incorrect guessing lightbulb sound matching)
- Data type: categorical



Let's carry out a simulation study to determine whether this patient who was suspected of

- Variable of interest: outcome (correct/incorrect guessing lightbulb sound matching)

- Data type: categorical



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	☐	11	Red Blue	☐
2	Red Blue	☐	12	Red Blue	☐
3	Red Blue	☐	13	Red Blue	☐
4	Red Blue	☐	14	Red Blue	☐
5	Red Blue	☐	15	Red Blue	☐
6	Red Blue	☐	16	Red Blue	☐
7	Red Blue	☐	17	Red Blue	☐
8	Red Blue	☐	18	Red Blue	☐
9	Red Blue	☐	19	Red Blue	☐
10	Red Blue	☐	20	Red Blue	☐
Total Number of Correct Results:					

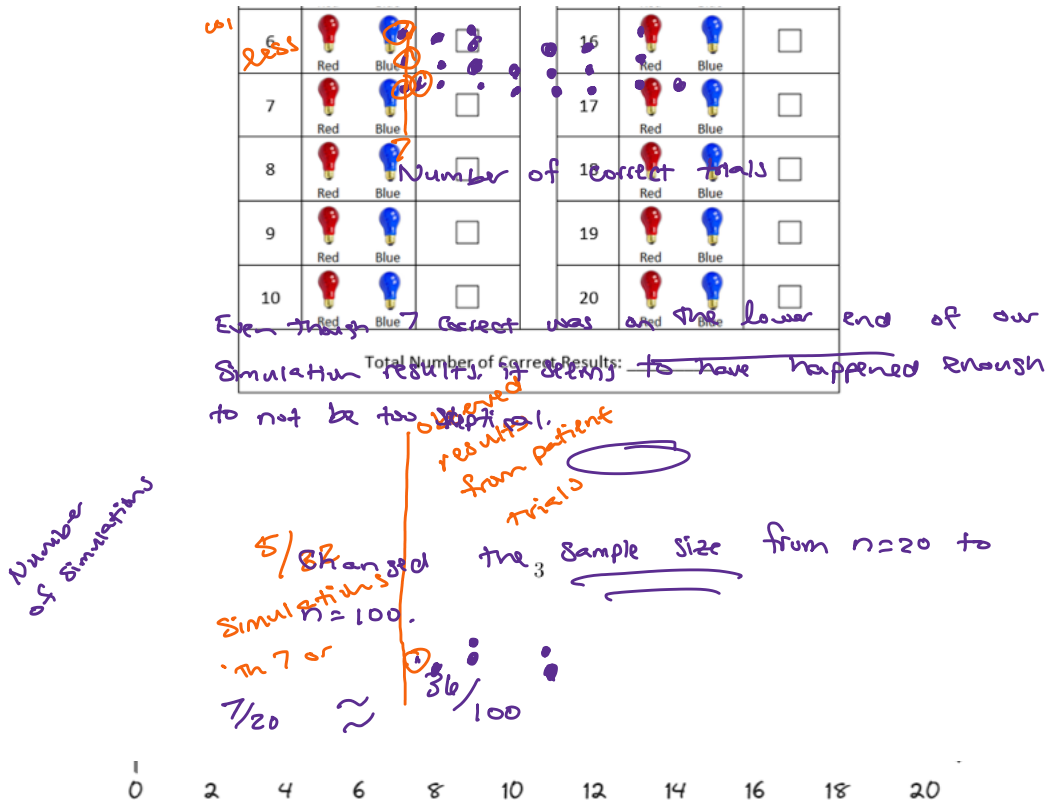
Even though 7 correct was at the lower end of our simulation results, it still ~~is~~ to have happened enough to not be too unusual.

Number of simulations

5/8 arranged Simulations n=100.

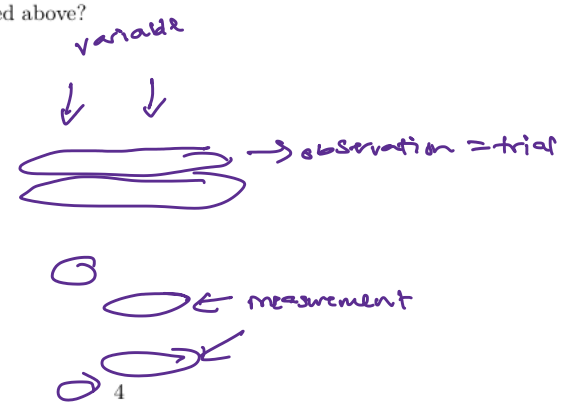
observed results from patient trials

the 3 sample size from n=20 to



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?

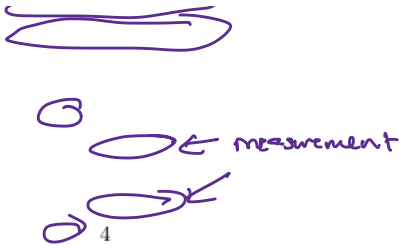
5. In reality, the subject was correct on 36 out of 100 trials when they were asked to identify whether the tone played with either the red or the blue lightbulb. How does this change the process we just conducted above?



17	Incorrect
18	Incorrect
19	Correct
20	Correct

Tidy Data

When working with statistical research questions, the information is usually stored in a data



17	Incorrect
18	Incorrect
19	Correct
20	Correct

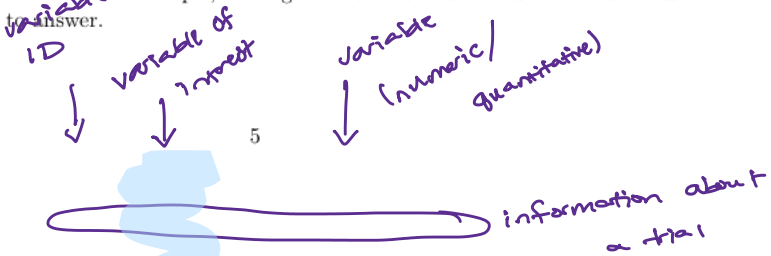
Tidy Data

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:

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.



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

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

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?

   Infants might have a preference for certain colors, shape, or sides. We want to eliminate this bias, to parse out if it is due to helper/hinder, not color/shape/side.

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

- Population of interest: All 10-month old infants
- Sample: n = 16 10-month old infants who participated in the study
- Variable of interest: character (helper/hinderer)
- Data type: categorical

Infant	Character
1	helper
2	helper
3	hinderer
4	
5	
6	

categorical

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

- Population of interest: All 10-month old infants
- Sample: n = 16 10-month old infants who participated in the study
- Variable of interest: character (helper/hinderer)
- Data type: categorical

Infant	Character
1	helper
2	helper
3	hinderer
4	
5	
6	

16

- 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.

about 8 infants to choose helper (50% of 16)

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?

10 infants out of 16 seems like it could happen by chance (close to 8).

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,

helper/hinderer

0.50

16

14 observed

10 observed

sample results

100 dots on this plot

# of simulated samples

somewhat surprising, or not so surprising if the population of all 10-month-old infants has no real preference.

5 14 helpers

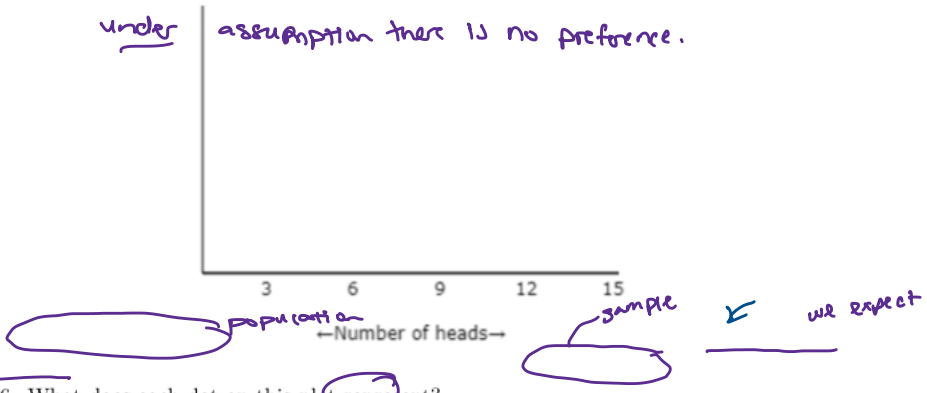
of all 10-month-old infants. To answer this, we will simulate the process of 16 infants simply choosing a toy at random over and over again. Each time we simulate the process, we'll keep track of how many infants out of the 16 chose the helper toy (note that you could also keep track of the number that 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:

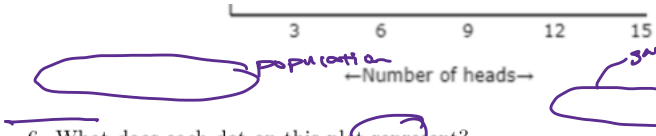
- What are the two possible outcomes on each of the trials?
- What is the probability that the helper is selected under the assumption that the population of all 10-month-old infants has no real preference for either toy? Change your `Probability of heads` accordingly.
- How many infants were used in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



6. What does each dot on this plot represent?

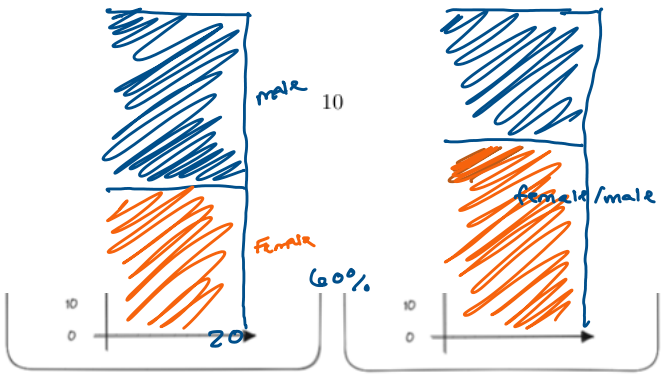
7. Suppose that in the actual study 10 out of 16 infants chose the helper toy. Would this convince you that the majority of the population of all 10-month-old infants had a preference for the helper toy? Why or why not?

40% male; however, since the management program began, 9 of the 20 employees chosen for management training were female (only 45%). Do the female employees have a valid statistical argument that they are being discriminated against?

? **Research Question:** *Is there statistical evidence for gender discrimination against females?*

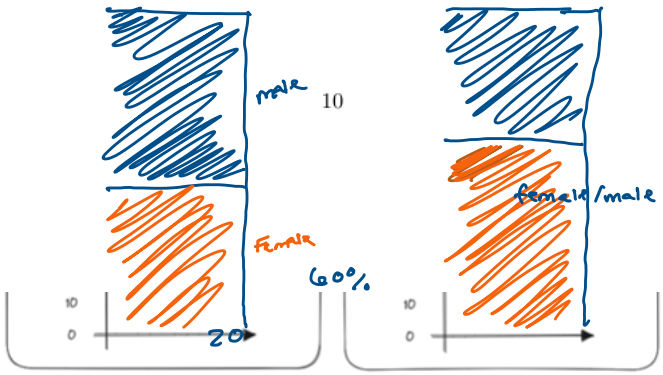
1. Identify the following in the context of this example:

- Population of interest:
- Sample:
- Variable of interest:
- Data type:



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

- Data type:



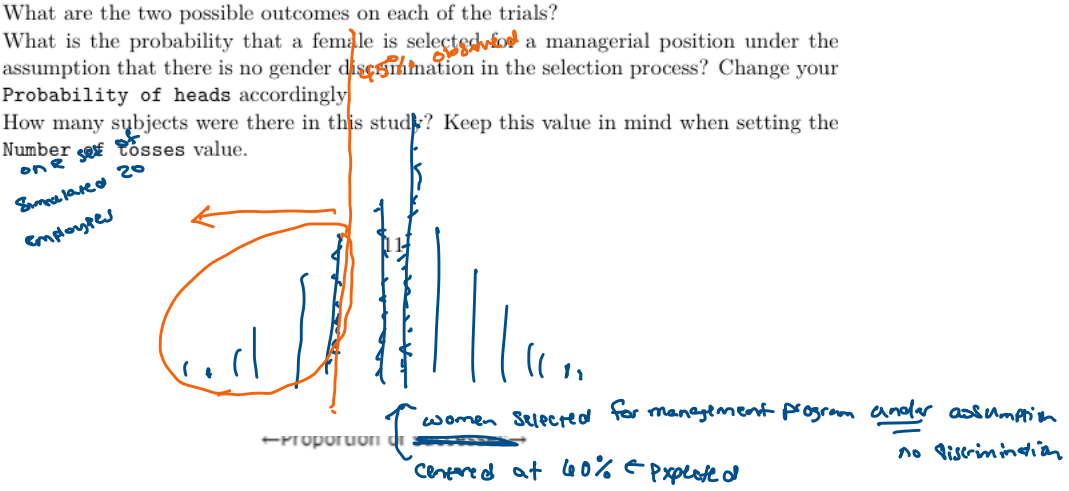
3. If the selection process was unbiased, how many of the 20 employees selected for management do you expect to be women? Explain.

12 to be female (20 × 0.6) or 60% of 20

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 female. 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 female is selected for a managerial position under the assumption that there is no gender 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.



4. What does each dot on this plot represent?

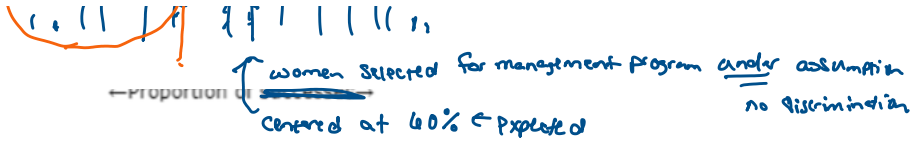
one set of 20 employees under the assumption no discrimination

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

Not so surprising, 45% or less occurs fairly often under the assumption no discrimination.

Example 1.4: Font Preferences

Researchers carried out a marketing field study in order to study preferences of all 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



4. What does each dot on this plot represent?

one set of 20 employees under the assumption no discrimination

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

Not so surprising, 45% or less occurs fairly often under the assumption no discrimination.

Example 1.4: Font Preferences

Researchers carried out a marketing field study in order to study preferences of all 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 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: All consumers
- Sample: n = 40 subjects in the study
- Variable of interest: Font (Salem/Signet)

  20, about half
- Data type: categorical

Salem / Signet

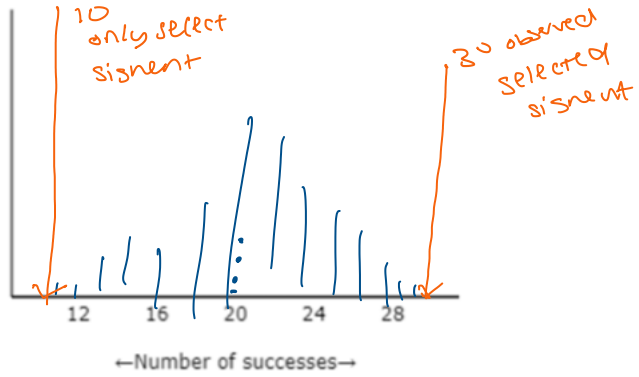
1. Identify the following in the context of this example:

- Population of interest: All consumers
- Sample: n = 40 subjects in the study
- Variable of interest: Font (Salem/Signet)
  20, about half
- Data type: Categorical

Salem / Signet

0.5

Signet on each of the simulated experiments. Sketch in your results below:



3. What does each dot on this plot represent?
   one set of 40 selections under the assumption there is no font preference.

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?

References

Binder, L. 1993. "Malingering Detected by Forced Choice Testing of Memory and Tactile

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?

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.

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.*