Section 2: Ten Tools for Applying Sociology

CHAPTER 2.8: ANALYZING DATA

QUICK START: In this chapter, you will learn

- How to perform the basics of quantitative and qualitative data analysis.
- How to organize data so it can be analyzed.
- To determine which quantitative statistics are appropriate for different levels of measurement.
- To calculate univariate statistics and frequency distributions.

TERMS

Data Analysis	Searching for patterns in the information you've collected so you can make sense out of social reality.
Qualitative Analysis	Ways to analyze a written, audio or video recording of people's words or behavior.
Quantitative Analysis	Ways to analyze numbers that represent individual or group characteristics, attitudes, values, beliefs or behaviors.
Frequency	The number of responses for each value of a variable in a sample.
Frequency Distribution	A table displaying the frequency and percentage of all of a variable's values.
Percentage	The frequency for each value divided by the total number of cases in the sample, multiplied by 100.
Descriptive Statistics	Providing the answer to "What is it?" and/or "Who and how many?" by using numbers.

Chapter 2.8: Analyzing Data

Mode	The most frequent response for a variable (sometimes there is more than one).
Median	The value of a variable that has 50% of the responses above it and 50% below it.
Mean	Sometimes called the average, it is the sum of the products of each value and its frequency (Σ x), divided by the total number of cases in the sample.
Univariate Statistics	Summary statistics on one variable.
Bivariate Statistics	Statistics that reflect the impact of one variable on another.
Crosstabulations	A way of displaying and analyzing two categorical variable's joint frequency distributions in a table, with one variable's values listed across the rows and the other variable's values listed across the columns.
Compare Means	Analyzing continuous data by comparing the mean of one group or sub sample with the mean of another.
Correlations	Analyzing continuous data by assessing the tendency for one variable to increase or decrease in value as another variable increases or decreases in value.
Dependent Variable	A variable that changes in value as a result of change in another variable.
Independent Variable	A variable that influences or causes a dependent variable to change in value.
Categorical	A variable with nominal or ordinal level of measurement.
Continuous	A variable with interval or ratio level of measurement.
Recode	To regroup a variable's values, such as regrouping an age variable measured in years into a variable measured in the categories of 'less than 21', '21-29', '30-39', etc.
Linear Relationship	A relationship between an independent and a dependent variable in which incremental changes across the independent variable result in consistent changes in the dependent variable.

Statistical Significance	Whether two estimates, such as means, percentages, or correlations, are statistically different. Statistical significance helps sociologists, and other social scientists, determine how different two estimates need to be before they are confident in stating that there is an important difference.
Multivariate Statistics	Analyzing data by assessing simultaneous correlations of several variables on another variable.
Statistical Control	The ability to keep one variable's values constant in an analysis so as to isolate the effect of a particular independent variable on a dependent variable. For example, being able to keep the influence of sex out of an analysis on the influence of education on occupational prestige.

TOPICS

In this chapter, we provide you with a quick introduction to the basic **data analysis** techniques. We think the basics below will help you both as citizens and as professionals. In general, applied sociologists use two approaches to analyze data that they collect on a problem: quantitative data analysis and qualitative data analysis. Let's look at each of these approaches.

Quantitative Analysis	**Qualitative Analysis**
Look for patterns in numbers through the use of statistics such as percentages, means and correlations.	Look for patterns in written text, notes, documents, audio recordings, or video recordings.

As an applied sociologist, whether you use quantitative or qualitative research methods in trying to understand and solve your client's problem, you will likely gather a boatload of information, otherwise known as data. How will you make sense of this information? How will you learn from the data you collect? The answer is data analysis. In this chapter we will introduce you to the world of data analysis. These data analysis tools will serve you in any social or physical science pursuit.

The first step in data analysis, regardless of data collection method, is to answer the question, "What's going on here?" This is basically a descriptive question. It asks you to describe the initial facts about the problem at hand. An answer would include the number of people included in your study, their characteristics such as demographic information, and a description of the setting (if applicable). To get started in answering this question, you should calculate and display the frequency of each value for all of your variables. A **frequency** refers to the number

of times something happens. Central to describing something is to count the number of times that something happens. A **frequency distribution** contains both the frequency of each of a variable's values, and each frequency expressed as a **percentage**. Suppose we have asked a random sample of Americans their opinion on whether the U.S. military should accept homosexuals. Let's compute a frequency distribution on this variable.

Compute a Frequency Distribution

1. Identify the target group you want to describe. Is it all the participants or respondents in your study? Or a sub-group, such as only the residents of a particular county, or only the men, or only people of a particular ethnicity? In this case, we want to use all respondents, but later we will look at only male respondents.

2. Determine the variable you want to describe? Out of one study you may have dozens of variables. Which one(s) is of interest now? How is it measured?

   Should U.S. military bar homosexuals?

   1. Yes
   2. No
   8. Don't know
   9. No answer

3. Count the number of people that responded with each response category:

Response	Frequency
1. Yes	573
2. No	301
8. Don't know	88
9. No answer	38
Total	1000

4. Translate the number of responses into percents (some people call these relative frequencies)

Response	Frequency	Percentage
1. Yes	573	57.3%
2. No	301	30.1%
8. Don't know	88	8.8%
9. No answer	38	3.8%
Total	1000	100%

5. Use the percentages to calculate a cumulative percentage column. This column adds each percentage together, moving down the column.

Response	Frequency	Percentage	Cumulative Percentage
1. Yes	573	57.3%	57.3%
2. No	301	30.1%	87.4%
8. Don't know	88	8.8%	96.2%
9. No answer	38	3.8%	100%
Total	1000	100%	

We can use a frequency distribution to easily calculate some **descriptive statistics**, such as the **mode**, **median** or **mode**. These simple statistics will help you answer the question, "What's going on here?" Descriptive statistics are also referred to as **univariate statistics**, because they describe one variable. Descriptive or univariate statistics are used in both quantitative and qualitative research. The table below summarizes which descriptive statistics are appropriate for different levels of measurement.

Level of Measurement	**Univariate Data Analysis Options**
Nominal	Mode
Ordinal	Mode, median
Interval	Mode, median, mean

So, what level of measurement is our variable on American's attitudes about homosexuals in the military? It is nominal, because the response categories can't be ranked as more or less, better or worse, or so on. We assigned a code of 1, 2, 8 or 9 to each response category. But these code numbers contain no inherent meaning. They are simply codes. This means that the only statistic, which is appropriate to summarize this variable, is the mode. To determine the mode, we can refer to the percentage column. Which category has the largest percentage? That is the mode. In this case it is response "Yes" with 57.3% of the responses. So, most respondents think that homosexuals should be barred from the U.S. military.

The median and mean are not appropriate for the variable above. For example, how could we calculate an average from yes's and no's? You could take the codes for those responses (1's and 2's) and average them, but then you would obtain a number between those, such as 1.3. What would that mean?

To illustrate how you calculate a mean and range from a frequency distribution, we need another variable, one measured on an interval level. Let's say, as part of the above study, we also collect data on the number of homosexual friends or family members that respondents had at the time of the survey. We collected this data because we thought it might influence whether

respondents think the U.S. military should bar homosexuals. With this data, we calculate the following frequency distribution.

Response	Frequency	Percentage	Cumulative Percentage
0	907	90.7%	90.7%
1	40	4%	94.7%
2	25	2.5%	97.2%
3	18	1.8%	99%
4	8	.8%	99.8%
9	1	.1%	99.9%
12	1	.1%	100%
Total	1000	100%	

To determine the median, we just look at the cumulative percentage column to see which category represents 50% of the respondents. In this case, it is the first category "0 homosexual friends." To calculate the mean, we simply multiply each frequency by the response value and add up those results. Then, divide that number by the sample size.

$$\text{Mean} = [0(907)+1(40)+2(25)+3(18)+4(8)+9(1)+12(1)]/1000 = .20$$

On average, the respondents have .20 homosexual friends or family members. In this case, a mean of .20 people does not make much sense. To obtain a more meaningful number, we can round down to 0 friends or family members. If our mean was .50 or higher, we would have rounded it up to 1.

Combined, we can use univariate statistics to describe the demographic characteristics of a sample. For example, in your qualitative study on organizational culture (from chapter 2.5), you might provide an initial description of the people you interviewed and observed as follows: "Of the 250 employees (frequency) participating in the study, roughly 80% are female (percentage), and 34% are ethnic or racial minorities (percentage). The average age of the responding employees is 52 (mean). Approximately half of the respondents make less than $25,000 a year, and a half make more than that (median). To calculate these statistics you would follow the same steps as we did above.

Overall, a frequency summarizes how a group of people thinks on a particular issue or their experiences on a particular issue. You answered the question, "What is going on?" This is an important place to start. But next you will want to answer, "Why are these things happening?" Answering this question requires a more thorough analysis of the data. To answer it, you will want to see how people compare on particular issues and experiences. This is where you need **bivariate statistics**. This second step in the data analysis process is also where quantitative data and qualitative data analysis divide into separate techniques. We introduce you to these techniques in the Tools section.

TOOLS

Qualitative and quantitative methods to answer, "Why are these things happening?" differ. Qualitative analysis is more conceptual, theoretical, and logical. The tool is the researcher and his or her ability to think about the data collected. In qualitative analysis, researchers systematically review their data over and over again, looking for patterns in the data. This is a rigorous, lengthy process. Remember the data here are transcribed interviews, observation notes, and interviewer notes, which add up to a large amount of data to review. Some sociologists use computer programs, such as *Nudist* and *Ethnograph,* to help them identify patterns in the data.

Quantitative data analysis is more technical. There are specific statistical tools such as **crosstabulations**, **means comparisons**, **correlations**, and other bivariate statistics to help the researcher analyze the data. These statistical tools quickly find the patterns in the data for the scientist. Sociologists use computer software packages such as *SPSS* (Statistical Packages for the Social Science) and *SAS* (Statistical Analysis System) to calculate these and other statistics. The difficult part of quantitative analysis is determining what statistical tools are appropriate for the task at hand, and knowing how to interpret the statistics.

Qualitative Analysis: In qualitative data analysis, you need to systematically and rigorously analyze all of the data collected. To keep this process from being overwhelming, we need to break it down into steps:

Steps to Qualitative Analysis
1. Read all of your data (transcribed interviews, interview notes, observation notes), making notes on any possible patterns you identify while reading, as well as any concepts that make sense of those developing patterns.
2. Study the notes and ideas that you wrote while reading your data. Take your knowledge of the topic and the project and try to make sense of the patterns and concepts identified in your notes. Let your knowledge inform your original list of patterns and concepts. Then revise your list, giving each pattern and concept an abbreviated name or code.
3. Refocus your analysis on the patterns and concepts developed at the end of step 2. Read all of your data again, circling, highlighting or in some way marking off the text illustrating each pattern and concept. Write the code name next to each circled segment of text.

Steps to Qualitative Analysis
4. Repeat step 3. Here you may code your data exactly as you did in step 3. If so, you have a sense of strong validity and reliability of your analysis. In contrast, you might change the codes on the same segments of text. Or you may code additional segments of text with the existing patterns and codes. You may even identify entirely new patterns and concepts and codes for them in this step. You may want to repeat step 4 again, after a week or two, as another check on the validity and reliability, especially if you make any changes during the first run through step 4.
5. Record how many times each pattern and concept is coded in step 4.
6. Pull out the coded segments of text. Read these again, looking for relationships between them. For example, what patterns or concepts occur in co-presence of others? Which usually come before or after others? Which only hold for sub-groups of your entire sample? Which patterns evolve differently for different sub-groups? It may help to develop new codes for these relationships, and then to code each segment and sort the segments accordingly. You may need to repeat this step several times. Make notes of your "findings" as you proceed through this step.
7. Record how many times each relationship between patterns and concepts occurs in step 7.
8. Write in paragraph form a description of the patterns and concepts you found in step 4. Then write in paragraph form a description of the relationships between the patterns and concepts that you found in step 6. This process will take several drafts. Your goal should be to "tell the story" of your findings. Include how often the patterns, concepts, and relationships occur.
9. Choose excerpts from within the segments of coded text to illustrate the patterns, concepts and relationships. Weave these excerpts into the story you wrote in step 8.

Many sociologists find it helpful to write about what they are finding at each step in the qualitative data analysis process. Writing what you are thinking and learning can clarify and push your reasoning.

Quantitative Analysis

Bivariate Statistics: Bivariate statistics look at the relationship between an independent variable and a dependent variable. An independent variable influences or predicts another variable. A dependent variable is the variable that is influenced or predicted. For example, in some societies age influences or predicts the amount of respect that people receive from others. Age is the independent variable, and respect is the dependent variable. Here is another example: In

our society, educational attainment usually predicts occupational prestige and income. People with higher educational attainment tend to have occupations with higher prestige and income. In this case, education is the independent variable, and prestige and income are dependent variables. Often times, variables can be both independent and dependent variables. That is, in one analysis, a variable might be an independent variable. Then, in another analysis, it could be a dependent variable. For example, you could have gone on to look at the influence of respect on self-esteem or health. So remember, you must determine whether a variable is an independent or dependent variable in each analysis.

The three most common bivariate statistics or techniques are crosstabulations, comparison of means, and correlations. Which one is appropriate to use depends on the level of measurement of the independent and dependent variables.

Bivariate Data Analysis Techniques

	Dependent Variable	
Independent Variable	**Categorical** (Nominal or Ordinal)	**Continuous** (Interval)
Categorical	Crosstabulation	Compare Means
Continuous	Crosstabulation*	Correlation

* Requires **recoding** the continuous variable into categories.

Crosstabulations: A crosstabulation is simply a Frequency Distribution that combines how people responded on two or more variables. Your purpose for doing this may be twofold, both of which are perhaps best understood through example. First, we might want to know how often married men and single men go to a doctor when they are ill. You might want this information for a project on men's health needs. Here we have two variables: marital status and doctor's visits. We use a crosstabulation to describe how many people in each of the categories of one variable (marital status) responded with each category of another variable (doctor's visits). In this case, we use a crosstabulation to further describe a sample.

A second reason for doing a crosstabulation (often called a crosstab) is a little more complicated. Your client may believe there's an association or relationship between two variables. Suddenly we have a hypothesis! You've got it, a presumed statement of fact. For example, we might hypothesize that marriage has some impact on whether men see a doctor when they are sick. You might suspect that women encourage their husbands to see a doctor when they are sick. We can use a crosstabulation to test this hypothesis. Regardless of whether we simply want to describe a group of people, or if we want to test a hypothesis, we will create a crosstabulation in the same way. Here's how you do it.

Chapter 2.8: Analyzing Data

How to Crosstabulate Two Variables

1. Identify the target group you want to describe. Is it all the participants or respondents in your study? Or a sub-group, such as only the men? In this case, we want to use only the male respondents' data.

2. Determine the variables you want to put in your table and how they are measured? In this case it is:

   Currently Married: Yes, No

   Visit a Doctor when Sick: Always, Often, Not Usually, Never

3. Create a frequency distribution for married men's doctor's visits. And a frequency distribution for single men's doctor's visits.

Married Men	Frequency	Percentage
Always	45	10%
Often	189	42%
Not Usually	124	27.5%
Never	92	20.4%
Total	450	99.9%*

*Sometimes percentages do not add to an even 100% due to rounding error.

Currently Unmarried Men	Frequency	Percentage
Always	21	4.7%
Often	30	6.7%
Not Usually	246	54.7%
Never	153	34%
Total	450	100.1%

4. Combine the two Frequency Distributions above into one table, with a frequency and percentage in each cell.

	Currently Married	Not Currently Married
Always	45	21
	10%	4.7%
Often	189	30
	42%	6.7%
Not Usually	124	246
	27.5%	54.7%
Never	92	153
	20.4%	34%

Now we can compare the response percentages of currently and not currently married men to assess whether marriage has an effect on how often men visit a doctor when they are sick. You can see by looking down the Currently Married column, that the most popular response was "often" (42%). Looking down the Not Currently Married column, you see that the most popular response was "not usually" (54.7%). This pattern suggests that married men visit the doctor more often than single men. Looking across the rows, you see that very few men, regardless of marital status, see the doctor every time they are ill (10% and 4.7%, respectively). Next, many more married men (42%) then single men (6.7%) report that they see the doctor often when they are sick. In a reverse pattern, many more single men (54.7%) than married men (27.5%) report that they do not usually see a doctor when they are sick. And, similar percentages of married (20.4%) and single men (34%) report never seeing a doctor when they are sick.

To crosstabulate two variables where the independent variable is continuous, you will follow the same steps as above, but first you must **recode** the continuous variable into categories. To recode a variable is to change the way the variable is measured. You can either collapse categories into each other, or delete categories. You can recode any variable, regardless of level of measurement. In the case of a crosstabulation with a continuous independent variable, you would first examine the frequency distribution of the original variable. For example, let's say your variable is age, which you measured as follows: "What is your age? ___", with people filling in their age in years. You could take this variable and recode it into categories, such as:

1. Younger than 25
2. 26-45
3. 46-65
4. 66 and older

Now you have an ordinal variable that you can include in a crosstabulation. You should always choose categories that make sense for the topic you are studying and the people participating in the study. Let's say we have survey data on recreational boaters' attitudes about new laws requiring a driver's license to operate a boat. In contrast, if we studied students' attitudes about sex education in school, we would need different age categories. On a related note, sometimes you have a variable that was measured continuously, but the actual responses do not represent a wide, or continuous, distribution. In this case, you can use the continuous variable in a crosstabulation without recoding it. For example, in our earlier study we asked people how many times they went to see a doctor last year. The responses may have been 0, 1, 2, 3 and 4. This range of responses can easily be incorporated into a crosstabulation.

Comparing Means: Wouldn't it be nice if those tables that we just looked at could be reduced to just one or two numbers? You've already learned tools to do this. A long time ago you learned how to calculate averages. An average, also called a mean, is one number that summarizes a particular characteristic of a group of people. For example, among a group of your friends, you might ask how many college applications they plan to submit. Your friends might respond: 5, 0, 2, 2, 1, 0, 4, 3, 0, and 1. If you then try to talk about the number of colleges your friends are applying to, it would prove cumbersome to have to list out all of these numbers. Instead, you

can compute a mean, which will summarize all of the numbers. The mean in this case is 1.8. This number describes the number of college applications that is closest to most of numbers that your friends reported. A mean is a way of expressing generally what's going on with a group of people.

Means are also handy when you need to compare two groups quickly. For example, let's say you are working for a county recreational department that recently implemented a new marketing campaign targeting senior citizens. It has been six months since they started the new program, and the recreational department wants you to tell them if the program is working. You can use means to help you with this evaluation. First you would calculate the mean number of seniors attending recreational activities before the program went into effect, and then the mean since the program was implemented. Then you can compare those two means. If the mean attendance before the new program was 6 and the mean after the program is in place is 15, that suggests that the new program is working.

As another example of comparing means, our travel company from Chapter 2.7 might want to expand their clientele base. Most of their current clients are couples without children. They hire you to help them plan vacation packages that fit the lifestyles of a more diverse population. You collect this data by using a telephone survey of a statewide sample. Among other questions, you ask how many vacations people plan to take next year, how long the vacations will be, and a series of demographic questions. You find that people with small children plan to take several short vacations of, on average, 4 days each. People without children or with older children plan to take one or two longer vacations of, on average, 9 days each. You could look at those two means and conclude that people with and without small children would be attracted to different types of vacations. The travel agency decides to create several reasonably priced, short vacation packages at family-friendly recreational sites across the state. And, they create a few more expensive vacation packages requiring air-travel to resort style exotic locations.

Correlations: With a correlation we can reduce the information learned from columns of data into just one number. A correlation is a statistic that describes the **linear relationship** between a continuous dependent variable and continuous independent variable. A linear relationship refers to the change that occurs in the dependent variable as a result of a change in the independent variable. So, with a linear relationship, if the independent variable increases in value, we would expect a change in the dependent variable. Correlations range from –1 to +1 with integers in between. A negative correlation means that the variables are negatively related. That means that as the independent variable increases in value, the dependent variable decreases. Or as the independent variable decreases in value, the dependent variable increases. With a negative correlation, the two variables change in opposite directions.

A positive correlation means that the variables have a positive relationship. That means that as the independent variable increases, the dependent variable increases too. Or as the independent variable decreases, the dependent variable decreases too. The two variables change in the same direction. A correlation of 0 means the two variables have no linear

relationship; they are not associated with each other. The closer the correlation is to 0, the weaker the relationship. The closer the correlation is to +1 or –1, the stronger the relationship.

As an example, let's look back at the relationship between educational attainment and starting income. A high school sociology teacher wants to present this information to the junior class in an assembly on applying to college. He obtains data from a national study that measures educational attainment with years of schooling and starting income in $1000's (for example $25,000 = $25, $32,400 = $32.4). He correlates the two variables to see how education influences income, finding a correlation of +.75. So, as education increases, so does starting income. And since .75 is pretty close to +1, he knows that the relationship is strong. In other words, a one year increase in education results in an almost $1000 increase in starting income.

Correlations can be calculated by hand but it is much easier to use a computer software package such as SPSS or SAS to calculate them. The same is true of frequencies, modes, medians, means, means comparisons, and crosstabulations. You can also use these software packages to calculate the **statistical significance** of the above statistics. Statistical significance is something you will learn more about in a statistics class. But in short, statistical significance helps you to determine whether, say, the mean starting income of people with a college degree is substantially higher than the mean starting income of people without a college degree. As another example, statistical significance could help you decide whether the percentage of women opposed to homosexuals in the U.S. military is lower than the percentage of men also in opposition. On the surface, it may seem that you don't need a statistical test to determine if two means or two percentages differ. You may think that you can simply look at the two estimates and tell if one is bigger than the other. However, how different do the two estimates need to be before you are confident in saying that it is an important difference, one with real impact on people's lives? This is where statistical significance can help.

Multivariate Statistics: So far we have discussed univariate and bivariate data analysis. However, there is an even larger set of options available to you in the form of multivariate analysis. The most popular use of multivariate analysis is to look at the combined influence of a group of independent variables on a dependent variable. Such an approach is very useful because it models what happens in the real world. In the real world many things simultaneously influence people's behavior, values, and attitudes.

Here's an example. A large city's crime rates are rising. The police chief is considering submitting a request to the mayor to hire more police officers. The police chief first hires you to help determine why the rates are increasing. You gather data on various crime rates across several decades, along with demographic information, including the number of police officers, and the age, race and gender of the citizen population. A multivariate analysis would allow you to examine the influence of all these independent variables together on the crime rate.

Multivariate analyses also allow you to do a number of other things too. They provide **statistical control**, which enables you to look at the unique affect of each independent variable on the dependent variable. For example, you might want to look at the influence of the number of police officers on the crime rate separate from the influence of population size. Sociologists

use computer software to perform multivariate analyses. Multivariate analyses are complicated to calculate and to explain -- too complicated to review here. We encourage you to learn more about them by reading additional books or taking a class in statistical analysis.

TASKS

1. A college president hires you to estimate the occurrence of hazing on campus. You interview small samples of freshman students living in different dorms. Below is data from 20 freshmen in one dorm on the number of hazing events they participated in this past year. Calculate a Frequency Distribution with this data and calculate the mode, median, and mean?

   0, 1, 1, 1, 4, 3 0, 0, 2, 0, 2, 1, 2, 3, 4 1, 1, 0, 3, 0

2. What analysis should you do for the following:

   - The influence of self-esteem on depression, with self-esteem measured on a scale from 0-10, where 0 means no self-esteem and 10 means high self-esteem, and depression measured on a scale from 0-10, where 0 means not at all depressed and 10 means very depressed.

   - The influence of religious affiliation on whether or not people volunteer in their communities.

   - The influence of income in dollars on political affiliation (Democratic or Republican).