Investigation Brief: Investigating Two Categorical Variables

Learning Objectives and Possible Standards

Learning how to explore and visualize two categorical variables begins in middle grades in conjunction with probability and looking for associations between variables. At the high school level we predominantly see categorical variables coming up in the context of probability in Math 2 (primarily with regards to conditional probability) and potentially in Math 4. As a note, this idea can build to the topic of the Chi-square test, which is covered in AP Statistics and AP Biology.

- **NC.M2.S-CP.4** Represent data on two categorical variables by constructing a two-way frequency table of data. Interpret the two-way table as a sample space to calculate conditional, joint and marginal probabilities. Use the table to decide if events are independent.
- **NC.M2.S-CP.5** Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations.
- **NC.M2.S-CP.6** Find the conditional probability of A given B as the fraction of B's outcomes that also belong to A, and interpret the answer in context.
- **NC.M2.S-CP.7** Apply the Addition Rule, P(A or B) = P(A) + P(B) – P(A and B), and interpret the answer in context.
- **NC.M2.S-CP.8** Apply the general Multiplication Rule P(A and B) = P(A)*P(B|A) = P(B)*P(A|B), and interpret the answer in context. Include the case where A and B are independent: P(A and B) = P(A)* P(B)
- **NC.M3.S-IC1** Understand the process of making inferences about a population based on a random sample from that population.
- **NC.M3.S-IC.5** Use simulation to determine whether observed differences between samples from two distinct populations indicate that the two populations are actually different in terms of a parameter of interest.
- **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.

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?
- How many outcomes are there for each categorical attribute?
- What variables might you expect a relationship between?

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 diploma 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

Investigative Brief: Two Categorical

- 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

Formulate Problem

When writing a question that involves the investigation of two categorical variables 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

1. ***Is there a relationship between the self identified sex and education level of people living in the U.S. in 2020?***
2.

Process Data

Processing the data may be necessary for some questions but not all. We can use various data moves to process data. One example for this question is that we need to create a new attribute that consolidates the original Education Recode variable into fewer options. For example, we could take the original 11 possible outcomes and group them into five as shown in Table 1. Figure 1 shows how to write a formula that regroups Education Level from the original groupings.

Some/No Schooling	• N/A or no schooling completed

	• Some schooling, no high school diploma • Completed Grade 12, diploma not identified
High School/GED	• High school diploma or GED
Some Postsecondary	• 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	• Bachelor's degree
Graduate	• Master's or professional degree • Doctoral degree

Table 1. *Organization for collapsing original education categories into the five categories we wanted for analysis.*

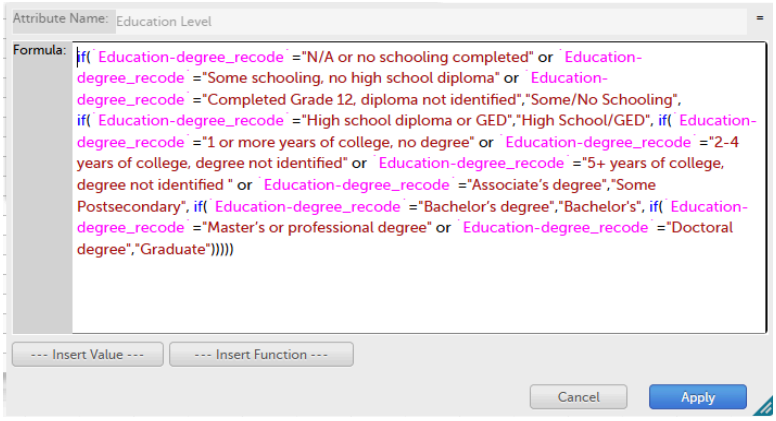

Figure 1. *Image from the CODAP workspace showing how to create a Formula for the Education attribute. If you would like to copy and paste the code, it is below:*

```
if(`Education-degree_recode`="N/A or no schooling completed" or
`Education-degree_recode`="Some schooling, no high school diploma" or
`Education-degree_recode`="Completed Grade 12, diploma not identified","Some/No Schooling",
if(`Education-degree_recode`="High school diploma or GED","High School/GED",
if(`Education-degree_recode`="1 or more years of college, no degree" or
`Education-degree_recode`="2-4 years of college, degree not identified" or
`Education-degree_recode`="5+ years of college, degree not identified " or
`Education-degree_recode`="Associate's degree","Some Postsecondary",
if(`Education-degree_recode`="Bachelor's degree","Bachelor's",
if(`Education-degree_recode`="Master's or professional degree" or
`Education-degree_recode`="Doctoral degree","Graduate")))))
```

Coding in CODAP Note:

The easiest way to condense the original education attribute into the five groups we want includes using nested if() functions and Boolean logic. Table 1 shows how we decided to regroup the original entries. To nest

the if() functions, we will need five for the new categories. Notice that once we open an if() function, we have to close it at the very end, so there will be five parentheses at the end of the code.

The if() function is reading the original education attribute to see if the new description in quotation marks “” matches. Each time we want it to check for a new description, we have to rewrite the attribute and put in quotations what we want the function to search for. It has to be exact! We can group all of the original categories into one instance of an if() function by connecting the commands with Boolean logic. Here, we want to use “or”. Thus, the first iteration of the if() function reads each entry in the education attribute to see if it contains the phrase “N/A or no schooling completed” OR “Some schooling, no high school diploma” OR “Completed Grade 12, diploma not identified”. If so, it returns the new category name “Some/No Schooling”. That ends the first run of the if() function.

Adding a comma will nest the next if() function which searches for instances of high school diploma or GED entries and recodes it as “High School/GED”. This pattern continues until we have recategorized all the entries from the original education attribute to the new Education Level attribute with fewer categories.

There is still one small problem. The apostrophe in the original coding for Associate’s, Bachelor’s and Master’s is read differently by the code than the apostrophe on our keyboard. So to fix this, copy and paste the original phrase into the formula for the new attribute. To access the workspace that has the attribute created already, click here.

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. ***Is there a relationship between the self identified sex and education level of people living in the U.S. in 2020?***

In the case of a categorical variable the main way we visualize the data is using bar graphs. 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.

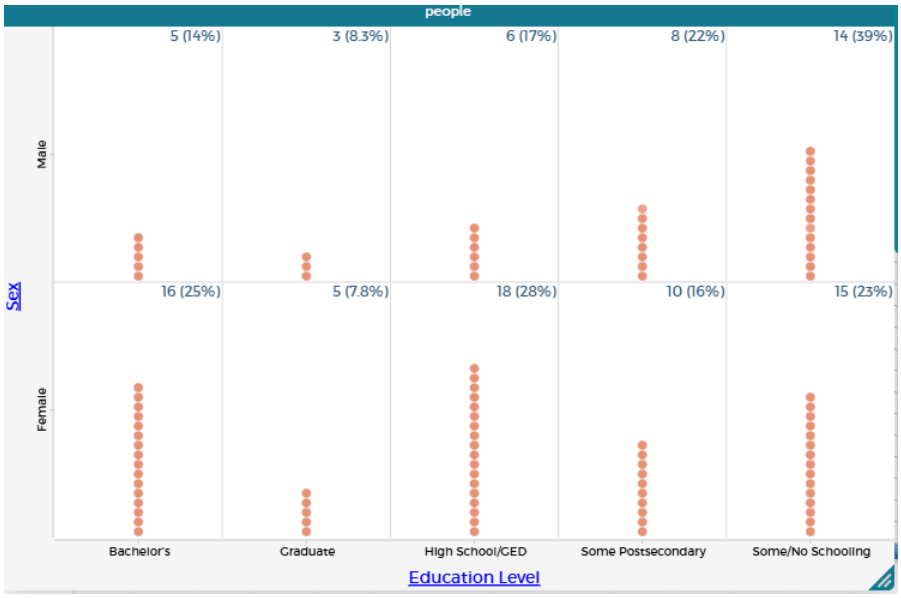

Figure 2. *Two-way table showing frequencies and relative frequencies partitioned by sex category and education category, with row percentages displayed.*

To answer this question, drag the Sex attribute and the variable we created in the Process section: Education Level, onto a graph. Click on the ruler to toggle counts and percentages on, and use row percentages to display the relative frequencies. Figure 2 shows row percentages while Figure 3 shows column percentages. Row percentages help compare across the educational categories in our example. So if we are looking at just males, we can quickly see that most males from this sample have some or no schooling.

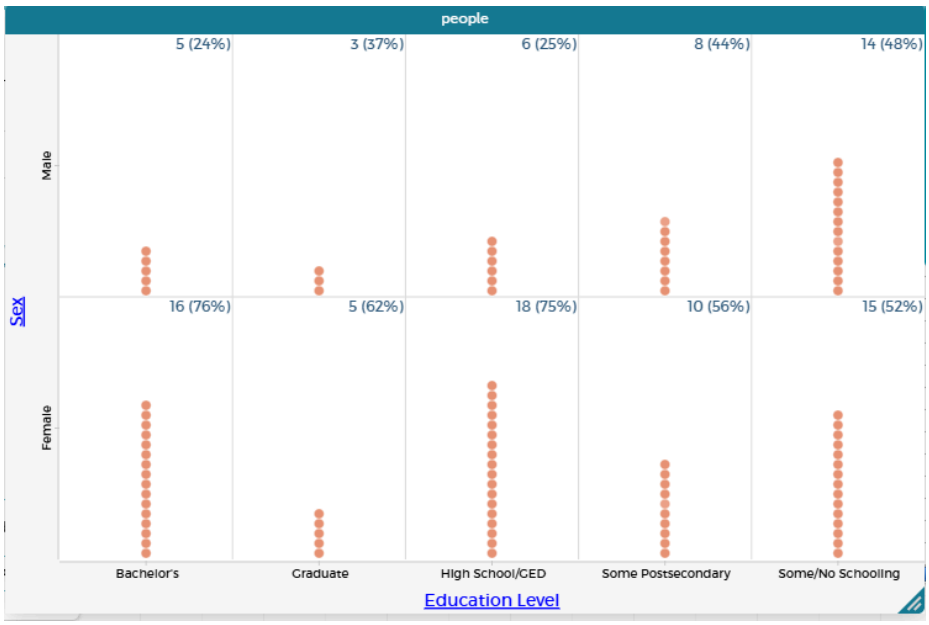

Figure 3. *Two-way table showing frequencies and relative frequencies by sex category and education category, with column percentages displayed.*

In our example, column percentages help compare between the sex categories. So 62% of females from the sample have a graduate degree while only 37% of men have a graduate degree. Figure 4 displays total cell

percentages. This version helps locate where the majority of the sample falls, so 16% of the sample is female with a graduate degree.

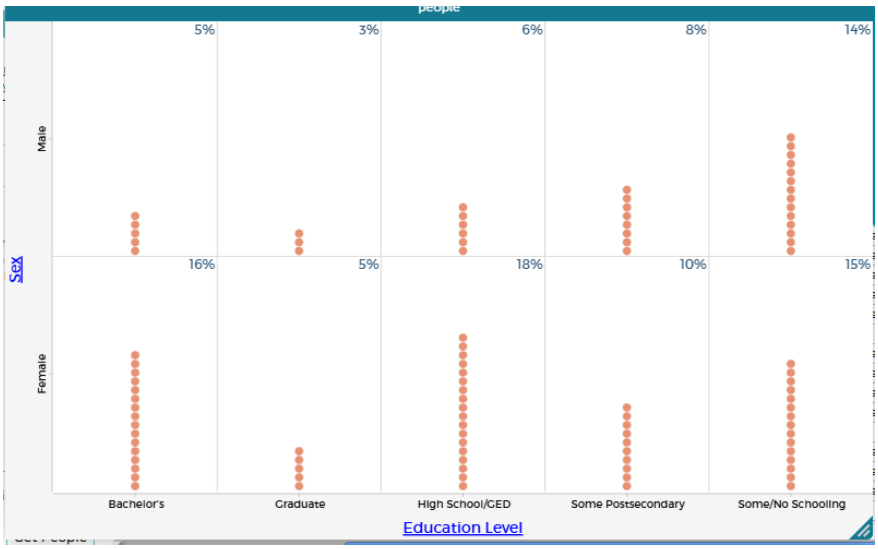

Figure 4. *Two-way table showing relative frequencies by sex category and education category, with total cell percentages displayed.*

Figure 5. *Stacked bar chart showing relative frequencies by sex category and education category with percentages displayed and the y-axis scaled by percent. This makes it easy to compare across the categories based on the proportion of the colors.*

Figure 5 displays a stacked bar graph scaled by percentages and Figure 6 shows a stacked bar graph scaled by counts. These charts help to visualize the proportion of males and females in each education category. From Figure 5 we can see that the proportion of males with some/no education and some post secondary is

larger than the proportion of females in those categories and we see an opposite relationship for the categories of bachelors and high school/GED. Counts would be preferred if you were looking for the number of people in that category. Percentages would be preferred if you were looking for a more proportional interpretation or in our case looking for relationships. In particular, when scanning from left to right of Figure 5 if there was no relationship the orange segments should be about the same height in each bar and likewise the purple segments. In our case it appears there is some evidence of a possible relationship due to the differing heights of the collared segments in each bar.

Figure 6. *Stacked bar chart showing relative frequencies by sex category and education category with percentages displayed and the y-axis scaled by count (frequency). This makes it easy to compare across the categories based on the proportion of the colors.*

Consider Models

1. ***Is there a relationship between the self identified sex and education level of people living in the U.S. in 2020?***

Informally, we want to look at the row percentages. Looking at row percentages from Figure 4, about 22% of males attained higher education (Bachelor's and Graduate), while about 33% of females attained higher education (Bachelor's and Graduate). This ties back into the ideas of conditional probabilities students will see in Math 2 and Math 3. Recall that conditional probability formula is P(A|B)=P(A and B)/P(B). Therefore, P(Bachelor's and Male + Graduate and Male | Males)=P(Higher Education)/P(Male). This becomes $\frac{\frac{5+3}{36}}{\frac{36}{100}} = \frac{8*100}{36*36} = \frac{800}{1296} = .6173$ or about 62%. Notice how in Figure 5, the column percentages show 24% and 37% for Bachelor's and Graduate for Males, respectively. This reinforces the conditional probability calculation above, and is interpreted as: Given a person identifies as Male, the probability that they obtained a Bachelor's degree or higher is about 61-62%.

Extension: Chi Square Test for Independence

AP Statistics looks at several Chi Square tests. The Test for Independence is the most applicable here. It determines if there is a significant relationship between the two variables in question. The null hypothesis is that they are independent of each other, while the alternative is that they are dependent on each other. Figure 7 shows the results of using the CODAP Testimate plug-in to run an independence test. Notice that the p value of .249 is large, greater than the significance level of .05. Therefore, the decision is to fail to reject the null hypothesis. Thus, there is evidence to believe that the variables are independent, or not related. You may consider asking yourself if this makes sense? Would you have expected education level and race to be related? Does this result surprise you? There are some underlying assumptions regarding count totals that would need to be discussed for further analysis.

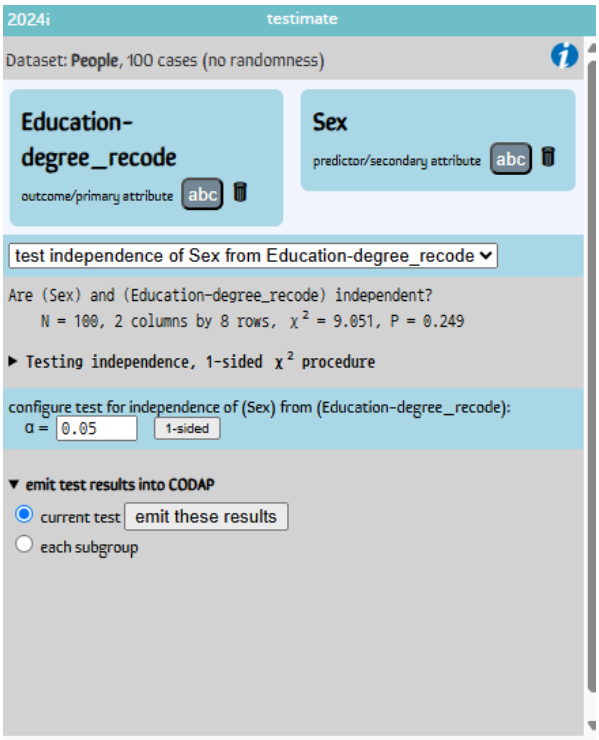

Figure 7. *Screenshot of using the Testimate app in CODAP to generate the test for independence results.*

Communicate and Propose Action

1. ***Is there a relationship between the self identified sex and education level of people living in the U.S. in 2020?***

From Figure 5 we can see that the proportion of males with some/no education and some post secondary is larger than the proportion of females in those categories and we see an opposite relationship for the categories of bachelors and high school/GED. Thus, there may be a relationship between sex and education level, with women more often than men being college educated.

Formally, a Two Sample Z-Test for Proportions hypothesis test could be conducted to determine if there is a difference in the proportions of education level based on sex. A Chi-Square test could also be conducted. Both of these topics are better suited for AP Statistics.

To cite this document
Weiland, T., & Ireland, C. (2025). Investigation brief: Investigating two categorical variables. Online resource through SLiCE. Available at:
https://docs.google.com/document/d/11aTFVLp_ACkCliO7nnvme28SENukCshY/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.