Social Data Analysis

Social Data Analysis

Qualitative and Quantitative Approaches

MIKAILA MARIEL LEMONIK ARTHUR AND ROGER CLARK

KALEIGH POIRIER

PROVIDENCE, RI

Contents

Section IV. Quantitative Data Analysis With SPSS

Section V. Qualitative and Mixed Methods Data Analysis with Dedoose

Acknowledgements

Dragan Gill worked tirelessly to gain access to the platform that made this text possible and provided technical and other support throughout the process of creating the text.

Kaleigh Poirier made an invaluable contribution to this book by translating many of the formulas it contains into LaTeX code.

This text was made possible, in part, by a Rhode Island College Committee for Faculty Scholarship Major Grant Award. Funding from Rhode Island College, the Rhode Island College Foundation, and the Rhode Island College Alumni Affairs Office provided essential editorial support.

Finally, thanks to the many students who have enrolled in Sociology 404 with Drs. Arthur and Clark over the years. Their questions, struggles, and successes have shaped the material presented in this text in vital ways.

How to Use This Book

This book is divided into four parts:

1. A conceptual section on conducting quantitative data analysis
2. A conceptual section on conducting qualitative data analysis
3. A practical section on conducting quantitative data analysis using SPSS
4. A practical section on conducting qualitative data analysis using Dedoose

Each part can be used separately by those interested in developing the relevant skills.

Each chapter includes suggested exercises at the end of the chapter for those seeking practice with the ideas, concepts, and skills introduced in the chapter. There is also a hyper-linked glossary of terms. Bibliographic information is available in a separate bibliography rather than in each individual chapter.

For users who prefer to download or print the text, go to the text homepage and click "Download This Book." It is available in PDF and other formats for use in ereaders; those who want to print can bring the download to a local print or office store. The text is designed to be compatible with screen readers; where applicable, image descriptions for software screenshots provide instructions for how to use keycodes to access key functions. Should users discover any screenreader compatibility problems, they are welcome to email Mikaila Mariel Lemonik Arthur to get them corrected. Note that while SPSS is basically screenreader compatible (plugins may be required depending on a user's specific system configuration), Dedoose is not.

In addition, as an Open Educational Resources text, the authors encourage others to develop equivalent practical sections using other software packages, like Atlas.ti, Nvivo, Stata, SAS, R, and Excel. This project can be forked to add such sections, or those interested in collaborating to incorporate new sections into this base text are welcome to reach out to Mikaila Mariel Lemonik Arthur to discuss.

SECTION I

INTRODUCING SOCIAL DATA ANALYSIS

1. Introducing Social Data Analysis

An Overview

ROGER CLARK

Social data analysis enables you, as a researcher, to organize the facts you collect during your research. Your data may have come from a questionnaire survey, a set of interviews, or observations. They may be data that have been made available to you from some organization, national or international agency or other researchers. Whatever their source, social data can be daunting to put together in a way that makes sense to you and others.

This book is meant to help you in your initial attempts to analyze data. In doing so it will introduce you to ways that others have found useful in their attempts to organize data. You might think of it as like a recipe book, a resource that you can refer to as you prepare data for your own consumption and that of others. And, like a recipe book that teaches you to prepare simple dishes, you may find this one pretty exciting. Analyzing data in a revealing way is at least as rewarding, we've found, as it is to cook up a yummy cashew carrot paté or a steaming corn chowder. We'd like to share our pleasure with you.

Quantitative or Qualitative Data Analysis?

Our book is divided into two parts. One part focuses on what researchers call quantitative data analysis; the other, on qualitative data analysis. These two types of analysis are often complementary: the same project can employ both of them. But for now we'd like to look at the main distinction between the two. In general, **quantitative data analysis** focuses on variables and/or the relationships among variables. This analysis involves the statistical summary of such variables and those relationships. Roger recently completed a study, with two students, of the relationship between Americans' gender and their party affiliation (Petit, Mellor and Clark, 2020). We were interested in the relationship between two variables: gender and party affiliation. Women are more likely to identify as Democrats in the United States than men. We found that we could largely explain the emergence and maintenance of this relationship since the 1970s in terms of three other variables: the increased participation in paid labor by women, the decreasing likelihood that both men and women are married, and the declining participation of Americans in labor unions. To examine the

relationship among these five variables (gender, political affiliation, labor force participation, marital status and attachment to labor unions) we relied almost exclusively on quantitative, or statistical, analysis.

Qualitative data analysis, on the other hand, focuses on the interpretation of action or the representation of meaning. Roger did another study with a student in which we watched YouTube recordings of Trump and Clinton rallies during the 2016 presidential campaign (Fernandez and Clark, 2019). A careful examination of these rallies led us to the conclusion that Trump rallies typically looked more like quasi-religious events—with participants displaying quasi-sacred objects (like Make America Great Again caps), participating in quasi-religious rituals (like shouting rhythmically and in unison, "Lock Her Up"), and cheering quasi-religious beliefs (like how valuable it would be to slow immigration)—than Clinton rallies did. In this study, we were focused on both interpreting the actions of rally participants and on trying to represent what they meant by those actions.

In practice, researchers often employ both quantitative and qualitative data analyses in the same study. For instance, Roger recently did another project with students[1] (Gauthier *et al.* 2020), one of whose goals was to discern the decades (one variable) since the 1930s in which children's picture books were most likely to depict characters in gender stereotyped ways (another variable). Put this way, the study looks like one that required quantitative data analysis, because it was examining the relationship between two variables—the time in which books were created and the degree to which they depicted characters in gender stereotyped ways. But to discern whether an individual book used gender stereotypes, we had to interpret the actions and thoughts of individual characters in terms of a number of characteristics we viewed as gender stereotyped. For instance, we had to decide whether a character was nurturing (a stereotypically feminine characteristic) and whether they seemed competitive (a stereotypically masculine characteristic). Such decisions are essentially qualitative in their nature.

Consequently, the distinction we've used to organize this text—quantitative vs. qualitative data analysis—is a little misleading. Researchers often employ both kinds of research in the same project. Still, it is conventional for teachers to teach quantitative and qualitative analyses as if they were distinct and who are we to defy convention? Thus, this text includes both chapters about quantitative data analyses and those about qualitative data analyses.

Exercises

1. Roger is actually familiar with research that others have done. He's getting on in years, though, and now most trusts himself not to misrepresent his own work.

1. For this exercise we'd like you to use data from the **General Social Survey** (**GSS**), a survey which has been executed about every other year (sometimes more frequently) since 1972. The GSS is a nationally representative survey of American adults. What we'd like you to do is to use the data it produces, made available for us to work with by the University of California, Berkeley, to check whether men or women have been more likely to participate in the GSS over the years. We'll be using this source a lot in this book, so getting a feel for its use is worthwhile here.

   The data are available at https://sda.berkeley.edu/ (you may have to copy and paste this address to request the website). What we'd first like you to do is connect to this link, then go down to the second full paragraph and click on the "SDA Archive" link you'll find there. Then scroll down to the section labeled "General Social Surveys" and click on the first link there: General Social Survey (GSS) Cumulative Datafile 1972-2018 –release.

   Now type "sex" into the "row" box and hit "run the table." What percentage of GSS respondents have been female? What kind of analysis—quantitative or qualitative—have you done? What makes you say so?

2. Watch the first commercial (about the Toyota Highlander) in this YouTube recording of the 10 Best Super Bowl Commercials of 2020:

   One or more interactive elements has been excluded from this version of the text. You can view them online here: https://pressbooks.ric.edu/socialdataanalysis/?p=76#oembed-1

   Which character in this commercial, would you say, is the main one? What one word, would you say, sums up the personality of this character best? What kind of analysis—quantitative or qualitative—have you done? What makes you say so?

SECTION II

QUANTITATIVE DATA ANALYSIS

2. Preparing Quantitative Data and Data Management

MIKAILA MARIEL LEMONIK ARTHUR

The process of research design and data collection is beyond the scope of this book, but it is worth spending some time on the steps required to get quantitative data ready for data analysis. Social science researchers who are working with quantitative data may have collected that data themselves, or they may have obtained that data from another researcher or from a data repository such as the General Social Survey, a national census bureau or other government data source (e.g. the U.S. Census Bureau), or the Institute for Social Research at the University of Michigan. Preparing data for analysis requires different steps depending on the initial source and format of the data.

When a researcher has collected their own data, they need to enter that data into a computer file in a machine-readable format. Some online survey software systems permit survey data to be downloaded in an appropriate format, but not all do—and if data was collected on paper or face-to-face, it needs additional processing. Typically, research teams enter data into a spreadsheet program like Microsoft Excel or Google Sheets. But doing so requires the creation of a codebook, or a document in which numerical codes are assigned to all answer choices or data entry elements.

Figure 1 provides an example of what a codebook for survey data entry might look like, drawing on a survey a group of students created and administered as part of a research methods course. Each question is assigned a column, and each answer choice is assigned a numerical code, with a special code for missing or unusable data (often 9, 99, 999, or -1). Note that in circumstances where a survey question asked respondents to "check all that apply," each answer choice must be converted into a separate question, with selected and not selected as the coded answer choices. This is one reason why downloaded survey data must often still be prepared for use, as survey software like Google Forms may not reliably process "check all that apply" questions or

1. What is your current employment status? **Column B**
   1__I work for pay full time (including multiple part-time jobs)
   2__I work for pay part time (select only if you work fewer than 30 hours a week at all jobs)
   3__I work seasonal or temporary jobs
   4__I am self-employed or freelance
   5__I am not currently employed (whether due to retirement, work as a homemaker, disability, unemployment, or any other status outside the paid workforce)

2. How long does it take you to get to work on an average day? **Column C**
   1__I work at home
   2__Less than 15 minutes
   3__15-29 minutes
   4__30-59 minutes
   5__60+ minutes
   9__I am not currently employed

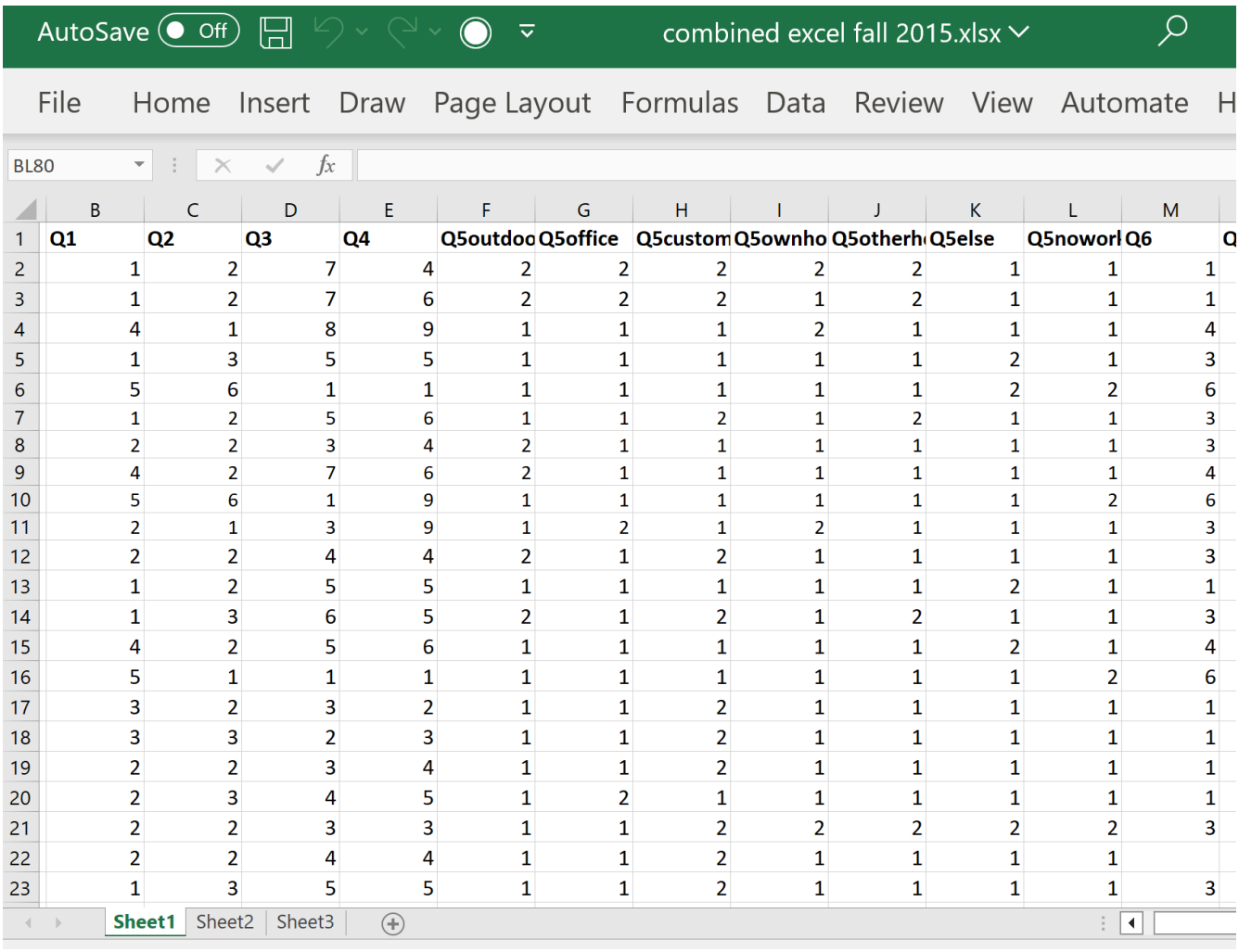
3. In an average week, how many hours do you work for pay? **Column D**
   1__I don't work for pay in an average week
   2__1-10 hours
   3__11-20 hours
   4__21-30 hours
   5__31-40 hours
   6__41-50 hours
   7__51 or more hours
   9__It varies too much for me to be able to say

4. In an average week, on how many days do you work for pay? **Column E**
   0__None
   1__1
   2__2
   3__3
   4__4
   5__5
   6__6
   7__7
   9__It varies too much for me to be able to say

Figure 1. An Example of a Codebook

automatically convert multiple-choice questions to the type of numeric answers statistical software requires.

Figure 2 shows what completed data entry might look like; it is taken from the same survey and shows the data after student survey-takers entered it into Excel. Each survey response, coded text, or other unit of analysis in the quantitative project has its data entered on a particular row. Note that without the codebook, it is not possible to understand the data displayed on the screen. When researchers perform data analysis directly in spreadsheet software, they may need to rely on the codebook to convert data back and forth from machine-readable (numerical) codes to human-language response categories. However, when data is imported into statistical analysis software, codebook information can be entered directly into the software, as will be discussed in the chapter Quantitative Analysis with SPSS: Data Management.



	B	C	D	E	F	G	H	I	J	K	L	M
1	Q1	Q2	Q3	Q4	Q5outdoo	Q5office	Q5custom	Q5ownho	Q5otherh	Q5else	Q5nowork	Q6
2	1	2	7	4	2	2	2	2	2	1	1	1
3	1	2	7	6	2	2	2	1	2	1	1	1
4	4	1	8	9	1	1	1	2	1	1	1	4
5	1	3	5	5	1	1	1	1	1	2	1	3
6	5	6	1	1	1	1	1	1	1	2	2	6
7	1	2	5	6	1	1	2	1	2	1	1	3
8	2	2	3	4	2	1	1	1	1	1	1	3
9	4	2	7	6	2	1	1	1	1	1	1	4
10	5	6	1	9	1	1	1	1	1	1	2	6
11	2	1	3	9	1	2	1	2	1	1	1	3
12	2	2	4	4	2	1	2	1	1	1	1	3
13	1	2	5	5	1	1	1	1	1	2	1	1
14	1	3	6	5	2	1	2	1	2	1	1	3
15	4	2	5	6	1	1	1	1	1	2	1	4
16	5	1	1	1	1	1	1	1	1	1	2	6
17	3	2	3	2	1	1	2	1	1	1	1	1
18	3	3	2	3	1	1	2	1	1	1	1	1
19	2	2	3	4	1	1	2	1	1	1	1	1
20	2	3	4	5	1	2	1	1	1	1	1	1
21	2	2	3	3	1	1	2	2	2	2	2	3
22	2	2	4	4	1	1	2	1	1	1	1	
23	1	3	5	5	1	1	2	1	1	1	1	3

Figure 2. Survey Data After Entry Into Excel

When obtaining data from elsewhere, many sites will provide the option of downloading data in a variety of file formats. In that case, researchers should choose—if possible—the appropriate file format for the software they are using, and should also download any codebook, readme, or help files that will explain the data and coding. Sometimes data is not available in a given file format and will need to be converted or imported, which will be discussed in the chapter Quantitative Analysis with SPSS: Data Management.

Note that most statistical analysis software is not cloud-resident, so it is important that researchers save their datasets after creating, importing, or modifying them; keep good backups; and keep records of all tests and procedures run, modifications made, etc. during the data analysis process.

Data Cleaning & Working With Data

Aside from preparing data for analysis, the other crucial step researchers need to take prior to beginning their analysis is data cleaning. Data cleaning is the process of examining data to find any errors, mistakes, duplications, corruptions, omissions, or other issues. Where possible, researchers can correct these issues; in other cases, certain data may need to be omitted from analysis.

Researchers may also need to modify variables or datasets in various ways. For example, many studies involve the creation of an **index variable**, or a composite measure created by combining information from multiple variables. For example, a study might involve administering a self-esteem inventory consisting of a number of different multiple-choice questions getting at various elements of self-esteem. Then, researchers combine the answers to all of these questions using a scoring system to create one variable representing the score on the self-esteem index. In other cases, researchers need to reduce the number of response categories a variable has or convert a **continuous** variable into an **ordinal** variable. Or a researcher might be working with a dataset that includes respondents of all ages, but for a study only interested in 18-29 year olds, and thus may need to filter the dataset. As one final example, researchers may have data from the same study stored in multiple spreadsheets and may need to combine or merge that data.

These are only a few examples of the tasks researchers face. The practical how-to of carrying out these tasks will be discussed in the chapter Quantitative Analysis with SPSS: Data Management — but before trying to carry them out, researchers need to take the time to think through their projects, determine which steps are necessary, and plan carefully.

Exercises

1. Write five basic multiple-choice survey questions (they do not have to be anything fancy–consider asking questions like age and favorite color). Create a codebook for your survey. Then, ask ten people you know to answer the questions, without using survey software. Finally, enter the data into Excel or another spreadsheet program of your choice, following your codebook.
2. Choose one of the data sources noted at the top of this chapter. Visit the website for the data source and learn as much as you can about it, then write a paragraph summarizing how the data is collected and what the data focuses on.

Media Attributions

- codebook-example-1 © Mikaila Mariel Lemonik Arthur, in conjunction with Fall 2015 students in Sociology 302 at Rhode Island College.
- survey data entry © Mikaila Mariel Lemonik Arthur, in conjunction with Fall 2015 students in Sociology 302 at Rhode Island College.

3. Univariate Analysis

ROGER CLARK

Univariate Analyses in Context

This chapter will introduce you to some of the ways researchers use statistics to organize their presentation of individual variables. In Exercise 1 of Introducing Social Data Analysis, you looked at one variable from the General Social Survey (GSS), "sex" or gender, and found that about 54 percent of respondents over the years have been female while about 46 percent have been male. You in fact did an analysis of one variable, sex or gender, and hence did an elementary univariate analysis.

Before we go further into your introduction to univariate analyses, we'd like to provide a somewhat larger context for it. In doing so, we begin with a number of distinctions. One distinction has to do with the number of variables that are involved in an individual analysis. In this book you'll be exposed to three kinds of analysis: *univariate*, *bivariate* and *multivariate* analyses. **Univariate analyses** are ones that tell us something about one variable. You did one of these when you discovered that there have been more female than male respondents to the GSS over the years. **Bivariate analyses**, on the other hand, are analyses that focus on the relationship between two variables. We have just used the GSS source we guided you to (Thomas 2020/2021) to discover that over the years men have been much more likely to work full time than women—roughly 63 percent of male respondents have done so since 1972, while only about 40 percent of female respondents have. This finding results from a bivariate analysis of two variables: gender and work status. **Multivariate analyses**, then, are ones that permit the examination of the relationship between two variables while investigating the role of other variables as well. Thus, for instance, when we look at the relationship between gender and work status for White Americans and Black Americans separately, we are involving a third variable: race. For White Americans, the GSS tells us, about 63 percent of males have held full time jobs over time, while only about 39 percent of females have done so. For Black Americans, the difference is smaller: 56 percent of males have worked full time, while 44 percent of females have done so. We thus did a multivariate analysis, in which we examined the relationship between gender and work status, while also examining the effect of race on that relationship.

Another important distinction is between **descriptive** and **inferential statistics**. This distinction calls into play another: that between *samples* and *populations*. Many times researchers will use data that have been collected from a **sample** of subjects from a larger

population. A *population* is a group of cases about which researchers want to learn something. These cases don't have to be people; they could be organizations, localities, fire or police departments, or countries. But in case of the GSS, the population of interest is in fact people: all adults in the United States. Very often, it is impractical or undesirable for researchers to gather information about every subject in the population. You can imagine how much time and money it would cost for those who run the GSS, for instance, to contact every adult in the country. So what researchers settle for is information from samples of the larger population. A *sample* is a number of cases drawn from a larger population. In 2018, for instance, the organization that runs the GSS collected information on just over 2300 adult Americans.

Now we can address the distinction between descriptive and inferential statistics. *Descriptive statistics* are statistics used to describe a sample. When we learned, for instance, that the GSS reveals that about 63 percent of male respondents worked full time, while about 40 percent of female respondents worked full time, we were getting a description of the sample of adult Americans who had ever participated in the GSS. (And you'd be right if you added that this is a case of bivariate descriptive statistics, since the percentages describe the relationship between two variables in the sample—gender and work status. You're so smart!) *Inferential statistics*, on the other hand, are statistics that permit researchers to make inferences about the larger populations from which the sample was drawn. Without going into too much detail here about the requirements for using inferential statistics[1] or how they are calculated, we can tell you that our analysis generated statistics that suggested we'd be on solid ground if we inferred from our sample data that a relationship between gender and work status not only exists in the sample, but also in the larger population of American adults from which the sample was drawn.

In this chapter we will learn something about both univariate descriptive statistics (statistics that describe single variables in a sample) and univariate inferential statistics (statistics that permit inferences about those variables in the larger population from which the sample was drawn).

Levels of Measurement of Variables

Now we can get down to basics. We've been throwing around the term **variable** as if it were second nature to you. (If it is, that's great. If not, here we go.) A *variable* is a characteristic that can vary from one subject or case to another or for one case over time. In the case of

1. Note that you can use many statistical methods to analyze data about populations, there are some differences in how they are employed, as will be discussed later in this chapter.

the GSS data we've presented so far, one variable characteristic has been gender or sex. A human adult responding to the GSS may indicate that they are male or female. (They could also identify with other genders, of course, but the GSS hasn't permitted this so far.) Gender is a variable because it is a characteristic that can vary from one human to another. If we were studying countries, one variable characteristic that might be of interest is the size of the population. Variables, we said, can also vary from one subject over time. Thus, for instance, your age is in one category today, but will be in another next year and in yet another in two years.

The nature of the kinds of categories is crucial to the understanding of the kinds of statistical analysis that can be applied to them. Statisticians refer to these "kinds" of categories as **levels of measurement**. There are four such levels or kinds of variables: *nominal level variables*, *ordinal level variables*, *interval level variables*, and *ratio level variables*. And, as you'll see, the term "level" of measurement makes sense because each level requires that an additional criterion is met for distinguishing it from the previous "level." The most basic level of measurement is that of the **nominal level variable**, or a variable whose categories have names. (The word "nominal" has the Latin root *nomen*, or name.) We say the nominal level is the most basic because every variable is at least a nominal variable. The variable "gender," when it has the two categories, male and female, has categories that have names and is therefore nominal. So is "religion," when it has categories like Protestant, Catholic, Jew, Muslim, and other. But so does the variable "age," when it has categories from 1 and 2 to, potentially, infinity. Each one of categories (1,2,3, etc.) has a name, even though the name is a number. In other words, again, every variable is a nominal level variable. There are some nominal level variables that have the special property of only consisting of two categories, like yes and no or true and false. These variables are called **binary** variables (also known as **dichotomous** variables).

To be an **ordinal level variable**, a variable must have categories can be ordered in some sensible way. (The word "ordinal" has the Latin root *ordinalis*, or order.) Said another way, an ordinal level variable is a variable whose categories have names *and* whose categories can be ordered in some sensible way. An example would be the variable "height," when the categories are "tall," "medium," and "short." Clearly these categories have names (tall, medium and short), but they also can be ordered: tall implies more height than medium, which, in turn, implies more height than short. The variable "gender," would not qualify as an ordinal level variable, unless one were an inveterate sexist, thinking that one gender is somehow a superior category to the others. Both nominal and ordinal level variables can be called **discrete variables**, which means they are variables measured using categories rather than numbers.

To be an **interval level variable**, a variable must be made up of adjacent categories that are a standard distance from one another, typically as measured numerically. Fahrenheit temperatures constitute an interval level variable because the difference between 78 and

79 degrees (1 degree) is seen as the same as the difference between 45 and 46 degrees. But because all those categories (78 degrees, etc.) are named and can be ordered sensibly, it's pretty easy to see that all interval level variables could be measured at the ordinal level—even while not all nominal and ordinal level variables could be measured at the interval level.

Finally, we come to **ratio level variables**. Ratio variables are like interval level variables, but with the addition of an absolute zero, a category that indicates the absence of the phenomenon in question. And while some interval level variables cannot be multiplied and divided, ratio level variables can be. Age is an example of a ratio variable because the category, zero, indicates a person or thing has no age at all (while, in contrast, "year of birth" in the calendar system used in the United States does not have an absolute zero, because the year zero is not the absence of any years). But, while interval and ratio variables can be distinguished from each other, we are going to assert that, for the purposes of this book, they are so similar that the distinction isn't worth insisting upon. As a result, for practical purposes, we could be calling all interval and ratio variables, interval-ratio variables, or simply interval variables. Both ratio and interval level variables can also be referred to as **scale** or **continuous** variables, as their (numerical) categories can be placed on a continuous scale.

But what are those practical purposes for which we need to know a variable's level of measurement? Let's just see . . .

Measures of Central Tendency

Roger likes to say, "All statistics are designed with particular levels of measurement in mind." What's this mean?[2] Perhaps the easiest way to illustrate is to refer to what statisticians call "**measures of central tendency**" or what we laypersons call "averages." You may have already learned about three of these averages before: the *mean*, the *median*, and the *mode*. But have you asked yourself why we need three measures of central tendency or average?

The answer lies in the level of measure required by each kind of average. The **mean** (which is what people most typically refer to when they use the term "average"), you may recall, is the sum of all the categories (or values) in your sample divided by the number of such categories (or values). Now, stop and think: what level of measurement (nominal, ordinal or interval) is required for you to calculate a mean?

If your answer was "interval," you should give yourself a pat on the back.[3] You need a vari-

2. Besides the fact that he's getting increasingly senile?
3. Something that's increasingly difficult for Roger to do as he gets up in years.

able whose categories may legitimately be added to one another in order to calculate a mean. You could do this with the variable "age," whose categories were 0, 1, 2, 3, etc. But you couldn't, say, with "height," if the only categories available to you were tall, medium, and short (if you had actual height in inches or centimeters, of course, that would be a different story).

But if your variable of interest were like that height variable—i.e., an ordinal level variable, statisticians have cooked up another "average" or measure of central tendency just for you: the **median**. The median is the middle category (or value) when all categories (or values) in the sample are arranged in order. Let's say your five subjects had heights that were classified as tall, short, tall, medium and tall. If you wanted to calculate the median, you'd first arrange these in order as, for instance, short, medium, tall, tall and tall. You'd then pick the one in the middle—i.e., tall—and that would be your median. Now, stop and think: could you calculate the median of an interval level variable, like the age variable we just talked about?

If your answer was "yes," you should give yourself a hardy slap on the knee.[4] The median can be used to analyze an interval level variable, as well as ordinal level variables, because all interval level variables are also ordinal. Right?

OK, you say, the mean has been designed to summarize interval level variables and the median has been fashioned to handle ordinal level variables. "I'll bet," you say, "the mode is for analyzing nominal level variables." And you're right! The **mode** is the category of a variable in a sample that occurs most frequently. This can be calculated for nominal level variables because nominal level variables, whatever else they have, have categories (with names). Let's say the four cars you were studying had the colors of blue, red, green and blue. The mode would be blue, because it's the category of colors that occurs most frequently. Before you take these averages out for a spin, we'd like you to try another question. Can a mode be calculated on an ordinal or an interval level variable?

If you answer "yes," you should be very proud. Because you've probably seen that ordinal and interval variables could also be treated like nominal level variables and therefore can have modes. (That is, categories that occur most frequently). Note, though, that the mode is unlikely to be a helpful measure in instances where continuous variables have many possible numerical values, like annual income in dollars, because in these cases the mode might just be some dollar amount made by three people in a sample where everyone else's income is unique.

Your Test Drive

4. Unless you've got arthritis there like you know who.

Examine the following sample data for five students (A through E). Calculate as many of the measures of central tendency (or average) as you can for each of the three variables: religion, height and age. (See this footnote[5] for the correct answer once you're done.)

Student	A	B	C	D	E
Religion	Catholic	Protestant	Jewish	Catholic	Catholic
Height	Tall	Short	Medium	Short	Short
Age	19	20	19	21	19

How do you know which measure of central tendency or average (mode, median or mean) to use to describe a given variable in a report? The first rule is a negative: do NOT report a measure that is not suitable for your variable's level of measurement. Thus, you shouldn't report a mean for the religion or height variables in the "test drive" above, because neither of them is an interval level variable.

You might well ask, "How could I possibly report a mean religion, given the data above?" This is a good question and leads us to mention, in passing, that when researchers set up computer files to help them analyze data, they will almost always code variable categories using numbers so that the computer can recognize them more easily. **Coding** is the process of assigning observations to categories—and, for computer usage, this often means changing the names of variables categories to numbers. Perhaps you recall doing Exercise 1 at the end of Introducing Social Data Analysis—the one that asked you to determine the percentage of respondents who were female over the years (about 54 percent). Well, to set up the computer to do this analysis, the folks who created the file (and who supplied us with the data) coded males as 1 and females as 2. So the computer was faced with over 34,000 1s and 2s rather than with over 34,000 "males" and "females." Computers like this kind of help. But computers, while very good at computing, are often a little stupid when it comes to interpreting their computations.[6] So when I went in and asked the computer to add just a few more statistics, including the mean, median and mode, about the sex or gender of GSS respondents, it produced this table. (Don't worry, I'll show you how to produce a table like this in Exercise 3 of this chapter.)

5. The mode of religion is Catholic. No other average is applicable. The median of height is short, and so is the mode. The mean of height can't be calculated. The mean height is 19.6. Its median is 19, as is its mode.
6. No offense to you, my faithful laptop, without which I couldn't bring you, my readers, this cautionary tale.

Table 1: Univariate Statistics Associated with "Sex" in the GSS

Summary Statistics					
Mean =	1.54	Std Dev =	.50	Coef var =	.32
Median =	2.00	Variance =	.25	Min =	1.00
Mode =	2.00	Skewness =	-.17	Max =	2.00
Sum =	99,993.48	Kurtosis =	-1.97	Range =	1.00

What this table effectively and quickly, tells us is that the mode of "sex" (really gender) is 2, meaning "female." Part of your job as a social data analyst is to translate codes like this back into English—and report that the mode, here, is "female," not "2". But another important part, and something the computer also cannot do, is recognizing the level of measure of the variable concerned—in this case, nominal—and realize which of the reported statistics is relevant given that level. And in terms of "sex," as reported in Table 1, only you can know how silly it would be to report that the mean "sex" is 1.54 (notice the computer can't see that silliness) or that its median is 2.00. When Roger was little,[7] Smoky the Bear used to tell kids "Only YOU can prevent forest fires." But Roger is here to tell you, "Only YOU can prevent statistical reporting travesties." So, again, you do not want to report statistics that aren't designed for the level of measure of your variables.

In general, though, when you ARE dealing with an interval variable, like age in years, you really have three choices about which to report: the mean, the median and the mode. For the moment, we're going to recommend that, in such case, you might consider that the reading public is likely to be most familiar with the mean and, for that reason, you might report the mean. (We'll get to qualifications of that recommendation a little later.)

Variation

Measures of central tendency are often useful for summarizing variables, but they can sometimes be misleading. Roger just[8] Googled the average life expectancy for men in the United States and discovered it was about 76.5 years. (Pretty clearly a mean, not a mode

7. Many years ago.
8. In 2020.

or median, right?) At this sitting, he is about 71.5 years old. Does this mean he has exactly 5 years left of life to live? Well, probably not. Given his health, educational level, etc., he's likely to live considerably longer…unless COVID-19 gets him tomorrow. The point is that for life expectancy, as for other variables, there's variation around the average. And sometimes knowing something about that variation is at least as important as the average itself—sometimes more important.

We can learn a lot about a variable, for instance, simply by showing how its cases are distributed over its categories in a sample. Exercise 1 at the end of Introducing Social Data Analysis actually told you the modal gender of respondents to the GSS survey. ("Modal" is the adjectival form of mode.) Do you recall what that was? It was "female," right? What this tells you is that the "average" respondent over the years has been a female. But the mode, being what it is, doesn't tell you whether 100 percent of respondents were female or 50.1 percent were female. And that's an important difference.

One of the most commonly used ways of showing variation is what's called a **frequency distribution**. A frequency distribution shows the number of times cases fall into each category in a sample. I've just called up the table you looked at in Exercise 1 of Introducing Social Data Analysis and plunked it down here as Table 2. What this table shows is that while about 35,179 females had participated in the GSS since 1972, 29,635 males had done so as well. The table further tells us that while about 54 percent of the sample is female, about 46 percent has been male. The distribution has been much closer to 50-50 than 100-0. And this extra information about the variable is a significant addition to the fact that modal "sex" was female.

Table 2. The Frequency Distribution Associated with "Sex" in the GSS as of 2018

Frequency Distribution		
Cells contain: –**Column percent** -Weighted N		**Distribution**
SEX	1: MALE	**45.7** 29,635.4
	2: FEMALE	**54.3** 35,179.1
	COL TOTAL	***100.0*** *64,814.4*

"Sex" is a nominal level variable, and frequency distributions have been designed for displaying the variation of nominal level variables. But, of course, because ordinal and interval variables are also nominal level variables, frequency distributions can be used to describe their variation as well. And this often makes sense with ordinal level variables. Thus, for instance, we used a frequency distribution of respondents' confidence in the military ("conarm") to show that there was relatively little variation in Americans' confidence in that institution in 2018 (Table 3, below). Almost 61 percent of respondents said they had a "great deal of confidence" in the military that year, while only about 39 percent said they had "only some" or "hardly any" confidence. In other words, at least in comparison with the variation in "sex," variation in confidence in the military, which, after all, has three categories, seems limited. In other words, this kind of confidence seems more concentrated in one category ("great deal of confidence") than you might expect.

Quiz at the End of the Paragraph: Can you see what the median and the mode of confidence in the military was?

Bonus Trick Question: What was its mean?

Table 3. The Frequency Distribution and Other Statistics Related to Americans' Confidence in the Military, 2018 General Social Survey Data

Frequency Distribution		
Cells contain: –**Column percent** -Weighted N		**Distribution**
CONARMY (Confidence in the U.S. Military)	1: A GREAT DEAL	**60.6** 940.7
	2: ONLY SOME	**32.5** 504.3
	3: HARDLY ANY	**7.0** 108.0
	COL TOTAL	***100.0*** *1,553.0*

Summary Statistics					
Mean =	1.46	Std Dev =	.62	Coef var =	.43
Median =	1.00	Variance =	.39	Min =	1.00
Mode =	1.00	Skewness =	1.00	Max =	3.00
Sum =	2,273.32	Kurtosis =	-.06	Range =	2.00

Measures of Variation for Interval Level Variables

Looking at frequency distributions is a pretty good way of getting a sense of the variation in nominal and ordinal variables. But it would be a fairly awkward way of doing so for interval variables, many of which, if you think about it, would have many categories. (Can you imagine a frequency distribution for the variable "age" of respondents in the GSS?) Statisticians have actually given us some pretty elegant ways of dealing with the description of variation in interval variables and we'd now like to illustrate them with simple examples.

Roger's daughter, Wendy, was a day care provider for several years and could report that variation in the ages of preschool children made a tremendous difference in the kinds of things you can do with them. Imagine, if you will, that you had two groups of four preschool children, one of which had four 3-year-olds in it and one of which had two 5-year-olds and two 1-year-olds. Can you calculate the mean age of each group?

If you found that the mean age of both groups was 3 years old, you did a fine job. Now, if you were inclined to think that any two groups with the same mean age were likely to be similar, think of these two from a day care provider's point of view. Figuring out what to do for a day with two 1-year-olds and two 5-year-olds would be a much more daunting task than planning for four 3-year-olds. Wouldn't it?

Statisticians have given us one particularly simple measure of spread or variation for interval level variables: the **range**. The range is simply the highest category in your sample minus the lowest category. For the group with four 3-year-olds, the range would be (3-3=) zero years. There is no variation in age for this group. For the group with two 1-year-olds and two 5-year-olds, the range would be (5-1=) four years. A substantial, and important difference, again especially if you, like my daughter, were a day care provider. Means don't always tell the whole story, do they?

Perhaps the more commonly used statistic for describing the variation or spread of an interval level variable, however, is the **standard deviation**. The range only gives you a sense of how spread out the extreme values or categories are in your sample. The **standard devi-**

ation is a measure of variation that takes into account every value's distance from the sample mean. The usefulness of such a measure can be illustrated with another simple example. Imagine, for instance, that your two groups of preschool children had the following ages: 1, 1, 5, 5, on the one hand, and 1, 3, 3, and 5, on the other.

The mean of these two groups is 3 years and the range is 4 years. But are they identical? No. You may notice that each of the individual ages in the first group is a "distance" of 2 away from the mean of 3. (The two 1s are each 2 away from 3 and the two 5s are also 2 away from 3.) So the average "distance" of each age from the mean is 2 for group 1. But that's not true for the second group. The 1 and the 5 are both 2 away from the mean of 3, but the two 3s are both no distance away. So the average distance of ages from the mean in this group is something less than 2. Hence, the average distance of ages from the mean in the first group is larger than the average distance in the second group. The standard deviation is a way of capturing a difference like this—one that is not captured by the range.

It does this by using a formula that essentially adds the individual "distances" of categories or values from the mean and then divides that number by the categories. We think of it as being very similar to the computation of the mean itself: a sum divided by the number of cases involved. The computational formula is:

$$SD_{sample} = \sqrt{\frac{\sum_{i=1}^{N}(x - \overline{x})^2}{N - 1}}$$

where SD_{sample} stands for the standard deviation

$\sqrt{\ }$ stands for the square root of the entire expression that follows

$\sum_{i=1}^{N}$ means to add up the sequence of numbers produced by the expression that follows

x stands for each value of category in the sample

$\overline{x}$ stands for the sample mean

N stands for the number of sample cases

The formula may look daunting, but it's not very difficult to compute with just a few cases—and we'll never ask you to use anything other than a computer to compute the standard deviation with more cases. Note that to calculate the standard deviation for an entire population, rather than a sample, we use N rather than N-1 in the denominator. And also note that the numerator—$Var(X) = \sum_{i=1}^{N}(x - \overline{x})^2$—is referred to as the *variance*.

Notice first that the formula asks you to compute the sample mean. For the second sample of ages above—the one with ages 1, 3, 3, 5—the mean is 3. It then asks you to take the difference between each category in the sample and the mean and square the differences.

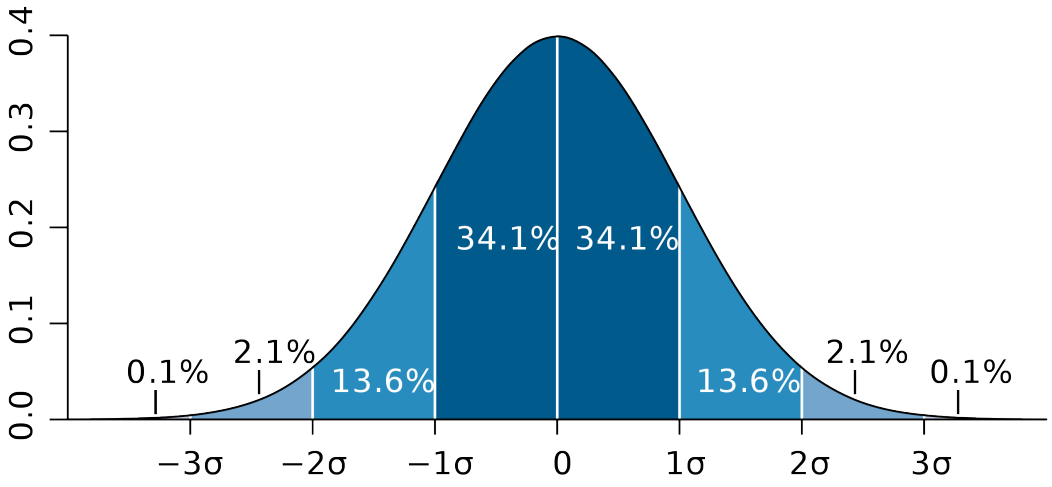
1-3, for instance, is -2 and its square is 4. 3-3 is 0 and its square is 0. And 5-3 is 2 and its square is 4. The formula then asks you to add these squared values up: 4 0 0 4=8. Then it says to divide by the number of cases, minus 1: 3. 8/3=2.67. It then asks you to take the square root of 2, or about 1.6. So the standard deviation of this sample is about 1.6 years.

> Can you calculate the standard deviation for the second sample of ages above: 1, 1, 5, 5?
>
> Did you get 2.3? If so, give yourself another pat on the back.[9]

Measures of Deviation from the Normal Distribution

We've suggested that, other things being equal, the mean is a good way of describing the central tendency or average of an interval level variable. But other things aren't always equal. The mean is an excellent measure of central tendency, for instance, when the interval level variable conforms to what is called a **normal distribution.** A normal distribution of a variable is one that is symmetrical and bell-shaped (otherwise called a **bell curve**), like the one in Figure 2.1. This image suggests what is true when the distribution of a variable is normally distributed: that 68 percent of cases fall within one standard deviation on either side of the mean; that 95 percent of the cases fall within two standard deviations on either side; and that 99.7 percent of the cases fall within three standard deviations on either side. Note that the symbol σ is used to indicate standard deviation in many statistical contexts.

Figure 2.1 The Normal Curve



9. You already know Roger can't do this for himself.

One example that is frequently cited as a normally distributed variable is height. For American men, the average height in 2020 is about 69 inches,[10] where "average" here refers to the mean, the median and the mode, if, in fact, height is normally distributed. The peak of the curve (can you see it in your mind?) would be at 69 inches, which would be the most frequently occurring category, the one in the middle of the distribution of categories and the arithmetic mean.

But what happens when a variable is not normally distributed? We asked the Social Data Archive to use GSS data from 2010 to tell us what distribution of the number of children respondents had looked like, and we got these results (see Table 4):

Table 4. Number of Children Reported by General Social Survey Respondents (2010)

Summary Statistics					
Mean =	1.91	Std Dev =	1.73	Coef var =	.91
Median =	2.00	Variance =	2.99	Min =	.00
Mode =	.00	Skewness =	1.05	Max =	8.00
Sum =	3,894.30	Kurtosis =	1.39	Range =	8.00

As you might have expected, the greatest number of respondents said they had zero, one or two children. But, then the number of children tails off pretty quickly as you get into categories that represent respondents with 3 or more children. This variable, then, is not normally distributed. Most of the cases are concentrated in the lowest categories. When an interval level variable looks that this, it is said to have right, or positive skewness, and this is reflected in the report that "number of children" has a skewness of positive 1.05. **Skewness** refers to an asymmetry in a distribution in which a curve is distorted either to the left or the right. The skewness statistic can take on values from negative infinity to positive infinity, with positive values indicating right skewness (with "tails" to the right) and negative values indicating left skewness (when "tails" are to the left). A skewness statistic of zero would indicate that a variable is perfectly symmetrical.

Our rule of thumb is that when the skewness statistic gets near to 1 or near -1, the variable has more than enough skewness (either to the right or to the left) to be disqualified as a normally distributed variable. And in such cases, it's probably useful to report both the mean and the median as measures of central tendency, since the relationship of the two

10. 5 feet, 9 inches

will give some idea to readers of the nature of the variable's skewness. If the median is greater than the mean (as it is in the case of "number of children"), it's a sign that the author means to convey that the variable is right skewed. If it's less than the mean, the implication is that it's left skewed.

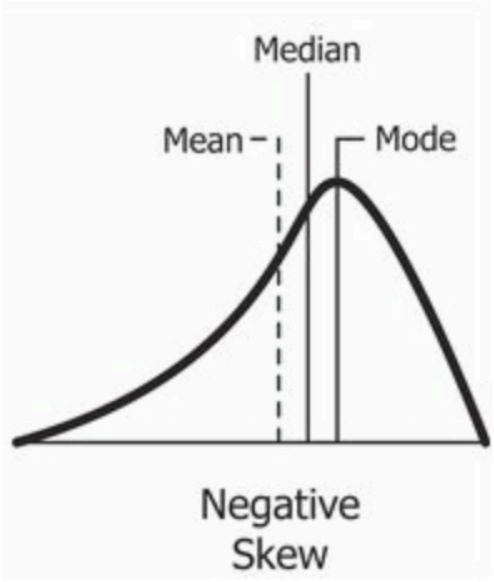


Figure 2.2 Negative Skew

Kurtosis refers to how sharp the peak of a frequency distribution is. If the peak is too pointed to be a normal curve, it is said to have positive kurtosis (or "**leptokurtosis**"). The kurtosis statistic of "number of children" is 1.39, indicating that the variable's distribution has positive kurtosis (or leptokurtosis). If the peak of a distribution is too flat to be normally distributed, it is said to have negative kurtosis (or **platykurtosis**), as seen in Figure 2.3.

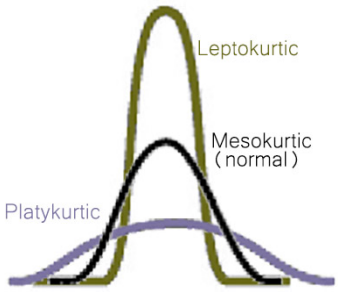


Figure 2.3. Kurtosis

A rule of thumb for the kurtosis statistic: if it gets near to 1 or near -1, the variable has more

than enough kurtosis (either positive or negative) to be disqualified as a normally distributed variable.

For a fascinating, personal lecture about the importance of being wary about reports using only measures of central tendency or average (e.g., means and medians), however, we encourage you to listen to the following talk by Stephen Jay Gould:

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://pressbooks.ric.edu/socialdataanalysis/?p=99#oembed-1

A Word About Univariate Inferential Statistics

Up to this point, we've only talked about univariate descriptive statistics, or statistics that describe one variable in a sample. When we learned that 54 percent of GSS respondents over the years have been women, we were simply learning about the (large) sample of people who have responded to the GSS over the years. And when we learned that the mean number of children that respondents had in 2010 was about 1.9 and the median was 2.0, those too were descriptions of the sample that year. One of the purposes of sampling, though, is that it can provide us some insight into the population from which the sample was drawn. In order to make inferences about such populations from sample data we need to use inferential statistics. *Inferential statistics*, as we said before, are statistics that permit researchers to make inferences about the larger population from which a sample is drawn.

We'll be spending more time on inferential statistics in other chapters, but now we'd like to introduce you a statistical concept that frequently comes up in relation to political polls: *margin of error*. To appreciate the concept of the margin of error, we need to understand the difference between two important concepts: *statistics* and *parameters*. A **statistic** is a description of a variable (or the relationship between variables) in a sample. The mean, median, mode, range, standard deviation and skewness are all types of statistics. A **parameter**, on the other hand, is a description of a variable (or the relationship between variables) in a population; many (but not all) of the same tools used as statistics when analyzing data from samples can be used as parameters when analyzing data on populations. A **margin of error**, then, is a suggestion of how far away from the actual population parameter a statistic is likely to be. Thus political polling can tell you precisely what percentage of the sample say they are going to vote for a candidate, but it can't tell you precisely what percentage would say the same thing in the larger population from which the sample was drawn.

BUT, when a sample is a *probability sample* of the larger population, we can estimate how close the population percentage is likely to be to the sample percentage. A full discussion of the different kinds of samples is beyond the scope of this book, but let's just say that a **probability sample** is one that has been drawn to give every member of the population a known (non-zero) chance of inclusion. Inferential statistics of all kinds assume that one is dealing with a probability sample of the larger population to which one would like to generalize (though, sometimes, inferential statistics are calculated even when this fundamental assumption of inferential statistics has not been met).[11]

Most frequently, a margin of error is a statement of the range around the sample percentage in which there is a 95 percent chance that the population percentage will fall. The pre-election polls before the 2016 election are frequently criticized for how badly they got it wrong when they predicted Hillary Clinton would get a higher percentage of the vote than Donald Trump—and win the election. But in fact most of the national polls came remarkably close to predicting the election outcome perfectly. Thus, for instance, an ABC News/Washington Post poll, collected between November 3rd and November 6th (two days before the election), and involving a sample of 2,220, predicted that Clinton would get 49 percent of the vote, plus or minus 2.5 percentage points (meaning that she'd likely get somewhere between 46.5 percent and 51.5 percent of the vote), and that Trump would get 46 percent, plus or minus 2.5 percentage points (meaning that he'd likely get somewhere between 43.5 percent and 48.5 percent of the vote). The margin of error in this poll, then, was plus or minus 2.5 percentage points. And, in fact, Clinton won 48.5 percent of the actual vote (well within the margin of error) and Trump won 46.4 percent (again, well within the margin of error) (CNN Politics, 2020). This is just one poll that got the election precisely right with respect to the total vote (if not the crucial electoral vote) count in advance of the election.

We haven't shown you how to calculate a margin of error here but, as you'll see in Exercise 4 at the end of the chapter, they are not hard to get a computer to spit out. One thing to keep in mind is that the size of a margin of error is a function of the size of the sample: the larger the sample, the smaller the margin of error. In fact all inferences using inferential statistics become more accurate as the sample size increases.

So, welcome to the world of univariate statistics! Now let's try some exercises to see how they work.

Exercises

11. And we hope you'll always say "naughty, naughty," when you know this has been done.

1. Which of the measures of central tendency has been designed for nominal level variables? For ordinal level variables? For interval level variables? Why can all three measures be applied to interval level variables?
2. Which way of showing the variation of nominal and ordinal level variables have we examined in this chapter? What measures of variation for interval level variables have we encountered?
3. Return to the Social Data Archive we explored in Exercise 1 of Introducing Social Data Analysis. The data, again, are available at https://sda.berkeley.edu/ . Again, go down to the second full paragraph and click on the "SDA Archive" link you'll find there. Then scroll down to the section labeled "General Social Surveys" and click on the first link there: General Social Survey (GSS) Cumulative Datafile 1972-2018 release. Now type "religion" in the row box, hit "output options," click on "summary statistics," then click on "run the table." See if you can answer these questions:

    - What level of measurement best characterizes "religion"? What is this variable measuring?
    - What's the only measure of central tendency you can report for "religion"? Report this measure, in English, not as a number.
    - What's a good way you can describe "religion"'s variation? Describe its variation.

    Now type "happy" in the row box, hit "output options," click on "summary statistics," then click on "run the table." See if you can answer these questions:

    - What level of measurement best characterizes "happy"? What is this variable measuring?
    - What are the only measures of central tendency you can report for "happy"? Report these measures, in English, not as a number.
    - What's a good way you can use to describe "happy"'s variation? Describe its variation.

    Now type "age" in the row box, hit "output options," click on "summary statistics," then click on "run the table." See if you can answer these questions:

    - What level of measure best describes "age"? What is this variable measuring?
    - What are all the measures of central tendency you could report for "age"? Report these measures, in English, not simply as numbers.
    - What are two good statistics for describing "age"'s variation? Describe its variation.
    - Is it your sense that "age" is essentially normally distributed? Why or why not? (What statistics did you check for this?)

4. Return to the Social Data Archive. The data, again, are available at https://sda.berkeley.edu/ (You may have to copy and paste this address to request the website.) Again, go down to the second full paragraph and click on the "SDA Archive" link you'll find there. Then scroll down to the section labeled "American National Election Studies (ANES)" and hit on the first link there: American National Election Study (ANES) 2016. These data come from a survey done after the 2016 election. Type "Trumpvote" in the row, hit "output options," and hit "confidence intervals," then hit "run table." What percentage of respondents, after the election, said they had voted for Trump? What was the "95 percent" confidence interval for this percentage? Check the end of this chapter for

the actual percentage of the vote that Trump got. Does it fall within this interval?

Media Attributions

- Standard_deviation_diagram.svg © M. W. Toews is licensed under a CC BY (Attribution) license
- Negative Skew © Diva Dugar adapted by Roger Clark is licensed under a CC BY-SA (Attribution ShareAlike) license
- Kurtosis © Mikaila Mariel Lemonik Arthur

4. Bivariate Analyses: Crosstabulation

Crosstabulation

ROGER CLARK

In most research projects involving variables, researchers do indeed investigate the central tendency and variation of important variables, and such investigations can be very revealing. But the typical researcher, using quantitative data analysis, is interested in testing hypotheses or answering research questions that involve at least two variables. A **relationship** is said to exist between two variables when certain categories of one variable are associated, or go together, with, certain categories of the other variable. Thus, for example, one might expect that in any given sample of men and women (assume, for the purposes of this discussion, that the sample leaves out nonbinary folks), men would tend to be taller than women. If this turned out to be true, one would have shown that there is a relationship between gender and height.

But before we go further, we need to make a couple of distinctions. One crucial distinction is that between an **independent variable** and a **dependent variable**. An *independent variable* is a variable a researcher suspects may affect or influence another variable. A *dependent variable*, on the other hand, is a variable that a researcher suspects may be affected or influenced by (or *dependent upon*) another variable. In the example of the previous paragraph, gender is the variable that is expected to affect or influence height and is therefore the independent variable. Height is the variable that is expected to be affected or influenced by gender and is therefore the dependent variable. Any time one states an expected relationship between two (or more) variables, one is stating a **hypothesis**. The hypothesis stated in the second-to-last sentence of the previous paragraph is that men will tend to be taller than women. We can map two-variable hypotheses in the following way (Figure 3.1):

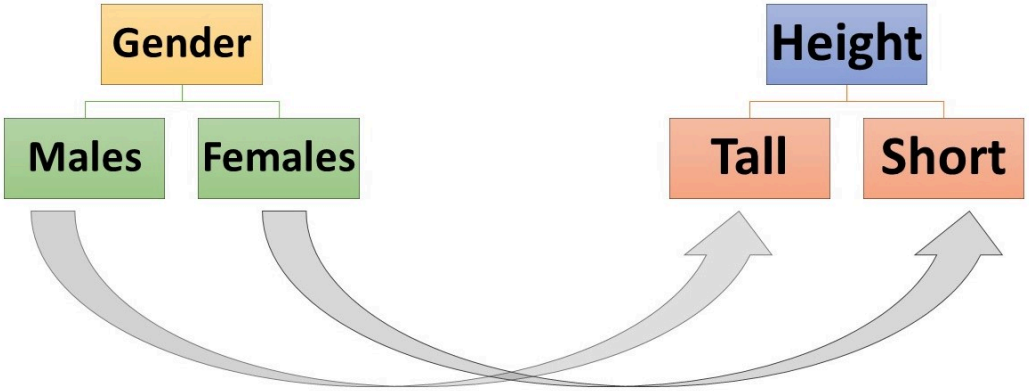


Figure 1. A Mapping of the Hypothesis That Men Will Tend To Be Taller Than Women

When mapping a hypothesis, we normally put the variable we think to be affecting the other variable on the left and the variable we expect to be affected on the right and then draw arrows between the categories of the first variable and the categories of the second that we expect to be connected.

Quiz at the End of The Paragraph

Read the following report by Annie Lowrey about a study done by two researchers, Kearney and Levine. What is the main hypothesis, or at least the main finding, of Kearney and Levine's study on the effects of Watching *16 and Pregnant* on adolescent women? How might you map this hypothesis (or finding)?

https://www.nytimes.com/2014/01/13/business/media/mtvs-16-and-pregnant-derided-by-some-may-resonate-as-a-cautionary-tale.html

We'd like to say a couple of things about what we think Kearney and Levine's major hypothesis was and then introduce you to a way you might analyze data collected to test the hypothesis. Kearney and Levine's basic hypothesis is that adolescent women who watched *16 and Pregnant* were less likely to become pregnant than women who did not watch it. They find some evidence not only to support this basic hypothesis but also to support the idea that the ones who watched the show were less likely to get pregnant because they were more likely to seek information about contraception (and presumably to use it) than others. Your map of the basic hypothesis, at least as it applied to individual adolescent women, might look like this:

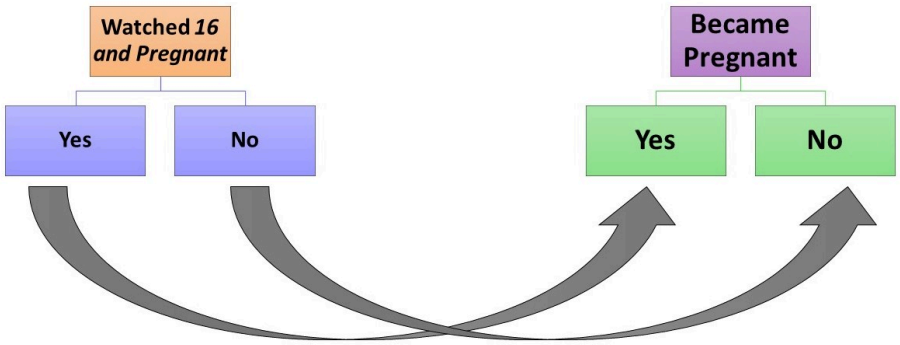


Figure 3.2: A Mapping of Kearney and Levine's Hypothesis

Let's look at a way of showing a relationship between two nominal level variables: *crosstabulation*. **Crosstabulation** is process of making a bivariate table for nominal level variables to show their relationship. But how does crosstabulation work?

Suppose you collected data from 8 adolescent women and the data looked like this:

Table 1: Data from Hypothetical Sample A

	Watched *16 and Pregnant*	Got Pregnant
Person 1	yes	no
Person 2	yes	no
Person 3	yes	no
Person 4	yes	yes
Person 5	no	yes
Person 6	no	yes
Person 7	no	yes
Person 8	no	no

Quick Check: What percentage of those who have watched *16 and Pregnant* in the sample have become pregnant? What percentage of those who have NOT watched *16 and Pregnant* have become pregnant?

If you found that 25 percent of those who had watched the show became pregnant, while 75 percent of those who had not watched it did so, you have essentially done a crosstabulation in your head. But here's how you can do it more formally and more generally.

First you need to take note of the number of categories in your independent variable (for "Watched *16 and Pregnant*" it was 2: Yes and No). Then note the number of categories in your dependent variable (for "Got Pregnant" it was also 2: again, Yes and No). Now you prepare a "2 by 2" table like the one in Table 3.2,[1] labeling the columns with the categories of the independent variables and the rows with the categories of the dependent variable. Then decide where the first case should be put, as we've done, by determining which cell is where its appropriate row and column "cross." We've "crosstabulated" Person 1's data by putting a mark in the box where the "Yes" for "watched" and the "No" for "got pregnant" cross.

Table 2. Crosstabulating Person 1's Data from Table 3.1 Above

		Watched *16 and Pregnant*	
		Yes	*No*
Got Pregnant	*Yes*		
	No	I	

We've "crosstabulated" the first case for you. Can you crosstabulate the other seven cases? We're going to call the cell in the upper left corner of the table cell "A," the one in the upper right, cell "B," the one in the lower left, cell "C," and the one in the lower right, cell "D." If you've finished your crosstabulation and had one case in cell A, 3 in cell B, 3 in cell C, and 1 in cell D, you've done great!

In order to interpret and understand the meaning of your crosstabulation, you need to take one more step, and that is converting those tally marks to percentages. To do this, you add up all the tally marks in each column, and then you determine what percentage of the

1. If one of your variables had three categories, it might be a "2 by 3" table. If both variables had 3 categories, you'd want a 3 by 3 table, etc.

column total is found in each cell in that column. You'll see what that looks like in Table 3 below.

Direction of the Relationship

Now, there are three characteristics of a crosstabulated relationship that researchers are often interested in: its *direction*, its *strength*, and its *generalizability***.** We'll define each of these in turn, and as we come to it. The **direction** of a relationship refers to how categories of the independent variable are related to categories of the dependent variable. There are two steps involved in working out the direction of a crosstabulated relationship... and these are almost indecipherable until you've seen it done:

1. Percentage in the direction of the independent variable.
2. Compare percentages in one category of the dependent variable.

The first step actually involves three substeps. First you change the tally marks to numbers. Thus, in the example above, cell A would get a 1, B, a 3, C, a 3, and D, a 1. Second, you'd add up all the numbers in each category of the independent variable and put the total on the side of the table at the end of that column. Third, you would calculate the percentage of that total that falls into each cell along that column (as noted above). Once you'd done all that with the data we gave you above, you should get a table that looks like this (Table 3.3):

Table 3 Crosstabulation of Our Imaginary Data from a *16 and Pregnant* Study

		Watched *16 and Pregnant*	
		Yes	*No*
Got Pregnant	*Yes*	1 (25%)	3 (75%)
	No	3 (75%)	1 (25%)
Total		4 (100%)	4 (100%)

Step 2 in determining the direction of a crosstabulated relationship involves comparing percentages in one category of the dependent variable. When we look at the "yes" category, we find that 25% of those who watched the show got pregnant, while 75% of those who did NOT watch the show got pregnant. Turning this from a percentage comparison to plain English, this crosstabulation would have shown us that those who *did* watch the show were less likely to get pregnant than those who did not. And **that** is the direction of the relationship.

Note: because we are designing our crosstabulations to have the independent variable in the columns, one of the simplest ways to look at the direction or nature of the relationship is to compare the percentages across the rows. Whenever you look at a crosstabulation, start

by making sure you know which is the independent and which is the dependent variable and comparing the percentages accordingly.

Strength of the Relationship

When we deal with the ***strength*** of a relationship, we're dealing with the question of how reliably we can predict a sample member's value or category of the dependent variable based on knowledge of that member's value or category on the independent variables, just knowing the direction of the relationship. Thus, for the table above, it's clear that if you knew that a person had watched *16 and Pregnant* and you guessed she'd not gotten pregnant, you'd have a 75% (3 out of 4) chance of being correct; if you knew she hadn't watched, and you guessed she had gotten pregnant, you'd have a 75% (3 out of 4) chance of being correct. Knowing the direction of this relationship would greatly improve your chances of making good guesses...but they wouldn't necessarily be perfect all the time.

There are several measures of the strength of association and, if they've been designed for nominal level variables, they all vary between 0 and 1. When one of the measures is 0.00, it indicates that knowing a value of the independent variable won't help you at all in guessing what a value of the dependent variable will be. When one of these measures is 1.00, it indicates that knowing a value of the independent variable and the direction of the relationship, you could make perfect guesses all the time. One of the simplest of these measures of strength, which can only be used when you have 2 categories in both the independent and dependent variables, is *the absolute value* of ***Yule's Q***. Because the "absolute value of Yule's Q" is so relatively easy to compute, we will be using it a lot from now on, and it is the one formula in this book we would like you to learn by heart. We will be referring to it simply as |Yule's Q|—note that the "|" symbols on both sides of the 'Yule's Q' are asking us to take whatever Yule's Q computes to be and turn it into a positive number (its absolute value). So here's the formula for Yule's Q:

$$|\text{Yule's Q}| = \frac{|(A \times B) - (B \times C)|}{|(A \times D) + (B \times C)|}$$

$$\text{Where}$$

$$A \text{ is the number of cases in cell } A$$
$$B \text{ is the number of cases in cell } B$$
$$C \text{ is the number of cases in cell } C$$
$$D \text{ is the number of cases in cell } D$$

For the crosstabulation of Table 3,

$$|\text{Yule's Q}| = \frac{|(1 \times 1) - (3 \times 3)|}{|(1 \times 1) + (3 \times 3)|} = \frac{|1 - 9|}{|1 + 9|} = \frac{|-8|}{|10|} = \frac{8}{10} = .80$$

In other words, the Yule's Q is .80, much close to the upper limit of Yule's Q (1.00) than it is to its lower limit (0.00). So the relationship is very strong, indicating, as we already knew,

that, given knowledge of the direction of the relationship, we could make a pretty good guess about what value on the dependent variable a case would have if we knew what value on the independent variable it had.

Practice Exercise

Suppose you took three samples of four adolescent women apiece and obtained the following data on the *16 and Pregnant* topic:

Sample 1		**Sample 2**		**Sample 3**	
Watched	*Pregnant*	*Watched*	*Pregnant*	*Watched*	*Pregnant*
Yes	No	Yes	No	Yes	Yes
Yes	No	Yes	Yes	Yes	Yes
No	Yes	No	Yes	No	No
No	Yes	No	No	No	No

See if you can determine both the *direction* and *strength* of the relationship between having watched "16 and Pregnant" in each of these imaginary samples. In what ways does each sample, other than sample size, differ from the Sample A above? Answers to be found in the footnote.[2]

Roger now wants to share with you a discovery he made after analyzing some data that two now post-graduate students of his, Angela Leonardo and Alyssa Pollard, have made using crosstabulation. At the time of this writing, they had just coded their first night of TV commercials, looking for the gender of the authoritative "voice-over"—the disembodied voice that tells viewers key stuff about the product. It's been generally found in gender studies that these voice-overs are overwhelmingly male (e.g., O'Donnell and O'Donnell 1978; Lovdal 1989; Bartsch *et al.* 2000), even though the percentage of such voice-overs that were male had dropped from just over 90 percent in the 1970s and 1980s to just over 70 percent in 1998. We will be looking at considerably more data, but so far things are so interesting that Roger wants to share them with you...and you're now sophisticated enough about

2. Answers: In Sample 1, the direction of the relationship is the same as it was in Sample A (those who watched the show were less likely than those who didn't), but its strength is greater (Yule's Q= 1.00, rather than 0.80). In Sample 2, there is no direction of the relationship (those who watched the show were just as likely to get pregnant as those who didn't) and its strength is as weak as it could be (Yule's = 0.00). In Sample 3, the direction of the relationship is the opposite of what it was in Sample A. In this case, those who watched the show were more likely to get pregnant than those who didn't. And the strength of the relationship was as strong as it could be (Yule's Q= 1.00).

crosstabs (shorthand for crosstabulations) to appreciate them. Thus, Table 3.4 suggests that things have changed a great deal. In fact the direction of the relationship between the time period of the commercials and the gender of the voice-over is clearly that more recent commercials are much more likely to have a female voice-over than older ones. While only 29 percent of commercials in 1998 had a female voice-over, 71 percent in 2020 did so. And a Yule's Q of .72 indicates that the relationship is very strong.

Table 3.4 Crosstabulation of Year of Commercial and Gender of the Voice-Over

		Year of Commercial	
		1998	*2020*
Gender of Voice-Over	*Male*	432 (71%)	14 (29%)
	Female	177 (29%)	35 (71%)
Notes: |Yule's Q| = 0.72; 1998 data from Bartsch et al., 2001.			

Yule's Q, while relatively easy to calculate, has a couple of notable limitations. One is that if one of the four cells in a 2 x 2 table (a table based on an independent variable with 2 categories and a dependent variable with 2 categories) has no cases, the calculated Yule's Q will be 1.00, even if the relationship isn't anywhere near that strong. (Why don't you try it with a sample that has 5 cases on cell A, 5 in cell B, 5 in cell C, and 0 in cell D?)

Another problem with Yule's Qis that it can only be used to describe 2 x 2 tables. But not all variables have just 2 categories. As a consequence, there are several other measures of strength of association for nominal level variables that can handle bigger tables. (One that we recommend for sheep farmers is lambda. Bahhh!) But, we most typically use one called Cramer's V, which shares with Yule's Q (and lambda) the property of varying between 0 and 1. Roger normally advises students that values of Cramer's V between 0.00 and 0.10 suggests that the relationship is weak; between 0.11 and 0.30, that the relationship is moderately strong; between 0.31 and and 0.59, that the relationship is strong; and between 0.60 and 1.00, that the relationship is very strong. Associations (a fancy word for the strength of the relationship) above 0.59 are not so common in social science research.

An example of the use of Cramer's V? Roger used statistical software called the Statistical Package for the Social Sciences (SPSS) to analyze the data Angela, Alyssa and he collected about commercials (on one night) to see whether men or women, both or neither, were more likely to appear as the main characters in commercials focused on domestic goods (goods used inside the home) and non-domestic goods (goods used outside the home). Who (men or women or both) would you expect to be such (main) characters in commercials involving domestic products? Non-domestic products? If you guessed that females might be the major characters in commercials for domestic products (e.g., food, laundry detergent, and home remedies) and males might be major characters in com-

mercials for non-domestic products (e.g., cars, trucks, cameras), your guesses would be consistent with findings of previous researchers (e.g., O'Donnell and O'Donnell, 1978; Lovdal, 1989; Bartsch et al., 2001). The data we collected on our first night of data collection suggest some support for these findings (and your expectations), but also some support for another viewpoint. Table 3.5, for instance, shows that women were, in fact, the main characters in about 48 percent of commercials for domestic products, while they were the main characters in only about 13 percent of commercials for non-domestic products. So far, so good. But males, too, were more likely to be main characters in commercials for domestic products (they were these characters about 24 percent of the time) than they were in commercials for non-domestic products (for which they were the main character only about 4 percent of the time). So who were the main product "representatives" for non-domestic commercials? We found that in these commercials at least one man *and* one woman were together the main characters about 50 percent of the time, while men and women together were the main characters in only about 18 percent of the time in commercials for domestic products.

But the analysis involving gender of main character and whether products were domestic or non-domestic involved more than a 2 x 2 table. In fact, it involved a 2 x 4 table because our dependent variable, gender of main character, had four categories: female, male, both, and neither. Consequently, we couldn't use Yule's Q as a measure of strength of association. But we could ask, and did ask (using SPSS), for Cramer's V, which turned out to be about 0.53, suggesting (if you re-examine Roger's advice above) that the relationship is a strong one.

Table 3.5 Crosstabulation of Type of Commercial and Gender of Main Character

		Type of Commercial	
		For Domestic Product	*For Non-Domestic Product*
Gender of Main Character	*Female*	18 (47.4%)	3 (12.5%)
	Male	9 (23.7%)	1 (4.2%)
	Both	7 (18.4%)	12 (50%)
	Neither	4 (10.4%)	8 (33.3%)
Notes: Cramer's V = 0.53			

Generalizability of the Relationship

When we speak of the **generalizability** of a relationship, we're dealing with the question of whether something like the relationship (in direction, if not strength) that is found in

the sample can be safely generalized to the larger population from which the sample was drawn. If, for instance, we drew a probability sample of eight adolescent women like the ones we pretended to draw in the first example above, we'd know we have a sample in which a strong relationship existed between watching "16 and Pregnant" and not becoming pregnant. But how could one tell that this sample relationship was likely to be representative of the true relationship in the larger population?

If you recall the distinction we drew between *descriptive* and *inferential statistics* in the Chapter on Univariate Analysis, you won't be surprised to learn that we are now entering the realm of inferential statistics for bivariate relationships. When we use percentage comparisons within one category of the dependent variable to determine the direction of a relationship and measures like Yule's Q and Cramer's V to get at its strength, we're using descriptive statistics—ones that describe the relationship in the sample. But when we talk about **Pearson's chi-square** (or X^2), we're referring to an inferential statistic—one that can help us determine whether we can generalize that something like the relationship in the sample exists in the larger population from which the sample was drawn.

But, before we learn how to calculate and interpret Pearson's chi-square, let's get a feel for the logic of this inferential statistic first. Scientists generally, and social scientists in particular, are very nervous about inferring that a relationship exists in the larger population when it really doesn't exist there. This kind of error—the one you'd make if you inferred that a relationship existed in the larger population when it didn't really exist there—has a special name: a **Type I error**. Social scientists are so anxious about making Type 1 errors that they want to keep the chances of making them very low, but not impossibly low. If they made them impossibly low, then they'd risk making the opposite of a Type 1 error: a **Type 2 error**—the kind of error you'd make when you failed to infer that a relationship existed in the larger population when it really did exist there. The chances, or probability, of something happening can vary from 0.00 (when there's no chance at all of it happening) to 1.00, when there's a perfect chance that it will happen. In general, social scientists aim to keep the chances of making a Type 1 error below .05, or below a 1 in 20 chance. They thus aim for a very small, but not impossibly small, chance of making the inference that a relationship exists in the larger population when it doesn't really exist there.

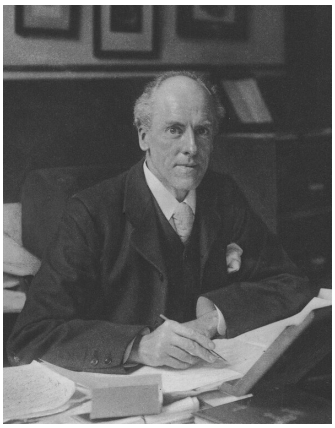


Figure 3.3: Karl Pearson in 1910

Karl Pearson, the statistician whose name is associated with Pearson's chi-square, studied the statistic's property in about 1900. He found, among other things, that crosstabulations of different sizes (i.e., different numbers of cells) required a different chi-square to be associated with a .05 chance, or probability (*p*), of making a Type 1 error or less. As the number of cells increase, the required chi-square increases as well. For a 2 x 2 table, the critical chi-square is 3.84 (that is, the computed chi-square value should be 3.84 or more for you to infer that a relationship exists in the larger population with only a .05 chance, or less, of being wrong); for a 2 x 3 table, the critical chi-square is 5.99, and so on. Before we were able to use statistical processing software like SPSS, statistical researchers relied on tables that outlined the critical values of chi-quare for different size tables (degrees of freedom, to be discussed below) and different probabilities of making a Type 1 error. A truncated (shortened) version of such a table can be seen in Table 6.

Table 6: Table of Critical Values of the Chi-Square Distribution

	Probability less than the critical value			
Degrees of Freedom	**0.90**	**0.95**	**0.99**	**0.999**
1	2.706	3.841	5.024	10.828
2	4.605	5.991	7.378	13.816
3	6.251	7.815	9.384	16.266
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	20.515
6	10.645	12.592	14.449	22.458
7	12.017	14.067	17.013	24.458

And so on...

Now you're ready to see how to calculate chi-square. The formula for chi-square (X^2) is:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where

χ means "the sum of"

O = the number of observed number of cases in each cell in the sample

E = the expected number in each cell, if there were no relationship between the two variables

Let's see how this would work with the example of the imaginary data in Table 3.3. This table, if you recall, looked (mostly) like this:

Table 7 (Slightly Revised) Crosstabulation of Our Imaginary Data from a "16 and Pregnant" Study

		Watched *16 and Pregnant*		Row Marginals
		Yes	*No*	
Got Pregnant	*Yes*	1	3	4
	No	3	1	4
Column Marginals		4	4	*N=8*

How do you figure out what the expected number of cases would be in each cell? You use the following formula:

$$E = \frac{M_r \times M_c}{N}$$

Where

M_r is the row marginal for the cell

M_c is the column marginal for the cell

N is the total number of cases in the sample

A **row marginal** is the total number of cases in a given row of a table. A **column marginal** is the total number of cases in a given column of a table. For this table, the N is 8, the total number of cases involved in the crosstabulation. For cell A, the row marginal is 4 and the column marginal is 4, which means its expected number of cases would be 4 x 4 = 16/8 = 2. In this particular table, all the cells would have had an expected frequency (or number of cases) of 2. So now all we have to do to compute χ^2 is to make a series of calculation columns:

Cell	Observed Number of Cases in Cell	Expected Number of Cases in Cell	(O-E)	$(O-E)^2$	$(O-E)^2/E$
A	1	2	-1	1	½
B	3	2	1	1	½
C	1	2	-1	1	½
D	3	2	1	1	½

And the sum of all the numbers in the $(0-E)^2/E$ column is 2.00. This is less than the 3.84 that χ^2 needs to be for us to conclude that the chances of making a Type 1 error are less than .05 (see Table 3.6), so we cannot safely generalize that something like the relationship in this small sample exists in the larger population. Aren't you glad that these days programs like SPSS can do these calculations for us? Even though they can, it's important to go through the process a few times on your own so that you understand what it is that the computer is doing.

Chi-square varies based on three characteristics of the sample relationship. The first of these is the number of cells. Higher chi-squares are more easily achieved in tables with more cells; hence the 3.84 standard for 2 x 2 tables and the 5.99 standard for 2 x 3 tables. You'll recall from Table 3.6 that we used the term **degrees of freedom** to refer to the calculation of table size. To figure out the degrees of freedom for a crosstabulation, you simply count the number of columns in the table (only the columns with data in them, not columns with category names) and subtract one. Then you count the number of rows in the table, again only those with data in them, and subtract one. Finally, you multiply the two numbers you have computed. Therefore, the degrees of freedom for a 2×2 table will be 1 [(2-1)*(2-1)], while the degrees of freedom for a 4×6 table will be 15 [(4-1)*(6-1)].

Higher chi-squares will also be achieved when the relationship is stronger. If, instead of the 1, 3, 3, 1 pattern in the four cells above (a relationship that yields a Yule's Q of 0.80, one had a 0, 4, 4, 0 pattern (a relationship that yields a Yule's Q of 1.00), the chi-square would be 8.00,[3] considerably greater than the 3.84 standard, and one could then generalize that something like the relationship in the sample also existed in the larger population.

But chi-square also varies with the size of the sample. Thus, if instead of the 1, 3, 3, 1 pattern above, one had a 10, 30, 30, 10 pattern—both of which would yield a Yule's Q of 0.80 and are therefore of the same strength, and both of which have the same number of cells (4)—the chi-square would compute to be 20, instead of 2, and give pretty clear guidance to infer that a relationship exists in the larger population. The message of this last co-variant of chi-square—that it grows as the sample grows—implies that researchers who want to find generalizable results do well to increase sample size. A sample that tells us that the relationship under investigation *is* generalizable is said to be **significant**—sometimes a desirable and often an interesting thing.[4] Incidentally, SPSS computed the chi-square for the crosstabulation in Table 3.5, the one that showed the relationship between type of product advertised (domestic or non-domestic) and the gender of the product representative, to be 17.5. Even for a 2 x 4 table like that one, this is high enough to infer that a relationship exists in the larger population, with less than a .05 chance of being wrong. In fact, SPSS went even further, telling us that the chances of making a Type 1 error were less than .001. (Aren't computers great?)

3. Can you double-check Roger's calculation of chi-square for this arrangement to make sure he's right? He'd appreciate the help.
4. Of course, with very large samples, like the entire General Social Survey (GSS) since it was begun, it is sometimes possible to uncover significant relationships—i.e., ones that almost surely exist in the larger population—that aren't all that strong. Does that make sense?

Crosstabulation with Two Ordinal Level Variables

We've introduced crosstabulation as a technique designed for the analysis of the relationship between two nominal level variables. But because all variables are at least nominal level, one could theoretically use crosstabulation to analyze the relation between variables of any scale.[5] In the case of two interval level variables, however, there are much more elegant techniques for doing so and we'll be looking at those in the chapter on correlation and regression. If one were looking into the relationship between a nominal level variable (say, gender, with the categories male and female)[6] and an ordinal level variable (say, happiness with marriage with the three categories: very happy, happy, not so happy), one could simply use all the same techniques for determining the direction, strength, and generalizability we've discussed above.

If we chose to analyze the relationship between two ordinal level variables, however, we could still use crosstabulation, but we might want to use a more elegant way of determining direction and strength of relationship than by comparing percentages and seeing what Cramer's V tells us. One very cool statistic used for determining the direction and strength of a relationship between two ordinal level variables is **gamma**. Unlike Cramer's V and Yule's Q, whose values only vary between 0.00 and 1.00, and therefore can *only* speak to the strength of a relationship, gamma's possible values are between -1.00 and 1.00. This one statistic can tell us about *both* the direction *and* the strength of the relationship. Thus, a gamma of zero still means there is no relationship between the two variables. But a gamma with a positive sign not only reveals strength (a gamma of 0.30 indicates a stronger relationship than one of 0.10), but it also says that as values of the independent variable increase, so do values of the dependent variable. And a gamma with a negative sign not only reveals strength (a gamma of -0.30 indicates a *stronger* relationship than one of -0.10), but also says that as values of the independent variable increase, values of the dependent variable decrease. But what exactly do we mean by "values," here?

Let's explore a couple of examples from the GSS (via the Social Data Archive, or SDA). Table 8 shows the relationship between the happiness of GSS respondents' marriages (HAPMAR) and their general happiness (HAPPY) over the years. Using our earlier way of determining direction, we can see that 90 percent of those that are "very happy" generally are also happy in their marriages, while only 19.5 percent of those who are "not too happy"

5. You would generate some pretty gnarly tables that would be very hard to interpret, though.
6. While there are clearly more than two genders, we are at the mercy of the way the General Social Survey asked its questions in any given year, and thus for the examples presented in this text only data for males and females is available. While this is unfortunate, it's also an important lesson about the limits of existing survey data and the importance of ensuring proper survey question design.

generally are pretty happy in their marriages. Pretty clear that marital happiness and general happiness are related, right?

Table 8. Crosstabulation of Marital Happiness and General Happiness, GSS data from SDA

Frequency Distribution					
Cells contain: –**Column percent** -N of cases		HAPPY			
		1 very happy	2 pretty happy	3 not too happy	***ROW TOTAL***
HAPMAR	1: very happy	**90.0** 11,666	**46.5** 7,938	**35.0** 894	***63.0*** *20,498*
	2: pretty happy	**9.6** 1,237	**51.0** 8,617	**45.5** 1,120	***34.0*** *10,974*
	3: not too happy	**.4** 51	**2.4** 433	**19.5** 503	***2.9*** *987*
	COL TOTAL	***100.0*** *12,954*	***100.0*** *16,988*	***100.0*** *2,517*	***100.0*** *32,459*
Means		1.10	1.56	1.84	1.40
Std Devs		.32	.54	.72	.55
Unweighted N		12,954	16,988	2,517	32,459

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0

N in each cell: Smaller than expected Larger than expected

Summary Statistics						
Eta* =	.46	Gamma =	.75	Rao-Scott-P: F(4,2360) =	1,807.32	(p= 0.00)
R =	.46	Tau-b =	.45	Rao-Scott-LR: F(4,2360) =	1,709.73	(p= 0.00)
Somers' d* =	.42	Tau-c =	.35	Chisq-P(4) =	8,994.28	
				Chisq-LR(4) =	8,508.63	

*Row variable treated as the dependent variable.

The more elegant way is to look at the statistics at the bottom of the table. Most of these statistics aren't helpful to us now. But one, gamma, certainly is. You'll note that gamma is 0.75. There are two important attributes of this statistic: its sign (positive) and its magnitude (0.75). The former tells you that as coded values of marital happiness—1=very happy; 2 happy; 3=not so happy—go up, values of general happiness—1=very happy; 2=happy; 3=not so happy—tend to go up as well. We can interpret this by saying that respondents who are

less happy with their marriages are likely to be less happy generally than others. (Notice that this also means that people who are happy in their marriages are also likely to be more generally happy than others.) But the 0.75, independent of the sign, means that this relationship is very strong. By the way, you might also notice that there is a little parenthetical expression at the end of the row gamma is on in the statistics box—(p=0.00). The "p" stands for the chances (probability) of making a Type 1 error, and is sometimes called the "**p value**" or the significance level. The fact that the "p value" here is 0.00 does NOT mean that there is zero chance of making an error if you infer that there is a relationship between marital happiness and general happiness in the larger population. There will always be such a chance. But the SDA printouts of such values give up after two digits to the right of the decimal point. All one can really say is that the chances of making a Type 1 error, then, are less than 0.01 (which itself is less than 0.05)—and so researchers would conclude that they could reasonably generalize.

To emphasize the importance of the sign of gamma (+ or -), let's have a look at Table 9, which displays the relationship between job satisfaction, whose coded values are 1=very dissatisfied; 2=a little dissatisfied; 3= moderately satisfied; 4=very satisfied, and general happiness, whose codes are the same as they were in Table 3.7. You can probably tell from looking at the internal percentages of the table that as job satisfaction increases so does general happiness—as one might expect. But sign of the gamma of -0.43 might at first persuade you that there is a negative association between job satisfaction and happiness, until you remember that what it's really telling you is that when the coded values of job satisfaction go up, from 1 (very dissatisfied) to 4 (very satisfied), the coded values of happiness go down, from 3 (not so happy) to 1 (very happy). Which really means that as job satisfaction goes up, happiness goes up as well, right? Note, however, that if we reversed the coding for the job satisfaction variable, so that 1 represented being very satisfied with your job while 4 represented being very dissatisfied, the direction of gamma would reverse. Thus, it is essential that data analysts do not stop by looking at whether gamma is positive or negative, but rather also ensure they understand the way the variable is coded (its **attributes**).

Also note here that the 0.43 portion of the gamma tells you how strong this relationship is—it's strong, but not as strong as the relationship between marital happiness and general happiness (which had a gamma of 0.75). The "p value" here again is .00, which means that it's less than .01, which of course is less than .05, and we can infer that there's very probably a relationship between job satisfaction and general happiness in the larger population from which this sample was drawn.

Table 9. Crosstabulation of Job Satisfaction and General Happiness, GSS data from SDA

Frequency Distribution

Cells contain: –**Column percent** -Weighted N		satjob2 1 Very Dissatisfied	2 A Little Dissatisfied	3 Moderately Satisfied	4 Very Satisfied	*ROW TOTAL*
happy	1: very happy	**15.1** 283.0	**15.6** 722.3	**23.9** 4,317.4	**44.9** 10,134.3	***32.8*** *15,457.0*
	2: pretty happy	**51.1** 955.7	**62.1** 2,877.8	**64.9** 11,716.0	**48.7** 10,982.8	***56.3*** *26,532.3*
	3: not too happy	**33.8** 631.3	**22.3** 1,034.4	**11.3** 2,032.6	**6.4** 1,448.8	***10.9*** *5,147.1*
	COL TOTAL	***100.0*** *1,870.0*	***100.0*** *4,634.5*	***100.0*** *18,066.0*	***100.0*** *22,566.0*	***100.0*** *47,136.4*
Means		2.19	2.07	1.87	1.62	1.78
Std Devs		.67	.61	.58	.60	.62
Unweighted N		1,907	4,539	17,514	22,091	46,051

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0

N in each cell: Smaller than expected Larger than expected

Summary Statistics

Eta* =	.28	Gamma =	-.43	Rao-Scott-P: F(6,3396) =	584.48	(p= 0.00)
R =	-.28	Tau-b =	-.26	Rao-Scott-LR: F(6,3396) =	545.83	(p= 0.00)
Somers' d* =	-.25	Tau-c =	-.23	Chisq-P(6) =	4,310.95	
				Chisq-LR(6) =	4,025.87	

*Row variable treated as the dependent variable.

We haven't shown you the formula for gamma, but it's not that difficult to compute. In fact, when you have a 2 x 2 table gamma is the same as Yule's Q, except that it can take on both positive and negative values. Obviously, Yule's Q could do that as well, if it weren't for the absolute value symbols surrounding it. As a consequence, you can use gamma as a substitute for Yule's Q for 2 x 2 tables when using the SDA interface to access GSS data—as long as you remember to take the absolute value of gamma that is calculated for you. Thus, in Table 10, showing the relationship between gender and whether or not a respondent was married, the absolute value of the reported gamma—that is, |-0.11|=0.11—is the Yule's Q for the relationship. And it is clearly weak. By the way, the p value here, 0.07, indicates that we cannot safely infer that a similar relationship existed in the larger population in 2010.

Table 10. Crosstabulation of Gender and Marital Status in 2010, GSS data from SDA

Frequency Distribution

Cells contain: –**Column percent** -Weighted N		**sex** 1 male	 2 female	 ***ROW TOTAL***
	0: not married	**45.4** 420.9	**50.7** 565.9	***48.3*** *986.8*
married	1: married	**54.6** 506.1	**49.3** 549.5	***51.7*** *1,055.6*
	COL TOTAL	***100.0*** *927.0*	***100.0*** *1,115.4*	***100.0*** *2,042.4*
Means		.55	.49	.52
Std Devs		.50	.50	.50
Unweighted N		891	1,152	2,043

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 **Z**

N in each cell: Smaller than expected Larger than expected

Summary Statistics

Eta* =	.05	Gamma =	-.11	Rao-Scott-P: F(1,78) =	3.29	(p= 0.07)
R =	-.05	Tau-b =	-.05	Rao-Scott-LR: F(1,78) =	3.29	(p= 0.07)
Somers' d* =	-.05	Tau-c =	-.05	Chisq-P(1) =	5.75	
				Chisq-LR(1) =	5.76	

*Row variable treated as the dependent variable.

One problem with an SDA output is that none of the statistics reported (not the Eta, the R, the Tau-b, etc.) are actually designed to measure the strength of relationship between two purely nominal level variables—Cramer's V and Yule's Q, for instance, are not provided in the output. All of the measures that are provided, however, do have important uses. To learn more about these and other measures of association and the circumstances in which they should be used, see the chapter focusing on measures of association.

Exercises

1. Write definitions, in your own words, for each of the following key concepts from this chapter:
   - independent variable
   - dependent variable

- crosstabulation
- direction of a relationship
- strength of a relationship
- generalizability of relationship
- Yule's Q
- Cramer's V
- Type 1 error
- Type 2 error
- Pearson's chi-square
- gamma
- hypothesis
- null hypothesis

2. Use the following (hypothetical) data, meant to test the hypothesis (with a hypothetically random sample) that adults tend to be taller than children. Create a crosstabulation of the data that enables you to determine the direction, strength and generalizability of the relationship, as well as what determinations you can make in relation to the null and research hypotheses. Present the statistics that permit you to describe these characteristics:

Case	Gender	Height
Person 1	Child	Short
Person 2	Adult	Tall
Person 3	Child	Short
Person 4	Adult	Tall
Person 5	Child	Short
Person 6	Adult	Tall
Person 7	Child	Short
Person 8	Adult	Tall

3. Return to the Social Data Archive we've explored before. The data, again, are available at https://sda.berkeley.edu/. Go down to the second full paragraph and click on the "SDA Archive" link you'll find there. Then scroll down to the section labeled "General Social Surveys" and click on the first link there: General Social Survey (GSS) Cumulative Datafile 1972-2021 release.

    - Now type "hapmar" in the row box and "satjob" in the column box. Hit "output options" and find the "percentaging" options and make sure "column" is clicked. (Satjob will be our independent variable here, so we want column percentages.) Now click on "summary statistics," under "other options." Hit on "run the table," examine the resulting printout and write a short paragraph in which you use gamma and the p-value to evaluate the hypothesis that people who are more satisfied with their jobs are more likely to be happily married than those who are less satisfied with their jobs. Your paragraph should mention the direction, strength and generalizability of the relationship as well as what determinations you can make in terms of the null and research hypotheses.

Media Attributions

- A Mapping of the Hypothesis that Men Will Tend to be Taller than Women © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- A Mapping of Kearney and Levine's Hypothesis © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

5. Hypothesis Testing in Quantitative Research

MIKAILA MARIEL LEMONIK ARTHUR

Statistical reasoning is built on the assumption that data are **normally distributed**, meaning that they will be distributed in the shape of a **bell curve** as discussed in the chapter on Univariate Analysis. While real life often—perhaps even usually—does not resemble a bell curve, basic statistical analysis assumes that if all possible random samples from a population were drawn and the **mean** taken from each sample, the distribution of sample means, when plotted on a graph, would be normally distributed (this assumption is called the **Central Limit Theorem**). Given this assumption, we can use the mathematical techniques developed for the study of probability to determine the likelihood that the relationships or patterns we observe in our data occurred due to random chance rather than due some actual real-world connection, which we call statistical significance.

***Statistical* significance** is not the same as *practical* significance. The fact that we have determined that a given result is unlikely to have occurred due to random chance does not mean that this given result is important, that it matters, or that it is useful. Similarly, we might observe a relationship or result that is very important in practical terms, but that we cannot claim is statistically significant—perhaps because our sample size is too small, for instance. Such a result might have occurred by chance, but ignoring it might still be a mistake. Let's consider some examples to make this a bit clearer. Assume we were interested in the impacts of diet on health outcomes and found the statistically significant result that people who eat a lot of citrus fruit end up having pinky fingernails that are, on average, 1.5 millimeters longer than those who tend not to eat any citrus fruit. Should anyone change their diet due to this finding? Probably not, even those it is statistically significant. On the other hand, if we found that the people who ate the diets highest in processed sugar died on average five years sooner than those who ate the least processed sugar, even in the absence of a statistically significant result we might want to advise that people consider limiting sugar in their diet. This latter result has more practical significance (lifespan matters more than the length of your pinky fingernail) as well as a larger effect size or association (5 years of life as opposed to 1.5 millimeters of length), a factor that will be discussed in the chapter on association.

While people generally use the shorthand of "the likelihood that the results occurred by chance" when talking about statistical significance, it is actually a bit more complicated than that. What statistical significance is *really* telling us is the likelihood (or **probability**)

that a result equal to or more "extreme[1]" is true in the real world, rather than our results having occurred due to random chance or **sampling error**. Testing for statistical significance, then, requires us to understand something about probability.

A Brief Review of Probability

You might remember having studied probability in a math class, with questions about coin flips or drawing marbles out of a jar. Such exercises can make probability seem very abstract. But in reality, computations of probability are deeply important for a wide variety of activities, ranging from gambling and stock trading to weather forecasts and, yes, statistical significance.

Probability is represented as a proportion (or decimal number) somewhere between 0 and 1. At 0, there is absolutely no likelihood that the event or pattern of interest would occur; at 1, it is absolutely certain that the event or pattern of interest will occur. We indicate that we are talking about probability by using the symbol p. For example, if something has a 50% chance of occurring, we would write $p = 0.5$ or $\frac{1}{2}$. If we want to represent the likelihood of something *not* occurring, we can write $1 - p$.

Check your thinking: Assume you were flipping coins, and you called heads. The probability of getting heads on a coin flip using a fair coin (in other words, a normal coin that has not been weighted to bias the result) is 0.5. Thus, in 50% of coin flips you should get heads. Consider the following probability questions and write down your answers so you can check them against the discussion below.

- Imagine you have flipped the coin 29 times and you have gotten heads each time. What is the probability you will get heads on flip 30?
- What is the probability that you will get heads on *all* of the first five coin flips?
- What is the probability that you will get heads on *at least one* of the first five coin flips?

1. One way to think about this is to imagine that your result has been plotted on a bell curve. Statistical significance tells us the probability that the "real" result—the thing that is true in the real world and not due to random chance—is at the same point as or further along the skinny tails of the bell curve than the result we have plotted.

There are a few basic concepts from the mathematical study of probability that are important for beginner data analysts to know, and we will review them here.

Probability over Repeated Trials: The probability of the outcome of interest is the same in each trial or test, regardless of the results of the prior test. So, if we flip a coin 29 times and get heads each time, what happens when we flip it the 29th time? The probability of heads is still 0.5! The belief that "this time it must be tails because it has been heads so many times" or "this coin just wants to come up heads" is simply superstition, and—assuming a fair coin—the results of prior trials do not influence the results of this one.

Probability of Multiple Events: The probability that the outcome of interest will occur repeatedly across multiple trials is the product[2] of the probability of the outcome on each individual trial. This is called the **multiplication theorem**. Thinking about the multiplication theorem requires that we keep in mind the fact that when we multiply decimal numbers together, those numbers get *smaller*—thus, the probability that a series of outcomes will occur is *smaller than* the probability of any one of those outcomes occurring on its own. So, what is the probability that we will get heads on all five of our coin flips? Well, to figure that out, we need to multiply the probability of getting heads on each of our coin flips together. The math looks like this (and produces a very small probability indeed):

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.03125$$

Probability of One of Many Events: Determining the probability that the outcome of interest will occur on at least one out of a series of events or repeated trials is a little bit more complicated. Mathematicians use the **addition theorem** to refer to this, because the basic way to calculate it is to calculate the probability of each sequence of events (say, heads-heads-heads, heads-heads-tails, heads-tails-heads, and so on) and add them together. But the greater the number of repeated trials, the more complicated that gets, so there is a simpler way to do it. Consider that the probability of getting *no* heads is the same as the probability of getting *all tails* (which would be the same as the probability of getting all heads that we calculated above). And the only circumstance in which we would not have at least one flip resulting in heads would be a circumstance in which *all* flips had resulted in tails. Therefore, what we need to do in order to calculate the probability that we get at least one heads is to subtract the probability that we get *no* heads from 1—and as you can imagine, this procedure shows us that the probability of the outcome of interest occurring at least once over repeated trials is *higher* than the probability of the occurrence on any given trial. The math would look like this:

$$1 - (\frac{1}{2})^5 = 0.9688$$

So why is this digression into the math of probability important? Well, when we test for

2. In other words, what you get when you multiply.

statistical significance, what we are really doing is determining the probability that the outcome we observed—or one that is more extreme than that which we observed—occurred by chance. We perform this analysis via a procedure called Null Hypothesis Significance Testing.

Null Hypothesis Significance Testing

Null hypothesis significance testing, or **NHST**, is a method of testing for **statistical significance** by comparing observed data to the data we would expect to see if there were no relationship between the variables or phenomena in question. NHST can take a little while to wrap one's head around, especially because it relies on a logic of double negatives: first, we state a hypothesis we believe *not* to be true (there is no relationship between the variables in question) and then, we look for evidence that disconfirms this hypothesis. In other words, we are assuming that there is no relationship between the variables—even though our research hypothesis states that we think there *is* a relationship—and then looking to see if there is any evidence to suggest there is *not* no relationship. Confusing, right?

So why do we use the null hypothesis significance testing approach?

- The null hypothesis—that there is no relationship between the variables we are exploring—would be what we would generally accept as true in the absence of other information,
- It means we are assuming that differences or patterns occur due to chance unless there is strong evidence to suggest otherwise,
- It provides a benchmark for comparing observed outcomes, and
- It means we are searching for evidence that *disconforms* our hypothesis, making it less likely that we will accept a conclusion that turns out to be untrue.

Thus, NHST helps us avoid making errors in our interpretation of the result. In particular, it helps us avoid **Type 2 error**, as discussed in the chapter on Bivariate Analyses. As a reminder, Type 2 error is error where you accept a hypothesis as true when in fact it was false (while **Type 1 error** is error where you reject the hypothesis when in fact it was true). For example, you are making a Type 1 error if you decide not to study for a test because you assume you are so bad at the subject that studying simply cannot help you, when in fact we know from research that studying does lead to higher grades. And you are making a Type 2 error if your boss tells you that she is going to promote you if you do enough overtime and you then work lots of overtime in response, when actually your boss is just trying to make you work more hours and already had someone else in mind to promote.

We can never remove all sources of error from our analyses, though larger sample sizes help reduce error. Looking at the formula for computing **standard error**, we can see that the standard error (SE) would get smaller as the sample size (N) gets larger. Note: σ is the symbol we use to represent standard deviation.

$$SE = \frac{\sigma}{\sqrt{N}}$$

Besides making our samples larger, another thing that we can do is that we can choose whether we are more willing to accept Type 1 error or Type 2 error and adjust our strategies accordingly. In most research, we would prefer to accept more Type 1 error, because we are more willing to miss out on a finding than we are to make a finding that turns out later to be inaccurate (though, of course, lots of research does eventually turn out to be inaccurate).

Performing NHST

Performing NHST requires that our data meet several assumptions:

1. Our sample must be a random sample—statistical significance testing and other inferential and explanatory statistical methods are generally not appropriate for non-random samples[3]—as well as representative and of a sufficient size (see the Central Limit Theorem above).
2. Observations must be independent of other observations, or else additional statistical manipulation must be performed. For instance, a dataset of data about siblings would need to be handled differently due to the fact that siblings affect one another, so data on each person in the dataset is not truly independent.
3. You must determine the rules for your significance test, including the level of uncertainty you are willing to accept (significance level) and whether or not you are interested in the direction of the result (one-tailed versus two-tailed tests, to be discussed below), in advance of performing any analysis.
4. The number of significance tests you run should be limited, because the more tests you run, the greater the likelihood that one of your tests will result in an error. To make this more clear, if you are willing to accept a 5% probability that you will make the error of accepting a hypothesis as true when it is really false, and you run 20 tests, one of those tests (5% of them!) is pretty likely to have produced an incorrect result.

3. They also are not appropriate for censuses—but you do not need inferential statistics in a census because you are looking at the entire population rather than a sample, so you can simply describe the relationships that do exist.

If our data has met these assumptions, we can move forward with the process of conducting an NHST. This requires us to make three decisions: determining our **null hypothesis**, our **confidence level** (or acceptable significance level), and whether we will conduct a one-tailed or a two-tailed test. In keeping with Assumption 3 above, we must make these decisions before performing our analysis. The null hypothesis is the hypothesis that there is no relationship between the variables in question. So, for example, if our research hypothesis was that people who spend more time with their friends are happier, our null hypothesis would be that there is no relationship between how much time people spend with their friends and their happiness.

Our confidence level is the level of risk we are willing to accept that our results could have occurred by chance. Typically, in social science research, researchers use $p<0.05$ (we are willing to accept up to a 5% risk that our results occurred by chance), $p<0.01$ (we are willing to accept up to a 1% risk that our results occurred by chance), and/or $p<0.001$ (we are willing to accept up to a 0.1% risk that our results occurred by chance). P, as was noted above, is the mathematical notation for probability, and that's why we use a p-value to indicate the probability that our results may have occurred by chance. A higher p-value increases the likelihood that we will accept as accurate a result that really occurred by chance; a lower p-value increases the likelihood that we will assume a result occurred by chance when actually it was real. Remember, what the p-value tells us is not the probability that our own research hypothesis is true, but rather this: assuming that the null hypothesis is correct, what is the probability that the data we observed—or data more extreme than the data we observed—would have occurred by chance.

Whether we choose a one-tailed or a two-tailed test tells us what we mean when we say "data more extreme than." Remember that normal curve? A two-tailed test is agnostic as to the direction of our results—and many of the most common tests for statistical significance that we perform, like the Chi square, are two-tailed by default. However, if you are only interested in a result that occurs in a particular direction, you might choose a one-tailed test. For instance, if you were testing a new blood pressure medication, you might only care if the blood pressure of those taking the medication is significantly *lower* than those not taking the medication—having blood pressure significantly *higher* would not be a good or helpful result, so you might not want to test for that.

Having determined the parameters for our analysis, we then compute our test of statistical significance. There are different tests of statistical significance for different variables (for example, the **Chi square** discussed in the chapter on bivariate analyses), as you will see in other chapters of this text, but all of them produce results in a similar format. We then compare this result to the p value we already selected. If the p value produced by our analysis is lower than the confidence level we selected, we can reject the null hypothesis, as the probability that our result occurred by chance is very low. If, on the other hand, the p value produced by our analysis is higher than the confidence level we selected, we fail to reject the

null hypothesis, as the probability that our result occurred by chance is too high to accept. Keep in mind this is what we do even when the p value produced by our analysis is quite close to the threshold we have selected. So, for instance, if we have selected the confidence level of $p<0.05$ and the p value produced by our analysis is $p=0.0501$, we still fail to reject the null hypothesis and proceed as if there is not any support for our research hypothesis.

> I actually like to think of the null hypothesis as 'innocent until proven guilty': the null hypothesis (innocence) is assumed to be true as long as there isn't enough evidence to reject it. –Patrick Altmeyer @paltmey via twitter, 09/13/2022, 3:55 pm.

Thus, the process of null hypothesis significance testing proceeds according to the following steps:

1. Determine the null hypothesis
2. Set the confidence level and whether this will be a one-tailed or two-tailed test
3. Compute the test value for the appropriate significance test
4. Compare the test value to the critical value of that test statistic for the confidence level you selected
5. Determine whether or not to reject the null hypothesis

Your statistical analysis software will perform steps 3 and 4 for you (before there was computer software to do this, researchers had to do the calculations by hand and compare their results to figures on published tables of critical values). But you as the researcher must perform steps 1, 2, and 5 yourself.

Confidence Intervals & Margins of Error

When talking about statistical significance, some researchers also use the terms **confidence intervals** and **margins of error**. Confidence intervals are ranges of probabilities within which we can assume the true population parameter lies. Most typically, analysts aim for 95% confidence intervals, meaning that in 95 out of 100 cases, the population parameter will lie within the upper and lower levels specified by your confidence interval. These are calculated by your statistics software as well. The margin of error, then, is the range of values within the confidence interval. So, for instance, a 2021 survey of Americans conducted by the Robert Wood Johnson Foundation and the Harvard T.H. Chan School of Public Health found that 71% of respondents favor substantially increasing federal spending on public health programs. This poll had a 95% confidence interval with a +/- 3.6 margin

of error. What this tells us is that there is a 95% probability (19 in 20) that between 67.4% (71-3.6) and 74.6% (71+3.6) of Americans favored increasing federal public health spending at the time the poll was conducted. When a figure reflects an overwhelming majority, such as this one, the margin of error may seem of little relevance. But consider a similar poll with the same margin of error that sought to predict support for a political candidate and found that 51.5% of people said they would vote for that candidate. In that case, we would have found that there was a 95% probability that between 47.9% and 55.1% of people intended to vote for the candidate—which means the race is total tossup and we really would have no idea what to expect. For some people, thinking in terms of confidence intervals and margins of error is easier to understand than thinking in terms of p values; confidence intervals and margins of error are more frequently used in analyses of polls while p values are found more often in academic research. But basically, both approaches are doing the same fundamental analysis—they are determining the likelihood that the results we observed or a similarly-meaningful result would have occurred by chance.

What Does Significance Testing Tell Us?

One of the most important things to remember about significance testing is that, while the word "significance" is used in ordinary speech to mean importance, significance testing does *not* tell us whether our results are important—or even whether they are interesting. A full understanding of the relationship between a given set of variables requires looking at statistical significance *as well as* association and the theoretical importance of the findings. Table 1 provides a perspective on using the combination of significance and association to determine how important the results of statistical analysis are—but even using Table 1 as a guide, evaluating findings based on theoretical importance remains key. So: make sure that when you are conducting analyses, you avoid being misled into assuming that significant results are sufficient for making broad claims about the importance and meaning of results. And remember as well that significance only tells us the likelihood that the pattern of relationships we observe occurred by chance—not whether that pattern is causal. For, after all, quantitative research can never eliminate all plausible alternative explanations for the phenomenon in question (one of the three elements of causation, along with association and temporal order).

Table 1. Significance and Association

		Significance	
		Significant	***Not Significant***
Strength of Association	***Strong***	Something's happening here!	Could be interesting, but might have occurred by chance
	Weak	Probably did not occur by chance, but not interesting	Nothing's happening here

Exercises

1. Using the approach described in this chapter, calculate the probability of the following coin flip scenarios:
    - Getting 7 heads on 7 coin flips
    - Getting 5 heads on 7 coin flips
    - Getting 1 head on 10 coin flips

    Then check your work using the Coin Flip Probability Calculator.
2. Write the null hypothesis for each of the following research hypotheses:
    - As the advertised hourly pay for a job goes up, the number of job applicants increases.
    - Teenagers who watch more hours of makeup tutorial videos on TikTok have, on average, lower self-esteem.
    - Couples who share hobbies in common are less likely to get divorced.
3. Assume a research conducted a study that found that people wearing green socks type on average one word per minute faster than people who are not wearing green socks, and that this study found a p value of $p<0.01$. Is this result *statistically* significant? Is this result *practically* significant? Explain your answers.
4. If we conduct a political poll and have a 95% confidence interval and a margin of error of +/- 2.3%, what can we conclude about support for Candidate X if 49.3% of respondents tell us they will vote for Candidate X? If 24.7% do? If 52.1% do? If 83.7% do?

6. An In-Depth Look At Measures of Association

MIKAILA MARIEL LEMONIK ARTHUR

Measures of **association** are statistics that tell analysts the strength of the relationship between two (or more) variables, as well as in some cases the direction of that relationship. There are a variety of measures of association; choosing the correct one for any given analysis requires understanding the nature of the variables being used for that analysis. This chapter will detail a number of measures of association that are used by quantitative analysts, though there are others that will not be covered here. While the chapter will not provide full instructions for calculating most measures of association, it aims to give those who are new to quantitative analysis a general understanding of how calculations of measures of association work, how to interpret and understand the results, and how to choose the correct measure of association for a given analysis.

To start, then, what do measures of association tell us? Remember that they do *not* tell us whether a result is **statistically significant**, as discussed in the chapter on statistical significance. Instead, they are designed to tell us about the nature and strength of the observed relationship between the variables, whether or not that relationship is likely to have occurred by chance. There are different ways of thinking about what association means: for instance, two variables that are strongly associated are those in which the values of one variable tend to co-occur with the values of the other variable. Or we might say that strongly associated variables are those in which variation in one variable can explain much of the variation in another variable. In addition, for analyses using only ordinal and/or continuous variables, some measures of association can tell us about the **direction** of the relationship—are we observing a direct (positive) relationship, where as the value of x goes up the value of y also goes up, or are we observing an inverse (indirect or negative) relationship, where as the value of x goes up the value of y goes down?

Keep in mind that it is possible for a relationship to appear to have a moderate or even strong association but for that association to not be meaningful in explaining the world. This can occur for a variety of reasons—the relationship may not be significant, and thus the likelihood that the observed pattern occurred by chance could be high. Note that even a $p<0.001$ there is a one in one-thousand likelihood that the result occurred by chance! Or the relationship may be **spurious**, and thus while it *appears* that the two variables are associated, this apparent association is only a reflection of the fact that each variable is separately associated with some other variable. Or the strong association may be due to the fact that both variables are basically measuring the same underlying phenomena, rather than

measuring separate but related phenomena (for instance, one would observe a very strong relationship between year of birth and age).

There is one other important difference between statistical significance and measures of association: while the computation of statistical significance assumes that data has been collected using a random sample, measures of association do not necessarily require that the data be from a random sample. Thus, for instance, measures of association can be computed for data from a census.

Preparing to Choose a Measure of Association

When choosing a measure of association, analysts must begin by ensuring that they understand how their variable is measured as well as the nature of the question they are asking about their data so that they can choose the measure of association that is best suited to these variables and this question. There are a number of relevant factors to consider.

First, the **levels of measurement** of the variables that are being used: different measures of association are appropriate for variables of different levels of measurement.

Second, whether information about the direction of the relationship is important to the research question. Some measures of association provide direction and others do not.

Third, whether a symmetric or an asymmetric measure is required. Symmetric measures consider the impact of each variable upon the other, while asymmetric measures are used in circumstances where the analyst wants to use an independent variable to explain or predict variation in a dependent variable. Note that when producing asymmetric measures of association in statistical software, the software will typically produce multiple versions, and the analyst must ensure that they use the one for the correct independent/dependent variable.

Fourth, the number of attributes of each variable (for non-continuous variables). Some measures of association are only appropriate for variables with few attributes—or for crosstabulations in which the resulting tables are relatively small—while others are appropriate for greater numbers of attributes and larger tables.

There are also specific circumstances that are especially suited to particular measures of association based on the nature of the research question or characteristics of the variables being used. And, as will be discussed below, it is essential to understand the way attributes are coded. It is especially important in the case of ordinal and continuous variables to understand whether increasing numerical values of the variable represent an increase or a decrease in the underlying concept being measured. Finally, there are a variety of factors other than the actual relationship between the variables that can impact the strength of association, including the sample size, unreliable measurements, the presence of outliers,

and data that are restricted in range.[1] Analysts should explore their data using descriptive statistics to see if any of these issues might impact the analysis.

Keep in mind that while it is sometimes appropriate to produce more than one measure of association as part of an analysis, it is not appropriate to simply run all of them and select the one that provides the most desirable result. Instead, the analyst should carefully consider the variables, their question, and the options and choose the one or two most appropriate to the situation to produce and interpret.

General Interpretation of Measures of Association

When interpreting measures of association, there are two piece of information to look for: (1) strength and (2) direction.

Table 1. Strength of Association

Strength	Value
None	0
Weak/Uninteresting	±0.01-0.09
Moderate	±0.10-0.29
Strong	±0.30-0.59
Very Strong	±0.60-0.99
Perfect Identity	±1

The strength of nearly all measures of association ranges from 0 to 1. Zero means there is no observed relationship at all between the two (or more) variables in question—in other words, their values are distributed completely randomly with respect to each other. One would represent what we call a complete identity—in other words, the two variables are measuring the exact same thing and all values line up perfectly. This would be the situation, for instance, if we looked at the association between height in inches and height in centimeters, which are after all just two different ways of measuring the same value. While different researchers do use different scales for assessing the strength of association, Table 1 provides one approach for doing so. Note that very strong values are quite rare in social science, as most social phenomena are too complex for the types of simple explanations where one variable explains most of the variation in another.

The direction of association, where applicable, is determined by whether the measure of association is a positive or negative number–whether the number is positive or negative does not tell us anything about strength (in other words, +0.5 is not bigger than -0.5—they are the *same strength* but a *different direction*). Positive numbers mean a direct associ-

1. For instance, a study looking at the relationship between age and health that only included people between the ages of 23 and 27 would be restricted in range in terms of age.

ation, while negative numbers mean an inverse relationship. Direction cannot be determined when examining relationships involving nominal variables, since nominal variables themselves do not have direction. Keep in mind that it is essential to understand how a variable is coded in order to interpret the direction. For example, imagine we have a variable measuring self-perceived health status. That variable could be coded as 1:poor, 2:fair, 3:good, 4:excellent. Or it could be coded as 1:excellent, 2:good, 3:fair, 4:poor. If we looked at the relationship between the first version of our health variable and age, we might expect that it would be negative, as the numerical value of the health variable would decline as age increased. And if we looked a the relationship between the second version of our health variable and age, we might expect that it would be positive, as the numerical value of the health variable would increase as age increased. The actual health data could be exactly the same in both cases—but if we change the direction of how our variable is coded, this changes the direction of the relationship as well.

Details on Measures of Association

In this section, we will review a variety of measures of association. For each one, we will provide information about the circumstances in which it is most appropriately used and other information necessary to properly interpret it.

Phi

Phi is a measure of association that is used when examining the relationship between two binary variables. Cramer's V and Pearson's r, discussed below, will return values identical to Phi when computed for two binary variables, but it is still more appropriate to use Phi. It is a symmetric measure, meaning it treats the two variables identically rather than assuming one variable is the independent variable and the other is the dependent variable. It can indicate direction, but given that binary variables are often assigned numerical codes somewhat at random (should yes be 0 and no 1, or should no be 0 and yes 1?), interpretation of the direction may not be of much use. The computation of Phi is the square root of the Chi square value divided by the sample size. While Phi is the most commonly used measure of association for relationships between two binary variables in social science data, there are other measures used in other fields (for instance, risk ratios in epidemiology) that are asymmetric. Yule's Q, discussed in several other chapters, is another example. These will not be discussed here.

Cramer's V

If there is any "default" measure of association, it is probably Cramer's V. Cramer's V is used in situations involving pairs of nominal, ordinal, or binary variables, though not in situations with two binary variables (then Phi is used) and it is less common in situations where both variables are ordinal. It is symmetric and non-directional. The size of the table/number of attributes of each variable does not matter. However, if there is a large difference between the number of columns and the number of rows, Cramer's V may overestimate the association between the variables. It is calculated by dividing the Chi square by the sample size multiplied by whichever is smaller, the number of rows in the table minus one or the number of columns in the table minus one, and then taking the square root of the resulting number.

Contingency Coefficient

The Contingency Coefficient is used for relationships in which at least one of the variables is nominal. It is symmetric and non-directional, and is especially appropriate for large tables (those 5×5 or larger—in other words, circumstances in which both variables have more than five attributes). This is because, for smaller tables, the Contingency Coefficient is not mathematically able to get close to one. It is computed by dividing the Chi square by the number of cases plus the Chi square, and then taking the square root of the result.

Lambda and Goodman & Kruskall's Tau

Lambda is a measure of association used when at least one variable is nominal. It is asymmetric and nondirectional. Some statisticians believe that Lambda is not appropriate for circumstances in which the dependent variable's distribution is skewed. Unlike measures based on the Chi square, Lambda is based on calculating what is called "the proportional reduction in error" (PRE) when one uses the values of the independent variable to predict the values of the dependent variable. The formula for doing this is quite complex, and involves the number of columns and rows in the table, the number of observations in a given row and column, the number of observations in the cell where that row and column intersect, and the total number of observations.

Goodman & Kruskall's Tau works according to similar principles as Lambda, but without consideration of the number of columns and rows. Thus, it is generally advised to use it only for fairly small tables. Like Lambda, it is asymmetric and non-directional. In some statistical

software packages (including SPSS), Goodman & Kruskall's Tau is produced when Lambda is produced rather than it being possible to select it separately.

Uncertainty Coefficient

The Uncertainty Coefficient is also used when at least one variable is nominal. It is asymmetric and directional. Conceptually, it measures the reduction in prediction error (or uncertainty) that occurs when one variable is used to predict the other. Some analysts prefer it to Lambda because it better accounts for the entire distribution of the variable, though others find it harder to interpret. As you can imagine, this makes the formula even more complicated than the formula for Lambda; it relies on information about the total number of observations in each row, each column, and each cell.

Spearman

Spearman is used when both variables are ordinal. It is symmetric and directional and can be used for large tables. In SPSS, it can be found under "correlations." Computing Spearman requires converting values into ranks and using the difference in ranks and the sample size in the formula. Note that if there are tied values or if the data is truncated or reduced in range Spearman may not be appropriate.

Gamma and Kendall's Tau (b and c)

The two Kendall's Tau measures are both symmetric and directional and are used for relationships involving two ordinal variables. However, Kendall's Tau b is used when tables are square, meaning that they have the same number of rows and columns, while Kendall's Tau c is used when tables are not square. Like Spearman, Kendall's Tau is based on looking at the relationship between ranks. After converting values to ranks, one counts pairs of values that are in agreement below a given rank (concordant pairs) and how many are not in agreement (discordant pairs). The formula, then, involves subtracting the number of discordant pairs from the number of concordant pairs, then dividing this number by the number of discordant pairs plus the number of concordant pairs.

Gamma is similar–also symmetric and directional and used for relationships involving two ordinal variables, and with a similar method of calculation, except using same-order and different-order (ranking high or low on both variables versus ranking high on one and

low on the other) instead of concordant and discordant pairs. Gamma is preferred when many of the observations in an analysis are tied, as ties are discounted in the computation of Kendall's tau and thus Kendall's tau will produce a more conservative (in other words, lower) value in such cases. However, Gamma may overestimate association for larger tables.

Kappa

Kappa is a measure of association that is especially likely to be used for testing interrater reliability, as it is designed for use when both variables are ordinal with the same categories. It measures agreement between the two variables and is symmetric. Kappa is calculated by subtracting the degree of agreement between the variables that would be expected by chance from the degree of agreement that is observed; subtracting the degree of agreement that would be expected by chance from one, and dividing the former by the latter.

Somers' D

Somers' D is designed for use in examining relationship involving two ordinal variables and is directional, but unlike the other ordinal x ordinal measures of association discussed above, Somers' D is asymmetric. As such, it measures the extent to which our ability to predict values of the dependent variable is improved by knowing the value of the independent variable. It is a conservative measure, underestimating the actual extent to which two variables are associated, though this underestimation declines as table size increases.

Eta

Eta is a measure of association that is used when the independent variable is discrete and the dependent variable is continuous. It is asymmetric and non-directional, and is primarily used as part of a statistical test called ANOVA, which is beyond the scope of this text. In circumstances where independent variables are discrete but not binary, many analysts choose to recode those variables to create multiple dummy variables, as will be discussed in the chapter on multivariate regression, and then use Pearson's R as discussed below.

Pearson's r

Pearson's r is used when examining relationships between two (or more) continuous variables and can also be used in circumstances where an independent variable is binary and a dependent variable is continuous. It is symmetric and directional. The calculation of Pearson's r is quite complex, but conceptually, what this calculation involves is plotting the data on a graph and then finding the line through the graph that best fits this data, a topic that will be further explored in the chapter on Correlation and Regression.

Other Situations

Attentive readers will have noticed that not all possible variable combinations have been addressed above. In particular, circumstances in which the independent variable is continuous and the dependent variable is *not* continuous have not been addressed. For beginning analysts, the most straightforward approach to measuring the association in such relationships is to recode the continuous variable to create an ordinal variable and then proceed with crosstabulation. However, there are a variety of more advanced forms of regression that are beyond the scope of this book, such as logistic regression, that can also handle relationships between these sorts of variables, and there are various pseudo-R measures of association that can be used in such analyses.

Exercises

1. Determine the strength and direction for each of the following measure of association values:
    - -0.06
    - 0.54
    - 0.13
    - -0.27
2. Select the most appropriate measure of association for each of the following relationships, and explain why it is the most appropriate:
    - Age, measured in years, and weight, measured in pounds
    - Opinion about the local police on a 5-point agree/disagree scale and highest educational degree earned
    - Whether or not respondents have health insurance (yes/no) and whether or not they have been to a doctor in the past 12 months (yes/no)
    - Letter grade on Paper 1 and letter grade on Paper 2 in a first-year composition class
3. Explain, in your own words, the difference between association and significance.

7. Multivariate Analysis

ROGER CLARK

We saw, in our discussion of bivariate analysis, how crosstabulation can be used to examine bivariate relationships, like the one Kearney and Levine discovered between watching *16 and Pregnant* and becoming pregnant for teenaged women. In this chapter, we'll be investigating how researchers gain greater understanding of bivariate relationships by controlling for other variables. In other words, we'll begin our exploration of **multivariate analyses**, or analyses that enable researchers to investigate the relationship between two variables while examining the role of other variables.

You may recall that Kearney and Levine claim to have investigated the relationship between watching *16 and Pregnant* and becoming pregnant, and thought it might have been at least partly due to the fact that those who watched were more likely to seek out information about (and perhaps use) contraception. Researchers call a variable that they think might affect, or be implicated in, a bivariate relationship a **control variable.** In the case of Kearney and Levine's study, the control variable they thought might be implicated in the relationship between watching *16 and Pregnant* and becoming pregnant was seeking out information about (or using) contraception.

Before we go further we'd like to introduce you to three kinds of control variables: *intervening*, *antecedent*, and *extraneous* control variables. An **intervening** control variable is a variable a researcher believes is affected by an independent variable and in turn affects a dependent variable. The Latin root of "intervene" is *intervener*, meaning "to come between"—and that's what intervening variables do. They come between, at least in the researcher's mind, the independent and dependent variables.

For Kearney and Levine, seeking information about contraception was an intervening variable: it's a variable they thought was affected by watching *16 and Pregnant* (their independent variable) and in turn affected the likelihood that a young woman would be become pregnant (their dependent variable). More precisely, their three-variable hypothesis goes something like this: a young woman who watched *16 and Pregnant* was more likely to seek information than a woman who did not watch it, and a woman who sought information about contraception was less likely to get pregnant than a woman who did not seek information about contraception. One quick way to map such a hypothesis is the following:

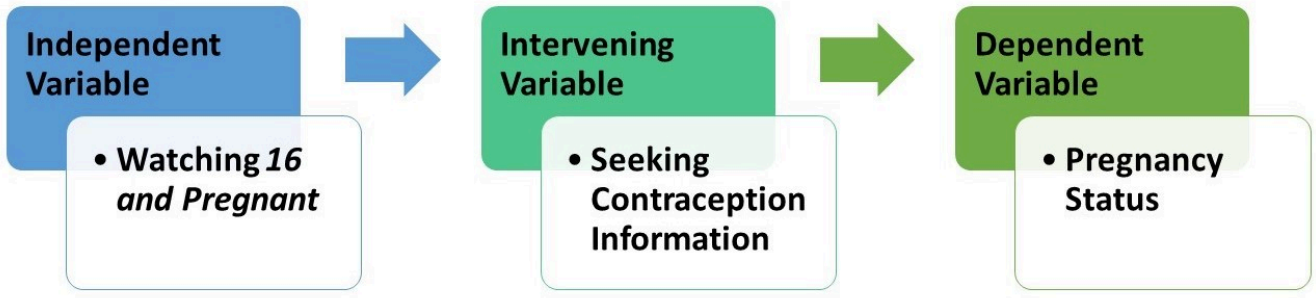


Figure 1. A Depiction of Kearney & Levine's Hypothesis

Importantly, researchers who believe they've found an intervening variable linking an independent variable and a dependent variable don't believe they are challenging the possibility that the independent variable may be a cause of variation in the dependent variable. (More about "cause" in a second.) They are simply pointing to a possible way, or mechanism through which, the independent variable may cause or affect variation in the dependent variable.

A second kind of control variable is an antecedent variable. An **antecedent variable** is a variable that a researcher believes affects both the independent variable and the dependent variable. Antecedent has a Latin root that translates into "something that came before." And that's what researchers who think they've found an antecedent variable believe: that they've found a variable that not only comes before and affects both the independent variable and the dependent variable, but also, in some real sense, causes them to go together, or to be related.

Example

For an example of a researcher/theorist who thinks he may have found an antecedent variable that explains a relationship, think about what Robert Sternberg is saying about the correlation between the attractiveness of children and the care their parents give them in this article on the research by W. Andrew Harrell.

Quiz at the end of the article: What two variables constituted the independent and dependent variables of the finding announced by researchers at the University of Alberta? How did they show these two variables were related? What variable did Robert Sternberg suspect might have been an antecedent variable for the independent and dependent variables found to be related by the U. of Alberta researchers? How did he think this variable might explain the relationship?

If you said that the basic relationship discovered by the University of Alberta researchers was that ugly children get poorer care from their parents than pretty children, you were right on the money. (It's back-patting time!) Here the proposed independent variable was the attractiveness of children and the dependent variable was the parental care they received.

If you said that the socioeconomic status or wealth of the parents was what Sternberg thought might be an

antecedent variable for these two variables (attractiveness and care), then you should glow with pride. Sternberg suggested that wealthier parents can both make their children look more attractive than poorer parents can and give their children better care than poorer parents can. One quick way to map such a hypothesis is like this:

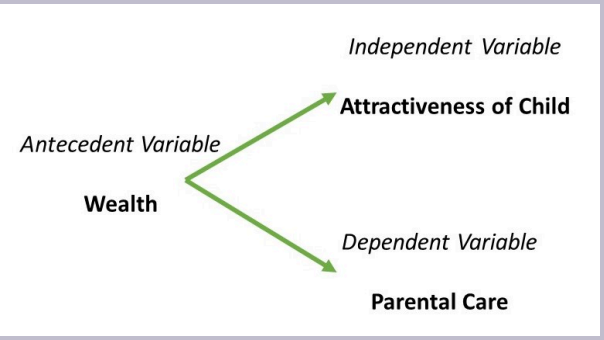


Figure 2. Sternberg's Hypothesis

A Word About Causation

Importantly, a researcher who thinks they have found an antecedent variable for a relationship implies that that have found a reason why the original relationship might be non-causal. **Spurious** is a word researchers use to describe non-causal relationships. Philosophers of science have told us that in order for a relationship between an independent variable and a dependent variable to be causal, three conditions must obtain:

1. The independent and dependent variables must be related. We demonstrated ways, using crosstabulation, that such relationships can be established with data. The Alberta researchers did show that the attractiveness of children was associated with how well they were treated (cared for or protected) in supermarkets. This condition is sometimes called **association**.

2. Instances of the independent variable occurring must come before, or at least not after, instances of the dependent variables. The attractiveness of the children in the Alberta study almost certainly preceded their treatment by their parents during the shopping expeditions observed by the researchers. This factor is often called **temporal order**.

3. There can be NO antecedent variable that creates the relationship between the independent variable and the dependent variable. This is the really tough condition for researchers to demonstrate, because, in principle, there could be an infinite numbers of antecedent variables that create such a relationship. This factor is often called **elimination of alternatives**. There is one research method—the controlled laboratory experiment—that theoretically eliminates this difficulty, but it is beyond the scope of this book to show you how. Yet it is not beyond our scope to show you how an antecedent variable might be

shown, with data, to throw real doubt on the notion that an independent variable causes a dependent variable. And we'll be doing that shortly.

Back to Our Main Story

A third kind of control variable is an **extraneous variable**. An extraneous variable is a variable that has an effect on the dependent variable that is separate from the effect of the independent variable. One can easily imagine variables that would affect the chances of an adolescent woman's getting pregnant (the dependent variable for Kearney and Levine) that have nothing to do with her having watched, or not watched, the TV show *16 and Pregnant*. Whether or not friends are sexually active, and whether or not she defines herself as a lesbian, are two such variables. Sexual experience of her friendship group and sexual orientation, then, might be considered extraneous variables when considering the relationship between watching *16 and Pregnant* and becoming pregnant. One might map the relationship among these four variables in the following way:

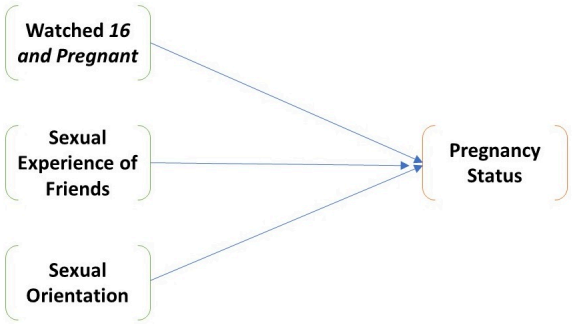


Figure 3. A Hypothetical Extraneous Variable Relationship

What Happens When You Control for a Variable and What Does it Mean?

You may be wondering how one could confirm any three-variable hypothesis with data. Let's look at an example using data from eight imaginary adolescent women, whether they watched *16 and Pregnant*, got pregnant, and sought information about contraception:

Case	Watched Show	Got Pregnant	Sought Contraception Information
1	Yes	No	Yes
2	Yes	No	Yes
3	Yes	No	Yes
4	Yes	Yes	Yes
5	No	Yes	No
6	No	Yes	No
7	No	Yes	No
8	No	No	No

Checking out a three-variable hypothesis requires, first, that you determine the relationship between the independent and dependent variables: in this case, between having watched *16 and Pregnant* and pregnancy status. Do you recall how to do that? In any case, we've done it in Table 1.

Table 1. Crosstabulation of Watching *16 and Pregnant* and Pregnancy Status

		Watched *16 and Pregnant*	
		Yes	*No*
Got Pregnant	*Yes*	1 (25%)	3 (75%)
	No	3 (75%)	1 (25%)
|Yule's Q|=0.80			

You'll note that the direction of this relationship, as expected by Kearny and Levine, is that women who had watched the show were less likely to get pregnant than those who had not. And a Yule's Q of 0.80 suggests the relationship is strong.

What **controlling a relationship** for another variable means is that one looks at the original relationship (in this case between watching the show and becoming pregnant) after eliminating variation in the control variable. We eliminate such variation by separating out the cases that fall into each category of the control variable, and examining the relationship between the independent and dependent variables in each category. In this case, what this means is that we first look at the relationship between watching and getting pregnant for those who have sought contraceptive information and then look at it for those who have not sought such information. To do this we create two more tables that have the same form as Table 1, one into which only those who fell into the "yes" category of having sought contraceptive information are put, the other into which only those cases that fell into the "no" category of having sought contraceptive information are put. Doing this, we've created two

more tables, Tables 2 and 3. Table 2 looks at the relationship between having watched the show and having become pregnant only for the four cases that sought contraceptive information; Table 3 does this only for the four cases that didn't seek contraceptive information.

Table 2 Crosstabulation of Watching *16 and Pregnant* and Pregnancy Status For Those Who Sought Contraceptive Information

		Watched *16 and Pregnant*	
		Yes	*No*
Got Pregnant	*Yes*	1	0
	No	3	0
|Yule's Q|=0.00			

Table 3 Crosstabulation of Watching *16 and Pregnant* and Pregnancy Status For Those Who Did Not Seek Contraceptive Information

		Watched *16 and Pregnant*	
		Yes	*No*
Got Pregnant	*Yes*	0	3
	No	0	1
|Yule's Q|=0.00			

We call relationship between an independent and dependent variable for the part of a sample that falls into one category of a control variable a **partial relationship** or simply a **partial**. What is notable about the partial relationships in both Table 4.2 and 4.3 is that they are as weak as they could possibly be (both Yule's Qs are equal to 0.00); both are much weaker than the original relationship between watching the show and becoming pregnant. In fact, in the context of controlling a relationship between two variables for a third, the relationship between the independent variable and the dependent variable, before the control, is often called an **original relationship**.

It may not surprise you to learn that controlling a relationship for a third variable does not always yield partials that are all weaker than the original. In fact, a famous methodologist, Paul Lazarsfeld (see Rosenberg, 1968), identified four distinct possibilities and others have called the resulting typology **the elaboration model**. **Elaboration**, in fact, is the term used by researchers for the process of controlling a relationship for a third variable. Table 4 outlines the basic characteristics of Lazarsfeld's four types of elaboration, with one more thrown in because, as we'll show, this fifth one is not only a logical, but also a practical, possibility.

Table 4. The Elaboration Model: Five Kinds of Elaboration

Type of Elaboration	Kind of Control Variable	Relationship of Partials to the Original
Interpretation	Intervening	All partials weaker than the original
Replication	Doesn't Matter	All partials about the same as the original
Explanation	Antecedent	All partials weaker than the original
Specification	Doesn't Matter	Some, but not all, partials different (stronger and/or weaker) than the original
Revelation	Doesn't Matter	All partials stronger than the original

Quiz at the end of the table: What kind of elaboration is demonstrated, in your view, in Tables 4.1 to 4.3?

You may recall that Kearney and Levine saw seeking contraceptive information as an intervening variable between the watching of *16 and Pregnant* and pregnancy status. Moreover, the partial relationships (shown in Tables 2 and 3) are both weaker than the original (shown in Table 1), so the elaboration shown in Tables 1 through 3 is an *interpretation*. If this kind of elaboration occurred in the real world, one could be pretty sure that seeking contraceptive information was indeed a mechanism through which watching the show affected a teenage woman's pregnancy status. Note: while the quantitative result in cases of interpretation and explanation are the same, the explanations for the processes at work are different, and this means the researcher must rely on their own knowledge of the variables at hand to determine which is at work. In cases of interpretation, an intervening variable is at work, and thus the relationship between the independent and dependent variables is a real relationship—it's just that the intervening variable is the mechanism through which this relationship occurs. In contrast, for cases of explanation, an antecedent variable is responsible for the apparent relationship between the independent and dependent variables, and thus this apparent relationship does not really exist. Rather, it is spurious.

In the real world, things don't usually work out quite as neatly as they did in this example, where an original relationship completely "disappears" in the partials. If one finds evidence of an *interpretation*, it's likely to be more subdued. Tables 5 and 6 demonstrate this point. Here, the researcher's (Roger's) hypothesis had been that people who are more satisfied with their finances are generally happier than people who are less satisfied. Table 5 uses General Social Survey (GSS) data to provide support for this hypothesis. Comparing percentages, about 47.5 percent of people who are satisfied with their finances claimed to be very happy, while only 17.3 percent who have claimed to be not at all satisfied with their finances said they were very happy. Moreover, a gamma of 0.42 suggests this relationship is in the hypothesized direction and that it is strong.

Table 5 Crosstabulation of Satisfaction with Finances and General Happiness, GSS Data from SDA

Frequency Distribution

Cells contain: –**Column percent** -Weighted N		**satfin** 1 pretty well satisfied	2 more or less satisfied	3 not satisfied at all	***ROW TOTAL***
happy	1: very happy	**47.5** 8,964.6	**30.4** 8,731.4	**17.3** 2,831.3	***32.1*** *20,527.3*
	2: pretty happy	**47.0** 8,874.7	**60.1** 17,228.1	**57.2** 9,345.6	***55.5*** *35,448.4*
	3: not too happy	**5.4** 1,023.9	**9.5** 2,721.7	**25.4** 4,148.7	***12.4*** *7,894.3*
	COL TOTAL	***100.0*** *18,863.2*	***100.0*** *28,681.2*	***100.0*** *16,325.6*	***100.0*** *63,870.0*
Means		1.58	1.79	2.08	1.80
Std Devs		.59	.60	.65	.64
Unweighted N		18,761	28,254	16,824	63,839

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 **Z**

N in each cell: Smaller than expected Larger than expected

Summary Statistics

Eta* =	.29	Gamma =	.42	Rao-Scott-P: F(4,2420) =	1,287.60	(p= 0.00)
R =	.29	Tau-b =	.26	Rao-Scott-LR: F(4,2420) =	1,232.07	(p= 0.00)
Somers' d* =	.25	Tau-c =	.24	Chisq-P(4) =	6,058.38	
				Chisq-LR(4) =	5,797.08	

*Row variable treated as the dependent variable.

Roger also introduced a control variable, the happiness of respondents' marriages, believing that this variable might be an intervening variable for the relationship between financial satisfaction and general happiness. In fact, he hypothesized that people who are more satisfied with their finances would be happier in their marriages than people who were not satisfied with their finances, and that happily married people would be more generally happy than people who are not happy in their marriages. In terms of the elaboration model, he was expecting that the relationships between financial satisfaction and general happiness for each part of the sample defined by a level of marital happiness (i.e., the partials) would be weaker than the original relationship between financial satisfaction and general happiness. And (hallelujah!) he was right. Table 6 shows that the relationship between financial satisfaction and general happiness for those with very happy marriages yielded a gamma of 0.35; for those with pretty happy marriages, 0.33; and for those with not too happy marriages, 0.31. All three of the partial relationships were weaker than the original, which we showed in Table 5 had a gamma of 0.43.

Table 6. Gammas for the Relationship Between Satisfaction with Finances and General Happiness for People with Different Degrees of Marital Happiness

Very Happy	Pretty Happy	Not Too Happy
0.34	0.33	0.31

Because the partials for each level of marital happiness are only somewhat weaker than the original relationship between financial satisfaction and general happiness, they don't suggest that marital satisfaction is the only reason for the relationship between financial satisfaction and general happiness, but they do suggest it is probably part of the reason. A curious researcher might look for others. But you get the idea: data can be used to shed light on the meaning of basic, two-variable relationships.

Perhaps more interesting still is that data can be used to resolve disputes about basic relationships. To illustrate, let's return to the "Ugly Children" study and Alberta, discussed in the chapter on bivariate analysis. One of the Alberta researchers, a Dr. Harrell, essentially said the fact that prettier children got better care than uglier children was causal: parents with prettier children are propelled by evolutionary forces, in his view, to protect their children (and, one assumes, parents of uglier children are not). A Dr. Sternberg, however, didn't see this relationship as causal. Instead, he saw it as the spurious result of wealth: wealthier parents can feed and clothe their kids better than others and are more likely to be caught up on supermarket etiquette associated with child care than others. Who's right?

One way one could check this out is by collecting and analyzing data. Suppose, for instance, that a researcher replicated the Alberta study (following parent/child dyads around supermarkets to determine the attractiveness of the children and how well they were cared for), but added observations about the cars the parent/child couples came in.

Late-model cars might be used as an indicator of relative wealth; beat-up dentmobiles (like Roger's), of relative poverty. Then one could see how much of the relationship between attractiveness and care "disappeared" in the parts of the sample that were defined by wealth and poverty. Suppose, in fact, the data so collected looked like this:

Table 7. Hypothetical Data to Test Dr. Sternberg's Hypothesis

Case	Attractiveness	Care	Wealth
1	Pretty	Good	Rich
2	Pretty	Good	Rich
3	Pretty	Good	Rich
4	Pretty	Bad	Rich
5	Ugly	Bad	Poor
6	Ugly	Bad	Poor
7	Ugly	Bad	Poor
8	Ugly	Good	Poor

Quiz about these data: Can you figure out the direction and strength of the relationship between the attractiveness of children and their care in this sample? What is the strength of this relationship within each category (rich and poor) of the control variable? What kind of elaboration did you uncover?

If you found that the original relationship was that pretty children got better care than ugly children (75% of the former did so, while only 25% of the latter did), you should be glowing with pride. If you found that the strength of the relationship (Yule's Q=0.80) was strong, your brilliance is even more evident. And if you found that this strength "disappeared" (Yule's Qs = 0.00) within each category of the wealth, you're a borderline genius. If you decided that the elaboration is an "explanation," because the partials are both weaker than the original and you've got an antecedent variable (at least according to Sternberg), you've crossed the border into genius.

Now referring back to the criteria for demonstrating causation (above), you'll note that the third criterion was that there must not be any antecedent variable that creates the relationship between the independent and dependent variables. What this means in terms of data analysis is that there can't be any antecedent variables whose control makes the relationship "disappear" within each of the parts of the sample defined by its categories. But that's exactly what has happened above. In other words, one can show that a relationship is non-causal (or spurious) by showing, through data, that there is an antecedent variable whose control, as in the example we've just been working with, makes the relationship "disappear." Pretty cool, huh?

On the other hand, while it's *impossible* to use data to show that a relationship *is* causal,[1] it is possible to show that any single third variable that others hypothesize is creating the relationship between the relevant independent and dependent variables isn't really creating that relationship. Thus, for example, Harrell and his Alberta team might have heard Sternberg's claim that the wealth of "families" is the real reason why the attractiveness and care of children are related. And if they'd collected data like the following, they could have shown this claim was false. Can you use the data to do so? See if you can analyze the data and figure out what kind of elaboration Harrell *et al.* would have discovered.

Table 8. Hypothetical Data to Test Dr. Harrell

1. The reason you can never show, through data analysis, that a two-variable relationship *is* causal is that for every two-variable relationship there are an infinite number of possible antecedent variables, and we just don't live long enough to test all the possibilities.

Case	Attractiveness	Care	Wealth
1	Pretty	Good	Rich
2	Pretty	Good	Rich
3	Ugly	Good	Rich
4	Ugly	Bad	Rich
5	Pretty	Good	Poor
6	Pretty	Good	Poor
7	Ugly	Good	Poor
8	Ugly	Bad	Poor

If you found that these data yielded a "replication," you're clearly on the brink of mastering the elaboration model. The original relationship between attractiveness and care was that pretty kids got better care than ugly kids (100% of pretty kids got it, compared to 50% of ugly kids who did) and this relationship was strong (Yule's Q= 1.00). But each of the partials was just as strong (Yule's Qs = 1.00), and had the same direction, as the original. What a replication shows is that the variable that was conceived of as an antecedent variable (wealth) does not "explain" the original relationship at all. The relationship is just as strong in each part of the sample defined by categories of the antecedent variable as it was before variation in this variable was controlled.

A Quick Word About Significance Levels

This chapter's focus on the elaboration model and controlling relationships has been all about making comparisons: primarily about comparing the strength of partial relationships to the strength of original relationships (but sometimes, as you'll soon see, comparing the strength of partials to one another). We haven't said a thing about comparing inferential statistics and the resulting information about whether one dare generalize from a sample to the larger population from which the sample has been drawn. This has been intentional. You may recall (from the chapter on bivariate analyses) that the magnitude of chi-square is directly related to the size of the sample: the larger the sample, given the same relationship, the greater the chi-square. When one controls a relationship between an independent and dependent variable, however, one is dividing the sample into at least two parts, and, depending on the number of categories of the control variable, potentially more. So comparing the chi-squares, and therefore the significance levels, of partials to that of an original is hardly a fair fight. The originals will always involve more cases than the partials. So we usually limit our comparisons to those of strength (and sometimes direction), though if a relationship loses its statistical significance when examining the partials this does mean that the relationship cannot necessarily be generalized in its partial form.

Having made this important point, however, we'll let you loose on the two quizzes that will end this chapter, each of which will introduce you to a new kind of elaboration.

Quiz #1 at the End of the Chapter

Show that the (hypothetical) sample data, below, conceivably collected to test Kearney and Levine's three-variable hypothesis (that adolescents who watched the show were more likely to seek contraceptive information than others, and that those who sought informa-

tion were less likely to get pregnant than others) are illustrative of a "specification." For which category of the control variable (sought contraceptive information) is the relationship between having watched *16 and Pregnant* and having gotten pregnant stronger? For which is it weaker? Why would such data NOT support Kearney and Levine's hypothesis?[2]

Case	Watched Show	Got Pregnant	Sought Contraceptive Information
1	Yes	No	Yes
2	Yes	No	Yes
3	No	Yes	Yes
4	No	Yes	Yes
5	Yes	Yes	No
6	Yes	No	No
7	No	Yes	No
8	No	No	No

Quiz #2 at the End of the Chapter

Suppose the data you collected to test Sternberg's hypothesis (that the relationship between the attractiveness of children and their care is a result of family wealth or social class) really looked like this ⇒

What kind of elaboration would you have uncovered? What makes you say so? (Doesn't it seem odd that partial relationships can be stronger than original relationships? But they sure can. That what Roger calls a "revelation.")

Case	Attractiveness	Care	Wealth
1	Pretty	Good	Rich
2	Pretty	Good	Rich
3	Ugly	Bad	Rich
4	Ugly	Bad	Rich
5	Pretty	Bad	Poor
6	Pretty	Bad	Poor
7	Ugly	Good	Poor
8	Ugly	Good	Poor

2. The original relationship, in this case, would be strong |Yule's Q|= 0.80. The partial relationship for those who had sought contraception, however, would be stronger (|Yule's Q|= 1.00, while that for those who had not sought contraception would be very weak (|Yule's Q|= 0.00). You can specify, therefore, that the original relationship is particularly strong for those who'd sought contraception and particularly weak for those who had not. Kearney and Levine's hypothesis had anticipated an "interpretation," but this data yield a specification. So the data would prove their hypothesis wrong.

Exercises

1. Write definitions, in your own words, for each of the following key concepts from this chapter:
    - multivariate analysis
    - antecedent variable
    - intervening variable
    - control variable
    - extraneous variable
    - spurious
    - original relationship
    - partial relationship
    - elaboration
    - interpretation
    - replication
    - explanation
    - specification
    - revelation
2. Below are real data from the GSS. See what you can make of them.

    Who do you think would be more fearful of walking in their neighborhoods at night: males or females? Recalling that gamma= Yule's Q for 2 x 2 tables, what does the following table, and its accompanying statistics, tell you about the actual direction and strength of the relationship? Support your answer with details from the table.

    Table 9. Crosstabulation of Gender (Sex) with Whether Respondent Reports Being Fearful of Walking in the Neighborhood at Night (Fear), GSS Data from SDA

Frequency Distribution

Cells contain: –Column percent -Weighted N		sex 1 male	2 female	*ROW TOTAL*
	1: yes	**22.2** 4,468.4	**51.6** 12,027.9	***38.0*** *16,496.2*
fear	2: no	**77.8** 15,620.2	**48.4** 11,294.1	***62.0*** *26,914.3*
	COL TOTAL	***100.0*** *20,088.6*	***100.0*** *23,322.0*	***100.0*** *43,410.6*
Means		1.78	1.48	1.62
Std Devs		.42	.50	.49
Unweighted N		19,213	24,158	43,371

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 **Z**

N in each cell: Smaller than expected Larger than expected

Summary Statistics

Eta* =	.30	Gamma =	-.58	Rao-Scott-P: F(1,590) =	1,776.84	(p= 0.00)
R =	-.30	Tau-b =	-.30	Rao-Scott-LR: F(1,590) =	1,828.11	(p= 0.00)
Somers' d* =	-.29	Tau-c =	-.29	Chisq-P(1) =	3,936.97	
				Chisq-LR(1) =	4,050.56	

*Row variable treated as the dependent variable.

We controlled this relationship for "race," a variable that had three categories: Whites, Blacks, and others. Suppose you learned that the gamma for this relationship among Whites was -0.61, among Blacks was -0.53 and among those identifying as members of other racial groups was -0.44. What kind of elaboration, in your view, would you have uncovered? Justify your answer.

3. Please read the article by Robert Bartsch *et al*., entitled "Gender Representation in Television Commercials: Updating an Update" (*Sex Roles,* Vol. 43, Nos. 9/10, 2000: 735-743).[3] What is the main point of the article, in your view? What is the significance, according to Bartsch *et al.,* of the gender of the voice-over in a commercial? Please examine Table 1 on page 739. Describe the overall gender breakdown of the voice-overs in 1998. Which gender was more represented in the voice-overs? Now look at the gender breakdown for voice-overs for domestic products and nondomestic products separately. Which of these is the stronger relationship: the one for domestic or the one for nondomestic products? What kind of elaboration would you say Bartsch *et al*. uncovered when they controlled the gender of voice-over for type of product (domestic or nondomestic)? How might you account for this finding?

3. If the link below doesn't work, perhaps you can hunt down an electronic copy of the article through your college's library service.

Media Attributions

- Figure 4.1
- Figure 4.2
- Diagramming an Extraneous Variable Relationship © Mikaila Mariel Lemonik Arthur

8. Correlation and Regression

ROGER CLARK

The chapter on bivariate analyses focused on ways to use data to demonstrate relationships between nominal and ordinal variables and the chapter on multivariate analysis on controling these relationships for other variables. This chapter will introduce you to the ways scholars show relationships between interval variables and control those relationships with other interval variables.

It turns out that the techniques presented in this chapter are by far the most likely ones you'll see used in research articles in the social sciences. There are a couple of reasons for this popularity. One is that the techniques we'll show you here are much less clumsy than the ones we showed you in prior chapters. The other is that, despite what we led you to believe in the chapter on univariate analysis (in our discussion of levels of measurement), all variables, whatever their level of measurement, can, via an ingenious method, be converted into interval-level variables. This method may strike you at first as having a very modest name for an ingenious method: **dummy variable** creation. Until you realize that *dummy* does not always refer to a dumb person—a dated and offensive expression in any case. Sometimes *dummy* refers to a "substitute for," as it does in this case.

Dummy Variables

In fact, a *dummy variable* is a two-category variable that is used as an ordinal or interval level variable. To understand how any variable, even a nominal-level variable can be treated as an ordinal or interval level variable, let's recall the definitions of ordinal and interval level variables.

An ordinal level variable is a variable whose categories can be ordered in some sensible way. The General Social Survey (GSS) measure of "general happiness" has three categories: very happy, happy, and not too happy. It's easy to see how these three categories can be ordered sensibly: "very happy" suggests more happiness than "happy," which in turn implies more happiness than "not too happy." But we'd normally say that the variable "gender," when limited to just two categories (female and male), is merely nominal. Neither category seems to have more of something than the other.

Not until you do a little conceptual blockbusting and think of the variable gender as a measure of either how much "maleness" or "femaleness" a person has. If we coded, as the GSS does, males as 1 and females as 2 we could say that a person's "gender," really "female-

ness," is greater any time a respondent gets coded 2 (or female) than when s/he gets coded 1 (or male).[1] Then one could, as we've done in Table 1, ask for a crosstabulation of *sex* (really "gender") and *happy* (really level of unhappiness) and see that females, generally, were a little happier than males in the U.S. in 2010, either by looking at the percentages or the gamma—a measure of relationship generally reserved for two ordinal level variables. The gamma for this relationship is -0.08, indicating that, in 2010, as femaleness went up, unhappiness went down. Pretty cool, huh?

Table 1. Crosstabulation of Gender (Sex) and Happiness (Happy), GSS data from SDA, 2010

Frequency Distribution

Cells contain:
–**Column percent**
-Weighted N

		sex 1 male	2 female	***ROW TOTAL***
happy	1: very happy	**26.7** 247.0	**29.8** 330.8	***28.4*** *577.8*
	2: pretty happy	**57.5** 532.3	**57.5** 639.3	***57.5*** *1,171.5*
	3: not too happy	**15.9** 146.9	**12.7** 141.0	***14.1*** *287.8*
	COL TOTAL	***100.0*** *926.1*	***100.0*** *1,111.0*	***100.0*** *2,037.1*
Means		1.89	1.83	1.86
Std Devs		.64	.63	.64
Unweighted N		890	1,149	2,039

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 **Z**

N in each cell: Smaller than expected Larger than expected

1. Professional statistical analysts usually use 0 and 1 rather than 1 and 2 when making dummy variables. This is due to the fact that the numbers used can impact the interpretation of the regression constant, which is not something beginning quantitative analysts need to worry about. Therefore, in this text, both approaches are used interchangeably.

Summary Statistics						
Eta* =	.05	Gamma =	-.09	Rao-Scott-P: F(2,156) =	1.95	(p= 0.15)
R =	-.05	Tau-b =	-.05	Rao-Scott-LR: F(2,156) =	1.94	(p= 0.15)
Somers' d* =	-.05	Tau-c =	-.05	Chisq-P(2) =	5.31	
				Chisq-LR(2) =	5.30	

*Row variable treated as the dependent variable.

We hope it's now clear why and how a two-category (dummy) variable can be used as an ordinal variable. But why and how can it be used as an interval variable? The answer to this question also lies in a definition: this time, of an interval level variable. An interval level variable, you may recall, is one whose adjacent categories are a standard or fixed distance from each other. For example, on the Fahrenheit temperature scale, we think of 32 degrees being the same distance from 33 degrees as 82 degrees is from 83 degrees. Returning to what we previously might have said was only a nominal-level variable, gender (using here just two categories: female and male), statisticians now ask us to ask: what is the distance between categories here. They answer: who really cares? Whatever it is, it's a standard distance because there's only one length of it to be covered. Every male, once coded, say, as a "1," is as far from the female category, once coded as a "2," as every other male. And *every* two-category (dummy) variable similarly consists of categories that are a "fixed" distance from each other. We hope this kind of conceptual blockbusting isn't as disorienting for you as it was for us when we first had to wrap our heads around it.

But this leaves the question of how "every" nominal-level variable can become an ordinal or interval level variable. After all, some nominal level variables have more than two categories. The GSS variable "labor force status" (wrkstat) has eight usable categories: working full time, working part time, temporarily not working, unemployed, retired, school, keeping house, and other.

But even this variable can become a two-category variable through recoding. Roger, for instance, was interested is seeing whether people who work fulltime were happier than other people, so he recoded so that there were only two categories: working full time and not working full time (wrkstat1). Then, using the Social Data Archive facility, he asked for the following crosstab (Table 2):

Table 2. Crossbulation of Whether a Respondent Works Fulltime (Wrkstat1) and Happiness (Happy), GSS data vis SDA

Frequency Distribution

		wrkstat1		
Cells contain: –**Column percent** -Weighted N		1 Working Full Time	2 Not Working Full Time	***ROW TOTAL***
happy	1: very happy	**33.2** 9,963.5	**32.8** 9,860.2	***33.0*** *19,823.8*
	2: pretty happy	**57.7** 17,300.1	**53.2** 15,986.4	***55.4*** *33,286.5*
	3: not too happy	**9.1** 2,738.6	**14.0** 4,223.8	***11.6*** *6,962.4*
	COL TOTAL	***100.0*** *30,002.2*	***100.0*** *30,070.5*	***100.0*** *60,072.7*
Means		1.76	1.81	1.79
Std Devs		.60	.66	.63
Unweighted N		29,435	30,604	60,039

Color coding: <-2.0 | <-1.0 | <0.0 | >0.0 | >1.0 | >2.0 **Z**

N in each cell: Smaller than expected Larger than expected

Summary Statistics

Eta* =	.04	Gamma =	.06	Rao-Scott-P: F(2,1132) =	120.24	(p= 0.00)
R =	.04	Tau-b =	.03	Rao-Scott-LR: F(2,1132) =	121.04	(p= 0.00)
Somers' d* =	.04	Tau-c =	.04	Chisq-P(2) =	368.93	
				Chisq-LR(2) =	371.39	

*Row variable treated as the dependent variable.

The gamma here (0.06) indicates that those working full time do tend to be happier than others, but that the relationship is a weak one.

We've suggested that dummy variables, because they are interval-level variables, can be used in analyses designed for interval-level variables. But we haven't yet said anything about analyses aimed at looking at the relationship between interval level variables. Now we will.

Correlation Analysis

The examination of relationships between interval-level variables is almost necessarily dif-

ferent from that of nominal or ordinal level variables. Doing crosstabulations of many interval level variables, for one thing, would involve very large numbers of cells, since almost every case would have its own distinct category on the independent and on the dependent variables.[2] Roger hypothesizes, for instance, that as the percentage of residents who own guns in a state rises, the number of gun deaths in a year per 100,000 residents would also increase. He just went to a couple of websites and downloaded information about both variables for all 50 states in 2017. Here's how the data look for the first six cases:

Table 3. Gun Ownership and Shooting Deaths by State, 2016[3]

State	Percent of Residents Who Own Guns	Gun Shooting Deaths Per 100,000 Population
Alabama	52.8	21.5
Alaska	57.2	23.3
Arizona	36	15.2
Arkansas	51.8	17.8
California	16.3	7.9
Colorado	37.9	14.3
...	...	...

One could of course recode all this information, so that each variable was reduced to two or three categories. For example, we could say any state whose number of gun shooting per 100,000 was less than 13 fell into a "low gun shooting category," and any state whose number was 13 or more fell into a "high gun shooting category." And do something like this for the percentage of residents who own guns as well. Then you could do a crosstabulation. But think of all the information that's lost in the process. Statisticians were dissatisfied with this solution and early on noticed that there was a better way than crosstabulation to depict the relationship between interval level variables.

They discovered a better way was to use what is called a *scatterplot*. A **scatterplot** is a visual depiction of the relationship between two interval level variables, the relationship between which is represented as points on a graph with an x-axis and a y-axis. Thus, Figure 1 shows the "scatter" of the states when plotted with the percent of residents who are gun owners along the x-axis and the gun shooting death per 100,000 along the y-axis.

2. Note that this would impact statistical significance, too, since there would be many cells in the table but few cases in each cell.
3. Gun ownership data from Schell *et al.* 2020; death data from National Center for Health Statistics 2022.

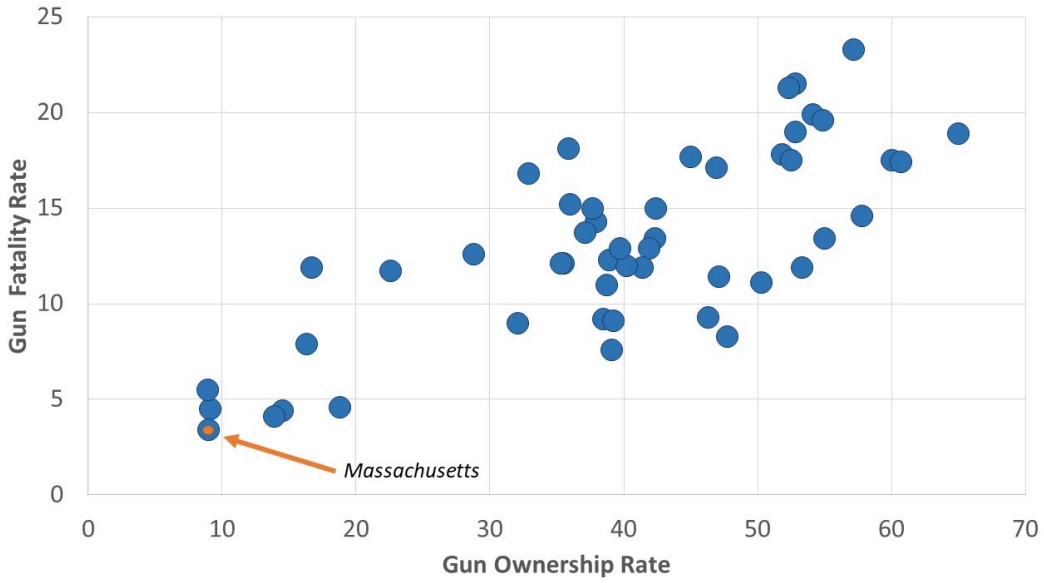


Figure 1 Scatterplot of Gun Ownership Rates and Per Capita Gun Deaths by State

Each state is a distinct point in this plot. We've pointed in the figure to Massachusetts, the state with the lowest value on each of the variables (9% percent of residents own guns and there are 3.4 gun deaths per 100,000 population), but each of the 50 states is represented by a point or dot on the graph. You'll note that, in general, as the gun ownership rate rises, the gun death rate does as well.

Karl Pearson (inventor of chi-square, you may recall) created a statistic, **Pearson's r**, which measures both the strength and direction of a relationship between two interval level variables, like the ones depicted in Figure 1. Like gamma, Pearson's r can vary between 1 and -1. The farther away Pearson's r is from zero, or the closer it is to 1 or -1, the stronger the relationship. And, like gamma, a positive Pearson's r indicates that as one variable increases, the other tends to as well. And this is the kind of relationship depicted in Figure 1: as gun ownership rises, the gun death rates tend to rise as well.

When the sign of Pearson's r (or simply "r") is negative, however, this means that as one variable rises in values, the other tends to fall in values. Such a relationship is depicted in Figure 2. Roger had expected that drug overdose death rates in states (measured as the number of deaths due to drug overdoses per 100,000 people) would be negatively associated with the percentage of states' residents reporting a positive sense of overall well being in 2016. Figure 2 provides visual support for this hypothesis. Note that while in Figure 1 the plot of points tends to move from bottom left to upper right on the graph (typical of positive relationships), the plot in Figure 2 tends to move from top left to bottom right (typical of negative relationships).

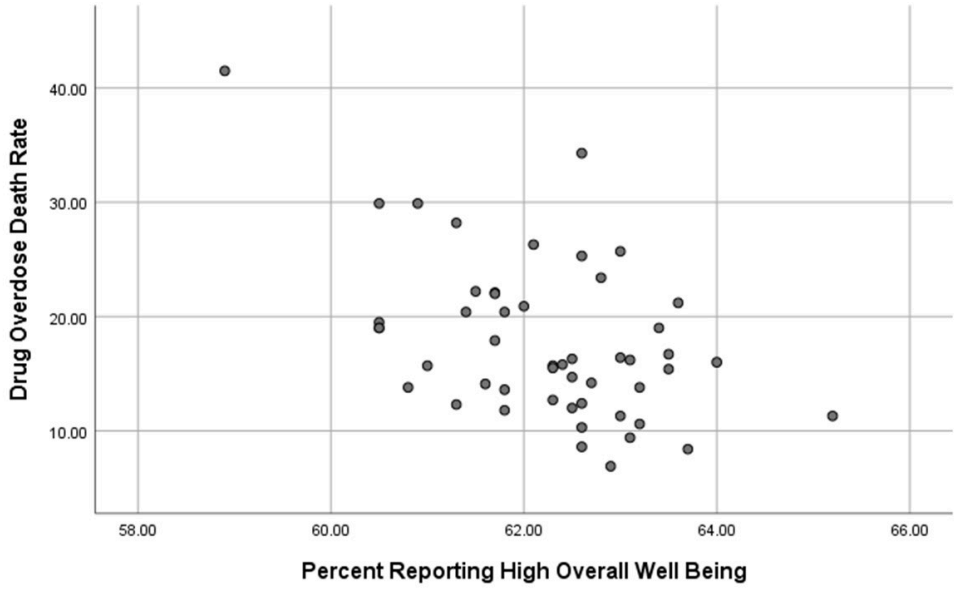


Figure 2. Scatterplot of the Relationship Between Drug Overdose Death Rates and Population Wellbeing, By State

A Pearson's r of 1.00 would not only mean that the relationship was as strong as it could be, that as one variable goes up, the other goes up, but also that all points fall on a line from bottom left to top right. A Pearson's r of -1.00 would mean that the relationship was as strong as it can be, that as one variable goes up, the other goes down, and that all points fall on a line from top left to bottom right. Figure 3 illustrates what various graphs producing various Pearson's r values would look like.

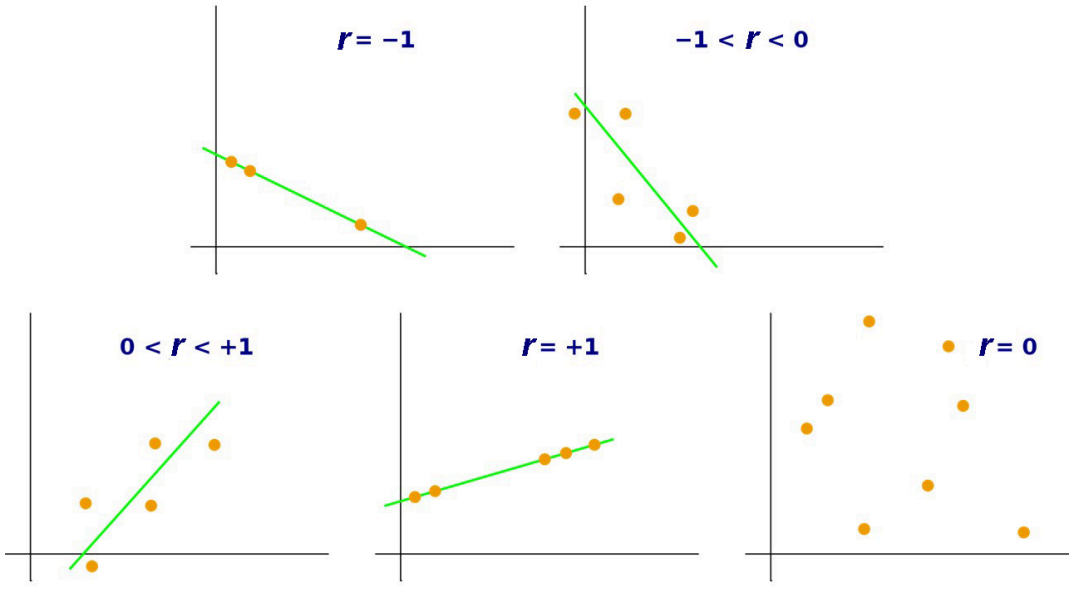


Figure 3. Examples of Scatterplots with Different Values of Pearson Correlation coefficient (r)

The formula for calculating Pearson's r is not much fun to use, but all contemporary com-

puters have no trouble adapting to systems that calculate it rapidly. The computer calculated the Pearson's r for the relationship between gun ownership and gun death rates for 50 states and it is indeed positive. Pearson's r is equal to 0.76. So it's a strong positive relationship. In contrast, the r for the relationship between the overall wellbeing of state residents and their overdose deaths rates is negative. The r turns out to be -0.50. So it's a strong negative relationship...though not quite as strong as the one for gun ownership and gun shootings. 0.76 is farther from zero than -0.50.

Most statistics packages will quickly calculate the several correlations very quickly as well. Roger asked SPSS to calculate the correlations among three variable characteristics of states: drug overdose death rates, percent of residents saying they have high levels of overall well being, and whether a state is in the southeast or southwest of the country. (Roger thought states in the southeast and southwest—the South, for short—might have higher rates of drug overdose deaths than other states.) The results of this request are shown in Table 3. This table shows what is called a correlation matrix and it's worth a moment of your time.

One reads the correlation between two variables by finding the intersection of the column headed by one of the variables and seeing where it crosses the row headed by the other variable. The top number in the resulting box is the Pearson correlation for the two variables. Thus, if one goes down the column in Table 3 headed by "Drug Overdose Death Rate" and sees where it crosses the row headed by "Percent Reporting High Overall Well Being" one see that their correlation is "-0.495," which rounds to -0.50. (Research reports always round correlation coefficients to two digits after the decimal point.) This is what we reported above.

Table 3. Correlations Among Drug Overdose Death Rates, Levels of Overall Well Being and Whether a State is in the American Southeast or Southwest

Correlations

		Drug Over-dose Death Rate	Percent Reporting High Overall Well Being	South or Other
Drug Overdose Death Rate	Pearson Correlation	1	-.495**	.094
	Sig. (2-tailed)		.000	.517
	N	50	50	50
Percent Reporting High Overall Well Being	Pearson Correlation	-.495**	1	-.351*
	Sig. (2-tailed)	.000		.012
	N	50	50	50
South or Other	Pearson Correlation	.094	-.351*	1
	Sig. (2-tailed)	.517	.012	
	N	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Quiz at the end of the table: What is the correlation between the drug overdose death rate and whether or not a state is in the South of the United States? And what does it mean?

If you answered, "I'm not sure," you're right! Whether a state is the South is a dummy variable: the state can be either in the South, on the one hand, or in the rest of the country, on the other. But since we haven't told you how this variable is coded, you couldn't possibly know what the Pearson's r of 0.09 means. But once we tell you that Southern states were coded 1 and all others were coded 0, you should be able to see that Southern states tended to have higher drug overdose rates than others, but that the relationship isn't very strong. Then you'll also realize that the Pearson's r relating region and overall well being (-0.35) suggests that overall well being tends to be lower in Southern states than in others.

One other thing is worth mentioning about a correlation matrix yielded by SPSS...and about Pearson's r's. If you look at the box (there are actually two such boxes; can you find

the other?) telling you correlation between overdose rates and overall well being, you'll see two other numbers in it. The bottom number (50) is of course the number of cases in the sample (there are, after all, 50 states). But the one in the middle (0.000) gives you some idea of the generalizability of the relationship (if there were, in fact, more states). A significance level or p-value of 0.000 does NOT mean there is no chance of making a Type 1 error (i.e., the error we make when we infer that a relationship exists in the larger population from which a sample is drawn when it does not), just that it's lower than can be shown in an SPSS printout. It does mean it is lower than 0.001 and therefore than 0.05, so inferring that such a relationship would exist in a larger population is reasonably safe. Karl Pearson was, after all, the inventor of chi-square and was always looking for inferential statistics. He found one in Pearson's r itself (imagine his surprise!) and figured out a way to use it to calculate the probability of making a Type 1 error (or p value) for values of r with various sample sizes. We don't need to show you how this is done but we do want you to marvel at this: Pearson's r is a measure of direction, strength, *and* generalizability of the relationship all wrapped into one.

There are several assumptions one makes when doing a correlation analysis of two variables. One, of course, is that both variables are interval-level. Another is that both are normally distributed. One can, with most statistical packages, do a quick check of the **skewness** of both variables. If the skewness of one or both is greater than 1.00 or less than -1.00, it is advisable to make a correction. Such corrections are pretty easy, but showing you how to do them is beyond our scope here. Roger did check on the variables in Table 5.3, found that the drug overdose rate was slightly skewed, corrected for the skewness, and found the correlations among the variables was very little changed.

A third assumption of correlation is that the relationship between the two variables is *linear*. A **linear relationship** is one in which a good description of its scatterplot is that it tends to conform to a straight line, rather than some other figure, like a U-shape or an upside-down U-shape. This seems to be true of the relationships shown in Figures 1 and 2. One can almost imagine that a line from bottom left to top right, for instance, is a pretty good way of describing the relationship in Figure 1, and we've done so in Figure 4. It's certainly not easy to see that any curved line would "fit" the points in that figure any better than a straight line does.

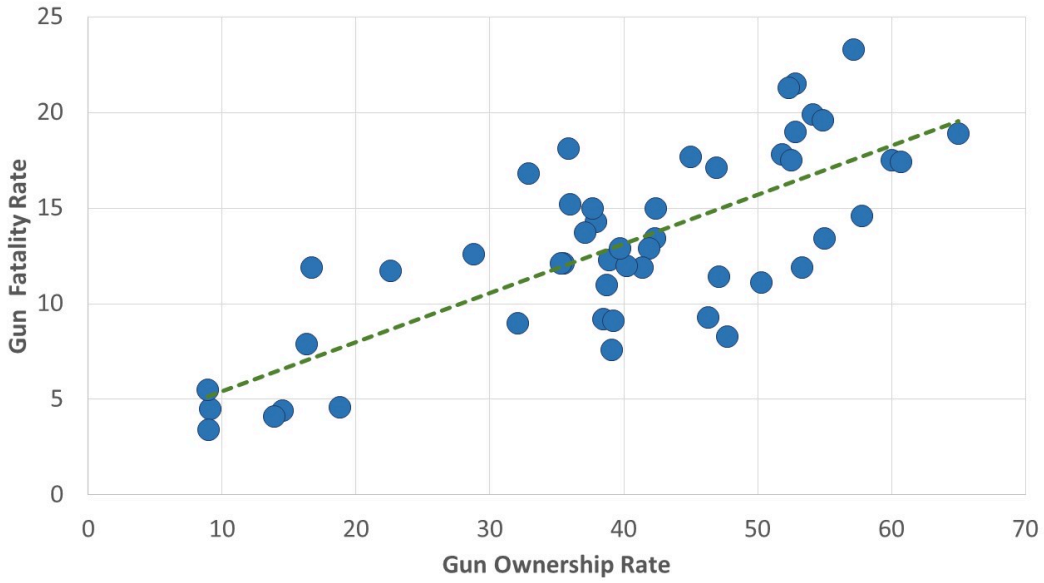


Figure 4. Scatterplot of Gun Ownership Rates and Per Capita Gun Deaths by State, with Trendline

Regression Analysis

But the assumption of a linear relationship raises the question of "Which line best describes the relationship?" It may not surprise you to learn that statisticians have a way of figuring out what that line is. It's called *regression*. **Regression** is a technique that is used to see how an interval-level dependent variable is affected by one or more interval-level independent variables. For the moment, we're going to leave aside the very tantalizing "or more" part of that definition and focus on the how regression analysis can provide even more insight into the relationship between two variables than correlation analysis does.

We call regression **simple linear regression** when we're simply examining the relationship between two variables. It's called **multiple regression** or **multivariate regression** when we're looking at the relationship between a dependent variable and more than one independent variable. Correlation, as we've said, can tell us about the strength, direction, and generalizability of the relationship between two interval level variables. Simple linear regression can tell us the same things, while adding information that can help us use an independent variable to predict values of a dependent variable. It does this by telling us the formula for the *line of best fit* for the points in a scatterplot. The **line of best fit** is a line that minimizes the distance between itself and all of the points in a scatterplot.

To get the flavor of this extra benefit of regression, we need to recall the formula for a line:

$$y = a + bx$$

where, in the case of regression, y refers to values of the dependent variable
x refers to values of the independent variable
a refers to the y-intercept, or where the line crosses the y-axis
b refers to the slope, or how much y increases every time x increases 1 unit

What simple linear regression does, in the first instance, is find the line that comes closest to all of the points in the scatterplot. Roger, for instance, used SPSS to do a regression of the gun shooting death rate by state on the percentage of residents who own guns (this is the vernacular used by statisticians: they regress the dependent variable on the independent variable[s]). Part of the resulting printout is shown in Table 4.

Table 4. Partial Printout from Request for Regression of Gun Shooting Death Rate Per 100,000 on Percentage of Residents Owning Guns

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	
		B	Std. Error	Beta		Sig.
1	(Constant)	2.855	1.338		2.134	.038
	Gun Ownership	.257	.032	.760	8.111	<.001

a. Dependent Variable: Gun Shooting Death Rate

Note that under the column labeled "B" under "Unstandardized Coefficients," one gets two numbers: 2.855 in the row labeled (Constant) and 0.257 in the row labeled "Gun Ownership." The (Constant) 2.855,[4] rounded to 2.86, is the y-intercept (the "a" in the equation above) for the line of best fit. The 0.257, rounded to 0.26, is the slope for that line. So what this regression tells us is that the line of best fit for the relationship between gun shooting deaths and gun ownership is:

*Gun shooting death rate = 2.86 + 0.26 * (% of residents owning guns)*

Correlation, we've noted, provides information about the strength, direction and generalizability of a relationship. By generating equations like this, regression gives you all those things (as you'll see in a minute), but also a way of predicting what (as-yet-incompletely-known) subjects will score on the dependent variable when one has knowledge of the their

4. This constant is the number that is impacted by whether we choose to code our dummy variable as 0 and 1 or as 1 and 2. As you can see, this choice impacts the equation of the line, but otherwise does not impact our interpretation of these results.

values on the independent variable. It permitted us, for instance, to draw the line of best fit into Figure 4, one that has a y-intercept of about 2.86 and a slope of about 0.26. And suppose you knew that about 50 percent of a state's residents owned guns. You could predict the gun death rate of the state by substituting "50" for the "% of residents owning guns" and get a prediction that:

Gun shooting death rate = 2.86 + 0.26 (50)= 2.86 + 13 = 15.86

Or that 15.86 per 100,000 residents will have experienced gun shooting deaths. Another output of regression is something called *R squared*. **R squared** x 100 (because we are converting from a decimal to a percentage) tells you the approximate percentage of variation in the dependent variables that is "explained" by the independent variable. "Explained" here is a slightly fuzzy term that can be thought of as referring to how closely points on a scatterplot comes to a line of best fit. In the case of the gun shooting death rate and gun ownership, the R squared is 0.578, meaning that about 58 percent of the variation in the gun shooting death rate can be explained by gun ownership rate. This is actually a fairly high percentage of variance explained, by sociology and justice studies standards, but would mean one's predictions using the regression formula are likely to be off by a bit, sometimes quite a bit.

The prediction bonus of regression is very profitable in some disciplines, like finance and investing. And predictions can even get pretty good in the social sciences if more variables are brought into play. We'll show you how this gets done in a moment, but first a word about how regression, like correlation, also provides information about the direction, strength and generalizability of a two-variable relationship. If you return to Table 5.4, you'll find in a column labeled "standardized coefficient" or *beta* (sometimes represented as β), the number 0.760. You may recall that the Pearson's r of the relationship between the gun shooting death rate and the percentage of residents who own guns was also 0.76, and that's no coincidence. The beta in simple regression is always the same as the Pearson's r for the bivariate relationship. Moreover, you'll find at the end of the row headed by "Gun Ownership" a significance level (<0.001)—which was exactly the same as the one for the original Pearson's r. In other words, through beta and the significance level associated with an independent variable we can, just as we could with Pearson's r, ascertain the direction, strength and generalizability of a relationship.

But beta's meaning is just a little different from Pearson's r's. **Beta** actually tells you the correlation between the relevant independent variable and the dependent variable when all other independent variables in the equation or model are controlled. That's a mouthful, we know, but it's a magical mouthful, as we're about to show you. In fact, the reason that the beta in the regression above is the same as the relevant Pearson's r is that there are no other independent variables involved. But let's now see what happens when there are....

Multiple Regression

Multiple regression (also called multivariate regression), as we've said before, is a technique that permits the examination of the relationship between a dependent variable and several independent variables. But to put it this way is somehow to diminish its magic. This magic is one reason that, of all the quantitative data analytic technique we've talked about in this book, multiple regression is probably the most popular among social researchers. Let's see why with a simple example.

Roger's taught classes in the sociology of gender and has long been interested in the question of why women are better represented in some countries' governments than in others. For example, why are women better represented in the national legislatures in many Scandinavian countries than they are, say, in the United States? In 2020, the United States achieved what was then a new high in female representation in the House of Representatives—23.4 percent of the House's seats were held by women after the election of the year before—while in Sweden 47 percent of the seats (almost twice as many) were held by women (Inter-Parliamentary Union 2022)[5]. He also knew that women were better represented in the legislatures of certain countries where some kind of quota for women's representation in politics had been established. Thus, in Rwanda, where a bitter civil war tore the country apart in 1994, a new leader, Paul Kagame, felt it wise to bring women into government and established a law that women should constitute at least 30 percent of all government decision-making bodies. In 2019, 61 percent of the members of Rwanda's lower house were women—by far the greatest percentage in the world and more than two and a half times as many as in the United States. Most Scandinavian countries also have quotas for women's representation.

In any case, Roger and three students (Rebecca Teczar, Katherine Rocha, and Joseph Palazzo), being good social scientists, wondered whether the effect of quotas might be at least partly attributable to cultural beliefs—say, to beliefs that men are better suited for politics than women. And, lo and behold, they found an international survey that measured such an attitude in more than 50 countries: the 2014 World Values Survey. They (Teczar *et al*. found that for those countries, the correlation between the presence of some kind of quota and the percentage of women in the national legislature was pretty strong (r =0.31), but that the correlation between the percentage of the population that thought men were better suited for politics and the presence of women in the legislature was even stronger (r= -0.46). Still, they couldn't be sure that the correlation of one of these independent variables with

5. When the U.S. Congress goes into session in 2023, the House of Representatives will be 28.5% women (Center for American Women in Politics 2022). Sweden has stayed about the same, while in Rwanda, women now make up 80% of members of the lower house (Inter-Parliamentary Union 2022).

the dependent variable wasn't at least partly due to the effects of the other variable. (Look out: we're about to use the language of the elaboration model outlined in the chapter on multivariate analysis.)

One possibility, for instance, was that the relationship between the presence of quotas and women's participation in legislatures was the spurious result of attitudes about women's (or men's) suitability for office on both the creation of quotas promoting their access to them and on the access itself. If this position had been proven correct, they would have discovered that there was an "explanation" for the relationship between quotas and women's representation. But they would have had to see the correlation between the presence of quotas and women's participation in legislatures drop considerably when attitudes were controlled for this position to be borne out.

On the other hand, it might have been that attitudes about women's suitability made it more (or less) likely that countries would adopt quotas, which in turn made it more likely that women would be elected to parliaments. Had the data supported this view, if, that is, the controlled association between attitudes and women's presence in parliaments dropped when the presence of quotas was controlled, we would have discovered an "interpretation" and might have interpreted the presence of quotas as the main way in which positive attitudes towards women in politics affected their presence in parliaments.

As it turns out, there was support, though nowhere near complete support, for both positions. Thus, the beta for the attitudinal question (-0.41) is slightly weaker than the original correlation (-0.46), suggesting that some of effect of cultural attitudes on women's parliamentary participation may be accounted for by their effects on the presence of quotas and the quotas' effects on participation. But the beta for the presence of quotas (0.23) is also weaker than its original correlation with women in parliaments (0.31), suggesting that some of its association with women in parliament may be due to the direct effects of attitudes on both the presence of quotas and on women in parliament. The R squared for this model (0.26) involving the two independent variables is considerably greater than it was for models involving each independent variable alone (0.20 for the attitudinal variable; 0.09 for the quota variable), so the two together explain more of the variance in women's presence in parliaments than either does alone.

But an R squared of 0.26 suggests that even if we used the formula that multiple regression gives us for predicting women's percentage of a national legislature from knowledge of whether a country had quotas for women and the percentage agreeing that men are better at politics, our prediction might not be all that good. That formula, though, is again provided by the numbers in the column headed by "B" under "Unstandardized Coefficients." That column yields the formula for a line in three-dimensional space, if you can imagine:

Fraction of Legislature that is Female = 0.286 + 0.067 (Presence of Quota) – 0.002 (Percentage Thinking Men More Suitable)

If a country had a quota and 10% of the population thought men were better suited for politics, we would predict that the fraction of the legislature that was female would be 0.286 + 0.067 (1) – 0.002 (10) = 0.333 or that 33.3 percent of the legislature would be female. Because such a prediction would be so imperfect, though, social scientists usually wouldn't make too much of it. It's frequently the case that sociologists and students of justice are more interested in multiple regression for its theory testing, rather than its predictive function.

Table 5. Regression of Women in the Legislature by Country on the Presence of a Quota for Women in Politics and The Percent of the Population Agreeing that Men Are More Suitable for Politics than Women

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.286	.049		5.828	.000
	Presence of a quota for women in politics	.067	.037	.228	1.819	.075
	Percent agreeing that men are more suitable for politics than women	-.002	.001	-.412	-3.289	.002

a. Dependent Variable: women in legislature 2017

These are the kinds of lessons one can learn from multiple regression. Two things here are worthy of note. First, the variable "presence of a quota for women in parliament" is a dummy variable treated in this analysis just as seriously as one would any other interval level variable. Second, we could have added any number of other independent variables into the model, as you'll see when you read the article referred to in Exercise 4 below. And any of them could have been a dummy variable. (We might, for instance, have included a dummy variable for whether a country was Scandinavian or not.) Multiple regression, in short, is a truly powerful, almost magical technique.

Exercises

1. Write definitions, in your own words, for each of the following key concepts from this chapter:
    - dummy variable

- scatterplot
- Pearson's r
- linear relationship
- regression
- line of best fit
- simple linear regression
- multiple regression
- R squared
- beta

2. Return to the Social Data Archive we've explored before. The data, again, are available at https://sda.berkeley.edu/ . (You may have to copy this address and paste it to request the website.) Again, go down to the second full paragraph and click on the "SDA Archive" link you'll find there. Then scroll down to the section labeled "General Social Surveys" and click on the first link there: General Social Survey (GSS) Cumulative Datafile 1972-2021 release.

   For this exercise, you'll need to come up with three hypotheses:
   1) Who do you think will have more offspring: older or younger adults?
   2) People with more education or less?
   3) Protestants or other Americans?

   Now you need to test these hypotheses from the GSS, using correlation analysis. To do this, you'll first need to make a dummy variable of religion. First, put "relig" in the "Variable Selection" box on the left and hit "View." How many categories does "relig" have? This is how to reduce those categories to just two. First, hit the "create variables" button at the upper left. Then, on the right, name the new variable something like "Protestant." (Some other student may have done this first. If so, you may want to use "their" variable.) The label for the new variable could be something like "Protestant or other." Then put "relig" in the "Name(s) of existing variables" box and click on the red lettering below. There should be a bunch of boxes down below. Put a "1" in the first box on the left, give the category a name like "Protestant," and put "1" for the Protestant category of "relig" on the right. Then go down one row and put "0" in the first box on the left in the row, label the category "other," and put "2-13" in the right-hand box of the row. This will put all other religions listed in "relig" in the "other" category of "Protestant." Then go to the bottom and hit "Start recoding." If no one else has done this yet, you should see a frequency distribution for your new variable. If someone else has done it, you may use their variable for the rest of this exercise.

   Now hit the "analysis" button at the upper left. Choose "Correl. Matrix" (for "correlation matrix") for the kind of analysis. Now put the four variables of interest for this exercise ("childs," "age," "educ," and "Protestant") in the first four "Variables to Correlate" boxes. Now go to the bottom and hit "Run correlations."

   Report the correlations between the three independent variables (age, educ and Protestant) and your dependent variable (childs). Do the correlations support your hypotheses? Which hypothesis receives the strongest support? Which the weakest? Were any of your hypotheses completely falsified by the analysis?

3. Now let's use the same data that we used in Exercise 1 to do a multiple regression analysis. You'll first need to leave the Social Data Archive and get back in again, returning to the GSS link. This time, instead of hitting "Correl. Matrix," hit "Regression." Then put "Childs" in the "Dependent" variable box and "Age," "Educ," and "Protestant" in three of the "Independent variables" boxes. Hit "Run regression." Which of the

independent variables retains the strongest association with the number of children a respondent has when all other variables in the model are controlled? What is that association? Which has the weakest when other variables are controlled?

4. Please read the following article:

Teczar, Rebecca, Katherine Rocha, Joseph Palazzo, and Roger Clark. 2018. "Cultural Attitudes towards Women in Politics and Women's Political Representation in Legislatures and Cabinet Ministries." *Sociology Between the Gaps: Forgotten and Neglected Topics* 4(1):1-7.

In the article, Teczar *et al*. use a multiple regression technique, called stepwise regression, which in this case only permits those variables that have a statistically significant (at the 0.05 level) controlled association into the model.

a. What variables do Teczar *et al*. find have the most significant controlled associations with women in national parliaments? Where do you find the relevant statistics in the article?

b. What variables do Teczar *et al*. find have the most significant controlled association with women in ministries? Where do you find the relevant statistics in the article?

c. Which model—the one for parliaments or the one for ministries (or cabinets)—presented in the article has the greater explanatory power? (i.e., which one explains more of the variation in the dependent variable?) How can you tell?

d. Do you agree with the authors' point (at the end) that political attitudes, while tough to change, are not unchangeable? Can you think of any contemporary examples not mentioned in the conclusion that might support this point?

Media Attributions

- Scatterplot of Gun Ownership Rates and Per Capita Gun Deaths by State © Mikaila Mariel Lemonik Arthur
- Scatterplot of the Relationship Between Drug Overdoes Death Rates and Population Wellbeing, By State © Roger Clark
- correlation-coefficients-1 © Kiatdd adapted by Mikaila Mariel Lemonik Arthur is licensed under a CC BY-SA (Attribution ShareAlike) license
- Scatterplot of Gun Ownership Rates and Per Capita Gun Deaths by State, with Trendline © Mikaila Mariel Lemonik Arthur

9. Presenting the Results of Quantitative Analysis

MIKAILA MARIEL LEMONIK ARTHUR

This chapter provides an overview of how to present the results of quantitative analysis, in particular how to create effective tables for displaying quantitative results and how to write quantitative research papers that effectively communicate the methods used and findings of quantitative analysis.

Writing the Quantitative Paper

Standard quantitative social science papers follow a specific format. They begin with a title page that includes a descriptive title, the author(s)' name(s), and a 100 to 200 word abstract that summarizes the paper. Next is an introduction that makes clear the paper's research question, details why this question is important, and previews what the paper will do. After that comes a literature review, which ends with a summary of the research question(s) and/or hypotheses. A methods section, which explains the source of data, sample, and variables and quantitative techniques used, follows. Many analysts will include a short discussion of their descriptive statistics in the methods section. A findings section details the findings of the analysis, supported by a variety of tables, and in some cases graphs, all of which are explained in the text. Some quantitative papers, especially those using more complex techniques, will include equations. Many papers follow the findings section with a discussion section, which provides an interpretation of the results in light of both the prior literature and theory presented in the literature review and the research questions/hypotheses. A conclusion ends the body of the paper. This conclusion should summarize the findings, answering the research questions and stating whether any hypotheses were supported, partially supported, or not supported. Limitations of the research are detailed. Papers typically include suggestions for future research, and where relevant, some papers include policy implications. After the body of the paper comes the works cited; some papers also have an Appendix that includes additional tables and figures that did not fit into the body of the paper or additional methodological details. While this basic format is similar for papers regardless of the type of data they utilize, there are specific concerns relating to quantitative research in terms of the methods and findings that will be discussed here.

Methods

In the methods section, researchers clearly describe the methods they used to obtain and analyze the data for their research. When relying on data collected specifically for a given paper, researchers will need to discuss the sample and data collection; in most cases, though, quantitative research relies on pre-existing datasets. In these cases, researchers need to provide information about the dataset, including the source of the data, the time it was collected, the population, and the sample size. Regardless of the source of the data, researchers need to be clear about which variables they are using in their research and any transformations or manipulations of those variables. They also need to explain the specific quantitative techniques that they are using in their analysis; if different techniques are used to test different hypotheses, this should be made clear. In some cases, publications will require that papers be submitted along with any code that was used to produce the analysis (in SPSS terms, the syntax files), which more advanced researchers will usually have on hand. In many cases, basic descriptive statistics are presented in tabular form and explained within the methods section.

Findings

The findings sections of quantitative papers are organized around explaining the results as shown in tables and figures. Not all results are depicted in tables and figures—some minor or null findings will simply be referenced—but tables and figures should be produced for all findings to be discussed at any length. If there are too many tables and figures, some can be moved to an appendix after the body of the text and referred to in the text (e.g. "See Table 12 in Appendix A").

Discussions of the findings should not simply restate the contents of the table. Rather, they should explain and interpret it for readers, and they should do so in light of the hypothesis or hypotheses that are being tested. Conclusions—discussions of whether the hypothesis or hypotheses are supported or not supported—should wait for the conclusion of the paper.

Creating Effective Tables

When creating tables to display the results of quantitative analysis, the most important goals are to create tables that are clear and concise but that also meet standard conven-

tions in the field. This means, first of all, paring down the volume of information produced in the statistical output to just include the information most necessary for interpreting the results, but doing so in keeping with standard table conventions. It also means making tables that are well-formatted and designed, so that readers can understand what the tables are saying without struggling to find information. For example, tables (as well as figures such as graphs) need clear captions; they are typically numbered and referred to by number in the text. Columns and rows should have clear headings. Depending on the content of the table, formatting tools may need to be used to set off header rows/columns and/or total rows/columns; cell-merging tools may be necessary; and shading may be important in tables with many rows or columns.

Here, you will find some instructions for creating tables of results from descriptive, crosstabulation, correlation, and regression analysis that are clear, concise, and meet normal standards for data display in social science. In addition, after the instructions for creating tables, you will find an example of how a paper incorporating each table might describe that table in the text.

Descriptive Statistics

When presenting the results of descriptive statistics, we create one table with columns for each type of descriptive statistic and rows for each variable. Note, of course, that depending on level of measurement only certain descriptive statistics are appropriate for a given variable, so there may be many cells in the table marked with an — to show that this statistic is not calculated for this variable. So, consider the set of descriptive statistics below, for occupational prestige, age, highest degree earned, and whether the respondent was born in this country.

Table 1. SPSS Ouput: Selected Descriptive Statistics

Statistics

		R's occupational prestige score (2010)	Age of respondent
N	Valid	3873	3699
	Missing	159	333
Mean		46.54	52.16
Median		47.00	53.00
Std. Deviation		13.811	17.233
Variance		190.745	296.988
Skewness		.141	.018
Std. Error of Skewness		.039	.040
Kurtosis		-.809	-1.018
Std. Error of Kurtosis		.079	.080
Range		64	71
Minimum		16	18
Maximum		80	89
Percentiles	25	35.00	37.00
	50	47.00	53.00
	75	59.00	66.00

Statistics

R's highest degree

N	Valid	4009
	Missing	23
Median		2.00
Mode		1
Range		4
Minimum		0
Maximum		4

R's highest degree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	less than high school	246	6.1	6.1	6.1
	high school	1597	39.6	39.8	46.0
	associate/junior college	370	9.2	9.2	55.2
	bachelor's	1036	25.7	25.8	81.0
	graduate	760	18.8	19.0	100.0
	Total	4009	99.4	100.0	
Missing	System	23	.6		
Total		4032	100.0		

Statistics

Was r born in this country

N	**Valid**	3960
	Missing	72
Mean		1.11
Mode		1

Was r born in this country

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**yes**	3516	87.2	88.8	88.8
	no	444	11.0	11.2	100.0
	Total	3960	98.2	100.0	
Missing	**System**	72	1.8		
Total		4032	100.0		

To display these descriptive statistics in a paper, one might create a table like Table 2. Note that for discrete variables, we use the value label in the table, not the value.

Table 2. Descriptive Statistics

	Occupational Prestige Score	Age	Highest Degree Earned	Born in This Country?
Mean	46.54	52.16	—	1.11
Median	47	53	1: Associates (9.2%)	1: Yes (88.8%)
Mode	—	—	2: High School (39.8%)	—
Standard Deviation	13.811	17.233	—	—
Variance	190.745	296.988	—	—
Skewness	0.141	0.018	—	—
Kurtosis	-0.809	-1.018	—	—
Range	64 (16-80)	71 (18-89)	Less than High School (0) – Graduate (4)	—
Interquartile Range	35-59	37-66	—	—
N	3873	3699	4009	3960

If we were then to discuss our descriptive statistics in a quantitative paper, we might write something like this (note that we do not need to repeat every single detail from the table, as readers can peruse the table themselves):

> This analysis relies on four variables from the 2021 General Social Survey: occupational prestige score, age, highest degree earned, and whether the respondent was born in the United States. Descriptive statistics for all four variables are shown in Table 2. The median occupational prestige score is 47, with a range from 16 to 80.

50% of respondents had occupational prestige scores scores between 35 and 59. The median age of respondents is 53, with a range from 18 to 89. 50% of respondents are between ages 37 and 66. Both variables have little skew. Highest degree earned ranges from less than high school to a graduate degree; the median respondent has earned an associate's degree, while the modal response (given by 39.8% of the respondents) is a high school degree. 88.8% of respondents were born in the United States.

Crosstabulation

When presenting the results of a crosstabulation, we simplify the table so that it highlights the most important information—the column percentages—and include the significance and association below the table. Consider the SPSS output below.

Table 3. R's highest degree * R's subjective class identification Crosstabulation

			R's subjective class identification			
			lower class	working class	middle class	upper class
R's highest degree	less than high school	Count	65	106	68	7
		% within R's subjective class identification	18.8%	7.1%	3.4%	4.2%
	high school	Count	217	800	551	23
		% within R's subjective class identification	62.9%	53.7%	27.6%	13.9%
	associate/junior college	Count	30	191	144	3
		% within R's subjective class identification	8.7%	12.8%	7.2%	1.8%
	bachelor's	Count	27	269	686	49
		% within R's subjective class identification	7.8%	18.1%	34.4%	29.5%
	graduate	Count	6	123	546	84
		% within R's subjective class identification	1.7%	8.3%	27.4%	50.6%
Total		Count	345	1489	1995	166
		% within R's subjective class identification	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	819.579[a]	12	<.001
Likelihood Ratio	839.200	12	<.001
Linear-by-Linear Association	700.351	1	<.001
N of Valid Cases	3995		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.22.

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Interval by Interval	**Pearson's R**	.419	.013	29.139	<.001[c]
Ordinal by Ordinal	**Spearman Correlation**	.419	.013	29.158	<.001[c]
N of Valid Cases		3995			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Table 4 shows how a table suitable for include in a paper might look if created from the SPSS output in Table 3. Note that we use asterisks to indicate the significance level of the results: * means p < 0.05; ** means p < 0.01; *** means p < 0.001; and no stars mean p > 0.05 (and thus that the result is not significant). Also note than N is the abbreviation for the number of respondents.

		Respondent's Subjective Class Identification				
		Lower Class	**Working Class**	**Middle Class**	**Upper Class**	**Total**
Highest Degree Earned	**Less than High School**	18.8%	7.1%	3.4%	4.2%	6.2%
	High School	62.9%	53.7%	27.6%	13.9%	39.8%
	Associate's / Junior College	8.7%	12.8%	7.2%	1.8%	9.2%
	Bachelor's	7.8%	18.1%	34.4%	29.5%	25.8%
	Graduate	1.7%	8.3%	27.4%	50.6%	19.0%
	N: 3995 Spearman Correlation 0.419***					

If we were going to discuss the results of this crosstabulation in a quantitative research paper, the discussion might look like this:

> A crosstabulation of respondent's class identification and their highest degree earned, with class identification as the independent variable, is significant, with a Spearman correlation of 0.419, as shown in Table 4. Among lower class and working class respondents, more than 50% had earned a high school degree. Less than 20% of poor respondents and less than 40% of working-class respondents had earned

more than a high school degree. In contrast, the majority of middle class and upper class respondents had earned at least a bachelor's degree. In fact, 50% of upper class respondents had earned a graduate degree.

Correlation

When presenting a correlating matrix, one of the most important things to note is that we only present half the table so as not to include duplicated results. Think of the line through the table where empty cells exist to represent the correlation between a variable and itself, and include only the triangle of data either above or below that line of cells. Consider the output in Table 5.

Table 5. SPSS Output: Correlations

		Age of respondent	R's occupational prestige score (2010)	Highest year of school R completed	R's family income in 1986 dollars
Age of respondent	**Pearson Correlation**	1	.087**	.014	.017
	Sig. (2-tailed)		<.001	.391	.314
	N	3699	3571	3683	3336
R's occupational prestige score (2010)	**Pearson Correlation**	.087**	1	.504**	.316**
	Sig. (2-tailed)	<.001		<.001	<.001
	N	3571	3873	3817	3399
Highest year of school R completed	**Pearson Correlation**	.014	.504**	1	.360**
	Sig. (2-tailed)	.391	<.001		<.001
	N	3683	3817	3966	3497
R's family income in 1986 dollars	**Pearson Correlation**	.017	.316**	.360**	1
	Sig. (2-tailed)	.314	<.001	<.001	
	N	3336	3399	3497	3509

**. Correlation is significant at the 0.01 level (2-tailed).

Table 6 shows what the contents of Table 5 might look like when a table is constructed in a fashion suitable for publication.

Table 6. Correlation Matrix

	Age	Occupational Prestige Score	Highest Year of School Completed	Family Income in 1986 Dollars
Age	1			
Occupational Prestige Score	0.087***	1		
Highest Year of School Completed	0.014	0.504***	1	
Family Income in 1986 Dollars	0.017	0.316***	0.360***	1

If we were to discuss the results of this bivariate correlation analysis in a quantitative paper, the discussion might look like this:

> Bivariate correlations were run among variables measuring age, occupational prestige, the highest year of school respondents completed, and family income in constant 1986 dollars, as shown in Table 6. Correlations between age and highest year of school completed and between age and family income are not significant. All other correlations are positive and significant at the $p<0.001$ level. The correlation between age and occupational prestige is weak; the correlations between income and occupational prestige and between income and educational attainment are moderate, and the correlation between education and occupational prestige is strong.

Regression

To present the results of a regression, we create one table that includes all of the key information from the multiple tables of SPSS output. This includes the R^2 and significance of the regression, either the B or the beta values (different analysts have different preferences here) for each variable, and the standard error and significance of each variable. Consider the SPSS output in Table 7.

Table 7. SPSS Output: Regression

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.395[a]	.156	.155	36729.04841

a. Predictors: (Constant), Highest year of school R completed, Age of respondent, R's occupational prestige score (2010)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	805156927306.583	3	268385642435.528	198.948	<.001[b]
	Residual	4351948187487.015	3226	1349022996.741		
	Total	5157105114793.598	3229			

a. Dependent Variable: R's family income in 1986 dollars

b. Predictors: (Constant), Highest year of school R completed, Age of respondent, R's occupational prestige score (2010)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	**(Constant)**	-44403.902	4166.576		-10.657	<.001		
	Age of respondent	9.547	38.733	.004	.246	.805	.993	1.00
	R's occupational prestige score (2010)	522.887	54.327	.181	9.625	<.001	.744	1.34
	Highest year of school R completed	3988.545	274.039	.272	14.555	<.001	.747	1.33

a. Dependent Variable: R's family income in 1986 dollars

The regression output in shown in Table 7 contains a lot of information. We do not include *all* of this information when making tables suitable for publication. As can be seen in Table 8, we include the Beta (or the B), the standard error, and the significance asterisk for each variable; the R^2 and significance for the overall regression; the degrees of freedom (which tells readers the sample size or N); and the constant; along with the key to p/significance values.

Table 8. Regression Results for Dependent Variable Family Income in 1986 Dollars

	Beta & SE
Age	0.004 (38.733)
Occupational Prestige Score	0.181*** (54.327)
Highest Year of School Completed	0.272*** (274.039)
R^2	*0.156***
Degrees of Freedom	3229
Constant	-44,403.902

** p<0.05 **p<0.01 ***p<0.001*

If we were to discuss the results of this regression in a quantitative paper, the results might look like this:

> Table 8 shows the results of a regression in which age, occupational prestige, and highest year of school completed are the independent variables and family income is the dependent variable. The regression results are significant, and all of the independent variables taken together explain 15.6% of the variance in family income. Age is not a significant predictor of income, while occupational prestige and educational attainment are. Educational attainment has a larger effect on family income than does occupational prestige. For every year of additional education attained, family income goes up on average by $3,988.545; for every one-unit increase in occupational prestige score, family income goes up on average by $522.887.[1]

Exercises

1. Choose two discrete variables and three continuous variables from a dataset of your choice. Produce appropriate descriptive statistics on all five of the variables and create a table of the results suitable for inclusion in a paper.
2. Using the two discrete variables you have chosen, produce an appropriate crosstabulation, with significance and measure of association. Create a table of the results suitable for inclusion in a

1. Note that the actual numberical increase comes from the B values, which are shown in the SPSS output in Table 7 but not in the reformatted Table 8.

paper.

3. Using the three continuous variables you have chosen, produce a correlation matrix. Create a table of the results suitable for inclusion in a paper.
4. Using the three continuous variables you have chosen, produce a multivariate linear regression. Create a table of the results suitable for inclusion in a paper.
5. Write a methods section describing the dataset, analytical methods, and variables you utilized in questions 1, 2, 3, and 4 and explaining the results of your descriptive analysis.
6. Write a findings section explaining the results of the analyses you performed in questions 2, 3, and 4.

SECTION III

QUALITATIVE DATA ANALYSIS

10. The Qualitative Approach

The Qualitative Approach

MIKAILA MARIEL LEMONIK ARTHUR

At the most basic level, **qualitative** research is research that emphasizes data that is not numerical in nature, data like words, pictures, and ideas. In contrast, **quantitative** data emphasizes numbers, or at least variables that can relatively easily be translated into numerical terms. In other words, quantitative data is about quantities, while qualitative data is about qualities. Beyond this basic distinction, qualitative research can look very similar to quantitative research or it can take a very different approach. Later in this chapter, you will learn more about different ways of thinking about data and how they might apply to qualitative data analysis. When people talk about qualitative approaches to research, however, they are often focused on those approaches that are distinct from what quantitative researchers do.

So what are some of the unique features of qualitative research? First of all, qualitative research tends to rely on the use of rich, thick description. In other words, qualitative research does not just provide summaries of data and findings, it really takes the reader or consumer of research there and lets them explore the situation and make conclusions for themselves by drawing on extended descriptions and excerpts from the data. Qualitative research also leaves room for focus on feelings and emotions, elements of the social world that can be harder to get at with quantitative data. For the qualitative researcher, data should not just depict specific actions or occurrences, but rather the contexts and backgrounds that lead up to what happened. More broadly, qualitative research tends to focus on a deep understanding of a specific place or organization or of a particular issue, rather than providing a wider but shallower understanding of an area of study.

Among the strengths of qualitative research are that it provides for the development of new theories and the exploration of issues that people do not know much about. It is very high in validity since it is so connected to real life. And it permits the collection and analysis of more detailed, contextual, and complex kinds of data and information. Of course, with strengths come limitations. The higher validity of qualitative data is matched with lower reliability due to the unique circumstances of data collection and the impact of interviewer effect. The greater ability to develop new theories is matched with a greater difficulty testing existing theories, especially causal ones, given the impossibility of eliminating alternative explanations for the phenomena under investigation. The ability to collect more detailed and complex information comes in large part due to the focus on a much smaller number of participants or cases, which in turn limits generalizability and in some cases can

limit representativeness. And while there is no reason to conclude that any of these factors make qualitative research more prone to bias than quantitative research, which after all can be profoundly impacted by slight variations in survey question wording or sample design, those who are not well informed about research methodology may discount the strengths of qualitative research by suggesting that the lack of numbers or the close interaction between participants and researchers bias the results.

In their classic text on qualitative data analysis, Miles and Huberman (1994) present the following as among the key elements of qualitative data analysis:

- It involves more prolonged contact with more ordinary aspects of human life;
- It has a holistic rather than a particularistic focus, aiming to keep data and findings in context;
- Multiple interpretations and understandings of data are possible, and researchers should preserve respondents' own understandings of their worlds and lives;
- There is a lack of standardization and measurement, with the researcher themselves becoming the primary measurement instrument; and
- Analysis is done primarily with words.

For the purposes of this text on data analysis, which focuses on what we do after we collect data rather than on how we go about obtaining the data in the first place, the last of these elements is most important. However, other scholars would argue that qualitative analysis is not limited to words—it may also involve visuals ways of engaging with and presenting data.

Types of Qualitative Data

The data that we analyze in qualitative research consist primarily of words and images drawn from observation, interaction, interviewing, or existing documents. In particular, the types of data collection that tend to result in qualitative data include interviews and focus groups, ethnography and participant observation, and the analysis of existing documents. These different data collection strategies imply a variety of analytical strategies as well, and indeed qualitative data analysis relies on a breadth of techniques. Thus, part of the process of formulating and selecting qualitative data is selecting the right kinds of strategies to apply to the particular data being utilized.

One of the most common ways in which qualitative data is collected is through talking to people. We often refer to these people as **respondents** or **participants**. Sometimes, they may be called subjects, though many qualitative researchers find that term to be inappro-

priate. In contrast to respondents or participants, subjects implies a more passive kind of relationship to the research process, a relationship in which research is *done to* a person rather than one in which a person is a party to the research process.

Research involving talking to people usually involves **interviews** of various kinds, whether they be in-person or via video chat, short and structured or long oral histories, of an individual or of a larger **focus group**. The data collected from interviews may include interview notes and audio or video recordings. Alternatively, researchers may conduct **observational research** or **participant-observation** (often called **ethnography**). In this method, researchers observe real social life in all its detail, either with or without participating in it. Typically, the data collected from observation and ethnography entails detailed **fieldnotes** recording what has been encountered in the setting.

It is beyond the scope of this text to discuss the process of data collection. However, the next chapter will detail some of the strategies that researchers may want to consider in designing their studies and collecting their data in order to ensure that data is obtained in a form that is useful for analysis.

There are other kinds of qualitative data that do not involve talking to people. These include **trace analysis**, or observing the traces of life that people have left behind (this is what archeologists do), as well as the use of existing documents or images as a data source. For example, researchers might collect social media posts, photographs of social events, newspaper articles, or **archival** materials like letters, journals, and meeting minutes.

Paradigms of Research

Researchers approach their research from different perspectives or paradigms. A **paradigm** is a set of assumptions, values, and practices that shapes the way that people see, understand, and engage with the world, and thus the particular paradigm that a researcher inhabits shapes the fashion in which they carry out their research. Philosophers use the term **epistemology** to refer to the study of the nature of knowledge, and thus we can take an epistemological perspective to understanding how paradigms of research might vary.

Two paradigms that commentators often juxtapose are the **positivist** and **interpretivist** approaches. Positivism assumes that there is a real, verifiable reality and that the purpose of research is to come as close to it as possible. Thus, a positivist would argue that we can understand the world, subject it to prediction and control, and—through the processes of research and data analysis—empirically verify our claims. Positivist research projects can utilize a variety of methods, but experimental and quantitative survey data are especially likely. Among qualitative approaches, positivism is often associated with the type of observational study once common in anthropology, which aimed at uncovering the "real" social

practices of a group. These methods tend to involve keeping some degree of distance between the researcher and the participants and positioning the researcher as the expert on both research methods and the participants' own lives. From a positivist perspective, standards of rigor like reliability, validity, and generalizability are important and attainable markers of good research as they contribute to the likelihood that the research arrives at the right answer. As this suggests, **objectivity** is an essential goal of positive research. Good research, to a positivist, is that which is valid, reliable, generalizable, and has strong, significant results.

In contrast, interpretivism suggests that our knowledge of the world is created by our own individual experiences and interactions, and thus that reality cannot be understood as existing on its own in a form separate from our distinct existences. Thus, an interpretivist would argue that understandings are always based in a particular time and on a particular interpreter and are always open to reinterpretation. Interpretivist research projects utilize naturalistic research methods that are rooted in real social contexts, especially in-depth interviewing and participant-observation. These methods tend to involve a closer and more reciprocal relationship between the researcher and the participants, with a greater concern for ethical treatment and in some cases an emphasis on possibilities for social change. Interpretivist researchers also value participants' expertise and their understandings of their own lives rather than assuming the researcher's perspective is necessarily more accurate. From an interpretivist perspective, validity may not be attainable due to the fact that truth is not certain, and in any case standards of rigor are far less important than considerations like ethics, morality, the degree to which biases are made clear, and what the world can learn from the research. As this might suggest, interpretivists would tend to believe that objectivity is probably not attainable, and that even it is, the pursuit of it may not be worthwhile. To an interpretivist, good research is that which is done in a careful, respectful manner, contributes to knowledge, is reflective, and takes appropriate political and ethical considerations into account.

Lisa Pearce (2012) has outlined a paradigm she calls **pragmatist**. This approach is sometimes understood as a kind of middle position between positivist and interpretivist ways of thinking. Thus, its proponents neither believe that strict objectivity is possible nor abandon efforts to seek objectivity at all, instead engaging in reflexivity as they consider how researchers influence both research participants and research findings. While pragmatist approaches can be used with various methods of data collection, they tend to be employed by those using mixed-methods approaches, especially those combining quantitative and quantitative strategies.

Another paradigm of research is **feminist** in nature. While there are of course many ways to do research from a feminist perspective, one of the most important elements of feminist epistemology is the idea that everyone comes to research—whether they are a researcher, a research participant, or a consumer of research—from their own **standpoint**.

In other words, each person's individual life experiences and social positions shape their point of view on the world, and this point of view will in turn impact how the individual understands and interprets phenomena they encounter, including those that are part of research. Dorothy Smith (1987), one of the figures associated with feminist standpoint approaches, notes that this approach to methods requires that we be able to describe the social "in ways that can be checked back to how it actually is" (1987:122). Such approaches are powerful not only for understanding the experiences of women, but also for understanding the experiences of other minoritized, marginalized, and/or oppressed groups, including people who are Black, Indigenous, or of color, and those living with disabilities. Feminist research has much in common with the broader paradigm of interpretivist research, but it pays greater attention to the importance of standpoints and of inequality and oppression in shaping the dynamics of research.

While the discussion of paradigms here is not exhaustive—there are many other approaches to research, many other epistemologies—it does provide an overview of some of the possible ways to think about research and data analysis. One important thing to remember is that while there *are* criteria for good research, criteria that will be further outlined in subsequent chapters of this text, there are no objective or empirical standards for which paradigm is "correct." In other words, individual researchers or research teams approach their research from the perspective or philosophy that makes sense to them, and while others may have reasons for disapproving, they cannot say that such a choice is right or wrong. Researchers must make these sorts of decisions for themselves.

Inductive and Deductive Approaches

Another question we might ask about the epistemology of research processes is whether our data *emerges from* our analysis or whether our data *generates* our analysis. If you argue that data emerges from analysis, you are suggesting that you begin the research process with a theory and then look to the data you have collected to see whether or not you can find support for your theory. This approach enables the testing of theories. It is typically understood as a **deductive** approach to research. In deductive approaches, researchers develop a theory, collect data, analyze the data, and use their analysis to test their theory. Positivist research is often deductive in its approach.

Instead, if you argue that data generates analysis, you are suggesting that you begin the research process by collecting data and you then look to see what you can find within it. This approach enables the building of theories. It is typically understood as an **inductive** approach. In inductive approaches, researchers begin by collecting data. Then they analyze

that data and use that analysis to build new understandings. Interpretivist and feminist research are often inductive in their approach.

While qualitative research can be conducted using both deductive and inductive approaches, it is a bit more common for qualitative researchers to use inductive approaches. Such approaches are far less possible in quantitative analysis because of the need for more precisely-designed data collection techniques. Thus, one advantage of qualitative research is that it permits for an inductive approach and is thus especially useful in contexts in which very little is already know or where new explanations need to be uncovered. It is also possible to conduct research using what some call **abduction**, or an interplay between deductive and inductive approaches (Pearce 2012). Such an approach may also be found in mixed-methods research. This text will focus primarily on inductive approaches to qualitative data analysis, given that they are far more common. But deductive approaches do exist. For example, consider a researcher who is interested in what sorts of circumstances give rise to nonprofit organization boards deciding to replace the organization's director. More typically, a qualitative researcher with this question would interview a wide variety of non-profit board members and, based on the responses, would build a theory—an inductive approach. In contrast, the researcher could choose to conduct her study deductively. Then, she would read the prior literature on management and organizational decision-making and develop one or more hypotheses about the circumstances that give rise to leadership changes. She would then interview board members looking specifically for the constellation of circumstances she hypothesized to test whether these circumstances were associated with the decision to replace the director.

Research Standards

As researchers design and carry out their data collection and data analysis strategies, there are a variety of issues they must consider in terms of ensuring their research meets appropriate disciplinary and professional standards for quality and rigor. These include considerations of **generalizability** or **representativeness**, **reliability** and **validity**, and **ethics** and **social responsibility**. It is also important to note that researchers must be attentive to ensuring that they are not overstating the degree to which their research can demonstrate evidence of **causation**. It is only possible for research to demonstrate causation if it meets three essential criteria:

- **Association**, which means that there must be clear empirical evidence of a relationship between the factor understood as the cause and the factor understood as the effect,

- **Temporal order**, which means that it must be known that the causal factor happened earlier in time than the effect, and
- **Elimination of alternatives**, which means that the research must have eliminated *all possible alternative explanations* for the effect.

It is not generally possible to eliminate **all** possible alternative explanations—even if a research project is able to eliminate all the ones the researcher thought of, there are still other possibilities. Thus, research can only make true causal claims if its finding come from a properly-controlled laboratory experiment in which the only element that could possibly have change the outcome was the one under examination. If research does not involve a properly-controlled laboratory experiment, researchers must be cautious about the way they describe their findings. They cannot say that their study has proven anything or that it shows that A causes B. Instead, they can say something like "these findings are consistent with the hypothesis that A causes B." Qualitative research cannot conclusively show causal relationships, even though it can be suggestive of them.

Generalizability refers to whether the research findings from a particular study can be assumed to hold true for the larger population. Research can only be generalized if it is the result of a properly-conducted **random sample**, also called a **probability sample**, and then only to the population that was sampled from. In other words, if I conduct a random sample of students at a particular college, I can only assume my findings will hold true for students at that college—I cannot assume they accurately reflect dynamics at other colleges or among people who are not college students. Furthermore, because probability sampling can involve what is called sampling error, even it can not guarantee generalizability to the population from which the sample has been drawn. It simply optimizes the chance that such generalizability exists.

And if my sample was not random, I cannot assume that my findings reflect the broader dynamics of that college. This is because the randomization that is part of developing a random sample is designed to eliminate the potential for sample bias that might shape the results. For example, if I conduct a non-random sample of college students by posting an ad on a social media site, then my participants will only be those who saw or heard about the ad, and they may be different in some way from students who did not see or hear about the ad.

While it is possible to conduct qualitative research using a random sample, a considerable portion of qualitative research projects do not use random sampling. This is because it is only possible to develop a random sample if you have a list of all possible people in the population (or can use sampling methods like cluster sampling that allow you to randomize without such a list). Clearly, if I want to study students at a particular college, I can get a list of all possible students at that college. But what if I wanted to study people who play the video game *Fortnite*? Or individuals who enjoy using contouring makeup? Or parents who

have a child with autism as well as a child who is neurotypical? There are no lists of people in these categories, and thus a random sample is not possible. In addition, it can be hard to use random sampling for studies in which the researcher will ask participants for more lengthy time commitments, such as in-depth interviewing and ethnographic observation.

Where generalizability is not possible, researchers can instead strive for *representativeness*. Having a representative sample means having a sample that includes a sufficient number of people from various subgroups within the population such that the research can understand whether the dynamics it uncovers are applicable broadly across groups or whether they only apply to specific subgroups. Which characteristics must be reflected to ensure representativeness will vary depending on the study in question. A study of students' participation in extracurricular activities probably should consider both residential students and those who commute to campus. A study of retail employees might need to include both full-time and part-time workers as well as those who do and do not hold managerial positions. Race, gender, and class, as well as other axes of inequality, are very common subgroups used to ensure representativeness. Note that it is entirely ok to exclude various subgroups from a study, *as long as the study makes clear who and what it is studying*. In other words, it would be reasonable to conduct a study of mothers of children with autism. It would not be acceptable to conduct a study of *parents* with autism but only include mothers in the sample.

Reliability refers to the extent to which repeated measures produce consistent results. Usually, discussions of reliability refer to the consistency of specific measures. For instance, if I ask you what you ate for breakfast on Wednesday in a conversation on Wednesday evening and then on Friday morning, will you give me the same answer? Or if I administer two different self-esteem scales, do you come out with similar results? Changes in the way questions are asked, the context in which they are asked, or who is doing the asking can have remarkable impacts on the responses, and these impacts mean reliability is reduced. Some concerns about reliability, such as that illustrated with the self-esteem scale, refer to consistency between different approaches for measuring the same underlying idea. Others have to do with **repeatability**, **replicability**, or **reproducibility** (Plesser 2017). An example of the issue of repeatability is the question about what you ate for breakfast—if the same researcher repeats the same measurement, do they get the same results? Replicability refers to situations in which a different researcher uses the same measurement approaches on the same type of population, though the research may take place in a different location. While a researcher can never ensure that their research will be replicable, researchers who strive to ensure replicability do endeavor to make their research process as clear as possible in any publications so that others will be able to take the same exact steps in trying to replicate it. However, this can be difficult in qualitative studies as the impact the researcher has on the context through phenomena such as interviewer effect may mean that a different researcher or research team cannot exactly replicate the original conditions of data

collection. Finally, reproducibility refers to whether a different research team can develop its own methodological approach to answering the research question but still find results consistent with those in the original study. It is always possible that a study fails to reproduce not because the findings are inherently irreproducible but rather because some variation in the population or setting is responsible for the different results. Another element of reliability is **inter-rater reliability**. To understand inter-rater reliability, consider a study in which a researcher is trying to determine whether the degree of sexism displayed in advertisements differs depending on the type of product being advertised. In order to collect this data, a team of research assistant has to examine each advertisement and rate, on a scale of 1 to 5, how sexist the advertisement is. It's not surprising that different research assistants might judge the same advertisement differently–and this can impact the results of the study. Measuring inter-rater reliability helps determine how different these multiple raters' ratings are from one another, and if the differences are large, the researcher can go back and retrain the research assistants so they can more consistently apply the intended rating scale.

Validity refers to the extent to which research measurements accurately reflect the underlying reality. Well-designed qualitative approaches, especially in-depth interviewing and participant-observation, tend to be high in validity. This is because such methods come the closest of all social science methods to reflecting real life in all of its complexity. Validity can be increased by careful attention to research design, the use of method **triangulation** (multiple research methods or approaches), and deep reflection on process and findings.

While a full treatment of research *ethics* is beyond the scope of this book, it is essential to remember that good research always attends to the highest ethical standards. People who talk about research ethics often focus their primary attention to the treatment of human subjects, or the people who actually participate in the research project. Ethical treatment of participants includes ensuring that any risks they face are limited, that they have given fully-informed consent to their participation in research, that their identity will be protected,[1] and that they do not experience coercion to participate. An interesting example of the kinds of issues that a commitment to research ethics raises has to do with the legal risks inherent in research. Shamus Khan, a researcher studying sexual assault, has written about an instance in which he became embroiled in the court process after his research materials were subpoenaed in a lawsuit. The subpoena would have entitled the litigants to materials that would have disclosed confidential personal information, information research participants were assured would remain confidential. Khan details the lengths that he had to go to in order to protect participants' information and the complex ethical questions

1. However, there are research participants who wish to disclose their real identity, and some qualitative researchers argue that truly ethical research gives participants the option to make informed decisions about such disclosure.

his case raises, ultimately concluding that a real commitment to research ethics requires some changes in how the institutions that sponsor research think about and manage their responsibilities (Khan 2019).

Many commentators who discuss research ethics suggest that researchers' ethical responsibility goes much further. For example, feminist researchers often suggest that research participants be given the opportunity to review interview transcripts for errors, omissions, or statements they would have preferred not to make and issue corrections, even if their words and experiences will be used anonymously. Attention should also be paid to ensuring that people and communities who participate in research are able to share in the benefits of that research. For example, if a program is developed through research on a particular community of homeless people, those people should be among the first to be able to access the new program. If researchers profit financially from the research they have done, they might consider sharing the profits with those they have studied.

While traditional treatments of research ethics consider only the researcher's responsibility to research participants, a broader treatment of ethics—in keeping with interpretivist and feminist paradigms—would also include social responsibility as an ethical touchstone. Researchers concerned with social responsibility might consider whether their approach to publication or the content of their publications might have harmful impacts on the populations they have studied, stigmatizing them or exposing them to disadvantageous policy consequences. For example, Robert Putnam, a political scientist, conducted a study that examined the impact of neighborhood diversity on social cohesion and trust. When he found that diversity can reduce trust, he worried that his findings would be used as a political weapon by those opposed to diversity, racial equity, and immigration. Thus, while he made some data available to other researchers, he withheld publication for several years while he developed policy proposals designed to mitigate the potential harm of his findings. Some commentators felt that withholding publication was itself unethical, while Putnam felt that publishing without due consideration of the impact of his findings was the unethical thing. A commitment to social responsibility might also include attention to ensuring equity in citation practices, an issue that has been brought to the fore by the social media campaign #CiteBlackWomen, which urges scholars and teachers to ensure that they read publications by Black women, acknowledge Black women's scholarship through citation as well as inclusion in course reading lists, and ensure that Black women are represented as speakers at conferences, among other things (Cite Black Women Collective n.d.).

As noted above, research paradigms influence the particular qualities that researchers value in their research. In addition, it is not always realistic or even possible to maximize all of these qualities in a given project. Thus, most research, including most excellent research, will emphasize some of these standards and not others. This does not mean the research is lacking in rigor. Good research, however, is always explicit about its own limitations. Thus, researchers should indicate whether or not their results can be generalized, and if so, to

whom. They should be clear on which subgroups they included in their efforts to ensure representativeness.

The Process of Qualitative Research

So, how does one go about conducting an inductive qualitative research project? Well, there are a series of steps researchers follow. However, it is important to note that qualitative research and data analysis involve a high degree of fluidity and are typically **iterative**, meaning that they involve repeatedly returning to prior steps in the process.

First, researchers design their data collection process, which includes developing any data collection instruments such as interview guides and locating participants. Then, they collect their data. To collect data researchers might conduct interviews, observations, or ethnography, or they might locate documents or other sources of textual or visual data. While deductive quantitative approaches require researchers collect all their data and only then analyze it, inductive qualitative approaches provide the opportunity for more of a cyclical process in which researchers collect data, begin to analyze it, and then use what they have found so far to reshape their further data collection.

Once data is collected, researchers need to ensure that their data is usable. This may require the transcription of audio or video recordings, the scanning or photocopying of documents, typing up handwritten fieldnotes, or other processes designed to move raw data into a more manipulable form.

Next, researchers engage in **data reduction**. Research projects typically entail the collection of really large quantities of data, more data that can possibly be managed or utilized in the context of one paper. This is especially likely in the case of qualitative research because of the richness and complexity of the data that is collected. Therefore, once data collection is completed, researchers use strategies and techniques to reduce the hundreds or thousands of pages of **fieldnotes** or interview transcripts or documents into a manageable form. Activities involved in data reduction, which will be taken up in a later chapter, include coding, summarization, the development of data displays, and categorization.

Once data reduction has made data more usable, researchers can develop conclusions based on their data. Remember, however, that this process is iterative, which means that it is a continuing cycle. So, when researchers make conclusions, they also go back to earlier stages to refine their approaches. In addition, the process of developing conclusions also requires careful consideration to limitations of the data and analytical approaches, such as those discussed earlier in this chapter.

Finally, researchers present their findings. During each project, researchers must determine how best to disseminate results. Factors influencing this determination include the

research topic, the audience, and the intended use of the results—for instance, are these the results of **basic research**, designed to increase knowledge about the phenomena under study, or are they the results of **applied research**, conducted for a specific audience to inform the administration of a policy or program? Findings might be disseminated in a graphical form like an infographic or a series of charts, a visual form like a video or animation, an oral form like a lecture, or a written form like a scholarly article or a report. Of course, many projects incorporate multiple forms of dissemination.

While this chapter is titled "The Qualitative Approach," it is actually inaccurate to suggest that there is just one overall approach to qualitative research. As this chapter has shown, there are some core characteristics that qualitative approaches to research have in common, such as data that relies on words or images rather than numbers and a richer, more contextual understanding of the phenomena under study. But there are also many ways in which qualitative approaches to research vary. They use different methods of data collection. They take place within different paradigms and epistemologies. They focus their attention on emphasizing different standards for research quality. And, as the following chapters will show, they utilize different methods for preparing and managing data, analyzing that data, and disseminating their findings.

Exercises

1. Find a few grocery store circulars from your area. The ones that get delivered with your mail are fine, or you can locate them online on the website of your local grocery stores. Spend some time examining the circulars. Look at the words and images, the types of items represented, the fonts and layouts, anything that catches your eye, and then answer two questions: first, what do the circulars tell you about the lives of people today who live in your area, and second, what did you do, cognitively, to figure that out?
2. Locate a recent scholarly journal article in your field of study and read it. Do you think this article used a more positivist or more interpretivist paradigm of knowledge? Explain how you know, drawing on the key elements of these paradigms.
3. What do you think it means to do good research? Which of the various standards for good research do you think are most important to the topics or issues you are interested in? And what are some of the strategies you might employ to be sure your research lives up to these standards?

11. Preparing and Managing Qualitative Data

MIKAILA MARIEL LEMONIK ARTHUR

When you have completed data collection for a qualitative research project, you will likely have voluminous quantities of data—thousands of pages of **fieldnotes**, hundreds of hours of interview recordings, many gigabytes of images or documents—and these quantities of data can seem overwhelming at first. Therefore, preparing and managing your data is an essential part of the qualitative research process. Researchers must find ways to organize the voluminous quantities of data into a form that is useful and workable. This chapter will explore data management and data preparation as steps in the research process, steps that help facilitate data analysis. It will also review methods for data reduction, a step designed to help researchers get a handle on the volumes of data they have collected and coalesce the data into a more manageable form. Finally, it will discuss the use of computer software in qualitative data analysis.

Data Management

Even before the first piece of data is collected, a data management system is a necessity for researchers. Data management helps to ensure that data remain safe, organized, and accessible throughout the research process and that data will be ready for analysis when that part of the project begins. Miles and Huberman (1994) outline a series of processes and procedures that are important parts of **data management**.

First, researchers must attend to the formatting and layout of their data. Developing a consistent template for storing fieldnotes, interview transcripts, documents, and other materials, and including consistent **metadata** (data about your data) such as time, date, pseudonym of interviewee, source of document, person who interacted with the data, and other details will be of much use later in the research process.

Similarly, it is essential to keep detailed records of the research process and all research decisions that are made. Storing these inside one's head is insufficient. Researchers should keep a digital file or a paper notebook in which all details and decisions are recorded. For instance, how was the sample conducted? Which potential respondents never ended up going through with the interview? What software decisions were made? When did the dig-

ital voice recorder fail, and for how long? What day did the researcher miss going into the field because they were ill? And, going forward, what decisions were made about each step in the analytical process?

As data begin to be collected, it is necessary to have appropriate, well-developed physical and/or digital filing systems to ensure that data are safely stored, well-organized, and easy to retrieve when needed. For paper storage, it is typical to use a set of file folders organized chronologically, by respondent, or by some other meaningful system. For digital storage, researchers might use a similar set of folders or might keep all data in a single folder but use careful file naming conventions (e.g. RespondentPseudonym_Date_Transcript) to make it easy to find each piece of data. Some researchers will keep duplicate copies of all data and use these copies to begin to sort, mark, and organize data in ways that enable the presence of relationships and themes to emerge. For instance, researchers might sort interview transcripts by the way respondents answered a particular key question. Or they might sort fieldnotes by the central activities that took place in the field that day. Activities such as these can be facilitated by the use of index cards, color-coding systems, sticky notes, marginal annotations, or even just piles. Cross-referencing systems may be useful to ensure that thematic files can be connected to respondent-based files or to other relevant thematic files. Finally, it is essential that researchers develop a system of backups to ensure that data is not lost in the event of a catastrophic hard drive failure, a house fire, lack of access to the office for an extended period, or some other type of disaster.

One more issue to attend to in data management is research ethics. It is essential to ensure that confidential data is protected from disclosure; that identifying information (including signed consent forms) are not kept with or linkable to data; and that all researchers, analysts, interns, and administrative personnel involved in a study sign statements of confidentiality to ensure they understand the importance of nondisclosure (Berg 2009). Note that such documents will not protect researchers and research personnel from subpoena by the courts—if research documents will contain information that could expose participants to criminal or legal liability, there are additional concerns to consider and researchers should do due diligence to protect themselves and their respondents (see, e.g., Khan 2019), though the methods and mechanisms for doing so are beyond the scope of this text. Researchers must attend to data security protocols, many of which were likely agreed to in the IRB submission process. For example, paper research records should be locked securely where they cannot be seen by visitors or by personnel or accessed by accident. Digital records should be securely stored in password protected files that meet current standards for strong passwords. Cloud storage or backups should have similar protections, and researchers should carefully review the terms of service to ensure that they continue to own their data and that the data are protected from disclosure.

Preparing Data

In most cases, data are not entirely ready for analysis at the moment at which they are collected. Additional steps must be taken to prepare data for analysis, and these steps are somewhat different depending on the form in which the data exists and the approach to data collection that was used: fieldnotes from observation or ethnography, interviews and other recorded data, or documentary data like texts and images.

Fieldnotes

When researchers conduct ethnographic or observational research, they typically do not have the ability to maintain verbatim recordings. Instead, they maintain fieldnotes. Maintaining fieldnotes is a tricky and time-consuming process! In most instances, researchers cannot take notes—at least not too many—while present in the research site without making themselves conspicuous. Therefore, they need to limit themselves to a couple of jotted words or sentences to help jog their memories later on, though the quantity of notes that can be taken in the field is higher these days because of the possibility of taking notes via smartphone, a notetaking process largely indistinguishable from the socially-ubiquitous practices of text messaging and social media posts. Immediately after leaving the site, researchers use the skeleton of notes they have taken to write up full notes recording everything that happened. And later, within a day or so, many researchers go back over the fieldnotes to edit and refine the fieldnotes into a useful document for later analysis. As this process suggests, analysis is already beginning even while the research is ongoing, as researchers make notes and annotations about theoretical ideas, connections to explore, potential answers to their research questions, and other things in the process of refining their fieldnotes.

When fleshing out fieldnotes, researchers should be attentive to the distinctions between recollections they believe are accurate, interpretations and reflections they have made, and analytical thoughts that develop later through the process of refining the fieldnotes. It is surprisingly easy for a slight mistake in recording, say, which people did what, or in what sequence a series of events occurred, to entirely change the interpretation of circumstances observed in the field. To demonstrate how such issues can arise, consider the following two hypothetical fieldnote excerpts:

Excerpt A	Excerpt B
Sarah walked into the living room and before she knew what happened, she found Marisol on the floor in tears, surrounded by broken bits of glass. “What did you do?” Sarah said, her voice thick with emotion. Marisol covered her face and cried louder.	Her voice thick with emotion, Sarah said, “What did you do?” Before she knew what happened, she found Marisol on the floor in tears, surrounded by bits of broken glass. Sarah walked into the living room. Marisol covered her face and cried louder.

In Excerpt A, the most reasonable interpretation of events is probably that Sarah walked into the room and found Marisol, the victim of an accident, and was concerned about her. In Excerpt B, in contrast, Sarah probably caused the accident herself. Yet the words are exactly the same in both excerpts—they have just been slightly rearranged. This example highlights how important careful attention to detail is in recording, refining, and analyzing fieldnotes (and other forms of qualitative data, for that matter).

Fieldnotes contain within them a vast array of different types of data: records of verbal interactions between people, observations about social practices and interactions, researchers’ inferences and interpretations of social meanings and understandings, and other thoughts (Berg 2009). Therefore, as researchers work to prepare their fieldnotes for analysis, they may need to work through them again to organize and categorize different types of notes for different uses during analysis. The data collected from ethnographic or observational research can also include documents, maps, images, and recordings, which then need to be prepared and managed alongside the fieldnotes.

Interviews & Other Recordings

First of all, interview researchers need to think carefully about the form in which they will obtain their data. While most researchers audio- or video-record their interviews, it is useful to keep additional information alongside the recordings. Typically, this might include a form for keeping track of themes and data from each interview, including details of the context in which the interview took place, such as the location and who was present; biographical information about the participant; notes about theoretical ideas, questions, or themes that occur to the researcher during the interview; and reminders of particularly notable or valuable points during the interview. These information sheets should also contain the same pseudonym or respondent number that is used during the interview recording, and thus can be helpful in matching biographical details to participant quotes at the time of ultimate writeup. Interviewers may also want to consider taking notes throughout the interview, as notes can highlight elements of body language, facial expression, or more subtle comments that might not be picked up on audio recordings. While video recordings

can pick up such details, they tend to make participants more self-conscious than do audio recordings.

Once the interview has concluded, recordings need to be transcribed. While automated transcription has improved in recent years, it still falls far short of what is needed to make an accurate transcript. Transcription quality is typically assessed using a metric called the Word Error Rate—basically, dividing the number of incorrect words by the number of words that should appear in the passage—there are other, more complex assessment metrics that take into consideration individual words' importance to meaning. As of 2020, automated transcription services still tended to have Word Error Rates of over 10%, which may be sufficient for general understanding (such as in the case of apps that convert voicemails to text) but which is definitely too high of an error rate for use in data analysis. And error rates increase when audio recordings contain background noise, accented speech, or the use of dialects other than Standard American English (SAE). There can also be ethical concerns about data privacy when automated services are used (Khamsi 2019). However, automated services can be cost-effective, with a typical cost of about 25 cents per minute of audio (Brewster 2020). For a typical study involving 40 interviews averaging 90 minutes each, this would come to a total cost of about $900, far less than the cost of human transcription, which averages about $1 per minute these days. Human transcription is far more accurate, with extremely low Word Error Rates, especially for words essential to meaning. But human transcribers also suffer from increased error when transcribing audio with noisy backgrounds, where multiple speakers may be interrupting one another (for instance in recordings of focus groups), or in cases where speakers have stronger accents or speak in dialects other than Standard American English. For example, a study examining court reporters—professional transcribers with special experience and training at transcribing speech in legal contexts—working in Philadelphia who were assigned to transcribe African American English had average Word Error Rates of above 15%, and these errors were significant enough to fundamentally alter meaning in over 30% of the speech segments they transcribed (Jones *et al*. 2019).

Researchers can, of course, transcribe their recordings themselves, an option that vastly reduces cost but adds an enormous amount of time to the data preparation process. The use of specialized software or devices like foot-pedal controlled playback can facilitate the ease of transcription, but it can easily take up to four hours to complete the transcription of one hour of recordings. This is because people speak far faster than they type—a typical person speaks at a rate of about 150 words per minute and types at a rate more like 30-60 words per minute. Another possibility is to use a kind of hybrid approach in which the researcher uses automated transcription or voice recognition to get a basic—if error-laden—transcript and then corrects it by hand. Given the time that will be invested in correcting the transcript by listening to the recording while reviewing the transcript, even lower-quality transcription services may be acceptable, such as the automated captioning

video services like YouTube offer, though of course these services also present data privacy concerns. Alternatively, researchers might use voice-recognition software. The accuracy of such software can typically be improved by training it on the user's voice. This approach can be especially helpful when interview respondents speak with accents, as the researcher can re-record the interview in their own voice and feed it into software that is already trained to understand the researcher's voice.

Table 1 below compares different approaches to transcription in terms of financial cost, time, error rate, and ethical concerns. Costs for transcription by the researcher and hybrid approaches are typically limited to the acquisition of software and hardware to aid the transcription process. For a new researcher, this might entail several hundred dollars of cost for a foot pedal, a good headset with microphone, and software, though these costs are often one-time costs not repeated with each project. In contrast, even automated transcription can cost nearly a thousand dollars per project, with costs far higher for the hired human transcriptionsts who have much better accuracy. In terms of time, though, automated and hired services require far less of the researchers' time. Hired services will require some time for turnaround, more if the volume of data is high, but the researcher can work on other things during that time. For self and hybrid transcription approaches, researchers can expect to put in much more time on transcription than they did conducting interviews. For a typical project involving 40 interviews averaging 90 minutes each, the time required to conduct the interviews and transcribe them—not including time spent preparing for interviews, recruiting participants, traveling, analyzing data, or any other task—can easily exceed 300 hours. If you assume a researcher has 10 hours per week to devote to their project, that would mean it would take over 30 weeks just to collect and transcribe the data before analysis could begin. And after transcription is complete, most researchers find it useful to listen to the recordings again, transcript in hand, to correct any lingering errors and make notes about avenues for exploration during data analysis.

Table 1. Comparing Transcription Approaches for a Typical Interview-Based Research Project

Automated	$900	A few hours turnaround	High	High
Hired	$3,600	At least several days turnaround	Low for SAE	Probably Low
Self	Minimal	About 240 hours active	Varies	Low
Hybrid	Minimal	Varies, likely at least 120 hours active	Low for SAE	Varies
Note: this table assumes a project involving 40 interviews, all conducted by the main researcher, averaging 90 minutes in length. Time costs do not include interviewing itself, which would add an additional 60 hours to the time required to complete the project.				

Documents and Images

Data preparation is far different when data consists of documents and images, as these already exist in textual form. Here, concerns are more likely to revolve around storage, filing, and organization, which will be discussed later in this chapter. However, it can be important to conduct a preliminary review of the data to better understand what is there. And for visual data, it may be especially useful to take notes on the content in and the researcher's impressions of each visual as a starting point to thinking about how to further work with the materials (Saldaña 2016).

There are special concerns about research involving documents and images that are worth noting here. First all, it is important to remember the importance of sampling issues in relation to the use of documents. Sampling is not always a concern—for instance, research involving newspaper articles may involve a well-conducted random sample, or photographs may have been taken by the researcher themselves according to a clear purposive sampling process—but many projects involving textual data have used sampling procedures where it remains unclear how representative the sample is of the universe of data. Researchers must keep careful notes on where the documents and images included in their data came from and what sorts of limitations may exist in the data and include a discussion of these issues in any reporting on their research.

When writing about interview data, it is typical to include excerpts from the interview transcripts. Similarly, when using documents or visual materials, it is preferable to include some of the original data. However, this can be more complex due to copyright concerns. When using published works, there are real legal limits on the quantity of text that you can include without getting permission from the copyright owner, who may make you pay

for the privilege. This is not an issue for works that were created or published more than 95 years ago, as their copyrights have expired. For works more recent than that, the use of more than a small portion of the work typically violates copyright, and the use of an image is almost never permitted unless it has been specifically released from copyright (or created by the researcher themselves). Archival data may be subject to specific usage restrictions imposed by the archive or donor. Copyright can make the goal of providing the data in a form useful to the reader very difficult, so you might need to get the copyright clearance or find other creative ways of providing the data.

Data Reduction

In qualitative data analysis, data collection and data analysis are often not two distinct research phases. Rather, as researchers collect data, they begin to develop themes, ask analytical questions, write theoretical memos, and otherwise begin the work of analysis. And when researchers are analyzing data, they may find they need to go back and collect more to flesh out certain areas that need further elaboration (Taylor, Bogdan, and DeVault 2016). But as researchers move further towards analysis, one of the first steps is reading through all of the data they have collected. Many qualitative researchers recommend taking notes on the data and/or annotating it with simple notations like circles or highlighting to focus your attention on those passages that seem especially fruitful for later focus (Saldaña 2016). This is often called "pre-coding." Other approaches to pre-coding include noting hypotheses about what might emerge elsewhere in the data, summarizing the main ideas of each piece of data and annotating it with details about the respondent or circumstances of its creation, and taking preliminary notes about concepts or ideas that emerge.

This sort of work is often called "preliminary analysis," as it enables researchers to start making connections and working with themes and theoretical ideas, but before you get to the point of making actual conclusions. It is also a form of **data reduction**. In qualitative analysis, the volume of data collected in any given research project is often enormous, far more than can be productively dealt with in any particular project or publication. Thus, data reduction refers to the process of reducing large volumes of data such that the more meaningful or important parts are accessible. As sociologist Kristen Luker points out in her text *Salsa Dancing into the Social Sciences* (2008), what we are really trying to do is recognize patterns, and data reduction is a process of sifting through, digesting, and thinking about our data until we can see the patterns we might not have seen before. Luker argues that one important way to help ourselves see patterns is to talk about our data with others—lots of others, and not just other social scientists—until what we are explaining starts to make sense.

There are a variety of approaches to data reduction. Which of these are useful for a particular project depends on the type and form of data, the priorities of the researcher, and the goals of the research project, and so each researcher must decide for themselves how to proceed. One approach is **summarization**. Here, researchers write short summaries of the data—summaries of individual interview transcripts, of particular days or weeks of fieldnotes, or of documents. Then, these summaries can be used for preliminary analysis rather than requiring full engagement with the larger body of data. Another approach involves writing memos about the data in which connections, patterns, or theoretical ideas can be laid out with reference to particular segments of the data. A third approach is annotation, in which marginal notes are used to highlight or draw attention to particularly important or noteworthy segments of the data. And Luker's suggestion of conversations about our data with others can be understood as a form of data reduction, especially if we record notes about our conversations.

One of the approaches to data reduction which many analysts find most useful is the creation of **typologies**, or systems by which objects, events, people, or ideas can be classified into categories. In constructing typologies, researchers develop a set of mutually-exclusive categories—no one can be placed into more than one category of the typology (Berg 2009)—that are, ideally, also exhaustive, so that no one is left out of the set of categories (an "other" category can always be used for those hard to classify). They then go through all their pieces of data or data elements, be they interview participants, events recorded in fieldnotes, photographs, tweets, or something else, and place each one into a category. Then, they examine the contents of each category to see what common elements and analytical ideas emerge and write notes about these elements and ideas.

One approach to data reduction which qualitative researchers often fall back on but which they should be extremely careful with is quantification. **Quantification** involves the transformation of non-numerical data into numerical data. For example, if a researcher counts the number of interview respondents who talk about a particular issue, that is a form of quantification. Some limited quantification is common in qualitative analysis, though its use should be particularly rare in ethnographic research given the fact that ethnographic research typically relies on one or a very small number of cases. However, the use of quantification should be constrained to those circumstances where it provides particularly useful or illuminating descriptive information about the data, and not as a core analytical tool. In addition, given that it is exceptionally uncommon for qualitative research projects to produce generalizable findings, any discussion of quantified data should focus on numbers rather than percents. Numbers are descriptive—"35 out of 40 interview respondents said they had argued with housemates over chores in the past week"—while percents suggest broader and more generalizable claims ("87.5% of respondents said they had argued with housemates over chores in the past week").

Qualitative Data Analysis Software

As part of the process of preparing data for analysis and planning an analysis strategy, many—though not all—qualitative researchers today use software applications to facilitate their work. The use of such technologies has had a profound impact on the way research is carried out, as have many technological changes over history. Take a much older example: the development of technology permitting for the audio recording of interviews. This technology made it possible to develop verbatim transcripts, whereas prior interview-based research had to rely on handwritten notes conveying the interview content—or, if the interviewer had significant financial resources, perhaps a stenographer. Recordings and verbatim transcripts also made it possible for researchers to minutely analyze speech patterns, specific word choices, tones of voice, and other elements that would not previously have been able to be preserved.

Today's technologies make it easier to store and retrieve data, make it faster to process and analyze data, and provide access to new analytical possibilities. On a basic level, software can allow for more sophisticated possibilities for linking data to memos and other documents. And there are a variety of other benefits (Adler and Clark 2008) to the use of software-aided analysis (often referred to as **CAQDAS**, or computer-aided qualitative data analysis software). It can allow for more attention to detail, more systematic analysis, and the use of more cases, especially when dealing with large data sets or in circumstances where some quantification is desirable. The use of CAQDAS can enhance the perception of rigor, which can be useful when bringing qualitative data to bear in settings where those using data are more used to quantitative analysis. When coding (to be discussed further in the chapter on qualitative coding), software enhances flexibility and complexity, and may enliven the coding process. And software can provide complex relational analysis tools that go well beyond what would be possible by hand.

However, there are limitations to the use of CAQDAS as well (Adler and Clark 2008). Software can promote ways of thinking about data that are disconnected from qualitative ideals, whether through reductions in the connection between data and context or the increased pressure to quantify. Each individual software application creates a specific model of the architecture of data and knowledge, and analysis may become shaped or constrained by this architecture. Coding schemes, taxonomies, and strategies may reflect the capacities available in and the structures prioritized by the software rather than reflecting what is actually happening in the data itself, and this can further homogenize research, as researchers draw from a few common software applications rather than from a wide variety of personal approaches to analysis. Software can also increase the psychic distance between the researcher or analyst and their data and reduce the likelihood of researchers understanding the limitations of their data. The tools available in CAQDAS applications

tend to emphasize typical data rather than unusual data, and so outliers or negative cases may be missed. Finally, CAQDAS does not always reduce the amount of time that a research project takes, especially for newer users and in cases with smaller sets of data. This is because there can be very steep learning curves and prolonged set-up procedures.

The fact that this list of limitations is somewhat longer than the list of positives should not be understood as suggesting that researchers avoid CAQDAS-based approaches. Software truly does make forms of research possible that would not have been without it, speeds data processing tasks, and makes a variety of analytical tasks much easier to do, especially when they require attention to detail. And digital technologies, including both software applications and hardware devices, facilitate so much about how qualitative researchers work today. There are a wide variety of types of technological aids to the qualitative research purpose, each with different functions.

First of all, digital technologies can be used for capturing qualitative data. This may seem obvious, but as the example of audio recording above suggests, the development of technologies like audio and film recording, especially via cellphone or other small personal devices, led to profound changes in the way qualitative research is carried out as well as an expansion in the types of research that are possible. Other technologies that have had similar impacts include the photocopier and scanner, and more recently the possibility to use a cell phone to capture photographs of documents in archives (without the flash on to avoid damaging delicate items). Finally, videoconferencing software makes it possible to interview people who are halfway around the world, and most videoconferencing platforms have a built-in option to save a video record of the conversation, and potentially autocaption it. It's also worth noting that digital technologies provide access to sources of data that simply did not exist in the past, whether interviewing via videoconferencing, content analysis of social media, or ethnography of massively-multiplayer online games or worlds.

Software applications are very useful for data management tasks. The ability to store, file, and search electronic documents makes the management of huge quantities of data much more feasible. Storing metadata with files can help enormously with the management of visual data and other files. Word processing programs are also relevant here. They help us produce and revise text and reports, compile and edit our fieldnotes and transcriptions, write memos, make tables, count words, and search for and count specific words and phrases. Graphics programs can also facilitate the creation of graphs, charts, infographics, and other data displays. Finally, speech recognition programs aid our transcription process and, for some of us, our writing process.

Coding programs fall somewhere between data reduction and data analysis in their functions. Such software applications typically provide researchers with the ability to apply one or more codes to specific segments of text, search for and retrieve all segments that have had particular codes applied to them, and look at relationships between different codes. Some also provide data management features, allowing researchers to store memos, doc-

uments, and other materials alongside the coded text, and allow for **interrater reliability** testing (to be discussed in another chapter). Finally, there are a variety of data analysis tools. These tools allow researchers to carry out functions like organizing coded data into maps or diagrams, testing hypotheses, merging work carried out by different researchers, building theory, utilizing formal comparative methods, creating diagrams of networks, and others. Many of these features will be discussed in subsequent chapters.

Choosing the Right Software

There are so many programs out there that carry out each of the functions discussed above, with new ones appearing constantly. Because the state of technology changes all the time, it is outside the scope of this chapter to detail specific options for software applications, though online resources can be helpful in this regard (see, e.g., University of Surrey n.d.). But researchers still need to make decisions about which software to use. So, how do researchers choose the right qualitative software application or applications for their projects? There are four primary sets of questions researchers should ask themselves to help with this decision.

First, what functions does this research need and does this project require? As discussed above, programs have very different functions. In many cases, researchers may need to combine multiple programs to get access to all the functions they need. In other cases, researchers may need only a simple software application already available on their computers.

Second, researchers should consider how they use technology. There are a variety of questions that are relevant here. For example, what kind of device will be used, a desktop computer, laptop, tablet, or phone? What operating system, Windows, Mac/iOS, Chrome, or Android? How much experience and skill do researchers have with computers—do they need software applications that are very easy to use, or can they handle command-line interfaces that require some programming skills? Do they prefer software that is installed on their devices or a cloud-based approach? And will the researcher be working alone or as part of a team where multiple people need to contribute and share access to the same materials?

What type of data will be used? Will it be textual, visual, audio, or video? Will data come from multiple sources and styles or will it all be consistent? Is the data organized or free-form? What is the magnitude of the data that will be analyzed?

Finally, what resources does the researcher already have available? What software can they access, whether already available on their personal computing devices or via licenses provided by their employer or college/university? What degree of technical support can they access, and are technical support personnel familiar with CAQDAS? And how much

money do they have available to pay for software on a one-time or ongoing basis? Note that some software can be purchased, while other software is provided as a service with a monthly subscription fee. And even when software is purchased, licenses may only provide access for a limited time period such as a year. Thus, both short-term and long-term financial costs and resource availability should be assessed prior to committing to a software package.

Exercises

1. Transcribe about 10 minutes of an audio interview—one good source might be your local NPR station's website. Be sure that your transcription is an *exact record* of what was said, including any pauses, laughter, vulgarities, or other kinds of things you might not typically write in an academic context, and that you transcribe both questions and responses. What was it like to complete this exercise?
2. Use the course listings at your college or university as a set of data. Develop a typology of different types of courses—*not* based on the department or school offering them or the course number alone—and classify courses within this typology. What does this exercise tell you about the curriculum at your college or university?
3. Review the notes, documents, and other materials you have already collected from this course and develop a new system of file management for them, with digital or physical folders, subfolders, and labels or file names that make items easy to locate.

12. Qualitative Coding

MIKAILA MARIEL LEMONIK ARTHUR

Codes are words or phrases that capture a central or notable attribute of a particular segment of text or visual data (Saldaña 2016). **Coding**, then, is the process of applying codes to texts or visuals. It is one of the most common strategies for data reduction and analysis of qualitative data, though many qualitative projects do not require or use coding. This chapter will provide an overview of approaches based in coding, including how to develop codes and how to go through the coding process.

In order to understand coding, it is essential to think about what it means for something to be a code. To analogize to social media, codes might function a bit like tags or hashtags. They are words or phrases that convey content, ideas, perspectives, or other key elements of segments of text. Codes are not the same as themes. **Themes** are broader than codes—they are concepts or topics around which a discussion, analysis, or text focuses. Themes are more general and more explanatory—often, once we code, we find themes emerge as ideas to explore in our further analysis (Saldaña 2016). Codes are also different from descriptors. **Descriptors** are words or phrases that describe characteristics of the entire text and/or the person who created it. For example, if we note the profession of an interview respondent, whether an article is news or opinion, or the type of camera used to take a photograph, those would be descriptors. Saldaña (2016) instead calls these **attributes**. The term attributes more typically refers to the possible answer choices or options for a variable, so it is possible to think about descriptors as variables (or perhaps their attributes) as well.

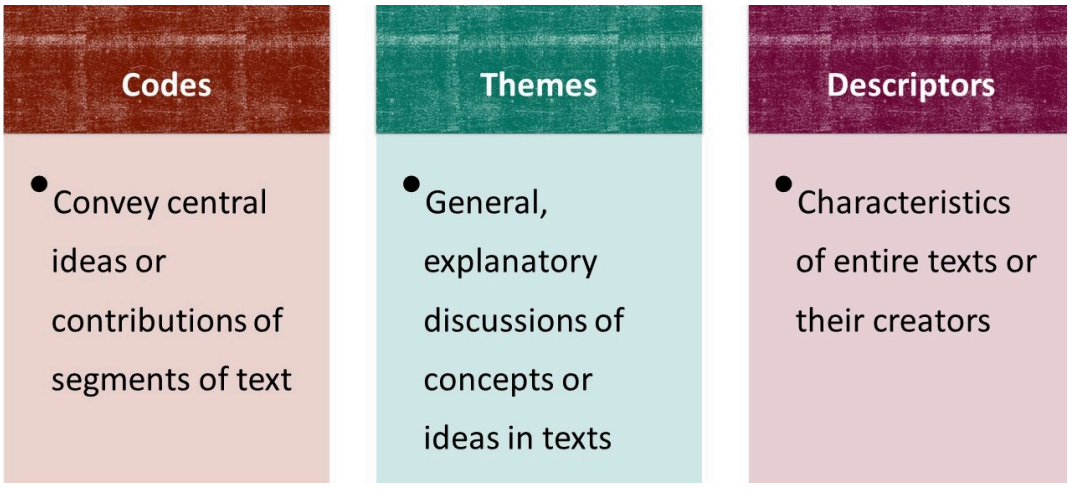


Figure 1. Codes vs. Themes v.s. Descriptors

Let's consider an example. Imagine that you were conducting an interview-based study looking at minor-league athletes' workplace experiences and later-life career plans. In this study, themes might be broad ideas like "aspirations" or "work experiences." There would be a vast array of codes, but they might include things like "short-term goals," "educational plans," "pay," "team bonding," "travel," "treatment by managers," "family demands," and many more. Descriptors might include the athlete's gender and what sport they play.

Developing a Coding System

While all approaches to coding have in common the idea that codes are applied to segments of text or visuals, there are many different ways to go about coding. These approaches differ in terms of when they occur during the research process and how codes are developed. First of all, there is a distinction between first- and second-cycle coding approaches (Saldaña 2016). **First-cycle coding** happens early in the research process and is really a bridge from data reduction to data analysis, while **second-cycle coding** occurs later in the research process and is more analytical in nature. Another version of this distinction is the comparison between rough, analytic, and focused coding. **Rough coding** is really part of the process of data reduction. It often involves little more than putting a few words near each segment of text to make clear what is important in that segment, with the approach being further refined as coding continues. In contrast, **analytic coding** involves more detailed techniques designed to move towards the development of themes and findings. Finally, **focused coding** involves selecting ideas of interest and going back and re-coding your texts to orient your approach more specifically around these ideas (Bergin 2018).

A second set of distinctions concerns whether the data drives the development of codes or whether codes are instead developed in advance. If codes are determined in advance, or predetermined, researchers develop a set of codes based on their theory, hypothesis, or research question. This sort of coding is typically called **deductive coding** or **closed coding**. In contrast, **open coding** or **inductive coding** refers to a process in which researchers develop codes based on what they observe in their data, grounding their codes in the texts. This second approach is more common, though by no means universal, in qualitative data analysis. In both types of coding, however, researcher may rely upon ideas generated by writing theoretical memos as they work through the connections between concepts, theory, and data (Saldaña 2016).

Finally, a third set of distinctions focuses on *what* is coded. **Manifest coding** refers to the coding of surface-level and easily observable elements of texts (Berg 2009). In contrast, **latent coding** is a more interpretive approach based on looking deeply into texts for the meanings that are encoded within or symbolized by them (Berg 2009). For example, con-

sider a research project focused on gender in car advertisements. A manifest approach might count the number of men versus women who appear in the ads. A latent approach would instead focus on the use of gendered language and the extent to which men and women are depicted in gender-stereotyped ways.

Researchers need to answer two more questions as they develop their coding systems. First, *what* to code, and second, *how many* codes. When thinking about what to code, researchers can look at the level of individual words, characters or actors in the text, paragraphs, entire textual items (like complete books or articles), or really any unit of text (Berg 2009), but the most useful procedure is to look for chunks of words that together express a thought or idea, here referred to as "segments of text" or "textual segments," and then code to represent the ideas, concepts, emotions, or other relevant thoughts expressed in those chunks.

How many codes should a particular coding system have? There is no simple answer to this question. Some researchers develop complex coding systems with many codes and may have over a hundred different codes. Others may use no more than 25, perhaps fewer, even for the same size project (Saldaña 2016). Some researchers nest codes into code trees, with several related "child" codes (or subcodes) under a single "parent" code. For example, a code "negative emotions" could be the parent code for a series of codes like "anger," "frustration," "sadness," and "fear." This approach enables researcher to use a smaller or larger number of codes in their analysis as seems fit after coding is complete. While there is no formula for determining the right number of codes for a particular project, researchers should be attentive to overgrowth in the number of codes. Codes have limited analytical value if they are used only once or twice—if a coding system includes many codes that are applied only a small number of times, consider whether there are larger categories of codes that might be more useful. Occasionally, there are codes worth keeping but applying rarely, for example when there is a rare but important phenomenon that arises in the data. But for the most part, codes should be used with some degree of frequency in order for them to be useful for uncovering themes and patterns.

Types of Codes

A wide variety of different types of codes can be used in coding systems. The discussion below, which draws heavily on the work of Saldaña (2016), details a variety of different approaches to coding and code development. Researchers do not need to choose just one of these approaches—most researchers combine multiple coding approaches to create an overall system that is right for the texts they are coding and the project they are conducting. The approaches detailed here are presented roughly in order of the degree of complexity they represent.

At the most basic level is **descriptive coding**. Descriptive codes are nouns or phrases describing the content covered in a segment of text or the topic the segment of text focuses on. All studies can use descriptive coding, but it often is less productive of rich data for analysis than other approaches might be. Descriptive coding is often used as part of rough coding and data reduction to prepare for later iterations of coding that delve more deeply into the texts. So, for instance, that study of sexism in advertisements might involve some rough coding in which the researcher notes what type of product or service is being advertised in each advertisement.

Structural coding, in contrast, attends more closely to the research question rather than to the ideas in the text. In structural coding, codes indicate which specific research question, part of a research question, or hypothesis is being addressed by a particular segment of text. This may be most useful as part of rough coding to help researchers ensure that their data addresses the questions and foci central to their project.

In vivo coding captures short phrases derived from participants' own language, typically action-oriented. This is particularly important when researchers are studying subcultural groups that use language in different ways than researchers are accustomed to and where this language is important for subsequent analysis (Manning 2017). In this approach, researchers choose actual portions of respondents' words and use those as codes. In vivo coding can be used as part of both rough and analytical coding processes.

A related approach is **process coding**, which involves "the use of gerunds to label actual or conceptual actions relayed by participants" (Saldaña 2016:77). (**Gerunds** are verb forms that end in -ing and can function grammatically as if they are nouns when used in sentences). Process coding draws researchers' attention to actions, but in contrast to in vivo coding it uses the researcher's vocabulary to build the coding system. So, for instance, in the study of minor league athletes discussed earlier in the chapter, process codes might include "traveling," "planning," "exercising," "competing," and "socializing."

Concept coding involves codes consisting of words or short phrases that represent broader concepts or ideas rather than tangible objects or actions. Sticking with the minor league athletes example, concept codes might include "for the love of the game," "youth," and "exploitation." A combination of concept, process, and descriptive coding may be useful if researchers want their coding system to result in an inventory of the ideas, objects, and actions discussed in the texts.

Figure 2. A Selection of "Emoticodes"

Emotion codes are codes indicating the emotions participants discuss in or that are evoked by a segment of text. A more contemporary version of emotion codes relies on "emoticodes" or the emoji that express specific kinds of emotions, as shown in Figure 2.

Values coding involves the use of codes designed to represent the "perspectives or worldview" of a respondent by conveying participants' "values, attitudes, and beliefs" (Saldaña 2016:131). For example, a project on elementary school teachers' workplace satisfaction might include values codes like "equity," "learning," "commitment," and "the pursuit of excellence." Do note that choices made in values coding are, even more so than in other forms of coding, likely to reflect the values and worldviews of the coder. Thus, it can be essential to use a team of multiple coders with different backgrounds and perspectives in order to ensure a values coding approach that reflects the contents of the texts rather than the ideas of the coders.

Versus coding requires the construction of a series of binary oppositions and then the application of one or the other of the items in the binary as a code to each relevant segment of text. This may be a particularly useful approach for deductive coding, as the researcher can set out a series of hypothesized binaries to use as the basis for coding. For example, the project on elementary school teachers' workplace satisfaction might use binaries like feeling supported vs. feeling unsupported, energized vs. tired, unfulfilled needs vs. fulfilled needs, kids ready to learn vs. kids needing services, academic vs non-academic concerns, and so on.

Evaluation coding is used to signify what is and is not working in the policy, program, or endeavor that respondents are discussing or that the research focuses on. This approach is obviously especially useful in evaluation research designed to assess the merit or functioning of particular policies or programs. For example, if the project about elementary school teachers was part of a mentoring program designed to keep new teachers in the education profession, codes might include "future orientation" to flag portions of the text in which teachers discuss their longer-term plans and "mentor/mentee match" to flag portions in which they explore how they feel about their mentors, both key elements of the program and its goals.

There are a variety of other approaches more common outside of sociology, such as **dramaturgical coding**, which is a coding approach that treats interview transcripts or fieldnotes as if they are scripts for a play, coding such things as actors, attitudes, conflicts, and

subtexts; coding approaches relying on terms and ideas from literary analysis; and those drawn from communications studies, which focus on facets of verbal exchange. Finally, some researchers have outlined very specific coding strategies and procedures such that someone else could pick up their methods and apply them exactly. This sort of approach is typically deductive, as it requires the advance specification of the decisions that will be made about coding.

Some coding strategies incorporate measures of weight or intensity, and this can be combined with many of the approaches detailed above. For example, consider a project collecting narratives of people's experiences with losing their jobs. Respondents might include a variety of emotional content in their narratives, whether sadness, fear, stress, relief, or something else. But the emotions they discuss will vary not only in type, they will also vary in extent. A worker who is fired from a job they liked well enough but who knows they will be able to find another job soon may express sadness while a worker whose company closed after she worked there for 20 years and who has few other equivalent employment opportunities in the region may express devastation. **Code weights** help account for these differences.

A final question researchers must consider is whether they will apply only one code per segment of text or will permit overlapping codes. Overlapping codes make data analysis more complex but can facilitate the process of looking for relationships between different concepts or ideas in the data.

Codebooks

As a coding system is developed and certainly upon its completion, researchers create documents known as **codebooks**. As is the case with survey research, codebooks lay out the details of how the measurement instrument works to capture data and measure it. For surveys, a codebook tells researchers how to transform the multiple-choice and short-answer responses to survey questions into the numerical data used for quantitative analysis. For qualitative coding, codebooks instead explain when and how to use each of the codes included in the project. Codebooks are an important part of the coding process because they remind the researcher, and any other coders working on the project, what each code means, what types of data it is meant to apply to, and when it should and should not be used (Luker 2008). Even if a researcher is coding without others, it is easy to lose sight of what you were thinking when you initially developed your coding system, and so the codebook serves as an important reminder.

For each code, the codebook should state the name of the code, include a couple of sentences describing the code and what it should be used for, any information about when the code should *not* be used, examples of both typical and atypical conditions under which the

code would be used, and a discussion of the role the code plays in analysis (Saldaña 2016). Codebooks thus serve as instruction manuals for when and how to apply codes. They can also help researchers think about taxonomies of codes as they organize the code book, with higher-level ideas serving as categories for groups of child, or more precise, codes.

The Process of Coding

So, what does the process of coding look like? While qualitative research can and does involve deductive approaches, the process that will be detailed here is an inductive approach, as this is more common in qualitative research. This discussion will lay out a series of steps in the coding process as well as some additional questions researchers and analysts must consider as they develop and carry out their coding.

The first step in inductive coding is to completely and thoroughly read through the data several times while taking detailed notes. To Saldaña (2016), the most important question to ask during this initial read is what is especially interesting or surprising or otherwise stands out. In addition, researchers might contemplate the actions people take, how people go about accomplishing things, how people use language or understand the world, and what people seem to be thinking. The notes should include anything and everything—objects, people, emotions, actions, theoretical ideas, questions—really anything, whether it comes up again and again in the data or only once, though it is useful to flag or highlight those concepts that seem to recur frequently in the data.

Next, researchers need to organize these notes into a coding system. This involves deciding which coding approach(es) to incorporate, whether or not to use parent and child codes, and what sort of vocabulary to use for codes. Remember that readers will not see the coding system except insofar as the researcher chooses to convey it, so vocabulary and terms should be chosen based on the extent to which they make sense to the research team. Once a coding system has been developed, the researcher must create a codebook. If paper coding will be used, a paper codebook should be created. If researchers will be using CAQDAS, or computer-aided qualitative data analysis software, to do their coding, it is often the case that the codebook can be built into the software itself.

Next, the researcher or research team should rough code, applying codes to the text while taking notes to reflect upon missing pieces in the coding system, ways to reorganize the codes or combine them to enhance meaning, and relevant theoretical ideas and insights. Upon completing the rough coding process, researchers should revise the coding system and codebook to fully reflect the data and the project's needs.

At this point, researchers are ready to engage in coding using the revised codebook. They should always have someone else code a portion of the texts—usually a minimum of

10%—for interrater reliability checks, and if a larger research team is used, 10% of the texts should be coded in common by all coders who are part of the research team. Even in cases where researchers are working alone, it truly strengthens data analysis to be able to check for interrater reliability, so most analysts suggest having a portion of the data coded by another coder, using the codebook. If at all possible, additional coding staff should not be told what the hypothesis or research question is, as one of the strengths of this approach is that additional coding staff will be less likely to be influenced by preexisting ideas about what the data should show (Luker 2008). There are various quantitative measures, such as Chronbach's alpha and **Kappa**, that researchers use to calculate interrater reliability, the measure of how closely the ratings of multiple coders correspond. All coders should keep detailed notes about their coding process and any obstacles or difficulties they encounter.

How do researchers know they are done coding? Not just because they have gone through each text once or twice! Researchers may need to continue repeating this process of revision and re-coding until additional coding does not reveal anything more. This repetition is an essential part of coding, as coding always requires refinement and rethinking (Saldaña 2016). In Berg's (2009:354-55) words, it is essential to "code minutely," beginning with a rough view of the entire text and then refining as you go until you are examining each detail of a text. Then, researchers think about why and how they developed their codes and what jumps out at them as important from the research as they delve into findings, making sure that nothing has been left out of the coding process before they move towards data analysis.

One interesting question is whether the identities and standpoints (as discussed in the chapter "The Qualitative Approach") of coders matter to the coding process. Eduardo Bonila-Silva (Zuberi and Bonilla-Silva 2008:17) has described how, after a presentation discussing his research on racism, a colleague asked whether the coders were White or Black—and he responded by asking the colleague "if he asked such questions across the board or only to researchers saying race matters." As Bonilla-Silva's question suggests, race (like other aspects of identity and experience, such as gender, immigration status, disability status, age, and social class, just to name a few) very well might shape the way coders see and understand data, functioning as part of a particular coding filter (Saldaña 2016). But that shaping extends broadly across all issues, not just those we might assume are particularly salient in relationship to identities. Thus, it is best for research teams to be diverse so as to ensure that a variety of perspectives are brought to bear on the data and that the findings reflect more than just a narrow set of ideas about how the world works.

Coding and What Comes After

If researchers will code by hand, they will need multiple copies of their data, one for reference and one for writing on (Luker 2008). On the copy that will be written on, researchers use a note-taking system that makes sense to them—whether different-colored markers, Roman numerals in the margins, a complex series of sticky notes, or whatever—to mark the application of various codes to sections of your data. You can see an example of what hand coding might look like in Figure 3 below, which is taken from a study of the comments faculty members make on student writing. Segments of text are highlighted in different colors, with codes noted in the margins next to the text. You can see how codes are repeated but in different combinations. Once the initial coding process is complete, researchers often cut apart the pieces of paper to make chunks of text with individual codes and sort the pieces of paper by code (if multiple codes appear in individual chunks of text, additional copies might be needed). Then, each pile is organized and used as the basis for writing theoretical memos. Another option for coding by hand is to use an index sheet (Berg 2009). This approach entails developing a set of codes and categories, arranging them on paper, and entering transcript, page, and paragraph information to identify where relevant quotes can be found.

For more complex analytical processes, researchers will likely want to use software, though there are limitations to software. Luker (2008), for instance, argues that when coding manually, she tends to start with big themes and only breaks them into their constituent parts later, while coding using software leads her to start with the smallest possible codes. (One solution to this, offered by some software packages, is upcoding, where a so-called "parent" code is simultaneously applied to all of the "child" codes under it. For instance, you might have a parent code of "activism" and then child codes that you apply to different kinds of activism, whether protest, legislative advocacy, community organizing, or whatever.)

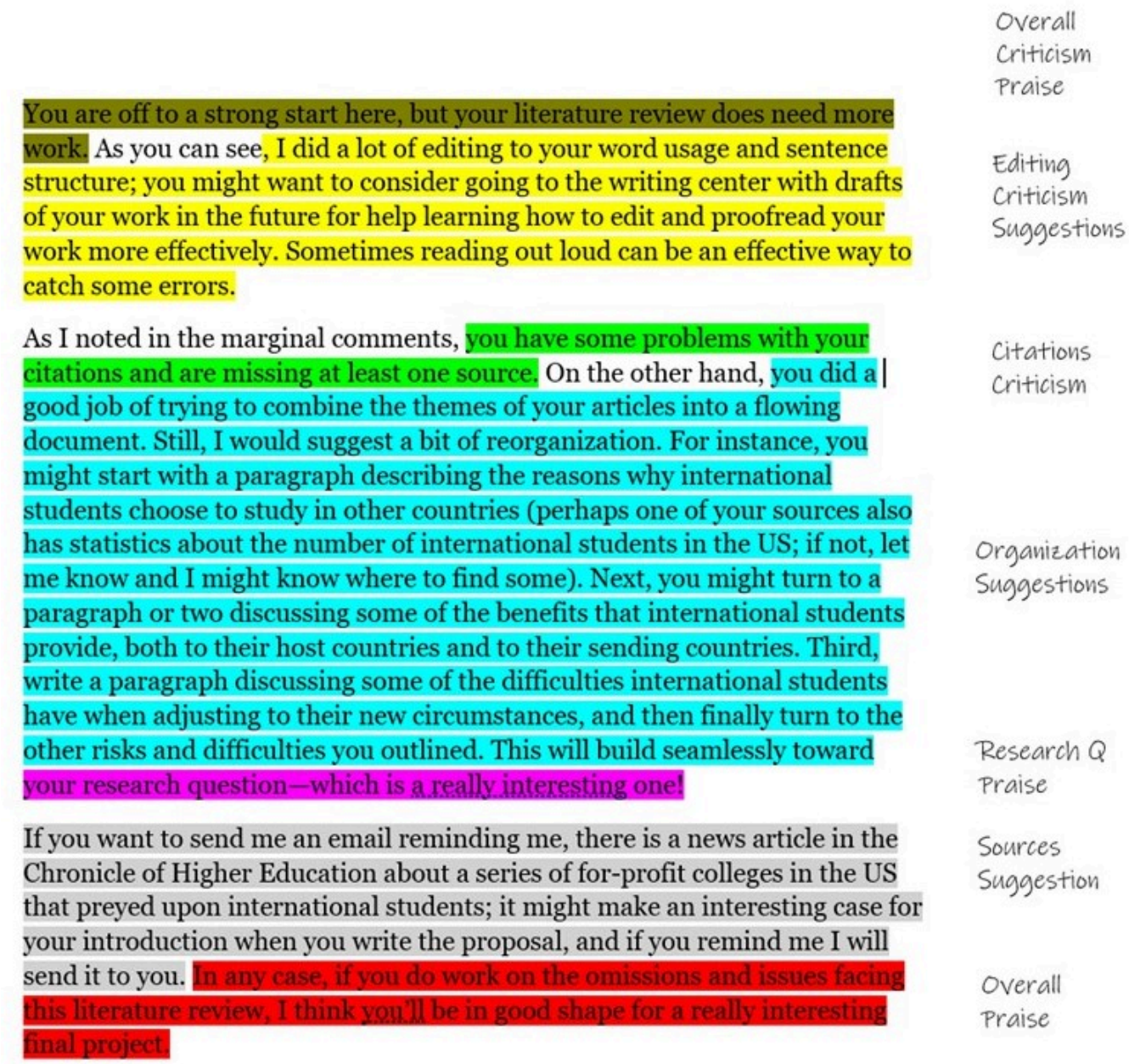


Figure 3. An Example of Hand Coding

Coding does not stand on its own, and thus simply completing the coding process does not move a research project from data to analysis. While the analysis process will be discussed in more detail in a subsequent chapter, there are several steps researchers take alongside coding or immediately after completing coding that facilitate analysis and are thus useful to discuss in the context of coding. Many of these are best understood as part of the process of data reduction. One of the most important of these is categorizing codes into larger groupings, a step that helps to enable the development of themes. These larger

groupings, sometimes called "parent" codes, can collapse related but not identical ideas. This is always useful, but it is especially useful in cases where researchers have used a large number of codes and each one is applied only a few times. Once parent codes have been created, researchers then go back and ensure that the appropriate parent code is assigned to all segments of text that were initially coded with the relevant "child" codes (a step that can be automated in CAQDAS). If appropriate, researchers may repeat this process to see if parent codes can be further grouped. An alternative approach to this grouping process is to wait until coding is complete, and then create more analytical categories that make sense as thematic groupings for the codes that have been utilized in the project so far (Saldaña 2016).

There are a variety of other approaches researchers may take as part of data reduction or preliminary analysis after completing coding. They may outline the codes that have occurred most frequently for specific participants or texts, or for the entire body of data, or the codes that are most likely to co-occur in the same segment of text or in the same document. They may print out or photocopy documents or segments of text and rearrange them on a surface until the arrangement is analytically meaningful. They may develop diagrams or models of the relationships between codes. In doing this, it is especially helpful to focus on the use of verbs or other action words to specify the nature of these relationships—not just stating that relationships exist, but exploring what the relationships do and how they work.

In inductive coding especially, it is often useful to write theoretical and analytical memos while coding occurs, and after coding is completed it is a good time to go back and review and refine these memos. Here, researchers both clearly articulate to themselves how the coding process occurred and what methodological choices they made as well as what preliminary ideas they have about analysis and potential findings. It can be very useful to summarize one's thinking and any patterns that might have been observed so far as a step in moving towards analysis. However, it is extremely important to remember the data and not just the codes. Qualitative researchers always go back to the actual text and not just the summaries or categories. So a final step in the process of moving toward analysis might be to flag quotes or data excerpts that seem particularly noteworthy, meaningful, or analytically useful, as researchers need these examples to make their data come alive during analysis and when they ultimately present their results.

Becoming a Coder

This chapter has provided an overview of how to develop a coding system and apply that system to the task of conducting qualitative coding as part of a research project. Many new

researchers find it easy—if sometimes time-consuming and not always fascinating—to get engaged with the coding process. But what does it take to become an effective coder? Saldaña (2016) emphasizes personality attributes and skills that can help. Some of these are attributes and skills that are important for anyone who is involved in any aspect of research and data analysis: organization, to keep track of data, ideas, and procedures; perseverance, to ensure that one keeps going even when the going is tough, as is often the case in research; and ethics, to ensure proper treatment of research participants, appropriate data security behaviors, and integrity in the use of sources. In most aspects of data analysis, creativity is also important, though there are some roles in quantitative data analysis that require more in the way of technical skills and ability to follow directions. In qualitative data analysis, creativity remains important because of the need to think deeply and differently about the data as analysis continues. Flexibility and the ability to deal with ambiguity are much more important in qualitative research, as the data itself is more variable and less concrete; quantitative research tends to place more emphasis on rules and procedures. A final strength that is particularly important for those working in qualitative coding is having a strong vocabulary, as vocabulary both helps researchers understand the data and enhances their ability to create effective and useful coding systems. The best way to develop a stronger vocabulary is to read more, especially within your discipline or field but broadly as well, so researchers should be sure to stay engaged with reading, learning, and growing.

Reading, learning, and growing, along with a lot of practice, is of course how researchers enhance their data collection, coding, and data analysis skills, so keep working at it. Qualitative research can indeed be easy to get started with, but it takes time to become an expert. Put in the time, and you, too, can become a skilled qualitative data analyst.

Exercises

1. For each of the following words or phrases, consider whether it is most likely to represent a code, a theme, or a descriptor. Explain your response.
    - Female respondent
    - Energized
    - The relationship between poverty and social control
    - Creative
    - A teacher
    - The process of divorce
    - Social hierarchies
    - Grief

2. Pick a research topic you find interesting and determine which of the approaches to coding detailed in this chapter might be most appropriate for your topic, then write a paragraph about why this approach is the best.
3. Sticking with the same topic you used to respond to Exercise 2, brainstorm some codes that might be useful for coding texts related to this topic. Then, write appropriate text for a codebook for each of those codes.
4. Select a hashtag of interest on a particular social media site and randomly sample every other post using that hashtag until you have selected 15 tweets. Then inductively code those posts and engage in summarization or classification to determine what the most important themes they express might be.
5. Create a codebook based on what you did in Exercise 4. Exchange codebooks and tweets with a classmate and code each other's tweets according to the instructions in the codebook. Compare your results—how often did your coding decisions agree and how often did they disagree? What does this tell you about interrater reliability, codebook construction, and coder training?

Media Attributions

- codes themes descriptors © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC (Attribution NonCommercial) license
- Emoticodes © AnnaliseArt is licensed under a CC BY (Attribution) license
- Hand Coding Example © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-ND (Attribution NonCommercial NoDerivatives) license

13. From Qualitative Data to Findings

MIKAILA MARIEL LEMONIK ARTHUR

So far in this text, you have learned about various approaches to managing, preparing, reducing, and otherwise interacting with qualitative data. Because of the iterative and cyclical nature of the qualitative research process, it is not accurate to say that these steps come *before* analysis. Rather, they are an integral *part of* analysis. Yet there are procedures and methods for moving from the data to findings that are essential to completing a qualitative data analysis project. This chapter will outline three basic strategies for analysis of qualitative data: theoretical memos, data displays, and narratives; discuss how to move towards conclusions; and suggest approaches for testing these conclusions to ensure that they hold up to scrutiny.

But before the final stages of analysis occur, researchers do need to take a step back and ensure that data collection really is finished—or at least finished enough for the particular phase of analysis and publication the researcher is working on, in the case of a very long-term project. How do researchers know their data collection has run its course? Well, in some cases they know because they have exhausted their sample. If a project was designed to include interviews of forty respondents or the collection of 500 social media posts, then it is complete when those interviews have been conducted or those social media posts have been saved. In other cases, researchers know that data collection is complete when they reach **saturation**, or the point in the research process where continuing to engage in data collection no longer yields any new insights. This way of concluding data collection is more common in ethnographic work or work with archival documents.

In addition, since qualitative research often results in a truly enormous amount of data, one of the key tasks of analysis is finding ways to select the most central or important ideas for a given project. Keep in mind that doing so does not mean dismissing other ideas as unimportant. Rather, these other ideas may become the basis for another analysis drawing on the same data in the future. But one project, even an entire book, often cannot contend with the full body of data that a researcher or research team has collected. That is why it is important to engage in data reduction before or alongside the analysis process.

As researchers move from data towards findings, it is essential that they remember that, unlike much quantitative research, most qualitative research draws on small or otherwise unrepresentative samples, the findings also cannot be generalized. Thus, while the findings of qualitative research may be suggestive of general patterns, they must be regarded as just that: only suggestive.

Similarly, qualitative research cannot demonstrate **causation**. Demonstrating causation requires three elements:

- **Association**, or the ability to show a clear relationship between the phenomena, concepts, or processes in question,
- **Temporal order**, or the ability to show that the supposed cause came earlier in time than the supposed effect, and
- **Elimination of alternatives**, or the ability to show that there is *no possible alternative explanation* that could account for the phenomena in question.

While qualitative research *can* demonstrate association and temporal order, it cannot eliminate all alternative explanations—only well-designed and properly-controlled laboratory experiments can do so. Therefore, qualitative researchers (along with any quantitative researchers who are not relying on data from controlled laboratory experiments) need to take care to stay away from arguments suggesting that their analysis has proven anything or shown a causal relationship. However, qualitative researchers can locate evidence that *supports* the argument that a relationship is causal, leading to sentences like "This provides evidence suggestive of a causal relationship between X and Y."

Theoretical Memos

Memo-writing has been discussed in prior chapters as a strategy for data reduction, but memos (or theoretical notes) can be a key element in the process of moving from data to findings. Memos and theoretical notes are texts the researcher writes for themselves in which they work through ideas from the data, connect examples and excerpts to themes and theories, pose new questions, and illuminate potential findings. Memos can serve as a way to move from the rich, contextual detail of qualitative data—detail that can sometimes be overwhelming—towards the broader issues and questions that motivate a study. Initial memos are often drafted while data collection is still going on. For instance, a researcher might write reflective memos that integrate preliminary thoughts and ideas about data, help clarify concepts central to the research project, or pull together disparate chunks of data. But additional, and more specifically analytical, memos come later in the process.

These memos may focus on a variety of topics and ideas, including reflecting on the researcher's role and thought processes; contemplating the research question, potential answers, and shifts in the research focus; noting choices about coding strategies, the coding process, and choices made during coding; clarifying ethical or methodological issues that have arisen in the course of the research; considering what further research may need

to be done in the future; and trying out ideas that may become part of the final analysis or write-up (Saldaña 2016). What is integral to the memo-writing approach is the recognition that writing is a key part of thinking (Roberts 1993). Thus, it is in drafting memos that researchers can come to better understand the ideas shaping their study and what their data is saying in response to their research question.

Saldaña (2016:274-76) suggests a variety of techniques for analytical processes that use theoretical memos as thinking tools:

- Select the ten most interesting passages—a paragraph to half a page in length—from your data. Arrange, rearrange, and comment on these passages.
- Choose the three "codes, categories, themes, and/or concepts" (Saldaña 2016:275) that most stand out to you. Write about how they relate to one another.
- Write a narrative in as few sentences as possible that incorporates all of the most important key words, codes, or themes from your study.
- Make sure your analysis is based in concepts instead of nouns. Saldaña advises the use of what he calls the "touch test"—if you can touch it, it is a noun rather than a concept. For instance, you can touch a formerly incarcerated person, but you cannot touch their reentry strategy for life after prison. Thus, if the data still seems to be focused on or organized around nouns, write about the concepts that these nouns might exemplify or illuminate.

Alternatively, you can deconstruct your research question into its central parts and write memos laying out what the data seems to be telling you about the answer(s) to each one of these parts. Then, you can go back to the data to see if your impressions are supported or if you are missing other key elements. But note that theoretical memos cannot be simply lists of anecdotes or ways to confirm pre-existing ideas. They *must* be explicitly and thoroughly grounded in the data.

Another option, related to the crafting of theoretical memos but distinct in its approach, is to create what Berg (2009) calls "short-answer sheets." These function as a kind of open-ended survey, except instead of administering the survey directly to a respondent, researchers use an interview transcript or a text as a data source to complete the survey themselves. The short-answer sheet might summarize respondents' answers to key portions of the research question or detail how texts address particular issues of concern to the project, for instance. Such an approach can help categorize different themes or types of responses by making it easier to see commonalities and patterns.

Data Displays

Another tool for moving from data to findings is called a **data display**. Data displays are diagrams, tables, and other items that enable researchers to visualize and organize data so that it is possible to clearly see the patterns, comparisons, processes, and themes that emerge. These patterns, comparisons, processes, and themes then enable the researcher to articulate conclusions and findings. Data displays can be used as analytical tools, as part of the presentation process (to be discussed elsewhere), or to serve both purposes simultaneously. While the discussion of data displays in this chapter will of necessity be introductory, researchers interested in learning more about the development and use of data displays in qualitative research can consult Miles and Huberman's (1994) thorough and comprehensive sourcebook.

Researchers who use data displays need to remember that even as the displays enable the drawing of conclusions through looking for patterns and themes, this is not sufficient to support analysis and writeup. The display is simply a tool for analysis and understanding and cannot fully encapsulate the findings or the richness of the data. Thus, the analytical process always needs to return to the data itself, and whatever researchers write up or present needs to include stories, quotes, or excerpts of data to bring the concepts and ideas that are part of the study alive. The display is just that and does not itself contain or encapsulate the conclusions/analysis.

There are a variety of types of data displays, and this chapter will only present two common types: first, process and network diagrams, and second, matrixes, including tables, chronologies or timelines, and related formal methods. This summary should provide researchers with a sense of the possibilities. Of course, researchers can craft new approaches that make sense to them and for their specific projects—they should not feel bound to using only those types of data displays they have previously encountered. It is also worth noting here than many CAQDAS software packages integrate their own proprietary approaches to data display, and these can be very helpful for researchers using such software.

Process or Network Diagrams

Drawing various kinds of diagrams can help make sense of the ideas and connections in the data (Taylor, Bogdan, and DeVault 2016). Process and network diagrams are both key ways of visualizing relationships, thoughthe type of relationships they visualize differ. **Process diagrams** visualize relationships between steps in a process or procedure, while **network diagrams** visualize relationships between people or organizations. There is also

a specialized kind of network diagram called a **cognitive map** that shows how ideas are related to one another (Miles and Huberman 1994). Process diagrams are particularly useful for policy-related applied research, as they can help researchers understand where an intervention would be helpful or where a current policy or program is falling short of its goals, as well as whether the number or complexity of steps in a process may be getting in the way of optimal outcomes.

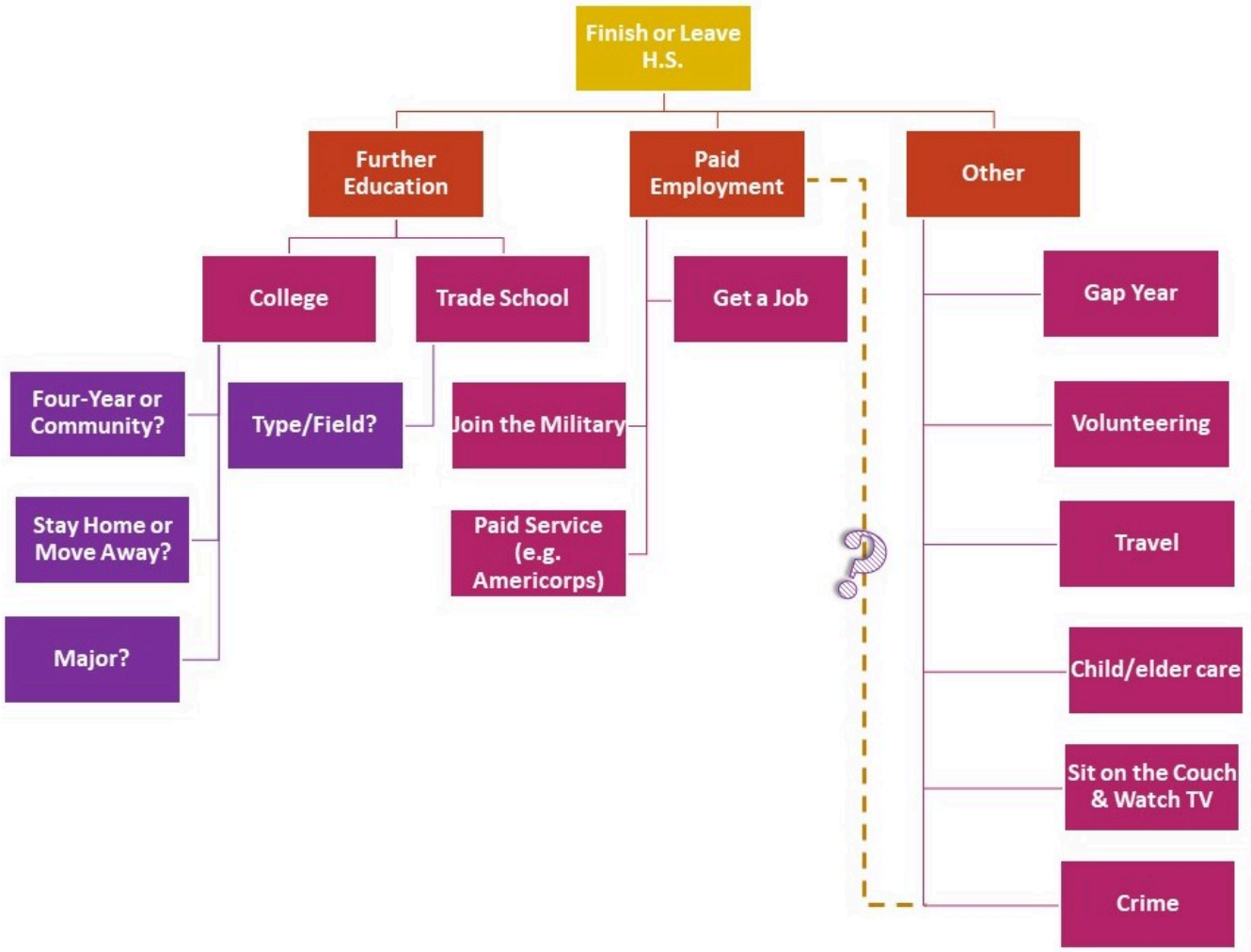


Figure 1. A Decision Tree or Process Diagram of Options for After High School

The category of visualizations that we call process diagrams also includes decision trees and flow charts. **Decision trees** are diagrams that specifically lay out the steps that people or organizations take to make decisions. For example, consider Figure 1, which is a hypothetical decision tree that could have emerged from a study of high school students considering their post-high school plans. Such a diagram can allow researchers to uncover unexpected sticking points or organize an analytical narrative around stages in the deci-

sion-making process. For example, maybe students are not aware of the option to participate in paid service programs or consider a gap year, options depicted in Figure 1. And, as the question mark suggests, those entering into criminal activity, such as drug sales, may see what they do as a kind of under-the-table employment—or as a rejection of employment. Similarly, **flow charts** can be used to diagram stages in a process and can depict complex, multidimensional relationships between these stages (flow charts can also be used to diagram personal relationships—one common use of flow charts is to diagram organizational relationships within corporations or other organizations, but more on this later).

To develop any type of process diagram, researchers should review their interview transcripts, fieldnotes, documents, and other data for instances were decisions are made (or avoided), events occurred that could have had a different outcome, or steps are taken or not taken that lead or could have led in a particular direction. All of these occurrences should then be listed and categorized, and researchers should work out which decisions or steps seem to depend on earlier decisions or steps, to be associated with other co-occurring decisions or steps, or to shape later decisions or steps. It is essential to ensure that no decisions or steps are missed when the diagram is created. For example, if one were to create a diagram of the process of making a peanut butter and jelly sandwich, it would not be sufficient to say "put peanut butter and jelly on bread." Rather, the diagram would have to account for choosing and obtaining the peanut butter, jelly, and bread; the utensils needed spread the peanut butter and jelly; how to go about spreading it; and the fact that the sandwich needs to be closed up after the spreading is complete.

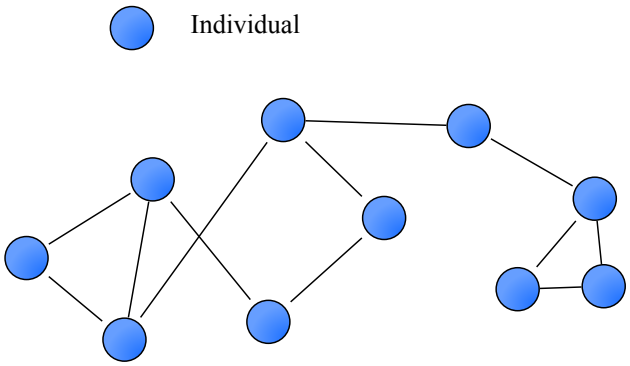


Figure 2. A Hypothetical Network Diagram

Network diagrams, as noted above, visualize the connections between people, organizations, or ideas. There is an entire subfield of sociology concerned with network analysis, and it has developed a specialized language for talking about these diagrams. In this specialized language, the individual people, organizations, or ideas included in the diagram are called **nodes** (represented by the blue circles in Figure 2), while the lines connecting them are called **edges** (represented by the black lines in Figure 2). Analysts can choose to use double-headed arrows, which indicate reciprocal relationships, or single-headed arrows, which indicate one-way relationships—or lines without arrows, if the direction of relationships is unclear, as in Figure 2. While the field of network analysis typically relies on quantitative and computational methods to draw conclusions from these diagrams, which can be large and complex enough that generating them requires specialized software, network diagrams can also be useful

tools for smaller-scale qualitative research because of the way they enable visualization of complicated webs of relationships.

To develop a network diagram, researchers must begin by combing through their data to find all of the individual nodes (people, organizations, ideas, or other entities) that will be included in the diagram. Then, they must find the relationships between the nodes, which can require a lot of time spent reviewing codes and categories produced as part of data reduction as well as many memoing tasks designed to help clarify these relationships and understand where each node fits in the broader scheme of things. Analysts can then draw diagrams by hand, by using computer graphics software, or by using specialized network analysis software. Alternatively, researchers may wish to create a flow chart. As noted above, flow charts can be extremely useful for diagramming relationships—the difference is that in a flow chart the relationships tend to be more formalized, like supervisory relationships in a corporation or volunteer organization (this is often referred to as an **organizational chart**) or kinship ties in a family tree. Once researchers create an initial draft of their network diagram, they need to look back at the evidence and data they have collected, especially searching for disconfirming bits of data that might suggest alternative network arrangements, in order to be sure they are on the right track.

Once network diagrams are created, they can be used in a variety of ways to help researchers build towards conclusions and findings. As noted above, they can be especially useful for understanding processes or policies, as they enable analysts and viewers to see whether something is working, where the social breakdowns might be, or where a procedure might benefit from changes. Of course, network diagrams also help make relationships and patterns of interaction clear, and by following the path laid out in the diagrams, it is possible to uncover insight into what the extended network surrounding an individual might look like—a network the individual themselves may not even be fully conscious of. Examining the shape and structure of the network can also answer a variety of questions. For example: are lots of entities clustered closely together in one area? Are there parts of the diagram where it seems like there are very few connections, or where it takes a lot of steps to get from one person to another? Are there **cliques**, or portions of the network where everyone is close with everyone else in their subgroup, or are ties spread more evenly? Are there ties that seem like they should exist but do not? And how does the diagram map on to social processes in the real world, such as the spread of ideas, behaviors, or disease? The possibilities go on. The point, though, is that these approaches allow researchers to see the way relationships interact and intersect as part of building their analysis.

Matrices and Tables

Another type of data display arranges data in a grid. We call these displays **matrices** (the

plural of matrix) or **tables**. Such displays let researchers more clearly see patterns of similarities and differences and observe comparisons across cases or types of cases. In developing a matrix-based data display, researchers have a variety of choices to make. The most important concern is that the matrix be usable for analysis and display purposes. Thus, researchers should strive to create matrixes that they can see all at once—those fitting on a single sheet of paper or a single computer screen. As a simple heuristic or rule designed to help guide matrix development, try to ensure that a matrix has no more than 10-12 rows or columns, and a number closer to six may be preferable. If a matrix cannot be adjusted to usably fit into one view, researchers should consider whether it is possible to break the matrix into component parts, each of which fits onto a screen. For example, a study of racism in the residential real estate industry might, rather than including one overall matrix, include separate matrixes detailing the experiences of real estate agents, home buyers, and home sellers, or alternatively separate matrixes for rental markets and sales markets. While many researchers draw matrixes by hand, word processing programs have excellent table construction features that enable the development of clear, organized, and well-formatted matrixes.

In order to populate the matrix, it is necessary for researchers to already have completed some level of data reduction and preliminary analysis, such as coding and/or the development of typologies or categories. These steps will then form the basis of matrix development. However, there are many different ways to organize the data in a matrix, and thus there are a variety of questions researchers must think through as they develop this type of data display.

First, will the matrix be used to explore one case at a time, with the grid being used to illuminate elements of the case and with a new copy of the matrix being completed for each case covered in the data? Or alternatively, will it be used to make cross-case comparisons or look for patterns?

Second, will this matrix be descriptive or explanatory? A descriptive matrix is especially helpful in circumstances where researchers are trying to categorize or classify data or develop a typology with key characteristics for each type or category. In contrast, a matrix designed to facilitate explanation needs to go well beyond the detailed characteristics of each case to instead think about comparisons, processes, potentially-causal mechanisms, and other dynamics, enabling researchers to see relationships among the data.

Third, which axes of variation—variables, if you will—will be incorporated into the table? Remember, a typical matrix has two dimensions, the columns and the rows, though modern computer applications can create matrixes that are multidimensional and thus include one or more additional axes. These axes might include individual respondents, types or categories of people, settings or contexts, actions people take, events that have occurred, distinct perspectives, organizations, processes and procedures, or portions of the research question, just to name a few.

Fourth, does the order of the data in the table show anything important about the data? In many cases it does not, and the order of table columns or rows is, for all intents and purposes, random. But this need not be the case—order can be a valuable part of both the development of the data display and the analytical power it presents. Order can be used to convey information about participants' social roles, timing and process, or the magnitude or strength of ideas and emotions, just to give a few examples.

Fifth, what information will be placed within each cell of the matrix or table? Will the table cells include summaries of data, events, or stories? Vignettes developed from multiple accounts? Cross-references to specific interview transcripts or documents? Key quotes? Explanations of what the data shows about the intersection of the ideas noted in the columns and rows? Images? Or perhaps there is another idea that is a better fit for a particular research project.

Once the design and structure of the matrix have been developed, researchers then return to their data to complete the matrix, filling in headers for the rows and columns and then entering information of the specified kind into each table cell. To do so properly, researchers need to be sure they are returning to the raw data frequently to check that details are correct and that nothing has been inappropriately excluded.

After the matrix has been fully completed, it can be used to facilitate analysis. The design of a matrix makes it particularly useful for observing patterns or making comparisons, but matrixes can also be helpful in making themes or categories clear or looking for clusters of related ideas or cases. For example, consider the matrix shown in Table 1. Note that this matrix was constructed based on an already-published article (Edin 2000), but it reflects the kind of approach a researcher might take as part of the creation of data displays for analysis. The study on which Table 1 was based was an inductive in-depth interview study of low-income single mothers who identified as Black or White in three United State cities, looking at how low-income women think about marriage. While the author, Kathryn Edin, does not detail her analysis strategy in the article, she likely used a coding-based approach to uncover five key factors shaping women's thoughts about marriage: affordability, respectability, control, trust, and domestic violence. By selecting a series of quotes that illustrate each theme and categorizing them by the race of the respondent, it is possible to see whether there are important racial differences in any of the themes. In fact—as may be clear from Table 1 itself—there are not major racial differences in the way Black and White respondents discussed each theme, except that White respondents were more likely to think "marrying up" was a realistic possibility; there were, however, differences in the *frequency* with which respondents discussed each theme, which Edin discusses in the text of the article. As this might suggest, some researchers, including Edin, use counting as part of the matrix process, whether to count rows and column with specific characteristics or to insert numbers into table cells to quantify the frequency of particular types of responses,

but care should be taken with numbers to avoid risks related to over-quantification (as discussed elsewhere in this text).

Table 1. Themes and Quotes from Edin 2000

	Black	White
Affordability	"Men simply don't earn enough to support a family. This leads to couples breaking up." "You can't get married and go on living with your mother. That's just like playing house." "We were [thinking about marriage] for a while, but he was real irresponsible. I didn't want to be mean or anything, [but when he didn't work] I didn't let him eat my food."	"I couldn't get him to stay working....It's hard to love somebody if you lose respect...."
Respectability	"I am not going to get married and pay rent to someone else. When we save up enough money to [buy] an acre of land and [can finance] a trailer, then we'll marry." I want to marry "up or not at all."	I plan to "marry out of poverty" and become a housewife.
Control	"[I won't marry because] the men take over the money. I'm too afraid to lose control of my money again. "I'm the head of the household right now, and I make the [financial] decisions. I [don't want to give that up]." "One thing my mom did teach me is that you must work some and bring some money into the household so you can have a say in what happens. If you completely live off a man, you are help-less. That is why I don't want to get married until I get my own [career] and get off of welfare."	"If we were to marry, I don't think it would be so ideal. [Husbands] want to be in charge, and I can't deal with that." "[Marriage isn't an option] right now. I don't want any man thinking that he has any claim on my kids or on how I raise them." "I don't want to depend on nobody. It's too scary." "You know, I feel better [being alone] because I am the provider, I'm getting the things that I want and I'm getting them for myself, little by little."
Trust	"All those reliable guys, they are gone, they are gone. They're either thinking about one of three things: another woman, another man, or dope...." "I would like to find a nice man to marry, but I know that men cannot be trusted." "I'm frustrated with men, period. They bring drugs and guns into the house, you take care of their kids, feed them, and then they steal your rent money out of your purse."	"I was married for three years before I threw him out after discovering that he had another woman. I loved my husband, but I don't [want another one]. This is a wicked world we are living in." "I want to meet a man who will love me and my son and want us to grow together. I just don't know if he exists." "Love is blind. You fall in love with the wrong one sometimes. It's easy to do."

Domestic Violence	"My daughter's father, we used to fight. I got to where nobody be punching on me because love is not that serious. And I figure somebody is beating on you and the only thing they love is watching you go the emergency room. That's what they love."	"...after being abused, physically abused, by him the whole time we were married, I was ready to [kill him]. He put me in the hospital three times. I was carrying our child four and a half months, he beat me and I miscarried.
	"So [we got in the car] and we started arguing about why he had to hang around people like that [who do] drugs and all that sort of stuff. One thing led to another and he kind of tossed me right out of the car."	"I was terrified to leave because I knew it would mean going on welfare.... But that is okay. I can handle that. The thing I couldn't deal with is being beat up."

A specialized type of matrix that works somewhat differently is the **timeline** or **chronology**. In this type of matrix, one axis is time (whether times of day, days of the week, dates, years, ages, or some other time scale), and the researcher fills in details about events, interactions, or other phenomena that occurred or that could be expected to occur at each given time. Researchers can create separate timelines for individual respondents or organizations, or can use common timelines to group and summarize data.

Matrixes also form the basis of a set of more complex analytical methods, including truth tables and related formal methods, such as Charles Ragin's Qualitative-Comparative Analysis (QCA). While a full exploration of such approaches is beyond the scope of this text, new data analysts should be aware of the existence of these techniques. They are designed to formalize qualitative analysis (Luker 2008) using a rigorous procedure focusing on the existence of necessary and sufficient factors. Ragin (2000; 2008) uses Boolean algebra and the logic of sets to determine which factors or groups of factors are necessary and which factors or groups of factors are sufficient for each important outcome. Boolean algebra is the mathematical expression of logical relationships based in the principles of syllogism—deductive reasoning based on a series of premises (Boole 1848). Sets simply represent groups of similar objects that are classified together. Thus, in QCA and other techniques based on this approach, factors and combinations of factors are evaluated to see which are necessary—without them, the outcome cannot happen—and which are sufficient—with them, the outcome must always happen.

There are a variety of other types of formal methods as well. For example, in case study research, researchers might use quasi-experimental approaches to compare cases across key axes of variation (Gerring 2007). Another approach is called process tracing, in which data, "with the aid of a visual diagram or formal model" is used to "verify each stage of this model" (Gerring 2007:184), and there are many others not discussed here. Researchers who seek formal approaches for developing within-case or cross-case comparisons can turn to the robust literature on case study research as they enhance their methodological toolkits.

Narrative Approaches

In many cases, researchers rely on the crafting of narratives as a key part of the analytical process. Such narratives may be used only for analysis or they may be integrated into the ultimate write-up of the project. There are a variety of different narrative approaches (Grbich 2007). One approach is the case study, in which the researcher tells the story of a specific individual or organization as part of an extended narrative discussion or a short summary. A given project may include multiple case studies. Case studies can be used to holistically highlight the dynamics and contexts under exploration in a project, seeking to describe and/or explain the dynamics of the case or cases under examination. Another approach is to use vignettes, or "small illustrative stor[ies]" (Grbich 2007:214) created by summarizing data or consolidating data from different sources. Vignettes can be used to bring attention to particular themes, patterns, or types of experiences. Similarly, anecdotes are short tales that are used to illustrate particularly central ideas within the discussion. Narratives can be descriptive—simply using the data to tell a story—or theoretical and analytical, where data is used to illustrate concepts and ideas (Taylor, Bogdan, and DeVault 2016).

Narrative approaches are used most frequently for ethnographic and archival data. Here, the common strategy is to craft what is called a **thick description**, a detailed account that is rich with details about the observations, actors, and contexts, with analytical points highlighted in the course of the narrative (Berg 2009). Researchers using such a strategy often also employ metaphor to help them make sense of the ideas that they are working with. Berg (2009:236), for example, uses "revolving-door justice," a phrase used to refer to a situation in which people rapidly move into and out of the criminal justice system, as an example of a metaphor which is analytically useful.

Grounded theory is a particular approach to both data collection and analysis in which researchers collect data, identify themes in the data, review and compare the data to see if it fits the themes and theoretical concepts, collect additional data to continue to refine the analytical ideas, and ultimately build theory that reflects what the data has to say about the world (Taylor, Bogdan, and DeVault 2016). When utilizing a grounded theory approach, researchers must take care to ensure that theories they build are in fact grounded in data, rather than in prior knowledge or scholarship (Bergin 2018). Thus, grounded theory is a highly inductive approach to research and data analysis. A grounded theory approach to data analysis involves open coding, the writing of theoretical memos, further selective coding to validate categories and ideas, and integration of the empirical data with the theoretical memos (Grbich 2007).

Narrative approaches can also center the use of quotes from participants. Grbich (2007) discusses several ways to do this:

- *Layering*, or the interweaving of different perspectives to show how views and experiences may diverge;
- *Pastiche*, in which many voices are brought together to create an overall picture of experiences; and
- *Juxtaposition*, in which opposing views are contrasted to one another.

Most common is an approach in which a series of quotes exemplifying a particular theme or idea are presented together. Each quote must be analytically explained in the body of the text—they cannot simply stand on their own. For example, Figure 3 below presents an excerpt from the section of Kathryn Edin's paper on motherhood and marriage among poor women that focuses on control. In it, you can see how a theme is discussed and a series of quotes showcasing respondents' words relating to this theme are presented, with the meaning and relevance of each quote explained before or after the quote.

Whether to choose memo-writing, data display, or narrative approaches to analysis, or some combination of two or more of these approaches, is determined both by researchers' personal styles of research and analysis and by the nature, type, and complexity of the data. For instance, a multifaceted case study of an organization with multiple departments may have a story which is too complex for a narrative approach and thus requires the researcher to find other ways of simplifying and organizing the data. An ethnography of a girls' soccer team over the course of a single season, though, might lend itself well to a narrative approach utilizing thick description. And interviews with people recovering from surgery about their experiences might best be captured through a narrative approach focusing on quotes from participants.

In a non-marital relationship, women often felt they had more control than they would have had if they married. Even if the couple cohabited, they nearly always lived with her mother or in an apartment with her name on the lease. Thus, mothers had the power to evict fathers if they interfered with child rearing, or tried to take control over financial decision-making. Mothers said that fathers who knew they were "on trial" could do little about this state of affairs, especially since they needed a place to live and could not generally afford one on their own. One African American Philadelphia-area respondent's partner quipped, "her attitude is like, 'it's either my way or the highway. "

Why was control, not power, such an important issue for these women? Most mothers said they thought their children's fathers had very traditional notions of sex roles—notions that clashed with their more egalitarian views. One white cohabiting mother from Charleston said, "If we were to marry, I don't think it would be so ideal. [Husbands] want to be in charge, and I can't deal with that." Regardless of whether or not the prospective wife worked, mothers feared that prospective husbands would expect to be "head of the house," and make the "final" decisions about child rearing, finances, and other matters. Women, on the other hand, felt that since they had held the primary responsibility for both raising and supporting their children, they should have an equal say.

When we asked single mothers what they liked best about being a single parent, their most frequent response was "I am in charge," or "I am in control." Mothers seemed willing to take on the responsibilities of child rearing if they were also able to make and enforce the rules. In most mothers' views, the presence of fathers often interfered with their parental control, particularly if the couple married. Most women also felt that the presence of a husband might impede their efforts to discipline and spend time with their children. Mothers criticized men for being "too demanding" of their time and attention. A white Chicago mother

Figure 3. An Example of the Use of Quotes in a Narrative Approach (Edin 2000:121)

Making Conclusions

Theoretical memos, data displays, and narratives are not themselves conclusions or findings. Rather, they are tools and strategies that help researchers move from data towards findings. So how do researchers use these tools and strategies in service of their ultimate goal of making conclusions? Most centrally, this occurs by looking for patterns, comparisons, associations, or categories. And patterns are probably the most common and useful of these. There are a variety of types of patterns that researchers might encounter or look for. These include finding patterns of similarities, patterns of predictable differences, patterns of sequence or order, patterns of relationship or association, and patterns that appear to be causal (Saldaña 2016).

Making comparisons across cases is one way to look for patterns, and doing so also enhances researchers' ability to make claims about the representativeness of their data.

There are a variety of ways to make comparisons. Researchers can make predictions about what would happen in different sets of circumstances represented in the data and then examine cases to see whether they do or do not it with these predictions. They can look at individual variables or codes across cases to see how cases are similar or different. Or they can focus on categorizing whole cases into typologies. Categorization is an especially important part of research that has used coding. Depending on the approach taken, researchers tend to collapse groups of codes into categories either during the development of the coding strategy, as part of the coding process, or once it concludes, and it is these broader categories that then provide the core of the analysis (Saldaña 2016). To take some examples of approaches to analysis that might involve comparison, a researcher conducting ethnographic research in two different prisoner-reentry programs with different success rates might compare a series of variables across the two programs to see where the key differences arise. Or a project involving interviews of marketing professionals might categorize them according to their approach to the job to see which approaches have most helped people move forward in their careers.

Another strategy for developing conclusions, one that must be more tightly integrated into the data collection process, involves making predictions and then testing whether they come true by checking back with the site or the participant and seeing what happens later (Miles and Huberman 1994). For instance, consider an interview-based study looking at adult women returning to college and asking how they are adjusting. The researcher might learn about different strategies student respondents use and then might develop a hypothesis or prediction about which of these strategies will be most likely to lead to students staying enrolled in college past their first year. Then, the researcher can follow up with participants a year later to see if the strategies worked as predicted.

A final approach to developing conclusions involves the use of negative or deviant case methodologies. Deviant case methodologies are usually positioned as a sampling strategy in which researchers sample cases that seem especially likely to present problems for existing theory, often selecting on the dependent variable. However, deviant case methodologies can also be used long after sampling is completed. To do this, researchers sift through their data, looking for cases that do not conform to their theory, that do not fit the broader patterns they have observed from the body of their data, or that are as different as possible from other cases, and then they seek to understand what has shaped these differences. For instance, a project looking at why countries experience revolutions might collect data on a variety of countries to see if common theories hold up, zoning in on those cases that do not seem to it the theories. Or a project involving life-history interviews with men in prison for murder might devote special attention to those men whose stories seem least like the others.

While qualitative research does not have access to heuristics as simple as the null hypothesis significance testing approach used by quantitative researchers, qualitative researchers

can still benefit from the use of an approach informed by the idea of the null hypothesis (Berg 2009). In simple terms, the **null hypothesis** is the hypothesis that there is no relationship or association between the variables or concepts under study. Thus, qualitative researchers can approach their data with the assumption of a null hypothesis in mind, rather than starting from an assumption that whatever they are hoping to find will be displayed in their data. Such an approach reduces the likelihood that data becomes a self-fulfilling prophecy. **Deviant case** methodology—looking for examples within the data that do not fit the explanations researchers have developed and considering how the analysis can account for these aberrations (Warren and Karner 2015)—has a particular strength in this regard.

Testing Findings

Settling in on a set of findings does not mean a research project has been completed. Rather, researchers need to go through a process of testing, cross-checking, and verifying their conclusions to be sure they stand up to scrutiny. Researchers use a variety of approaches to accomplish this task, usually in combination.

One of the most important is consulting with others. Researchers discuss their findings with other researchers and other professional colleagues as well as with participants or other people similar to the participants (Warren and Karner 2015). These conversations give researchers the opportunity to test their logic, learn about questions others may have in regards to the research, refine their explanations, and be sure they have not missed obvious limitations or errors in their analysis. Presenting preliminary versions of the project to classmates, at conferences or workshops, or to colleagues can be particularly helpful, as can sharing preliminary drafts of the research write-up. Talking to participants or people like the participants can be especially important. While it is always possible that a research project will develop findings that are valid but that do not square with the lived experiences of participants, researchers should take care in such circumstances to responsibly address objections in ways that uphold the validity of both the research and the participants' experiences and to listen carefully to criticisms to be sure all potential errors or omissions in the analysis have been addressed. Feminist, critical social science, and action research perspectives especially value participants' expertise and encourage researchers to give participants final control over how they are portrayed in publication and presentation. For instance, feminist researchers often seek to ensure that relationships between researchers and participants are non-exploitative and empowering for participants (Grbich 2007), which may require that researchers do not position themselves as experts about participants' lives but rather provide participants with the opportunity to have their own lived experience shine through.

However, it is essential to be attentive to how participants' feedback is used. Such feedback can be extremely important to a project, but it can also misdirect analysis and inclusions in problematic ways. Participants vary greatly in their degree of comprehension of social scientific methods and language. This means researchers must strive to present their results to participants in ways that make sense to them. Participants' critiques of methods and language—while potentially illuminating—also could, if incorporated into the project, weaken its scientific strengths. In addition, there can be disagreements between different participants or groups of participants, as well as between participants and researchers, about explanations for the phenomena under consideration in the study.

While many qualitative researchers emphasize the important of participants' own knowledge about their social worlds, sometimes individuals are not the best at analyzing and understanding their own social circumstances. For example, consider someone you know whose romantic relationship is ending and ask them what happened. There is a good chance that the explanations offered by each partner are different, and maybe even that neither explanation matches what you as an outside observer see. Similarly, participants' narratives and explanations are a vital part of conclusion-building in qualitative research, but they are not the only possible conclusions a project can draw. In addition, attention to participants' views can sometimes lead researchers to self-censor, especially when researchers have ongoing relationships with participants or contexts and when participants' priorities and understandings are opposed to researchers'. Similarly, participants may use various discursive strategies—or ways of using language intended to have particular effects—that researchers may wish to critically interrogate. For example, what researchers call "race talk," or the discourse strategies people use to talk about and around race (Twine and Warren 2000; Van Den Berg, Wetherell, and Houtkoop-Steenstra 2003), can shed light on patterns of thought that participants may not be willing to openly admit.

Returning to participants to discuss findings and conclusions can also lead to new ethical, privacy, and even legal concerns as participants may be exposed to information gathered from others. While an interview-based study using appropriate care for confidentiality and in which participants do not know one another is not likely to raise these concerns, as long as data has been handled appropriately, in the case of ethnographic research or interviewing in which participants are acquainted, even careful attention to confidentiality may still leave the possibility that participants recognize one another in the narrative and analysis. Thus, it may be necessary to share only those sections of the data and analysis talking about a participant with that participant.

But researchers need not rely only on others to help them assess their work. There are a variety of steps researchers can take as part of the research process to test and evaluate their findings. For instance, researchers can critically re-examine their data, being sure their findings are firmly based on stronger data: data that was collected later in the research process after early weaknesses in collection methods were corrected, first-hand observa-

tions rather than occurrences the researcher only heard about later, and data what was collected in conditions with higher trust. They should also attend to the issue of **face validity**, the type of validity concerned with whether the measures used in a study are a good fit for the concepts. Sometimes, in the course of analysis, face validity falls away as researchers focus on exciting and new ideas, so returning to the core concepts of a study and ensuring the ultimate conclusions are based on measures that fit those concepts can help ensure solid conclusions.

Even if a study does not specifically use a deviant case methodology (as discussed above), researchers can take care to look for evidence that is surprising or that does not fit the model, theory, or predictions. If no such evidence appears—if it seems like all of the data conforms to the same general perspective—remember that the absence of such evidence does not mean the researcher's assumptions are correct. For example, imagine a research project designed to study factors that help students learn more math in introductory math classes. The researcher might interview students about their classes and find that the students who report that there were more visual aids used in their class all say that they learned a lot, while the students who report that their classes were conducted without visual aids say they did not learn so much. In this analysis, there may not have been any responses that did not fit this overall pattern. Clearly, then, something is going on within this population of students. But it is not necessarily the case that the use of visual aids impacts learning. Rather, the difference could be due to some other factor. Students might have inaccurate perception of how much they have learned and the use or lack of use of visual aids could impact these perceptions. Or visual aids might have a **spurious** relationship with students' perceptions of learning given some other variable, like the helpfulness of the instructor, that correlates with both perception and use of visual aids. Remember that a spurious relationship is a relationship in which two phenomena seem to vary in association with one another, but the observed association is not due to any causal connection between the two phenomena. Rather, the association is due to some other factor that is related to both phenomena but that has not been included in the analysis.

Careful attention to logic can also help researchers avoid making conclusions that turn out to be spurious. If an association is observed, then the researcher should consider whether that association is plausible or whether there might be alternative explanations that make more sense, turning back to the data as necessary. Indeed, researchers should always consider the possibility of alternative explanations, and it can be very helpful to ask others to suggest alternative explanations that have not yet been considered in the analysis. Not only does doing this increase the odds that a project's conclusions will be reliable and valid, it also staves off potential criticism from others who may otherwise remain convinced that their explanations are more correct.

Researchers should always make clear to others how they carried out their research. Providing sufficient detail about the research design and analytical strategy makes it possible

for other researchers to **replicate** the study, or carry out a repeat of the research designed to be as similar as possible to the initial project. Complete, accurate replications are possible for some qualitative projects, such as an analysis of historical newspaper articles or of children's videos, and thus providing the level of detail and clarity necessary for replication is a strength for such projects. It is far less possible for in-depth interviewing or ethnography to be replicated given the importance of specific contextual factors as well as the impact of interviewer effect. However, providing as much detail about methodological choices and strategies as possible, along with why these choices and strategies were the right ones for a given project, keeps the researcher and the project more honest and makes the approach more clear, as the goals of good research should include transparency.

Additionally, research projects involving multiple coders should have already undergone **inter-rater reliability** checks including at least 10% of the texts or visuals to be coded, and, if possible, even projects with only one coder should have conducted some inter-rater reliability testing. A discussion of the results of inter-rater reliability testing should be included in any publication or presentation drawing on the analysis, and if inter-rater reliability was *not* conducted for some reason this should be explicitly discussed as a limitation of the project. There are other types of limitations researchers must also clearly acknowledge, such as a lack of **representativeness** among respondents, small sample size, any issues that might suggest stronger-than-usual interviewer or **Hawthorne effects**, and other issues that might shape the reliability and validity of the findings.

There are a variety of other cautions and concerns that researchers should keep in mind as they build and evaluate their conclusions and findings. The term **anecdotalism** refers to the practice of treating anecdotes, or individual stories or events, as if they themselves are sufficient data upon which to base conclusions. In other words, researchers who are engaging in anecdotalism present snippets of data to illustrate or demonstrate a phenomenon without any evidence that these particular snippets are representative. While it is natural for researchers to include their favorite anecdotes in the presentation of their results, this needs to be done with attention to whether the anecdote illustrates a broader theme expressed throughout the data or whether it is an outlier. Without this attention, the use of anecdotes can quickly mislead researchers to misplace their focus and build unsupported conclusions. One of the most problematic aspects of anecdotalism is that it can enable researchers to focus on particular data because it supports the research hypothesis, is aligned with researchers' political ideals, or is exotic and attention-getting, rather than focusing on data that is representative of the results. The practice of anecdotalism is at the foundation of some people's perceptions that qualitative research is not rigorous or methodologically sound. In reality, bad or sloppy qualitative research, including that suffering from anecdotalism, is not rigorous, just as bad quantitative research is also not rigorous.

Qualitative researchers must take care to present excerpts and examples from their data. When researchers do not do this and instead focus their write-up on summaries (or

even numbers), readers are not able to draw their own conclusions about whether the data supports the findings. Respondents' actual words, words or images from documents, and ethnographers' first-hand observations are the true strength of qualitative research and thus it is essential that these things come through in the final presentation. Plus, if researchers focus on summaries or numbers, they may miss important nuances in their data that could more accurately shape the findings. On the other hand, researchers also must take care to avoid making overconclusions, or conclusions that go beyond what the data can support. Researchers risk making overconclusions when they assume data are representative of a broader or more diverse population than that which was included in the study, when they assume a pattern or phenomenon they have observed occurs in other types of contexts, and in similar circumstances when limited data cannot necessarily be extended to apply to events or experiences beyond the parameters of the study.

Another risk in qualitative research is that researchers might underemphasize theory. The role of theory marks one of the biggest differences between social science research and journalism. By connecting data to theory, social scientists have the ability to make broader arguments about social process, mechanisms, and structures, rather than to simply tell stories. Remember that one common goal in social science research is to focus on ordinary and everyday life and people, showing how—for instance—social inequality and social organizations structure people's lives, while journalism seeks stories that will draw attention.

Thinking Like a Researcher

This chapter has highlighted a variety of strategies for moving from data to conclusions. In the quantitative research process, moving from data to conclusions really is the analysis stage of research. But in qualitative research, especially inductive qualitative research, the process is more iterative, and researchers move back and forth between data collection, data management, data reduction, and analysis. It is also important to note that the strategies and tools outlined here are only a small sampling of the possible analytical techniques qualitative researchers use—but they provide a solid introduction to the qualitative research process. As you practice qualitative research and develop your expertise, you will continue to find new approaches that better fit your data and your research style.

So what ties all of these approaches to qualitative data analysis together? Among the most important characteristics is that the data needs to speak for itself. Yes, qualitative researchers may engage in data reduction due to the volume and complexity of the data they have collected, but they need to stay close enough to the data that it continues to shape the analysis and come alive in the write up.

Another very important element of qualitative research is **reflexivity**. Reflexivity, in the

context of social science research, refers to the process of reflecting on one's own perspective and **positionality** and how this perspective and positionality shape "research design, data collection, analysis, and knowledge production" (Hsiung 2008:212). The practice of reflexivity is one of the essential habits of mind for qualitative researchers. While researchers should engage in reflexivity throughout the research process, it is important to engage in a specifically reflexive thought process as the research moves towards conclusions. Here, researchers consider what they were thinking about their project, methodology, theoretical approach, topic, question, and participants when they began the research process, how these thoughts and ideas have or have not shifted, and how these thoughts and ideas—along with shifts in them—might have impacted the findings (Taylor, Bogdan, and DeVault 2016). They do this by "turn[ing] the investigative lens away from others and toward themselves" (Hsiung 2008:213), taking care to remember that the data they have collected and the data reduction and analysis strategies they have pursued result in records of interpretations, not clear, objective facts. Some feminist reflexive approaches involve talking through this set of concepts with participants; reflexive research may also involve having additional researchers serve as a kind of check on the research processes to ensure they comport with researchers' goals and ethical priorities (Grbich 2007).

Adjusting to the qualitative way of thinking can be challenging. New researchers who are accustomed to being capable students are generally used to being good at what they do—getting the right answers on tests, finding it easy to write papers that fulfill the professor's requirements, and picking up material from lectures or reading without much difficulty. Thus, they may end up thinking that if something is hard, they are probably falling short. And those who are accustomed to thinking o themselves as not such good students are often used to finding many of these academic activities hard and assuming that the fault lies within themselves. But one of the most important lessons we can learn from doing social science research is that neither of these sets of assumptions is accurate. In fact, doing research is hard by definition, and when it is done right, researchers are inevitably going to hit many obstacles and will frequently feel like they do not know what they are doing. This is not going to be because there is something wrong with the researcher! Rather, this is because that is how research works, because research involves trying to answer a question no one has answered before by collecting and analyzing data in a way no one has tried before. In Martin Schwartz's (2008:1771) words, faculty like those of us writing this book and teaching your class have been doing students a disservice by not making students "understand how hard it is to do research" and teaching them how we go about "confronting our absolute stupidity," the existential ignorance we all have when trying to understand the unknown. Schwartz says this kind of ignorance, which we choose to engage in when we pursue research, is highly productive, because it drives us to learn more. And that, after all, is the real point of research.

Exercises

1. Ask three friends or acquaintances to tell you what steps they took to get their most recent job. Create a process diagram of the job-searching process.
2. Create a network diagram of your class, with nodes representing each student and edges reflecting whether students knew one another before the semester began (or perhaps whether they are taking multiple courses together this semester).
3. Using a textual or video interview with a celebrity as your data (be sure it is an interview and not an article summarizing an interview), write a narrative case study of that celebrity's life, being sure to reference sociological concepts where appropriate.
4. Locate a work of long-form journalism about a topic of social science interest. Good publications to explore for this purpose include *The Atlantic*, *The New Yorker*, *The New York Times Magazine*, *Vanity Fair*, Slate, and longreads.com, among others. Summarize how the article might be different if it were an example of social science rather than journalism—what theory or theories might it draw on? What types of scholarly sources might it cite? How might its data collection have been different? How might data analysis have been conducted? What social science conclusions might it have reached?
5. Drawing on the "Conceptual Baggage Inventory Chart," (Hsiung 2008:219), identify your own research interests and goals; biographical characteristics; beliefs, values, and ideologies; and position in structures of stratification (including, but not limited to, race, gender, class, sexuality, age, and disability). Then consider how each of these might serve as potential advantages *and* as potential disadvantages in carrying out research design, data collection, and data analysis.

Media Attributions

- Process Diagram of Post-High-School Pathways © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Social-network © By Wykis - Own work is licensed under a Public Domain license
- Edin 2000 Excerpt © Kathryn Edin is licensed under a All Rights Reserved license

14. Presenting the Results of Qualitative Analysis

MIKAILA MARIEL LEMONIK ARTHUR

Qualitative research is not finished just because you have determined the main findings or conclusions of your study. Indeed, disseminating the results is an essential part of the research process. By sharing your results with others, whether in written form as scholarly paper or an applied report or in some alternative format like an oral presentation, an infographic, or a video, you ensure that your findings become part of the ongoing conversation of scholarship in your field, forming part of the foundation for future researchers. This chapter provides an introduction to writing about qualitative research findings. It will outline how writing continues to contribute to the analysis process, what concerns researchers should keep in mind as they draft their presentations of findings, and how best to organize qualitative research writing

As you move through the research process, it is essential to keep yourself organized. Organizing your data, memos, and notes aids both the analytical and the writing processes. Whether you use electronic or physical, real-world filing and organizational systems, these systems help make sense of the mountains of data you have and assure you focus your attention on the themes and ideas you have determined are important (Warren and Karner 2015). Be sure that you have kept detailed notes on all of the decisions you have made and procedures you have followed in carrying out research design, data collection, and analysis, as these will guide your ultimate write-up.

First and foremost, researchers should keep in mind that writing is in fact a form of thinking. Writing is an excellent way to discover ideas and arguments and to further develop an analysis. As you write, more ideas will occur to you, things that were previously confusing will start to make sense, and arguments will take a clear shape rather than being amorphous and poorly-organized. However, writing-as-thinking cannot be the final version that you share with others. Good-quality writing does not display the workings of your thought process. It is reorganized and revised (more on that later) to present the data and arguments important in a particular piece. And revision is totally normal! No one expects the first draft of a piece of writing to be ready for prime time. So write rough drafts and memos and notes to yourself and use them to think, and then revise them until the piece is the way you want it to be for sharing.

Bergin (2018) lays out a set of key concerns for appropriate writing about research. First, present your results accurately, without exaggerating or misrepresenting. It is very easy to overstate your findings by accident if you are enthusiastic about what you have found,

so it is important to take care and use appropriate cautions about the limitations of the research. You also need to work to ensure that you communicate your findings in a way people can understand, using clear and appropriate language that is adjusted to the level of those you are communicating with. And you must be clear and transparent about the methodological strategies employed in the research. Remember, the goal is, as much as possible, to describe your research in a way that would permit others to replicate the study. There are a variety of other concerns and decision points that qualitative researchers must keep in mind, including the extent to which to include quantification in their presentation of results, ethics, considerations of audience and voice, and how to bring the richness of qualitative data to life.

Quantification, as you have learned, refers to the process of turning data into numbers. It can indeed be very useful to count and tabulate quantitative data drawn from qualitative research. For instance, if you were doing a study of dual-earner households and wanted to know how many had an equal division of household labor and how many did not, you might want to count those numbers up and include them as part of the final write-up. However, researchers need to take care when they are writing about quantified qualitative data. Qualitative data is not as generalizable as quantitative data, so quantification can be very misleading. Thus, qualitative researchers should strive to use raw numbers instead of the percentages that are more appropriate for quantitative research. Writing, for instance, "15 of the 20 people I interviewed prefer pancakes to waffles" is a simple description of the data; writing "75% of people prefer pancakes" suggests a generalizable claim that is not likely supported by the data. Note that mixing numbers with qualitative data is really a type of mixed-methods approach. Mixed-methods approaches are good, but sometimes they seduce researchers into focusing on the persuasive power of numbers and tables rather than capitalizing on the inherent richness of their qualitative data.

A variety of issues of scholarly ethics and research integrity are raised by the writing process. Some of these are unique to qualitative research, while others are more universal concerns for all academic and professional writing. For example, it is essential to avoid plagiarism and misuse of sources. All quotations that appear in a text must be properly cited, whether with in-text and bibliographic citations to the source or with an attribution to the research participant (or the participant's pseudonym or description in order to protect confidentiality) who said those words. Where writers will paraphrase a text or a participant's words, they need to make sure that the paraphrase they develop accurately reflects the meaning of the original words. Thus, some scholars suggest that participants should have the opportunity to read (or to have read to them, if they cannot read the text themselves) all sections of the text in which they, their words, or their ideas are presented to ensure accuracy and enable participants to maintain control over their lives.

Audience and Voice

When writing, researchers must consider their audience(s) and the effects they want their writing to have on these audiences. The designated audience will dictate the **voice** used in the writing, or the individual style and personality of a piece of text. Keep in mind that the potential audience for qualitative research is often much more diverse than that for quantitative research because of the accessibility of the data and the extent to which the writing can be accessible and interesting. Yet individual pieces of writing are typically pitched to a more specific subset of the audience.

Let us consider one potential research study, an ethnography involving participant-observation of the same children both when they are at daycare facility and when they are at home with their families to try to understand how daycare might impact behavior and social development. The findings of this study might be of interest to a wide variety of potential audiences: academic peers, whether at your own academic institution, in your broader discipline, or multidisciplinary; people responsible for creating laws and policies; practitioners who run or teach at day care centers; and the general public, including both people who are interested in child development more generally and those who are themselves parents making decisions about child care for their own children. And the way you write for each of these audiences will be somewhat different. Take a moment and think through what some of these differences might look like.

If you are writing to academic audiences, using specialized academic language and working within the typical constraints of scholarly genres, as will be discussed below, can be an important part of convincing others that your work is legitimate and should be taken seriously. Your writing will be formal. Even if you are writing for students and faculty you already know—your classmates, for instance—you are often asked to imitate the style of academic writing that is used in publications, as this is part of learning to become part of the scholarly conversation. When speaking to academic audiences outside your discipline, you may need to be more careful about jargon and specialized language, as disciplines do not always share the same key terms. For instance, in sociology, scholars use the term diffusion to refer to the way new ideas or practices spread from organization to organization. In the field of international relations, scholars often used the term cascade to refer to the way ideas or practices spread from nation to nation. These terms are describing what is fundamentally the same concept, but they are different terms—and a scholar from one field might have no idea what a scholar from a different field is talking about! Therefore, while the formality and academic structure of the text would stay the same, a writer with a multidisciplinary audience might need to pay more attention to defining their terms in the body of the text.

It is not only other academic scholars who expect to see formal writing. Policymakers

tend to expect formality when ideas are presented to them, as well. However, the content and style of the writing will be different. Much less academic jargon should be used, and the most important findings and policy implications should be emphasized right from the start rather than initially focusing on prior literature and theoretical models as you might for an academic audience. Long discussions of research methods should also be minimized. Similarly, when you write for practitioners, the findings and implications for practice should be highlighted. The reading level of the text will vary depending on the typical background of the practitioners to whom you are writing—you can make very different assumptions about the general knowledge and reading abilities of a group of hospital medical directors with MDs than you can about a group of case workers who have a post-high-school certificate. Consider the primary language of your audience as well. The fact that someone can get by in spoken English does not mean they have the vocabulary or English reading skills to digest a complex report. But the fact that someone's vocabulary is limited says little about their intellectual abilities, so try your best to convey the important complexity of the ideas and findings from your research without dumbing them down—even if you must limit your vocabulary usage.

When writing for the general public, you will want to move even further towards emphasizing key findings and policy implications, but you also want to draw on the most interesting aspects of your data. General readers will read sociological texts that are rich with ethnographic or other kinds of detail—it is almost like reality television on a page! And this is a contrast to busy policymakers and practitioners, who probably want to learn the main findings as quickly as possible so they can go about their busy lives. But also keep in mind that there is a wide variation in reading levels. Journalists at publications pegged to the general public are often advised to write at about a tenth-grade reading level, which would leave most of the specialized terminology we develop in our research fields out of reach. If you want to be accessible to even more people, your vocabulary must be even more limited. The excellent exercise of trying to write using the 1,000 most common English words, available at the Up-Goer Five website (https://www.splasho.com/upgoer5/) does a good job of illustrating this challenge (Sanderson n.d.).

Another element of voice is whether to write in the first person. While many students are instructed to avoid the use of the first person in academic writing, this advice needs to be taken with a grain of salt. There are indeed many contexts in which the first person is best avoided, at least as long as writers can find ways to build strong, comprehensible sentences without its use, including most quantitative research writing. However, if the alternative to using the first person is crafting a sentence like "it is proposed that the researcher will conduct interviews," it is preferable to write "I propose to conduct interviews." In qualitative research, in fact, the use of the first person is far more common. This is because the researcher is central to the research project. Qualitative researchers can themselves be understood as research instruments, and thus eliminating the use of the first person in

writing is in a sense eliminating information about the conduct of the researchers themselves.

But the question really extends beyond the issue of first-person or third-person. Qualitative researchers have choices about how and whether to foreground themselves in their writing, not just in terms of using the first person, but also in terms of whether to emphasize their own subjectivity and reflexivity, their impressions and ideas, and their role in the setting. In contrast, conventional quantitative research in the positivist tradition really tries to eliminate the author from the study—which indeed is exactly why typical quantitative research avoids the use of the first person. Keep in mind that emphasizing researchers' roles and reflexivity and using the first person does not mean crafting articles that provide overwhelming detail about the author's thoughts and practices. Readers do not need to hear, and should not be told, which database you used to search for journal articles, how many hours you spent transcribing, or whether the research process was stressful—save these things for the memos you write to yourself. Rather, readers need to hear how you interacted with research participants, how your standpoint may have shaped the findings, and what analytical procedures you carried out.

Making Data Come Alive

One of the most important parts of writing about qualitative research is presenting the data in a way that makes its richness and value accessible to readers. As the discussion of analysis in the prior chapter suggests, there are a variety of ways to do this. Researchers may select key quotes or images to illustrate points, write up specific case studies that exemplify their argument, or develop vignettes (little stories) that illustrate ideas and themes, all drawing directly on the research data. Researchers can also write more lengthy summaries, narratives, and thick descriptions.

Nearly all qualitative work includes quotes from research participants or documents to some extent, though ethnographic work may focus more on thick description than on relaying participants' own words. When quotes are presented, they must be explained and interpreted—they cannot stand on their own. This is one of the ways in which qualitative research can be distinguished from journalism. Journalism presents what happened, but social science needs to present the "why," and the why is best explained by the researcher.

So how do authors go about integrating quotes into their written work? Julie Posselt (2017), a sociologist who studies graduate education, provides a set of instructions. First of all, authors need to remain focused on the core questions of their research, and avoid getting distracted by quotes that are interesting or attention-grabbing but not so relevant to the research question. Selecting the right quotes, those that illustrate the ideas and argu-

ments of the paper, is an important part of the writing process. Second, not all quotes should be the same length (just like not all sentences or paragraphs in a paper should be the same length). Include some quotes that are just phrases, others that are a sentence or so, and others that are longer. We call longer quotes, generally those more than about three lines long, **block quotes**, and they are typically indented on both sides to set them off from the surrounding text. For all quotes, be sure to summarize what the quote should be telling or showing the reader, connect this quote to other quotes that are similar or different, and provide transitions in the discussion to move from quote to quote and from topic to topic. Especially for longer quotes, it is helpful to do some of this writing before the quote to preview what is coming and other writing after the quote to make clear what readers should have come to understand. Remember, it is always the author's job to interpret the data. Presenting excerpts of the data, like quotes, in a form the reader can access does not minimize the importance of this job. Be sure that you are explaining the meaning of the data you present.

A few more notes about writing with quotes: avoid patchwriting, whether in your literature review or the section of your paper in which quotes from respondents are presented. Patchwriting is a writing practice wherein the author lightly paraphrases original texts but stays so close to those texts that there is little the author has added. Sometimes, this even takes the form of presenting a series of quotes, properly documented, with nothing much in the way of text generated by the author. A patchwriting approach does not build the scholarly conversation forward, as it does not represent any kind of new contribution on the part of the author. It is of course fine to paraphrase quotes, as long as the meaning is not changed. But if you use direct quotes, do not edit the text of the quotes unless how you edit them does not change the meaning and you have made clear through the use of ellipses (...) and brackets ([])what kinds of edits have been made. For example, consider this exchange from Matthew Desmond's (2012:1317) research on evictions:

> The thing was, I wasn't never gonna let Crystal come and stay with me from the get go. I just told her that to throw her off. And she wasn't fittin' to come stay with me with no money...No. Nope. You might as well stay in that shelter.

A paraphrase of this exchange might read "She said that she was going to let Crystal stay with her if Crystal did not have any money." Paraphrases like that are fine. What is not fine is rewording the statement but treating it like a quote, for instance writing:

> The thing was, I was not going to let Crystal come and stay with me from beginning. I just told her that to throw her off. And it was not proper for her to come stay with me without any money...No. Nope. You might as well stay in that shelter.

But as you can see, the change in language and style removes some of the distinct meaning of the original quote. Instead, writers should leave as much of the original language as

possible. If some text in the middle of the quote needs to be removed, as in this example, ellipses are used to show that this has occurred. And if a word needs to be added to clarify, it is placed in square brackets to show that it was not part of the original quote.

Data can also be presented through the use of data displays like tables, charts, graphs, diagrams, and infographics created for publication or presentation, as well as through the use of visual material collected during the research process. Note that if visuals are used, the author must have the legal right to use them. Photographs or diagrams created by the author themselves—or by research participants who have signed consent forms for their work to be used, are fine. But photographs, and sometimes even excerpts from archival documents, may be owned by others from whom researchers must get permission in order to use them.

A large percentage of qualitative research does not include any data displays or visualizations. Therefore, researchers should carefully consider whether the use of data displays will help the reader understand the data. One of the most common types of data displays used by qualitative researchers are simple tables. These might include tables summarizing key data about cases included in the study; tables laying out the characteristics of different taxonomic elements or types developed as part of the analysis; tables counting the incidence of various elements; and 2×2 tables (two columns and two rows) illuminating a theory. Basic network or process diagrams are also commonly included. If data displays are used, it is essential that researchers include context and analysis alongside data displays rather than letting them stand by themselves, and it is preferable to continue to present excerpts and examples from the data rather than just relying on summaries in the tables.

If you will be using graphs, infographics, or other data visualizations, it is important that you attend to making them useful and accurate (Bergin 2018). Think about the viewer or user as your audience and ensure the data visualizations will be comprehensible. You may need to include more detail or labels than you might think. Ensure that data visualizations are laid out and labeled clearly and that you make visual choices that enhance viewers' ability to understand the points you intend to communicate using the visual in question. Finally, given the ease with which it is possible to design visuals that are deceptive or misleading, it is essential to make ethical and responsible choices in the construction of visualization so that viewers will interpret them in accurate ways.

The Genre of Research Writing

As discussed above, the style and format in which results are presented depends on the audience they are intended for. These differences in styles and format are part of the **genre** of writing. Genre is a term referring to the rules of a specific form of creative or productive

work. Thus, the academic journal article—and student papers based on this form—is one genre. A report or policy paper is another. The discussion below will focus on the academic journal article, but note that reports and policy papers follow somewhat different formats. They might begin with an executive summary of one or a few pages, include minimal background, focus on key findings, and conclude with policy implications, shifting methods and details about the data to an appendix. But both academic journal articles and policy papers share some things in common, for instance the necessity for clear writing, a well-organized structure, and the use of headings.

So what factors make up the genre of the academic journal article in sociology? While there is some flexibility, particularly for ethnographic work, academic journal articles tend to follow a fairly standard format. They begin with a "title page" that includes the article title (often witty and involving scholarly inside jokes, but more importantly clearly describing the content of the article); the authors' names and institutional affiliations, an **abstract**, and sometimes keywords designed to help others find the article in databases. An abstract is a short summary of the article that appears both at the very beginning of the article and in search databases. Abstracts are designed to aid readers by giving them the opportunity to learn enough about an article that they can determine whether it is worth their time to read the complete text. They are written about the article, and thus not in the first person, and clearly summarize the research question, methodological approach, main findings, and often the implications of the research.

After the abstract comes an "introduction" of a page or two that details the research question, why it matters, and what approach the paper will take. This is followed by a literature review of about a quarter to a third the length of the entire paper. The literature review is often divided, with headings, into topical subsections, and is designed to provide a clear, thorough overview of the prior research literature on which a paper has built—including prior literature the new paper contradicts. At the end of the literature review it should be made clear what researchers know about the research topic and question, what they do not know, and what this new paper aims to do to address what is not known.

The next major section of the paper is the section that describes research design, data collection, and data analysis, often referred to as "research methods" or "methodology." This section is an essential part of any written or oral presentation of your research. Here, you tell your readers or listeners "how you collected and interpreted your data" (Taylor, Bogdan, and DeVault 2016:215). Taylor, Bogdan, and DeVault suggest that the discussion of your research methods include the following:

- The particular approach to data collection used in the study;
- Any theoretical perspective(s) that shaped your data collection and analytical approach;
- When the study occurred, over how long, and where (concealing identifiable details as

needed);
- A description of the setting and participants, including sampling and selection criteria (if an interview-based study, the number of participants should be clearly stated);
- The researcher's perspective in carrying out the study, including relevant elements of their identity and standpoint, as well as their role (if any) in research settings; and
- The approach to analyzing the data.

After the methods section comes a section, variously titled but often called "data," that takes readers through the analysis. This section is where the thick description narrative; the quotes, broken up by theme or topic, with their interpretation; the discussions of case studies; most data displays (other than perhaps those outlining a theoretical model or summarizing descriptive data about cases); and other similar material appears. The idea of the data section is to give readers the ability to see the data for themselves and to understand how this data supports the ultimate conclusions. Note that all tables and figures included in formal publications should be titled and numbered.

At the end of the paper come one or two summary sections, often called "discussion" and/or "conclusion." If there is a separate discussion section, it will focus on exploring the overall themes and findings of the paper. The conclusion clearly and succinctly summarizes the findings and conclusions of the paper, the limitations of the research and analysis, any suggestions for future research building on the paper or addressing these limitations, and implications, be they for scholarship and theory or policy and practice.

After the end of the textual material in the paper comes the bibliography, typically called "works cited" or "references." The references should appear in a consistent citation style—in sociology, we often use the American Sociological Association format (American Sociological Association 2019), but other formats may be used depending on where the piece will eventually be published. Care should be taken to ensure that in-text citations also reflect the chosen citation style. In some papers, there may be an appendix containing supplemental information such as a list of interview questions or an additional data visualization.

Note that when researchers give presentations to scholarly audiences, the presentations typically follow a format similar to that of scholarly papers, though given time limitations they are compressed. Abstracts and works cited are often not part of the presentation, though in-text citations are still used. The literature review presented will be shortened to only focus on the most important aspects of the prior literature, and only key examples from the discussion of data will be included. For long or complex papers, sometimes only one of several findings is the focus of the presentation. Of course, presentations for other audiences may be constructed differently, with greater attention to interesting elements of the data and findings as well as implications and less to the literature review and methods.

Concluding Your Work

After you have written a complete draft of the paper, be sure you take the time to revise and edit your work. There are several important strategies for revision. First, put your work away for a little while. Even waiting a day to revise is better than nothing, but it is best, if possible, to take much more time away from the text. This helps you forget what your writing looks like and makes it easier to find errors, mistakes, and omissions. Second, show your work to others. Ask them to read your work and critique it, pointing out places where the argument is weak, where you may have overlooked alternative explanations, where the writing could be improved, and what else you need to work on. Finally, read your work out loud to yourself (or, if you really need an audience, try reading to some stuffed animals). Reading out loud helps you catch wrong words, tricky sentences, and many other issues. But as important as revision is, try to avoid perfectionism in writing (Warren and Karner 2015). Writing can *always* be improved, no matter how much time you spend on it. Those improvements, however, have diminishing returns, and at some point the writing process needs to conclude so the writing can be shared with the world.

Of course, the main goal of writing up the results of a research project is to share with others. Thus, researchers should be considering how they intend to disseminate their results. What conferences might be appropriate? Where can the paper be submitted? Note that if you are an undergraduate student, there are a wide variety of journals that accept and publish research conducted by undergraduates. Some publish across disciplines, while others are specific to disciplines. Other work, such as reports, may be best disseminated by publication online on relevant organizational websites.

After a project is completed, be sure to take some time to organize your research materials and archive them for longer-term storage. Some Institutional Review Board (IRB) protocols require that original data, such as interview recordings, transcripts, and field notes, be preserved for a specific number of years in a protected (locked for paper or password-protected for digital) form and then destroyed, so be sure that your plans adhere to the IRB requirements. Be sure you keep any materials that might be relevant for future related research or for answering questions people may ask later about your project.

And then what? Well, then it is time to move on to your next research project. Research is a long-term endeavor, not a one-time-only activity. We build our skills and our expertise as we continue to pursue research. So keep at it.

Exercises

1. Find a short article that uses qualitative methods. The sociological magazine *Contexts* is a good place to find such pieces. Write an abstract of the article.
2. Choose a sociological journal article on a topic you are interested in that uses some form of qualitative methods and is at least 20 pages long. Rewrite the article as a five-page research summary accessible to non-scholarly audiences.
3. Choose a concept or idea you have learned in this course and write an explanation of it using the Up-Goer Five Text Editor (https://www.splasho.com/upgoer5/), a website that restricts your writing to the 1,000 most common English words. What was this experience like? What did it teach you about communicating with people who have a more limited English-language vocabulary—and what did it teach you about the utility of having access to complex academic language?
4. Select five or more sociological journal articles that all use the same basic type of qualitative methods (interviewing, ethnography, documents, or visual sociology). Using what you have learned about coding, code the methods sections of each article, and use your coding to figure out what is common in how such articles discuss their research design, data collection, and analysis methods.
5. Return to an exercise you completed earlier in this course and revise your work. What did you change? How did revising impact the final product?
6. Find a quote from the transcript of an interview, a social media post, or elsewhere that has not yet been interpreted or explained. Write a paragraph that includes the quote along with an explanation of its sociological meaning or significance.

SECTION IV

QUANTITATIVE DATA ANALYSIS WITH SPSS

This portion of this text provides details on how to perform basic quantitative analysis with SPSS, a statistical software package produced by IBM. Students and faculty can access discounted versions of the software via a variety of educational resellers, with lower-cost limited-term licenses for students that last six months to cover the time spent in a course (see IBM's list of resellers here). For most users, the GradPack Standard is the right option. Many colleges and universities also make SPSS available in campus computer labs or via a virtual lab environment, so check with your campus before assuming you need to pay for access. SPSS for non-students can be very expensive, though a 30-day free trial is available from IBM and should provide sufficient time to learn basic functions. Note that SPSS does offer screenreader capabilities, but users may need to install an additional plugin and may wish to seek technical support in advance for accomplishing this. Those looking for free, open-source statistical analysis software may want to consider R instead, though it does have a steeper learning curve. Hopefully, R supplements to this book will be available at some point in the future.

The examples and screenshots provided throughout this section of the book utilize data from the 2021 General Social Survey. The standard 2021 GSS file has been imported into SPSS and modified and simplified to produce an SPSS file that is available for download so users of this book can follow along with the examples. The number of variables has been reduced to 407, with most duplicated and survey-experiment variables removed as well as those that are difficult to use or that were responded to by only a very small number of people. Variable information has been adjusted and variables have been reordered to further simplify use. Finally, survey weights[1] have been removed from this dataset, as the proper use of survey weights is beyond the scope of this text. The dataset is thus designed only for learning purposes. Researchers who want to conduct actual analyses will need to download the original 2021 GSS file, import it into SPSS, and apply the survey weights. To learn more about survey weighting in the GSS, read this FAQ, and for instructions about applying survey weights in SPSS, see this handy guide from Kent State.

1. Survey weights are adjustments made to survey data to correct for the fact that certain populations have been under- or oversampled. For instance, because in the GSS only one person per household is sampled, individuals living in larger households have a lower chance of being selected into the survey. Thus, the survey weights adjust for household size in the calculation of results.

A simplified codebook is also available as part of this book (see Modified GSS Codebook for the Data Used in this Text). The codebook is an edited version of the 2021 GSS Codebook, with some technical detail removed and the variable list edited and simplified to match the dataset. Users of this book should take some time to familiarize themselves with the codebook before beginning to work with the data.

15. Quantitative Analysis with SPSS: Getting Started

MIKAILA MARIEL LEMONIK ARTHUR

This chapter focuses on getting started with SPSS. Note that before you can start to work with SPSS, you need to get your data into an appropriate format, as discussed in the chapter on Preparing Quantitative Data and Data Management. It is possible to enter data directly into SPSS, but the interface is not conducive to data entry and so researchers are better off entering their data using a spreadsheet program and then importing it.

Importing Data Into SPSS

In some cases, existing data will be able to be downloaded in SPSS format (*.sav is the file extension for an SPSS datafile), in which case it can be opened in SPSS by going to File → Open → Data and then locating the location of the file. However, in most cases, researchers will need to import data stored in another file format into SPSS. To import data, go to the file menu, then select import data. Next, choose the type of data you wish to import from the menu that appears. In most cases, researchers will be importing Excel or CSV data (when they have entered it themselves or are downloading it from a general-purpose site like the Census Bureau) or SAS or Stata data (when they are downloading it from a site that makes prepared statistical data files available).

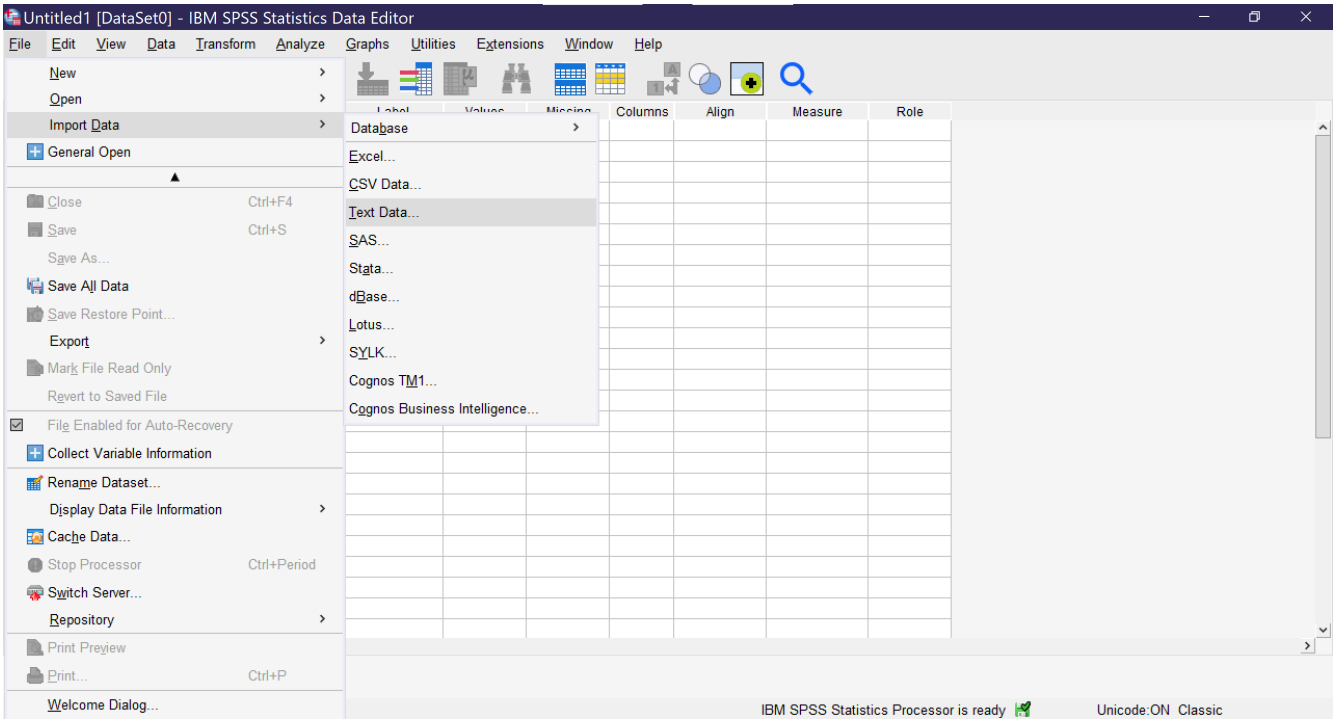


Figure 1. The Import Data Menu in SPSS

Once you click on a data type, a window will pop up for you to select the file you wish to import. Be sure it is of the file type you have chosen. If you import a file in a format that is already designed to work with statistical software, such as Stata, the importation process will be as seamless as opening a file. Researchers should be sure that immediately after importing, they save their file (File → Save As) so that it is stored in SPSS format and can be opened in SPSS, rather than imported, in the future. It is essential to remember that SPSS is not cloud-resident software and does not have an autosave function, so any time a file is changed, it must be manually saved.

If you import a file in Excel, CSV (comma-separated values) or text format, SPSS will open an import wizard with a number of steps. The steps vary slightly depending on which file type you are importing. For instance, to import an Excel file, as shown in Figure 2, you first need to specify the worksheet (if the file has multiple worksheets—SPSS can only import one worksheet at a time). You can choose to specify a limited range of cells. Checking the checkbox next to "Read variable names from first row of data" will replace the V1, V2, V3, and so on column headers with whatever appears in the top row of data in the Excel file. You can also choose to change the percentage of values that are used to determine data type, remove leading and trailing spaces from **string** values, and—if your Excel file has hidden rows or columns—you can choose to ignore them. Below the options, a preview of your Excel file will be shown; you can scroll through the preview to see that data is being displayed correctly. Clicking OK will finalize the import.

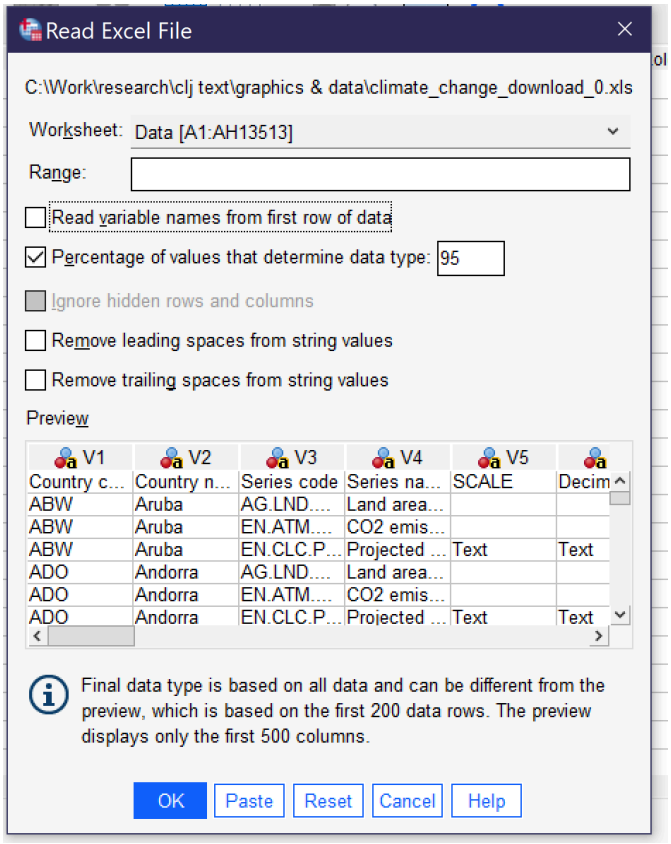


Figure 2. The Import Data Window for an Excel File

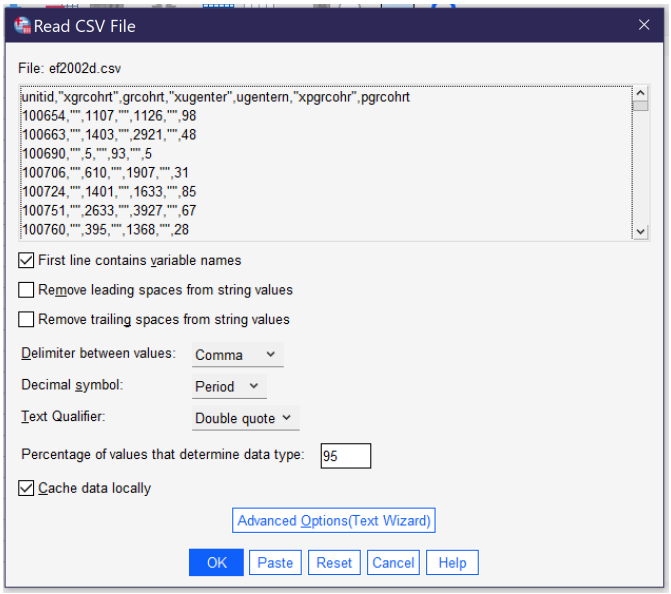


Figure 3. Window for Importing CSV Files

A different set of options appears when you import a CSV file, as shown in Figure 3. The top of the popup window shows a preview of the data in CSV format. While toggles related to whether the first line contains variable names, removing leading and trailing spaces, and indicating the percentage of values that determine the data type are the same as for importing data from Excel, there are additional options that are important for the proper importing of CSV data. First of all, the user must specify whether values are delimited by a comma, a semicolon, or a tab. While commas are the most common delimiters in

CSV files, the other delimiters are possible, and looking at the preview should make clear which of the delimiters is being used in a given file, as shown in the example below.

Comma-delimited:	`1,2312,"Yes",984`
Semicolon-delimited:	`1;2312;"Yes";984`
Tab-delimited:	`1 2312 "Yes" 984`

Second, the user must specify whether the period or the comma is the decimal symbol. Data produced in the United States typically uses the period (as in 1238.67), as does data produced in many other English-speaking countries, while most of Europe and Latin America use the comma. Third, the user must specify the text qualifier (single quotes, double quotes, or none). This is the character used to note that the contents of a particular entry in the CSV file are textual (string variables) in nature, not numerical. If your data includes text, it should be clear from the preview which qualifier is being used. Users can also toggle whether data is cached locally or not; caching locally speeds the importation process.

Finally, there is a button for Advanced Options (Text Wizard). The text wizard offers the same window and options that users see if they are importing a text file directly, and this wizard offers more direct control over the importation process over a series of six steps. First, users can specify a predefined format if they have a *.tpf file on their computers (this is rare) and see a preview of what the data in the file looks like. In step two, they can indicate if the file is delimited (as above) or fixed-width (where values are stored in columns of constant size specified within the file); which—if any—row contains the variable names; and the decimal symbol. Note that some forms of fixed-width files may not be supported. Third, they indicate which line of the file contains the first line of data, whether each line represents a case or a specific given number of variables represents a case, and how many cases to import. This last choice includes the option to import a random sample of cases. Fourth, users specify the delimiter and the text qualifier and determine how to handle leading and trailing spaces in string values. Fifth, users can double-check variable names and formats. Finally, before clicking the “Finish” button, users can choose to save their selections as a *.tpf file to be reused or to paste the syntax (to be discussed later in this chapter).

In all cases, once the importation options have been selected and OK or Finish has been clicked, the data is imported. An output window (see Figure 4) may open with various warnings and details about the importation process, and the Data View window (see Figure 5) will show the data, with variable names at the top of each column. At this point, be sure to save the dataset in a location and with a name you will be able to locate later.

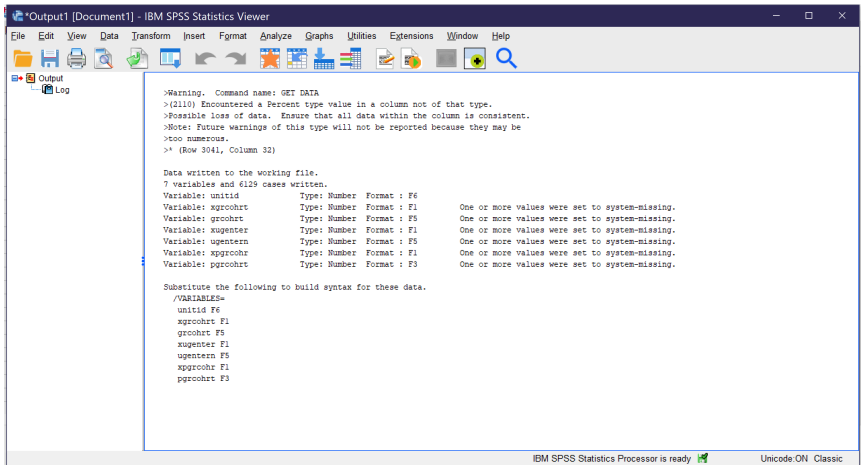


Figure 4. The SPSS Output Window

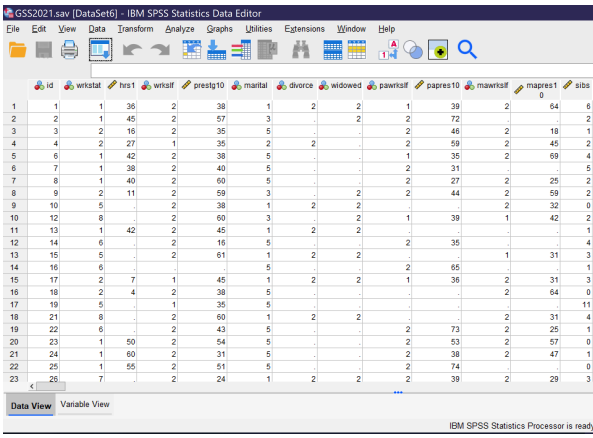


Figure 5. SPSS Data View

Before users are done setting up their dataset, they must be sure that appropriate variable information is included. When datasets are imported from other statistical programs, they will typically come with variable information. But when they are imported from Excel or CSV files, the variable information must be manually entered, typically from a codebook or related document. Variable information is entered using Variable View. Users can switch between Data View and Variable View by clicking the tabs at the bottom of the screen or using the Ctrl+T key combination. As you can see in Figure 6, a screenshot of a completed dataset, Variable View shows each variable in a row, with a variety of information about that variable. When a dataset is imported, each of these pieces of information need to be entered by hand for each variable. To move between columns by key commands, use the tab key; to open variable information that requires a menu for entry, click the space bar twice.

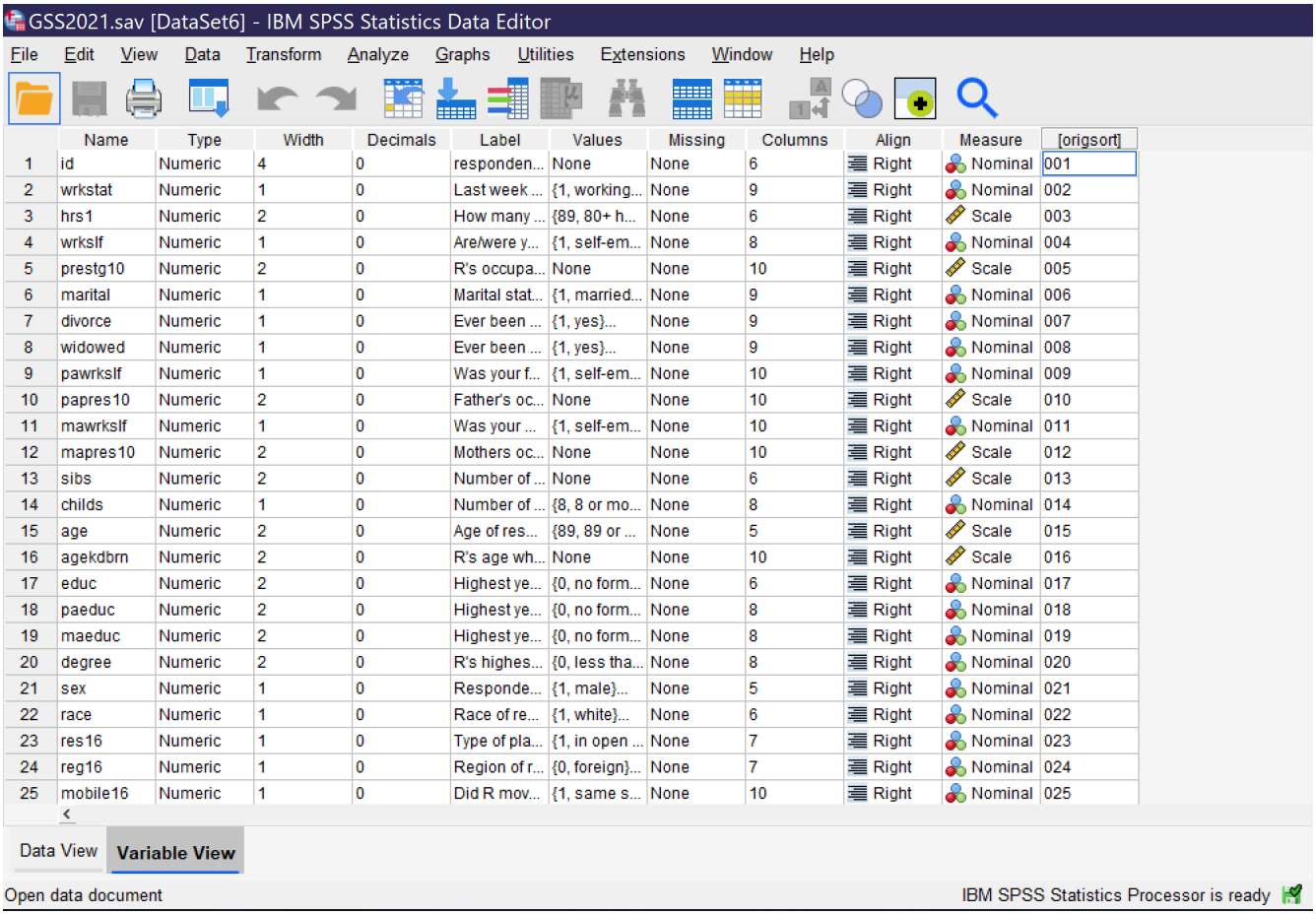


Figure 6. SPSS Variable View

- **Name** requires that each variable be given a short name, without any spaces. There are additional rules about names, but in short, names should be primarily alphanumeric in nature and cannot be words or use symbols that have meaning for the underlying computer processing. Names can be entered directly.
- **Type** specifies the variable type. To open up the menu allowing the selection of variable types, click on the cell, then click on the three dots [....] that appear on the right side of the cell. Users can then choose from among numeric, dollar, date, numeric with leading zeros, string, and other variable types.
- **Width** specifies the number of characters of width for the variable itself in data storage, while **decimals** specifies how many decimal places the variable will have. These can both be entered or edited directly or in the dialog box for Type.
- **Label** provides space for a longer variable name that spells out

more completely what the variable is measuring. It can be entered directly.

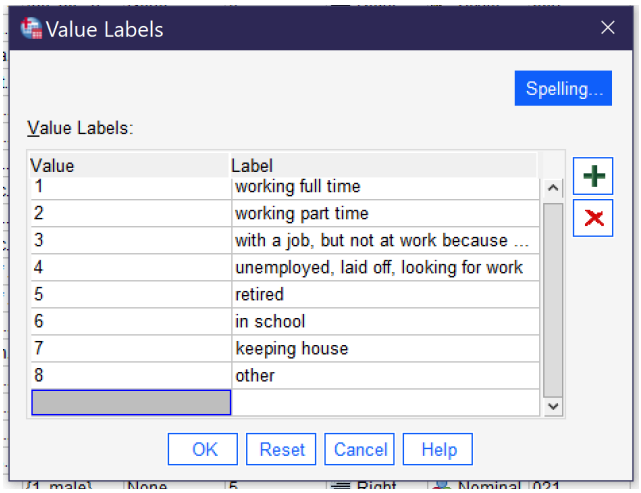


Figure 7. Value Labels Popup Window

- **Values** is where the **attributes** or value labels for a variable are specified. Clicking the three dots [....]—remember, they are not visible until you click in a cell—opens a dialog box in which values and their labels can be entered, as shown in Figure 7. To enter a value and its label, click on the green plus sign. Then enter the numerical value under the "Value" column and the value label under the "Label" column, and continue doing this until all values are labled. Labels can be long, but the beginning portions should be easily distinguishable so analysts can work with them even when the entire label is not displayed. There is a "Spelling..." button for spell-checking your work. Use the red X to delete a value and its label.

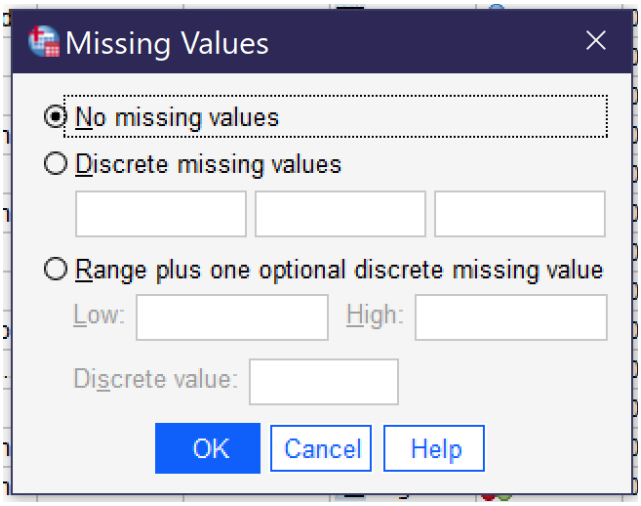


Figure 8. Missing Values Popup

- **Missing** provides for the indication that particular values—like "refused to answer"—should be treated by the SPSS software as missing data rather than as analytically useful categories. Clicking the three dots [....] opens a dialog box for specifying missing values. When there are no missing values, "no missing values" should be selected. Otherwise, users can select "discrete missing values" and then enter three specific missing values—the numerical values, not the value labels—or they can elect "range plus one optional discrete missing value" to specific a range from low to high of missing values, optionally adding an additional single discrete value.
- **Columns** specifies the width of the display column for the variable. It can be entered directly.
- **Align** specifies whether the variable data will be aligned right, center, or left. Users can click in the cell to make a menu appear or can press spacebar twice and then use arrows to select the desired alignment.
- **Measure** permits the indication of level of measurement from among nominal, ordinal, and scale variables. Users can click in the cell to make a menu appear or can press

spacebar twice and then use arrows to select the desired level of measurement. Note that measure is often wrong in datasets and analysts should not rely on it in determining the level of measurement for selection of statistical tests; SPSS does not use this characteristic when running tests.

- Some datasets will have additional criteria. For example, the dataset shown in Figure 6 has a column called origsort which displays the original sort order of the dataset, so that if an analyst sorts the variables they can be returned to their original order.

When entering variable information, it is especially important to include Name, Label, and Values and be sure Type is correct and any Missing values are specified. Other variable information is less crucial, though clearly it is better to fully specify all variable information. Once all variable information is entered and double-checked and the dataset has been saved, it is ready for use.

Using SPSS

When a user first opens SPSS, they are greeted with the "Welcome Dialog" (see figure 9). This dialog provides tips, links to help resources, and options for creating a new file (by selecting "new dataset") or opening recently used files. There is a checkbox for turning off the Welcome Dialog so that it will not be shown in the future.

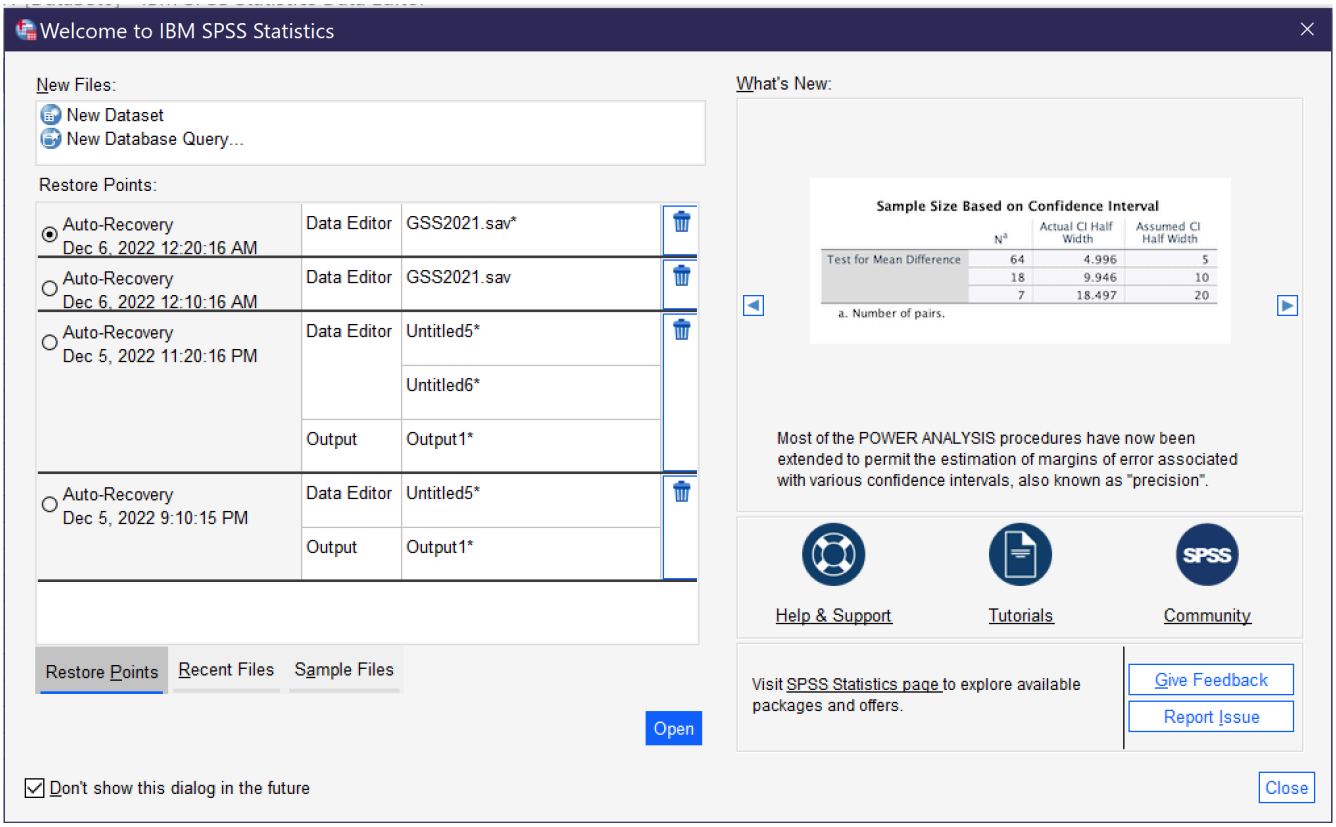


Figure 9. SPSS Welcome Dialog

When the Welcome Dialog is turned off, SPSS opens with a blank file. Going to File → Open → Data (Alt+F, O, D) brings up the dialog for opening a data file; the Open menu also provides for opening other types of files, which will be discussed below. Earlier in this chapter, the differences between Data View and Variable view were discussed; when you open a data file, be sure to observe which view you are using.

It can be useful to be able to search for a variable or case in the datafile. There are two main ways to do this, both under the Edit menu (Alt+E).[1] The Edit menu offers Find and Go To. Find, which can also be accessed by pressing Ctrl+F, allows users to search for all or part of a variable name. Figure 10 displays the Search dialog, with options shown after clicking on the "show options" button. (Users can also use the Replace function, but this carries the risk of writing over data and so should be avoided in almost all cases.) Be sure to select the column you wish to search—the Find function can only examine one column in Variable View at a time. Most typically, users will want to search variable names or labels. The checkbox for Match Case toggles whether or not case (in other words, capitalization) matters to the search. Expanding the options permits users to specify how much and which part of a cell must be matched as well as search order.

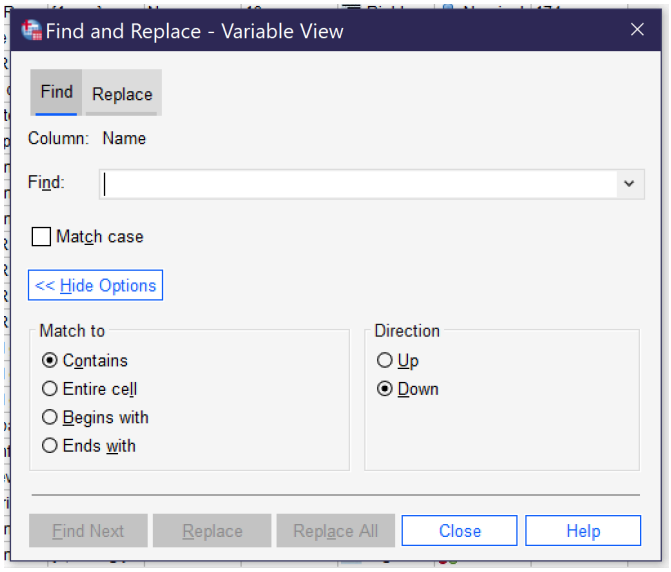


Figure 10. Find and Replace Dialog in SPSS

Users can also navigate to specific variables by using the Edit → Go to Case (to navigate to a specific case—or row in data view) and Edit → Go to Variable (to navigate to a specific variable—a row in variable view or a column in data view). Users can also access detailed variable information via the tool Utilities → Variables.

Another useful feature is the ability to sort variables and cases. Both types of sorting can be found in the data menu. Variables can be sorted by any of the characteristics in variable view; when sorting, the original sort order can be saved as a new characteristic. Cases can be sorted on any variable.

SPSS Options

The Options dialog can be reached by going to Edit → Options (or Alt+E, Alt+N). There are a wide variety of options available to help users customize their SPSS experience, a few of which are particularly important. First of all, using various dialogs and menus in the program is much easier if the options Variable List—Display Names (Alt+N) and Alphabeti-

1. Note that "Search," another option under the Edit menu, does not search variables or cases but instead launches a search of SPSS web resources and help files.

cal (Alt+H) are selected under General. You can also change the display language for both the user interface and for output under Language, change fonts and colors for output under Viewer, set number options under Data; change currency options under Currency; set default output for graphs and charts under Charts; and set default file locations for saving files under File locations. While most of these options can be left on their default settings, it is really important for most users to set variables to display names and alphabetical before use. Options will be preserved if you use the same computer and user account, but if you are working on a public computer you should get in the habit of checking every time you start the program.

Getting More Out of SPSS

So far, we have been working only with Data View and Variable View in the main dataset window. But when researchers produce the results of an analysis, these results appear in a new window called Output—IBM SPSS Statistics Viewer. New Output windows can be opened from the File menu by going to Open → Output or from the Window menu by selecting "Go to Designated Viewer Window" (the later command also brings the output window to the foreground if one is already open). Output will be discussed in more detail when the results of different tests are discussed. For now, note that output can be saved in *.spv format, but this format can only be viewed in SPSS. To save output in a format viewable in other applications, go to File → Export, where you can choose a file location and a file format (like Word, PowerPoint, HTML, or PDF). Individual output items can also be copied and pasted.

SPSS also offers a Syntax viewer and editor, which can also be accessed from both the File and Window menus. While syntax is beyond the scope of this text, it provides the option for writing code (kind of like a computer program) to control SPSS rather than using menus and buttons in a graphical user interface. Experienced users, or those doing many similar repetitive tasks, often find working via syntax to be faster and more efficient, but the learning curve is quite steep. If you are interested in learning more about how to write syntax in SPSS, Help → Command Syntax Reference brings up a very long document detailing the commands available.

Finally, the Help menu in SPSS offers a variety of options for getting help in using the program, including links to web resource guides, PDF documentation, and help forums. These tools can also be reached directly via the SPSS website. In addition, many dialog boxes contain a "Help" button that takes users to webpages with more detail on the tool in question.

Exercises

Go to https://www.baseball-reference.com/ and select 10 baseball players of your choice. In an Excel or other spreadsheet, enter the name, position, batting arm, throwing arm, weight in pounds, and height in inches, as well as, from the Summary: Career section, HR (home runs) and WAR (wins above replacement). Each player should get one row of the Excel spreadsheet. Once you have entered the data, import it into SPSS. Then use Variable View to enter the relevant information about each variable—including value labels for position, batting arm, and throwing arm. Sort your cases by home runs. Finally, save your file.

Media Attributions

- import menu
- import excel © IBM SPSS is licensed under a All Rights Reserved license
- import csv © IBM SPSS is licensed under a All Rights Reserved license
- output window © IBM SPSS is licensed under a All Rights Reserved license
- spss data view © IBM SPSS is licensed under a All Rights Reserved license
- variable-view © IBM SPSS is licensed under a All Rights Reserved license
- value labels © IBM SPSS is licensed under a All Rights Reserved license
- missing values © IBM SPSS is licensed under a All Rights Reserved license
- welcome dialog © IBSM SPSS is licensed under a All Rights Reserved license
- find and replace © IBM SPSS is licensed under a All Rights Reserved license

16. Quantitative Analysis with SPSS: Univariate Analysis

MIKAILA MARIEL LEMONIK ARTHUR

The first step in any quantitative analysis project is **univariate** analysis, also known as **descriptive statistics**. Producing these measures is an important part of understanding the data as well as important for preparing for subsequent bivariate and multivariate analysis. This chapter will detail how to produce **frequency distributions** (also called frequency tables), **measures of central tendency**, **measures of dispersion**, and graphs in SPSS. The chapter on Univariate Analysis provides details on understanding and interpreting these measures. To select the correct measures for your variables, first determine the **level of measurement** of each **variable** for which you want to produce appropriate descriptive statistics. The distinction between **binary** and other **nominal** variables is important here, so you need to determine whether each variable is binary, nominal, **ordinal**, or **continuous**. Then, use Table 1 to determine which descriptive statistics you should produce.

Table 1. Selecting the Right Univariate/Descriptive Statistics

	Measures of Central Tendency	Measures of Dispersion	Graphs
Binary	Mean; Mode	Frequency distribution	Pie Chart; Bar graph
Nominal	Mode	Frequency distribution	Pie Chart; Bar Graph
Ordinal	Median; Mode	Range (min/max); Frequency distribution; occasionally Percentiles	Bar Graph
Continuous	Mean; Median	Standard deviation; Variance; Range (min/max); Skewness; Kurtosis; Percentiles	Histogram

Producing Descriptive Statistics

Other than graphs, all of the univariate analyses discussed in this chapter are produced by going to Analyze → Descriptive Statistics → Frequencies, as shown in Figure 1. Note that SPSS also offers a tool called Descriptives; avoid this unless you are specifically seeking to pro-

duce **Z scores**, a topic beyond the scope of this text, as the Descriptives tool provides far fewer options than the Frequencies tool.

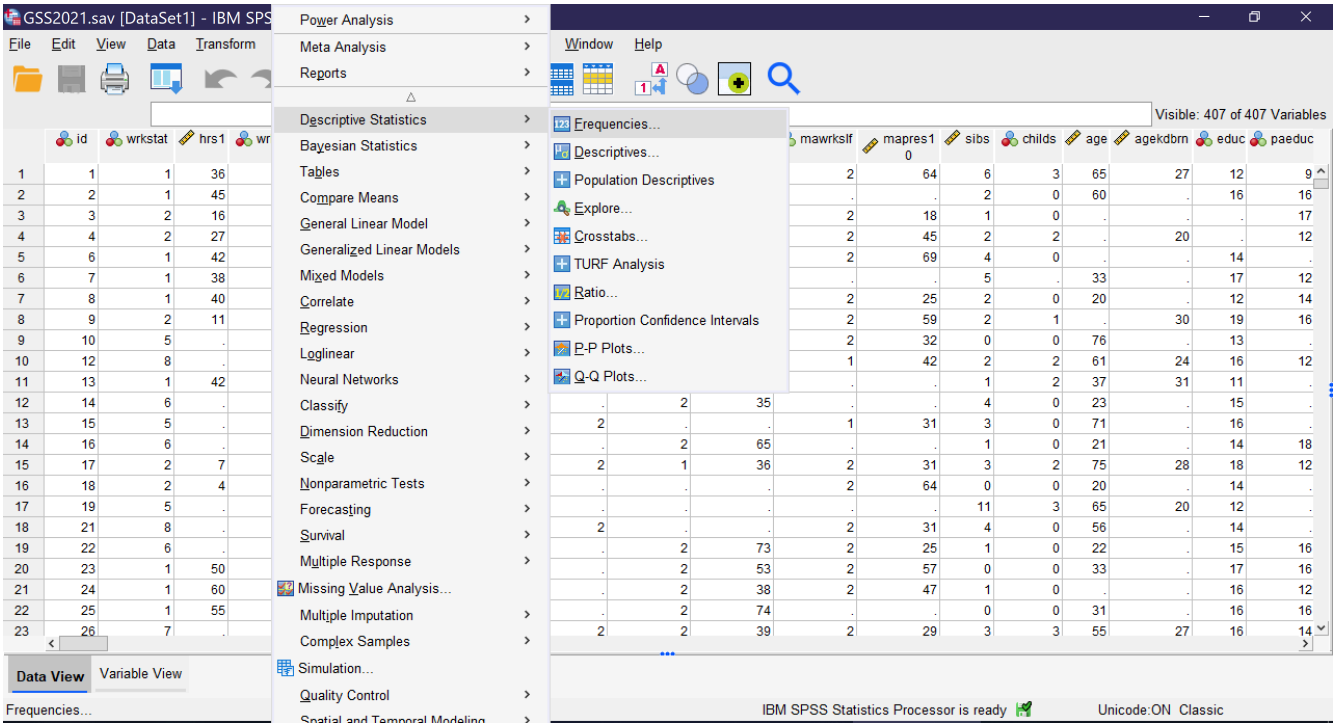


Figure 1. Running Descriptive Statistics in SPSS

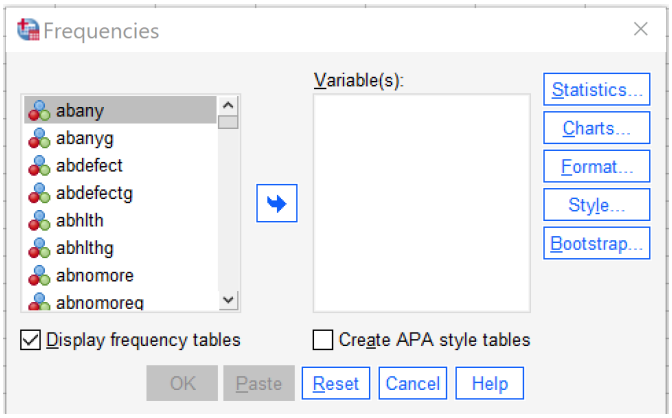


Figure 2. The Frequencies Window

Selecting this tool brings up a window called “Frequencies” from which the various descriptive statistics can be selected, as shown in Figure 2. In this window, users select which variables to perform univariate analysis upon. Note that while univariate analyses can be performed upon multiple variables as a group, those variables need to all have the same level of measurement as only one set of options can be selected at a time.

To use the Frequencies tool, scroll through the list of variables on the left side of the screen, or click in the list and begin typing the variable name if you remember it and the list will jump to it. Use the blue arrow to move the variable into the Variables box or grab and drag it over. If you are performing analysis on a binary, nominal, or ordinal variable, be sure the checkbox next to “Display frequency tables” is checked; if you are performing analysis on a continuous variable, leave that box unchecked. The checkbox for “Create

APA style tables" slightly alters the format and display of tables. If you are working in the field of psychology specifically, you should select this checkbox, otherwise it is not needed. The options under "Format" specify elements about the display of the tables; in most cases those should be left as the default. The options under "Style" and "Bootstrap" are beyond the scope of this text.

It is under "Statistics" that the specific descriptive statistics to be produced are selected, as shown in Figure 3. First, users can select several different options for producing percentiles, which are usually produced only for continuous variables but occasionally are used for ordinal variables. Quartiles produces the 25th, 50th (median), and 75th percentile in the data. Cut points allows the user to select a specified number of equal groups and see at which values the groups break. Percentiles allows the user to specify specific percentiles to produce—for instance, a user might want to specify 33 and 66 to see where the upper, middle, and lower third of data fall.

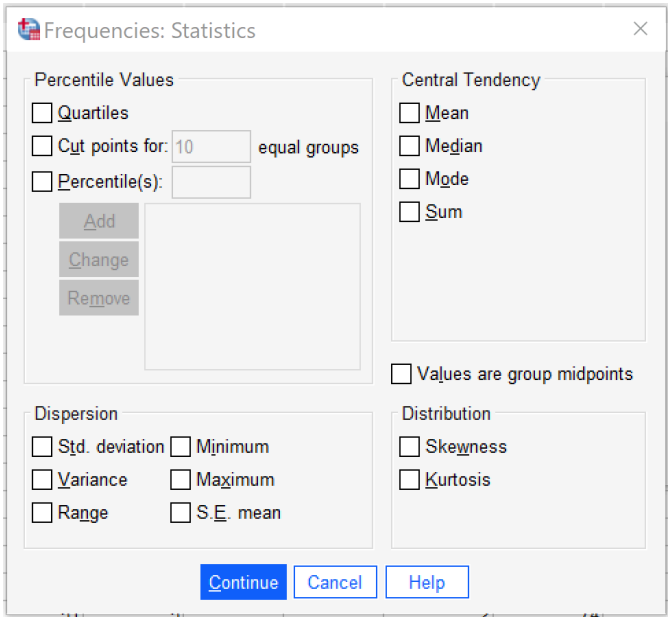


Figure 3. The Dialog Box for Selecting Descriptive Statistics

Second, users can select measures of central tendency, specifically the **mean** (used for binary and continuous variables), the **median** (used for ordinal and continuous variables), and the **mode** (used for binary, nominal, and ordinal variables). Sum adds up all the values of the variable, and is not typically used. There is also an option to select if values are group midpoints, which is beyond the scope of this text.

Next, users can select measures of dispersion and distribution, including the **standard deviation** (abbreviated here Std. deviation, and used for continuous variables), the **variance** (used for continuous variables), the **range** (used for ordinal and continuous variables), the minimum value (used for ordinal and continuous variables), the maximum value (used for ordinal and continuous variables), and the standard error of the mean (abbreviated here as S.E. mean, this is a measure of sampling error and beyond the scope of this text), as well as **skewness** and **kurtosis** (used for continuous variables).

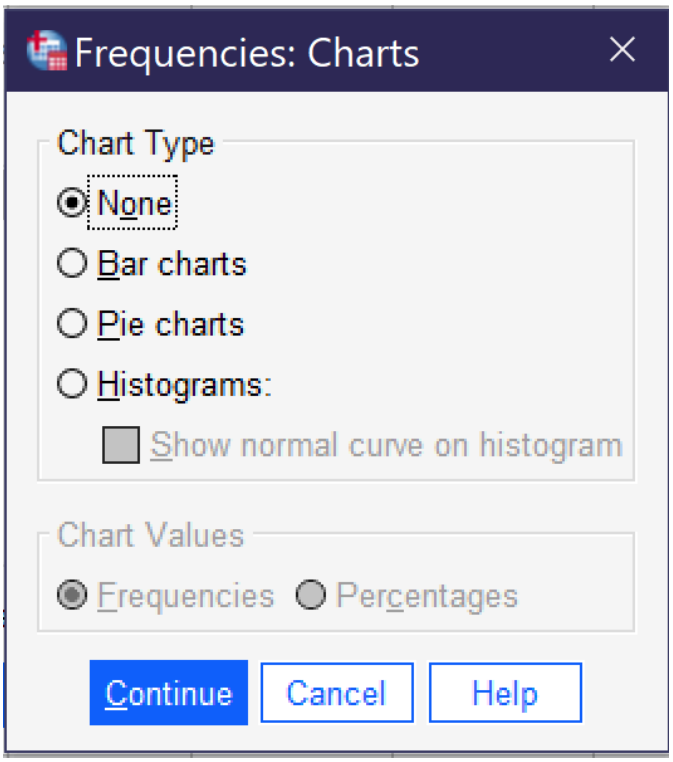


Figure 4. Making Graphs from the Frequencies Dialog

Once all desired tests are selected, click "Continue" to go back to the main frequencies dialog. There, you can also select the Chart button to produce graphs (as shown in Figure 4), though only one graph can be produced at a time (other options for producing graphs will be discussed later in this chapter). **Bar charts** are appropriate for binary, nominal, and ordinal variables. **Pie charts** are typically used only for binary variables and nominal variables with just a few categories, though they may at times make sense for ordinal variables with just a few categories. **Histograms** are used for continuous variables; there is an option to show the **normal curve** on the histogram, which can help users visualize the distribution more clearly. Users can also choose whether their graphs will be displayed in terms of frequencies (the raw count of values) or percentages.

Examples at Each Level of Measurement

Here, we will produce appropriate descriptive statistics for one variable from the 2021 GSS file at each level of measurement, showing what it looks like to produce them, what the resulting output looks like, and how to interpret that output.

A Binary Variable

To produce descriptive statistics for a binary variable, be sure to leave Display frequency tables checked. Under statistics, select Mean and Mode and then click continue, and under graphs select your choice of bar graph or pie chart and then click continue. Using the variable GUNLAW, then, the selected option would look as shown in Figure 5. Then click OK, and the results will appear in the Output window.

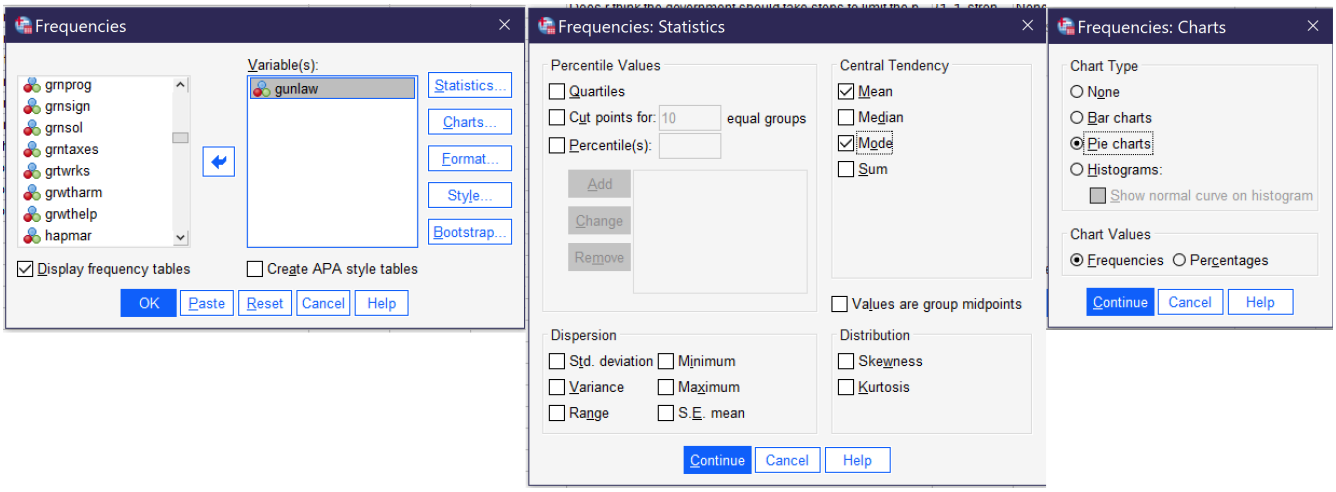


Figure 5. SPSS Dialogs Set Up for Descriptive Statistics for the Binary Variable GUNLAW

The output for GUNLAW will look approximately like what is shown in Figure 6. GUNLAW is a variable measuring whether the respondent favors or opposes requiring individuals to obtain police permits before buying a gun.

Statistics

Favor or oppose requiring gun permits

N	Valid	3992
	Missing	40
Mean		1.33
Mode		1

Favor or oppose requiring gun permits

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	favor	2686	66.6	67.3	67.3
	oppose	1306	32.4	32.7	100.0
	Total	3992	99.0	100.0	
Missing	System	40	1.0		
Total		4032	100.0		

Pie Chart
Favor or oppose requiring gun permits

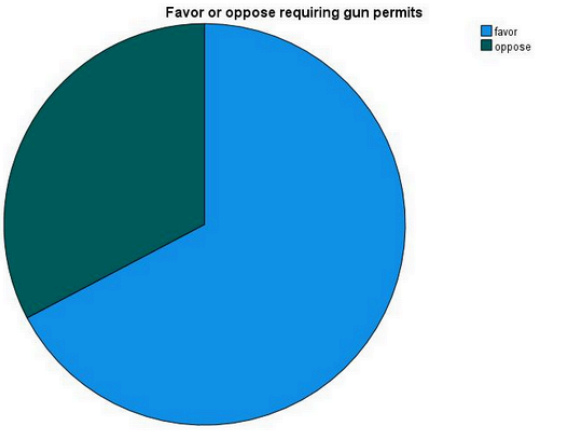


Figure 6. SPSS Output for Descriptive Statistics on GUNLAW

The output shows that 3,992 people gave a valid answer to this question, while responses for 40 people are missing. Of those who provided answers, the mode, or most frequent response, is 1. If we look at the value labels, we will find that 1 here means "favor;" in other words, the largest number of respondents favors requiring permits for gun owners. The mean is 1.33. In the case of a binary variable, what the mean tells us is the approximate proportion of people who have provided the higher-numbered value label—so in this case, about ⅓ of respondents said they are opposed to requiring permits.

The frequency table, then, shows the number and proportion of people who provided each answer. The most important column to pay attention to is Valid Percent. This column tells us what percentage of the people who answered the question gave each answer. So, in this case, we would say that 67.3% of respondents favor requiring permits for gun ownership, while 32.7% are opposed—and 1% are missing.

Finally, we have produced a pie chart, which provides the same information in a visual format. Users who like playing with their graphs can double-click on the graph and then right-click or cmd/ctrl click to change options such as displaying value labels or amounts or changing the color of the graph.

A Nominal Variable

To produce descriptive statistics for a nominal variable, be sure to leave Display frequency tables checked. Under statistics, select Mode and then click continue, and under graphs select your choice of bar graph or pie chart (avoid pie chart if your variable has many categories) and then click continue. Using the variable MOBILE16, then, the selected option would look as shown in Figure 7. Then click OK, and the results will appear in the Output window.

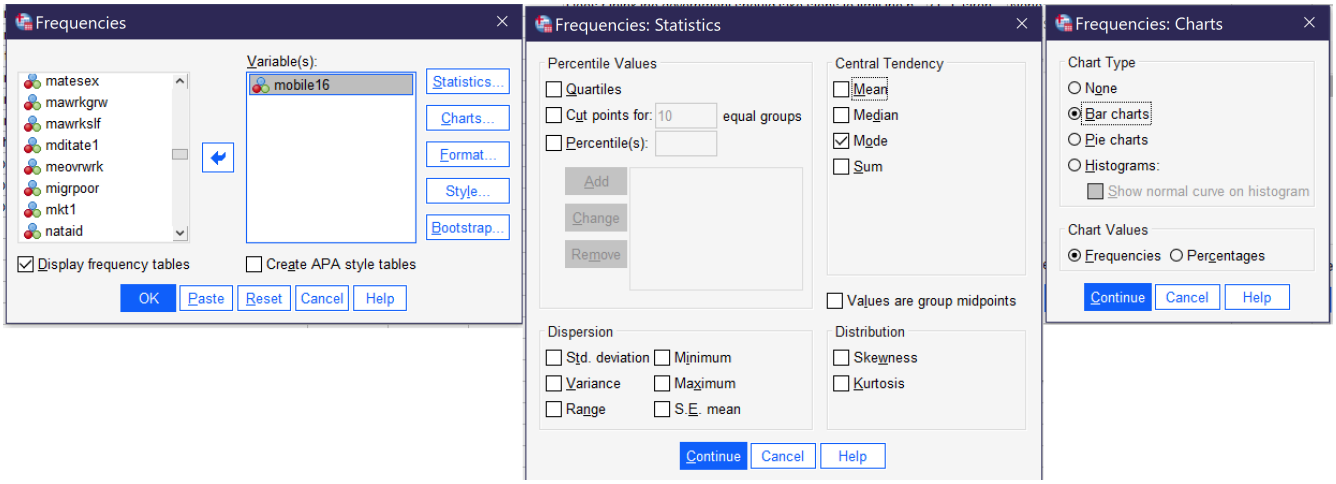


Figure 7. SPSS Dialogs Set Up for Descriptive Statistics for the Nominal Variable MOBILE16

The output will then look approximately like the output shown in Figure 8. MOBILE16 is a variable measuring respondents' degree of geographical mobility since age 16, asking them if they live in the same city they lived in at age 16; stayed in the same state they lived in at age 16 but now live in a different city; or live in a different state than they lived in at age 16.

Statistics

Did R move cities since age 16

N	Valid	3608
	Missing	424
Mode		2

Did R move cities since age 16

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	same state, same city	1087	27.0	30.1	30.1
	same state, different city	1294	32.1	35.9	66.0
	different state	1227	30.4	34.0	100.0
	Total	3608	89.5	100.0	
Missing	System	424	10.5		
Total		4032	100.0		

Bar Chart
Did R move cities since age 16

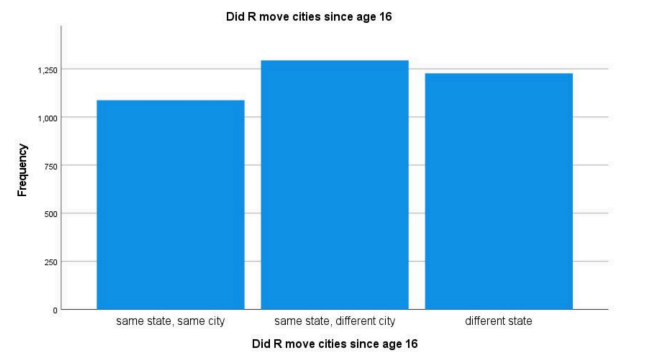


Figure 8. SPSS Output for Descriptive Statistics on MOBILE16

The output shows that 3608 respondents answered this survey question, while 424 did not. The mode is 2; looking at the value labels, we conclude that 2 refers to "same state, different city," or in other words that the largest group of respondents lives in the same state they lived in at age 16 but not in the same city they lived in at age 16. The frequency table shows us the percentage breakdown of respondents into the three categories. Valid percent is most useful here, as it tells us the percentage of respondents in each category *after those who have not responded to the question are removed*. In this case, 35.9% of people live in the same state but a different city, the largest category of respondents. Thirty-four percent live in a different state, while 30.1% live in the same city in which they lived at age 16. Below the frequency table is a bar graph which provides a visual for the

information in the frequency table. As noted above, users can change options such as displaying value labels or amounts or changing the color of the graph.

An Ordinal Variable

To produce descriptive statistics for an ordinal variable, be sure to leave Display frequency tables checked. Under statistics, select Median, Mode, Range, Minimum, and Maximum, and then click continue, and under graphs select your choice of bar graph and then click continue. Then click OK, and the results will appear in the Output window. Using the variable CARSGEN, then, the selected option would look as shown in Figure 7.

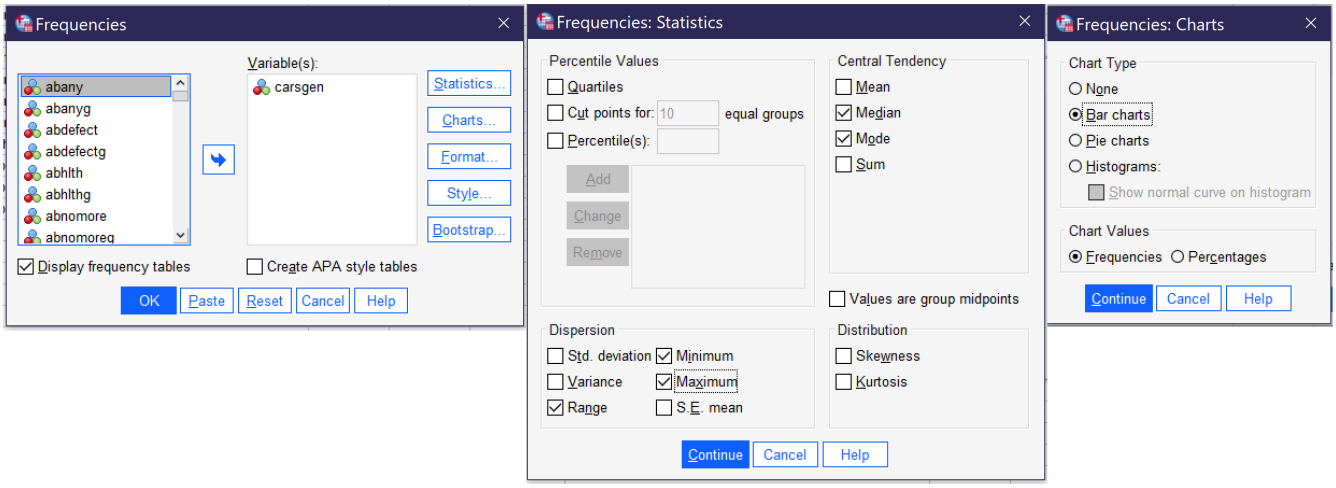


Figure 9. SPSS Dialogs Set Up for Descriptive Statistics for the Ordinal Variable CARSGEN

The output will then look approximately like the output shown in Figure 10. CARSGEN is an ordinal variable measuring the degree to which respondents agree or disagree that car pollution is a danger to the environment.

Statistics

Does r think car pollution is a danger to the environment

N	Valid	1778
	Missing	2254
Median		3.00
Mode		3
Range		4
Minimum		1
Maximum		5

Does r think car pollution is a danger to the environment

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	extremely dangerous	232	5.8	13.0	13.0
	very dangerous	559	13.9	31.4	44.5
	somewhat dangerous	814	20.2	45.8	90.3
	not very dangerous	151	3.7	8.5	98.8
	not dangerous at all	22	.5	1.2	100.0
	Total	1778	44.1	100.0	
Missing	System	2254	55.9		
Total		4032	100.0		

Bar Chart

Does r think car pollution is a danger to the environment

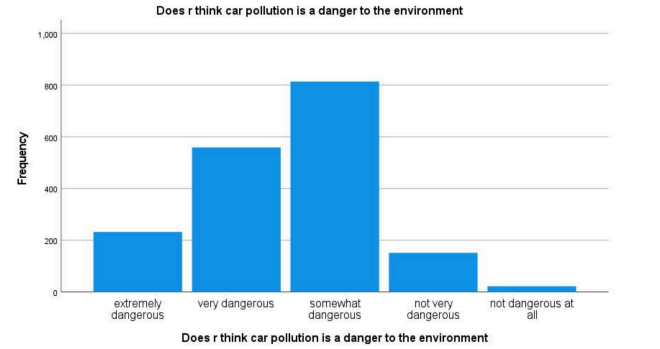


Figure 10. SPSS Output for Descriptive Statistics on CARSGEN

First, we see that 1778 respondents answered this question, while 2254 did not (remember that the GSS has a lot of questions; some are asked of all respondents while others are only asked of a subset, so the fact that a lot of people did not answer may indicate that many were not asked rather than that there is a high degree of nonresponse). The median and mode are both 3. Looking at the value labels tells us that 3 represents "somewhat dangerous." The range is 4, representing the maximum (5) minus the minimum (1)—in other words, there are five ordinal categories.

Looking at the valid percents, we can see that 13% of respondents consider car pollution extremely dangerous, 31.4% very dangerous, and 45.8%—the biggest category (and both the mode and median)—somewhat dangerous. In contrast only 8.5% think car pollution is not very dangerous and 1.2% think it is not dangerous at all. Thus, it is reasonable to conclude that the vast majority—over 90%—of respondents think that car pollution presents at least some degree of danger. The bar graph at the bottom of the output represents this information visually.

A Continuous Variable

To produce descriptive statistics for a continuous variable, be sure to uncheck Display frequency tables. Under statistics, go to percentile values and select Quartiles (or other percentile options appropriate to your project). Then select Mean, Median, Std. deviation, Variance, Range, Minimum, Maximum, Skewness, and Kurtosis and then click continue, and under graphs select Histograms and turn on Show normal curve on histogram and then click continue. Using the variable EATMEAT, then, the selected option would look as shown in Figure 11. Then click OK, and the results will appear in the Output window.

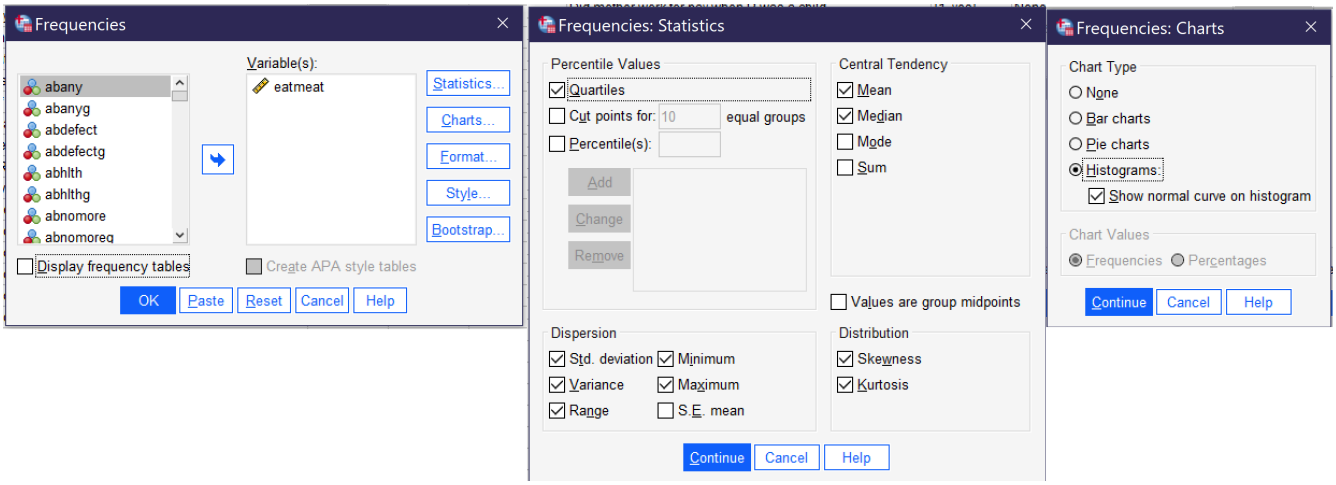


Figure 11. SPSS Dialogs Set Up for Descriptive Statistics for the Nominal Variable EATMEAT

The output will then look approximately like the output shown in Figure 12. EATMEAT is a continuous variable measuring the number of days per week that the respondent eats beef, lamb, or products containing beef or lamb.

Statistics

In a typical week, on how many days does r eat beef, lamb, or products containing them

N	Valid	1795
	Missing	2237
Mean		2.77
Median		3.00
Std. Deviation		1.959
Variance		3.838
Skewness		.541
Std. Error of Skewness		.058
Kurtosis		-.462
Std. Error of Kurtosis		.115
Range		7
Minimum		0
Maximum		7
Percentiles	25	1.00
	50	3.00
	75	4.00

Histogram

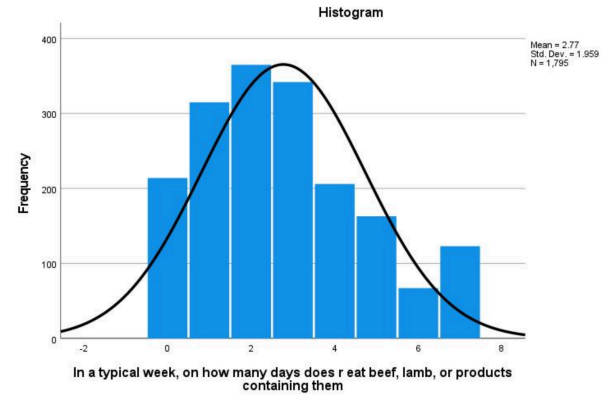


Figure 12. SPSS Output for Descriptive Statistics on EATMEAT

Because this variable is continuous, we have not produced frequency tables, and therefore we jump right into the statistics. 1795 respondents answered this question. On average, they eat beef or lamb 2.77 days per week (that is what the mean tells us). The median respondent eats beef or lamb three days per week. The standard deviation of 1.959 tells us that about 68% of respondents will be found within ±1.959 of the mean of 2.77, or between 0.811 days and 4.729 days. The skewness of 0.541 tells us that the data is mildly skewed to the right, with a longer tail at the higher end of the distribution. The kurtosis of -0.462 tells us that the data is mildly platykurtic, or has little data in the outlying tails. (Note that we have ignored several statistics in the table, which are used to compute or further interpret the figures we are discussing and which are otherwise beyond the scope of this text). The range is 7, with a minimum of 0 and a maximum of 7—sensible, given that this variable is measuring the number of days of the week that something happens. The 25th percentile is at 1, the 50th at 3 (this is the same as the median) and the 75th at 4. This tells us that one quarter of respondents eat beef or lamb one day a week or fewer; a quarter eat it between one and three days a week; a quarter eat it between three and four days a week; and a quarter eat it more than four days per week. The histogram shows the shape of the distribution; note that while the distribution is otherwise fairly normally distributed, more respondents eat beef or lamb seven days a week than eat it six days a week.

Graphs

There are several other ways to produce graphs in SPSS. The simplest is to go to Graphs → Legacy Dialogs, where a variety of specific graph types can be selected and produced, including both univariate and bivariate charts. The Legacy Dialogs menu, as shown in Fig-

ure 13, permits users to choose bar graphs, 3-D bar graphs, line graphs, area charts, pie charts, high-low plots, boxplots, error bars, population pyramids, scatterplots/dot graphs, and histograms. Users are then presented with a series of options for what data to include in their chart and how to format the chart.

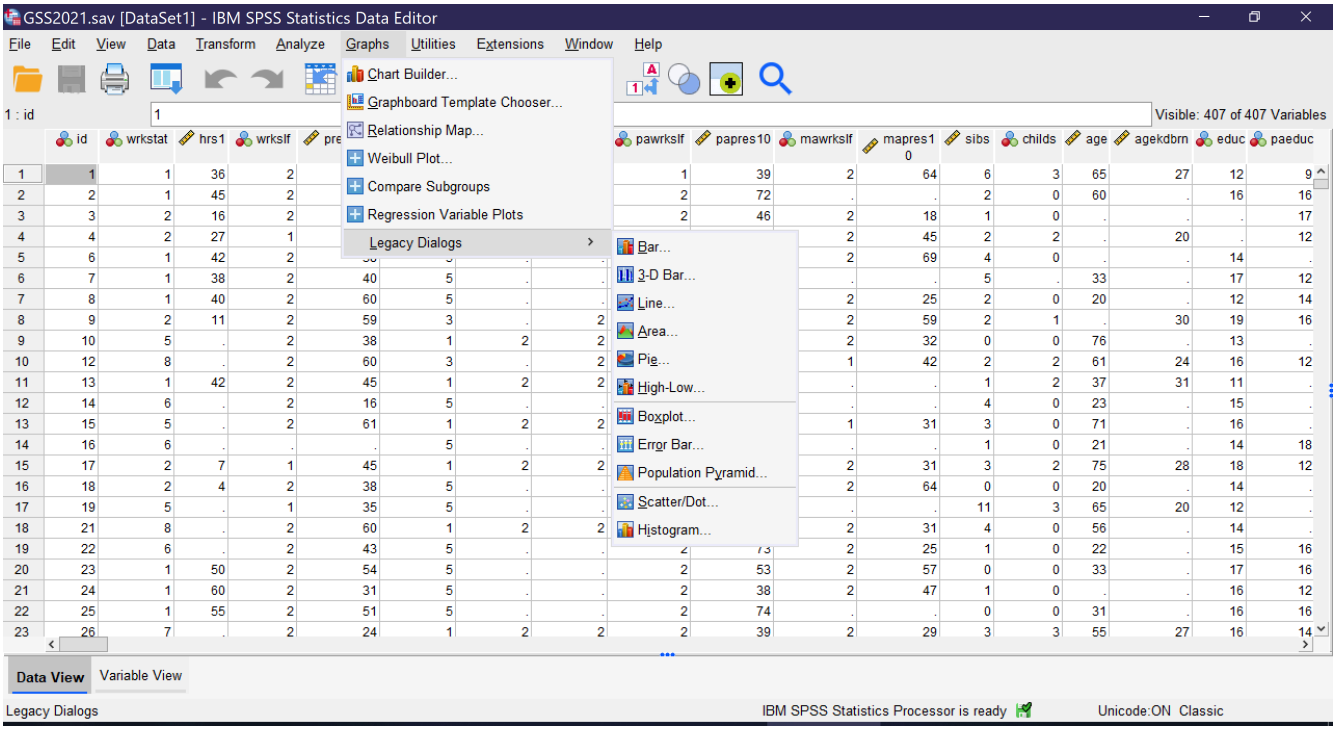


Figure 13. The Legacy Dialogs/Graphs Menu in SPSS

Here, we will review how to produce univariate bar graphs, pie charts, and histograms using the legacy dialogs. Other graphs important to the topics discussed in this text will be reviewed in other chapters.

Bar Graphs

To produce a bar graph, go to Graphs → Legacy Dialogs → Bar. For a univariate graph, then select Simple, and click Define. Then, select the relevant binary, nominal, or ordinal variable and use the blue arrow (or drag and drop it) to place it in the "Category Axis" box. You can change the options under "Bars represent" to be the number of cases, the percent of cases, or other statistics, if you choose. Once you have set up your graph, click OK, and the graph will appear in the Output Viewer window. Figure 14 shows the dialog boxes for creating a

bar graph, with the appropriate options selected, as well as a graph of the variable NEWS, which measures how often the respondent reads a newspaper.

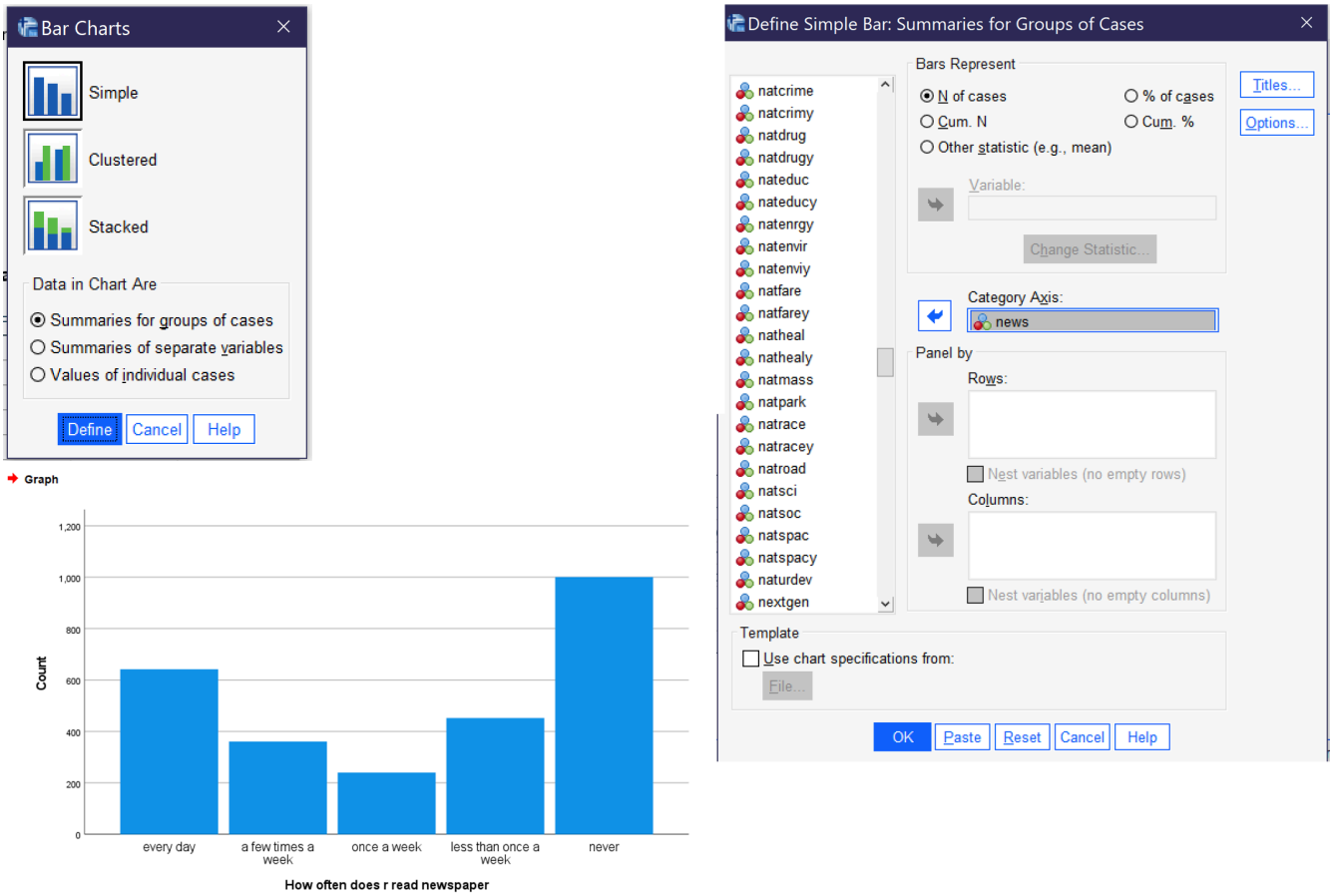


Figure 14. Bar Graph Dialog and Resulting Bar Graph for NEWS

Pie Charts

To produce a pie chart, go to Graphs → Legacy Dialogs → Pie. In most cases, users will want to select the default option, “Summaries for groups of cases,” and click define. Then, select the relevant binary, nominal, or ordinal variable (remember not to use pie charts for variables with too many categories) and use the blue arrow (or drag and drop it) to place it in the “Define Slices By” box. You can change the options under “Slices represent” to be the number of cases or the percent of cases. Once you have set up your graph, click OK, and the graph will appear in the Output Viewer window. Figure 15 shows the dialog boxes for creating a pie chart, with the appropriate options selected, as well as a graph of the variable BORN, which measures whether or not the respondent was born in the United States.

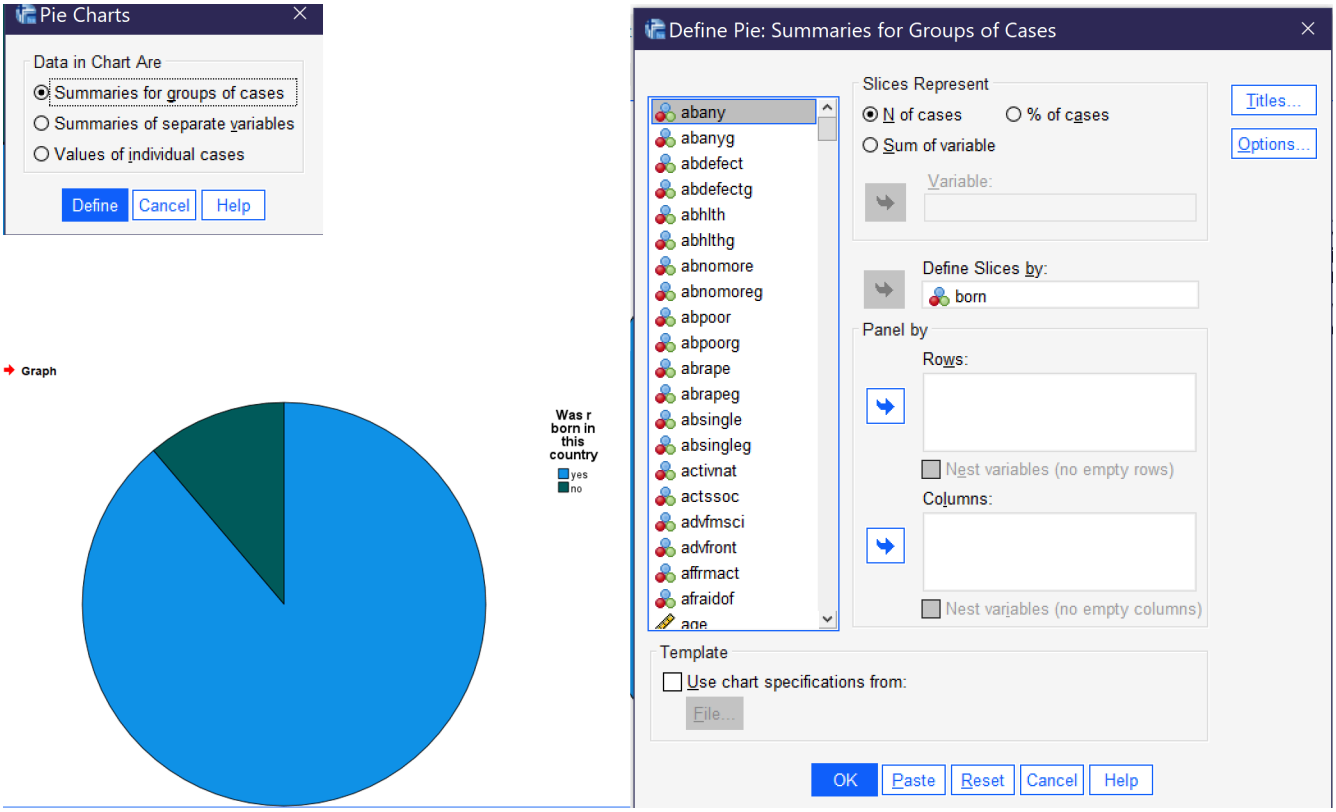


Figure 15. Pie Chart Dialog and Resulting Pie Chart for BORN

Histograms

To produce a histogram, go to Graphs → Legacy Dialogs → Histogram. Then, select the relevant continuous variable and use the blue arrow (or drag and drop it) to place it in the "Variable" box. Most users will want to check the "Display normal curve" box. Once you have set up your graph, click OK, and the graph will appear in the Output Viewer window. Figure 16 shows the dialog boxes for creating a histogram, with the appropriate options selected, as well as a graph of the variable AGE, which measures the respondent's age at the time of the survey. Note that when histograms are produced, SPSS also provides the mean, standard deviation, and total number of cases along with the graph.

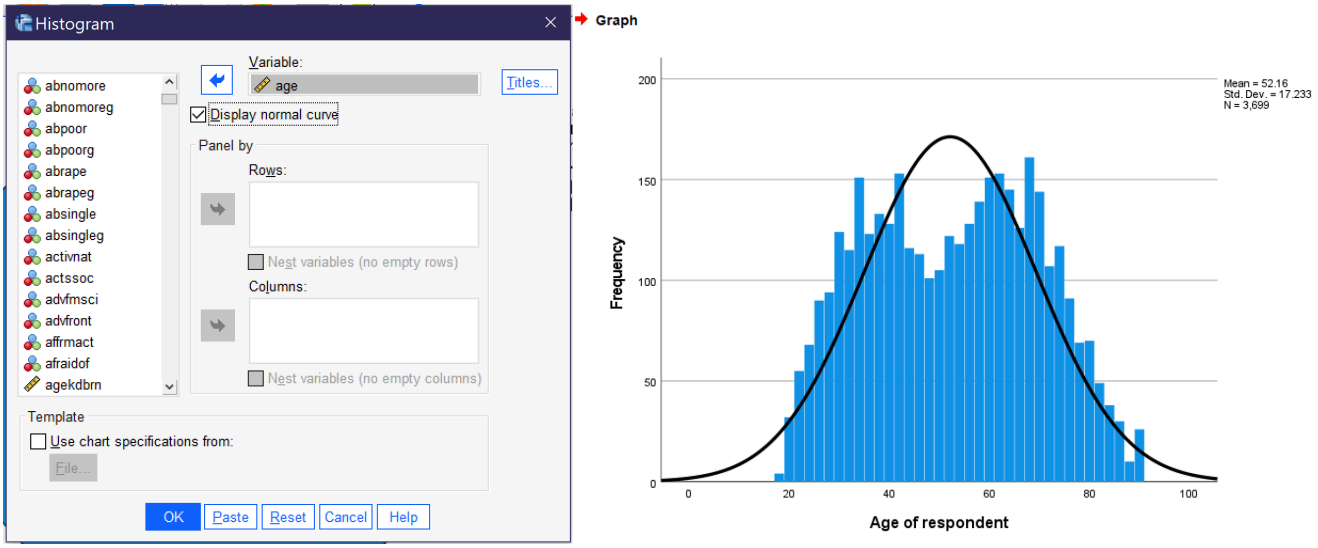


Figure 16. Histogram Dialog and Resulting Histogram for AGE

Other Ways of Producing Graphs

Other options include the Chart Builder and the Graphboard Template Chooser. In the Graphboard Template Chooser, users select one or more variables and SPSS indicates a selection of graphs that may be suitable for that combination of variables (note that SPSS simply provides options, it cannot determine if those options would in fact be appropriate for the analysis in question, so analysts must take care to evaluate the options and choose which one(s) are actually useful for a given analysis). Then, users are able to select from among a set of detailed options and provide titles for their graph. In chart builder, users first select from among a multitude of univariate and bivariate graph formats and drag and drop variables into the graph, then setting options and properties and changing colors as desired. While both of these tools provide more flexibility than the graphs accessed via Legacy Dialogs, advanced users designing visuals often move outside of the SPSS ecosystem and create graphs in software more directly suited to this purpose, such as Excel or Tableau.

Exercises

To complete these exercises, load the 2021 GSS data prepared for this text into SPSS. For each of the following variables, answer the questions below.

- ZODIAC
- COMPUSE
- SATJOB
- NUMROOMS
- Any other variable of your choice

---

1. What is the variable measuring? Use the GSS codebook to be sure you understand.
2. At what level of measurement is the variable?
3. What measures of central tendency, measures of dispersion, and graphs can you produce for this variable, given its level of measurement?
4. Produce each of the measures and graphs you have listed and copy and paste the output into a document.
5. Write a paragraph explaining the results of the descriptive statistics you've obtained. The goal is to put into words what you now know about the variable—interpreting what each statistic means, not just restating the statistic.

Media Attributions

- descriptives frequencies © IBM SPSS is licensed under a All Rights Reserved license
- frequencies window © IBM SPSS is licensed under a All Rights Reserved license
- frequencies-statistics © IBM SPSS is licensed under a All Rights Reserved license
- frequencies charts © IBM SPSS is licensed under a All Rights Reserved license
- binary descriptives © IBM SPSS is licensed under a All Rights Reserved license
- gunlaws output © IBM SPSS is licensed under a All Rights Reserved license
- nominal descriptives © IBM SPSS is licensed under a All Rights Reserved license
- mobile16 output © IBM SPSS is licensed under a All Rights Reserved license
- ordinal descriptives © IBM SPSS is licensed under a All Rights Reserved license
- carsgen output © IBM SPSS is licensed under a All Rights Reserved license
- continuous descriptives © IBM SPSS is licensed under a All Rights Reserved license
- eatmeat output © IBM SPSS is licensed under a All Rights Reserved license
- graphs legacy dialogs © IBM SPSS is licensed under a All Rights Reserved license
- bar graphs © IBM SPSS is licensed under a All Rights Reserved license
- pie charts © IBM SPSS is licensed under a All Rights Reserved license
- histogram © IBM SPSS is licensed under a All Rights Reserved license

17. Quantitative Analysis with SPSS: Data Management

MIKAILA MARIEL LEMONIK ARTHUR

This chapter is designed to introduce a variety of ways to work with datasets and variables that facilitate analysis. None of the approaches in this chapter themselves produce results, but rather are designed to enable analysis that might not be possible if datasets are used in their default form. First, it will show how to perform analysis on more limited subsets of data. Then, it will show how to transform variables to change their level of measurement, reduce attributes, create index variables, and otherwise combine variables. One quick note about a topic that is not covered in this text: the application of survey weights. More advanced quantitative analysts will want to learn to properly weight their data before performing analysis.

Working With Datasets

In some cases, analysts may wish to use a smaller subset of their dataset or to analyze different groups within the dataset separately. This section of the chapter will review select cases and split file, approaches for doing just this.

Select Cases

The Select Cases tool permits analysts to choose a subset of cases upon which to perform analysis. It can be found at the bottom of the Data menu (Alt+D, Alt+S); the Select Cases dialog is shown in Figure 1. Select cases offers the option of selecting cases based on satisfying a certain condition (e.g. cases with a specific value for a specific variable), selecting a random sample of a percentage or number of cases, selecting a specific range of cases, or using a filter variable to select only those cases with a value other than 0 or missing on that variable.

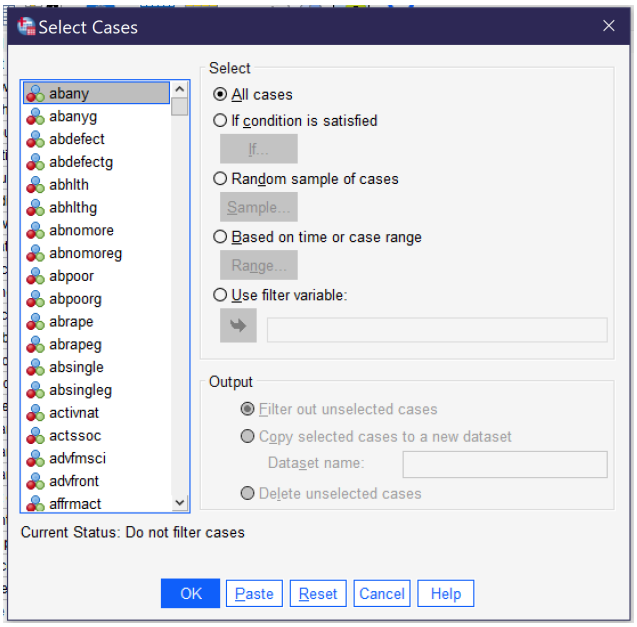


Figure 1. The Select Cases Dialog

When using the "If condition is satisfied" option, click the "If..." button and use variable names and logical or mathematical operators to write an expression (either by clicking or just typing the expression). For instance, one might select only those who have a bachelor's degree or higher by writing $degree = 3 \mid degree = 4$,[1] as shown in Figure 2.

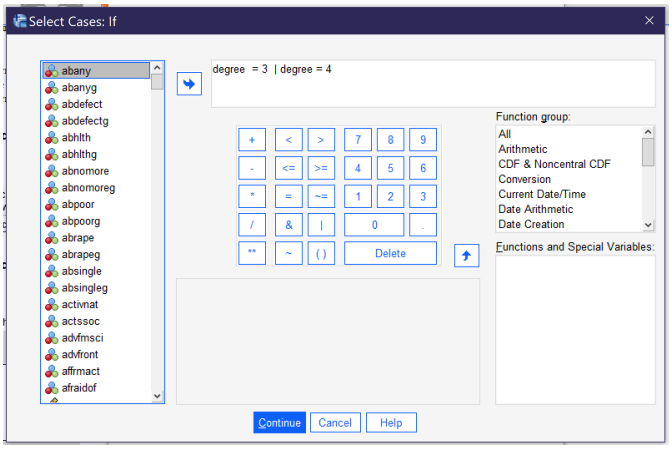


Figure 2. The "Select Cases If" Dialog

Once an option has been selected, analysts then need to determine what should happen to the selected cases. They can choose to filter out the unselected cases or to copy the selected cases into a new file with a given filename. SPSS also permits the option of deleting unselected cases, but since this permanently alters the original dataset, it is not recommended. If "Filter Out Unselected Cases" is chosen, it is important to remember to return to the Select Cases dialog when the portion of the project relying on the selected subset of cases is completed. When returning to Select Cases, "All Cases" should be selected in order to revert to the original dataset with all cases available for analysis.

1. Note: the symbol | means *or* in mathematical notation.

Split File

The split file tool allows analysts to produce output that is separated according to the attributes of a variable. For instance, analysts could perform descriptive statistics or crosstabulations and generate separate output for different race, sex, educational, or other categories. Split file can be accessed via the Data menu (Alt+D, Alt+F); the dialog is shown in Figure 3. Analysts can choose to analyze all cases (not splitting the file) or to split the file and either compare groups or organize output by groups. Using each of these options, all analyses performed appear in the output in multiple copies, one for each of the attributes of the selected variable. So, for instance, if the file were split by SEX (in the 2021 GSS, SEX only has the attributes of male and female), separate descriptive statistics, crosstabs, graphs, or whatever other output is desired will be produced. The difference between "Compare groups" and "Organize output by groups" is that "Compare groups" produces a stack of output—say, the frequencies tables for male and female—right on top of each other, while "Organize output by groups" produces all output requested in a single procedure separated for each attribute.

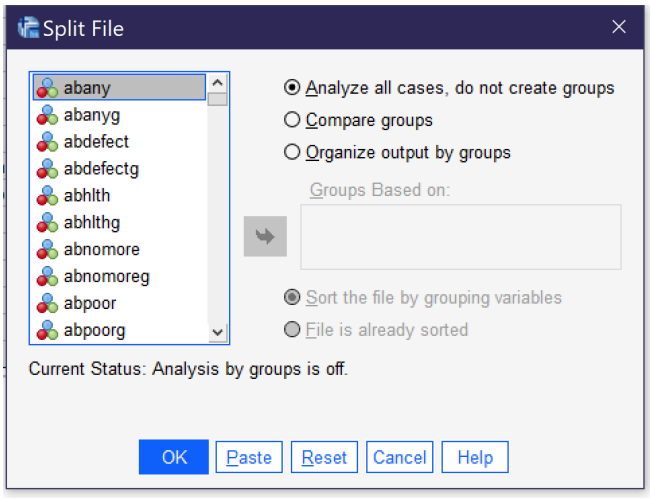


Figure 3. The Split File Dialog

Once the analyst has selected one of these options, they select the variable and use the blue arrow to put it in the Groups Based on box; in most cases, the option for Sort the file by grouping variables should be selected as well. Then click on and proceed to perform the desired analysis. Once the analysis is completed, return to the Compare Groups dialog and select "Analyze all cases, do not compare groups" and click OK so that the split file is turned off.

Working With Variables

Analysts may wish to use variables differently than the way they were originally collected. This section of the chapter will address three approaches for transforming variables: first, recoding variables, which permits transforming a continuous variable into an ordinal one or reducing the number of attributes in a variable to fewer, larger categories; second, creating indexes by combining variables using the count function; and third, using compute to manipulate or combine variables in other ways, such as creating averages.

Recoding

Recoding is a procedure that permits analysts to change the way in which the attributes of a variable are set up. It can be used for a variety of purposes, among them:

- Converting a continuous variable into an ordinal one by grouping numerical values into categories,
- Simplifying a nominal or ordinal variable with many categories by collapsing those categories into a smaller number of categories,
- Changing the direction of a variable, for instance taking an ordinal variable with 5 categories ranging from 1: strongly disagree to 5: strongly agree and turning it into an ordinal variable with 5 categories ranging from 1: strongly agree to 5: strongly disagree, and
- Creating **dummy variables**, as will be discussed in the chapter on multivariate regression.

This section of the chapter will provide examples of how to conduct the first two types of recoding. Note that before proceeding to recode any variable, it is essential to first produce complete descriptive statistics for the variable in question and study them carefully. If you are recoding a continuous variable, you may wish to use the "cut points" option under Analyze → Descriptive Statistics → Frequencies → Statistics, being sure to specify the number of equal groups you are considering creating. If you are recoding a discrete variable it is also essential to understand what the attributes (value labels) are for that variable and how they are coded (values). Take good notes on both the descriptive statistics and the attributes so that you have the information available to help you decide how to set up your recode.

The recode dialog is found under Transform (Alt+T). Note that there are several different recoding options. You should **never** use Recode into Same Variables (Alt+S), as this writes over your original data. The Automatic Recode (Alt+A) option is most useful when data is in non-numeric form and needs to be converted to numeric form. Most frequently, as a quantitative analyst, you will use Recode into Different Variables (Alt+R).

Recoding a Continuous Variable into an Ordinal Variable

Let's say we would like to recode AGE, changing it from a continous variable to a discrete variable. We might use our own understanding of ages to come up with categories, like 18-25, 26-35, 36-45, 46-55, 56-65, 66-75, and 75 and older. But these categories, it turns out, might not be so useful, as not very many people in our dataset are at the youngest or oldest ends of the distribution—something we find out if we look at the descriptive statistics. In fact, if we produce descriptive statistics using cut points for five equal groups—as shown in Table 1, we would find out that we have to get all the way to age 35 to have 20% of the people in our sample fall into one age category. We might not want to just use the cut points our descriptive statistics found, though, as they do not necessarily make sense as theoretical groupings. Perhaps instead we would choose 18-35, 36-45, 46-59, 60-69, and 70 and older. These grouping would be approximately equal in size, but make more sense numerically. Once we determine our groups, we also need to decide which numerical value we will assign each group—perhaps 1:18-35, 2:36-45; 3:46-59, 4:60-69; 5:70+. Now that we have decided how we will recode our variable, we are ready to proceed with actually recoding the variable.

Table 1. Descriptive Statistics for AGE, 2021 GSS

Age of respondent		
N	**Valid**	3699
	Missing	333
Mean		52.16
Median		53.00
Std. Deviation		17.233
Variance		296.988
Skewness		.018
Std. Error of Skewness		.040
Kurtosis		-1.018
Std. Error of Kurtosis		.080
Range		71
Minimum		18
Maximum		89
Percentiles	**20**	35.00
	40	46.00
	60	59.00
	80	69.00

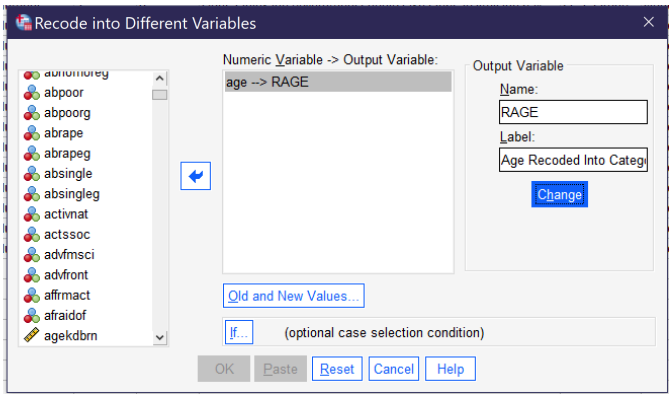


Figure 4. Recode Into Different Dialog Box Set Up to Recode Age

To begin the process of recoding, go to Transform → Recode Into Different. Select the variable you wish to recode and move it into the box using the blue arrow. Then, give the variable a new name and label. Many analysts use the convention of adding an R to the original variable name, thus here we are giving our variable the new name RAGE and the label "Age Recoded Into Categories." Click the Change button. There is an If... option for more complicated recoding procedures, but in most cases all that needs to be done now is clicking Old and New Values to put in the old and new values we already decided upon.

In the Old and New Values dialog, there are a variety of ways to indicate the original value (old) and the new value. We always begin by selecting old value: System or user-missing and new value: System-missing, to ensure that missing values remain missing. We then put in the rest of our categories using the Range ____ through _____ option, except for the final category, where we use Range, value through highest to ensure we don't accidentally leave out those 89-year-olds. Once an old value and its respective new value have been entered, we click the Add button so that they appear in the Old → New box. In other cases, analysts might change individual values, use Range, lowest through value, or combine all other values. If it is necessary to edit or delete something that has already been added, use the Change (to edit) or Remove (to delete) buttons. When all of the old and new values have been added, the Old and New Values dialog should look as it does in Figure 5, with the following test in the Old → New box:

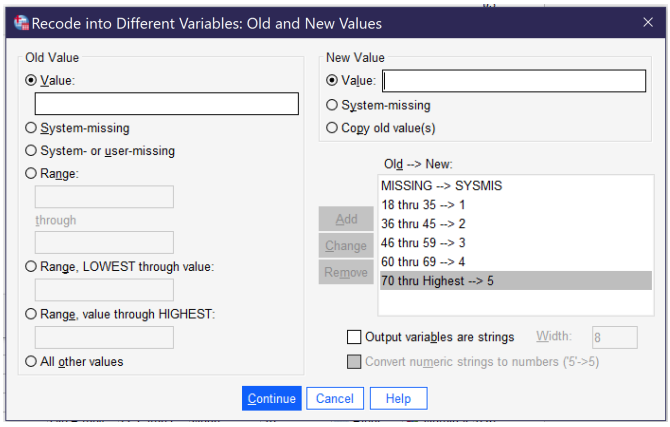


Figure 5. Old and New Values Dialog for Recoding Age

```
MISSING-->SYSMIS
 18 thru 35-->1
 36 thru 45-->2
 46 thru 59-->3
 60 thru 69-->4
70 thru highest-->5
```

When everything is set up, click Continue and then OK. To see your new variable, scroll to the bottom of the screen in Variable View.

There is one more step to recoding, and that is to add the value labels. To do this, go to Variable View; you will most likely find your new variable at the very bottom of the list of variables. If you click in the Values box for the row with your new variable in it, as shown in Figure 6, you will see a box with … in it. Click the … and the value labels dialog will come up.

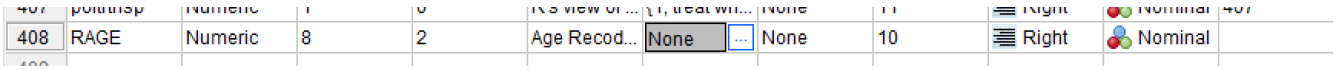


Figure 6. Preparing to Enter Value Labels

To enter the value labels, click on the green plus sign, and then enter a numerical value and its associated value label. Click the plus sign again to enter the next value and value label, and so on until all have been entered. If you need to remove one that has been entered incorrectly, use the red X. There is a spellchecker if useful. When you are done, the Value Labels dialog should look as it does in Figure 7. Then click OK, and remember to save your dataset so that you don't lose your new variable.

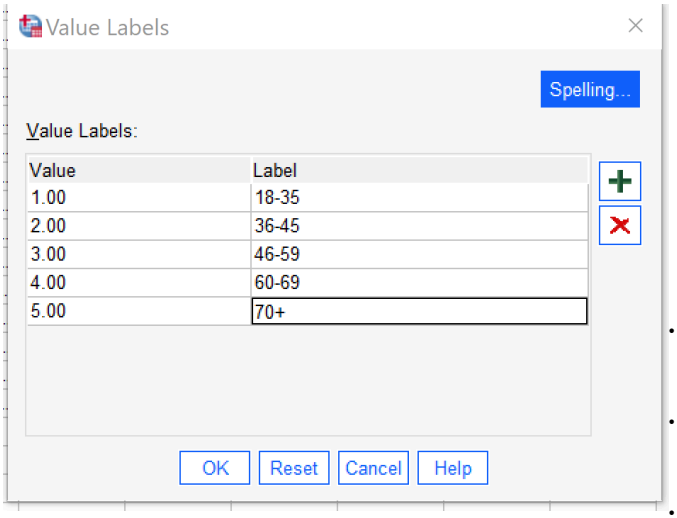


Figure 7. Value Labels for Recoded Age Variable

Finally, it is time to check that the recoding is proceeding correctly before using the new variable in any analysis. To check the variable, produce a frequency distribution. Assess the frequency distribution for evidence of any errors, such as:

- Old values that did not get caught in the recoding process,
- Categories that are missing from the new values,
- More missing values than would be expected, and
- Unexpected discrepancies between the descriptive statistics from the original variable and those produced now.

In the case of our RAGE variable, we can observe in Table 2 that we did a relatively good job of keeping the proportion of respondents in each of our categories pretty even. In the absence of theoretical or conceptual reasons for choosing a particular recoding strategy, making categories of relatively consistent size can be a good way to proceed.

Table 2. Frequency Distribution for Age Recoded Into Categories

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	**18-35**	795	19.7	21.5	21.5
	36-45	643	15.9	17.4	38.9
	46-59	855	21.2	23.1	62.0
	60-69	728	18.1	19.7	81.7
	70+	678	16.8	18.3	100.0
	Total	3699	91.7	100.0	
Missing	**System**	333	8.3		
Total		4032	100.0		

Reducing the Attributes of a Discrete Variable

As noted above, recoding can also be used to condense categories in the case of an ordinal or nominal variable with many categories. The example here uses the variable POLVIEWS, an ordinal variable measuring respondents' political views on a seven-point scale from extremely liberal to extremely conservative. In recoding this variable, we might want to reduce the seven points to three: liberal, moderate, and conservative. But which values belong in which categories? We might say that extremely liberal and liberal make up the liberal category; extremely conservative and conservative make up the conservative category; and slightly liberal, slightly conservative, and moderate/middle of the road make up the moderate category. So we produce a frequency table to see what our data looks like. This frequency table is shown in Table 3.

Table 3: Think of Self as Liberal or conservative

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	extremely liberal	207	5.1	5.2	5.2
	liberal	623	15.5	15.7	20.9
	slightly liberal	490	12.2	12.4	33.3
	moderate, middle of the road	1377	34.2	34.7	68.0
	slightly conservative	476	11.8	12.0	80.0
	conservative	617	15.3	15.6	95.6
	extremely conservative	174	4.3	4.4	100.0
	Total	3964	98.3	100.0	
Missing	System	68	1.7		
Total		4032	100.0		

Table 3 might cause us to think that our original idea about how to combine these categories is not the best, given how few people would end up in the liberal and conservative categories and how many in the moderate category. Instead, we might conclude that it would make more sense to group extremely liberal, liberal, and slightly liberal together; keep moderate on its own; and then group extremely conservative, conservative, and slightly conservative together. And we might decide that 1 will be liberal, 2 will be moderate, and 3 will be conservative. We also need to write down the value labels from our existing variable so that we can use them in the recoding.

Once we've made these decisions, it's time to proceed with the recoding, which we do much the same way as we did for the recoding of the continuous variable above, by going to Transform → Recode Into Different. Give the variable a new name, RPOLVIEWS, and a new label, Is R Liberal, Moderate, or Conservative? Click Change, and then click Old and New Values. Note that if you are recoding right after recoding a prior variable, you will need to use the Remove button to remove the old and new values from the prior recoding process. The old and new values we will be entering are shown below; Figure 8 shows what the recode dialogs should look like.

```
MISSING-->SYSMIS
1 thru 3-->1
4-->2
5 thru 7-->3
```

Once the old and new values are entered, click Continue and then OK. Then, scroll to the

new variable in Variable View and add the value labels: 1 Liberal, 2 Moderate, 3 Conservative, as shown in Figure 8.

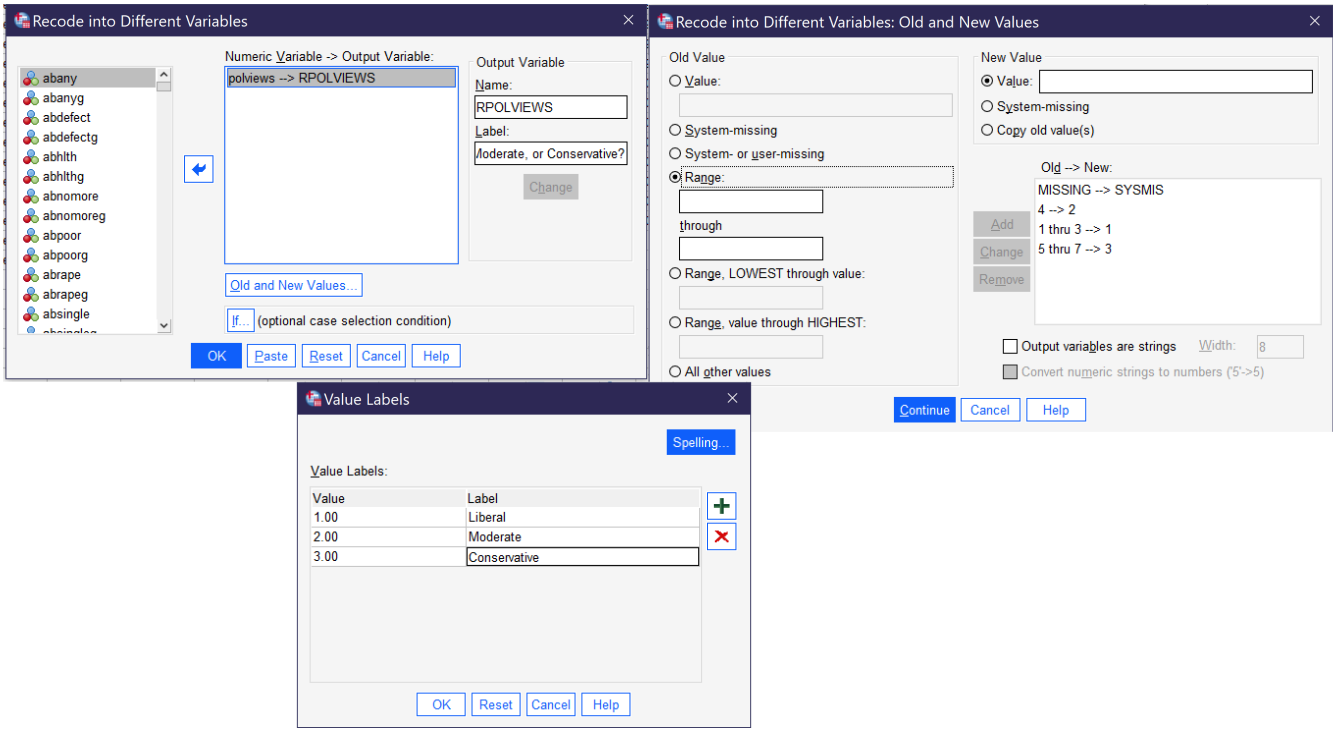


Figure 8. Recoding POLVIEWS

Finally, produce a frequency table to check for errors before using the new RPOLVIEWS variable in an analysis. As Table 4 shows, the recoding strategy we chose happened to distribute respondents quite evenly, though of course it is based on conceptual concerns rather than simply the distribution of respondents.

Table 4. Is R Liberal, Moderate, or Conservative?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Liberal	1320	32.7	33.3	33.3
	Moderate	1377	34.2	34.7	68.0
	Conservative	1267	31.4	32.0	100.0
	Total	3964	98.3	100.0	
Missing	System	68	1.7		
Total		4032	100.0		

Creating an Index

In the course of many research and data analysis projects, researchers may seek to create **index variables** by combining responses on multiple related variables. While it may seem that simply adding the values together might work, one of the main reasons simply adding does not work is that it cannot distinguish between circumstances where a respondent did not answer one or more questions and circumstances in which respondents gave answers with lower numerical values. Therefore, the process to be detailed here requires two steps: first, collecting missing responses, and then, creating an index while excluding those respondents who did not answer all questions included in the index.

The example index detailed here is an index of the seven variables in the 2021 GSS that ask respondents their views on whether abortion should be legal in a variety of circumstances: in the case of fetal defect (ABDEFECT), risks to health (ABHLTH), rape (ABRAPE), if the pregnant person is single (ABSINGLE), if the pregnant person is poor (ABPOOR), if the pregnant person already has children and does not want any more (ABNOMORE), and for any reason (ABANY). Note that in the 2021 GSS there are a separate set of variables asking about abortion that end in G. These variables reflect differences in wording as part of a survey experiment; one could use them instead of the non-G abortion opinion variables, but the two sets should not be combined as they were asked of different people.

Our first task is to look at the value labels for our variables. Each of these abortion variables is coded with 1:Yes and 2:No; we need to determine whether we wish to make an index of yeses or nos, and in this case we will use the Yeses. The second task is to collect the missing responses so that we can exclude them. Both this part of the process and the ultimate task of creating the index utilize Transform → Count Values Within Cases... (Alt+T, Alt+O). Once the Count dialog is open, we need to give our new variable a name (in the Target Variable box) and Label (in the Target Label box). We will call this variable ABMISSING with the label Count of Missing Variables for Abortion Variables, given that's what we are counting at the moment. We then move all of the variables (seven, in this case) we are including in our index into the Variables box using the blue arrow. Next, click on the Define Values button. Select the radio button next to System- or user-missing and click Add, then click Continue, then click OK. Figure 9 shows how this Count procedure should be set up.

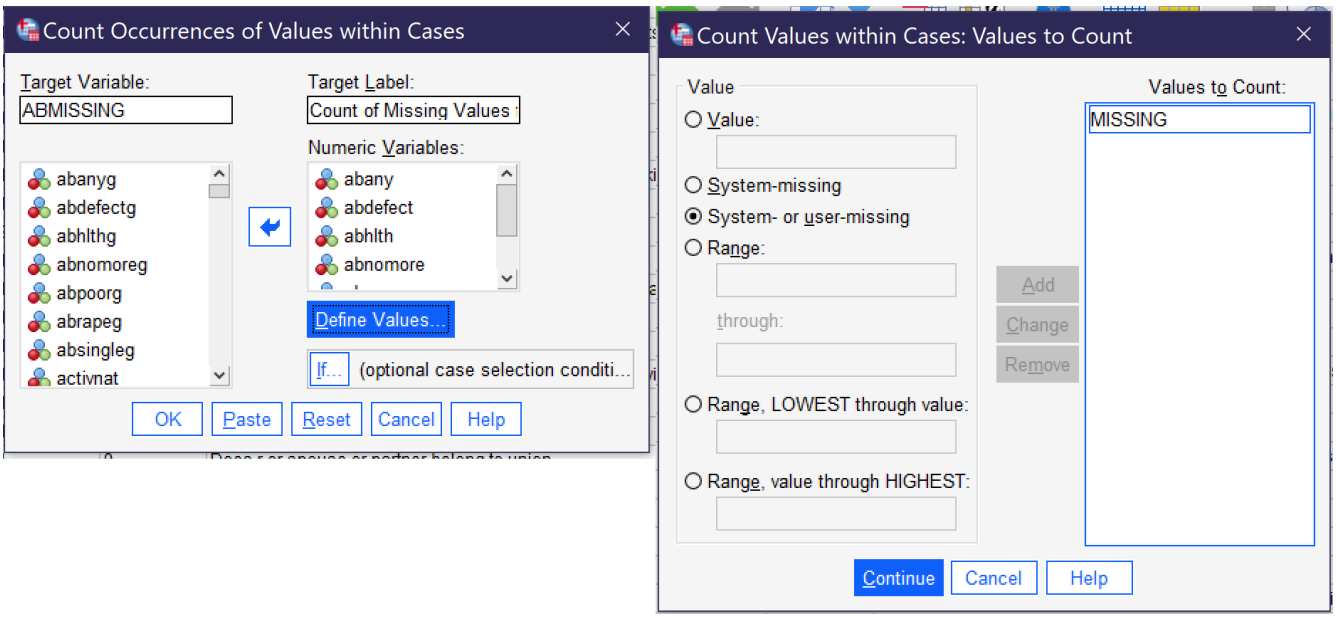


Figure 9. Setting Up the Count for Missing Values

Table 5. Count of Missing Values for Abortion Variables

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	1284	31.8	31.8	31.8
	1.00	118	2.9	2.9	34.8
	2.00	26	.6	.6	35.4
	3.00	8	.2	.2	35.6
	4.00	9	.2	.2	35.8
	5.00	1	.0	.0	35.9
	6.00	11	.3	.3	36.1
	7.00	2575	63.9	63.9	100.0
	Total	4032	100.0	100.0	

It is a good practice to produce a frequency table at this stage to check for errors and ensure that a reasonable number of cases remain for use in producing the desired index variable. In this example, the frequency table for the missing values count should appear as in Table 5. This shows that 31.8%, or 1284 cases, answered all seven abortion questions and thus will be able to be included in our index variable. 63.9% did not answer any of the seven questions (presumably because they were not asked them), while a far smaller percent answered only some of the questions—and thus also will be excluded from our index.

The next step is to create the index variable while excluding those who have missing values. To do this, we again go to Transform → Count Values Within Cases.... This time, we will

call our new variable ABINDEX and give it the label Index Variable for Abortion Questions. Under Define Values, we remove MISSING and, in its place, add 1 to the Values to Count box, and then click Continue. Next, we click If... and select the radio button next to Include if case satisfies condition. In the box, we say ABINDEX = 0 (so that cases in which *any* of the included variables have missing values are excluded) and click Continue. Figure 10 below shows what all of the dialogs should look like when everything is set up to produce the index. Once everything is ready, click OK.

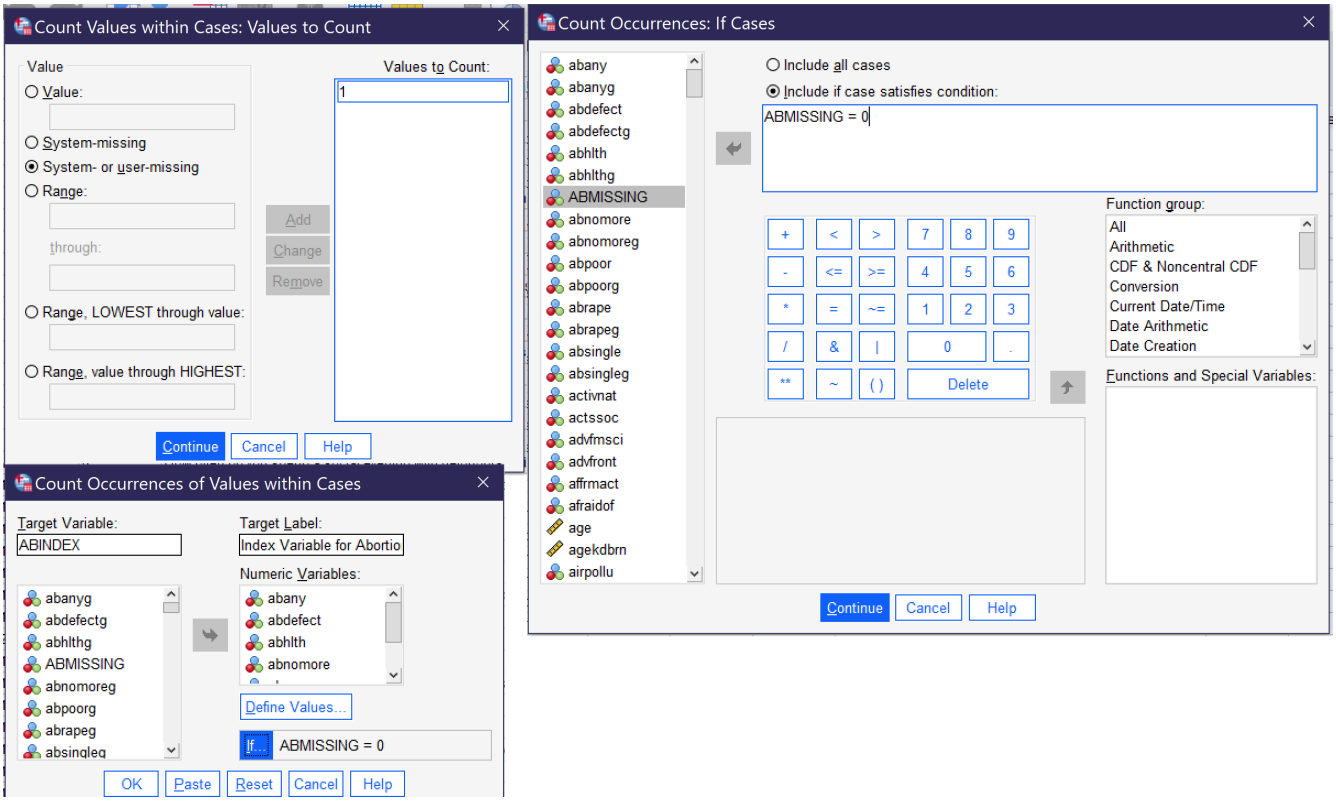


Figure 10. Creating an Index

Finally, produce a frequency distribution for the new index variable. (Note: some analysts treat this type of index variable as ordinal, while others argue that because it is a count of numbers it might better be understood as continuous. Either approach is acceptable for producing descriptive statistics.) Table 6 shows the results, which many people might find surprising: 50.4% of respondents—just a tad more than half—agree that abortion should be legal in *all* of the cases about which they were asked, while only 7.2% believe abortion should be legal in none of them.

Table 6. Frequencies for Index Variable for Abortion Questions

		Frequency	Percent	Valid Percent	Cumulative Percent
	.00	93	2.3	7.2	7.2
	1.00	82	2.0	6.4	13.6
	2.00	104	2.6	8.1	21.7
	3.00	172	4.3	13.4	35.1
Valid	**4.00**	69	1.7	5.4	40.5
	5.00	43	1.1	3.3	43.8
	6.00	74	1.8	5.8	49.6
	7.00	647	16.0	50.4	100.0
	Total	1284	31.8	100.0	
Missing	**System**	2748	68.2		
Total		4032	100.0		

Computing Variables

Sometimes, analysts want to combine variables in ways other than by making an index (for instance, adding two continuous variables together or taking their average) or otherwise wish to perform mathematical functions on them. These types of operations can be conducted by going to Transform → Compute Variable (Alt+T, Alt+C). Here, we will try two examples, one in which we take the average of the two variables measuring parental occupational prestige (MAPRES10 and PAPRES10) to determine respondents' parents' average occupational prestige, and one in which we take a continuous variable collected on a weekly basis (WWWHR, which measures how many hours respondents spend on the Internet per day) and divide it by seven, the number of days in a week, to produce a variable on a daily basis (the average number of hours respondents spend on the internet per day).

To create the computed variable for average parental occupational prestige, we go to Transform → Compute Variable. We then indicate the name for our new variable in the Target Variable box; we will call it PARPRES10 (for parental prestige). Once we enter this, we can click on the Type & Label box to provide our variable with a label and be sure it is classified as a numeric variable. Next, under the Numeric Expression box, we set up the formula that will produce the outcome we want. Here, we are averaging two variables, so we want to add them together (remember the parentheses, for order of operations) and then divide by two, like this: $(mapres10 + papres10) / 2$. The If... (optional case selection) can be used to only include or to make sure to exclude certain kinds of cases; we do not need to use it

to exclude missing values, as the Compute function already excludes them from computation.

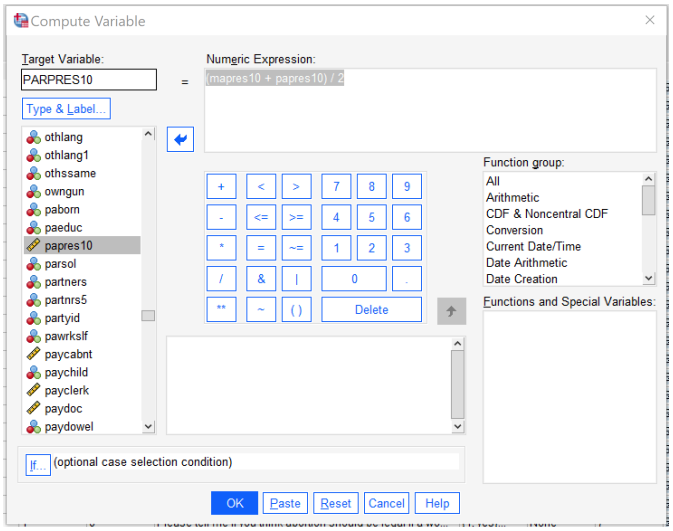


Figure 11. Computing an Average Variable

Figure 11 shows what the Compute Variable dialog should look like to produce the desired variable, the average of mother's and father's occupational prestige. When it is set up, click OK. The resulting variable is continuous, so the last step is to produce descriptive statistics for the continuous variable. Here, we will produce descriptive statistics for all three variables–the two original occupational prestige scores and our new average. Table 7 shows the results.

Table 7. Descriptive Statistics on Occupational Prestige

		Average of Mother's and Father's Occupational Prestige	Mothers occupational prestige score (2010)	Father's occupational prestige score (2010)
N	**Valid**	2232	2767	3349
	Missing	1800	1265	683
Mean		44.1633	42.66	45.16
Median		42.5000	42.00	44.00
Std. Deviation		10.58686	13.168	13.148
Variance		112.082	173.387	172.869
Skewness		.493	.389	.622
Std. Error of Skewness		.052	.047	.042
Kurtosis		-.317	-.672	-.305
Std. Error of Kurtosis		.104	.093	.085
Range		60.00	64	64
Minimum		20.00	16	16
Maximum		80.00	80	80
Percentiles	**25**	36.0000	32.00	35.00
	50	42.5000	42.00	44.00
	75	51.5000	50.00	52.00

Let's take one last example: starting with a variable that measures the number of hours respondents spend on the Internet per week and adjusting it so it measures the number of hours they spend per day. We will call the new variable WWWHRDAY, and the expression to produce it is simply $wwwhr / 7$, as shown in Figure 13. Then click OK.

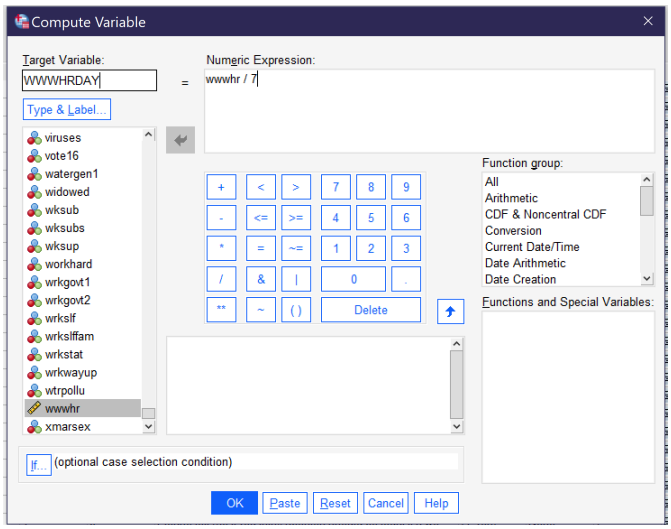


Figure 13. Another Example of Computing a Variable

Again, the final step is to compute descriptive statistics for the new variable, as shown in Table 8. These descriptive statistics show that the average person spends nearly 15 hours a week online, or just over 2 hours per day—but that this is pretty skewed by folks who spend quite a bit of time online, as the median time online is just 9 hours a week or somewhat over 1 hour per day. The range shows that there are people in the dataset who spend no time online at all, and others who claim to spend every hour of the day online (how is this possible? Don't they sleep?). Looking at the percentiles, we can observe that a quarter of our respondents spend less than half an hour a day online, while another quarter claim to spend more than 2.8 hours per day online.

Table 8. Descriptive Statistics For Time Online

		Hours per week r spends on the Internet	Average Number of Hours Online Per Day
N	**Valid**	2466	2466
	Missing	1566	1566
Mean		14.80	2.1146
Median		9.00	1.2857
Std. Deviation		17.392	2.48459
Variance		302.487	6.173
Skewness		2.575	2.575
Std. Error of Skewness		.049	.049
Kurtosis		10.257	10.257
Std. Error of Kurtosis		.099	.099
Range		168	24.00
Minimum		0	.00
Maximum		168	24.00
Percentiles	**25**	3.00	.4286
	50	9.00	1.2857
	75	20.00	2.8571

Exercises

1. Select cases to select only those who identify as poor (from the variable CLASS). Produce a histogram of working hours (from the variable HRS1). Then select all cases and produce the same histogram. Compare your results.
2. Split file by RACE or SEX. Choose any variable of interest and perform appropriate descriptive statistics. Write a paragraph explaining the differences you observe between the racial or sex categories in your analysis.
3. Recode HRS1, creating no more than 5 categories. Be sure you can explain why your categories are grouped the way they are, and look at the descriptive statistics as you determine them. Then produce descriptive statistics after recoding and summarize your results.
4. Recode ENPRBUS, creating no more than 4 categories. Be sure you can explain why your categories are grouped the way they are, and look at the descriptive statistics as you determine them. Then produce descriptive statistics after recoding and summarize your results.
5. Create an index of the six variables asking about whether people with various views should be allowed

to speak in your community (SPKATH, SPKRAC, SPKCOM, SPKMIL, SPKHOMO, SPKMSLM), being sure to create the missing value index first. Produce appropriate descriptive statistics and summarize your results.

6. Create a new variable for the average number of hours per day the respondent spends in a car or other vehicle by using the Compute function to divide CARHR (the number of hours in a vehicle per week) by 7. Produce descriptive statistics for the original CARHR variable and your new computed variable.

Media Attributions

- select cases © IBM SPSS is licensed under a All Rights Reserved license
- select cases if © IBM SPSS is licensed under a All Rights Reserved license
- split file © IBM SPSS is licensed under a All Rights Reserved license
- recode age © IBM SPSS is licensed under a All Rights Reserved license
- recode age values © IBM SPSS is licensed under a All Rights Reserved license
- to enter value labels © IBM SPSS is licensed under a All Rights Reserved license
- age value labels © IBM SPSS is licensed under a All Rights Reserved license
- recoding polviews © IBM SPSS is licensed under a All Rights Reserved license
- count of missing values © IBM SPSS is licensed under a All Rights Reserved license
- creating an index © IBM SPSS is licensed under a All Rights Reserved license
- compute average © IBM SPSS is licensed under a All Rights Reserved license
- compute math © IBM SPSS is licensed under a All Rights Reserved license

18. Quantitative Analysis with SPSS: Bivariate Crosstabs

MIKAILA MARIEL LEMONIK ARTHUR

This chapter will focus on how to produce and interpret bivariate crosstabulations in SPSS. To access the crosstabs dialog, go to Analyze → Descriptive Statistics → Crosstabs (Alt+A, Alt+E, Alt+C). Once the Crosstabs dialog opens, the independent variable should be placed in the Columns box and the dependent variable in the Rows box. There is a checkbox that will make a clustered bar chart appear, as well as one that will suppress the tables so that *only* the bar chart appears (typically one would not want to suppress the tables; whether or not to produce the clustered bar chart is a matter of personal preference). In the analysis shown in Figure 1, SEX is the independent variable (here permitting only male and female as answers) and DISRSPCT is the dependent variable (how often the respondent feels they are treated with "less courtesy or respect" than other people are).

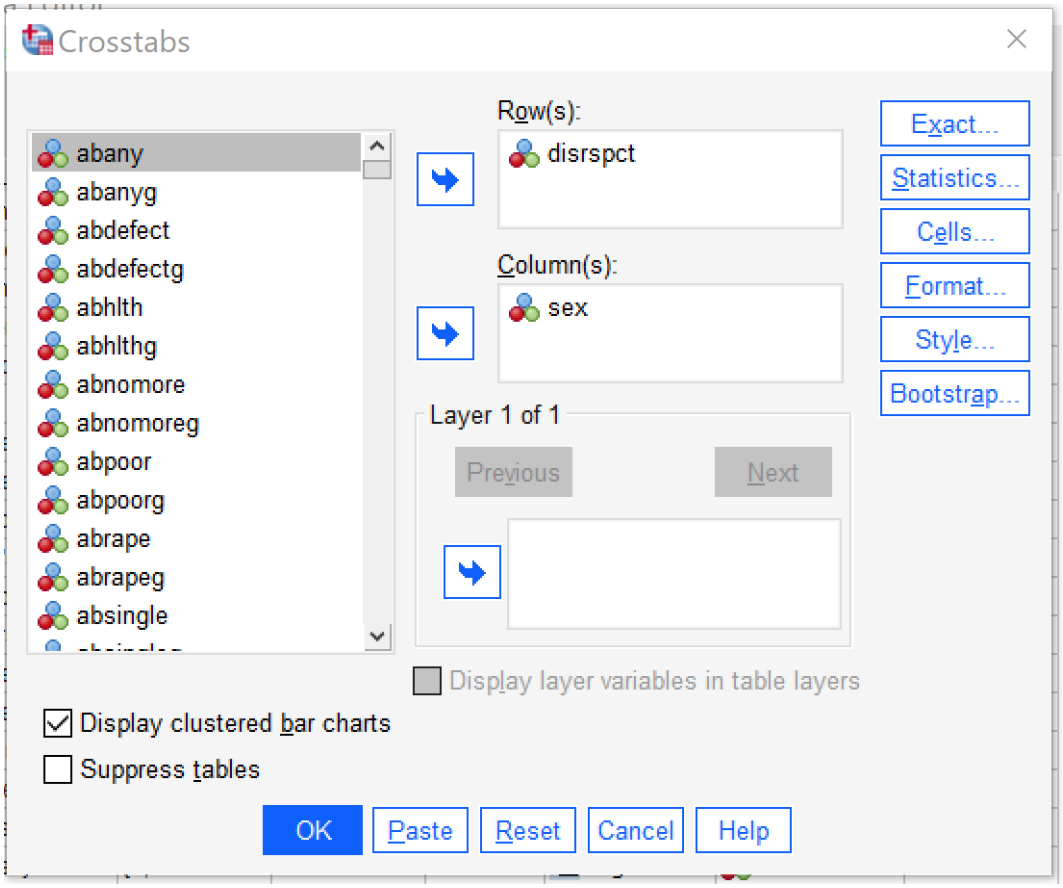


Figure 1. Setting Up a Bivariate Crosstab

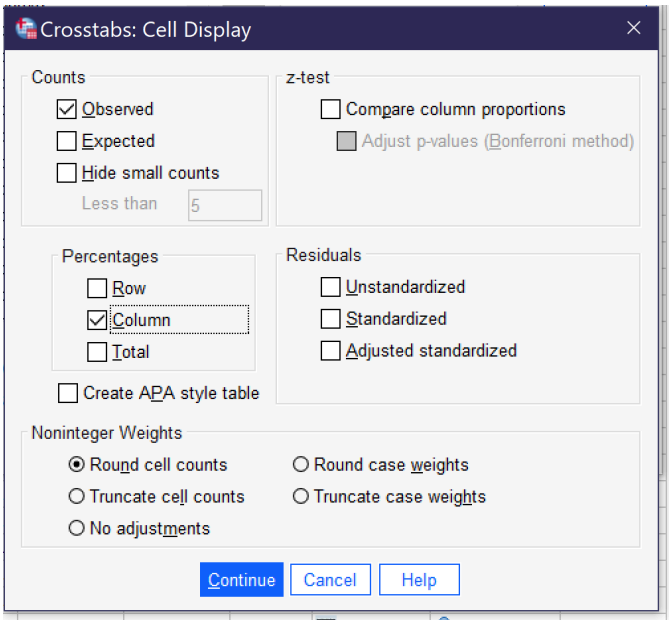


Figure 2. Cell Display Dialog for Crosstabs

Next, click the Cells button (Alt+E) and select Column under Percentages. It is important to be sure to select column percentages when the independent variable is in the columns; this is necessary for proper interpretation of the table. Observed will already be checked under Counts; if it is not, you may want to select it as well. There are a variety of other options in this dialog, but they are beyond the scope of this chapter. To see what the Cell Display dialog should look like before proceeding, take a look at Figure 2. Once the appropriate options are selected from Cell Display, click Continue. If one is interested only in producing the crosstabulation table and/or clustered bar charts, OK can be pressed after returning to the main crosstabs dialog.

However, in most cases analysts will want to obtain statistical significance and association measures as well. These can be located by clicking the Statistics button. Chi-square should be checked in order to produce statistical significance; analysts should then select the appropriate measure of association for their analysis from among those displayed in the box. In the absence of information suggesting a different measure of association, Phi and Cramer's V is a reasonable default, with Phi being used for 2×2 tables and Cramer's V for larger tables, though this may not be appropriate for Ordinal x Ordinal tables. For more information on selecting an appropriate measure of association, see the chapter on measures of association. The default options are shown in Figure 3.

Some measures of association that SPSS can compute are not listed in the dialog but instead are produced by selecting a different option: Goodman and Kruskal tau can be found under Lambda, while both Pearson's r and Spearman Correlation are found under correlations.

Note that not all of the statistics SPSS can produce are frequently used by beginning social science data analysts, and thus some are not addressed in the chapter on measures of association. And remember to select only one or two appropriate options—it is never the right answer to produce all, or many, of the statistics that are available, especially not if the analyst is simply searching for the strongest possible association.

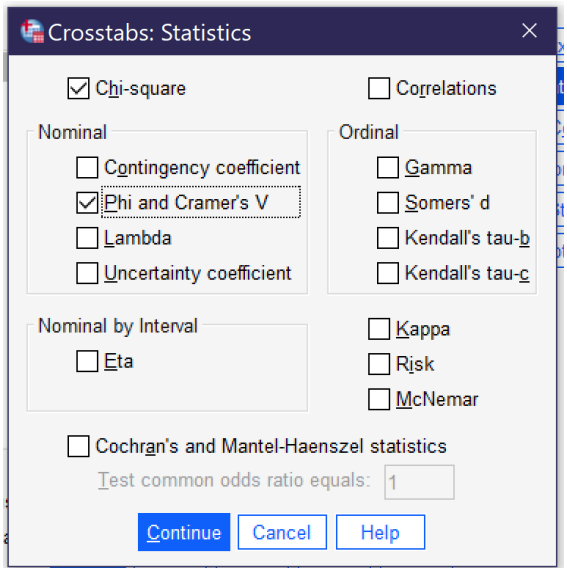


Figure 3. Statistics Dialog for Crosstabs

Once the appropriate statistics are selected, click continue to go back to the main Crosstabs dialogue, and then OK to proceed with producing the results (which will then appear in the output).

The output for this analysis is shown in Figure 4. Below Figure 4, the text will review how to interpret this output.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
How often r is treated with less courtesy or respect than others * Respondent's sex	2588	64.2%	1444	35.8%	4032	100.0%

How often r is treated with less courtesy or respect than others * Respondent's sex Crosstabulation

			Respondent's sex		Total
			male	female	
How often r is treated with less courtesy or respect than others	almost every day	Count	59	76	135
		% within Respondent's sex	5.1%	5.3%	5.2%
	at least once a week	Count	91	139	230
		% within Respondent's sex	7.9%	9.7%	8.9%
	a few times a month	Count	145	181	326
		% within Respondent's sex	12.6%	12.6%	12.6%
	a few times a year	Count	322	476	798
		% within Respondent's sex	27.9%	33.2%	30.8%
	less than once a year	Count	269	282	551
		% within Respondent's sex	23.3%	19.7%	21.3%
	never	Count	268	280	548
		% within Respondent's sex	23.2%	19.5%	21.2%
Total		Count	1154	1434	2588
		% within Respondent's sex	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	16.320[a]	5	.006
Likelihood Ratio	16.344	5	.006
Linear-by-Linear Association	7.532	1	.006
N of Valid Cases	2588		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.20.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.079	.006
	Cramer's V	.079	.006
N of Valid Cases		2588	

Figure 4. SPSS Output for a Crosstabulation of SEX and DISRSPCT

The first table shown is the Case Processing Summary, which simply shows the proportion of valid cases included in the analysis versus those missing from the analysis (which are those where there is no response to at least one of the variables).

The second table is the main crosstabulation table. To read this table, compare the percentages across the rows. So, for instance, we can see that very similar proportions of males and females feel disrespected almost every day or a few times a month, though females are somewhat more likely to feel disrespected at least once a week. Females are also more likely to feel disrespected a few times a year, while males are more likely to feel disrespected less than once a year or never. These conclusions are made simply by comparing percentage across the rows and noting which are bigger and which are smaller. Ignore the count (the raw number) as this is heavily impacted by the total number of people in each category of the independent variable and thus is not analytically useful. For example, in this analysis, there are 1434 women and 1154 men (see the total row in the crosstabulation table) and thus there are more women than men in *every* category of the dependent variable—even those where men are more likely to have selected that answer choice than women! Thus, it is necessary to focus on the percentages, not the raw numbers.

The third table presents the results of the Chi-square significance test. A variety of figures are provided in this table, including the value and degrees of freedom used to compute the Chi square. However, in most cases you need only pay attention to the figure under Asymptotic Significance (2-Sided). You will note there are several rows in that column, all of which provide the same figure. It will almost always be the case that the same figure appears in each row under the significance column; if it does not, attend to the first significance figure. In this case, the significance figure presented is 0.006, well under both the $p<0.05$ and $p<0.01$ confidence levels though above the $p<0.001$ level.

The fourth table presents the measures of association, in this case Phi and Cramer's V. Sometimes as in this case, these figures are the same, while in other cases they are different. If they are different, be sure you know which one you should be looking at given the level of measurement of your variables. Here, they are 0.079, which the strength chart in the measures of association chapter would tell us means there is a weak association.

Finally, at the bottom, is a clustered bar chart. Bivariate bar graphs can also be produced using the Graphs menu. Under Graphs → Legacy Dialogs → Bar Charts, both clustered and stacked bar charts are available. Those interested in displaying their data in bivariate graphs may wish to play around with the different options to see which presents the data in the most useful form.

Exercises

Select two variables of interest. Answer the following questions:

- Which is the independent variable and which is the dependent variable?
- What is the research hypothesis for this analysis?
- What is the null hypothesis for this analysis?
- What confidence level (p value) have you chosen?
- Which measure of association is most appropriate for this relationship?

Next, use SPSS to produce a crosstabulation according to the instructions in this chapter. Interpret the crosstabulation, being sure to answer the following questions:

- Is the relationship statistically significant?
- Can the null hypothesis be rejected?
- How strong is the association between the two variables?
- Looking at that pattern of percentages across the rows, what can you determine about the nature of the relationship between the two variables?
- Is there support for your research hypothesis?

Repeat this exercise for two additional pairs of variables, choosing new variables each time.

Media Attributions

- crosstabs dialog bivariate © IBM SPSS is licensed under a All Rights Reserved license
- cell display crosstabs © IBM SPSS is licensed under a All Rights Reserved license
- statistics dialog crosstabs © IBM SPSS is licensed under a All Rights Reserved license
- crosstabs output © IBM SPSS is licensed under a All Rights Reserved license

19. Quantitative Analysis with SPSS: Multivariate Crosstabs

MIKAILA MARIEL LEMONIK ARTHUR

Producing a multivariate crosstabulation is exactly the same as producing a bivariate crosstabulation, except that an additional variable is added. Note that, due to the limitations of the crosstabulation approach, you are not actually looking at the relationships between all three variables simultaneously (and this approach is limited to three variables). Rather, you are looking at how controlling for a third variable—your "Layer" or control variable—changes the relationship between the independent and dependent variable in your analysis. What SPSS produces, then, is basically a stack of crosstabulation tables with your independent and dependent variables, one for each category of your control variable, along with statistical significance and association values for each category of your control variable. This chapter will review how to produce and interpret a multivariate crosstabulation. It uses variables with fairly few categories for ease of interpretation. Do note that when using variables with many categories, results can become quite complex and lengthy, and due to small numbers of cases left in each cell of the very lengthy tables, statistical significance is likely to be reduced. Thus, analysts should take care to consider whether the relationship(s) they are interested in are suitable for this type of analysis, and may want to consider recoding variables (see the chapter on data management) with many categories into somewhat fewer categories to facilitate analysis.

To produce a multivariate crosstabulation, follow the same steps as you

would follow to produce a bivariate crosstabulation—put the independent variable in the columns box, the dependent variable in the rows box, select column percentages under cells, and select chi square and an appropriate measure of association under statistics. Note that the measure of association you choose should be the same one that you would choose for a bivariate analysis with the same independent and dependent variables, as the third variable is a control variable and does not alter the criteria upon which the decision about measures of association is made. The one thing you need to add in order to produce a multivariate crosstabulation is that you add your third variable, the control variable, to the Layer box in the crosstabs dialog. Figure 1 shows what this would look like for a crosstabulation with the independent variable SEX, the dependent variable HAPMAR, and the control variable DIVORCE. In other words, this analysis is exploring whether being male or female influences respondents' feelings of happiness in their marriages, controlling for whether or not they have ever been divorced. Below are the tables SPSS produces for this analysis. After the tables, the text will continue, with an explanation of how one would go about interpreting these results.

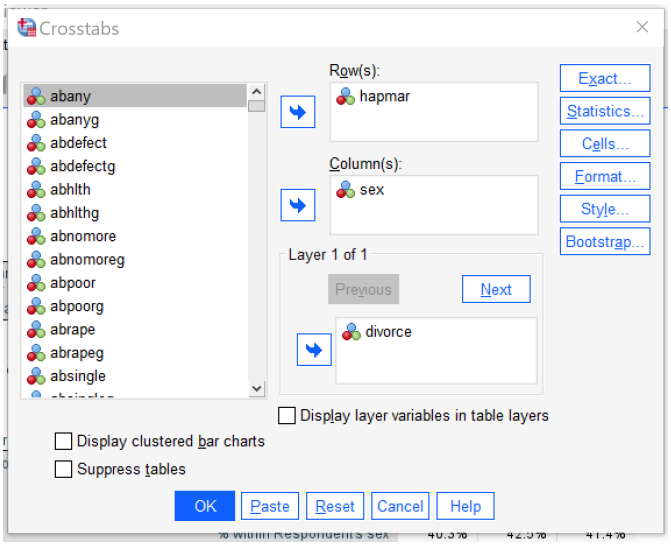


Figure 1. Crosstabs Dialog for an Analysis for SEX as Independent Variable, HAPMAR as Dependent Variable, and DIVORCE as Control Variable

Happiness of R's marriage * Respondent's sex * Ever been divorced or separated Crosstabulation

Ever been divorced or separated				Respondent's sex		Total
				male	female	
yes	Happiness of R's marriage	very happy	Count	136	133	269
			% within Respondent's sex	57.1%	52.8%	54.9%
		pretty happy	Count	96	107	203
			% within Respondent's sex	40.3%	42.5%	41.4%
		not too happy	Count	6	12	18
			% within Respondent's sex	2.5%	4.8%	3.7%
	Total		Count	238	252	490
			% within Respondent's sex	100.0%	100.0%	100.0%
no	Happiness of R's marriage	very happy	Count	467	439	906
			% within Respondent's sex	65.6%	59.8%	62.7%
		pretty happy	Count	224	260	484
			% within Respondent's sex	31.5%	35.4%	33.5%
		not too happy	Count	21	35	56
			% within Respondent's sex	2.9%	4.8%	3.9%
	Total		Count	712	734	1446
			% within Respondent's sex	100.0%	100.0%	100.0%
Total	Happiness of R's marriage	very happy	Count	603	572	1175
			% within Respondent's sex	63.5%	58.0%	60.7%
		pretty happy	Count	320	367	687
			% within Respondent's sex	33.7%	37.2%	35.5%
		not too happy	Count	27	47	74
			% within Respondent's sex	2.8%	4.8%	3.8%
	Total		Count	950	986	1936
			% within Respondent's sex	100.0%	100.0%	100.0%

Chi-Square Tests

Ever been divorced or separated		Value	df	Asymptotic Significance (2-sided)
yes	Pearson Chi-Square	2.231[b]	2	.328
	Likelihood Ratio	2.269	2	.322
	Linear-by-Linear Association	1.649	1	.199
	N of Valid Cases	490		
no	Pearson Chi-Square	6.710[c]	2	.035
	Likelihood Ratio	6.748	2	.034
	Linear-by-Linear Association	6.524	1	.011
	N of Valid Cases	1446		
Total	Pearson Chi-Square	8.772[a]	2	.012
	Likelihood Ratio	8.840	2	.012
	Linear-by-Linear Association	8.200	1	.004
	N of Valid Cases	1936		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 36.31.

b. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.74.

c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 27.57.

Symmetric Measures

Ever been divorced or separated			Value	Approximate Significance
yes	Nominal by Nominal	Phi	.067	.328
		Cramer's V	.067	.328
	N of Valid Cases		490	
no	Nominal by Nominal	Phi	.068	.035
		Cramer's V	.068	.035
	N of Valid Cases		1446	
Total	Nominal by Nominal	Phi	.067	.012
		Cramer's V	.067	.012
	N of Valid Cases		1936	

First, consider the crosstabulation table. As you can see, this table really consists of three tables stacked on top of each other. Each of these three tables considers the relationship between sex and the happiness of the respondent's marriage, but there is one table for

those who have ever been divorced, one table for those who have never been divorced, and one table for everyone. Comparing the percentages across the rows, we can make the following observations:

- Among those who *have* ever been divorced, males are slightly more likely to be very happy in their marriage, while females are somewhat more likely to be not too happy.
- Among those who *have not* ever been divorced, males are more likely to be very happy in their marriage, while females are more likely to be pretty happy and are somewhat more likely to be not too happy.
- Among the entire sample, males are more likely to be very happy in their marriages, while females are more likely to be pretty happy and somewhat more likely to be not too happy.
- Overall, then, the results suggest men are happier in their marriages than women.

Next, we turn to statistical significance. At the $p<0.05$ level, we can observe that this analysis produces significant results for those who have never been divorced and for the entire sample, but *not* for those who *have* been divorced. Turning to the association, we find a weak association—the figures for those who have been divorced, those who have not been divorced, and the entire population are quite similar.

Thus, we can conclude that women who have never been divorced are, on average, less happy in their marriages than men who have never been divorced, but that among those who *have been* divorced, the relationship between sex and marital happiness is not statistically significant.

Exercises

Select three variables of interest. Answer the following questions:

- Which is the independent variable, which is the dependent variable, and which is the control variable?
- What is the research hypothesis for this analysis? What do you predict will be the relationship between the independent variable and the dependent variable, and how will the control variable impact this relationship?
- What is the null hypothesis for this analysis?
- What confidence level (p value) have you chosen?
- Which measure of association is most appropriate for this relationship?

Next, use SPSS to produce a multivariate crosstabulation according to the instructions in this chapter. Interpret the crosstabulation. First, answer the following questions for each of the stacked crosstabula-

tions of your independent and dependent variable (one for each category of the control variable, plus one for everyone):

- Is the relationship between the independent and dependent variables statistically significant?
- Can the null hypothesis be rejected?
- How strong is the association between the two variables?
- Looking at that pattern of percentages across the rows, what can you determine about the nature of the relationship between the two variables?

Then, compare your results across the different categories of the control variable.

- What does this tell you about how the control variable impacts the relationship between the independent and dependent variables?
- Is there support for your research hypothesis?

Media Attributions

- crosstabs dialog multivariate © IBM SPSS is licensed under a All Rights Reserved license

20. Quantitative Analysis with SPSS: Comparing Means

MIKAILA MARIEL LEMONIK ARTHUR

In prior chapters, we have discussed how to perform analysis using only **discrete variables**. In this chapter, we will begin to explore techniques for analyzing relationships between discrete independent variables and continuous dependent variables. In particular, these techniques enable us to compare groups. For example, imagine a college course with 400 people enrolled in it. The professor gives the first exam, and wants to know if majors scored better than non-majors, or if first-year students scored worse than sophomores. The techniques discussed in this chapter permit those sorts of comparisons to be made. First, the chapter will detail descriptive approaches to these comparisons—approaches that let us observe the differences between groups as they appear in our data without performing statistical significance testing. Second, the chapter will explore statistical significance testing for these types of comparisons. Note here that the Split File technique, discussed in the chapter on Data Management, also provides a way to compare groups.

Comparing Means

The most basic way to look at differences between groups is by using the Compare Means command, found by going to Analyze → Compare Means → Means (Alt+A, Alt+M, Alt+M). Put the independent (discrete) variable in the Layer 1 of 1 box and the dependent (continuous) variable in the Dependent List box. Note that while you can use as many independent and dependent variables as you would like in one Compare Means command, Compare Means does not permit for multivariate analysis, so including more variables will just mean more paired analyses (one independent and one dependent variable at a time) will be produced. Under Options, you can select

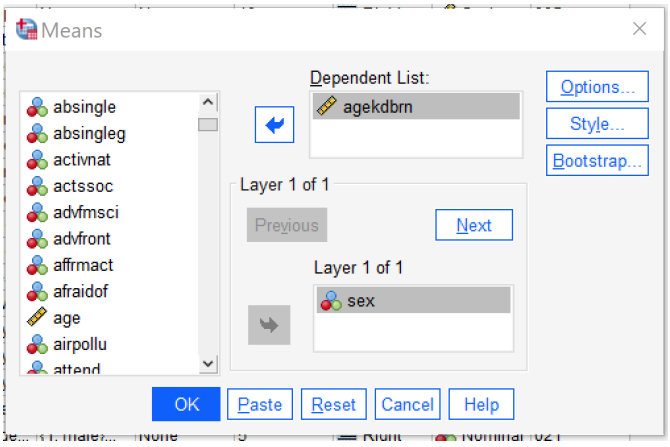


Figure 1. The Compare Means Dialog in SPSS

additional statistics to produce; the default is the mean, standard deviation, and number of cases, but other descriptive and explanatory statistics are also available. The options under Style and Bootstrap are beyond the scope of this text. One the Compare Means test is set up, click ok.

The results that appear in the Output window are quite simple: just a table listing the statistics of the dependent variable that were selected (or, if no changes were made, the default statistics as discussed above) for each category or attribute of the independent variable. In this case, we looked at the independent variable SEX and the dependent variable AGEKDBRN to see if there is a difference between the age at which men's and women's first child was born.

Table 1: Comparing Mean Age At Birth of First Child By Sex

R's age when their 1st child was born

Respondent's sex	Mean	N	Std. Deviation
male	27.27	1146	6.108
female	24.24	1601	5.917
Total	25.51	2747	6.179

Table 1, the result of this analysis, shows that the average male respondent had their first child at age 27.27, while the average female respondent had her first child at the age of 24.24, or about a three year difference. The higher standard deviation for men tells us that there is more variation in when men have their first child than there is among women.

The Compare Means command can be used with discrete independent variables with any number of categories, though keep in mind that if the number of respondents in each group becomes too small, means may reflect random variation rather than real differences.

Boxplots

Boxplots provide a way to look at this same type of relationship visually. Like Compare Means, boxplots can be used with discrete independent variables with any number of categories, though the graphs will likely become illegible when the independent variable has more than 10 or so categories. To produce a boxplot, go to Graphs → Legacy Dialogs → Boxplot (Alt+G, Alt+L, Alt+X). Click Define. Then put the discrete independent variable in the Category Axis box and the continuous dependent variable in the Variable box. Under Options it is possible to include missing values to see if those respondents differ from those who did respond, but this option is not usually selected. Other options in the Boxplot dialog generally increase the complexity of the graph in ways that may make it harder to use, so just click OK once your variables are set up.

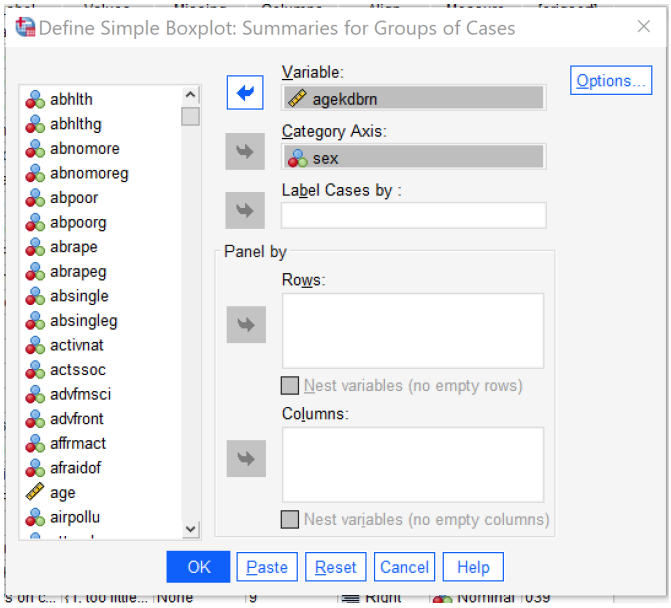


Figure 2. The Boxplot Dialog

Figure 3 displays the boxplot that is produced. It shows that the median (the thick black line), the 25th percentile (the bottom of the blue box), the 75th percentile (the top of the blue box), the low end extreme (the ⊥at the bottom of the distribution) and the high end extreme before outliers (the T at the top of the distribution) are all higher for men than women, while the most extreme outlier (the *) is pretty similar for both. Outliers are labeled with their case numbers so they can be located within the dataset. As you can see, the boxplot provides a way to describe the differences between groups visually.

T-Tests For Statistical Significance

But what if we want to know if these differences are statistically significant? That is where T-tests come in. Like the Chi square test, the T test is designed to determine statistical significance, but here, what the test

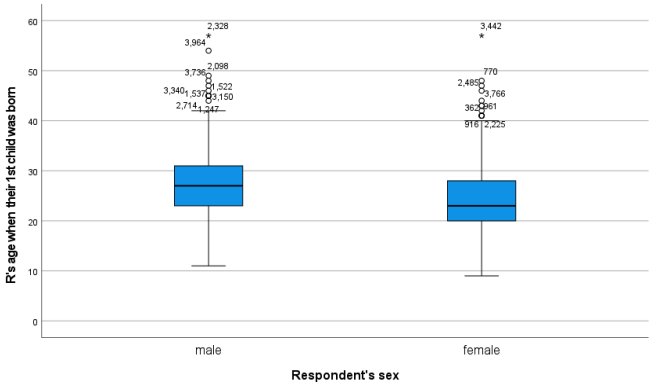


Figure 3. A Boxplot of Sex and Age at the Birth of First Child

is examining is whether there is a statistically significant difference between the means of two groups. It can only be used to compare two groups, not more than two. There are multiple types of T tests; we will begin here with the independent-samples T test, which is used to compare the means of two groups of different people.

The computation behind the T test involves the standard deviation for each category, the number of observations (or respondents) in each category, and taking the mean value for each category and computing the difference between the means (the mean difference). Like in the case of the Chi square, this produces a calculated T value and degrees of freedom that are then compared to a table of critical values to produce a statistical significance value. While SPSS will display many of the figures computed as part of this process, it produces the significance value itself so there is no need to do any part of the computation by hand.

To produce an independent samples T test, go to go to Analyze → Compare Means → Independent-Samples T Test (Alt+A, Alt+M, Alt+T). Put the continuous dependent variable in the Test Variable(s) box. Note that you can use multiple continuous dependent variables at once, but you will only be looking at differences in each one, one at a time, not at the relationships between them. Then, put the discrete independent variable in the Grouping Variable box. Click the Define Groups button, and specify the numerical values of the two groups you wish to compare[1]—keep in mind that any one T test can only compare two values, not more, so if you have a discrete variable with more than two categories, you will need to perform multiple T tests or choose another method of analysis.[2] In most cases, other options should be left as they are. For our analysis looking at differences in the age a respondent's first child was born in terms of whether the respondent is male or female, the Independent-Samples T Test dialogs would look as shown in Figure 4. AGEKDBRN is the test variable and SEX is the grouping variable, and under Define Groups, the values of 1 and 2 (the two values of the SEX variable) are entered. If we were using a variable with more than two groups, we would need to select the two groups we were interested in comparing and input the numerical values for just those two groups.

1. Yes, you will need to check variable view for this first, before proceeding to produce your T-test.
2. It is also possible to use a continuous variable for this type of analysis, with the "Cut Point" automatically dividing people into two categories.

After clicking OK to run the test, the results are produced in the output window. While multiple tables are produced, the ones most important to the analysis are called Group Statistics and Independent Samples Test, and for our analysis of sex and age at the birth of the first child, they are reproduced below as Table 2 and Table 3, respectively.

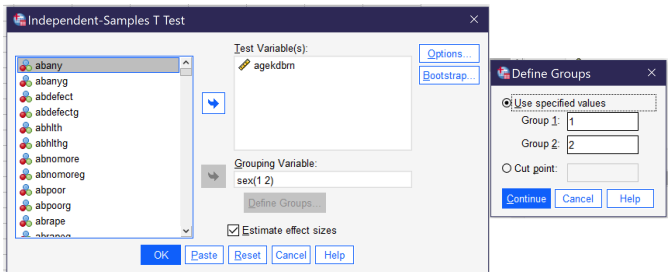


Figure 4. The Independent-Samples T Test Dialogs

Table 2. Group Statistics, Age at Birth of First Child Grouped by Sex

	Respondent's sex	N	Mean	Std. Deviation	Std. Erro
R's age when their 1st child was born	**male**	1146	27.27	6.108	.180
	female	1601	24.24	5.917	.148

Table 2 provides the number of respondents in each group (male and female), the mean age at the birth of the first child, the standard deviation, and the standard error of the mean, statistics much like those we produced in the Compare Means analysis above.

Table 3. Independent Samples Test Results, Sex by Age at Birth of First Child

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
						One-Sided p	Two-Sided p			Lower
R's age when their 1st child was born	**Equal variances assumed**	1.409	.235	13.026	2745	<.001	<.001	3.023	.232	2.568
	Equal variances not assumed			12.957	2418.993	<.001	<.001	3.023	.233	2.565

Table 3 shows the results of the T test, including the T test result, degrees of freedom, and confidence intervals. There are two rows, one for when equal variances are assumed and one for when equal variances are not assumed. If the significance under "Sig." is below 0.05, that means we should assume the variances are not equal and proceed with our analysis

using the bottom row. If the significance under "Sig." is 0.05 or above, we should treat the variances as equal and proceed using the top row. Thus, looking further at the top row, we can see the mean difference of 3.023 (which recalls the mean difference from our compare means analysis above) and the significance. Separate significance values are produced for one-sided and two-sided tests, though these are often similar. One-sided tests only look for change in one direction (increase or decrease), while two-sided tests look for any change or difference. Here, we can see both significance values are less than 0.001, so we can conclude that the observed mean difference of 3.023 does represent a statistically significant difference in the age at which men and women have their first child.

There are a number of other types of T tests. For example, the Paired Samples T Test is used when comparing two means from the same group—such as if we wanted to compare the average score on test one to the average score on test two given to the same students. There is also a One-Sample T Test, which permits analysts to compare observed data from a sample to a hypothesized value. For instance, a researcher might record the speed of drivers on a given road and compare that speed to the posted speed limit to see if drivers are going statistically significantly faster. To produce a Paired Samples T Test, go to Analyze → Compare Means → Paired-Samples T Test (Alt+A, Alt+M, Alt+P); the procedure and interpretation is beyond the scope of this text.

To produce a One-Sample T Test, go to Analyze → Compare Means → One-Sample T Test (Alt+A, Alt+M, Alt+S). Put the continuous variable of interest in the Test Variables box (Alt+T) and the comparison value in the Test Value box (Alt+V). In most cases, other options should be left as is. The results will show the sample mean, mean difference (the difference between the sample mean and the test value), and confidence intervals for the difference, as well as statistical significance tests both one-sided and two-sided. If the results are significant, that tells us that there is high likelihood that our sample differs from our predicted value; if the results are not significant, that tells us that any difference between our sample and our predicted value is likely to have occurred by chance, usually because the difference is quite small.

ANOVA

While a detailed discussion of **ANOVA**—Analysis of Variance—is beyond the scope of this text, it is another type of test that examines relationships between discrete independent variables and continuous dependent variables. Used more often in psychology than in sociology, ANOVA relies on the statistical F test rather than the T test discussed above. It enables analysts to use more than two categories of an independent variable—and to look at multiple independent variables together (including by using interaction effects to look at

how two different independent variables together might impact the dependent variable). It also, as its name implies, includes an analysis of differences in variance between groups rather than only comparing means. To produce a simple ANOVA, with just one independent variable in SPSS, go to Analyze → Compare Means → One Way ANOVA (Alt+A, Alt+M, Alt+O); the independent variable in ANOVA is called the "Factor." You can also use the Compare Means dialog discussed above to produce ANOVA statistics and eta as a measure of association by selecting the checkbox under Options. For more on ANOVA, consult a more advanced methods text or one in the field of psychology or behavioral science.

Exercises

1. Select a discrete independent variable of interest and a continuous dependent variable of interest. Run appropriate descriptive statistics for them both and summarize what you have found.
2. Run Compare Means and a Boxplot for your pair of variables and summarize what you have found.
3. Select two categories of the independent variable that you wish to compare and determine what the numerical codes for those categories are.
4. Run an independent-samples T test comparing those two categories and summarize what you have found. Be sure to discuss both statistical significance and the mean difference.

Media Attributions

- compare means © IBM SPSS is licensed under a All Rights Reserved license
- boxplot dialog © IBM SPSS is licensed under a All Rights Reserved license
- boxplot © IBM SPSS is licensed under a All Rights Reserved license
- independent samples t test dialog © IBM SPSS is licensed under a All Rights Reserved license

21. Quantitative Analysis with SPSS: Correlation

MIKAILA MARIEL LEMONIK ARTHUR

So far in this text, we have only looked at relationships involving at least one discrete variable. But what if we want to explore relationships between two continuous variables? Correlation is a tool that lets us do just that.[1] The way correlation works is detailed in the chapter on Correlation and Regression; this chapter, then, will focus on how to produce scatterplots (the graphical representations of the data upon which correlation procedures are based); bivariate correlations and correlation matrices (which can look at many variables, but only two at a time); and partial correlations (which enable the analyst to examine a bivariate correlation while controlling for a third variable).

Scatterplots

To produce a scatterplot, go to Graphs → Legacy Dialogs → Scatter/Dot (Alt+G, Alt+L, Alt+S), as shown in Figure 13 in the chapter on Quantitative Analysis with SPSS: Univariate Analysis. Choose "Simple Scatter" for a scatterplot with two variables, as shown in Figure 1.

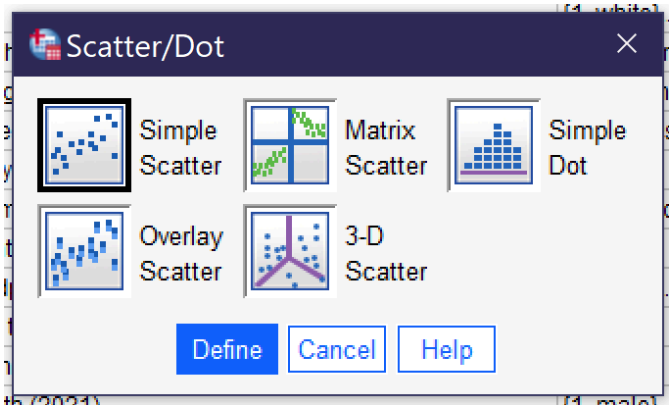


Figure 1. Scatter/Dot Graph Selection Dialog

1. Note that the bivariate correlation procedures discussed in this chapter can also be used with ordinal variables when appropriate options are selected, as will be detailed below.

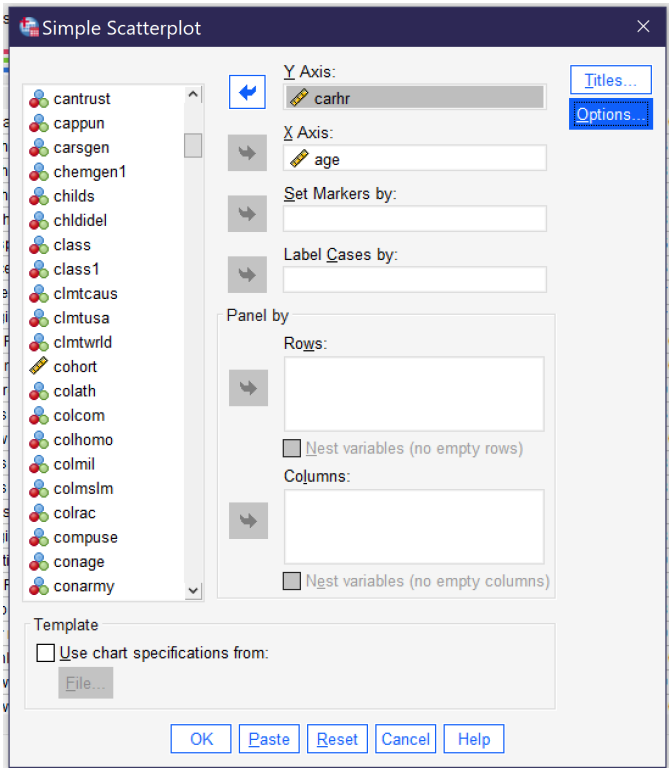


Figure 2. Simpler Scatter Dialog

This brings up the dialog for creating a scatterplot, as shown in Figure 2. The independent variable is placed in the X Axis box, as it is a graphing convention to always put the independent variable on the X axis (you can remember this because X comes before Y, therefore X is the independent variable and Y is the dependent variable, and X goes on the X axis while Y goes on the Y axis). Then the dependent variable is placed in the Y Axis box.

There are a variety of other options in the simple scatter dialog, but most are rarely used. In a small dataset, Label Cases by allows you to specify a variable that will be used to label the dots in the scatterplot (for instance, in a database of states you could label the dots with the 2-letter state code).

Once the scatterplot is set up with the independent and dependent variables, click OK to continue. The scatterplot will then appear in the output. In this case, we have used the independent variable AGE and the dependent variable CARHR to look at whether there is a relationship between the respondent's age and how many hours they spend in a car per week. The resulting scatterplot is shown in Figure 3.

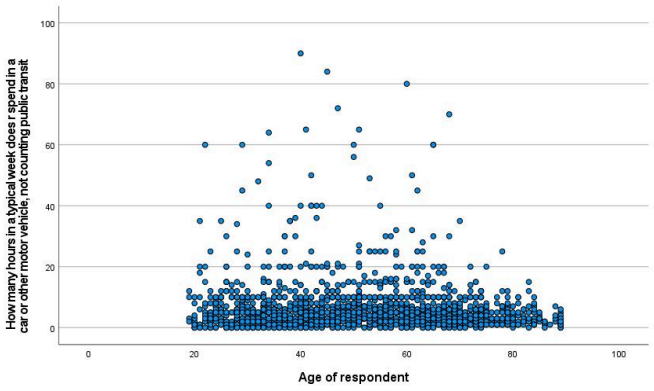


Figure 3. A Scatterplot of Age and Hours Spent in a Car Per Week

In some scatterplots, it is easy to observe the relationship between the variables. In others, like the one in Figure 3, the pattern of dots is too complex to make it possible to really see the relationship. A tool to help analysts visualize the relationship is the **line of best fit**, as discussed in the chapter on Correlation and Regression. This line is the line mathematically calculated to be the closest possible to the greatest number of dots. To add the line of best fit, sometimes called the regression line or the fit line, to your scatterplot, go to the scatterplot in the output window and double-click on it. This will open up the Chart Editor window. Then go to Elements → Fit Line at Total, as shown in Fig-

ure 4. This will bring up the Properties window. Under the Fit Line tab, be sure the Linear button is selected; click apply if needed and close out.

Doing so will add a line with an equation to the scatterplot, as shown in Figure 5.[2] From looking at the line, we can see that age age goes up, time spent in the car per week goes down, but only slightly. The equation confirms this. As shown in the graph, the equation for this line is $y = 9.04 - 0.05x$. This equation tells us that the line crosses the y axis at 9.04 and that the line goes down 0.05 hours per week in the car for every one year that age goes up (that's about 3 minutes).

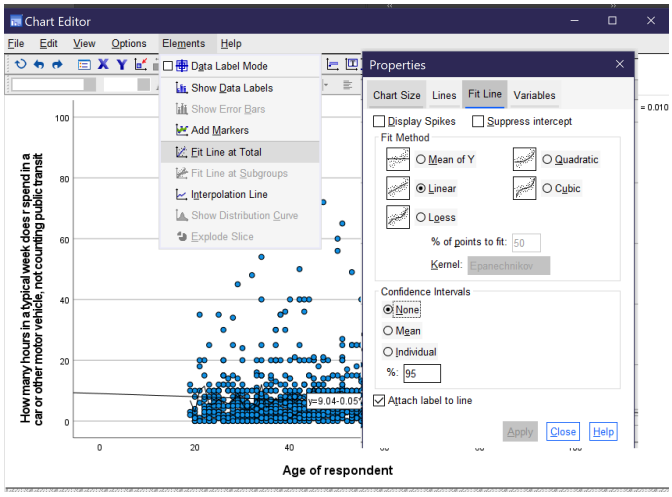


Figure 4. Adding a Fit Line to a Scatterplot

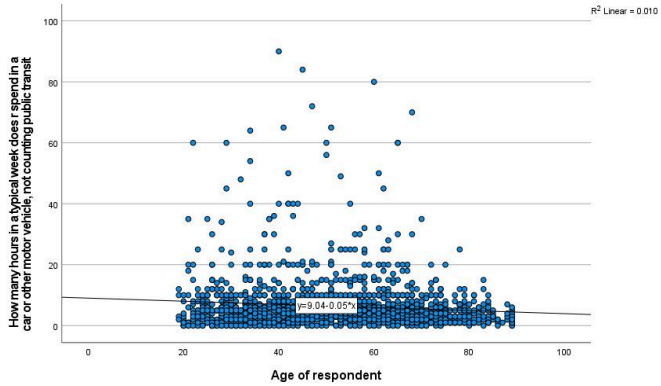


Figure 5. Scatterplot of Age and Hours Spent in the Car Per Week with Fit Line

2. It will also add the R^2; see the chapter on Correlation and Regression for more on how to interpret this.

What if we are interested in a whole bunch of different variables? It would take a while to produce scatterplots for each pair of variables. But there is an option for producing them all at once, if smaller and a bit harder to read. This is a scatterplot matrix. To produce a scatterplot matrix, go to Graphs → Legacy Dialogs → Scatter/Dot (Alt+G, Alt+L, Alt+S), as in Figure 1. But this time, choose Matrix from the dialog that appears.

In the Scatterplot Matrix dialog, select all of the variables you are interested in and put them in the Matrix Variables box, and then click OK. The many other options here, as in the case of the simple scatterplot, are rarely used.

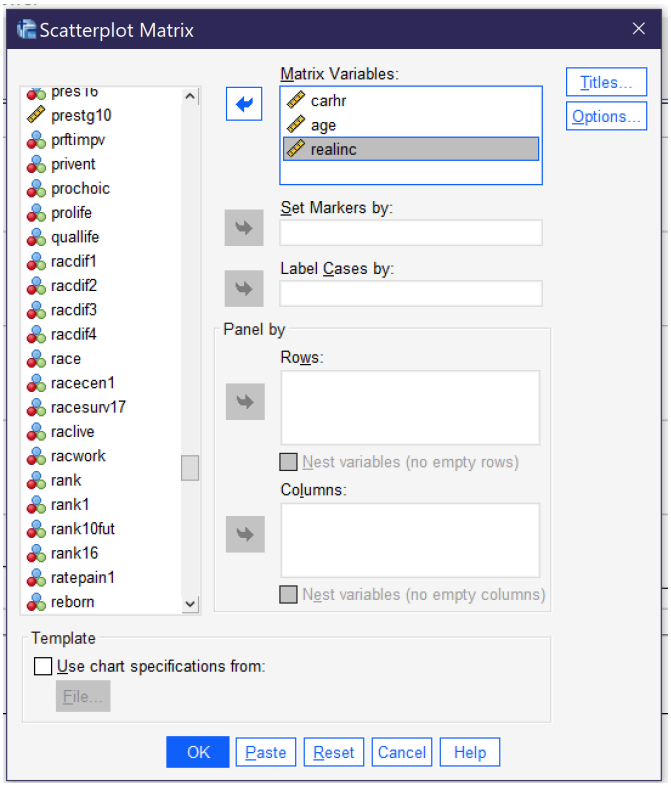


Figure 6. The Scatterplot Matrix Dialog

The scatterplot matrix will then be produced. As you can see in Figure 7, the scatterplot matrix involves a series of smaller scatterplots, one for each pair of variables specified. Here we specified CARHR and AGE, the two variables we were already using, and added REALINC, the respondent's family's income in real (inflation-adjusted) dollars. It is possible, using the same instructions detailed above, to add lines of best fit to the little scatterplots in the scatterplot matrix. Note that each little scatterplot appears twice, once with the variable on the x-axis and once with the variable on the y-axis. You only need to pay attention to one version of each pair of scatterplots.

Keep in mind that while you can include discrete variables in a scatterplot, the resulting scatterplot will be very hard to read as most of the dots will just be stacked on top of each other. See Figure 8 for an example of a scatterplot matrix that uses some binary and ordinal variables so you are aware of what to expect in such circumstances. Here, we are looking at the relationships between pairs of the three variables real family income, whether the respondent works for themselves or someone else, and how they would rate their family income from the time that they were 16 in comparison to that of others. As you can see,

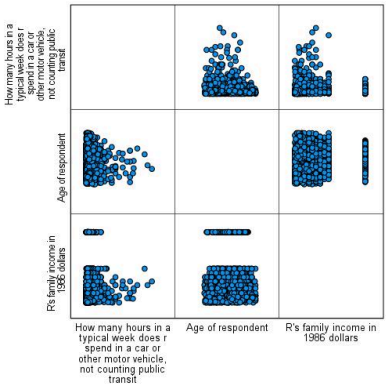


Figure 8. A Scatterplot Matrix

including discrete variables in a scatterplot produces a series of stripes which are not very useful for analytical purposes.

Correlation

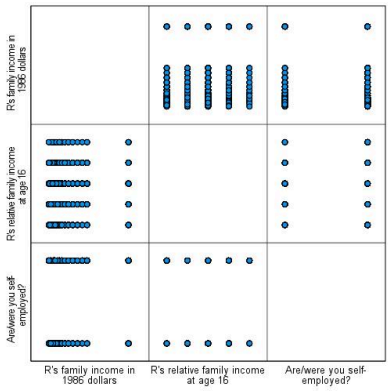


Figure 7. A Scatterplot Matrix Including an Ordinal and a Binary Variable

Scatterplots can help us visualize the relationships between our variables. But they cannot tell us whether the patterns we observe are statistically significant—or how strong the relationships are. For this, we turn to correlation, as discussed in the chapter on Correlation and Regression. Correlations are bivariate in nature—in other words, each correlation looks at the relationship between two variables. However, like in the case of the scatterplot matrix discussed above, we can produce a correlation matrix with results for a series of pairs of variables all shown in one table.

To produce a correlation matrix, go to Analyze → Correlate → Bivariate (Alt+A, Alt+C, Alt+B). Put all of the variables of interest in the Variables box. Be sure Flag significant correlations is checked and select your correlation coefficient. Note that the dialog provides the option of three different correlation coefficients, Pearson, Kendall's tau-b, and Spearman. The first, Pearson, is used when looking at the relationship between two continuous variables; the other two are used when looking at the relationship between two ordinal variables.[3] In most cases, you will want the two-tailed test of significance. Under options, you can request that means and standard deviations are also produced. When your correlation is set up, as shown in Figure 8, click OK to produce it. The results will be as shown in Table 1 (the order of variables in the table is determined by the order in which they were entered into the bivariate correlation dialog).

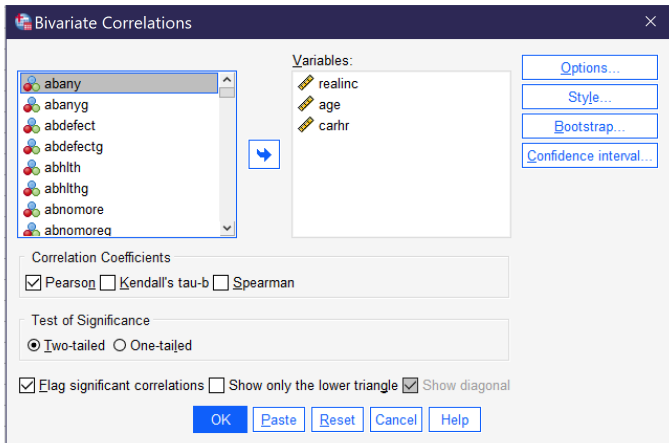


Figure 8. Bivariate Correlation Dialog

3. A detailed explanation of each of these measures of association is found in the chapter An In-Depth Look At Measures of Association.

Table 1. Bivariate Correlation Matrix

		R's family income in 1986 dollars	Age of respondent	How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit
R's family income in 1986 dollars	**Pearson Correlation**	1	.017	-.062*
	Sig. (2-tailed)		.314	.013
	N	3509	3336	1613
Age of respondent	**Pearson Correlation**	.017	1	-.100**
	Sig. (2-tailed)	.314		<.001
	N	3336	3699	1710
How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit	**Pearson Correlation**	-.062*	-.100**	1
	Sig. (2-tailed)	.013	<.001	
	N	1613	1710	1800

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

As in the scatterplot matrix above, each correlation appears twice, so you only need to look at half of the table—above or below the diagonal. Note that in the diagonal, you are seeing the correlation of each variable with itself, so a perfect 1 for complete agreement and the number of cases with valid responses on that variable. For each pair of variables, the correlation matrix includes the N, or number of respondents included in the analysis; the Sig. (2-tailed), or the p value of the correlation; and the Pearson Correlation, which is the measure of association in this analysis. It is starred to further indicate the significance level. The direction, indicated by a + or – sign, tells us whether the relationship is direct or inverse. Therefore, for each pair of variables, you can determine the significance, strength, and direction of the relationship. Taking the results in Table 1 one variable pair at a time, we can thus conclude that:

- The relationship between age and family income is not significant. (We could say there is a weak positive association, but since this association is not significant, we often do not comment on it.)
- The relationship between time spent in a car per week and family income is significant at the $p<0.05$ level. It is a weak negative relationship—in other words, as family income goes up, time spent in a car each week goes down, but only a little bit.
- The relationship between time spent in a car per week and age is significant at the

p<0.001 level. It is a moderate negative relationship—in other words, as age goes up, time spent in a car each week goes down.

Partial Correlation

Partial correlation analysis is an analytical procedure designed to allow you to examine the association between two continuous variables while controlling for a third variable. Remember that when we control for a variable, what we are doing is holding that variable constant so we can see what the relationship between our independent and dependent variables would look like without the influence of the third variable on that relationship.

Once you've developed a hypothesis about the relationship between the independent, dependent, and control or intervening variable and run appropriate descriptive statistics, the first step in partial correlation analysis is to run a regular bivariate correlation with all of your variables, as shown above, and interpret your results.

After running and interpreted the results of your bivariate correlation matrix, the next step is to produce the partial correlation by going to Analyze → Correlate → Partial (Alt+A, Alt+C, Alt+R). Place the independent and dependent variables in the Variables box, and the control variable in the Controlling for box, as shown in Figure 9. Note that the partial correlation assumes continuous variables and will only produce the Pearson correlation. The resulting partial correlation Table 2 will look much like the original bivariate correlation, but will show that the third variable has been controlled for, as shown in Table 2. To interpret the results of the partial correlation, begin by looking at the significance and association displayed and interpret them as usual.

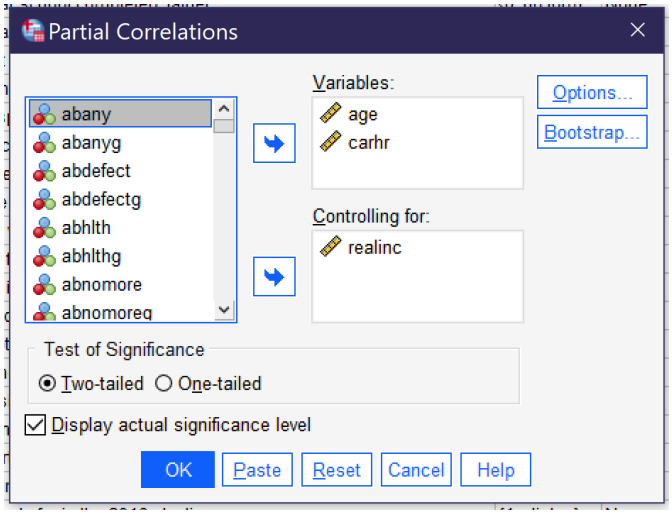


Figure 9. The Partial Correlation Dialog

Table 2. Partial Correlation

Control Variables			Age of respondent	How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit
R's family income in 1986 dollars	Age of respondent	Correlation	1.000	-.106
		Significance (2-tailed)	.	<.001
		df	0	1547
	How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit	Correlation	-.106	1.000
		Significance (2-tailed)	<.001	.
		df	1547	0

To interpret the results, we again look at significance, strength, and direction. Here, we find that the relationship is significant at the $p<0.001$ level and it is a weak negative relationship. As age goes up, time spent in a car each week goes down.

After interpreting the results of the bivariate correlation, compare the value of the measure of association in the correlation to that in the partial correlation to see how they differ. Keep in mind that we ignore the + or – sign when we do this, just considering the actual number (the absolute value). In this case, then, we would be comparing 0.100 from the bivariate correlation to 0.106 from the partial correlation. The number in the partial correlation is just a little bit higher. So what does this mean?

Interpreting Partial Correlation Coefficients

To determine how to interpret the results of your partial correlation, figure out which of the following criteria applies:

- If the correlation between x and y is **smaller** in the bivariate correlation than in the partial correlation: the third variable is a suppressor variable. This means that when we don't control for the third variable, the relationship between x and y seems smaller than it really is. So, for example, if I give you an exam with a very strict time limit to see if how much time you spend in class predicts your exam score, the exam time limit might suppress the relationship between class time and exam scores. In other words,

if we control for the time limit on the exam, your time in class might better predict your exam score.

- If the correlation between x and y is **bigger** in the bivariate correlation than in the partial correlation, this means that the third variable is a mediating variable. This is another way of saying that it is an **intervening variable**—in other words, the relationship between x and y seems larger than it really is because some other variable z intervenes in the relationship between x and y to change the nature of that relationship. So, for example, if we are interested in the relationship between how tall you are and how good you are at basketball, we might find a strong relationship. However, if we added the additional variable of how many hours a week you practice shooting hoops, we might find the relationship between height and basketball skill is much diminished.
- It is additionally possible for the **direction** of the relationship to change. So, for example, we might find that there is a direct relationship between miles run and marathon performance, but if we add frequency of injuries, then running more miles might reduce your marathon performance.
- If the value of Pearson's r is the **same or very similar** in the bivariate and partial correlations, the third variable has little or no effect. In other words, the relationship between x and y is basically the same regardless of whether we consider the influence of the third variable, and thus we can conclude that the third variable does not really matter much and the relationship of interest remains the one between our independent and dependent variables.

Finally, remember that **significance still matters**! If neither the bivariate correlation nor the partial correlation is significant, we cannot reject our null hypothesis and thus we cannot conclude that there is anything happening amongst our variables. If both the bivariate correlation and the partial correlation are significant, we can reject the null hypothesis and proceed according to the instructions for interpretation as discussed above. If the original bivariate correlation was not significant but the partial correlation was significant, we *cannot reject* the null hypothesis in regards to the relationship between our independent and dependent variables alone. However, we *can reject* the null hypothesis that there is no relationship between the variables as long as we are controlling for the third variable! If the original bivariate correlation was significant but the partial correlation was not significant, we *can reject* the null hypothesis in regards to the relationship between our independent and dependent variables, but we *cannot reject* the null hypothesis when considering the role of our third variable. While we can't be sure what is going on in such a circumstance, the analyst should conduct more analysis to try to see what the relationship between the control variable and the other variables of interest might be.

So, what about our example above? Well, the number in our partial correlation was higher, even if just a little bit, than the number in our bivariate correlation. This means that

family income is a suppressor variable. In other words, when we do not control for family income, the relationship between age and time spent in the car seems smaller than it really is. But here is where we find the limits of what the computer can do to help us with our analysis—the computer cannot explain *why* controlling for income makes the relationship between age and time spent in the car larger. We have to figure that out ourselves. What do you think is going on here?

Exercises

1. Choose two continuous variables of interest. Produce a scatterplot with regression line and describe what you see.
2. Choose three continuous variables of interest. Produce a scatterplot matrix for the three variables and describe what you see.
3. Using the same three continuous variables, produce a bivariate correlation matrix. Interpret your results, paying attention to statistical significance, direction, and strength.
4. Choose one of your three variables to use as a control variable. Write a hypothesis about how controlling for this variable will impact the relationship between the other two variables.
5. Produce a partial correlation. Interpret your results, paying attention to statistical significance, direction, and strength.
6. Compare the results of your partial correlation to the results from the correlation of those same two variables in Question 3 (when the other variable is not controlled for). How have the results changed? What does that tell you about the impact of the control variable?

Media Attributions

- scatter dot dialog © IBM SPSS is licensed under a All Rights Reserved license
- simple scatter dialog © IBM SPSS is licensed under a All Rights Reserved license
- scatter of carhrs and age © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-ND (Attribution NonCommercial NoDerivatives) license
- scatter fit line © IBM SPSS is licensed under a All Rights Reserved license
- scatter with line © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-ND (Attribution NonCommercial NoDerivatives) license
- scatterplot matrix dialog © IBM SPSS is licensed under a All Rights Reserved license
- matrix scatter © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-ND (Attribution NonCommercial NoDerivatives) license

- scatter binary ordinal © Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- bivariate correlation dialog © IBM SPSS is licensed under a All Rights Reserved license
- partial correlation dialog © IBM SPSS is licensed under a All Rights Reserved license

22. Quantitative Analysis with SPSS: Bivariate Regression

MIKAILA MARIEL LEMONIK ARTHUR

This chapter will detail how to conduct basic bivariate linear regression analysis using one continuous independent variable and one continuous dependent variable. The concepts and mathematics underpinning regression are discussed more fully in the chapter on Correlation and Regression. Some more advanced regression techniques will be discussed in the chapter on Multivariate Regression.

Before beginning a regression analysis, analysts should first run appropriate descriptive statistics. In addition, they should create a scatterplot with regression line, as described in the chapter on Quantitative Analysis with SPSS: Correlation & descriptive statistics. One important reason why is that linear regression has as a basic assumption the idea that data are arranged in a linear—or line-like—shape. When relationships are weak, it will not be possible to see just by glancing at the scatterplot whether it is linear or not, or if there is no relationship at all.

However, there are cases where it is quite obvious that there *is* a relationship, but that this relationship is not line-like in shape. For example, if the scatterplot shows a clear curve, as in Figure 1, one that could not be approximated by a line, then the relationship is not sufficiently linear to be detected by a linear regression.[1] Thus, any results you obtain from linear regression analysis would considerably underestimate the strength of such a relationship and would not be able to discern its nature. Therefore, looking at the scatterplot before running a regression allows the analyst to determine if the particular relationship of interest can appropriately be tested with a linear regression.

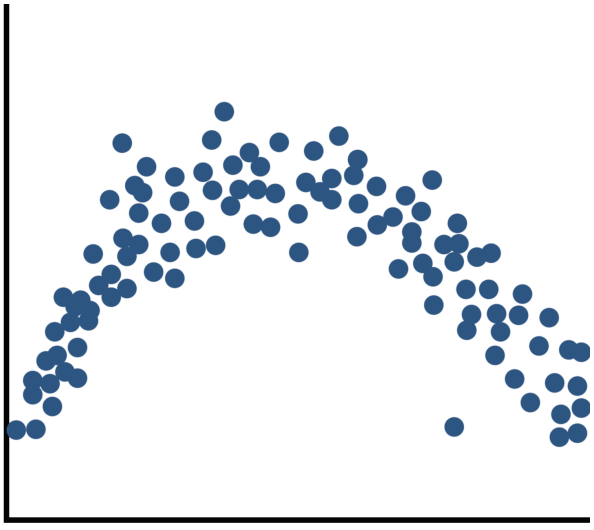


Figure 1. A Curvilinear Scatterplot

1. There are other regression techniques that *are* appropriate for such relationships, but they are beyond the scope of this text.

Assuming that the relationship of interest *is* appropriate for linear regression, the regression can be produced by going to Analyze → Regression → Linear[2] (Alt+A, Alt+R, Alt+L). The dependent variable is placed in the Dependent box; the independent in the "Block 1 of 1" box. Under Statistics, be sure both Estimates and Model fit are checked. Here, we are using the independent variable AGE and the dependent variable CARHR. Once the regression is set up, click OK to run it.

The results will appear in the output window. There will be four tables: Variables Entered/Removed; Model Summary; ANOVA; and Coefficients. The first of these simply documents the variables you have used.[3] The other three contain important elements of the analysis. Results are shown in Tables 1, 2, and 3.

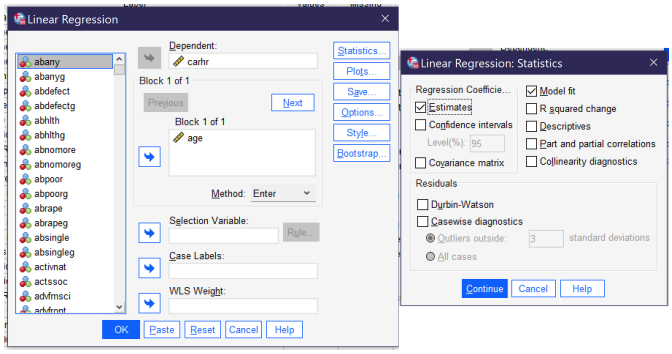


Figure 2. The Linear Regression Dialog

Table 1. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.100[a]	.010	.009	8.656

a. Predictors: (Constant), Age of respondent

Table 2. ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	1290.505	1	1290.505	17.225	<.001[b]
	Residual	127966.152	1708	74.922		
	Total	129256.657	1709			

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit

b. Predictors: (Constant), Age of respondent

2. You will notice that there are many, many options and tools within the Linear Regression dialog; some of these will be discussed in the chapter on Multivariate Regression, while others are beyond the scope of this text.
3. The Variables Entered/Removed table is important to those running a series of multivariate models while adding or removing individual variables, but is not useful when only one model is run at a time.

Table 3. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.037	.680		13.297	<.001
	Age of respondent	-.051	.012	-.100	-4.150	<.001

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit

When interpreting the results of a bivariate linear regression, we need to answer the following questions:

- What is the *significance* of the regression?
- What is the *strength* of the observed relationship?
- What is the *direction* of the observed relationship?
- What is the *actual numerical relationship*?

Each of these questions is answered by numbers found in different places within the set of tables we have produced.

First, let's look at the significance. The significance is found in two places among our results, under "Sig." in the ANOVA table (here, table 2) and under "Sig." in the Coefficients table (here, table 3). In the Coefficients table, look at the significance number in the row with the independent variable; you will also see a significance number for the constant, which will be discussed later. In a bivariate regression, these two significance numbers are the same (this is not true for multivariate regressions). So, in these results, the significance is $p<0.001$, which means we can conclude that the results are significant.

Next, we look at the strength. Again, we can look in two places for the strength, under R in the Model Summary table (here, table 1) and under Beta in the Coefficients table. Beta refers to the Greek letter β, and beta and β are used interchangeably when referring to the standardized coefficient. R, here, refers to Pearson's r, and both it and Beta are interpreted the same way as measures of association usually are. While the R and the Beta will be the same in a bivariate regression, the sign (whether the number is positive or negative) may not be; again, in multivariate regressions, the numbers will not be the same. This is because Beta is used to look at the strength of the relationship each individual independent variable has with the dependent variable. Here, the R/Beta is 0.100, so the relationship is moderate in strength.

The direction of the relationship is determined by whether the Beta is positive or negative. Here, it is negative, so that means it is an inverse relationship. In other words, as age goes up, time spent in cars each week goes down. And the B value, found in the Coeffi-

cients table, tells us by how much it goes down. Here we see that for every one year of additional age, time spent in cars goes down by about 0.051 hours (a little more than three minutes).

One final thing to look at is the R squared (R^2) in the Model Summary table. The R^2 tells us how much of the variance in our dependent variable is explained by our independent variable. Here, then, age explains 1% (0.010 converted to a percent by multiplying it times 100) of the variance in time spent in a car each week. That might not seem like very much, and it is not very much. But considering all the things that matter to how much time you spend in a car each week, it is clear that age is contributing somehow.

The numbers in the Coefficients table also allow us to construct the regression equation (the equation for the line of best fit). The number under B for the constant row is the y intercept (in other words, if X were 0, what would Y be?), and the number under B for the variable is the slope of the line. We apply asterisks to indicate significance, giving us the following equation: $y = 9.037 - 0.051x^{***}$. Note that whether or not the constant/intercept is statistically significant is just telling us whether the constant/intercept is statistically significantly different from zero, which is not actually very interesting, and thus most analysts do not pay much attention to the significance of the constant/intercept.

So, in summary, our results tell us that age has a significant, moderate, inverse relationship with time spent in a car each week; that age explains 1% of the variance in time spent in the car each week, and that for every one year of additional age, just over 3 more minutes per week are spent in the car.

Exercises

1. Choose two continuous variables of interest. Write a hypothesis about the relationship between the variables.
2. Create a scatterplot for these two variables with regression line (line of best fit). Explain what the scatterplot shows.
3. Run a bivariate regression for these two variables. Interpret the results, being sure to discuss significance, strength, direction, and the actual magnitude of the effect.
4. Create the regression equation for your regression results.

Media Attributions

- curvilinear © Mikaila Mariel Lemonik Arthur is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- linear regression dialog © IBM SPSS is licensed under a All Rights Reserved license

23. Quantitative Analysis with SPSS: Multivariate Regression

MIKAILA MARIEL LEMONIK ARTHUR

In the chapter on Bivariate Regression, we explored how to produce a regression with one independent variable and one dependent variable, both of which are continuous. In this chapter, we will expand our understanding of regression. The regressions we produce here will still be linear regressions with one continuous dependent variable, but now we will be able to include more than one independent variable. In addition, we will learn how to include discrete independent variables in our analysis.

In fact, producing and interpreting multivariate linear regressions is not very different from producing and interpreting bivariate linear regressions. The main differences are:

1. We add one or more additional variables to the Block 1 of 1 box (where the independent variables go) when setting up the regression analysis,
2. We check off one additional option under Statistics when setting up the regression analysis, **Collinearity** diagnostics, which will be explained below,
3. We interpret the strength and significance of the entire regression and then look at the strength, significance, and direction of each included independent variable one at a time, so there are more things to interpret, and
4. We can add or remove variables and compare the R^2 to see how those changes impacted the overall predictive power of the regression.

Each of these differences between bivariate and multivariate regression will be discussed below, beginning with the issue of collinearity and the tools used to diagnose it.

Collinearity

Collinearity refers to the situation in which two independent variables in a regression analysis are closely correlated with one another (when more than two independent variables are closely correlated, we call it multicollinearity). This is a problem because when the correlation between independent variables is high, the impact of each individual variable on the dependent variable can no longer be separately calculated. Collinearity can occur in a variety of circumstances: when two variables are measuring the same thing but using differ-

ent scales; when they are measuring the same concept but doing so slightly differently; or when one of the variables has a very strong effect on the other.

Let's consider examples of each of these circumstances in turn. If a researcher included both year of birth and age, or weight in pounds and weight in kilograms, both of the variables in each pair are measuring the exact same thing. Only the scales are different. If a researcher included both hourly pay and weekly pay, or the length of commute in both distance and time, the correlation would not be quite as close. A person might get paid $10 an hour but work a hundred hours per week, or get paid $100 an hour but work ten hours per week, and thus still have the same weekly pay. Someone might walk two miles to work and spend the same time commuting as someone else driving 35 miles on the highway. But overall, the relationships between hourly pay and weekly pay and the relationship between commute distance and commute time are likely to be quite strong. Finally, consider a researcher who includes variables measuring the grade students earned on Exam 1 and their total grade in a course with three exams, or one who includes variables measuring families' spending on housing each month and their overall spending each month. In these cases, the variables are not measuring the same underlying phenomena, but the first variable likely has a strong effect on the second variable, resulting in a strong correlation.

In many cases, the potential for collinearity will be obvious when considering the variables included in the analysis, as in the examples above. But it is not always obvious. Therefore, researchers need to test for collinearity when performing multivariate regressions. There are several ways to do this. First of all, before beginning to run a regression, researchers can check for collinearity by running a correlation matrix and a scatterplot matrix to look at the correlations between each pair of variables. The instructions for these techniques can be found in the chapter on Quantitative Analysis with SPSS: Correlation. A general rule of thumb is that if a Pearson correlation is above 0.8, this suggests a likely problem with collinearity, though some suggest scrutinizing those pairs of variables with a correlation above 0.7.

In addition, when running the regression, researchers can check off the option for Collinearity diagnostics (Alt+l) under the statistics dialog (Alt+S), as shown in Figure 1. The resulting regression's Coefficients table will include two additional pieces of information, the VIF and the Tolerance, as well as an additional table called Collinearity diagnostics. The VIF, or Variance Inflation Factor, calculates the degree of collinearity present. Values of around or close to one suggest no collinearity; values around four or five suggest that a deeper look at the variables is needed, and values at ten or above definitely suggest collinearity great enough to be problematic for the regression analysis. The Tolerance measure calculates the extent to which other independent variables can predict the values of the variable under consideration; for tolerance, the smaller the number, the more likely that collinearity is a problem. Typically, researchers performing relatively straightforward regressions such as those detailed in this chapter do not need to rely on the Collinearity diagnostics table, as they will be able to determine which variables may be correlated with one another by simply considering the variables and looking at the Tolerance and VIF statistics.

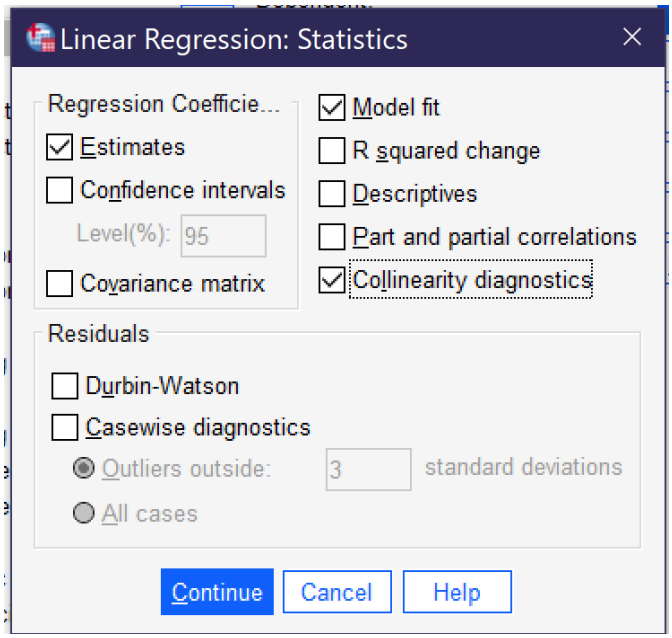


Figure 1. Using Collinearity Diagnostics in Regression

Producing & Interpreting Multivariate Linear Regressions

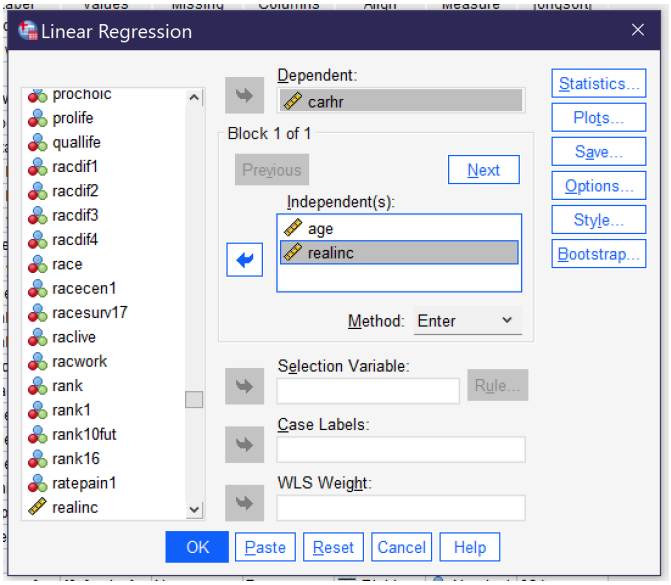


Figure 2. The Linear Regression Dialog Set Up With CARHR as Dependent and AGE and REALRINC as Independent

Producing multivariate linear regressions in SPSS works just the same as producing bivariate linear regressions, except that we add one or more additional variables to the Block 1 of 1 box and check off the Collinearity diagnostics, as shown in Figure 2. Let's continue our analysis of the variable CARHR, adding the independent variable REALINC (inflation-adjusted family income) to the independent variable AGE. Figure 2 shows how the linear regression dialog would look when set up to run this regression, with CARHR in the Dependent box and AGE and REALINC in the Independent(s) box under Block 1 of 1. Be sure that Estimates, Model fit, and Collinearity diagnostics are checked off, as shown in Figure 1. Then click OK to run the regression.

Tables 1, 2, and 3 below show the results (excluding those parts of the output unnecessary for interpretation.

Table 1. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.124[a]	.015	.014	8.619

a. Predictors: (Constant), R's family income in 1986 dollars, Age of respondent

Table 2. ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1798.581	2	899.290	12.106	<.001[b]
	Residual	114913.923	1547	74.282		
	Total	116712.504	1549			

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit

b. Predictors: (Constant), R's family income in 1986 dollars, Age of respondent

Table 3. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	9.808	.752		13.049	<.001		
	Age of respondent	-.055	.013	-.106	-4.212	<.001	1.000	1.00
	R's family income in 1986 dollars	-1.356E-5	.000	-.064	-2.538	.011	1.000	1.00

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not coun public transit

So, how do we interpret the results of our multivariate linear regression? First, look at the Collinearity Statistics in the Coefficients table (here, Table 3). As noted above, to know we are not facing a situation involving collinearity, we are looking for a VIF that's lower than 5 and a Tolerance that is close to 1. Both of these conditions are met here, so collinearity is unlikely to be a problem. If it were, we would want to figure out which variables were overly correlated and remove at least one of them. Next, we look at the overall significance of the regression in the ANOVA table (here, Table 3). The significance shown is <0.001, so the regression is significant. If it were not, we would stop there.

To find out the overall strength of the regression, we look at the R in the Model Summary (here, Table 1). It is 0.124, which means it is moderately strong. The R^2 is 0.015, which—converting the decimal into a percentage by multiplying it by 100—tells us that the two independent variables combined explain 1.5% of the variance in the dependent variable, how much time the respondent spends in the car. And here's something fancy you can do with that R^2: compare it to the R^2 for our prior analysis in the chapter on Bivariate Regression, which had just the one independent variable of AGE. That R^2 was 0.10, so though our new regression still explains very little of the variance in hours spent in the car, adding income does enable us to explain a bit more of the variance. Note that you can only compare R^2 values among a series of models with the same dependent variable. If you change dependent variables, you can no longer make that comparison.

Now, let's turn back to the Coefficients table. When we interpreted the bivariate regression results, we saw that the significance and Beta values in this table were the same as the significance value in the ANOVA table and the R values, respectively. In the multivariate regression, this is no longer true—because now we have *multiple* independent variables, each with their *own* significance and Beta values. These results allow us to look at each independent variable, *while holding constant (**controlling for**) the effects of the other independent variable(s).* Age, here, is significant at the p<0.001 level, and its Beta value is -0.106, showing a moderate negative association. Family income is significant at the p<0.05 level, and its Beta value is -0.064, showing a weak negative association. We can compare the Beta

values to determine that age has a larger effect (0.106 is a bigger number than 0.064; we ignore sign when comparing strength) than does income.

Next, we look at the B values to see the actual numerical effect of each variable. For every year of additional age, respondents spend on average 0.055 fewer hours in the car, or about 1.65 minutes less. And for every dollar of additional family income, respondents spend -1.356E-5 fewer hours in the car. But wait, what does -1.356E-5 mean? It's a way of writing numbers that have *a lot* of decimal places so that they take up less space. Written the long way, this number is -0.00001356—so what the E-5 is telling us is to move the decimal point five spaces over. That's a pretty tiny number, but that's because an increase of $1 in your annual family income really doesn't have much impact on, well, really anything. If instead we considered the impact of an increase of $10,000 in your annual family income, we would multiply our B value by $10,000, getting -0.1356. In other words, an increase of $10,000 in annual family income (in constant 1986 dollars) is associated with an average decrease of 0.1356 hours in the car, or a little more than 8 minutes.

Our final step is to create the regression equation. We do this the same way we did for bivariate regression, only this time, there is more than one x, so we have to indicate the coefficient and significance of each one separately. Some people do this by numbering their xs with subscript numerals (e.g. x_1, x_2, and so on), while others just use the short variable name. We will do the later here. Taking the numbers from the B column, our regression equation is

$$y = 9.808 - 0.055AGE^{***} - 1.356E-5REALINCOME^{*}.$$

Phew, that was a lot to go through! But it told us a lot about what is going on with our dependent variable, CARHR. That's the power of regression: it tells us not just about the strength, significance, and direction of the relationship between a given pair of variables, but also about the way adding or removing additional variables changes things as well as about the actual impact each independent variable has on the dependent variable.

Dummy Variables

So far, we have reviewed a number of the advantages of regression analysis, including the ability to look at the significance, strength, and direction of the relationships between a series of independent variables and a dependent variable; examining the effect of each independent variable while controlling for the others; and seeing the actual numerical effect of each independent variable. Another advantage is that it is possible to include independent variables that are discrete in our analysis. However, they can only be included in a very specific way: if we transform them into a special kind of variable called a **dummy variable** in which a single value of interest is coded as 1 and all other values are coded as 0. It is

even possible to create multiple dummy variables for different categories of the same discreet variable, so long as you have an excluded category or set of categories that are sizable. It is important to leave a sizeable group of respondents or datapoints in the excluded category because of **collinearity**.

Consider, for instance, the variable WRKSLF, which asks if respondents are self-employed or work for someone else. This is a binary variable, with only two answer choices. We could make a dummy variable for self-employment, with being self-employed coded as 1 and everything else (which, here, is just working for someone else) as 0. Or we could make a dummy variable for working for someone else, with working for someone else coded as 1 and everything else as 0. But we cannot include both variables in our analysis because they are, fundamentally, measuring the same thing.

Figuring out how many dummy variables to make and which ones they should be can be difficult. The first question is theoretical: what are you actually interested in? Only include categories you think would be meaningfully related to the outcome (dependent variable) you are considering. Second, look at the descriptive statistics for your variable to be sure you have an excluded category or categories. If all of the categories of the variable are sufficiently large, it may be enough to exclude one category. However, if a category represents very few data points—say, just 5 or 10 percent of respondents—it may not be big enough to avoid collinearity. Therefore, some analysts suggest using one of the largest categories, assuming this makes sense theoretically, as the excluded category.

Let's consider a few examples:

Table 4. Examples of Dummy Variables

GSS Variable	Answer Choices & Frequencies	Suggested Dummy Variable(s)
RACE	White: 78.2% Black: 11.6% Other: 10.2%	Option 1. 2 variables: Back 1, all others 0 & Other 1, all others 0 Option 2. Nonwhite 1, all others 0 Option 3. White 1, all others 0
DEGREE	Less than high school: 6.1% High school: 38.8% Associate/junior college: 9.2% Bachelor's: 25.7% Graduate: 18.8%	Option 1. Bachelor's or higher 1; all others 0 Option 2. High school or higher 1; all others 0 Option 3. 4 variables: Less than high school 1, all others 0; Associate/junior college 1, all others 0; Bachelor's 1, all others 0; Graduate 1, all others 0 Option 4. Use EDUC instead, as it is continuous
CHILDS	0: 29.2% 1: 16.2% 2: 28.9% 3: 14.5% 4: 7% 5: 2% 6: 1.3% 7: 0.4% 8 or more: 0.5%	Option 1. 0 children 1, all others 0 Option 2. 2 variables: 0 children 1, all others 0; 1 child 1, all others 0 Option 3. 3 variables: 0 children 1, all others 0; 1 child 1, all others 0; 2 children 1, all others 0 Option 4. Ignore the fact that this variable is not truly continuous and treat is as continuous anyway
CLASS	Lower class: 8.7% Working class: 27.4% Middle class: 49.8% Upper class: 4.2%	The best option is to create thee variables: Lower class 1, all others 0; Working class 1, all others 0; Upper class 1, all others 0 (however, you could instead include Working class and have a variable for Middle class if that made more sense theoretically)
SEX	Male: 44.1% Female: 55.9%	Option 1. Male 1, all others 0 Option 2: Female 1, all others 0

So, how do we go about making our dummy variable or variables? We use the Recode technique, as illustrated in the chapter on Quantitative Analysis with SPSS: Data Management. Just remember to Recode into different and to make as many dummy variables as needed: maybe one, maybe more. Here, we will make one for SEX. Because we are continuing our analysis of CARHRS, let's assume we hypothesize that, on average, women spend more time in the car than men because women are more likely to be responsible for driving children to school and activities. On the basis of this hypothesis, we would treat female as the included category (coded 1) and male as the excluded category (coded 0) since what we are interested in is the *effect of being female*.

As a reminder, to recode, we first make sure we know the value labels for our existing original variable, which we can find out by checking Values in Variable View. Here, male is 1 and female is 2. Then we go to Transform → Recode into Different Variables (Alt+T, Alt+R). We add the original variable to the box, and then give our new variable a name, here generally

something like the name of the category we are interested in (here, Female) and descriptive label, and click Change. Next, we click "Old and New Values." We set system or user missing as system missing, our category of interest as 1, and everything else as 0. We click continue, then go to the bottom of variable view and edit our value labels to reflect our new categories. Finally, we run a frequency table of our new variable to be sure everything worked right. Figure 3 shows all of the steps described here.

Dummy Variables in Regression Analysis

After creating the dummy variable, we are ready to include our dummy variable in a regression. We set up the regression just the same way as we did above, except that we add FEMALE to the independent variables REALINC and AGE (the dependent variable will stay CARHR). Be sure to check Collinearity diagnostics under Statistics. Figure 4 shows how the linear regression dialog should look with this regression set up. Once the regression is set up, click ok to run it.

Now, let's consider the output, again focusing only on those portions of the output necessary to our interpretation, as shown in Tables 5, 6, and 7.

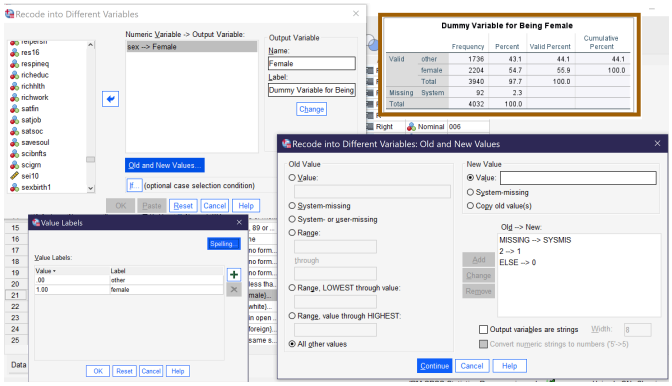


Figure 3. The Process of Recoding Sex to Create the Dummy Variable Female

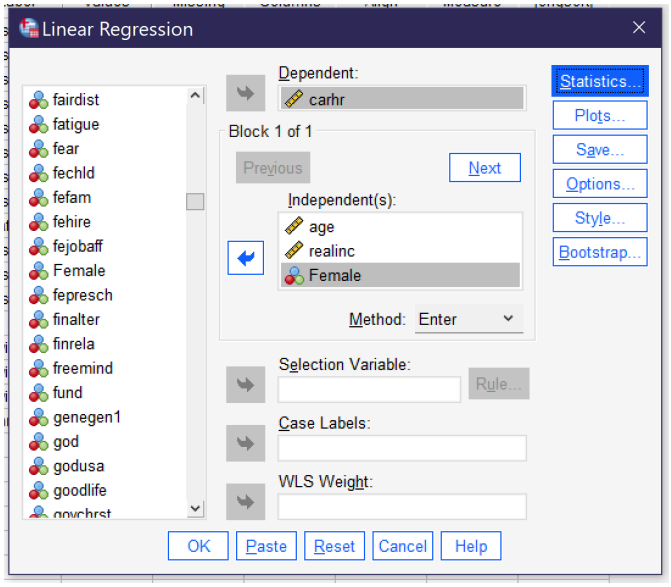


Figure 4. The Multivariate Linear Regression Window with our Dummy Variable Added

Table 5. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.143[a]	.020	.018	8.602

a. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars

Table 6. ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2372.292	3	790.764	10.686	<.001[b]
	Residual	114259.126	1544	74.002		
	Total	116631.418	1547			

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit

b. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars

Table 7. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	10.601	.801		13.236	<.001		
	Age of respondent	-.056	.013	-.108	-4.301	<.001	1.000	1.00
	R's family income in 1986 dollars	-1.548E-5	.000	-.073	-2.882	.004	.986	1.014
	Dummy Variable for Being Female	-1.196	.442	-.069	-2.706	.007	.986	1.015

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not coun public transit

First, we look at our collinearity diagnostics in the Coefficients table (here, Table 7). We can see that all three of our variables have both VIF and Tolerance close to 1 (see above for a more detailed explanation of how to interpret these statistics), so it is unlikely that there is a collinearity problem.

Second, we look at the significance for the overall regression in the ANOVA table (here, Table 6). We find the significance is <0.001, so our regression is significant and we can continue our analysis.

Third, we look at the Model Fit table (here, table 5). We see that the R is 0.143, so the regression's strength is moderate, and the R^2 is 0.02, meaning that all of our variables together explain 2% of the variance (0.02 * 100 converts the decimal to a percent) in our dependent variable. We can compare this 2% R^2 to the 1.5% R^2 we obtained from the earlier regression without Female and determine that adding the dummy variable for being female helped our regression explain a little bit more of the variance in time respondents spend in the car.

Fourth, we look at the significance and Beta values in the Coefficients table. First, we find that Age is significant at the p<0.001 level and that it has a moderate negative relationship with time spent in the car. Second, we find that income is significant at the p<0.01

level and has a weak negative relationship with time spent in the car. Finally, we find that being female is significant at the p<0.01 level and has a weak negative relationship with time spent in the car. But wait, what does this mean? Well, female here is coded as 1 and male as 0. So what this means is that when you move from 0 to 1—in other words from male to female—the time spent in the car goes down (but weakly). This is the opposite of what we hypothesized! Of the three variables, age has the strongest effect (the largest Beta value).

Next, we look at the B values to see what the actual numerical effect is. For every one additional year of age, time spent in the car goes down by 0.056 hours (3.36 minutes) a week. For every one additional dollar of income, time spent in the car goes down by -1.548E-5 hours per week; translated (as we did above), this means that for every $10,000 additional dollars of income, time spent in the car goes down by 0.15 hours (about 9 minutes) per week. And women, it seems, spend on average 1.196 hours (about one hour and twelve minutes) fewer per week in the car than do men.

Finally, we produce our regression equation. Taking the numbers from the B column, our regression equation is $y = 10.601 - 0.056AGE^{***} - 1.584E-5REALINCOME^{**} - 1.196FEMALE^{**}$.

Regression Modeling

There is one more thing you should know about basic multivariate linear regression. Many analysts who perform this type of technique systematically add or remove variables or groups of variables in a series of regression models (SPSS calls them "Blocks") to look at how they influence the overall regression. This is basically the same as what we have done above by adding a variable and comparing the R^2 (the difference between the two R^2 values is called the **R^2 change**). However, SPSS provides a tool for running multiple blocks at once and looking at the results. When looking at the Linear regression dialog, you may have noticed that it says "Block 1 of 1" just above the box where the independent variables go. Well, if you click "next" (Alt+N), you will be moved to a blank box called "Block 2 of 2". You can then add additional independent variables here as an additional block.

Just below the Block box is a tool called "Method" (Alt+M). While a description of the options here is beyond the scope of this text, this tool provides different ways for variables in each block to be entered or removed from the regression to develop the regression model that is most optimal for predicting the dependent variable, retaining only those variables that truly add to the predictive power of the ultimate regression equation. Here, we will stick with the "Enter" Method, which does not draw on this type of modeling but instead

simply allows us to compare two (or more) regressions upon adding an additional block (or blocks) of variables.

So, to illustrate this approach to regression analysis, we will retain the same set of variables for Block 1 that we used above: age, income, and the dummy variable for being female. And then we will add a Block 2 with EDUC (the highest year of schooling completed) and PRESTIG10 (the respondent's occupational prestige score)[1]. Remember to be sure to check the collinearity diagnostics box under statistics. Figure 5 shows how the regression dialog should be set up to run this analysis.

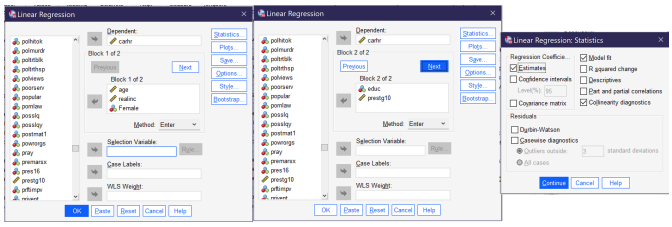


Figure 5. Setting Up a Linear Regression With Blocks

The output for this type of analysis (relevant sections of the output appear as Tables 8, 9, and 10) does look more complex at first, as each table now has two tables stacked on top of one another. Note the output will first, before the relevant tables, include "Variables Entered/Removed" table that simply lists which variables are included in each block. This is more important for the more complex methods other than Enter in which SPSS calculates the final model; here, we already know which variables we have included in each block.

Table 7. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.155[a]	.024	.022	8.201
2	.200[b]	.040	.037	8.139

a. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars

b. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars, R's occupational prestige score (2010), Highest year of school R completed

1. Occupational prestige is a score assigned to each occupation. The score has been determined by administering a prior survey in which respondents were asked to rank the prestige of various occupations; these rankings were consolidated into scores. Census occupational codes were used to assign scores of related occupations to those that had not been asked about in the original survey.

Table 8. ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2486.883	3	828.961	12.326	<.001[b]
	Residual	101353.178	1507	67.255		
	Total	103840.061	1510			
2	Regression	4144.185	5	828.837	12.512	<.001[c]
	Residual	99695.876	1505	66.243		
	Total	103840.061	1510			

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not counting public transit

b. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars

c. Predictors: (Constant), Dummy Variable for Being Female, Age of respondent, R's family income in 1986 dollars, R's occupational prestige score (2010), Highest year of school R completed

Table 9. Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	10.642	.783		13.586	<.001		
	Age of respondent	-.055	.013	-.111	-4.344	<.001	1.000	1.00
	R's family income in 1986 dollars	-1.559E-5	.000	-.077	-3.003	.003	.986	1.014
	Dummy Variable for Being Female	-1.473	.426	-.089	-3.456	<.001	.986	1.015
2	(Constant)	16.240	1.411		11.514	<.001		
	Age of respondent	-.053	.013	-.107	-4.216	<.001	.995	1.00
	R's family income in 1986 dollars	-4.286E-6	.000	-.021	-.762	.446	.827	1.20
	Dummy Variable for Being Female	-1.576	.424	-.095	-3.722	<.001	.983	1.017
	Highest year of school R completed	-.279	.092	-.092	-3.032	.002	.691	1.44
	R's occupational prestige score (2010)	-.040	.018	-.067	-2.225	.026	.712	1.40

a. Dependent Variable: How many hours in a typical week does r spend in a car or other motor vehicle, not coun public transit

You will notice, upon inspecting the results, that what appears under Model 1 (the rows with the 1 at the left-hand side) is the same as what appeared in our earlier regression in this

chapter, the one where we added the dummy variable for being female. That is because, in fact, Model 1 is the same regression as that prior regression. Therefore, here we only need to interpret Model 2 and compare it to Model 1; if we had not previously run the regression that is shown in Model 1, we would also need to interpret the regression in Model 1, not just the regression in Model 2. But since we do not need to do that here, let's jump right in to interpreting Model 2.

We begin with collinearity diagnostics in the Coefficients table (here, Table 9). We can see that the Tolerance and VIF have moved further away from 1 than in our prior regressions. However, the VIF is still well below 2 for all variables, while the Tolerance remains above 0.5. Inspecting the variables, we can assume the change in Tolerance and VIF may be due to the fact that education and occupational prestige are strongly correlated. And in fact, if we run a bivariate correlation of these two variables, we do find that the Pearson's R is 0.504—indeed a strong correlation! But not quite so strong as to suggest that they are too highly correlated for regression analysis.

Thus, we can move on to the ANOVA table (here, Table 8). The ANOVA table shows that the regression is significant at the $p<0.001$ level. So we can move on to the Model Summary table (here, Table 7). This table shows that the R is 0.200, still a moderate correlation, but a stronger one than before. And indeed, the R^2 is 0.040, telling us that all of our independent variables together explain about 4% of the variance in hours spent in the car per week. If we compare this R^2 to the one for Model 1, we can see that, while the R^2 remains relatively small, the predictive power has definitely increased with the addition of educational attainment and occupational prestige to our analysis.

Next, we turn our attention back to the Coefficients table to determine the strength and significance of each of our five variables. Income is no longer significant now that education and occupational prestige have been included in our analysis, suggesting that income in the prior regressions was really acting as a kind of proxy for education and/or occupational prestige (it is correlated with both, though not as strongly as they are correlated with one another). The other variables are all significant, age and being female at the $p<0.001$ level; education at the $p<0.01$ level; and occupational prestige is significant at the $p<0.05$ level. Age of respondent has a moderate negative (inverse) effect. Being female has a weak negative association, as do education and occupational prestige. In this analysis, age has the strongest effect, though the Betas for all the significant variables are pretty close in size to one another.

The B column provides the actual numerical effect of each independent variable, as well as the numbers for our regression equation. For every one year of additional age, time spent in the car each week goes down by about 3.2 minutes. Since income is not significant, we might want to ignore it; in any case, the effect is quite tiny, with even a $10,000 increase in income being associated with only a 2.4 minute decrease in time spent in the car. Being female is associated with a decrease of, on average, just over an

hour and a half (94.56 minutes). A one year increase in educational attainment is associated with a decrease of just under 17 minutes a week in the car, while a one-point increase in occupational prestige score[2] is associated with a decline of 24 minutes spent in the car per week. Our regression equation is $y = 16.240 - 0.053AGE^{***} - 4.286E-6REALINCOME - 1.567FEMALE^{***} - 0.279EDUC^{**} - 0.040PRESTG10^{*}$.

So, what have we learned from our regression analysis in this chapter? Adding more variables can result in a regression that better explains or predicts our dependent variable. And controlling for an additional independent variable can sometimes make an independent variable that looked like it had a relationship with our dependent variable become insignificant. Finally, remember that regression results are generalized average predictions, not some kind of universal truth. Our results suggest that folks who want to spend less time in the car might benefit from being older, being female, getting more education, and working in a high-prestige occupation. However, there are plenty of older females with graduate degrees working in high-prestige jobs who spend lots of time in the car–and there are plenty of young men with little education who hold low-prestige jobs and spend no time in the car at all.

Notes on Advanced Regression

Multivariate linear regression with dummy variables is the most advanced form of quantitative analysis covered in this text. However, there are a vast array of more advanced regression techniques for data analysts to use. All of these techniques are similar in some ways. All involve an overall significance, an overall strength using Pearson's r or a pseudo-R or R analog which is interpreted in somewhat similar ways, and a regression equation made up of various coefficients (standardized and unstandardized) that can be interpreted as to their significance, strength, and direction. However, they differ as to their details. While exploring all of those details is beyond the scope of this book, a brief introduction to **logistic regression** will help illuminate some of these details in at least one type of more advanced regression.

2. In the 2021 General Social Survey dataset, occupational prestige score ranges from 16 to 80 with a median of 47.

Logistic regression is a technique used when dependent variables are binary. Instead of estimating a best-fit line, it estimates a best-fit logistic curve, an example of which is shown in Figure 6. This curve is showing the odds that an outcome will be one versus the other of the two binary attributes of the variable in question. Thus, the coefficients that the regression analysis produces are themselves odds, which can be a bit trickier to interpret. Because of the different math for a logistic rather than a linear equation, logistic regression uses pseudo-R measures rather than Pearson's r. But logistic regression can tell us, just like linear regression can, about the significance, strength, and direction of the relationships we are interested in. And it lets us do this for binary dependent variables.

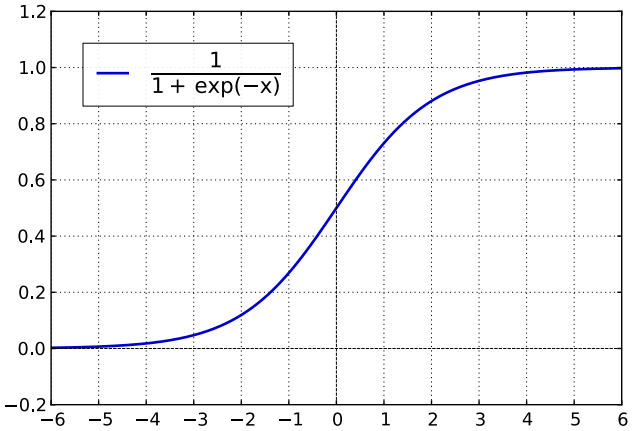


Figure 6. A Plot of a Logistic Function

Besides using different regression models, more advanced regression can also include **interaction terms**. Interaction terms are variables constructed by combining the effects of two (or more) variables so as to make it possible to see the combined effect of these variables together rather than looking at their effects one by one. For example, imagine you were doing an analysis of compensation paid to Hollywood stars and were interested in factors like age, gender, and number of prior star billings. Each of these variables undoubtedly has an impact on compensation. But many media commentators suggest that the effect of age is different for men than for women, with starring roles for women concentrated among the younger set. Thus, an interaction term that combined the effects of gender and age would make it more possible to uncover this type of situation.

There are many excellent texts, online resources, and courses on advanced regression. If you are thinking about continuing your education as a data analyst or pursuing a career in which data analysis skills are valuable, learning more about the various regression analysis techniques out there is a good way to start. But even if you do not learn more, the skills you have already developed will permit you to produce basic analyses–as well as to understand the more complex analyses presented in the research and professional literature in your academic field and your profession. For even more complex regressions still rely on the basic building blocks of significance, direction, and strength/effect size.

Exercises

1. Choose three continuous variables. Produce a scatterplot matrix and describe what you see. Are there

any reasons to suspect that your variable might not be appropriate for linear regression analysis? Are any of them overly correlated with one another?

2. Produce a multivariate linear regression using two of your continuous variables as independent variables and one as a dependent variable. Be sure to produce collinearity diagnostics. Answer the following questions:
    - Are there any collinearity problems with your regression? How do you know?
    - What is the significance of the entire regression?
    - What is the strength of the entire regression?
    - How much of the variance in your dependent variable is explained by the two independent variables combined?
    - For each independent variable:
        - What is the significance of that variable's relationship with the dependent variable?
        - What is the strength of that variable's relationship with the dependent variable?
        - What is the direction of that variable's relationship with the dependent variable?
        - What is the actual numerical effect that an increase of one in that variable would have on the dependent variable?
    - Which independent variable has the strongest relationship with the dependent variable?
3. Produce the regression equation for the regression you ran in response to Question 2.
4. Choose a discrete variable of interest that may be related to the same dependent variable you used for Question 2. Create one or more dummy variables from this variable (if it has only two categories, you can create only on dummy variable; if it has more than two categories, you may be able to create more than one dummy variable, but be sure you have left out at least one largeish category which will be the excluded category with no corresponding dummy variable). Using the Recode into Different function, create your dummy variable or variables. Run descriptive statistics on your new dummy variable or variables and explain what they show.
5. Run a regression with the two continuous variables from Question 2, the two dummy variables from Question 4, and one additional dummy or continuous variable as your independent variables and the same dependent variable as in Question 2.
6. Be sure to produce collinearity diagnostics. Answer the following questions:
    - Are there any collinearity problems with your regression? How do you know?
    - What is the significance of the entire regression?
    - What is the strength of the entire regression?
    - How much of the variance in your dependent variable is explained by the two independent variables combined?
    - For each independent variable[3]:
        - What is the significance of that variable's relationship with the dependent variable?
        - What is the strength of that variable's relationship with the dependent variable?
        - What is the direction of that variable's relationship with the dependent variable?
        - What is the actual numerical effect that an increase of one in that variable would have on the dependent variable?

3. Be sure to pay attention to the difference between dummy variables and continuous variables in interpreting your results.

    - Which independent variable has the strongest relationship with the dependent variable?

7. Produce the regression equation for the regression that you ran in response to Question 6.
8. Compare the R^2 for the regression you ran in response to Question 2 and the regression you ran in response to Question 6. Which one explains more of the variance in your dependent variable? How much more? Is the difference large enough to conclude that adding more additional variables helped explain more?

Media Attributions

- collinearity diagnostics menu © IBM SPSS is licensed under a All Rights Reserved license
- multivariate reg 1 © IBM SPSS is licensed under a All Rights Reserved license
- recode sex dummy © IBM SPSS is licensed under a All Rights Reserved license
- multivariate reg 2 © IBM SPSS is licensed under a All Rights Reserved license
- multivariate reg 3 © IBM SPSS is licensed under a All Rights Reserved license
- mplwp_logistic function © Geek3 is licensed under a CC BY (Attribution) license

SECTION V

QUALITATIVE AND MIXED METHODS DATA ANALYSIS WITH DEDOOSE

While researchers can and do analyze qualitative data by hand or with the use of basic computer software like word processing programs and spreadsheets, most qualitative projects involving a moderate to large volume of data today do rely on qualitative data analysis software packages. There are a variety of such packages, all with different strengths and limitations, and those who intend to perform qualitative data analysis regularly as part of their research or professional responsibilities should explore the options to find out which program is the best fit for their research style and priorities. This text features Dedoose. To get started with Dedoose, visit https://dedoose.com/ — the website has helpful guides, an explanation of pricing, and other resources. Users can sign up for an account at https://dedoose.com/signup (there are instructions about a student discount there as well) and can download the software at https://www.dedoose.com/resources/articledetail/dedoose-desktop-app for Windows, Mac, Chromebook, or Linux. Note there is both a regular installation and a "portable" option for Windows users who do not have administrative privileges. Unfortunately, Dedoose is **not** screenreader compliant at the time of this writing and has some limitations in terms of screen zoom applications.

The chapters here on how to use Dedoose include screenshots from a dataset created by students in the Sociology 404 class in Spring 2020, just before the onslaught of COVID. Students were asked to write a paragraph about their first day on campus here at Rhode Island College and answer a few questions about their graduation year, living arrangements, gender, major, and what they had been doing before coming to our campus. Students then did the work, collectively, of developing a code tree and coding the excerpts. Care has been taken to keep all of their participation, both their responses to the writing prompt and their work on the project, confidential, but let me take this moment to express my deepest appreciation for their excellent work and their contributions, which also helped inspire me to write this book.

24. Qualitative Data Analysis with Dedoose: Data Management

MIKAILA MARIEL LEMONIK ARTHUR

While researchers generally refer to the software they use to facilitate qualitative research as qualitative data analysis software, such software programs also play a very important role in data management and data reduction. Indeed, without employing the data management capabilities of qualitative data analysis software, the software itself is unlikely to be functional. Thus, before we start analyzing our data using tools like Dedoose, we need to feed the data we have collected into the tool of our choice and take various steps to set it up to be usable. This chapter will provide an overview of how to get started with a project in Dedoose and add data to the project, as well as how to view and manipulate the data. This chapter assumes you have already created a user account with Dedoose and downloaded and installed the software, and that you can successfully log into your account in the program. Note that if you run into technical difficulties while using Dedoose, their support can be reached at support@dedoose.com.

Getting Started With a New Project

The first step in getting started with Dedoose is to either create a new project or be added to an existing one. This chapter will assume you are starting with a new project; if you are working with an existing project, someone who has administrative privileges on the existing project will need to add you to it. "Projects," in Dedoose's terminology, are workspaces that store complete collections of data from a particular research study.

To create a new project, first click on "projects" on the menu bar in the top right corner of the screen, as shown in Figure 1.

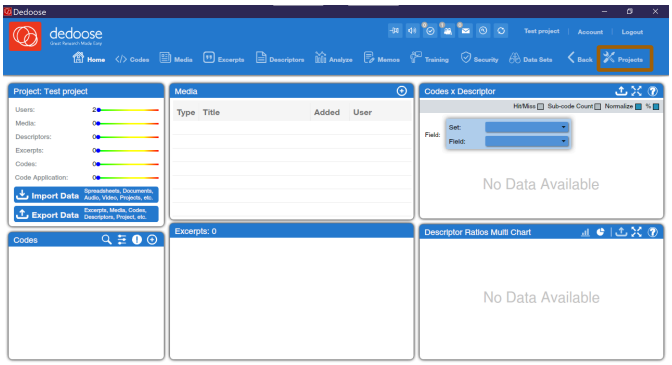


Figure 1. Finding the Projects Button on the Dedoose Home Screen

This will bring up a window that lists all of the projects your account has access to. If you are a new Dedoose user, you will likely have far fewer projects in your account than the examples here will show; as you develop additional projects, they will be added to your list. Next, click on the "Create Project" button at the bottom of the screen.

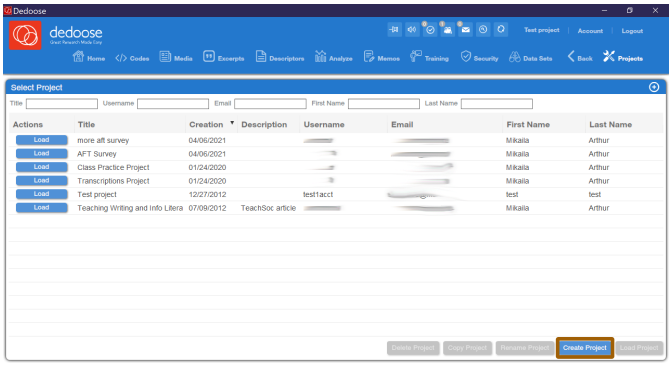


Figure 2. Creating a New Project in Dedoose

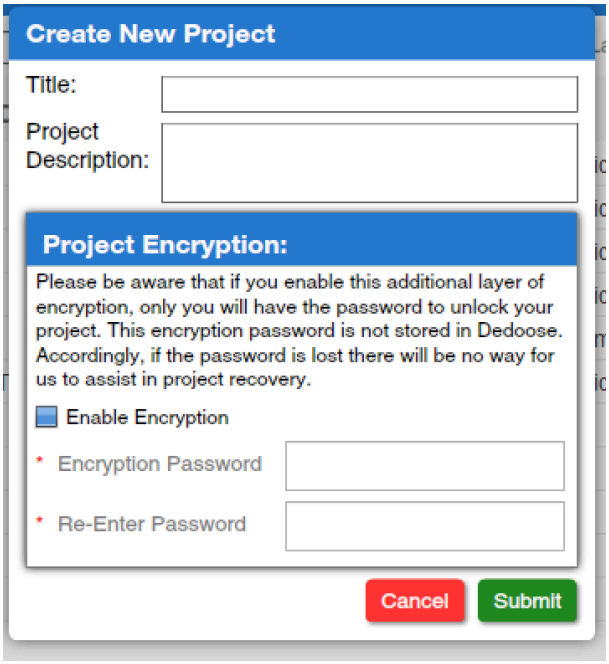


Figure 3. Popup Window for Creating a New Project

The final step in creating a project is to set up that project, using the "create project" popup window shown in Figure 3, by giving it a title (this should be short but descriptive, so you can remember which project is which) and a brief description of the project. If you wish, you can assign the project an encrypted password, but note that this makes it impossible for Dedoose to help you recover the project. Once you have set up the project, click submit.

Once you have created your project, you may need to load it by first selecting it from the list of projects in the project window, then clicking the "load" button next to the project name, as shown in Figure 4. You will see that the information you provided in setting up your project now appears in the list of projects. The project screen also provides buttons for deleting, renaming, and copying projects.

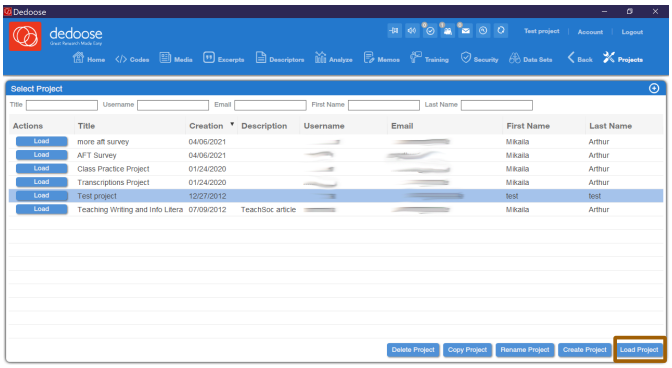


Figure 4. Loading a Project

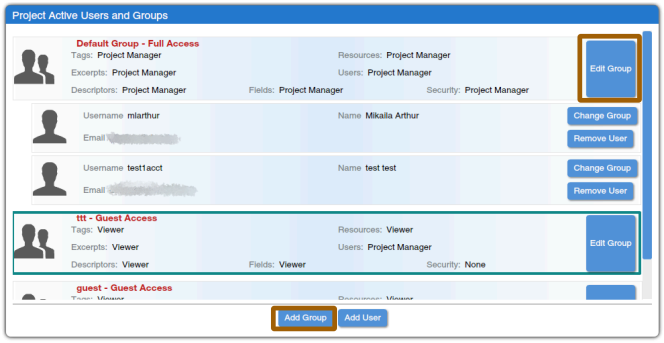


Figure 6. Dedoose Security Center

Figure 5. Security Icon

If you would like to add another user to your project, like your professor, a classmate, or a research collaborator, click on "security" in the top menu. Of course, if you are planning on working alone, you do not need to add anyone. The security center, shown in Figure 6, is not always the most intuitive feature of Dedoose, but luckily you do not need to use it too often.

The first step in the Security Center is to use the Add Group button to add a new group, or the Edit Group button to edit an existing group. A "group," in this context, refers to a set of users who have similar security privileges—for instance, you might have a project director with full privileges to do anything with the project, two lead researchers who can use most tools in the project but cannot add other users, five research assistants who can code and use analysis tools but cannot delete anything, and an intern who can view the project but not edit it. Each of these types of people—project director, lead researcher, research assistant, and intern—would be a group, and the security center enables you to set specific security privileges for them.

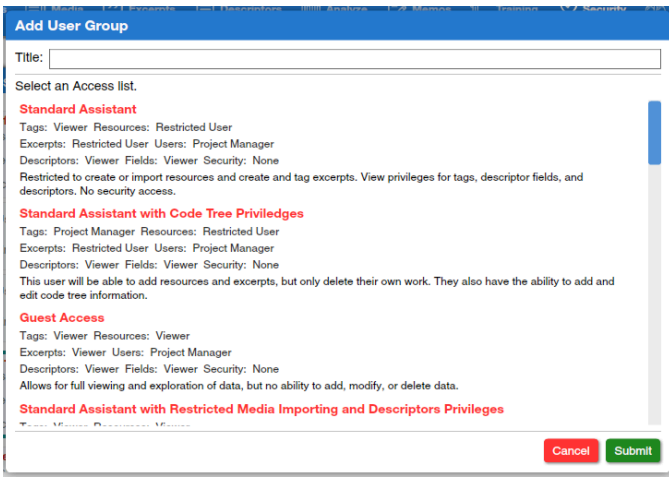


Figure 7. Security Privileges

When you click on Add Group or Edit Group, a popup window opens, as shown in Figure 7, with a list of the various types of security privileges that are available. These vary in terms of the extent to which users in those groups can use various features, ranging from "Full Access" for our project manager to "Guest Access" for our intern. It may take a little while to explore the options and select the right one for your project. When you have done so, select that option and click the green submit button.

To add users to the user group you have now defined, use the "add user" button (next to the "add group" button highlighted in Fig-

ure 6. First, you will need to indicate which of the security groups you have defined this user should be added to. The system will then ask you for the user's email address, and depending on whether the user is already an active Dedoose user or not will ask you whether you want to invite them to Dedoose, add them to your account, or just add them to the project.

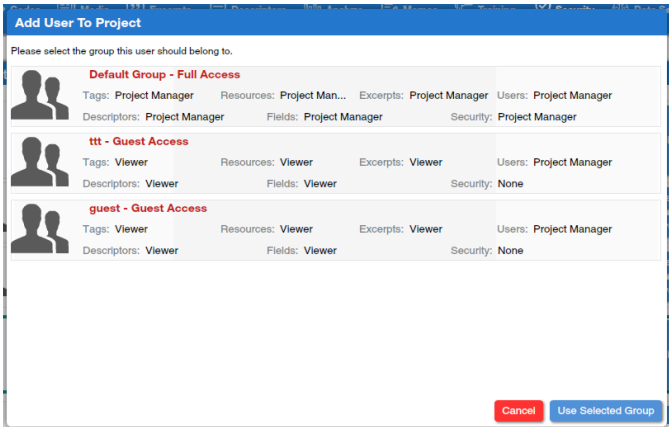


Figure 8. Adding a User

Once you have finished working with the security center—or, if this is an individual project and thus you did not need to use the security center—your project should be ready for you to start working.

Working With Data

The first step in the qualitative data analysis project, as discussed in the chapter on Preparing & Managing Qualitative Data, is to prepare your data for analysis. In Dedoose, a key part of this process is adding your data to the application. To do this, click on the plus sign in a circle at the top of the "Media" box on the home screen, as shown in Figure 9.

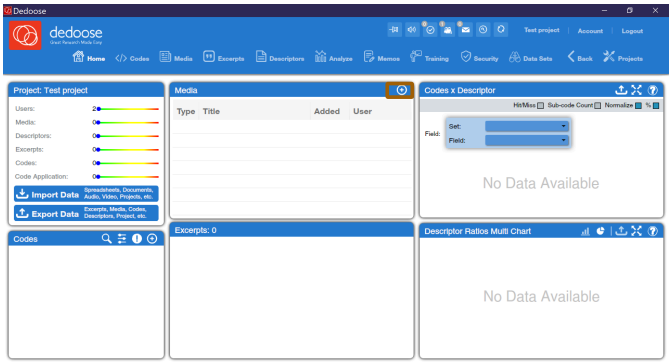


Figure 9. Opening the Dialogue for Adding Media

This will open a pop-up window with a variety of options for importing media, as shown in Figure 10, depending on how your data is stored. On the most basic level, you can create a blank document and type or copy-paste content into Dedoose, but that becomes cumbersome with more than a small volume of data. If you have stored fieldnotes in document (*.doc, *.docx, *.rtf, *.txt, *.htm, or *.html) or in PDF format, you can choose the "Import Text" or "Import PDFs" options, respectively. These options will allow you to select one or more files and have them all imported at once. Using this option is especially handy when you have a separate, single file for each interview, respondent, or case, as Dedoose will then store each file as a separate instance of data in ways that facilitate analysis. It is also possible to import image files (*.jpg, *.png, *.bmp, or *.gif), which is a helpful option for visual sociology. Dedoose can handle audio and video files but at additional cost, and discussing these features is beyond the scope of this text. Finally, you can import a spreadsheet. However, this method of importation requires careful spreadsheet construction and formatting. The Dedoose User Guide (scroll down to "Survey Importer") provides the necessary details about how to format your file.

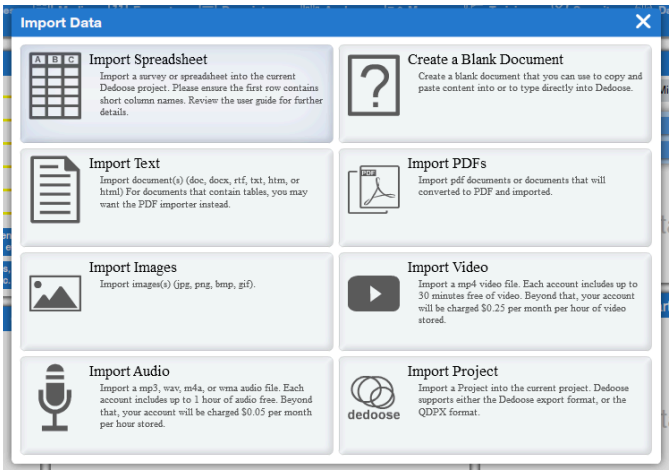


Figure 10. Methods of Importing Data into Dedoose

Figure 11. Media Icon

Once your data is imported, you can view and manipulate it using the Media tools by clicking on the Media icon at the top of the Dedoose screen. The media icon takes users to a window that displays a list of all documents, graphics, or other texts that are part of a project, with their title, the user that imported them, the date they were imported, their length, and several other features that will be discussed later on in this text (including descriptors, memos, and excerpts), as shown in Figure 12.

If you click on any media item, Dedoose opens that particular media item, as shown in Figure 13.

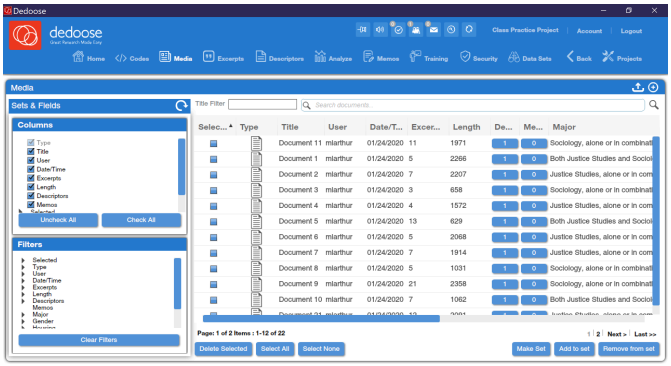


Figure 12. Dedoose Media Tool

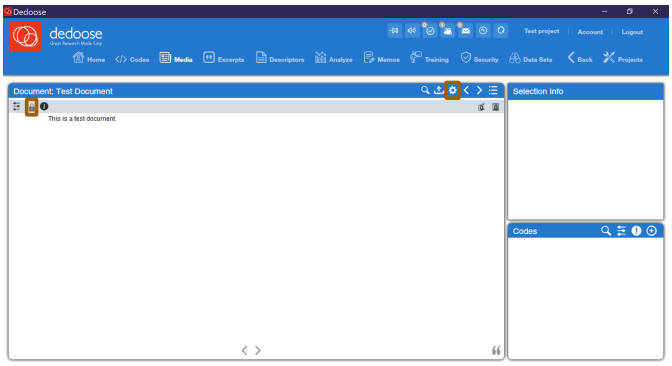


Figure 13. Media Editing Window

By clicking on the gearshift highlighted in Figure 13, you can edit the title or description of a media item. And by clicking on the lock icon highlighted in Figure 13, you can edit the text in the media item itself. Other features of the media window will be discussed later, after those tools are introduced.

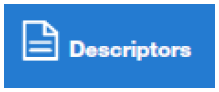


Figure 14. Descriptors Icon

There are two additional important features that are part of setting up your data in Dedoose. The first is called "**Descriptors**." Descriptors are category labels that apply to an entire text or piece of media or to its author or creator—they are not codes, which are applied to specific segments of text. For instance, if we were keeping track of the gender, age, or race of the author of a narrative or the participant in an interview, that would be

a descriptor. Similarly, if our study involved collecting advertisements, the magazine, website, or television program on which the advertisement appeared might be a descriptor. Descriptors are created and edited in Dedoose by using the Descriptors tool in the toolbar at the top of the program.

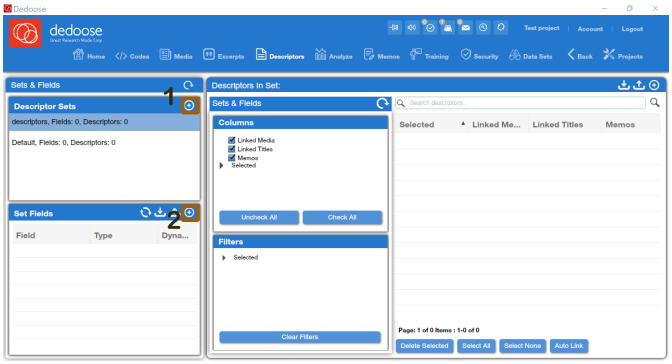


Figure 15. The Descriptors Window

To create descriptors, load the descriptors tool using the icon shown in Figure 14, and then click on the plus sign in the circle highlighted in Figure 15 next to the number 1, in the Descriptor Sets section of the screen. Give your descriptor set a name (the name does not matter, it is just part of the data storage system) and click submit. Then, click on the plus sign in the circle highlighted in Figure 15 next to the number 2. This will bring up a window that permits you to develop a set of descriptors.

For each descriptor you wish to add, as shown in Figure 16, you can provide a name (something like "gender" or "age" or "magazine appeared in") and a longer description explaining the descriptor so you can remember what you did. Then you can choose a field type from four options: an option list, which is basically the same as a multiple-choice question; a number, which permits the free entry of any numerical value; a date; or a text field, in which any short text can be entered. There is an option for dynamic fields, which are those where data might change over time, as in a longitudinal study, but we will leave those aside for the purposes of this discussion. If you select option list for field type, you can then use "Add Options" under Field Options to add each of your multiple-choice options to the list. When you are done, click submit. For example, Figure 16 shows an option list descriptor called "Housing," used in a study of student life, in which respondents indicated whether they lived with their family, lived in on-campus housing, or lived off-campus but not with their family. You can use the X icon to delete an option from the list, the icon that looks like a piece of paper to edit the option, and the ∧ and ∨ icons to reorder your options.

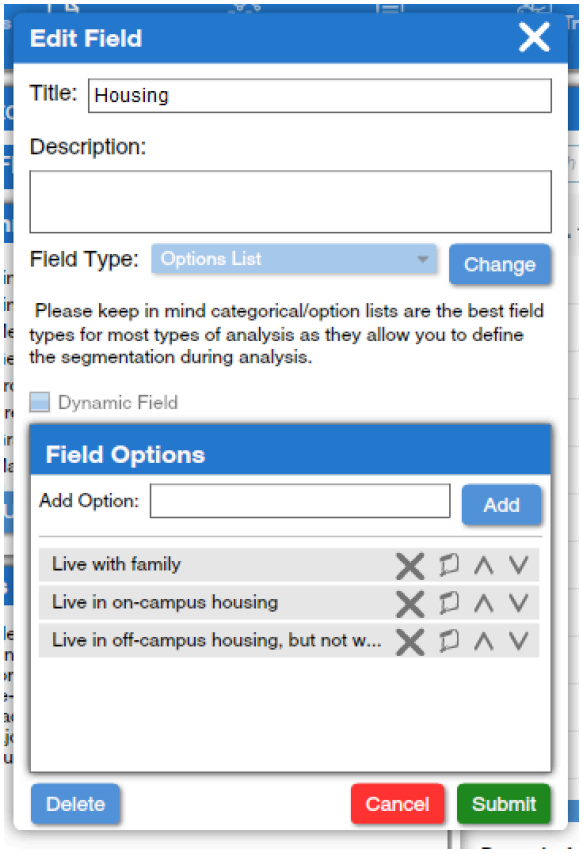


Figure 16. Adding or Editing Descriptor Fields

The next step in using descriptors is to link your descriptors to your media. There are several ways to do this. You can go to the Media window and click the blue box under the Descriptors heading, or you can use the tiny (and hard to see) descriptor icon on an individual media item. Both options are shown in Figure 17.

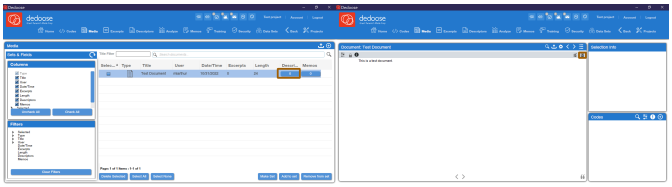


Figure 17. Adding Descriptors to Media

Clicking on either option will bring up a popup window called Descriptor Links. Then click "Create and Link Descriptor," as shown in Figure 18.

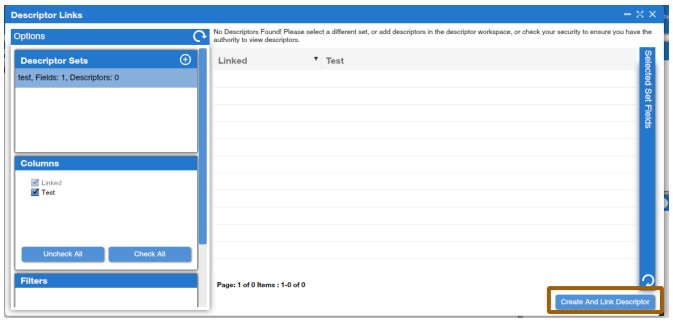


Figure 18. Linking Descriptors to Media

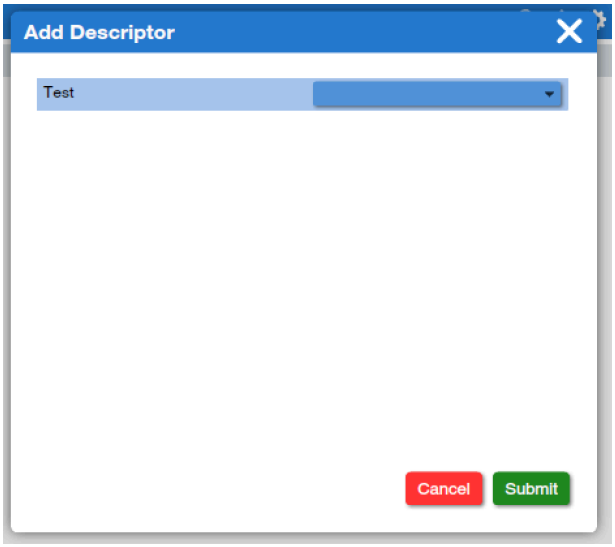


Figure 19. Editing Descriptors for a Particular Media Item

This will bring up a pop-up window in which you can select from drop-down menus (for option lists) or enter (for other types of descriptor fields) as shown in Figure 19 to apply descriptors to the selected media item. You may have far more descriptors than are shown here; however many there are, select the appropriate options for each one, and then click submit.

You will need to do this individually for each media item in your project. Once you have done this, if you return to the Media tool, you will be able to preview all of your media items and their linked descriptors. In addition, if you return to the Descriptors tool, you will see all of your descriptors, but without the associated media.

The final tool you should know about as you get started with Dedoose is the memo tool. To create a memo, first open a specific media item, then click on the memo icon, as shown in Figure 20.

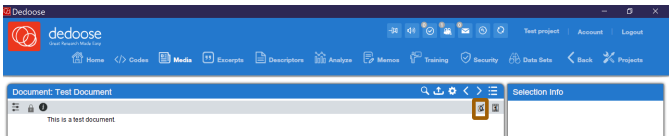


Figure 20. Opening the Memo Tool

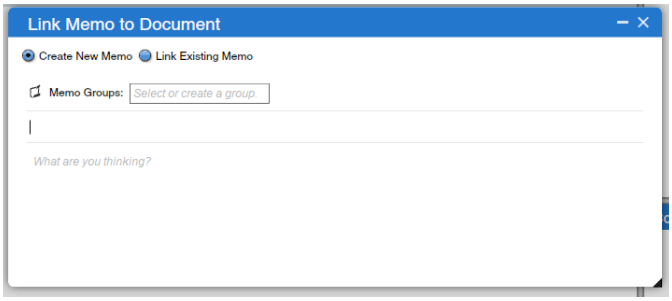


Figure 21. Adding a Memo

This will open a pop-up window, as shown in Figure 21, in which you can create a memo. You should enter a title, and then you can type or copy-paste your text where the screen says "What are you thinking?" Using the memo groups box at the top, you can also create a memo group or add a memo to an existing group, if you have many memos and want to classify or categorize them. Once you begin typing, a "Save" button will appear at the bottom of the memo screen—be sure to save when you are done.

Once you have created one or more memos, you can use the memos icon, as shown in Figure 22, to load a window that shows all of your memos and allows you to work with them.

Figure 22. Memo Icon

Backing Up Your Data & Managing Dedoose

As Dedoose uses cloud-based storage to keep your data, you do not need to save—all data is automatically saved. However, you may wish to download your data, either to back it up, to keep a local copy, or to import it into a different software package in the future. You can use the Export button on the Home screen, as shown in Figure 23, to export all or a portion of a project.

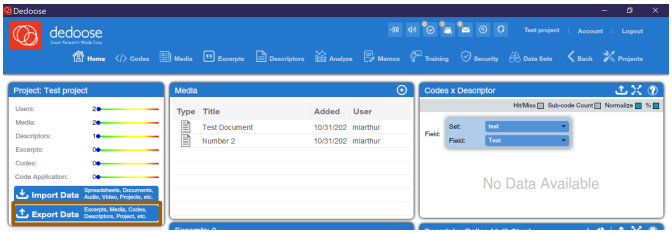


Figure 23. The Export Button

The export button brings up a pop-up window, as shown in Figure 24, with a variety of options. You can export just the codes, descriptors, media information with linked descriptors, or excerpts to a spreadsheet by selecting the option of your choice. Excerpts, as well as memos, can be exported to a document. Keep in mind that none of these options involves exporting the full volume of original media—Dedoose strongly encourages you to keep your original media, in the form it was prior to uploading to Dedoose, intact and backed up outside of Dedoose. If you have the media plus these exports, you can load data into other programs or applications. There is also an option to export the entire project, but this type of file is hard to use outside of Dedoose itself, so it best serves as a backup of your work.

Figure 24. Export Options Window

If Dedoose gets a little slow or nonresponsive, you may wish to log out of the application, close it, and then reopen and log back in. There is also a refresh icon (◈)[1] at the top of the screen that can be helpful if you are working on a project with another researcher and want to be sure their changes have loaded into your view.

Exercises

1. Create a project in Dedoose. If you are doing this work as part of a class, give your instructor access to the project.
2. Download five oral histories from the COVID-19 Archive (you can do this at https://covid-19archive.org/s/oralhistory/item). Import them into your project.
3. Create a descriptor set with at least three descriptors you find relevant to the oral histories you selected. Link the descriptor set to your oral history media, being sure to correctly select any options.
4. Read one of the oral histories you downloaded. Write a memo of at least 250 words summariz-

1. This icon might not display correctly in some versions of this text. Click here to see what it looks like.

ing the most important insights in that oral history, and add your memo to that media item in Dedoose.

Media Attributions

- Dedoose Home Screen with Projects Button © Dedoose is licensed under a All Rights Reserved license
- Creating a New Project © Dedoose is licensed under a All Rights Reserved license
- Creating a Project Popup Window © Dedoose is licensed under a All Rights Reserved license
- Loading a Project © Dedoose is licensed under a All Rights Reserved license
- The Security Icon © Dedoose is licensed under a All Rights Reserved license
- The Security Center © Dedoose is licensed under a All Rights Reserved license
- Security Privileges Popup © Dedoose is licensed under a All Rights Reserved license
- Adding User to Project © Dedoose is licensed under a All Rights Reserved license
- Adding Media © Dedoose is licensed under a All Rights Reserved license
- Importing Data Popup Window © Dedoose is licensed under a All Rights Reserved license
- The Media Icon © Dedoose is licensed under a All Rights Reserved license
- The Dedoose Media Tool © Dedoose is licensed under a All Rights Reserved license
- Editing Media © Dedoose is licensed under a All Rights Reserved license
- The Descriptors Icon © Dedoose is licensed under a All Rights Reserved license
- The Descriptors Window © Dedoose is licensed under a All Rights Reserved license
- Editing Descriptor Fields © Dedoose is licensed under a All Rights Reserved license
- Adding Descriptors to Media © Dedoose is licensed under a All Rights Reserved license
- Descriptor Links © Dedoose is licensed under a All Rights Reserved license
- Editing Descriptors © Dedoose
- Opening the Memo Tool © Dedoose is licensed under a All Rights Reserved license
- Adding a Memo © Dedoose is licensed under a All Rights Reserved license
- The Memo Icon © Dedoose is licensed under a All Rights Reserved license
- The Export Button © Dedoose is licensed under a All Rights Reserved license
- Export Options Window © Dedoose is licensed under a All Rights Reserved license

25. Qualitative Data Analysis with Dedoose: Coding

MIKAILA MARIEL LEMONIK ARTHUR

The process of **coding** in Dedoose begins with the creation of a **code tree**. Once the code tree is created, then codes can be applied to segments of text or other media. This chapter will step users through the process of creating a code tree and applying codes, as well as other issues in working with codes and coding in Dedoose.

The Code Tree

In order to begin coding, analysts must first develop a code tree, as discussed in the chapter on Qualitative Coding. Once the code tree has been developed, it can be added to Dedoose so that it is ready to use. Codes can be added using the Codes section of the Dedoose home page or by selecting "Codes" from the menu at the top of the screen, and then clicking the ⊕ (plus sign in a circle) symbol.

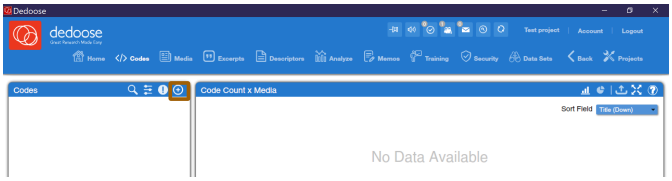


Figure 1. The Code Screen in Dedoose

Clicking the ⊕ icon brings up a pop-up window into which users can enter a code. The window asks for a title, which is the term used to identify that code in the list of codes. This title should be clear and brief. For example, rather than writing "An instance of the use of stereotyping by the respondent" you would write "stereotyping."

The description box provides space to enter a longer explanation of the code and the circumstances in which it should be used, which is especially useful when coding with multiple coders or over a longer period of time so everyone can keep track of the meanings behind codes. Users can also change the color of the code (note that using this feature also requires changing a setting, which will be discussed below). Finally, if code weighting is part of the project, users can enable it and specify minimum, maximum, and default weights. Weighting is used when analysts wish to not only code the text, but also indicate the degree to which a code applies. For example, if a study involved codes representing a variety of emotions—happy, sad, angry, excited, satisfied, etc.—weighting could be used to distinguish pleased from ecstatic and gloomy from devastated by applying a 1 to the more mild emotion, a 2 for a moderate emotion, and a 3 for a more intense emotion. While the rest of this text will proceed without involving code weighting, you may wish to explore its use in projects you are completing.

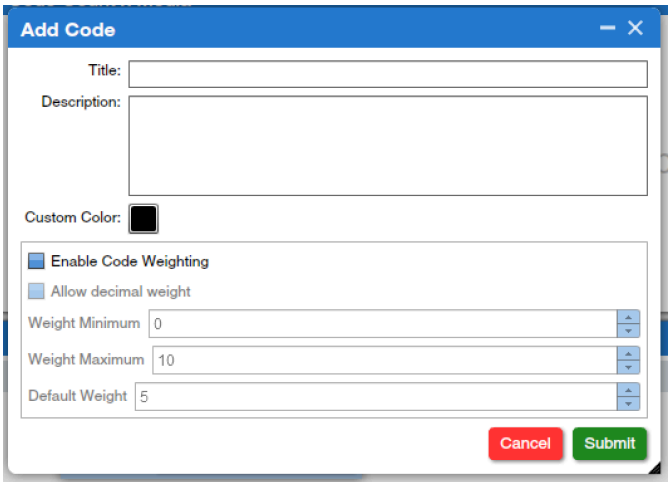


Figure 2. New Code Window

In many cases, analysts develop code trees that have multiple levels of nested codes. For instance, in the example above of emotion codes, the code tree might look something like this:

```
Emotions
        Happy
                Pleased
                Joyous
                Ecstatic
        Sad
                Gloomy
                Dejected
                Devastated
```

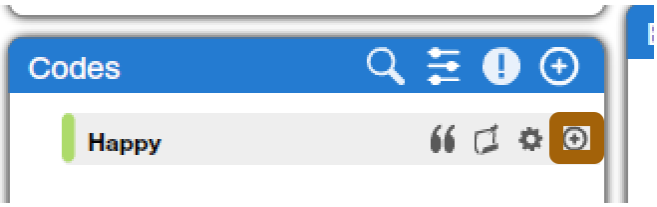


Figure 3. Adding a Child Code

To set the code tree up in this way, Dedoose uses the language of "parent codes" (those at the top or first level of the tree, like emotions in this example) and "child codes" (those at lower levels of the tree). First, analysts need to enter a parent code into the coding system, as shown above. Then, they can use the ⊕ symbol next to that code to add a child code, using just the same dialog box shown in Figure 2 above. Figure 3 also shows an icon of a gear shift, next to the ⊕ symbol—this gear shift can be used to re-open the dialog box for editing the code.

Once codes are added, the code tree will look something like the one shown in Figure 4. The little triangles can be used to open and close parent codes, making child codes visible or hiding them. Note that there can be multiple levels of codes, so an analyst could add additional child codes under, say, anxiety or confusion.

The magnifying glass at the top of the codes window can be used to search all of the codes used in a project, which can be helpful when the code tree gets very lengthy. There are also two other icons at the top of the code tree, one that looks like little slider bars and one that looks like an exclamation point in a circle. The sliders allow the analyst to set options for how coding will proceed.

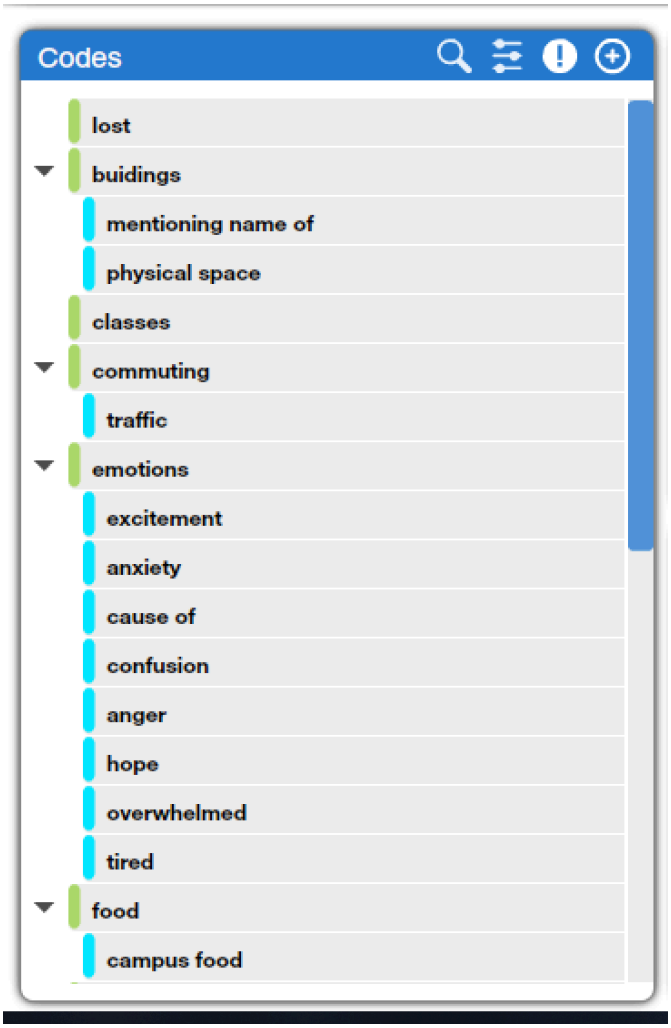


Figure 4. A Complete Code Tree in Dedoose

- Automatic upcoding: When automatic upcoding is turned on, any time that a child code is used while coding, the parent code will also be applied to the same segment of text.
- Sort alphabetically: Just as it sounds, this option reorders codes in alphabetical order, which can make it easier to find them in a lengthy code tree.
- Code counts: The code counts option displays the number of times each code has

been used in a project right in the code tree next to the code itself. There are two ways to implement code counts. In the first, "explicit code count," just instances in which the code itself has been used are counted, while in the second, "child sum code count," the sum of all uses of all child codes is displayed next to each parent code.
- Color scheme: Changing the color scheme to "custom" allows for the use of colors designated in the process of adding codes to be used in the display.

The exclamation point icon provides a number of useful tools:

- Collapse/Expand: This tool is the equivalent of going through and clicking all of the little black triangles one at a time—when clicked, it toggles the code tree between having all parent codes closed, such that child codes are hidden, and having all parent codes open, such that all child codes are visible.
- Retroactive upcode: This tool is used when, having not turned on "automatic upcoding" (as discussed above) at the beginning of a coding process, the analyst decides later that they would like the parent code applied to all instances where the child code is used.
- Reorder codes: This tool allows the analyst to reorganize the code tree into a different order.
- Import codes: This tool permits for the importation of codes from a Microsoft Excel or comma-separated file.
- Export codes: This tool permits the analyst to export codes to Microsoft Excel or Microsoft Word.

Once all settings and options have been set to the analysts' preference and the code tree has been added, it is time to start coding. Note that it is possible to change settings and add codes during the coding process. However, it is very important that, if a new code is added during the coding process, the analyst goes back to all texts that have already been coded and re-codes them. Otherwise, that new code will be used for only part of the dataset, which will introduce errors into the data analysis process.

Coding in Dedoose

In order to apply codes to texts, the first step is to create an excerpt. To create an excerpt, load a media item, highlight a segment of text to which one or more codes should be applied, and click the quotation mark in the corner of the document screen, as shown in

Figure 5. If you have made a mistake in your selection, you can click the X next to the quotation mark in the “Selection Info” box to delete it.

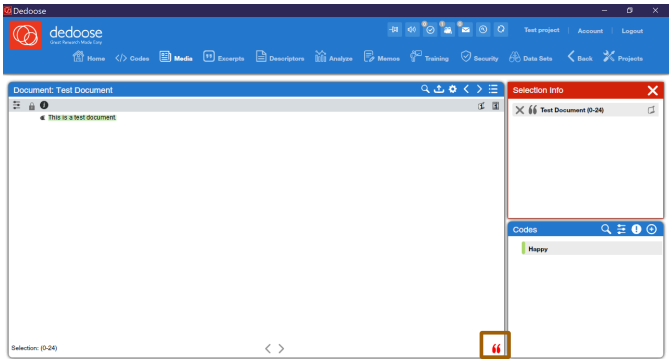


Figure 5. Creating an Excerpt

Once you have created an excerpt, you can then apply codes by either dragging each individual code from the “codes” box to the “selection info” box or by double-clicking on the code in the “codes” box. If you want to remove a particular code you have added to an excerpt, just click the X next to that code in the “selection info” box. When you are done applying codes to a given excerpt, click the X next to “selection info” to exit the editing mode and move on to create your next code. If you want to re-open a particular excerpt, you can click on the black bracket next to the excerpt, and this will permit you to add additional codes or delete the excerpt. When you are done with a given text, you can use the < and > icons at the bottom of the screen to move on to the next text.

Figure 6 provides an example of what it might look like after a complete (short) text is coded. You can see how each excerpt appears highlighted in color, with a black bracket in the margin. One excerpt is currently selected, and the “selection info” box shows the codes that the coder applied to that excerpt. Do note the typical length of the excerpts—when selecting an excerpt, analysts should strive to select a unit of text that represents a complete idea or utterance, whether that complete idea/utterance is just a few words or whether it is a paragraph or more in length.

Working with Codes

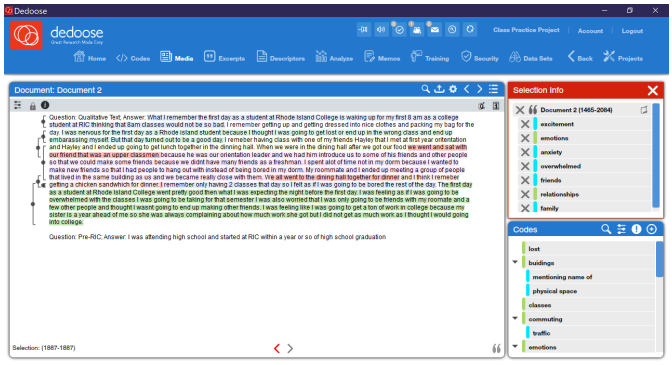


Figure 6. A Coded Text

While the process of coding primarily involves moving through each text, creating excerpts, and applying relevant codes (and sometimes repeating this process as the code tree evolves or additional texts are added to the dataset), Dedoose does offer some tools that can further enhance the coding process.

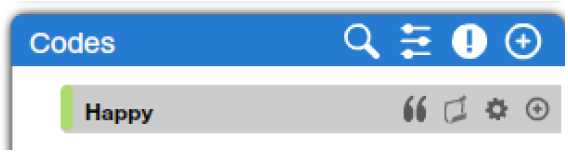


Figure 7. Code Tools

In the chapter on Data Management with Dedoose, you learned about linking memos to texts. But what if the memo you write is less connected to a given text and instead is generated by observations made while applying a particular code? In that case, you may wish to link a memo to a code. The icon that looks a bit like a scroll of paper next to the code, as shown in Figure 7—the one after the quotation mark icon—allows analysts to link memos to individual codes.

The quotation mark icon brings up a window, as shown in Figure 8, that includes all of the excerpts to which a given code has been applied. The "view text excerpts in full" button shows the complete text of each excerpt and all codes that have been applied to it. You can also export all of the excerpts. If you double-click on a specific excerpt, you can copy and paste the text of that excerpt, which is useful when you need to include a quote in the paper or talk you are preparing. After double-clicking, there is also a "view in context" button, which loads the text in question in the background such that after you close the various pop-up windows you will be able to view it.

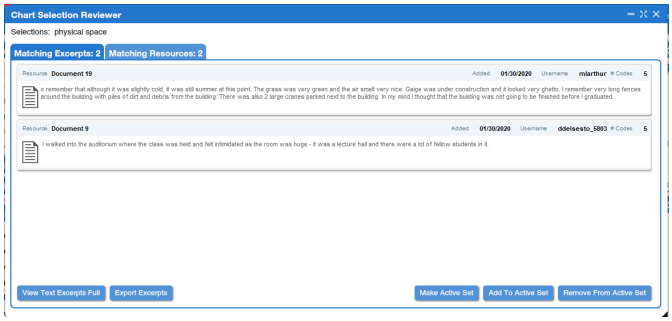


Figure 8. Excerpt Viewer Tool

Similar information is available by clicking on the "Excerpts" tool at the top of the screen, as shown in Figure 9. The excerpts tool brings up a list of all the excerpts in a given project and shows, for each one, which text it is part of, when it was created, who created it, how long it is, how many codes were applied to it, which codes, and—if applicable—any memos or descriptors. The list of excerpts can be sorted or filtered by any of these columns, and double-clicking on any row will bring up the specific excerpt in a pop-up window.

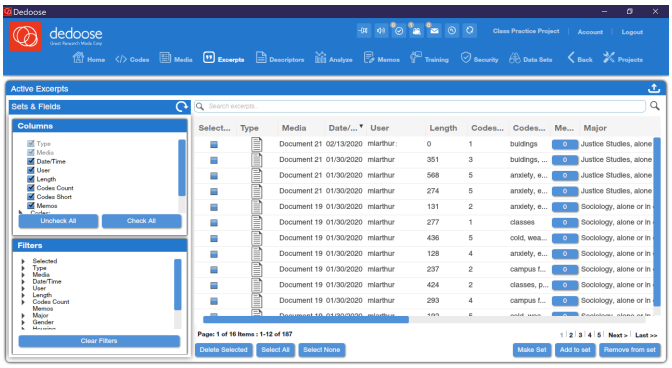


Figure 9. The Excerpts Window in Dedoose

A final tool worth noting for those who are coding in teams is the Training Center. While it is beyond the scope of this text to detail the workings of the Training Center, it is designed to help coding teams enhance their interrater reliability. In short, team leads can select a variety of excerpts and codes as part of a coding test that all members of a coding team then take. After coders complete the test, they and their team leads can see an overall interrater reliability score comparing their coding to the codes applied by the initial coder to those excerpts selected for the test, and they can also delve deeper by looking at agreement and disagreement rates and **Kappa** scores for each individual code. Reports can be excerpted, and team members can also view specific code applications to see how their coding compares to the initial coding seeding the test.

A final note: the analysis tools in Dedoose do rely on completed coding, so finish coding all your texts before delving into analysis.

Exercises

1. Return to the five oral history transcripts you selected in the exercises for the chapter Qualitative Data Analysis with Dedoose: Data Management. Read through the transcripts and develop a code tree including at least five parent codes and additional child codes as relevant. Enter the code tree into Dedoose.
2. Choose one of your transcripts and code that transcript completely.
3. Write a memo focusing on what the transcript you coded tells you in relation to one or two of the codes you selected.

Media Attributions

- code screen © Dedoose is licensed under a All Rights Reserved license
- new code window © Dedoose is licensed under a All Rights Reserved license
- add child code © Dedoose is licensed under a All Rights Reserved license

- code tree © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- create excerpt © Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- example coding © Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- code tools © Dedoose is licensed under a All Rights Reserved license
- view excerpts © Dedoose is licensed under a All Rights Reserved license
- excerpts window © Dedoose is licensed under a All Rights Reserved license

26. Qualitative Data Analysis with Dedoose: Developing Findings

MIKAILA MARIEL LEMONIK ARTHUR

When working with very small datasets, analysts may find it sufficient to use Dedoose to code texts and then use the coding tools to just explore each code. But Dedoose offers a variety of tools for digging more deeply into the data, exploring relationships, and moving towards findings. In this chapter, we will review the each of the analysis tools, showing what each tool can provide and how to work with the analysis tools in developing findings.

Using the Analysis Tools

To access the analysis tools in Dedoose, click on the "Analyze" button in the toolbar at the top of the screen, as shown in Figure 1. This will bring up the Analyze window. In this window, the sidebar contains a list of all of the analysis tools that Dedoose provides, categorized by type. The main window displays the results after a particular analysis tool is selected.

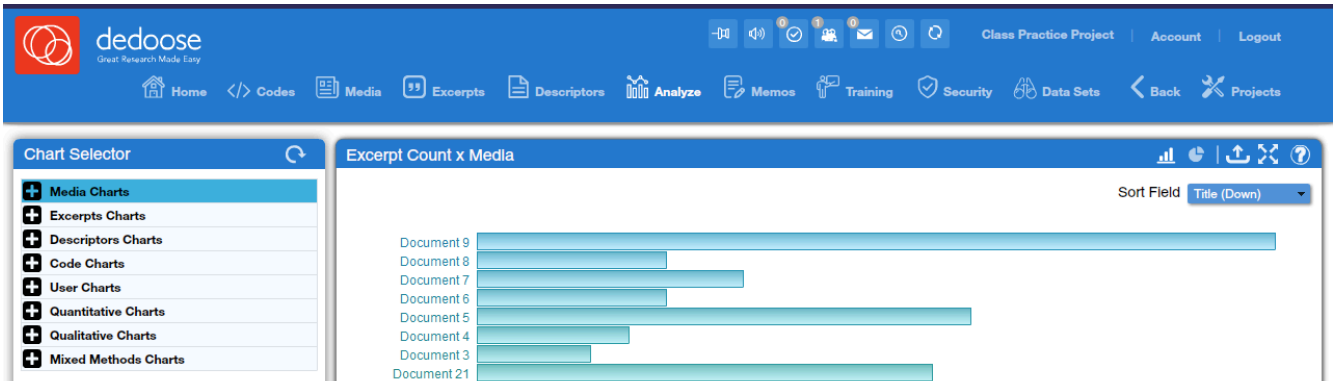


Figure 1. The Analyze Window in Dedoose

In the top corner of the screen, you will observe a variety of icons for interacting with the results of the selected analysis tool. Note that not all of these icons are available for all tools—for instance, the icons that let users switch between bar graph and pie chart views are only available for tools where the results are displayed in bar graphs and pie charts. Next to the pie chart you will see a tool with an up arrow; this tool is used to export results. Most

results are exported in Microsoft Excel format (which can be easily viewed in Google Sheets and other spreadsheet programs and includes both data and visuals), though in one or two cases a PDF file is produced. Sometimes there are options users can select when determining how to export their results. After that, the icon with four arrows is used to enter full-screen view, and the icon with the question mark is used to access very brief information about the currently-selected tool. Many tools will also provide specific options for formatting or data processing, and these will be explained along with the explanation of each tool.

Before getting into the specific tools, it is also important to note that tools provide direct access to relevant excerpts. While tools are loaded, you can typically put your cursor over any data element to bring up a popup with more detailed information. Then, if you click on the data element, you will bring up a window that provides all excerpts that meet the given criteria specified by the data element you have clicked on. For instance, if an analyst clicked on one of the bars shown in the bar graph in Figure 1 (which represent texts), they would be taken to a window showing all of the excerpts in the text they clicked on. Clicking on an excerpt then brings up more detail about that specific excerpt in a new window, and the text of that excerpt can then be selected, copied, and pasted into a working document when quotes are desired.

The Dedoose Analysis Toolkit

Below, each of the analysis tools in Dedoose will be explored. There are a few more advanced tools that will only be touched upon briefly. Note that many tools can be found under multiple tool categories in the sidebar, but provide the same information regardless of which category the tool has been selected under.

Exploring Media and Users

Dedoose offers a few tools that are rather limited in terms of analytical power but that do offer some useful ways to explore the texts that are part of a project and to track the work of multiple coders on the project. The first two tools discussed in this section are for exploring texts, while the final three are for looking at the work of coders.

Excerpt Count x Media (found under Media Charts, Excerpt Charts, and Quantitative Charts) provides an overview of how many excerpts have been created in each text. This is the tool shown in Figure 1 above. Clicking on the bar representing any given text provides a window with all of the excerpts from that text. The dropdown menu in the corner provides the option of changing the sort order.

Code Count x Media (found under Media Charts, Code Charts, and Quantitative Charts) shows how many codes were applied to each text, displaying a bar graph (which can be changed to a pie chart) as shown in Figure 2. This bar graph is often quite similar to the one found under Excerpt Count x Media, except the numbers are typically higher as more than one code is typically applied per excerpt. Similarly, clicking on a bar brings up all relevant excerpts, and the dropdown menu in the corner provides the option of changing the sort order.

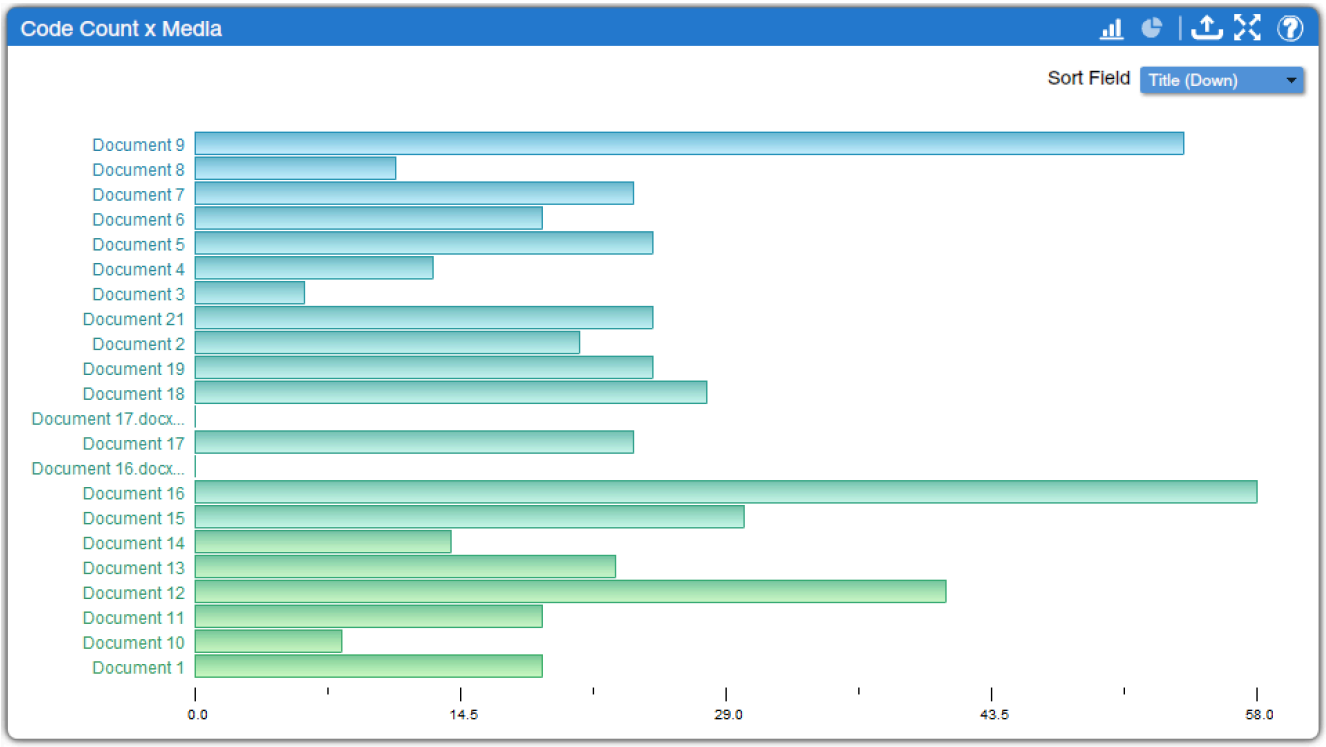


Figure 2. Code Count x Media Tool

User Excerpts (found under Excerpt Charts, User Charts, and Quantitative Charts) and **User Code Application** (found under Code Charts, User Charts, and Quantitative Charts) provide bar graphs similar to those discussed above, except instead of displaying the number of excerpts and codes by text, they display the number of excerpts and codes by user. These tools, then, can be a useful way of tracking engagement with a project when multiple coders are working together. **User Media** (found under Media Charts, User Charts, and Quantitative Charts) provides a bar graph showing how many media were uploaded by each user. Except in large projects with many researchers, this tool is less likely to be useful.

Code Tools

The tools that are most important for the standard forms of qualitative data analysis discussed in this text are among the code tools. These are a set of tools that allow analysts to explore what they can learn from the way they have coded within their project. The most basic of these is **Code Presence** (found under Media Charts, Code Charts, and Qualitative Charts), which simply displays whether or not a particular code has been applied at least once to a given text. As shown in Figure 3, this tool provides a grid or matrix in which the entire code tree developed in the project is arrayed across the top, while the document titles are listed down the side. When a code appears in a particular document or text, a red square with the numeral 1 marks the intersection; when the code does not appear, the intersecting cell is empty. If you click on the red square, a window will pop up with all of the excerpts from that text to which that code was applied.

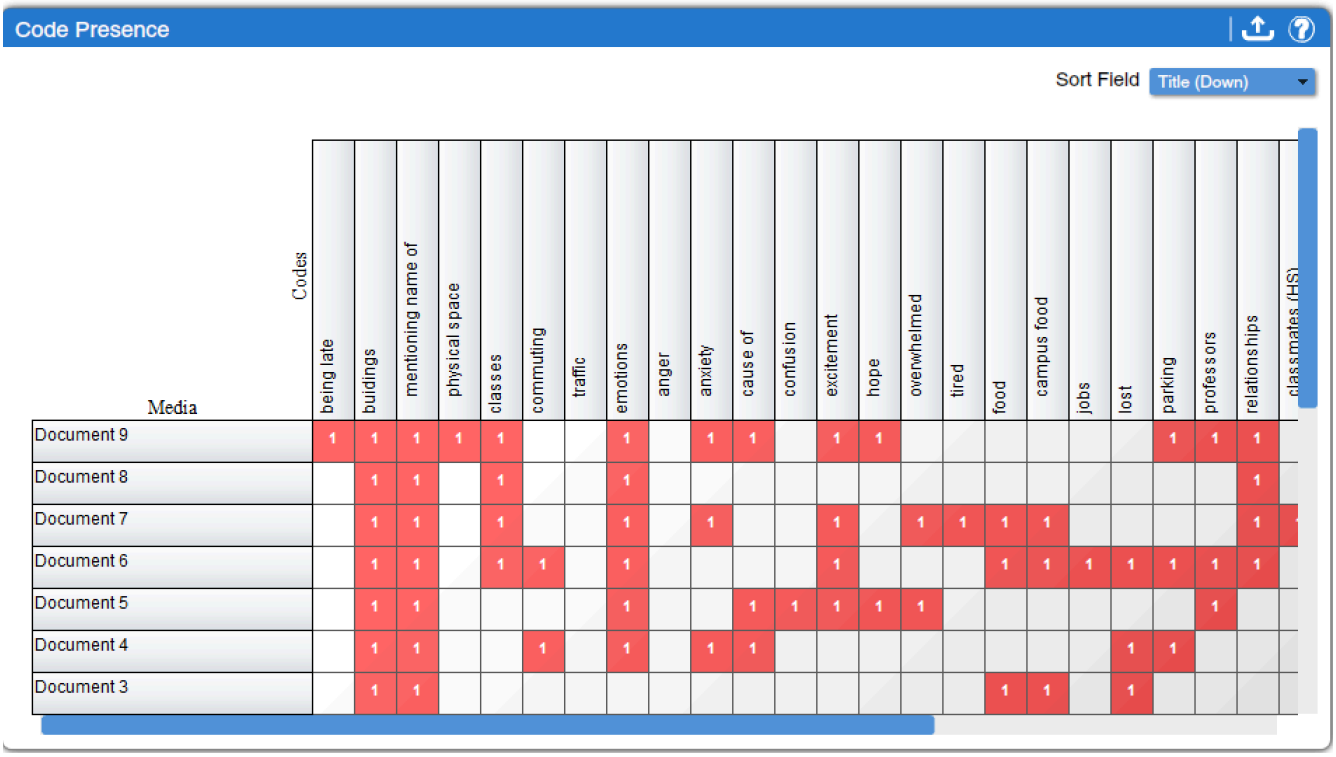


Figure 3. Code Presence Tool

Code Application (found under Media Charts, Code Charts, and Qualitative Charts) is a somewhat more useful way to view the same data. In this tool, instead of just displaying the presence or absence of each code in each text, the number of times each code was applied to each text is displayed, as shown in Figure 4. Color coding helps users quickly spot the most frequently used codes, which are displayed in orange and red (while rarely used codes

are displayed in blue) with the number of times they were applied indicated. As in the case of other tools, clicking on a table cell brings up all applicable excerpts.

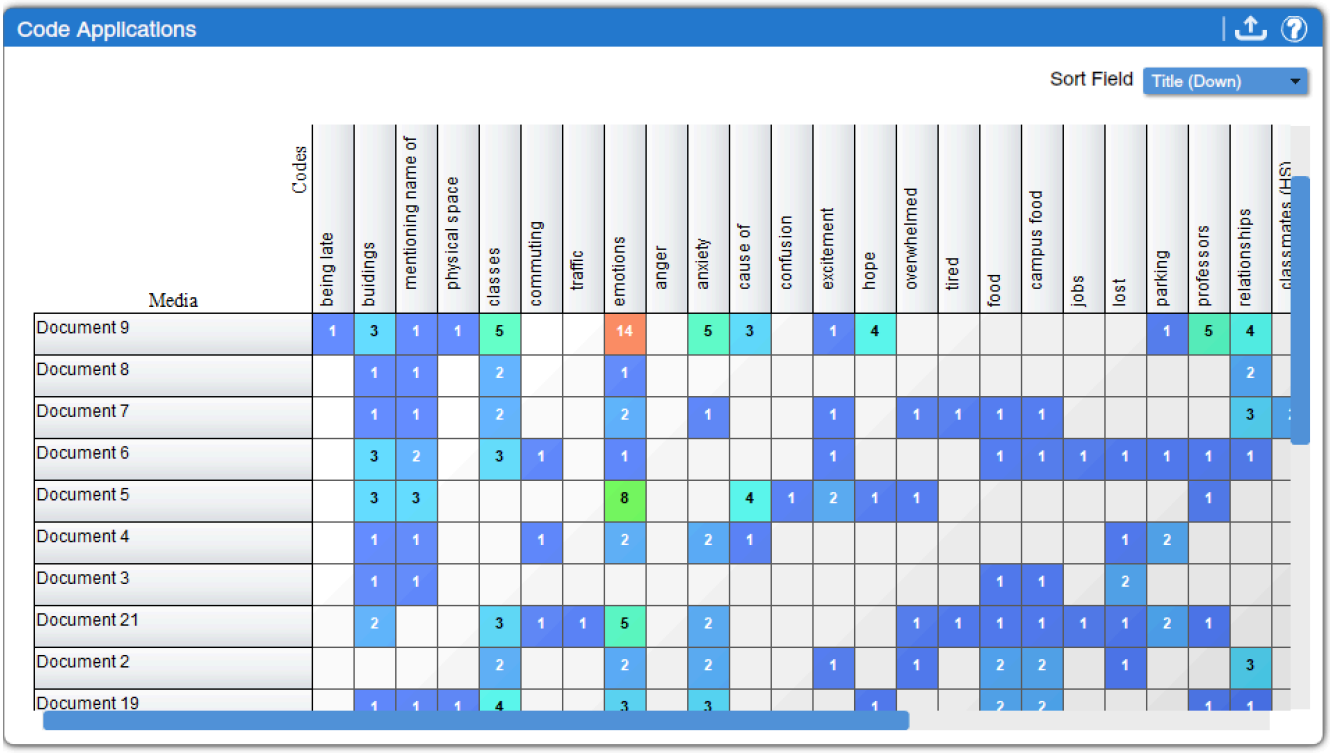


Figure 4. Code Application Tool

Code Co-Occurrence (found under Code Charts and Qualitative Charts) is arguably the most useful of the code tools. Rather than simply documenting the presence or extent of given codes in particular texts, the Code Co-Occurrence tool lets analysts explore the relationships between codes. It flags and tallies excerpts in which the same codes appear. For example, in investigating the sample data displayed in Figure 5, we can observe that "emotions" tends to co-occur with relationships and classes, and anxiety is a particular emotion frequently occurring in relation to discussions of classes. The same color scheme as noted above in relation to the Code Application tool helps viewers see, at a glance, which codes co-occur most frequently, and clicking on table cells brings up relevant excerpts. The checkbox in the corner toggles whether overlapping excerpts—or multiple excerpts that have some sections of text in common—are included. As is the case with other tools, the resulting chart can be exported to a spreadsheet format, though excerpts are not included in the export.

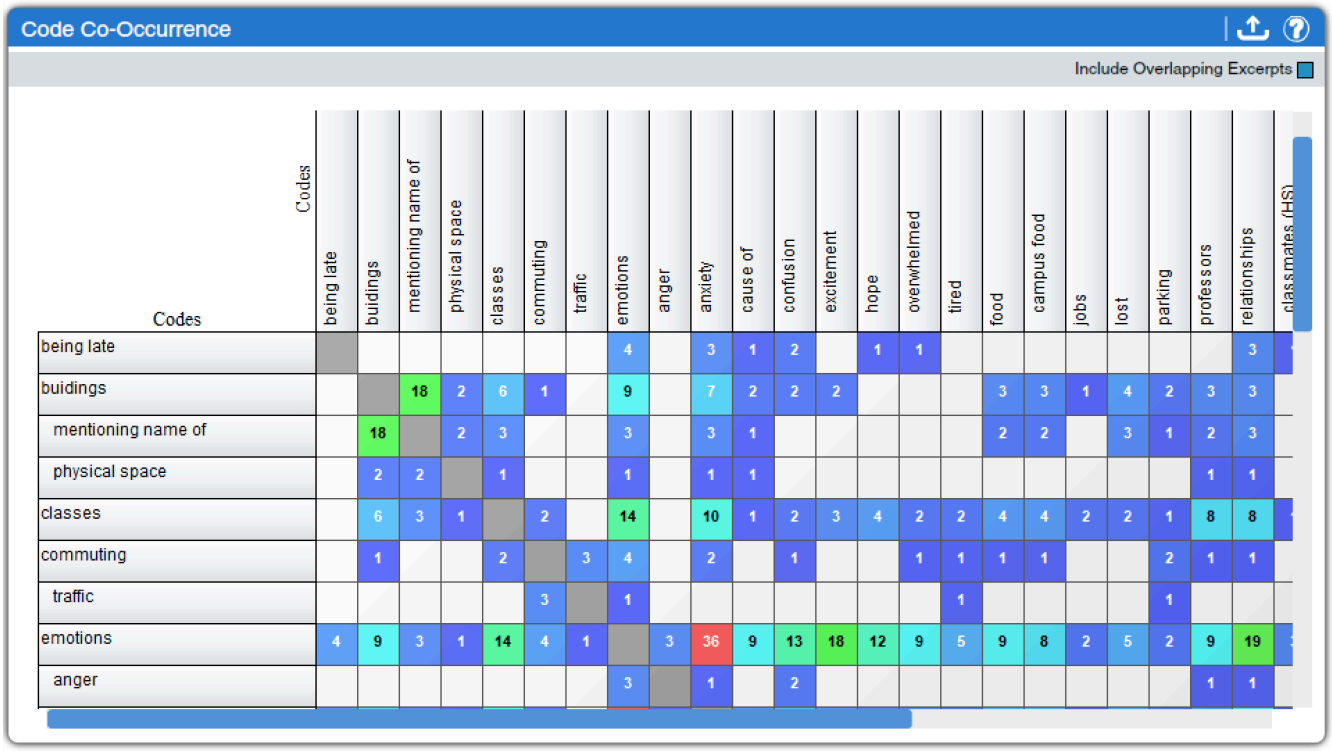


Figure 5. The Code Co-Occurrence Tool

Less useful, but perhaps more fun, are two tools that produce code clouds. Code clouds are a type of **word clouds** used specifically to display the frequency at which a given code has been applied within the body of text in a project. In code clouds, codes that have been applied more frequently are displayed in larger and bolder text than codes that are used less frequently. These tools can be used to get an at-a-glance sense of which codes are most predominant in the text, and can also be useful in creating visuals to use in presentations when discussing coding. The **Packed Code Cloud** (found under Code Charts and Qualitative Charts), as shown in Figure 4, provides a static visual which can be modified in a variety of ways. The "Sub-code Count" checkbox toggles whether or not child codes are included in the counts driving the sizes of the codes. Under the "Colors" drop-down menu, analysts can choose from the default color scheme shown in Figure 4 or color schemes that are more blue, red/yellow/orange, or pastel in color. The "Layout" drop-down menu offers a fast scheme or one that places each code into the visual one at a time. The "Direction" drop-down menu offers options for how horizontal or vertical the display is, as well as an option called "Wiggly" for a more dynamic diagonal display of terms. Finally, the "Redraw" button refreshes the display once options have been changed; even if options aren't changed, slightly different presentations will occur each time Redraw is hit, so that the analyst can chose the one that is most visually appealing. Hovering the mouse over a code provides the

number of times that code was applied in the project, and clicking on it brings up a window with all instances of the code. The export tool creates a PDF of the code cloud, which is necessary if you need a print-quality version of the resulting image, though for including in a presentation or other digital document many users might prefer to take a screenshot. Note that the Packed Code Cloud is most useful as a visual representation of coding data—to actually carry out data analysis, other tools will be more helpful.

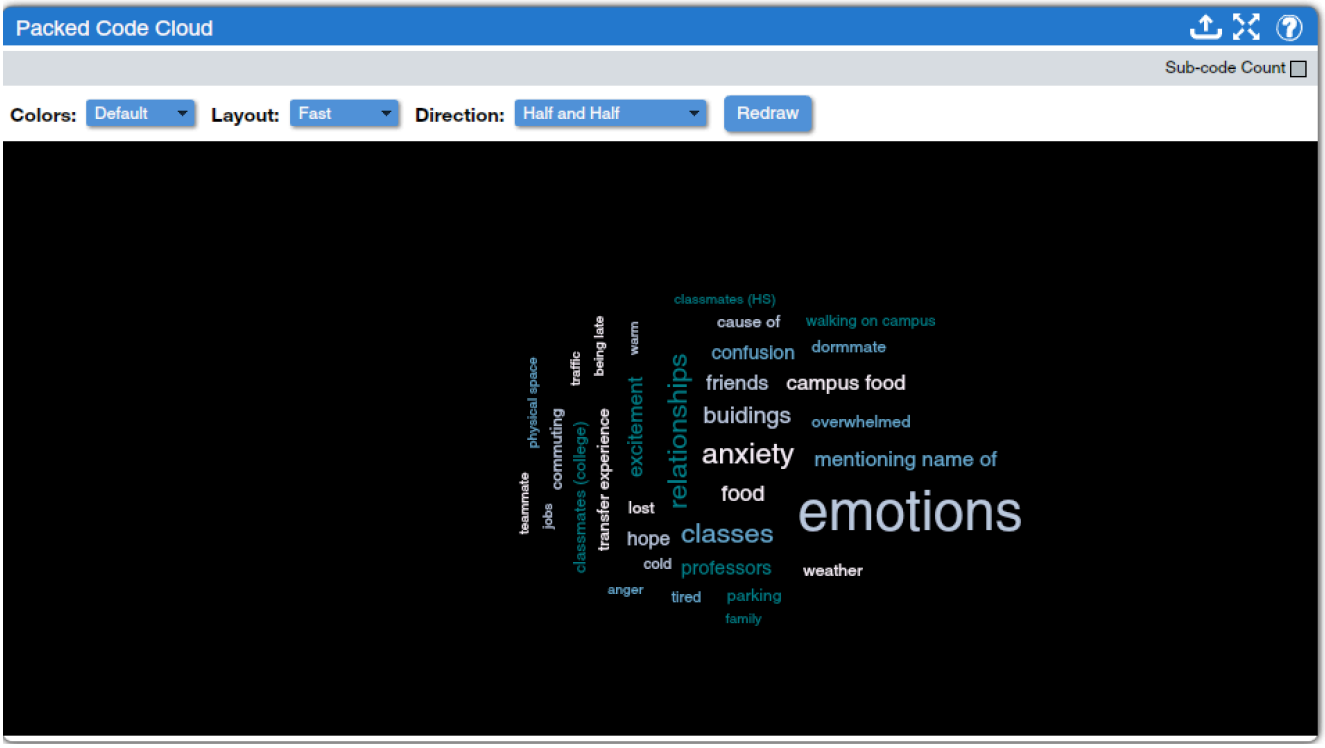


Figure 6. Packed Code Cloud Tool

Similarly, the **3D Code Cloud** (found under Code Charts and Qualitative Charts) presents word cloud data, but in a simulated three-dimensional format, as shown in the (silent) video clip below. A checkbox toggles the inclusion or exclusion of sub-codes, while sliders on the side of the window allow users to adjust the zoom and the minimum frequency of code applications for inclusion. Note that the 3D Code Cloud tool does not provide an export option, so users will need to have a screen recorder in order to use these visualizations elsewhere. Just as in the Packed Code Cloud tool, users can click on individual codes to bring up matching excerpts. However, as not all codes can be clearly seen at once, the 3D Code Cloud is even less useful as an analytical tool.[1]

1. For screenreader users: there is a screencapture video of this tool below; it is silent. Tab to move through toggle buttons and options; the screenreader is unable to read the moving graphic shown.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://pressbooks.ric.edu/socialdataanalysis/?p=62#video-62-1

Tools for both Descriptors and Codes

An additional set of tools is designed to help analysts investigate the relationships between codes and descriptors. These tools would permit, for example, an investigation of gender or racial differences in how respondents discuss a particular topic, or an analysis of whether different age groups use different emotion words when talking about their pets. All of the tools provide similar basic information, but display it differently or permit deeper dives.

Codes x Descriptor (found under Descriptors Charts, Code Charts, and Mixed Methods Charts, as well as under the Codes tab in Dedoose) creates bar graphs that show how often a code is applied to texts with a given descriptor, as shown in Figure 7. All codes in the code tree are shown in the visualization, and a box at the top of the window with two drop-down menus lets the analyst select the descriptor they wish to investigate (and the descriptor set in which that descriptor appears, if more than one descriptor set is in use in a given project). Other options, shown in the grey bar at the top of the window, allow for configuration of the results:

- Hit/Miss: when selected, the display will show the number of texts with a given descriptor to which a particular code is applied. When unselected, the display will show the total number of times a particular code is applied to texts with a given descriptor. This option cannot be selected at the same time as the Normalize option.
- Sub-code Count: as in other tools, this toggles on or off the inclusion of child codes when parent codes are presented.
- Normalize: this option applies a mathematical calculation to the figures presented in the tool to adjust them in light of the overall number of texts with a given descriptor. For instance, if a dataset had 23 nurses and 5 doctors, it might not be reasonable to just examine how many times nurses versus doctors discussed status at work—there are so many more nurses that their figures would just seem inflated. Normalizing helps correct for this.
- %: toggles between displaying data as counts (raw numbers) and percentages.

As in other tools, clicking a bar in one of the graphs brings up a window with relevant excerpts.

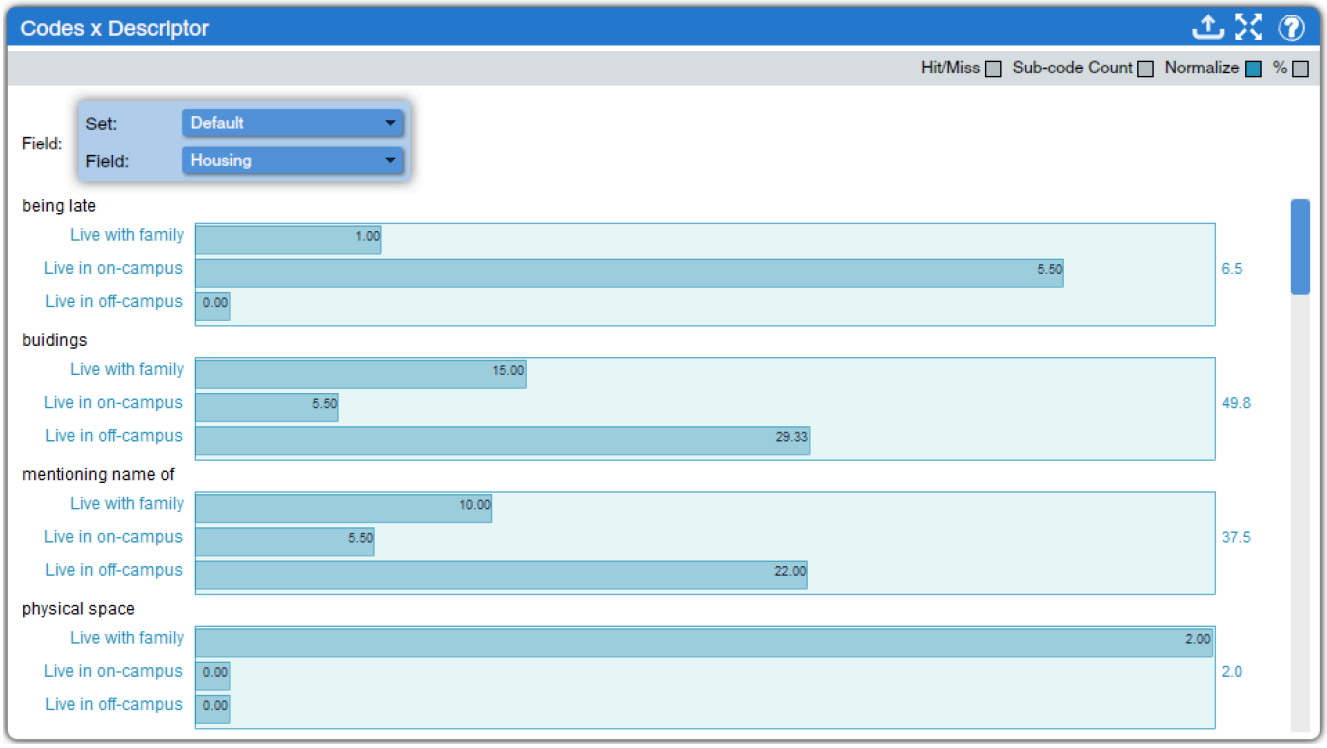


Figure 7. The Code x Descriptor Tool

Descriptor x Code (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) is, in a way, the inverse of Code x Descriptor. Here, all descriptors employed in the project are displayed in the window, and a drop-down menu permits the analyst to select which code they wish to investigate. The same options for adjusting the output are provided. In the example displayed in Figure 8, for instance, we can see that only those texts produced by individuals who were recently in high school are coded as involving high school classmates, though not too many texts discuss high school classmates at all.

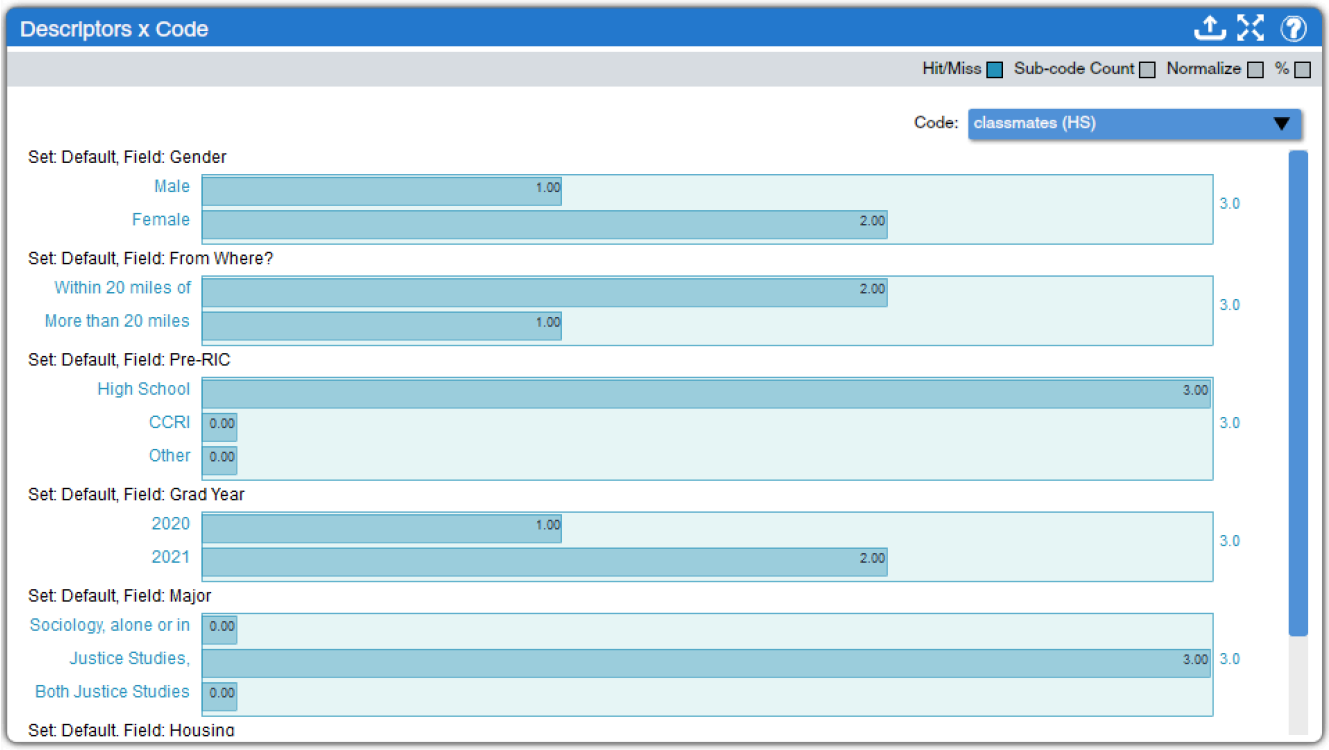


Figure 8. The Descriptor x Code Tool

The next two tools provide an overview of code applications by descriptor. Both have the same options and the same basic format, but they provide slightly different data. The **Descriptor x Code Case Count Table** (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) shows the number of excerpts across all texts with a given descriptor to which a particular code has been applied. In contrast, the **Descriptor x Code Count Table** (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) shows the number of texts with a given descriptor to which a particular code has been applied. Thus, it is unsurprising that the numbers are generally higher in the former than in the latter. Which is more appropriate to use depends on the research question and goals of a given project. Figure 9 shows the Descriptor x Code Count Table; the Descriptor x Code Case Count Table would look similar but with higher numbers in many cells.

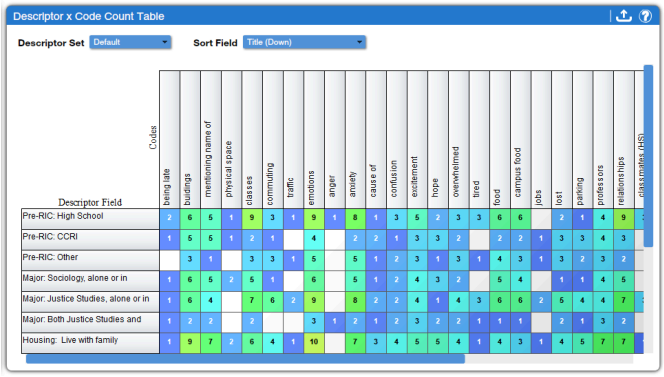


Figure 9. The Descriptor x Code Count Table

Descriptor x Descriptor x Code (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) provides a way to look at the relationship between two descriptors and a code. Analysts choose two descriptors, and the tool produces a set of bar graphs, one graph for each category of the first descriptor with one bar in each graph for each category of the second descriptor. Then analysts choose a code, and the applications of that code determine the length of each bar. For instance, in Figure 10, we can see the relationship between gender, living arrangements, and the application of the anxiety code. The results show that anxiety is discussed most frequently among male respondents who live on campus, but among female respondents who live off campus but not with family. Note that switching which descriptor is in Field 1 and which is in Field 2 does change the results, especially when the normalize option is switched on as this makes the data quite susceptible to alteration based on small numbers of texts with a given descriptor. Thus, it is essential that analysts think carefully about their research question and how to set up any Descriptor x Descriptor X Code analysis to address that question. This tool can be used with code weights in projects that have applied them. Options for the inclusion of child codes, for normalizing figures, and for toggling between percentages and raw numbers are also available.

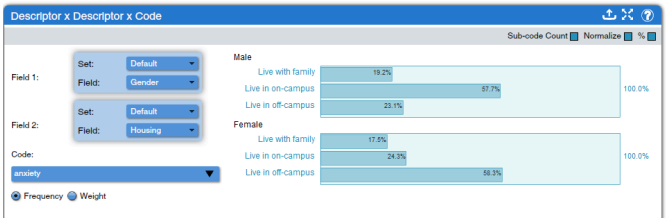


Figure 10. The Descriptor x Descriptor x Code Tool

Code Frequency Descriptor Bubble Plot (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) creates a visual display incorporating data from three codes and one descriptor. The frequency of applications of one code is displayed on the X axis, of a second code on the Y axis, and of a third code in the size of the bubble. Note that rearranging which code is in the X axis, Y axis, or size drop-down box will alter the display, so analysts should think carefully about how they wish to set up their display. Then, the descriptor selected from the Field drop-down box creates different bubbles, with a color key, for each category of the selected descriptor. For example, in Figure 11, we can see a plot looking at anxiety, being lost, and talking about classes, with the descriptor of where students were prior to their first day on my campus: high school, the local community college, or another college. The results show that anxiety is most prevalent in the discussions of students starting directly from high school, who are also more likely to talk about classes. In contrast, students coming from the local community college wrote little about anxiety, but a lot about getting lost.

As in other tools, including sub-code counts and normalizing can be toggled on and off; clicking a bubble will bring up a list of included excerpts, and results can be exported, in this case as both spreadsheet and PDF files.

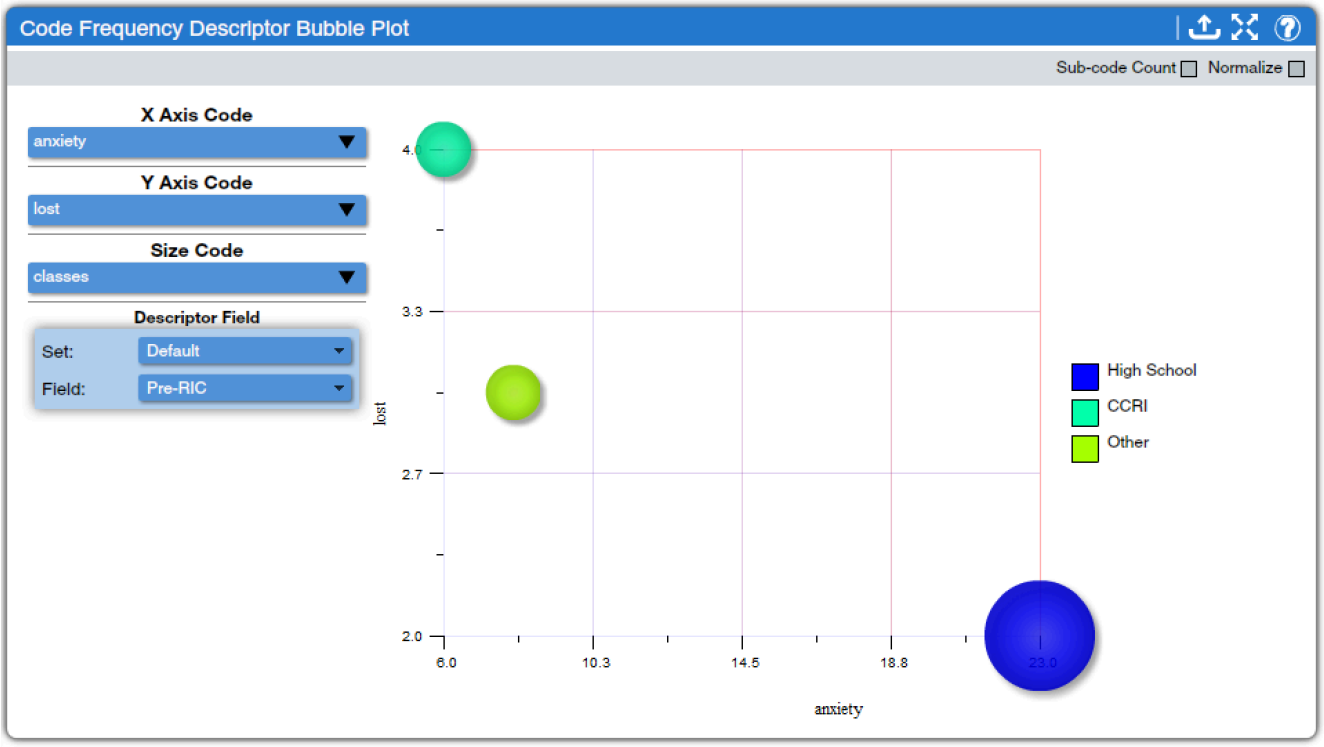


Figure 11. The Code Frequency Descriptor Bubble Plot

Descriptor Fields by Codes Grid Chart (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) is a particularly flexible tool that lets analysts look at various combinations of descriptors and codes. First, analysts select as many descriptor fields as they wish to include from the Field drop-down menu, clicking "Add Field" after each one to add it to the list of Descriptor Fields in the top corner. Then, they select the checkboxes next to the codes they wish to include from the Codes list. Counts or weights can be displayed in projects that use weights. These selections generate a grid chart that shows all possible combinations of the selected descriptor categories and the number of code applications of each selected code in texts with those combinations of descriptors. For instance, in Figure 12, we can see that female students coming directly from high school made more statements that were coded with Anxiety than did other groups.

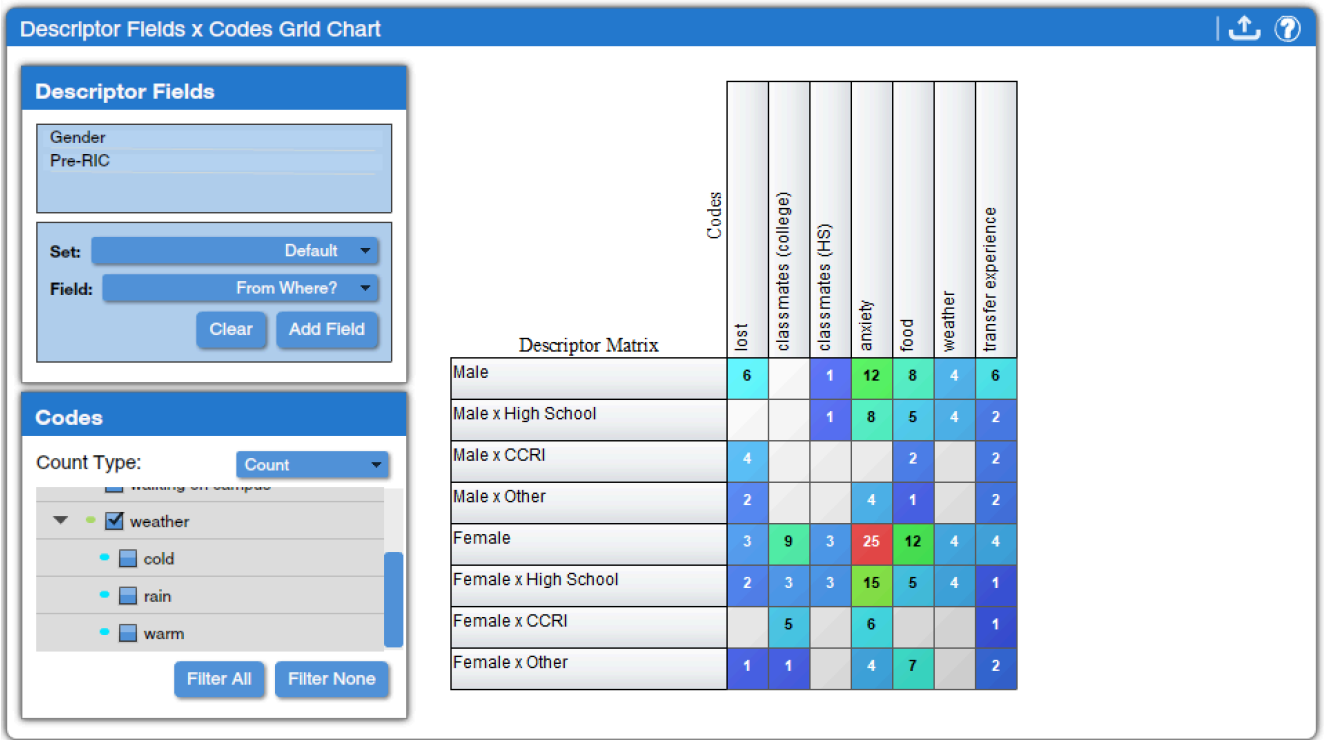


Figure 11. Descriptor Fields x Codes Grid Chart

Descriptor Tools

The descriptor tools are tools designed to provide summary data, including descriptive statistics and some basic explanatory statistics, using descriptors as variables. Unlike other tools in Dedoose, the descriptor tools produce primarily quantitative data. As such, they are generally most useful either for presenting basic descriptive data about participants in a study or when the study is designed as a mixed-methods study. However, in most cases, if more than the most basic descriptive data is desired, it would make more sense to export the relevant descriptor data from Dedoose (which can be done under the Descriptors tab) and load it into appropriate statistical analysis software.

Descriptor Ratios Multi Chart (found under Descriptors Charts and Quantitative Charts) provides a choice of pie or bar graphs that display the number of texts associated with each category for each descriptor field. This tool, as shown in Figure 12, is a good way to quickly familiarize oneself with the distribution of descriptor data in a project.

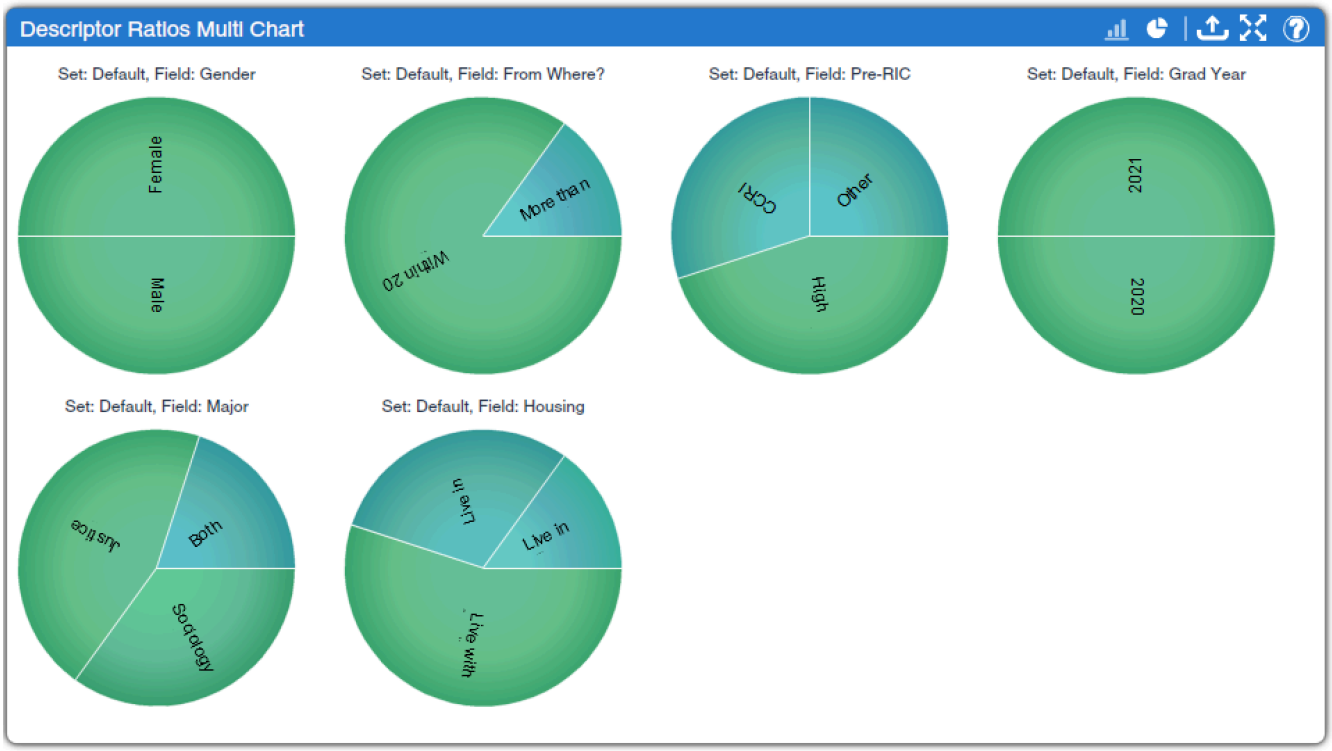


Figure 12. Descriptor Ratios Multi Chart

Descriptor Ratios Grid Chart (found under Descriptors Charts and Quantitative Charts) provides a way to view **crosstabulations** of descriptor field categories. Analysts choose one descriptor from the Target Field drop-down menu and that descriptor is used as the independent variable in a stack of crosstabulations, with the other descriptors as the dependent variables. A toggle is available to turn on or off the inclusion of descriptors that have not been linked to a text. For example, Figure 13 shows students' status prior to coming to our campus as the target field; crosstabulations with gender and how far away students lived from campus prior to becoming students as the other included variables. More information on how to interpret crosstabulations can be found in the chapter on Bivariate Analyses: Crosstabulation, but in short, researchers compare the percentages across the rows. Doing so for the data displayed in Figure 13 shows that there is no notable gender difference in students' educational status before coming to our campus, but that there is a difference in terms of how far away their homes are from campus—students who came to our campus from the local community college are more likely to live within 20 miles of campus.

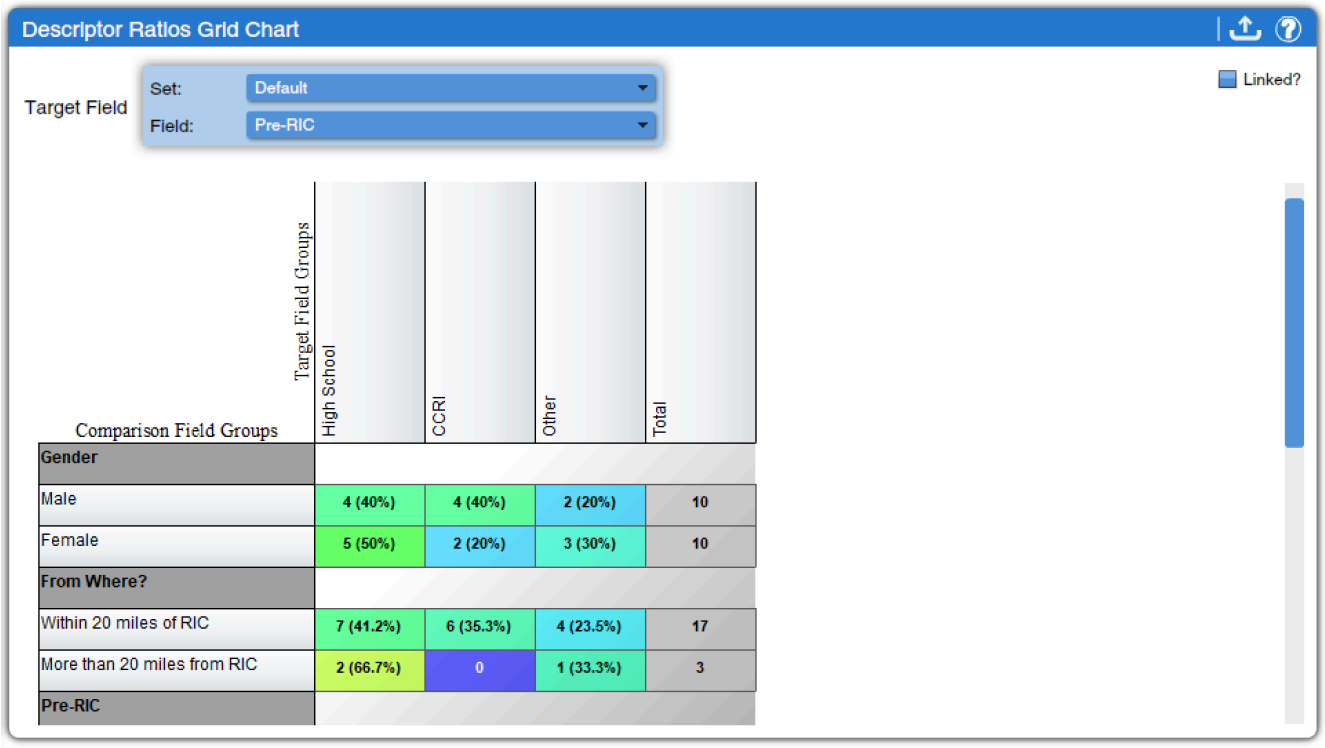


Figure 13. Descriptor Ratios Grid Chart

Descriptor Field x Descriptor Field (found under Descriptors Charts and Quantitative Charts) provides a different way to examine crosstabulation data. To use this tool, researchers select two descriptors, one for the outer field and one for the inner field. A separate bar graph is produced for each category of the inner field, with bars for each category of the outer field. Options toggle whether the count is of descriptors or of excerpts, whether only linked descriptors are included, and whether categories with no data are included. At the bottom of the screen, the critical value of the **Chi square** and the degrees of freedom (df) are presented; clicking on the question mark in a circle at the bottom of the screen brings up a webpage that provides the table of critical values of Chi square, as discussed in the chapter on Bivariate Analyses: Crosstabulation. For example, in Figure 14, an analysis of the relationship between housing one's first year on campus and how far away one lived prior to enrolling is presented. The data shows that the vast majority of students lived within 20 miles of our campus prior to enrolling, and that those students were more likely to continue to live with family, while students from further away were more likely to live in on-campus housing. However, the Chi square calculation is such that this observed relationship is not statistically significant. To determine this, note the Chi square and df values and click on the question mark in a circle at the bottom of the screen, and follow the directions to use the table that is presented.

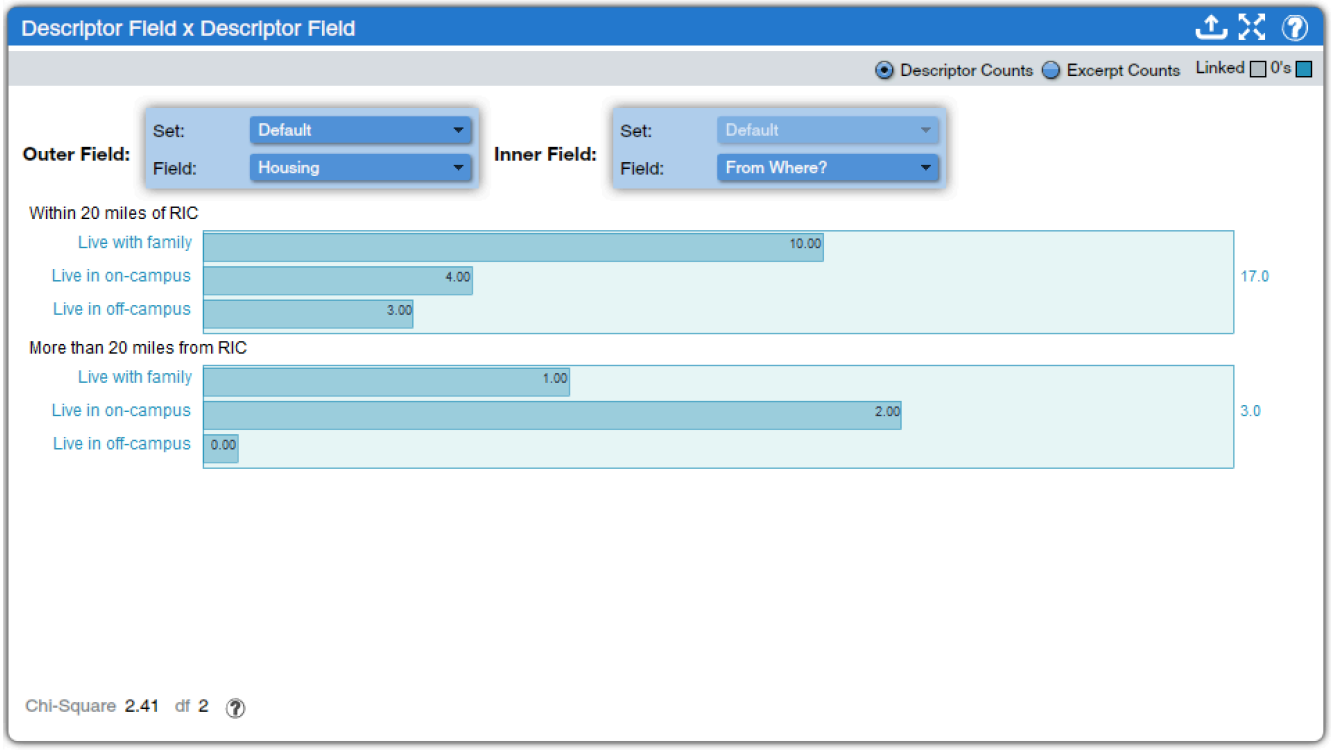


Figure 14. The Descriptor Field x Descriptor Field Tool

Because this text recommends that statistical analysis be performed using statistical analysis software, discussion of the remaining more-quantitative tools will be limited–especially as these tools rely heavily on the inclusion of continuous, numerical variables among the descriptors, which is generally less common in qualitative and perhaps even mixed-methods projects. The **Descriptor Number Distribution Plot** (found under Descriptors Charts and Quantitative Charts) provides a way to obtain basic descriptive data for large descriptors in larger datasets. However, its use is not possible with smaller datasets such as are commonly used for qualitative analysis. The **Descriptor Field T-Test** (found under Descriptors Charts, Code Charts, Quantitative Charts, and Mixed Methods Charts) produces an independent-samples T-test. To use this tool, you can select under the drop-down for "primary field" any descriptor with categories, and this will be the independent variable. The dependent variable must be **continuous** rather than **discrete**. The tool then produces a plot, mean and median differences, and the T-value and degrees of freedom to look up in the chart that loads when the "Critical Values Table" button is pressed. Similarly, the **Descriptor ANOVA** (found under Descriptors Charts, Code Charts, Quantitative Charts, and Mixed Methods Charts) produces an ANOVA output for a discreet independent variable and a continuous dependent variable. The **Descriptor Field Correlation** (found under Descriptors Charts and Quantitative Charts) produces a kind of scatterplot with descriptors that are continuous variables in nature presented in both the X-axis and the Y-axis. As in the other

tool, the correlation and degrees of freedom (here called DoF) are provided, along with a button to bring up the relevant critical value table.

Code Weight Tools

Code weighting is generally beyond the scope of this text, but it is worth briefly noting that there are four analysis tools designed specifically for projects that use code weight. In addition, some of the tools discussed above have options that allow the incorporation of data on code weights into the analysis process, as you have seen.

Code Weight Statistics (found under Code Charts and Qualitative Charts) displays the minimum, maximum, mean, median, and count of code weights applied for each code, while **Code Weight Distribution Plot** (found under Code Charts and Quantitative Charts) provides more ways to examine descriptive statistics about code weights for a particular code.

Code Weight Descriptor Bubble Plot (found under Descriptors Charts, Code Charts, and Mixed Methods Charts) lets users create four-dimensional visualizations of the relationship between three different codes with code weights and one descriptor. Codes can be assigned to the X axis, Y axis, and bubble size, while the categories of the descriptor become different bubbles on the graph. **Code Weight Frequency x Field** (found under Code Charts and Mixed Methods Charts) permits the analyst to see how the weight of codes varies across different categories of a descriptor.

Conducting and Concluding Analysis

While Dedoose provides powerful analytical tools, it is important to remember that it remains up to the analyst to make the right choices and use the tools appropriately in service of the project. Analysts need to make sure to choose the tools that fit with the research question and approach, not just those that are appealing or easy to use. For instance, many students of Dedoose are drawn to the code cloud tools because they are attractive and simple—but they provide far less analytical power than does the code co-occurence tool or even the code application tool.

In addition, in qualitative research, it is not sufficient to simply report what a tool tells you. And it is very unlikely that many of the graphs, tables, and other visuals Dedoose produces will find their way into publications or presentations. Instead, the tools should be used as a guide to determine what findings are worth investigating further or focusing on. Analysts will then need to return to the texts to choose appropriate quotes for illustrating these find-

ings and making sure the findings are sensible in the context of the data. Dedoose provides tools to help researchers make sense of their data, but it does not itself provide answers.

Exercises

1. Return to the Dedoose project involving oral history transcripts. Click though all of the analysis tools. Select two tools you find useful (*not just easy or pretty*) and write a paragraph summarizing the findings from each tool.
2. Use the tools you have selected to locate and copy two quotes that illustrate each of the findings you have summarized.

Media Attributions

- analyze tools © Dedoose is licensed under a All Rights Reserved license
- codecountmedia © Dedoose is licensed under a All Rights Reserved license
- code presence © Dedoose is licensed under a All Rights Reserved license
- code application © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- code coocurrance © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- packed code cloud © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- code x descriptor © Dedoose is licensed under a All Rights Reserved license
- descriptor x code © Dedoose is licensed under a All Rights Reserved license
- descriptor x code count table © Dedoose is licensed under a All Rights Reserved license
- descriptor x descriptor x code © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- code descriptor bubble © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- descriptor codes grid © Dedoose adapted by Mikaila Mariel Lemonik Arthur is licensed under a All Rights Reserved license
- descriptor ratios multi © Dedoose is licensed under a All Rights Reserved license
- descriptor ratios grid © Dedoose is licensed under a All Rights Reserved license
- descriptor field x descriptor field © Dedoose

Glossary

abduction

An approach to research that combines both inductive and deductive elements.

abstract

A short summary of a text written from the perspective of a reader rather than from the perspective of an author.

addition theorem

The theorem addressing the determination of the probability of a given outcome occurring at least once across a series of trials; it is determined by adding the probability of each possible series of outcomes together.

analytic coding

Coding designed to move analysis towards the development of themes and findings.

anecdotalism

When researchers choose particular stories or incidents specifically to illustrate a point rather than because they are representative of the data in general.

ANOVA

A statistical test designed to measure differences between groups.

antecedent variable

A variable that is hypothesized to affect both the independent variable and the dependent variable.

applied research

Research designed to address a specific problem.

archive

A repository of documents, especially those of historical interest.

association

The situation in which variables are able to be shown to be related to one another.

attributes

The possible levels or response choices of a given variable.

bar chart

Also called bar graphs, these graphs display data using bars of varying heights.

basic research

Research designed to increase knowledge, regardless of whether that knowledge may have any practical application.

bell curve

A graph showing a normal distribution—one that is symmetrical with a rounded top that then falls away towards the extremes in the shape of a bell

beta

The standardized regression coefficient. In a bivariate regression, the same as Pearson's r; in a multivariate regression, the correlation between the given independent variable and the dependent variable when all other variables included in the regression are controlled for.

binary

Consisting of only two options. Also known as dichotomous.

bivariate analyses

Quantitative analyses that tell us about the relationship between two variables.

block quote

A quotation, usually one of some length, which is set off from the main text by being indented on both sides rather than being placed in quotation marks.

CAQDAS

An acronym for "computer-aided qualitative data analysis software," or software that helps to facilitate qualitative data analysis.

causation

A relationship between two phenomena where one phenomenon influences, produces, or alters another phenomenon.

central limit theorem

The theorem that states that if you take a series of sufficiently large random samples from the population (replacing people back into the population so they can be reselected each time you draw a new sample), the distribution of the sample means will be approximately normally distributed.

chronology

A list or diagram of events in order of their occurrence in time.

cliques

An exclusive circle of people or organizations in which all members of the circle have connections to all other members of the circle.

closed coding

Coding in which the researcher developed a coding system in advance based on their theory, hypothesis, or research question.

code tree

A hierarchically-organized coding system.

code weights

Elements of a coding strategy that help identify the intensity or degree of presence of a code in a text.

codebooks

Documents that lay out the details of measurement. Codebooks may be used in surveys to indicate the way survey questions and responses are entered into data analysis

software. Codebooks may be used in coding to lay out details about how and when to use each code that has been developed.

codes

Words or phrases that capture a central or notable attribute of a particular segment of textual or visual data.

coding

The process of assigning observations to categories.

coding (in quantitative methods)

Assigning numerical variables to replace the names of variable categories.

cognitive map

Visualizations of the relationships between ideas.

collinearity

The condition where two independent variables used in the same analysis are strongly correlated with one another.

column marginal

The total number of cases in a given column of a table.

concept coding

Coding using words or phrases that represent concepts or ideas.

confidence interval

A range of estimates into which it is highly probable that an unknown population parameter falls.

confidence level

The probability that the sample statistics we observe holds true for the larger population.

continuous variable

A variable measured using numbers, not categories, including both interval and ratio variables. Also called a scale variable.

control variable

A variable that is neither the independent variable nor the dependent variable in a relationship, but which may impact that relationship.

controlling a relationship

Examining a relationship between two variables while eliminating the effect of variation in an additional variable, the control variable.

crosstabulation

An analytical method in which a bivariate table is created using discrete variables to show their relationship.

data cleaning

The process of examining data to find any errors, mistakes, duplications, corruptions, omissions, or other issues, and then correcting or removing data as is appropriate.

data display

Tables, diagrams, figures, and related items that enable researchers to visualize and organize data in ways that permit the perception of patterns, comparisons, processes, or themes.

data management

The process of organizing, preserving, and storing data so that it can be used effectively.

data reduction

The process of reducing the volume of data to make it more usable while maintaining the integrity of the data.

decision tree

A diagram that lays out the steps taken to reach decisions.

deductive

An approach to research in which researchers begin with a theory, then collect data and use that data to test their theory.

deductive coding

Coding in which the researcher developed a coding system in advance based on their theory, hypothesis, or research question.

degrees of freedom

The number of cells in a table that can vary if we know something about the row and column totals of that table, calculated according to the formula (# of columns-1)*(# of rows-1).

denominator

The expression below the line in a fraction; the entity used to divide another entity in a formula.

dependent variable

A variable that is affected or influenced by (or depends on) another variable; the effect in a causal relationship.

descriptive coding

Coding that relies on nouns or phrases describing the content or topic of a segment of text.

descriptive statistics

Statistics used to describe a sample.

descriptor

A category in an information storage system; more specifically in Dedoose, a characteristic of an author or entire text. Also, the word used to indicate that category or characteristic.

deviant case

A case that appears to be an exception to commonly-understood patterns or explanations.

dichotomous

Consisting of only two options. Also known as binary.

direction

How categories of an independent variable are related to categories of a dependent variable.

discrete variable

A variable measured using categories rather than numbers, including binary/dichotomous, nominal, and ordinal variables.

dramaturgical coding

Coding that treats texts as if they are scripts for a play.

dummy variable

A two-category (binary/dichotomous) variable that can be used in regression or correlation, typically with the values 0 and 1.

edge

The line connecting nodes in a network diagram; such lines represent real-world relationships or linkages.

elaboration

A term used to refer to the process of controlling for a variable.

elimination of alternatives

In relation to causation, the requirement that for a causal relationship to exist, all possible explanations other than the hypothesized independent variable have been eliminated as the cause of the dependent variable.

emotion codes

Codes indicating emotions discussed by or present in the text, sometimes indicated by the use of emoji/emoticons.

empirical

That which could hypothetically be shown to be true or false; statements about reality rather than opinion.

epistemology

The philosophical study of the nature of knowledge.

ethics (in research)

Standards for the appropriate conduct of research that seek to ensure researchers treat human participants in research appropriately and do not harm them and that scientific misconduct is avoided.

ethnography

A research method in which the researcher is a participant in a social setting while simultaneously observing and collecting data on that setting and the people within it.

evaluation coding

A coding system used to indicate what is or is not working in a program or policy.

exhaustive

The property of a variable which has a category for everyone.

extraneous variable

A variable that impacts the dependent variable but is not related to the independent variable.

face validity

The extent to which measures appear to measure that which they were intended to measure.

feminism

A perspective rooted in the idea that explorations and understandings of gendered power relations should be at the root of inquiry and action.

fieldnotes

Qualitative notes recorded by researchers in relation to their observation and/or participation of participants, social circumstances, events, etc. in which they document occurrences, interactions, and other details they have observed in their observational or ethnographic research.

first-cycle coding

Coding that occurs early in the research process as part of a bridge from data reduction to data analysis.

flow chart

A diagram of a sequence of operations or relationships.

focus group

A research method in which multiple participants interact with each other while being interviewed.

focused coding

Selective coding designed to orient an analytical approach around certain ideas.

frequency distribution

An analysis that shows the number of cases that fall into each category of a variable.

gamma

A measure of the direction and strength of a crosstabulated relationship between two ordinal-level variables.

General Social Survey

A nationally-representative survey on social issues and opinions which has been carried out roughly every other year since 1972. Also known as the GSS.

generalizability

The degree to which a finding based on data from a sample can be assumed to be true for the larger population from which the population was drawn.

genre

A classification of written or artistic work based on form, content, and style.

gerunds

Verb forms that end in -ing and function grammatically in sentences as if they are nouns.

grounded theory

An inductive approach to data collection and data analysis in which researchers strive to generate a conception of how participants understand their own lives and circumstances.

Hawthorne effect

When research participants modify their behavior, actions, or responses due to their awareness that they are being observed.

histogram

A graph that looks like a bar chart but with no spaces between the bars, it is designed to display the distribution of continuous data by creating rectangles to represent equally-sized groups of values.

hypothesis

A statement of the expected or predicted relationship between two or more variables.

In vivo coding

Coding that relies on research participants' own language.

independent variable

A variable that may affect or influence another variable; the cause in a causal relationship.

index variable

A composite variable created by combining information from multiple variables.

inductive

A research approach in which researchers begin by collecting data and then use this data to build theory.

inductive coding

Coding in which the researcher develops codes based on what they observe in the data they have collected.

inferential statistics

Statistics that permit researchers to make inferences (or reasoned conclusions) about the larger populations from which a sample has been drawn.

inter-rater reliability

The extent to which multiple raters or coders assign the same or a similar score, code, or rating to a given text, item, or circumstance.

interaction term

A variable constructed by multiplying the values of other variables together so as to make it possible to look at their combined impact.

interpretivism

A philosophy of research that assumes all knowledge is constructed and understood by human beings through their own individual and cultural perspectives.

interval variable

A variable with adjacent, ordered categories that are a standard distance from one another, typically as measured numerically.

intervening variable

A variable hypothesized to intervene in the relationship between an independent and a dependent variable; in other words, a variable that is affected by the independent variable and in turn affects the dependent variable.

interview

A research method in which a researcher asks a participant open-ended questions.

iterative

A process in which steps are repeated.

Kappa

A measure of association especially likely to be used for testing interrater reliability.

kurtosis

How sharp the peak of a frequency distribution is. If the peak is too pointed to be a normal curve, it is said to have positive kurtosis (or "leptokurtosis"). If the peak of a distribution is too flat to be normally distributed, it is said to have negative kurtosis (or platykurtosis).

latent coding

Interpretive coding that focuses on meanings within texts.

leptokurtosis

The characteristic of a distribution that is too pointed to be a normal curve, indicated by a positive kurtosis statistic.

levels of measurement

Classification of variables in terms of the precision or sensitivity in how they are recorded.

line of best fit

The line that best minimizes the distance between itself and all of the points in a scatterplot.

linear relationship

A relationship in which a scatterplot will produce a reasonable approximation of a straight line (rather than something like a U or some other shape).

logistic regression

A type of regression analysis that uses the logistic function to predict the odds of a particular value of a binary dependent variable.

manifest coding

Coding of surface-level and/or easily observable elements of texts.

margin of error

A suggestion of how far away from the actual population parameter a sample statistic is likely to be.

matrices

Tables with rows and columns that are used to summarize and analyze or compare data.

mean

The sum of all the values in a list divided by the number of such values.

measures of central tendency

A measure of the value most representative of an entire distribution of data.

measures of dispersion

Statistical tests that show the degree to which data is scattered or spread.

median

The middle value when all values in a list are arranged in order.

metadata

Data about other data.

mode

The category in a list that occurs most frequently.

multiple regression

Regression analysis looking at the relationship between a dependent variable and more than one independent variable.

multiplication theorem

The theorem in probability about the likelihood of a given outcome occurring repeatedly over multiple trials; this is determined by multiplying the probabilities together.

multivariate analyses

Quantitative analyses that explores relationships involving more than two variables or examines the impact of other variables on a relationship between two variables.

multivariate regression

Regression analysis looking at the relationship between a dependent variable and more than one independent variable.

mutually exclusive

The characteristic of a variable in which no one can fit into more than one category, such as age categories 5-10 and 11-15 (rather than 5-10 and 10-15, as this would mean ten-year-olds fit into two categories).

network diagram

A visualization of the relationships between people, organizations, or other entities.

NHST

Null hypothesis significance testing.

nodes

Points in a network diagram that represents an individual person, organization, idea, or other entity of the type the diagram is designed to show connections between.

nominal variable

A variable whose categories have names that do not imply any order.

normal distribution

A distribution of values that is symmetrical and bell-shaped.

null hypothesis

The hypothesis that there is no relationship between the variables in question.

null hypothesis significance testing

A method of testing for statistical significance in which an observed relationship, pattern, or figure is tested against a hypothesis that there is no relationship or pattern among the variables being tested

objectivity

The ability to evaluate something without individual perspectives, values, or biases impacting the evaluation.

observational research

A research method in which the researcher observes the actions, interactions, and behaviors of people.

open coding

Coding in which the researcher develops codes based on what they observe in the data they have collected.

ordinal variable

A variable with categories that can be ordered in a sensible way.

organizational chart

A diagram, usually a flow chart, that documents the hierarchy and reporting relationships within an organization.

original relationship

The relationship between an independent variable and a dependent variable before controlling for an additional variable.

p value

The measure of statistical significance typically used in quantitative analysis. The lower the p value, the more likely you are to reject the null hypothesis.

paradigm

A set of assumptions, values, and practices that shapes the way that people see, understand, and engage with the world.

partial

Shorter term for a partial relationship.

partial relationship

A relationship between an independent and a dependent variable for only the portion of a sample that falls into a given category of a control variable.

participant-observation

A research method in which the researcher observes social interaction while themselves participating in the social setting.

participants

People who participate in a research project or from or about whom data is collected.

Pearson's chi-square

A measure of statistical significance used in crosstabulation to determine the generalizability of results.

Pearson's r

A measure of association that calculates the strength and direction of association between two continuous (interval and/or ratio) level variables.

Pie charts

Circular graphs that show the proportion of the total that is in each category in the shape of a slice of pie.

platykurtosis

The characteristic of a distribution that is too flat to be a normal curve, indicated by a negative kurtosis statistic.

population

A group of cases about which researchers want to learn something; generally, members of a population share common characteristics that are relevant to the research, such as living in a certain area, sharing a certain demographic characteristic, or having had a common experience.

population parameter

A quantitative measure of data from a population.

positionality

An individual's social, cultural, and political location in relation to the research they are doing.

positivism

A view of the world in which knowledge can be obtained through logic and empirical observation and the world can be subjected to prediction and control.

pragmatism

A philosophy that suggests that researchers can adapt elements of both objectivist and interpretivist philosophies.

probability

How likely something is to happen; also, a branch of mathematics concerned with investigating the likelihood of occurrences.

probability sample

A sample that has been drawn to give every member of the population a known (non-zero) chance of inclusion.

process coding

Coding in which gerunds are applied to actions that are described in segments of text.

process diagrams

Visualizations that display the relationships between steps in a process or procedure.

qualitative data analysis

Data analysis in which the data is not primarily numeric, for instance based on words or images.

quantification

The transformation of non-numerical data into numerical data.

quantitative data analysis

Data analysis in which the data is numerical.

R squared

The square of the regression coefficient, which tells analysts how much of the variation in the dependent variable has been explained by the independent variable(s) in the regression.

R2 change

The change in the percent of the variance of the dependent variable that is explained by all of the independent variables together when comparing two different regression models

random sample

A sample in which all members of the population have an equal probability of being selected.

range

The highest category in a list minus the lowest category.

ratio level variable

A numerical variable with an absolute zero which can also be multiplied and divided.

reflexivity

A continual reflection on the research process and the researcher's role within that process designed to ensure that researchers are aware of any thought processes that may impact their work.

regression

A statistical technique used to explore how one variable is affected by one or more other variables.

regression line

The line that is the best fit for a series of data, typically as displayed in a scatterplot.

relationship (between variables)

When certain categories of one variable are associated, or go together, with certain categories of the other variable(s).

reliability

The extent to which multiple or repeated measurements of something produce the same results.

repeatability

The extent to which a researcher can repeat a measurement and get the same result.

replicability

The extent to which a research study can be entirely redone and yet produce the same overall findings.

replicate

Repeating a research study with different participants.

representativeness

The degree to which the characteristics of a sample resemble those of the larger population.

reproducibility

The extent to which a new study designed to test the same hypothesis or answer the same question ends up with the same findings as the original study.

respondents

People who participate in a research project or from or about whom data is collected.

rough coding

Coding for data reduction or as part of an initial pass through the data.

row marginal

The total number of cases in a given row of a table.

sample

A subset of cases drawn or selected from a larger population.

sample statistics

Quantitative measures of data from a sample.

sampling error

Measurement error created due to the fact that even properly-constructed random samples are do not have precisely the same characteristics as the larger population from which they were drawn.

saturation

The point in the research process where continuing to engage in data collection no longer yields any new insights. Can also be used to refer to the same point in the literature review process.

scale variable

A variable measured using numbers, not categories, including both interval and ratio variables. Also called a continuous variable.

scatterplot

A visual depiction of the relationship between two interval level variables, the relationship between which is represented as points on a graph with an x-axis and a y-axis.

second-cycle coding

Analytical coding that occurs later in the data analysis process.

significance (statistical)

A statistical measure that suggests that sample results can be generalized to the larger population, based on a low probability of having made a Type 1 error.

simple linear regression

A regression analysis looking at a linear relationship between one independent and one dependent variable.

skewness

An asymmetry in a distribution in which a curve is distorted either to the left or the right, with positive values indicating right skewness and negative values indicating left skewness.

social data analysis

The analysis of empirical data in the social sciences.

social responsibility (in research)

The extent to which research is conducted with integrity, is trustworthy, is relevant, and meets the needs of communities.

spurious

The term used to refer to relationship where variables seem to vary in relation to one another, but where in fact no causal relationship exists.

standard deviation

A measure of variation that takes into account every value's distance from the sample mean.

standard error

A measure of accuracy of sample statistics computed using the standard deviation of the sampling distribution.

standpoint

The particular social position in which a person exists and in which their understandings of the world are rooted.

strength (of relationship)

A measure of how well we can predict the value or category of the dependent variable for any given unit in our sample based on knowing the value or category of the independent variable(s).

string

A data type that represents non-numerical data; string values can include any sequence of letters, numbers, and spaces.

structural coding

Coding that indicates which research question or hypothesis is being addressed by a given segment of text.

subjects

People who participate in a research project or from or about whom data is collected.

summarization

The process of creating abridged or shortened versions of content or texts that still keep intact the main points and ideas they contain.

table

A display that uses rows and columns to show information.

temporal order

The order of events in time; in relation to causation, the fact that independent variables must occur prior to dependent variables.

the elaboration model

A typology developed by Paul Lazarsfeld for the possible analytical outcomes of controlling for a variable.

themes

Concepts, topics, or ideas around which a discussion, analysis, or text focuses.

thick description

A detailed narrative account of social action that incorporates rich details about context and meaning such that readers are able to understand the analytical meaning of the description.

timeline

A diagram that lays out events in order of when they occurred in time.

trace analysis

Research that uses the traces of life people have left behind as data, as in archeology.

triangulation

The use of multiple methods, sites, populations, or researchers in a project, especially to validate findings.

type 1 error

The error made if one infers that a relationship exists in a larger population when it does not really exist; in other words, a false positive error.

type 2 error

The error you make when you do not infer a relationship exists in the larger population when it actually does exist; in other words, a false negative conclusion.

typologies

Classification systems.

univariate

Using one variable.

univariate analyses

Quantitative analyses that tell us about one variable, like the mean, median, or mode.

validity

The degree to which research measurements accurately reflect the real phenomena they are intended to measure.

values coding

Coding that relies on codes indicating the perspective, worldview, values, attitudes, and/or beliefs of research participants.

variable

A characteristic that can vary from one subject or case to another or for one case over time within a particular research study.

variance

A basic statistical measure of dispersion, the calculation of which is necessary for computing the standard deviation.

versus coding

Coding that relies on a series of binary oppositions, one of which must be applied to each segment of text.

voice

The style or personality of a piece of writing, including such elements as tone, word choice, syntax, and rhythm.

word cloud

Visual display of words in which the size and boldness of each word indicates the frequency with which it appears in a body of text.

Yule's Q

A measure of the strength of association use with binary variables

Z score

A way of standardizing data based on how many standard deviations away each value is from the mean.

Modified GSS Codebook for the Data Used in this Text

The General Social Survey

2021 GSS (Cross-section study) Documentation and Public Use File Codebook (Release 2)

Edited and Modified for use with this text

Citation of This Document

In publications, please acknowledge the original source. The citation for this Public Use File is:

Davern, Michael; Bautista, Rene; Freese, Jeremy; Morgan, Stephen L.; and Tom W. Smith. General Social Survey 2021 Cross-section. [Machine-readable data file]. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Stephen L. Morgan, and Tom W. Smith. NORC ed. Chicago, 2021. 1 datafile (68,846 cases) and 1 codebook (506 pages).

The codebook is provided "as is," without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of, or in connection with the codebook or the use or other dealings in the codebook.

Please contact the GSS team at gss@norc.org with any questions or requests.

Notes on the modified version:

A modified version of the 2021 GSS was created for use with the Open Educational Resources text *Social Data Analysis*. This version is simplified for use by undergraduate students learning SPSS software; as such, survey weights and certain variables have been removed from the dataset and it has been converted to SPSS format. This codebook includes only important information about the 2021 GSS and information about those variables included in the modified dataset. Other information has been removed to shorted and simply the codebook. –Mikaila Mariel Lemonik Arthur

2021 GENERAL SOCIAL SURVEY CROSS-SECTION CODEBOOK, RELEASE 1

(Codebook for the Machine-Readable Data File 2021 General Social Survey Cross-section)

Principal Investigators	Michael Davern
Co-Principal Investigator	René Bautista
Co-Principal Investigator	Jeremy Freese
Co-Principal Investigator	Stephen L. Morgan
Co-Principal Investigator	Tom W. Smith
Senior Advisor	Colm O'Muircheartaigh
Research Associates	Jaesok Son Benjamin Schapiro Jodie Smylie Beth Fisher Katherine Burda Ned English Steven Pedlow Amy Idhe María Sánchez Eyob Moges Hans Erickson Abigail Norling Ruggles Rachel Sparkman

Produced by NORC at the University of Chicago

This project was supported by the National Science Foundation

INTRODUCTION

Introduction to the General Social Survey (GSS)

The General Social Survey (sometimes, General Social Surveys) is a series of nationally representative cross-sectional interviews in the United States that have occurred since 1972. The GSS collects data on contemporary American society to monitor and explain trends in opinions, attitudes, and behaviors. The GSS has adapted questions from earlier surveys, thereby allowing researchers to conduct comparisons for up to 80 years. Originally proposed and developed by James A. Davis, the GSS has been administered by NORC at the University of Chicago (NORC) and funded by the National Science Foundation (NSF) since its inception. Currently, the GSS is designed by a set of Primary Investigators (PIs), with input from the GSS Board, comprised of notable researchers within the scientific community. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-

being, social mobility, and stress and traumatic events. Altogether, the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general, as well as the role played by relevant subgroups and to compare the United States to other nations. The GSS aims to make high-quality data easily accessible to scholars, students, policymakers, and others, with minimal cost and waiting.

The GSS has been tracking trends in public opinion since 1972. Throughout, the GSS has taken great care to keep the survey methodology as comparable over time as possible, which includes everything from keeping the same sampling approach to not changing question wording. This is done to minimize potential changes due to changes in methodology and support the study of trends in public opinion in the United States over time. However, due to the global COVID19 pandemic, the 2021 GSS Cross-section[1] implemented significant methodological adaptions for the safety of respondents and interviewers. Since its inception, the GSS has traditionally used in-person data collection as its primary mode of data collection. However, the 2021 GSS Cross-section used an address-based sampling with push to web and a web self-administered questionnaire. This new survey design and methodology bring numerous changes, which are discussed in this codebook.

The GSS comprises a core set of items (the Replicating Core) that are repeated every round, as well as topical modules, which may or may not be repeated. The GSS is currently composed of three separate ballots (A, B, and C), as well as two separate forms (X and Y), which allow for up to six different paths through the interview itself (in addition to paths determined by respondent answers, such as questions about spouses or partners, or questions on employment). Not every question in the Replicating Core is asked of every respondent; most only appear on two of the three ballots. However, every item in the Replicating Core overlaps on at least one ballot with every other item in the Replicating Core, ensuring that researchers can estimate inter-item correlations. Forms are used for experiments such as wording variations within questions, ensuring that half of the respondents on each ballot see the experimental or control conditions of each relevant variable. Within the GSS, these form experiments are usually assigned mnemonics that end in -Y.

Topical modules are typically assigned to either two full ballots (e.g., A and B) or one full ballot and one half-ballot (e.g., A and BX), covering two-thirds or half of sample respondents, respectively. However, some topical modules are included on all ballots. Modules are usually assigned to specific ballots based on one of two conditions: overlap with other key questions (either ensuring that respondents to specific items also receive specific modules or that respondents to specific items *do not* receive specific modules), or time constraints. The GSS tries to balance the length of all six paths to be approximately equal. Topical modules

1. See Study Overview for an explanation for why this is the 2021 GSS, rather than the 2020 GSS.

may be administered via interviewer in any mode or completed by self-administered questionnaire, depending on the sensitivity of the items included.

Topical modules come from several different sources. While the GSS broadly is funded by NSF, individual modules may be sponsored by other government agencies, universities, research institutes, or individuals. The GSS typically includes modules every round that are related to the *International Social Survey Programme* (ISSP), a consortium of national-level studies like the GSS (for more information, see Introduction to the International Social Survey *Programme*, below). Finally, modules may be solicited by the GSS Scientific Advisory Board or the Principal Investigators and can be included based on scientific merits and available time in the interview. The number of GSS modules varies by year.

Additionally, the GSS has implemented experimental designs over time or through collaborations (for instance, supporting other studies such as the National Organization Studies, National Congregations Study, National Voluntary Associations Study, and the 2016-2020 General Social Survey-American National Election Studies (GSS- ANES) Panel), which have led to several ancillary datasets.

Introduction to the International Social Survey Programme (ISSP)

The ISSP is a consortium of nationally representative research studies, like the GSS, who have all agreed to ask questions on the same topics on an annual basis. It emerged out of bilateral collaboration with NORC and the German organization Zentrum für Umfragen, Methoden, und Analysen (ZUMA; now part of GESIS-Leibniz Institute of the Social Sciences). Starting in 1982, each organization devoted a small segment of their national surveys, ALLBUS and GSS, to a common set of questions. The ISSP was formally established in 1984 by Australia, Germany, Great Britain, and the United States, and it now has 42 member countries across five continents and collects data in 70 countries. NORC represents the United States as a country in the ISSP.

ISSP modules have several defining criteria:

- Modules are developed in English and translated to every administered language.
- Modules must contain approximately 60 questions and can be supplemented by optional items, as well as around 40 items on background and demographic characteristics.
- Modules must contain the same items across all languages and studies, asked in the same order, with minor variations to account for mode differences or regional differences.[2]
- Each topical module must be administered within a relatively narrow time frame, usu-

ally about 18 months from the start of the relevant year.

ISSP Modules are currently replicated every 10 years, allowing for topics to be studied as both a multinational snapshot at a single point in time as well as a slowly evolving repeated cross-section of national opinions. While not every topic is repeated, the longest-running module is up to its fifth replication. ISSP rules require that when a topic is repeated, at least two-thirds of the questions must be repeated from a previous round, while up to one- third can be new questions.

Study Overview

Due to the onset of the COVID-19 pandemic in the United States in the early months of 2020, GSS staff redesigned the GSS in several ways to ensure the safety and well-being of all the people who participate in and administer the study.

In 2020, we conducted the GSS as two studies: 1) a panel reinterview of past respondents from the 2016 and 2018 cross-sectional GSS studies (referred to as the 2016-2020 GSS Panel), and 2) an independent fresh cross-sectional address-based sampling push-to-web study (referred to in this document as the 2021 GSS Cross-section but also known as the 2020 cross-sectional survey in previous documents). This codebook provides details of the second study—namely, the 2021 GSS Cross-section, where newly selected respondents answered a GSS questionnaire from December 2020 to May 2021. We refer to the second study as the 2021 GSS Cross-section because the majority of the data was collected in 2021. Documentation for the first study (the 2016-2020 GSS Panel) is provided separately. During the spring and summer of 2020, GSS staff redesigned both the panel and the cross-section to be administered primarily via web self-administered questionnaire, instead of face-to-face interviews, with phone interviews as a secondary mode.

Each of these major changes had several ramifications for sampling, fielding, questionnaire design, data cleaning, response rates, and weights.

Cross-section Overview

This codebook focuses on the 2021 GSS Cross-section survey. While this iteration of the Cross-section has meaningful changes from previous editions, there is much that remains consistent. Just as was done since 2004, the GSS Cross-section survey administers a full-

2. For example, asking about Congress or Parliament, or asking about the European Union or NAFTA.

probability sample approach with samples created from an adapted form of the United States Postal Service (USPS) metropolitan statistical area (MSA)/county frame area. More on the GSS conventional design can be found on pages 3177–3178 of the legacy cumulative codebook, available at the GSS website.

The GSS has been conducted since 1972 and currently functions as a social indicators program, which highly values historical trends and continuity. To that end, the GSS's replicating core contains items that have been asked since its inception. In some cases, these items were asked on even older surveys, allowing for continuous measurement of concepts since the 1940s.

Table 1. Key Aspects of the 2021 GSS Cross-section

KEY ASPECTS	2021 GSS Cross-section
Sample	Adults 18 or older in the United States who live in noninstitutional housing at the time of interviewing
Invitation	Mailing materials that show a web link to invite people to participate on the web (i.e., pushing respondent to a web survey first). Phone option was also provided.
Survey mode	Web (supplemented by phone)
Incentive	Both non-contingent pre-paid incentive and contingent post- paid incentive
Final sample size	4,032 completes from 27,591 lines of sample
Response Rate (AAPOR RR3)	17.4%
Fielding period	December 1, 2020, to May 3, 2021
Administration	Mail push to web as primary mode, supplemented with phone
Respondent selection within household	Last birthday method
Language	English and Spanish
Paradata derived from instrument	Paradata were recorded but are not available in release 1.

Occasionally, the cumulative datafile is updated with newly cleaned or derived variables, and error corrections. Please see Release Notes for changes since the initial release of the 1972-2021 Cumulative file.

NOTE ON MEASUREMENT AND INTERPRETATION

The GSS has been tracking trends in public opinion since 1972. Over this time, the GSS has taken great care to keep the survey methodology as comparable over time as possible, which includes everything from keeping the same sampling approach to not changing question wording. This is done to minimize potential changes due to methodology and support the study of changes in public opinion in the United States. However, due to the global COVID-19 pandemic, the 2021 GSS8 needed to implement significant methodological adaptations for the safety of respondents and interviewers. Since its inception, the GSS has traditionally used in-person data collection as its primary mode of data collection. In response to the COVID-19 pandemic, the GSS altered its methodology for the first time to be primarily an address-based sampling push-to-web methodology in 2021. As a result, when examining changes in trends over time, we caution GSS data users to carefully examine how a change they are observing in a trend may have been impacted by the methodological differences employed in 2021. The 2021 GSS Cross-section codebook provides documentation of the methodological changes and adaptations that were necessary to transition the GSS survey from in-person to a web survey.

Total Survey Error Perspective for GSS Trend Estimates in 2021

The GSS was collected in 2021 to provide vital opinion data to the research community at a critical time in U.S. history. While the data will contribute to our understanding of society, any changes in public opinion seen in the 2021 GSS data could be due to either changes in actual opinion and/or changes the GSS made in the methodology to adapt to COVID-19. When evaluating the GSS for trend changes over time, we caution our users to carefully consider changes in the GSS methodology from a total survey error perspective. Total survey error is a way of comprehending the impact on estimates due to measurement, nonresponse, coverage, and sampling error. Below, we provide a high-level summary of the components of the total survey error in the 2021 GSS Cross-section, but we invite the user to carefully review the details in the present document.

Measurement Error: Changes in how survey questions were administered can impact the answers. The GSS has traditionally been conducted in person and administered by an inter-

viewer. Due to the COVID-19 pandemic, the GSS needed to administer survey questions primarily over the web without any interviewer assistance. Some cases were collected over the phone as well. To adapt to the primary mode of administration (web), some changes were needed in the measures. For example, "Don't Know" response categories were included for factual questions, but they were not displayed for opinion questions (only for factual questions). In the past, Don't Knows could be recorded by interviewers regardless of whether they were factual or opinion questions.

Non-Response Error: Historically, the GSS has achieved high response rates well above 50 percent, mostly because in-person surveys can attain higher response rates. The 2021 GSS was conducted using a mail invite to push the respondent to the web. The 2021 GSS Cross-section response rate is 17 percent (which is still high for web surveys). Differential participation rates across selected GSS participants in 2020 relative to previous years could also contribute to a change in estimates. To help control for this concern, the GSS implemented an adjustment to known population totals for the 2021 Cross-section round based on a raking approach (i.e., post-stratification weighting) to ensure the weighted totals in the 2021 GSS Cross-section sample are as close as possible to the control totals from the U.S. Census Bureau estimates by education, sex, age, region of the country, race, and ethnicity.

Coverage Error: In 2021, the GSS had to change how it rostered and ultimately estimated adults residing in a household. Typically, household rosters have been used in the in-person methodology to randomly select one adult in the household to complete the survey. That is, in previous years the GSS has instructed interviewers to complete an initial household enumeration form to collect some basic data on everyone residing in a household and then randomly chose one adult with whom to complete the main interview. In the 2021 GSS such a household enumeration was not possible up front, and the GSS instructed selected households to identify the adult with the most recent birthday and to report the number of adults living in the household. It is possible that respondent selection may have happened at the household level or missed some household residents (for instance, people abroad, adult children living at home, etc.).

Sampling Error: The 2021 GSS Cross-section relies on scientific sampling to allow for the calculation of sampling error (i.e., margin of error). As in any scientific sample, any one trend from year to year can be impacted by the fact we only observed a sample and not the entire population. The dataset contains survey design variables (clusters, strata, and weights) to account for the complex survey sample design. And it is possible that our trend estimates are off by a sampling or "margin of error" between any given set of years. It is essential, therefore, to control for sampling error by conducting tests of significance for trend differences estimates.

We recommend our users include the one of the following statements when reporting on the GSS 2021 Cross-section data:

*Total Survey Error Summary Perspective for the 2021 GSS Cross-sectio*n: Changes in opin-

ions, attitudes, and behaviors observed in 2021 relative to historical trends may be due to actual change in concept over time and/or may have resulted from methodological changes made to the survey methodology during the COVID-19 global pandemic.

Suggested Statement to Include in Articles and Reports That Use GSS Data

To safeguard the health of staff and respondents during the COVID-19 pandemic, the 2021 GSS data collection used a mail-to-web methodology instead of its traditional in-person interviews. Research and interpretation done using the data should take extra care to ensure the analysis reflects actual changes in public opinion and is not unduly influenced by the change in data collection methods. For more information on the 2021 GSS methodology and its implications, please visit https://gss.norc.org/Get-The-Data

Screenshots of Changes

The 2021 GSS Cross-section comprised primarily a self-administered web questionnaire. The following typical item from this questionnaire displays both the typical layout of a self-administered web questionnaire item and the change to Don't Know response options discussed in Don't Know and No Answer Responses.

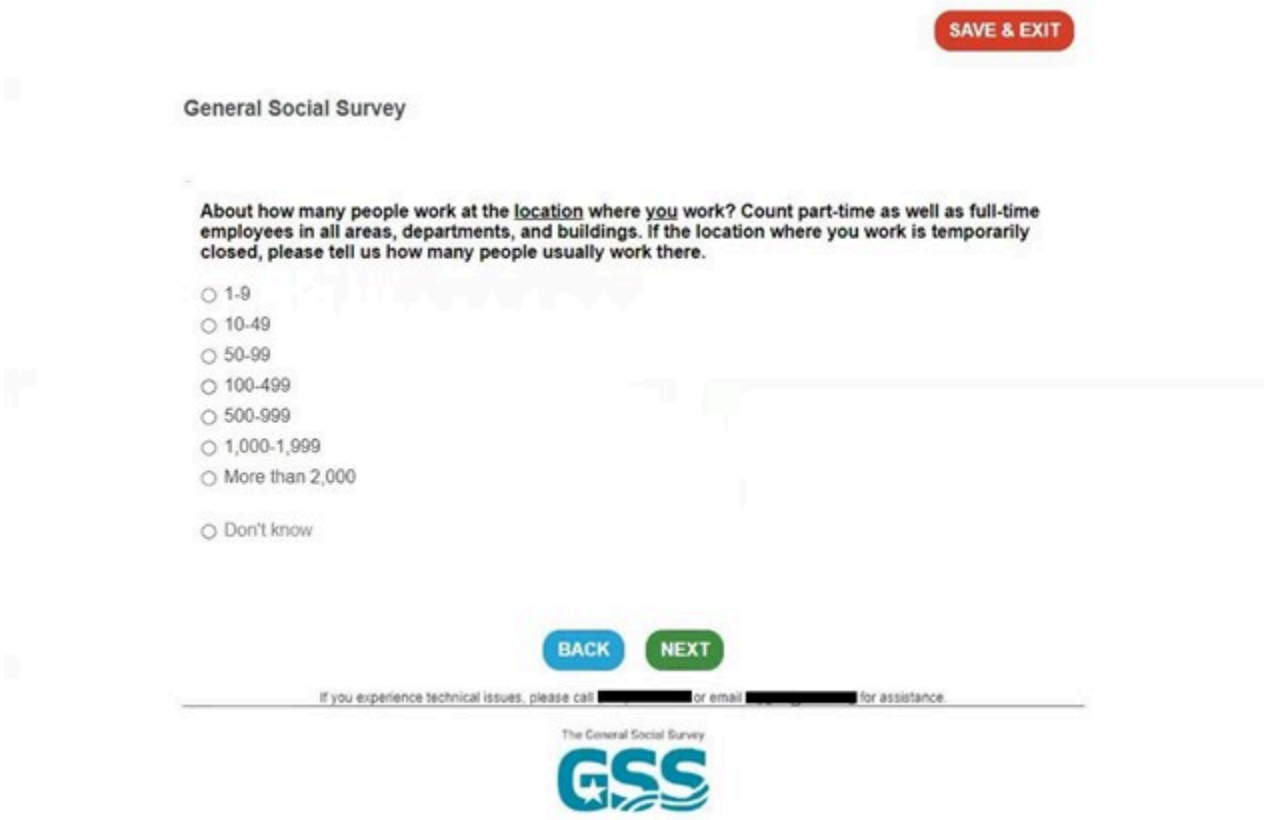


Figure 1: Visual Display of a Survey Question in the 2021 GSS Cross-section

Figure 1: Visual Display of a Survey Question in the 2021 GSS Cross-section

Appendix A: 2021 GSS Cross-section Outcomes

Table A1: 2021 GSS Cross-section Response Rate Calculation

Status	Frequency
Complete	4,032
Eligible cases (partial or NIR)	5,589
Non-interviewed respondent – Eligibility not known	19,670
Total	27,591

Completes were defined as those cases that completed the full interview or met the data threshold predefined by the research team to be included with the Completes category. Partial completes were cases that started the GSS interview but did not meet the threshold to be included with Completes (i.e., completed two-thirds of the core section). Non-interviewed respondent (NIR) cases were split into two categories: eligibility known and eligibility not known. “Eligibility known” cases were those in which the sampled address was confirmed as being an occupied residence. “Eligibility not known” cases were those in which the address was not confirmed as being an occupied residence. Finally, out-of-scope cases were those in which the sampled address was identified as vacant, a business, or not an address. Out-of-scope cases were also those in which nobody in the sampled household spoke English or Spanish, or the selected respondent was ill or incapacitated and could not complete the interview. A total of 3,561 (88 percent) of completes were completed by web SAQ, with the remainder (471) completed by phone (12 percent).

APPENDIX B: Other GSS Documentation

GSS Codebook

This is the legacy codebook that combines all of the notes, appendices, and frequency tables of the GSS since its inception in 1972 up through 2018. This is useful to see how trends have changed over the decades, find question-level wording, and find when questions were added or discontinued.

GSS Data Explorer

A web-based interactive tool to access and analyze GSS data, question wording, and ballot timing. The Data Explorer is helpful to visualize individual variables’ trend lines or extract

data into various software packages. The Data Explorer tool is going through renovations and it is expected to be upgraded in December of 2021.

The Codebook Guide to GSS Variables

1) ID
Respondent's ID number
RANGE: 1 to 4471
N Mean Std. Deviation
Total 4032 2221.813 1289.407

2) WRKSTAT
Last week were you working full time, part time, going to school, keeping house, or what?
RANGE: 1 to 8
N Mean Std. Deviation
Total 4024 3.111 2.254
1) Working full time 1776 44.1
2) Working part time 365 9.1
3) With a job, but not at work because of temporary illness, vacation, strike 104 2.6
4) Unemployed, laid off, looking for work 265 6.6
5) Retired 993 24.7
6) In school 88 2.2
7) Keeping house 315 7.8
8) Other 118 2.9
Missing 8

3) HRS1
How many hours did you work last week, at all jobs?
Responses greater than 89 were coded 89.
RANGE: 0 to 89
N Mean Std. Deviation
Total 2116 39.98 13.199
Missing 1916

4) WRKSLF
(Are/were) you self employed or (do/did) you work for someone else?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3925 1.89 0.313
1) Self-employed 432 11.0
2) Someone else 3493 89.0
Missing 107

5) PRESTG10
Prestige of respondent's occupation
RANGE: 16 to 80
N Mean Std. Deviation
Total 3873 46.544 13.811

6) MARITAL
Are you currently - married, widowed, divorced, separated, or have you never been married?
RANGE: 1 to 5
N Mean Std. Deviation
Total 4023 2.438 1.655
1) Married 1999 49.7
2) Widowed 301 7.5
3) Divorced 655 16.3
4) Separated 96 2.4
5) Never married 972 24.2
Missing 9

7) DIVORCE
IF CURRENTLY MARRIED OR WIDOWED: Have you ever been divorced or legally separated?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2300 1.737 0.44
1) Yes 605 26.3
2) No 1695 73.7

Missing 1732

8) WIDOWED
IF CURRENTLY MARRIED, SEPARATED, OR DIVORCED: Have you ever been widowed?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2752 1.983 0.131
1) Yes 48 1.7
2) No 2704 98.3
Missing 1280

9) PAWRKSLF
Was your [father/stepfather/male relative you were living with when you were 16] an employee, self-employed without employees, or self-employed with employees?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3349 1.758 0.428
1) Self-employed 810 24.2
2) Someone else 2539 75.8
Missing 683

10) PAPRES10
Prestige of respondent's father's occupation
RANGE: 16 to 80
N Mean Std. Deviation
Total 3349 45.157 13.148

11) MAWRKSLF
At this job, was [mother/stepmother/female relative you were living with when you were 16] an employee, self-employed without employees, or self-employed with employees?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2797 1.897 0.304
1) Self-employed 287 10.3
2) Someone else 2510 89.7

Missing 1235

12) MAPRES10
Prestige of respondent's mother's occupation
RANGE: 16 to 80
N Mean Std. Deviation
Total 2767 42.664 13.168

13) SIBS
How many brothers and sisters did you have? Please count those born alive, but no longer living, as well as those alive now. Also include stepbrothers and stepsisters, and children adopted by your parents.
RANGE: 0 to 35
N Mean Std. Deviation
Total 3968 3.13 2.646
Missing 64

14) CHILDS
How many children have you ever had? Please count all that were born alive at any time (including any you had from a previous marriage).
RANGE: 0 to 8
N Mean Std. Deviation
Total 3983 1.7 1.526
0) None 1163 29.2
1) One 646 16.2
2) Two 1152 28.9
3) Three 578 14.5
4) Four 277 7.0
5) Five 79 2.0
6) Six 52 1.3
7) Seven 17 0.4
8) Eight or more 19 0.5
Missing 49

15) AGE
Respondent's age
RANGE: 18 to 89

N Mean Std. Deviation
Total 3699 52.165 17.233
Missing 333

16) AGEKDBRN
How old were you when your first child was born?
RANGE: 9 to 57
N Mean Std. Deviation
Total 2803 25.47 6.192

17) EDUC
Respondent's education
RANGE: 0 to 20
N Mean Std. Deviation
Total 3966 14.769 2.8
Missing 66

18) PAEDUC
What is the highest grade in elementary school or high school that your father finished and got credit for?
RANGE: 0 to 20
N Mean Std. Deviation
Total 3090 12.546 3.809
Missing 942

19) MAEDUC
What is the highest grade in elementary school or high school that your mother finished and got credit for?
RANGE: 0 to 20
N Mean Std. Deviation
Total 3613 12.504 3.294
Missing 419

20) DEGREE
Respondent's degree
RANGE: 0 to 4
N Mean Std. Deviation

Total 4009 2.116 1.283
0) Less than high school 246 6.1
1) High school 1597 39.8
2) Associate/junior degree 370 9.2
3) Bachelor's 1036 25.8
4) Graduate 760 19.0
Missing 23

21) SEX
Respondent's sex
RANGE: 1 to 2
N Mean Std. Deviation
Total 3940 1.559 0.497
1) Male 1736 44.1
2) Female 2204 55.9
Missing 92

22) RACE
What race do you consider yourself?
RANGE: 1 to 3
N Mean Std. Deviation
Total 3978 1.32 0.649
1) White 3110 78.2
2) Black 4 63 11.6
3) Other 405 10.2
Missing 54

23) RES16
Which of these categories comes closest to the type of place you were living in when you were 16 years old?
RANGE: 1 to 6
N Mean Std. Deviation
Total 4029 3.824 1.5
1) In open country but not on a farm 406 10.1
2) Farm 204 5.1
3) In a small city or town (under 50,000) 1208 30.0
4) In a medium-size city (50,000-250,000) 777 19.3

5) In a suburb near a large city 742 18.4
6) In a large city (over 250,000) 692 17.2
Missing 3

24) REG16
In what state or foreign country were you living when you were 16 years old?
RANGE: 1 to 9
N Mean Std. Deviation
Total 4018 5.128 2.621
1) New England 176 4.4
2) Middle Atlantic 543 13.5
3) East North Central 812 20.2
4) West North Central 335 8.3
5) South Atlantic 536 13.3
6) East South Atlantic 235 5.8
7) West South Central 356 8.9
8) Mountain 242 6.0
9) Pacific 783 19.5
Missing 14

25) MOBILE16
IF STATE NAMED IS SAME STATE R. LIVES IN NOW, ASK MOBILE16: When you were 16 years old, were you living in this same (city/town/county)?
RANGE: 1 to 3
N Mean Std. Deviation
Total 3608 2.039 0.8
1) Same state, same city 1087 30.1
2) Same state, different city 1294 35.9
3) Different state 1227 34.0
Missing 424

26) MAWRKGRW
Did your mother ever work for pay for as long as a year, while you were growing up?
RANGE: 1 to 2
N Mean Std. Deviation

Total 3770 1.242 0.429
1) Yes 2856 75.8
2) No 914 24.2
Missing 262

27) INCOM16
Thinking about the time when you were 16 years old, compared with American families in general then, would you say your family income was: far below average, below average, average, above average, or far above average?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3826 2.739 0.952
1) Far below average 421 11.0
2) Below average 1013 26.5
3) Average 1625 42.5
4) Above average 679 17.7
5) Far above average 88 2.3
Missing 206

28) BORN
Were you born in this country?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3960 1.112 0.316
1) Yes 3516 88.8
2) No 444 11.2
Missing 72

29) GRANBORN
(Were all of your four grandparents born in this country?...) IF NO: how many were born outside the United States?
RANGE: 0 to 4
N Mean Std. Deviation
Total 3633 0.979 1.52
0) None 2379 65.5
1) One 220 6.1
2) Two 350 9.6

3) Three 99 2.7
4) Four 585 16.1
Missing 399

30) MABORN
Was(your mother/ your stepmother/ the female relative you were living with at 16) born in this country?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3939 1.159 0.366
1) Yes 3312 84.1
2) No 627 15.9
Missing 93

31) PABORN
Was(your father/your stepfather/the male relative you were living with at 16) born in this country?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3918 1.166 0.372
1) Yes 3269 83.4
2) No 649 16.6
Missing 114

32) SEXBIRTH1
Was your sex recorded as male or female at birth?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3928 1.56 0.497
1) Male 1730 44.0
2) Female 2198 56.0
Missing 104

33) REGION
Region of interview
RANGE: 1 to 9
N Mean Std. Deviation

Total 4032 5.192 2.454
1) New England 203 5.0
2) Middle Atlantic 414 10.3
3) East North Central 676 16.8
4) West North Central 314 7.8
5) South Atlantic 800 19.8
6) East South Atlantic 270 6.7
7) West South Central 426 10.6
8) Mountain 345 8.6
9) Pacific 584 14.5

34) PARTYID
Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?
RANGE: 0 to 7
N Mean Std. Deviation
Total 4000 2.776 2.135
0) Strong Democrat 822 20.6
1) Not very strong Democrat 541 13.5
2) Independent, close to Democrat 471 11.8
3) Independent (neither, no response) 817 20.4
4) Independent, close to Republican 327 8.2
5) Not very strong Republican 384 9.6
6) Strong Republican 524 13.1
7) Other party 114 2.9
Missing 32

35) VOTE16
In 2016, you remember that Hillary Clinton ran for president on the Democratic ticket against Donald Trump for the Republicans. Do you remember for sure whether or not you voted in that election?
RANGE: 1 to 3
N Mean Std. Deviation
Total 3703 1.28 0.567
1) Voted 2886 77.9
2) Did not vote 596 16.1
3) Ineligible 221 6.0

Missing 329

36) PRES16
Did you vote for Hillary Clinton or Donald Trump?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2764 1.551 0.69
1) Clinton 1509 54.6
2) Trump 1037 37.5
3) Other candidate (please specify) 169 6.1
4) Didn't vote for President 49 1.8
Missing 1268

37) IF16WHO
Who would you have voted for, for president, if you had voted?
RANGE: 1 to 3
N Mean Std. Deviation
Total 761 1.859 0.809
1) Clinton 310 40.7
2) Trump 248 32.6
3) Other 203 26.7
Missing 3271

38) POLVIEWS
We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal - point one - to extremely conservative - point seven. Where would you place yourself on this scale?
RANGE: 1 to 7
N Mean Std. Deviation
Total 3964 3.968 1.536
1) Extremely liberal 207 5.2
2) Liberal 623 15.7
3) Slightly liberal 490 12.4
4) Moderate, middle of the road 1377 34.7
5) Slightly conservative 476 12.0

6) Conservative 617 15.6
7) Extremely conservative 174 4.4
Missing 68

39) NATSPAC
(Let's begin with some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Space exploration program
RANGE: 1 to 3
N Mean Std. Deviation
Total 1969 1.975 0.685
1) Too little 488 24.8
2) About right 1043 53.0
3) Too much 438 22.2
Missing 2063

40) NATENVIR
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Improving and protecting the environment
RANGE: 1 to 3
N Mean Std. Deviation
Total 1979 1.401 0.658
1) Too little 1376 69.5
2) About right 413 20.9
3) Too much 190 9.6
Missing 2053

41) NATHEAL
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Improv-

ing and protecting the nation's health
RANGE: 1 to 3
N Mean Std. Deviation
Total 1966 1.418 0.644
1) Too little 1314 66.8
2) About right 483 24.6
3) Too much 169 8.6
Missing 2066

42) NATCITY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Solving the problems of big cities
RANGE: 1 to 3
N Mean Std. Deviation
Total 1949 1.661 0.763
1) Too little 1009 51.8
2) About right 591 30.3
3) Too much 349 17.9
Missing 2083

43) NATCRIME
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Halting the rising crime rate
RANGE: 1 to 3
N Mean Std. Deviation
Total 1969 1.449 0.644
1) Too little 1249 63.4
2) About right 556 28.2
3) Too much 164 8.3
Missing 2063
44) NATDRUG

(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Dealing with drug addiction
RANGE: 1 to 3
N Mean Std. Deviation
Total 1964 1.466 0.647
1) Too little 1216 61.9
2) About right 581 29.6
3) Too much 167 8.5
Missing 2068

45) NATEDUC
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Improving the nation's education system
RANGE: 1 to 3
N Mean Std. Deviation
Total 1976 1.349 0.618
1) Too little 1439 72.8
2) About right 384 19.4
3) Too much 153 7.7
Missing 2056

46) NATRACE
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Improving the condition of Blacks
RANGE: 1 to 3
N Mean Std. Deviation
Total 1959 1.654 0.759
1) Too little 1020 52.1

2) About right 596 30.4
3) Too much 343 17.5
Missing 2073

47) NATARMS
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) The military, armaments, and defense
RANGE: 1 to 3
N Mean Std. Deviation

Total 1974 2.068 0.764
1) Too little 513 26.0
2) About right 814 41.2
3) Too much 647 32.8
Missing 2058

48) NATAID
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) For-eign aid
RANGE: 1 to 3
N Mean Std. Deviation
Total 1969 2.472 0.655
1) Too little 177 9.0
2) About right 686 34.8
3) Too much 1106 56.2
Missing 2063

49) NATFARE
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending

too much money on it, too little money, or about the right amount.) Welfare
RANGE: 1 to 3
N Mean Std. Deviation
Total 1970 2.023 0.797
1) Too little 604 30.7
2) About right 717 36.4
3) Too much 649 32.9
Missing 2062

50) NATROAD
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) High-ways and bridges
RANGE: 1 to 3
N Mean Std. Deviation
Total 4004 1.485 0.599
1) Too little 2281 57.0
2) About right 1504 37.6
3) Too much 219 5.5
Missing 28

51) NATSOC
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Social Security
RANGE: 1 to 3
N Mean Std. Deviation
Total 4002 1.478 0.59
1) Too little 2286 57.1
2) About right 1518 37.9
3) Too much 198 4.9 Missing 30

52) NATMASS

(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Mass transportation
RANGE: 1 to 3
N Mean Std. Deviation
Total 3996 1.675 0.62
1) Too little 1628 40.7
2) About right 2039 51.0
3) Too much 329 8.2 Missing 36

53) NATPARK
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Parks and recreation
RANGE: 1 to 3
N Mean Std. Deviation
Total 4002 1.684 0.545
1) Too little 1428 35.7
2) About right 2412 60.3
3) Too much 162 4.0 Missing 30
54) NATCHLD
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Assis-tance for childcare
RANGE: 1 to 3
N Mean Std. Deviation
Total 3984 1.518 0.641
1) Too little 2244 56.3
2) About right 1418 35.6
3) Too much 322 8.1
Missing 48

55) NATSCI
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Supporting scientific research
RANGE: 1 to 3
N Mean Std. Deviation
Total 3996 1.62 0.622
1) Too little 1820 45.5
2) About right 1873 46.9
3) Too much 303 7.6
Missing 36

56) NATENRGY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Developing alternative energy sources
RANGE: 1 to 3
N Mean Std. Deviation
Total 4008 1.54 0.678
1) Too little 2265 56.5
2) About right 1321 33.0
3) Too much 422 10.5
Missing 24

57) NATSPACY
(Let's begin with some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Space exploration
RANGE: 1 to 3
N Mean Std. Deviation

Total 2030 2.018 0.687
1) Too little 461 22.7
2) About right 1071 52.8
3) Too much 498 24.5
Missing 2002

58) NATENVIY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Improving and protecting the environment
RANGE: 1 to 3
N Mean Std. Deviation
Total 2040 1.399 0.648
1) Too little 1410 69.1
2) About right 447 21.9
3) Too much 183 9.0
Missing 1992

59) NATHEALY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Health
RANGE: 1 to 3
N Mean Std. Deviation
Total 2025 1.47 0.711
1) Too little 1333 65.8
2) About right 432 21.3
3) Too much 260 12.8
Missing 2007

60) NATCITYY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending

too much money on it, too little money, or about the right amount.) Assistance to big cities
RANGE: 1 to 3
N Mean Std. Deviation
Total 2014 2.058 0.76
1) Too little 526 26.1
2) About right 846 42.0
3) Too much 642 31.9
Missing 2018

61) NATCRIMY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Law enforcement
RANGE: 1 to 3
N Mean Std. Deviation
Total 2037 1.806 0.778
1) Too little 853 41.9
2) About right 727 35.7
3) Too much 457 22.4
Missing 1995

62) NATDRUGY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Drug rehabilitation
RANGE: 1 to 3
N Mean Std. Deviation
Total 2023 1.495 0.671
1) Too little 1225 60.6
2) About right 595 29.4
3) Too much 203 10.0
Missing 2009

63) NATEDUCY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Education
RANGE: 1 to 3
N Mean Std. Deviation
Total 2031 1.321 0.607
1) Too little 1532 75.4
2) About right 346 17.0
3) Too much 153 7.5
Missing 2001

64) NATRACEY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Assistance to Blacks
RANGE: 1 to 3
N Mean Std. Deviation
Total 2006 1.728 0.762
1) Too little 930 46.4
2) About right 692 34.5
3) Too much 384 19.1
Missing 2026

65) NATARMSY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.)
National defense
RANGE: 1 to 3
N Mean Std. Deviation

Total 2033 2.046 0.74
1) Too little 512 25.2
2) About right 915 45.0
3) Too much 606 29.8
Missing 1999

66) NATAIDY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Assistance to other countries
RANGE: 1 to 3
N Mean Std. Deviation
Total 2031 2.501 0.67
1) Too little 202 9.9
2) About right 609 30.0
3) Too much 1220 60.1 Missing 2001

67) NATFAREY
(We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount.) Assistance to the poor
RANGE: 1 to 3
N Mean Std. Deviation
Total 2038 1.406 0.648
1) Too little 1392 68.3
2) About right 464 22.8
3) Too much 182 8.9
Missing 1994

68) EQWLTH
Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Oth-

ers think that the government should not concern itself with reducing this income difference between the rich and the poor. Here is a card with a scale from one to seven. Think of a score of one as meaning that the government ought to reduce the income differences between rich and poor, and a score of seven meaning that the government should not concern itself with reducing income differences. What score between one and seven comes closest to the way you feel?
RANGE: 1 to 7
N Mean Std. Deviation
Total 2661 3.385 2.2
1) The government should reduce income differences 882 33.1
2) 242 9.1
3) 325 12.2
4) 389 14.6
5) 249 9.4
6) 153 5.7
7) The government should not concern itself with reducing income differences 421 15.8
Missing 1371
69) TAX
Do you consider the amount of federal income tax which you have to pay as too high, about right, or too low?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2634 1.467 0.552
1) Too high 1478 56.1
2) About right 1082 41.1
3) Too much 74 2.8
Missing 1398

70) SPKATH
There are always some people whose ideas are considered bad or dangerous by other people. For instance, somebody who is against all churches and religion... If such a person wanted to make a speech in your (city/town/community) against churches and religion, should he be allowed to speak, or not?
RANGE: 1 to 2

N Mean Std. Deviation
Total 1312 1.186 0.389
1) Yes, allowed to speak 1068 81.4
2) Not allowed 244 18.6
Missing 2720

71) COLATH
(There are always some people whose ideas are considered bad or dangerous by other people. For instance, somebody who is against all churches and religion...) Should such a person be allowed to teach in a college or university, or not?
RANGE: 4 to 5
N Mean Std. Deviation
Total 2647 4.304 0.46
4) Yes, allowed to teach 1842 69.6
5) Not allowed 805 30.4
Missing 1385

72) LIBATH
(There are always some people whose ideas are considered bad or dangerous by other people. For instance, somebody who is against all churches and religion...) If some people in your community suggested that a book he wrote against churches and religion should be taken out of your public library, would you favor removing this book, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1310 1.844 0.363
1) Remove 204 15.6
2) Not remove 1106 84.4
Missing 2722

73) SPKRAC
Or consider a person who believes that Blacks are genetically inferior. If such a person wanted to make a speech in your community claiming that Blacks are inferior, should he be allowed to speak, or not?
RANGE: 1 to 2
N Mean Std. Deviation

Total 1310 1.512 0.5
1) Yes, allowed to speak 639 48.8
2) Not allowed 671 51.2
Missing 2722

74) COLRAC
(Or consider a person who believes that Blacks are genetically inferior...) Should such a person be allowed to teach in a college or university, or not?
RANGE: 4 to 5
N Mean Std. Deviation
Total 2633 4.672 0.47
4) Yes, allowed to teach 864 32.8
5) Not allowed 1769 67.2
Missing 1399

75) LIBRAC
(Or consider a person who believes that Blacks are genetically inferior...) If some people in your community suggested that a book he wrote which said Blacks are inferior should be taken out of your public library, would you favor removing this book, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1301 1.613 0.487
1) Remove 503 38.7
2) Not remove 798 61.3
Missing 2731

76) SPKCOM
Now, we would like to ask you some questions about a man who admits he is a Communist. Suppose this admitted Communist wanted to make a speech in your community. Should he be allowed to speak, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1305 1.266 0.442
1) Yes, allowed to speak 958 73.4
2) Not allowed 347 26.6

Missing 2727

77) COLCOM
(Now, we would like to ask you some questions about a man who admits he is a Communist...) Suppose he is teaching in a college. Should he be fired, or not?
RANGE: 4 to 5
N Mean Std. Deviation
Total 1293 4.704 0.457
4) Yes, fired 383 29.6
5) Not fired 910 70.4
Missing 2739

78) LIBCOM
(Now, we would like to ask you some questions about a man who admits he is a Communist...) Suppose he wrote a book which is in your public library. Somebody in your community suggests that the book should be removed from the library. Would you favor removing it, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1303 1.787 0.41
1) Remove 278 21.3
2) Not remove 1025 78.7
Missing 2729

79) SPKMIL
Consider a person who advocates doing away with elections and letting the military run the country. If such a person wanted to make a speech in your community, should he be allowed to speak, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1301 1.381 0.486
1) Yes, allowed to speak 805 61.9
2) Not allowed 496 38.1
Missing 2731

80) COLMIL

(Consider a person who advocates doing away with elections and letting the military run the country...) Should such a person be allowed to teach in a college or university, or not?
RANGE: 4 to 5
N Mean Std. Deviation
Total 2637 4.529 0.499
4) Yes, allowed to teach 1243 47.1
5) Not allowed 1394 52.9
Missing 1395

81) LIBMIL
Suppose he wrote a book advocating doing away with elections and letting the military run the country. Somebody in your community suggests that the book be removed from the public library. Would you favor removing it, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1302 1.699 0.459
1) Remove 392 30.1
2) Not remove 910 69.9
Missing 2730

82) SPKHOMO
And what about a man who admits that he is homosexual... Suppose this admitted homosexual wanted to make a speech in your community. Should he be allowed to speak, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1304 1.083 0.276
1) Yes, allowed to speak 1196 91.7
2) Not allowed 108 8.3
Missing 2728

83) COLHOMO
Should such a person be allowed to teach in a college or university, or not?
RANGE: 4 to 5

N Mean Std. Deviation
Total 2656 4.069 0.253
4) Yes, allowed to teach 2473 93.1
5) Not allowed 183 6.9
Missing 1376

84) LIBHOMO
(And what about a man who admits that he is homosexual...) If some people in your community suggested that a book he wrote in favor of homosexuality should be taken out of your public library, would you favor removing this book, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1308 1.867 0.34
1) Remove 174 13.3
2) Not Remove 1134 86.7
Missing 2724

85) SPKMSLM
Now consider a Muslim clergyman who preaches hatred of the United States. If such a person wanted to make a speech in your community preaching hatred of the United States, should he be allowed to speak, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1307 1.519 0.5
1) Yes, allowed 629 48.1
2) Not allowed 678 51.9 Missing 2725
86) COLMSLM
Should such a person be allowed to teach in a college or university, or not?
RANGE: 4 to 5
N Mean Std. Deviation
Total 2646 4.67 0.47
4) Yes, allowed to teach 874 33.0
5) Not allowed 1772 67.0
Missing 1386

87) LIBMSLM
(Now consider a Muslim clergyman who preaches hatred of the United States...) If some people in your community suggested that a book he wrote which preaches hatred of the United States should be taken out of your public library, would you favor removing this book, or not?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1304 1.564 0.496
1) Remove 569 43.6
2) Not remove 735 56.4
Missing 2728

88) CAPPUN
Do you favor or oppose the death penalty for persons convicted of murder?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3957 1.438 0.496
1) Favor 2222 56.2
2) Oppose 1735 43.8 Missing 75

89) GUNLAW
Would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3992 1.327 0.469
1) Favor 2686 67.3
2) Oppose 1306 32.7
Missing 40

90) RACLIVE
Are there any ('Whites' for Black respondents, 'Blacks' for non-Black respondents) living in this neighborhood now?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3587 1.176 0.381
1) Yes 2955 82.4

2) No 632 17.6
Missing 445

91) RELIG
What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?
RANGE: 1 to 13
N Mean Std. Deviation
Total 3951 2.774 2.358
1) Protestant 1590 40.2
2) Catholic 824 20.9
3) Jewish 75 1.9
4) None 1121 28.4
5) Other 55 1.4
6) Buddhism 47 1.2
7) Hinduism 30 0.8
8) Other Eastern religions 2 0.1
9) Muslim/Islam 25 0.6
10) Orthodox Christian 37 0.9
11) Christian 124 3.1
12) Native American 3 0.1
13) Inter/nondenominational 18 0.5
Missing 81

92) FUND
Fundamentalism/liberalism of respondent's religion
RANGE: 1 to 3
N Mean Std. Deviation
Total 3742 2.255 0.719
1) Fundamentalist 612 16.4
2) Moderate 1565 41.8
3) Liberal 1565 41.8
Missing 290

93) ATTEND
How often do you attend religious services? (USE CATEGORIES AS PROBES, IF NECESSARY.)

RANGE: 0 to 8

N Mean Std. Deviation

Total 3962 2.853 2.756

0) Never 1178 29.7

1) Less than once a year 565 14.3

2) About once or twice a year 453 11.4

3) Several times a year 403 10.2

4) About once a month 122 3.1

5) Two to three times a month 200 5.0

6) Nearly every week 331 8.4

7) Every week 532 13.4

8) Several times a week 178 4.5

Missing 70

94) PRAY

About how often do you pray? (USE CATEGORIES AS PROBES.)

RANGE: 1 to 6

N Mean Std. Deviation

Total 3955 3.255 1.968

1) Several times a day 1139 28.8

2) Once a day 656 16.6

3) Several times a week 542 13.7

4) Once a week 175 4.4

5) Less than once a week 560 14.2

6) Never 883 22.3

Missing 77

95) SPREL

What is your (SPOUSE'S) religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?

RANGE: 1 to 7

N Mean Std. Deviation

Total 1639 2.422 1.451

1) Protestant 590 36.0

2) Catholic 465 28.4

3) Jewish 27 1.6

4) None 486 29.7

5) Other 26 1.6
6) Buddhism 20 1.2
7) Hinduism 25 1.5
Missing 2393

96) AFFRMACT
Some people say that because of past discrimination, Blacks should be given preference in hiring and promotion. Others say that such preference in hiring and promotion of Blacks is wrong because it discriminates against Whites. What about your opinion? Are you for or against preferential hiring and promotion of Blacks? IF FAVORS: Do you favor preference in hiring and promotion strongly or not strongly? IF OPPOSES: Do you oppose preference in hiring and promotion strongly or not strongly?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2628 3.063 1.041
1) Strongly favors 349 13.3
2) Not strongly favors 299 11.4
3) Not strongly opposes 818 31.1
4) Strongly opposes 1162 44.2
Missing 1404

97) WRKWAYUP
Do you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly with the following statement (HAND CARD TO RESPONDENT) Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without special favors.
RANGE: 1 to 5
N Mean Std. Deviation
Total 2688 2.878 1.418
1) Agree strongly 609 22.7
2) Agree somewhat 536 19.9
3) Neither agree nor disagree 640 23.8
4) Disagree somewhat 380 14.1
5) Disagree strongly 523 19.5
Missing 1344

98) HAPPY
Taken all together, how would you say things are these days - would you say that you are very happy, pretty happy, or not too happy?
RANGE: 1 to 3
N Mean Std. Deviation
Total 4014 2.035 0.651
1) Very happy 783 19.5
2) Pretty happy 2308 57.5
3) Not too happy 923 23.0
Missing 18

99) HAPMAR
(IF CURRENTLY MARRIED, ASK HAPMAR) Taking things all together, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1986 1.427 0.565
1) Very happy 1211 61.0
2) Pretty happy 701 35.3
3) Not too happy 74 3.7
Missing 2046

100) HEALTH
Would you say your own health, in general, is excellent, good, fair, or poor?
RANGE: 1 to 4
N Mean Std. Deviation
Total 4023 2.06 0.739
1) Excellent 835 20.8
2) Good 2264 56.3
3) Fair 773 19.2
4) Poor 151 3.8
Missing 9

101) LIFE

In general, do you find life exciting, pretty routine, or dull?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2669 1.691 0.562
1) Exciting 962 36.0
2) Routine 1571 58.9
3) Dull 136 5.1
Missing 1363

102) CONFINAN
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Banks and financial institutions
RANGE: 1 to 3
N Mean Std. Deviation
Total 2660 2.039 0.633
1) A great deal 482 18.1
2) Only some 1592 59.8
3) Hardly any 586 22.0
Missing 1372

103) CONBUS
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Major companies
RANGE: 1 to 3
N Mean Std. Deviation
Total 2656 2.048 0.62
1) A great deal 450 16.9
2) Only some 1628 61.3
3) Hardly any 578 21.8
Missing 1376

104) CONCLERG
(I am going to name some institutions in this country. As far as the peo-

ple running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Organized religion
RANGE: 1 to 3
N Mean Std. Deviation
Total 2650 2.183 0.669
1) A great deal 394 14.9
2) Only some 1377 52.0
3) Hardly any 879 33.2
Missing 1382

105) CONEDUC
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Education
RANGE: 1 to 3
N Mean Std. Deviation
Total 2658 2.052 0.619
1) A great deal 443 16.7
2) Only some 1633 61.4
3) Hardly any 582 21.9
Missing 1374

106) CONFED
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Executive branch of the federal government
RANGE: 1 to 3
N Mean Std. Deviation
Total 2658 2.318 0.686
1) A great deal 336 12.6
2) Only some 1140 42.9
3) Hardly any 1182 44.5
Missing 1374

107) CONLABOR
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Organized labor
RANGE: 1 to 3
N Mean Std. Deviation
Total 2648 2.151 0.594
1) A great deal 297 11.2
2) Only some 1655 62.5
3) Hardly any 696 26.3
Missing 1384

108) CONPRESS
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Press
RANGE: 1 to 3
N Mean Std. Deviation
Total 2654 2.358 0.679
1) A great deal 306 11.5
2) Only some 1093 41.2
3) Hardly any 1255 47.3
Missing 1378

109) CONMEDIC
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Medicine
RANGE: 1 to 3
N Mean Std. Deviation
Total 2662 1.69 0.631
1) A great deal 1070 40.2
2) Only some 1346 50.6
3) Hardly any 246 9.2

Missing 1370

110) CONTV
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) TV
RANGE: 1 to 3
N Mean Std. Deviation
Total 2660 2.341 0.617
1) A great deal 207 7.8
2) Only some 1340 50.4
3) Hardly any 1113 41.8
Missing 1372

111) CONJUDGE
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) U.S. Supreme Court
RANGE: 1 to 3
N Mean Std. Deviation
Total 2662 1.943 0.676
1) A great deal 689 25.9
2) Only some 1437 54.0
3) Hardly any 536 20.1
Missing 1370

112) CONSCI
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Scientific community
RANGE: 1 to 3
N Mean Std. Deviation
Total 2654 1.563 0.616
1) A great deal 1337 50.4

2) Only some 1141 43.0
3) Hardly any 176 6.6
Missing 1378

113) CONLEGIS
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Congress
RANGE: 1 to 3
N Mean Std. Deviation
Total 2661 2.485 0.597
1) A great deal 141 5.3
2) Only some 1089 40.9
3) Hardly any 1431 53.8
Missing 1371

114) CONARMY
(I am going to name some institutions in this country. As far as the people running this institution are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?) Military
RANGE: 1 to 3
N Mean Std. Deviation
Total 2656 1.624 0.653
1) A great deal 1254 47.2
2) Only some 1147 43.2
3) Hardly any 255 9.6
Missing 1376

115) OBEY
(If you had to choose, which thing on this list would you pick as the most important for a child to learn to prepare him or her for life? Which comes next in importance? Which comes third? Which comes fourth?) To obey
RANGE: 1 to 5
N Mean Std. Deviation
Total 2573 3.875 1.027

1) First 127 4.9
2) Second 160 6.2
3) Third 294 11.4
4) Fourth 1318 51.2
5) Fifth 674 26.2
Missing 1459

116) POPULAR
(If you had to choose, which thing on this list would you pick as the most important for a child to learn to prepare him or her for life? Which comes next in importance? Which comes third? Which comes fourth?) To be well-liked or popular
RANGE: 1 to 5
N Mean Std. Deviation
Total 2573 4.649 0.64
1) First 21 0.8
2) Second 20 0.8
3) Third 48 1.9
4) Fourth 663 25.8
5) Fifth 1821 70.8
Missing 1459

117) THNKSELF
(If you had to choose, which thing on this list would you pick as the most important for a child to learn to prepare him or her for life? Which comes next in importance? Which comes third? Which comes fourth?) To think for himself or herself
RANGE: 1 to 5
N Mean Std. Deviation
Total 2573 1.93 1.111
1) First 1268 49.3
2) Second 581 22.6
3) Third 411 16.0
4) Fourth 262 10.2
5) Fifth 51 2.0
Missing 1459

118) WORKHARD
(If you had to choose, which thing on this list would you pick as the most important for a child to learn to prepare him or her for life? Which comes next in importance? Which comes third? Which comes fourth?) To work hard
RANGE: 1 to 5
N Mean Std. Deviation
Total 2573 2.201 0.865
1) First 624 24.3
2) Second 934 36.3
3) Third 894 34.7
4) Fourth 116 4.5
5) Fifth 5 0.2
Missing 1459

119) HELPOTH
(If you had to choose, which thing on this list would you pick as the most important for a child to learn to prepare him or her for life? Which comes next in importance? Which comes third? Which comes fourth?) To help others when they need help
RANGE: 1 to 5
N Mean Std. Deviation
Total 2573 2.345 0.926
1) First 533 20.7
2) Second 878 34.1
3) Third 926 36.0
4) Fourth 214 8.3
5) Fifth 22 0.9
Missing 1459

120) SOCREL
(Would you use this card and tell me which answer comes closest to how often you do the following things...) Spend a social evening with relatives?
RANGE: 1 to 7
N Mean Std. Deviation
Total 2705 3.912 1.615
1) Almost daily 191 7.1

2) Once or twice a week 438 16.2
3) Several times a month 478 17.7
4) About once a month 481 17.8
5) Several times a year 694 25.7
6) About once a year 275 10.2
7) Never 148 5.5
Missing 1327

121) SOCOMMUN
(Would you use this card and tell me which answer comes closest to how often you do the following things...) Spend a social evening with someone who lives in your neighborhood?
RANGE: 1 to 7
N Mean Std. Deviation
Total 2705 5.277 1.78
1) Almost daily 39 1.4
2) Once or twice a week 235 8.7
3) Several times a month 272 10.1
4) About once a month 322 11.9
5) Several times a year 438 16.2
6) About once a year 323 11.9
7) Never 1076 39.8
Missing 1327

122) SOCFREND
(Would you use this card and tell me which answer comes closest to how often you do the following things...) Spend a social evening with friends who live outside the neighborhood?
RANGE: 1 to 7
N Mean Std. Deviation
Total 2701 4.473 1.508
1) Almost daily 34 1.3
2) Once or twice a week 253 9.4
3) Several times a month 467 17.3
4) About once a month 554 20.5
5) Several times a year 777 28.8
6) About once a year 273 10.1

7) Never 343 12.7
Missing 1331

123) SOCBAR
(Would you use this card and tell me which answer comes closest to how often you do the following things...) Go to a bar or tavern?
RANGE: 1 to 7
N Mean Std. Deviation
Total 2702 5.688 1.465
1) Almost daily 5 0.2
2) Once or twice a week 91 3.4
3) Several times a month 190 7.0
4) About once a month 271 10.0
5) Several times a year 513 19.0
6) About once a year 461 17.1
7) Never 1171 43.3
Missing 1330

124) JOBLOSE
Thinking about the next 12 months, how likely do you think it is that you will lose your job or be laid off--very likely, fairly likely, not too likely, or not at all likely?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1495 3.395 0.706
1) Very likely 34 2.3
2) Fairly likely 92 6.2
3) Not too likely 619 41.4
4) Not likely 750 50.2
Missing 2537

125) JOBFIND
About how easy would it be for you to find a job with another employer with approximately the same income and fringe benefits you now have? Would you say very easy, somewhat easy, or not easy at all?
RANGE: 1 to 3
N Mean Std. Deviation

Total 1489 2.269 0.703
1) Very easy 221 14.8
2) Somewhat easy 647 43.5
3) Not easy 621 41.7
Missing 2543

126) SATJOB
On the whole, how satisfied are you with the work you do—would you say you are very satisfied, moderately satisfied, a little dissatisfied, or very dissatisfied?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2734 1.755 0.809
1) Very satisfied 1198 43.8
2) Moderately satisfied 1119 40.9
3) A little dissatisfied 305 11.2
4) Very dissatisfied 112 4.1
Missing 1298

127) RICHWORK
(IF CURRENTLY WORKING OR TEMPORARILY NOT AT WORK, ASK RICHWORK.) If you were to get enough money to live as comfortably as you would like for the rest of your life, would you continue to work or would you stop working?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1624 1.382 0.486
1) Continue to work 1004 61.8
2) Stop working 620 38.2
Missing 2408

128) CLASS
If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?
RANGE: 1 to 4
N Mean Std. Deviation
Total 4018 2.495 0.713

1) Lower class 349 8.7
2) Working class 1501 37.4
3) Middle class 1999 49.8
4) Upper class 169 4.2
Missing 14

129) RANK
(In our society there are groups which tend to be toward the top and those that are toward the bottom. Here we have a scale that runs from top to bottom...) Where would you put yourself on this scale?
RANGE: 1 to 10
N Mean Std. Deviation
Total 1966 4.58 1.718
1) Top 76 3.9
2) 89 4.5
3) 361 18.4
4) 365 18.6
5) 685 34.8
6) 142 7.2
7) 134 6.8
8) 64 3.3
9) 20 1.0
10) Bottom 30 1.5
Missing 2066

130) SATFIN
We are interested in how people are getting along financially these days. So far as you and your family are concerned, would you say that you are pretty well satisfied with your present financial situation, more or less satisfied, or not satisfied at all?
RANGE: 1 to 3
N Mean Std. Deviation
Total 4016 1.927 0.739
1) Pretty well satisfied 1254 31.2
2) More or less satisfied 1800 44.8
3) Not satisfied at all 962 24.0
Missing 16

131) FINALTER
During the last few years, has your financial situation been getting better, worse, or has it stayed the same?
RANGE: 1 to 3
N Mean Std. Deviation
Total 4021 1.991 0.894
1) Getting better 1623 40.4
2) Getting worse 811 20.2
3) Stayed the same 1587 39.5
Missing 11

132) FINRELA
Compared with American families in general, would you say your family income is far below average, below average, average, above average, or far above average? (PROBE: Just your best guess.)
RANGE: 1 to 5
N Mean Std. Deviation
Total 4020 2.938 0.962
1) Far below average 287 7.1
2) Below average 978 24.3
3) Average 1603 39.9
4) Above average 1000 24.9
5) Far above average 152 3.8
Missing 12

133) WKSUB
We have some more questions about [your/your spouse's] job. At work, [do you/does your spouse] have a supervisor to whom [you/he or she] are directly responsible?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2500 1.169 0.375
1) Yes 2077 83.1
2) No 423 16.9
Missing 1532

134) WKSUBS
IF YES: Does that supervisor have a supervisor to whom he or she is directly responsible?
RANGE: 3 to 4
N Mean Std. Deviation
Total 2049 3.113 0.317
3) Yes 1817 88.7
4) No 232 11.3
Missing 1983

135) WKSUP
At work, [do you/does your spouse] supervise anyone who is directly responsible to [you/your spouse]?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2548 1.669 0.471
1) Yes 844 33.1
2) No 1704 66.9 Missing 1484

136) UNION
Do you (or your spouse) belong to a labor union? (Who?)
RANGE: 1 to 4
N Mean Std. Deviation
Total 2652 3.679 0.874
1) Yes, respondent belongs 212 8.0
2) Yes, spouse belongs 88 3.3
3) Yes, both belong 39 1.5
4) No, neither belong 2313 87.2
Missing 1380

137) UNION1
Do you (or your spouse or partner) belong to a labor union? (Who?)
RANGE: 1 to 4
N Mean Std. Deviation
Total 2652 3.678 0.867
1) Yes, respondent belongs 202 7.6
2) Yes, spouse or partner belongs 100 3.8

3) Yes, both belong 49 1.8
4) No, neither belong 2301 86.8
Missing 1380

138) PARSOL
Compared to your parents when they were the age you are now, do you think your own standard of living now is much better, somewhat better, about the same, somewhat worse, or much worse than theirs was?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2653 2.404 1.114
1) Much better 656 24.7
2) Somewhat better 835 31.5
3) About the same 703 26.5
4) Somewhat worse 353 13.3
5) Much worse 106 4.0
Missing 1379

139) ABDEFECT
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If there is a strong chance of serious defect in the baby?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1435 1.215 0.411
1) Yes 1126 78.5
2) No 309 21.5
Missing 2597

140) ABNOMORE
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If she is married and does not want any more children?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1424 1.409 0.492
1) Yes 841 59.1

2) No 583 40.9
Missing 2608

141) ABHLTH
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the woman's own health is seriously endangered by the pregnancy?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1431 1.099 0.298
1) Yes 1290 90.1
2) No 141 9.9
Missing 2601

142) ABPOOR
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the family has a very low income and cannot afford any more children?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1425 1.421 0.494
1) Yes 825 57.9
2) No 600 42.1
Missing 2607

143) ABRAPE
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If she becomes pregnant as a result of rape?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1430 1.153 0.36
1) Yes 1211 84.7
2) No 219 15.3 Missing 2602

144) ABSINGLE
(Please tell me whether or not you think it should be possible for a preg-

nant woman to obtain a legal abortion if...) If she is not married and does not want to marry the man?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1425 1.429 0.495
1) Yes 813 57.1
2) No 612 42.9
Missing 2607

145) ABANY
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the woman wants it for any reason?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1328 1.436 0.496
1) Yes 749 56.4
2) No 579 43.6
Missing 2704

146) ABDEFECTG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If there is a strong chance of serious defect in the baby?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1176 1.207 0.406
1) Yes 932 79.3
2) No 244 20.7
Missing 2856

147) ABNOMOREG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If she is married and does not want any more children?
RANGE: 1 to 2
N Mean Std. Deviation

Total 1165 1.467 0.499
1) Yes 621 53.3
2) No 544 46.7
Missing 2867

148) ABHLTHG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the woman's own health is seriously endangered by the pregnancy?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1188 1.106 0.308
1) Yes 1062 89.4
2) No 126 10.6
Missing 2844

149) ABPOORG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the family has a very low income and cannot afford any more children?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1160 1.437 0.496
1) Yes 653 56.3
2) No 507 43.7
Missing 2872

150) ABRAPEG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If she becomes pregnant as a result of rape?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1179 1.173 0.378
1) Yes 975 82.7
2) No 204 17.3
Missing 2853

151) ABSINGLEG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If she is not married and does not want to marry the man?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1158 1.461 0.499
1) Yes 624 53.9
2) No 534 46.1 Missing 2874
152) ABANYG
(Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if...) If the woman wants it for any reason?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1300 1.401 0.49
1) Yes 779 59.9
2) No 521 40.1
Missing 2732

153) PROCHOIC
(We hear a lot of talk these days about abortion. Please indicate to what extent you agree or disagree with each of the following statements.) I consider myself pro-choice.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3551 2.448 1.306
1) Strongly agree 1053 29.7
2) Agree 1004 28.3
3) Neither agree nor disagree 729 20.5
4) Disagree 379 10.7
5) Strongly disagree 386 10.9
Missing 481

154) PROLIFE
(We hear a lot of talk these days about abortion. Please indicate to what

extent you agree or disagree with each of the following statements.) I consider myself pro-life.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3537 2.888 1.329
1) Strongly agree 660 18.7
2) Agree 791 22.4
3) Neither agree nor disagree 941 26.6
4) Disagree 574 16.2
5) Strongly disagree 571 16.1
Missing 495

155) CHLDIDEL
What do you think is the ideal number of children for a family to have?
RANGE: 0 to 8
N Mean Std. Deviation
Total 2693 4.13 2.674
0) 49 1.8
1) 49 1.8
2) 1076 40.0
3) 501 18.6
4) 160 5.9
5) 22 0.8
6) 9 0.3
7) Seven or more 3 0.1
8) As many as you want 824 30.6
Missing 1339

156) PILLOK
Do you strongly agree, agree, disagree, or strongly disagree that methods of birth control should be available to teenagers between the ages of 14 and 16 if their parents do not approve?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2691 2.099 1.028
1) Strongly agree 947 35.2
2) Agree 885 32.9

3) Disagree 505 18.8
4) Strongly disagree 354 13.2
Missing 1341

157) SEXEDUC
Would you be for or against sex education in the public schools?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2688 1.089 0.284
1) Favor 2450 91.1
2) Oppose 238 8.9
Missing 1344

158) PREMARSX
There's been a lot of discussion about the way morals and attitudes about sex are changing in this country. If a man and a woman have sexual relations before marriage, do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2680 3.306 1.098
1) Always wrong 390 14.6
2) Almost always wrong 162 6.0
3) Wrong only sometimes 367 13.7
4) Not wrong at all 1761 65.7
Missing 1352

159) TEENSEX
What if they are in their early teens, say 14 to 16 years old? In that case, do you think sex relations before marriage are always wrong, almost always wrong, wrong only sometimes, or not wrong at all?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2679 2.048 1.12
1) Always wrong 1196 44.6
2) Almost always wrong 581 21.7
3) Wrong only sometimes 479 17.9

4) Not wrong at all 423 15.8
Missing 1353

160) XMARSEX
What is your opinion about a married person having sexual relations with someone other than the marriage partner—is it always wrong, almost always wrong, wrong only sometimes, or not wrong at all?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2650 1.553 0.831
1) Always wrong 1667 62.9
2) Almost always wrong 606 22.9
3) Wrong only sometimes 272 10.3
4) Not wrong at all 105 4.0
Missing 1382

161) HOMOSEX
What about sexual relations between two adults of the same sex—do you think it is always wrong, almost always wrong, wrong only sometimes, or not wrong at all?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2611 3.047 1.313
1) Always wrong 693 26.5
2) Almost always wrong 112 4.3
3) Wrong only sometimes 185 7.1
4) Not wrong at all 1621 62.1
Missing 1421

162) PORNLAW
Which of these statements comes closest to your feelings about pornography laws?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2652 1.793 0.518
1) There should be laws against the distribution of pornography whatever the age 686 25.9

2) There should be laws against the distribution of pornography to persons under 18 1828 68.9
3) There should be no laws forbidding the distribution of pornography 138 5.2
Missing 1380

163) XMOVIE
Have you seen an X-rated movie in the last year?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2656 1.696 0.46
1) Yes 808 30.4
2) No 1848 69.6
Missing 1376

164) SPANKING
Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hard spanking?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2684 2.49 0.952
1) Strongly agree 424 15.8
2) Agree 977 36.4
3) Disagree 827 30.8
4) Strongly disagree 456 17.0
Missing 1348

165) LETDIE1
When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1316 1.29 0.454
1) Yes 934 71.0
2) No 382 29.0
Missing 2716

166) SUICIDE1
(Do you think a person has the right to end his or her own life if this person...) Has an incurable disease?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1212 1.268 0.443
1) Yes 887 73.2
2) No 325 26.8
Missing 2820

167) SUICIDE2
(Do you think a person has the right to end his or her own life if this person...) Has gone bankrupt?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1298 1.881 0.324
1) Yes 155 11.9
2) No 1143 88.1
Missing 2734

168) SUICIDE3
(Do you think a person has the right to end his or her own life if this person...) Has dishonored his or her own family?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1317 1.89 0.313
1) Yes 145 11.0
2) No 1172 89.0
Missing 2715

169) SUICIDE4
(Do you think a person has the right to end his or her own life if this person...) Is tired or living and ready to die?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1231 1.795 0.404

1) Yes 252 20.5
2) No 979 79.5
Missing 2801

170) POLHITOK
Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1300 1.315 0.465
1) Yes 890 68.5
2) No 410 31.5
Missing 2732

171) POLABUSE
(Would you approve of a policeman striking a citizen who...) Had said vulgar and obscene things to the policeman?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1304 1.923 0.266
1) Yes 100 7.7
2) No 1204 92.3
Missing 2728

172) POLMURDR
(Would you approve of a policeman striking a citizen who...) Was being questioned as a suspect in a murder case?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2634 1.787 0.409
1) Yes 560 21.3
2) No 2074 78.7
Missing 1398

173) POLESCAP
(Would you approve of a policeman striking a citizen who...) Was attempting to escape from custody?

RANGE: 1 to 2
N Mean Std. Deviation
Total 2639 1.402 0.49
1) Yes 1579 59.8
2) No 1060 40.2
Missing 1393

174) POLATTAK
(Would you approve of a policeman striking a citizen who...) Was attacking the policeman with his fists?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1303 1.214 0.41
1) Yes 1024 78.6
2) No 279 21.4
Missing 2729

175) FEAR
Is there any area right around here—that is, within a mile—where you would be afraid to walk alone at night?
RANGE: 1 to 2
N Mean Std. Deviation
Total 4022 1.65 0.477
1) Yes 1409 35.0
2) No 2613 65.0
Missing 10

176) OWNGUN
Do you happen to have in your home (IF HOUSE: or garage) any guns or revolvers?
RANGE: 1 to 3
N Mean Std. Deviation
Total 3922 1.65 0.482
1) Yes 1383 35.3
2) No 2529 64.5
3) Refused 10 0.3
Missing 110

177) HUNT1
Do you (or does your [husband/wife/partner]) go hunting?
RANGE: 1 to 4
N Mean Std. Deviation
Total 4025 3.658 0.887
1) Yes, respondent does 325 8.1
2) Yes, spouse or partner does 155 3.9
3) Yes, both do 93 2.3
4) No, neither respondent nor spouse or partner does 3452 85.8
Missing 7

178) NEWS
How often do you read the newspaper—every day, a few times a week, once a week, less than once a week, or never?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2696 3.3 1.628
1) Every day 642 23.8
2) A few times a week 361 13.4
3) Once a week 240 8.9
4) Less than once a week 452 16.8
5) Never 1001 37.1
Missing 1336
179) TVHOURS
On the average day, about how many hours do you personally watch television?
RANGE: 0 to 24
N Mean Std. Deviation
Total 2683 3.458 3.109
Missing 1349

180) FECHLD
Please read the following statements and indicate whether you strongly agree, agree, disagree, or strongly disagree with each statement. For example, here is the statement: A working mother can establish just as warm and secure a relationship with her children as a mother who does not work.

RANGE: 1 to 4
N Mean Std. Deviation
Total 2714 1.817 0.793
1) Strongly agree 1062 39.1
2) Agree 1173 43.2
3) Disagree 394 14.5
4) Strongly disagree 85 3.1
Missing 1318

181) FEPRESCH
A preschool child is likely to suffer if his or her mother works.
RANGE: 1 to 4
N Mean Std. Deviation
Total 2707 2.97 0.766
1) Strongly agree 105 3.9
2) Agree 520 19.2
3) Disagree 1433 52.9
4) Strongly disagree 649 24.0
Missing 1325

182) FEFAM
It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.
RANGE: 1 to 4
N Mean Std. Deviation
Total 2708 3.132 0.841
1) Strongly agree 131 4.8
2) Agree 410 15.1
3) Disagree 1137 42.0
4) Strongly disagree 1030 38.0
Missing 1324

183) RACDIF1
On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are...
Mainly due to discrimination?
RANGE: 1 to 2

N Mean Std. Deviation
Total 2681 1.432 0.495
1) Yes 1524 56.8
2) No 1157 43.2
Missing 1351

184) RACDIF2
(On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are...) Because most (Negroes/Blacks/African-Americans) have less in-born ability to learn?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2689 1.948 0.223
1) Yes 141 5.2
2) No 2548 94.8
Missing 1343

185) RACDIF3
(On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are...) Because most (Negroes/Blacks/African-Americans) don't have the chance for education that it takes to rise out of poverty?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2673 1.435 0.496
1) Yes 1511 56.5
2) No 1162 43.5
Missing 1359

186) RACDIF4
(On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are...) Because most (Negroes/Blacks/African-Americans) just don't have the motivation or willpower to pull themselves up out of poverty?
RANGE: 1 to 2
N Mean Std. Deviation

Total 2666 1.735 0.441
1) Yes 706 26.5
2) No 1960 73.5
Missing 1366

187) HELPPOOR
Next, here are issues that some people tell us are important. Some people think that the government in Washington should do everything possible to improve the standard of living of all poor Americans, they are at Point One on the scale below. Other people think it is not the government's responsibility, and that each person should take care of himself, they are at Point Five. Where would you place yourself on this scale, or haven't you made up your mind on this?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2633 2.682 1.254
1) Government should help 675 25.6
2) 335 12.7
3) Agree with both 1035 39.3
4) 327 12.4
5) People should help themselves 261 9.9
Missing 1399

188) HELPNOT
Some people think that the government in Washington is trying to do too many things that should be left to individuals and private businesses. Others disagree and think that the government should do even more to solve our country's problems. Still others have opinions somewhere in between. Where would you place yourself on this scale, or haven't you made up your mind on this?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2609 2.859 1.293
1) Government should do more 539 20.7
2) 386 14.8
3) Agree with both 977 37.4
4) 319 12.2

5) Government does too much 388 14.9
Missing 1423

189) HELPSICK
In general, some people think that it is the responsibility of the government in Washington to see to it that people have help in paying for doctors and hospital bills. Others think that these matters are not the responsibility of the federal government and that people should take care of these things themselves. Where would you place yourself on this scale, or haven't you made up your mind on this?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2627 2.376 1.267
1) Government should help 917 34.9
2) 471 17.9
3) Agree with both 793 30.2
4) 227 8.6
5) People should care for themselves 219 8.3
Missing 1405
190) HELPBLK
Some people think that Blacks have been discriminated against for so long that the government has a special obligation to help improve their living standards. Others believe that the government should not be giving special treatment to Blacks. Where would you place yourself on this scale, or haven't you made up your mind on this?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2602 3.027 1.457
1) Government should help 560 21.5
2) 407 15.6
3) Agree with both 679 26.1
4) 315 12.1
5) No special treatment 641 24.6
Missing 1430

191) GOD
Please look at this card and tell me which statement comes closest to

expressing what you believe about God.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2615 4.657 1.682
1) Don't believe 173 6.6
2) Don't know, no way to find out 237 9.1
3) Higher power 341 13.0
4) Believe sometimes 124 4.7
5) Believe with doubts 427 16.3
6) No doubts 1313 50.2
Missing 1417

192) REBORN
Would you say you have been born again or have had a born again experience—that is, a turning point in your life when you committed yourself to Christ?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2600 1.661 0.474
1) Yes 882 33.9
2) No 1718 66.1
Missing 1432

193) SAVESOUL
Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3949 1.63 0.483
1) Yes 1461 37.0
2) No 2488 63.0
Missing 83

194) LIVEBLKS
Now, I'm going to ask you about different types of contact with various groups of people. In each situation would you please tell me whether you would be very much in favor of it happening, somewhat in favor, neither in

favor nor opposed to it happening, somewhat opposed or very much opposed to it happening? Living in a neighborhood where half of your neighbors were Black?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2695 2.722 0.943
1) Strongly favor 454 16.8
2) Favor 234 8.7
3) Neither favor nor oppose 1694 62.9
4) Oppose 233 8.6
5) Strongly oppose 80 3.0
Missing 1337

195) MARBLK
How about having a close relative marry a Black person? Would you be very in favor of it happening, somewhat in favor, neither in favor nor opposed to it happening, somewhat opposed, or very opposed to it happening?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2698 2.624 0.944
1) Strongly favor 546 20.2
2) Favor 196 7.3
3) Neither favor nor oppose 1747 64.8
4) Oppose 144 5.3
5) Strongly oppose 65 2.4
Missing 1334

196) MARASIAN
How about having a close relative marry an Asian American person? Would you be very in favor of it happening, somewhat in favor, neither in favor nor opposed to it happening, somewhat opposed, or very opposed to it happening?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2689 2.588 0.88
1) Strongly favor 513 19.1
2) Favor 242 9.0

3) Neither favor nor oppose 1816 67.5
4) Oppose 76 2.8
5) Strongly oppose 42 1.6
Missing 1343

197) MARHISP
How about having a close relative marry a Hispanic person? Would you be very in favor of it happening, somewhat in favor, neither in favor nor opposed to it happening, somewhat opposed, or very opposed to it happening?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2690 2.58 0.882
1) Strongly favor 517 19.2
2) Favor 257 9.6
3) Neither favor nor oppose 1794 66.7
4) Oppose 82 3.0
5) Strongly oppose 40 1.5
Missing 1342

198) MARWHT
What about having a close relative marry a White person? Would you be very in favor of it happening, somewhat in favor, neither in favor nor opposed to it happening, somewhat opposed, or very opposed to it happening?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2694 2.461 0.922
1) Strongly favor 666 24.7
2) Favor 235 8.7
3) Neither favor nor oppose 1709 63.4
4) Oppose 54 2.0
5) Strongly oppose 30 1.1
Missing 1338

199) RACWORK
IF EMPLOYED: Are the people who work where you work all White, mostly White, about half and half, mostly Black, or all Black?
RANGE: 1 to 5

N Mean Std. Deviation
Total 1371 2.253 0.722
1) All White 163 11.9
2) Mostly White 757 55.2
3) About half and half 397 29.0
4) Mostly Black 49 3.6
5) All Black 5 0.4
Missing 2661

200) DISCAFF
What do you think the chances are these days that a white person won't get a job or promotion while an equally or less qualified Black person gets one instead? Is this very likely, somewhat likely, or not very likely to happen these days?
RANGE: 1 to 3
N Mean Std. Deviation
Total 3948 2.367 0.688
1) Very likely 476 12.1
2) Somewhat likely 1549 39.2
3) Not very likely 1923 48.7
Missing 84

201) FEJOBAFF
Some people say that because of past discrimination, women should be given preference in hiring and promotion. Others say that such preference in hiring and promotion of women is wrong because it discriminates against men. What about your opinion—are you for or against preferential hiring and promotion of women? IF FOR: Do you favor preference in hiring and promotion strongly or not strongly? IF AGAINST: Do you oppose the preference in hiring and promotion strongly or not strongly?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1323 2.796 1.13
1) Strongly favor 257 19.4
2) Not strongly favor 235 17.8
3) Not strongly oppose 352 26.6
4) Strongly oppose 479 36.2

Missing 2709

202) DISCAFFM
What do you think the chances are these days that a man won't get a job or promotion while an equally or less qualified woman gets one instead. Is this very likely, somewhat likely, somewhat unlikely, or very unlikely these days?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1336 2.719 0.898
1) Very likely 119 8.9
2) Somewhat likely 422 31.6
3) Not very likely 511 38.2
4) Very unlikely 284 21.3
Missing 2696

203) FEHIRE
Now I'm going to read several statements. As I read each one, please tell me whether you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree. Because of past discrimination, employers should make special efforts to hire and promote qualified women.
RANGE: 1 to 5
N Mean Std. Deviation
Total 2708 2.449 1.134
1) Strongly agree 622 23.0
2) Agree 898 33.2
3) Neither agree nor disagree 673 24.9
4) Disagree 381 14.1
5) Strongly disagree 134 4.9
Missing 1324

204) RELPERSN
To what extent do you consider yourself a religious person? Are you...
RANGE: 1 to 4
N Mean Std. Deviation
Total 3954 2.72 1.038
1) Very religious 518 13.1

2) Moderately religious 1284 32.5
3) Slightly religious 939 23.7
4) Not religious at all 1213 30.7
Missing 78

205) SPRTPRSN
To what extent do you consider yourself a spiritual person? Are you...
RANGE: 1 to 4
N Mean Std. Deviation
Total 3934 2.319 1.007
1) Very spiritual 961 24.4
2) Moderately spiritual 1362 34.6
3) Slightly spiritual 1006 25.6
4) Not spiritual at all 605 15.4
Missing 98

206) OTHLANG
Can you speak a language other than English [Spanish]?
RANGE: 1 to 2
N Mean Std. Deviation
Total 3937 1.709 0.454
1) Yes 1144 29.1
2) No 2793 70.9
Missing 95

207) OTHLANG1
What other language(s) do you speak? First response. (2016)
RANGE: 1 to 176
N Mean Std. Deviation
Total 1064 11.837 22.84
Missing 2968
See Value Labels in SPSS for the 176 possible answer choices. Most common answers:
2) Spanish 46.6
1) English 9.8
4) French 8.9
12) German 6.8

8) Chinese 3.0
10) Italian 2.3
32) Japanese 1.9
19) Korean 1.8
6) Russian 1.5
33) Portuguese 1.5
46) Sign language 1.3
22) Arabic 1.1

208) SPKLANG
How well do you speak that language? [IF SPEAKS 2 OR MORE, ASK ONLY OF THE MOST FLUENT LANGUAGE] READ CATEGORIES
RANGE: 1 to 4
N Mean Std. Deviation
Total 1066 1.957 0.931
1) Very well 441 41.4
2) Well 275 25.8
3) Not well 305 28.6
4) Poorly/hardly at all 45 4.2
Missing 2966

209) COMPUSE
Do you personally ever use a computer at home, at work, or at some other location?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2651 1.092 0.29
1) Yes 2406 90.8
2) No 245 9.2
Missing 1381

210) EMAILHR
About how many minutes or hours per week do you spend sending and answering electronic mail or email?
RANGE: 0 to 124
N Mean Std. Deviation
Total 2322 6.918 10.038

211) WWWHR
Not counting email, about how many minutes or hours per week do you use the web? (Include time you spend visiting regular websites and time spent using interactive internet services like social media, streaming services, chat rooms, online conferencing services, discussion boards or forums, and the like.)
RANGE: 0 to 168
N Mean Std. Deviation
Total 2466 14.803 17.392

212) DISRSPCT
(In your day-to-day life how often have any of the following things happened to you?) You are treated with less courtesy or respect than other people.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2601 4.178 1.409
1) Almost every day 136 5.2
2) At least once a week 231 8.9
3) A few times a month 327 12.6
4) A few times a year 801 30.8
5) Less than once a year 552 21.2
6) Never 554 21.3
Missing 1431

213) POORSERV
(In your day-to-day life how often have any of the following things happened to you?) You receive poorer service than other people at restaurants or stores.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2592 4.831 1.116
1) Almost every day 24 0.9
2) At least once a week 82 3.2
3) A few times a month 155 6.0
4) A few times a year 672 25.9

5) Less than once a year 774 29.9
6) Never 885 34.1
Missing 1440

214) NOTSMART
(In your day-to-day life how often have any of the following things happened to you?) People act as if they think you are not smart.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2602 4.611 1.35
1) Almost every day 101 3.9
2) At least once a week 129 5.0
3) A few times a month 202 7.8
4) A few times a year 684 26.3
5) Less than once a year 619 23.8
6) Never 867 33.3
Missing 1430
215) AFRAIDOF
(In your day-to-day life how often have any of the following things happened to you?) People act as if they are afraid of you.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2604 5.267 1.14
1) Almost every day 39 1.5
2) At least once a week 56 2.2
3) A few times a month 120 4.6
4) A few times a year 352 13.5
5) Less than once a year 426 16.4
6) Never 1611 61.9
Missing 1428

216) THREATEN
(In your day-to-day life how often have any of the following things happened to you?) You are threatened or harassed.
RANGE: 1 to 6
N Mean Std. Deviation
Total 2604 5.267 0.997

1) Almost every day 20 0.8
2) At least once a week 44 1.7
3) A few times a month 84 3.2
4) A few times a year 322 12.4
5) Less than once a year 738 28.3
6) Never 1396 53.6
Missing 1428

217) QUALLIFE
In general, would you say your quality of life is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 3632 2.483 0.915
1) Excellent 477 13.1
2) Very good 1449 39.9
3) Good 1243 34.2
4) Fair 399 11.0
5) Poor 64 1.8
Missing 400

218) HLTHPHYS
In general, how would you rate your physical health?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3630 2.674 1.038
1) Excellent 526 14.5
2) Very good 1012 27.9
3) Good 1361 37.5
4) Fair 582 16.0
5) Poor 149 4.1
Missing 402

219) HLTHMNTL
In general, how would you rate your mental health, including your mood and your ability to think?
RANGE: 1 to 5
N Mean Std. Deviation

Total 3637 2.519 1.016
1) Excellent 579 15.9
2) Very good 1313 36.1
3) Good 1147 31.5
4) Fair 476 13.1
5) Poor 122 3.4
Missing 395

220) SATSOC
In general how would you rate your satisfaction with your social activities and relationships?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3622 2.782 1.034
1) Excellent 362 10.0
2) Very good 1125 31.1
3) Good 1283 35.4
4) Fair 645 17.8
5) Poor 207 5.7
Missing 410

221) ACTSSOC
In general, please rate how well you carry out your usual social activities and roles. (This includes activities at home, at work and in your community, and responsibilities as a parent, child, spouse, employee, friend, etc.)
RANGE: 1 to 5
N Mean Std. Deviation
Total 3624 2.507 0.912
1) Excellent 431 11.9
2) Very good 1457 40.2
3) Good 1290 35.6
4) Fair 358 9.9
5) Poor 88 2.4
Missing 408

222) PHYSACTS

To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair?
RANGE: 1 to 5

	N	Mean	Std. Deviation
Total	3623	1.689	0.995
1) Completely	2190	60.4	
2) Mostly	667	18.4	
3) Moderately	507	14.0	
4) A little	219	6.0	
5) Not at all	40	1.1	
Missing	409		

223) EMOPROBS
In the past seven days, how often have you been bothered by emotional problems such as feeling anxious, depressed or irritable?
RANGE: 1 to 5

	N	Mean	Std. Deviation
Total	3614	2.491	1.092
1) Never	765	21.2	
2) Rarely	1092	30.2	
3) Sometimes	1126	31.2	
4) Often	478	13.2	
5) Always	153	4.2	
Missing	418		

224) FATIGUE
In the past seven days, how would you rate your fatigue on average?
RANGE: 1 to 5

	N	Mean	Std. Deviation
Total	3615	2.314	0.916
1) None	684	18.9	
2) Mild	1491	41.2	
3) Moderate	1124	31.1	
4) Severe	253	7.0	
5) Very severe	63	1.7	
Missing	417		

225) WKRSLFFAM
(Do/did) you work in your own family business or farm?
RANGE: 1 to 2
N Mean Std. Deviation
Total 429 1.55 0.498
0) Inapplicable
1) Yes 193 45.0
2) No 236 55.0
Missing 3603

226) NEXTGEN
I'm going to read to you some statements like those you might find in a newspaper or magazine article. For each statement, please tell me if you strongly agree, agree, disagree or strongly disagree. Because of science and technology, there will be more opportunities for the next generation.
RANGE: 1 to 4
N Mean Std. Deviation
Total 1859 1.712 0.668
1) Strongly agree 734 39.5
2) Agree 951 51.2
3) Disagree 149 8.0
4) Strongly disagree 25 1.3
Missing 2173

227) TOOFAST
I'm going to read to you some statements like those you might find in a newspaper or magazine article. For each statement, please tell me if you strongly agree, agree, disagree or strongly disagree. Science makes our way of life change too fast.
RANGE: 1 to 4
N Mean Std. Deviation
Total 1862 2.605 0.798
1) Strongly agree 163 8.8
2) Agree 617 33.1
3) Disagree 875 47.0
4) Strongly disagree 207 11.1
Missing 2170

228) ADVFRONT
I'm going to read to you some statements like those you might find in a newspaper or magazine article. For each statement, please tell me if you strongly agree, agree, disagree or strongly disagree. Even if it brings no immediate benefits, scientific research that advances the frontiers of knowledge is necessary and should be supported by the federal government.
RANGE: 1 to 4
N Mean Std. Deviation
Total 1855 1.822 0.704
1) Strongly agree 615 33.2
2) Agree 994 53.6
3) Disagree 207 11.2
4) Strongly disagree 39 2.1 Missing 2177

229) SCIBNFTS
Now for another type of question. People have frequently noted that scientific research has produced benefits and harmful results. Would you say that, on balance, the benefits of scientific research have outweighed the harmful results, or have the harmful results of scientific research been greater than its benefits?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1838 1.447 0.561
1) Benefits greater 1078 58.7
2) About equal (if volunteered) 698 38.0
3) Harmful results greater 62 3.4
Missing 2194

230) VIRUSES
Now, I would like to ask you a few short questions like those you might see on a television game show. For each statement that I read, please tell me if it is true or false. If you don't know or aren't sure, just tell me so, and we will skip to the next question. Remember true, false or don't know. Antibiotics kill viruses as well as bacteria.
RANGE: 1 to 2
N Mean Std. Deviation

Total 1834 1.683 0.466
1) True 582 31.7
2) False 1252 68.3
Missing 2198

231) INTEDUC
Are you very interested, moderately interested or not at all interested in local school issues?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1878 1.929 0.707
1) Very interested 541 28.8
2) Moderately interested 930 49.5
3) Not at all interested 407 21.7
Missing 2154

232) INTSCI
Are you very interested, moderately interested or not at all interested in issues about new scientific discoveries?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1877 1.638 0.642
1) Very interested 849 45.2
2) Moderately interested 858 45.7
3) Not at all interested 170 9.1
Missing 2155

233) INTECON
Are you very interested, moderately interested or not at all interested in economic issues and business conditions?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1867 1.653 0.634
1) Very interested 812 43.5
2) Moderately interested 891 47.7
3) Not at all interested 164 8.8
Missing 2165

234) INTTECH
Are you very interested, moderately interested or not at all interested in issues about the use of new inventions and technologies?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1874 1.677 0.639
1) Very interested 782 41.7
2) Moderately interested 915 48.8
3) Not at all interested 177 9.4
Missing 2158

235) POSSLQ
Which of these statements applies to you?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1970 2.324 1.385
1) I am married and living in the same household as my husband or wife 945 48.0
2) I am living as married and my partner and I together live in the same household 169 8.6
3) I have a husband or wife or steady partner, but we don't live in the same household 129 6.5
4) I don't have a steady partner 727 36.9
Missing 2062

236) POSSLQY
Which of these statements applies to you?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2040 2.26 1.354
1) I am married and living in the same household as my husband or wife 990 48.5
2) I am living as married and my partner and I together live in the same household 213 10.4
3) I have a husband or wife or steady partner, but we don't live in the same household 153 7.5

4) I don't have a steady partner 684 33.5
Missing 1992

237) MARCOHAB
Marriage and cohabitation status
RANGE: 1 to 3
N Mean Std. Deviation
Total 4031 1.913 0.947
1) Married 1997 49.5
2) Not married, cohabitating partner 386 9.6
3) Not married, no cohabitating partner 1648 40.9
Missing 1

238) ENDSMEET
Thinking of your household's total income, including all the sources of income of all the members who contribute to it, how difficult or easy is it currently for your household to make ends meet?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1763 3.264 1.147
1) Very difficult 129 7.3
2) Fairly difficult 321 18.2
3) Neither easy nor difficult 548 31.1
4) Fairly easy 485 27.5
5) Very easy 280 15.9
Missing 2269

239) OPWLTH
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is coming from a wealthy family?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1815 3.149 1.091
1) Essential 118 6.5
2) Very important 407 22.4
3) Fairly important 580 32.0
4) Not very important 506 27.9

5) Not important at all 204 11.2
Missing 2217

240) OPPARED
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is having well-educated parents?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1845 2.685 0.937
1) Essential 159 8.6
2) Very important 642 34.8
3) Fairly important 738 40.0
4) Not very important 233 12.6
5) Not important at all 73 4.0
Missing 2187

241) OPEDUC
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is having a good education yourself?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1853 1.883 0.781
1) Essential 623 33.6
2) Very important 875 47.2
3) Fairly important 313 16.9
4) Not very important 32 1.7
5) Not important at all 10 0.5 Missing 2179
242) OPHRDWRK
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is hard work?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1857 1.666 0.718
1) Essential 856 46.1
2) Very important 797 42.9
3) Fairly important 176 9.5
4) Not very important 24 1.3

5) Not important at all 4 0.2
Missing 2175

243) OPKNOW
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is knowing the right people?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1845 2.543 0.913
1) Essential 250 13.6
2) Very important 603 32.7
3) Fairly important 763 41.4
4) Not very important 199 10.8
5) Not important at all 30 1.6
Missing 2187

244) OPCLOUT
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is having political connections?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1793 3.524 1.021
1) Essential 76 4.2
2) Very important 203 11.3
3) Fairly important 499 27.8
4) Not very important 735 41.0
5) Not important at all 280 15.6
Missing 2239

245) OPRACE
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is a person's race?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1800 3.908 1.128
1) Essential 44 2.4
2) Very important 184 10.2

3) Fairly important 420 23.3
4) Not very important 398 22.1
5) Not important at all 754 41.9
Missing 2232

246) OPRELIG
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is a person's religion?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1806 4.185 0.996
1) Essential 42 2.3
2) Very important 97 5.4
3) Fairly important 217 12.0
4) Not very important 579 32.1
5) Not important at all 871 48.2
Missing 2226

247) OPSEX
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is being born a man or woman?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1793 3.83 1.173
1) Essential 69 3.8
2) Very important 198 11.0
3) Fairly important 403 22.5
4) Not very important 422 23.5
5) Not important at all 701 39.1
Missing 2239

248) OPBRIBES
(For each of these, please tell me how important you think it is for getting ahead in life...) How important is giving bribes?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1718 4.588 0.82

1) Essential 20 1.2
2) Very important 51 3.0
3) Fairly important 96 5.6
4) Not very important 282 16.4
5) Not important at all 1269 73.9
Missing 2314

249) GOODLIFE
The way things are in America, people like me and my family have a good chance of improving our standard of living—do you agree or disagree?
RANGE: 1 to 5
N Mean Std. Deviation
Total 2664 2.741 1.073
1) Strongly agree 290 10.9
2) Agree 944 35.4
3) Neither agree nor disagree 750 28.2
4) Disagree 526 19.7
5) Strongly disagree 154 5.8
Missing 1368

250) PAYDOC
About how much do you think a doctor in general practice earns?
RANGE: 0 to 999996
N Mean Std. Deviation
Total 1828 228682.544 188107.26

251) PAYCLERK
How much do you think a salesclerk earns?
RANGE: 0 to 999996
N Mean Std. Deviation
Total 1823 38771.885 76630.982

252) PAYEXEC
How much do you think a chairman of a large national corporation earns?
RANGE: 0 to 999996
N Mean Std. Deviation
Total 1818 673620.721 371536.782

253) PAYUNSKL
How much do you think an unskilled worker in a factory earns?
RANGE: 0 to 999996
N Mean Std. Deviation
Total 1824 36354.363 66898.405

254) PAYCABNT
How much do you think a cabinet minister in the federal government earns?
RANGE: 0 to 999996
N Mean Std. Deviation
Total 1812 257485.826 232526.024

255) INCGAP
To what extent do you agree or disagree with the following statements? Differences in income in America are too large.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1822 1.96 1.071
1) Strongly agree 793 43.5
2) Agree 545 29.9
3) Neither agree nor disagree 301 16.5
4) Disagree 130 7.1
5) Strongly disagree 53 2.9
Missing 2210

256) GOVEQINC
It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1806 3.122 1.257
1) Strongly agree 190 10.5
2) Agree 434 24.0
3) Neither agree nor disagree 471 26.1
4) Disagree 387 21.4
5) Strongly disagree 324 17.9

Missing 2226

257) GOVUNEMP
The government should provide a decent standard of living for the unemployed.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1784 2.786 1.213
1) Strongly agree 288 16.1
2) Agree 508 28.5
3) Neither agree nor disagree 455 25.5
4) Disagree 363 20.3
5) Strongly disagree 170 9.5
Missing 2248

258) TAXRICH
Generally, how would you describe taxes in America today for those with high incomes? Taxes are...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1704 3.601 1.168
1) Much too high 98 5.8
2) Too high 240 14.1
3) About right 329 19.3
4) Too low 614 36.0
5) Much too low 423 24.8
Missing 2328

259) TAXSHARE
Do you think people with high incomes should pay a larger share of their income in taxes than those with low incomes, the same share, or a smaller share?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1748 2.011 0.822
1) Much larger share 512 29.3
2) Large 760 43.5

3) The same share 432 24.7
4) Smaller 32 1.8
5) Much smaller share 12 0.7
Missing 2284

260) CONWLTH
(In all countries, there are differences or even conflicts between different social groups. In your opinion, in America how much conflict is there between...) Poor people and rich people?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1675 2.216 0.726
1) Very strong conflicts 255 15.2
2) Strong conflicts 849 50.7
3) Not very strong conflicts 526 31.4
4) There are no conflicts 45 2.7
Missing 2357

261) CONCLASS
(In all countries, there are differences or even conflicts between different social groups. In your opinion, in America how much conflict is there between...) The working class and the middle class?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1701 2.768 0.668
1) Very strong conflicts 69 4.1
2) Strong conflicts 416 24.5
3) Not very strong conflicts 1057 62.1
4) There are no conflicts 159 9.3
Missing 2331

262) CONUNION
(In all countries, there are differences or even conflicts between different social groups. In your opinion, in America how much conflict is there between...) Management and workers?
RANGE: 1 to 4
N Mean Std. Deviation

Total 1687 2.434 0.683
1) Very strong conflicts 140 8.3
2) Strong conflicts 721 42.7
3) Not very strong conflicts 780 46.2
4) There are no conflicts 46 2.7
Missing 2345

263) CONAGE
(In all countries, there are differences or even conflicts between different social groups. In your opinion, in America how much conflict is there between...) Young people and older people?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1716 2.594 0.733
1) Very strong conflicts 138 8.0
2) Strong conflicts 537 31.3
3) Not very strong conflicts 925 53.9
4) There are no conflicts 116 6.8
Missing 2316
264) CONIMM
(In all countries, there are differences or even conflicts between different social groups. In your opinion, in America how much conflict is there between...) People born in America and people from other countries who have come to live in America?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1700 2.236 0.721
1) Very strong conflicts 244 14.4
2) Strong conflicts 854 50.2
3) Not very strong conflicts 558 32.8
4) There are no conflicts 44 2.6
Missing 2332

265) LDCGAP
(Turning to international differences, to what extent do you agree or disagree with the following statements?) Present economic differences between rich and poor countries are too large.

RANGE: 1 to 5
N Mean Std. Deviation
Total 1700 2.311 0.953
1) Strongly agree 365 21.5
2) Agree 619 36.4
3) Neither agree nor disagree 587 34.5
4) Disagree 81 4.8
5) Strongly disagree 48 2.8
Missing 2332

266) LDCTAX
(Turning to international differences, to what extent do you agree or disagree with the following statements?) People in wealthy countries should make an additional tax contribution to help people in poor countries.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1722 3.177 1.162
1) Strongly agree 150 8.7
2) Agree 335 19.5
3) Neither agree nor disagree 548 31.8
4) Disagree 438 25.4
5) Strongly disagree 251 14.6
Missing 2310

267) RICHHLTH
Is it just or unjust—right or wrong—that people with higher incomes can buy better health care than people with lower incomes?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1771 3.655 1.231
1) Very just, definitely right 121 6.8
2) Somewhat just, right 191 10.8
3) Neither just nor unjust, mixed feelings 452 25.5
4) Somewhat unjust, wrong 421 23.8
5) Very unjust, definitely wrong 586 33.1 Missing 2261

268) RICHEDUC

Is it just or unjust—right or wrong—that people with higher incomes can buy better education for their children than people with lower incomes?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1769 3.617 1.254
1) Very just, definitely right 119 6.7
2) Somewhat just, right 241 13.6
3) Neither just nor unjust, mixed feelings 422 23.9
4) Somewhat unjust, wrong 404 22.8
5) Very unjust, definitely wrong 583 33.0
Missing 2263

269) PAYRESP
(In deciding how much people ought to earn, how important should each of these things be, in your opinion...) How much responsibility goes with the job—how important do you think that ought to be in deciding pay?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1767 1.989 0.733
1) Essential 436 24.7
2) Very important 950 53.8
3) Fairly important 357 20.2
4) Not very important 12 0.7
5) Not important at all 12 0.7 Missing 2265

270) PAYEDTRN
(In deciding how much people ought to earn, how important should each of these things be, in your opinion...) The number of years spent in education and training?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1771 2.395 0.855
1) Essential 242 13.7
2) Very important 759 42.9
3) Fairly important 622 35.1
4) Not very important 125 7.1
5) Not important at all 23 1.3

Missing 2261

271) PAYCHILD
(In deciding how much people ought to earn, how important should each of these things be, in your opinion...) Whether the person has children to support?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1729 3.577 1.161
1) Essential 91 5.3
2) Very important 242 14.0
3) Fairly important 413 23.9
4) Not very important 544 31.5
5) Not important at all 439 25.4
Missing 2303

272) PAYDOWEL
(In deciding how much people ought to earn, how important should each of these things be, in your opinion...) How well he or she does the job?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1760 1.701 0.726
1) Essential 763 43.4
2) Very important 796 45.2
3) Fairly important 171 9.7
4) Not very important 24 1.4
5) Not important at all 6 0.3
Missing 2272

273) MARHOMO
(Do you agree or disagree?) Homosexual couples should have the right to marry one another.
RANGE: 1 to 5
N Mean Std. Deviation
Total 2658 2.149 1.371
1) Strongly agree 1265 47.6
2) Agree 500 18.8

3) Neither agree nor disagree 419 15.8
4) Disagree 179 6.7
5) Strongly disagree 295 11.1
Missing 1374

274) MEOVRWRK
Family life often suffers because men concentrate too much on their work.
RANGE: 1 to 5
N Mean Std. Deviation
Total 2714 2.806 0.985
1) Strongly agree 202 7.4
2) Agree 886 32.6
3) Neither agree nor disagree 1001 36.9
4) Disagree 487 17.9
5) Strongly disagree 138 5.1
Missing 1318

275) RELACTIV
How often do you take part in the activities and organizations of a church or place of worship other than attending services?
RANGE: 1 to 10
N Mean Std. Deviation
Total 3949 2.865 2.278
1) Never 1650 41.8
2) Less than once a year 614 15.5
3) About once or twice a year 488 12.4
4) Several times a year 449 11.4
5) About once a month 150 3.8
6) Two to three times a month 159 4.0
7) Nearly every week 137 3.5
8) Every week 241 6.1
9) Several times a week 26 0.7
10) Once a day 35 0.9
Missing 83

276) CANTRUST
Generally speaking, would you say that people can be trusted or that you

can't be too careful in dealing with people?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1815 2.724 0.867
1) People can almost always be trusted 71 3.9
2) People can usually be trusted 714 39.3
3) You usually can't be too careful in dealing with people 750 41.3
4) You almost always can't be too careful in dealing with people 205 11.3
Missing 2286

277) RELIGINF
(Please indicate to what extent you agree or disagree with each of the following statements.) The U.S. would be a better country if religion had less influence.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3559 3.095 1.25
1) Strongly agree 514 14.4
2) Agree 565 15.9
3) Neither agree nor disagree 1069 30.0
4) Disagree 890 25.0
5) Strongly disagree 521 14.6
Missing 473

278) PRIVENT
(How much do you agree or disagree with each of these statements) Private enterprise is the best way to solve America's economic problems.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1754 2.845 1.065
1) Strongly agree 225 12.8
2) Agree 366 20.9
3) Neither agree nor disagree 730 41.6
4) Disagree 322 18.4
5) Strongly disagree 111 6.3
Missing 2278

279) POSTMAT1
Looking at the list on the hand card, please tell me the one thing you think should be America's highest priority, the most important thing it should do.
RANGE: 1 to 4
N Mean Std. Deviation
Total 1610 2.196 1.103
1) Maintain order in the nation 509 31.6
2) Give people more say in government decisions 618 38.4
3) Fight rising prices 141 8.8
4) Protect freedom of speech 342 21.2
Missing 2422

280) SCIGRN
(How much do you agree or disagree with each of these statements)? Modern science will solve our environmental problems with little change to our way of life.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1772 3.53 0.978
1) Strongly agree 35 2.0
2) Agree 241 13.6
3) Neither agree nor disagree 525 29.6
4) Disagree 691 39.0
5) Strongly disagree 280 15.8
Missing 2260

281) GRNECON
(How much do you agree or disagree with each of these statements)? We worry too much about the future of the environment and not enough about prices and jobs today.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1795 3.403 1.162
1) Strongly agree 114 6.4
2) Agree 310 17.3
3) Neither agree nor disagree 453 25.2

4) Disagree 574 32.0
5) Strongly disagree 344 19.2
Missing 2237
282) HARMSGRN
(How much do you agree or disagree with each of these statements?) Almost everything we do in modern life harms the environment.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1798 2.804 0.993
1) Strongly agree 135 7.5
2) Agree 639 35.5
3) Neither agree nor disagree 520 28.9
4) Disagree 451 25.1
5) Strongly disagree 53 2.9
Missing 2234

283) GRNPROG
(How much do you agree or disagree with each of these statements)? People worry too much about human progress harming the environment.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1772 3.42 1.071
1) Strongly agree 72 4.1
2) Agree 312 17.6
3) Neither agree nor disagree 459 25.9
4) Disagree 658 37.1
5) Strongly disagree 271 15.3
Missing 2260

284) GRWTHELP
(And please tell me for each of these statements, how much you agree or disagree with it.) Economic growth always harms the environment.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1754 2.712 0.955
1) Strongly agree 143 8.2
2) Agree 637 36.3

3) Neither agree nor disagree 617 35.2
4) Disagree 297 16.9
5) Strongly disagree 60 3.4
Missing 2278

285) GRWTHARM
(And please tell me for each of these statements, how much you agree or disagree with it.) In order to protect the environment America needs economic growth.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1771 3.501 0.835
1) Strongly agree 30 1.7
2) Agree 160 9.0
3) Neither agree nor disagree 619 35.0
4) Disagree 817 46.1
5) Strongly disagree 145 8.2
Missing 2261

286) GRNPRICE
How willing would you be to pay much higher prices in order to protect the environment?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1778 2.918 1.142
1) Very willing 135 7.6
2) Fairly willing 625 35.2
3) Neither willing nor unwilling 476 26.8
4) Fairly unwilling 334 18.8
5) Very unwilling 208 11.7
Missing 2254

287) GRNTAXES
And how willing would you be to pay much higher taxes in order to protect the environment?
RANGE: 1 to 5
N Mean Std. Deviation

Total 1775 3.208 1.25
1) Very willing 129 7.3
2) Fairly willing 479 27.0
3) Neither willing nor unwilling 433 24.4
4) Fairly unwilling 361 20.3
5) Very unwilling 373 21.0
Missing 2257

288) GRNSOL
And how willing would you be to accept cuts in your standard of living in order to protect the environment?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1778 3.15 1.196
1) Very willing 111 6.2
2) Fairly willing 505 28.4
3) Neither willing nor unwilling 487 27.4
4) Fairly unwilling 356 20.0
5) Very unwilling 319 17.9
Missing 2254

289) TOODIFME
How much do you agree or disagree with each of these statements? It is just too difficult for someone like me to do much about the environment.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1783 3.392 0.999
1) Strongly agree 62 3.5
2) Agree 305 17.1
3) Neither agree nor disagree 474 26.6
4) Disagree 756 42.4
5) Strongly disagree 186 10.4
Missing 2249

290) IHLPGRN
(How much do you agree or disagree with each of these statements?) I do what is right for the environment, even when it costs more money or takes

more time.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1781 2.495 0.813
1) Strongly agree 123 6.9
2) Agree 868 48.7
3) Neither agree nor disagree 602 33.8
4) Disagree 161 9.0
5) Strongly disagree 27 1.5
Missing 2251

291) CARSGEN
In general, do you think that air pollution caused by cars is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1778 2.534 0.869
1) Extremely dangerous for the environment 232 13.0
2) Very dangerous 559 31.4
3) Somewhat dangerous 814 45.8
4) Not very dangerous 151 8.5
5) Not dangerous at all for the environment 22 1.2
Missing 2254

292) RECYCLE
How often do you make a special effort to sort glass or cans or plastic or newspapers and so on for recycling?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1700 1.769 1.065
1) Always 1006 59.2
2) Often 278 16.4
3) Sometimes 218 12.8
4) Never 198 11.6
Missing 2332

293) GRNGROUP
Are you a member of any group whose main aim is to preserve or protect the

environment?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1820 1.898 0.303
1) Yes 186 10.2
2) No 1634 89.8
Missing 2212

294) GRNSIGN
In the last five years, have you signed a petition about an environmental issue?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1802 1.752 0.432
1) Yes, I have 447 24.8
2) No, I have not 1355 75.2
Missing 2230

295) GRNMONEY
In the last five years, have you given money to an environmental group?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1814 1.764 0.425
1) Yes, I have 428 23.6
2) No, I have not 1386 76.4
Missing 2218

296) GRNDEMO
In the last five years, have you taken part in a protest or demonstration about an environmental issue?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1817 1.951 0.216
1) Yes, I have 89 4.9
2) No, I have not 1728 95.1
Missing 2215

297) IMPGRN
(How much do you agree or disagree with each of these statements?) There are more important things to do in life than protect the environment.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1788 3.45 1.004
1) Strongly agree 54 3.0
2) Agree 262 14.7
3) Neither agree nor disagree 553 30.9
4) Disagree 664 37.1
5) Strongly disagree 255 14.3
Missing 2244

298) OTHSSAME
(How much do you agree or disagree with each of these statements?) There is no point in doing what I can for the environment unless others do the same.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1791 3.539 1.028
1) Strongly agree 67 3.7
2) Agree 257 14.3
3) Neither agree nor disagree 376 21.0
4) Disagree 826 46.1
5) Strongly disagree 265 14.8
Missing 2241

299) GRNEXAGG
(How much do you agree or disagree with each of these statements?) Many of the claims about environmental threats are exaggerated.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1777 3.547 1.187
1) Strongly agree 106 6.0
2) Agree 282 15.9
3) Neither agree nor disagree 354 19.9
4) Disagree 604 34.0

5) Strongly disagree 431 24.3
Missing 2255

300) TOPPROB1
Which of these issues is the most important for America today?
RANGE: 1 to 9
N Mean Std. Deviation
Total 1690 3.723 2.546
1) Health care 551 32.6
2) Education 216 12.8
3) Crime 53 3.1
4) The environment 224 13.3
5) Immigration 68 4.0
6) The economy 363 21.5
7) Terrorism 50 3.0
8) Poverty 99 5.9
9) None of these 66 3.9
Missing 2342

301) GRNCON
Generally speaking, how concerned are you about environmental issues? Please tell me what you think, where one means you are not at all concerned and five means you are very concerned.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1823 3.889 1.156
1) Not at all concerned 81 4.4
2) 142 7.8
3) 419 23.0
4) 438 24.0
5) Very concerned 743 40.8
Missing 2209

302) ENPRBUS
Here is a list of some different environmental problems. Which problem, if any, do you think is the most important for America as a whole?
RANGE: 1 to 10

N Mean Std. Deviation
Total 1642 5.988 2.612
1) Air pollution 141 8.6
2) Chemicals and pesticides 140 8.5
3) Water shortage 76 4.6
4) Water pollution 131 8.0
5) Nuclear waste 40 2.4
6) Domestic waste disposal 87 5.3
7) Climate change 642 39.1
8) Genetically modified foods 115 7.0
9) Using up out natural resources 177 10.8
10) None of these 93 5.7
Missing 2390

303) GRNEFFME
(How much do you agree or disagree with each of these statements?) Environmental problems have a direct effect on my everyday life.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1768 2.819 0.997
1) Strongly agree 132 7.5
2) Agree 607 34.3
3) Neither agree nor disagree 542 30.7
4) Disagree 423 23.9
5) Strongly disagree 64 3.6 Missing 2264

304) TEMPGEN1
In general, do you think that a rise in the world's temperature caused by climate change is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1734 2.251 1.072
1) Extremely dangerous for the environment 528 30.4
2) Very dangerous 501 28.9
3) Somewhat dangerous 493 28.4
4) Not very dangerous 166 9.6
5) Not dangerous at all for the environment 46 2.7

Missing 2298

305) BUSGRN
Which of these approaches do you think would be the best way of getting business and industry in America to protect the environment?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1643 1.826 0.745
1) Heavy fines for businesses that damage the environment 623 37.9
2) Use the tax system to reward businesses that protect the environment 683 41.6
3) More information and education for businesses about the advantages of protecting the environment 337 20.5
Missing 2389

306) PEOPGRN
Which of these approaches do you think would be the best way of getting people and their families in America to protect the environment?
RANGE: 1 to 3
N Mean Std. Deviation
Total 1689 2.3 0.693
1) Heavy fines for people that damage the environment 228 13.5
2) Use the tax system to reward people that protect the environment 727 43.0
3) More information and education for people about the advantages of protecting the environment 734 43.5
Missing 2343

307) NOBUYGRN
And how often do you avoid buying certain products for environmental reasons?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1821 2.761 0.873
1) Always 146 8.0
2) Often 525 28.8
3) Sometimes 769 42.2

4) Never 381 20.9
Missing 2211

308) IMPORTS
America should limit the import of foreign products in order to protect its national economy.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1800 2.683 0.99
1) Strongly agree 195 10.8
2) Agree 617 34.3
3) Neither agree nor disagree 607 33.7
4) Disagree 326 18.1
5) Strongly disagree 55 3.1
Missing 2232

309) POWRORGS
International organizations are taking away too much power from the American government.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1746 2.99 1.082
1) Strongly agree 174 10.0
2) Agree 377 21.6
3) Neither agree nor disagree 618 35.4
4) Disagree 446 25.5
5) Strongly disagree 131 7.5
Missing 2286

310) LETIN1A
Do you think the number of immigrants to America nowadays should be
RANGE: 1 to 5
N Mean Std. Deviation
Total 2670 3.121 1.133
1) Increased a lot 227 8.5
2) Increased a little 476 17.8
3) Remain the same as it is 1128 42.2

4) Reduced a little 426 16.0
5) Reduced a lot 413 15.5
Missing 1362

311) PARTNERS
How many sex partners have you had in the last 12 months?
RANGE: 0 to 9
N Mean Std. Deviation
Total 2313 0.966 1.038
0) No partners 599 25.9
1) One partner 1504 65.0
2) Two partners 83 3.6
3) Three partners 44 1.9
4) Four partners 27 1.2
5) Five to 10 partners 40 1.7
6) 11-20 partners 4 0.2
7) 21-100 partners 2 0.1
8) More than 100 partners 4 0.2
9) One or more partners (unspecified) 6 0.3
Missing 1719
312) MATESEX
Was one of the partners your husband or wife or regular sexual partner?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1697 1.077 0.266
1) Yes 1567 92.3
2) No 130 7.7
Missing 2335

313) SEXSEX
Have your sex partners in the last 12 months been...
RANGE: 1 to 3
N Mean Std. Deviation
Total 1689 1.944 0.994
1) Exclusively male 884 52.3
2) Both male and female 16 0.9
3) Exclusively female 789 46.7

Missing 2343

314) SEXFREQ
About how often did you have sex during the last 12 months?
RANGE: 0 to 6
N Mean Std. Deviation
Total 2157 2.254 1.914
0) Not at all 633 29.3
1) Once or twice 251 11.6
2) Once a month 283 13.1
3) Two or three times a month 351 16.3
4) About once a week 287 13.3
5) Two or three times a week 268 12.4
6) More than three times a week 84 3.9
Missing 1875

315) NUMWOMEN
Now thinking about the time since your 18th birthday (including the past 12 months), how many female partners have you had sex with?
RANGE: 0 to 996
N Mean Std. Deviation
Total 2207 14.745 92.634
Missing 1825

316) NUMMEN
Now thinking about the time since your 18th birthday (including the past 12 months), how many male partners have you had sex with?
RANGE: 0 to 997
N Mean Std. Deviation
Total 2190 13.307 90.957
Missing 1842

317) PARTNRS5
Now thinking about the past five years - the time since February/March 2015, and including the past 12 months, how many sex partners have you had in that five-year period?
RANGE: 0 to 9

N Mean Std. Deviation
Total 2314 1.723 2.025
0) No partners 336 14.5
1) One partner 1416 61.2
2) Two partners 151 6.5
3) Three partners 83 3.6
4) Four partners 64 2.8
5) Five to 10 partners 109 4.7
6) 11-20 partners 40 1.7
7) 21-100 partners 25 1.1
8) More than 100 partners 7 0.3
9) One or more partners (unspecified) 83 3.6
Missing 1718

318) SEXSEX5
(Now thinking about the past five years—the time since February/March 2015, and including the past 12 months), have your sex partners in the last five years been...
RANGE: 1 to 3
N Mean Std. Deviation
Total 1925 1.925 0.983
1) Exclusively male 1007 52.3
2) Both male and female 55 2.9
3) Exclusively female 863 44.8
Missing 2107

319) EVPAIDSX
Thinking about the time since your 18th birthday, have you ever had sex with a person you paid or who paid you for sex?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2275 1.937 0.243
1) Yes 143 6.3
2) No 2132 93.7
Missing 1757

320) EVSTRAY

Have you ever had sex with someone other than your husband or wife while you were married?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2677 2.252 0.641
1) Yes 297 11.1
2) No 1408 52.6
3) Never married 972 36.3
Missing 1355

321) CONDOM
The last time you had sex, was a condom used? By 'sex' we mean vaginal, oral, or anal sex.
RANGE: 1 to 2
N Mean Std. Deviation
Total 2157 1.835 0.371
1) Used last time 356 16.5
2) Not used 1801 83.5
Missing 1875

322) RELATSEX
The last time you had sex, was it with someone you were in an ongoing relationship with, or was it with someone else? Remember that by 'sex' we mean only vaginal, oral, or anal sex.
RANGE: 1 to 2
N Mean Std. Deviation
Total 2186 1.088 0.284
1) Yes, the last time I had sex, it was with someone I was in an ongoing relationship with 1993 91.2
2) No, the last time I had sex, it was not with someone I was in an ongoing relationship with 193 8.8
Missing 1846

323) EVIDU
Have you ever, even once, taken any drugs by injection with a needle (like heroin, cocaine, amphetamines, or steroids)? DO NOT include anything you took under a doctor's orders.

RANGE: 1 to 2
N Mean Std. Deviation
Total 2259 1.973 0.162
1) Yes 61 2.7
2) No 2198 97.3
Missing 1773

324) EVCRACK
Have you ever, even once, used 'crack' cocaine in chunk or rock form?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2277 1.952 0.214
1) Yes 110 4.8
2) No 2167 95.2
Missing 1755

325) HIVTEST
Have you ever been tested for HIV? Do not count tests you may have had as part of a blood donation. Include oral test (where they take a swab from your mouth).
RANGE: 1 to 2
N Mean Std. Deviation
Total 2216 1.667 0.471
1) Yes 737 33.3
2) No 1479 66.7
Missing 1816
326) SEXORNT
Which of the following best describes you?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2260 2.89 0.406
1) Gay, lesbian, or homosexual 76 3.4
2) Bisexual 96 4.2
3) Heterosexual or straight 2088 92.4
Missing 1772

327) REALINC

Family income in 1972-2006 surveys in constant dollars (base=1986)
RANGE: 218 to 144835.4286
N Mean Std. Deviation
Total 3509 40053.127 40147.485
Missing 523

328) CONINC
Inflation-adjusted family income
RANGE: 336 to 168736.29696
N Mean Std. Deviation
Total 3509 55955.939 47370.022
Missing 523

329) COHORT
Birth cohort of respondent
RANGE: 1932 to 9999
N Mean Std. Deviation
Total 4032 2632.041 2210.722

330) VETYEARS
Have you ever been on active duty for military training or service for two consecutive months or more? IF YES: What was your total time on active duty?
RANGE: 0 to 3
N Mean Std. Deviation
Total 3934 0.24 0.714
0) No active duty 3491 88.7
1) Yes, less than two years 87 2.2
2) Yes, two to four years 212 5.4
3) Yes, more than four years 144 3.7
Missing 98

331) DWELOWN16
When you were 16 years old, did your family own your own home, pay rent, or something else?
RANGE: 1 to 3
N Mean Std. Deviation

Total 2636 1.252 0.463
1) Owned or was buying 2007 76.1
2) Paid rent 595 22.6
3) Other 34 1.3
Missing 1396

332) DWELOWN
(Do you/Does your family) own your (home/apartment), pay rent, or what?
RANGE: 1 to 3
N Mean Std. Deviation
Total 2645 1.324 0.494
1) Own or is buying 1821 68.8
2) Pays rent 791 29.9
3) Other 33 1.2
Missing 1387

333) SEI10
Respondent's socioeconomic index (2010)
RANGE: 10.6 to 93.7
N Mean Std. Deviation
Total 3873 52.409 23.173

334) HISPANIC
Are you Spanish, Hispanic, or Latino/Latina? IF YES: Which group are you from?
RANGE: 1 to 50
N Mean Std. Deviation
Total 3998 1.873 4.628
1) Not Hispanic 3544 88.6
2) Mexican, Mexican America, Chicano/a 245 6.1
3) Puerto Rican 51 1.3
4) Cuban 21 0.5
5) Salvadorian 9 0.2
6) Guatemalan 2 0.1
7) Panamanian 4 0.1
8) Nicaraguan 2 0.1
9) Costa Rican 4 0.1

10) Central American 5 0.1
11) Honduran 3 0.1
15) Dominican 17 0.4
20) Peruvian 4 0.1
21) Ecuadorian 8 0.2
22) Colombian 10 0.3
23) Venezuelan 5 0.1
24) Argentinian 1 0.0
25) Chilean 1 0.0
30) Spanish 46 1.2
35) Filipino/a 1 0.0
41) South American 3 0.1
46) Latino/a 2 0.1
47) Hispanic 7 0.2
50) Other, not specified 3 0.1
Missing 34

335) RACECEN1
(What is your race? Indicate one or more races that you consider yourself to be.) First mention
RANGE: 1 to 15
N Mean Std. Deviation
Total 21 3.524 3.816
1) White 3110 77.1% 78.2%
2) Black or African American 2 463 11.5% 11.6%
3) American Indian or Alaska Native 44 1.1% 1.1%
4) Asian Indian 45 1.1% 1.1%
5) Chinese 39 1.0% 1.0%
6) Filipino 15 0.4% 0.4%
7) Japanese 13 0.3% 0.3%
8) Korean 23 0.6% 0.6%
9) Vietnamese 5 0.1% 0.1%
10) Other Asian 20 0.5% 0.5%
11) Native Hawaiian 2 0.0% 0.1%
12) Guamanian or Chamorro 2 0.0% 0.1%
13) Samoan 2 0.0% 0.1%
14) Other Pacific Islander 3 0.1% 0.1%

15) Some other race 56 1.4% 1.4%
16) Hispanic 136 3.4% 3.4%
Missing 54

336) ZODIAC
Astrological sign of respondent
RANGE: 1 to 12
N Mean Std. Deviation
Total 3676 6.816 3.337
1) Aries 237 6.4
2) Taurus 281 7.6
3) Gemini 253 6.9
4) Cancer 270 7.3
5) Leo 305 8.3
6) Virgo 327 8.9
7) Libra 319 8.7
8) Scorpio 346 9.4
9) Sagittarius 339 9.2
10) Capricorn 390 10.6
11) Aquarius 330 9.0
12) Pisces 279 7.6
Missing 356

337) WRKGOVT1
(Are/Were) you employed by the government? (Please consider federal, state, or local government.)
RANGE: 1 to 2
N Mean Std. Deviation
Total 3916 1.778 0.416
1) Yes 869 22.2
2) No 3047 77.8
Missing 116

338) WRKGOVT2
(Are/Were) you employed by a private employer (including non-profit organizations)?
RANGE: 1 to 2

N Mean Std. Deviation
Total 3919 1.408 0.491
1) Yes 2322 59.2
2) No 1597 40.8
Missing 113

339) IMMLIMIT
America should limit immigration in order to protect our national way of life.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1813 3.069 1.2
1) Strongly agree 194 10.7
2) Agree 433 23.9
3) Neither agree nor disagree 471 26.0
4) Disagree 484 26.7
5) Strongly disagree 231 12.7
Missing 2219

340) TRRESRCH
(On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely.) University research centers
RANGE: 0 to 10
N Mean Std. Deviation
Total 1788 6.247 2.383
0) No trust 58 3.2
1) 22 1.2
2) 64 3.6
3) 83 4.6
4) 93 5.2
5) 389 21.8
6) 151 8.4
7) 280 15.7
8) 343 19.2
9) 207 11.6
10) Complete trust 98 5.5

Missing 2244

341) TRMEDIA
(On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely.) The news media
RANGE: 0 to 10
N Mean Std. Deviation
Total 1820 3.61 2.732
0) No trust 372 20.4
1) 149 8.2
2) 199 10.9
3) 179 9.8
4) 150 8.2
5) 291 16.0
6) 153 8.4
7) 172 9.5
8) 105 5.8
9) 31 1.7
10) Complete trust 19 1.0
Missing 2212

342) TRBUSIND
(On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely.) Business and industry
RANGE: 0 to 10
N Mean Std. Deviation
Total 1805 4.672 1.975
0) No trust 67 3.7
1) 45 2.5
2) 134 7.4
3) 218 12.1
4) 249 13.8
5) 571 31.6
6) 210 11.6
7) 193 10.7

8) 80 4.4
9) 16 0.9
10) Complete trust 22 1.2
Missing 2227

343) TRLEGIS
(On a scale of 0 to 10, how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely.) The U.S. Congress
RANGE: 0 to 10
N Mean Std. Deviation
Total 1818 3.236 2.418
0) No trust 339 18.6
1) 195 10.7
2 197 10.8
3) 255 14.0
4) 222 12.2
5) 332 18.3
6) 108 5.9
7) 88 4.8
8) 46 2.5
9) 11 0.6
10) Complete trust 25 1.4
Missing 2214

344) CLMTCAUS
There has been a lot of discussion about the world's climate and the idea it has been changing in recent decades. Which of the following statements comes closest to your opinion?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1770 3.308 0.777
1) The world's climate has not been changing 35 2.0
2) The world's climate has been changing mostly due to natural processes 240 13.6
3) The world's climate has been changing about equally due to natural processes and human activity 639 36.1

4) The world's climate has been changing mostly due to human activity 856 48.4 Missing 2262

345) CLMTWRLD
On a scale from 0 to 10, how bad or good do you think the impacts of climate change will be for the world as a whole? 0 means extremely bad, 10 means extremely good.
RANGE: 0 to 10
N Mean Std. Deviation
Total 1737 2.876 2.359
0) Extremely bad 428 24.6
1) 137 7.9
2) 243 14.0
3) 239 13.8
4) 184 10.6
5) 346 19.9
6) 56 3.2
7) 44 2.5
8) 20 1.2
9) 15 0.9
10) Extremely good 25 1.4
Missing 2295
346) CLMTUSA
On a scale from 0 to 10, how bad or good do you think the impacts of climate change will be for America? 0 means extremely bad, 10 means extremely good.
RANGE: 0 to 10
N Mean Std. Deviation
Total 1722 3.029 2.327
0) Extremely bad 372 21.6
1) 124 7.2
2) 242 14.1
3) 263 15.3
4) 198 11.5
5) 349 20.3
6) 60 3.5
7) 52 3.0

8) 19 1.1
9) 24 1.4
10) Extremely good 19 1.1
Missing 2310

347) NATURDEV
How willing would you be to accept a reduction in the size of America's protected nature areas, in order to open them up for economic development?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1771 4.036 1.051
1) Very willing 28 1.6
2) Fairly willing 159 9.0
3) Neither willing nor unwilling 303 17.1
4) Fairly unwilling 513 29.0
5) Very unwilling 768 43.4
Missing 2261

348) INDUSGEN1
In general, do you think that air pollution caused by industry is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1779 2.142 0.863
1) Extremely dangerous for the environment 461 25.9
2) Very dangerous 688 38.7
3) Somewhat dangerous 557 31.3
4) Not very dangerous 63 3.5
5) Not dangerous at all for the environment 10 0.6
Missing 2253

349) CHEMGEN1
And do you think that pesticides and chemicals used in farming are...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1773 2.27 0.906
1) Extremely dangerous for the environment 419 23.6
2) Very dangerous 578 32.6

3) Somewhat dangerous 667 37.6
4) Not very dangerous 97 5.5
5) Not dangerous at all for the environment 12 0.7
Missing 2259

350) WATERGEN1
And do you think that pollution of America's rivers, lakes and streams is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1780 1.982 0.849
1) Extremely dangerous for the environment 591 33.2
2) Very dangerous 691 38.8
3) Somewhat dangerous 442 24.8
4) Not very dangerous 51 2.9
5) Not dangerous at all for the environment 5 0.3
Missing 2252

351) GENEGEN1
And do you think that modifying the genes of certain crops is...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1655 2.857 1.05
1) Extremely dangerous for the environment 198 12.0
2) Very dangerous 367 22.2
3) Somewhat dangerous 649 39.2
4) Not very dangerous 355 21.5
5) Not dangerous at all for the environment 86 5.2
Missing 2377

352) NUKEGEN1
And do you think that nuclear power stations are...
RANGE: 1 to 5
N Mean Std. Deviation
Total 1743 2.714 1.114
1) Extremely dangerous for the environment 309 17.7
2) Very dangerous 386 22.1

3) Somewhat dangerous 631 36.2
4) Not very dangerous 328 18.8
5) Not dangerous at all for the environment 89 5.1
Missing 2289

353) ENJOYNAT
How much, if at all, do you enjoy being outside in nature?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1792 4.07 0.876
1) Not at all 11 0.6
2) To a small extent 61 3.4
3) To some extent 380 21.2
4) To a great extent 679 37.9
5) To a very great extent 661 36.9
Missing 2240

354) ACTIVNAT
In the last 12 months how often, if at all, have you engaged in any leisure activities outside in nature, such as hiking, bird watching, swimming, skiing, other outdoor activities, or just relaxing?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1774 2.628 1.137
1) Daily 328 18.5
2) Several times a week 530 29.9
3) Several times a month 475 26.8
4) Several times a year 356 20.1
5) Never 85 4.8
Missing 2258

355) PLANETRP
In the last 12 months, how many trips did you make by plane? Count outward and return journeys, including transfers, as one trip.
RANGE: 0 to 200
N Mean Std. Deviation
Total 1806 1.118 7.11

Missing 2226

356) CARHR
In a typical week, about how many hours do you spend in a car or another motor vehicle, including motorcycles, trucks, and vans, but not counting public transport? Do not include shared rides in buses, collective taxis, or carpooling services.
RANGE: 0 to 90
N Mean Std. Deviation
Total 1800 6.313 8.605
Missing 2232

357) EATMEAT
In a typical week, on how many days do you eat beef, lamb, or products that contain them?
RANGE: 0 to 7
N Mean Std. Deviation
Total 1795 2.77 1.959
Missing 2237

358) NUMROOMS
How many rooms are there in your home (apartment or house)? Do not count any separate kitchens, bathrooms, garages, balconies, hallways or cup-boards.
RANGE: 0 to 20
N Mean Std. Deviation
Total 1812 5.343 2.795
Missing 2220

359) AIRPOLLU
(Thinking about your neighborhood, to what extent, if at all, was it affected by the following things over the last 12 months?) Air pollution
RANGE: 1 to 5
N Mean Std. Deviation
Total 1717 2.015 1.034
1) Not at all 661 38.5
2) To a small extent 563 32.8

3) To some extent 348 20.3
4) To a great extent 96 5.6
5) To a very great extent 49 2.9
Missing 2315

360) WTRPOLLU
(Thinking about your neighborhood, to what extent, if at all, was it affected by the following things over the last 12 months?) Water pollution
RANGE: 1 to 5
N Mean Std. Deviation
Total 1699 1.809 0.982
1) Not at all 854 50.3
2) To a small extent 449 26.4
3) To some extent 285 16.8
4) To a great extent 88 5.2
5) To a very great extent 23 1.4
Missing 2333

361) EXWEATHR
(Thinking about your neighborhood, to what extent, if at all, was it affected by the following things over the last 12 months?) Extreme weather events (such as severe storms, droughts, floods, heat waves, cold snaps, etc.)
RANGE: 1 to 5
N Mean Std. Deviation
Total 1756 2.574 1.079
1) Not at all 301 17.1
2) To a small extent 554 31.5
3) To some extent 591 33.7
4) To a great extent 212 12.1
5) To a very great extent 98 5.6
Missing 2276

362) MKT1
It is the responsibility of private companies to reduce the differences in pay between their employees with high pay and those with low pay.
RANGE: 1 to 5

N Mean Std. Deviation
Total 1790 2.402 1.086
1) Strongly agree 384 21.5
2) Agree 680 38.0
3) Neither agree nor disagree 425 23.7
4) Disagree 224 12.5
5) Strongly disagree 77 4.3
Missing 2242

363) RESPINEQ
From the following list, who do you think should have the greatest responsibility for reducing differences in income between people with high incomes and people with low incomes?
RANGE: 1 to 6
N Mean Std. Deviation
Total 1535 2.542 1.701
1) Private companies 515 33.6
2) Government 539 35.1
3) Trade unions 87 5.7
4) High-income individuals themselves 97 6.3
5) Low-income individuals themselves 122 7.9
6) Income differences do not need to be reduced 175 11.4
Missing 2497

364) GOVINEQ1
To what extent do you agree or disagree with the following statement: Most politicians in America do not care about reducing the differences in income between people with high incomes and people with low incomes.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1749 2.011 0.942
1) Strongly agree 593 33.9
2) Agree 692 39.6
3) Neither agree nor disagree 338 19.3
4) Disagree 103 5.9
5) Strongly disagree 23 1.3
Missing 2283

365) INEQMAD

Some people feel angry about differences in wealth between the rich and the poor, while others do not. How do you feel when you think about differences in wealth between the rich and the poor in America? Please place yourself on a scale of 0 to 10, where 0 means not angry at all and 10 means extremely angry.

RANGE: 0 to 10

	N	Mean	Std. Deviation
Total	1778	4.764	2.986
0) Not angry at all	266	15.0	
1)	76	4.3	
2)	109	6.1	
3)	124	7.0	
4)	125	7.0	
5)	356	20.0	
6)	168	9.4	
7)	229	12.9	
8)	143	8.0	
9)	50	2.8	
10) Extremely angry	132	7.4	
Missing	2254		

366) MIGRPOOR

(Turning to international differences, to what extent do you agree or disagree with the following statements?) People from poor countries should be allowed to work in wealthy countries.

RANGE: 1 to 5

	N	Mean	Std. Deviation
Total	1742	2.376	0.944
1) Strongly agree	280	16.1	
2) Agree	770	44.2	
3) Neither agree nor disagree	498	28.6	
4) Disagree	145	8.3	
5) Strongly disagree	49	2.8	
Missing	2290		

367) CONTPOOR
How often do you have any contact with people who are a lot poorer than you when you are out and about? This might be in the street, on public transport, in shops, in your neighborhood, or at your workplace.
RANGE: 1 to 7
N Mean Std. Deviation
Total 1704 4.623 1.87
1) Never 80 4.7
2) Less than once a month 240 14.1
3) Once a month 151 8.9
4) Several times a month 378 22.2
5) Once a week 128 7.5
6) Several times a week 377 22.1
7) Every day 350 20.5
Missing 2328

368) CONTRICH
How often do you have any contact with people who are a lot richer than you when you are out and about? This might be in the street, on public transport, in shops, in your neighborhood, or at your workplace.
RANGE: 1 to 7
N Mean Std. Deviation
Total 1706 3.978 1.901
1) Never 158 9.3
2) Less than once a month 349 20.5
3) Once a month 188 11.0
4) Several times a month 399 23.4
5) Once a week 110 6.4
6) Several times a week 294 17.2
7) Every day 208 12.2
Missing 2326

369) CLASS1
Most people see themselves as belonging to a particular class. Please tell me which social class you would say you belong to?
RANGE: 1 to 6
N Mean Std. Deviation

Total 1806 3.295 1.165
1) Lower class 95 5.3
2) Working class 461 25.5
3) Lower middle class 336 18.6
4) Middle class 666 36.9
5) Upper middle class 226 12.5
6) Upper class 22 1.2
Missing 2226

370) RANK1
In our society there are groups which tend to be toward the top and groups which tend to be toward the bottom. On the handcard is a scale that runs from top to bottom. Where would you put yourself now on this scale?
RANGE: 1 to 10
N Mean Std. Deviation
Total 1812 5.047 1.684
1) Top 34 1.9
2) 44 2.4
3) 228 12.6
4) 267 14.7
5) 730 40.3
6) 197 10.9
7) 165 9.1
8) 78 4.3
9) 29 1.6
10) Bottom 40 2.2
Missing 2220

371) RANK16
(In our society there are groups which tend to be toward the top and groups which tend to be toward the bottom. On the handcard is a scale that runs from top to bottom.) If you think about the family that you grew up in, where did they fit in then?
RANGE: 1 to 10
N Mean Std. Deviation
Total 1808 5.482 1.866
1) Top 34 1.9

2) 44 2.4
3) 160 8.8
4) 226 12.5
5) 599 33.1
6) 243 13.4
7) 237 13.1
8) 146 8.1
9) 64 3.5
10) Bottom 55 3.0
Missing 2224

372) RANK10FUT
(In our society there are groups which tend to be toward the top and groups which tend to be toward the bottom. On the handcard is a scale that runs from top to bottom.) And thinking ahead 10 years from now, where do you think you will be on a scale of 1 to 10, where 1 is the top and 10 the bottom?
RANGE: 1 to 10
N Mean Std. Deviation
Total 1793 4.732 1.917
1) Top 77 4.3
2) 106 5.9
3) 282 15.7
4) 317 17.7
5) 554 30.9
6) 170 9.5
7) 129 7.2
8) 82 4.6
9) 27 1.5
10) Bottom 49 2.7
Missing 2239

373) FAIRDIST
How fair or unfair do you think the income distribution is in America?
RANGE: 1 to 4
N Mean Std. Deviation
Total 1652 2.887 0.797

1) Very fair 74 4.5
2) Fair 406 24.6
3) Unfair 805 48.7
4) Very unfair 367 22.2
Missing 2380

374) ENDSME12
And during the next 12 months, how difficult or easy do you think it will be for your household to make ends meet?
RANGE: 1 to 5
N Mean Std. Deviation
Total 1752 3.244 1.162
1) Very difficult 132 7.5
2) Fairly difficult 349 19.9
3) Neither easy nor difficult 507 28.9
4) Fairly easy 488 27.9
5) Very easy 276 15.8
Missing 2280

375) SKIPMEAL
How often do you or other members of your household skip a meal because there is not enough money for food?
RANGE: 1 to 7
N Mean Std. Deviation
Total 1767 1.421 1.119
1) Never 1471 83.2
2) Less than once a month 107 6.1
3) Once a month 46 2.6
4) Several times a month 82 4.6
5) Once a week 18 1.0
6) Several times a week 31 1.8
7) Every day 12 0.7
Missing 2265

376) RATEPAIN1
On a scale from 0 to 10, with 0 meaning no pain and 10 being the worst imaginable pain, how would you rate your pain on average?

RANGE: 0 to 10
N Mean Std. Deviation
Total 3613 2.7 2.481
0) No pain 828 22.9
1) 620 17.2
2) 600 16.6
3) 430 11.9
4) 258 7.1
5) 302 8.4
6) 211 5.8
7) 173 4.8
8) 114 3.2
9) 56 1.5
10) The worst imaginable pain 21 0.6
Missing 419

377) RELIGIMP
How important is religion in your life—very important, somewhat important, not too important, or not at all important?
RANGE: 1 to 4
N Mean Std. Deviation
Total 3609 2.369 1.163
1) Very important 1124 31.1
2) Somewhat important 923 25.6
3) Not too important 668 18.5
4) Not at all important 894 24.8
Missing 423

378) RELIDIMP
When thinking about religion, how important is being (a Christian/a Catholic/a Jew/a Buddhist/a Hindu/a Muslim/an atheist/an agnostic/someone who does not identify with a religion/ a member of your religion) to you?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3511 2.797 1.416
1) Extremely important 855 24.4
2) Very important 776 22.1

3) Somewhat important 711 20.3
4) Not too important 563 16.0
5) Not at all important 606 17.3
Missing 521

379) RELIDESC
How well does the term (Christian/Catholic/Jew/Buddhist/Hindu/Muslim/atheist/agnostic/non-religious/N/A) describe you?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3450 2.497 1.086
1) Extremely well 689 20.0
2) Very well 1085 31.4
3) Somewhat well 1130 32.8
4) Not very well 363 10.5
5) Not well at all 183 5.3
Missing 582

380) RELIDWE
When talking about (Christians/Catholics/Jews/Buddhists/Hindus/Muslims/atheists/agnostics/people who do not identify with a religion/your religion), how often do you say 'we' instead of 'they'?
RANGE: 1 to 5
N Mean Std. Deviation
Total 3432 2.859 1.352
1) Never 785 22.9
2) Rarely 586 17.1
3) Some of the time 857 25.0
4) Most of the time 735 21.4
5) All of the time 469 13.7 Missing 600
381) RELIDINS
If someone criticized (Christians/Catholics/Jews/Buddhists/Hindus/Muslims/Atheists/Agnostics/people who do not identify with a religion/your religion), to what extent would it feel like a personal insult?
RANGE: 1 to 4
N Mean Std. Deviation
Total 3496 2.682 1.024

1) A great deal 483 13.8
2) Somewhat 1115 31.9
3) Very little 927 26.5
4) Not at all 971 27.8
Missing 536

382) SPRTCONNCT
(Some people say they have experiences of being personally moved, touched, or inspired, while others say they do not have these experiences at all. How often, if at all, do you experience each of the following?) Felt particularly connected to the world around you.
RANGE: 1 to 7
N Mean Std. Deviation
Total 3523 3.811 1.859
1) At least once a day 324 9.2
2) Almost every day 828 23.5
3) Once or twice a week 514 14.6
4) Once or twice a month 503 14.3
5) A few times per year 654 18.6
6) Once a year or less 279 7.9
7) Never 421 12.0
Missing 509

383) SPRTLRGR
(Some people say they have experiences of being personally moved, touched, or inspired, while others say they do not have these experiences at all. How often, if at all, do you experience each of the following?) Felt like you were part of something much larger than yourself.
RANGE: 1 to 7
N Mean Std. Deviation
Total 3544 4.095 1.951
1) At least once a day 334 9.4
2) Almost every day 693 19.6
3) Once or twice a week 442 12.5
4) Once or twice a month 421 11.9
5) A few times per year 693 19.6
6) Once a year or less 409 11.5

7) Never 552 15.6
Missing 488

384) SPRTPURP
(Some people say they have experiences of being personally moved, touched, or inspired, while others say they do not have these experiences at all. How often, if at all, do you experience each of the following?) Felt a sense of a larger meaning or purpose in life.
RANGE: 1 to 7
N Mean Std. Deviation
Total 3537 3.996 1.949
1) At least once a day 371 10.5
2) Almost every day 708 20.0
3) Once or twice a week 460 13.0
4) Once or twice a month 428 12.1
5) A few times per year 672 19.0
6) Once a year or less 390 11.0
7) Never 508 14.4
Missing 495

385) MDITATE1
How often do you meditate?
RANGE: 1 to 7
N Mean Std. Deviation
Total 3574 4.803 2.185
1) At least once a day 336 9.4
2) Almost every day 402 11.2
3) Once or twice a week 461 12.9
4) Once or twice a month 343 9.6
5) A few times per year 374 10.5
6) Once a year or less 204 5.7
7) Never 1454 40.7
Missing 458

386) GRTWRKS
(Please indicate to what extent you agree or disagree with each of the following statements.) The great works of philosophy and science are the best

source of truth, wisdom, and ethics.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3547 2.796 0.942
1) Strongly agree 297 8.4
2) Agree 959 27.0
3) Neither agree nor disagree 1624 45.8
4) Disagree 506 14.3
5) Strongly disagree 161 4.5
Missing 485

387) FREEMIND
(Please indicate to what extent you agree or disagree with each of the following statements.) To understand the world, we must free our minds from old traditions and beliefs.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3556 2.893 1.05
1) Strongly agree 345 9.7
2) Agree 902 25.4
3) Neither agree nor disagree 1334 37.5
4) Disagree 740 20.8
5) Strongly disagree 235 6.6
Missing 476

388) DECEVIDC
(Please indicate to what extent you agree or disagree with each of the following statements.) When I make important decisions in my life, I rely mostly on reason and evidence.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3564 2.078 0.847
1) Strongly agree 833 23.4
2) Agree 1888 53.0
3) Neither agree nor disagree 626 17.6
4) Disagree 167 4.7
5) Strongly disagree 50 1.4 Missing 468

389) ADVFMSCI
(Please indicate to what extent you agree or disagree with each of the following statements.) All of the greatest advances for humanity have come from science and technology.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3558 2.734 0.997
1) Strongly agree 366 10.3
2) Agree 1135 31.9
3) Neither agree nor disagree 1271 35.7
4) Disagree 652 18.3
5) Strongly disagree 134 3.8
Missing 474

390) GODUSA
(Please indicate to what extent you agree or disagree with each of the following statements.) The success of the United States is part of God's plan.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3553 3.285 1.264
1) Strongly agree 326 9.2
2) Agree 607 17.1
3) Neither agree nor disagree 1238 34.8
4) Disagree 493 13.9
5) Strongly disagree 889 25.0
Missing 479

391) GOVCHRST
(Please indicate to what extent you agree or disagree with each of the following statements.) The federal government should advocate Christian values.
RANGE: 1 to 5
N Mean Std. Deviation
Total 3559 3.367 1.246
1) Strongly agree 287 8.1
2) Agree 598 16.8

3) Neither agree nor disagree 1087 30.5
4) Disagree 696 19.6
5) Strongly disagree 891 25.0
Missing 473

392) TRDUNIO1
To what extent do you agree or disagree with the following statements?
Workers need strong trade unions to protect their interests.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1697 2.666 1.109
1) Strongly agree 246 14.5
2) Agree 555 32.7
3) Neither agree nor disagree 542 31.9
4) Disagree 227 13.4
5) Strongly disagree 127 7.5
Missing 2335

393) BOARDREP
(To what extent do you agree or disagree with the following statements?)
Workers should be represented on the board of directors at major companies.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1701 2.098 0.888
1) Strongly agree 427 25.1
2) Agree 820 48.2
3) Neither agree nor disagree 342 20.1
4) Disagree 84 4.9
5) Strongly disagree 28 1.6
Missing 2331

394) UPWAGES
(To what extent do you agree or disagree with the following statements?)
The government should ensure that the wages of low-paying jobs increase as the economy grows.
RANGE: 1 to 5
N Mean Std. Deviation

Total 1714 2.127 1.075
1) Strongly agree 537 31.3
2) Agree 708 41.3
3) Neither agree nor disagree 254 14.8
4) Disagree 144 8.4
5) Strongly disagree 71 4.1
Missing 2318
395) LIMITPAY
(To what extent do you agree or disagree with the following statements?) The government should take steps to limit the pay of executives at major companies.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1697 2.864 1.215
1) Strongly agree 259 15.3
2) Agree 432 25.5
3) Neither agree nor disagree 459 27.0
4) Disagree 374 22.0
5) Strongly disagree 173 10.2
Missing 2335

396) PRFTIMPV
(To what extent do you agree or disagree with the following statements?) Allowing business to make good profits is the best way to improve every-one's standard of living.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1690 2.649 1.007
1) Strongly agree 184 10.9
2) Agree 627 37.1
3) Neither agree nor disagree 555 32.8
4) Disagree 246 14.6
5) Strongly disagree 78 4.6
Missing 2342

397) GOVFNANC
(To what extent do you agree or disagree with the following statements?)

The government should finance projects to create new jobs, even if it might require a tax increase to pay for it.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1698 2.756 1.057
1) Strongly agree 180 10.6
2) Agree 566 33.3
3) Neither agree nor disagree 540 31.8
4) Disagree 312 18.4
5) Strongly disagree 100 5.9
Missing 2334

398) DCLINDUS
(To what extent do you agree or disagree with the following statements?) The government should support declining industries to protect jobs, even if it might require a tax increase to pay for it.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1677 3.291 0.991
1) Strongly agree 54 3.2
2) Agree 311 18.5
3) Neither agree nor disagree 582 34.7
4) Disagree 553 33.0
5) Strongly disagree 177 10.6
Missing 2355

399) GOVFNAID
(To what extent do you agree or disagree with the following statements?) The government should give financial assistance to college students from low-income families, even if it might require a tax increase to pay for it.
RANGE: 1 to 5
N Mean Std. Deviation
Total 1713 2.521 1.084
1) Strongly agree 281 16.4
2) Agree 681 39.8
3) Neither agree nor disagree 419 24.5
4) Disagree 242 14.1

5) Strongly disagree 90 5.3
Missing 2319

400) HIVAFRAID
Please indicate how much you agree or disagree with each of the following comments. I would be afraid to be around a person with HIV because I would be worried I could get infected.
RANGE: 1 to 4
N Mean Std. Deviation
Total 2174 1.685 0.933
1) Strongly disagree 1249 57.5
2) Disagree 508 23.4
3) Agree 270 12.4
4) Strongly agree 147 6.8
Missing 1858

401) HIVIMMRL
(Please indicate how much you agree or disagree with each of the following comments.) People who have HIV have participated in immoral activities.
RANGE: 1 to 4
N Mean Std. Deviation
Total 2032 1.653 0.875
1) Strongly disagree 1163 57.2
2) Disagree 500 24.6
3) Agree 280 13.8
4) Strongly agree 89 4.4
Missing 2000

402) HIVDSCRM
(Please indicate how much you agree or disagree with each of the following comments.) There is a lot of discrimination against people with HIV in this country today.
RANGE: 1 to 4
N Mean Std. Deviation
Total 1976 2.781 0.848
1) Strongly disagree 161 8.1
2) Disagree 490 24.8

3) Agree 945 47.8
4) Strongly agree 380 19.2
Missing 2056

403) STRVBIAS
Which do you think should be the bigger priority for the U.S. criminal justice system today?
RANGE: 1 to 2
N Mean Std. Deviation
Total 1149 1.665 0.472
1) Strengthening law and order through more police and greater enforcement of the laws 385 33.5
2) Reducing bias against minorities in the criminal justice system by reforming court and police practices 764 66.5
Missing 2883

404) RACESURV17
How much, if at all, do you think the legacy of slavery affects the position of Black people in American society today?
RANGE: 1 to 4
N Mean Std. Deviation
Total 2340 2.168 1.064
1) A great deal 802 34.3
2) A fair amount 704 30.1
3) Not much 474 20.3
4) Not at all 360 15.4
Missing 1692

405) DEFUND
Do you favor or oppose reducing funding for police departments, and moving those funds to mental health, housing, and other social services?
RANGE: 1 to 2
N Mean Std. Deviation
Total 2344 1.594 0.491
1) Favor 952 40.6
2) Oppose 1392 59.4
Missing 1688

406) POLTRTBLK

The following questions are about police and law enforcement. In general, do the police (treat Whites better than Blacks, treat them both the same, or treat Blacks better than Whites/treat Blacks better than Whites, treat them both the same, or treat Whites better than Blacks)?

RANGE: 1 to 7

	N	Mean	Std. Deviation
Total	2328	2.382	1.344
1) Treat Whites much better than Blacks	914	39.3	
2) Treat Whites moderately better than Blacks	452	19.4	
3) Treat Whites a little better than Blacks	167	7.2	
4) Treat both the same	765	32.9	
5) Treat Blacks a little better than Whites	19	0.8	
6) Treat Blacks moderately better than Whites	5	0.2	
7) Treat Blacks much better than Whites	6	0.3	
Missing	1704		

407) POLTRTHSP

In general, do the police (treat Whites better than Latinos, treat them both the same, or treat Latinos better than Whites/treat Latinos better than Whites, treat them both the same, or treat Whites better than Latinos)?

RANGE: 1 to 7

	N	Mean	Std. Deviation
Total	2315	2.513	1.318
1) Treat Whites much better than Latinos	761	32.9	
2) Treat Whites moderately better than Latinos	509	22.0	
3) Treat Whites a little better than Latinos	174	7.5	
4) Treat both the same	856	37.0	
5) Treat Latinos a little better than Whites	3	0.1	
6) Treat Latinos moderately better than Whites	6	0.3	
7) Treat Latinos much better than Whites	6	0.3	
Missing	1717		

Media Attributions

- gss screenshot from codebook © NORC is licensed under a All Rights Reserved license

Works Cited

Adler, Emily Stier, and Roger Clark. 2008. *How It's Done: An Invitation to Social Research*, 3rd. Belmont, CA: Wadsworth.

American Sociological Association. 2019. *American Sociological Association Style Guide*, 6th Edition. Washington, D.C.: American Sociological Association.

Bartsch, Robert A., Teresa Burnett, Tommye R. Diller, and Elizabeth Rankin-Williams. 2000. "Gender Representation in Television Commercials: Updating an Update." *Sex Roles* 43(9): 735-43.

Berg, Bruce L. 2009. *Qualitative Research Methods for the Social Sciences*, 7th. Boston: Allyn and Bacon.

Bergin, Tiffany. 2018. *An Introduction to Data Analysis: Quantitative, Qualitative, and Mixed Methods*. London: Sage Publications.

Boole, George. 1848. "The Calculus of Logic." *Cambridge and Dublin Mathematical Journal* III:183-98.

Brewster, Signe. 2020. "The Best Transcription Services." Wirecutter. Website accessed 07/28/2020 (https://www.nytimes.com/wirecutter/reviews/best-transcription-services/).

Center for American Women and Politics. 2022. "New Records for Women in the U.S. Congress and House." *Rutgers Eagleton Institute of Politics*. Website accessed 12/12/2022 (https://cawp.rutgers.edu/news-media/press-releases/new-records-women-us-congress-and-house).

Cite Black Women Collective. n.d. "Cite Black Women." Website accessed 06/24/2020, (https://www.citeblackwomencollective.org/).

Davern, Michael; Bautista, Rene; Freese, Jeremy; Morgan, Stephen L.; and Tom W. Smith. General Social Survey 2021 Cross-section. [Machine-readable data file]. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Stephen L. Morgan, and Tom W. Smith. NORC ed. Chicago, 2021. 1 datafile (68,846 cases) and 1 codebook (506 pages).

Desmond, Matthew. 2012. "Disposable Ties and the Urban Poor." *American Journal of Sociology* 117(5):1295-1335.

Edin, Kathryn. 2000. "What Do Low-Income Single Mothers Say about Marriage?" *Social Problems* 47(1):112-33.

Fernandez, Mary Ellen. 2019. "Trump and Clinton Rallies: Are Political Campaigns Quasi-Religious in Nature?" *Sociology Between the Gaps: Forgotten and Neglected Topics* 4(1):1-10.

Gauthier, Jessica, Madeline MacKay, Madison Mellor, and Roger Clark. 2020. "Ebbs and Flows in the Feminist Presentation of Female Characters Among Caldecott Award-Win-

ning Picture Books for Children." *Sociology Between the Gaps: Forgotten and Neglected Topics* 5(1): 1-20.

Gerring, John. 2007. *Case Study Research: Principles and Practices*. Cambridge, UK: Cambridge University Press.

Grbich, Carol. 2007. *Qualitative Data Analysis: An Introduction*. London: Sage.

Hsiung, Ping-Chun. 2008. "Teaching Reflexivity in Qualitative Interviewing." *Teaching Sociology* 36(July):211-26.

Inter-Parliamentary Union. 2022. "Monthly Ranking of Women in National Parliaments." IPU Parline. Website accessed 12/12/20222 (https://data.ipu.org/women-ranking).

Jones, Taylor, Jessica Rose Kalbfeld, Ryan Hancock, and Robin Clark. 2019. "Testifying While Black: An Experimental Study of Court Reporter Accuracy in Transcription of African American English." *Language* 95(2):e216-52.

Kearney, Melissa S. and Phillip B. Levine. 2015. "Media Influences on Social Outcomes: The Impact of MTV's *16 and Pregnant* on Teen Childbearing." 105(12):3597-3632.

Khamsi, Roxanne. 2019. "Say What? A Non-Scientific Comparison of Automated Transcription Services." The Open Notebook. Website accessed 07/28/2020 (https://www.theopennotebook.com/2019/12/17/
say-what-a-non-scientific-comparison-of-automated-transcription-services/).

Khan, Shamus. 2019. "The Subpoena of Ethnographic Data." *Sociological Forum* 34(1):253-63.

Lovdal, Lynn T. 1989. "Sex Role Messages in Television Commercials: An Update." *Sex Roles* 21(11):715-24.

Luker, Kristin. 2008. *Salsa Dancing into the Social Sciences: Research in an Age of Info-glut*. Cambridge, MA: Harvard University Press.

Manning, Jimmie. 2017. "In Vivo Coding." In *The International Encyclopedia of Communication Research Methods*, edited by Jörg Matthes, Christine S. Davis, and Robert F. Potter. New York: Wiley-Blackwell.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd Edition. Thousand Oaks, CA: Sage.

National Center for Health Statistics. 2022. "Firearm Mortality by State." Centers for Disease Control and Prevention. Website accessed 09/19/2022 (https://www.cdc.gov/nchs/pressroom/sosmap/firearm_mortality/firearm.htm).

O'Donnell, William J., and Karen J. O'Donnell. 1978. "Update: Sex-role Messages in TV Commercials." *Journal of Communication* 28(1): 156-58.

Pearce, Lisa D. 2012. "Mixed Methods Inquiry in Sociology." *American Behavioral Scientist* 56(6):829-48.

Plesser, Hans E. 2017. "Reproducibility vs. Replicability: A Brief History of a Confused Terminology." *Frontiers in Neuroinformatics* 11(76), accessed online (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115/).

Posselt, Julie. 2017. “4 tips for using quotes in #qualitative findings, after a month of much reviewing -journal articles, conference proposals, & student writing.” Twitter thread accessed 08/13/2020 (https://web.archive.org/web/20221213180729/https://twitter.com/JuliePosselt/status/883050014024450049).

Petit, Lindsay, Madison Mellor, and Roger Clark. “The Gender Gap in Political Affiliation: Understanding Why It Emerged and Maintained Itself Over Time.” *Sociology Between the Gaps: Forgotten and Neglected Topics* 5(1):1-13.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Roberts, Keith A. 1993. “Toward a Sociology of Writing.” *Teaching Sociology* 21(4):317-324.

Rosenberg, Morris. 1968. *The Logic of Survey Analysis*. New York: Basic Books.

Saldaña, Johnny. 2016. *The Coding Manual for Qualitative Researchers*. London: Sage.

Sanderson, Theo. n.d. “The Up-Goer Five Text Editor.” Website accessed 08/13/2020 (https://www.splasho.com/upgoer5/).

Schell, Terry L. et al. 2020. “State-Level Estimates of Household Firearm Ownership.” RAND Corporation report TL-354-LJAF. Website accessed 09/19/2022 (https://www.rand.org/pubs/tools/TL354.html).

Schwartz, Martin A. 2008. “The Importance of Stupidity in Scientific Research.” *Journal of Cell Science* 121(11):1771.

Smith, Dorothy E. 1987. *The Everyday World as Problematic: A Feminist Sociology*. Boston, MA: Northeastern University Press.

Taylor, Steven J., Robert Bogdan, and Marjorie L. DeVault. 2016. *Introduction to Qualitative Research Methods: A Guidebook and Resource*, Fourth edition. Hoboken, NJ: John Wiley & Sons.

Teczar, Rebecca, Katherine Rocha, Joseph Palazzo, and Roger Clark. 2018. “Cultural Attitudes towards Women in Politics and Women’s Political Representation in Legislatures and Cabinet Ministries.” *Sociology Between the Gaps: Forgotten and Neglected Topics* 4(1):1-7.

Thomas, Charlie. 2020/2021. “SDA: Survey Documentation and Analysis Archive.” Social Data Archive. Website accessed December 12, 2022 (https://sda.berkeley.edu/archive.htm).

Twine, France Winddance, and Jonathan W. Warren. 2000. *Racing Research, Researching Race: Methodological Dilemmas in Critical Race Studies*. New York: New York University Press.

University of Surrey. n.d. “Computer Assisted Qualitative Data Analysis (CAQDAS) Networking Project.” Website accessed 7/30/2020 (https://www.surrey.ac.uk/computer-assisted-qualitative-data-analysis).

Van Den Berg, Harry, Margaret Wetherell, and Hanneke Houtkoop-Steenstra. 2003. *Analyz-*

ing Race Talk: Multidisciplinary Approaches on the Research Interview. Cambridge, UK: Cambridge University Press.

Warren, Carol A. B., and Tracy Xavia Karner. 2015. *Discovering Qualitative Methods: Ethnography, Interviews, Documents, and Images*. Oxford: Oxford University Press.

Zuberi, Tukufu, and Eduardo Bonilla-Silva. 2008. *White Logic, White Methods: Racism and Methodology*. Lanham, MD: Rowman & Littlefield.

About the Authors

Mikaila Mariel Lemonik Arthur is Professor of Sociology at Rhode Island College, where she has taught a wide variety of courses including Social Research Methods, Social Data Analysis, Senior Seminar in Sociology, Professional Writing for Justice Services, Comparative Law and Justice, Law and Society, Comparative Perspectives on Higher Education, and Race and Justice. She has written a number of books and articles, including both those with a pedagogical focus (including *Law and Justice Around the World*, published by the University of California Press) and those focusing on her scholarly expertise in higher education (including *Student Activism and Curricular Change in Higher Education*, published by Routledge). She has expertise and experience in academic program review, translating research findings for policymakers, and disability accessibility in higher education, and has served as a department chair and as Vice President of the RIC/AFT, her faculty union. Outside of work, she enjoys reading speculative fiction, eating delicious vegan food, visiting the ocean, and spending time with amazing humans.

Roger Clark is Professor Emeritus of Sociology at Rhode Island College, where he continues to teach courses in Social Research Methods and Social Data Analysis and to coauthor empirical research articles with undergraduate students. He has coauthored two textbooks, *An Invitation to Social Research* (with Emily Stier Adler) and *Gender Inequality in Our Changing World: A Comparative Approach* (with Lori Kenschaft and Desirée Ciambrone). He has been ranked by the USTA in its New England 60- and 65-and-older divisions, shot four holes in one on genuine golf courses, and run multiple half and full marathons. Like the Energizer Bunny, he keeps on going and going, but, given his age, leaves it to your imagination where.