

UNDERSTANDING DATA: CRITICAL CONCEPTS

Before analyzing inequalities with survey data, it is important to be familiar with some of the basic concepts in survey research and data analysis. This chapter begins with an introduction to the General Social Survey (GSS), which is one of the most widely analyzed surveys in the United States. Following this introduction, the chapter presents some basic techniques for finding and interpreting variables in the GSS.

Data from the 1972–2014 General Social Surveys can be easily accessed and analyzed through the Institute for Scientific Analysis's Survey Documentation and Analysis (SDA) website (<http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14>). You can also access this same page using this link: <http://tinyurl.com/GSS72to14>. The SDA is free for anyone to use and can be used on any computer, tablet, or mobile device.

OVERVIEW OF THE GSS¹

The General Social Survey is a national survey of adults in the United States that has been conducted regularly since 1972. From 1972 to 1993, the survey was administered almost annually, and it contained approximately 1,400 interviews each year. From 1994 on, the survey has been conducted every other year. It contains numerous questions about respondents' sociodemographic characteristics (e.g., respondents' age, sex, income, and occupation); a range of questions about respondents' behaviors (e.g., whether respondents voted in the last election, whether they participate in voluntary organizations, how frequently they use email, and how frequently they attend religious services);

Learning objectives

By the end of this chapter, you should be able to:

1. Describe the General Social Survey.
2. Use the SDA website to identify variables of interest to you.
3. Use the SDA website to identify the mode and, where applicable, the median and mean of each variable.
4. Describe a range of variables, including the survey question associated with each variable, the response categories for each variable, and level of measurement.
5. Interpret a univariate frequency table.

and events that may have occurred in respondents' lives (e.g., if they have been arrested, if they are married or divorced, and if they have experienced discrimination).

The GSS also includes a number of questions about what respondents believe (e.g., if they believe the government should spend more money to help the poor, or if they believe abortion should be legal under all circumstances). It also includes questions designed to tap respondents' knowledge—there is a small vocabulary test, and recent surveys have included questions assessing respondents' knowledge of science and perceptions of scientific careers (e.g., “Did the universe begin with a big bang?” “Is astrology a science?” and “If you had a daughter, how would you feel if she wanted to be a scientist?”). There are also a range of questions concerning respondents' physical health, psychological health, and feelings (e.g., how often respondents feel in poor mental health; how often respondents feel in poor physical health; how happy respondents are; and—for those who are married—how happy their marriages are).

Some of the GSS questions (such as respondents' age, marital status, and educational attainment) are asked every year to all respondents. Other survey questions are included in only a handful of years, and still others are included in only one or two years and are asked to a randomly selected subsample of respondents. For this reason, whenever analyzing data from the GSS, it is important to make sure you understand when the survey questions you are analyzing were included in the survey. When sharing your results in a paper, presentation, or other type of project, you should be sure to explain this to your audience.

Who Is Included in the GSS Sample?

The GSS includes respondents who are 18 years of age or older and living in households at the time of the survey. It includes both citizens and noncitizens of the US. The requirement that people be living in households means that individuals who live in institutions such as college dorms, nursing homes, prisons, or jails, and also those who are homeless, are not included in the GSS sample. From 1972 to 2004, the GSS included only individuals who were able to speak English. Since 2006, the GSS has been administered in Spanish also, so now both Spanish and English speakers are represented in the survey.

How Are the Data Collected?

As described on its website, the GSS relies primarily on face-to-face interviews, though in some cases, when it is difficult to arrange an in-person interview, interviews are conducted by telephone. All interviewers undergo extensive training before conducting any interviews.

What Are Variable Weights?

Throughout this book, data will be analyzed using variable weights. In brief, variable weights are values assigned to each observation (in the GSS, each person who takes

the survey can be thought of as an “observation”), and the use of these weights helps to ensure that the sample of respondents matches the overall population from which the sample was drawn. For example, during the years 1975 to 2002, the GSS used a sampling design that gave each household in the US an equal probability of being included in the GSS. Because only one adult per household is allowed to be interviewed, those who live in larger households had a slightly lower probability of being included in the GSS sample. Weights are used to adjust for this small discrepancy. In the 1982 and 1987 surveys, the sample included an “oversample” of Black respondents. Survey designers often deliberately include oversamples of minority groups so that they can be sure to have a sufficient number of minority respondents in the final sample. If one were to analyze data from these years without using variable weights, the responses of Black respondents would be disproportionately represented. Using the variable weights is an easy way to adjust for the disproportionately higher number of Black respondents or other groups that are oversampled in particular years.

In this book, all analyses are conducted using the weighting variable COMPWT, which is the default selection in the Survey Documentation and Analysis (SDA) program. For the most part, if you were to perform basic analyses of data with and without this weight, the results would be similar. Below is an example of the variable HAPPY, which corresponds to the survey question “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?” Figure 2.1 shows the frequency distribution for this variable when no weights are used. Figure 2.2 shows the frequency distribution using the weight COMPWT.

Figure 2.1	
UN-Weighted Frequency Distribution (HAPPY)	
Cells contain: -Column percent -N of cases	Distribution
1: VERY HAPPY	31.6 17,316
2: PRETTY HAPPY	55.9 30,655
3: NOT TOO HAPPY	12.5 6,880
COL TOTAL	100.0 54,851

Figure 2.2

<i>Weighted Frequency Distribution (Happy)</i>	
Cells contain: -Column percent -Weighted N	Distribution
1: VERY HAPPY	33.2 18,253.3
2: PRETTY HAPPY	55.4 30,416.8
3: NOT TOO HAPPY	11.4 6,247.3
COL TOTAL	100.0 54,917.4

The top numbers in each of these represent the percentage of respondents who select each response. Without the weights (Figure 2.1), it appears that 31.6% of respondents describe themselves as “very happy.” With the weights (Figure 2.2), we see that 33.2% of respondents describe themselves as “very happy.” The bottom number in each cell corresponds to the number of cases in each category of the variable. In Figure 2.1, for example, we see that 6,880 people described themselves as “not too happy” (corresponding to 12.5% of respondents who provided valid answers to the survey question). In Figure 2.2, the bottom number in each cell still represents the number of observations that fall into each category of the variable, but rather than the *raw number of cases*, the weighted frequency distribution produces a *weighted number of cases*. When interpreting the 6,247.3, we could say that “about 6,247 people reported being ‘not too happy.’” An even better interpretation would be “11.4% of respondents reported being ‘not too happy,’ which corresponds to approximately 6,247 people.” More information on constructing and interpreting frequency distributions is provided in the latter part of this chapter and throughout the rest of this book.

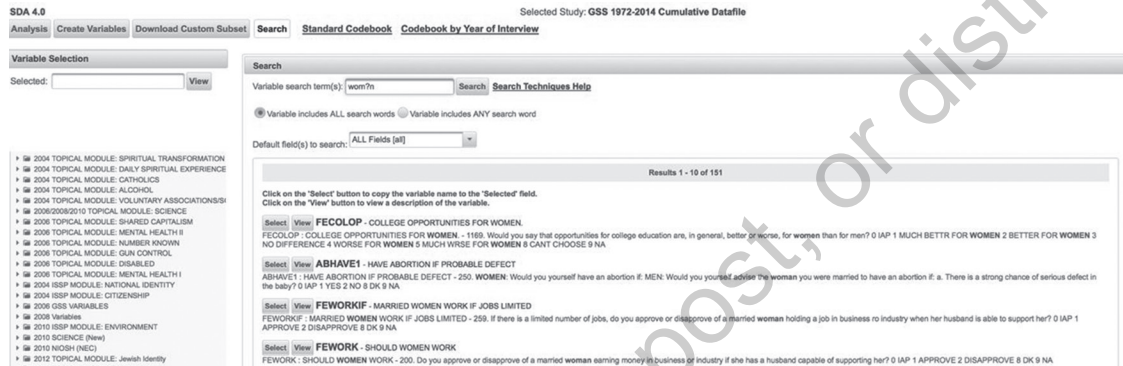
FINDING DATA IN THE GSS

Searching for Variables

Starting from the homepage of the GSS SDA website, click on the “Search” button toward the top of the screen. This brings us to a new screen with a prominent field on the right for entering search terms. Capitalization doesn’t matter here, so don’t worry about that. When searching for variables, you can also use * and ? as “wildcard characters.” As shown in Figure 2.3, typing in wom?N will produce a search for all

variables that include the words *woman* or *women*, for example. Typing in *WOMEN**, on the other hand, will produce a search for all variables that include the term *women* or *women's* but won't search for variables that simply use the singular *woman*.

Figure 2.3



To get started, type *gender* in the “Variable search term(s)” box, and then click the “Search” button. The results appear on the bottom portion of the screen and show us that our search produced 25 hits. These results show us that, since 1972, there have been 25 unique variables that have used the term *gender* somewhere. The term *gender* might appear in the variable name itself, in the wording of the question asked to the respondent, or in the response categories.

To get more information about any of the 25 variables, simply click on the “View” button next to the variable that interests you. If we look at the variable *WKSEXISM*, for example, clicking on the “View” button brings up a new screen that shows us the precise wording of the survey question: “Do you feel in any way discriminated against on your job because of your gender?”

“Viewing” a variable in this way also produces a frequency table—a table displaying the response categories for each variable as well as the number of responses that fall into each category. To exit this screen and to return to the search results, simply click outside the pop-up window. Alternatively, click on the small “x” in the bottom right-hand corner of the pop-up window. More information on interpreting a frequency table is provided under “Viewing the Survey Question, Response Categories, and Variable Distribution” below.

Browsing for Variables

In addition to searching for variables with the search function, a good resource for identifying variables related to gender or any other issue is the SDA virtual codebook, located in the left-hand side of the SDA interface. The codebook is a topical tree, where headings are used to group variables on similar topics. The first subject heading is

SEARCH TIP

When searching for variables, remember that the search engine does not recognize slang, abbreviations, acronyms, or texting language!

- Avoid abbreviations and acronyms.
 - For example, type out VETERANS rather than VETS.
 - Search for GAY, LESBIAN, HOMOSEXUAL, OR BISEXUAL rather than LGBT.
- Try to use more formal words like CHILDREN and MOTHER rather than KIDS and MOM.
- If searching for variables about respondents' sexual behaviors, for example, search for SEX or even CASUAL SEX rather than HOOK-UP.
- When searching for variables about the number of children a respondent has or would like to have, type in NUMBER OF CHILDREN rather than # OF KIDS.

“Case Identification and Year.” The next is “Respondent Background Variables,” which includes a range of information on respondents’ race, sex, age, socioeconomic status, and other topics. Clicking on the small triangle to the left of “Respondent Background Variables” will expand the topic and show a list of subheadings. Clicking again on the small triangle to the left of any of these subheadings (e.g., “Education”) will finally bring us to a list of variables, as shown in Figure 2.4.



















Each variable name is presented in capital letters and is followed by a brief description of the variable. For example, the variable EDUC is described as “highest year of school completed”; DEGREE is the respondent’s highest degree (RS is an abbreviation for *respondent’s*); and the variable MAEDUC is described as “highest year school completed, mother.”

In addition to browsing for variables using the topical tree, you can browse for variables using the “Standard Codebook,” the link to which is appears in blue text at the top of the SDA page.

Viewing the Survey Question, Response Categories, and Variable Distribution

To gain more precise information about the meaning of a variable, you can easily view the variable in one of two ways. Double-clicking on either the name or short description of a variable inside the topical tree will add the variable name into the “Variable Selection” box in the upper left-hand corner of the screen. Alternatively, you can simply type the name of the variable into this field. Capitalization doesn’t matter, but spelling does, so be sure to spell it correctly! Regardless of how you get the variable name to appear in the “Variable Selection” field, once you do, clicking the “View” button will open up a new window with detailed information

Figure 2.4

- ▶  CASE IDENTIFICATION AND YEAR
- ▼  RESPONDENT BACKGROUND VARIABLES
 - ▶  Age, Gender, Race, and Ethnicity
 - ▼  Education
 - ▢ EDUC - HIGHEST YEAR OF SCHOOL COMPLETED
 - ▢ PAEDUC - HIGHEST YEAR SCHOOL COMPLETED, FATHER
 - ▢ MAEDUC - HIGHEST YEAR SCHOOL COMPLETED, MOTHER
 - ▢ SPEDUC - HIGHEST YEAR SCHOOL COMPLETED, SPOUSE
 - ▢ DEGREE - RS HIGHEST DEGREE
 - ▢ PADEG - FATHERS HIGHEST DEGREE
 - ▢ MADEG - MOTHERS HIGHEST DEGREE
 - ▢ SPDEG - SPOUSES HIGHEST DEGREE
- ▶  2012 CORE
- ▶  Military Service
- ▶  Respondent and Spouse Work Week
- ▶  Respondent's Dwelling
- ▶  Word Association
- ▶  R's Activities
- ▶  Religious Attendance and Identity
- ▶  Respondent's Household Composition
- ▶  Socio-Economic and Status Indicators
- ▶  Other Respondent Background Variables
- ▶  PERSONAL AND FAMILY INFORMATION
- ▶  ATTITUDINAL MEASURES - NATIONAL PROBLEMS
- ▶  PERSONAL CONCERNS
- ▶  SOCIETAL CONCERNS

about the variable you selected. You can find the variable HRS1 in the topical tree under “Personal and Family Information,” under the subheading for “Respondent’s Employment.” Viewing the variable HRS1 tells us that this variable corresponds to the question “If working, full or part time: how many hours did you work last week, at all jobs?” Viewing the variable in this way tells us that this question was asked only to those respondents who indicated (previously in the survey) that they were currently working either full or part time.

Toward the end of the codebook are a series of “Topical Modules” that correspond to variables included in specific years of the GSS—for example, “2000 Topical Module: Health Status” and “2002 Topical Module: Quality of Working Life.” As explained earlier, some of the variables included in the GSS (e.g., AGE, SEX) have

been included in every year of the survey, but other variables have been included only one or two years. The topical modules that appear in the bottom portion of the SDA codebook include variables that are not asked every year. The date of the topical module indicates the first year in which these survey questions were asked. *It does not necessarily mean that these survey questions were asked only in this year, however.* The “2002 Topical Module: Quality of Working Life,” for example, includes the variable *WKHARSEX*, which corresponds to the question “In the last 12 months, were you sexually harassed by anyone while you were on the job?” As we shall see below, this question was asked to a subset of respondents in four years of the GSS: 2002, 2006, 2010, and 2014.

IN WHICH YEARS WAS THIS VARIABLE INCLUDED IN THE GSS?

When “viewing,” “searching,” and “browsing” for variables in the SDA, the results are pooled from across the more than four decades in which the GSS has been administered (1972–present). When analyzing survey data, it is always important for the researcher to understand when the data that she or he is analyzing was collected.

The easiest way to determine the years in which a particular variable was included in the GSS is to make a cross-tab of the variable of interest with the variable YEAR.

To do this, first make sure you are on the “Analysis” tab, which is on the top left of the SDA page, and the “Tables” tab, which is on the right portion of the page. Then, type the name of the variable that you are curious about (e.g., *WKSEXISM*) in the “Row” field, and the variable YEAR in the “Column” field. The resulting cross-tab will show you the frequency distribution for the variable *WKSEXISM* within each year for which there are data (in this case, 2002, 2006, 2010, and 2014). Producing and interpreting cross-tabs is covered at greater length in Chapters 3 and 4.

PRODUCING AND INTERPRETING A FREQUENCY TABLE

In addition to showing the wording of the survey question, “viewing” a variable also produces a frequency table showing the distribution of responses for the variable. Here it is important to remember that, when we examine a variable using the “View” option, we are examining the totality of information about this variable since the beginning of the GSS. In other words, the data that we see when “viewing” a variable are the aggregated data from 1972 to 2014. As explained later in this chapter, in many cases we may want to look at a smaller subset of data, and there are easy ways to do this using the “Filter” option.

Figure 2.5 shows the frequency table for the variable WKSEXISM, which assesses whether respondents feel in any way discriminated against in their workplace because of their gender. Focusing on the middle portion of the table and reading from right to left, the “Label” heading tells us the names of the different categories for the variable. In this case, respondents are asked if they feel discriminated against on the job because of their gender, and they are offered two options: yes, they have felt discriminated against, or no, they haven’t.

Figure 2.5

WKSEXISM R FEELS DISCRIMINATED BECAUSE OF GENDER				
Description of the Variable				
902. Do you feel in any way discriminated against on your job because of your gender?				
Percent	N	Value	Label	
6.2	365	1	YES	
93.8	5,528	2	NO	
	53,632	0	IAP	
	15	8	DONT KNOW	
	59	9	NO ANSWER	
100.0	59,599		Total	
Properties				
Data type:	numeric			
Missing-data codes:	0,8,9			
Mean:	1.94			
Std Dev:	.24			
Record/column:	1/3820			

Missing-Data and “Invalid” Responses: IAP, DK, NA

The values of “Inapplicable” (IAP), “Don’t know” (DK), and “No answer” (NA) are all considered to be invalid responses and are examples of “missing data.” While

analyzing patterns of missing data can be both important and insightful, social scientists generally focus only on data with valid responses and disregard cases with missing data.²

“Don’t know” means that the respondent indicated that she or he wasn’t sure or didn’t know if they felt discriminated against on the basis of their gender, and “No answer” means that the respondent didn’t provide an answer—not even an answer of “I don’t know.” “Inapplicable” means that this question was not included as part of the survey for some respondents. Recall that the “View” screen is showing us an overall picture of the combined data from 1972 to 2014 and that not every question is included every year. In addition, in some years the GSS uses a “split-ballot” design, where some questions are asked to a randomly selected group of respondents but not to others. If respondents were not given the opportunity to provide a response to a question because it was not included in the survey they were administered, then their responses are coded as “IAP.”

The response categories for each variable are always assigned a number. Taken together, the “Value” and “Label” headings show us which response corresponds to which number. So, here, “Yes” corresponds to the number 1 and “No” corresponds to the number 2.³ Missing-data categories are also given numbers: In this case, 0 is IAP, 8 is DK, and 9 is NA.

The heading “N” refers to the number of respondents whose answers fell into each category of the variable. Reading across the top row, we see that 365 respondents indicated that yes, they had perceived gender discrimination on the job, and 5,528 respondents indicated that no, they hadn’t perceived gender discrimination. The majority of responses (53,632) fall into the “IAP” category, suggesting that WKSEXISM is a question that has not been included regularly in the GSS. The total number, 59,599 refers to the number of respondents who took the survey from 1972 to 2014. This number will be the same for any variable you view.

The “Percent” column tells us the percentage of valid responses that fell into each category of the variable. When calculating the percentages, the SDA program takes into consideration only those cases where respondents provided “valid” data. The 6.2% in the top line is calculated by dividing the number of respondents who answered “Yes, I have felt discriminated against at work on the basis of my gender” (N = 365) by the total number of respondents who provided valid data for this question, and then multiplying this number by 100. In this case, a total of 5,893 respondents provided valid answers (365 + 5528 = 5893). The percentage is calculated as follows:

$$(365/5893) * 100 = 6.2\%$$

Taken as a whole, the frequency table tells us that most individuals (93.8%) who answered the survey question responded that they had not felt discriminated against at work on the basis of their gender. But this table likely raises more questions for you than it answers. In what year or years was WKSEXISM included in the survey?

Are these data from 2014, 1972, or somewhere in between? Is this a question that is asked to all respondents, or is it asked only to women? And is it asked only to individuals who are currently working for pay? Finally, what factors might influence the likelihood of individuals perceiving gender discrimination at work? Are men and women equally likely to perceive gender discrimination at work? What about women with various levels of education, women of different racial and ethnic groups, and women working in different occupations? These questions are easily answered with the GSS data and will be addressed in later chapters.

DESCRIBING VARIABLES: LEVELS OF MEASUREMENT

In addition to knowing the precise wording of the survey question and the year in which the data were collected, it is also important to know the level of measurement for each variable. When analyzing data from the GSS, it is useful to consider three levels: nominal, ordinal, and interval-ratio.⁴

Nominal-Level Variables

Nominal-level variables, also known as categorical variables, are variables that have no meaningful order to their categories. With nominal-level variables, it is impossible to speak of “high” or “low” values, because there is no order to the values. The variable categories would make just as much sense if they were completely reordered. Examples include RELIG, which corresponds to respondents’ religious preference; SEX, which corresponds to respondents’ gender; RACEHISP, which is one of many variables describing respondents’ racial/ethnic group; DWELLING, which represents the type of situation in which the respondent lives (e.g., a trailer, an apartment, a single-family house); and MARITAL, which corresponds to respondents’ marital status (i.e., whether they are currently married, widowed, divorced, never married, etc.)

Let’s look more closely at the variable RELIG. By viewing the variable RELIG, you should see a chart identical to that in Figure 2.6 below. Respondents are asked, “What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?” You can also see that, when respondents answer this question, their responses are put into one of 13 categories. A person who answers “Protestant” will be assigned a 1 on this variable, someone who responds “Catholic” will be assigned a 2, someone who answers “Jewish” will be assigned a 3, and so forth.

What makes RELIG a nominal-level variable is that the ordering of the religious categories is arbitrary. The variable would make just as much sense if the categories were put in a different order and each religious group was associated with a different number.

Someone who describes themselves as “Orthodox-Christian” scores a 10 on this variable, and someone who describes themselves as “Catholic” scores a 2, but this does not mean that Orthodox-Christians are five times as religious, or have five times

Figure 2.6

RELIG			
RS RELIGIOUS PREFERENCE			
Description of the Variable			
104. What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?			
Percent	N	Value	Label
58.3	34,596	1	PROTESTANT
24.5	14,532	2	CATHOLIC
2.0	1,195	3	JEWISH
11.2	6,635	4	NONE
1.7	1,025	5	OTHER
0.3	156	6	BUDDHISM
0.1	76	7	HINDUISM
0.1	34	8	OTHER EASTERN
0.2	117	9	MOSLEM/ISLAM
0.2	105	10	ORTHODOX-CHRISTIAN
1.2	723	11	CHRISTIAN
0.0	26	12	NATIVE AMERICAN
0.2	128	13	INTER-NONDENOMINATIONAL
	23	98	DK
	228	99	NA
100.0	59,599		Total
Properties			
Data type:	numeric		
Missing-data codes:	0,98,99		
Mean:	1.90		
Std Dev:	1.67		
Record/columns:	1/480-481		

more religious preference than Catholics. It doesn't even mean that "Orthodox-Christians" have a "higher" level of religiosity than Catholics. While the variable categories themselves have meaning, the ordering of the categories does not.

INTERPRETING A MEAN OR STANDARD DEVIATION FOR NOMINAL-LEVEL DATA

Note that, when viewing a variable in SDA, the bottom portion of the resulting chart provides a mean and standard deviation. For the variable RELIG, for example, the mean is 1.90 and the standard deviation (abbreviated "Std Dev") is 1.67. As explained below, the mean and standard deviation for nominal-level variables is nonsensical. Because the ordering

of the categories as well as the number associated with each category is arbitrary, both the mean and the standard deviation are meaningless. It makes no sense to talk about an average religious preference of 1.90. Even though the SDA reports this mean, you should pay it no attention when analyzing a nominal-level variable.

Dummy Variables

Dummy variables are a special kind of nominal-level variable, where responses are coded into just two categories. The variable SEX (1 = male, 2 = female) is a dummy variable, as is the variable CITIZEN, based on the question "Are you a citizen of America?" (1 = yes, 2 = no). Other examples include XMOVIE, the variable that corresponds to the question "Have you seen an X-rated movie in the last year?" (1 = yes, 2 = no); WKRACISM, "Do you feel in any way discriminated against on your job because of your race or ethnic origin?" (1 = yes, 2 = no); and CONVICTD, "Not counting minor traffic offenses, have you ever been convicted of a crime?" (1 = yes, 2 = no).

Ordinal-Level Variables

Ordinal-level variables have a meaningful order to their response categories. Unlike interval-ratio-level variables (discussed below), however, the number associated with each category doesn't provide a meaningful indication of relative distance between each category. In other words, in ordinal-level variables, the interval between categories is undefined. The size of the categories themselves is often uneven.

The variable DEGREE provides a good example.⁵ This variable represents respondents' highest educational degree and is shown in Figure 2.7. We see that respondents who have less than a high school education ("LT High School") score a 0 on this variable and that those who have a high school degree—but no degree beyond that—score a 1. Those whose highest degree is an associate's degree would score a 2, and those with a bachelor's degree would score a 3.

The categories of this variable clearly have an underlying order to them. Higher scores on this variable represent higher levels of education. But the categories themselves are uneven and the distance between them is undefined. A score of 0 could represent someone with 0 to 11 years of education, but a score of 1 likely represents someone who has between 12 and 13 years of education. Someone who scores a 4 on this variable could have a one-year master's degree, a three-year law degree (JD), or a master's and a doctoral degree.

With ordinal-level variables, caution must be used when making claims about the categories in relation to one another. It would be appropriate to say that those who score a 4 on this variable report having a higher degree than those who score a 2 on this variable. It would not, however, be accurate to say that someone who scores a 4 on this variable has twice as much education, or twice as high a degree, as someone who scores a 2.

Figure 2.7

DEGREE		RS HIGHEST DEGREE	
Description of the Variable			
19. If finished 9th-12th grade: Did you ever get a high school diploma or a GED certificate?			
Percent	N	Value	Label
21.9	12,997	0	LT HIGH SCHOOL
51.4	30,556	1	HIGH SCHOOL
5.5	3,256	2	JUNIOR COLLEGE
14.3	8,474	3	BACHELOR
7.0	4,151	4	GRADUATE
	30	8	DK
	135	9	NA
100.0	59,599		Total
Properties			
Data type:	numeric		
Missing-data codes:	7,8,9		
Mean:	1.33		
Std Dev:	1.17		
Record/column:	1/152		

The GSS contains hundreds of ordinal-level variables. Respondents' personal income (as measured by the variable RINCOM06) and respondents' family income (as measured by the variable INCOME06) are both good examples of ordinal-level variables. Examples also include all of the attitudinal and opinion questions involving Likert scales, such as ETHIGNOR and RESPECT:

ETHIGNOR: Here are some opinions some people have expressed in connection with ethnic issues in the United States. To what extent do you agree or disagree with each one? a. Harmony in the United States is best achieved by down-playing or ignoring ethnic differences. (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree)

RESPECT: Now I'm going to read you a list of statements that might or might not describe your main job. Please tell me whether you strongly agree, agree, disagree, or strongly disagree with each of these statements . . . At the place where I work, I am treated with respect. (1 = strongly agree, 2 = agree, 3 = disagree, 4 = strongly disagree)

The GSS also includes some "feeling thermometer" variables, and these too are measured at the ordinal level. Examples include FEELBLKS ("In general, how warm or cool do you feel towards African Americans?") and FEELWHTS ("In general, how warm or cool do you feel towards white or Caucasian Americans?"). Both variables have nine categories, where scores of 1 indicate feeling "very warm" toward the group and scores of 9 indicate feeling "very cool" toward the group. Respondents who score a 1 on the variable RESPECT strongly agree with the notion that they are treated with respect at work. Higher numeric scores on this variable (such as a 2, 3, or 4) are associated with increased disagreement. Similarly, on the variable FEELBLKS, we can say that low scores indicate respondents who report having "very warm" feelings toward African Americans, and higher scores represent increasingly cool feelings.

Interval-Ratio-level Variables

Like ordinal variables, **interval-ratio variables** have an intrinsic order to the variable response categories. Unlike ordinal variables, however, interval-ratio variables have a defined space between the categories, which is uniform throughout the range of the variable. Variables like AGE (respondent's age in years), HRS1 (the number of hours the respondent worked last week), and SIBS (respondent's number of brothers and siblings) are all interval-ratio variables. Not only are the categories ordered (for example, someone who scores a 4 on the variable SIBS has more siblings than someone who scores a 1), but the interval between the categories is constant: each one-unit increase in the variable SIBS represents an additional sibling. Similarly, in the variable HRS1, each one-unit increase represents an additional hour worked last week.

INTERVAL-RATIO VARIABLES AND CROSS-TABS

Interval-ratio level variables often have a high number of response categories. For example, SIBS ranges from 0 to 68, AGE ranges from 18 to 89 (89 indicates respondents who are 89 and older), and EDUC (years of formal schooling) ranges from 0 to 20. With so

many categories, interval-ratio variables can be difficult to analyze using cross-tabs unless the variables are recoded into a smaller number of categories. Recoding variables is easy in SDA and is discussed in detail in Chapters 5 and 6.

DESCRIBING VARIABLES: MEASURES OF CENTRAL TENDENCY

As discussed above, the first steps in analyzing survey data are (1) determining what the variable of interest to you actually measures (that is, identifying the survey question from which the variable was created) and (2) determining what the variable categories represent (that is, understanding what the response categories are and determining whether the variable is a categorical, ordinal, or interval-ratio level variable). After doing so, the next logical step is to take a look at how respondents actually answered the question. Researchers are often interested to see if there is a particular response or range of responses that are given more frequently than others.

The Mode

The mode of the variable is the response category with the highest number of responses. For example, in Figure 2.7, we see that in the variable DEGREE, the category for “high school degree” has 30,556 responses—more than any other response category. Figure 2.6 provides the distribution for the variable RELIG and shows that 34,596 people described themselves as Protestants. Because the number of people who describe their religious preference as Protestant is greater than the number of people who select any other religion, we can say that the mode of the variable RELIG is Protestant. While the mode can be found for any variable, whether categorical, ordinal, or interval-ratio, it is most meaningful for interpreting categorical-level and ordinal-level variables.

The Median

When a variable is an ordinal or interval-ratio level, it is often useful to determine the median. To find the median for a particular variable, the individual responses first are arranged from greatest to smallest, or smallest to greatest. Once they are so arranged, the median is simply the value of the middle observation. More specifically, in situations where there is an odd number of observations, the median is the value

of the middle observation. When there are an even number of cases and thus no single middle observation, the median is found by taking the average of the two most middle observations. The median also corresponds to the 50th percentile.

Figure 2.8, for example, shows the distribution for the variable IMMCULT, which was first asked in the 2014 survey. The variable corresponds to a survey question that asks, “How much do you agree or disagree with the following statements? American culture is generally undermined by immigrants.” In total, 1,207 (36 + 184 + 267 + 606 + 114 = 1207) people provided valid responses to this question. Imagine asking these 1,207 people to line themselves up so that people who strongly agree that American culture is generally undermined by immigrants were on the left and people who strongly disagreed with this sentiment were on the right. Because the people were lined up in order, we would know that, moving from left to right, we would first see 36 people who “strongly agree,” then 184 people who “agree,” and then 267 people who “neither agree nor disagree.” Taken together, there would be 487 people in these first three categories. The next 606 people would be people who responded “disagree,” and finally there would be 114 people who “strongly disagree.” The person standing in the exact center of this line would have 603 people on her or his left and 603 people on her or his right. The middle person, the 604th person, is someone who “disagrees” with this survey question.

Figure 2.8			
IMMCULT			
IMMIGRANTS UNDERMINE AMERICAN CULTURE			
Description of the Variable			
How much do you agree or disagree with the following statements? B. American culture is generally undermined by immigrants.			
Percent	N	Value	Label
3.0	36	1	Agree Strongly
15.2	184	2	Agree
22.1	267	3	Neither Agree nor Disagree
50.2	606	4	Disagree
9.4	114	5	Strongly disagree
	58,325	0	IAP
	64	8	Don't know
	3	9	No answer
100.0	59,599		Total

Because an underlying order to the response categories is necessary to identify the median, it is not possible to calculate a meaningful median for categorical-level data. Again, for categorical data, the best measure of central tendency is the mode.

The Mean

The mean of a variable is the arithmetic average, calculated by adding up all the scores and dividing that sum by the number of observations. Consider the example below, which examines a random sample of 10 respondents from the 2014 GSS, and their responses for the variable COLSCINM. The survey first asks respondents, “Have you ever taken any college-level science courses?” Those who respond that they *have* taken a college-level science course are then asked, “How many college-level science courses have you taken?” COLSCINM is an interval-ratio level variable with responses ranging from 1 to 90 (those who have taken no courses are not asked the question).

Figure 2.9

	COLSCINM: How many college-level science courses have you taken?
Respondent #1	1
Respondent #2	2
Respondent #3	3
Respondent #4	2
Respondent #5	8
Respondent #6	4
Respondent #7	2
Respondent #8	1
Respondent #9	2
Respondent #10	1

To calculate the mean for this variable, using this small sample of 10 respondents, simply add up the individual responses and then divide by the number of observations (which in this case is 10):

$$\text{Mean number of science courses} = \frac{1 + 2 + 3 + 2 + 8 + 4 + 2 + 1 + 2 + 1}{10} = 2.6$$

To interpret this number, we can say that in our sample, which includes only those who have taken at least one college-level science course, the mean number of courses

taken is 2.6. The modal number of science courses is 2, since 2 is the number of science courses that occurs most frequently among this sample of respondents.

The median number of science courses from this small sample is calculated by ordering the responses from smallest to largest (ordering them largest to smallest will also work) and by then identifying the most central observation. When put into order, the responses are:

1, 1, 1, 2, 2, 2, 2, 3, 4, 8

With an even number of responses, there is no one single central value. In this case, the fifth and sixth responses represent the middle-most values. Since they are both 2, we can say that the median is 2. If the values were different from each other, the median would be calculated by taking the average of the fifth and sixth responses.

PRODUCING THE MEAN, MEDIAN, AND MODE IN SDA

Using the SDA, it is easy to find the mean, median, and modes for any variable. On the “Analysis” page, simply type the variable name into the “Row” field. Underneath, click on “Output Options,” and under “Other Options” click the box for “Summary Statistics.” When you run the table, a new window will open that shows you the frequency table for your variable, and beneath that will be a table of summary statistics, as shown in Figure 2.10.

Figure 2.10 Summary Statistics for COLSCINM

Mean =6.07	Std Dev =8.85	Coef var =1.46
Median =3.00	Variance =78.29	Min =1.00
Mode =2.00	Skewness =3.50	Max =90.00
Sum =15,303.67	Kurtosis =15.69	Range =89.00

When examining means, medians, and modes in this way, the default in SDA is to provide these measures for all cases, for all years in which the question was asked. If you would prefer to see the mean, median, and mode for a subgroup of cases, then you can use the “Filter” field to restrict the cases included in the analysis. By typing YEAR (2010–2014) in the “Filter” field and then running the table, the median, mean, and mode presented will be calculated by including only those cases included in the 2010, 2012, and 2014 surveys. More information on using the “Filter” option is included in Chapters 4 and 6.

Note: SDA and other computer programs like SPSS and STATA will often provide a median and mean for categorical-level data, but it is important for you, the

researcher, to be able to determine when this value is meaningful and when it isn't. Because the variable for the number of science courses taken is an interval-ratio level variable, the mean, the median, and the mode all have meaning. Notice in Figure 2.6, however, that SDA provides the mean for the variable RELIG—a categorical-level variable assessing respondents' preference. The mean provided by SDA is 1.90, but since the numbers assigned to each religious preference are entirely arbitrary and there is no underlying order to the response categories, this number is completely meaningless!

MEASURES OF CENTRAL TENDENCY AND MISSING DATA

Missing-data and "invalid" responses such as "Don't know," "No answer," and "Inapplicable" should not be included when calculating

means, medians, and modes. The SDA program will automatically exclude these values.

CONCLUSION

After reading this chapter, you should have a sense of how the General Social Survey has been conducted, including who has been eligible for inclusion in the sample, and the broader population to which the results from analysis of the GSS can be generalized. The GSS website contains a much more in-depth discussion of survey design as well as of how and why this design has changed over time.

This chapter also served as an introduction to the Survey Documentation and Analysis (SDA) website. SDA provides an easy way to identify and analyze data from the GSS. After reading this chapter, you should have a good sense of how to search and browse for variables using the SDA website. You should also feel comfortable viewing variables to determine the precise wording of individual survey questions.

Finally, this chapter provided an introduction to describing variables. Before analyzing any variable, it is crucial to determine its level of measurement: is it categorical, ordinal, or interval-ratio? It is also important to consider the number of valid responses for each variable and the extent of missing data. After determining the variable's level of measurement and examining the extent of missing data, it is useful to examine the central tendency of the variable: the modal category if the variable is nominal level, or the median and mean if the variable is ordinal or interval-ratio level.

The following chapters use the concepts and techniques presented in this chapter, along with the concepts presented in Chapter 1, to analyze a range of inequalities. So

far, we have focused on univariate statistics—that is, statistics about a single variable. The next several chapters focus on assessing the relationship between two variables—bivariate analyses—using cross-tabulations and comparisons of means. In addition, these chapters introduce filters and control variables, with which it is possible to show even more complex relationships and to highlight the intersection of multiple inequalities.

EXERCISES

1. View the variable CLASS. What, precisely, does this variable assess?
 - a. respondent's perception of herself or himself as a member of the lower class, the working class, the middle class, or the upper class
 - b. respondent's occupational status
 - c. respondent's level of wealth
 - d. respondent's beliefs about class inequality
2. View the variable SPEDUC. What, precisely, does this variable assess?
 - a. respondent's formal education, measured in years
 - b. respondent's formal education, measured in terms of highest degree earned
 - c. respondent's spouse's formal education, measured in years
 - d. respondent's spouse's formal education, measured in terms of highest degree earned
3. View the variable SPPOORKD. In what year was this variable included in the GSS?
 - a. 1980
 - b. 1990
 - c. 2000
 - d. 2010
4. View the variable UNEMP. This variable is best described as:
 - a. a dummy variable.
 - b. an ordinal-level variable.
 - c. an interval-ratio-level variable.
5. View the variable SPANKING. This variable is best described as:
 - a. a nominal-level variable.
 - b. a dummy variable.
 - c. an ordinal-level variable.
 - d. an interval-ratio-level variable.
6. View the variable ETHNIC. This variable is best described as:
 - a. a nominal-level variable.
 - b. a dummy variable.
 - c. an ordinal-level variable.
 - d. an interval-ratio-level variable.
7. View the variable NUMPROBS. This variable is best described as:
 - a. a nominal-level variable.
 - b. a dummy variable.
 - c. an ordinal-level variable.
 - d. an interval-ratio-level variable.
8. In what year was the variable NUMPROBS included in the GSS?
 - a. 2012
 - b. 2002
 - c. 1992
 - d. 1982
9. Using the variable NUMPROBS, approximately what percentage of respondents reported having no close friends?
 - a. 4
 - b. 8

- c. 12
 - d. 25
10. Using the summary statistics for the variable NUMPROBS, what is the mean number of close friends that respondents report having? Be sure to keep the default of COMPWT in the “Weight” field.
- a. 8.17
 - b. 11.01
 - c. 5
 - d. 2

ANALYSES & ESSAYS

1. Identify two variables related to inequalities of race or ethnicity. For each variable, identify the precise wording of the survey question, the level of measurement, and the response categories. How might analyzing these variables help to advance a social justice project?
 2. Identify two variables related to inequalities of gender or sexuality. For each variable, identify the precise wording of the survey question, the level of measurement, and the response categories. How might analyzing these variables help to advance a social justice project?
- For questions 3, 4, and 5 below, choose a specific social justice issue that is important to you and find three variables in the GSS that are related to this issue.
3. View each of these variables (separately) and identify the precise wording of the survey question, the level of measurement, and the response categories. Then, describe the resulting univariate frequency tables, including the extent of missing data.
 4. For each of these variables, identify the most appropriate measure of central tendency. Explain why you chose the measure of central tendency you did.
 5. Describe how an analysis of these three variables could contribute to a better understanding of the social justice issue you have chosen.

NOTES

1. Much of this information comes from the GSS website (<http://www3.norc.org/GSS+Website/>) and from the FAQs about the GSS (<http://www3.norc.org/GSS+Website/FAQs/>). More detailed information, including a complete description of the survey methodology, can be found here as well.
2. One major exception here is when survey questions have a high percentage of “Don’t know” or “No answer” responses. Particularly when there are patterns to the missing data, this can result in biased findings. A classic example here concerns missing data related to income. If people with high levels of education, for example, are less likely than people with low levels of education to provide information about their income, then analyses that simply disregard cases with missing values on education will result in an analytic sample that is biased with respect to both education and income. In other words, the conclusions that we draw from our analyses could be incorrect.
3. Because the variable WKSEXISM is nominal level (also called a categorical variable), the decision to assign the value of 1 to “Yes” and 2 to “No” is completely arbitrary.
4. In using the term “interval-ratio” level variables, I am drawing from Frankfort-Nachmias and Leon-Guerrero (2015), who include “interval”- and “ratio”-level variables as a single category

for purposes of learning the basics of social statistics. Frankfort-Nachmias, Chava, and Anna Leon-Guerrero. 2015. *Social Statistics for a Diverse Society*. 7th ed. Thousand Oaks, CA: SAGE.

5. As discussed in Chapter 5, “viewing” in SDA almost always shows the precise wording of the survey question that was asked to respondents. DEGREE is one of the exceptions. In the survey,

respondents are asked a number of questions about their educational attainment, and their answers provide information that is later combined into the variable DEGREE. In other words, while viewing DEGREE seems to suggest that this variable is associated with the question “If finished 9th–12th grade: Did you ever get a high school diploma or a GED certificate?” the variable draws from multiple survey questions.

Draft Proof - Do not copy, post, or distribute

Draft Proof - Do not copy, post, or distribute