Program Evaluation Toolkit
Module 7, Chapter 2: Data Analysis Examples

Regional Educational Laboratory Central

From the National Center for Education Evaluation at IES

Speaker 1:

Welcome to the second chapter of module 7. In this chapter, you will explore some examples of common data analysis methods.

This chapter focuses on two relatively straightforward examples of data analysis, presented in relation to the fictitious AMMP! evaluation. Because the field of data analysis is extensive and significant expertise is often required for the various methods, covering the full spectrum of methods is beyond the scope of this toolkit. Instead, the toolkit focuses on some of the more common and accessible methods for evaluating education programs.

In the first example, you will choose quantitative methods to answer one of the AMMP! evaluation questions. In the second example, you will use qualitative methods to answer another evaluation question.

Let's start with a quantitative example. Consider the following AMMP! evaluation question: "How do AMMP! participants' scores on high school math placement tests compare to nonparticipants' scores?" This question involves examining test scores; therefore, a quantitative method is more suitable to address the question.

Let's imagine you are a member of the AMMP! evaluation team and see how you might address this question by using descriptive statistics. To get started, open the Microsoft Excel workbook titled *Descriptive Statistics Activity*, available on the resources page of the website. The first sheet in the workbook is a codebook that defines the variables used in the following activities. You will work across each of the sheets in this workbook over the next several slides.

Start by cleaning the data. You may want to review the handout *Common Sources of Data Errors and Error-Checking Techniques*, available on the resources page of the website, before you begin this activity.

In the *Descriptive Statistics Activity* workbook, click on the "Messy Data" tab. Consider the following questions:

- Do the data contain duplicates?
- Do the data contain missing values?
- Are there entry errors?
- Are there outliers?

Pause this video to review the data in the "Messy Data" sheet while considering these questions. You will review the data in more detail over the next several slides.

Begin by checking whether the data contain duplicates. In this case, you want to be sure that you aren't counting a student more than once. To check for duplicates in the "Messy Data" sheet, follow these steps: highlight the "student_id" column, choose "Conditional Formatting" on the Home ribbon, select "Highlight Cell Rules," and then click "Duplicate Values." Student IDs 13, 19, and 72 are then highlighted. Now the "Messy Data" sheet should look like the next sheet in the workbook, titled "Removing Duplicates." You can see in these sheets that the AMMP! status and test scores are the same for each pair of duplicates. If this were an actual evaluation, you would double-check the original dataset to make sure the test scores were not entered incorrectly. In this example, though, you can safely say that these students have been entered into the dataset twice, so the duplicates should be deleted.

Next, check whether the dataset contains any missing values. In cell D2 of the "Removing Duplicates" sheet, type the formula "=ISBLANK(C2)" and then press Enter. The result will be "TRUE" if a value is missing from column C or "FALSE" if no value is missing. Apply the formula to the entire column by placing your cursor over the bottom right corner of cell D2 so that the cursor turns into a plus sign. Then click and drag the plus sign down to cell D101. This copies the formula to all cells below C2. Once you have done this, the "Removing Duplicates" sheet should look like the next sheet, titled "Examining Missing Values." You can see in these sheets that students 10 and 99 have missing values. If this were an actual evaluation, you would double-check the original dataset to see if there was an entry error. If there was an entry error, you would reenter the data correctly. For this example, let's say you were able to locate test score data for these students and entered a score of 88 for student 10 and a 78 for student 99.

In this simple example, you can easily find the missing values by scanning the data yourself in Microsoft Excel. However, when you work with large datasets with multiple variables, using more sophisticated software tools such as SAS, Stata, SPSS, or R can improve efficiency and accuracy of identifying and correcting missing values.

Next, look for possible data entry errors. One way is to check for values in the dataset that are outside the theoretical range of possibilities. You might identify entry errors by finding the minimum and maximum in the dataset. Let's start by looking at the minimum possible test score. Because the lowest possible test score is 0, any negative test score must be an entry error.

Go to the "Cleaning Entry Errors" sheet. To find the minimum test score in the dataset, type the formula "=MIN(C2:C101)" in cell D2. The result shows negative 79 in row 61. If this were an actual evaluation, you would return to the original dataset to find the source of entry error. In this example, let's say that the negative was an error and the student's test score was actually 79. Let's make that change. Now, find the maximum value by typing the function "=MAX(C2:C101)" in cell E2. The maximum is 186 in row 83, which is more than the highest possible score of 100. Let's say you went back to the original tests and found that this student actually scored an 86. So you make that change. You can continue this process until the minimum and maximum values are no longer outside the range of possible values.

Alternatively, you can use the "Highlight Cells Rules" under "Conditional Formatting" on the Home ribbon to highlight all cells with values above 100 and all cells with values below 0. This alternative might be preferable if a dataset contains many such values.

If it is impossible to determine the correct value of a data entry error, then simply delete the incorrect value from the cell, but do not remove the row from the Excel sheet. By deleting only the incorrect value, the cell will be empty, showing that this data value is missing.

Now that the data have been cleaned, you can begin to answer the AMMP! evaluation question by computing descriptive statistics. You can use measures of central tendency and variation to describe the data. To understand the central tendency of a variable, calculate its mean, median, and mode. To understand the variation of a variable, calculate its standard deviation, minimum, maximum, and interquartile range.

Because you want to compare AMMP! students and non-AMMP! students, first sort the data. In the "Measures of Central Tendency" sheet in the *Descriptive Statistics Activity* workbook, highlight all the data. Click the "Sort & Filter" option on the Home ribbon, and then select "Sort Smallest to Largest." This will put all non-AMMP! students in rows 2 to 51 and all AMMP! students in rows 52 to 101. In this sheet, non-AMMP! students are denoted by a 0 in the ammp_status column, and AMMP! students are denoted by a 1 in the column.

Now that the data are sorted, you can find the mean, or average, test score for non-AMMP! students. In cell D2, type the formula "=AVERAGE(C2:C51)" and press Enter. Then, find the mean test score for AMMP! students by typing the formula "=AVERAGE(C52:C101)" in cell E2 and then press Enter. The mean test scores are 76.4 for non-AMMP! students and 83.6 for AMMP! students. Comparing the two means, you can see that the average test score for AMMP! students was 7.2 points higher than the average test score for non-AMMP! students.

Now look at the median test scores, or the middle scores. For non-AMMP! students, type the formula "=MEDIAN(C2:C51)" in cell D3 and press Enter. Then, for AMMP! students, type "=MEDIAN(C52:C101)" in cell E3 and press Enter. The median test scores are 76.5 for non-AMMP! students and 84.0 for AMMP! students.

If you compare the mean test scores to the median test scores for each group, you find that they are quite similar. This shows that the average scores are roughly the same as the middle scores. This comparison suggests that relatively few AMMP! or non-AMMP! students scored very high on high school math placement tests, as many high scores would pull the mean well above the median. At the same time, the comparison suggests that relatively few AMMP! or non-AMMP! students scored very low on the tests, which would pull the mean well below the median.

Finally, find the mode, or the most common test score, for non-AMMP! and AMMP! students. Type the formula "=MODE(C2:C51)" in cell D4 for non-AMMP! students and press Enter. Then, type "=MODE(C52:C101)" in cell E4 for AMMP! students and press Enter. The mode for non-AMMP! students is 76.0, and the mode for AMMP! students is 86.0.

This calculation tells you that the most common test score for non-AMMP students is very close to both the average test score and the middle test score. For AMMP! students, the mode of 86.0 is slightly higher than the mean and median test scores. These values give you a sense of the range in students' ability between the AMMP! and have non-AMMP! students.

Measures of variation tell you about the spread of a distribution. One measure is standard deviation, which indexes how far away from the mean a randomly selected observation is likely to be. Open the "Measures of Variation" sheet in the *Descriptive Statistics Activity* workbook to find the standard deviations for the AMMP! example in this chapter.

First, calculate the standard deviation separately for non-AMMP! and AMMP! students, as you did earlier for the measures of central tendency: mean, median, and mode. For non-AMMP! students, type the formula "=STDEV(C2:C51)" in cell D2 and press Enter. For AMMP! students, type "=STDEV(C52:C101)" in cell E2 and press Enter. The standard deviation for non-AMMP! students' test scores is 8.7, and the standard deviation for AMMP! students' test scores is 6.2. This shows that non-AMMP! students have a wider spread, or more variability, in their test scores than AMMP! students do.

Next, look at the range of scores for non-AMMP! and AMMP! students by determining the maximum and minimum values. You can use the same formulas you used for cleaning data. To find the minimum test score for non-AMMP! students, type the formula "=MIN(C2:C51)" in cell D3 and press Enter. To find the maximum test score for non-AMMP! students, type the formula "=MAX(C2:C51)" in cell D4 and press Enter. Then, perform the same calculations for AMMP! students. In cell E3, type "=MIN(C52:C101)" and press Enter. In cell E4, type "=MAX(C52:C101)" and press Enter.

These calculations show that the minimum test score for non-AMMP! students is 51.0 and the maximum score is 90.0. The range is therefore 39 points. For AMMP! students, the minimum test score is 72.0 and the maximum score is 96.0, which is a range of 24 points. Thus, you find that non-AMMP students have both a higher standard deviation, or greater variability, and a greater range in their test scores than AMMP! students do.

Another useful measure of variation is interquartile range, which tells you the range between students scoring at the 75th and 25th percentiles. Interquartile range shows the spread of scores for students in the middle of the distribution and is therefore not influenced by outliers.

First, calculate the interquartile range for non-AMMP! students by subtracting the 75th percentile from the 25th percentile. Type the formula "=QUARTILE(C2:C51,3)-QUARTILE(C2:C51,1)" into cell D5 and press Enter. In this formula, 3 refers to the third quartile, or the 75th percentile, and 1 refers to the first quartile, or the 25th percentile. Next, to calculate the interquartile range for AMMP! students, type the formula "=QUARTILE(C52:C101,3)-QUARTILE(C52:C101,1)" into cell E5 and press Enter.

You can see that non-AMMP! students have a greater interquartile range than AMMP! students do. This suggests that the higher standard deviation in test scores for non-AMMP! students is not merely due to outliers because this group also has a greater range of middle values than AMMP! students do.

Most of the descriptive statistics in this chapter can be computed simultaneously by using the Program Evaluation Toolkit Calculator, found on the resources page of the website. To practice open the *Descriptive Statistics Activity* workbook in Excel. To enable the Program Evaluation

Toolkit Calculator add-in, follow the directions in the *Program Evaluation Toolkit Calculator: User's Guide*, available on the resources page of the website.

Once the Program Evaluation Toolkit Calculator is enabled in Excel, click the "Descriptive Statistics" button on the "Program Evaluation Toolkit Calculator" ribbon. Then, uncheck all boxes except for those next to "test_score" and "ammp_status," and click "OK." The resulting values will look like the screenshot on this slide.

Using descriptive statistics is an important first step in answering the evaluation question "How do AMMP! participants' scores on high school math placement tests compare to nonparticipants' scores?" However, if you want to more deeply understand relationships between variables of interest—including causal relationships—you can apply inferential statistics. Regression analysis is one of the most common methods in inferential statistics. Again, imagine you are a member of the AMMP! evaluation team and want to further address the evaluation question by using inferential statistics.

To get started, download the *Inferential Statistics Activity* workbook, available on the resources page of the website. This Excel workbook includes all the clean variables from the *Descriptive Statistics Activity* workbook as well as four additional variables: homework completion percentage, eligibility for the National School Lunch Program, sex, and race/ethnicity. Assume that these additional variables have already been cleaned.

Begin by identifying variables. The first sheet in the *Inferential Statistics Activity* workbook is a codebook that describes all variables included. High school math placement test score is the dependent variable, and AMMP! participation is the main independent variable of interest (that is, the main objective is to understand the causal effect of AMMP! participation on math placement test scores). Possible covariates include homework completion percentage, eligibility for the National School Lunch Program, sex, and race/ethnicity.

Confounds are not in the dataset but might also impact math achievement. For example, outside math tutoring could influence conclusions. AMMP! students might also have access to other community or home resources that could impact their math achievement, and the effects could be attributed to AMMP!. Conversely, there may be confounding variables that hide any positive effect of AMMP!. For example, non-AMMP! students might have more access to outside opportunities that improve math achievement.

One important point: When adjusting for potential confounds using multiple regression analysis, only adjust for variables that were measured before student exposure to the program. Otherwise, the regression may be adjusting away some of the impact of the program that you are trying to measure. In the AMMP! example, homework completion percentage is measured after students start to participate in AMMP!. Therefore, you should not include this variable in the regression analysis. On the other hand, if homework completion percentage were measured in the year before AMMP! began, it would be fine to use that variable in the regression equation.

First, let's focus on simple linear regression analysis. If you have not done so already, open the *Inferential Statistics Activity* workbook and double-click the Program Evaluation Toolkit

Calculator add-in to add it to your Excel menu bar. Make sure you are in the "Data" sheet of the workbook.

Simple linear regression analysis compares a single independent variable with a single dependent variable—for example, participation in AMMP! and achievement in math. For all the analyses in this chapter, you will use "test_score" as the dependent variable. For this simple linear regression analysis, use "ammp_status" as the only independent variable.

To run this analysis, click the "Linear Regression" button on the Program Evaluation Toolkit Calculator ribbon in Excel. Next, select "test_score" as the dependent variable. Then select "ammp_status" and add it to the independent variables box. Click "OK." You will see a new sheet in your workbook with the title "Model 1."

Let's focus on a few aspects of the output in the "Model 1" sheet. This sheet provides additional output if you want to perform more advanced analyses. There is a table with the row headings "Intercept" and "ammp_status." Look at the "Coefficients" column. The intercept coefficient is the predicted value of the dependent variable (or the average test score in this case) when the independent variables all have a value of 0. The coefficients for the independent variables indicate the predicted change in the dependent variable when an independent variable increases by one unit.

This linear regression analysis has only one independent variable, "ammp_status," which equals 0 for non-AMMP students and 1 for AMMP students. So, for this example, the intercept coefficient represents the average test score for non-AMMP! students (76.44), and the "ammp_status" coefficient represents the difference (7.18) between the average test score for non-AMMP! students and the average test score for AMMP! students.

For information on additional outputs of the Program Evaluation Toolkit Calculator, see the *Program Evaluation Toolkit Calculator: User's Guide*, available on the resource page of the website.

The previous simple comparison of mean differences is not a very credible estimate of the average change in student test scores due to AMMP! participation. This is because simple linear regression analysis does not account for the effects of other independent variables that have a relationship to both the dependent and independent variables. For instance, a *covariate* is an independent variable that has a relationship to the dependent variable that should be considered but that is not directly related to the program, such as student race/ethnicity. A *confound* is an independent variable that could result in misleading interpretations of a relationship between the independent and dependent variable. In the AMMP! example in chapter 1, students exposed to an additional supplemental math curriculum might see improvements in achievement, so the evaluation team must consider that this confound may actually lead to any changes in student performance.

Multiple regression analysis accounts for this by including the single independent variable AMMP! status and potential covariates, including eligibility for the National School Lunch Program, sex, and race/ethnicity. Remember that homework completion percentage is measured after students start to participate in AMMP!. Therefore, do not include this variable in the

regression analysis. On the other hand, if homework completion percentage were measured in the year before AMMP! began, it would be fine to use that variable in the regression equation. So, in the AMMP! example, you can add "school_lunch_status," "sex," and "race/ethnicity" to the regression model for the multiple regression analysis. Because there is usually very little downside to including additional variables in a regression model, it's usually best practice to include all potential covariates in the model.

To generate an accurate regression model, first transform the "sex" and "race/ethnicity" variables. You can do this by clicking the "Dummy Variables" button on the Program Evaluation Toolkit Calculator ribbon. First, select "sex" and click "Create Dummy Variable." Then, select "race/ethnicity" and click "Create Dummy Variable" again. Finally, click "OK."

For each unique value of the variable in question (for example, the variable "race/ethnicity" and the value "Asian"), this procedure creates a new column in the dataset that contains a 1 if the variable equals the unique value and a 0 if the variable does not. For example, because the "race/ethnicity" variable has seven categories, the "Dummy Variables" procedure creates seven new columns in the dataset. The "race/ethnicity.Eq. American Indian/Alaskan Native" variable is 1 if a student's race/ethnicity is "American Indian/Alaskan Native" or 0 otherwise.

Include these dummy variables as independent variables in the regression model instead of the original variables with alphabetic values. However, for each of the original variables, you need to omit one of the dummy variables, which then serves as the reference category. For example, you might use "female" as the reference category for the variable "sex" and "Black" as the reference category for the variable "race/ethnicity."

Other variables can also be transformed into dummy variables based on meaningful criteria. For example, you might transform raw values for the "test_score" variable into a pass or fail value where pass is equal to 1 and fail is equal to 0. You would do this before loading your data into the Program Evaluation Toolkit Calculator by using the SUBSTITUTE function in Excel. Directions on how to use this function are in the handout *Microsoft Excel Functions for Data Cleaning*, available on the resources page of the website.

The results of the multiple regression analysis allow you to estimate the difference in average test scores between AMMP! and non-AMMP! students while controlling for the potential effects of the eligibility for the National School Lunch Program, sex, and race/ethnicity covariates. As it turns out, the estimate did not change much from the linear regression analysis, from 7.18 to 7.30. This small difference suggests that the three variables had minimal effects.

However, there is still the possibility of unmeasured covariates (such as race/ethnicity) and confounds (such as other implemented programs) that could undermine your conclusions. Randomized controlled trials, or RCTs, help with the problem of unmeasured covariates or confounds, which is why results from RCTs are considered the highest tier of evidence when determining whether interventions are evidence-based. Without randomization, AMMP! and non-AMMP! students might differ in motivation or math aptitude. Using random assignment, however, would greatly reduce the chance that the two groups of students differ in those

characteristics and would likely lead to a more accurate measurement of the effects of the program.

Next, let's consider a qualitative example, using the AMMP! evaluation question "What barriers exist that prevent students from completing homework?"

This question likely involves complex, subjective judgments for which complete quantitative data do not exist or cannot be easily collected. It is more appropriate to address the question with qualitative data.

One way to collect qualitative data to answer the evaluation question is conducting interviews. In this activity, imagine you are a member of the AMMP! evaluation team and want to collect data from both students and teachers. You create one interview question for students and another for teachers.

The question for students is "What are some things that might keep you from completing homework?"

The question for teachers is "What do you think are barriers that prevent students from completing homework?"

You can use interview transcripts to find themes that arise during the interviews. Open the handout *Qualitative Analysis Activity*, available on the resources page of the website, to identify some common themes in the mock responses to the interview questions. Pause the video now and complete this activity.

What common themes did you identify in the *Qualitative Analysis Activity*? The example in the handout highlighted a lack of time, and perhaps you also noted that both teachers and students mentioned unsupervised time after school, the difficulty of homework, and students not seeing the value of homework. These all might be barriers to completing homework. Although qualitative analysis is often more complex, this activity may give you a sense of how you can use qualitative analysis to answer certain evaluation questions more completely than you could by using quantitative analysis.

If you need additional support in conducting an RCT, the tool *Statistical Theory for the* RCT-YES *Software: Design-Based Causal Inference for RCTs*, available on the resources page of the website. For example, the tool provides guidance on examining clustering and covariances between subgroups, which might be valuable if you plan to conduct an evaluation that is implemented and analyzed at the district, school, or classroom level.

Another helpful resource is the free online tool Evidence to Insights (e2i) Coach, also available on the resources page of the website. This tool can help you answer more complex questions related to your program and can support you in assessing the impacts of programs and practices using more advanced procedures and methods.

This concludes chapter 2. In the next chapter, you will consider how to move from the results of data analysis to interpretation and recommendations.

This handout was prepared under Contract ED-IES-17-C-0005 by Regional Educational Laboratory Central, administered by Marzano Research. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.