

Module #5. Quantitative Data Analysis

At this point, you will likely have collected quantitative data through administering surveys, coding qualitative data, or gathering available data, such as district-level datasets. You are now ready to undertake quantitative data analysis. This process will help you to identify trends, relationships, and anomalies within your dataset. In keeping with Carnegie's design principles for continuous improvement of your school model, your findings will inform your design.

In this module, we review the skills and steps involved in quantitative data analysis. We assume the reader has a basic familiarity with Microsoft Excel (2007 or later; images and instructions are for Excel for PC), including how to use basic formulas. The example we use in this section is grounded in the Carnegie design principle of positive youth development. We focus specifically on caring, consistent student-adult relationships that communicate high expectations for student learning and behavior. For the purposes of this section, we draw from a follow-up survey that focuses on characteristics students say they value in their counselors. The sequence of sections in this module appears below:

1. Gather and merge available data
2. Prepare survey data for analysis
3. Prepare qualitative data for analysis
4. Get to know your data and sub-groups through descriptive statistics
5. Examine relationships through correlations
6. Explore more complex data analysis options

1. GATHER AND MERGE AVAILABLE DATA

KEY INFORMATION

Using available demographic, assessment, and behavioral data. Every school and district collects some demographic, assessment, and behavioral information about their students. Oftentimes, with some thoughtful planning and perseverance, you can acquire that information.

Gathering available data

There are generally two types of available student data:

- Demographic data: This data typically consists of relatively static information about students, such as gender, race/ethnicity, whether the student has an IEP, whether the student is an English Language Learner, whether the student receives free or reduced-price lunch, and home address information.
- Assessment/behavioral data: Assessment and behavioral data is collected as students progress through their academic careers, and typically include scores on standardized tests, attendance records, grades, course passing data, and discipline records.

If it makes sense, simply ask students for this information. In some cases, it is much easier to simply ask students for this information, rather than attempt to gather it from an outside source. If the question is fairly straightforward and students' self-report can be trusted, this route will streamline the process of preparing your data set for analysis. Students, for example, can generally reliably report on their gender, race/ethnicity, and home language.

If directly asking is not feasible, develop a plan to gather data. In other cases, students may not be able to provide a reliable self-report. For instance, if a student is young, she may not know whether she receives free lunch, and almost no student could tell you their exact score on a fourth-grade state ELA exam. In addition, students may feel compelled to lie about questions regarding topics like IEP or free lunch status. In these instances, consider gathering the information from the district or school.

Different states, districts, and schools have different available data. The path to obtaining this data will vary by school and district, and is considered further in the worksheet in the tools appendix. If you do not already know who in your district or a particular school can access the right data, find someone who deals with data regularly at the high school level and talk with them. Be sure that the data you want exists before you count on it.

In New York City, where Eskolta has done most of its work, the people who know the most about available data are typically: guidance counselors, who are responsible for student intake and programming; school administrative support staff, who are responsible for gathering and generating reports; and, oftentimes, math or science teachers who formally or informally take on responsibilities related to internal data analysis.

Ensure that any data you collect comes with a student number you can match up with your own data. Do not rely on student names: simple spelling discrepancies can throw off the matching process, and you will spend a great deal of time “cleaning” your data. If you have a large data set, this task quickly becomes time-consuming. (Note that if you plan to combine survey data with district data, this means you must ask students or adults to provide student ID numbers.)

Merging Available Data

Before you pull any data into Excel, begin by creating a tab and naming it “Notes.” In this tab, write down where data in other tabs comes from and note where there are cells that you will have to update in the future. This becomes your cheat sheet when you forget (and you will forget!) what steps were involved in putting your data together.

Use a VLOOKUP function to merge data. If you gather data from an outside source, request it in a CSV (comma-separated value) or tab-delimited format, as these are practically universal in their availability and can be easily imported into Microsoft Excel. You will likely receive a number of different files with different pieces of information. For example, you might receive one spreadsheet with attendance data, one with standardized test scores, and another with grades. In addition to merging all of these into one data set, you will then need to merge that data with your primary data set (e.g., student responses to a survey). You can merge all these data sources by hand, but it is not recommended—there are too many opportunities for error. The Excel appendix contains details about using the VLOOKUP function to merge data, as does this page on the Microsoft web site: <http://office.microsoft.com/en-us/excel-help/vlookup-HP005209335.aspx>

Set up a separate tab in Excel for each piece of “raw data” you are importing and give the tab a name that makes it clear where the data came from. Specifically, you should plan to have various tabs that hold your “raw data” (i.e., original survey responses and district data). Maintain a direct correspondence between tabs and data files—one tab for each data file. Do not alter these tabs! It is very common to have to revert to raw data when your data manipulation efforts go awry or when you update data. Annotate in your Notes tab where you got the data in each of your Raw Data tabs.

Then, create a tab called “Master Data.” This tab is where you pull data from every other tab into one place. The first column of this tab should always be student ID (assuming you are using student-level data). Other columns may simply draw directly from a column in another tab (using the VLOOKUP function to find that data based on student

ID) or may use a calculation to modify data drawn from other tabs. Never draw from separate files, as this inevitably causes problems in updating.

2. PREPARE SURVEY DATA FOR ANALYSIS

KEY INFORMATION AND APPLICATION TO EXAMPLE

Entering data

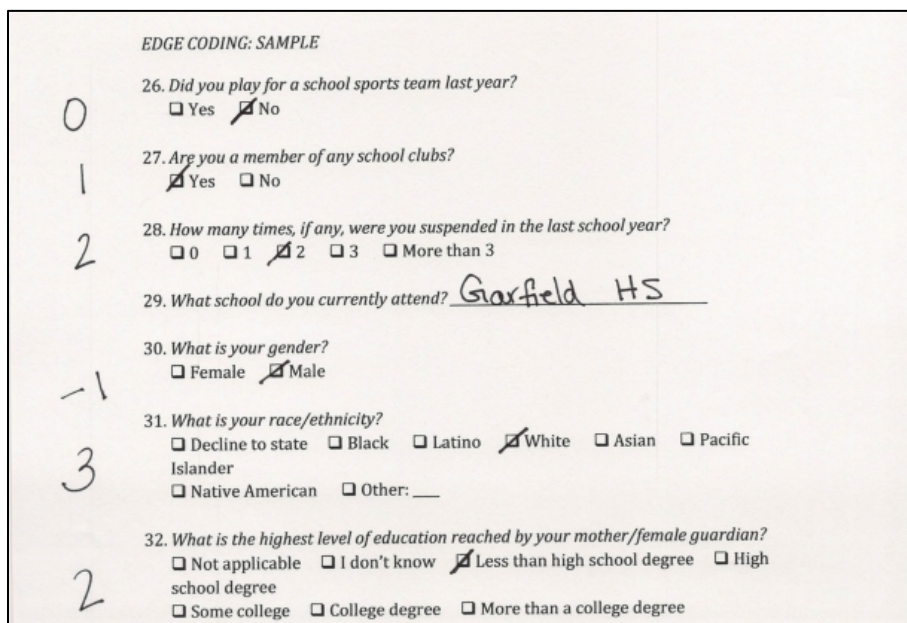
Manual and automatic data entry. When you administer a survey online, you can download data into Excel. If administering a survey online is not feasible, you will need to manually enter responses into a spreadsheet.

Survey administration

We recommend using SurveyMonkey, SurveyGizmo, or Google Forms to administer your survey. Hand-entering survey responses is both boring and rife with opportunities for error. It is also typically unnecessary: increasingly students have Internet access at school, and some students have Internet access at home.

Hand-entering data. Data entry is inherently boring, and therefore always involves a great deal of error. There are two strategies to reduce this error and ensure that you have the best chance of discovering interesting elements of your data.

- 1) Edge coding. We suggest using this strategy, regardless of whether you also undertake step 2 (double-entering). Edge coding involves writing the numerical code for the response to each item on the edge of the paper to reduce the need for your eye to travel as you enter the data. An example appears below:



EDGE CODING: SAMPLE

0 26. Did you play for a school sports team last year?
 Yes No

1 27. Are you a member of any school clubs?
 Yes No

2 28. How many times, if any, were you suspended in the last school year?
 0 1 2 3 More than 3

29. What school do you currently attend? Garfield HS

-1 30. What is your gender?
 Female Male

3 31. What is your race/ethnicity?
 Decline to state Black Latino White Asian Pacific Islander
 Native American Other: ___

2 32. What is the highest level of education reached by your mother/female guardian?
 Not applicable I don't know Less than high school degree High school degree
 Some college College degree More than a college degree

We encourage that you do this regardless of how many responses you have, but it becomes particularly important when you have 100 or more responses to enter.

- 2) Double entering.** If at all feasible, you should have two people independently enter the data. This strategy allows you to run a comparison on the two columns to chase down and fix any errors. For instance, if one person entered results in the tab MikeSurveys and another person entered them in the tab JessSurveys, then you can check that they match by entering into any given cell the formula: =MikeSurveys!A1=JessSurveys!A1. This will yield a TRUE if the entries match and a FALSE if they don't.

Preparing data for analysis

Once data has been entered, prepare it for analysis by creating dedicated tabs for different types of data. If you need to do further manipulation of your data, remember not to ever do it directly in the Raw Data tabs. Instead, create three more tabs:

Lists: You should create one tab that holds all the standard lists you will be drawing upon in your data. For example, if you have a codebook that translates survey choices into numbers (for example, a five point Likert scale from “Strongly disagree” to “Strongly agree” translated into 1-5), put that codebook in your Lists tab. If you have a sub-group based on a range of attendance data (as described in step 5), name those ranges in your Lists tab. Identify the key lists in your Codebook tab. The codebook section below contains more detail on this important tab.

Modified data tabs: These are the tabs where you take your raw data and modify it, such as converting survey responses to numbers using your codebook or a VLOOKUP table. See the below section on creating a codebook for more information.

Pivot Tabs: Sometimes you will have raw data with multiple rows for each student. In these instances, you will need to consolidate the data using PivotTables in Excel. Create a tab in which you place such pivot tables so that you can easily find them, and take careful notes on which need to be updated and when in your Notes tab. For instructions on creating PivotTables, refer to the Microsoft Office web site: <http://office.microsoft.com/en-us/excel-help/pivottable-reports-101-HA001034632.aspx>

Creating a codebook

As described in the section above, when entering survey data, a critical step is to create a codebook so you know what numbers to assign to each survey responses, and can then recode your data.

A codebook is the key to translating answer choices (such as “Strongly Disagree”) to numbers (such as “1”). When preparing a codebook, it is a good idea to use a simple convention for assigning numbers to codes, such as greater agreement yields higher numbers. A sample codebook appears below, using survey items from Module #3, Surveys:

Variable	Item	0	1	2	3	4	5
Same Neighborhood	Someone who is from my neighborhood		Not at all interested	Not very interested	Somewhat interested	Interested	Very interested

SameRace	Someone who is the same race/ethnicity as me.		Not at all interested	Not very interested	Somewhat interested	Interested	Very interested
----------	---	--	-----------------------	---------------------	---------------------	------------	-----------------

Recode your data

Once you have created a codebook, you can then recode your data to reflect the values ascribed to each response in the codebook. To recode your data, follow the steps below:

1. Place your codebook in your Lists tab. Much like the one above, the codebook should provide variable names in the rows and the codes you have used for your data (typically numbers from zero up) in the columns, and the translations of those values in the cells. (See the example above.)
2. Using the Define Name option in Excel, name the codebook something simple (like Codebook). Highlight the array starting in the top left corner and going down to the bottom right. It is a good idea to extend beyond the bottom of your codebook in case it gets larger.
3. Pick a variable that you will be recoding and count the number of the row of that column in your codebook. For instance, if we wanted to recode the SameRace variable, our row number is 3. Now pick a column in your Modified Data tab where your recoded data will appear.
4. Use the HLOOKUP function to pull the appropriate recoding from your codebook into your new column. For example: =HLOOKUP(SurveyData!B3, codebook, 3, false) would provide the translation for the value in B3 for the SameRace variable. More information about the HLOOKUP function is available on the Microsoft web site: <http://office.microsoft.com/en-us/excel-help/hlookup-HP005209114.aspx>

3. PREPARE QUALITATIVE DATA FOR ANALYSIS

KEY INFORMATION

Attach numbers to qualitative data. As discussed in Module #4, Qualitative Data Analysis, when you code using a rigorous, established coding scheme, you can be confident enough in your data to run quantitative analyses on them. To prepare for the quantitative data analysis process, create a row for each student and a column for each code as well as a column for each characteristic you will use to demarcate subgroups, such as gender, grade level, etc.

What numbers should you place in the code columns? Typically, you include a frequency or a proportion for each participant for each code:

- Frequencies: Frequencies can be either a count for each person for each code (for example, the exact number of times each respondent mentioned peers or a rating of the intensity with which they mentioned peers), or a 0/1 (for example, did the respondent mention peers at all?).
- Proportions: When the number of codes applied is very different across transcripts/artifacts it is best to use proportions rather than raw counts. For example, if one person wrote a 500-word response to an open-ended

question, resulting in 50 codes, and another person wrote a 50-word response that resulted in 5 codes, you might find it more accurate to say that the coding appeared as a 10% proportion of writing for both, rather than 50 for one and 5 for the other.

APPLICATION TO EXAMPLE

To prepare for the quantitative data analysis process of the interviews they conducted, the Lucretia Mott design team creates a row for each student and column for each code. They tally the frequencies of occurrence of each code for each respondent. They also include two columns for sub-groups they expect to be relevant to our analysis: the student’s counselor and the student’s gender:

	Counselor	Gender	Parent relations	Sibling relations	Romantic relations	Peer relations	Improving academics	Specific assignments	Applying to college	Importance of college
Alan	Wendy	M	3	0	0	2	0	2	0	0
Betsy	Wendy	F	2	0	3	0	2	1	0	1
Carlos	Joe	M	1	5	2	1	5	4	0	1
Davina	Joe	F	0	1	2	3	0	0	2	0
Edwina	Joe	F	1	0	3	2	0	0	0	5

4. GET TO KNOW YOUR DATA THROUGH DESCRIPTIVE STATISTICS, INCLUDING SUB-GROUP ANALYSIS

KEY INFORMATION

Identifying overall trends. When you have a set of data, often the most obvious thing you want to know is: What things did the most people say? This is a good first step of analysis and can provide a valuable sense of overall trends. In addition, as you dig into data, you may want to know how two or more groups differ from each other in terms of what they said or in other characteristics. This may sound like a small discovery, but it will inform much of your analyses. For instance, you may wonder how boys’ and girls’ attitudes vary with respect to their counselors; to answer this question, you can compare data by gender.

Start with descriptive statistics. To undertake this initial analysis, you will use descriptive statistics to learn more about the frequency of different responses in your data sets, and how these responses vary across groups. In the “Application to Example” section that follows, we focus on descriptive statistics you can generate in Microsoft Excel.

Think strategically about sub-groups. Examining differences across sub-groups can be highly revealing in the design process. For example, we might ask: How does male performance compare to female performance on a given assessment? How do ELL students perform compared to non-ELL students? These subgroups are based on easy-to-create groupings as long as the data already exists.

However, other subgroups can often provide more actionable insight, and are easily created so long as you ask the right questions of your group of interest. For example, these sub-groups might be defined by questions like: How do students who worked with one counselor compare to students who worked with another? How do students who used a specific blended learning platform differ from those who did not? How do students with high, middle, and low attendance respond differently to their counselor relationships? How do students with a history of high, middle, and low grades work differently with computers?

Consider which groups might provide actionable insight and plan to look at a few sub-group differences in addition to your whole sample. For these sub-groups, you need to have somewhere between three and five categories, define those categories, and then convert your data into sub-groups using the VLOOKUP function in Excel (see the Excel appendix for more information about how to use VLOOKUP). If those categories are discrete—for instance, five different counselors—then simply create a list in your Lists tab and look up values from there much as you would from a codebook: =VLOOKUP(SurveyData!A1, CounselorNames, 2, false). If those categories are continuous—for instance, attendance—then decide where your natural cut points are and create a list with those cut points in your Lists tab and name it appropriately. For example, the table below could be named attendanceRates:

YTD Attendance	Category Name
0%	Very Low Attendance
60%	Low Attendance
85%	Moderate Attendance
92%	Strong Attendance
100.1%	Error

With a list like the one above, you can instead use the following function to turn any level of attendance into one of 5 codes: =VLOOKUP(Attendance!A1, attendanceRates, 2, true). In this example, any student with attendance below 60% will be labeled “Very Low Attendance” while students with attendance above 92% will be labeled “Strong Attendance.”

Analyze your data in steps. After thinking strategically about sub-groups, examine your data in steps:

- 1) Run frequencies on your data for the whole group
- 2) Run the same frequencies for your subgroups of interest
- 3) Look at the means for each item for your whole group and compare to subgroups

As you analyze your data, think of the story you would tell to explain your findings. What are possible reasons that different subgroups have different values? What questions do these possible stories raise for you? What are the explanations that you do not immediately arrive at but that could be reflected in the data?

APPLICATION TO EXAMPLE

Earlier, after conducting their first survey, the Lucretia Mott design team had found that students overwhelmingly preferred having a counselor of the same gender. With this result in mind, they decided to conduct a follow-up survey to investigate the impact of a counselor's gender, as well as other factors, on students' relationships with adults. In their survey, they ask a variety of questions about students' beliefs about their relationships with their counselor, along with questions to help categorize the data in order to allow them to see how these responses compare with each other. Ultimately, they are interested in investigating whether beliefs vary across students of different genders. The sample survey appears below:

	<p>1. First name: _____ 2. Last name: _____ 3. ID number: _____</p>
Categorizing variables	<p>4. What is the gender of your current counselor? <input type="checkbox"/> Male <input type="checkbox"/> Female</p>
	<p>5. What is the race of your current counselor? <input type="checkbox"/> White <input type="checkbox"/> Black <input type="checkbox"/> Latino <input type="checkbox"/> Asian/Pacific Islander <input type="checkbox"/> Native American</p>
Belief variables	<p>6. If you could choose your own counselor, how important would each of the following be in your decision?</p>
	<p>A. Someone who is from my neighborhood <input type="checkbox"/> Very important <input type="checkbox"/> Important <input type="checkbox"/> Somewhat important <input type="checkbox"/> Not very important <input type="checkbox"/> Not important at all</p>
	<p>B. Someone who is the same race/ethnicity as me <input type="checkbox"/> Very important <input type="checkbox"/> Important <input type="checkbox"/> Somewhat important <input type="checkbox"/> Not very important <input type="checkbox"/> Not important at all</p>
	<p>C. Someone who has the same gender as me <input type="checkbox"/> Very important <input type="checkbox"/> Important <input type="checkbox"/> Somewhat important <input type="checkbox"/> Not very important <input type="checkbox"/> Not important at all</p>
	<p>D. Someone who will talk to me about my home life <input type="checkbox"/> Very important <input type="checkbox"/> Important <input type="checkbox"/> Somewhat important <input type="checkbox"/> Not very important <input type="checkbox"/> Not important at all</p>
	<p>E. Someone who will talk to me about my future career <input type="checkbox"/> Very important <input type="checkbox"/> Important <input type="checkbox"/> Somewhat important <input type="checkbox"/> Not very important <input type="checkbox"/> Not important at all</p>
	<p>7. If you could choose your own counselor, rank the following from 1-5 in order of importance for your decision. "1" represents the most important, and "5" represents the least important. ___ Someone who is from my neighborhood ___ Someone who is the same race/ethnicity as me</p>

Someone who has the same gender as me
 Someone who will talk to me about my home life
 Someone who will talk to me about my future career

8. How important are each of the following adults in terms of supporting your success at school?

A. Teachers
 Very important Important Somewhat important Not very important Not important at all

B. Mother, father, or adults you live with
 Very important Important Somewhat important Not very important Not important at all

C. Guidance counselors
 Very important Important Somewhat important Not very important Not important at all

Categorizing variables {

9. What is your gender?
 Male Female

10. What is your race?
 White Black Latino Asian/Pacific Islander Native American

Check frequencies

The team begins by checking the frequency of each response. Consider Question 7 on the survey, which asks students to rank the importance of the various characteristics a guidance counselor could have. The table below shows a sample of how two students ranked the items from 1 (“most important”) to 5 (“least important”).

StudentID	SameHood	SameRace	SameGender	HomeLife	Career
100	1	2	3	4	5
101	4	3	2	5	1

As a starting point, the team wants to get a sense of which responses were most often highly ranked. To do this, they use the COUNTIF function in Excel. Using the COUNTIF function requires defining the following arguments:

- RANGE: The range of cells in which Excel should look for a given response (in this case, the cells contained in each column).
- CRITERIA: The criteria Excel uses to “count” a given cell (in this case, whether the cell value is equal to 1 (or 2, or 3, etc.)).

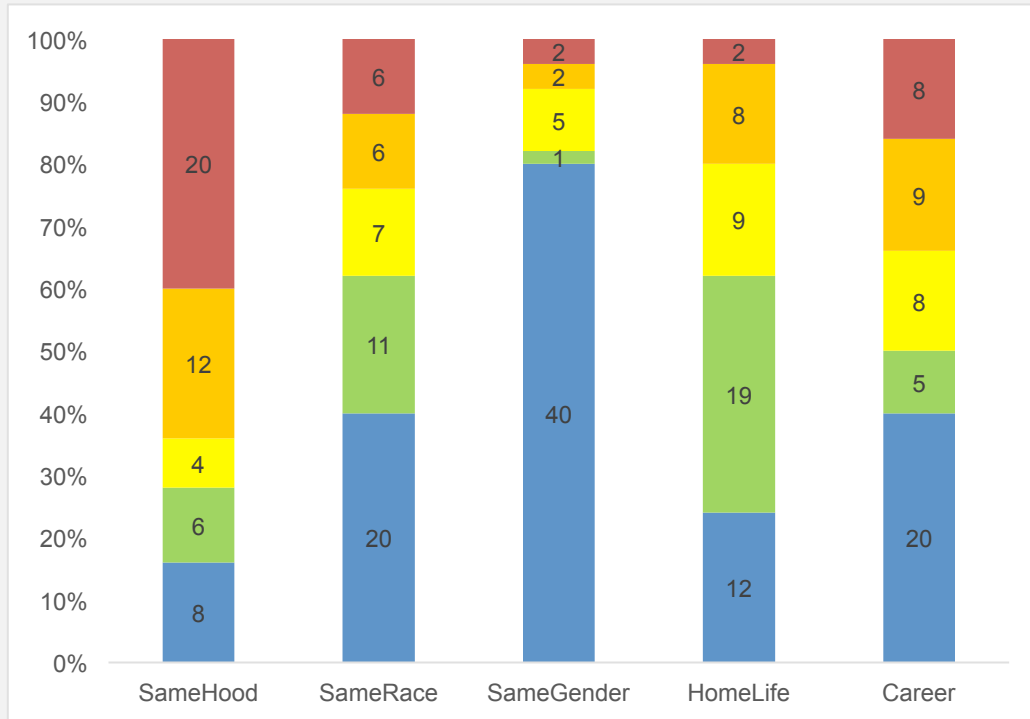
A sample of how to use the COUNTIF function appears below:

	A	B	C
1	Personcode	SameHood	SameRace
2	100	1	2
3	101	4	3
200	1	=countif(b\$2:b\$101, \$a200)	=countif(c\$2:c\$101, \$a200)
201	2	=countif(b\$2:b\$101, \$a201)	=countif(c\$2:c\$101, \$a201)
205	5	=countif(b\$2:b\$101, \$a205)	=countif(c\$2:c\$101, \$a205)

This analysis yields the following frequencies for each response:

Personcode	SameHood	SameRace	SameGender	HomeLife	Career
100	1	2	3	4	5
101	4	3	2	5	1
1	8	20	40	12	20
2	6	11	1	19	5
3	4	7	5	9	8
4	12	6	2	8	9
5	20	6	2	2	8

With these figures calculated, the team uses the Excel capability to quickly create a stacked bar chart to see which items had the highest ranks. By coloring the bars following the colors of the rainbow with colors like red and orange for the lowest ranks and green blue for the highest, it is easy for the eye to immediately see what stands out as highly ranked.

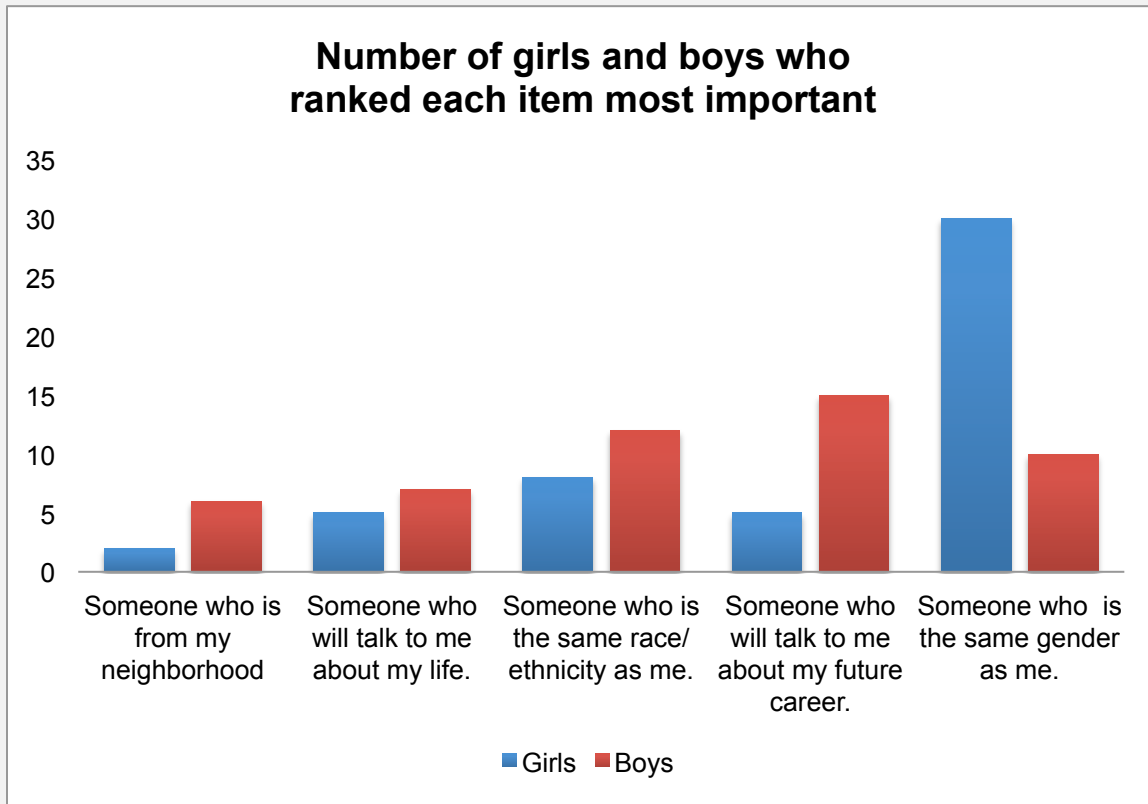


As they examine the histograms, the team considers whether any frequencies for specific responses appear to be unusually high or low. A helpful rule of thumb for this process is: A response that appears more than 1.5 to 2 times as many times as any other option is likely high. If you have five response options, each response constitutes 1/5 of the total options (20%). If a response appears more than 2/5 of the time (40%), this would be a high frequency. Similarly, a response option that appears less 1/10 of the time (10%) would be a low frequency.

Examine whether frequencies vary across sub-groups

As the team examines these results, they wonder whether there are differences hiding in this data. Could it be that gender tops the list more often for girls, or that race tops the list more often for boys? Examining data in this way will ultimately help the team ensure they are effectively personalizing student learning.

Using a filter (learn more about this functionality on the Microsoft web site here: <http://office.microsoft.com/en-us/excel-help/filter-data-in-a-range-or-table-HP010073941.aspx>), they choose only the girls and rerun the same analyses for boys and girls separately, yielding the following results:



When the team examines the data in this manner, they see that girls cluster around valuing a counselor of the same gender. While there are female respondents who endorsed other options, they are fewer and further between. Boys, on the other hand, are more evenly spread between options. The two most frequent top choices are talking about career and race/ethnicity; gender, talking about home life, and neighborhood are farther behind.

Check means

For survey items that use a Likert-scale response set easily translated into numbers, the team can look at means, in addition to frequencies, both overall and by group. They can also examine the range of responses. As such, after examining frequencies, they also consider means, or averages. For columns with numerical codes, it is simple to take the mean of each column using the AVERAGE function in Excel. As you examine means, consider whether any results seem unusually high or low, or are otherwise surprising. For example, on a five-point Likert scale, an average over 4.0 would likely be high, while an average under 2.0 would be low.

Compare means across sub-groups

When the team examined frequencies, they found that girls seemed to especially value having a counselor of the same gender, while boys tended to rank “someone who talks about career” first. They considered that this could have an important impact on how they would not only match students but also design the curriculum. Perhaps boys’ groups should focus on career? They want to see whether this pattern is also apparent in the survey items that ask students to assess the overall importance of counselors of the same gender, and of counselors who talk about career. They therefore compare means for these two survey items across genders.

Using Excel, they take the average of each survey item for girls and boys (to break down by gender, they use the “Filter” function). In doing so, they find the following:

	Mean: Gender, importance	Mean: Someone who talks about career, importance
Boys	3.1	4.2
Girls	4.3	4

As suggested by the above table, it appears that boys do not value having a counselor of the same gender as much as girls. However, based on a comparison of means, girls do appear to value talking about careers about the same amount as boys do—it is just not their top choice.

Consider creating sub-groups from available data

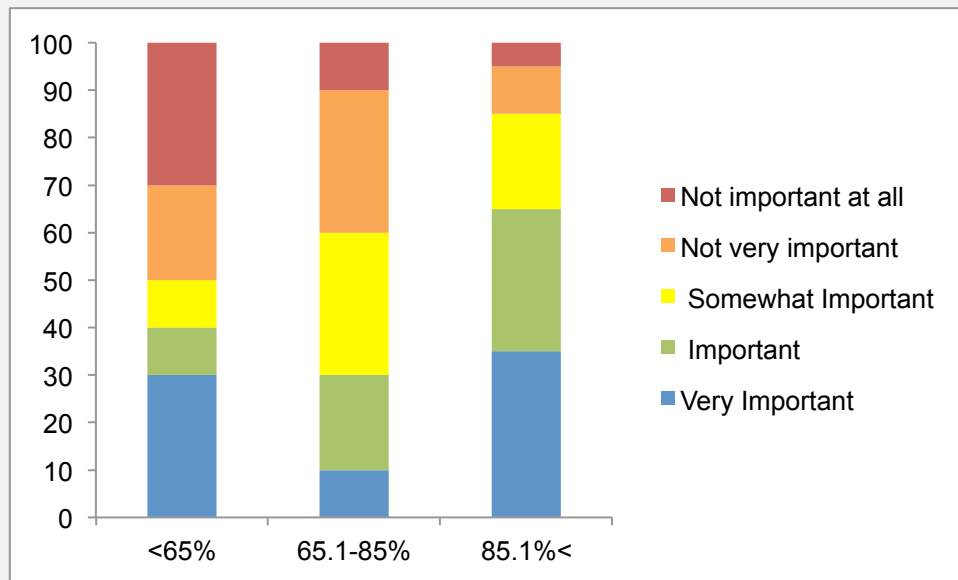
As noted in the previous sub-sections, there may be instances in which you wish to use available data in conjunction with survey data. For example, the Lucretia Mott design team might want to know if rates of attendance affect the importance students assign to their relationships with their counselors.

In their survey, the team asked how important guidance counselor relationships are. They also have access to attendance data for the students who took our survey, so they merged the two data sources together.

First, they create sub-groups out of the attendance data, dividing students into three groups: students with year-to-date attendance below 65%, students with 65-85% attendance, and students with 85.1% attendance or better.

Once they have separated attendance data into these groups, they examine the means and histograms for each group. Creating a stacked bar chart comparing importance of counselor relationships across the three attendance groups proves illustrative:

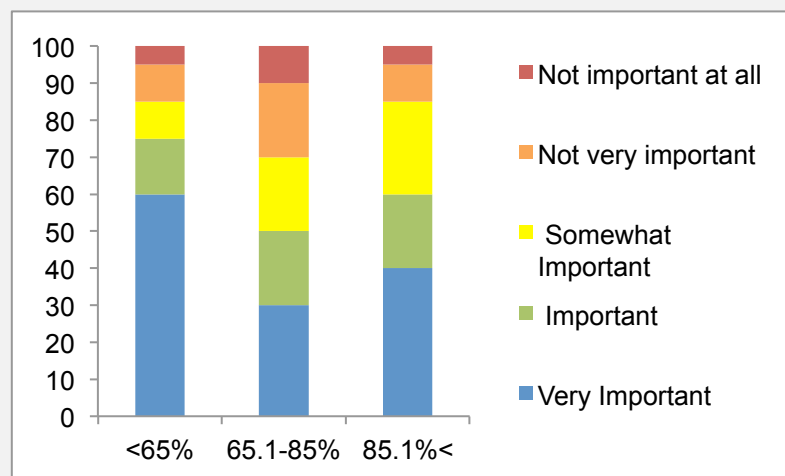
Importance of Counselors by Attendance Rate



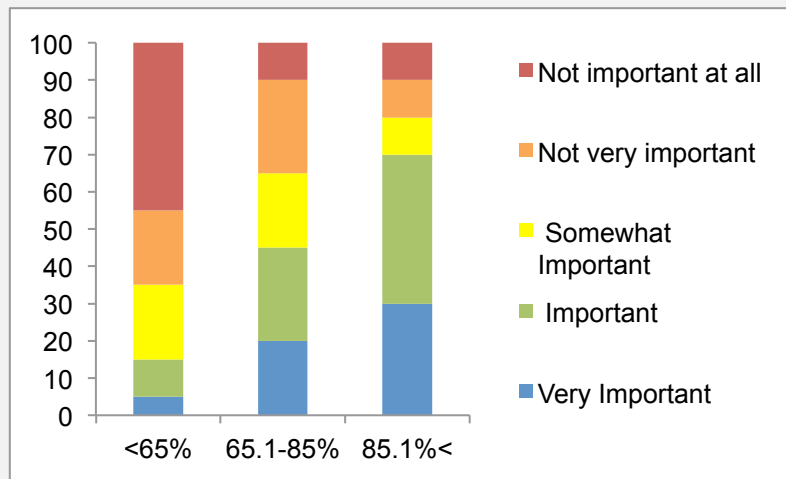
This analysis shows that students with high attendance find counselor relationships very important, while those with moderate attendance say counselor relationships are “somewhat” to “not very” important. Students with low attendance either highly value counselor relationships or say they do not value them at all.

Noting this relationship, the team checks what these different groups of students say about the importance of the two other relationships addressed in the survey: teachers and adults at home. They find that students with low attendance perceive adults at home as very important, but assign very little importance to their relationships with teachers, as shown on the following page. This leads them to the hunch that if counselors can help disengaged students to develop relationships with one teacher, they may see improvements in attendance, leading to further ideas and sub-questions about how the counseling match and counseling curriculum can best engage students.

Importance of Adults at home by Attendance Rate



Importance of Teachers by Attendance Rate

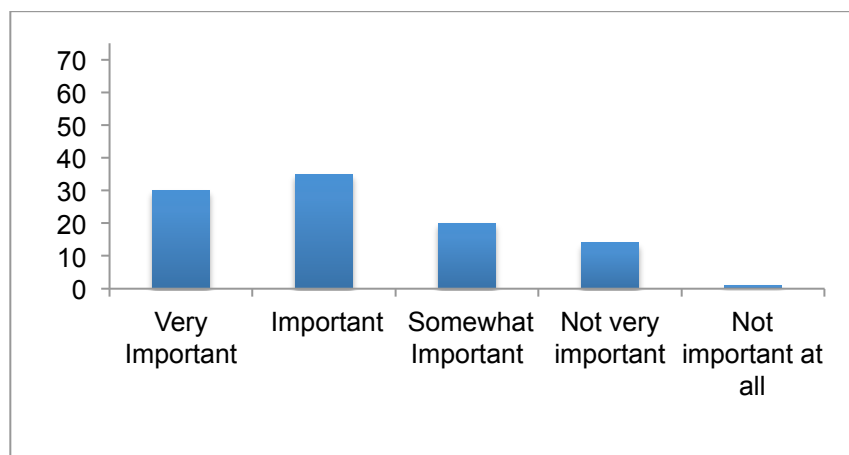


Means can be deceptive!

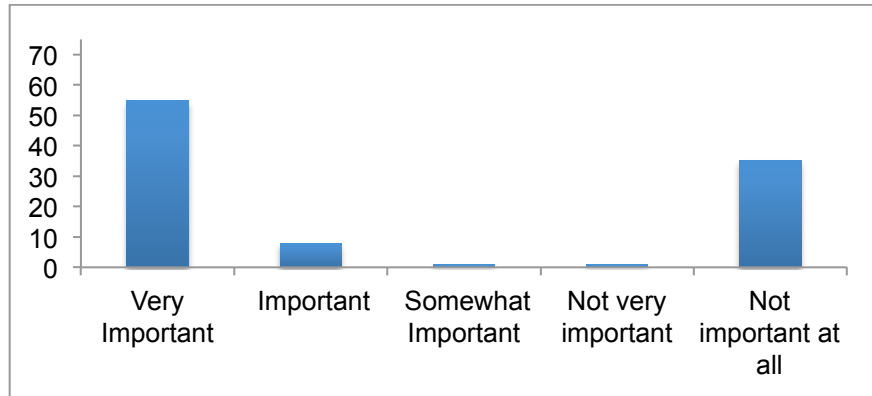
While means are a useful statistic, it is important to still take frequencies into account when looking at means.

Why? Means are very helpful because they give a sense of the “central tendency” of your data. However, they are also potentially misleading: they almost always obscure a larger story. In some cases, this happens because the data is not normally distributed—that is, the scores do not spread out evenly from a central point. In other cases, even if the data is normally distributed, looking only at the mean can obscure important differences between sub-groups.

When examining means, you should also look at the histogram of responses to see what patterns emerge, as very different patterns can yield similar means. For example, for a graph with a mean of 3.7 on a scale of 1 to 5, we might see a histogram that looks like this:



In this instance, we would know that very few people endorsed lower options on the scale, suggesting that every respondent found the item at least somewhat important. However, consider how our analysis would change if the bar chart for the same mean of 3.7 looked like this:



Despite having the same mean as the previous graph, the interpretation would be quite different: people either felt that the item was very important, or not important at all, with little middle ground.

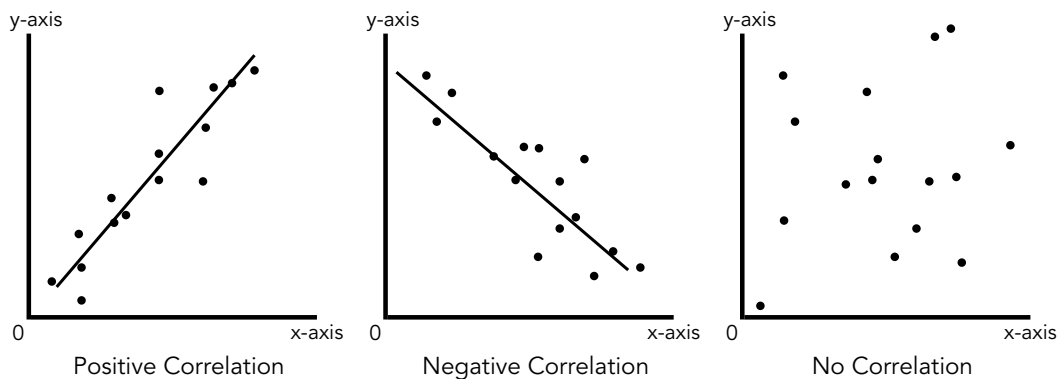
With this in mind, when examining means, make sure to also look at histograms. If your answer choices are unevenly distributed in the histogram, consider why this might be.

5. EXAMINE RELATIONSHIPS THROUGH CORRELATIONS

KEY INFORMATION

Look at how variables change together. Data analysis can help you see whether and how two variables change together. You can examine correlations—a numerical representation of the strength of a relationship between two variables—using Excel.

Types of correlations. Correlations run from -1.00 to $+1.00$. A zero correlation means there is no discernible relationship between the two variables. A positive correlation, in turn, means that as one variable increases, so does the other. Conversely, a negative correlation means that as one variable decreases, the other increases. Images of types of correlations appear below:



Interpreting correlations. When interpreting correlations, you are able to hone in on the interaction between pieces of information that might be revealing or unexpected. Take the absolute value of the correlation and consider the following ranges:

0.00-0.09 (Zero correlation): This is a potentially interesting result. This shows you that two variables are not related. If you expected that students with strong attendance would report strong relationships with adults, but the correlation is only 0.07, you would have to revise your understanding. When you see a correlation at this level, you should immediately ask yourself: What can explain why these two items do not go together?

0.10-0.30 (Weak correlation): This signifies a weak correlation and is usually too vague to be revealing.

0.31-0.60 (Moderate correlation): This suggests that two variables are related. If you were not sure about the relationship between attendance and students' sense of connection to school and saw a 0.52 relationship, you would have evidence that these variables are connected.

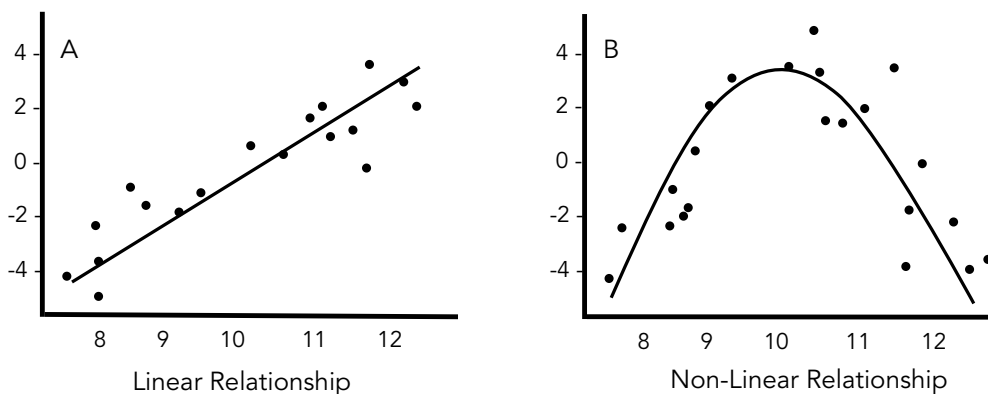
0.61-0.90 (Strong correlation): This is almost always an interesting result. This signifies a strong correlation and gives good evidence that two variables are connected. When you see a correlation at this level you should immediately ask yourself: What can explain why these two items go together?

0.90-1.00 (Redundant variables): This signifies a correlation so strong as to suggest that two variables are identical. Sometimes this result is revealing because it tells you that two questions asked in a survey were really asking (or perceived to be asking) the exact same thing.

Formatting data. To calculate correlations, be sure that your data is laid out so that each student has a row, and in each column is the student's numerical response to the given variable of interest.

Correlations and linear relationships

Correlations are only relevant for linear relationships, meaning that the effect of one variable on the other is expected to change in a constant fashion and would appear as a straight line if graphed:



Consider Image B above. If you tried to fit a straight line to the points in that image, you would obscure the truth about the data: that the relationship between the two variables is curvilinear. Up to a certain point, there is a positive relationship between the two variables, but then the correlation flattens and becomes negative.

Producing a scatterplot (one of the available graph types in most recent versions of Excel) for two variables for which you want to calculate a correlation can give you a good sense of whether the relationship is linear or non-linear.

APPLICATION TO EXAMPLE

After analyzing their data with descriptive statistics, the Lucretia Mott design team wants to know: Does feeling that counselors are important to school success correlate to feeling that parents and/or teachers are important? That is, do students simply find all the adults in their lives important or unimportant together (suggested a high positive correlation); do some students find one adult more important in a way that bears no relation to how they think of other adults (suggested by a low correlation); or do students substitute one adult for another so that, for instance, students who find adults at home more important express less importance for counselors, and vice versa (suggested by a high negative correlation).

To explore this question, the team looks at correlations across students' responses to the questions asking them to rate the importance of adults, parents, and counselors in their success at school. They use the CORREL function in Excel (see the Microsoft web site for more on this function: <http://office.microsoft.com/en-us/excel-help/correl-HP005209023.aspx>). A subset of the data appears below, along with the formula to calculate the correlation of the importance of counselors and teachers:

StudentID	CounselorImp	TeacherImp	ParentImp
001	4	5	4
002	2	2	1
100	3	4	5
=CORREL(B2:B101,C2:C101)			

Through this analysis, the team finds:

	Correlation between value placed on teacher and counselor relationships	Correlation between value placed on parent and counselor relationships
Boys	-0.62	+0.15
Girls	+0.07	+0.58

This result indicates that girls place the same kind of value on parent and counselor relationships—if they value one, they value both. If they do not value one, they do not value the other. The correlation is moderate, so it is not a perfect relationship: there is room for exceptions to this rule. Conversely, the team sees that boys do not have a strong relationship between how they value counselors and how they value parents. With respect to the relationship between teacher and counselor relationships, the design team finds that boys have a strong negative relationship, meaning that if they value one, they do not value the other.

The team considers this finding, and wonders whether, if they are trying to set up at least one positive adult relationships for every student at the school, they might look into making this a counselor-student or teacher-student relationship for boys (for whom the data suggests one relationship is key), while, for girls, they might plan to put more effort into engaging adults at home to try to help navigate that relationship (since the data seems to suggest a high connection between parent and counselor relationship for girls).

6. EXPLORE MORE COMPLEX DATA ANALYSIS OPTIONS

KEY INFORMATION

There are many other, more sophisticated statistical tests and tools available to you, which you can use to make additional inferences about your data. These tests require more statistical knowledge than the descriptive techniques described in the previous sub-sections, and should only be undertaken with the support of someone familiar with inferential statistics:

Standard Deviation

The standard deviation for a set of data (STDEV function in Excel) gives you a sense of how variable the data is. Almost everyone in a population (to be precise, 95.45% if your group is normally distributed) fall within two standard deviations of the mean. For instance, if you have a group of students whose average course passing rate is 85% with a standard deviation of 2%, then you know that almost all of your students have passing rates between 81% and 89%. Compare this to another group with an average of 85% and a standard deviation of 15%. With this group, you know less: passing rates range from 55% all the way up to 100%.

Tests of Group Differences

Tests of group differences assess whether the differences in means across groups for a given variable are statistically significant—that is, whether the difference in means is greater than would arise by chance. If a difference in means is statistically significant, that indicates you are 90-95% sure it did not arise merely by chance. There are three basic tests of group difference:

- T-test: Compare means of a continuous variable (such as a score on a test) across two groups (TTEST function in Excel).
- One-way ANOVA: Compare means of a continuous variable (such as a score on a test) across three or more groups (easiest to do with a statistical software package).
- Chi-Square tests: Compare whether a group has more of a certain discrete trait or variable (such as a responses to a question of ethnicity or gender or a yes/no variable) than you would expect by chance. For example, if we see that one group has more people who receive free or reduced-price lunch than another group, we could use a Chi-Square test to determine whether this difference is significant, or if it arose by chance (CHISQ.DIST function in Excel).

Model-Building

T-tests, one-way ANOVAs, and Chi-Square tests examine the relationship of two variables at a time. They do not, however, recognize the inter-relationships between other variables of interest. Models can take into account how multiple variables affect a single outcome variable. For example, if you wanted to know whether positive adult relationships affect grades, we might also want to include socioeconomic status in our analysis. By including socioeconomic status, we could examine the impact of positive relationships over and above socioeconomic status on grades. As such, models can be built to control for many variables at once. With the correct control variables in the model, it is possible to isolate the unique contribution of the variable of interest.

Person-Centered Analysis

Person-centered analysis is a way to analyze how variables of interest tend to co-occur in individuals. For example, we might find that there tend to be three types of people: people who benefit from positive relationships with adults, people who do not benefit from these relationships, and people who are negatively affected by these relationships.

You can think of this process as the inverse of sub-group analysis. In person-centered analysis you let differences emerge from the data to create groupings. Often you can descriptively interpret the groups that emerge and then use those groups to understand other data points.