Investigation Brief: Investigating A Single Quantitative and Categorical Variable

Learning Objectives and Possible Standards

Learning how to explore and visualize a single quantitative variable begins in elementary grades in conjunction with representing data using concrete objects, pictures, tables and graphs. At the high school level we predominantly see quantitative variables coming up in the context of describing distributions and representing data in Math 1. In Math 3 and Math 4, we begin drawing inferences from one quantitative variable. We also can utilize sampling distributions to help infer from the sample to the population.

- **NC.M1.S-ID.1** Use technology to represent data with plots on the real number line (histograms, and box plots).
- **NC.M1.S-ID.2** Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets. Interpret differences in shape, center, and spread in the context of the data sets.
- **NC.M1.S-ID.3** Examine the effects of extreme data points (outliers) on shape, center, and/or spread.
- **NC.M3.S-IC.1** Understand the process of making inferences about a population based on a random sample from that population.
- **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.1.3** Organize large datasets of real-world contexts (i.e. datasets that include 3 or more measures and have sample sizes >200) using technology (e.g., spreadsheets, dynamic data analysis tools) to determine: types of variables in the data set, possible outcomes for each variable, statistical questions that could be asked of the data, and types of numerical and graphical summaries could be used to make sense of the data.
- **NC.M4.SP.2.1** Design a simulation to make a sampling distribution that can be used in making informal statistical inferences.

Consider Data

Investigative Brief: One Quantitative

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 quantitative variable

- What types of quantitative data do we have?
- What are the possible outcomes for the quantitative data? Can you talk about the range of the data?
- Is this a sample or a population?
- What attributes in the dataset are quantitative?
- How were the quantitative attributes measured? What are the units of measurement?

In the case of our sample dataset there are two quantitative attributes we could consider:

- **Age:** reports the Individual's age in years as of the last birthday. Values range from 0 (less than 1 year old) to 90 and above.

  Special codes:
  All years: 0 = less than 1 year old
  90 = 90 years old and older
  115 = 115 years old and older

- **Income:** reports on each respondent's total pre-tax income.Total income is the sum of the amounts reported for multiple types of income, including wage or salary income; net self-employment income; interest, dividends, or net rental or royalty income or income from estates and trusts; Social Security or Railroad Retirement income; Supplemental Security Income (SSI); public assistance or welfare payments; retirement, survivor, or disability pensions; and all other income. Amounts are expressed in contemporary dollars.

  Income-total is a 7-digit figure with the following special codes:

  0000000 = None
  0000001 = $1 or break even (2000, 2005-onward ACS)
  9999999 = N/A

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



Formulate Problem

Investigative Brief: One Quantitative

When writing a question that involves the investigation of a single quantitative variable we generally focus on the distribution of common statistics we can consider of this type of variable which are: mean, median, mode, range, interquartile range, standard deviation, and variance. 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 mean.

Investigative Questions

Some possible questions

1. ***What is the typical income for the sample of people living in the U.S. in 2020?***
2. ***How are the ages of a sample of people living in the U.S. in 2020 distributed?***
3. ***What is the estimated mean age for the U.S. population in 2020?***



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 is the typical income for the sample of people living in the U.S. in 2020?*** One data move we need to carry out is to filter out the individuals that have the value 9999999 for their income, which is a special code for Not Applicable. We can filter out these individuals because if we look in the metadata file that comes with this dataset we can see that that value is not an income, but it is used to signify missing data, which we do not want to include in our statistics. This could be done by hiding selected cases using a graph. This allows us to remove outliers to see the spread of the rest of the distribution better. We might also want to filter out people on the younger end of the age range as they are not old enough to work legally in the workforce. It is up to the investigator to decide what data to include for analysis or exclude, but all those decisions should be described transparently later on.



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. Note that in CODAP, the red line indicates the median and the blue line represents the mean. The shaded regions denote one standard deviation above and below the mean.

1. What is the typical income for the sample of people living in the U.S. in 2020?

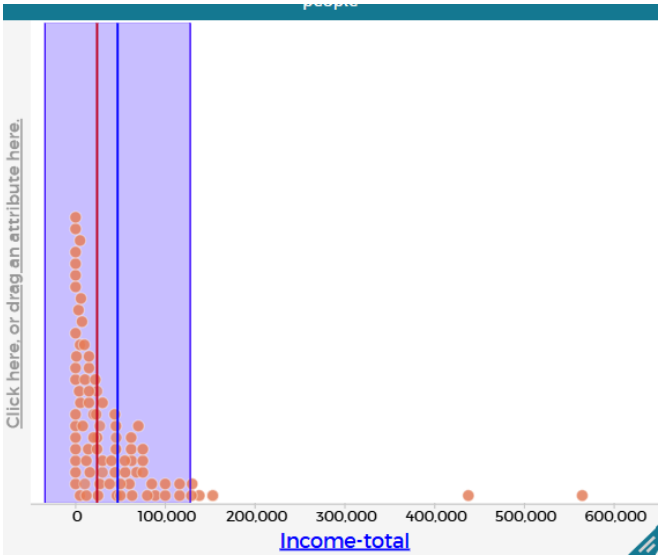


Figure 1. *Dot plot of the total incomes of a sample of about 100 people who completed the census in 2020. The blue line marks the mean, the red line the median and the blue shaded region is one standard deviation to the left and right of the mean.*

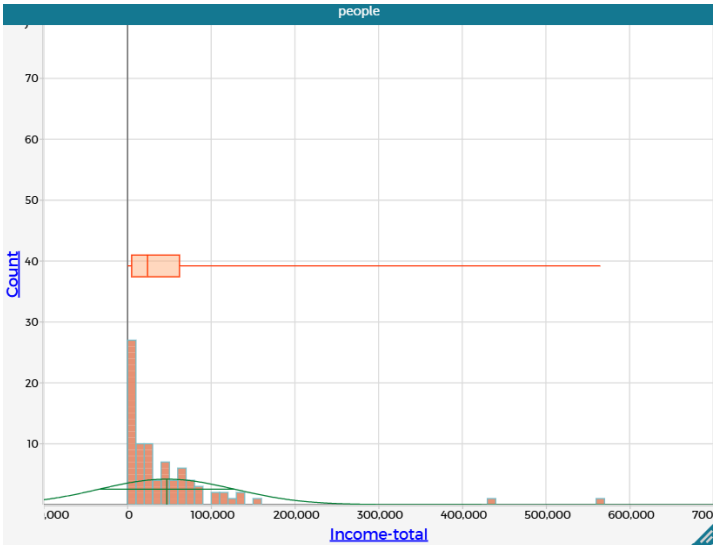


Figure 2. *Histogram with a normal distribution overlay in green and box plot of the total incomes of a sample of about 100 people who completed the census in 2020.*

In the case of a quantitative variable, the main way we visualize the data is using a dot plot or histogram. We can also include a boxplot to help visualize the five number summary which includes: minimum value, lower quartile (Q1), median, upper quartile (Q3), and maximum value.

Table 1. Income distribution descriptive statistics							
Minimum	Q1	Median	Mean	Q3	Maximum	Range	Standard Deviation
0	$5000	$24000	$46933	$62000	$565000	$565000	$80964

Investigative Brief: One Quantitative

One important issue to consider for this exploration is which measure of center more accurately represents the center of the distribution or what is typical? Because the distribution is skewed to the right, the mean is skewed to the right of the center. In such cases a median may be a more appropriate measure of center to consider in relation to the spread of the distribution. Notice how the histogram does not follow the shape of the Normal distribution curve overlay.

2. How are the ages of a sample of people living in the U.S. in 2020 distributed?

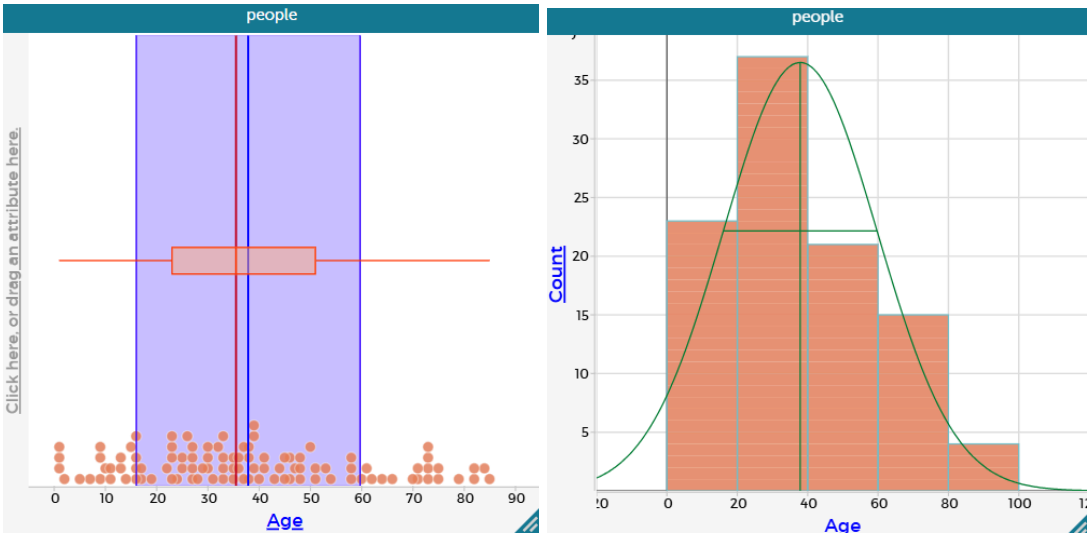


Figure 3. *Left, dot plot of the ages of a sample of 100 people who completed the census in 2020. The blue line marks the mean, the red line the median and the blue shaded region is one standard deviation to the left and right of the mean. Right, histogram with a normal distribution overlay in green and box plot of the ages of a sample of 100 people who completed the census in 2020.*

This question is very similar in terms of exploration and visualization as the last question. The main types of descriptive statistics are mean, median, range, and standard deviation, which can also be included in the graph or in a separate table depending on preference.

Table 2. Age distribution descriptive statistics

Minimum	Q1	Mean	Median	Q3	Maximum	Range	Standard Deviation
1	23	35.5	37.9	51	85	84	21.9

Since the question asks about the sample, we do not need to go much beyond this exploration for this question.

3. What is the estimated mean age for the U.S. population in 2020?

This question is very similar in terms of exploration and visualization as the previous questions; however, this time we create a sampling distribution from the sample mean, using a simulation that gathers 1000 sample means. By simulating a sample distribution, we can begin to infer the true mean of the population based on our sample, which we will consider in the modeling section.



Consider Models

1. What is the typical income for the sample of people living in the U.S. in 2020?
2. How are the ages of a sample of people living in the U.S. in 2020 distributed?
In descriptive questions, our models are generally just descriptions of the distribution of a variable using data visualizations and descriptive statistics. In our model measures of central tendency (such as mean) generally serve as the signal and measures of spread (such as standard deviation) describe the noise or error in the model. Depending on the question or data, you may want to report more detailed measures (such as interquartile range, or variance for measures of spread) to provide a clearer picture of the distribution of the data. Or, you may consider discussing the comparison between mean and median, and how they are impacted by the skewness of the distribution. In such cases the median may be a better model in conjunction with the IQR. in other words the model is essentially a description of the distribution based on the analysis from the previous section and communicate in the next section.

3. What is the estimated mean age for the U.S. population in 2020?
The main approach to creating a model for this question is using a confidence interval that could be found by simulations or resampling (i.e. bootstrapping). Because we are making an inference from a sample to a population we do not know exactly what the population mean is, but we can use a sample to make an educated estimate. To do so we need a sense of how much variation in the sample means we might expect from sample to sample. In other words, we have one sample and one sample mean and if we went out and collected another random sample we would likely not get the exact same set of people or sample mean, but it should be close. We need to figure out what variation we might see by chance from collecting lots of samples so that we can then make an educated estimate of what is likely the population mean.

To carry out repeated sampling we can use CODAP, which allows us to quickly collect random samples (with replacement) from a sample (that we treat as a population for this process) to then see what would happen by chance (see Figure 4). We can then create a new attribute in the sample level of the table created from the sampler and calculate the mean for each sample to then create a dotplot of the sample means (see Figure 5). From our dotplot we can see the mean of the sampling distribution of sample means is 38.32 years and the standard deviation is 2.37. We can then add movable lines to determine how far to the left and right of the mean we need to go to capture the middle 95% of the sample means to create an informal 95% confidence interval, which in the case of our simulation comes out to be approximately (34.07,42.59) as shown in Figure 5. You could also use the empirical rule assuming the sampling distribution could be modeled using the normal distribution which would then tell us 95% of the observations are about 2 standard deviations from the mean so (33.58, 43.06). As a note different simulations will create slightly different results by chance.

Investigative Brief: One Quantitative

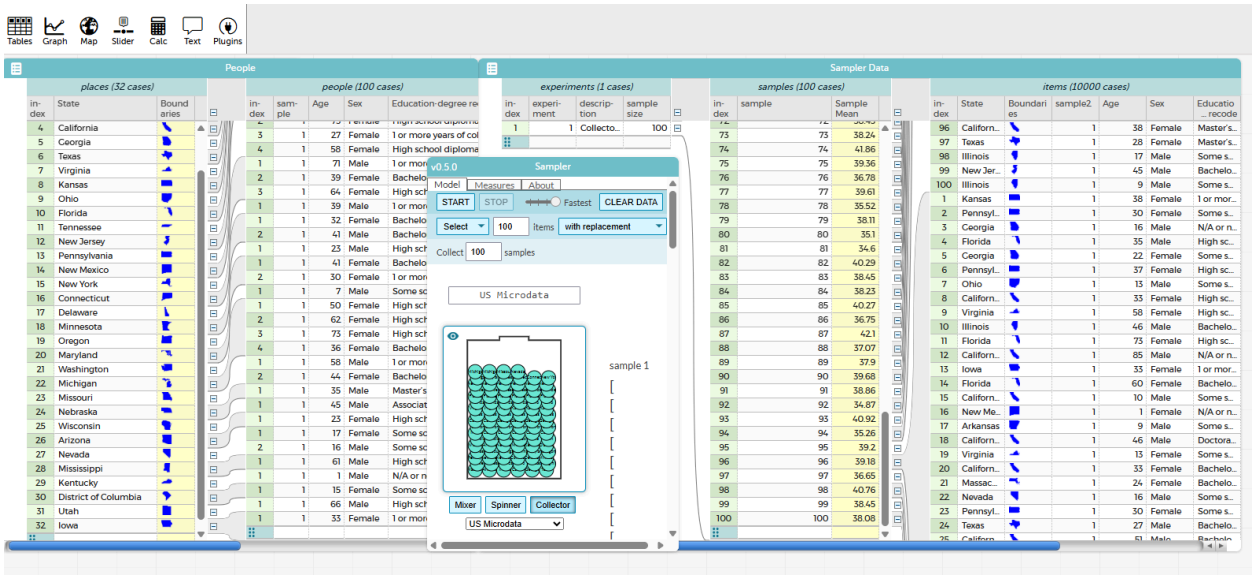


Figure 4. *View of the workspace in CODAP with the sampler to collect 100 samples of 100 individuals sampled with replacement.*

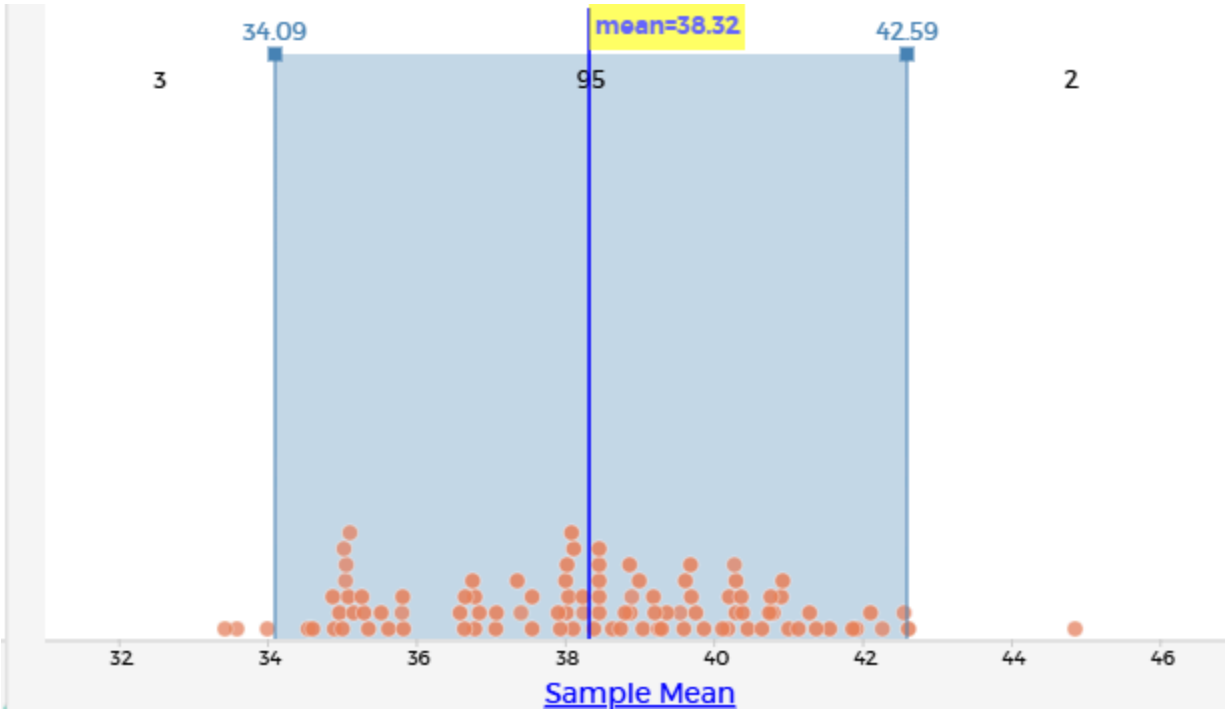


Figure 5. *Dot plot of a sampling distribution of sample means of the ages of 100 individuals who lived in the U.S. in 2020 from running the collector sampler in CODAP, shown in Figure 4. Each point represents a sample mean of 100 individuals. The blue line in the center makes the mean of the sampling distribution and the light blue lines to the left and right of the mean are movable lines that were moved to mark the center 95 sample means.*

Looking Ahead

Though it is not in the Math 1-4 standards in AP Statistics and beyond one of the topics discussed is calculating a confidence interval similar to what we do cover for population proportions. We share that here for

this example so you have a sense of where this learning goes but it is not something you are expected to teach in the math 1-4 standards. In this approach, we use the t-distribution curve and the model $\bar{x} \pm t^*(s/\sqrt{n})$ *where t** represents the critical value from the t-distribution corresponding to the desired confidence level and degrees of freedom. Since the sample size is 100, the degrees of freedom are 99. In other words for a 95% confidence interval, I would obtain a t critical value of 1.984 and use the formula: $38.3 \pm 1.984(21.3/\sqrt{100})$ = 38.3 +/- 4.226 = (34.074, 42.526) to get an estimate of the true mean for the age of the population.



Communicate and Propose Action

1. What is the typical income for the 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 85 individuals who completed the census in 2020, the distribution of income is skewed right, with a mean of $46,933 and a median of $24,000, and a standard deviation of $80,964. The middle 50% of incomes range from $5,000 to $62,000 with the whole distribution ranging from $0 to $565000. There are a few large outliers contributing to the skewing of the distribution.*

You may choose to add some context to this response depending on the question and your knowledge of the context. You may also choose to discuss that the median is a better representation of center in this case since the distribution is skewed and discuss the impacts of the outliers on measures of spread as well.

2. How are the ages of a sample of people living in the U.S. in 2020 distributed?

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, the distribution of age is approximately symmetric, with a mean of 37.9 years old, and a standard deviation of 21.9 years. There are no obvious outliers.*

You may choose to add some context to this response depending on the question and your knowledge of the context. You may also choose to discuss the median being slightly lower at 35.5, or include range or IQR instead of standard deviation.

3. What is the estimated mean age for the U.S. population in 2020?

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, the mean age was 37.9. Using a simulation of resampling 100 samples of size 100 with replacement from our original sample we can create an informal confidence interval to estimate that the true mean age of the US population is between 34.09 and 42.57. If we were to repeat this study many times, 95% of the confidence intervals created would contain the true mean age of the population of the U.S.*

For the t-distribution model approach:

Investigative Brief: One Quantitative

In our representative sample of 100 individuals who completed the census in 2020, the mean age was 37.9. Using a t-distribution to model the variability from sample to sample we created a 95% confidence interval of the mean age of the population to be (34.074, 42.526). This interval suggests that the true mean age of people in the U.S. in 2020 could be as low as 34.074 or as high as 42.526, based on a 95% confidence level.

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

Basic Interpretation:

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

Addressing the Margin of Error (if applicable):

- "The mean 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 mean 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 mean of [population]." (Focuses on the process, not a probability about the true mean itself.)
- "This interval suggests that the true mean 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.)

Inference in CODAP

By clicking on Plug-Ins, you can access Inference tools in CODAP. Click “Testimate”, and drag and drop Age into the applet. You should see the corresponding Confidence Interval, test statistic, critical value and p value. These results are very similar to the calculations we did in the previous sections.

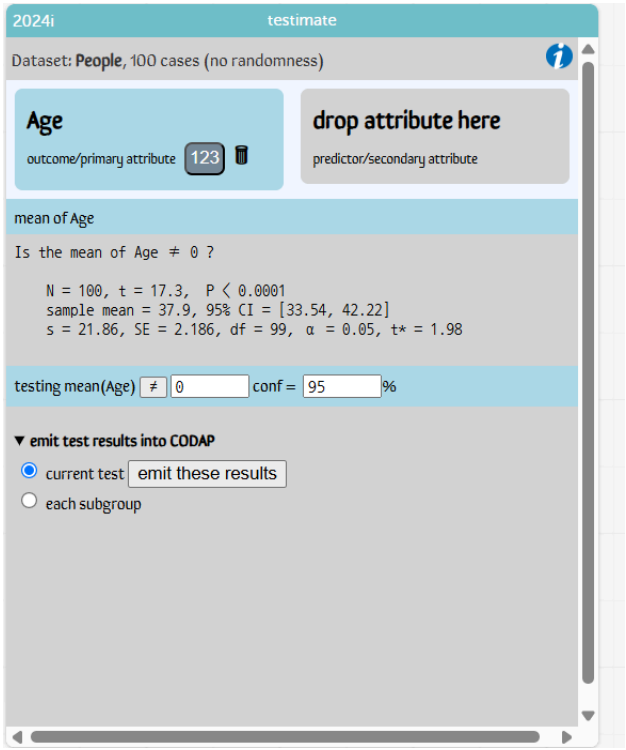


Figure 6. *Screenshot of using the Testimate app in CODAP to generate the confidence interval.*

This work is licensed under a CC-BY-NC-SA 4.0.
To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

To cite this document

Weiland, T., & Ireland, C. (2025). Investigation brief: Investigating a single quantitative variable. Online resource through SLiCE. Available at: https://docs.google.com/document/d/1pHW9PWkCpQFGkYcwC8_VtkJqRYgiBHsl/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.