Investigation Brief: Investigating A Single Categorical Variable

Learning Objectives and Possible Standards

Learning how to explore and visualize a single categorical variable begins in elementary grades in conjunction with sorting, categorizing and counting things. At the high school level we predominantly see categorical variables coming up in the context of probability and inference particularity in Math 3 and Math 4.

- **NC.M3.S-IC.4:** Use simulation to understand how samples can be used to estimate a population mean or proportion and how to determine a margin of error for the estimate.
- **NC.M4.SP.2.1:**Design a simulation to make a sampling distribution that can be used in making informal statistical inferences.
- **NC.M4.SP.2.2:** Construct confidence intervals of population proportions in the context of the data.
- **NC.M4.SP.2.3:** Implement a one proportion z-test to determine if an observed proportion is significantly different from a hypothesized proportion.
- **NC.M4.SP.3.3:** Recognize from simulations of sampling distributions of sample means and proportions that a normal distribution can be used as an approximate model in certain situations.

Consider Data

To model how you could link together the parts of the data investigative process for this type of analysis we will start by considering the data we will use. This is publicly available data that you could use with your students. The specific CODAP workspace I used to create the examples in this brief can be found here.

Here are some questions you can use to help guide your consideration of the data specific to investigating a single categorical variable

- What things (living or nonliving) are being measured? (i.e. what are the observational units?)
- How many things are being measured?
- Is this a sample or a population?
- What attributes in the dataset are categorical?
- How were the categorical attributes measured?

Investigative Brief: One Categorical

- How many outcomes are there for each categorical attribute?

In the case of our sample dataset there are four categorical attributes we could consider:

- **Sex:** reports each individual's biological sex. Categories are:
    - Male
    - Female
- **Educational-degree recode:** reports on the individual's level of educational attainment based on the highest degree completed, in years for which this information is available. Recodes a detailed version of the Education-years variable into new categories:
    - N/A or no schooling completed
    - Some schooling, no high school diploma
    - Completed Grade 12, diploma not identified
    - High school degree or GED
    - 1 or more years of college, no degree
    - 2-4 years of college, degree not identified
    - 5+ years of college, degree not identified
    - Associate's degree
    - Bachelor's degree
    - Master's or professional degree
    - Doctoral degree
- **Race_ethnicity-multi_recode:** reports each respondent's combined race and Hispanic ethnicity status, grouped into 7 primary categories. This variable draws on the Race-multi variable and therefore includes multiple race categories. Categories are:
    - Hispanic
    - Non-Hispanic White
    - Non-Hispanic Black
    - Non-Hispanic American Indian/Alaska Native
    - Non-Hispanic Asian or Pacific Islander
    - Non-Hispanic Other race
    - Non-Hispanic two or more major races
- Veteran_status indicates whether individuals served in the military forces of the U.S. (Army, Navy, Air Force, Marine Corps, or Coast Guard) in time of war or peace. Service includes active duty for any length of time and at any place at home or abroad. Categories are:
    - 0- N/A or not applicable
    - 1- Not a veteran
    - 2- Veteran
    - 3- Unknown

These descriptions can be found in the metadata file that comes with this dataset

Investigative Brief: One Categorical

Formulate Problem

When writing a question that involves the investigation of a single categorical variable we generally focus on the distribution of common statistics we can consider of this type of variable which are, frequencies and relative frequencies. If we have a sample that is representative of a population, which in this case we do, then you could also ask a question about estimating the population proportion.

Investigative Questions

Some possible questions

1. ***What is the racial/ethnic make-up of a sample of people living in the U.S. in 2020?***
2. ***What proportion of the population of the U.S. in 2020 identifies as female?***
3. ***What proportion of people who identify as veterans in the U.S. in 2020 have attained a bachelor's degree or higher?***

Process Data

Processing the data may be necessary for some of these questions but not all. We can use various data moves to process data. One example in this case is for the question, “What proportion of people who identify as veterans in the U.S. in 2020 have attained a bachelor's degree or higher?” One data move we need to carry out is to filter out the individuals that do not identify as veterans. This could be done by hiding selected cases using a graph. We also need to possibly create a new variable that is binary for educational attainment at a bachelor's degree or above or not.

Explore/Visualize

When we explore and visualize we are beginning to make sense of the data by looking at visualizations of the data as well as finding descriptive statistics.

1. What is the racial/ethnic make-up of a sample of people living in the U.S. in 2020?

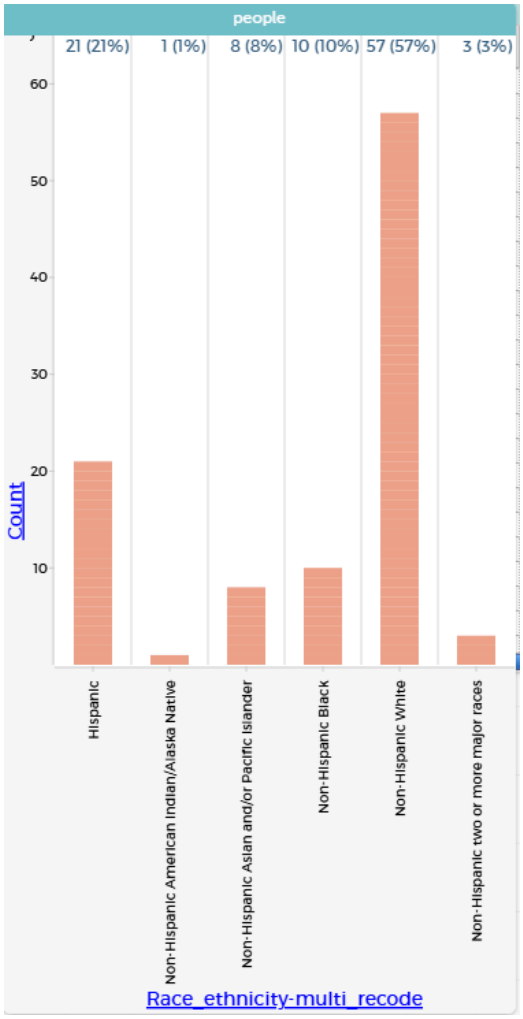

Figure 1. *Bar graph of frequencies of Race/Ethnicity from a sample of about 100 people who completed the census in 2020.*

In the case of a categorical variable the main way we visualize the data is using a bar graph. Some people like to use pie charts but they are problematic in terms of interpretation. The two main types of descriptive statistics are frequencies (counts) and relative frequencies (proportions), which can also be included in the graph or in a separate table depending on preference. Since the question asks about the sample we do not need to go much beyond this exploration for this question.

2. What proportion of the population of the U.S. in 2020 identifies as female?

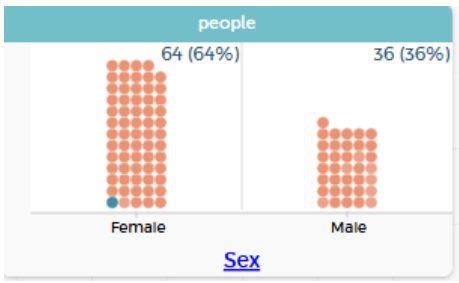

Figure 2. *Bar graph of relative frequencies of Sex from a sample of about 100 people who completed the census in 2020.*

Investigative Brief: One Categorical

This question is very similar in terms of exploration and visualization as the last question. We will also need to think inferentially about this question when we start to model it.

3. What proportion of people who identify as veterans in the U.S. in 2020 have attained a bachelors degree or higher?

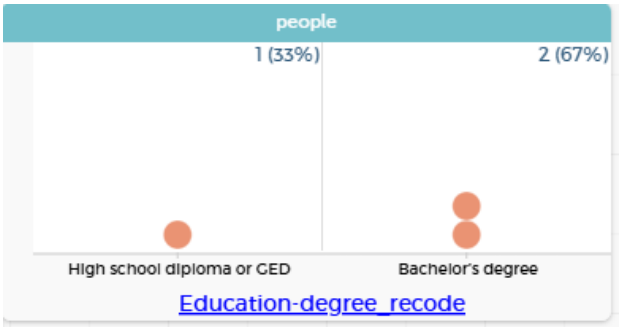

Figure 3. *Bar graph of relative frequencies of veterans' Education degrees (only high school and bachelors) from a sample of about 100 people who completed the census in 2020.*

This question is very similar in terms of exploration and visualization as the last question; however there are very few observations to consider after the filtering done during the process data stage. Something to consider at this point is whether it might be best to go back and collect more data as modeling with a sample this small is problematic.

Consider Models

1. What is the racial/ethnic make-up of a sample of people living in the U.S. in 2020?
In descriptive questions our models are generally just descriptive statistics of the distribution of a variable. Relative frequencies are commonly what we report when describing the distribution of a categorical variable but at times if it makes more sense for the question or data you may want to report frequencies. You may organize this information into a table to more clearly communicate it.

2. What proportion of the population of the U.S. in 2020 identifies as female?
We can take two different approaches to model this data. One is using a confidence interval that could be found by simulations or resampling (i.e. bootstrapping). For example in the simulation of resampling in the figure below I collected 100 samples of 100 individuals with replacement to get a sense of the variability in the proportion of individuals identifying as female from sample to sample.

Investigative Brief: One Categorical

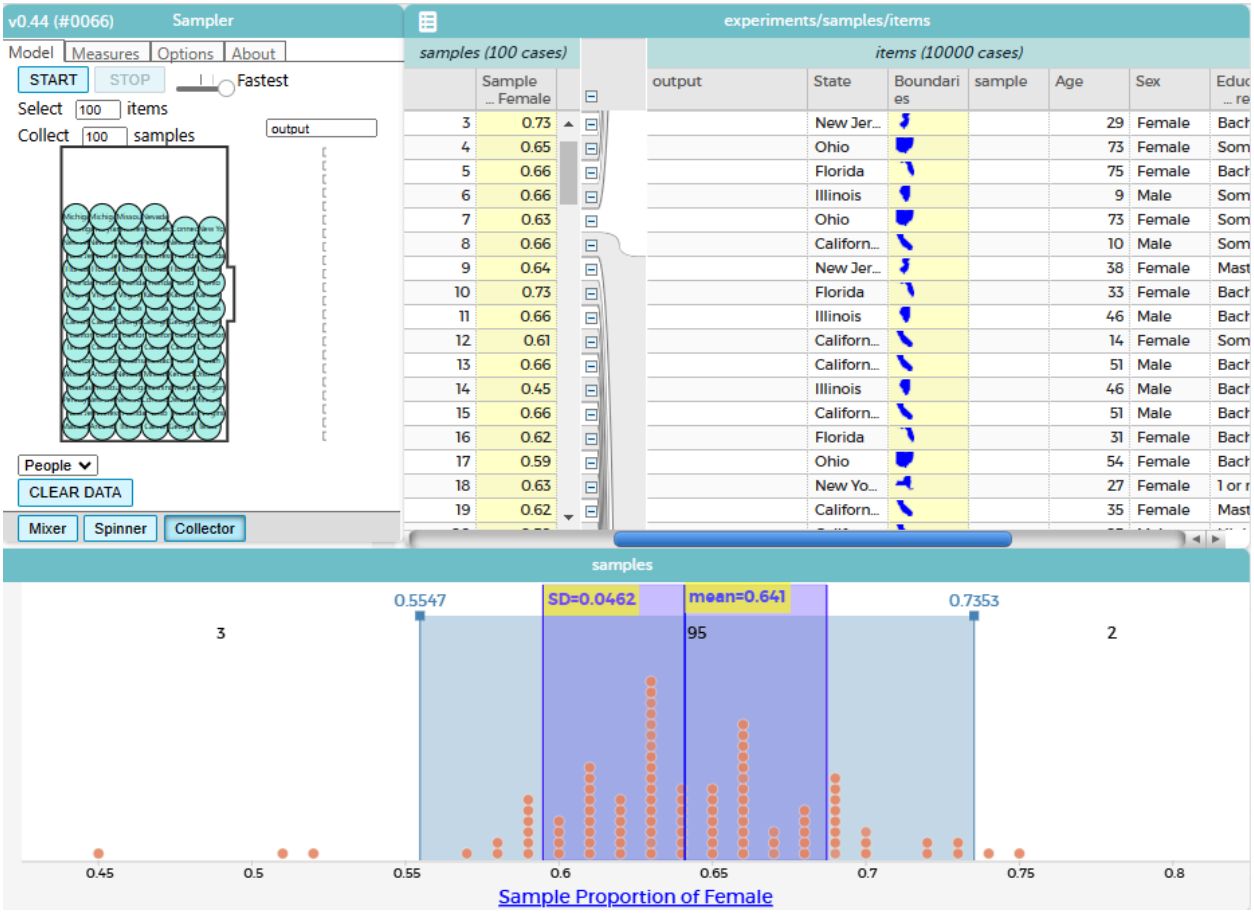

Figure 4. *Dot plot showing the distribution of Sample Proportion of Female using the bootstrap method in CODAP's sampler plug-in.*

Another approach is by using the normal distribution and the model $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ *where z** represents the critical value from the standard normal distribution corresponding to the desired confidence level. In other words for a 95% confidence interval I could do $.64 \pm 1.96\sqrt{\frac{.64(1-.64)}{100}}$ $=.64 \pm .094 = (.546, .734)$.

Inference in CODAP

By clicking on Plug-Ins, you can access Inference tools in CODAP. Click "Testimate", and drag and drop Sex into the applet. You should see the corresponding Confidence Interval, test statistic, critical value and p value. This confirms the calculations we did in the previous sections.

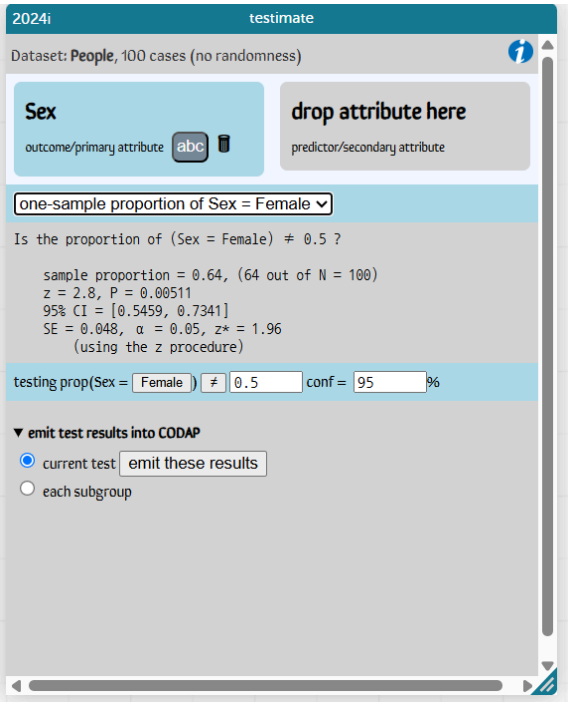

Figure 5. *Screenshot of using the Testimate app in CODAP to generate the confidence interval.*

3. What proportion of people who identify as veterans in the U.S. in 2020 have attained a Bachelor's degree or higher?

Ideally we would model this data using a confidence interval that could be found either by simulations or resampling (i.e. bootstrapping) or by using the normal distribution and the model $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ *where z** represents the critical value from the standard normal distribution corresponding to the desired confidence level. However, with only 3 observations a normal distribution does not model the data well. A simulation could provide some estimate but it also problematic in this case.

Communicate and Propose Action

1. What is the racial/ethnic make-up of a sample of people living in the U.S. in 2020?

In descriptive questions our communication is generally just descriptive statistics of the distribution of a variable. In this case using the graph about we could say something like,

> *In our sample of 100 individuals who completed the census in 2020 there are six different possible racial/ethnic identities in the dataset with the largest proportion of the sample identifying as non-Hispanic white (57%), followed by Hispanic (21%), then non-Hispanic Black (10%) and non-Hispanic Asian and/or Pacific Islander (8%), and small proportions of non-Hispanic two or more races (3%) and non-Hispanic American Indian (1%).*

You may choose to add some context to this response depending on the question and your knowledge of the context.

Investigative Brief: One Categorical

2. What proportion of the population of the U.S. in 2020 identifies as female?

Depending on what model you chose to use for making an estimate towards the population you would describe it slightly differently.

For the simulation/bootstrapping approach:

> *In our representative sample of 100 individuals who completed the census in 2020 64% identified as female. In carrying out a simulation of taking 100 samples of 100 individuals from the original sample with replacement we found that about 95% of the samples fell between 55% and 73% identifying as female.*

For the normal distribution model approach:

> *In our representative sample of 100 individuals who completed the census in 2020 64% identified as female. Using a normal distribution to model the variability from sample to sample we created a 95% confidence interval of the proportion of the population identifying as female to be (.546, .734). This interval suggests that the true proportion of people in the U.S. in 2020 identifying as female could be as low as 54.6% or as high as 73.4%, based on a 95% confidence level.*

3. What proportion of people who identify as veterans in the U.S. in 2020 have attained a bachelors degree or higher?

Because of the small number of outcomes associated with people identifying as a veteran there are not good models for this size. Therefore the main action we would recommend is to collect more data.

Here are some sentence stems for interpreting a confidence interval for a sample proportion:

Basic Interpretation:

"We are [confidence level]% confident that the true proportion of [population] is between [lower bound] and [upper bound]."

Addressing the Margin of Error (if applicable):

- "The proportion of [population] is estimated to be [point estimate] with a margin of error of [margin of error] at a [confidence level]% confidence level." (This can be followed by the interval calculation.)
- "At a [confidence level]% confidence level, we estimate that the proportion of [population] is [point estimate] plus or minus [margin of error]."

More nuanced interpretations (avoiding common mistakes):

- "If we were to repeat this study many times, [confidence level]% of the confidence intervals created would contain the true proportion of [population]." (Focuses on the process, not a probability about the true proportion itself.)
- "This interval suggests that the true proportion of [population] could be as low as [lower bound] or as high as [upper bound], and we are [confidence level]% confident in this range." (Acknowledges uncertainty within the interval.)

To cite this document

Weiland, T., & Ireland, C. (2025). Investigation brief: Investigating a single categorical variable. Online resource through SLiCE. Available at: https://docs.google.com/document/d/1EOkErbat2GUqNjoeEMgkuR2ElMVfDUg2/edit?usp=sharing&ouid=105451572967577654701&rtpof=true&sd=true

Acknowledgement

This material is based on work supported by the National Science Foundation under DRK-12 Grant No. 2143816 & 2517085. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.